wrangle_report

November 8, 2022

1 Reporting: wragle_report By Ashim Sharma

2 Introduction:

In order to record the project's data wrangling efforts, this wrangle report is a component of the Wrangle and Analyze Data project. The WeRateDogs account on Twitter, @dog rates, provided the dataset for this project. A Twitter account called WeRateDogs awards stars to users' pets along with amusing comments about them. The three processes of data wrangling—data collection, assessment, and cleaning—are documented in the wrangle report.

2.1 1. Data Collection

- I must acquire information for my assignment from several sources and in a variety of forms. The archive of tweets from WeRateDogs. The project provides the file, which is available for direct download from the Udacity website.
- The predicted tweet images. The file is stored on servers owned by Udacity. I automated the download of this file using Python's Requests package.
- Retrieve from another file the retweet and favorite counts that are absent from the Twitter archive. Since I don't have a Twitter account, I decided to obtain the tweet JSON file programmatically using the Requests package.

2.2 2. Evaluating data

After collecting the data, I evaluated it programmatically and visually to find any problems with data quality or organization. Tidiness pertains to data structure, whereas quality relates to content. Requirements for orderly data: each variable should be represented by a column, each observation by a row, and each sort of observational unit by a table. The files weren't that big, so I could load them in EXCEL and skim through the data to look for any obvious problems. To view particular subsets and summaries of the data, I also utilized code in Jupyter Notebook, such as the pandas info, head, sample, value counts, duplicated, query, and description methods. While analyzing the data, I took notes on the findings so that I could address the problems later.

3 Quality Issues:

3.1 Findings and Summaries:

3.1.1 Twitter Archieve Table:

Only original ratings with photos are required; retweets and responses entries should be eliminated, along with related columns. The following columns in the twitter archive table should all be str: in reply to status id, in reply to user id, retweeted status id, and retweeted status user id. Later, we'll fix the picture component. Remove +0000 from timestamp aberrant numbers in rating denominator, such as 170, 150, 130, etc., and timestamp is str, should be datetime. Most frequently, the rating denominator There are 10 aberrant values in the rating numerator that are illogical, such as 1776, 960, 666, 204, 165, etc. ### Image Prediction table: - data sources redundant and challenging to read table with predicted images p1, p2, and p3 columns have erroneous capitalisation and duplicate jpg urls - There are numerous entries that aren't dogs, such as a jaguar, postbox, peacock, cloak, etc. - For this investigation, only the most certain dog breed predictions will do. ### Tweet json Table: - missing data in twitter archive perhaps as a result of retweets ### Uniformity and Tidiness of data: - Twitter archive organization: doggo, floofer, pupper, and puppo are all dog stage names, and they should all be in one column.

- According to the standards for clean data, the three tables should be consolidated into one because they are all connected to the same kind of observational unit.

3.2 3.Removing data

- While assessing, I cleaned each of the concerns listed. Even though the entire dataset has a lot of problems, fixing them all would take a lot of time. I therefore concentrated only on those relevant to my analysis. The three steps of the programmatic data cleaning process are define, code, and test. Before cleaning, it's also crucial to make copies of the original data. Because certain concerns will vanish when we clear them one at a time, it's crucial to address them in a logical order. For instance, once we fixed the problem that many photos in the image prediction are not dogs, several of the odd rating numerators and denominators we earlier observed vanished.
- Some of the anomalous ratings automatically disappeared when I eliminated posts that weren't about dogs. Another illustration is the datatype problems with in reply to status id, in reply to user id, retweeted status id, and retweeted status user id. All of the variables should be typed as ints, but they are all typed as floats. Retweets and replies will be removed because the project only needs original tweets. After eliminating those columns, the datatype issue was once more no longer a problem. The majority of cleanings were completed using programmatic tools, however some cleanings required manual intervention, such as rectifying the erroneous rating denominators and numerators. To discover the correct ratings, I had to read through the language and filter out odd ratings.
- Reevaluate the dataset and iterate as necessary after all the issues have been resolved. Next, keep clean data in a CSV file.

In [ ]: