## STEPS:

1. Objective
2. Introduction
3. Data for analysis -- list all the variables
4. EDA analysis →
   1. Structure of dataset, then summary
   2. Univariate plot analysis → histogram, box plot
   3. Bivariate plot analysis → Scatter plot, Correlation matrix
   4. Multivariate plot analysis
   5. Final plot
   6. Correlation matrix
5. Final conclusion

→ This was just a part of assignment

# EDA

## Objective

The objective of the assignment is to use statistical techniques for the given dataset HR and find the reason behind the leaving of the employees. For that purpose, need to build a model which will help to predict either an employee will stay in the company or leave it. The main aim is to build model which will best predict that employees will leave the company or not.

## Introduction

In a company the predicting whether an employee will leave the company or not will depend on many factors. As the reason may be different for different for different person. The task is to build a model which will help to predict this based on several other reasons. To avoid the leaving of the employees, the model will help to predict.

## Data for analysis

There 10 variables and 14999 observations are present. Following are the variables present in the dataset:

* Satisfaction Level
* Last evaluation
* Number of projects
* Average monthly hours
* Time spent at the company
* Whether they have had a work accident
* Whether they have had a promotion in the last 5 years
* Departments (column sales)
* Salary
* Whether the employee has left

10 Variables 14999 Observations

-----------------------------------

satisfaction\_level

n missing distinct

14999 0 92

Info Mean Gmd (Gini's mean difference)

1 0.6128 0.2823

.05 .10 .25

0.11 0.21 0.44

.50 .75 .90

0.64 0.82 0.92

.95

0.96

lowest : 0.09 0.10 0.11 0.12 0.13

highest: 0.96 0.97 0.98 0.99 1.00

-----------------------------------

last\_evaluation

n missing distinct

14999 0 65

Info Mean Gmd

1 0.7161 0.1973

.05 .10 .25

0.46 0.49 0.56

.50 .75 .90

0.72 0.87 0.95

.95

0.98

lowest : 0.36 0.37 0.38 0.39 0.40

highest: 0.96 0.97 0.98 0.99 1.00

-----------------------------------

number\_project

n missing distinct

14999 0 6

Info Mean Gmd

0.945 3.803 1.367

Value 2 3 4 5

Frequency 2388 4055 4365 2761

Proportion 0.159 0.270 0.291 0.184

Value 6 7

Frequency 1174 256

Proportion 0.078 0.017

-----------------------------------

average\_montly\_hours

n missing distinct

14999 0 215

Info Mean Gmd

1 201.1 57.48

.05 .10 .25

130 137 156

.50 .75 .90

200 245 267

.95

275

lowest : 96 97 98 99 100

highest: 306 307 308 309 310

-----------------------------------

time\_spend\_company

n missing distinct

14999 0 8

Info Mean Gmd

0.905 3.498 1.43

Value 2 3 4 5

Frequency 3244 6443 2557 1473

Proportion 0.216 0.430 0.170 0.098

Value 6 7 8 10

Frequency 718 188 162 214

Proportion 0.048 0.013 0.011 0.014

-----------------------------------

Work\_accident

n missing distinct

14999 0 2

Info Sum Mean

0.371 2169 0.1446

Gmd

0.2474

-----------------------------------

left

n missing distinct

14999 0 2

Info Sum Mean

0.544 3571 0.2381

Gmd

0.3628

-----------------------------------

promotion\_last\_5years

n missing distinct

14999 0 2

Info Sum Mean

0.062 319 0.02127

Gmd

0.04163

-----------------------------------

sales

n missing distinct

14999 0 10

accounting (767, 0.051), hr (739,

0.049), IT (1227, 0.082),

management (630, 0.042), marketing

(858, 0.057), product\_mng (902,

0.060), RandD (787, 0.052), sales

(4140, 0.276), support (2229,

0.149), technical (2720, 0.181)

-----------------------------------

salary

n missing distinct

14999 0 3

Value high low medium

Frequency 1237 7316 6446

Proportion 0.082 0.488 0.430

-----------------------------------

## Exploratory Data Analysis

str(dataset)

'data.frame': 14999 obs. of 10 variables:

$ satisfaction\_level : num 0.38 0.8 0.11 0.72 0.37 0.41 0.1 0.92 0.89 0.42 ...

$ last\_evaluation : num 0.53 0.86 0.88 0.87 0.52 0.5 0.77 0.85 1 0.53 ...

$ number\_project : int 2 5 7 5 2 2 6 5 5 2 ...

$ average\_montly\_hours : int 157 262 272 223 159 153 247 259 224 142 ...

$ time\_spend\_company : int 3 6 4 5 3 3 4 5 5 3 ...

$ Work\_accident : int 0 0 0 0 0 0 0 0 0 0 ...

$ left : int 1 1 1 1 1 1 1 1 1 1 ...

$ promotion\_last\_5years: int 0 0 0 0 0 0 0 0 0 0 ...

$ sales : Factor w/ 10 levels "accounting","hr",..: 8 8 8 8 8 8 8 8 8 8 ...

$ salary : Factor w/ 3 levels "high","low","medium": 2 3 3 2 2 2 2 2 2 2 ...

## Univariate Analysis

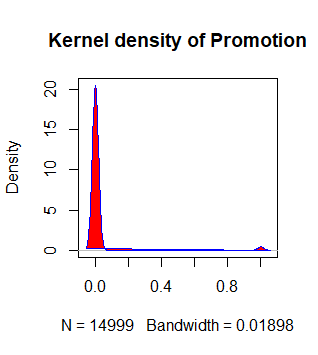


Fig: Kernel density plot of promotion

From the above plot, we can conclude that there are very less number of promotions in the company. As the plot is peaked near the region of 0.0 and there is very less density around 0.9.

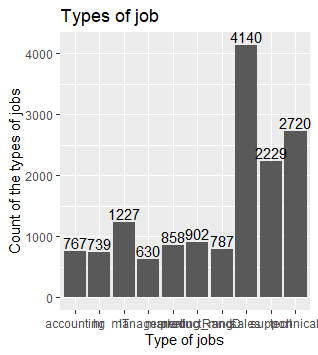


Fig: Types of jobs in the company and their count

From the above, I conclude that sales job is the job where many number of people work. So, top three types of job where larger number of people work are sales, support and technical.

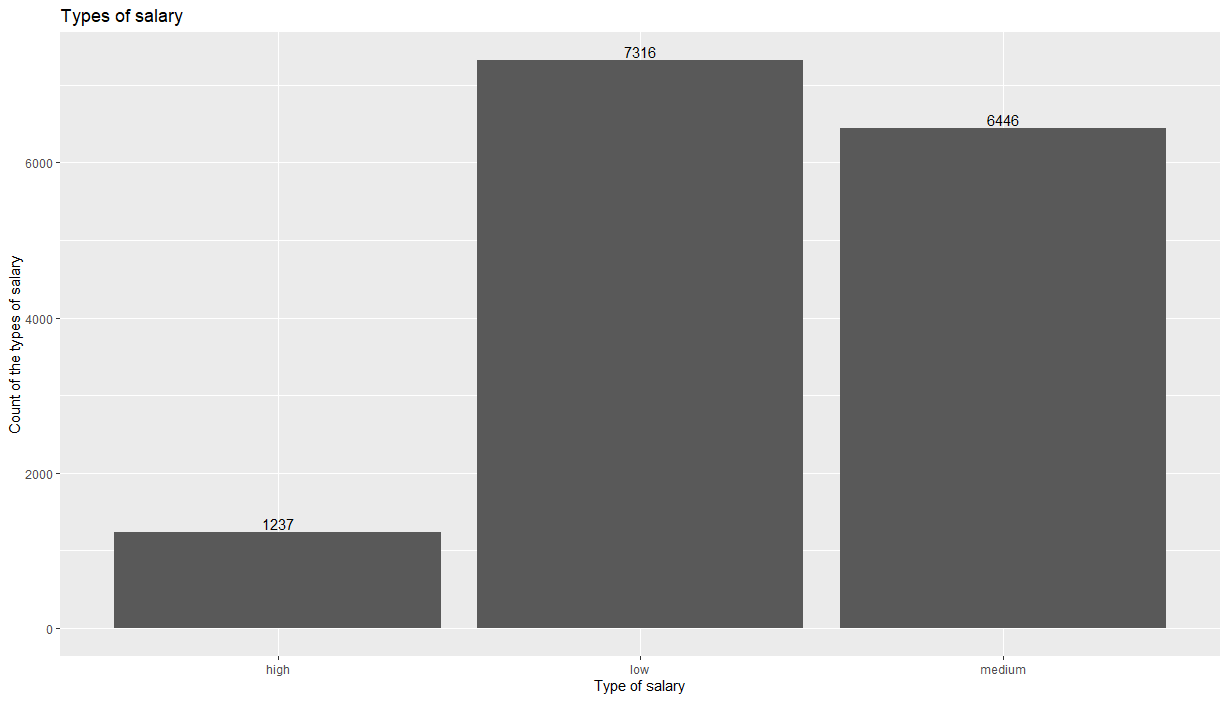


Fig: Count of the types of salary given to the employee

From the above plot, I can conclude that larger number of people are working under low salary and very less number are paid high by the company.

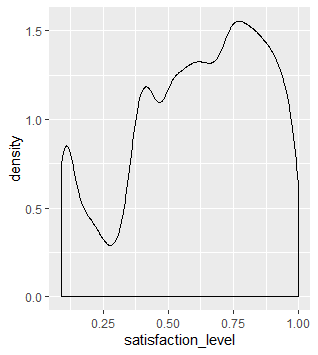


Fig: Density plot of satisfaction level

From the above, I can conclude that more number of people are satisfied and also there is fall near 0.25 to 0.75, which means there are a group of people who are not satisfied with their job.

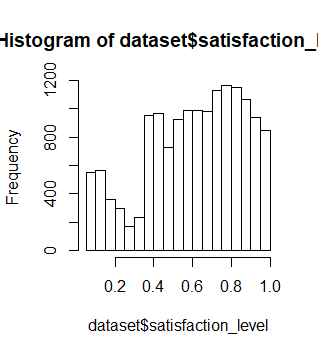


Fig: Histogram plot of Satisfaction level

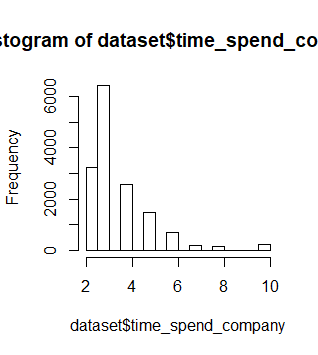


Fig: Histogram plot of time spend in the company

From the above plot, I can infer that a large number of people are spending less time in the company. Thus, the most common time spend in the company is in between 2 and 4.

## Bivariate/Multivariate Analysis

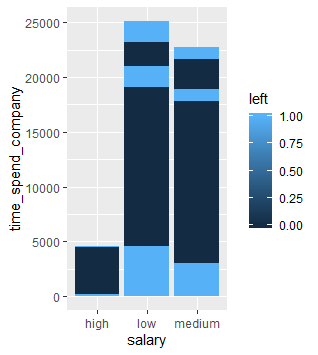


Fig: Relation between salary, time spent in the company and people left.

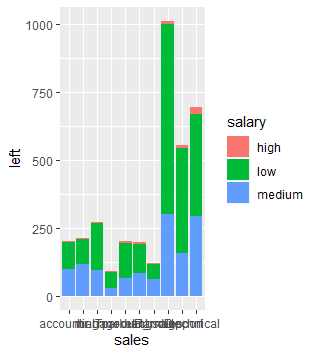


Fig: Relation between sales, left and salary

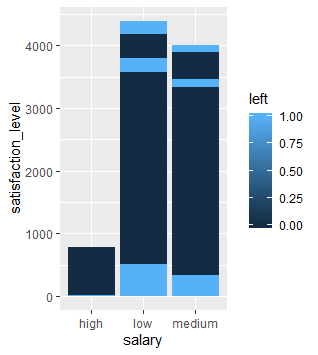


Fig: Relation between salary, satisfaction level and left

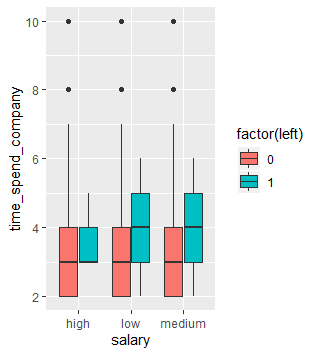


Fig: Relation between salary, time spend in company and left

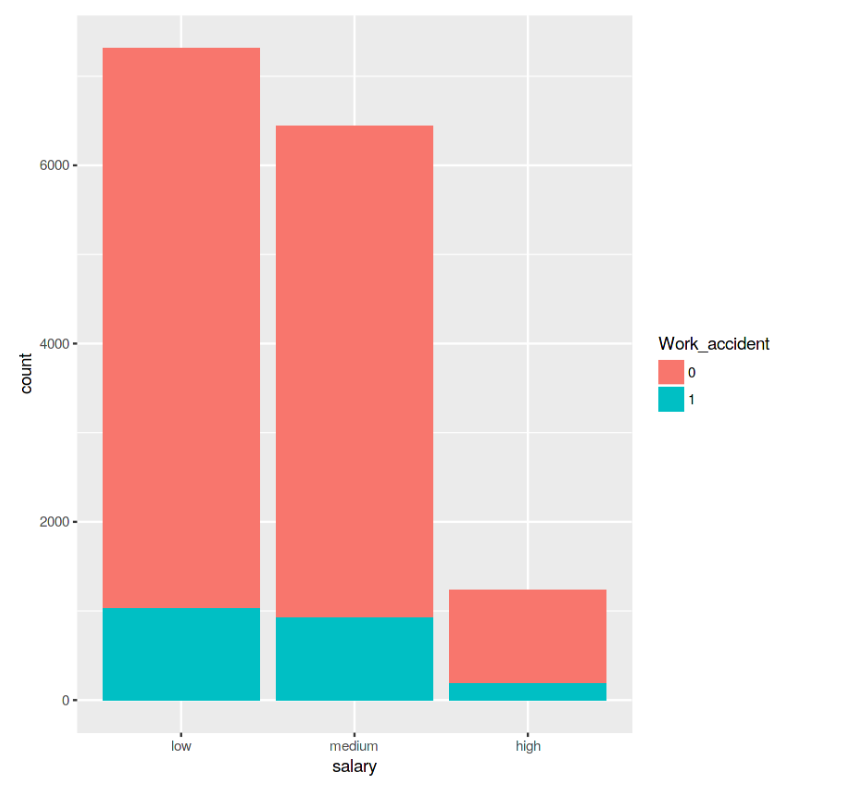


Fig: Relation between the count of types of salary and work accidents

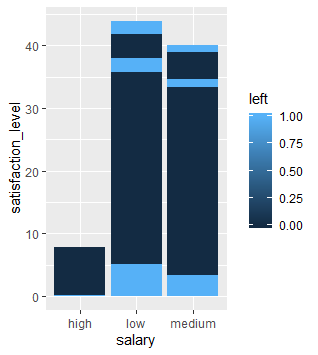


Fig: Relation between salary, satisfaction level and left

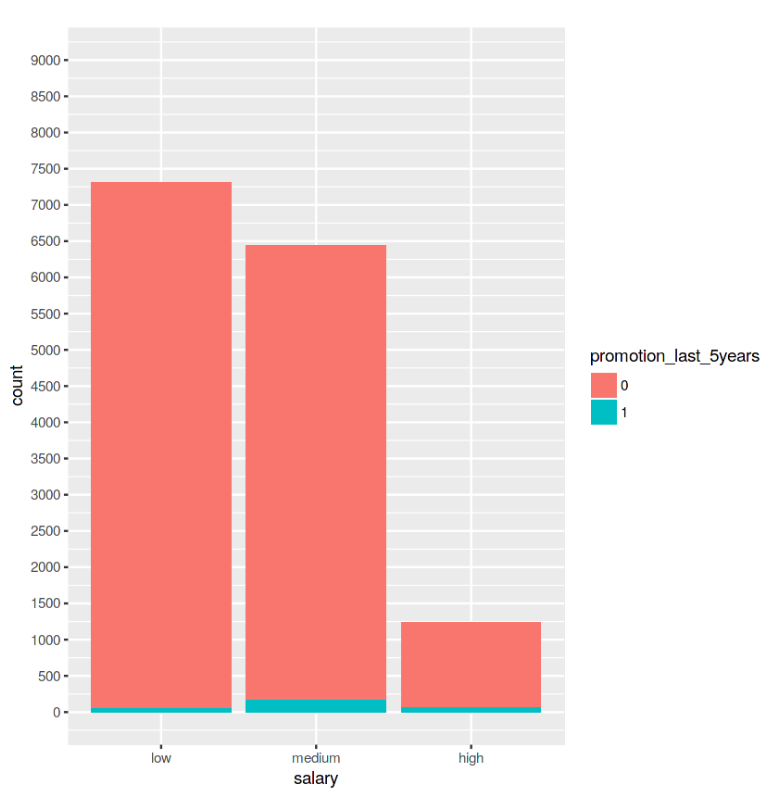


Fig: Relation between and salary and promotion last 5 years

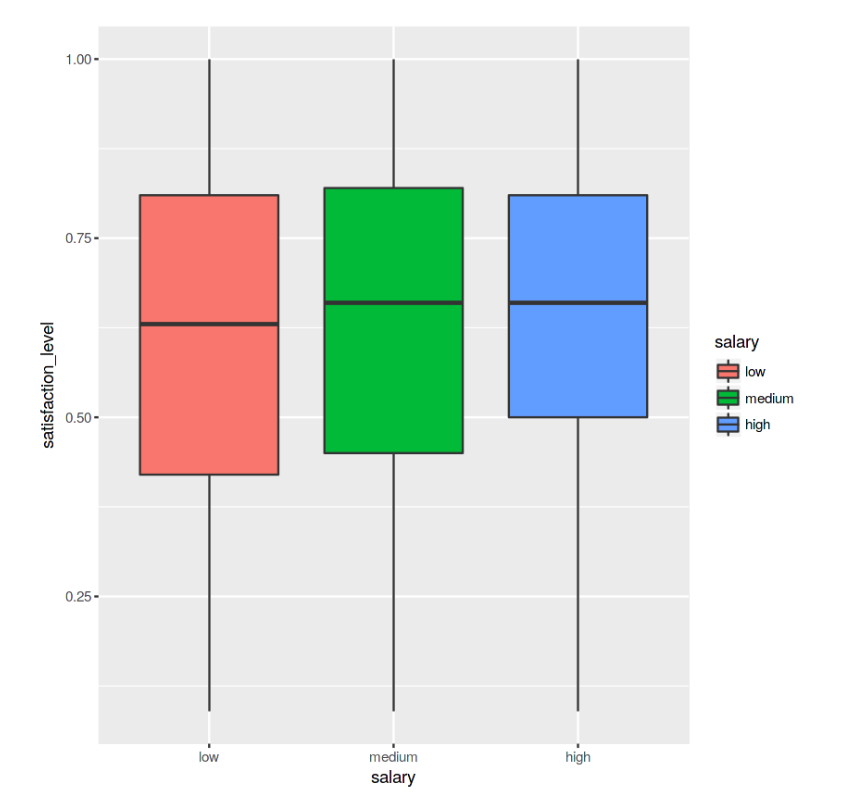


Fig: Relation between satisfaction level and types of salary

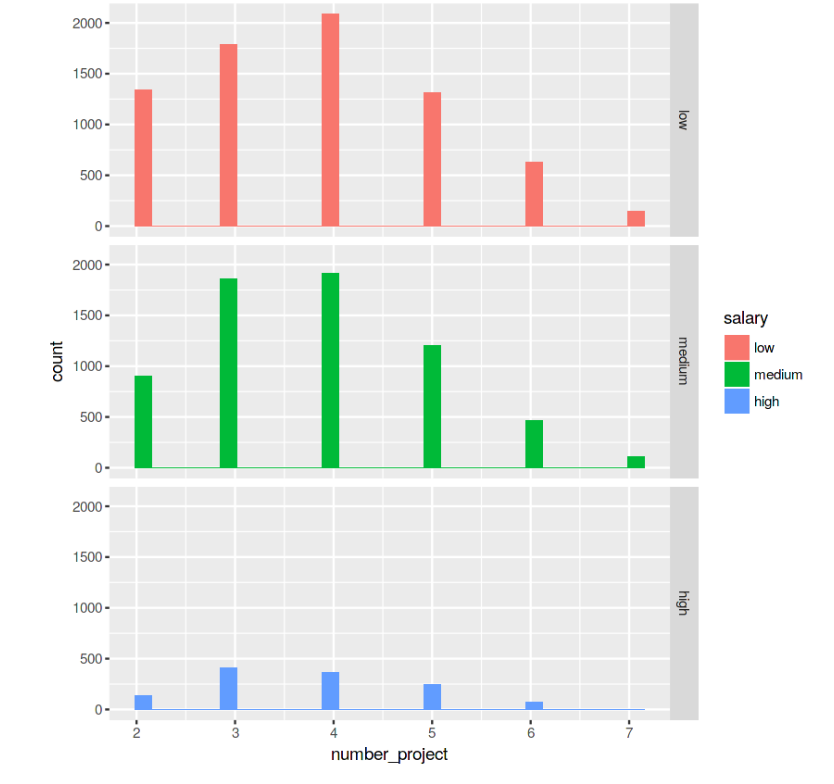


Fig: Relationship between the number project and handled by different salary group.

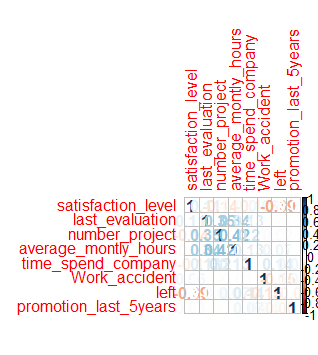


Fig: Correlation table

In this table we can see that there are many features which are not positively or negatively related.

**Conclusion**

From above univariate, bivariate and multivariate analysis we can conclude that Salary, Satisfaction\_level, Number\_project, Work\_accident, Promotion\_last\_5years are the major factors responsible for the people leaving their job. The average Satisfaction level of Medium and High salary employees is a bit higher compared to Low salary employees. As employees who have low salary are working for longer duration in comparison to those employees who have a higher salary. And also the majority of work accidents are seen among Low and Medium salaried employees. Also, low Salary employee category have least promotions in last 5 years.

So, we have to consider all these features for building our model to predict that employees will stay or leave the job.