**ADTA 5240 HARVESTING, STORING, AND RETRIEVING DATA**

**FINAL PROJECT OFFICIAL REPORT**

|  |  |
| --- | --- |
| **ADITYA KESHAVAPATNAM** | **11705737** |
| **AESHALA RUDRA TAPASWI** | **11686609** |
| **EZEANYAGU IFEANYICHUKWU** | **11616966** |
| **HARSHA VARDHAN CHAVVAKULA** | **11646088** |
| **ZAKIYA AL JABRI** | **11727401** |

**Topic:** The Evaluation of Stock Performance of the Top 10 Largest US Banks For the Pre-Covid-19, COVID-19, and post-Covid-19 Period from January 2015 to November 2023.

**Description of Dataset:** New York Stock Exchange (NYSE) Ticker names.

**Data Set:** We got the stock information for the following top 10 largest US Banks with the tickers below as listed on NYSE by using the Yahoo Finance Python API.

**The Top 10 Largest US Banks in total assets are:**

1. JPMorgan Chase (JPM)

2. Bank of America (BAC)

3. Wells Fargo (WFC)

4. Citibank (C)

5. US Bancorp/US Bank (USB)

6. PNC Financial Services Group Inc. (PNC)

7. Goldman Sachs Group Inc. (GS)

8. Truist Financial Corp. (TFC)

9. Capital One (COF)

10 Toronto-Dominion Bank, (TD)

Source: US Federal Reserve Statistical Release (As of September 30, 2023)

**Every bank ticker name has 6 columns:**

Open: This is the daily opening stock price

Close: This is the daily closing stock price

High: This is the daily highest stock price

Low: This is the daily lowest stock price

Adj Close: This is the daily stock adjusted closing price aftermarket hours concerning other stock attributes such as dividends, stock splits, and new stock offerings.

Volume: This is the daily stock volume traded on the floor of the NYSE.

For this Project, we will use only the Adj Close price to get the daily returns. The adjusted close price is most often used to estimate the value of the stock because it is considered a more technically accurate reflection or indicator of the stock value.

**Data Collection:** The open-source Yahoo Finance Python API is known as yfinance.

**Data Preparation:** The downloaded data was divided into 3 periods below:

- Pre-Covid19 : January 5, 2015 - January 21, 2020

- Covid19 : January 22, 2020 - August 31, 2022

- Post- Covid19 : September 1, 2022 - November 24, 2023.

**EXPLORATORY DATA ANALYSIS(EDA):**

**Steps Involved:**

* Imported Python Libraries like NumPy, Pandas, Series, and Data Frame.
* Imported Libraries and modules for data visualization.
* Imported Python Libraries Yahoo Finance API, Datetime, and Data Reader to get up-to-date data.
* Downloaded the bank stocks from Yahoo Finance API.
* We grouped the column with column header names to do multi-level indexing based on the header names.
* We checked the last five rows to make sure that it is within the specified period with the column header names.

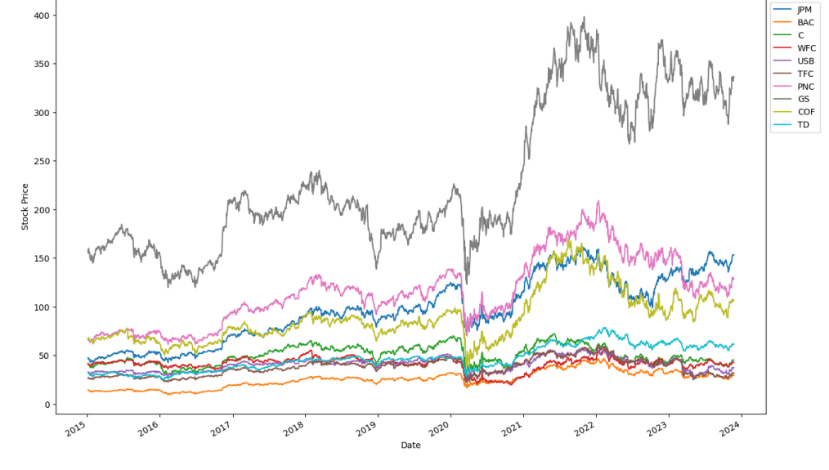
****

Fig. The Adjusted closed stock price for the entire period from Jan 5, 2015 to Nov 24, 2023

**BOXPLOTS:**

* A Boxplot has been visualized to find out if there are any possible outliers
* It is not a good idea to remove outliers in stock analysis because in most cases it indicates huge gains or loss based on market dynamics such as the release of monthly or quarterly economic data, outliers in stock analysis indicates market volatility.
* We can see that three of the banks’ data contain some outliers. But we just kept them.

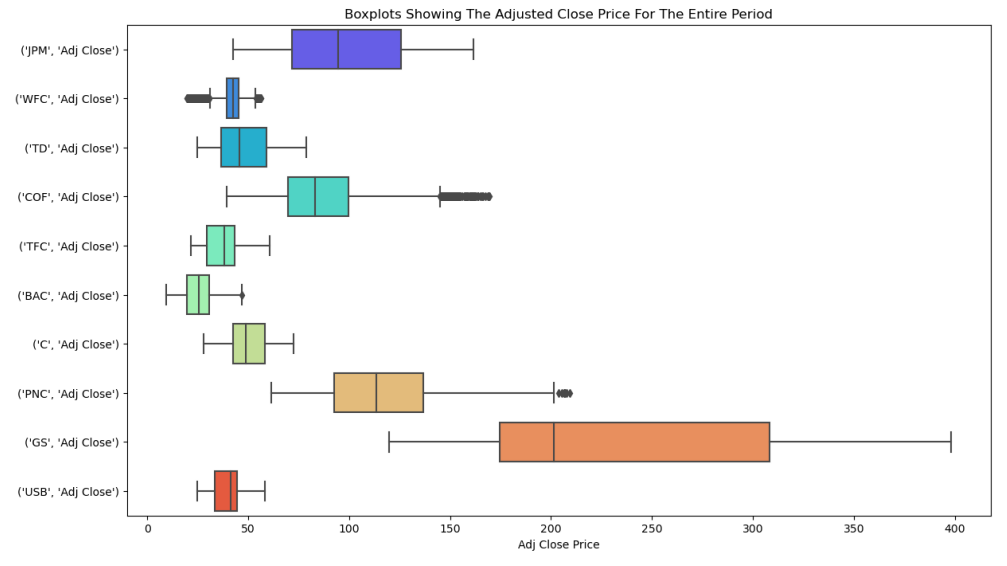


Fig. Boxplots showing the adjusted close price for the entire period

**HISTOGRAMS:**

* We have created histograms for all ten banks so that We can obtain a thorough understanding of the variability and stability demonstrated by these banks by looking at the histograms, which allowed us to see the frequency and size of stock price fluctuations. Most notably, the visualization makes it easier to spot any patterns, anomalies, and similarities among the banks.
* The distinctive features of each bank are highlighted by the histograms, which graphically display the variation in stock price distributions. For investors, financial analysts, and other stakeholders, this study becomes crucial since it makes it easier to comprehend the subtle differences in market behavior that are unique to each institution.
* The visual story reveals as we move through the histograms, showing the rise and fall of stock values over time. In the end, the EDA provides a strong basis for deriving important conclusions and directing financial plans in a knowledgeable and data-driven way.

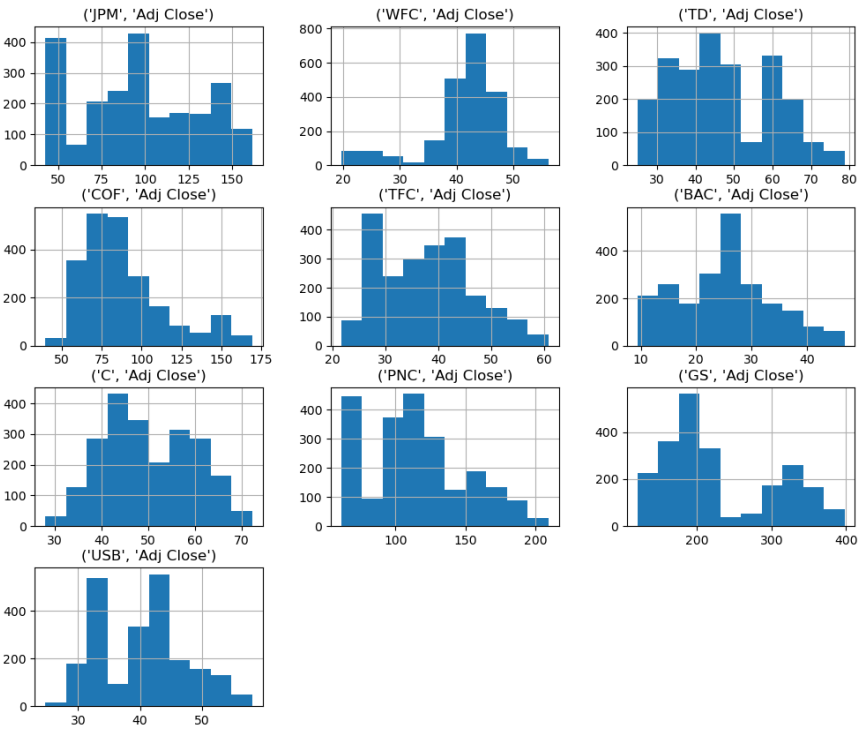


Fig. Histogram of each numeric variable in the data set

**DENSITY PLOTS:**

* We used density plots for each of the top 10 banks' stock prices to conduct Exploratory Data Analysis (EDA) as part of our thorough examination of the industry. With a graphic representation of each bank's market dynamics, these insightful density plots give a comprehensive view of the distribution and concentration of stock values inside each bank. We want to identify trends, and patterns using this method that might affect banks' financial decision-making. Our dedication to providing a data-driven narrative and providing stakeholders with a concise and clear summary of the financial ecosystem for the chosen banks is demonstrated by the inclusion of these density charts in our report.

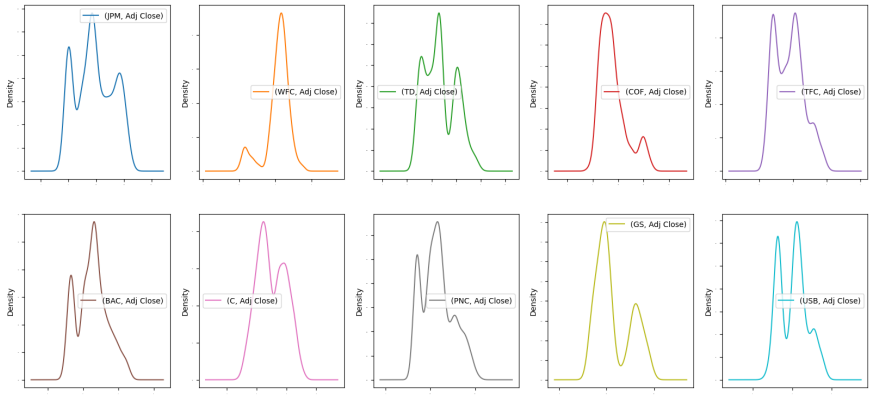


Fig. Density plots of each numeric variable in the data set

**Adjusted closing price for the pre-covid period:**

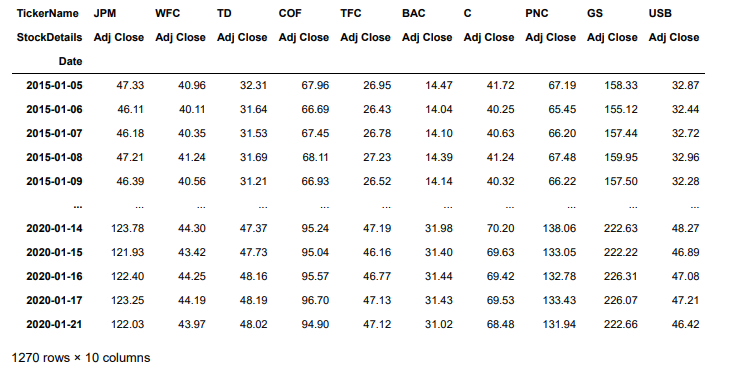
****

Fig. The adjusted closing price for the bank stocks in the pre - covid period from Jan 5, 2015, to Jan 21, 2020

**Adjusted closing price during the Covid period:**

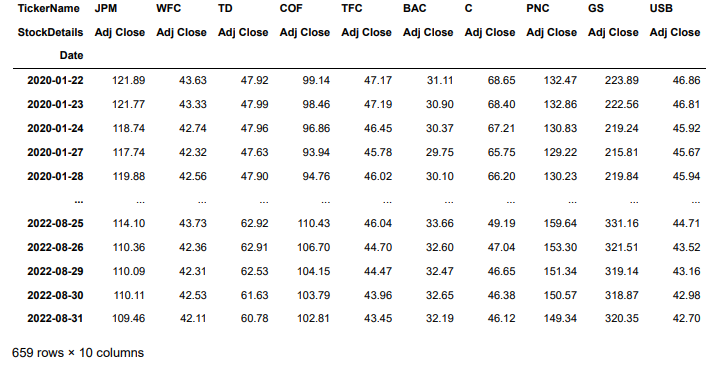
****

Fig. The adjusted closing price for the bank stocks during the pre - covid period from Jan 22, 2020, to Aug 31, 2022

**Adjusted closing price in the post-Covid period:**

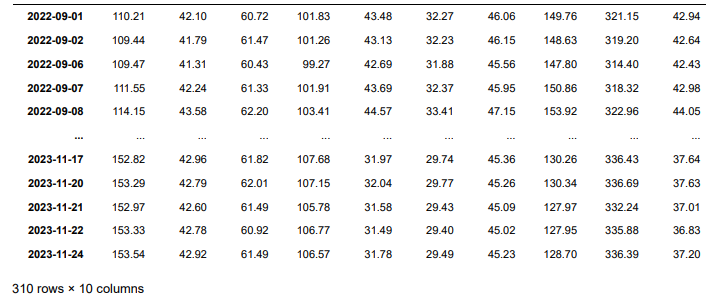
****

Fig. The adjusted closing price for the bank stocks during the pre - covid period from Sep 1, 2020, to Nov 24, 2023

**THE CORRELATION OF ALL THE STOCK RETURNS FOR THE ENTIRE PERIOD:**

* All the stock returns are positively correlated and have very strong to strong correlations. The most correlated stocks are JP Morgan Chase's Return and Bank of America's Return of 0.91 followed by Bank of America's Return and Citi Bank's Return of 0.88.

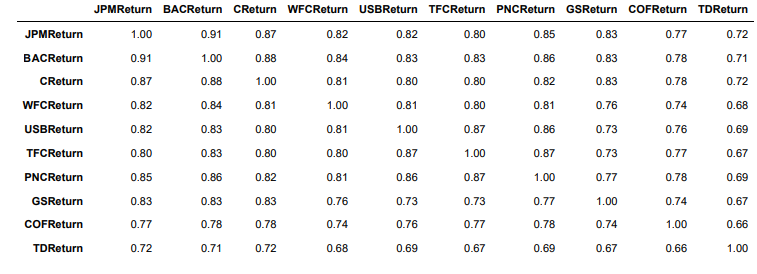
****

Fig. Returns of all the Bank’s stocks for the entire period.

**DATA VISUALIZATION FOR THE CORRELATION OF THE RETURNS:**

* We carried out an insightful Exploratory Data Analysis (EDA) as part of our thorough examination of the top 10 US banks to reveal the complex linkages between their stock returns. One of our report's graphic outstanding works, the correlation heatmap, effectively conveys the dynamic interactions between these heavyweights of the finance world. With each color variation on the heatmap denoting the degree and direction of association, it provides a broad picture of how market forces interact with various financial institutions.

The very strong to strong correlation between these returns could be a result of the equity instruments of these banks tending to move in the same direction altogether and secondly, the very strong correlation between JP Morgan Chase’s Return and Bank of America's Return could be a result of belonging to the same institutional holders known as JP Morgan Investment Management.

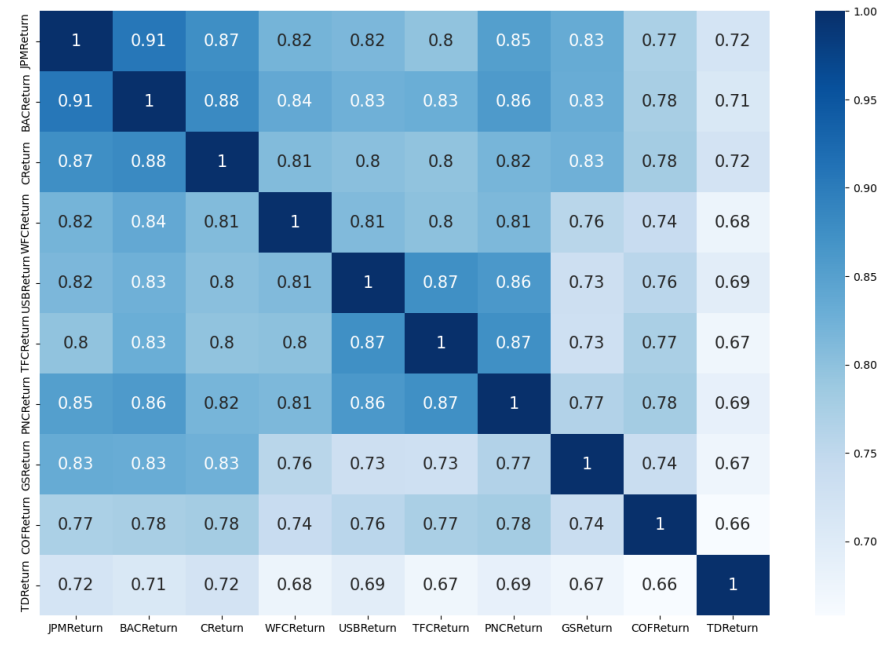


Fig. Heat map that shows the correlation of the stock returns of the top 10 banks

**The Descriptive Statistics and Visualization of The Stock Returns for the Pre-Covid Period:**

* Goldman Sachs had the highest day return before Covid of about 9.54% and the lowest return during this period was Capital One at about -13.12% as shown in the time-series chart below.

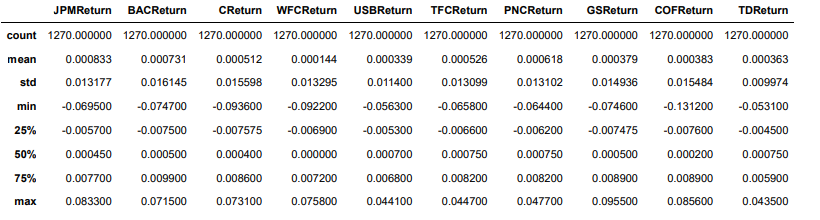
****

Fig. The Descriptive Statistics of The Stock Returns of top 10 banks for Pre-Covid Period

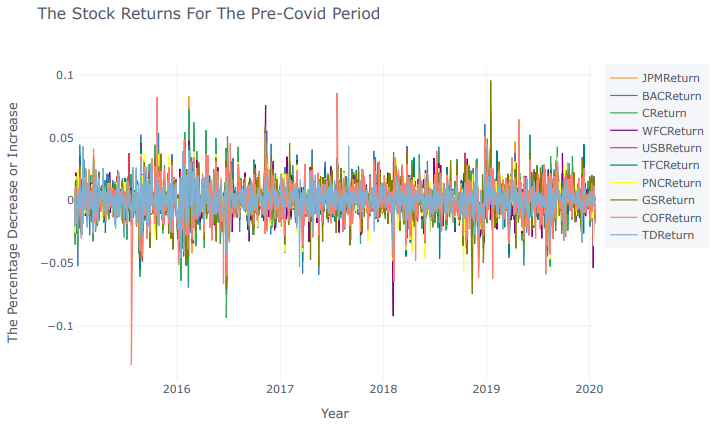


Fig. Visualization of the stocks for the pre covid period

**The Descriptive Statistics and Visualization of The Stock Returns for the Covid Period:**

* Capital One (COF Return) had the highest day return during Covid of about 18.76% and the lowest return during this period was also Capital One at about -23.87% as shown in the time-series chart below. The high standard deviation 0.033 of Capital One also indicates the risky nature of the stock.

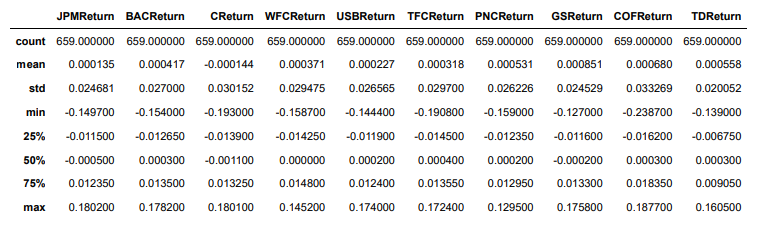
****

Fig. The Descriptive Statistics of The Stock Returns of top 10 banks for the Covid Period

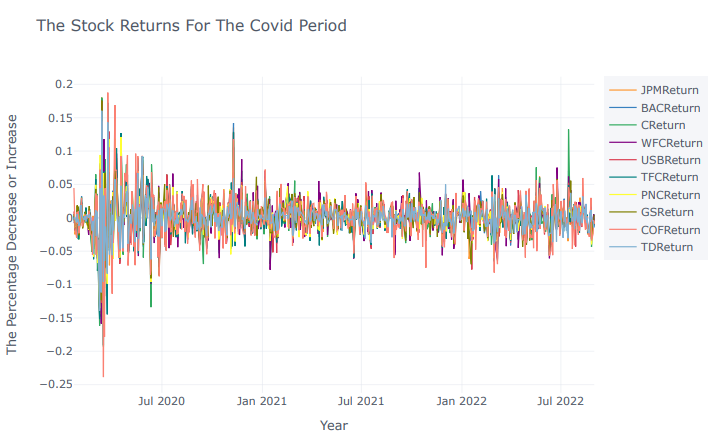


Fig. Visualization of the stocks for the Covid period

**The Descriptive Statistics and Visualization of Stock Returns for the Post-Covid Period:**

* Capital One (COF Return) had the highest day return after Covid of about 11.78% and the lowest return during this period was Truist Financial (TFC Return) at about -16.99% is shown in the time-series chart below. The high standard deviation of 0.026 of TFC Return also indicates high volatility.

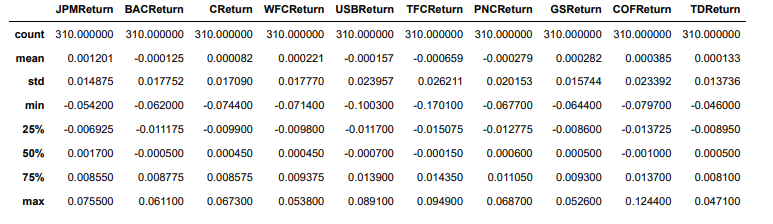
****

Fig. The Descriptive Statistics of The Stock Returns of top 10 banks for the Post-Covid Period

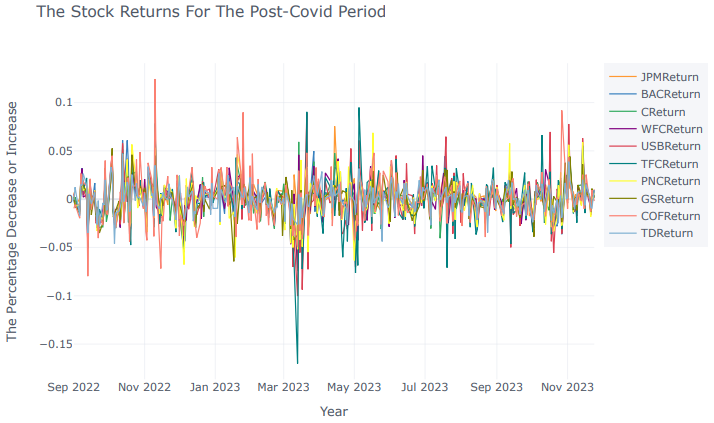


Fig. Visualization of the stocks for the post-Covid period

**VOLATILITY OF ALL THE STOCK RETURNS BEFORE COVID:**

* Volatility of all the stock returns indicates the degree to which the stock return values for a collection of stocks change. In the world of financial markets, volatility is where the price of a financial object, such as a stock, changes over time. While high volatility suggests larger and more frequent price movements, low volatility indicates more consistent and predictable market fluctuations.
* When we talk about the "volatility of all the stock returns," we are evaluating the total variability of the returns across several stocks. Statistical tools like variance and standard deviation might be used to evaluate the total risk attached to the group of stocks and estimate the size of price swings. To effectively manage risk and make well-informed judgments regarding investing strategies, analysts and investors must have a thorough understanding of volatility.
* It shows that the most volatile stock before Covid is Capital One (COF)

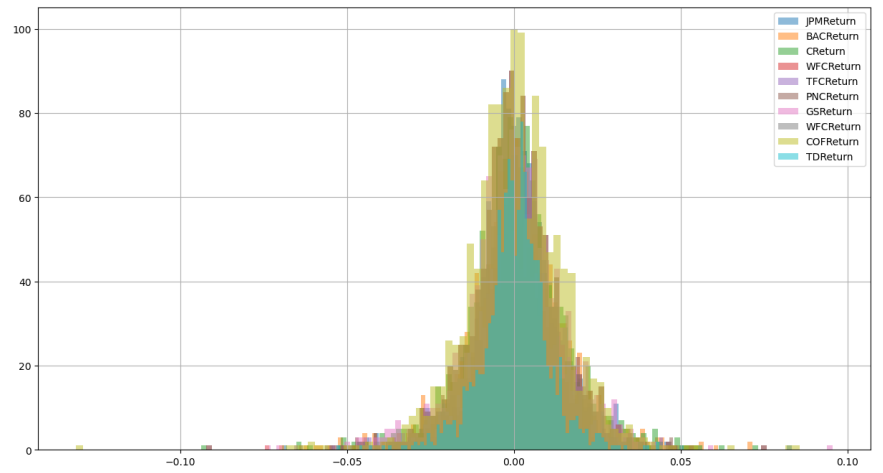


Fig. Volatility of bank stocks before COVID

**VOLATILITY OF ALL THE STOCK RETURNS DURING COVID:**

* It shows that the most volatile stock during Covid is Toronto-Dominion Bank (TD)

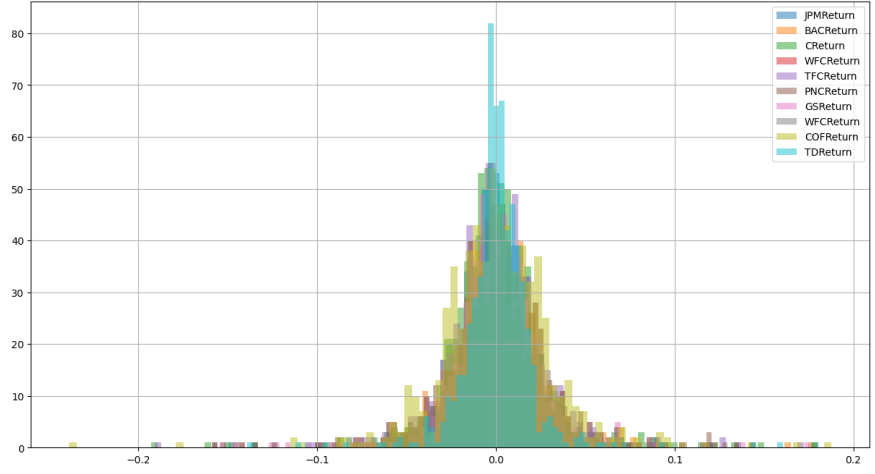
****

Fig. Volatility of bank stocks During COVID

**VOLATILITY OF ALL THE STOCK RETURNS AFTER COVID:**

* It shows that the most volatile stock after Covid is Capital One

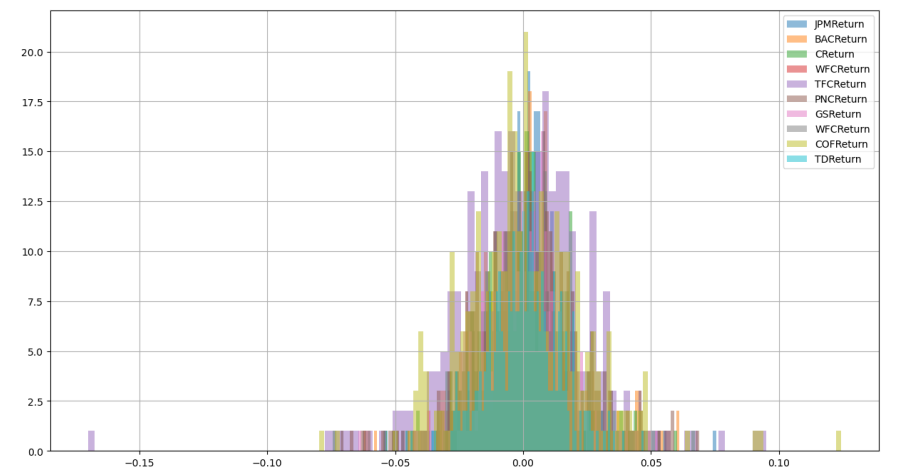
****

Fig. Volatility of bank stocks After COVID

**GOOGLE CLOUD PLATFORM:**

* In this project we have used the Google Cloud platform as a tool to complete our final project.

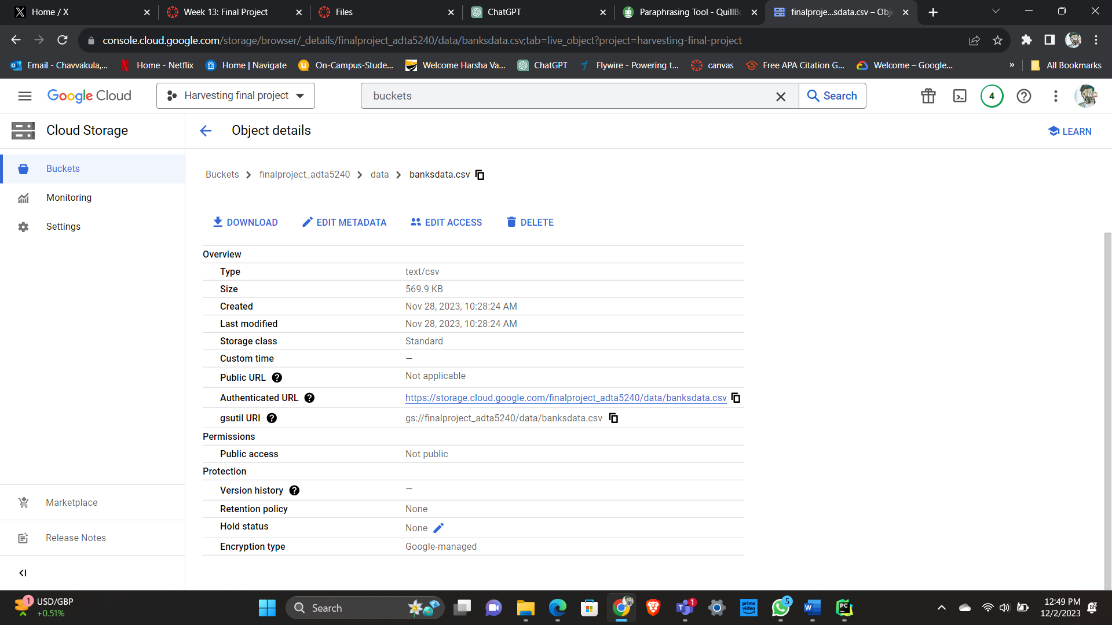
****

Fig. Dataset uploaded in GCP

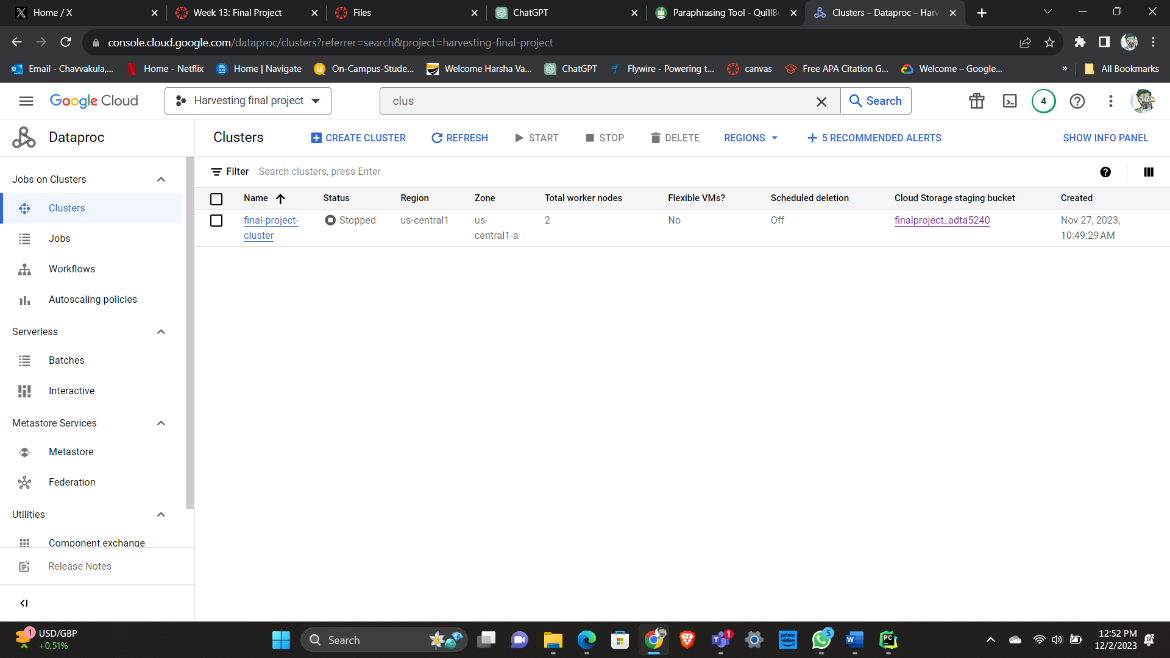


Fig. Created a Cluster with our Dataset

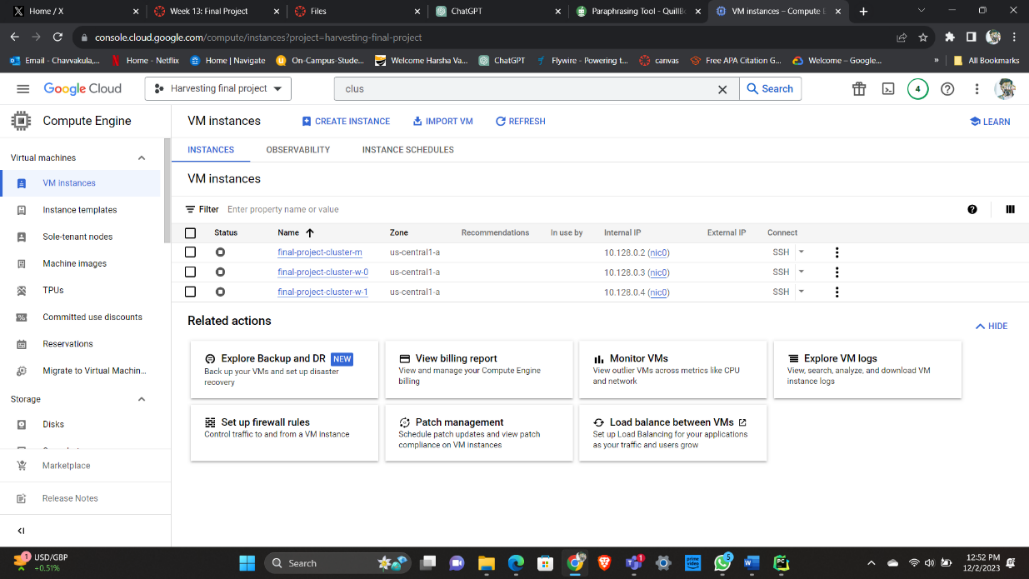
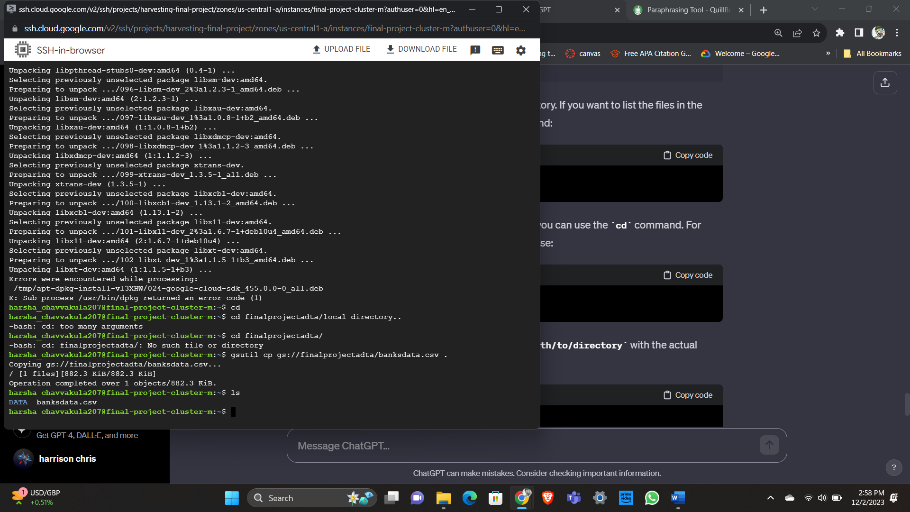


Fig. VM Instances of our Dataset

**LINUX:**

* We have used a Linux command in the SSH terminal of the Manager node to retrieve files from our path where we stored our file in the local directory of the Manager node.
* First, we installed ‘gsutil’ tool which can do various functions like uploading, downloading, setting access…etc.
* We have used the command ‘gsutil cp gs://finalprojectadta/banksdata.csv .’ to copy the dataset file from our GCP bucket to the local directory. Now this command downloads the dataset file from the Bucket and places it in our current local directory.
* After successfully retrieving it, we have run the command ‘ls’ to confirm that the file has been successfully downloaded.
* Now we have successfully retrieved the dataset file from the Google Cloud storage bucket to our local directory on the manager node.



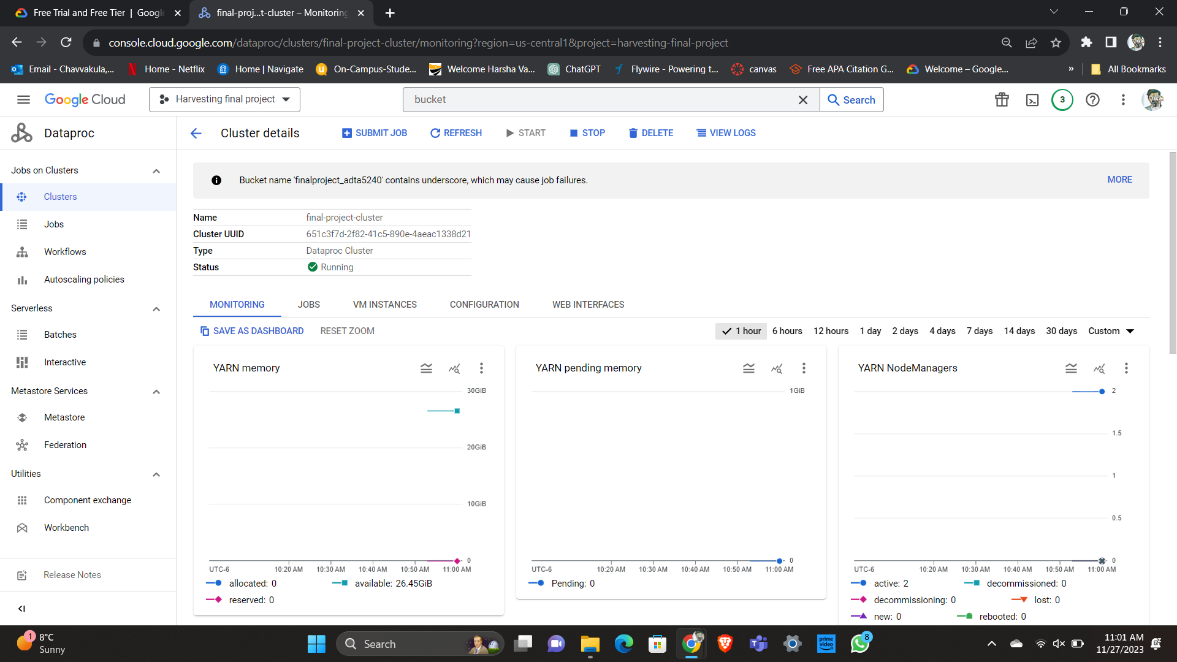


Fig. YARN Memory, YARN pending memory, YARN Capacity

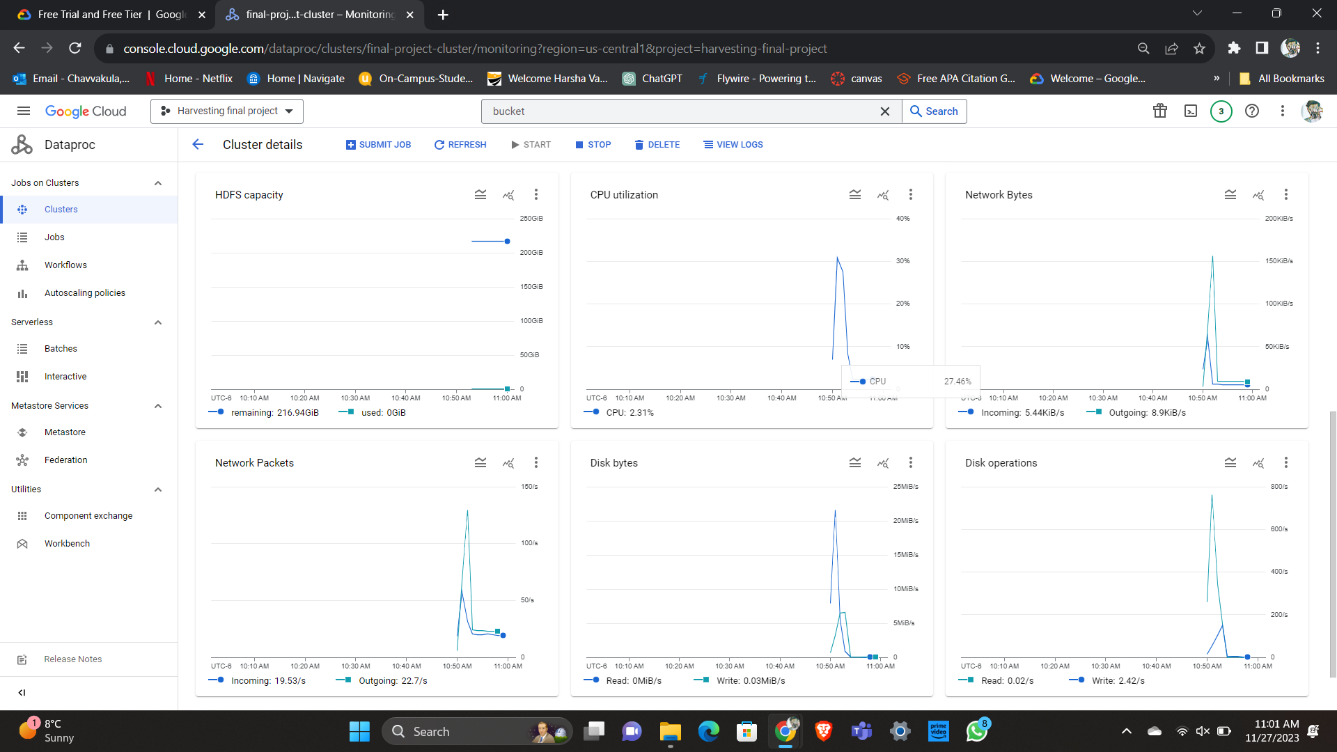


Fig. HDFS Capacity

**BIG QUERY:**

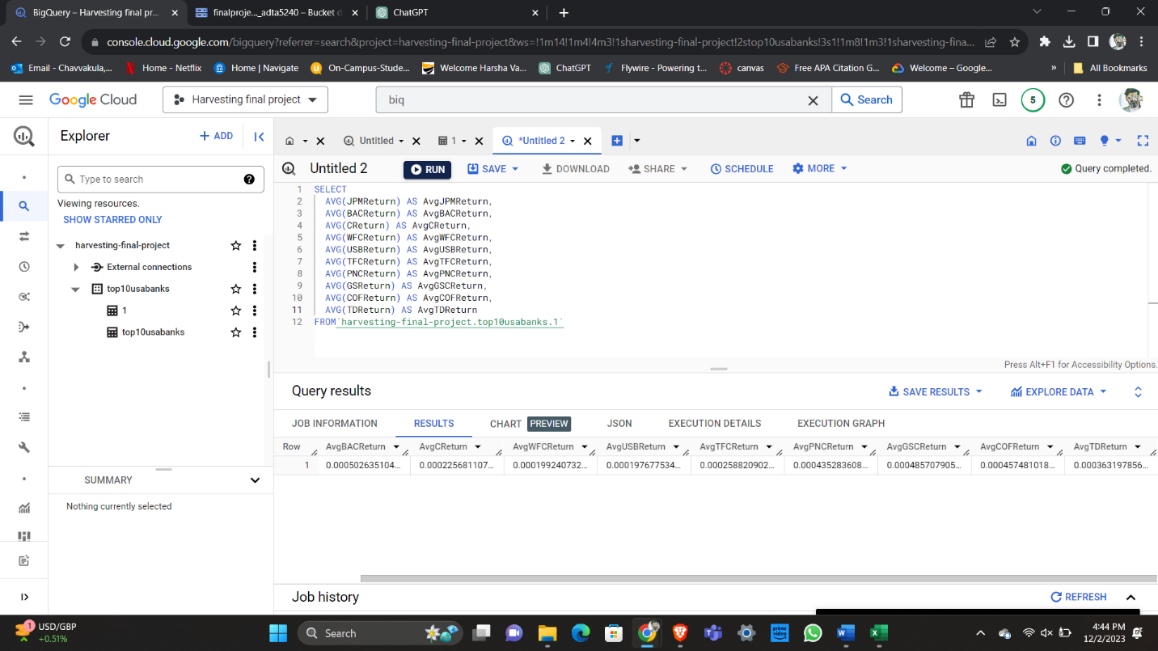
* We have created a storage bucket moved our data set file uploaded it in Big Query.
* We have run a Query in which we found out the Average returns of the Ten Banks. In that, we have come to know that Bank of America has the highest Average Number of Returns which is 0.0005026 and the US Bank has the lowest returns which is 0.000197. These returns were calculated for the whole period which was from 2015 – 2023.
* ****

Fig. Average calculated for each bank Returns

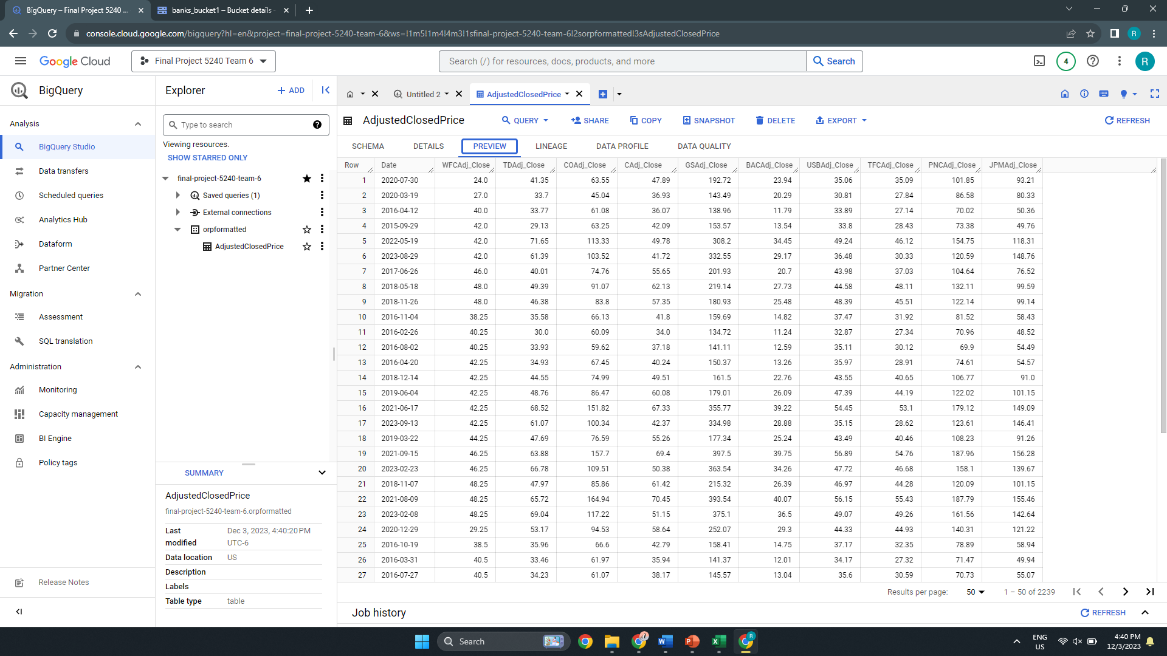
****

Fig. Adjusted stock price dataset loaded in to Big Query

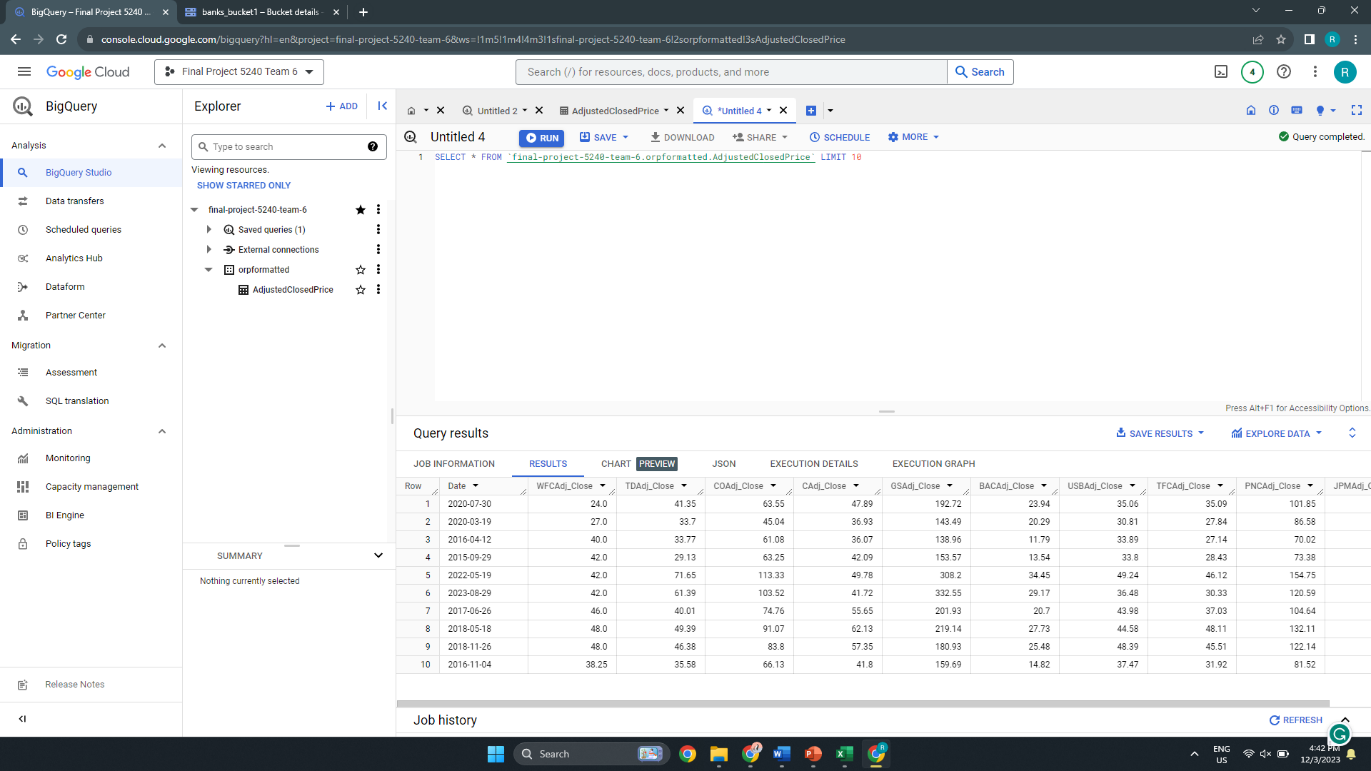


Fig. Retrieving top 10 rows from our dataset

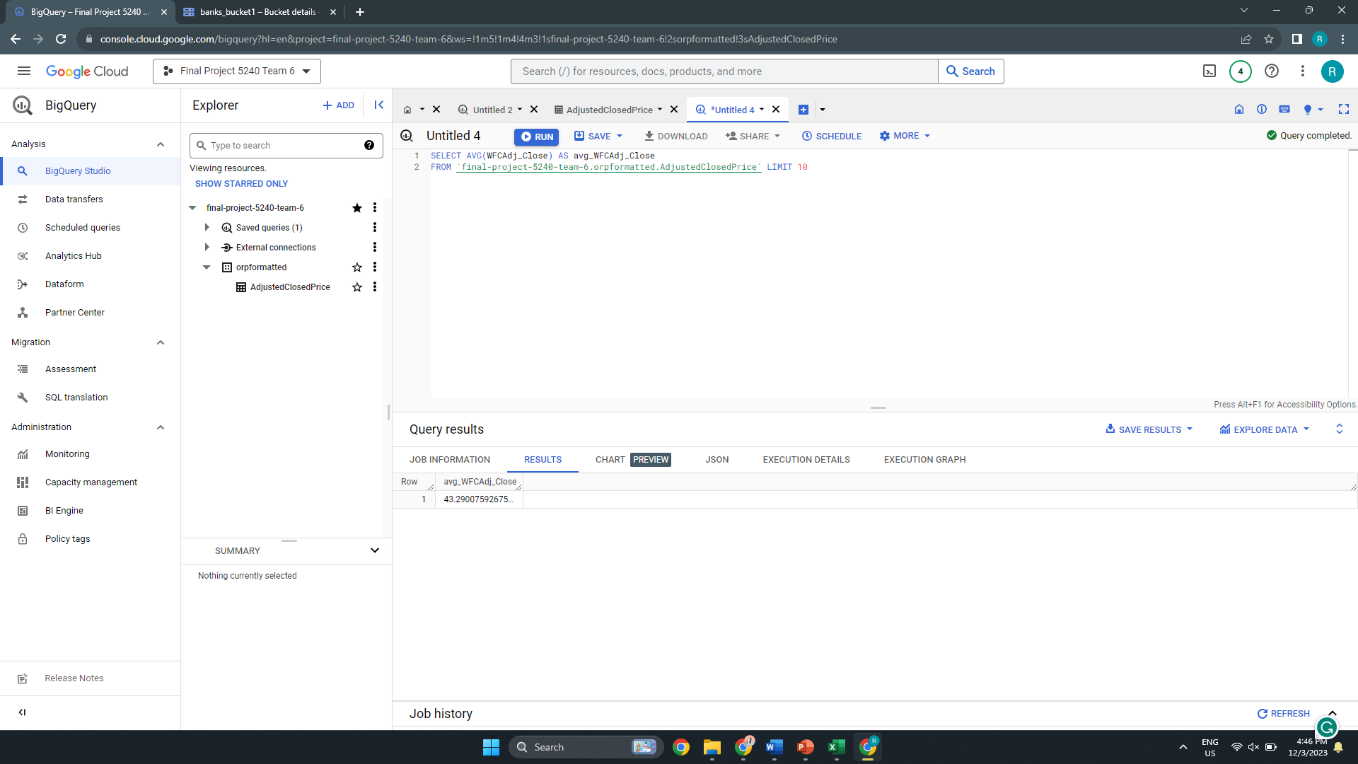


Fig. Average Stock price for Wells Fargo Bank

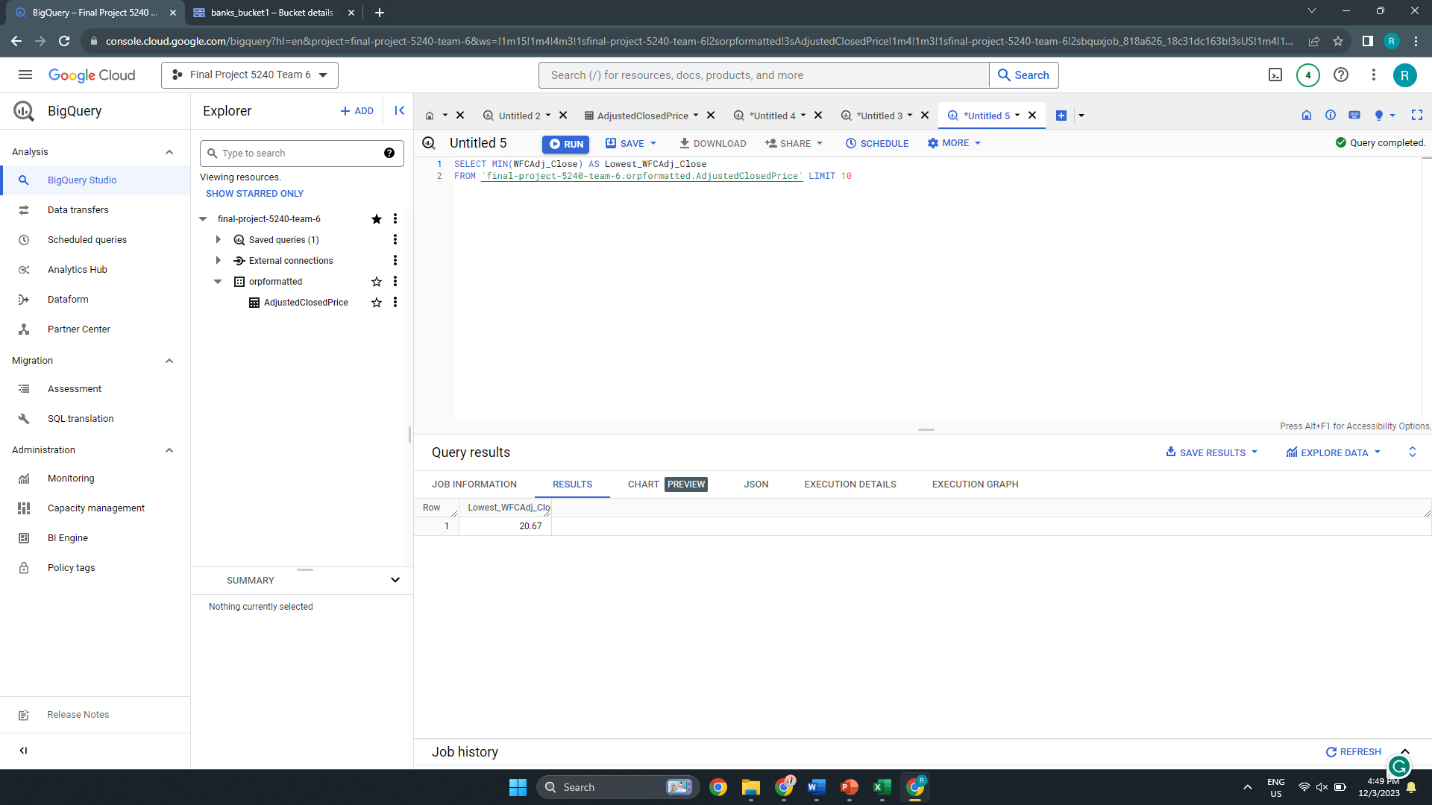


Fig. Retrieving Lowest stock price for Wells Fargo Bank

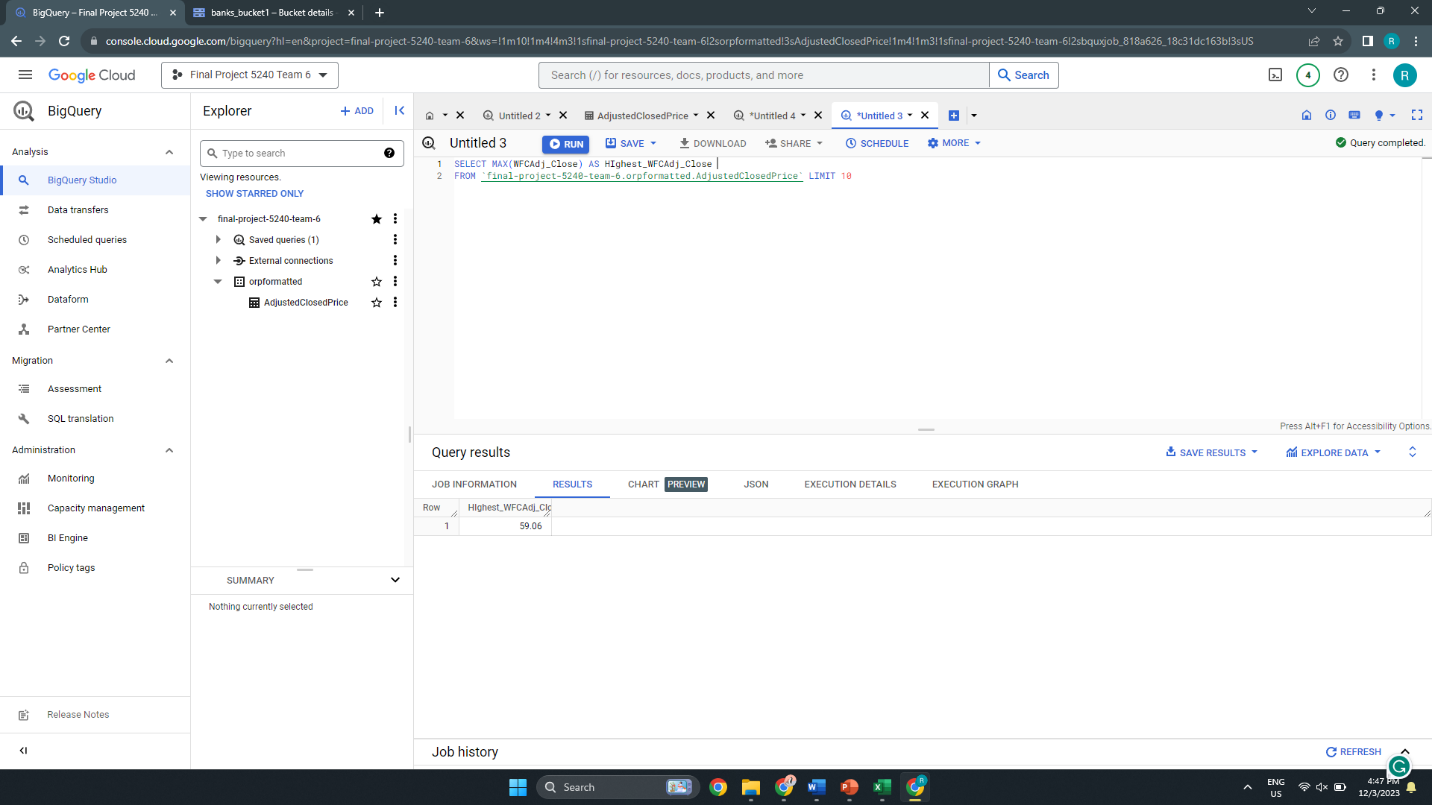


Fig. Retrieving Highest stock price for Wells Fargo Bank

**OPEN REFINE:**

* The utilization of Open Refine, an effective tool for data translation and cleansing, was essential to improving our project's effectiveness. We were able to efficiently eliminate unnecessary columns from our dataset, making it more manageable for analysis thanks to its user-friendly interface. In addition to making data manipulation simpler, this open-source application guaranteed that our reports were accurate and consistent. Open Refine proved to be a priceless resource, improving data quality and enabling perceptive analysis in the process. Its versatility and simplicity in handling massive datasets made it an indispensable tool.

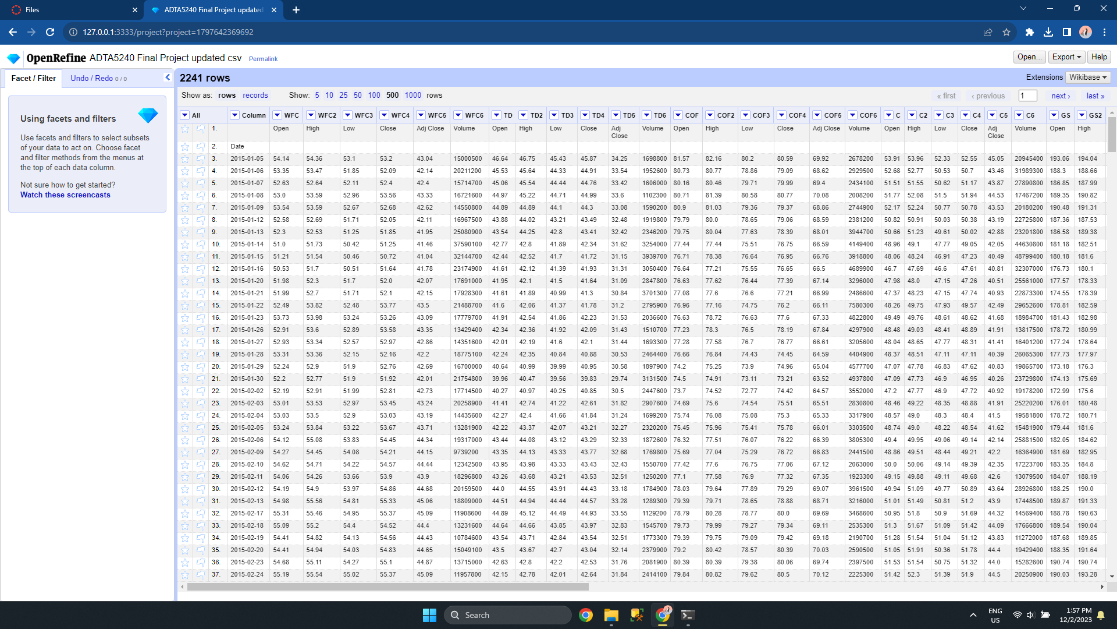


Fig. Dataset before Cleaning

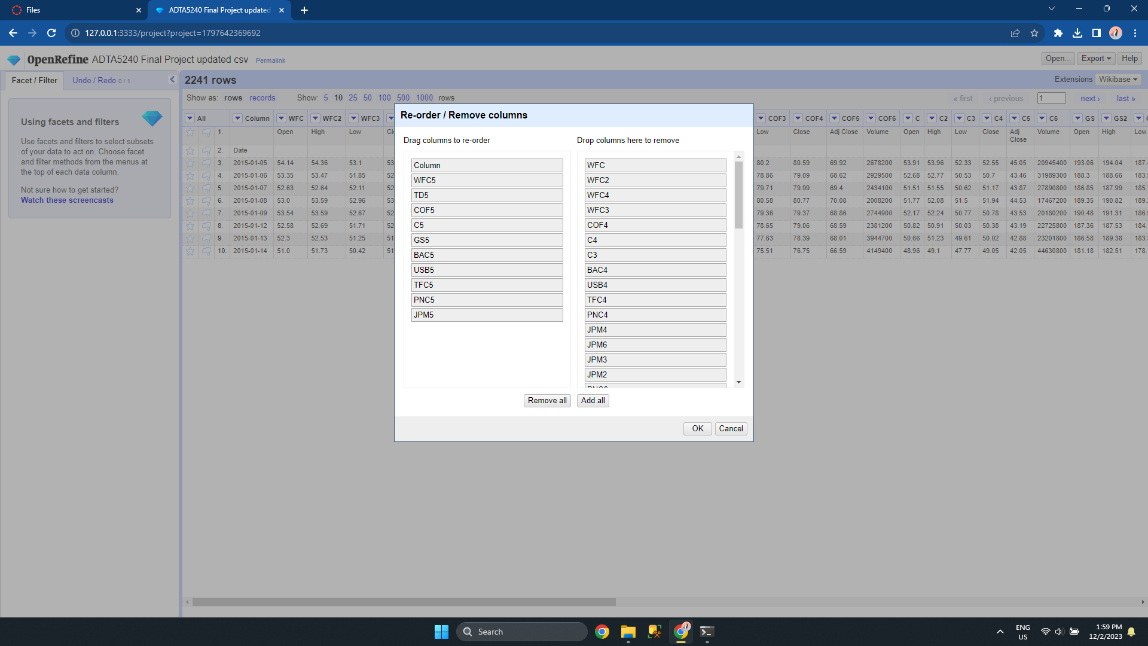


Fig. Selecting the Columns that needed to be dropped

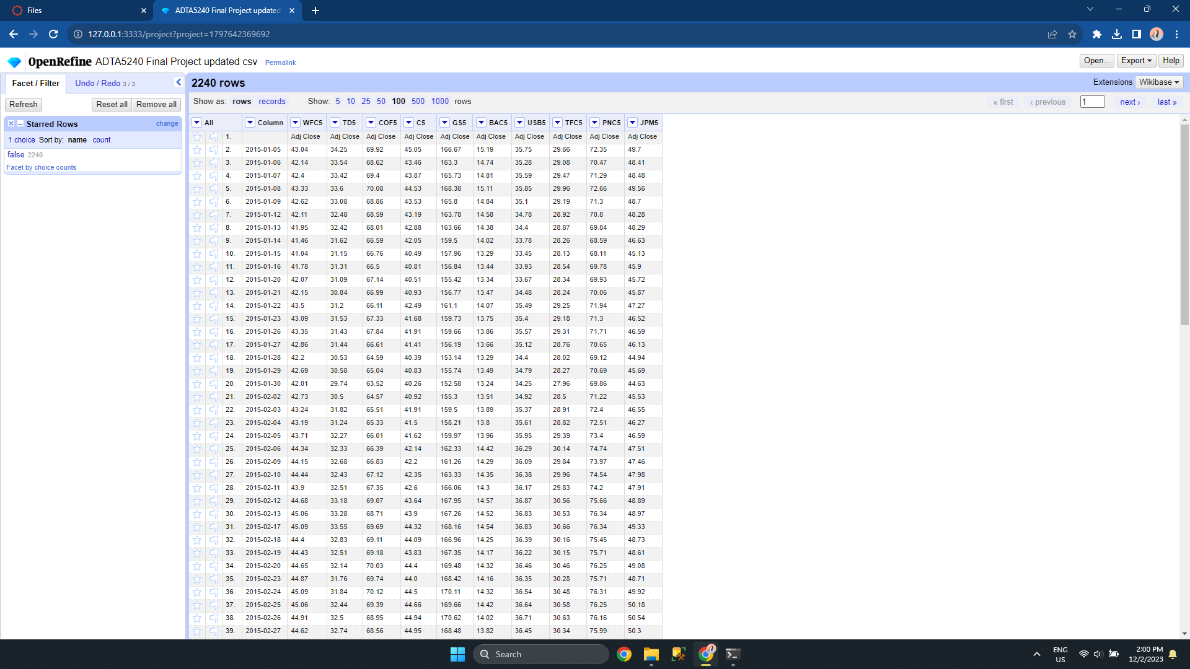


Fig. Dataset after cleaning

**HIVE AND SPARK SQL:**

Spark SQL and Hive are strong technologies for large data analytics. Large datasets may be queried and data warehoused using a SQL-like interface thanks to Hive, is developed on top of Hadoop. It allows complex queries and manages schema via a Metastore. Conversely, Spark SQL, a component of the Apache Spark ecosystem, provides a unified data processing platform by facilitating the easy integration of SQL queries with Spark applications. With the help of Spark SQL's expanded functionality and support for Hive's SQL syntax, we may effectively handle data by utilizing Spark's distributed computing capabilities. Both technologies are extensively utilized for high-performance and scalable data processing in big data contexts. In this project, we used these tools to make new directories on HDFS, make a folder on the Manager Node, move files from the GCP storage bucket to the Manager Node, and finally onto the HDFS directories and also create tables and retrieve data from our dataset using some queries.

These are some of the command lines we used,

hdfs dfs -ls /  
hdfs dfs -ls /user  
hdfs dfs -mkdir /user/ harsha\_chavvakula207

hdfs dfs -mkdir /user/ harsha\_chavvakula207/data

hdfs dfs -mkdir /user/ harsha\_chavvakula207/data/adjustednew

gsutil cp gs://finalprojectharsha/data/adjustednew.csv adjustednew.csv

hdfs dfs -put adjusted.csv /user/harsha\_chavvakula207/data/adjustednew

hdfs dfs -ls /user/harsha\_chavvakula207 /data/adjustednew

* The below command was used to create an external table:

CREATE EXTERNAL TABLE IF NOT EXISTS adjustednew\_1  
(adjusted string)  
ROW FORMAT DELIMITED  
LOCATION '/user/harsha\_chavvakula207/data/adjustednew/';

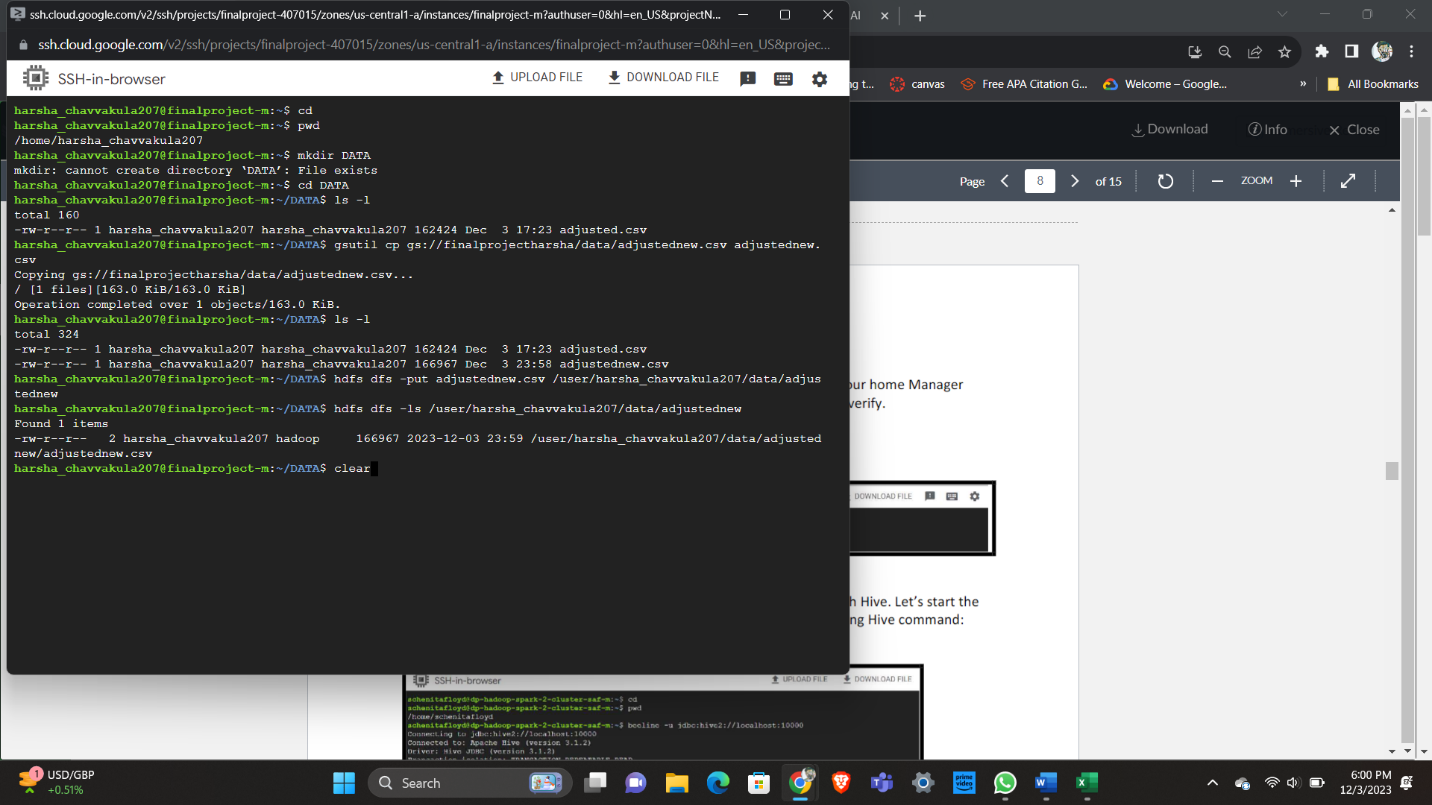


Fig. Moving dataset file from Manager node to Hadoop Ecosystem

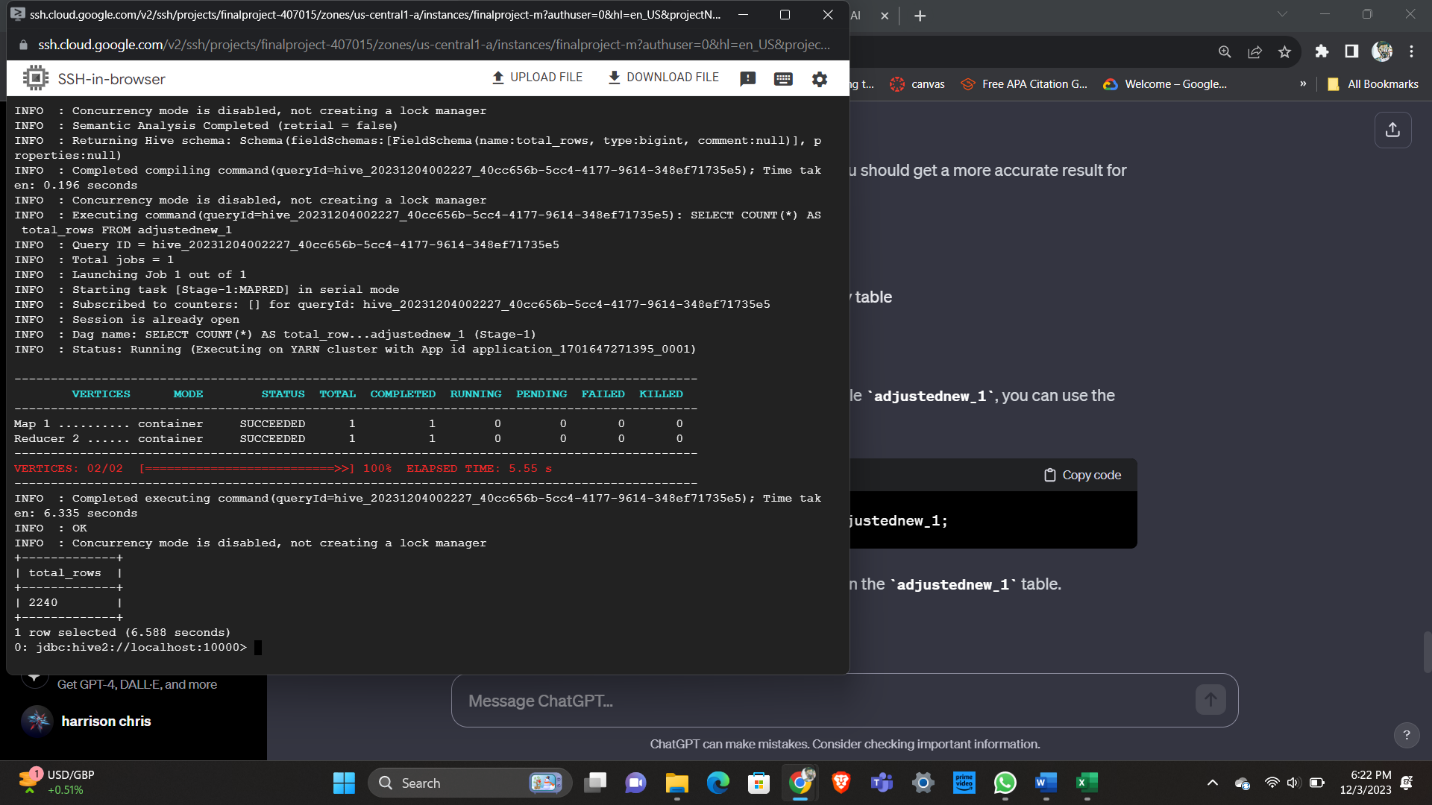


Fig. Query used to retrieve the number of rows in the Hive

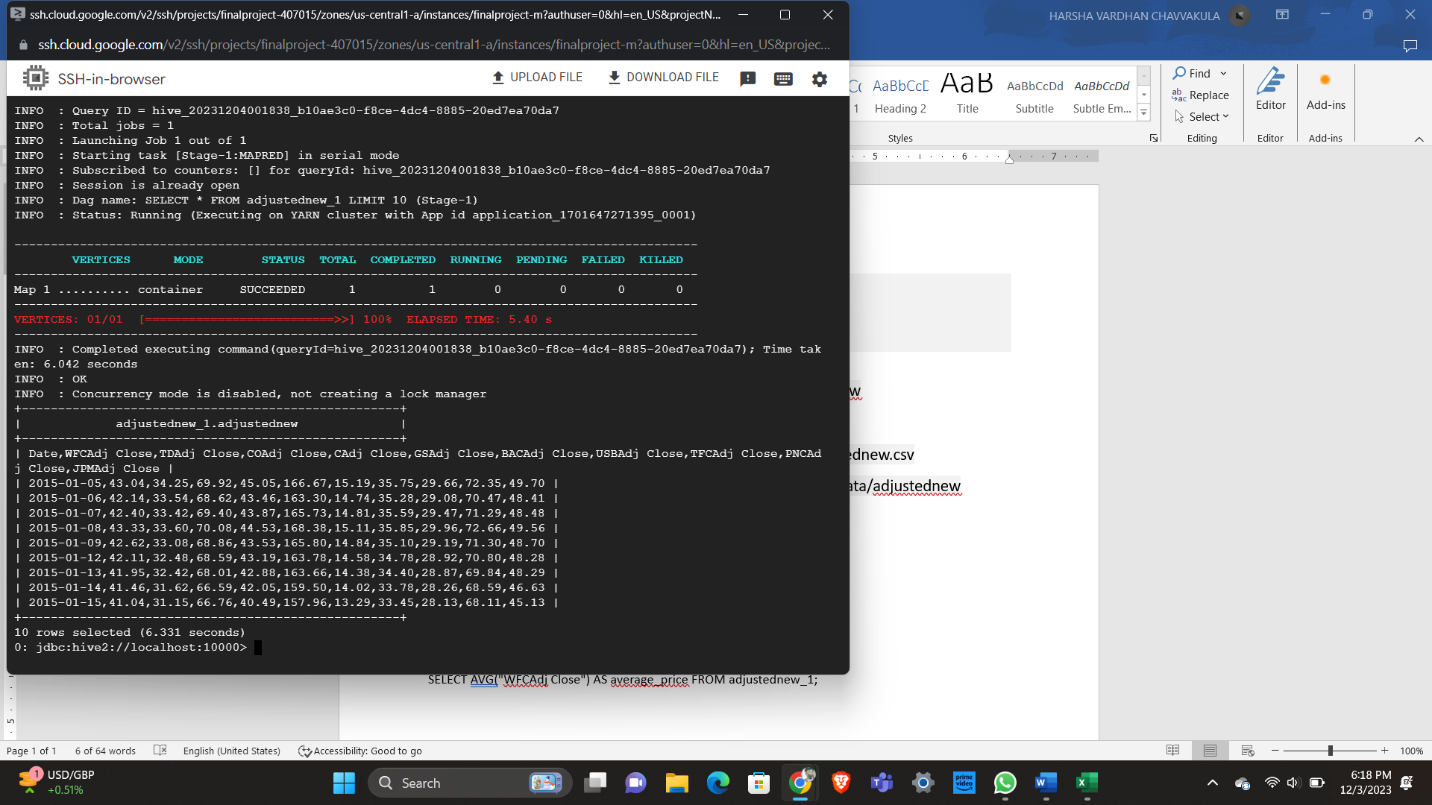


Fig. Query used to retrieve the first 10 rows from our dataset in Hive

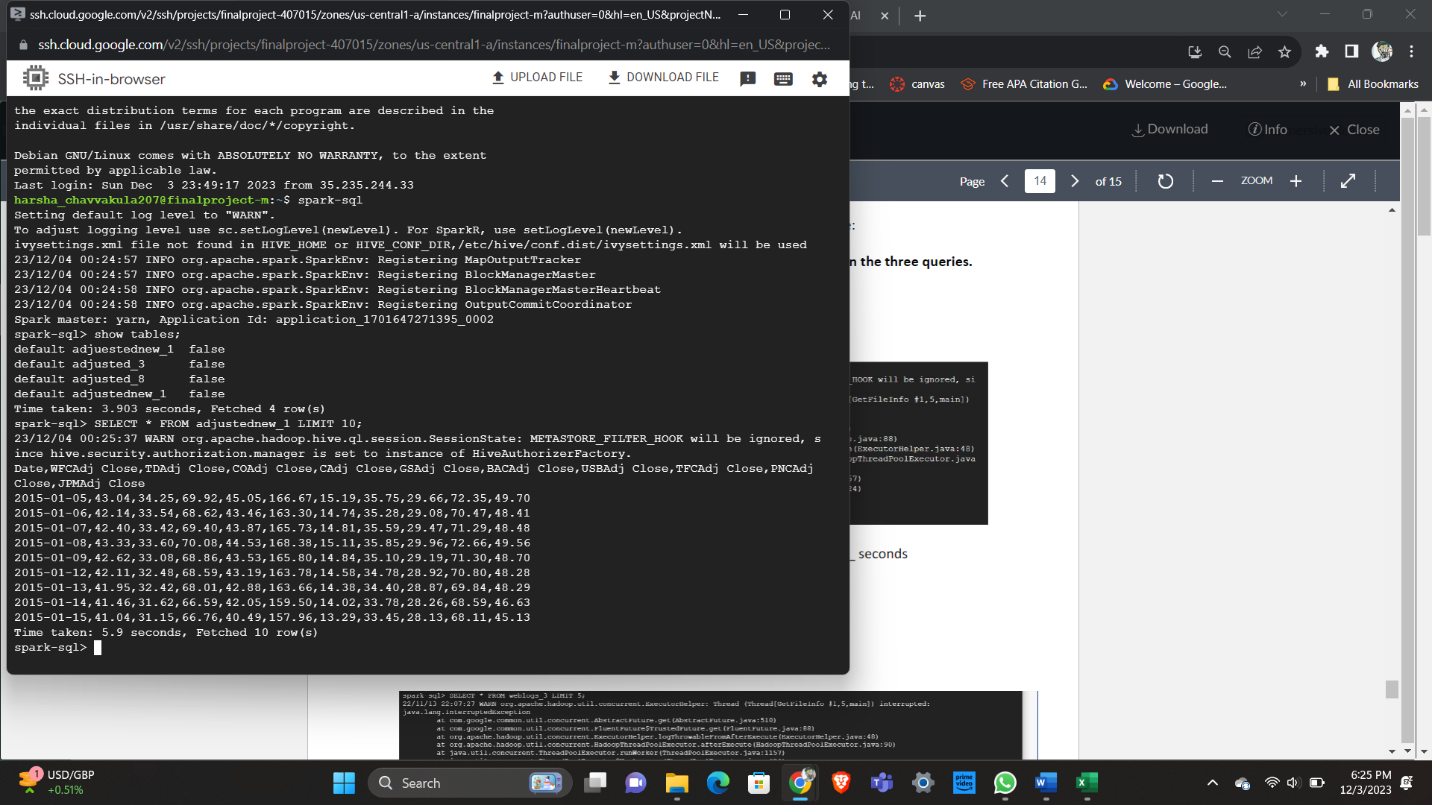


Fig. Query was used to retrieve the first 10 rows from our dataset using Spark SQL

* Here if we compare the time taken to complete the query, we can observe a difference between the two tools Spark and Hive. To retrieve 10 Rows from our dataset Hive took 5.99 sec as Spark took only 6.3 seconds.