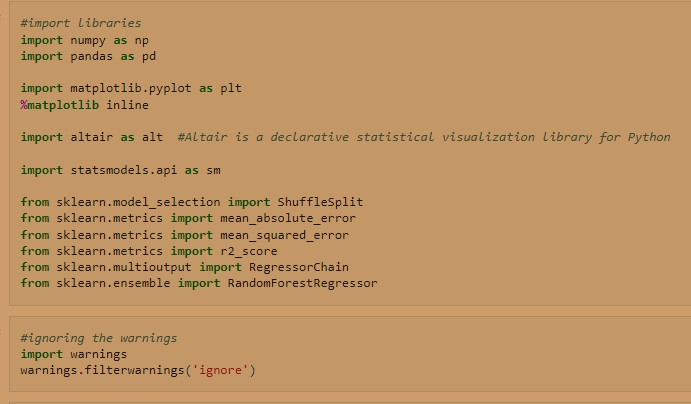
*Investment Predictions*

***Wireframe Documentation***

Homepage

Create a hybrid model for stock price/performance prediction using numerical analysis of historical stock prices and sentimental analysis of news headlines.

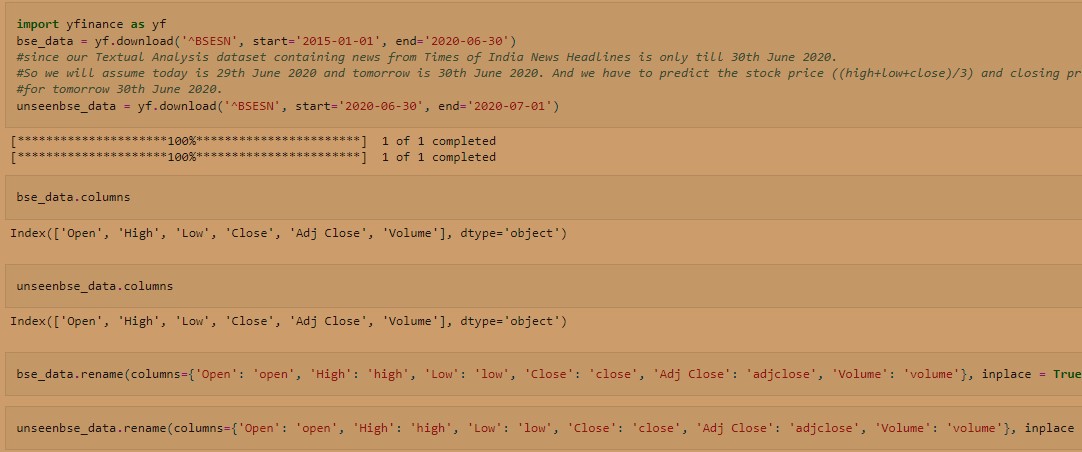
1. First we have imported all libraries like panda, numpy, matplotlib, seaborn. Imported statmodels.api as sm and sklearn metrics. Then we have imported warnings.



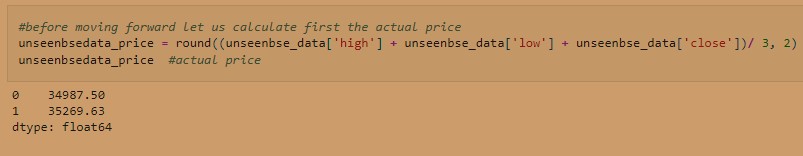
Imported yfinance dataset

#since our Textual Analysis dataset containing news from Times of India News Headlines is only till 30th June 2020.

#So we will assume today is 29th June 2020 and tomorrow is 30th June 2020. And we have to predict the stock price ((high+low+close)/3) and closing price of BSE index for tomorrow 30th June 2020.

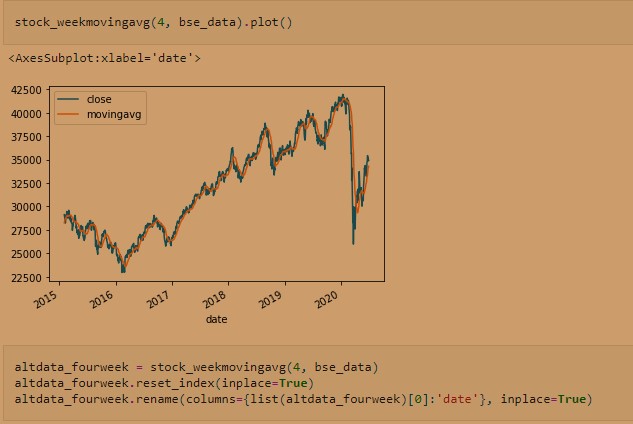


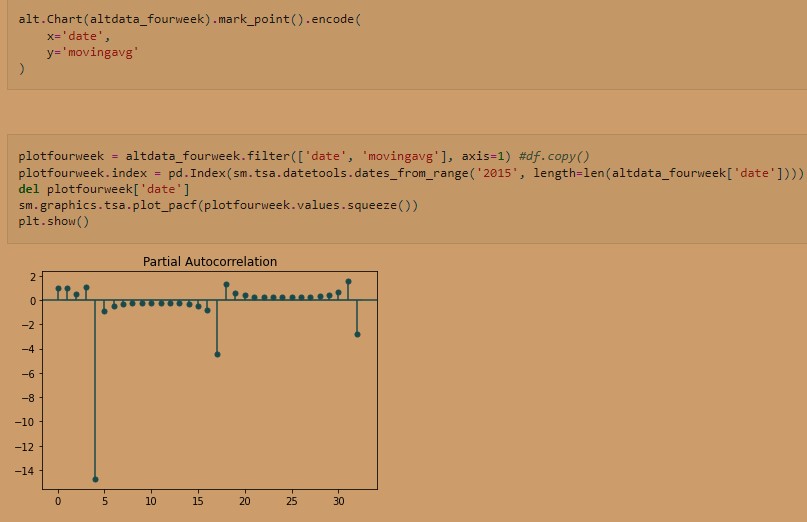
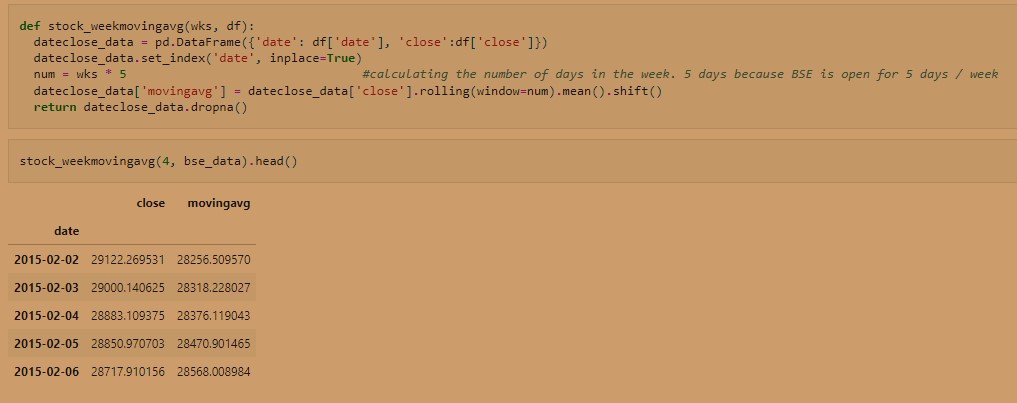
Before moving to rolling window analysis of time series we first calculated the actual price



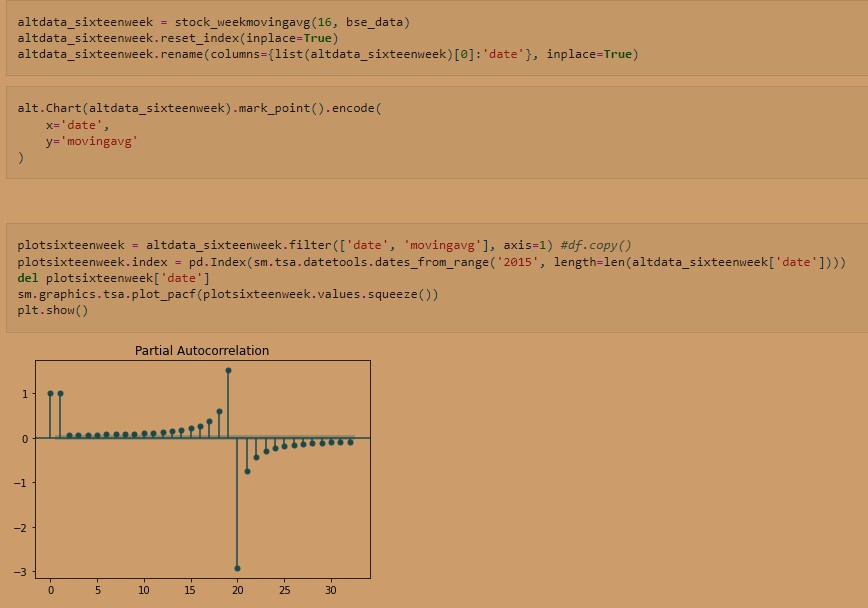
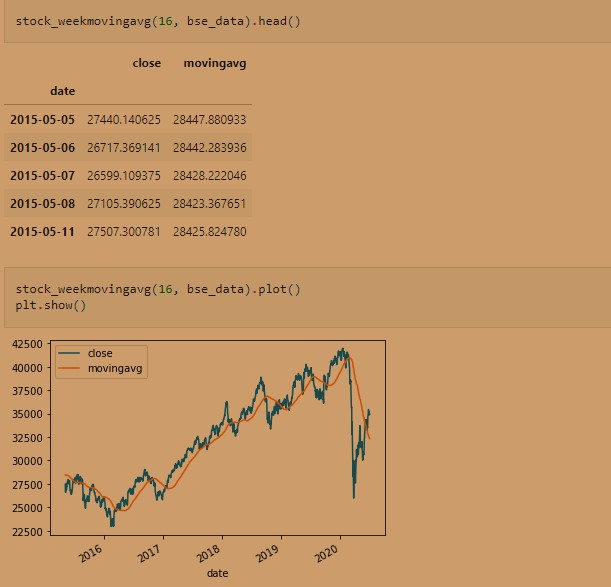
# Rolling window analysis of time series

Creating 4,16, 52 week moving average of closing price of BSE index

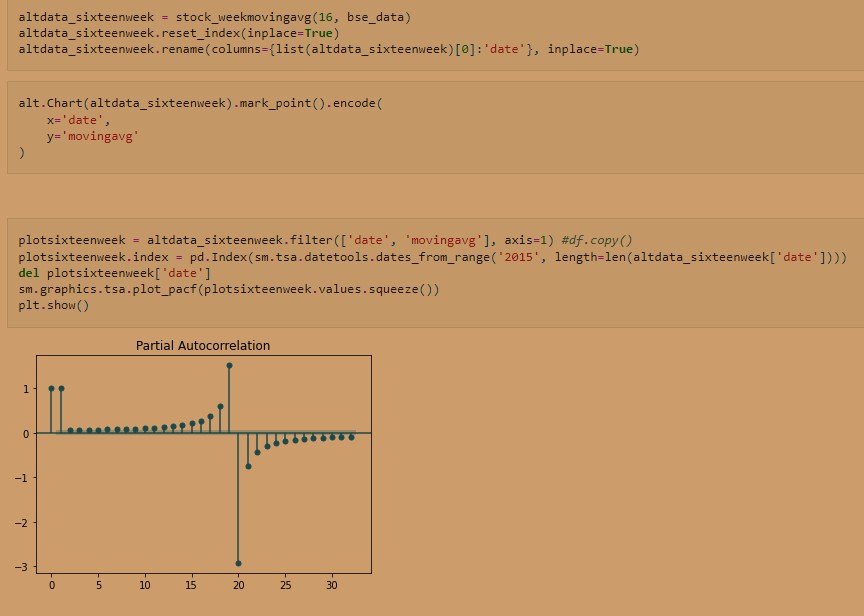
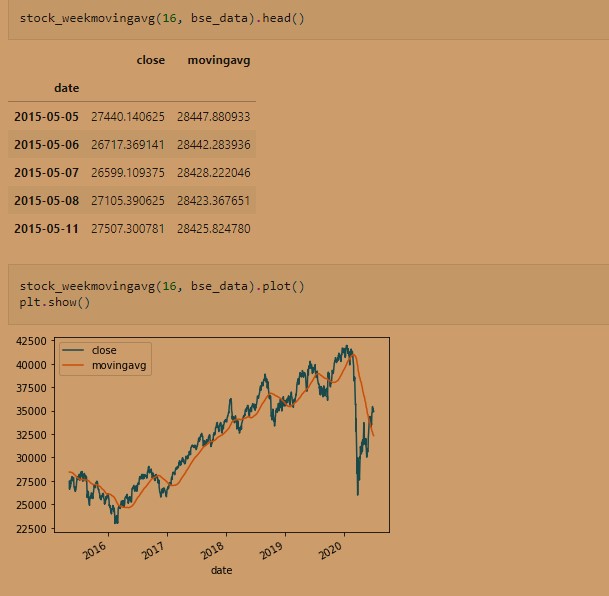




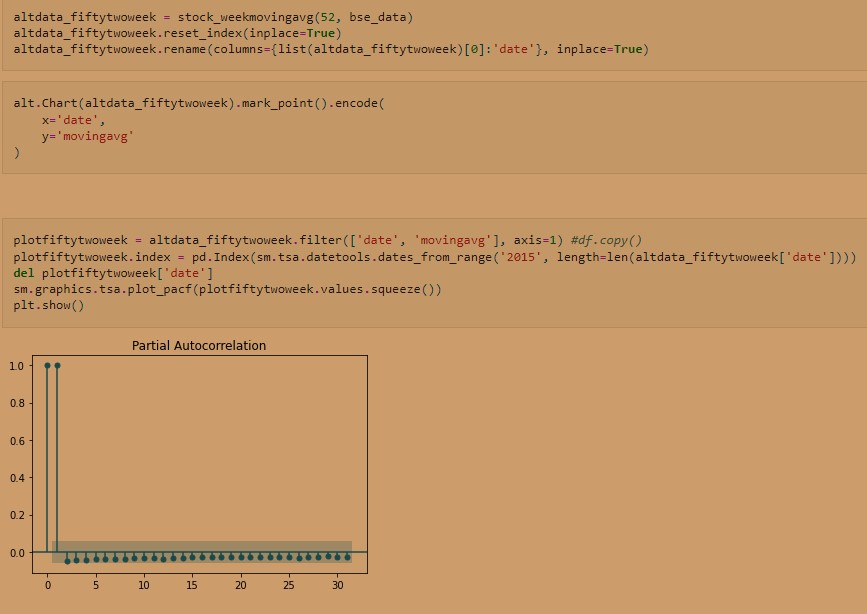
In the partial autocorrelation plot below, we have statistically significant partial autocorrelations at lag values 4 and 32. Since it is less than 0 and more than -1 so 4 and 32 represents a perfect negative correlation. While the rest of values are very close to 0 and under the confidence intervals, which are represented as blue shaded regions (which is not vividly seen in the below plot)

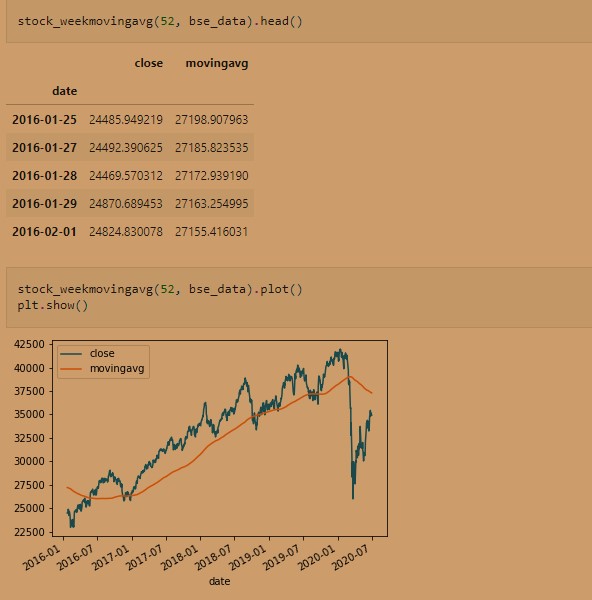


# In the partial autocorrelation plot below, we have statistically significant partial autocorrelations at lag values 0, 1, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28. Where 0, 1, 19 represents a perfect positive correlation and 20 represents a perfect negative correlation. While the rest of values are very close to 0 and under the confidence intervals, which are represented as blue shaded regions (which is not vividly seen in the below

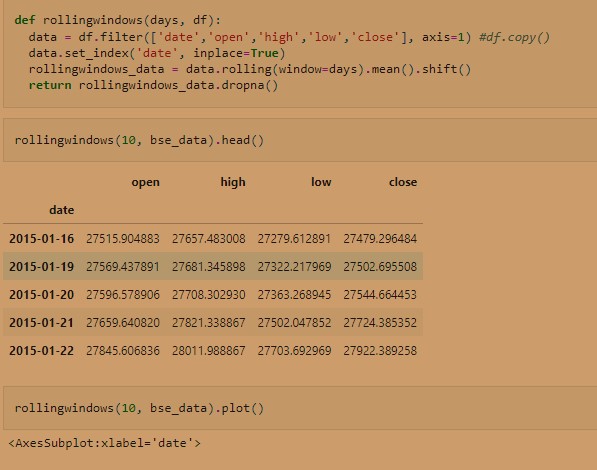
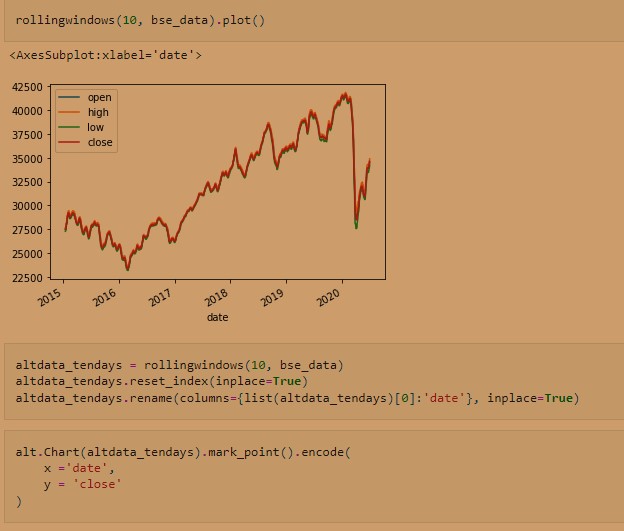


In the partial autocorrelation plot below, we have statistically significant partial autocorrelations at lag values 0, 1 representing a perfect positive correlation. While the rest of values are very close to 0 and under the confidence intervals, which are represented as blue shaded regions





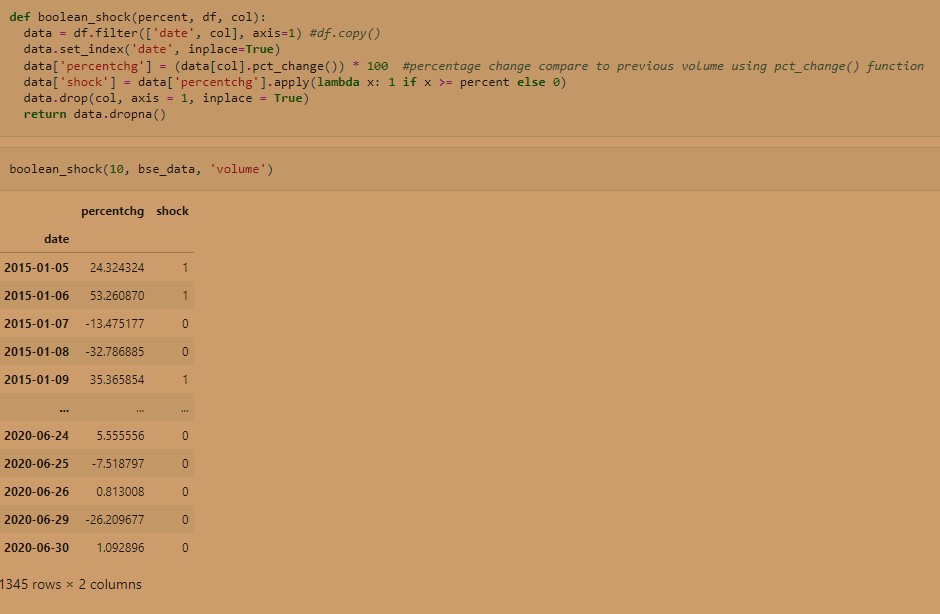
# Creating a rolling window of size 10 and 50 of the BSE index





# Creating the dummy time series:

Volume shocks : we will be creating a 0/1 dummy-coded boolean time series for shock, based on whether volume traded is 10% higher/lower than previous day. ( 0/1 dummy-coding is for direction of shock)

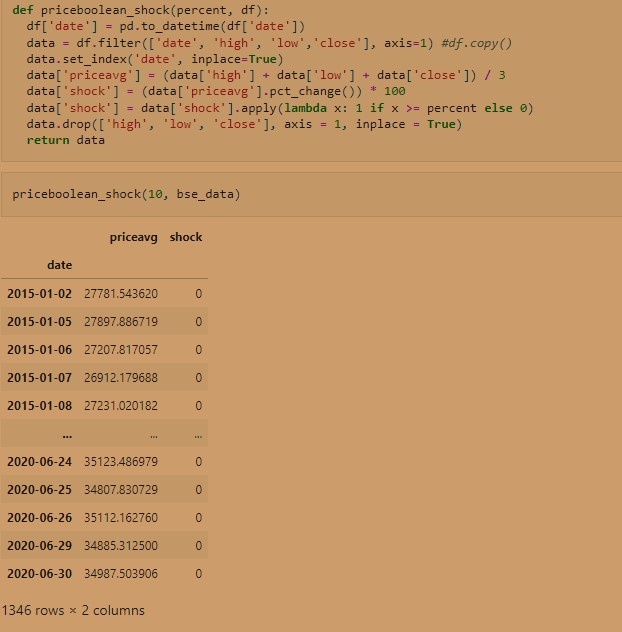


In the partial autocorrelation plot above, we have statistically significant partial autocorrelations at lag values 0, 3, 4, 5, 8, 9. 10, 12, 13, 15, 16, 18, 19, 20, 22, 23, 29, 30, 32. And lag value 0 represents a perfect positive correlation. While the rest of values are very close to 0 and under the confidence intervals, which are represented as blue shaded regions



In the partial autocorrelation plot above, we have statistically significant partial autocorrelations at lag values 0, 5, 6, 7, 10, 11, 24. And lag value 0 represents a perfect positive correlation. While the rest of values are very close to 0 and under the confidence intervals, which are represented as blue shaded regions.

# Pricing shock without volume shock

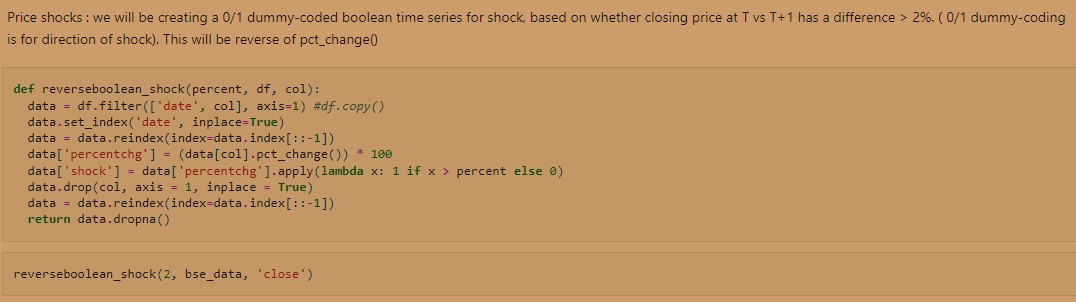


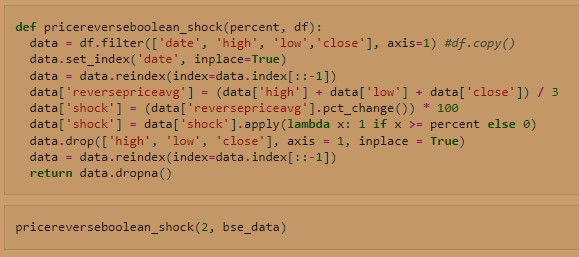


In the partial autocorrelation plot above, we have statistically significant partial autocorrelations at lag values 0, 1, 2, 4, 6, 7, 8, 15, 16, 21, 22, 25, 26. And lag values 0, 1 represents a perfect positive correlation. While the rest of values are very close to 0 and under the confidence intervals, which are represented as blue shaded regions

# Creating the reverse dummy time series:

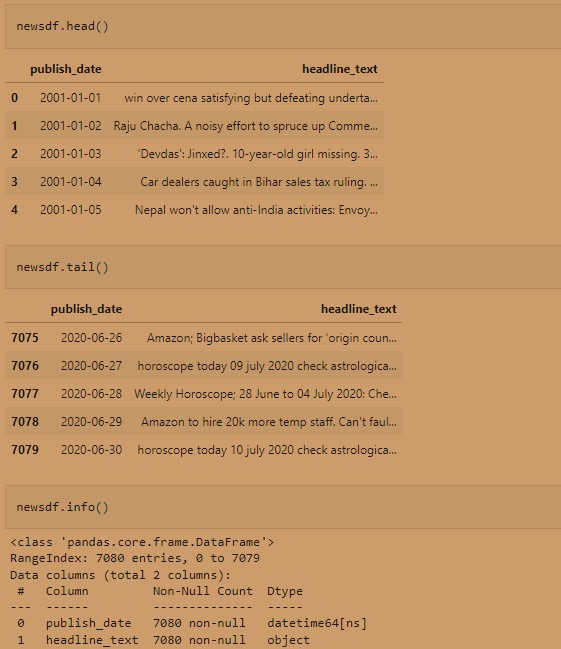
Price shocks : we will be creating a 0/1 dummy-coded boolean time series for shock, based on whether closing price at T vs T+1 has a difference > 2%. ( 0/1 dummy-coding is for direction of shock). This will be reverse of pct\_change()





Pricing shock without volume shock : Now we will be creating a time series for pricing shock without volume shock based on whether price at T vs T+1 has a difference > 2%. ( 0/1 dummy-coding is for direction of shock). This will be reverse of pct\_change()

**Textual Analysis of news from Times of India News Headlines**



We can calculate the sentiment using TextBlob. Based on the polarity, we determine whether it is a positive text or negative or neutral. For TextBlog, if the polarity is more than 0, it is considered positive, if it is less than 0 then it is considered negative and if it ia=s equal to 0 is considered neutral. Subjectivity quantifies the amount of personal opinion and factual information contained in the text. The higher subjectivity means that the text contains personal opinion rather than factual information.

# Preparing the dataset for machine learning

In [132]:

#adding new row for 30th June 2020 (price to be predicted of this day) to main dataset to get average values of all the columns for this day

#taking average because we don't know the values of all the columns for tomorrow so to predict we need average for independent variable.

#We will separate this row later from this main dataset so we can use this as prediction of unseen data for tomorrow.

#And then tally it with actual data from unseenbse\_data dataset which we have downloaded too for 30th June 2020 actual values

