CS 5010 – Project

Important dates:

- **SUBMISSION**: Code and write-up (**all groups**) due on Wednesday, **July 31 by 11:30pm**
- **PRESENTATIONS**: **Wednesday, July 31, 2019** and **Friday, August 2, 2019**

Grading: ***100 points*** *total*

- Grading rubric will be provided and posted soon on Collab Resources

Submission:

- In one Zip file, submit:
  - Code
  - Write-up
  - Your presentation visuals (e.g. PowerPoint presentation or otherwise, if used)
- Submission location: under Assignments on Collab

Project Specifications:

- Choose a data set to work with. Pick a data set that you and your group finds interesting (Example source: UC Irvine Machine Learning Repository -http://archive.ics.uci.edu/ml/, feel free to select your data from any other source as appropriate)
- Perform data pre-processing, data cleaning, outlier removal, etc to sanitize your data, if necessary
- Save your data in a .csv file (or other format as appropriate for your data set and project scenario)
- Read in data to your program from the .csv file
- (*Optional* - do as appropriate) Process the data or perform any calculations or statistics on it before storing the data into a data frame (see next step)
- Save the data into one or more data frame data structure (or other structure as appropriate) (*review lecture notes/examples*)
- Once your data is stored query your data to reveal interesting/useful information based on your project scenario
  - Query your data using **at least 4 different queries**
- Capture the results of the queries appropriately (either write results to a file, or store into another data structure, or do something else with the results as appropriate based on your project scenario)
- (*Do as appropriate*) Process the results, or submit additional queries to the obtained results (if results were saved to a file or another data structure)
- Display final results in a presentable way (use tables and/or appropriate ***visualizations***)
- Perform adequate testing (TDD and/or Unit testing) Submit this in a separate .py file

- Submit a write-up along with your project (see details in the next section below), in your write-up be sure to display the results and describe what you have learned, as well as how these results can be used by others (mention the relevance/significance of the results you've obtained)
- **Group size**: 4. *No individuals or pairs permitted.*

Write-up Details:

Be sure to include the following in your write-up:

- **INTRODUCTION**: describe your project scenario. Starting out, what did you hope to accomplish/learn?
- **THE DATA**: Describe your data set and its significance. Where did you obtain this data set from? Why did you choose the data set that you did? Indicate if you carried out any pre-processing/data cleaning/outlier removal, etc… to sanitize your data
- **BEYOND THE ORIGINAL SPECIFICATIONS**: highlight clearly what things you did that went beyond the original specifications. That is, discuss what you implemented that would count towards the ***extra-credit*** portion of this project (see section below.)
- **RESULTS**: Display and discuss the results. Describe what you have learned and mention the relevance/significance of the results you have obtained. Be sure to use appropriate visualizations to display your results.
- **TESTING**: Describe what testing you did. Describe the Unit tests that you wrote. Show a sample run of one or two of your tests (screen caps or copy-and-paste is fine)
- **CONCLUSIONS**: Summarize your findings, explain how these results could be used by others (if applicable), describe ways you could improve your program or ways you might like to expand the functionality of your program if given more time

Extra Credit Opportunities For Those Students Who Want A Challenge:

- Web-scraping to obtain your data set instead of downloading a ready-made data set from a source (*see **Homework #3*** ☺)
- Have some user-interaction where you may obtain some more data from a user (if appropriate)
- Have some user-interaction where the user may choose the kinds of queries to perform on the data. Retrieve/display only the appropriate result
- Use advanced queries or manipulate the data in another way (other data manipulation methods, etc..) and display the results. If you choose to do this, mention in your write-up how this goes beyond the basic/general queries you initially used
- **IF YOU THINK OF SOMETHING ELSE YOU WOULD LIKE TO DO, AND FEEL IT COULD BE CONSIDERED AS EXTRA-CREDIT, PLEASE LET ME KNOW.**