Wrangle & Cleaning of WeRateDogs

Introduction

The purpose of this project is to put into practice what I've learned in Data wrangling data course which is part of Udacity Data Analysis Nanodegree program.

The dataset wrangles in this project is the tweet archive of Twitter user @dog_rates, also known as **WeRateDogs.** It rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. (11/10, 12/10, 13/10).

Process

This project focus on the aspects of

- Gathering
- Assessing
- Cleaning

Gathering Data

The data is gathered from different sources and finally merged to analyse and visualize the data.

Twitter archive file (udacity resources)

Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, I have filtered for tweets with ratings only (there are 2356). The file is named as **twitter-archive-enhanced.csv**

Tweet Image prediction(download from the URL)

This file contains top 3 predictions of for each dog image from the twitter archive file. Data is downloaded from the URL provided by the Udacity using the request library and loaded into **image-predictions.tsv.** This data consist of 2075 rows and 11 columns

Twitter API file(udacity resources)

The file contains tweet id, favorite count and retweet count. Data has downloaded manually **(tweet-json.txt).** This data consist 2354 rows and 2 columns.

Assessing Data

After the data has been gathered it is assessed to find the Null values, missing values and inapproriate measure or data entry.

A function **data_info** has created that will analyse the data and output the following

- Shape
- Duplicates
- Index are unique?
- Missing values
- Datatype of each column

Twitter archive file

- With the hep of the function the basis analysis on the data is completed.
- The data is assessed to check if the dog belogns to mulitple category and visualized.
- The **rating_denominator** is assessed to check the number that are valid.
- The **name** column is assessed to check the popular dog and the highest valid rating in the dataframe.

- The **rating_denominator** is checked if there are any float rating and are extracted incorrectly.
- **expanded_url** column is checked if it contains multiple url for a tweet.

Image prediction file

- With the hep of the function the basis analysis on the data is completed.
- The data is assessed to see the average confidence and correct predictions (i.e., is it dog?) for each prediction columns.

Twitter API file

- With the hep of the function the basis analysis on the data is completed.

Cleaning Data

The issues identified during assessing the data are cleaned and merged to get a clean dataframe.

Twitter archive file

- Removed retweets and inappropriate columns.
- Combining classification (doggo, pupper, floofer, puppo) to one column (Handling tweets with one or more category).
- Correcting the datetype of timestamp column.
- Correct rating when there is a float in the numerator.
- New column rating (numerator / denominator)
- Removing unwanted rows (denominator != 10) and columns (source, doggo, puppo, pupper, fluffer)
- Fixing the expanded URL's columns

After cleaning the data is analysed with data_info function and found

- **Shape** - 2158 X 8
- **Duplicates** – 0
- **Index** – Unique
- **Missing values** – 0
- **Datatype of each column is appropriate**

Image prediciton file

- Dropped the duplicate rows.
- Rename the columns with appropriate names.
- Remove "_" in the p1,p2 and p3.
- Wrangle the prediction and summarize to new column predicted_breed & prediction_confidence
- Drop inappropriate rows and columns.

After cleaning the data is analysed with data_info function and found

- **Shape** - 1691 X 3
- **Duplicates** – 0
- **Index** – Unique
- **Missing values** – 0
- **Datatype of each column is appropriate**

Twitte API file

- Dropped the duplicate rows.

After cleaning the data is analysed with data_info function and found

- **Shape** - 2353 X 1
- **Duplicates** – 0
- **Index** – Unique
- **Missing values** – 0
- **Datatype of each column is appropriate**