Udacity Project 5

Wrangle And Analyze Data

August 13, 2019

Wrangle Act Report

Data Gathering

Data is gathered from 3 resources and saved as 3 DataFrames: *df, df2, df3*.

1. Gathering the data from file on hand

Use pd.read_csv() to read data from existing file *twitter-archive-enhanced.csv* and save it as *df*.

Extract the tweet_id from url The column *tweet_id* in *df* has wrong value and datatype. Extract *tweet_id* from *expanded_urls*.

2. Download file using Requests library and URL

Download file *image_prediction.tsv* programmatically from the Internet and store data in *df2*.

3. Gather data from twitter API using Python's Tweepy library and store data

Get *retweet_count* and *favorite_count* from twitter API for records with *tweet_id* from *df*. Save data as text file *tweet_json.txt*, then read the file and store data in *df3*.

Assess Data

Quality

In *df*, the *tweet_ID* is not the right data type and value. I extracted the *tweet_ID* from *expanded_urls*, but still some *tweet_ID* values are missing.
Erroneous datatypes and values for *in_reply_to_status_id*, *in_reply_to_user_id*.
In *df*, we only want original ratings (no retweets).So the retweets shouldn't be there.
We only want ratings with images. Not all ratings have images.

In *df*, some ratings are wrong.
In *df*, erroneous datatype for *timestamp*.
In *df*, nulls represented as 'None' in columns *name*, *doggo*, *floofer*, *pupper*,*puppo*.
In *df*, some dog names are not correct.
In *df2*, some predictions are not dogs, there is no column for the most possible breed of a dog.

Tidiness

1. In *df*, the columns *retweeted_status_id*, *retweeted_status_user_id*, and *retweeted_status_timestamp* are not useful after we get rid of retweets.
2. in df1, the columns *doggo*, *floofer*, *pupper*,*puppo* show one variable.
3. *df3* should be part of *df1*.
4. *rating_numerator* and *denominator* should be one variable rating.

Clean Data

Copy *df*, *df2*, *df3* as *df1_clean*, *df2_clean*, *df3_clean*.

Issue 1
Some oberservations don't have *tweet_id* value.
In df1, the columns *retweeted_status_id*, *retweeted_status_user_id*, and *retweeted_status_timestamp* are not useful after we get rid of retweets.

Define
Delete retweets and observations without ID
Delete columns: *retweeted_status_id*,*retweeted_status_user_id*,*retweeted_status_timestamp*

Issue 2
We only want ratings with images. Not all ratings have images.

Define
Delete observations without image in *df1_clean*

Issue 3
One variable in four columns in *df*.
Nulls represented as 'None' in columns *name*, *doggo*, *floofer*, *pupper*,*puppo*.

Define
Create column *stage* to show dog stage, drop columns *doggo*,*floofer*,*pupper*,*puppo*. Replace 'None' with np.nan.

Issue 4
df3 should be part of *df*.

Define
Join *df3_clean* table to *df3_clean* table, joining on *tweet_id*.

Issue 5
Erroneous datatype for *timestamp*

Define
Convert *timestamp* to datetime data type.

Issue 6
In *df1*, nulls represented as 'None' in columns *name*, some values are wrong in *name*. Names that aren't capitalized are wrong.

Define
Set wrong names to 'None' and replace 'None' with np.nan.

Issue 7
In *df1*, some ratings are wrong.
Rating_numerator and *denominator should* be one variable rating.

Define
Change the *rating_numerator* and *rating_denominator* for oberservations with wrong value
Oberservations with tweet_id *810984652412424192* doesn't have a valid rating, so drop this row.
Create new column *rating = rating_numerator/rating_denominator*. Drop *rating_numerator* and *rating_denominator*.

Drop oberservations with extreme ratings.

Issue 8
In *df2*, some predictions are not dogs, there is no column for the most possible breed of a dog and the confidence.

Define
Create new columns *predicted_breed* and *predicted_conf* for the most possible breed of a dog and the confidence.

Store Data

Store the clean DataFrame *df1_clean* in a CSV file named ‘twitter_archive_master.csv’ and *df2_clean*
in additional file ‘twitter_image_predictions.csv’.