Data Analist Nanodegree
Andrés Pozuelo

Project 4. Wrangle and Analyze Data

Data wrangling

- Gathering data

1 I have downloaded the file twitter_archive_enhanced.csv manually in workspaces.
2 Programmatic download using the library Request
3 Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting

- Assessing data

Evaluating data for this project
After collecting each of the above data, evaluate them visually and programmatically to detect problems of quality and order. Detect and document at least eight (8) quality problems and two (2) cleaning problems on your wrangle_act.ipynbupport Jupyter. To meet the specifications, problems that satisfy the motivation of the project must be evaluated (see the heading Key points on the previous page).
eight (8) quality issues
contributors, coordinates and geo They have 0 data. Those columns are eliminated as unnecessary.
created_at it does not have to be on datetime64. Create 3 columns with the year, month and day.
Line breaks in full_text. Remove.
Aislar el nombre del perro de full_text.
Extract the url from full_text in another column.
Extract the vote full_text in another column.
place has only one record. You can ignore and delete column.

Cleaning Data and tidiness issues

Quality problems / Order problems

Rename 'id' to match it with the other df's and change it to str
Eliminate retweets
Delete columns that we do not use
Create year, month and day
Eliminate page breaks in 'full_text'
change '_' for ' '
change dog name by its correct name or none
Separate the 'stage' of each dog in a new column
Create rating column concating rating_numerator with rating_denominator

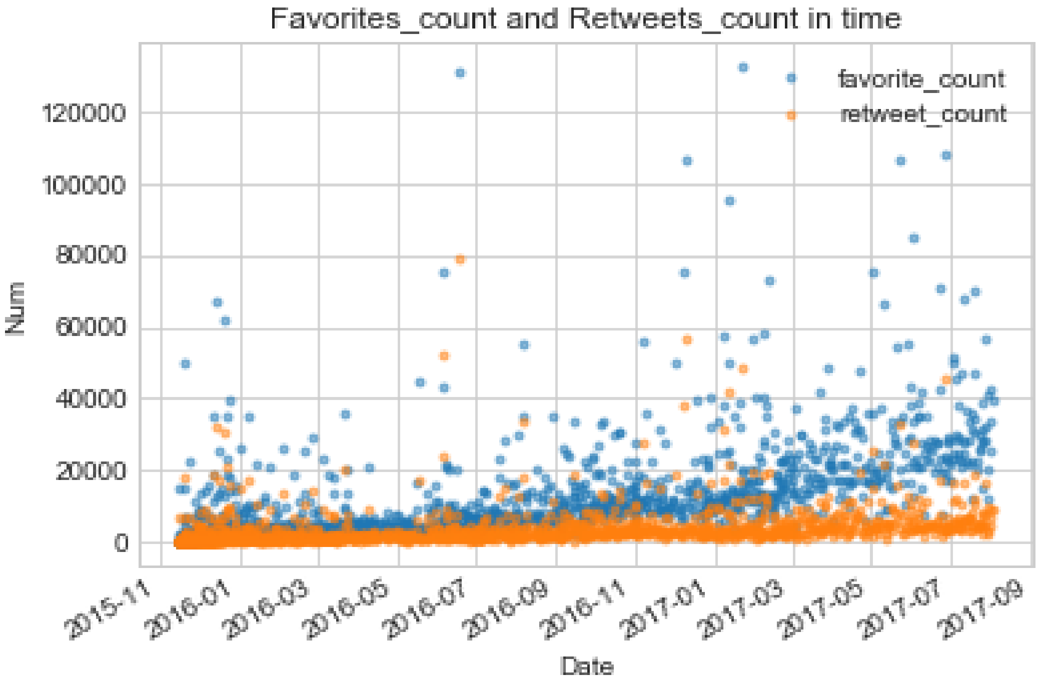
Storage, visualization and analisys of data for this project

As requested, I record in the file twitter_archive_master.csv the merged data of the 3 datasets, df_clean, df_tweet_clean and photo_dogs_clean. I also create a csv file for each of the aforementioned datasets.

Analisys

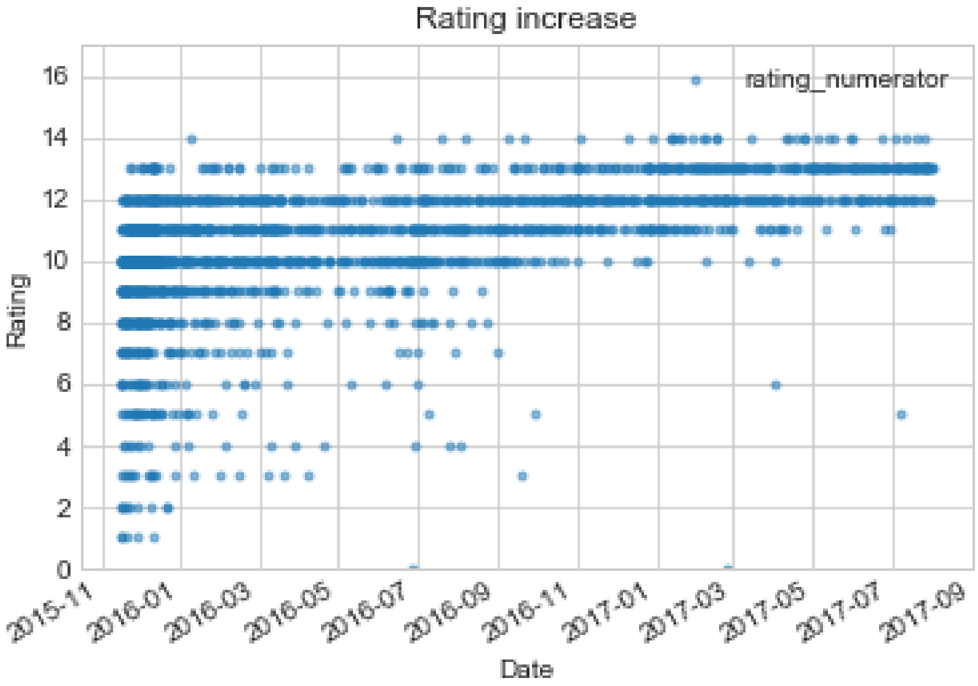
Analyze and visualize your unordered data in your wrangle_act.ipynb Jupyter notebook. At least three (3) ideas and one (1) visualization must be produced.

Favorites_count and Retweets_count in time.¶



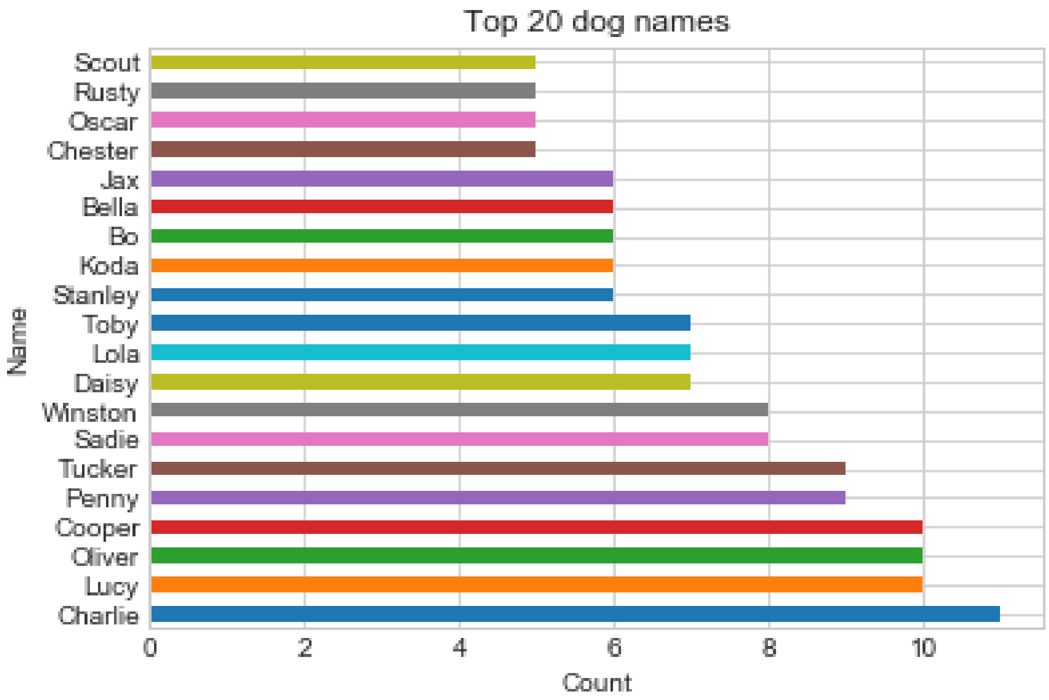
The number of tweets and retweets increases as we get closer to today. It shows that there are more and more users and they are more active.

Rating

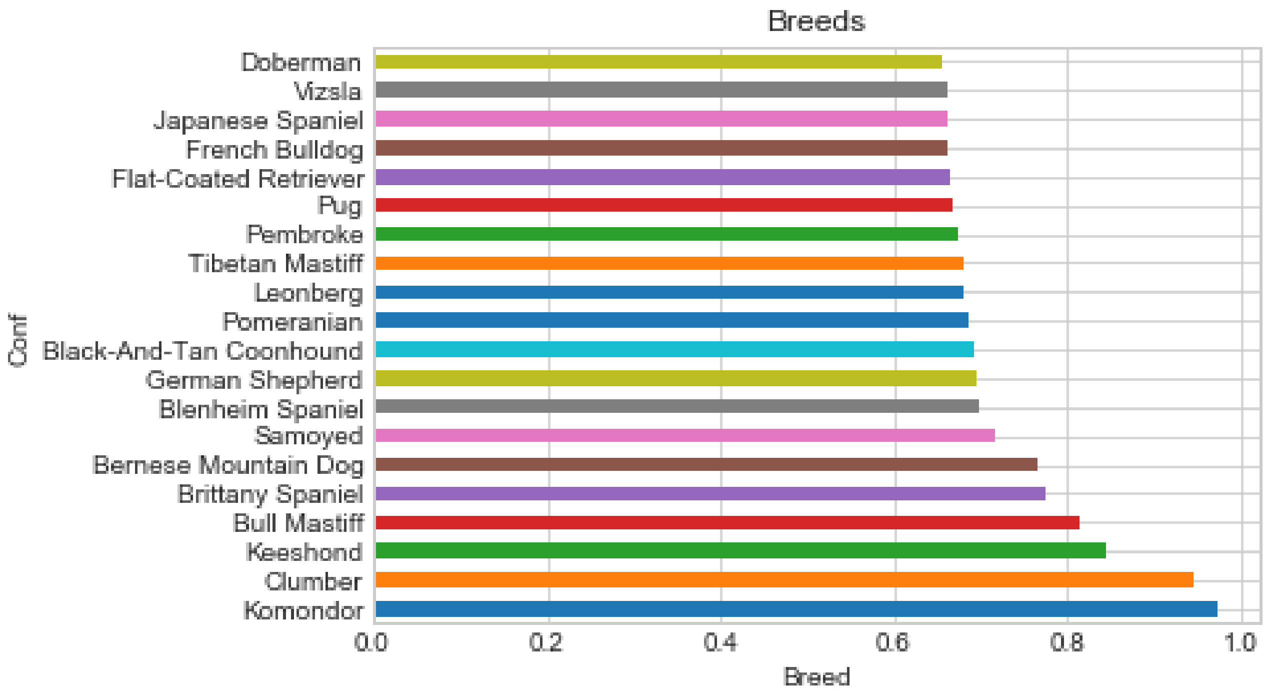


We also verify that the qualifications are increasing.

Top 20 dog names



Top 20 Breeds



In this graph you can see the top 20 breeds with the highest score

Not the most punctuated ones are the most tweeted. Possibly the most "nice" are more successful.