EZE IFEANYI

6/13/2022

WERATEDOGS WRANGLE REPORT

INTRODUCTION:

The WERATEDOGS data analysis project on Udacity consists of gathering three data's from different data source, assessing, cleaning and also merging the three different dataset as one. It also required me providing insights about my findings and also visualizing them.

Gathering:

The gathering of the data's where from three different source which are;

- **Twitter archive file**: This file was downloaded manually by clicking the following link: twitter_archive_enhanced.csv, this link was provided by the Udacity instructors in the classroom.
- **The tweet image predictions:** This file contained 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 was downloaded programmatically using the Requests library.
- **Twitter API & JSON**: Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas Data Frame with (at minimum) tweet ID, retweet count, and favorite count.
  NOTE: Due to personal reasons I had some verification issues, due to this issue, Udacity tutors had already provided the json.txt file in the classroom.

ASSESSMENT

Assessing the data's were done in two ways which are:

1. Visually
2. Programmatically

Visually

This had to do with scanning the data without the use of any code or function. This can be done by looking at them in an excel sheet or in a jupyter notebook.

Programmatically

This deals with assessing or looking at the data with a better in depth with the use of codes. This is also the use of pandas' functions and/or methods to assess the data.

After viewing the data, I made down notes of the quality issues of each dataset and also made some points about the tidiness issues.

CLEANING

This part of the data wrangling was divided in three parts: Define, code and test the code. These three steps were on each of the issues described in the assess section. First and very helpful step was to create a copy of the three original data frames. I wrote the codes to manipulate the copies. If there was an error, I could create a new copy from the original. Whenever I made a mistake, I could create another copy of the data frames and continue working on the cleaning part.

There were a couple of cleaning steps that were very challenging. One of my greatest challenge in cleaning this data was understanding the dataset, once I understood it cleaning the data became easier. These are the issues I handled while cleaning;

Quality Issues

- Dropping columns that won't be used for analysis
- Erroneous datatypes (doggo, floofer, pupper and puppo columns)
- Editing the timestamp
- Creating a ratings column(numerator/denominator)
- Creating a column for the month and year
- Replacing wrong words for name as "None"
- Create 1 column for dog type and 1 column for probability rate
- Dog stage to a categorical column

Tidiness issues

- Merging all three tables on the tweet_id
- Renaming the id column in the tweet data to tweet_id to merge with other dataset

MERGING

After cleaning what was left was merging the datasets for my analysis.