Stock Analysis

Le Qin

Presentation Structure

- Research Question
- Dataset
- Preliminary Analysis
- Time Series
- Methods
  - Method 1: simple linear regression
  - Method 2: ARIMA
  - Method 3: LSTM (Nope, still researching)
- Reference

Research Question

How (well) can we predict stock prices in future?

Dataset

New York Stock Exchange

https://www.kaggle.com/dgawlik/nyse

date	symbol	open	close	low	high	volume
2016-01-05	WLTW	123.430000	125.839996	122.309998	126.250000	2163600.0
2016-01-06	WLTW	125.239998	119.980003	119.940002	125.540001	2386400.0
2016-01-07	WLTW	116.379997	114.949997	114.930000	119.739998	2489500.0
2016-01-08	WLTW	115.480003	116.620003	113.500000	117.440002	2006300.0
2016-01-11	WLTW	117.010002	114.970001	114.089996	117.330002	1408600.0

Dataset (Google)

- Clean up

```
priceDF.isnull().any()
```

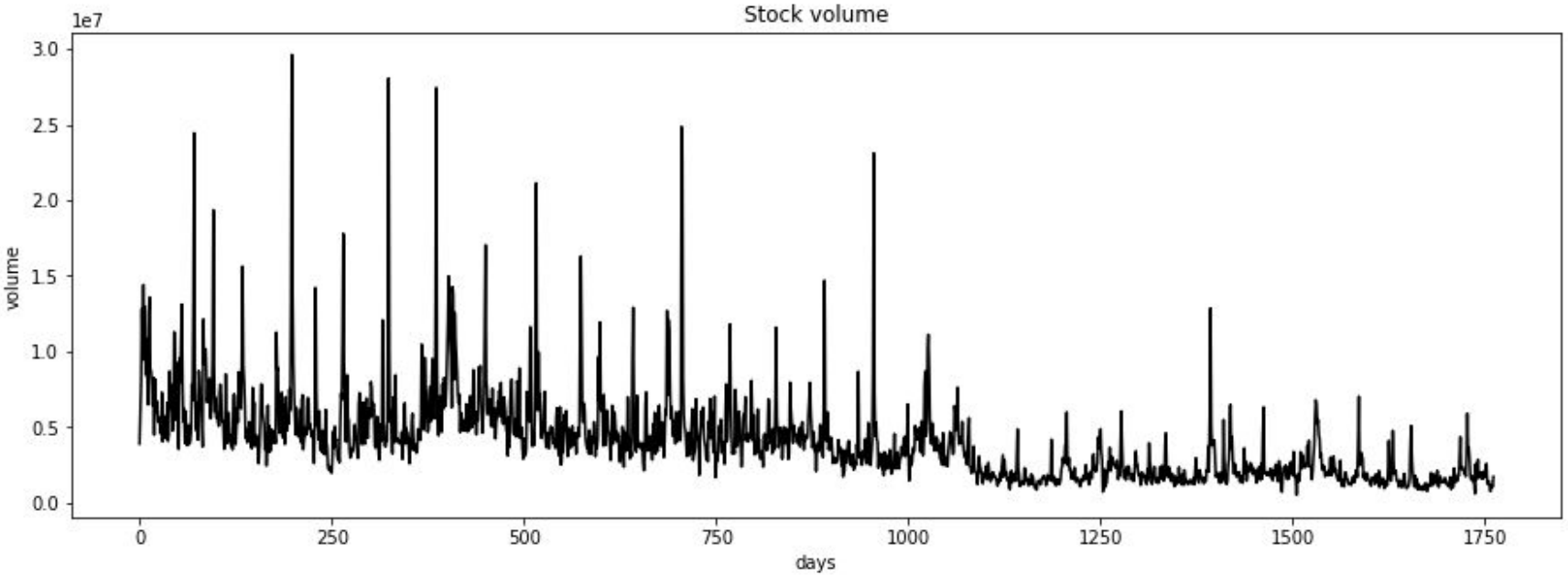
```
symbol    False
open      False
close     False
low       False
high      False
volume    False
dtype: bool
```

- Summary

	open	close	low	high	volume
count	1762.000000	1762.000000	1762.000000	1762.000000	1.762000e+03
mean	467.296599	467.088977	463.037583	471.042921	4.096043e+06
std	181.343840	181.223168	179.767145	182.608562	2.884423e+06
min	219.374377	218.253253	217.032031	221.361361	5.206000e+05
25%	299.672184	299.807316	297.302311	302.253514	2.004075e+06
50%	438.628637	438.786291	436.166174	440.785803	3.670550e+06
75%	587.770005	587.598039	583.287491	591.472856	5.171750e+06
max	838.500000	835.739990	829.039978	839.000000	2.961990e+07

Preliminary Analysis

Preliminary Analysis



Time Series Analysis

- Systematic pattern
  - Trend
    - Smoothing
  - Seasonality
    - Autocorrelation
- Error

Method 1: Simple Linear Regression

date	open	close	low	high	volume	label
2016-12-09	799.299988	809.450012	798.049988	809.950012	1894000.0	807.799988
2016-12-12	804.820007	807.900024	804.530029	811.349976	1627300.0	809.929993
2016-12-13	812.390015	815.340027	811.940002	824.299988	2103300.0	804.570007
2016-12-14	815.919983	817.890015	812.780029	824.260010	1769700.0	802.880005
2016-12-15	817.359985	815.650024	812.000000	823.000000	1768500.0	792.450012
2016-12-16	818.309998	809.840027	808.119995	819.200012	2589100.0	NaN
2016-12-19	809.280029	812.500000	804.500000	816.219971	1259600.0	NaN
2016-12-20	813.369995	815.200012	811.000000	816.489990	1270200.0	NaN
2016-12-21	815.719971	812.200012	805.099976	815.719971	1454500.0	NaN
2016-12-22	809.099976	809.679993	806.030029	811.070007	1131600.0	NaN
2016-12-23	808.010010	807.799988	805.109985	810.969971	764100.0	NaN
2016-12-27	808.679993	809.929993	805.799988	816.000000	974400.0	NaN
2016-12-28	813.330017	804.570007	802.440002	813.330017	1199700.0	NaN
2016-12-29	802.330017	802.880005	798.140015	805.750000	1056500.0	NaN
2016-12-30	803.210022	792.450012	789.619995	803.289978	1728300.0	NaN

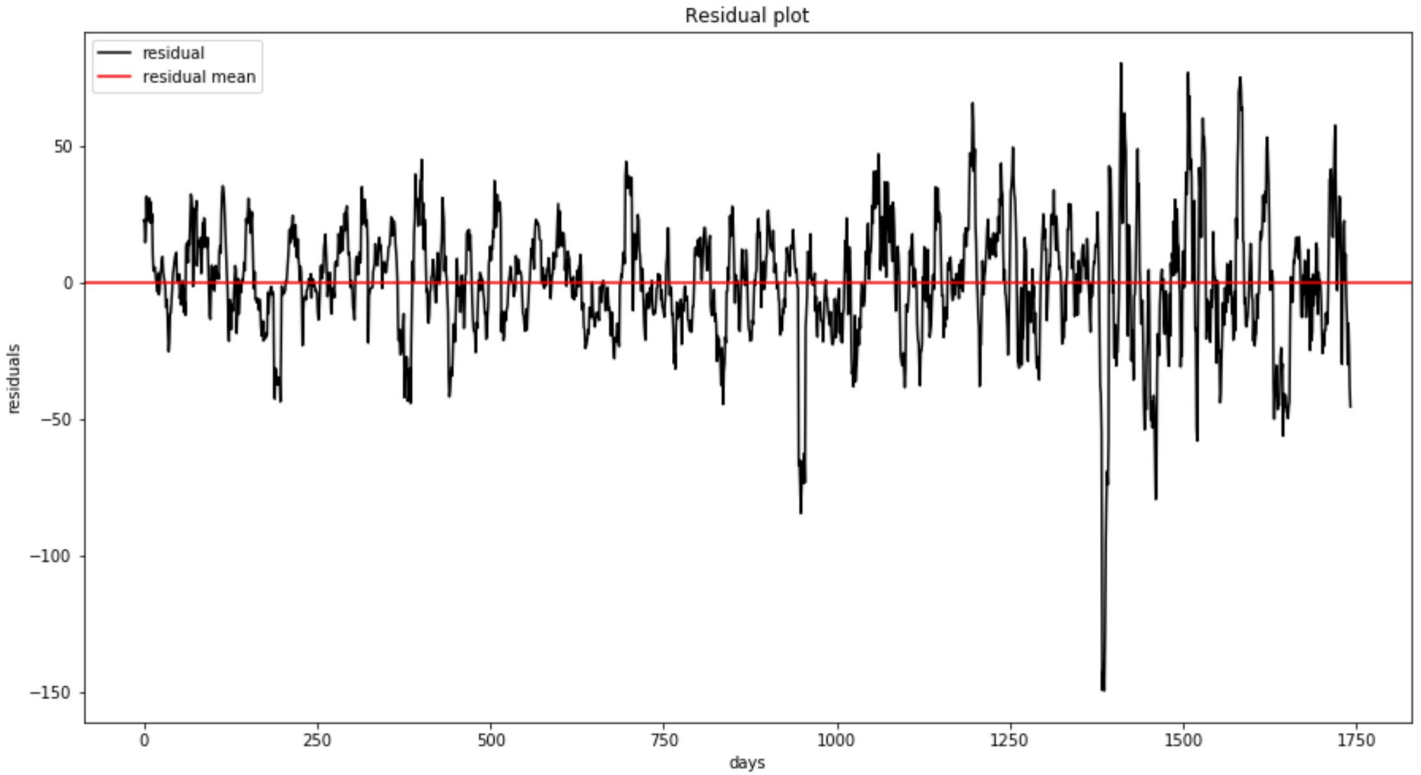
Method 1: Simple Linear Regression

rss

863717.9327

R^2

0.98731



Method 1: Simple Linear Regression

- Prediction

date	label	predict
2016-12-02	809.840027	767.158566
2016-12-05	812.500000	779.774439
2016-12-06	815.200012	779.226163
2016-12-07	812.200012	792.965314
2016-12-08	809.679993	796.514695
2016-12-09	807.799988	810.038786
2016-12-12	809.929993	808.631644
2016-12-13	804.570007	818.488540
2016-12-14	802.880005	819.940604
2016-12-15	792.450012	817.631714

Method 2: ARIMA

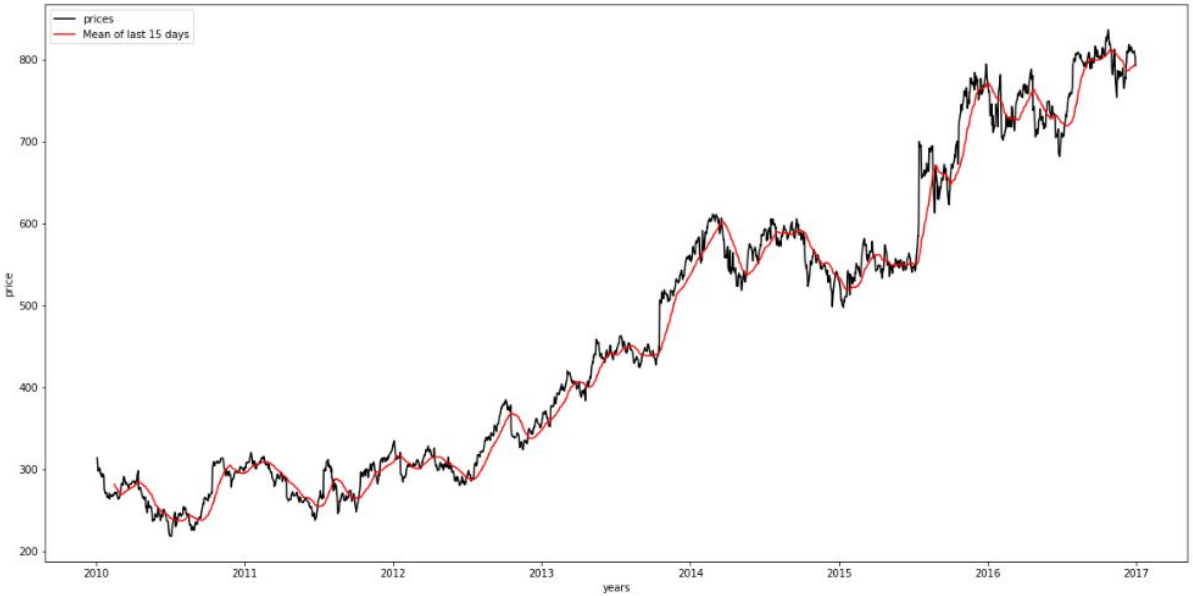
- Autoregressive integrated moving average
- Requirements:
  - Stationarity
- Autoregressive
- Moving average

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

$$\begin{aligned} \overline{p}_{\mathrm{SM}} &= \frac{p_M + p_{M-1} + \cdots + p_{M-(n-1)}}{n} \\ &= \frac{1}{n} \sum_{i=0}^{n-1} p_{M-i} \end{aligned}$$

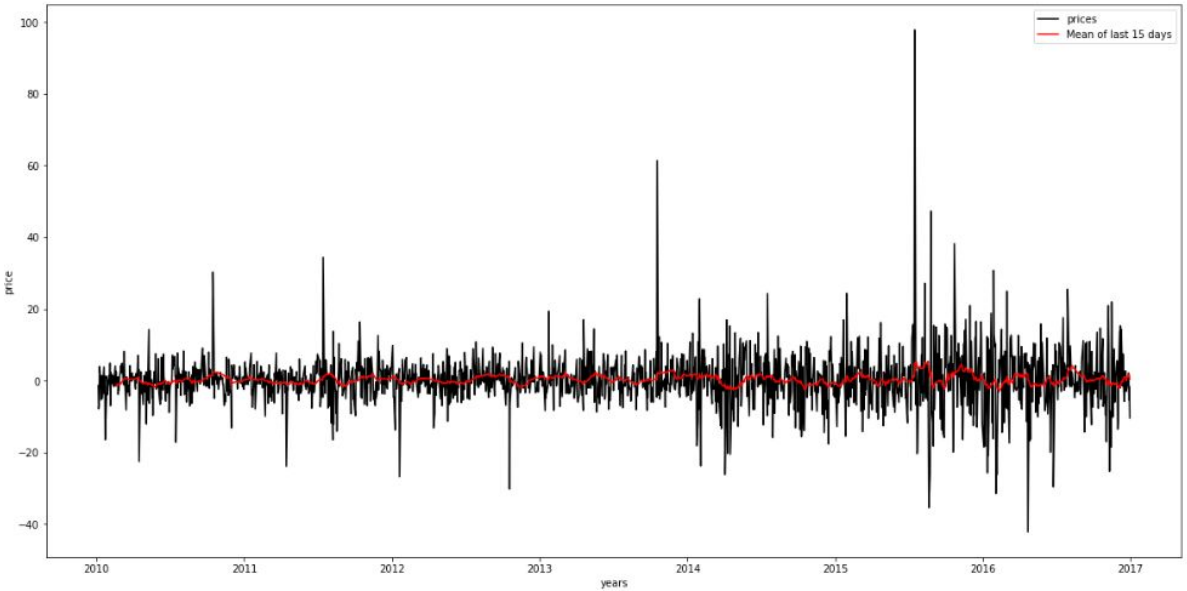
Method 2: ARIMA

```
Dickey-Fuller test results

Test Statistic      -0.131880
p-value              0.946169
# of lags           3.000000
# of obs          1758.000000
dtype: float64
Critical value at 1%: -3.43408
Critical value at 5%: -2.86319
Critical value at 10%: -2.56765
```



Method 2: ARIMA

```
Dickey-Fuller test results

Test Statistic     -25.262277
p-value              0.000000
# of lags           2.000000
# of obs         1758.000000
dtype: float64
Critical value at 1%: -3.43408
Critical value at 5%: -2.86319
Critical value at 10%: -2.56765
```



Method2: ARIMA

ARIMA(0, 0, 0) MSE=98.969

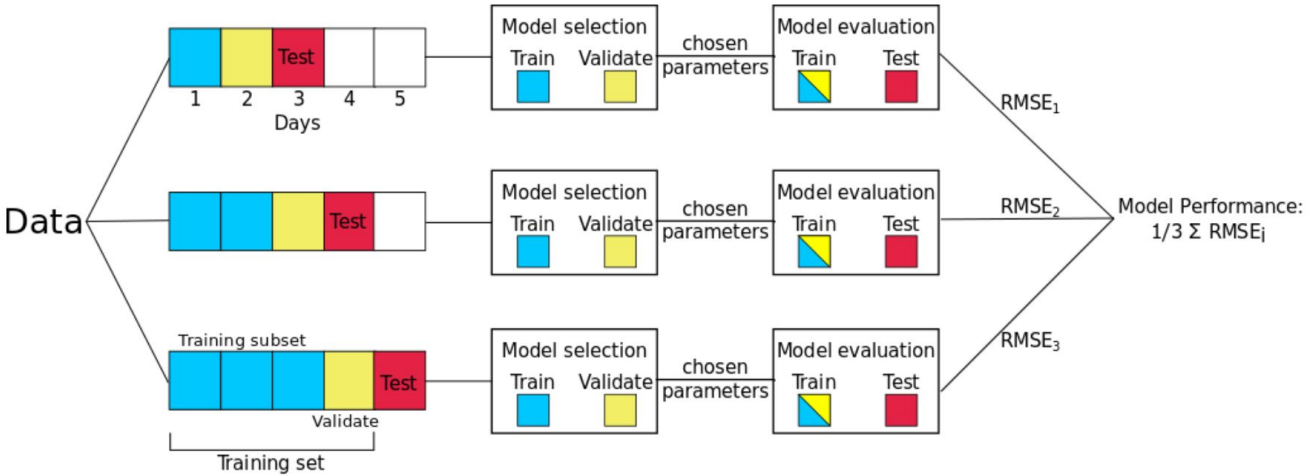
ARIMA(0, 0, 1) MSE=98.803

ARIMA(0, 1, 0) MSE=183.706

ARIMA(0, 1, 1) MSE=98.927

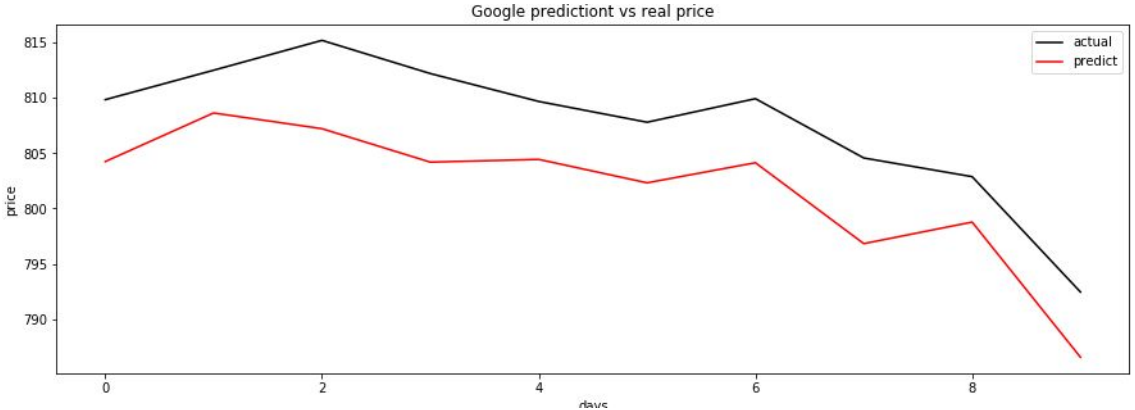
ARIMA(1, 0, 0) MSE=98.805

ARIMA(1, 1, 0) MSE=144.196



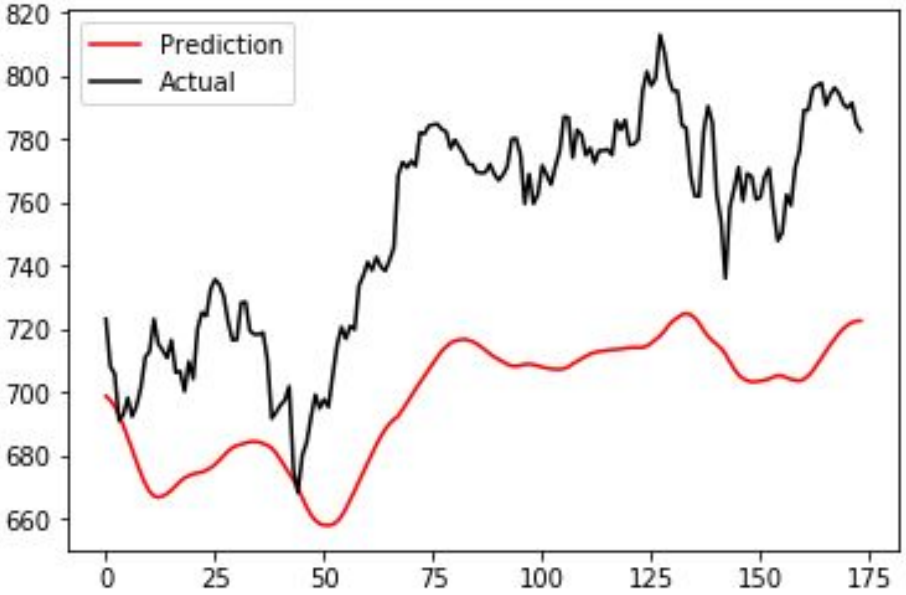
Method2: ARIMA

- Prediction



date	close	predict
2016-12-16	809.840027	804.246481
2016-12-19	812.500000	808.638771
2016-12-20	815.200012	807.217218
2016-12-21	812.200012	804.194317
2016-12-22	809.679993	804.442232
2016-12-23	807.799988	802.325425
2016-12-27	809.929993	804.140819
2016-12-28	804.570007	796.827717
2016-12-29	802.880005	798.775944
2016-12-30	792.450012	786.558157

Method 3: LSTM



Reference

How To Identify Patterns in Time Series Data: Time Series Analysis

http://www.statsoft.com/Textbook/Time-Series-Analysis

NYSE Stock Data - ARIMA Mode

https://www.kaggle.com/ravishankars/nyse-stock-data-arima-model#ARIMA-model-for-NYSE-stock-data

How to Grid Search ARIMA Model Hyperparameters with Python

https://machinelearningmastery.com/grid-search-arima-hyperparameters-with-python/

Time Series Nested Cross-Validation

https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9