Wrangle Report

The Gathering Process

The data for this project was provided by different sources. Each source required a different method to obtain the data

- By the Twitter API
- downloading the Data from an external server in the tsv format
- importing the data from a provided csv file

Description of the data scources

1. Enhanced Twitter Archive

This dataset was initialy provided by Udecity and provided a baseline for my analysis as it contained 5000+ data entries with a Tweet ID and other meta information.

2. Twitter API

To obtain the retwwet count and favorit count for each tweet, a script was used to fetch the information from Twitter. The script used my private Twitter account in combination with the Developer account to fetch the data with the Tweepy libary and store the results as a JSON file.

3. Image Prediction Dataset

The Image Prediction Dataset was a collection of results created by a ML model which had classified each dog in the picture by breed. It was programaticaly downloaded from the Udecity server and was in the tsv format.

Quality Issues

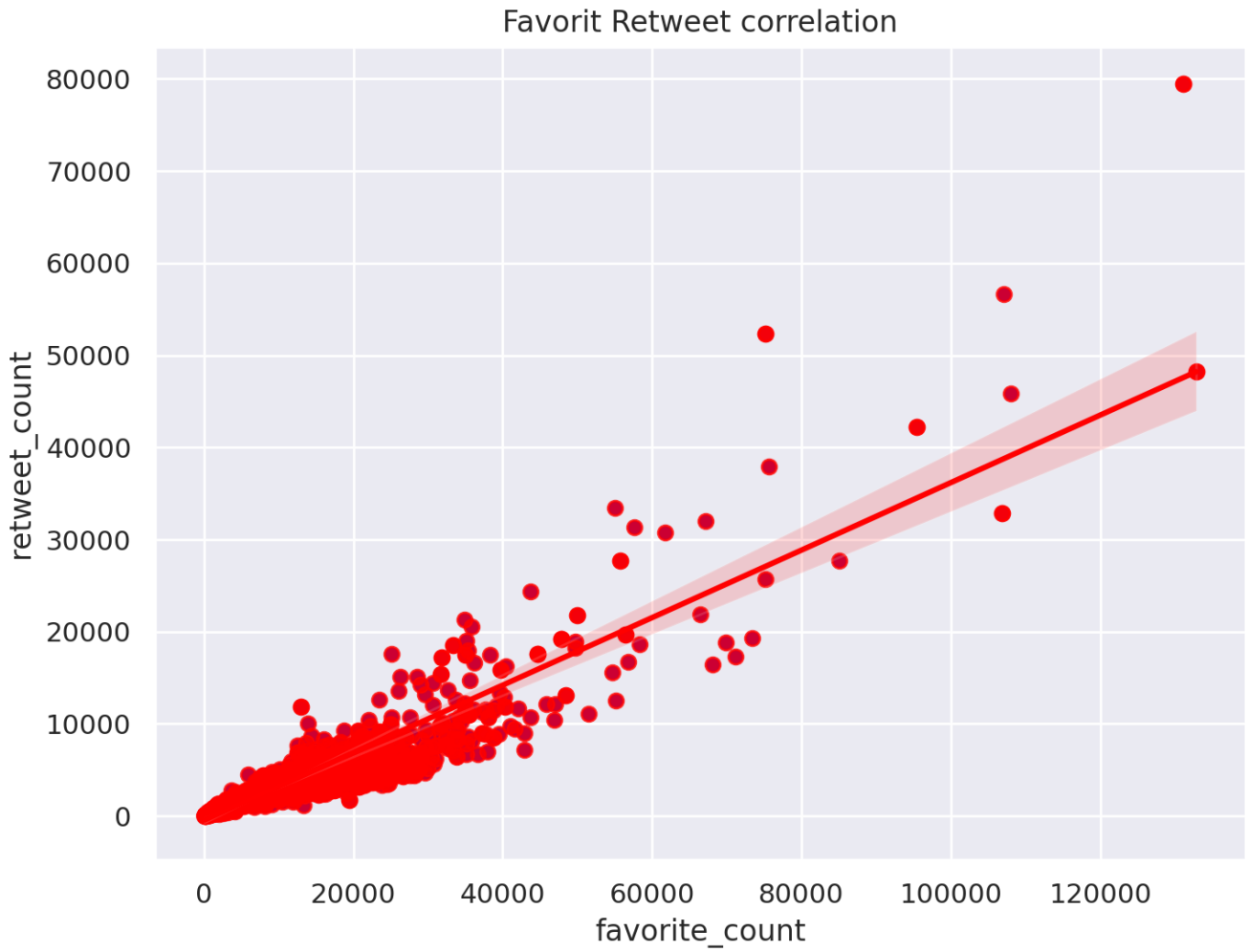
- Twitter Archive Enhanced Dataframe
  - Drop `retweeted_status_id`, `in_reply_to_user_id`, `retweeted_status_id` and `retweeted_status_user_id`
  - `source` colum should be categorical
  - `timestamp` colum should be renamed `archive_timestamp`
- Image Prediction Dataframe
  - rows there no Dog can be identified because `*_dog` was false should be droped
  - `p1`, `p2` and `p3` should be categorical as well as the colums should be descriptiv
  - the confidence level `p1_conf`, `p2_conf` and `p3_conf` should be mearged together
  - colums `*_dog` should be merged
- Tweets Json Dataframe
  - Drop Colums like `id_str` and any other `*_str` column
  - Cast `created_at` as a datetime object instead of string

Tidiness Issues

- In the image_prediction Dataframe we should drop `img_num` dataframe after everything else was merged
- all the Dataframes (Twitter Archive, Twitter Json, Image Prediction) should be merged for easy analisys
- merge the doggo, floofer, pupper and puppo colums together

Insights gained about the dataset

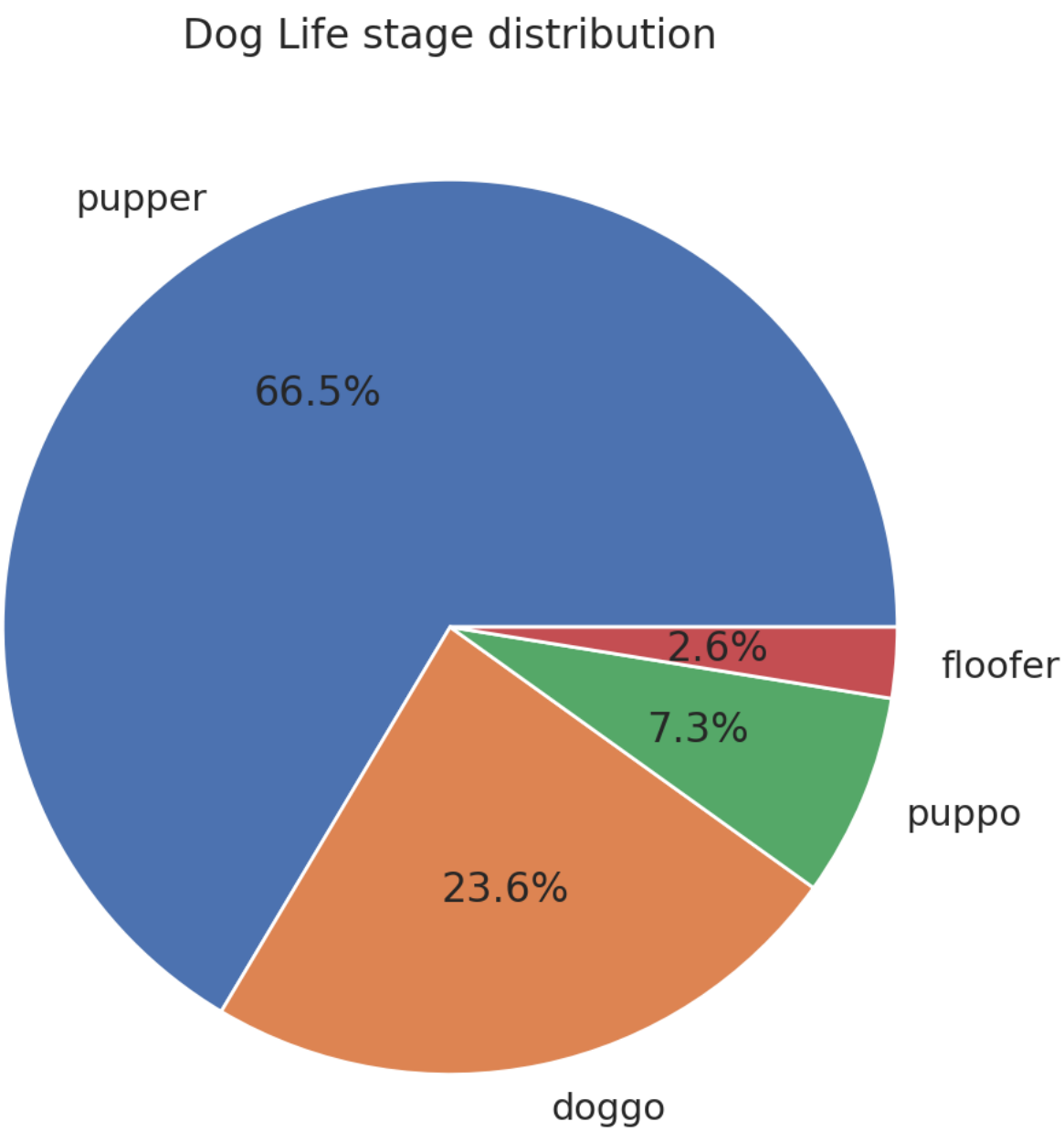
The first question I asked myself then working with the dataset was: How is the favorit count correlating to the retweets? Here is my result.



It is Interesting to see we we have a lot of tweets which are retweetetd but not marked as a favorite

---

After what I wanted to identify in what kind of stage most of the Dogs are then presented on WeRateDogs.



The conclusion is that most of our Dogs are pupper! The smallest amount of dogs the dataset includes are floofer which is realy sadge_distribution.png

---

Last but not leasted I wanted to know what are the 5 most common names of the dogs and here is the Top 5:

Type *Markdown* and LaTeX: α^2