Tyrone Ekhator

Machine Learning COM 624

Introduction

All businesses depend on data to function. Making decisions based on data-driven insights might be the difference between staying competitive or falling farther behind. In order to harness the value of corporate and customer data and make decisions that keep a business ahead of the competition, machine learning may be the answer (NetApp,2023). Now, the question is what is machine learning? Machine learning is a subfield of artificial intelligence, which is broadly defined as the capability of a machine to imitate intelligent human behavior (NetApp,2023). Moreover, the model that I used for my project is the LTSM model that basically uses neural networks to predict future share prices in the stock dataset. This report shall be discussing the problem definition; how the data was collected; understanding the machine Learning data via visualization; the preprocessing techniques used and the explanation and evaluation of the model.

Problem Definition

Whenever one wants to solve a problem in Artificial intelligence, it must have a proper problem definition in order to understand how to go about it. Additionally, an abhorrent problem definition can lead to being unable to create a proper model to solve the problem at hand.

Now, the task given by the instructor is to create an AI that can predict future stock prices for multiple companies and can also indicate to the user the worth of each stock in specific time period (for instance, ones ROI for investing 50 pounds in a month). Furthermore, the probability of success at the initial point is quite high reason being the AI can use past share price trends as training data in order to predict future shares.

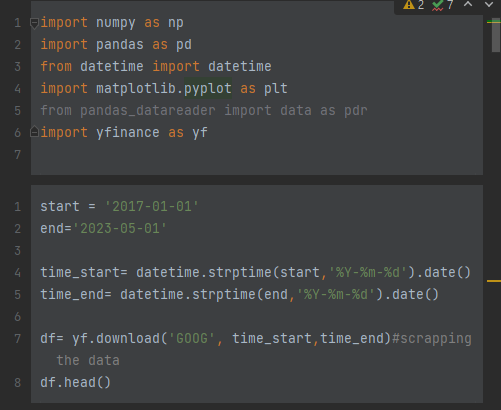
Collecting appropriate data for your ML Problem

This section of the report explains the definition of data collection; how the dataset was gathered, techniques used; quality of the dataset and the dimension of the dataset.

Data collection is the process of capturing past records of events in order to find recurring patterns and from those patterns one can build predictive models based on the collection of the past data (DataRobot,2023).

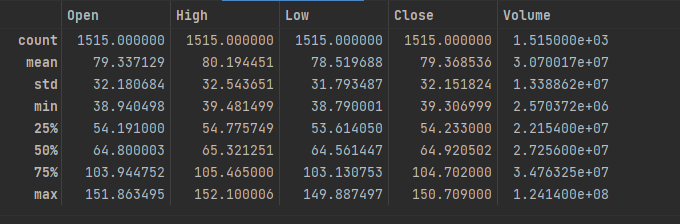
## Now, the dataset was gathered from yahoo finance which is reliable source to get stock and financial reports in real time.

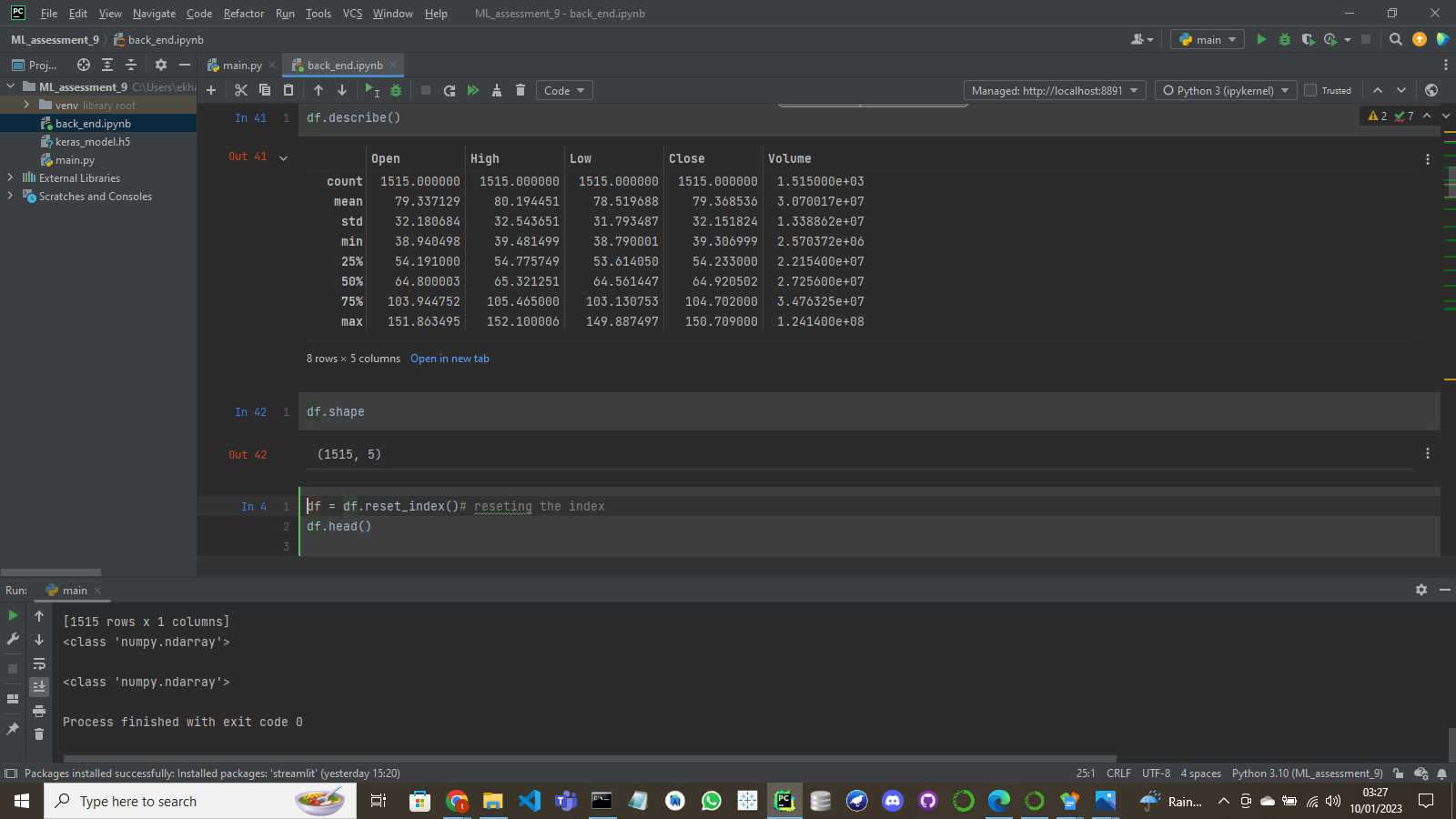
## Moreover, the data gathering technique used to get the stock data from yahoo finance is known as the automated data collection technique where one gets data from online resources (like yahoo finance); with the web crawling method allowing one to get specific data (like a stock dataset form yahoo finance).

Additionally, looking at the code snippet below in order to web crawl for the data the code firstly; imports yahoo finance and then simply downloads the dataset within a specific date range (and yes yahoo finance provides real time data from the stock market). Then the code simply displays the first five rows of the dataset.

Features of the dataset

In order to know the characteristics of the dataset all one has to do is the ‘data.describe()’ and data.shape() function.





understanding the machine Learning data

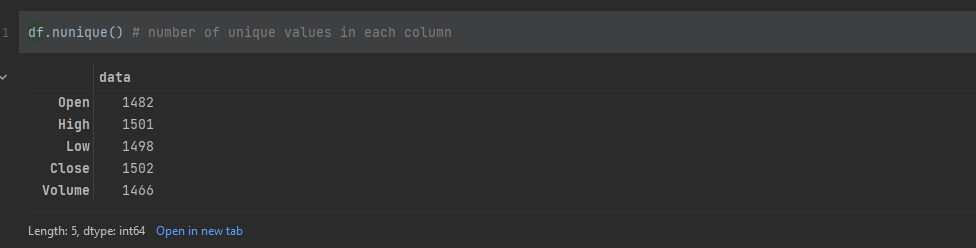
In order to understand the machine learning data, one must use the process known as data exploration.

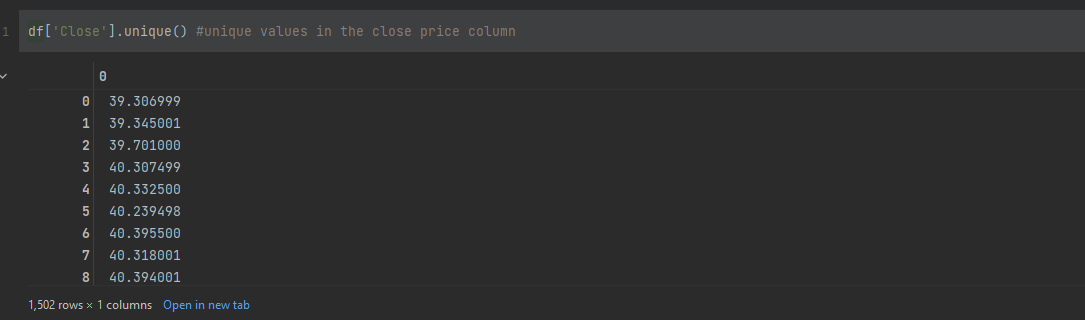
**Data exploration is the act of surveying and visualizing data in order to see trends and relationships in the dataset. Furthermore, the reason why data exploration is so important is because it aids users to make better decisions with their data and businesses better understanding of the next move after exploring the data.**

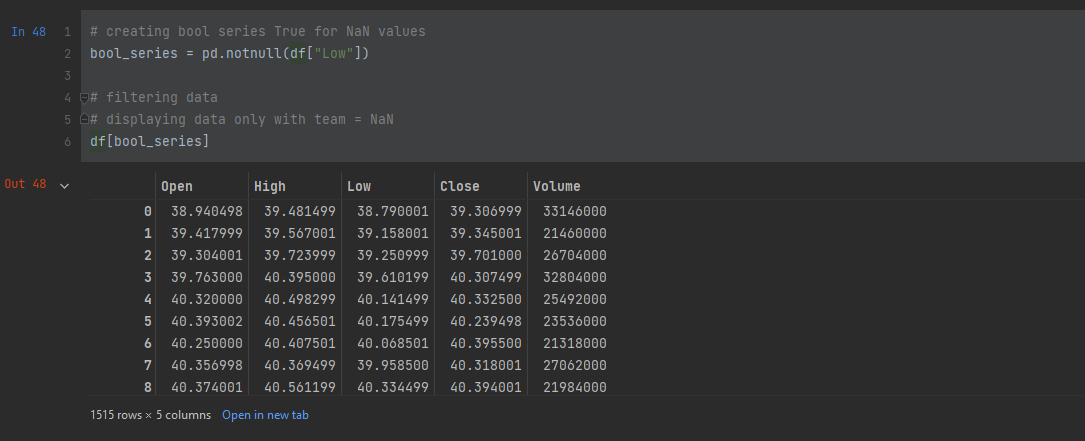
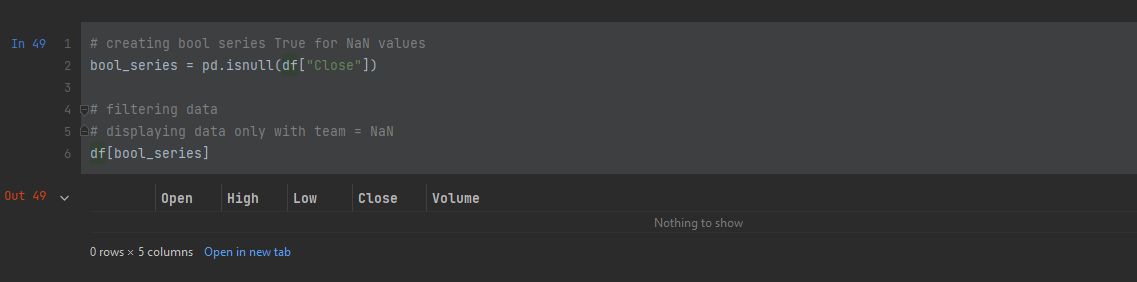
**Now, the best way to explore the dataset is by using** Exploratory Data Analyses (EDAs). EDA (E**xploratory Data Analysis)** can be defined as an approach of examining datasets to epitomize their main features, frequently using statistical graphs and other data visualization methods (wikipedia,2023). Additionally, this report will discuss the precautions taken before performing an EDA as well as explaining the types of EDA and how they were used with the dataset.

Precautions taken to explore the data

1. Check for unique values: Unique values are simply the non-recuring values in the dataset. The code snippets below show the number of unique values for each column as well as all the unique values present in a specific column.



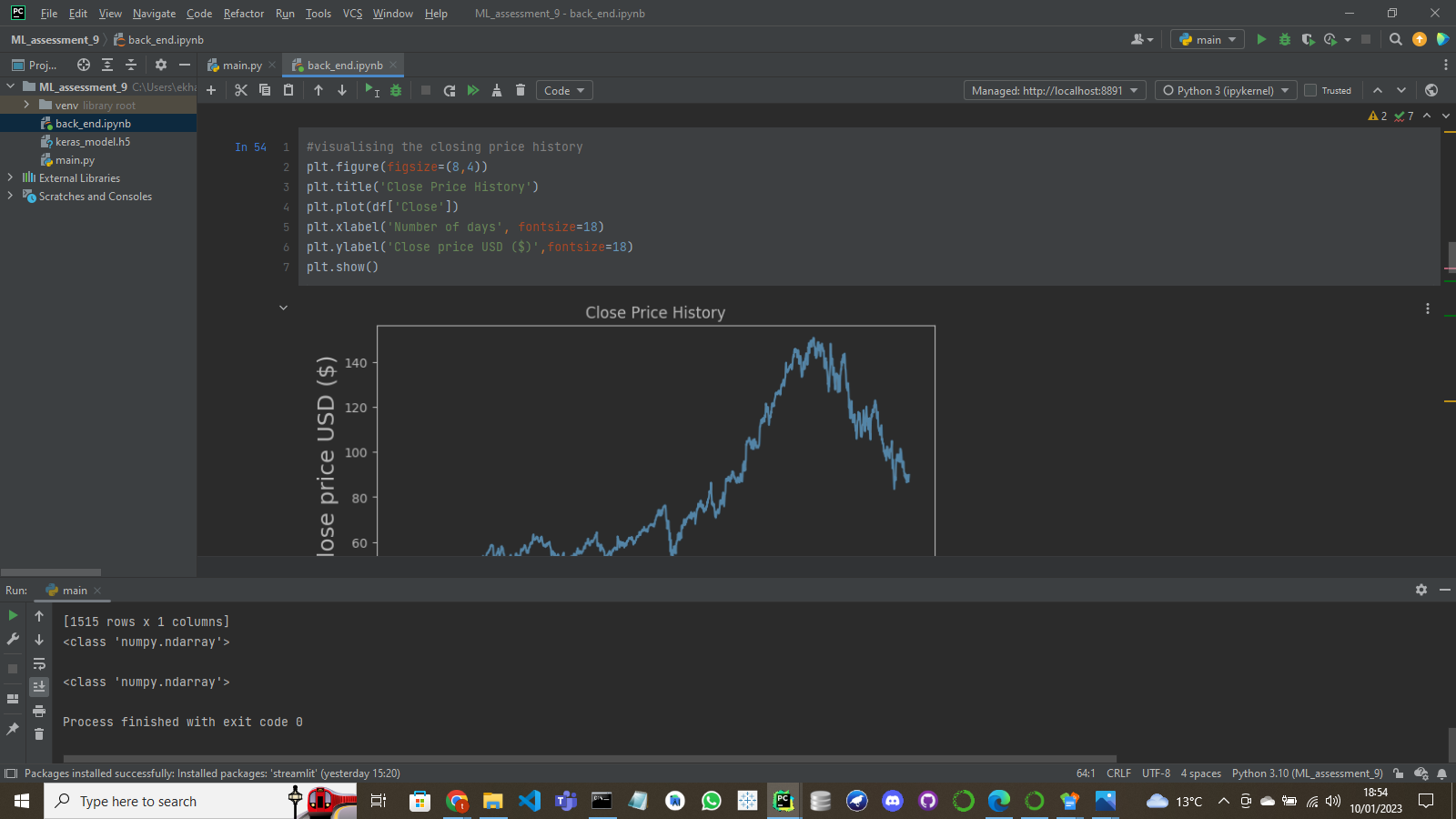
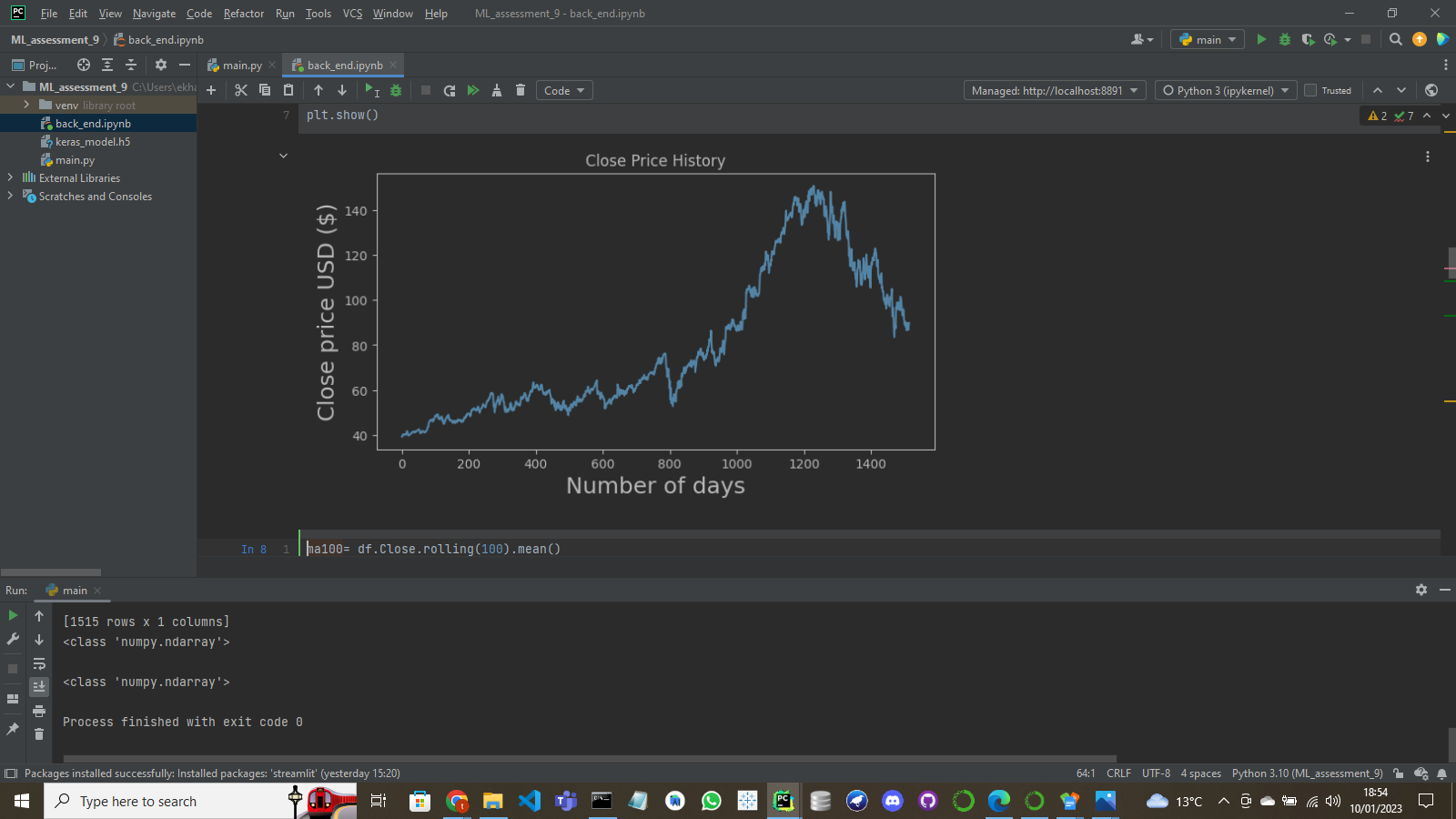
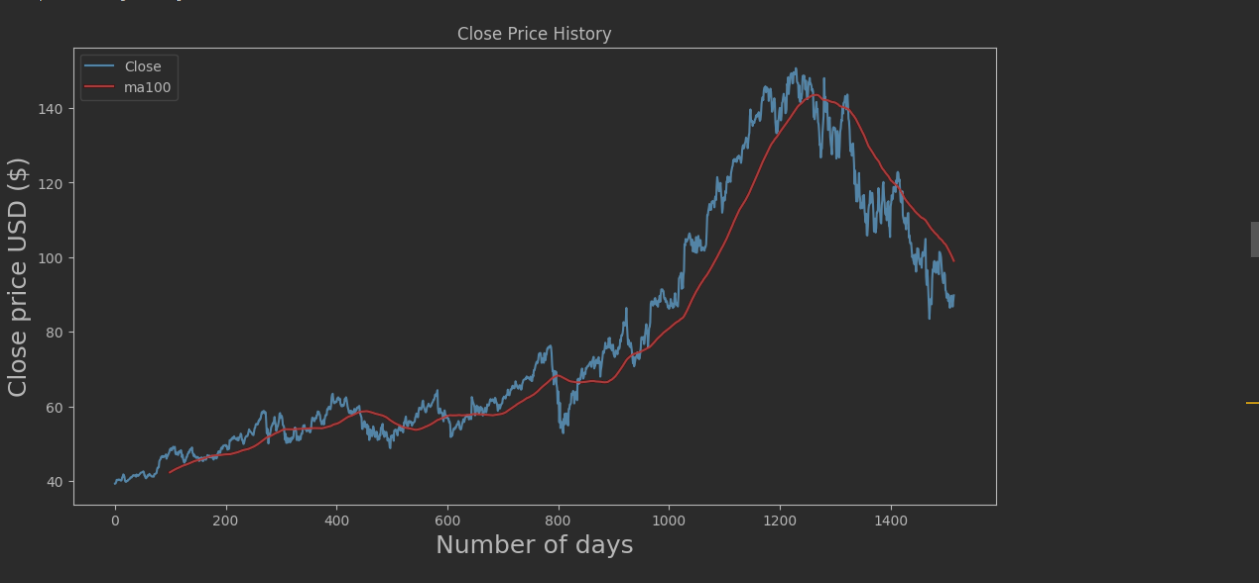
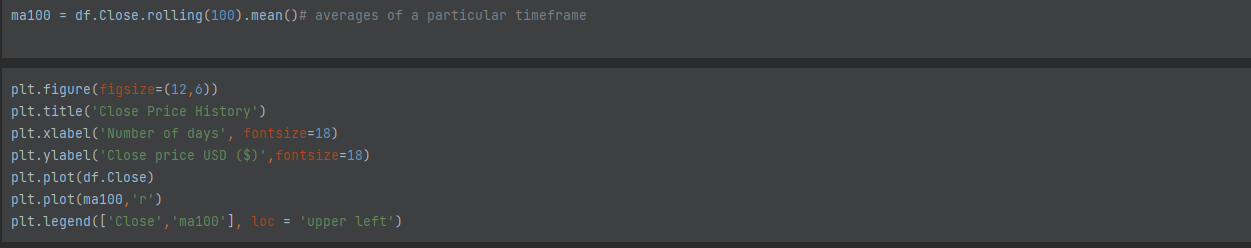
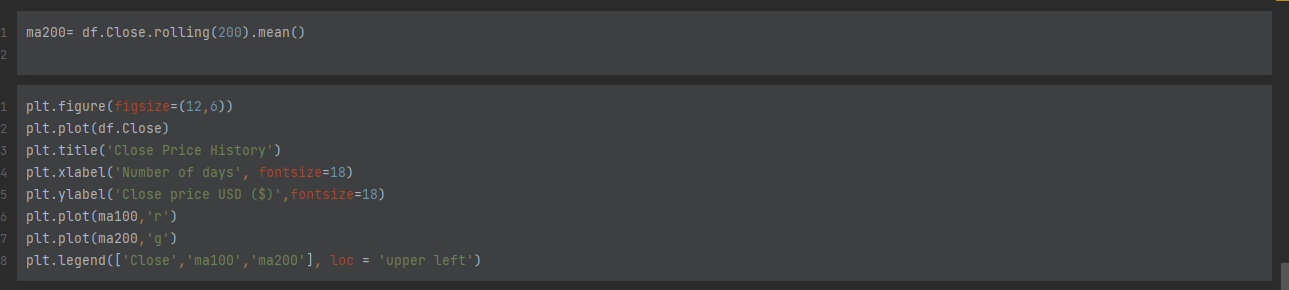


1. Check null values: Before performing EDAs, one must check for null values in the dataset
   1. Isnull(): aids in showing all the columns that have null values. Likewise, the code written displays an empty table meaning that there are no null values in the dataset.
   2. Notnull(): aids in showing all the columns that have no null values which ends up displaying a full dataset.
   3. Reset the index: Reason for this so that the user can be able to drop the date column.
   4. Drop the columns: the final precaution is dropping the two columns seen in the code snippet in order to plot the EDAs.

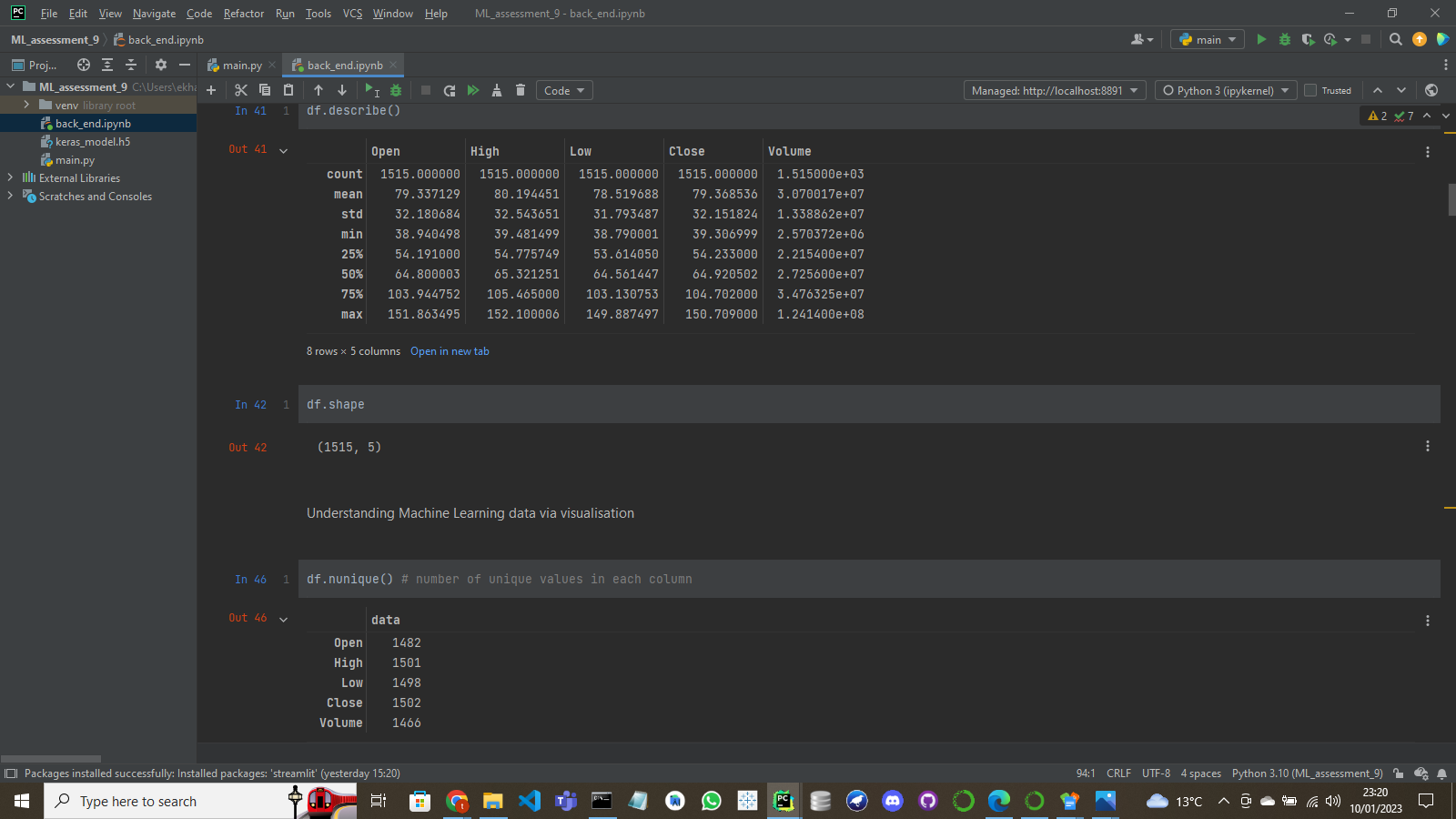
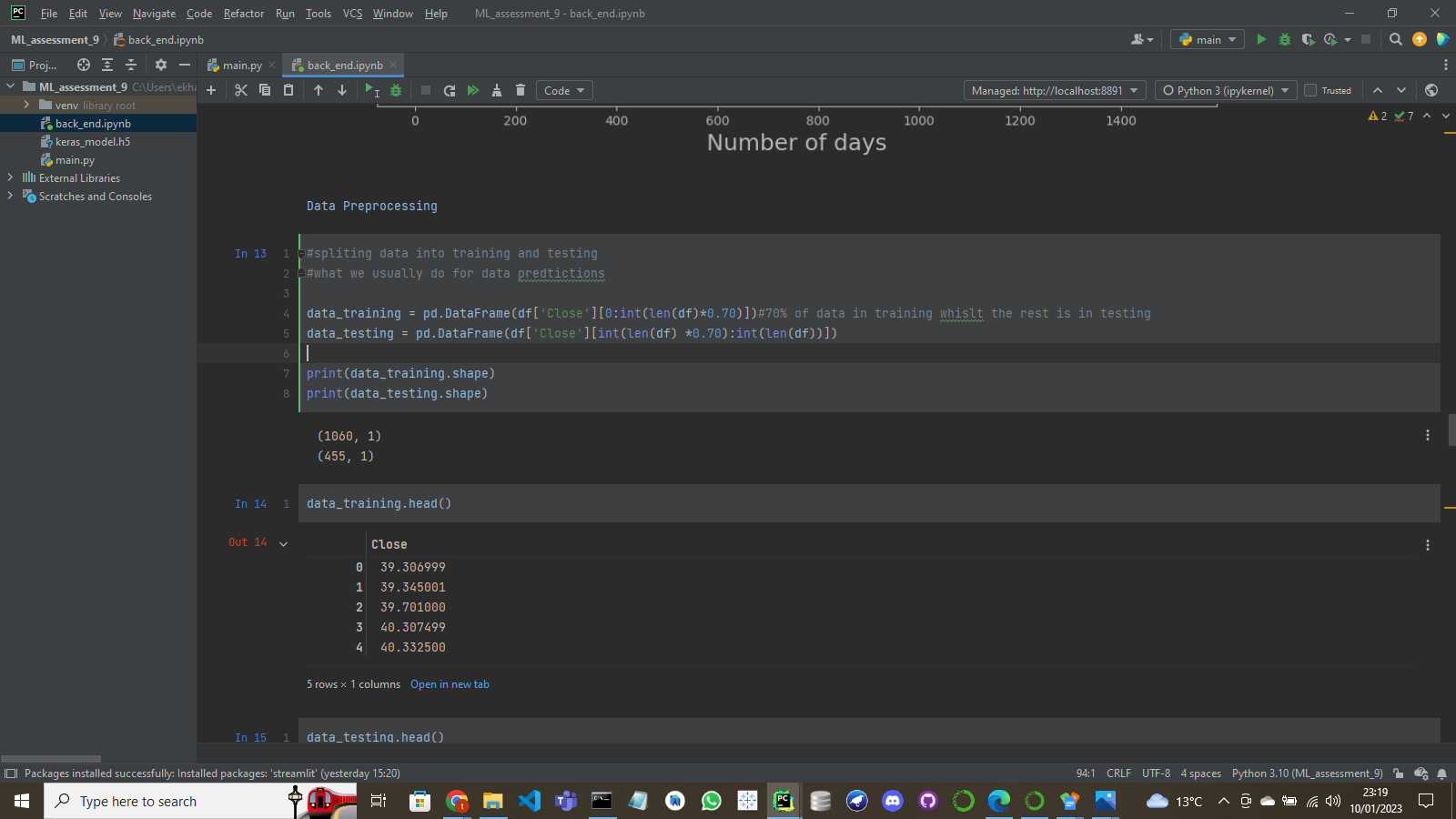
Performing EDAs on the dataset

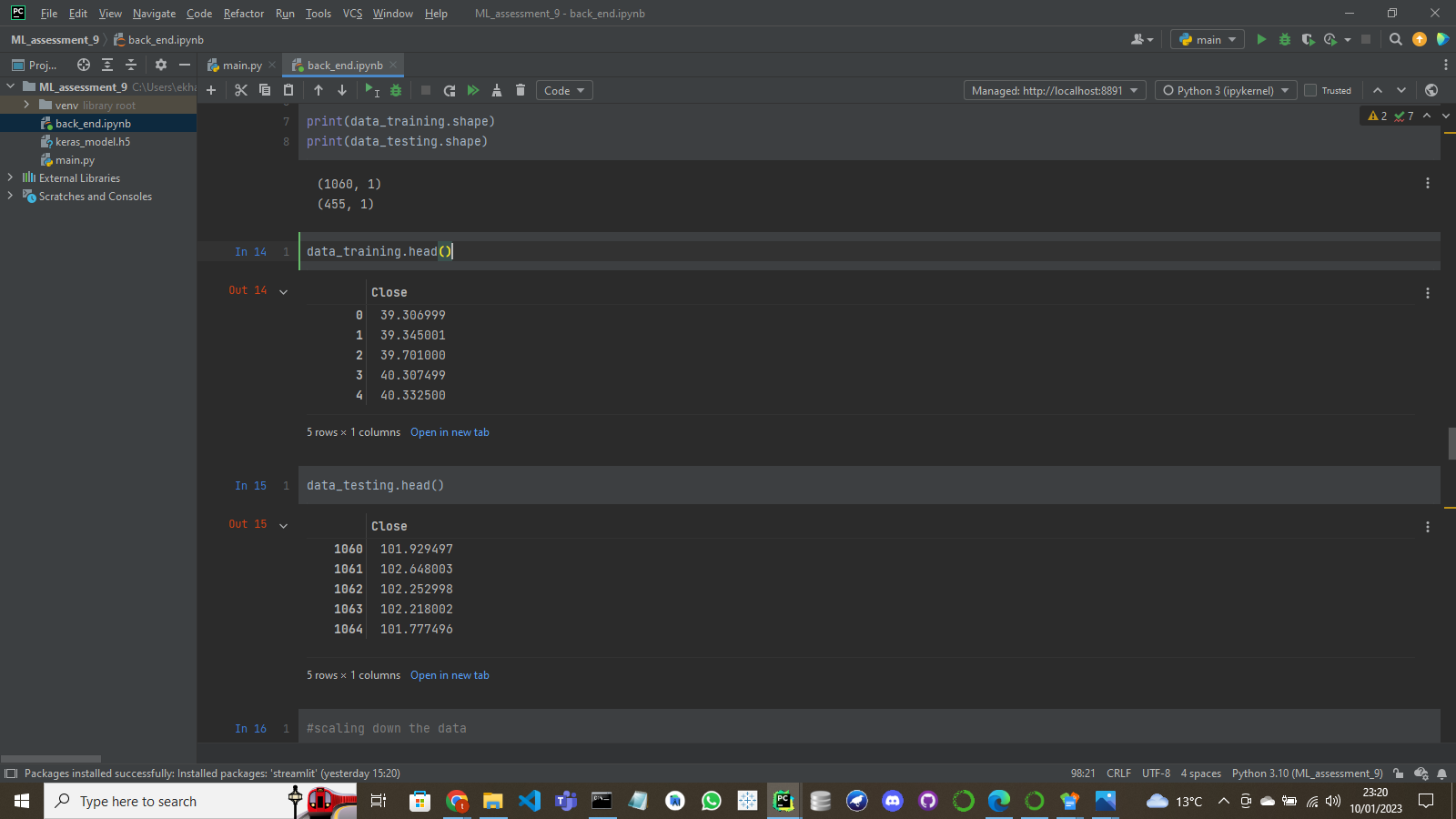
Now the question is what is Exploratory Data Analysis? Exploratory data analysis (EDA) can be defined as a data exploration technique to understand various aspects of data.

Types of EDAs

1. Graphical Univariate analysis: This basically has to do with exploring one variable which is not appropriate for the dataset.
2. Graphical Multivariant analysis: This is essentially the study of the relationship between two or more variables; the dataset being discussed will be subjected to this analysis in order to identify any patterns that may be present. An example of multivariant analysis is the time series chart.
   1. A time series chart presents a series of data points collected over a specified period of time. Moreover, the x-axis plots time and the y-axis plots data points. Likewise, the code snippet below basically plots a graph of the close share prices against the number of days. Lastly, research has proven that in predicting the stock market, the time series chart is a very common visualization method used to recognize patterns, and forecast future trends based on these said patterns (Rakesh Agrawal et al,1995).
   2. Time series chart with simple mean averages: The simple mean average is basically a summation of all close prices over a time period and dividing it by said time period. Furthermore, SMAs react to current markets and cannot predict future share prices. Additionally, when the mean average goes above or below the share price line that is an indication to buy shares or sell said shares respectively.
   3. SMA for 100 days:
   4. SMA for 200 days:

Data preprocessing

Before the data is preprocessed it must first be split into two groups which aid in forming the predictions. These groups known as the training set and test set; with training set gotten from the first 70% of values from the dataset with remaining 30% being used to test the data. Furthermore, the code snippet below shows the shapes of both sets when added together add to the total number of the original dataset. 



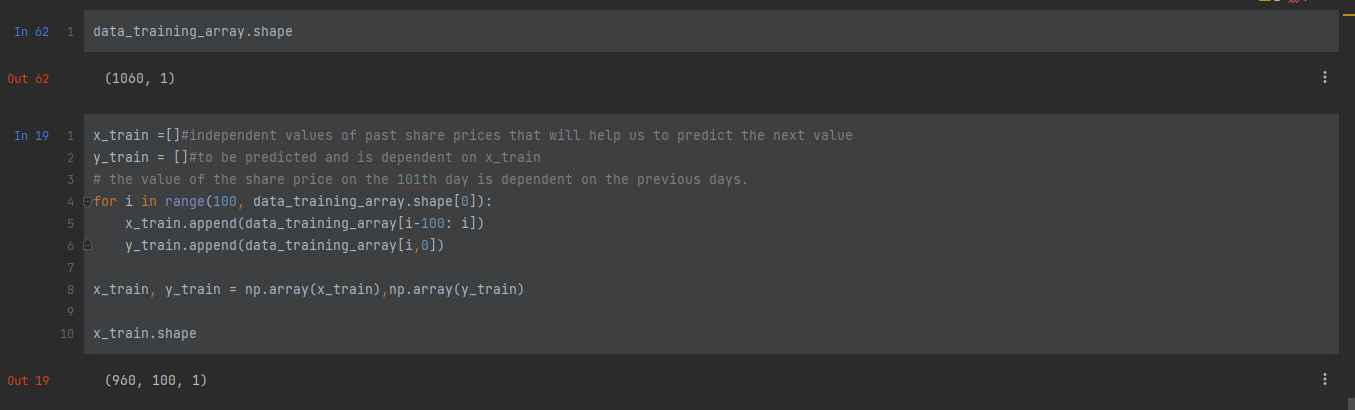
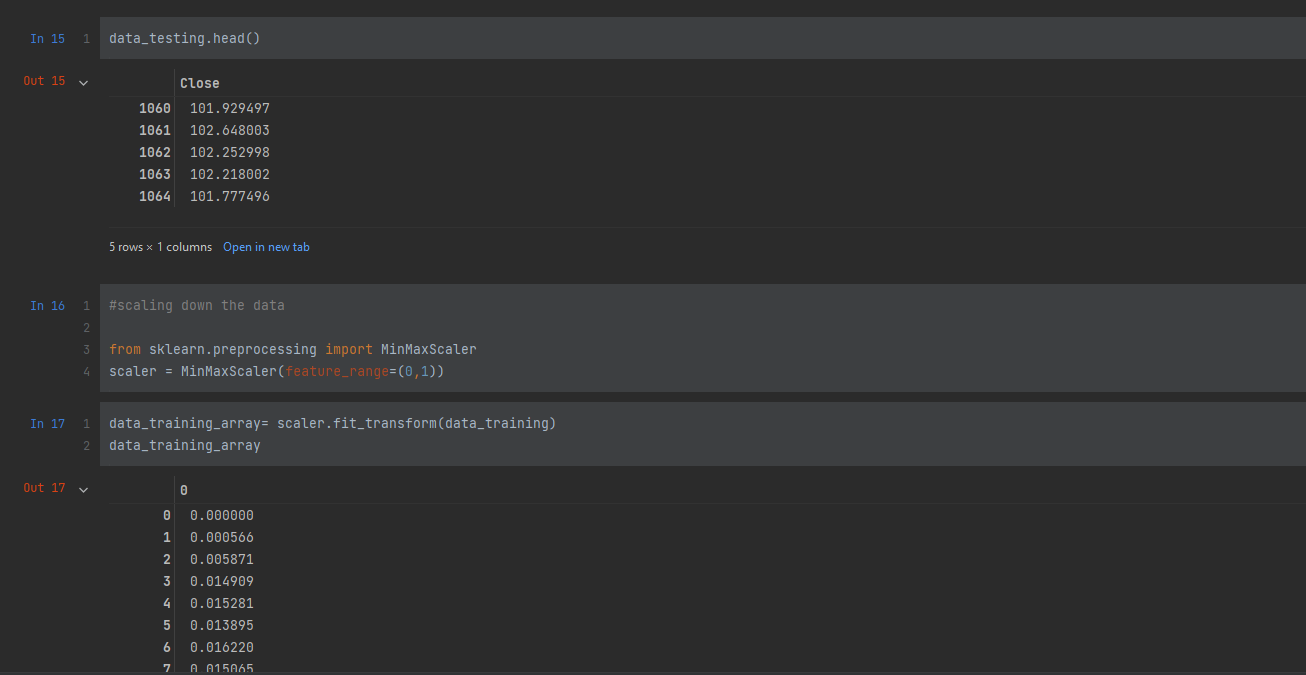
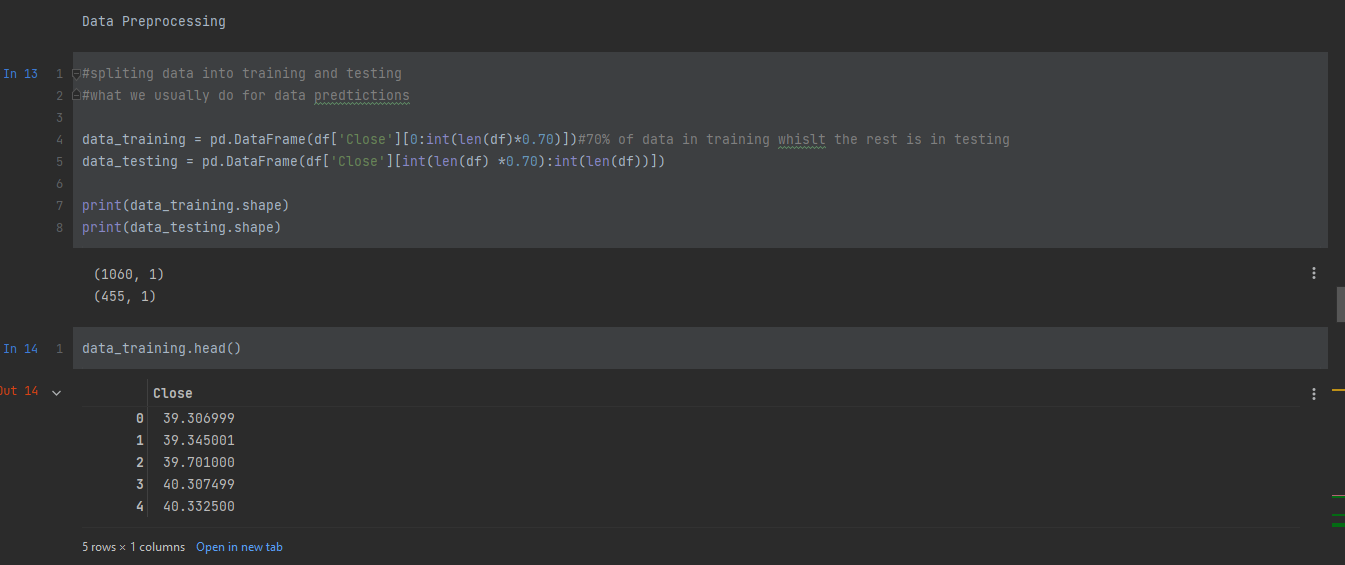
Data preprocessing is simply a method used to remove unnecessary data from the dataset and to make it suitable to be used for a model. The next paragraph shall simply explain the data preprocessing procedures that I made for the model.

The technique I used for my dataset was feature scaling. This is because the data being used has different independent variables of different numerical sizes; and if scaling was not used it would have affected the accuracy of the data.

Furthermore, there are four types of feature scaling but the only type that was used was Min-Max Scaling (Scaling). It varies from normalization in that the only goal here is to alter the data's range, as opposed to normalization and standardization, which have the goal of perfecting the distribution's shape curve into a Gaussian curve. The scale of the data is set to a defined range, typically 0 to 1. ([Kumar,2020](https://medium.com/@nishant61988?source=post_page-----c920ed3637e7--------------------------------))

Now the preprocessing method has been explained. The next step is scaling the trained data after which the training data is split into two lists. The ‘x\_train’ list which stores all the past 100-day share price values enabling the model to predict the next possible value in the dataset; and also, y-train list which basically stores the predicted values.

Consequently, because the model being used is an LSTM model the values in our lists must converted to NumPy arrays in order to be compatible with said model. Code snippets below:



Model development/adoption

A file that has been taught to detect particular patterns is known as a machine learning model. Moreover, a model is trained using a set of data and an algorithm that allows it to analyze and learn from the data (a model doesn’t have to be perfect! Just efficient) (Microsoft,2023).

Now, the model that was used for predicting future shares in the stock dataset is known as LSTM (long short-term memory network). Long short-term memory networks, or LSTMs, are employed in deep learning; where by recurrent neural networks (RNNs) are able to learn long-term dependencies, particularly in tasks involving sequence prediction(intellipaat,2023). Therefore, proving to be the suitable model for the task.

However, research has shown that other models such as the linear regression model can be used to solve stock prediction. Moreover, in this paper it talked about how regression can be used to predict continuous values (e.g., share prices) when given independent values (the training data). The project made in the paper is based on a linear regression algorithm for forecasting correct values by minimizing the error. The algorithm is known as gradient descent; it also uses a linear function to predict the values unlike LSTM(Nordin,2012).

Lastly, another possible model is the support vector regression-based model. Support Vector Machines are supervised learning models with related learning algorithms used in machine learning that examine data used for solving regression and classification problems. The straight line needed to fit the data is referred to as the hyperplane (line of best fit) in Support Vector Regression (Raj,2020).

Moreover, a paper was done using the SVR model to solve a stock prediction problem; with the gamma and Radial Basis Function (RBF) being the parameters of the model. Furthermore, within the paper, it talks about how the value of gamma largely influences the model; with a large gamma causing overfitting and a small gamma giving rise to the model being finite making it unable to capture the shape of the data. Finally, once the data is split into training and testing the resulting are then plotted against the original price like other models (Meesad,2013).

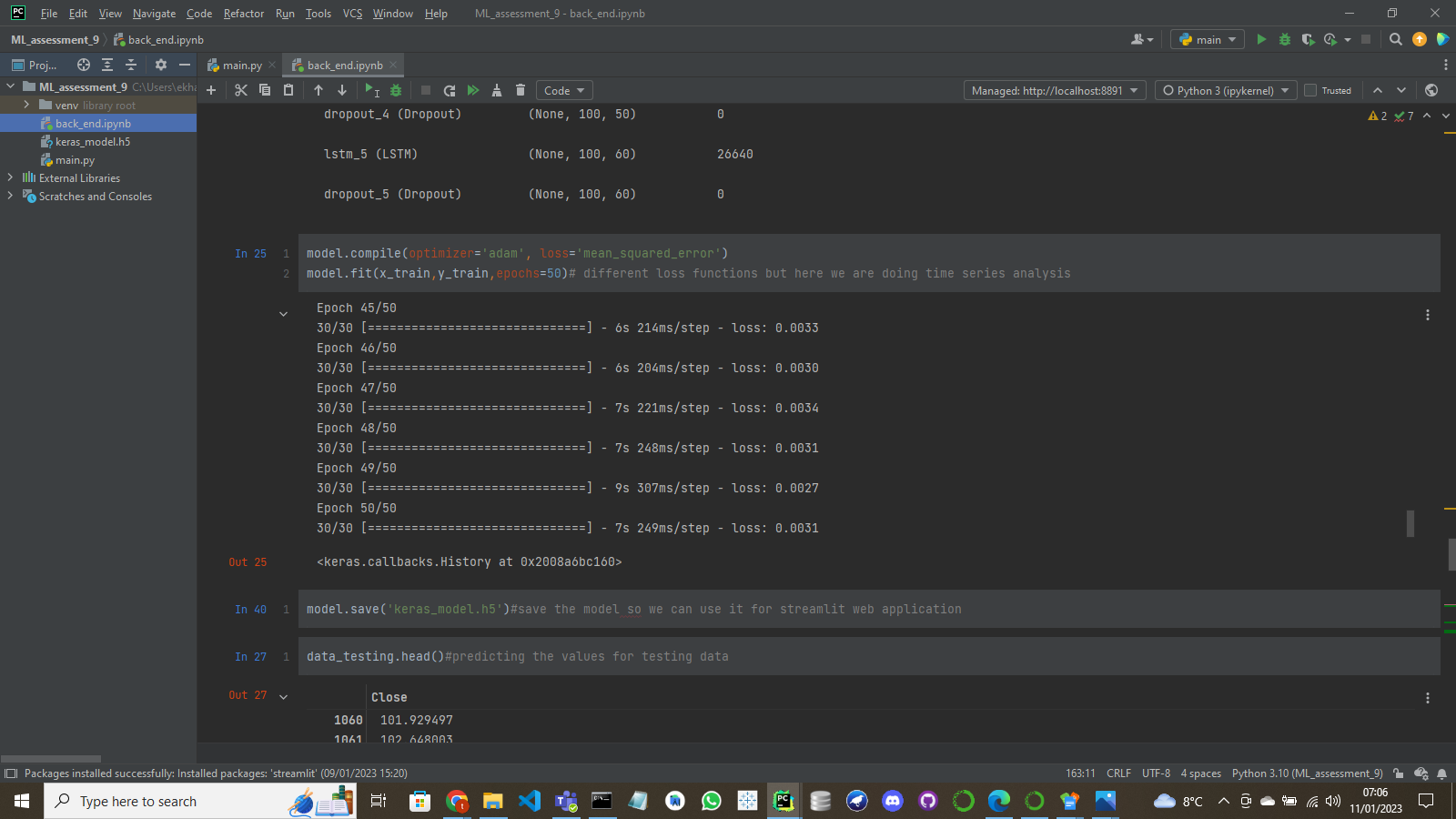
Creating the model

As seen by the code snippet above the layers of the is created using the model.add(LSTM) with also the model.add(Dense) putting it all together. Furthermore, the summary of the model is also printed out.

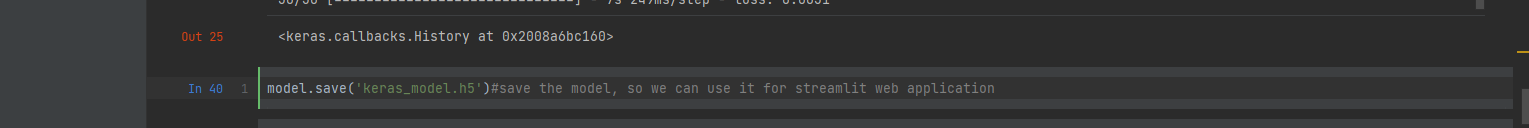
Compiling and fitting the model

Now that the model has been built it must then be compiled and trained. This is done by using the model.compile and model.fit functions respectively.

1. Model.compile: Enables one to compile the model with the optimizer improving the loss function; and loss function simply measuring the performance of the model.
2. Model.fit: to train the model with epochs being the number of iterations in a neural network.

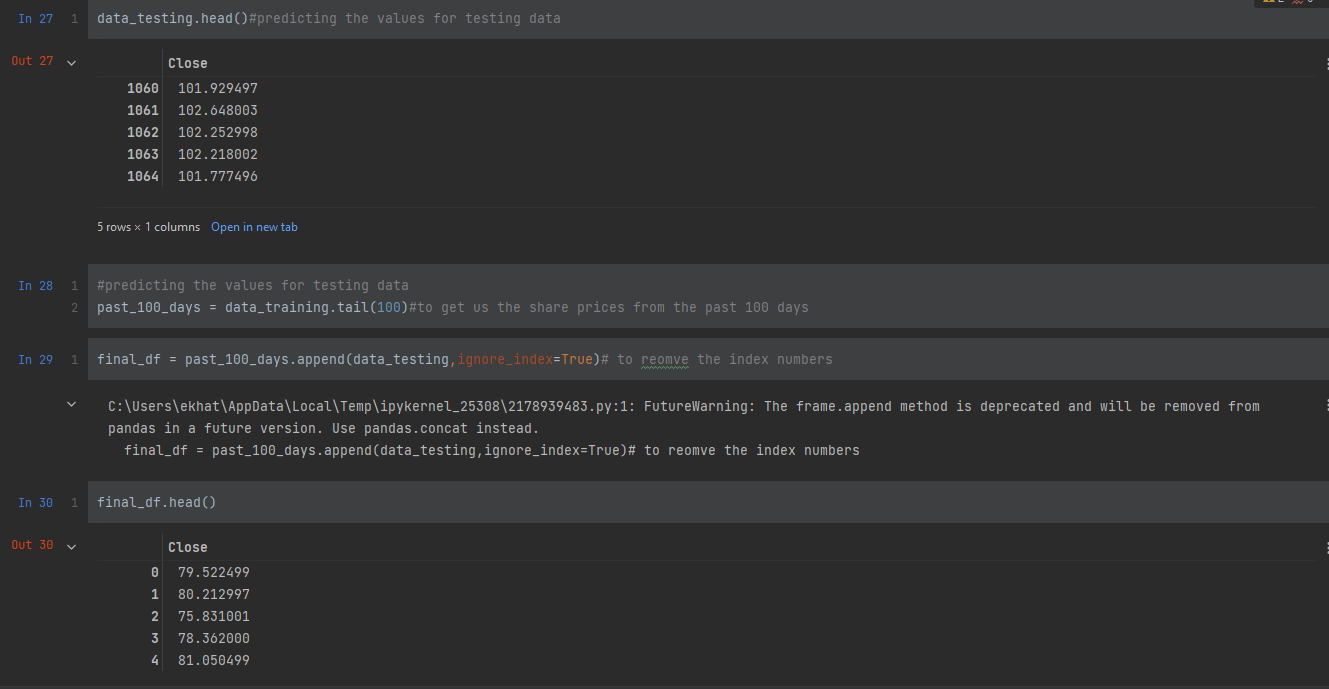


Lastly the model is saved using keras so that It can be used on streamlit.

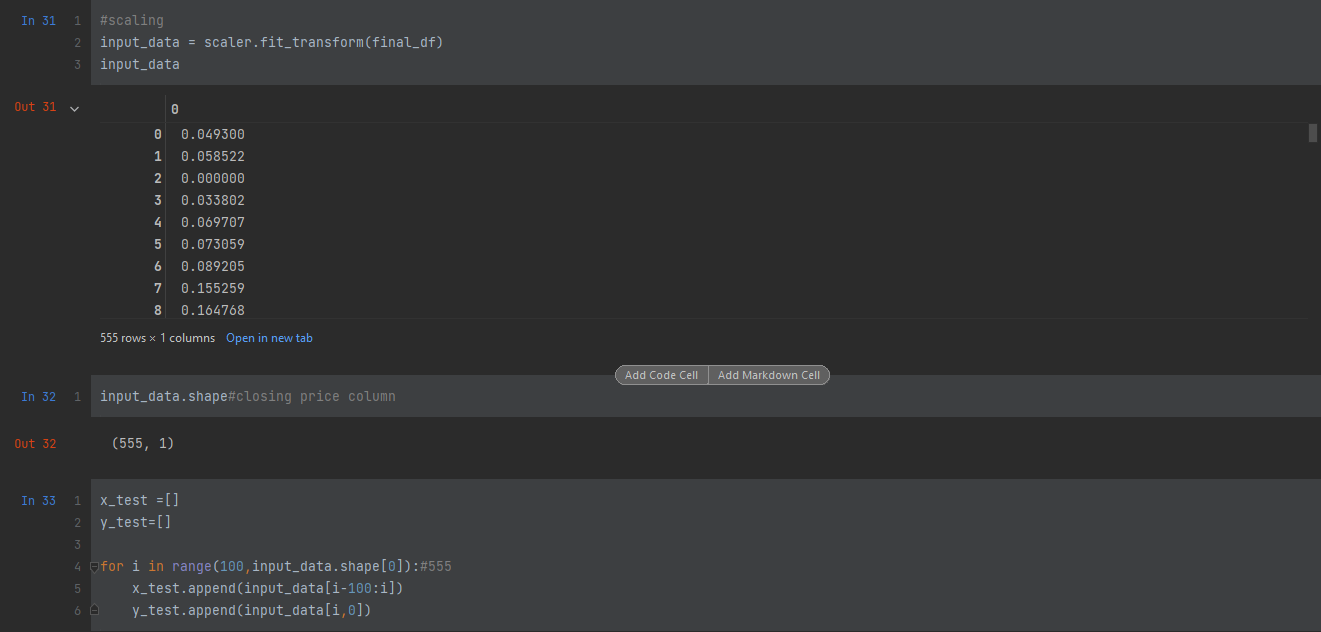


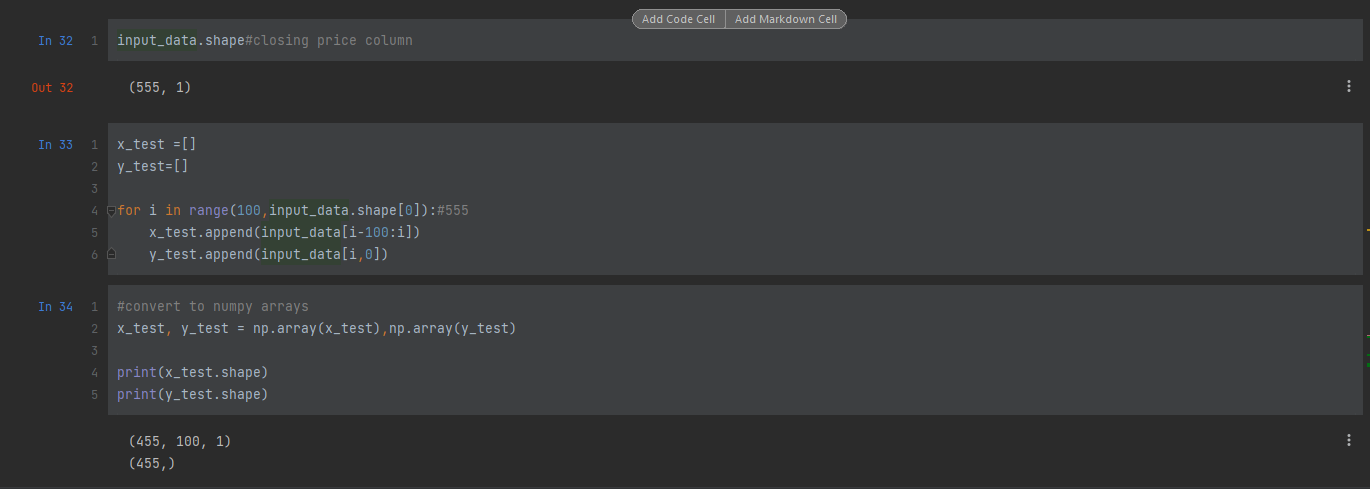
The test data

The testing data helps in checking the accuracy of the prediction. However, in order to predict the it needs the previous 100 days in order to know the next trend.

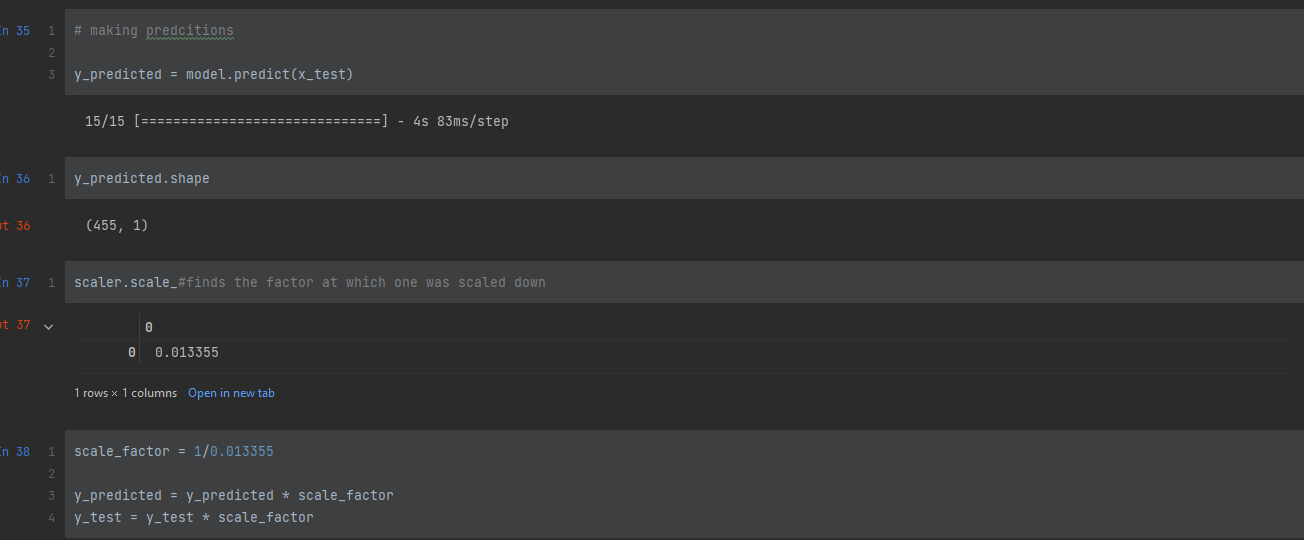


So according to the snippet below, the past 100 days are appended to data testing. The values in data testing are then scaled down.Afterwards, data testing is split into two lists(x\_tests and y\_tests).Then a for loop is ran to append the first 100 days to the first list and the predicted values in the y\_test list. The two lists are than converted into NumPy arrays.





Model prediction



According to the code snippet above the model.predict allows the model to predict the values using X\_test as a guide.After the model is predicted the predicted values are unscaled.

A graph is made where the predicted values are plotted against the original values (note: only one ticker was used as a template for other tickers).



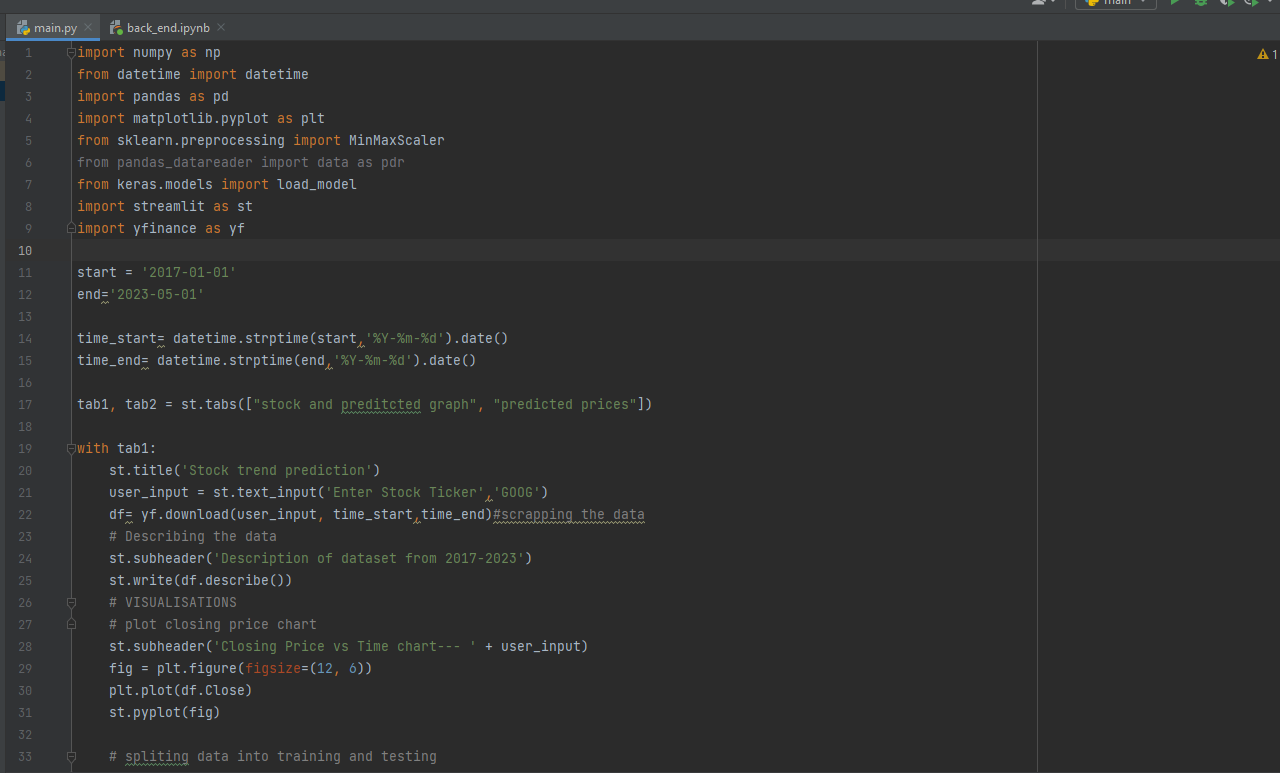
Model evaluation

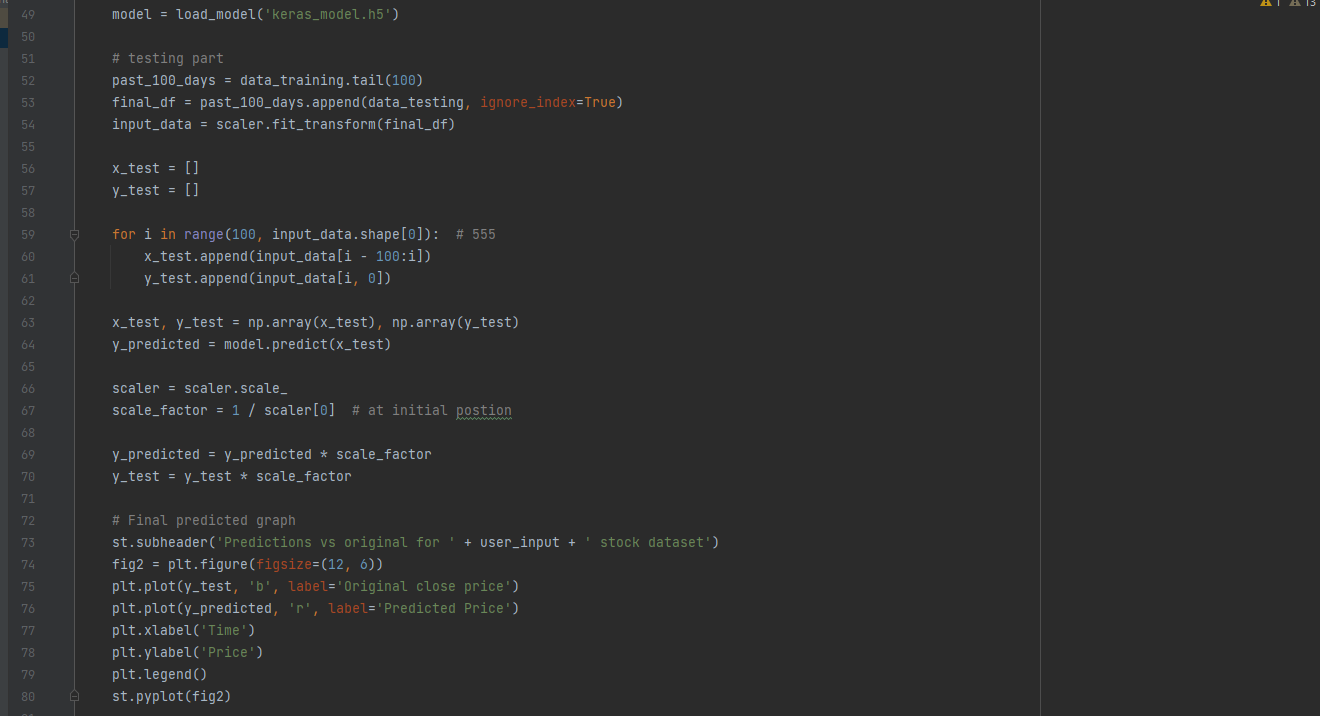
#Evaluate the model via the Root Mean Squared Error(RMSE)  
rmse = np.sqrt(np.mean(y\_predicted - y\_test)\*\* 2)  
print("the root mean square error",rmse)

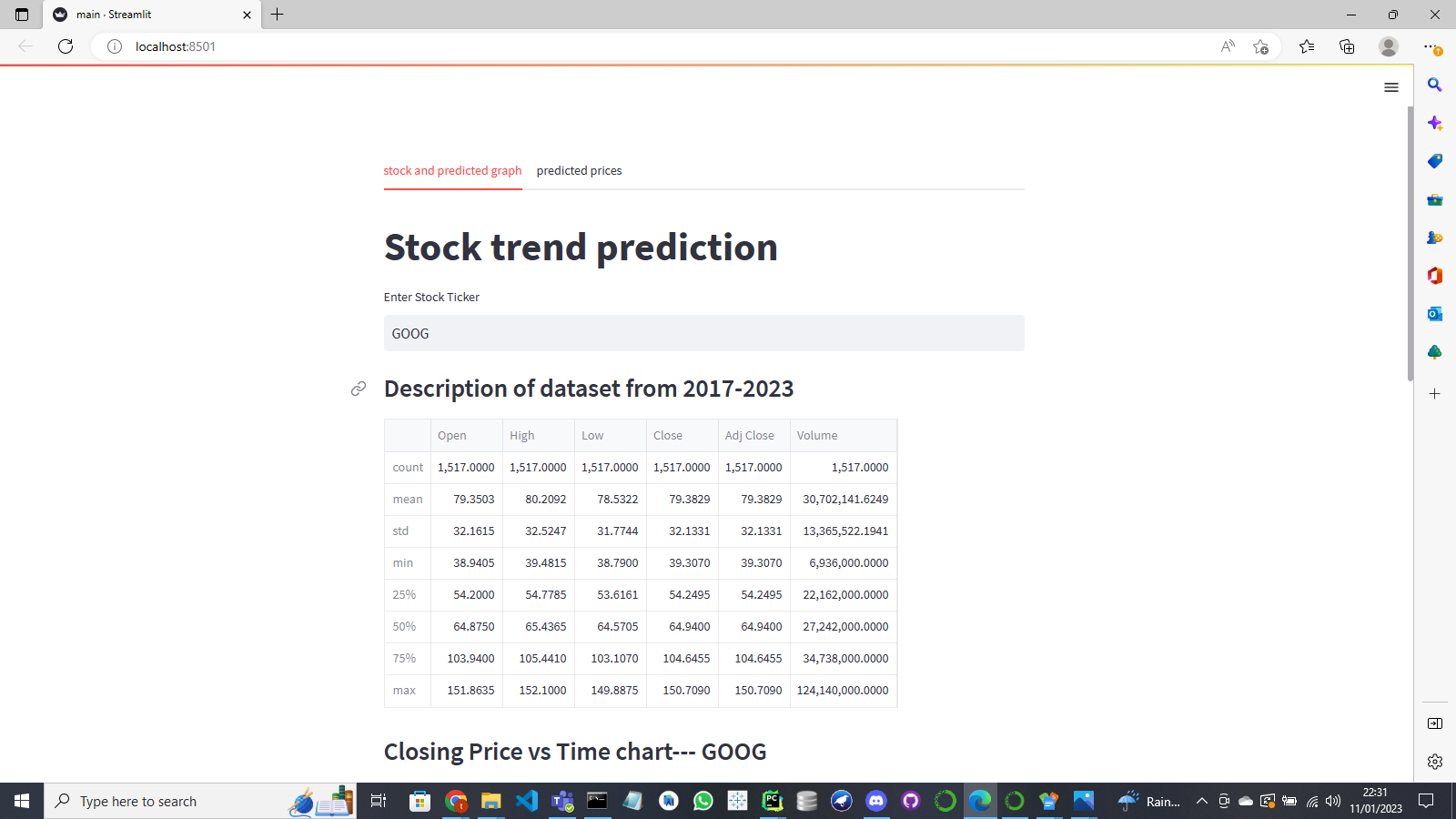
The model is evaluated using the root mean square error(rmse). This tells us the difference in accuracy between accurate values and predicted values. Moreover, the lower rmse score is the more accurate it is.

The front end

The front end basically uses stream lit and when looking at the code snippets our front end has two tabs with the first tab showing the close price graph and the predicted vs original graph for the ticker the user has searched for (a graph where the original and predicted shares are plotted side by side). Whilst the second tab shows the dataset and also a table containing the original and predicted prices.







References

# DataRobot(2023*) Data Collection. Available at:* [*https://www.datarobot.com/wiki/data-collection/#:~:text=Collecting%20data%20allows%20you%20to,trends%20and%20predict%20future%20changes*](https://www.datarobot.com/wiki/data-collection/#:~:text=Collecting%20data%20allows%20you%20to,trends%20and%20predict%20future%20changes)*. (*Accessed : 11 January 2023)

# Wikipedia (2023*)* Exploratory data analysis *Available at:* [*https://en.wikipedia.org/wiki/Exploratory\_data\_analysis*](https://en.wikipedia.org/wiki/Exploratory_data_analysis) *(*Accessed: 11 January 2023)

# NetApp(2023) *What Is Machine Learning?* Available at: <https://www.netapp.com/artificial-intelligence/what-is-machine-learning/> *(*Accessed: 11 January 2023)

Agrawal et al, `Fast Similarity Search in the presence of Noise, Scaling and Translation in Time-Series Databases,' Proceedings of the 21st VLDB Conference, Zurich, Switzerland 1995.

# Microsoft(2023)***What is a machine learning model?* Available at:** [**https://learn.microsoft.com/en-us/windows/ai/windows-ml/what-is-a-machine-learning-model**](https://learn.microsoft.com/en-us/windows/ai/windows-ml/what-is-a-machine-learning-model) **(Accessed: 11 January 2023)**

# IBM (2023) *Time Series Charts****?* Available at:** [**https://www.ibm.com/docs/en/tnpm/1.4.5?topic=charts-time-series**](https://www.ibm.com/docs/en/tnpm/1.4.5?topic=charts-time-series) **(Accessed: 11 January 2023)**

# Raj (2023)***Unlocking the True Power of Support Vector Regression* Available at:** [**https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0#:~:text=Support%20Vector%20Regression%20is%20a,the%20maximum%20number%20of%20points**](https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0#:~:text=Support%20Vector%20Regression%20is%20a,the%20maximum%20number%20of%20points)**. (Accessed: 11 January 2023)**

TIBCO (2023) *What is Data Exploration?* **Available at:** [**https://www.tibco.com/reference-center/what-is-data-exploration**](https://www.tibco.com/reference-center/what-is-data-exploration) **(Accessed: 11 January 2023)**

Intellipaat (2023)***What is LSTM? Introduction to Long Short-Term Memory.*Available at** <https://intellipaat.com/blog/what-is-lstm/#:~:text='%20LSTM%20stands%20for%20long%20short,especially%20in%20sequence%20prediction%20problems>.  **(Accessed: 11 January 2023)**

# Kumar (2023)**Feature Scaling:- Normalization, Standardization and Scaling !**

**Available at:** [**https://medium.com/analytics-vidhya/feature-scaling-normalization-standardization-and-scaling-c920ed3637e7**](https://medium.com/analytics-vidhya/feature-scaling-normalization-standardization-and-scaling-c920ed3637e7) **(Accessed: 11 January 2023)**

Phayung Meesad and Risul Islam Rasel. 2013. Predicting stock market price using support vector regression. In2013 International Conference on Informatics, Electronics and Vision (ICIEV). IEEE, 1–6.

H. L. Siew and M. J. Nordin, "Regression techniques for the prediction of stock price trend," 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE), Langkawi, 2012, pp. 1-5.