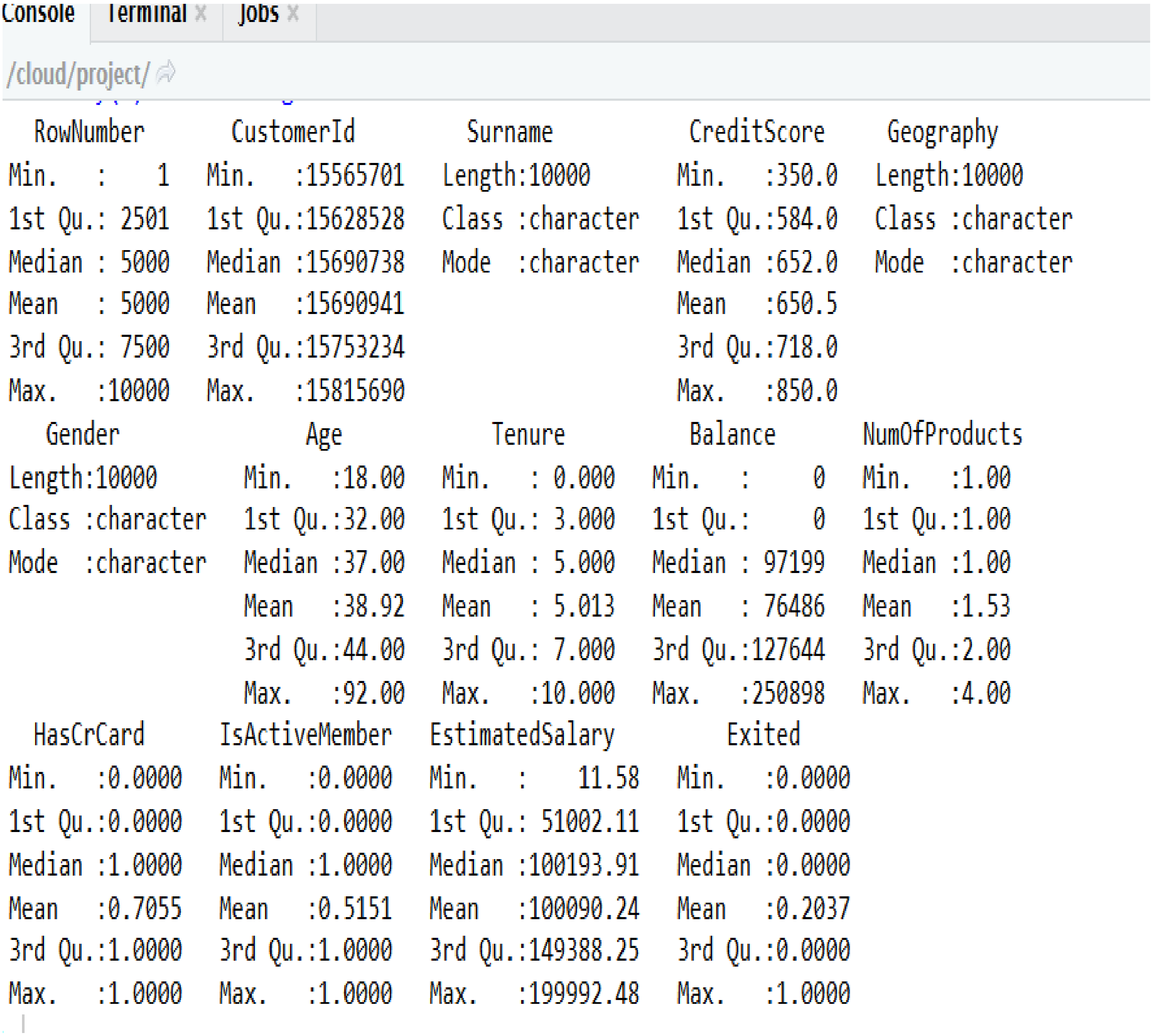
Assessment

File: - Distribution of variables.R

Summary of data



```
   RowNumber       CustomerId         Surname           CreditScore     Geography        
 Min.   :    1   Min.   :15565701   Length:10000       Min.   :350.0   Length:10000      
 1st Qu.: 2501   1st Qu.:15628528   Class :character   1st Qu.:584.0   Class :character  
 Median : 5000   Median :15690738   Mode  :character   Median :652.0   Mode  :character  
 Mean   : 5000   Mean   :15690941                      Mean   :650.5                     
 3rd Qu.: 7500   3rd Qu.:15753234                      3rd Qu.:718.0                     
 Max.   :10000   Max.   :15815690                      Max.   :850.0                     
    Gender               Age            Tenure          Balance       NumOfProducts 
 Length:10000       Min.   :18.00   Min.   : 0.000   Min.   :     0   Min.   :1.00  
 Class :character   1st Qu.:32.00   1st Qu.: 3.000   1st Qu.:     0   1st Qu.:1.00  
 Mode  :character   Median :37.00   Median : 5.000   Median : 97199   Median :1.00  
                    Mean   :38.92   Mean   : 5.013   Mean   : 76486   Mean   :1.53  
                    3rd Qu.:44.00   3rd Qu.: 7.000   3rd Qu.:127644   3rd Qu.:2.00  
                    Max.   :92.00   Max.   :10.000   Max.   :250898   Max.   :4.00  
   HasCrCard      IsActiveMember   EstimatedSalary         Exited      
 Min.   :0.0000   Min.   :0.0000   Min.   :    11.58   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 51002.11   1st Qu.:0.0000  
 Median :1.0000   Median :1.0000   Median :100193.91   Median :0.0000  
 Mean   :0.7055   Mean   :0.5151   Mean   :100090.24   Mean   :0.2037  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:149388.25   3rd Qu.:0.0000  
 Max.   :1.0000   Max.   :1.0000   Max.   :199992.48   Max.   :1.0000  
```

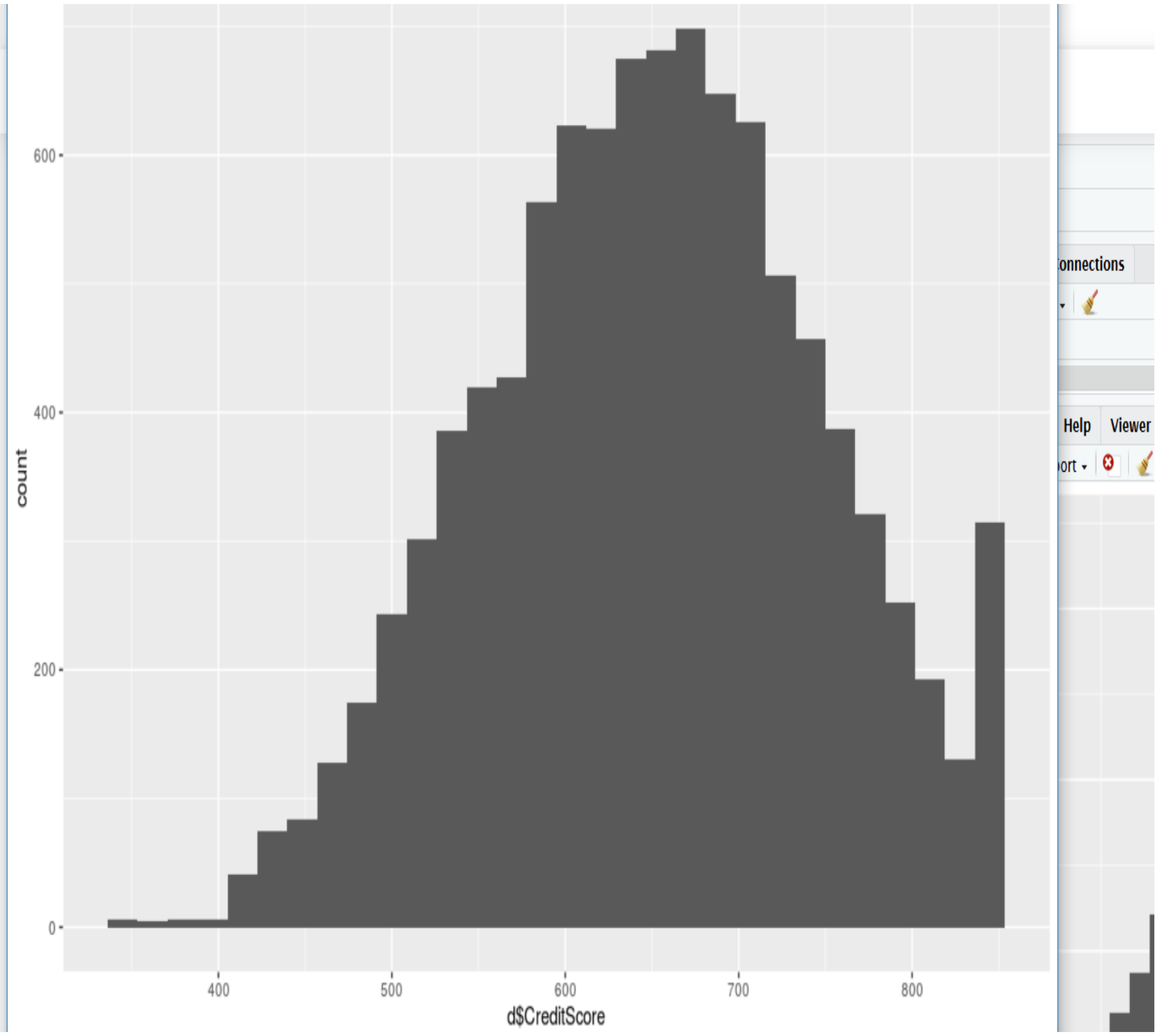
a> Missing values: - None as displayed in the summary of the dataset.

b> Categorical variables: - Geography, Gender, HasCrCard, IsActiveMember, Exited.

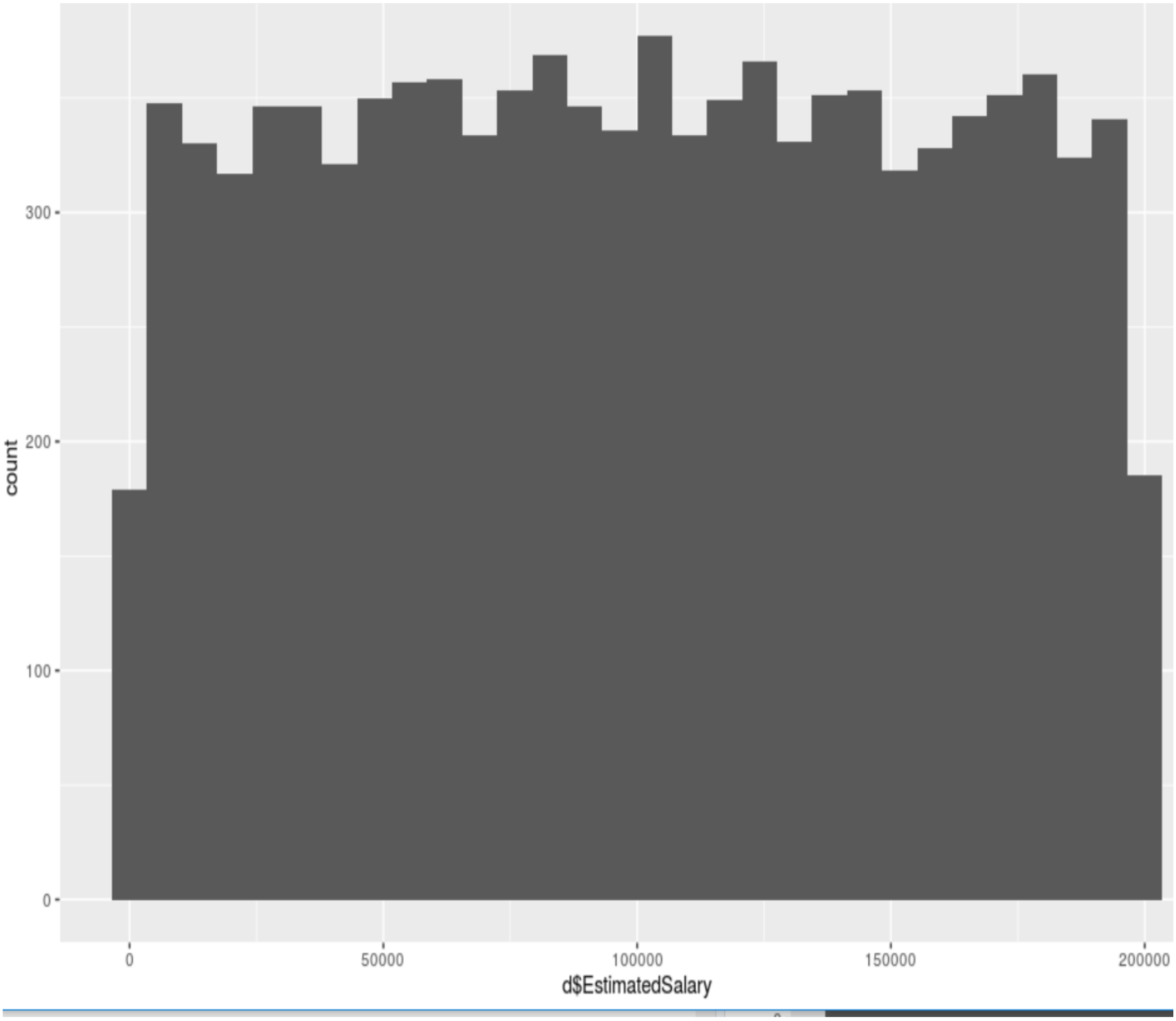
Graphs and insights (EDA)

Plots and observations:-

Uni-variate Analysis :-



Credit-Score - Negatively-skewed distribution

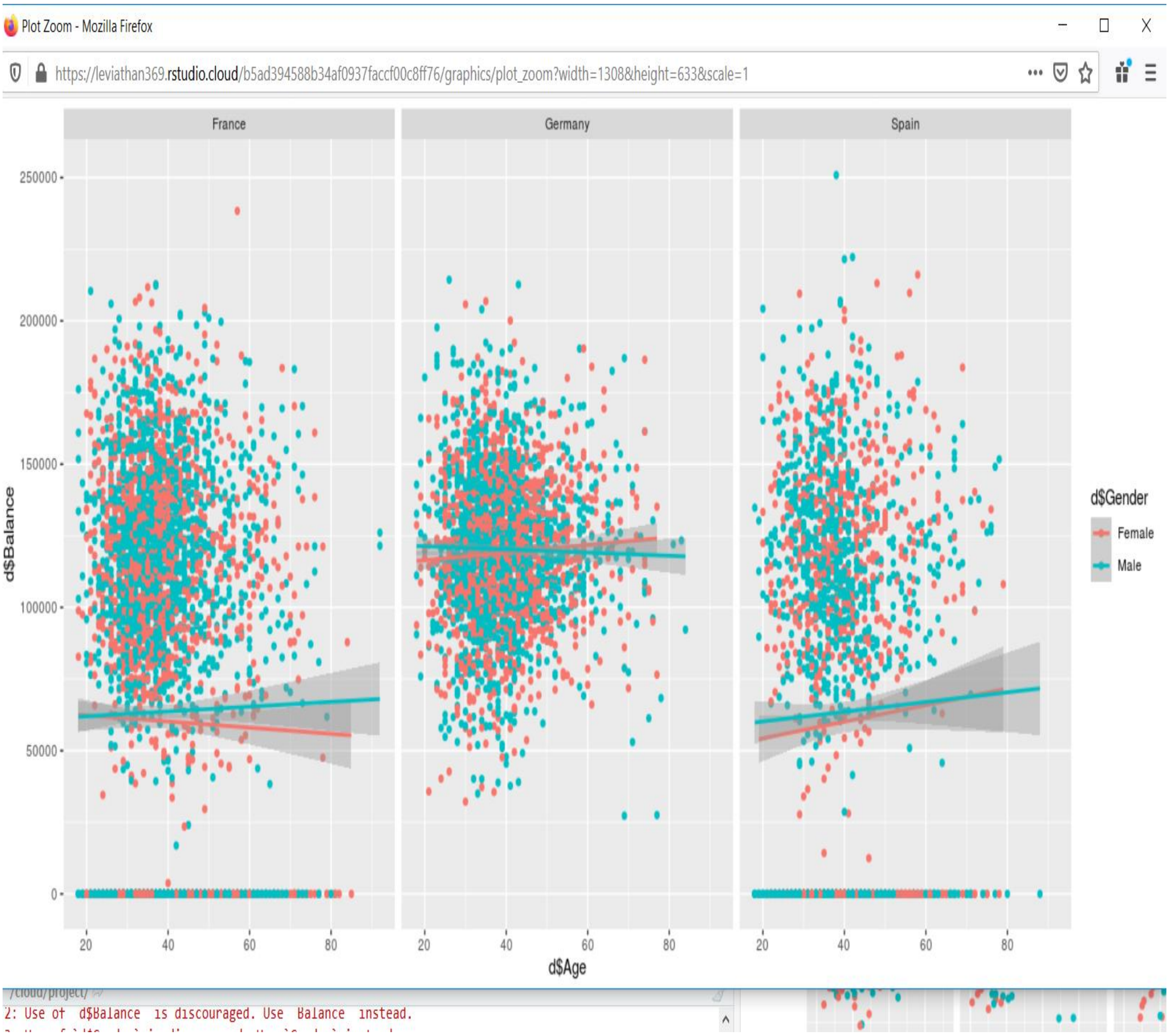


Estimated-Salary - Normal Distribution

File: - Scatter plot and Data story telling.R

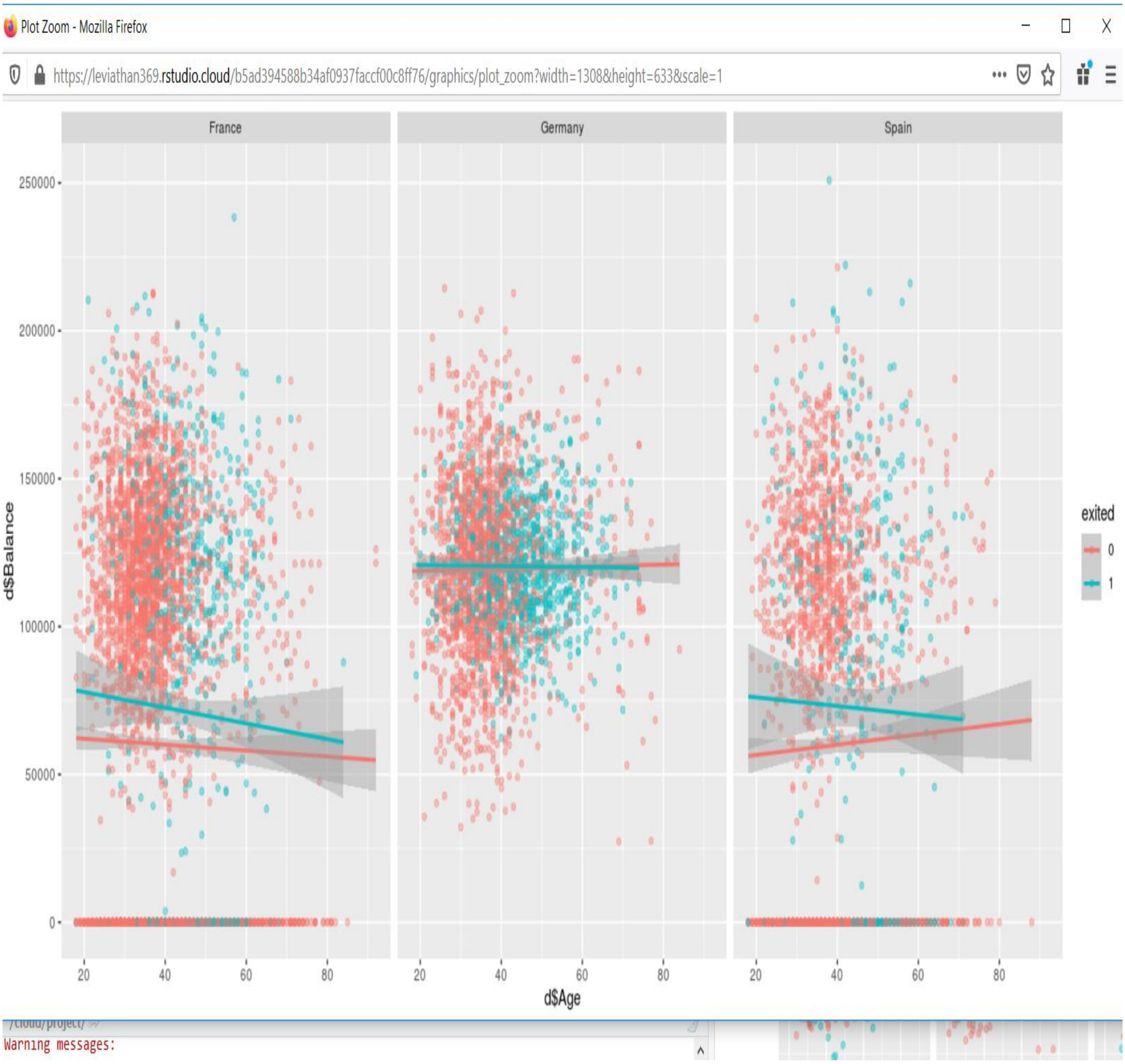
Multivariate analysis :-

Scatter plots :-


Plot Zoom - Mozilla Firefox
https://leviathan369.rstudio.cloud/b5ad394588b34af0937faccf00c8ff76/graphics/plot_zoom?width=1308&height=633&scale=1
France
Germany
Spain
250000
200000
150000
100000
50000
0
d$Balance
20
40
60
80
d$Age
d$Gender
Female
Male

```
~/cloud/project/
2: Use of  d$Balance  is discouraged. Use  Balance  instead.
```

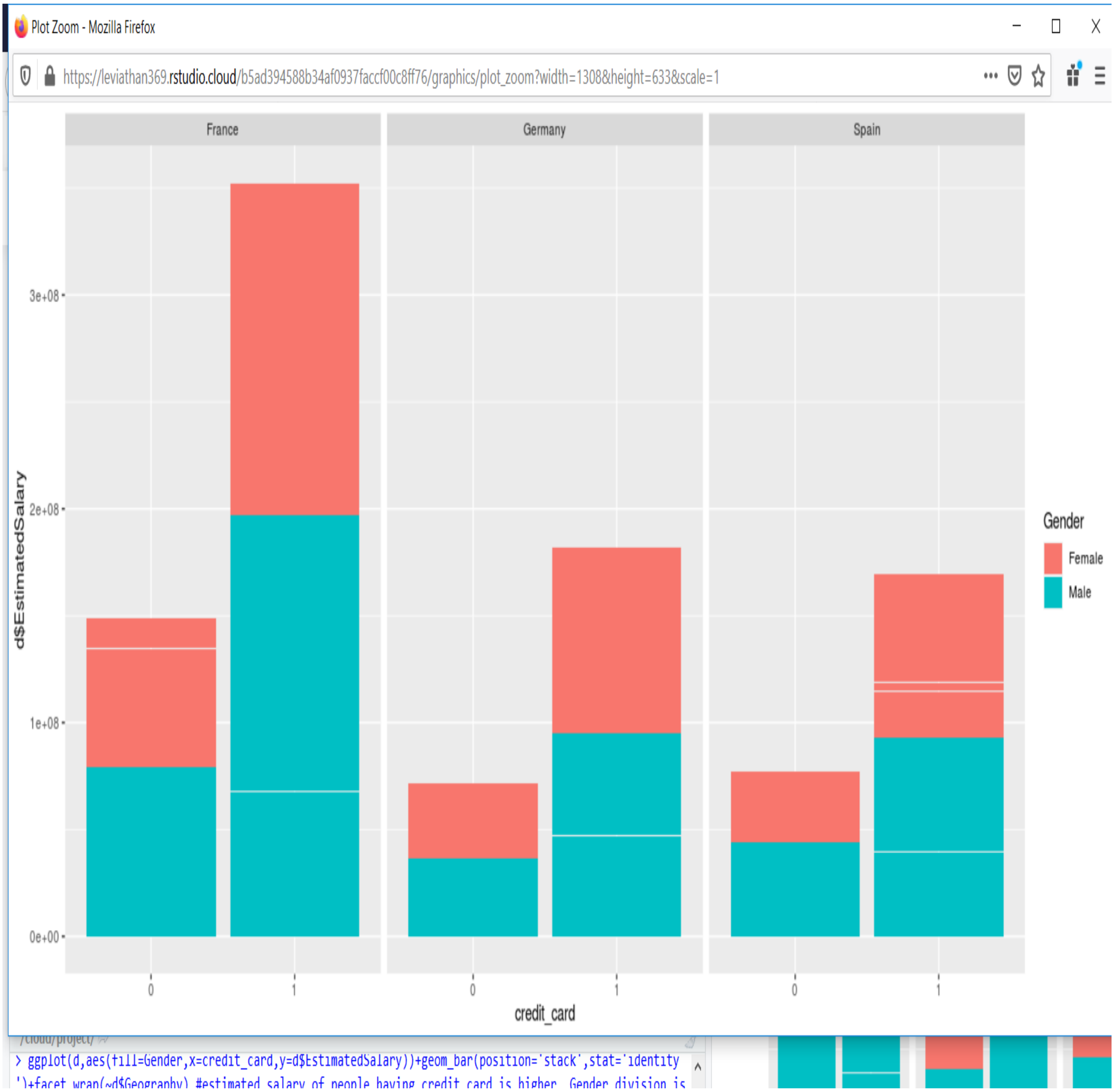
Trend line varies w.r.t countries. (x= Age, y=Balance, colour= Gender, distinction = countries). No major trend observed between these two variables.



```
Warning messages:
```

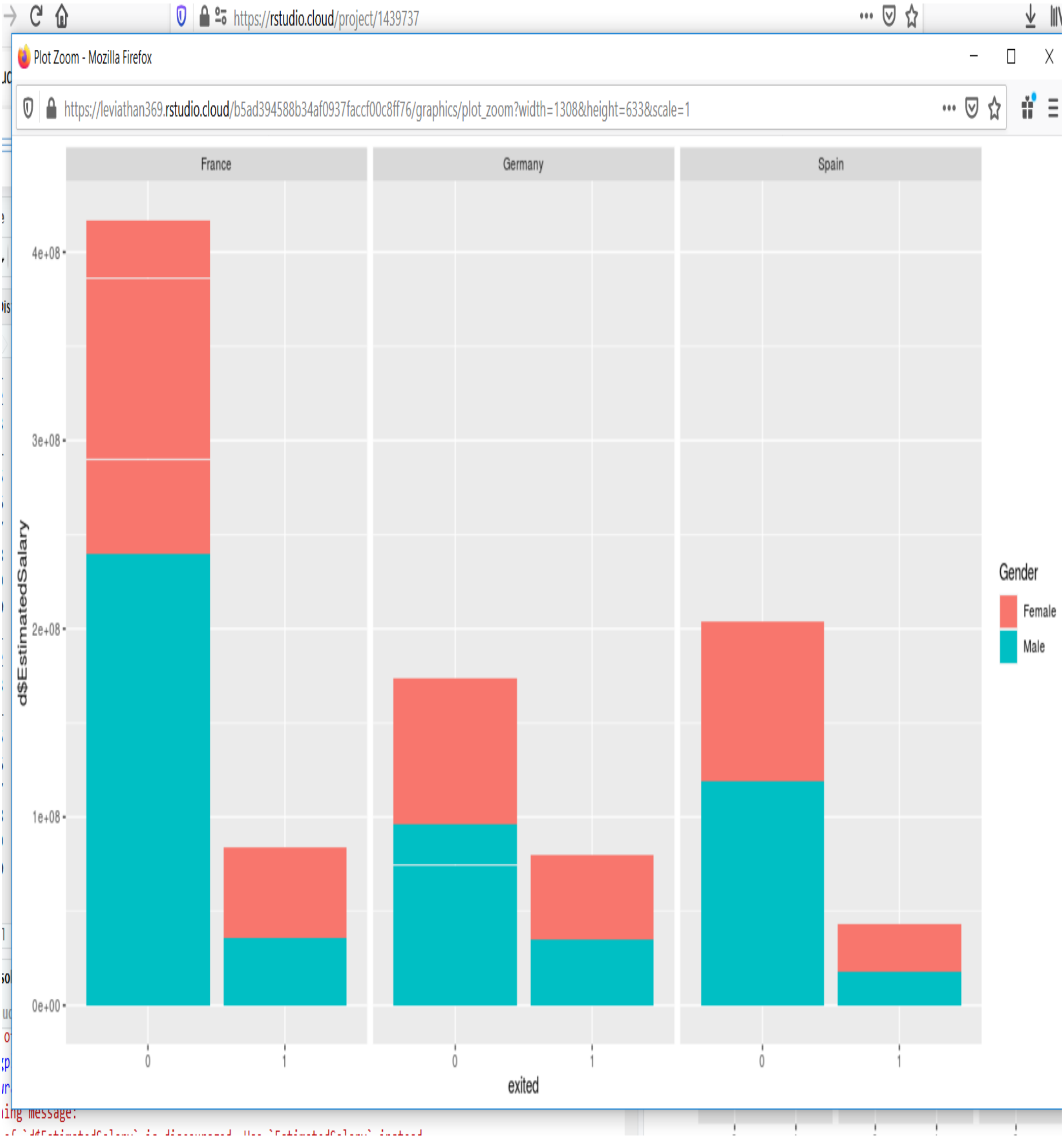
Age vs. balance for categorical variable exited to be predicted. As observed number of people not exiting (0) are between (20-40) yrs. in age and has balance between (100000-150000). No major observation from trend line.

Bar plots :-


Plot Zoom - Mozilla Firefox
https://leviathan369.rstudio.cloud/b5ad394588b34af0937faccf00c8ff76/graphics/plot_zoom?width=1308&height=633&scale=1
France
Germany
Spain
3e+08
2e+08
1e+08
0e+00
d$EstimatedSalary
0
1
0
1
0
1
credit_card
Gender
Female
Male
/cloud/project/

```
> ggplot(d,aes(fill=Gender,x=credit_card,y=d$EstimatedSalary))+geom_bar(position='stack',stat='identity
')+facet_wrap(~d$Geography) #estimated salary of people having credit card is higher .Gender division is
```

Estimated salary of people having credit card is higher .Gender division is almost equal.

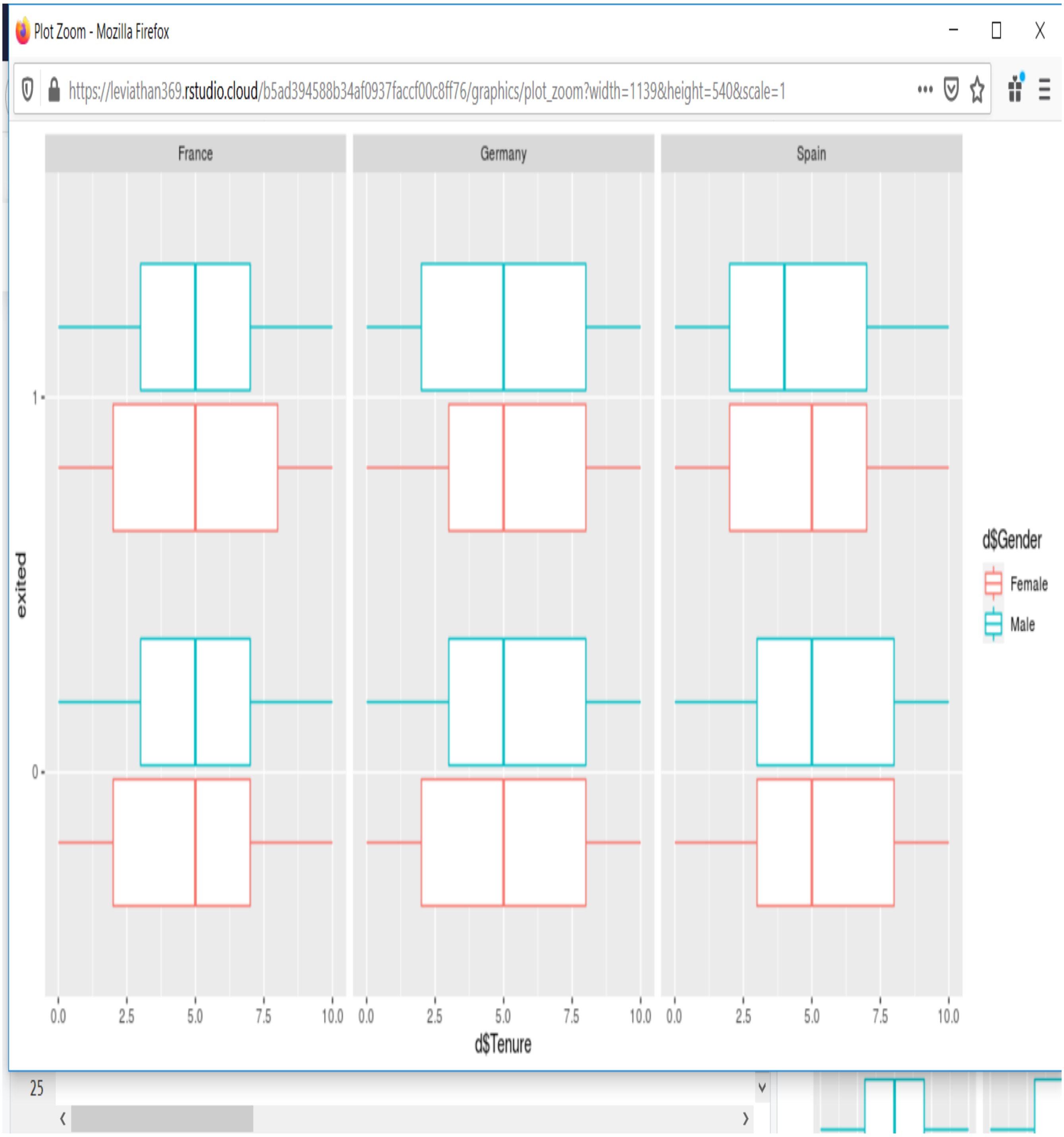


```
ing message:
of `d$EstimatedSalary` is discouraged. Use `EstimatedSalary` instead.
```

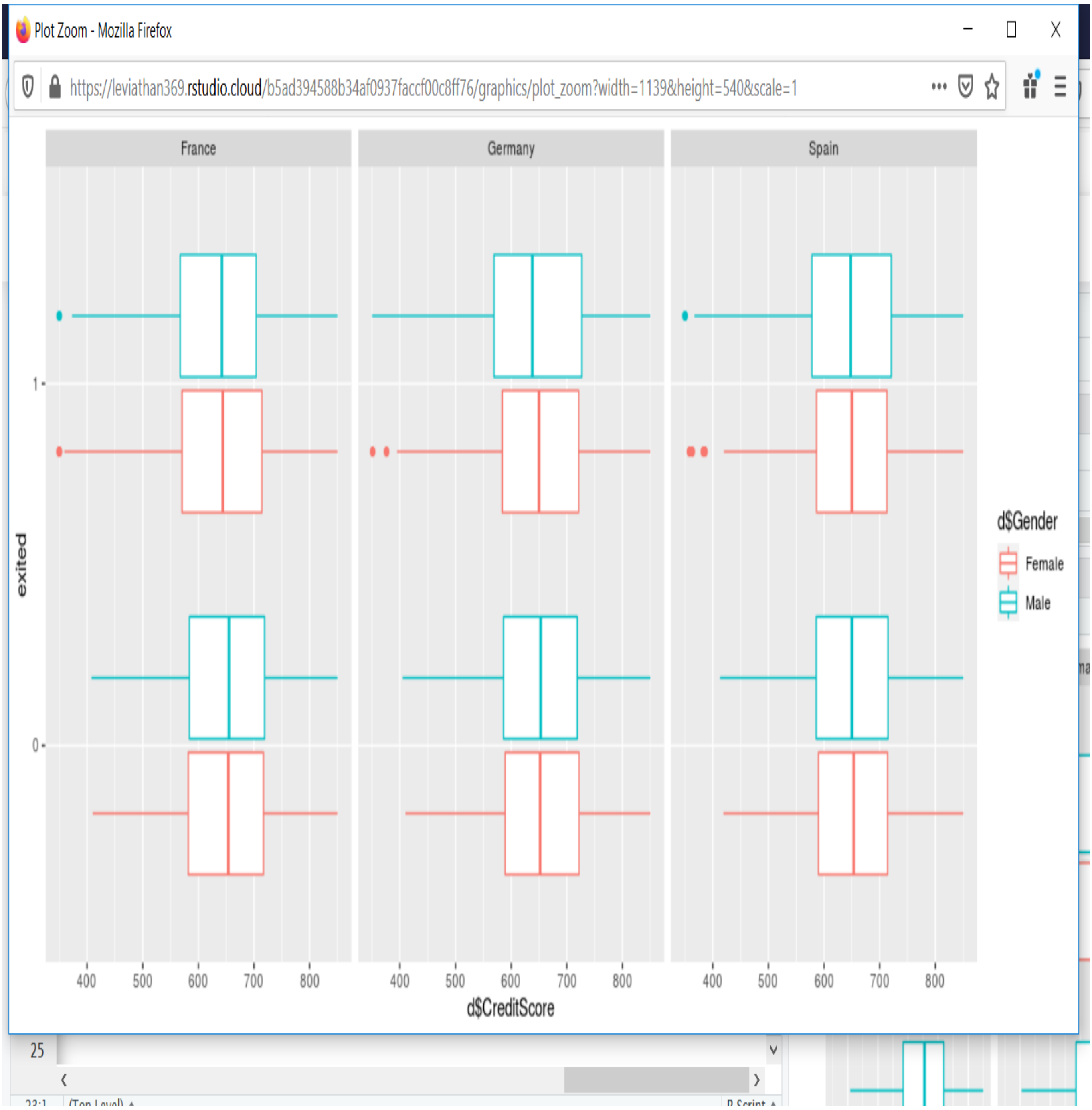
People with higher estimated salary has not exited. Distribution of male and female are almost equal.

Similar trend can be observed with balance and credit score for the categorical variable (exited).

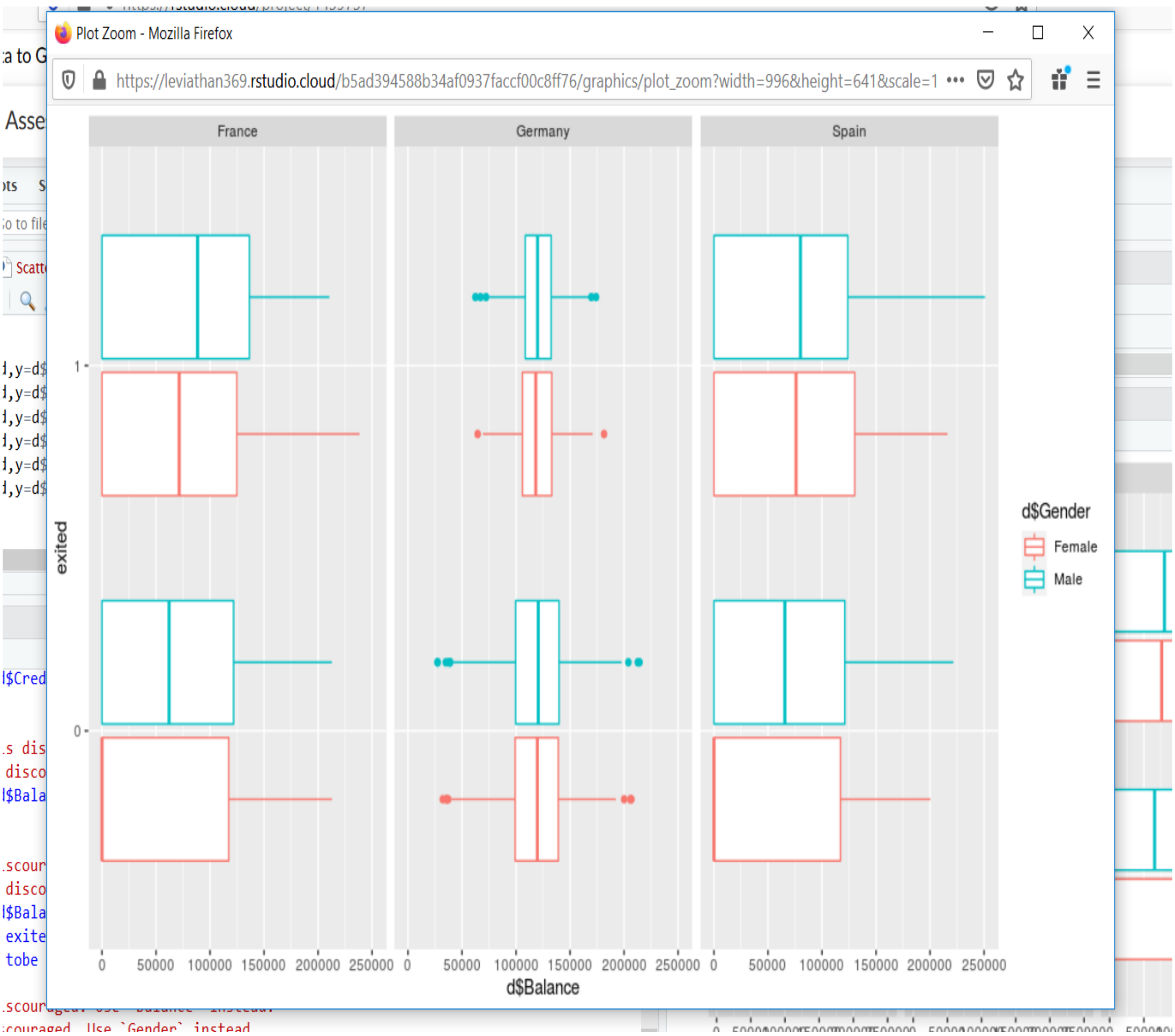
Box plots



The median tenure is almost same in case of people who exited and a little lower in the case of males in Spain who exited. But the difference can be spotted region wise.


Plot Zoom - Mozilla Firefox
https://leviathan369.rstudio.cloud/b5ad394588b34af0937faccf00c8ff76/graphics/plot_zoom?width=1139&height=540&scale=1
France
Germany
Spain
exited
1
0
400
500
600
700
800
d$CreditScore
d$Gender
Female
Male
25

The median credit score is almost same in case of people who exited and a little lower in the case of males who exited. No major difference region wise. Outliers detected which are [< Q1-1.5*(Q3-Q1)]



Plot Zoom - Mozilla Firefox

https://leviathan369.rstudio.cloud/b5ad394588b34af0937faccf00c8ff76/graphics/plot_zoom?width=996&height=641&scale=1

For the variable balance outliers are detected on both sides of the plot for Germany. Outliers detected are:

>Q3+ (1.5*[Q3-Q1])

<Q1-(1.5*[Q3-Q1])

File: - Data preparation (missing value treatment).R

Data Preparation (Outlier and missing value treatment, normalizing data)

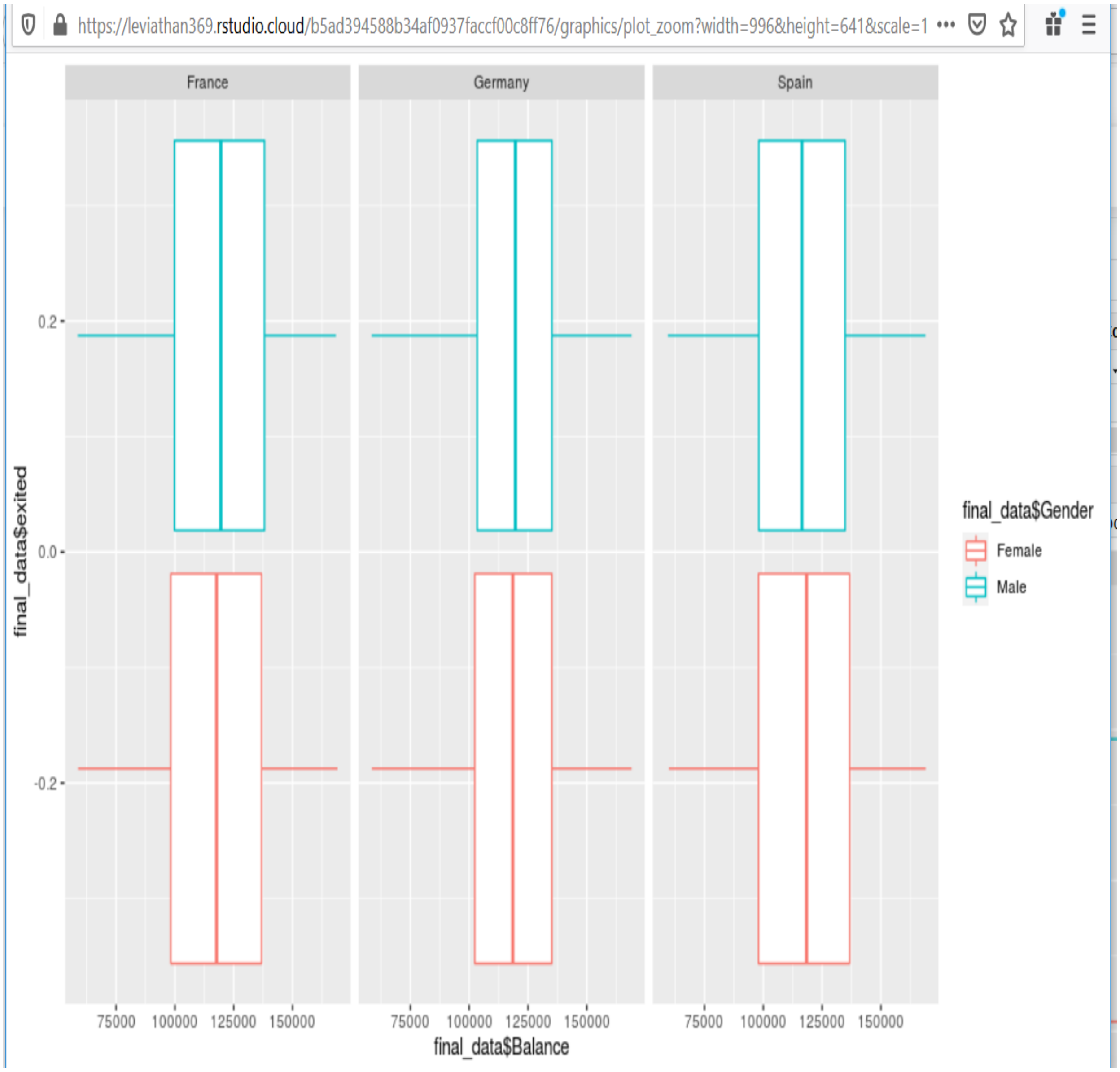
1> Removing the columns unnecessary for the model. Using select function in deplyr.

Final data<-select (d,-'Row Number',-'Customer ID',-'Surname')

2> Removing outliers :-

Balance:-

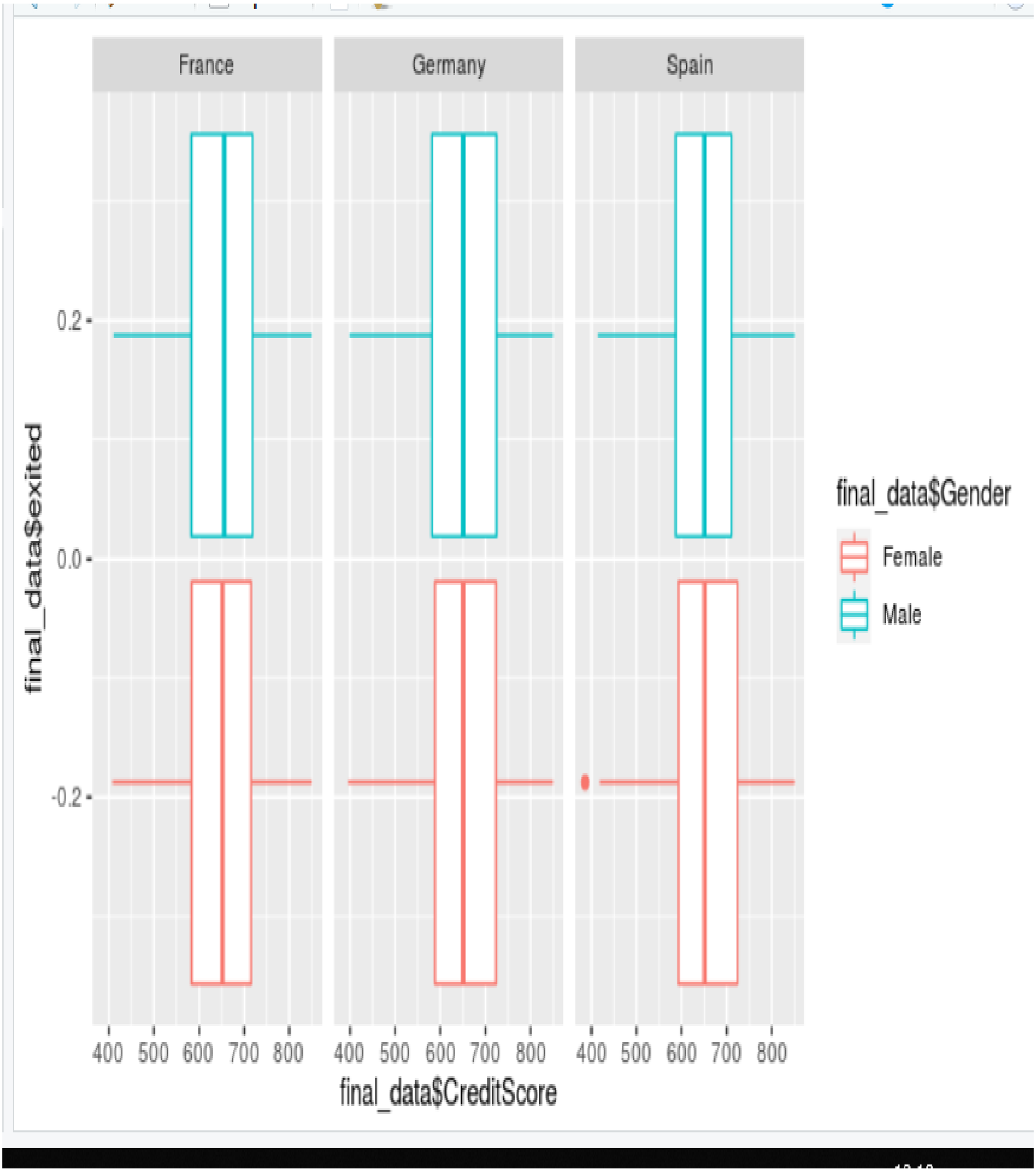
final_data<-filter (final_data, final_data$Balance> (100000-(1.5*27644))&final_data$Balance<(127644+(1.5*27644)))



https://leviathan369.rstudio.cloud/b5ad394588b34af0937faccf00c8ff76/graphics/plot_zoom?width=996&height=641&scale=1

Credit score:-

```
final_data<- filter (final_data,final_data$CreditScore>(584.0-(1.5*134))
```



3> After outlier removal the variable credit _score follows a nearly normal distribution. (median : 652 mean : 651)

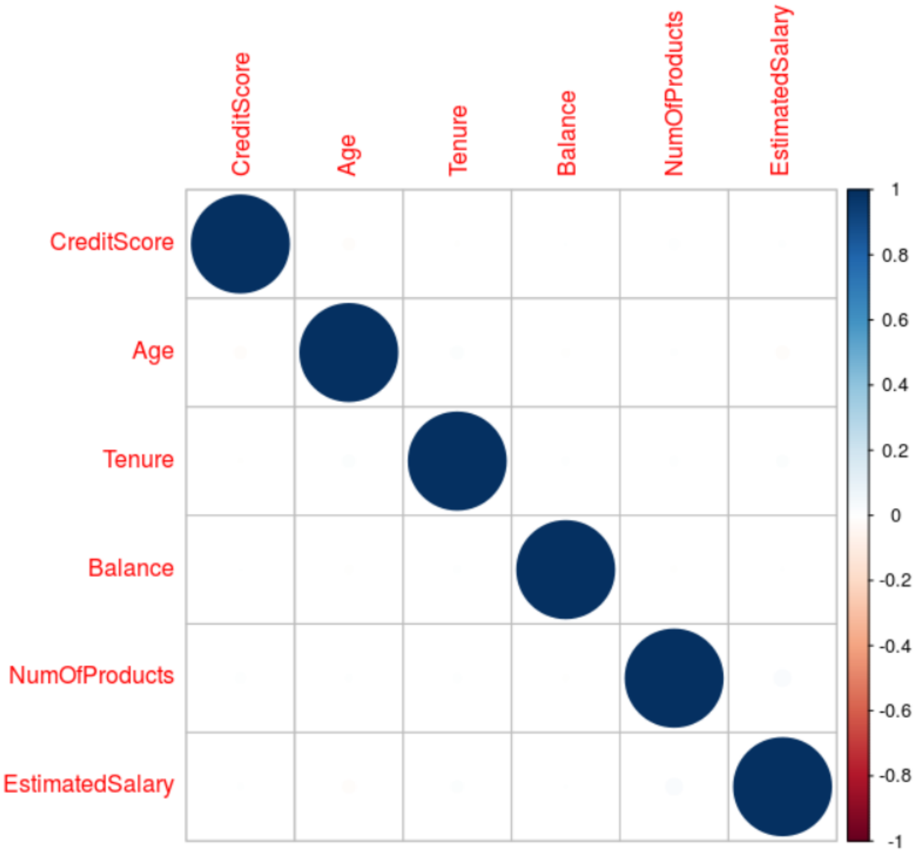
Correlation plot (final_Data)

After outlier removal our final dataset is ready. Stored in “final_dataset” object.

To figure out the correlation between the variables in the final dataset we need a correlation plot.

We will consider numerically continuous variables for this

corr_var<-select (final_data,-'IsActiveMember',-'HasCrCard',-'Gender',-'Geography',-'Exited')



No significant correlation between variables as depicted from the corr plot. Method used = Pearson correlation.

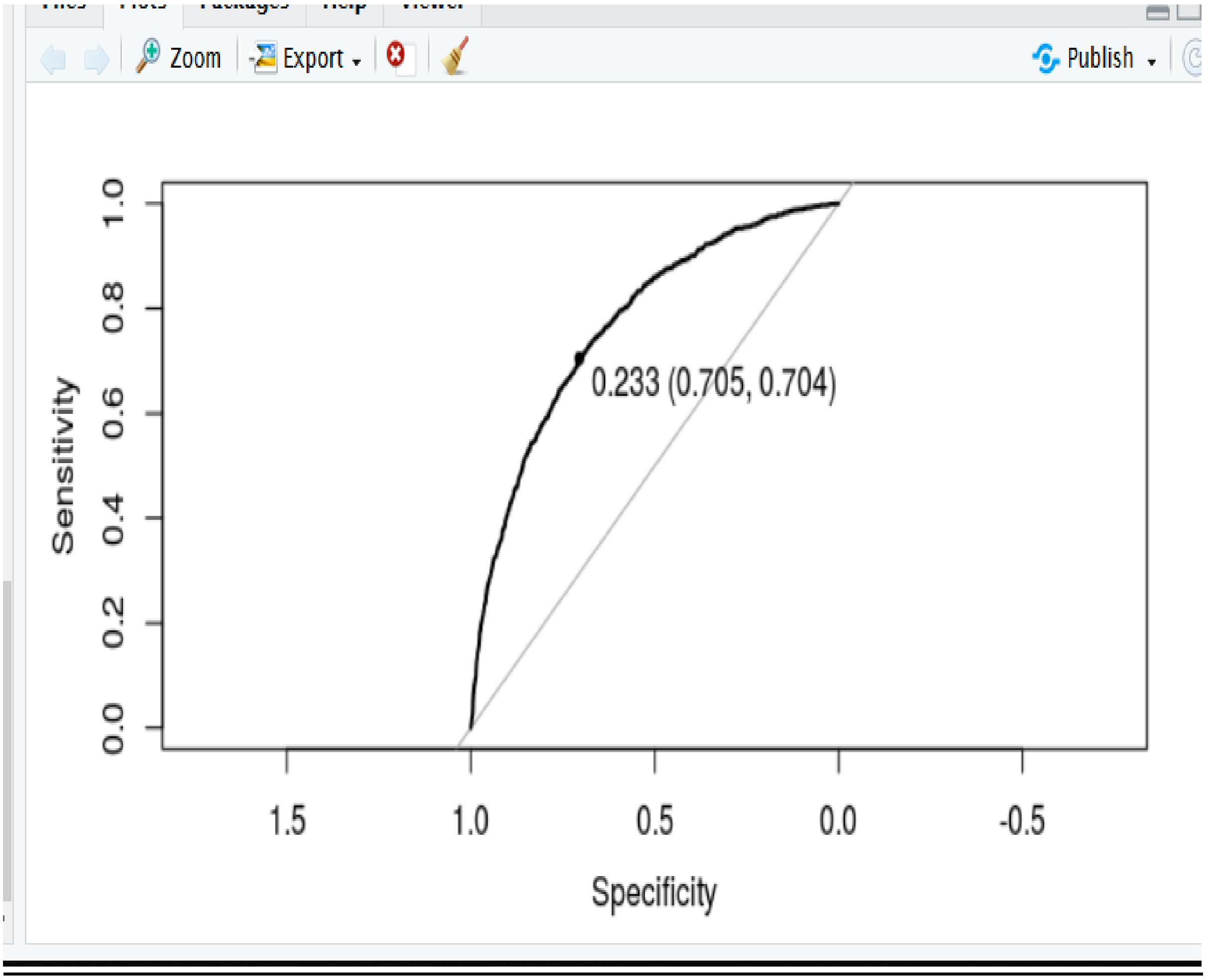
Models

File :- Logistic_Regression.R

Logistic regression

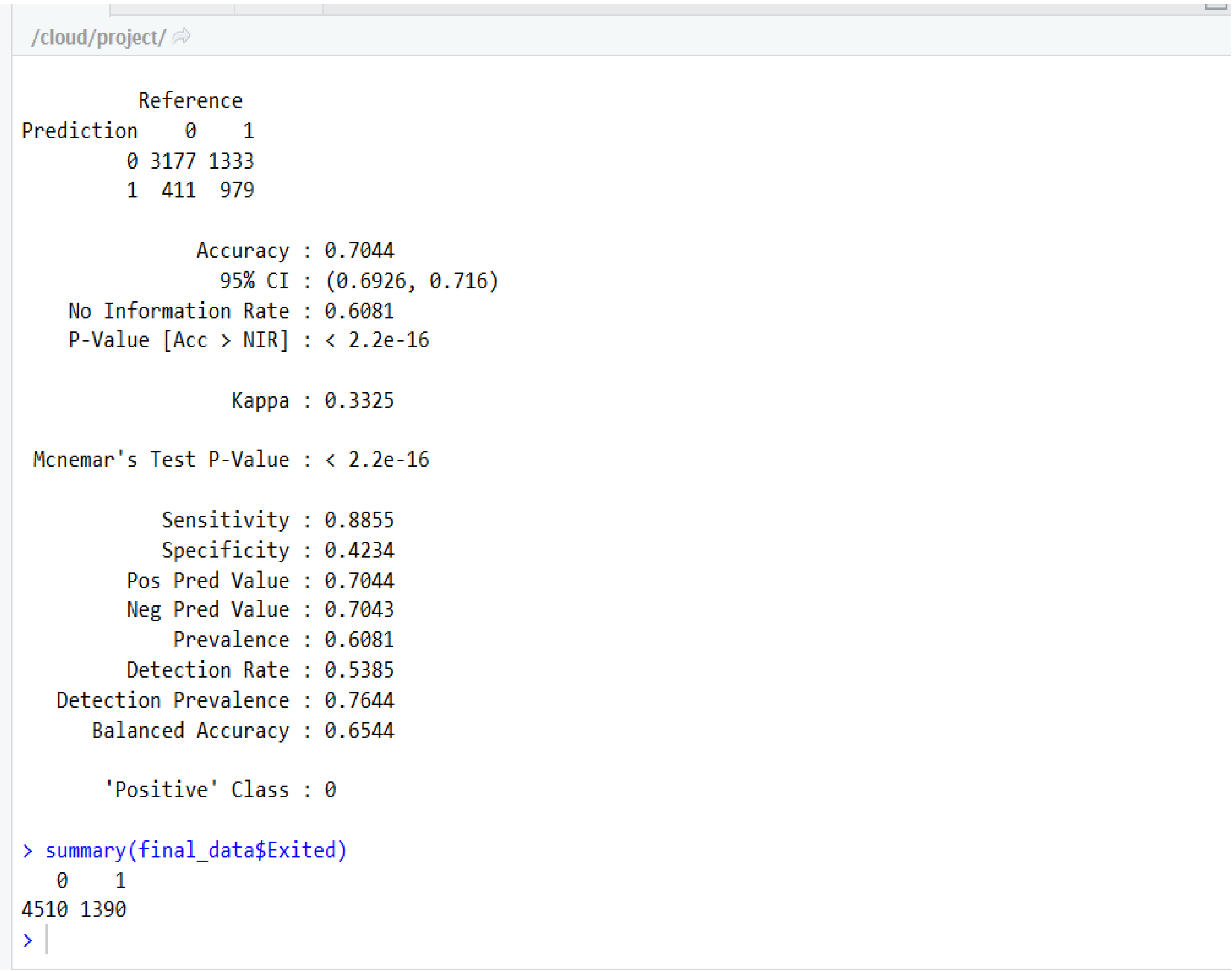
AIC = 5421.1

ROC CURVE



Threshold – 0.233

Model summary (Confusion matrix and insights)



```
          Reference
Prediction    0    1
         0 3177 1333
         1  411  979
                                          
               Accuracy : 0.7044          
                 95% CI : (0.6926, 0.716) 
    No Information Rate : 0.6081          
    P-Value [Acc > NIR] : < 2.2e-16       
                                          
                  Kappa : 0.3325          
                                          
 Mcnemar's Test P-Value : < 2.2e-16       
                                          
            Sensitivity : 0.8855          
            Specificity : 0.4234          
         Pos Pred Value : 0.7044          
         Neg Pred Value : 0.7043          
             Prevalence : 0.6081          
         Detection Rate : 0.5385          
   Detection Prevalence : 0.7644          
      Balanced Accuracy : 0.6544          
                                          
       'Positive' Class : 0               
                                          
> summary(final_data$Exited)
   0    1 
4510 1390 
> 
```

As observed the model is statistically significant in explaining variations (low P-VALUE).

Over all accuracy = 0.7044

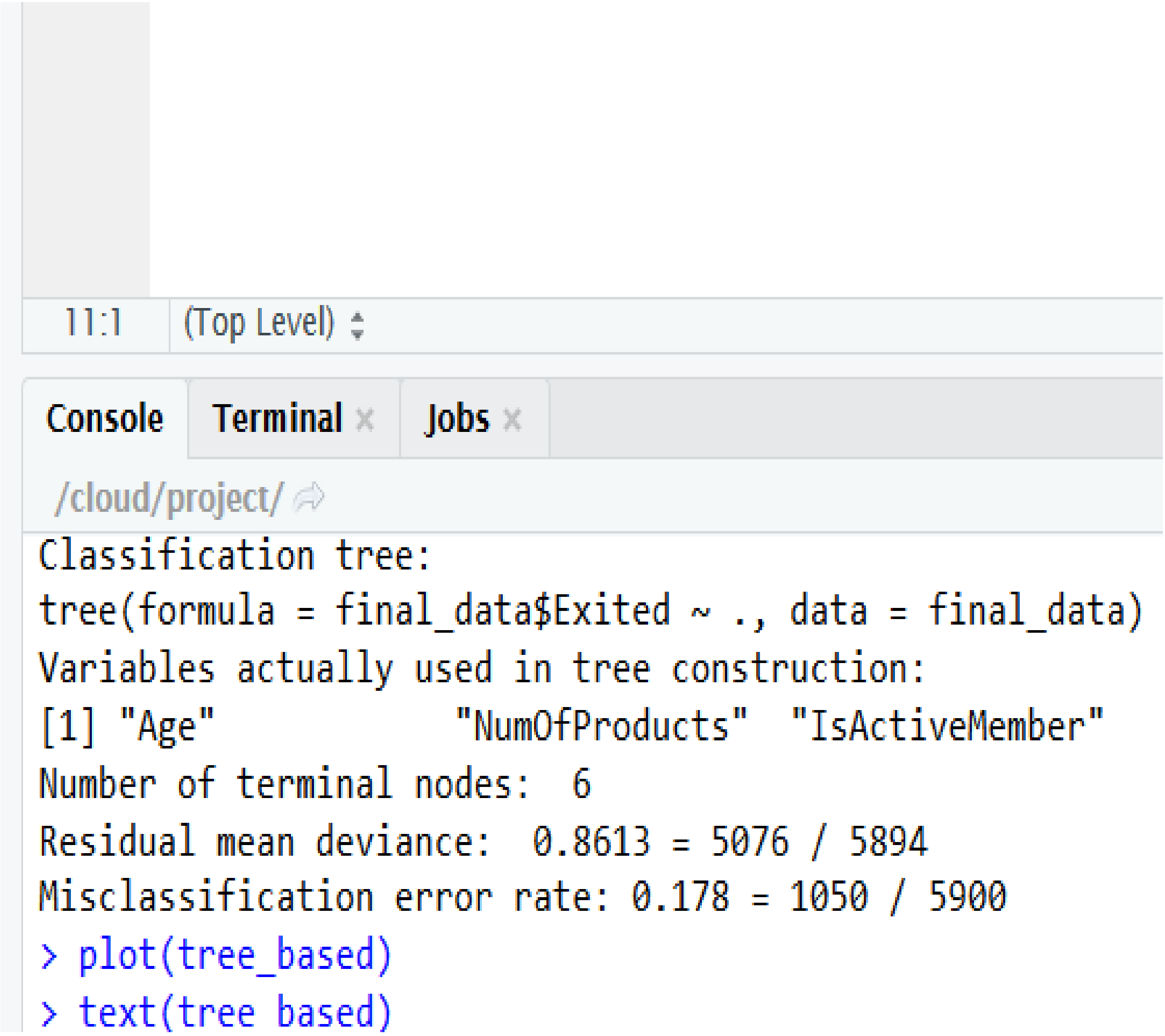
High sensitivity and low specificity indicates that the model is biased towards predicting ‘1’. (Over fitting)

This is due to the imbalance in dataset as observed from summary function result.

TREE BASED MODEL

FILE:- TREE MODEL.R

Model_summary



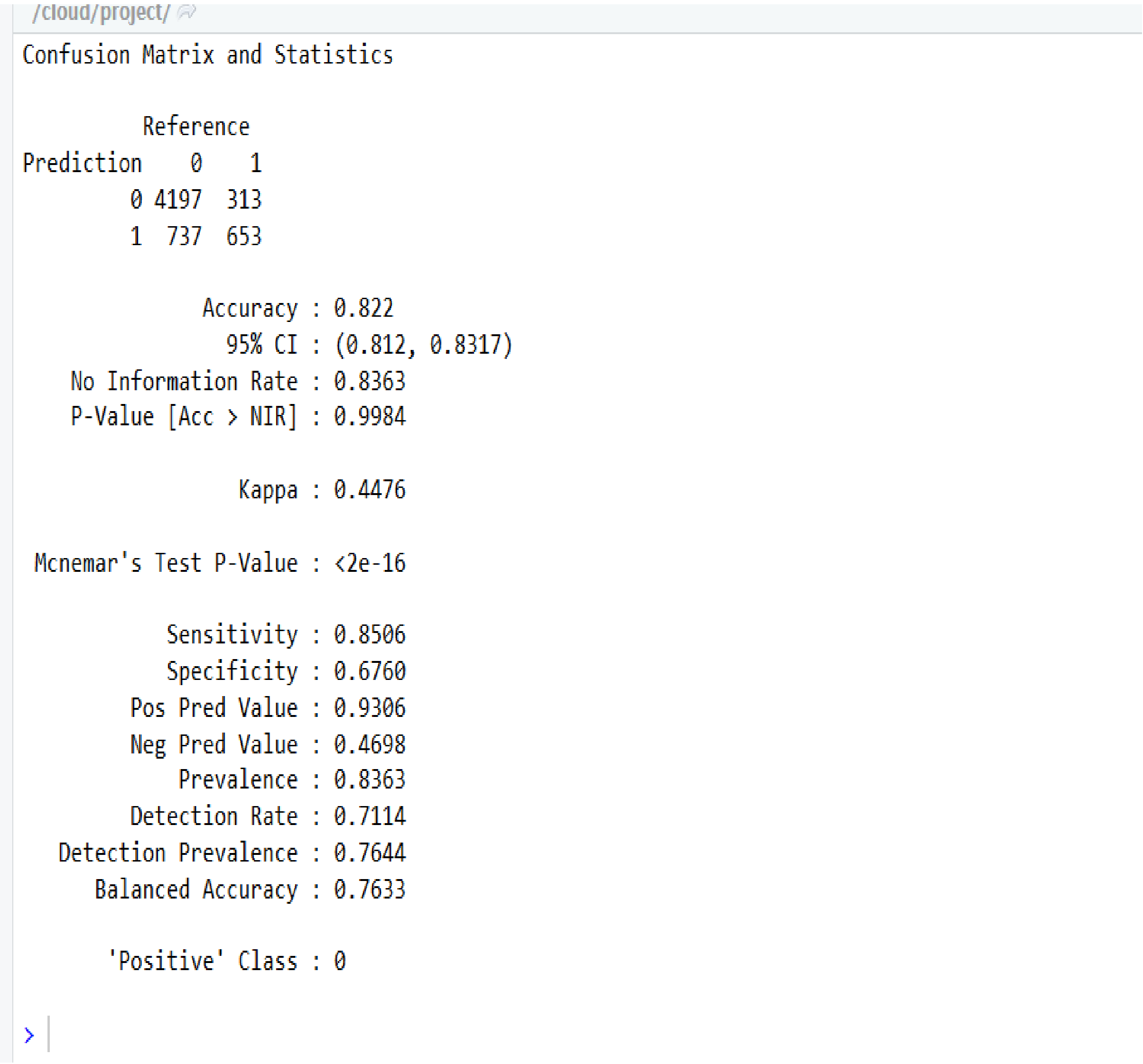
```
Classification tree:
tree(formula = final_data$Exited ~ ., data = final_data)
Variables actually used in tree construction:
[1] "Age"           "NumOfProducts"  "IsActiveMember"
Number of terminal nodes:  6
Residual mean deviance:  0.8613 = 5076 / 5894
Misclassification error rate: 0.178 = 1050 / 5900
> plot(tree_based)
> text(tree_based)
```

Misclassification error is 0.17 so the accuracy is higher.

Statistically significant variables with highest order of priorities:-

Age > NumOfProducts > IsActiveMember

Confusion Matrix :-



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4197  313
         1  737  653

               Accuracy : 0.822
                 95% CI : (0.812, 0.8317)
    No Information Rate : 0.8363
    P-Value [Acc > NIR] : 0.9984

                  Kappa : 0.4476

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8506
            Specificity : 0.6760
         Pos Pred Value : 0.9306
         Neg Pred Value : 0.4698
             Prevalence : 0.8363
         Detection Rate : 0.7114
   Detection Prevalence : 0.7644
      Balanced Accuracy : 0.7633

       'Positive' Class : 0

>
```

Model is statistically significant -> lower p value

Accuracy of the model -> 0.82

The specificity of this higher than in logistic regression model. Thus classification bias towards ‘1’ class is reduced as compared to logistic regression model.

Comparison between tree and Logistic regression model

Accuracy of tree based model was significantly higher (82%) than in logistic regression model (70.44%).

Logistic regression	**final_data**
SPECIFICITY – 0.4234	0: 4510
SENSITIVITY – 0.8855	1: 1390

Decesion tree

SPECIFICITY – 0.6760

SENSITIVITY – 0.8560

Conclusion

Despite of the high class imbalance the decision tree model has higher accuracy and statistical significant.