Wrangle-act Report

TWITTER API DATA

About the data

This project utilizes the Twitter dataset of the twitter user @dog_rates or WeRateDogs. This account is a platform that people can publish their dog' photos and the account administration gave these dogs rating. The rating system is out of 10 but the numerator can be greater than 10 to reflect the cuteness of the dogs.

The main aim of this project is to put the skills that is gained throughout this course into practice and gain insights of the dataset by applying the various processes on this data. So, below is a demonstration of the project details and the efforts that is done to fullfil the requirements.

Project details

The main processes that are applied on this- project are:

- Gathering data
- Assessing data
- Cleaning data
- Analyses & visualization of data
- Build prediction model

Gathering data

The dataset that is processed of this project consist of three dataset:

- **Twitter archive file:** This file is downloaded manually by the link provided in Udacity classroom. The file name is ***twitter_archive_enhanced.csv*** and contains some information about the tweets such as the tweet id and the dog stages. This file is imported as dataframe named (***twitter_arc)*.**
- **Image-predictions file:** this is a table of image predictions that is generated by a neural network according to the tweet. This file is downloaded programmatically using Requests library and URL information and stored locally to ***image_predictions.tsv*** file. This file is imported as a dataframe named ***(image_prediction_df).***
- **Twitter API & JSON:** by using the *tweet id* column values in ***Twitter archive file***, Twitter API is queried for each tweet's JSON data using Python's *Tweepy* package and each tweet's JSON data is stored in a text file called ***tweet_json.txt***. Pandas

dataframe is made from the selected variables in the text file such as (*tweet id, favorite count, retweet count, retweeted status and url*) and named ***(tweet_json).***

Assessing data

Data in this project is assessed using two approaches

Visual assessment

Visual assessment is used by printing the whole dataset and notice any errors that need to be cleaned. Throughout visual assessment, 1 quality issue and two tidiness issues were found.

- The quality issue was the html tags in the *Source* column in *twitter archive table* which is not necessary.
- The first tidiness issue is that the dog stages are separated in four columns instead of being values for dog stage column,
- The second issue is that the columns related to retweets will have empty cells for the original tweets, which can be dropped.

Programmatic assessment

by using the different methods (e.g. value_counts, info, sample, duplicated) and various types of indexing and selecting data such as (loc and bracket notation with/ without Boolean indexing and iloc). The issues that are listed below.

- Erroneous datatypes(*in_reply_to_status_id, in_reply_to_user_id, timestamps*), and found by using info method.
- The retweet columns and columns that will not be used for analysis.
- Correct the type of *rating_numerator* and *rating_denominator* to allow for decimals.
- Erroneous dog *name* that starts with lowercase characters
- The column of *dog_stage* should has a type of categorical
- The 66 duplicated values of jpg_url column
- Erroneous datatypes(confidence_level)
- The rows that do not have retweeted status of `Original tweet` (duplicated)

After that, these issues are separated into issues related to quality and others related to the tidiness.

Cleaning data

The cleaning process is divided into three sub processes:

- ***Define*** the issue
- ***Code*** to solve the issue
- ***Test*** the code

These steps are applied on all the issues that were located in the datasets. The first step was to create a copy of the original data and apply cleaning process on it, as it allow us to return to the original data at any time to correct the code which saves time and effort.

The most challenging parts in this project are:

- Locating some issues such as the existence of decimals as it exists in small part of the data (5 records). This issue was solved after some searches as other Udacity student faced the same issue and was solved by Udacity reviewer.
- Melting the four columns of the dog stages into one column without any errors, and this issue was solved by testing many codes and some searches.

Analyses & visualization of data

After the cleaning stage is finished, the data is stored in a csv file called ***twitter_archive_master.csv*** and then some analysis and visualization is applied in this file in order to gain some insights about the dataset such as the relationship between the rating of dogs and the number of retweets they get.

Build prediction model

As an extra step of our process, we build a prediction model that aims to predict the breed of the dog based on the data provided to it about these dogs such as:

- Dog stage
- Rating of the dog
- Number of retweets
- Number of favorites
- Confidence level