5/22/2023

Wrangle Report

Udacity Wrangle and Analyze

Written by: Danielle Lamke

Contents

Introduction

WeRateDogs is a popular Twitter account that is known for its humorous reviews of dogs along with their pictures. In this project, I was tasked with wrangling the data from WeRateDogs, cleaning and assessing it, and then analyzing it to gain insights into this popular Twitter account's behavior and trends. In this report I will discuss what I did during these steps.

Data Gathering

The data was gathered from three different sources:

twitter-archive-enhanced.csv

The WeRateDogs Twitter archive csv file was provided by Udacity; this file was downloaded and then loaded to Jupyter notebook using pandas.

image-predictions.tsv

A file containing tweet image predictions of the dog breed from the tweets was downloaded using the request library and link provided by Udacity.

tweet_json.txt

Data on tweet engagement and retweet count was to be obtained using the Twitter API. Since my Twitter account lacked access, I used the file provided by Udacity for this data.

Assessing Data

Quality issues

1. The timestamp column was in string format, not datetime.
2. The dataset contains retweets that are not needed.
3. The 'name' column has invalid names (e.g., 'a', 'such', 'the', 'just', 'getting', etc.).
4. Dog breeds in columns p1, p2, and p3 are not capitalized consistently.
5. The source column contains HTML tags.
6. The ‘tweet_id’ column is not in a consistent format across all datasets.
7. Some jpg_url values are duplicated, which could indicate that some images were tweeted multiple times.
8. Columns ‘p1_conf’, ‘p2_conf’, and ‘p3_conf’ contain extended decimal confidence levels which can be difficult to interpret.

Tidiness issues

1. The ‘doggo’, ‘floofer’, ‘pupper’, and ‘puppo’ columns represent the same variable and should be combined into a single ‘stage’ column.
2. Columns ‘retweeted_status_id’, ‘retweeted_status_user_id’, and ‘retweeted_status_timestamp’ should be dropped since they are empty after removing the retweets.

Cleaning Data

To prevent any loss of information in case of an error, a duplicate copy of the datasets was created before making any changes. The above items were then corrected as described below:

Quality issues

1. The timestamp column was converted to datetime format.
2. The retweets were removed.
3. The invalid names were removed from the 'name' column.
4. Capitalization of predicted dog breeds in the p1, p2, and p3 columns were changed to lowercase.
5. The HTML tags were removed from the source column.
6. The tweet_id column was converted to string format for all datasets.
7. The duplicated jpg_url values were removed.
8. The p1_conf, p2_conf, and p3_conf columns confidence levels for each prediction were rounded to 2 decimal places.

Tidiness issues

1. The 'doggo', 'floofer', 'pupper', and 'puppo' columns were combined into a single 'stage' column.
2. The 'retweeted_status_id', 'retweeted_status_user_id', and 'retweeted_status_timestamp' columns were dropped.

Storing Data

The data was merged on 'tweet_id' into a master dataset called 'tw_master.' This master dataset was then stored as a csv file called twitter_archive_master.csv.

Analyzing and Visualizing Data

After obtaining a clean dataset, my initial analysis focused on the favorite count of different dog stages. 'Pupper' was found to have significantly higher favorite counts than any other stage. To conduct this analysis, I filtered out rows where the 'stage' column was null, grouped the data by stage, viewed the total favorite counts stage, then created a bar chart for a visual.

Next, I looked at the source to see where most posts came from. Twitter for iPhone was the most used method of posting. To find this information, I retrieved the value counts of the 'source' column, viewed total post counts by source, then created a bar chart for a visual.

Finally, I looked at the relationship between favorited tweets and retweets. Using a scatter plot followed by the calculation of the correlation coefficient; what I discovered is that as favorite count increases, retweet count also tends to increase. This means that the more a tweet is favorited it also tends to get more retweets.

Conclusion

In conclusion, this project involved the gathering, cleaning, and analysis of data from WeRateDogs, a popular Twitter account known for its unique rating system for dogs. Through my analysis, I was able to gain insights into the behavior and trends of this account, including the most favored stage, the most common source for posting, and the correlation between favorites versus retweets. Overall, this project provided valuable experience in data wrangling and analysis using real-world data.