Introduction

In this project we use real world data to gather, assess, and clean data, a process which is called Data Wrangling.

There are three pieces of data that was gathered, assessed, cleaned, and then transform into pandas Dataframe to analyse.

The project data was gotten from the Twitter account called 'WeRateDogs' (@dog_rates) which "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."

Data Wrangling Effort.

The Data Wrangling process involves:

1. Gathering Data
2. Assessing Data
3. Cleaning Data
4. Storing, Analysing and Visualizing Data

Gathering Data

Data was gathered from three source

i. Directly downloading a CSV file from Udacity classroom (twitter_archive_enhanced.csv)
ii. Programmatic download from Udacity's server (image_predictions.tsv) url: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
iii. Using Twitter API to Query tweet count in the Twitter archive and save it as a JSON file (tweet_json.txt)

Assessing Data

The collected dataset where first carefully observed visually in Excel. We then accessed it programmatically using python in Jupyter Notebook with pandas.

Tidiness issues was observed:

In the twitter archive data, we discover that:

- 1. Since Project Motivation indicated that we only want original ratings (no retweets) that have images.We Remove the rows with retweets
- 2. Missining values (NAN) In in columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_timestamp, and expanded_urls.
- 3. Not all denominators are exactly 10 and there are a lot of large numerators

- 4. Missing values and messy names in column name
- 5. There should have been a column for type or stages of dogs so that doggo, floofer, pupper and puppo can be accomodated in
- 6. Data type issues were noticed in twitter_id and datetime columns-which also have extra 0000 that need to be removed
- 7. Check in the text column to determine dog sex
- 8. Source column look messy and clumsy, was removed as it wont be necessary for analysis

In the twitter image prediction data, we discover that:

- Columns p1, p2, and p3 consists of a mixture of uppercases and lowercases in the values

Looking at the twitter count data (tweet_df) we discover that:

- 1. The twitter archive and twitter count data looks similar, this should be merged
- 2. Column id_str should be changed to tweet_id and convert to object to enable us merge tables.

Tidiness Issues

- Inaccurate Data Type for tweet_id and Datetime columns
- Inappropriate column name for Column id_str in the tweet_df dataset

Cleaning Data

Data was cleaned using the define, code, test format

Twitter Archive Data Set

- ✓ Remove the rows with retweets using python isnull() function
- ✓ Remove all columns with missing values that is (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_timestamp, and expanded_urls.)
- ✓ Scale numerator to 10.
- ✓ Rename column to rating and change type.
- ✓ Check for inappropriate dog name and change to NaN
- ✓ Create a column for doggo, floofer, pupper and puppo and change data type to categorical
- ✓ Change data type for twitter_id from (int64) to object using the astype() function and datetime from (object) to datetime using pandas to_datetime() function.
- ✓ Use str.strip() to remove extra 0000 from datetime
- ✓ Determing dog sex using code to check for she and he pronoun

Twitter image prediction data set

- ✓ Using str.lower() we convert all columns to lowercase
- ✓ Change tweet_id type from int to object

Twitter count data set

- ✓ Change Column id_str should be changed to tweet_id and convert to object

<u>Storing, Analysing and Visualizing Data</u>

Store the clean master copy in a csv file named: twitter_archive_master