Created by

Mohamed Ramadan

Ahmed Abd Elmoneam

Shimaa Ashraf

Mai Abu Elmagd

Menna Ismail

S&P 500 Analysis project report

ITI Ismailia branch | Data Analytics Track

2022

Contents

[Data sources: 1](#_Toc109500341)

[Data: 1](#_Toc109500342)

[Data cleansing and preparation: 1](#_Toc109500343)

[Database: 3](#_Toc109500344)

[Our Schema Contains: 5](#_Toc109500345)

[1. 3 Dimensions: 5](#_Toc109500346)

[2. 2 Facts: 6](#_Toc109500347)

[3. 1 fact\_less Fact: (S&P index stocks) 7](#_Toc109500348)

[Queries: 8](#_Toc109500349)

[Data visualization: 8](#_Toc109500350)

[Conclusion: 8](#_Toc109500351)

# Introduction:

This project tracks the S&P 500 stocks. The project tracks the changes of stock price for each company and the change of S&P 500 index. This will help in drawing insights about the market change, identifying the top gainers and top losers.

# Data sources:

The data source used were:

* Investing.com (our main Data source)

<https://www.investing.com/>

* Wikipedia

<https://en.wikipedia.org/wiki/List_of_S%26P_500_companies>

* Datahub

[https://datahub.io/core/s-and-p-500-companies-financials#resource-constituents-financials](https://datahub.io/core/s-and-p-500-companies-financials%23resource-constituents-financials)

* Kaggle

# Datasets:

The data sets downloaded consist of:

* *Tesla Reasons:* contains the events that were the cause of the stoc price change for Tesla.
* *Cash Flow:* contain the cash flow statement for Tesla.
* *Company VS industry:* contains the ratios and KPIs for Tesla in comparison with its industry.
* *Forecast VS actual:* this contain revenue and eps that was estimated for each quarter and the actual values.
* *The S&P 500 companies list:* this data set contains the list of the companies in S&P 500 and their sectors and sub-sectors.
* *S&P 500 index daily stocks:* this index is a summarization to the performance of the 500 companies and you can better know and estimate the performance of any company compared with the other companies in S&P 500.
* *The S&P 500 companies’ daily stocks:* this data set contains the daily stocks of the 500 companies. This data can track the performance of each company through the years.
* *constituents-financials:* this data sets contains the annual ratios and KPIs of the stock market for each company in S&P 500.

# Data cleansing and preparation:

* The data range used was set.
* The S&P 500 companies’ daily stocks data set was too large to be processed in excel so Python was used as in the following steps:
  + Data frame was created.

stocks = pd.read\_csv('sp500\_stocks.csv')

* + Checking nulls

stocks.isnull().sum()

* + Removing Nulls

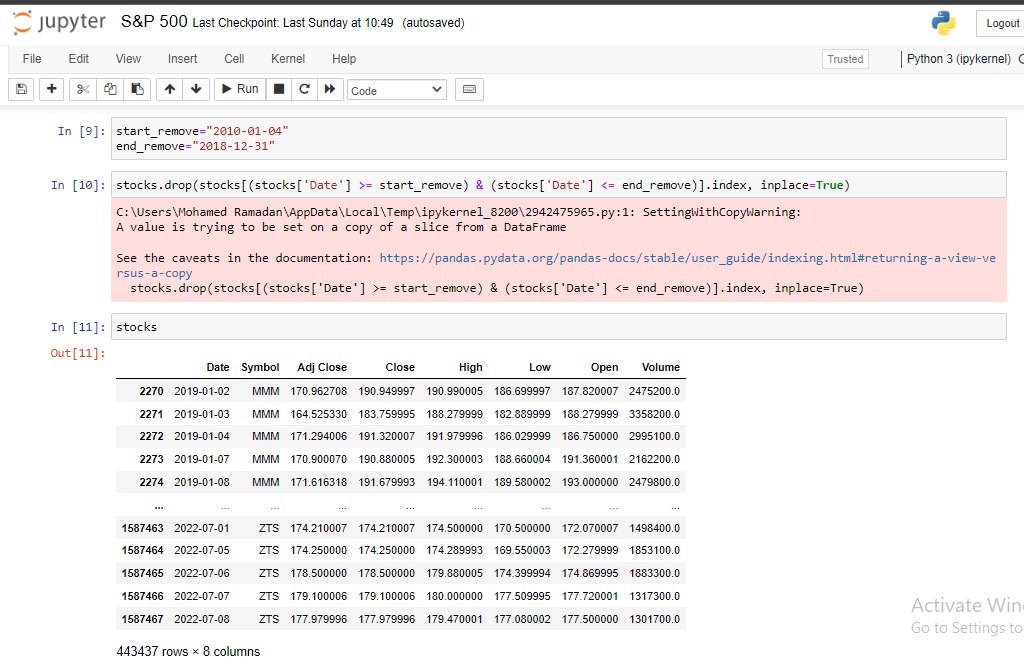
stocks=stocks.dropna()

* + Removing the unwanted data range using the date column

start\_remove="2010-01-04"

end\_remove="2018-12-31"

stocks.drop(stocks[(stocks['Date'] >= start\_remove) & (stocks['Date'] <= end\_remove)].index, inplace=True)



* The rest of data sets were cleansed and processed using excel.
* The range of data used was set and the unwanted data was removed.
* The data was cleansed from nulls and empty data.
* Some null data were filled from reliable online data sources manually like:

<https://www.investing.com/>

* The data was edited using excel and some columns were added from a data set to another using:

VLOOKUP function.

* The data was edited to be compatible with each other so that no missing data appears.
* The format was edited and transformed with the function:

=LEFT(cell,LEN(cell)-1)\*10^(SEARCH(RIGHT(cell),"kmbt")\*3)

So that all data would be in a numeric format.

* Some data in the constituents-financials dataset were derived and added to the dataset.
* The S&P 500 companies’ daily stocks was processed in excel after it was reduced in size that fit excel processing as the following:
  + The daily %chg was calculated then the performance was derived according to the %chg using:

=SUNIFS(sum range,symbol range,symbol cell,date range,”>=start date”,date range,”<=end date”)

* + Using this function the performance: %1 week, %1 month, %1 year, %3 years were calculated for each company.
  + **Note:** This dataset (>440,000 rows) is too large for the excel to process on it as a whole so the data range (columns) must be a fixed range not the whole column so that the excel sheet won’t crash.

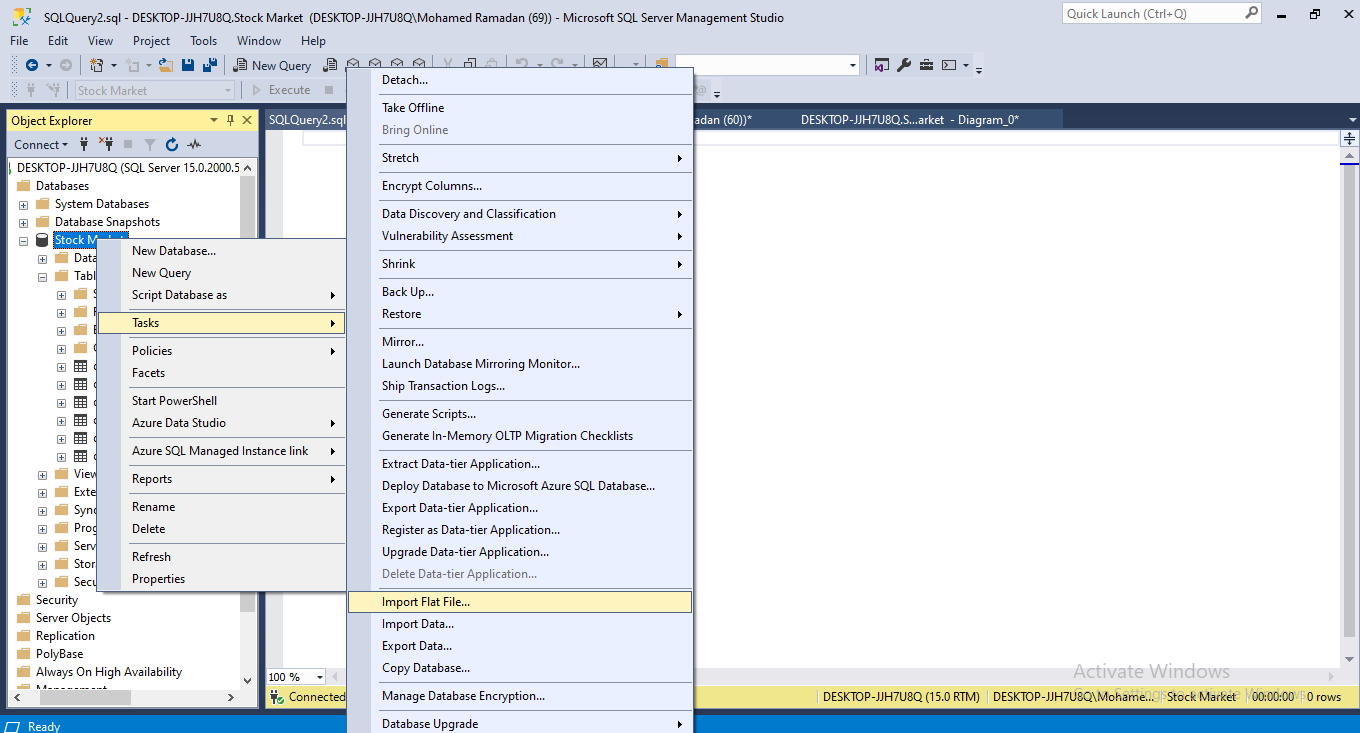
# Database:

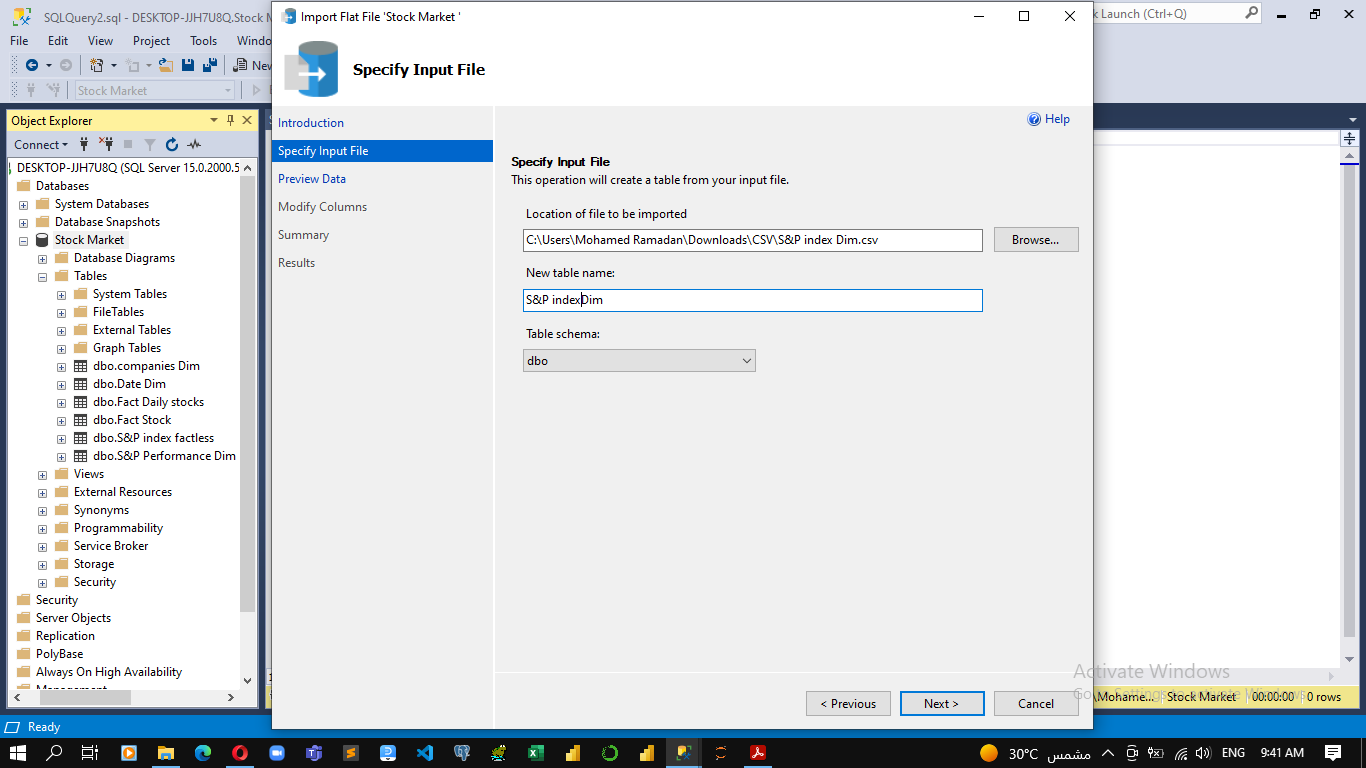
* At first, Database was built on SQL Server under the name Stock Market.

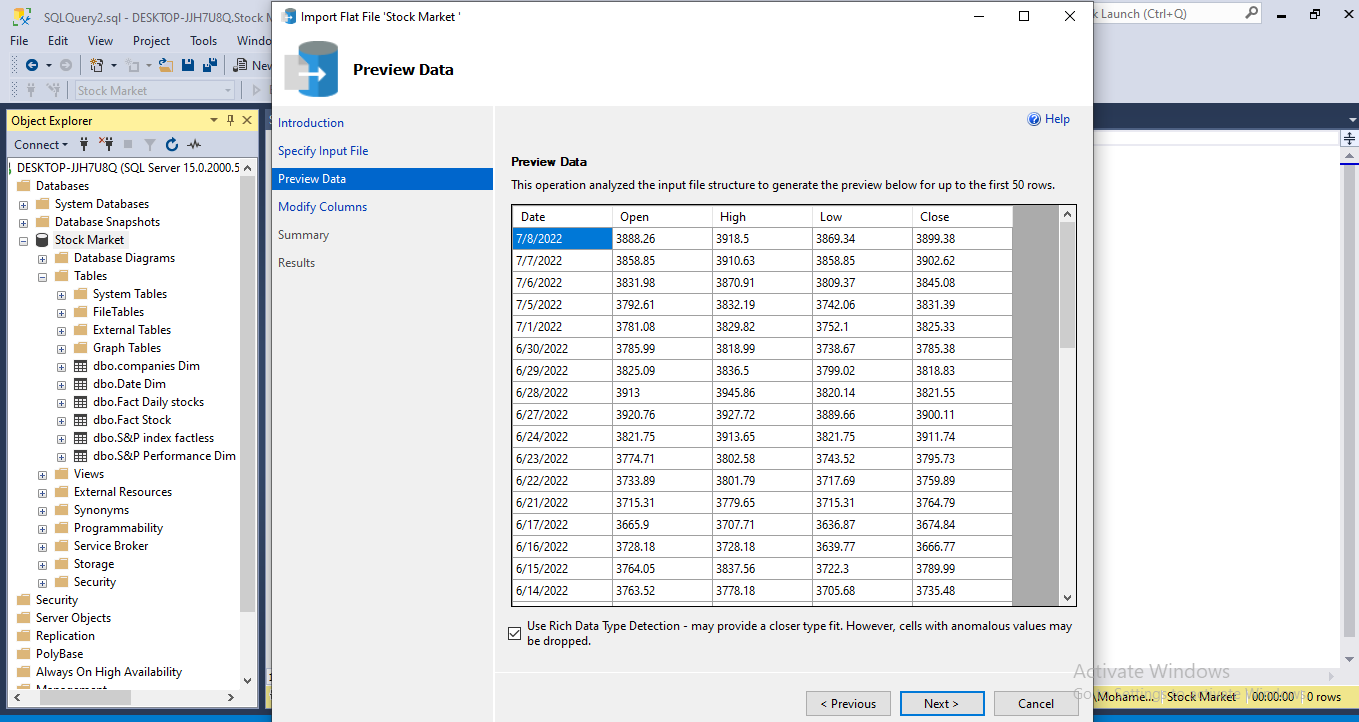
A picture containing text, screenshot, indoor

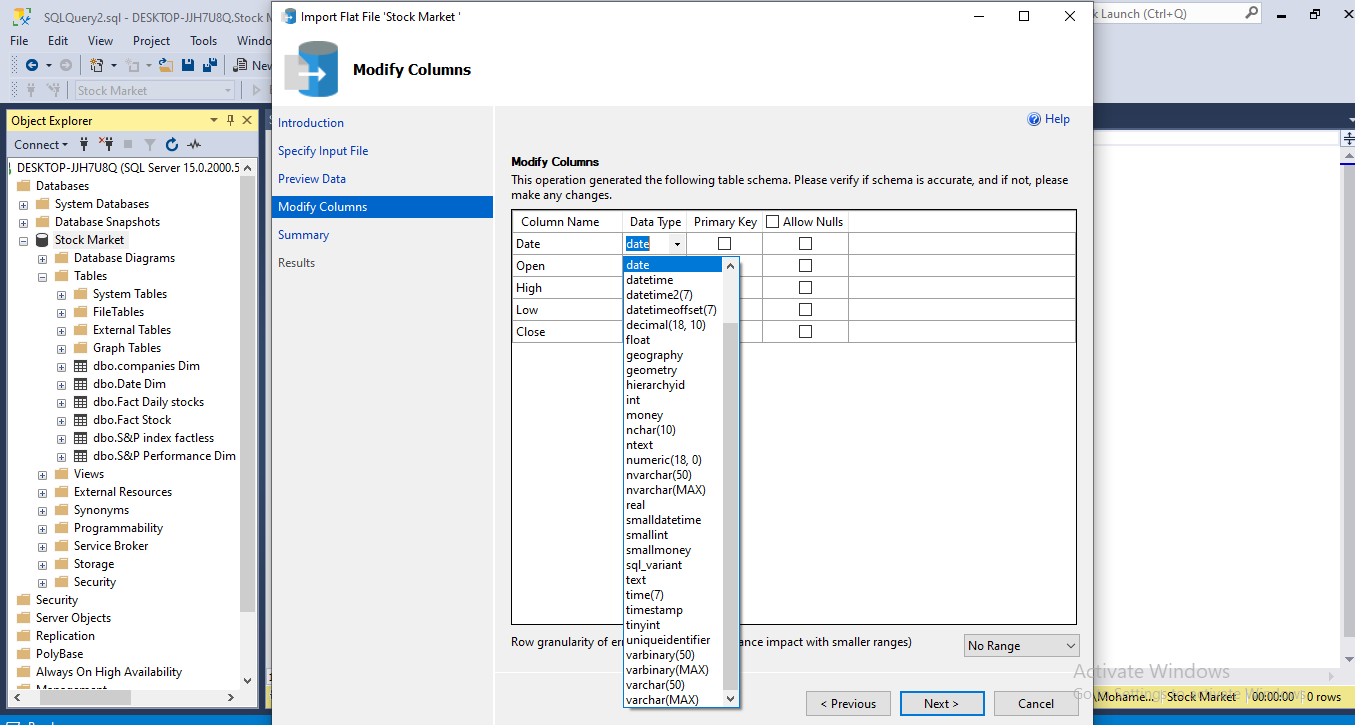
Description automatically generated

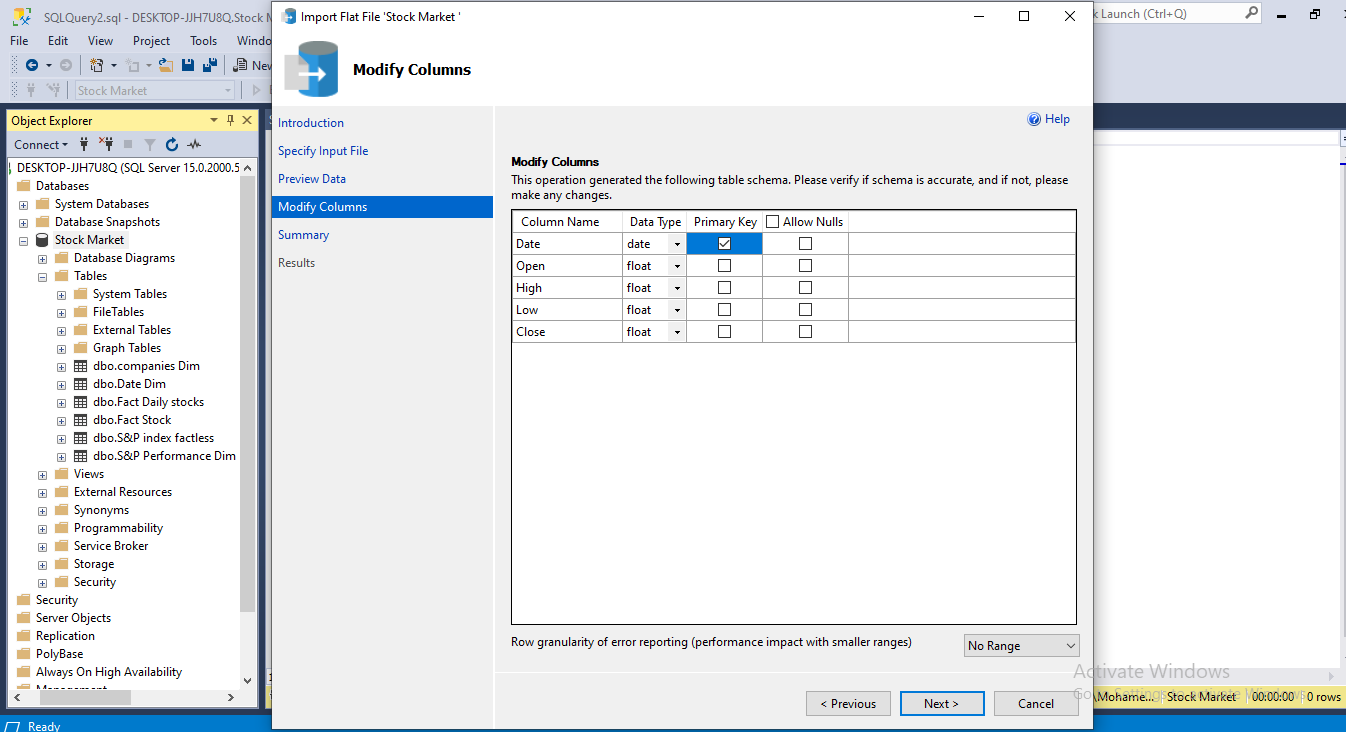
* Then the tables were uploaded to Database after cleaning all the data sources and dividing the tables using Excel and Python, as will be explained in the following pictures.

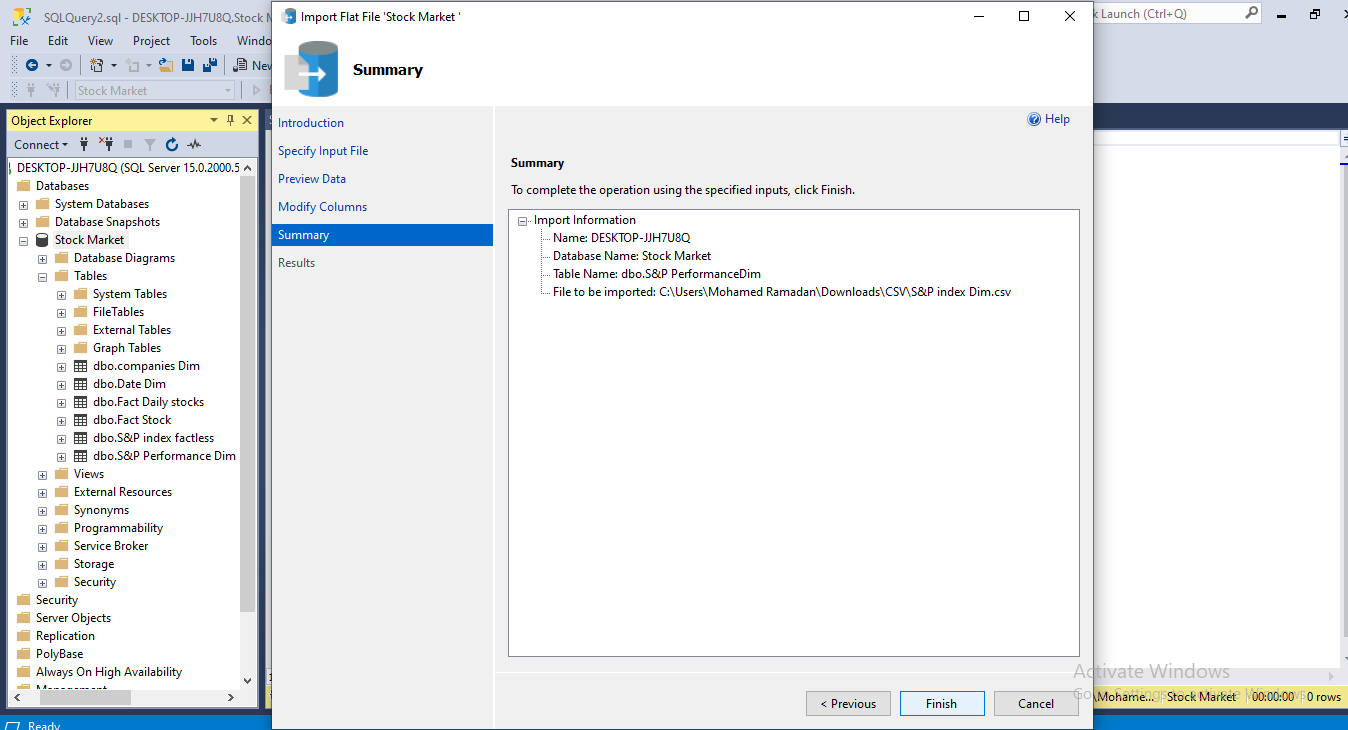












# Our Schema:

## 3 Dimensions:

* + **Companies Dim**: (List of S&P 500 companies)

This dimension was created to contain the name of each company and its symbol, in addition to the public sector to which this company belongs, as well as the sub-sector.

* + **Date Dim**:

This dimension was created using DAX Query on Power BI and then filtered its content to suit the extension of the existing data

(= List.Dates(#date(2010,01,01), 365.25\*13, #duration(1,0,0,0))).

* + **S&P Performance Dim**:( S&P Performance)

This dimension includes the performance of companies over more than one range, so that it is possible to know the company's performance during a week, a month, a year, or even three years combined.

## 2 Facts:

* + Daily Stocks: (The S&P 500 companies’ daily stocks)

This table contains a daily follow-up of the stocks,

As it contains the stock price at the beginning of the stock exchange day and the stock price at the end of the day at the close. It also contains the highest value reached by the stock on the day and the lowest value, in addition to the trading and purchase volume of this stock.

* + Annual Stocks:( constituents-financials)

This table contains an annual report on the stock. The table shows Market capitalization, Shares outstanding (TOTAL SHARES), Earnings per share (EPS), The price-to-earnings ratio (P/E ratio), Dividends, The dividend yield, The 52-week high/low and The price-to-sales (P/S) ratio and Book value per common share

Which is calculated by the following equation

Text

Description automatically generated

**price-to-book ratio (P/B ratio)**

Text

Description automatically generated with medium confidence

## 1 fact\_less Fact: (S&P index stocks)

* + S&P 500 index daily stocks: this index is a summarization to the performance of the 500 companies and you can better know and estimate the performance of any company compared with the other companies in S&P 500.
* After loading the tables on the Database, relationships were created between the tables to form our model.

Text

Description automatically generated

* The final form of the model (Galaxy Schema)

Diagram

Description automatically generated

# Queries:

Select date, open from s&p index factless

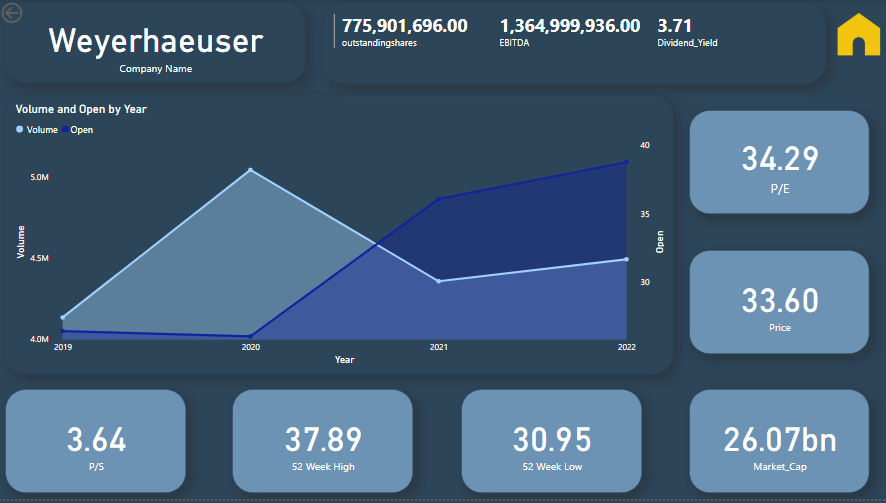
Select sector, earning/share,revenue,netincome,market cap,price\_book from fact stock , companies dim where fact\_stock.symbol = companies\_dim.symbol

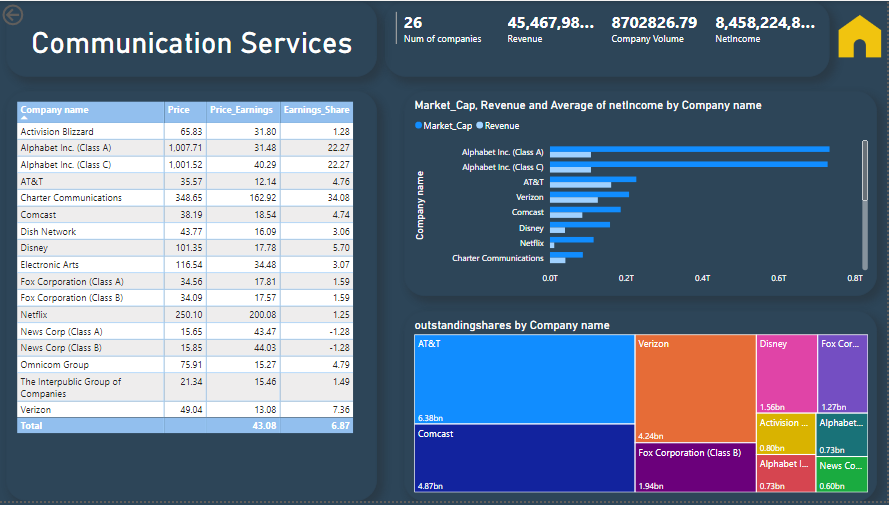
Select price/earning,earning/share,sector,revenue,outstanding shares,market cap,netincome from companies\_dim,fact\_stock where companies\_dim.symbol=fact\_stock.symbol

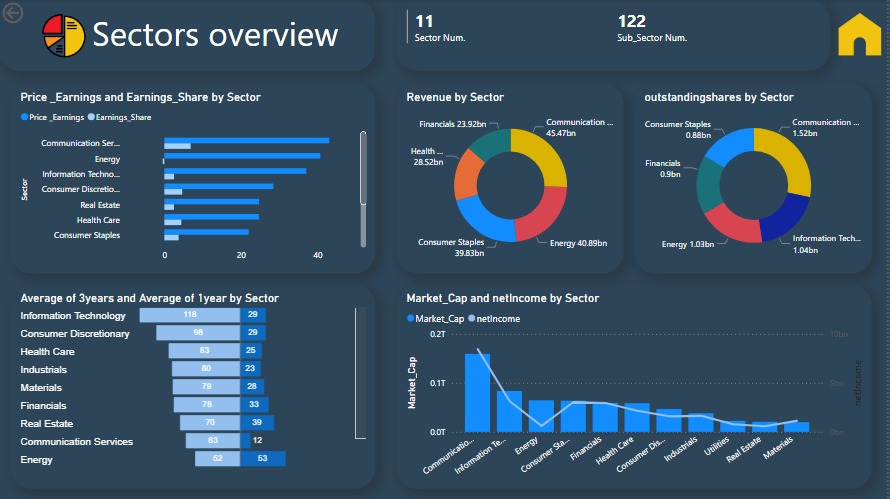
Select 1year,3year,sector from fact\_daily\_stock,companies\_dim where fact\_daily\_stock.symbol=companies\_dim.symbol

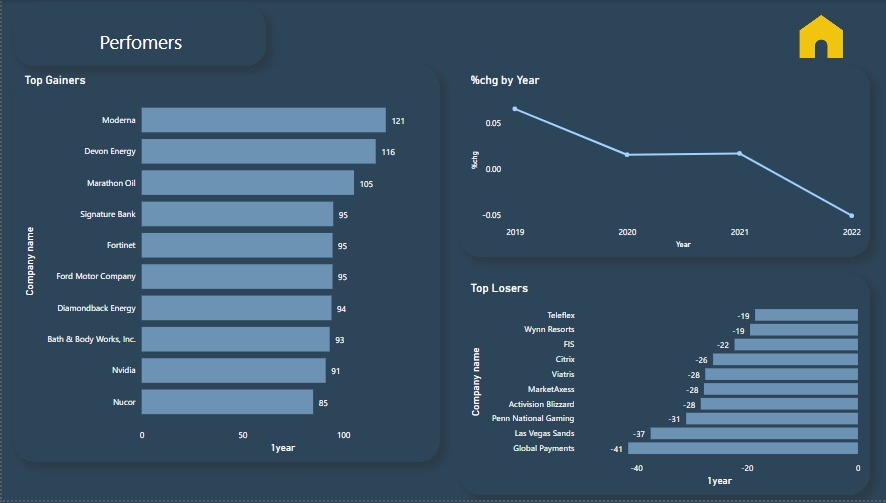
# Data visualization:











# Conclusion:

The visualization in the first dashboard clearly showed the drop in stock market in 2020 due to the pandemic and shows how it strongly recovered by the end of the year. The top gainers, top losers and avg chg. in price were identified.