Wrangle Report

1.Project Details

this project is as follows:

- Data wrangling, which consists of:
  - Gathering data (downloadable file in the Resources).
  - Assessing data
  - Cleaning data
- Storing, analyzing, and visualizing your wrangled data
- Reporting on 1) my data wrangling efforts and 2) my data analyses and visualizations

2.Gathering Data

Gathering Data for this Project

Gather each of the three pieces of data as described below in a Jupyter Notebook titled `wrangle_act.ipynb`:

1. The WeRateDogs Twitter archive. Download this file (`twitter_archive_enhanced.csv`
2. The tweet image predictions, i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network. This file (`image_predictions.tsv`) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. Twitter API for each tweet's JSON library and store in a file called `tweet_json.txt` file.

3.Assessing

Visually and programmatically find out:

- there are many no useful data
- there is an error in the time date-time "timestamp."
- some NaN be as NONE
- some column stores the same values such as Puppo and pupper etc

4. Cleaning

Quality

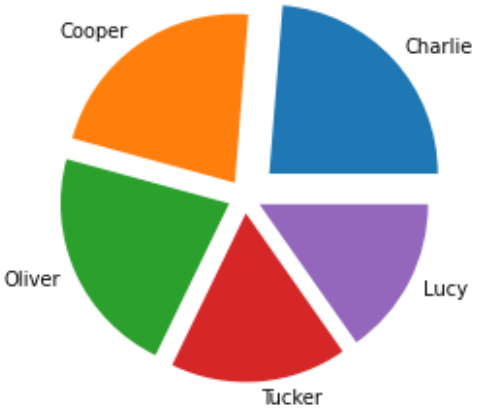
1. Clean the source column makes it readable.
2. Refine p1, p2 and p3 columns and confidence as the probapility of p2_conf and p3_conf both is lower than 0.5 we will drop theses unuesful coulums
3. rename p1 , p1_dog and p1-conf makes it readable
4. drop unusefull columns (in_reply_to_status_id), (in_reply_to_user_id),(retweeted_status_timestamp) ,(retweeted_status_user_id), (retweeted_status_id)
5. dropping duplicates columns refar to datatime (created_at ,timestamp which is object not in a good format .
6. check and Delete duplicated tweet_id
7. change any none to NaN
8. ues text extract gthe dog name adds it to new coulme dog_name and drop the column name

Tidiness

1. Fusion in one column and mix the four columns of the dog type: "puppe," "pump" and "floofer" Modify categorical data sort
2. Fusion table with API data table archive table
3 Fuse tables into the data structure of Archive Clean

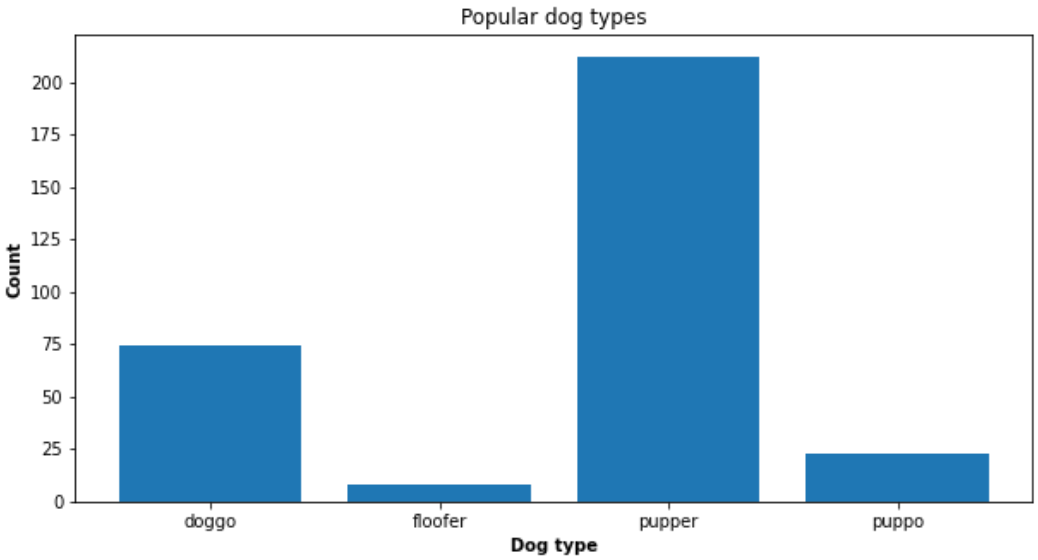
4. visualizations and analyses

Visualization 1



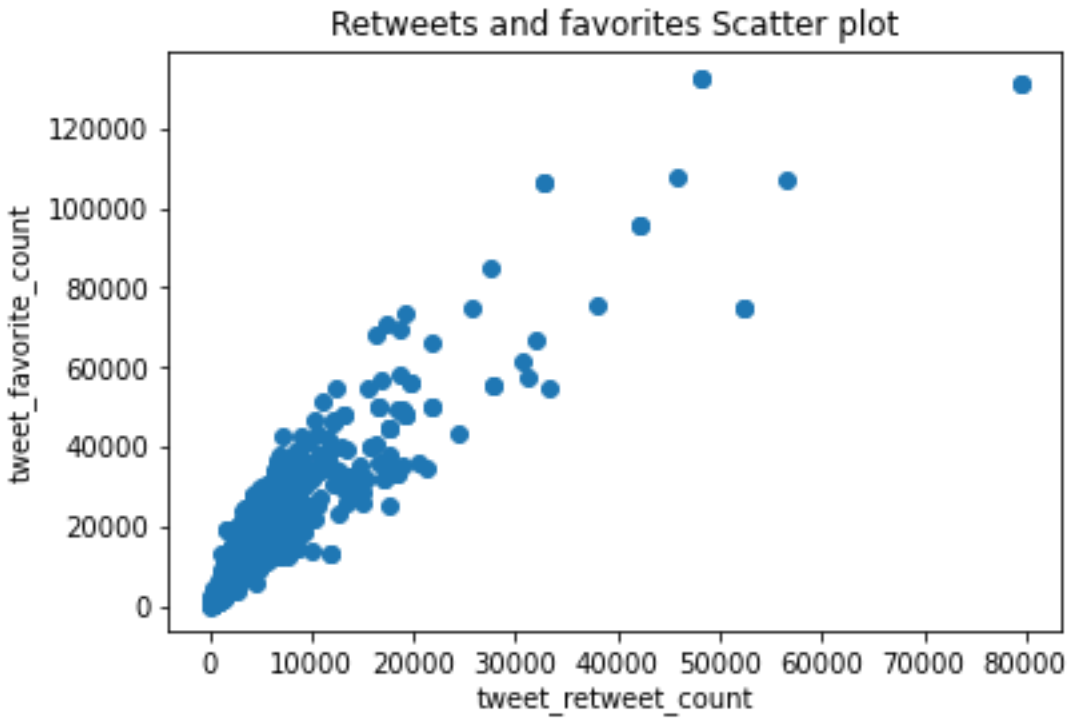
Analyses: this is the most frequent dog name(charlie)

Visualization 2



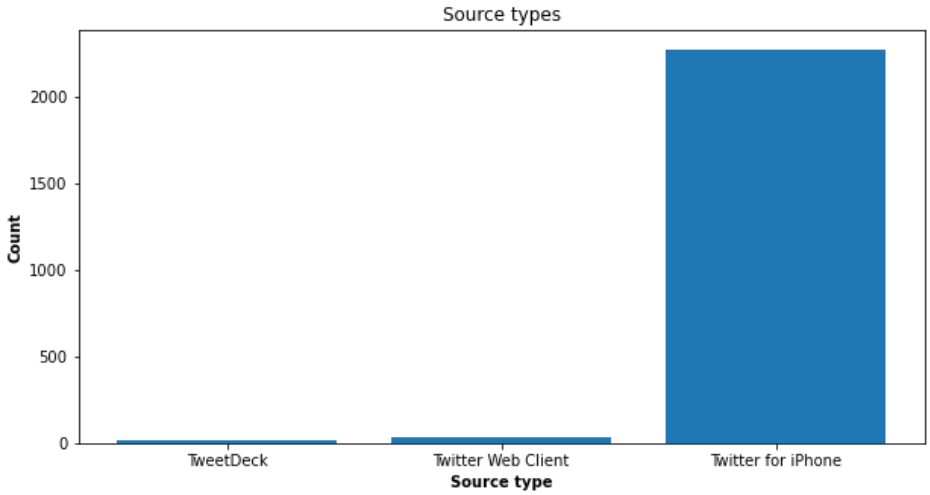
Analyses: this is the most frequent Dog stage of life (pupper)

Visualization 3



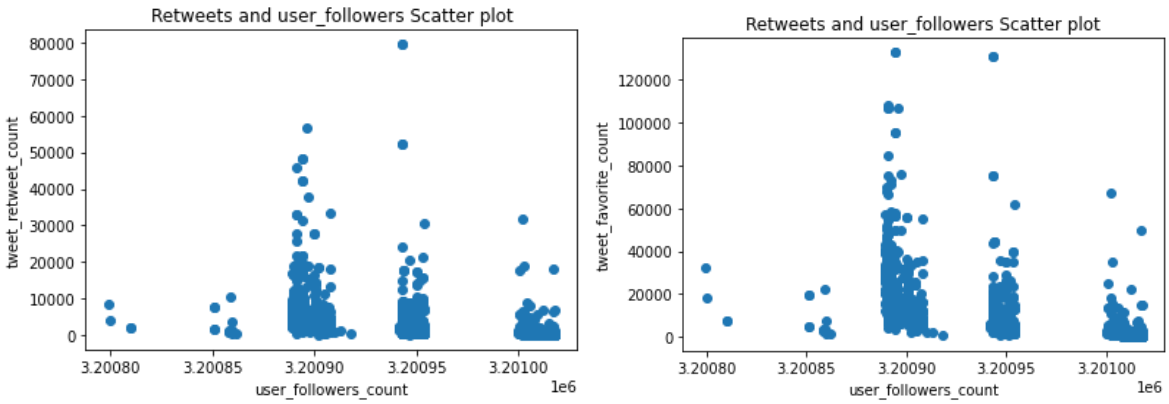
Analyses: there is a massive correlation between the tweet (retweets and favorites).

Visualization 4



Analyses: the most tweets are frequent tweets use (Twitter for iPhone)

Visualization 5



Analyses: the tweets(tweet_retweet_count,tweet_favorite_count) not effected by user_followers_count

5.Conclusion

Data wrangling consists of collecting data (file downloadable in resources): data cleaning storage, data analysis, and visualization of discordant data report in the following fields: data collection (downloadable file in resources).