Wrangle and Analyse: WeRateDogs

May 25, 2020
Udacity – Data Analyst Nanodegree

Saikiran Bikumalla
Project- 4 (Wrangle and Analyse Data)
Hyderabad, India

Introduction :

Real-world data rarely comes clean. The dataset wrangled for this project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. This project works through the data wrangling process, focusing on the gathering, assessing, and cleaning of data. There are visualizations and observations from the analysis provided as well.

Goals:

- Data wrangling, which consists of:
    1. Gathering data
    2. Assessing data
    3. Cleaning data
- Storing, analyzing, and visualizing your wrangled data
- Reporting on data wrangling efforts , data analyses and visualizations

Gathering Data:

This project involved gathering data from three different sources:

1. The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file was provided to Udacity students which can be downloaded directly.
   This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and has been downloaded programmatically using the Requests library and the following URL:
   https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3. Using Twitter API (which is authorised) and extracting the contents from twitter API such as tweet_id , retweet count and favorites count using python tweepy library and storing it in tweet_json.txt file. Later this file is used to read data in json format (pd.read_json( )) and converted into a dataframe to perform cleaning and analysis.

Assessing data:

Assess data involves the evaluation and assessment of the dataset to get to know the quality and tidiness issues of the dataset. The issues that I have assessed are listed below.

Quality Issues :

- **archive table**:
  1. tweet_id is an int
  2. timestamp is an object and retweeted_status_timestamp is also an object
  3. name column has dog names with None, a, an.
  4. html tags in source column which have link of source
  5. This dataset includes retweets, which means there is duplicated data (as a result, these columns will be empty: retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp)
  6. missing data in many columns
  7. rating_denominator has values other than 10

- **images table**:
  1. tweet_id is an int
  2. p1, p2 and p3 columns have multiword dog breeds with (- or _)
  3. p1, p2 and p3 columns have names sometimes with lower case and other times sentence case.

Tidiness Issues:

- Last four columns in archive table belong to the same variable.
- Unwanted columns or duplicate columns for retweets data in archive table.
- All the three dataframes belong to the same observational unit. so merge into one.

Cleaning Data:

Cleaning data is tedious, and often iterative. Just when an analyst believes they have found all quality and tidiness issues, there are often additional issues that arise. The cleaning process involves three steps:

1. Define: determine exactly what needs to be cleaned, and how
2. Code: programmatically clean the code
3. Test: evaluate the code to ensure the data set was cleaned properly

By using the above means i have cleaned the data in following way.

- Drop the columns retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_to_status_id,in_reply_to_user_id, expanded_urls.
- Create a new column called dog_type and store the type. drop the last four columns.

- Merge all the three data frames into a single data frame.
- Convert tweet_id from int64 to string(object) using astype( ).
- Convert timestamp from object(string) to datetime using to_datetime by removing last 5 digits from timestamp as they depict time zone.
- Gather all the faulty names present in the name column and replace those names with None
- Extract the text related to source (name of source) from the link.
- Assign all the values in rating_denominator to 10
- Replace _,- in p1,p2 and p3 columns with " " using str.replace() function
- Use str.capitalize() method to convert all values of p1,p2 and p3 columns to sentence case.
- Iterate through each row of the dataframe using iterrows() function and then select the dogbreed from p1, p2 and p3 columns using boolean indexing from p1_dog , p2_dog , p3_dog columns
- Drop img_num column from weratedogs dataframe

Storing , Analysis and Visualization:

After the dataset is cleaned properly, it is stored in a csv file named twitter_archive_master.csv

I have analysed the following five insights on this data:

- Golden retriever is the most popular dog breed.
- Correlation between favorite_count and retweet_count with a correlation coefficient of 0.86 which means there exits a strong positive correlation
- The most retweeted dog breed is Golden retriever, with Labrador retriever the next most and pembroke later
- Golden retriever breed has highest number of likes, with Labrador retriever next most.`
- The most popular dog names are Oliver, Cooper, Charlie, Tucker and Penny for a tie at 10 each.

Conclusions:

This write up provides a straight forward look for the data wrangling process essentials and data analysis and visualisation insights. There are so many more insights that can be assessed and analysed from this dataset, I highly encourage to deep dive into this data set to get what else we can find from it!