**ASSIGNMENT**

**Introduction to R Programming**

**Rakshan Kumar**

**PGID:2002025**

**Prof. Savio Saldanha**

**14.03.2021**

**Step 1**: Importing the data using url as csv file and by using chain operations cleaning the unnecessary data.

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

d1<- read.csv(url)

library(dplyr)

library(ggplot2)

***#Removing rownumber and customer id columns and then creating converting into factor data the columns for gender, is\_active\_member, has\_credit\_card, exited and tenure***

d2<-d1 %>%

dplyr::select(-RowNumber, -CustomerId, -Surname) %>% #remove unwanted column

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))

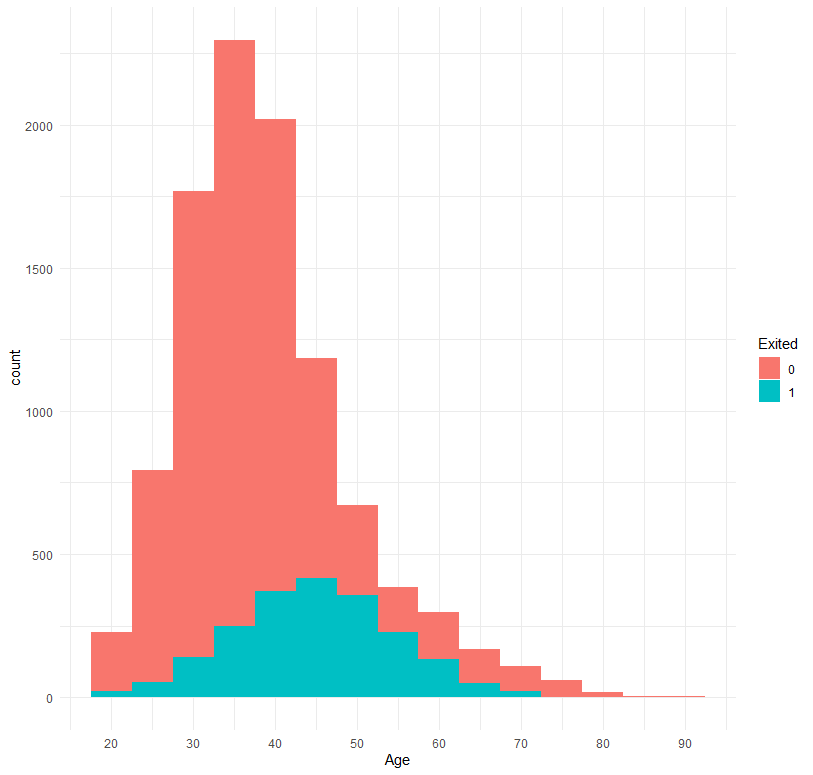
**1. Plotting the churn numbers with respect to age**

a. ggplot(d2, aes(x = Age, fill = Exited)) +

geom\_histogram(binwidth = 5) +

theme\_minimal() +

scale\_x\_continuous(breaks = seq(0,100,by=10), labels = comma)



Based on the age am trying to identify different age group are exited and not exited.

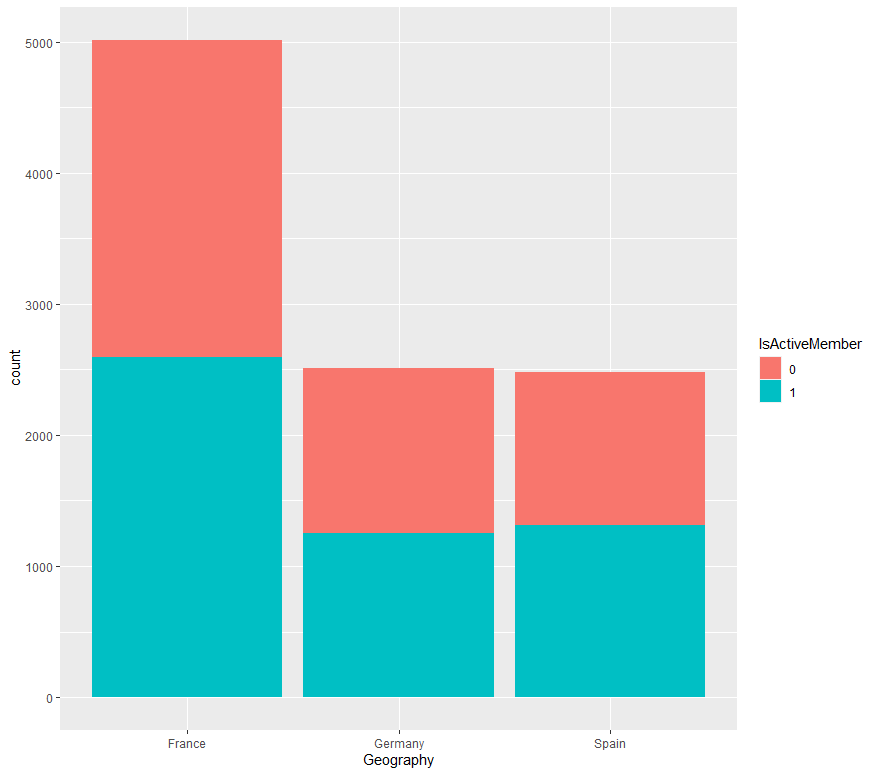
If 0 is exited there are many people who are exited for the products and the age group between 30-50 are more exited.

Then coming to 1 the people with the age between 20-30 are high who are not exited. But the range 40-65 are increased to exited.

**b. The number of active members country-wise**

ggplot(d2, aes(Geography, fill = IsActiveMember)) +

geom\_bar()



As looking at geography of three countries France, Germany, Spain. There are more active and non-active members in France compare to remaining two. If we see there are more in-active members in France rather than two.

**c. Box plot of the customers to identify how many have exited based on age**

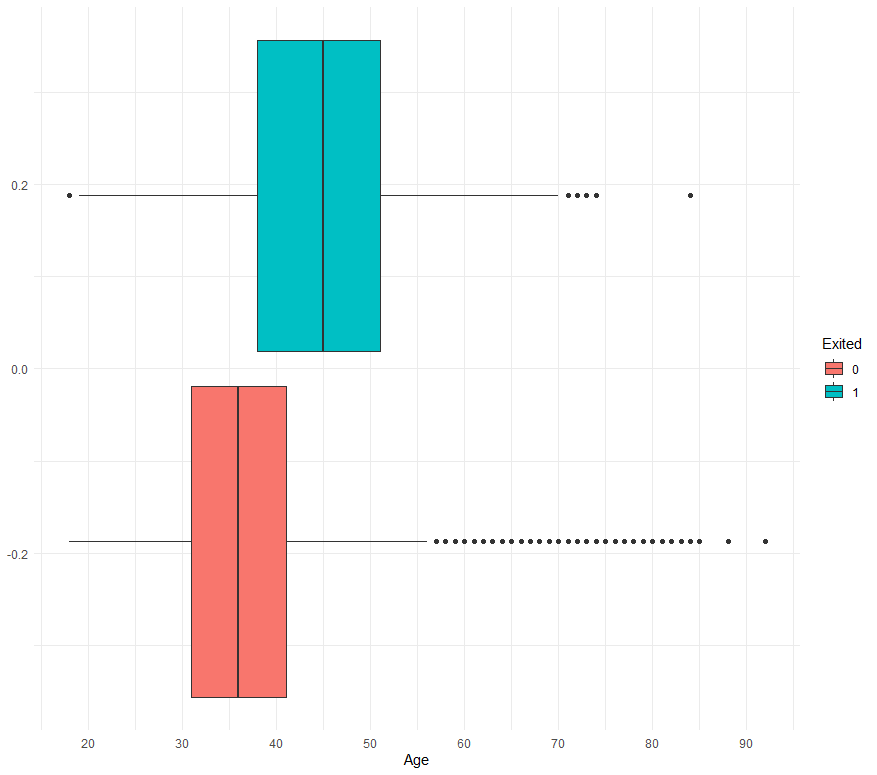
library(scales)

ggplot(d2, aes(x = Age, fill = Exited)) +

geom\_boxplot(binwidth = 5) +

theme\_minimal() +

scale\_x\_continuous(breaks = seq(0,100,by=10), labels = comma)



Here in the above graph there more non exited people in the age group of 40-50 compare to the exited people in the age group 30-40.

**d. Plotting the churn ratio of the active members**

library(tidyr)

library(tidyselect)

library(tidyverse)

d2 %>%

dplyr::select(IsActiveMember, Exited) %>%

table(.) %>%

as.data.frame() %>%

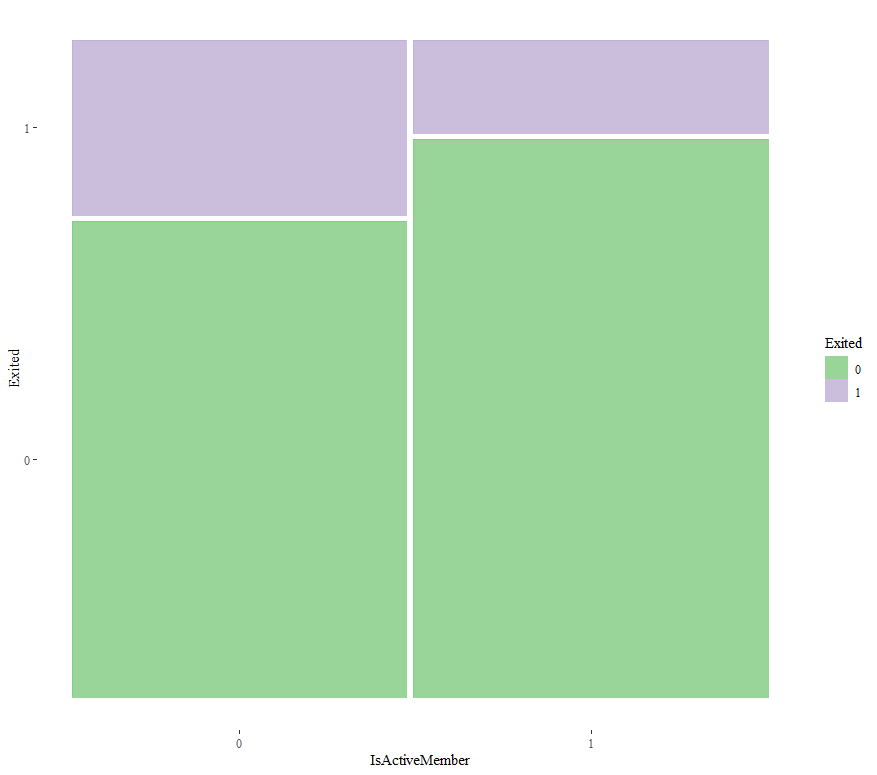
ggplot(.) +

ggmosaic::geom\_mosaic(aes(weight = Freq, x = product(IsActiveMember), fill = Exited)) +

ggthemes::theme\_tufte() +

scale\_fill\_brewer(type = "qual") +

labs(x = 'IsActiveMember')



In the above graph there are more active members who are exited compare the who are non-exited. There are active members who are more exited and less non exited people in the group 1.

**e. Plotting tenure versus churn ratio**

library(ggmosaic)

library(ggthemes)

d2 %>%

dplyr::select(Tenure, Exited) %>%

table(.) %>%

as.data.frame() %>%

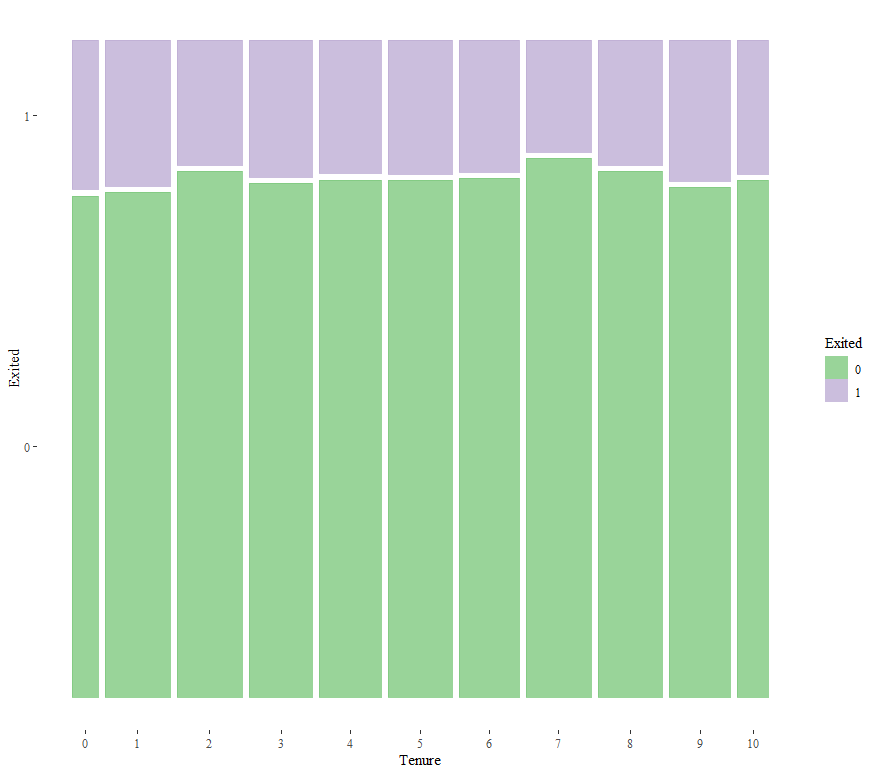
ggplot(.) +

ggmosaic::geom\_mosaic(aes(weight = Freq, x = product(Tenure), fill = Exited)) +

ggthemes::theme\_tufte() +

scale\_fill\_brewer(type = "qual") +

labs(x = 'Tenure')



In the above graph there are more exited people in the tenure of 7 and more or less are exited with the same average and there are few people who are non-exited.

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

d2 %>% select(CreditScore, Gender, Geography) %>% filter(Geography == "France") %>%

dplyr::group\_by(Gender) %>%

dplyr::summarise(Gender\_Average = mean(CreditScore))

**Output:**

Gender Gender\_Average

<fct> <dbl>

1 Female 649.

2 Male 650.

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

d2 %>% select(CreditScore, Age) %>% mutate(agegroup = case\_when(Age >= 41 & Age <= 50 ~ '3', Age >= 31 & Age <= 40 ~ '2', Age >= 20 & Age <= 30 ~ '1')) %>%

filter(agegroup == "1" | agegroup == '2' | agegroup == '3') %>%

dplyr::group\_by(agegroup) %>%

dplyr::summarise(Age\_Average = mean(CreditScore))

**Output:**

# A tibble: 3 x 2

agegroup Age\_Average

<chr> <dbl>

1 1 651.

2 2 651.

3 3 649.

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

d2 %>% select(CreditScore, EstimatedSalary) %>% cor()

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

# Create the relationship model.

model <- lm(CreditScore ~Gender+Age+EstimatedSalary, data = d2)

# Show the model.

print(model)

summary(model)

**Output:**

Call:

lm(formula = CreditScore ~ Gender + Age + EstimatedSalary, data = d2)

Coefficients:

(Intercept) GenderMale Age EstimatedSalary

6.525e+02 -5.785e-01 -3.739e-02 -2.416e-06

> summary(model)

Call:

lm(formula = CreditScore ~ Gender + Age + EstimatedSalary, data = d2)

Residuals:

Min 1Q Median 3Q Max

-300.630 -66.880 1.262 66.930 201.174

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.525e+02 4.254e+00 153.382 <2e-16 \*\*\*

GenderMale -5.785e-01 1.942e+00 -0.298 0.766

Age -3.739e-02 9.221e-02 -0.405 0.685

EstimatedSalary -2.416e-06 1.681e-05 -0.144 0.886

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 96.67 on 9996 degrees of freedom

Multiple R-squared: 2.659e-05, Adjusted R-squared: -0.0002735

F-statistic: 0.0886 on 3 and 9996 DF, p-value: 0.9663.