Project Title: Analyzing the tweet archive of Twitter user @dog_rates

Date: Dec. 2018

Data time period: from 2015-11-15 22:32:08 to 2017-08-01 00:17:27

All the project had be done in 2 steps: wrangling and analyzing and visualization.

1. Data Wrangling

1.1. Data Gathering:

Data from this project, collected from 3 different sources:

Data Source Name	Location, Gathering method	Description and Note
`tweeter-archive-enhanced.csv`	Project page on class, download manually	
`image_predictions.tsv`	hosted on Udacity's servers ,downloaded programmatically using the `Requests` library from the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv	This file is top 3 result prediction of dog breed for each tweet based on a neural network algorithm. The file
`df_tweet_scrap.csv`	Scrap the twitter account, using tweepy API	There are some missing complementary data from "WeRateDogs" account which help for better analysis result. API, because of error 88: 'Rate limit exceeded',run 3 time to download complete, save each tweet as JASON in txt file and then return txt file to csv

1.2. Data Assessing:

1.2.1. Quality:

`df_tweet_scrap` dataframe: (most validity issues)	• most of columns not valid to use for analysis because simply, they not data or they are data but have not any role in analysis. We are going to save 3 columns:
	  • 'id' to use as key for joining data sets
	  • 'favotire_count' and 'retweet_count' inorder to use in analysis
	In file named: `df_clean_scrap.csv`

`twitter-archive-enhanced` dataframe:	• consolidate dogs stages in one column (consistency issue) • `rating_denominator` some of denominator are bigger than 10, so accuracy, validity and consistency are problem of this column • `rating_nummerator` some numbers are really huge like 150, re-extract rating from original text and then drop wrong numbers • removing tweets rows which they are reply to followers, (validity issue) because we want to analysis just original tweets • dropping unnecessary columns for further analysis and saving clean data frame in file named: `df_clean_enh.csv`
`image_predictions` dataframe:	• removing rows which all 3 prediction results are not a breed of dog • for remaining rows with at least one True on `p*_dog`,for each tweet, prediction result with bigger confident in `p*_conf` has been selected as breed of dog's picture of that tweet • aggregate all predications in one column: `breed`

	• dropping none necessary columns • dropping unnecessary columns and saving the rest in file named: `image_clean.csv`

1.2.2. Tidiness:

joint up `df_tweet_scrap` with `twitter-archive-enhanced` and making `df_clean_twitter.csv` file
joint up `df_clean_twitter.csv` with `image_clean.csv` and creating `twitter_archive_master.csv`
in `df_tweet_scrap`: change type of 'favorite_count' and 'retweet_count' from string to integer
in `twitter-archive-enhanced`: change timestamp format to date time
in `twitter-archive-enhanced`: mark as `multi` name for those tweet IDs with more than one stage
in `image_clean` dataframe replacing `'_'` with a space and then capitalized all first letter

1.3. Data Cleaning:

Cleaning process, based on challenges what described in data assessing, split to 3 steps of define, code and test. During the process, sometimes defining and testing for both quality and tidy of data happen together. The final result was a clean data frame, named: `twitter_archive_master.csv`

2. Data Analyzing and Visualization:

For digging inside of `twitter_archive_master.csv`, finding the answers for these questions has been followed:

1. Based on time period which data has been collected from WeRateDogs twitter account, what is the relation between favorite and retweet counts?
2. What time during a day, tweets got more seen by followers based on more retweet and favorite counts?
3. Which `stage` was more attractive for follower based on `retweet_count` and `favorite_count`?
4. Which breed get more retweet and favorite?
5. Does image prediction algorithm, work out with more confident with more image?
6. Do those tweets with higher `rating_denominator` (10x) get more `favorite_count` and `retweet_count` than other with standard `rating_denominator` (10)?

3. Project Limitation:

1. Wrangling step was very time consuming and take more than 90% of project time, so time remaining for analyzing and visualization was really tight
2. Data source lonely relying on twitter account and this cause we tolerate bias in our findings. For example, we missed influence of foreign triggers on tweets like and retweets