**Project-1**

Paresh Patil

Data Analysis

Exploratory Data Analysis with ‘R’

**Question 1.1**

1. Create a table to show the total income by each category of marital status.

Output:

|  |
| --- |
| M\_Status Total\_Income |
| 1 divorced 47305889 |
| 2 married 72416616 |
| 3 never 31039000 |
| 4 widowed 8788122 |

1. Which status has the highest total income?

Output:

|  |
| --- |
| M\_Status Total\_Income |
| married 72416616 |

**Question 1.2**

1. Calculate the mean age of respondents born in Asia.

Output:

Asia

40.96185

1. Calculate the mean age of respondents born in Asia weighted by the number of children they have.

Output:

[1] 40.61

**Question 1.3**

1. Create a table to show the mean score on the political awareness test for males compared to females

Output:

|  |
| --- |
| Gender mean\_score |
| female 0.0068007663 |
| male 0.0007063197 |

1. Which has a higher score?

Output:

|  |
| --- |
| Gender mean\_score |
| female 0.006800766 |

**Question 1.4**

1. Calculate the 34th and 63rd percentiles of percentage of time taken on the test.

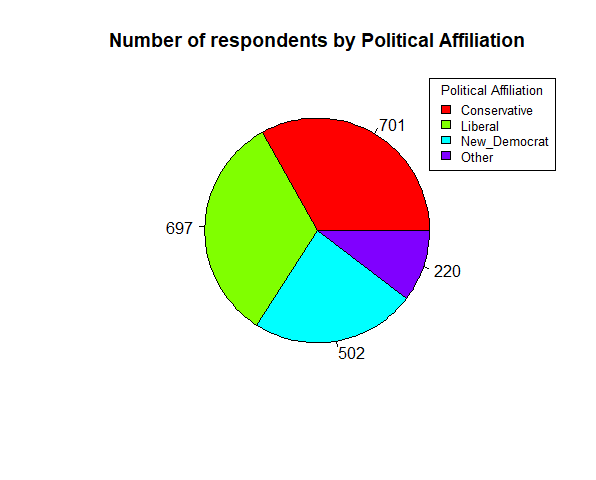
Output:

|  |
| --- |
| 34% 63% |
| 0.20 0.61 |

**Question 2.1**

1. Create a pie chart showing the number of respondents by Political Affiliation.

Output:



1. Which Political Affiliation contains the most respondents (remember each row of your study file represents one respondent)?

Output:

|  |
| --- |
| Political pol\_percent |
| Conservative 701 |

1. Which Political Affiliation has the fewest respondents?

|  |
| --- |
| Output: |
| Political pol\_percent |
| Other 220 |

**Question 2.2**

1. Create a table that shows the percentage of respondents from each Region that are in the Treatment group.

Output:

|  |
| --- |
| nation group percentage |
| Asia treat 22.51163 |
| Europe treat 15.16279 |
| North America treat 50.79070 |
| Southern treat 11.53488 |

1. Which region has the highest percentage of people in the Treatment group?

Output:

|  |
| --- |
| Nation group percentage |
| North America treat 50.7907 |

1. Which region has the lowest percentage of people in the Treatment group?

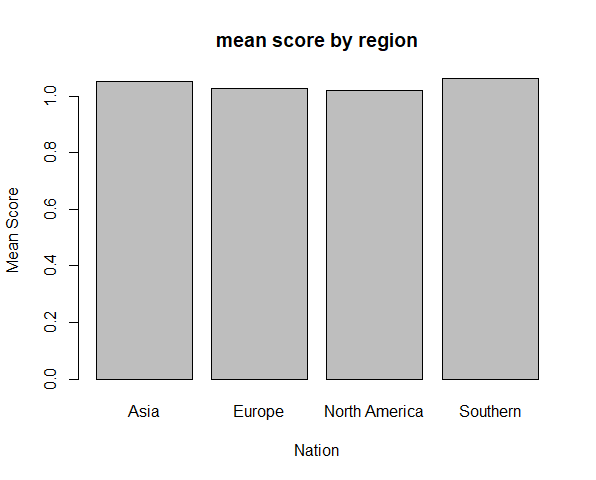
Output:

|  |
| --- |
| Nation group percentage |
| Southern treat 11.53488 |

**Question 2.3**

1. Create a bar chart showing the mean Standardized Test Score on the Political Awareness Test for each Region.

Output:



1. Which Region has the lowest mean score?

Output:

|  |
| --- |
| Nation Percentage |
| North America 1.019343 |

1. Which Region has the highest mean score?

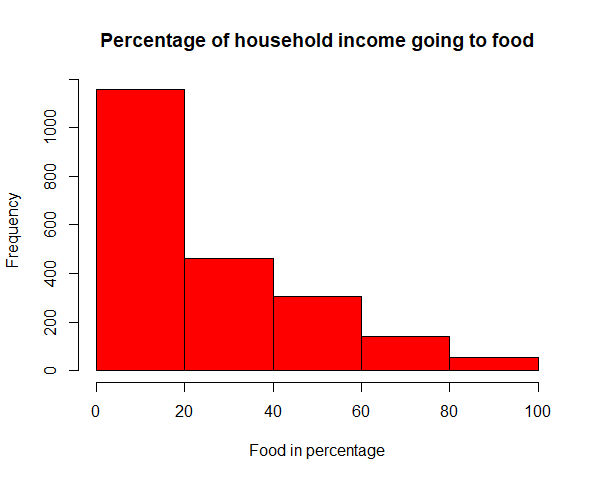
Output:

|  |
| --- |
| Nation Percentage |
| Southern 1.062424 |

**Question 2.4**

1. Create a histogram with 5 bins showing the distribution of the percentage of household income going to food.

Output:



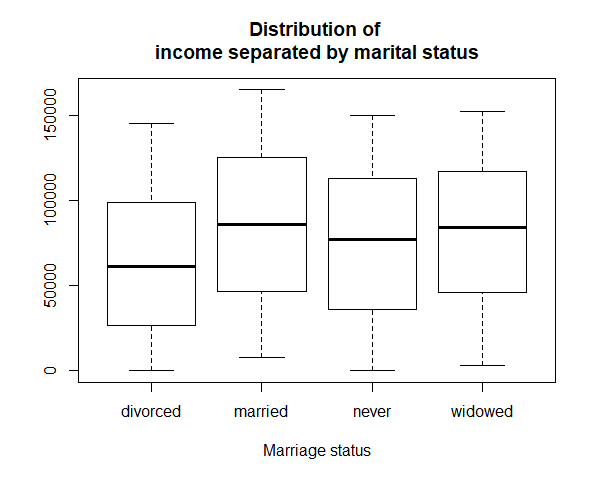
1. Which range of values has the highest frequency?

Output: 0% to 20% : 1155

**Question 2.5**

1. Create a sequence of box plots showing the distribution of income separated by marital status.

Output:



1. According to the charts, which martial status has the highest average income?

Output: Married: 81730.63

1. Which marital status has the lowest average income?

Output: Divorced: 57065.78

1. Which marital status has the greatest variability in income?

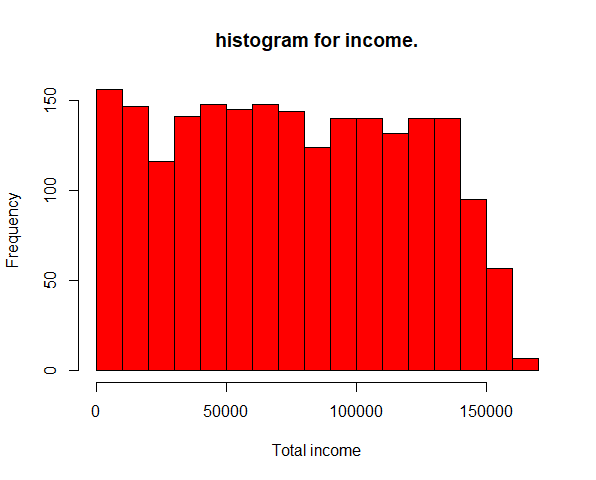
Output: Married status has the greatest variability in income.

(Distance greater between high and median )

**Question 2.6**

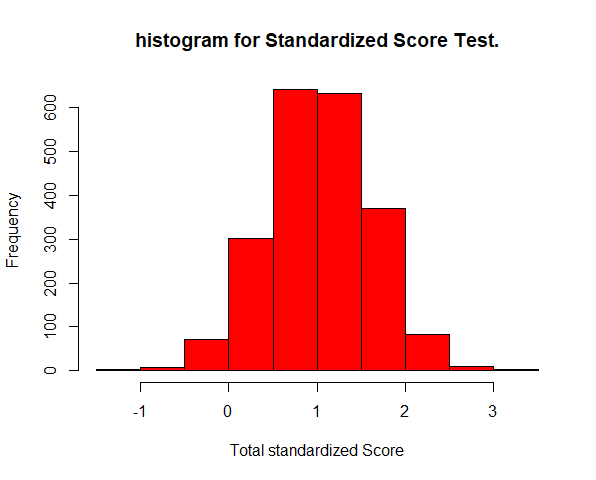
1. Create a histogram for income

Output:



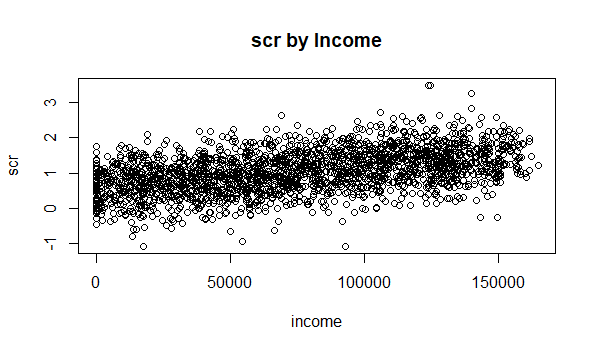
1. Create a histogram for standardized score

Output:



1. Create a scatter plot showing the relationship between the income and standardized score. (Note: income should be on the x-axis, standardized score should be the y-axis)

Output:



1. What conclusions, if any, can you draw from the chart?

Output: In above plot, there is still a trend in the data, Thus, we would say that the data has a weak or lower correlation.

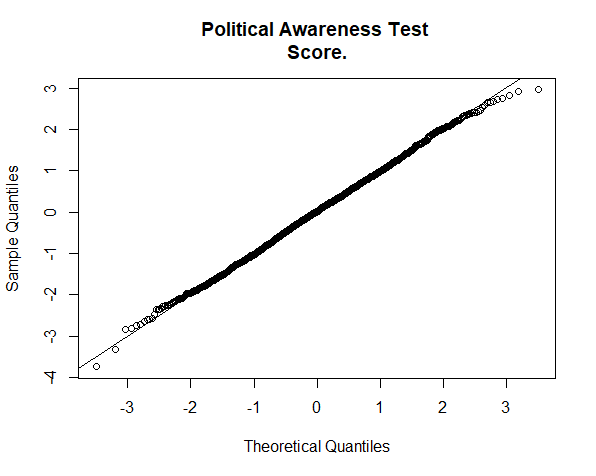
1. Calculate a correlation coefficient between these two variables. What conclusion you draw from it?

Output: r = 0.4568418 (0.25 ≤ 𝑟 < 0.50 Weak linear relationship)

**Question 3.1**

1. Create a QQ Normal plot of the Political Awareness Test Score

Output:



1. Conduct a statistical test for normality on the Political awareness Test Score.

Output: Shapiro-Wilk normality test

data: PROG8430\_Assign\_Explore$score

W = 0.99921, p-value = 0.5068

1. Are the Political Awareness Test Scores normally distributed? What led you to this conclusion?

Output: Yes, it is very close to normal distribution.

Because the points are on straight line (it should fall on straight line through first and third quartiles) according to the value of W=0.99 approximately 1 and p value is greater than 0.05 .

**Question 3.2**

1. Compare Political Awareness Test Scores between the treatment and control group using a suitable hypothesis test

Output: Two Sample t-test

data: score by group

t = -1.4694, df = 2118, p-value = 0.1419

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.14831720 0.02125633

sample estimates:

mean in group control mean in group treat

-0.02850718 0.03502326

Additional test we must do before t.test

F test to compare two variances

data: score by group

F = 0.9583, num df = 1044, denom df = 1074, p-value = 0.4887

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.8495287 1.0811154

sample estimates:

ratio of variances

0.958302

b) Explain why you chose the test you did ?

Output: Because the data is continuous, independent, normally distributed and Variance is unknown, but equal as 1(below is the variance test).

F test to compare two variances

data: score by group

F = 0.9583, num df = 1044, denom df = 1074, p-value = 0.4887

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.8495287 1.0811154

sample estimates:

ratio of variances

0.958302

c)Do you have strong evidence that the average test scores are different between the treatment and control groups?

Output: Strong Evidence that differences are not significant.

mean in group control mean in group treat

-0.02850718 0.03502326

**Question 3.3**

1. Determine if the Score on the Political Awareness Test varies by region using ANOVA (statistical) and a sequence of boxplots (graphical).

Output: nation Residuals

Sum of Squares 2.5041 2097.5034

Deg. of Freedom 3 2116

Residual standard error: 0.9956197

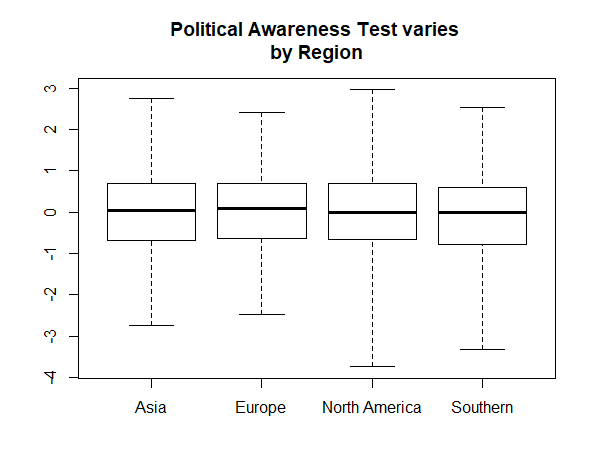
Estimated effects may be unbalanced

Summary:

Df Sum Sq Mean Sq F value Pr(>F)

nation 3 2.5 0.8347 0.842 0.471

Residuals 2116 2097.5 0.9913



b) Determine if the Measure of Political Involvement (Pol) varies by Political Affiliation using ANOVA and a sequence of boxplots

Output: political Residuals

Sum of Squares 565.029 3582.719

Deg. of Freedom 3 2116

Residual standard error: 1.301213

Estimated effects may be unbalanced

Summary:

Df Sum Sq Mean Sq F value Pr(>F)

political 3 565 188.34 111.2 <2e-16 \*\*\*

Residuals 2116 3583 1.69

