Wrangle and Analyze data Report

Udacity act report

Introduction:

This report is part of the wrangle and analyze data project from Udacity Data Analyst Course. Here I am going to explain how the process was for gathering the datasets, cleaning, making it tidy and create visualizations and insights.

The project was built in a Jupyter Notebook, and although the project is in a linear sequence in there, it is important to explain that it was made in several circles of accessing, cleaning and organizing.

Gathering:

There were 3 files to gather, the twitter-archive-enhanced I just made the normal download via Udacity, although the other 2 files was necessary to gather programmatically.

To gather the image-prediction file, I just used a simple request along with the respectively URL.

The twitter reactions file was the most complicated one, I did on a Sublime Text Editor because it was asked for not showing the twitter API on the submission. To collect this data I used the tweepy inside a for loop, where each tweet_id collected were from the twitter-archive-enhanced. Some of the tweets were not working so inside of the for there was a try statement, to gather only the working tweets and write it in a JSON file on my directory.

Accessing:

The twitter-archive and image-predictions were accessed using pandas, but for the twitter-reactions I first had to read the JSON file, separating per line and headers. After that was possible to read it as pandas Data Frame.

Cleaning:

In overall the data was not very dirty, although it was very messy, and to take clear results from it I had to make it all tidy first.

The main issues are explained in the table below

DataFrame	Messy	Dirty
Twitter Archive	• Columns with missing values • Each dog stage in one column • Timestamp as string	• Wrong names on name column • Wrong denominators • Wrong numerators
Image Predictions	• Non uniform breeds names • Too many columns • Predictions with False, False and False	
Twitter Reactions	• Timestamp column as string	
All Data Frames	• The 3 datasets could be concentrated in one only.	

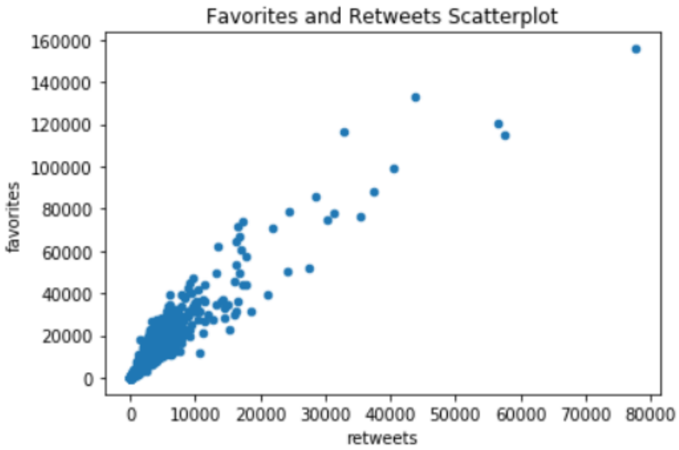
Visualization and Insights:

After cleaning all of the issues showed in the table above, I could explore one master dataset with reduced information but tidy and clean. So, I asked myself some questions like:

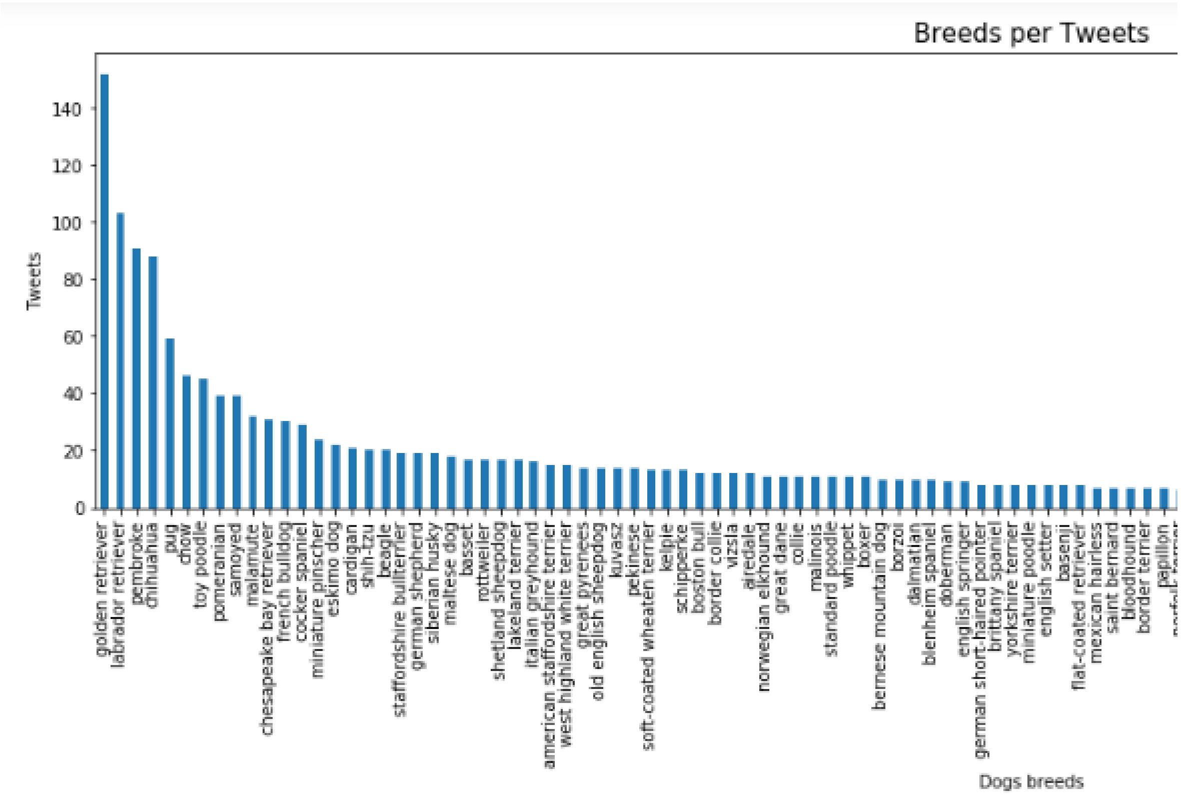
1) Favorites and retweets are correlated?
2) What is the most popular dog breed?
3) How are the stages distributed along the tweets?
4) Which tweet had the higher numerator rate?
5) Which tweet had more likes?

Answers:

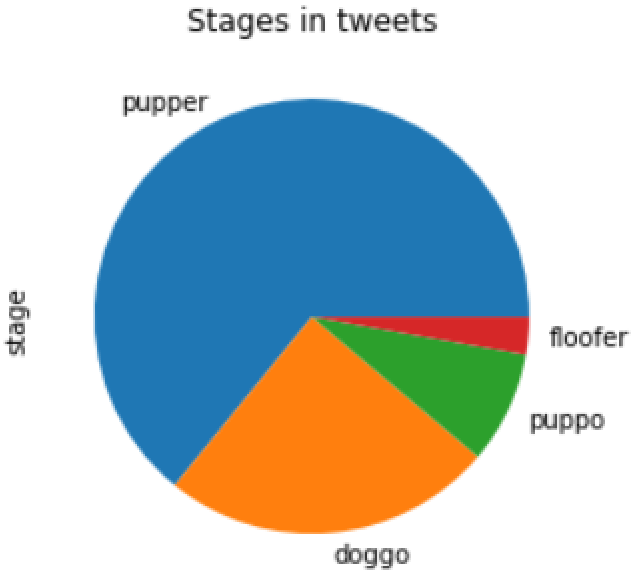
1) YES, as common sense, favorites and retweets are correlated.



2) The most popular breed in our dataset is the Golden Retriever

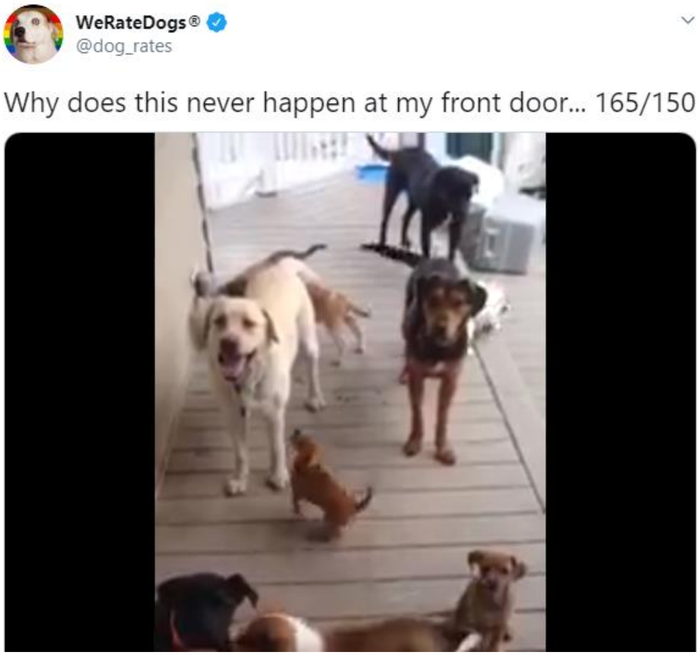


3) Considering only tweets with dog stage declared on it, that’s how it looks like.



4) The tweet with highest rate was 165/150, with a video full of cute dogs:

https://twitter.com/dog_rates/status/758467244762497024

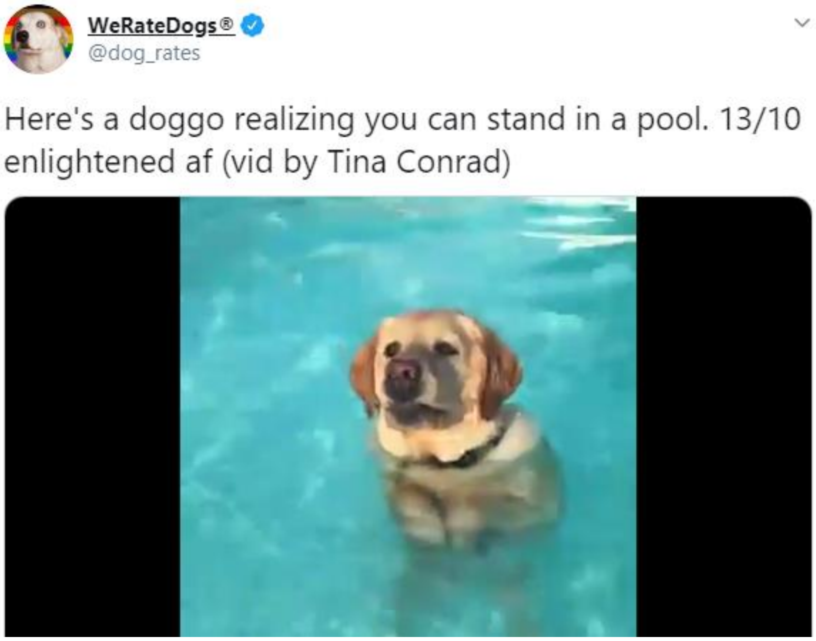


WeRateDogs®
@dog_rates

Why does this never happen at my front door... 165/150

5) The twitter with more likes and retweets was this funny doggo Labrador swimming on the pool. 155624 Likes and 77560 retweets.

https://twitter.com/dog_rates/status/744234799360020481



WeRateDogs®
@dog_rates

Here's a doggo realizing you can stand in a pool. 13/10 enlightened af (vid by Tina Conrad)