wrangle_report.pdf

July 24, 2022

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.

0.1.1 WRANGLE REPORT

This report briefly describes the data wrnagling efforts exercised in this project

The dataset that was wrangled (and analyzed and visualized) 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 comments about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

This project was completed on the udacity project workspace.

The wrangling process includes the following: 1. Gathering Data 2. Accessing Data 3. Cleaning Data

1. GATHERING DATA The datasets used in this project was gathered from three diiferent sources;

A) Enhanced Twitter Archive

This dataset was provided by Udacity for this project so it was downloaded manually into the workspace. The data provides us with the rating of the dogs, dog name, etc.

B) Image Prediction Dataset

This dataset was present in each tweet according to a neural network. It was hosted on Udacity's servers and it was downloaded programmatically using the Requests library

C) Additional data from Twitter API

This dataset was gotten via the one provided by udacity, that is, the tweet json txt file provided by Udacity. This was because I could not get a developer acc from twitter to have access to twitter API, so I was advised to use the file provided by Udacity to carry out the project

2. ASSESSING DATA After gathering the datasets needed for this project, the data sets were assessed visually and programmatically for quality and tidiness issues. Therefore the following findings were made;

A) Quality Issues

Enhanced Twitter Archive

1. Tweet_id column is integer datatype instead of string datatype
2. Timestamp column is string datatype instead of datetime datatype
3. The rating_denominator has some values higher than 10
4. Some dog names are incorrect e.g. a, an, the, etc
5. There are 181 retweets and they are not needed, just tweets are needed

Tweet Image Prediction 6. There is inconsistency of upper/lower case in the p1, p2, p3 columns 7. Underscores are used in names instead of spaces in the p1, p2, p3 columns 8. There are some missing photos for some tweet_ids (2075 rows instead of 2356) 9. Tweet_id column datatype is integer

Tweet Data From Twitter API 10. The tweet_id column datatype is integer

B) Tidiness Issues

1. doggo, floofer, pupper, puppo have their own column each and they can all be in one column
2. All the 3 dataframes are related but they are seperated

3. CLEANING DATA The cleaning process was done after listing out the issues and the Quality of the datasets were increased

```
In [ ]:
```