**Title of Project**

**Project-Based Internship 2020 Report**

Submitted

To

**DataRitz Technologies**

**Duration : Six Weeks**

**By**

**TUSHAR KUMAR**

**KAVIN KUMAR SINGH**

**1803210165**

**1803210079**

**ABES ENGINEERING COLLEGE**

**AKTU**

**Under the guidance of:-**

**1)GOPAL GUPTA**

**2)SHASHANK SHEKHAR**

**TABLE OF CONTENTS**

Page No.

DECLARATION ii

CERTIFICATE iii

ACKNOWLEDGEMENTS iv

ABSTRACT v

LIST OF TABLES vii

LIST OF FIGURES viii

**CHAPTER-1 INTRODUCTION 9**

1.1 Problem Definition 9

1.2 Motivation

1.3 Objective of the Project

1.4 Scope of the Project

1.5 Need of Work 9

**CHAPTER 2 RELATED WORK 10**

**CHAPTER 3 PROPOSED METHODOLOGY 11**

3.1 Dataset Description 11

3.2 Methods

3.3 Hardware / Software Requirements 11

3.3 Our Methodology 11

**CHAPTER 4 EXPERIMENT AND RESULT ANALYSIS 17**

**CHAPTER 5 CONCLUSION**

**5.1 Discussion**

**5.2 Future Work**

**REFERENCES**

## CERTIFICATE

This is to certify that Project Report entitled “Stock Price Prediction” which is submitted by TUSHAR KUMAR and KAVIN KUMAR SINGH in partial fulfillment of the requirement for the summer internship of Data Analysis and Machine Learning using Python in Department of Computer Science and Engineering of ABES Engineering College, is a record of the candidate own work carried out by him under my/our supervision.

**Supervisor 1: Gopal Gupta**

**Supervisor 2: Shashank Shekhar**

**Date: 7July, 2020**

ACKNOWLEDGEMENT

*It gives us a great sense of pleasure to present the report of the Project Based Internship 2020 undertaken during B-tech 2nd year. We owe special debt of gratitude to Gopal Gupta/Shashank Shekhar Lead Technical Architect/Project Consultant, DataRitz Technologies for his constant support and guidance throughout the course of our work. His constant motivation have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.*

*We also take the opportunity to acknowledge the contribution of team members of DataRitz Technologies for their full support and assistance during the development of the project.*

*We also do not like to miss the opportunity to acknowledge the motivation of Computer Science and Engineering to provide us the opportunity to undergo training at DataRitz Technologies.*

*Signature: TK*

*Name : Tushar Kumar*

*Roll No : 1803210165*

*Date : 8th July , 2020*

*Signature: KK*

*Name : Kavin Kumar Singh*

*Roll No : 1803210079*

*Date : 8th July , 2020*

**ABSTRACT**

Predicting stock price is always a challenging task. In this project we are trying to predict the next day’s closing price for one hundred sixty different companies individually .The development of our model for analyzing and predicting stock market prices is a basic tool aimed at accelerating the rate of ‘investors’ interest in stock market . The first thing we have taken into account is the dataset which we have acquired from ‘kaggle’ website that contains the stock market prices from previous year. The dataset was pre-processed and tuned up for real analysis. The development and implementation of a stock price prediction is explained in this project and for this purpose we have applied multiple regression algorithm. In most of the economies trading in shares is big business. From the information gathered from their website, it appears that the stock brokers do not have any intelligent tool which can help them advice their clients on which stocks are proper for them to buy or sell. The prevailing methods show a trend on future movement of stocks and not the likely price for any stock in the future. It is therefore preferable to have a tool that does not just point a direction towards price movement, but also indicates the most likely price of the stock itself.

Project Summary

|  |  |
| --- | --- |
| 1. **Region/Unit** | 1. DataRitz Technologies |
| 1. **Location** | 1. Ghaziabad |
| 1. **Program** | 1. DataRitz Technologies.<<programcode>>.<<version>> |
| 1. **Project Number** | 1. DataRitz Technologies. <<projectcode>>.<<version>> |
| 1. **Project Description** | 1. <<project description>> |

Document Control

|  |  |
| --- | --- |
| 1. Prepared by: | 1. Tushar Kumar, Kavin Kumar Singh |
| 1. Title: | Stock Price Prediction |
| 1. College: | ABES Engineering College |
| 1. Department: | 1. Computer Science and Engineering |
| 1. Location: | 1. Ghaziabad |
| 1. Version date: | 1. 8th July 2020 |
| 1. Status: | 1. **<<Initial Draft**/Consultation Draft/Approved Document/Minor Revision/Major Revision>> |

Version history

|  |  |  |  |
| --- | --- | --- | --- |
| **Version no.** | **Date** | **Changed by** | **Nature of amendment** |
| 1. **1.0** |  |  |  |
| 1. **1.1** |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Endorsement and Approval

Project Customer

I approve the business requirements specifications in this document.

|  |  |  |  |
| --- | --- | --- | --- |
| 1. Name | <<customer name>> | | |
| 1. Position | <<customer position>> | | |
| 1. Signature |  | 1. Date |  |

The following officers have **endorsed** this document

Project Sponsor

|  |  |  |  |
| --- | --- | --- | --- |
| 1. Name | 1. <<sponsor name>> | | |
| 1. Position | 1. <<sponsor position>> | | |
| 1. Signature |  | 1. Date |  |

Project Manager (= Component Project Customer)

|  |  |  |  |
| --- | --- | --- | --- |
| 1. Name | 1. Mr. Gopal Gupta / Mr. Shashank Shekhar | | |
| 1. Position | 1. Lead Technical Architect / Project Consultant | | |
| 1. Signature |  | 1. Date |  |

Component Project Sponsor

I accept the business requirements specifications in this document.

|  |  |  |  |
| --- | --- | --- | --- |
| 1. Name | 1. Dr B P Sharma | | |
| 1. Position | 1. Country Head – Delivery | | |
| 1. Signature |  | 1. Date |  |
| 1. **Comments** | | | |
|  | | | |

The following officers have **endorsed** this document

Component Program Manager

|  |  |  |  |
| --- | --- | --- | --- |
| 1. Name | 1. Mr. Gaurav Kansal | | |
| 1. Position | 1. Chief Operating Officer | | |
| 1. Signature |  | 1. Date |  |

**LIST OF TABLES**

**LIST OF FIGURES**

Description of fig pgno.

1. Dataset description showing

no. of rows of data and how 13

much files are there in range.

1. Displaying our raw dataset. 14
2. Displaying the conversion of 15

null values.

1. Displaying our processed data. 16
2. Showing the accuracy of our 18

Machine.

**CHAPTER 1**

**INTRODUCTION**

* 1. **: Problem Definition**

In this project we will try to predict the upcoming closing price of a stock of particular company using data analysis and machine learning.In Stock Market Prediction, the aim is to predict the future value of the financial stocks of a company based on the values of current stock market indices by training on their previous values.

* 1. **Motivation**

In most of the economies trading in shares is a big business. From the information gathered from their website, it appears that the stock brokers do not have any intelligent tool which can help them advice their clients on which stocks are proper for them to buy or sell. The prevailing methods show a trend on future movement of stocks and not the likely price for any stock in the future. Therefore we want to show a way so that they can predict the future price of stock at some extent.

* 1. **Objective of the Project:**

The objective of the proposed work is create a predictive model, to **predict** the future value of the financial **stocks** of a company. It must calculate the estimated **price** of **stock** based on the historical data.

* 1. **Scope of the Project:**

To predict the future stock price we will use the past prices to find out a pattern in which the prices revolving and with the help of that pattern we will find the next stock price.

A pattern can be made between predictor and target values by establishing relations through regressions. By using the historical data we can prepare the input data . The input data is grouped into 2 sets as training data set and testing data set in this model. To estimate the unknown coefficients of the regression equation and to train a model the training data set is used. To predict the future price of a stock, the estimated coefficients are used.

* 1. **Need of Work**

In today’s era money is just not a requirement but a necessity to live your life happily and investing in stocks is a great way to earn money but also a great way to lose it as well because it’s not in someone’s hand due to which we just most of the end up losing money in stock market. It is therefore preferable to have a tool that does not just point a direction towards price movement, but also indicatesthe most likely price of the stock itself to a numerical value that is predicted by regression. Where the target values have been defined already, regression performs operations on a data set .By adding new information the result can be extended. Therefore we are making this machine so that probability of losing money can be reduced.

**CHAPTER 2**

**RELATED WORK**

In the early research related to stock market prediction, Fama, E. F. (1970) proposed the Efficient Market Hypothesis (EMH) and Horne, J. C., & Parker, G. G. (1967) proposed the Random Walk theory. These theories proposed that market prices are affected by information other than historical prices and thus market price cannot be predicted.

The EMH theory suggests that the price of a stock depends completely on market information and thus any new information will lead to a price change as a reaction of the newly released information. This theory also claimed that stocks are always traded on their fair value, where traders cannot buy nor sell stocks in a special price undervalued or inflated and therefore the only way a trader can increase her profits is by increasing her risk. EMH discusses three different variations that affect market price: Weak Form, where only historical data is considered, semi- Strong Form, which incorporates current public data in addition to historical data, and Strong Form, which goes farther to incorporate private data. EMH states that any price movement is either a result of new released information or a random move that would prevent prediction models from success.

The Random Walk Hypothesis by Horne, J. C., & Parker, G. G. (1967) states that the stock prices are randomly changed and argue that past price movements are independent of current movements. This is slightly different from EMH as it focuses on short-term pattern of stock market.

Based on the above two hypotheses by Horne, J. C. et al. (1967) and Fama, E. F. (1970), the stock market will follow a random move and the accuracy of predicting such movement cannot exceeds 50%.

As opposed to these theories, many recent studies have shown that stock market price movement can be predicted to some degree. These studies depend on two different types of financial analysis to predict stock market prices:

• **Fundamental Analysis:** it is based on the health of the company and this includes qualitative and quantitative factors such as interest rate, return on assets, revenues, expenses and price to earnings among others. The aim of this analysis is to check the longterm sustainability and strength of the company for the purpose of long-term investment.

• **Technical analysis:** It is based on time series data. Traders analyze historical price movements and chart patterns and consider time as a crucial parameter in the prediction. Technical analysis can rely on three main keys: stock prices movement although many times the movement seems to be random, historical trends which are assumed to repeat as time passes, and all relevant information about a stock.

In most recent studies, different machine learning techniques have been used to predict stock prices. Machine learning was proven to be a good tool used in price predictions tasks due to the techniques it uses in analyzing data to drawing generalized pattern. Different machine learning models and risk strategies have been applied to stock market prediction task trying to predict mainly the direction of the price for different time frames and using different features that would affect market prices.

Arévalo, A. et al. (2016) used four main features as input to a Depp Neural Network (DNN) model. These features can be considered as technical analysis features for the stock market as they are based on mathematical calculations as described below:

**• Log return**: a finance term that represents the logarithmic difference between the close price at time t and close price at time t-1

**• Pseudo-log-return:** the logarithmic difference between average prices of consecutive minutes

**• Trend Indicator:** a linear model applied on 1- minute tick data to generate a linear equation with a certain slope. A negative slope implies a decrease in the price while a positive slope implies an increase and a slope close to zero implies that the price is almost stable.

Arévalo, A. et al. 2016 formalize the input data as follows: the time feature which is included in the inputs as minutes and hours parameters, and a variable window size (n) which is used for the other inputs. Thus, the input file will include last n pseudo-log-return, last n standard deviations and last n trend indicators. The output of the model was “next one-minute pseudo-log-ret. Then after having the input data file ready, it was given to a DNN with one input layer, five hidden layers and one output layer. The data was fragmented into training and testing data. The model was trained during 50 epochs with different window sizes and the results show that window size 3 can show the best performance of the model with accuracy 66% and 0.07 MSE.

Weng, B. et al. (2017) attempted to predict oneday ahead price movement using disparate sources of data, where combining data from online sources with prices and indicators can enhance the prediction of the stock market state. This study was tested on Apple Inc. (APPL) stock information gathered over 3 years with multiple inputs and different output targets. The target was a binary value (0 or 1) which represent a fall or rise of variation between prices. Four datasets were gathered from disparate sources: first dataset includes the public information available at yahoo finance online for stock prices; second dataset includes number of unique page visits to Wikipedia per visitor per day; third dataset includes count of data published on google related to a company on a specific date; forth dataset includes three technical indicators (Stochastic Oscillator, Larry William, Relative Strength index) that represent the variation of stock price over time. Additional features were generated from the four datasets to provide a meaningful parameter for the model. Twenty features were selected as input. A common observation was drawn, that for any target, all the datasets were represented by at least one feature. Different AI techniques: Artificial Neural Network (ANN), Support Vector Machines (SVM) and Decision Trees (DT) were applied to predict stock price movement and compared to each other. After the evaluation on the three different models listed above, the output comparing open price of day i+1 to open price of day i achieves the best prediction accuracy with around 85% using SVM model.

Schumaker, R. P. et al. (2009) tried to predict direction of the price movement based on financial news. The study was done in 2009 as market prediction was and still facing difficulties due to the ill-defined parameters. In order to use the financial news articles in the prediction model, news should be represented as numerical value. Several techniques have been known to analyze articles related to certain stock to label these articles with sentiments or use them as vectors for the input features. These techniques could be bag of words, noun phrases, named entities and proper nouns. Proper noun technique is a combination of noun phrases and named entities. The proposed technique outperformed other techniques based on a comparison study.

AZFin Text is another system built by (Schumaker, R. P. et al 2009) that predicts price changes after 20 minutes of news release. The main component of this system is the financial news articles collected from yahoo finance and represented as noun phrases; all the collected noun phrases are represented as vector of binary values indicating the presence or absence of a phrase in the article. The second main component of this system is the stock price data collected in oneminute time frame. Then, the final major task after collecting the data and formalizing the inputs was building and training the AI model. To finalize the input of the model, stock price quotation at the same minute news was released, have been added to the input matrix, in addition to that +20 minutes price which will be the output of the system. The data was then fed to different models. Support Vector Regression (SVR) model was built to predict the price after 20 minutes of news release. Only the data during market time was included leaving 1 hour for opening of the market to show the effect of news released during the closure of the market. Moreover, a new constraint was added to the model where only one article could be used for 20 minutes. If two articles were released during the same 20-minute period, both will be discarded. The results show that the average directional accuracy established was 71.18%.

It is evident that released news and published articles affect the market. Most of the existing studies analyzing news rely on shallow features such as bag-of-words, named entities and noun phrases. A newer representation was introduced by (Ding, X. et al. 2014) which represents news as structured events to predict the daily stock price movement. Unlike the previous approaches, this representation can show the relation between events since representing phrases as vectors or bag of words cannot show the actor, action, and the actor which the action was applied on, thus trivial representations cannot show the relation between event and stock. To evaluate the performance of this new representation, news articles data were collected from Reuters and Bloomberg, in addition to the daily close prices of S&P index.

Two different models were built to test the representation: a linear SVM model which have news document as input and +1 or -1 as output indicating increase or decrease in the price for different time frames (1 day, 1 week and 1 month). A non-linear Deep neural network model is also implemented to learn hidden relations between events.

Input features for both linear and nonlinear models were the same: bag-of-words features which use the trivial TFIDF representation after removing stop words and event features represented by different combination of the tuple (𝑜1, P, 𝑜2, 𝑜1+ P, P +𝑜2, 𝑜1+ P + 𝑜2 ) where 𝑜1 is the first object to the left of extracted sentence above and 𝑜2 is the nearest object to the right, and P represents the verb. This feature representation is used to reduce the sparseness of the representation in addition to verb classes.

To evaluate the models, different scenarios were applied. When comparing the results of the models with the bag-of-words articles representation, structured events showed a better performance. From another perspective, when comparing the models, DNN performed better than SVM due to its ability to learn hidden relationships. Moreover, it was distinguished from different timeframes used (1 day, 1 week, 1 month); the shorter the frame the better the results. Thus, the best model was DNN with structured event features for daily prediction with accuracy around 60%.

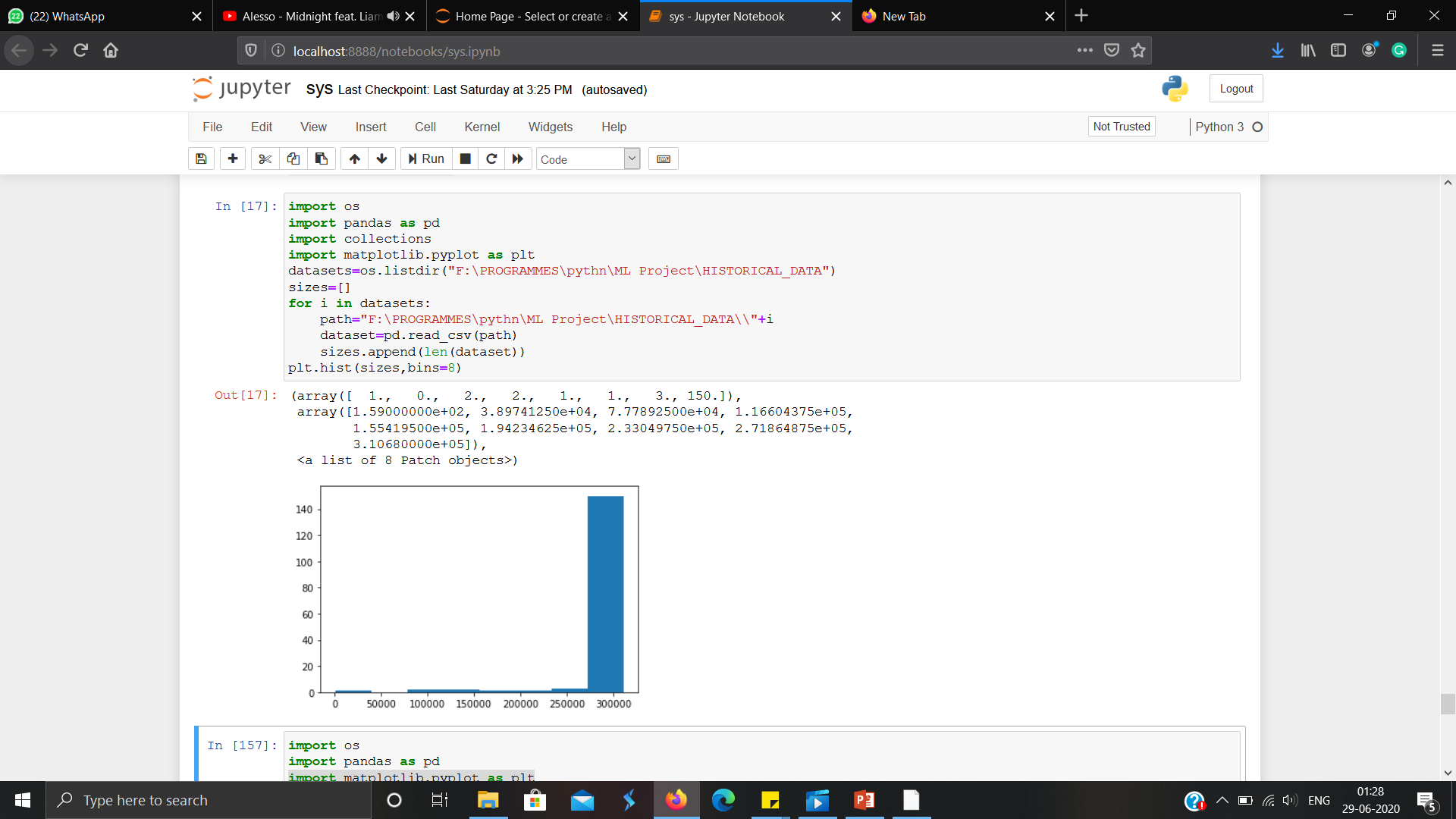
As shown from the above recent studies based on machine learning, stock price movement can be predicted with an accuracy more than 50% which opposed the EMH and Random walk theory using different timeframes, features, and models. In the next section, we detail our proposed prediction models and highlight its improved performance over the existing models.

**CHAPTER 3**

**PROPOSED METHODOLOGY**

* 1. **Dataset Description**

We have downloaded our dataset from Kaggle website. Our dataset is a 3-Dimensional dataset in which the outermost dimensional consist of 160 files of .csv file format. Each file is named after the company whose data it contains. That means our dataset contains data of 160 different companies. There are approx. 3lakhs of rows in each file which from Jan-2017 till Mar-2020 and the time duration of each data is 1minute meaning each data present in a row is recorded after a duration of 1 minute.



In this diagram you can see how is the files are distributed in terms of size. And by this diagram only we can also state that more than 140 files (>90% of files) contains around 3lakhs of rows.

Now there are 6 columns in each of these files which are Timestamp, High, Low, Open, Close, Volume. Description of each of these columns are as follows

**Timestamp** - date when was stock price this one

time in minutes for the stock

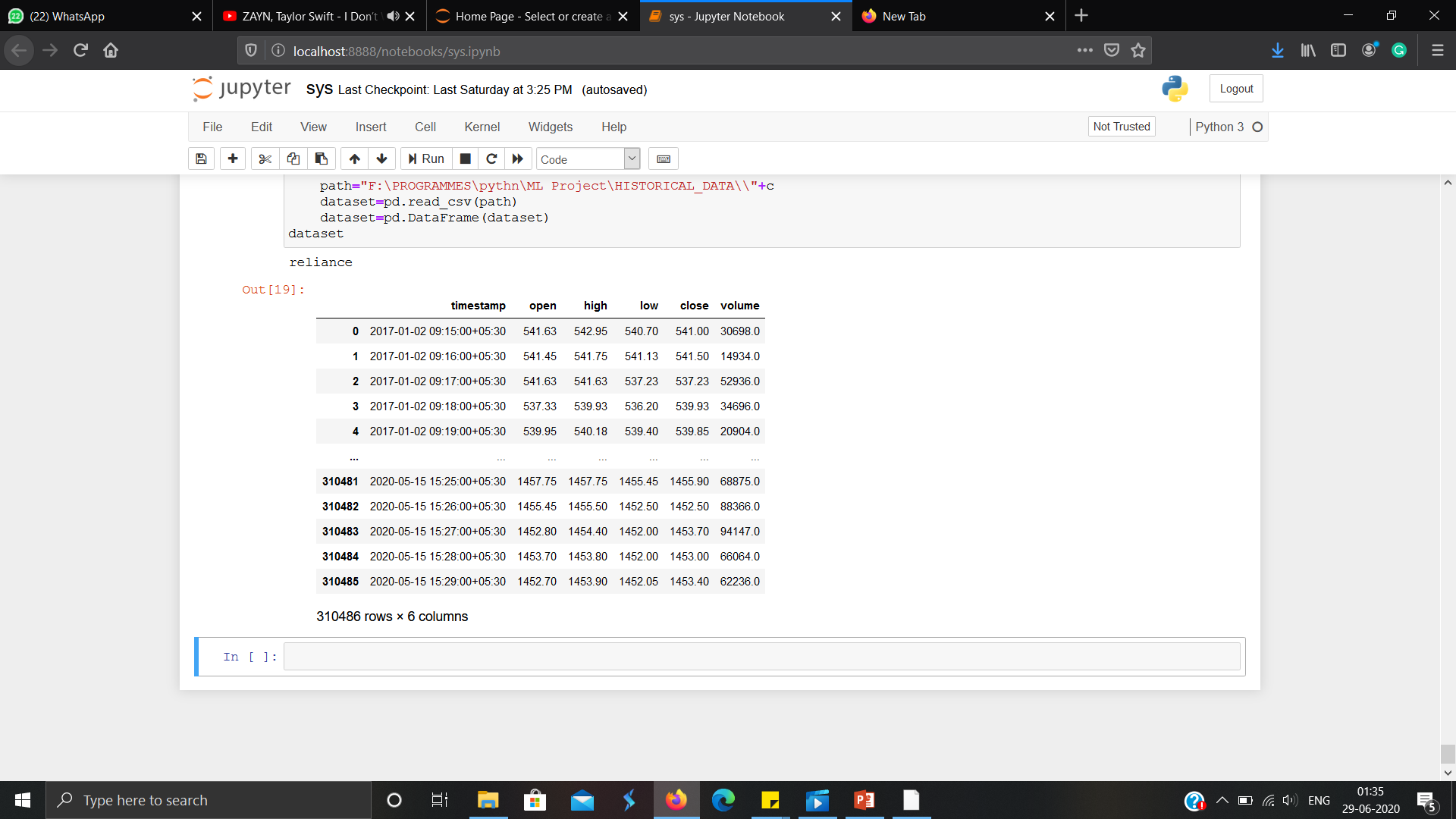
**High** - Means the highest price in a given period of time.

**Low** - Means the lowest price in a given period of time.

**Open** - in a stock market refer to the opening price for the stock

**Close** - the closing price of the stock.

**Volume** - the total number of people buying the stock.



This is the dataset of reliance.

* 1. **Methods**

In this we will be using various python libraries for manipulating data and performing various mathematical calculations on it.

Firstly we will analysis the dataset that what we will be using for prediction and data cleaning consisting of removing of unrequired columns, null values, etc then preprocessing means manipulating data according to our needs and lastly applying the algorithm.

* 1. **Hardware / Software Requirements**

**Hardware:**

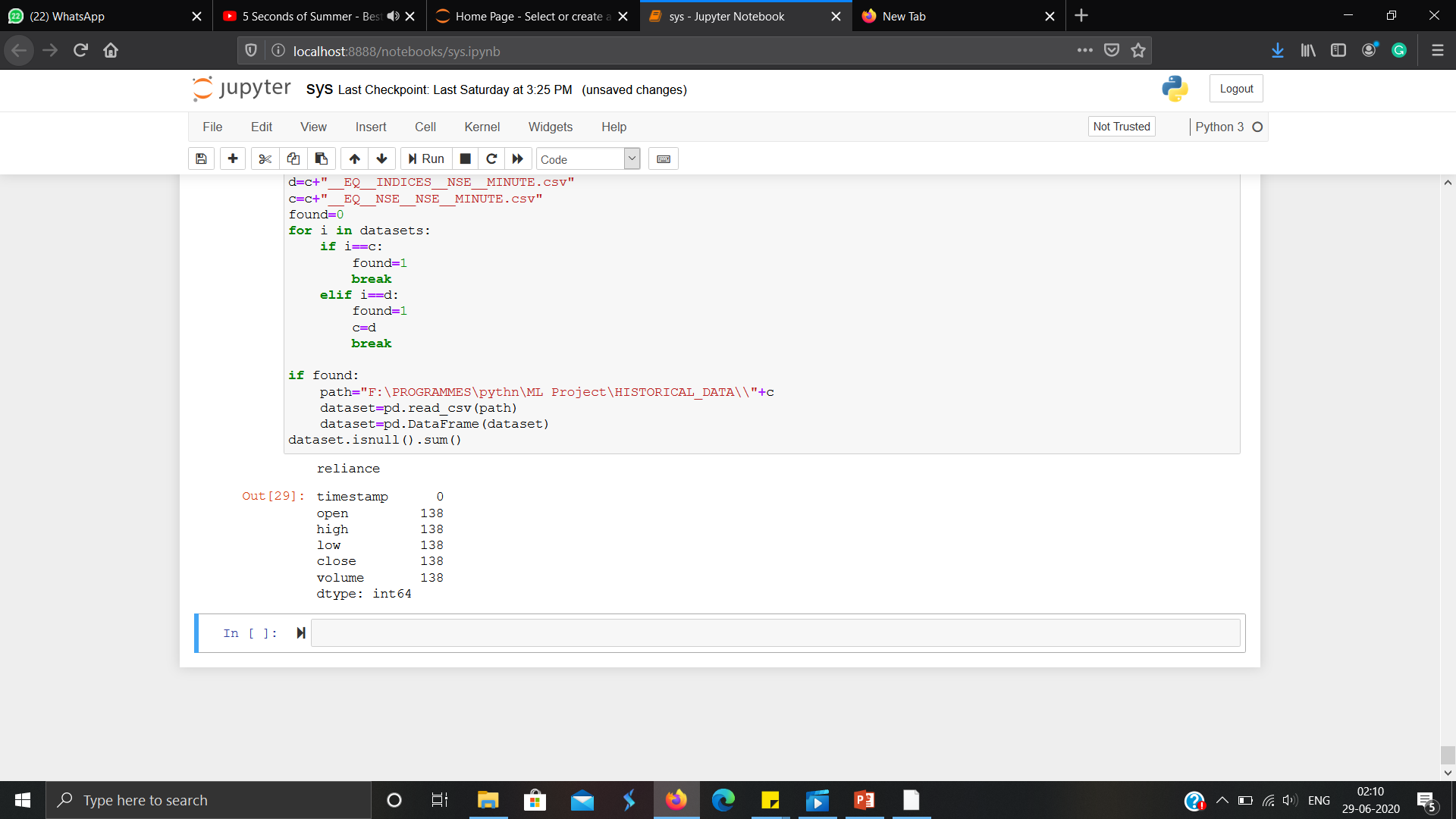
* Memory and disk space required per user: 1GB RAM + 1GB of disk + 5 CPU core.
* Server overhead: 2-4GB or 10% system overhead (whatever is larger), . 5 CPU cores, 10GB disk space.

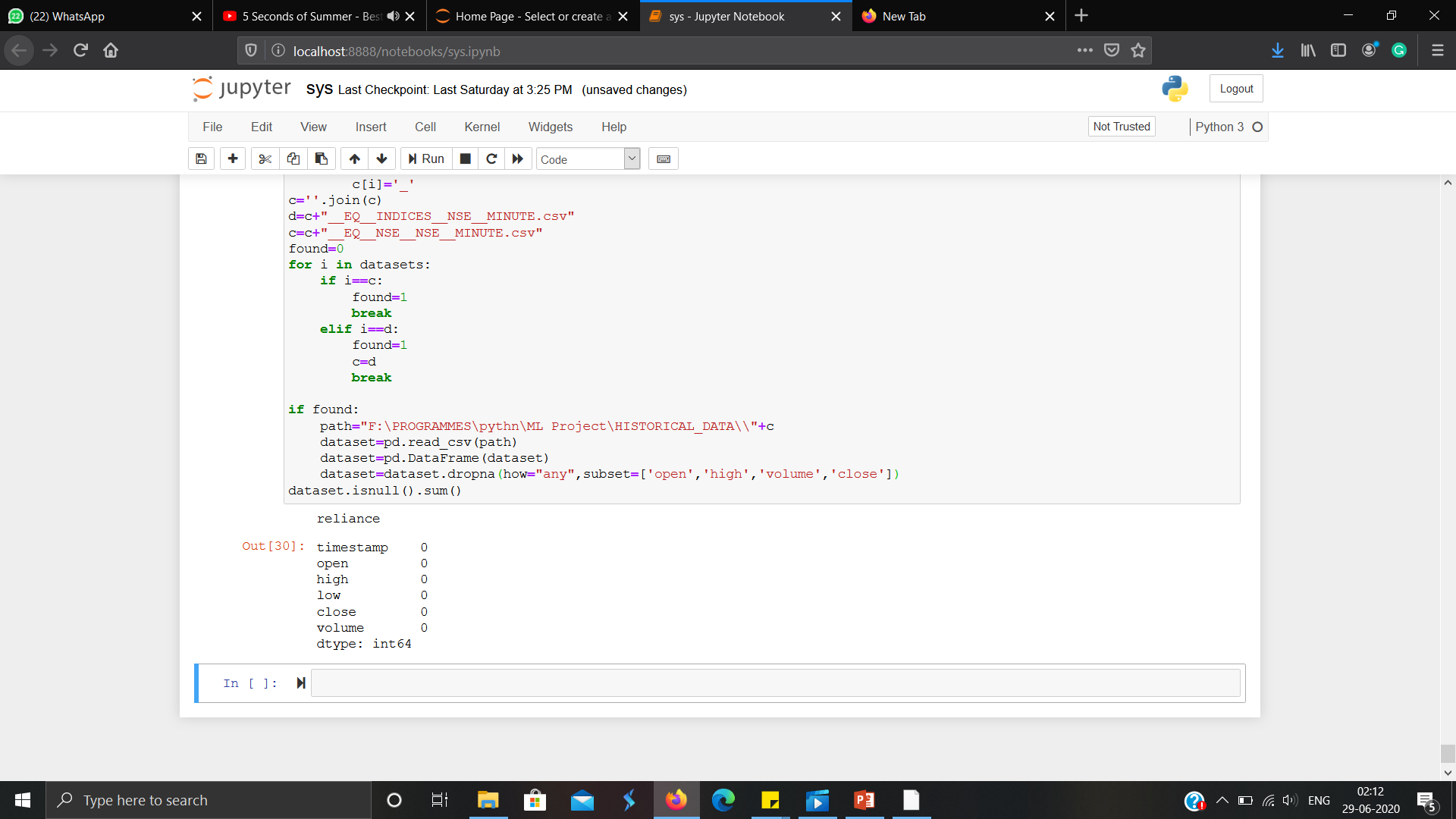
**Software:**

* Jupyter Notebook.
* Chrome/Firefox.

* 1. **Our Methodology**
* **Data Cleaning:**

In this we will remove all the null values, duplicates and unwanted columns from our data because these values can eventually through errors.





Here you can see in first list there are 138 null values and then we removed them.

* **Data Preprocessing:**

Here we made a new column of date from timestamp. We did so because we want to predict next price of next day and we have dataset of mintues. To do so we need to group all the rows of same date together here we use this new column.

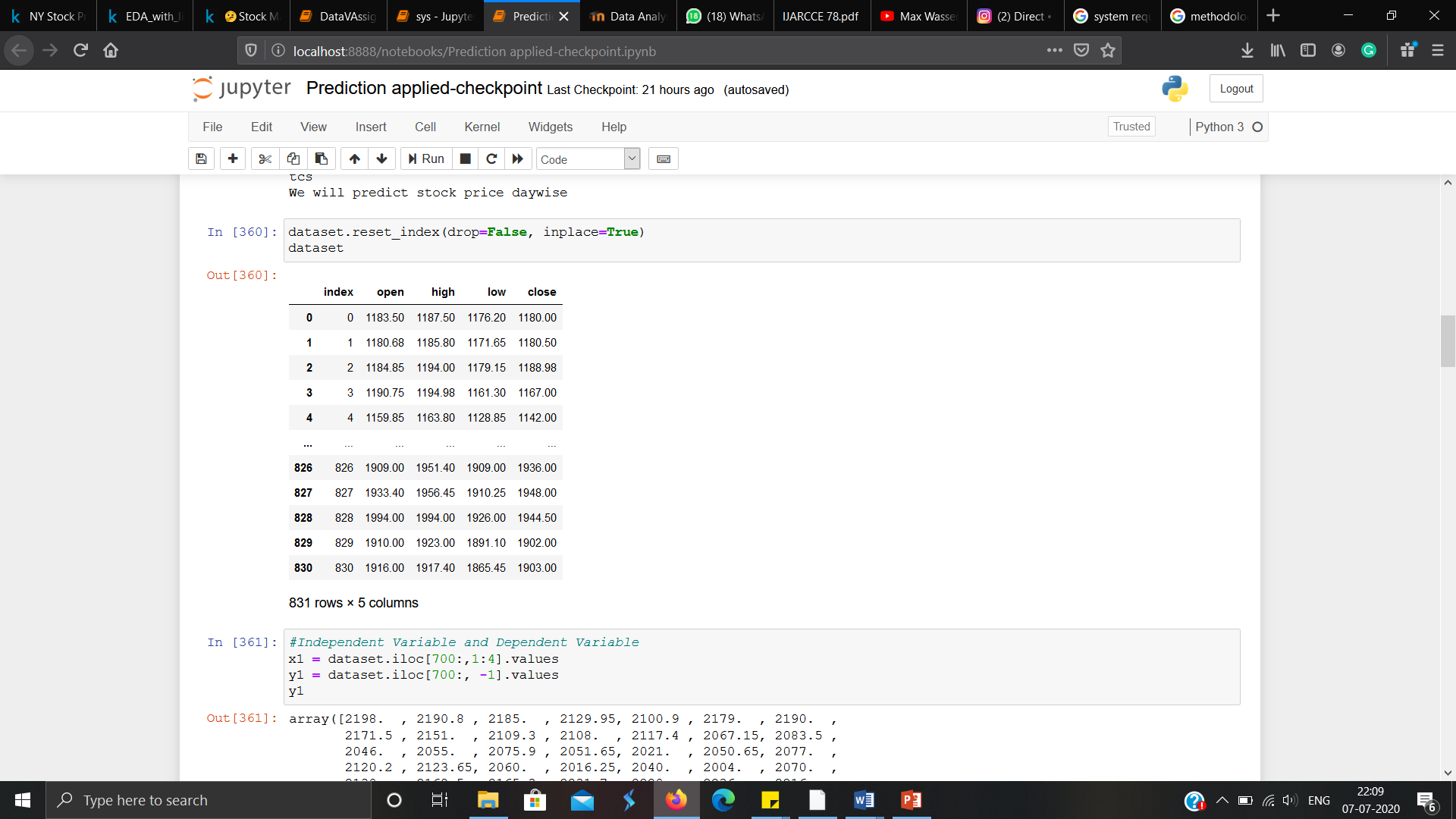
After grouping we take the first price of open column

Last price of close column

Highest price of high column

And lowest price of lowest column.

Reason behind this is as open is the first one meaning on that day the stock price opens. Close is last means on that day stock close on that price and so on.And finally we added index column representing as days.



For this we have taken close as dependent variable and open high and low as independent variables.

The predicted value of Y is a linear transformation of the X variables. Multiple Regression Equation having “K” independent variables is given by:

Y= b0+ b1X1+ b2X2+ b3X3+...... bnXn

For three independent variables, the prediction of Y is expressed by the following equation:

Y= b0+ b1X1+ b2X2+ b3X3

In our case,

X1= High value of a stock of previous day,

X2=Low value of a stock of previous day,

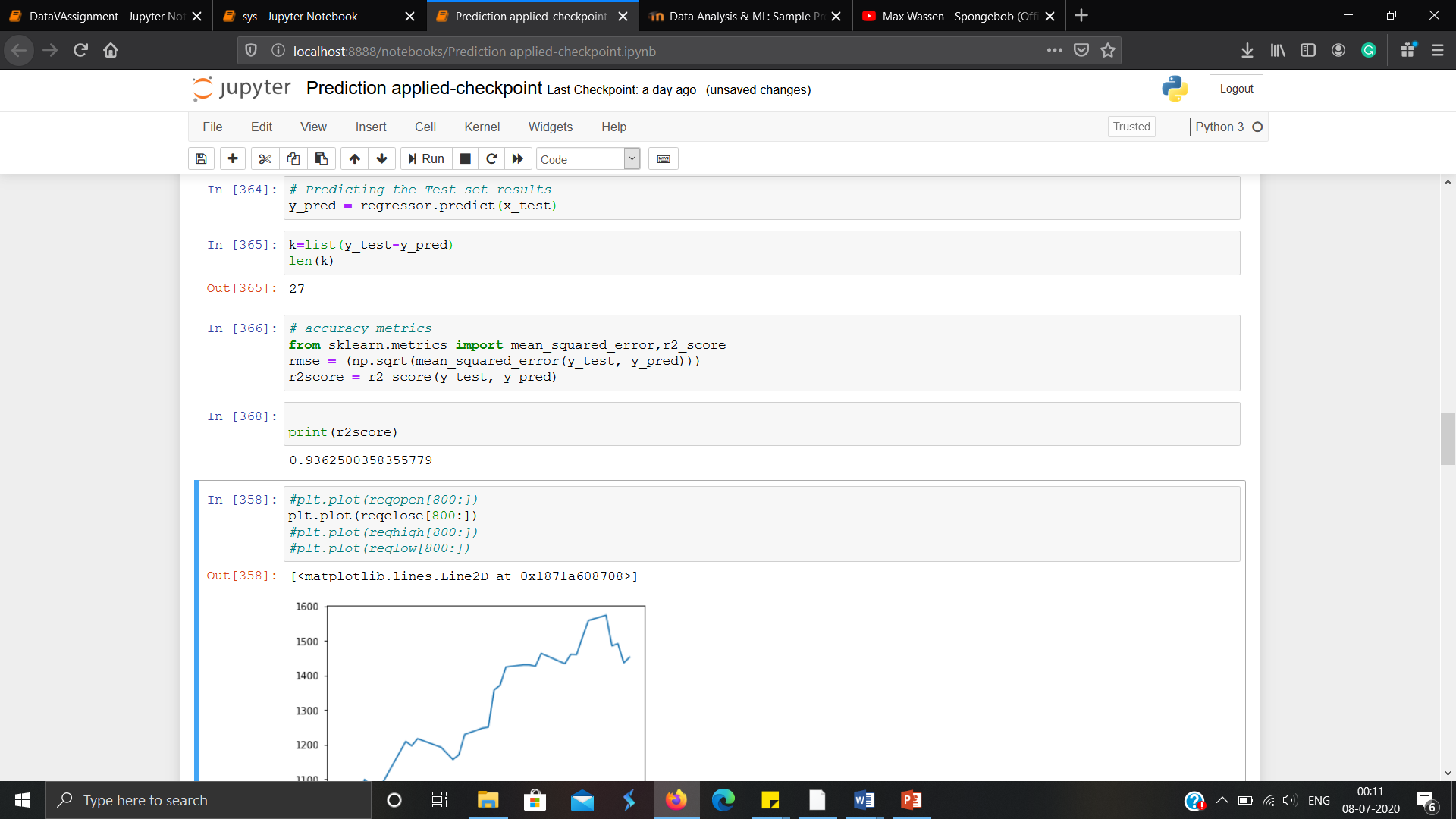
X3=Close value of a stock of previous day.

**CHAPTER 4**

**EXPERIMENT AND RESULT ANALYSIS**

We divide our data into two part one to train our machine and other for the test. For training we have given 80% of our data to our model and after that 20% is used for testing the result.

After training when we applied multiple regression on our model on test data to get some predictions and when we compare that predicted data from from the original data we found that our accuracy is revolving around 93%-98%.



**CHAPTER 5**

**CONCLUSION**

* 1. **Discussion**

In this paper we studied some well-known prediction algorithms concerned with regression . The accuracy yielded by them and the various parameters used for prediction have been stated.It is observed that prediction using multiple regression yields better results than linear regression. Further the use of neural networks for prediction sounds to be a promising field in the future and can be used for real time trading in stock market. We even learn more about data analysis how to manipulate data and handle different types of files and about python.

* 1. **Future Work**

The efficiency of this model can be improved by incorporating the concept of Artificial Neural Network to account for the dynamic factors like present day happenings around the world.

**REFERENCES**

<https://www.kaggle.com/yedukrishnansm/eda-with-linear-regression-multiple-datasets>

<https://www.kaggle.com/spidy20/stock-market-prediction-with-linear-regression/>

<https://www.kaggle.com/raoulma/ny-stock-price-prediction-rnn-lstm-gru/comments#5.-Predictions->

<https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html>

<https://www.tutorialspoint.com/python_pandas/python_pandas_groupby.htm>

<https://www.geeksforgeeks.org/pandas-groupby/>

<https://github.com/>