Wrangle Report

WeRateDogs Twitter archive

Nisrein Sada Dec,21th, 2018

In this report I looked into WeRateDogs Twitter Archive which is a twitter account were people rate people's dogs. Data wrangling consist of 3 steps: Gathering, Assessing and cleaning.

Gathering

The data was gathered from 3 sources. The first source was the twitter_archive_enhanced.csv which was a given file. This file contained the tweets that contained ratings and the following tweet information: tweet text, tweet id, rating numerator, rating denominator, dog stage(doggo, pupper, floofer and puppo). The second source was the tweets image predication which was downloaded programmatically using the requests library. This file had the the breed predication of each dog image. The third source was the Twitter API(Tweepy) to extract the missing information about tweets such as the retweet count and the favorites count.

Assessing

In this part the data was assessed to look into possible issues programmatically and visually. I looked into quality and tidiness issue. Quality issues are related to missing or inconsistent data while tidiness issue is the structure of the data.

First I looked into the twitter_archive_enhanced.csv. The Quality issues found include: wrong names of the dogs that is present such as a,an, such, the,.., erroneous datatypes for dog stage that should be a category and timestamp that should be a date type instead of the string type for both, Source of the tweets was encapsulated in a tag which made it more difficult to read and the data contained retweets that needs to be removed for the analysis.

The twitter image predication p1, p2, p3 had erroneous datatype of a string instead of a category. And extra tweet ids from the first dataframe.

The last source had extra tweets that were retweets which are not of interest for the current analysis. And there were missing tweets that are no longer available.

The tidiness issue were the multiple columns for dogs stage could be merged in one column dog stage, the last issue is to merge all the dataset in one dataframe and multiple variable information contained in the same column like the text column which is contained in the rating, dog stages and name columns.

Cleaning

how the above issues were tackled:

- Existing retweets in the dataset: removal of all the rows that had column values for the following: in_reply_to_user_id, retweeted_status_id and retweeted_user_id then this means we have replys to tweets or retweets so we need to remove these.

- Incorrect or missing dog names issue: read the tweet text and extract the dog name that followed “name is” 0r “named “. Also some tweets had two types of dog stages which was tackled by creating multiple rows for the same tweet ID.

- Source column difficult to read: extract the content of the a tag and use it instead of the full tag.

- Missing values for dog stages: this issue arises from the existence of some plural form of the dog stages in the tweet text. So look for the plural form and add the dog stage to a new column instead of the current existing form columns. And then drop the 4 columns the previous columns for dog stages to solve the tidiness issue.

- Erroneous datatypes: change the columns data type to the appropriate type.

- Missing tweets IDs: This was solved by merging the dataframes together so only the tweet ids present in the main source(twitter archive) while removing anything else.