WeRateDogs Data Wrangling Report

This project was very challenging, but it has allowed me to get a good grasp of data wrangling techniques and also added to my experience in python programming. I'm now more confident of my data wrangling and python skills.

In the following few lines, I'm going to summarize the work done and what I have learned. Starting by data gathering, I have learned how to use the twitter API to collect tweet data using tweepy library. It took me some time to figure out how to setup the API configuration for the first time, until I came across this page https://www.digitalocean.com/community/tutorials/how-to-authenticate-a-python-application-with-twitter-using-tweepy-on-ubuntu-14-04 , which explains the necessary steps in a very straight forward way. Collecting the tweets using the API took more than an hour and finally revealed only 1758 entries, so I decided to use the original tweet-json.txt file provided by Udacity to have more data records for better analysis.

After gathering the needed data, which consists of three main sources, the WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, the JSON file collected using tweepy where the retweet_counts and favored_counts were extracted, and the Image predictions file that contains the predicted breed alongside each tweet ID and image URL. I stored them in dataframes and I started assessing them using both visual and programmatical methods, where I pointed out several quality and tidiness issues. However, iterating over the code, I discovered more issues, but fixing them all would require more time.

Before cleaning the dataframes, I made a backup copy to keep the original data as a reference point. In the cleaning phase I utilized many of the pandas methods, like merge, extract along with regular expressions, apply, iloc and loc to set specific cells values. Using loc and iloc was also tricky in the beginning but then I mastered it. Using regular expressions to extract specific patterns was a lot of fun. Finally, I used seaborn library in plotting the scatter matrix to explore the relations between the different variables and plot summary visualizations.

In a nutshell, this project was one of the most comprehensive projects I worked on and put a lot of efforts in, but the knowledge I gained through completing it was so rewarding.