Wrangle Report

Ekaterina Kuznetsova, 21.12.21

Introduction:

WeRateDogs is a Twiiter account that rates the dogs of people with a humorous comment about the dog, that let people to rate dogs with a funny comment and we analyzed an archive of tweets from Twitter user @dog_rates

WeRateDogs Project consist from:

1 .Data Wrangling includes 3 main parts:

1. Gathering Data
2. Accessing Data
3. Cleaning Data

2. Analysing and Visualising Data

3. Reporting

Step 1. Gathering Data:

Gathering for this Project consisted 3 pieces, which were gathering and represented as Data frames:

1. WeRateDogs archive “twitter-archive-enhanced.csv”, which we manually downloaded. This File was provided to Udacity Data Analytics students

2. Tweet image prediction “image_predictions.tsv”. This File was downloaded programmatically with help of Request library from a provided from Udacity URL.

3. Querying an API “tweet_json.json”. Twitter API and Python's Tweepy library to gather each tweet, for getting JSON objects of all the tweet_ids using Tweepy and importing this data into Jupiter Notebook.

Step 2 Accessing Data

We are assessing the data (all 3 datasets) on Quality and Tidiness issues.

Quality issues including: Missing, Invalid ,Inaccurate and Inconsistent data

Quality Issues:

1. 'twitter_archive_enhanced' table:

- remove not needed for analysis columns *(Inconsistent Data)*

- dog names: some dogs have 'None' as a name, or 'a', or 'an.' *(Validity)*
- change timestamp from string to date *(Invalid Data)*

2. 'twitter_archive_enhanced' table:

- p1, p2 and p3 has uppercase and lowcase and should be Uppercase *(Inconsistent Data)*

- drop duplicated jpg_url (Validity)
- rename columns in clear names (Accuracy)

3. 'tweet_json' table:

- id should be rename to tweet_id for further merging (Inaccurate Data)
- id and id_str are the same (Inaccurate Data)

Tidiness Issues:

1. Make part of image_predictions a part of the twitter_archive table
2. Create one table "rating" from rating_numerator and rating_denominator
3. Create new column 'breed' from 4 merged 'doggo', 'floofer', 'pupper' and 'puppo' columns

Step 3 Cleaning Data

I cleaned the data according the scheme: Define, Code and Test

Quality Solutions:

1. Convertes timestramp column datatype to datetipe
2. Dropped duplicates in jpg_urls
3. Rename columns “source, expanded_urls and timestramp” with help of rename function.
4. Removing ['retweeted_status_user_id', 'retweeted_status_timestamp','in_reply_to_status_id', 'in_reply_to_user_id','retweeted_status_id']
5. Replaced wrong dog names (stopwords) into Nan
6. Dropout rows with P1_confidence ['p1_conf'] less than coefficient 0.5 in image_prediction.
7. Convert type of name column to string and normilize later with title method. Drop p2,p3 and rename p1 into “breed” .
8. Convert the null values to None type in dogs name with replacing wrong dog names into Nan
9. Merge the clean versions of archive, images, and twitter datasets into new one correct dataframe.

Tidiness Solutions:

1. Merge image_prediction and twitter_archive table

2. ‘rating_numerator’ and ‘rating_denominator’ should be meged into one column “rating”

3. 'doggo', 'floofer', 'pupper' and 'puppo' should be merged into one column

4. Merge 3 cleaned data sets into one dataset.

Solution: Drop 'doggo', 'floofer', 'pupper' and 'puppo' columns and with help of regular expression concentrate them in one column.

Result:

Received dataset “twitter_archive_master.csv” consists all information and used for analyzing and visualization .