COSC-587
Fall 2021
Project Assignment 1
Due on September 24th at 5 pm

PROCEDURES AND LATE POLICY REMINDER

- **Turn-in:** Please turn in your work in the github classroom. Once your groups are determined, you will be given access to the room. All written components of your assignment should be in pdf format. I encourage latex, but it is not a requirement. All code should be in python 3.
- **Deadline**: The on-time deadline for all students is 5 pm on the due date.
- **Late policy**: All written work is to be turned in at 5 pm on the day that it is due. Written work turned in after the deadline will be accepted but penalized 50% per day. Once an assignment has been returned, a late assignment will not be accepted.

Overview

This project asks you to identify a data science problem of interest to the entire group and gather the data necessary to conduct a data science analysis in subsequent project assignments. The data science problem can be descriptive, predictive (or both), but keep in mind that through the course of the semester, you will work on both types of analyzes. You can change the specifics of the data science question that you ask, but you need to have a general direction that you plan to explore to ensure that you collect data that will be reasonable. You will complete this project in groups of four.

Data Science Problem (5%)

Explain the problem you plan to investigate. Provide sufficient context and background information about why this problem is meaningful or adds insight. Have a citation or two to give context to the problem – why is it meaningful to study? What is different between what you are doing and what has been done before? Your problem must be a scientific study. You cannot use IMDB data or sports data without special permission.

Potential Analyzes that Can Be Conducted Using Collected Data (10%)

You should first briefly describe the data you plan to collect and why these data are meaningful for your data science problem. You should have data from three different sources. What will your data set contribute to answering your question? What are the variables and why are they useful? Then write a brief explanation of possible directions / hypotheses that you may be able to investigate with the data you collected. Ideas here may not end up being your final question. At this stage, you are generating possible directions.

Collecting New Data (55%)

Your main task is to collect data for your analysis. You need to write automated scripts to collect three different data sets that you can combine in future projects, e.g. newspaper data and climate data. One of the data sets must include text that will be useful for text analysis in later projects. You can choose to collect more than three data sets. You must use python (or an approved language) to collect them. One of your datasets can be a simple download that does not involve an API or scraping. For two data sets, you should write a python script that uses an API to collect data or you can scrape data from different web pages. Between the three data sets, you should have **at least** 15 attributes that are different. You may not have less. Of course, it is expected that the data may contain noise or missing values. You must have at least 20,000 records of data from each data set, but it is fine if some of the attributes are null. If you have an interesting problem that has less data, please come and talk to me. I may let you use it.

Data Issues (5%)

For this part, please explain the issues that you see with the data, e.g. noise, missing values, etc. Make a detailed list of the different issues for each variable.

Data Cleanliness (25%):

Some of you will download data that is fairly clean. Others will not. In either case, you should have a program that checks the level of cleanliness of your data. You should develop a program that looks at your attributes and quantifies how ‘clean’ the attribute is. Specifically, you should identify missing and incorrect values. You can then record:

- The fraction of missing values for each attribute.
- The fraction of noise values, e.g. gender = ‘fruit’.

Can you use this information to generate a data quality score? Based on this data quality metric, how clean is your data? Your code should be easily adaptable, putting constraints in an input file, no hard-coding, etc.

A few final notes:

- All your code should be well commented with reasonable variables names, etc.
- We run all the code you write, so make sure it works. You will get deductions if it does not run properly.
- You will code in a github classroom – your group will have a private repo.
- The directory should have a README.txt file that explains what each of the files are.
- There will be a peer evaluation conducted after you submit the project.