Wrangle Report

Gathering Data

In this project, we attempted to work with three different datasets:

1. Tweet archives from the Twitter account “WeRateDogs”
2. Prediction dataset that contains the predictions of dog breed based on an image
3. A twitter dataset containing the retween and favourite count of a tweet

These datasets come from different sources, mainly a comma-separated values file that we could download, prediction dataset, which was retrieved using an API request from a cloud storage url and finally, a twitter dataset which was obtained using the twitter API (or by udacity if unable to obtain the access keys).

Data Visualization

Once the data has been gathered, we accessed it for a plethora of quality and tidiness issues. In our assessment, we identified the following issues.

Quality Issues:

1. Consistency – Instances where the rating denominator is lower than the rating nominator, which should not have been the case
2. Consistency – Some columns which represented the same thing but with slight variation (e.g., source & expanded URL)
3. Consistency – Some variable (e.gm Tweet_id) was not standardized across all data types (int64, int64, object)
4. Accuracy – “Name” column in the dataset, which we assume refers to the dog has quite several incorrect inputs (e.g., ‘A’ as a dog name) which we assumed was an error
5. Validity – Columns named p1, p2 and p3 which are all different predictions of objects, animals or dogs
6. Accuracy – datatypes not declared correctly for some columns (e.g, timestamp was an object type instead of datetime)
7. Consistency – Rows of tweets in the twitter dataset without a tweet_id
8. Validity – Redundant columns (e.g., rating nominator/denominator) after we fix these columns

Tidiness/Structural Issues:

1. Last 4 columns (I.e, dogoo, floofer, pupper & puppo) does not make sense at this point of time. Unless we are doing data transformation later on for our machine learning algorithm to take in dummy binary variables only, otherwise, we will remove it for housekeeping's sake.
2. Redundant columns and we can join these tables together to make a master dataset

Data Cleaning

Having accessed the following issues, we did the following :

1. Merged the 3 tables together into one master dataframe on tweet_id
2. “normalized” or fixed the dog rating columns
3. Merged the p1, p2 and p3 columns into 1 single column and preserved the prediction with the highest confidence score
4. Dropped redundant columns
5. Assigned the correct datatypes for the respective variables (e.g, timestamp > datetime64)

After fixing the dataset, we moving on to visualization and analysis.