wrangle_report

June 23, 2022

0.1 Reporting: wrangle_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.

0.1.1 Data Gathering

I gathered 3 different file formats of data from 3 different sources. -twiitter_archived_enhanced.csv which was already provided; just had to download and read into a dataframe -image-predictions.tsv(the tweet image prediction) which was gotten from its url using Requests method -tweet-json.txt- which was gotten alternatively from the notebook as i have not been granted access to twitter api yet then read into a data frame(dog_count_list).

0.1.2 Data Assessment

I discovered 8 data quality issues and 3 tidiness issues via visual and programmatic assessment(using .info method). #### Quality Issues: ##### Detected Visually -Missing values represented by NaN and None.

-Mixture of upper and lower case letters in p1, p2, and p3 columns.

Detected Programmatically -Date and time columns in string type.
-Extract correct ratings from text column.
-Some denominator ratings are less than 10
-Tweet Id is integer type.
-Name called a and an.
-Source of tweets between anchor tags.

Tidy Issues: -tweet_id duplicated across all 3 tables.
-Retweeted columns.
-Date and time in a column
-Dog Stages in multiple columns.

0.1.3 Data Cleaning

Prior to cleaning, I made a copy of each dataframes then began cleaning with the missing values represented by None, changing it to NaN.

Then moved on to the tidy issues dropping the retweeted columns , splitting timestamp into separate date and time columns and converting the dog stages into a single column.

Prior to addressing the quality issues, I merged all 3 dataframes on tweet_id, then converted date and time columns to datetime type, extracted the actual numerator ratings(fraction_numerator) from the text column, changed all the denominator ratings to 10, converted tweet_id to string type, replaced names called a and an with NaN(missing values) , extracted the source of tweets from the anchor tags and changed all upper and lower case letters in the p1, p2 and p3 columns to lower case.

Data Storage I stored the cleaned data to a csv file named twitter_archive_master.csv.

```
In [ ]:
```