This report is created with the purpose of documenting the data wrangle efforts made for 1475 tweets in 2017 from Twitter user @dog_rates, also known as WeRateDogs.

The dataset wrangled is extracted from Twitter account WeRateDogs. They rate people's dogs with a funny caption about the dog. These ratings almost always have a denominator of 10, and the numerators almost always greater than 10 such as 11/10, 12/10, 13/10, etc. This absurd rating has been questioned and they answered, in quote: "they're good dogs Brent." WeRateDogs is a famous on social media with over 4 million followers on Twitter and over 2 million followers on Instagram. They also received international media coverage.

There are 3 sources of data that need to be gathered. The first one is Enhanced Twitter Archive which has basic tweet data for 5000+ of their tweets. This dataset includes 16 columns. The second source is Twitter API which I extracted the retweet count and favorite count of tweets. The third source is Image Predictions File in which the dog breed is predicted based on its image. These dataframes have many columns and not all of them are necessary for the analysis. There are also many quality issues and tidiness issues in those datasets such as wrong values or datatypes.

After gathering enough data, each data source is accessed one by one to find critical issues that directly affect the analysis results and other issues. For this step, both Visual assessment and Programmatic assessment are used to find out the issues to be cleaned. After the while, 10 issues are identified:

(1) Absurd name for dogs such as 'a', 'None', 'this';

(2) Retweets are taken as tweet entries;

(3) Column retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, source (in df_tweets df) and p2, p2_conf, p2_dog, p3, p3_conf, p3_dog (in image_predictions df) are not relevant for the analysis;

(4) Wrong datatype of timestamp column in df_tweets table;

(5) Wrong datatype of tweet_id in tables,

(6) Duplicated values in Expaned URLs column;

(7) Some tweets are not dogs;

(8) Rating_numerator and rating_denominator are not extracted correctly,

(9) One variable in four collumns in df_tweets table (pupper, floofer, puppo, doggo);

(10) One observation unit belongs to three tables.

These issues are tackled one by one using Python, resulting in a gathered, assessed, and cleaned master dataset call Twitter_archive_master. Now the dataset is ready for analysis.