wrangle_report

February 7, 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.

In this report, I will describe my wrangling efforts for the Udacity data wrangling project. The data for this project was sourced from the WeRateDogs Twitter account, which rates dogs and provides humorous commentary on the ratings. The data consisted of three different sources:

1. A twitter archive file that contained basic information about tweets such as tweet ID, timestamp, and text.

2. A file of image predictions that contained predictions of what was in the images of tweets, such as dog breeds.

3. Additional data that was gathered via the Twitter API, such as the number of retweets and favorites for each tweet.

My first step in the wrangling process was to assess the data for quality and tidiness issues. I found several issues in the data, including:

Missing data in the image predictions file and additional data from the Twitter API.

Incorrect data types for some columns, such as the timestamp column being a string rather than a datetime object.

Inaccurate dog ratings in the text column, as some ratings were not in the format of "x/10".

Inconsistencies in the dog breed predictions, as some predictions were not the name of a breed, but rather a phrase such as "not a dog".

To address these issues, I performed a series of cleaning actions. For missing data, I dropped the rows that had missing values in the image predictions file and additional data from the Twitter API. For incorrect data types, I converted the timestamp column to a datetime object. To handle inaccurate dog ratings, I extracted the rating numerator and denominator from the text column and created new columns for them, then I dropped the text column. And to fix the breed predictions I replaced it with a breed if the breed has a high confidence or if not I put not a dog or other.

Next, I tackled the tidiness issues in the data. The main tidiness issue I found was that the dog "stage" (e.g. "pupper", "doggo") was spread across multiple columns in the twitter archive file. To fix this, I created a new column called "stage" and consolidated the information from the other columns into it.

Finally, I exported the cleaned data to a new csv file for further analysis.

Overall, the wrangling process for this project was challenging, but it allowed me to gain a deeper understanding of the data and improve its quality and tidiness. Through the cleaning process I was able to remove unnecessary columns and data, fix inaccuracies, and make the data more usable for analysis.