**Group 22**

* Syed Suleman Maqsood (20100164)
* Zain Haider (20020537)
* Bilal Masood (20100034)
* Shahzeb Faheem Khan (20100157)

**HR Data Analysis Report to Yelp Management**



**Table of Contents**

[Chapter 1: Exploratory Data Analysis](#_Chapter_1:_Exploratory)Page 1

[Chapter 2: Cluster and Outlier Analysis](#_Chapter_2:_Cluster)Page 9

[Chapter 3: Recommendations & Summary](#_Chapter_3:_Recommendations)Page 13

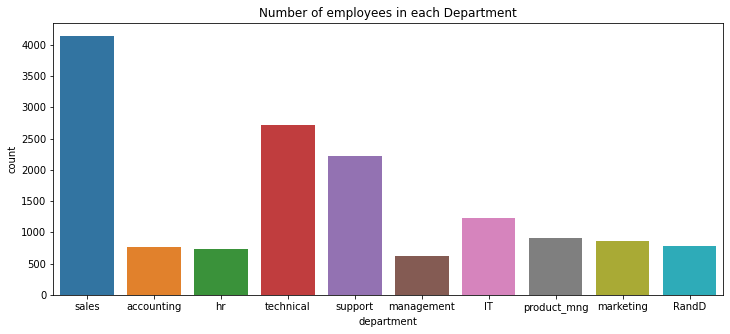
# 

# **Chapter 1: Exploratory Data Analysis**

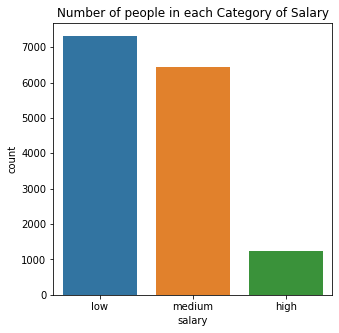


This data set consists of observations about 10 variables and 14,999 employees of the company. The first and foremost step of doing data analysis is **data cleaning**. For doing this, we consider various factors. Some of these include seeing the total number of observations that the data set has, checking the data types of variables involved, checking for null values and replacing them appropriately, correcting any spelling errors in the data set, seeing summary statistics and exploring data via bar charts. After doing a preliminary analysis, we find out that our data is free of any such errors as it has **no missing values**, duplicates and not many spelling errors. The names of some columns had errors which were corrected.

In addition to this, exploring the given data set via bar charts may help us understand the type of data we have. Some of the **bar charts** below help us see the department or salary range from which most employees in our data set are.

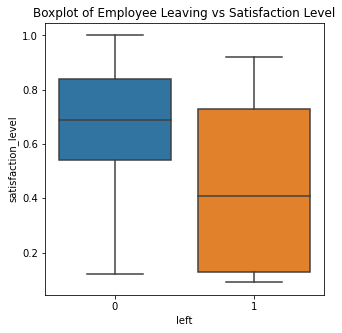


The above bar chart shows the number of employees working in each department. The Sales department has the greatest number of employees while the Management department has the least number of employees.

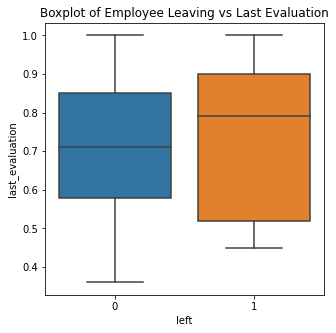


The bar chart above depicts the number of employees in each category of salary. There are a substantial number of employees with a low and medium salary but only a few employees with a high salary.

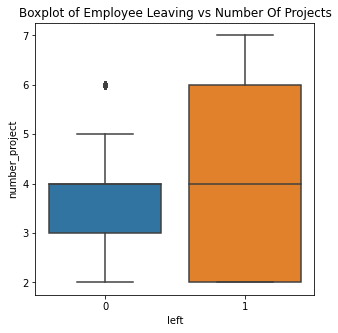
To analyze the summary statistics better, we take help from **box plots** by plotting the summary statistics of all variables relative to employees leaving the company. This helps see the breakdown of summary statistics (min, max, mean, 25th quartile and 75th quartile) of each variable for who left the company (indicated by 1) and those who did not leave the company (indicated by 0). Following are the various box plots that we drew, where the explanation for each of these figures is given below the boxplot.



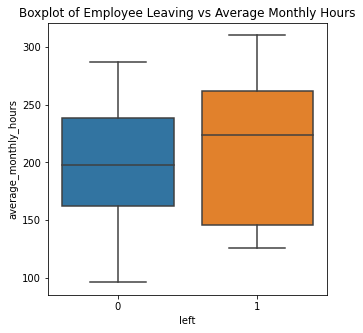
This box plot shows that the employees were less likely to leave if they were satisfied with their job.



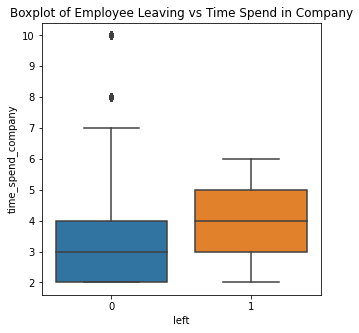
This box plot shows that the employees were more likely to leave if they were a valuable employee to the company and had been given a good evaluation score.



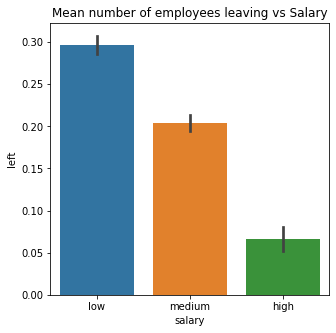
This box plot shows that the employees who were overburdened with projects were more likely to leave the company.



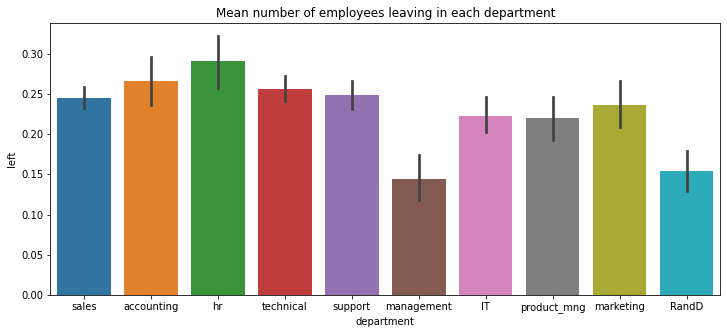
This box plot shows that the employees were more likely to leave if they were being asked to work extra hours.



This box plot shows that some employees were more likely to leave if they had been with the company for a long time but there were some extremely loyal employees who were with the company for a long time and did not think of leaving.



This bar chart shows that on average, employees were more likely to leave if they had a low salary. There was a greater chance that an employee would stay if his salary was relatively high.



In the bar chart above, the average number of employees that left each department can be seen. According to it, the HR department lost the highest number of employees while the Management department lost the least number of employees.

A screenshot of a cell phone

Description automatically generated

This bar chart above shows the average salary of employees with respect to each department. Here the employees in the management department received much better salaries on average than their counterparts. Therefore, it can be deduced that one of the reasons why the management department lost the least number of employees was due to its better pay package than other departments.

To see which factor most significantly impacts employee retention, a **correlation table** is made. This correlation table states the correlation values among all the variables in our data set. For more easily and quickly understanding this table, we make a **heat map** which can help analyze the correlation results visually. The colour scheme of the heat map suggests that the more darkly coloured a box is, the higher is the correlation between the two variables associated with that box.



According to the colour coded scheme in the figure above, the three most important factors in employee retention are satisfaction level, work accident and time spent in company.

Satisfaction level is negatively correlated with the left variable. The correlation value of -0.39 indicates that as the satisfaction level of an employee goes up by one unit, then the chances of that employee leaving the company go down by 39%. Similarly, work accidents are also negatively correlated with the left variable and reduce the chances of employees leaving the company by 15%. Hence, we can see that work accidents do not significantly increase the chances of employees leaving the company.

On the other hand, time spent in company is positively correlated with the left variable. The correlation value of 0.14 indicates that as time spent in a company by an employee goes up by one unit, then the chances of that employee leaving the company go up by 14%. Hence, we can see that more time spent in a company can greatly increase the chances of employees leaving the company.

Moreover, for seeing how our categorical variables (department and salary) affect our numerical variables, we shall make dummy variables for our categorical variables. After this, another correlation matrix is made for all the variables in our data set. The results of this are shown in the form of a heatmap below.

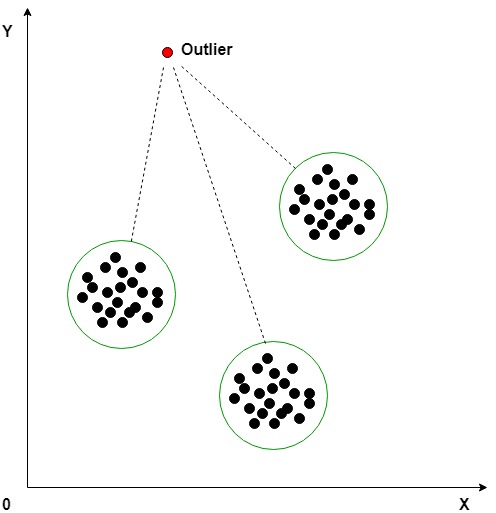


In the table above, we are primarily concerned with the upper right portion of the heatmap. This is where we have correlation values for the dummies of department /salary variables with other numerical variables in our data set. If we analyze the values in this part, they are very small and close to zero. Hence, we can deduce that our categorical variables have minimal effect on the numeric variables.

However, if we see how departments and salary are related to employees leaving the company, we can see that employees in the research and development, IT and management department are least likely to leave. This is because they have the largest negative correlation values. On the other hand, HR department has the largest positive correlation indicating that employees in the HR department are the ones most likely to leave. Similarly, when comparing how different categories of salaries affect employee retention, results show that higher salaries are also likely to lower the ratio of employees that end up leaving the company.

# 

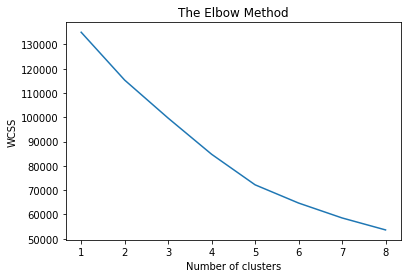
# **Chapter 2: Cluster and Outlier Analysis**



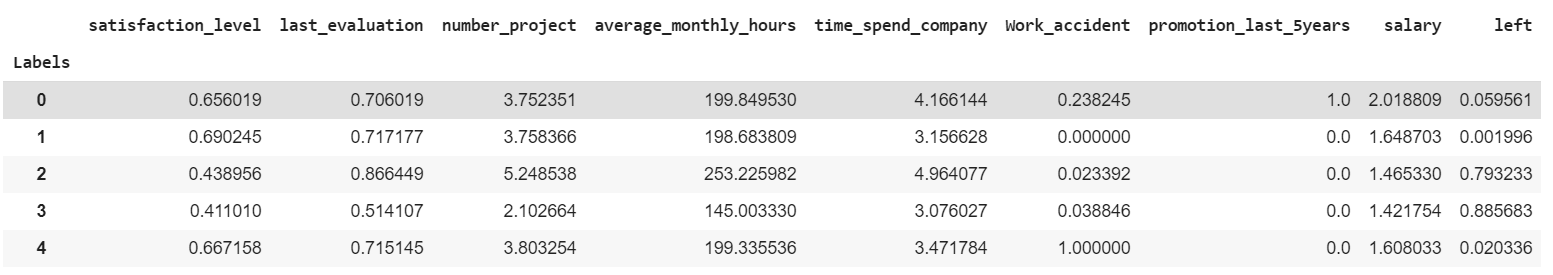
**Clustering Procedure**

To see further about which factors affect employee retention the most, we perform cluster analysis. For this purpose, we shall be using a partitional clustering method **(K-Means)** on our data set as it shall help group our data into clusters based on the various characteristics, i.e. the variables present in our dataset.

The optimal number of clusters is decided by using the **Elbow method** which looks at the total within-cluster sum of square (WSS) as a function of the numberofclusters. We shall be choosing the number of clusters so that adding another cluster does not improve much better the total WSS. According to the graph we obtained, we shall be using k=5, which is the value obtained at the kink.



The results for K-Means clustering are then obtained and our dataset is partitioned into 5 clusters. Label values from 0 to 4 are assigned to each of the 5 clusters. Afterwards, the resultant data values are grouped according to the labels assigned to them and the mean values for each group are obtained as follows:



**Clustering Results Analysis**

Looking at the results obtained in the table above, we can see that clusters with label ‘2’ and ‘3’ are the ones where on average the employees are most likely to leave, as the value for left variable is the highest. Similarly, clusters ‘0’, ‘1’ and ‘4’ are the ones where employees are not likely to leave as left variables has almost zero value. As cluster ‘1’ has the highest probability that an employee shall not leave and cluster ‘3’ has the greatest probability of an employee leaving, we shall be comparing these 2 clusters.

While **comparing cluster 1 with cluster 3**, we can see that the employees who leave the company have a greater mean level of satisfaction and salary. Moreover, they also have greater chances of work accident and a lower last evaluation (indicating that these employees were not regarded as being as capable by their employer).

On the other hand, when we **compare cluster 1 with cluster 2** (which also indicates employees that are likely to leave the company), we get to see an interesting trend. Here, the employees who left the company have on average a greater workload than those remain in the company. This is because these employees have a greater number of projects assigned to them and have greater average monthly working hours. Additionally, these employees are also amongst the oldest ones in the company as they have greater values for time spent in the company. The satisfaction level is low, and the work accident probability is greater among these people just like before. What is interesting is that despite working for more hours and on more projects, these employees have a lesser mean salary than those who eventually remain in the company. This finding can have significant policy implications for the company.

**Work accidents** do not seem to cause employees to leave the company. In fact, the clusters 0 and 4 where the employees are least likely to leave have the highest probability of work accidents - 23% and 100% respectively. Even though the clusters 2 and 3 have a greater chance for work accident as compared to 1, however, these values are very small and almost negligible as they are close to zero. Hence, we can see that work accidents are not a significant factor for employee retention.

**Outlier Procedure and Analysis**

Outliers are data points that are notably different from other data points. They are problematic as they can cause distortion in results. The **z-score metric** can be used to detect outliers in the data set. The z-score for every data point is calculated and if it is greater than 3, then this indicates that the data point is an outlier. This is because z-score is an indicator for standard deviations and 99.7% of the data lies within 3 standard deviations.

We calculated the z-scores for every variable and concluded that none of the variables contained any outliers except for the “time\_spend\_company” variable that had 376 outliers. Considering the size of our data, 376 outliers out of almost 15,000 observations does not seem very large. It's only a small fraction of the whole data and hence can be overlooked.

The z-score parameter was used to evaluate the outliers since it was more suitable for our data. Since our data set was large, other metrics like Interquartile Range (IQR) might not have performed as well as z-score did in detecting the outliers. IQR might even have wrongly over reported the outliers in the data.

# 

# **Chapter 3: Recommendations & Summary**



After performing a rigorous data analysis exercise on a data set consisting of about 15,000 yelp employees, the findings suggest that there is room for some vital policy alternatives to be implemented. These policy alternatives can help the company retain its capable employees and hence, can help improve both its productivity and profitability.

According to our results, satisfaction level is the foremost factor that can make or break an employee’s decision to leave the company. Hence, it should be the utmost focus of the company. For improving satisfaction levels, a variety of measures may be implemented. These might include carrying out surveys to see what the employees think they like or dislike about the company. Moreover, policies should be introduced that aim at improving the working experience for the employee. This might include introducing manageable work hours, recreational activities/trips, more off days, performance-based bonuses, along with numerous other measures.

Even though our findings suggest a not so significant relationship between work accidents and employee retention, but still a safer workplace should be ensured for employees in order to increase their satisfaction levels. Another important finding is that employees in the research and development, IT and management department are least likely to leave. Hence, targeted incentivized packages can be introduced for other departments as well to increase employee retention in them.

A crucial finding of our cluster analysis was that there are some employees that have a higher last evaluation by the company. But still these employees on average had lower salaries than those who were evaluated as being lower than them. This meant that these potential capable and experienced employees might have left the company due to monetary reasons. Hence, a proper performance appraisal system should be formulated and implemented to ensure that this does not happen. Furthermore, the employees that have been productively serving the company for many years should be given just reward for their services to the company.

**Recommendations Summary**



* Launch targeted incentivized packages for departments with lowest employee retention.
* Improved salary packages to increase satisfaction levels among employees.
* Ensure better workplace safety at work to reduce work accidents.
* Manage workload by regulating work hours of employees.
* Increase holidays and paid leaves via mutual understanding with employees.
* Organize team recreational activities or trips to help employees relax.
* Introduction of a vivid performance appraisal procedure to ensure performance is rewarded aptly.
* Tailored packages and incentives for older and loyal employees.