Wrangle Report

This report is about the wrangling process I did through this DAND project. I did several things

1. Gathered the data.
2. Accessed it.
3. Cleaned it.
4. Analyzed and Storing it.

1. Gathered the data.

There are 3 sources for data gathering:

1. ***twitter_archive_enhanced.csv***: Directly download CSV file

Use *pd.read_csv* import into pandas data frame.

2. ***image_predictions.tsv***: Programmatic download from Udacity's server

The tweet image predictions is present in each tweet according to a neural network. This file is hosted on Udacity's servers and downloaded programmatically using the *requests* library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image predictions/image-predictions.tsv

3. ***tweets_df***: Query from Twitter API

Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's *tweepy* library and store each tweet's entire set of JSON data in a file named *tweet_json.txt* file.

5. Accessed it.

Quality

- All the values in the denominator and numerator tha are less than 10 will be removed, because there are not useful because of their low quality.

- Some of the columns names will be renamed and will be illustrated after.

- Entries that don't state any type pf dog will be removed

- P1, P2, and P3 will be renamed

- No need for the jpg_url and it will be removed

Tidy

181 retweets and 78 tweet replies will be removed due to its repetition and to make sure all of the data are tidy and have unique values.

3. Cleaned it.

Summary of Arch cleaning:

- I recognized the need for removing useless coloumns as it is clear on the last cleaning step I did.

- Howevery others I stated before will do them.

Summery Of Image Preds:

- I droped the clean image and the coloums that aren't needed for me like the type of image.

- I changed two names to be more clear like, p1_conf to prediction_confidence.

4. Analyzed and Storing it.

Store the clean df in CSV file with name using *.to_csv(Wrangle_Dog_Analysis.csv')*