In this project, we would like to track the stock price change starting from January 1, 2019 to August 5, 2019, and further group the change to evaluate the performance by sector. In performing the ETL, we grabbed data in two types of formats from data sources, transformed and merge in Python environment and eventually manipulated and output in pgAdmin 4.

Extract: We extracted data from two different sources and format. CSV file listed all stock tickers, August 5, 2019 last price, sectors and other information in New York Exchange (source link see Appendix) and Json file through API retrieved from Financial Modeling Prep (source link see Appendix).

Transform: We transformed data in Python environment. For CSV. file cleaning, we performed dropna in important columns Sectors and Last price, drop extraneous columns such as industry, IPO year and output the cleaned data in CSV. format for further processing. On the other hand, we filtered only stocks last price on December 31, 2018 in Json file through API, and joined into cleaned CSV. data. After that we did calculations to show the stock price percentage change during time period.

Load: We loaded transformed data in pgAdmin 4 using Python psycopg2 and sqlalchemy packages. In pgAdmin 4, we queried to group the average percentage change in sectors to show our conclusion. pgAdmin 4 is a relational database that it is easier to aggregate data according to our needs.

Graph 1. Performance record grouped by Sectors
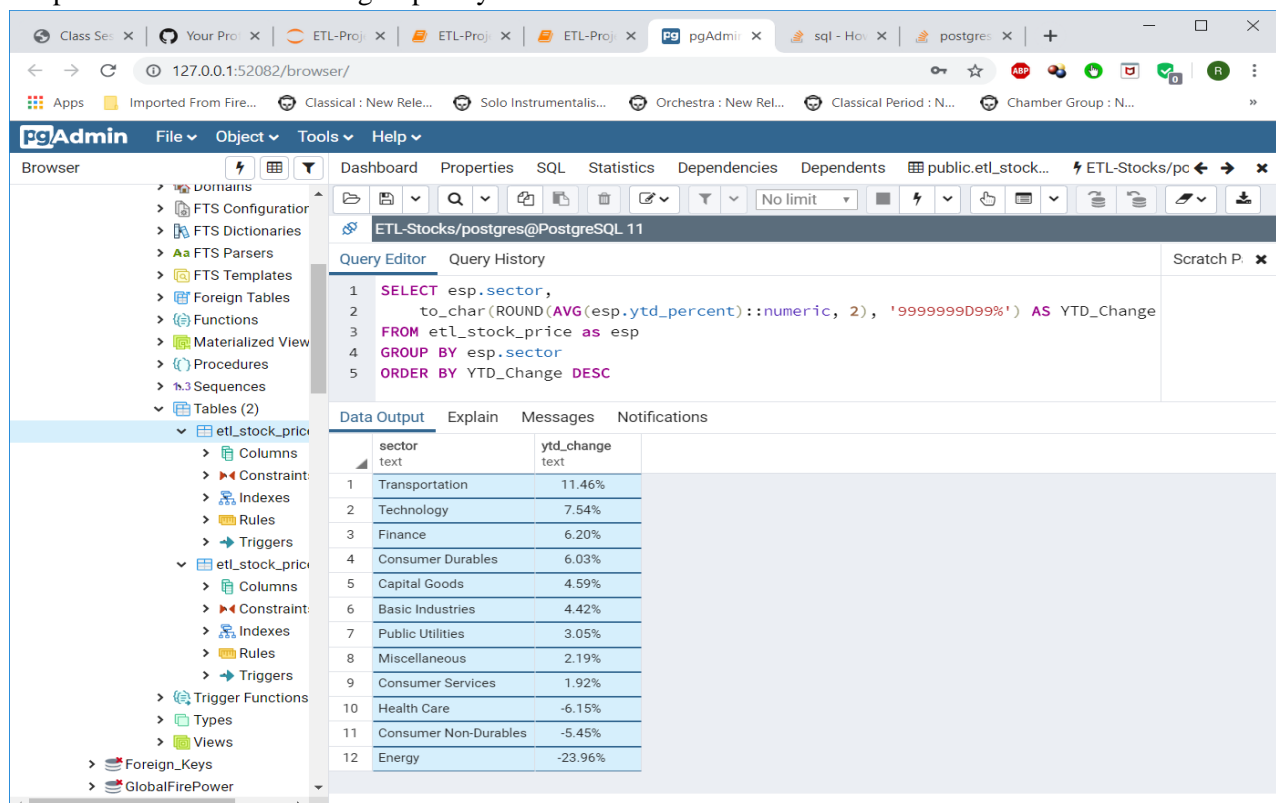
After ETL process, we find that the best three performed sectors from January 1, 2019 to August 5, 2019 are Transportation (11.46%), Technology (7.54%), Finance (6.20%). For other sector performance please refer to Graph 1.

Appendix:

New York Exchange stock data, source link https://www.nasdaq.com/screening/company-list.aspx

Financial Modeling Prep, source link: https://financialmodelingprep.com/api/v3/historical-price-full/