**Analysis of the U.S. Stock Market during**

**Health Crises**

**By**

**Kush Patel**

**Katherine Thai**

**Jonathan Perez**

**Ron Nathaniel**

**Justin Molina**

**Dhrumil Rana**

**a) Overview:**

The overall goal of this project is to better understand the US economy through stock trends and events that affect trends. We intend to achieve this goal by studying the S&P 500 index, crude oil price, and gold price per ounce. In this analysis, gold prices, oil prices, and S&P 500 index will be maintained as our representation of the overall stock market. We will look at time periods of the first 6 months specifically during health crises such as the SARS outbreak in 2002, the Swine Flu pandemic in 2009, and the current COVID-19 pandemic. Through this analysis we intended to learn: the effect health has on the US economy by analyzing how the current pandemic compares, economically, to past health crise; compare short versus long investment in crude oil price, gold price per ounce, and the S&P 500 index during health crises; how does the governmental policies tackling health crises affect the stock market. In conclusion, based on the trends of these attributes from the previous health crises, we believe that we can predict the behaviour of the stock market of the current COVID-19 pandemic.

**b) Problem Definition:**

For this analysis, there are three main goals:

1. Determine the similarity of different health crises based off of gold and oil prices and S&P 500 index? How did the introduction of the treatment affect these attributes?

Through this question we want to analyze how similar these events were and how the introduction of a possible vaccine for COVID-19 can potentially affect said attributes, hence affecting the stock market.

1. Will a short or long-term investment be better suited during a health crisis?

Through this question we want to identify the profit or loss probability during a health crisis. Within this question we would also like to explore other lurking variables that affect the conclusive yield.

1. How do the policies of the Presidential administrations (Bush, Obama, and Trump) affect the stock market during their respective health crises?

Through this question we will see how each administration’s actions affected the stock market during each health crisis.

**c) Period of Analysis:**

This analysis lasted from 20201009- 20201203.

**d) Contact Information(vendor of Data sets):**

[kt352@njit.edu](mailto:kt352@njit.edu)

Yahoo Finance: <https://help.yahoo.com/kb/finance-for-web>

**e) Documentation (usually given by Vendor):**

Crude Oil (CL=F): <https://finance.yahoo.com/quote/CL%3DF/history?period1=1036108800&period2=1598745600&interval=1mo&filter=history&frequency=1mo&includeAdjustedClose=true>

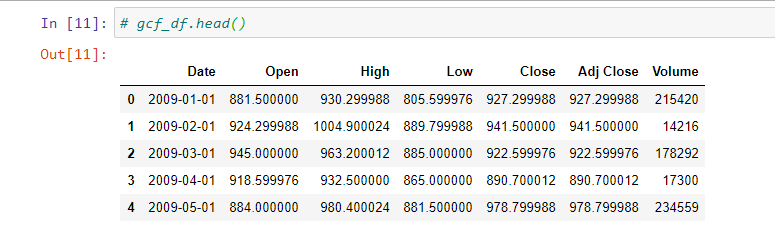
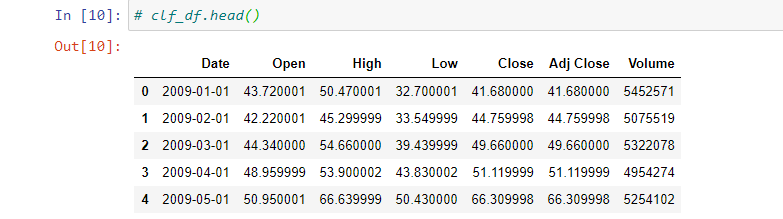
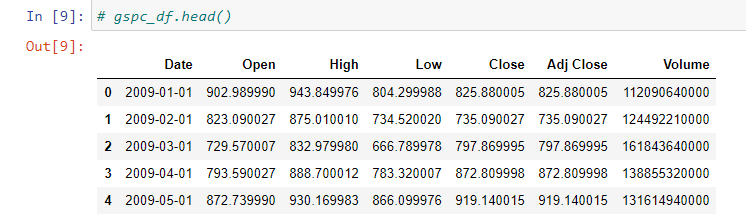
S&P 500 (^GSPC): <https://finance.yahoo.com/quote/%5EGSPC/history?period1=1036108800&period2=1598745600&interval=1mo&filter=history&frequency=1mo&includeAdjustedClose=true>

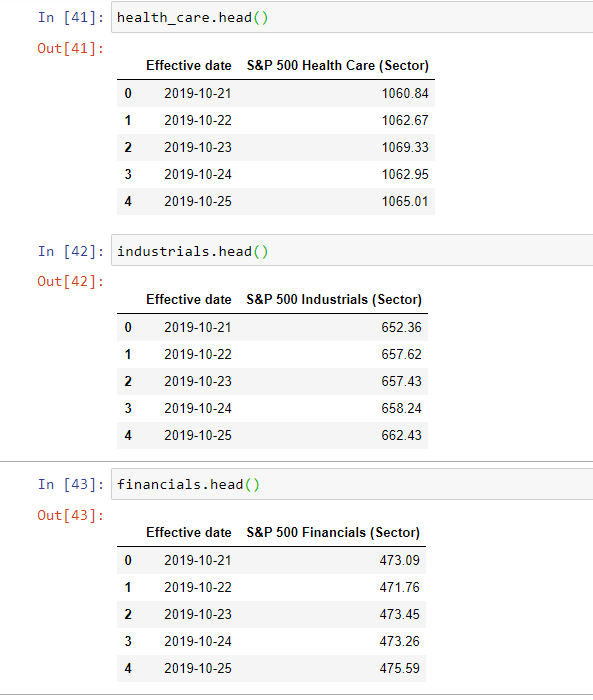
Gold (GC=F): <https://finance.yahoo.com/quote/GC%3DF/history?period1=1036108800&period2=1598745600&interval=1mo&filter=history&frequency=1mo&includeAdjustedClose=true>

**f) Sample data with Links:**

Sample data shared through drive.

<https://tinyurl.com/cs301g5>



****

**g) Code:**

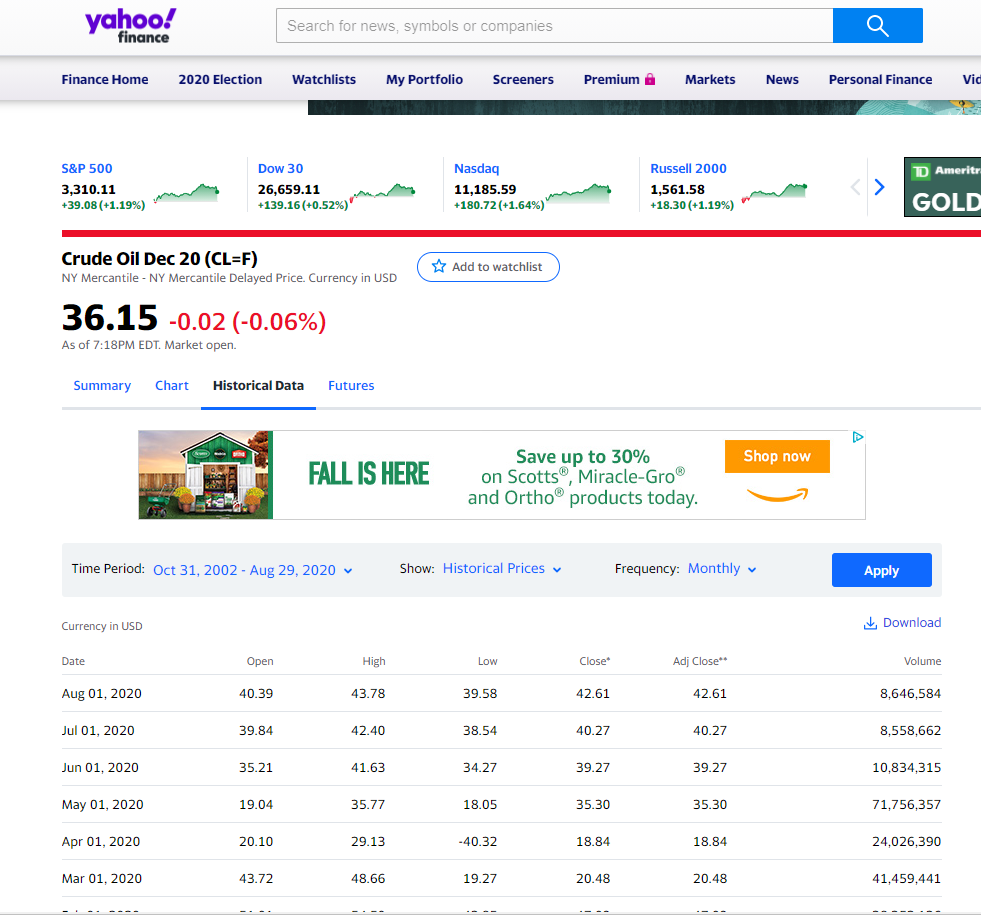
GitHub repository: https://github.com/cs301-star/CS301.git

**Section 2: Data Exploration**

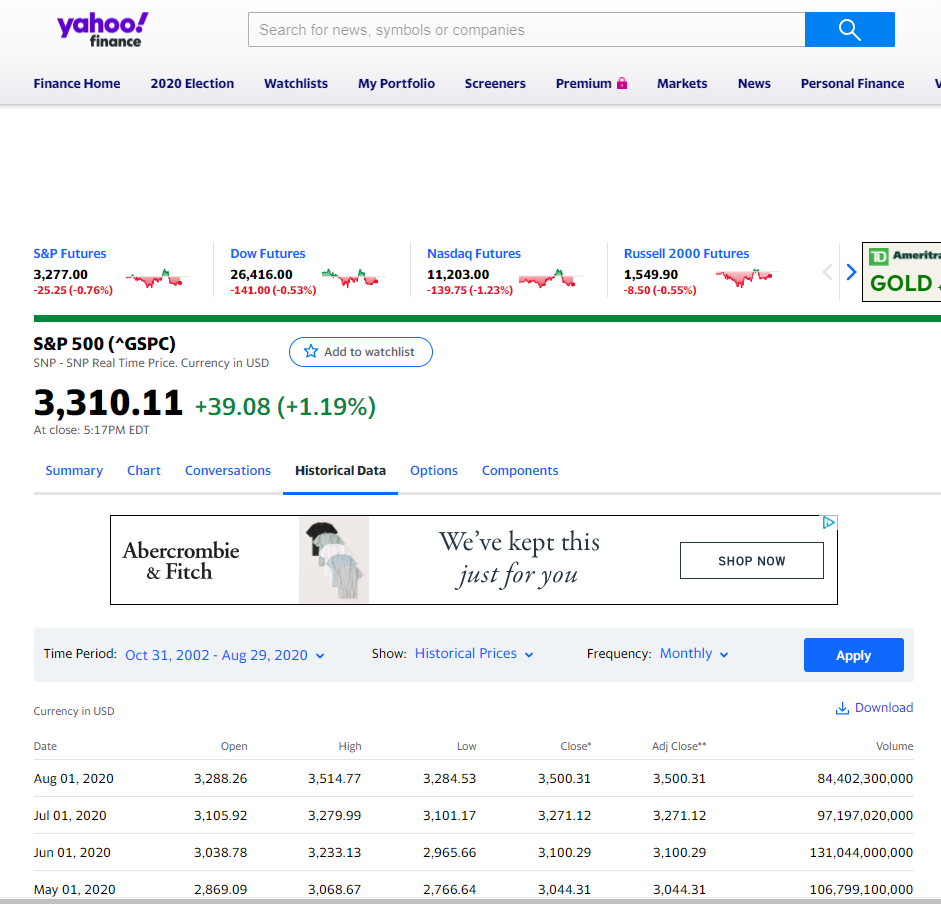
1. **Data Collection:**

All data was downloaded from Yahoo Finance. We had to filter through the historical data of the gold and oil price, and the S&P 500 index. The filter was fixed upon historical data, the time period (October 31, 2020 to August 29, 2020), and frequency (monthly). This allowed us to collect data from the first day of each month through that timespan. Yahoo Finance allowed us to download this data as a CSV file.

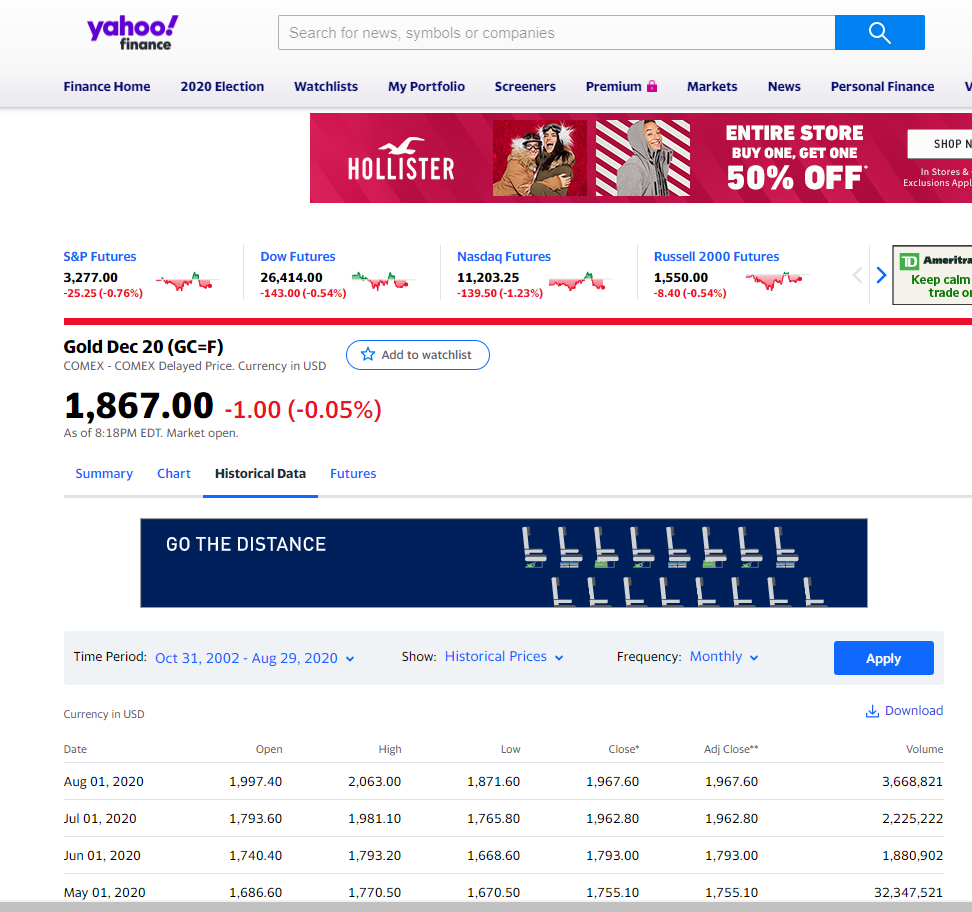
Crude Oil (CL=F):



S&P 500 (^GSPC):



Gold (GC=F):



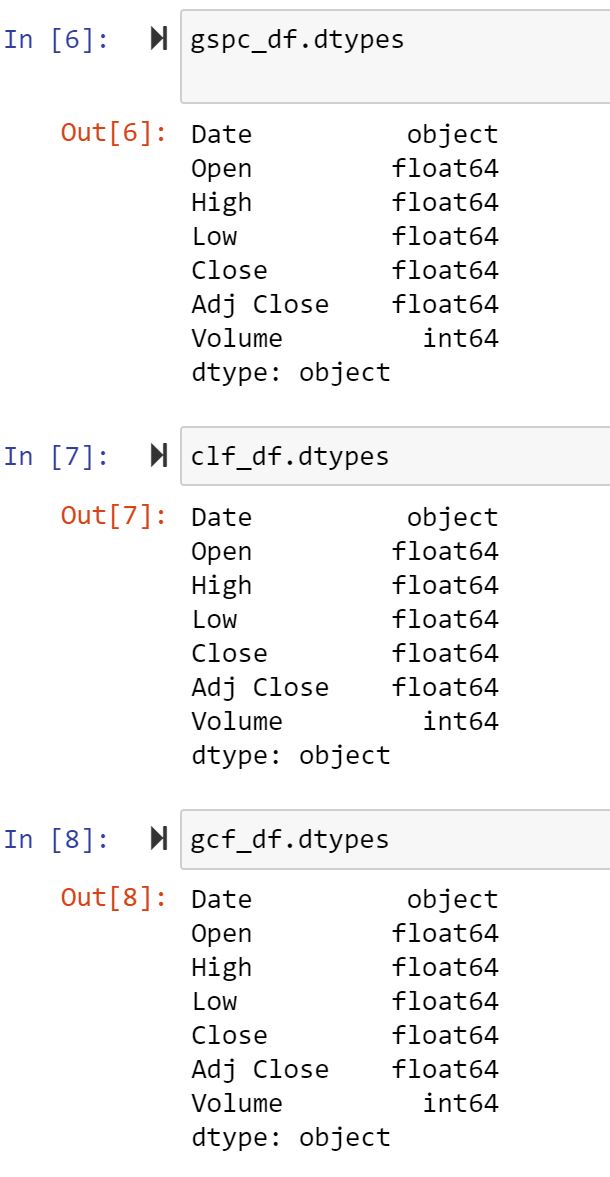
1. **Column Description:**

Note: None of the columns have missing data.

1. **Date:** Ordinal column that represents when the data has been recorded. Measurements are done monthly. It has a timestamp in format yyyy-mm-dd.
2. **Open:** Numerical column that represents the opening price of the stock on the first day of the month. This can range from [0, ∞). It has a format of 6 decimals (0.000000).
3. **High:** Numerical column that represents the stock price at its peak value on the first day of the month. This can range from [0, ∞). It has a format of 6 decimals (0.000000).
4. **Low:** Numerical column that represents the stock price is at its lowest on the first day of the month. This can range from [0, ∞). It has a format of 6 decimals (0.000000).
5. **Close:** Numerical column that represents the closing price of the stock on the first day of the month. This can range from [0, ∞). It has a format of 6 decimals (0.000000).
6. **Adjusted close:** Numerical column that represents the closing price after adjustments for all applicable split and dividend distributions. It is the adjusted closing price at the end of the first day of the month. This can range from [0, ∞).
7. **Volume:** Quantitative column that represents the number/quantity of stocks traded per day. This can range from [0, ∞).

* **What percentage of the data in the column gets populated ?**

All the data from the data frame gets populated and we got the expected data types of each column. We used the isNaN method to check the invalid data, but we got zero rows. We attached the picture of data types of each column.

****

* **Data type of each column:**

Date: Ordinal

Open: quantitative

High: quantitative

Low: quantitative

Close: quantitative

Adj close: quantitative

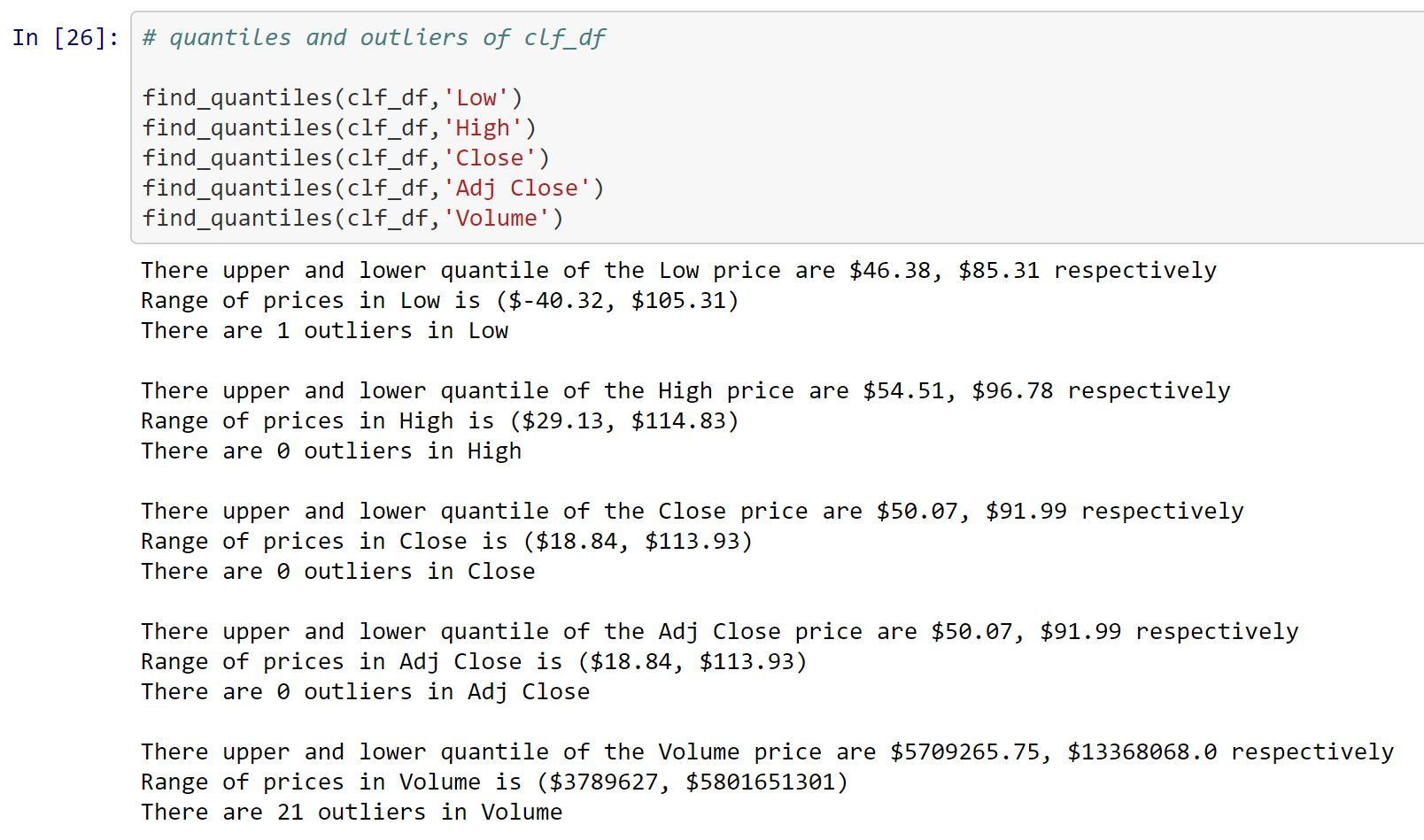
Volume: quantitative

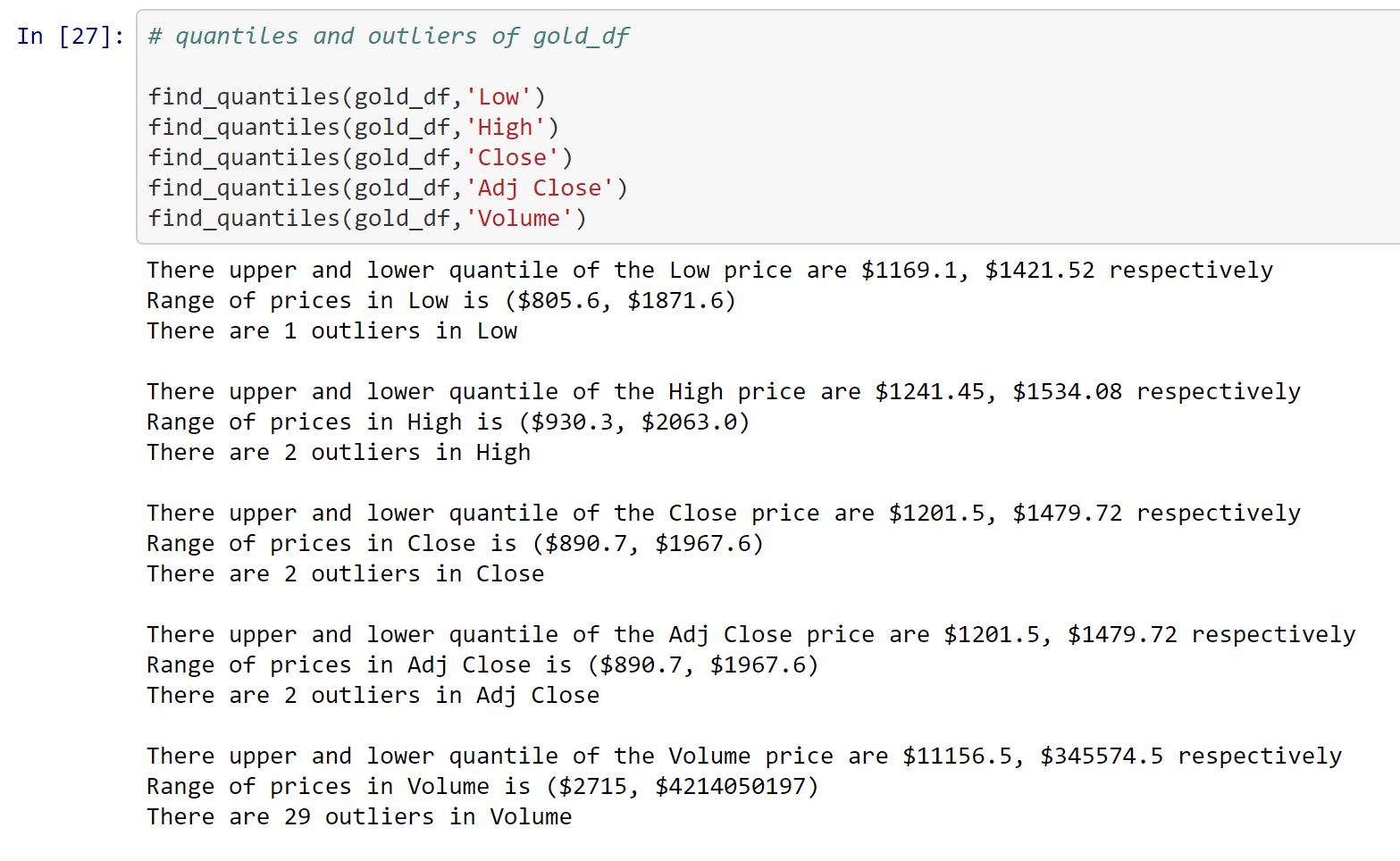
* **Describe the set of values in the column:**

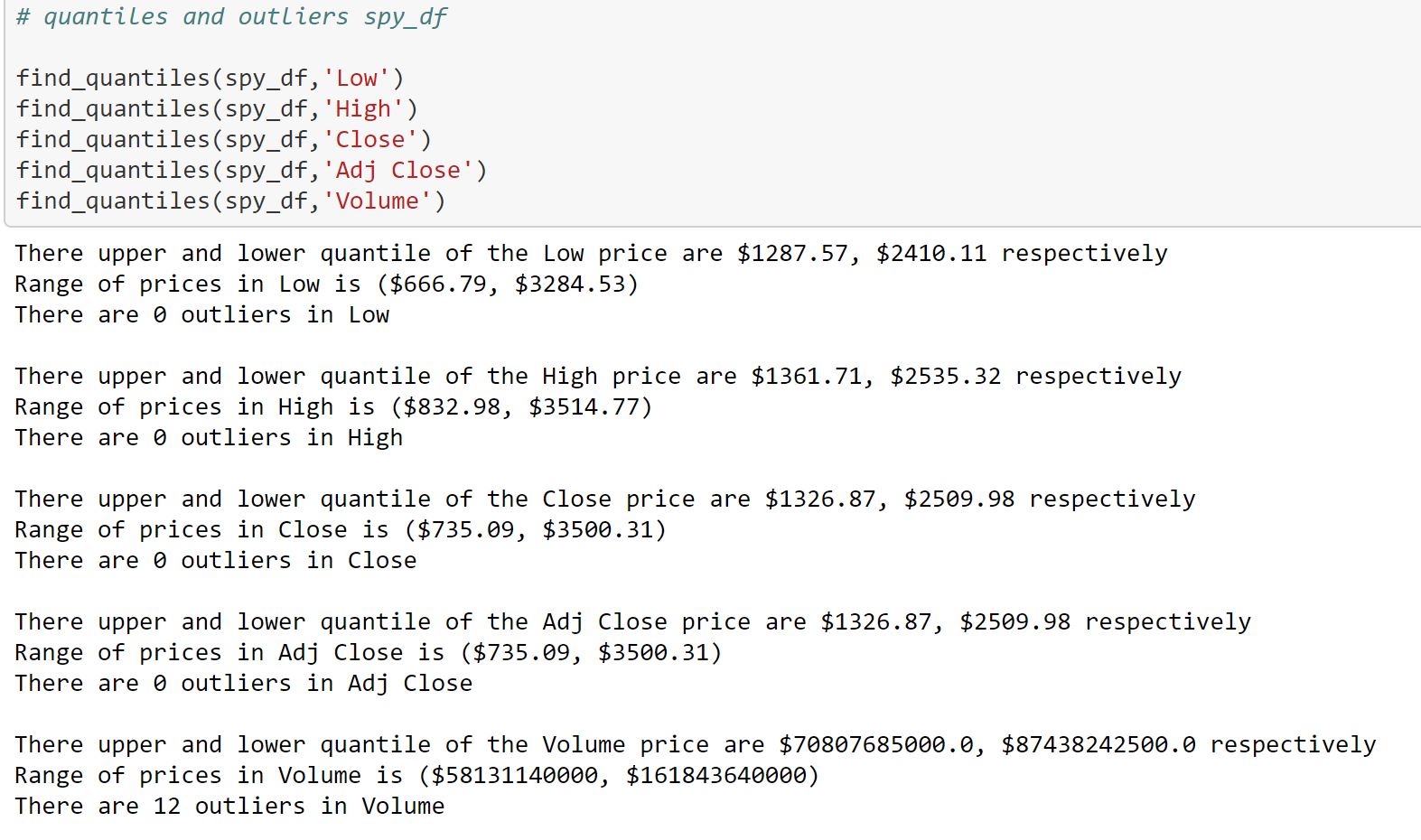
**Date:**

Date range is 2009/01/01 - 2020/08/01

The measurement is monthly

****

****

****

**Github link to notebook file:** https://github.com/cs301-star/CS301/blob/main/Analysis%20of%20the%20U.S.%20Stock%20Market.ipynb

1. **Data Processing:**

* There is a potential incorrect value in the crude oil data frame which shows low column range (-40.32, 105.31). It has to be a positive 40.32 instead of negative value. The incorrect value is the outlier for the low column.
* The gold price is fluctuating a lot, so the range of each column in the gold dataframe is larger compared to the other dataframe. Due to anomaly in the range there are times when the gold value is lower than the quantile and becomes an outlier.
* The volume of the stock depends on the value of the stock when it is at a high price or when the price of the stock is low. The investor will buy the stocks when the price is low; therefore we will see outliers in the volume accordingly.
* The data was downloaded as a csv table from Yahoo Finance. There is fortunately no missing data for Yahoo Finance supplied us with each month’s first day data values. In addition, there are no duplicate entries, for these values are unique to a specific day and a specific stock.