Project 5: Wrangle and Analyze Data

WRANGLE_REPORT

Sara Alazeri .

Introduction:

We will wrangle (analyze and visualize) a dataset which is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Note: I did not use Twitter API because they Couldn not support me with an access to developer account.

Gathering Data:

The data for this project consist on three different dataset that were obtained as following:

- **Twitter archive file**: the twitter_archive_enhanced.csv was provided by Udacity and downloaded manually.

- **The tweet image predictions:** i.e., what breed of is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information.

- **Twitter API & JSON:** by using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file. I read this .txt file line by line into a pandas dataframe with tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and url.

Assessing Data:

Tidiness:

1. rating_numerator and rating_denominator can be in one column.
2. doggo , floofer , pupper , and puppo are all dog type so we need to merge them in one column.

Quality:

1- Wrong dog names sometimes "a".

2- Empty columns in twitter_archive

3- Some duplicated records in twitter archive because of retweets.

4- 'timestamp' has wrong datatype. It is a object and should be datetime .

5- 'id' name should be change to 'tweet_id' in (tweet) table .

6- 'tweet_id' has wrong datatype it is int and should be a string in (tweet_archive) table .

7- 'tweet_id' has wrong datatype it is int and should be a string in (image_predictions) table .

8- Drop the duplicated rows in (image_predictions) table.

9- "jpg_url" information that don't benefit any thing to our analysis so we don't need it.

Cleaning Data:

In this part i create a copy of the original datasets and I work with it without effect the real data .Also the Cleaning Data was divided into three parts: Define, code and test the code. These three steps were on each of the issues described in the assess section .After that I use the cleaned data to make the Visualization process.