Data Wrangling, Analysis and Visualisazion using Twitter API for WeRateDogs Dataset - Wrangling Report

By Sumukha K

Wrangling is about gathering the data, accessing it and cleaning the unwanted data also fixing some errors in data. In this WeRateDogs Dataset i have done some wrangling as a part of the Udacity curriculum. WeRateDogs twitter account posts pictures and video's of dogs. It also give ratings and also describe dogs in its own fashion.

Step 1: Gathering the data

In this step, i have gathered data from Udacity "twitter_archive_enhanced.csv" through manual download and 'image_predictions.tsv' through programmatically. Also have gathered data from twitter api. Gathering data from various sources were challenging especially using twitter as i was not exposed to it. To gather data from twitter api, i had to login to my twitter account and had to create an app to get the credentials. Once credentials were acquired, code had to be written to gather data of twitter id's that were present "twitter_archive_enhanced.csv".

Step 2: Accessing data

In this step, Accessing data was made both visually using spreadsheets and programatically. This step is all about finding errors in dataset and noting down to correct it in the next steps. The issues found in the datasets include

Quality

1. t_archive dataframe tweet_id, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, in_reply_to_user_id should be an object but its shown as integer, float, float, float, float respectively
2. t_image dataframe tweet_id should be an object but its shown as integer
3. t_df retweet_count and favorite_count should be an integer, its shown to be an object
4. In t_archive, None is treated to non-null values in column names names,doggo,floofer,pupper,puppo, to change none to NaN value
5. Column headers are not discriptive
6. Ratings given in t_archive dataframe (rating_numerator and rating_denominator doesn't mathamatically make sense. because cannot be bigger that denominator. Also if i thought of considering only numerator, there are many numbers that are unexpected like 420, 1776,960, 666 etc.,
7. To remove all rows with non_null retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp in t_archive dataframe
8. The timestamp datatype is referred as object which should have been datetime in t_archive dataframe
9. In name column of t_archive dataframe, some names start with lowercase letters which doesnot look to be names of person (some examples include my, one, his, him etc) , so converting those names to NaN values

10. There are some mistakes rating_numerator column of t_archive dataframe.

Tideness

1. t_image and t_df dataframe should have been murged in t_archive
2. columns such as doggo,floofer,pupper,puppo could have been in 1 column instead of 4

Step 3: Cleaning

In this step, the problems encountered in the access step is fixed one by one. The step included defining the problem, coding and testing it to know if the problem was fixed.

Step 4: Analysis and Visualisation

Once the problems encountered were fixed, i did some basic analysis like Most favorite dog, most retweeted dog and less retweeted dog, Also i did visualise the to find patterns with retweets and favorite verses timestamp.