**DSCI 505: Project report**

Descriptive analysis using Python.

**Problem statement**

This project will use Python to conduct descriptive analysis on an HR employee dataset. The objective of this analysis is mainly to assist the business in understanding why its employees leave based on certain criteria such as the level of their salary, satisfaction, promotion, etc. The results of this analysis will shed light on those areas that the business needs to improve/focus on to retain their employees and as a result reduce costs associated with employee departure.

To solve the above problem, it is critical to have a deep understanding of how each factor is related to employees leaving the job. Furthermore, it’s important to quantify and visualize how each variable affects satisfaction level. As this HR dataset is mostly numeric data along with some categoric variables, Python is an appropriate programming language to achieve the above goal since it has powerful libraries such as NumPy and matplotlib that allow to effectively analyze numeric data through various relevant statistical measures such as mean and to visualize trends that are crucial to uncover relationships between variables. Finally, Python works better for large datasets like the HR employee dataset than other software like Excel, making it suitable for this problem.

Link to the dataset: <https://www.kaggle.com/datasets/colara/human-resource/data>

**Dataset description**

The data consists of 14999 observations and 11 variables/columns namely Emp\_ID, satisfaction\_level, last\_evaluation, number\_project, average\_monthly\_hours, time\_spend\_company, Work\_accident, left, promotion\_last\_5years, Department, salary. The datatypes and levels of measurement are as follows:

**Numerical variables- Float64 and interval**: satisfaction\_level and last\_evaluation.

**Numerical variables- Ratio and int64:** number\_project, average\_monthly\_hours, time\_spend\_company.

**Int64 and binary**: Work\_accident, left, and promotion\_last\_5years.

**Object- Categoric variables**: Salary (ordinal) and Department (nominal)

**Object- String**: Emp\_ID.

The dataset is complete and does not have missing values. The boxplot of the variables suggested that only time\_spend\_company variable has outliers as shown below:

A graph with a blue rectangle

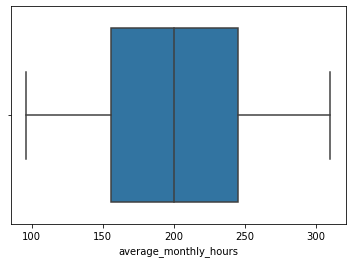
Description automatically generatedA blue rectangular object with white lines

Description automatically generatedA graph with a blue rectangle

Description automatically generated

A blue rectangular object with black lines

Description automatically generated



A graph with a rectangle and a line

Description automatically generatedFurther inspection revealed that there are 1282 outliers for that column. To handle them, a box plot showing ‘time\_spend\_company’ for those working in ‘management’ was created. The result can be seen on the right. This showed that although 6,7,8, and 10 are outliers when visualizing time\_spend\_company alone, only 10 is considered an outlier when visualizing it in the context of departments, specifically management. Furthermore, to confirm that 6,7, and 8 are valid values and not errors, a pivot table was created to show the salaries for each ‘department’. Then the mode function was used to see the most common salary level for managers. These were high and medium, which confirms that these managers likely spend more time in the company since they are paid well, suggesting that 6,7, and 8 are valid and should be included in the dataset.

To handle number 10, the mean satisfaction level for each time spent at the company was calculated at approximately 66% (0.655327). This was compared to the mean satisfaction level for all the employees which is at 61% (0.612834). The result was that those who spent 10 years achieved an above average satisfaction level. Hence the data was considered valid and was included in the analysis since it is expected for employees to stay longer if they are well satisfied.

**Descriptive analysis**

Firstly, almost all the variables are considered key variables since they tend to have a relationship with/impact employee turnover (leaving). Hence, the relationship between these and the ‘left’ variable was analyzed. The key descriptive statistics that were used were mean, median, mode, and measures of variability standard deviation. The below table provides an overview of the means for each variable.

|  |  |
| --- | --- |
| **satisfaction\_level** | **0.612834** |
| **last\_evaluation** | **0.716102** |
| **number\_project** | **3.803054** |
| **average\_monthly\_hours** | **201.050337** |
| **time\_spend\_company** | **3.498233** |
| **Work\_accident** | **0.144610** |
| **left** | **0.238083** |
| **promotion\_last\_5years** | **0.021268** |

Currently, about 24% (0.238083) of the *employees have left* the company.

***Satisfaction level and employee departure***

A graph of a graph

Description automatically generated with medium confidenceThe mean satisfaction level for employees who left the company is approximately 44% (below average) while for those who are in the company is approximately 67% (above average). The visual distribution are as follows:

A graph of a number of people

Description automatically generated

The above distribution shows that most of the employees who left the company had a below-average satisfaction level and those who did not leave mostly had an above-average satisfaction level. Additionally, for the employees who stayed, the distribution is left-skewed, suggesting that many of the employees had high levels of satisfaction. Although satisfaction level has a relationship with the left variable as shown, it is not the only reason. Some employees had below-average satisfaction levels but stayed, necessitating further analysis.

***Department, satisfaction level, and employee departure***

Next, it is crucial to see the distribution of satisfaction levels and the frequency of employees leaving each department to help the company effectively target those departments that are highly impacted. The mean satisfaction levels of each department were above the average except for ***accounting*** and ***HR*,** which were slightly lower at approximately 58% (0.582151) and 60% (0.598809) respectively. The ***technical*** department had a slightly below-average satisfaction level of 0.607897 or approximately 61%. Although the ***sales*** department had a slightly above average satisfaction level of 0.614447, it had the highest number of employees who left followed by ***technical*** and ***support***, suggesting that other factors also play a role in determining whether employees will stay or leave. The below chart shows the number of employees who left and those who stayed in each department.

A graph of blue and orange bars

Description automatically generated

Additionally, the standard deviation of the satisfaction level for each department was calculated and it can be summarized as follows:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| IT | RandD | accounting | hr | management | marketing | product\_mng | sales | support | technical |
| 0.249903 | 0.245269 | 0.255235 | 0.247929 | 0.227747 | 0.244280 | 0.242279 | 0.250251 | 0.246406 | 0.254193 |

The ***accounting*** department had the highest variability in satisfaction levels compared to the rest, followed by the ***technical*** and ***sales*** department. ***Management*** had the lowest variability, followed by the ***support*** and ***product management*** department. The departments with low variability are considered to have a distribution close to normal distribution since the values are clustered around the mean. For those with high variability, the data points are spread out from the mean, suggesting that some employees might have a substantially lower or higher satisfaction level compared to the mean value. Although sales have above average satisfaction levels, it has high variability. The measure of variability gives the business important insights especially when there is high variability. It allows the business to focus its efforts on those employees with lower satisfaction levels.

***Employee departure, number of projects, and last evaluation***

A graph of blue and orange bars

Description automatically generatedA graph of a number of blue squares

Description automatically generated with medium confidenceThe plot below (left) shows that there isn’t any noticeable trend between the number of projects and the number of employees who left/stayed. However, most of the employees who are currently in the company got engaged in 3 and 4 projects. Additionally, only a few of those who left worked on 3 projects while most of them worked on only 2 projects, suggesting that 3 projects are optimal. Furthermore, the plot is right skewed, implying that many of the employees work on few projects. Assigning only 2 projects causes the highest number of employee departures. Finally, the plot on the right shows that as expected, the employees who stayed tended to have high evaluation rates. Additionally, the distribution is left skewed suggesting that many of the employees had high evaluation/performance level.

Furthermore, as the plot below suggests, most of those who left had low evaluation percentages. However, a surprisingly large number of those with good evaluations were also among those who left, suggesting that there are other factors that can explain why employees leave.

A graph of a number of blue bars

Description automatically generated

Lastly, the standard deviation of the last evaluation for those who left is 0.197673 and for the other group is 0.162005.

***Salary, department, last evaluation, and employee departure***

The below plot shows that almost all the departments receive low/medium salary except for managers who receive medium/high salary. Low salaries can be one reason why ***sales*** have a slightly above average percentage of employees who leave the department. The sales department is at 0.244928. The median performance percentage for each department is slightly below/above the overall median performance percentage for all employees (72%) while for some departments it was on par. *However, the sales department had the lowest median evaluation percentage of 70% compared to other departments.*The support department has the highest median performance percentage (74%).

A graph of different colored bars

Description automatically generated

***Work accident and employee departure***

The below count plot shows that work accidents are not one of the reasons why employees leave because those who left had lower work accidents than those who stayed.

A graph with blue and orange bars

Description automatically generated

A graph of a number of hours

Description automatically generatedA graph of blue and white bars

Description automatically generated***Average monthly hours and employee departure***

The above chart (left) illustrates that although employees who work for long average hours leave the company, it’s evident that most of those who left were those who worked *below 160 hours*, implying they might be underworked. The chart on the right shows that many of those who stayed tend to work more average hours, suggesting that lack of work-life balance is not a reason why employees leave. Additionally, the plot suggests that the data is close to normal distribution. Finally, the standard deviation for each group is 45.682731 for those who stayed and 61.202825 for those who left.

**Summary**

After analyzing the relationship between the relevant variables, the key findings can be summarized as follows:

* Employees with a low satisfaction level had the highest departures.
* The business is advised to focus on improving its sales department since comparatively, it had the lowest employee retention rate and performance level as well as a high variability in satisfaction level.
* Accounting has the highest variability in satisfaction, suggesting some employees might have dramatically high/low satisfaction levels. Hence, the business should also focus on stabilizing it.
* Management had the lowest variability and turnover.
* The business should aim to assign 3 projects to reduce the turnover rate.
* Most employees who left had low performance and vice versa.
* The company is losing many of its high-performing employees.
* Most employees receive low/medium salaries while managers receive medium/high salaries.
* It is recommended that employees work above 160 hours and below 210 hours on average per month since these had a lower number of employees who left and a high number of those who stayed.
* Overall, satisfaction level, average monthly hours, and the number of projects present potential reasons as to why employees stay/leave.

**Results**

The analysis revealed interesting information about the data, it allowed me to understand which variables had a relationship with the **left** variable, which gave me insights about how the key variables were linked with whether employees stayed or left. Additionally, the dataset presented the complexities associated with understanding why exactly employees leave. For instance, although it’s expected dissatisfied employees to leave, this dataset showed that some employees chose to stay.

This exercise allowed me to have hands-on experience with working with a large dataset. It allowed me to gain a deeper understanding of how functions work together to present useful information in data analytics. One such function is the pivot () which helped me create datasets containing parts of the HR dataset to analyze and visualize. Furthermore, since this dataset had outliers, it allowed me to carefully think about their inclusion/exclusion, allowing me to improve my analytical skills and become a better data analyst.

Some of the biggest challenges were in deciding whether to include or delete outliers. However, with some analysis, it was clear that those were valid data. Since the dataset had both qualitative and quantitative variables which require different ways of visual representation, one of the challenges was selecting appropriate visuals for the data to clearly show any hidden trends/information.

Python is a great tool to use for this kind of analysis as it has many benefits. Python has useful libraries, such as pandas, matplotlib, and seaborn, that are usually used for this type of analysis. It allows for effective analysis and visualization of the data which are crucial for descriptive analysis.