Data Wrangling Report - WeRateDogs

Data wrangling includes three parts:

1. Gathering Data

2. Assessing Data

3. Cleaning Data

Project objectives

The project main objectives were:

- Perform data wrangling (gathering, assessing and cleaning) on the provided sources of data.
- Store, analyze, and visualize the wrangled data.
- Reporting on
    1. data wrangling efforts.
    2. data analyses and visualizations.

Gathering Data

- In this phase, the three pieces of data were gathered:

- The archive of the enhanced twitter data file was downloaded manually, which including variables for each tweet like timestamp, tweet id, rating numerator, text, rating denominator, name, dog type, and etc…

- Using twitter API additional data are provided including favorite count, retweet count.

- Also the tweet image predictions was downloaded programmatically using import requests library

Assessing Data

- I assessed the data visually and programmatically using .info() , .head() , .describe(), .value_counts()

Tissues were cleaned:

- Combined all the data fames together if they are containing data related the same tweet

- Combined 4 stages of dog into one stage and one column dog_stage

- Removing useless column

- assign 10 to rating denominator column

Cleaning Data

Using code methods and test:

Capitalize(), drop(), replace(), info(), head(), value_counts(), rename(), merge()

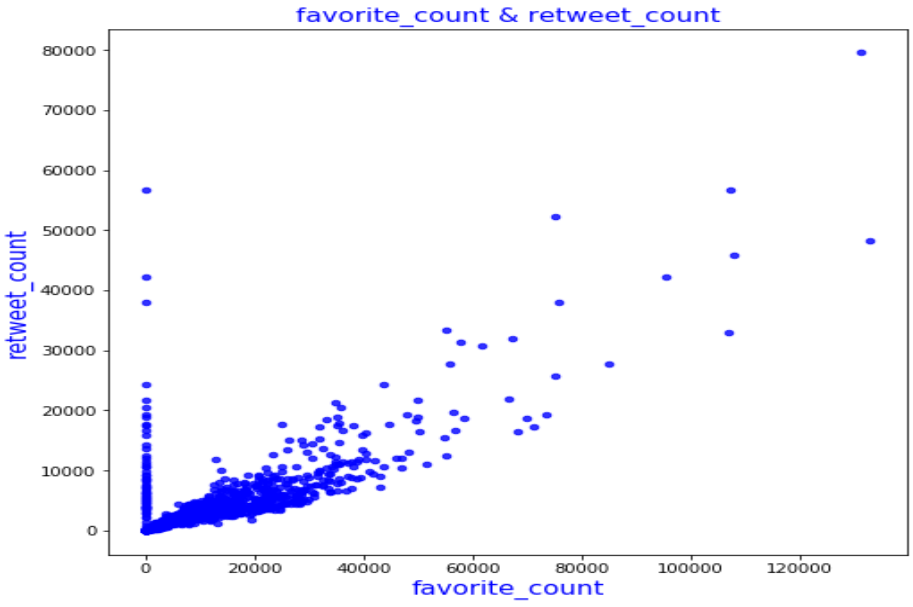
Mergeing data to new fcsv file called ”twitter_archive_enhanced_new”

Analysis and Visualization

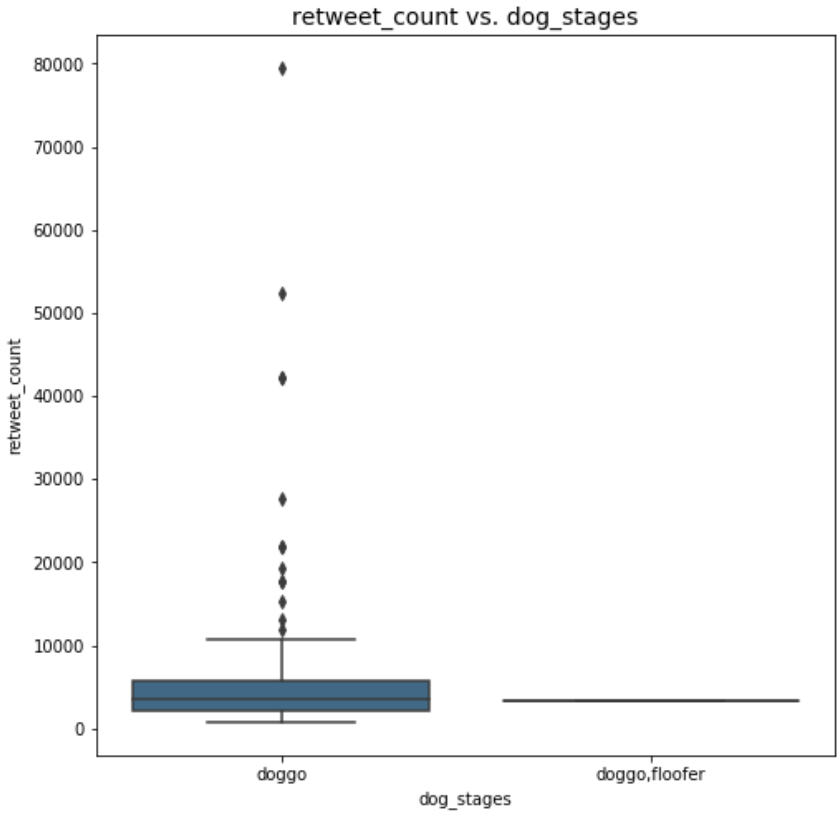
Highest retweeted dog:

Atticus dog wearing USA flag hat and tie and also in front of USA flag.

The favorite rate dog puppo:



You can see that number of retweet_count increasing and number of favorite_count also increasing and that's called linear coordination



We can see here that doggo has the highest rate retweets.

Sources:

https://pypi.org/project/requests/

https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

http://docs.tweepy.org/en/latest/getting_started.html#api

https://github.com/siznax/wptools/wiki

https://stackoverflow.com/questions/28384588/twitter-api-get-tweets-with-specific-id

https://wiki.python.org/moin/HandlingExceptions

https://stackabuse.com/reading-and-writing-json-to-a-file-in-python/

https://pythonprinciples.com/blog/converting-integer-to-string-in-python/

https://pandas.pydata.org/pandas-docs/stable/user_guide/merging.html

https://stackoverflow.com/questions/52267654/jupyter-notebook-python-nameerror

https://stackoverflow.com/questions/51762938/seaborn-color-palette-not-working-appropiate-with-lineplot