wrangle_report

September 1, 2022

0.1 Reporting: wrangle_report

0.1.1 Data Gathering

In the wrangling part of this project, I **firstly** manually downloaded The WeRateDogs Twitter archive and then uploaded it to the project workspace. Once uploaded, I used pandas library to read the csv file. **Second**, I have downloaded the tweet image pedictions programmatically using the Requests library and stored the TSV file in a new folder in the workspace. **Third**, I used Tweepy to query the Twitter API after generating the Consumer API keys,the Access Token and Access Token Secret. Then, I stored the JSON data in tweet_json.txt file. The text file was afterward read using the open() and readlines() functions. Finally, the tweet_id, retweet_count and favorite_count were extracted from the file and stored into a dataframe df_tweets.

0.1.2 Data Assessing

In the assessing part, I used both visual and programmatic assessments that led to eight quality issues and five tidiness issues.

Visually, the data was displayed in excel to get a better view for assessment. Five issues were spotted:

1. The values in retweeted_status_timestamp and timestamp columns contained four zero digits at the end.
2. Invalid dog names in 'name' column such as 'a', 'an', 'the'. Along the way, some wrong names were also found and were corrected such as 'Johm' to 'John' and 'Jessiga' to 'Jessica'.
3. Inconsistency of denominator_rating values (not 10 in some cases).
4. Inaccurate numerator_rating values (higher than 100 and 1000).
5. Missing dog names (not extracted from the text).

Programmatically, I used the info() function to detect basically erroneous datatypes. Three issues were detected:

1. Datatype is object for timestamp,retweeted_status_timestamp instead of datetime.
2. Datatype is float for status_id and user_ids columns instead of int.
3. Datatype of dog 'stages' is object instead of category.

0.1.3 Data Cleaning

For data cleaning, I went through the issues one by one. For the issues detected **visually**, firstly, I removed the unwanted characters (four zeros) from the retweeted_status_timestamp and timestamp columns using 'rstrip' function.

For issue #2, I created a function 'extract_name' that extracts the dog name from the text and when possible replace the prepositions 'a', 'an, 'the' with the right name, else the value is set to None. As I was fixing this issue, I found some names that needed further correction such as "Johm" to "John" and "Jessiga" to "Jessica". Afterwards, I set the values of denominator_rating column to 10 as the rating system of WeRateDogs is usually value/10 (issue #3). For issue #4, I used data query to return the records where numerator_rating is higher than 20 => 25 rows had invalid values that were either wrongly extracted from the text, or given collectively (rate more than one dog). These values were corrected one by one. However, some were not possible to be replaced as they were given by the user that way (considered as outliers). For issue #5,unlike what's done in the second issue, here, I looked for the missing names (value = None), and tried to extract the dog name if possible. Otherwise, the row is dropped.

Programmatically, the main used funtion is info() where all the erroneous datatypes were converted to the right ones: retweeted_status_timestamp and timestamp columns from object to datetime, status_id and user_ids from float to integer and finally dog_stage from object to category.

In [ ]: