wrangle_report

September 18, 2021

0.1 Reporting: wragle_report

- Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

I started wranling by gathering information from 3 different sources. First by using the given file of tweet archives, and then collecting data from Udacity's predictions by using the python requests library and inserting the data into a pandas DataFrame. I then used the given link to extract the data that is support to be gathered from Twiter's tweepy library.

After gathering the data i started investigating the data set both visually and programmatically as I understood the columns, their values and the reason behind their collection. During this investigation I detected 10 issues in my data 2 of which were issues linked to data tidiness and 8 of which that has to do with the data quality and its dirtiness. I started by fixing the missing values by removing values that have missing links. Then I started tackling tidiness issues by detecting an issue where multiple URLs (varialbles) were stores in on column (expanded_urls), and this had to be stored in a different dataframe due to a rule where one column can only be resnonsible for one viriable, and creating extra rows in the dataframe just for the expanded URLs didnt make sense. Thenmthat there were multiple tables that are used for the same reason and that have shared 'tweet_id' then I gathered them all in one DataFrame (in addition to the URLs DataFrame). After fixing tidiness issues I started fixing quality issues, starting by namely dog breeds being stored in different cases, I unified the cases for dog breeds. Aditionally, all the tweets had negative for the retweeted column, so I went ahead and removed this column as it was not necessary for our application and visualisation. I also noticed that about 745 dogs are missing names and are stored as 'None' strings which is a quality issue as these names should have been stored as nulls. Moreover, dog sizes were stored as 'None' Strings as well, which was fixed by giving them nulls. Note that strictly speaking, the previous part could have also been a tidfiness issue to be fixed by creating one column for this data and storing the size of the dog. I also noticed that there were a number of dogs with random (not 10) denominators, so i went ahead and set them all to 10. I also noticed some dogs with innacurate breed types, so I found a dog breed list from the internet and used it to filter out all non-dog breeds in our data. We also noticed that there were many tweets that has 0 retweet count but given that the account has a large number of interactions and favorites, it seems that this is innacurate, so I removed all rows with retweet count 0.

In [ ]: