Feisal Hasham 660473
IST3015 END SEM

Question 1

a)

Column names should not be empty

The data type that is stored in a data frame can be of numeric, factor or character type

Column names shouldn't be empty

Special type of list where every element of the list has same length

b)

We combine the data frames horizontally using rbind

df3 <-rbind(df1,df2)

df3

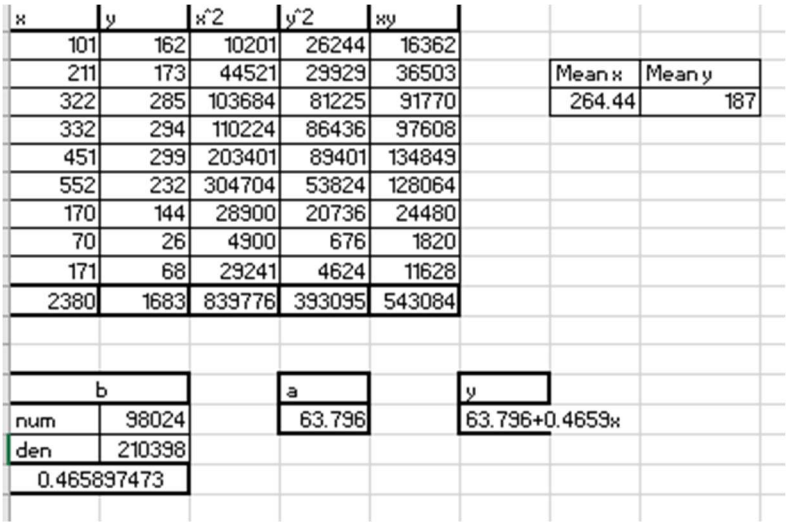
```
df1 <- data.frame (
  Class =c("Male","Female","Total")
)

df2 <- data.frame (
  Class =c("18","19","37")
)

df3 <-rbind(df1,df2)
df3
```

```
> df3
   Class
1   Male
2 Female
3  Total
4     18
5     19
6     37
> |
```

c)

i)

Feisal Hasham 660473
IST3015 END SEM



x	y	x^2	y^2	xy
101	162	10201	26244	16362
211	173	44521	29929	36503
322	285	103684	81225	91770
332	294	110224	86436	97608
451	299	203401	89401	134849
552	232	304704	53824	128064
170	144	28900	20736	24480
70	26	4900	676	1820
171	68	29241	4624	11628
2380	1683	839776	393095	543084

Mean x	Mean y
264.44	187

b	
num	98024
den	210398
0.465897473	

a
63.796

y
63.796+0.4659x

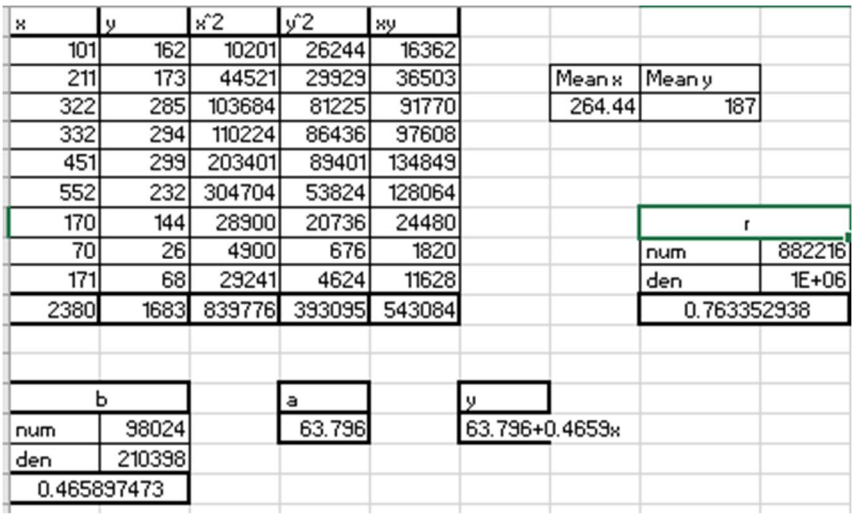
63.796+0.4659x

ii)

=63.796+(0.4659*397)

=248.76

iii)



x	y	x^2	y^2	xy
101	162	10201	26244	16362
211	173	44521	29929	36503
322	285	103684	81225	91770
332	294	110224	86436	97608
451	299	203401	89401	134849
552	232	304704	53824	128064
170	144	28900	20736	24480
70	26	4900	676	1820
171	68	29241	4624	11628
2380	1683	839776	393095	543084

Mean x	Mean y
264.44	187

r	
num	882216
den	1E+06
0.763352938	

b	
num	98024
den	210398
0.465897473	

a
63.796

y
63.796+0.4659x

0.763 being our r shows us that there's a strong positive linear relationship between variable x and y such that if x increases there's a very strong chance that y will increase too

Feisal Hasham 660473
IST3015 END SEM

Question 2

a)

Using Single Bracket [ ]

```
> my_list <- list(1,2,3,4,5,6,7,8,9, a = "Feisal")
> my_list[1]
[[1]]
[1] 1
```

Using double Bracket [[ ]]

```
> my_list <- list(1,2,3,4,5,6,7,8,9, a = "Feisal")
> my_list[["a"]]
[1] "Feisal"
>
```

Using the dollar sign $ operator

```
> my_list$a
[1] "Feisal"
>
```

b)

```
library(readxl)
library(dplyr)
library(tidyverse)
library(skimr)
BankChurners<-read_excel(file.choose())
View(BankChurners)
str(BankChurners)
glimpse(BankChurners)
summary(BankChurners)
dim(BankChurners)
skim(df)
```

```
> glimpse(BankChurners)
Rows: 10,127
Columns: 18
$ CLIENTNUM              <dbl> 768805383, 818770008, 713982108, 769911858, 709106358, 713061558, 810347208, 818906208, 710930508, 719661558, 7087908…
$ Attrition_Flag         <chr> "Existing Customer", "Existing Customer", "Existing Customer", "Existing Customer", "Existing Customer", "Existing Cu…
$ Customer_Age           <dbl> 45, 49, 51, 40, 40, 44, 51, 32, 37, 48, 42, 65, 56, 35, 57, 44, 48, 41, 61, 45, 47, 62, 41, 47, 54, 41, 59, 63, 44, 4…
$ Gender                 <chr> "M", "F", "M", "F", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "F", "M", "M", "M", "M", "F", "M", "F", "M", "F…
$ Dependent_count        <dbl> 3, 5, 3, 4, 3, 2, 4, 0, 3, 2, 5, 1, 1, 3, 2, 4, 4, 3, 1, 2, 1, 0, 3, 4, 2, 3, 1, 1, 3, 4, 3, 2, 4, 2, 0, 1, 3, 4, 1, …
$ Education_Level        <chr> "High School", "Graduate", "Graduate", "High School", "Uneducated", "Graduate", "Unknown", "High School", "Uneducated…
$ Marital_Status         <chr> "Married", "Single", "Married", "Unknown", "Married", "Married", "Married", "Unknown", "Single", "Single", "Unknown",…
$ Income_Category        <chr> "$60K - $80K", "Less than $40K", "$80K - $120K", "Less than $40K", "$60K - $80K", "$40K - $60K", "$120K +", "$60K - $…
$ Card_Category          <chr> "Blue", "Blue", "Blue", "Blue", "Blue", "Blue", "Gold", "Silver", "Blue", "Blue", "Blue", "Blue", "Blue", "Blue", "Bl…
$ Months_on_book         <dbl> 39, 44, 36, 34, 21, 36, 46, 27, 36, 36, 31, 54, 36, 30, 48, 37, 36, 34, 56, 37, 42, 49, 33, 36, 42, 28, 46, 56, 34, 4…
$ Total_Relationship_Count <dbl> 5, 6, 4, 3, 5, 3, 6, 2, 5, 6, 5, 6, 3, 5, 5, 5, 6, 4, 2, 6, 5, 2, 4, 3, 4, 6, 4, 3, 5, 6, 3, 2, 4, 5, 6, 4, 6, 2, 3, …
$ Months_Inactive_12_mon <dbl> 1, 1, 1, 4, 1, 1, 1, 2, 2, 3, 3, 2, 6, 1, 2, 1, 2, 4, 2, 1, 2, 3, 2, 3, 2, 1, 1, 3, 2, 0, 2, 5, 1, 2, 2, 2, 2, 3, 3, …
$ Contacts_Count_12_mon  <dbl> 3, 2, 0, 1, 0, 2, 3, 2, 0, 3, 2, 3, 0, 3, 2, 2, 3, 1, 3, 2, 0, 3, 1, 2, 3, 2, 2, 2, 2, 0, 3, 1, 2, 3, 2, 1, 3, 3, 2, …
$ Credit_Limit           <dbl> 12691.0, 8256.0, 3418.0, 3313.0, 4716.0, 4010.0, 34516.0, 29081.0, 22352.0, 11656.0, 6748.0, 9095.0, 11751.0, 8547.0,…
$ Total_Revolving_Bal    <dbl> 777, 864, 0, 2517, 0, 1247, 2264, 1396, 2517, 1677, 1467, 1587, 0, 1666, 680, 972, 2362, 1291, 2517, 1157, 1800, 0, 6…
$ Avg_Open_To_Buy        <dbl> 11914.0, 7392.0, 3418.0, 796.0, 4716.0, 2763.0, 32252.0, 27685.0, 19835.0, 9979.0, 5281.0, 7508.0, 11751.0, 6881.0, 1…
$ Total_Amt_Chng_Q4_Q1   <dbl> 1.335, 1.541, 2.594, 1.405, 2.175, 1.376, 1.975, 2.204, 3.355, 1.524, 0.831, 1.433, 3.397, 1.163, 1.190, 1.707, 1.708…
$ Total_Trans_Amt        <dbl> 1144, 1291, 1887, 1171, 816, 1088, 1330, 1538, 1350, 1441, 1201, 1314, 1539, 1311, 1570, 1348, 1671, 1028, 1336, 1207…
```

Feisal Hasham 660473
IST3015 END SEM

```
> summary(BankChurners)
   CLIENTNUM         Attrition_Flag      Customer_Age       Gender          Dependent_count Education_Level    Marital_Status     Income_Category
 Min.   :708082083   Length:10127       Min.   :26.00   Length:10127       Min.   :0.000   Length:10127       Length:10127       Length:10127
 1st Qu.:713036770   Class :character   1st Qu.:41.00   Class :character   1st Qu.:1.000   Class :character   Class :character   Class :character
 Median :717926358   Mode  :character   Median :46.00   Mode  :character   Median :2.000   Mode  :character   Mode  :character   Mode  :character
 Mean   :739177606                      Mean   :46.33                      Mean   :2.346
 3rd Qu.:773143533                      3rd Qu.:52.00                      3rd Qu.:3.000
 Max.   :828343083                      Max.   :73.00                      Max.   :5.000
 Card_Category       Months_on_book  Total_Relationship_Count Months_Inactive_12_mon Contacts_Count_12_mon  Credit_Limit   Total_Revolving_Bal
 Length:10127       Min.   :13.00   Min.   :1.000           Min.   :0.000          Min.   :0.000          Min.   : 1438   Min.   :   0
 Class :character   1st Qu.:31.00   1st Qu.:3.000           1st Qu.:2.000          1st Qu.:2.000          1st Qu.: 2555   1st Qu.: 359
 Mode  :character   Median :36.00   Median :4.000           Median :2.000          Median :2.000          Median : 4549   Median :1276
                    Mean   :35.93   Mean   :3.813           Mean   :2.341          Mean   :2.455          Mean   : 8632   Mean   :1163
                    3rd Qu.:40.00   3rd Qu.:5.000           3rd Qu.:3.000          3rd Qu.:3.000          3rd Qu.:11068   3rd Qu.:1784
                    Max.   :56.00   Max.   :6.000           Max.   :6.000          Max.   :6.000          Max.   :34516   Max.   :2517
 Avg_Open_To_Buy Total_Amt_Chng_Q4_Q1 Total_Trans_Amt
 Min.   :    3   Min.   :0.0000       Min.   :  510
 1st Qu.: 1324   1st Qu.:0.6310       1st Qu.: 2156
 Median : 3474   Median :0.7360       Median : 3899
 Mean   : 7469   Mean   :0.7599       Mean   : 4404
 3rd Qu.: 9859   3rd Qu.:0.8590       3rd Qu.: 4741
 Max.   :34516   Max.   :3.3970       Max.   :18484
```

```
> dim(BankChurners)
[1] 10127    18
```

```
> skim(BankChurners)
── Data Summary ────────────────────
                           Values
Name                       BankChurners
Number of rows             10127
Number of columns          18
_________________________
Column type frequency:
  character                6
  numeric                  12
_________________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────
  skim_variable   n_missing complete_rate min max empty n_unique whitespace
1 Attrition_Flag          0             1  17  17     0        2          0
2 Gender                  0             1   1   1     0        2          0
3 Education_Level         0             1   7  13     0        7          0
4 Marital_Status          0             1   6   8     0        4          0
5 Income_Category         0             1   7  14     0        6          0
6 Card_Category           0             1   4   8     0        4          0

── Variable type: numeric ────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────
   skim_variable           n_missing complete_rate          mean          sd          p0          p25          p50          p75         p100 hist
 1 CLIENTNUM                       0             1 739177606.    36903783.    708082083   713036770.   717926358    773143533    828343083 ▇▁▁▁▂
 2 Customer_Age                    0             1        46.3         8.02          26          41           46           52           73 ▁▃▇▃▁
 3 Dependent_count                 0             1         2.35        1.30           0           1            2            3            5 ▇▇▇▇▂
 4 Months_on_book                  0             1        35.9         7.99          13          31           36           40           56 ▁▂▇▂▁
 5 Total_Relationship_Count        0             1         3.81        1.55           1           3            4            5            6 ▇▇▇▇▇
 6 Months_Inactive_12_mon          0             1         2.34        1.01           0           2            2            3            6 ▇▇▇▁▁
 7 Contacts_Count_12_mon           0             1         2.46        1.11           0           2            2            3            6 ▃▇▇▂▁
 8 Credit_Limit                    0             1      8632.        9089.       1438.        2555         4549        11068.       34516 ▇▂▁▁▁
 9 Total_Revolving_Bal             0             1      1163.         815.          0         359          1276         1784         2517 ▅▃▃▇▆
10 Avg_Open_To_Buy                 0             1      7469.        9091.          3        1324.        3474         9859        34516 ▇▁▁▁▁
11 Total_Amt_Chng_Q4_Q1            0             1         0.760       0.219         0           0.631        0.736        0.859        3.40 ▂▇▁▁▁
12 Total_Trans_Amt                 0             1      4404.        3397.        510        2156.        3899         4741        18484 ▇▃▁▁▁
>
```

The mean customer age is 46.3 meaning on average most of the customers lie around that age

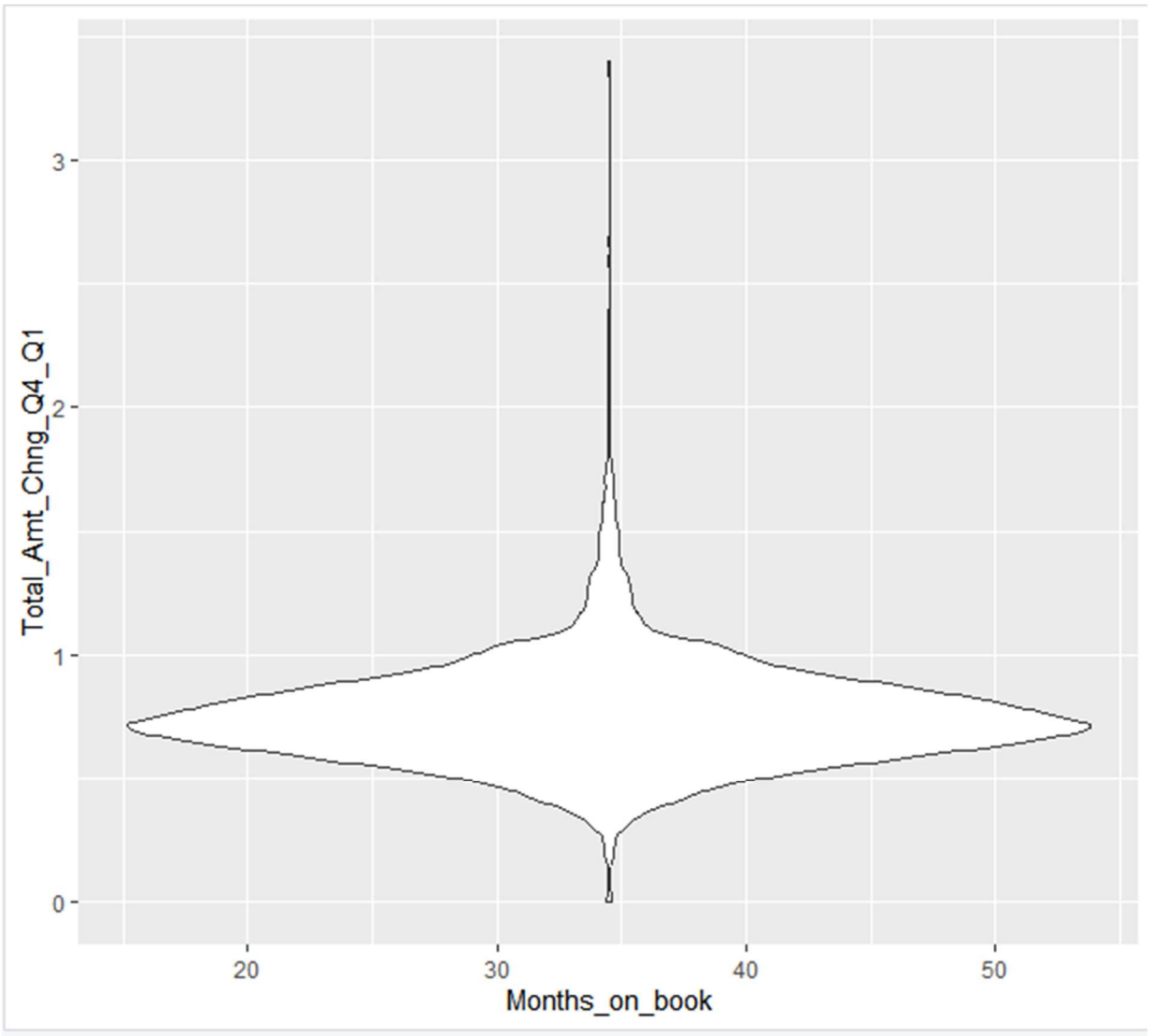
```
> sum(is.na(BankChurners))
[1] 0
```

There are no empty/null spaces meaning the bank has all the details that it needs

Feisal Hasham 660473
IST3015 END SEM

C)

```
install.packages('tidyverse')
install.packages('readxl')
library(tidyverse)
library(readxl)
BankChurners<-read_excel(file.choose())

ggplot(BankChurners, aes(x=Months_on_book, y=Total_Amt_Chng_Q4_Q1)) +
  geom_violin()
```



Feisal Hasham 660473
IST3015 END SEM

Question 3

a)

Define the process: Start by defining the process that you want to analyze.

Identify the data sources that contain information about the process.

Collect the relevant data from the identified sources and clean the data to remove any errors, inconsistencies, or duplicates.

Since we went through Tableau in class, we can use it to create visualizations such as graphs that provide insights into the process.

Use the insights from the analysis to identify areas where the process can be improved

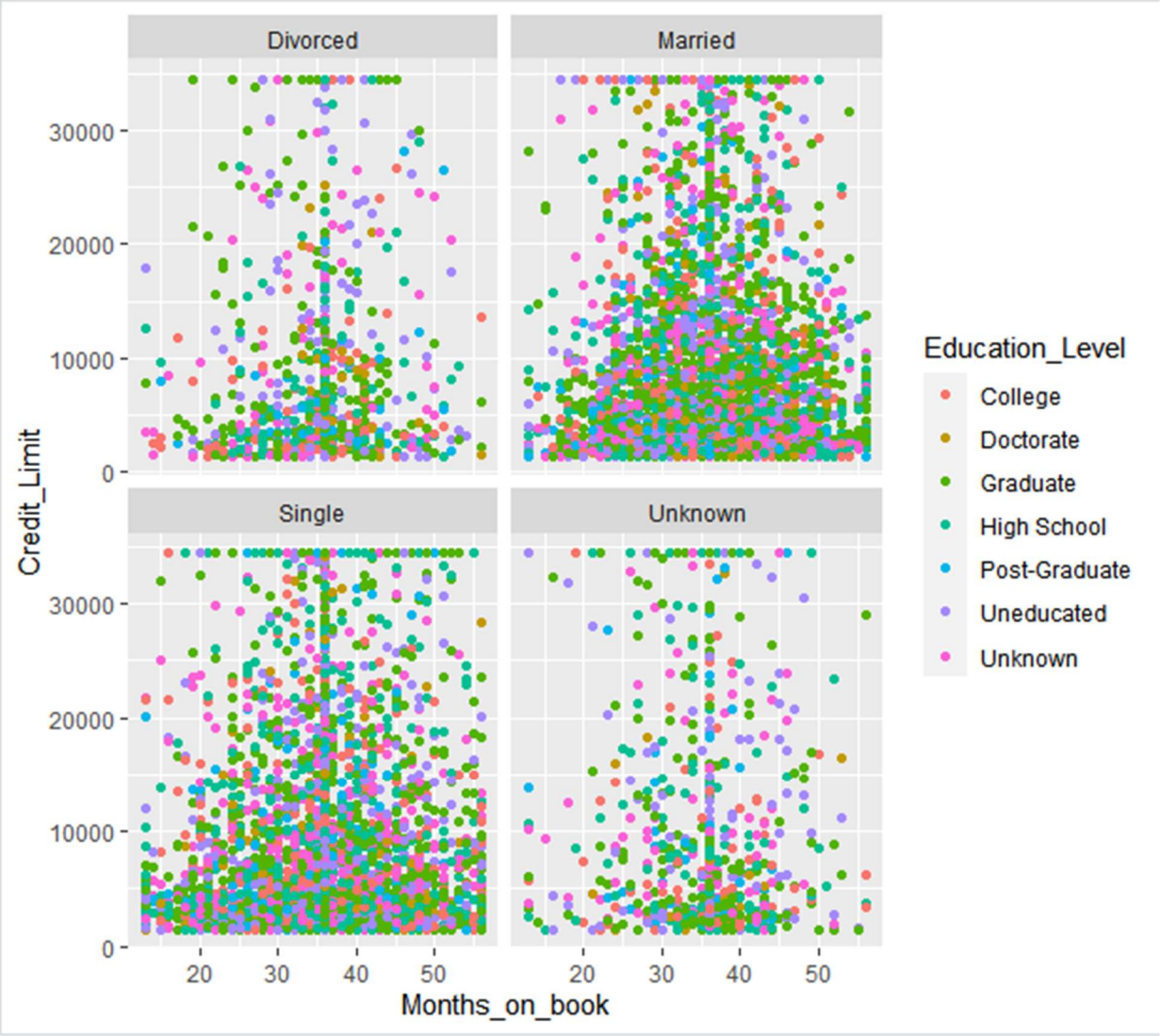
Implement the changes we have deemed necessary.

Monitor the process over time to ensure that the changes have resulted in the desired improvements and that the changes have not created any problems.

b)

```
install.packages('tidyverse')
install.packages('readxl')
library(tidyverse)
library(readxl)
BankChurners<-read_excel(file.choose())

ggplot(BankChurners, aes(x = Months_on_book, y = Credit_Limit, color = Education_Level)) +
  geom_point() +
  facet_wrap(~ Marital_Status)
```

Feisal Hasham 660473
IST3015 END SEM



c)

```
install.packages('tidyverse')
install.packages('readxl')
library(tidyverse)
library(readxl)
BankChurners<-read_excel(file.choose())
cont_table <- table(BankChurners$Credit_Limit, BankChurners$Total_Revolving_Bal)
chi_sq_test <- chisq.test(cont_table)
chi_sq_test
```

```
> chi_sq_test

        Pearson's Chi-squared test

data:  cont_table
X-squared = 12228088, df = 12240492, p-value = 0.9939
```

The p-value is 0.9939

Since the p-value is very high (close to 1), it suggests that the evidence against the null hypothesis is weak. There is not enough evidence to support the alternative hypothesis and we fail to reject the null hypothesis.

d)

```
install.packages('tidyverse')
install.packages('readxl')
library(tidyverse)
library(readxl)
BankChurners<-read_excel(file.choose())
ggplot(BankChurners, aes(x = "", y = Total_Trans_Amt, fill = Card_Category)) +
  geom_bar(stat = "identity") +
  labs(title = "Total AMount") +
  theme(legend.position = "right")
```