Individual Project STA5373 (Alexander Mantzaris, alexander.mantzaris@ucf.edu)

In this assignment you will conduct work individually. The scope is to go through the major steps of studying a dataset to produce 'insight' for data driven decision making. There are 2 parts of this assignment, the data analytics and investigation with a follow up presentation of the results. The results should be backed up by code which is available on a public repository such as github or Kaggle.

*please try to select a topic which will not make the audience uncomfortable

Part I (10 points)

- Find a dataset of your choice to analyze. You can search through sites such as 'www.kaggle.com', 'data.gov', or even through Google's data research tool 'https://datasetsearch.research.google.com'. There are other sites as well you are free to search and there are also forums where as a member you can request community support to find a dataset you have in mind which cannot be found via searches such as 'https://opendata.stackexchange.com/'.
- Analyze the dataset with 2-4 different statistical models/methods/tools. These can be linear regression, logistic regression, decision trees, XGBoost, clustering, random forests etc. Use these tools to make a set of at least 10 'predictions' on data not used in the 'training' process. You will be reporting on the accuracy/ROC you find using each methodology or on the different MSE based on model choices.
- Produce plots and print outs of the results to describe your findings in your presentation.
- The dataset should have at least 200 data points to it, and be less than 10 years old since it was produced. (eg stock data since 2014 and not before)
- The dataset should also have at least 2 independent variables for the dependent variable.
- Look at reducing the model and report how this was done.

Part II (10 points)

You will produce a presentation (slides) which will be given in front of your class that should last approximately 5 minutes. The angle of the presentation is to showcase your analytics findings and not use the analytics to drive the 'decision making' by providing opinions and interpretations. The structure of the presentation will be as follows:

- Title slide
- Question slide; what question does the investigation seek to answer? Eg. 'Can this dataset X be used to train a model to predict Y?'
- Data section should be around 2-3 slides describing the data. Having an image screen shot of the data is ok and you should be 'listing' the variables contained in the data stating which you choose to present, which you studied and why you made those choices.
- Methodology, present an overview of the methods used and why you chose them. Why they are suitable for this task and dataset. Also discuss the programming language selected the libraries used.

- Results, present your results in the form of plot outputs and printed numerical outputs. The figures should also have a caption to each one for the audience to clearly understand the axis meanings.
- Discussion/Conclusion, discuss whether your investigation can be used to reliably predict Y or not. Not being able to is also a valid answer if the investigation is thorough.

[Submit the slides and the code]