**DSCI-549: Introduction to**

**USC** Viterbi School of Engineering

**Computational Thinking and**

**Data Science**

**Term: Fall 2021**

**Project – Part 1**

**Date due: FRIDAY, November 5th 2021 - 5pm PT**

**Goals**

The purpose of the project is to familiarize yourself with the steps involved in most data science projects, at least in the conceptualization phase.

**Overview**

Choose a data science project of interest and answer the questions in the detailed assignment section below. Answer each question individually. **DO NOT** submit an essay or report-like assignment (you will be marked down).

Remember that **ALL ANSWERS MUST BE JUSTIFIED and MUST APPLY TO YOUR DATASET.** e.g., when describing the algorithm, ensure that you explain which columns/image… in the dataset the analysis will be applied to. Even though you are not writing code, you are expected to provide a close recipe.

You are allowed to work in groups. **HOWEVER,** this is an individual assignment, and your data science question/dataset/analysis must be unique. See the Course Overview slides for examples.

This is the first part of the assignment. You are allowed to change you question/dataset for the second part; in which case you will have to redo this assignment for context. Note that it will not be regraded.

**Assignment**

Part 1: Data (50 points)

1. Define an interesting problem that could be answered through data analysis. Describe the problem to a non-informed audience (10 points)
2. Find one or more datasets on the Web that are relevant to these questions, and that are: 1) accessible, 2) released with an open license, and 3) machine processable. For one of the datasets, specify:
   1. A brief description of what the data represents, with a small excerpt of the data as an example (10 points)
   2. A discussion on how the dataset satisfies the three criteria above (20 points)
   3. The **details** for how you can access the data (10 points)

**Use** the technical terms that you have learned in class to describe data.

Part 2: Hypothesis (30 points)

1. Is the study you want to conduct experimental or observational? Why? How does that affect your conclusions? (10 points)
2. State the null and alternate hypothesis (10 points)
3. What are the dependent and independent variables? (10 points)

Part 3: Data exploration (50 points)

1. List and describe three data exploration steps that you will likely need. (note: visualization is acceptable as three separate steps as long as they allow you to explore three different characteristics of the dataset) (30 points)
2. For each of the steps, describe how it will help you with pre-processing and analysis. (i.e., if my visualization indicates outliers, then I can remove them). (20 points)

Part 4: Pre-processing (50 points)

1. List and describe three processing steps that you think are involved in the data science pipeline. (30 points)
2. For each of the steps, what do you expect to achieve and how does it help your analysis? (20 points)

Part 5: Analysis (50 points)

1. In your own words, give an overview of the analysis you would need to perform to answer your data science question. (20 points)
2. Can you think of an algorithm learned in class that would be best to test your hypothesis? (30 points)
   1. If so, which one? Describe the algorithm
   2. If not, write down the algorithm

Part 6: Workflow (20 points)

1. Sketch the workflow for your data science problem, omitting the data exploration steps but including pre-processing and analysis. The workflow needs to include the data flow. The analysis step should be one component.

***IMPORTANT NOTES***

*Plagiarism – presenting someone else’s ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences.  Please familiarize yourself with the discussion of plagiarism in SCampus in Section 11, Behavior Violating University Standards* [*https://scampus.usc.edu/1100-behavior-violating-university-standards-and-appropriate-sanctions*](https://scampus.usc.edu/1100-behavior-violating-university-standards-and-appropriate-sanctions/)*.  Other forms of academic dishonesty are equally unacceptable.  See additional information in SCampus and university policies on scientific misconduct,* [*http://policy.usc.edu/scientific-misconduct*](http://policy.usc.edu/scientific-misconduct/)*.*

*A number of USC’s schools provide support for students who need help with scholarly writing.  Check with your advisor or program staff to find out more.  Students whose primary language is not English should check with the American Language Institute* [*http://dornsife.usc.edu/ali*](http://dornsife.usc.edu/ali)*, which sponsors courses and workshops specifically for international graduate students.*

*For more information, see the class syllabus and the USC web site.*