wrangle_report

April 20, 2023

1 Reporting: Wrangle_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.

1.1 Introduction

This project's purpose was to wrangle WeRateDogs Twitter data in order to make engaging and trustworthy analysis and visualisations. The challenge was in the fact that the Twitter archive was great, but it only contained very basic tweet information. Additional gathering, assessing, and cleaning was required to achieve the desired outcomes.

1.2 Data Gathering

Three pieces of data were gathered for this project:

- **Twitter archive data** (twitter_archive_enhanced.csv) was downloaded directly.
- **tweet image predictions** (image_predictions.tsv) were downloaded programmatically using the Requests library from a provided URL.
- **Additional data**, such as retweet count and favorite count, were gathered using the Tweepy library to query the Twitter API (tweet_json.txt). However, since the Twitter API access was not granted, the tweet-json.txt file was downloaded directly and parsed to extract the relevant data.

1.3 Data Assessing

The data was assessed both visually and programmatically to identify quality and tidiness issues. In total, eight quality issues and two tidiness issues were identified.

1.3.1 Quality Issues

Twitter Archive Table - Timestamp column was in string format instead of datetime. - The dataset contained 181 retweets and 78 replies, which were not needed for the analysis. - Missing values in the expanded_urls column. - Extreme numbers in rating denominators, with 23 denominators not being 10. - Extreme numbers in rating numerators. - Some dogs' names were invalid (e.g., None, a, such, etc.).

Image Predictions Table - Some photos were not identified as dogs (e.g., orange, bagel, banana), with p#_dog = False. - Inconsistent capitalization in names of dog breeds.

1.3.2 Tidiness Issues

- Dog stage data in the tweet archive table was divided into columns doggo, floofer, pupper, and puppo.
- Some columns related to analysis in the image predictions and tweet JSON tables needed to be merged into the main archive table.

1.4 Data Cleaning

Before cleaning the data, copies of the original datasets were created. Each identified issue was then addressed one by one, following the Define-Code-Test cycle.

- Retweets and replies were removed from the dataset, and the corresponding columns were dropped.
- The timestamp column's data type was converted to datetime.
- Rows with missing values in the expanded_urls column were dropped.
- Invalid dog names were replaced with 'None'.
- A new column 'dog_stage' was created to consolidate the four separate dog stage columns.
- Rating numerators and denominators were extracted from the text, and extreme values were adjusted accordingly.
- The highest confidence prediction for dog breeds was used to create a 'breed' column, and photos not identified as dogs were filtered out.
- The capitalization in the names of dog breeds was normalized.

Finally, the cleaned datasets were merged into a single, tidy master dataset for further analysis and visualization.

1.5 Conclusion

Data wrangling is an important step in the data analysis process. It involves gathering, assessing, and cleaning data to ensure that it is both high-quality and tidy. Multiple datasets were obtained from various sources and formats for this project, analyzed for quality and tidiness issues, and cleaned accordingly. The clean dataset that resulted allowed for the production of informative analysis and visualizations.

```
[ ]:
```