Intro to Data Science, Fall 2018: Homework 2
The Data Science Process – a Case Study

Due at start of class on Sep 17

---

5. *This problem is appearing again with clear questions for you to answer.*

   The EPA publishes air quality data on a continuous basis. It is possible to download data on fine particulate matter air pollution, also referred to as PM2.5, which refers to the amount of particulate matter that is smaller than 2.5 micrometers in the air. Higher this number, more is the pollution. The fields are called the following: *RD, Action.Code, State.Code, County.Code, Site.ID, Parameter, POC, Sample.Duration, Unit Method, Date, Start.Time, Sample.Value*. The air quality index is in the column titled *Sample.Value*. More details on the descriptions of the data can be found at: `https://aqs.epa.gov/aqsweb/airdata/FileFormats.html`. Your homework is to follow the first few steps of the *Data Science Process* and to perform appropriate analysis comparing these two data files.

   The two data files are in the directory `https://users.cs.fiu.edu/~giri/teach/5768/F18/epaData/`. Note that this has been corrected from before. The two files for the two years, 199 and 2012, are called:

   1. `RD_501_88101_1999-0.txt`, and
   2. `RD_501_88101_2012-0.txt`

   To guide you through the analysis, consider the following questions and answer as many as possible with Python code. Some questions use tools that may not have been covered in class. Feel free to attempt answering all. You are also encouraged to be creative by contributing additional questions and attempting to answer them.

   (a) How many rows and columns are there in the two data files?

   (b) How many missing (null) values for air quality index are in the two data files?

   (c) What fraction of the air quality index values are missing?

   (d) How does the average air quality for the whole country compare for the two years?

   (e) What is the standard deviation for the two data sets?

   (f) What kind of broad inferences can you draw about the efficacy of the *Clean Air Act* across the nation? Are there improvements? Is there less variation?

   (g) What about the average air quality and the standard deviation for Florida?

   (h) What about the counties of Miami-Dade, Broward, Monroe and Palm Beach?

   (i) Which states, counties in the country had the max values, min values, best increase, worst decrease, etc.? If you look at the best/worst indicators, do they look clustered or random?

(j) What kind of visualizations would you use to perform the various analyses?

(k) Are the changes statistically significant? How did you infer this?

(l) Do you have explanations for differences in the four counties?

(m) How would you go about identifying factors that differentiate the counties? Does this involve looking at other data?

All the analysis should be done in Python and using the Python Notebook. You will submit a soft copy of your notebook. The name of your notebook should be **Firstname-Lastname-hw1.ipynb**

Mail your notebook before the deadline: **start of class on September 17**.

6. Read the information on the dataset from Kaggle described under "Predicting BTC Price Using RNN". An analysis by a user called Mengran Tang is provided at `https://www.kaggle.com/microtang/predicting-btc-price-using-rnn`.

   After reading the descriptions and learning about Bitcoins, insert the code into a Python notebook and reconstruct the analysis provided on the site. Submit the working python notebook. What suggestions do you have to improve the approach?