INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Analytical Framework to Discover Churn Rate Insights

Patrícia Raquel Gil Lisboa Saúde

Master in Integrated Business Intelligence Systems

Supervisor:
Doctor João Carlos Amaro Ferreira, Assistant Professor with habilitation,
Iscte - Instituto Universitário de Lisboa

Co-Supervisor:
Doctor Vitor Manuel Basto Fernandes, Assistant Professor with habilitation,
Iscte - Instituto Universitário de Lisboa

October, 2020

TECNOLOGIAS
E ARQUITETURA

Analytical Framework to Discover Churn Rate Insights

Patrícia Raquel Gil Lisboa Saúde

Master in Integrated Business Intelligence Systems

Supervisor:
Doctor João Carlos Amaro Ferreira, Assistant Professor with habilitation,
Iscte - Instituto Universitário de Lisboa

Co-Supervisor:
Doctor Vitor Manuel Basto Fernandes, Assistant Professor with habilitation,
Iscte - Instituto Universitário de Lisboa

October, 2020

Acknowledgments

First, I would like to thank all my family, friends and work colleagues who encouraged me not to give up and start this project, while at the same time as I was developing another one in the company. My eternal gratefulness to my family and close friends, for the patience and encouragement during all the work.

I am grateful to the company for making the data available. Unfortunately, the support that I had inside the company leaving the company, so I had to develop the ideas alone.

I want to express my gratitude for my supervisor João Ferreira, who tried to give me the biggest support and incentive, always being available and giving guidance to finish this dissertation.

Patrícia Lisboa Saúde

Resumo

A rotatividade de colaboradores é um problema que pode ter um enorme custo anual para qualquer empresa, monetário ou temporal. Quando um colaborador sai da empresa, pode exigir que essa vaga seja preenchida de imediato. Essa substituição resulta em uma nova contratação que custa tempo (recrutamento, aquisição, formações) e que pode diminuir o impacto de uma empresa no crescimento da receita e expansão. Neste sentido, a retenção dos colaboradores torna-se uma estratégia crucial para a empresa, uma vez que pode reduzir custos em larga escala.

De forma a mitigar este problema, as organizações devem criar planos de estratégia para a retenção dos seus colaboradores. Para a criação destas estratégias, existe a necessidade em primeira instância de compreender e avaliar os motivos dos colaboradores a sair da empresa. Estes motivos podem ser validados com base na visualização/análises de dados históricos, existentes em uma organização referentes aos colaboradores.

As razões de um colaborador sair da empresa, podem se classificar em dois eventos. Por vontade própria, neste caso rotatividade voluntária, ou rotatividade involuntária, caracterizado por ser iniciativa da empresa.

Esta dissertação compreende a criação de uma *framework* para a análise dos motivos de saída de um colaborador de uma empresa. A *framework* proposta é criada com base em um projeto real de uma organização e, pretende desmistificar conceitos/premissas conhecidas, com base em revisão de literatura, referentes aos motivos de rotatividade dos colaboradores. Pretendendo validar se as premissas são aplicáveis ao contexto empresarial e se são verdadeiras. Adicionalmente a estas premissas, esta *framework* permite também estudar os principais motivos que levam à decisão de sair de uma organização (progressão de carreira) sugerindo algumas variáveis chave para a aplicação da *framework*.

Palavras-Chave: Rotatividade de colaboradores, Premissas, Visualização de dados, Correlações, Framework

Abstract

Employee turnover is a costly problem for any business. For every position left vacant, the time and cost associated with identifying, acquiring, and training the right person can lessen a businesses impact on revenue growth and brand expansion [1].

Employee retention becomes a crucial strategy for the company since it can reduce costs. To mitigate this problem, organizations must create strategy plans as a way to retain their employees. To create these strategies, in the first instance, there is a need to understand and evaluate the reasons employees leave the company. These reasons can be validated based on existing reasons, or by creating analyses of the organization's existing data regarding employees. The reasons an employee leaves the company can be classified into two events. The first can be initiated by the employee. In this case, it is a voluntary turnover. The second one is an involuntary turnover, characterized by the initiative of company.

This dissertation includes the creation of a framework for the acknowledgement of the reasons for an employee to leave the company. The framework is created based on a real project of an organization and intends to demystify common misconceptions/premises, based on literature review, regarding the reasons for an employee turnover. It proposes to validate if the misconceptions apply to the business context and if they are true. Additionally, this framework also allows the study of the main reasons that lead to the decision to leave an organization (e.g. search for career progress) by suggesting some key variables for the application of these analyses to understand how they affect employee churn.

Keywords: Employee Churn, Misconception, Visualization, Correlation, Framework

Index

Tables Index

Figures Index

List of abbreviations

CRISP-DM	Cross-Industry Standard Process for Data Mining
HR	Human Resources
KNN	K-Nearest Neighbors
MYR	Middle Year Review
MYR12	Middle Year Review Twelve
MYR13	Middle Year Review Thirteen
MRY14	Middle Year Review Fourteen
MRY15	Middle Year Review Fifteen
MRY16	Middle Year Review Sixteen
MRY17	Middle Year Review Seventeen
MRY18	Middle Year Review Eighteen
MRY19	Middle Year Review Nineteen
SVM	Support Vector Machine
YER	Year End Review
YER12	Year End Review Twelve
YER13	Year End Review Thirteen
YER14	Year End Review Fourteen
YER15	Year End Review Fifteen
YER16	Year End Review Sixteen
YER17	Year End Review Seventeen
YER18	Year End Review Eighteen
YER19	Year End Review Nineteen

Chapter 1 – Introduction

1.1. Overview

Employee churn is an occurrence in which an employee leaves the company (also known as employee turnover, employee retention or attrition).

Employee turnover is painful for a company and can have a very high cost. The employee turnover rate is approximately 12-15% for IT service organizations, in particular [2]. This churn rate is very high and, assuming a lower churn rate of 5%, the cost of an employee leaving a company is around 1.5 times an employee's annual salary.

When employees leave, it costs the organization time and money. Some example costs associated with turnover are, accrued paid time off, staffing costs such as time and effort expended in recruitment, selection, orientation and training [3]. Customer trust may also be directly affected, causing client dissatisfaction and work delays [2].

Many companies are also highly concerned with their ability to maintain key employees (e.g. high performers and high-demand or hard-to-replace skill sets of employees) since these have a much higher cost.

Understanding when employees are more likely to leave will lead to steps to boost employee retention as well as potentially scheduling new recruiting in advance.

While shared characteristics and results can be correlated with each turnover occurrence, there are different types of turnover, each one with it is own effects. It is possible to define the sorts of churn across three: categories, involuntary, voluntary and retirement.

While voluntary turnover is started by the employee, involuntary turnover is a decision initiated by the company, often related to unsatisfactory job results or organizational restructuring. Typically, retention management focuses on voluntary turnover, because these employees are often people that the organization would prefer to retain.

However there is an important distinction between functional and dysfunctional turnover [4]. Dysfunctional turnover, such as the departure of high performers or employees who have difficult-to-replace skill sets, is detrimental to the company.

Although it has costs and it is disruptive, functional turnover is not always harmful. The exit of easy-to-replace employees or bad performers, may even be helpful.

The decision to voluntarily leave the organization can be centered around multiple factors, so it make sense to analyze and to understand the reasons for an employee to leave a company.

With factors ranging from age, tenure, salary, work satisfaction to education, recognition, burnout and several other causes, there is no consensus about what are the main triggers for churn. Because each company has different people, it is likely that no company is similar, and only historical data can be used to better assess the HR management decision plan [5]. Research is crucial for the performance and good management of the organization. By understanding the causes that drive an employee to leave, it is possible to create mitigation plans and anticipate their exit.

The need for an automated and algorithm-power solution, based on Machine Learning (ML), is requested to address talent retention, provide explanations for churn and insights to engage employees. This provides the possibility to create a framework to discover generic insights to employee churn reasons that can also be applied to multiple organizations' contexts.

1.2. Motivation

Employee churn has become an important part of any organization's strategy due to the associated high costs and the negative impact it has on productivity.

Each company has different costs and employees with different skills, so it's necessary to study and analyse all history a company has of its employees, in order to understand which factors are applicable and which lead the employee to leave. To this end, it is possible to create a transversal framework, applied to any company, as long as it has a set of historical data of the employee's permanence in the company over several years.

It should be taken into consideration the project overview in the previous section, that proposes creating a transversal framework applicable for any company with specific historical data of the employee, such as skills employees, performance evaluation, age, hire date, date of leaving the company, promotions, work hours, dissatisfied indicators, among others.

The framework should guide analyses based on the standard dataset, to form conclusions regarding the reasons for an employee leaving each company.

Having the opportunity to act towards maintaining established potential churners proactively makes it possible to translate the high cost associated with these events into a high return on investment (ROI) [6].

New advances in Data Science, along with public libraries available in tools like python, propose to explore crucial assumptions/decisions results and try to discover reasons and new solutions to the problem of employee retention. This solution aims to make processes more powerful and decisive, in order to reduce the churn rate of the company and allowing to create HR plans that anticipate and offer some benefits. This way the employee feels more comfortable in his company and does not leave and consequently reduces the costs.

1.3. Objectives

Taking into consideration the motivation proposed, which is based on a real case-of -study, a data-driven approach is used to understand collected data of a real-problem of an organization. The purpose is to discover the reasons why employees leave an organization. The project priorities split into two main subjects.

The first objective is to gain a general understanding of a data set associated to a historical data of employees over several years, through numerous exploratory data analysis. This can helps to identify the variables in order to analyze the risk of employee churn and also retain them once identified. The main goal is to determine a set of "key variables" to create a framework across all organizations that can be reused by any company. Meaning, the objective is to identify a collection of main variables that must be available in historical databases associated with employee information of each company. After establishing the set of variables, the goal is that any organization with real data can apply the analyses that the framework developed throughout this real-case-of-study detail.

The second objective with this real case of study is to develop an extended generic methodology to identify the different reasons for churning employees, based on the CRISP-DM methodology and process model.

Based on research [1] that describes several misconceptions focused on a shared understanding of employee turnover reasons, knowledge of cause-and-effect relationships, and the ability to adapt this knowledge, the objective is to understand if these misconceptions are

applicable to disparate companies' contexts and if they are true (e.g. growing job dissatisfaction, quit because of pay, etc.).

After analysing the misconceptions presented in the research described above, the last objective, is to understand the data and create rules that allow to demystify new assumptions applied to the real case of study in question, such as: if the employee leaves due to not being promoted or due to an event happening in recent months, or if the decision to leave the organization is due to an external factor to the data.

1.4. Methodology approach

The goal is to find answers regarding the reasons for an employee to leave an organization, creating an analytical framework that can be applied transversally to any company.

When building up a framework, it is very significant to start with a known methodology. During the development of this methodology, it is necessary to ensure that it follows a set of rigorous, systematic and general processes and stages that allow to reuse the framework in the context of any company, based on key information related to employees.

One of the most popular methodologies for increasing the success of DM projects is the Cross-Industry Standard Process for Data Mining (CRISP-DM) [7]. This approach is adopted by several researchers in the past to get expected results and focusing on delivering real business value.

The methodology [7] consists of a non-rigid six steps sequence that enables the development and implementation of a DM model to be used in a real environment to support business decision-making.

Figure 1 displays the life cycle of the CRISP-DM model and following it is described the six phases of the cyclic process [7]:

1. **Business understanding:** Focuses on identifying the priorities and requirements of the project from a business perspective, and then translating this information into a description of the problem of data mining and a preliminary plan.
2. **Data Understanding:** Start with an initial collection of data and continue with activities to get acquainted with the data, to identify issues with data quality, to discover first

insights into the data, or to identify interesting subsets to shape hypotheses for hidden details.

3. **Data Preparation**: The data preparation phase covers all activities to collect and construct the final dataset from the initial raw data.
4. **Modeling:** techniques are selected and applied. For the same form of data mining problem, there are several methods. There are unique criteria for certain techniques in the form of data. Therefore, return to the data preparation stage always if it is necessary.

Figure 1 – Phases of the CRISP-DM process model (Adapted from [8])*[1]*

5. **Evaluation:** When one or more models have been created that seem to have high quality based on whatever loss functions have been chosen, they need to be checked to ensure that they are generalized against unseen data and that all main business concerns have been properly taken into account. The end result is the champion model(s) range**.**
6. **Deployment:** The implementation phase may be as easy as producing a report or as complex as implementing a repeatable data mining method across the enterprise, depending on the requirements. The features and outcomes of the previous steps are described.

[1] Image from site: https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome

The above steps describe only the scientific definition of the CRISP-DM process model. In Chapter 3, the framework intends to describe in detail, taking into consideration the real case of study and also each phase of the model previously described.

1.5. Structure and Organization

This dissertation is composed of 6 chapters, including Introduction chapter and the structure is organized the following way:

Chapter 1 (current chapter), introduces a context work, motivation and a set of objectives to be achieved.

Chapter 2 presents the literature review. This chapter consists of research for similar projects in academic or HR contexts. Try to extract popular misconceptions/assumptions related to the reasons for employees churn.

Chapter 3 describes the entire stage of the CRISP-DM model process applied to the real problem is detail.

Chapter 4 visualization of data of the most important variables, in order to understand the dataset and identify indicators to be considered in the next chapter.

Chapter 5 in this chapter, the author attempts to discuss misconceptions derived from the literature review of chapter 2, contrasting whether the misconceptions are relevant and validating whether they are true reasons of employee churn for this real case of research.

Chapter 6 has the work's conclusions, starting with the work done and goals, then going to the author's suggested discussion of the outcomes and future work.

Chapter 2 – Literature Review

Based on the research of many articles, this chapter aims to analyze the solutions found in the literature review, that indicate the most well-known reasons for an employees' leaving a company and also, to justify the options taken in the development of the methodology.

The author S. Harrison and P. A. Gordon [9] has a scientific and mathematical basis, consisting of a study of several known concepts to help companies to retain their employees. This article describes several turnover myths with recommendations for evidence-based retention management strategies focusing on mutual turnover awareness, knowledge of relationships of cause and effect.

First, the author introduces the definition and causes of turnover, then focuses on the purposes of the dissertation, that is the research of three misconceptions. The misconceptions are explained based on the literature review.

Then, is introduced other possible approaches that help to understand and extract new premises if the dataset has key-variables.

Finally, displays similar papers to the real-problem of this dissertation, including the explanation of machine learning algorithms and analyses performed in other projects.

2.1. Turnover definition

Employees are leaving companies for several different reasons [10], such as taking a better paying job, leaving an abusive boss, returning to school, pursuing a relocating partner, or being fired. While similar features and outcomes can be correlated with each turnover occurrence, there are different forms of turnover, each with its own consequences.

The authors in papers [9] and [10] argue that employee churn can be voluntary or involuntary. The organization initiates involuntary turnover, while the employee initiates voluntary turnover [11]. Involuntary turnover happens mostly because of poor work results or organizational restructuring.

Typically, retention management focuses on voluntary attrition, and these employees are also workers whom the company would like to maintain.

Nevertheless, also voluntary turnover situations, there is a major gap between dysfunctional and functional turnover [12]. Dysfunctional turnover, such as the departure of high performers or employees who have difficulty-to-replace a set of skills, is detrimental to the organization.

Although disturbing, efficient turnover, such as leaving employees who are easy to replace, can even be useful, as leaving poor performers. Retention management strategies generally focus more on dysfunctional turnover.

2.1.1. Turnover costs

Voluntary turnover, it can bring many costs to the organization, such as time and money [13].

The basic methods of measuring turnover costs [14] concentrate on calculating the expense of replacing an employee in terms of the percentage of the annual salary added, along with the cost of benefits.

There are a wide variety of other direct and indirect costs associated, additionally to the obvious direct costs associated with turnovers, such as unpaid paid time off and staffing costs associated with recruiting a replacement [15], [16].

The company may have costs like [9], [3], [14]:

- HR operate time (e.g., benefits enrolment, selection and recruitment)
- Recruitment
    - Hiring inducements (e.g., bonus, relocation, perks)
- Selection
- Hiring time for managers (e.g., feedback on new hire decisions, orientation, training)
- Orientation and Training
    - Orientation program time and resources
    - Formal and informal training (e.g. Time, materials, equipment, mentoring)

Additionally, associated with the costs, [1] many organizations are also concerned about their ability to retain key employees (e.g., high performers and employees with high-demand or difficult-to-replace skill sets).

2.2. Main reasons for churn

There are many different explanations pointed out in the literature as drivers for employee turnover. There are also authors E. Ribes, K. Touahri, and B. Perthame [17] and S. Harrison and P. A. Gordon [9] argue that churn factors depend by company, i.e. every organization is different having its own different drivers for churn, and there may be no consensus for what the main drivers of turnover.

The explanations for the churn found in the literature differ from different grades. And the author in paper [17] recognize that the key causes for churn were employee behavior and low performance and that new-joiners and collaborators who have been in the same grade for more than four years were also at risk of churn.

Additionally, job design and the work environment are moderately linked to turnover, such as, job satisfaction, job variety, advancement opportunities, communication, and decision-making participation; organizations that can design jobs, and the environment that complies with these findings can realize enhanced retention.

The focus of this dissertation is the misconceptions following detailed and, that is extracted to the scientific paper [1].

- **Misconception #1: "People Quit Because of Pay"**

When asking about why an employee leaves the company, the pay is almost always the first or second reason given. Some people do quit because they are unhappy with their pay. The fact that people often stop taking higher-paying jobs elsewhere is also true [13] and [18]. However, the author [1] also defend that pay may not matter as much as expected by many executives.

Although compensation is significant, pay level and pay satisfaction are generally relatively weaker predictors of individual turnover decisions; pay raises may not always be the most successful way to tackle turnover problems.

- **Misconception #2: "People quit because they are dissatisfied with their jobs"**

The authors [1] and [9] argues that there are various reasons for employee dissatisfaction and that there are also various ways of retaining them, according to each reason.

It is possible to derive the reasons for dissatisfaction [19] from work-related (e.g. quit because for a promotion) or non-job-related (e.g., spouse offered an opportunity in another place). Also, it can be positive, negative or neutral. A positive example is receiving a good job offer, and a negative receiving a negative performance evaluation.

Additionally, have employees who leave despite being relatively satisfied. These are likely impulsive quits, generally in response to adverse shocks such as passing over for promotion.

Job stress and job satisfaction are generally correlated with employee turnover in other jobs domains. The author S. Harrison and P. A. Gordon [9] correlates stress as a cause of burnout. When there is a disparity between any area of the work environment and the perception of the employees about their work, employees can experience burnout.

Workload, culture, power, reward, fairness, and values are the vital areas of the work environment where mismatches can occur. The workload is the amount of work that can be done by an employee during a given time.

The greater the difference between some part of the work environment and the person, the more likely they will experience burnout.

Finally, M. Phil C. Bryant [13] believes that the key aspects and consistent predictors of employee turnover decisions are job satisfaction and organizational commitment.

Organizations should recognize both employee satisfaction and work engagement to be measured and controlled.

T. Y. Park and J. D. Shaw [11] argues that perform a built-in reward system for managers who keep good people can be a plan to reduce the high costs of employee leaves a company.

- **Misconception #3: "There is little Managers can do to directly influence turnover decisions"**

Many executives are of the view that most voluntary turnover is inevitable. It is true that some turnover cases are inevitable; however, there is proof of unique cause-effect relationships and strategies of human resource management that can help companies handle turnover [1].

In terms of the recruiting, training, and socialization of new workers joining the company, there are clear evidence-based strategies managers may employ. The risk of eventual turnover

is reduced by recruiting strategies that offer candidates the most detailed image of the company, such as accurate job previews and referrals by current employees.

Selection strategies that determine candidates who suit the job and organization, as well as the use of weighted application blanks, allow people who are more likely to stay with the organization to be employed.

In the crucial first year after organizational entry, socialization activities that include links to others, constructive feedback, and consistent knowledge also minimize the risk of turnover [1]. Given the essential role of supervisors in many turnover decisions, it is also possible to minimize turnover by delivering appropriate leadership preparation, integrating retention measures into manager reviews, and handling toxic or abusive supervisors effectively.

Some approaches include the provision of autonomy and diversity of responsibilities, the promotion of a team atmosphere [13], the provision and promotion of unique demanding objectives, and the appreciation of employee contributions. Socialization operations that involve ties to others, positive feedback, and clear information often mitigate the risk of turnover in the critical first year after organizational entry [13].

- **Other approaches**

S. M. Abbasi and K. W. Hollman [16] believes that employee motivation [20] is the contrast between good and poor organizations. When employees are motivated with your job in your company, their personal growth and long-term potential, they can commit to the company. Maintaining a client's good relationships and, develop new products or services. In order to give reasons for an employee to continue and to stay and achieve their career aspirations [19], the organization should build programs to enhance their employees' competencies, skills and intellectual development.

The author also introduces indicators as an organization requires the employee to choose between a life and a career; organizations look at its employees and sees them as benefits, costs, salaries and, overhead and do not see their employees as assets. Despite the inevitability of turnover, many experts agree that managers should implement appropriate plans to minimize turnover [16]. Managers must show a character-based approach to the employer-employee relationship to promote successful employee relationships. Additionally, at all times, they must display dignity and fairness.

- **Variables:**

R. Cheripelli and P. V. Ajitha [15] shows that if the company has a dataset with a set of data, such as the attributes like employee's salary, current position, promotion, performance evaluation, number of hours worked, dissatisfied indicators, project indicators, among others [21]. They can show conclusions such as employees with no promotion in last five years tend to leave; Employees having high salary but no promotion and high working hours also tend to leave; Employees getting promotion but not hike in salary also tend to leave.

In the paper J. A. Grissom, S. L. Viano, and J. L. Selin [22], it is possible to extract key-attributes based on real-case-of-study applied to the turnover in the public sector. The author in the scientific journal, extract insights from teacher mobility and identify some characteristics to an employee are more likely to leave than others. The author in the paper [22] reinforce the importance of salary variable, since it is crucial to understand if the employee leaves an organization because of pay salary.

Through the papers [23], [22], [15], [21], [24], [25], [6] and [18] studies about what is commonly known as HR Analytics and Employees Analytics are presented, to understand the motivation for employee leaving a company. Also, it is possible to extract the key-indicators of an employee leaving the company, based on many cases-of-study applied to an organization in different industries. The variables extracted are a salary indicator, promotion, education, job, performance evaluation, gender, career opportunities, age and experience, a distance from home, work/life balance, benefits, marital status, current manager, overtime, churn, business travel frequency, years at the company, years with current role, environment satisfaction, hourly rate, relationship satisfaction indicators and, among others.

2.3. Data Analytics

The author in paper [26] shows the importance of advanced data analytics that allows an organization to have a 360 overview of their employees and operations. The insight extracted such analyses is then used to guide, optimize and automate their decision-making to achieve their organizational objectives effectively.

The paper [6] applied correlations between 26 variables studied relate to turnover. Author correlated turnover variable with other different variables. It is classified the correlations in

external factors, personal characteristics and structural or work-related factors. The first factor, external correlates, contains variables such as employee perceptions, vacancy rate, organizational size, and work-until size. In the Work-related correlates includes salary variable, performance evaluation, grade, overall job satisfaction, satisfaction with pay and among others. Finally, as personal correlates, it has age, tenure, gender, education, marital status, number of dependents, skills, and among others. The review displays, such as factors as the employee frequently will have a significant impact on turnover. It is a challenge to compare various correlates, but some conclusions are extracted from the meta-analyses.

Also, demonstrated that hard to examine turnover across several different industries. However, it is possible to apply a framework with analyses, correlations and results, if have variables clearly documented.

In the paper [24] is applied support vector machine, that is a predictive model of the supervised machine learning algorithm. First 22 features are collected from Human Resource databases of three IT companies, then categorize the features in continuous (Age, Distance from home, Education, among other), categorical features (attrition, business travel, department, gender, grade, among others) and a target value (Attrition). After that, performed analyses between churn variable and the other variables. Lastly, applied the SVM model and showed that the accuracy results from the confusion matrix for the model is of 85 per cent. The results explain that the SVM model predicting better who is leaving than who is not leaving the company.

Chapter 3 – Proposed Framework & Methodology

The purpose of the current chapter is to define the phases of a transversal framework to discover the reasons of employee churn in management in the IT services industry, based on CRIPS-DM model process. The first bulk of work is to correctly understand features between data exploration through historical data of employees churn and features associated with the employee on the real case of an organization. To be converting into generic and transversal outputs for any actual organizational context problem. Meaning, in each sub-chapter, Business Understanding, Data Understanding and Data Preparation seek to explain the real cause of study associated to IT organization-related research and try to extract generic features that can be implemented cross-cuttingly in any organizational, provided them follow the suggested steps.

3.1. Business Understanding

In this chapter and as suggested as the first phase of the CRISP-DM model process, the entire understanding of the project should be described. From the identification of key questions to the research of features under study.

What is proposed is to create an analytical framework that can be reused by any company within its business industry. The key task is to choose and understand the research's business intent, through exploratory analyses so that the data can be understood. Another task in this process is to come up with the questions about the findings & impact the study is going to make. The purpose is the validation of misconceptions to the context of each organization, taken from a scientific paper. Meaning, intendeds to validate whether the following premises are true and applicable to any business context: “People quit because of pay” and “People quit because they are dissatisfied with their jobs”.

A second objective is to explore and create rules to discover new assumptions for sharing with the company the common reasons for an employee's departure, based on historical data.

This research would enable companies to establish action plans and strategies and make informed decisions for the retention of their employees. Not only that, but the aim is to provide the company with a study over the years of what were the most frequent reasons for the employee to leave the company, directing the research to analysis to understand the real reasons.

3.2. Data Understanding

According to CRISP-DM model process, this sub-section is intended to formulate framework steps to data understanding that can be applied for each company context.

Additionally, the author tries to apply the framework in a real problem, according to the attributes available in their dataset.

3.2.1. Data Understanding - Framework

In this section, intended to define some key variables for the description of this problem and analyses that can be made.

First, needs to identify in their dataset the number of attributes have and observations. Some variables must exist in order to analyse churn employee reasons. Among these variables, the papers [25] and [18] suggest the following variables: Age; Distance from home; Education; Employee Count; Employee Number; Environment Satisfaction; Hourly rate; Job level; Job satisfaction; Salary; Churn; Gender; Marital Status; Overtime; Business Travel Frequency; Performance evaluation; Training times last year; Number of companies worked; Total Working Hours per day; Years since last promotion; Years with the current manager.

Second, based on data and histograms for numerical features, a few observations can be made. Before fitting a model to the data, the following 3.3.1 describes the needs of data transformation methods that may be required to approach a normal distribution.

Once the variables are known, the next step should be to identify the target value. In this case, the target value is the “churn” variable. Then develop analysis with the churn variable and with other variables. Following are some analyses that can be developed:

- Distributions for employees churn and no-churn per age. To understand the range of age that employees leave the company.
- Try to understand the relationship between churn and business travel frequency, to understand whether the case that the employee travels a lot can have an influence on employee leaving the company.
- Analyses such as checking marital attribute status and gender can be performed to see if you are more likely to leave the company if you are married and; which gender may be more likely to leave.

- Years with Current Manager is a good variable to try to understand misconception "There is little managers can do to directly influence turnover decisions". It allows a relationship to be made of whether being with the same manager can influence the retention of the employee in the company.
- Overtime is a variable that allows verifying the dissatisfaction of the employee, the possibility of burnout, workload, etc.
- Distribution for Years since last promotion and Churn variables allows to understanding if employee leaving the company because they are not promoted.
- Environment Satisfaction, Total Working Hours per day and Job satisfaction variables, allows demystifying the applicability misconception "People quit because they are dissatisfied with their jobs" in any organization.

3.2.2. Data Understanding applied to the project

The first important objective to complete identified in subchapter 1.4 is to understand the features better. An extensive data exploration process was carried out to this end.

First, various plots were created to understand the population, to analyze the data stored in the dataset and the volume of information. According to the previous chapter, these variables must be compared with "churn" variable. However, according to the dataset, the author performed these analyses in Chapter 4. In this chapter, focus on understanding the variables that dataset contains.

The original information contained in the dataset was about the performance assessments of the employee along his journey in the company. Having indicators such as: the number of intermediate assessments; the period during which each evaluation took place; the ID of the appraiser; the grade of the assessment received; the professional category in each period; among others. It is important to incorporate some principles to be able to understand some of the analysis during this research.

Throughout the analyses, many references are made to the YER and MYR terminology which correspond to the evaluation periods occurring in this case of study. The moment of performance evaluation is called by Performance Review and occurs in two distinct moments, the first evaluation occurs in the middle year (MYR) review and the second one in the year-end

review (YER). The rate of PR determines the employee promotion decision at the Year-End-Review (YER) moment.

The following Figure 2 shows a clear overview of the data that the dataset contains, relating to the number of employees with an evaluation performance and the corresponding evaluation period.

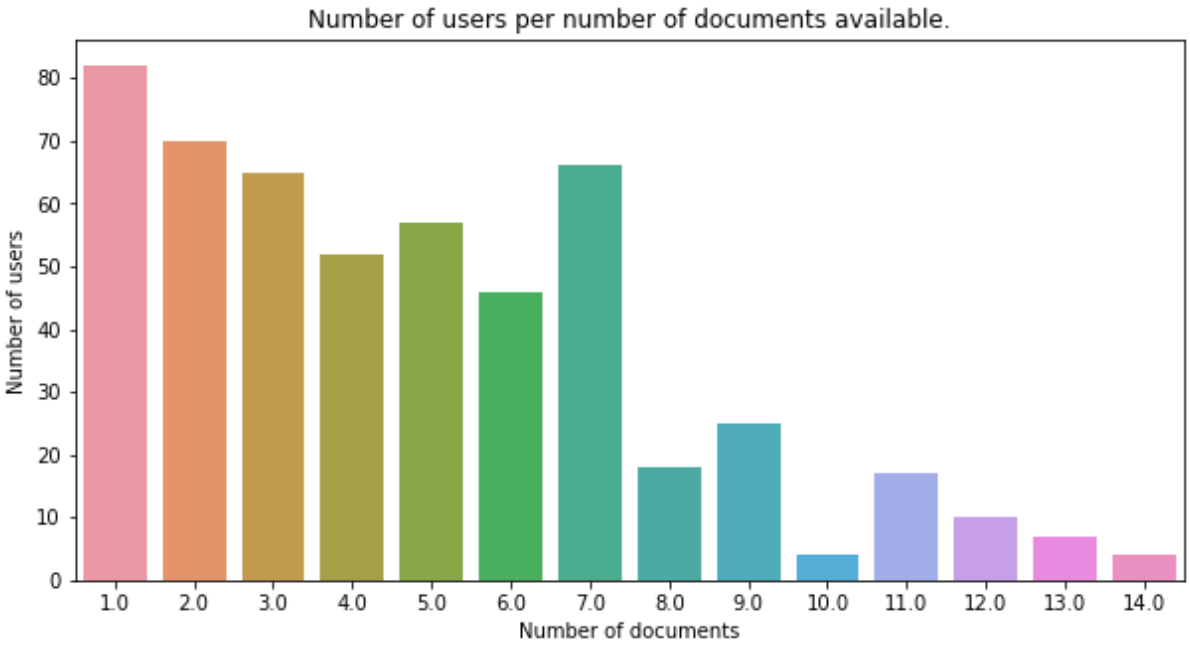

Figure 2 - Number of employees and number of documents per employee

According to graphic and other outputs, it is possible to observe around 2771 documents, corresponding to performance evaluation over a number of periods and have a population of about 563 employees between churn and no-churn. In other words, the total number of employees that leaves and stay of the company are 563.

After this explanation, it is possible to analyze the following plot Figure 3 that displays the number of evaluations per period.

It should be mentioned that the visualization of this information consists of historical data of employees who could stay or be outside the organization. No knowledge or metrics of turnover have yet been taken into account.

According to the following plot (Figure 3), the number of performance assessments focuses more on periods between YER15 and MYR18, with the highest peak in MYR17.

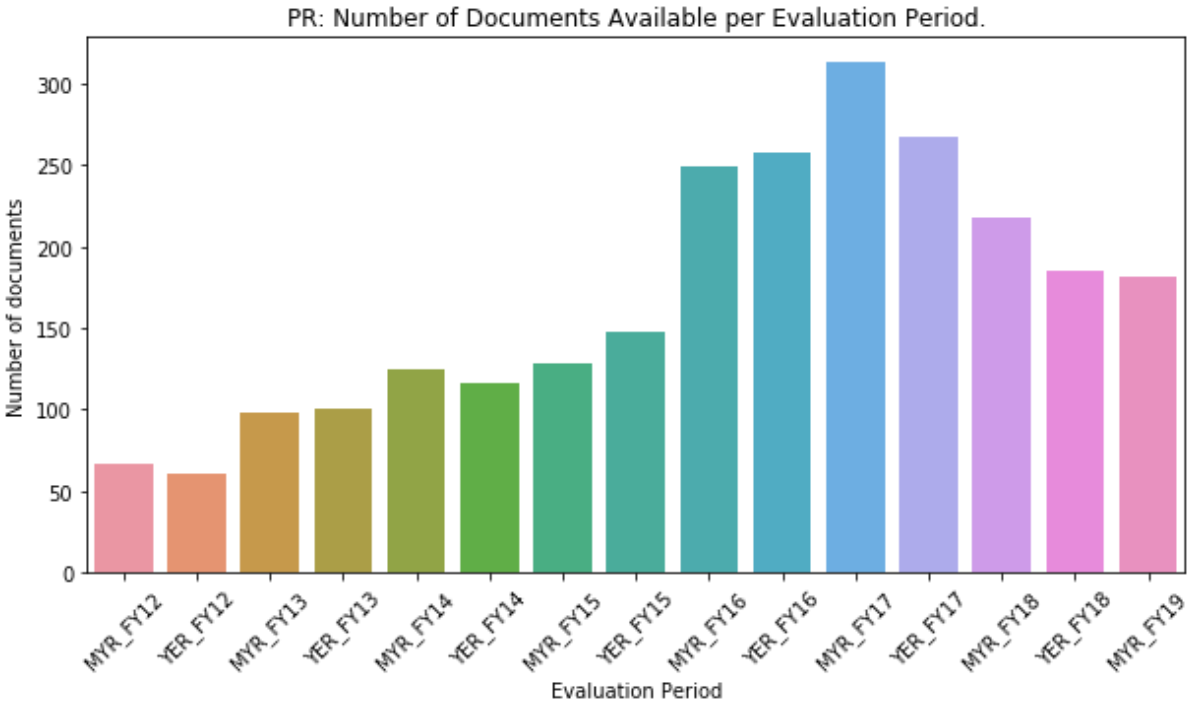

Figure 3 - Number of documents available per Evaluation Period

The relevance of these outcomes is that there is more occurrence between these periods, it means there is more information regarding the organization employees.

Maybe it makes sense to be particularly focused on this period of data because there may be a lack of features or even missing values on the dataset, in other years. Other studies were prepared, such as awareness of information gaps that may occur in the dataset. The reasons for this can be linked to employees who have taken unpaid leave or whether information may be missing.

Figure 4 displays a heat map correlation between periods of evaluation and employees with a gap of information performance evaluation.

As it is possible to observe in Figure 4, the black color represents the correlation of period with no-continuous evaluation documents and through other functions was observed that there exist around 40 employees with non-continuous performance evaluation.

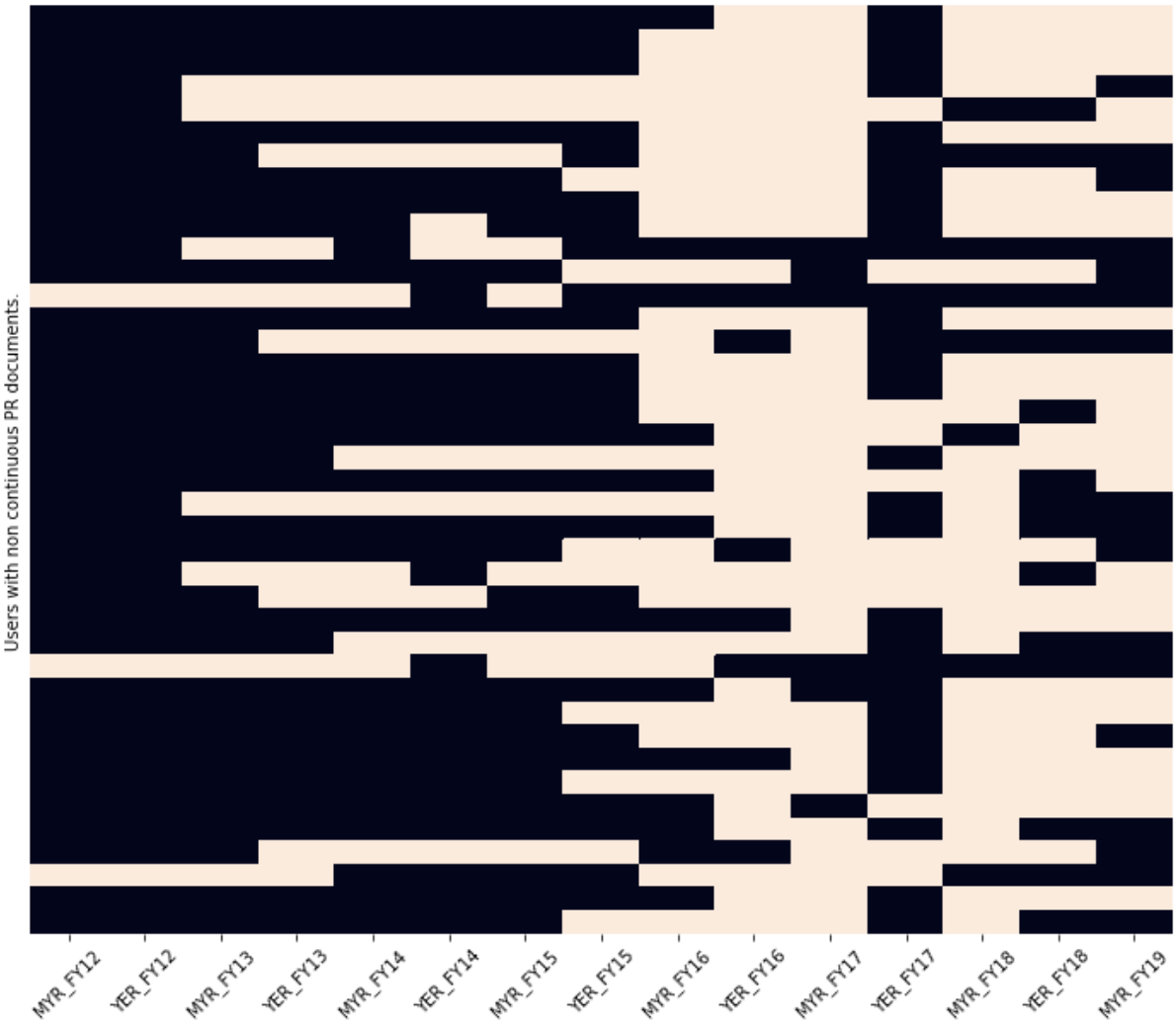

Figure 4 - Employees with non-continuous performance reviews

The last plot has been developed in Figure 5, as an exploratory analysis, where its attempt to verify between which periods of time there is a greater incidence of active employee assessment. Meaning, the y-axis corresponds to the interval since the first to the last evaluation period of employee evaluation.

As an initial inference, it is possible to draw that the focus is on these years since this is where the highest amount of employee information. Meaning, years after the indicated periods, cannot have been full information about all employees of the firm or even missing or poorly annotated information.

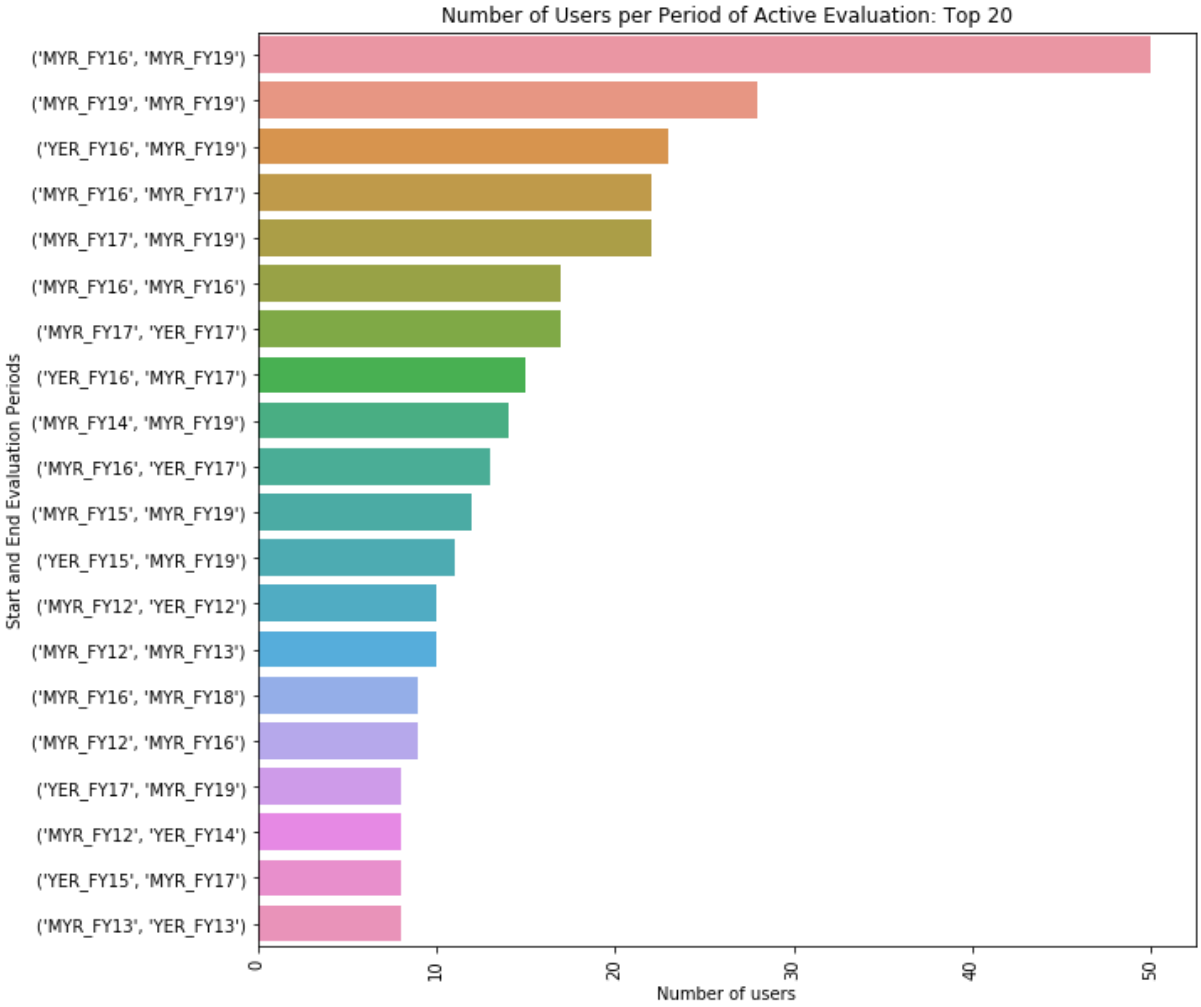

Figure 5 - Number of employees per Period of Active Evaluation

Taking into account these initial analyses it is possible to understand that the dataset needs to be processed so that the data are adapted as much as possible to the actual context of the problem.

It is also important to explain that, in the first instance, only a dataset with information from the evaluations among other attributes associated with employee detail was being considered.

With the development of the project, it was possible to acquire a new dataset. This new dataset reflects all employees who left the company. Additionally, it contains features such as hire date, exit date, whether to hire again, among others.

3.3. Data Preparation

Data pre-processing and cleaning are essential tasks that normally need to be performed before datasets can be used effectively for machine learning. Raw data is often noisy and inconsistent, with values that may be missing. It can yield misleading results by using such data for modeling. The typical issues with data quality that occur are:

- Incomplete: Data lacks attributes or values that are missing.
- Noisy: Data contains erroneous records or outliers.
- Inconsistent: Data contains conflicting records or inconsistencies.

3.3.1. Data Preparation – Framework

Following chapter described the steps that should be taken into account for data cleaning and which should be applied in real cases:

1) **Remove missing values** [27]: Since machine learning algorithms learn from data, the better the data, the better the results. Therefore, if your dataset has data instances that are incomplete or suspect may be corrupt, these instances need to be corrected. Possible options to deal with this problem is:
    - Delete the rows containing the missing values. This is particularly useful when having the luxury of hand-picking high-quality data, or having a large dataset, since it is the easiest solution.
2) **Replace missing values** [27]: Use a mathematical attribute such as means, mode and minimum/maximum to replace the missing values. Missing data can also be replaced by a value stating that the data is missing.
3) **Handling Categorical & Continuous variables** [28]: Categorical and continuous features cannot be entered directly into some algorithms and be meaningfully interpreted. In order to use the information in each of the features, it is possible to a categorical feature with k levels transformed into k-1 features each with two levels. In dummy coding, is allocated a value of 0 for each code variable to the reference group, a value of 1 is assigned to the interest group for comparison to the reference group for its defined code variable, while all other groups are assigned 0 for that particular code variable. One of the disadvantages of creating dummy code is that the data set

dimensionality will increase a lot (depending of the number of categorical features and the different values inside each of them).

4) **Outliers** [28]: If data set has variables, such as age, marital status, job position etc. Outliers are data points that are clearly separate from the rest of the data. Outliers, if present in their dataset, can cause problems by distorting their predictive model that may result in an unreliable prediction of the data. It is a safe idea to clip or eliminate the outliers in certain situations. As the dimensionality of the data increases more difficult it is to identify a record as an outlier.
5) **Normalization** [28]: Normalize samples to the unit norm individually. It is the method of getting all the information on the same scale; since their data has different scales, normalization is beneficial, and the algorithm is used does not allow assumptions about the distribution of their data, such as k-nearest neighbours and artificial neural networks.

3.3.2. Data Preparation applied to the project

As mentioned, the dataset was poorly adjusted to the project's needs, leading to the need for many visualizations, the development of functions that would enable the collection of more data, data as variables that would indicate whether the employee was promoted.

One need that existed was the joining of the two data sets, having as keys the employee ID and the year of entry into the company. Actually, a feature that converts the entry day into periods had to be created on top of the hire date, since the year YER / MYR is one of the main keys of the project. After all the conditions for exploitation have been created, missing values were replaced with numerical variables.

Finally, repeated views of the data were made, and incoherent information was identified. Then this information was removed so as not to create conflict in the data.

All the improvements that have been made to the data are presented below so that they can act as a guide and be applied in other real contexts.

Dataset has 511 entries for each column match the number of rows (554) of the dataset, there are no missing values. Dataset is composed by two variables as *Datetime* converted into object (Hire Date and "*Último dia de trabalho"*), twenty-five variables as objects and two variables as integer.

Below are the tasks that were performed in this phase for this research:

- **Cleaning data:**

All removals made to the dataset are listed below in detail:

- Users with dates misaligned between datasets: Initially, information such as the "date of the first evaluation" was misaligned with the entry date of the employee into the company. For instance, the original dataset had employees performance evaluation filled in. However, when joining with "Exit Interview" dataset, this dataset displayed that the same employee had already left the company.
- From the analysis of graphs shown previously, it is possible to verify information gaps between intervals/periods associated with an employee, so were removed from the dataset.
- Missing values from employees who had the exit report filled in with dates behind the years under study. Meaning, the original dataset contains employees that have a hire date or exit date before the first period of study.
- A rule was created to remove the employee who has information starting in 2007, 2008 or 2011 and leaving in MYR 12, 13, 14. If considering this data, the analysis would examine that employee leaves at the year an employee begins working. Or even that was in the company for a short period, when he had been in the organization for about four years, actually.
- Employees who leave the company in the period prior to the study.
- Employees who had information gaps, as mentioned above, are about 40, were removed.
- Employees who do not have a performance evaluation.
- The focus of this research is on analysts, consultants and senior consultants, considered "Staff", thus other grades above these were not considered for the investigation, removing them.

- **Data preparation:**

Since there was a shortage of variables for the research, some conditions were developed to make some analyses possible.

A pivot table was developed through the employee's assessments over the periods to explore the entire journey of the employee. With this information, it was possible to create a condition

that extracts the last period in which the employee had an evaluation. Enabled to explore other analyses, for instance, understanding in which year the user left the company. Although the dataset already had this indication on the exit interview. The same way that it is possible to create the condition for the exit year through the panda's libraries, it was possible to establish this condition for the first period in which the employee had the first evaluation.

Creation of gap counting, i.e. counting the number of periods when the employee has been outside the company. The conclusion at this point was that the maximum number of periods that the employee could be out was 4. The real context of the problem would correspond to 2 years. Since it was not possible to establish whether it was missing information; or whether it was an unpaid leave; or whether the employee had left and joined the company after two years; It ended up removing this information as it was inconclusive, and there were no facts to analyses it.

It was necessary to establish a condition to extract a promoted or not promoted variable considering the "Churn" employees and the analysis of the detailed employee grade variables in the last year of evaluation and the year they left.

Regarding professional categories, it was initially defined that the research would focus on grades from Analyst, Consultant and Senior Consultant. There are several professional careers, as a technical career; a career as a data science specialist; and a functional career; a function was developed to group all these professional categories.

The reasons for employee churn in the exit interview of the employee were explored. Through some searching regarding popular reasons for employee churn, it was trying to answer the question "what are the reasons that make sense to group" and developed the reasons aggregation takes into consideration the reasons searched.

- **Categorical & Continuous variables:**

One of the main difficulties in the project was to work with categorical variables rather than Continuous. Using the methods mentioned in the previous chapter, some variables were converted into classes and then, columns with dummies values were created. When correlations were explored, the goal is to use these variables.

- **Variables obtained:**

The final dataset, considered, was created based on the joining of two datasets. Where the first dataset contained information from the evaluation of employees over several years, and the second dataset only contains data on employees who left, with the respective exit information, such as reasons, days, etc.

With the join of this dataset, the present sample had to be considered and which of those employees had already left the company. Meaning, with this research, many attributes which were not relevant, i.e. were not giving any useful information, like evaluation detailed by objectives, first appraiser, second appraiser, place of work, etc. Hence these attributes and attributes detailed above were removed in the process of data cleaning. The below Table 1 shows all the final attributes of this dataset which were transformed after the three phases detailed above. Additionally, a column has been created where identifying if that attribute is crucial to perform the analysis. Meaning, whether the variables to applying this framework is essential.

Attribute	Meaning of each attribute	Key Attribute (Y/N)
User_key	Identification ID of the employee	Y
PerformanceEvaluation	Identification marks of employee	N
Year_period	Created based on the concatenation of 'ANO FISCAL' and 'PERIODO' variables and order by chronological time	Y Applicable to each context
NrPeriodosFora	Variable created to identify the number of periods that when it already leaves. These variable is created based on evaluate and period, created pivot table and a function that counters each position [user_key x year_period] of pivot table	
UltimoPeriodoAvaliacao	Based on the last variable NrPeriodosFora, this variable identifies the last period of evaluation employee	Y
NrPeriodosQueEntrou	Count of number of period that employee join to company	
PrimeiroPeriodoAvaliacao	Identification the period that employee join to company	Y
Categoria Actual	Professional Category that employee has in the most recent period	N

CATEGORIA AGREGADA	Aggregate professional category taking into account the detailed category	N
CATEGORIA DETALHADA	Professional Category detailed	Y
Hire Date	Hire Date of employee	Y
Último dia trabalho	Last day of employee work	Y
Voltaria a contratar	Indicates whether the company intends to hire the employee again	Y
JobCd Desc	Professional category when the employee leave	Y
Escritório	Office in which the churn employee was associated	N
Última avaliação	Last churn employee evaluation	N
Razões da Saída	Grouping of reasons of employee churn (filled in by the company)	Y
Razões do colaborador:	Employee reasons	Y
Iniciativa	Initiative (company/employee)	Y
Churn	Churn employee information, converted into dummies columns: 'Churn_0', 'Churn_1'	**Y**
Categorias	Grouping of professional categories, co nverted into classes and then into dumm ies values. ('_Analyst'; '_Consultant';'_Senior Cons ultant')	Y
PeriodoSaida:	Variable processed based on a function (Hire Date)	Y
PeriodoEntrada	Variable processed based on a function (Last work day)	Y
Promoção	Variable created based on a function and dummies columns ('Promovido_0'; 'Promovido_1')	N
GroupMotivoSaida	Grouping of employee reasons	N
MotivoClass	Grouping of employee reasons, convert ed into classes and then into dummies c olumns ('Motivo_0', 'Motivo_1', 'Motiv o_2', 'Motivo_4', 'Motivo_5', 'Motivo_6' 'Motivo_7', 'Motivo_8')	N
IniciativaClass	Class created based on ‘Iniciativa’	N
VoltariaAContratarClass	Class and dummies values, created based on ‘Voltaria’ variable 'Voltaria_1' - Yes 'Voltaria_2' - No	N
RazaoColaboradorClass	Class created, note that only select one reason of employee 'RazaoColaborador_2' – Work like-balance	N

Table 1 – Final Attributes used on the research

It is possible to conclude that are no missing values in the current dataset. Since this dataset had a variety of categorical variables, such as the Yes and No values of the Churn variable, and all the rating attributes had a Likert scale ranging from 1 to 5, all of these attributes were type-cast to variables. After the tasks performed, the dataset was ready to analyse further.

Chapter 4 – Data Insights - Visualization

In this chapter is displayed, all exploratory analyses performed for a better understanding of the data. All of these analyses are performed based on the use of advanced libraries such as pandas and *NumPy*.

As mentioned above, many transformations were necessary to make the data to have a real and clear interpretation. All analyses that explored consider the variable "year period". One period duration equals to 6 months. That means, in one year there are two evaluation periods, one referring to the middle of the year and the other at the end of the year.

The periods considered in the dataset begin with a period of MYR12 and end with a period of MYR19. It is possible to deduce in a previously (Figure 5) demonstrated analysis that the period time in which had the highest proportion of employee data would be between MYR16-MYR19. Notice that these intervals can differ when adding the Churn employees, by joining the answers of the exit interview. It is this sort of research that was going to understand and evaluate. Typically, the new employees join the company in periods as MRY.

In the business context, the professional category of the analyst, who are beginning their careers, typically begins to work in a company in MYR periods.

The specialist groups that considered for this research are those known as the company's "Staff". Analysts, consultants and senior consultants are among these.

4.1. Churn/No Churn Employees

This subchapter aims to visualize analyses of the churn and No-churn variables compared to other types of variables, from performance evaluation periods, to exit periods and professional categories. The main target is to visualize within each variable, where there is a higher Churn rate.

When joining the churn employees to those who are no-churn, it is possible observed that in total are more churn employees than churn (Figure 6).

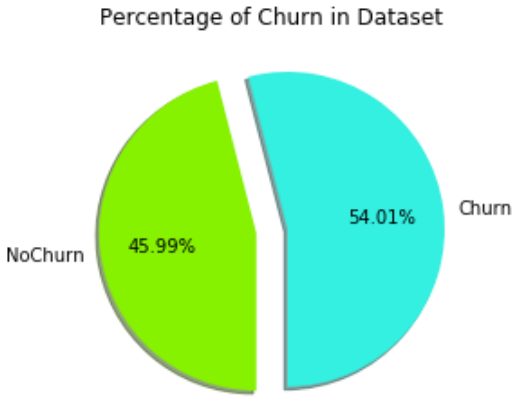

Figure 6 - Percentage of Churn & No-Churn Employee

Whereas that it would not be possible to say that more employees are leaving than those who are in the company. Because, in a total of 511 employees (Table 2), have about 276 who are churn and 235 no-churn.

Churn	No Churn	Total	Churn Rate
276	235	511	54.01%

Table 2 - Total of employees Churn & No-Churn

Churn employees per period of the evaluation were the best approach decided to understand and visualize the data to resolve the problem in the study.

Through analyses developed from counting the number of employees churn per period, it was seen that there exist more Churn employees in periods like YER16 with 47 employees churn, YER17 and MYR17 with 40 employees churn.

During the development of the project, and through the visualization of plots, it was possible to draw some conclusions. Such as, what sort of variable would make more sense to use to achieve the intended purposes. The major difficulty was to understand which variable "*PeriodoEntrada*" (Entry Period) "*PrimeiroPeriodoAvaliacao*" (First Period of evaluation) should be used. The purpose of the analysis is to understand the number of Churns per period evaluation. The following variables "*PeriodoEntrada*" (Entry Period) and "*PeriodoSaida*" (Exit Period) were created based on the exit interview report, meaning this variable only contains information related to churn employee.

Considering the previous assumption, to research variables Churn/No-Churn, considered the variables First period that employee is evaluated and the last period that has an evaluation were considered.

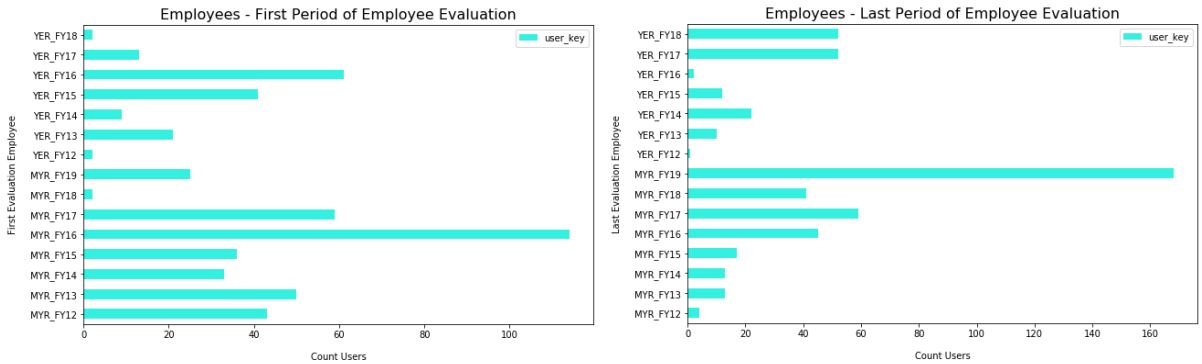

Figure 7 - Comparing Employees with First and Last Evaluation

As displayed in Figure 7, it possible to compare the number of employees that entering and leaving the company. As observed by the two plots, it has possible to say that there are more employees with first evaluation than last in the following periods: MYR12, MYR13, YER13, MYR14, MYR15, YER15, MYR16, YER16 and MYR17.

Unlike, it has more employees with the last evaluation period in the following periods: YER17, MYR18 and YER18.

It does not make sense to consider the last evaluation employees, since are the latest year that dataset has records, and it can be churn or no-churn employees.

The conclusions that can draw is that have more employees with a first assessment than with the last evaluation, which drives to the suspiciousness that the churn rate could decrease by period since that have more employees with the first performance evaluation per period than with the last performance evaluation.

The following plot (Figure 8) intends to analyse churn/no-churn for the first evaluation period.

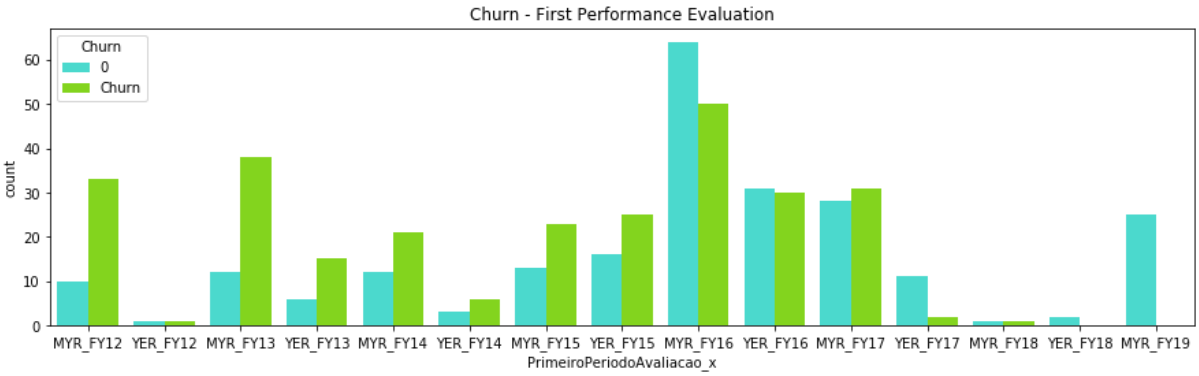

Figure 8 - Churn/NoChurn vs First Performance Evaluation

As it is possible to observe the Figure 8, have around 38 employees who joined the year MYR13 and are churn employee. From the analysis of the graph, it seems that employee entering the year MYR16 is still in the company, also in the year YER16 but with a slightly lower number. Finally, employees who join the YER17 tend to stay in the organization, have approximately eleven No-churn and two Churn employees.

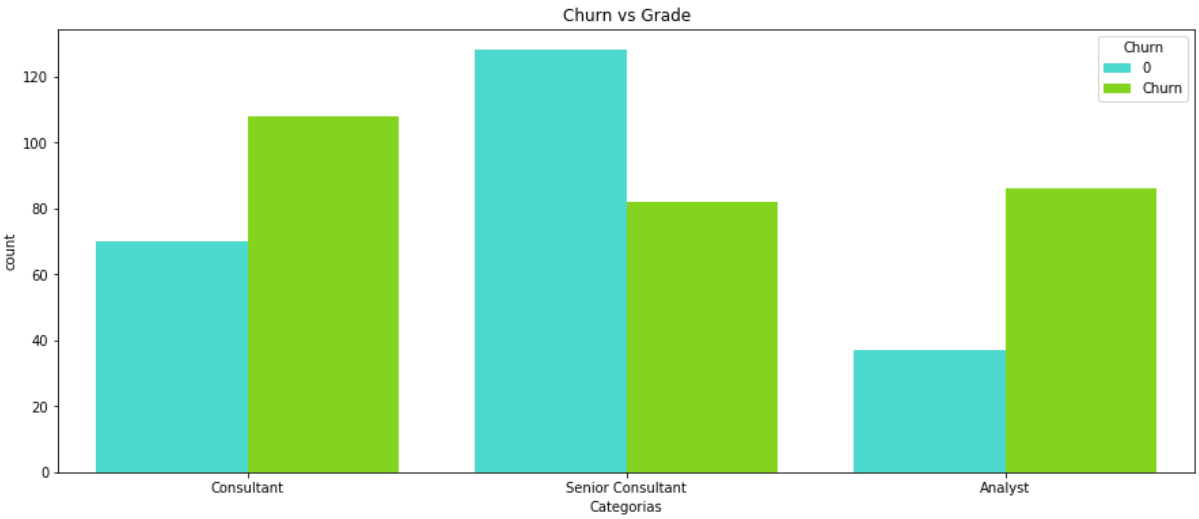

Figure 9 - Churn/No-Churn vs Professional Category

As observed in Figure 9, exists more employees to doing churn in categories professional like Analyst and Consultant. The interpretation that can draw from this observation is that analysts probably do not identify with the company and leaving the company, or they leave at the company's option. These conclusions are verified later, through more specific analysis capable of answering this type of question.

4.2. Churn Employee

The current section presents some analyses developed specifically to conclude the employees who are churn what kind of characteristics they may have.

The following Figure 10 displays the percentage of churn/No-Churn Employees and Percentage of employees Churn per professional Categories (Analyst, Consultant and Senior Consultant).

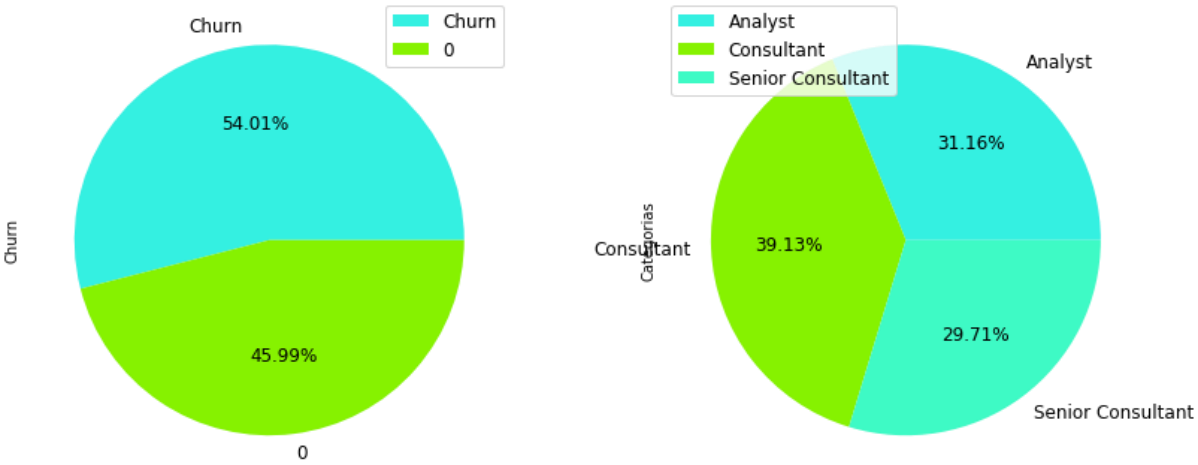

Figure 10 - Churn Professional Categories

Total of the 276 Churn employees, about 108 are consultants, 86 analysts and 82 senior consultants. Meaning, exist a higher percentage of consultants leaving the company. However, it was not necessarily mean that have more consultants leaving the company than analysts, have a smaller sample of analysts versus consultants. In the dataset, have about 178 consultants of which 108 are churn. Nevertheless, have a total of 123 analysts of whom about 86 are churn.

The following Figure 11, displays three graphics that present the employees churn per professional category compared with three features such as, Period of employee leaves, Period of the employee starting in the company and, the first period of employee performance evaluation.

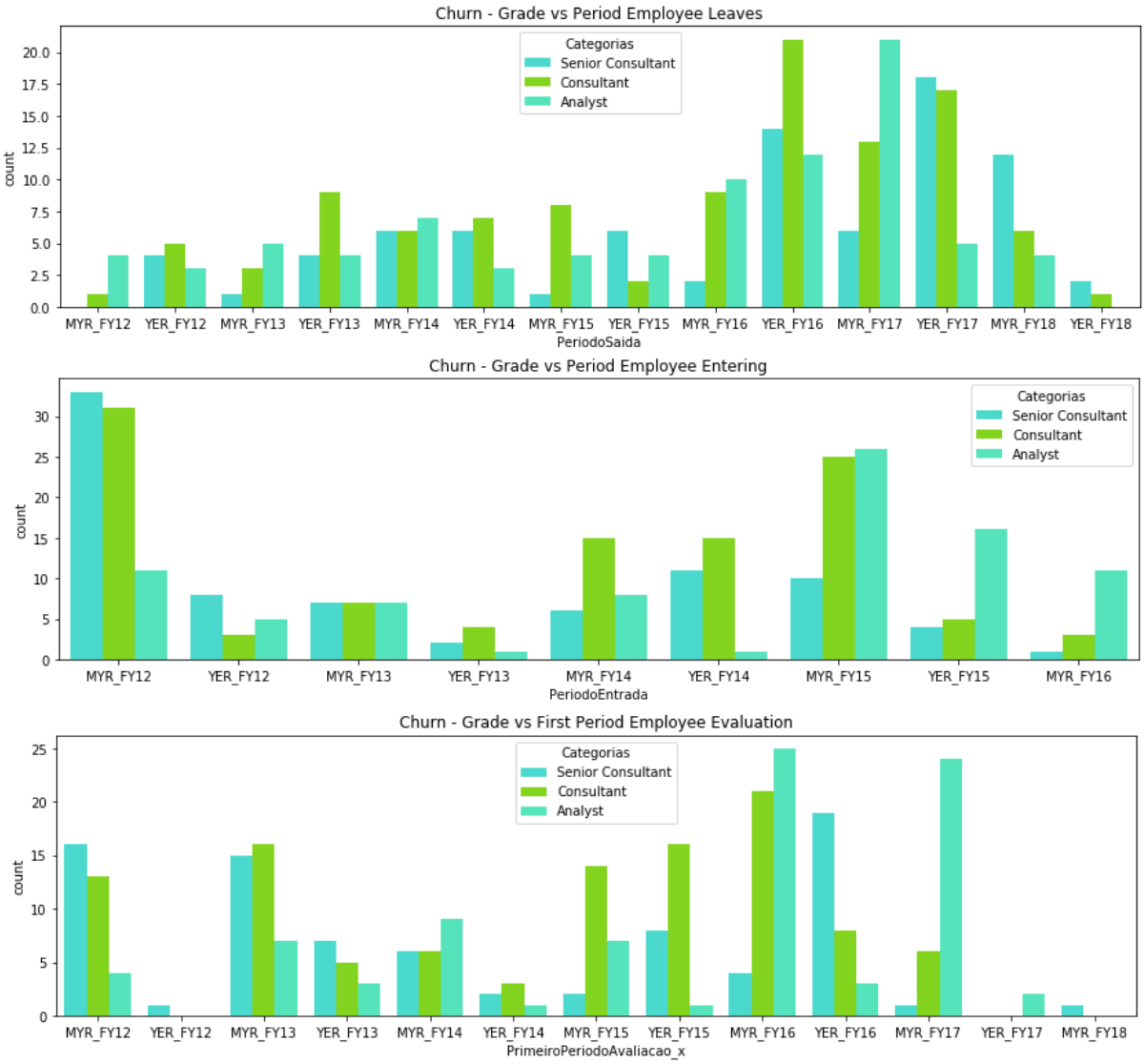

Figure 11 – Professional Categories Churn Compared to Exit Period, Entry Period, First Period of Employee Evaluation

By analysing the first plot, it was possible to observe that where the professional-grade of analyst had the highest churn rate was in the YER16, YER17 and MYR18 periods. In the Consultant category, have the highest churn rate in the periods YER16, MYR17 and YER17. Finally, the Senior Consultants were more in the YER16 and MYR17 periods.

The highest occurrence of data is concentrated in the 16-19 periods, as seen in previous studies, which explains why this highest incidence is on these data.

From the second plot analysis, it can observe that several employees who joined the MYR12 period are leaving. Additionally, in the MYR15 period, existing many analysts and consultants

who entered that year and left. By comparing the first graph, it can speculate that these analysts leave the company on MYR16, YER16 or MYR17. Perhaps means that analysts do not remain in the organization for more than six months, end the contract and leave.

Looking across the three graphs, the MYR17 period, it can observe that many analysts are having their first assessment and leaving immediately, i.e. employee leaves after six months. The same happens in the year MYR16, existing many analysts with first evaluation in the year MYR16, and employee leaves after six months (YER16). From the grade of analysts, it is possible to conclude that employee left after the first evaluation.

As concerns other professional categories, cannot conclude much, the senior consultant appears to stay longer in the company. While consultants, they end up leaving. Hence, it can investigate at which moment of consultant and senior consultant these grades leaving the company the most, further.

With the analysis performed, the next question is "Did the analysts leave at the intention of the company?", "Would the company hire again?". The following visualization of data intends to answer these questions.

The following Figure 12, displays the professional category compared with initiative to leaving the company (initiated by Employee/Company).

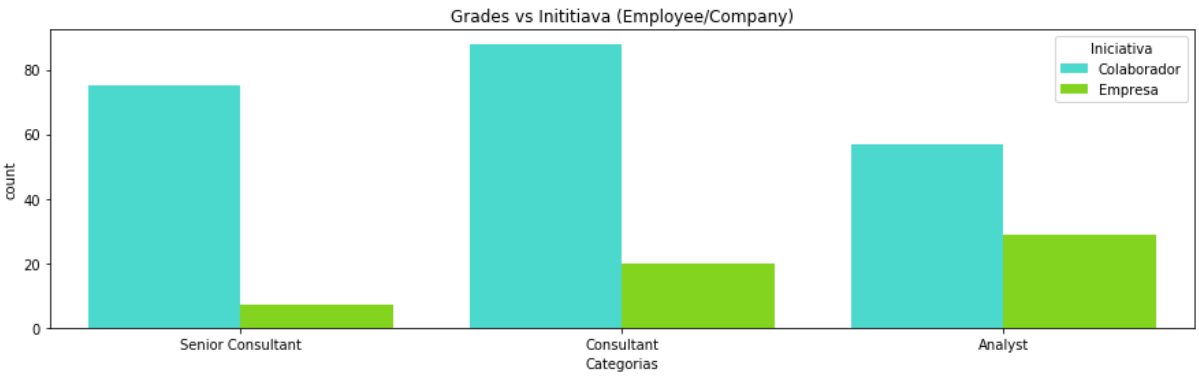

Figure 12 - Churn - Professional Categories compared to Initiative (Employee/Company)

As it is possible to observe in Figure 12, the initiative to leave the company is higher by an initiative of the employee rather than employee motivation. It should note that this initiative is less apparent in the analyst grade. Observe at the second issue of whether the company would hire again.

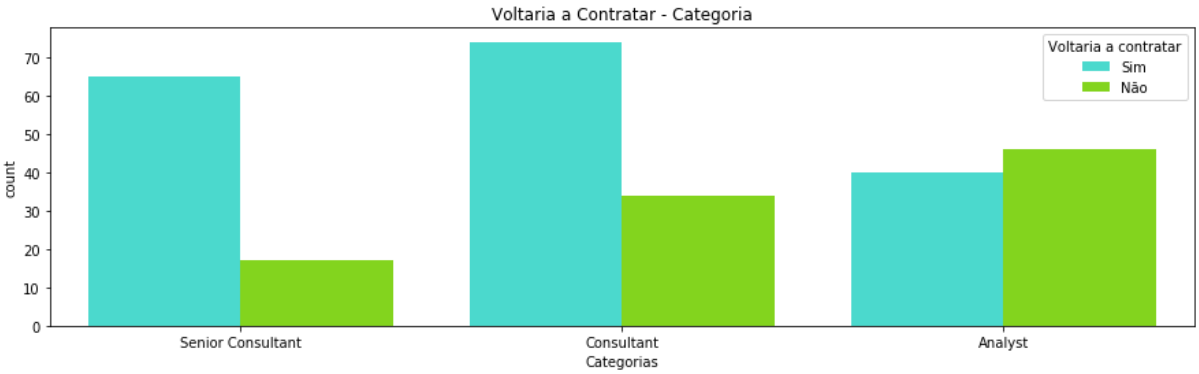

Figure 13 - Churn - Professional Categories compared to Company would hire again

As it is possible to see in Figure 13, for the analyst category, although the percentage of the variable per employee's will is higher than that of the company, the percentage of the variable "would hire the employee back" for the analyst category is higher in the answer "No".

This can be an indicator that analysts leave of their own free intention, but the company also does not consider them potential targets to be retained in the company.

The Figure 14 present the counting of churn employees between each professional category.

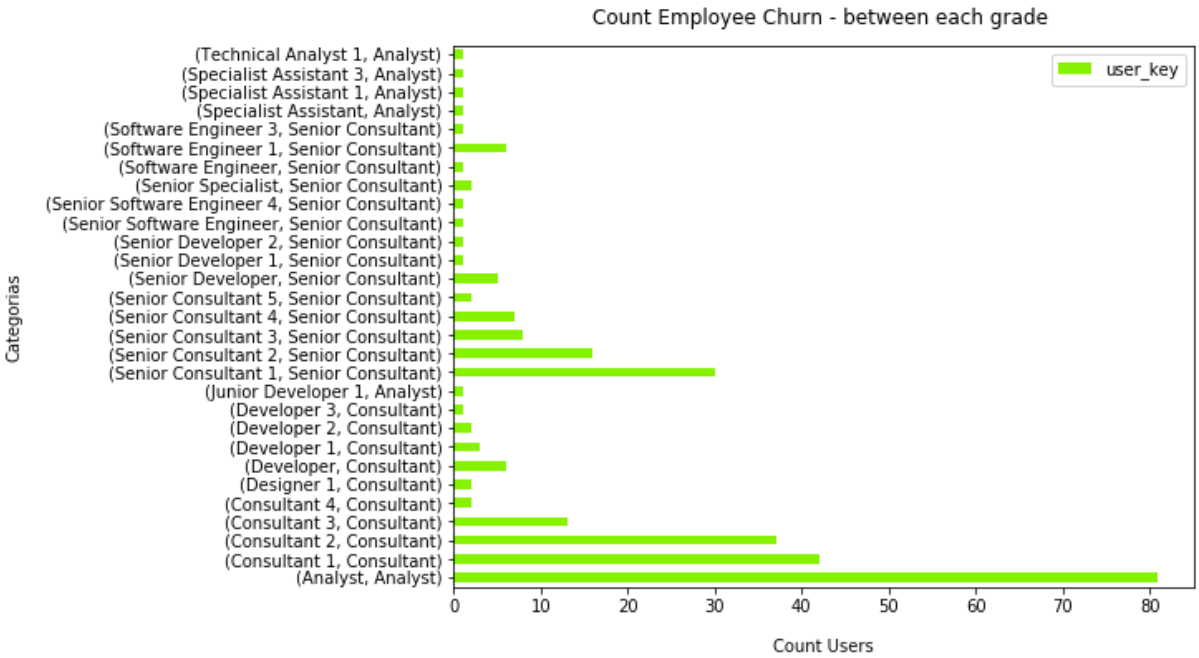

Figure 14 - Churn - Professional Category detailed

With the analysis chart displayed in Figure 14, intends to visualize at which stage of the career the employee leaves.

For instance, in Senior Consultant the normal is four years in this grade, the objective is to see if he leaves more in the first year, second, third or fourth.

As it possible to observe in Figure 14, the consultant's grade leaves more in the first year and in the second year. Concerning when promoted to senior consultant, they do not leave the company as much. However, in the senior consultant professional category, the year in which it is most common to leave the company is the first and second years.

Regarding professional categories like "Consultant 3", "Consultant 4" or "Senior Consultant 5" is normal that these employees leave because it means that these employees are not promoted. However, it does not have many employees in which cases.

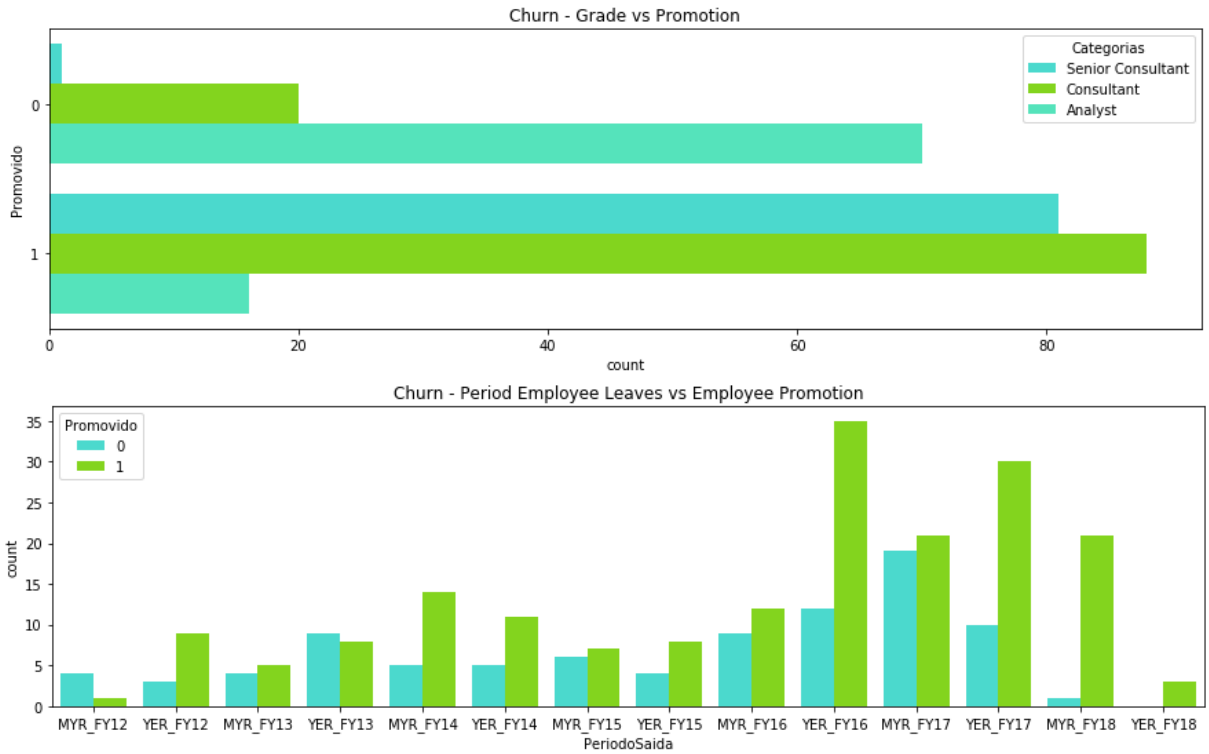

Figure 15 - Churn – Promotion compared to Professional Category and Exit Period

The plots in Figure 15, displays the promotion variables by professional categories and the exit period of churn employees.

Within each professional category, exist more analyst leaving, because employee no-promoted. In truth, it already observed that have analysts that left because they complete the

contract or by choice. Additionally, are some Consultants who leave the company, because they are not promoted, but it does not appear to be a variable that influences much the exit.

When analysing the first plot, it can see that the fact that the employee does not get promoted does not influence a lot the outcome.

GroupMotivoSaida	user_key
0	53
Conduta profissional inadequada	1
Extinção do posto de trabalho/Fim de contrato/Estágio	3
Insatisfação: Progressão de carreira; Tipo de função; Projectos ; Carga de trabalho; Incompatibilidade com equipa	38
Insuficiências de competências técnicas para progredir na carreira	51
Motivos pessoais e/ou familiares	69
Oferta de uma função/carreira mais atractiva; Oportunidade internacional	24
Remuneração superior	37

Figure 16 - Churn - Reasons for Employees Leaves

Figure 16 shows the count of the reasons for employees leaving, filled in by the company. In the plot, it is possible to see that, the zero answers representing the counting answers that are not filled in. Also, it can observe that the reason that is most common for employee leaving are: “Personal and/or family reasons"; “Insufficient technological abilities to advance in a career”; “Dissatisfaction: Career progress; Function/Job Type; Projects; Workload; Conflict Team”; and the last one “salary higher”.

Finally, Figure 17 displays each reason for employee leaves for each professional categories.

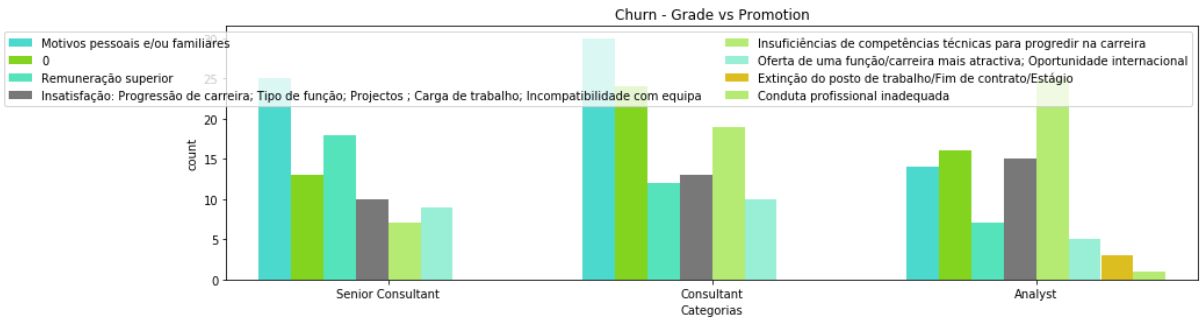

Figure 17 - Churn - Reasons for Employee Leaves compared to Professional Categories

It is possible to see that the reason for personal reason is presenting in all category. Particularly, in senior consultant have the "salary higher" reason a little higher than the others.

It is a reasonable assumption that one could draw from this analysis "the employees when they reach the senior consultant category, are more tending to leave the company because of their low salary or because they have better proposals".

Another important observation is that have the reason "Insufficient technical skills to progress in the career" with a very high percentage for analysts. Meaning that the conclusions suspecting may be close to confirmation, analysts leave the company because they do not have the skills the company is looking for.

In the career as a consultant, this reason is also slightly high, and just below have the dissatisfaction. It is possible to suppose that in terms of exit reasons, the reasons are aligned between analysts and consultants. However, there are more evident in analyst grade.

Generally, it can conclude that the most common reason in the three professional categories is the dissatisfaction of the employee.

At this moment, examined the most critical variables in the dataset. Based on this knowledge, in the next chapter, the goal is to try to answer some questions. Gives more focus to assumptions that were drawn from analyses and then, performed correlations of misconceptions/assumptions with other features. To discover the real reason for an employee leaving.

4.3. Visualization approach

Initially, this type of analysis should be carried out, with the help of the available Python libraries, such as the pandas, *seaborn*, among others, for the creation of graphs such as *countplot*.

The intention is to understand and make analysis between variables to understand the data and, what set of assumptions that can extract as output reason. For instance, have data such as promotion, exiting initiative and professional category, it is possible to assumptions extract.

As is the case above, where suspect that analysts leave because they have no skills. In the following chapter, can be an assumption to be analysed and, must try to confirm through correlations.

Chapter 5 – Misconceptions & Assumptions – Discussion

The purpose of this chapter can be divided into two main objectives. First, based on misconceptions taken from paper [1], intended to correlate the variables that were selected as the most important to obtain answers to the following reason leaving of employee:

1. “People quit because of pay”
2. “People quit because they are dissatisfied with their jobs”

Additionally, through the visualization performed in the previous chapter, it was possible to draw more assumptions that can also be true in the real project of any organization. In other words, the goal is to try to create an approach that draws near proves the misconceptions in a real project context. Meaning, that approach must validate if these misconceptions are faithful and also, applicable to any organization.

The following Table 3 displays the variables that used to create a correlation of heat map:

Attribute	Attribute Description
Churn_0	No-Churn
Churn_1	Churn
Promovido_0	No-Promoted
Promovido_1	Promoted
Motivo_0	No answer
Motivo_1	Professional Conduct
Motivo_2	Job termination/End of contract/Internship
Motivo_3	Dissatisfaction: Career progress; Function/Job Type; Projects; Workload; Conflict Team
Motivo_4	Insufficient technological abilities to advance in a career
Motivo_5	Personal and/or family reasons
Motivo_6	Offering a more attractive function/career; International Opportunity
Motivo_7	salary higher
_Anlyst	Analyst Professional Category
_Consultant	Consultant Professional Category
_SeniorConsultant	Senior Consultant Professional Category
Iniciativa_1	Company initiative “Involuntary turnover” from [1]

Iniciativa_2	Employee initiative "Voluntary turnover" from [1]
Voltaria_1	Company would hire again: Yes
Voltaria_2	Company would hire again: No
RazaoColaborador_1	Employee Reason: work/life balance
RazaoColaborador_2	Employee Reason: professional growth and development

Table 3 - Features selected to correlation

Finally, all company can apply the correlation that demonstrated in the following sub-chapter. Also, to prove the questions "People quit because of pay", it needs to have features as, reasons pay/salary employee information or, salaries indicators that can transform for this purpose. Additionally, to find the answer to the question "People quit because they are dissatisfied with their jobs", it needs to have attributes such as work/life balance score; pay/salary employee information; distance for home; performance evaluation; or employee satisfaction from [25].

5.1. Correlation variables

The correlation used to observe the relationship between the employee churn and other features like employee satisfaction and the attributes previous introduced. Correlation is a very useful technique to discover the relationships between the predictor and the forecasting variable, as well as between the different predictor variables in a dataset.

Find Correlation between Features and choose the features which are highly correlated and remove redundant features.

Figure 18 is a heat map that looks at the correlations among all variables in our dataset. There are Machine Learning algorithms that do not handle well highly correlated variables so when making clusters, this has to be taken into account.

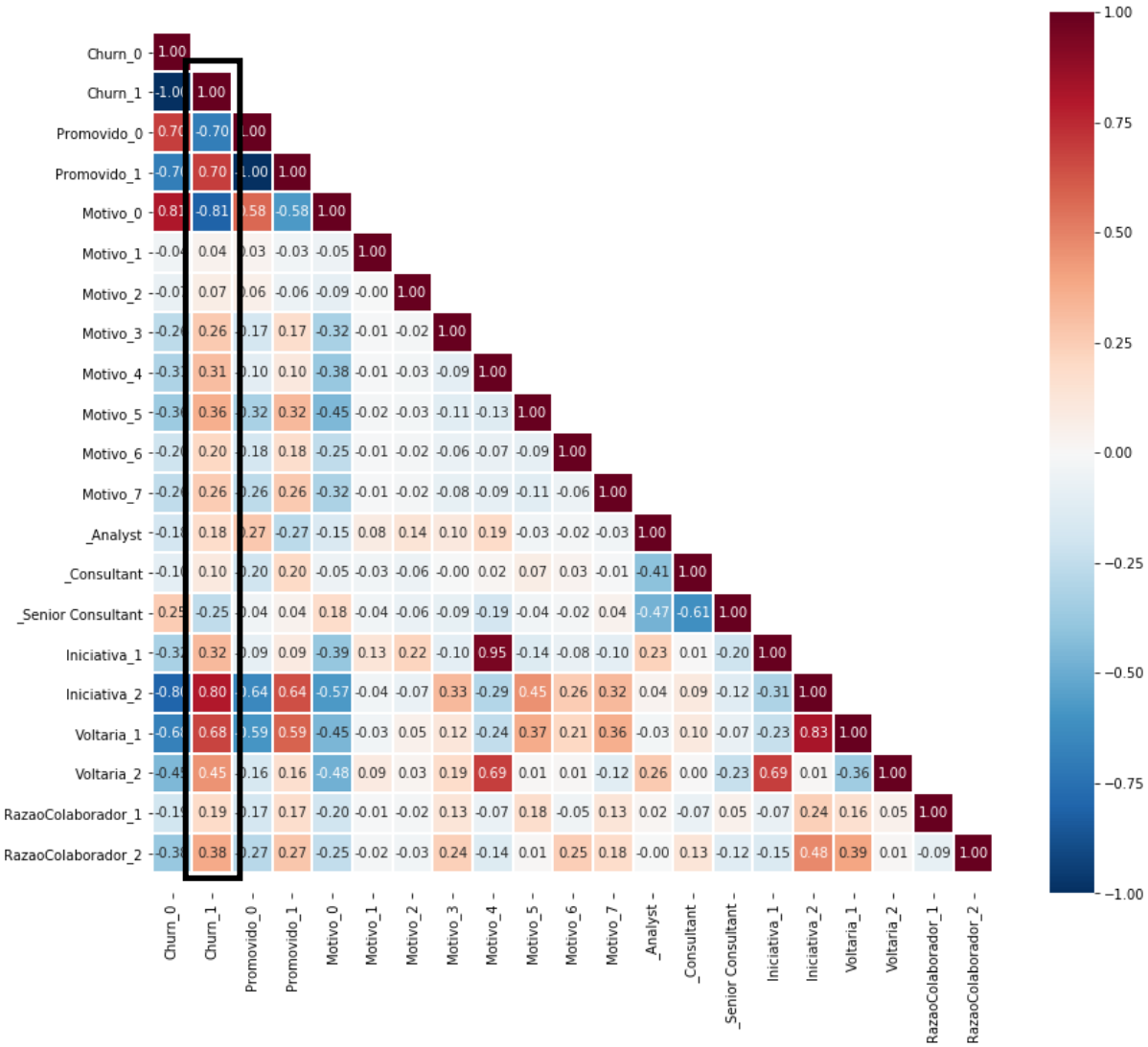

Figure 18 - Correlation Matrix

In this case, it is possible to observe that there are variables that have a correlation of 1 that are on the matrix diagonal. This happens because one of the variables was calculated by the other, so it expected that they highly depend on each other.

The opposite happens when observing, for instance, "Churn_0" and "*RazaoColaborador_2*" (correlation = -1) because as the "Churn_0" increases, "*RazaoColaborador_2*" decreases. Therefore, it must only consider the correlation between positive features.

About the relation between the variables should always consider "churn_1" with other variables since that is more relevant for research, to understand the employee churn with other variables.

- **Promoted:** It has a strong correlation of 0.7, between the churn variables and is promoted. However, it makes no sense to consider why saying "officials resign because they are promoted" is not right. Additionally, it was excluded promoted variable the research, for the reason that does not enough information to explore the obvious assumptions that are “People quit because they are no promoted”.
- **Motivo_1 and Motivo_2**: These variables are clear that have a lower correlation with churn_1 variable. It means that “Professional Conduct” and “Job termination/End of contract/Internship” reasons do not have a relation with employee churn.
- **Motivo_3 and Motivo_7**: Corresponding respective to **“Dissatisfaction: Career progress; Function/Job Type; Projects; Workload; Conflict Team”** and **“salary higher”** reasons, having both the correlations’ at 0.26 with the variable Churn_1. Therefore, it makes sense to select each of these variables to be studied in detail. "Salary higher" is a reason that exists in the dataset and which states that the employees left because they had a higher remuneration offer. Hence, this variable purposes attending in trying to explain the misconception **“People quit because of pay”** [1]. Also, **"Dissatisfaction: Career progress; Function/Job Type; Projects; Workload; Conflict Team"** is another reason in the dataset, and intends to reflect the dissatisfaction of the employee, spreading out for various reasons such as, working with people who do not like, working on projects that are not satisfied, workload, among others. This variable intends to help to demystify the issue of paper "**People quit because they are dissatisfied with their jobs”**.
- **Motivo_4:** “**Insufficient technological abilities to advance in a career**” there is a high correlation between the reasons proposes to research, with a correlation of 0.31. Through this reason, it is possible to draw from one assumption as the reason for the employee's churn the company: **“People quite because they do not have enough skills in line with company standards”**.
- **Motivo_5**: Correspond to **“Personal and/or family reasons”** reason, and it is clear that it the highest correlation between employee churn reasons. Having a correlation about 0.36 with variable of employee churn. However, given the context of the company, that cannot create any assumption around this, because is a redundant reason. Since that, personal reasons can be many and with a different purpose, it can be because of dissatisfaction; burnout; or because they do not want to expose the real reason. For this reason, does not create any assumption around this variable.

- **Motivo_6: "Offering a more attractive function/career; International Opportunity"** has a correlation between churn around 0.20. This could be considered a reason for dissatisfaction since they are not liking their current positions and are looking for better or want an international experience that the company can even provide but did not do so for any needs of this employee in other projects. Also, it is possible to extract the assumption: **"People quit because they have more attractive job offers"**.
- **Professional Category:** Among the three professional categories, there is no question that the analyst grade is the one most correlated with churn. Meaning, this variable is interesting to study within each misconception/assumption.
- **Initiative (Employee/Company):** "*Iniciativa_2*" is the highest correlation in our data, having a correlation at 0.8 with "churn_1" variable. Meaning, are observing a voluntary turnover. It makes sense that, extract the assumption **"Voluntary turnover is the most common reason for employee churn"**.
- **Voltaria_1 and Voltaria_2:** "Company would hire again: Yes" have a correlation between churn around 0.68. Meaning, most of the company's employees would be hired again. This variable becomes interesting to be compared to professional categories.
- **RazaoColaborador_1:** With this variable, the purpose is to investigate if the employees leave because they do not have a work/life balance. Unfortunately, the variables are not strong correlations that justify the statement of this assumption.
- **RazaoColaborador_2: "Professional growth and development"** is directly linked to reason 6. These variables have a correlation of 0.38. So, also reflects in the analysis of the assumption mentioned above **"People give up because they have more attractive job offers".**

In the following section, developed more correlations around the assumptions created with the first correlation analysis. The aim is to interpret these assumptions between other variables, assuming that was only correlating the variables based on employees churn variable.

5.2. Misconceptions

Based on the previous correlations, already observed that there is strong evidence that the two misconceptions selected from the paper [1], apply to the business context. The premises

selected were "People quit because of pay" and "People quit because they are dissatisfied with their jobs".

The misconception "There is little managers can do to directly influence turnover decisions" could not be proved because the data was not sufficient to answer these questions. However, if other companies have data like the project manager ID, it is possible to investigate this misconception.

5.2.1. People quit because of pay

Through the variable "salary higher" (Motivo_7) it is possible to verify the validity of the misconception "People quit because of pay" in the business context.

The next heat map (Figure 19) contains correlations considering only "churn_1" feature. The aim is to try to understand "*Motivo_7*" variable with other variables and try to analyses assumptions such as "the employees when they reach the senior consultant category, are more tending to leave the company because of their low salary or because they have better proposals".

It is possible to discern that this reason has a negative correlation between the professional categories of analyst and consultant. Therefore, in the context of the problem, is the most common reason for a senior consultant grade. As observed before, it is not the reason that has the highest correlation, but there are people leaving for this reason. Consequently, it can infer that this assumption is valid in the business context, given the dataset.

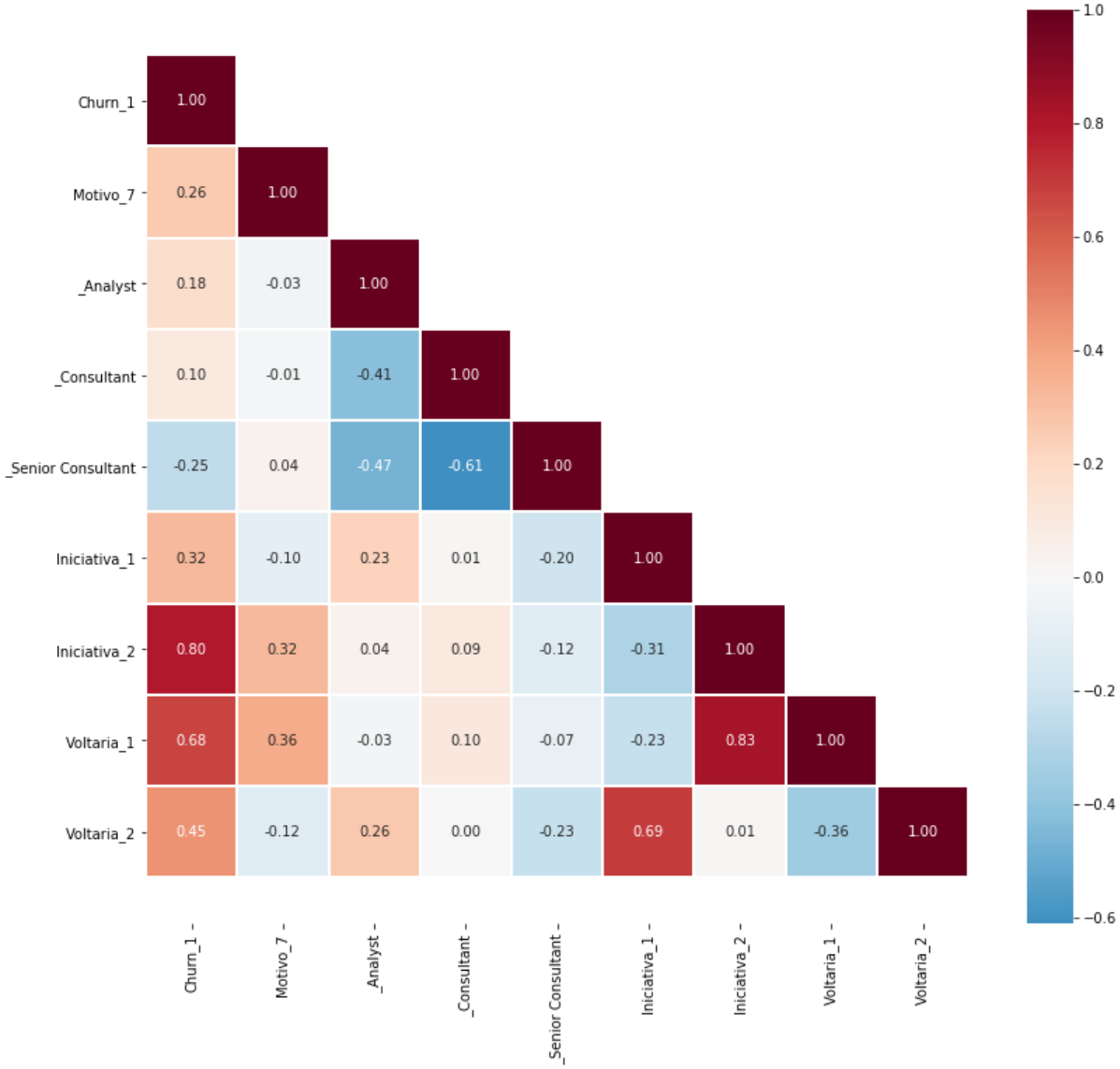

Figure 19 - Correlation Matrix - People quit because of pay

5.2.2. People quit because they are dissatisfied with their jobs

The variable aimed at studying the misconception set out in the title of the subchapter has undergone some transformation. The dissatisfaction of an employee can be defined by several factors. The factors considered are the following: Working conditions (includes workload and pressure); Incompatibility with some elements of the team; Career progression, type of function; Projects. Then groups in a single reason established as "Dissatisfaction: Career progress; Function/Job Type; Projects; Workload; Conflict Team".

The following Figure 20 displays the correlation of this reason with other dataset variables.

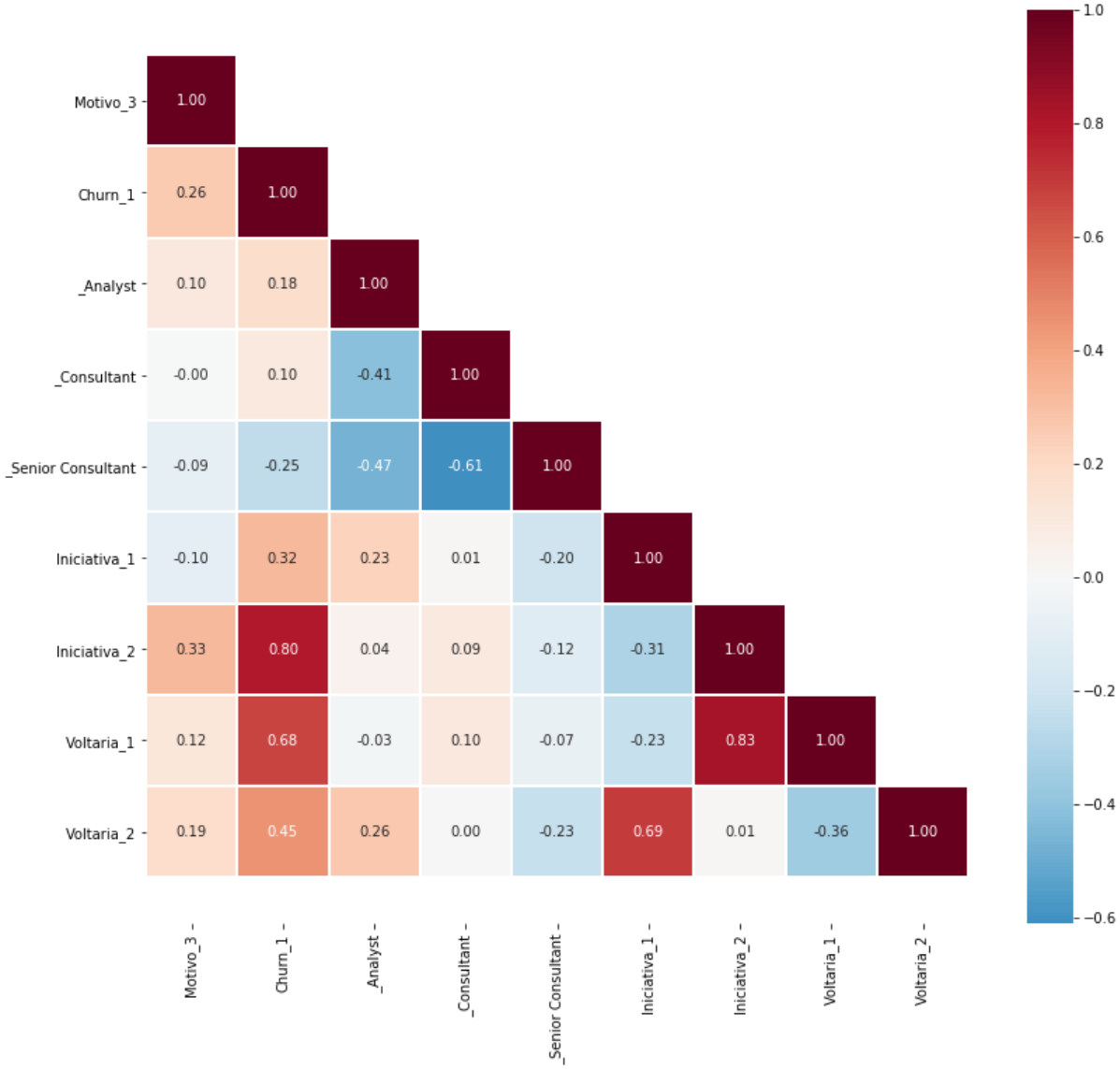

Figure 20 - Correlation Matrix - People quit because they are dissatisfied with their jobs

According to Figure 20, this reason is more common in the career of analyst since it has a positive correlation. It is a reason ("*Motivo_3*") of voluntary turnover (Iniciativa_2). Curiously, existing a higher correlation between "*Voltaria_2*" and "*Motivo_3*" than "*Voltaria_1*".

It can assume that people who are dissatisfied leave of their own free intention, but the company has no desire to hire again.

From the analysis of Figure 13 (4.2 Churn Employee), already verified that probably the analysts leave and the company would not hire them again. In other words, putting all this information together, this correlation may be higher due to the category of the analyst.

5.3. Assumptions

The present chapter reflects the second objective proposed, which was to find new assumptions as reasons for churn of an employee.

Previously, were revealed how arrived at these premises, in the following chapter the assumptions “People quit because they search for career progress” and “People quit because they do not have enough skills in line with company standards” are analyzed.

5.3.1. People quit because they search for career progress

The variables to support the creation of the assumption “People quit because they search for career progress” are the following "Offering a more attractive function/career; International Opportunity" (“*Motivo_6*”) and "professional growth and development" (“*RazaoColaborador_2*”).

The following Figure 21 intends to cross “*Motivo_6*” and “*RazaoColabrador_2*” with other variables in order to prove the assumption that people quit because they want to grow in their career.

As it can observe in the above plot, is possible draw some conclusions such as, both reasons “Offering a more attractive function/career; International Opportunity" and "professional growth and development” are strongly correlated. The consultant grade is the one that seeks further career progression when it shows the intention to leave the company. This assumption can be drawn from the analysis of the correlation between the variable “*Consultant_*” and “*RazaoColaborador _2*” and also, the interpretation of the “*Iniciativa_2*” with “*RazaoColaborador _2*”. Finally, should note that although the employees feel they need to progress, the company would hire again.

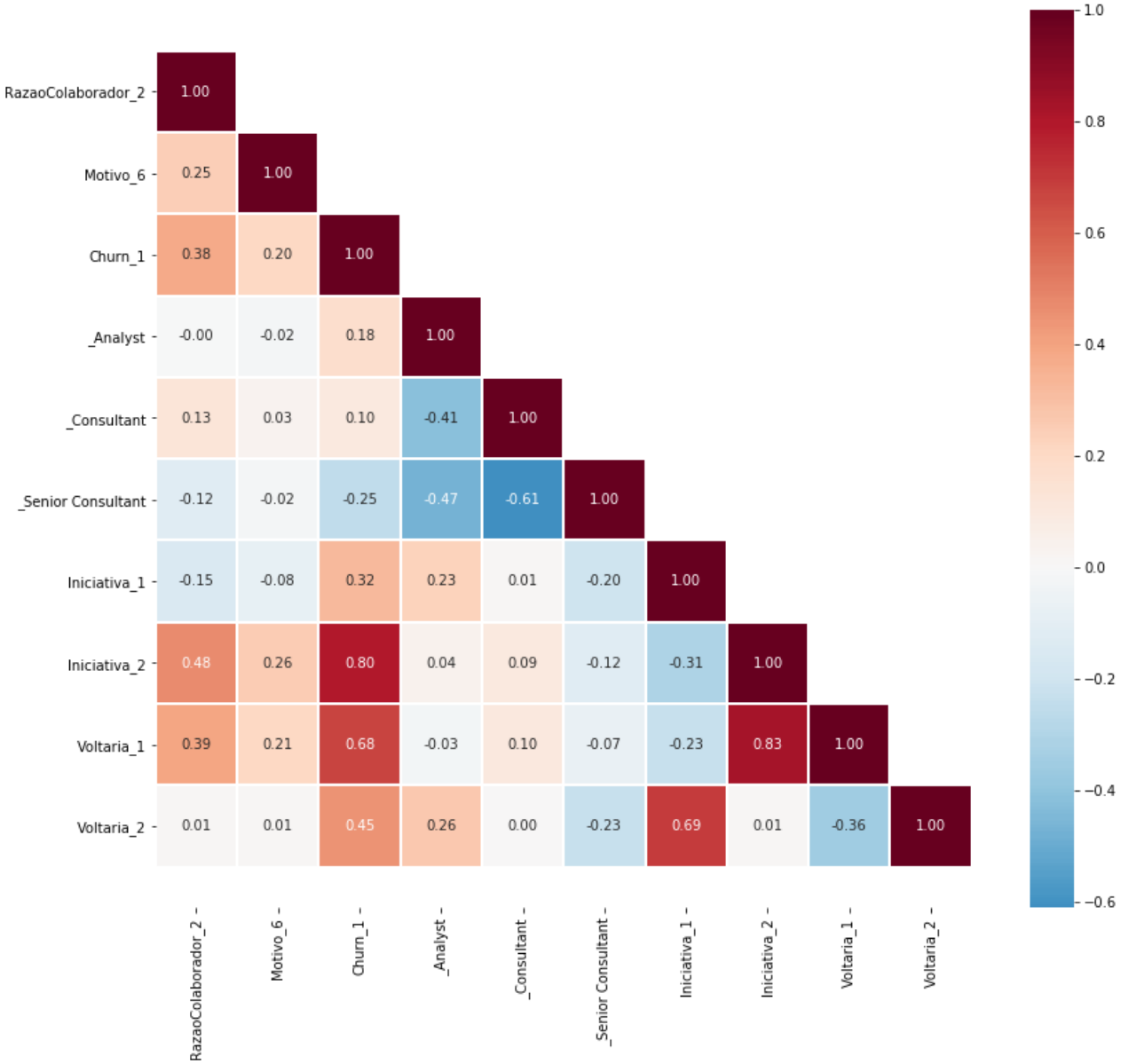

Figure 21 – Correlation Matrix - People quit because they search for career progress

5.3.2. People quit because they do not have enough skills in line with company standards

“People quit because they do not have enough skills in line with company standards" is an assumption, according to the definition in the introduction chapter, that reason of leaving is associated with involuntary turnover, i.e. the initiative of leaving is initiated by the company.

The focus is to confirm here is that the employee's motives are not always in the most common. Some reasons for employee leaving are also initiated by the company's decision.

The interest in researching involuntary turnover, too, is that it is possible to build mitigation plans. For instance, create more elaborate recruitment phases, with three steps in which different skills are testing at each phase.

As display in Figure 22, existing a very high correlation between "*Iniciativa _1*" and "*Motivo _4*", with a correlation of 0.95. The interpretation of these variables mentioned above, this reason is initiated by the intention of the company. Also, it can observe that this reason is more common in analyst grade, with a correlation between the variables of 0.19. While for consultant the correlation is very low, correlation approximately to zero. The senior consultant has a negative correlation.

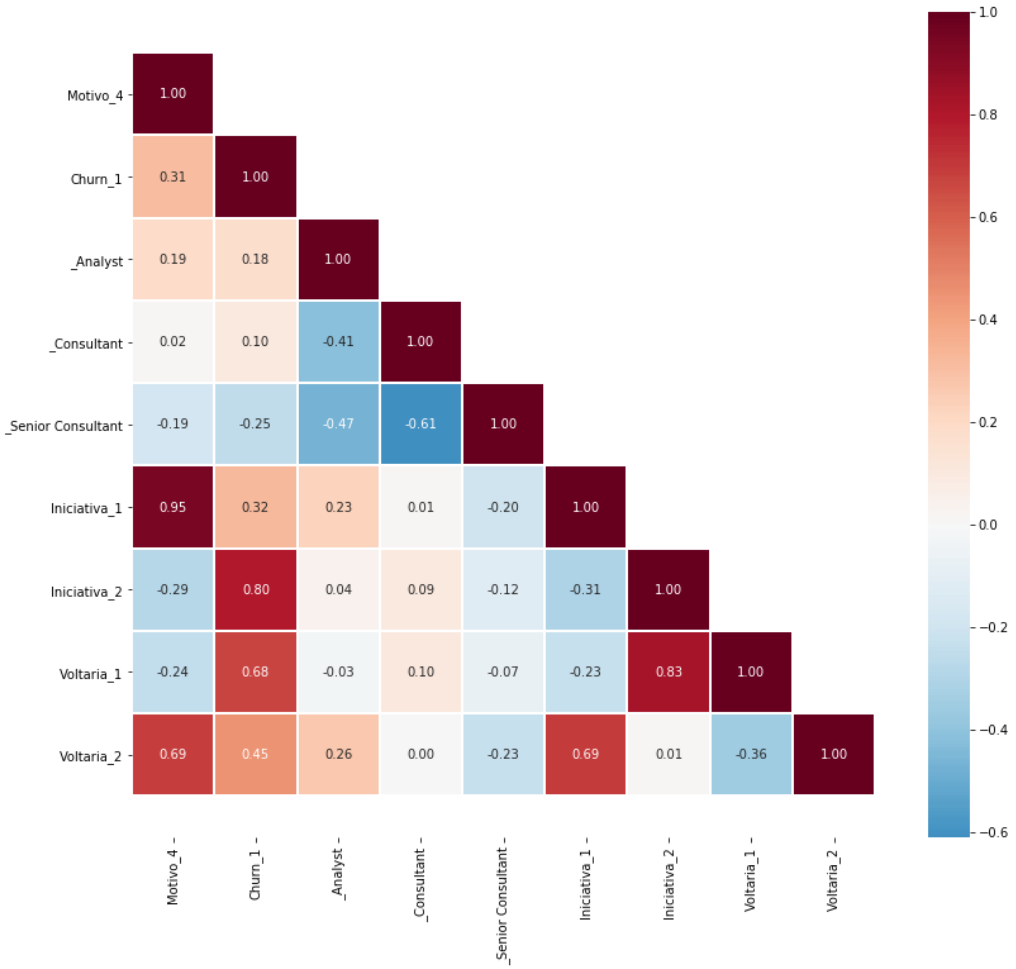

Figure 22 - Correlation Matrix - People quit because they do not have enough skills in line with company standards

It was possible to conclude that the recruiting phase is the next action plan that should review for better retention of employees.

The last evidence that proves the needs of the previous assumption is that the correlation with the variables "if it would hire again" is higher in the answer "No".

Chapter 6 – Conclusions and Further Work

6.1. Conclusion

The focus of this research was to build a transversal framework based on the CRISP-DM model that applies to any company. The purpose of the framework is to acknowledge the main reasons for the employees to leave the organization, in order to create employee retention plans. This analysis is built-it on the investigation of a paper [1] with common misconceptions, where the framework tries to identify key-variables, analyses to demystify if the misconceptions are true and applicable to the context of the problem.

Based on records and observations of a real project of a company, many analyses and correlations are performed to understand the most common reasons for an employee to leave a company.

The most common misconceptions are presented on paper [1] and the main results of the research are “People quit because of pay”, “People quit because they are dissatisfied with their jobs” and “There is little Managers can do to directly influence turnover decisions”.

A set of attributes in the dataset is presented in the framework, that is able to analyse and achieve the reasons for the employee's churn. First, it is relevant to evaluate the dataset, where the key attributes are identified, evaluated and compared to the target value of the employee churn. Many plots are observed to reach the most significant variables. After identifying the variables, correlations are performed based on attributes extracted from the visualization. First, the key-variables with the churn variables are analysed, and then all the variables that are strongly correlated with the churn variable are extracted. Finally, within the most correlated variables, those that allow validating each misconception are extracted to realize that other variables can be more correlated with the variable identified as key to prove the misconception.

After misconceptions validated, the framework also suggests, based on dataset attributes, possible assumptions that can be extracted.

6.2. Discussion

In this dissertation was applied the framework to the data of a real problem of a company and it was analysed a dataset with about 511 employees, where 275 are churn and 235 are no-churn employees.

According to the framework, there was an in-depth analysis of the data considering all employees, to try understand by evaluation period in which year more employees leave a company. Then through these analyses, we tried to understand which variables influenced more the churn variable. Thus, by visualization of graphics, compared variables such as professional category; if it would hire again; if it was at the will of the company or employee; and also some reasons filled by the company; all features are compared with the churn variable.

After the visualization of these variables and knowing the dataset, correlations performed to reinforce the visualization previously made. First, all the variables were correlated to identify the variable that allowed the response to each misconception. Some rules created that allows withdrawing assumptions about the reasons for employee leaving a company. The assumptions extracted are the following “People quit because they search for career progress” and “People quit because they do not have enough skills in line with company standards”. One assumption classified as voluntary turnover and another by the initiative of the company, the meaning is involuntary turnover.

In a group of the three misconceptions listed in the paper, only two were possible to be studied. Justified by, the lack of data proving the third as true.

In conclusion, it was possible to verify that misconception “People quit because of pay” and “People quit because they are dissatisfied with their jobs” apply to a business context and besides, it is possible to understand the following: The category that most motivates an employee leaving the company because of pay is the category of senior consultant, where this category usually search for higher salary progression; Although the reasons for dissatisfaction may incorporate n-factors, the grade that most leads to leaving for dissatisfaction, is the category of analyst and consultant; in the assumption “People quit because they search for career progress” the consultants are those who search for career progression; And at last, the company considers that some analysts may not have the skills or, do not consider employees with the standards that the company search, and end up leaving the company when the contract ends.

Additionally, to these conclusions, also find that one of the reasons for leaving derived from personal/family reasons, however, do not consider it as an analysis for the research, since it is a vast reason and can lead to false conclusions.

6.3. Research limitations and Future work

The dataset in research suffered many data transformation. The initial data did not have the necessary and enough variables for the study of the reasons for the employee's leave. It was essential to understand the data that was wanted to analyse to understand what kind of data was needed to create. Based on simple data like the employee's evaluation over the years, the data was converted into a single variable of which year the employee started, and what year the employee left the company. To apply the framework proposed, the dataset was very limited and also, had a very small sample of data.

It could become more interesting if it was known data such as: project satisfaction; project manager ID; gender (to analyse by gender, to see if there is gender discrimination, for example, if there are evidence that female gender employees, leaves for intention of company or employee); the number of travels (if it influences the exit); among many others that were not available in the dataset.

Additionally, since this dataset is limited to small findings, it is helpful to perform research with a larger dataset with more characteristics in the future to provide more precision about the turnover of employees and to understand if there is any discrepancy in the results depending on the size of the dataset.

For the future scope of this research, the variables collected should be used to make predictions. Different machine learning algorithms should be investigated (Logistic Regression, Random Forest, SVM, KNN, Decision Tree Classifier, Gaussian Naïve Bayes) and applied to the dataset.

References

[1] D. G. Allen, P. C. Bryant, and J. M. Vardaman, “Retaining Talent : Replacing Misconceptions With,” *Acad. Manag. Perspect.*, vol. 24, no. 2, pp. 48–64, 2010.

[2] V. V. Saradhi and G. K. Palshikar, “Employee churn prediction,” *Expert Syst. Appl.*, vol. 38, no. 3, pp. 1999–2006, 2011, doi: 10.1016/j.eswa.2010.07.134.

[3] J. P. Hausknecht and C. O. Trevor, “Collective turnover at the group, unit, and organizational levels: Evidence, issues, and implications,” *J. Manage.*, vol. 37, no. 1, pp. 352–388, 2011, doi: 10.1177/0149206310383910.

[4] D. r. Dalton, W. D. Todor, and D. M. Krackhardt, “Turnover Overstated: The Functional Taxonomy,” *The Academy of Management Review*, vol. 7, no. 1, pp. 117–123, 1982.

[5] S. N. Mishra, D. R. Lama, and Y. Pal, “Human Resource Predictive Analytics (HRPA) For HR Management In Organizations,” *International Journal of Scientific & Technology Research*, vol. 5, no. 5, pp. 33–35, 2016.

[6] J. L. Cotton and J. M. Tuttle, “Employee Turnover: A Meta-Analysis and Review with Implications for Research.,” *Acad. Manag. Rev.*, vol. 11, no. 1, pp. 55–70, 1986, doi: 10.5465/amr.1986.4282625.

[7] C. S. and R. W. Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, *CRISP-DM 1.0 - Step-by-step data mining guide*. CRISP-DM Consortium, 2000.

[8] W. Vorhies, “CRISP-DM – a Standard Methodology to Ensure a Good Outcome - Data Science Central,” *Data Science Central*, 2016. [Online]. Available: https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome. [Accessed: 09-Sep-2020].

[9] S. Harrison and P. A. Gordon, “Misconceptions Of Employee Turnover: Evidence-Based Information For The Retail Grocery Industry,” *J. Bus. Econ. Res. Quart.*, vol. 12, no. 2, 2014.

[10] H. Ongori, “A review of the literature on employee turnover,” *Am. J. Bus. Manag.*, vol. 5, no. 6, pp. 226–228, 1977, doi: 10.1177/036354657700500601.

[11] T. Y. Park and J. D. Shaw, “Turnover rates and organizational performance: A meta-analysis,” *J. Appl. Psychol.*, vol. 98, no. 2, pp. 268–309, 2013, doi: 10.1037/a0030723.

[12] C. R. Williams, “Reward contingency, unemployment, and functional turnover,” *Hum. Resour. Manag. Rev.*, 2000.

[13] U. of M. Phil C. Bryant, Assistant Professor of Management and Marketing, Columbus State University; and David G. Allen, Distinguished Professor of Management, “Compensation, Benefits and Employee Turnover: HR Strategies for Retaining Top Talent,” *Compens. Benefits Rev.*, vol. 45, no. 3, pp. 171–175, 2013, doi: 10.1177/0886368713494342.

[14] W. R. Mckinney, K. R. Bartlett, and M. A. Mulvaney, “Measuring the costs of employee turnover in illinois public park and recreation agencies: An exploratory study,” *J. Park Recreat. Admi.*, vol. 25, no. 1, pp. 50–74, 2007.

[15] R. Cheripelli and P. V. Ajitha, “Evaluation of machine learning models for employee churn prediction,” *Test Eng. Manag.*, vol. 83, no. Icici, pp. 18–22, 2020.

[16] S. M. Abbasi and K. W. Hollman, “Turnover: The real bottom line,” *Public Pers. Manage.*, vol. 29, no. 3, pp. 333–342, 2000, doi: 10.1177/009102600002900303.

[17] E. Ribes, K. Touahri, and B. Perthame, “Employee turnover prediction and retention policies design: a case study,” *2010 Math. Subj. Classif.*, pp. 1–12, 2017.

[18] T. Attri, “Why an Employee Leaves: Predicting using Data Mining Techniques MSc Research Project Data Analytics,” School of Computing National College of Ireland, 2018.

[19] J. F. Lawrence, E. S. Nielsen, and I. M. Mackerras, “Motivational Factors of Employee Retention and Engagement in Organizations,” *Int. J. Adv. Manag. Econ.*, pp. 1–32, 1991.

[20] E. P. and A. D. D. Prodromos D. Chatzoglou, Eftichia Vraimaki, Eleni Komsiou, “Factors Affecting Accountants’ Job Satisfaction and Turnover Intentions: A Structural Equation Model”, 8thInternational Conference on Enterprise Systems, Accounting and Logistics,” vol. 8, no. July, pp. 130–147, 2011.

[21] R. Punnoose and P. Ajit, “Prediction of Employee Turnover in Organizations using Machine Learning Algorithms,” *International Journal of Advanced Research in Artificial Intelligence*, vol. 5, no. 9, pp. 22–26, 2016.

[22] J. A. Grissom, S. L. Viano, and J. L. Selin, “Understanding Employee Turnover in the Public Sector: Insights from Research on Teacher Mobility,” *Public Adm. Rev.*, vol. 76, no. 2, pp. 241–251, 2016, doi: 10.1111/puar.12435.

[23] P. Stamolampros, N. Korfiatis, K. Chalvatzis, and D. Buhalis, “Job satisfaction and employee turnover determinants in high contact services: Insights from Employees’Online reviews,” *Tour. Manag.*, vol. 75, no. April, pp. 130–147, 2019, doi: 10.1016/j.tourman.2019.04.030.

[24] S. N. Khera and Divya, “Predictive Modelling of Employee Turnover in Indian IT Industry Using Machine Learning Techniques,” *Vision*, vol. 23, no. 1, pp. 12–21, 2019, doi: 10.1177/0972262918821221.

[25] H. Bendemra and D. Ph, “Building an Employee Churn Model in Python to Develop a Strategic Retention Plan,” *Towards Data Science*, 2020. [Online]. Available: https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d. [Accessed: 16-Aug-2020].

[26] R. Bose, “Advanced analytics: opportunities and challenges,” *Ind. Manag. Data Syst.*, vol. 109, no. 2, pp. 155–172, 2009, doi: 10.1108/02635570910930073.

[27] T. H. and R. T. Gareth James, Daniela Witten, *An Introduction to Statistical Learning with Applications in R*, Vol. 6. Ne., vol. 64, no. 9–12. Springer Series in Statistics, 2013.

[28] J. Hastie, Trevor, Tibshirani, Robert and Friedman, *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*, NY Springe. Springer Series in Statistics, 2001.