Wrangle and Analyse: WeRateDogs

May 29, 2020
Udacity – Data Analyst Nanodegree

Somasekhar Goud Addakula
Project- 4 (Wrangle and Analyse Data)

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:

1.twitter-archive-enhanced.csv

- ID fields: The ID fields, like tweet_id, in_reply_to_status_id etc. should be objects, not integers or floats because they are not numeric and aren't intended to perform calculations
- remove the columns which are having more than 90% of null values as they arent required
- convert the datatype of timestamp :to_datetime
- From the anchor tag extract only what is the actual source for the source attribute
- convert the rating_numerator and rating_denominator columns to float. Then iterate through the dataset (text) and find the records having decimal rating in numerator. store the indices and the decimal rating and relocate these values with faulty values.
- Name attribute has irrelevant values so replace them with None

2.image-predictions.tsv

- In p1_conf we find 100% confidence in prediction might be technically true but its not possible in general so remove it
- p1, p2 and p3 columns have names sometimes with lower case and other times sentence case.so convert into lower case
- p1, p2 and p3 columns have multiword dog breeds with (- or _) instead of white spaces
- p1, p2 and p3 columns have breeds some times with lowercase and other times with sentence case
- Create a new dog_breed column to store the breed of dog

3.from 'tweet_json.txt' in dataframe twitter_counts3(twitter_id,favorite_count,retweet_count)

- tweet_id is an integer convert into object

Tidiness issues:

1.twitter-archive-enhanced.csv

- For the various stages of dog:doggo,floofer,pupper,puppo create a column named 'dog_stages' for the multiple dog_stages.

*** 2)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.

Quality issues:

1).ID fields: The ID fields, like tweet_id, in_reply_to_status_id etc. should be objects, not integers or floats because they are not numeric and aren't intended to perform calculations
2).Remove the columns which are having more than 90% of null values as they arent required
they are
‘in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp'
3)convert the datatype of timestamp :to_datetime as it is a standard
4)From the anchor tag extract only what is the actual source for the source attribute that is nothing but the text related to source
5)convert the rating_numerator and rating_denominator columns to float. Then iterate through the dataset (text) and find the records having decimal rating in numerator. store the indices and the decimal rating and relocate these values with faulty values.
6)Name attribute has irrelevant values so replace them with “None” irrelevant values like these in the
['such', 'a', 'quite', 'one', 'incredibly', 'an', 'very', 'just',
'my', 'not', 'his', 'getting', 'this', 'unacceptable', 'all',
'infuriating', 'the', 'actually', 'by', 'officially', 'light',
'space']

7)In p1_conf we find 100% confidence in prediction might be technically true but its not possible in general so remove the related row
8)p1, p2 and p3 columns have multiword dog breeds with (- or _) instead of white spaces.
9)p1, p2 and p3 columns have breeds some times with lowercase and other times with sentence case using Capitalize method
10)Create a new dog_breed column to store the breed of dog

Tidiness issues:

For the various stages of dog:doggo,floofer,pupper,puppo create a column named 'dog_stages' for the multiple dog_stages.

*** 2)All the three dataframes belong to the same observational unit. so merge into one.**

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 six insights on this data:

1.correlation between favorite_count and retweet_count with a correlation coefficient of 0.86 which means there exits a - -strong positive correlation

2.Golden retriever is the most popular dog breed.

3.The most retweeted dog breed is Golden retriever, with Labrador retriever the next most and pembroke later

4.The most popular dog names are Charlie,Tucker,Cooper,Penny,Oliver with count of 10

5.Labrador retriever is the breed of dog in p1 which has the most favorite_count followed by the rest nine dog_breeds

6.Newfoundland is the dog breed of p2 with least retweet count and followed by the rest nine dog_breeds

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!