Project: Wrangle and Analyze Data

Data Wrangling Report

By Ahmed Ebaid

This project for the data analyst nanodegree and the aim of it is to gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it.

The dataset used in this project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

In the First part, data was gathered using three different sources.

The first source is downloaded manually using the link provided.

The second source is downloaded programmatically using the Requests library.

The third one is gathered using Twitter API and Tweepy library . This file was stored in JSON format.

In doing assessment both visual and programmatic assessment was used. In visual assessment, google spreadsheet was to scroll and check the validity of the data. In programmatic, diffrent functions was used to assess the data and check missing values. Several quality and tideness issues were found. listed as follows:

Quality

1-drop rows that are retweets and replays using is.null

2-Select the columns related to retweets and drop them as it is of no use further.

3-Select the column 'timestamp' and change the DataType of timestamp from string to datetime.

4-Select rows with missing values of expand urls and remove them.

5-convert 'none' in (name) column to NAN in df_archive DataFrame.

6-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.

7- Select the source column and extract the text between anchor tags.

8-Select the columns for which dog breed classifier is true and remove the images which are not dogs.

9-Select the dog breed prediction columns that is p1, p2 and p3 and then replace underscore in dog breed's name with space.

Tidiness

1-merge the last 4 culomns (doggo, floofer, pupper and puppo) to one titled as dog type, and create new categories to fill multiple types.

2-Merge numerator and denominator in one table named rating

3- Merging the three datasets in one table.

In the cleaning section the issues were managed. Firstly, a copy was made fore each dataset. Then, rows that contains retweets and replays they were removed because they are not original tweet by the account so they will effect the quality of analysis. After that, I dropped unnecessary columns from twitter_achive dataset such as the mentiond earlier the retweets and replays colums. To handle the data type issues in some columns I changed the tweet_id to object in the three data sets. In the twitter archieve table the stages columns has many missing values but its regarded as None and this will effect the analysis by having the most common stage as none.

Regarding Tidiness issues, most probelems were with the seperate columns. Firstly, rating numerators and denominators are in seperate columns and they be used in analysis so it better to have the rating in one column. Also, in the twitter archive dataset the dog stages were merged in one column labled 'dog type'. Finally, the three datasets were merged in one table by the tweet id.