Wrangle Report

Omar Abuhassan

#Project Overview

The dataset that I will be wrangling then analyzing is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

#Steps of Wrangling Data

1. Gathering data
2. Assessing data
3. Cleaning data
4. Storing the wrangled data

#Data Gathering

Data was gathered from three different resources:

1. The WeRateDogs Twitter archive (Local)
2. The tweet image predictions (URL)
3. Additional Data (Twitter API)

#Assessing Data

We looked for issues in **quality** and **tidiness** using two types of assessment:

1. Visual assessment
2. Programmatic assessment

#Assessing Data

Quality issues

1.[Programmatically] In archive, there are some retweets.

2.[visually] In archive, 'source' values in unneeded template.

3.[Programmatically] In archive, some tweets does not have a photo.

4.[Programmatically] In archive, incorrect dog names.

5.[Programmatically] In archive, erroneous datatypes for some columns.

6.[visually] In archive, there are unneeded columns.

7.[Programmatically] In image_predictions, there are duplicated images.

8.[Programmatically] In archive, rating_denominator have different values.

9.[visually & Programmatically] In api_data, id column should be string and named 'tweet_id’.

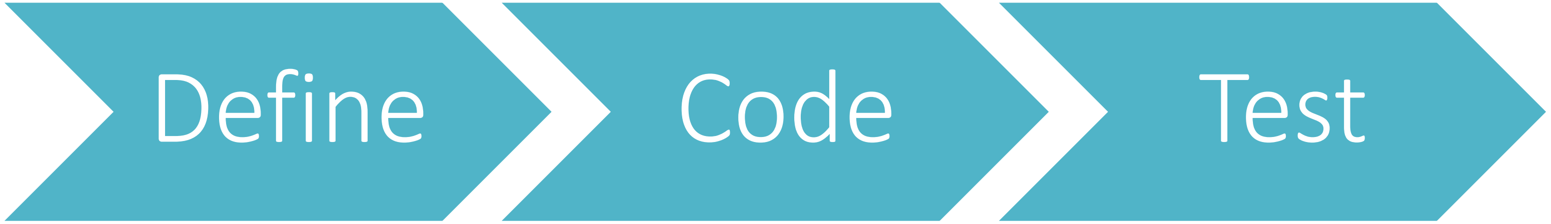
#Assessing Data

Tidiness issues

1. [visual] In archive, columns (doggo, floofer, pupper, puppo) should be one column "dog_stage" with different values.
2. [visual] In api_data, (retweet_count, favorite_count) should be part of archive table.

#Cleaning Data

We resolved the issues that observed in the assessing section using this cycle:



Define

Code

Test

#Storing the wrangled data

Finally, we store the cleaned data in “twitter_archive_master.csv” file.