Project Proposal

Student: Simran Nirajkumar Shah
ID: 989439437

# R File:

https://drive.google.com/file/d/1DszEnqSBaMLSUCqvS8MZTO-6covhdW2q/view?usp=sharing

# R output File:

https://docs.google.com/document/d/1mDmcrvTq0bKMhAECI01UQdNg_TwuPXjR/edit?usp=sharing&ouid=104785859357775399979&rtpof=true&sd=true

# R RMD File:

https://drive.google.com/file/d/1OrcbhwWirFVKJoFmdshNk260QVwq0iM7/view?usp=sharing

# R HTML file:

https://drive.google.com/file/d/1_CBw6iBE66rTA2LGICMKPKQCfiqPzJRf/view?usp=sharing

# Dataset used by adding new columns:

https://drive.google.com/file/d/1XTRHK-KhBEJ6rOFQ3VWTbyQBlCqoffzU/view?usp=sharing

# R Code:

```
library(tidyverse)
library(corrplot)
library(dplyr)
customers <- read.csv("Customers.csv")
view(customers)
str(customers)
summary(customers)
```

```
#----------------------------Scatter Plot---------------------------------------------
ggplot(Customers,
      aes(FamilySize,
         SpendingScore,
         color = Gender)) +
  geom_point(fill ="turquoise2", size = 2)
ggsave('P1.png')

ggplot(Customers,
      aes(Age, WorkExperience,
         color = Profession)) +
  geom_point(fill= "palegreen1")
ggsave('p2.png')

ggplot(Customers,
      aes(SpendingScore, AnnualIncome,
         color = Profession)) +
  geom_point(fill= "skyblue1")
ggsave('p3.png')

#--------------------------Box Plot-----------------------------------------------
ggplot(Customers,
      aes(AnnualIncome,
         Profession)) +
  geom_boxplot(fill= "thistle2")
ggsave('p4.png')

ggplot(Customers,
      aes(SpendingScore,
         Profession,
          color = Profession)) +
  geom_boxplot(fill ="turquoise2")
ggsave('p5.png')

#--------------------Bar Chart----------------------------

ggplot(Customers,
      aes(SpendingScore,
          color = Gender)) +
  geom_bar(fill ="plum")
ggsave('p6.png')

#-----------------Histogram--------------------------------
ggplot(Customers,
      aes(SpendingScore,
          color = Profession)) +
  geom_histogram(fill="snow")
ggsave('p7.png')

# ------------------- Co-relation -----------------------------------
customers <- read.csv("Customers.csv")
cor.test(customers$Age, customers$FamilySize, method = "spearman")
cor(select(customers, Age, FamilySize))
cus<- pairs(select(customers, Age, FamilySize,SpendingScore))
summary(cus)

customers <- read.csv("Customers.csv")
#customers <- filter(customers, customers$Profession_numeric == 8)
cor.test(customers$WorkExperience, customers$FamilySize, method =
"spearman")
```

```
cor(select(customers, WorkExperience, FamilySize))
pairs(select(customers, WorkExperience, FamilySize,SpendingScore))

# ---------------- Linear regression ------------------------------
customers <- read.csv("Customers.csv")
linear_customer <- lm(customers$FamilySize ~
customers$AnnualIncome, data = customers)
summary(linear_customer)
ggplot(customers, aes(FamilySize, AnnualIncome)) +
  geom_point() +
  geom_smooth(method=lm)
ggsave("Linear_regression1.png")

#customers <- filter(customers, customers$Profession_numeric == 8 &
customers$Gender_numeric == 1)
linear_customer <- lm(customers$AnnualIncome ~
customers$WorkExperience, data = customers)
summary(linear_customer)
ggplot(customers, aes(AnnualIncome, WorkExperience)) +
  geom_point() +
  geom_smooth(method=lm)
ggsave("Linear_regression2.png")

#-------------------- T-test-----------------------------------------
customers <- read.csv("customers.csv")
males <- subset(customers, Gender == "Male")
females <- subset(customers, Gender == "Female")
t.test(males$SpendingScore, females$SpendingScore, var.equal =
FALSE)
Age_A <- subset(customers, customers$Age > 50)
```

Age_B <- subset(customers, customers$Age <= 50)
t.test(Age_A$SpendingScore, Age_B$SpendingScore, var.equal = FALSE)

# R output:

```
> library(tidyverse)
── Attaching packages ──────────────────────────────────────── tidyverse 1.
3.2 ──
✔ ggplot2 3.4.1     ✔ purrr   1.0.1
✔ tibble  3.1.8     ✔ dplyr   1.1.0
✔ tidyr   1.3.0     ✔ stringr 1.5.0
✔ readr   2.1.3     ✔ forcats 1.0.0
── Conflicts ───────────────────────────────────────── tidyverse_conflict
s() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
> library(corrplot)
corrplot 0.92 loaded
> library(dplyr)
> 
> customers <- read.csv("Customers.csv")
> view(customers)
> str(customers)
'data.frame':	2000 obs. of  10 variables:
 $ CustomerID        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Gender            : chr  "Male" "Male" "Female" "Female" ...
 $ Age               : int  19 21 20 23 31 22 35 23 64 30 ...
 $ Age_numeric       : chr  "B-value" "B-value" "B-value" "B-value" ...
 $ AnnualIncome      : int  15000 35000 86000 59000 38000 58000 31000 84000
97000 98000 ...
 $ SpendingScore     : int  39 81 6 77 40 76 6 94 3 72 ...
 $ Profession        : chr  "Healthcare" "Engineer" "Engineer" "Lawyer" ...
 $ Profession_numeric: int  1 2 2 3 4 5 1 1 2 5 ...
 $ WorkExperience    : int  1 3 1 0 2 0 1 1 0 1 ...
 $ FamilySize        : int  4 3 1 2 6 2 3 3 3 4 ...
> summary(customers)
   CustomerID        Gender               Age         Age_numeric
 Min.   :   1.0   Length:2000        Min.   : 0.00   Length:2000
 1st Qu.: 500.8   Class :character   1st Qu.:25.00   Class :character
 Median :1000.5   Mode  :character   Median :48.00   Mode  :character
 Mean   :1000.5                      Mean   :48.96
 3rd Qu.:1500.2                      3rd Qu.:73.00
 Max.   :2000.0                      Max.   :99.00
  AnnualIncome     SpendingScore     Profession         Profession_numeric
 Min.   :     0   Min.   :  0.00   Length:2000        Min.   : 1.00
 1st Qu.: 74572   1st Qu.: 28.00   Class :character   1st Qu.: 2.00
 Median :110045   Median : 50.00   Mode  :character   Median : 5.00
 Mean   :110732   Mean   : 50.96                      Mean   : 4.43
 3rd Qu.:149093   3rd Qu.: 75.00                      3rd Qu.: 5.00
 Max.   :189974   Max.   :100.00                      Max.   :10.00
 WorkExperience     FamilySize
 Min.   : 0.000   Min.   :1.000
 1st Qu.: 1.000   1st Qu.:2.000
 Median : 3.000   Median :4.000
 Mean   : 4.103   Mean   :3.768
 3rd Qu.: 7.000   3rd Qu.:5.000  
 Max.   :17.000   Max.   :9.000  
> 
> #----------------------------Scatter Plot---------------------------------
--------
> ggplot(Customers,
+        aes(FamilySize,
+            SpendingScore,
+            color = Gender)) +
+   geom_point(fill ="turquoise2", size = 2)
> ggsave('P1.png')
Saving 6.52 x 5.54 in image
> 
> ggplot(Customers,
+        aes(Age, WorkExperience,
+            color = Profession)) +
+   geom_point(fill= "palegreen1")
> ggsave('p2.png')
Saving 6.52 x 5.54 in image
> 
> ggplot(Customers,
+        aes(SpendingScore, AnnualIncome,
+            color = Profession)) +
+   geom_point(fill= "skyblue1")
> ggsave('p3.png')
Saving 6.52 x 5.54 in image
> 
> #--------------------------Box Plot---------------------------------------
----
> ggplot(Customers,
+        aes(AnnualIncome,
+            Profession)) +
+   geom_boxplot(fill= "thistle2")
> ggsave('p4.png')
Saving 6.52 x 5.54 in image
> 
> ggplot(Customers,
+        aes(SpendingScore,
+            Profession,
+            color = Profession)) +
+   geom_boxplot(fill ="turquoise2")
> ggsave('p5.png')
Saving 6.52 x 5.54 in image
> 
> #---------------------Bar Chart------------------------------
> 
> ggplot(Customers,
+        aes(SpendingScore,
+            color = Gender)) +
+   geom_bar(fill ="plum")
> ggsave('p6.png')
Saving 6.52 x 5.54 in image
> 
> #------------------Histogram----------------------------------
> ggplot(Customers,
+        aes(SpendingScore,
+            color = Profession)) +
+   geom_histogram(fill="snow")
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
> ggsave('p7.png')
Saving 6.52 x 5.54 in image
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
> 
> # ------------------- Co-relation ------------------------------------
> customers <- read.csv("Customers.csv")
> cor.test(customers$Age, customers$FamilySize, method = "spearman")

	Spearman's rank correlation rho

data:  customers$Age and customers$FamilySize
S = 1280876395, p-value = 0.07857
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho 
0.03934246 

Warning message:
In cor.test.default(customers$Age, customers$FamilySize, method = "spearman"
) :
  Cannot compute exact p-value with ties
> cor(select(customers, Age, FamilySize))
                  Age FamilySize
Age        1.00000000 0.03825438
FamilySize 0.03825438 1.00000000
> cus<- pairs(select(customers, Age, FamilySize,SpendingScore))
> summary(cus)
Length  Class   Mode 
     0   NULL   NULL 
> 
> customers <- read.csv("Customers.csv")
> #customers <- filter(customers, customers$Profession_numeric == 8)
> cor.test(customers$WorkExperience, customers$FamilySize, method = "spearma
n")

	Spearman's rank correlation rho

data:  customers$WorkExperience and customers$FamilySize
S = 1316329885, p-value = 0.5687
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho 
0.01275234 

Warning message:
In cor.test.default(customers$WorkExperience, customers$FamilySize,  :
  Cannot compute exact p-value with ties
> cor(select(customers, WorkExperience, FamilySize))
               WorkExperience FamilySize
WorkExperience     1.00000000 0.01187302
FamilySize         0.01187302 1.00000000
> pairs(select(customers, WorkExperience, FamilySize,SpendingScore))
> 
> # ----------------- Linear regression --------------------------------
> customers <- read.csv("Customers.csv")
> linear_customer <- lm(customers$FamilySize ~ customers$AnnualIncome, data
= customers)
> summary(linear_customer)

Call:
lm(formula = customers$FamilySize ~ customers$AnnualIncome, data = customers
)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0860 -1.6894 -0.0106  1.4621  5.5550 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)            3.325e+00  1.150e-01  28.916  < 2e-16 ***
customers$AnnualIncome 4.007e-06  9.597e-07   4.175  3.1e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.963 on 1998 degrees of freedom
Multiple R-squared:  0.00865,  Adjusted R-squared:  0.008154 
F-statistic: 17.43 on 1 and 1998 DF,  p-value: 3.103e-05

> ggplot(customers, aes(FamilySize, AnnualIncome)) +
+   geom_point() +
+   geom_smooth(method=lm)
`geom_smooth()` using formula = 'y ~ x'
> ggsave("Linear_regression1.png")
Saving 6.52 x 5.54 in image
`geom_smooth()` using formula = 'y ~ x'
> 
> #customers <- filter(customers, customers$Profession_numeric == 8 & custom
ers$Gender_numeric == 1)
> linear_customer <- lm(customers$AnnualIncome ~ customers$WorkExperience, d
ata = customers)
> summary(linear_customer)

Call:
lm(formula = customers$AnnualIncome ~ customers$WorkExperience, 
    data = customers)

Residuals:
    Min      1Q  Median      3Q     Max 
-108665  -36320   -1933   38272   83478 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)              106467.4     1474.7    72.2  < 2e-16 ***
customers$WorkExperience   1039.5      259.9     4.0 6.56e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45570 on 1998 degrees of freedom
Multiple R-squared:  0.007945,	Adjusted R-squared:  0.007449 
F-statistic:    16 on 1 and 1998 DF,  p-value: 6.559e-05

> ggplot(customers, aes(AnnualIncome, WorkExperience)) +
+   geom_point() +
+   geom_smooth(method=lm)
`geom_smooth()` using formula = 'y ~ x'
> ggsave("Linear_regression2.png")
Saving 6.52 x 5.54 in image
`geom_smooth()` using formula = 'y ~ x'
> 
> #--------------------- T-test------------------------------------------
> customers <- read.csv("customers.csv")
> males <- subset(customers, Gender == "Male")
> females <- subset(customers, Gender == "Female")
> t.test(males$SpendingScore, females$SpendingScore, var.equal = FALSE)

	Welch Two Sample t-test

data:  males$SpendingScore and females$SpendingScore
t = -0.023614, df = 1756.4, p-value = 0.9812
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.520631  2.460656
sample estimates:
```

```
mean of x mean of y 
 50.94472  50.97470 

> 
> 
> Age_A <- subset(customers, customers$Age > 50)
> Age_B <- subset(customers, customers$Age <= 50)
> t.test(Age_A$SpendingScore, Age_B$SpendingScore, var.equal = FALSE)

	Welch Two Sample t-test

data:  Age_A$SpendingScore and Age_B$SpendingScore
t = -1.2209, df = 1990.7, p-value = 0.2223
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.9743111  0.9246314
sample estimates:
mean of x mean of y 
 50.16806  51.69290 
```

# Original Dataset Link

Dataset: https://www.kaggle.com/datasets/datascientistanna/customers-dataset

Dataset features:

- 2000 rows
- 10 variables
- Quantitative variables: Annual income, Age
- Categorical variables: Gender, Profession

#Dataset Observation Representation

Each observation in the dataset represents the customer’s characteristics for membership in the shop. Here the analysis is done for the most common characters customer contains for membership.

#Dataset Variable Representation

Variables characteristics:

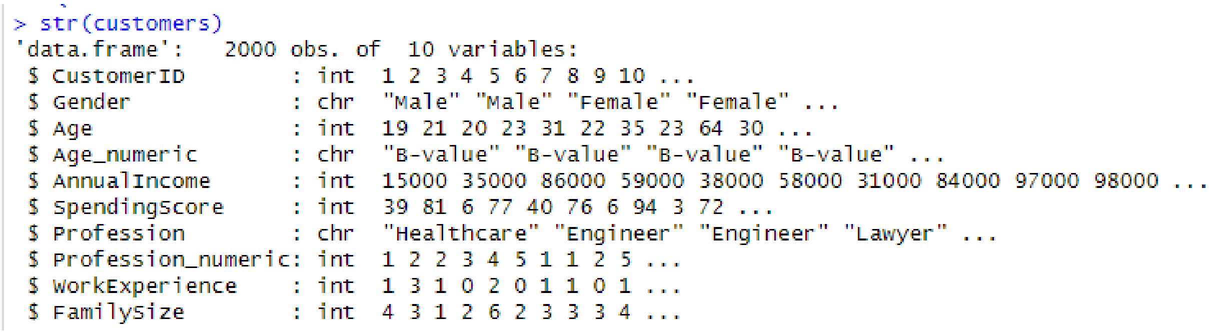
1. Customer ID: ID given to the customers.
2. Gender: The gender of the customer
3. Age: The age of the customer
4. Age_numeric: Age classification based on >50 and <=50 **(Newly Added)**
5. Annual Income: Income of the customer on an annual basis

6. Spending Score: scores of the customer given by the shop to the customer on certain points.
7. Profession: the profession of the customer.
8. Profession_numeric: Profession divided into numeric numbers. **(Newly Added)**
9. Work Experience: work experience of the customer (in years).
10. Family Size: The family size of the customers.

#Dataset Motive

The Dataset contains all the types of customer characteristics. These characteristics help data analysts better describe the relationship between getting products from the shop. Better visualization of the characteristics of the customer will help in more marketing in that area. This marketing is then helpful in suggesting people on different social media holding such features. Customers will be more likely to come to shop and get products.

#Number of columns and rows in the dataset from str() Function

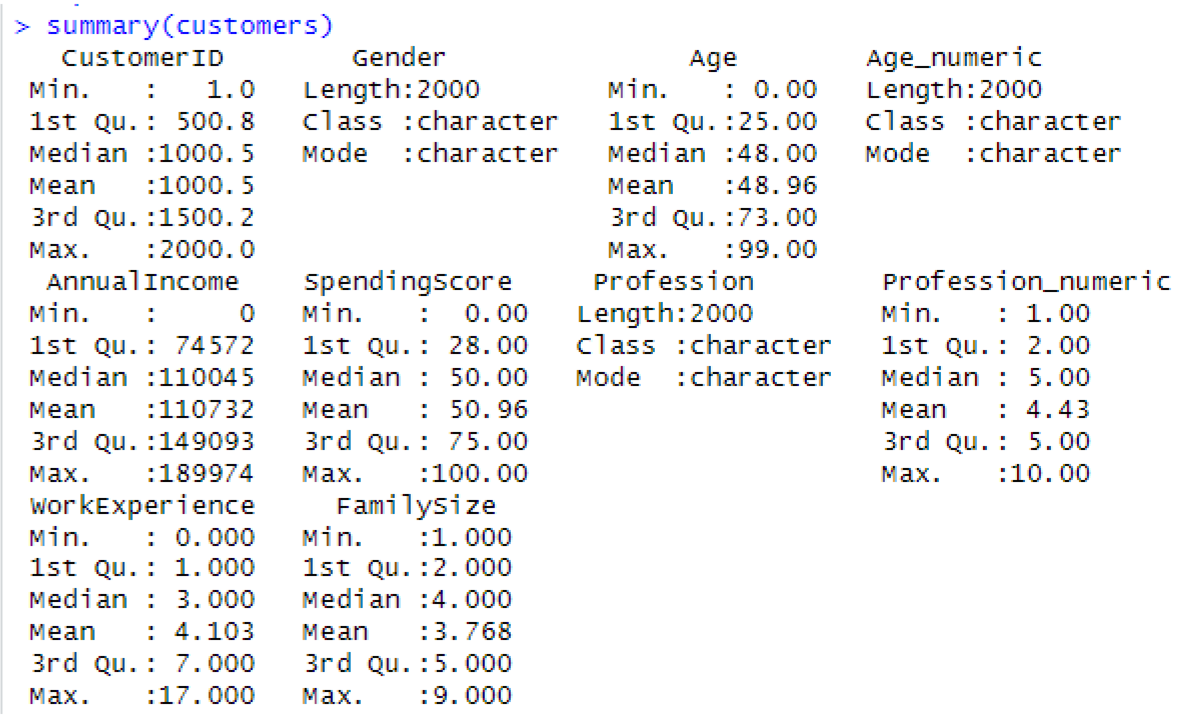


```
> str(customers)
'data.frame':	2000 obs. of  10 variables:
 $ CustomerID        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Gender            : chr  "Male" "Male" "Female" "Female" ...
 $ Age               : int  19 21 20 23 31 22 35 23 64 30 ...
 $ Age_numeric       : chr  "B-value" "B-value" "B-value" "B-value" ...
 $ AnnualIncome      : int  15000 35000 86000 59000 38000 58000 31000 84000 97000 98000 ...
 $ SpendingScore     : int  39 81 6 77 40 76 6 94 3 72 ...
 $ Profession        : chr  "Healthcare" "Engineer" "Engineer" "Lawyer" ...
 $ Profession_numeric: int  1 2 2 3 4 5 1 1 2 5 ...
 $ WorkExperience    : int  1 3 1 0 2 0 1 1 0 1 ...
 $ FamilySize        : int  4 3 1 2 6 2 3 3 3 4 ...
```

Observations: 2000

Variable: 10

# Summary of the Dataset using Summary() Function



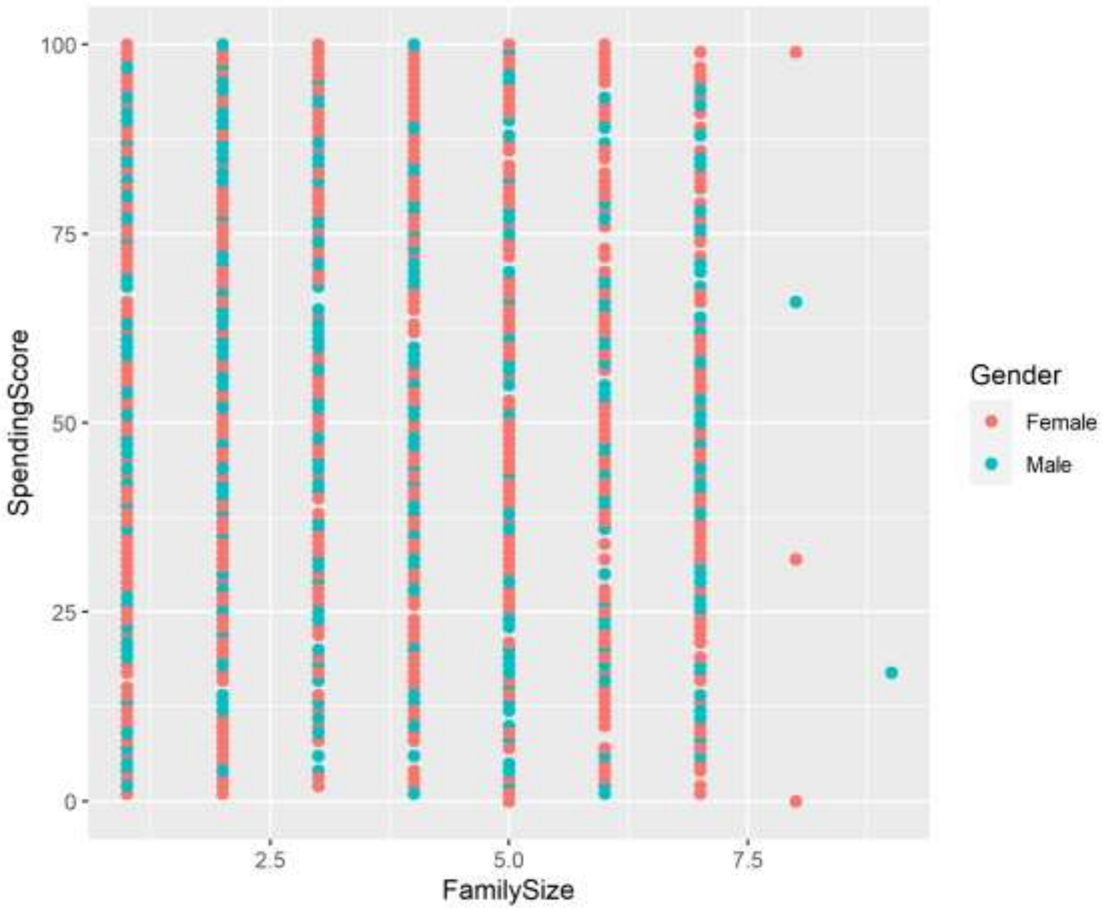
```
> summary(customers)
   CustomerID       Gender               Age         Age_numeric       
 Min.   :   1.0   Length:2000        Min.   : 0.00   Length:2000       
 1st Qu.: 500.8   Class :character   1st Qu.:25.00   Class :character  
 Median :1000.5   Mode  :character   Median :48.00   Mode  :character  
 Mean   :1000.5                      Mean   :48.96                     
 3rd Qu.:1500.2                      3rd Qu.:73.00                     
 Max.   :2000.0                      Max.   :99.00                     
  AnnualIncome    SpendingScore     Profession        Profession_numeric
 Min.   :     0   Min.   :  0.00   Length:2000        Min.   : 1.00     
 1st Qu.: 74572   1st Qu.: 28.00   Class :character   1st Qu.: 2.00     
 Median :110045   Median : 50.00   Mode  :character   Median : 5.00     
 Mean   :110732   Mean   : 50.96                      Mean   : 4.43     
 3rd Qu.:149093   3rd Qu.: 75.00                      3rd Qu.: 5.00     
 Max.   :189974   Max.   :100.00                      Max.   :10.00     
 WorkExperience     FamilySize   
 Min.   : 0.000   Min.   :1.000  
 1st Qu.: 1.000   1st Qu.:2.000  
 Median : 3.000   Median :4.000  
 Mean   : 4.103   Mean   :3.768  
 3rd Qu.: 7.000   3rd Qu.:5.000  
 Max.   :17.000   Max.   :9.000  
```

Exploratory Data Analysis

The imaginary Shop dataset contains Several variables. These variables are counted in increasing the number of customers. Using exploratory Data analysis, some insights are been taken out. This insight shows customers’ characteristics and their way of buying the products.

#Scatter Plot

1st Scatterplot graph shows the relationship between family Size Vs spending score on basis of Gender.



Based on the scatter plot, we can see that there is a relationship between Family Size and Spending Score, with a moderate positive correlation between the two variables.

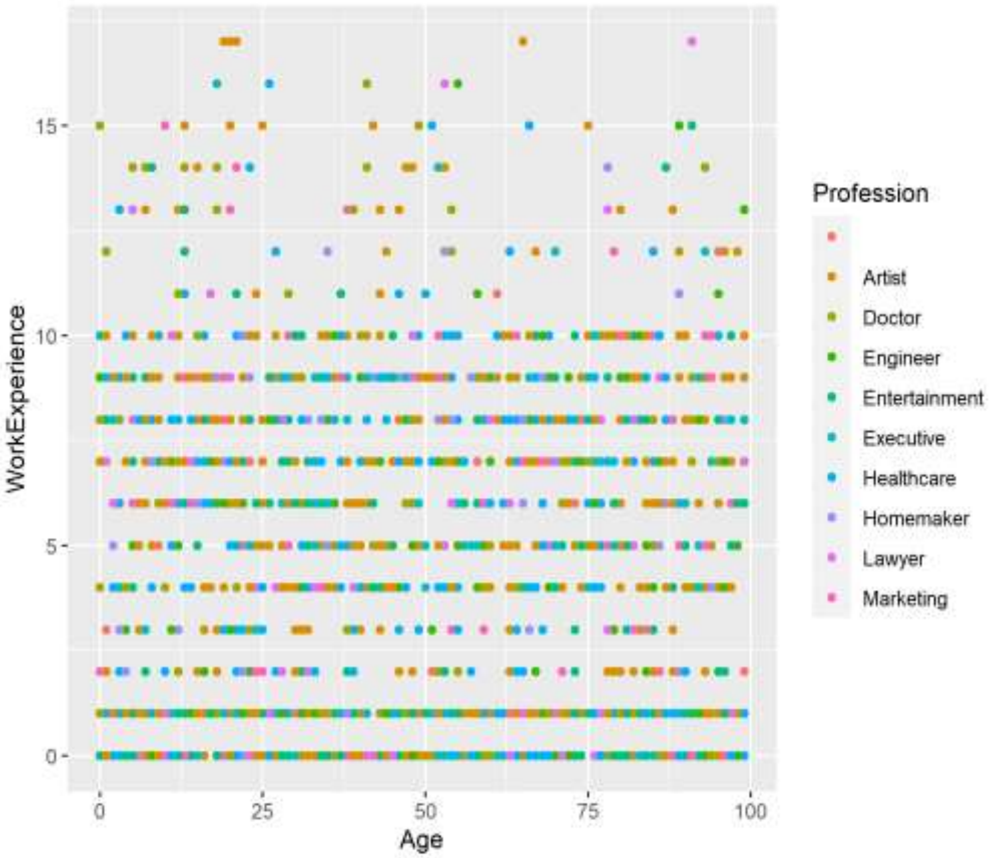
We can also see that the data is split by gender, with male observations represented by blue points and female observations represented by orange points. Looking at the clusters of points, we can see that female observations tend to have a higher Spending Score than male observations for the same Family Size. This means that, on average, females tend to spend more than males with the same family size.

Additionally, we can see that there are several clusters of observations for each gender, indicating that there may be distinct subgroups within the data. For example, for both males and females, there is a cluster of observations with low Family Size and low Spending Scores, indicating that there are many individuals or families with low spending habits.

Similarly, there are clusters of observations with high Family Sizes and high Spending Scores, indicating that there are some individuals or families with high spending habits.

Overall, this scatters plot suggests that Family Size and Spending Score are related, and that gender may also play a role in spending habits.

2nd Scatterplot graph shows the relationship between Age Vs work experience on basis of profession.

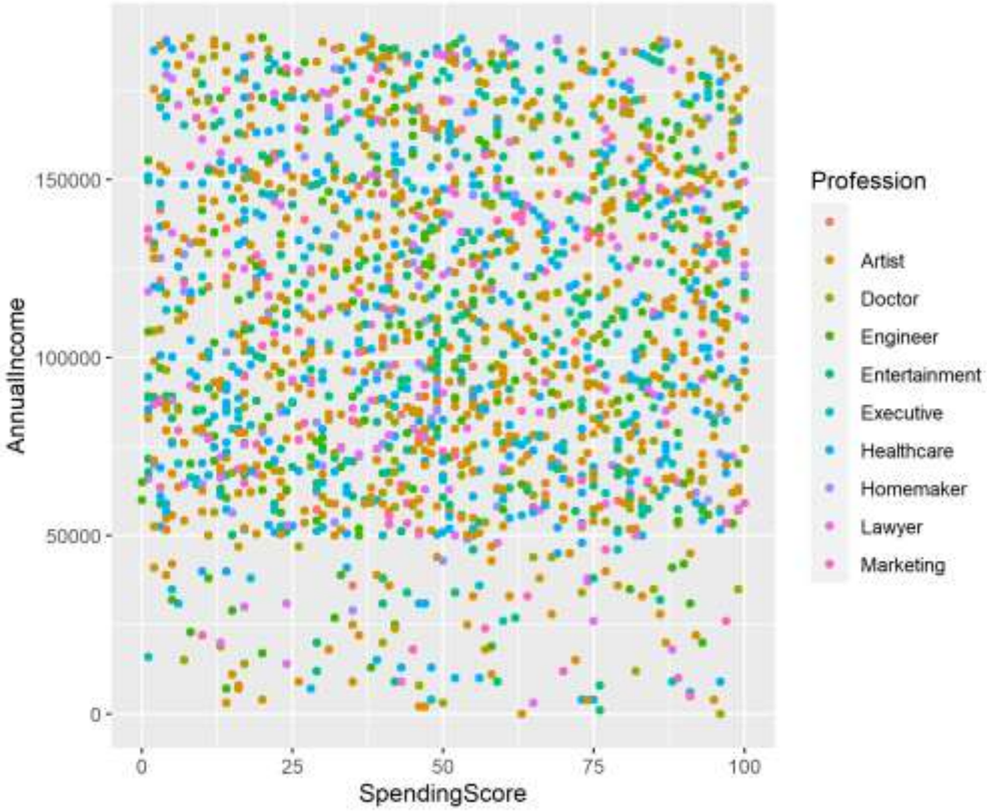


Based on the scatter plot, we can see that there is a positive correlation between age and work experience, which means that as age increases, work experience tends to increase as well. This relationship holds true for all professions represented in the graph. However, we can also see that there is a significant amount of variation in the data, indicating that there are individuals in each profession with varying levels of work experience at any given age.

We can also see that the data is separated by profession, with different colors representing different professions. Looking at the clusters of points, we can see that certain professions tend to have higher work experience levels for the same age compared to others. For example, the purple points representing engineers tend to have higher work experience levels for the same age compared to the blue points representing healthcare professionals.

Overall, this scatters plot provides insight into the relationship between age and work experience for different professions.

3rd Scatterplot graph shows the relationship between spending score Vs Annual Income on basis of profession.



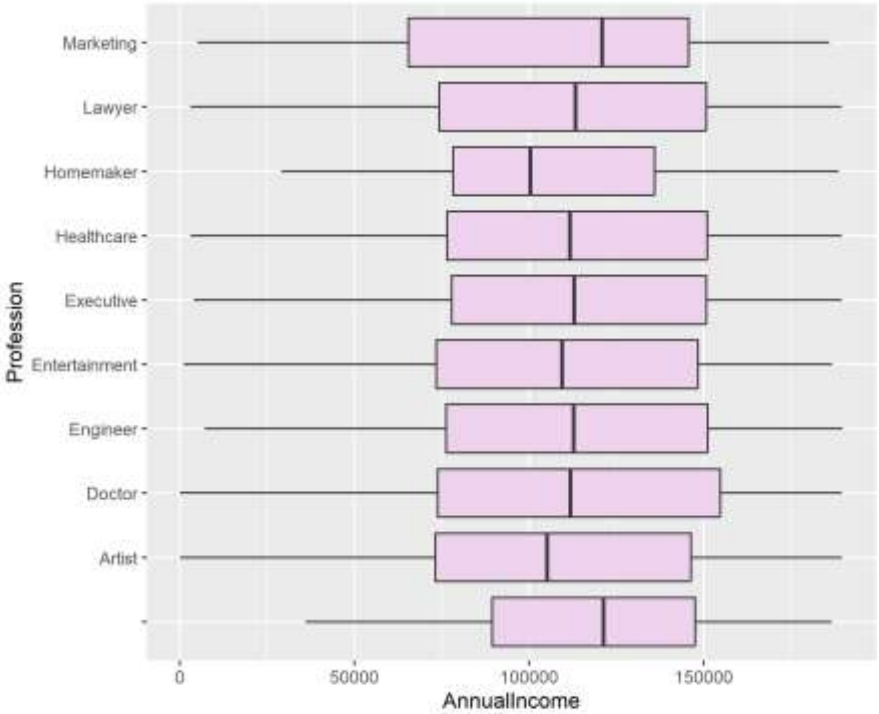
We can see that there is no clear relationship between annual income and spending score for any profession. Instead, the data appears to be clustered into distinct groups, indicating that there are different types of customers with varying spending habits and income levels.

Looking at the data, we can see that there are three main clusters of data points: one at the top of the graph, one in the middle, and one at the bottom. The top cluster represents customers with high annual incomes and high spending scores, indicating that they have the financial means to spend more. The middle cluster represents customers with moderate annual incomes and moderate spending scores, while the bottom cluster represents customers with lower annual incomes and lower spending scores.

We can also see that different professions are represented in each cluster, with some professions being overrepresented in certain clusters. For example, the blue points representing healthcare professionals are primarily clustered in the middle group, while the purple points representing engineers are primarily clustered in the top group.

#Box Plot

1st Boxplot graph shows the relationship between Annual Income Vs Profession.

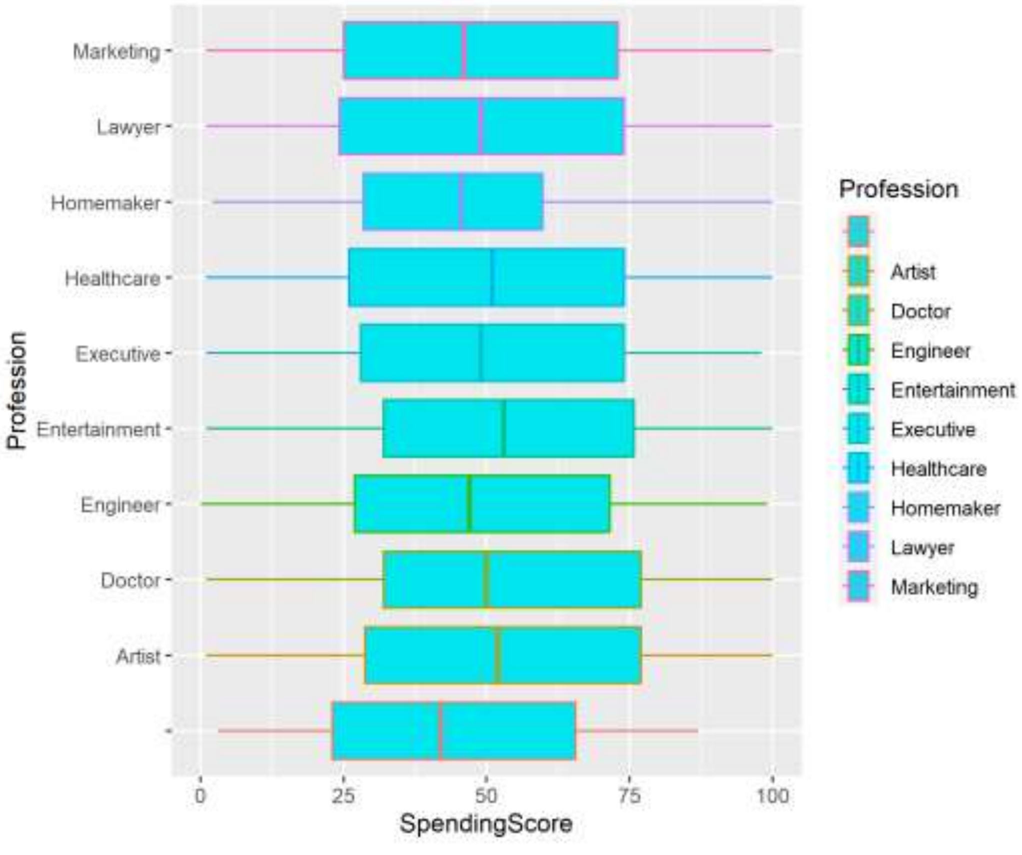


From the plot, we can see that the median annual income for doctors, engineers, executives, and lawyers is higher than other professions. The median annual income for artists, entertainment, healthcare, homemakers, and marketing is relatively lower. The no profession category has the lowest median annual income.

We can also see that the range of annual income is wider for executive, lawyer, and doctor professions, indicating that there is more variability in their annual income. The range of annual income is narrower for artists, entertainment, healthcare, and homemakers, indicating less variability in their annual income.

Overall, this plot gives us a good idea of the distribution of annual income for different professions, which can help us identify trends and make informed decisions.

2nd Boxplot graph shows the relationship between spending Score Vs Profession.

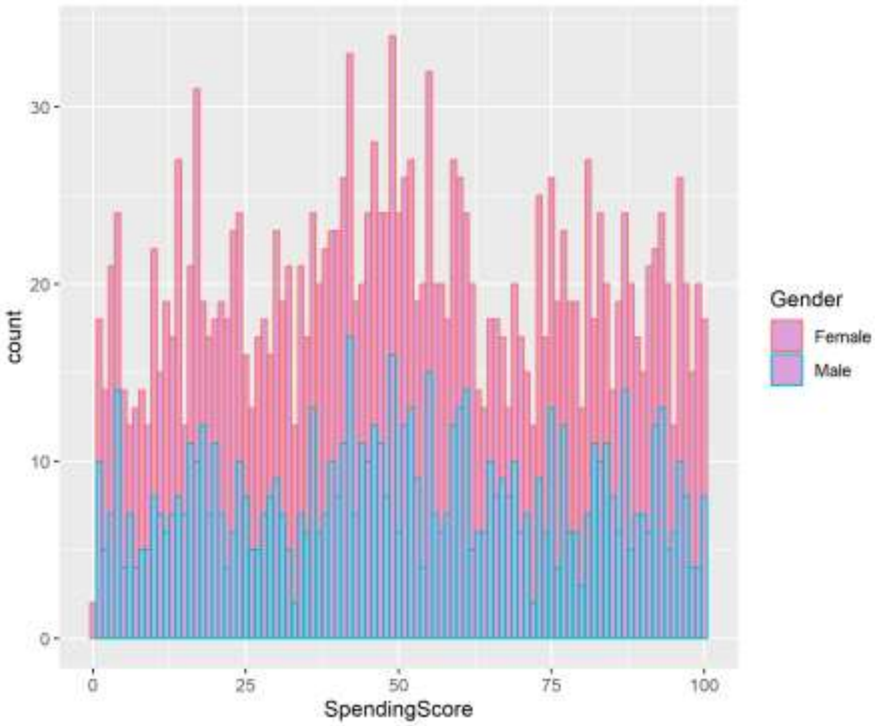


Based on the plot, it appears that the median spending score is higher for the Entertainment and Executive professions compared to the other professions. The Healthcare and Homemaker professions have the smallest interquartile range, indicating that the spending scores in those professions are relatively similar.

There are also some outliers in the plot, particularly for the Entertainment, Marketing, and No profession categories, which indicates that there are some individuals in those professions who spend significantly more than the rest of the group.

#Bar chart

Here Bar chart Shows the relationship between Spending Score and Gender.

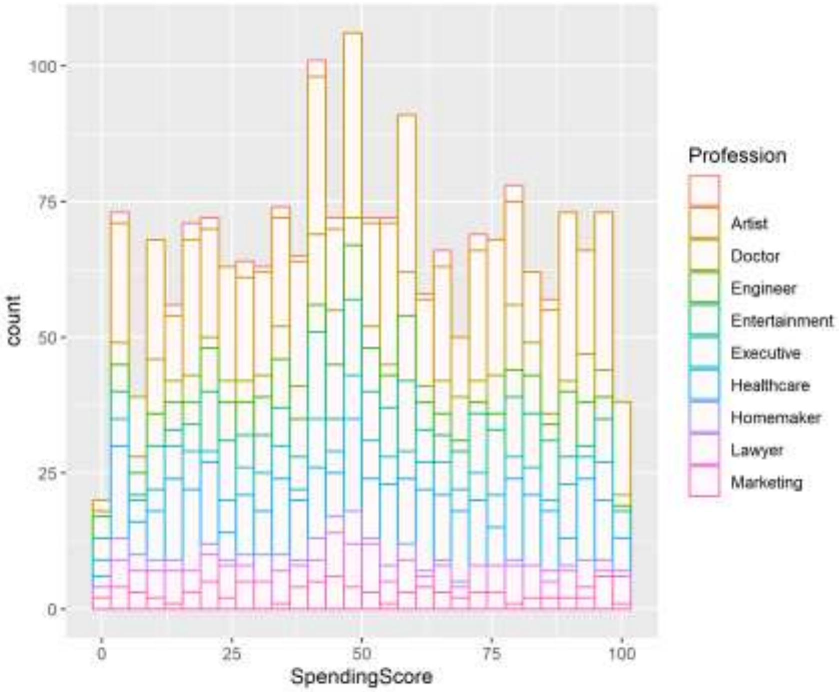


From the graph, we can observe that there are more female customers than male customers for every spending score level. The highest count of female customers is observed for the spending score level of 50, while the highest count of male customers is observed for the spending score level of 60.

Overall, the graph suggests that female customers tend to spend more than male customers, as there are more female customers across all spending score levels.

#Histogram

Here Histogram shows the relationship between spending Score and Profession.



From the histogram, we can observe the following:

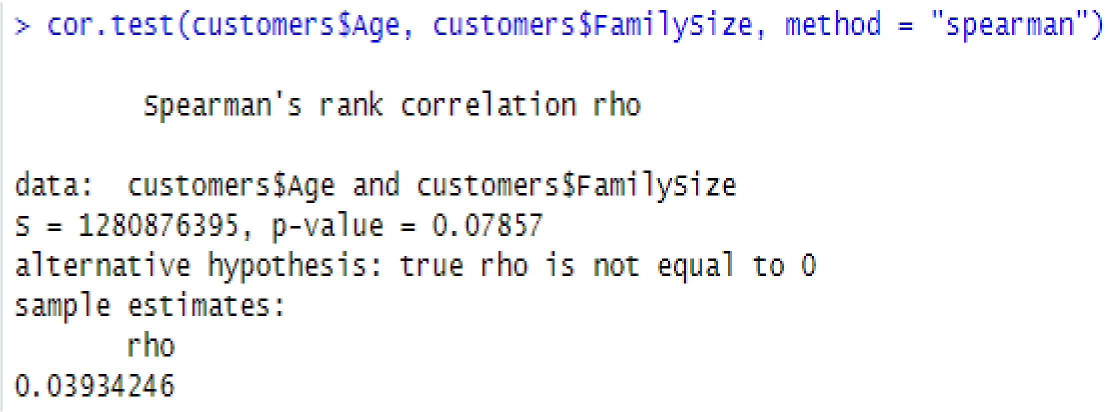
- The majority of people have a spending score in the range of 40-60.
- The highest count of people with a spending score in the range of 50-60 belongs to the healthcare and marketing professions.
- The highest count of people with a spending score in the range of 60-70 belongs to the artist and lawyer professions.
- The highest count of people with a spending score in the range of 70-80 belongs to the doctor and entertainment professions.

- The executive, homemaker, and people with no profession have a lower spending score than people in other professions.

Overall, the histogram indicates that the spending scores of people in different professions vary, with doctors and people in the entertainment industry having the highest spending scores, followed by artists and lawyers. The spending scores of people in healthcare and marketing are also relatively high, while people with no profession and homemakers have a lower spending score.

#Correlation

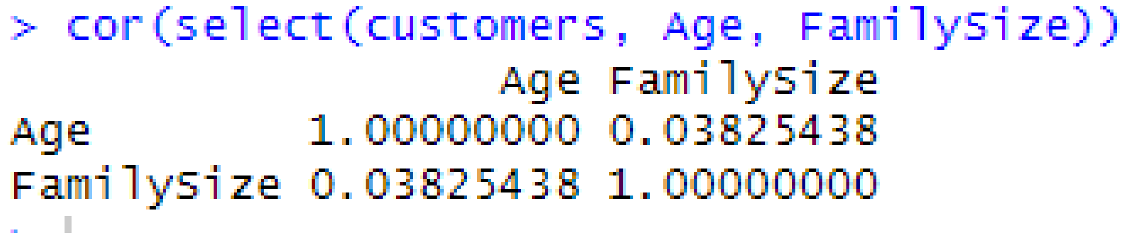
1st correlation shows the relationship between Age and Family Size.



```
> cor.test(customers$Age, customers$FamilySize, method = "spearman")

	Spearman's rank correlation rho

data:  customers$Age and customers$FamilySize
S = 1280876395, p-value = 0.07857
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho 
0.03934246 
```

The cor.test() function was used to calculate the Spearman's rank correlation coefficient (rho) between the variables customers$Age and customers$FamilySize. The computed value of rho is 0.03934246, indicating a very weak positive correlation between the two variables.

The p-value obtained from the test is 0.07857, which is higher than the typical level of significance of 0.05. This implies that there is insufficient evidence to reject the null hypothesis that there is no correlation between the two variables. Hence, we cannot draw a conclusion that there is a statistically significant correlation between age and family size based on this analysis.

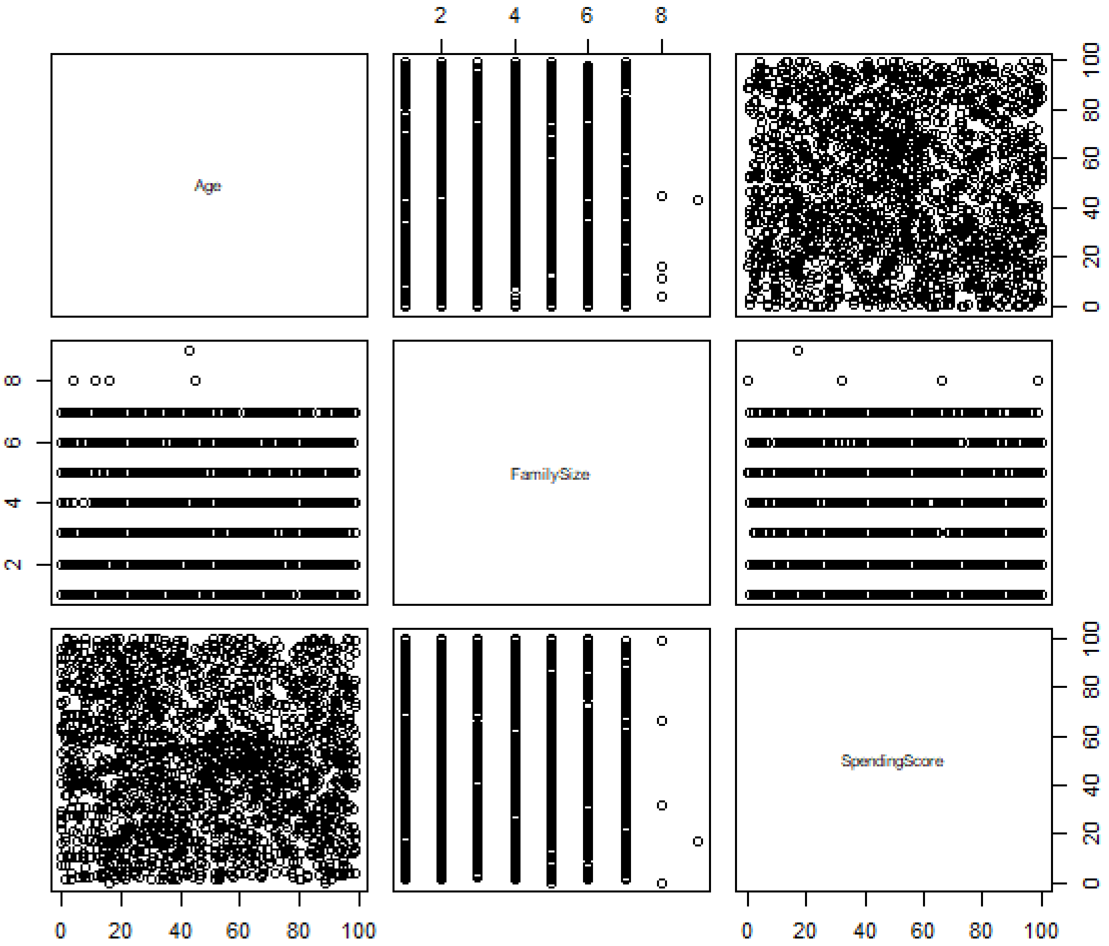


```
> cor(select(customers, Age, FamilySize))
                  Age FamilySize
Age        1.00000000 0.03825438
FamilySize 0.03825438 1.00000000
```

The output above displays the correlation matrix for the variables Age and FamilySize in the customer’s dataset. The computed correlation coefficient between Age and FamilySize is 0.03825438, which implies a very weak positive correlation between the two variables.

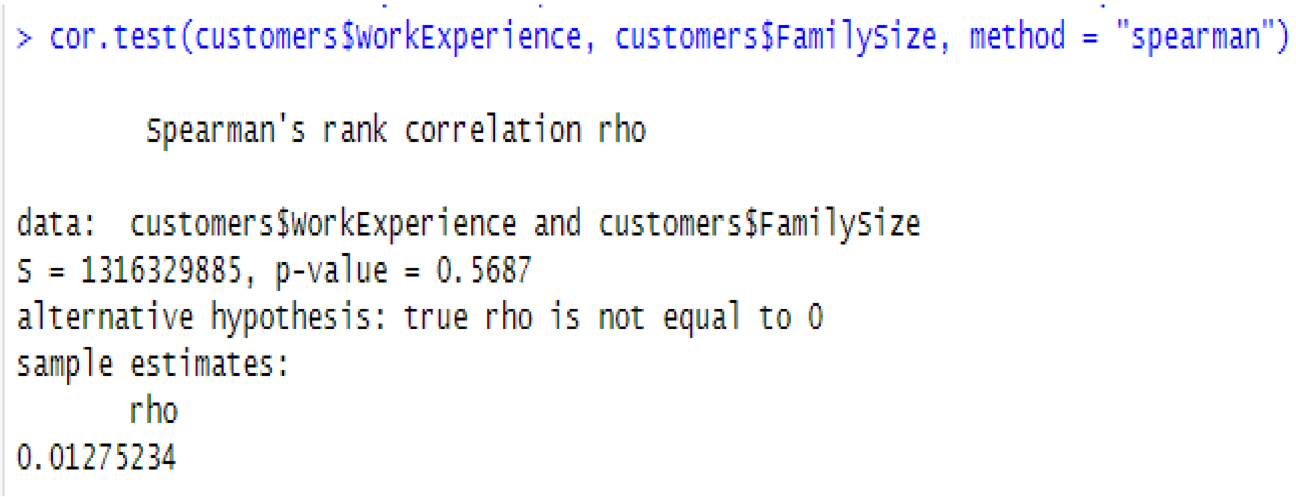
Furthermore, the correlation coefficient between Age and itself is 1, as expected, since any variable has a perfect correlation with itself. The same holds true for FamilySize and itself, which also has a correlation coefficient of 1.

In conclusion, this output suggests that there is little to no meaningful correlation between Age and FamilySize in the customer's dataset.



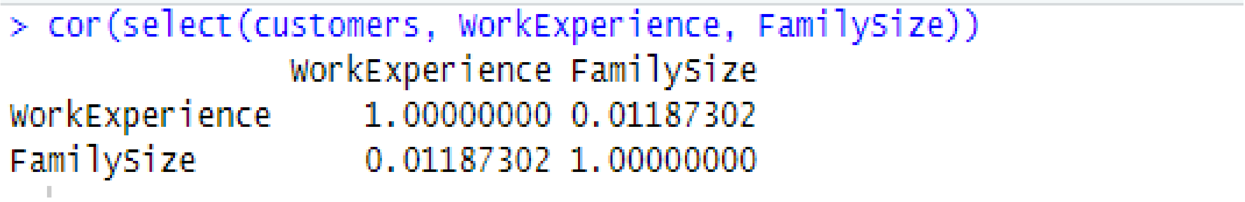
From the image it is seen that there is no relation between the 3 variables.

2nd correlation exists between work experience and Family Size.



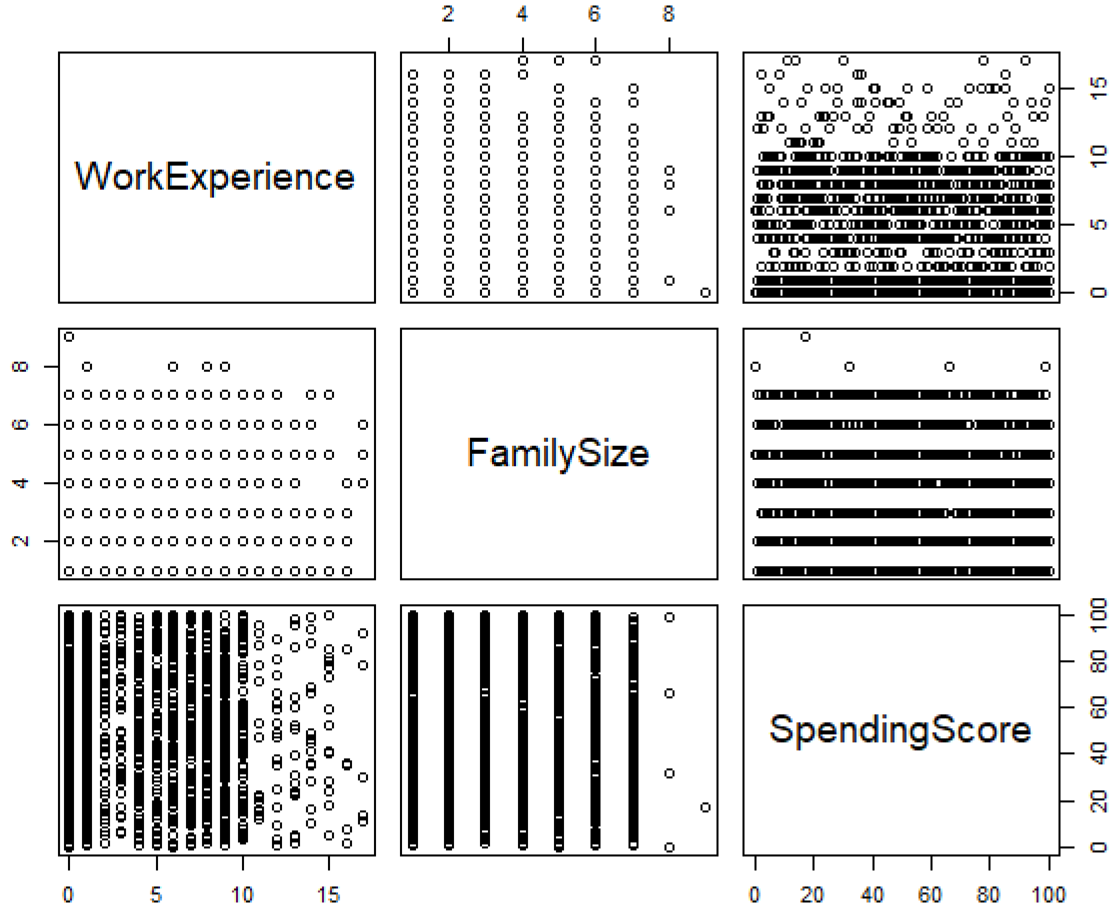
```
> cor.test(customers$WorkExperience, customers$FamilySize, method = "spearman")

	Spearman's rank correlation rho

data:  customers$WorkExperience and customers$FamilySize
S = 1316329885, p-value = 0.5687
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho 
0.01275234 
```

This output is the result of performing a Spearman's rank correlation test between customers$WorkExperience and customers$FamilySize. The correlation coefficient between the two variables is 0.01275234, indicating a very weak positive relationship. The p-value of 0.5687 suggests that there is no statistically significant evidence to reject the null hypothesis of zero correlation between the two variables at the 5% significance level. Therefore, we can conclude that there is no significant correlation between customers$WorkExperience and customers$FamilySize based on this test.



```
> cor(select(customers, WorkExperience, FamilySize))
               WorkExperience FamilySize
WorkExperience     1.00000000 0.01187302
FamilySize         0.01187302 1.00000000
```

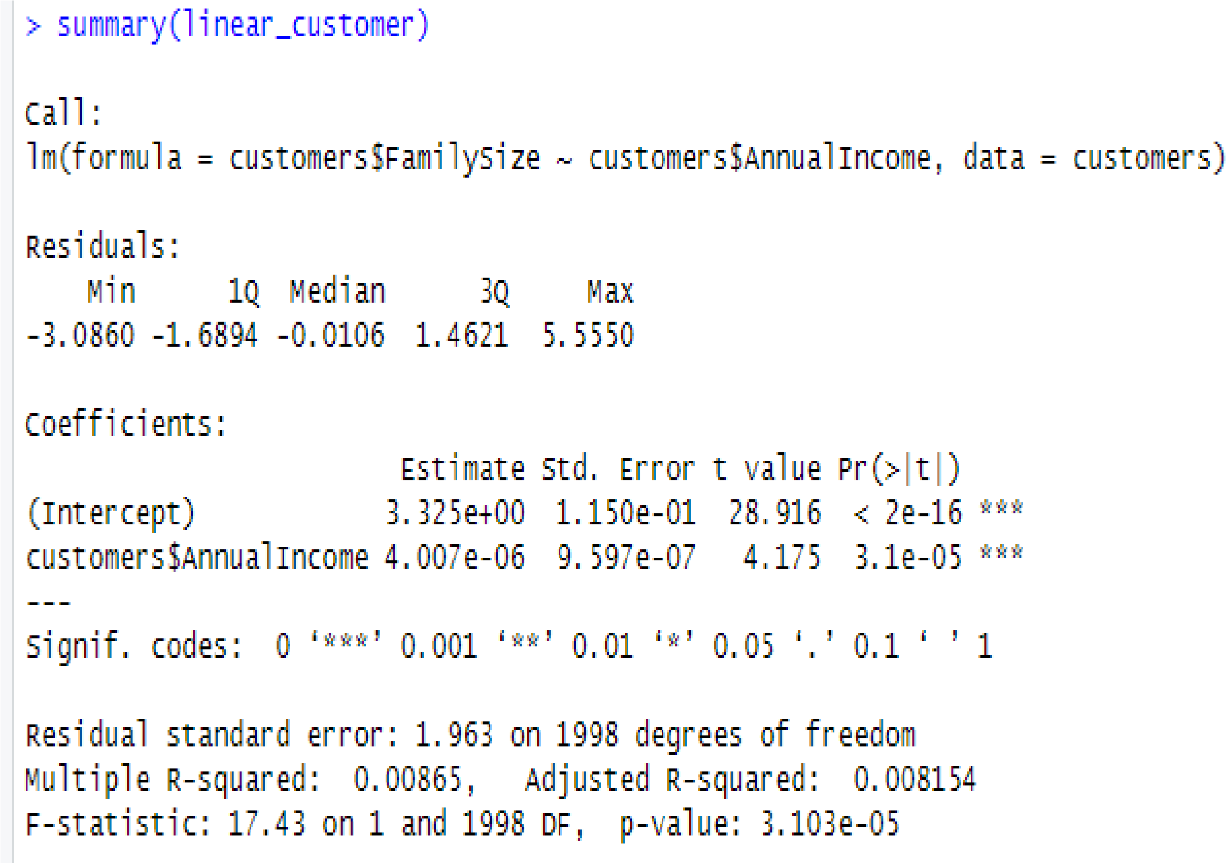
This output shows the result of a Pearson's correlation coefficient calculation between WorkExperience and FamilySize variables in customers’ data frame. The correlation coefficient between the two variables is 0.01187302, which indicates a very weak positive relationship. As the correlation coefficient is close to zero, there is no significant linear correlation between the two variables. We can conclude that there is no significant linear association between WorkExperience and FamilySize in the customer’s data frame based on this test.



From the image it is seen that there is very week relation between the 3 variables.

#Linear Regression

1st Linear Regression exists between the Family Size and annual Income.

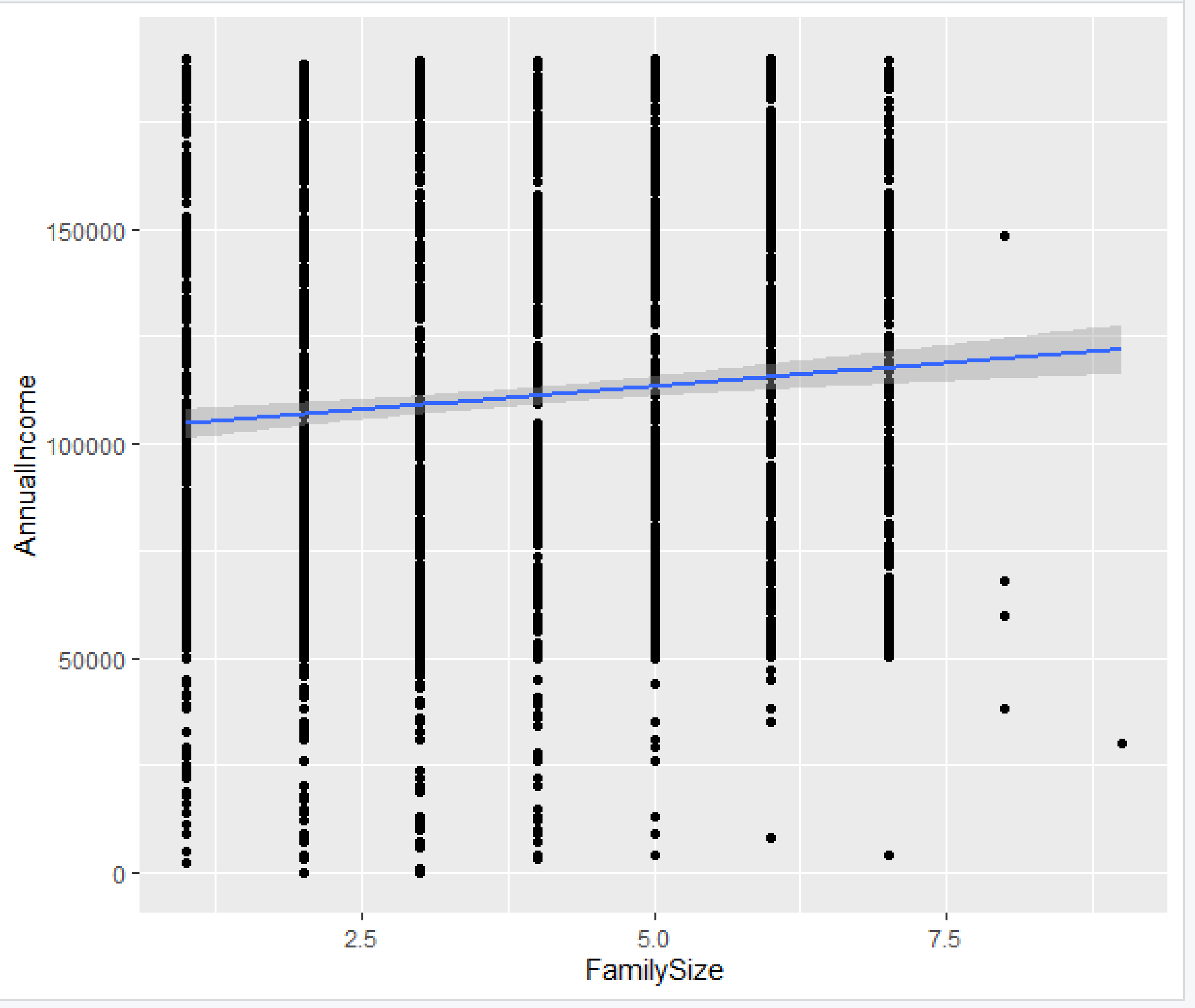


```
> summary(linear_customer)

Call:
lm(formula = customers$FamilySize ~ customers$AnnualIncome, data = customers)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0860 -1.6894 -0.0106  1.4621  5.5550 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            3.325e+00  1.150e-01  28.916  < 2e-16 ***
customers$AnnualIncome 4.007e-06  9.597e-07   4.175  3.1e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.963 on 1998 degrees of freedom
Multiple R-squared:  0.00865,	Adjusted R-squared:  0.008154 
F-statistic: 17.43 on 1 and 1998 DF,  p-value: 3.103e-05
```

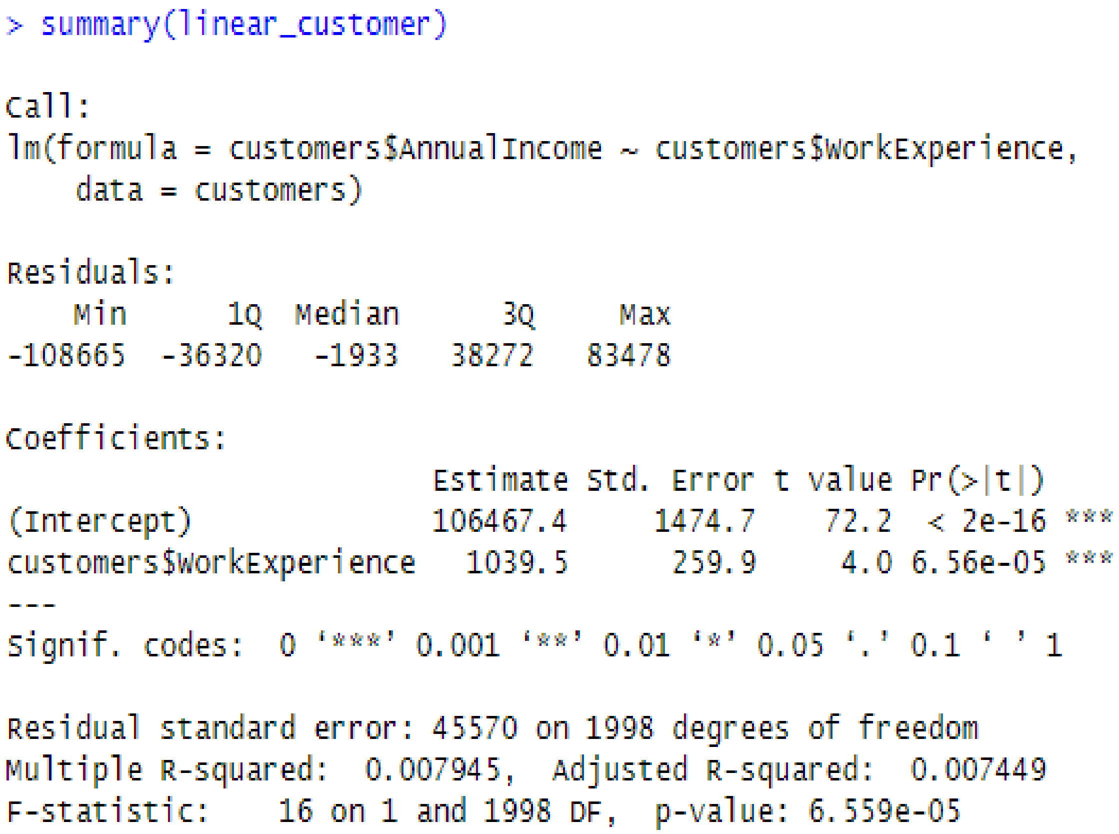
This output displays the outcome of a linear regression analysis between the customers$FamilySize and customers$AnnualIncome variables from the customers data frame. The regression coefficient for customers$AnnualIncome is 4.007e-06, indicating that a unit increase in customers$AnnualIncome is associated with a 4.007e-06 increase in customers$FamilySize. The intercept value of the model is 3.325, which means that when customers$AnnualIncome is zero, customers$FamilySize is expected to be 3.325.

The p-value associated with customers$AnnualIncome is below 0.05, indicating that the coefficient is statistically significant, suggesting that there is a significant positive relationship between customers$AnnualIncome and customers$FamilySize. However, the multiple R-squared values of 0.00865 suggests that only 0.87% of the variance in customers$FamilySize is explained by customers$AnnualIncome.

Overall, the results of the linear regression model suggest that customers$AnnualIncome and customers$FamilySize have a statistically significant positive relationship, but the model has limited explanatory power, explaining only a small proportion of the variation in customers$FamilySize.



2nd Linear Regression exists between the Work Experience and annual Income.

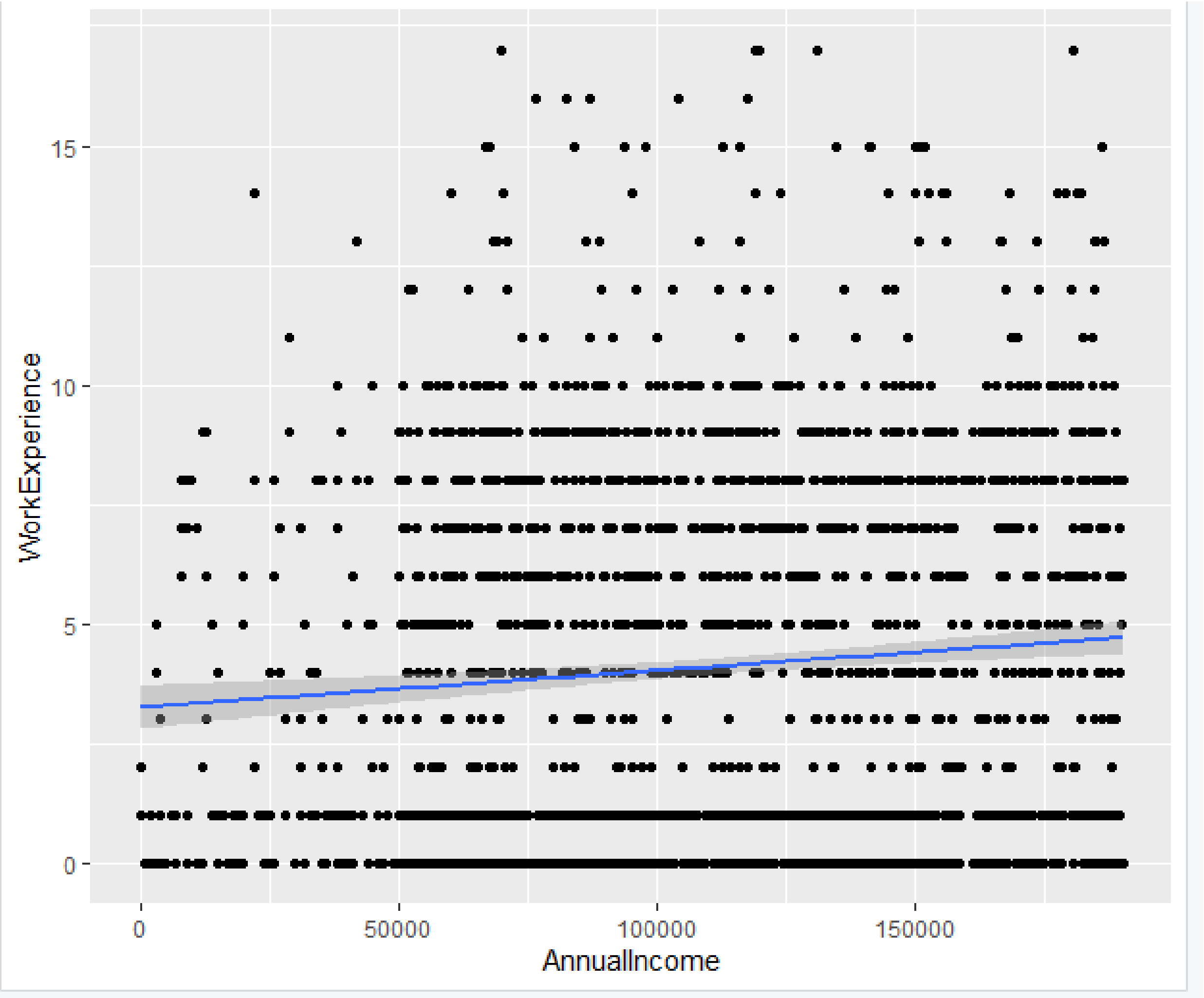


```
> summary(linear_customer)

Call:
lm(formula = customers$AnnualIncome ~ customers$WorkExperience,
    data = customers)

Residuals:
    Min      1Q  Median      3Q     Max 
-108665  -36320   -1933   38272   83478 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)              106467.4     1474.7    72.2  < 2e-16 ***
customers$WorkExperience   1039.5      259.9     4.0 6.56e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45570 on 1998 degrees of freedom
Multiple R-squared:  0.007945,	Adjusted R-squared:  0.007449 
F-statistic:    16 on 1 and 1998 DF,  p-value: 6.559e-05
```

This output presents the result of a linear regression model between customers$AnnualIncome and customers$WorkExperience variables in the customers data frame. The regression coefficient for customers$WorkExperience is 1039.5, indicating that for every unit increase in customers$WorkExperience, there is an increase of 1039.5 in customers$AnnualIncome. The intercept of the model is 106467.4, which suggests that when customers$WorkExperience is zero, customers$AnnualIncome is expected to be 106467.4.

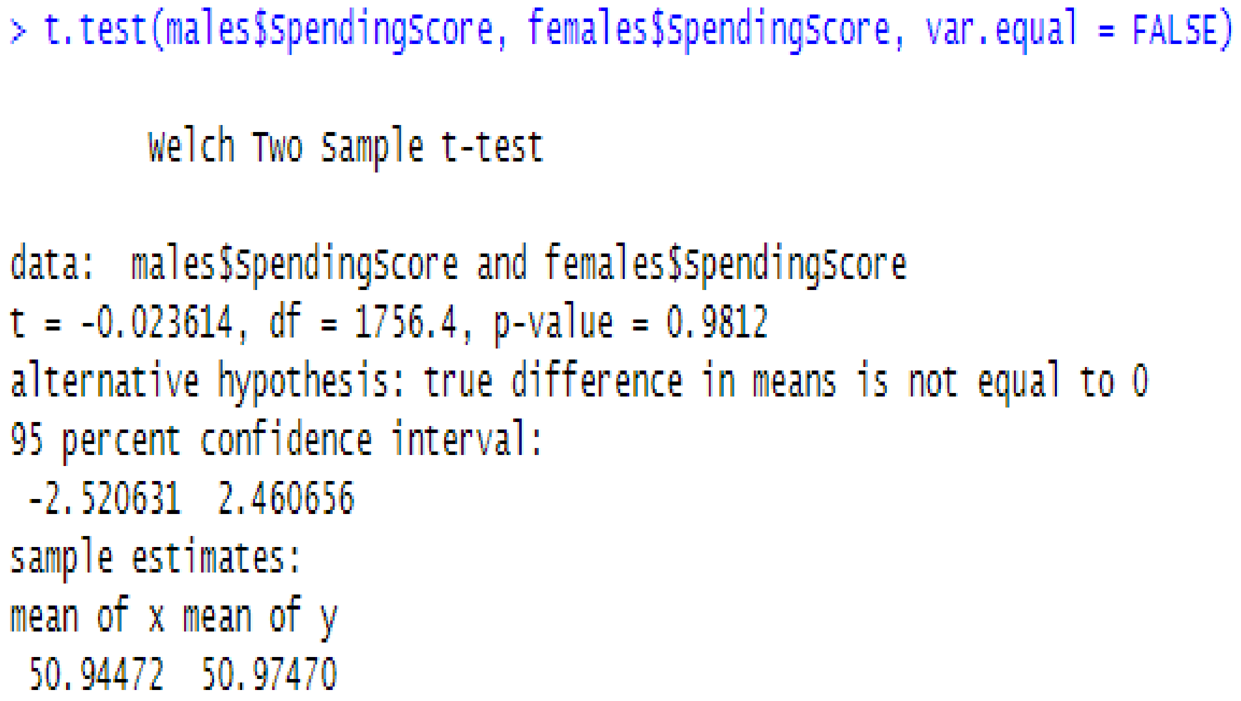
The p-value associated with customers$WorkExperience is less than 0.05, which means that the coefficient is statistically significant. This indicates that there is a significant positive relationship between customers$WorkExperience and customers$AnnualIncome. The multiple R-squared value of 0.007945 suggests that only 0.79% of the variability in customers$AnnualIncome can be explained by customers$WorkExperience.

In summary, the linear regression model indicates that there is a statistically significant positive relationship between customers$WorkExperience and customers$AnnualIncome, but the model explains only a small proportion of the variation in customers$AnnualIncome.



#T-Test

1st t-test exists between male and female spending scores.

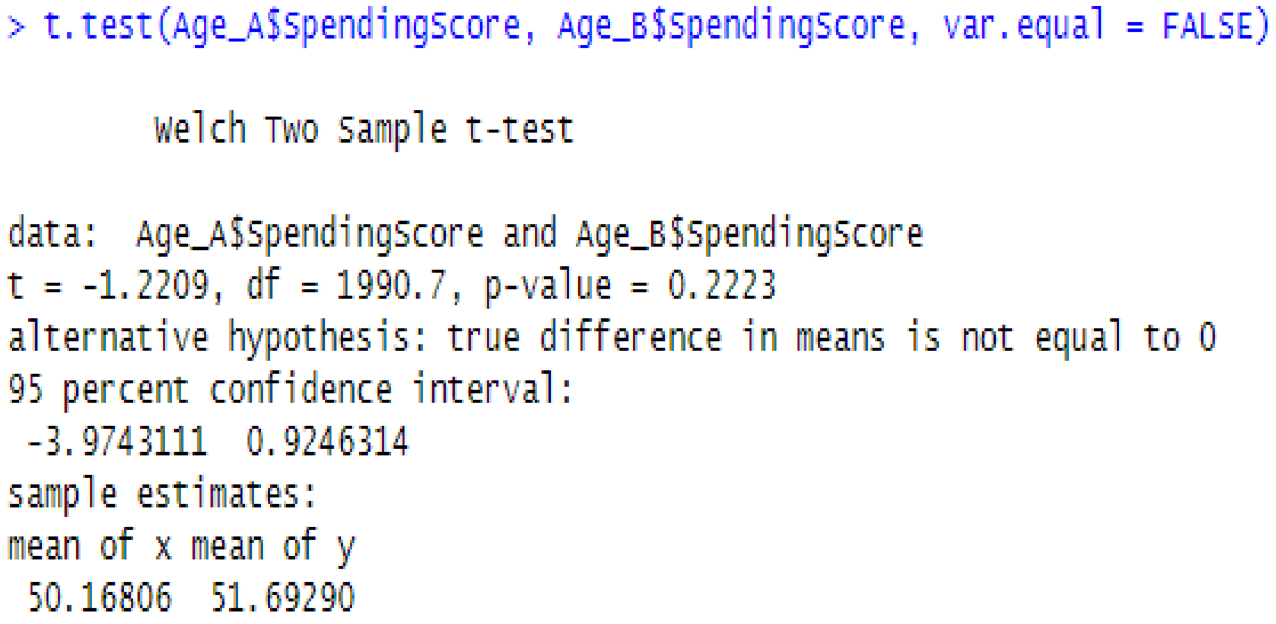


```
> t.test(males$SpendingScore, females$SpendingScore, var.equal = FALSE)

	Welch Two Sample t-test

data:  males$SpendingScore and females$SpendingScore
t = -0.023614, df = 1756.4, p-value = 0.9812
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.520631  2.460656
sample estimates:
mean of x mean of y 
 50.94472  50.97470 
```

The output shows the results of a two-sample t-test to compare the spending scores between males and females. The null hypothesis is that there is no difference in the mean spending scores between males and females. The alternative hypothesis is that the mean spending scores are different for males and females.

The t-statistic is -0.023614 and the degrees of freedom are 1756.4. The p-value is 0.9812, which is greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis and conclude that there is no significant difference in the mean spending scores between males and females. The 95% confidence interval for the difference in means is [-2.520631, 2.460656], which contains zero, further supporting the null hypothesis. The sample means for males and females are 50.94472 and 50.97470, respectively.

2nd t-test exists between male and female spending scores.



```
> t.test(Age_A$SpendingScore, Age_B$SpendingScore, var.equal = FALSE)

	Welch Two Sample t-test

data:  Age_A$SpendingScore and Age_B$SpendingScore
t = -1.2209, df = 1990.7, p-value = 0.2223
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.9743111  0.9246314
sample estimates:
mean of x mean of y 
 50.16806  51.69290 
```

This output shows the results of a two-sample t-test to compare the spending scores between two groups labeled Age_A and Age_B. The null hypothesis is that there is no difference in the mean spending scores between these two groups. The alternative hypothesis is that the mean spending scores are different between the two groups.

The t-statistic is -1.2209 and the degrees of freedom are 1990.7. The p-value is 0.2223, which is greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis and conclude that there is no significant difference in the mean spending scores between Age_A and Age_B. The 95% confidence interval for the difference in means is [-3.9743111, 0.9246314], which contains zero, further supporting the null hypothesis. The sample means for Age_A and Age_B are 50.16806 and 51.69290, respectively.