Sector Regrouping for Invesco QQQ ETF based on K-Mean — Clustering Algorithm—

Authors: Topu Saha, Weikang Kong, Nan Wang

Introduction

Invesco QQQ is an ETF based on the Nasdaq-100 index and is allocated to seven sectors based on the company's nature. With the growth and expansion of the company businesses, the nature of the company and the performance of their stocks might be changed consistently. The goal is to re-cluster stocks from QQQ into sectors by the performance of their stocks using the K-Mean clustering algorithm on the holdings in the QQQ exchange-traded fund. The correlations between the predicted sectors and their stocks vs. the original sectors and their stocks will be compared, analyzed, and visualized in this research.

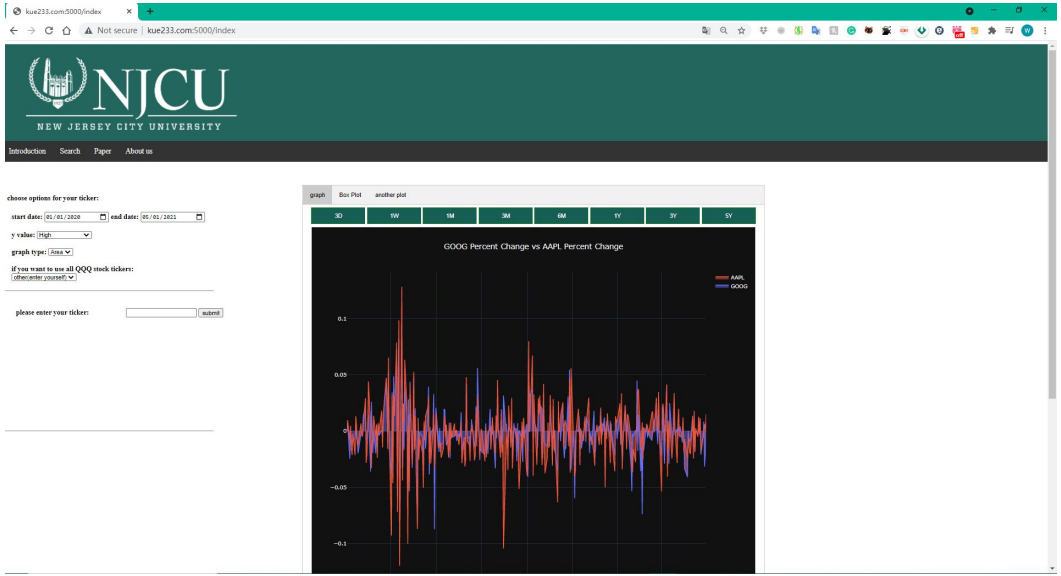
Methodology

- Pre-Process Data
  - Create a Pandas DataFrame with all the needed information
    - Ticker, Monthly Closing Prices, Original Sector
  - Delete null values
  - Generate Monthly Percent Change from the closing prices
  - Transform into a Pivot plot
- Run K-Means algorithm on Data
  - Specify how many clusters
  - Fed the model data from the stocks
  - Run the Model on the stocks to classify them
- Visualize Results into scatterplot, and heatmaps

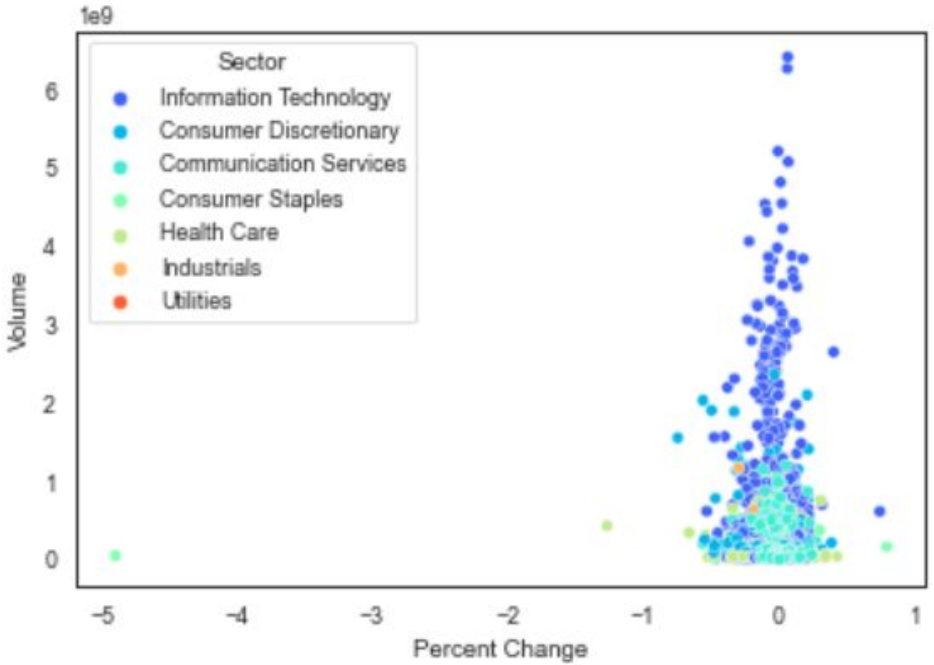
Website

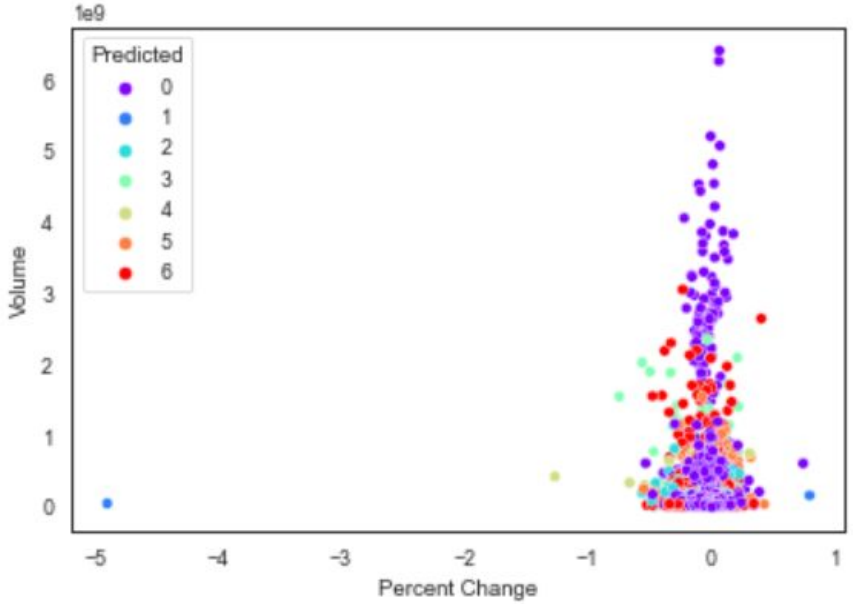
Our Website Includes

- Visualization of Stocks information
- Copy of our research paper
- About us page

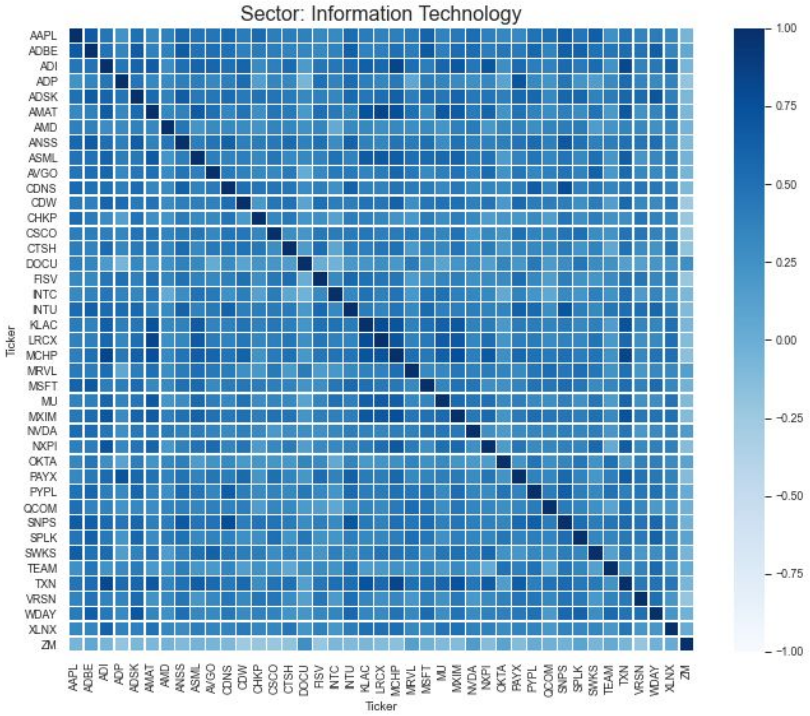


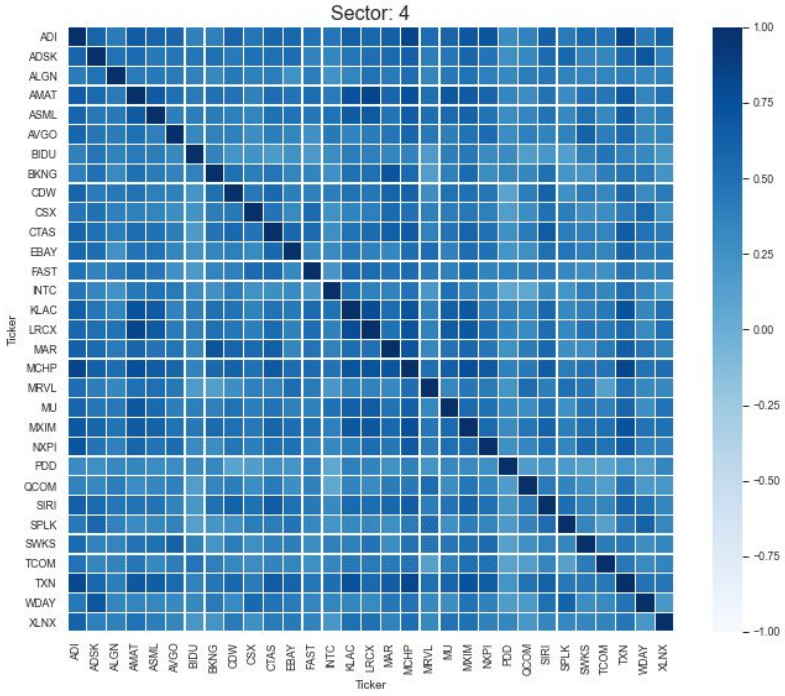
Results: Scatterplots



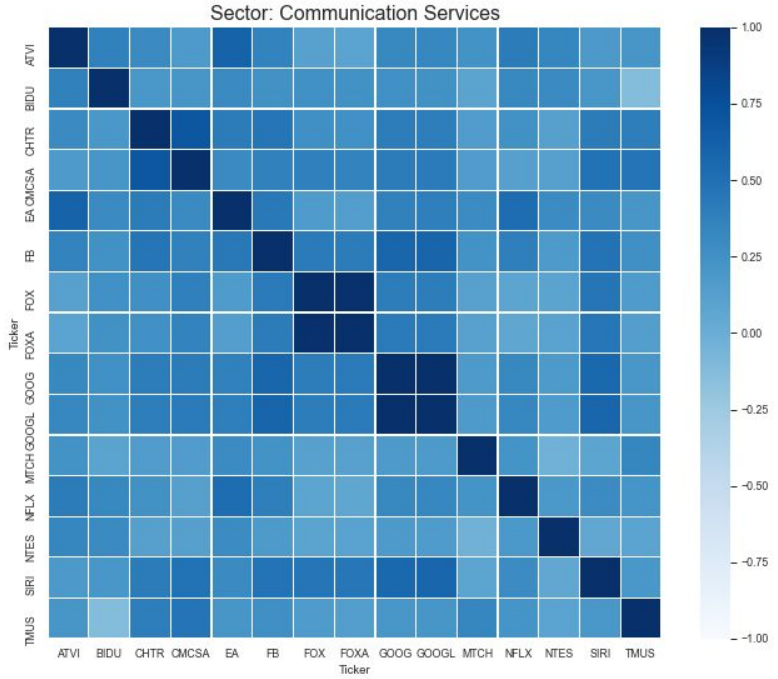


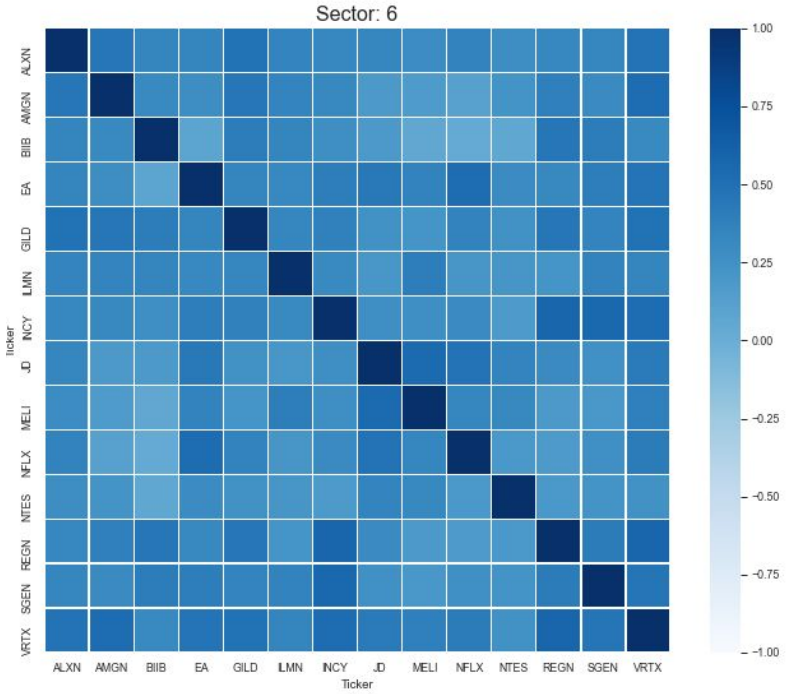
Results: Information Technology



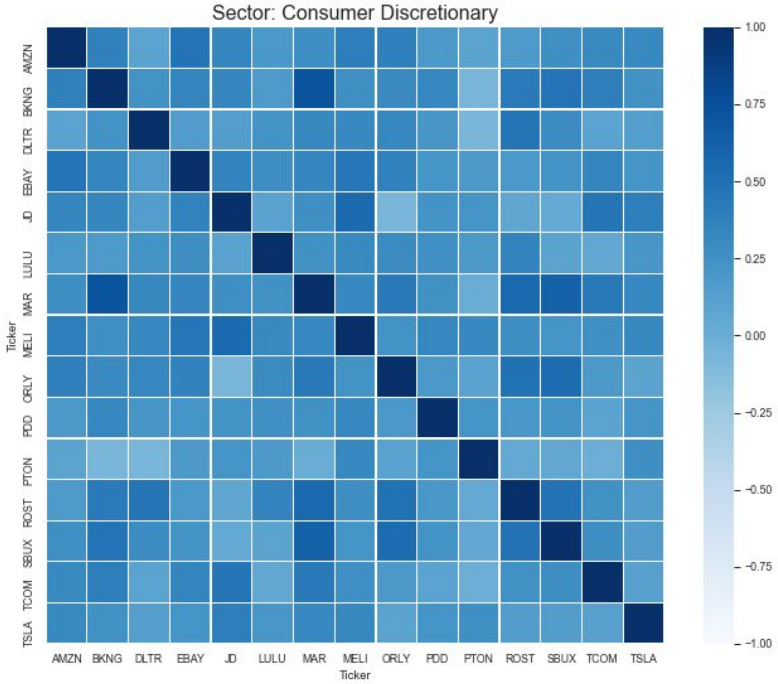


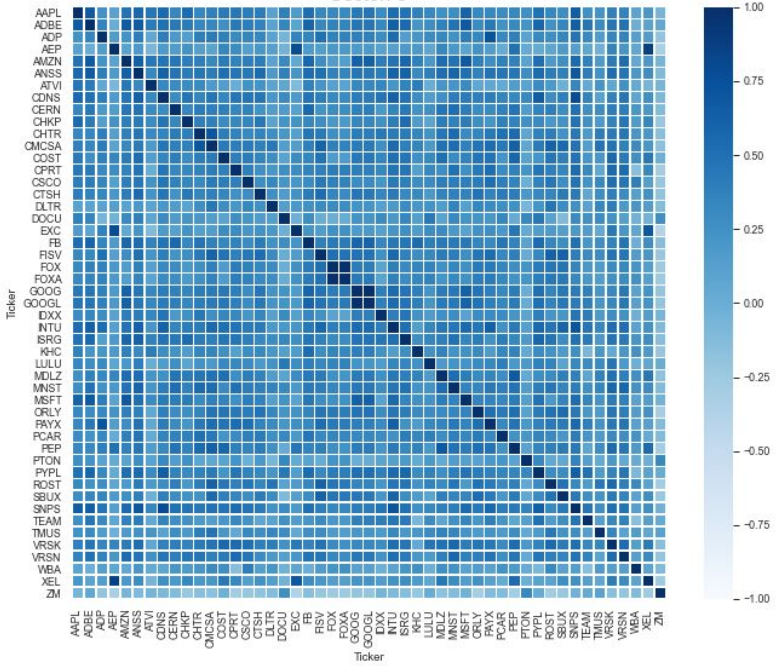
Results: Communication Services





Results: Consumer Discretionary





Conclusion

Our k-means algorithm was able to cluster the stocks in the QQQ fund differently than what it originally had. When comparing the heatmaps of the three largest sectors in the original and predicted data, we only see a slight positive difference between the Information Technology sector and its corresponding predicted sector. The same is true for the Consumer Discretionary sector, however false for the Communication Services sector. In future testing, we plan to fed more data into our model for which we need a more powerful computer to do. With more data in our model, we hope to see a stronger difference in correlations between sectors. We can also use machine learning to predict the stock price in the future and compare the returns between predicted and original sectors.

Acknowledgment

I like to dedicate this portion to express my gratitude to all parties involved in helping with this research. Firstly, thank you New Jersey City University for allowing me to be involved in the Summer Stem Internship. I would also like to give thanks to my mentor Dr. Nan Wang for guiding me through this experience and Weikang Kong for developing the website for this research project. Last but not least, I would like to acknowledge the **US Education Department Title III Part F HSI-STEM Grant # P031C160155 and US Education Department Title V DHSI Grant # P031S200124 for their generous grant offers.**