Human Resource Exploratory Analysis

Yun-Hsiang Chan

The human resource dataset is a simulated dataset from Kaggle. By exploring this dataset, we are able to extract good insights for problems that the Human Resource department deals daily. This project mainly focuses on explortary analysis.

```
## -- Attaching packages ---------------------------------------------------------------------------------
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## -- Conflicts ------------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## Parsed with column specification:
## cols(
##   satisfaction_level = col_double(),
##   last_evaluation = col_double(),
##   number_project = col_double(),
##   average_montly_hours = col_double(),
##   time_spend_company = col_double(),
##   Work_accident = col_double(),
##   left = col_double(),
##   promotion_last_5years = col_double(),
##   sales = col_character(),
##   salary = col_character()
## )
```

Data Cleaning

variables	descriptions
satisfaction_level	Satisfaction Level
last_evaluation	Last evaluation
number_project	Number of projects
average_montly_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

The descpitions of the variables are above. I have done some data cleaning before the analysis.

1. The variable name "sales"" is not the correct name for the variable. Rename the variable as "department".
2. There is a typo in the variable name for average monthly hours. Correct it.
3. Since the descrpition doesn't include the more details about 'promotion_last_5years', I don't think they will be useful in my analysis. So I decide to take it out from the dataset.
4. To further investigate the impact of salary, create a new variable called salary_factor. The number in this variable corresponds to the levels in variable salary.

This is the fundamental structure of the dataset after the data cleaning.

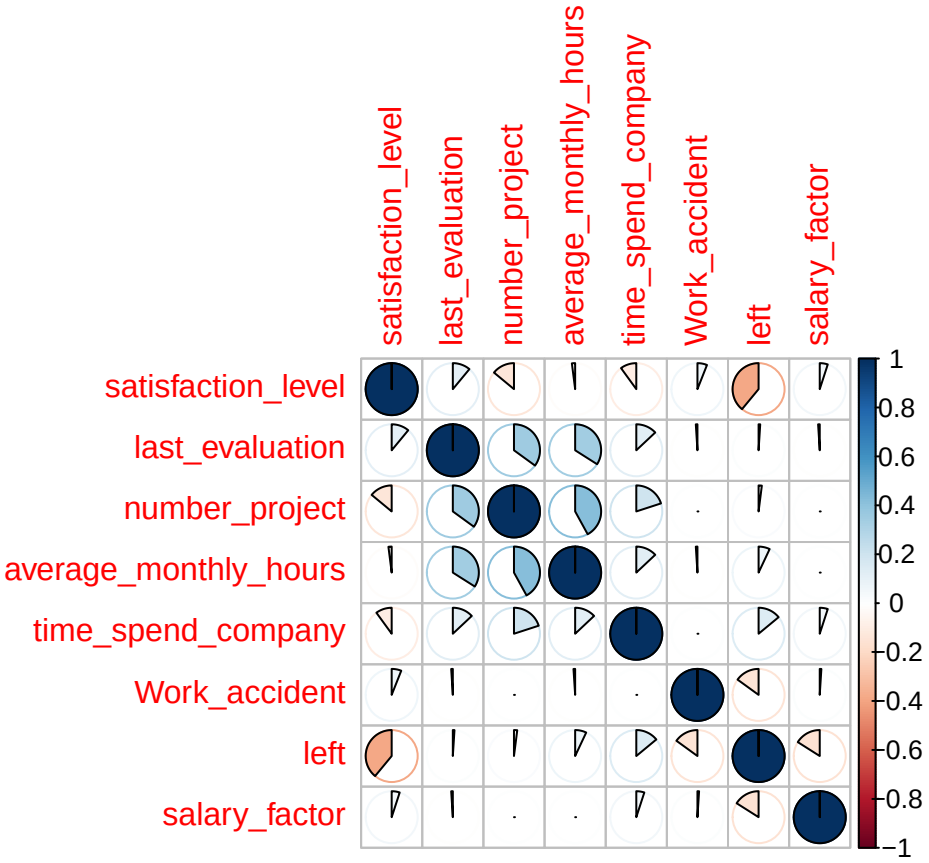
```
human_resource[sample(nrow(human_resource), 10), ]
```

```
## # A tibble: 10 x 10
##    satisfaction_le~ last_evaluation number_project average_monthly~
##               <dbl>           <dbl>          <dbl>             <dbl>
##  1             0.78            0.71              4               296
##  2             0.99            0.68              4               190
##  3             0.97            0.89              3               264
##  4             0.91            0.98              5               135
##  5             0.62            0.68              3               124
##  6             0.8             0.95              4               272
##  7             0.5             0.91              3               240
##  8             0.21            0.580             7               203
##  9             0.86            0.7               5               160
## 10             0.92            0.69              3               139
## # ... with 6 more variables: time_spend_company <dbl>, Work_accident <dbl>,
## #   left <dbl>, department <chr>, salary <chr>, salary_factor <dbl>
```

Correlation Analysis

To gain some insights of the dataset,I would like to see whether there is any relationship between the variables. The tool I use is correlation matrix.

```
## corrplot 0.84 loaded
```



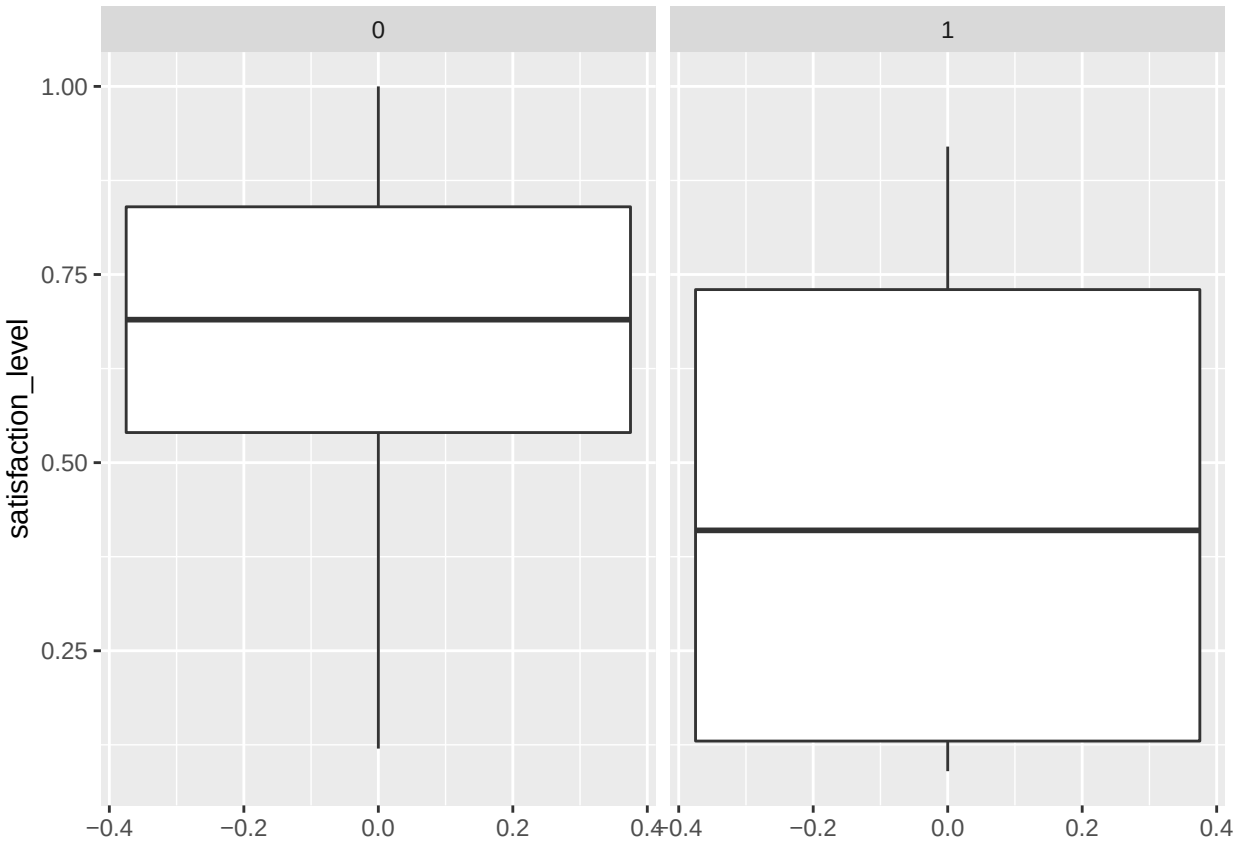
From this correlation plot we can know that:

1. The highest positive correlation is 0.42 (between **number_project** and **average_monthly_hours**), the highest negative correlation is -0.39 (between **left** and **satisfaction_level**).
2. The remaining correlations greater than 0.3 is 0.35 (between number_project and last_evaluation) and 0.34 (between last_evaluation and average_monthly hours). Both of them are related to **last_evaluation**.
3. It is noteworthy that left has low negative correlation with both work_accident and salary_factor, i.e the number of work accidents and salary level has a negative influence to employees' decisions of leaving.
4. We cannot ignore that time spend company is postively correlated to number of project, monthly working hours and left. In addition, it is negatively correlated with satisfaction level.

Question: Why employees leave the company?

We discover three factors (satisfaction level, salary, and work accident) having negative correlations with left and one factor (time spend company) having positive correlation with left. Well, are they the main reason for employees to leave the company?

Let's first investigate the satisfaction level.



Without doubt, the boxplot implies that unsatisfaction is one of the main reasons for employees to leave.

Then we can look at the salary.

By comparing the barplots, we can know that among employees who didn't leave, 7.7% of them receive high salary. But among all employees who leave, only 2.3% of them receive high salary and the proportion of employees receiving low salary increaes by 28% (from 32.3% to 60.8%). Therefore, it is highly likely that salary is one of the main reasons for employees to leave.

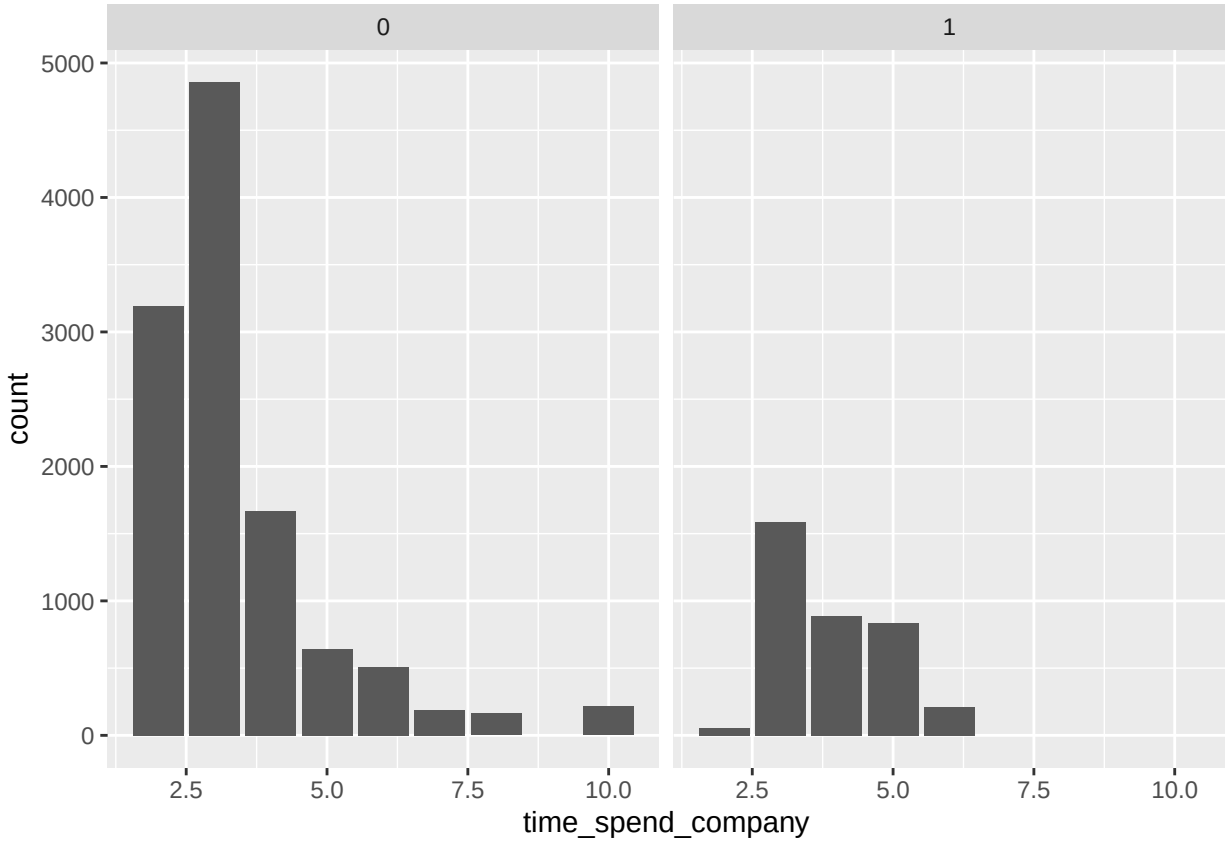
Next step, we test if work accident is related to leave of the employees.

left	n_employee	work_accident_prop
0	11428	0.1750088

left	n_employee	work_accident_prop
1	3571	0.0473257

It is obvious that the proportion of people who suffer work accident and leave is much lower than the proportion of people who suffer work accident and stay. Hence, it is unlikely that work accident is one of the main reason for empoyees to leave.

Lastly, let's check if people who stay longer are more likely to leave.

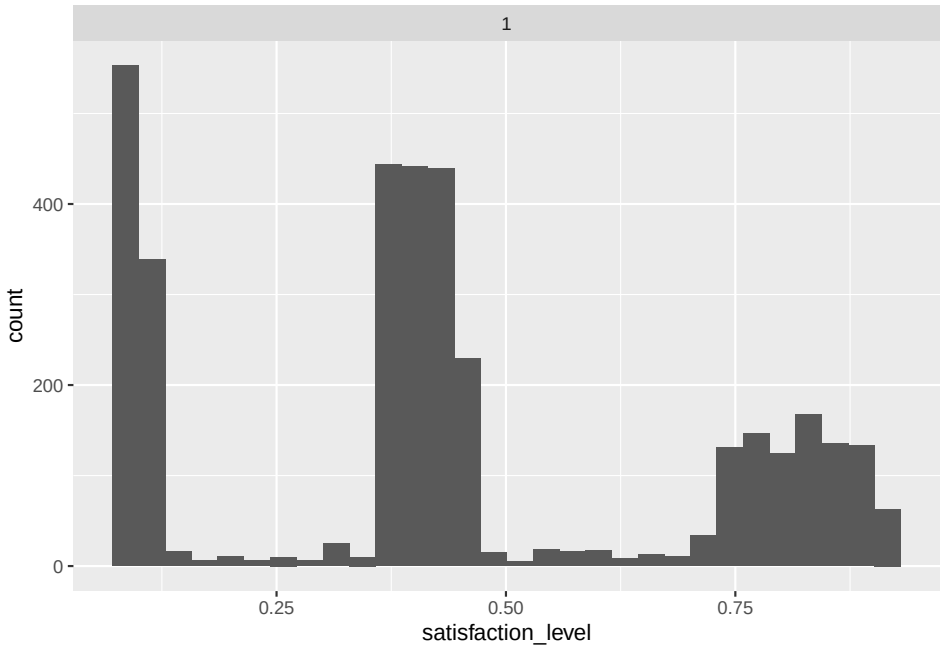



```
## # A tibble: 2 x 2
##    left time_stay_company
##   <dbl>             <dbl>
## 1     0              3.38
## 2     1              3.88
```

Although the distributions of time stay in the firm are quite different between people who leave and people who stay, the average time stay in the company are similar between these two groups of people. So it is unlikely to say the time spend in the company is a reason for employees to leave.

In conclusion, we believe that the employees leave mainly due to unsatisfaction and low salary.

Question: Performance difference between two groups of people who choose to leave.



There is one more phenomenon we need to notice. If we visualize the histogram of satisfaction level of employees who left, we can see three modes. It is understandable that people with low satisfaction and medium satisfaction level choose to leave. **But why many employees left even if they were highly satisfied?**

Is it because of the salary or working hours? Should we retain those people?

Now we split the people who left into two groups. The first group is the people who have less than 0.7 satisfaction level. The second group is the people who have more than 0.7 satisfaction level. We call them low satisfaction group and high satisfaction group.

group	mean_accident_prop	low_salary_prop	medium_salary_prop	high_salary_prop
low satisfaction group	4.68	60.84	36.61	2.55
high satisfaction group	4.88	60.76	37.65	1.59

group	average_last_eva	average_num_project	average_hours	average_time_spend_company
low satisfaction group	0.65	3.61	194.70	3.44
high satisfaction group	0.91	4.53	242.86	5.08

We summarize some statisitcs for these two groups and find some facts.

1. It seems that the accident proportion and low, medium, high salary proportion are similar in these two groups, which means **work accident and the salary is not the reason why high satisfaction group choose to leave.**
2. The average last evaluation score, average number of projects, average working hours and average time stay in the company are different between two groups. In the correlation analysis, we have already known that the first three factors are correlated. The logic behind it is, as you work for more hours,

you will be able to work on more projects and get a higher evaluation score. Besides, people who stay in the company for longer time can work on more projects.

By these facts, we can know that **highly satisfied employees are valuable to the firm**. They are more familiar with the work, generally work harder, accumulate more experience and get a higher evaluation score.

So why they left? I think most of them left for getting promoted. Because of their recent outstanding performances, other firms would try to cut corner and provide better offers.

Therefore, they worth us to retain. In order to avoid losing these outstanding workers, I would recommend the company promotes the employees who stay for a long time, work hard and recieve a high evaluation score recently.

Question: How to retain expereienced employees.

Last part, we mention that most of high statisfaction workers who left perform pretty well. In this part, I will further investigate the condition of experienced employees, try to explain why they choose to leave and what can we do in order to retain them.

First, we need to define experienced employees. Only employees who accmpolish more than 2 projects and stay for more than 2 years can be recorded in this dataset. On the basis of that, I set three stricter standards.

1. Stay in the company for more than 3 years.
2. Accomplish more than 4 projects.
3. The last evaluation score must be greater than half of all employees.

These numbers are selected out of the quantiles of the variables.

```
##   0%  25%  50%  75% 100%
##    2    3    4    5    7
```

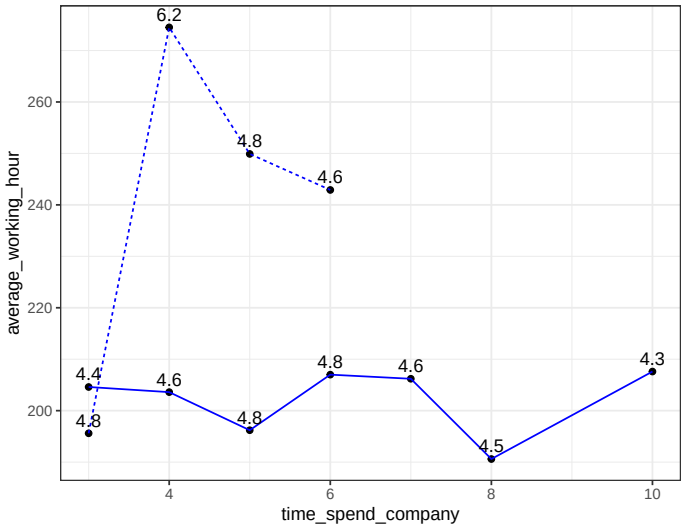
```
##   0%  25%  50%  75% 100%
##    2    3    3    4   10
```

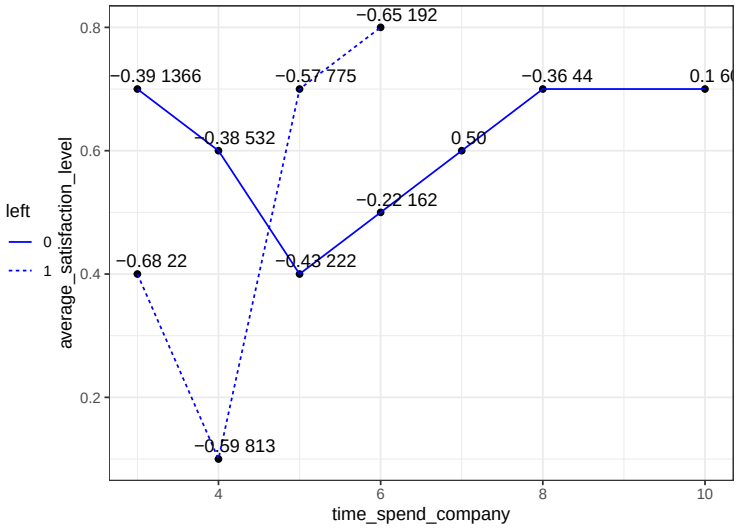
```
##   0%  25%  50%  75% 100%
## 0.36 0.56 0.72 0.87 1.00
```

To analyze the performance, I group these employees based on whether they leave and the time they stay in the company.

I create two time series plots:

1. The first plot is about average working hours, average number of project and the status (leave or stay).
2. The second plot is about average satisfaction level, average salary, the number of workers and the status (leave or stay).





In the first plot we get several facts:
1. The average working hours of people who left skyrockets in the forth year. The average number of projects increases drastically in the meanwhile.
2. After the forth year, the average working hour and number of projects of people who left slide down, but the average working hour of them is still much higher than the working hour of those who stay.
3. After the sixth year, no experienced employees left.

In the second plot we also get some information:
1. In the third year, not much experienced employees want to leave the company. Those who leave in the third year do that because of low salary.
2. In the forth year, the average satisfaction level of people who left plummets, a large amount of experienced employees choose to leave in this year. Those who choose to stay have relatively lower salary than those who choose to leave.
3. In the fifth and sixth year, although the average satisfaction level of people who left skyrocketted, many of them still choose to leave the company.
4. After the fifth year, even if the average salary of those who stay is still unstable, the average satisfaction level keeps on growing and maintain at 0.7 from the eighth year.

Now let's summarize the information we get and describe the overall situation.

My conclusion: Most of experienced employees left in the forth, fifth and sixth year are college graduate. They left for higher salary.

I claim this because the average working hour of these three groups are very high, while the average salary of them is very low. This is a typical phenomenon. Those graudate just commence their career in the company, it is hard for them to get high paid from the beginning. However, after spending 4 to 6 years for accumulating experience, they would expect a higher salary.

The employees who left in the fourth year are the most valuable employees, and I will describe them as 'first-class' experienced employees. They have the highest average working hour and number of projects, but their salary doesn't match their performance. This result leads to those employees' leave.

As for those who left in the fifth and sixth year, they are not as valuable as thoes who left in the forth year, and I will descibe them as 'second-class' experienced employees. These people also work hard, but they are not as hard-working and outstanding as those who left in the fourth year. They need more time to accmulate experience and therefore they are more satisifed even if the salary is low. They eventually choose to leave because they gain enough experience and other firms provide relativelty competitve offers. But they are still not as great as those who left in the forth year.

Recommendation

Based on the analysis above, I believe we need a new policy.

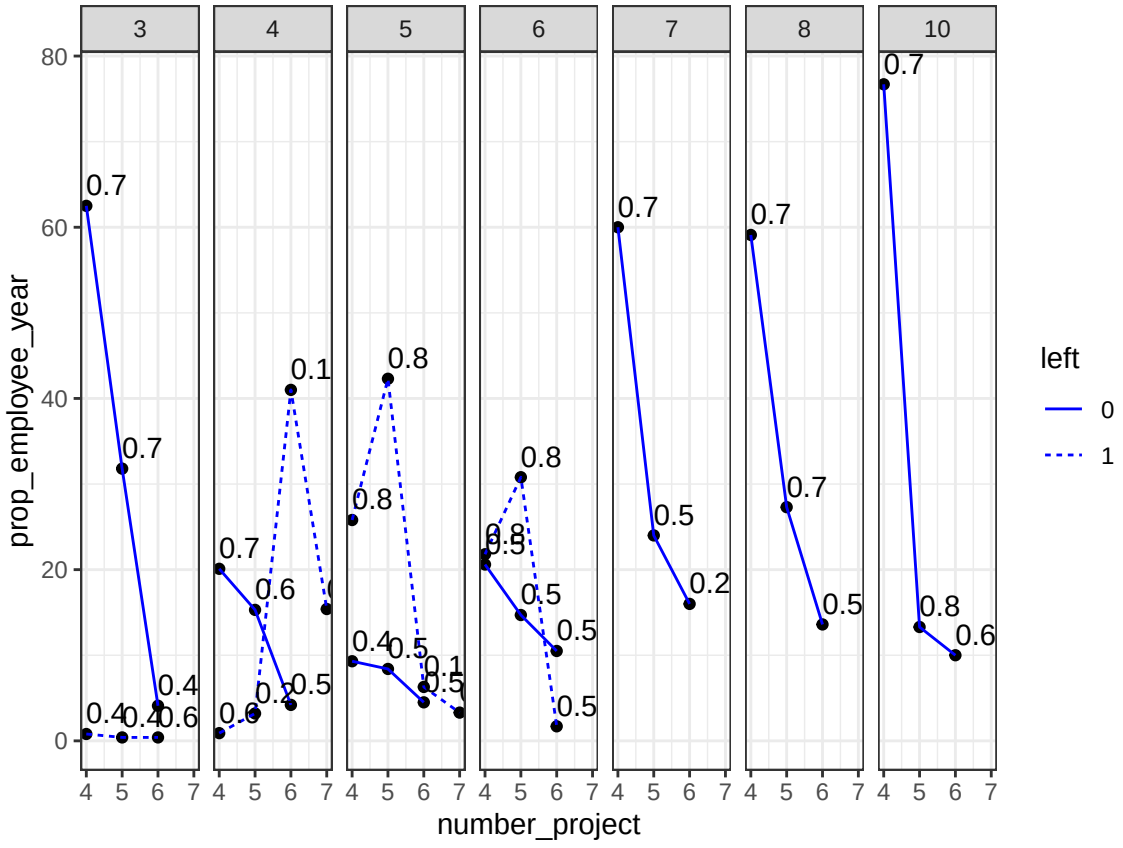
In order to retain those first-class and second-class experienced employees, we need to raise their salary before they leave. For the first-class experienced employees who work harder and complete more projects, we raise their salary at the forth year. And for the second-class experienced employees who are relatively less competitive than the first-class, we gradually raise their salary in the fifth and sixth year. The level of pay raise should be different between the first-class and the second-class, because the first-class deserves a higher pay.

However, we need to consider one more thing, the loyalty. The evaluation of pay raise should not only depend on the performance. It should be related to each individual employee's career plan and their loyalty. The additional money pay by the firm would become meaningless if we pay a person who will eventually leave the firm or who doesn't good at collaboration. The HR department needs to consider the invidiual situation thoroughly and provide a corresponding raise for each person.

In-depth Analysis by the number of projects

In this section, we further group them by the number of projects they do.

I create a series of plots based on these new groups. The plots are about the proportion of employees leave/stay in a specific year. (i.e if the total number of employee that spend 3 years in the company is 1000, and the number of people who choose to left in the third year with 4 project completed is 200, the proportion is 0.2.) The number next to the point is the average satisfaction level of this group of employees.



Now we can set more details for the policy.

1. What is the the best criteria to raise the salary for first-class experienced employees?

The best criteria is 3-to-4-year experience and 5 projects. After the forth-year, many of the experienced employees completed 6 projects are unsatisfied with the company and choose to leave. So the best solution is raising their salary in advance, as soon as they completed 5 projects.

2. What is the best criteria to raise the salary for second-class experienced employees?

In addition, the best criteria is 5-year experience and 5 projects.

Compared to the previous one, it is much harder to decide the criteria for second-class experienced employees. I will explain my criteria.

We must realize that people who left in the fifth year with 4 and 5 projects were not leaving because of unsatisfaction. Actually, their satisfaction level is pretty high. We have already discuss this phenomenon and conlude that they leave mainly because of better offers coming from other firms.

Without doubt, these people are needed by the firms, but I believe no firm could stop the leave of all employees and we must balance the cost and the benefit. If we raise their pay in advance, when they complete 4 projects, they might still leave if any better offer is provided.

We need to spend the money on the right person. It is accpetable to lose part of the experienced employees, as long as we persuade the best groups to stay.

Therefore, we only award those who can complete 5 projects in 5 years. This group of people worths us to pay more.

Conclusion for all questions

- The employees leave mainly due to unsatisfaction and low salary.

- Many employees left even if they are satisifed with the firm. We believe it is because of better offers provided by other companys. Since those highly satisfied employees perform very well, it is better to raise their pay and promote them.
- Most of the employees who left are college graduate, it is hard for them to earn a lot at the beginning. So they are likely to leave after they gain enough experience.

- The best cirteria for raising the salary to retain the greatest experienced employees, those who left in the forth year, is 3-to-4-year experience and 5 projects.

- The best cirteria for raising the salary to retian those second-class experienced employees, those who left in the fifth or sixth year, is 5-year experience and 5 projects.

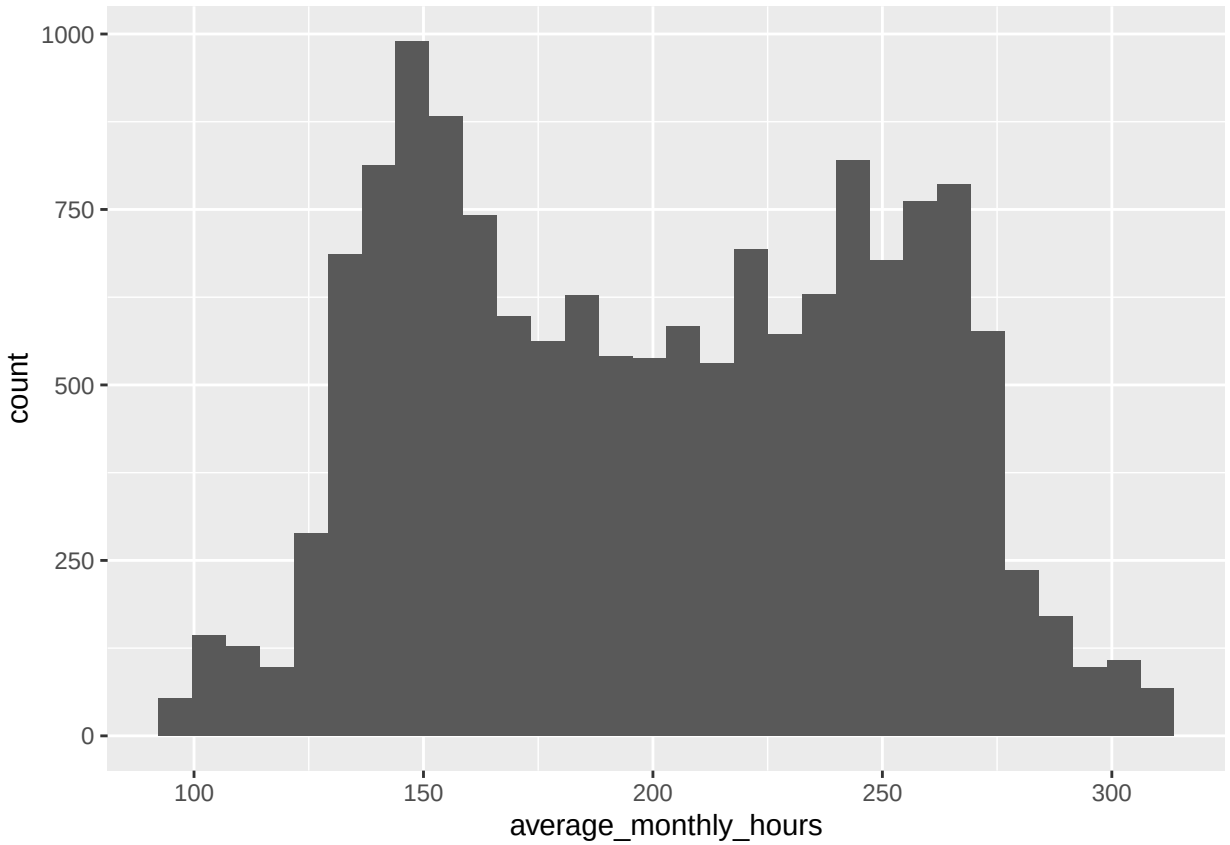
Appendix: Visualization of Employee and department information

Employee information

1. The salary structure of the company.

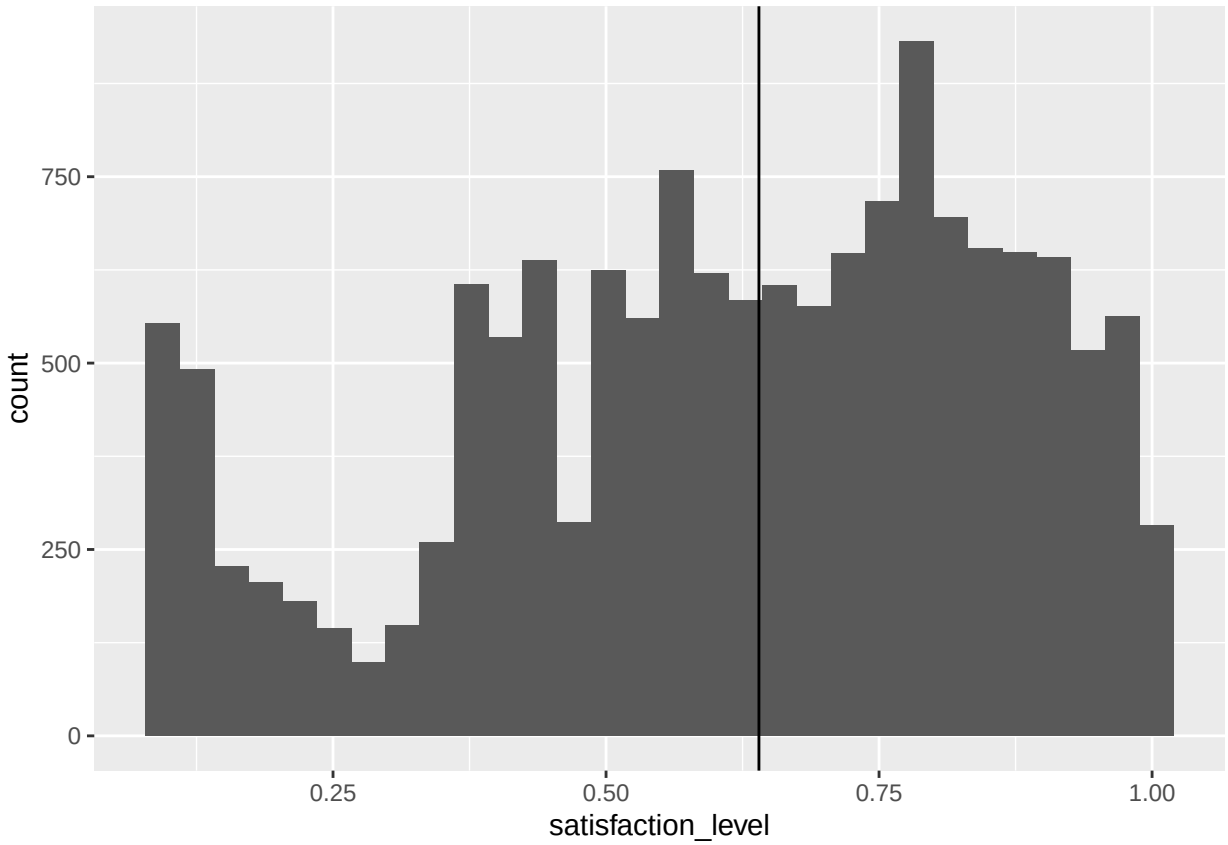
It is a very typical firm structure. Most of the employees receive low and medium level of salary. Only few people can receive high salary.

2. The condition of employees' monthly working hours.



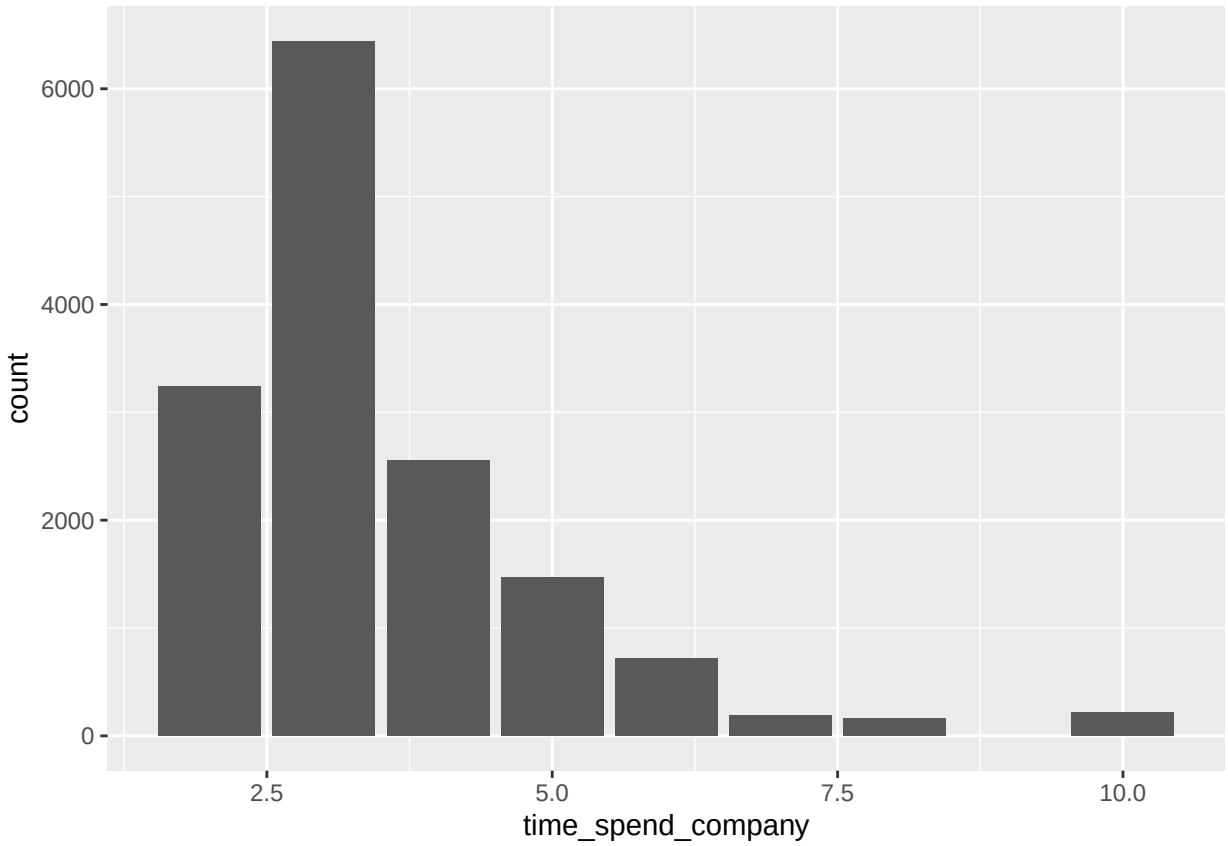
In this company, the median of the average monthly working hours is around 200. The distribution of monthly working hours is bimodal. One mode is around 150 hours and the other mode is around 260 hours.

3. The condition of employees' satisfaction level.



By histogram, we know that a small amount of people are extremely unsatisfied. Apart from them, the other part of the distribution is quite uniform. We don't know the reason of unsatisfaction, but we might find it in the latter part.

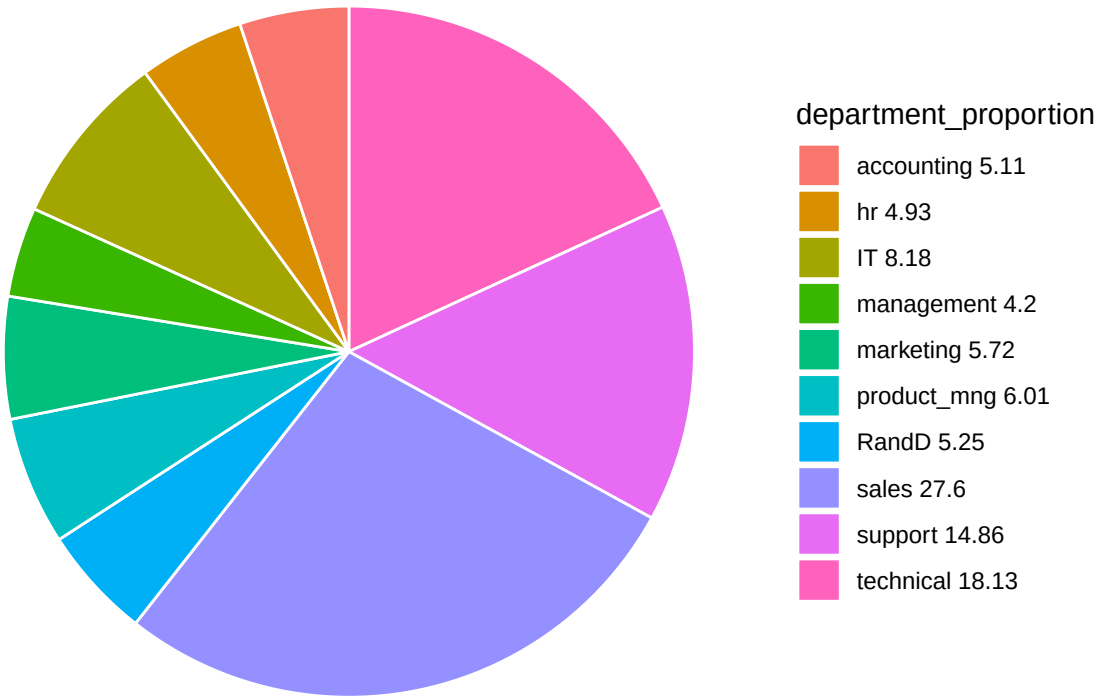
At last, we can check the distribution of the time employees stay in the company.



Most of the employees stay for less than 4 years. The most usual length of stay is 3 yeras. Only a few employees stay for more than 7 years. It indicates that this is a company open for less than 10 years.

Department Information

1. The size of the departments



From the pie chart, we can know that in this firm, the biggest department is sales department. Besides, most of the employees are in sales, support and technical department. Apart from these three departments, each of the remaining departments only has less than 10% of the employees.

2. Statistics for each department

department	mean_average_monthly_working_hours	mean_satisfaction_level	mean_last_evaluation_score	mea
accounting	201.16	0.58	0.72	
hr	198.68	0.60	0.71	
IT	202.22	0.62	0.72	
management	201.25	0.62	0.72	
marketing	199.39	0.62	0.72	
product_mng	199.97	0.62	0.71	
RandD	200.80	0.62	0.71	
sales	200.91	0.61	0.71	
support	200.76	0.62	0.72	
technical	202.50	0.61	0.72	

department	left_prop	salary_high_prop	salary_medium_prop	salary_low_prop
accounting	0.27	9.65	43.68	46.68
hr	0.29	6.09	48.58	45.33
IT	0.22	6.76	43.60	49.63
management	0.14	35.71	35.71	28.57
marketing	0.24	9.32	43.82	46.85
product_mng	0.22	7.54	42.46	50.00
RandD	0.15	6.48	47.27	46.25
sales	0.24	6.50	42.80	50.70
support	0.25	6.33	42.26	51.41
technical	0.26	7.39	42.17	50.44

To make it clear, I split the statistics into two tables.

The first table includes the variables indicating nothing important. These variables (mean average monthly working hours, mean satisfaction level, ...) are similar in each department.

The second table are the variables that show differences. Compared to other departments, management and RandD (an unknown department that dataset information doesn't explain, so I will skip the analysis on it) have relatively low leaving proportions. In addition, the management department has an extremely high salary proportion. This might be the main cause of the low leaving proportion in management department.