Udacity Project

We Rate Dogs Data Wrangling and Analysis:

As an Udacity taking the data analyst nanodegree, it was expected of us to gather, wrangle, and plot some beautiful visuals with a given data in one of the hands on project. Turned out it was some twitter archive data on dogs rating. Lovely right? Ok let's dive in as I take you through the beautiful world of dogs with We Rate dogs' project.

Introduction and background

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.

This project focuses more on the data wrangling process because it is rare to see an almost clean data in the real world. Maybe because some of this data are documented by humans which we know is prone to errors. We are all human right? Haha.

ThereforeWeRateDogs is one of this dirty data everywhere. Also we will be assessing, cleaning and analyzing this data.

1. Data gathering

Steps taken for the data gathering

- First I downloaded the data which is a given CSV file and named as twitter-archive-enhanced.csv.
- Next I downloaded the JSON file named tweet_json.txt and placed them in my working directory
- Next I downloaded the file image predictions file which is in the tsv format.
- Then i saved them all into a pandas dataframe

2. Assessing the data

The data was well assessed using programmatic and visual process and i discovered some quality and tidiness issues.

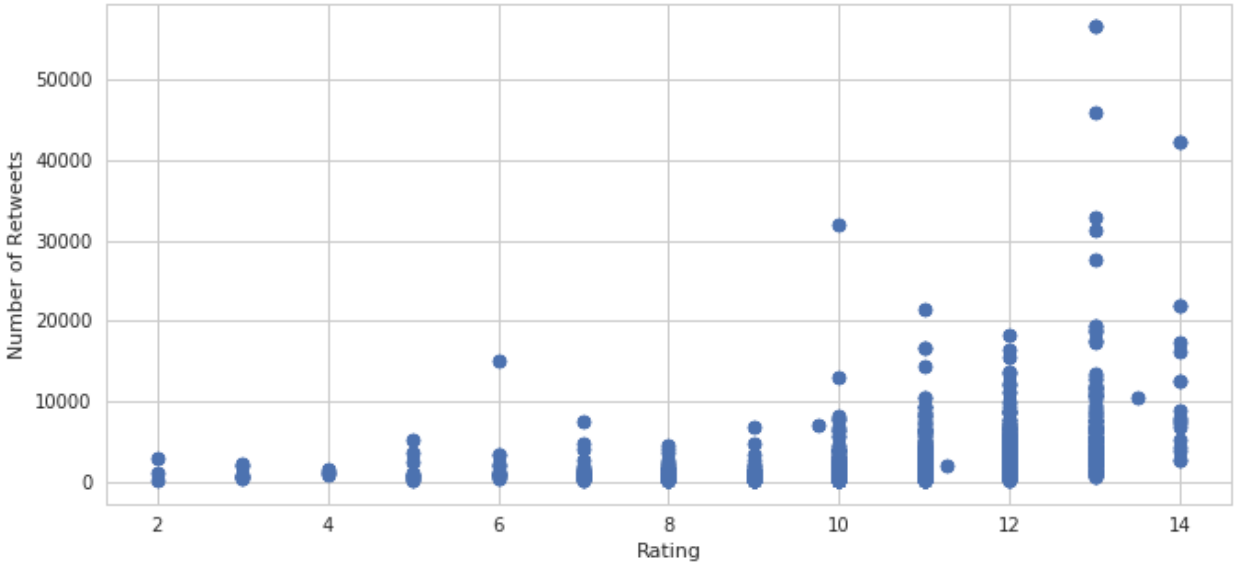
3. Data Wrangling

All quality and tidiness issues where cleaned and a new clean dataset prepared and saved for analysis

4. Analysis and Visualization

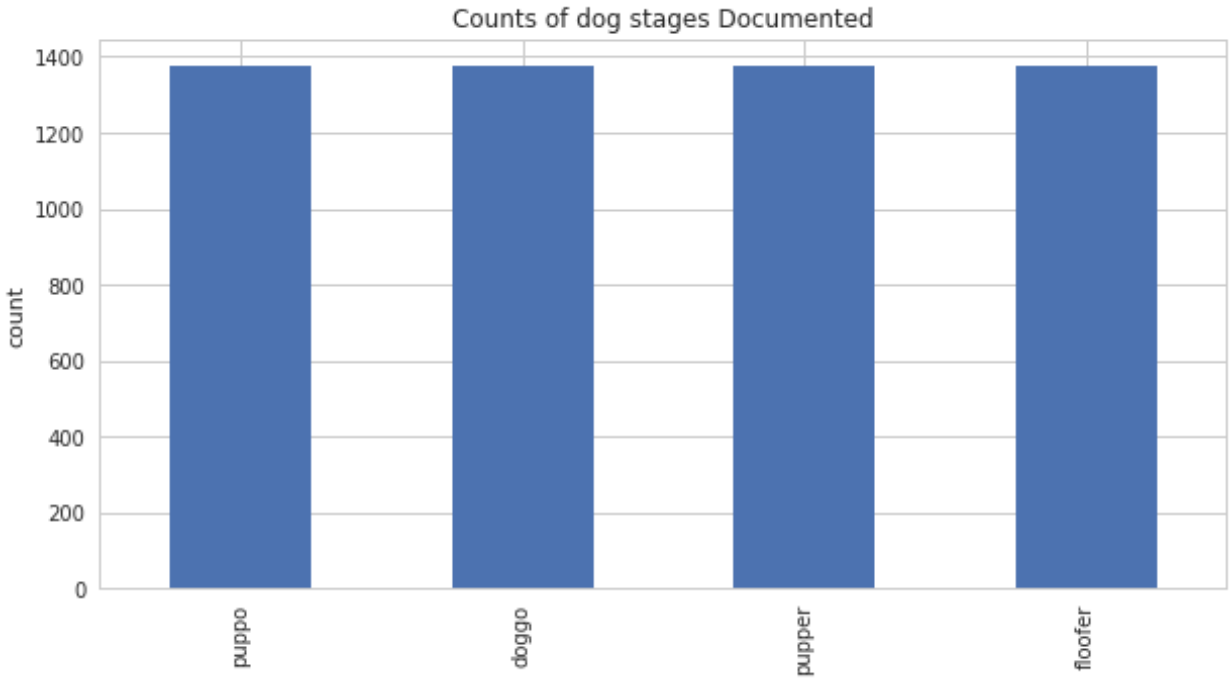
Let's consider some questions to guide our visual analysis

- **How well does rating correlates with retweets?**



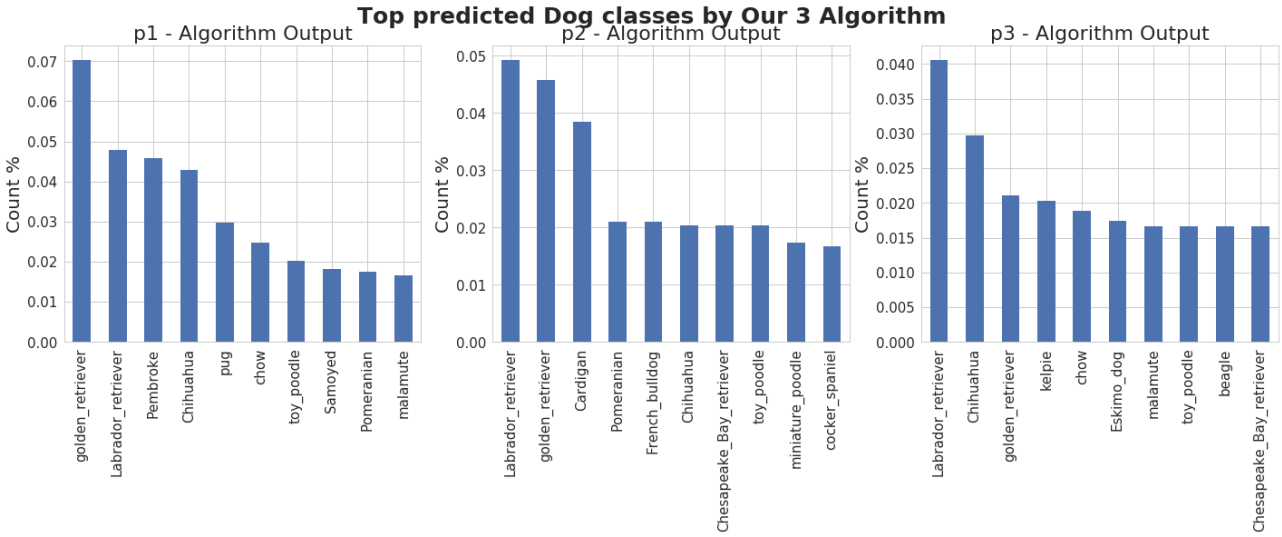
The insight is clear, the higher the rating the more retweet that dog get,

- **How many dog stages is present in the data**



So this clear visual shows that we have four stages and each stage well represented in our data

- **Lastly we have a machine learning algorithm that predicts classes of dogs. Let's see which class has the top prediction**



Labrador Retriever is the most predcited dogs overall.

Conclusion

It was a great project for me overall. I hope you understood a thing or two reading this