Project Name: Wrangle and Analyze Data
Farhan Mohammad
December 2018

The project on Data Wrangling was very intuitive and challenging and I gained valuable skills from working on this project. The steps required to complete project includes gathering the data from multiple data sources in different formats, assessing the data for quality and tidiness issues, and cleaning the data programmatically. The final step performed after cleaning the data was to perform Exploratory Data Analysis to generate valuable insights.

Data for the project was gathered from three sources:

1. The WeRateDogs Twitter archive – This file was provided by Udacity to download in a CSV format.
2. The tweet image predictions – This file was hosted on Udacity's servers and I used Python's Request library to download the file programmatically.
3. Finally querying the Twitter API using tweet IDs in the WeRateDogs Twitter archive and storing each tweet data in JSON format in a text file. This text file was then converted to a Pandas dataframe for further assessment and cleaning purpose.

The next step in the Data Wrangling process was Assess phase in which I looked at the data in detail to check for quality and tidiness issues. I used Panda's info function to get a concise summary of individual data files. I specific issues I looked for in the data sets included missing data, incorrect data, duplicate data, data types and redundant data. Finally I have identified and documented several issues with data quality and tidiness which is listed as follows:

Quality Issues¶

1. Includes retweeted status values which is not required. Only keep retweeted_status_id with values as NaN
2. Datatype for timestamp columns need to be changed from str(object) to datatime format
3. The name column has lot of incorrect or incomplete values. The name used most often is a
4. Several columns have empty values, such as in_reply_to_status, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp
5. Replace URLs in the source column to the source text
6. Remove redundant data such as duplicated tweet_id and tweets with no pictures
7. Change the data type of ratings_numerator and ratings_denominator to float

8. The columns p1, p1_conf, p2, p2_conf etc can be categorized into a single column of dog breed

Tidiness Issues

1. Similar data (tweet_id) in 3 different tables
2. Dog stages in multiple columns. It needs to be combined to a single dog stage variable

As part of Data Cleaning phase, I fixed the Quality and Tidiness issues which were identified programmatically. I combined the three tables by joining with tweet IDs into a master DataFrame so the structure of the data will be easier to work with and all the necessary data cleaning can be performed on this single master dataframe. The issues were fixed in three steps namely Define, Code and Test.

The final master dataframe was used for exploratory data analysis. I produced multiple insights and visualization from the data which were very interesting.