Assignment 1 - Machine Learning

Snehitha Anpur

2022-09-11

The dataset considered here is Bank Churn Data downloaded from the below website: https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset

```
library(readr)
BankCustomerChurnPrediction <- read_csv("BankChurnData/BankCustomerChurnPrediction.csv")
```

```
## Rows: 10000 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): country, gender
## dbl (10): customer_id, credit_score, age, tenure, balance, products_number, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
View(BankCustomerChurnPrediction)
attach(BankCustomerChurnPrediction)
```

Here are the descriptive statistics and of all the variables in the dataset:

```
summary(BankCustomerChurnPrediction)
```

```
##   customer_id        credit_score     country             gender
##  Min.   :15565701   Min.   :350.0   Length:10000       Length:10000
##  1st Qu.:15628528   1st Qu.:584.0   Class :character   Class :character
##  Median :15690738   Median :652.0   Mode  :character   Mode  :character
##  Mean   :15690941   Mean   :650.5
##  3rd Qu.:15753234   3rd Qu.:718.0
##  Max.   :15815690   Max.   :850.0
##       age            tenure          balance       products_number
##  Min.   :18.00   Min.   : 0.000   Min.   :     0   Min.   :1.00
##  1st Qu.:32.00   1st Qu.: 3.000   1st Qu.:     0   1st Qu.:1.00
##  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
##   credit_card     active_member    estimated_salary        churn
##  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
```

Here are the frequency and proportion tables of categorical variables:

```
table1 <- table(country,gender)
table1
```

```
##          gender
## country   Female Male
##   France    2261 2753
##   Germany   1193 1316
##   Spain     1089 1388
```

```
prop.table(table1)
```

```
##          gender
## country   Female   Male
##   France  0.2261 0.2753
##   Germany 0.1193 0.1316
##   Spain   0.1089 0.1388
```

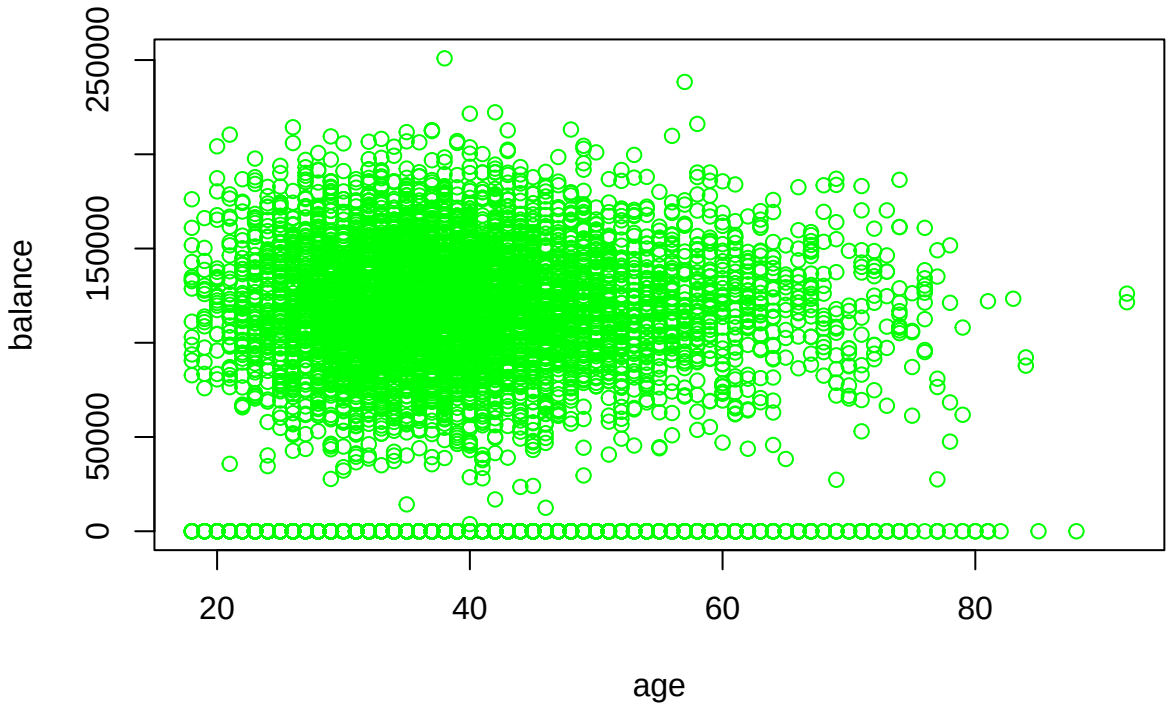
Here are the data transformations: - log transformation of a variable “age” - square root of the variable “balance”

```
BankCustomerChurnPrediction$log_age <- log10(age)
BankCustomerChurnPrediction$sqrt_bal <- sqrt(balance)
head(BankCustomerChurnPrediction)
```

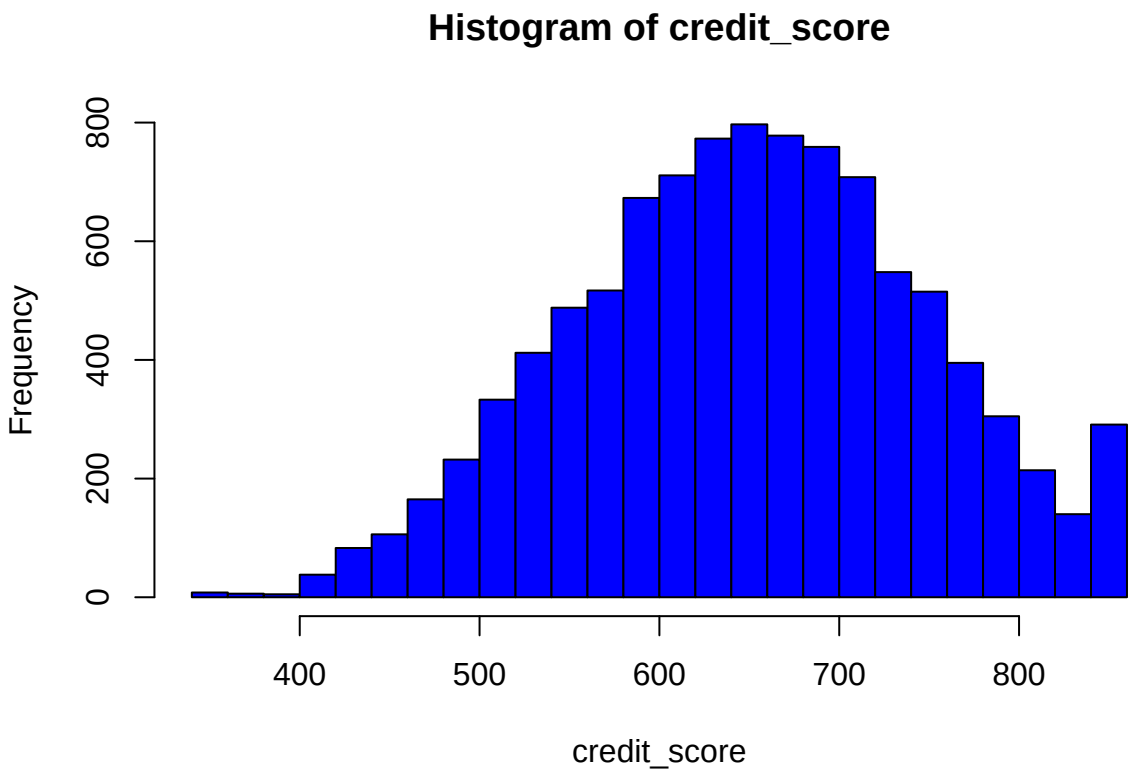
```
## # A tibble: 6 x 14
##   customer~1 credi~2 country gender   age tenure balance produ~3 credi~4 activ~5
##        <dbl>   <dbl> <chr>   <chr>  <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1   15634602     619 France  Female    42      2      0        1       1       1
## 2   15647311     608 Spain   Female    41      1  83808.       1       0       1
## 3   15619304     502 France  Female    42      8 159661.       3       1       0
## 4   15701354     699 France  Female    39      1      0        2       0       0
## 5   15737888     850 Spain   Female    43      2 125511.       1       1       1
## 6   15574012     645 Spain   Male      44      8 113756.       2       1       0
## # ... with 4 more variables: estimated_salary <dbl>, churn <dbl>,
## #   log_age <dbl>, sqrt_bal <dbl>, and abbreviated variable names
## #   1: customer_id, 2: credit_score, 3: products_number, 4: credit_card,
## #   5: active_member
```

Below are the plots for quantitative variables: # Scatter Plot of Age vs Balance # Histogram of Credit_Score # Density graph of credit_score # Density graph of sqrt_bal # Density graph of log_age

```
plot(age,balance,col = 'green') # Scatter Plot of Age vs Balance
```



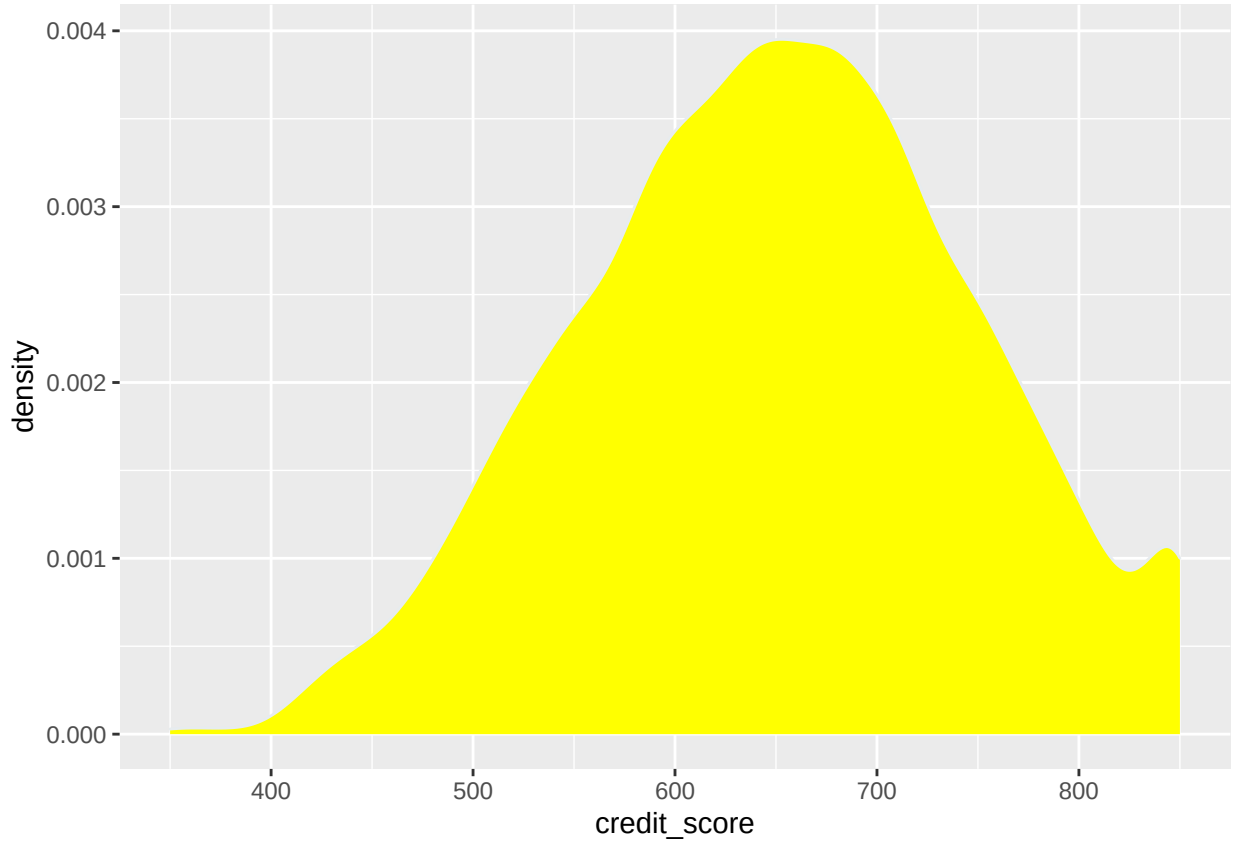

```
hist(credit_score,breaks = 20,col = 'blue') # Histogram of Credit_Score

#install.packages("ggplot2") # Install package ggplot2
library(ggplot2) # Initializing ggplot2 library
```



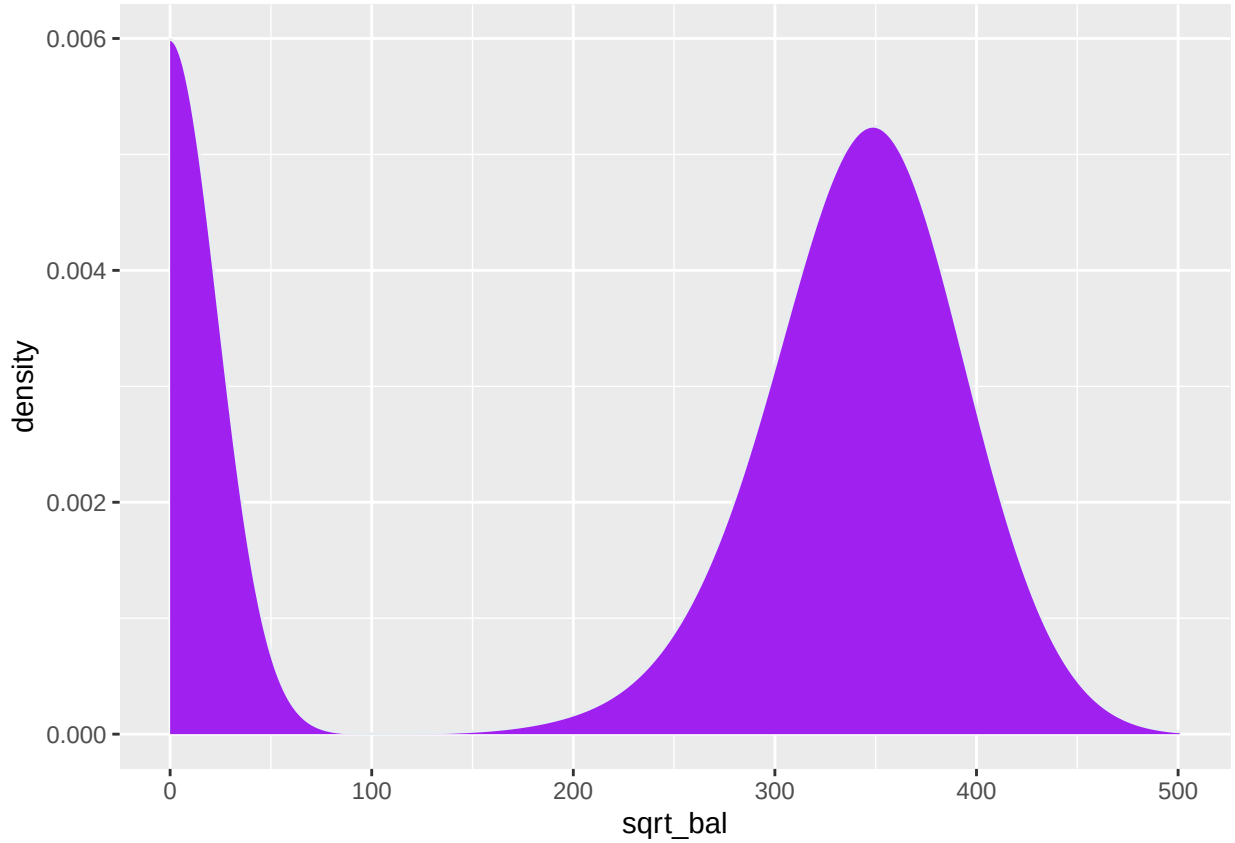

```
#install.packages("dplyr") # Install package dplyr
library(dplyr) # Initializing dplyr library
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

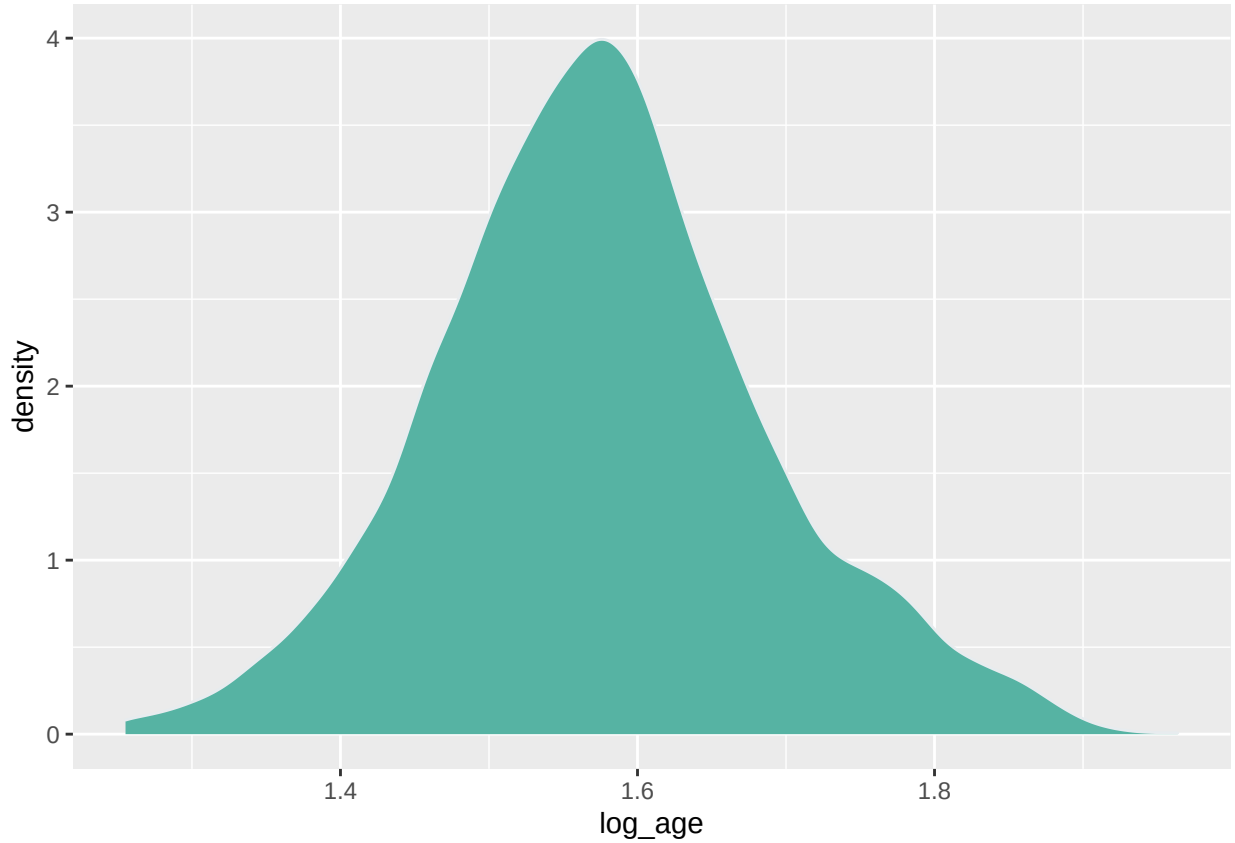
```
BankCustomerChurnPrediction %>%
  ggplot( aes(x=credit_score)) +
    geom_density(fill="yellow", color="#e7ecef") # Density graph of credit_score
```




```
BankCustomerChurnPrediction %>%
  ggplot( aes(x=sqrt_bal)) +
    geom_density(fill="purple", color="#e7ecef") # Density graph of sqrt_bal
```

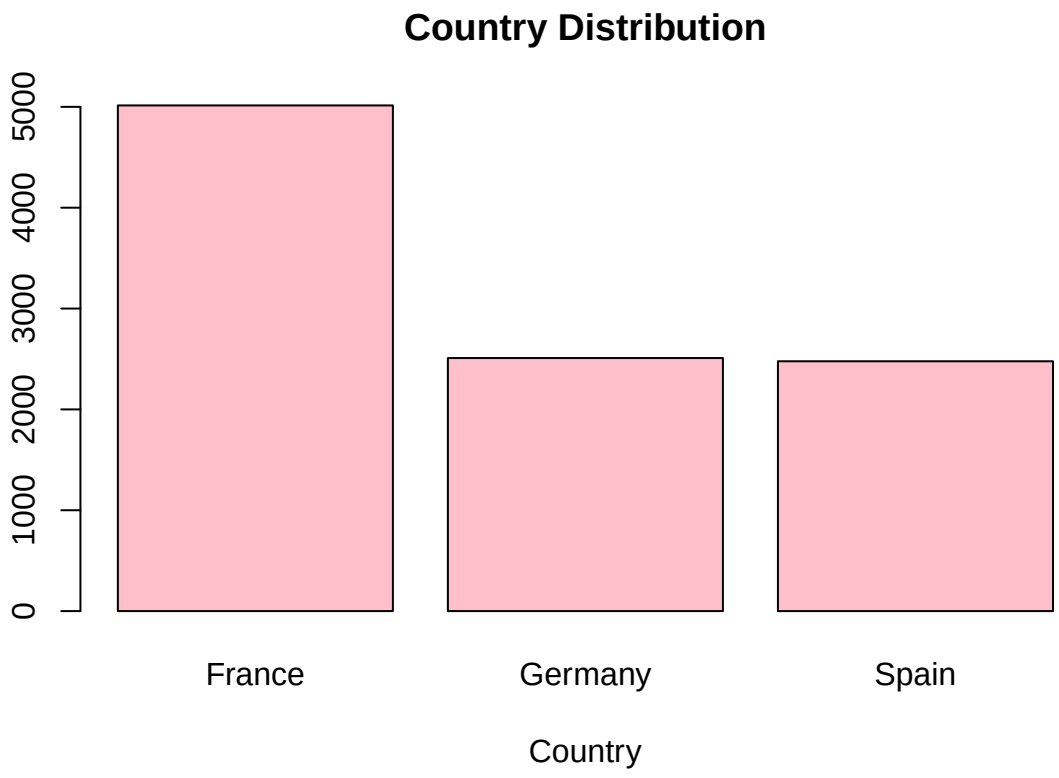



```
BankCustomerChurnPrediction %>%
  ggplot( aes(x=log_age)) +
    geom_density(fill="#56b3a3", color="#e7ecef") # Density graph of log_age
```

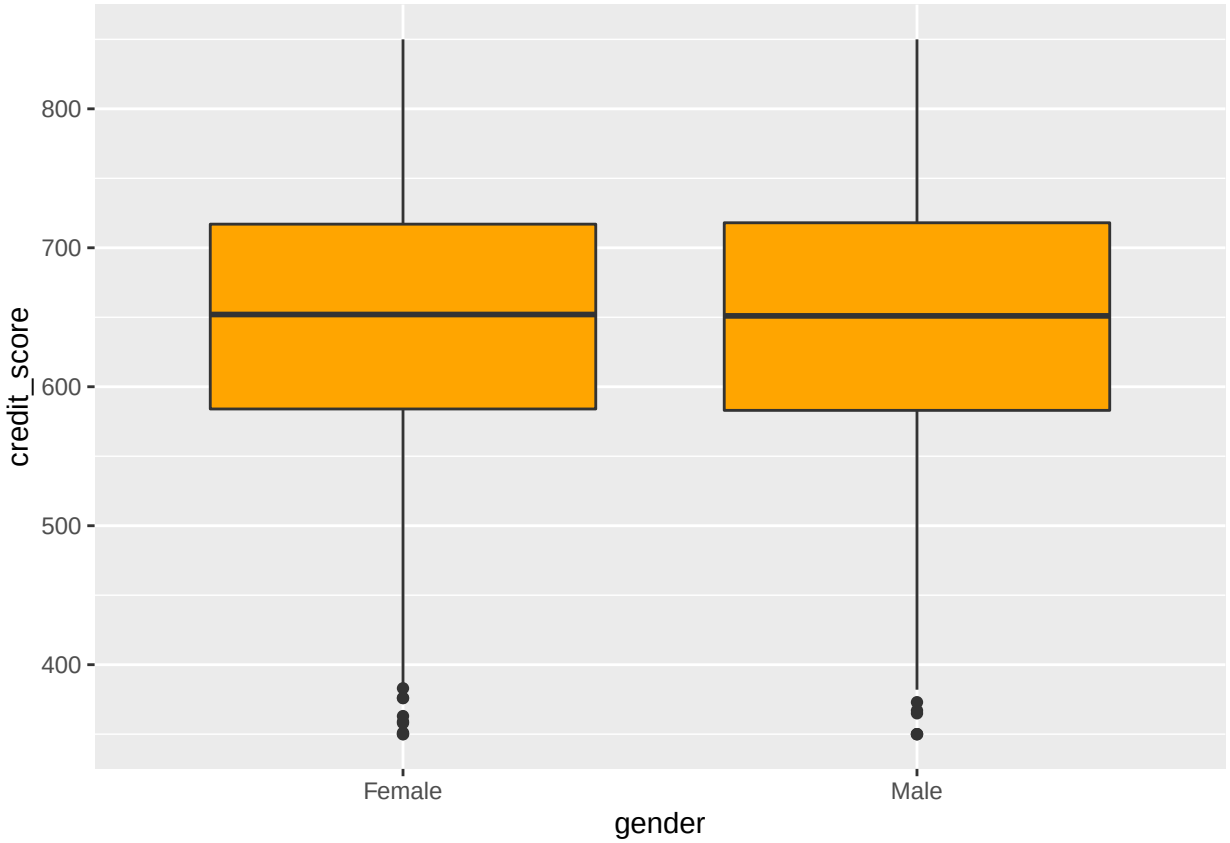


Below are the plots for categorical variables: # Bar plot of Country # Box plot of credit_score over gender

```
counts <- table(country)
barplot(counts, main = "Country Distribution", xlab = "Country", col = "pink") # Barplot of Country
```




```
ggplot(BankCustomerChurnPrediction, aes(x=gender, y=credit_score)) +  geom_boxplot(fill='orange') # Box
```


800
700
600
500
400
credit_score
Female
Male
gender