**Project Assignment 2 - Fall 2021**
**Due: October 20 at 9 pm**

**PROCEDURES AND LATE POLICY REMINDER**

- **Turn-in:** Please turn in your work in the github classroom. Once your groups are determined, you will be given access to the room. All written components of your assignment should be in pdf format. I encourage latex, but it is not a requirement. All code should be in python 3.
- **Deadline**: The on-time deadline for all students is 9 pm on the due date.
- **Late policy**: All written work is to be turned in at 9 pm on the day that it is due. Written work turned in after the deadline will be accepted but penalized 50% per day.

**Overview**

This project asks you to clean and conduct a descriptive analysis using the data you collected in Project 1 You will explore it, and will begin developing support for your data science questions. Five analyses requirements are listed below. In addition to conducting each analysis, you should also explain what each analysis means (provide an interpretation). The interpretation will be in the Project report. All the analyses must be conducted in Python3.

Code examples for different parts are part of your class notes/slides, etc. Please refer to those to get additional help on this project.

**Data Cleaning (15%):**

- For this part, you should make sure that all the variables you are using in your descriptive analysis are clean (handle missing values and duplicate). Use python to clean them. 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. Explain your cleaning decisions.

**Basic Statistical Analysis and data cleaning insight (20%)**

- Determine the mean (mode if categorical), median, and standard deviation of at least 10 attributes in your data sets. Use Python to generate these results and use the project report to show and explain each.

- Use a statistical, LOF (try 3 values for k), or more robust unsupervised outlier/anomaly detection method to identify outliers in your attributes.
    - Construct new variables that do not contain the outliers if necessary.
    - Explain how you detected the outliers, and how you made the decision to keep or remove them.
    - If you find that your data needs to be further cleaned or differently cleaned based on analyses, include explanations here. Be specific about what you did and why.

- For at least one of the numeric variables in one of the datasets, write code to bin the data. This will create a new column. Use the binning strategy that is most intuitive for your data. Explain your decision. Include why you chose to bin the specific attribute selected, the binning method used, and why that method makes sense for your data.

**Histograms and Correlations (15%)**

- Use a histogram to plot at least three (3) of the variables (attributes) in either dataset. Discuss the insight generated by the histograms. What do they show or suggest?
- Identify three (3) quantitative variables from either data set. Find the correlation between all the pairs of these quantity variables. Include a table of the output in your report, and

explain your findings – what does this indicate about your data? Use scatterplots to display the results. Ideally, create a set of scatterplot subplots.

**Cluster Analysis (30%)**

- Conduct four (4) cluster analyses on your data. Include a hierarchical clustering method (such as Ward), a partition clustering method (use k-means), a density method (dbscan), and a statistical method (GMM). Explain your findings.
- Use the Silhouette or Calinski-Harabaz procedure to assess the quality of the clusters. Ref: http://scikit-learn.org/stable/modules/clustering.html
- Plot the clusters or if the dimensionality is too high, plot a PCA projection of the clusters. Does the plot give you additional insight about the clustering – explain.

REF: http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
REF: http://scikit-learn.org/stable/auto_examples/decomposition/plot_pca_iris.html

**Sentiment Analysis or Topic Modeling (15%)**

Below are **options** for different potential analyses. **You must choose one.**

- Sentiment analysis – If you have textual data, you may be interested in determining how positive or negative the data is. Make sure you explain what you are finding sentiment for and what your sentiment results mean. You should also show the accuracy of your sentiment analysis on some manually labeled data as well. How can you meaningfully summarize the findings?
- Topic modeling – Topic modeling allows you to looks at your different text features and see (1) what topics are prevalent in your data set, (2) the distribution of topics for each "document" and (3) what text is similar to other text based on similar topics. You can present your results using text or visualizations. Make sure you explain what you are finding topics for and what the results mean. You can use different algorithms, but LDA is sufficient.

**Writeup (10%)**

You should begin to have an overall story coming together. Please integrate this with Project 1 to begin to pull together the overall story.

*A Note About Your Data:*

You should not add more data without discussing it with me. Data collection was the goal of Project 1.

*A few final notes:*

- All your python code should be well commented, well-structured and easy to read and understand visually; with reasonable variables names, well organized functions, sufficient comments, etc.
- All code must run. Once you submit, download your submission, and re-run it. This will assure that your submission was successful and what you intended it to be.
- Your final submission should contain: (1) All Code, (2) The Project Report Document that address and discusses all noted requirements and elements (this should be a pdf file), (3) Any needed files and/data that your code will read from, (4) a README.txt to explain basic code use if you feel this is needed.

**Extra Credit** – This project is eligible for up to 3% extra credit.