Data Wrangling Report on WeRateDogs twitter archive Report

Gathering of Data:

For this project data has been gathered from three different sources. And data has been gathered using the different methods mentioned below.

- Importing data via csv
- Using requests to download data off internet
- Scrape data from an API

The Three Data Sources

Enhanced Twitter Archive

The WeRateDogs provided by the Udacity. This contains 17 columns and more than 2000 tweets. This file is download manually from Udacity which is in csv format and loaded using the pd.read_csv() command.

Image prediction file

The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet. This file image_prediction.tsv is hosted Udacity's servers and should be downloaded programmatically using the request library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

Data using Twitter API

Each tweet's retweet count and favourite ("like") count at minimum, and any additional data you find interesting. 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 called tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas Data Frame with (at minimum) tweet ID, retweet count, and favourite count.

Assessing the Data

After gathering all the 3 files, their data is stored into a data frame for easier assessment and cleaning. In order to assess the data, I examined it visually and programmatically using python's pandas library. First, we printed out all the data frames entirely, used the info() function to assess the datatypes, used describe() function to summarize the quantitative variables in the datasets, etc. Then examined the data frames more specifically by examining each variable separately and found out the following issues.

Quality

- missing and incorrected dog names and the most popular dog name is 'a' which is not a name given by owner
- TimeStamp is in string format.
- source not extracted from hyper link tag.

- Columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_user_id, retweeted_status_id and retweeted_status_timestamp, have a lot of null values.
- Gender of dog could be extracted from text.

Tidiness

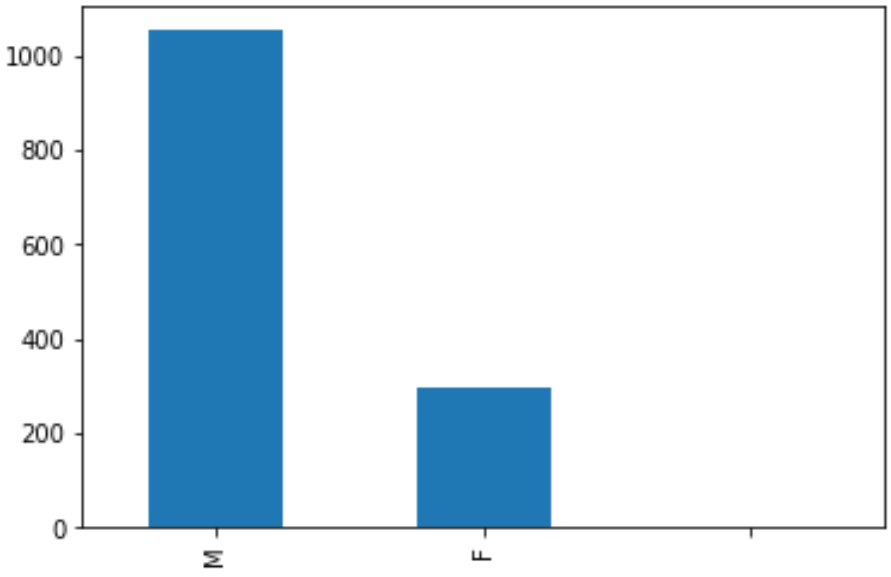
- The dog stage columns in twitter_archive can be arranged into a single column.
- The image predictions could be condensed to show just the most confident dog breed prediction.
- All three dataframes can be combined into one single dataframe.

Cleaning process

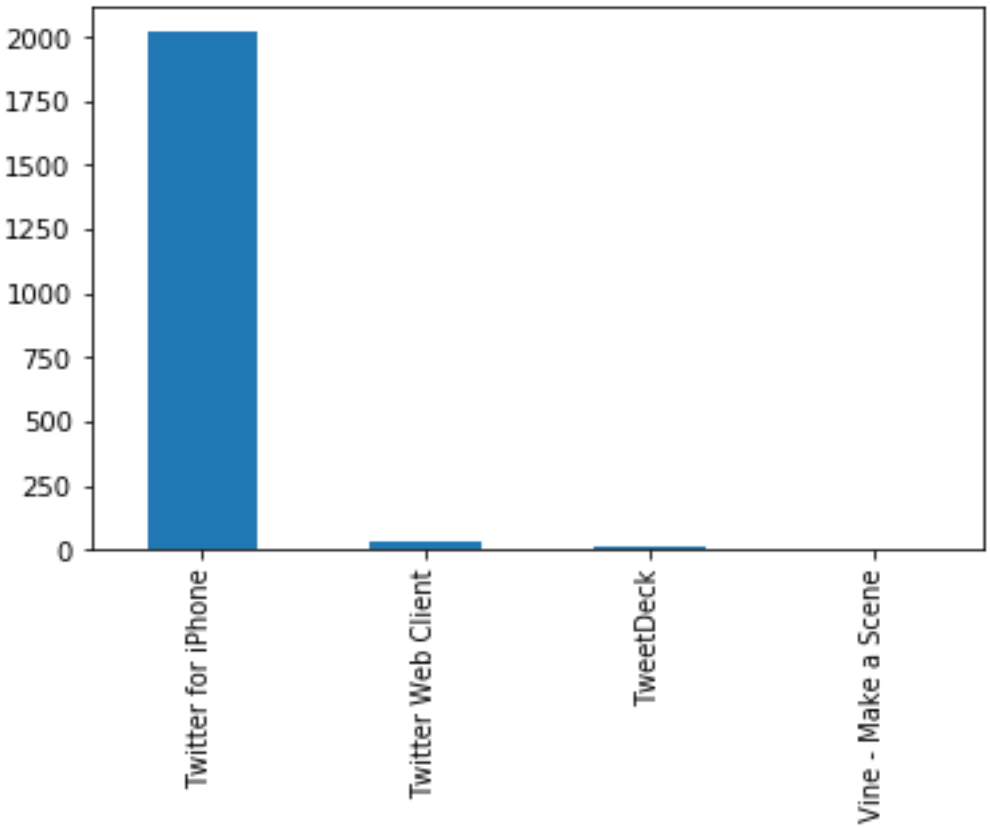
Cleaning process consists of three steps: Define, code & Test. First, we define how to tackle the issue. Then, we code to resolve the issue and finally we test our code to see if the issues with the data have been resolved. So, in order to clean these 3 data frames, these 3 steps are carried out for each of the issues and was finally able to achieve a clean data frame. The issues which has been mentioned in Assessing the Data is carried out in cleaning process

Visualization

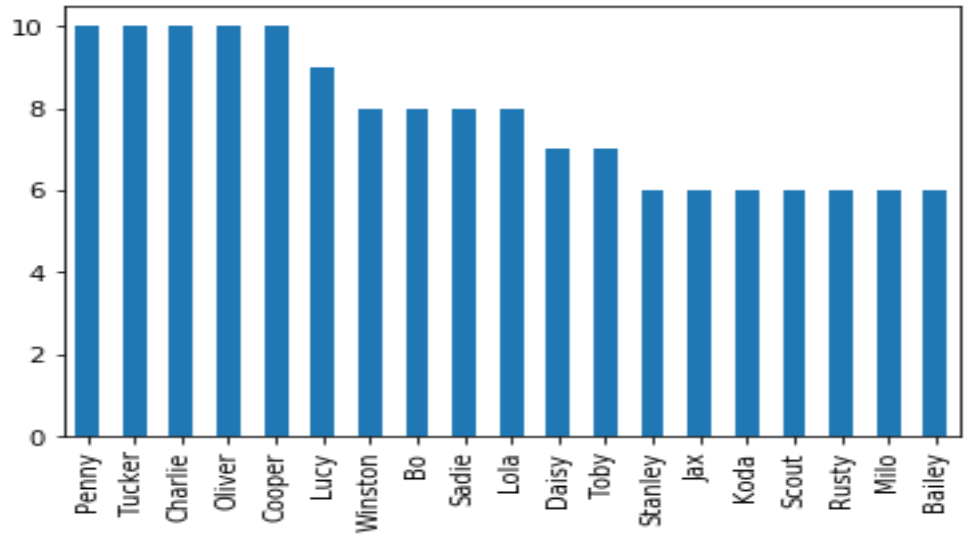
Gender Analysis: from the Graph we can clearly analysis that male breed is more preferred by the owners.



Top Sources: Out of the 4 sources, Twitter for iPhone is clearly the most widely used source to share tweets



Most popular names Given to dogs: Penny, copper, Charlie and Oliver are the most popular names can be seen from the below fig.



Average Retweet and Favourite counts for dog breeds : Standard_poodle had the highest average retweet count while Saluki had the highest favourite count.