R – PROGRAMMING ASSIGNMENT

Name:Abhinav Reddy Cherlakola

Pgid:2002002

1) Develop 5 different visuals using GGPLOT with descriptions of the insights they convey.

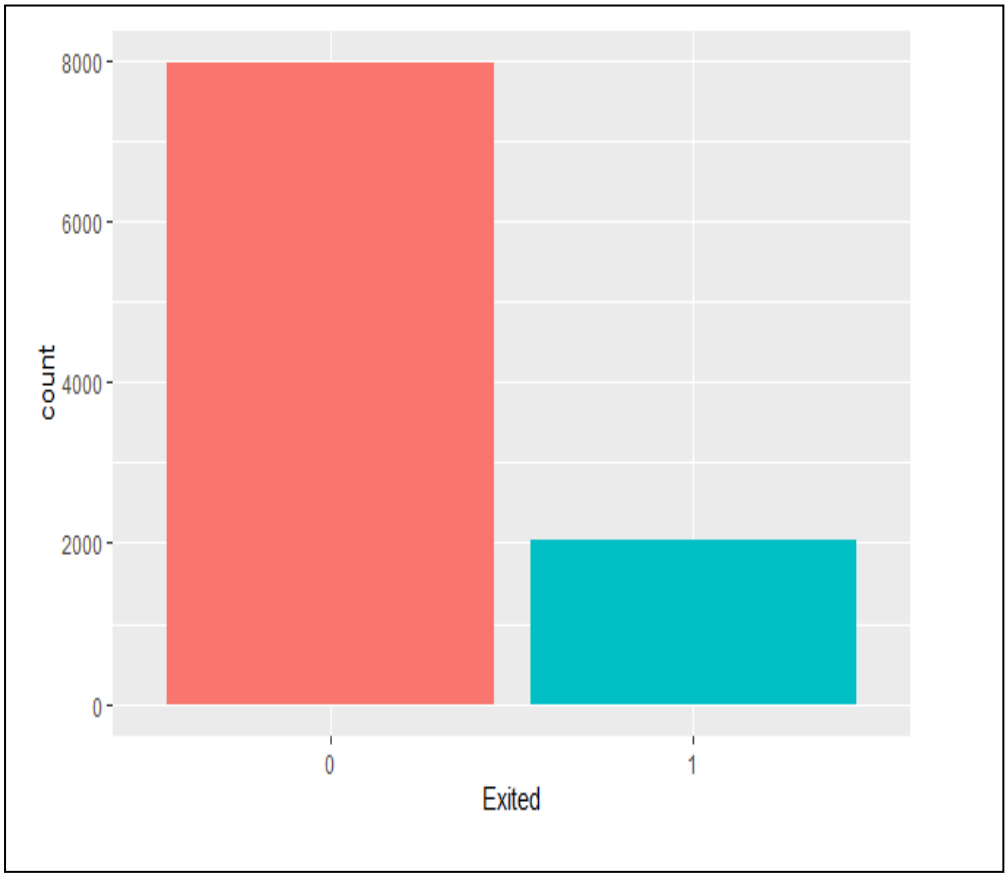
```
url <- "https://github.com/SavioSal/datasets/raw/master/Bank%20Churn_Modelling.csv"
data_1 <- read.csv(url)
data_1
library(dplyr)
library(ggplot2)
#remove unwanted columns
data_2<-data_1 %>%
  dplyr::select(-RowNumber, -CustomerId, -Surname) %>%
  mutate(Geography = as.factor(Geography),
         Gender = as.factor(Gender),
         HasCrCard = as.factor(HasCrCard),
         IsActiveMember = as.factor(IsActiveMember),
         Exited = as.factor(Exited),
         Tenure = as.factor(Tenure),
         NumOfProducts = as.factor(NumOfProducts))
```

I have removed unwanted columns.

Graph no.1

```
ggplot(data_2, aes(Exited, fill = Exited)) +
    geom_bar() +
    theme(legend.position = 'none')
```

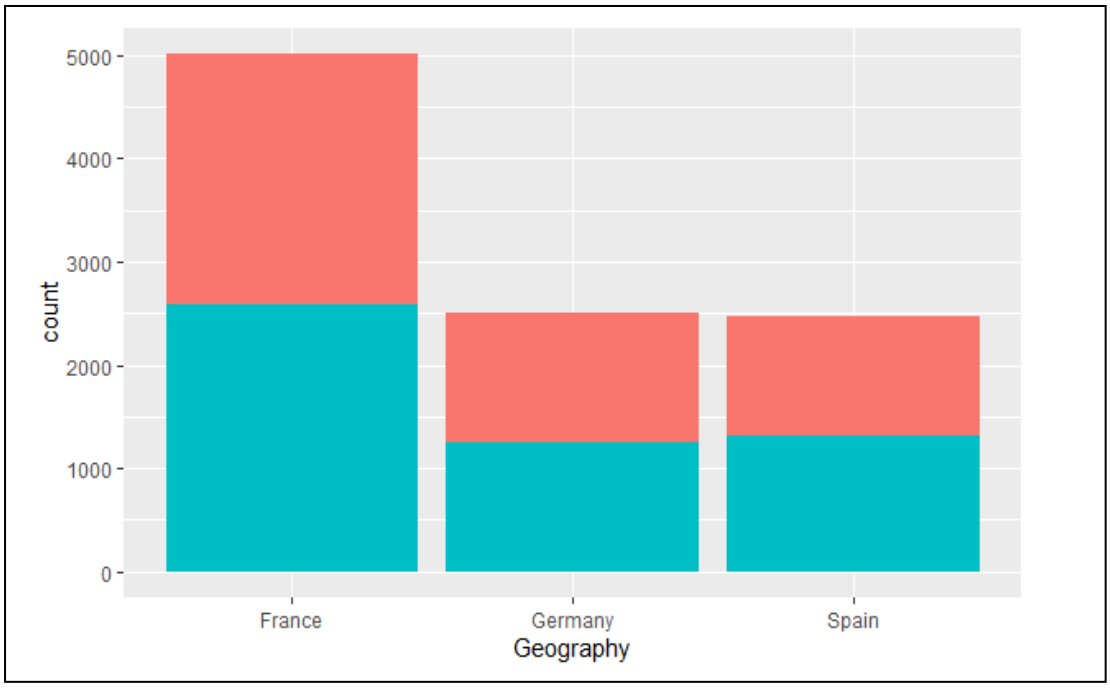
From this graph we can see how many are exited



Graph no.2

```
ggplot(data_2, aes(Geography, fill =IsActiveMember)) +
    geom_bar() +
    theme(legend.position = 'none')
```

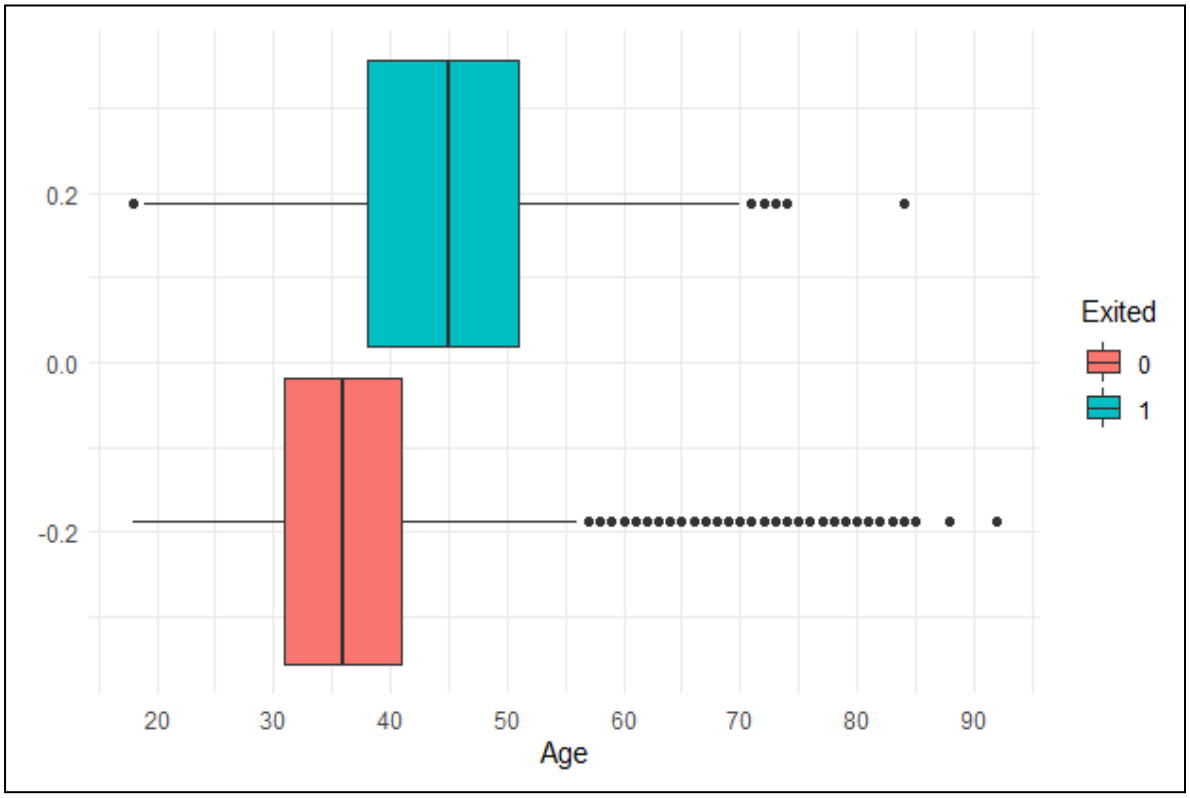
it shows the equal distribution of active and non active members in the countries.



Graph no 3

```
library(scales)
ggplot(data_2, aes(x = Age, fill = Exited)) +
    geom_boxplot(binwidth = 5) +
    theme_minimal() +
    scale_x_continuous(breaks = seq(0,100,by=10), labels = comma)
```

Exited customers are mostly around 40 to 50.

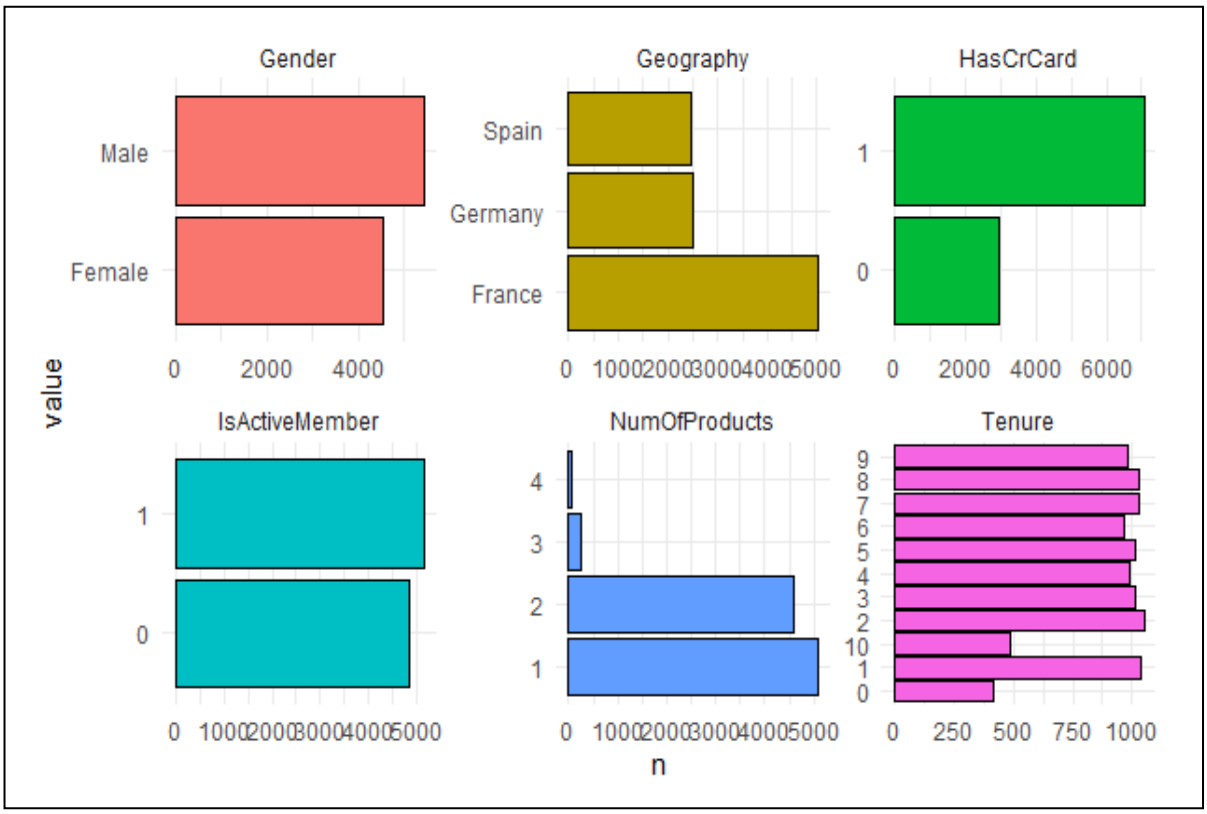


Graph no 4

```
library(tidyr)
library(tidyselect)
library(tidyverse)
data_2 %>%
  dplyr::select(-Exited) %>%
  keep(is.factor) %>%
  gather() %>%
  group_by(key, value) %>%
  summarize(n = n()) %>%
  ggplot() +
  geom_bar(mapping=aes(x = value, y = n, fill=key), color="black", stat='identity') +
  coord_flip() +
  facet_wrap(~ key, scales = "free") +
  theme_minimal() +
  theme(legend.position = 'none')
```

Information we can get from the plots:

- We have more male customers than females.
- Customers from France (most), Germany and France.
- Most of the customers have the bank's credit card
- We have an almost equal number of active and non-active members, not a very good sign
- Most of the customers use one or two kinds of products, with a very few use three or four products
- Almost equal number of customers in different tenure groups, except 0 and 10.



Graph no 5

```
library(ggcorrplot)
data_3 <- names(which(sapply(data_2, is.numeric)))
corr <- cor(data_2[,data_3], use = 'pairwise.complete.obs')
ggcorrplot(corr, lab = TRUE)
```

I don’t see any high correlation between the continuous variables

2. Develop answers to the following. Use dplyr wherever necessary

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

```
data_2 %>% select(CreditScore, Gender, Geography) %>% filter(Geography == "France") %>%
  dplyr::group_by(Gender) %>%
  dplyr::summarise(Gender_Average = mean(CreditScore))
```

Female :649

Male :650

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

```
data_2 %>% 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))
```

20-30: 651

31-40:651

41-50:649

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

```
data_2 %>% select(CreditScore, EstimatedSalary) %>% cor()
```

	CreditScore
CreditScore	1.000000000
EstimatedSalary	-0.001384293

	EstimatedSalary
CreditScore	-0.001384293
EstimatedSalary	1.000000000

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 = data_2)


# Show the model.
print(model)
summary(model)
```

Call:

lm(formula = CreditScore ~ Gender + Age + EstimatedSalary, data = data_2)

Coefficients:

(Intercept)	GenderMale
6.525e+02	-5.785e-01

Age	EstimatedSalary
-3.739e-02	-2.416e-06

Call:

lm(formula = CreditScore ~ Gender + Age + EstimatedSalary, data = data_2)

Residuals:

Min	1Q	Median	3Q	Max
-300.630	-66.880	1.262	66.930	201.174

Coefficients:

	Estimate	Std. Error
(Intercept)	6.525e+02	4.254e+00
GenderMale	-5.785e-01	1.942e+00

```
Age              -3.739e-02  9.221e-02
EstimatedSalary -2.416e-06  1.681e-05
                t value Pr(>|t|)
(Intercept)     153.382   <2e-16 ***
GenderMale       -0.298    0.766
Age              -0.405    0.685
EstimatedSalary  -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
```