Analyzing Customer Churn and Credit Scores: A Case Study in Banking

Background: Bank customer churn, also known as customer attrition, refers to the phenomenon where customers stop doing business with a bank or switch to another bank. Churn is a critical metric for banks as it directly impacts their customer base and revenue. The dataset represents bank customer information for churn analysis. Each row in the dataset corresponds to a specific customer and contains several features or attributes that describe them.

Importing Libraries

```
library(e1071)
library(dplyr)
```

1. Import “Bank Churn” data and check dimension, top 5 rows and bottom 5 rows of the data frame.

```
masterdata<-read.csv("Bank Churn.csv",header=T)
str(masterdata)
```

```
'data.frame':   10000 obs. of  14 variables:
 $ RowNumber      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ CustomerId     : int  15634602 15647311 15619304 15701354 15737888 15574012 15592531 15656148 15792365 15592238
9 ...
 $ Surname        : chr  "Hargrave" "Hill" "Onio" "Boni" ...
 $ CreditScore    : int  619 608 502 699 850 645 822 376 501 684 ...
 $ Geography      : chr  "France" "Spain" "France" "France" ...
 $ Gender         : chr  "Female" "Female" "Female" "Female" ...
 $ Age            : int  42 41 42 39 43 44 50 29 44 27 ...
 $ Tenure         : int  2 1 8 1 2 8 7 4 4 2 ...
 $ Balance        : num  0 83808 159661 0 125511 ...
 $ NumOfProducts  : int  1 1 3 2 1 2 2 4 2 1 ...
 $ HasCrCard      : int  1 0 1 0 1 1 1 0 1 ...
 $ IsActiveMember : int  1 1 0 0 1 0 1 0 1 ...
 $ EstimatedSalary: num  101349 112543 113932 93827 79084 ...
 $ Exited         : int  1 0 1 0 0 1 0 1 0 0 ...
```

```
head(masterdata,5)
```

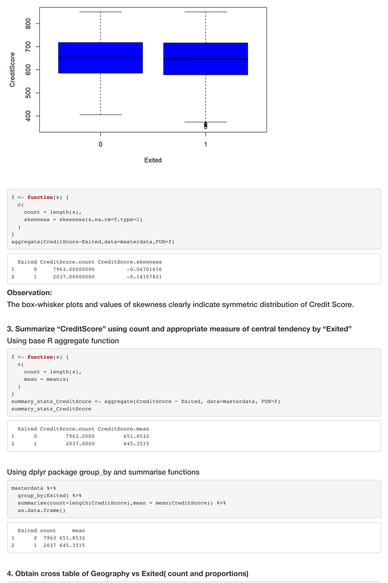
```
  RowNumber CustomerId  Surname CreditScore Geography Gender Age Tenure
1         1   15634602 Hargrave         619    France Female  42      2
2         2   15647311     Hill         608     Spain Female  41      1
3         3   15619304     Onio         502    France Female  42      8
4         4   15701354     Boni         699    France Female  39      1
5         5   15737888 Mitchell         850     Spain Female  43      2
    Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
1      0.00             1         1              1        101348.88      1
2  83807.86             1         0              1        112542.58      0
3 159660.80             3         1              0        113931.57      1
4      0.00             2         0              0         93826.63      0
5 125510.82             1         1              1         79084.10      0
```

```
tail(masterdata,5)
```

```
      RowNumber CustomerId   Surname CreditScore Geography Gender Age Tenure
9996       9996   15682229  Obijiaku         771    France   Male  39      5
9997       9997   15656092 Johnstone         516    France   Male  35     10
9998       9998   15584532       Liu         709    France Female  36      7
9999       9999   15682355 Sabbatini         772   Germany   Male  42      3
10000     10000   15628319    Walker         792    France Female  28      4
       Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
9996      0.00             2         1              0       96270.64      0
9997  75075.31             1         1              1      101699.77      0
9998      0.00             1         0              1      42085.58      1
9999  75075.31             2         1              0      92888.52      1
10000 130142.79             1         1              0      38190.78      0
```

2. Check if the distribution of “CreditScore” is symmetric for Exited=1 and Exited=0. Obtain box-whisker plot and estimate the values of skewness.

```
boxplot(CreditScore~Exited,data = masterdata,col="blue")
```




```
f <- function(x) {
  c(
    count = length(x),
    skewness = skewness(x,na.rm=T,type=2)
  )
}
aggregate(CreditScore~Exited,data=masterdata,FUN=f)
```

```
  Exited CreditScore.count CreditScore.skewness
1      0             7963          -0.04791616
2      1             2037          -0.14107821
```

Observation:

The box-whisker plots and values of skewness clearly indicate symmetric distribution of Credit Score.

3. Summarize “CreditScore” using count and appropriate measure of central tendency by “Exited”

Using base R aggregate function

```
f <- function(x) {
  c(
    count = length(x),
    mean = mean(x)
  )
}
summary_stats_CreditScore <- aggregate(CreditScore ~ Exited, data=masterdata, FUN=f)
summary_stats_CreditScore
```

```
  Exited CreditScore.count CreditScore.mean
1      0             7963.0000      651.8532
2      1             2037.0000      645.3515
```

Using dplyr package group_by and summarise functions

```
masterdata %>%
  group_by(Exited) %>%
  summarise(count=length(CreditScore),mean = mean(CreditScore)) %>%
  as.data.frame()
```

```
  Exited count     mean
1      0  7963 651.8532
2      1  2037 645.3515
```

4. Obtain cross table of Geography vs Exited(count and proportions)

```
Exited_geo<-table(masterdata$Geography,masterdata$Exited)
colnames(Exited_geo)<-c("Stayed","Exited")
Exited_geo
```

```
        Stayed Exited
France    4204    810
Germany   1695    814
Spain     2064    413
```

```
Exited_geo2<-round(prop.table(Exited_geo,1)*100,2)
Exited_geo2
```

```
        Stayed Exited
France   83.85  16.15
Germany  67.56  32.44
Spain    83.33  16.67
```

Observation:

Churn rates vary significantly: France and Spain have similar rates, while Germany's is notably higher.

5. Obtain Correlation Coefficient between CreditScore and Estimated Salary and interpret.

```
round(cor(masterdata$CreditScore,masterdata$EstimatedSalary),4)
```

```
[1] -0.0014
```

Observation:

The correlation coefficient of approximately -0.0014 suggests a very weak, near-zero correlation between CreditScore and Estimated Salary. These variables appear largely unrelated.

6. Derive a new variable as CreditScore_Cat=1 if >=650;0 if <650

```
masterdata$CreditScore_Cat<-ifelse(masterdata$CreditScore<650,1,0)
head(masterdata,5)
```

```
  RowNumber CustomerId  Surname CreditScore Geography Gender Age Tenure
1         1   15634602 Hargrave         619    France Female  42      2
2         2   15647311     Hill         608     Spain Female  41      1
3         3   15619304     Onio         502    France Female  42      8
4         4   15701354     Boni         699    France Female  39      1
5         5   15737888 Mitchell         850     Spain Female  43      2
    Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
1      0.00             1         1              1        101348.88      1
2  83807.86             1         0              1        112542.58      0
3 159660.80             3         1              0        113931.57      1
4      0.00             2         0              0         93826.63      0
5 125510.82             1         1              1         79084.10      0
  CreditScore_Cat
1               0
2               1
3               0
4               1
5               1
```

7. Obtain cross table of CreditScore_Cat vs Exited

```
Exited_cat<-table(masterdata$CreditScore_Cat,masterdata$Exited)
colnames(Exited_cat)<-c("Stayed","Exited")
Exited_cat
```

```
   Stayed Exited
0    3851   1049
1    4112    988
```

```
Exited_cat2<-round(prop.table(Exited_cat,1)*100,2)
Exited_cat2
```

```
   Stayed Exited
0   78.59  21.41
1   80.63  19.37
```

Observation:

Customers with a CreditScore_Cat 0 have a slightly higher exit rate (21.4%) compared to those with a CreditScore_Cat 1, who have a lower exit rate (19.4%).

8. Create a subset of 300 customers with highest Credit Score and check how they are spread over Geography

```
top_300_customers <- masterdata[order(masterdata$CreditScore,decreasing = T),]
top_300_customers<-head(top_300_customers,300)

table(top_300_customers$Geography)
```

```
 France Germany   Spain
    150      88      70
```

Observation:

Among the top 300 customers with the highest Credit Scores, the majority are from France, followed by Germany and Spain.

9. Summarize “CreditScore” using count, mean and median by Geography+Gender

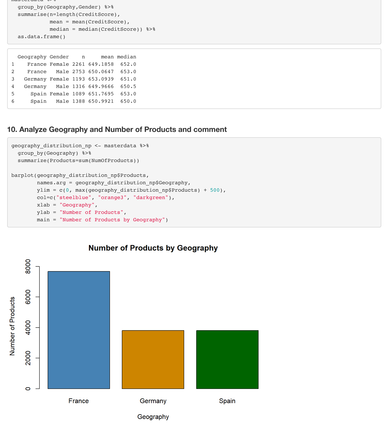
```
masterdata %>%
  group_by(Geography,Gender) %>%
  summarise(n=length(CreditScore),
            mean = mean(CreditScore),
            median = median(CreditScore)) %>%
  as.data.frame()
```

```
  Geography Gender    n     mean median
1    France Female 2261 649.1858  652.0
2    France   Male 2753 650.0647  653.0
3   Germany Female 1193 653.0939  651.0
4   Germany   Male 1314 649.9666  650.5
5     Spain Female 1089 651.7695  653.0
6     Spain   Male 1388 650.9921  650.0
```

10. Analyze Geography and Number of Products and comment

```
geography_distribution_np <- masterdata %>%
  group_by(Geography) %>%
  summarise(Products=sum(NumOfProducts))

barplot(geography_distribution_np$Products,
        names.arg = geography_distribution_np$Geography,
        ylim = c(0, max(geography_distribution_np$Products) + 500),
        col=c('steelblue', 'orange3', 'darkgreen'),
        xlab = "Geography",
        ylab = "Number of Products",
        main = "Number of Products by Geography")
```



Observation:

France has the highest number of products, while Spain and Germany have the same count.