## Source Data:

The input data I used is on HDFS and in /source\_data directory named “Data.csv”.

## Directories:

* data\_ingest:

I upload Data.csv on Dataproc and put it into the HDFS with HDFS commands in HDFS.txt. Then, I create a table named data in Hive with Hive commands in Hive.txt. Lastly, I count the amount of data in Presto with Presto commands in Presto.txt.

* etl\_code:

I used MR in java to clean data. I drop the columns which have too many null values and format the Text data by making them all caps for normalization. The codes are in .java files for data cleaning and other .class, .jar files are used for MR jobs. The commands to implement it is in HDFS.txt.

* profiling\_code:
* I put cleaned data into the HDFS with HDFS commands in HDFS.txt. Then, I create a table named data in Hive with Hive commands in Hive.txt. Lastly, I do data profiling in Presto with Presto commands in Presto.txt.
* screenshots:

All the results and process to implement commands or code in previous directories can be seen in screenshots in this directory.