Report on the wrangling process

Introduction:

the dataset wrangled in this project is the weratedogs.WeRateDogs is a twitter account that rates people’s dogs with humorous comment about the dog.

This project works through the data wrangling process which includes gathering, assessing and cleaning of data. There are also analysis and visualizations involved in the project.

GATHERING:

the WeRateDogs Twitter project data was gathered from multiple sources which are:

The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file was provided by Udacity. This archive contains basic tweet data for all 5000+ of their tweets as they stood on August 1, 2017.

The tweet image predictions, i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network. This file was provided by Udacity.

• Twitter API and Python's Tweepy library to gather each tweet's retweet count and favorite ("like") count at minimum, and any additional data.

Assessing:

After the gathering process, I began to assess the data and look out for both quality and tidiness issues

There are four main issue in quality dimensions:

1. Completeness: -Missing data

-tweet_id is an int

1. Validity: -dog names: some dogs have 'None' as a name, or 'a', or 'an

-the dataset includes retweets

1. Accuracy: -Time-stamp is a string datatype

2. Consistency: -In p1, p2 and p3 columns there is an underscore for multi-word dog breeds

There are three main requirements for tidiness:

1. Each variable forms a column:

-(dogoo, floofer,pupper, puppo) should be under one column

1. Each observation forms a row
2. Each type of observation unit forms a table

Cleaning:

Cleaning data is tedious and iterative.Therefore, After assessing, I cleaned the data using the framework below:

1. Define: Determine exactly what needs to be clean and how.
2. Code: Programmatically clean the code
3. Test: Evaluate the code to ensure the data set was cleaned properly.

In [ ]: