Wrangle_Report
January 2, 2019
Wrangle Report
Author: Neelam Agarwal

Data wrangling process was carried out for “Weratedogs” twitter page by using multiple files
The following steps were carried out
1) Gather
2) Assess
3) Cleaning

Gather: This was the trickiest part of the project as the data has to be culled out from multiple sources including web (twitter), learning which was time taking. In this I read the archived date provided along with the Image predictions of the tweets along with tweets from the online

Assess: Data for every dataset was observed for missing values, outliers, formats, unwanted columns and so on.

Following points were observed and worked upon:

Quality

1) Timestamp is in "Object" type.
2) Tweet_id should be objects, not integers or floats because they are not numeric and aren't intended to perform calculations.
3) Rename id to tweet_id in twitter data to maintain consistency across the multiple files.
4) Check for the names of dog with None to No Name.
5) Rename the best prediction column names.
6) Proper case of the best prediction dog breeds.
7) Consider the first probable case from the Image predictions file (i.e., P1).
8) Should keep only the original tweets and remove row with retweets.
9) Keep only the observations which are having in either of three predictions to be a "Dog" only (324) from twitter feed.
10) Keep only wanted columns from the twitter downloaded data (api).
11) Most appropriate datatype for Rating_numerator must be float

Tidiness

1) To create a single dog stage column in stead of four different columns.
2) Remove missing values from Dog stages.
3) Need to create one single and master table with all these three tables information pulled together

Cleaning: Following changes were made

Created a master dataframe.
Retweets - Deletion
Replys - Deletion

Timestamp format changed to datetime
Tweet_id datatype changed to object
Rating_numerator changed to float from integer
Cleaning up the data frames - removal of columns
Cleaning of the names - proper case
Keep only the observations which are having in either of three predictions to be a "Dog" only
Considered the first probable case from the Image predictions file (i.e., P1)
Keep only wanted columns from the twitter downloaded data (API)

Hence the data was gathered, Assessed and Cleaned for all three datasets.