Report on Data Wrangling

Objectives: The main objectives of the project were to perform data wrangling, store, analyze, and visualize the cleaned data, and report on the data wrangling efforts, data analyses, and visualizations.

Step 1: Gathering Data

To gather the data, three different sources were used, and the data was represented as pandas dataframes. The WeRateDogs Twitter archive was already available as a file, while the tweet image predictions file was downloaded programmatically using the Requests library from a provided URL. For each tweet's entire set of JSON data, Twitter API and Python's Tweepy library were used to store the data in a file called 'tweet_json.txt'. Each tweet's JSON data was written to its own line.

Step 2 and 3: Assessing and Cleaning Data

During the data wrangling process, several observations were made about the data that required cleaning. Below are some of the observations made and the corresponding actions taken during the cleaning step.

Quality issues

Twitter archive dataframe

1/ tweet_id should be string not int

2/ The time stamp should be date time not object

3/ We should delete unecessary columns in the dataset

4/ drop rating denominator that's different than 10

5/ delete data after 1st of Aug 1st, 2017

Image predictions dataframe

1/ tweet_id should be string not int

Tweet data dataframe

1/ id should be string not int

2/id should be named tweet_id

Tidiness issues

1/ Columns doggo,floofer, pupper and puppo should be combined in one column

2/ combine all dataframes into one dataframe

Result

A new dataframe made and ready for analyzing.