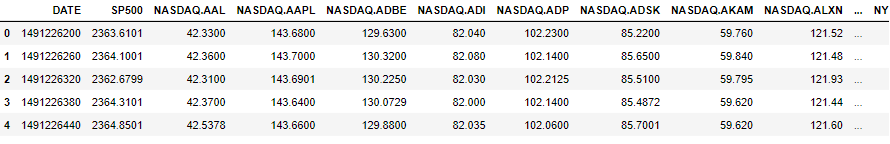
**Stock Market Analysis and Prediction**

**Introduction:**

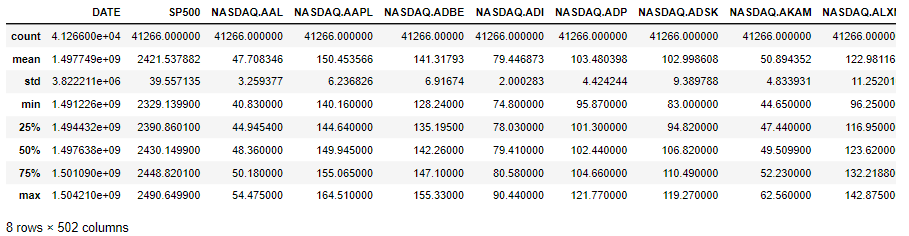
Time series data is a type of data that is collected over time at regular intervals. It is characterized by its chronological order and can exhibit trends and seasonality. Time series data can be analyzed using various statistical and machine learning techniques to make predictions or forecasts about future events based on patterns observed in past data.

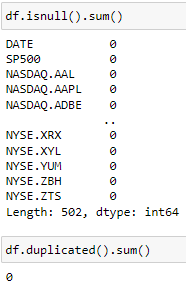
**Dataset:**

My dataset consist of many columns, I interested in DATE and SP500 columns



Description of data

**Preprocessing:**



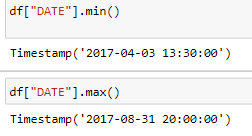
Conclusion

very good dataset since -> no missing values or duplicate values.

Apply preprocessing on DATE column, since this number 1491226200 represents

the Unix timestamp, we need convert this number to datetime object

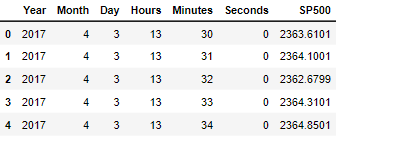
Like(1491226200 to 2017-04-03 13:30:00)



#### the conclusion of this data that the date start from 2017-04-03 13:30:00 and end 2017-08-31 20:00:00, i have same year but difference in month and day and time

Split column DATE into year, month,day,hours,minutes, seconds

make new data frame with column Year,Month,Day,Hours,Minutes,Seconds,sp500



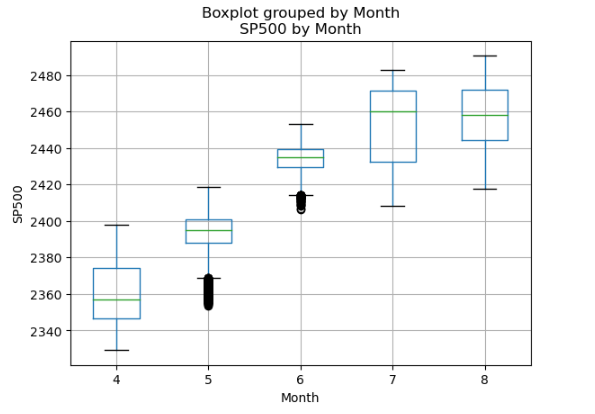
## Exploratory Data Analysis:

## plot the SP500 over time

## 

Conclusion that SP500 price increasing over time from 2017-04 to 2017-09

**plot a box plot of the SP500 by month**



#### From the boxplot, we can get the following information.

#### Month 4

#### The max value of the SP500 in month 4 is around 2400 and the minimum is more than 2340. There are no outliers in the data. The median of the data is around 2395 and the 75th percentile is around 2370 and the 25th percentile is around 2330.[¶](http://localhost:8888/notebooks/stock-prediction.ipynb#From-the-boxplot,-we-can-get-the-following-information.-The-max-value-of-the-SP500-is-more-than-2480-and-the-minimum-is-less-than-2340.-There-are-no-outliers-in-the-data.-The-median-of-the-data-is-around-2430-and-the-75th-percentile-is-around-2448-and-the-25th-percentile-is-around-2390.)

#### Month 5

#### The max value of the SP500 in month 5 is around 2420 and the minimum is around 2370. There are outliers in the data. The median of the data is around 2350 and the 75th percentile is around 2400 and the 25th percentile is around 2390.[¶](http://localhost:8888/notebooks/stock-prediction.ipynb#From-the-boxplot,-we-can-get-the-following-information.-The-max-value-of-the-SP500-is-more-than-2480-and-the-minimum-is-less-than-2340.-There-are-no-outliers-in-the-data.-The-median-of-the-data-is-around-2430-and-the-75th-percentile-is-around-2448-and-the-25th-percentile-is-around-2390.)

#### Month 6

#### The max value of the SP500 in month 6 is around 2455 and the minimum is around 2415. There are outliers in the data. The median of the data is around 2435 and the 75th percentile is around 2440 and the 25th percentile is around 2430.[¶](http://localhost:8888/notebooks/stock-prediction.ipynb#From-the-boxplot,-we-can-get-the-following-information.-The-max-value-of-the-SP500-is-more-than-2480-and-the-minimum-is-less-than-2340.-There-are-no-outliers-in-the-data.-The-median-of-the-data-is-around-2430-and-the-75th-percentile-is-around-2448-and-the-25th-percentile-is-around-2390.)

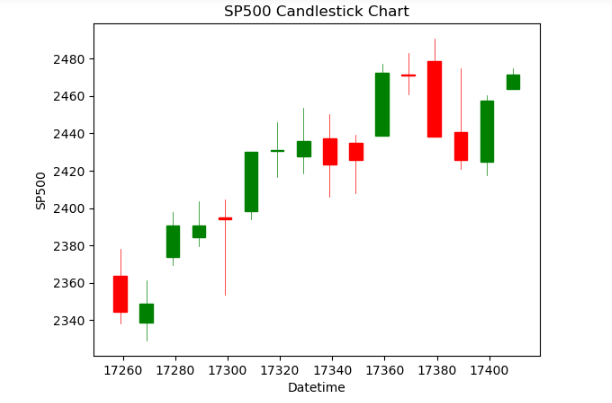
#### Month 7

#### The max value of the SP500 in month 7 is around 2485 and the minimum is around 2410. There are no outliers in the data. The median of the data is around 2460 and the 75th percentile is around 2470 and the 25th percentile is around 2435.[¶](http://localhost:8888/notebooks/stock-prediction.ipynb#From-the-boxplot,-we-can-get-the-following-information.-The-max-value-of-the-SP500-is-more-than-2480-and-the-minimum-is-less-than-2340.-There-are-no-outliers-in-the-data.-The-median-of-the-data-is-around-2430-and-the-75th-percentile-is-around-2448-and-the-25th-percentile-is-around-2390.)

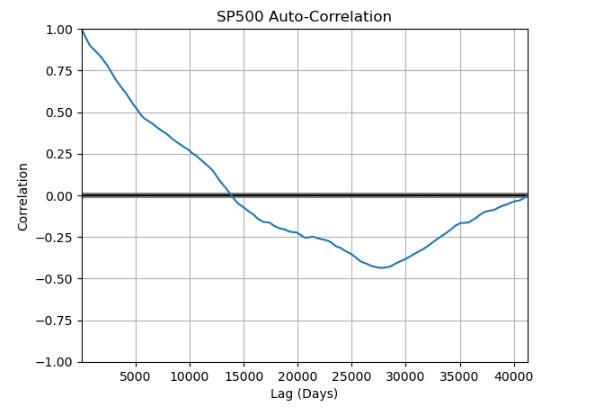
#### Month 8

#### The max value of the SP500 in month 8 is around 2490and the minimum is around 2420. There are outliers in the data. The median of the data is around 2470 and the 75th percentile is around 2470 and the 25th percentile is around 2445.[¶](http://localhost:8888/notebooks/stock-prediction.ipynb#From-the-boxplot,-we-can-get-the-following-information.-The-max-value-of-the-SP500-is-more-than-2480-and-the-minimum-is-less-than-2340.-There-are-no-outliers-in-the-data.-The-median-of-the-data-is-around-2430-and-the-75th-percentile-is-around-2448-and-the-25th-percentile-is-around-2390.)

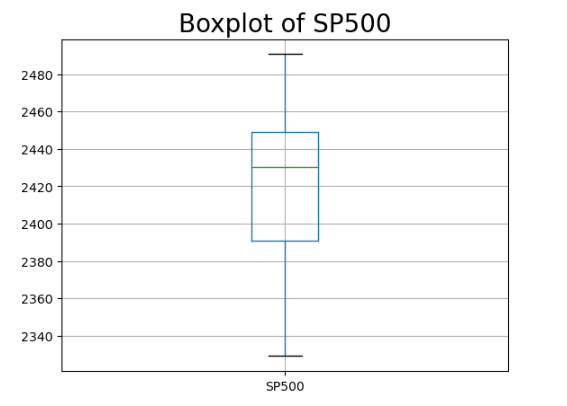
**plot a candlestick chart of the SP500**



**plot an auto-correlation plot of the SP500**



**Display boxplot of SP500 values**



#### From the boxplot, we can get the following information. The max value of the SP500 is more than 2480 and the minimum is less than 2340. There are no outliers in the data. The median of the data is around 2430 and the 75th percentile is around 2448 and the 25th percentile is around 2390.

# Models:

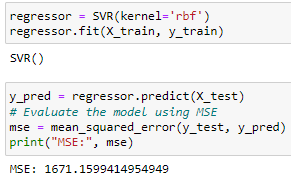
#### apply machine learning models and deep learning to predict SP500 column, i will use

* SVR
* XGBRegressor
* RandomForestRegressor
* CNN
* LSTM

First split data into training data and test data, I am split data into 80% training and 20% test

# Support Vector Regression (SVR) model

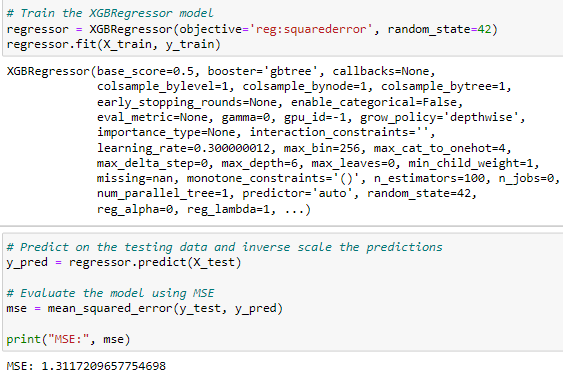
Using SVR with hyperparameter kernel =”rbf”



Mean Square Error on test data is :1671.1599

# XGBRegressor

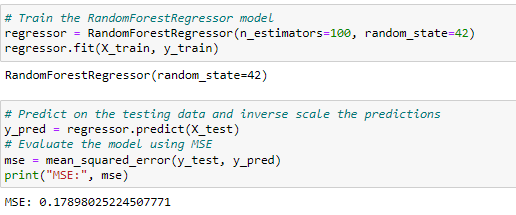
Using XGBRegressor with hyperparameter objective =”reg:squarederror”,random\_state=42



Mean Square Error on test data is :1.3117

# RandomForestRegressor

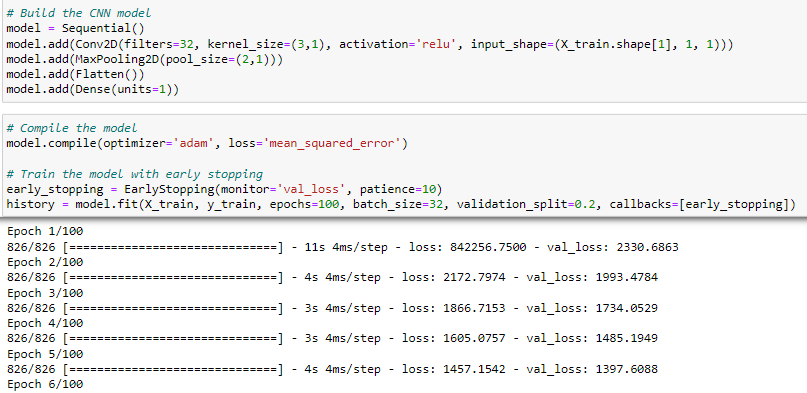
Using RandomForestRegressor with hyperparameter n\_estimator=100 random\_state=42



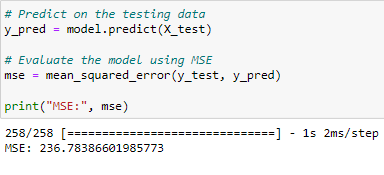
Mean Square Error on test data is :0.178980

**CNN**

training



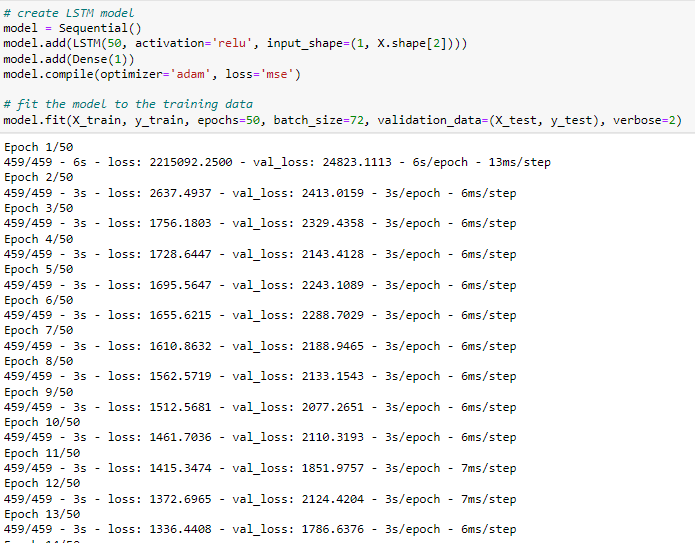
Evaluate



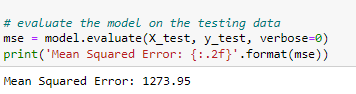
Mean Square Error on test data is :236.7838

**LSTM**

Training



Evaluate



Mean Square Error on test data is :1273.95

**Conclusion:**

## the best model is RandomForestRegressor with MSE: 0.17898025224507771