Wrangle and analyze twitter data

By Hana AL-alawi
25 February 2019

Twitter is social networking on which users post and interact with messages known as "tweets". Twitter has a feature which is a twitter developer account. It is a self-serve tool that developer can manage their API access and Twitter apps to use Twitter data to explore and learn a lot about people, countries, foods, and dog too. Today, many people using Twitter to share news or write about their life and days.

In this project, I'll use my twitter developer account to gathering data about a @weRateDogs account to create interesting and trustworthy analyses and visualizations by using data wrangling steps.
To wrangle data there are three steps should work with: **Gathering data, assessing data and Cleaning data.**

Gathering data is one of the most important skills that any data analyst and data scientist should learn and working with. In this project I gathered data from three different sources the first data frame gathered by reading 'Enhanced Twitter Archive' is CSV file which contains basic tweet data for all 5000+ of their tweets, The second data frame gathered from TSV file 'Image Predictions File' by reading this file from URL, and The last data frame gathered from Twitter.
All these data frames are about dogs, this is a twitter account (@weRateDogs) working to rate dogs by a special rating system to rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10.

Assessing data, in this step I worked to investigate and explore the data frame to find issues. These data frames contain many issues some of them are Quality issues which are accuracy, validity, consistency, completeness and other is Tidiness issues which is structural issues.

Quality issues:

- Timestamp in twitter_archive_df is an object, change it to datetime type.
- Text (tweets) contains #, extract the hashtags from the text column and store it in the tweet_hashtag column.
- Name have some names with (a, an, not..) which is not names.
- Calculate denominator and nominator to rating_number.
- Delete the columns that I don't use it.
- Sperate timestamp into three columns, (day, month, year).
- Remove any denominator does not equal 10 (1,50,80,20,2,16,40,70,15,90,110,120, 130,150,7,0).
- Remove outliers from nominator column.
- Calculate the rating in a new column and store the percentage value.
- Remove all retweets from dog_images_df and keep tweets.
- Some tweets without images because the number of rows in twitter_archive_df and dog_images_df does not match.
- None values to NAN in Name, doggo, fFloofer, pupper, and Puppo.
- Combine Puppo, Floofer, doggo, and Pupper into one column 'dog_stages'

- Remove duplicated values in twitter_archive_df.

Tidiness Issues:

- Merge all data frames together
- jpg_url contains duplicated URLs.
- There are four different columns of dog stages (Doggo, flooder, popper, and puppy) merge it into one column and remove nulls.

After Assessing data, should **cleaning data** frames to analyze it. In this step, I solve the issues above
And store three data frames into one to CSV file, and then create a visualization.