We_Rate_Dogs Wrangling Project

Wrangle Report

Wrangling is a three step process. The first step is collecting our data so that we can analyse it to make decisions, this step is called as GATHERING. The second and most important step is to point out the problems and issues present in the data, this process of listing out the issues is known as assessing. The last and important step is cleaning our data and correcting the issues present in our data.

GATHERING

For this project I gather data from the provided .CSV file. There was also a link provided from which I downloaded a. TSV file and finally extracted the data of retweet_count and favorite_count from a Jason file which I stored in a text file tweet_jason by using twitter api tweepy.

After all, steps of gathering I got the below mentioned files:

1. twitter_archive.csv

2. image_predictions.tsv

3. count.csv (containing data of retweet_count and favorite_count).

ASSESSING

Using .info() and .value_counts() I was able to list the issues present in each of the table.

Assessing twitter_archive_enhanced :

QUALITY

- Remove retweeted rows.
- 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', -retweeted_status_timestamp' are of no use.
- Change tweet_id datatype to str.
- Convert timestamp to datetime.
- There is only 2075 rows in image_predictions table, therefore there might be some missing data in it or may be duplicate data in twitter_archive.
- Problem with denominator greater than 10.
- Problem with numerator greater than 20.

- a , an , this is not the name and they are starting with lowercase alphabet therefore names starting with lowercase are not valid change them to None

TIDINESS

- Merge image_predictions and twitter_archive to twitter_image so that we can have only rows which have image.
- Melt doggo, floofer, pupper and puppo in one column i.e. type_of_dog.

Assesseing image_predictions:

QUALITY

- Some breeds starts with lower case and some with upper case in columns p1, p2, p3.
- Change name of columns p1, p2, p3.
- Tweet_id to str.

Assesseing data:

QUALITY

- Tweet_id to str.

TIDINESS

- Merge all the three tables.

CLEANING

Cleaning further is a 3 step process of define, code and test. Some of the numerators and denominators were incorrect and were corrected looking at the text. Expanded url in twitter archive contains url of every tweet twice but I left it unchanged . Cleaning is done keeping all the issues in mind which were raised during assessing.

And finally after cleaning all the three tables I merged them into we_rate_dogs and stored it as a .CSV file with the name twitter_archive_master.csv.