Data Wrangling (Twitter Account WeRateDogs)

Wrangle report

A data wrangling project on a twitter account dataset called WeRateDogs was gathered, cleaned, stored and assessed in this report.

Gathering Data

The data was gathered by downloading three files in different ways:

1) twitter-archive-enhanced.csv that is an access to the twitter account archive dataset and has many data such as timestamp, tweet id, name and etc. also the downloading was manually using pandas library.
2) image_predictions.tsv that is an image prediction dataset that has many data such as tweet id, img_num ,jpg_url and etc. also the downloading was through a link and using request library.
3) tweet_json.txt that is a dataset for the retweet count, favorite count and tweet id that was queried through Twitter API and tweepy library and this one was not easy to do like others.

Assessing Data

The assessing was through discovering a number of quality and tidiness issues that must be solved in cleaning data.
Quality issues:
1) removing retweets columns and retweets data
2) change timestamp type to datetime and tweet id type to string
3) change rating numerators and denominators types to float
4) remove rows where there are no images
5) remove unused columns from the dataset
6) rename the id column to tweet id to be similar with other datasets
7) Change values in name from None to NaN
8) remove the underscore in p1,p2 and p3
Tidiness issues:
1) merging df, df_img and tweetData after cleaning
2) combine doggo, floofer, pupper and puppo into one column called dogStage

Cleaning Data

The above quality and tidiness issues was solved in this section. In the beginning a copy of each dataset was made and step by step all issues were cleaned and checked. At the end a merge between all three datasets was made.

Storing Data

The merging dataset of the three datasets was stored to a csv file named twitter_archive_master.csv.