Wrangle Report

By Mohammed Al Mujadib

Introduction:

In this Project I wrangle dataset from weRateDog from Twitter Through 3 steps which are:
- Step1: I collected data from multiple sources which are required in this project.
-Step2: I took a look to dataset that I collected for quality a tidiness issue.
- Step3: I solved quality and tidiness issues that I found in previous step.

1-Gather Data:

I took Data from 3 Sources Which are:
A- CSV file that contains archive tweets for weRateDog account, given by Udacity.
B- TSV file that contains detailed data about images in each tweet, given by Udacity.
C- JSON file that contains specific data about the count of the favorite and retweets, I extracted the file using twitter API.

2-Assess Data:

In this part I passed through datasets to find quality and tidiness issues using Two ways:
A- Visually.
B- Programmatically.

Some of Quality Issues:

1-There are rows that does not have expanded.

2-img_num column should be converted to category date type in Image Prediction Dataset.

3-There are too many duplicate tweet.

4-In twitter archive dataset we should remove the (a) tag and make the column contain the (URL) for the source of tweet only.

5-There are (55) dog who have no name and their name recorded as letter (a) in twitter archive dataset. 6-Timestamp should be converted to datetime.

7-The names and type that is unavailable should be (null).

Some of Tidiness Issue:

1-In twitter archive dataset we should have one column insted of having four columns for each type and we should call that column is (type).

2-Create master dataframe that merge all the dataframes togther.

3-Clean Data:

First, I cleaned the missing values. Then, I cleaned quality issues. After That, I merged columns in the dataset to be more tidiness. Finally, I Organized it.