Data Wrangling Project

Introduction:

Data wrangling is a process required through data analysis and it is a skill that any data scientist must be familiar with, this process is divided into three parts:

1- Gather
2- Assess
3- Clean

This report contains how I have used these three parts to complete data wrangling process.

Gathering:

In this project I have gathered data from three different sources:

1- **The WeRateDogs archive given to us by Udacity.**
I have imported this file on the jupyter notebook the used pandas (read_csv) function to open this file such that it is ready for assessing.
2- **Image prediction file downloaded programmatically.**
By using Requests library the get methods I have downloaded 'image_predictions' then also using pandas I have stored it in a dataframe using (read_csv).
3- **Using tweepy**
I haven't receive The permsission to use the Twitter's API tweepy so I have done as written in the TwitterAPI part in the project section and downloaded the tweet_json.txt which is the file that is ready for analysis with out the need for collecting more data so I downloaded itand Imported it directly on my jupyter notebook to be ready for the next stage which is the assessing.

Assess

The second step in data wrangling is the asses and in this step I check each one of the three sources of data given for improving its quality and tidiness and I have written below each one of the three gathered data sources theirqualoty and tidiness issues in order to simplify the cleaning process and these issues were:

1- In the **The WeRateDogs archive:**

Quality:

- tweet id is float must be string or object
- time stamp is not a datetime variable
- (_) in colomn names should be replaced by white spaces
- rating denominator must be 10 in all rows
- Faulty names must be removed

Tidiness

- dog stages needs to be represented in one column

2- In the image predictions:

Quality:

- tweet_id is float not string or object
- column names is not describable
- p1,p2 and p3 contain (_) instead of white space
- remove unwanted columns

Tidiness:

- All tables should be merged

3- In the Tweepy(JSON) Data:

No issues found

Clean

It this process I have fixed the quality and tidiness issues that I have stated in the assess stage but before doing that I have created a copy for the three resources such that if something went wrong when cleaning the original version is still safe.

Storing

This is the step where I've stored the final data frame that contains the cleaned data and I have stored it in 'twitter_archive_master.csv' and then the data is ready for the final step which is the Visualization and Insight.