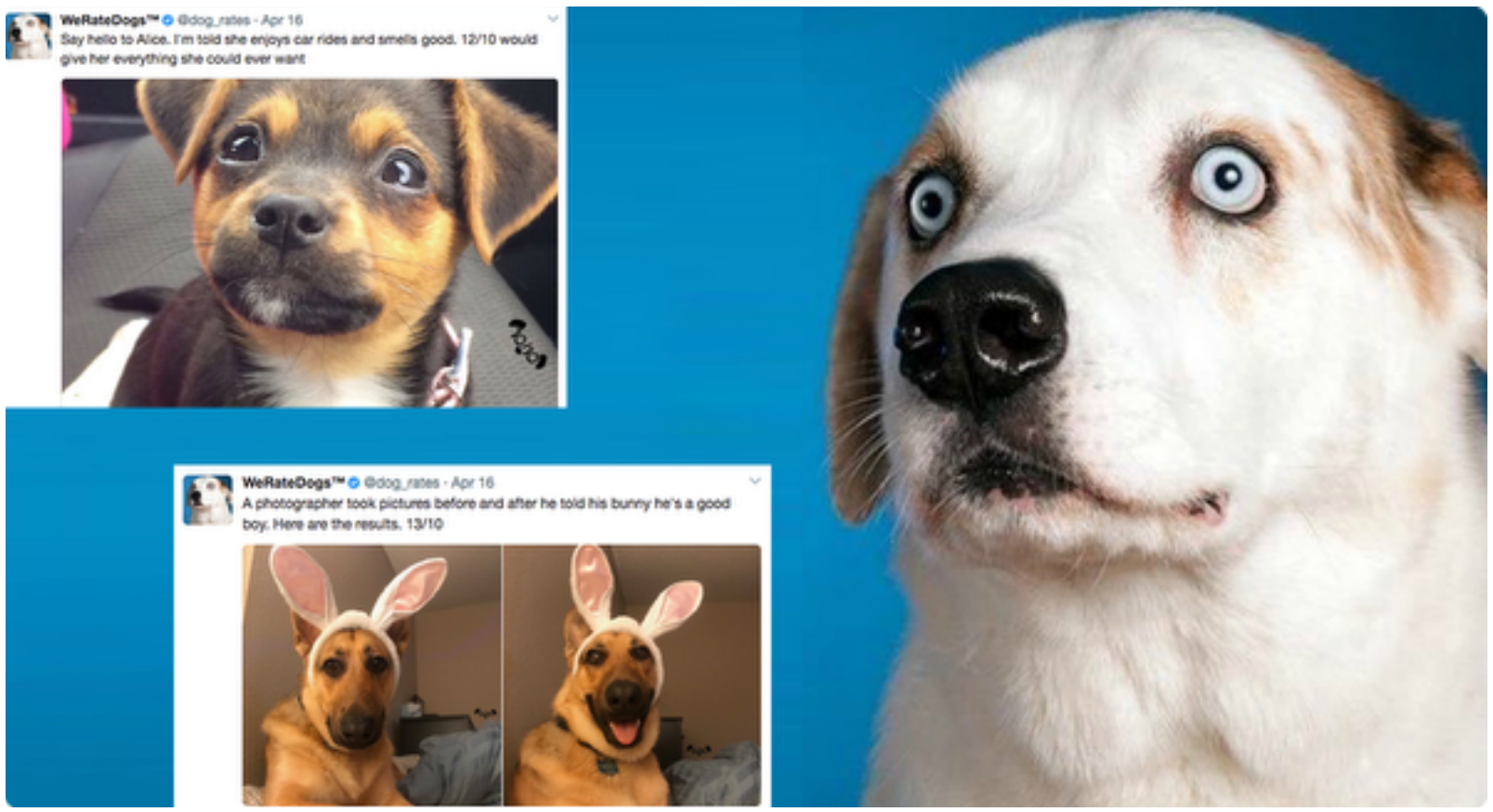
Udacity

Data Wrangling Project



Introduction

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. You will document your wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL.

The dataset that you will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user **@dog_rates**, also known as **WeRateDogs**. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "**they're good dogs Brent**." WeRateDogs has over 4 million followers and has received international media coverage.

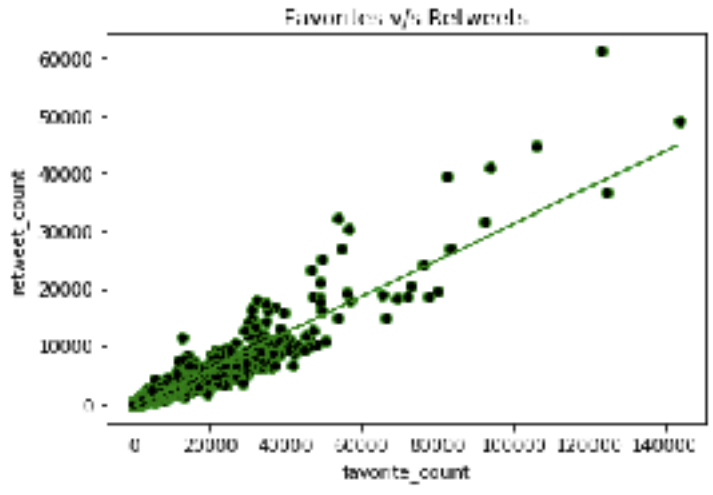
WeRateDogs **downloaded their Twitter archive** and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. More on this soon.

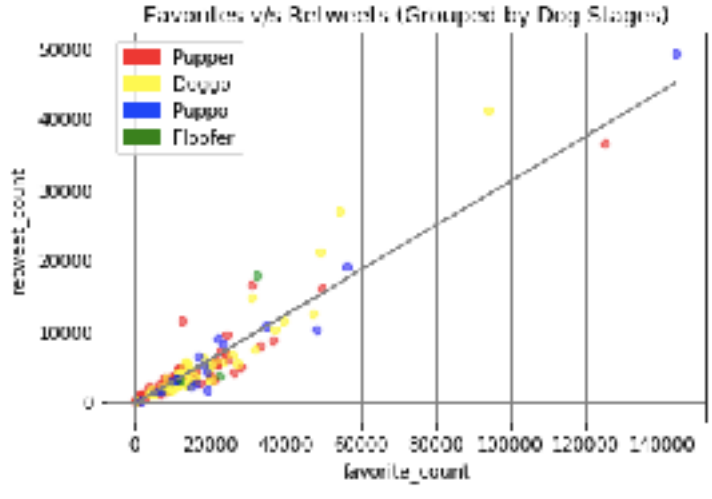
Udacity

Data Visualization Highlights

The cleaned and enhanced data was further acted upon to visualise and infer important takeaways about the tweet data set.

Sample Graph 1





Plotting a scatterplot of all tweets from final dataframe

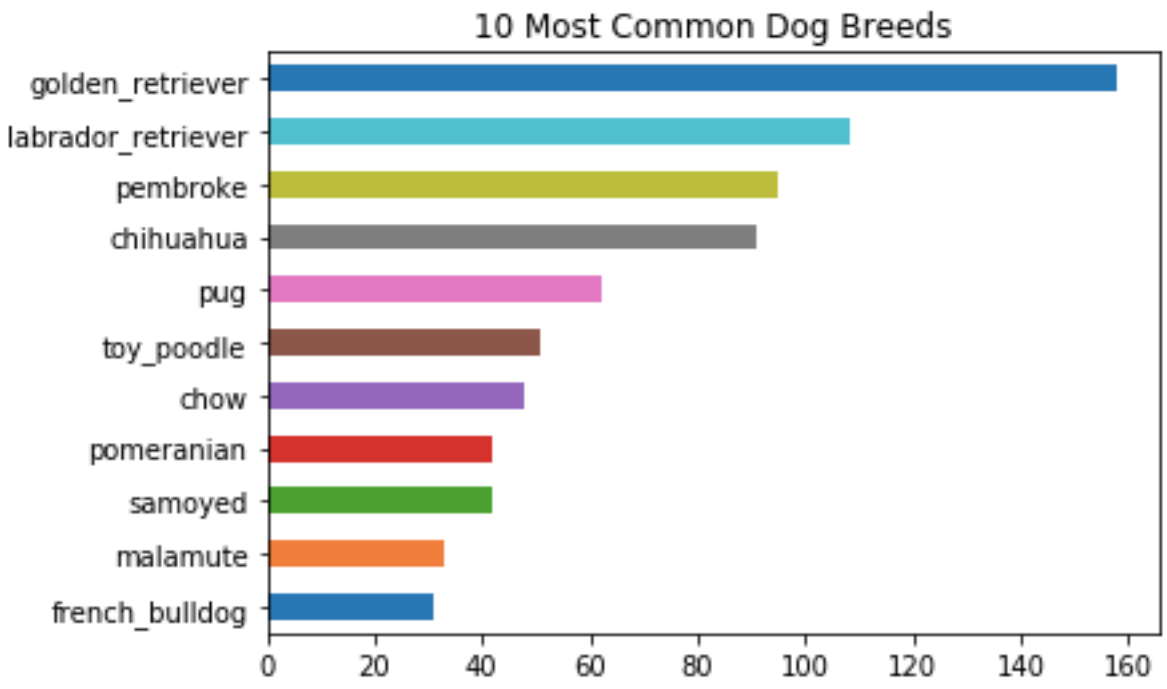
Few of the observations from this plot are as follows:

- Bulk of the tweets happens to have less than 40,000 favorites.

- Bulk of the tweets happen to have less than 10,000 retweets.

- From the line of best fit, the ratio of favorites to retweets tends to be 3:1 or more in majority cases. For the larger values, this may not hold though.

Sample Graph 2



The plot above illustrates the following:

- Close to 500 tweets were classified as not a dog, and hence excluded from this plot.

- Possibble reason for that maybe either fudgy images, image recognition algorithm failing or just objects other than dog being prominent in pictures.

- Among the detected, Golden Retriever is by far the most common breed seen in the tweet database.

- The top 5 was made up by Golden Retriever, Labrador, Pembroke, Chihuahua and Pug.

Note - As we have relied on our algorithm for this classification of images, further analysis should be done before concluding this ranking order to be the perfect analysis. Indicatively though, the data does clearly show which breed has received more mentions from users in their twitter activity.