CIS 5400 – Topics in Comput Info Sys.
Topic: Data Analysis Methods

Spring 2020

Instructor: Fitzroy Nembhard

Homework 8: Numerical Tools (NumPy and Pandas)

Total Points: 45

Date Assigned: Wednesday, Apr 8, 2020 **Due Date:** Wednesday, Apr 15, 2020

Submission Instructions: Please submit your work on Canvas as a Jupyter Notebook `ipynb` file named `yourname_cis5400hw8.ipynb`. Make sure to add comments/headings to your code showing especially the question numbers.

Key Data Analysis Methods Demonstrated

- Preprocessing
  - Parsing CSV files
  - Merging two CSV files
- Working with Time Series Data
  - Resampling data based on a certain frequency
  - Computing data coverage/duration
  - Plotting time series data
  - Computing ROI
  - Testing a statistical hypothesis based on time series data

Given the following two datasets that can be found on Canvas, please answer the following questions:
Dataset 1: ***apple_stocks_1980_2014.csv***
Dataset 2: ***apple_stocks_2015_2020.csv***

Preprocessing

1. (5 points)
   (a) Use Pandas to read the two datasets and merge them into one dataset using the following line of code:

   ```
   full_dataset = df1.append(df2, ignore_index=True)
   ```

   (b) Convert the ***Date*** column to a ***datetime*** object and set it as the index of the dataset; display the head of the dataset to confirm the changes.

   (c) Sort the dataset in ascending order by date and display the head to confirm the changes.

Data Analysis

2. (10 points) Filter the data for records that appear on the *business month end*. (Hints: Use Pandas *resample* function to resample according to *business month end*. Locate the frequency codes on slide 66 of the lecture for week 12. You may also read section 6 of chapter 11 to learn more about the resample function).
3. (5 points) How many months of data do we have in the resampled data?
4. (10 points) On a single chart, plot both the ***stock closings*** and the return on investment ***(ROI) based on stock closings*** for the full dataset. Use *calendar day* for the frequency. (Hint: see slide 77-78 of the lecture for week 12).

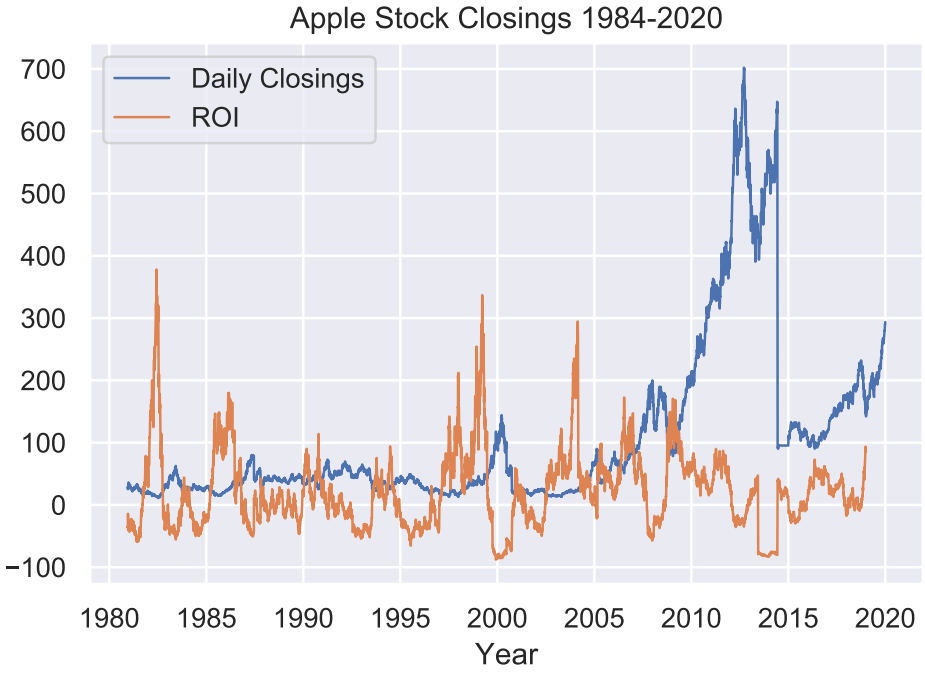

Figure 1: Expected plot for question 4

5. (15) It has been touted that Apple traded the same **volume** of stocks for 2018 and 2019. Using a random sample of $N = 15$ from these two years of data (**sample_2018** and **sample_2019**), determine whether we can apply the t test to test this hypothesis. If yes, conduct an independent-samples t test and state your conclusion. Use $\alpha = 0.05$ for all your tests.