Wrangle Report

Godwin Akpa

Project Goal

Wrangle WeRateDogs Twitter data to create interesting and trustworthy analysis and visualization.

Data Gathering

Data for this project was gathered from three (3) sources:

i. **archive: Directly downloaded as twitter-archive-enhanced.csv using the of read_csv()**
ii. **image: Downloaded as image-predictions.tsv using the Request library**
iii. **tweet: This additional data was gotten via the Twitter API as tweet_json.txt using the Tweepy library.**

Access Data

Quality issues

a. *archive*

i. **Archive and tweet tables contain unnecessary columns.**
ii. **Wrong data type for **timestamp** and **tweet_id**.**
iii. **Most rows have wrong rating**

b. *image*

4. **Inconsistent column name for tweet table compared to archive and image table.**
5. **The types of dogs in columns **p1**, **p2** and **p3** had some uppercase and lower case letters**
6. **Wrong datatype (**p1_conf**, **p2_conf**, **p3_conf**) and (**p1_dog**, **p2_dog**, **p3_dog**)**

c. *tweet*

7. **Very few values recorded in (**contributors**, **coordinates**, **extended_entities**, **geo**, **place**, **retweeted_status**, and **place**)**
8. **Wrong datatype for **tweet_id**.**
9. **Values in **Sources** are not readable.**

Tidiness issues

1. **Missing information for dog stage resulting from many instances of 'dog stage' in many columns - archive**
2. **The tweet, image, and archive table can be merged as they contain related fields**

Clean Data

Quality

Make a copy of all the three dataframes using the .copy()

. archive_copy = archive.copy()
. image_copy = image copy()
. tweet_copy = tweet.copy()

1. **Issues #1: archive_copy, tweet_copy -> Unnecessary columns**
   Define: Drop Unnecessary columns

2. **Issue #2, #6, #8: archive_copy, image_copy -> Wrong datatypes**

   Define:

- ***In the `archive_copy` table, change the dtype of column **timestamp** from object to datetime using pandas `to_datetime()` function.***
- ***In the `archive_copy` table, change the dtype of column **tweet_id** from int64 to object using the `astype()` function.***
- ***In the `image_copy` table, change the dtype of column **tweet_id** from int64 to object using the `astype()` function.***
- ***In the `image_copy` table, Convert (p1_conf,p2_conf,p3_conf) from string to float using the `astype()` function.***
- ***In the `image_copy` table, Convert (p1_dog,p2_dog,p3_dog) from string to bool using the `astype()` function.***

3. **Issue #3: archive_copy -> Some dog names are wrong**

   Define: Convert wrong names into nan values

4. **Issue #4: tweet_copy -> Inconsistent column names.**

   Define: Change the column name 'id_str' to 'tweet_id' using the .rename()

5. **Issue #5: image_copy -> The types of dogs in P1, p2 and p3 had some upper-case and lower-case letters.**

   Define: Convert all the names of dogbreeds in the p1, p2, and p3 to lowercase letters in the image_pred_copy table.

6. **Issue #7: tweet_copy -> Very few values recorded for 'contributors', 'coordinates', 'extended_entities', 'geo', 'place', and 'retweeted_status'.**

   Define: Drop column ('contributors', 'coordinates', 'extended_entities', 'geo', 'place', and 'retweeted_status')

7. **Issue #9: archive_copy, tweet_copy -> 'sources' not readable.**

   Define: Extract the four (4) main sources categories from the 'source' columns

Tidiness

1. **Issue #1: archive_copy -> Missing information for dog stages resulting from many instances of 'dog_stage' in separate columns.**

2. **Define:**

   i. ***Create a separate column 'dog_stage'***

   ii. ***Extract all not null value from each column***

   iii. ***Drop ('doggo', 'floofer', 'pupper', 'puppo')***

2. Issue #2: The tweet, image, and archive table can be merged as they contain related fields.

Define:

i. ***Merge both the archive_copy and tweet_copy tables into one (merge_1) using the merge() on the common column ('tweet_id')***

ii. ***Merge the resulting merge_1 table to the image_copy table using the merge() on the 'tweet_id' column.***