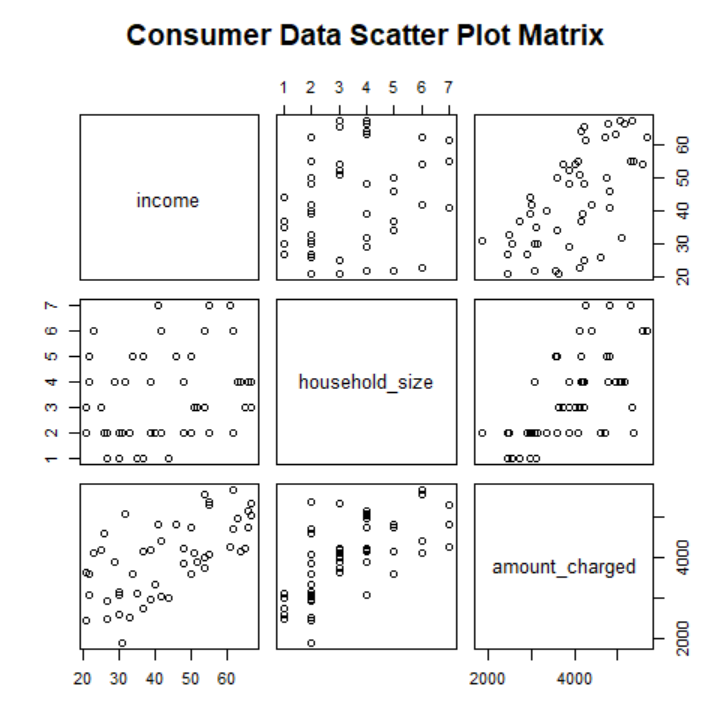
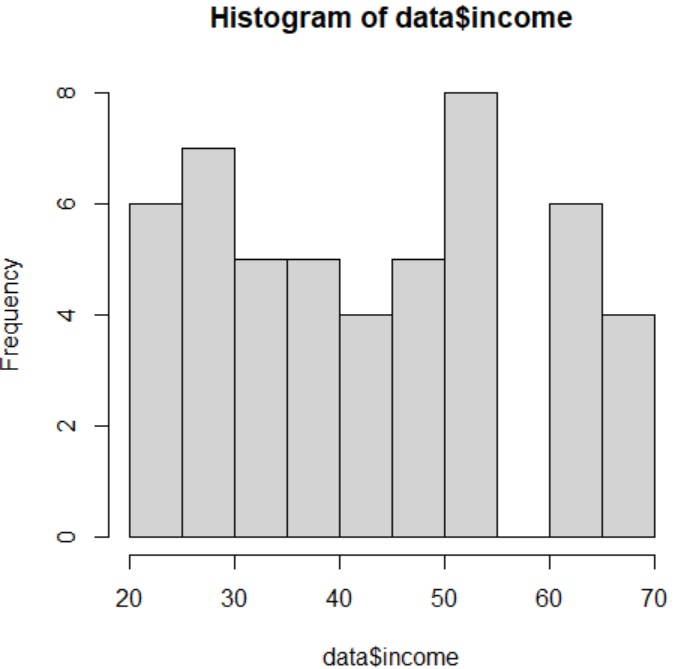
**QUESTION 1**

Scatter plot: Plot(data)

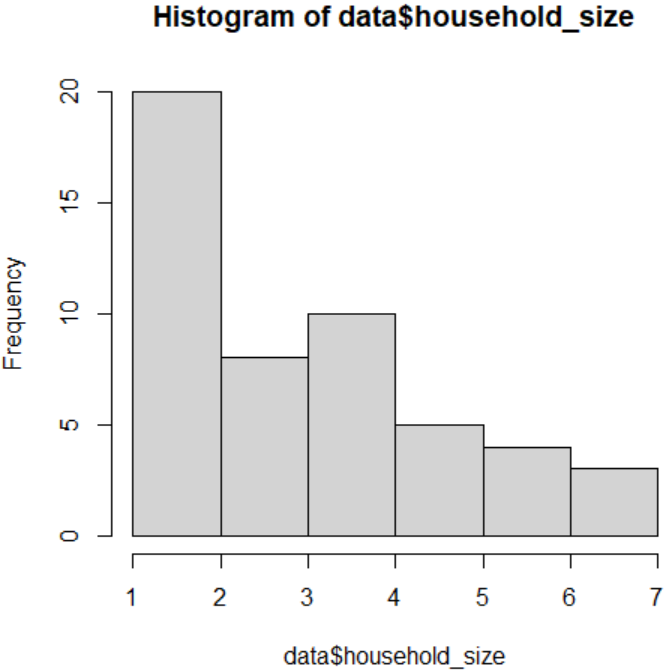


Normal Distribution through Histogram:

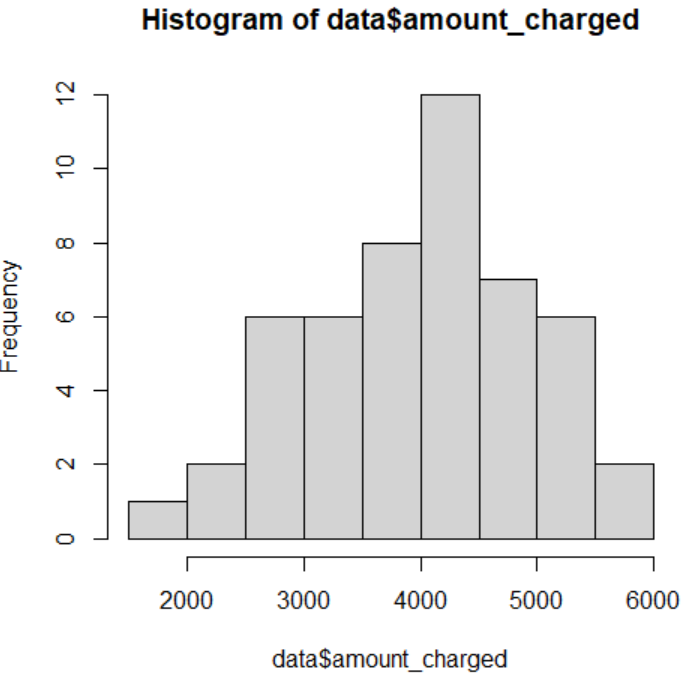
hist(data$income)



hist(data$household\_size)

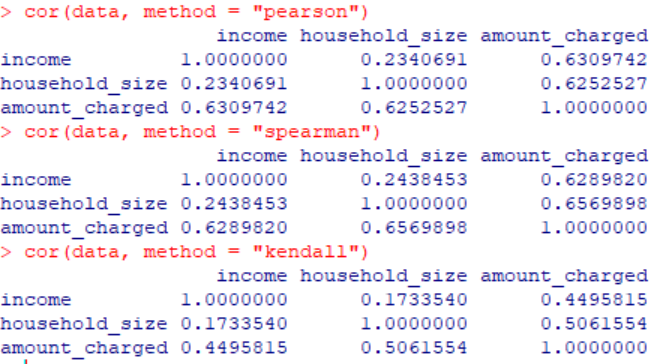


hist(data$amount\_charged)



Since not all the data follows normal distribution we use non-parametric tests for correlation.

Correlation:



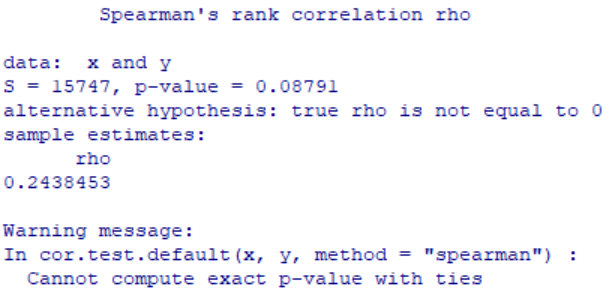
Correlation Test:

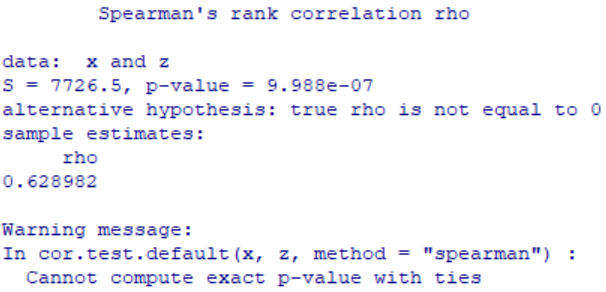
Spearman:

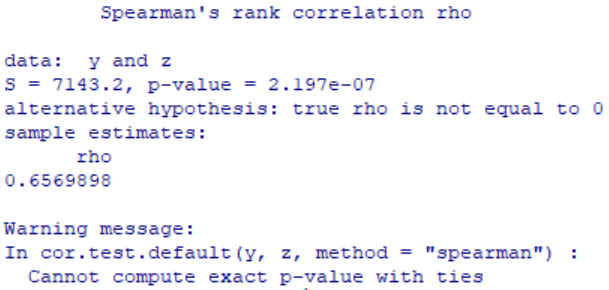
cor.test(x,y, method = "spearman")

cor.test(x,z, method = "spearman")

cor.test(y,z, method = "spearman")







For alpha = 0.05, x and y have a low correlation and are insignificant as p value is greater

than alpha. x and y have a high correlation but as the p value is lesser than alpha, it is

significant. Lastly, y and z have a high correlation but as the p value is lesser than alpha,

it is significant.

**QUESTION 2**

data <- data.frame(

income = c(54, 30, 32, 50, 31, 55, 37, 40, 66, 51, 25, 48, 27, 33, 65, 63, 42, 21, 44, 37, 62, 21, 55, 42, 41, 54, 30, 48, 34, 67, 50, 67, 55, 52, 62, 64, 22, 29, 39, 35, 39, 54, 23, 27, 26, 61, 30, 22, 46, 66),

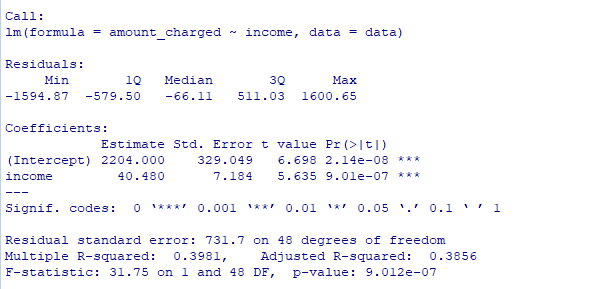
household\_size = c(3, 2, 4, 5, 2, 2, 1, 2, 4, 3, 3, 4, 1, 2, 3, 4, 6, 2, 1, 5, 6, 3, 7, 2, 7, 6, 1, 2, 5, 4, 2, 3, 2, 3, 2, 4, 5, 4, 2, 1, 4, 3, 6, 2, 2, 7, 2, 4, 5, 4),

amount\_charged = c(4016, 3159, 5100, 4742, 1864, 4070, 2731, 3348, 4764, 4110, 4208, 4219, 2477, 2514, 4214, 4965, 4412, 2448, 2995, 4171, 5678, 3623, 5301, 3020, 4828, 5573, 2583, 3866, 3586, 5037, 3605, 5345, 5370, 3890, 4705, 4157, 3579, 3890, 2972, 3121, 4183, 3730, 4127, 2921, 4603, 4273, 3067, 3074, 4820, 5149)

)

lm\_income <- lm(amount\_charged ~ income, data = data)

summary(lm\_income)

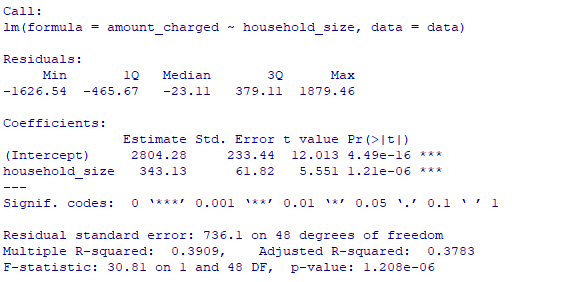


EQUATION:

Amount Charged = 2204 + 40.480 \* Income

lm\_household\_size <- lm(amount\_charged ~ household\_size, data = data)

summary(lm\_household\_size)



EQUATION:

Amount Charged = 2804.28 + 343.13 \* Household Size

By examining the coefficient estimates and p-values, we can determine which variable is a better predictor of annual card charges.

1. Examining the coefficient estimates: Household size has larger coefficient estimates.
2. Examining the p-values: Household size has smaller p value.

Hence, household size is better predictor of annual card charges.

**QUESTION 3**

data <- data.frame(

income = c(54, 30, 32, 50, 31, 55, 37, 40, 66, 51, 25, 48, 27, 33, 65, 63, 42, 21, 44, 37, 62, 21, 55, 42, 41, 54, 30, 48, 34, 67, 50, 67, 55, 52, 62, 64, 22, 29, 39, 35, 39, 54, 23, 27, 26, 61, 30, 22, 46, 66),

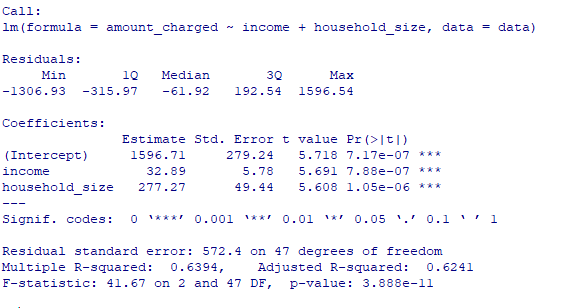
household\_size = c(3, 2, 4, 5, 2, 2, 1, 2, 4, 3, 3, 4, 1, 2, 3, 4, 6, 2, 1, 5, 6, 3, 7, 2, 7, 6, 1, 2, 5, 4, 2, 3, 2, 3, 2, 4, 5, 4, 2, 1, 4, 3, 6, 2, 2, 7, 2, 4, 5, 4),

amount\_charged = c(4016, 3159, 5100, 4742, 1864, 4070, 2731, 3348, 4764, 4110, 4208, 4219, 2477, 2514, 4214, 4965, 4412, 2448, 2995, 4171, 5678, 3623, 5301, 3020, 4828, 5573, 2583, 3866, 3586, 5037, 3605, 5345, 5370, 3890, 4705, 4157, 3579, 3890, 2972, 3121, 4183, 3730, 4127, 2921, 4603, 4273, 3067, 3074, 4820, 5149)

)

lm\_model <- lm(amount\_charged ~ income + household\_size, data = data)

summary(lm\_model)



EQUATION:

Amount Charged = 1596.71 + 32.89 \* Income + 277.27 \* Household Size

**QUESTION 4**

Household size=3

Annual income=40,000

Amount Charged = 1596.71 + 32.89 \* 40000 + 277.27 \* 3

Amount Charged=$1318028.52

**QUESTION 5**

To enhance the predictive power and accuracy of the regression model, it is worth considering the inclusion of additional independent variables that could potentially influence the amount charged.

Age: The age of individuals could potentially impact their spending habits and financial situations

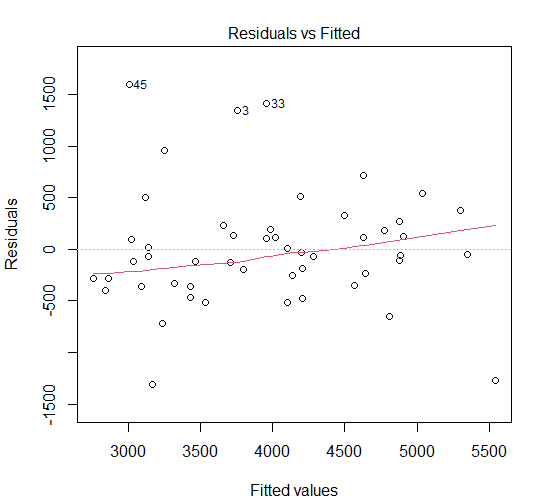
Employment Status: Whether individuals are employed, unemployed, self-employed, or retired can have an impact on their income and spending behavior.

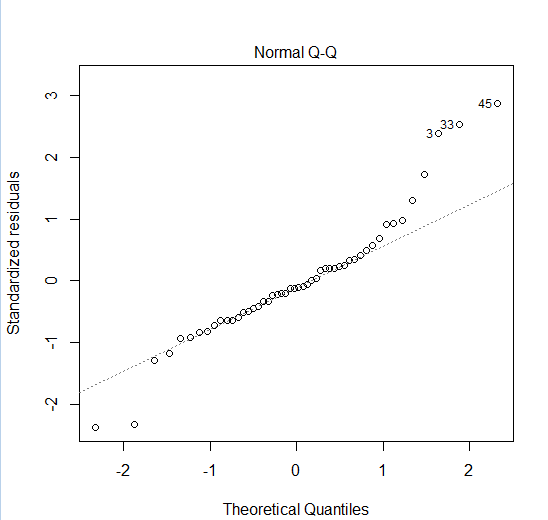
Marital Status: Marital status may be related household financial management, which can impact the amount charged.

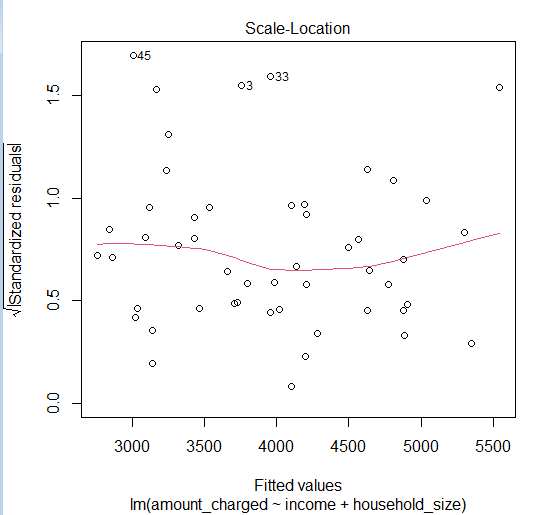
**QUESTION 6**

res=residuals(lm\_model)

plot(lm\_model)







The model vs fitted graph shows no clear pattern hence it shows that linear regression is a good fit for the data. It also suggests that there is a positive relationship between household size + annual income and the amount charged by credit card users. The data doesn't follow normal distribution. The R-squared value of the model is 0.94 this suggests that annual income and household size are important predictors of the amount charged by credit card users.