HR DATA ANALYSIS PROJECT

DATA MINING

Group 16

Maira Abid	Syed Usman Samee	Muhammad Shahrukh	Muhammad Asad
20100291	20100067	20100287	20100216

CHAPTER 1 - EXPLORATORY DATA ANALYSIS

Introduction

The management has provided us with employees' data in order to have meaningful visualizations and information which can help employees retention and facilitate managerial decisions.

Finding the missing values

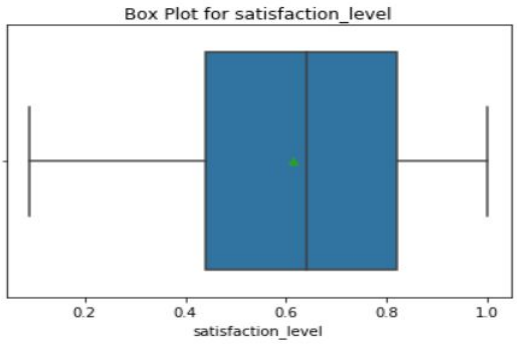
```
train_data.describe()
```

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	promotion_last_5years	left
count	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000
mean	0.612834	0.716102	3.803054	201.050337	3.498233	0.144610	0.021268	0.238083
std	0.248631	0.171169	1.232592	49.943099	1.460136	0.351719	0.144281	0.425924
min	0.090000	0.360000	2.000000	96.000000	2.000000	0.000000	0.000000	0.000000
25%	0.440000	0.560000	3.000000	156.000000	3.000000	0.000000	0.000000	0.000000
50%	0.640000	0.720000	4.000000	200.000000	3.000000	0.000000	0.000000	0.000000
75%	0.820000	0.870000	5.000000	245.000000	4.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	7.000000	310.000000	10.000000	1.000000	1.000000	1.000000

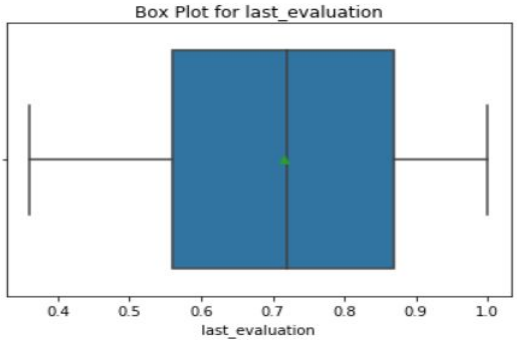
When loading the data and then outputting the summary of the data, we saw that none of the attributes had any missing values. Each attribute contained 14,999 entries.

Calculating and visualization of Summary Statistics

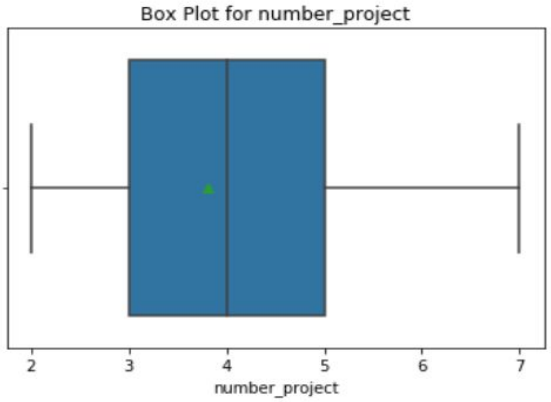
Using the seaborn library, we visualized the boxplots for all attributes except the ones with binary data values because visualizing the boxplots for binary attributes alone do not give significant information.



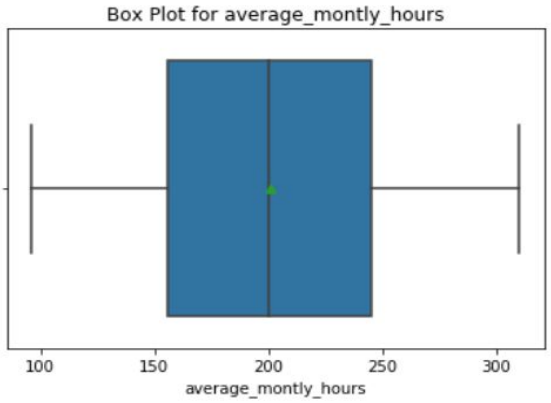
More than 75% of the employees have satisfaction levels above 0.4. The average satisfaction level of employees is nearly 0.6.



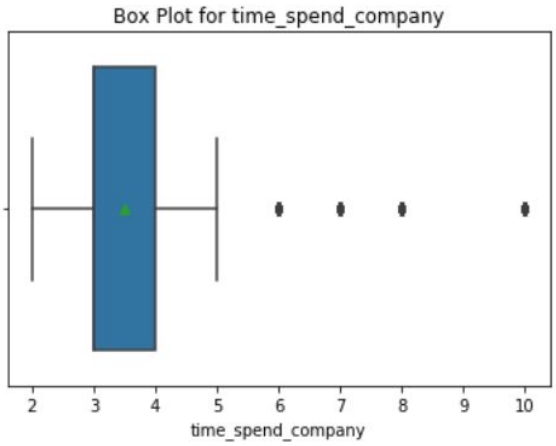
More than 75% of the employee scored above 0.55 in their last evaluation. The average evaluation score of employees is nearly 0.7.



Almost 75% of the employees worked for 3 and above projects. Almost 50% of the employees worked between 3 and 5 projects. The average number of projects per employee is almost 4.

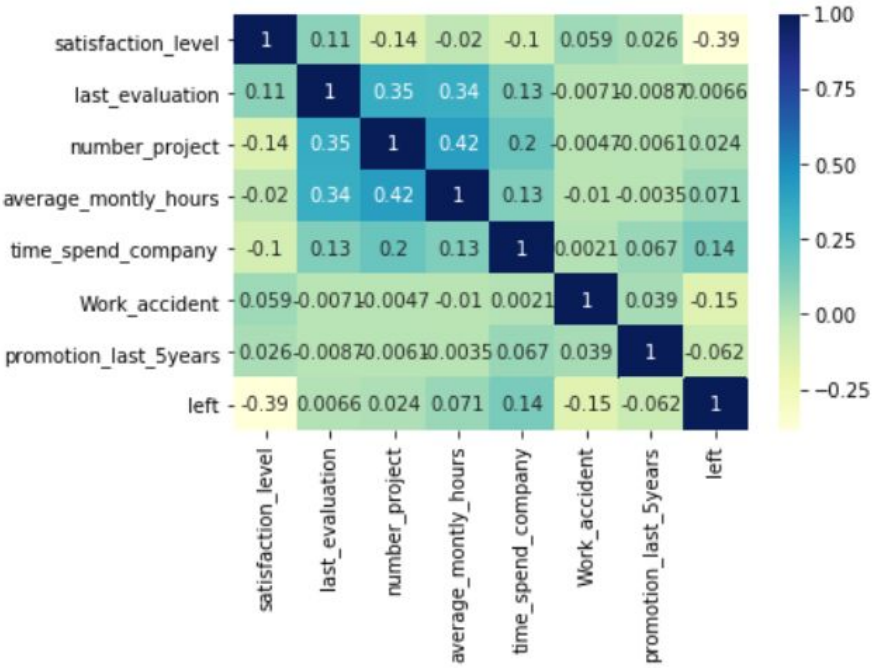


More than 75% of the employees spend above 150 hours monthly in the office. The average monthly number of hours spent by the employees is nearly 200.



Apart from 4 employees, everyone in the company has spent between 2 to 5 years with the company. Almost 50% of the employees spent 3 to 4 years with the company. The average is almost 3.5 years that the employees have spent with the company.

Correlation between attributes



Using the correlation matrix above, some of the interesting observations include:

1. There is somewhat a positive correlation (of 0.35) between last_evaluation and the number_project. This somewhat explains the company's way of evaluating its employees. The more you work, the probability you will be evaluated higher is more.

2. The satisfaction level is somewhat negatively correlated with the number of projects done by an employee.

3. There is somewhat a positive correlation (of 0.2) between the time spent in the company and the number of projects. It could mean the more experience you gain, the more likely the company is willing to trust you with projects.

4. There is somewhat a positive correlation (of 0.13) between the time spent in the company and the average monthly hours. It could mean the more time you spend with the company, the more responsibilities are at your disposal.

5. There is somewhat a negative correlation (of -0.39) between the satisfaction level of the employee and whether the employee left the company. It makes sense for the employee to leave the company given he/she is not satisfied with the job.

Finding the attribute that has the highest influence on employee retention

To investigate the reasons why a large number of employees left, we split the dataset into 2; one in which employees left the company and the second in which employees stayed. The summaries of the datasets were outputted.

```
leftEmployees.describe()
```

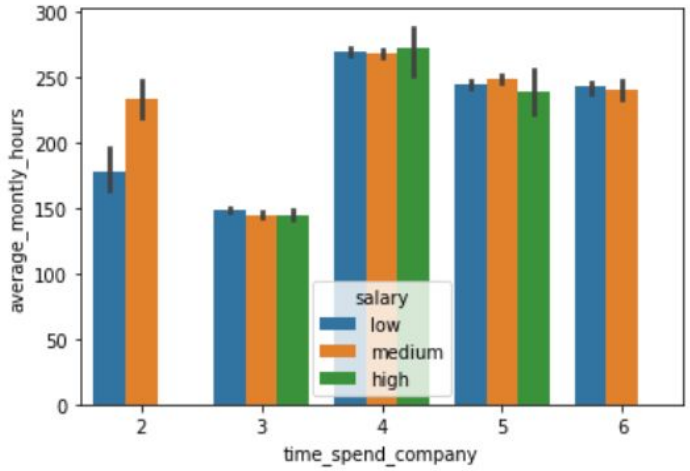
	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	promotion_last_5years	left
count	3571.000000	3571.000000	3571.000000	3571.000000	3571.000000	3571.000000	3571.000000	3571.0
mean	0.440098	0.718113	3.855503	207.419210	3.876505	0.047326	0.005321	1.0
std	0.263933	0.197673	1.818165	61.202825	0.977698	0.212364	0.072759	0.0
min	0.090000	0.450000	2.000000	126.000000	2.000000	0.000000	0.000000	1.0
25%	0.130000	0.520000	2.000000	146.000000	3.000000	0.000000	0.000000	1.0
50%	0.410000	0.790000	4.000000	224.000000	4.000000	0.000000	0.000000	1.0
75%	0.730000	0.900000	6.000000	262.000000	5.000000	0.000000	0.000000	1.0
max	0.920000	1.000000	7.000000	310.000000	6.000000	1.000000	1.000000	1.0

```
stayedEmployees.describe()
```

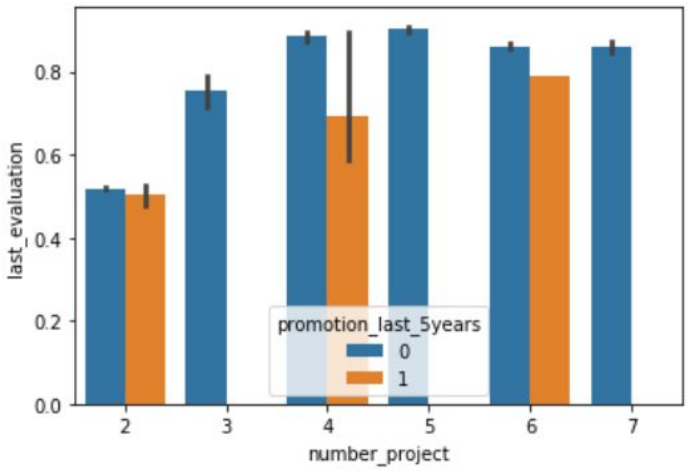
	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	promotion_last_5years	left
count	11428.000000	11428.000000	11428.000000	11428.000000	11428.000000	11428.000000	11428.000000	11428.0
mean	0.666810	0.715473	3.786664	199.060203	3.380032	0.175009	0.026251	0.0
std	0.217104	0.162005	0.979884	45.682731	1.562348	0.379991	0.159889	0.0
min	0.120000	0.360000	2.000000	96.000000	2.000000	0.000000	0.000000	0.0
25%	0.540000	0.580000	3.000000	162.000000	2.000000	0.000000	0.000000	0.0
50%	0.690000	0.710000	4.000000	198.000000	3.000000	0.000000	0.000000	0.0
75%	0.840000	0.850000	4.000000	238.000000	4.000000	0.000000	0.000000	0.0
max	1.000000	1.000000	6.000000	287.000000	10.000000	1.000000	1.000000	0.0

From comparing the two summaries, we can see that the satisfaction levels of employees who left (0.44) are much lower than the employees who stayed (0.67). Also, employees who left had to work for more number of hours per month (207.4 hours) than employees who stayed (199 hours). Furthermore, employees who left were much less promoted (0.005) in the last 5 years in the company compared to the employees who continued to stay (0.026).

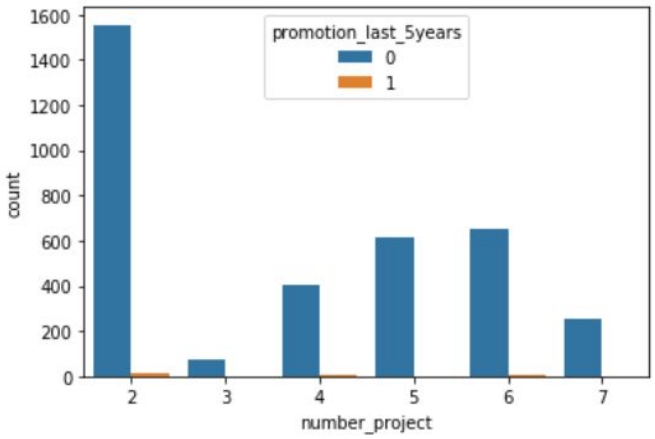
To explore how salary played its part in making employees leave the company, further bar charts were plotted.

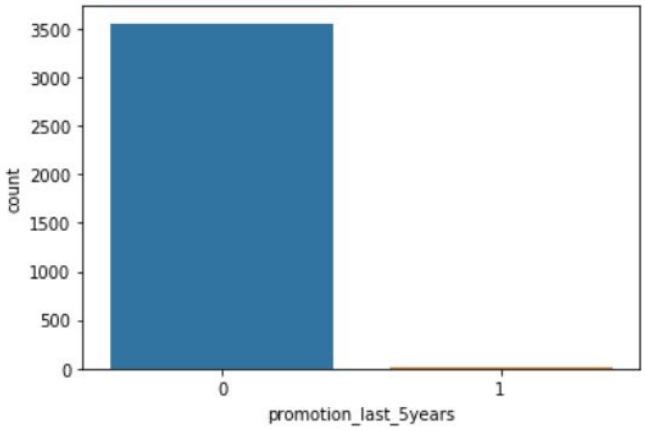


From the above graph, we can see that no employees who had worked for 6 years had a high salary. The number of high salaries being paid to employees who left does not increase based on the time they spent with the company. This might be another reason why employees chose to leave the company.

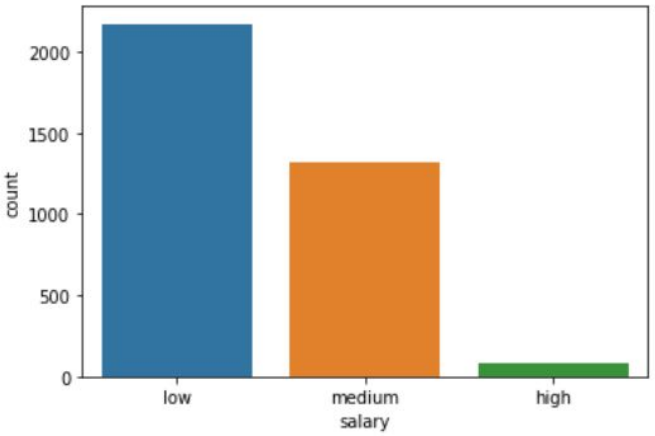


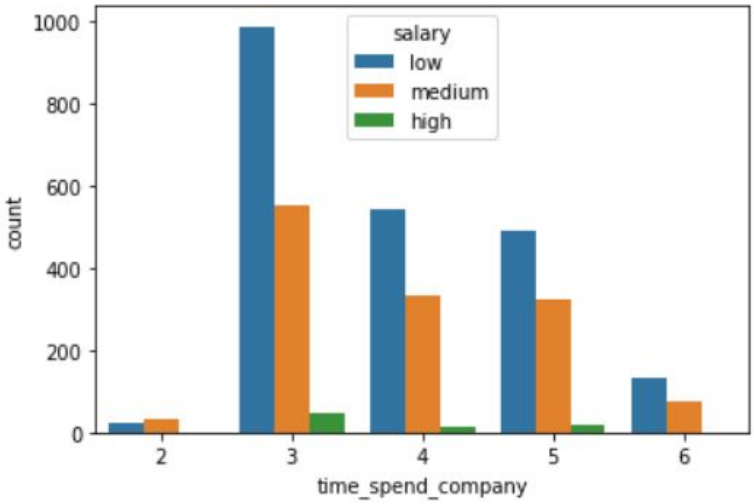
In this graph, no employees who had worked on 3, 5, or 7 projects were promoted in the last 5 years even though they had scored above 75% in evaluations. This could be another reason why these employees left.





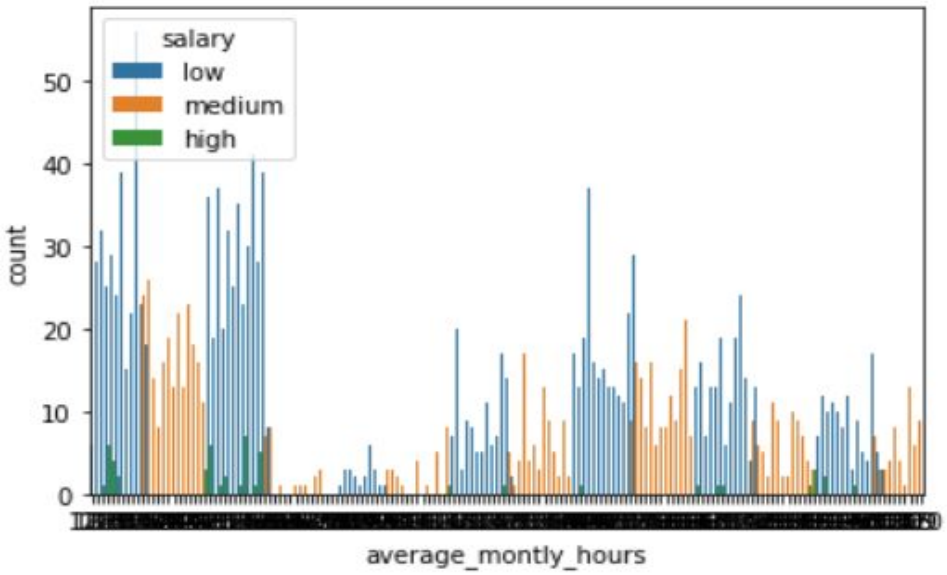
The above graphs show people who worked on even 7 projects failed to get a promotion in the last 5 years. Additionally, the number of employees who were promoted in the last 5 years is very close to 0. Maybe the employees got tired of waiting to be promoted. Hence, they chose to leave the company. From the number of projects, we see that those with a higher number of projects were more likely to leave and those with a *lesser* number of projects also left. This could be due to feeling they could do more work but was not being given the work which led to frustration with the company.





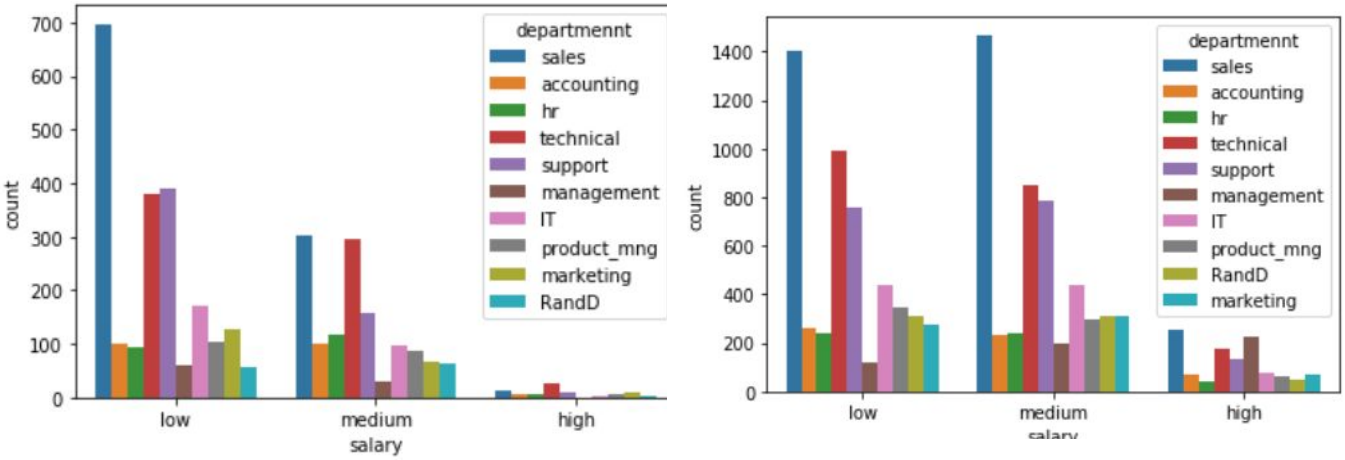
The above graphs show that out of the 3571 employees who left, less than 100 of them had high salaries while more than 2000 of them had low salaries. Additionally, we see that after spending 2 years in the company, the number of employees who had high salaries decreased. It is important to take note of the fact that no employee had a high salary if he/she had spent 6 years. This shows that the company does not reward experience with high salaries. Another interesting thing to note is that most employees left after spending three years with the

company. There could be several reasons such as the ones explained above, I.e. a lack of promotions, getting overworked, or not receiving a higher salary.



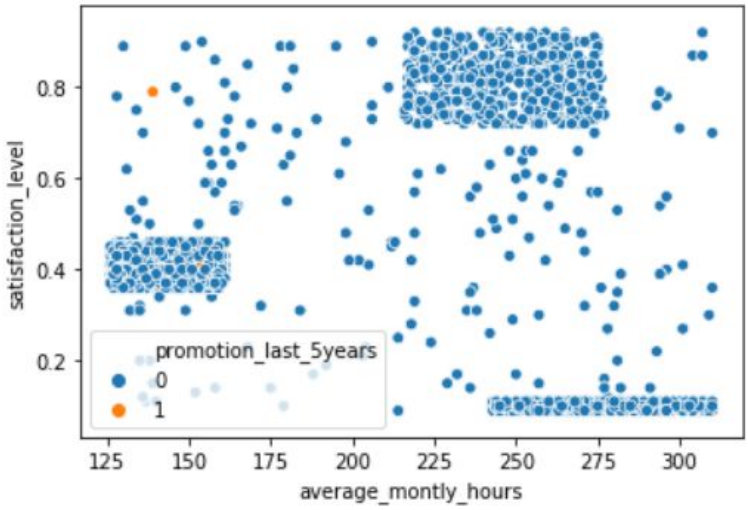
In this figure, it is highlighted that putting more hours in the work does not guarantee high salaries as very few employees obtain high salaries by putting in more hours of work.

Let's take a look department wise. The left bar graph is for people who have left the company and the right side bar graph is for people who stayed in the company.

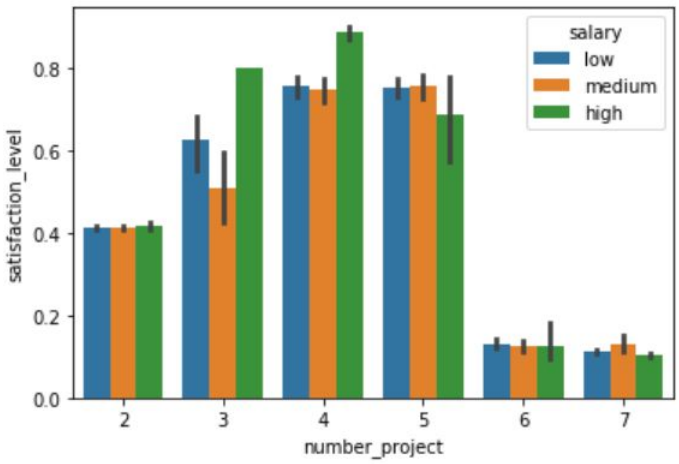


We see that people left the company because of the overall difference in salaries. The employees who left had much lesser high salaries and medium salaries compared to employees who stayed in the company. Furthermore, we see that the sales department has the highest

number of employees so it is normal to see more people from sales leaving the company as well. Therefore, we can disregard department-wise analysis as not having much impact on employee retention when compared with low salaries, the number of promotions.



The above scatterplot visualizes the employees that left, some had low satisfaction levels because they were getting too *little* work so they needed to put fewer hours while some left because of *too much* work so they had to put in more hours. Whilst others were content with their work but let's look at satisfaction levels and promotions to explain why those employees left. We see in the above plot that even those with HIGH satisfaction levels (above 0.7) eventually left due to a lack of promotion. We can barely see any orange dots above that highlight an employee who got promoted.

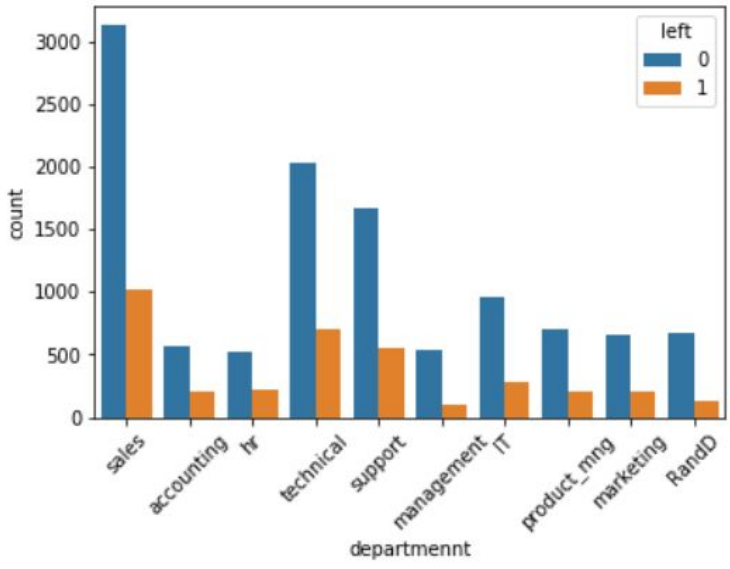


In this bar graph, one can conclude that projects between 3 and 5 are the right number to give to employees as this results in greater satisfaction levels and the number of employees who obtain

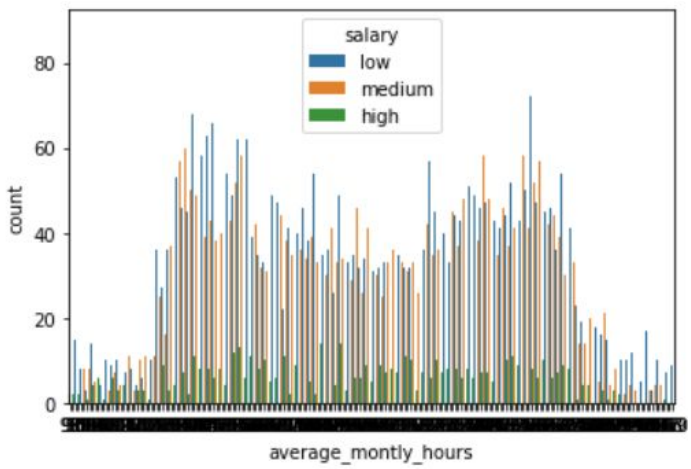
medium and high salaries is also greater. Giving 6 or 7 projects to employees is not a good idea since the satisfaction level drops greatly as it is too much workload for them.

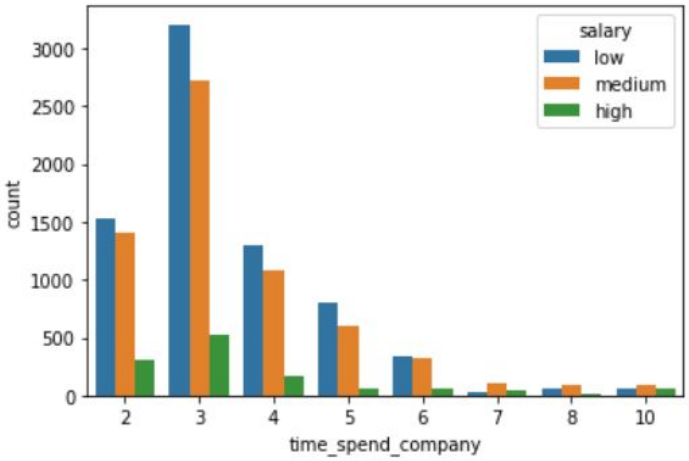
Summary of attributes having an impact on employee retention

- Employees are getting overworked, they have a high number of projects, high working hours.
- Some employees are under-working, they should work on more projects. It seems 3-5 seems to be the right number of projects to give.
- Most people that have left have low-medium tier salaries.
- Perhaps the ***most important attribute is the promotion in the last 5 years***. Employees are less likely to leave if they get promoted in the last 5 years.
- Employees are more likely to leave around the 3-year mark.
- High satisfaction is not enough, promotion matters. Even if an employee is rated highly by the employee and is doing good work, if they don't get promoted or get paid relatively well they will leave.
- The ratio of employees that say and leave department-wise is the same so there seems to be no relation to employee retention when it comes to department (as seen in the bar graph below).

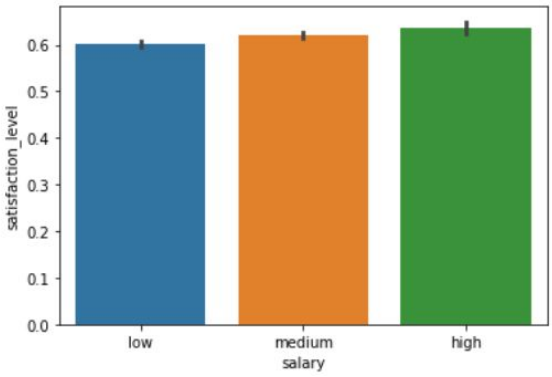


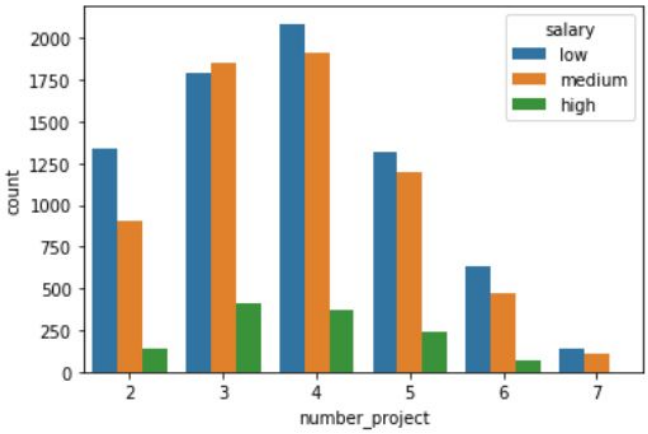
Dependency between categorical and numerical attributes



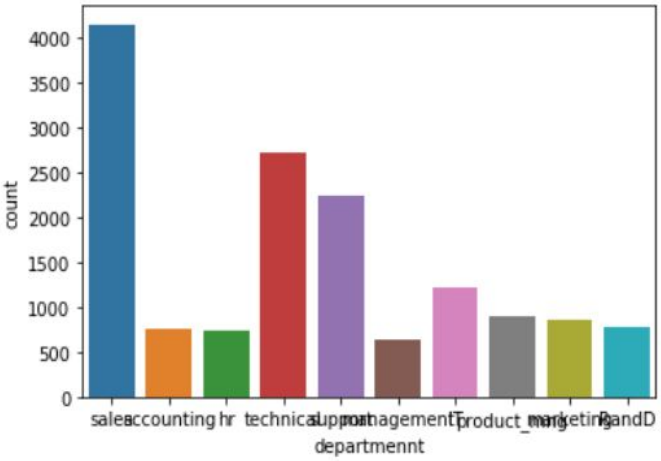


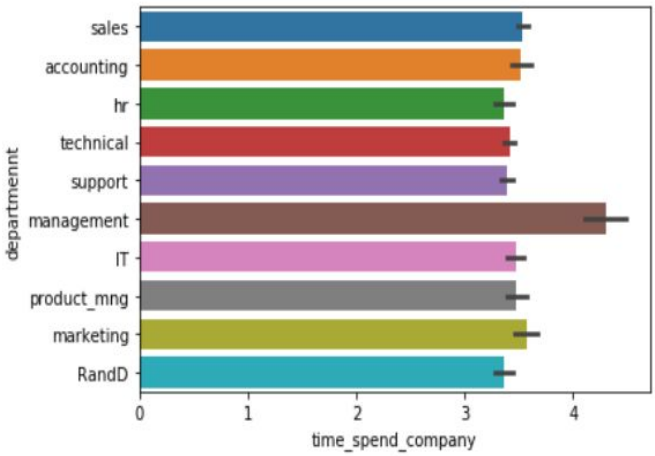
From the above graphs, it can be concluded that putting higher monthly hours does not lead to high salaries since the number of people getting high salaries by putting lower hours is almost the same as the number of people getting high salaries by putting greater hours of work. Additionally, the employees who get paid high and medium salaries are those who have worked for 3 years. After 3 years of spending in the company, the number of employees getting high and medium salaries get lower year by year.





There is a very tiny difference between the satisfaction levels among the employees who get low, salaries, medium salaries, and high salaries. The more interesting observation is employees are more likely to get high or medium salaries if they take projects between the number 3 and 5 inclusive.





Most numbers of employees are hired in the sales department and the least the number of employees is hired in the management department. However, despite a small management department, the employees in this department tend to stick around for more years than employees of all other departments.