EDA for HR dataset

getwd()

dataset = read.csv("HR.csv")

#I check that there are 14999 observations in the dataset which have 10 variables

View(dataset)

#I can see that there are 10 variables along with their data types in the given dataset

[Plot Analysis](https://drive.google.com/open?id=1f22eKWIIr9rQbgklYeek6_DPgrmydhbLb-eE92CiRSY)

plot\_str(dataset)

#It will show the five point summary of each of the column along with their mean for numerical values

summary(dataset)

#Observing this I can say that there are 9 types of sales job in the given dataset

table(dataset$sales)

head(dataset, n=5)

glimpse(dataset)

#Seeing this I can say that there are no missing data in the given dataset

sum(is.na(dataset))

#Univariate analysis

dim(dataset)

str(dataset)

[Plot Analysis](https://drive.google.com/open?id=1f22eKWIIr9rQbgklYeek6_DPgrmydhbLb-eE92CiRSY)

hist(dataset$satisfaction\_level)

hist(dataset$last\_evaluation)

hist(dataset$number\_project)

hist(dataset$average\_montly\_hours)

hist(dataset$time\_spend\_company)

hist(dataset$Work\_accident)

hist(dataset$left)

hist(dataset$promotion\_last\_5years)

d = density(dataset$promotion\_last\_5years)

plot(d, main = "Kernel density of Promotion")

polygon(d, col = "red", border = "blue")

[Plot Analysis](https://drive.google.com/open?id=1f22eKWIIr9rQbgklYeek6_DPgrmydhbLb-eE92CiRSY)

plot(sales ~ group, data = dataset)

#hist(dataset$sales)

#hist(dataset$salary)

library(ggplot2)

ggplot(data = dataset , aes(x=sales)) + geom\_histogram(color="black", fill="WHITE", bins = 10, stat = "count")

ggplot(data = dataset , aes(x=salary)) + geom\_histogram(color="black", fill="WHITE", bins = 10, stat = "count")

[Plot Analysis](https://drive.google.com/open?id=1f22eKWIIr9rQbgklYeek6_DPgrmydhbLb-eE92CiRSY)

ggplot(dataset, aes( x = satisfaction\_level)) + geom\_density( )

ggplot(dataset, aes(x=left , color=satisfaction\_level)) + geom\_density( )

p = ggplot(dataset, aes(x= salary)) + geom\_bar()

p = p + labs(title = "Types of salary", x = "Type of salary", y = "Count of the types of salary")

p + geom\_text(stat='count', aes(label=..count..), vjust=-0.25)

#Bivariate

dataset$satisfaction\_level = as.numeric(dataset$satisfaction\_level / 100)

p2 <- ggplot(data=dataset, aes(x=salary, y=sales, fill=left))

p2 + geom\_bar(stat = "identity")

ggplot(dataset, aes(salary, time\_spend\_company, fill = factor(left))) + geom\_boxplot()

The plot and analysis of the plot

[Plot Analysis](https://drive.google.com/open?id=1f22eKWIIr9rQbgklYeek6_DPgrmydhbLb-eE92CiRSY)

Ans.

After all the analysis, I found that there is a relation between the people who are working more and leaving the job. There are other factors also, as the people who are paid less or medium salary.