**Title: New York Stock Exchange Dataset Analysis**

## OVERVIEW

The New York Stock Exchange (NYSE) is a stock exchange located in New York City that is the largest equities-based exchange in the world, based on the total market capitalization of its listed securities. Formerly run as a private organization, the NYSE became a public entity on March 8, 2006. The New York Stock Exchange is the biggest marketplace for investors to buy and sell shares of stock in the world. Located on Wall Street in downtown Manhattan in New York City and owned by Intercontinental Exchange, the NYSE has a history that goes back more than 200 years, and most of the largest, best-known, and most prestigious businesses in the world choose to list their shares on the stock exchange.

### OBJECTIVES

1. How many categories of ticker symbols are there? (A ticker symbol or stock symbol is an abbreviation used to uniquely identify publicly traded shares of a particular stock on a particular stock market. A stock symbol may consist of letters, numbers or a combination of both. E.g., ABDE is ticker symbol of Adobe.)
2. What were the earnings of the company before paying the tax?
3. What is the effect of exchange rate on the companies?
4. Do any companies have some net borrowings over the years?
5. How many companies have invested in long term investments?

### COLUMNS AND DATA TYPE

|  |  |  |
| --- | --- | --- |
| * Sno *INT* * Ticker *STRING* * AccountsPayable *BIGINT* * AccountsReceivable *BIGINT* * AfterTaxROE *BIGINT* * EarningsBeforeTax *BIGINT* | * LongTermInvestments *BIGINT* * NetBorrowings *BIGINT* * NetIncome *BIGINT* * ProfitMargin *BIGINT* * TotalRevenue *BIGINT* * ForYear *BIGINT* | |
| * EffectofExchangeRate *BIGINT* * GrossMargin *BIGINT* * GrossProfit *BIGINT* * IncomeTax *BIGINT* * Investments *BIGINT* | * EarningsPerShare *BIGINT* |

● LongTermDebt *BIGINT*

**PROBLEM STATEMENTS:**

1. The total number of distinct ticker symbols in the dataset.
2. Listing the companies which have zero income tax to be paid.
3. Ticker symbols of the companies which had no profit overall.
4. Companies going into total loss and have debts to be paid.
5. Total number of stocks filings each year had (2012-16).
6. Maximum total revenue over different years.
7. Top 5 companies with highest gross profit.
8. Average net borrowings done by companies over the years.

**HIVE QUERIES**

1. **Using the Database, we created a database named as bde\_project.**

Entering the hive shell.

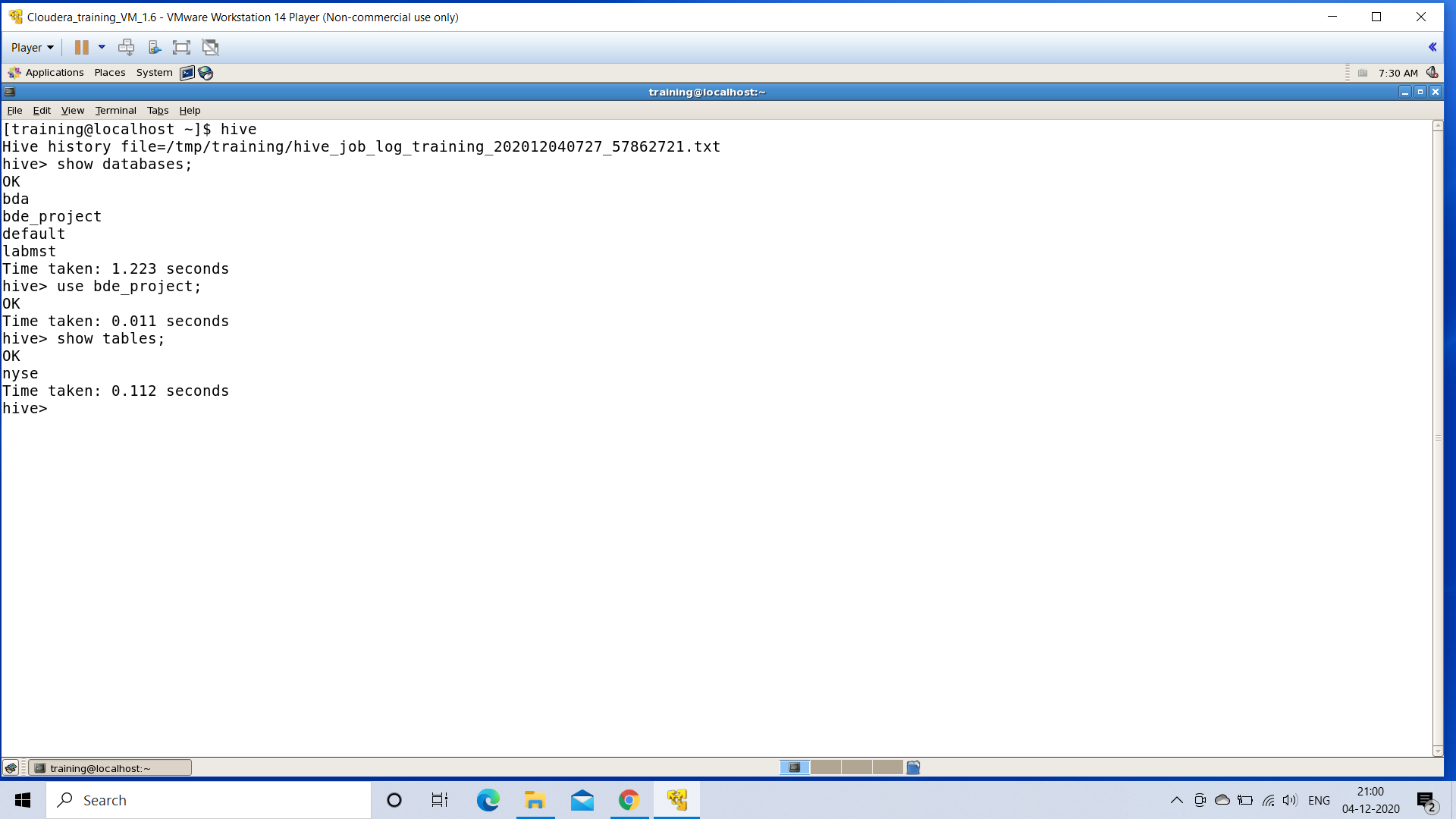
***Syntax:*** hive

Looking up all the databases that we have in the hive

***Syntax:*** show databases;

Using the database of our choice to work in

***Syntax:*** use <databasename>;



1. **Creating table named NYSE**

After entering the database to work. We create a **Hive Table** on which our queries and problem statements will be executed.

***Syntax:*** CREATE TABLE <tablename> (col\_name1 data\_type, col\_name2 data\_type, col\_name3 data\_type)

Row Format Delimited

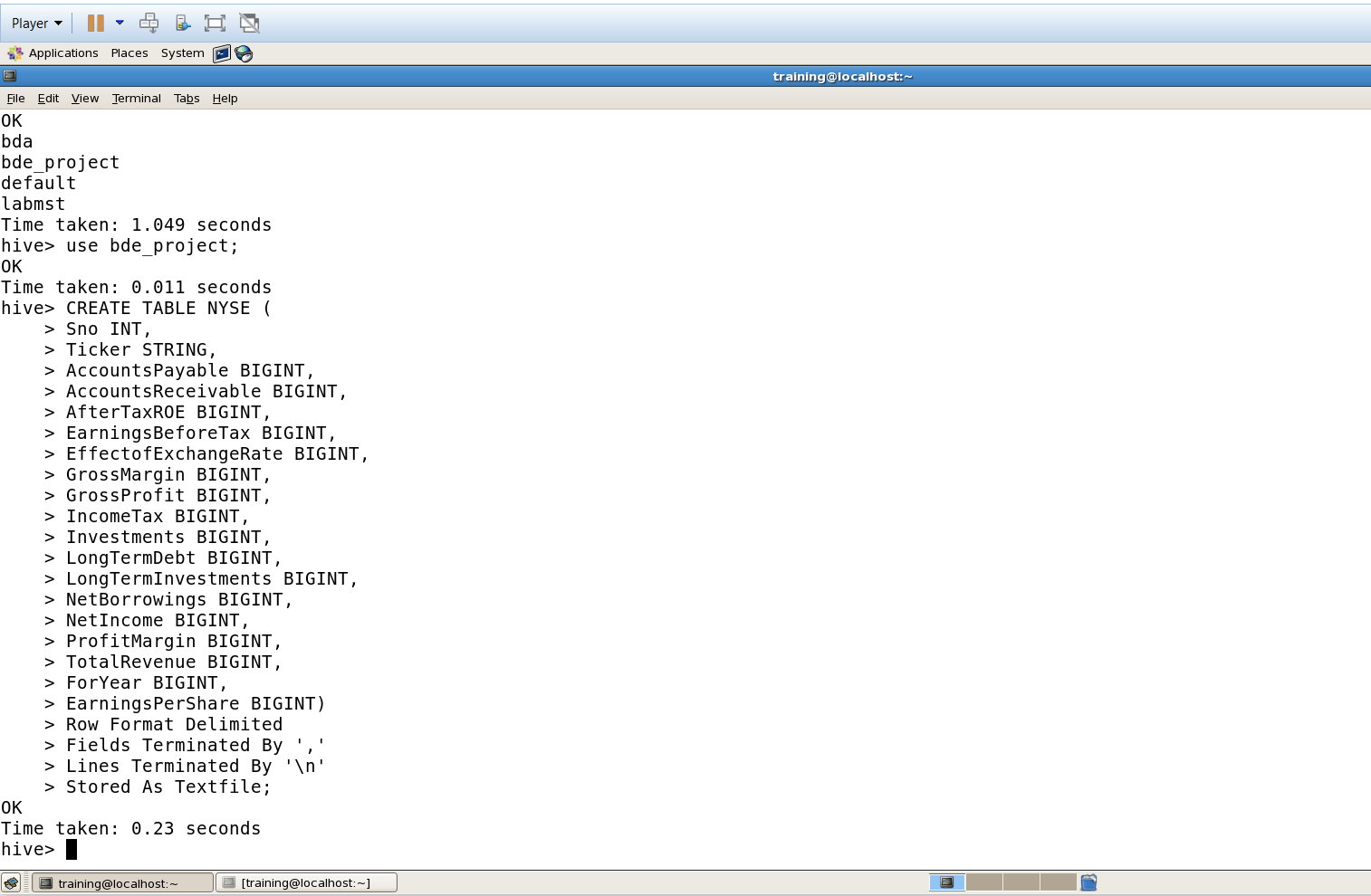
Fields Terminated by ‘,’

Lines Terminated by ‘\n’

Stored as Textfile;

The name of our created **Hive Table** is **NYSE**

* *A thing to keep in mind is that, while creating the table make sure to keep the column names same as the column names of our CSV file from where data will be loaded.2*

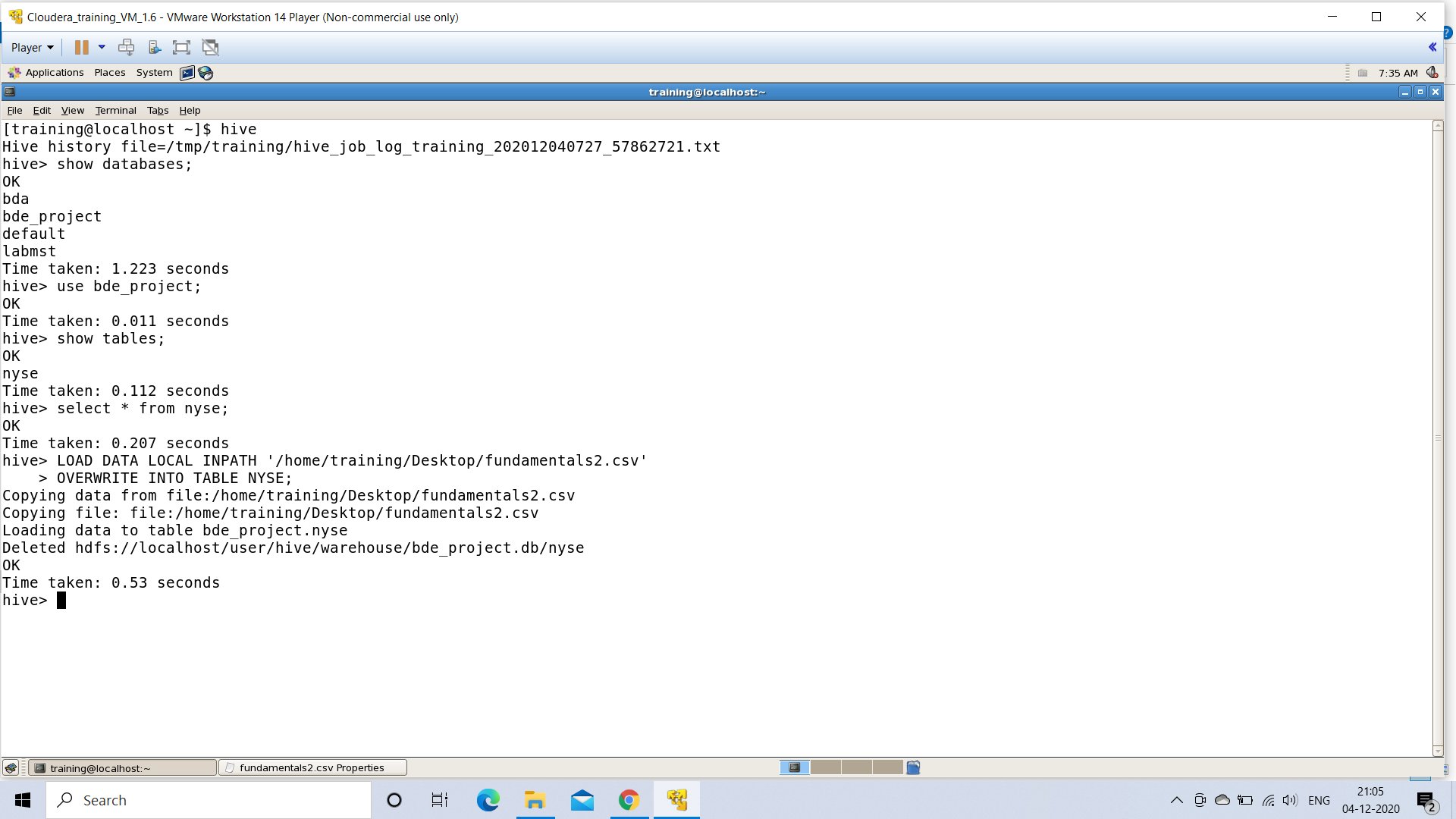


1. **Loading data into the table named NYSE.**

Now that the **Hive table** is created, we need to load our .csv format dataset into this newly created Hive table

***Syntax:*** LOAD DATA LOCAL INPATH ‘location\_of\_our\_csv\_file ’

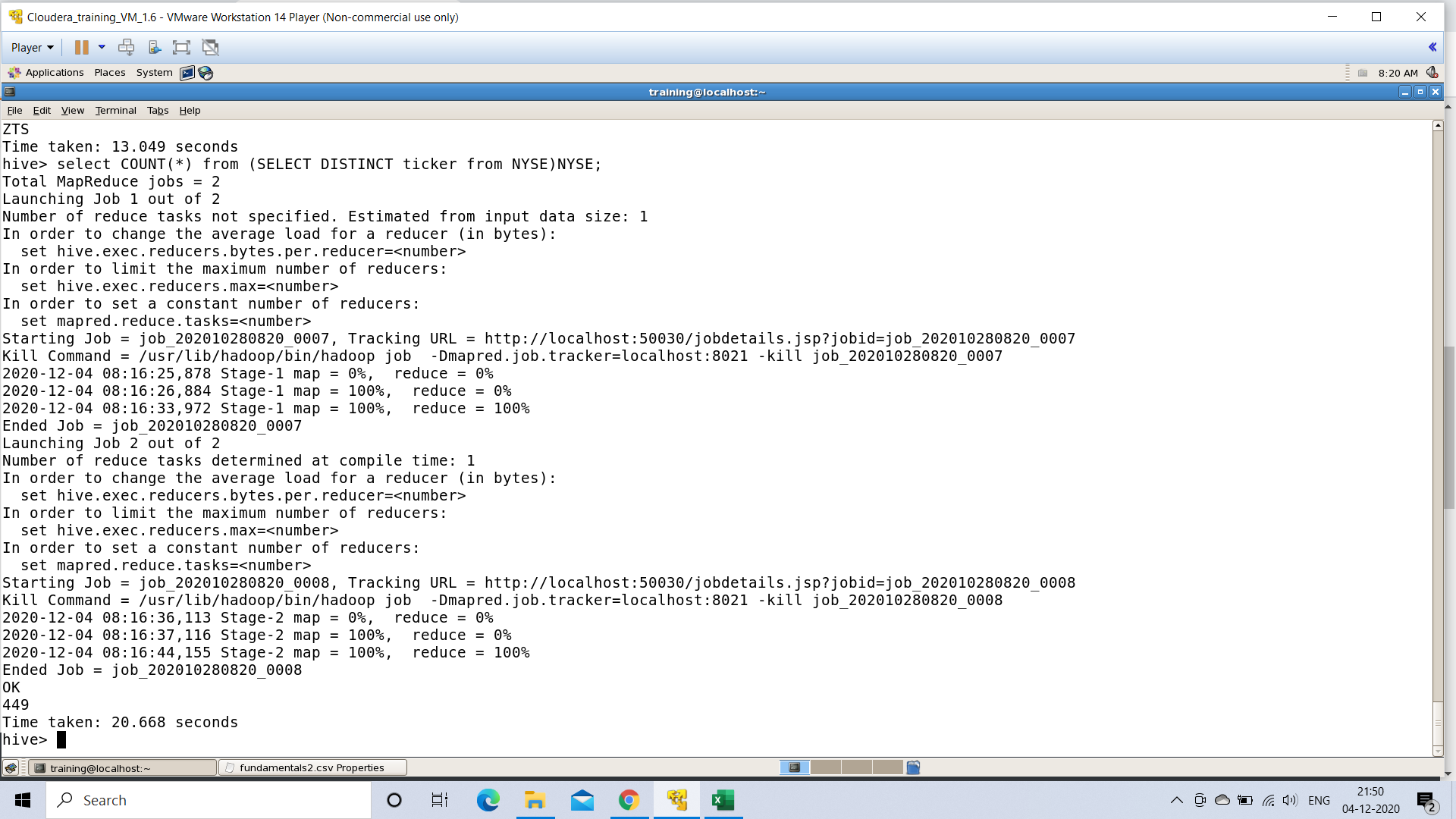
OVERWRITE INTO TABLE <tablename>;



**Executing the Problem Statements**

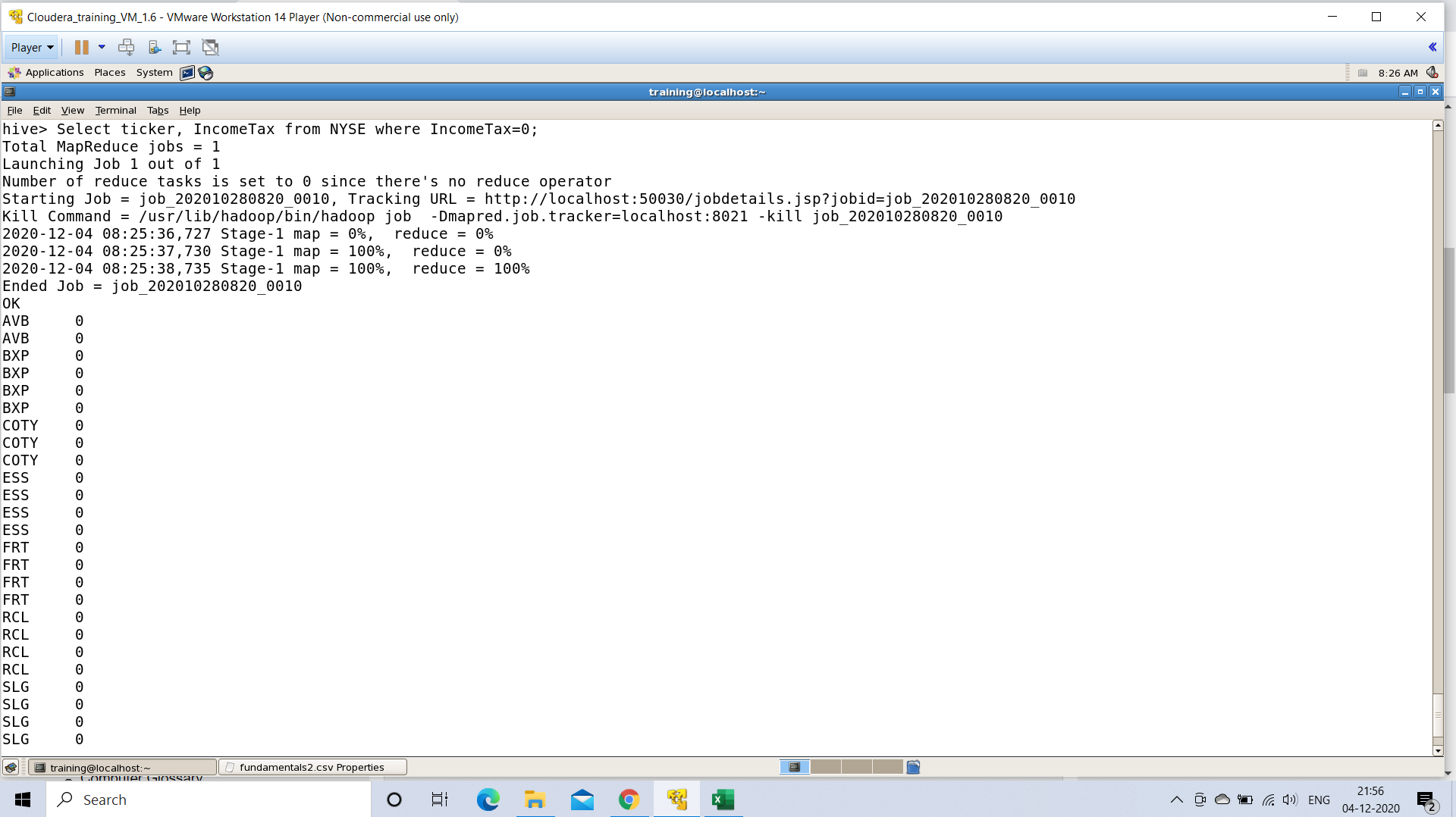
**PROBLEM STATEMENT 1: The total number of distinct ticker symbols in the dataset.**

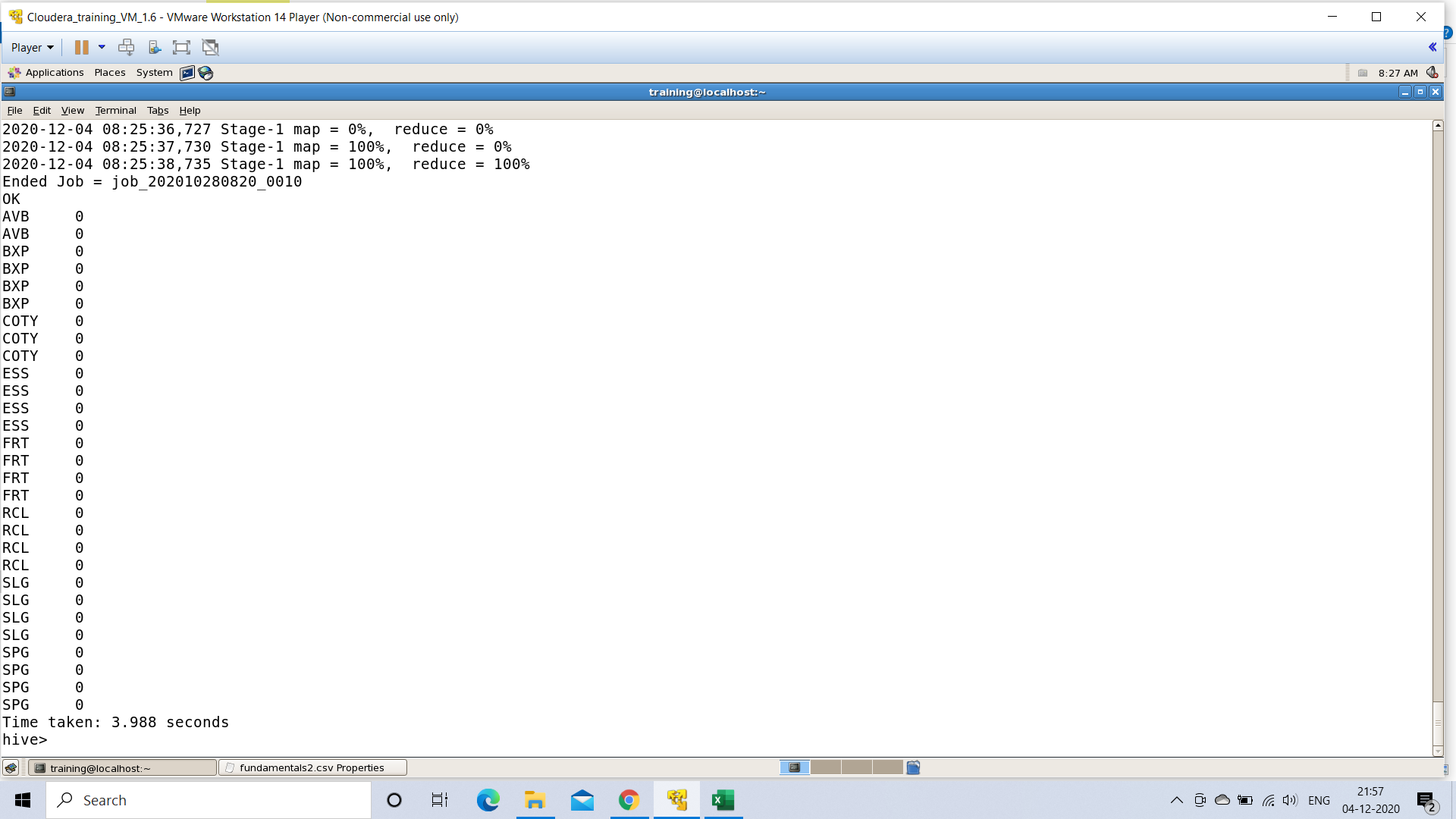
***Hive Command used:*** select COUNT(\*) from (SELECT DISTINCT ticker from NYSE)NYSE;



**PROBLEM STATEMENT 2: Listing the companies which have zero income tax to be paid.**

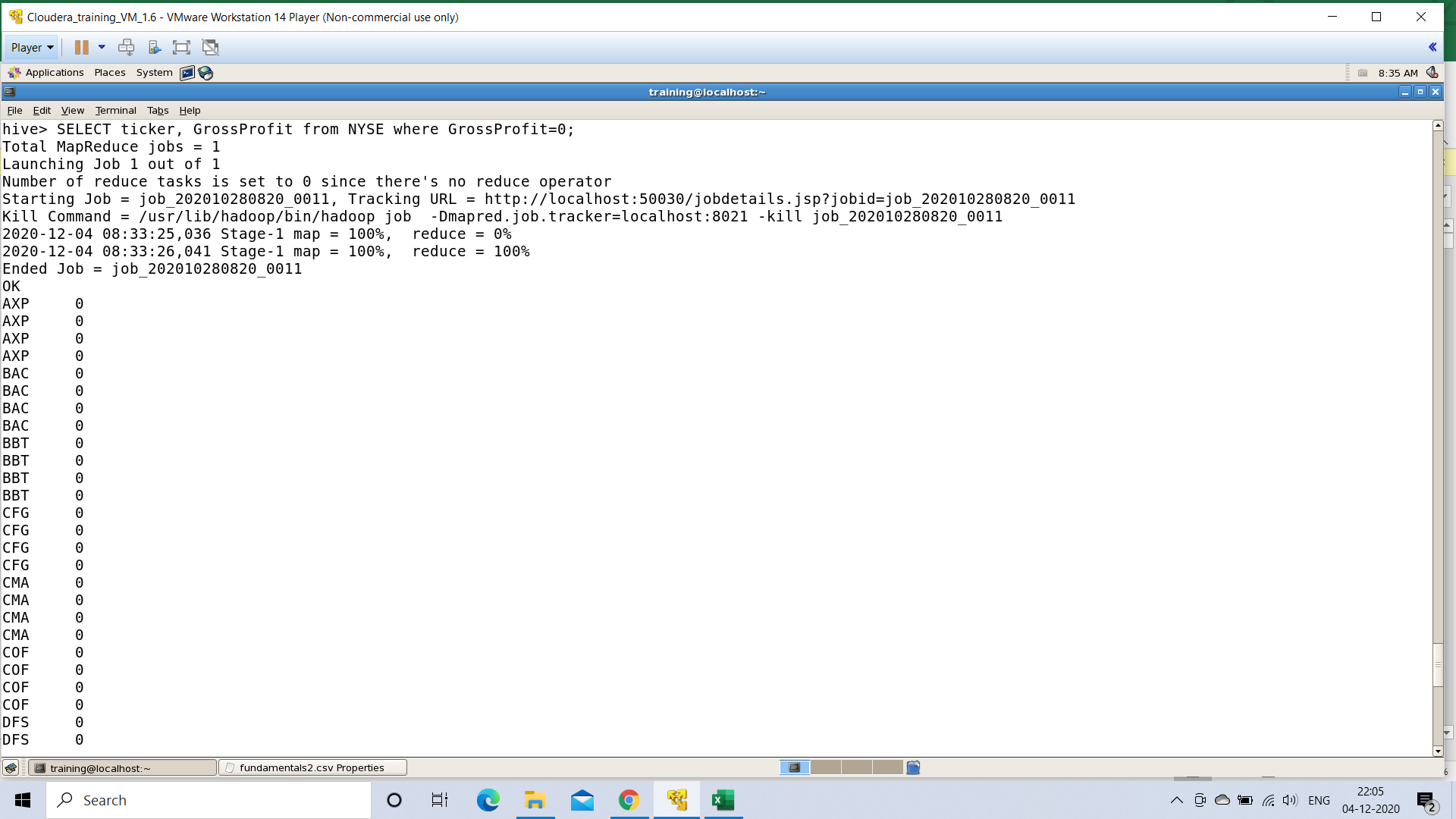
***HIVE COMMAND USED:*** Select ticker, IncomeTax from NYSE where IncomeTax=0;

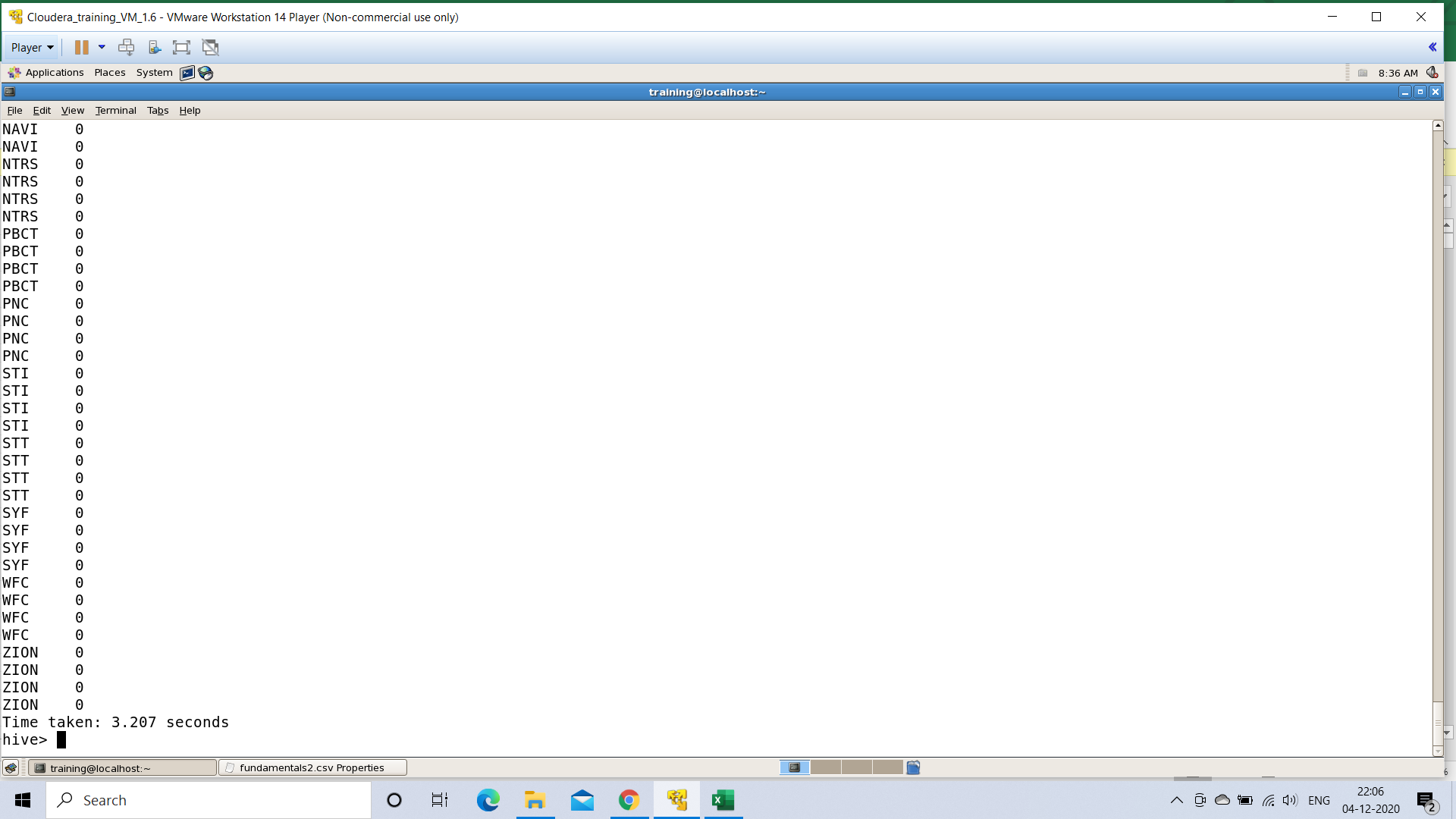




**PROBLEM STATEMENT 3:** **Ticker symbols of the companies which had no profit overall.**

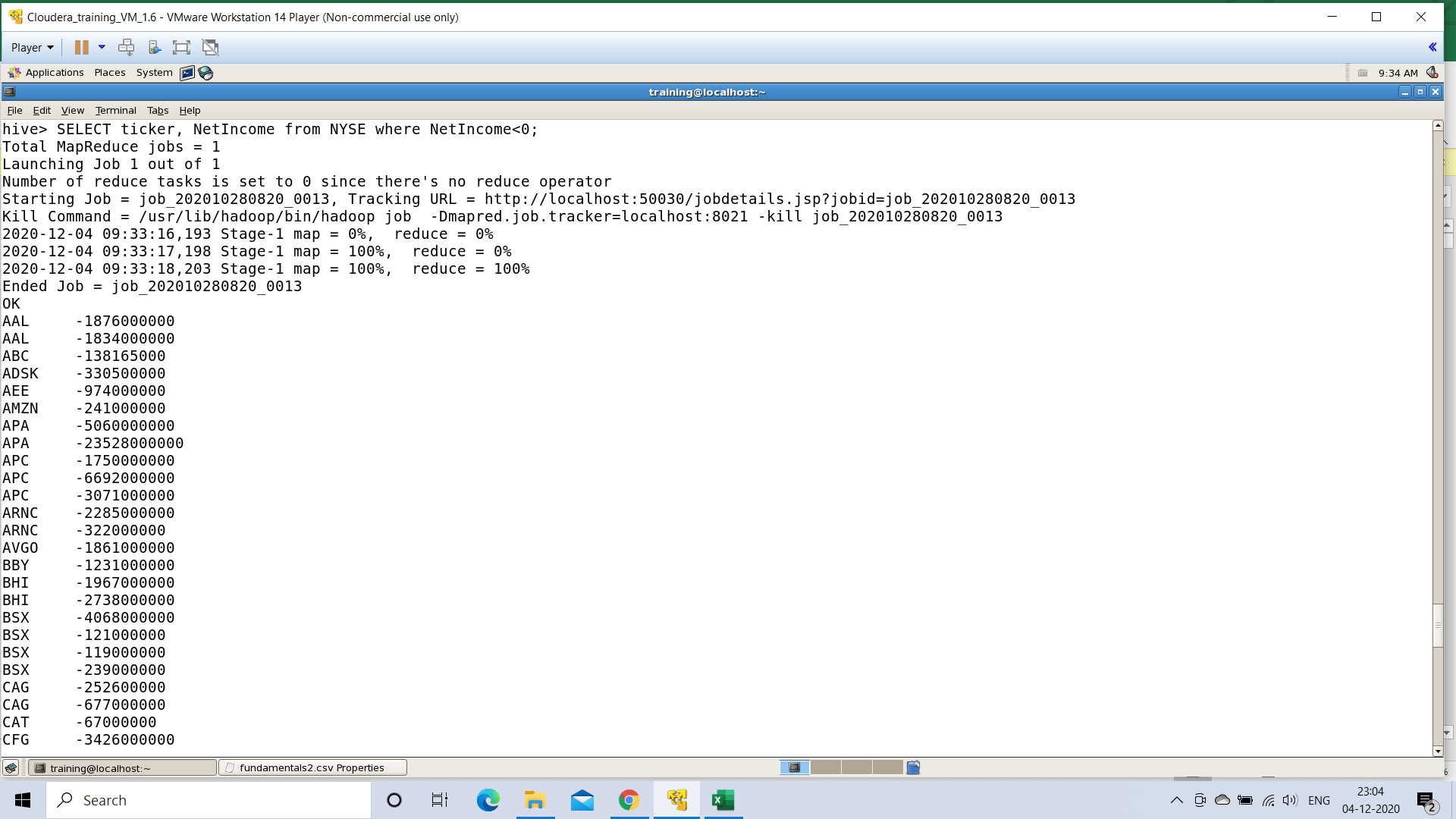
***HIVE COMMAND USED:*** select ticker, GrossProfit from NYSE where GrossProfit=0;





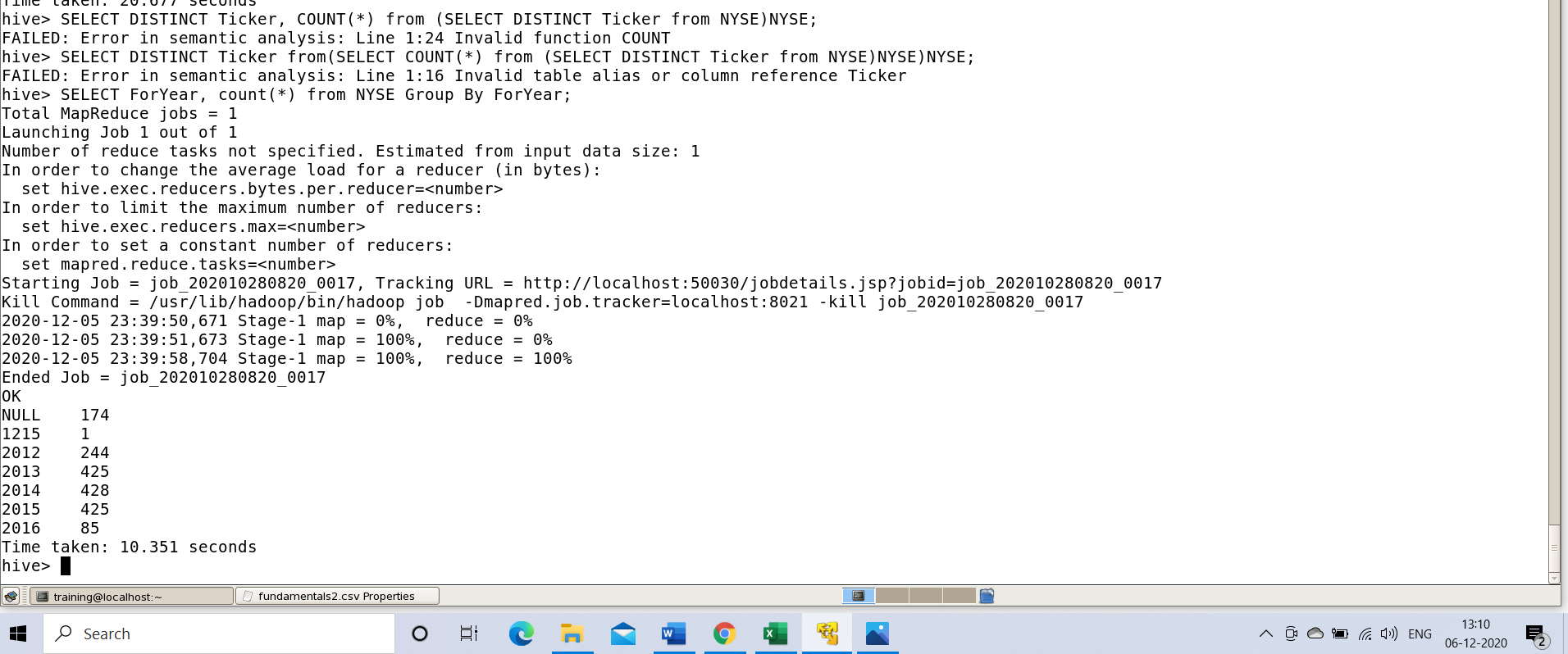
**PROBLEM STATEMENT 4:** **Companies going in total loss and have debts to be paid.**

***HIVE COMMAND USED:*** Select ticker, NetIncome from NYSE where NetIncome<0;



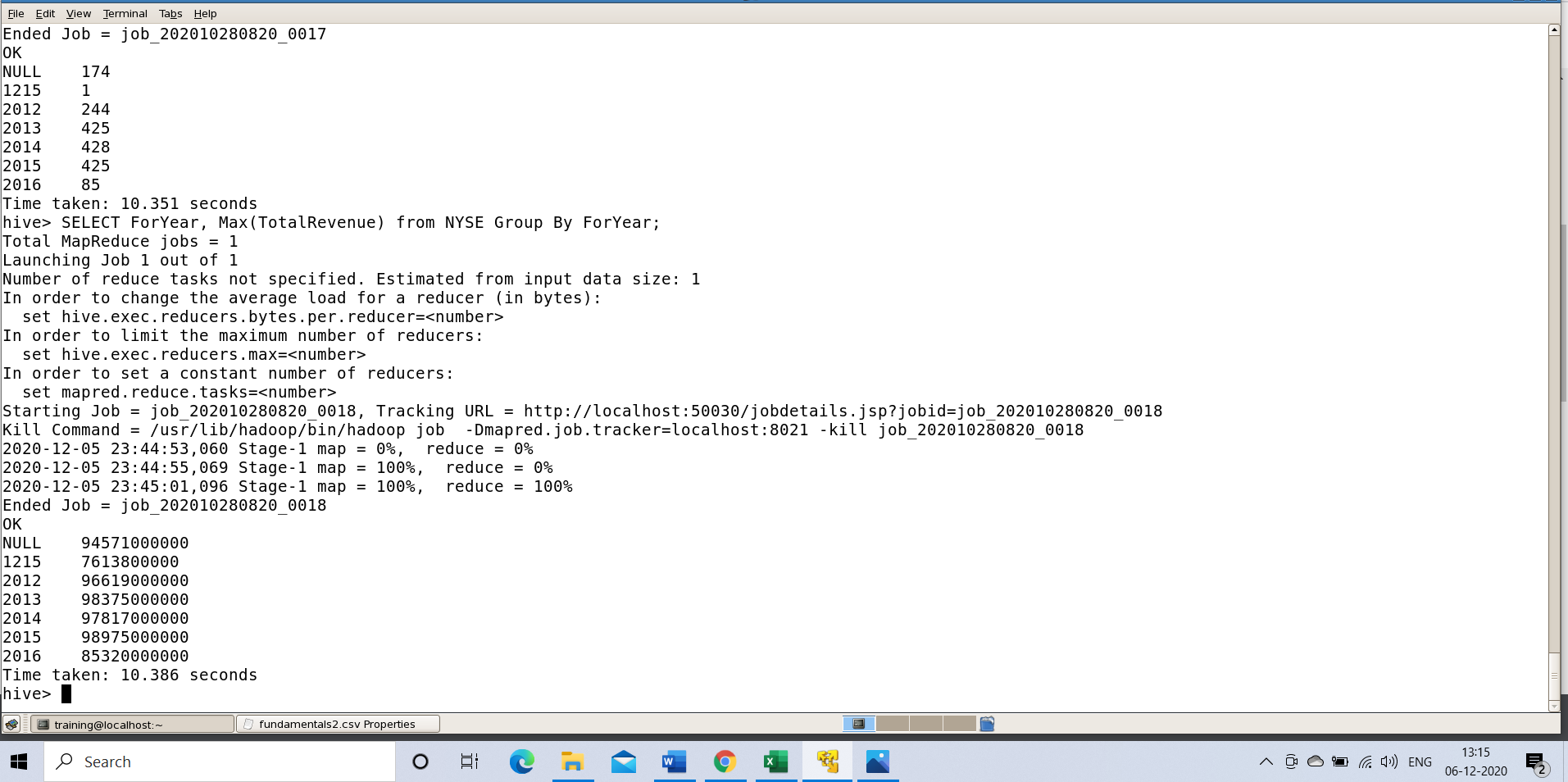
**PROBLEM STATEMENT 5: Total number of stocks filings each year had (2012-16).**

***HIVE COMMAND USED:*** SELECT ForYear, count(\*) from NYSE Group By ForYear;



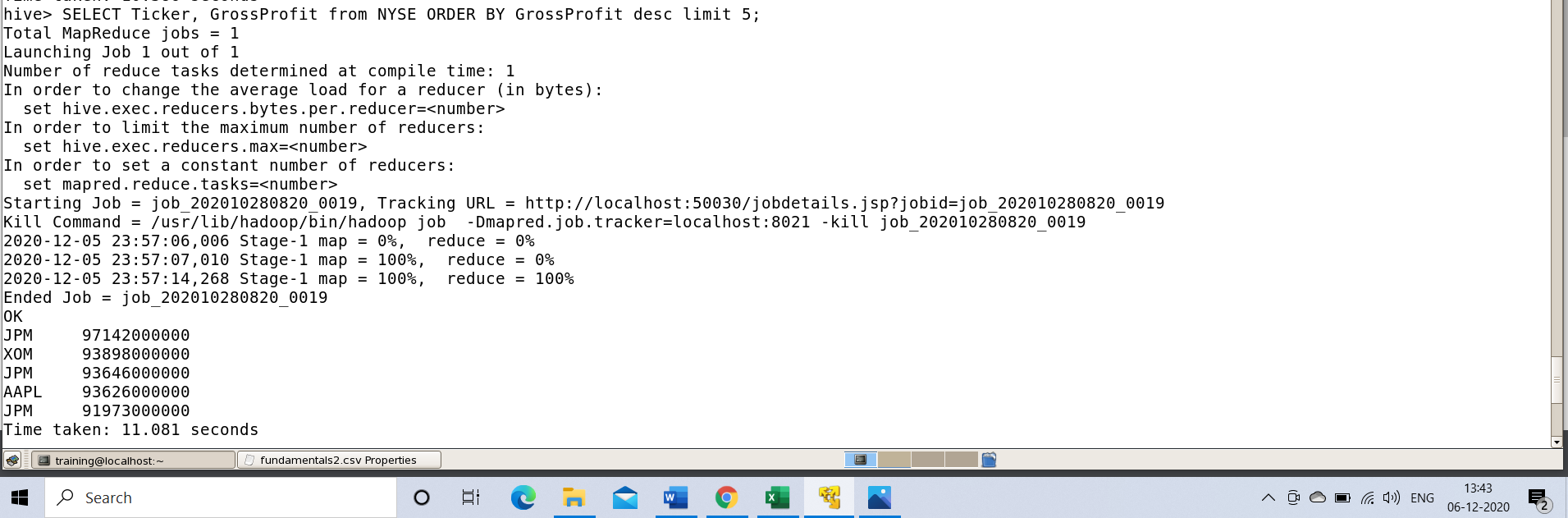
**PROBLEM STATEMENT 6: Maximum total revenue over different years.**

***HICE COMMAND USED:*** SELECT ForYear, Max(TotalRevenue) from NYSE Group By ForYear;



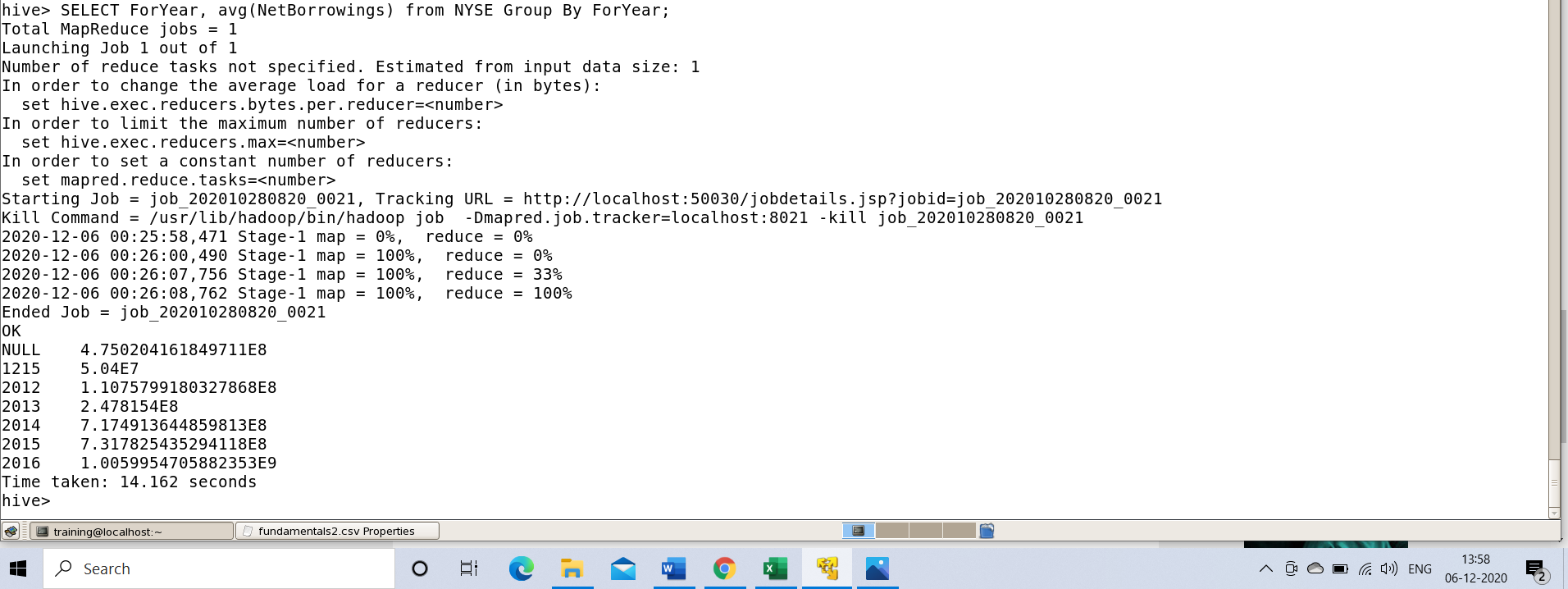
**PROBLEM STATEMENT 7: Top 5 companies with highest gross profit.**

***HIVE COMMAND USED:*** SELECT Ticker, GrossProfit from NYSE ORDER BY GrossProfit desc limit 5;



**PROBLEM STATEMENT 8: Average net borrowings done by companies over the years.**

***HIVE COMMAND USED:*** SELECT ForYear, avg(NetBorrowings) from NYSE Group By ForYear;



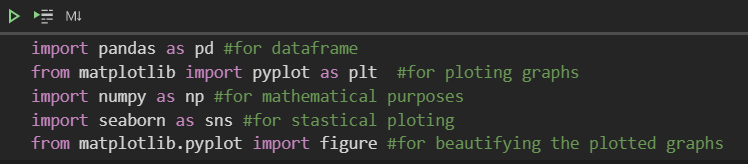
**Visualization of the Data Set**

After obtaining the required Data Set from Kaggle. We needed to understand the data that was to be analyzed.

In order to understand the dataset, the relationship between all the columns of the data. We are using Python3 in Jupyter in VS code to find the relationship among the data elements of our dataset and visualize the same in graphs and charts.

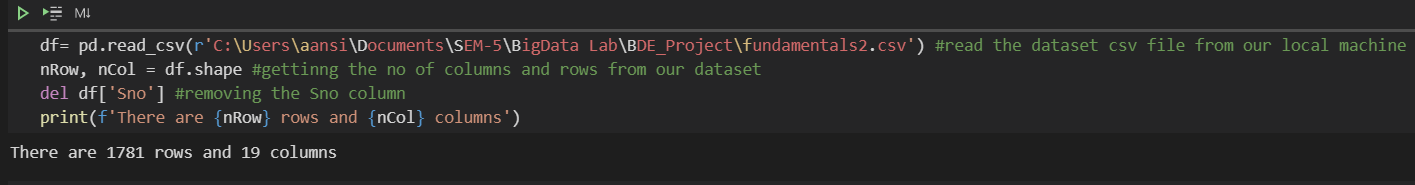
***Library Used***

We are using the inbuilt python libraries for visualizations. This will be done by importing all the required libraries in our **Jupyter Notebook.**



***Preparing the dataframe for Visualization***

We use **Pandas** for preparing our dataframe for visualization.

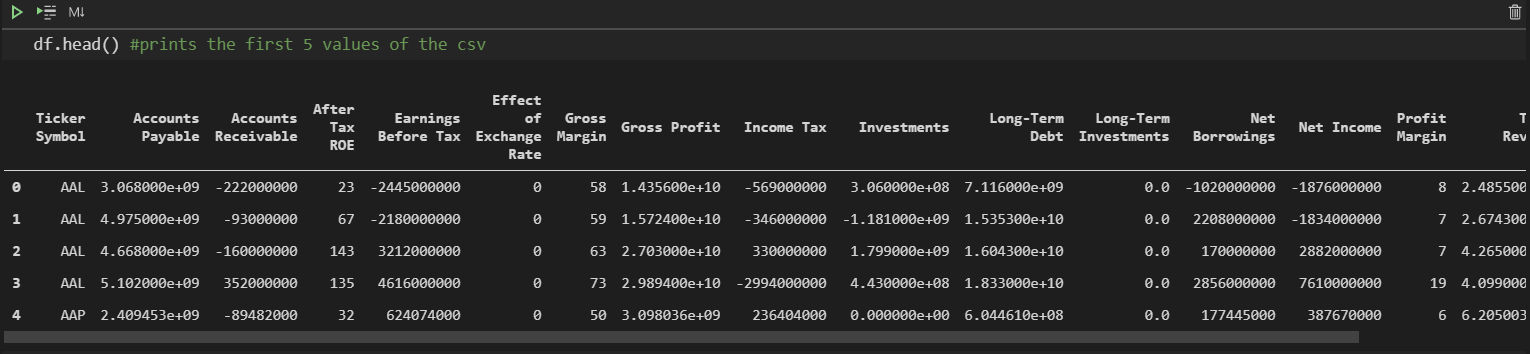


***Visualizing the Dataframe***

Before starting Analysis of the Dataset on our **Hive tool.** We need to understand the dataset that we are dealing with. To understand the dataset, we need to understand all the different relationships that the columns of the dataset have with each other.

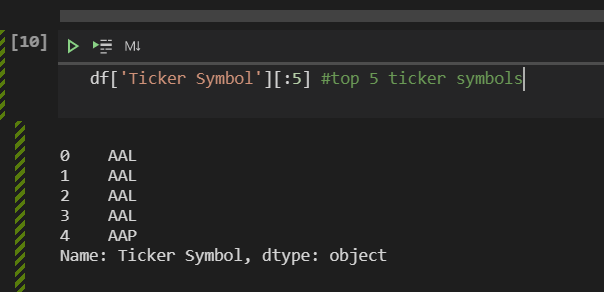
In order to do that we are visualizing our dataset in Jupyter.

***Checking the loaded dataframe***

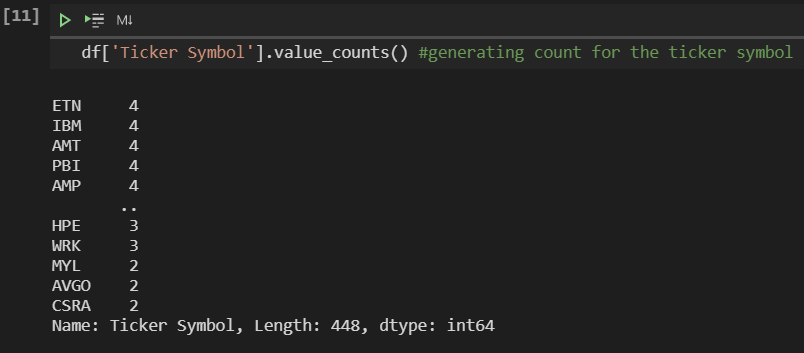


Now, to actually visualize onto the Dataframe.

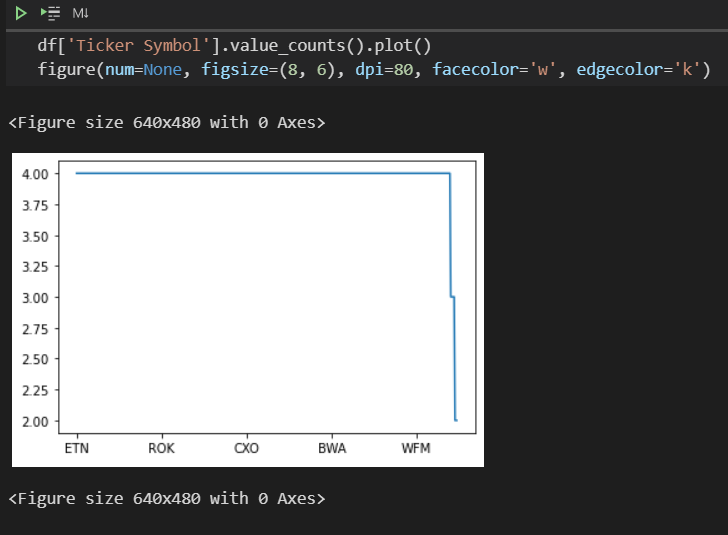
***Checking the top 5 ticker symbols of the dataframe***



***Getting the count of all the ticker symbols***



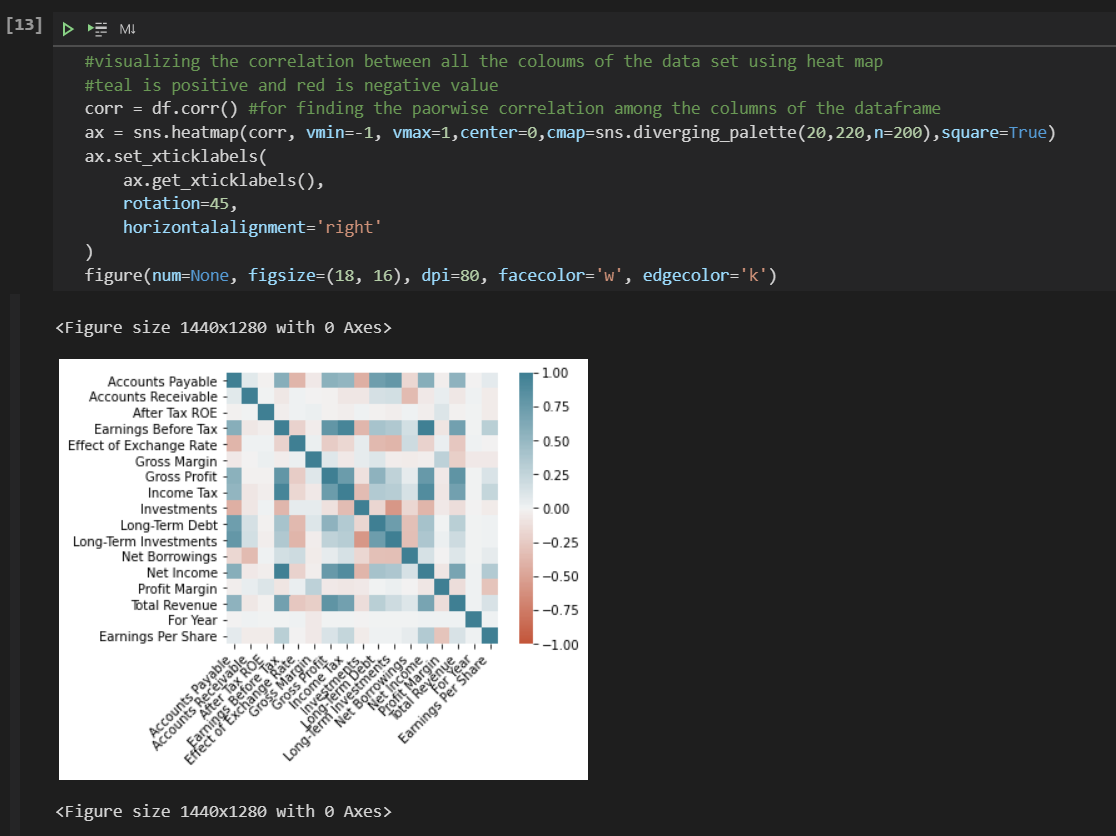
***Plotting a chart for the counts of the Ticker Symbol***



***Finding the behavior of the various dataframe columns***

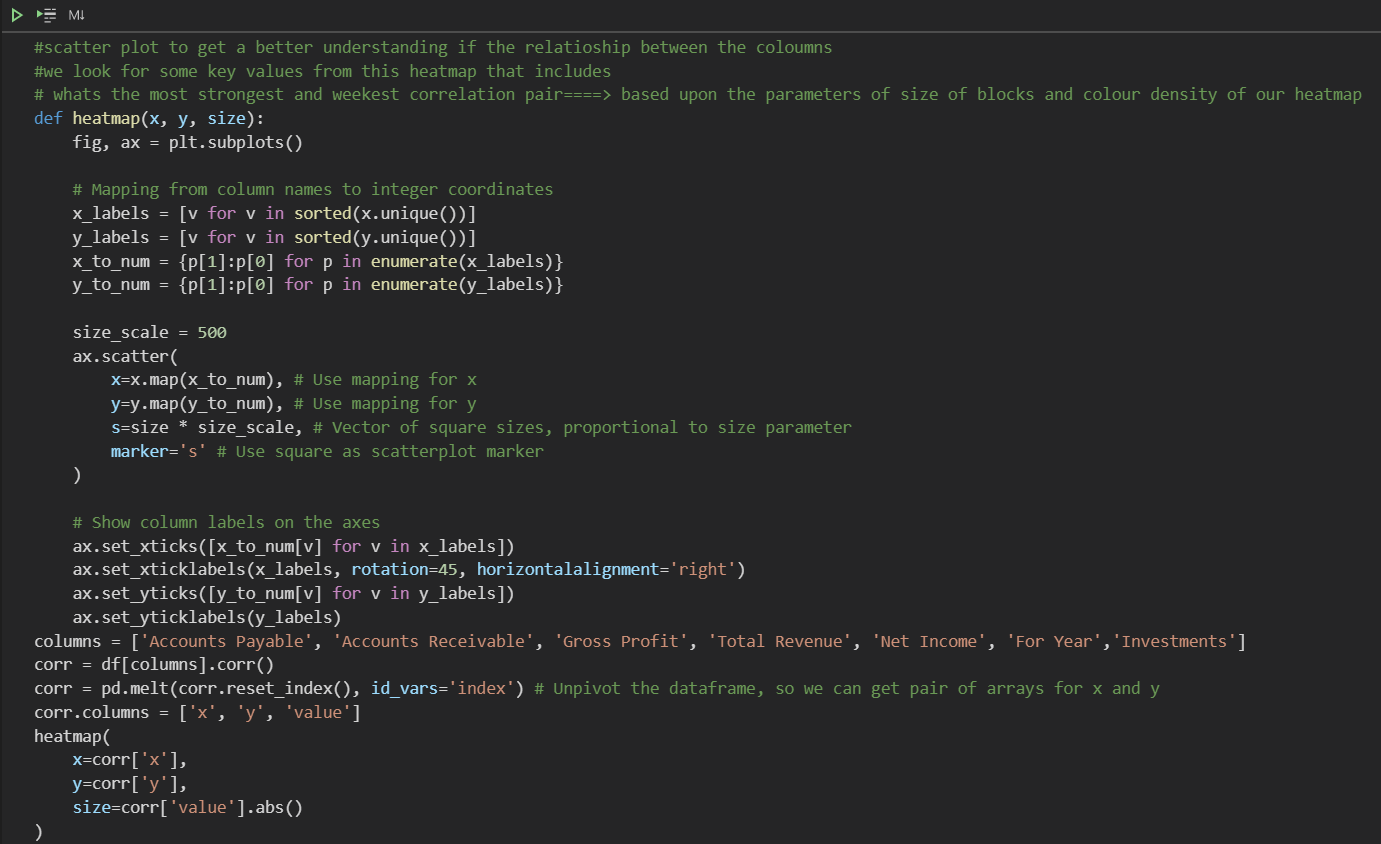
We will be using **Heatmap** of **Matplotlib** for this.

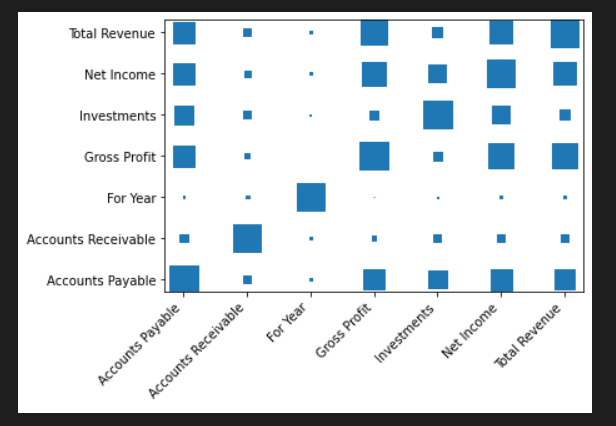
A **Heatmap** is a data visualization technique that shows magnitude of a phenomenon as color in two dimensions. The variation in color may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space.



***Analyzing the dataframe to get understanding of the relationship among columns***

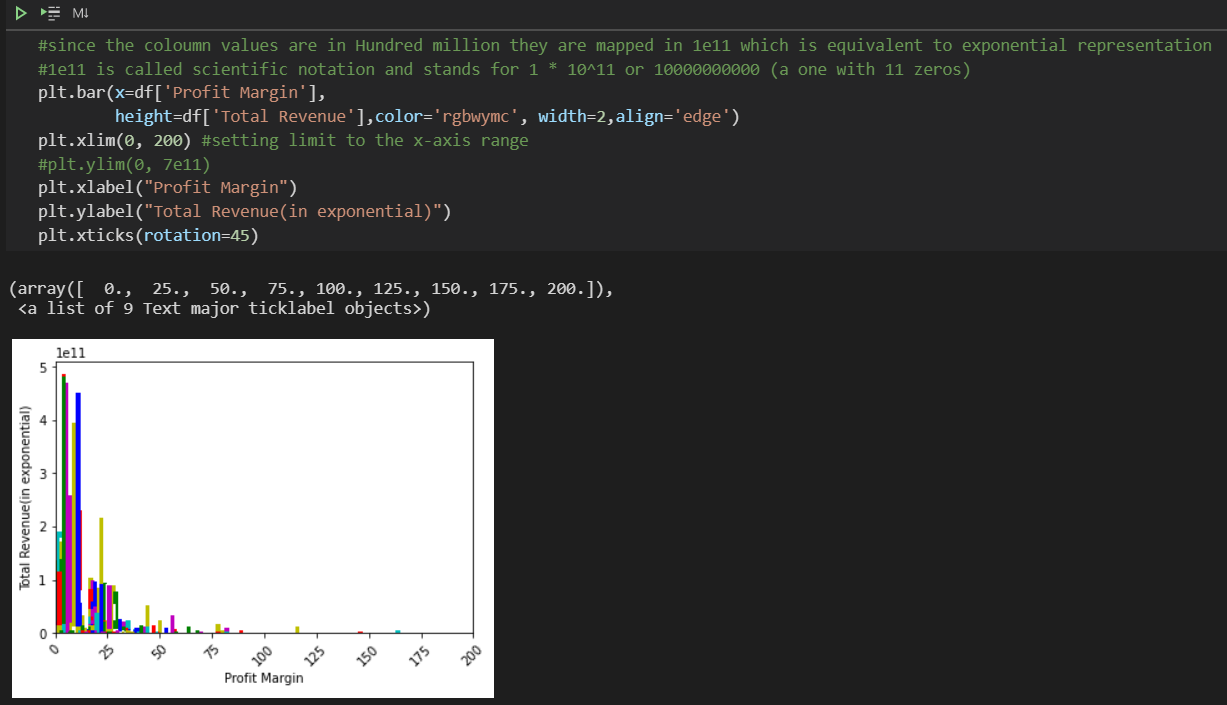
*Making the previous Heatmap even more clear and understandable by plotting the graph between a few select properties.*





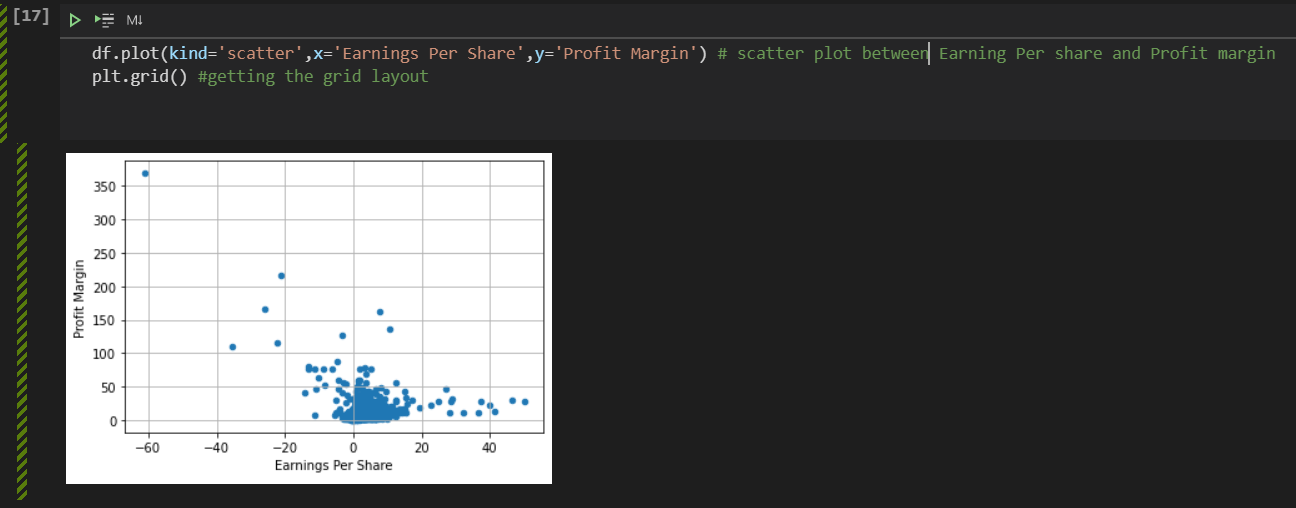
***Finding the Proportional relationship***

Since it’s a Financial Dataset its imperative that we look for a relationship between Revenue and Profit Margin. So we plot a **Bar Graph** to present the comparison between the categorical column values of ‘Profit Margin’ and ‘Total Revenue’



***Scatter plot between ‘Earning Per Share’ and ‘Profit Margin’***

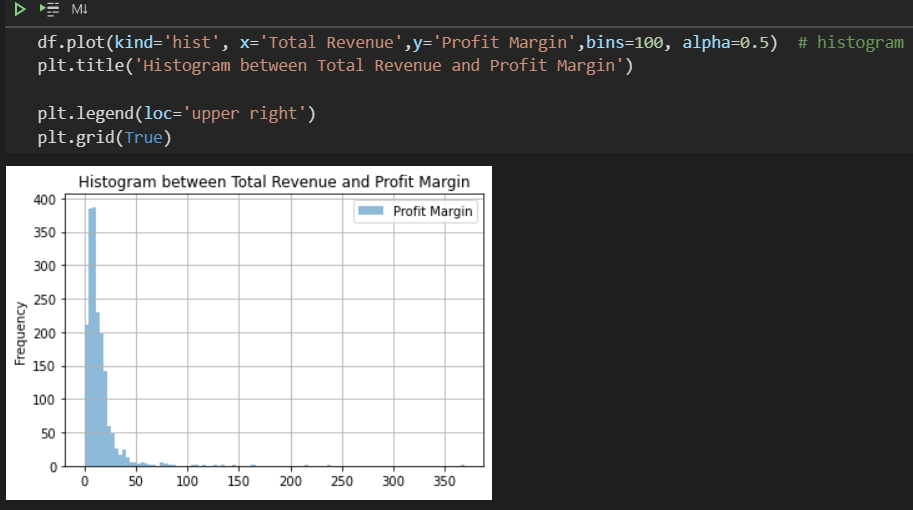
The **Scatter** **Plot** is used to represent relation among **Earning Per Share** and **Profit Margin** how change in one affects the other.



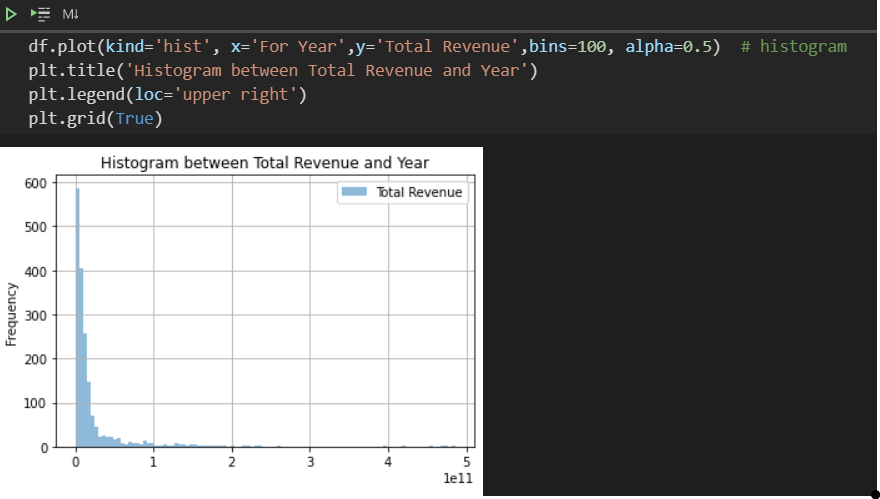
***Plotting a Histogram to get the Frequency of a column value corresponding to another column value***

A **histogram** is basically used to represent data provided in a form of some groups. It is a type of bar plot where **X-axis** represents the **bin ranges** while **Y-axis** gives information about **frequency**

**Histogram between Total Revenue and Profit Margin**

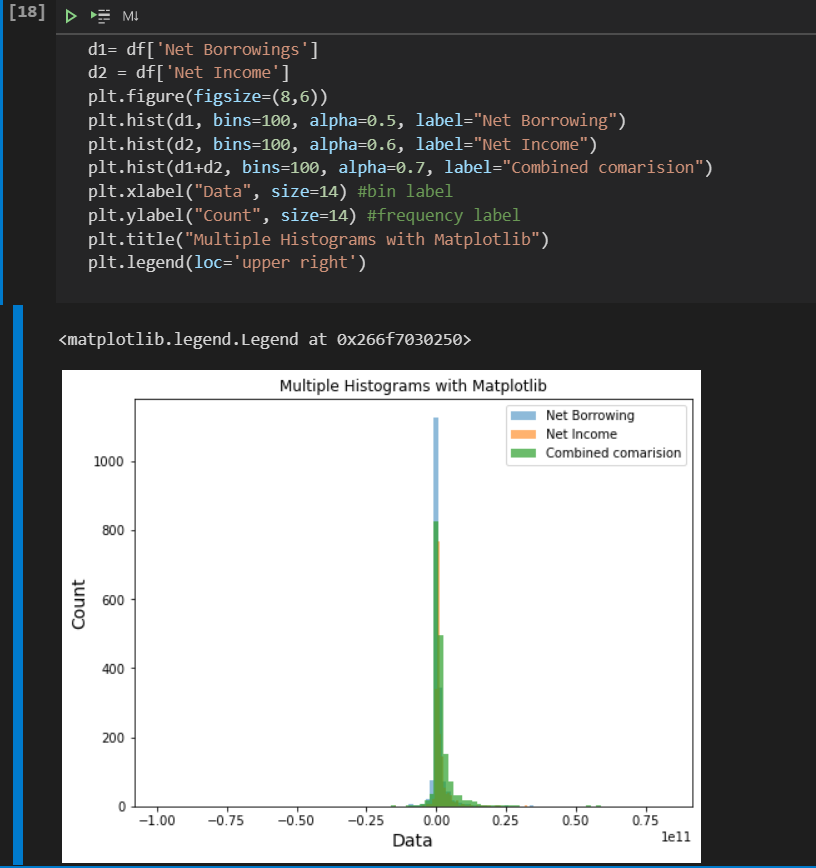


**Histogram between Year and Total Revenue**



An advanced version of these Histograms would be if we can plot several individual histograms in the same graph and then compare the values.

**Individual and Combined comparison of Histogram showing Net Borrowing and Net Income**



**Final Conclusion**

***We have successfully visualized the ‘New York Stock Exchange Data’ and on the basis of this analysis, we have been able to come up with some interesting Problem Statements to make further concrete analysis of this data using Hive Tool.***