Johnpaul Ebubechukwu Onuora

Wrangle Report

August 19, 2022

1 Data Wrangling Report

1.0.1 Project Description

WeRateDogs is a twitter account that rates people dogs with humorous comment about the dog by sharing the dog's image and brief comments about it. These ratings almost always have a denominator of 10. The numerators though? Almost greater than 10. 11/10, 12/10,13/10 etc. Why? Because "they're good dogs Brent".

WeRateDogs has over 4 million followers and has received international media coverage. They have shared 5000+ of their tweets containing some basic data such as there breeds and brief comments.

1.0.2 Project Objective

The objective of this project is to gather these data from different sources, thoroughly assess them and clean them to raise the tidiness and quality of the data in order to create a correct analysis.

1.0.3 Data Gathering

The data used for this analysis was gathered from three different sources

1. **twitter_archive_enhanced.csv** WeRateDogs twitter archive file, which was downloaded manually.
2. **image_predictions.txt** an image prediction file downloaded programmatically from its Url using request library.
3. **twitter** This last data was generated from Twitter API using Tweepy library as **tweet_json.txt**.

1.1 Assessing Data

The above pieces of data was assessed both **Visually** and **Programmatically**. During the assessment, the following **quality** and **tidiness** issues were found;

1.1.1 Quality issues

1. Archive table.

- Contains Irrelevant columns (**in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id, retweeted_status_user_id,retweeted_status_timestamp, expanded_url**).
- Wrong datatype for **tweet_id.** (str).
- Wrong datatype for **timestamp** and **retweeted_status_timestamp.** The datatypes are objects instead of datetime.
- Over 745 names in **name** column is None.
- **doggo,floofer,pupper and puppo** columns contains None values.
- All names in **name** column are not capitalized.
- **source** column too lengthy and not readable.
- **name** column has over 654 duplicates.

2. **Image Table**.

- **p1,p2,p3** contains uppercase and lowercase letters.
- Wrong datatype for **tweet_id**. (str)
- **jpg_Url** column name makes no sense.

3. **Tweet table**.

- The **source** column is not readable and too lengthy.
- **id_str** has a wrong datatype.
- **Id_str** is the same with **Id** contents are the same and have to be changed to tweet_id.
- **created_at** column has a wrong datatype.
- **place,geo,contributors,coordinates and retweeted_status** etc.. columns contains zero (0) and few records.

1.1.2 Tidiness Issues

1. **Archive Table**
    - **doggo,floofer,pupper and puppo** columns are developmental stages of dog information.
2. **General**
    - All tables should be merged as one table as the all contain corresponding fields.

1.2 Data Cleaning

Before cleaning this data from the above issues, a copy of each dataset was made to keep the original data. the copied data were cleansed from the quality and tidy issues following the steps below;

Step 1: Making copies of data

- **arch_copy = arch_df.copy()**
- **image_copy = image_df.copy()**
- **tweet_copy = tweet_df.copy()**

Step2: Handling Quality issues

Issue #1: Irrelevant columns contained in archive and tweet tables

Define:

- Irrelevant columns conatined in archive and tweet tables will be dropped using .drop() function.

Issue #2: Capitalize words in name column in archive table, p1,p2,p3 column in image table

Define:

- Capitalize the first letters in name column, p1,p2 and p3 column using capitalize() function.

Issue #3: Handling Wrong datatypes

Define:

- Change timestamp datatype to datetime using pd.to_datetime() function.
- Change id_str datatype to integer using astype(int) function.
- Change created_at datatype to datetime using pd.to_datetime() function.

Issue #4: Rename jp_url

Define:

- Rename jpg_url to image_url to be more understandable using the .rename() function.

Issue #5: Rename id_str to tweet_id because of its unique values

Define:

- Rename id_str to tweet_id using .rename() function to enable us merge these datasets together.

Issue #6: Change the position of tweet_id in tweet_copy table

Define:

- We need to remove the column from our dataset using .pop() function and save it in a variable called f_column.
- Insert the the column at the first position using .insert() function.
- Preview our dataset and check if it worked.

Issue #7: None values

Define:

We need to change all None values in name, doggo,floofer,pupper and puppo columns to NaN. Because, np.nan allows for vectorized operations; its a float value, while None, by definition,forces object type, which basically disables all efficiency in numpy.

Issue #8: Source column values in tweet and archive table is lengthy

Define:

- We need to get the unique source values in these tables and replace them with something shorter and understandable using .replace() function

Step3: Handling Tidiness issues

Issue #1: doggo, pupper, floofer and puppo are different developmental stages of dogs

Define:

- We need to create a new column that will contain the four different dog stages with name dog_stages.
- We would drop doggo, pupper, floofer and puppo column from archive table.
- Check if our code worked perfectly.

Issue #2: Merge the three tables to one table

Define:

- We would merge these three tables to one using the pd.merge() function on its unique identifers tweet_id.