By Mohammed El Garhi January 2021

Wrangle Report

This report describes the data wrangling efforts involved in completing the data wrangling project “WeRateDogs” for the Udacity Data Analyst Nanodegree Program.

The report will cover the following data wrangling steps:

1. Gathering Data
2. Assessing Data
3. Cleaning Data
4. Storing Data
5. Analyzing and Visualizing Data

1. Gathering Data

The Data for this project comes from three different sources:

1. The WeRateDogs Twitter archive **twitter_archive_enhanced.csv**
   Downloaded as a CSV file from the project page and imported into the Jupyter notebook as **df_archive** using the pandas.read_csv function.
2. The tweet image predictions **image_predictions.tsv**
   Programmatically downloaded the tsv file from Udacity’s server using the Requests library and the following URL:
   https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
   Then opened the file using read_csv and the separator \t and saved it as **df_ip**.

3. The Twitter API Data (Tweepy) **tweet_json.txt**
   Downloaded each tweet’s retweet count and favorite count using
   the Tweepy library to get the Tweets JSON data and storing them as a txt file to avoid having to run this time-consuming code again.
   Opening the txt file by reading it line by line and creating a dataframe **api_df** containing the columns id, retweet_count and favorite_count.

2. Assessing Data

The Data was assessed using visual assessment by using the head(), tail() and sample() functions as well as a variety of programmatic assessment.

Following quality and tidiness issues were detected:

Quality Issues

In df_archive

- Retweets should be dropped
- Remove columns in_reply_to_status_id, in_reply_to_user_id
- Timestamp is not in datetime
- Some Items in Name,doggo, floofer, pupper,puppo are None not NaN
- Urls are written twice in expanded_url
- Some expanded urls are missing
- Remove html code from source/ use more descriptive names
- In the rating_denominator there are different values than 10, Either People posting for more than one dog as an accumulative or one that wrote something not related to the dog entirely
- Problems with nominators, some have wrong ratings, the ratings are wrong and
- (new issue after making changes) changing of ratings has lead to them being of the type float, they need to be changed to int
- Other words have been used for floofer e.g. archive_clean['433']
- names 'a', 'an' and 'the' to NaN

In df_ip

- Remove row if p1_dog, p2_dog and p3_dog are false
- Remove duplicates jpg_url
- Drop img_num
- Names of the columns are not clear

Tidiness Issues

- Doggo, Floofer, Pupper, Puppo should be in one Column
- Multiple Breed predictions might be confusing
- All the three Dfs should be combined in one dataframe

3. Cleaning Data

I created copies of the dataframes (archive_clean, ip_cean, api_clean) in order to have a clean version of them at the end of the cleaning process.

Then I first cleaned the Quality issues in archive_clean and then in ip_clean. After that I cleaned all the tidiness issues, such as the combination of the three clean Dataframes into one master Dataframe.

While cleaning I first stated the Issue, then defined the issue, then cleaned it and finally tested whether it has worked or not.

Storing Data

After merging the three dataframes into one master dataframe I saved the master dataframe using the following code:

master_df.to_csv('twitter_archive_master.csv', index=False)

4. Analyzing and Visualizing Data

I created a new dataframe called **df_analyzing** by reading the stored master csv file.

I choose to analyze and visualize the following:

- The 10 most common dog breeds → Pie chart
- The 10 most retweeted dog breeds → bar chart
- The 10 most favorited dog breeds → bar chart
- The 10 most common dog names → horizontal bar chart