WeRateDogs Wrangle Report

Introduction

This project is aimed to use data analysis and data wrangling techniques learned in the Udacity Data Analysis Nanodegree course.
The main dataset used in this project is the archived tweets of the WeRateDogs account on twitter in addition to other sources of data.

Project Details

The steps included in this project are :

- Data Gathering
- Data Assessing
- Data Cleaning

- Gathering Data

Twitter_archive_enhanced.csv file which is a csv file with 5000 tweet information like tweet_id, timestamp and text . The file was given by Udacity and it was manually downloaded.

Image_predictions.tsv which holds what breed of dog (or other object, animal, etc.) is present in each tweet, top three predictions of what type of dog is in the image and the image URL. The file was hosted on Udacity's servers and was downloaded programmatically.

Twitter API file that has each tweet's retweet count and favorite ("like") count at minimum. Data was in a file called tweet_json.txt given by Udacity.

All files were imported using a different technique into a pandas dataframe.

- Assessing Data

After gathering and importing all 3 files some assessing was needed to extract quality and tidiness issues in the dataset.

- Enhanced Twitter Archive

By visual assessment it was noticed that columns doggo, floofer, pupper and puppo should be one column .Some columns were unnecessary to the analysis and should be removed.Using programmatic techniques it was noticed that timestamp columns is of type object, rating_numerator has not been correctly extracted from tweet, Some rating_denominator value are greater than 10 so they should be changed .Also some dog names are incorrectly entered like :"None", "the" and "a" .Lastly, Expanded_urls column has 639 rows with multiple URLs separated with a ',' .

- Image Prediction

By visual assessment it was noticed that columns p1, p2 and p3 are not all lowercase, some are uppercase they also have an underscore instead of a whitespace between the names .Also some columns needed to be renamed. Using programmatic assessment it was noticed that columns p1_dog, p2_dog and p3_dog in some rows are all False indicating no dog in the image.jpg_url column was checked for duplicates. Finally,324 tweets do not contain any dogs .

- Twitter API

No problems were noted for this dataframe.

- Cleaning Data

All three data frames were merged into one dataframe with tweet_id in all three being the index. A copy of the resulted dataframe was generated to use throughout the cleaning and analysis process.

First we removed the rows in column jpg_url that hold NaN values and tweets that do not contain any dogs(p1_dog, p2_dog and p3_dog are all False). Then columns : in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp and retweeted are not needed so they will be dropped.

Some data types are imported wrong and some were changed due to the merge. Timestamp column will be converted to datetime instead of string.Retweet_count, favorite_count and img_num columns will be converted from float to int.

Using the text column we will extract the correct number for rating_numerator that were wrongly imputed. We will do the same for rating_denominator. Columns doggo, floofer, pupper and puppo will be merged to one column called dog_class.

Replace incorrect urls and NaN values with twitter home url concatenated with the twitter ID . Columns p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3 , p3_conf and p3_dog will be changed to be more descriptive, remove wrong data with NaNs and convert underscores between the names to spaces.

- Storing Data

This is the last step before analysis. The final data frame will be saved as a csv file called : twitter_archive_master.