wrangle_report

July 25, 2019

0.0.1 Project: Wrangle and Analyze Data

Which consists of:

1. Gathering data (downloadable file in the Resources tab in the left most panel of Udacity classroom and linked in step 1 below).
2. Assessing data
3. Cleaning data
4. Storing, analyzing, and visualizing your wrangled data
5. Reporting on 1) your data wrangling efforts and 2) your data analyses and visualizations

0.0.2 Gathering Data for this Project

Although, tried to gather data with the help of data scraping, but could'nt do it, so, carried out the Gathering data with the help of downloadable file in the Resources tab in the left most panel of Udacity classroom.

0.1 Assessment for data 'twitter-archive-enhanced.csv'

Cleaning Issues:

1. The retweeted status id should be an Integer and not a float.
2. The retweeted_status_user_id column should be an Integer and not a float.
3. Rating Numerator Contains Highest rating 1776 which is exceptionally high and lowest rating 0 which is not in sync with rating pattern. Hence, we either need to fiugure out correct rating or remove this rating. Similarly, some ratings are less than 10 hence we need to remove such ratings.(Apart from this, lets not carry out any scaling for the rating pattern which is greater than 10 as it is unique in its own way.)
4. Rating Denominator contains Highest rating 170 which should be 10 as per the rating system as it seems and lowest rating is 0 which should be 10.

5. In name column some names are a, an, the etc which is not correct, hence we need to replace it with either name or None.
6. 'In reply to status id' column are type float and in decimals which should be an integer.
7. 'In reply to user id' column are type float, instead it should be an integer.

Tidiness Issues:

1. Doffo, pupper and puppo should come in column with with respective stings instead of four different columns, as the are related to age of the dog.As it can be sipmly descirbed as pupper(young) < puppo(adolescend) < Doffo(matured).(Note not Floofer, because floofer is a different attribute and not much related to age of the dogs.)
2. Unwanted Columns like timestamp, source etc. needs to be removed.

0.1.1 Assessment for 'tweet_json.txt' data:

Cleanliness Issues:

1. Lenght of the text can be determined from 'display_text_range'
2. From 'Full Text' info we can try figure out the male or female information of the dog but, to a certain extent only.

Tidiness Issues:

1. We need to delete some of the unwanted columns like ' contributors', 'coordinates', 'geo' and such 10 - 12 columns are not required. (This task we will perform after joining the three data sets.)

Assesment for data 'image-predictions.tsv': Cleanliness Issue:

1. The names breed/species in column p1, p2 and p3 are written either starting with lower case or upper case. which is not uniform. Instead, the names should start with Upper Case.

0.1.2 Assessment for Combined dataset for all above three Cl data:

0.1.3 Tidiness Issue:

1. We Join All the three Data Sets One by one. First we join df1_clean (i.e.'twitter-archive-enhanced.csv') with df_cleaned (i.e.'tweet_json.txt'). And then we join it with our third dataframe df_1_cleaned (i.e. 'image-predictions.tsv').

2. We make a copy of the above formed data set and remove certain not so important columns (as we discussed before) like for example contributors', 'coordinates', 'geo', 'place', 'id' etc.

==========

In [ ]: