WeRateDogs – Insights into the @dog_rates Twitter page

Introduction

Real-world data rarely comes clean. The dataset wrangled for this project 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. Here's an example:

This project works through the data wrangling process, focusing on the gathering, assessing, and cleaning of data. There are visualizations and observations from the analysis provided as well.

Gathering

This project gathered data from the following sources:

1. The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file was provided to Udacity students.
2. WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for 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.
3. 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 was used to gather each tweet's retweet count and favorite ("like") count.

Assessing

Assessing data requires data analysts to evaluate a data set on quality and tidiness issues. The four main data quality dimensions in order of importance are:

1. Completeness: missing data?
2. Validity: does the data make sense?
3. Accuracy: inaccurate data? (wrong data can still show up as valid)
4. Consistency: standardization?

There are three requirements for tidiness according to Hadley Wickham:

1. Each variable forms a column
2. Each observation forms a row
3. 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?

Cleaning

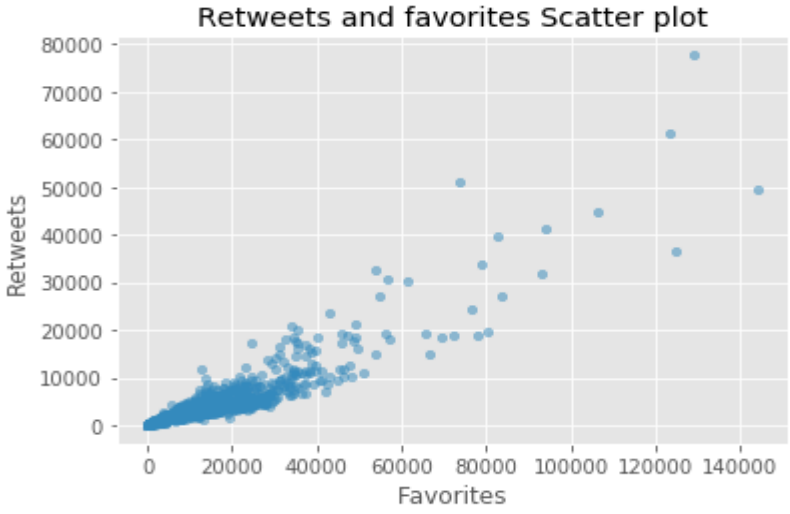
Cleaning data is tedious, and often iterative. Just when an analyst believes they have found all quality and tidiness issues, there are often additional issues that arise. The cleaning process involves three steps:

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

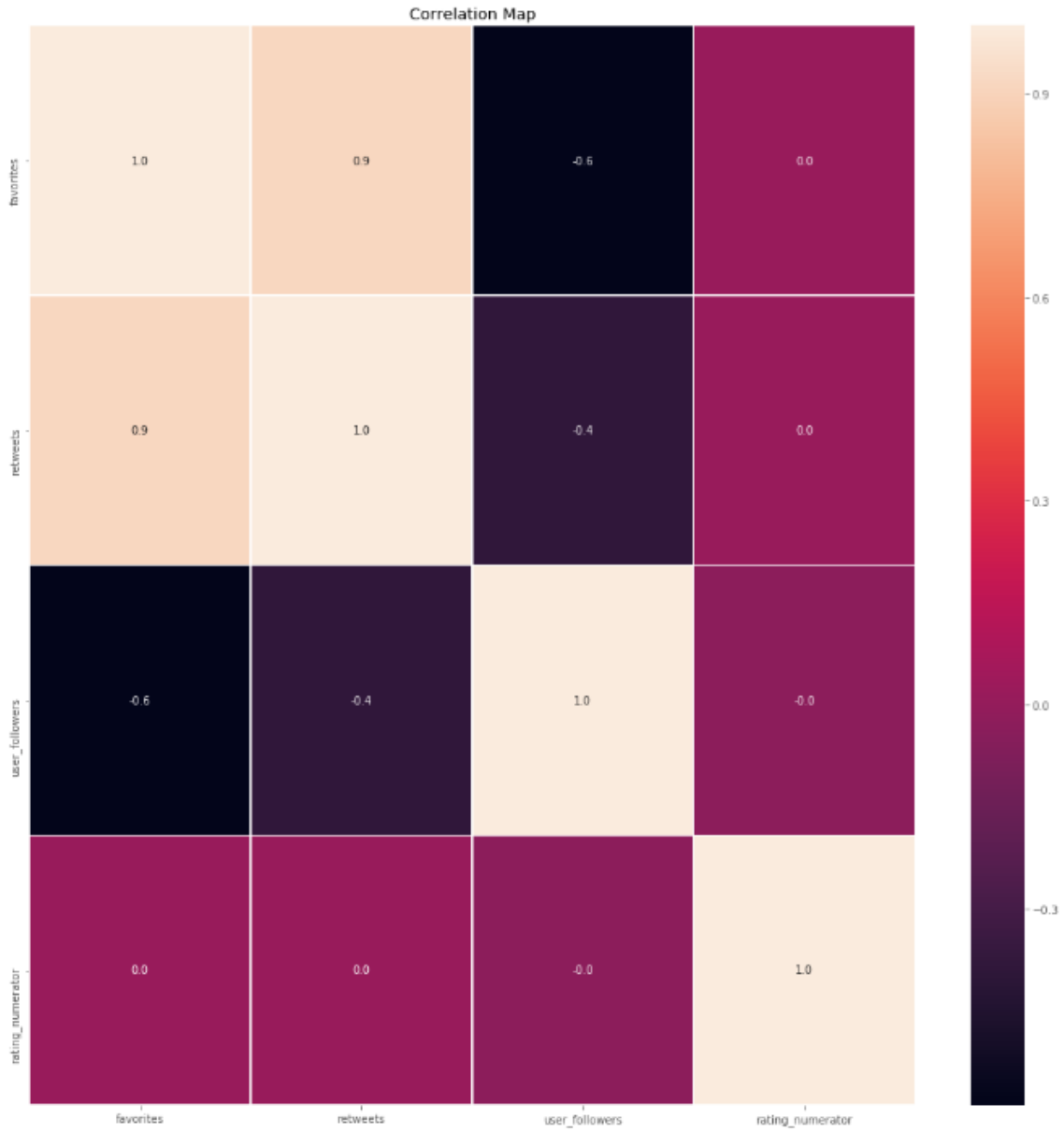
I chose to analyze and present on four different pieces of the WeRateDogs data set.

Retweets Vs Favorites



As seen from the scatter plot above, the number of retweet is highly correlated with the number of favorites

Correlation between Variables in the dataset

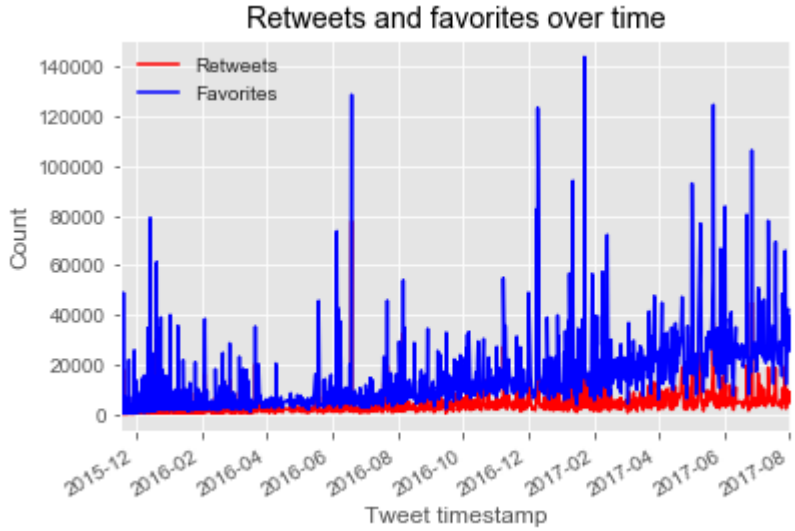


1. The only strong correlation we see here is between favorites and retweet. This is normal as more favorites translates to more retweets
2. User followers and retweet have a weak negative correlation of -0.4
3. Ratings do not get affected by any other variable asides from the ones plotted

Rating System

The twitter page starts with very small ratings. With time, they adopted the system of rating numerators more than the denominators.

Retweets and Favorites Timeplot



A twitter user, Brent, had called out this page for unnecessarily giving dogs overly high ratings. Maybe Brent has all the right to get mad as the ratings were getting higher for no specific reasons.

Conclusion

This write-up offers a straightforward look at the data wrangling process. There is so much more that can be done with this data set, but I encourage aspiring data analysts to dive deep into this data set and see what else can be found!