Wrangle and Analyze Data

Data wrangling report

By Ahmed Essam

Data gathering

I gathered the data from three different sources. The dataset that I gathered is the tweet archive of Twitter user @dog_rates.

First Source(the local file**):**

twitter_archive_enhanced.csv

The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets but here we have filtered around 2000+ tweets with ratings.

Second Source –URL Image predictions

The tweet image predictions. What breed of dog is present in each tweet according to a neural network. This file is hosted on Udacity's servers and we will be downloading it programmatically using the Requests library and the given URL.

Third Source - Twitter API

I used file in the class room.

Data assessing

Throughout assessing the data both visually and programmatically I ended up with the following observations.

Quality Issues:

1. Many rows in the twitter enhanced dataset did not mention the stage of dog that is all the four stages in many rows are None.
2. There are 1976 rows with no definition of the dog's stage**.**
3. DataType of columns in the twitter enhanced dataset such as 'timestamp','retweeted_status_timestamp' are defined as String whereas it should be datetime.
4. There are missing expanded urls in the twitter enhanced dataset.
5. There are 181 retweeted_status_id which means that our dataset contains retweets as well.
6. We do not need retweets in our dataset for analysis so I removed retweet_user_id and other columns related to retweets.
7. Some of the names are 'a', 'an', 'the' which are not invalid.
8. The common numerator ratings given by @weratedogs are 11,12,13,16 so on. But, here we find that most of the ratings are too high such as 1776,960,666 etc.
9. We know that @WeRateDogs keep their denominator as 10 always while rating dogs but here some of the ratings are 11,50,2,7,0,110 etc.
10. Some of the names of dog breed are not defined, like “bookshop”, “bakery”, “book_jacket” and “orange”.
11. The Image Urls are same for some images.
12. The names of dog in Image prediction Dataset are separated by underscore instead of space.

Tidiness Issues:

1. There are four columns namely doggo, floofer, puppo and pupper for the stages of a particular dog. We don't need four columns for the stage, only one column will be enough.
2. We only need one master dataset for our analysis and visualizations, so we will merge all the three datasets collected from different sources.

Data Cleaning

To start the process of cleaning I made copies of the original datasets. I First defined the process to clean the data, then I wrote it in code and eventually tested it.

Steps made to clean the data are:

1. Removed the rows with null retweeted_status_id
2. Dropped the columns I won’t be using
3. Selected the main four colmns and created new dataframe
4. Add a new column 'Stage' to the new dataframe.
5. Add the new column 'Stage' to our original dataset.
6. Drop the four columns 'Doggo', 'Floofer', 'Pupper', 'Puppo' from original dataset.
7. Changed the datatype of the timestamp column to datetime.
8. Turned each invalid name to none.
9. Set the numerator rating in terms of denominator as most of the times denominator is 10 and then remove the denominator column with ratings not equal to 10.
10. Merging all the datasets using join and make tweet_id as main key as it unique for everyone.
11. Merged the three datasets into twitter_archive_master file.