wrangle_report

October 2, 2022

1 wrangle_report

1.1 Introduction

Real life data doesn't usually come clean, they are always dirty and with some quality issues.

This project is not an exeption as it requires getting data from three different sources in three different format.

1.2 Gathering Data

This project requires three datasets, from different sources and in different formats.

The first dataset is the twiter_enhanced_archive.csv which I downloaded and upload into my workspace. This doesn't require much, I downloaded the file, upload it into the Jupyter Notebook and read it into the DataFrame using pd.read_csv.

The second dataset is a .tsv file named image_prediction.tsv, hosted on the Udacity's server. I downloaded this using the Requests library directly into my workspace and read it into to the DataFrame.

The third and final dataset is a JSON file hosted on Twitter API where I had to extract the number of retweets and likes/favorites.

1.3 Assessing Data

After gathering all the dataset, the next step is to assess the one after the other, both programatically and visually.

These dataset were assessed and the following issues were discovered;

1.3.1 Quality issues

1. Retweet data are included in the twitter_archive_enhanced dataset.

2. **timestamp** is string and should be datetime

3. Some rows have denominators greater than 10

4. Index numbber 979 (tweet_id:749981277374128128) has a rating numerator way too far from 10, that is 1776

5. **name** is inconsistent, as some names start with lower case letters while others start with Upper case and some are invalid names, such as a, an, the and so on

6. **doggo**, **floofer**, **pupper** and **puppo** columns contain 'None' value where 'NaN' should be used.

7. **p1**, **p2** and **p3** are not consistent with naming the dog breeds in terms of capitalising

8. There are different number of observations in the datasets. That is there are 2356 entries in twitter_archive_enhanced.csv, while there are 2075 observations for the image_prediction and the data from twitter api contains 2354 observations.

1.3.2 Tidiness issues

1. **doggo**, **floofer**, **pupper** and **puppo** ought to be one column as they represent the same thing i.e dog stage

2. The three datasets are about only one item, that is tweets. So, they ought to be one.

1.4 Cleaning Data

In an effort to present clean datasets which could be used for analysis, the issues metioned in the assessing data stage were treated one after the other.

This includes; 1. Removing retweets and retweet related information

2. Changing the datatype of columns with wrong datatype,

3. Creating new columns from existing ones, to satisfy our need, and dropping off the ones that are not required,

At the end, the clean datasets were merged together to form a complete dataset and was saved as **twitter_archive_master.csv**