**R Programming Assignment**

**Nukavarapu Lalita Aditya Patel**

**MBA (BA. AI. ML)**

**2002020**

**# Reading the GitHub file**

url <- "https://github.com/SavioSal/datasets/raw/master/Bank%20Churn\_Modelling.csv"

bank\_churn\_modelling <- read\_csv(url) # reads the url file

view(bank\_churn\_modelling) # using this command gives us an overview towards what is present in the downloaded dataset

str(bank\_churn\_modelling)

names(bank\_churn\_modelling) # this command will help us in knowing what the names of columns are present in the dataset.

head(bank\_churn\_modelling) # using this command will display the data that is present in the dataset

**# Data Cleaning**

bank\_churn\_modelling <- bank\_churn\_modelling %>%

dplyr::select(-RowNumber, -CustomerId, -Surname) %>% #removes the unwanted columns

mutate(Geography = as.factor(Geography),

Gender = as.factor(Gender),

HasCrCard = as.factor(HasCrCard),

IsActiveMember = as.factor(IsActiveMember),

Exited = as.factor(Exited),

Tenure = as.factor(Tenure),

NumOfProducts = as.factor(NumOfProducts))

**# Checking for Missing and Null Values**

sapply(bank\_churn\_modelling, function(x) sum(is.na(x)))

**# Data Overview**

summary(bank\_churn\_modelling)

**1. Develop 5 different visuals using GGPLOT with descriptions of the insights they convey. 10m.**

**### Categorical Distribution of the variables from the given Dataset**

bank\_churn\_modelling %>%

dplyr::select(-Exited) %>%

keep(is.factor) %>%

gather() %>%

group\_by(key, value) %>%

summarize(n = n()) %>%

ggplot() +

geom\_bar(mapping=aes(x = value, y = n, fill=key), color="black", stat='identity') +

coord\_flip() +

facet\_wrap(~ key, scales = "free") +

theme\_minimal() +

theme(legend.position = 'none')

Chart, bar chart

Description automatically generated

From the above plotting’s I have observed the points which are mentioned below as follows,

1. Here from the above plotting’s it was clear that there were more number of male customers than female customers.
2. There are more number of customers from France when compared with Germany.
3. From the above plot it is noted that most of the customers utilizes and had access to bank’s credit card services.
4. From the above plot’s, there is one important thing which I have noticed is there are equal number of active as well as non-active customers.

**### Continuous Distribution of the Variables from the given dataset**

bank\_churn\_modelling %>%

keep(is.numeric) %>%

gather() %>%

ggplot() + geom\_histogram(mapping = aes(x=value,fill=key), color="black") +

facet\_wrap(~ key, scales = "free") + theme\_minimal() + theme(legend.position = 'none')

Chart, histogram

Description automatically generated

From the above histogram plotting, it is observed that most credit scores of the customers is above 600 and there are high chances of most customers will leave the bank.

**### Who are the churned customers and non-churned customers**

ggplot(bank\_churn\_modelling, aes(Exited, fill = Exited)) +

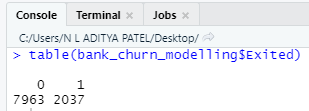
geom\_bar() +

theme(legend.position = 'none')

Chart, bar chart

Description automatically generated

table(bank\_churn\_modelling$Exited)



From the above bar graphs and above image, we can clearly see that the churn percentage of customers in the bank is very low which means more customers did not churn from the bank.

**### Finding of the Correlation Matrix for Continuous Variables**

numericVarName <- names(which(sapply(bank\_churn\_modelling, is.numeric)))

corr <- cor(bank\_churn\_modelling[,numericVarName], use = 'pairwise.complete.obs')

ggcorrplot(corr, lab = TRUE)

The correlation matrix image is attached in the below and from the attached correlation matrix it is observed that there is no high correlation between continuous variables.

Chart

Description automatically generated

**### which age group category of customers fall under more churn**

ggplot(bank\_churn\_modelling, aes(Exited, Age, fill = Exited)) +

geom\_boxplot() +

theme\_minimal() +

theme(legend.position = 'none')

Chart, box and whisker chart

Description automatically generated

From the above boxplot, it was observed that churned customers lies between the age group 40 – 50 whereas non-churned customers tend to be young as they because in the boxplot it is observed that they lies in between 30 – 40.

**### Account balance comparison between churned customers and non - churned customers**

ggplot(bank\_churn\_modelling, aes(x = Exited, y = Balance, fill = Exited)) +

geom\_boxplot() +

theme\_minimal() +

theme(legend.position = 'none')

Chart, box and whisker chart

Description automatically generated

From the above boxplot, it is observed that churned customers had good bank balance in their accounts when compared with the non – churned customers who belong to the same banks and it is also noted that non – churned customers doesn’t have good and adequate balance in their bank accounts.

**### Estimated Salary of churned and non - churned customers**

ggplot(bank\_churn\_modelling, aes(x = Exited, y = EstimatedSalary, fill = Exited)) +

geom\_boxplot() +

theme\_minimal() +

theme(legend.position = 'none')

Chart, box and whisker chart

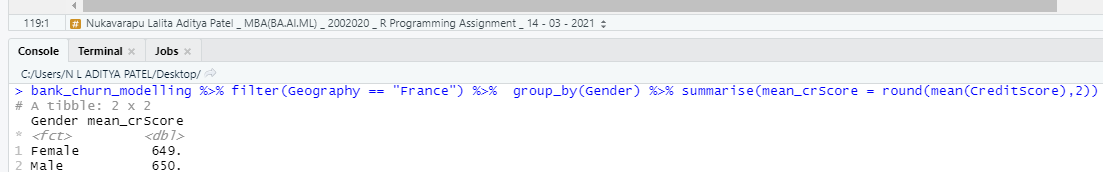
Description automatically generated

From the above box plotting it was observed that both churned customers as well as non – churned customers are having approximate level of salary.

**2. Develop answers to the following. Use dplyr wherever necessary: 10m.**

**A. What is the average credit score of females and males in France?**

bank\_churn\_modelling %>% filter(Geography == "France") %>% group\_by(Gender) %>% summarise(mean\_crScore = round(mean(CreditScore),2))



**B. What is the average credit score of people in the age brackets 20-30,31-40,41-50?**

bank\_churn\_modelling %>%

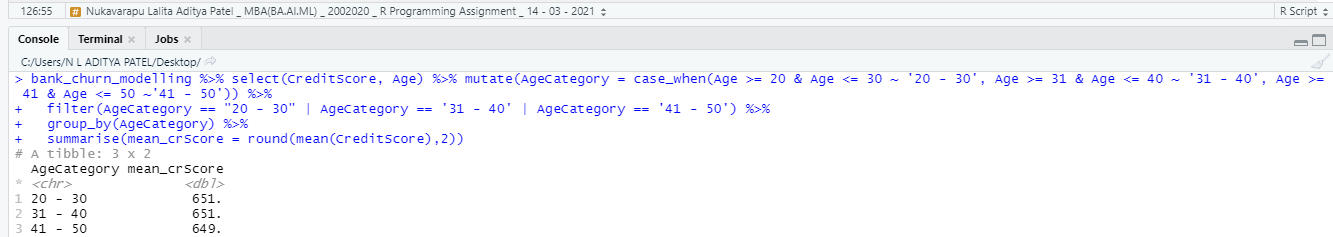
select(CreditScore, Age) %>%

mutate(AgeCategory = case\_when(Age >= 20 & Age <= 30 ~ '20 - 30', Age >= 31 & Age <= 40 ~ '31 - 40', Age >= 41 & Age <= 50 ~'41 - 50')) %>%

filter(AgeCategory == "20 - 30" | AgeCategory == '31 - 40' | AgeCategory == '41 - 50') %>%

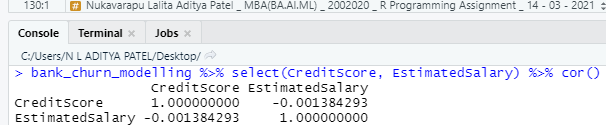
group\_by(AgeCategory) %>%

summarise(mean\_crScore = round(mean(CreditScore),2))



**C. What is the correlation between credit score and estimated salary?**

bank\_churn\_modelling %>% select(CreditScore, EstimatedSalary) %>% cor()



plot(bank\_churn\_modelling$CreditScore, bank\_churn\_modelling$EstimatedSalary, main="Correlation between Credit Score and Estimated Salary", xlab="Credit Score", ylab="Estimated Salary")

Chart, scatter chart

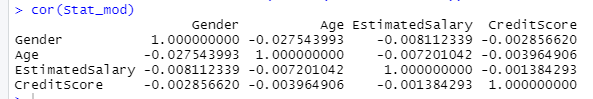
Description automatically generated

**D. Develop a statistical model to explain and establish a mathematical relationship between credit score (dependent) and gender, age, estimate salary.**

Stat\_mod <- bank\_churn\_modelling %>% select(Gender, Age, EstimatedSalary, CreditScore)

Stat\_mod$Gender <- as.numeric(factor(Stat\_mod$Gender))

cor(Stat\_mod)



corrplot(cor(Stat\_mod), method = "number")

Chart, scatter chart

Description automatically generated

set.seed(123)

train = sample(1:nrow(Stat\_mod), nrow(Stat\_mod)\*0.2)

stat\_mod\_train <- Stat\_mod[train,]

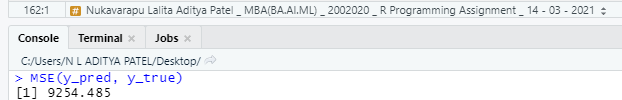
stat\_mod\_test <- Stat\_mod[-train,]

stat\_mod\_lm = lm(formula = CreditScore ~ Age + EstimatedSalary, data = bank\_churn\_modelling)

y\_pred <- predict(stat\_mod\_lm, stat\_mod\_test)

y\_true <- stat\_mod\_test$CreditScore

MSE(y\_pred, y\_true)



summary(stat\_mod\_lm)

