Wrangle Report

Project #4 Data Wrangling

(WeRateDogs.)

Introduction:

In this project I wrangled and visualize/analyze the tweet archive of @dog_rates, also known as WeRateDogs. We can get helpful and useful insights based on the information of archive tweets, that is will be done through different steps starting with gathering and extracting data ending up with the analysis and visualization.

In this report, I will illustrate the process I went through in order to extract knowledge and insights.

Steps:

1. Data gathering

In this step I gathered and extracted data from different sources which are: tweets archive of @dog_rates and other .csv file, all given by Udacity, also data has been gathered through twitter API and a jSON file.

2. Data Assessing

In this step, I solved some problems occurs from gathering data from different sources, which are:

Quality problems:

Twitter archive data frame :

1. Change datatype of 'tweet_id' , 'in_reply_to_status_id' , 'in_reply_to_user_id' , 'retweeted_status_id', 'retweeted_status_user_id' to String.

2. Change 'timestamp' datatype to datetime.

3. Create 'rating' column that contains numerator rating divided by denominator rating.

4. Drop unimportant columns : expanded_url_denominator rating , numerator, name, doggo , floofer , pupper , and puppo

5. Simplify the url in 'source' column to make it more neat/clear and readable.

6. Replace 'None' with 'NaN'.

7. Deal with incorrect dogs name like: "a", "an", "such","the"..

6. Remove retweeted tweets to identify only original tweets.

Image prediction data frame :

1. Change 'tweet_id' datatype to String

Tweets data frame :

1. Change 'id' datatype to String

Tidiness:

Twitter archive data frame :

1. Merge the following columns: "doggo , floofer , pupper , and puppo" into one column "stage"
2. Merge all data frames together since their information are related.

3. Data Cleaning

After assessing the data there was multiple things need to be cleaned based on tidiness and quality issues, so I divided each data frame needed to be cleaned into two parts.

1. **Code**: shows the code of cleaning.
2. **Test:** shows and ensure the result of written code

In twitter archive data frame, first have copied it into new data frame so that I can make the cleaning process for the new version, then I have changed datatypes into suitable ones, then I merged doggo , floofer, pupper and puppo columns into one column named 'stage' and I dropped unnecessarily columns ('doggo', 'floofer', 'pupper', 'puppo' , 'expanded_urls' , 'name','rating_numerator','rating_denominator','retweeted_status_user_id','retweeted_status_id '), after that, I created new 'rating' column that equals to numerator rating divided by dominator rating, then I simplified 'source' column records/URLs into neat and clear version that does not contain an HTML code, and I have removed retweeted tweets from ‘text’ column to identify only original tweets. Also I have replaced incorrect dog names such as ‘very, the’..etc into Nan. Finally, I dropped nulls and replace each None record with NaN.

In image prediction data frame, first I will copy it into new variable, then I will change the datatype of ‘tweet_id’ into string to unify the id type for all data frames.

In tweets data frame, first I repeated some steps in image predictions data frame; first I copied the data frame into new version then I have changed the ‘tweet_id’ type into string.

Finally, I have merged the three data frames, first I chose the important columns then I put the primary key of the three data frames which is 'tweet_id' because it allows me to merge correctly after that I stored the merged data frames into cdv file: 'twitter_archive_master.csv'.