**ATOC5860 - Homework 6 - due April 21, 2022**

**Please send your homework to Jen/Prof. Kay on Slack as a direct message.**

***Please Name Your Homework Files: “ATOC5860\_HW6\_LastName.pdf, .html, .ipynb”***

**Your submissions should include: 1) A .pdf document with responses to the questions below, 2) Your code in both .ipynb and .html format.**

**Show all work including the equations used (e.g., by referring to the Barnes Notes).**

**Write in complete, clear, and concise sentences.**

**Eliminate spelling/grammar mistakes.**

**Label all graph axes. Include units.**

**Report values using appropriate rounding.**

**Problem I) K-means clustering of airport weather observations into “seasons” (40 points total)**

**Your data:** Hourly weather observations from the Local Climatological Data (LCD) dataset for 6 airports: 1) Denver, Colorado (DEN), 2) Boston, Mass (BOS), 3) Seattle, Washington (SEA), 4) Minneapolis-St. Paul, Minnesota (MSP), 5) Grand Junction, Colorado (GJT), and 6) San Juan, Puerto Rico (SJU). Your data have been munged to replace missing values.

**More details:** *“Local Climatological Data (LCD) are summaries of climatological conditions from airport and other prominent weather stations managed by NWS, FAA, and DOD. The product includes hourly observations and associated remarks, and a record of hourly precipitation for the entire month. Also included are daily summaries summarizing temperature extremes, degree days, precipitation amounts and winds. The tabulated monthly summaries in the product include maximum, minimum, and average temperature, temperature departure from normal, dew point temperature, average station pressure, ceiling, visibility, weather type, wet bulb temperature, relative humidity, degree days (heating and cooling), daily precipitation, average wind speed, fastest wind speed/direction, sky cover, and occurrences of sunshine, snowfall and snow depth. The source data is global hourly (DSI 3505) which includes a number of quality control checks.”*

[**https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00684/html**](https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00684/html)

**a) Pick an airport dataset to analyze. Look at your data. Plot the seasonal cycle of all of the weather observations at the airport. *Only keep the data values at Noon local time.* The Y-axis of your plots should be the weather variable, while the X-axis should be date. Make sure the axes of your plots are labeled, including units. Is there date-based seasonality of weather conditions at the airport? Which weather variables show strong date-based seasonality? Which weather variables show weak date-based seasonality? (10 points)**

**b) Based on a), speculate what you will find when you cluster the hourly weather observations at local noon. What “seasons” do you expect to find with data-based clustering? How many data-based clusters will be the best fit to the data? Will the data-based clusters correspond to the date-based seasons (i.e., Winter, Fall, Spring, Summer)? Why or why not? (10 points)**

**c) Scale your data such that each variable in the airport dataset has unit variance. Use k-means clustering to divide the airport data into 4 clusters. (5 points)**

**d) Visualize your results by making three plots. First, create X-Y scatterplots of weather variables with the cluster centroid identified to see which variables helping define the clusters. Second, create histograms of the assigned clusters in each of the date-based seasons to see if the data-based clusters match the date-based seasons. Third, make any plot of your own choice to analyze the data. Be sure to describe how you created the plot and why it is useful for analyzing the results. (15 points)**

**e) Time to Tinker! Experiment with different numbers of clusters. Based on your tinkering, how many data-based clusters do you think are the best fit for the airport? (5 points)**

**f) Summarize your results. Provide and describe the three plots from d) for your preferred k-means cluster analysis with the best-fit number of clusters identified in e). What did you learn from the clustering analysis? Did you learn anything physical from this analysis? Are the results consistent with your intuition of the “seasons” at this airport? (10 points)**

**Problem II) Apply machine learning to a dataset of your choice (60 points)**

1. **Identify a dataset that is of interest to you. Provide a thorough description of your data including the reference, the variable (including names and units), the sampling frequency, etc. Describe why you think machine learning may be useful to apply to your data. (5 points)**
2. **Choose a machine learning algorithm to apply to your dataset. Specify if it is a supervised or unsupervised technique. Explain in 3-5 sentences what you expect to see when you apply this machine learning to your dataset. Do this \*\*before\*\* you do the analysis. (There is not any penalty for it being wrong…!) (5 points)**
3. **Apply the machine learning algorithm to your dataset. Include at least 3 relevant figures that show your results (25 points)**
4. **Discuss your results. What did you find? If applicable: do you think there is added value from machine learning technique over a simpler technique? (10 points)**

*Note for grading Problem II. You are analyzing your own data and choosing your own machine learning method. Since there is no “right answer”, you will be largely graded on how well I can follow your description of the data, the methods, the results, and the conclusions. Keep your code and your explanations simple, clear, and easy to follow. Spend the time to make your code concise, clear, and well documented. Look at your code as an opportunity to re-enforce the understanding you have gained. Write enough that someone else can easily follow your analysis.*