SEPTEMBER 07, 2022

Data Wrangling Report

Gathering, Assessing, Cleaning and Visualizations

PRESENTED TO

Udacity

PRESENTED BY

Dania Alaa Adimi

Table of Contents

About the project

This project was all about Data Wrangling, which consists of data gathering, assessing and cleaning. In this project, I wrangled and analyzed 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.

In this project, I used three data sources in order to build a robust analysis. Twitter Archive dataset which was in the form of a .csv file, It was provided by Udacity, It contains some basic information about tweets such as tweet_id, date, etc. Image prediction dataset, that contained some predictions, I had to download it programmatically from Udacity's sever. And finally, I used Twitter API (along with tweepy) in order to extract more meaningful data about tweets such as retweet count and like count.

In the next pages, I will be documenting each step. For further details, you can check the notebook wrangle_act.ipynb

Data gathering

My wrangling efforts for the @WeRateDogs Twitter project included gathering data from the following sources:

The WeRateDogs Twitter archive: The twitter_archive_enhanced.csv file was provided by Udacity. This archive contains basic tweet information such as tweet ID, timestamp, text, etc, for over 5000+ tweets. These tweets were filtered and we only kept those who have ratings. So this dataset contains 2356 entry.

The tweet image predictions: i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network. This file was provided by Udacity and I had to download it programmatically.

Additional data from the Twitter API: Using Twitter API and Python's Tweepy library, I gathered more data about the tweets such as each tweet's retweet count and favorite ("like") count.

Data Assessing

Once our data is gathered, we need to assess it right? going through some visual and programmatic assessts here's what I could come up with:

Quality Issues

df_twitter_enhanced

- We need to remove retweets and columns that come with it.
- timestamp should be datetime not an object.
- Extra HTML text in the source column needs to be removed.
- In the name column, we have null objects are not declared as non-null objects. So we should turn None to NaN.
- Incorrect names or missing names in name column such as, a, an, the, very (entries that are not names), most of them in lowercase.
- puppo, floofer, pupper and doggo columns should be Boolean.
- The ratings is not standard since there are different denominators. We better add a new column that contains the resulted rating.

Data Assessing

df_image_predictions

- Remove duplicate jpg_url entries.
- Refine p1, p2 and p3 columns and confidence associated by combining.

df_image_predictions

- user_favourites, and user_followers values are identical for all rows, which is specious, we should deal with this.
- Drop the datetime column because we already have timestamp in df_twitter_enhanced.

Tidiness

- Three columns are present for variable dog stage - pupper, puppo and doggo. So we should merge them.
- Join archive, predictions and tweet_json tables to have a more concise dataframe for our analysis. And we remove redondante columns such as datetime in the df_tweets_json dataframe.

Now It's time for cleaning!

Data Cleaning

After the assessment, I went through cleaning the data through the following means "Define", "Code" and "Test". I first made a copy of dataset in order to keep the original versions in case we need them at certain point, then I cleaned each dataset seperately according to the previous step (assessing Data), then I merged them all together and stored them in a dataframe called df_archive_master.

Once our data was cleaned, I was able to run some analysis and display some visuals. I'll discussing that part in particular in another report called act_report.pdf.

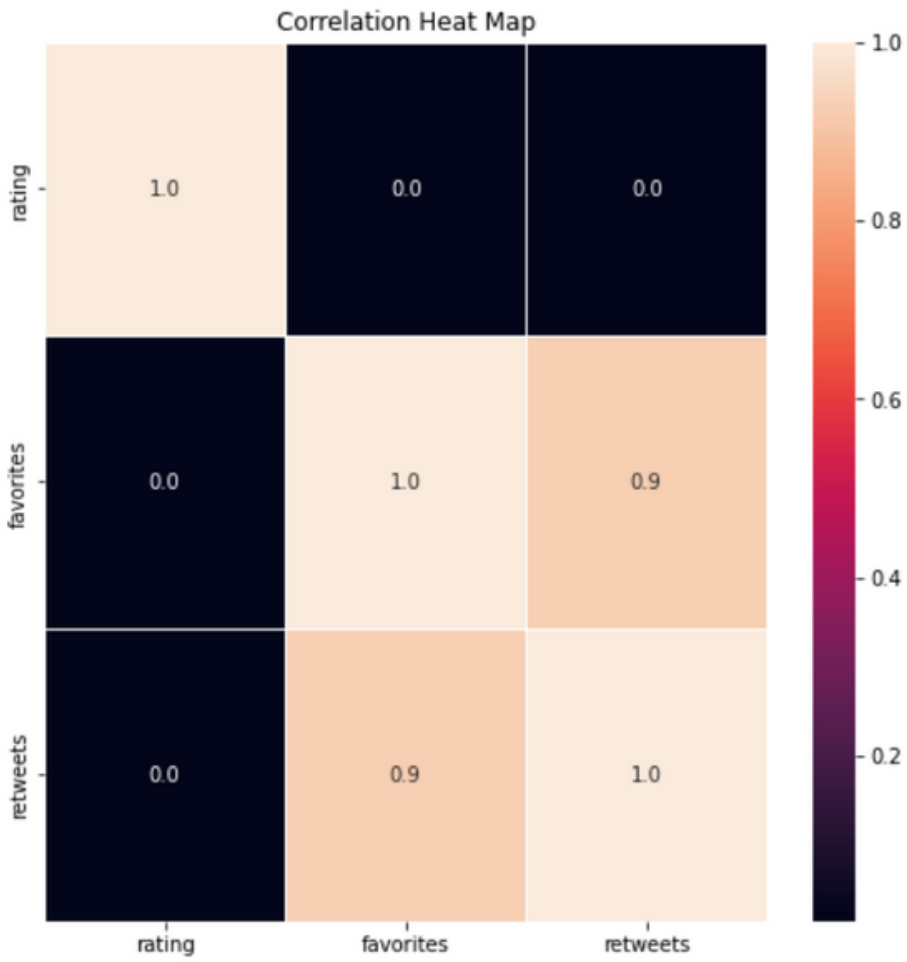
Analyzing and Visualizing

Once our Data is cleaned, we can start analyzing it and extracting insights from it.

In the project requirements. We were asked to do at least 3 visuals, and here are the ones that I chose.

Correlation Heatmap

In order to see the correlation between our variables, we can plot a heatmap.

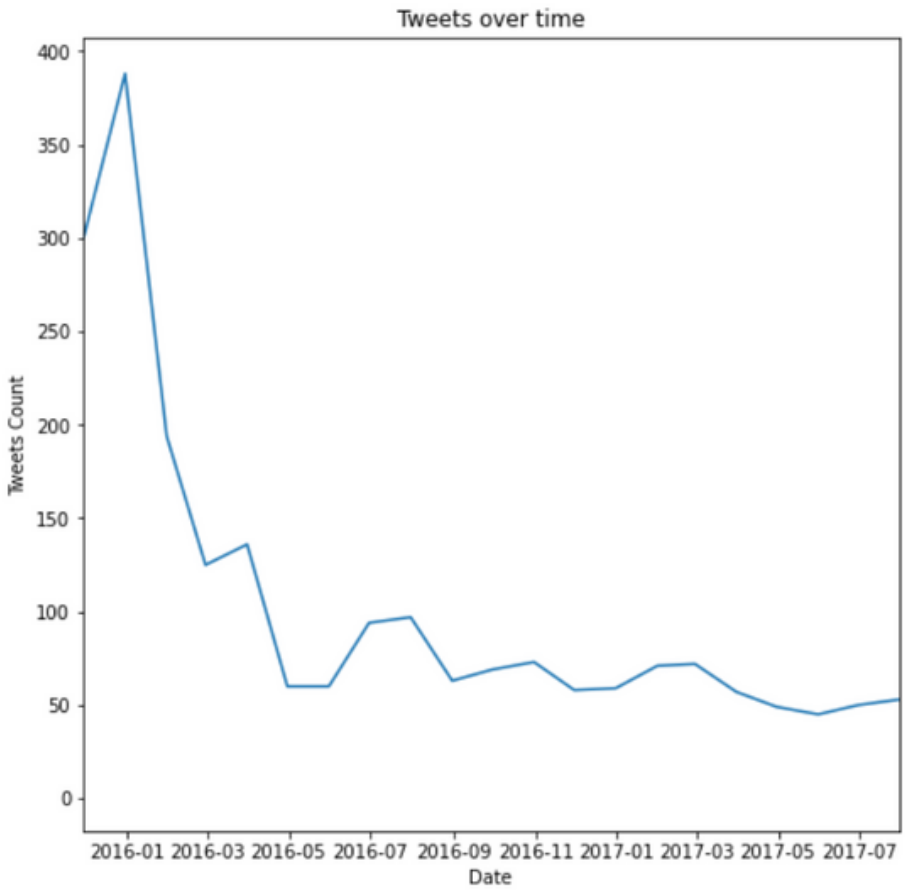


Insights

- We can clearly see that there's a strong correlation between favorites and retweets, which kind of expected.
- There's no correlation between retweets and rating, favorites and rating.

Tweets over time

In order to see the evolution of tweets over time, we can use a graph for better insights.

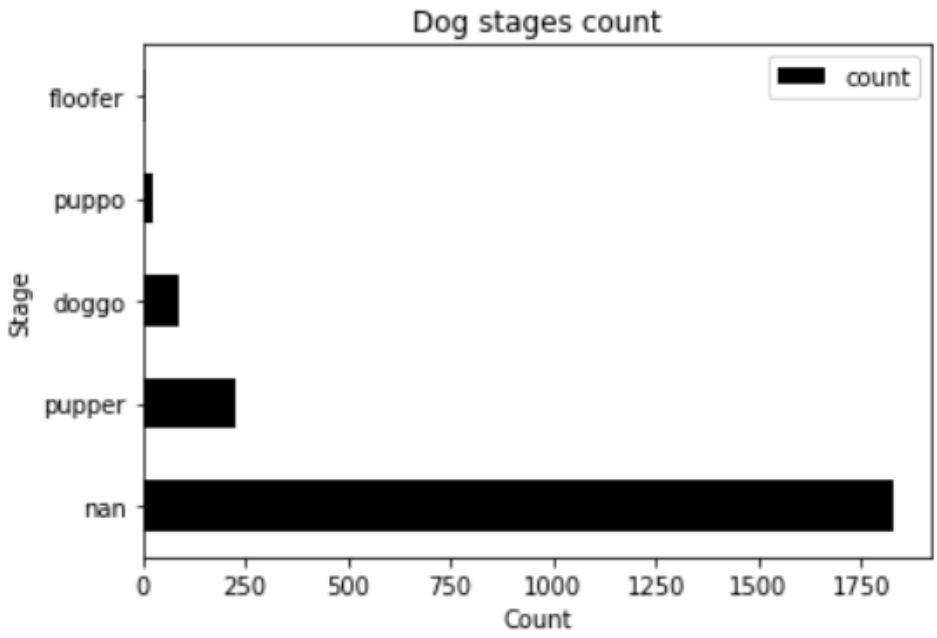


Insights

- We can clearly see that the tweets count was decreasing tremendously.

Dogs stages

Another interesting thing that we can do is to check which are the dogs stages that are more frequent in our dataset. We can do that using a bar plot.

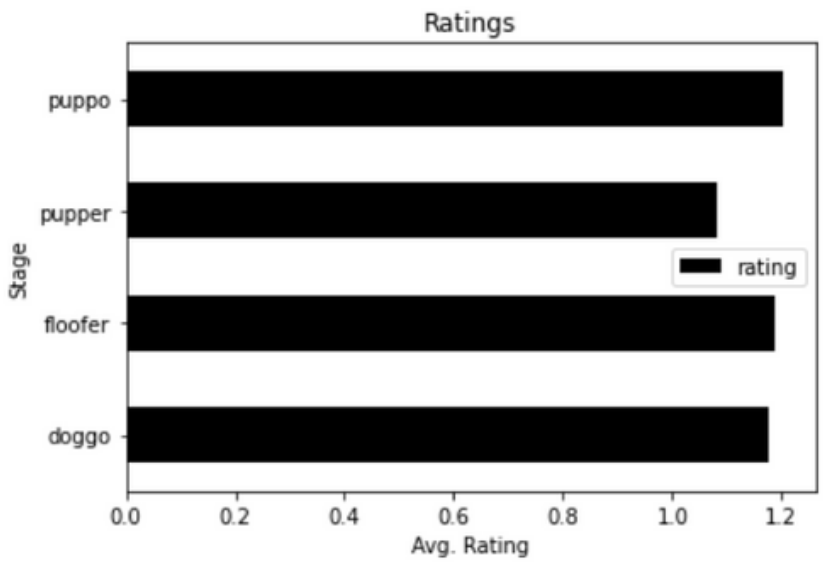


Insights

- Dogs in Pupper stage of dog life cycle get most tweets, which is expected.
- However, most dogs in our dataset are missing the stage.

Dogs stages with ratings

Even more interesting is to check which are the dogs stages that are have the highest avg. rating.



Insights

- We can clearly see that puppos are on the top of the list.