wrangle_report

September 14, 2022

1 WRANGLE REPORT FOR WERATEDOGS

This report contains my wrangling efforts for the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. I divided the wrangling into three parts; gathering, assessing and cleaning.

1.0.1 Gathering

I made use of the workspace provided by Udacity which is a jupyter notebook assessed through the course. I started by importing the libraries which I made use of during the course of the project such as pandas, numpy, requests, tweepy and others. I had to work with three datasets. The first dataset was twitter_archive_enhanced.csv, a comma separated variable file which was already made available on the workspace. The second dataset was image_predictions.tsv, a tab separated variable file which I had to query Udacity servers for and download programmatically using the requests library. The third dataset was an additional data I had to query twitter API for by applying for a twitter developer account and make use of a library named tweepy to download this data programmatically and save as tweet_json.txt. After gathering these datasets, I read the datasets into respective pandas dataframes using the pd.DataFrame(file) format.

1.0.2 Assessing

I assessed the datasets both visually and (primarily) programmatically. Visual assessment: I displayed the datasets and scrolled to get familiar with the columns and sight any quality/tidiness issue. I also made use of the .sample() function to visually assess random samples from the datasets. Programmatic assessment: I made use of various pandas functions such as .info(), .isnull(), .isnotnull(), .unique(), .nunique(), .duplicated(), .where() and others to assess the datasets. At the end of the assessment I was able to point out 8 quality issues and 2 tidiness issues. These are: #### Quality issues df_archive_tweets 1. table contained retweeted tweets

2. retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp are not useful and should be dropped
3. some decimal ratings have been incorrectly extracted
4. some ratings are not correct
5. some ratings are inconsistent
6. timestamp should be a datetime object

7. some tweets do not contain ratings
8. some rows have rating numerator as 0

Tidiness issues

1. doggo, floofer, pupper and puppo have been separated into different columns
2. all 3 datasets should be merged into one master dataset and all duplicate columns be dropped

1.0.3 Cleaning

I cleaned the datasets by tackling the quality and tidiness issues already mentioned.

- I made copies of the datasets before carrying out cleaning operations
- I dropped rows with unclean data(incomplete, invalid or not useful)
- I dropped off the columns that weren't useful
- I reextracted the ratings to account for decimal ratings also
- I cleaned the rating columns of incorrect, inconsistent and missing values
- I reassigned correct datatypes to columns with wrong datatypes
- I merged the doggo, floofer, pupper and puppo columns into a single column(stages) using the pandas.melt function
- I merged all three clean datasets into a single master dataset named "twitter_archive_master" and saved it in a file named "twitter_archive_master.csv"

At the end of the cleaning stage, my dataset was clean and ready for EDA.