Assignment 1: Data Cleaning and Summarising

Submit Assignment

Due Sunday by 23:59 **Points** 15 **Submitting** a file upload **File types** zip **Available** 19 Mar at 12:30 - 3 May at 23:59 about 2 months

Course Name: Practical Data Science
Course Code: COSC2670

Assignment 1: Data Cleaning and Summarising

Due: 23:59, 19th/April, 2020 (extended deadline)

This assignment is worth 15% of your overall mark.

- The assignment specification is available here, including
  - The data set.
  - The assignment specification document.
  - The template to develop your solution.

rubric for assignment 1- updated

Criteria	Ratings	Pts
Task 1.1: Data Retrieving Point 1: Load the CSV data from the file. You need to use an appropriate pandas function to load the csv data, and make use of the correct arguments including sep, decimal, header, names, if needed.		0.5 pts
Task 1.2: Check data types Point 2: Check whether the loaded data is equivalent to the data in the source (CSV) file. That is, you will need to ensure that the loaded data has appropriate data types assigned, or take steps to ensure that the appropriate types are used.		0.5 pts
Task 1.3: Typos Point 3: Check whether there are typos in the data. If there are any typos, correct them by using masks.		0.5 pts
Task 1.4: Extra-whitespaces Point 4: Check whether there are instances of extra whitespaces in the data, and if so, demonstrate how to remove them by calling on an appropriate function.		0.5 pts
Task 1.5: Upper/Lower-case Point 5: Cast all text data to upper-case by using an appropriate function.		0.5 pts
Task 1.6: Sanity checks Point 6: Design and run a small test-suite, consisting of a series of sanity checks to test for the presence of impossible values for each attribute.		1.0 pts
Task 1.7: Missing values Point 7: Check whether the loaded data has any missing values. If so, use an appropriate function to replace them with one of the following values: - a fixed value - the column-wise median value - the column-wise mean value - or ignoring all observations containing missing values.		1.5 pts
Task 2.1: Explore a survey question Explore the survey question: [Please rank the Star Wars films in order of preference with 1 being your favorite film in the franchise and 6 being your least favorite film. (Star Wars: Episode I The Phantom Menace; Star Wars: Episode II Attack of the Clones; Star Wars: Episode III Revenge of the Sith; Star Wars: Episode IV A New Hope; Star Wars: Episode V The Empire Strikes Back; Star Wars: Episode VI Return of the Jedi)], then analysis how people rate Star Wars Movies.		1.5 pts
Task 2.2: Relationships between columns Explore the relationships between columns. You may choose which pairs of columns to focus on, but you need to generate 3 visualisations for this subtask. These should address a plausible hypothesis for the data concerned. Please also format the graph as required in Task 2.1.		3.0 pts
Task 2.3: Explore a specific relationship Explore whether there are relationship between people's demographics (Gender, Age, Household Income, Education, Location) and their attitude to Start War characters.		0.5 pts
Task 3.1: Data Preparation Create a heading called "Data Preparation" in your report. For step 2 to 7 in Task 1 above, create a sub-section with corresponding numbering, and provide a brief explanation of how you addressed the task, and explain any choices that you made (if appropriate).		1.5 pts
Task 3.2.1: Data Exploration Create a heading called "Data Exploration" in your report. For each numbered step in Task 2 above, create a sub-section with corresponding numbering. E.g. in subsection 1, you need to explain how you explore, and why, and what you obtain.		1.5 pts
Task 3.2.2: Data Exploration In subsection 2, include your plots from Task 2, Step 2. With each plot, state the hypothesis that you are investigating. Then, briefly discuss any interesting relationships (or lack of relationships) that you can observe from your visualisation.		1.5 pts
Task 3.2.3: Data Exploration In subsection 3, you need to explain how you explore, and justify why you do that, and what you find.		0.5 pts
	Total points: 15.0	