wrangle_report

August 7, 2022

0.1 Reporting: wragle_report

Author: Mlungiseleli Notshokovu | Data Analyst

0.2 Project: Wrangling and Analyze Data

The focus of this project is on wrangling and analyzing data from the WeRateDogs twitter page. The data was collected from different sources(csv file and twitter api) using Python.

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for us to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

The goal of this project is to wrangle the WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The challenge lies in the fact that the Twitter archive is great, but it only contains very basic tweet information that comes in JSON format. I needed to gather, asses and clean the Twitter data for a worthy analysis and visualization.

The project followed these steps:

Step 1: Gathering data The data was collected from various sources with different formats including the following: - A .csv file with over 5000 @WeRateDogs tweets. Pandas library was used to create a dataframe of this twitter archive - @WeRateDogs tweets image predictions data downloaded using the Requests library - Additonal @WeRateDogs twitter data via the Twitter API using the Tweepy library, then data was then stored in JSON

Step 2: Assessing data

The following criteria was used to properly assess the data: **Quality issues**

- timestamp datatype is not correct. Datetime is the correct datatype
- Will need to remove all retweeted data and leave only the original tweets
- Source links still have HTML tags. If useful, we can simply extract the source kind, i.e Iphone/Android
- id format should be string, not numeric
- Some dog names do not make sense, and some are just letters: 'a', 'the'
- Remove rows with no breed predictions.
- There are tweets with no images. These should be removed, as the rating is derived from the image

- For those tweets with multiple breed predictions, take the one with the highest confidence level
- Remove columns that are not useful and those that are empty, i.e retweeted_status_id, retweeted_status_user_id

Tidiness issues

- The columns: doggo, floofer, pupper, puppo can be under a single column describing the dog stage. **dog_stage**
- combine rating columns to one column. convert to float and calculate rating_numerator/rating_denominator as 'rating'
- Merge all 3 dataframes into 1 master dataframe for better analysis and visualization

Step 3: Cleaning data
The process of cleaning/addressing all issues that were detected in the Assess step
Step 4: Storing data
Once the data was properly cleaned and merged together, it was then saved to a 'twitter_archive_master.csv' file to analyze and visualize.
Step 5: Analyzing, and visualizing data
Insights and Visuals that help to answer the following questions:

1. What is the top used source to publish tweets?
2. What is the most popular dog breed?
3. What is the most popular dog name?
4. Top 10 highest ratings?
5. Top 10 favourite dog breeds?

Step 6: Reporting
A summary of data wrangling efforts, data analysis, and visualizations

0.3 Summary & Conclusion

First of all, this was the most educational Capstone project I have done so far. Being taken through the process of gathering, assessing, cleaning, analyzing, and visualizing data has definitely strengthened my love for analytics.

The following findings/insights were derived: - The top favourite dog breed appears to be a Golden Retriever, followed by a Labrador Retriever. - The highest dog breed rating is 7.5 - 'Copper' appears to be the most popular dog name followed by 'Oliver', 'Charlie', 'Lucy', and 'Tucker' - The most popular dog breed appears to be a Golden Retriever, followed by a Labrador Retriever. - 'Twitter for Iphone' is the top used source to publish tweets

Given more time and resources, I believe more insights could have been derived from this data: the users who publish these tweets, the type of algorithm used to predict dog breeds, etc.