Data Wrangling Report for the WeRateDogs Data

1 Objective

The objective of this project is to wrangle and clean data from the WeRateDogs Twitter account.

2 Project Details

In this project we gather data from different sources, assess and clean it. To do this we use data from the WeRateDogs twitter account.

We are given a CSV file containing data from the twitter archive. This file contains data like rating, name and stage of a dog.
In addition to this file we use the twitter api to receive more data like the retweet and favorite count.

The last source is an image prediction file we load from the internet.

3 Gathering Data

The first step was to download the 'twitter_archive_enhanced.csv' file manually and read it into a DataFrame. Next i downloaded the 'image-predictions.tsv' file and created another DataFrame from this file. For the last gathering step i queried twitters api to get the retweet and favorites count for the tweets. For this purpose i used the tweepy library. The received tweets were stored in an empty list and after i received all tweets i converted them to dataframe which i saved on disk. The result was saved to a file named 'tweet_json.txt'.

4 Assessing Data

After the three tables were in place i assessed the data visually by using a Jupyter Notebook and programmatically by using Pandas built in methods like info, value_counts, duplicated etc.

Then i grouped the issues in quality and tidiness issues.

5 Cleaning Data

For each issue this part was divided intp three steps.

- Define how to clean the issue.
- Write the cleaning code.
- Test the result

Before starting any cleaning i created a copy of the three dataframes, So i kept the original ones and could easily revert any changes i made in case there was an error during my cleaning efforts.

After i finished my cleaning steps i stored the resulting dataframe as a csv file.

6 Conclusion

Data wrangling and cleaning is an import process and a skill everyone who has to deal with data should be familiar with.