https://www.youtube.com/watch?v=ea_5ERYSUF8WRANGLING REPORT
Ikuzwe Mbagatuzinde
16th September 2022

This is a report that shows my wrangling efforts in gathering, assessing, cleaning, storing and visualization on WeRateDogs, data from twitter API. This is a project from Udacity Data analyst ALX nano degree.

The Data Wrangling has four main processes:

- Gathering Data
- Assessing Data
- Cleaning Data
- Storing, Analyzing and Visualizing Data

Gathering Data

3 ways of gathering data were provided: twitter_archive_enhanced.csv, image_prediction.tsv, tweet.json, and twitter API but I was refused the twitter API and used the tweet.json.

- **twitter_archive_enhanced.csv**: A csv file directedly provided by Udacity.

  I used pd.read_csv to be able read and assess the file.

- **image_predictions.tsv**: Programmatically downloaded on Udacity server and stored it in a folder "image_predictions".

  Using python and os library I was able to programmatically download the file and save it in its folder.

- **tweets_json**: Provided by Udacity. I copied the tweet_api.py as requested but I didn't execute it.

  I just used pd.read_json to be able to read the json file.

Assessing Data

After gathering data, I was left with three data frames which I had to assess programmatically with the help of Jupyter Notebook, python and Pandas. With Google sheets I was able to assess visually.

While assessing data, I was able to find several issues which I am listing below:

Quality issues

Twitter archive table

1. some names are not relevant and should be replaced by NAN.
2. Timestamp should be timestamp datatype.
3. Remove retweets associated columns. There not original tweets and we won't be needing it.
4. The source should be simple source names instead of links in tags.
5. The numerator and Denominator should be correctly extracted.
6. There should be a column named rating.
7. The rating denominator should be float.

Image predictable table

1. wrong datatype on tweet_id column.
2. Remove duplicate jpg_url entries.

Tweet-json table

1. Columns can be renamed to make them more relevant.

Tidiness issues

1. Dog stages should be in one single column.
2. Remove image_num column because it is useless
3. All Datasets can be joined. ('twitter-archive', 'image-predictions', 'tweet-json').

Cleaning Data

After Assessing the data, just right before the cleaning process I had to make a copy of the original data as it is best practice.

Tweeter archive

1. Irrelevant names were replaced by NAN.
2. Timestamp column data type was corrected.
3. Some columns were dropped because there were not needed.
4. I changed the source which were like links into simple names.
5. The numerator and Denominator were extracted correctly.
6. I created the rating column to contain the rating information.

7. Rating column data type was changed into float.

Image predictions

8. tweet_id datatype was changed into string(object).
9. Duplicates were removed in image_prediction.

Tweet Json

10. Columns were named to have simple names.

Tidiness

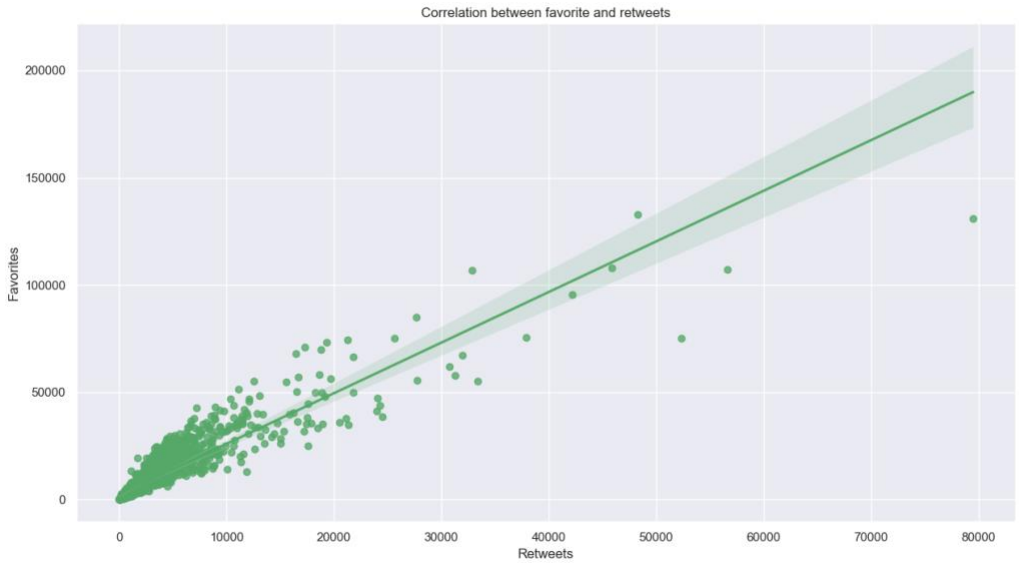
11. Dog stages were put into one single column under ‘stage’.
12. Remove image_num column because it is useless.
13. I joined all the cleaned data into one Data Frame.

Storing Data

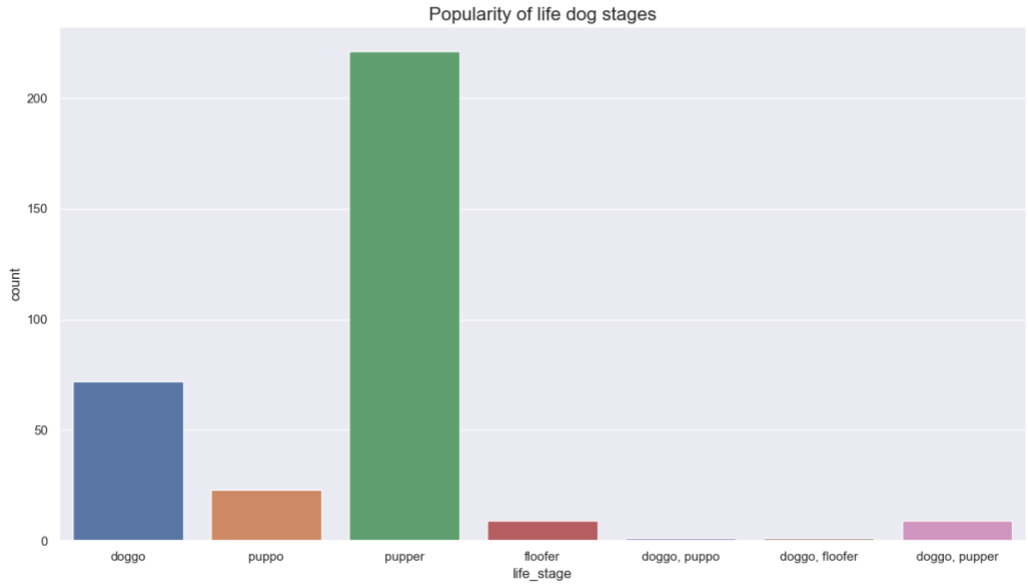
After cleaning the data I stored the clean Data frame in CSV file with name using *.to_csv('twitter_archive_master.csv')* .

INSIGHTS:

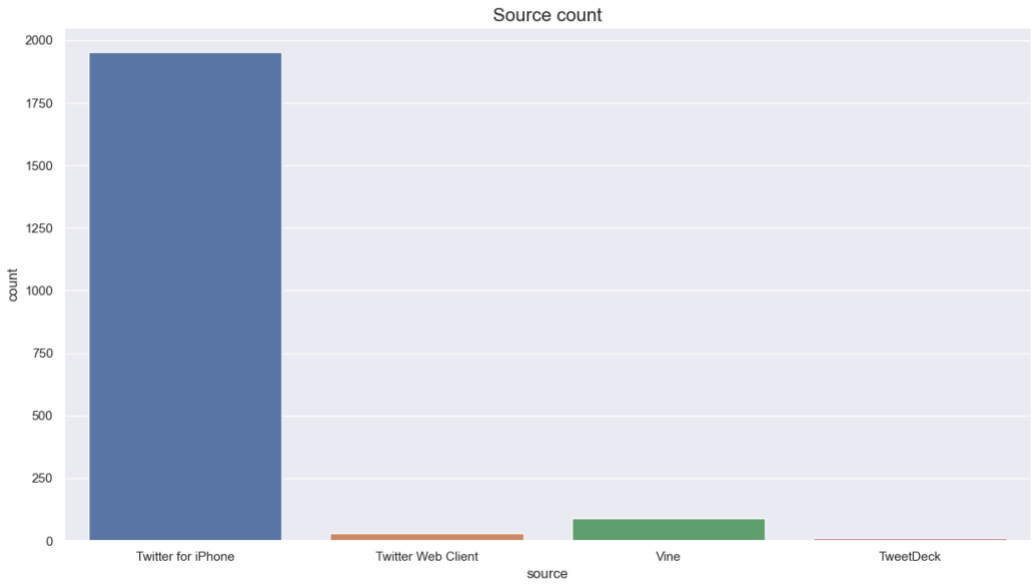
- I found positive correlation between favorites and retweets is positive which was expected.



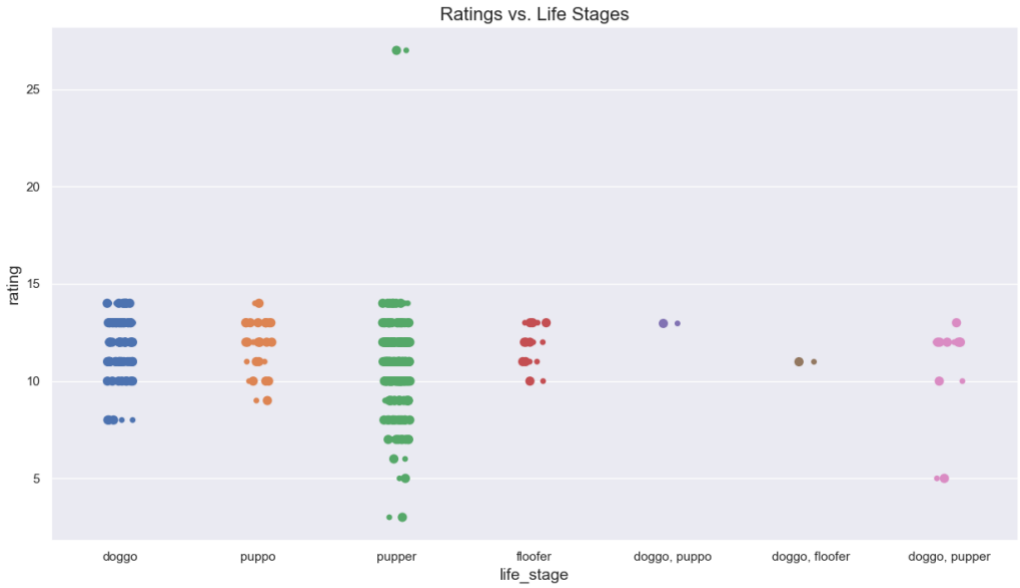
- I found that the most popular dog life stage is Pupper and doggo.



- This shows that most users used Twitter for iphone.



- This graph below shows that the rating is related to the popularity of the life_stage of the dog.



- This a world cloud that shows the most popular dogs. This world cloud shows us the most pupular dogs Which are Charlie, Lucy, Olivier, Penny, Tucker and Lola.