SERIAL NO CHAPITER PAGE NO

Chapter 1 Business Problem 2

Chapter 2 Business question 2

Chapter 3 Hypotheses test 2

Chapter 4 Data collection and exploration 2

4.1. Codes and output 3

Chapter 5 Feature engineering 6

5.1 Codes and output 6

Chapter 6 Chi square test 7

6.1 Satisfaction level impact on employee’s turnover 7

6.2 Salary impact on employee’s turnover 7

Chapter 7 Prediction model 7

7.1 Codes and output 8

Chapter 8 Conclusion 9

Chapter 9 GitHub repository link 10

## **Business Problem:**

The company is concerned about the high number of employees leaving. The HR department want to identify the factors causing employees to leave the company and make a statistic model to predict which employees will likely leave.

## **Business question:**

* What are the factor impacting employees decision when leaving the company
* does satisfaction level have a high impact on employees leaving the company
* does salary have a high impact on employees leaving the company

## **Hypotheses test:**

This project has two hypotheses to be tested

first

* H0: satisfaction level does not have a significant impact on employees leaving the company
* H1: satisfaction level has a significant impact on employees leaving the company

Second

* H0: salary does not have a significant impact on employees leaving the company
* H1: salary have a significant impact on employees leaving the company

## **Data collection and exploration:**

Dataset use for this project was collected by the HR department of the company. It is compose of 8 numerical column and two categorical column which are satisfaction level, last evaluation , number of project , average monthly hours, time spend company, work accident , left, promotion last 5 years for the numerical data and salary, department for the categorical data

**Link to dataset:** [**HR Analytics (kaggle.com)**](https://www.kaggle.com/datasets/giripujar/hr-analytics?resource=download)

* I assigned the dataset to a variable called “df”.
* I now checked the number of missing values contained in the dataset and the answer is 0
* After that I used the function summary to summarize the information about the dataset. By using Describe which gives the number of data point, the mean ,standard deviation, minimum values, maximum values, the kurtosis and other descriptive statistics about each column of the data set
* Convert salary to numeric
* Make descriptive statistic visualization of satisfaction level, salary and left

## ***Codes and output***



A white rectangular object with a black border

Description automatically generated

A screenshot of a computer screen

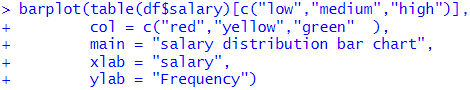
Description automatically generated

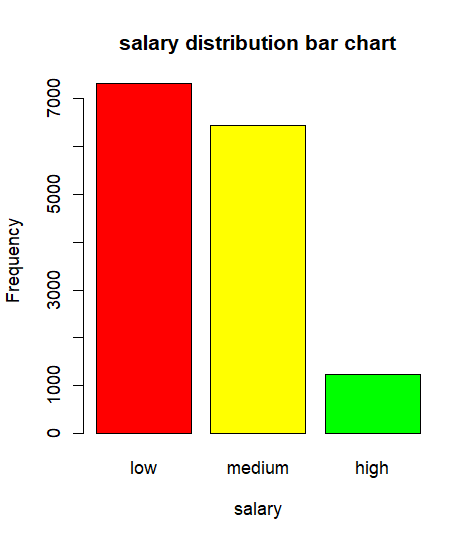
Visualise the distribution of satisfaction level, salary and left by using box plot for satisfaction level and bar chart for both salary and left column



A diagram of a box with a blue and black line

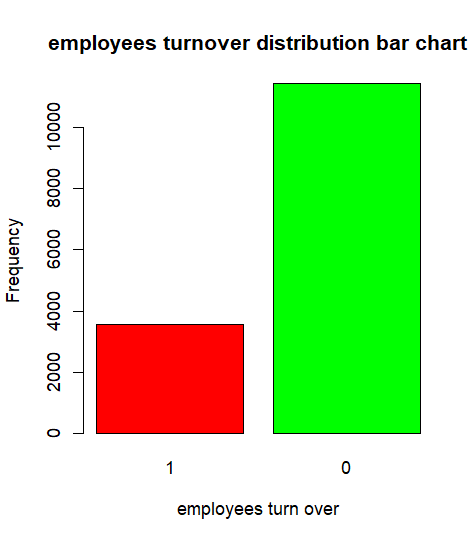
Description automatically generated





A white background with blue text

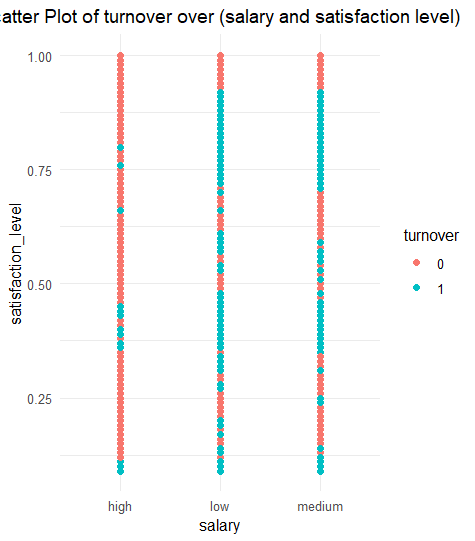
Description automatically generated



Make a scatter plot to check how employee’s turnover is distributed per salary level and satisfaction level. We observe that high salary employees tend to leave the company and there is no real relationship between satisfaction level and turnover

A computer code with red and blue text

Description automatically generated



Calculate the mean of average satisfaction of employees which leave vs employees which stay

A computer code with numbers and symbols

Description automatically generated

## **Feature engineering :**

* Convert salary to numerical by assigning 1 for low , 2 for medium and 3 for high because it is an ordinal data.
* Use one hot encoding for department as it is categorical
* Create dataset to make chi square test

## ***Codes and output***

Convert salary into numerical by assigning 1 for low 2 for medium and 3 for high

A close-up of a computer code

Description automatically generated

Make a matrix which have the number of low and high satisfaction level that leave and stay for chi square test. High satisfaction is satisfaction greater than 0.5 and low is satisfaction lower than 0.5

A computer screen shot of text

Description automatically generated

Make a matrix which have the number of low, medium and high salary that leaves and stays for chi square test .

A computer screen with blue text

Description automatically generated

## **Chi square test:**

* 1. **Satisfaction level impact on employee’s turnover**

I applied pearson chi-squared on the satisfaction level vs leave matrix and I got a p-value of 2.2e-16 which is less than 0.05 and a high X-squared this mean that the null hypothesis have to be rejected . in conclusion satisfaction level has a significant impact on employees leaving the company

.

A white background with black text

Description automatically generated

* 1. **Salary impact on employee’s turnover**

I applied pearson chi-squared on the salary vs leave matrix and I got a p-value of 2.2e-16 which is less than 0.05 this mean that the null hypothesis have to be rejected . in conclusion salary have a significant impact on employees leaving the company.

A black text on a white background

Description automatically generated

## **Prediction model :**

For model prediction I choose logistic regression as it is suitable for binary classification and easy to interpret.

* Split the data set into train and test
* Train the model
* Summarize the model
* Predict test data
* Evaluate the model

## ***Codes and output***

A screenshot of a computer program

Description automatically generated

I split my data into train and test where train was 80% of the data and test is 20% then I train my model to predict left base on salary and satisfaction level after doing that I summarized the model and this show that there is a negative relationship between both feature and turnover that is high salary employees and employees with high satisfaction level are more likely to leave the company it also show that satisfaction level is the factor that have the greatest impact on employees turnover its coefficient is -4.

A screenshot of a computer program

Description automatically generated

Now I tested my model on the test data and I changed the probabilities to binary as this task is a binary classification finally I created a confusion matrix with some metric to evaluate my model. For this task the most important metric is recall as it is more crucial for the company to determine employees that are likely to leave the company my model have a recall of 0.92 which is a very decent score this mean that my model can predict 92% of the employees that will be likely to leave the company and the second most important metric is the f1 score as it show how balance is the performance of my model, my model have a f1 score of 0.87 which is a very descent score .

## **Conclusion:**

In conclusion having this model will be very beneficial for the company as it will help them predict early the employees which are likely to leave the company .so now the company have more time to look for a backup plan and also the company knows in advance which employees to target to make them stay. The coefficient of each factor of the model also provide the company with enough information to know which factors have to be improve to make employees stay. The coefficient also chose that the satisfaction level is the strongest feature but it have a negative relationship with employees turnover this may mean that employees willing to leave the company are likely going to give a high rating because of lack of concern in the company no more so to improve that a survey format may be a better idea than a rating format.

## **GitHub repository link:** https://github.com/juniorskg/turnover\_statistical\_modelling

.