HR Data Analysis

Alina Majeed 2020-11-0305
Muaaz Ahmed Noor 2020-02-0507
Danish Nadeem 2020-02-0177
Khawaja Usman Ghani 2020-02-0280

Description and Categorization of Variables

There are several types of variables present in the dataset upon which we will conduct our analysis. They can roughly be broken down into numerical and categorical data types, with further categorizations as well. There are 5 numerical variables; 2 of them are continuous as they can occupy any value between a given range. These are *satisfaction_level* and *last_evaluation*. The other 3 numeric variables are discrete as they primarily take up integer values. These are *number_project, time_spend_company* and *average_monthly_hours.* Furthermore, there are also 3 binary categorical variables in the dataset that take up the value of 0 or 1. These are *work_accident, promotion_last_5_years* and *left.* A value of one indicates that the respective event has occurred, whereas 0 shows that it has not. Lastly, we have 2 other discrete categorical variables; *department* and *salary*. *Department* is a nominal categorical variable as there is no ordering between the various departments whereas, *salary* is ordinal as there is meaningful ordering between different salary levels.

Exploratory Data Analysis

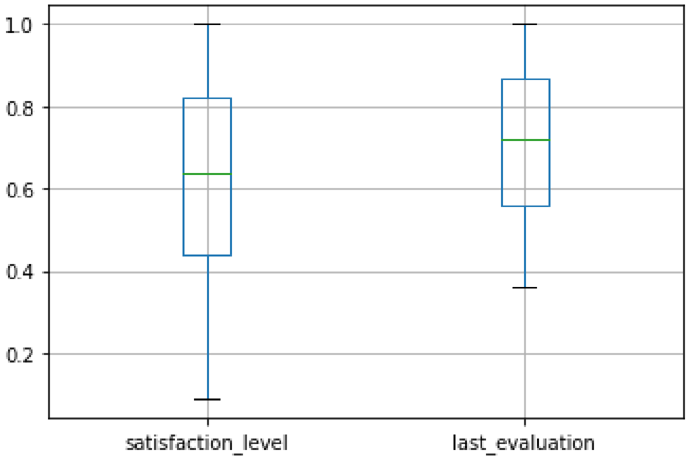
Missing Values

In order to carry out integral parts of our analysis, we needed to check the data for missing values and replace them with an appropriate technique. However, upon execution of the Python code, we discovered that there are no missing values present in this dataset hence, this step does not need to be carried out. There is a total of 14,999 values in this dataset.

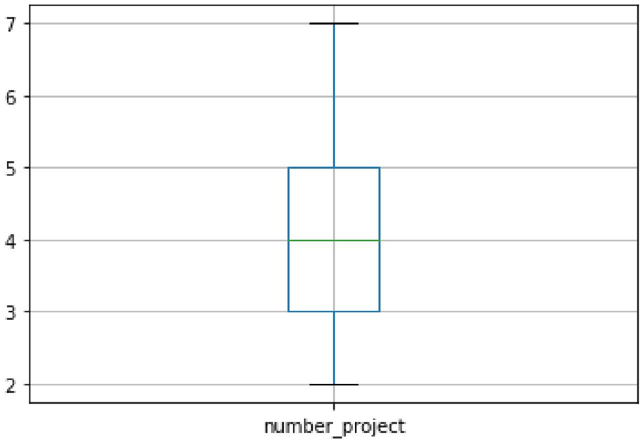
Summary Statistics

To better understand the data and its distribution, summary statistics were calculated, and their visualizations were made. The summary statistics included count, mean, standard deviation, minimum, 25th percentile, 50th percentile, 75th percentile and maximum. Based on the above calculations, boxplots were plotted for *satisfaction_level, last_evaluation, number_project, average_monthly_hours, time_spend_company*.

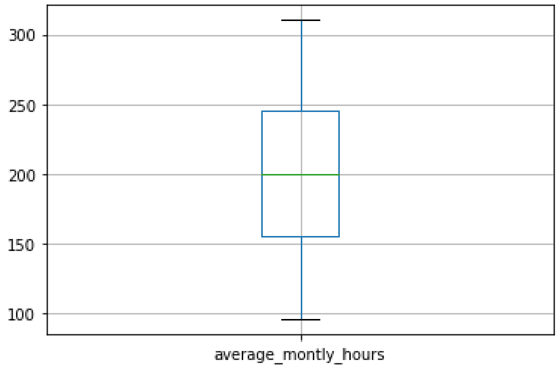
Satisfaction_level and *last_evaluation* both represent scores that can be plotted for analysis and comparison on one boxplot.

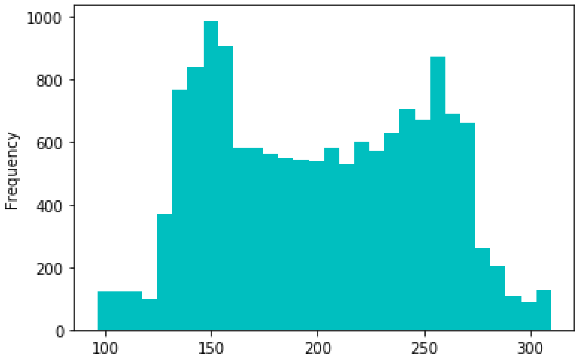


As shown above, *satisfaction_level* is more dispersed as its whiskers for minimum and maximum points extend further than that of *last_evaluation*. *Satisfaction_level* ranges from a minimum score of 0.09 to a maximum of 1.0, whereas, *last_evaluation* extends from 0.36 to 1.0. The interquartile range (Q1 – Q3) for *last_evaluation* works out to be 0.31 in comparison to 0.38 for *satisfaction_level*. Even though these metrics are calculated differently, perhaps this variation can indicate important differences between the employee and employers' perception of the company and the work.

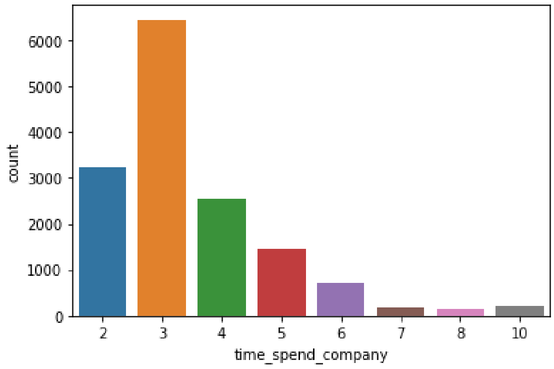


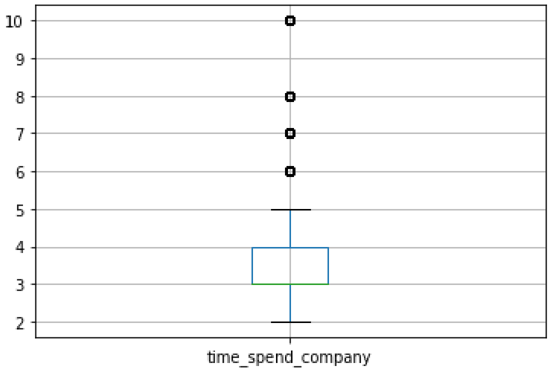
Moving on to the *number_projects* variable, we can see that the minimum number of projects done by an employee are 2 while the maximum is 7. Most of the values lie between 3 and 5 projects, making the interquartile range 2, which seems reasonable.



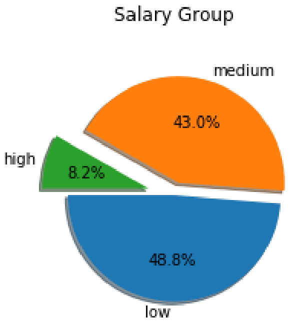


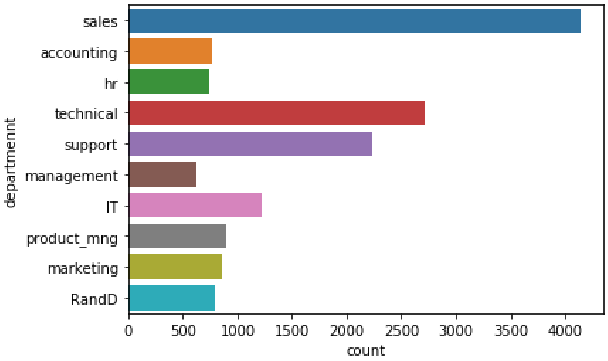
Additionally, we can observe the *average_monthly_hours* variable. Both the bar chart and the box plot show that majority of the employees work between 156 and 245 hours a month. If we assume an average of 20 working days in a month, this comes out to be between 7.8 and 12.25 hours per day. Additionally, it can be inferred from the histogram that this variable is bimodal.



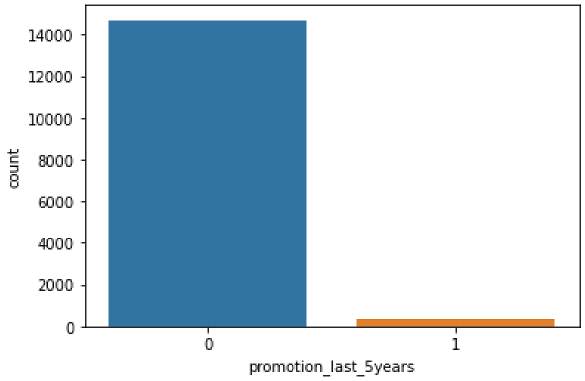


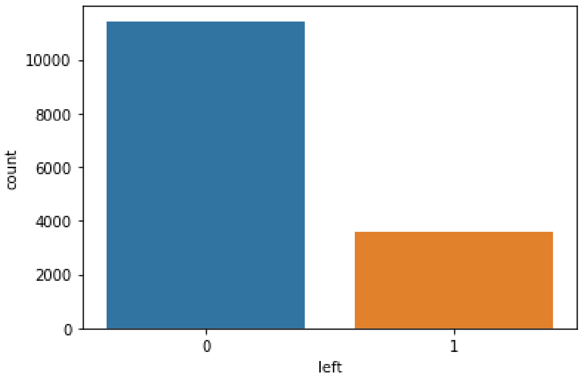
Furthermore, we move on to exploring the variable *time_spend_company.* As can be seen on the boxplot, the maximum time an employee has been in the company is ten years while the minimum is two years. Most the people have spent between 2 and 5 years at the company, which can be seen both from the bar chart and the boxplot. However, there are some outliers of employees who have spent seven, eight and ten years in the company. The interquartile range for this variable is 1 year. The bar graph represents a slightly clearly picture as we see that the highest count is that of 3 years of time spent in the company.



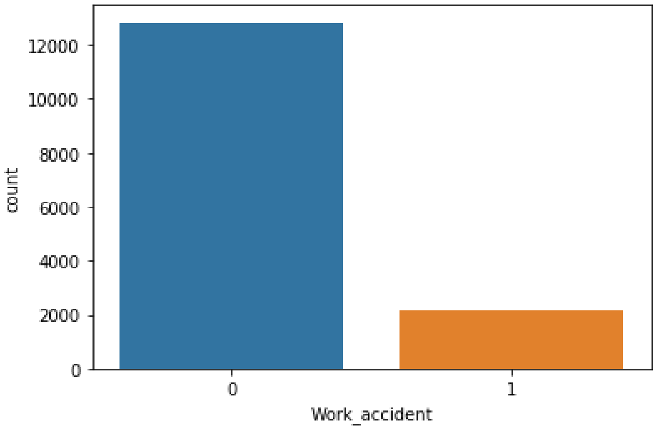


The *salary* variable also offers interesting insight into the company. We can tell that majority of the employees are paid “low” salary and only 8.2% fall into the “high” bracket. Since we do not have enough information on the ranges for these categories or the type of company, it would be difficult to draw out any conclusions, but it can be safe to say that majority of the company is paid between low and medium salaries. The *department* variable can also be observed to develop a better understanding of the company. The sales department has the highest number of employees followed by the technical and support departments. Perhaps the company belongs to the consumer-goods industry as it requires a large sales staff.



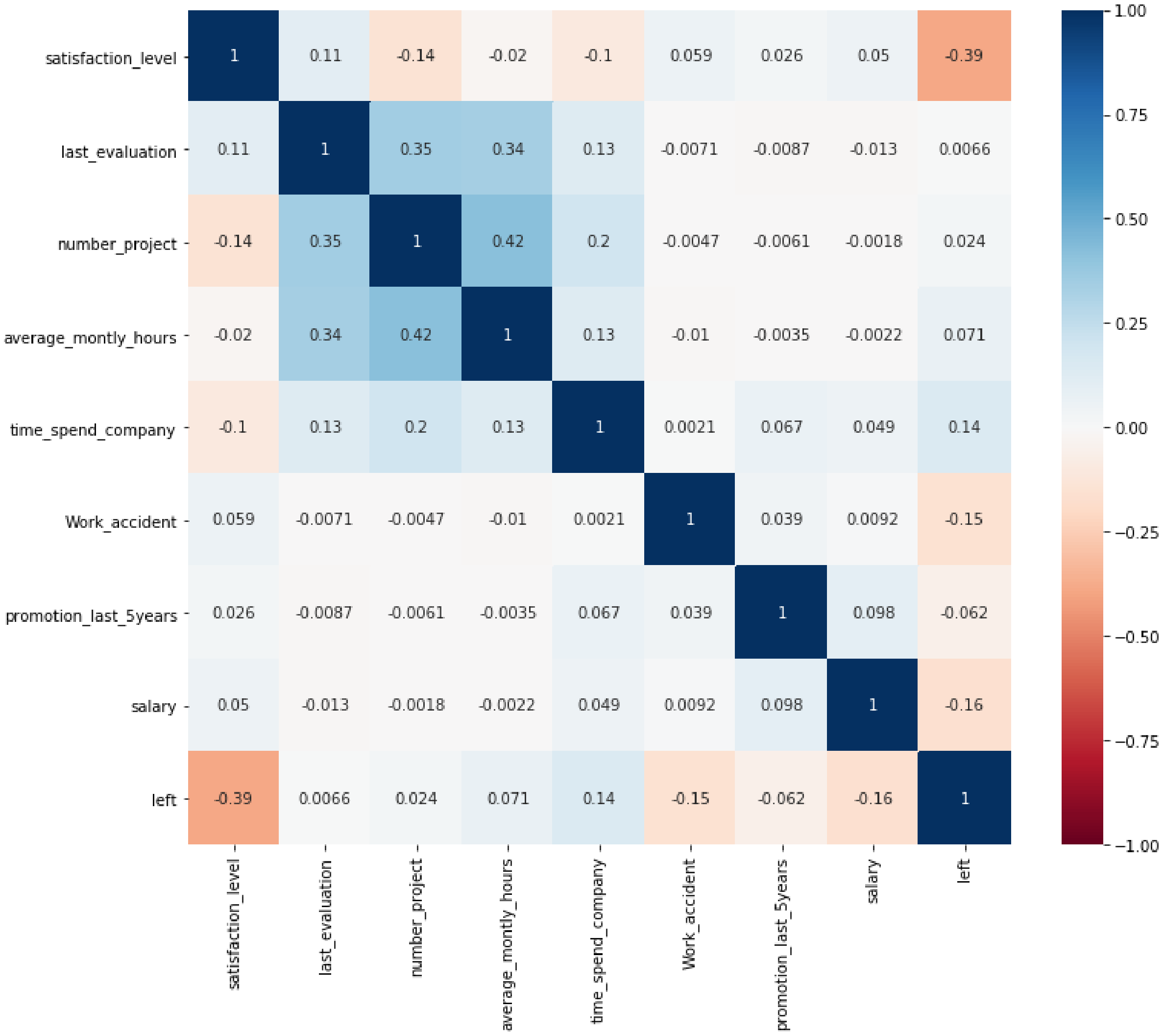


The binary variables help us understand the company employees even further. From the bar graph for *promotion_last_5_years*, we can see that more than 14,000 employees have not been promoted and only a very small minority of employees have been promoted. This gives an indication that employees might be unhappy with the company, and the *left* variable bar graph offers more insight into this. 3571 employees in total have left the company and 11,428 of the employees have been retained. These figures give a retention rate of 76.2%, which is not that good for a large company.



Overall, the working environment is also safe as there haven’t been too many work accidents, with a total of around 2000 employees reporting an accident.

Correlation between attributes



	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	promotion_last_5years	salary	left
satisfaction_level	1	0.11	-0.14	-0.02	-0.1	0.059	0.026	0.05	-0.39
last_evaluation	0.11	1	0.35	0.34	0.13	-0.0071	-0.0087	-0.013	0.0066
number_project	-0.14	0.35	1	0.42	0.2	-0.0047	-0.0061	-0.0018	0.024
average_montly_hours	-0.02	0.34	0.42	1	0.13	-0.01	-0.0035	-0.0022	0.071
time_spend_company	-0.1	0.13	0.2	0.13	1	0.0021	0.067	0.049	0.14
Work_accident	0.059	-0.0071	-0.0047	-0.01	0.0021	1	0.039	0.0092	-0.15
promotion_last_5years	0.026	-0.0087	-0.0061	-0.0035	0.067	0.039	1	0.098	-0.062
salary	0.05	-0.013	-0.0018	-0.0022	0.049	0.0092	0.098	1	-0.16
left	-0.39	0.0066	0.024	0.071	0.14	-0.15	-0.062	-0.16	1

In order to explore the relationship between various attributes, we plotted a correlation heatmap for all the variables. Focusing on Employee retention, we were able to gather the following relationships:

- <u>An employee is less likely to leave if they have reported a high satisfaction level score:</u> *Left* and *satisfaction_level* are negatively correlated by 0.39 which means that as the satisfaction score increases, the binary variable value for *left* approaches 0 which means they are less likely to leave. This finding makes intuitive sense, as an employee who is satisfied at the company would not want to leave. ***satisfaction_level* has the greatest effect on *left* as it has the highest correlation of any variable with *left*.**

- <u>An employee is less likely to leave if they have a higher salary:</u> *salary* and *left* have a negative correlation of 0.16, indicating that as an employee moves up the salary groups, he is less likely to leave and the value of *left* approaches 0.

- An employee who has spent more time in the company is more likely to leave: *time_spent_company* and *left* have a positive correlation of 0.14. This finding is counterintuitive at first sight. However, at greater inspection, we see that the employees who've spent the most time in the company are very less and are classified as outliers. This finding makes sense for the spectrum of 0-5 years spent at the company since most employees have spent lesser than 5 years at the company.

We also gathered some other interesting dependencies among other variables. These findings are listed below:

- An employee with a higher last evaluation score is working for longer hours: *last_evaluation* and *average_montly_hours* are positively correlated by 0.34 which means that as the evaluation score for an employee increases, longer working hours are also noted.

- An employee with a higher last evaluation score is likely to have more projects: *last_evaluation* and *number_projects* are positively correlated by 0.35 which means that as evaluation score for an employee increases, they also seem to have a greater number of projects.

- An employee working on more projects works higher number of hours: *number_project* and *average_monthly_hours* are positively correlated by 0.42 which means that as the number of projects an employee has increases, we also see an increase in their monthly working hours.

Cluster and Outlier Analysis

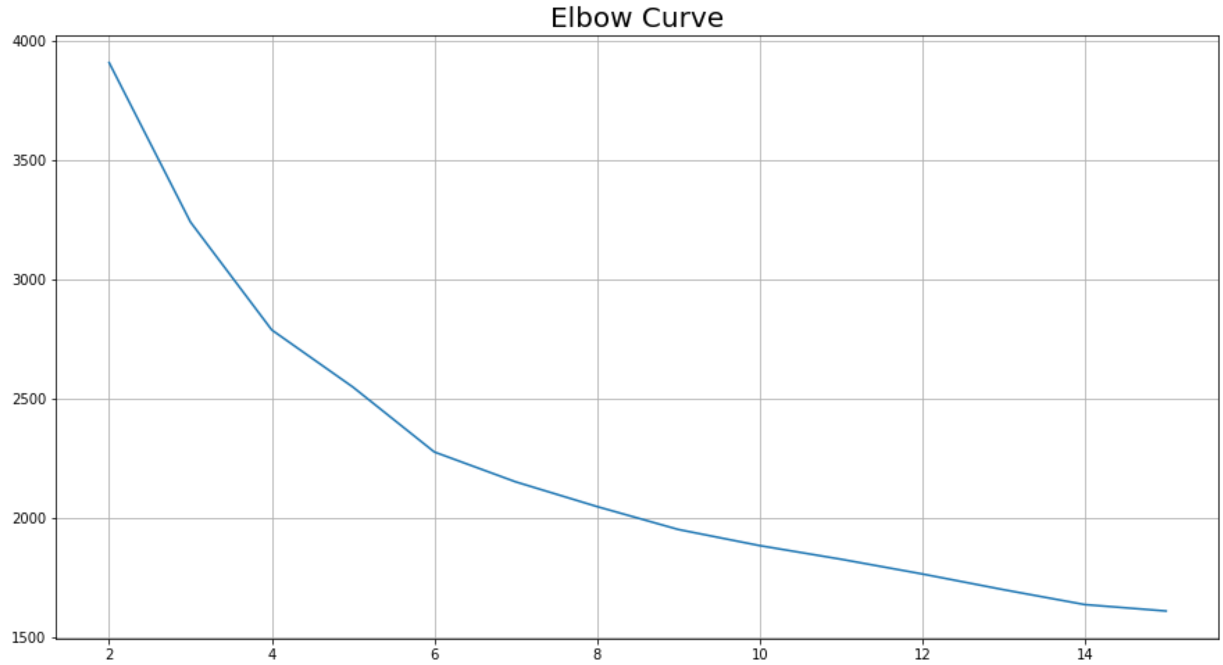
Cluster Analysis

The categorical attributes *department* and *salary* were converted to numeric values. Afterwards, all attributes were normalized so they are on the same scale. For cluster analysis, we used the K-means algorithm as it proved to work better on our dataset. The initial task of finding the optimum number of k required:

a) Carrying out a suitable method based on statistical computation

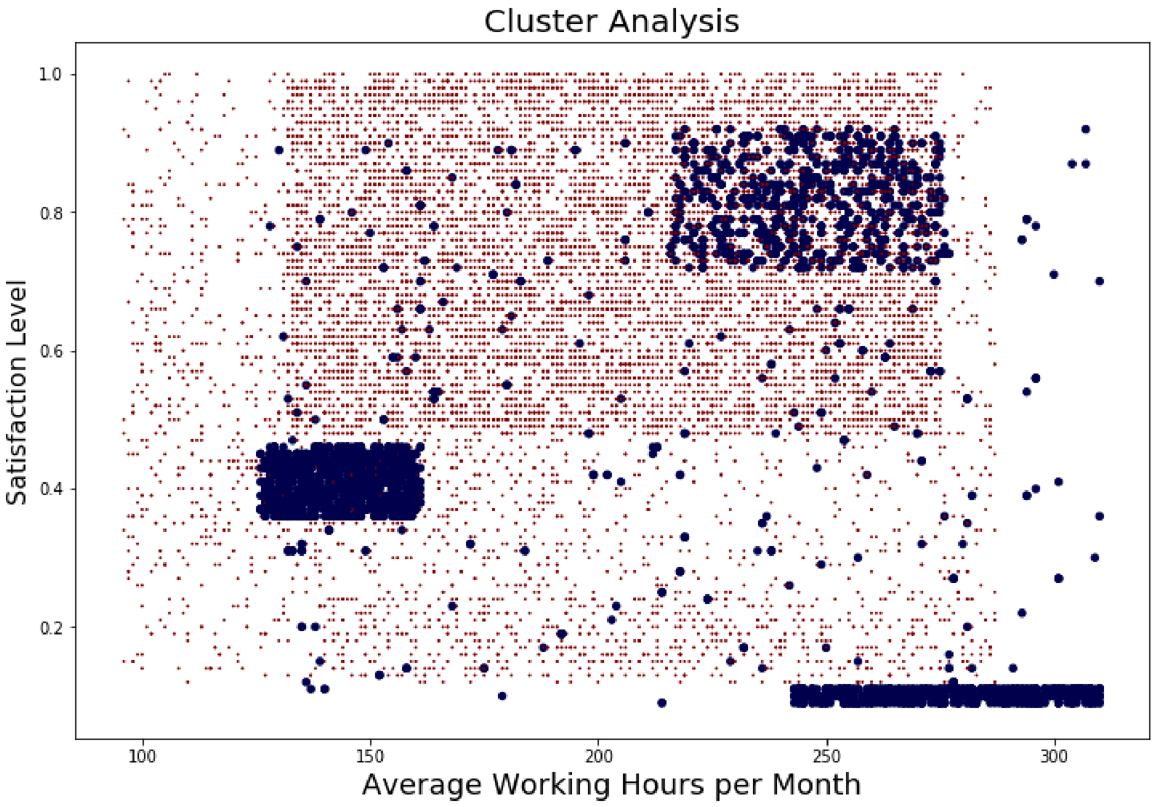
b) Using this data along with the domain knowledge required for the task at hand.

We first implemented the infamous Elbow Method by testing the K-means algorithm on our dataset with different values of K. These values ranged from 2 to 15. The within cluster variations for each value of K were calculated and plotted against K. This gave us the following result:

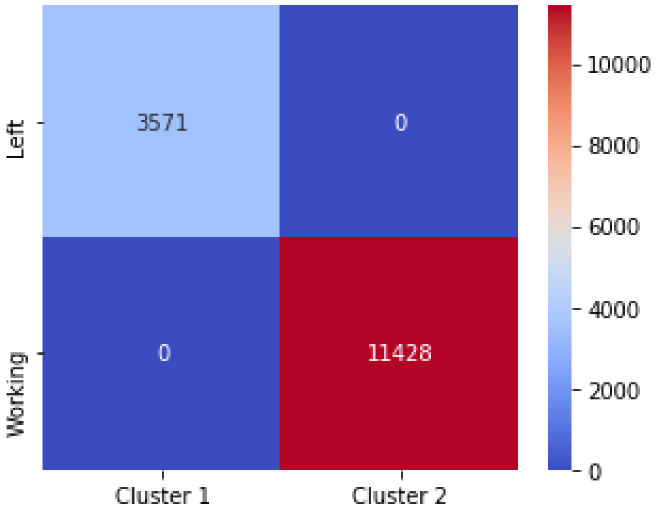


Our results show a fairly smooth curve which does not seem to show a clear ‘elbow’ as the decrease in variance is not considerably larger for any one step on k. Therefore, this step proved to be redundant.

Since our primary task for the project was to analyze employ retention, clustering the employees based on whether they are still working in the company or not seemed to be the optimal choice. To test this, we used the K-means algorithm with K=2. Upon visualizing the results, we could clearly identify that the clusters were formed based on the *left* attribute.



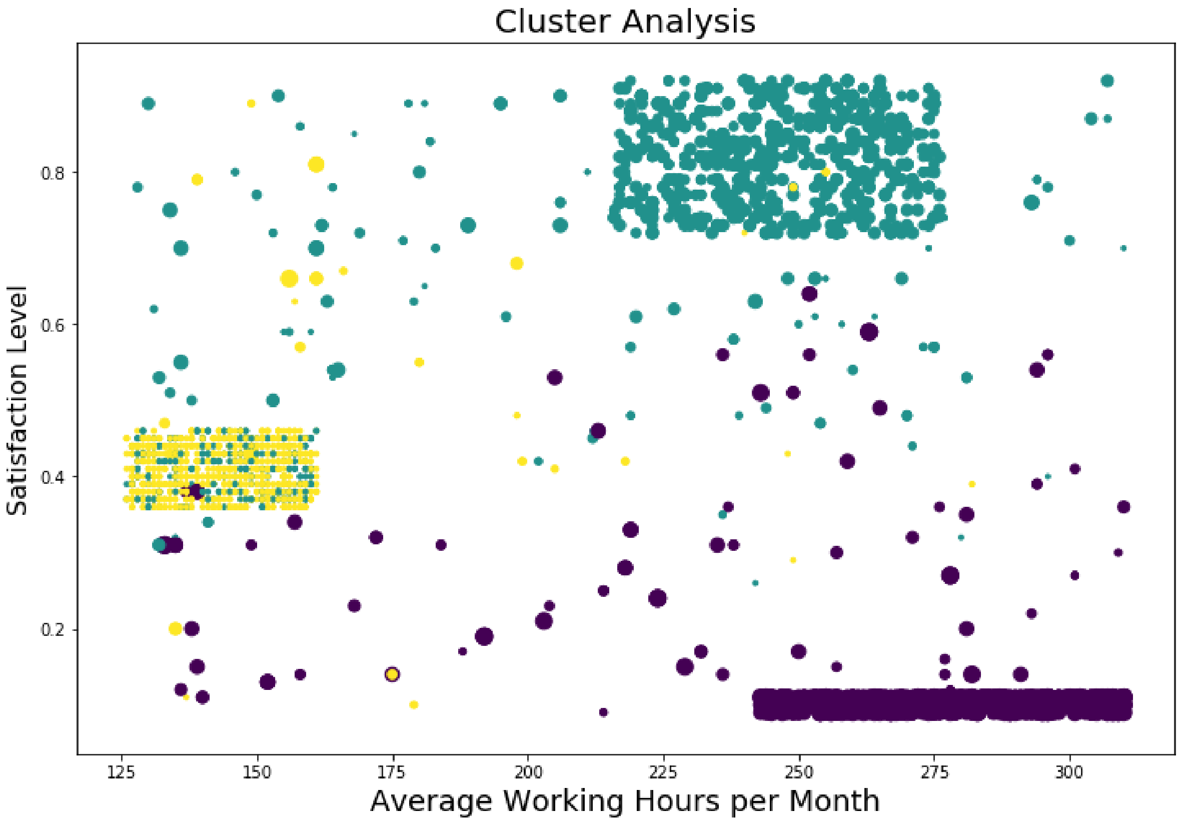
The above scatter plot was used for visualizing the resulting clusters. Average working hours were plotted against the Satisfaction Level for the employees, whereas the radius of each data point is used to indicate whether the employee still works in the company or not. A bigger radius (and therefore a larger point) indicates the employees who have left the company. We can clearly see that all blue data points are larger in size than the red data points. To further substantiate our claim, we created a confusion matrix to see how the clusters were made. The confusion matrix heatmap shows that all 3571 employees who left the company were clustered in the blue cluster while the 11428 employees still working in the company were part of the red cluster.



The scatterplot also provided insight into the characteristics of different groups of employees within each cluster. The employees who are still working in the company do not show any sub clusters. However, a greater chunk of these employees has a higher satisfaction level and works for a greater average number of hours and are therefore concentrated on the top right corner of the scatterplot. Another relatively smaller chunk of employees have a low satisfaction level and high average working hours. However, we did not feel like there was any need to further explore these groups, as they will not provide any valuable insight to our primary goal.

Employees who have left the company clearly showed a set of similar characteristics, apart from a handful of outliers. We can clearly see three different groups within this cluster, and this paves the way for further analysis on this cluster.

To make our analysis easier, we created a separate dataframe of employees who have left the company and performed cluster analysis on this new dataframe. We used K=3 as we could identify three different groups from the previous scatterplot. The results were visualized again using a scatterplot which is shown below. The radius of the data points now represents the number of projects the employee was working on.



Upon visualizing the results, the cluster characteristics are identified as follows:

(i) Employees with low satisfaction levels who were working low hours and on a considerably smaller number of projects.

(ii) Employees with extremely low satisfaction levels who were working a very high number of hours and on a considerably larger number of projects.

(iii) Employees with high satisfaction levels who were working a high number of hours and on an average number of projects.

To get a better understanding of the characteristics of each cluster, we grouped the attributes based on the said clusters. The mean of each attribute was computed, and the results are shown in the following table.

Cluster	Satisfaction Level	Last Evaluation	Projects	Monthly Hours	Time at the Company	Salary*
1	0.124	0.854	6.09	271.06	4.05	1.43
2	0.413	0.514	2.06	145.31	3.03	1.57
3	0.688	0.795	3.78	214.70	4.47	1.28
Overall**	0.613	0.716	3.80	201.05	3.49	1.59

*Group-wise salary: 1-Low, 2-Medium, 3-High **Represents the original dataframe of all 14,999 employees

Following on with the initial characteristics, we were able to identify even more interesting findings. These could indicate why employees from a certain cluster left the company.

For example, employees in cluster 1 had the greatest number of projects, the most working hours, and the best scores on their evaluations. However, they were still very low on the satisfaction level charts. The salary of these employees was below the overall average. This gives us reason to believe that even though these employees were performing fairly well, they were extremely overburdened with work, and probably did not feel like they were properly rewarded for that.

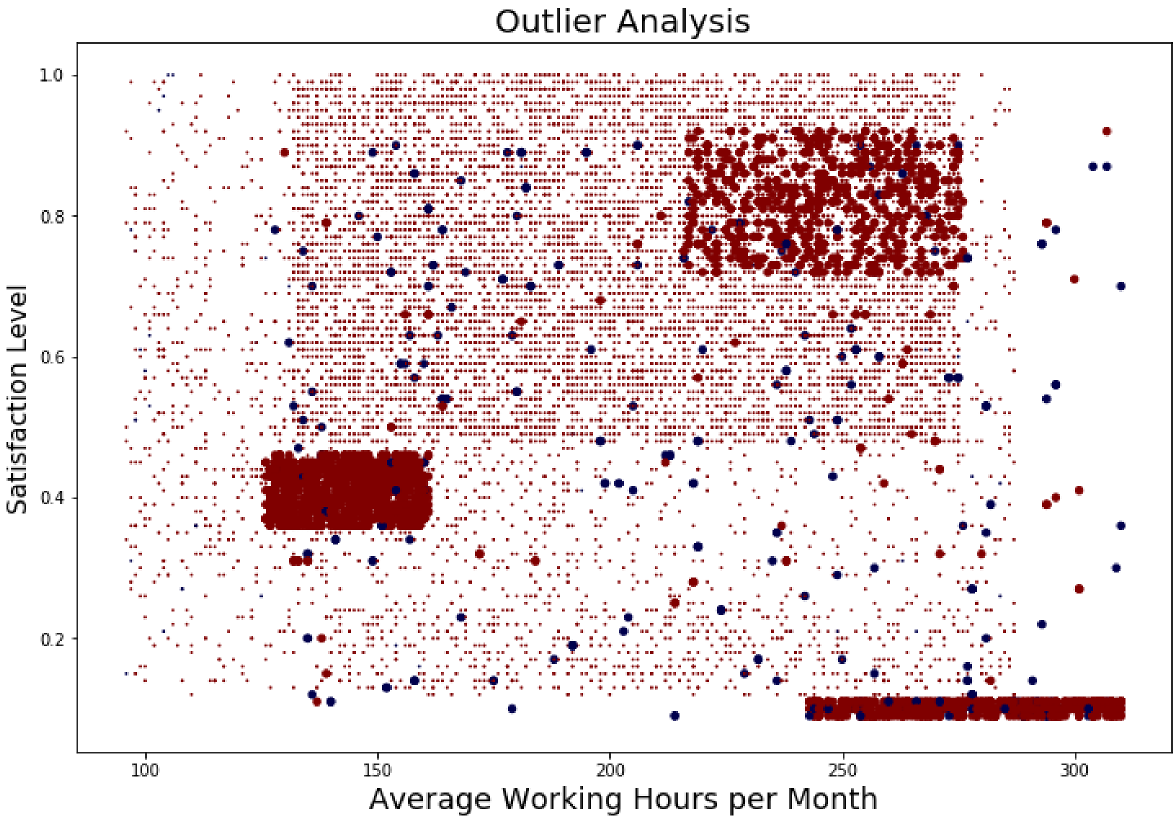
Following on with our analysis, we now look at employees from cluster 2. These employees had low satisfaction levels, and their evaluation scores were also below par and the lowest amongst all clusters. They were working for very less hours and had around 2 projects on average to their name. This shows us that they were quite low on their motivation and were not working up to the mark. However, they had the highest average salary group amongst the clusters. This probably took away their motivation to work, or another reason could be the low average time they had spent at the company – around 3 years.

Cluster 3 showed the oddest behavior according to our analysis. The performance of these employees was on par as they had a handful number of projects they were working on, and their working hours were also better than the overall mean. They had performed well on their evaluations as well and had above par satisfaction levels. However, they had spent 4.47 years on average at the firm, which is very high compared to the overall average. The average salary of these employees was also below average. This gives reason to believe that even though they were quite satisfied with their work, they probably felt underpaid or that they had been at the same firm for too long and needed a change.

We feel that the insights provided by cluster analysis can help the firm in making the right managerial decisions. Details of these suggestions will be provided later. Focusing on the right fronts can prove pivotal in sustaining employee retention.

Outlier Analysis

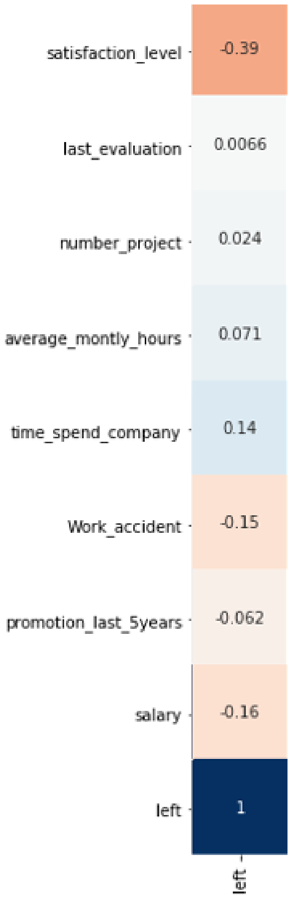
We saw throughout the cluster analysis that our plots had several outliers which did not belong to a specific identified group. Now we will go further and try to identify such points and single out such characteristics. We used Local Outlier Factor, which is a density-based approach, to detect these outliers. We set the k nearest neighborhood at 20 for the data points. Running the algorithm resulted in the detection of **364** outliers. The results were visualized using a scatter plot as before, which is shown below. The radius indicates whether an employee has left the firm or not.



The scatterplot shows that the majority of these outliers are employees who have left the firm but do not fall in any of the identified groups. Therefore, we cannot find out their reason to leave the firm as it might involve other personal factors. However, some of the outliers are also employees who still work at the company. Most of these employees seem to be the ones who do not fall in the major chunk in the top right of the scatterplot identified earlier. To better substantiate these claims, we again looked at the mean attributes of the outliers. The mean of *left* came out to be 0.87 for these outliers, indicating that there is an 87% chance of the outliers being an employee who left the firm. The average satisfaction level of these employees was 0.43, which means that most of these outliers were employees who were unsatisfied with their jobs. The rest of the attributes did not give any interesting results, or were close to the population mean of the specific attribute.

Recommendations:

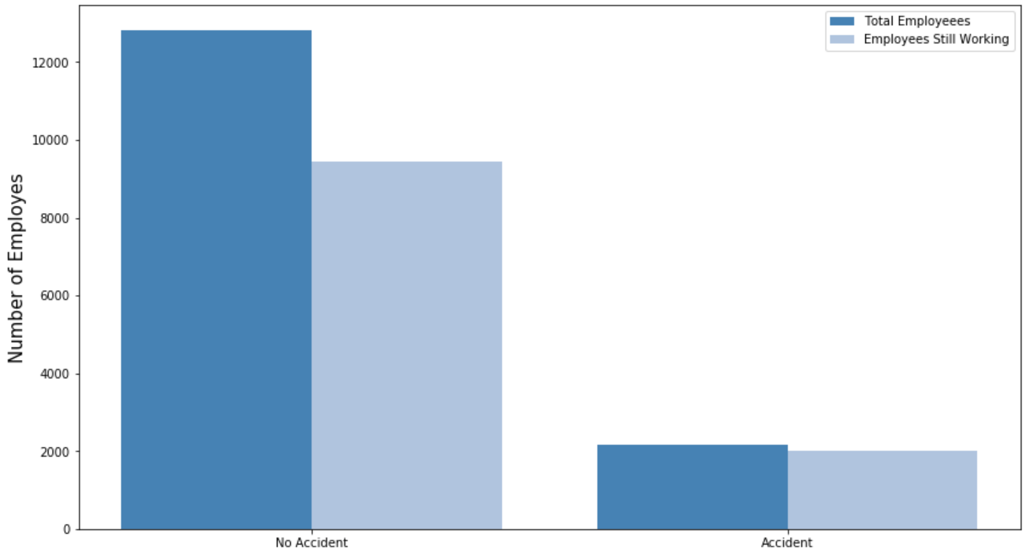
Analyzing the data through different techniques proved to be extremely insightful to recommending managerial decisions that could potentially affect employee retention. We were able to identify various reasons as to why the employees were leaving. An unsurprising, but rather intuitive, relationship was with the satisfaction level of the employees. Correlation analysis on the data also showed that the attribute *left* had the greatest correlation with *satisfaction_level.* To explain in simpler words, an employee who is not satisfied at the firm has a much higher chance of leaving. Another factor that will affect

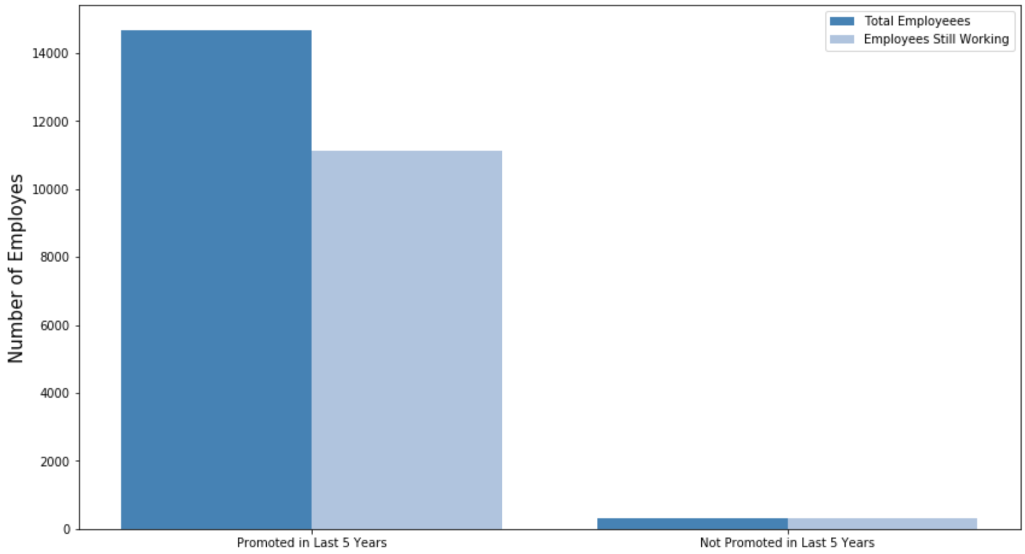


the employees' decision is the salary they are paid. This result is highly intuitive as well. We found a correlation of -0.16 with *salary*, indicating that the employees are more likely to leave if they are paid less. Since the salary was divided into three groups, we expect the actual correlation to be even greater had numerical values for salary had been used. The company can focus on making the office environment safer as this will likely reduce some of the numbers of left employees. Employees also seem to stick to the company if they have worked long enough, as correlation with *time_spend_company* is 0.14. The company can grant incentives to employees which encourage them to work for more years. This would in turn bolster their chances to keep working in the company. All the mentioned changes will also positively affect the satisfaction level of employees, making it more probable for them to not look for a career change, and stick to their current job. Although other attributes also show a slight correlation, but the magnitude of the relationships is too less to provide any useful insight, or too help us in finding any further suggestions.

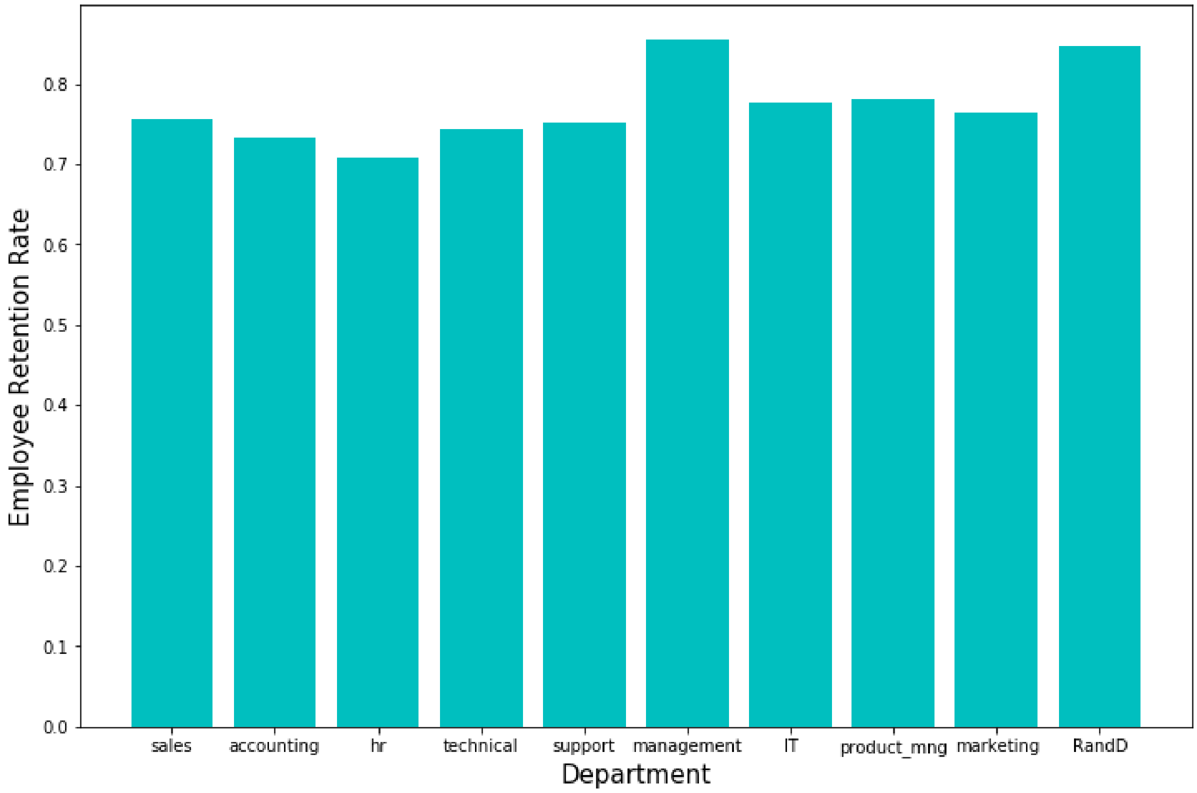
The findings of cluster analysis also gave us insights into why specific groups of employees leave the firm. As noted before, a cluster of employees were leaving the firm because they were extremely overworked. They had been working for over 270 hours per month on average. The mean projects they had been taking up was over 6. On the other hand, one other cluster had employed who had been working for significantly lower hours and below average number of projects. What the company could focus on would be a more balanced division of workload among its employees. Although this workload might differ among departments, it could be incorporated at each department level depending on their requirements. This would lessen the burden on the employees who likely leave due to overburden, as well as encourage and motivate employees who seem to have lost interest in the job. The overall affect would likely reduce the number of people leaving their jobs from these clusters.

Dwelling deeper into the analysis on each attribute, we found a very interesting result. As shown in the graph below, a larger portion of employees seem to stay at the company among the people who experienced accidents, compared to the ones who didn't. Although this seems oddly counter-intuitive, this finding is backed by our correlation analysis as well, which shows that *left* has a correlation of -0.15 with *Work_accident.* However, since the size of both portions differ considerably, we cannot say that there is a causal relationship. Similar analysis on *promotion_last_5years* showed that 94.04% of the employees who were promoted in the last 5 years were still working at the company. On the contrary, this percentage was merely 75.8% for employees who weren't promoted in the same period.

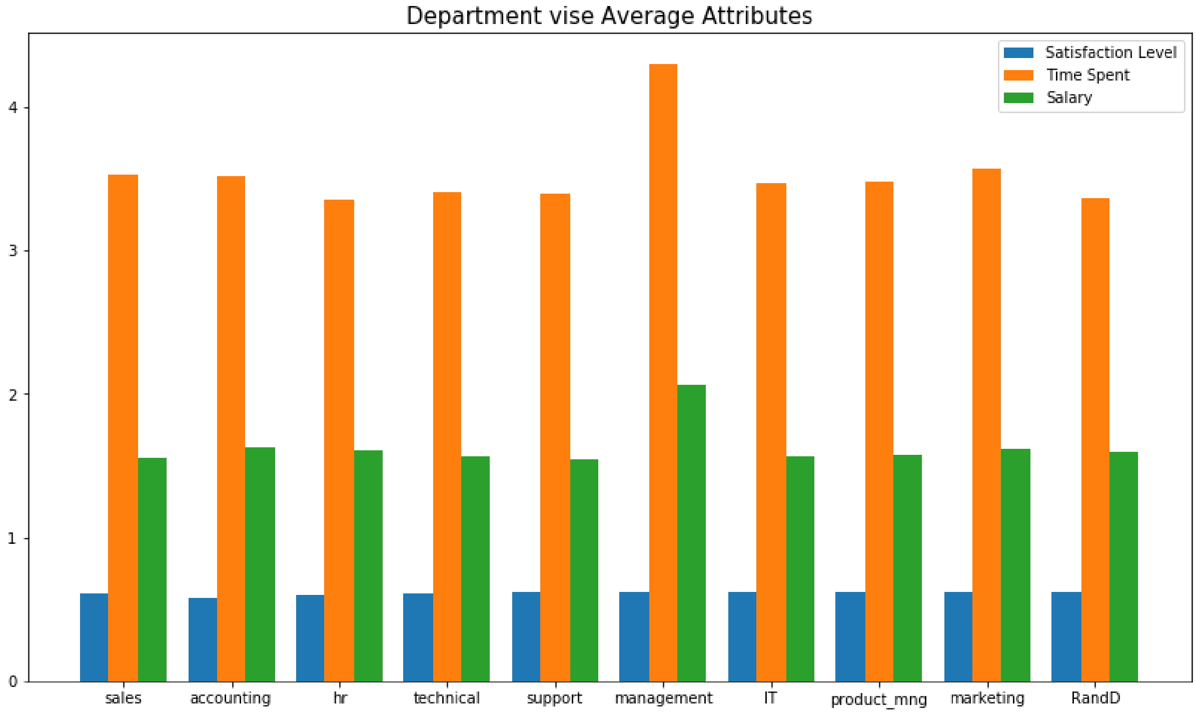




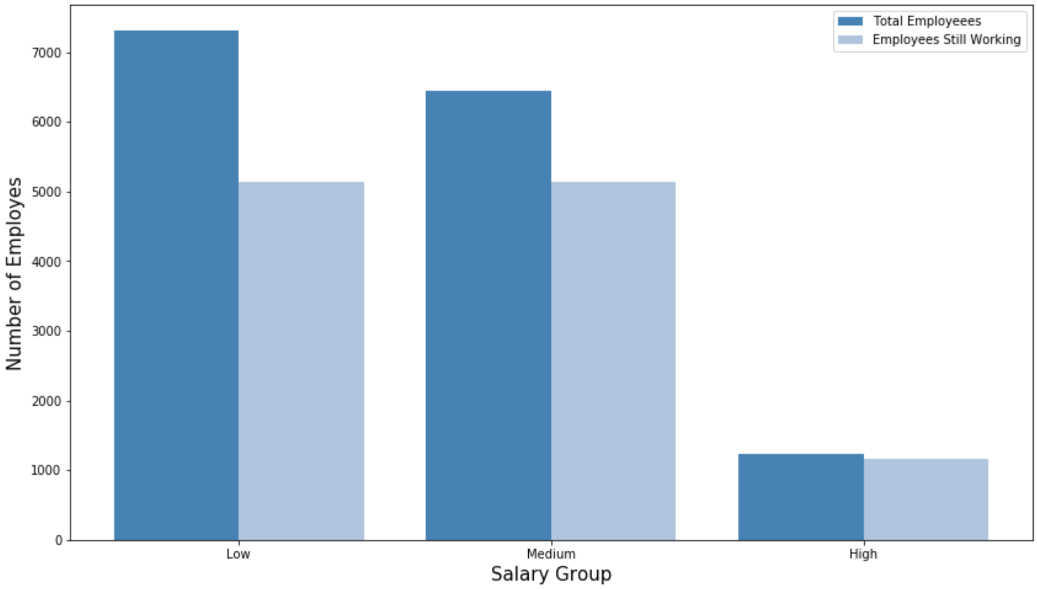
In order to better understand which variables have a high impact on the employee retention rate in different settings, we make some visualizations that would help us further explore some relationships. In the first visualization, we broke down employee retention rate by department. As can be seen in the bar graph below, there isn't too much variation between all the departments, however, retention is highest in management followed by R&D.



Upon further investigation, we note that while the satisfaction level is almost similar across all departments, the salary and the time spent are also the highest in the management department. This would explain why their retention rate is the highest and also provide suggestions for areas that HR can work on.



Breaking down salary further, we see that it has an impact on the retention rate. For all the employees that were being paid a “high” salary, very few of them left. This can be seen as there is barely any difference between the bars for ‘Total Employees’ and ‘Employees still Working’ in the bar graph shown below. On the contrary, a significantly large number of employees left the company among the group of ‘low’ wagers. This number for the ‘medium’ group lies in between these two extremes.



The management can consider all the suggestions given above, as our analysis leads us to believe that it will positively affect the retention rate of the company.