**ANLY-501 Fall 2018**

**Class - this is now the official project assignment. I left in my notes and clarifications.**
**NOTES: Due Oct 7 by 11:59pm ET (-5% per HOUR if late - the first hour is waived)**

**All Teams will present this project in class on OCT 1. All members of each team must speak equally during the presentation. Time Range: 5 - 6 minutes. ** do not exceed 6 minutes). Focus on relevant information - we will talk about this in class.**

**Class - I have added notes to this project IN RED (and in bold) to help to clarify as many portions as possible. Please read every word in this document carefully :)**

**All code for the Project must be written in Python 3 only. (Python 2 and R will not be accepted). This was noted in class - but may have been missed.**

**Python 3 must be used for all class code and for all submissions, assignments, and projects in this class.**

**Project Assignment 1**

**Due Oct 7 by 11:59pm ET (-5% per HOUR if late - the first hour is waived)**

**PROCEDURES AND LATE POLICY REMINDER**

- **Turn-in:** Please turn in your work via Blackboard. 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

**which major is matched up with highest returned job**

api:
Linkedin
Glassdoor
Monster
payscale
USAjobs

1) This project asks you to identify a data science problem of interest to the entire group
2) Gather the data necessary to conduct a data science analysis in subsequent project assignments, and assess (determine using measures and calculations and code) the quality of the data you gathered.
3) The data science problem can be a descriptive or a predictive one (or both), but keep in mind that through the course of this project, you will work on both types of analyzes. In other words, all analysis methods will be included in your project. Think about this when you choose your data.

4) 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.
5) NOTES: You will "identify" a data science "problem". You do not yet need to create a set of question that you will ask about the problem you have identified.

# Data Science Problem (5%)

1) Explain the problem you plan to investigate.
2) Provide sufficient context and background information about why this problem is meaningful or adds insight.
3) Have a citation or two that supports your "problem ideas".

RULES: No projects can be on stock market data, real estate data, or movie data. These are over-done and so are not creative. Think about a more creative "outside-the-box" data science question. A data science question is a large idea that can be investigated using 10 - 15 smaller questions that support it.

EXAMPLES of ideas one can start with:

1) Is the climate changing because of us?
2) Will the human population become unsustainable and if so, what should be done?
3) Are guns more dangerous than cars?
4) Should all gas-powered items, such as cars, leaf-blowers, etc be illegal?
5) Are women treated equally in all aspects around the world?
6) What are the key contributing factors to cancer and are things getting better or worse?
7) Should telecommuting be required, at least one day per week, for all companies in the US?
8) Have advancements in technology made our school-age children less attentive and less able or willing to learn new things?
9) Should GMO labeling be required by law?
10) Should religion play a role in government decision making?

**More General Areas**

1) Bio and health
2) Climate
3) Science
4) Finance and Economics
5) Social and public policy
6) Transportation
7) Education
8) Crime and Punishment
9) Politics and government

10) Zoology and Botany
11) Sports

**Remember - a data science question will allow you to use all the analyses and answer at least 10 different supporting questions.**

# Potential Analyzes that Can Be Conducted Using Collected Data (5%)

1) You should first **briefly describe the data you plan to collect** and
2) **why these data are meaningful** for your data science problem.
3) Then write a brief explanation of possible directions / hypotheses that you may be able to investigate with the data you collected.
4) Ideas here may not end up being your final question. At this stage, **you are generating possible directions.**

# Data Issues (5%)

1) For this part, please explain the issues that you see with the data, e.g. noise, missing values, etc.
2) List all issues that you see - so that when you are ready to conduct your descriptive analysis **in the next assignment**, you will be able to clean the data accordingly.
3) You do not need to clean the data yet.

# Collecting New Data (45%)

1) **Your main task is to collect data for your analysis**.
2) 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.
3) You can choose to collect more than two data sets.
4) You **must use python** to collect all data - **YOU CANNOT CLICK to get data - all data at this point and for this project Part 1 MUST be retrieved via Python.**
5) Between the two data sets, you should have at **least 12 attributes (12 variables or columns**.
6) Yes - you can have more. No you cannot have less.
7) Of course, it is expected that the data may contain noise or missing values.

UPDATE: I will allow the average number of rows in both datasets to be 5000. Yes - this means that one can have 3000 and the other can have 7000 so that the mean is 5000. I will accept no further questions on this requirement. Please note that in future projects, it might be required

that you have two datasets (both at least 5000 rows). If this is the case, you will have to get and clean more data. I again recommend that you follow the instructions. However, for those of you set on two dataset of different sizes - that is fine.

8) **You must have at least 5000 records (rows) of data in each dataset you collect**, but it is fine if some of the attributes are null.
9) If you have an interesting problem that has less data, please come and talk to me. I may let you use it. - NO LONGER TRUE AT THIS POINT

**Data Cleaning (30%):**

1) Some of you will download data that is fairly clean. Others will not.
2) In either case, **you should have a script that checks the level of cleanliness of your data**.
3) You should develop a script that **looks at your attributes, and "quantify" how 'clean' the attribute is**.
4) Specifically, you should identify missing and incorrect values.
5) You can then record:

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

**6)** Use this information to generate a data quality score? SHOW ALL WORK and how you got your quality score and what it means.
7) Based on this score, how clean is your data?

**NOTES: You do not yet have to actually clean the data. Your scripts should "investigate" the data and determine its issues. Pretend that you cannot see all the data but that you need to if it is dirty, how dirty, what is making it dirty, etc. Create generalized scripts that will work with most data - rather than hardcoded just for your data. Yes there will be cases where your cleaning scripts will be just for specifics in your data - but not always - try to think about both cases.**
**In the next project (part 2) you will be cleaning your data. Not for this project.**

**Feature Generation (10%):**

1) For this part, you should construct **3 new features from the data** you have downloaded.
2) Write a script that takes in some of the original data and then constructs a new feature from that data.
For example, you may choose to bin continuous data or create binary variables out of a set of attribute values.

*A few final notes:*

- All your code should be well commented with reasonable variables names, etc.
- I highly encourage you to use git.hub or Google to share code - with each other. This is optional. When you submit you will use BlackBoard.
- You will submit an electronic version of your project through Blackboard.

Here is what you will submit to BlackBoard:

1) A **Project Word Document** that answers all questions asked in this assignment (except the code portions)
2) A Python script that gets your two datasets and stores them into csv files - separately for now so it's easy to see that you have two datasets.
    - In the same Python script, open the two datasets you have scraped (using an API or directly) and include all required Python code for investigating the cleanliness of the data.
    - In the **Project Word Document** include explanations of the code you created to find dirt in the data. In other words, did you write code for missing value detection? Did you include code for incorrect values (and if so - what were the ideas behind it - did you check ranges or did you check values, etc. ?)

3) You will also submit the two (uncleaned ) csv files that you have the data you got. These WILL NOT BE CLEANED YET. We will clean the data in the next project.

4) Note that the **Project Word Document** will also show your "data cleanliness metric", how you arrived at it, and an example of how it works. To explain how it works, choose one attribute and show and simple before and after with say "missing values" only to see how the metric will work in that case.

5) Note that this project requires code for several things. It is best to place all the code into one Python Program (as functions).

6) Do not "hardcode" things - be sure that if items need to be read - in that they are read in from files. You are welcome to prints things to the console, but also write results to files. For example, when checking for missing values, write the results to a MissingValues.txt (or whatever) file.

7) Feel free to also submit a README.txt file that briefly notes the items included in your submissions and how to run them.