Table of Contents

1. Introduction

In the last decades, having the best machines was enough to be competitive or to dominate an industrial sector. Nowadays, the company that has more engaged and productive employees will have a better chance of winning market competition. For this reason, companies can not lose important employees and when that begins to happen you need to understand why, to prevent this from happening. The Human Resource analytics data set is used to explain the first steps in the data analysis path. In this first part it is presented how to familiarize yourself with the data set by performing a descriptive analysis.
Techniques such as exploratory data analysis (EDA) allow us to present the data in a more meaningful way, applying general statistics methods and exploratory graphics, that allow a simpler interpretation before engaging a machine learning algorithm.

1.1 About the Data Set

The Human Resource Analytics is a simulated data set from kaggle and the focus is to understand why the best and most experienced employees are leaving the company. By the exploration of this data set its possible to extract good insights of a problem that the Human Resource department deals daily. In many Industries retaining the best employees is a question of a long term strategy, and can impact the companies growth or put in financial risk, mainly if the employees leave to work for the competitor.

1.2 Exploratory Data Analysis(EDA)

Exploratory data analysis employs a variety of techniques (mostly statistical graphics) before making inferences from data. It is essential to examine all variables in the data set to:

i. Catch Mistakes
ii. Generate Hypotheses
iii. See patterns in the data
iv. Extract important variables
v. Detect out-liers and anomalies
vi. Gain deep familiarity with the data set
vii. Refine selection of features that will be used to build the machine learning models.

It is Important to not skip the EDA process, because it can highlight inaccurate models or accurate models on the wrong data. This data set contains 14999 objects and 10 attributes described below:

Variables	Descriptions
satisfaction_level	Satisfaction Level
last_evaluation	Last evaluation
number_project	Number of projects
average_monthly_hours	Average monthly hours
time_spend_company	Time spent at the company
Work_accident	Whether they have had a work accident
left	Whether the employee has left
promotion_last_5years	Whether had a promotion in the last 5 years
sales	Departments (column sales)
salary	Salary

1.3 Preprocessing the data set

Before starting the process, it's important to answer if it's clear what kind of problems we are dealing with, because in many cases it isn't so simple to identify it. A good understanding of the problem will help to choose the right data mining and machine learning techniques to make the right predictions. Thus, the first step is preprocessing the data to look for missing, incomplete or noise values, because, in real world, the raw data can be collect from many sources like sensors, websites, public data and many others.

2. Descriptive Analysis

The descriptive Analysis is used to simplify and summarize the main characteristics of the data set. In other words, show what kind of information the data set has. The Pandas method described generates a descriptive statistic that summarizes the central tendency, dispersion and shape of the data set. By using this method in the Human Resource data set, the possible important insights are;

1. That approximately 24% of the employees left the company.
2. The satisfaction level is around 62% and performance is around 72%.
3. Employees work in average on 4 projects with 200 hours worked per month.

2.1 How many employees work in each department?

Depending on how many employees work on each department, you can learn more about the company segment.

Sales	4140
Technical	2720
Support	2229
IT	1227
Product_mng	902
Marketing	858
RandD	787
Accounting	767
HR	739
Management	630

2.2 How many employees per salary range?

The employee's salary is divided into Low (1), Medium (2) and High (3) distributed as follows:

1	7,316
2	6,446
3	1,237

2.3 How many employees per salary range and department?

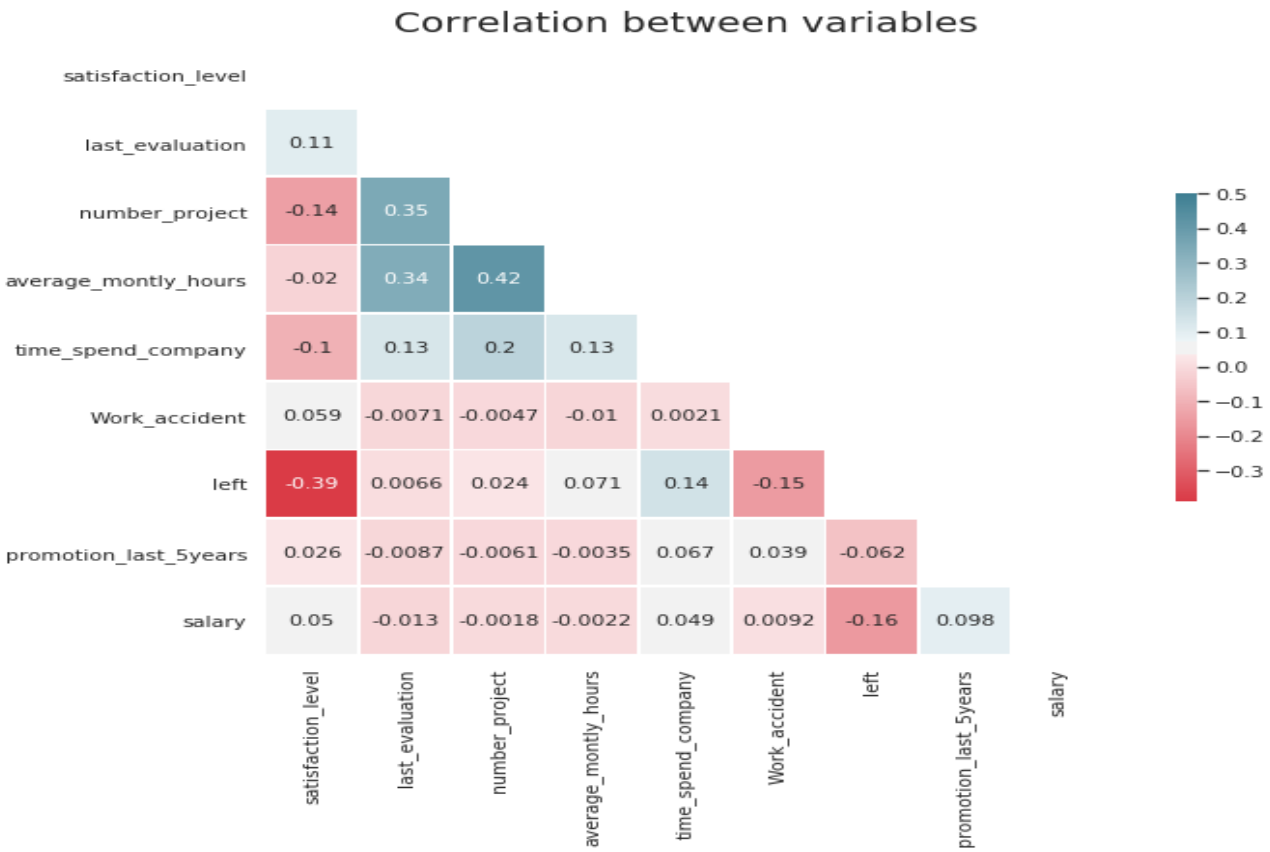
salary department	1	2	3
IT	609.0	535.0	83.0
RandD	364.0	372.0	51.0
accounting	358.0	335.0	74.0
hr	335.0	359.0	45.0
management	180.0	225.0	225.0
marketing	402.0	376.0	80.0
product_mng	451.0	383.0	68.0
sales	2099.0	1772.0	269.0
support	1146.0	942.0	141.0
technical	1372.0	1147.0	201.0

3. Correlation Analysis

Correlation is a very useful statistical analysis that describes the degree of relationship between two variables. The table below and the heat map to see what relationships are in the data.

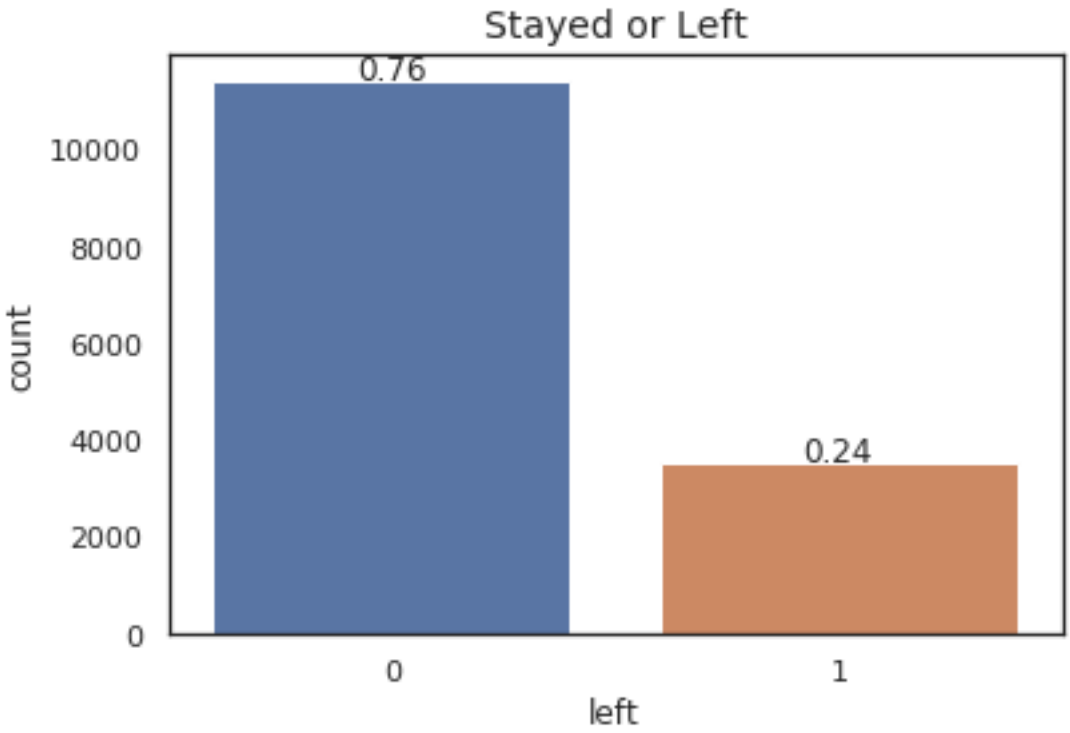
From the heat map it is possible to see:

1. Negative correlation of(-0.39) between satisfaction level and the employees that left the company.
2. The highest positive correlation is between number f=of projects and average monthly hours(0.42)
3. Last evaluation is high correlated to number of project(0.35)
4. and average monthly hours(0.34)
5. Work accident have a low negative correlation(-0.15) and salary(-0.16) with employees that left.



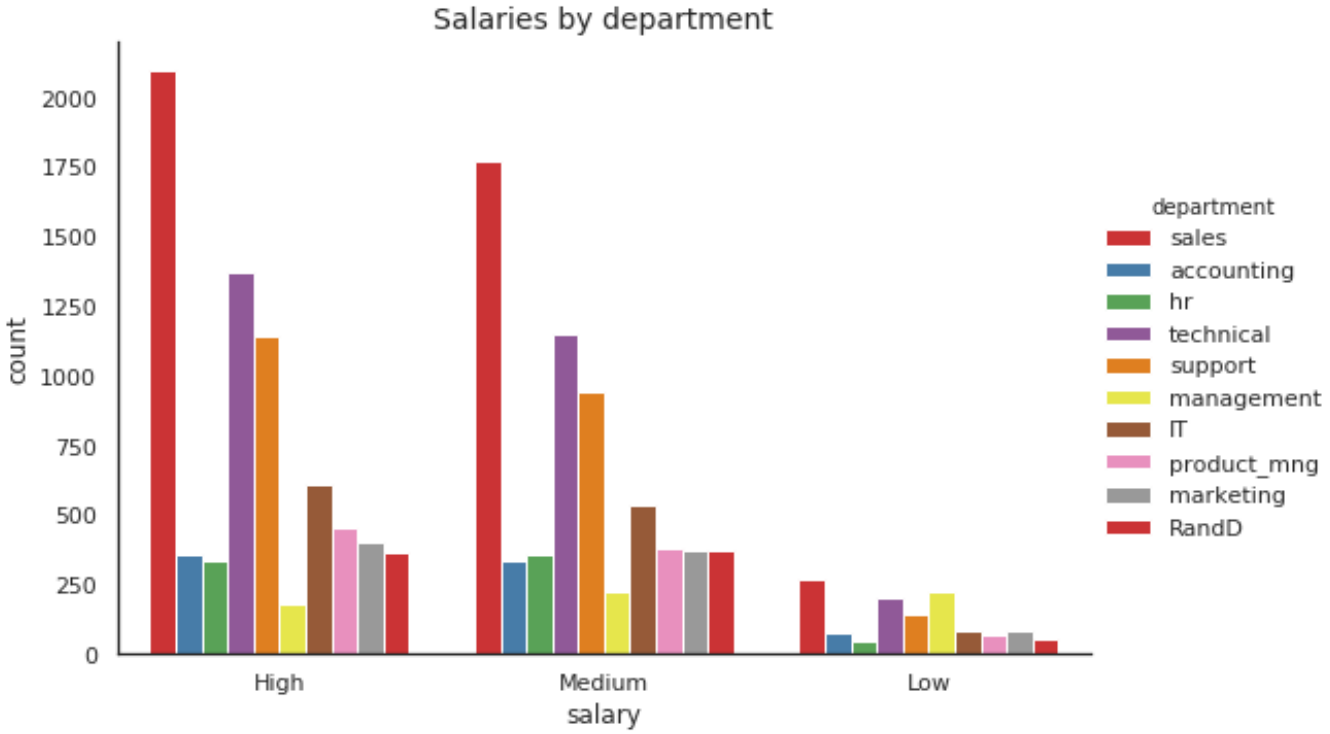
4.Hypothesis

Let's extract some information and test some hypothesis, to begin with 3,571 which is equal to 24% employees left the company.



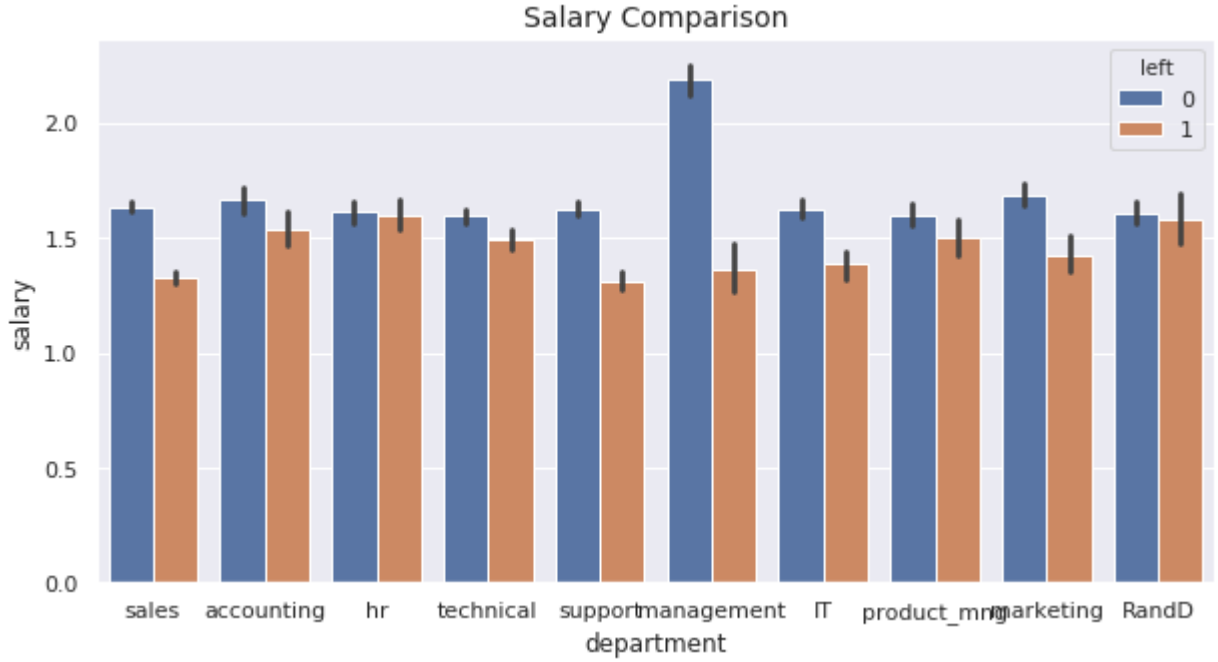
4.1 The first Hypotheses

Is Salary the reason why employees left the company?



From the graph above it is possible to see distribution of the salaries by department.

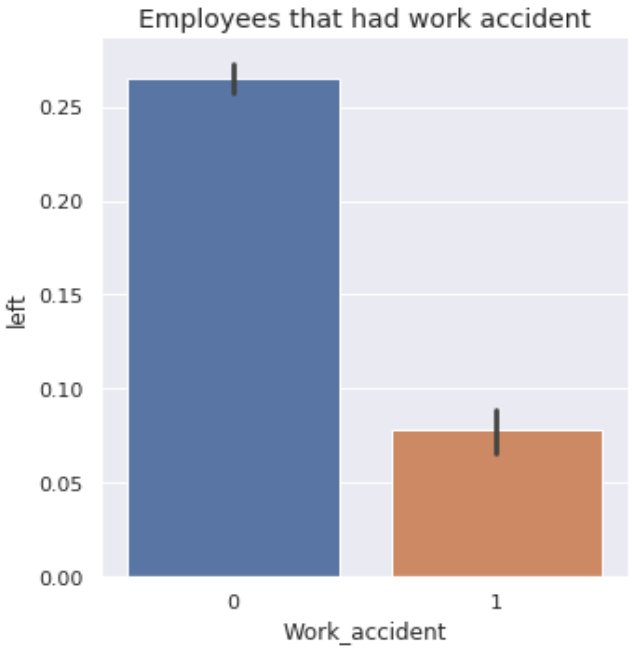
1. Most of the employees of the sales department have low or medium salaries, this may be due that in some companies the sales commission is paid separately.
2. Technical department is in the second place where most employees receives low and medium salaries.



In the graph salaries comparison:
The management has the biggest difference between the salary of the employees who stayed and those who left.
It is not possible to see huge difference in other departments.
The conclusion here therefore is that this hypothesis is weak to be the main reason why the employee left the company.

4.2 The Second Hypotheses

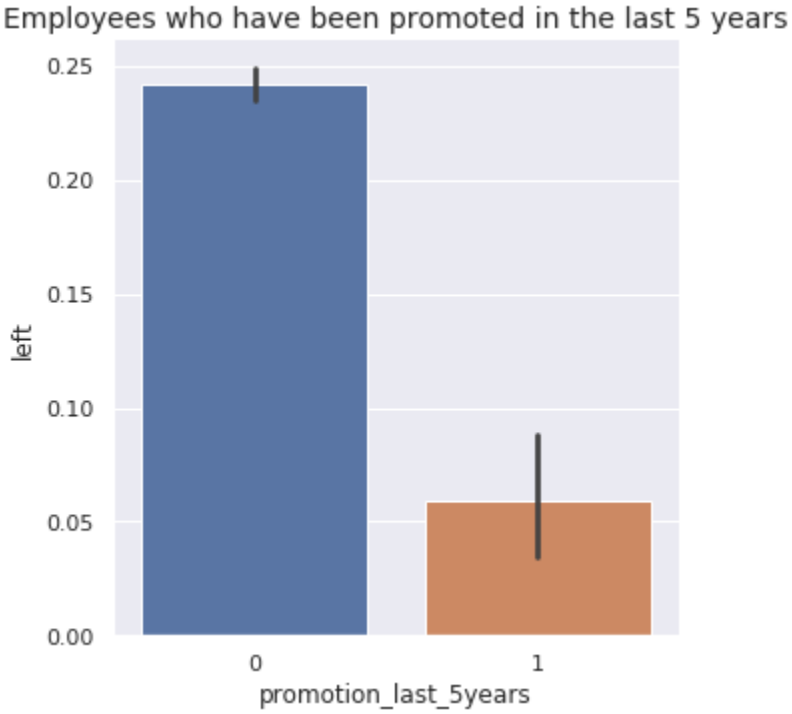
Is it that it is a dangerous job? one that is likely to accidents?
Do employees leave because it is not safe?



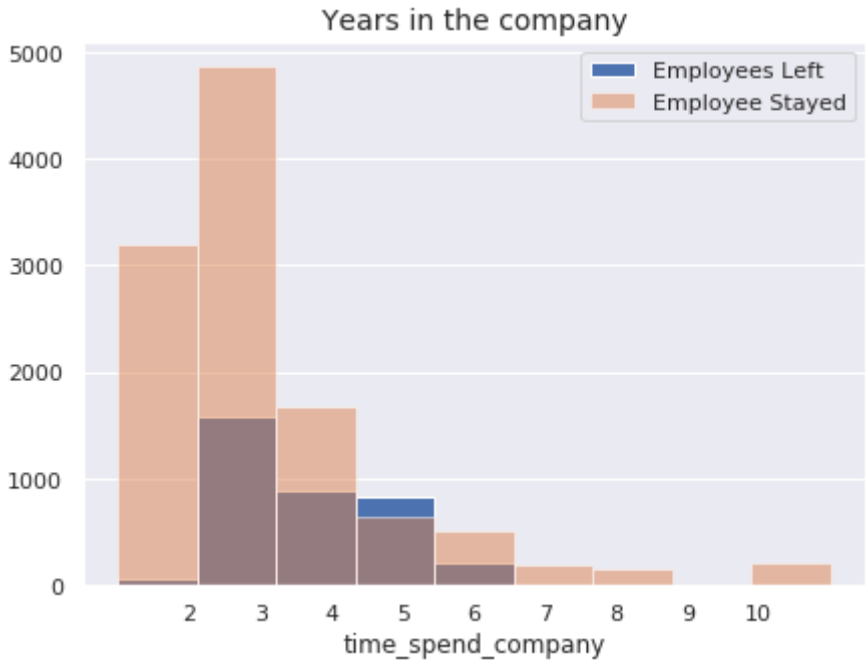
About 14% of the employees had a work accident,of this number only 169 employee left the company had accident, this therefore proves that this hypothesis needs to be carded as well.

4.3 The third Hypothesis

Is this company a good place to work professionally?



In the last five years only 319 employees had been promoted, this is equivalent to only 2% of all employees. This may be a problem because if it is difficult to get promoted many employees become unmotivated and start looking for new jobs.

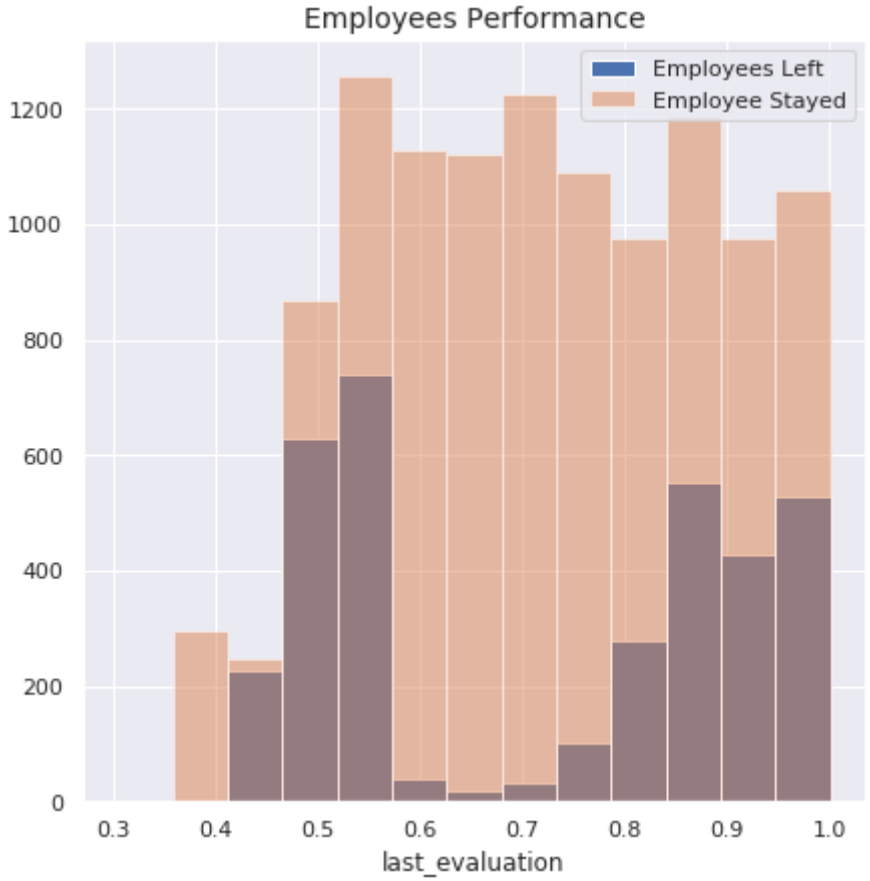


In the graph above we can identify the following important characteristics:

1. Employees with more than seven years did not leave the company, maybe because with the passing of years they are more comfortable and not so Interested in looking for a new challenge in another company.
2. The problem starts when the employee have more than 3 years and get worse when they achieve five years.
3. It is too early to say that the difficulty to get promoted is the main reason for the leaving of the employee but more research is needed.

5. Performance Analysis

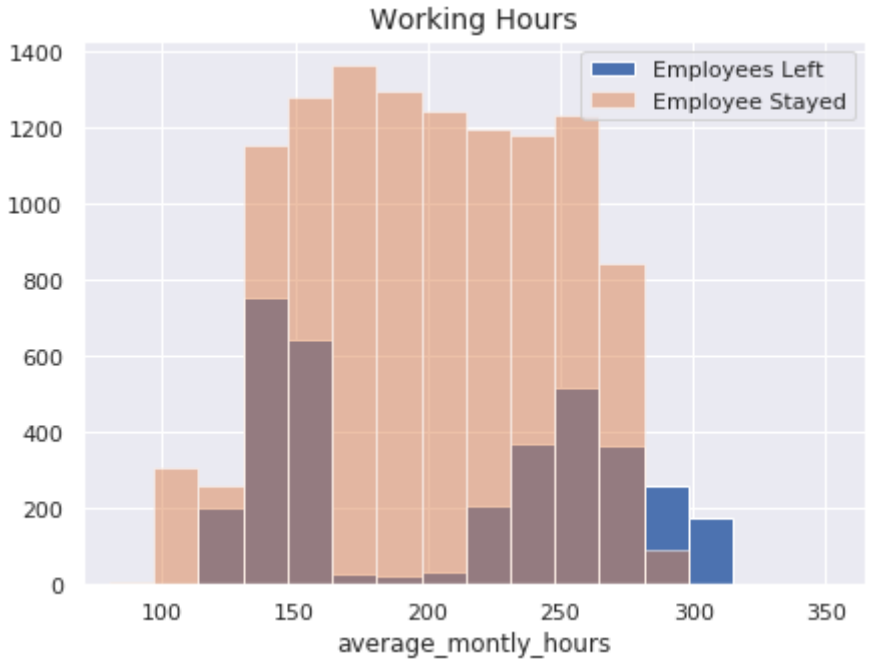
There are two distinct groups of employees. A group with poor performance and another with high performance employees. It is natural that employees that don't work well leave the company, but the main problem is that the high performance employees are leaving too and it's necessary to understand why.



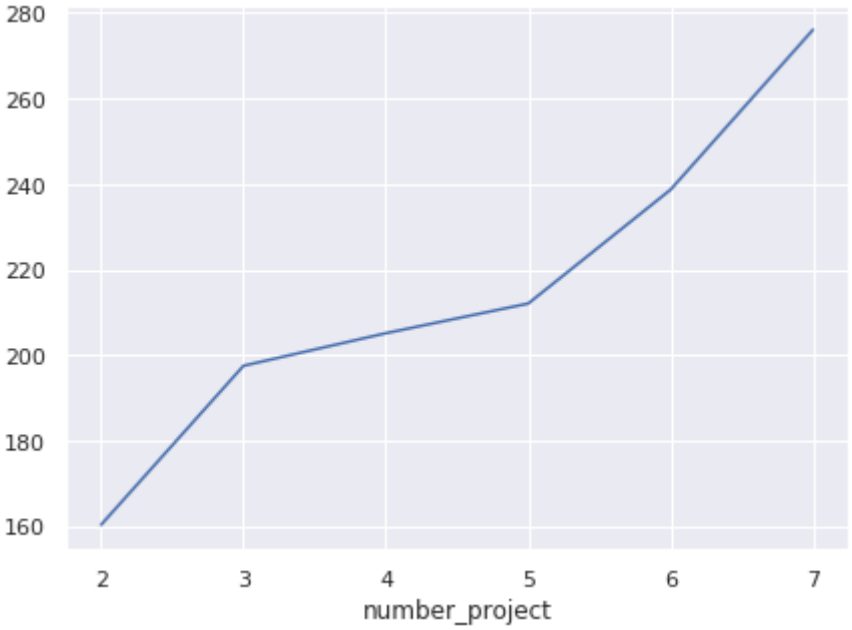
It is possible to see that 98% of employees with few projects that left also have poor performance. Also, 95% of the employees with 5 or more projects that left the company had the highest performance. Three or four are the best number of projects since where three or four projects are involved there is higher number of employees who stayed in the company.

6. Working Hours

There are two groups of employees, a group that woks fewer hours and another that works more hours compared to the average hours worked.



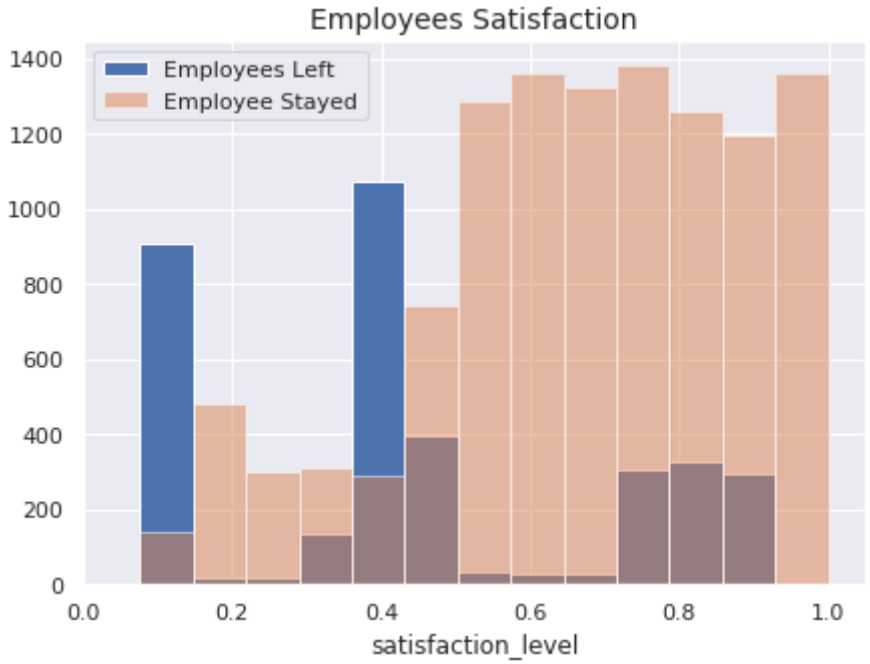
It is clearly possible to see that employees with six or more projects work on average 20% more hours.



The employees that left may be grouped as:
Employees with two projects and worked less than the average hours of the company.
Employee with five or more project that worked at least 20% more than average.

7. Satisfaction Level

It is possible to see three interesting peaks in the satisfaction levels of the employees that left the company.
We have a peak of employees who are totally disappointed.
Another peak at 0.4, representing another group with the satisfaction level below the average.
Another amount in the range 0.7 and 0.9, with employees that left, although the high satisfaction.



It is clear from the results a drop in satisfaction when employees are working on 6 or more projects.
From the employees that left with a high performance, four or more years in the company and working on five or more projects had:

1. Low satisfaction.
2. worked most hours.
3. Have not been promoted in the last five years.

Conclusion

This is a relatively young company, on average, employees have three or four years in the company and the oldest employees are working ten years, when it comes to salaries the biggest difference in the salary from who stayed and those who left, was found in the management department, in the others departments although the salaries of who stayed are higher in average, it is not a big difference.
In respect of work accidents, the number of employees that had a work accident is about 14%, of which only 169 employees left the company, so didn't seem to have a correlation with the employees leaving.

In five years only 2% of the employees were promoted, It Is possible that many employees get unmotivated and start planning to leave. Employees with seven years or longer in the company didn't leave while employees with 5 years have more chances to leaving.
There are two distinct groups of employee's performance that left, a group with poor performance with two projects and another group with high performance with five or more projects. It is not necessary to retain all the employees hence the focus is on keeping employees with high performance, the employees with four years in the company have the lowest average satisfaction level of all in the company with (0.47). The satisfaction drops when the employees are working in five or more projects thus a number of three or four projects seems to be ideal, independent of the time spent in the company. The employees with five or more projects that left also worked at least 20% more hours than the average hours of the company.
The satisfaction level of the employees that left is grouped in totally disappointed, below the average satisfaction and satisfied.