FWD-UDACITY

BY: Sara Abozeina

14-12-2020

Wrangle and Analyze Data Report

Contents :

This is an assignment for the Udacity Analyst Nonodegree

Page1

Wrangle and Analyze Data Report

Wrangling the data were gathered from Twitter API for WeRateDogs Dataset. WeRateDogs twitter account posts pictures and videos of dogs. It also gives ratings and also describe dogs in its own fashion. During that, three different data sources were gathered, assessed the quality and tidiness of the data, cleaned and finally the data analysis and visualization of the insights.

Data Gathering

There were three main sources for the data to deal with:

- The first data source " twitter_archive_enhanced" was given by Udacity in the form of a csv file.

- The second source "image_predictions" , were downloaded as a tsv file by its URL using Request Library and "pd.read_csv" pandas function.

- The final dataset was tweet's JSON data thar gathered from twitter RESET API via Tweepy library by quering the API to obtain extra information.

- After gathering all the three sources of data, the data assessment step was started.

Data Assessment

We investigate our imported_data visually using spreadsheets and programmatically regards to data quality and tidiness issues.

In this step we finding errors in dataset and noting down to correct it in the next steps.

Quality issues

A summary of the data quality and tidiness issues identified are as following:

twitter-archive table

- "tweet_id datatype is int64, require str(string)
- type of 'timestamp' and 'retweeted_status_timestamp' are str not Timestamp.
- Separate Date and Time from timestamp column and create two new columns.

- As per project description 'doggo' , 'floofer' , 'pupper' & 'puppo' columns have some missing values.
- Need to fix consistency issue for 'doggo','floofer','pupper'& 'puppo' columns and then assign NaN , if one of its corresponding columns already filled with non-null value.
- Remove rows where there are no images.
- Fix rating numerator and denominators that are not actually ratings.
- Fix rating numerator that have decimals.
- Change missing values in 'name' from 'None' to NaN.
- Remove extra characters after '&' in 'text' column.

image-prediction table

- 'tweet_id' datatype is int64, require str(string)
- A new column 'type' could be created which will represent dog types.

api_df table

- 'retweet_count' & 'favorite_count' column have some missing values :"Not Exist" .
- 'retweet_count' & 'favorite_count' datatype are str(string) , required int64.

Tidiness

twitter-archive table

- 'doggo' , 'floofer' , 'pupper' & 'puppo' are stage of dogs,so they should be under 1 column

image_predictions table.

- image_predictions should have been merged in twitter_archive api_df table.

Api_df table

- 'retweet_count' & 'favorite_count' features should be the part of the twitter-archive table.
- api_df dataframe should have been merged in twitter_archive.

Cleaning Data

The step included defining the problem, coding and testing it to know if the problem was fixed. The cleaned version of the three original datasets were stored in a folder called “Stored_Clean_Data”.

Analysis and Visualization

By applying some basic analysis, we can get the favorite dog, most retweeted dog and highest rating_numerator, … etc.

Finally I visualized data analyst for rating_numerator and favorite retweets.