WRANGLE REPORT

August 19, 2019

1 WRANGLE REPORT

1.1 Introduction

This Wrangle and Analyze Data Project is part of Udacity's Data Analyst Nanodegree. The project involves wrangling of data from various sources associated with tweets from the Twitter user @dog_rates, also known as WeRateDogs. After scraping together the data, quality and tidiness issues were assessed and then cleaned.

1.2 1- Gathering Data

Data were collected from three different sources. - First:- data was collected from the "twitter-archive-enhanced.csv" file which was in the same directory in which project notebook was located. The csv file was imported into pandas dataframe `twitter_archive`. - Second:- data was extracted programmatically from a URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image predictions/image-predictions.tsv. Python's request library was used to extract data from URL and save it to a file .this file was imported as a dataframe in pandas `image_predictions`. - Third:- data was extracted from Twitter API using python's tweepy library. I needed to extract the favourites and retweet counts for each tweet. This data was then saved as a JSON file.

1.3 2- Assessing Data

1.3.1 Quality

`twitter_archive`

- Pandas'.info() method showed that tweet_id is an integer not a string, timestamp column needed to be a datetime object instead of a string.
- df.name.value_counts() showed that there were a,an,the and by are used as names
- df.head() showed that Unnecessary html tags in source column.
- Nulls represented as "None" (str) for name, doggo, floofer, pupper, and puppo columns
- The numerator and denominator columns have unusual values.
- There are 2075 rows in the image_predictions, 2356 rows in twitter_archive dataframe and 2333 rows in the status_df.
- 137 duplicated rows in expanded_urls

`image_predictions`

- pred_img.jpg_url.duplicated().sum() showed that there were 66 duplicated jpg_url

```
status_df
```

1.3.2 Tidiness

- doggo, floofer, pupper and puppo columns in `twitter_archive` table should be merged into one column named "stage" and convert None to null
- Joining breed column with `twitter_archive` table
- retweet_count and favorite_count columns from `status_df` should be joined with twitter_archive table

1.4 3- Cleaning Data

- dropping unnecessary columns from `twitter_archive` dataframe
- doggo, floofer, pupper and puppo columns in twitter_archive table should be merged into one column named "stage".After i used `pandas.melt()` i found that `twitter_archive` was 4 times more than befor,so I used forloop to iterate through rows ,appended them to a list and created a dataframe. after that I used a forloop again to replace each value in stage column with one word.
- Condensing dog breed predictions by using function I created.
- Merging `breed` column with `twitter_archive` table by using `pandas.merge()`
- Merging `retweet_count` and `favorite_count` with`twitter_archive` table by `tweet_id`by using `pandas.merge()`.
- Using `astype` to convert integers to strings and object to date.
- Using `.str.lower()` to change the uppercase to lowercase.
- Using `.drop_duplicates()` to drop 66 duplicated rows in `jpg_url`and 137 duplicated rows in `expanded_urls`.
- Stripping all html anchor tags (i.e. <a..>) in `source` column and retain just the text in between the tags. Convert the datatype from string to categorical.
- Replacing a,an,the and by with `np.nan`
- Creating a function that identifies the value before the last / in the text and uses this in the rating_numerator column. Manually correct any ratings that are not covered by the function.

```
In [ ]:
```