Matthew Shinder

Capstone Three - Project Report

Introduction

The stock market is an aggregation of buyers and sellers of stocks. On normal business days, high volumes of stocks are bought and sold by various individuals / companies. Trying to predict where a stock's value might be heading based on various features is a tricky endeavor and a big financial risk for those who are involved. The goal of this project is to create a stock market prediction model using advanced machine learning techniques.

Datasets

To gather the data, we will use Yahoo's Python API in order to gather the data from a certain date range rather than download a database. The dates that we will choose from will range from January of 2021 to December of 2023, a three year span. For best price evaluation, we will use the closing price of each stock everyday as a better representation for the stock's evaluation and what current events might have taken place that day.

Links:

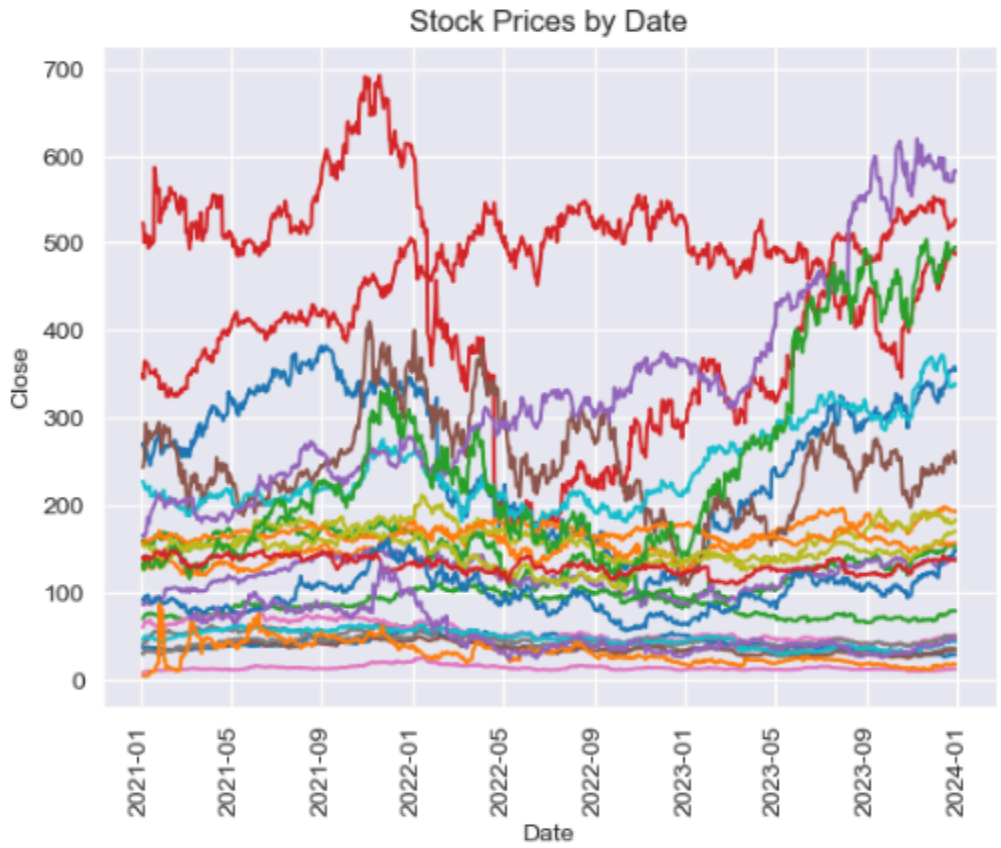
https://pypi.org/project/yfinance/

Cleaning/Wrangling

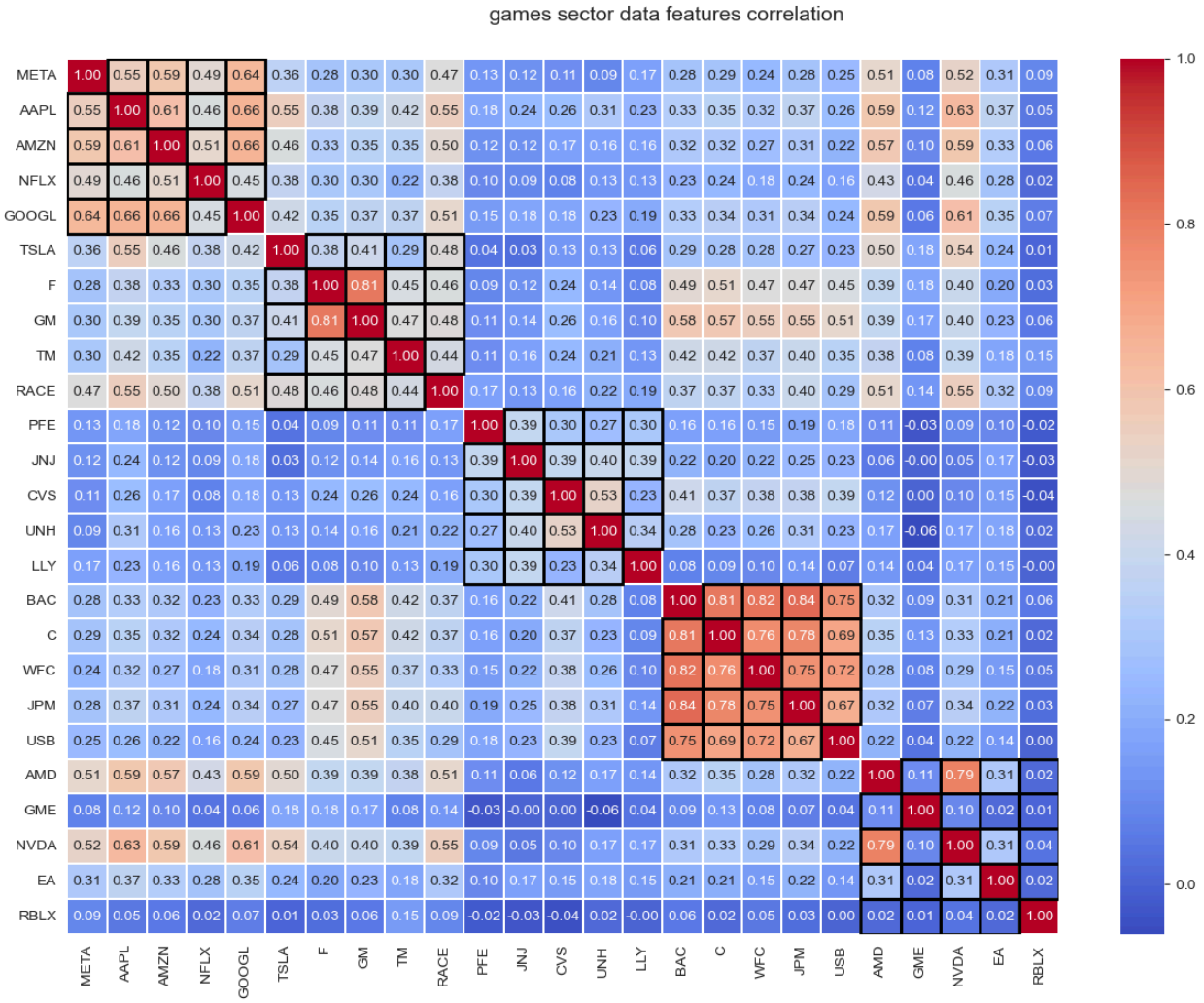
Not many steps were needed for cleaning / wrangling our data. The data taken from an API has been mostly set up in a desirable format and had no null values. A column of interest to make for our EDA later would be the return of the stock daily. The return is created by taking the percentage difference (positive or negative) change of the stock each day.

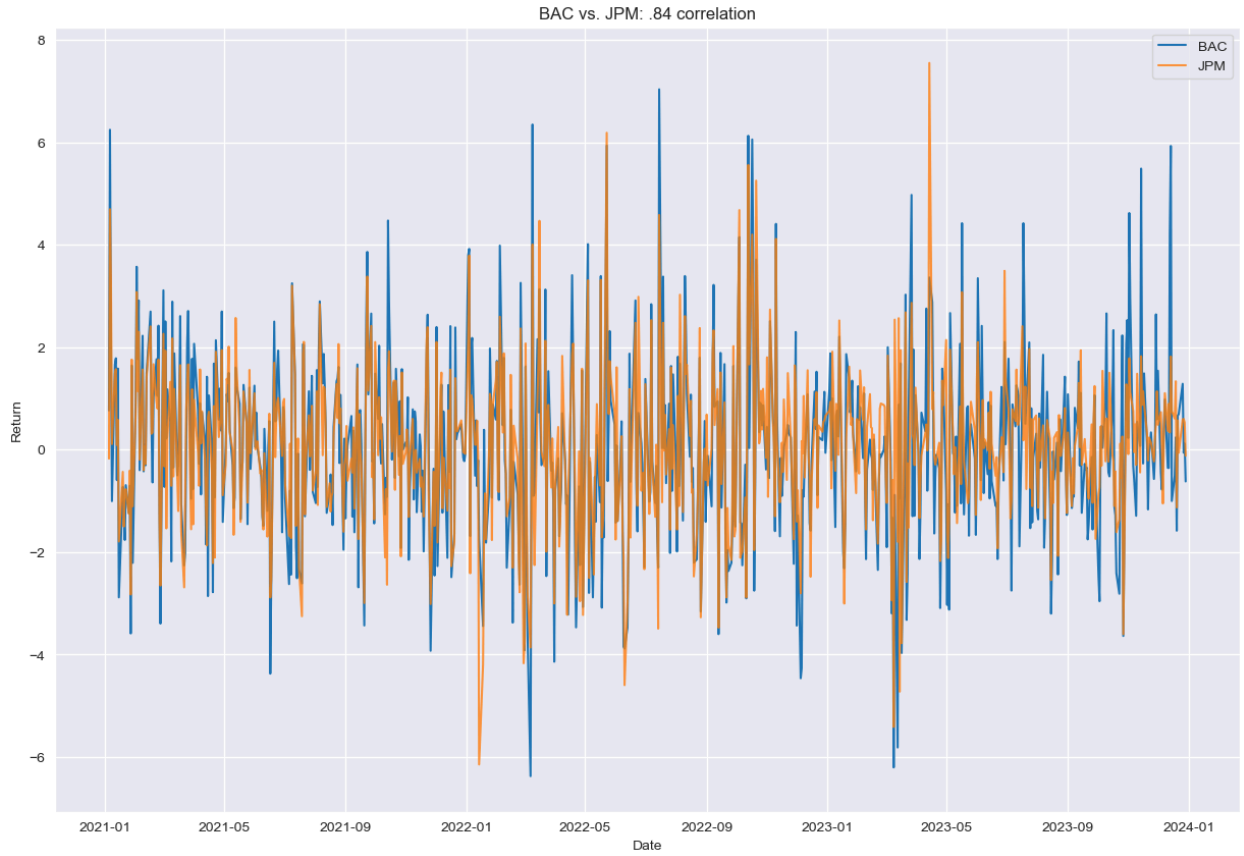
EDA

When looking at our beginning graphs below, comparing the stock prices of all 25 at once would prove not to be beneficial as the prices have high variance between each other. We then looked at our returns divided by the sector of the stocks.

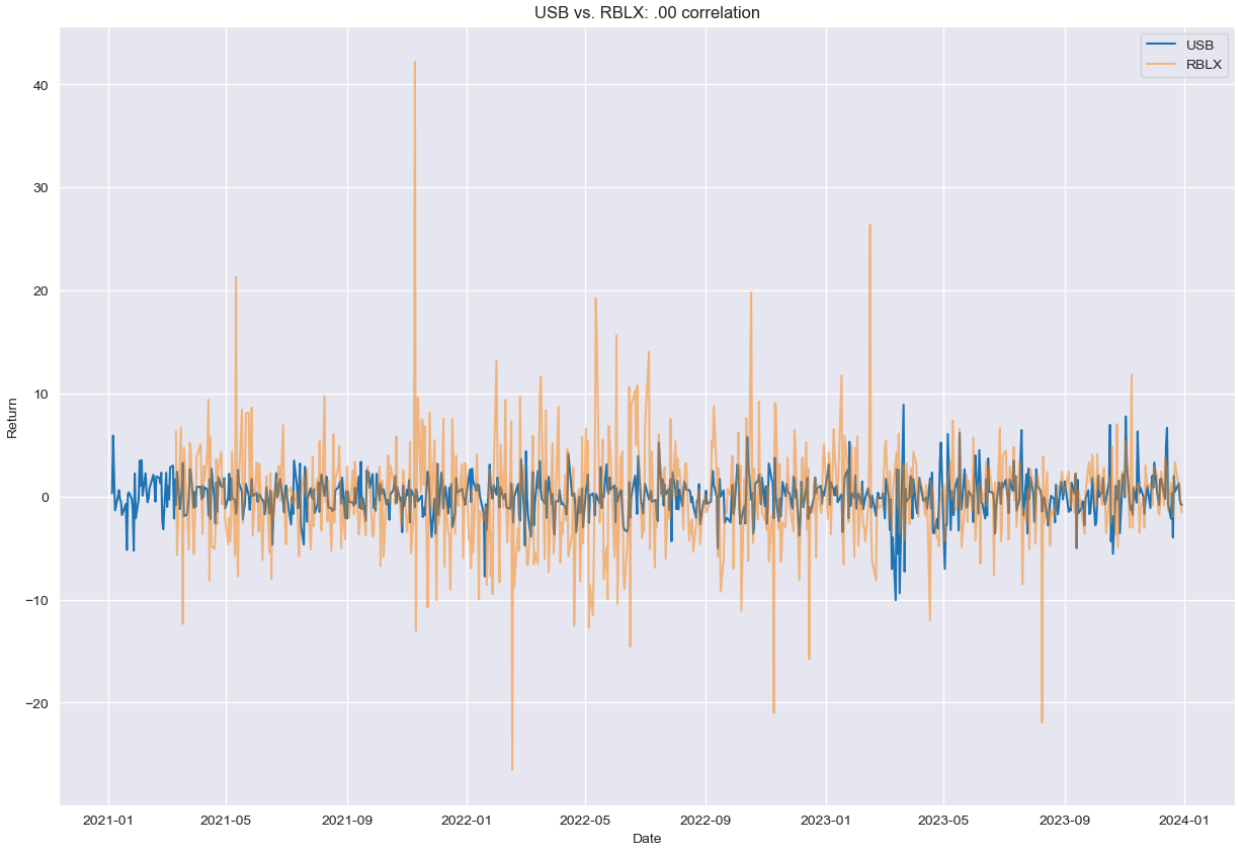


To see any early trends, we decided to inspect a heatmap showing the correlations between the returns of each stock between each other. The highlighted box groups show each sector.





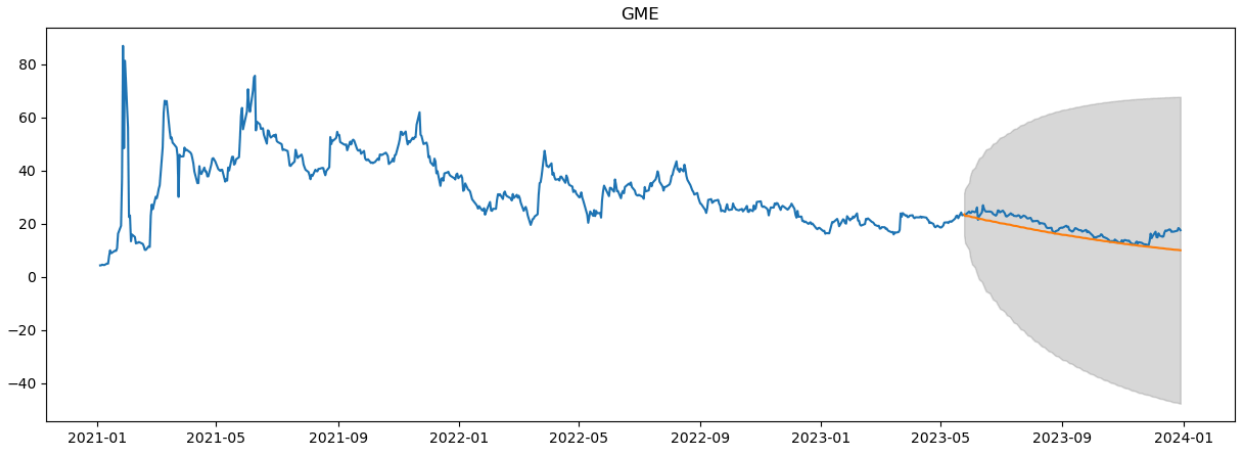
Bank stocks seemed to be highly correlated with each other but when compared to a different sector, the correlation was near 0.

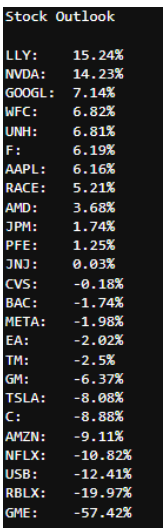


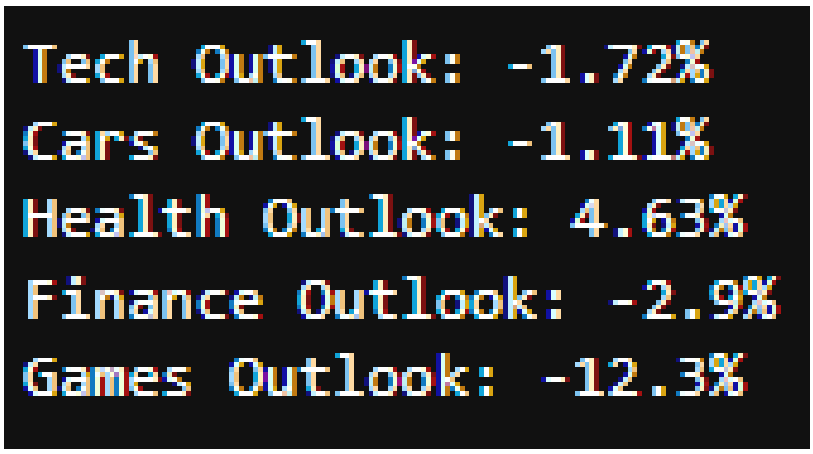
Preprocessing and Modeling

The goals of this step were to set up and model our data for time series forecasting by also being able to loop through our data in order to automate the process with modular code.

To do so, we ran an Auto ARIMA function in order to first find stationarity, and then find the optimal parameters of p and q in order to best forecast our models.







Analysis

Our models have slightly indicated that our health stocks overall have the best outlook in the future. This does not guarantee that all health sector stocks will go up though. LLY or Eli Lilly And Co seems to have the best future while GME, or GameStop, seems to be on the heavy decline.

Many of the stocks we made models for were not easy to predict. If they were easy to predict, everyone would be rich. With more fine tuning on each individual stock, we may be able to better predict our markets in the future.