Analysis and Insights

Data Analyst Nanodegree
By: Ankit Chaudhary

Introduction:

This Wrangle and Analyze Data Project is part of Udacity's Data Analyst Nanodegree. The project involves wrangling of data from various sources associated with tweets from Twitter user @dog_rates , also known as WeRateDogs. WeRateDogs rate's pictures of people's dpgs in humorous manner, most often giving ratings higher than 10/10. After scraping together the data, quality and tidiness issues were assessed and then cleaned.

WeRateDogs™ @dog_rates · Apr 10

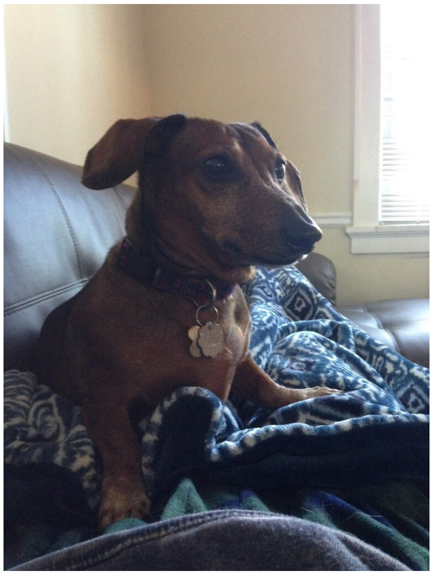
This is Ruby and Max. They pick up a piece of litter on every walk. Sometimes their ears get tangled, but the planet is worth it. Both 14/10

potter

Gather:

This project gathered data from the following sources:

- The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file was provided to Udacity students.
    - 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.
- The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file was provided to Udacity students.
- Twitter API and Python's Tweepy library to gather each tweet's retweet count and favorite ("like") count.



Did you get all the data we need??

(Source: https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg)

Assess:

Assessing data requires data analysts to evaluate a data set on quality and tidiness issues.

The four (4) main data quality dimensions are:

- Completeness: missing data?
- Validity: does the data make sense?
- Accuracy: inaccurate data? (wrong data can still show up as valid)
- Consistency: standardization?

And there are three (3) requirements for tidiness:

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table

As you look at the data gathered, keep the final product in mind – what kind of data should be presented visually vs. which portions of data only require programmatically analyzing in order to convey insights into the data set?

Clean:

Cleaning data is tedious, and often iterative. Just when an analyst believes they have found all quality and tidiness issues, there are often additional issues that arise. The cleaning process involves three steps:
1. Define: determine exactly what needs to be cleaned, and how
2. Code: programmatically clean the code
3. Test: evaluate the code to ensure the data set was cleaned properly

Analysis and Visualization:

Finally, two visualizations were created and insights can be found below.

1. Favorite vs Retweet count:

At the time this dataset was collected, WeRateDogs had over 4 million followers; therefore, their tweets are likely to get many favorites and retweets. In addition, there may be some tweets that are extremely popular if they become part of international news coverage or go viral. In Figure 1, it can be seen that favorite and retweet counts are highly positively correlated. The majority of data falls below 40,000 favorites and 10,000 retweets. The most popular tweets has 13,0000 favorites and 80,000 retweets.

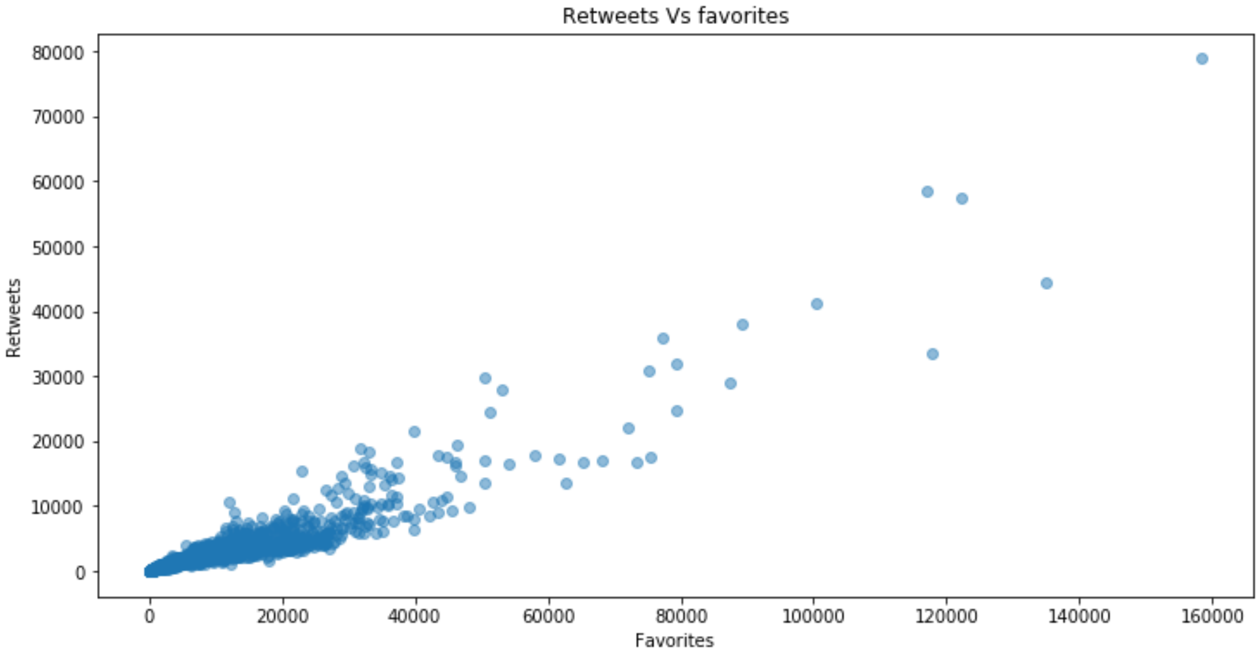
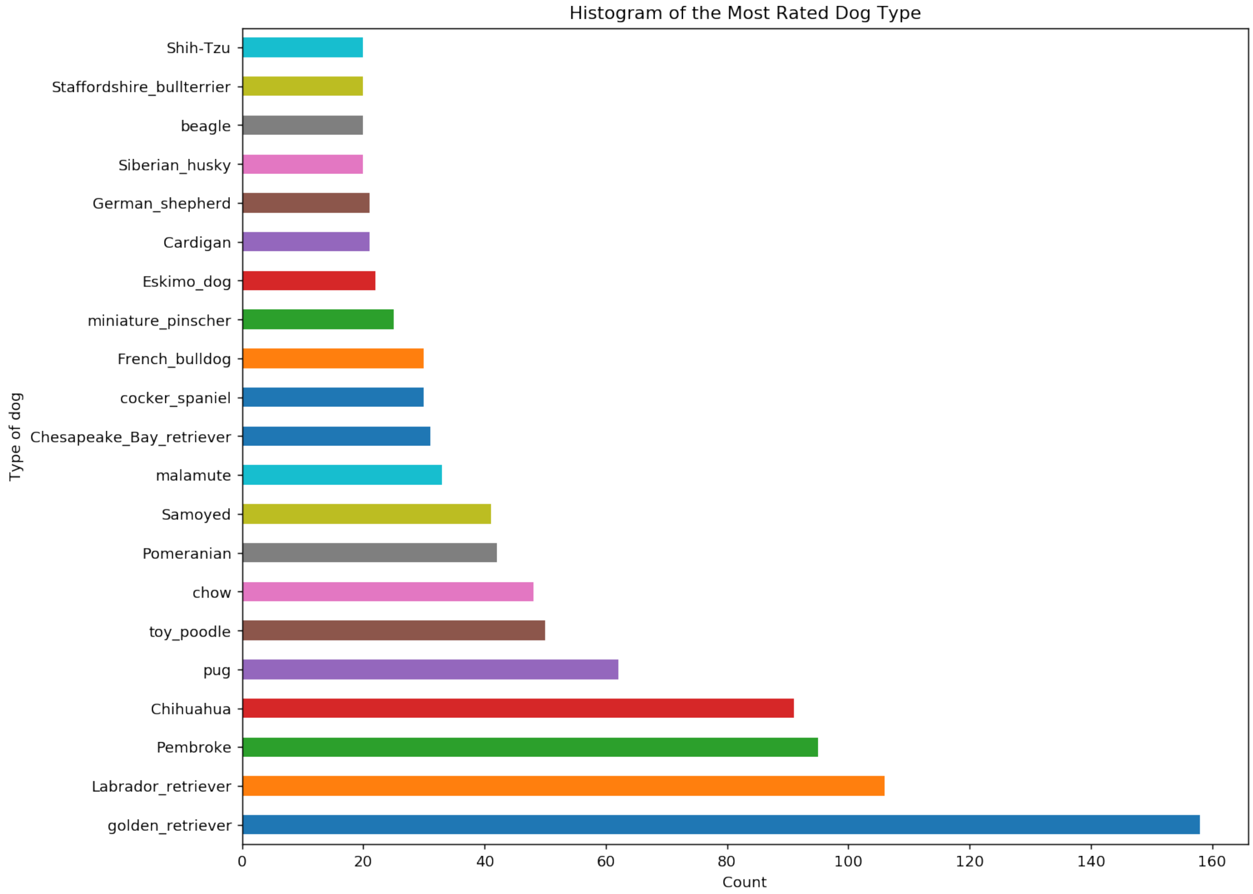


Figure 1: Favorites Vs Retweets

2. Most common Dog Type:

WeRateDogs has over 6000+ tweets. I was able to analyzed 1500+ tweets. The most rated dogs was golden retriever with more than 150 ratings.



Lowest Average Rating among dpg types:

Japanese Spaniel has the lowest rating.