wrangle_report

September 12, 2022

0.1 Project on WeRateDogs Twitter Data.

0.1.1 Wrangle Report

The dataset wrangled in the project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

The WeRateDogs Twitter project goals included: Wrangling the twitter data through the following processes: - Gathering data - Assessing data - Cleaning data - Storing, analyzing, and visualizing your wrangled data - Reporting on the data wrangling efforts and data analyses and visualizations

1. Gathering the Data Three sources of data were gathered to complete this project:

1. Twitter enhanced archive file - csv downloaded from udacity course site, which includes various information about tweets of WeRateDogs account
2. Tweet image predictions file - tsv requested from udacity site, which includes predictions of objects on images included in WeRateDogs tweets
3. Tweet details file - json file downloaded from twitter using API, which includes information missing from the enhanced archive file, namely retweets and likes counts

Data was gathered using different methods: - csv was read from a file and stored in archive - tsv was downloaded using the requests library, written to a file and stored in images - json was gathered using twitter API, stored as txt file, loaded using json library and finally saved as tweet_tables

2. Assessing the Data After gathering the above three dataframes we moved straight to assessing them for quality and tidiness issues, in which we uncovered nine quality issues and two tidiness issues. They are as follows:

The two types of Data Assessment performed

- Visual assessment: Each piece of gathered data is displayed in the Jupyter Notebook. Once displayed, data are additionally assessed in an external application (Excel, text file reader)
- Programmatic assessment: pandas' functions and/or methods are used to assess the data.

Quality Issues

- tweet_id is an integer

- timestamp is of 'object' datatype.
- name has values that are string 'None' instead of NaN and some values have unusual names of less than 3 characters such as 'a'.
- NaNs represented as 'None' (str) for name, doggo, floofer, pupper, and puppo columns.
- Archive data contains retweets.
- Some of the ratings are wrongly mentioned e.g. in one case, the rating should've been 13/10, not 960/00, while some tweets are not about dogs, so doesn't contain rating (tweets not containing dog images can be discarded), and some of the ratings contain decimal in the numerator.
- There are some missing rows in images dataset (2075 rows instead of 2356): either the rows are missing or some tweets didn't have dog images.
- There are some duplicate jpg_urls.
- p1, p2, and p3 contains underscores instead of spaces in the string.

Tidiness Issues

- The different dataframes should be merged into a single one.
- There are 4 different columns (doggo, floofer, pupper, and puppo) for dog stages.

3. Cleaning the Data First Step: I have copied all the three DataFrames using .copy() method,

- archive_clean = archive.copy()
- images_clean = images.copy()
- tweets_clean = tweets_table.copy()

Further Steps: - This is the action stage. here I put all of the above observations into action and used the the given cleaning mehtod to adreess all the above issues. - Taking each issue separately, I used the Define, Code and Test method to firstly tackle the tidiness issues as they are wider and cover the whole dataset, then follow up with each of the quality issues. - Each issue was handled separately using the define-code-test workflow. - The cleaning efforts finished with storing the final merged data set as twitter_archive_master.csv.