Project – Wrangle and Analyze Data.

This project, as the name suggests focused on data wrangling stage of data analysis. Over here I focused on wrangling data from multiple formats from varied sources with an objective of creating a unified data set akin to performing exploratory analysis on it. The unified data source had primary requirement to be in csv format. The data that was analyzed was of twitter archive of user @dog_rates.

Wrangling was performed on distinct pieces of data and a master dataframe was created

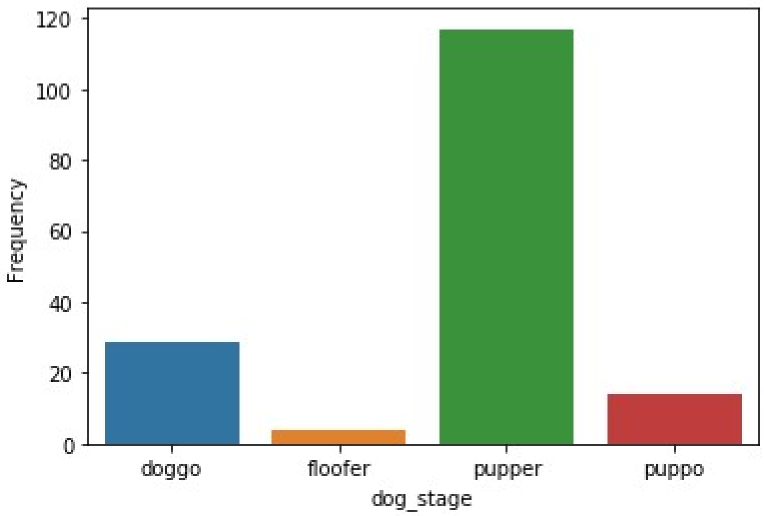
With the data ready for analysis there came the time to propose questions and discover several insights from and into the data.

There are several stages in dogs based on age group such as doggo, floofer, pupper, and puppo. I was curious about which stage represents higher value ratings? When is it most likely that the dog will be rated higher during these stages? After analysis I found out that the ranking of varied stage did not represent significant differences. Mean rankings in each stage were as follows:

dog_stage	mean score
1) doggo	11.586207
2) floofer	11.500000
3) pupper	10.931624
4) puppo	11.785714

Building up on the distribution of scores, I wanted to find out what was the distribution of number of dogs? In other words when is it most likely based on current data that a dog will be requested to be rated?

After performing required analysis, result plotted using barplot looks like this:



Obviously pupper outranks other stages by huge margin. ☺

Later I had an idea that it might be useful to look at the connection between popularity of the tweet and its contents. Therefore I proposed question to pull up full text of the most favourite tweet. It was “'This is Stephan. He just wants to help. 13/10 such a good boy https://t.co/DkBYaCAg2d'”.

Then on the similar line I pulled up text of least favorite tweet, which turned out to be “Oh my. Here you are seeing an Adobe Setter giving birth to twins!!! The world is an amazing place. 11/10 https://t.co/11LvqN4WLq”

Next step could be to pull up 5 most favorite tweets and 5 least favorite tweets then we can run further analysis on it and try to quantify potential factors that increase or decrease chances of a tweet gaining or losing popularity.

When I was a small kid (even today actually LOL), I had the most awesome time when I got a chance to play with Labradors and Alsatians. And I always ended up finding them while walking on the street no matter where I was or no matter what time it was. ☺. That got me thinking, based on the data is it possible that the most common breed is one of these 2 (Labradors and Alsatians)?

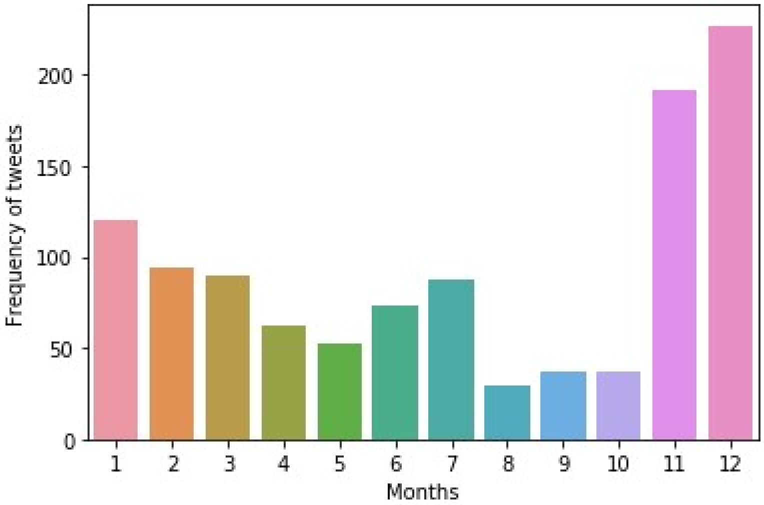
So I ran an analysis to answer that question and realized (little disappointedly) that most common dog breed was ‘golden retriever’. Little disappointed I sure was but I think after going out and interacting with couple of retrievers I think golden retriever has definitely cracked into my top 3 now.

Similarly I wanted to find out which was the least common dog breed.

That answer turned out to be Irish wolfhound. I had neither seen nor heard of that one but quick google search told me that it is a sight-hound dog and is also a strong candidate for guard duty when going in the wild. And now my mind is thinking that this one is like the best candidate to become my training partner in my athletic pursuits.

Moving on to next insight, does the season have an impact on dogs getting rated? Does winter break mean there will be less tweets? Will summer mean lots of dogs outside and by extension more tweet ratings? First step for analysis in these directions was to get distribution of number of tweets in each month.

After running the analysis it can be clearly seen that there is a big jump in the months of November and December, distribution of the same is as follows:



In conclusion, some very interesting and some unexpected insights were learned in this process. But HEY that is the whole point of rational analysis, you are never going to know what to expect.

Most of the questions' answers consist of derived data which can lead to further insights. If I had to pick one favorite insight I would say finding most and least common dog breed was a very positive experience, which opened some good memories from past and gave some plans for the future.

Happy Exploring!!!!!

--Saurabh Daphtardar