wrangle_report

August 26, 2022

Wrangle Report

0.1 WeRateDogs; Twitter Analysis of Dog Rating

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

This project aimed at generating insights from twitter data about dog ratings. Three wrangling processes are followed in achieving this:

1. Gather
2. Assess
3. Clean

0.2 Gather

Data was gotten from three sources:

WeRateDogs Twitter Archive Data: This is manually downloaded and uploaded to Jupyter Notebook. File name is twitter-archive-enhanced.csv.

Tweet Image Prediction: Though this is a tsv file, it is hosted on Udacity's servers and is to be programmatically downloaded. The Requests library is used to download the file ìmage_predictions.tsv from this URL here.

Additional data from Tweepy library via the Twitter API: Additional data is queried from the twitter API to add more details and analysis to the report. Retweet Count and Favorite Count by Twitter IDs will be queried.

0.3 Assess

After the data was gathered from the different sources, this is then visually and programmatically assessed to identify possible **quality** and **tidiness** issues.

The issues identifed are:

Quality issues

1. Since only original ratings are needed, the replies and retweets are not necessary for the analysis.

2. Tweets that have no image and even retweets/replies are included in the data.

3. Since there's no need for retweets and replies, some columns are extraneous in the data.

4. The 'timestamp' column has a consistent tailing '+0000' which is unnecessary.
5. The 'source' column in the dataset still has unneeded html elements which make the column not easy to read and use.
6. The timestamp field has the wrong datatype. This is to be converted to datetime.
7. Inconsistent letter case in the 'p1', 'p2' and 'p3' columns as some are lower case while some are sentence case.
8. Some columns do not have good descriptive names.

Tidiness issues

1. The analysis will be cleaner and easier if the three datasets (twitter_archive, image_prediction_file, tweet_json_data) were merged into a single DataFrame.
2. The doggo, floofer, pupper and puppo columns on the twitter_archive dataset could actually be Melt into a single column and called say, 'dog_type'.

0.4 Clean

Now we come to the part where we do the cleaning of the aforementioned issues so we can have a clean dataset and consequently a qualit and accurate analysis.

1. Since only original ratings are needed, the replies and retweets are not necessary for the analysis. **This was cleaned up by deleting records that are actually replies and retweets**
2. Tweets that have no image and even retweets/replies are included in the data. **Since tweets that do not have image are not needed, these were excluded from the data that is to be analysed**
3. Since there's no need for retweets and replies, some columns are extraneous in the data. **Columns that exclusively have to do with retweets and replies were removed since they serve no purpose to the data**
4. The 'timestamp' column has a consistent trailing '+0000' which is unnecessary. **The unnecessary trailing '+0000' after each value in the timestamp column was removed with regex expression**
5. The 'source' column in the dataset still has unneeded html elements which make the column not easy to read and use. **The source of tweet was cleaned-up with the use of regex expression**
6. The timestamp field has the wrong datatype. This is to be converted to datetime. **The timestamp column that had an object datatype was converted to a datetime datatype using the pandas' to_datetime**
7. Inconsistent letter case in the 'p1', 'p2' and 'p3' columns as some are lower case while some are sentence case. **The 'p1', 'p2' and 'p3' columns were converted to lowercase**
8. Some columns do not have good descriptive names. **The necessary columns were renamed appropraitely**

9. The analysis will be cleaner and easier if the three datasets (twitter_archive, image_prediction_file, tweet_json_data) were merged into a single DataFrame. **The three datasets were merged together using the merge function in pandas**

10. The doggo, floofer, pupper and puppo columns on the twitter_archive dataset could actually be Melt into a single column and called say, 'dog_type'. **These columns were converted into a single column using the melt function**