CSci 343 Foundations of Data Science
Class Project – Spring 2019

Due date: Sunday, May 5 at 11:59pm

The project counts for 10% of your overall grade. You will choose a data set from Kaggle to analyze. No two students will be allowed to use the same data set, so follow the instructions below to select your dataset.

Requirements:

- Write a 2-3 page summary of your analysis. It should include:
  - A basic description of the data set and its features
  - Simple statistics about the data set and features
  - EITHER some machine learning results for prediction or regression OR some sort of text processing (such as sentiment analysis, topic identification, etc) results.
  - Future (other) analysis that could be done with the dataset if you had more time.
  - At least 2 graphs must be used in your summary.
  - A list of web sites or references that you used.

It is perfectly fine to find references showing how to perform certain tasks, but don't just follow a tutorial or recreate someone else's analysis for that data set.

What to turn in: A single file (zip okay) that contains your summary (all items listed above) and all code you wrote to do the analysis. If you use Excel, for example, to compute the basic statistics, just indicate that in the summary (no need to upload the sheet or your data).

How to select your data set:

1. Go to kaggle.com and sign in (or sign up) so you can download datasets. Look through the data sets with an eye to what kind of ML or Text Analysis that you could do with that data. You can do keyword searches if you have some domain that you are interested in. Look at the attributes available and consider what you could do to analyze the data. A lot of the data sets are of images. There are many good tutorials on how to work with images to do ML if you want to work with one of those.
2. Select at least 3-4 data sets that you are interested in exploring.
3. Email me (dwikins@cs.olemiss.edu) an **ordered list** of the data sets you selected. Please cut/paste the URL from kaggle. For example: https://www.kaggle.com/jameslko/gun-violence-data. Also include your preference order on which day M 4/22, W 4/24 or F 4/26 that you would like to describe your data and your plan for analysis (3 minutes).
4. Data sets will be assigned first-come-first-served. So start exploring the data sets early to get your choice!! I'll reply back ASAP to let you know that you can proceed (and which of your top choices is acceptable). Every now and then I'll post/update a list on Blackboard so you can see which data sets are already taken.
5. **You MUST have an approved data set by Monday, April 22nd by 8:00 AM.**