UST Fall 2025
Data Science and Algorithmic Fairness
Homework Clustering
Submit via GitHub commit/push
Due Monday Nov. 17th 11:59 PM

Example and API can be found here:
https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_iris.html#sphx-glr-auto-examples-cluster-plot-cluster-iris-py
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMeans

Other resources can be found here:
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
https://github.com/scikit-learn/scikit-learn/blob/1495f6924/sklearn/cluster/k_means_.py#L772
https://stackoverflow.com/questions/41540751/sklearn-kmeans-equivalent-of-elbow-method

1. Use a dataset of the City of Houston that has information about traffic counts in different sections of the city. Can be downloaded from here: https://github.com/computingcelts/f25-ds-examples/blob/main/datasets/TrafficCounts_OpenData_wm.csv
   The metadata information can be found here: https://bit.ly/tfcounts Not all the metadata will be useful for this analysis, you can make decisions about what will be useful and which not.
2. Create a jupyter notebook by using as a template the one provided above for K-Means
3. The purpose is to do an exploratory analysis and clustering of traffic patterns, specifically areas with major traffic and those with minor, or anything in between.
4. If you think that there's a need to create new features, do some feature engineering and create new ones, be creative, and explain what did you create.
5. Explore the dataset and provide some descriptive statistics, histograms.
6. Create a K-Means with a cluster of 2 and plot the clusters
7. Create a K-Means with a cluster of 3 and plot the clusters
8. Create a loop to compute K-Means with clusters from 3 to 10
9. For each loop, compute the Sum of square distances
10. Plot an elbow graph
11. For each run indicate what are the characteristics of the centroids, what do they tell you about the dataset.