A picture containing timeline

Description automatically generated

Full name: Seif Ahmed Rami Esmat Abdelaziz Hassan

TKH ID: 2200251

Coursework: Data Science Project

Instructors: Dr.Batoul and Dr.Mostafa

Course ID: KH5004

Date of submission: 27/04/2025

Table of Contents

[Abstract](#Abstract)……………………………………………………………………………………………………………3

[Introduction](#Introduction)………………………………………………………………………………………………………3

[Data Scrape](#DataScraping)………………………………………………………………………………………………………5

[Research Questions](#RS)……………………………………………………………………….……………….…7

[Deployment code resilience](#DCS)………………………………………….……………….……………………9

[Conclusion and Future work](#CF)………………………………………….……………….……………………10

Table of Figures

[NineEleven](#Nine)……………………………………………………………………………………………………………8

[Covid](#Covid)……………………………………………….…………………………………………………………………9

[AppleReleases](#Releases)……………………………………………………………………………………………………10

Table of Tables

[Data Description](#DD)……………………………………………………………………………………………………6

Abstract

During this project, we went through Kaggle to find a suitable dataset with real data to use to analyze Apple stocks and create 3 research questions. Then we scrape data from the Apple stocks website to later concatenate it with the Kaggle one to have a complete and more updated dataset. We clean the data by removing duplicates, scaling the scraped data, and fix the data type of date to be datetime to be later used as feature engineering. We perform a t-test to see if the means of two years are significantly different or they are not and display a visualization for it. Later, we use a RandomForestRegression model train and test our dataset that will be later used as prediction tool during our deployment phase. Lastly in the deployment phase we added our three research questions and their visualizations as well as including our AI Model predictor that predicts the open price of future dates in our Streamlit where we deploy it with GitHub.

Introduction

Data science is one of the most essential and influential fields in today’s world as the importance of data has grown exponentially in the last few years. With the power of data science, we can gather data that will be cleaned and prepared for analysis which can provide us with insight such as pattern recognition, analyzing trends , and reaching conclusions. The data science life cycle consists of eight stages that contribute to a result that helps whether a stakeholder, marketing team ,or research, find their desired need from it. The start of a data science project is to find a research question that identifies a problem , explore opportunities , find insight , and much more. This helps create a goal to aim towards. The next step is gathering data for that desired goal whether you find data using primary sources such as interviews and questionnaires or secondary sources such as Kaggle datasets and web scraping. Having that data is the foundation where we build everything as without it there would be nothing to use. Next up is data cleaning and preprocessing where you deal with issues that occur in your dataset such as handling missing values, removing outliers, removing duplicates, and data types conversions. This stage though it may be optional based on your dataset but is crucial to ensure a successful result as neglecting this phase can really hinder your progress and cause unwanted results. The fourth stage is the exploratory stage where you apply statistical functions such as finding the mean, mode, and standard deviation to find patterns and understand the structure of the data. The next stage is an important stage as it helps with predictions and finding patterns. This is done by using models that split the data into two sections, one for training and the other for testing and based on your model selection your model prediction accuracy is altered. Known models such as Linear regression, k-means clustering, and classifications are used in machine learning, all with different perks for different varieties of data . The next stage is evaluating the model by checking its accuracy measurement using tools like R2 score and mean squared error, then apply Hyperparameter tuning if needed. The next stage is to deploy the model, whether you use an app or website like React or Streamlit . Lastly you monitor and maintain your app and model for effective usage. In our goal we aim to analyze three research questions about apple stocks and perform the data science life cycle on it.

Data Scraping

Data scraping is the gathering of data phase, where we go on the web and start finding a trusted source to scrape data from using tools such as Selenium and Beautiful Soup. We scraped data from a trusted source owned by Apple that displayed their stocks. The website called Apple Stocks had me input three components: day, month, and year, and then it generates the open, high, low, and close prices along with the volume of stocks for that date. I used Selenium WebDriver manager from Chrome to extract the data from the website and input it into a CSV file called NewData.csv. I did by selecting the years, and I want to scrape from and then iterate over the days and months of those years and reach the final day of the year. Using the class name of the HTML tags, I managed to input the data automatically to iterate over the days and extract the data I wanted from the output tags.

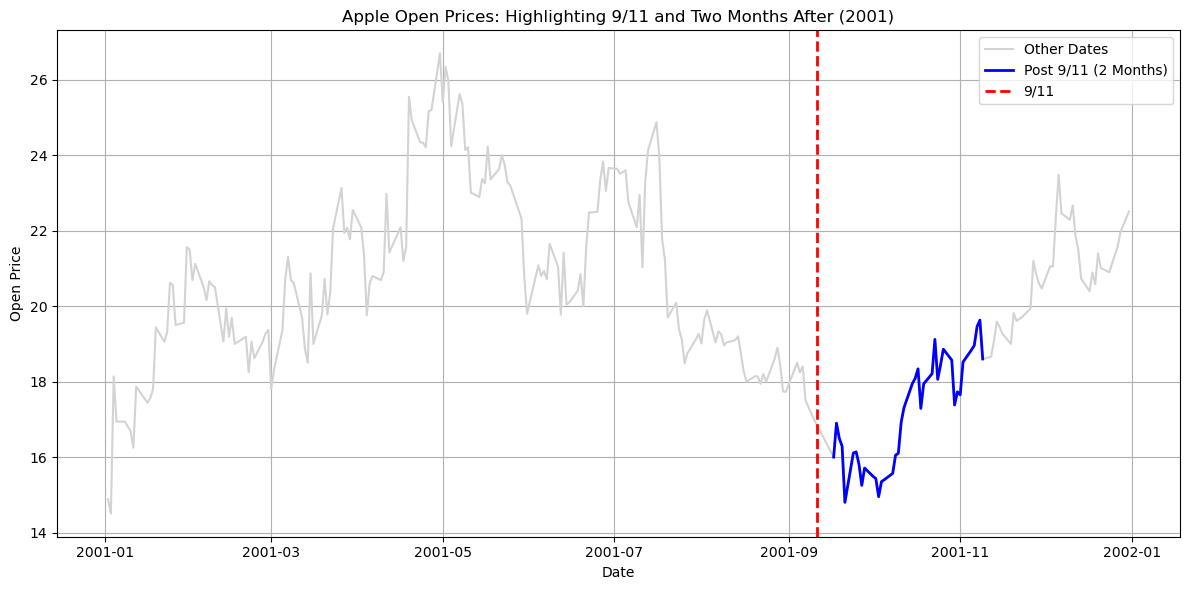
Data Description

|  |  |
| --- | --- |
| Data | Explanation |
| Date | The date of the record |
| Open price | **Opening price of stock** |
| High price | Highest price of stock for the day |
| Low price | **Lowest price of stock for the day** |
| Close price | Closing price of the stock |
| Volume | **Number of shares traded** |

Research questions

Did the event of 9/11 affect apple stocks ?

9/11 was a tragic event that occurred on September 11, 2001,where two hijacked planes flew into the twin tower of the world trade center causing the death thousands of civilians and security dilemma all over the world on airlines. Such an event had negative impact on the financial market, so I traced the 9/11/2001 on my dataset and created a line graph which a vertical red line highlighting the day and then we visualize the next 2 month after the tragic event by highlighting them in blue . Now we can see that during those two months the stock market of Apple did take a hit but managed to recover quickly.



Did the Covid pandemic have any effect on the apple stocks ?

Covid-19 also know corona virus was a virus that started to spread during 2019 and as it was quickly spreading it lead to full pandemic in the year 2020. National health organizations implemented rules like social distancing, which significantly disrupted financial markets in exchange for ensuring public safety. For apple that meant that they needed to sell their stocks quickly which led to a high uprise but followed by massive downfall that led to long-term financial damage. We analyze the damage by getting apple stocks records from the year 2017 till 2022. We loop over each year and plot a line graph that shows us the open price during those years and color each year differently to see all their differences. We can see that during 2020 the stocks prices went up and then plummeted down with long term damage.

A graph showing the growth of a stock market

AI-generated content may be incorrect.

Is there a massive difference between the release of the first iPhone and the last iPhone?

Apple is a worldwide known company for their prestigious technology. They’re mainly known for their iconic iPhone releases that always catches the eye of everyone . We wanted to see if their first iPhone release had more of an impact then their last release. We do that by making two subplots that each aim on the dates of their first iPhone release and their latest one. Then we check the month before each release to see if they are decaying or going down then we check the month to see if the release nourished them with high open prices. And as we see the first iPhone was inclined before the release and profited so much more after the release while the iPhone 16 was declining and then inclined right before the release then declined and inclined again making their growth seem unclear but based on the open price iPhone 16 gains much more money at a slow rate while the first iPhone gets less money that grows quickly over time

A screenshot of a graph

AI-generated content may be incorrect.

Deployment code resilience

During the deployment phase we take safety measure to ensure that the user does not mess with the inputs and break the app so we implement dropdown menus for the date section as he cannot enter anything except the date and if enters an invalid date then it shows an warning as for the null values if the user enters one then nothing will happen if the dates are null and if the volume is null then it resets to default value.

Conclusion and future work

In conclusion, we applied the whole lifecycle to our apple stocks project by getting, cleaning, analyzing, modeling ,evaluating, and deploying our project successfully. Any future improvements would be a better model that doesn’t overfit the data as much and dive into deep learning algorithm that would benefit the project more