University of Padova

Department of Mathematics "Tullio Levi-Civita"

Master Thesis in Data Science

Process Optimization and Automation in E-Commerce Business Operation

Supervisor
Prof. Giorgio Maria Di Nunzio
University of Padova

Co-supervisor
Matea Mutz
Mister Sandman GmbH

Master Candidate
Gaurav Anand

Student ID
1214766

Academic Year
2021-2022

"What we need to do is always lean into the future; when the world changes around you and when it changes against you — what used to be a tail wind is now a head wind — you have to lean into that and figure out what to do because complaining isn't a strategy."
— **Jeff Bezos**, founder & chairman : Amazon

Abstract

The E-commerce platform is a snowballing market and has an amazing reach among customers per year. With the immense popularity that E-commerce has gained in recent years comes the responsibility to deliver relevant results to provide a rich experience to users. Add to this the fact that nowadays companies use different E-commerce marketplace platforms to sell their products, making it difficult to synchronize the different product attributes across different categories and product-lines. To achieve this, the products and the corresponding attributes displayed to customers need to be synchronized with all platforms and correctly categorized. Incorrect product information leads to irrelevant results for users, which not only reflects poorly on the website, but can also lead to a loss of customers. Therefore, process optimization and automation can be of great benefit. This task of automation requires an understanding of the overall marketplace behavior across all the product-lines and for that we need various techniques to crawl and scrape the data from the live platforms. Once the data is collected, different types of analysis and visualization tasks are performed as needed. Analyzing prices, discounts, shipping, ratings & reviews, inventory, visibility, orders & sales helped to optimize and automate the process using various techniques and Python skills. Various reports for orders and sales - weekly and monthly - using different analysis and visualization tools helped to understand the development and improvement areas for the business to grow and continue to serve our customers in the best possible way.

Contents

Listing of figures

Listing of tables

Listing of acronyms

API ..... Application Programming Interface
CAPTCHA ..... Completely Automated Public Turing Test To Tell Computers and Humans Apart
CSS ..... Cascading Style Sheets
CSV ..... Comma Separated Values
DB ..... Database
DSI ..... Days Sales of Inventory
DRT ..... Daily Review Tracking
EAN ..... European Article Number
EDA ..... Exploratory Data Analysis
GDPR ..... General Data Protection Regulation
HTML ..... Hypertext Markup Language
JSON ..... JavaScript Object Notation
Mkpl ..... Marketplace
MoM ..... Month over Month
MSE ..... Mean Squared Error
NLP ..... Natural Language Processing
ODF ..... Order File
OOS ..... Out of Stock
QOS ..... Quality of Service
QTY ..... Quantity
SKU ..... Stock Keeping Unit
SQL ..... Structured Query Language
URL ..... Uniform Resource Locator
VA ..... Visibility Analysis
VC ..... Visibility Crawling
WoW ..... Week over Week
XML ..... Extensible Markup Language
XLSX ..... is a file extension for open XML Spreadsheet file created by Microsoft Excel
YoY ..... Year over Year

1 Introduction

The development of technology is impacting many aspects of life, including business and the economy. Internet technology is becoming an important aspect of life, changing the way people learn, find information, communicate, and even shop. Selling and buying, which was done in the traditional way by a meeting between seller and buyer, can now be done online through the Internet. The tremendous technological change in communications, software applications and computer hardware, web browsing technology, and multimedia facilitates the information search for specific goods and services that people need. The activity of buying and selling using electronic media and the Internet is called e-commerce [1]. E-commerce is chosen as the current selling system because it offers many benefits to customers and sellers. The benefits of e-commerce can be seen over both a short and long period of time. Through the use of E-commerce, the information about the products can be disseminated more widely, which allows consumers to choose the goods with the best price, provide unlimited, fast communication and information access, be time and cost efficient, and improve the quality of service.

One such platform is Mister Sandman, an E-commerce startup based in the heart of Berlin, Germany. It is an online mattress and bedding company that sells various product categories on both its own and 17 other marketplaces in Europe. The dataset on which this thesis is based is real-time business data from the e-commerce company whose daily operations are described in chapter 2. An overview of the company's business data and its B2C marketplaces is provided in Section 2.1. While Section 2.2 elaborates on the different categories of E-commerce business data in order to leverage the available expertise and obtain sufficient data for analysis and visualization, which ultimately helps to optimize and automate tasks.

The focus is on using various techniques to crawl and scrape data and product-related information from marketplaces [2]. Compare, review and analyse the data accordingly and use it for different purposes to optimise and automate the process for better results. Our goal in this work will be to analyse the above attributes in the marketplaces and product lines. One of the biggest challenges in the work was to overcome the CAPTCHA issues and web blocking during data collection on the different platforms. To overcome this challenge some parameters were introduced and for some marketplaces data was collected using their API back-end links [3].

The aim of this project is to understand the overall marketplace behavior across all the product-lines and for that we need various techniques to crawl and scrape the data from the live platforms [4] which is discussed in section 3.1. Once the data is collected, data cleaning, prepossessing and manipulation will be discussed in detail, which will be addressed in section 3.2. Different types of analysis and visualization tasks are performed as needed. Analyzing inventory & prices, discounts, shipping, ratings & reviews, visibility, orders & sales helped to optimize and automate the process using various techniques and Python skills which is discussed in section 3.3. Various reports for orders and sales - weekly and monthly - using different analysis and visualization tools discussed in section 3.4.

Finally, the results obtained are reported and discussed in chapter 4. It details the goals that were to be achieved and the importance of process optimization and automation according to the methods discussed. It also explains how this has helped to understand the areas of development and improvement for the company to grow and continue to serve our customers in the best possible way.

2

Data presentation

The dataset on which this paper is based is the real-time business data of an E-commerce company and its daily operations. An overview of the company's business keywords and its B2C marketplaces is given in Section 2.1. Section 2.2 elaborates on the different categories of E-commerce business data to leverage the existing expertise and provide enough data for analysis and visualization, which eventually helps to optimize and automate the tasks.

2.1 Big Business Data

To understand the E- commerce, it is always important to know the key words and their definitions. In order to better understand the structure of the data, perform various analyzes, create reports and derive meaning from them, it is important to follow a pattern. This chapter is divided into different sections to understand the basic pattern of E-commerce data which the company follows, as well as the behavior of the B2C marketplaces that the company is connected to in order to sell its products.

2.1.1 Company data

A category of keywords needed to simulate an E-commerce activities is the description of the company data itself. The entities involved are generally organized hierarchically. These entities are:

- **Product line :** A group of products that are closely related because they function in a similar manner, are sold to the same customer groups, are marketed through the same types of outlets, and fall within a certain price range. Our Company is streamlined to home pl;
- **Product Categories :** A product line has several product categories. Our company currently focuses on the following seven categories, as indicated in Table 2.1. They are arranged according to their respective order and sales value;

Sl. No.	Categories
1	Mattress
2	Protector
3	Topper
4	Baseslatt
5	Pillow
6	Duvet
7	Soft-topper

Table 2.1: Product Categories

- **SKU :** In full stock keeping unit, a unique code number, usually used as a machine-readable bar code, assigned to an individual inventory item. As part of an inventory control system, the SKU represents the smallest unit of a product that can be sold from, purchased from, or added to inventory;
- **Product Description :** A product description is the marketing text that explains what a product is and why it is worth buying. The purpose of a product description is to provide the customer with important information about the features and benefits of the product so that they are compelled to buy it;
- **Parent - Child Relationship:** There are three components of a parent-child relationship: *The parent products*: The products that appear in the search results. *The child products*: The products that are associated with each parent product. *The variation theme*: The relationship between the parent and child products categorised by different size, colour, or material;
- **Inventory :** The term inventory refers to all items, goods, wares, and materials that a company owns in order to sell them on the market to make a profit;
- **Customer orders:** The leading units of production that meet a customer request for a particular finished product;
- **Order files:** At the end of the day, all orders from different Mkpls are combined. All appropriate order details are assigned a unique order number and sent to the warehouse for shipment;
- **Sales :** Sales are activities that relate to the sale or number of goods sold in a specific, targeted period of time. The provision of a service for consideration is also considered a sale;
- **Customer Reviews :** An online review is an evaluation of a product or service on the web, by a customer who has purchased and used the product or service or has had experience with it. Online reviews are a form of customer feedback on e-commerce and online shopping websites;
- **Web Visibility :** Web visibility is defined as the extent to which a user is likely to encounter a reference to a company's website in their online or offline environment. It is the process of being identified on the Internet when your potential customers search for your products and services;
- **Marketplace :** An online marketplace is an e-commerce website that brings sellers and buyers together. It is often referred to as an electronic marketplace and all transactions are managed by the website owner. Businesses use online marketplaces to reach customers who want to buy their products and services;

2.1.2 B2C Marketplaces

The term business-to-consumer (B2C) refers to the sale of products and services directly between a company and consumers, who are the end users of its products or services. As mentioned earlier, the company sells its products across all categories on various B2C mkpls, as shown in Figure 2.1. Currently, sales are made across Europe via seventeen different B2C E-commerce platforms. Most of these platforms are active in the German market, with a few in other European regions.

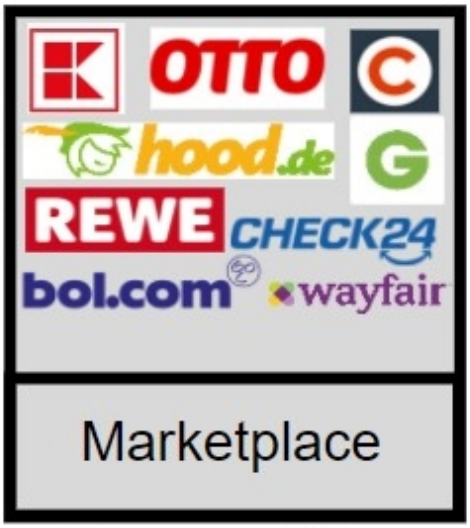

Figure 2.1: Marketplaces breakdown

The pattern of data is different on different platforms, and in this paper we will look at the top sellers. Subsection 2.3.2 deals with the differences and similarities between these mkpls. The corresponding data structures and their benefits and challenges. The three top-seller mkpls discussed here are listed in Table 2.2.

Marketplaces	Region
Kaufland.de	Germany
Otto.de	Germany
Cdiscount	France

Table 2.2: Top-Seller Mkpls

2.2 Dataset description

As mentioned in the opening paragraph of Chapter 2, the data we use is real-time business data from various Mkpls. Taking into account the diversity of E-commerce websites, the sequence of steps was designed to be applicable not only to a specific type or website, but to almost any e-commerce website with its different structures, content or formats.

More details about the website data model can be found in Figure 2.2, which shows the main tables of the database. As a summary overview of the process, we can consider three different phases, namely the data collection and pre-processing phase, the data analysis and pattern recognition phase, and finally the data visualization and reporting phase.

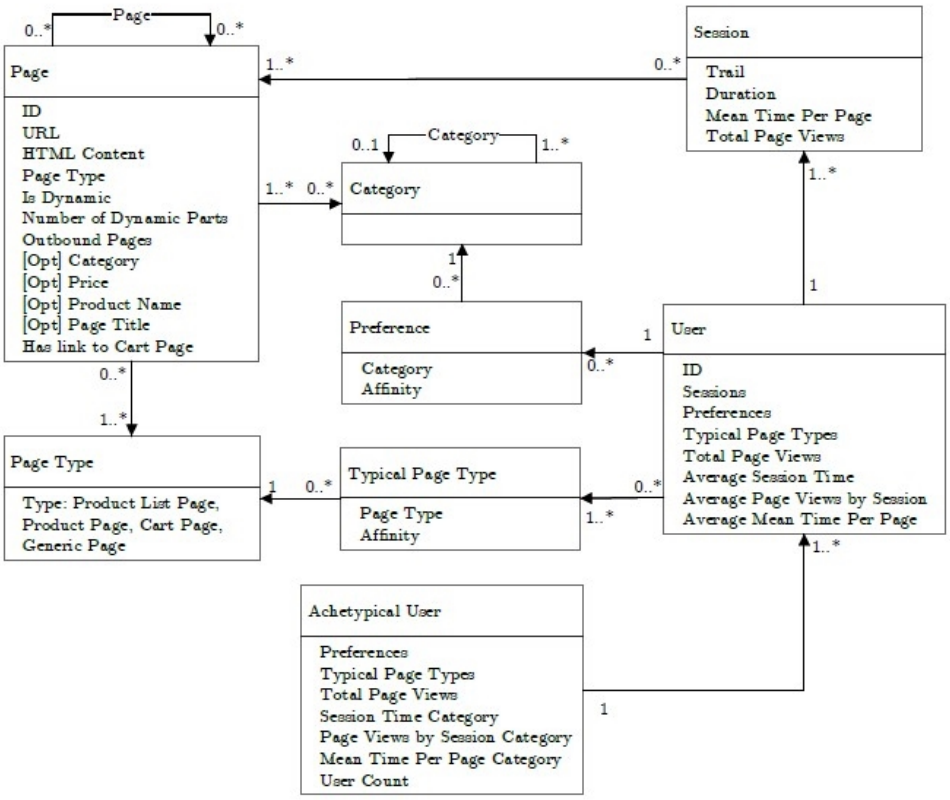

Figure 2.2: A representation of the website data model

2.2.1 Product page details

All the E-commerce platforms follow a general pattern for product pages that display information about a product. However, the layout, structure, format, and presentation are completely different. The following attributes are generally displayed on every product page, regardless of platform. Each of these attributes is important and directly visible to the customer. They are of great importance when it comes to persuading a customer to buy a particular product.

- **Title :** A product title is something where you list your product in an E-commerce store with a title that contains all the necessary information about the product, helping customers to better perceive your product;
- **Image :** It may refer to a diagram or photograph depicting a product for sale. Different types of product images, taken from different magnifications and angles, are sometimes used extensively by companies marketing their products on E-commerce websites and in online advertising to attract customer interest and make purchases;
- **Price :** Price indicates how much your product costs. How you price your product depends on your competitors, demand, the cost of producing the product, and the willingness of consumers to pay the price;
- **Availability :** Product availability includes the cost of developing, manufacturing, storing, and delivering different item variants. When product availability is high, customers can visit the website with confidence that all their purchasing needs can be met. Low availability, on the other hand, can lead to lost sales and low customer loyalty;
- **EAN :** The International Article Number (also known as the European Article Number or EAN) is a standard that describes a barcode symbology and numbering system used in global commerce to identify a specific type of product in a specific packaging configuration and from a specific manufacturer;
- **Product Description :** A product description is the marketing text that explains what a product is and why it is worth buying. The purpose of a product description is to provide the customer with important information about the features and benefits of the product so that they are compelled to buy it;
- **Dimensions :** Product dimensions are characteristics used to identify a product variant. The following product dimensions are available: Configuration, Color, Size, and Style. Each selection results in a unique product variant called a child;
- **Rating Score :** Product ratings are the star ratings that customers give to your products. These rating scores are important for building your E-commerce brand's online reputation because they show customers at a glance that the products have been highly rated by previous customers and that they trust them. These ratings and reviews help with product research and purchase decisions, bringing more qualified customers to your product pages;
- **Customer Reviews :** Every customer has the right to give a review on a product. It is based on the customer's experience with the product and can be positive or negative. These are in the form of text and also important for improving the product and related services;

2.2.2 Top-Sellers

As mentioned in Section 2.1.2, the focus is on the datasets of the top three sellers. As the name suggests, these mkpls have the highest number of orders and the highest sales value. Deep analysis is performed on these platforms in terms of number of sales, delivery status, price structure, and review management. This helps us understand and focus on the top selling Mkpls and improve growth.

The attributes and their corresponding fields are displayed in Table 2.3. The sales data is extracted based on product group, total SKU and total EAN. The following other attributes that directly impact sales are considered for analysis of sales declines or increases.

Attributes	**Field**
Sales Product Group	QTY
Sales SKU	QTY
Sales EAN	QTY
Price Level	Currency
Reviews	TXT
Ratings	Range
Availability	Time

Table 2.3: Top-Sellers Attributes

2.2.3 Google Shopping

Google Shopping is a service offered by Google. It allows customers to search for, view and compare products. These products are displayed when a customer uses Google to search for a product. They may appear on the main search engine results page or under the Shopping tab. The search results include a list of the products you sell with their attributes, including descriptive attributes such as color or brand, price, product tags, and basic information such as price, availability, ratings and number of reviews. The search is performed using the keywords described below -

- **Keyword :** A keyword is a term used in digital marketing to describe a word or group of words that a customer uses to perform a search in a search engine or search bar. Keywords are very important and should form the core of the title, content, and description. Keywords should be carefully placed so that any search for the appropriate product will provide a suggestion of similar products in the search listings.

The generation starts with a set of keywords as input which are designed, developed and carefully selected — after which the visibility analysis code is executed — the data is cleaned and formatted and updated in DB — after which the visibility analysis code is executed — namely, for each Mkpl the results are visualised based on the comparison parameters — after which the changes are made to the prices and stocks — which is inserted in DB. The algorithm is displayed in Algorithm 2.1.

Algorithm 2.1 Visibility pseudocode

For each simulation, execute the following actions:

1: Update the list of Keywords
2: Execute VC
3: Pre-process the data
4: Update the export in the DB
5: Execute VA
6: Simulate the progress of operations, namely for each Mkpl for the corresponding keyword results
7: Update price and inventory levels for each Mkpl accordingly
8: Export the current DB

To run a successful visibility campaign that delivers great results, regular optimization is required. This process is performed once a month and the changes are categorised on all mkpls.

2.2.4 Order File

As explained in detail in section 2.1.2 on our B2C mkpls, we currently use them to sell our products to customers. We create offers with our products that customers can find through search. The offers can be parents with many variants (sizes) or single offers with only one size.

The offers compete for visibility (ranking) with other offers from other companies. By applying SEO methods, we try to keep visibility high. Customers visit different platforms without specifically searching for our products, but a good ranking brings it into their view, so they consider it. If the customer decides to buy our product, they may also buy it from the marketplace.

After the sale, we collect the orders every morning as export from each marketplace. Now we have information about the customer, the product they bought and the address to which it must be sent. Our inventory is managed by a company called Bönning+Sommer (B+S), which acts as the manufacturer. B+S manufactures our products and ships them to customers on our behalf.

This information is obviously important for the creation of ODFs. We take the information from the marketplace exports and convert it into a standardised CSV format called an orderfile. A general order information contains the following features in the record:

- *Channel:* Name of the platform;
- *DO Number:* Order number generated by B+S. Different to Order number from Marketplace!;
- *Date:* Date when the order was sent out;
- *Partner:* Name of the customer;
- *Käuferadresse:* Customer Address;
- *Client Order Reference:* Order number from Marketplace for our matching;
- *Stückzahl:* How many units of this product did the customer buy;

- *Courier:* The carrier company that B+S used for this order;
- *Tracking Reference:* Number of reference to track the package on the carriers website;
- *Retour Number:* Return Tracking number for potential returns;
- *Code:* SKU;
- *Produkte:* Title of the offer at the platform;
- *Verkaufspreis:* Cost of one unit of the product;

The reader can see in Figure 2.3 the structure of order processing. The nodes are Company, B2C Marketplaces, Customers and Manufacturer(B+S). The interaction of these nodes explains the process from order creation to order delivery.

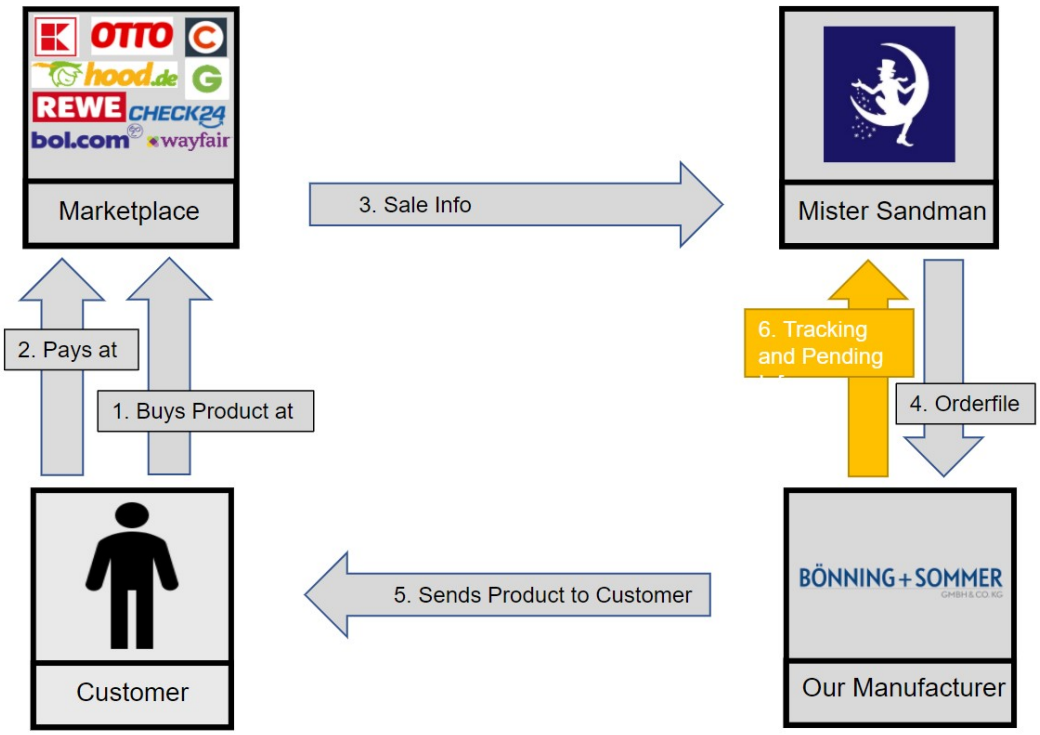

Figure 2.3: Order Processing Structure

3 Methods

In this chapter, the implemented methods are described in detail. In Section 3.1, the main focus is on collecting data in real time from various sources using different crawling and scraping techniques, because that is an important part of this thesis. Data cleaning, prepossessing and manipulation are discussed in detail in Section 3.2. Using this data, various categorical analysis are performed on inventory & prices, reviews, orders & sales, top-sellers and visibility as needed. The analysis is discussed in detail in Section 3.3, followed by a discussion of the reports in Section 3.4.

3.1 Data Collection

The data in the website content is mostly semi-structured or unstructured data. In order to understand, retrieve, and convert the useful information from this site into structured data formats, the use of certain automated techniques is required to extract, interpret, and present the data. This process is called web data extraction. The extraction of structured data from unstructured and semi-structured web data is very interesting, and the fact that most web pages have some structure that can be used to generate structured data.

A semi-structured document, such as web pages, is organized & grouped into semantic units that may or may not have same attributes. The order of attributes may not be important, and not all attributes are required. Also, the size & type of the same attributes in a group may be different and it is obvious that it is much more difficult to query and retrieve information from such sources than from structured information sources such as DBs.

In any case, semi-structured means that the document has some structure that can be identified and extracted. Web pages consist mostly of HTML code and plain text. The structure in these documents comes from the HTML tags used to build the page. If a set of similarly structured HTML documents describe similar content, the identifiable information may be semantically identical. The information can then be extracted and entered into a database. Thus, we have created structured data from the semi-structured web content[5].

One of the biggest challenges in this work was overcoming CAPTCHA issues and Denial of services during data collection on the various platforms.

- **CAPTCHA :** Although the word "CAPTCHA" is familiar to most, far fewer people know what it stands for: Computer Automated Public Turing test to tell Computers and Humans Apart. The unwieldy acronym hints at its rather unwieldy role in hindering otherwise perfectly usable web interfaces. There are several categories of CAPTCHA checks, depending on the structure of the website.
- **Denial of service :** Making too many requests to a website can cause the server to crash or delay responses for legitimate users. This is also known as a denial of service. So, too intensive scraping is a problem. Therefore, it is advisable to use the login credentials to access the websites to avoid this problem.

As an alternative to overcome this challenge, some parameters were introduced and data was collected for some marketplaces through their API back-end links. This is where application programming interfaces come in: they provide nice, convenient interfaces between several different applications. It does not matter if the applications were written by different programmers, with different architectures, or even in different languages - APIs are meant to serve as a lingua franca between different software components that need to exchange information with each other. Although using APIs is not considered web scraping by most people, both practices use many of the same techniques (sending HTTP requests) and produce similar results (getting information); they can often complement each other very well. The only thing that makes an API an API is the extremely regulated syntax it uses and the fact that APIs present their data as JSON or XML rather than HTML.

APIs are generally very easy to use, but unfortunately not every website offers them. Not every website can offer an API, because providing an API requires maintainability and proper structure, which in turn requires someone with the appropriate technical skills. Unfortunately, not every website owner can afford to hire a professional with the desired skills to maintain and operate an API offering. APIs offered by websites also usually come with pricing requirements and request limits, which adds to the problem of additional costs. Kaufland.de is one of the mkpl that offers an open API.

When it comes to storing the data, it can be exported in any of the popular formats - CSV, TXT, XLS/XLSX (Excel) etc. CSV format is one of the most reliable and compact file formats for storing data. It is supported by Microsoft Excel and many other applications and can be easily parsed. Finally, the extracted information is saved in the DB to be able to retrieve it later.

3.1.1 Web Crawling

"A web crawler starts with a list of URLs to visit. These initial URLs are called seeds. As the crawler visits these URLs, it identifies any hyperlinks in the retrieved web pages by communicating with web servers that respond to these URLs and adds them to the list of URLs to visit, called the crawl boundary. The URLs from the list are visited recursively according to a set of guidelines. When the crawler archives web pages (or web archives), it copies and stores the information during the search. The archives are usually stored so that they can be viewed, read, and navigated as if they were on the live web, but are preserved as *snapshots*"[6].

As shown in the Figure 3.1 the web crawler is usually fed a set of URLs as a starting point for the crawl. Then, the crawler parses through each seed page and collects hyperlinks that lead to other pages on the same website or possibly to pages on another website.

The collected hyperlinks are visited recursively according to a set of guidelines that determine how thorough the crawl should be. For most crawling projects, the crawl must be significantly limited and intelligently performed. The reason for this is the volume of web pages on the web, the available bandwidth and the time.

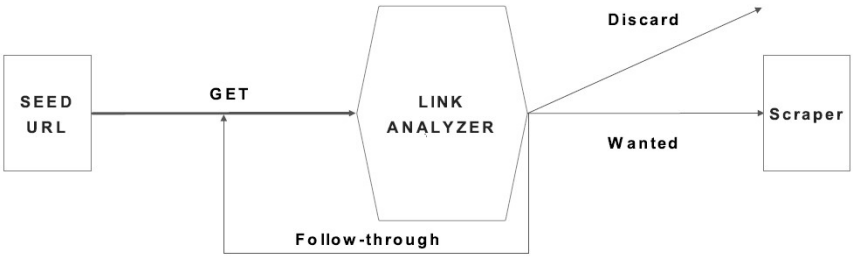

Figure 3.1: Web Crawling Process Flow Diagram

To extract data from websites, a discovery process is required. A web crawler is created that visits websites and performs this process. The way a web crawler works is that it loads the input list of links. Then it finds other links contained in those pages and adds them to a new list, called the crawl frontier, for further exploration. The crawler must identify whether a URL is relative or absolute. For relative URLs, the crawler must identify the base of the URL. The idea is to develop an intelligent crawler that identifies and recognizes circular references and slight variations of the same page so that it can extract and store data efficiently.

The Web is very dynamic by nature. Many events may have taken place, including content creation, updating, and deletion on the web pages. From the search engine's perspective, there is a cost associated with failing to detect an event, resulting in an outdated copy of a resource. The most common cost functions are freshness and age.

Freshness: This is a binary measure that indicates whether the local copy is accurate or not. The freshness of a page p in the repository at time t is defined as 3.1:

$$F_p(t) = \begin{cases} 1 & \text{if } p \text{ is equal to the local copy at time } t \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

Age: This is a measure that indicates how outdated the local copy is. The age of a page p in the repository, at time t is defined as 3.2:

$$A_p(t) = \begin{cases} 0 & \text{if } p \text{ is not modified at time } t \\ t - \text{modification time of } p & \text{otherwise} \end{cases} \tag{3.2}$$

The objective of the crawler is to keep the average freshness of pages in its collection as high as possible, or to keep the average age of pages as low as possible. These objectives are not equivalent: in the first case, the crawler is just concerned with how many pages are out-dated, while in the second case, the crawler is concerned with how old the local copies of pages are.

To limit a crawl, it is advised to apply a selection policy that specifies which links to follow and which pages to download as a result. This selection policy should be based on the purpose of the crawl. For example, if we want to analyse only HTML content and avoid other content types, we restrict the crawler to download only HTML content and omit all other types. If we want to collect pages from a particular website, we perform a path ascending crawl by starting with the index page of the website, identifying all links originating from that page, and following each of those links to find new links. If the goal is to gather information about a specific product page, we can perform a focused crawl. When performing a focused crawl, it would help identify the features without actually downloading them and solve a difficult problem.

It is important to visualize the HTML forms to understand the content of the page. One way to predict the content is to use the anchor text of the hyperlink as a hint to the content. It is critical to develop a crawler that crawls the Deep Web, i.e., the content hidden behind HTML forms. To capture the information behind such forms, it must be possible to submit a form with valid input values. Implementing this helps to cope with this complexity.

Websites also constantly keep updating the layout, pages are added, modified and deleted. Outdated information is less valuable to many systems, so pages must be revisited with some frequency for the maintenance of the crawlers. Two possible approaches are uniform and proportional re-visiting, where in the uniform approach all pages in a collection are re-visited with the same frequency. Or proportional re-visiting, where pages are re-visited in proportion to the update frequency of the page.

Because web crawlers can retrieve data at a much faster pace and in greater depth than humans browsing manually, they can overload a web server, crippling its performance. This can lead to a degradation in quality of service, which is unacceptable. Therefore, web crawlers must be designed to be courteous and adhere to crawling guidelines and courtesy norms that limit polling frequencies to acceptable levels. In general, the polling frequency for our web crawlers is 60 seconds. A lower frequency for crawlers is considered a more aggressive polling frequency. Dynamic polling frequencies are also used. For example, dynamic polling can be based on the download rate of the first page retrieved from a website.

The archive is called a repository and is used to store and manage the collection of web pages. The repository stores only HTML pages and these pages are stored as individual files. A repository is similar to any other system that stores data, like a modern database. The only difference is that a repository does not require all the features that a database system provides. The repository stores the most recent version of the web page retrieved by the crawler.

3.1.2 Web Scraping

"If programming is magic then web scraping is wizardry; that is, the application of magic for particularly impressive and useful-yet surprisingly effortless-feats" - Ryan Mitchell, [7]

Web scraping a web page involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when a user views a page). Therefore, web crawling as explained in section 3.1.1 is a main component of web scraping, to fetch pages for later processing. Once fetched, then extraction can take place. The content of a page may be parsed, searched, reformatted, its data copied into a spreadsheet or loaded into a database as shown in Figure 3.2. Web scrapers typically take something out of a page, to make use of it for another purpose somewhere else.

Figure 3.2: Web Scraping Process Flow Diagram

It is important to understand the components of the navigation patterns which are discussed below:

- **XPath :** The XPath language[8] is based on a tree representation of the XML document, and provides the ability to navigate around the tree, selecting nodes by a variety of criteria. The most important kind of expression in XPath is a location path. A location path consists of a sequence of location steps. Each location step has three components:

  1. an axis;
  2. a node test;
  3. zero or more predicates.

  An XPath expression is evaluated with respect to a context node. An Axis Specifier such as *child* or *descendant* specifies the direction to navigate from the context node. The node test and the predicate are used to filter the nodes specified by the axis specifier. A predicate can be used to specify that the selected nodes have certain properties, which are specified by XPath expressions themselves.

  The XPath syntax comes in two flavors -

  – the abbreviated syntax;
  – the full syntax.

- **CSS :** Cascading Style Sheets[9] is a style sheet language used for describing the presentation of a document written in a markup language such as HTML. CSS is designed to enable the separation of presentation and content, including layout, colors, and fonts. This separation can improve content accessibility; provide more flexibility and control in the specification of presentation characteristics; enable multiple web pages to share formatting by specifying the relevant CSS which reduces complexity and repetition in the structural content and enables to be cached to improve the page load speed between the pages that share its formatting.

  In CSS, selectors declare which part of the markup a style applies to by matching tags and attributes in the markup itself. Selectors may apply to the following:

  – all elements of a specific type;
  – elements specified by attribute;
  – id: an identifier unique within the document, identified with a hash prefix;
  – class: an identifier that can annotate multiple elements in a document.

As mentioned in Chapter 2, there are several ways a website can load and render its content. Most server-side rendered websites can be read by a simple HTTP request and the use of some XML parser. To extract meaningful data from the scraped data, web scrapers must use parsers. These are typically implemented to format and extract specific details from the data, such as the CV parser, which can extract person names and contact information from the contents of an email. Most web scraping libraries provide basic HTML parser support. There are also parsers for specialised data stored as PDF, CSV, QR code, or JSON. Real web browsers such as Firefox and Chrome have built-in parsers. Web scraping, which is done by controlling a real web browser, can also use the browser's built-in parser. Below are listed the python packages and libraries used for scraping the web pages:

- **Beautiful Soup :** is a python package for parsing HTML and XML documents (even with broken markup). It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping. It works with the parser to provide idiomatic ways to navigate, search and modify the parse tree. It automatically converts incoming documents to Unicode and outgoing documents to UTF-8.

  Beautiful Soup's *find()* and *findAll()* functions are the two functions you will probably use most often. They allow you to easily filter HTML pages to find lists of tags you want or a single tag based on its various attributes. The attributes argument takes a Python dictionary of attributes and finds tags that contain one of those attributes. The text argument is unusual in that it is searched based on the text content of the tags rather than the properties of the tags themselves. The limit argument is, of course, used only in the findAll method, where find is equivalent to the same findAll call with a limit of 1.

  Beautiful Soup and regular expressions go hand in hand when it comes to scraping the web. Most functions that accept a string argument also accept a regular expression. When you pass in a regular expression object, Beautiful Soup filters for that regular expression using its search() method. In the Beautiful Soup library, a distinction is made between children and descendants: similar to a human family tree, children are always exactly one tag below a parent, while descendants can be below a parent at any level in the tree. In general, Beautiful Soup functions always deal with the descendants of the currently selected tag.

- **Scrapy :** is a python library that handles much of the complexity of finding and evaluating links on a website and crawls domains or lists of domains effortlessly. It is an open-source framework that allows you to quickly and easily, yet extensibly, collect the data you need from websites. Scrapy uses Item objects to determine what information it should store from visited pages. This information can be stored by Scrapy in various ways, such as CSV, JSON or XML files. It is a powerful tool that solves many problems related to crawling the web. It automatically collects all URLs and compares them with predefined rules. It makes sure that all URLs are unique, normalizes relative URLs if needed and recurses to go deeper into pages.

  To put it in a simple context: Any website that displays information in HTML format. A web scraper is used, an automated program that sends requests to the servers of websites. The requests tell the servers that a particular web scraper wants to retrieve its data. The data provided by the particular website is then processed and depending on the properties specified by the web scraper, the web scraper extracts the desired properties. To simplify the process of data collection, the Scrapy framework is used. It allows the web scraper to efficiently extract the data it needs.

  To scrape the right properties, the web scraper extracts information based on the class name of an HTML element. Using the class name is more beneficial than traversing through tags, because even if a platform changes its layout, the likelihood that the class name will also change is minimal. Scrapy is a useful framework for platforms that primarily use HTML and CSS for the front-end portion of their website. However, due to popular JavaScript frameworks such as React.js and Angular.js, the rendered content of some platforms' websites is generated using JavaScript and therefore Scrapy cannot extract data from it. Nevertheless, when Scrapy is used with a testing framework called Selenium, it is possible to get all the data that is displayed in a normal web browser.

- **Selenium :** It is a powerful web scraping tool that was originally developed for testing websites. Nowadays, it is also used when it comes to rendering websites exactly as they appear in a browser. Selenium works by automating browsers to load the website, retrieve the necessary data and even take screenshots or claim that certain actions are taking place on the website. Selenium does not include its own web browser, but requires integration with third-party browsers to run.[10]

  Selenium is a library that provides an interface to automatically control real web browsers. There are some websites that rely heavily on JavaScript and can only be queried with a real browser. Even though Selenium allows scraping some complicated websites, it costs much more computer resources. Moreover, Selenium slows down the scraping process significantly, as it needs to open the browser and load the entire web page. The strategy is to try scraping without Selenium first and resort to it only when there is no other solution.

  To overcome this challenge, a tool called PhantomJS is integrated into Selenium, a so-called "headless" browser. It loads websites into memory and runs JavaScript on the page, but without graphical rendering of the website to the user. By combining Selenium with PhantomJS, you can run an extremely powerful web scraper that easily handles cookies, JavaScript, headers, and anything else you need.

  Selenium is a set of tools specifically for web browser automation. The Selenium library is an API that is called through a *WebDriver*. The WebDriver is a bit like a browser in that it can load websites, but it can also be used like a Beautiful Soup object to find page elements, interact with elements on the page (send text, click, etc.), and perform other actions to control the web scraper. Expected conditions can be many things in the Selenium library, including:

  - An alert box pops up
  - An element (such as a text box) is put into a "selected" state
  - The page's title changes, or some text is now displayed on the page or in a specific element
  - An element is now visible to the DOM, or an element disappears from the DOM

  Selenium is slower than Scrapy (in terms of performance) when it comes to extracting information. Therefore, using Scrapy is preferable, but if a website renders its content with Javascript, Selenium can help extract the desired information.

  Elements are specified using locators. A locator is an abstract query language, using the *By object*, which can be used in a variety of ways, including to make selectors.[11] The following locator selection strategies can used with the By object:

  - *CLASS_NAME :* Used to find elements by their HTML class attribute;
  - *CSS_SELECTOR :* Find elements by their class, id, or tag name;
  - *LINK_TEXT :* Finds HTML <a> tags by the text they contain;
  - *PARTIAL_LINK_TEXT :* Similar to LINK_TEXT, but matches on a partial string;
  - *NAME :* Finds HTML tags by their name attribute;
  - *TAG_NAME :* Finds HTML tags by their tag name;
  - *XPATH :* Uses an XPath expression to select matching elements;

The ethics of web crawling and scraping have been taken into account and implemented in compliance with all rules set forth in the GDPR laws.

3.2 Business Data Preprocessing

After the step of data extraction, an extensive work was done to properly clean, prepare and analyze in depth the data under study. It is important to clean and prepare the data in the standard format. In this section, we will focus on getting properly formatted data and using it to perform various analysis. To import data into Python, we will use the pandas package. Pandas is a powerful data analysis package and is widely-used for analysis and manipulation. It has several functions for reading data. The data can be in any of the common formats - CSV, TXT, XLS/XLSX (Excel), etc.

Data transformation is essentially about converting, cleaning and manipulating the data into a usable format, i.e., extracting, transforming, and loading the data into a database for analysis. The general code contains the following steps used for data preprocessing and transformation:

• *Rebuild Missing Data :* "Missing" simply means NA ("not available") or "not available for some reason." Many datasets simply arrive with missing data, either because they exist and were not collected or because they never existed. In Pandas, one of the most common ways to add missing data to a dataset is to re-index it. Pandas objects are equipped with several data manipulation methods to deal with missing data.

- Filling missing values: the fillna() function can "fill in" NA values with non-NA data in various ways, e.g. by replacing NA with a scalar value or by filling gaps forward or backward. If possible, missing information such as postal codes, states, countries, telephone area codes, gender, web addresses from e-mail addresses, etc. will be restored to the order files.

• *Standardize and Normalize Data :* The entries in the fields or categories of the given record must be homogeneous, i.e. all entries must have the same format for Title, Image_URLs, EAN, Dimensions, Ratings, etc. Also, this step ensures that similar information is stored overall in the same format or in their default format or as needed. As in this example, I made some basic type corrections, extracted and converted prices and currencies, and inserted the null values.

In addition to the usual steps (removing nulls, converting columns to the proper data types, etc.), there were some interesting steps I included in the code to preprocess the obtained data sets. As a first step, I go through each column using the Pandas unique() method. This way I can see if the values are consistent and make sense and identify any potential problems.

• *De-Duplicate data :* Identify potential duplicates, i.e., I check the data for duplicates by grouping rows by the column that serves as the unique identifier for a particular item - for example, I used the product ID. Look for highly accurate matches with a tolerance for misspellings, missing values, or different mkpl sequences. For business-critical data, manually check these results and update the database accordingly.

One of the things I observed when dealing with duplicate values is that some product pages were linked to multiple listings on the search results page. Although it was good to get rid of the duplicates, I decided to keep that information for analysis. So in such cases, it was better to first create a new column with the number of listings per item and then delete all but one of the copies.

• *Verification to enrich data :* Validate data against internal and external data sources to append value-added information. For example, information from the purchase order file can be validated before finalization to verify the current phone number and address. The same applies to various other fields such as profit, revenue, sales, orders, ratings and reviews, etc., which can be fetched for each Mkpl.

• *Merging data :* The merge or join operation combines records by linking rows with one or more keys. This is the most important function, as it is mostly needed for merging different Mkpls datasets. Depending on the requirements, some operations of the merge function are performed with the dataset:

– left: DataFrame to be merged on the left side.

– right: DataFrame to be merged on the right side.

– on: Column names to join on. Must be found on both DataFrame objects.

– left on: Columns in left DataFrame to use as a join keys.

– right on: Columns in right DataFrame to use as a join keys.

Next, I processed the preprocessed datasets to make them useful for different analysis and visualization tasks, which I discuss in detail in the next sections.[12]

3.3 Big Data Analysis

In recent years, the importance of Big Data Analytics in E-commerce has increased. Big Data Analytics [13] stands for the process of analyzing a set of data sets to extract unknown correlations, hidden patterns, market trends, customer preferences, and other useful information that enables companies to gain insights from data and improve their business decisions. [14] [15]

Big Data analytics applications allow us to examine growing amounts of data. It has become increasingly important in the E-commerce companies in recent years. The study of data is constantly generating a great deal of interest in business intelligence and analytics and offers a wide range of opportunities. As a data-centric approach, Big Data Analytics has its roots in established database management.

Big Data Analytics has emerged as a new frontier for innovation and competition in the broad spectrum of the E-commerce due to the challenges and opportunities created by the information revolution. Big Data Analytics increasingly provides value to E-commerce companies by leveraging the dynamics of people, processes, and technologies to transform data into insights for sound decision making and solving business problems. This is a holistic process that looks at data, sources, capabilities and systems to create a competitive advantage.

Big Data analytics and its scope are well defined by the parameters of web data analytics. Web analytics is an approach that involves collecting real-time information from Mkpls (see Section 3.1), extracting information from data processing (see Section 3.2), measuring, monitoring, analysing, and reporting web usage data to understand visitor experiences, and is discussed in detail in the next sections.

Analytics can help optimise websites to achieve business goals and/or improve customer satisfaction and loyalty and implement key business decisions. It refers to the measurement and analysis of the data, regardless of whether the company owns or maintains the website. It takes into account the visibility of all the mkpls.

Big Data Analytics establishes a relationship to communicate with the data and use it to find new opportunities. This leads to increasingly insightful business operations, higher benefits, profitable maneuvers, and energized customers. The idea is to soon share the business prospects in a better way and then use them with an investigation idea. Information is spreading at a rapid pace and the speed of information improvement is high. It is crucial to combine this data produced across the Mkpls.

If the various characteristics and types of Big Data are well understood and the challenges are properly addressed, the analytics process will maximize business value by facilitating end-to-end use and rapid delivery of insights at Mkpls. Therefore, the objective of this research is to identify various conceptual dimensions of Big Data in E-commerce and their importance to business value.[16]

3.3.1 Inventory and Pricing analysis

Inventory and price analysis offers numerous benefits. One of the most important is the significant reduction of lost sales and the increase of competition in the market. This is because the analysis indicates the quantities of products should be kept in stock that are most likely to sell. Another benefit is the reduction in working capital needed to support a company's inventory investments, as investment in inventory items with low turnover rates is reduced. A third benefit is increased customer loyalty, as inventory items that customers actually want to buy are delivered as promised. A fourth benefit is offering competitive prices and better service quality to the customers with promised delivery time.

As reported in Section 2.2.4 on Bönning+Sommer (B+S), which acts as the manufacturer and manages the inventory. It is important to perform the analysis on a bi-weekly basis to meet customer requirements, coordinating with B+S. In short, inventory analysis is about more than using a single calculation to determine inventory levels. Instead, a number of factors such as the company's strategy, production systems, financing and market requirements must be examined to determine the optimal inventory level.

Inventory analysis is the examination of inventory to determine the optimal time to ship products. Traditionally, this has been done using basic historical order and inventory data and setting up parameters based on orders and availability. However, inventory analysis must be performed to a significant degree to accommodate the philosophy of customer service and order fulfillment.

Inventory represents a current asset because a company usually intends to sell its finished goods within a short period of time, usually as early as possible. Inventory must be physically counted or measured before it can be included on the balance sheet. Companies typically have sophisticated inventory management systems that allow them to track inventory in real time. Inventories are accounted for using one of the following methods, as shown in Table 3.1, where our company uses the weighted average method:

The DSI is a common method of evaluating the average time it takes the company to convert its inventory into revenue. The DSI is calculated by taking the average annual inventory, dividing it by the cost of products sold for the same period, and multiplying the result by 365. The smaller the DSI, the more efficiently a company operates by quickly monetizing its inventory. The DSI can vary for different product categories and SKUs. For a company like us, it is an important factor to better organize inventory items and production. On average, the maximum DSI for some product categories is around 50 days, while for the top-selling category, i.e. mattresses, it is 10-15 days specifically for certain sizes and features.

Method	Description
Specific Cost Method	It is a method to control items individually by matching the item and its purchasing price.
FIFO Method	It is a method based on the assumption that the first purchased items are first used.
LIFO Method	It is a method based on the assumption that the last purchased items are first used.
Simple Average Method	It is a method using the average purchasing unit price for period
Weighted Average Method	It is a method using the average purchasing price for period by dividing by purchasing quantity.
Moving Average Method	It is a method where an issue unit cost is determined by every receive of purchased item.
Last Purchase Price Method	It is a method where the receive unit cost of last purchase is the issue unit cost.
Cost Percentage Method	It is a method where the evaluation is method by the percentage of the selling price.

Table 3.1: Inventory Analysis Methods

Inventory turnover indicates how quickly inventory is consumed in a given period. It is calculated by dividing the ending inventory by the annual cost of goods sold. If the ending inventory deviates significantly from the norm, the average annual inventory can be used instead. Based on the inventory turnover rate, we can assess whether the B+S company has excessive inventories compared to sales. Inventory turnover can fluctuate due to low sales or poor inventory management. Therefore, it is very important for us to keep track of this data through analysis in order to match it with the B+S company. Also, the orders of seventeen Mkpls have to be managed considering the importance of the customer. Any delay in delivery would directly affect the QOS, which in turn would lead to poor ratings and bad reviews and apparently loss of customers.

These methods of analyzing our company's inventory help us optimize pricing strategy and maximize revenue and profit margins. They help set constraints to provide boundaries for frequent price changes and incorporate additional costs and supplier incentives into the pricing model. The product lifecycle directly impacts and feeds into pricing decisions. These analysis supports dynamic pricing and event-driven price adjustments. In Quarter4, for example, it helped us lower prices for certain products during promotional periods such as Black Friday and increase sales. Thus, it helps us to implement ad-hoc price changes irrespective of Mkpls and product SKUs.

3.3.2 Review Management analysis

"An online review is the review of a product or a service made on the web, by a customer who has purchased and used or had experience with the product or the service. Online reviews are a form of customer feedback on electronic commerce and online shopping sites". [17]

To succeed in today's E-commerce business, it is increasingly important to understand the voice of the customer in order to develop better products and meet customers' needs. With the development of E-commerce, the large number of online reviews has significantly influenced product sales and the way customers make a purchase decision. This data could be a viable source to capture users' needs and preferences for product development, especially for all our Mkpl platforms, which help us improve our products and services in the competitive market.

Meanwhile, online reviews are updated in real time, so we can track changes in user preferences at any time. This unprecedented characteristic is basically summarized as the velocity of Big Data. It provides us with the ability to gain new insights about the various mkpls and the competitive landscape that are not possible with traditional methods of determining user preferences. It helps us to capture the changes and identify the trends of customer preferences and gain a strong competitive advantage in today's competitive market.

As described in Section 3.1 on web data collection methods through crawling and scraping, online review and rating data is collected daily from each Mkpl including our own Webshop platform. These datasets are real-time reviews and ratings scores provided by customers. The analysis focused on taking advantage of the velocity of the online review data. Therefore, this section discusses the parameters that provide a method to analyse the dynamic changes in user preferences in different time spans. The proposed parameters help the company to understand, evaluate, and develop product improvement strategies.

It is also important to understand the rating score, which in simple definition are closed-ended feedback in comparative form for specific products based on customer experience, product quality and services. They are visible on product pages in the form of star ratings that customers assign to each. It is one of the most common methods of rating a product and communicating one's satisfaction with it. The rating score on an E-commerce platform is considered direct feedback from the customer after an order is completed and is generally structured in the following format Table 4.1:

Rating Score	Feedback
5 star	Excellent service, exceptional satisfaction
4 star	Very good service, above average satisfaction
3 star	Decent service, average satisfaction
2 star	Dislike, Less than acceptable
1 star	Completely dissatisfied, unacceptable

Table 3.2: Rating score format on E-commerce platforms

The rating score and number of ratings after collection from the respective mkpl web pages, are merged together with respect to their corresponding product-id's. This data is cleaned and preprocessed and exported to DRT spreadsheet and stored in the DB for analysis.

An analysis of the ratings and number of ratings of all seventeen mkpls across seven product groups showed that the extremity of the rating, the impact, and the product type influenced the perceived improvement of the products. Fewer ratings with extreme ratings are less helpful than more ratings with moderate ratings. In addition, here the parent-child relationship is directly or indirectly affected by changes in the number of ratings. These analysis are performed considering the relationship, i.e., a poor rating for a child product will affect the rating of the corresponding parent product. To avoid this, the child product is excluded from the parent product and created as a separate single product ID. This gives the parent product a good rating again, and until the ratings of the child products improve, it is sold as a single product ID.

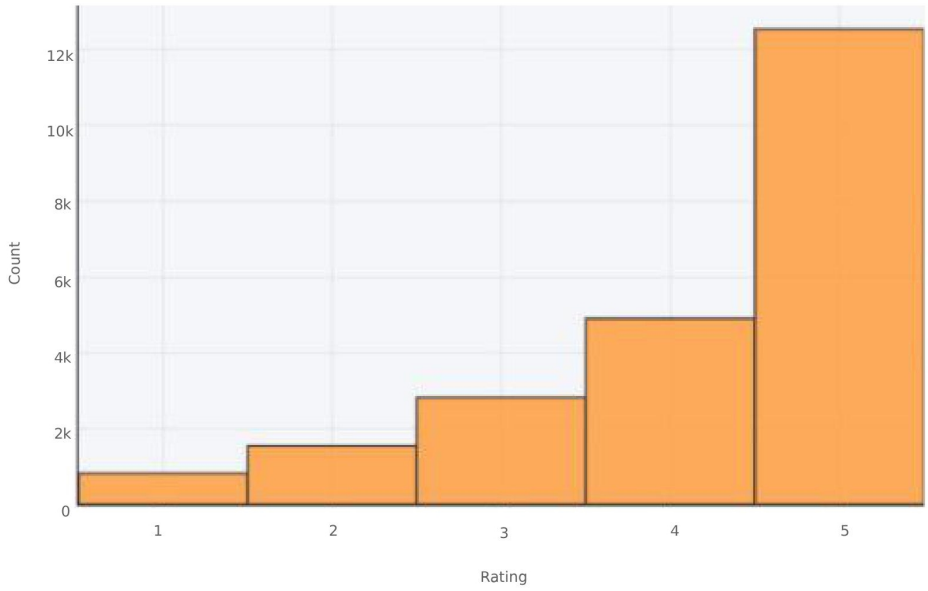

Figure 3.3: Distribution of total rating score across all mkpls and product categories including webshop

Figure 3.3 shows the distribution of total rating score across all mkpls and product categories including webshop. It implies that the ratings are consistent with the product and service quality offered by the company, i.e. most of the ratings are quite high at 4 or 5.

On the other hand, online text reviews provide our company with the opportunity to gather a large amount of information about customer requirements and preferences and cumulative feedback of any product. The large amount of easily accessible evaluation data allows us to capture the full spectrum of customer needs in a timely and efficient manner. The benefits of having customer reviews are immense, and also this widely available data in terms of feedback on any product makes customer reviews helpful to a customer in the process of making a purchase decision.

Through an analysis of the current state of the art in online review analysis, we have identified three challenges in online review analysis:

1) the challenge in data acquisition,

2) the challenge in data structuring, and

3) the challenge in data analytics.

As we enter the era of Big Data, certain data security challenges arise, as mentioned in Section 3.1. Crawling the web from various mkpls and scraping real-time data is becoming increasingly difficult nowadays. Therefore, when analysing online reviews, we implemented certain parameters based on the challenges to solve them accordingly. Since online reviews are text data, the unstructured nature is one of their characteristics. Customers can write about anything in text and they generally write about the things that interest them. For this reason, compared to other types of data, text data needs to be structured before further analysis.

Generally, a text review is divided into single sentences ("sentence-based") and words ("word-based") or very short texts from a single source. A data analysis method is proposed to capture the changes in customer preferences related to affordances based on the structured data.

Although today's NLP - natural language processing technology enables the machine to understand natural language to some degree, the variety in word usage, sarcasm, sentence ambiguity, etc., still prevent us from obtaining automated data structuring with 100 percent accuracy. Last but not least, data analysis requires translating the statistical features of the data into practical meaning, which requires the data analyst to have deep expertise. This analysis of challenges points to future improvements and advancements in analysis and implementation for the future.

Customer needs are measures of customer value that are controllable by product experience and service quality, independent of any solution or technology. A complete set of customer needs affects all aspects of innovation, the way markets are segmented and sized, the way product and pricing strategies are formulated, and the way ideas are developed, tested, and positioned. Thus, this analysis helps us improve our business and services on all platforms.

3.3.3 Top-Seller analysis

Top seller analysis is an in-depth analysis of specific mkpls that generate the highest sales across the platform. As described in Section 2.2.2, the analysis is performed based on various attributes. Sales are categorised and measured by product groups, total SKUs, and total EANs. The purpose of this analysis is to maintain and improve sales and review any measures that affect the growth of these mkpls in any way.

The top seller analysis is divided into different sections and the format of the report contains all the important information that can be extracted from the dataset. It is a WoW analysis process and the various parameters of the data required for the analysis are extracted from the DB using SQL. All these data are included in another spreadsheet to make the analysis process smoother and also to be able to check the progress compared to the last week or weeks. In general, the analysis is divided into the following segments:

1. General Sales analysis:

- Sales in Total, while assorting the offers from all the Parents shown in percentage and in numbers highlighting the changes in the value;
- Sales based on different product categories, showing percentage increase or decrease;
- Sales based on total assortment from EAN and SKU assortment, highlighting specific EANs and SKUs in case of major impact;

2. Root Cause analysis:

- *Pricing :* Observing the price changes and the significant impact on sales due the changes if any, the EANs and SKUs are highlighted if the sales increase or decrease;
- *Reviews and Score :* As mentioned in the previous section 3.3.2, the ratings are analysed daily and the percentage impact on sales is measured in terms of changes in rating score;
- *Availability and Inventory :* Availability on all mkpls is updated bi-weekly based on inventory analysis so that sales behaviour can be monitored based on product availability visible to customers;

3. Deep Dive analysis:

- It is a method in which an intensive, in-depth analysis of the impact on sales is observed. Deep Dive is conducted to explore the problem, identify and investigate the problem, and mitigate its negative impact in an efficient and cost-effective manner. It provides deeper and useful insights into potential bottlenecks impacting the business or sales, thus it helps to optimize the process and improve business standards.

4. Review to last week(s):

- This is a comparative analysis in terms of improvement or deterioration of sales or related factors based on the analysis reports of the last week or weeks. All information is provided to better understand the trend and identify the problem. This also helps us to identify the root cause and resolve it soon.

Figure 3.4 shows the MoM sales development for the individual product category mattress, over the three top seller mkpls in their respective sales region.

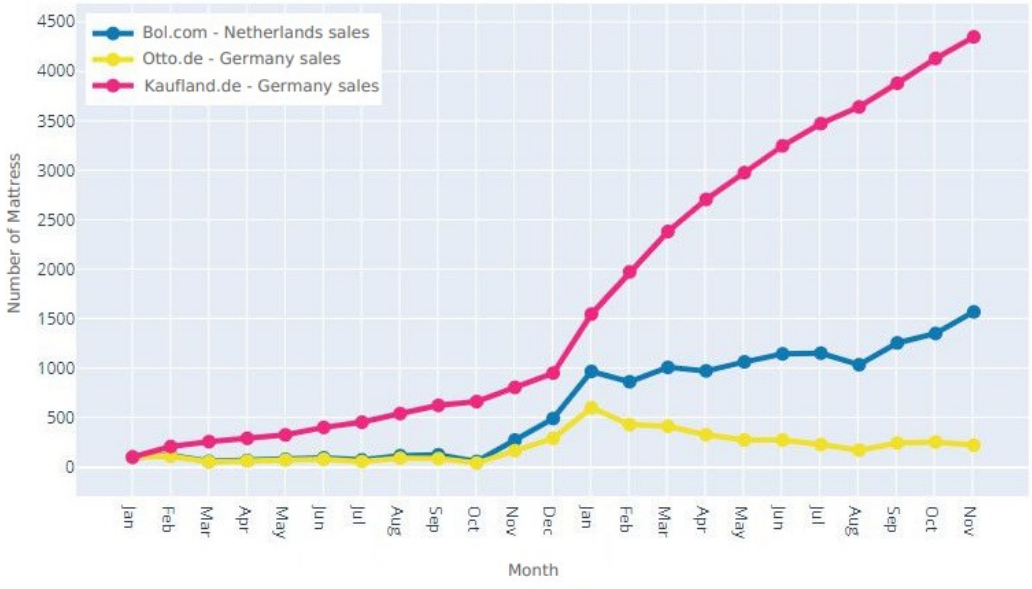

Figure 3.4: Top-seller Mkpls sales development for product category mattress, MoM

3.3.4 Visibility analysis

Search engines like Google use various ranking factors for websites to determine the position of a particular search result in a search engine results page. The visibility of mkpls products in search engines results from an algorithm that ranks and ranks websites according to their calculated ranking position. The original concept of ranking for the Google search engine is called PageRank, after one of the founders of Google. PageRank was invented and published in 1998. [18]

This concept takes into account inbound links and estimates ranking positions for websites and corresponding keywords based on their volume and quality. Nowadays, this topic is more researched and can be divided into onsite factors and offsite factors. Onsite factors are domain-related, website-related, and page-related while offsite factors are link-related, user action-related, rule based, brand-related, and spam-related.

In recent years, a customer who wants to buy any product searches for it on the Google Shopping platform with a relevant keyword related to the product of his interest. It allows customers to search, view and compare products. These products are displayed when a customer uses this platform to search for a product. They can appear on the main page of the search engine or under the Shopping tab. The search results contain a list of products with their attributes, including descriptive attributes such as color or brand, price, product tags, and basic information such as price, availability, ratings, and number of reviews.

Search engine visibility is measured by the number of keywords, position, and number of pages visible, which can be used to compare competing companies in a common field or industry. The comparison provides information on how high the market share of each compared company is. Based on this comparison, further analysis of Internet strategies in the areas of marketing, sales, advertising and publishing can be determined. Visibility in search engines is always subject to algorithms that sort and rank the results based on the type of content, metadata, and content creation model, or hide the results for various reasons.

With visibility analysis, we track our ranking (the position of our listings) on the Google shopping platform for various keywords to find our brand's products sold in all Mkpls. For example, we track how many products of our webshop appear in the top 100 regardless of the Mkpls when the customer types "Matratze 140x200" or any other specific keyword related to our product categories on Google Shopping as discussed in section 2.2.3 with the process algorithm.

The following parameters are the most important factors which are looked after as the outcomes of visibility analysis :

- Location specific keywords language
- Keywords coverage
- Optimized product title
- Good rating score
- High number of reviews
- Price Competitiveness
- Product availability and less time of delivery

If the keywords that one would expect do not appear in the query list, the website may not have enough useful content relevant to those keywords. High impression volume queries can help identify where titles and snippets for web search can be improved to match the customer's interests. The analysis reveals how many keywords show our products when the customer does or does not include a particular string, such as a brand name. This analysis makes it easier to identify issues with our content and specifically helps with title optimization. The results of these analyses can be reused by the marketing team to promote our brand and products.

The better our visibility, the greater the chance that a customer will find us and buy our products. We need to track the visibility to know if some changes and improvements had a positive or negative impact on the ranking (position, visibility). We track the visibility once a month and do the analysis of the results. If we are currently making improvements, we should check the visibility more frequently. In the future, we plan to implement visibility analysis on our top seller platforms and other mkpls as well and improve the visibility of our products for customers.

3.3.5 Order and Sales analysis

The two most important characteristics of any E-commerce business are the orders that customers place on the website or marketplace and the relative sales with respect to the cumulative orders. An end- to-end process that begins with the customer searching the website or mkpl for a product to find our brand that meets their requirements and placing an order until the customer receives the product at the specified address, completes an order. When this process is completed and the customer is satisfied with the delivered item and does not initiate a return, adds value to the sales.

Once a customer places an order on one of our platforms, we generate the ODF, as explained in Section 2.2.4, and initiate the process to fulfil the customer demand. An order is generally categorised into the following states:

- *Open\Processed :* An open order is any order placed by the Customer, regardless of mkpl, that has not yet been shipped. The information for each order is communicated to B+S in an ODF on a daily basis. As soon as B+S receives the ODF, it initiates the package shipment for delivery. Until delivery, the order remains in open or processed status.
- *Cancelled :* Until a product is delivered, it can be cancelled either by the customer or by the company for their respective reasons. A customer usually cancels the order due to changes in opinion or requirements, while an order is usually cancelled by the company due to the product OOS. In both cases, this means a loss for the company as it loses sales.
- *Sent\Completed :* An order is not considered complete until it has been successfully delivered to the customer. All payments will be generated by the mkpls only after an order has been completed. Any delay in delivery may result in cancellations by the customer.
- *Return :* A customer may initiate a return after receiving the ordered product if he is not satisfied with the quality of the product or service. Also, a deviation of dimensions or product specification from the ordered product can be a reason for return. So it is always important to validate the ODF with customer orders. This leads to lost sales, as the delivery and return costs add up with the refund of the product costs.
- *Announced :* This is a waiting period after an order is placed to confirm payment, which is usually made by credit card or in installments to avoid fraud attempts. Once the payment is confirmed, the announced status is changed to open status. This sometimes leads to customer dissatisfaction due to delivery delays.

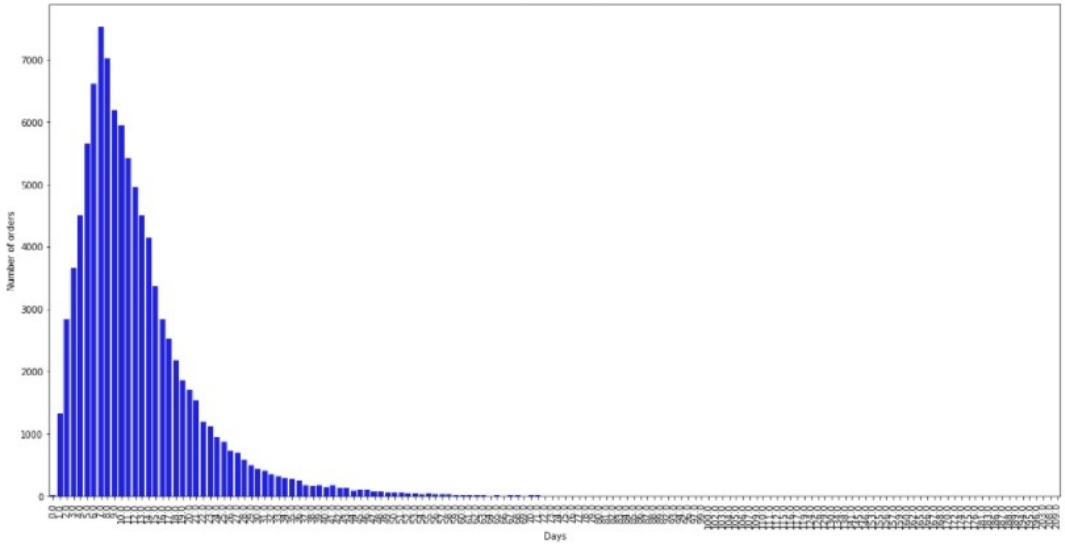

Figure 3.5: Delivery time, total orders across all mkpls and product categories

Delivery time is an important factor that directly and indirectly affects sales and customer values. Figure 3.5 shows that the majority of orders were delivered within a week, while a few orders took more than half a month. As an E-commerce company, our priority and practice is to deliver orders within the promised time unless there are any issues.

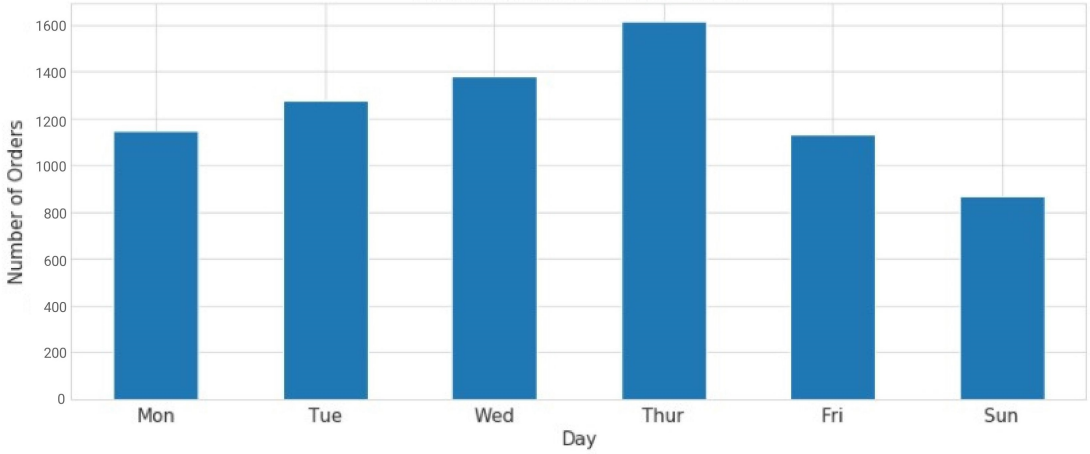

Figure 3.6: Total Orders across all mkpls and product categories, day-to-day

Figure 3.6 gives us insight into day-to-day orders across all mkpls and product categories. This is the data from the best sales week during Black Friday, when the number of customer orders jumped. It was a difficult time to fulfil all orders, even though the company and manufacturers were prepared. YoY analysis of orders and sales history helped us to fulfil about 90% of orders.

Orders and sales were analysed on our website in the last weeks of Q4, and a comparative analysis was performed based on the data to determine the reason for the decline in sales. Figure 3.7 shows total sales for each product category on Webshop and Figure 3.8 shows total orders across all product categories on Webshop, WoW.

The decline was observed due to high sales in the weeks before because of Black Friday. This trend has also been observed in some other mkpls. So it is important to understand customer behaviour in terms of orders and sales for periodic and uniform distribution of business. However, in this scenario, we were able to fulfil the orders and successfully include them in the revenue recognition.

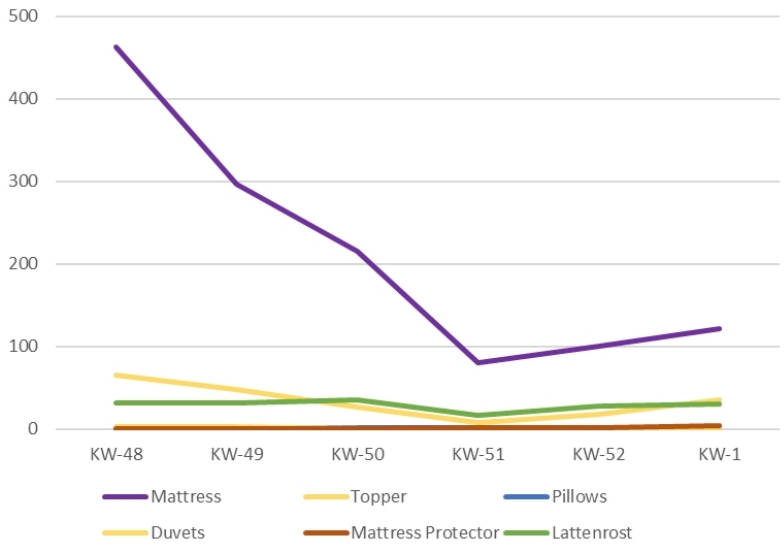

Figure 3.7: Total Sales for each product category on Webshop, WoW

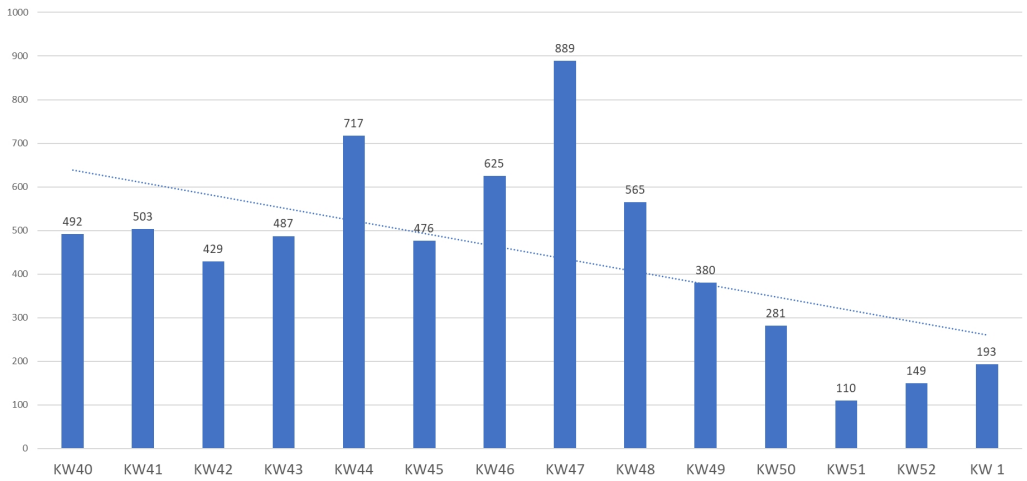

Figure 3.8: Total unit orders across all product categories on Webshop, WoW

As we talked about the requirements and the importance of analysing and understanding customer purchasing pattern, going forward we will be working across platforms, which can give us some insights into orders and sales. To incorporate this into our analysis, we need to implement Association Rule Learning [19] and Apriori Algorithm [20]. These algorithms will help us better understand customer behaviour and ordering patterns. It is important to understand the following concepts on which our analysis will be based:

- **Support :** Support for a given rule A->B (in other words if A is ordered, B is also ordered) is defined as the proportion of transaction in the data set in which both A and B appear. It shows the popularity of both A and B i.e.

$$\text{Support(A,B)} = \frac{\text{Number of Transactions in which A and B both appears}}{\text{Total Number of Transactions}} = \frac{freq(A,B)}{N} \tag{3.3}$$

- **Confidence :** Confidence for a given rule A->B is defined as the frequency of A and B together in a basket divided by the frequency of A i.e.

$$\text{Confidence} = \frac{freq(A,B)}{freq(A)} \tag{3.4}$$

  It shows the likelihood of item B being ordered when A is ordered. If product B is also equally popular as A then there will be a higher probability that a transaction containing A will also contain B thus increasing the confidence.

- **Lift :** Lift for a given rule A->B is defined as the ratio of the observed support to that expected if A and B were independent i.e.

$$\text{Lift} = \frac{Support(A,B)}{Support(A) * Support(B))} \tag{3.5}$$

  This signifies the likelihood of the product B being ordered when product A is ordered while taking into account the popularity of B. If the value of lift is greater than 1, it means that the product B is likely to be purchased along with product A, while a value less than 1 implies that product B is unlikely to be purchased if the product A is purchased.

- **Conviction :** Conviction for a given rule A->B is defined as the ratio of the expected frequency that A occurs without B if A and B were independent, divided by the observed frequency of incorrect predictions i.e.

$$\text{Conviction(A,B)} = \frac{1 - Support(B)}{1 - Confidence(A,B)} \tag{3.6}$$

- **Leverage :** Leverage for a given rule A->B is defined as

$$\text{Leverage(A,B)} = P(AandB) - P(A)P(B) \tag{3.7}$$

  Leverage measures the difference of A and B appearing together in the data set and what would be expected if A and B were statistically dependent. The rational in a sales situation is to find out how many more units (product A and B together) are sold than would be expected from the independent sales. The use of minimum thresholds for leverage simultaneously implies an implicit restriction on frequency.

3.4 Revenue Reports

Revenue is the total amount earned after completion of an order and all transactions associated with the order. It is the backbone of the business and an important component of any company's business model. The revenue report is created to track the growth of the business and improvement in sales. It is also important to see if the business is improving due to the various process optimization and automation tasks. The revenue of our company operates on the following models:

1. *Web-Shop Sales model :* In this model, a customer makes a purchase directly through our platform. As a result, the customer pays the company directly for the product, and no third party medium is required, and all services are provided to the customer directly by our company. All complaints and problems are handled and solved directly by our company.

2. *Channel Sales model :* The channel sales model includes all mkpls including the top seller platforms that we use to sell our products. This model is definitely a good complement to the Web-Shop Sales model as it helps us reach a large number of customers, establish our brand equity and operate globally. However, the channel receives a commission for each sale, which varies depending on the Mkpl. The channel acts as a medium between our company and customers. It can also impose penalties if any customer's dissatisfaction leads to damage the platform's image.

In our company, the revenue report is prepared on MoM and YoY basis. The yearly revenue report is an annual analysis of business growth and includes data for the current calendar week based on total units sold and sales achieved compared to performance in the previous two years for the same calendar weeks. Figure 3.9 shows the annual revenue graph for the 2nd calendar week for three fiscal years. From the graph, the growth of the company as compared to the the last two years has increased by approx. 39.2 %.

On the other hand, the monthly revenue report contains information segregated over WoW basic for that period. It records all units sold across all platforms and product categories for 6 weeks. Accordingly, WoW comparative analysis is performed to measure business performance and identify and address the issues if any decline in sales is observed. These reports directly reflect the performance of the business.

Figure 3.10, Figure 3.11, and Figure 3.12 are three monthly revenue reports for different months, showing fluctuations in units sold and sales generated in different periods, WoW. This fluctuations in the revenue and sales is due to different periods including Black-Friday.

Margin is the price difference between what our company and the customer pay for the same product. The margin for the sale of products may only include the actual cost difference and not overhead or other variable costs. Margin is the total difference between revenue and expenses. Margin ratios measure our company's ability to turn its sales into profits.[21]

Figure 3.13 shows the graph of the revenue report with the revenue above the margin. Our company maintains a difference of approximately 78% between margin and revenue. This includes the profit sharing of our manufacturer B+S. It can also be observed that in some weeks the number of units sold is lower, but the revenue generated is higher and vice versa. This is due to the fact that expensive product categories such as mattresses are sold in greater numbers than other, less expensive product categories. Similarly for less expensive product categories, even if the number of units sold are more it does not reflect incline in the revenue percentage.

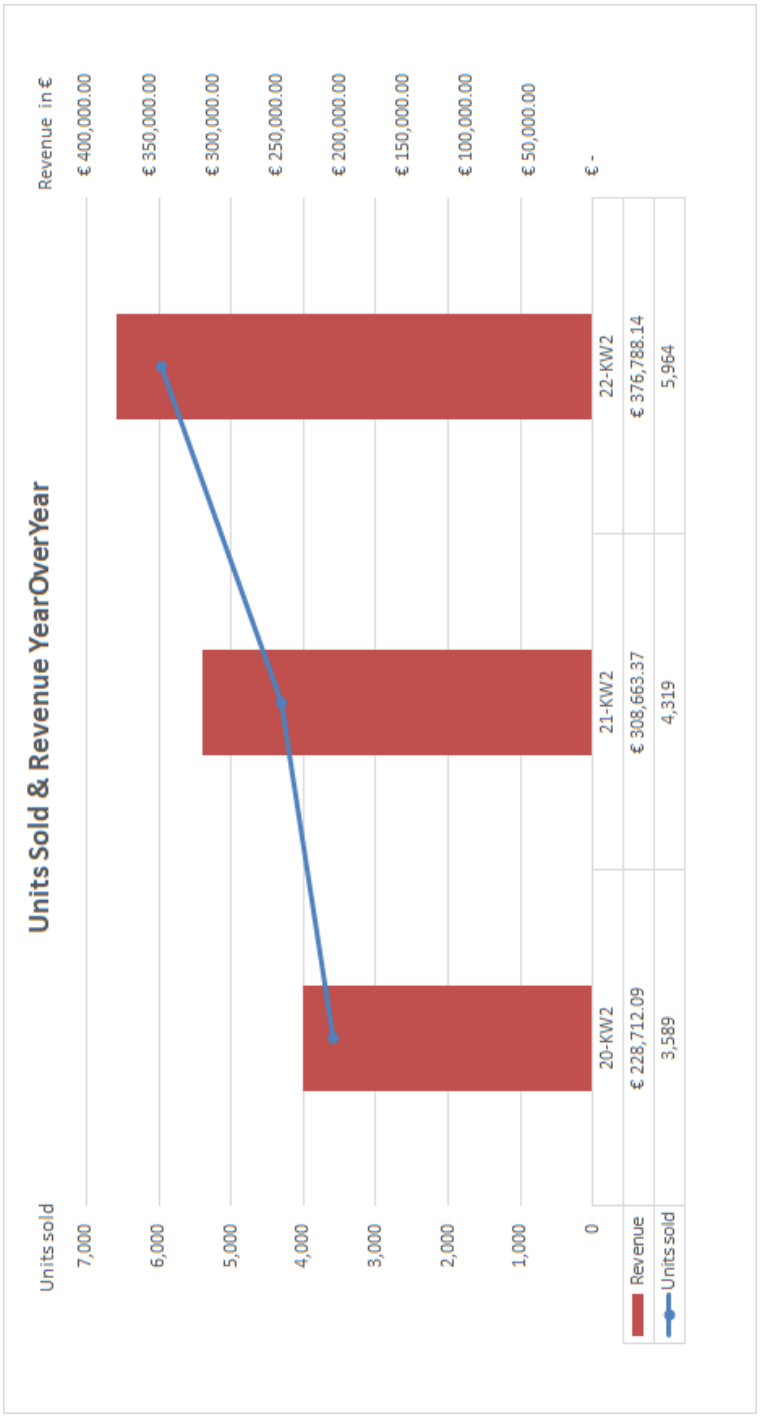

Figure 3.9: Revenue Report Graph, YoY

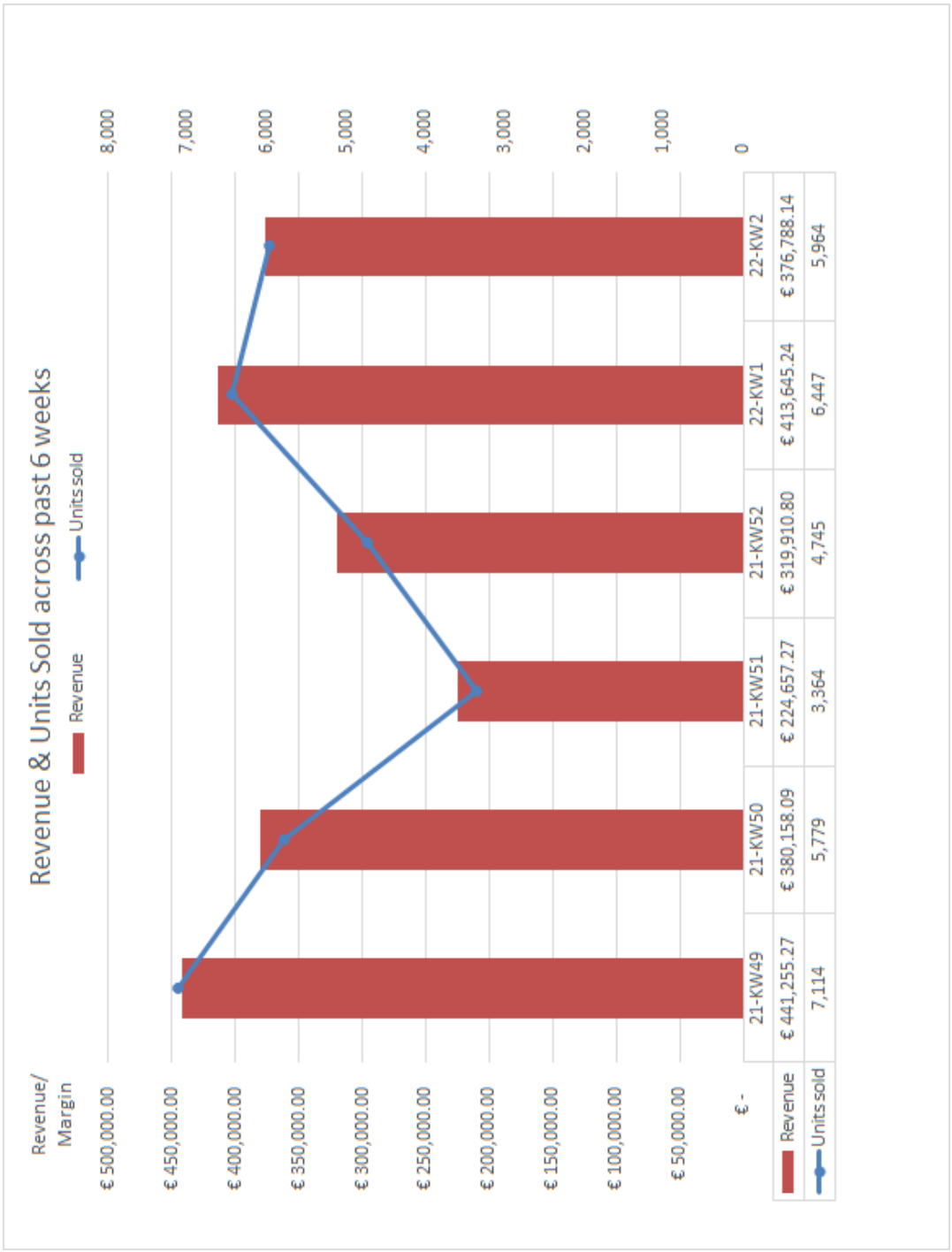

Figure 3.10: Revenue Report Graph, MoM W49 - W02

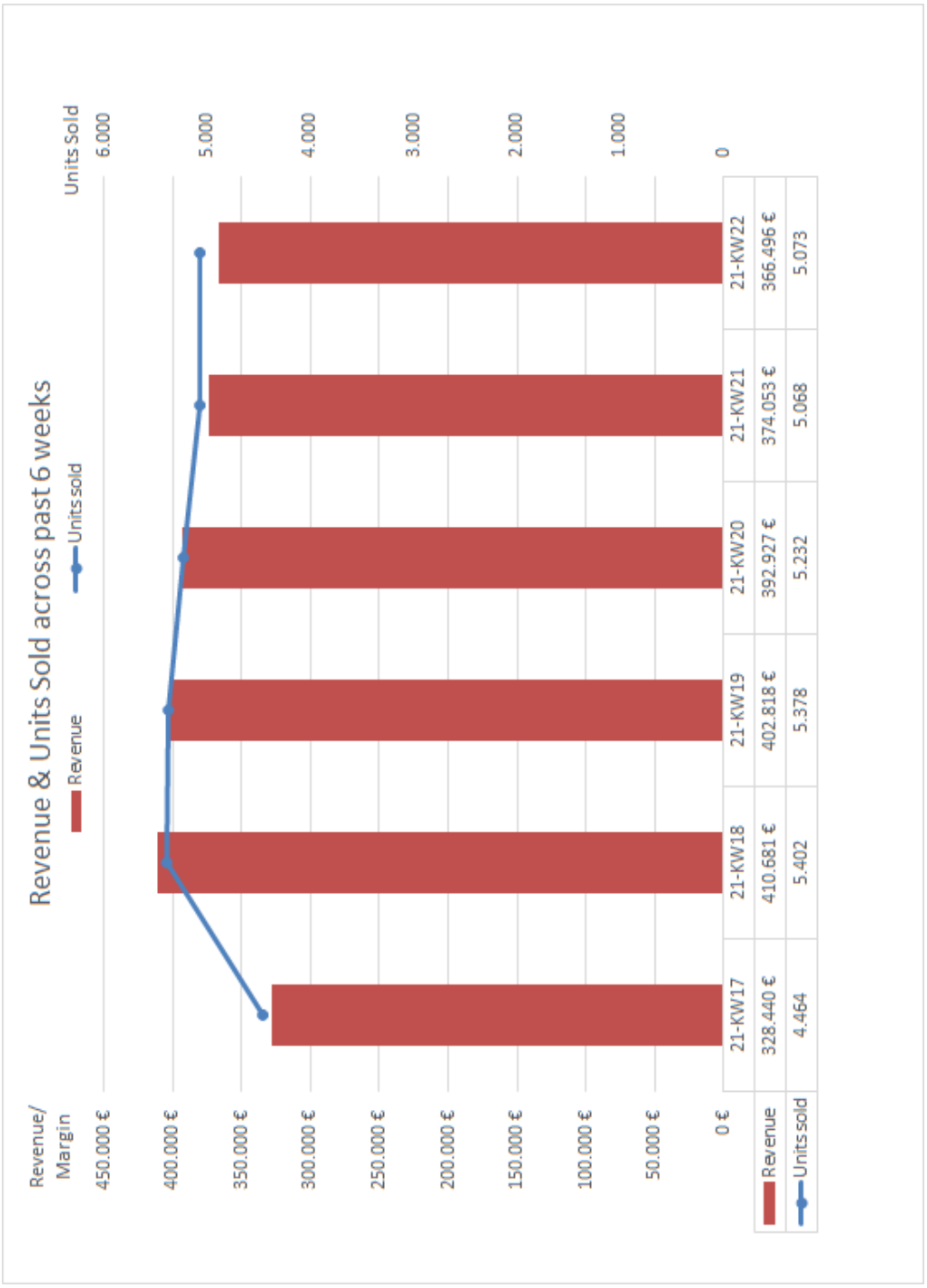

Figure 3.11: Revenue Report Graph, MoM W17 - W22

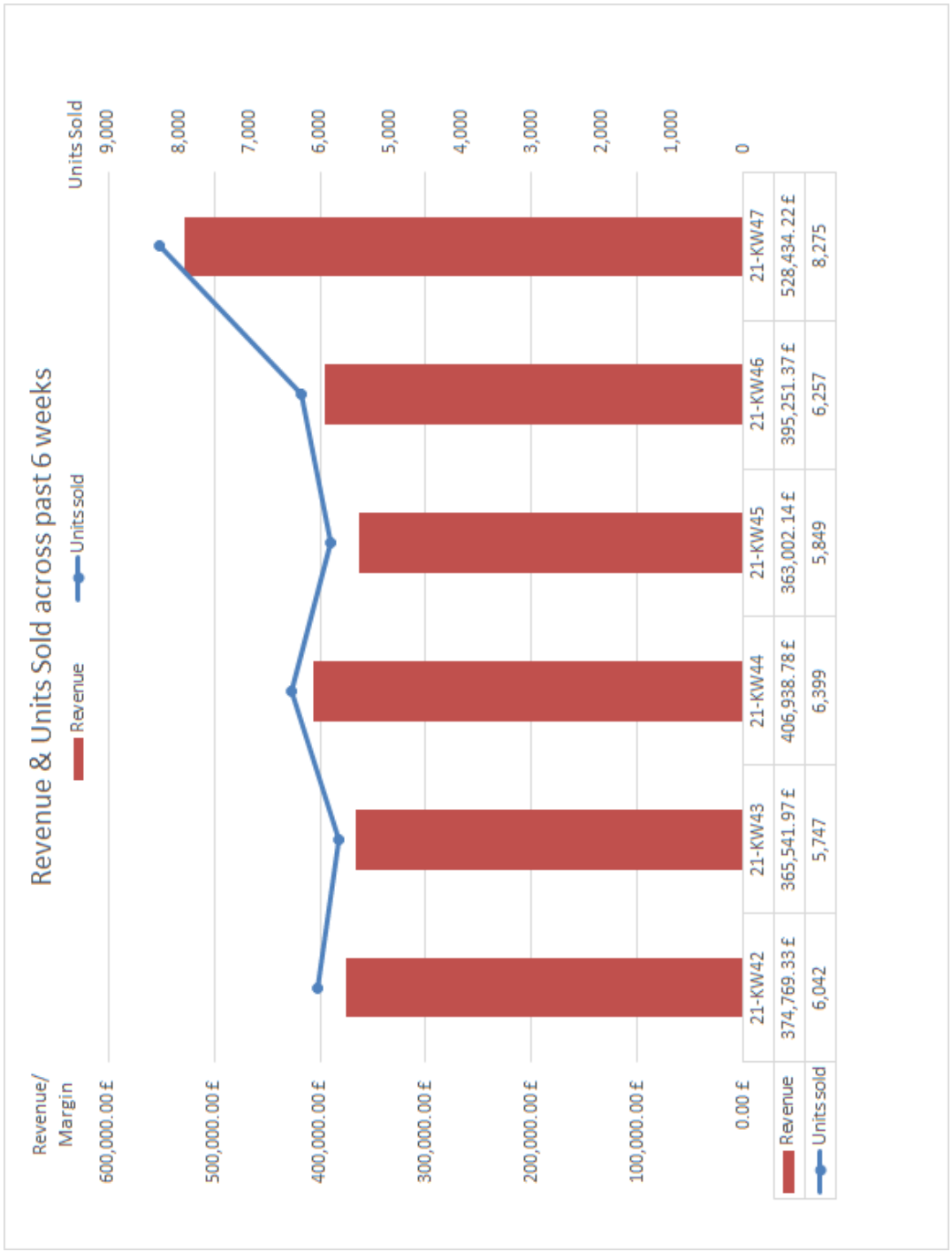

Figure 3.12: Revenue Report Graph, MoM W42 - W47

Figure 3.13: Revenue Report Graph with Margin, MoM W45 - W50

4 Results

4.1 Process Optimization

Process optimization leads to more efficient work. In today's world, process optimization is a prerequisite for managing complex business operations, providing better services and developing products. It is a comprehensive concept that provides solutions to all problems and improves the quality of business operation for it to have value and benefit to a customer, a company or other stakeholders.

As an E-commerce company, it is our duty to provide the best possible service to the customer, that's what it's all about. By optimizing processes, you become more effective as a company. You can act faster and become more flexible in terms of customer requirements. When you use the right software solutions, you have more complete and accurate information, so you can provide the best QOS to your customers. Of course, better customer service leads to more satisfied customers and is in the interest of the entire company.

Process optimization helped our company to work more efficiently by eliminating unnecessary steps and automating process steps to save time, reduce errors, and eliminate duplicate work. Especially for frequent processes that were daily tasks, it was successful in increasing efficiency through process optimization. It also helped the company to offer solutions or identify problems and prioritize customer requirements with the help of the right tools. Using the right software solutions helped to have more complete and accurate information so that the company was able to serve its customers in a better way.

Strictly organised and carefully documented processes form the basis for continuous improvement. The goal of this work was to optimise various processes that helped the company achieve outstanding improvements, as stated in Section 3.4. We saw that various data collection techniques were used and different techniques to analyse business data in different categories. Optimising these processes helped us achieve our business goals and establish our brand across Europe.

4.2 Process Automation

Automation plays an important role in process optimization. With the right software solutions, many processes can be made simple, fast and error-free. The first step to automating a process was to define the goal(s) and understand the objectives. This helped prioritise the tasks and workflows and create an automation strategy tailored to the business.

A common assumption about process automation is that the goal is always efficiency, but that is not the case here. The main reason for automation was:

· to add value to the product, service and company
· to have more predictable and accurate outcomes
· to improve the QOS for customers
· to create better outcomes and integrate all Mkpls
· to reduce repetitive work or tasks
· to reduce cost of the business

Some of the business processes were complicated because some subprocesses or certain related processes were implemented together to achieve the results. For example, the visibility analysis described in Section 3.3.4 involved automating the data collection from Google Shopping platform and then implementing the visibility analysis process. To automate this entire end-to-end process, it was important to understand the complexity of the task and create a step-by-step algorithm as described in Algorithm 2.1.

The following table will help you identify the tasks that can be automated based on the structure of the requirements:

Quality	Description
Repetitive	Involves the same steps or inputs each time;
Frequent	High volume task that must be completed multiple times in an hour, day, or week;
Recurring	Tasks must be completed according to a regular schedule;
Dependent	Triggered by another event or change in status;
Simple	Does not require complex information, decision making, or problem solving;
Predictable	Planned element of a workflow or process, occurs in every instance of the process;
Collaborative	Requires action or input from multiple stakeholders;

Table 4.1: Qualities of tasks that can be automated

From automating web data collection with various Python techniques and skills, as described in Section 3.1, to automating various analytics tasks performed on regularly operated business data sets, it helped the company overcome challenges and effectively reduce the cost of outsourcing these third-party services. It also helped secure customer data and prevent data theft.

Figure 4.1 and Figure 4.2 shows the business growth in terms of net sales and the overview of the periodic target inclination.

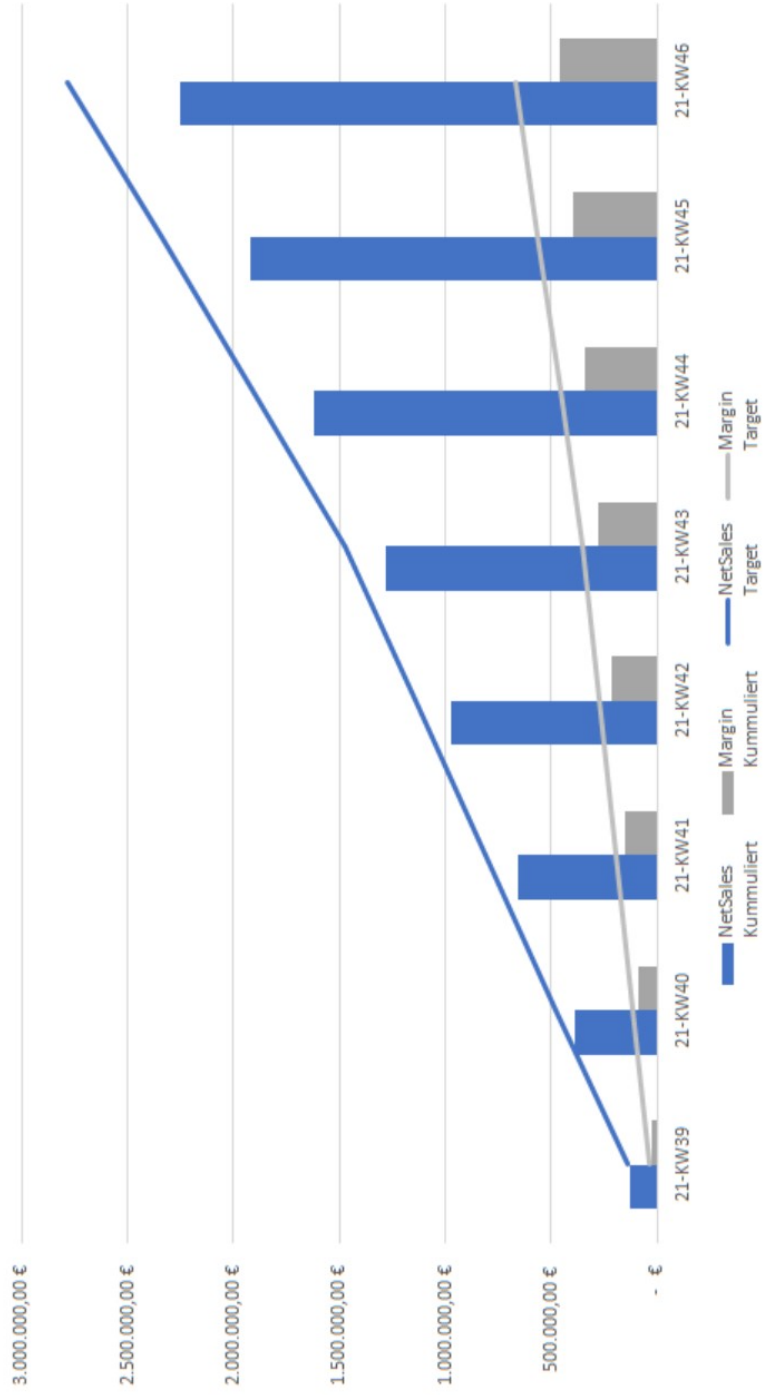

Figure 4.1: Overview - Net Sales Target, WoW

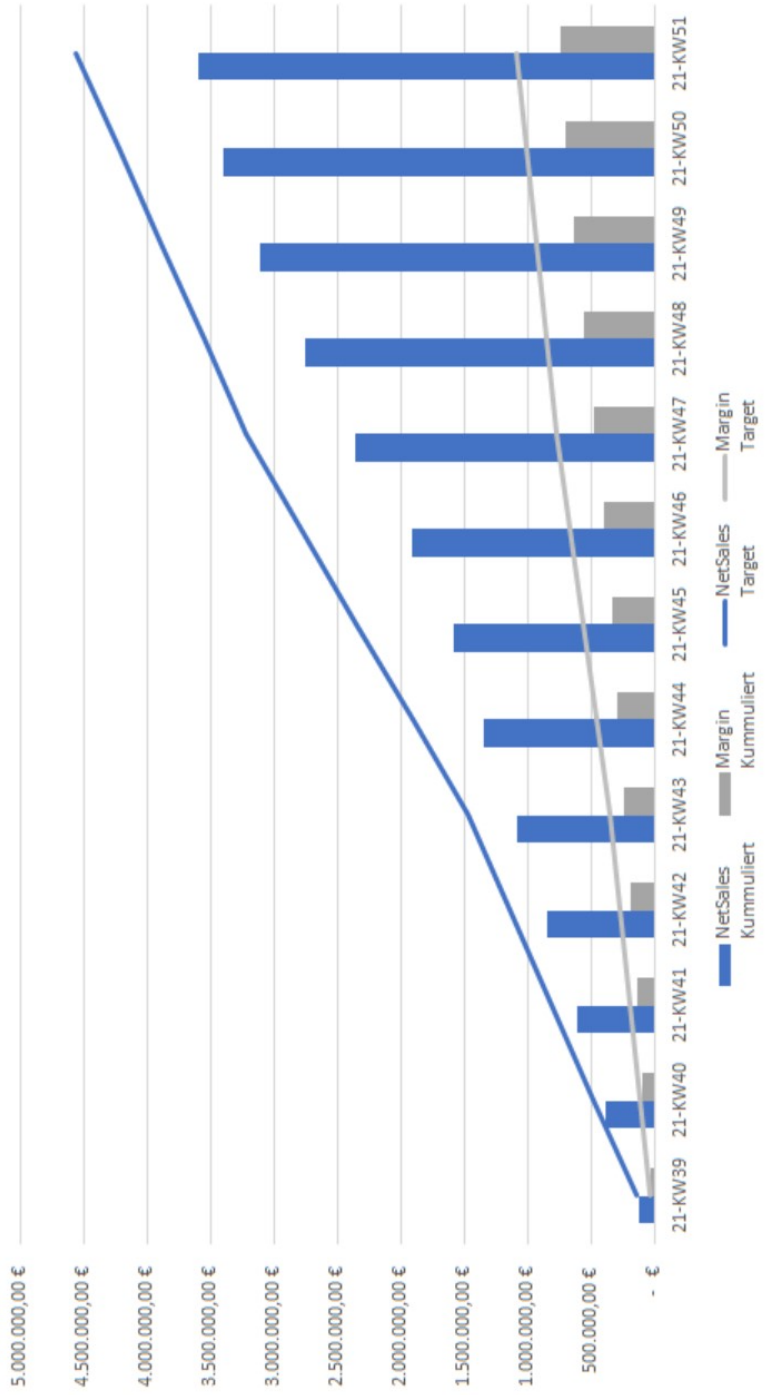

Figure 4.2: Overview - Net Sales Target growth, WoW

5 Conclusion

This thesis was about optimising and automating various business processes in an E-commerce company. The tasks mainly used Python, with additional tools like SQL, VBA and Macros. Python is a highly adaptable language that can be understood effectively and is also unusually groundbreaking. It helped us automate the process, achieve our goals, and meet the business needs of the company.

The first phase focused mainly on collecting unstructured data obtained from various mkpls websites. This raw data was crawled and scraped according to the requirements using different techniques and Python libraries and stored in the DB. There were many challenges in developing and maintaining some scrapers such as CAPTCHA and Denial of Service, which were overcome using APIs and other techniques as described in Section 3.1.

In the second phase, the stored data goes through a process of data integration, for example as in ODF creation, to homogenize this data with data from the company's platforms. The tedious process of structuring the external data requires the implementation of data preprocessing techniques and algorithms, as described in Section 3.2.

The third main phase considers the various data analytics and visualizations performed on these big business datasets to obtain a deep insight. The filtered, cleaned and structured data is used for various categorical analysis and is used to understand the business through graphical visualizations, periodic revenue reports to understand business growth and brand valuation. This helps us to develop appropriate business and marketing strategies to improve service quality to customers, increase profit margins and be competitive.

As discussed in Section 3.3.5 on the requirements and importance of analysing and understanding customer buying behaviour, it would be important in the future to gain some insights into orders and sales. We would be happy to implement Association Rule Learning and the Apriori algorithm. These algorithms can be applied at deeper levels and will help us better understand customer behaviour and order patterns.

References

[1] Understanding e-commerce and it's business operations. [Online]. Available: https://en.wikipedia.org/wiki/E-commerce

[2] Html. [Online]. Available: https://en.wikipedia.org/wiki/HTML

[3] Ghazvinian, Holbert, and Viswanathan, *Simple Web Scraping*, 1st ed. wordpress, 2015.

[4] R. Lawson, *Web scraping with python*. Packt Publishing Ltd, 2015.

[5] J. M. Patel, *Getting Structured Data from the Internet*. Apress Publishing, 2020.

[6] Web crawling and it's applications on online platforms. [Online]. Available: https://en.wikipedia.org/wiki/Web_crawler

[7] R. Mitchell, *Web Scraping with Python: Collecting Data from the Modern Web*. O'Reilly Media, Inc., 2015.

[8] Xpath. [Online]. Available: https://en.wikipedia.org/wiki/XPath

[9] Css selector. [Online]. Available: https://en.wikipedia.org/wiki/CSS

[10] S. S. Salunke, *Selenium Webdriver in Python*. Createspace Independent Pub, 2014.

[11] J. Duckett, *HTML and CSS: Design and Build*. Wiley, 2011.

[12] K. Dale, *Data Visualization with Python and JavaScript*. O'Reilly Media, Inc., 2016.

[13] Big data and analytics: top 5 python libraries for data science. [Online]. Available: https://www.simplilearn.com/top-python-libraries-for-data-science-article

[14] Phuong, Martin, Raman, and Ashish, *Python: Data Analytics and Visualization*. Packt Publishing Ltd, 2017.

[15] R. Story, *Python for Data Visualization*. O'Reilly Media, Inc., 2016.

[16] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, Inc., 2017.

[17] Customer review. [Online]. Available: https://en.wikipedia.org/wiki/Customer_review

[18] Page rank google shopping visibility. [Online]. Available: https://en.wikipedia.org/wiki/PageRank

[19] Association rule learning. [Online]. Available: https://en.wikipedia.org/wiki/Association_rule_learning

[20] Apriori algorithm. [Online]. Available: https://en.wikipedia.org/wiki/Apriori_algorithm

[21] E. Lewinson, *Python for Finance Cookbook*. Packt Publishing Ltd, 2020.

Acknowledgments

A big thank you to Matea, who has constantly and patiently supported me and always given me constructive feedback and encouraging suggestions. More generally, I thank all my colleagues at Mister Sandman for the family and supportive environment in which they welcomed me and which helped me overcome the difficult moments.

On the academic side, I thank my supervisor for his great availability and understanding, as well as for his rare teaching and communication skills, which he demonstrated in the course I attended. Nevertheless, I thank the entire faculty of the University of Padua for the knowledge and passion you have given me during these two years of Master's studies.

I thank my parents for giving me the opportunity to complete this cycle of studies. I am immensely grateful to my brother Gautam, to whom I am very attached. I thank my second family, my uncle and aunt for motivating and guiding me to achieve this growth, development and change for me. I believe I was able to see the distances as a feature of my social network and not as a barrier to relationships.

This dissertation is also a tribute to my grandparents and my aunt who left us at this time, but your blessings will always be with me.

Finally, I thank my colleagues at the University of Padua. I regret not having had the time I would have wished to connect with you. Nevertheless, you have left me with a good memory of my short stay in Padua.