A study on U.S. Stock Market Index and Stock Prices.

Elizabeth Peiwen Li

Content

- Project Overview
- Dataset Extraction & Cleansing
- Study the impact of economy recession to the prices of the chosen indices and stock prices.
- Descriptive analysis on one of Stock Market Index.
- Predictive analysis on the Stock Market Index.

Project Overview

Part 1 Extend previous assignment, further study the impact of economy recession to the monthly prices of the chosen indices and stock prices for the past 40 years.

- Chosen indexes: Standard Poor 500, Dow Jones' Index, 10 Years Treasury Note Yield Index and Nasdaq Composite.
- Chosen U.S.stocks : GOOG(Alphabet: google);AIG (American International Group); XOM(Exxon Mobil Corporation); UAL(United Airline).
- Chosen four C.N stocks listed in U.S. market: PTR(petrochina company limited); BIDU(baidu); CEA(china eastern airlines); LFC (China life insurance company limited).

Part 2 Study the impact of economy recession to the monthly prices of the chosen indices and stock prices.

- Study the impact of economy recession on the monthly closing price of chosen indexes and stock during the past 40 years.
- Study the impact of economy recession on the monthly price change percentage of chosen indexes and stock during the past 40 years.

Part 3: Choose one of the stock market index to conduct descriptive analysis and predictive analysis.

- Describe the dataset.
- Descriptive analysis.
- Predictive analysis.

Data Extraction & Cleaning

- Data Extraction: Yahoo Finance (import yfinance)

```
df_index_Dow.head()
```

Date	Open	High	Low	Close	Adj Close	Volume
1985-01-01	1277.719971	1305.099976	1266.890015	1286.770020	1286.770020	44450000
1985-02-01	1276.939941	1307.530029	1263.910034	1284.010010	1284.010010	207300000
1985-03-01	1285.339966	1309.959961	1242.819946	1266.780029	1266.780029	201050000
1985-04-01	1264.800049	1290.300049	1245.800049	1258.060059	1258.060059	187110000
1985-05-01	1257.180054	1320.790039	1235.530029	1315.410034	1315.410034	242250000

➔ Dataset Attributes (Columns): Date, Open Price, High Price, Low Price, Close Price, Adj Price, Volume

➔ Dataset Samples (Rows):
e.g. SPY - 480
DOW - 408
GOOG - 175

- Data Cleaning

Step1: Drop the rows with missing values.

e.g. df_index_SPY = df_index_SPY.dropna()

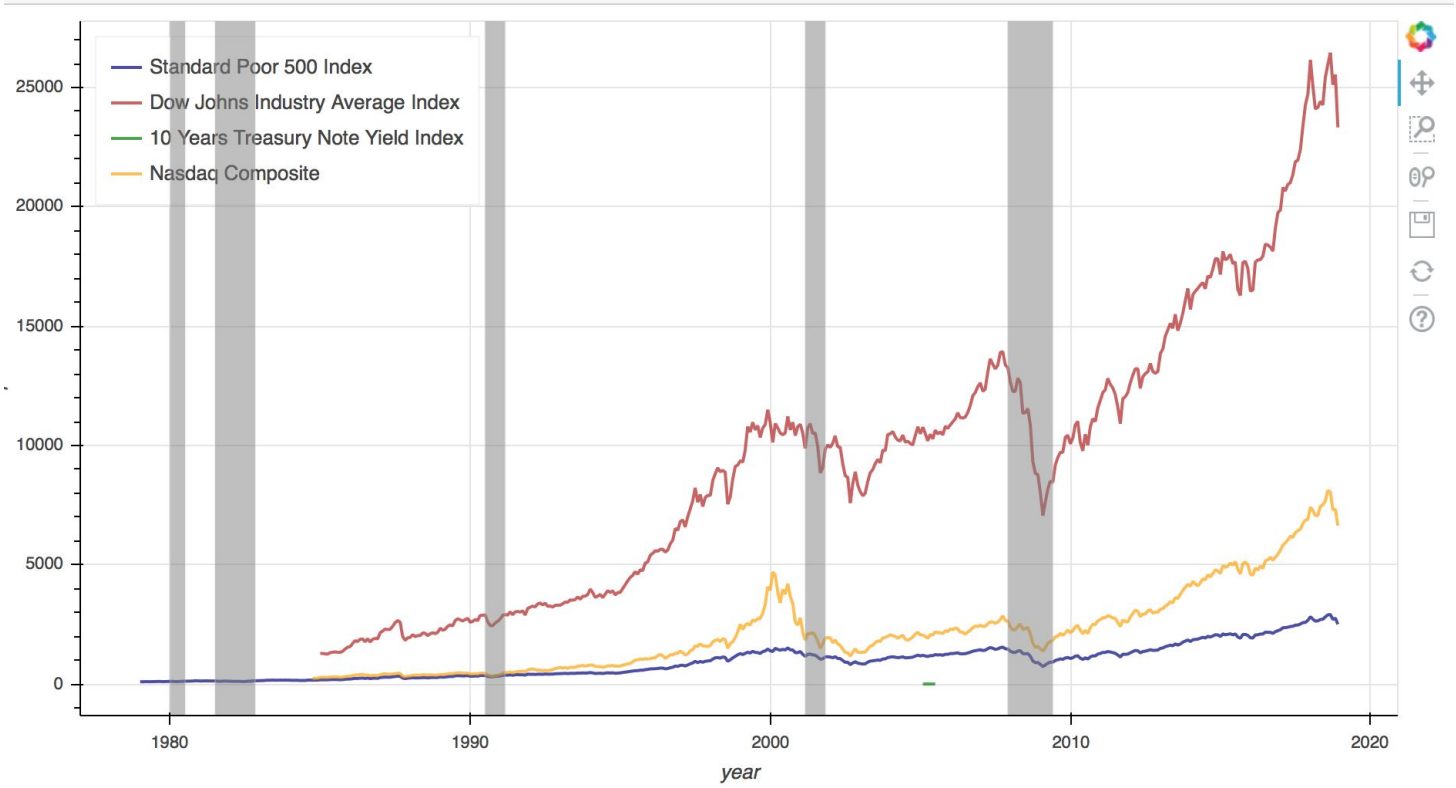
Step2: Drop all the rows with zero values.

df_index_SPY = df_index_SPY[~(df_index_SPY == 0).any(axis=1)]

Step 3:#Convert the index of the dataframe into a column

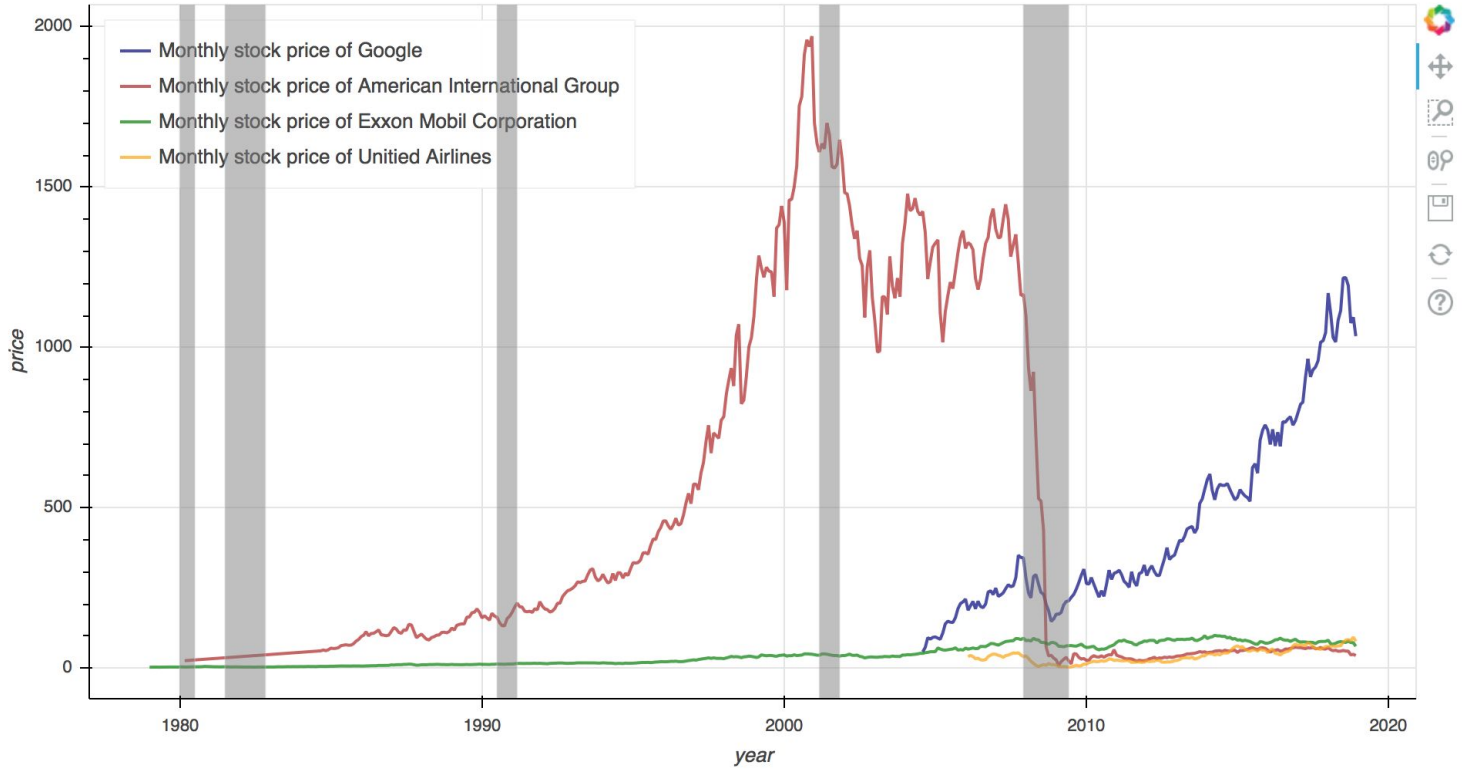
- **Impact of Economy Recession on Index Price**

Data Visualization: Interactive Line Chart with shaded recession duration.



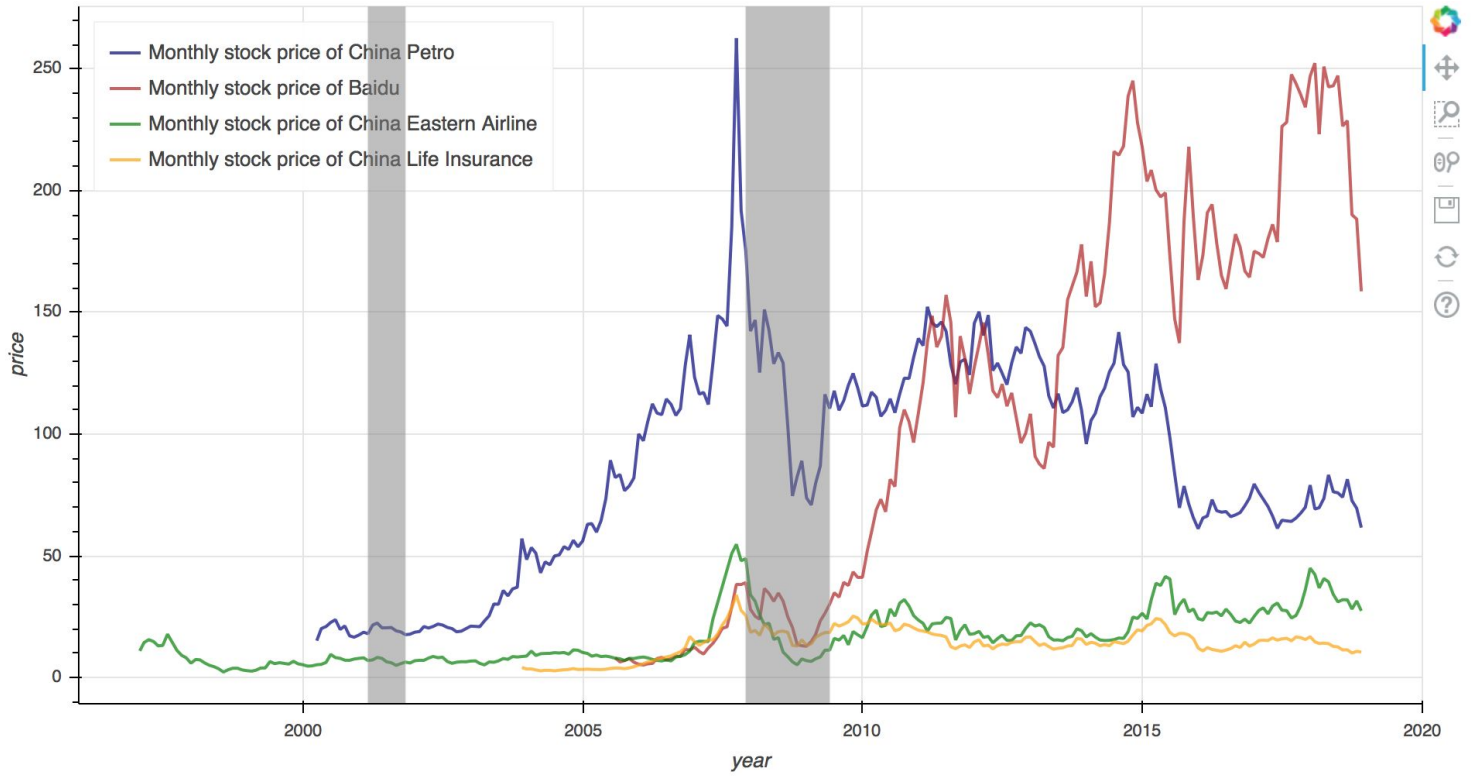
- # Impact of Economy Recession on Stock Price

Data Visualization: Interactive Line Chart with shaded recession duration.



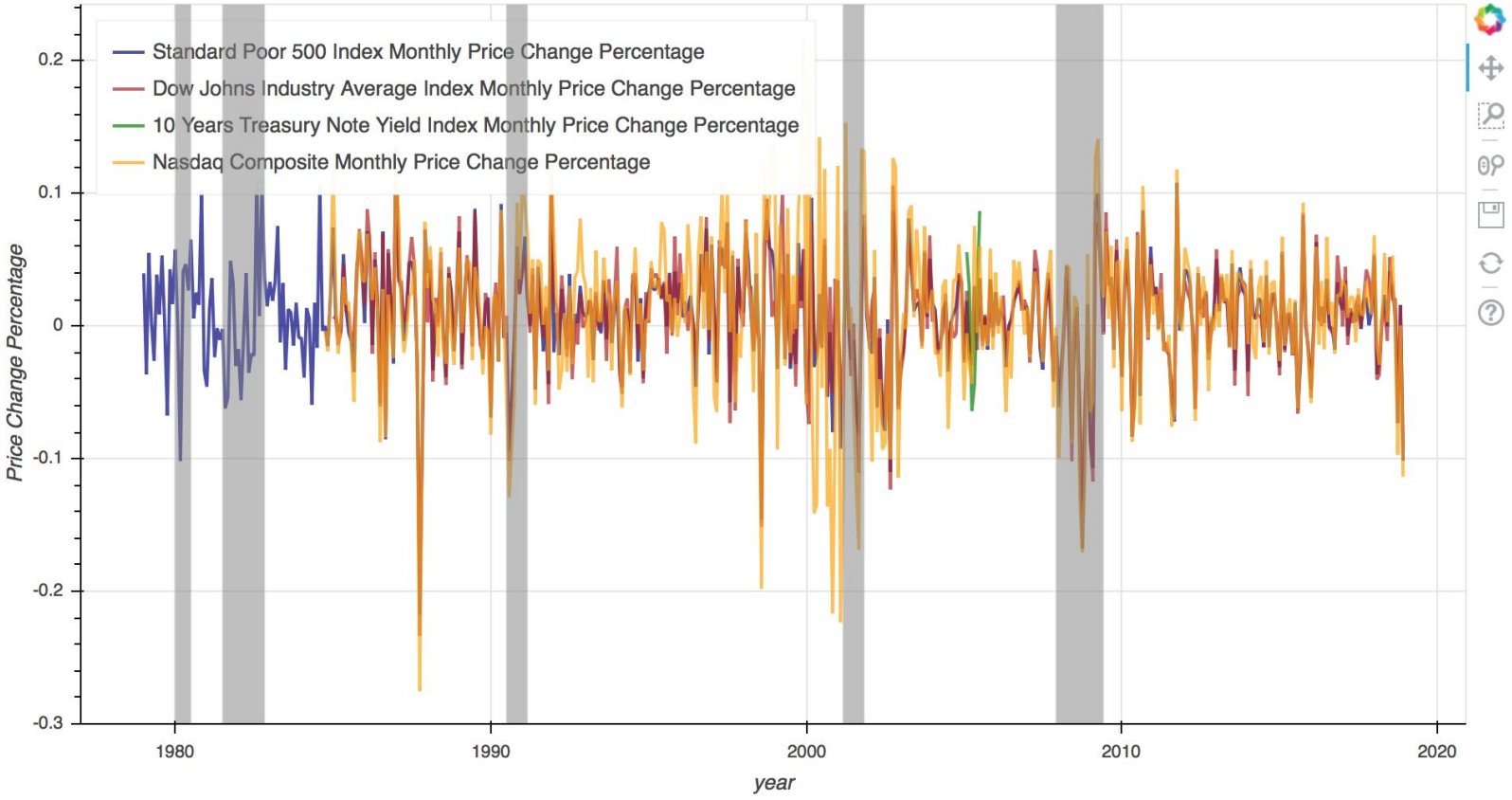
- **Impact of Economy Recession on Stock Price**

Data Visualization: Interactive Line Chart with shaded recession duration.



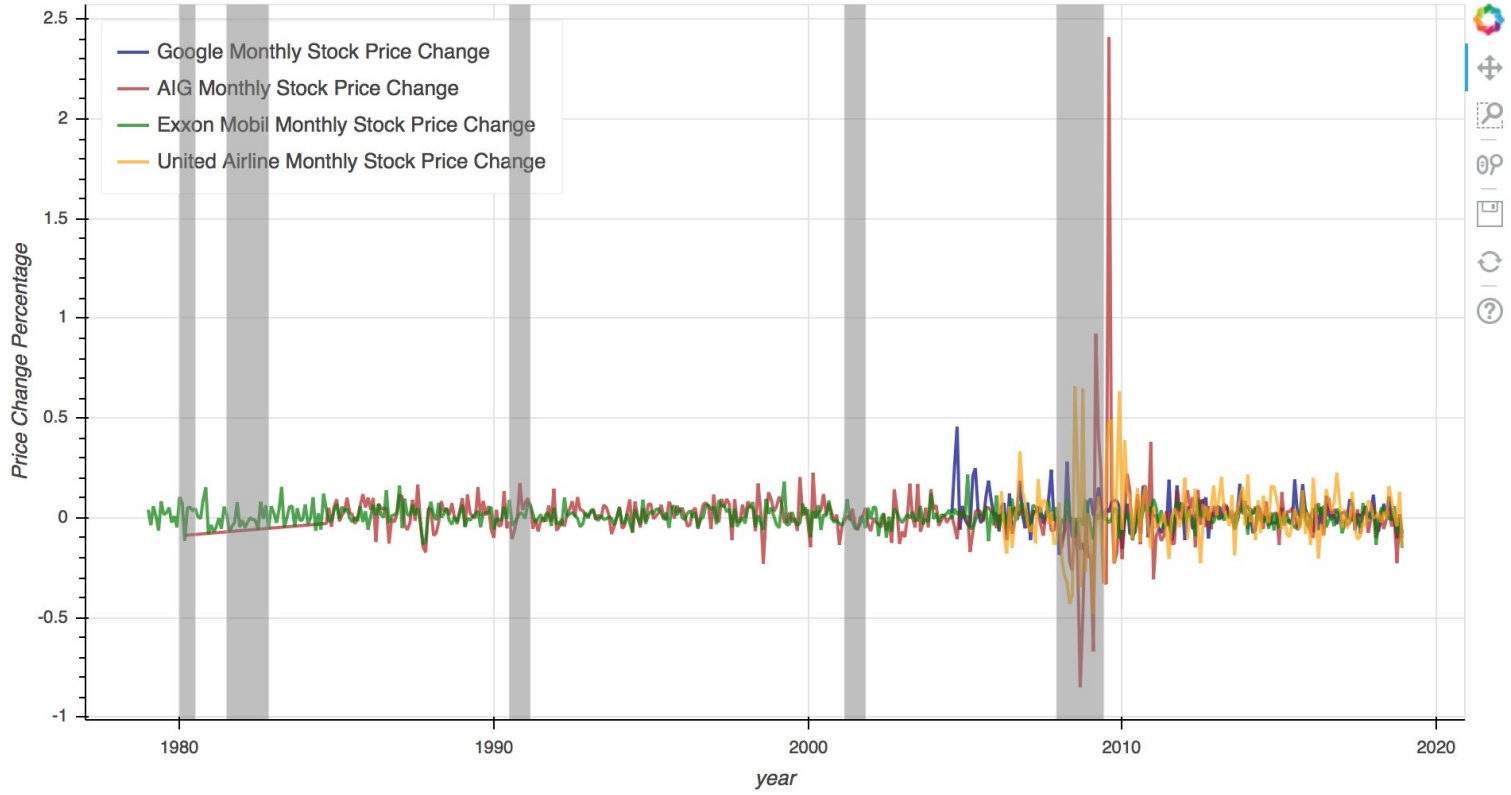
- **Impact of Economy Recession on Index Price Change Percentage**

Data Visualization: Interactive Line Chart with shaded recession duration.



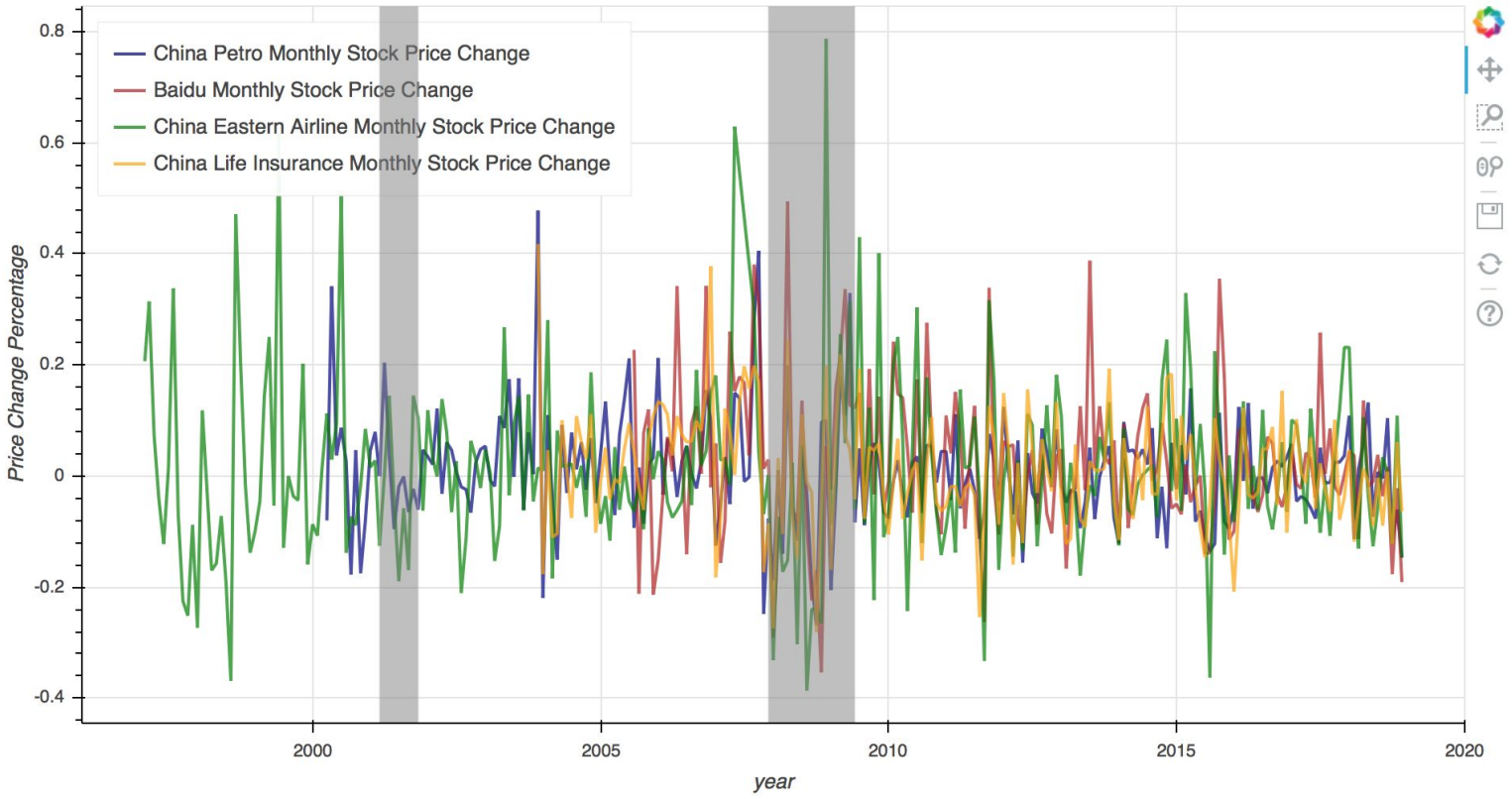
- **Impact of Economy Recession on Stock Price Change Percentage**

Data Visualization: Interactive Line Chart with shaded recession duration.



- **Impact of Economy Recession on Stock Price Change Percentage**

Data Visualization: Interactive Line Chart with shaded recession duration.



Interesting Findings

Findings from study on impact of recession on stock & indexes prices

1. The economic recession did affect the stock market inddex and each stock we chosen here.
2. Among the five economic recession happened during the past 40 years, we can see that 2008 global finance crisis had the most negative influences in the stock market, especially for the American International Group. The price almost hit the bottom. It reminded me the U.S. government used 182 billion dollars to bail it out from bankruptcy.
3. Compared with the traditional giants like energy, airline or insurance finance companies, the high tech company has a tremendous growth rate.
4. For Chinese companies listed in the U.S. stock market, it shared the similar pattern with the U.S. counterpart, meaning the market is fair. If one industry is booming, both stock prices of U.S. companies and Chinese companies in that industry are likely to grow.

Findings from study on impact of recession on stock & indexes prices changes

1. Compare with the major three stock market index (S&P 500, Dow and Nasdaq Composite), S&P 500 has been relatively stable than the other two.
2. From the 4 chose U.S. stocks, the energy company Exxon Mobil has been relatively stable.
3. Over the past 40 years the insurance tycoon AIG has also been relatively stable, except during the time in 2008 finance crisis. Similarly, for Chinese companies, both the energy and insurance company China Petro and China Life Insurance have relatively stable stock prices compared with the other 2.

Possible suggestion: if choose long term stock investment,to choose a stock, the traditional industry such as energy and life insurance still worth our time to pay attention for.

Descriptive Analysis on an index

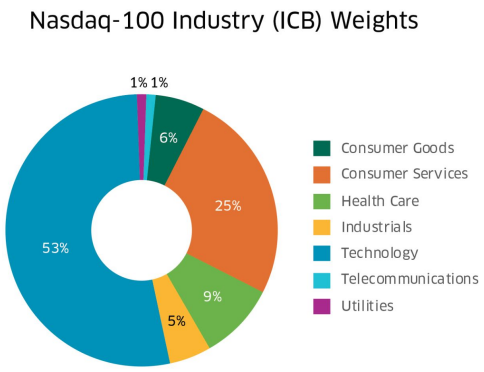
Final choice: Standard Poor 500

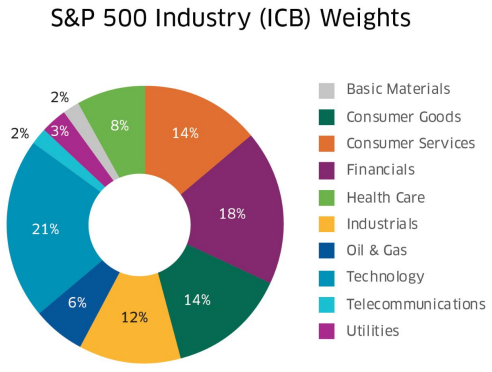
Why S&P 500?

— **Drop TNX:** The fluctuation of 10 years Treasury Note Yield Index has been in a relatively very small range.

— **Drop DOW:** Dow is a stock market index that measures the stock performance of 30 large companies listed on stock exchanges in the United States. Although it is one of the most commonly followed equity indices, since it only includes 30 companies and is not weighted by market capitalization and is not a weighted arithmetic mean.many investors consider the S&P 500 Index, which also includes the 30 components of the Dow, to be a better representation of the U.S. stock market.

— **Drop IXIC:** The Nasdaq-100 is heavily allocated towards top performing industries such as Technology, Consumer Services, and Health Care. The traditional industry like utility only weighted 1%, and industrials only weighted 5%. However, the varieties and weighted proportion of industries in S&P 500 are more balanced.And here, 6 out 8 chosen stocks in this study are from traditional industries.





- **Describe the Dataset**
  - Measure the central tendency & the variability

	Open	High	Low	Close	Adj Close	Volume	Price Change Pct
count	480.000000	480.000000	480.000000	480.000000	480.000000	4.800000e+02	480.000000
mean	927.330268	956.528356	896.323250	931.910812	931.910812	3.321781e+10	0.007568
std	690.643531	707.897561	669.697820	692.413605	692.413605	3.615204e+10	0.042564
min	96.110001	100.519997	94.230003	96.279999	96.279999	4.757100e+08	-0.217630
25%	302.359993	316.750008	292.495003	304.000000	304.000000	3.417065e+09	-0.016844
50%	918.869995	964.359985	869.385010	919.230011	919.230011	1.557574e+10	0.010440
75%	1328.909973	1376.079987	1281.195007	1329.197479	1329.197479	6.642231e+10	0.034857
max	2926.290039	2940.909912	2864.120117	2913.979980	2913.979980	1.618436e+11	0.131767

Describe & clean the Dataset

- Find the outliers of the dataset
  - Use IQR to find the outliers

```
# Find IQR for outliers of the dataset next step.
Q1 = df_index_SPY.quantile(0.25)
Q3 = df_index_SPY.quantile(0.75)
IQR = Q3 - Q1
print(IQR)
```

```
Open                1.026550e+03
High                1.059330e+03
Low                 9.887000e+02
Close               1.025197e+03
Adj Close           1.025197e+03
Volume              6.300524e+10
Price Change Pct    5.170080e-02
dtype: float64
```

  - Drop the outliers

```
# Drop the outliers if there's any.
df_new_index_SPY = df_index_SPY[~((df_index_SPY < (Q1 - 1.5 * IQR)) |(df_index_SPY > (Q3 + 1.5 * IQR))).any(axis=1)]
df_new_index_SPY.shape
```

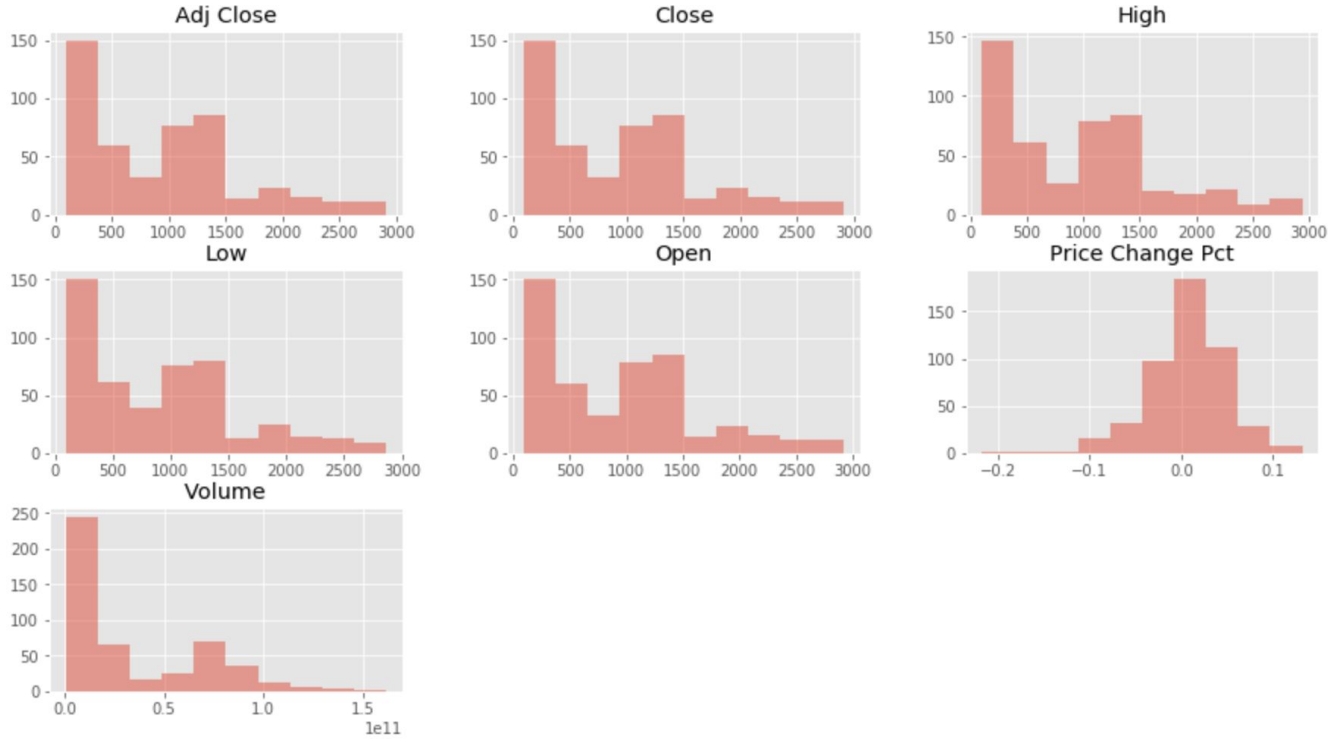
```
(468, 7)
```

  - 12 outliers were dropped.

- **Describe & clean the Dataset**

  - **Measure Kurtosis & Skew**

    — Plot histogram for each column



— Findings: Except price change percentage, all other 6 columns are obviously right skewed. Price change Percentage looks like Gaussian distribution, although not exactly centered around zero.

- **Predictive Analysis on S&P 500**

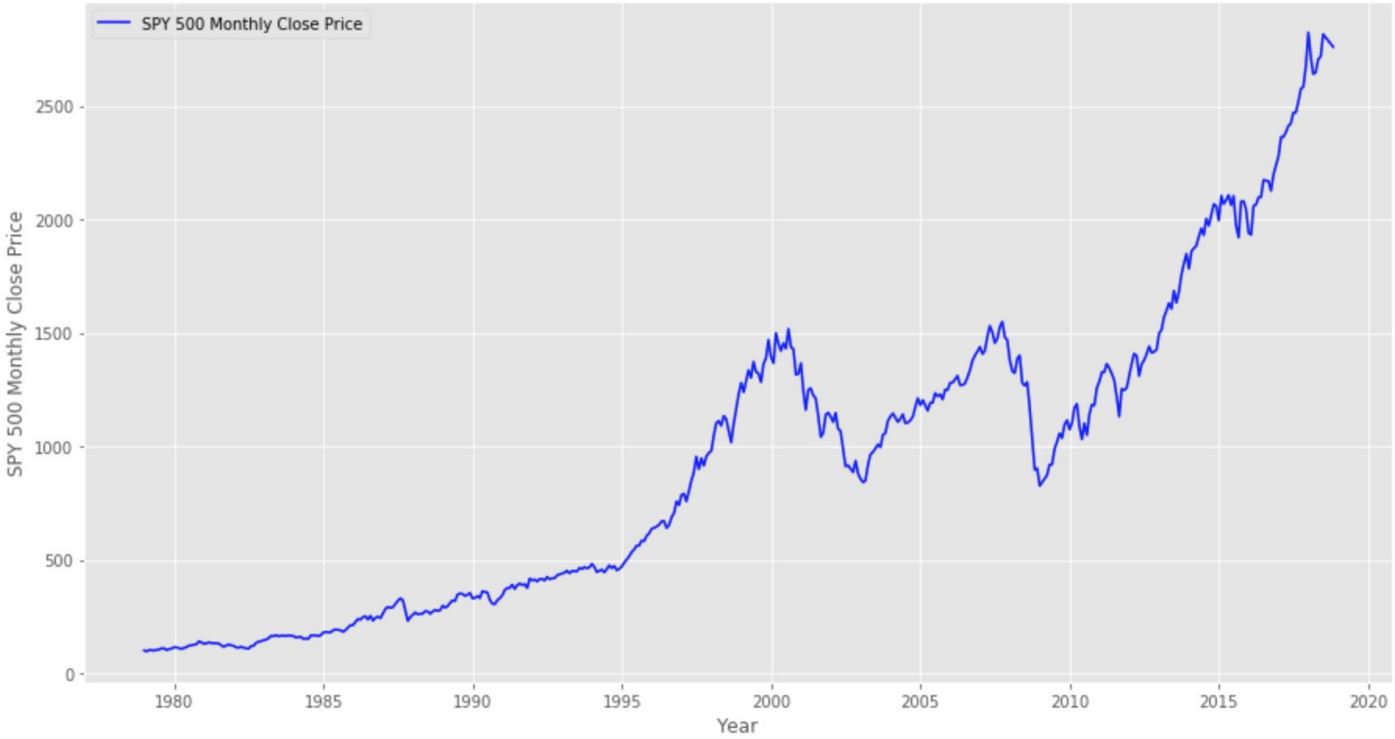
Part 1: Time Series Studies in S&P 500 Close Price

Part 2: Compare regression model for S&P 500 transaction volume and close price.

- **Time Series Studies in S&P 500 Close Price**

— Create a new subset Date- Close Price of S&P 500

— Data Visualization



— **Observation:** The components looks like jumps in magnitude overtime, and the pattern of seasonal variation is not very stable over the years. Therefore, it might be better to choose multiplicative model instead of addictive model. And compared with ARMA, for this case, it looks like it might be better to choose ARIMA model(Autoregressive Integrated Moving Average Model) .

- **Time Series Studies in S&P 500 Close Price**

— **Dataset Stationability Test**

- **Method 1 :**

```
X = df_cp_SPY['Close'].values
split = round(len(X) / 2)
X1, X2 = X[0:split], X[split:]
mean1, mean2 = X1.mean(), X2.mean()
var1, var2 = X1.var(), X2.var()
print('mean1=%f, mean2=%f' % (mean1, mean2))
print('variance1=%f, variance2=%f' % (var1, var2))
```

```
mean1=358.958462, mean2=1484.549528
variance1=61049.909513, variance2=222094.050799
```

- **Findings:** The mean and variance values are very different, so we can know that the time series sub dataset of SPY 500 Close price is not stationary. To further test it, we can also try a hypothesis testing to verify it.

• Time Series Studies in S&P 500 Close Price

— Dataset Stationability Test

- **Method 2 : ADF Test (Augmented Dickey-Fuller Test)**

➔ **Null Hypothesis (H0):** If failed to be rejected, it suggests the time series has a unit root, meaning it is non-stationary. It has some time dependent structure.

➔ **Alternate Hypothesis (H1):** The null hypothesis is rejected; it suggests the time series does not have a unit root, meaning it is stationary. It does not have time-dependent structure.

```
from statsmodels.tsa.stattools import adfuller
result = adfuller(X)
print('ADF Statistic: %f' % result[0])
print('p-value: %f' % result[1])
print('Critical Values:')
for key, value in result[4].items():
    print('\t%s: %.3f' % (key, value))
```

```
ADF Statistic: 1.600418
p-value: 0.997859
Critical Values:
        1%: -3.444
        5%: -2.868
        10%: -2.570
```

- **Findings:** The test statistic value is around 1.6, which is sitting inside the range is less than the critical value 2.57 (10%), and larger than the p-value 0.997859. This suggests that we can not reject the null hypothesis with a significance level of less than 10%,which verified our previous guessing. **Therefore, we'll use ARIMA model.**

- **Time Series Studies in S&P 500 Close Price**

— ARIMA Model

ARIMA Model Results

Dep. Variable:	D.Close	No. Observations:	467
Model:	ARIMA(5, 1, 0)	Log Likelihood	-2395.826
Method:	css-mle	S.D. of innovations	40.904
Date:	Wed, 11 Dec 2019	AIC	4805.652
Time:	15:18:04	BIC	4834.676
Sample:	1	HQIC	4817.074

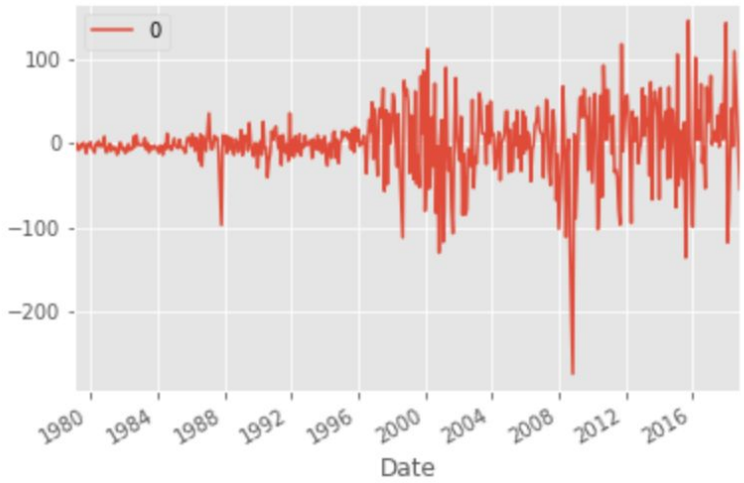
	coef	std err	z	P>|z|	[0.025	0.975]
const	5.7164	2.155	2.653	0.008	1.493	9.939
ar.L1.D.Close	0.0019	0.046	0.042	0.967	-0.088	0.092
ar.L2.D.Close	-0.0457	0.046	-0.989	0.323	-0.136	0.045
ar.L3.D.Close	0.0387	0.046	0.837	0.403	-0.052	0.129
ar.L4.D.Close	0.0157	0.046	0.339	0.735	-0.075	0.106
ar.L5.D.Close	0.1123	0.046	2.432	0.015	0.022	0.203

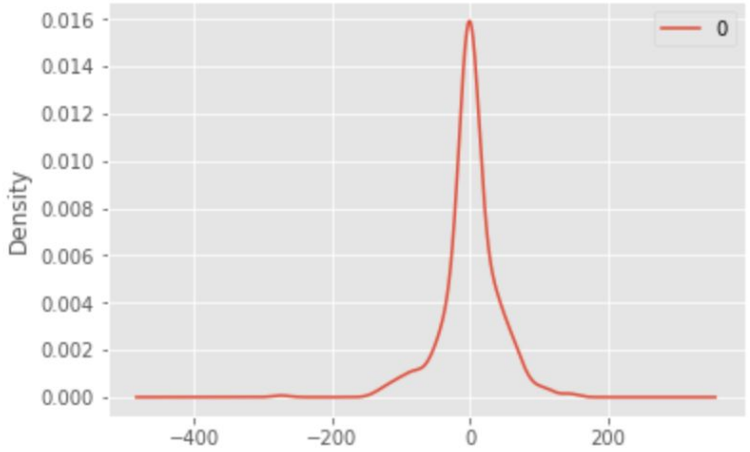
Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.5115	-0.0000j	1.5115	-0.0000
AR.2	-1.2337	-0.9707j	1.5698	-0.3939
AR.3	-1.2337	+0.9707j	1.5698	0.3939
AR.4	0.4082	-1.4911j	1.5460	-0.2075
AR.5	0.4082	+1.4911j	1.5460	0.2075

• Time Series Studies in S&P 500 Close Price

— ARIMA Model



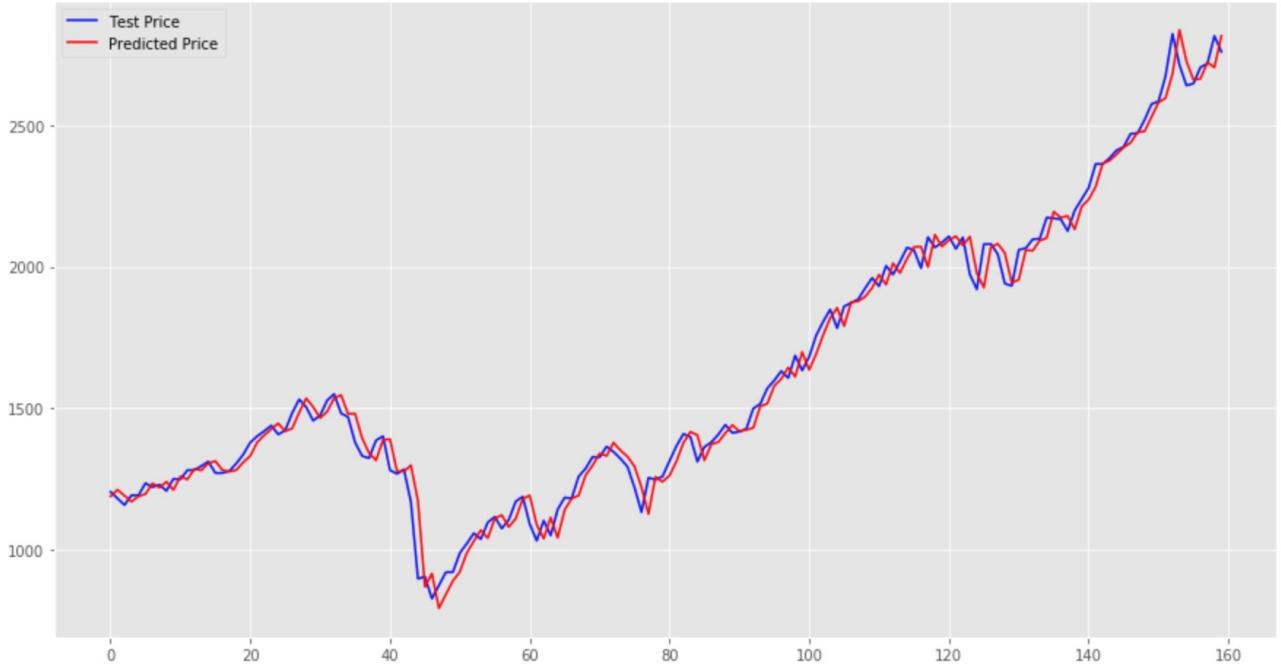


—Findings:

- From the residual errors line chart, there's no very obvious trend, hence we can say this model did capture the trend of the dataset.
- From the density plot of the residual error values, we can learn that the errors are Gaussian distribution and they are centered on zero.And from the residuals' descriptive statistics, we can also learn that the mean value is 0.007299, which is very close to 0, hence we can say that there's merely no bias in the prediction.

- **Time Series Studies in S&P 500 Close Price**

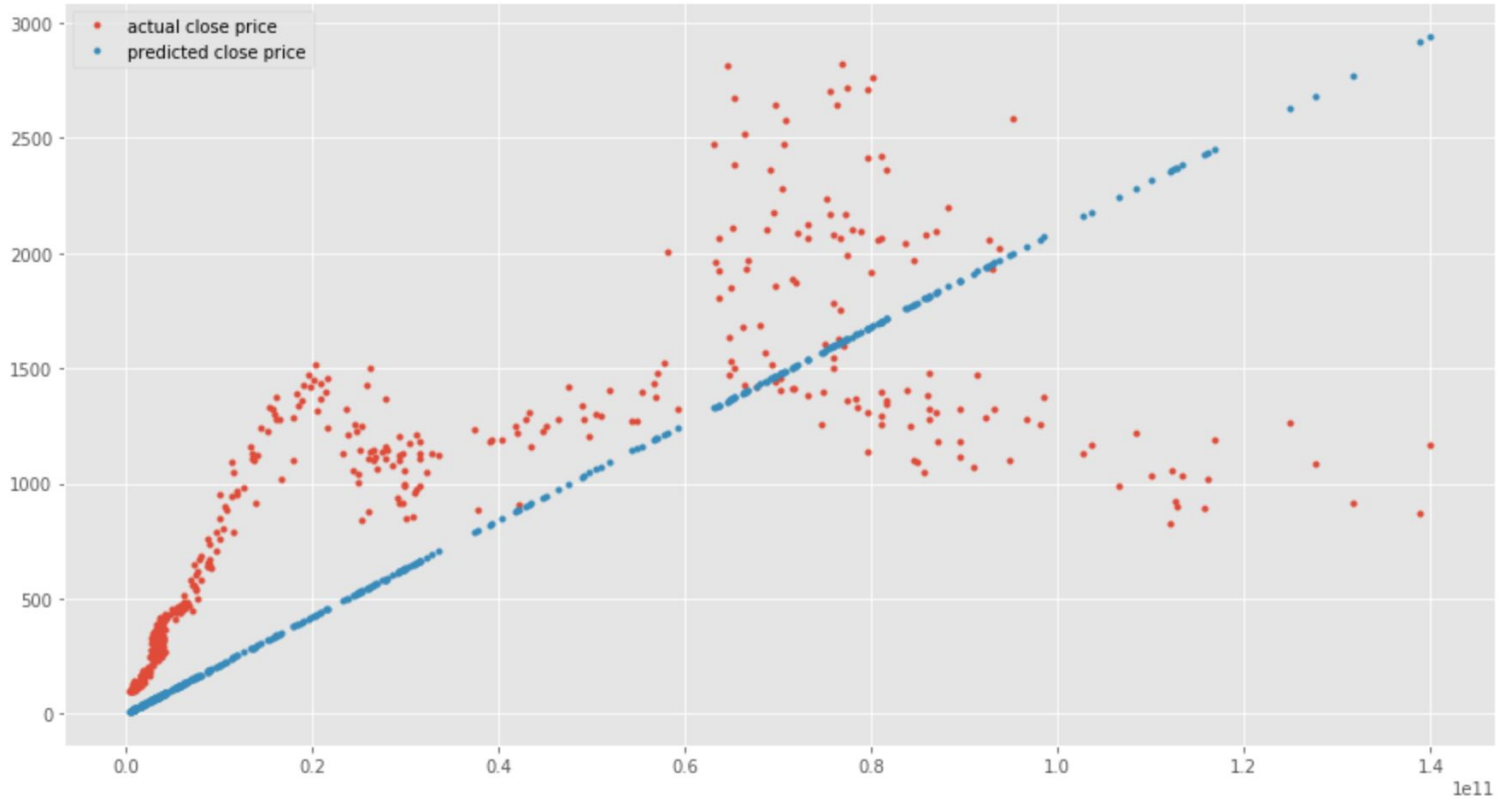
— Use the fit model to calculate the predictive values and plot the prediction using ARIMA model.



—Observations & Conclusion: From above chart, we can see that ARIMA model works well for the time series analysis and prediction, therefore, we can choose ARIMA model for time series studies for S&P 500 close price.

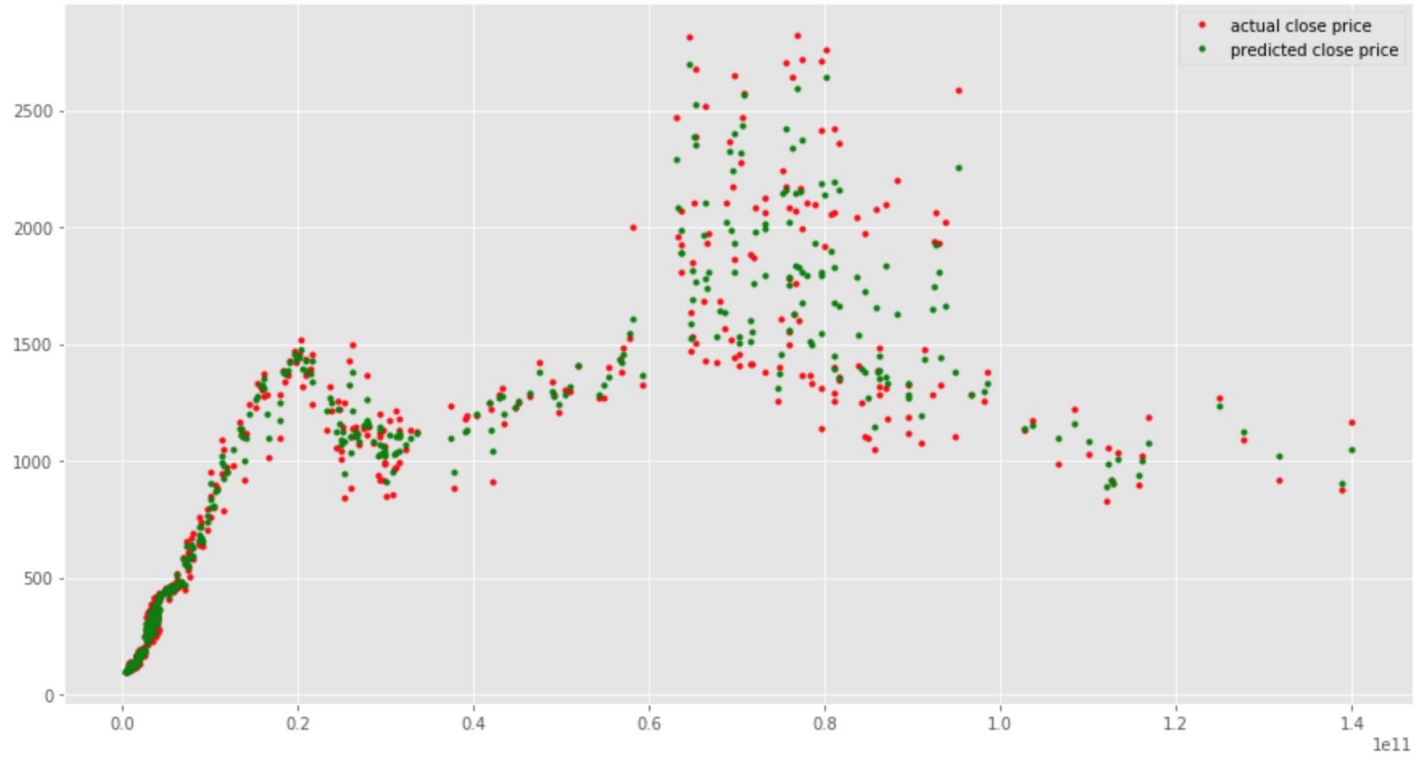
- **Use Sklearn methods for Regression Model Comparison for S&P 500 Transaction Volume and Close Price**

— **Linear Regression Model**



- **Use Sklearn methods for Regression Model Comparison for S&P 500 Transaction Volume and Close Price**

— Random Forest Regression Model



—Observation & Conclusion: The random forest model works better for the S&P 500 dataset for the predictive analysis. So to predict the close price by the transaction volume, we'll use the random forest regression model.

Thank you!