wrangle_report

November 2, 2022

0.1 Reporting: wrangle_report

INTRODUCTION In this section of the wrangling report, I'll briefly go over how I uncovered all the data quality and tidiness problems in the dataset and how I went about fixing them.

DISCRIPTION

After loading all three datasets required for this project, I proceeded to identify data quality and tidiness issues, where I found the following: I investigated the top 5 rows of the datasets using both programmatic and manual assessment. I noticed many columns in the data which contained missing values, while some missing values were misrepresented, such as "None". I solved these issues using the pandas replace() method to replace all instances of the "None" value with N/A in the dataset. By going through the details of the loaded Twitter archived data using programmatic assessment, I noticed the following issues: o Tweet id is loaded as integer type instead of string, and this happens because tweet id contains bunch of numbers. o The timestamp column is a string, instead of a datetime object. I solved the erroneous data type using both astype() and pandas.to_datetime() functions to change both the tweet id and timestamp column to their accurate format. Since the analysis did not require using retweeted and replies to tweets, I removed all the retweeted and replied tweets from the dataset.

Moving forward, I checked the descriptive and analytical summary of the dataset. This is where I identify the following issues: o Abnormal rating's numerator and denominator.. o Inconsistent dog breed names, i.e., separated with underscore instead of space, and some breed types were capitalized while others were lowercase.

I noticed that all dogs with unknown names have the same pattern, i.e., lowercase, so I dropped all dog names with lowercase. I solved the issue of inconsistent breed type by replacing underscore with space and also capitalizing all dogs' breeds.

CONCLUSION

A significant number of data observations were lost during the data cleaning step. The entirety of the dataset is clean and devoid of poor data quality issues. All issues were adequately addressed, and well-documented in the notebook.

In [ ]: