

# UK Bank Customer Analysis

Hitesh Bhamre

2022-12-23

## Dataset

Notes: The dataset is taken from a UK Bank's database. It includes details of their customer. The dataset contains 9 variables and 4014 rows. The data was processed using in R studio and the following steps were performed.

## Loading Packages

```
library(psych)
library(tidy)
library(lubridate)
library(dplyr)
library(tidyverse)
library(ggplot2)
library(plotrix)
library(plotly)
```

## Import Data

```
setwd("D:\\R Project\\UK_Bank_Customer")
bank = read.csv("UK_Bank_Customer.csv")
```

```
View(bank)
head(bank)
```

```
##   Customer.ID   Name Surname Gender Age      Region Job.Classification
## 1           1   Simon   Walsh  Male  21      England   White Collar
## 2           2  Jasmine   Miller Female 34 Northern Ireland   Blue Collar
## 3           3    Liam   Brown  Male  46      England   White Collar
## 4           4  Trevor   Parr  Male  32      Wales     White Collar
## 5           5 Deirdre Pullman Female 38      England   Blue Collar
## 6           6   Ava Coleman Female 30      Wales     Blue Collar
##
##   Date.Joined Balance
## 1 January 5, 2015    367
## 2 January 6, 2015    319
## 3 January 7, 2015    146
## 4 January 8, 2015    321
## 5 January 9, 2015    165
## 6 January 9, 2015    283
```

## Exploratory Data Analysis

```
dim(bank)
```

```
## [1] 4014    9
```

```
str(bank)
```

```
## 'data.frame':    4014 obs. of  9 variables:
##  $ Customer.ID      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name             : chr  "Simon" "Jasmine" "Liam" "Trevor" ...
##  $ Surname          : chr  "Walsh" "Miller" "Brown" "Parr" ...
##  $ Gender           : chr  "Male" "female" "Male" "Male" ...
##  $ Age              : int  21 34 46 32 38 30 34 48 NA 42 ...
##  $ Region           : chr  "England" "Northern Ireland" "England" "Wales" ...
##  $ Job.Classification: chr  "White Collar" "Blue Collar" "White Collar" "White Collar" ...
##  $ Date.Joined      : chr  "January 5, 2015" "January 6, 2015" "January 7, 2015" "January 8, 2015" ...
##  $ Balance          : int  367 319 146 321 165 283 361 433 39 113 ...
```

```
describe(bank)
```

```
##           vars      n   mean      sd median trimmed   mad min   max
## Customer.ID      1 4014 2007.50 1158.89 2007.5 2007.50 1487.79  1 4014
## Name*            2 4014   85.23   49.39   85.0   84.88   63.75  1 172
## Surname*         3 4014   75.26   42.99   75.5   75.25   54.11  1 150
## Gender*          4 4014   1.54    0.50    2.0   1.55    0.00  1   2
## Age             5 3999   38.69    9.83   37.0   38.21   19.38 15  64
## Region*         6 4014    2.09    1.16    1.0    1.88    0.09  1   4
## Job.Classification* 7 4014    2.23    0.84    2.0    2.28    1.48  1   3
## Date.Joined*     8 4014   170.54   90.07  188.5   173.02   113.42  1 307
## Balance          9 4014   250.76  140.85  248.5   249.92   178.65 10  500
```

```
##           range      skew kurtosis   se
## Customer.ID      4013  0.00   -1.20 18.29
## Name*            171  0.04   -1.17  0.78
## Surname*         149  0.00   -1.16  0.68
## Gender*           1 -0.16   -1.98  0.01
## Age              49  0.35   -0.42  0.16
## Region*           3  0.50   -1.38  0.02
## Job.Classification* 2 -0.44   -1.43  0.01
## Date.Joined*     306 -0.21   -1.30  1.42
## Balance          490  0.04   -1.19  2.22
```

```
summary(bank)
```

```
##   Customer.ID   Name      Surname      Gender
## Min.   < 1      Length:4014      Length:4014      Length:4014
## 1st Qu.:1004    Class :character    Class :character    Class :character
## Median :2008    Mode  :character    Mode  :character    Mode  :character
## Mean   :2008
## 3rd Qu.:3011
## Max.   :4014
##
##   Age      Region      Job.Classification Date.Joined
## Min.   <15.0      Length:4014      Length:4014      Length:4014
## 1st Qu.:31.0     Class :character    Class :character    Class :character
## Median :37.0     Mode  :character    Mode  :character    Mode  :character
## Mean   :38.6
## 3rd Qu.:45.0
## Max.   :64.0
## NA's   :15
##   Balance
## Min.   <10.0
## 1st Qu.:130.0
## Median :248.5
## Mean   :250.8
## 3rd Qu.:371.0
## Max.   :500.0
##
```

## Data Cleansing

### Change column Names

```
bank <- rename(bank, Customer_ID = Customer.ID,
               Job_Classification = Job.Classification,
               Date = Date.Joined,
               Deposit = Balance,
               First_Name = Name,
               Last_Name = Surname)
```

```
head(bank)
```

```
##   Customer_ID First_Name Last_Name Gender Age      Region
## 1           1    Simon   Walsh  Male  21      England
## 2           2  Jasmine   Miller Female 34 Northern Ireland
## 3           3    Liam   Brown  Male  46      England
## 4           4  Trevor   Parr  Male  32      Wales
## 5           5 Deirdre Pullman Female 38      England
## 6           6   Ava Coleman Female 30      Wales
##
##   Job_Classification Date Deposit Month
## 1   White Collar January 5, 2015    367   Jan
## 2   Blue Collar January 6, 2015    319   Jan
## 3   White Collar January 7, 2015    146   Jan
## 4   White Collar January 8, 2015    321   Jan
## 5   Blue Collar January 9, 2015    165   Jan
## 6   Blue Collar January 9, 2015    283   Jan
```

### Change Datatype

Notes: Changing Datatype of Date Variable (char) into Date format

```
bank$Date <- strptime(bank$Date, "%B %d, %Y")
bank$Age <- as.integer(bank$Age)
```

```
class(bank$Date)
```

```
## [1] "POSIXlt" "POSIXt"
```

```
class(bank$Age)
```

```
## [1] "integer"
```

### Checking NA Values

```
sum(is.na(bank$Age))
```

```
## [1] 15
```

### Replace NA values with Age mean base on gender

```
bank$Age[is.na(bank$Age)] =
  mean((bank$Age[bank$Gender == "Male"]), na.rm=TRUE)
bank$Age[is.na(bank$Age)] =
  mean((bank$Age[bank$Gender == "Female"]), na.rm=TRUE)
```

```
sum(is.na(bank$Age))
```

```
## [1] 0
```

### Extract Month From Date

```
bank$Month <- lubridate::month(bank$Date,
                              label = TRUE,
                              abbr = TRUE)
```

```
head(bank)
```

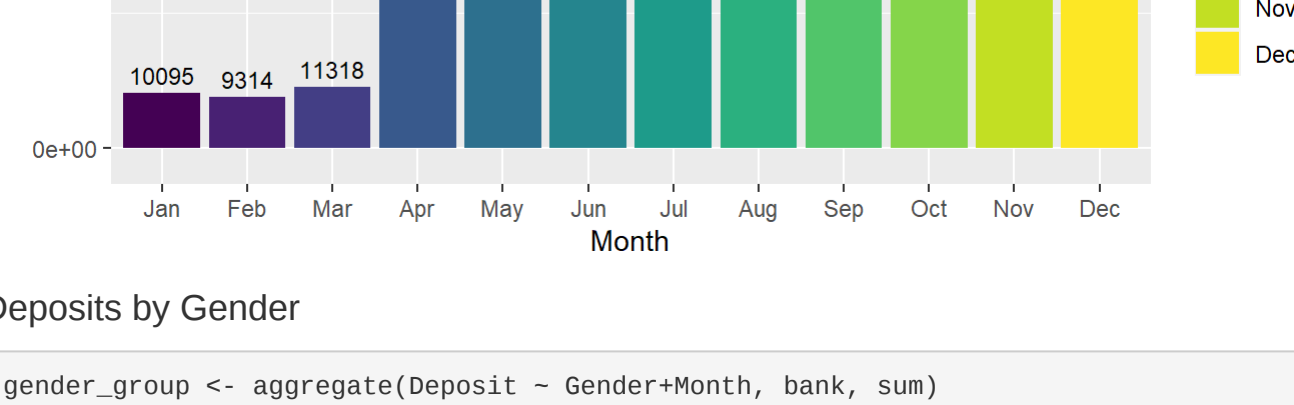
```
##   Customer_ID First_Name Last_Name Gender Age      Region
## 1           1    Simon   Walsh  Male  21      England
## 2           2  Jasmine   Miller Female 34 Northern Ireland
## 3           3    Liam   Brown  Male  46      England
## 4           4  Trevor   Parr  Male  32      Wales
## 5           5 Deirdre Pullman Female 38      England
## 6           6   Ava Coleman Female 30      Wales
##
##   Job_Classification Date Deposit Month
## 1   White Collar 2015-01-05    367   Jan
## 2   Blue Collar 2015-01-06    319   Jan
## 3   White Collar 2015-01-07    146   Jan
## 4   White Collar 2015-01-08    321   Jan
## 5   Blue Collar 2015-01-09    165   Jan
## 6   Blue Collar 2015-01-09    283   Jan
```

## Visualization Analysis

### Total Deposit by Month

```
month_group <- aggregate(Deposit ~ Month, bank, sum)
```

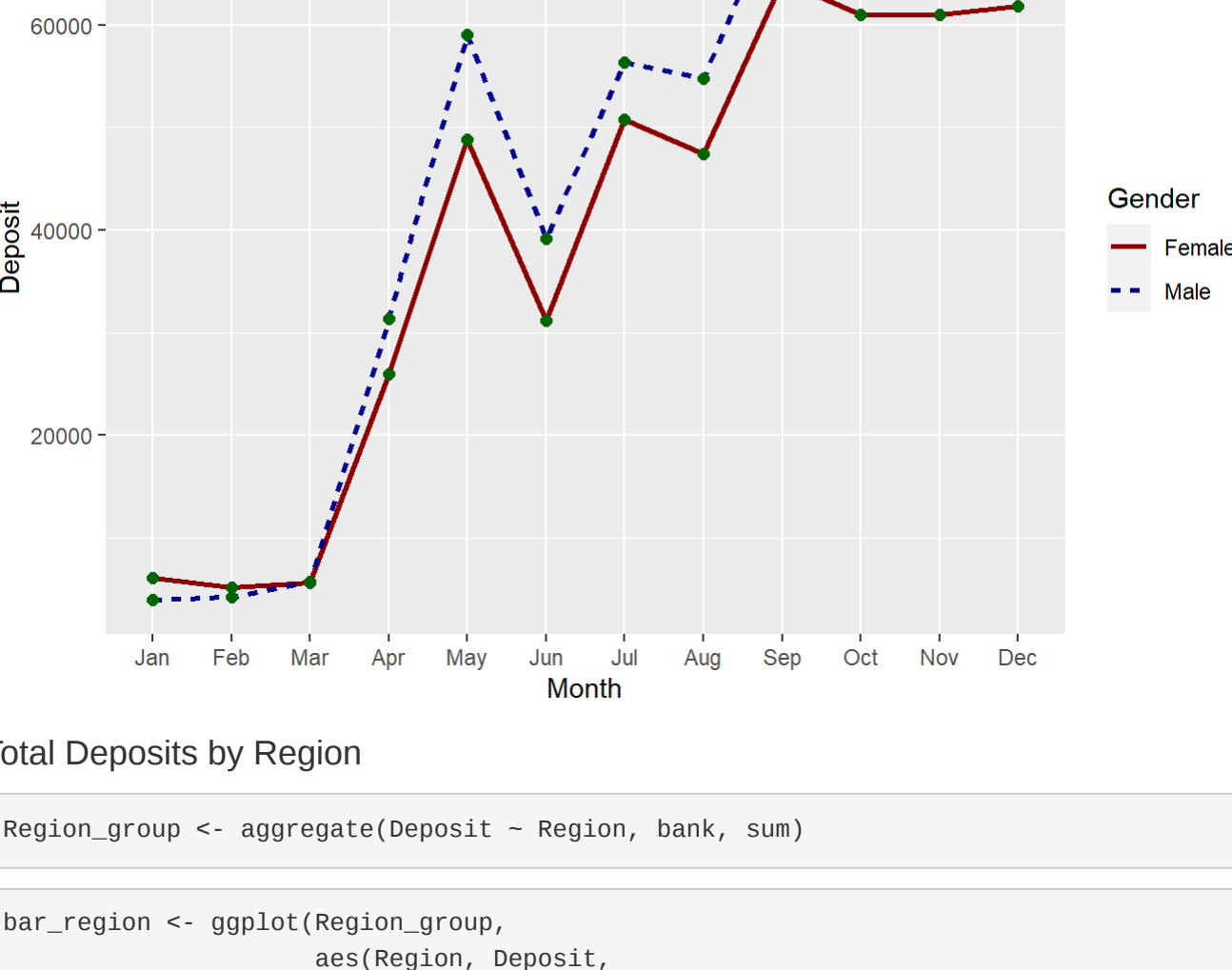
```
bar_chart_month <- ggplot(month_group,
                          aes(Month, Deposit,
                              fill = Month))+
  geom_bar(stat="identity")+
  geom_text(aes(label = Deposit),
            position=position_dodge(width=0.5),
            vjust=-0.50,
            size = 3)+
  theme_grey()+
  ggtitle("Deposits by Month")
bar_chart_month
```



### Deposits by Gender

```
gender_group <- aggregate(Deposit ~ Gender+Month, bank, sum)
```

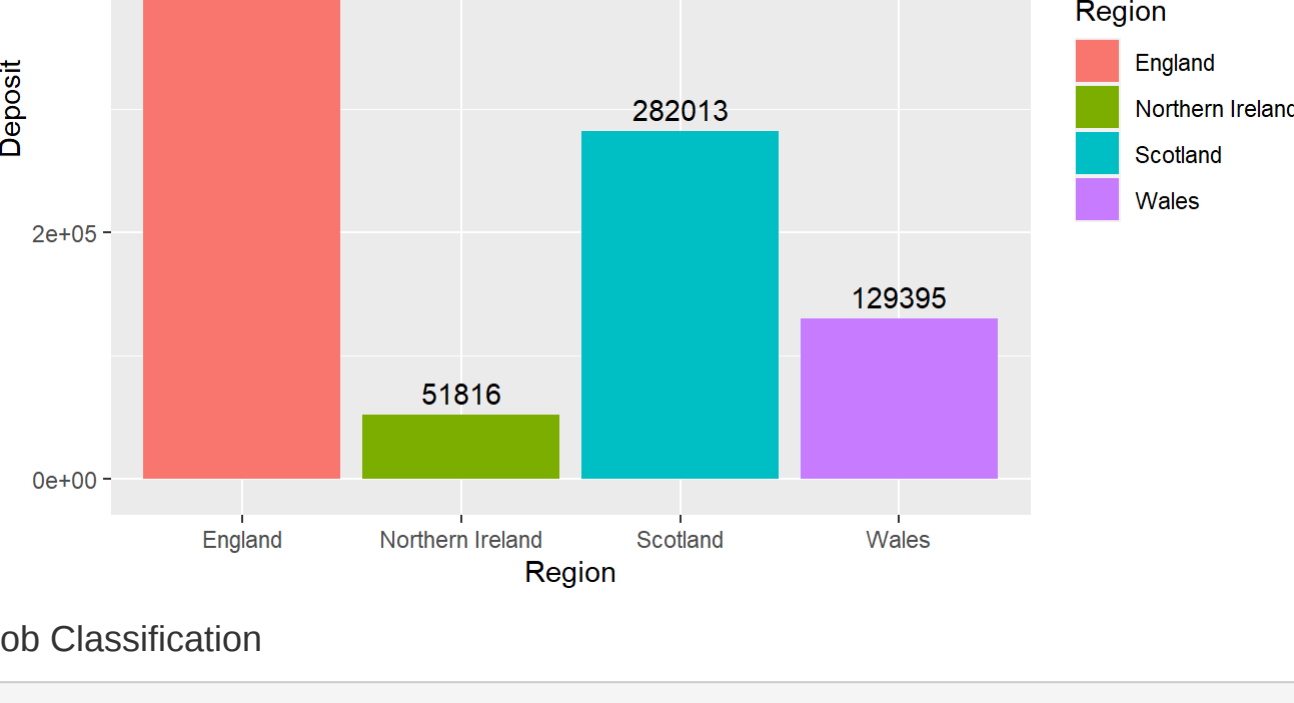
```
line_chart_month <- ggplot(gender_group, aes(x = Month,
                                              y = Deposit,
                                              group = Gender)) +
  geom_line(aes(color = Gender,
                 size = 1) +
            scale_color_manual(values = c("darkred",
                                           "darkblue")))+
  ggtitle("Deposits by Month")+
  geom_point(size = 2, color = "darkgreen")
line_chart_month
```



### Total Deposits by Region

```
Region_group <- aggregate(Deposit ~ Region, bank, sum)
```

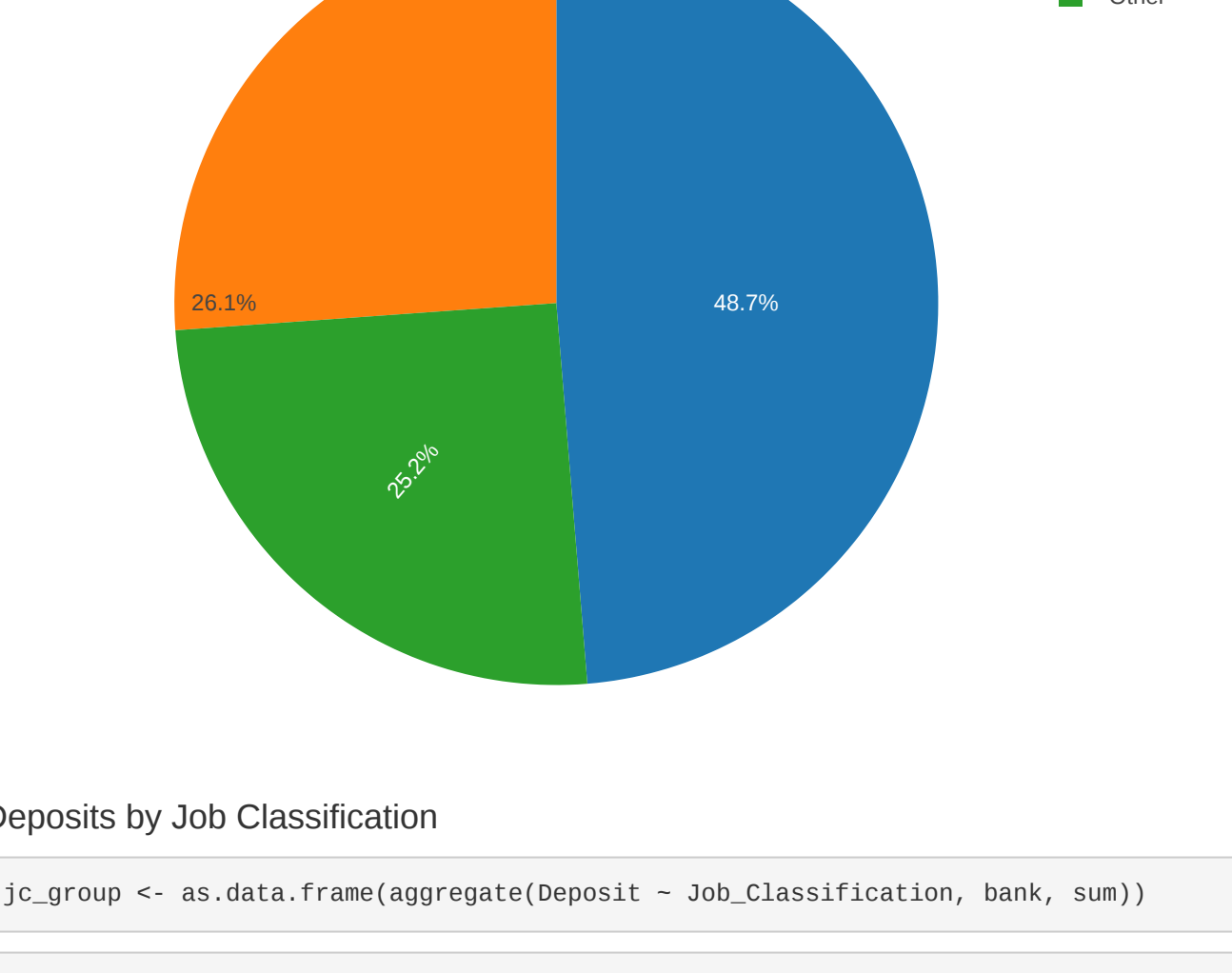
```
bar_region <- ggplot(Region_group,
                    aes(Region, Deposit,
                        fill = Region))+
  geom_bar(stat="identity")+
  geom_text(aes(label = Deposit),
            position=position_dodge(width=0.5),
            vjust=-0.50)+
  ggtitle("Region by Deposits")+
  ylin(0,600000)
bar_region
```



### Job Classification

```
jc_table <- table(bank$Job_Classification)
job <- as.data.frame(jc_table)
job<- rename(job, Job_Category = Var1)
```

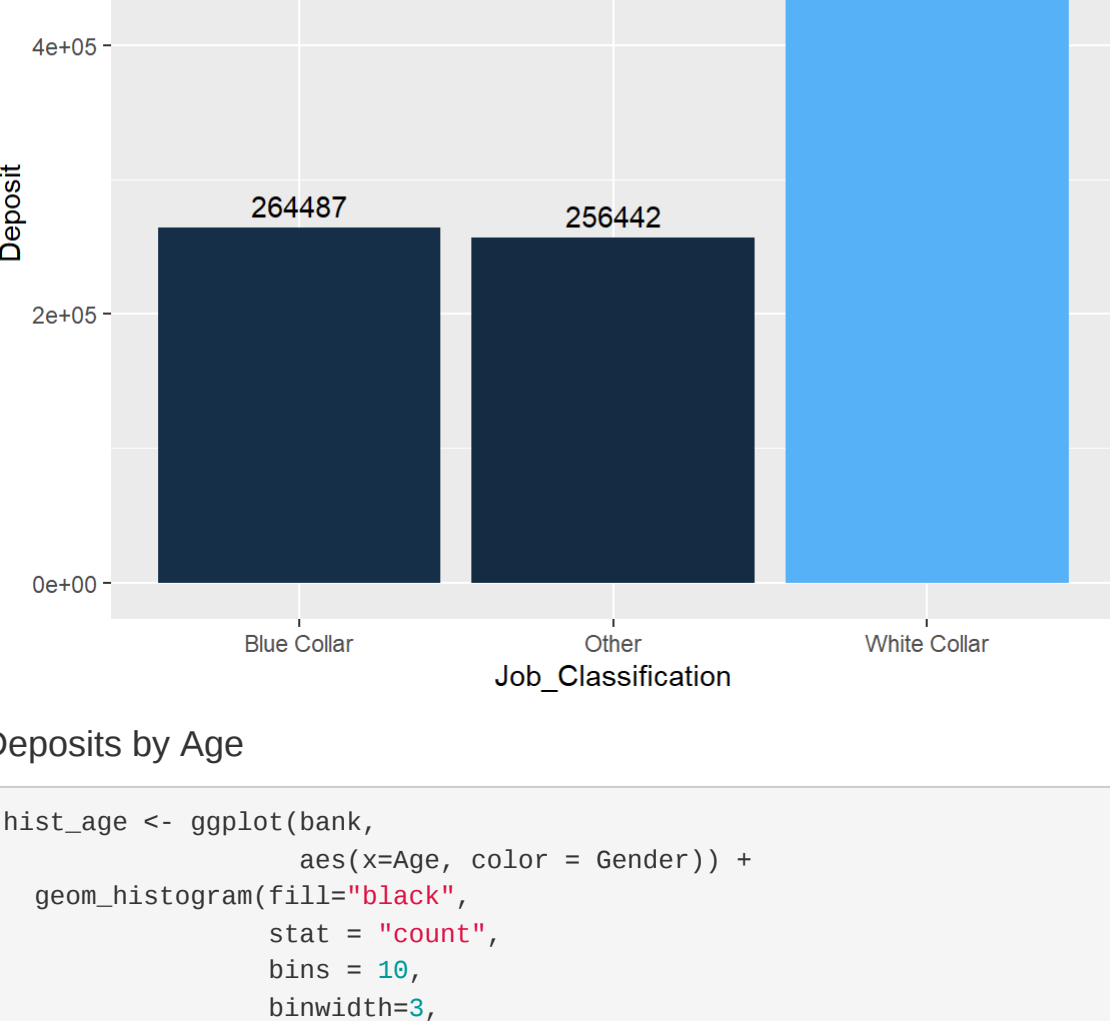
```
pie_chart_job <- plot_ly(data = job, title="Job Classification",
                        type='pie', labels=~Job_Category,
                        values=~Freq,
                        textinfo=~Freq,
                        insidetextorientation='radial')
pie_chart_job
```



### Deposits by Job Classification

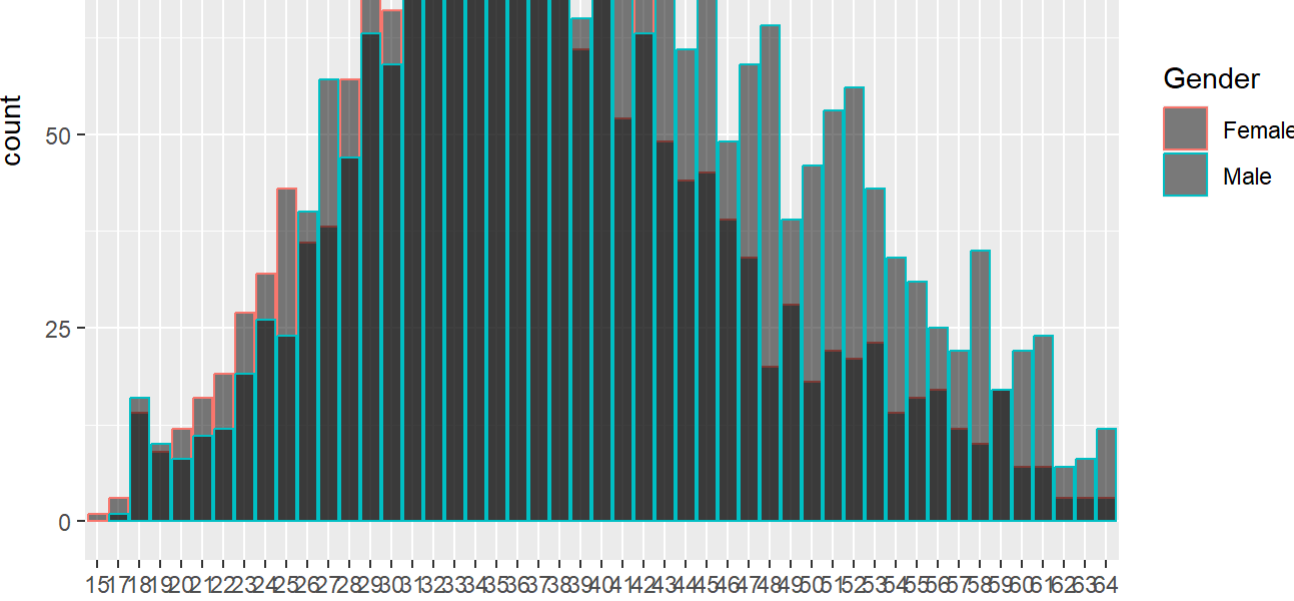
```
jc_group <- as.data.frame(aggregate(Deposit ~ Job_Classification, bank, sum))
```

```
bar_job <- ggplot(jc_group,
                  aes(x=Job_Classification,
                      y=Deposit,
                      fill = Job_Classification))+
  geom_bar(stat="identity")+
  geom_text(aes(label = Deposit),
            position=position_dodge(width=0.5),
            vjust=-0.50)+
  ggtitle("Job Category by Deposits")+
  ylin(0,550000)
bar_job
```



### Deposits by Age

```
hist_age <- ggplot(bank,
                   aes(x=Age, color = Gender)) +
  geom_histogram(fill="black",
                 stat = "count",
                 bins = 10,
                 binwidth=3,
                 alpha=0.5, position="identity")
hist_age
```



#bins not supported to the html file