Data Wrangling Report

In this report, we will review the steps took to analyze WeRateDog data. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The purpose of this project is to extract insight about Dogs data.

The method of work that was followed according to the following steps :

1- First is to Gather data from data sources,
2- Second is to Assess data that we have gathered and to find the quality and tidiness issues.
3- Third is to fix data issues to get consistent data.
4- Finally, is to extract insight about this data.

Gathering:

The data was gathered and collected from three sources:

1- the first source is the WeRateDog archive, that is available in Project Data.
2- the second source is the image prediction and also available in the Project Data.
3- and the third source was extracted using Twitter API, and is stored as JSON data.

After gathering data from the 3 sources, we moved to the next steps, which is assessing the data.

Assess:

In the Assessing stage, Python techniques were used to find issues in the data. Some of the issues were observed through observation using google sheets, and the other were observed through python programming Tanique, As result, ten quality problems and three Tidiness issues were noted and need to be clean.

Bellow is the issues found in data:

Quality	
twitter_archive table	1- tweet_id should be string, not integer.
	2- Sores with (.) read incorrectly, we need to read the correct value from text column.
	3- rating denominator values should be 10, there is values other than 10
	4- timestamp in archive is object, should convert it to datetime
	5- columns (doggo,floofer,pupper,puppo) have None instead of NaN
	6- source column have full html link, we are interested only in values (iphone, web client..etc)
	7- missing names (None), and invalid names (a, an, O, the ....)
	8- some dog has 2 stages (example tweet_id = '' has floofer and doggo)

image-prediction table:	9- missing records, 2075 instead of 2355.
twitter table	10- id columns should be rename to (twitter_id)

Tideness	
df_archive table	1- df_archive table: columns (doggo, floofer, pupper, puppo) are dogs stages, should be in one columns (dog_stage)
twitter table:	2- Columns (source , text) are also exists in df_archive table, so we can remove them, also we are only interested in only 3 columns (id, retweet_count, favorite_count), so we can remove all other columns
	3- all 3 tables should be combined into one table

Fixing data issues:

then we moved to the interesting part which is fixing these issues using Python programming techniques. we fix all noted quality and tidiness issues and as result we end with more consistent data and we save those data into (twitter_archive_master.csv) file.

Visualization and insights:

In the last steps, we analyzed the clean data, and we got 3 insights, I will review those insight in separate file (act_report.pdf)

Conclusion:

this project took time and effort and it is one of the interesting projects that I hope to continue working on similar projects.