Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

Aletta Tordai

# A case study of reporting business KPIs of gaming data through static and interactive visualizations

Master's Thesis
Espoo, September 20, 2021

Supervisor: Professor Tapio Takala, Aalto University
Advisor: Professor Tapio Takala

Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

ABSTRACT OF MASTER'S THESIS

**Author:** Aletta Tordai

**Title:**
A case study of reporting business KPIs of gaming data through static and interactive visualizations

| | | | |
|---|---|---|---|
| **Date:** | September 20, 2021 | **Pages:** | 108 |
| **Major:** | Master's Programme in ICT Innovation Data Science | **Code:** | SCI3095 |

**Supervisor:** Professor Tapio Takala
**Advisor:** Professor Tapio Takala

Gaming companies, especially due to the high volume of ingame event logs, generate terabytes of data on a daily basis, which not only need to be distributed in fast and reliable access but also need to be aggregated and visualized in a timely and useful way to help the production engineers make data-driven improvements and bring business decisions. As such a high volume of data is created every day with an astonishing speed, it becomes highly challenging to process, aggregate and showcases the data in a supporting manner for business decision making. For that reason, business intelligence approaches like visualizations, in order to explore and analyse game-related data, need to find an optimal way of capturing the most out of such diverse and voluminous dataset both from a data perspective as well as from the point of view of the end-user. This, in reality, is a rather complex and many times a seemingly irrelevant job for companies that need to keep up anyhow with the daily income of the terabytes. As a result, companies not only end up having Key Performance Indicators visualized in the simplest forms of tables but also face the burden to store more data than needed in expensive data warehouses.

In this thesis, we are going to propose a guideline for a business information visualization pipeline based on a practical use-case of a real-world gaming company. We present a pipeline that is not only built on top of the available data but also takes into consideration the cognitive processes involved in finding answers in visualizations and follows a task-driven structure. Furthermore, we are going to test which combination of view management and coordination helps better to deliver more accurate responses and a more acceptable system for decision making.

The purpose is not only to create effective visualization and provide a guideline for a pipeline but to find bottlenecks and points of failure that have an indirect effect on the efficiency of the outcome. It is critical to search for latent points of failures to improve upon pipelines that are near to a highly efficient and effective visualization that serves business intelligence with high turnover.

**Keywords:** gaming, business intelligence, business information visualization, KPI, dashboard, interactive, static, visualization pipeline

**Language:** English

# Abbreviations and Acronyms

| | |
|---|---|
| **General** | |
| KPI | Key Performance Indicator |
| BI | Business Intelligence |
| IQ | Intelligence Quotient |
| ETL | Extract Transform Load is a data pipeline for collecting data from various sources, transforming it according to business requirements, and loading it into a destination data storage space |
| EVT | Efficiently View Traversable |
| **Game related** | |
| Churn | A user who was active at a point in time but stopped playing and became inactive for at least the last 30 days |
| Retention | A measure of how the number of people that are still playing after a certain period of time from their installation/registration day |
| Ret 1d/Ret 3d/Ret 4d/Ret 14d/Ret 30d | Retention of day 1/3/7/14/30 in percentage |
| Cohort | A set of users that register/install on the same day, or over the same period of time |
| LTV | Life Time Value of user |

# Contents

# Chapter 1

# Introduction

The fast pace of change that the economy dictates to companies puts an external pressure on all businesses to perform better. In the performance rally against each other (and themselves) businesses rely on the data they generate as an in-house resource for understanding their current performance, shortcomings, and to also learn ways for improvement. They coin data as the "currency of the 21st century enterprise" [10].

Enterprises take their data very serious. They understand that their data does not only bring value in terms of financial sheets and reporting for bureaucratic duties but actually show what their services and product do good and where do they need more careful attention. Thus, leveraging statistics and performance indicators, businesses nowadays rely heavily on the reporting of key performance indicators for internal operations .

Businesses choose visualizing their data as a way to communicate because they understand the way human brain works. Moreover, they are aware of the importance of correct perception of performance indicators. A study showed that human brains can process visuals 60,000 times faster compared to text [6]. Another study underlined the effectiveness of visualized data in presenting concrete scientific information where the textually communicated claim received acceptance from 68% of the people whereas when a graph was added this acceptance rose to 97%[9].

It is of no surprise that companies prefer to have their data visualized in terms of performance indicator visualizations. However, with the revolution of data, companies are also facing the difficulties of making sense out of a high volume of data while rendering that into a fruitful product that has the potential to offer valuable insight and guide to further operations. Along these difficulties lies the lack of evaluating such visualizations in real case situations which in many cases also leads to a hard line of acceptance of visualization approaches across different teams within companies.

## 1.1 Problem statement

There is the general problem of technology advancing faster than human adaptability, which in the business data visualization domain gets accentuated as it does not only revolve around the scientific maturity of data visualizations and guidelines of best-practices but also highly involves the analyzer to eventually perform a number of meaningful actions in order to retrieve answers and valuable insights.

This human factor introduces on one side leverage to the development of the visualization tool, but it also has as side-effect the error prone quality of it as conclusions and arguments are no longer on the developed product's features.

Furthermore, practical observation based on the willingness of non-technical users of business visualizations show towards interacting with tools led us to wonder whether the type of construction one visualization can have carries significant influence in the efficacy and performance of the final tool. Understanding whether simple, report-like tools are more preferred in wider scale in businesses over interactive task-driven and human-supervised dashboards became a main question that steered the work. Based on initial experiences at the company of authors internship questions have arisen about whether data-driven interactive visualization tools are feared (because they are more likely to cause confusion and leave teams without answering business questions), and if this fear can be avoided by a task focused design that also delivers acceptability and usability on system and business level.

Thus, we came to question **how well could one data visualization pipeline be constructed in terms of *accuracy* and *task efficiency* in a way that does not only meet the business criteria but also understands the end users from the acceptability principle of *operational utility* and *operational effectiveness***.

The acceptability factor in many cases suffers or it becomes forgotten in the amount of requirements when the focus is on creating a visual solution that aims for accurate answers and strives for efficiency on cognitive tasks.

The greatest challenges in the visualization domain are defined by *operational utility* and *operational effectiveness.* One one hand, operational utility refers to the visualization being the right option to use to accomplish the business goal [2]. Such business goal is for example creating an unbiased data support for answering business questions. On another hand, operational effectiveness refers to the visualizations ability to deliver a support system well and accurate[2].

Another great challenge and problem is of the high dimensions in big data visualizations. Understanding how and which dimensions can be used as filters and interaction handles, an how these can have effect on a data visualization tool as a whole becomes hard to outline with the growth of the number of dimensions.

Moreover, visualization tools as business support tools also face problems when they are handed to users. Especially non-technical background users might have a harder time and more second thoughts when it comes to exploring data through data visualization tools and using it as support for arguments and decisions. These second thoughts can eventually lead to two directions with totally opposite outcomes. It can be that dashboards become forgotten under old employees personal credentials or it can eventually be the opposite and be a great work which will be used in business intelligence meeting to demonstrate findings and support arguments. As desired, we also sought at creating a data visualization pipeline that leads to the latter scenario of successful visualization tools with added value for the business.

However, with the binary quality of average visualization outcomes we formulated as ground hypothesis for our work. This, apart from aiming to deliver a constructive work by proposing a solution for a case study given by the company of authors internship, also strives to find out whether introducing interaction and stepping up from a general overview-to-detail styled static visual reporting can still perform on business questions and be accepted. Thus we formulated 2 key hypotheses based on these problems and on our question.

**Hypothesis 1** *Well-designed static dashboards with limited interaction possibilities and with view coordination limited to the necessities of tasks deliver more accurate answers for questions than the equivalent interactive version that offers a wide range of possibilities to filter, highlight and explore in depths.*

**Hypothesis 2** *Interactive dashboards are more accepted by a wider variety of users regardless of the outcome of the cognitive tasks.*

Therefore, understanding a specific company's needs, taking them as input scenarios, our motivation is to create in our work a data visualization pipeline that gives answers to business questions and tests our hypotheses about acceptability along task-driven performance. This would ultimately lead to a clear idea about what design choices should the company of the use case take into account when addressing Key Performance Indicator (KPI) based business decisions through visualizations.

## 1.2 Structure of the Thesis

This thesis is structured along the way the constructive work has been carried out. After setting the technical background, the proposed solution and methods are presented in consecutive chapters. These are then followed by the implementation and the evaluation of the implemented tool, which leads to the discussion on limitations of the system and further improvements.

Therefore, in Chapter 2 we focus on setting the thesis work in a clear technical background, both in terms of business context as well from the point of view of existing technical solutions and state-of-the-art of literature.

Next, in Chapter 3 a proposal is presented along with the study context. Thus, based on the observations and shortcomings noticed along the time of cooperation with the company a proposal of how the case study could be solved to deliver the desired business results is presented.

Following, Chapter 4 presents the methods that we implemented as well as the means through they were used. Then, Chapter 5 shows the implemented tools along the steps from the proposal using the methods previously described. This is followed by Chapter 6 which expands the used evaluation metrics and techniques as well as the performance of both of our visualizations proposed.

Lastly, in Chapter 7 we discuss about limitations and further improvement possibilities that could enable KPI visualization to be applicable on wider-scale.

# Chapter 2

# Background

## 2.1 Literature

It is, for the purpose of setting the common ground for our work, essential to be clear with the definitions, existing methods and techniques, current state of similar works. Thus, in the coming sections we aim to cover the business related value of our main element of work: the Key Performance Indicators, then to present the most relevant components and concepts in the information visualization domain. After touching base with the most essential concepts we present the contextual works related to data visualization especially the state-of-the-art proposals.

### 2.1.1 KPIs, metrics and measures

Businesses, regardless of their size, have been using for decades metrics to track their performance and their capabilitis towards achivieng their goals. For this purpose, they define a set of Key Performance Indicators (KPI) to use when tracking metrics of their business processes. Businesses rely on these KPIs, since it does not only offer an easy communication method to tell potential clients and competitors how well they are doing, but also it enables the companies to figure out ways to improve, serving an important role in internal decision making. Thus, companies not only measure KPIs, but also rely on visualizations of their KPIs for a better understanding.

Thus, based on the Oxford's Dictionary definition and the article of Humans of Data on KPI visualization, we define a KPI as a parameter of an organization that is measurable and serves as proof about the level of achievement of an intended target [3].

Therefore, it is highly relevant to be clear with the notion of KPIs, the current state-of-the-art status of KPI visualizations and the difference between measurable metrics and KPIs. Moreover, due to mandatory handling of financial chores within companies, KPIs are often produced and showcased through KPI reports, hence the need of clear rules to distinguish between KPI reports and KPI visualizations.

To start with, the definition of metric and measure needs to be cleared out as, especially within businesses, they are used interchangeably. In a bottom-up approach we can start with defining ***measure*** as the number value that can be aggregated: summed or/and averaged; and is the most fundamental and unit-specific term within business context. Whereas, a ***metric*** is quantifiable measure that is used to keep track and evaluate a specific process within the organization. Therefore, establishing the most important metrics within the specific organization as a key performance indicator enables to discover trends over time.

It is clear that depending on the organization, the KPIs will vary as they define the market the organization is operating in. It is important to understand as a starting point that, however the wide variety of measures in a company's dataset, and regardless the amount of relevant metrics being available to track, KPIs will only be the ones that are ”key” to the business. A metric would only translate into a KPI only when it can lead to clear decisions and actions that can help the organization achieve its goals.

KPIs are mostly presented through visualizations that are meant to serve 3 main objectives:

1. Project management
2. Investment management
3. Monitoring

These objectives are in line with the purpose of KPI tracking to improve business performance, measure the effectiveness of policies and decisions and enhance processes within the organization. Within such, quantitative, qualitative and process KPIs can serve the organizations processes.

To understand better, we can differentiate 3 types of KPIs that will serve as main components to our work:

1. **Quantitative**: these are the ones that help track measurable progress, them being measurable numeric metrics. For example we can think

of project that requests on a country-wide level higher education institutions to implement COVID safety measures that are defined in multiple steps. In this case, to track the progress of schools working on this project we can quantify the number of schools that implemented step one of placing masks at the buildings' entries and then communicate the number of such schools out of the total number of schools. This would give a clearly progress of step one on a country wide level.

2. **Qualitative**: as sometimes what we want to track is not as tangible as the number of schools complying with measures as in the previous example, therefore, the qualitative KPIs therefore comes in as more subjective metric, for instances in the previous example, there could also be that teachers are asked about how satisfied they are about how the process of implementing the required tools for becoming COVID-safety compliant is going. Therefore, in such scenario, the perception of teachers could be tracked with surveys and questionnaires. However great this may look like, there is still need to quantify somewhat the results from questionnaires, otherwise, the communication of such KPIs may fall victim of regrettable miscommunication. For this reason, qualitative data can be quantified with many various tools, such as the Likert scale [30].

3. **Process**: there are many processes within organization that are linked one after the other and as a whole can be measured too. However, getting an overall image of progress is not the same as getting a detailed image. For example, the project of schools becoming COVID-safe, there are multiple steps like placing masks out in front of entrances, then moving the desks and chairs 2 meters apart, then ensuring the right ventilation and as final step it could be to provide students with free quick tests. In such process, we can take every step individually and quantify the number of schools that already implemented them, but that will only tell about the projects progress amongst schools, and not about the processes quality. It can be the case that the steps chained after one another were planned by someone very optimistic who believed every school can do it all equally, however the steps might be in a wrong order, or simply, there can be a step that could cause bottleneck in the whole process. For such, a funnel view makes a lot of sense, to see which steps are easy to do and where schools are maybe dropping out of the project. Like this, it might be found that some schools cannot ensure the ventilation according to the guidelines and they might need extra help. Tracking thus the process KPI would then

allow in the end to get more schools finish with the project by also understanding the needs of schools per steps.

The field of Business Intelligence (BI) combines data analytics with data mining and data visualization as well as with data tools that enable for an infrastructure that provides support for data-driven decisions for a business [1]. There is a collection of different processes and tasks that are understood under the term of business intelligence which cover from the first phase of understanding the business' data through data mining and statistical analysis, through reporting of the analysis, comparing the performance to historical data, to querying the database for answering specific questions, to preparing and aggregating raw data to be used for higher level of representation to finally visualizing it.

### 2.1.2 Types of visualizations

In general, we refer to it as visualization, however, literature uses many related terms covering the wide use and relevance of it, such that data visualization, scientific visualization, information visualizations, visual analytics, and business visualization appear to be used for what we will mostly only refer to it as visualization in our work.

The field of data visualization became prominent along with computer graphics in the 1950s [23] being defined as the science of visual representation of data. Initially, they used the term scientific visualization when referring to visualization produced by a process of scientific computing that focused on showing hidden details of data as well as driving it in a way that it enriches other existing scientific methods [34].

The visualization field comprises of a plethora of different types that are all based on general visualization techniques. Namely

**T.1** information visualization

**T.2** scientific visualization

**T.3** business information visualization

**T.4** simulation

**T.5** illustration

are the main types that differ from each other in terms of the content they aim at presenting, the tools they are built with and the sole purpose of their existence.

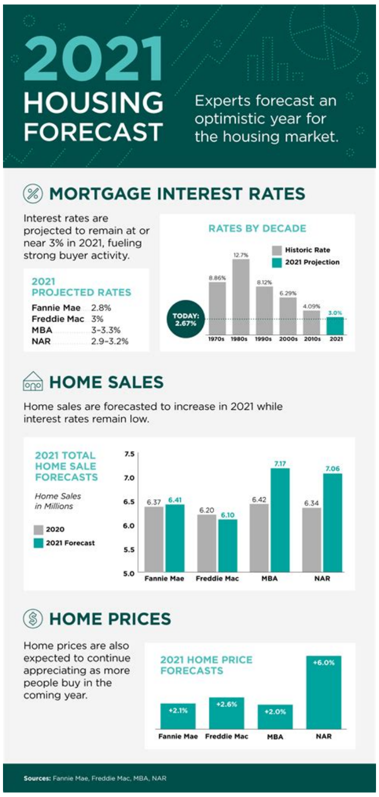


Figure 2.1: **T.1** Infographic about the housing marker[1].

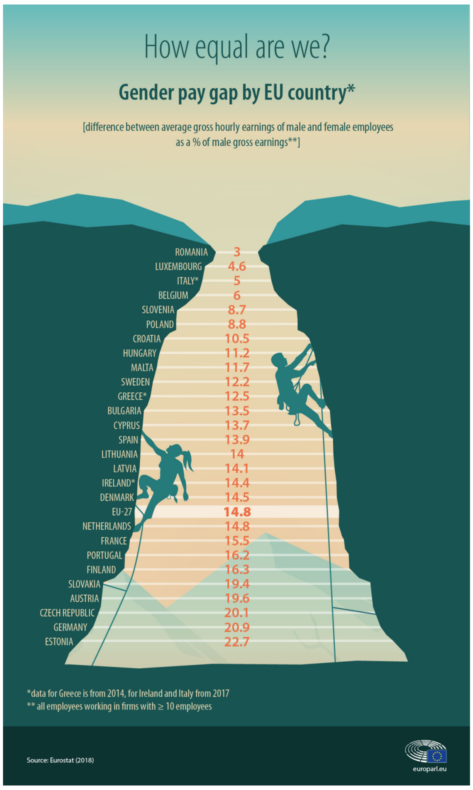


Figure 2.2: **T.1** Infographic about the gender pay gap[2].

In principle, **T.1** is also often used interchangeably with data visualisation when the data used belongs to a more generic content rather than to the business domain [18]. Despite the numerous similar techniques used, principles and features the two share, there is huge imbalance when it comes to whether it is more qualitative over quantitative the information, let alone the nature of the data structures. Therefore, when we are talking about information visualisation we are mostly talking about highly unstructured data, carrying more qualitative value, which for example can be part of an information about workflows, ideas or concepts. As a result the visual representation of it is also more free styled, leveraging the power of infographics and artistic illustration diagrams. Moreover, the reason of existence of them is also just

[1] https://images.app.goo.gl/A4W3S2NWFgzYYVkQA
[2] https://images.app.goo.gl/A4W3hS2NWFgzYYVkQA

as free styled as the tools used, as these are mainly used in casual communications and story telling. For example, a typical information visualization is the infographic with large font-size and long vertical orientation, containing key message as easily as possible and communicating it vividly for a wide audience in a highly engaging manner [13]. Figure 2.1 and Figure 2.2 are examples of such information visualizations.

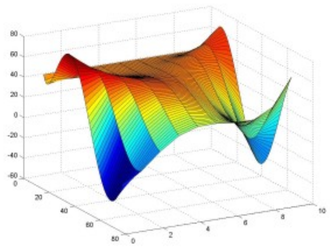


Figure 2.3: **T.2** Scientific visualization of a function[3].

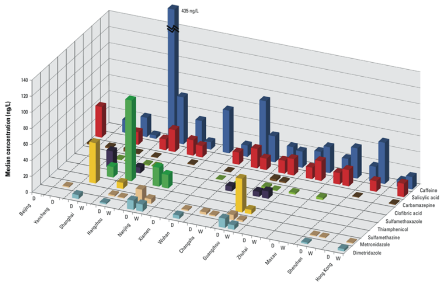


Figure 2.4: **T.2** Scientific visualization 3D barplots[4].

**T.2** on the other hand, is way more rigid and strict in terms that it conveys the visualizations of real-word objects or phenomenons as well as mathematical functions and formulas. Therefore these are visualized and presented through computer generated graphical elements and in 3D simulated virtual realities. For instances these can be computer generated replicas of such scientific phenomenons like meteorological events, medical items and shapes or architectural renderings. The main focus in scientific visualization is the realistic rendering of volumes and the vivid lighting and illumination of such volumes surfaces. Hence, the result is often just a replica of real world items in the computer graphics space, as well as well justified and data-supported imaginary creations as variations to the real-world objects. As an example, we can think about about a human brain being visualized in 3D space, volume rendered and color annotated per different sections that can help doctors and medical crew to visualize different impulses' and reactions' activation spots in a graphical environment. Figure 2.3 and Figure 2.4 are examples of such scientific visualizations.

**T.3**, having a complex purpose of delivering exploratory, analytical and decision supporting tasks uses more quantitative data, relying on metrics and key performance indicators. The main challenge the domain faces is represented by the variety of possible business domains and the increasing an volatility

[3] https://images.app.goo.gl/mGy3Rjg7htajCJJg6

[4] https://images.app.goo.gl/Akk68gfZestMr9Zp8

nature of data that is being used, however the quantitative quality of it balances some challenges out. Business information visualization, on the other hand relies on such basic and key elements of visualisations as charts and diagrams, but the essence of the comes in the context they are prepared in, more exactly the dashboard they are presented in. Figure 2.5 and Figure 2.6 are examples of such business information visualizations.

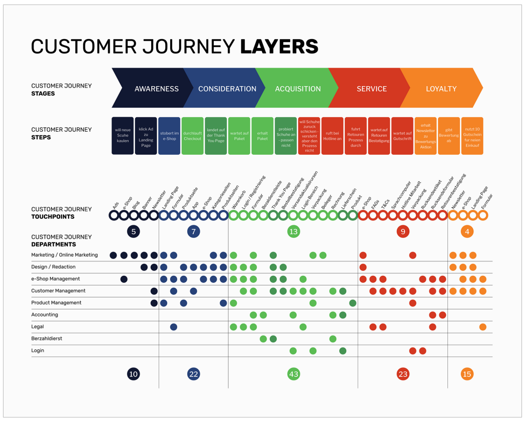


Figure 2.5: **T.3** Business information visualization about customer journey[5].

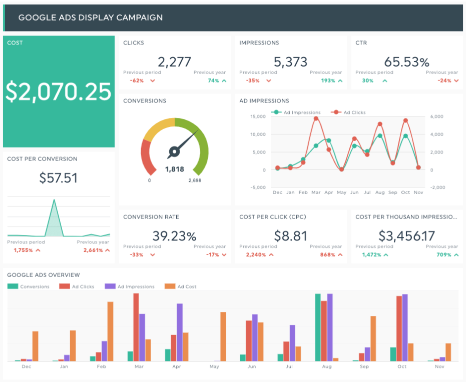


Figure 2.6: **T.3** Business information visualization about ad campaign performance[6].

https://images.app.goo.gl/tHTCwhyheQUCAY1z6

**T.4** relying heavily on computer graphics to generate real-world like scenarios it is closer to scientific visualisation compared to any other type, however in this case not a static replica is aimed to be visualised but rather a motion. For example, such a motion could be the flow of clouds where the volumes would be rendered constantly according to the changes of the volume's parameters. But a simulation doesn't necessarily have to be about heavy volume rendering, it can also be a simulation for example of urban traffic with the aim of optimizing traffic light scheduling. In this case the simulation still has parameters on how to be rendered, but the items can be visualized with less graphical resources as the shapes can be more regular and simplistic, compared to an irregular cloud shape. Figure 2.7 and Figure 2.8 are examples of such business information visualizations.

Lastly, **T.5** slightly differs from visualisation as a term, however the purpose of it is similar, explaining complex information, ideas with the help of the

[5] https://images.app.goo.gl/ufoWocK8uaRSBSPy5
[6] https://images.app.goo.gl/YJkYBQwvAhwzQWJv7

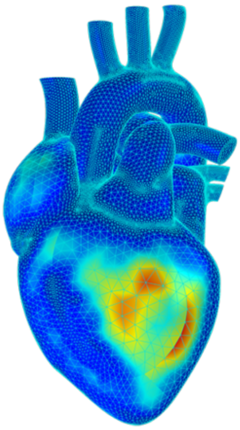


Figure 2.7: **T.4** Simulation of a heart with heatmap[7].

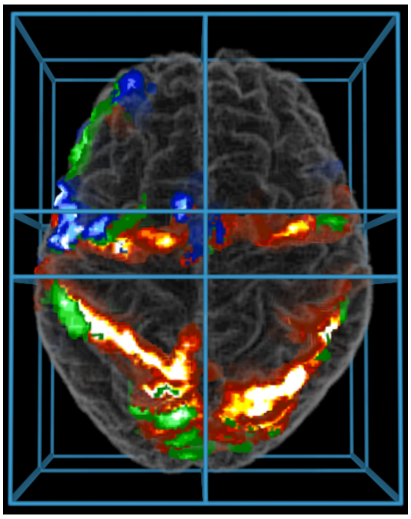


Figure 2.8: **T.4** Volume rendering of a brain with color-marked annotations[8].

materialization of abstract ideas through easily perceivable images or over-simplified diagrams. In a business context, illustrations are rather driven by the aim of explaining a concept than driven by data.

### 2.1.3 Visualization techniques

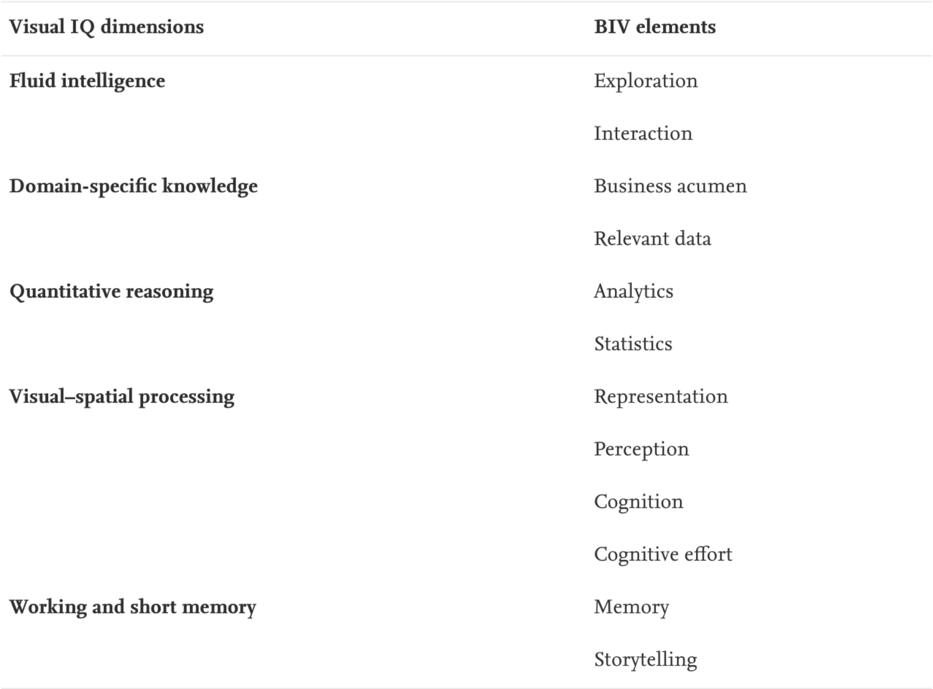
In early literature, a large amount of research indicates that a link between decision support technologies and human intelligence is rather important in delivering final visualization solution [16], [33], [32]. The goal of visualizations should always be to "help the decision maker achieve cognitive effectiveness and efficiency by shortening cognitive distance from visual representations and removing mediation for thinking" [34].

Therefore, Bačić et. al. propose a formal human intelligence-based map to business intelligence visualization focused business intelligence capabilities [5]. Their proposed framework leverages the Stanford-Binet Intelligence Quotient test as to measure human intelligence which includes the testing of nonverbal content too. Therefore, their work relies heavily on this mapping between the nonverbal IQ and the business intelligence visualization's components. Their framework, thus, relies on the mapping seen in Figure 2.9.

[7] https://images.app.goo.gl/ASwnShhph5dsAXJt6

[8] https://images.app.goo.gl/tHTCwhyheQUCAY1z6

Moreover, this framework set core ideas for our work by outlining that the business intelligence visual components of cognition, representation, perception and cognitive effort are tightly related to the visual spatial processing of humans, which naturally, highly varies between individuals. This work solidified in our process the need for a focus on how individuals discover and navigate in the visual-spatial dimension.



| Visual IQ dimensions | BIV elements |
|---|---|
| **Fluid intelligence** | Exploration |
| | Interaction |
| **Domain-specific knowledge** | Business acumen |
| | Relevant data |
| **Quantitative reasoning** | Analytics |
| | Statistics |
| **Visual–spatial processing** | Representation |
| | Perception |
| | Cognition |
| | Cognitive effort |
| **Working and short memory** | Memory |
| | Storytelling |

Figure 2.9: Visual IQ-based BIV element

According to Bačić et al., it would not keep with integrity to evaluate the performance of visual solutions, in terms of driver to decision making, while isolating the business intelligence visualization's components, neglecting the visual IQ [5]. Thus, the exploratory and interactive capabilities are linked to the fluid intelligence and they refer on top level to the ability of recognizing patterns [5]. Due to the complexity and uncertainty carried by the information that business intelligence developers face, it is common tasks to conceptualize and to understand the relationships within the data. For this a strong sense of pattern recognition is needed, and hence, to solve problems BI developers and users have to perform mental operations that are associated with the fluid intelligence. On higher levels of decision making fluid reasoning is extremely challenging without a properly designed visualization. This is the case especially when the data is multi-dimensional, however, visualization techniques can support the end-user in solving fluid analysis intense exploratory tasks [5].

The same mapping according to Bačić et al. links the domain-specific knowledge to business acumen and analytics, which is tied to the relevance of meaningful reduction in data dimension and complexity while maintaining a significant quality of possible new hypothesis. [5]. Without this key element, according to Bačić, nor the designer of the visualization nor the end-user would have the base knowledge to interact with the visualization in terms of understanding the terminology, using the correct filters. Hence the evaluation of the effectiveness of the resulting visualization also drops when deployed with an inaccurate or inappropriate data.

Furthermore, the link between quantitative reasoning intelligence and statistical and analytical capabilities in essence mean the gathered mathematical background knowledge and the ability for using key statistical metrics for reasoning and concluding [5].

Hence, as last pillar for his framework, Bačić et al. takes visual-spatial processing intelligence and links it to representation and informed design based on visual perception, cognition, and cognitive effort. This visual-spatial processing refers to the set of capabilities that revolve around information generation, storing, extraction and transformation. Thus, in order to enhance the visual–spatial processing ability of the end-user and consecutively to better drive the decision making based on the business intelligence visualization, one needs to implement properly key graphical representations and tabular reports. The key graphical representations being histograms, charts, bullet graphs with representation elements that serve the needed level of decor like colors and symbols, are all meant to be effective and work together with the human visual perception and the human cognition, rather than against it [5].

In another work, Zhang et al. underlines the setbacks traditional charts have when it comes to high dimensional and fast changing data. They also show throughout the evolution of scientific visualizations with the goal to meet the requirements of increased amount of data, that there are still crucial limitations in terms of applicability on wider scale in the managerial context [34].

According to Zhang "business information visualization is domain specific and dependent on user takes and the characteristics of the data to be visualized" [34]. Moreover, he states that the final visualization may not and should not be all the same, which leaves us with the concept of aiming to creating guideline and a framework to be followed when addressing business KPIs[34]. More, in his work he lines that not all geometric transformations and direct visual translations would be able to transfer procedures of scientific visualization in the wide spectrum of applications, especially the managerial ones [34].

Similarly to Bačić et al., Zhang also states that business information should be visualized taking into account human perception and cognition, just the same way as it is in the case of scientific visualizations, especially due to the consideration of the human problem-solving process [34].

Stressing that the purpose of business information visualization is the enhanced interaction for a relevant insight to the decision-maker, Zhang expresses the need for developing problem solving support system from the human-centered perspective, proposing a solution that can "enhance the interaction between humans and information from a "data representation - task fit" [34]. This led to the second most important pillar of our work next to the mapping of visual IQ dimensions to business information visualization elements. Thus, the data-representation and task-fit also became our focus when proposing our data visualization pipeline for the case study we will present in the following chapters.

Furthermore, Zhang also defines the core technical challenges when designing with aim for "data representation - task fit", and he identifies therefore the linking between data representation to tasks as one, the dimension of the data as another, and finally the configuration of geometric structures and relationships among the data as being the greatest difficulties. Hence, they propose a general model that consists of various iterations among different stages of the business information visualization for problem-solving support [34]. In their model they define a set of processes and techniques that followed can support the business domain in a task-centric approach allowing for problem space analysis, data and knowledge based information analysis, for pattern discovery and finally for image construction rendering. Each of the processes they use as building bricks have their independent difficulties and key points that need extra attention, hence the following guide represents they main take-away of this work, that is serving as basis to our work presented in the upcoming chapter.

The guideline of the processes is the following:

**P.1 Analysis of the domain problem space**: expressing the phase of defining tasks with the end-user which are the methods humans leverage when trying to solve some domain specific problems decomposed into tasks and sub-tasks.

**P.2 Domain specific data and knowledge collection**: is the step which structures the information, understands the relationships amongst the data. It walks through the analysis of the previously defined tasks and should be ideally inspected critically and through information analysis theory lenses.

**P.3** **Aggregation and pattern discovery**: is the process that has the highest risk carrying elements that can make the final result be either very effective or very blend. The process involves Knowledge Discovery in Databases (KDD) aiming for the potentially highest turnover of useful information extracted implicitly from large and noisy data.

**P.4** **Visual shape construction**: focuses on the proper placement of non-geometric data and relationships that express patterns. Laying out these correctly involves image creation, creativity and is highly determined by the human visual perception.

Therefore, the need for creating a framework of standards of KPI visualizations that apart from selecting the right set of metrics also show them in a truthful manner while considering cognition of a wider audience becomes essential and highly beneficial for businesses of all kinds. This is also supported by Zhang, as he also outlines the problem of most existing visual systems not being in line with how humans are solving problems and that this directly results in difficulty to use them to enhance decision making.

Apart from the issue of misalignment between the visual tools and the nature of problem solving by humans, the other point of failure when it comes to creating visualizations hides in the extensive time and effort needed for the data pre-processing and analytics. Bikakis et al. underline the need for additional effort devoted to the phase of visualization and explanation of the results [19]. One way to tackle this effort according to Bikakis et al. is to rely on efficient visualization metaphors and on smart visual interaction paradigms [19].

However, visualization systems must also offer customization capabilities to different user-defined exploration scenarios and preferences according to the analysis needs [19].

## 2.2 Business data visualization

This section aims at presenting the most common business data visualization types and forms, methods for choice selection and the most commonly used methods in dashboard building. Further, we present in this section the study case.

### 2.2.1 Types of data visualization

As discussed and learned from literature, the goal for a data visualization to be effective and helpful for the user. In order to achieve this, it needs to leverage the cognitive aspects of problem solving and to enable the user to solve problems related to predefined tasks.

Data visualization has various types and can help for numerous different tasks, however within the category of basic data visualizations the two most common types comprise of visualizations related to exploration and explanation tasks. In case of exploration, the pattern discovery is key as there is no initial hypothesis to be tested, rather a free, constraint-less lookup for potential hypothesis. On the other side, in case of explanation, the visualization is trying to solve the task of responding to a hypothesis [34].

Both exploration and explanation tasks can be presented visually in various different ways based on the nature of data. The nature of data and the nature of the cognitive task (Ct.) aimed to be supported by the visualisation leads to 4 main categories. These high level task categories are:

**Ct.1** comparison

**Ct.2** composition

**Ct.3** relationship

**Ct.4** distribution.

These high level, core tasks according to Simon et al. can be derived with the help of problem-solving models, which serve as guidelines to study and understand a domain-related problem and the tasks needed to tackle it [17].

**Ct.1 Comparison**

The cognitive task of **Ct.1** comprises of tasks of a large variety that are most commonly visualised with column charts, tables with embedded charts (trellis), bar charts, area chart and line chart. The comparison can be for example between the subject items, but it can also be over time. The most common data type in such case is categorical, as the cognitive task searches for answers to questions that have the form "Which one is more ...?"[34].

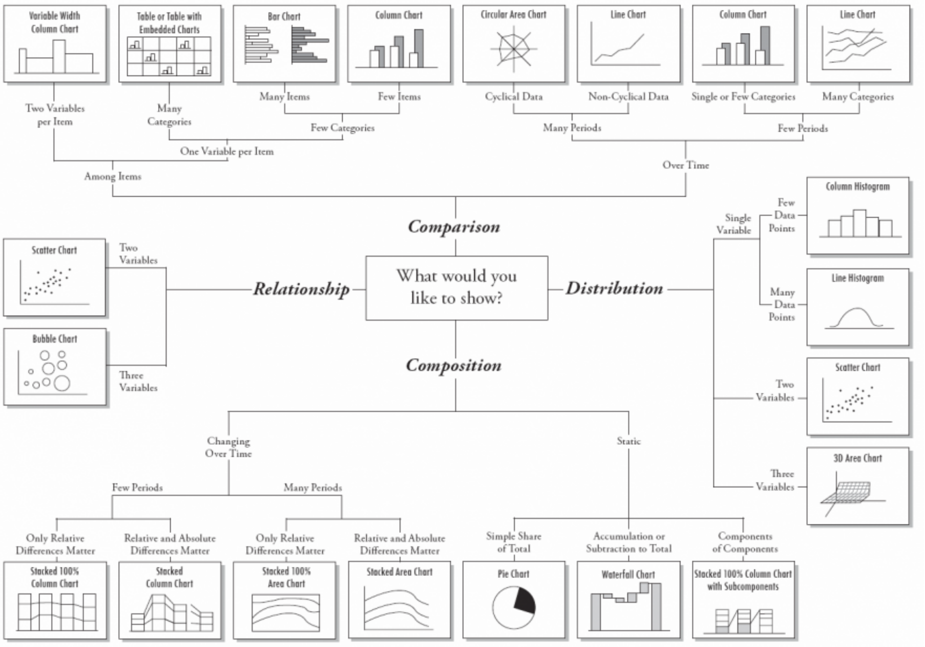


Figure 2.10: Types of data visualisation according to the task categories[9].

***Example Ct.1***:

One example for the comparative task would be, searching for the industry that declined the most in it's total revenue during the pandemic. In this case the core task is firstly to explore the data, and the specific task can go as much in depth as finding which month the industry that suffered the most started its decline. For this case we would have to use a line chart for multiple categories because the data under examination is categorical time series. Another example would be the case when a machine learning model has to select the features that are the most useful from the main subsets of the data, having to find the variables of the overall set of input variables to the model that carry the most information. In this case, the dataset could be about dating data, where the main subcategories would cover the different genders, and the features would be for example hair color, eye color etc. Therefore the number of categories is limited to a few, and the features can be compared across them easily. Looking for an answer to the question which feature is the most relevant when it comes to the gender = 'women', multiple bar charts next to each other are the ones that can be of the best help.

[9] https://images.app.goo.gl/xsdf5HiaFm7uGCMCA

### Ct.2 Composition

The cognitive task of **Ct.2**, similarly to **Ct.1** also comprises various tasks, and therefore covers a large amount a visualisation types. The question that it tries to answer is the type of question "How much of .. is in .. period/out of the total?"[34]. Tasks from this group can cover both static data as well as time series, and they both mainly focus on the exploration task. However, in some cases it can also serve as help for explanatory tasks, by supporting or negating hypothesis.

***Example Ct.2***: One great example for exploring the composition of the subject data would be to examine the nature of university students over the past two decades. In this case we would be interested to see how much the age groups attending higher education have changed, how much 50+ years old students are there in this period and how, much their number has grown or lessened throughout the years. To aid this, a stacked column chart that tracks the relative and absolute difference amongst its groups would be perfect, as the stacked columns in total could represent the total number of student, and the different stacks on the column would show each age group. Like this, for example we could see such phenomena like that a decade ago the total number of university attendees has grown and the most significant growth was in the age group 23-28.

### Ct.3 Relationship

The cognitive task of **Ct.3** covers the searching task which questions "How does .. relate to ..?". Because of the core principle that the visualisation aims at helping the human to solve a problem, the number of dimensions under investigation can be higher, but the one of plotting should not exceed 3. Therefore, looking for relationship between 2 or 3 variables is suggested even with the need of dimension reduction. This leads to easily understandable and less cluttered visualisation[34].

***Example Ct.3***: Looking for similar cities for a tourist city recommendation system we could take the number of historical points the city covers and the average rate of restaurant reviews. This would mean 3 variables under inspection for which we would choose a bubble chart with the x-axis of restaurant reviews and y-axis of tourist spots. The cities with their names would be indicated by colors and the closest bubbles would show the most similar cities, that the system could recommend to tourists that like to visit the most historical spots and like to have high rated restaurants in the nearby.

**Ct.4 Distribution** For the cognitive task of answering distribution related question the sub-tasks are also various and highly domain-specific. The question types of "How many .. are in the .. -th bucket?" are just one main the distribution covers. It purely and statistically shows the data, and enables the the user to find peaks at specific point of the variable under investigation. It also enables multiple variables to be investigated, but similarly to **Ct.3**, the maximum number of variables caps at three [34].

***Example Ct.4***: For example, a game publishing company has 50+ titles, each bringing a certain amount of revenue. Plotting on the x-axis the revenue, a column histogram of bucketed data would show which revenue bucket (i.e. 10k-30k) has the most games. If the largest bucket would be 100k euro+, then this would help see how much more/less games are in the highest revenue bucket than in the other buckets.

### 2.2.2 Forms of business data visualization

Business data visualization encompasses a variety of tools and applications that are meant to help BI developers to express their data. There are some typical forms related to the visual representation of business data and some main categories of visualization solutions that this section will present, moreover it will also focus on the tools that are at hand and are the most used in BI departments.

The three main categories of business data visualization: **reports**, **dashboards** and **standalone analytical tools**. Typically, BI result come in one of these forms, depending the nature of data and nature of insight aimed to deliver. The *reports* are generally more detailed and static, non-interactive, and serve as showing from raw to aggregated data in detailed view. Often accompanied by a rich filtering option, is usually used to show the user how different parts of the data look like. In most BI departments, reports show KPIs and metrics with the option to narrow the scope to specific locations/products/customer segment.

*Dashboards* can be thought as enriched reports, either through interactions or simply by an added layer of visual elements that make the report more comprehensible and more readable. These are also typically the most data driven, in terms that frequent updates of what data is shown based on the navigation and interaction by the user is usually what it makes it serve best for the exploratory task.

The third category of business data visualisations is represented by standalone analytical tools, such as Power BI or Looker. These tools are analytical dashboards, driven by both data and visualisation.

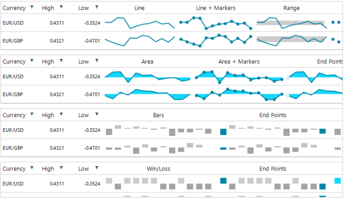


Figure 2.11: Sparkline example[10].

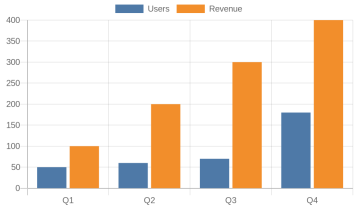


Figure 2.12: Block visual example[11].

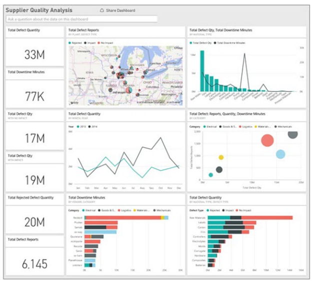


Figure 2.13: Standalone visual example[12].

When it comes to the visual forms that are used when building any of the previous three categories the option how to show the visualisations on the screen or canvas, we can name the following ones:

1. **Embedded visuals**: these are the ones that can go along the lines of texts, in-between table rows, being *Sparklines* or *formatted inline* charts, these are small visuals, serving as small hint for the user about the nature of data as seen in Figure 2.11. It also encompasses the direct formatting and styling of texts, shapes and other visual variables, thus, adding decorative feature of highlighting aspects of data and revealing insights is not directly intrusive on the content, but it can easily cause over-loaded visualisations, and can end up being more distracting than useful. Opposite to the conditional formatting, the Sparklines are more useful embedded visuals, they being minimized charts, usually showing trends, carry relevant insight to the data with less chance of distraction. Sparklines are meant to help cognitive tasks in supporting the interpretation of data, for example, in a data table with a greater number of columns, it can help showing the trend, that is anyhow noticed by the user, but supported and reinforced enables the user to focus on the next cognitive task. Being such a simple and effective visualization form it also easily follows Tufte's principles on effective visualisation[29].

2. **Block visuals**: are less threatening for the final visualisation in terms of being distracting chart junks. These block visuals are independent as shown in Figure 2.12, and they take up a more significant space on

[10] https://images.app.goo.gl/A3iGxX8sjzKhweNy5
[11] https://images.app.goo.gl/XtN37EpHd91mRSiT9
[12] https://images.app.goo.gl/Banx3j3mrK68qX5W8

the canvas, however, only in rare cases when they offer a significant amount of data points are on their own sufficient for a business data visualisation. Block visuals usually contain charts, diagrams, smaller maps and data tables, with the possibility to contain embedded visuals as well. However charts and diagrams are used interchangeably, when it comes to business data visualization there is a distinction between the two as for showing quantitative measures and business indicators a chart is more suitable, whereas for showing qualitative information or underlying structures and relationships a diagram is better. While charts focus on the level of abstraction of the information, diagrams are concrete and tangible visualisations for representing clearly the structures and relations in the data. Both take a huge role in business data visualizations, being fundamental building bricks they are used in most BI reports and visualizations.

Charts, and therefore the block visuals, cover eventually most of data visualization types as seen in subsection 2.2.1. It is also noticeable the distinction between the two in terms of showing more effectively quantitative or qualitative data fit with the 4 main purposes of comparison, composition, relation and distribution. On the other hand, there is still a plethora of different charts for more specific business use cases such as bullet charts, waterfall charts, Gantt charts, funnel charts or candlestick charts.

3. **Standalone visuals**: are applications on their own, serving the user through interaction in exploring and exploiting the data through a mix of different types of visuals as seen in Figure 2.13. Not only the type of visuals is various but also the type of interaction and the possibilities to use the different controls varies.

These are all relating to the physical position of the visualisation, which alone has the power to draw the attention of the user to specific parts of the visualisation.

### 2.2.3 Dashboards

By definition, a dashboard is “a visual display of the most important information needed to achieve one or more objectives; consolidated and arranged on a single screen so the information can be monitored at a glance” [18]. Originally the term is taken from the visual displays for operations status monitoring, but it got expanded to visualise business performance from available digital data while allowing for interactions.

The data driven standalone visualization primarily shows metrics and KPIs of the business, especially in case of performance-focused dashboards. Furthermore, trends, breakdowns and forecasts are the most common to have on dashboards of BI. The variety of visualizations that can be integrated into a dashboard has been discussed already in previous sections, but what makes it to be a dashboard solely lays in the richness of the userface and the interactions that the user is enabled to to do, this way genuinely contributing to helping in delivering answers to complicate cognitive tasks.

Dashboards are very different from reports, even though components can be the same, such as the use of charts embedded in block visuals. However, the greatest difference that sets dashboards apart from reports is the fact that dashboard balances through interaction the exploration, focusing on delivering a quick and insightful snapshot into what actually is happening in the business and what statuses key metrics report. Pappas et al. have also spent significant attention on creating guidance and framework to effectively building data driven dashboards for business purposes [22]. As they conclude, without a well designed dashboard the decision maker would have to go through a number of reports, use that in-depth data-offering and find the higher level resolution alone, which is time consuming and also more prone to errors.

### 2.2.4 Tools

The most commonly used tools to create reports and visualize data in forms of dashboards are Tableau, PowerBI, Kibana, Looker and SAS Visual Analytics. The key difference amongst these lays in the data source they allow to connect to and the flexibility to which extent they allow the creator to build reports and dashboards with them. For example, Kibana only allows for data source from the data collecting and log-parsing engine Logstash in Elasticsearch search engine. On the other hand, Tableau allows for a variety of sources including the custom connections that are made available through connections to Google Cloud, MySQL databases, or to Oracle database and many more.

## 2.3 Company case

Following the types and forms, reasons and frameworks of essential visualization aspects on which our work is based on, in this section we aim to present the study case, developed at the company of the internship, an online game publishing company. The company has around 30 licensed game titles, operates on more than 20 game portals and also covers more than 500 casual games. However, the main revenue comes from the 30 main titles from the 20 different game portals. The company acquired through subsidiaries other media companies that focused on marketing, therefore they also track marketing attributions related to their games. The group has acquired 3 game publishing companies that have their own structure to send data to the BI department, however for unification purposes the transaction data from all three have been transformed in the ETL pipelines to follow the same schema.

### 2.3.1 Data

When talking about gaming data, we have to first understand the two main data types: in-game data and platform data. By these two, we understand that there is difference between events that are triggered by players actions and users account behavior such as logging in into the game, or purchasing in-game coins with real money. Platform related data covers two main types of data, namely marketing attribution and platform engagement such as logins or purchases with real money. However, because marketing attributions are handled and tracked independently, we consider them as being a third source of data. Thus, on the highest level of abstraction the goal of the company's BI team is to create reports using these 3 main data sources, as show in Figure 2.14.

The biggest challenge the company faces in its reporting and data analytics is due to its data. Their raw data is highly volatile and dynamic, therefore hard to create visualisation that is easily scaleable, while still offering an efficient interaction level for exploration and summarizing.

The data sources for in-game event level data also vary depending on the back-end of every individual game, since the nature of games and genres means that different event types and actions are tracked on game level. Moreover, when it comes to the marketing data, there is also difference imposed by different marketing analysis tools, which in result report the same attributions in different ways depending on the device the games is running,

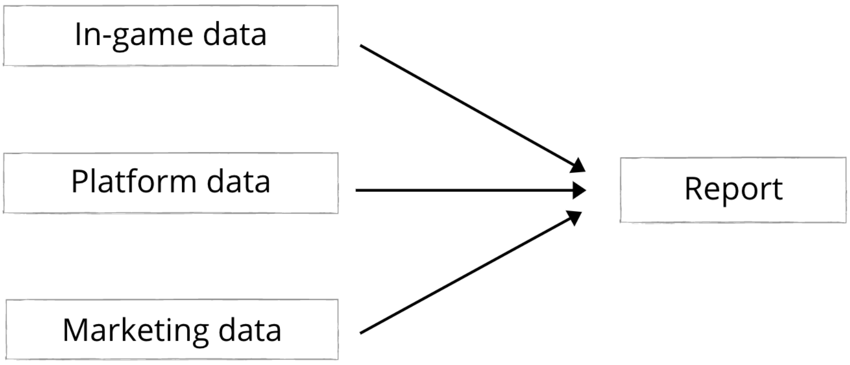


Figure 2.14: Abstract level of data sources

for example Flurry[13] is specifically meant for mobile game tracking, whereas Google Analytics is supported for both mobile and pc games marketing performance. On the other hand, platform data has been unified, therefore the logins and transactions on account level are coming in the same format for all games.

The company defines 3 main data categories:

1. **Game specific KPIs**: this encapsulates data from the in-game logs and covers battles, tournaments, quests and game character related actions such as acquiring skin, using experience points (xp-s).
2. **Core Game KPIs**: these are game related core metrics, such as virtual currency spending, specific items acquisition, bans etc.
3. **Core Company KPIs**: these are the high-level KPIs about the number of active players, registrations, churn players and revenue.

### 2.3.2 Reports

In terms of reporting the company does daily, weekly and monthly reports about the company's performance metrics using mostly data of the core company KPI group. These reports are created in Tableau, which allows for reports to be sent out in newsletter format to external partners too. The reports are of data tables type in a standalone format, and contain rows either

[13] https://www.flurry.com/

of a single day or a week or month. In general there is little to no interaction in these, as these are meant to solely serve the purpose of providing a snapshot of the company's performance to the higher level of management.

### 2.3.3 Dashboards

In terms of dashboards, the spectrum is richer as this entitles all Core Game KPI dashboards including the ones on currency usage, item frequency, quest statuses, as well as the Core Company KPI dashboards. The company uses both Kibana from the Elastic Stack to visualize data as well as Tableau.

Current method of dashboard creation for reporting purposes exhausts with simple report about a games performance in terms of Core Company KPIs visualized in the form of a block visual with data table visualization type. The report includes a game's performance per country, platform and source of user in terms of marketing attribution. For example, this report can tell the production team about how many users acquired through Facebook Ads registered from Germany. The data being temporal, the data table is broken down into days, that can be grouped into weeks, months and year. The columns showing the metrics are combined with attributes such as country, platform, and user source which can all individually and independently be always set to "Show All" or "Show individual" values, depending what the team wants to see. If the team want to see the number of Daily Active Users (DAU) in average from the month of June that they acquired organically they can collapse the month of report date to have monthly aggregations and set for country and platform "Show All", whereas keep "Show individual" for the Show platform.

### 2.3.4 General use-cases

General use cases in the company are provided by production teams who request majority of the dashboards and reports. The use-cases can differ depending on whether it is a new game that is being launched and information and indirect feedback is needed through data dashboards, or it can also be the case of testing hypothesis. The latter is the most common, however that is only due to the number of new releases being steady and number of production team members having thoughts and ideas to test on dashboards is higher. Although the use-cases are mostly for testing hypothesis, it is also common to request a dashboard that keeps track of the game performance according to the company KPIs (item 3 of company data categories) and allows for a in-game data analysis item 2 of company data categories).

The specific use case we are going to focus on for our case study is derived and given in the form of the following 4 scenarios:

**S1** Exploring the tutorial efficiency in making players understand the game

**S2** Finding the bottleneck in losing players after install (immediately and on long period)

**S3** Explore the purchasing willingness at different points in time

**S4** Finding the reason for starting/stopping to purchase

## 2.4 Evaluation methods

According to Bačić et al. it is common scenario to only use the most common measurements of decision performance such as decision speed, accuracy, and recall, however data visualization quality results are visible in terms of the impact on the outputs like decision confidence, trust and credibility [5]. Therefore, the fact that BI systems are complex decision making systems taking into consideration that the output of one visualization, being the input to another visualization, thus to a whole decision making support system, the traditional metrics of evaluation may fail to provide an optimal overall decision, which in business cases can result in financially costly decisions.

Therefore, the evaluations most often ran are done by monitoring the behaviour of independently selected stakeholders for example through trackers built in the visualization trying to measure the mouse movement and behaviour.

# Chapter 3

# Proposal

In the domain of applied decision making that involves visual-spatial uncertainty the dominant framework that is under research and at the same time well in effect is a dual-process account framework. Dual-process theories according to Padilla et al. differentiate between two processes regarding the type of decision-making, therefore one type is automatic and easy, whereas the other is more used for contemplative decision [21].

Understanding that humans in the business contexts make intuitive and strategic decisions with a degree of risk that is variable can be realized through the dual-process account of decision making [21]. The theory suggests that two types of processes cover easy, fast decisions by default, but are able to make decisions that require effort and contemplation. The processes are different, but the dual-process account theory proposes that the distinction between them does not necessarily mean completely separate cognitive and neurological systems action [21].

Therefore, the proposal to creating an effective visualization that steps out of the comfort of static reporting is going to be presented in the following subsections with arguments built on understanding existing work and literature. The proposal is also broken into two parts: one purely focusing on deriving the needed visual elements and the visual constraints that are induced by the tasks, and the other part focuses on proposing a layer of data preparation and aggregation that suits the visual components derived in the first part.

Hence, in the first part of the proposal the visualization cognition and the tasks that needed to be supported for the use-case presented in subsection 2.3.4 are going to be followed by ways for inducing freedom in the result for creative exploration. Then in the second part we are going to present way to unify the data and then how to understand the requirements on data level according to the visual components selected.

## 3.1 Framework for deriving visual components

Visual cognition covers the visual-spatial processing dimension of the visual IQ dimension mentioned in the chapter 2 in the work of [5] that proposed the mapping between the nonverbal IQ and the business intelligence visualization's components. This cognition is a part of visual-spatial reasoning, and covers the process of attaching meaning to an information visually delivered that carries spatial relations in itself.

In our company use-case, we need to consider that the production teams have prior knowledge to the games and to possess the knowledge of some graphic conventions, to which we are going to refer as mental schemes hereafter. Thus, as first step we link to the visual tasks the Cognitive Fit theory, thus we have to take into account that the team is going to compare the mental schemes to the visual representation that we are going to choose. This already excludes representations that would introduce error or a more intense use of the working memory in case of trying to match visualization that does not match with the mental schema. Excluding non-essential mental transformations from the final visualization saves the team some extra time and reduces the risk of their decisions being mistaken. However, to truly enable for the mental schema and final visual representation fit for a cognitive fit we need to first understand what questions the team has and what kind of answers they are looking for.

### 3.1.1 Abstraction and questions of use-case

In the case of the company's production team for the visualization of their new mobile game, there is a various set of questions on the table. On an abstract level the questions can be categorized according to the maturity of the game, therefore they can have initial questions looking for feedback about the game in a pre-launch phase that would encompass a whole multitude of small questions such as "How much do players understand the game from the tutorials?" or "Why some players never join a battle? Is the waiting time to long?", "At which point of the game can we win over a player to stay?". The other set of questions on an abstract level are about the long-term maintenance and game-loyalty establishment, with specific question like "What surprise did payers like the most that they kept coming back for?", "What in-game currency do players use the most?", or "When should we offer a personalized pack for the players so that they would spend real money?".

Thus, the visualization requested being for a newly developed game the questions are mostly about the fit between actual game performance and intended performance. This category of question does not only cover questions that can be answered by looking at KPIs but also induces the need for meaningful exploration. This exploration needs to be supported towards the production team and means that possible scenarios that are given by them need to be able to be played out in terms of exploration. This might seem to limit the true intention of exploration, however it actually introduces a constraint that is both constructive for us for the process of building the application as well as it is essential to the domain, as it does concern the gaming field, and not a general case exploration. With the scenarios given we can then list the questions the application will support in an abstract-to-detailed format. These scenarios and their questions in an abstract to detailed directions are:

**S1** Exploring the tutorial efficiency in making players understand the game

- **Q1** How does the dropout funnel of tutorial look like?
- **Q2** What percentage of players make it to the end of tutorial?
- **Q3** Which step of tutorial has the highest dropout rate?
- **Q4** How long in average do players stay in the tutorial step with the highest dropout?

**S2** Finding the bottleneck in losing players after install (immediately and on long period)

- **Q1** How does the retention rate look over time?
- **Q2** How significant is the retention rate drop from day 1 to day 3 to day 7 retention?
- **Q3** How long do players play on their first day of install?
- **Q4** What percentage of players plays again after installing on the same day?
- **Q5** On what platform do these players play?

**S3** Explore the purchasing willingness at different points in time

- **Q1** What is the revenue trend over time?
- **Q2** How long does it take for users to purchase after installing the game?

- **Q3** How likely is to have repeaters?
- **Q4** At what point in game is the repeater more likely to purchase again?

**S4** Finding the reason for starting/stopping to purchase

- **Q1** What game phases were the most popular before and after a purchase?
- **Q2** What are the biggest gaps between two purchases of users, and what could trigger the second one?

The abstract to detailed order of questions allows to break down the components of requirements and match them to possible visual representations.

### 3.1.2 Cognitive tasks supported

As mentioned earlier the main tasks of exploration and explanation both need to be supported, however the most important to start with is to understand the essence of a cognitive task, as it does not try to represent the "just do something". Cognitive task is any of the undertakings and operations that request human to mentally process new information and later on to recall that information and be able to use that information. Thus, this set of task require a varied set of skills that we have to understand differs per person. Skills related to attention, memory, logic and reasoning, as well as skills of visual and auditory processing are essential but various and different for individuals. Therefore, once we have the list of the tasks that need to be supported, according to the questions deduced from the scenarios presented in subsection 3.1.1, we can link them to the visual IQ dimensions and their matching BIV elements from the Figure 2.9 from chapter 2.

Hence, the list of cognitive tasks for the matching scenario and question is as follows:

**S1** **Q1** **Ct1** The cognitive task of checking the composition of data over time for exploration purpose.

- **Ct2** Cognitive task of comparing the steps of tutorial as parts of the composition over time.

**Q2** **Ct1** Cognitive task of exploring the composition of overall players who finish the tutorial.

**Ct2** The cognitive task of exploring the composition of players who finish the tutorial over changing periods of time.

**Q3** **Ct1** Cognitive task of comparing many items being the individual steps of tutorial to find the one with highest dropout rate.

**Q4** **Ct1** Cognitive task of checking the distribution of times player stay at a specific tutorial step.

**S2** **Q1** **Ct1** The cognitive task of comparing over time the retention rate.

**Q2** **Ct1** The cognitive task of comparing over time the tree retention rates (day 1,3 and 7).

**Q3** **Ct1** Cognitive task of understanding the distribution of a subset of players time spent on the specific day of their install.

**Q4** **Ct1** Cognitive task of learning the difference between players that contribute to the specific cohort that is examined, that being the examined day's installers.

**Ct2** Cognitive task of learning the growth in number of installs on the specific day relative to the total number of install.

**Q5** **Ct1** Cognitive task of exploring the composition of the selected cohort in terms which platform they belong to.

**S3** **Q1** **Ct1** The cognitive task of comparing over time the revenue trend.

**Ct2** The cognitive task of learning high level key metrics related to revenue.

**Q2** **Ct1** The cognitive task of understanding the time gap between install and first purchase.

**Ct2** Cognitive task of exploring the aggregation of different lengths over time.

**Q3** **Ct1** Cognitive task of understanding the distribution of the subset of purchasers that are eventually repeaters.

**Ct2** Cognitive task of exploring the distribution of repeaters to see the chance for having many upcoming repeaters.

**Ct3** Cognitive task of exploring the distribution of first time purchasers over time.

**Ct4** Cognitive task of exploring the different volumes of revenue by analyzing the users by their overall purchase behaviour.

**Q4** **Ct1** Cognitive task of comparing the game events where purchases of repeaters happened.

- **S4**
  - **Q1**
    - **Ct1** The cognitive task of understanding the two event phases that happened around the purchase before and after.
    - **Ct2** Cognitive task of exploring the most popular phase where purchase happened.
  - **Q2**
    - **Ct1** The cognitive task of finding the top biggest gaps between two purchases of the same user.
    - **Ct2** Cognitive task of matching the second purchase with the event.
    - **Ct3** Cognitive task of understanding the distribution of the events.

### 3.1.3 Degree of freedom

Just as in other fields, in this case too the degree of freedom refers to the options available at hand, in this case the option to get free hand on the final application. Incorporating degree of freedom in a business data visualization allows the production team and the users to find and explore the data from a new angle which is more powerful than filters alone.

In our proposal, we aim at creating and applying the guideline to crystallize the needs and the best fit for our application, but also project future scenarios that give the possibility for further exploration without the need of re-working the already delivered application. This aspect is mostly desired in production in companies where the need for such visualizations is high and the goal is to optimize the performance of team by not keeping them busy with reworking an application for every slightly new need.

Thus, what we propose is to take the measures and dimensions that are used in the final visualization and to create a separate view in which the user can combine them in the order they prefer with aggregations they select from a list. Therefore, the report builder feature is our incentive for inducing a degree of freedom.

## 3.2 Tools

The list of tools we propose are going to be the ones that we will continue to use in the following chapter of Methods and Implementation, however this can be also altered. We also provide some other tools that can be used for the same steps as we still aim at keeping the proposal also a guideline for other use-cases.

1. Google Kubernetes Engines: The game backend is run on Google Compute Engines which are component of the Google Cloud Platform (GCP). As alternative one can also use the Game Servers also provided by GCP or the Azure PlayFab game server provided by Microsoft Azure.

2. Elastic Stack: The game logs about every event happening in-game are loaded into Elastic Stack by taking the log files and shipping them into Logstash from where they are being parsed and transformed in Elasticsearch. The log data then can already be visualized in Kibana. As alternative the ELK as a whole can be replaced for instances to Loki, but also the components can be easily altered, for example Logstash could be switched to Logagent or Graylog, or Kibana to Grafana.

3. Tableau: is a business intelligence focused data visualization software tool that supports a large variety of data sources as data connections, therefore it allows for a flexible analysis while supporting relational databases, cloud databases and spreadsheets. Tableau allows for both static and interactive dashboard creation allowing for depicting the patterns and trends, variations and density of input data in the shape of graphs and charts. As alternative we can use seaborn, which is a statistical data visualization python library.

## 3.3 Preparing unified reporting data

The data side of the proposal requires many steps to achieve a unified data that can support the desired visualization, however, the reason why we started with the visual components is that they are the constraint along with the selected tools for the format of the data that is directly fed in.

In this section we will present the requirements and constraints on data per every visual component that is matching the cognitive tasks and scenarios listed in subsection 3.1.2. Then, we will present the data available from the company and present a proposal for a data pipeline that allows for the visualisations to be complete and integral both for the visualizations as well as for general usage of data which integrates and unifies the various sources.

In our case, we understand unification not necessarily only for the various data sources, but rather for the unification that needs to be maintained along the pipeline for visualization purpose. Thus, preparing unified reporting data means in our case working towards a unified data structure required for the visual components.

### 3.3.1 Requirements

The data, especially the location where they are stored creates a special and hefty procedure that is usually dealt by data engineers and refers to the extraction transfer and loading (ETL) procedure of it. Since presenting and optimizing an ETL is outside of the scope of our work, we will present the overview of data engineering that we performed and focus on the components that belong to our proposal based on the difficulties observed and faced during the first iteration of creating the visualization. As already mentioned, a core part of the visualization was requested at the company and prepared for the production team, however, that process' contribution to our work was to allow for understanding the need for a guideline and additional step in preparing the data for the visualization that supports the user needs.

The approach we built is inspired also by reverse engineering, since the iterations of the first versions of visualizations done within the company made it clear that the data should support the visualization and the visualization should lead and shape the data that is fed into, not the other way around.

### 3.3.2 Data

The data and the pipeline we propose is specifically for the one new game that was launched during the time of the internship. One part of the raw data lays on a number of game servers that are hosting the game and generate live logs on the fly about every action a player takes. The raw data, therefore, essentially consists of logs. The other part of raw data is called platform data, which is different from the game events, as it consists of transactions, registrations, logins, and account handling.

1. **In-Game data** This relates to everything that happens within the game, it constitutes of data generated by the game environment. The data is fired by the backend of the game and since it is highly unstructured belongs to the data lake and lays on Elastic clusters in json representation. e.g.

2. **Platform data** Platform data is a combination of data for online transactional processing and account events such as registration/ first_login,

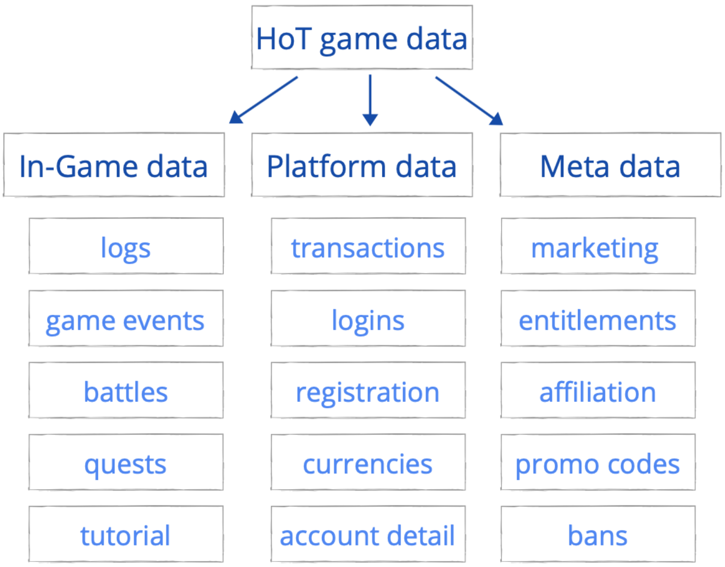


Figure 3.1: Game data categories

| **Raw data** | | | | | |
|---|---|---|---|---|---|
| **In-game data** | | | | | |
| | *timestamp* | *event_type* | *message* | *fields* | *region* |
| **logs** | 2021-07-12<br>T13:32:45+00:00 | WorldServerInfo | "<190>322 <14>\t2021-07-12T13:3245+00:00\thot-eu-test\t[d@0 p=\"M Armit(10.20.11.102) play Date(2021 07 12) Time(13 32 45) Id(401).log\"]\t2021-07-12T13:32:45.0000+00.00\t162378754\t[WorldServerInfo]\tSeq=8084.25876, \tLogD_ServerId=23, \teventDate=2021-07-12%2013%30A1%3A23, \teventId=1004, \tactionId=1912435757339898, \tlog=5]" | {"trade_item": [], "instancezonetitle_clean":None} | us |
| **game events** | 2021-08-03<br>T09:34:01+00:00 | CharacterSkillUsed | "..1629711241[CharacterSkillUsed]Seq=9656.1373162, LogD_ServerId=23, eventDate=2021-08-03%2009%3A34%3A01, eventId=203, actionId=59673141842105006, channelId=1, AID=349070, accountName=70528306, CID=633937, charaterName=Calipso, factionID=187, serverNum=23, zoneID=454, worldX=12471, worldY=28143, worldZ=98, race=8, raceTitle=warcage, gender=2, genderTitle=Female, charLevel=45, charExp=7784000, charHP=14557, charMP=12133, laborPower=981, ablility1=7, ablility1Level=55, ablility2=10, ablility2Level=55, ablility3=8, ablility3Level=55, jobTitle=Purifier(magic%3A55%2" | {"[]"} | us |
| **quests** | 2021-07-23<br>T09:45:5+00:00 | DailyQuestActivated | "<... 1629711953 [DailyQuestActivated] Seq=2016.2683334, LogD_ServerId=23, eventDate=2021-07-23%2009%3A45%353, eventId=423, actionId=58547293614334789, channelId=1, AID=356398, accountName=6500056, CID=662696, charaterName=Amarah, factionID=101, serverNum=23, zoneID=149, worldX=13100, worldY=10537, worldZ=118, race=1, raceTitle=nuian, gender=2, genderTitle=Female....ablility1=7, ablility1Level=55, ablility2=5, ablility2Level=55, ablility3=4, ablility3Level=55, jobTitle=Warlock(magic%3A55%2Cdeath%3A55%2Cwill%3A55), money=7134957" | {"[]"} | us |
| **tutorial** | 2021-07-13<br>T18:06:16+00.00 | payload.subtype:<br>Tutorial | "{ "userId": 1286, "log": {"logger": "hot.eu.test.core.netty.http.filter.ResponseDebugFilter","level": "TRACE" }, "payload": { "items": {"chestSubtype": "TUTORIAL", "softCurrency": 225, "tier": 1,"orbs": 1,"chestType": "GAME", "cards": [{"quantity": 1, "id": 501}, {"quantity": 1,"id": 23},{"quantity": 1,"id": 28}], "hardCurrency": 0}},"process": {"thread": {"name": "netty-http-executor-0007"}}, "@timestamp": "2021-07-13T18:06:16.122Z","@version": "1" }}" | "@timestamp": ["2021-07-13T18:06:16" ] }, " highlight": {"payload.items.chestSubtype": ["@kibana-highlighted-field@TUTORIAL@/kibana-highlighted-field@"]}" | us |
| **battles** | 2021-06-21<br>T10:27:59+00:00 | CombatEnded | "..M Armit(10.16.198.24) play Date(2021 06 21) Time(10 27 01) Id(401).log"] 2021-06-21T10:27:59.0000000+00:00 1629714479 [CombatEnded] Seq=9656.1616565, LogD_ServerId=23, eventDate=2021-06-21%2010%3A27%3A59, eventId=206, actionId=596731673232828, channelId=1, AID=349070, accountName=78765286, CID=692937, characterName=Meanhead, factionID=187, serverNum=53, zoneID=332, worldX=14016, world.." | {"[]"} | us |

Table 3.1: In-game data examples

logins/logouts, and currencies exchanged on the platform, meaning exchange of game coins purchased with real money into different currencies that can be used and spent in different game words. This data lays in the data warehouse hosted on Google Cloud Big Query. e.g.

<table>
<tr><th colspan="15">Raw data</th></tr>
<tr><th colspan="15">Platform data</th></tr>
<tr><td rowspan="6">transaction</td><td rowspan="3">timestamp</td><td rowspan="3">source</td><td>.amount</td><td rowspan="3">values</td><td>.amount</td><td rowspan="3">payment</td><td>.method</td><td rowspan="3">status</td><td rowspan="3">customer_id</td><td rowspan="3">user_id</td><td rowspan="3">item</td><td>.type</td><td>extra_dim<br>.name</td><td>extra_dim<br>.value</td></tr>
<tr><td rowspan="2">.currency</td><td>.currency</td><td>.provider</td><td>.name</td><td>description</td><td></td></tr>
<tr><td>.rate</td><td>.country</td><td>.id</td><td>RMT</td><td></td></tr>
<tr><td rowspan="3">2021-07-11<br>T20:05:12+00:00</td><td rowspan="3"></td><td>-9</td><td rowspan="3"></td><td>0</td><td rowspan="3">payment</td><td>null</td><td rowspan="3">Success</td><td rowspan="3">1932765</td><td rowspan="3">1236488</td><td rowspan="3"></td><td>collectible</td><td>.name</td><td>.value</td></tr>
<tr><td rowspan="2">Crowns</td><td>EUR</td><td>null</td><td>bags</td><td>description</td><td>medium<br>pouch</td></tr>
<tr><td>0.85</td><td>null</td><td>17538</td><td>RMT</td><td>none</td></tr>
<tr><td rowspan="5">registration/<br>first_login</td><td colspan="3">timestamp</td><td colspan="2">account_id</td><td>customer_id</td><td colspan="3">user_id</td><td colspan="3">extra_dim.name</td><td colspan="2">extra_dim.value</td></tr>
<tr><td colspan="3" rowspan="4">2021-07-22<br>T19:15:25+00:00</td><td colspan="2" rowspan="4">787846379221</td><td rowspan="4">1932765</td><td colspan="3" rowspan="4">1236488</td><td colspan="3">event_type</td><td colspan="2">registration</td></tr>
<tr><td colspan="3">shard</td><td colspan="2">Armit</td></tr>
<tr><td colspan="3">server</td><td colspan="2">EU</td></tr>
<tr><td colspan="3">shard_type</td><td colspan="2">live</td></tr>
<tr><td rowspan="2">account details</td><td>account_id</td><td>region</td><td colspan="2">times_purchased</td><td>entitlement_status</td><td colspan="2">days_active</td><td>banned</td><td colspan="2">registered_dt</td><td colspan="2">net_revenue</td><td>opt-in</td><td>status</td></tr>
<tr><td>787846379221</td><td>NL</td><td colspan="2">2</td><td>True</td><td colspan="2">79</td><td>False</td><td colspan="2">2020-03-09<br>T12:10:34+00:00</td><td colspan="2">53.6 €</td><td>True</td><td>Active</td></tr>
</table>

Table 3.2: Platform data examples

3. **Meta data** The metadata is one that expands with the game and helps the game grow in terms of revenue because it covers marketing tracking and gained entitlements that are used to invest in players. When the chances for user behaviour combined with entitlements gained are such, the company invests in players with the hope that they are going to purchase with real money, thus bring revenue to the company. e.g.

| Raw data | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Meta data | | | | | | | | | |
| *marketing* | *advertisment_id* | *type* | *user_id* | *placement* | *method* | *ab_testing* | *impressions* | *net_revenue* | *start_date* |
| | 82966RL34Ad23 | rewarded_video | 1932765 | Google | embedded | B | 97 | 0.89€ | 2021-07-01 |
| *promo codes* | *start_date* | *end_date* | *promo_code* | *count* | *discount_percent* | *region* | *segment* | *epic* | *game_id* |
| | 2021-07-01<br>T08:00:00+00:00 | 2021-08-01<br>T23:23:00+00:00 | FRIDAY_25 | 1 | 25 | 1236488 | US | spring_spree | 143 |
| *entitlements* | *account_id* | *region* | *entitlement_id* | | *entitlement_status* | *times_used* | *bot* | *inserted_dt* | |
| | 787846379221 | NL | 3742W325DZ | | ingame | 3 | False | 2021-02-01<br>T08:10:34+00:00 | |

Table 3.3: Meta data examples

#### 3.3.2.1 Data sources and formats

In previous section it was already mentioned where the 3 main data categories come from and where they reside physically. Thus, the Elastic Stack (ELK) represents the data lake which is populated with in-game data and Google Cloud Big Query is the data warehouse which holds both raw platform and meta data as well as staged aggregations of them.

The formats of data in the data warehouse therefore is historical, and is created in staging steps via mathematical operations such as aggregations. The data lake on the other hand contains data in log format in json files that are indexed and managed through policies that restrict them due to General Data Protection Regulations to be of a history of 3 months.

#### 3.3.2.2 Data pipeline proposal

The data pipeline that we propose came to be after few iterations. Through our discovery and focus on following the needed visual components and the desired final application, we did not pay extra attention to the data pipeline. The data pipeline initially we set up to be rather simplistic, extracting the raw logs from the game servers directly into ELK and the raw transactions from the game platform server into tables in Big Query. After extracting we moved on to the staging step of transforming the raw data. In this phase we created high lever aggregations and combinations from the in-game data with platform data to have tables and views that hold information about the game performance together, meaning that the aggregations are done such that daily metrics and performance indicators are calculated. The performance indicators are calculated first for the two separate parts, they being game KPIs and company KPIs in the next step are combined. The combined KPIs then are used for presenting it to the stakeholders, the production team, and therefore these are fed through another set of transformation into Tableau for visualization purposes.

Figure 3.2 shows the initial pipeline, when the final step of reporting and visualization was done after an extra set of transformations that happened on top of the already created aggregated KPIs. This extra set of transformations and the complexity they introduce can be clearly seen on the right hand side of the Figure 3.2.

The problem with this is the extra set of transformations. Since they initially seemed to be insignificant to separate, they were moved in the data source connection of Tableau, being ad-hoc small queries. These queries, however, grow the second the visualization creation starts off, as in the meantime realization that some views need differently structured data can intervene. Therefore, we understand that this complexity is not necessary with that it can be substituted with a schema that is defined by each main visualization component. This schema models the data that is required for a specific visual elem and is part of the main staging step along with the initial set of transformations. To make it clear we can take as an example our use case: the scenario-derived visual components are given along the human cognition needs, these visual components are put into a global configuration file which, when the pipeline reaches to the point that it took the data from the game's back end as well as the game platform and created the two core KPI sets (core game KPIs and core company KPIs) it reads the configuration file which it tells what visuals need to be supported and takes the predefined abstract schema from there to build the views-fit data format for these.

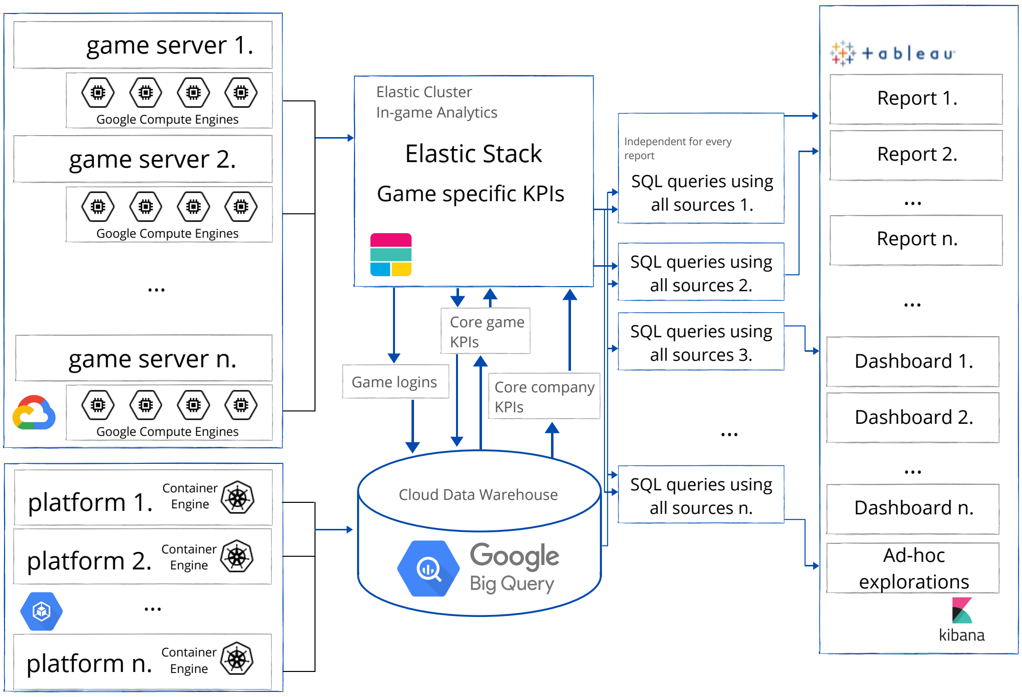


Figure 3.2: Data pipeline in iteration 1.

Figure 3.3 shows this data pipeline, where the cumbersome and hefty step of multiple complex queries in Tableau are replaced by the Visualization-Fit Format component.

## 3.4 Evaluation proposal

Information visualization is about reasoning in parts quantitative and parts qualitative manner, business information visualization is mostly of the quantitative type. Therefore, the main question to every reasoning derived from the visualization the question of ”Compared to what?” holds valid. In this terms, every argument built on top of a visualization, or derived from a visualization can be questioned whether standalone suffices the business requirements, or the ground of comparison would prove otherwise. One approach of benchmarking could possibly overcome some obstacles by setting ground values for generic KPIs in our case amongst many different companies within the same industry. However, since this would not report an overall evaluation we consider obsolete for our work.

Moreover, understanding that one evaluation only provides value and holds

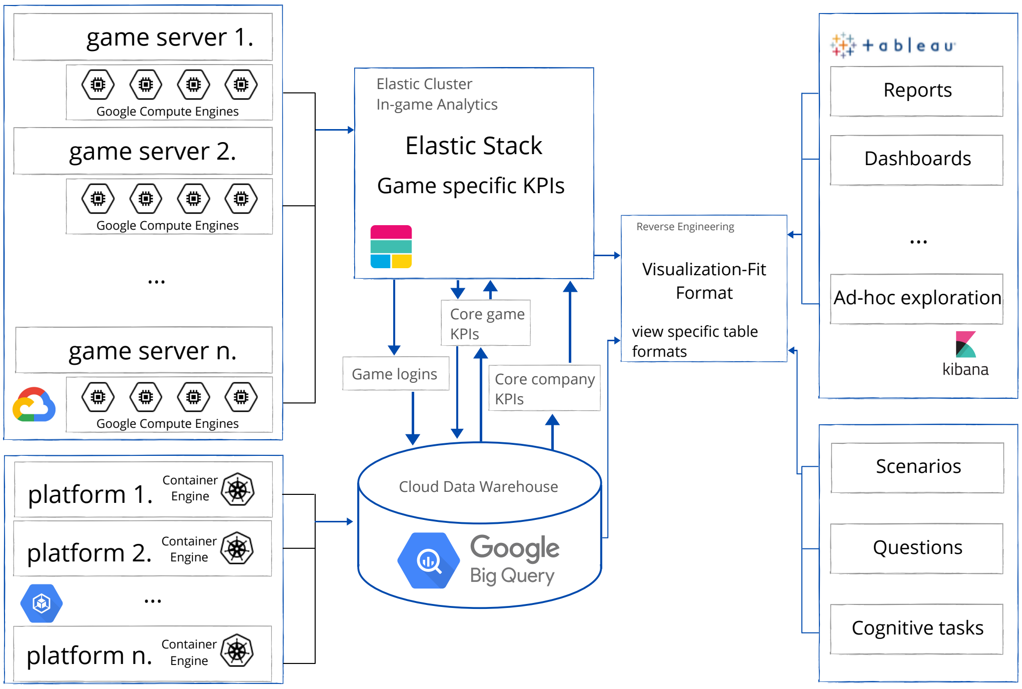


Figure 3.3: Visualization-Fit Format

true within a tight, well defined context is essential for being able to keep a critical mind about what actually is failing and what is successfully working in the final visual solution. Limitations of the context occur, and therefore, a global evaluation is almost impossible to achieve as use cases are specific and data varies as well as the tasks related too. Therefore, in our evaluation we propose to leverage cognitive walk-through, which is proven to be efficient in assessing the effectiveness of our visualization system, in terms of whether this is able to deliver the desired answers to the questions or not. Thus, we propose to get involved users that we supervise while performing cognitive processes on our visual solutions and based on the observations report the outcome:

1. of whether they try to reach to the correct answer
2. of whether they observe the availability of the correct action that they should perform
3. of whether they are able to map the right action with the right outcome
4. of whether they are aware after correct actions where they stay on the process of searching for answer.

This way of evaluating brings the possibility of easily noticing bottlenecks in the visualization tool and even can underline a bad design choice which then can be considered for change.

# Chapter 4

# Methods

To approach the case study and follow the proposal based on the observations and shortcomings at the company of the conducted internship, this chapter focuses on the methods used for our solutions with supporting arguments of literature for design choices.

## 4.1 Method for mapping data to visual representation

To the extent that our case study expands, the top level challenge requires solving the issue of handling and visualizing a vast amount of data which information visualization literature only refers to as Data Overload [28]. Driven by pipelines that map raw data to visual representation we propose a component that we are going to refer to as Visualization-Fit Format. The idea of this is supported by the visualization pipeline proposed by Card et al. that is tailored for information visualization and respects human interaction [7].

Therefore, the view specific table format is derived from the method that is presented in Figure 4.1. This includes 3 major steps: data transformation, visual mapping and view transformation. Encompassing data filtering and aggregation, then the creation of abstract visual structures and establishing the physical location or scale of the visual structures make the model of visualization pipelines robust and potentially highly effective.

For the preparation of data with reduction of data overload in mind the methods we used are filtering and aggregation by clustering. This is done by for example filtering out immediately data of types that are out of the scope

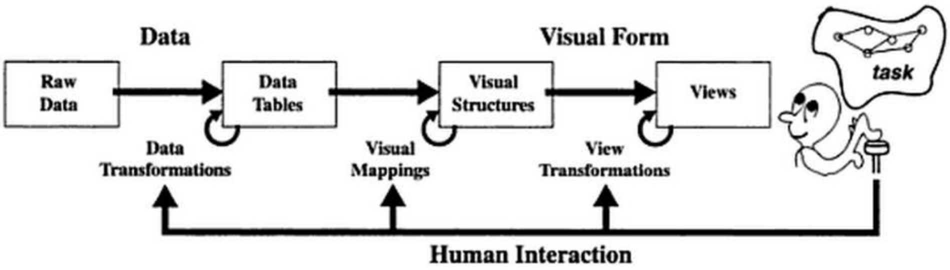


Figure 4.1: Visualization pipeline as proposed by Card et al. [7] Image taken from Card et al.[7]).

of the questions derived by the scenarios, such as eliminating all in-game event types that are not before and after a purchase or during the tutorial phase, since none of the other scenarios required information about ingame events. Moreover, the clustering is done by aggregating data in 2 stages as shown in Figure 4.2 where the Reporting layer refers to the model that fits the selected and required visuals.

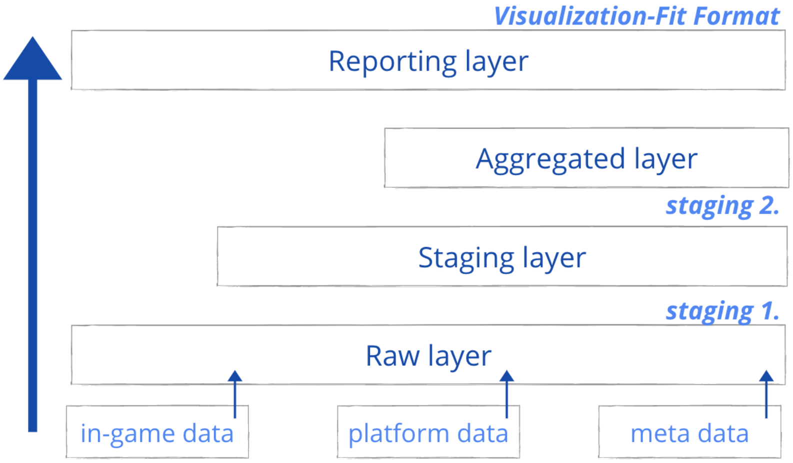


Figure 4.2: Clustering data by aggregation with staging method for data overload reduction

Staging is proposed not only to maintain a scalable and deployable big-data visualization solution by providing the necessary steps of an ETL pipeline but also for keeping integrity along the visualization pipeline proposed by Card et al. Hence, it is easy to use the staging steps for iterating backwards along the the visualization pipeline, allowing for tasks to drive the process of clustering data and preparing the reporting layer.

Staging layer number 1 takes the raw data and normalizes and cleans it, once done the data moves to the Staging layer. Data, such as logins, we normalized already in this stage, because it reduces the input data for further calculations with a great magnitude, since we only keep one login per user, instead of many. Moreover, we filtered out all irrelevant game events and server logs about server load and server performance that we did not need for any of our scenarios and questions are removed. The staging number 2. which is between the staging layer and the aggregated layer we defined as the one that performs the calculations and the one that based on the input from human interaction can take a function for aggregation and a visual component desired and create based on these 2 input the table that will serve the visual application.

We decided to stage our data 2 times as the amount of data that needs to be cleansed or normalised differs as it is greater form the amount of focus data that needs to be aggregated, even in case of task driven calculations. Thus, even in case of a new unseen task it becomes more data-efficient rolling back in rawness level as it might happen that the first staging from a previous scenario holds the data needed and only the aggregation function and the visual is new input.

### 4.1.1 Visual mapping methods

Transforming prepared data tables into visual structures that aim to combine spatial layers and marks with graphical properties while enabling as end-result for amplified cognition through visualization are the most important methods for achieving our proposed Visualization Fit Format. In literature referred to as cognitive map [7] is the careful process of mapping an intended representation of data by creating structures that support that representation.

The cognitive visual map consists of a few core elements that help structure the overall information and business information visualization. According to Card et al. these indicates the mapping of data relations into the following visual encoding:

1. Spatial substrate

2. Marks

3. Connection

4. Enclosure
5. Retinal properties
6. Temporal encoding [7]

This method also indicates that the most important and powerful is actually the way data gets mapped into the spatial layer, substrate, by selecting the attributes that are needed to be mapped and deciding on the how of the remaining variables' mapping. The decision of which attributes map into spatial position shows the importance of the chosen variable and leaves with a subset of remaining variables to be decided about their mapping.

| Data Classes and Visual Elements | | | | |
|---|---|---|---|---|
| Class | Data Classes | | Visual Classes | |
| | Description | Example | Description | Example |
| **U** | Unstructured:<br>absence or presence<br>to be identified | User Ban | Unstructured:<br>no axis needed as it can only indicate<br>absence or presence | Dot |
| **N** | Nominal | Shard | Nominal Grid: region along with division into subregions that vary in number starting with 0 | Colored square |
| **O** | Ordinal:<br>only tells about the value in relation to the context values | Spender type:<br>whale, free roller | Ordinal Grid: same as a nominal grid but the order of subregions is relevant | Alpha slider |
| **I** | Interval:<br>supports substraction of values but does not support ratios | Activity period:<br>2021.05.01 - 2021.06.12 | Interval Grid: region with a metric but no clear origin | Time axis |
| **Q** | Quantitative: supports arithmetics | | Quantitative Grid: region with a metric | |
| $\mathbf{Q_s}$ | Spatial | 0-20 km | Spatial grid | |
| $\mathbf{Q_m}$ | Similarity | True/False | Similarity space | Time slider |
| $\mathbf{Q_g}$ | Geographical coordinate | 30°N–50°N | Geographical coordinate | |
| $\mathbf{Q_t}$ | Time | 5-10 minutes | Time grid | |

Table 4.1: Classes of Data and Visual Elements based on [7] with examples of case-study dataset

This classification of the variables according to their scale types allows for further classification of the space properties related to the scale type of a space axis [31]. Therefore, the axes are, as indicated in Table 4.1:

1. Unstructured - implies no axis as it can only indicate absence or presence
2. Nominal grid - region along with division into subregions that vary in number, absence of subregion is also allowed
3. Ordinal grid - same as a nominal grid but the order of subregions is relevant
4. Quantitative grid - region with a metric that supports arithmetic functions

Therefore, based on Card et al., by using Table 4.1 we identify in our dataset the data classes and the visual classes along with them [7]. Table 4.2 shows the data classification done for item **S1**. The table shows that for example that *tutorial_started* is of unstructured type as it can be either absent or present by being set to True or False.

| | | | | | |
|---|---|---|---|---|---|
| **date** | $\mathbf{Q}_\mathbf{t}$ | 2021-07-03 | 2021-06-21 | ... | 2021-07-03 |
| **user_id** | **N** | 1932765 | 1684331 | ... | 1274121 |
| **tutorial_started** | **U** | True | False | ... | True |
| **tutorial_completed** | **U** | True | False | ... | False |
| **tutorial_max_completed_step** | **O** | 8 | 0 | ... | 4 |
| **tutorial_dropout_step** | **O** | - | - | ... | 5 |
| **length** | **Q** | 3.1 min | 0.2 min | ... | 1.3 min |

Table 4.2: Data Classification for Scenario 1 of "Exploring the tutorial efficiency in making players understand the game".

Table 4.3, Table 4.4 and Table 4.5 shows the same data classification for the scenarios of case study **S2**, **S3**, **S4** respectively.

| | | | | | |
|---|---|---|---|---|---|
| **date_registered** | ***Q_t*** | 2021-07-03 | 2021-06-21 | ... | 2021-07-03 |
| **user_id** | **N** | 1932765 | 1684331 | ... | 1274121 |
| **played_on_day_1** | **U** | True | True | ... | True |
| **played_on_day_3** | **U** | True | False | ... | False |
| **played_on_day_7** | **U** | False | True | ... | True |
| **played_on_day_14** | **U** | False | True | ... | False |
| **played_on_day_30** | **U** | False | False | ... | False |
| **avg_play_time** | **Q** | 3.1 min | 0.2 min | ... | 1.3 min |
| **last_played_event_type** | **N** | quick_match | training | ... | battle |

Table 4.3: Data Classification for Scenario 2 of "Finding the bottleneck in losing players after install"

Based on the previous 4 data tables and the axes defined in them we can develop the visual structures by combining the axes. For example, taking **number_of_times_purchased** and **avg_time_between_purchases** from Table 4.5:

$$\begin{aligned} &Avg\ time\ between\ purchases \longrightarrow Q_t \\ &Number\ of\ times\ purchased \longrightarrow Q \end{aligned} \tag{4.1}$$

Thus, the build of a visual components is done by taking two orthogonal quantities, for example variables number_of_times_purchased and avg_time_bet

| | | | | | |
|---|---|---|---|---|---|
| **report_date** | ***Q_t*** | 2021-07-03 | 2021-06-21 | ... | 2021-07-03 |
| **user_id** | **N** | 1932765 | 1684331 | ... | 1274121 |
| **first_purchaser** | **U** | True | False | ... | False |
| **first_date_purchased** | **Q** | 2021-07-03 | 2021-01-23 | ... | 2021-04-12 |
| **lifetime_revenue_of_user** | **Q** | 14.99 | 45.99 | ... | 24.99 |
| **last_event_before_purchase** | **O** | quick_match | daily_quest | ... | quick_match |
| **first_event_after_purchase** | **O** | daily_quest | tournament | ... | battle |
| **number_of_purchases** | **Q** | 1 | 2 | ... | 1 |

Table 4.4: Data Classification for Scenario 3 of "Explore the purchasing willingness at different points in time"

| | | | | | |
|---|---|---|---|---|---|
| **date_transaction** | $\mathbf{Q_t}$ | 2021-07-03 | 2021-06-21 | ... | 2021-07-03 |
| **user_id** | **N** | 1932765 | 1684331 | ... | 1274121 |
| **first_purchaser** | **U** | True | False | ... | False |
| **first_date_purchased** | **Q** | 2021-07-03 | 2021-01-23 | ... | 2021-04-12 |
| **last_date_purchased** | $\mathbf{Q_t}$ | 2021-07-03 | 2021-03-08 | ... | 2021-04-13 |
| **amount** | **Q** | 4.99 | 25.99 | ... | 7.99 |
| **status** | **N** | SUCCESS | FAILURE | ... | SUCCESS |
| **payment.status** | **N** | Settles | Success | ... | Chargeback |
| **payment.type** | **N** | payment | payment | ... | wallet |
| **channel** | **N** | mastercard | visa | ... | maestro |
| **time_between_purchases** | **Q** | 3 days | 4 days | ... | 12 days |
| **last_played_event_before_purchase** | **O** | quick_match | daily_quest | ... | quick_match |
| **first_played_event_after_purchase** | **O** | daily_quest | tournament | ... | battle |

Table 4.5: Data Classification for Scenario 4 of "Finding the reason for starting/stopping to purchase"

ween_purchases, and we map these to quantitative X and quantitative Y axis respectively as the equation above also indicates. Moreover, we can take other axes too from the data classification table, for example the spender_type which can take up a nominal axis used for coloring the marks in the final visual element.

Furthermore, we define the remaining 5 visual encodings apart from the spatial layer, depending on the nature of data. For the marks, we aim at sticking to the simplest of points and lines, wherever possible as volumes and ares might introduce an extra difficulty to cognition.

As per what relates the connections and enclosures when needed we chose them to enhance the structural hierarchy and the relationships between data points. Moreover, leveraging retinal properties we stick to color schemes that carry an extra level of information for example we use them to indicate the spender_type.

Lastly, the temporal encoding is considered by literature to be not in every case effective, and since our scenarios have a timely manner in their history we decided to keep them as they are, without any temporal change in marks positions or any of the marks retinal properties, as we believe that it might cause disturbing visual junk.

The presented visual mapping model, that we followed and gave example for our scenario data, was driven by the data and therefore the visualization pipeline followed a trend from data to visual component. This, however conflicts with the proposal of reverse engineering the desired visuals and data into preparation layer between the two. However, the method what we propose is not fully about deciding the visual and then building up the data, but meets traditional data to visual mapping methods with visual mapping derived data preparation. This means, in our case that the visual components might be directly derived from the data classes, but the scenarios can drive the mutation of data into a different format which after the transformation goes through the data classification and mapping procedure again. This is clearly represented in Figure 4.3.

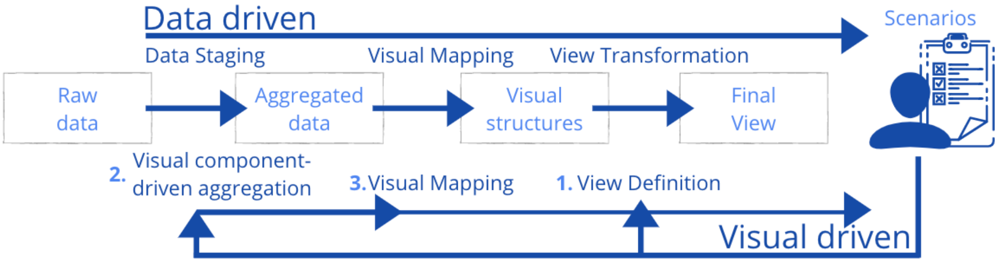


Figure 4.3: Visualization pipeline proposal combining data driven approach with visual driven approach

### 4.1.2 Visual component-driven aggregation methods

What in this thesis we call visual component-driven aggregation methods is the collection of methods that aggregate data based on arithmetical functions according to some general rule derived from the scenarios and questions of case study and also methods that are for dimensionality reduction[14].

#### 4.1.2.1 Dimensionality reduction

Firstly, Jugel et al. proposed visualization-driven data aggregation which not only reduces data overload by two orders of magnitude but also preserves pixel-perfect visualizations, as producible from the raw data [14]. This method belongs to the later that we refer to under visual component-driven aggregation as even though differs from state-of-the art time series dimensionality reduction as their complexity is of $O(n)$ it still provides a high pixel-density visualization. In their work, they created set of Visualization Driven Data Aggregation (VDDA) operators for the most common chart types, such as scatter plots and bar charts [14].

In order for us to model the pixel-level visual aggregation as query-level data aggregation we first defined our time series data model according to [14], $T(a, b)$ with the numeric columns having binary relation between them, meaning that they either relate to each other or are completely unrelated attributes. Sticking to their definition and encoding, we take the time series data as the relation $T(t, v)$ where $t$ is the timestamp and $v$ is the value, or set of values at the specific timestamp $v \in R^n$. For example, relations with multiple numerical values as in case of scenario **S1** in Data Table 4.2 where for every date timestamp there are 6 variables we can derive 6 separate time series relations by means of projection and renaming using relational algebra. As a result we will get the geometrical transformation functions $x = f_x(t)$ and $y = f_y(v), x, y \in R$ that project each timestamp and their corresponding value into the visualization's coordinate system.

The next step is then to take the parameters of the canvas where the data is going to be projected and take these geometric transformation functions to project each datetime entry $t$ and set of values onto the visualization's coordinate system. For example such geometric transformation functions according to Jugel et al. is:

$$f_x(t) = width \times (t - t_{start})/(t_{end} - t_{start})$$
$$f_y(v) = height \times (v - v_{min})/(v_{max} - v_{min})$$

In this setup the match of the datetime with start and end as well as the minimum and maximum of the selected value are used for plotting within the boundaries of the canvas.

Our goal is to display as many dimensions as needed for understanding the context, this including categorical and numerical variables. However, as seen in the Data Table 4.2 of scenario item **S1**, the number of dimensions on the variables 6 which is of high complexity. The other scenarios include up to 10 variables outside of the timestamp.

As indicated in [14], having 5 dimensions is already of a highest complexity among variables, therefore, we will need to do multiple subsets of the initial variables to be able to plot them in a way that does not introduce additional complexity in cognition, but improves that.

It is important to note that we select this method for dimensionality reduction where is needed, therefore, this method applies mostly in case of scenario item **S2**. In this case, there is a data overload in terms of user activity, logins and logouts, however reducing the dimension by aggregating this timeseries with a maximum function transforming the dataset of many datapoints over time for every user into datapoints where the maximum number of times of login activity is not zero leads to a much smaller input data. Another scenario that requires this method for dimensionality reduction is scenario **S1**, but only for the variable tutorial_max_completed_step. In this case the data model of $T(t, v)$ has $v$ as tutorial_max_completed_step and the function applied is also of the maximum.

#### 4.1.2.2 Arithmetical aggregations

After understanding the importance of dimensionality reduction the visual components also can demand arithmetical transformations, further aggregations, however, in this section we only consider them with the constraints that the selected tools bring. Some tools allow for more flexibility on data when visualizing, however in our case the method is presented with Tableau in mind.

Thus, aggregations are essential in our case and appear to be needed in all four scenarios as the general approach of handling and processing the data starts on user level and then requires to be grouped to understand the metrics on the game level. This means that the attributes presented previously are taken and transformed with maximum, average, and custom functions. Custom functions include selection of purchase times and averaging the days between all of them across all users, or the summing of all registration on one day and all logins from that day's users on day n. While most of the functions are simple arithmetic functions, few rely on the combination of them, and some on more specific filtering and aggregation with additional summing.

### 4.1.3 View management methods

View management methods aim to realize clutter-free, informative layouts with the possibility for the end user to explore and coherently examine the data to respond to questions.

Information visualization and business information visualization are built on top of separate views that are aligned together with some logic that represents the relationship they share [8]. In information visualization, overview and detail techniques were developed to compensate for the viewpoint limitation of users exploring a large data set [8]. Therefore, we chose models that build multiple and coordinated views while offering overview and detail as well to deliver the final visual application.

Roberts defines multiple views as "any instance where data is represented in multiple windows" [24], where based on the relationship between the multitude of windows organizes the visual application.

Due to the nature of the scenarios and questions we chose to use the following methods for view management:

1. **Overview & detail views:** in this case we use one view that shows the top level aggregation of data, presents it as a whole. This presents the data at more than one level while indicating where the finer grain view would fit in the large grain (overview) canvas [7]. For example in case of scenario **S2**, the overview, since the data class in Table 4.3 shows user level data that need to be further aggregated to get overview level data, would show retention in percentage of play day 1,3,7,14,30 for a filterable timeframe. For the detailed view it is then sufficient to go back one, navigate back on the aggregation layer and take for example the retention percentage of play day 3 in a line chart.

2. **Focus & context views:** are very similar to overview & detail views however, they start with the detailed view before proceeding onto an overall view management. This focus view does not have to carry as much information or cover as much data as in case of overview & detail, but it is used to put focus on a specific part of data and show similar data complexity in the surrounding. For example, with the **S2** scenario this could be interpreted as showing in prime view the retention of day 3 with a linechart, while showing retention of day 1, 7, 14, and 30 just around the main view to give context. Moreover, since the data of the case study is highly versatile and the nature of the tasks drive the visual components, we decided to also use another techniques of

focus & context which is called selective aggregation. This allowed us to create the new cases in the Data Tables that are aggregates of other cases.

3. **Difference views:** are best suited for data of the same type but that are different in a way that it brings additional value to further examine them. For example, in **S2** the perfect way to show and compare the different retentions would be through difference views that would focus on underlining the the difference in a view that merges together the separate views for each retention.

### 4.1.4 Interaction handling methods

To tackle the challenge introduced by big data with 4 scenarios of the case study we use view coordination to map the required views and understand the link between events and then show the derivation of efficiently view traversable structures.

#### 4.1.4.1 View coordination

The way views drive each other or lack the link between each other is called relationship between views. Baldonado states that this link is expected to be represented in an intuitive manner by coordination of views while interacting with them [4]. This coordination is built on the mapping that represents the changes and dependencies between the views. This mapping, that is specified by coupling functions, describes the way changes in views are affecting other views [4]. The timing and conditions that trigger this coupling functions have to be determined in a propagation model [4].

Coordination is the most obvious with user interaction. For example one coordination is brushing, which upon selection of elements in one view highlights the same or related elements in other linked views [24]. This is especially powerful in case of multiform views when the goal is to find similarities and anomalies in the data [24] .

Apart from the linking of data across views, another interaction technique that might be very useful is called navigational slave. This is for example synchronized scrolling in side-by-side difference views which in case of scenario **S1** would be a comparison of the maximum completed steps from scenario **S1**.

Figure 4.4, Figure 4.5, Figure 4.6, show the links between the views for scenarios **S1**, **S2**, **S3** and **S4** combined respectively.

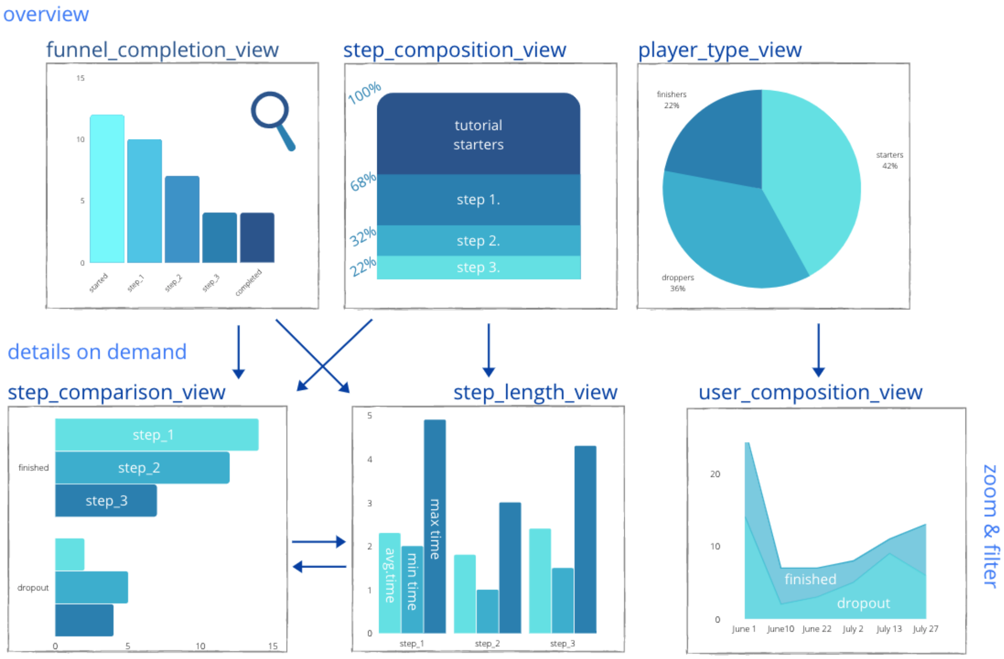


Figure 4.4: View hierarchy and links for scenario **S1**.

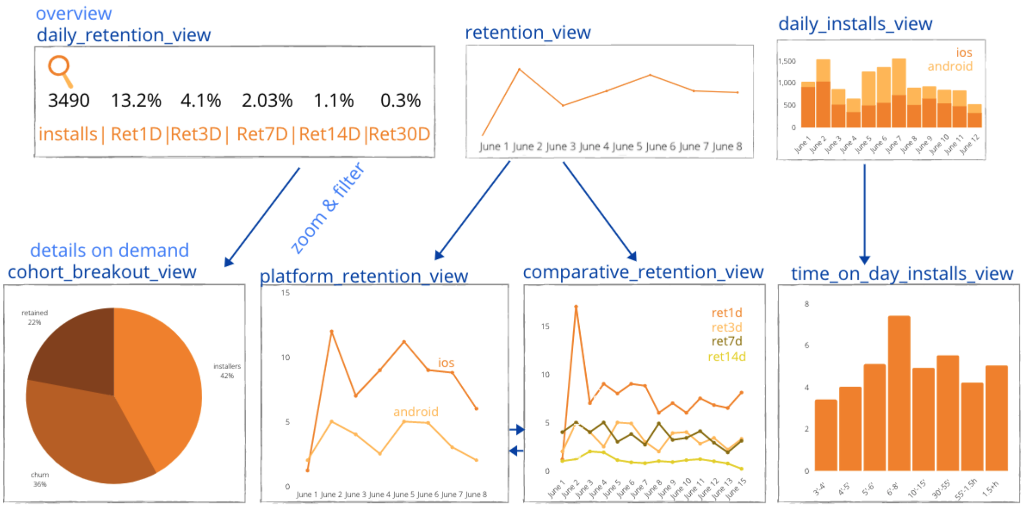


Figure 4.5: View hierarchy and links for scenario **S2**.

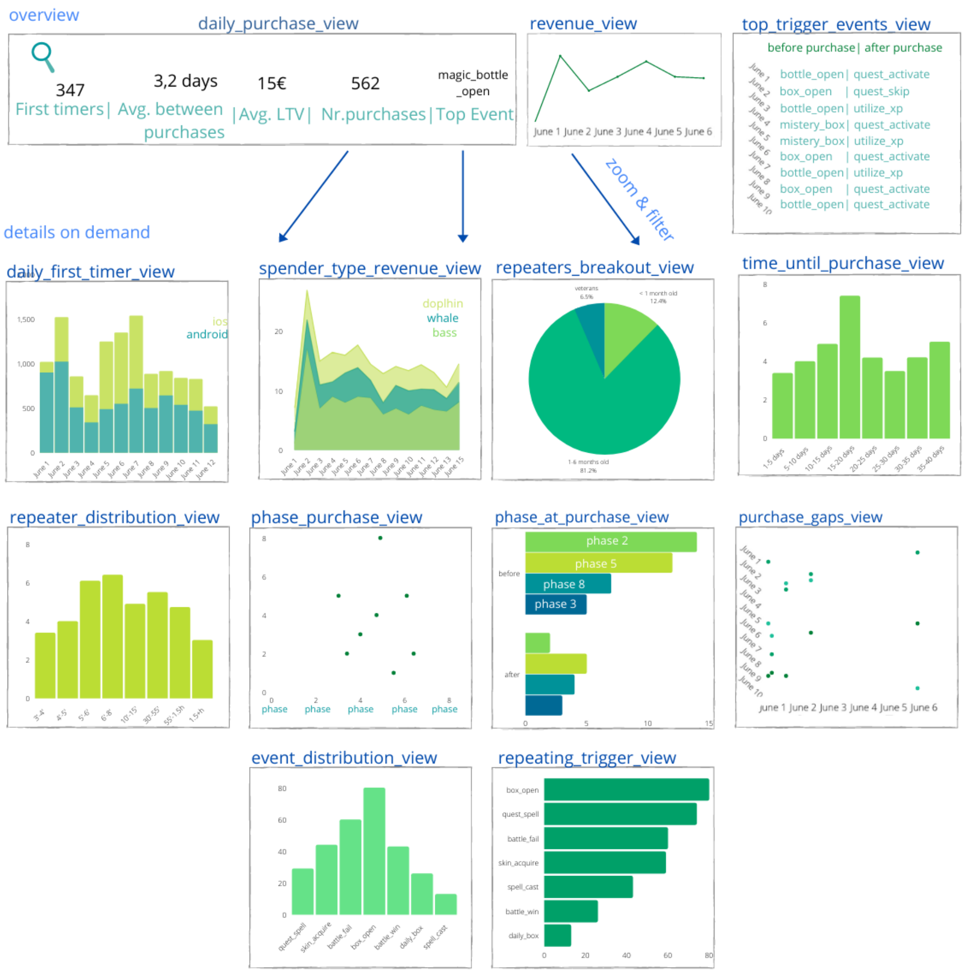


Figure 4.6: View hierarchy and links for scenario **S3** and scenario **S4**

.

There are two different approaches for directing the coordination of views. One of these is based on Shneiderman's information seeking mantra which leads to an "overview, zoom and filter, details-on-demand" coordination line [26]. Another proposal to the same multiple and coordinated views (MCV) abstract model is the "analyse first - show the important - zoom, filter and analyse further - details on demand" [25]. The latter is also called the visual analytics mantra. Either way, an effective view traversability needs to be created for an effective information representing visual tool. The common way to represent this is by trees that map the views and the relationships between them. Because our scenarios are of a number and therefore our input data is large we also decided to leverage the notion of jump-and-show (implemented in format of buttons) which is the efficiently view traversable structures additional property.

#### 4.1.4.2 Efficiently view traversable structures

The same way as elements of the information are organized in logical format characterized by a logical structure graph the views are structured in a viewing graph. These are useful representations to connect the views and elements to their logical neighbors [12]. Efficient view traversability is a characteristic of these view graphs if it's requirements are met. Efficient view traversable graphs have 2 requirements that need to be fulfilled which are related to the space and to the time as resource of the users while traversing the views even in the case of large structures. These requirements are derived from two assumptions: one of these relates to the largeness of the structure and the other one relates to the time limitation of traversing willingness of end users [12]. Thus, assuming a large structure and limited screen the user can only digest a small part of the structure from the current location the first requirement states that the number of out-going links of nodes from the view graph has to be relatively small to the size of the structure. Furthermore, to derive the second requirement the length of the paths has to be also small compared to the size of the structure [12].

## 4.2 Methods for evaluating the visualization tool

Visualization systems are designed to support high-level cognitive tasks which are by their nature difficult to quantify. Moreover, most of the visualization papers tend to have a significantly lower rate of evaluation than papers in the broader discipline of human-computer interaction. Therefore, we aim to evaluate our methods indirectly, validating our hypotheses through empirical qualitative analysis by involving the potential end users to our proposed visual solutions. Moreover, we also apply a pattern-based approach to visualization evaluation which is a proven solution to a common problem encountered when evaluating a visualization system [11]. Elmqvist et al. propose 5 categories of patterns to apply for different visualization tools, namely: exploration pattern, control-, generalization-, validation- and presentation pattern. From the aforementioned ones we decided to use patterns belonging to the exploration and the presentation pattern categories.

- **Exploration pattern:** focuses on exploring and exploiting the design space of the evaluation asking questions such as "Are the correct independent and dependent variables depicted?", "Is it confident that the study is appropriate?" or "Are the right questions asked?" [11].

  Exploration patterns are deployed at a seed stage of a visualization project, being an early evaluation for the experimenter for finding right tasks, questions and datasets and thus, solidifying a baseline for an evaluation. In order to gain a confidence that the evaluation is appropriate we decided to deploying this evaluation after deriving our questions from the scenarios [11].

  - **Our evaluation:** Attempting evaluation in early stage in our case focused on the questions and tasks that were derived from the scenarios. We decided to run a so called Do-It-Yourself (DIY) pattern for evaluation which involved at the point a single individual that is keeping a continuous evaluation on the design of a visualization system or technique. This means that the researcher of this thesis kept on asking the questions of scenarios when deciding on the visual components and kept trying the cognitive tasks. We decided to do this evaluation because the experience gained at the company of internship involved many iterative visualization creations for different use cases. This allowed for some experience and for an outside point of view when creating a proposal based on the experienced deficiencies.

- **Presentation pattern:** conserves the quality of the evaluation by reporting results in a correct and economical way. Asks questions such as "Are the results presented in an easily understandable manner?", "How to evaluate higher-level tasks and scenarios?" and "How can be the result communicated?" Evaluations to gain true meaning requires presentation to an external audience [11]. Presentation patterns helped us in the way to communicate the evaluation results clearly and efficiently. Furthermore, there is the approach of action research which was pioneered by Lewin [15]. In action research, the case study method involves the researcher more explicitly in the work aiming to steer the direction of the work ti find ways for improvement of processes.

  - **Our evaluation:** Having at hand already the case-study from the company we decided to stick with the case-study pattern for evaluation too, since this meant involving people from across different team at the company who are familiar with the scenarios and are empathetic with asking the questions or similar questions to what we derived. Since the environment thus became uncontrollable, the resulting insights could not be fully generalized. However, generalization might be a loss, ecological validity is still given at a very high rate as the context used in the particular case study is what gives rich details to individual level evaluation outcome report.

# Chapter 5

# Implementation

The implementation follows the proposal along with the scenarios given by the case study. After presenting which are the matching visual components that are derived from the cognitive tasks of the questions from the scenarios, we show the raw data and how the staging aggregates it. This is then followed by deriving the visual component-requested data format and matching that with the staged data format to see what further data classes need to be generated. As final step we show the Tableau implementation through ”overview - details - on - demand” format calling it the static solution and then combine the acquired visuals in an interactive solution which enhances the “analyse first - show the important - zoom, filter and analyse further - details on demand”.

## 5.1 Deriving the visual components

The proposal in chapter 3 started with deriving first questions and further the cognitive tasks assigned to each scenario. Matching cognitive tasks with visual IQ based business intelligence elements, such as exploration, perception or cognition were done by leveraging the types of data visualization grouped along the cognitive tasks of comparison, composition, relationship and distribution.

### 5.1.1 Scenario 1

Therefore, in case of scenario **S1**, which is about exploring the efficiency of tutorial, the following visual components were derived based on the cognitive tasks listed:

| Cognitive task | Visual component | Supporting view id |
|---|---|---|
| checking the composition of data along tutorial time | column chart with relative values | funnel_completion_view |
| comparing the steps of tutorial as parts of the composition | with relative difference matters | step_composition_view |
| exploring the composition of overall tutorial finishers | simple share of total pie chart | player_type_view |
| composition of finishers over changing periods of time | stacked area chart | user_composition_view |
| comparing steps to find the highest dropout rate | simple row chart | step_comparison_view |
| checking the time lengths players stay at specific steps | multiple row charts histogram with min/max | step_length_view |

Table 5.1: Deriving visual components for **S1** and the view ids that support the task.

After deriving the visual components we drew the view graph with EVT in mind. Thus, to comply with the 2 constraints of EVT about the limitation on the number of outgoing links and about the length of the path we drew the view graph as seen in Figure 5.1.

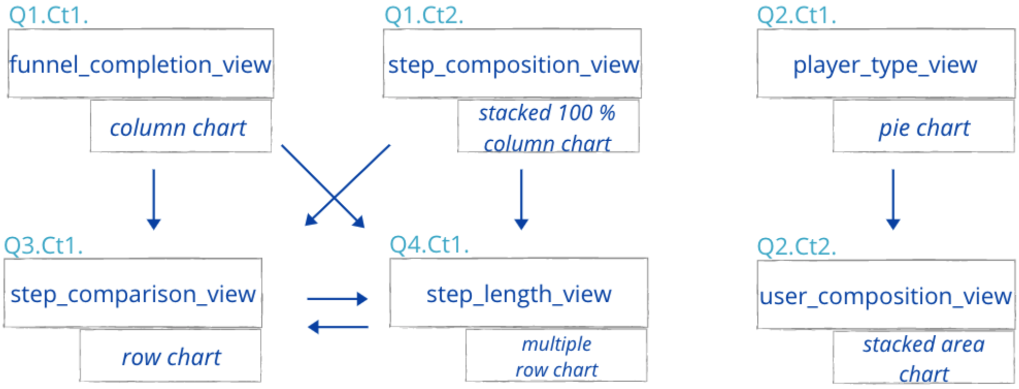


Figure 5.1: View graph with links and supported tasks complying the EVT requirements for scenario **S1** in static solution.

This further indicates the supported cognitive tasks and also shows that the highest number of outgoing links is 2 in the cases of the ***funnel_completion_view*** and ***step_composition_view***. Moreover, the longest path traversable is of 3 on ***funnel_completion_view*** - ***step_length_view*** - ***step_comparison_view***. These 2 arguments support that the view graph is efficiently view traversable (EVT).

Furthermore, the view graph is also drawn in the way that the views in the first row are of the overview and the second row represents the views used for details on demand. This is the case of information seeking in the "overview-details-on-demand" way, our static solution. In case of "analyse first - show the important - zoom, filter and analyse further - details on demand" the view graph modifies as seen in Figure 5.2, into our interactive solution.

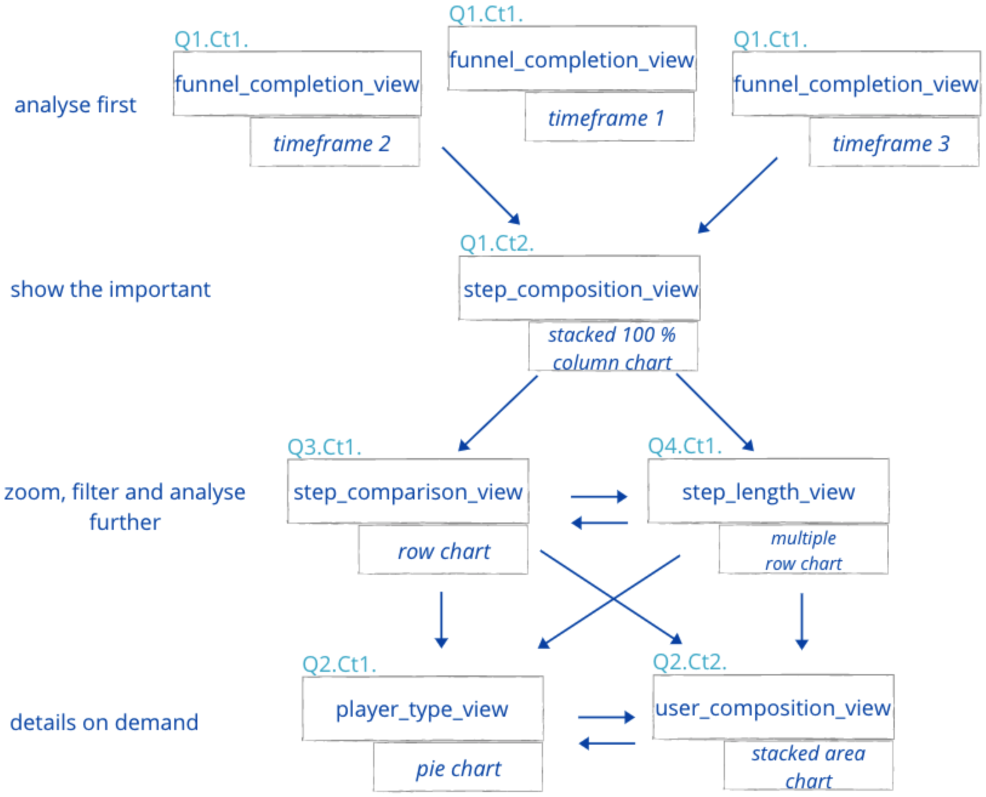


Figure 5.2: View graph with links and supported tasks complying the EVT requirements for scenario **S1** in interactive solution.

The interactive solution, how we call the "analyse first - show the important - zoom, filter and analyse further - details on demand" due to the difference in involving the user very interactively compared to the "overview-details-on-demand", also meets both EVT requirements. The number of outgoing links tops with 3 in case of ***step_comparison_view*** and ***step_length_view***, while the longest path is of 4 on ***funnel_completion_view*** - ***step_composit ion_view*** - ***step_comparison_view*** - ***user_composition_view*** - ***player_ type_view***.

**Filters:** Both the information seeking mantra as well as the visual analytics mantra involve in their view coordination technique the zooming and filtering. In our implementation for scenario **S1** we added filters only on the starting date of the tutorial. Due to the consideration of the limited number of views used for this dashboard and the fact that the overview level has to show all steps at the same time to give contextual meaning, we decided not to add a filter on the tutorial step.

**View management:** We implemented different view management for the static and interactive dashboards. For the static dashboard we followed the overview & detail view management by indicating on the overview level multiple levels of data with the direction to go into the more detailed view. On the other hand, in case of the interactive dashboard we implemented the difference views view management as we started with the funnel completion over time shown in two ways. These two ways show the same data but accentuate the difference of tutorial steps completion on an overall level compared to a daily level.

### 5.1.2 Scenario 2

Scenario **S2** focuses on understanding the retention rate of the players. This involves understanding various daily retention, overall retention and aims to perform a cohort analysis to see how a specific day's players keep sticking to the game.

Thus, similarly to subsection 5.1.1, in this case too, the implementation started after deriving the questions from the abstract scenario's description and then listing of the cognitive tasks. This was then followed by deriving the following visual:

The visual component mapping, as previously, was done referring to the Figure 2.10, which based on categorizing data visualization types according to high level cognitive tasks allows for easy mapping. This step was followed by drawing the view graph satisfying the EVT requirements. Figure 5.3 shows how the view graph looks. To prove that this is fulfills EVT constraints we can look at the highest number of outgoing links on node ***retention_view***, where this number is of 2. This, and the fact that the longest path on ***retention_view*** - ***platform_retention_view*** - ***comparative_retention_view*** and ***retention_view*** - ***comparative_retention_view*** - ***platform_retenti on_view*** is of 3 fulfills the EVT criteria of not having a higher outgoing link than 2 and a short path of 3.

Figure 5.3 is the view graph of information seeking in static, "overview-

| Cognitive task | Visual component | Supporting view id |
|---|---|---|
| comparing over time the retention rate | table view | daily_retention_view |
| comparing over time the main retention rates | multiple line chart | comparative_retention_view |
| understanding the distribution of time spent on the specific day of their install | histogram | time_on_day_installs_view |
| learning the difference between players that contribute to the specific cohort that is examined | pie chart | cohort_breakout_view |
| exploring the composition of the selected cohort in terms which platform they belong to | multiple_line_chart | platform_retention_view |
| checking the time lengths players stay at specific steps | multiple row charts histogram with min/max | step_length_view |
| learning the relative growth in number of installs on the specific day | stacked 100% line chart | daily_installs_view |
| comparing over time the aggregated retention rate | line chart | retention_view |

Table 5.2: Deriving visual components for **S2** and the view ids that support the task.

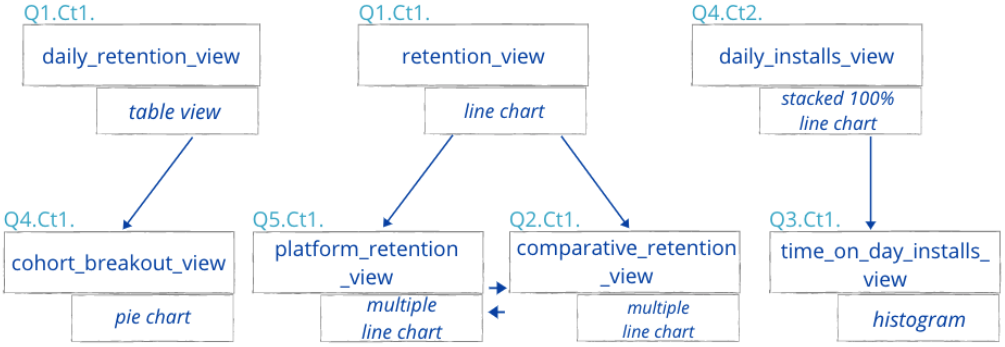


Figure 5.3: View hierarchy with links and supported tasks for scenario **S2**.

details-on-demand" way, thus, the views in the first row of the figure are of the overview and the second row represents the views used for details on demand.

As next step, similarly to how we did in case of scenario **S1**, we modified this static view graph to create the "analyse first - show the important - zoom, filter and analyse further - details on demand" view graph for the interactive solution by modifying it as seen in Figure 5.4.

The interactive "analyse first - show the important - zoom, filter and analyse further - details on demand" solution shown in Figure 5.4 also meets both EVT requirements. Hence, the number of outgoing links tops with 3 in case of ***retention_view***, while the longest path is of 4 on ***daily_installs_view*** -

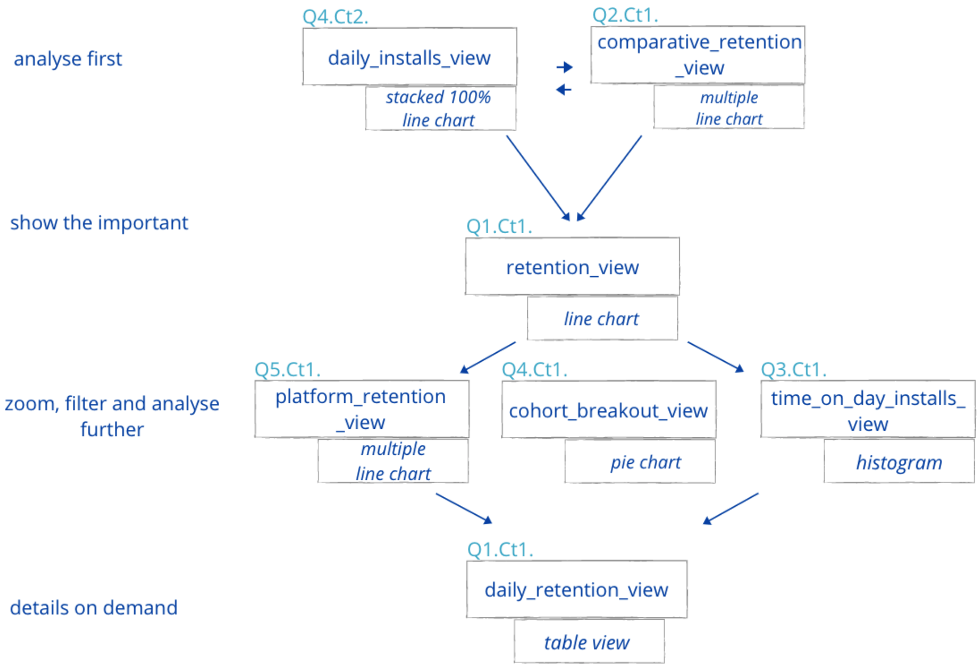


Figure 5.4: View hierarchy with links and supported tasks for scenario **S2**.

***comparative_retention_view*** - ***retention_view*** - ***platform_retention_view*** - ***daily_retention_view*** and length similarly 4 on ***daily_installs_view*** - ***comparative_retention_view*** - ***retention_view*** - ***cohort_breakout_view*** - ***daily_retention_view*** and ***daily_installs_view*** - ***comparative_retention_view*** - ***retention_view*** - ***time_on_day_installs_view*** - ***daily_retention_view***.

**Filters:** In our implementation of r scenario **S2** we added the installation/registration day and platform filters. Considering the questions derived from the scenario we saw that these two will satisfy both the information seeking and exploration to the extent to produce relevant answers.

**View management:** In case of scenario **S2** we again implemented different view management techniques for the interactive and static dashboards. In case of the static dashboard we used the overview & detail as the starting overview layer presents many dimensions of the data while indicating the direction to the specific finer granularity of them if demand exists. On the other hand, in case of the interactive dashboard we implemented the focus & context view management as we started with the detailed views about comparative retention and only afterwards moved to provided more context in terms of what the total aggregate retention is or how the cohort breakout shapes.

### 5.1.3 Scenario 3&4

Both scenario **S3** and scenario **S4** try to exploit the monetary aspect of the data, focusing on purchasing willingness and the trigger behind purchases these two scenarios are contextually so intertwined that we decided this is the best represented if we implement them in one dashboard.

Regardless, the process done was similar to previous subsection 5.1.1 and subsection 5.1.2, in this case too, the implementation started right after the question and cognitive task listing. Therefore, in this case the following visual components were derived:

| **Cognitive task** | **Visual component** | **Supporting view id** |
|---|---|---|
| comparing over time the revenue trend | line chart | revenue_view |
| learning high level key metrics related to revenue | data table | daily_purchase_view |
| understanding the time gap between install and first purchase | data table | top_trigger_events_view |
| exploring the aggregation of different lengths over time | histogram with buckets of variable size | time_until_purchase_view |
| understanding the distribution of the subset of purchasers that are eventually repeaters | histogram | repeater_distribution_view |
| exploring the distribution of repeaters to see the chance for having many upcoming repeaters | pie chart | repeaters_breakout_view |
| comparing the game events where purchases of repeaters happened | line chart | repeating_trigger_view |
| understanding the two events phases that happened around the purchase before and after | line chart | phase_at_purchase_view |
| exploring the distribution of first time purchasers over time | bar chart | daily_first_timer_view |
| exploring the different volumes of revenue by analyzing the users by their overall purchase behaviour | area chart | spender_type_revenue_view |
| exploring the most popular phase where purchase happened | scatter plot | phase_purchase_view |
| finding the top biggest gaps between two purchases of the same user | data table | user_purchase_gap_view |
| matching the second purchase with the event | line chart + event rank filter | repeating_trigger_view |
| understanding the distribution of the events | line chart | event_distribution_view |

Table 5.3: Deriving visual components for **S3**, **S4** and the view ids that support the task.

The visual component mapping, referring to the Figure 2.10 is shown in Table 5.3. This step was followed by drawing the view graph satisfying the EVT requirements. The view graph is shown in Figure 5.5. This view graph also satisfies both EVT constraints since highest number of outgoing links on nodes ***daily_purchase_view revenue_view*** and ***top_trigger_event_view***, where this number is of 2. Along with that the longest path on ***daily_purchase_view - spender_type_revenue_view - phase_at_purchase_view*** and ***phase_purchase_view*** is of 3 fulfilling the EVT criteria of not having a higher outgoing link than 2 and a short path of 3.

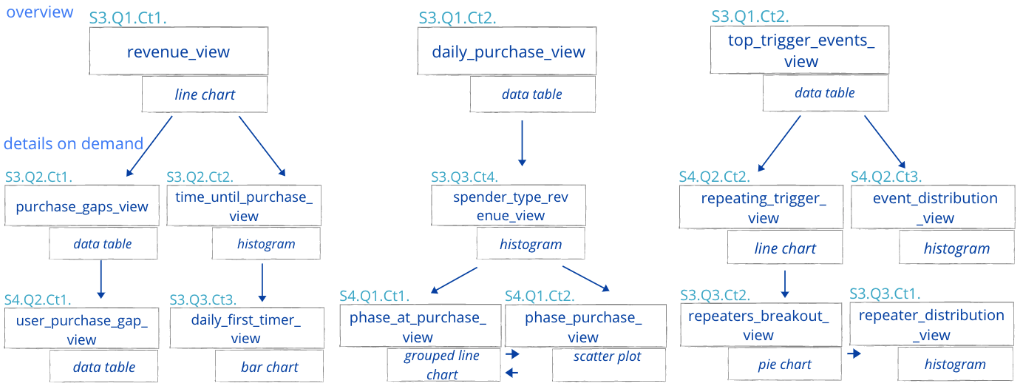


Figure 5.5: View hierarchy with links and supported tasks for scenario **S3** and **S4**.

The presented view graph in 5.5 is of the ”overview-details-on-demand”. Thus, in the next step we drew the interactive information seeking view graph of ”analyse first - show the important - zoom, filter and analyse further - details on demand” as seen in 5.6. Similar to its static version this view graph also satisfies the EVT constraints with that the highest number of outgoing links is of 3 on ***daily_purchase_view*** and the longest possible path is of 3 on any possible direction starting from ***daily_purchase_view*** downwards.

**View management:**

**Filters:** For the combined dashboard for scenarios **S3** and **S4** we implemented the filters of report date, filter of number of days before purchase on the triggering event for a purchase and on time until purchase, filter of game phase, filter of spender type.

**View management:** Only in the case of the combined solution for scenario **S3** and **S4** we implemented the same view management in both static and interactive dashboards. This being the overview & detail we decided to start

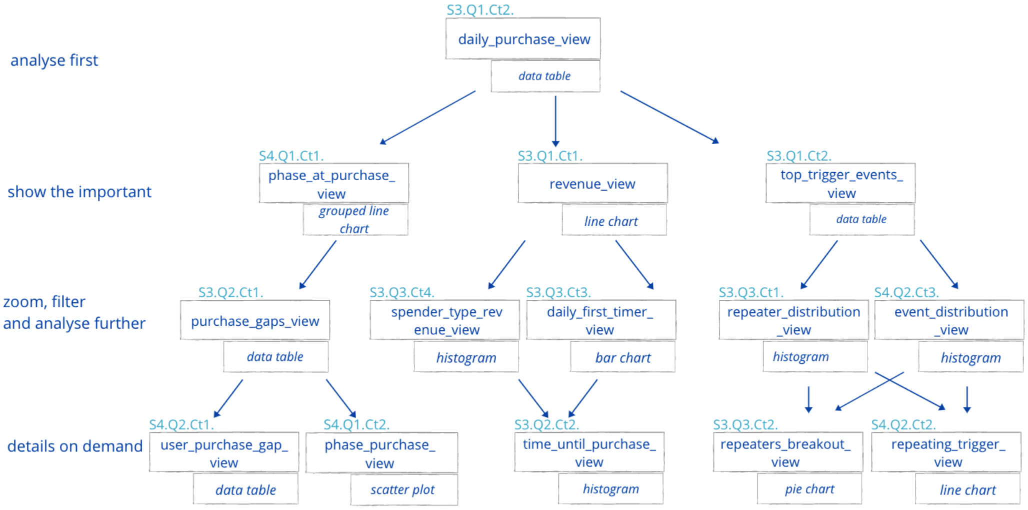


Figure 5.6: View hierarchy with links and supported tasks for scenario **S3** and **S4**.

with an overview layer that presents many layers of the data while indicating the direction to the specific finer granularity.

## 5.2 Staging the data

### 5.2.1 Scenario 1

In case of scenario **S1**, the data follows the pipeline of the two-step staging demonstrated in the chapter 3: normalizing and cleaning the data and then aggregating.

The following aggregations are done on user level expressed in SQL syntax:

```
SELECT
DISTINCT user_id,
MIN(DATE(tutorial_started)) date, % selecting the
first time user started tutorial
TRUE as tutorial_started,
MAX(step) as tutorial_max_completed_step,
(CASE WHEN tutorial_max_completed_step = 4 THEN True
else False END) as tutorial_completed,
(CASE WHEN tutorial_max_completed_step = 4 THEN null
else tutorial_max_completed_step+1 END)
```

```
as tutorial_dropout_step,
step_lengths as lengths_array
FROM
'staging_02.tutorials_data'
GROUP BY 1,5,6;
```

This arithmetic aggregations build the data structure which in the chapter 4 we showed in the data classification Table 4.2.

### 5.2.2 Scenario 2

Similarly to **S1** scenario **S2** also follows the data pipeline of two-step staging demonstrated in the chapter 3: normalizing and cleaning the data and then aggregating.

Afterwards, the user level aggregations that were done are the following expressed in SQL syntax:

```
SELECT
DISTINCT user_id,
MIN(DATE(datetime)) date_registered,
CASE WHEN play_day = 1 and played is TRUE
THEN TRUE ELSE FALSE END as played_on_day_1,
CASE WHEN play_day = 3 and played is TRUE
THEN TRUE ELSE FALSE END as played_on_day_3,
CASE WHEN play_day = 7 and played is TRUE
THEN TRUE ELSE FALSE END as played_on_day_7,
CASE WHEN play_day = 14 and played is TRUE
THEN TRUE ELSE FALSE END as played_on_day_14,
CASE WHEN play_day = 30 and played is TRUE
THEN TRUE ELSE FALSE END as played_on_day_30
AVG(play_time) as avg_play_time,
last_event as last_played_event_type
FROM
'staging_02.retention_aggregated'
GROUP BY 1,3,4,5,6,7,9;
```

This arithmetic aggregations build the data structure which in the chapter 4 we showed in the data classification Table 4.3.

### 5.2.3 Scenario 3

Even though scenario **S3** and **S4** are contextually very close and we considered them together in the visual component derivation, the data aggregation we present separately as in the final version we used two data sources on the dashboard for scenario **S3** and **S4**. The reason for this is because the data for scenario **S3** is aggregated on user level, while the data for scenario **S4** is transaction level aggregation. This means that in case of scenario **S3** the values that we select from the staged tables are considering aggregation on every user's historical data. This means that the result includes every user with its id once with the corresponding aggregated metric values.

```
SELECT
DISTINCT t.user_id,
(DATE(datetime)) t.report_date,
first_purchase_t.first_purchaser,
first_purchase_t.first_date_purchased
first_purchase_t.nr_purchases
first_purchase_t.revenue
as lifetime_revenue_of_user,
first_event_after_purchase
as first_event_after_last_purchase,
last_event_before_purchase
as last_event_before_last_purchase
FROM
'staging_02.purchase_data' t
left join
(SELECT CASE WHEN COUNT(*) = 1 THEN TRUE ELSE FALSE
END AS first_purchaser,
MIN(datetime) first_date_purchased,
COUNT(*) nr_purchases,
SUM (amount) as revenue
FROM 'staging_02.transactions'
WHERE status = 'SUCCESS'
and payment.type = 'payment'
and payment.status = 'Completed'
GROUP BY datetime) first_purchase_t on t.user_id =
first_purchase_t.user_id;
```

Therefore, the above listed query is used to generate one of the two data sources for the dashboard of the joint scenarios **S3**-**S4**.

### 5.2.4 Scenario 4

Unlike the high user level aggregation seen for **S3**, for **S4** the aggregation is done on transactions level, contributing to the second data source of the joint dashboard.

```
SELECT
user_id,
(DATE(datetime)) date_transaction,
first_purchase_t.first_purchaser,
first_purchase_t.first_date_purchased,
first_purchase_t.last_date_purchased,
amount,
status,
payment.status,
payment.type,
channel,
DATE_DIFF(first_purchase_t.last_date_purchased,
first_purchase_t.first_date_purchased, DAY)
as time_between_purchases,
first_event_after_purchase
as first_played_event_after_purchase,
last_event_before_purchase
as last_played_event_before_purchase
FROM
'staging_02.transactions' t
left join
(SELECT CASE WHEN COUNT(*) = 1 THEN TRUE ELSE FALSE
END AS first_purchaser,
MIN(datetime) first_date_purchased,
MAX(datetime) last_date_purchased,
COUNT(*) nr_purchases,
SUM (amount) as revenue
FROM 'staging_02.transactions' WHERE status = 'SUCCESS'
and payment.type = 'payment'
and payment.status = 'Completed'
GROUP BY datetime)
first_purchase_t
on t.user_id = first_purchase_t.user_id;
```

## 5.3 Augmenting aggregations for Visualization-Fit format

The augmentation for the aggregations to achieve the Visualization-Fit format was done by:

1. Taking the Table 5.1, Table 5.2, Table 5.3 that present the derivation of the visual components,
2. Taking the Data classes from Table 4.2, Table 4.3, Table 4.4 and Table 4.5 that show the available columns after the aggregation layer,
3. Defining which columns are missing from the data's aggregation layer that are needed as inputs for the geometrical transformation function to plot the visual component to the canvas.

**Example:**

In case of scenario **S1** the column of Visual components from Table 5.1 can not be created alone from the data structure that holds: user_id, date, tutorial_started, tutorial_max_completed_step, tutorial_completed, tutorial_drop out_step and lengths _array.

The following additional columns are calculated:

1. **player_type**:

```
    if [Tutorial Completed] > DATE('2021-01-01')
    then 'Completer'
    else if [Dropout Step] > 0 then 'Droppers'
    ELSE 'Starters'
    end END
```

2. **count_started**:

```
    COUNT([Tutorial Started])
```

3. **droppers**:

```
    if [Player type]= 'Droppers'
    then 1
    END
```

4. **Step &param (&param = 1,2,3,4)**:

```

    if [Tutorial Max Completed Step] >= \&param
    then 'Step␣\&param'
    end
```

5. **Step &param% (&param = 1,2,3,4)**:

```
    COUNT([Step &param])/[count_started]
```

6. **Step &param droupout (&param = 1,2,3)**:

```

    if [tutorial\_dropout\_step] == &param then
    'Drop&param' end
```

The same steps were followed for augmenting for Visualization-Fit format for scenario **S2**, **S3** and **S4** and are attached in the A.

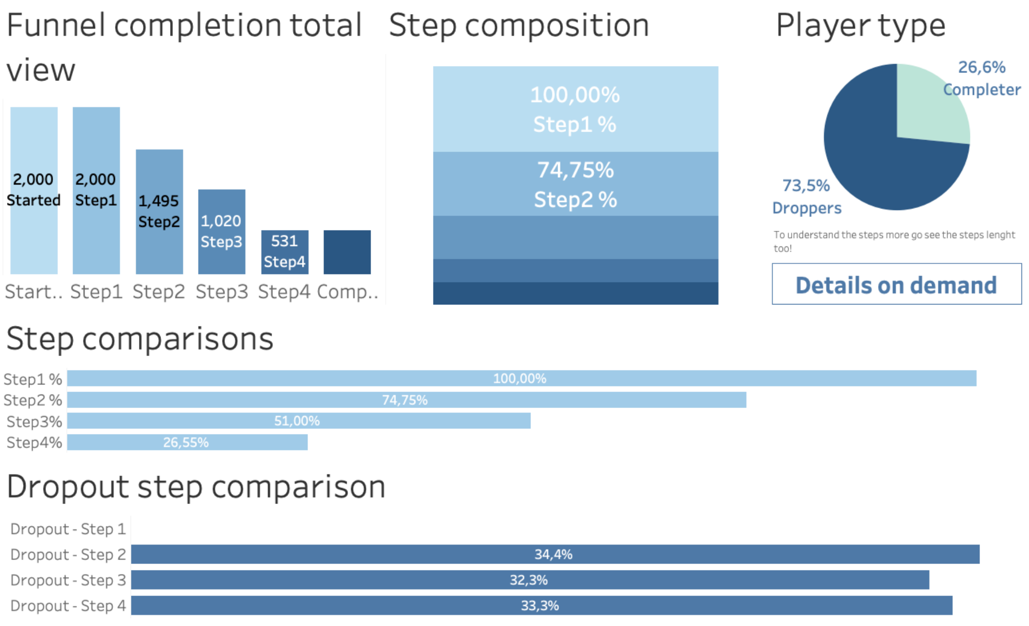
# 5.4 Creating the Dashboard

## 5.4.1 Scenario 1

After careful derivation of the visuals components and the aggregation of data as well as augmentation of it we built the dashboards using the Tableau software and the core of the view trees for view coordination that were proposed in chapter 3.

Along the information seeking ”overview - details on demand” mantra we started with creating the dashboard that we just refer to as the static dashboard of the specific scenario. This static dashboard with the two parts, overview and details, is shown in Figure 5.7 and Figure 5.8 respectively.

The final goal being testing our hypothesis about whether static or interactive visual solutions perform and work better, we continued the implementation by following the visual analytics mantra of “analyse first - show the important - zoom, filter and analyse further - details on demand”. This interactive dashboard is shown in Figure 5.9 and Figure 5.10 respectively.



To understand the steps more go see the steps lenght too!

Figure 5.7: Overview part of the static dashboard for **S1** following the view graph drawn on Figure 5.1.

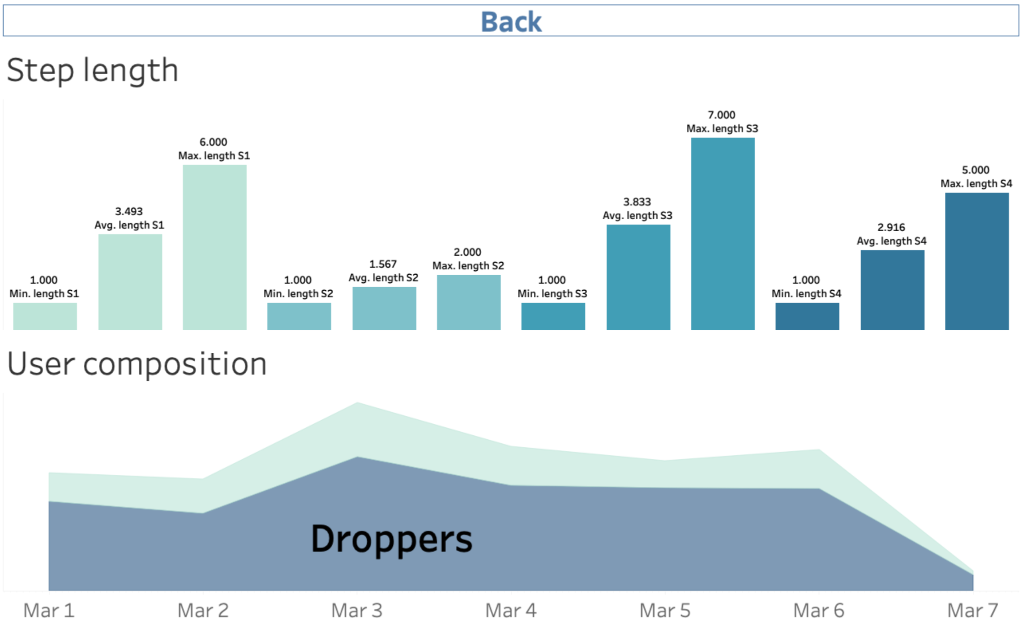


Figure 5.8: Details on demand for the static dashboard of **S1** following the view graph drawn on Figure 5.1.

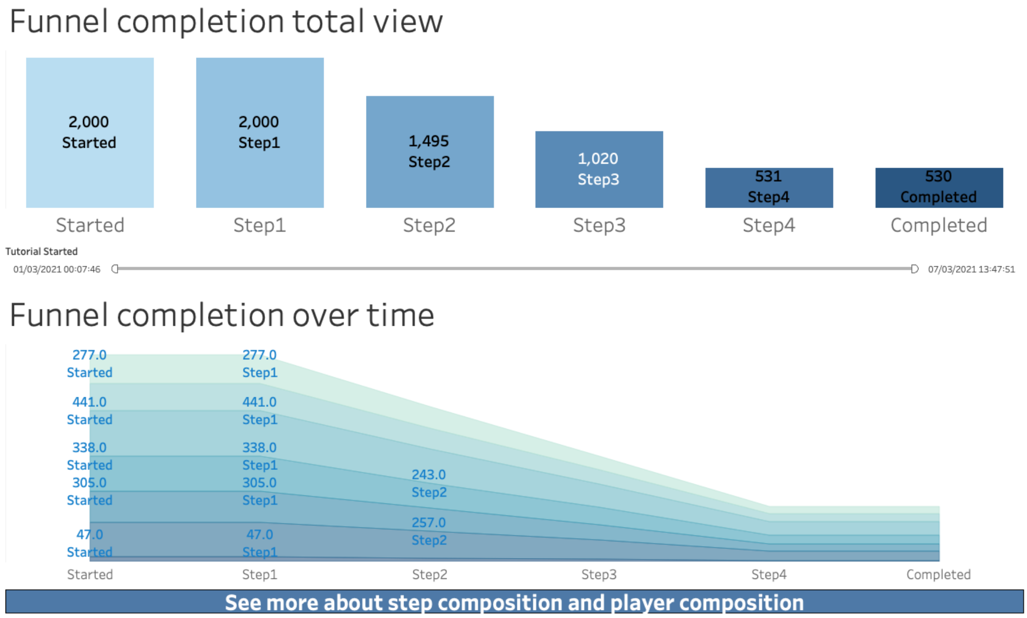


Figure 5.9: "Analyse first" part of the interactive dashboard for **S1** following the view graph drawn on Figure 5.2.

### 5.4.2 Scenario 2

As done in case of scenario **S1**, for implementing the dashboard by following the proposal steps we started with the static version as seen in Figure 5.13 and Figure 5.14

On the other hand, the interactive dashboard to scenario **S2** is shown in Figure 5.15 and Figure 5.16 respectively.

### 5.4.3 Scenario 3&4

As mentioned before, due to the contextual similarities between scenarios **S3** and **S4** the dashboard creation has been merged and only the data pipeline was kept separately. Thus, both the static and interactive versions to the dashboard were created by merging the two scenarios in one solutions including two separate data sources.

Therefore, the static, information seeking version of the dashboard resulted in the solution shown in Figure 5.17 and Figure 5.18.

The interactive dashboard to scenario **S3** and **S4** is shown in Figure 5.19, Figure 5.20, Figure 5.21 and Figure 5.22 respectively.

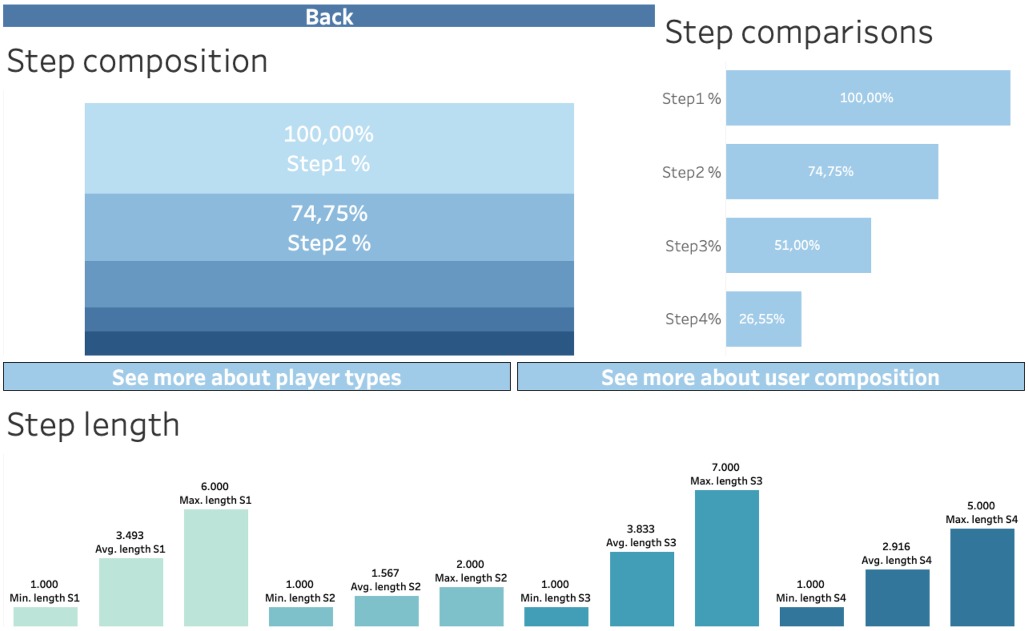


Figure 5.10: ”Show the important” followed by the ”analyse further” part of the interactive dashboard for **S1** following the view graph drawn on Figure 5.2.

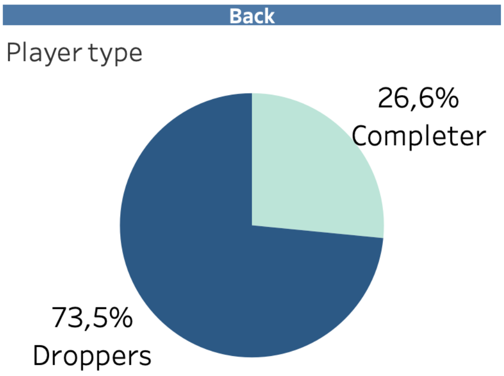


Figure 5.11: ”Details on demand” about player type of the interactive dashboard for **S1** following the view graph drawn on Figure 5.2.

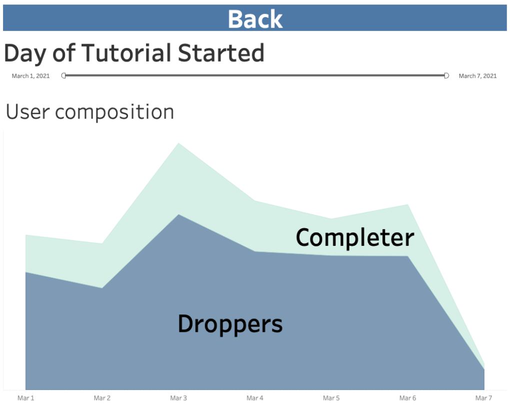
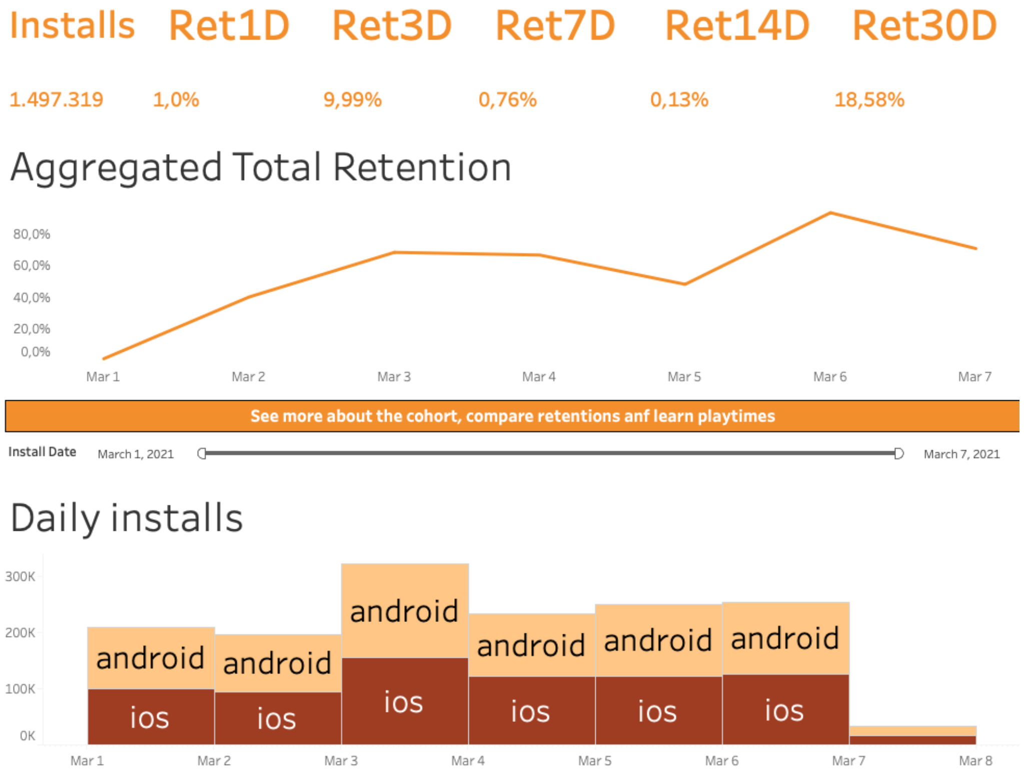


Figure 5.12: ”Details on demand” about user composition of the interactive dashboard for **S1** following the view graph drawn on Figure 5.2.



| Installs | Ret1D | Ret3D | Ret7D | Ret14D | Ret30D |
|---|---|---|---|---|---|
| 1.497.319 | 1,0% | 9,99% | 0,76% | 0,13% | 18,58% |

Aggregated Total Retention

See more about the cohort, compare retentions anf learn playtimes

Install Date March 1, 2021 March 7, 2021

Daily installs

Figure 5.13: "Overview" part of the static dashboard for **S2** following the view graph drawn on Figure 5.3.

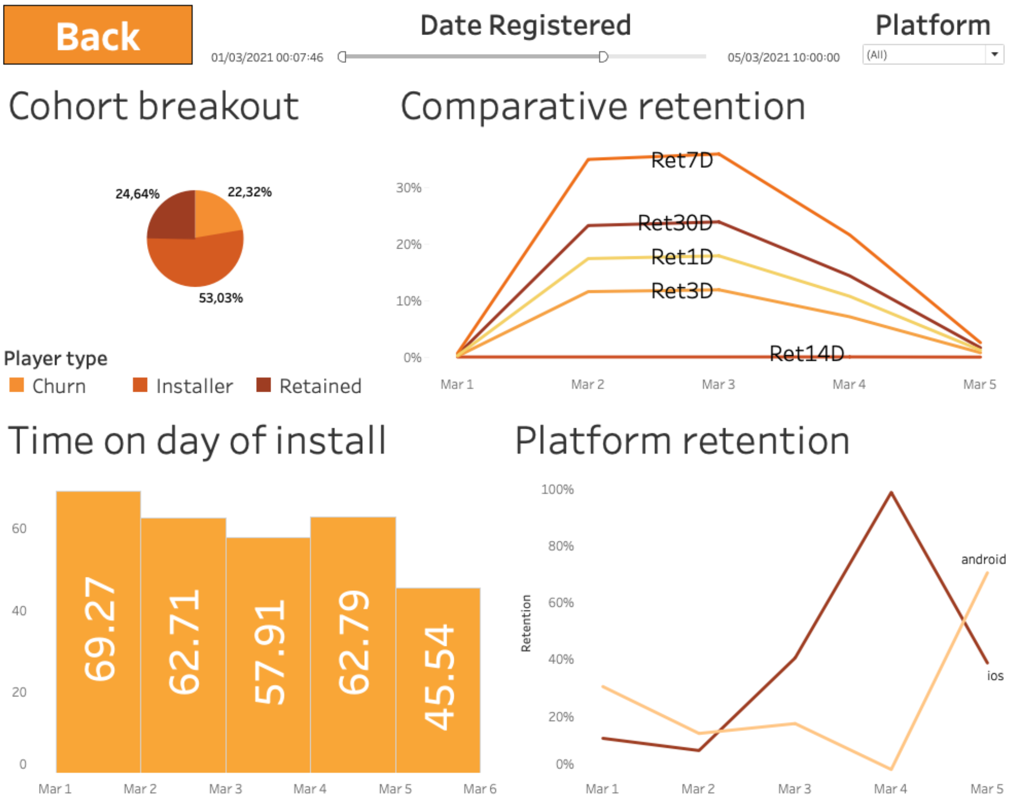


Figure 5.14: "Details on demand" part of the static dashboard for **S2** following the view graph drawn on Figure 5.3.

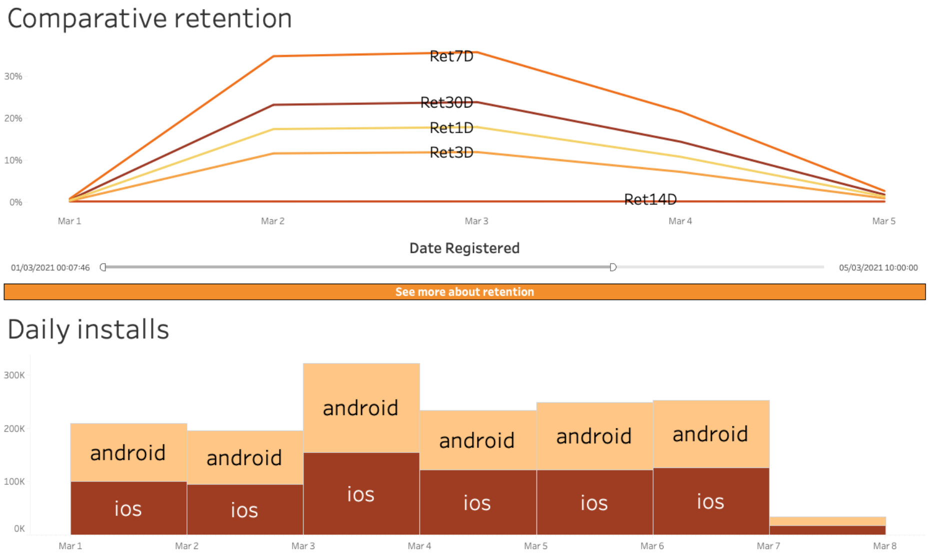
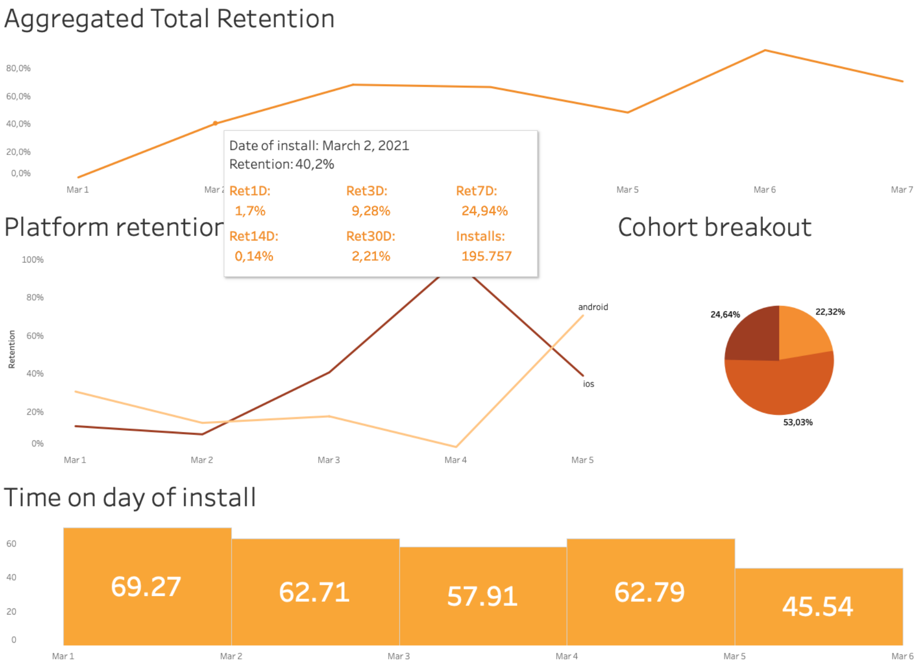
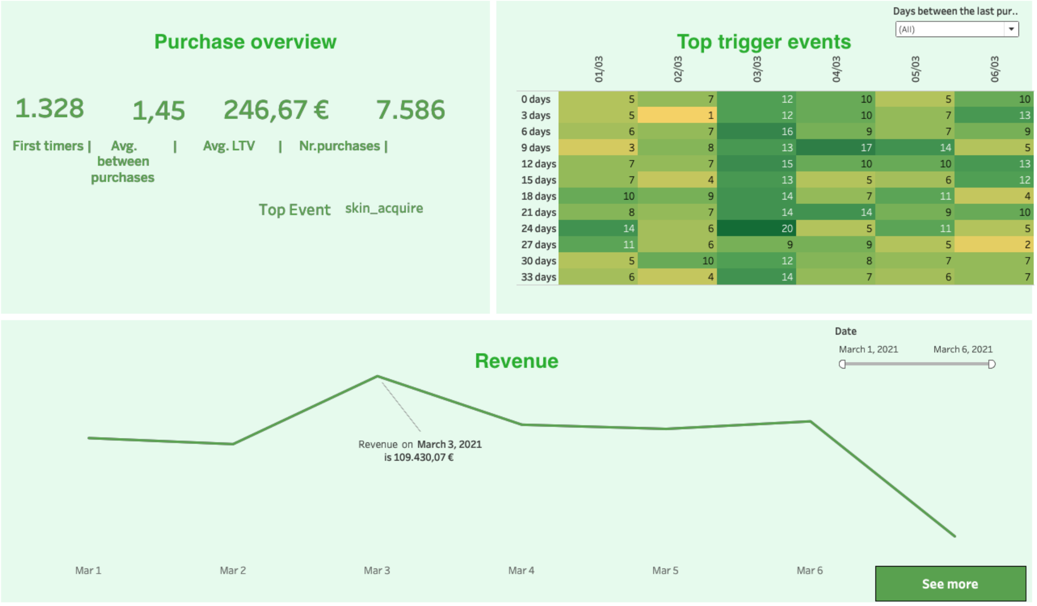


Figure 5.15: ”Analyse” part of the interactive dashboard for **S2** following the view graph drawn on Figure 5.4.


Date of install: March 2, 2021
Retention: 40,2%

| | | |
|---|---|---|
| Ret1D:<br>1,7% | Ret3D:<br>9,28% | Ret7D:<br>24,94% |
| Ret14D:<br>0,14% | Ret30D:<br>2,21% | Installs:<br>195.757 |

Figure 5.16: ”Show important - analyse further” part of the interactive dashboard for **S2** following the view graph drawn on Figure 5.4 with the ”details on demand” in tooltip.



Purchase overview

1.328 | 1,45 | 246,67 € | 7.586
First timers | Avg. between purchases | Avg. LTV | Nr.purchases |

Top Event skin_acquire

Top trigger events

Days between the last pur..

(All) ▾

| | 01/03 | 02/03 | 03/03 | 04/03 | 05/03 | 06/03 |
|---|---|---|---|---|---|---|
| 0 days | 5 | 7 | 12 | 10 | 5 | 10 |
| 3 days | 5 | 1 | 12 | 10 | 7 | 13 |
| 6 days | 6 | 7 | 16 | 9 | 7 | 9 |
| 9 days | 3 | 8 | 13 | 17 | 14 | 5 |
| 12 days | 7 | 7 | 15 | 10 | 10 | 13 |
| 15 days | 7 | 4 | 13 | 5 | 6 | 12 |
| 18 days | 10 | 9 | 14 | 7 | 11 | 4 |
| 21 days | 8 | 7 | 14 | 14 | 9 | 10 |
| 24 days | 14 | 6 | 20 | 5 | 11 | 5 |
| 27 days | 11 | 6 | 9 | 9 | 5 | 2 |
| 30 days | 5 | 10 | 12 | 8 | 7 | 7 |
| 33 days | 6 | 4 | 14 | 7 | 6 | 7 |

Revenue

Date

March 1, 2021 March 6, 2021

Revenue on March 3, 2021 is 109.430,07 €

Figure 5.17: "Overview" part of the static dashboard for **S3** and **S4** following the view graph drawn on Figure 5.5.

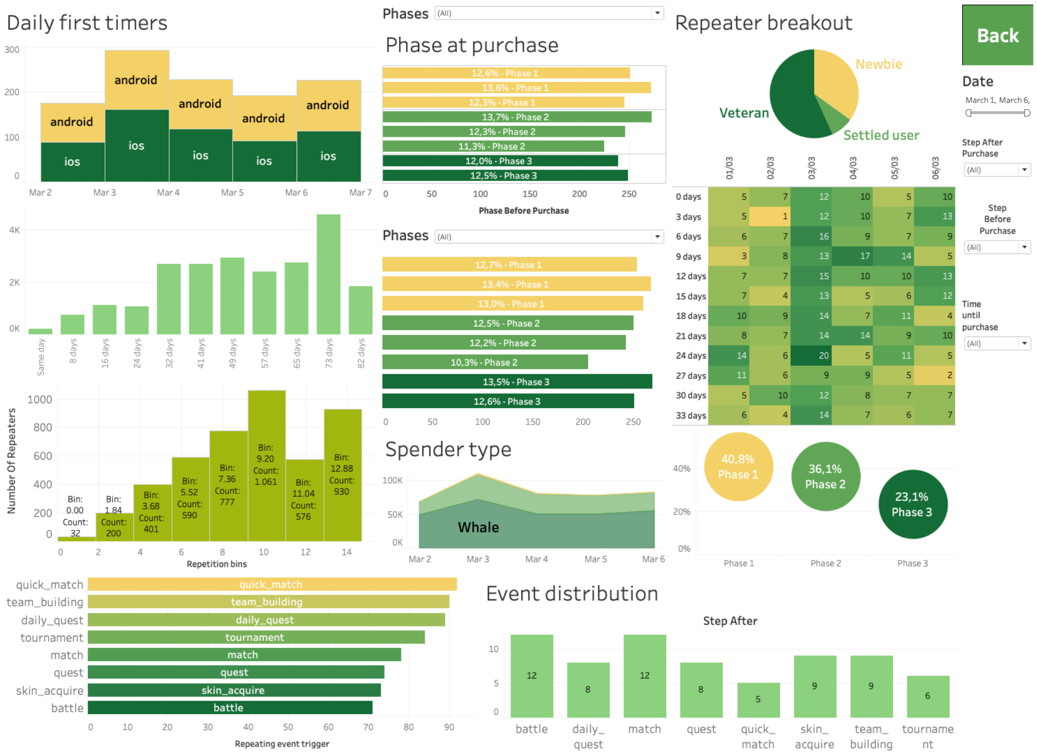


Figure 5.18: "Details on demand" part of the static dashboard for **S3** and **S4** following the view graph drawn on Figure 5.5.

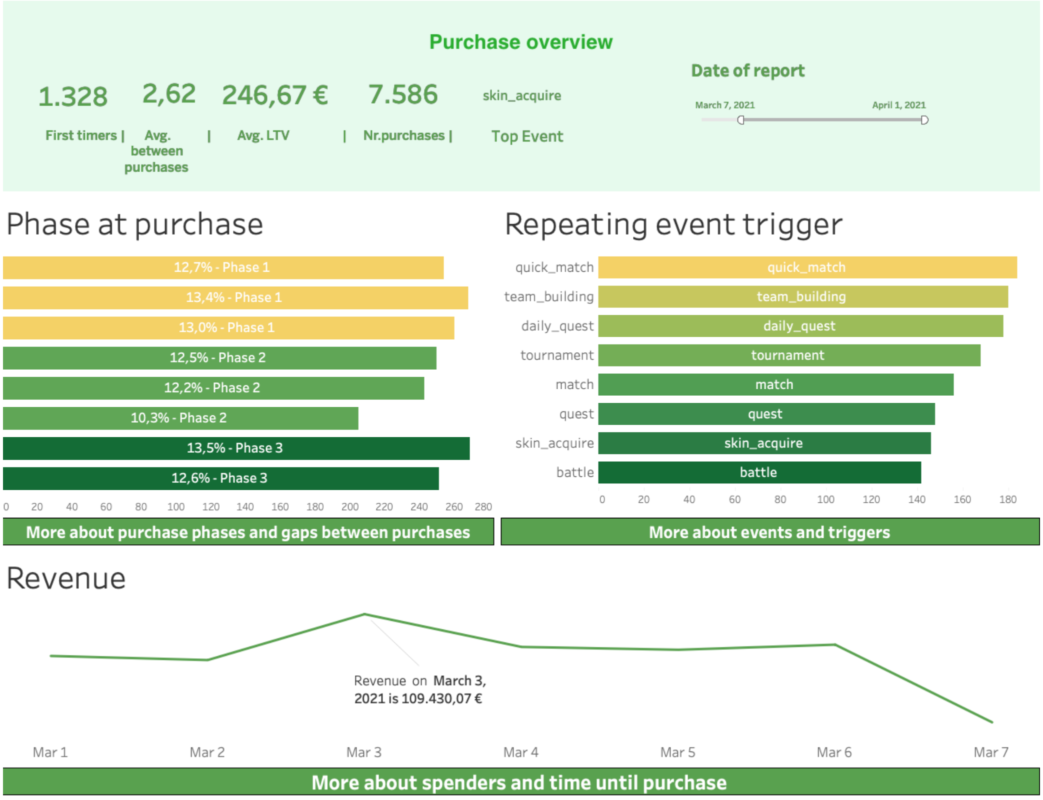
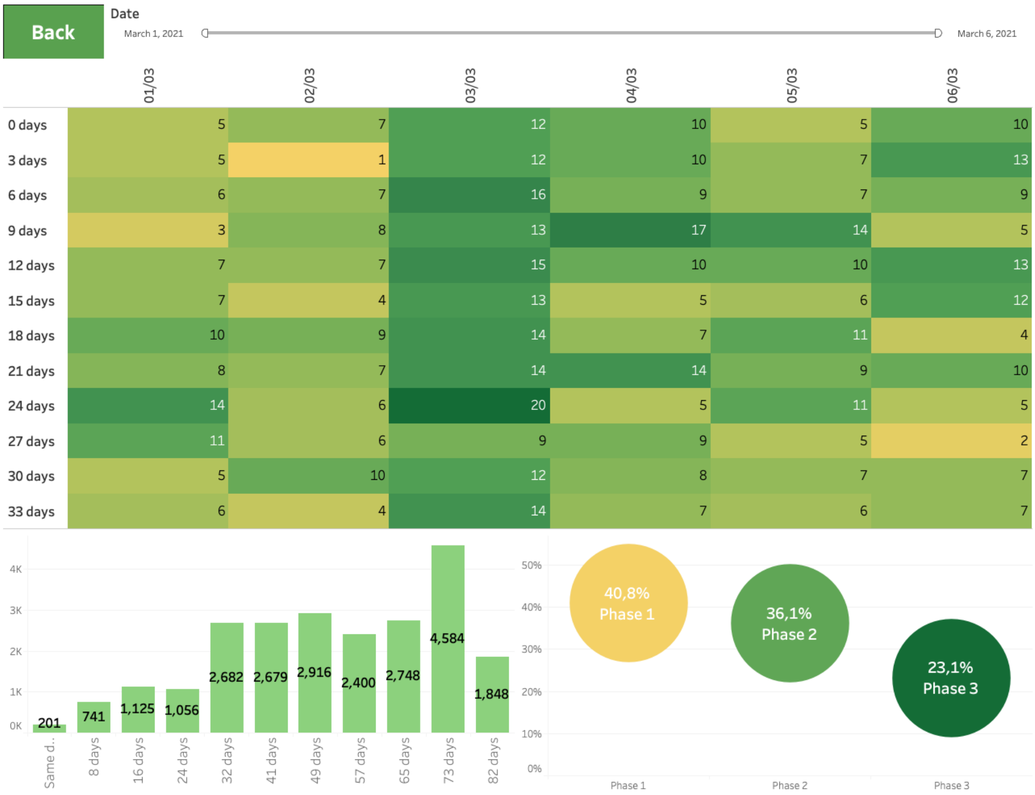


Figure 5.19: "Analyse - show the important" part of the interactive dashboard for **S3** and **S4** following the view graph drawn on Figure 5.6 including subsequently the "show the important" part too.



| | 01/03 | 02/03 | 03/03 | 04/03 | 05/03 | 06/03 |
|---|---|---|---|---|---|---|
| 0 days | 5 | 7 | 12 | 10 | 5 | 10 |
| 3 days | 5 | 1 | 12 | 10 | 7 | 13 |
| 6 days | 6 | 7 | 16 | 9 | 7 | 9 |
| 9 days | 3 | 8 | 13 | 17 | 14 | 5 |
| 12 days | 7 | 7 | 15 | 10 | 10 | 13 |
| 15 days | 7 | 4 | 13 | 5 | 6 | 12 |
| 18 days | 10 | 9 | 14 | 7 | 11 | 4 |
| 21 days | 8 | 7 | 14 | 14 | 9 | 10 |
| 24 days | 14 | 6 | 20 | 5 | 11 | 5 |
| 27 days | 11 | 6 | 9 | 9 | 5 | 2 |
| 30 days | 5 | 10 | 12 | 8 | 7 | 7 |
| 33 days | 6 | 4 | 14 | 7 | 6 | 7 |

Figure 5.20: ”Analyse further” part with ”details on demand” of the interactive dashboard for **S3** and **S4** following the view graph drawn on Figure 5.6 showing the sub-tree for *phase_at_purchase_view* node.

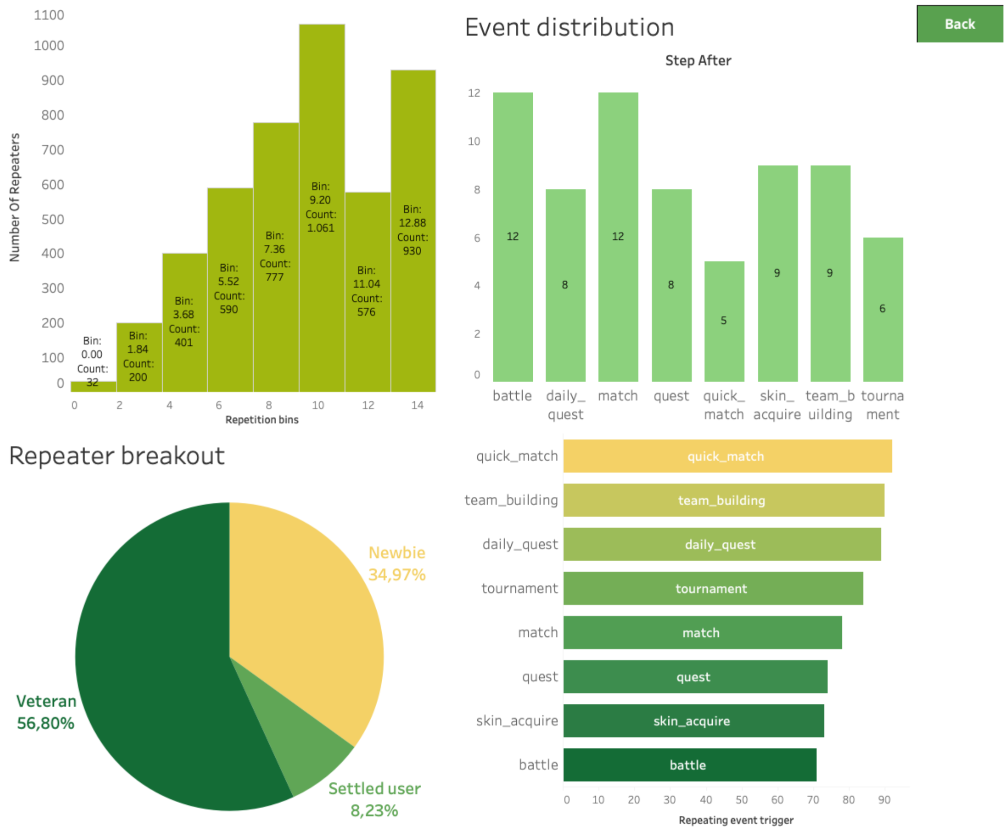
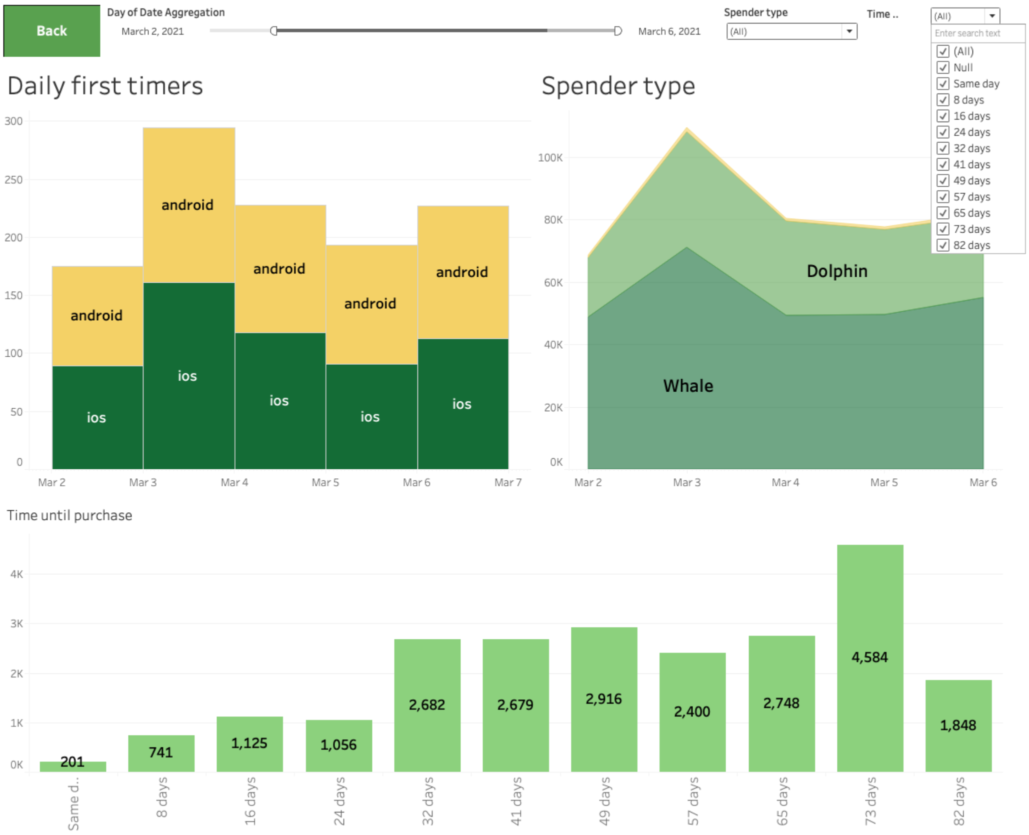


Figure 5.21: "Analyse further" part with "details on demand" of the interactive dashboard for **S3** and **S4** following the view graph drawn on Figure 5.6 showing the sub-tree for *revenue_view* node.



Back

Day of Date Aggregation
March 2, 2021 — March 6, 2021

Spender type (All)

Time .. (All)

Enter search text

☑ (All)
☑ Null
☑ Same day
☑ 8 days
☑ 16 days
☑ 24 days
☑ 32 days
☑ 41 days
☑ 49 days
☑ 57 days
☑ 65 days
☑ 73 days
☑ 82 days

Figure 5.22: "Analyse further" part with "details on demand" of the interactive dashboard for **S3** and **S4** following the view graph drawn on Figure 5.6 showing the sub-tree for *top_trigger_events_view* node.

# Chapter 6

# Evaluation

To evaluate our solution and implemented system we already proposed evaluation methods in the chapter 4. Evaluating visualization tools is considered a two-sworded topic due to the cases' veracity and intangibility that every different visualization is created for. Indeed, there are several guidelines to follow and best practices to implement, the evaluation becomes in any case subjective and leads to a low incidence rate of papers in the human-computer interaction topic with any kind of evaluation included.

In section 4.2 along with our proposal to tackle the case study we presented which evaluation methods we think are well suited for our visualization system on method level. Through this evaluation we intended to test the choices of methods and the work progress overall, however apart from these 2 evaluations done through the pattern based approach that are system level evaluations, we also conducted a task focused evaluation too including test users.

## 6.1 System level evaluation

The method for our system level evaluation was presented in section 4.2, particularly the pattern based approach as proposed by Elmqvist et al in [11]. This approach, through the exploration and presentation pattern, enabled us to run a continuous and iterative evaluation of the work as wells as to keep a conservative mind with regards to the final presentation of the work.

### 6.1.1 System evaluation through exploration pattern

Therefore, with exploration pattern we focused on the design space in the early stage of visual component derivation. The exploration pattern raises questions for evaluation such as "Are the correct independent and dependent variables depicted?" and "Are the right questions asked?".

Hence, our evaluation with exploration pattern first focused on evaluating whether the derivation of questions align the scenarios. For this the evaluation is rather subjective. As only the scenarios were given to the use case the questions were derived from a top-to-bottom approach with a limit on the number of them for the feasibility of implementation during the work-time. Thus, the way the questions take shape are of that they always start with trying to get a general idea about the scenario. This approach was chosen as intuitive and as an intended opposite direction to the general bottom-up model which assembles questions for higher-level tasks from many low-level ones and which is not proven to be a valid and effective approach [20]. Evaluating this is impossible by quantitative measures, but we believe that if a question cannot be answered after walking through the visualization and finding the right view to look at then the question is not correctly derived or not correctly formulated. In order to test this, we had to integrate this aspect in the user level task-focused evaluation which done by supervising the users clears the doubts around how well questions are formulated. Therefore, we report the questions that raised further questions and doubts:

1. Question **Q2** from scenario **S2**, introduced doubt and even fear about failing to support the answer with reasonable argument for the significance level. This question being more open and not looking for a numerical value shook a user who asked how should the significance be measured. The question in that case clearly failed to deliver to the user the message and deteriorated the cognitive task performance. The answer as a result was telling that there's not enough knowledge for an answer to be reported, even though the dashboard indicated in multiple views the same trend of retention drop of magnitudes. However, the answer could have been a "It is really significant as the game looses more than half of the players over time, and game developers should try to work for making their game more attractive or more marketing should be done", this question failed to lead the user in this direction.
2. Question **Q3** from scenario **S2**, raised some doubts, but compared to the previous case it was not as severe, as even though it raised doubts, it still lead to accurate response. However, the fact that 2 users got

confused on the ”how long” part of the sentence, as it did not include time unit (time unit was shown on the view only) made it already clear that not even this extent of doubt should be allowed on question level as in both cases this made the chain of action taken longer than needed.

3. Question **Q4** from scenario **S2**, raised doubts and further question about the percentage asked in the original question, as it was not clarified. A user asked whether the percentage should be calculated with all the users that ever played and then taking the subset of players that ever match the constraint of having played on the same day, or it should be calculated with the daily installers as total value and then the ones that match the constraint. Even though the visualization supported this question with clear readable value and there was no need for calculation, the user got confused and distracted as it immediately tried to read the two specific values and calculate the percentage. We assume this happened as the user lacked the domain knowledge of what eventually is attempted to find when learning game player behaviour with regards to the frequency of playing. It appeared as the lack of domain knowledge excluded business value from the thought process and steered it to a non-contextual chain of actions.

4. Question **Q3** from scenario **S3**, is similar to the **Q2** from scenario **S2** as it is more open, not looking for a clear numerical value as response, these have the same nature as questions and carry a high business value within themselves. However, the question made a user deteriorate from a clear cognitive process and lead to an interaction with the dashboard to the most extent possible, even on parts that did not serve any detail or insight to answering this question. Even though the question received useful answers in both static and interactive case, there was a difference in the confidentiality level of responses. As a result we saw that such open question with a good design can be still responded, but with a good interaction coordination and well thought out view graph can increase the confidence level of the response.

### 6.1.2 System evaluation with presentation pattern

The presentation pattern for visualization evaluation focuses on providing guidance for a clear and efficient way to deliver the communication message. Shneiderman et al. in their work underline the efficiency of delivering visual solutions through case studies as they have the power to provide in-depth insights about how the visualization techniques and proposed methods can be used in different or even day-to-day situation [27].

In our case the case study was already given as the proposal was also presented along with it. From the presentation pattern point of view we can say that our final visual solutions are in line with what this pattern tries to accomplish. Hence, our case study presentation allows listener to understand the pipeline provided and see that eventually it can be applied with slight modifications in different contexts too.

From another perspective of the visualization pattern we can evaluate our solution along Tufte's design principles. According to this, our solution meets great evaluation on the principle of maximizing data ink ratio, showing data above everything else and showing the truth in data without chartjunk. These are supported by the design process of every individual view, and as the results also show, redundant axis names (for example of date on x axis) are dropped, unnecessary grid-lines are removed, colors and shapes are kept simple and data is shown at a very high density (it is also outcome of visualizing big data).

## 6.2 Task focused evaluation

In order to asses whether the tasks could indeed be carried out we asked 8 people with different backgrounds to use one of our solutions for each of the scenarios and respond to a question that we derived. These 8 subjects have different background varying from mechanic through accountant manager to engineer and to data scientists. This selection was done such that not only tech savvy or data people get to use the tool, therefore really allowing a stark focus on cognitive task evaluation.

The way the assessment was performed was such that each individual received for each scenario one question but whether the solution they got to use was the static or the interactive depended on mutual exclusivity within the other individuals' received questions and dashboard types. This can be seen in Table 6.1 along with the outcome of the tasks, whether they could perform it and the response was acceptable or the response was definitely not correct.

### 6.2.1 Evaluating the support for tasks

The visual components were chosen mostly based on Figure 2.10, with minor changes whenever was possible towards a more simplistic option, for example bar chart if possible. This was done aiming that the precision in delivering results might be enhanced, however this did nor show clear results or indications in the evaluation with the users.

As a result to the user involved, task focused evaluation the Table 6.2 shows how the different solutions performed amongst the users involved in the evaluation.

| **Users** | **Scenarios** | | S1 | S2 | S3 | S4 |
|---|---|---|---|---|---|---|
| | M. | Q | Q2 | Q3 | Q4 | Q2 |
| | | i/s | s | s | i | i |
| | | resp. | y | y | y | n |
| | D. | Q | Q4 | Q4 | Q1 | Q1 |
| | | i/s | i | i | s | i |
| | | resp. | y | n | y | y |
| | E. | Q | Q1 | Q2 | Q3 | Q1 |
| | | i/s | s | i | s | s |
| | | resp. | y | n | n | y |
| | C. | Q | Q3 | Q4 | Q1 | Q1 |
| | | i/s | s | s | s | s |
| | | resp. | y | y | y | y |
| | A. | Q | Q4 | Q3 | Q3 | Q2 |
| | | i/s | s | i | i | s |
| | | resp. | y | y | y | y |
| | N. | Q | Q2 | Q2 | Q2 | Q2 |
| | | i/s | i | s | s | i |
| | | resp. | y | n | n | y |
| | O. | Q | Q3 | Q1 | Q2 | Q1 |
| | | i/s | i | i | i | i |
| | | resp. | y | y | y | y |
| | A. | Q | Q1 | Q1 | Q1 | Q2 |
| | | i/s | i | s | i | s |
| | | resp. | y | y | y | y |

Table 6.1: Evaluating whether the users could perform the cognitive tasks of the respective question and scenario. S: scenario, Q: question, I: interactive dashboard, S: static dashboard, resp.: response, y/n: yes/no. The users with their initials coloured have a technical background: blue - technical background, green - technical background + experience in business analytics and Tableau.

Table 6.2 shows an overall performance of the visualizations created for the

| **Scenario 1** | | **Scenario 2** | | **Scenario 3** | | **Scenario 4** | |
|---|---|---|---|---|---|---|---|
| i | s | i | s | i | s | i | s |
| 0 | 0 | 2 | 1 | 0 | 2 | 1 | 0 |

Table 6.2: Evaluating the performance difference of the interactive and static solutions.

scenarios, where under the *i* column we count for how many questions the interactive version of the specific scenario did not provide clear answer. On the other hand, the *s* column counts the times the static visualization failed in providing response.

The points where the **interactive** dashboard failed:

1. On scenario **S1** and question **Q2**, the task could be performed and users where well aware what to look for and what action to take, but the *accuracy* dropped as some reported the value of another day than the one that was randomly asked.
2. On scenario **S2** and question **Q4**, the lack of abbreviation meaning of *Ret1D* mislead users to believe that is the retention of the first day of install.
3. On scenario **S4** and question **Q2**, the column labeling and the visual element having the *frequency* of gaps color coded mislead the users to say the times the the specific gap happened rather than the size of the gap.

The points where the **static** dashboard failed:

1. On scenario **S2** and question **Q2**, user failed to find the view that is the one that could help them due to the availability of retentions in percentages in data table in comparative view and in aggregated view. Thus the response given to the question was imprecise.
2. On scenario **S3** and question **Q2**, user failed to walk through the system and ask find this information that has to be derived from the view and not explicitly read out. As it failed finding in first place the view or source from where to deduct an answer the dashboard failed to meet effective cognition.
3. On scenario **S3** and question **Q3**, user failed to find a view that can support whether the users are likely to become repeaters, leading to a vague guess as answer with low confidence.

## 6.3 User focused evaluation

Table 6.1 serves more purposes than reporting the task level performance, it also shows how individuals along with their background meet the request to answer the business questions.

Thus, we can see that there is a user (letter N.) who fails to respond using the static dashboards but find responses in the interactive ones. This can simply indicate that the user, based on the technical background, is more acquainted with interacting with tools and systems, and is prone to overlook answer that is given in a more straightforward way.

Another user (letter M.) fails to answer a question on the same interactive version for scenario **S3** and **S4**, while being able to answer another one on the same dashboard. This shows that the count of failures on each dashboard might not reflect the true accuracy of them as the accuracy can be deteriorate on question and view level, not necessarily on dashboard level. Based on this, we cannot say that the interactive dashboard failed for user M. therefore as a global solution is not good, but rather say that a particular question was not easy to answer either due to the view choice or to the view management implemented.

Moreover, we can also see in case of user A. coloured with green the fact that the domain knowledge and experience translated in correct navigation on the interactive dashboards and to a fully correct answer on both solutions. This just further supports the need for definitions and context setting with a focus on business value understanding to improve the quality of answers.

Moreover, another observation is that regardless of the background, bot technical and non-technical users can fail on the interactive dashboards. Even more, they also reported after finishing the evaluation a preference for the static dashboards.

Furthermore, when the users who received both interactive and static versions (6 users out of 8, exceptions are users C. and O.) 5 out of 6 said that they preferred using the static dashboards.

Lastly, the user level evaluation also shows that the most errors were related to scenario **S2**, which indicates that the topic of retention is a more difficult and harder to understand for various users. This is further supporting the need for implementing in dashboards components for achieving common base of domain knowledge.

# Chapter 7

# Discussion

In our work we aimed at testing interactive and static visual solution that were created along our proposed visualization pipeline on a given use case. Along our work of implementation as well as evaluation there were points where we had to choose methods over others and we also had to understand the way the dashboards perform in the hands of users. This naturally led to discussion about the limitations our system has. In this section we present along some general limitations the ones that were derived from the user involved evaluation and then propose ideas or points for improvement.

## 7.1 General limitations

As mentioned before, we noticed limitations to our system already during the implementation. Some happened as side effect to method choices and some happened unintended. The greatest limitation of our implemented system is that is focused on testing our hypotheses in a very specific context of the use cases. Even though we aimed at presenting in the thesis our work in a structured way that can be taken as it is and applied for other use cases, we do not know whether following the steps of our proposal as a guideline would result in efficient and well preforming visual solutions on different domains or topics.

Furthermore, an example of limitation induced by a particular method, the choice of tool, in our case Tableau, introduces limitations by not allowing for custom visuals. Even though we tried to select the needed visual component along a predefined mapping, we also tried to substitute these wherever possible with more simplistic visuals. This substitution could also be accounted for the confusion of views to rely on when answering some questions. This is especially due to that in the first phase when we selected the visual components we did that based on the individual cognitive tasks and then aimed for choosing the more simplistic option if available. In this phase we considered the cognitive tasks and questions independent from each other, but in the final solution we had to realize that there is overlap in some cases and there might be better combined visuals that in one could support multiple questions or cognitive tasks. Since we did not cover dependence between questions, we believe this could have been the cause why some users confused two similar views, namely the one for the gaps between two purchases and the one about the time until the first purchase. If these two would have been in a single view, providing answer for two questions the confusion could have been avoided.

On another hand, we believe that by focusing on testing a hypothesis and placing interactive and static solutions in parallel we could not have achieved a perfect or highly efficient of any of the two versions, but rather show which is more prone to errors. This, even though it was needed to test our hypotheses, could deteriorate the design choices individually, since these became intertwined rather then version focused. By this we mean that if we had only presented a proposal for an interactive-only dashboard then some design choices could have been different. For example, in this case we could have focused on what filters and in what way are they proven to work the best rather then just using the needed ones and implementing them with the goal of seeing whether they are going to be used at all or not.

## 7.2 Limitations derived from user evaluation

Amongst others the evaluation with test users allowed some insights to the performance of dashboards from another point of view rather then pure visual or task based approach. This relates mostly to the background knowledge and knowledge of the terminologies used in gaming, which seem to need more clarification. Moreover, a big learning point is for example when the word ”comparative” is used the question of what it compares to will be raised. The word ”comparative” appeared in a label along with many other lines of the same metric that is represented with percentage value over a time period. In this case the question whether the comparison goes along the individual line of days and compares to the previous day, or if it compares to the other lines around will for sure be asked. Such questions about the context and knowledge of the domain if are raised and the visual solution lacks to provide hints of such are going to deteriorate the accuracy of answers given to the tasks. This is caused by the doubts that arise and it leads to drop in accuracy of responses or even business decisions. Furthermore, accuracy drop happens with a higher chance when doubt arises even though this does not exclude the correct performance of the cognitive task.

Therefore, we can derive that in the case when background knowledge might not be on the same level amongst the various users the visualization's performance is limited and the evaluation of views, coordination and layouts become obsolete as the bottleneck in the cognition processes are induced by insufficient background knowledge of the domain.

Another limitation derived from the evaluation relates to a specific scenario, namely the scenario **S2**. As the outcome of the evaluation showed, the most failed and confusion introducing questions belong to scenario **S2**. It is also true that opposed to the other scenarios, this was the most domain specific, asking about game retention which itself is alone considered a tough topic among game analyst. This is further supporting the argument that knowledge of background and context, terminology and the understanding of the potential business value of such questions are vital to the performance of the visualization, regardless of it's static or interactive quality.

Lastly, the fact that some views were derived as duplicates, meaning that one view could have supported 2 cognitive tasks, introduced confusion between the users. Therefore, the derivation of views along cognitive tasks first should examine the relationship between data attributes and between cognitive tasks to avoid redundant view generation.

## 7.3 Points of improvement

The limitation of our solution being case specific can be overcome and considered applicable in a more general setting by taking into account the dependency between questions and similarities between cognitive tasks. This could improve not only in creating a more straightforward and efficient visualization but also for creating a more accurate and more acceptable one in terms of operations.

Another limitation of the background knowledge can be overcome with the use of hints and clear definitions indicated in a prime view, either as a hint box or as text boxes around the specific view this can be overcome. This, however should be accompanied by a clear scenario presentation and introduction of the business context since the purpose of the business information visualization is to deliver support for valuable business decision making. In case the end user lacks the domain knowledge it should be put in prime focus the introduction and understanding of the business value that the visualization carries. This focus can help deepen in the end user the relevance and importance of their response that they will give.

# Chapter 8

# Conclusions

In our work we presented a proposal for creating a business information visualization pipeline that supports both interactive and static dashboard creation for scenarios of our use case. We did this by understanding the related literature and focused our pipeline on considering human perception and cognition similarly to how it is done in scientific visualization [5]. Moreover, we designed the pipeline with ”data representation - task fit” [34] in mind. Apart from proposing in our constructive work a solution for the specific use case, we also took into account the visual intelligence and followed the mapping of business information visualization elements to visual IQ dimensions as Bačić et al. [5] showed.

The goal was to test hypotheses about whether static or interactive dashboards deliver more accurate responses, and therefore more reliable business decisions, and whether interactive or static dashboards are more accepted in a wide variety of users.

In order to find answers to these hypotheses we conducted an evaluation involving eight end-users with different backgrounds varying from data scientist to account manager.

**A1** From our evaluation, we learned that the proposed static dashboards can fail just as much as the interactive ones. However, we have also seen that the reason for failure was different in every scenario. This highlighted the fact that the answer to **H1** is that both solutions can work, however it is primarily question dependent and secondly view management dependent whether the outcome will be accurate or will fail.

**A2** For **H2** we found out that static dashboards are preferred over the

> interactive ones regardless of the outcome of the answers, and we also saw a very low tendency of interaction with the visualization from the users involved in our evaluation.

Therefore, we learned that when a visualization solution needs to be implemented, to support business decision making, then careful consideration should be on: 1.) deriving clear and understandable questions, 2.) deriving the visual components using mappings that take cognition into account, 3.) checking dependencies between data attributes to avoid duplicate view creation for similar purposes, 4.) understanding the target end-user and the domain knowledge level.

In this thesis we contributed, in a constructive way, by showing what additional points should be considered in designing efficient, accurate and company-wide accepted visual solutions.

We recommend companies to prioritize operational effectiveness and utility, because as seen with our proposed pipeline, accuracy and task efficiency only do not imply acceptability of the visualization. However, when acceptability is met, accuracy and task efficiency are easier to achieve along the way.

To conclude with, we can definitely state that the literature about interactive visualization is holding significant value, but the implementation of it appears to be more error-prone in the business intelligence environment. Thus, further research topics to this thesis could focus on interaction management in business information visualization and acceptability in business information visualization and hopefully provide answers to why this error-prone tendency is present in this environment.

# Bibliography

[1] Business Intelligence: What It Is & Its Importance — Tableau.

[2] *System Acceptability.* John Wiley  Sons, Ltd, 2005, ch. 6, pp. 46–58.

[3] Ayalasomayajula, A. V. Everything you need to know about kpi visualization - atlan — humans of data, 2016.

[4] Baldonado, M., Woodruff, A., and Kuchinsky, A. Guidelines for using multiple views in information visualization. *Proceedings of the Workshop on Advanced Visual Interfaces* (05 2000).

[5] Bačić, D., and Fadlalla, A. Business information visualization intellectual contributions: An integrative framework of visualization capabilities and dimensions of visual intelligence. *Decision Support Systems 89* (9 2016), 77–86.

[6] Brenner, M. Visual impact: Why you need to visualize your next presentation - visual matters, 8 2016.

[7] Card, S. *Information Visualization.* 01 2008, pp. 509–543.

[8] Cockburn, A., Karlson, A., and Bederson, B. B. A review of overview+detail, zooming, and focus+context interfaces. *ACM Comput. Surv. 41*, 1 (Jan. 2009).

[9] CSG. How and why visualizing data makes it actionable.

[10] Eggers, W. D., Hamill, R., Ali, A., and Hersey, J. Data as the new currency government's role in facilitating the exchange.

[11] Elmqvist, N., and Yi, J. S. Patterns for visualization evaluation.

[12] Furnas, G. W. Effective view navigation. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (New

York, NY, USA, 1997), CHI '97, Association for Computing Machinery, p. 367–374.

[13] Jason Lankow, Josh Ritchie, R. C. *Infographics - The Power Of Visual Storytelling.* Hoboken, N.J. : John Wiley  Sons, Inc., New Jersey, 2012.

[14] Jugel, U., Jerzak, Z., Hackenbroich, G., and Markl, V. Vdda: automatic visualization-driven data aggregation in relational databases. *VLDB Journal 25* (2 2016), 53–77.

[15] Lewin, K., Lewin, K., Cartwright, D., Cartwright, e., for Group Dynamics, M. U. R. C., of Michigan. Research Center for Group Dynamics, U., and Dynamics, A. *Field Theory in Social Science: Selected Theoretical Papers.* Harper torchbooks. Harper, 1951.

[16] Licklider, J. C. Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics HFE-1* (1960), 4–11.

[17] Mintrom, M. *Herbert A. Simon, Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization.* 01 2016.

[18] Munoz, J. M. *Global Business Intelligence.* Routledge, 711 Third Avenue, New York, NY 10017, 2018.

[19] Nikos Bikakis, George Papastefanatos, O. P. Big data exploration, visualization and analytics. *Big Data Research 18* (Dec. 2019).

[20] North, C. Toward measuring visualization insight. *IEEE Comput. Graph. Appl. 26*, 3 (May 2006), 6–9.

[21] Padilla, L. M., Creem-Regehr, S. H., Hegarty, M., and Stefanucci, J. K. Decision making with visualizations: a cognitive framework across disciplines. *Cognitive Research: Principles and Implications 2018 3:1 3* (7 2018), 1–25.

[22] Pappas, L., and Whitman, L. Lncs 6771 - riding the technology wave: Effective dashboard data visualization. *LNCS 6771* (2011), 249–258.

[23] Post, F. H., Nielson, G., and Bonneau, G.-P. Data visualization: The state of the art.

[24] Roberts, J. C. Kent academic repository full text document (pdf) enquiries citation for published version link to record in kar state of the art: Coordinated multiple views in exploratory visualization. 61–71.

[25] Scherr, M. Multiple and coordinated views in information visualization.

[26] Shneiderman, B. The eyes have it: A task by data type taxonomy for information visualizations.

[27] Shneiderman, B., and Plaisant, C. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization* (New York, NY, USA, 2006), BELIV '06, Association for Computing Machinery, p. 1–7.

[28] Spence, R. *Information visualization*, vol. 1. Springer, 2001.

[29] Stasko, J. Tufte's design principles learning objectives.

[30] van Alphen, A., Halfens, R., Hasman, A., and Imbos, T. Likert or rasch? nothing is more applicable than good theory. *Journal of Advanced Nursing 20* (1994), 196–201.

[31] von Engelhardt, J., Janssen, T., Scha, R., et al. The visual grammar of information graphics.

[32] West, T. G. Forward into the past. *ACM SIGGRAPH Computer Graphics 29* (11 1995), 14–19.

[33] Zannetos, Z. S. Toward intelligent management information systems.

[34] Zhang, P. Business information visualization: Guidance for research and practice, 2001.

# Appendix A

# First appendix

## A.1 Implementation of Visualization-Fit format

### A.1.1 Scenario 2

1. **retention_aggregate**:

```
    [Ret1D]+[Ret3D]+[Ret7D]+[Ret14D]+[Ret30D]
```

2. **player_type**:

```
    if [Played On Day 1] = True then 'Installer'

    else if [Played On Day 1] = True
    and [Played On Day 3] = False
    and [Played On Day 7] = False
    and [Played On Day 14] = False
    and [Played On Day 30] = False
    then 'Churn'

    ELSEIF [Played On Day 3] = True
    and [Played On Day 7] = False
    and [Played On Day 14] = False
    and [Played On Day 30] = False
    then 'Churn'

    ELSEIF [Played On Day 7] = True
    and [Played On Day 14] = False
    and [Played On Day 30] = False
    then 'Churn'

    ELSEIF [Played On Day 14] = True
    and [Played On Day 30] = False
    then 'Churn'

    ELSEIF [Played On Day 30] = True
    then 'Retained'

    end end
```

3. **player_percent**:

```
    COUNT([player_type]) / COUNTD(user_id)
```

4. **Ret &param D (&param = 1,3,7,14,30)**:

```

    COUNT([Played On Day ])/SUM([installers])
```

### A.1.2 Scenario 3&4

1. **player_type**:

```
    if DATE_DIFF([First Date Purchased],TODAY())>=60
    then 'Veteran'
    ELSEIF DATE_DIFF([First Date Purchased],TODAY())<15
    then 'Rookie'
    ELSEIF DATE_DIFF([First Date Purchased],TODAY())<60
    and DATE_DIFF([First Date Purchased],TODAY())>=15
    then 'Settled␣user'
    else 'Newbie' end
```

2. **count_steps_after**:

```
    COUNT([steps_after])
```

3. **count_steps_before**:

```
    COUNT([steps_before])
```

4. **nr_purchases_sofar_binned**:

```

    (MAX(Number Of Purchases Sofar) -
    MIN(Number Of Purchases Sofar))
    /COUNTD(Number Of Purchases Sofar)
```