wrangle_report

February 21, 2023

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.

0.2 Data Gathering

In `Data Gathering`, i have to load data from severals sources using multiple steps,i.e, read csv data using `pd.read_csv()` , Request data content by using `request.get(url)` then write into some files by using `with open('filename.txt', mode="wb") as file :`, and retrive data from twitter API by using `tweepy` even though i got verification issues for accessing `Auth` Twitter API.

0.3 Assessing Data

After the Data is already gathering as Data Frame, we have to assessing the data by checking data type, missing values, and the readiness of the data. after, i was already assessing the data, i got 8 quality issuess. i.e: ### Quality issues 1. Tweet ID , id_str, and Status ID have to be formated as object

2. All timestamp column have to be formated as datetime64 format
3. Convert 'None' into None/NaN values using `np.nan()`
4. exclude retweet id
5. Missing Expanded URL
6. Replace "a" and "an" to `None` in name column of twitter archive
7. Merge columns Doggo, Floofer, Pupper and Puppo into single column
8. Timestamp heve to be filltered before 1st Aug 2017

0.3.1 Tidiness issues

1. Melt column doggo, puppo, pupper, and floofe into single column
2. Change tweet ID into Object data type for joining data with other tables

0.4 Data Insight

After doing quality issuess and tidiness of the data. i got some insight and visualization ### Insights: 1.golden retriever's dog has the most favorited pic in twitter's dog tweet

2.pupper's dog has the most count of values in twitter tweet

3.p1_conf and p2_conf have corelation with rating_ratio