Data Wrangling : We Rate Dogs

This project was divided in 3 parts:

1. Data Gathering
2. Data Assessing
3. Data Cleaning
4. Data Analysing and Visualizing

I will describe shortly my work in each of this part.

1. Data Gathering

After loading librairies, I have to gather data from 3 differents sources :

- Twitter archive dataset, downloaded from a link provided by Udacity ('Archive dataframe'),
- Image predictions dataset, hosted on Udacity's servers ('Images predictions dataframe'),
- Others twitter data, obtained from the Twitter API ('API dataframe').

2. Data Assessing and Cleaning

I have looked for missing values, inconsistent formats, outliers values, etc on each dataset.
For finding them, I have observed data and try to obtain results by coding.
I have organized them into 2 categories :

Quality

- Archive dataframe

1. Data type: tweet_id is int64 → I have changed format.
2. Data type: timestamp is an object → I have changed format.
3. text: some dogs name are included in this cell → I have extracted dogs. names from cell in "text" column and add them in "name" column
4. Data consistency: some names are not consistent with dogs' names
   - 'little' words such as "an", "a", "the", etc.
   - lowercase

   → I have remove this word of the "name" column as there are not corresponding to names.
5. Inconsistent data : there are some retweets → I have excluded these rows.

- API dataframe

6. Column name: id corresponding to tweet_it (to be consistent across the multiple dataframes) → I have changed column name.
7. Data type: id is int64 → I have changed format.

- Images predictions dataframe

8. Data type: tweet_id is int64 → format has to be changed.

9. Wording consistency: lowercase and uppercase first letter in names in p1, p2, p3 → I have change names to be consistent.
10. Missing data: some images are not corresponding to dogs (p1_dog, p2_dog, p3_dog are 'FALSE') → I have excluded these rows.

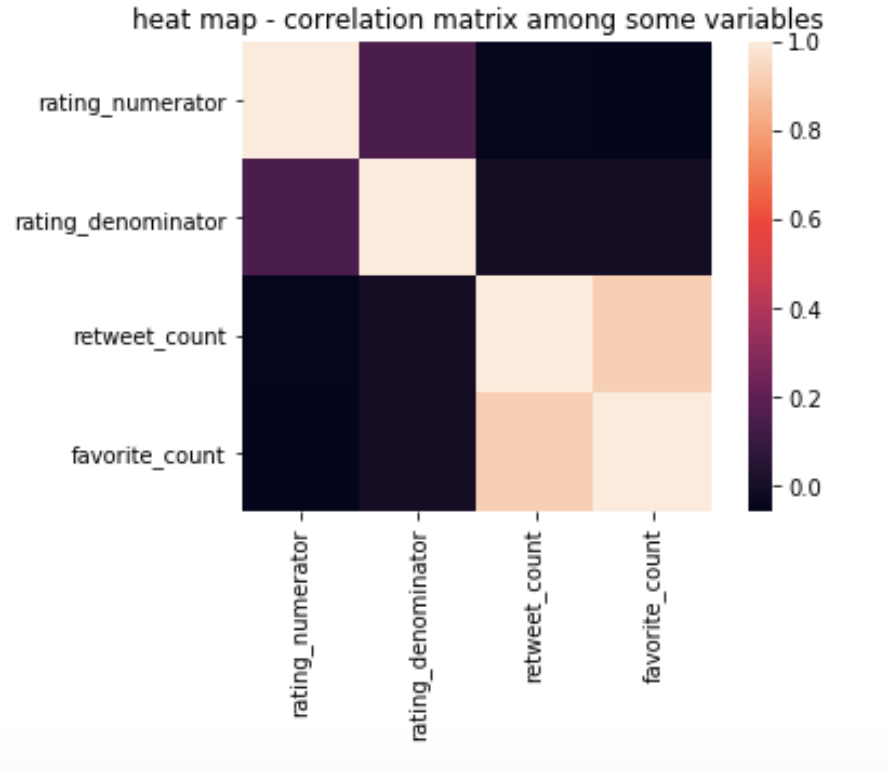
Tidiness

11. Create a single column instead of 4: doggo, floofer, pupper, puppo → I have add a column “stage” regrouping these 4 first columns.
12. Compute the 3 dataframes into one → I have created a new dataframe “df” which data from the 3 dataframes.

3. Data Analysing and Visualizing

After this, I have analyzed the data.
I have found 3 insights:

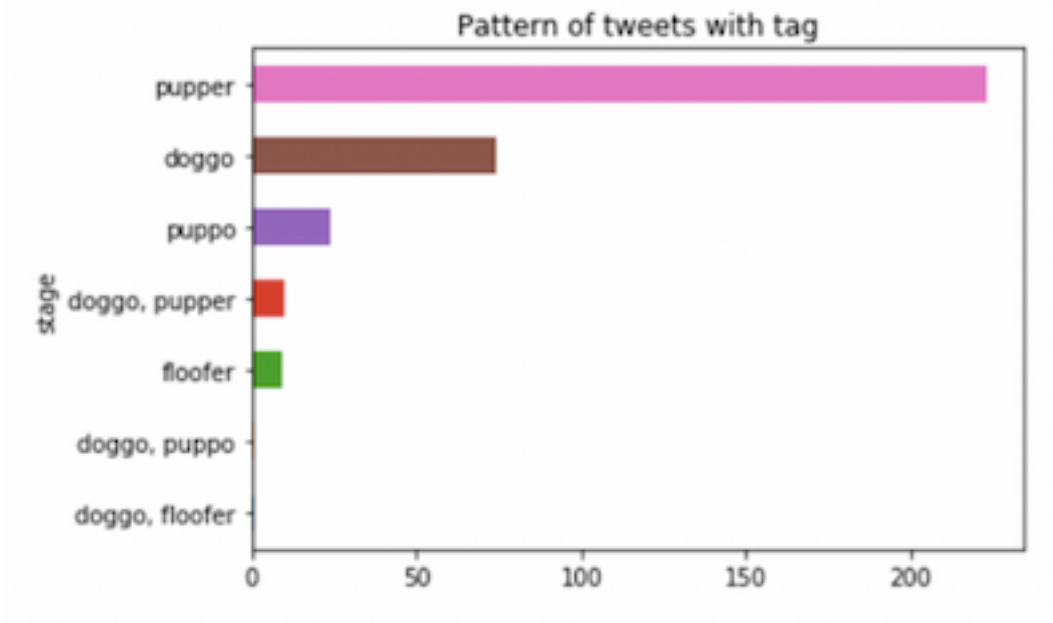
1. There is a high correlation between retweet_count and favorite_count.



2. Most popular names are Charlie and Lucy, followed by Oiver and Cooper.

```
Charlie     11
Lucy        11
Oliver      10
Cooper      10
Tucker       9
Penny        9
Winston      8
Lola         8
Daisy        7
```

3. Pupper are more tweeted than other dog stages.



I have visualized the number of retweets and favorites tweets by dogs stages.