Reporting for project

This project is on WeRateDogs twitter. I have some archive data, but this data is not complete. I aim to add data from other sources and makes this information qualitative.

Gathering Data for this Project

I had data from three sources. With CSV-file and TSV-file I didn't have any problem.

The source number three was Twitter API. First I used tweet IDs from the CSV-archive to query JSON-data and to write them in JSON-file. Then I rewrite the code to use the difference between IDs from the CSV-archive and IDs from stored JSON-file and then to add new lines to JSON-file. Only five tweet IDs didn't have a response from Twitter API.

When I was reading JSON-file to Data Frame, I had a problem with integer numbers. If the column with integer numbers have some missed data, then the type of column will be changed to float numbers and some long integers will be converted to float number with losing of information.

I had to write a code for download JSON-file to Data Frame.

Assessing Data for this Project

I created two service variables for recording my thinks about assessing data. It helped me documenting my assessments during exploring my data and print all records at the end.

I made of assessment only Data Frame from CSV-file because other two sources have given good data.

Cleaning Data for this Project

First I deleted some columns that include corrupted information and this data I can take from Twitter API.

I decided to merge all data in one Data Frame because all rows in data had an ID of tweets and this ID was unique in own Data Frame. In the end, I fixed all issues that I documented while assessing.

Storing, Analyzing, and Visualizing Data for this Project

I stored the clean master Data Frame in a CSV-file.

Also, I made some analysis of my data.

This twitter is WeRateDogs, so I started with the rating. I looked at activity on this account. I pictured scatter plot for retweets and likes. Also, I made clouds of dogs names and showed top ten.

In the end, I played with a prediction the breed data.