WRANGLE REPORT

Project#5: Wrangle and Analyze data.

Data analyst nanodegree program

REEM ALRASHOUD

July , 2020

Introduction:

This project points to urge quick data approximately puppy evaluations whereas illustrating progressed information wrangling and visualization strategies utilizing different Python libraries. I will assemble information from a assortment of sources and in a assortment of designs, evaluate its quality and tidiness, at that point clean it. Typically called information wrangling. The dataset that I will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, moreover known as WeRateDogs.

WeRateDogs may be a Twitter account that rates people's mutts with a amusing comment approximately the canine. The page has since developed greatly in popularity , with numerous clients sharing its substance and asking their pooches being appraised aswell. These appraisals nearly continuously have a denominator of 10. in spite of the fact that? Nearly continuously more prominent than 10. Since "they're great dogs."

Note: I did not use TWITTER API because they Couldn not support me with an access to developer account.

Step 1: Gathering Data

Here we will focus on three points of data:

1- `The WeRateDogs Twitter archive:` "Existing file" the WeRateDogs Twitter file contains fundamental tweet information for all 5000+ of their tweets. (manual download of 'twitter-archive-enhanced-2.csv').

2- `The tweet image predictions:` "Downable record" i.e., what breed of puppy or other obj, creature, etc. is display in each tweet agreeing to a neural network. This record (image_predictions.tsv) is facilitated on Udacity's servers and ought to be downloaded programmatically utilizing the Demands library and the taking after URL:https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv.

3- `The twitter API data "JSON API file":` with at least tweet ID, retweet tally, and favorite number in a record called 'tweet_json.txt' ,query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's whole set of JSON information in a record . Each tweet's JSON information ought to be composed to its possess line.

Step 2: Assessing Data

* Tidiness: (Messy Data)

1- In the dataset of (twitter-archive-enhanced-2.csv) , we will find : doggo , floofer , pupper , and puppo are all dog stages for this reason we need to reduce the duplication by merge them in one column.

2- rating_numerator and rating_denominator can be combined in one column.

3- We have three Separate tabels which are (Twitter archive data + Image predictions data + Tweet json data) I will merge them in one table to be more specific.

* Quality: (Dirty Data)

1- Timestamp column has type object , I want to change it to datetime type because it more suitable.

2- We need to rename p1 , p2 , and p3 Columns to be clearer. Because it is not organized in a way of capital and small letters.

3- doggo, floofer, pupper, puppo columns contain 'None' value where NaN should be used.

4- jpg_url contains two different path patterns to jpg files,This seems not to have any impact if we don't remove the duplication.

5- There are no calculations so 'tweet_id' columns should be string "or object" types Instead of int.

6- 179 rows missing data for favorites_count , that means we have more retweets than favourite tweets (Unusual situation).

7- There are many unneeded retweets, so we have to remove any (Retweets).

8- Change sources to have only the specific type of device "Valid source".

9- Remove entries where p1_dog, p2_dog, and p3_dog are all "False" in tweet_image.

10- Proper extraction of ratings.

Step 3: Cleaning Data

In this part , I did not work on the original data so I make a copy of the dataset that can be helpful to make any editions without effect the real data. In the data cleaning we have to follow three steps with every quality issue which are (Define – Code – Test). When I worked on the data I made the steps in a simple way to be more clearer. After that I use the cleaned data to make the Visualization process.

Visualization Process:

I worked on three graphs, which are: (Rating Distribution, Top 10 dogs names , and Relation between timestamp and status id)

Summary :

To accomplish the objectives of the examination, information have been assembled from diverse sources, counting the twitter as well as other web information. The possibilities of Python's pandas library has been utilized broadly on the appraisal and information cleaning parts.At the end I visualize the data by using specific sights