WeRateDogs - Wrangle Report

By Ugochukwu Uche-Agada
July, 2022

This is a report based on the data wrangling activity carried out on the Tweet archives of WeRateDogs Twitter account.
It covers the 3 actions in data wrangling: Gather,Assess and Clean.

Data Gathering

Data required for analysis sometimes is not always available at one location or data set. In this case, data had to be sourced from three separate locations and with three different methods of gathering.

1. From file on hand

   The file **twitter_archive_enhanced.csv** was already provided for so all that was done was to load the csv file into a dataframe.

2. Using the Requests Library

   The **image-predictions.tsv** was downloaded programmatically using the requests library and a link provided by Udacity.

3. Using Python's Tweepy library

   The last of the three files was downloaded using Tweepy library. This was made possible by the elevated access granted to my Twitter Developer Account. The `tweet_json.txt` was downloaded here.

Assess Data

When assessing, you're inspecting your data set for two things: Data quality issues: Data that has quality issues have issues with content like missing, duplicate, or incorrect data. This is called dirty data. Lack of tidiness: Data that has specific structural issues that slow you down when cleaning and analyzing, visualizing, or modeling your data later.

Source: Udacity

After using visually and programmatically assessing the datasets, **quality** and **tidiness** issues were identified.

Quality Issues

tweet_arch DataFrame

1. row values in retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp are not needed in the analysis
2. values in in_reply_to_status_id and in_reply_to_user_id are not needed in the analysis
3. timestamp has wrong datatypes
4. There are outliers in the numerator and denominator columns
5. Identical values in expanded_url
6. Source column is noisy since it can be shortened.
7. Incorrect values in name column

image_predict DataFrame

8. Inconsistent names in P1, P2 and P3

tweet_df DataFrame

9. wrong column datatypes for retweets and favorites

Tidiness issues

1. Retweets are not needed since only original tweets are needed.
2. Replies are not needed since only original tweets are needed.
3. From the Tweet_arch data frame, Puppo,Floofer,pupper and doggo should be values on a single column
4. Timestamp should be seperated to capture day, month and year.
5. 'img_num','p1','p1_conf','p1_dog','p2','p2_conf','p2_dog','p3','p3_conf','p3_dog' does not fulfil requirements of being different columns.
6. tweets_df should be merged with tweet_arch data frame

Clean Data

All the issues detected while assessing the data was cleaned using code and tested afterwards

The following actions were taken to clean the quality issues:

Quality Issues Cleaning

tweet_arch DataFrame

1. row values in retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp are not needed in the analysis
2. values in in_reply_to_status_id and in_reply_to_user_id are not needed in the analysis
3. timestamp has wrong datatypes
4. There are outliers in the numerator and denominator columns
5. Identical values in expanded_url
6. Source column is noisy since it can be shortened.
7. Incorrect values in name column

image_predict DataFrame

8. Inconsistent names in P1, P2 and P3

tweet_df DataFrame

9. wrong column datatypes for retweets and favorites

Tidiness issues Cleaning

1. Columns relating to retweets were dropped .
2. Columns relating to replies were dropped.
3. Puppo,Floofer,pupper and doggo were put in a single column **dog_stage** using the melt() function
4. Timestamp was separated to capture day, month and year.
5. 'img_num','p1','p1_conf','p1_dog','p2','p2_conf','p2_dog','p3','p3_conf','p3_dog' does not fulfil requirements of being different columns hence the named columns were melted into the **breed** and **confidence** columns.
6. The three data frames were merged into one master data.

At the end of this wrangling process, the final data set was saved in the project folder ready for analysing.