2021

Data Analysis Report

DINA EL-SOKARY

Introduction

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. You will document your wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL.

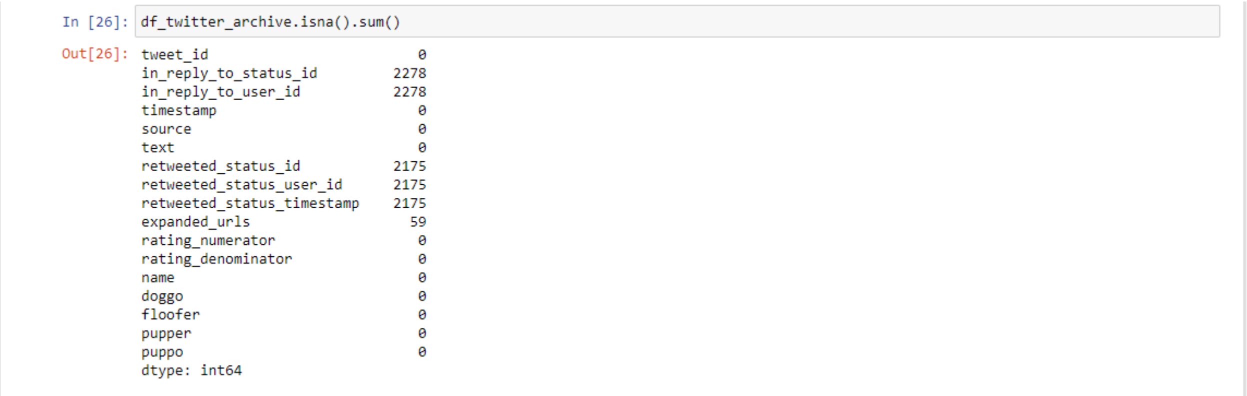
The dataset that you will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to 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.

Data Insights

I tried to gain some insights from the data.

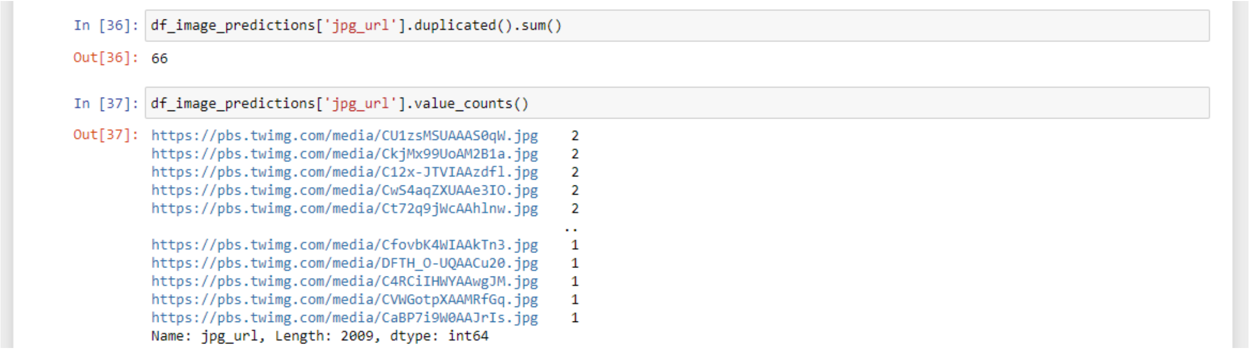
- ✓ The most columns with the NaN values



```
In [26]: df_twitter_archive.isna().sum()

Out[26]: tweet_id                         0
         in_reply_to_status_id         2278
         in_reply_to_user_id           2278
         timestamp                        0
         source                           0
         text                             0
         retweeted_status_id           2175
         retweeted_status_user_id      2175
         retweeted_status_timestamp    2175
         expanded_urls                   59
         rating_numerator                 0
         rating_denominator               0
         name                             0
         doggo                            0
         floofer                          0
         pupper                           0
         puppo                            0
         dtype: int64
```

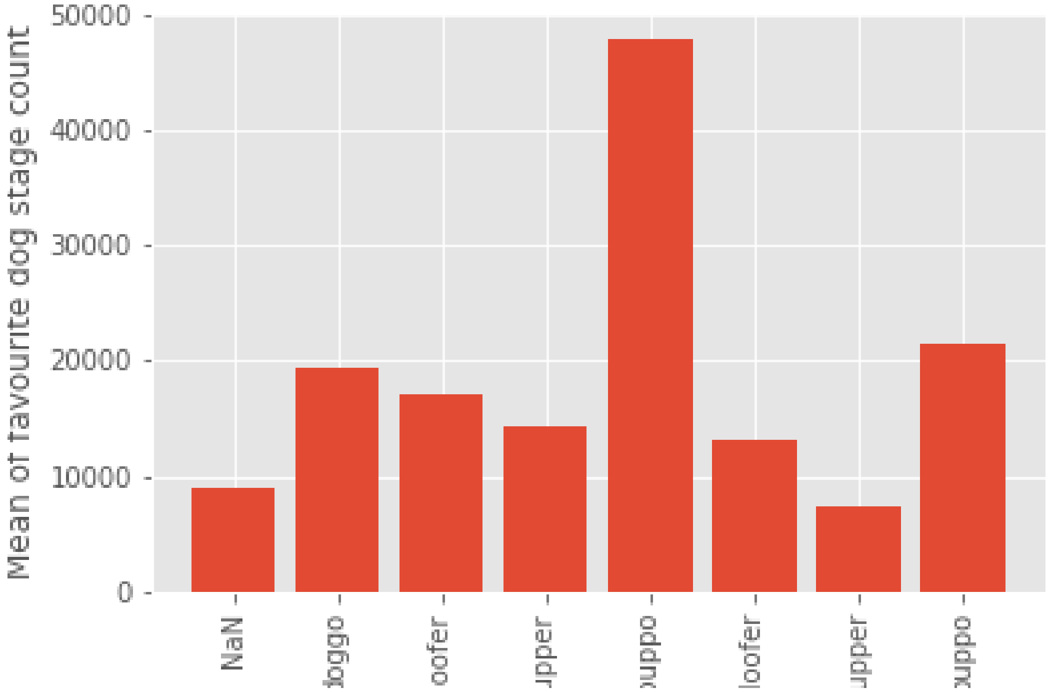
In_reply_status_id, In_reply_user_id,retweeted_status_id, retweeted_user_id, retweeted_status_timestamp indicates that most of our tweets are not retweets/replies

✓ Found duplicates in the image_predections dataframe

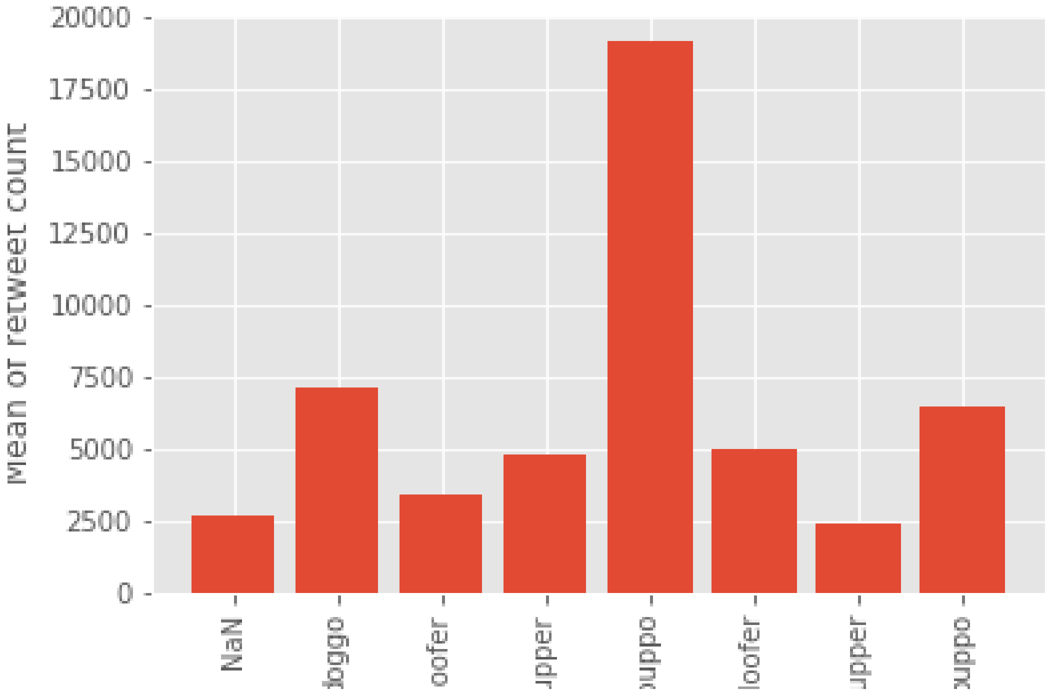


```
In [36]: df_image_predictions['jpg_url'].duplicated().sum()

Out[36]: 66

In [37]: df_image_predictions['jpg_url'].value_counts()

Out[37]: https://pbs.twimg.com/media/CU1zsMSUAAAS0qW.jpg    2
         https://pbs.twimg.com/media/CkjMx99UoAM2B1a.jpg    2
         https://pbs.twimg.com/media/C12x-JTVIAAzdfl.jpg    2
         https://pbs.twimg.com/media/CwS4aqZXUAAe3IO.jpg    2
         https://pbs.twimg.com/media/Ct72q9jWcAAhlnw.jpg    2
                                                           ..
         https://pbs.twimg.com/media/CfovbK4WIAAkTn3.jpg    1
         https://pbs.twimg.com/media/DFTH_O-UQAACu20.jpg    1
         https://pbs.twimg.com/media/C4RCiIHWYAAwgJM.jpg    1
         https://pbs.twimg.com/media/CVWGotpXAAMRfGq.jpg    1
         https://pbs.twimg.com/media/CaBP7i9W0AAJrIs.jpg    1
         Name: jpg_url, Length: 2009, dtype: int64
```

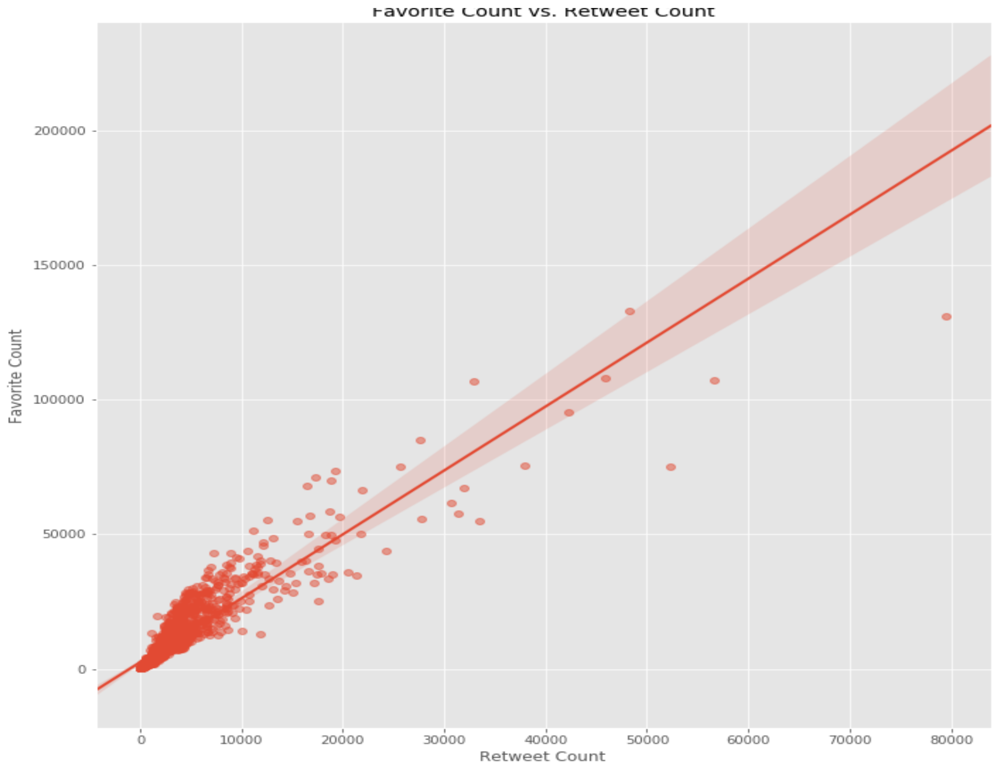
✓ The images that are classified as doggo-puppo have the highest mean favourite count.



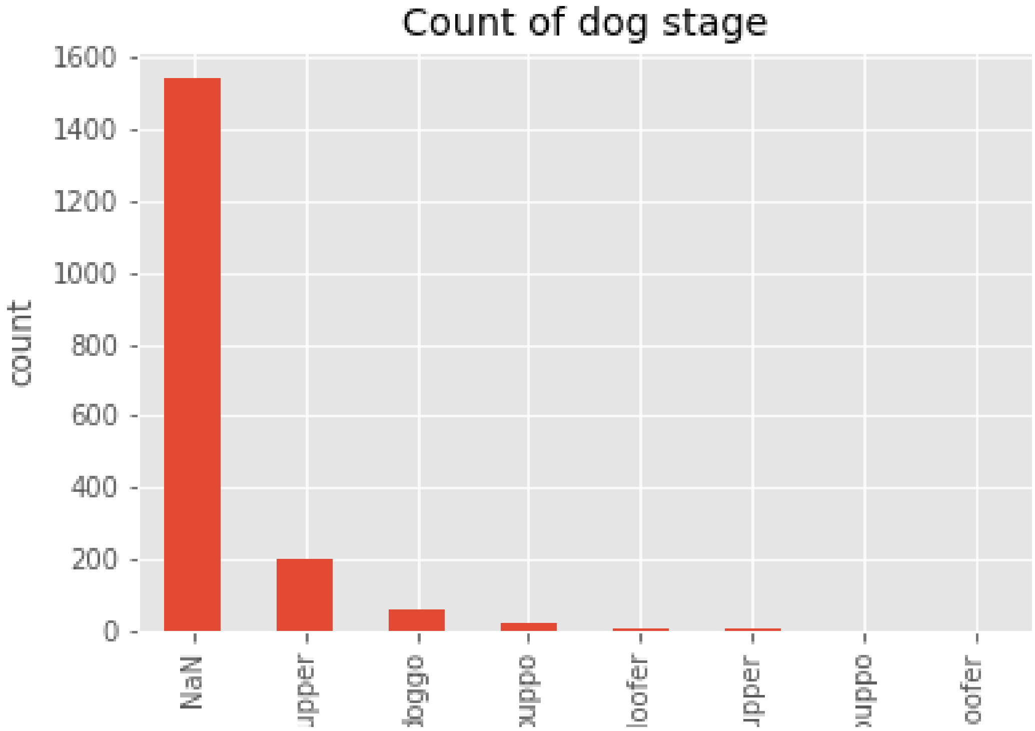
✓ The images that are classified as doggo-puppo have the highest mean retweet count and comes in second place doggo.



- ✓ From the scatter plot the favorite count is positively correlated with retweet count which means by any increase of favorite count retweet count increases.



- ✓ Aside from the NaN values pupper has the highest count among all dogs.


Count of dog stage
count
NaN
upper
oggo
uppo
loofer
pper
uppo
oofer