PROG8435 – Data Analysis, Modeling and Algorithms

Assignment 2

Statistical Inference and Comparisons

DUE BEFORE FEB 18; 11:59PM

1. Submission Guidelines

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

SUBMISSIONS

In the Assignment 2 Folder submit:

1. Your *R file. This file must have all your comments and code to the questions. I will *not* be running your code to generate it. Imay, however run the code to verify the results.
2. The *.pdf or *.doc file that is produced from your code.

DO NOT PUT THE DOCUMENTS IN TO A ZIP FILE!

PLEASE NOTE: The marks on the assignment are generally awarded 50% for the actual R code and calculations and 50% for interpretation and demonstration that you understand what you have done.

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.

THIS IS AN INDIVIDUAL ASSIGNMENT. UNAUTHORIZED COLLABORATION IS AN ACADEMIC OFFENSE AS IS DIRECT 'CUTTING AND PASTING' FROM OTHER SOURCES. Please see the Conestoga College Academic Integrity Policy for details. ALL REFERENCES MUST BE SPECIFIED.

Remember the discussion forums on eConestoga are a great place to ask questions.

2. Grading

This assignment will be marked out of 40 and is worth 10% of your total grade in the course.

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:

PROG8435_Assign02_24W.txt

Appendix one contains a data dictionary for the study file.

4. Background

In this assignment you will be working with synthesized data based on a dataset created to help the estimation of obesity levels in individuals based on their eating habits and physical condition[1].

The following tasks will seek to describe and explore some of the data which has been gathered. Each row represents one individual. Appendix 1 contains the data dictionary for the data set.

Examples of all of the tasks have been completed in class so a careful review of your notes from the lectures should give you everything you need to complete these tasks.

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

5. Assignment Tasks

Nbr	Description	Marks
1	Data Transformation and Preparation 1. Initial Transformation a. Rename all variables with your initials appended (just as was done in Assignment 1) b. Transform character variables to factor variables.	4
	2. Reduce Dimensionality a. Drop any variables that do not contribute any useful analytical information at all. b. Apply the Missing Value Filter to remove appropriate columns of data.	10

[1] Estimation of obesity levels based on eating habits and physical condition . (2019). UCI Machine Learning Repository. https://doi.org/10.24432/C5H31Z.

	c. Apply the Low Variance Filter to remove appropriate columns of data. d. Apply the High Correlation Filter to remove appropriate columns of data. e. Based on our discussions in class, what are some specific benefits of reducing the dimensionality of *this particular dataset*? Be specific. For example, if it increases computational efficiency, specify *how much of an improvement*. 3. Outliers a. Use an appropriate technique (or techniques) demonstrated in class to identify outliers. b. Comment on any outliers you see and deal with them appropriately. Make sure you explain ***why*** you dealt with them the way you decided to.	4
2	Organizing Data 1. Scatter Plots a. Create a histogram for Height. b. Create a histogram for Weight. c. Create a scatter plot showing the relationship between SMBR and SMBT. (*note: SMBR should be on the x-axis, SMBT should be the y-axis*) d. What conclusions, if any, can you draw from the chart? e. Calculate a correlation coefficient between these two variables. Why did you choose the correlation coefficient you did? What conclusion you draw from it?	6
3	Inference 1. Normality a. Create a QQ Normal plot of for Red Blood Cell Count. b. Conduct a statistical test for normality on Red Blood Cell Count. c. Is Red Blood Cell Count normally distributed? What led you to this conclusion? 2. Statistically Significant Differences a. Compare Red Blood Cell count between Genders in your dataset using a suitable hypothesis test. b. Explain why you chose the test you did. c. Do you have strong evidence that Red Blood Cell count is different between genders? 3. Multiple Statistical Differences a. Determine if Weight varies by method of transportation using ANOVA (statistical) and a sequence of boxplots (graphical). b. Determine if red blood count varies by method of transportation using ANOVA and a sequence of boxplots.	4 5 4
4	Professionalism and Clarity (Format, spelling, etc)	3

APPENDIX ONE: STUDY FILE DATA

Variable	Description
Index	Counter of Data
Gender	Gender
Age	Age (in years)
Height	Height (in metres)
Hgt	Height (in feet)
Weight	Weight (in Kg)
City	City of Residence
family_history_with_overweight	Has a family member suffered or suffers from overweight?
FAVC	Do you eat high caloric food frequently?
FCVC	Do you usually eat vegetables in your meals?
NCP	How many main meals do you have daily?
CAEC	Do you eat any food between meals?
SMOKE	Do you smoke?
CH2O	How much water do you drink daily?
FAF	How often do you have physical activity?
TUE	How much time do you use technological devices such as cell phone, videogames, television, computer and others?
CALC	How often do you drink alcohol?
MTRANS	Which transportation do you usually use?
NObeyesdad	Obesity level
RBC	Red Blood Cell count (million cells per microliter (mcL))
WBC	White Blood Cell count (Cells per mm3)