CSE 482 Exercise 4 (Date: October 16, 2020)

The purpose of this exercise is to help you learn how to use some of the preprocessing functions available in Python. Follow the instructions below to complete the exercise. Save your IPython notebook as exercise4.ipynb.

1. Download the data msft.csv from the class Web page. The data contains the stock prices for Microsoft from January 2007 to December 2016.
2. In this exercise, you will look for the days in which there are significant increase/drop in the closing price of the stock. To do this, write an ipython notebook that implements the following steps:
   a. Use pandas to load the CSV file into a DataFrame object named "data" (see lecture 2 as an example). After the file is successfully loaded, type data.head() to display the first 5 rows of the table.
   b. Type data.describe() to display summary statistics about the different columns of the table.
   c. Extract and display the first five lines of the closing price of the stock as follows:

```
closing = data['Close']
closing.head()
```

   d. Compute the change in the closing stock price of the stock compared to its previous trading day. Note that the rows in the CSV file have been ordered in such a way that the first row corresponds to the latest closing price (December 30, 2016) and the last row corresponds to the earliest closing price (January 3, 2007). You also need to convert the index of the Series from integers to the date of the stock price:

```
cdate = data['Date'].values
cdate = cdate[1:]                    # this will return the date for your index
N = closing.size
change = closing[:N-1].values-closing[1:].values      # this calculates the price difference
changeData = Series(change, index=cdate)              # this creates a new Series
changeData.head()                                     # display the first 5 rows of the Series
```

   Note: If there is an error in the code, make sure you have included Series as one of the classes imported from Pandas.

   e. Plot a histogram of the changeData time series (see lecture slides). Does it look more like a Gaussian distribution (bell-curve)?
   f. Suppose we are interested to identify abnormally large increase in the closing price of the stock. Standardize the time series (i.e., compute the Z-score) by subtracting its mean and dividing by its standard deviation. After standardization, select the rows in which the Z scores are above 4 (i.e., the change is more than 4 standard deviations from the mean).
   g. Repeat the previous step, except you should select the rows that have Z-scores below -4. These correspond to the dates in which they is a substantial drop in the price of the stock.

3. For this question, you will apply different discretization methods to the “closing” Series object you had created in step 2(c).
    a. Apply equal width discretization to produce 5 bins. Display the first 5 discretized values by using the head() function:

        bins = ….

        bins.head()

    b. Apply equal frequency discretization to produce 5 bins. (Note that the quantiles selected in the lecture slides will produce only 4 bins. Therefore, you need to choose the appropriate quantiles that will produce 5 bins).

        bins = …

        bins.head()

Deliverables: Submit (via d2l) the file **exercise4.ipynb** created.