**Information technology institute**

**Data visualization Track**

**DWBI Project**

**Architect, Populate and Explore**

**Data Warehouse For Stock Market Analysis**

**(Graduation Project Documentation)**

**Submitted By:**

**Abdelrahman Gamal**

**Mazen Magdy**

**Ahmed Osama Maher**

**Nov. 2022**

**Table of Contents**

[**1 Introduction 3**](#_Toc118309044)

[**1.1 Introduction 3**](#_Toc118309045)

[**1.2 Business Process 3**](#_Toc118309046)

[**2 Data Sources 4**](#_Toc118309047)

[**2.1 S&P 500 Component Stocks 4**](#_Toc118309048)

[**2.2 S&P 500 Index Stock Data 6**](#_Toc118309049)

[**2.3 S&P 500 Historical Data 7**](#_Toc118309050)

[**2.4 Constituents Financials 8**](#_Toc118309051)

[**2.5 Gold Price Dataset 10**](#_Toc118309052)

[**3 Data Warehouse Data Model 11**](#_Toc118309054)

[**3.1 Logical Model Structure : 11**](#_Toc118309055)

**3.2 Bus Matrix and Model:…………………...……………………………………………………12**

[**4 Physical Model & Importing Data 13**](#_Toc118309060)

[**5 BI Queries 15**](#_Toc118309061)

**6 Visualization……………….…………………………………………………………………………17**

[**Conclusion 20**](#_Toc118309062)

[**Appendix 21**](#_Toc118309063)

**Tasks………………………………………………………………………………………………….…25**

# **1. Introduction**

## 1.1 Introduction

Stock analysis refers to the method that an investor or trader uses to evaluate and investigate a particular trading instrument, investment sector, or the stock market as a whole. Stock analysis is also called equity analysis or market analysis. Investors or traders make buying or selling decisions based on stock analysis information.

Also, the Stock analysis helps traders to gain an insight into the economy, stock market, or securities. It involves studying the past and present market data and creating a methodology to choose appropriate stocks for trading. Stock analysis also includes the identification of ways of entry into and exit from investments.

The S&P 500 tracks the stock performance and market capitalization of the roughly 500 companies included in the index which reflects almost 80% of the whole USA economy, measuring the value of the stock of those companies.

The index is one of the factors in the computation of the conference board leading economic index, used to forecast the direction of the economy.

## 1.2 Business process

* Reporting Stock Market Daily Statistics
* Reporting Top And Lowest Performers
* Tracking Stock Market Changes

# **2. Data Sources**

## 2.1 S&P 500 Component Stocks

**Link Source**: [List of S&P 500 companies - Wikipedia](https://en.wikipedia.org/wiki/List_of_S%26P_500_companies)

**Description:**

The [**S&P500**](https://en.wikipedia.org/wiki/S%26P_500) [stock market index](https://en.wikipedia.org/wiki/Stock_market_index) is maintained by [S&P Dow Jones Indices](https://en.wikipedia.org/wiki/S%26P_Dow_Jones_Indices). It comprises **503** [common stocks](https://en.wikipedia.org/wiki/Common_stock) , which are issued by **500** [large-cap](https://en.wikipedia.org/wiki/Market_capitalization) companies traded on American stock exchanges (including the 30 companies that compose the [Dow Jones Industrial Average](https://en.wikipedia.org/wiki/Dow_Jones_Industrial_Average)). The index includes about 80 percent of the American equity market by capitalization. It is weighted by [free-float](https://en.wikipedia.org/wiki/Capitalization-weighted_index#Free-float_weighting) market capitalization, so more valuable companies account for relatively more weight in the index. The index constituents and the constituent weights are updated regularly using rules published by S&P Dow Jones Indices. Although called the S&P 500, the index contains 503 stocks because it includes two share classes of stock from 3 of its component companies.

**Attributes:**

* Symbol: Stock symbol is an [abbreviation](https://en.wikipedia.org/wiki/Abbreviation) used to uniquely identify publicly traded shares of a particular on a particular stock market.
* Security: The Term “Security” Refers to a Financial Instrument that holds some type of monetary value. However, In this context it is sufficient to say it is just the name of the corporation.
* Sec Fillings : The SEC filing is a [financial statement](https://en.wikipedia.org/wiki/Financial_statements) or other formal document submitted to the [U.S. Securities and Exchange Commission](https://en.wikipedia.org/wiki/U.S._Securities_and_Exchange_Commission) (SEC). [Public companies](https://en.wikipedia.org/wiki/Public_company), certain [insiders](https://en.wikipedia.org/w/index.php?title=Insider&action=edit&redlink=1), and [broker-dealers](https://en.wikipedia.org/wiki/Broker-dealer) are required to make regular SEC filings. [Investors](https://en.wikipedia.org/wiki/Investor) and financial professionals rely on these filings for information about companies they are evaluating for investment purposes.
* GICS Sector : The Global Industry Classification Standard (GICS) is an [industry taxonomy](https://en.wikipedia.org/wiki/Industry_taxonomy) developed in 1999 by [MSCI](https://en.wikipedia.org/wiki/MSCI) and [Standard & Poor's](https://en.wikipedia.org/wiki/Standard_%26_Poor%27s) (S&P) for use by the global financial community. The GICS structure consists of 11 sectors, 24 industry groups, 69 industries and 158 sub-industries into which S&P has categorized all major [public companies](https://en.wikipedia.org/wiki/Public_company).
* GICS Sub-Industry: It’s the same as GICS Sector attributes, Except that the sub-industry has a higher level of detailed attribute.
* Headquarters Location: The main physical location of the corporation.
* Date First Added: The date that this stock was firstly into duce d to be traded in the public stock market.
* CIK : A Central Index Key or CIK number is a number given to an individual, company, or foreign government by the [United States Securities and Exchange Commission](https://en.wikipedia.org/wiki/United_States_Securities_and_Exchange_Commission). The number is used to identify its filings in several online databases, including [EDGAR](https://en.wikipedia.org/wiki/EDGAR).
* Founded: Contains the year the corporation was first founded

**Preparation and Cleansing:** Using Power Query:  
 Checked null values  
 Columns like SEC Fillings, CIK, Founded and Date Added have been removed  
 Separated Headquarters column to City and State.

## 2.2 S&P 500 Stock Data

**Link Source**: <https://www.kaggle.com/camnugent/sandp500>

**Description**: This data set includes daily changes for the stock for the past 5 years for all companies currently found on the S&P 500 index. This data set was first published in 2017 and was lately updated on Feb.2018.

It consists of 619,400 rows; Each row represents the daily numbers of a single stock.   
There are a total of 505 corporations that were collected over 1825 days between 08/02/2013 and 07/02/2018 represented with USD currency.

**Attributes:**

* **Date**: The date on which a transaction was conducted.
* **Open**: The price at which a stock started on a specific day.
* **High**: The highest price reached by a certain stock on a specific day.
* **Low**: The lowest price reached by a certain stock on a specific day.
* **Volume**: The number of shares traded by a certain stock on a specific day.
* **Name**: The symbol name of a stock.

**Preparation and Cleansing:** Using Power Query:  
 No Cleaning process needed, Data was ready to use.

## 2.3 S&P 500 Historical Data

**Link Source**: [S&P 500 Historical Data | Kaggle](https://www.kaggle.com/datasets/henryhan117/sp-500-historical-data)

**Description**: This data set includes the historical data of S&P 500 index from 30/12/1927 to 04/11/2020.This data set was both published and updated in Nov.2020. This data set included the same attributes mentioned in the source above having the attributes aggregated on whole 505 corporate totals.

**Attributes:**

* **Date**: The date on which a transaction was conducted.
* **Open**: The price at which a stock started on a specific day overall the market
* **High**: The highest price reached by a certain stock on a specific day overall the market
* **Low**: The lowest price reached by a certain stock on a specific day overall the market
* **Volume**: The number of shares traded by a certain stock on a specific day overall the market

**Preparation and Cleansing:** Using Power Query:  
 Removing the unwanted data range, and set the range from 2013 to 2018.

## 2.4 Constituents financials

**Link source:** [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)

**Description:** It contains statistics about each stock at the end of one year. Although it was not clear which year these statistics were reported, the data set was uploaded 4 years ago. It was assumed that the indicators belonged to 2018.

**Attributes:**

* **Symbol:** It is the of stock name abbreviation.
* **Name**: The symbol name of a stock.
* **Sector:** The Global Industry Classification Standard is an [industry.](https://en.wikipedia.org/wiki/Industry_taxonomy)
* **Price:** The price at which a stock started on a specific day.
* **Price/ Earnings:** It is the ratio for valuing a company that measures its current share price relative to its Earnings/Share. A high P/E ratio could mean that a company's stock is over-valued, or else that investors are expecting high growth rates in the future.

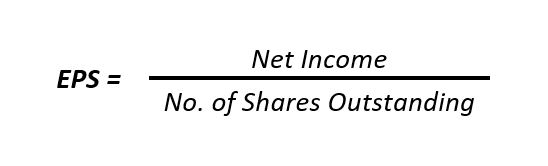
Diagram

Description automatically generated

* **Dividend Yield:** It is the amount of money a company pays shareholders for owning a share of its stock divided by its current stock price.



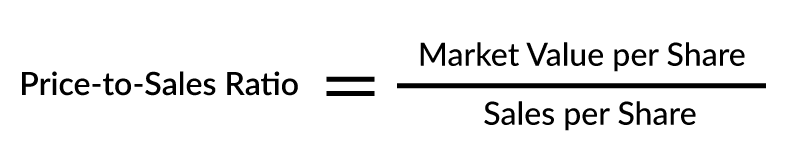
* **Earnings/Share**: It is a company's net profit divided by the number of common shares it has outstanding. The resulting number serves as an indicator of a company's profitability.



* **52 Week Low:** It is the highest price at which a stock has traded during the year.
* **52 Week High:** It is the lowest price at which a stock, has traded during the year.
* **Market Cap:** It is the total value of all a company's shares of stock. It is calculated by multiplying the price of a stock by its total number of outstanding shares.



* **EBITDA:** It is the company’s earnings before Interest, Taxes, Depreciation, and Amortization.
* **Price/Sales:** It shows how much investors are willing to pay per dollar of sales for a stock. It is typically calculated by dividing the stock price by the underlying company's sales per share.



* **Price/Book**: It measures the market's valuation of a company relative to its book value.

Text

Description automatically generated

## 

**Preparation and Cleansing:** Using Power Query:  
 Removed the unwanted Columns like SEC Fillings

## 2.5 Daily Gold Price Historical Data

**Link Source:** [Gold Price Prediction Dataset | Kaggle](https://www.kaggle.com/datasets/psycon/daily-gold-price-historical-data)

**Description**: Historically gold coinage was widely used as currency; when paper money was introduced, it typically was a receipt redeemable for gold coin or bullion. In a monetary system known as the gold standard, a certain weight of gold was given the name of a unit of currency. For a long period, the United States government set the value of the US dollar so that one troy ounce was equal to $20.67 ($0.665 per gram), but in 1934 the dollar was devalued to $35.00 per troy ounce ($0.889/g). By 1961, it was becoming hard to maintain this price, and a pool of US and European banks agreed to manipulate the market to prevent further currency devaluation against increased gold demand.

This data was first added on May 2022 and was last up dated on Sep. 2022

**Attributes:**

* **Date**: The date on which a transaction was conducted.
* **Open**: The price at which gold price started on a specific day**.**
* **High**: The highest price reached by gold price on a specific day.
* **Low**: The highest price reached by a certain stock on a specific day.
* **Close**:  The price at which a gold price ended on a specific day.
* **Volume**: The number of traded gold on a specific day.

**Preparation and Cleansing:** **Using Power Query:**  
 Removing the unwanted data range, and set the range from 2013 to 2018.

# **3. Data Warehouse Data Model**

## 3.1 Logical Model Structure

## We used Galaxy Schema to implement DWH model

## Business Process: Stock market analysis and portfolio recommendations.

## Granularity: Per day.

## Dimensions:

## Date: Contains full date, Year, Month, Quarter, Week, Day

## Company: keeps information about corporations that are components of the S&P500 index.

## Facts:

## Stock Market: Has information as price, price/earnings, dividend yield, earning/share, 52-week high/low, market cap, Ebitda and price/sale.

## Daily Monitoring: Contains information about the daily change for every stock like: open price, low, high, close, and volume

## SP500 Index: Contains information about the daily change for every overall companies open price, low, high, close, and volume

## Gold: Contains the change in gold price for every day

## 

## Diagram Description automatically generated3.2 Bus Matrix and Model

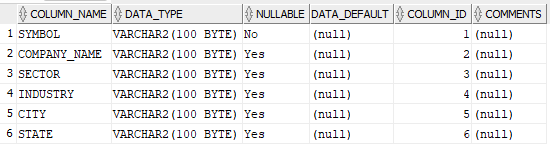
|  |  |  |
| --- | --- | --- |
| Facts\Dimensions | Date Dimension | Company Dimension |
| Stock Market Fact |  | **✓** |
| SP500 Index Fact | **✓** |  |
| Daily Monitoring Fact | **✓** | **✓** |
| Gold Fact | **✓** |  |

## 4. Physical model & Importing data

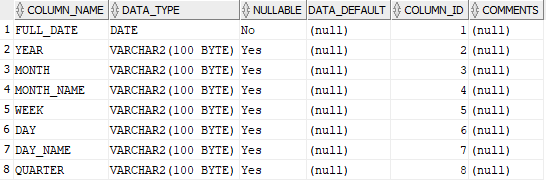
We created tables on SQL developer, we have Six tables in our data model.

First, we create a new Database called Stock, then we created the tables.

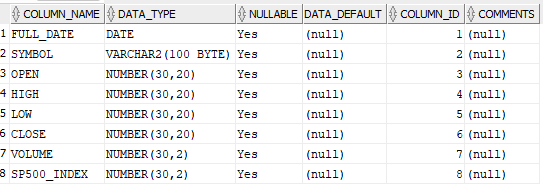
**First table:** Company dimension table



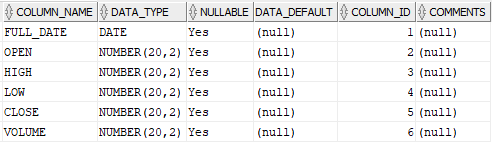
**Second table:** Date dimension table



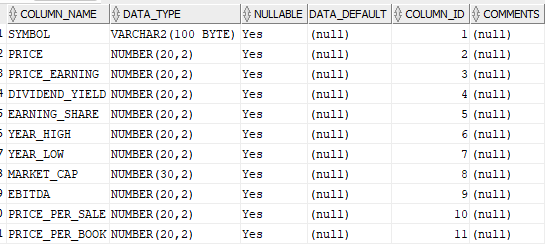
**Third table:** Daily monitoring fact table



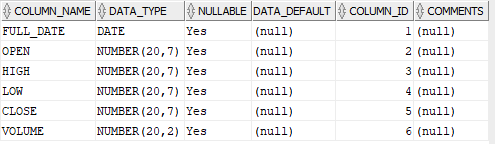
**Forth table:** Gold fact table



**Fifth table**: Stock market fact



**Sixth table:** S&P index fact

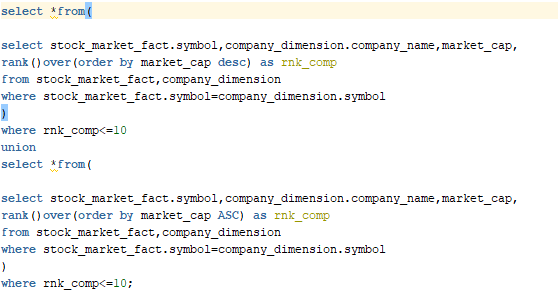


After creating tables on SQL developer we data was imported from Excel CSV files to SQL developer program.

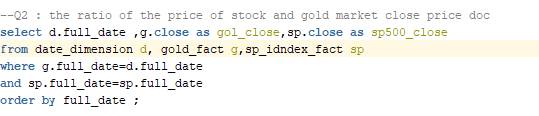
# 

# **5. BI Queries**

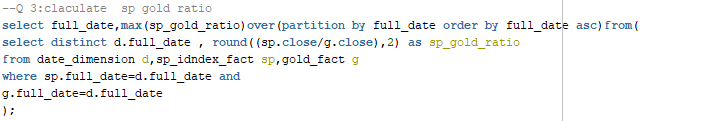
**1: Top and Bottom 10 stocks of market cap**



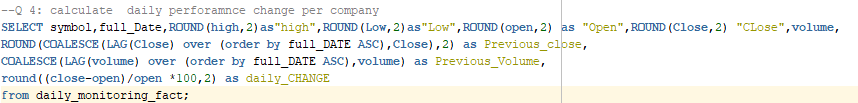
**2: Relation between the price of stock and gold market**



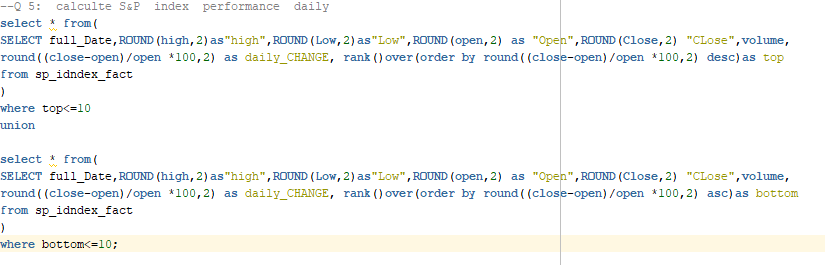
**3: S&P To Gold Correlation Ratio**



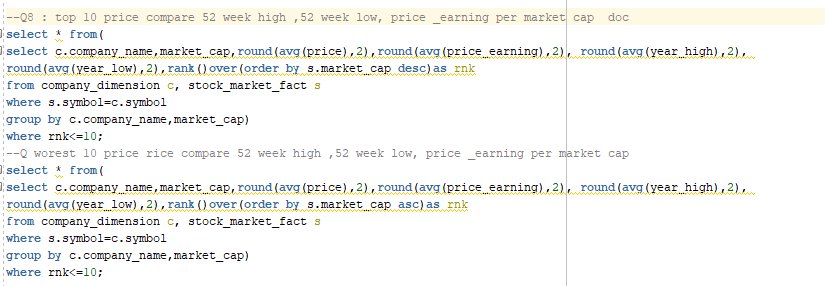
**4: Daily Performance Change Per Company**



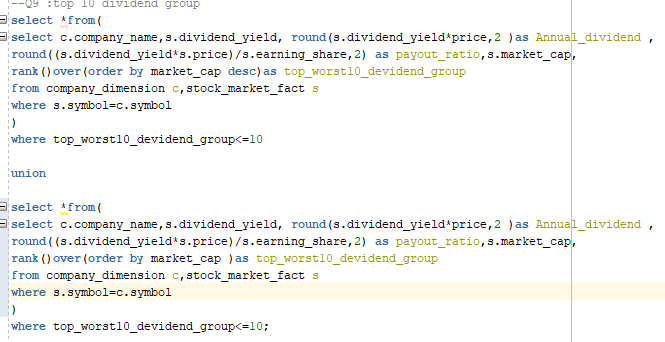
**5: S&P500 Index Daily Performance**



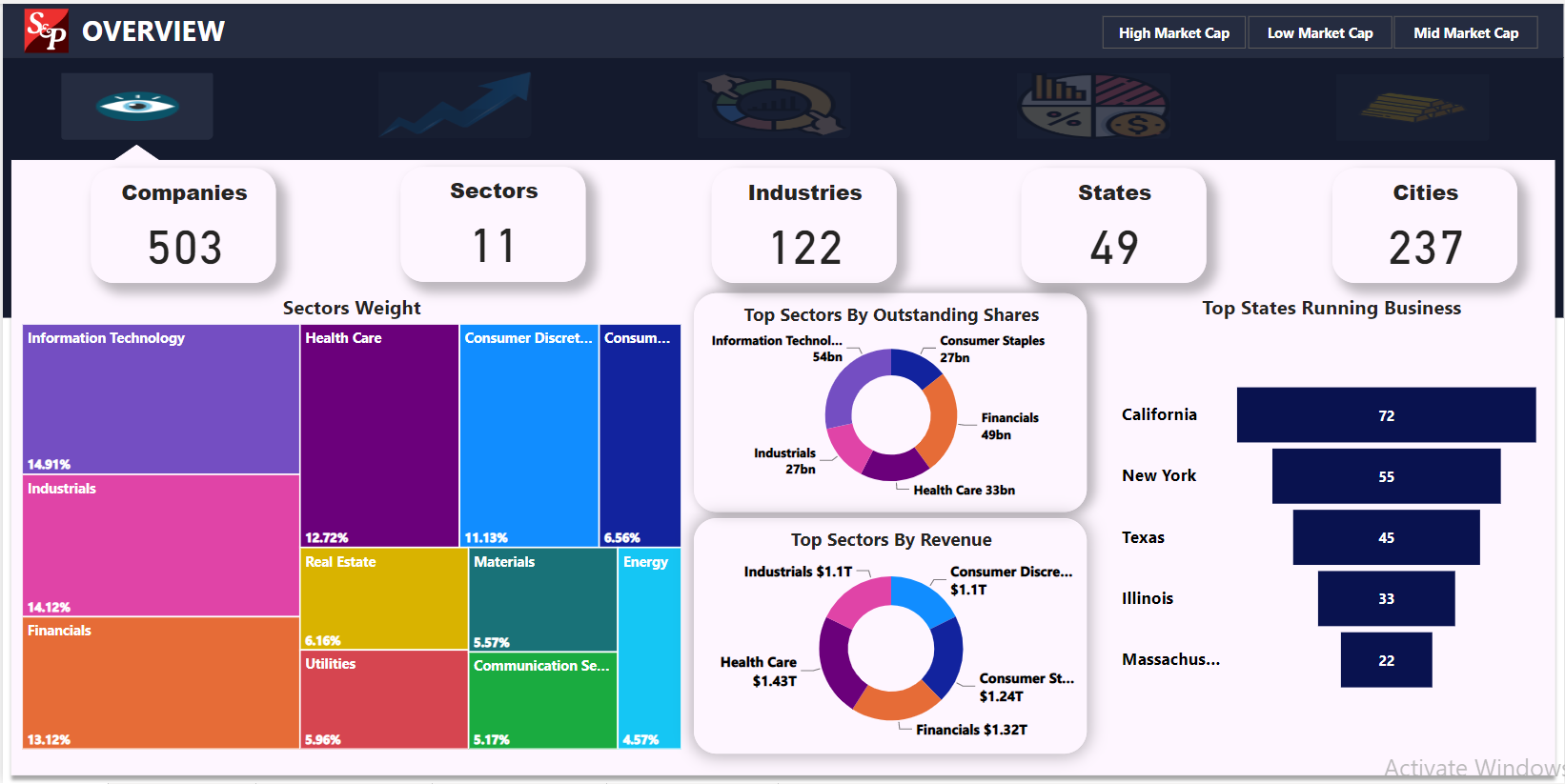
**6: Top And Bottom Prices Compared to 52 Week High, 52 Week Low**

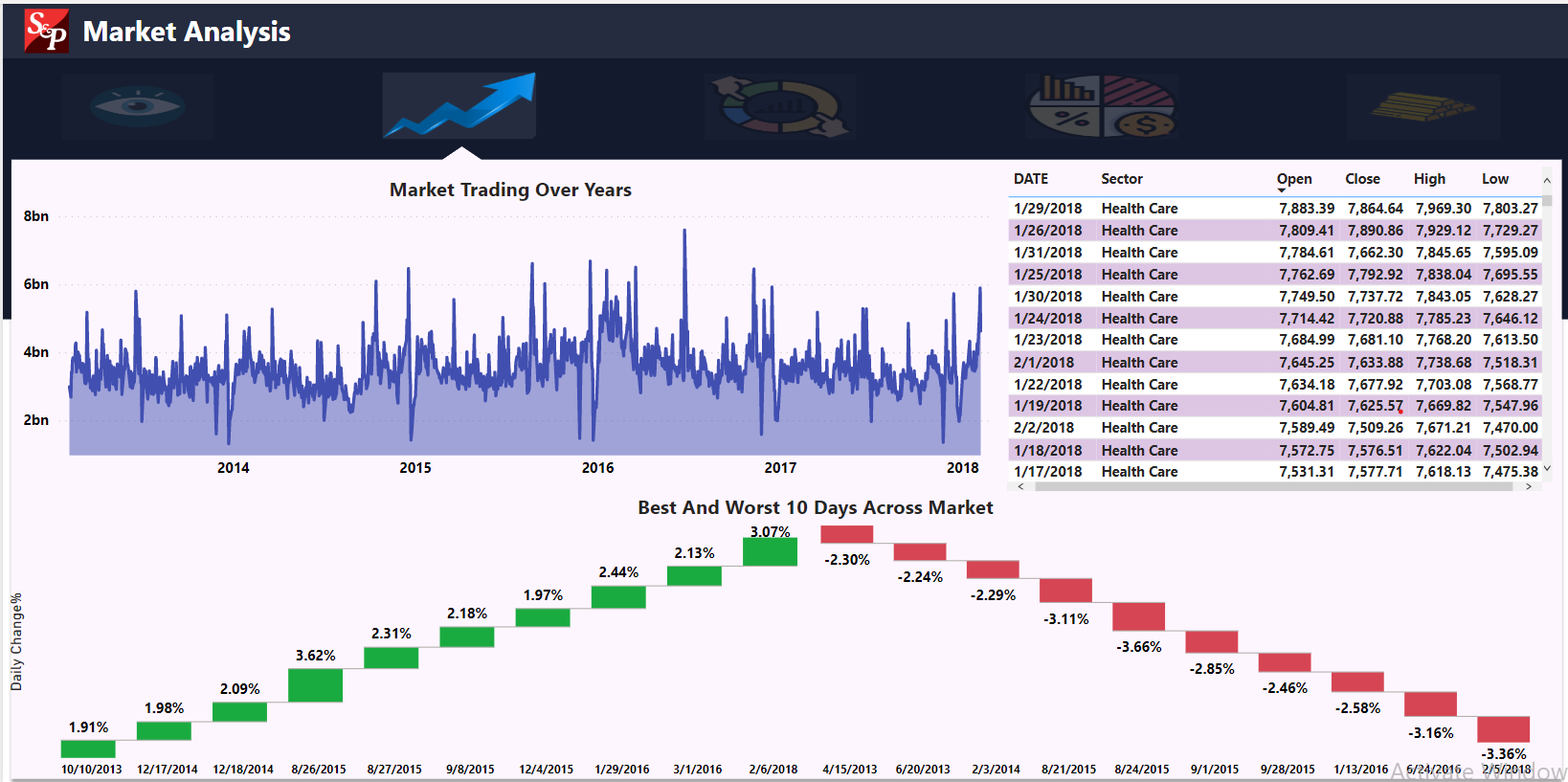


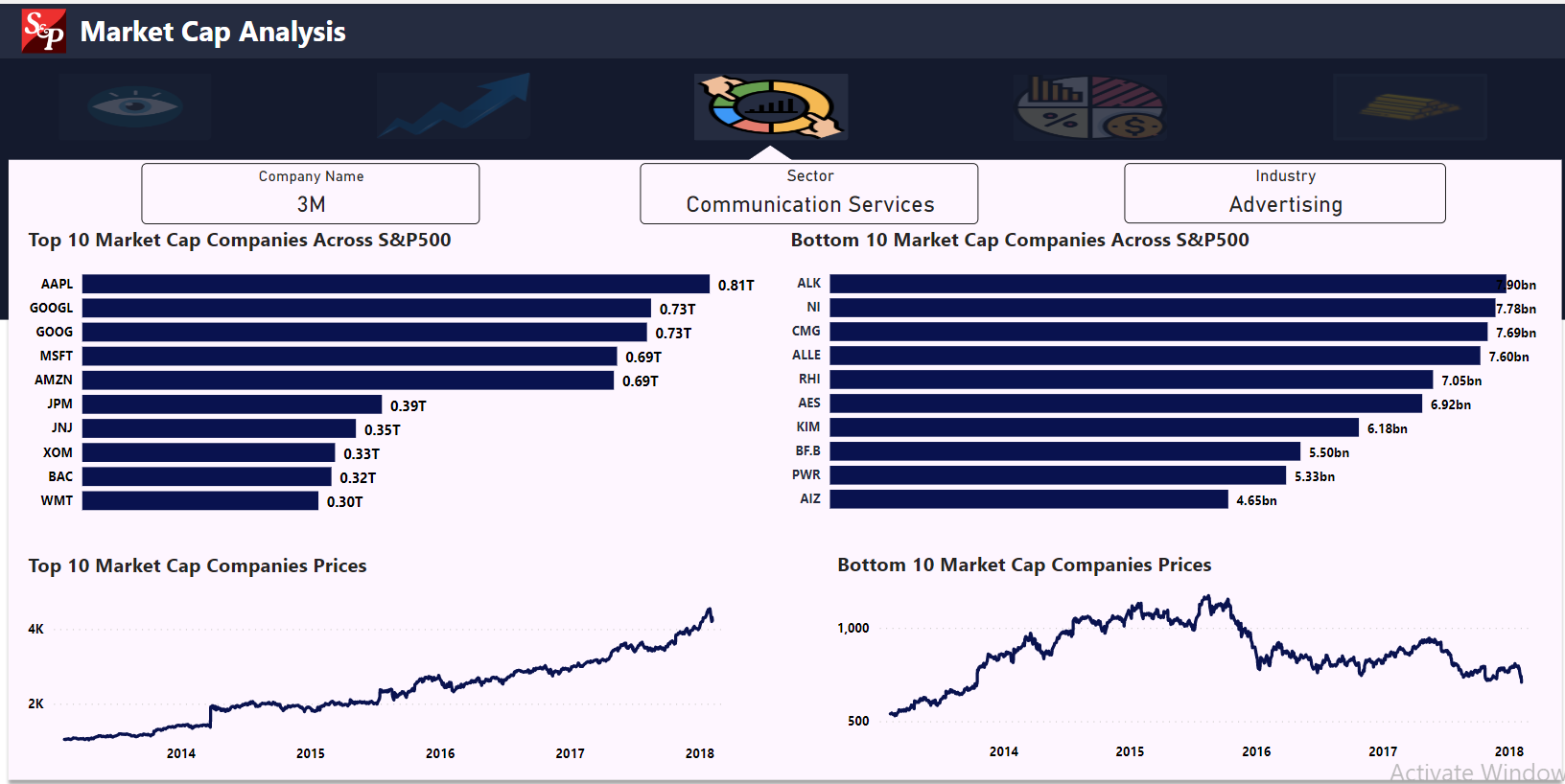
**7: Top And Bottom Companies Paying Dividends**

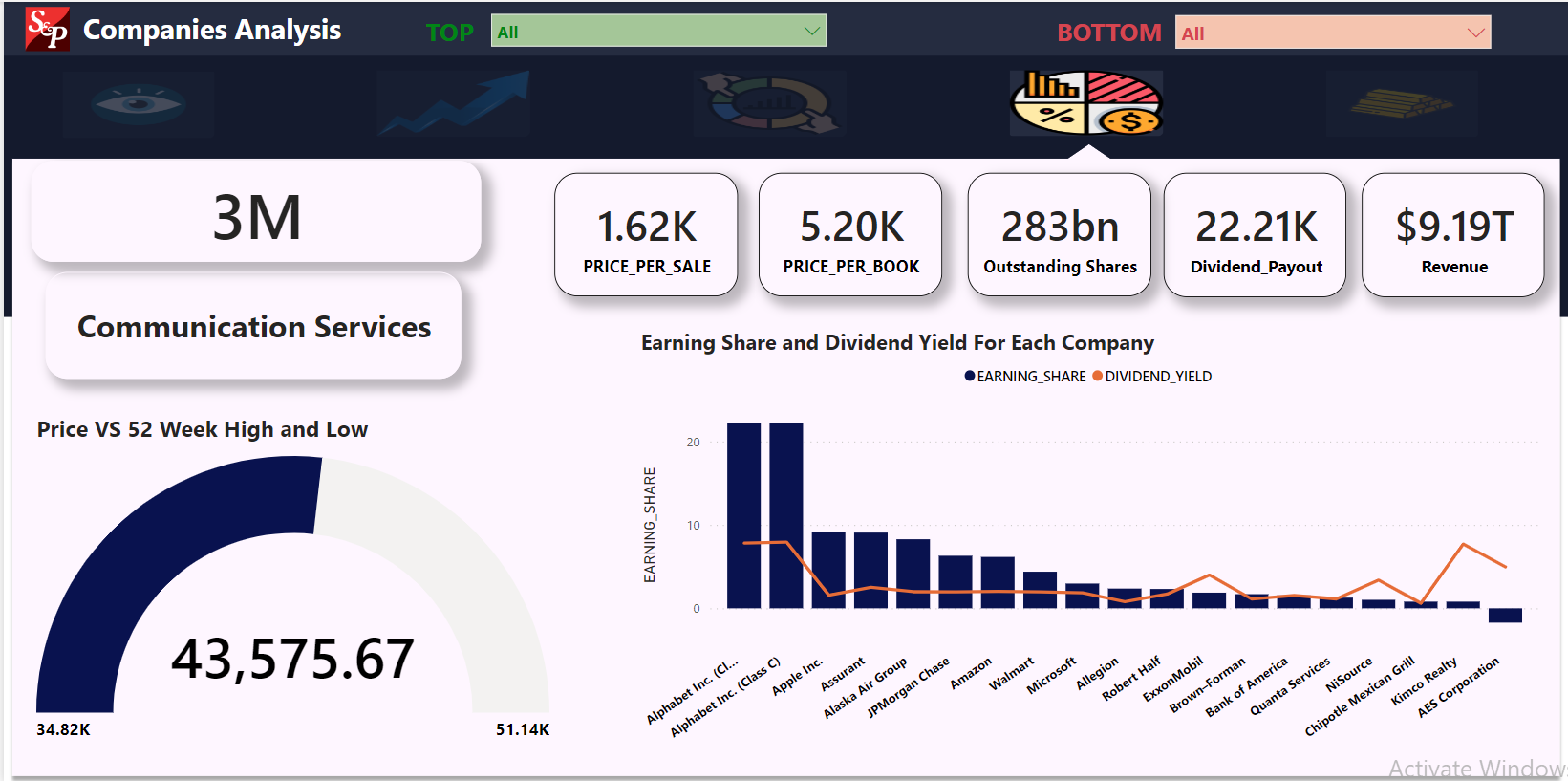


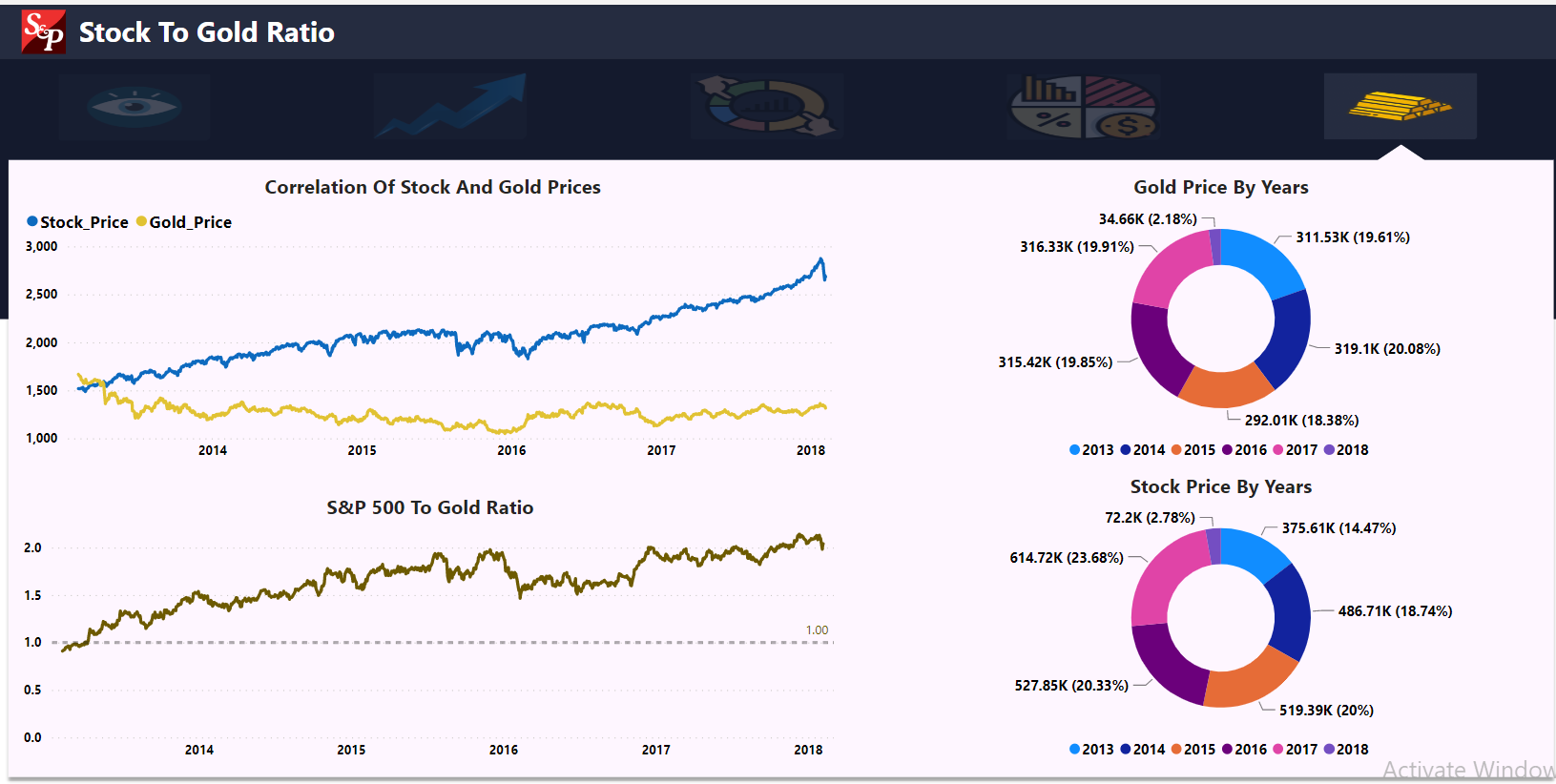
**6. Visualization Using PowerBI:**

****









# **Conclusion**

Project walkthrough:

1. Data understanding
2. Data gathering
3. Creating logical design (DWH model)
4. Physical design (Create tables on DBMS)
5. Import data on database
6. Building queries to answer the questions
7. Import files into power BI to build interactive dashboard

**Technologies Used In Project :**

* **Microsoft Excel** (using to import data from web and cleaning data)
* **SQL developer** (to create physical design for tables and queries)
* **Power BI** (Data visualization)

# **Appendix**

1. **SQL queries for creation tables**

**Company Dimension:**

CREATE TABLE COMPANY\_DIMENSION

(

SYMBOL VARCHAR2(100 BYTE) primary key,

COMPANY\_NAME VARCHAR2(100 BYTE),

SECTOR VARCHAR2(100 BYTE),

INDUSTRY VARCHAR2(100 BYTE),

CITY VARCHAR2(100 BYTE),

STATE VARCHAR2(100 BYTE)

)

**Date Dimension:**

CREATE TABLE DATE\_DIMENSION

(

FULL\_DATE DATE primary key,

YEAR VARCHAR2(100 BYTE),

MONTH VARCHAR2(100 BYTE),

MONTH\_NAME VARCHAR2(100 BYTE),

WEEK VARCHAR2(100 BYTE),

DAY VARCHAR2(100 BYTE),

DAY\_NAME VARCHAR2(100 BYTE),

QUARTER VARCHAR2(100 BYTE)

)

**Daily Monitoring Fact:**

CREATE TABLE DAILY\_MONITORING\_FACT

(

Full\_date date,

SYMBOL varchar2(100 BYTE),

OPEN NUMBER(20,2),

HIGH NUMBER(20,2),

LOW NUMBER(20,2),

CLOSE NUMBER(20,2),

VOLUME NUMBER(20,2),

);

**S&P Index Fact:**

CREATE TABLE S&P\_Index\_Fact

(

Full\_date date,

OPEN NUMBER(20,2),

HIGH NUMBER(20,2),

LOW NUMBER(20,2),

CLOSE NUMBER(20,2),

VOLUME NUMBER(20,2),

);

**Gold Fact**

CREATE TABLE GOLD\_FACT

(

Full\_date date,

OPEN NUMBER(20,2),

HIGH NUMBER(20,2),

LOW NUMBER(20,2),

CLOSE NUMBER(20,2),

VOLUME NUMBER(20,2)

)

**Stock Market Fact**

CREATE TABLE STOCK\_MARKET\_FACT

(

SYMBOL varchar2(100 Byte),

PRICE NUMBER(20,2),

PRICE\_EARNING NUMBER(20,2),

DIVIDEND\_YIELD NUMBER(20,2),

EARNING\_SHARE NUMBER(20,2),

year\_high number(20,2),

year\_low number(20,2),

MARKET\_CAP NUMBER(20,2),

EBITDA NUMBER(20,2),

PRICE\_per\_sale NUMBER(20,2),

price\_per\_book number(20,2)

)

ALTER TABLE DAILY\_MONITORING\_FACT ADD (

CONSTRAINT DATE\_FK

FOREIGN KEY (full\_date)

REFERENCES DATE\_DIMENSION(full\_date),

CONSTRAINT SYMB\_FK

FOREIGN KEY (SYMBOL)

REFERENCES COMPANY\_DIMENSION(symbol));

ALTER TABLE GOLD\_FACT ADD (

CONSTRAINT DAT\_FK

FOREIGN KEY (full\_date)

REFERENCES DATE\_DIMENSION(full\_date));

ALTER TABLE STOCK\_MARKET\_FACT ADD (

CONSTRAINT SYM\_FK

FOREIGN KEY (SYMBOL)

REFERENCES COMPANY\_DIMENSION(symbol));

**DAX Formulas:** Change% = ((Daily\_Monitoring\_fact[Close\_Stock] -   
 Daily\_Monitoring\_fact[Open\_Stock])/Daily\_Monitoring\_fact[Open\_Stock])

SP500 To Gold Ratio = DIVIDE(MAX(SP500\_index\_fact[Close]),MAX('gold\_Fact'[Close\_Gold]))

Revenue = ROUNDUP(Stock\_market\_fact[MARKET\_CAP]/Stock\_market\_fact[PRICE\_PER\_SALE],0)

Outstanding Shares = ROUNDUP(Stock\_market\_fact[MARKET\_CAP]/Stock\_market\_fact[PRICE],0)

Dividend\_Payout = ROUND(Stock\_market\_fact[Annual\_Dividend]/  
 Stock\_market\_fact[EARNING\_SHARE],2)

Annual\_Dividend = ROUND(Stock\_market\_fact[DIVIDEND\_YIELD]\* Stock\_market\_fact[PRICE],2)

Market cap group = SWITCH(

                    TRUE(),

                            Stock\_market\_fact[MARKET\_CAP] <= 10000000000, "Low Market Cap",

                            Stock\_market\_fact[MARKET\_CAP] > 10000000000 &&  Stock\_market\_fact[MARKET\_CAP] <= 20000000000 , "Mid Market Cap",

                            Stock\_market\_fact[MARKET\_CAP] > 20000000000, "High Market Cap",

                            BLANK())

**Tasks:**

**- Data Modeling: All Team members**

**- Data Preparation & Creation: Abdelrahman and Mazen**

**- Data Loading in SQL Developer: Ahmed Osama**

**- Ideas of Queries and analysis: Abdelrahman and Mazen  
-Implementing Queries: Ahmed Osama**

**- Visualization --> Mazen and Abdelrahman**

**-Writing Documentation: Mazen and Abdelrahman**

**-Presentation: Mazen Magdy**