UDACITY

Data Wrangling Project

ALX-T Data Analyst Nanodegree Program

Elaborated by : Ferdawes Haouala

Title : Act Report for project 2

Date : 06/09/2022

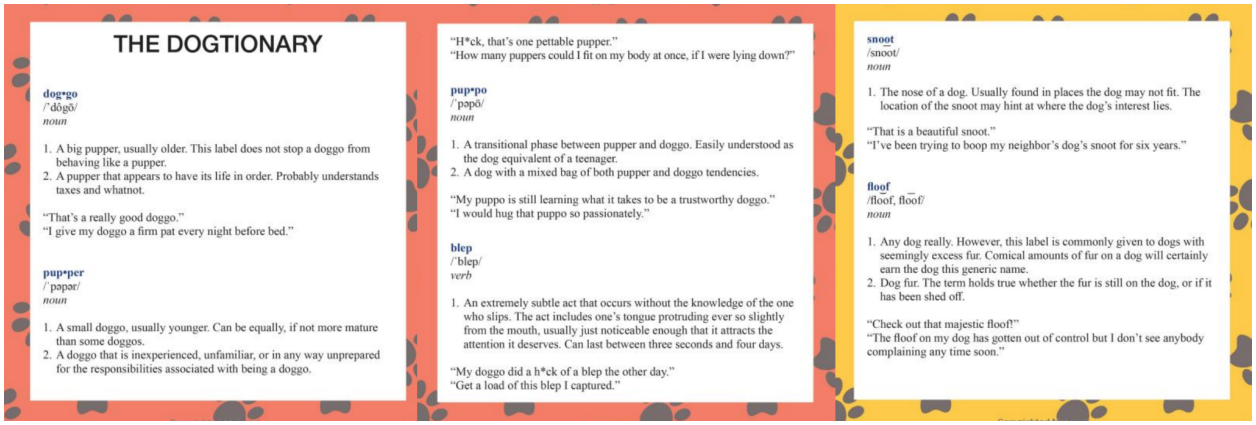
Table of content

1. Introduction : why and what is this project

- Through this project, I aim to wrangle data
- Data wrangling is the process of gathering data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it.
-
- The dataset that I will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as “WeRateDogs”. “WeRateDogs” is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, almost always greater than 10.

Name of the data set that will be used	Content
`image_predictions.TSV`	3 predictions along with their confidence interval and a boolean test
`twitter_archive_enhanced.csv`	Twitter data of the account “WeRateDog”
`tweet_json.txt`	Additional twitter data (retweet and favorite count)

- Additional resources :



- Names of tables:

Original name of the file	Name of the table in Jupyter notebook	Copy of the data set for cleaning
`image_predictions.TSV`	prediction	clean_prediction
`twitter_archive_enhanced.csv`	ratings	clean_ratings
`tweet_json.txt`	twitter_data	clean_twitter_data

New names of tables To solve tidiness issues :
Clean_twitter_data+clean_rating = twitter_data_dog table

clean_prediction+clean_rating = dog_table

2. Insights and visualizing

After cleaning the data set, I have :

Name of the data set	Columns
dog_table	['tweet_id', 'name', 'race_dog', 'jpg_url', 'img_num', 'rating_numerator', 'rating_denominator', 'p1', 'p1_conf', 'p1_dog', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog']
twitter_data_dog	['tweet_id', 'in_reply_to_status_id', 'in_reply_to_user_id', 'source', 'text', 'expanded_urls', 'date', 'time', 'retweet_count', 'favorite_count', 'retweeted']
`df`	dog_table+twitter_data_dog

2.1. Most dogs got 10/10 rating:

```
In [133]: dog_table.rating_numerator.value_counts()

Out[133]: 10      1573
          0        493
          20         2
          34         1
          70         1
          1770       1
          13         1
          11         1
          420        1
          9          1
          Name: rating_numerator, dtype: int64
```

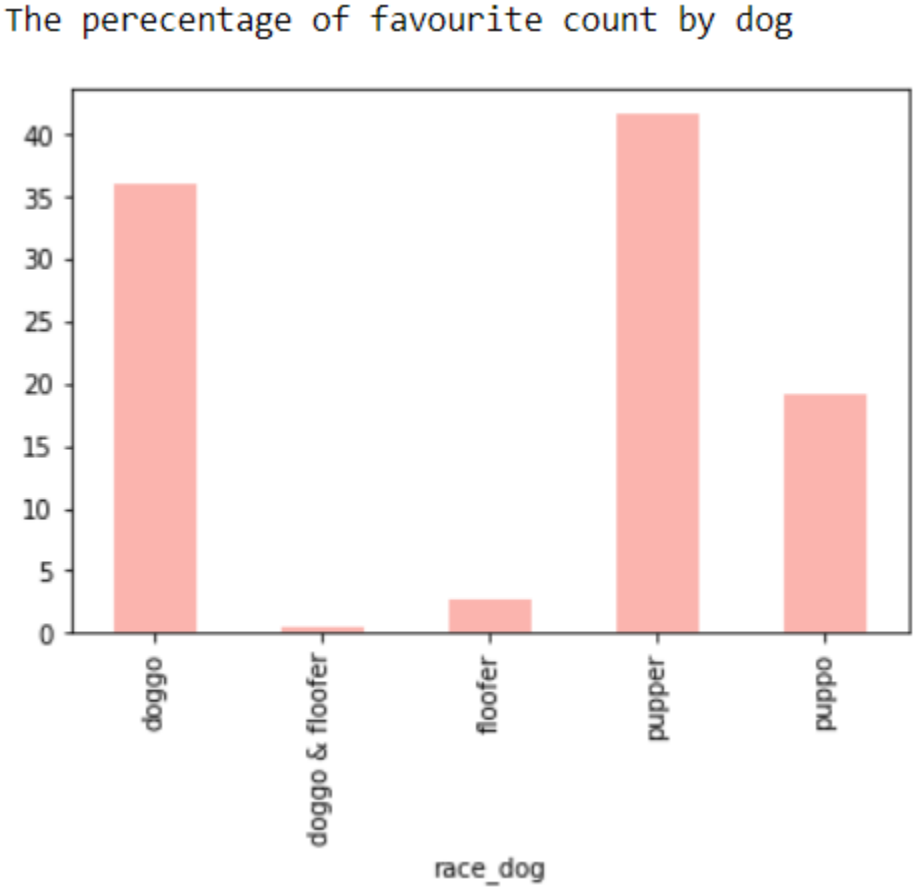
2.2. The number of tweets is highest in 2016:

```
In [137]: twitter_data_dog.year.value_counts()

Out[137]:  2016    1088
           2015     688
           2017     399
          -1        160
          Name: year, dtype: int64
```

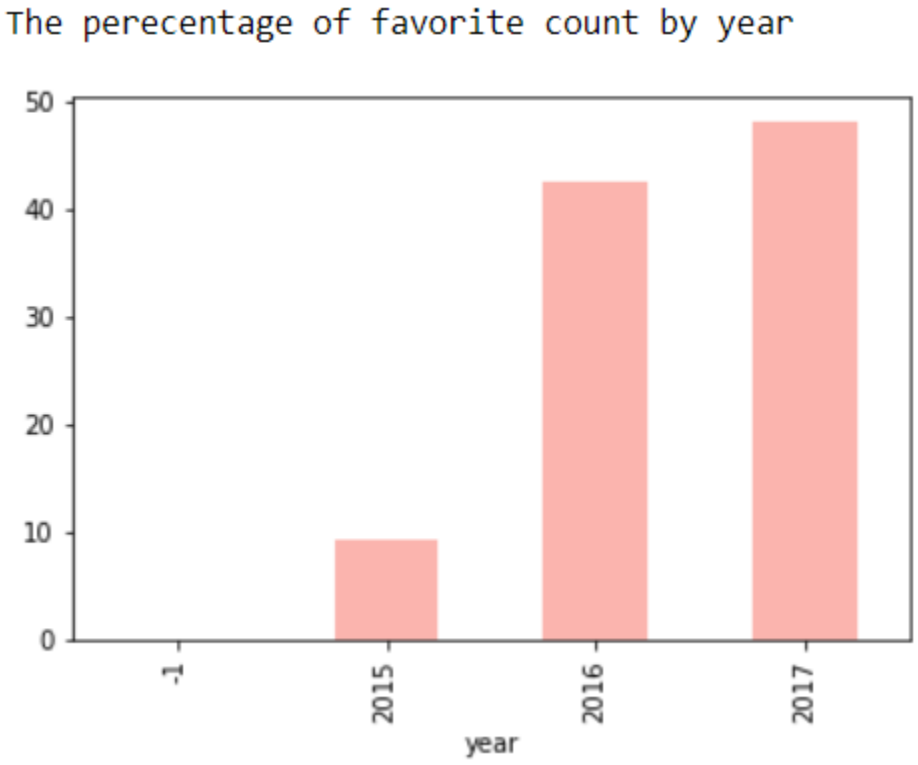
2.3. The highest type of dogs is Pupper:

```
In [138]: total_count = dog_table.race_dog.value_counts()
          total_count

Out[138]: pupper            203
          doggo              63
          puppo              32
          floofer             7
          doggo & floofer     1
          Name: race_dog, dtype: int64
```

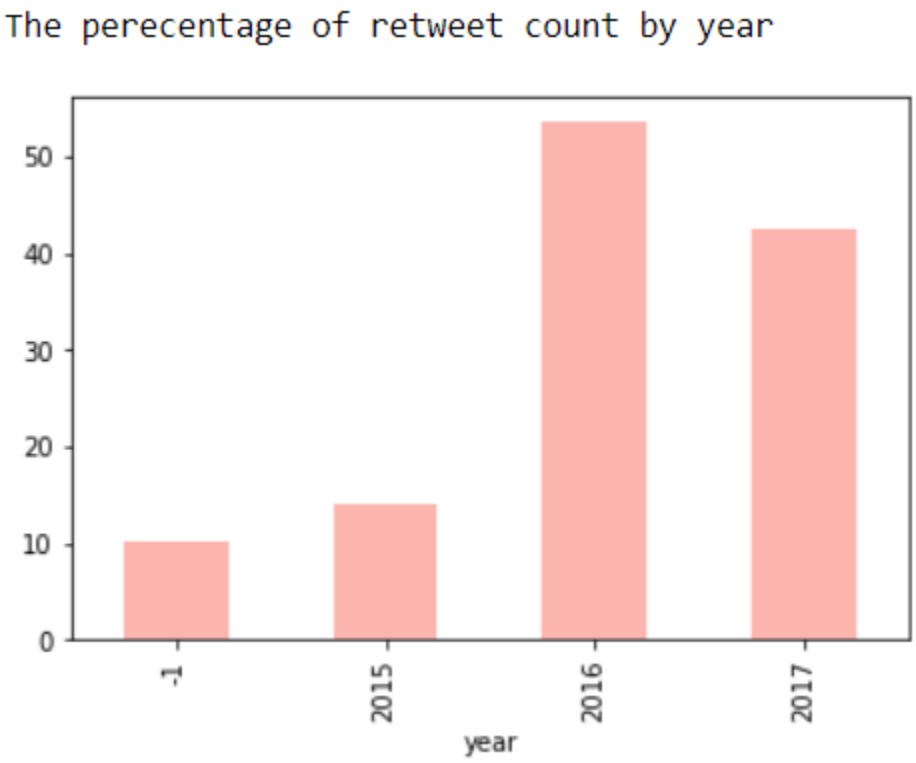
2.4. For each type of dog, how many favorite reactions were received:



2.5. For each year, how many favorite reactions were received:



2.6. For each year, how many retweets were received:



3. Insights :

- Fom the visualization 2.4 we can conclude that Pupper (smaller and not prepared for responsability)are the most dominant dogs and they are more prefered by users than other (higher favorite count of 1,254,680).

- Fom the visualization 2.5 we derive that even though "WeRateDogs" was most active at 2016 (corresponding to the highest number of tweets) and its activity was reduced since then but it started getting popular to reach its highest favorite count of 7,338,584 at 2017.
- Fom the visualization 2.6 we can confirm that the highest retweet rate was at 2016 with 2,154,752 (53.6% of total retweet count during the 3 years time) which is the year at which "WeRateDogs" was most active but not the year when it got the highest favorite count (even though the percentage of 2017 is colse to that of 2016 : 42.5%)

3. Conclusions and limitations

From the previous results, it may be that the high activety of that twitter account on 2016 made the retweet rate grow and that growth in retweet may

have contributed the growth of the favorite count to reach its highest value the following year

I couldn’t further investigate the most popular name of dog and other columns because they have missing values.