PROG8430 – Data Analysis, Modeling and Algorithms

Assignment 1

Exploratory Data Analysis with 'R'

DUE BEFORE OCTOBER 5; 10PM

1. Submission Guidelines

All assignments must be submitted via the econestoga course website before the due date in to the assignment folder.
You may make multiple submissions, but only the most current submission will be graded.

SUBMISSIONS

In the Assignment 1 Folder submit:

1. Your R Code
2. A Word document containing your answers to the questions following the format provided in the assignment (i.e. using the example assignment).

DO NOT PUT THE DOCUMENTS IN TO A ZIP FILE!

PLEASE NOTE: Your Word document forms the basis of your evaluation and will be what is marked. The R Code is there for me to verify your results and therefore your commentary and output *must* be included in your Word document to be evaluated.

EXAMPLES: The example output provided is simply to demonstrate what a typical submission might look like. You can use it as a basis, but your submission must be in your own words. Submissions that simply "cut and paste" my example commentary will be marked 0.

All variables in your code must abide by the naming convention [variable_name]_[intials]. For example, my variable for State would be State_DM. Follow the example in the into video from week 1.

You may only use base R (i.e. no additional packaged may be used)

THIS IS AN INDIVIDUAL ASSIGNMENT. UNAUTHORIZED COLLABORATION IS AN ACADEMIC OFFENSE. Please see the Conestoga College Academic Integrity Policy for details.

2. Grading

This assignment is worth 15% of your total grade in the course and you can expect it to take five to eight hours.

It is out of 32 marks overall.

Assignments submitted after 10pm will be reduced 20%. Assignments received after 8:00am the morning after the due date will receive a mark of 0%.

Assignments which do not follow the submission instructions may have marks deducted.

3. Data

Each student will have access to the study dataset.

STUDY DATASET:

PROG8430_Assign_Explore_21F.Rdata

Appendix one contains a data dictionary for the study file.

4. Background

A survey of 2120 residents of Canada was conducted to determine the key factors associated with political engagement. A variety of variables were measured and recorded including some tests they were asked to complete. Appendix 1 contains the data dictionary for the data set. One group of respondents ("Treat") were given additional education on political matters while the other ("Control") were not.

All of the tasks have been completed using the examples presented in class. A careful review of your notes from the lectures should give you everything you need to complete these tasks. Additionally, example output for most of the questions has been provided in the Appendix. NOTE – The sample output is to demonstrate only and your output data will not precisely match it.

All of your charts, tables and graphs should be properly labelled.

5. Assignment Tasks

Nbr	Description	Marks
1	Summarizing Data 1. Summary Table a. Create a table to show the total income by each category of marital status. b. Which status has the highest total income? 2. Calculate the mean (rounded to two decimal places) a. Calculate the mean age of respondents born in Asia.	8

	b. Calculate the mean age of respondents born in Asia weighted by the number of children they have. 3. Table Comparison a. Create a table to show the mean score on the political awareness test for males compared to females. b. Which has a higher score? 4. Calculate the 34th and 63rd percentiles of percentage of time taken on the test.	
2	Organizing Data **NOTE:** All charts should have appropriated scaled axes and should be properly labelled. 1. Pie Chart a. Create a pie chart showing the number of respondents by Political Affiliation. b. Which Political Affiliation contains the most respondents (remember each row of your study file represents one respondent)? c. Which Political Affiliation has the fewest respondents? 2. Summary table a. Create a table that shows the percentage of respondents from each Region that are in the Treatment group. b. Which region has the highest percentage of people in the Treatment group? c. Which region has the lowest percentage of people in the Treatment group? 3. Bar Chart a. Create a bar chart showing the mean Standardized Test Score on the Political Awareness Test for each Region. b. Which Region has the lowest mean score? c. Which Region has the highest mean score? 4. Histogram a. Create a histogram with 5 bins showing the distribution of the percentage of household income going to food. b. Which range of values has the highest frequency? 5. Box Plots a. Create a sequence of box plots showing the distribution of income separated by marital status. b. According to the charts, which martial status has the highest average income? c. Which marital status has the lowest average income? d. Which marital status has the greatest variability in income? 6. Scatter Plots a. Create a histogram for income. b. Create a histogram for standardized score. c. 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*)	12

	d. What conclusions, if any, can you draw from the chart? e. Calculate a correlation coefficient between these two variables. What conclusion you draw from it?	
3	Inference 1. Normality a. Create a QQ Normal plot of the Political Awareness Test Score. b. Conduct a statistical test for normality on the Political Awareness Test Score. c. Are the Political Awareness Test Scores normally distributed? What led you to this conclusion? 2. Statistically Significant Differences a. Compare Political Awareness Test Scores between the treatment and control group using a suitable hypothesis test. b. Explain why you chose the test you did. c. Do you have strong evidence that the average test scores are different between the treatment and control groups? 3. Multiple Statistical Differences a. Determine if the Score on the Political Awareness Test varies by Region using ANOVA (statistical) and a sequence of boxplots (graphical). b. Determine if the Measure of Political Involvement (Pol) varies by Political Affiliation using ANOVA and a sequence of boxplots.	10
4	Professionalism and Clarity	2

APPENDIX ONE: STUDY FILE DATA

Variable	**Description**
id	UserID (unique to each respondent)
group	Treatment or Control group
hs.grad	Graduated High School (Y or N)
nation	Nationality (Region)
gender	M/F
age	Age in Years
m.status	Marital Status
Political	Political Affiliation
n.child	Number of Children
Income	Annual Household Income
Food	Pct of Income to Food
Housing	Pct of Income to Housing
Other	Pct of Income to Other Expenses
Score	Score on Political Awareness Test
Scr	Standardized Score Test
time1	Time Taken on Test
time2	Time Taken on Section 1 (Standardized)
time3	Time Taken on Section 2 (Standardized)
Pol	Measure of Political Involvement

APPENDIX TWO: EXAMPLE OUTPUT – *Numbers will be different for the Study File*

Question 1.1

	Martial_Status	Income
1	divorced	
2	married	
3	never	
4	widowed	

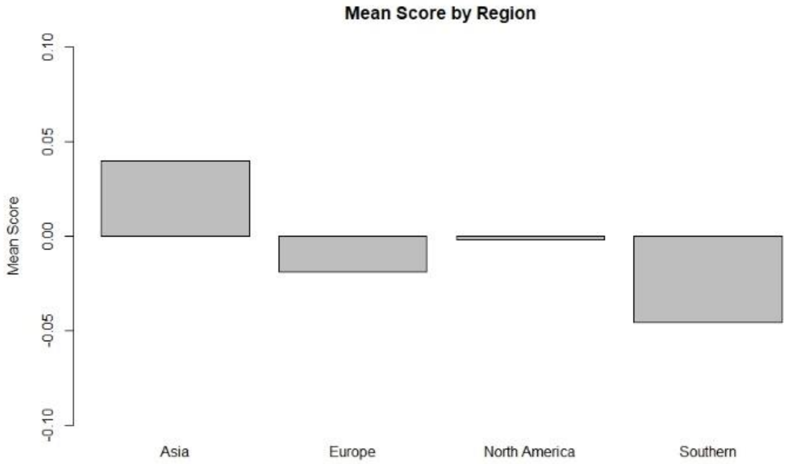
Question 1.3

	HighSchool	Score
1	no	
2	yes	

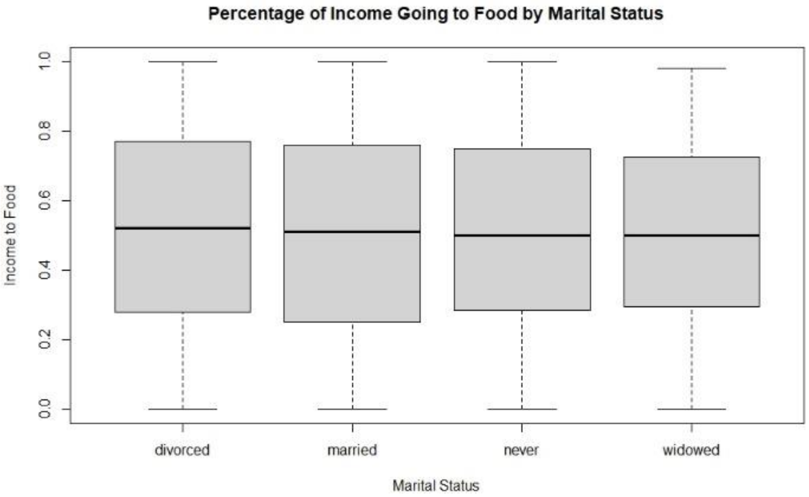
Question 2.2

	High_School	
Region	no	yes
Asia		
Europe		
North America		
Southern		

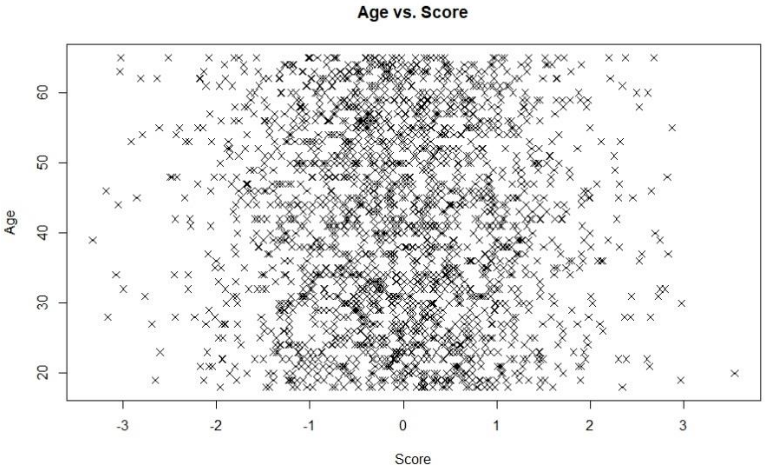
Question 2.3



Question 2.5



Question 2.6


Age vs. Score
Age
Score