Project 4

Wrangle and Analyze Data
Wrangle Report

By: Alhanoof Alnajashi

Goal of project:

Wrangle and cleaning data from Twitter account @dog_rates to get the most dog images that has the highest retweet and favorites.

Project is divided into 3 part:

- **Gathering**
- **Assessing**
- **Cleaning**

Gathering Data:

Done in 3 formats TSV, CSV, and from JSON file.

Assessing Data:

Identify some issues in quality and tidiness to clean the data set and try to get some insight about it, which are:

Quality:

- Twitter Archive table

1- Drop columns that will not use in the analysis to get fast response in the analysis process.
2- Change timestamp data type to date-time.
3- Change tweet id data type from integer to sting or object. 4- Incorrect dogs name, some dogs have a litter instead of real name.
5- Link source is in a HTML code which can't really know the source of the tweet from iPhone or web.

6- The rating_denominator has some values greater than 10 need to be removed snice all the rating based on 10. 7- None values in name attributes need to be removed.

- Image Prediction table

8- Drop img_num column, since we don't need the number of images in each tweet.
9- Rename some columns to be more readable.
10- Some images for other animals not a dog like turtles.
11- Change tweet id data type from integer to sting or object.

- Tweet table

12- Change id data type from integer to sting or object.

Tidiness:

1- The 4 categories of the dogs (doggo, floofer, pupper, and puppo).
2- Companied and merge table between Twitter_Atchive_df and image_predictions_df.
3- Companied and merge table between Twitter_Atchive_df and Tweet_df.

Data Cleaning:

Cleaning process done by using some function in python to get the information, then fix all the issues mentioned above in assessing step in order to get the insight and make some visualization about the data.