In-Class Exam 2

- Time: Nov 30, 13:30pm - 15:20pm (110 min)
- Open book, do whatever you can (by yourself!)
- Submit your ipynb file through the dropbox link: https://www.dropbox.com/request/c6QwwbsgkmDM8Pq93JOI
- Early submission is allowed
- **Labeling, figure and plotting styles matter**
- Submission Grace period: 3:20 - 3:25pm (5 min)
- After 3:25pm, grade drops 5% for every 5 min delay (e.g., if the ipynb file is submitted at 3:30pm, your grade will be downscaled by 5%, and so on)

1. Analyze one of the following two datasets (30%). If you finish both problem sets, you may get an extra 15%

1a. Ocean tides

The data file 'h329b.csv' contains hourly sea-level data measured at stations in Hong Kong between 1986 and 2018. The four columns are "Year", "Month", "Day", "Hour" and "Water Level".

- Generate a plot showing the variation of the sea level measured at Hong Kong in December, 2018 (10%)
- Generate a periodogram for the sea level data in December, 2018 (10%)
- Identify the main periodicities in the sea level data in the periodogram of December 2018. What's the main period for the ocean tides in Hong Kong based on your analysis? (10%)

1b. Local Air Quality

The data file 'HK_air_quality1.xlsx' contains local air quality measurements (pm2.5, pm10, o3, no2, so2 and co) between 2013/12 and 2020/4. Since the air quality measurement cannot be negative, **negative values** in the data file are regarded as **invalid** measurements and should be ignored in the following analysis.

- plot the measured daily PM2.5 and PM10 in year 2019 as a function of time (10%)
- plot the yearly-averaged PM2.5 as a function of year (i.e., horizontal axis is year 2014 - 2019). What is the general trend of PM2.5 measured at Hong Kong? (10%)
- plot the monthly-averaged PM2.5 as a function of month (i.e., horizontal axis is January to December) for each year (2014 - 2019). What kind of variation(s) do you see in the monthly average PM2.5 data? Which month of the year has the best air quality in terms of PM2.5? (10%)

2. Record of sunspot numbers (70%)

The data file 'sunspot_d_total.csv' contains the daily Sunspot number in the past two hundred years (from 1818 to 2020). The columns with the time information in the data file are "Year", "Month", "Day", "frac_Date", and the data column you're going to analyze is "Daily_smooth_number", which is the measured daily sunspot number.

- Load the file into a pandas dataframe, make sure that you skip the necessary number of lines. Show the first ten lines of the dataframe (10%)
- What are the maximum, minimum and standard deviation of the sunspot number in the past 200 years? (10%)
- Plot the daily sunspot number as a function of time; (10%)
- Generate a histogram of the daily sunspot nunmber. Is it a normal distribution? (10%)
- Create a new column to the data frame (e.g., named "monthly_mean") by performing a 30-day moving average to the "Daily_smooth_number" column. This is roughly the monthly averaged sunspot number. Plot the "monthly mean" sunspot number as a function of time. (10%)
- Based on the 30-day smoothed data, generate a periodogram of the daily sun spot number, identify the leading periodicity of the variation in terms of days (or years). This is basically the so-called "solar cycle". (10%)
- Now choose the data from three solar cycles between year 1976 and 2007:
    - Use the monthly-smoothed sunspot number (the "monthly mean" column calculated before) to perform a non-linear curve fitting to the daily sunspot number using the following function:

      $$y = a + c\sin^2(bx)$$

      here x is time and y is sunspot numbers. [hint: you may need to specify an initial guess for the curve-fitting function] (10%)