PROG8435: Data Analysis Mathematics, Algorithms and Modeling

ASSIGNMENT 1

EXPLORATORY DATA ANALYSIS WITH ‘R’

NAME: MOHAMMED ADEEN SHAIK

STUDENT ID: 8969152

1. TASK 0

Prerequisites:

LIBRARIES:

Must have installed library **ggplot2**

SETTING WORK DIRECTORY

Change the work directory to your local file path to run the r file and see all the outputs and charts.

You can change the work directory file path inside this:

**setwd ("C:/Users/Adeen/Desktop/R scripts/Assignment 1")**

LOADING THE DATASET:

The dataset **PROG8435\_Assign\_Explore\_24S.csv** must be loaded into a dataframe in the r file using **read.csv**

**Overview of the data:** use head(dataframe\_name) to get an idea of the dataset.

1. TASK 1

Summarizing the Data

Question 1a:

A screenshot of a computer

Description automatically generated

Question 1b:

A screenshot of a computer screen

Description automatically generated

Question 2a:

A screenshot of a computer program

Description automatically generated

Question 2b:

A screenshot of a computer program

Description automatically generated

Question 3a:

A screenshot of a computer screen

Description automatically generated

Question 3b:

A screenshot of a computer program

Description automatically generated

Question 4:

A screenshot of a computer program

Description automatically generated

1. TASK 2:

Organizing data

Question 1a: (Pie chart)

A screenshot of a computer

Description automatically generated

Question 1b: (since it is difficult to use the pie chart to determine the affiliation with most respondents I will use which.max() function to get it the answer to this question.)

A black screen with orange text

Description automatically generated

Question 1c

A black screen with orange text

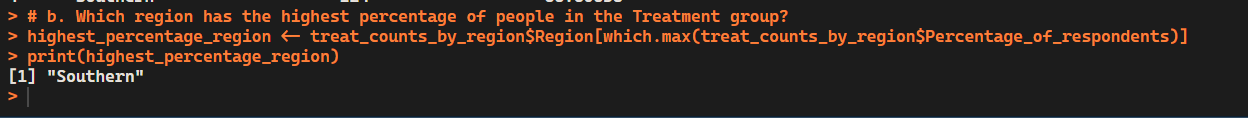
Description automatically generated

Question 2a: The summary table has the region(nation), Number of respondents (from the Treatment group ) , percentage of respondents (belonging to treatment group) from the region

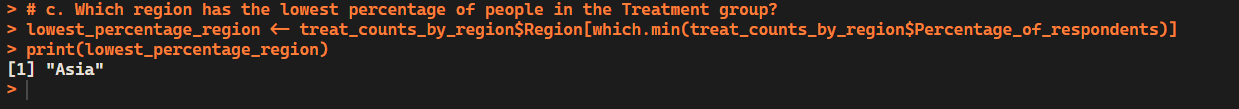
A screenshot of a computer

Description automatically generated

Question 2b:



Question 2c:



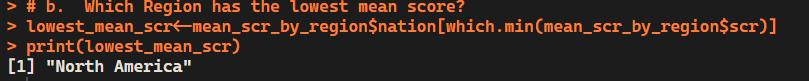
Question 3a: (Bar chart)

A screenshot of a computer

Description automatically generated **note:** I had to adjust the y-axis (1.0 to 1.2) to get the right visuals.

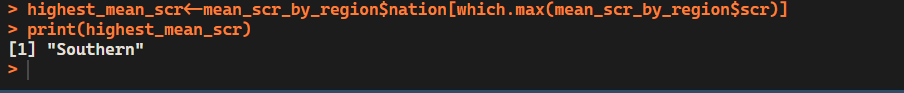
3b: Which Region has the lowest mean score?

North America / Answer provided from both observing the chart and coded output:



Question 3c: Which Region has the highest mean score?

Southern / code and output:



Question 4a: (Histogram)

A graph of a graph

Description automatically generated with medium confidence

Question 4b: Which range of values has the highest frequency?

By observing the histogram, I conclude that the between 0 to 25 (0,25) has the highest frequency among the intervals. And further deeper range is from 5 to 10% where 10% could be the highest frequency.

**Descriptive Analysis:**

**Conclusion:** The distribution is Right-skewed, meaning that most households spend a lower percentage of their income on food, while fewer households spend a higher percentage.

Question 5a:

A screenshot of a computer

Description automatically generated

Question 5b: According to the charts, which marital status has the highest average income?

*‘Married’ marital status has the highest average. income*

Question 5c: Which marital status has the lowest average income?

*‘Divorced’ marital status has the lowest average. income*

Question 5d: Which marital status has the greatest variability in income?

‘Married’ marital status has the greatest variability. My reasoning is that, it has the highest IQR as it has a wide box in the boxplot.

Question 6a: Create a histogram for income

A graph of blue bars

Description automatically generated with medium confidence

Question 6b: Create a histogram for Standardized score

A graph with blue bars

Description automatically generated with medium confidence

Question 6c: Create a scatter plot showing the relationship between the income and standardized score

A screen shot of a graph

Description automatically generated

Question 6d: Conclusion from the scatter plot

A descriptive analysis of the scatter plot:

There seems to be a wide variability in scores across all income levels. And their relationship could be said to be weak as they hardly affect each other.

There are also some outliers that can be pointed out, mostly from the higher income ranges, these outliers stand out by having either too high or too low standardized scores compared to the majority of the data points.

Question 6e: Calculate a correlation coefficient between these two variables. What conclusion do you draw from it?

A screenshot of a computer

Description automatically generated

**CONCLUSION:**

The relationship between income and the standardized score is moderate positive relationship. As income rises, the standardized score also tends to rise. Their strength of relationship and dependence is not strong but its not weak either.