EDA Report

(HUMAN RESOURCE ANALYTICS)

OBJECTIVE:

The objective of our project is to focus on the use of Data Analytics in the field of Human Resource Development and its ability to reap organizational benefits by a recommendation for retention of talented work force. With analysis and approach through different process flows in the Data Science which includes statistical inference and exploratory data analysis. The main goal is to understand the reasoning behind employee turnover and to come up with a model to classify an employee risk of attrition.

INTRODUCTION:

The major function of Human Resource is to evaluate talent management and development techniques and identify mechanisms to more effectively manage human capital. As human behaviour is much more complex and less predictable than machinery or other tangible assets, the optimization of human capital allocation and retaining is major task.

DATA FOR ANALYSIS:

The Variables in the dataset include:

- Satisfaction Level
- Last evaluation
- Number of projects
- Average monthly hours
- Time spent at the company
- Whether they have had a work accident
- Whether they have had a promotion in the last 5 years
- Departments (column sales)
- Salary
- Whether the employee has left

EXPLORATORY DATA ANALYSIS

```
str(hr)
'data.frame':          14999 obs. of  10 variables:
 $ satisfaction_level   : num  0.38 0.8 0.11 0.72 0.37 0.41 0.1 0.92 0.89 0.42 ...
 $ last_evaluation      : num  0.53 0.86 0.88 0.87 0.52 0.5 0.77 0.85 1 0.53 ...
 $ number_project       : int  2 5 7 5 2 2 6 5 5 2 ...
 $ average_montly_hours : int  157 262 272 223 159 153 247 259 224 142 ...
 $ time_spend_company   : int  3 6 4 5 3 3 4 5 5 3 ...
 $ Work_accident        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ left                 : int  1 1 1 1 1 1 1 1 1 1 ...
 $ promotion_last_5years: int  0 0 0 0 0 0 0 0 0 0 ...
 $ sales                : Factor w/ 10 levels "accounting","hr",..: 8 8 8 8 8 8 8 8 8 8 ...
 $ salary               : Factor w/ 3 levels "high","low","medium": 2 3 3 2 2 2 2 2 2 2 ...
```

Converting the work_accident and promotion_last_5years as factor variables.

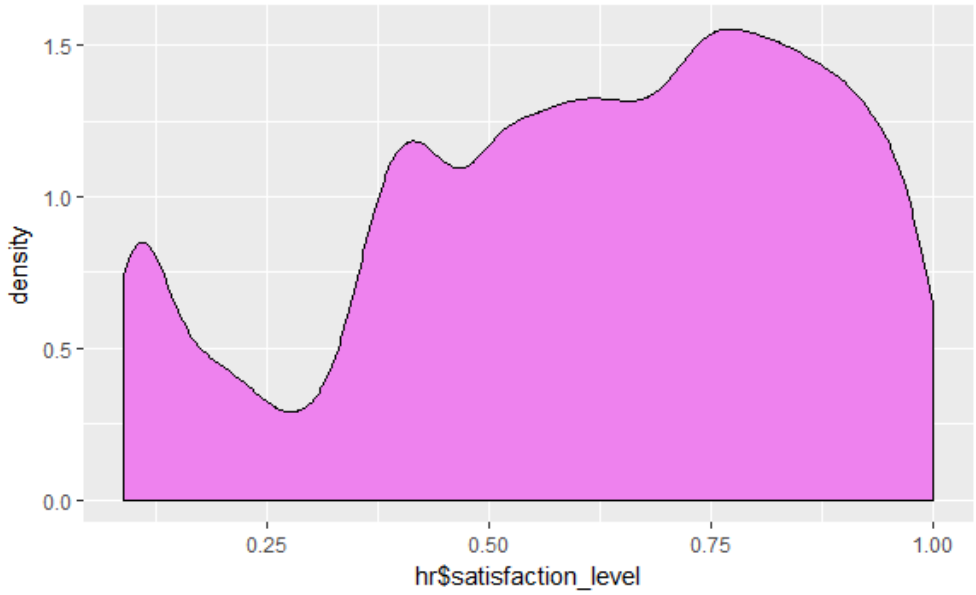
```
hr$Work_accident = as.factor(hr$Work_accident)
hr$promotion_last_5years = as.factor(hr$promotion_last_5years)
```

```
> str(hr)
'data.frame':          14999 obs. of  10 variables:
 $ satisfaction_level   : num  0.38 0.8 0.11 0.72 0.37 0.41 0.1 0.92 0.89 0.42 ...
 $ last_evaluation      : num  0.53 0.86 0.88 0.87 0.52 0.5 0.77 0.85 1 0.53 ...
 $ number_project       : int  2 5 7 5 2 2 6 5 5 2 ...
 $ average_montly_hours : int  157 262 272 223 159 153 247 259 224 142 ...
 $ time_spend_company   : int  3 6 4 5 3 3 4 5 5 3 ...
 $ Work_accident        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ left                 : int  1 1 1 1 1 1 1 1 1 1 ...
 $ promotion_last_5years: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ sales                : Factor w/ 10 levels "accounting","hr",..: 8 8 8 8 8 8 8 8 8 8 ...
 $ salary               : Factor w/ 3 levels "high","low","medium": 2 3 3 2 2 2 2 2 2 2 ...
```

UNIVARIATE ANALYSIS

1. SATISFACTION_LEVEL

With the data, we can see that there are two numeric variables, satisfaction level and last evaluation; these two are scores from 0 to 1, with 0 the worst score and 1 the best score. The two factor variables: sales and salary; this two are category variables related to the salary (low, medium and high) and the department of the employee. The other six variables are integer related to the years spent in the company, the average monthly hours, the number of projects (from 2 to 7), if the employee have a work accident (0 = NO, 1 = YES), if the employee has left the company (0 = NO, 1 = YES) and if the employee has a promotion in the last 5 years (0 = NO, 1 = YES).



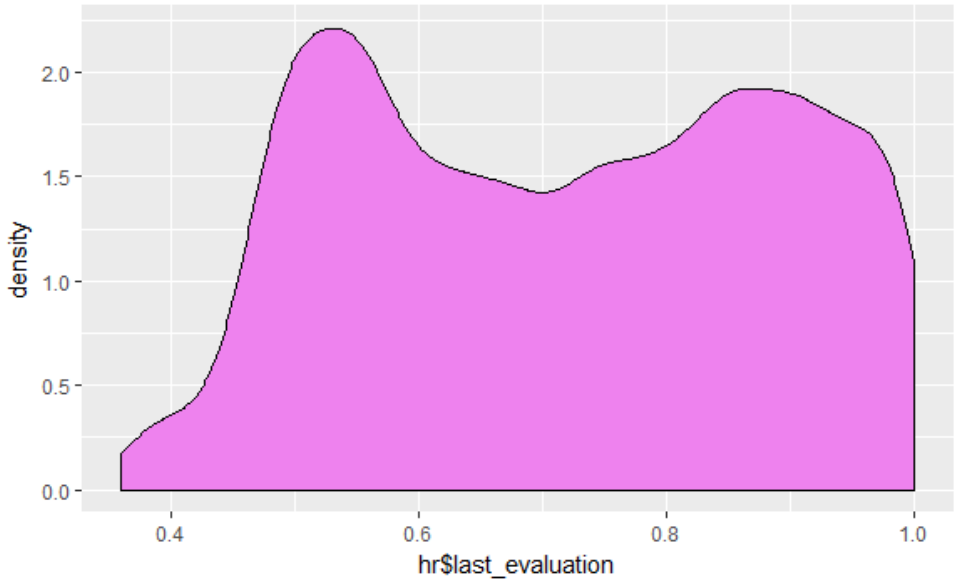
In this graph the density of the satisfaction level has a bimodal distribution, with two peaks, the first around 0.12 and the second around 0.76. In Bivariate and Multivariate plots section the two peaks are analysed in order to find the causes and the effects of this distribution.

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0900  0.4400  0.6400  0.6128  0.8200  1.0000
```

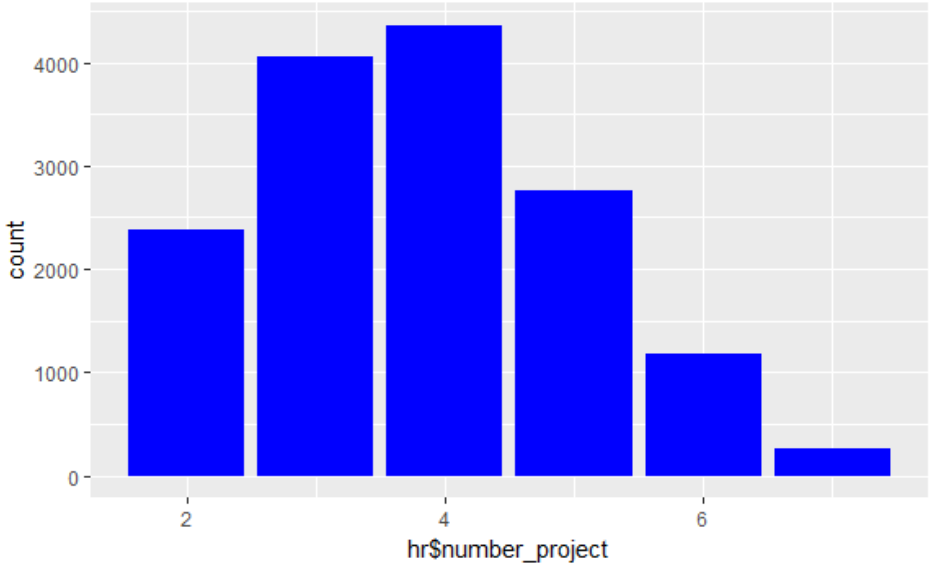
With the summary, instead, seems to have a normal distribution with a median value of 0.640.

2. LAST EVALUATION

Like the satisfaction level's graph, the last evaluation's graph has a bimodal distribution with a peak around 0.53 and another peak around 0.86. The initial thought was last evaluation distribution is a normal distribution with a peak around 0.7 and two decreasing tails for excellent and worst employees. The summary in this case, don't describe correctly the visualization distribution and may lead to an incorrect analysis.



3. NUMBER OF PROJECTS

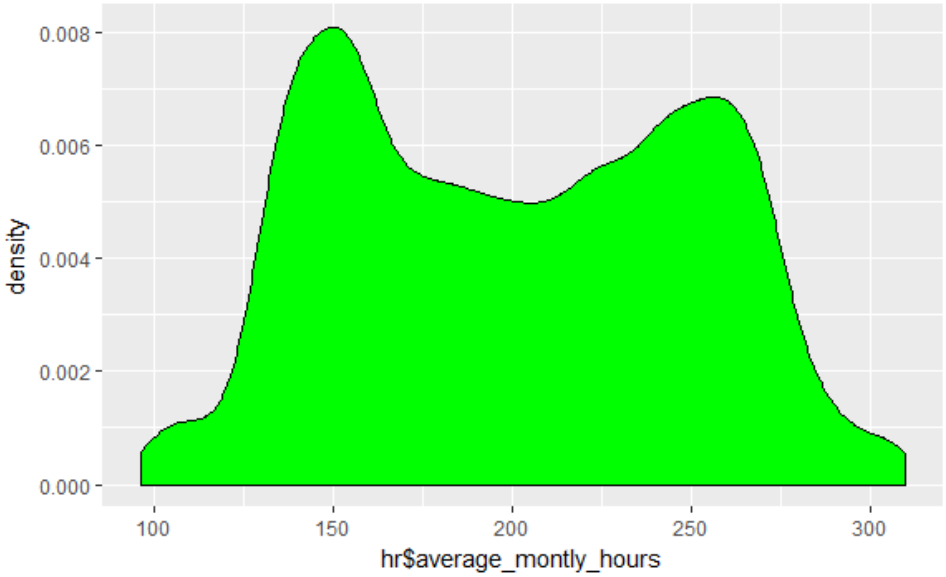


There is a peak in 3 and 4 number of projects for every employee, a long right tail on the right (the number of employees with number projects over 5 are very small) and a short left tail. This graph has a positive skew distribution.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 2.000   3.000   4.000   3.803   5.000   7.000
```

With this variable, the visualization fits perfectly the summary.

4. AVERAGE MONTHLY HOURS



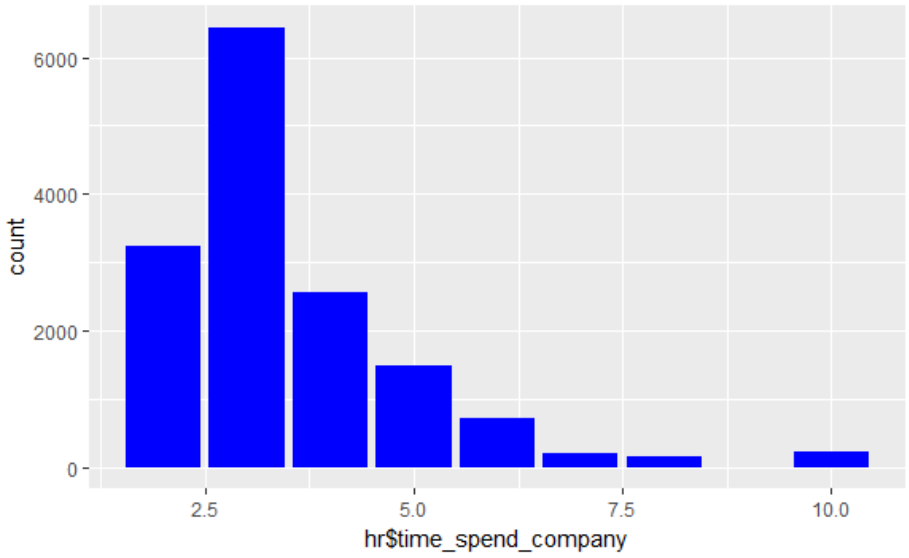

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   4.000   3.803   5.000   7.000
```

Like the first two variables, this variable as a bimodal distribution with two peaks, one in 150 and one around 270. These monthly hours are probably connected with the number of projects, but if there is a correlation in the Bivariate plot section has to be analysed.

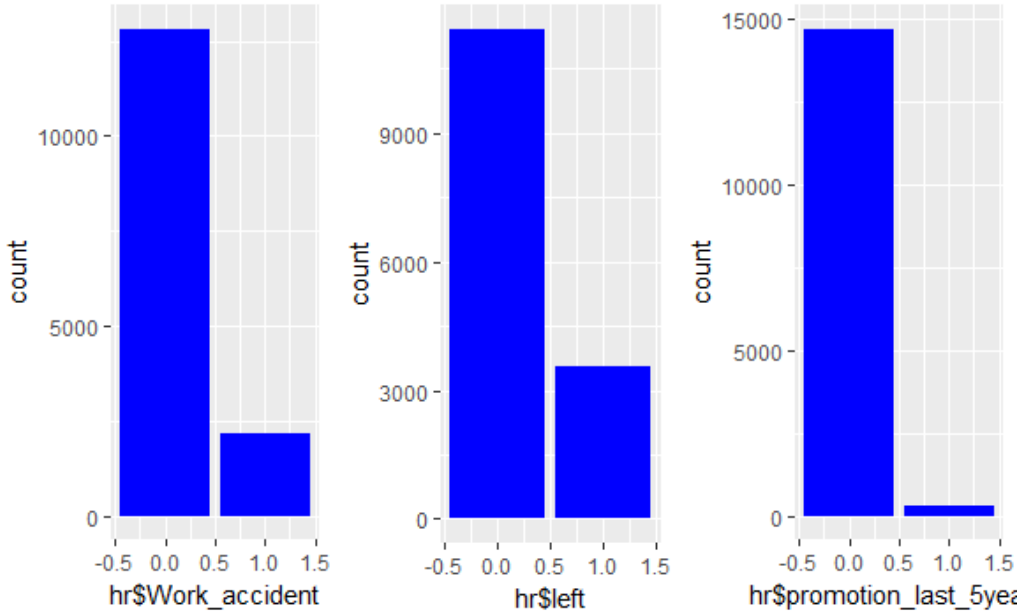
Next variable is time spent in the company

5. TIME SPENT IN THE COMPANY

This plot is quite interesting because it has a vision of the distribution of the years in company; this distribution could hypothetically show that there are a lot of people who left after 3 years in company or the company analysed has hired the majority of the employee from 2 to 4 years ago. The first hypothesis in the Bivariate plot section, but the second hypothesis, isn't possible to analyse because the year of hiring is required to better understand this results.



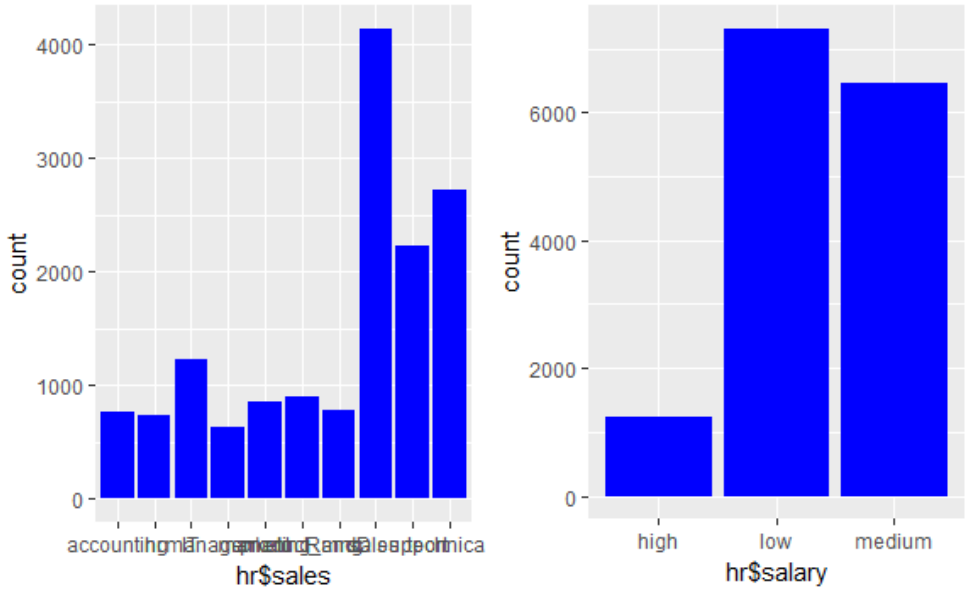
6. WORK-ACCIDENT-LEFT-PROMOTION LAST 5 YEARS



In the first plot (Work accident), in the company the number of employees with accident is more in number and in the bivariate plots section analysis of the department with the higher number and if it's correlated with the leaves. The second plot (Left), describe the flow of leaves and with this EDA to understand better the causes for the employee's leaves. Last but not least, the plot for the promotions. This plot describes a situation with a little percent of the employees with a promotion in the last 5 years and to explore if this variable is correlated with the leaves.

7. SALARY--SALES

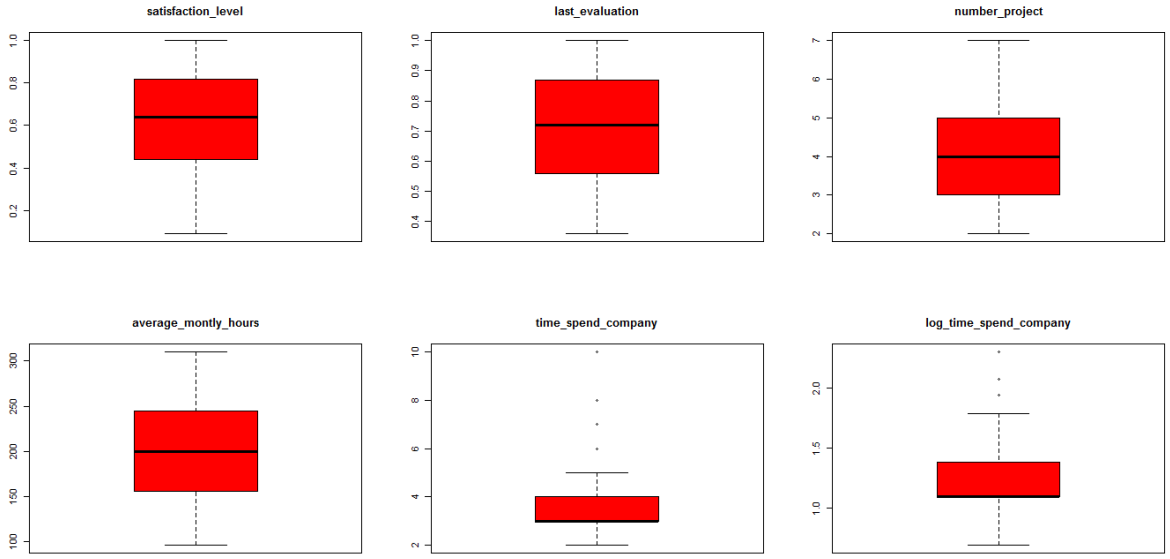
The last two variables are two categorical variables: sales and salary. Sales describe the departments in the company and salary, the salary of the employees.



In this two graph that there are departments with an elevate number of employees like sales, support and technical, and that the salary in the company are principally low and medium.

BOXPLOTS of the Quant variables:

Quantitative(numeric) variables are plotted using box plots to understand the data skewness.

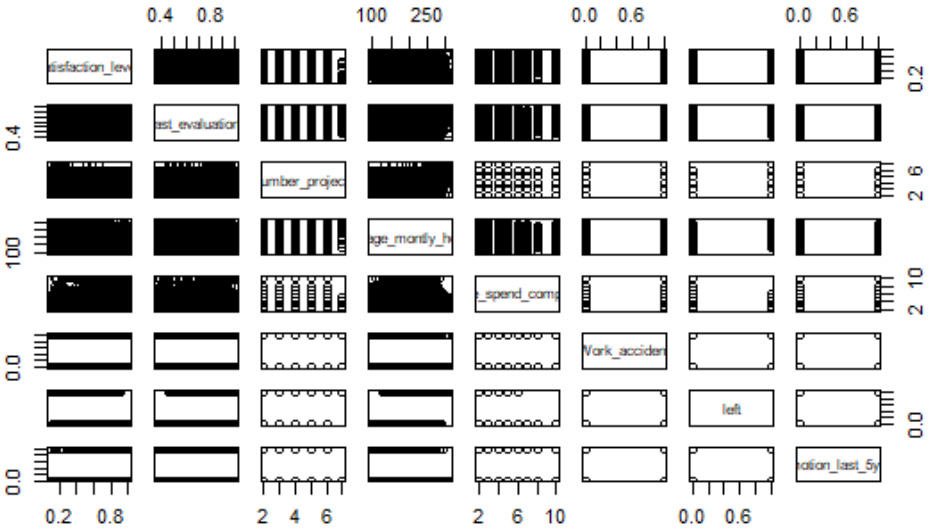


Only time_spent_company is right skewed and applying log transformation also didn't reduce the skewness in data.

BIVARIATE ANALYSIS:

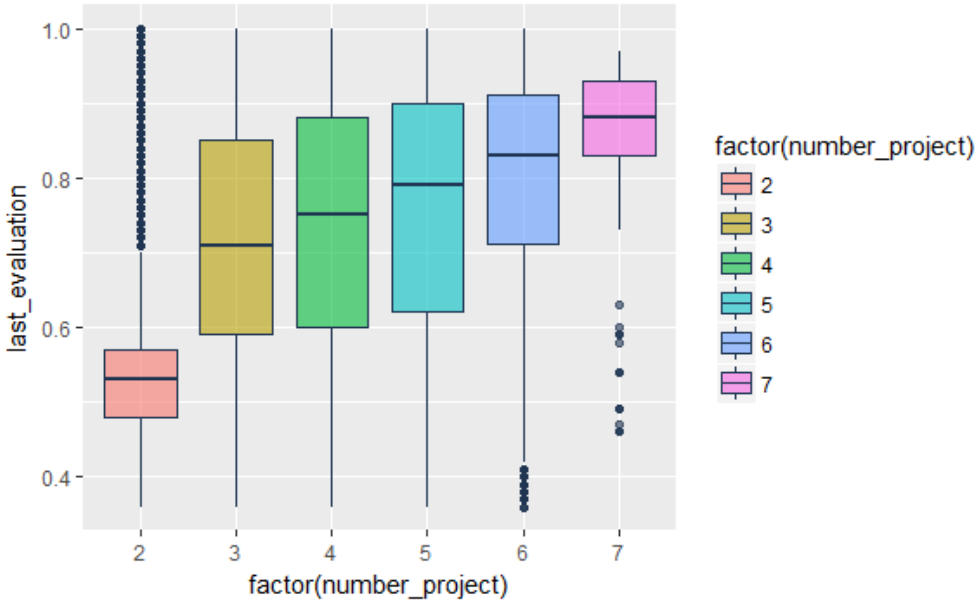
	satisfaction_level	last_evaluation	number_project
satisfaction_level	1.00000000	0.105021214	-0.142969586
last_evaluation	0.10502121	1.000000000	0.349332589
number_project	-0.14296959	0.349332589	1.000000000
average_montly_hours	-0.02004811	0.339741800	0.417210634

```
time_spend_company          -0.10086607     0.131590722    0.196785891
Work_accident               0.05869724    -0.007104289   -0.004740548
left                       -0.38837498     0.006567120    0.023787185
promotion_last_5years       0.02560519    -0.008683768   -0.006063958
                     average_montly_hours time_spend_company Work_accident       left
satisfaction_level           -0.020048113      -0.100866073   0.058697241 -0.38837498
last_evaluation               0.339741800       0.131590722  -0.007104289  0.00656712
number_project                0.417210634       0.196785891  -0.004740548  0.02378719
average_montly_hours          1.000000000       0.127754910  -0.010142888  0.07128718
time_spend_company            0.127754910       1.000000000   0.002120418  0.14482217
Work_accident                -0.010142888       0.002120418   1.000000000 -0.15462163
left                          0.071287179       0.144822175  -0.154621634  1.00000000
promotion_last_5years        -0.003544414       0.067432925   0.039245435 -0.06178811
                     promotion_last_5years
satisfaction_level           0.025605186
last_evaluation             -0.008683768
number_project              -0.006063958
average_montly_hours        -0.003544414
time_spend_company           0.067432925
Work_accident                0.039245435
left                        -0.061788107
promotion_last_5years        1.000000000
```



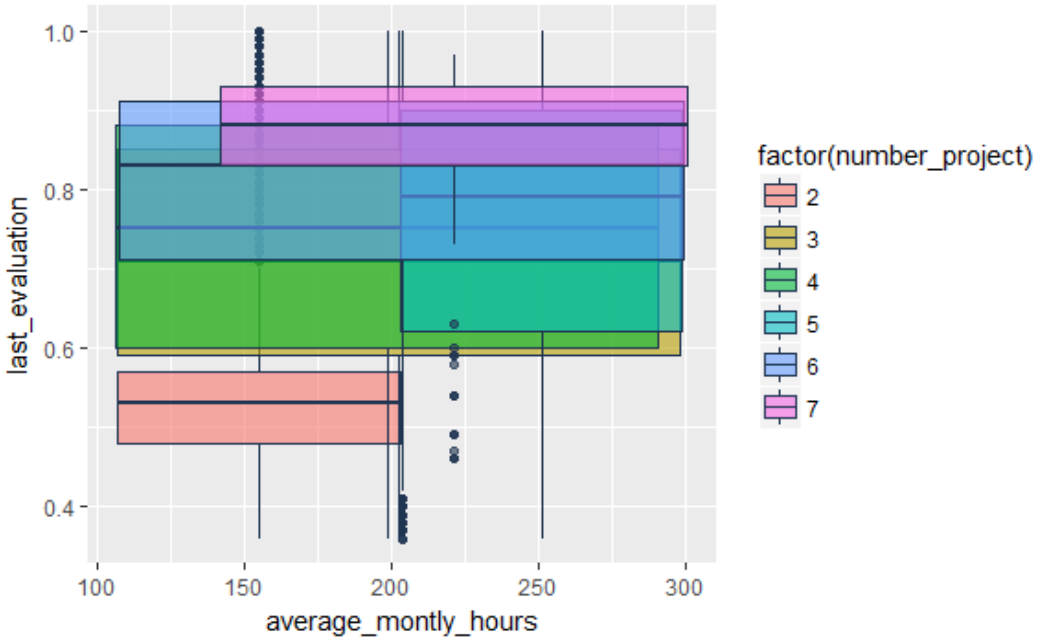
With this plot the variables with considerable correlation: LAST EVALUATION: NUMBER OF PROJECTS.

1. NUMBER OF PROJECTS---LAST EVALAUATION



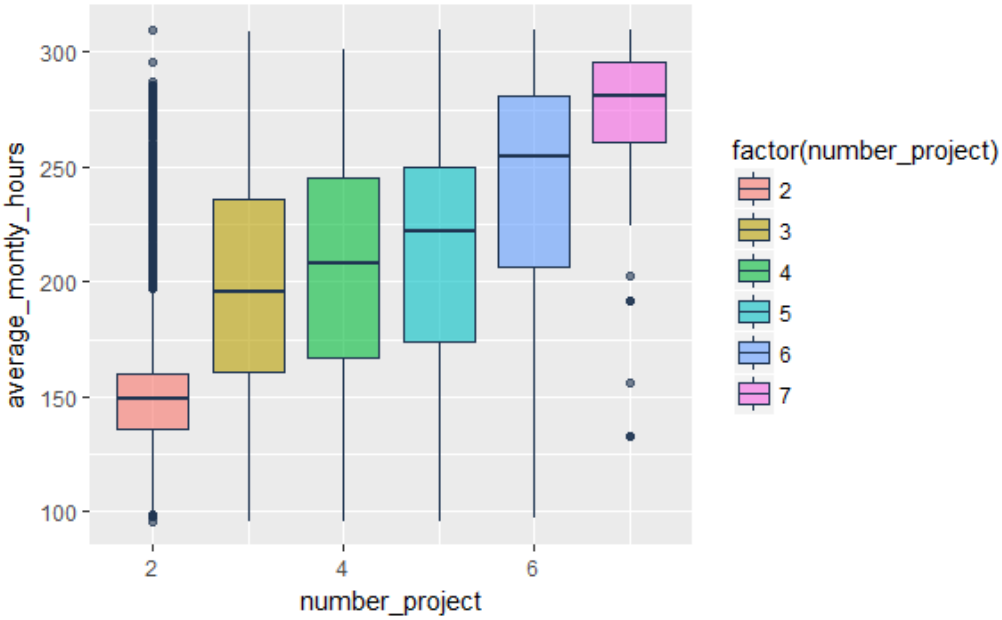
In this plot median value of the last evaluation level increase with the increasing in the number of projects.

2. AVERAGE MONTHLY HOURS---LAST EVALUATION

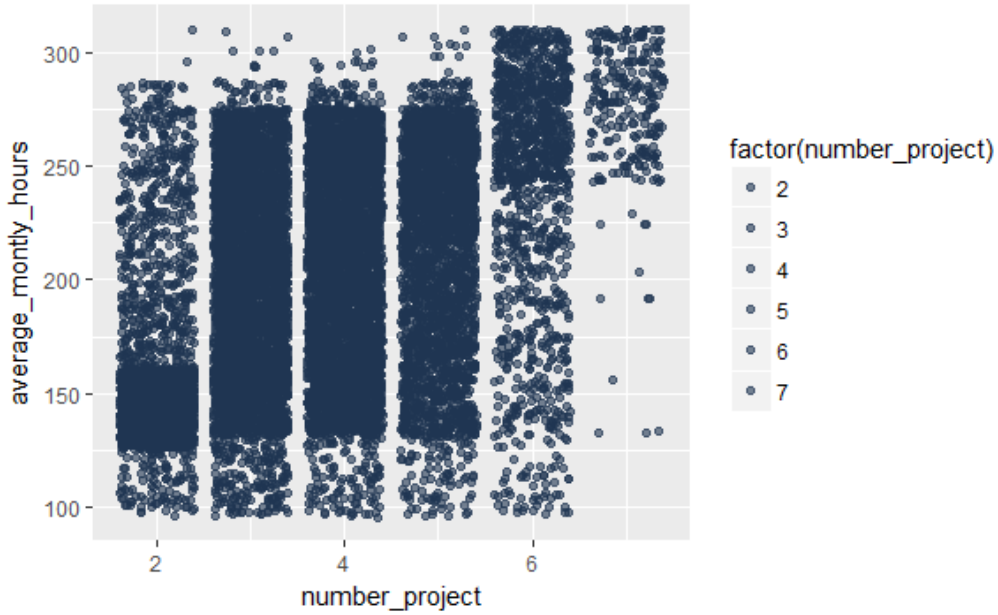


In this plot huge data in the left represents the low evaluation level correlated to the low number of hours and huge amount of data at top right represents high evaluation relating to high number of hours. Other data is scattered and hence we cannot notice any pattern between two variables.

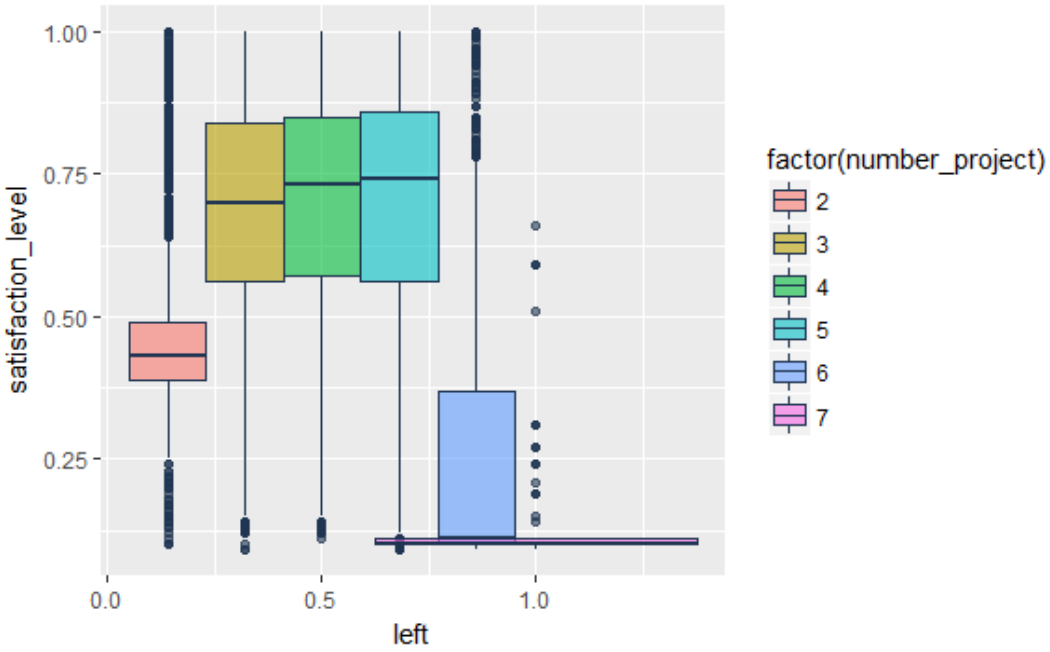
3. NUMBER OF PROJECTS----AVERAGE MONTHLY HOURS

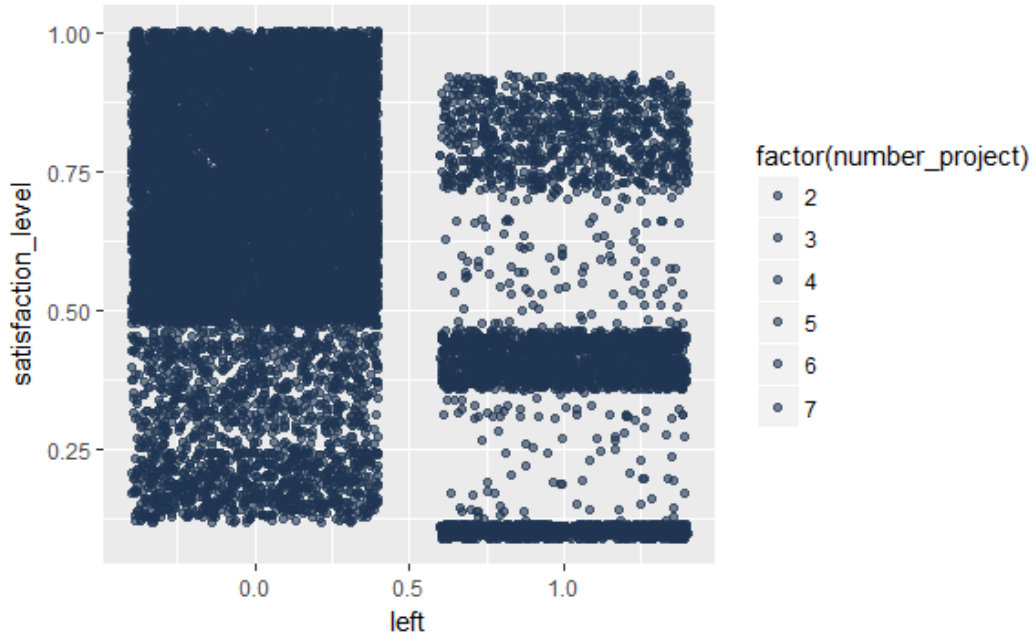


As the number of hours for employees is highly correlated with the number of projects in particular with 2,6,7 projects.



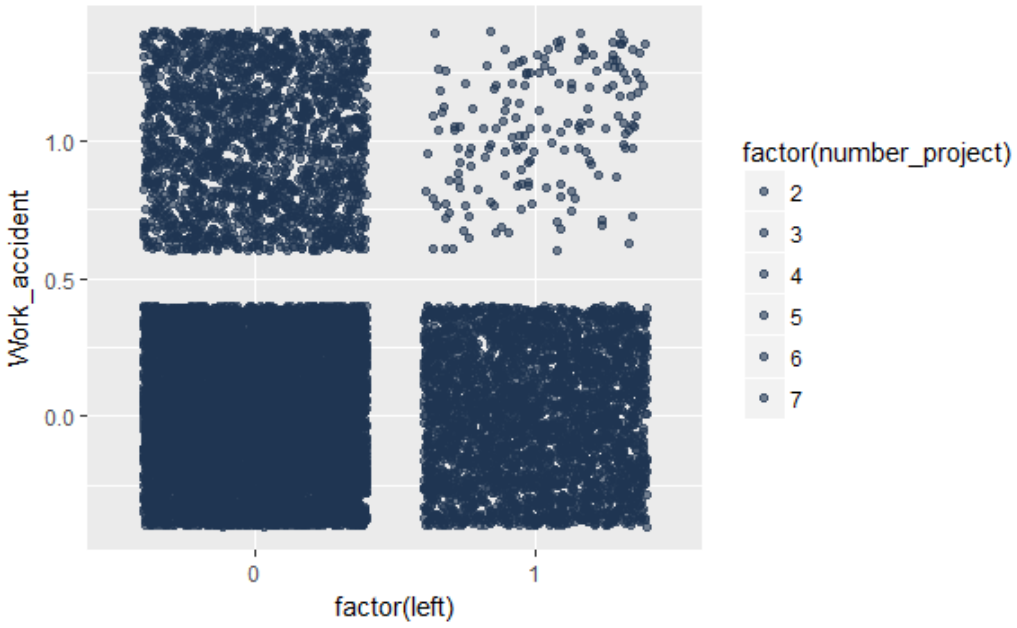
4. LEFT----SATISFACTION LEVEL





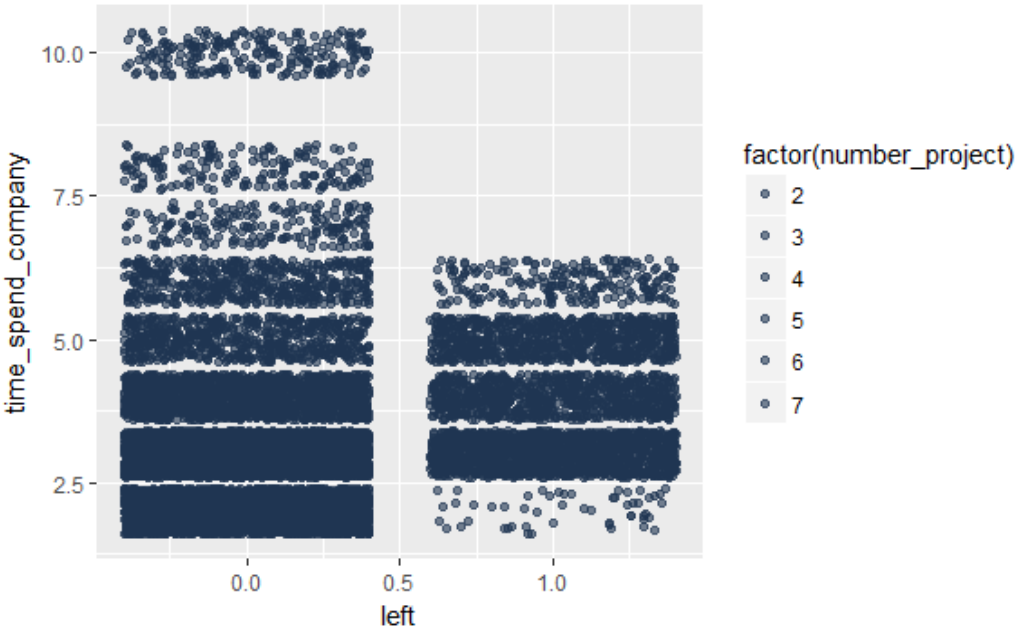
The first comparison between satisfaction level with the left of an employee (0 = No, 1 = Yes). This scatter plot is more readable than a box plot as there is no pattern in it., In the scatter plot, instead, we can notice that the people who stay in the company have a medium/high Satisfaction level, while the people who leaves the company, can be classified in 3 classes: the first are the employee with a low satisfaction level that probably change company in order to find a company that stimulate they in a better way; the second class is composed with the people under the 0.5 in the satisfaction level and the third class composed by the employee with an high satisfaction level may leave to get better career growth.

5. ACCIDENT-LEFT

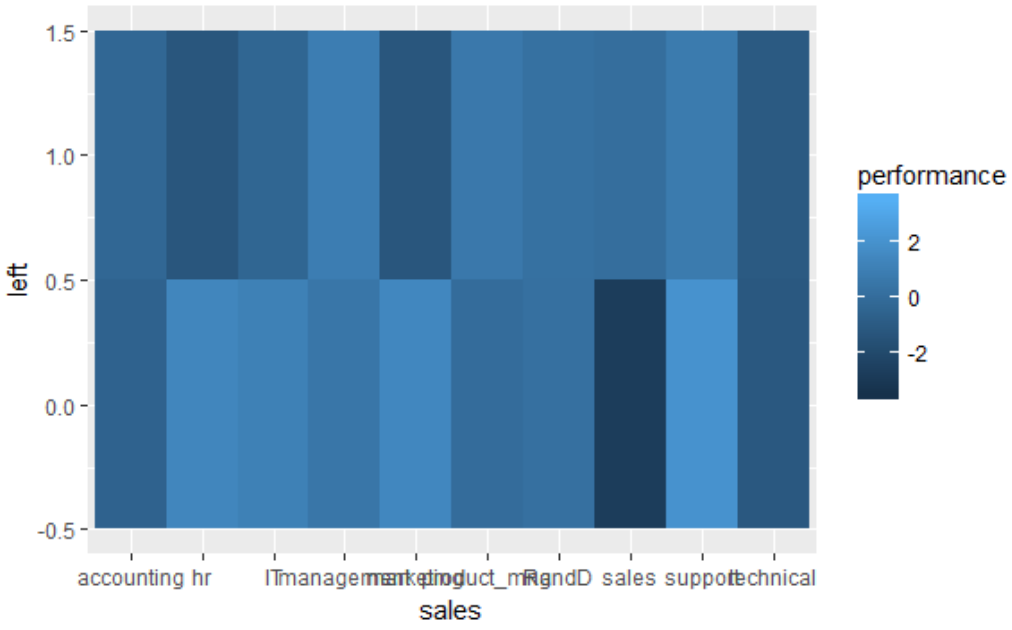


In this plot people don't leave company because they have an accident; there are few points in the top right box that shows this isn't the way to better understand the cause of the leave.

6. LEFT---TIME-SPEND IN THE COMPANY

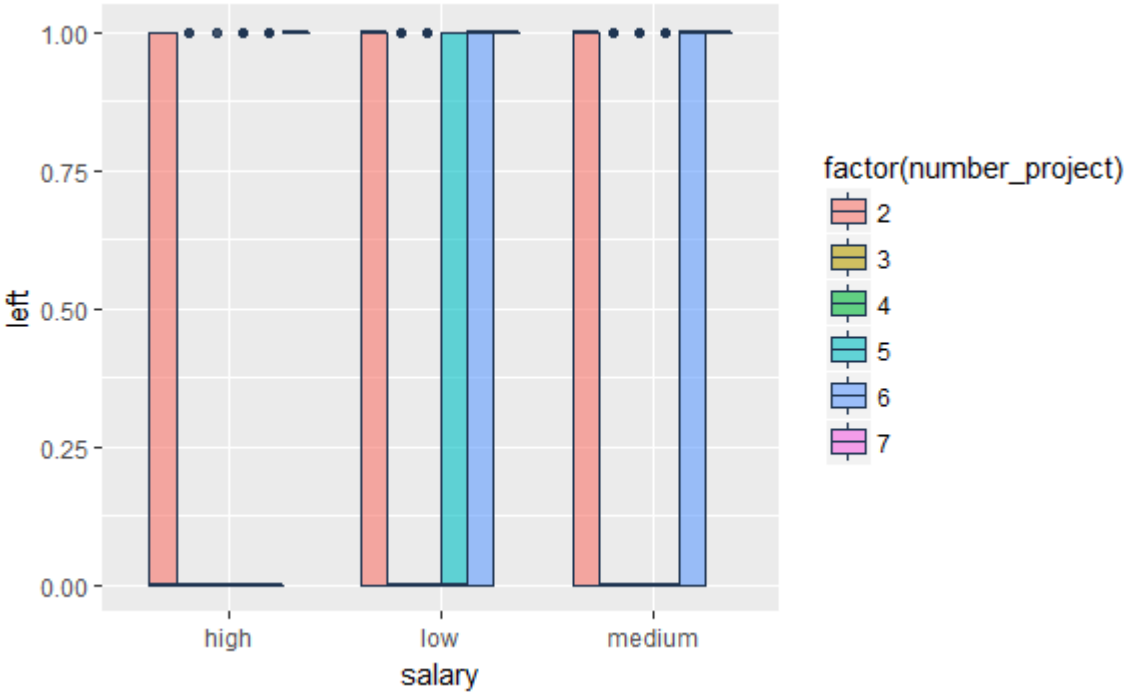


The higher percentage of employee who are leaving are the people who worked 4,5,6(years) from the plot.



The percentage of employee who leave the company is the same value for most of the department, with the exception of management and R and D.

7. SALARY--LEFT



Employees leave for every salary department, are the low salary's employees, followed by the medium salary's and the high salary's

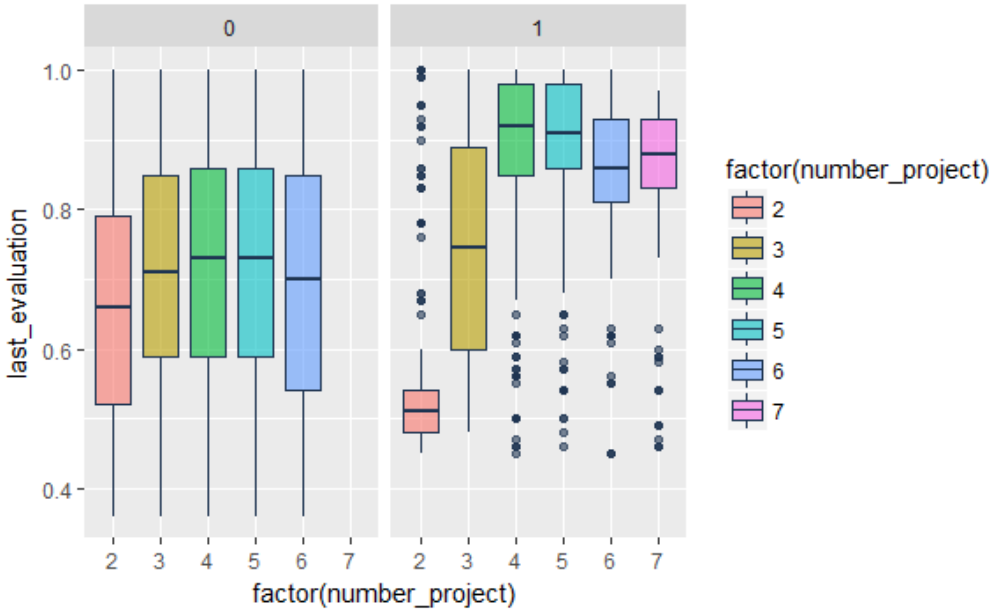
Bivariate Analysis Inference

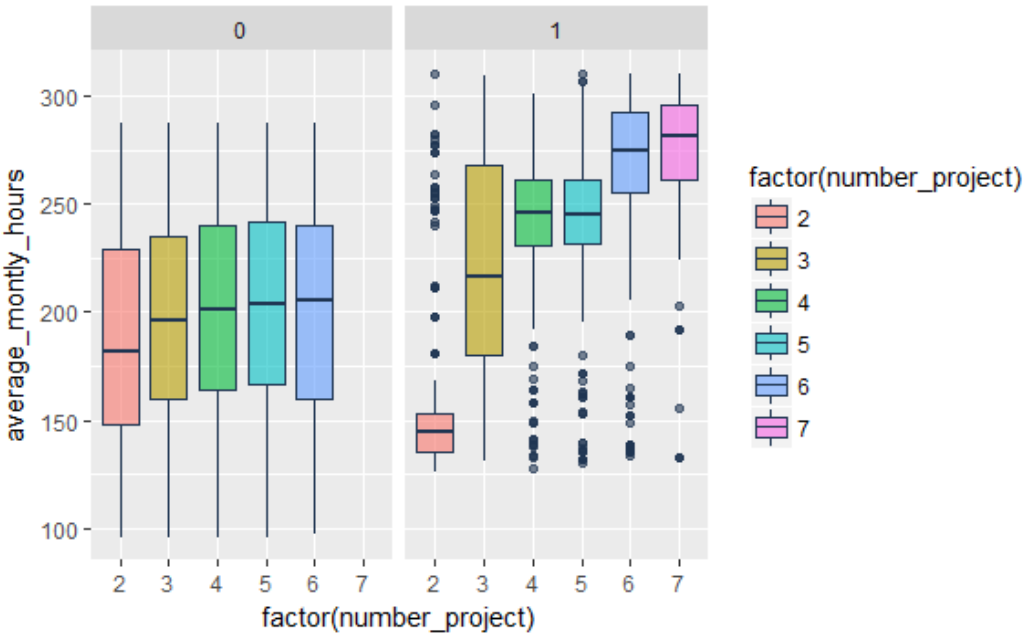
In this plot's section majorly the left variable compared with other variables. The left variable is strictly correlated with satisfaction level and lower is the level of satisfaction, higher is the number of people that left the company. The last evaluation level, instead, is correlated with

the number of projects (if number of projects is higher, than the last evaluation score is higher) and the average monthly hours (the hard workers are awarded with a high last evaluation score).

Multivariate Plots Analysis

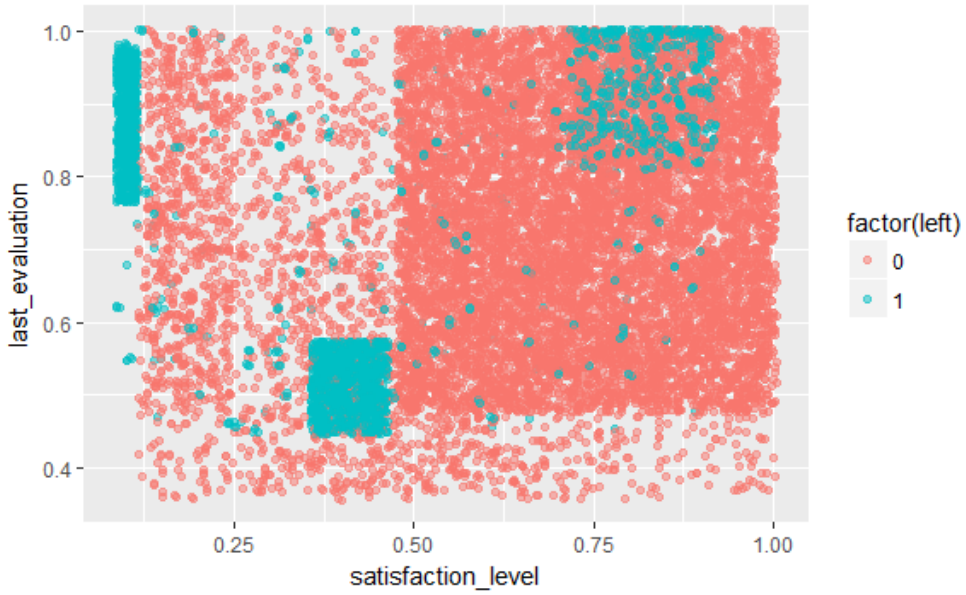
In this section to analyse four plots that describes very much this dataset. The first to create is the last evaluation on number of projects with the left variable.





With this plot workers, with a high number of average hours, tend to leave the company and the same for the people that are less engaged. Instead, the employees that stay in company, have around the same monthly hours.

LEFT ON SATISFACTON LEVEL FOR LAST EVALUATION LEVEL



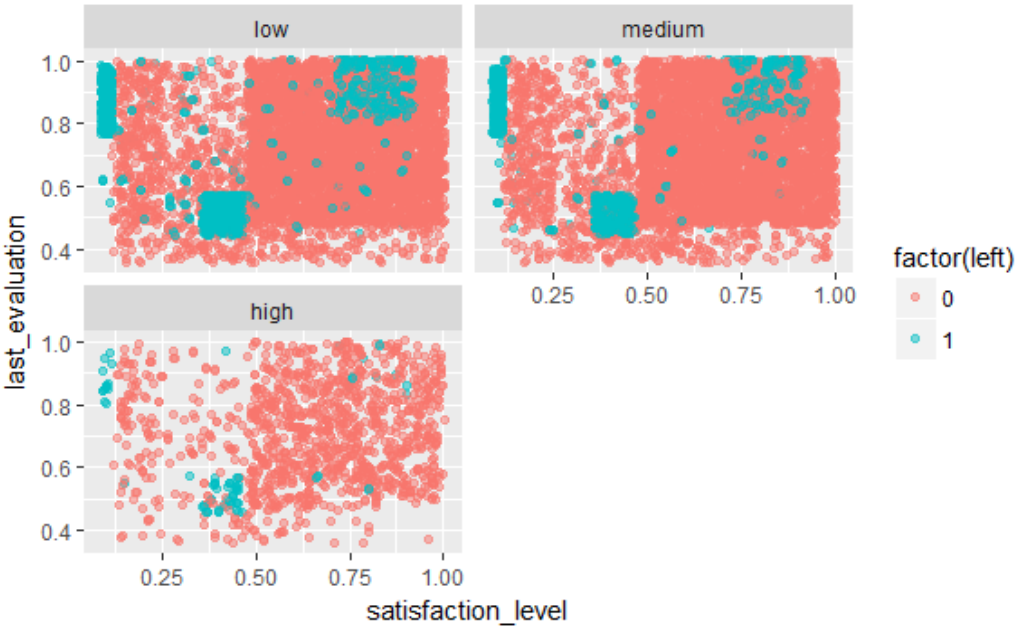
This plot doesn't explain the cause why employee leave as there are three types of people that leave the company: the first one are the employee that have an high evaluation score with a low satisfaction level, the second are the employee that have a low evaluation level and a low satisfaction level and the third one are the employee with an high satisfaction level and an high evaluation level.

LEFT ON SATISFACTON LEVEL FOR LAST EVALUATION LEVEL FOR NUMBER OF PROJECTS

The people leaving with the less number of project, two, have a low satisfaction level and a low evaluation level. In the employee that follow three project there are only few data of people who left and without a specific pattern. For the employee with four and five projects, the people who left the company have a high satisfaction level and a high evaluation level. Last but not least, the employee with ix or seven projects that left have a low satisfaction level, and

high last evaluation level. The number of project could be an interesting variable to better understand the future left.

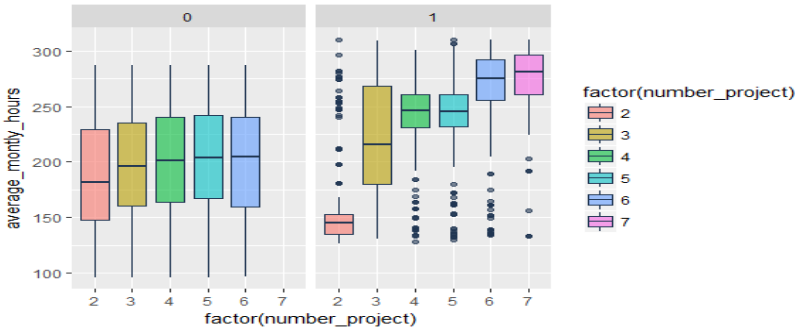
LEFT ON SATISFACTON LEVEL FOR LAST EVALUATION FOR SALARY



Here in this plot its great plot to explain the cause why employee leave.

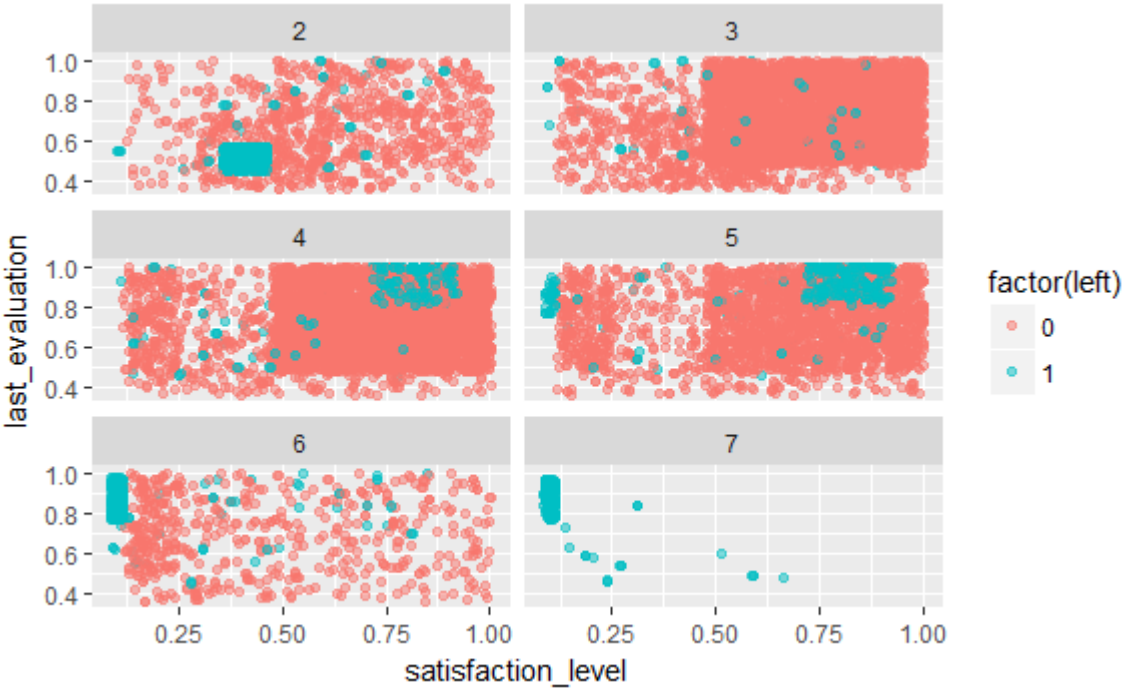
FINAL PLOT:

AVERAGE MONTHLY HOURS---------NUMBER OF PROJECTS-----PLOT1



In this plot, the people who stay in the company, have similar average monthly hours independently of The number of projects. These employees probably doing the job well, but not perfectly. Instead the employee who left the company, for the same number of projects are hard workers and spend more time on the projects probably to do better; there is only one exception, because the people who have left the company with two projects, are not good workers, with time spent monthly very above of the average monthly hours. In this plot the people with a hard work ethic, probably in the future will leave their job in order to search an advance in their careers.

NUMBER OF HOURS---SATISFACTION LEVEL---LAST EVALUATION LEVEL—PLOT2



In these plots the employees that leave with a number of projects in their hands equal to two, have a low satisfaction level and a low last evaluation level, probably because they are not good workers and they are less engaged than the others employees. Instead, there is a very low number of employees that leave the company with three projects in their hands. The workers that leave the company, with four and five number of projects, have high satisfaction level and a high evaluation level; probably they are changing their job to advance in their career. In the last, the workers that leave with high number of projects (6 or 7) have an high evaluation level, because are good workers, but a very low satisfaction level, probably because they are overwhelmed by the job; The only workers that have 7 projects on their hand, all leave the company. These plots help to better understand what are the causes of a low satisfaction level and consequently the left of the company.

Correlation matrix:

Correlation matrix and correlation plot below depicts that there is no strong correlation between the quantitative/numeric variables in the data and so collinearity measures may not be required unless the model requires the same during the logistic regression.

Correlation plot: