wrangle_report

October 10, 2019

1 Wrangle Report

The following report provides a general and concise description of the data wrangling process related to WeRateDogs dataset. This process can be divided into three main steps: - Gathering data - Assessing data - Cleaning data

1.1 Gathering Data

For this project, I worked with three separate data sources each with a different method of gathering. 1. Enhanced Twitter Archive

This csv file was directly given to me and I only needed to import it into the working environment using pandas `read_csv()` function.

2. Image prdictions

for this tsv file, I was given a url and had to download it programmatically using the requests library.

3. Twitter api data

Using the tweet IDs in the WeRateDogs Twitter archive, I had to query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Then I read tweet_json.txt file line by line into a pandas DataFrame.

1.2 Assessing Data

The three datasets were assessed for quality and tidiness both visually and programmatically, and the following data quality and data tidiness issues were found.

1.2.1 Quality Issues

`twitter _archive` **table**

1. there are 181 retweets
2. Erroneous data type (*timestamp* column)
3. two variables in one column (the *text* column contains both the tweet text and url)
4. missing values in the *name* column
5. inaccurate values in the *name* column ('the', 'an', 'a', 'quite')
6. inaccurate ratings in the (*rating_numerator* column and the *rating_denominator* column)
7. missing total rating column (*rating_numerator / rating_denominator*)

`image_predictions` **table**

1. the *p1*, *p2*, *p3* columns have '_' instead of spaces
2. some predictions are not for dogs
3. non descriptive column name (*jpg_url*)

`twitter_api` **table** n/a

1.2.2 Tidiness Issues

`twitter _archive` **table**

1. unnecessary columns (*source*, *in_reply_to_status_id*, *in_reply_to_user_id*, *retweeted_status_id*, *retweeted_status_user_id*, *retweeted_status_timestamp*)
2. column headers are values and not a variable name (*doggo*, *floofer*, *pupper*, *puppo*)

`image_predictions` **table**

1. unnecessary columns
2. this table should be part of the twitter_archive table to filter tweets

`twitter_api` **table**

1. the table should be part of twitter_archive table

1.3 Cleaning Data

The data quality and tidiness issues ranged from mild to challenging, and it took a lot of effort to get the data in the best possible format available.

many of the columns had wrong data types that had to be converted to the correct one using `astype()` function. Other columns were inaccurately extracted, which resulted in a lot of incorrect and missing values, so I had to re-extract them again and drop the inaccurate values.

some columns were either unnecessary that I had to drop to facilitate analysis or had undescriptive names which I had to change to a more interpretable one.

Even merging the three tables together for easier analysis resulted in some columns being casted with a wrong data type and I had to convert them back to the correct one.

1.4 Saving Data

The final step of the wrangling process was to save the data into a csv file `twitter_archive_master.csv`, as well as a sqlite database that will be used later for analysis and visualization.