WeRateDogs – Insights into the twitter archive of Twitter user @dog_rates

Introduction and Background:

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dogs. WeRateDogs has over 4 million followers and has received international media coverage.

So, through this project, we are looking to gain insights from this twitter data to identify trends common to our dog lovers, what makes them tick?

The Project focuses mainly on mastering the three steps of the Data Wrangling process:

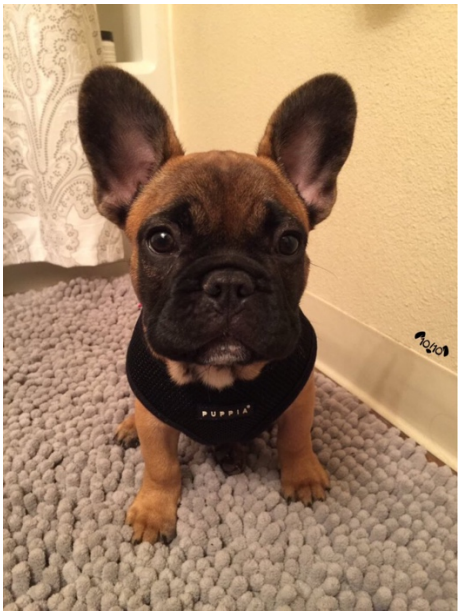
1) Gathering Data
2) Assessing Data
3) Cleaning Data

Finally, using the Cleaned Data to draw insights into trends common to our dog lovers.

This analysis can provide actionable insights into some real-life questions like:

1) What just makes a dog all the more lovable that people retweet, retweet and retweet it?
2) Wowie! What dog should I buy next??

3) Should I buy a pupper, a puppo, a doggo or a floofer? Which stage is more adorable??
4) Just at what times of the day are people like me (aka, the dog lovers), who share my passion more active, so I can connect with them?



Please feel free to reference my Jupyter notebook titled 'WeRateDogs_Project' on my GitHub Account https://github.com/Rajtra to follow the data wrangling process.

Image Source:
https://pbs.twimg.com/media/DE0BTnQUwAApKEH.jpg

Gathering Data:

The data for this project has been gathered from three different sources:

1) Manually downloading the file twitter_archive_enhanced.csv that was collected from the WeRateDogs Twitter Archive by WeRateDogs and sent to Udacity to be used in this project.
2) Programmatically downloading the file image_predictions.csv using hosted on Udacity's servers using the Requests Library using the URL: **https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv**
3) Using the tweet IDs in the WeRateDogs Twitter archive and querying the Twitter API for each tweet's JSON data using Python's Tweepy library and storing each tweet's entire set of JSON data in a file called tweet_json.txt file.

Assessing Data

Data Analysts then assess the data for both Quality Issues and Tidiness Issues.

The 4 dimensions of Data Quality that is assessed are:

1) Completeness – Is the dataset complete?
2) Validity – Does the data make sense?
3) Accuracy – Is the data accurate or inaccurate?
4) Consistency – Is the data standard?

The 3 dimensions of Tidy Data that is assessed are:

1) Each variable forms a column
2) Each observation forms a row
3) Each type of observational unit forms a table

It is important to perform both visual and programmatic assessment of data as this helps identify issues that could be easily overseen just through visual assessment. Ultimately, our goal is to assess the data to understand how to clean the data to prepare a sensible dataset that we can use to analyze and draw insights upon.

Cleaning Data:

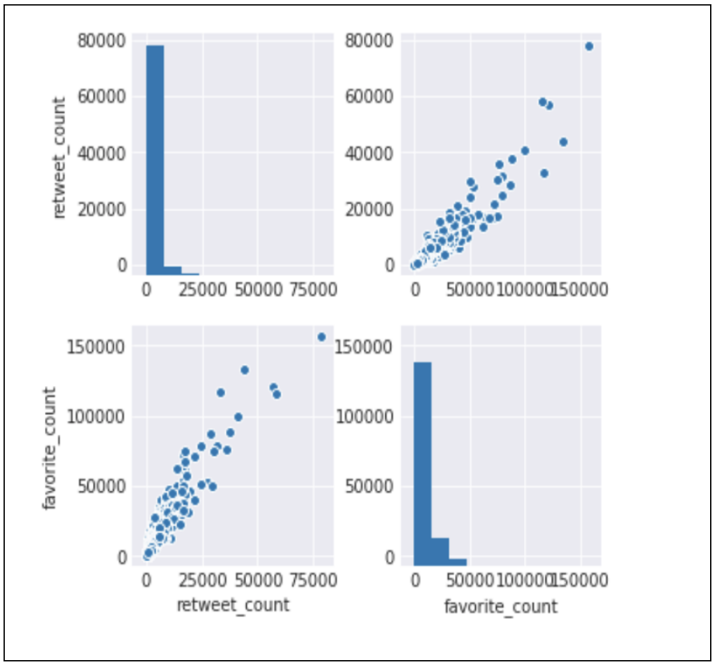
There are three steps to cleaning data. They are:

1) Define – Determine the issue and the resolution, how is it to be cleaned
2) Code – Write code to programmatically clean the data
3) Test – Test to ensure that we have properly cleaned the data

Analysis and Visualization of Data:

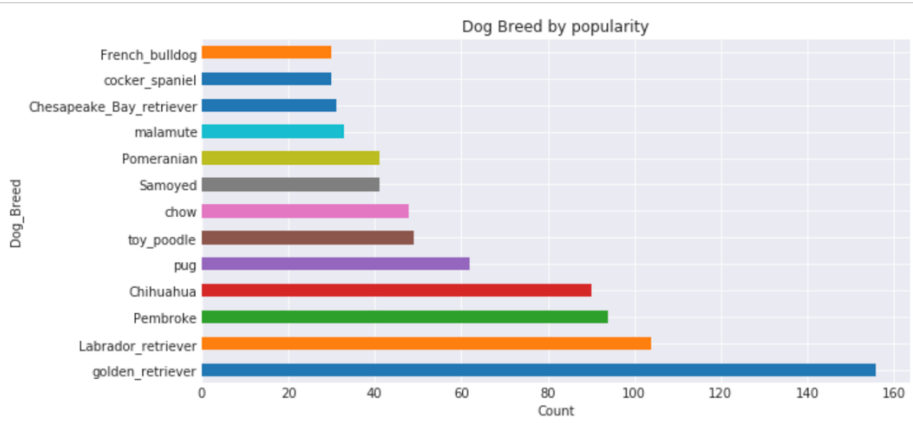
1) What just makes a dog all the more lovable that people retweet, retweet and retweet it?

This question can be answered in a variety of ways. But through this analysis we use the variables retweet_count and favorite_count to understand what a contributing factor could be to retweet a certain post/certain dog post in this example. Simply put, we are trying to identify if there is a correlation between the retweet_count and favorite_count.



2) Wowie! What dog should I buy next??

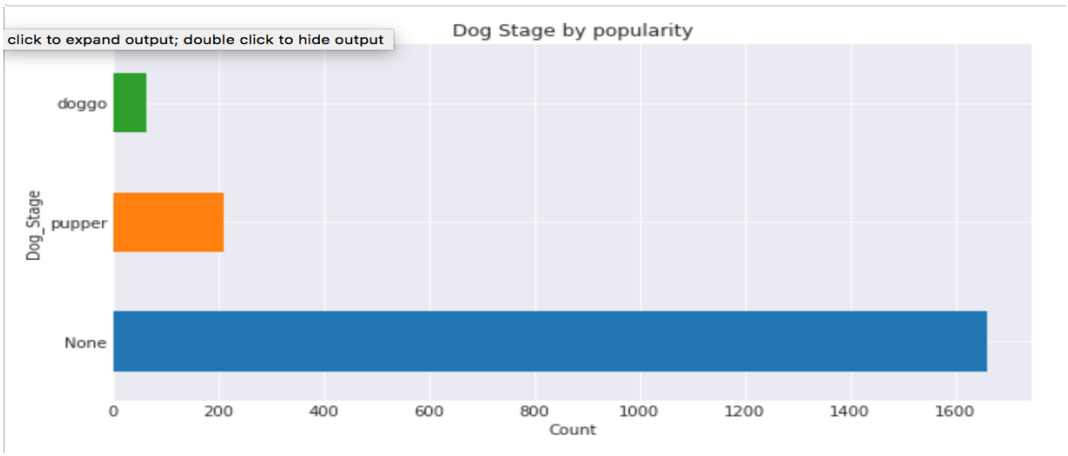
This analysis helps us understand what the most popular breed among dog lovers is. For a person who is new to the 'Dog World', this data is an amazingly useful resource to answer a few questions like, what are the most popular breeds out there? Which one should I buy?
The **Golden Retriever** has captured the hearts of hundreds of dog lovers, so if you are looking to buy a lovable dog anytime soon, you should definitely checkout a beautiful Golden Retriever.



#The most popualar dog breed seems to be Golden Retriever

3) Should I buy a pupper, a puppo, a doggo or a floofer? Which stage is more adorable??

Well, I agree Golden Retriever is a great dog to buy, but just at which stage is a Golden Retriever most adorable? This analysis helps us answer this question, and going by the data we have on hand, it seems like 'pupper' is the most adorable dog stage amongst our dog lovers.

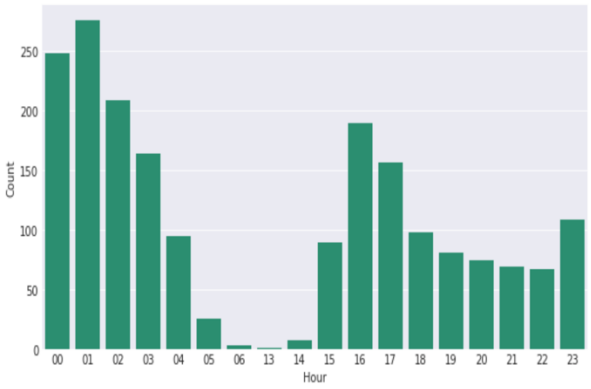


The most popular dog stage is pupper

4) Just at what times of the day are people like me (aka, the dog lovers), who share my passion more active, so I can connect with them?

Our analysis shows us that our dog lovers seem to be less active between 6am and 1pm. While they are most active at 1am in the morning and pretty active during the course of the day. So, if

you are looking to connect with people who share your passion, you know just when to get on your twitter!



Our dog lovers seem to be less active between 6 am and 1pm. They are most active at 1am in the morning and pretty active during the course of the day.

Conclusion:

This report briefly outlines the data wrangling process along with some of the ways we can analyze data to gain actionable insights and answers to some real-life questions. There is more than can be done and learnt from the data than just this, this analysis sets the pace to in-depth analysis of data that could possibly done on the dataset. I greatly enjoyed working on this project and I hope you enjoyed reading through my enough to jump on twitter and tweet away some dog lover posts!