WeRateDogs data wrangling

I divided data wrangling process to 4 stages: Gathering the data, assessing the data, cleaning the data and analyzing the data.

At first stage I started to download all the necessary files. After downloading the “twitter_archive_enhanced.csv” csv file manually, I created a folder named “Project” and downloaded “image_predictions_tsv” file programmatically using requests library to this folder. Then through the tweepy library I downloaded Twitters’ JSON data and created “twitter_data” file.

Second stage was assessing the data, I have identified several quality and tidiness issues.

Quality issues are:

- "twitter_archive" and "image_prediction" - twitter_id column should be string
- "twitter_archive" - timestamp column's data type should be datetime
- "twitter_archive" - in dog stage columns null values is not correctly displayed
- "twitter_archive" - Delete retweets
- "twitter_archive" - some columns miss lot of values (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_timestamp)
- "twitter_archive" - keep original tweets with images
- "twitter_archive" - there are inappropriate names in dog name column
- "twitter_archive" - Unnecessary columns
- "twitter_archive" text column includes links
- "twitter_archive" The standard denominator for "rating_denominator" is 10
- "image_prediction" - Capitalize the dog names in p1, p2, p3 columns

Tidiness issues are:

- "twitter_archive" - dog stages should be under one column
- “Image_prediction” and “twitter_data” tables should be merged to “twitter_archive” table

At the third stage of the wrangling process, I cleaned the data through solving the quality and tidiness issues. Finally, I saved the new data frame as “twitter_archive_master.csv” file.

Finally, I started to analysis the last version of the data. First, I looked how the number of retweets changed from the beginning of the account. Then relation between the favorite and retweet counts are displayed with scatter plot. Third insight was to identify the most popular dog breeds in dataset. Next step was to find out the dog breeds which are the most favorable. These results are showed with bar chart. The result of Top five dog names is also visualized with bar chart. At the last stage of analysis most rated dog are identified.