rd576, bl755, jg2262 **ORIE 4741 Midterm Report** Project

Part 1. Stock Dataset Description

1. Summary
   In this project, 3 datasets are used: the fundamentals dataset, the securites dataset, and the prices dataset.
   The dataset is downloaded from Kaggle and the original sources are: Yahoo Finance (stock prices), Nasdaq Financials (fundamentals) and extended by some fields from EDGAR SEC databases.

2. Fundamentals dataset
   The fundamentals dataset contains 77 explanatory variables which represent the fundamentals of 501 public listed stocks in NYSE according to their SEC 10K annual fillings. The time frame of this dataset is from 2012 to 2016. The number of observations is 1781.

3. Securities dataset
   The securities dataset contains 5 explanatory variables that gives general descriptions about the companies such as industry sectors. The number of observations is 505, however, we will exclude the stocks which do not appear in the fundamentals dataset or prices dataset.

4. Prices dataset
   The prices dataset contains daily historical prices data of each stock from 2010 to 2016, including open, close, low, high and volume. At this time, the closing price is selected to be the response variable but other variable may also be used later in the project. The number of observations is 851,264.

   Since the fundamentals data is annual and the prices data are daily, we will need to convert the daily prices into annual average prices in order to match the time frame.

   According to our experience, we may not use the full 77 explanatory variables in our model since some of them may be redundant and a complicated model will be difficult for interpretation. Variable selection will be performed to select the variables that are significant to our model.

Part 2. Visualization

1. Volume:

   From this plot, we observed the volume of stock traded each day against its price. The volume could be engineered into a variable that described the market liquidity of the stock we are interested in.

   In addition, we realized the stock prices and returns involve many fluctuations and noises. One way to smooth the time series is to taking a moving average. We plotted the half-month and one-month windows below.

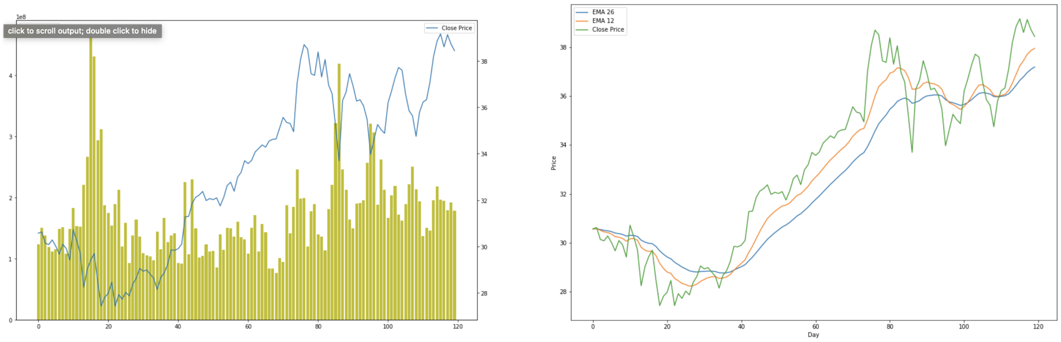
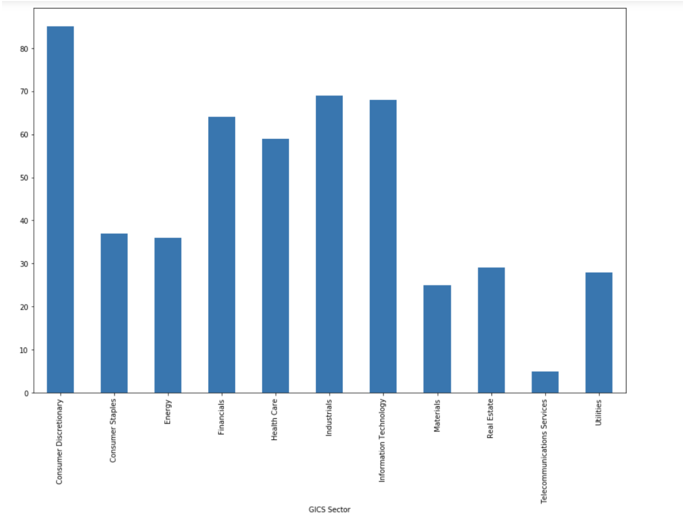


Figure 1: Volume and Moving average

2. Sector breakdown:

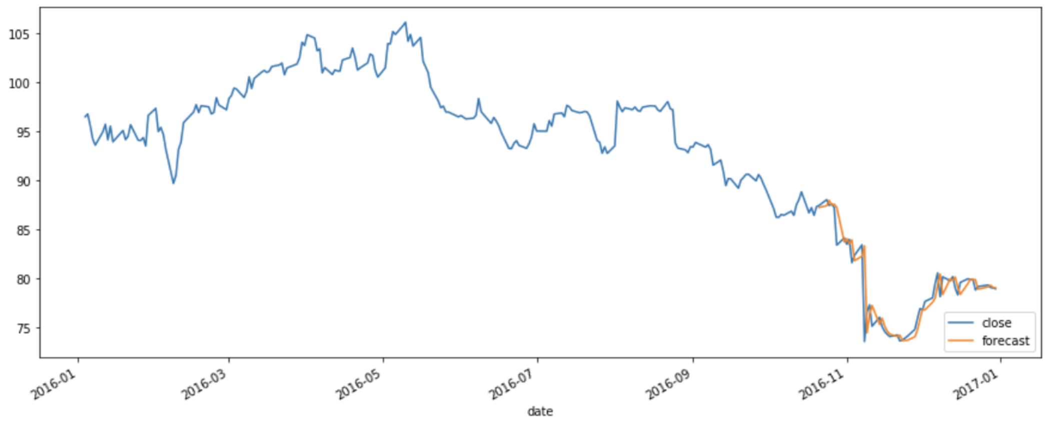
   On the fundamental analysis side, we understood that companies in different sectors would have unique business models and financial behaviors. The sectors breakdown is plotted below. Based on the plot, we are considering to focus on the Financials, Industrials, Healthcare, and Info Technology sectors. Since we have similar number of companies in those sectors, it would be easier for us to compare results.



Part 3. Technical Analysis Modeling (Auto-regressive)

Since we learned that the stock price is a time series, we could make the assumption that the historical stock price had predicative power over future prices. To model this relationship, we could use an auto-regressive model from Python's time series analysis package. Taking CVS's stock data for example:

```
model_CSV = ARIMA(trail, order = (4, 0, 0))
```



In this trail run, we observed that CSV's stock price is downward sloping. Although the model has the ability to fit the last few sessions of the stock traded fairly well, it has limited power to forecast long-term trend. And it would be hard for us to back any of such prediction with proper interpretations. Consequently, the limits of the auto-regressive model and technical analysis based approaches provide strong incentives for us to include fundamental information in our data set in our model.

Part 4. Future Planning

1. We will perform variable selection and exclude insignificant explanatory variables when fitting the fundamentals model.

2. We would explore models alternative to auto-regressions. For example, we could use k-means clustering to classify similar stocks.

3. We could perform more feature engineering. In addition to using data in a single company, we could use data from companies under the same sector to perform our analysis.