COSC-587 & ANLY-501
Fall 2018
Project Assignment 1
Due on Oct 5th 11:59 pm

PROCEDURES AND LATE POLICY REMINDER

- **Turn-in:** Please turn in your work via Canvas. Please put your names on your assignment. We will setup Blackboard groups for you so that you can submit as a group.
- **Deadline**: The on-time deadline for all students is 11:59 pm on the due date.
- **Late policy**: All written work is to be turned in at 11:59 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, gather the data necessary to conduct a data science analysis in subsequent project assignments, and assess the quality of the data you gathered. 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. At this stage, you do not have to know the precise data science question you plan to ask. However, 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 (there will be one group of three).

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?

Potential Analyzes that Can Be Conducted Using Collected Data (5%)
You should first briefly describe the data you plan to collect and why these data are meaningful for your data science problem. 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.

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 so that you will be able to clean the data accordingly.

Collecting New Data (45%)
Your main task is to collect data for your analysis. You need to write automated scripts to collect two different data sets that you can combine in future projects, e.g. Twitter data and stock data. You can choose to collect more than two data sets. You must use python (or an approved language) to collect them. You cannot just download a CSV file for this task. You can write a python 3 script that uses an API to collect data or you can scrape data from different web pages. Between the two data sets, you should have **at least** 12 attributes. You may not have less. Of course, it is expected that the data may contain noise or missing values.

ANAYLTICS: You must have at least 5000 records of data from each data set, but it is fine if some of the attributes are null.

COMPUTER SCIENCE: You need to have three data sets, as opposed to two. 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 Cleanliness (25%):
Some of you will download data that is fairly clean. Others will not. In either case, you should have a script that checks the level of cleanliness of your data. You should develop a script 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?

Note: You should make sure that your code is modular.

COMPUTER SCIENCE: Your code should be easily adaptable, putting constraints in an input file, no hard-coding, etc.

Data Cleaning (15%):
For this part, you should take at least three of your attributes that had a poor data quality/cleanliness score and write a python program that cleans these attributes/variables. For example, if some of your variables have bad values, you may choose to replace them with good values. If some of the values are text and you want numeric values, you may choose to change the data type. After you clean the data, re-run the data cleanliness program. Is your data cleaner? Explain.

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.
- I highly encourage you to use git.hub to share code.
- You will submit a zip file that contains a directory. The directory should have a README.txt file that explains what each of the files are. If you have a large data set, you can give us a link to it on git.hub. The zip electronic version of your project should be submitted through Blackboard.