Data Wrangling Report

Introduction

Data wrangling is the main objective of this project, which involves fixing the quality and tidiness of the data.

Data Gathering

1. Twitter_archive: I've got the data from Udacity and then I used pd.read_csv() to import them into Dataframe.

2. image_predictions: The tweet image predictions present in each tweet according to a neural network.
I used The file -'image_predictions.tsv' from Udacity and used the requests library and the provided URL.

3. `tweet_data`: Using the tweet IDs in the WeRateDogs Twitter archive, 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.

Data Assessing

We separate the data issues into 2 categories: data quality issues and tidiness issues.

• Quality Issues mean content issues like duplicate, missing or incorrect data
• Untidy Data has specific structural issues, structuring datasets to facilitate analysis.

quality Issues

1. In the text column, there are URLs, and these URLs also appear in the expanded_urls section. We don’t need those urls to appear twice.

2. The timestamp is an object instead of DateTime. So I change it to DateTime.

3. There are not all dog-related tweets.
There are some tweets that are not related to dogs so we can remove them.

4. Some tweets with rating_denominator are not 10. So I remove then and I left only 10.

5. Remove all the unnecessary columns: 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp'.

6. The column "name" contains 546 none values and 55 incorrect "a" names.
There are 546 missing dog names and there are 55 that call ‘a’ that are need to be under none because there are also missing names.

7.Column source has three values arranged as HTML code.

8.The retweet count and favorite count are float datatypes, and the source is an object datatype and not category.

9.Optimize the source content by ‘Twitter for iphone’, ‘Vine - Make a Scene’, ‘Twitter Web Client’, and ‘TweetDeck’.

Tidiness Issues

1. Columns 'doggo', 'floofer', 'pupper', 'puppo' should be in one column-dog_class
2. Create the rating numerator column more understandable to look like this :

rating_numerator
13.50
9.75
11.27
11.26

Instead of this:

rating_numerator	rating_denominator
12	10

3. Merge the tweet_data into the twitter_archive by using inner join.