Udacity Project # 2
Data Wrangling: We Rate Dogs
By: Maryam Mohammed

Project Overview

Welcome to my second project with udacity in the Data Analysis Nanodegree. In this project, I applied what I have learnt throughout the course. I gathered the needed data. I assessed the data and spotted some issues that I have worked on resolving them later in the cleaning process. My project consists of 5 parts. Gathering, Assessing, Cleaning, Storing, Visualizations and Conclusions.

The project's objective is to provide a standard and clean reference for the tweets about the dogs ratings for the twitter's account We Rate Dogs. Therefore, conclusions and insights can be drawn easily. This twitter account rates people's dogs and the ratings have a denominator of 10. However, you will find in the numerator numbers above 10 as I allowed this up to 15 considering people think that these dogs are very good.

This project wants to answer the following questions:

- Are the dogs who receive high ratings also favorited by a high number of other users?
- What is the most common dog stage in its life cycle that has the most attention from the users?
- What does the distribution look like for the dog ratings?
- How are favorites and retweets correlated together?

The following are steps to what I have done throughout the project:

Part one

1. I imported all needed modules and libraries needed for the project.
2. Downloading the tweets archive csv file manually and adding the data to a new dataframe.
3. Downloading the image predictions file programmatically and adding data to a new data frame
4. Downloading the json txt file manually (instead of getting data from the twitter API), adding data into a list of dictionaries, then finally constructing a df from this list.

Part two

1. I started assessing data visually by reviewing the data frames.
2. I checked them using info, describe, sample, duplicated, value counts, isnull and shape methods.
3. I spotted the following issues:

   Quality issues:
   • api_df id column rename to tweet_id for consistency across dfs
   • Having missing values in the dog types columns.
   • Retweets and replies should be removed to eliminate duplication.
   • remove invalid values in the name column in twitter archive df and replace None with np.nan
   • for dog ratings numerator there are zeros (twitter archive)
   • for dog ratings numerator there are values above 10 and typos (twitter archive)
   • for dog ratings denominator there are typos, zeros and extreme values (twitter archive)
   • timestamp column type is not datetime

   Tidiness issues:
   • Dog types can be merged into one column (twitter archive)
   • api_df can be merged with the archive df
   • Not all the columns are needed for our analysis
   • Text column in the twitter archive df has rankings which we can extract.

Part three

1. I started the cleaning process and tackled every issue mentioned in part two.
2. I created copies of each data frame at first.
3. After taking care of the issues I merged the data frames together in one df.

Part four

1. I stored the combined data frame into a file.

Part five

1. I started creating visualizations to answer the questions mentioned above.
2. Conclusions were also drawn after each visualization.