DATA MINING OBJECTIVES & UNDERSTANDING Some of our data mining objectives include:

1. To Load the Data and Access it using 3 methods. API, CSV & TSV.
2. Combine the files to make 1 data frame.
3. Tidy up the Data to get a cleaner Dataframe.
4. Detect and correct 8 quality issues.
5. Visualize the data

DATA UNDERSTANDING
Gather
The data used in this project were provided as follows:
● Dataset Source - Provided by ALX-T - CSV
● API on Tweepy
● TSV The project is done on Jupyter Notebook .
On this platform, the data needs to be loaded before we proceed.
a) Importing the necessary libraries into Jupyter Notebook. Next, we proceeded to load the dataset. - Using the CSV file is easy as the preloaded libraries make it so.
b) The TSV file of data isn't that hard either, as using the tab delimiter makes it easy to gather information from the TSV as a CSV.
c) Operating the API information is the challenging part. The process requires opening up a Twitter Developer's Account such that one can be able to get the credentials to load the files from twitter. Once one loads the credentials, a code is employed to load the tweets and then verify that we have gotten the required tweets from the API.
Once we confirm the same, the next process is to have this information loaded as a JSON file and loaded into the data workload.
Seeing as the instructions were to gather this information from 3 sources, we now have all the data that we require and, as such, can start the cleaning process.

DATA WRANGLING
Before analyzing the data, there is a need to understand the nature of the data presented.
We are getting to display the shape of the data to know how many rows and columns of data we are dealing with. From this, we can see that the columns have the correct data types, and we do not convert any of the rows. - This is a check on the correctness of the data.
Then knowing the information of the data is essential to understand the datatypes of the dataset.
We check for duplicated columns, and while the code results show there are no

repeated columns, an in-depth analysis shows that there are 2 repeated columns, and as such, we drop them.
Next, we proceeded to check on the missing/null values. The dataset had some null values. Since there were quite a number, the threshold code was used to ensure it hastens the process. The decision made was to drop them as they would interfere with the work while bringing no analysis benefits. There were some duplicated columns, and as such, the columns were dropped.
We try to clean up the tweets. To remove the tenses from the tweets, we can get a cleaner file.
We then proceed to check if the columns have any outliers that we need to deal with. While there are some outliers, the information that is an outlier is what should be in the rate of 1-10; hence it is illogical to remove them, as we are trying to make the rest of the ratings conform to this standard.