Wrangle Report

Introduction

In this project I used twitter API to access tweets from a twitter account called WeRateDogs, this account posts images of doges and rates them, I collected a random dataset of 2000 tweets, the WeRateDogs Twitter archive from Udacity, and the tweet image predictions from Udacity's servers. Then I started cleaning and analyzing them.

Gathering

- The first dataset : I import [twitter_archive_enhanced.csv] to Jupiter notebook and read it using pandas.
- The second dataset : it was an online file saving in Udacity's servers so I used requests library to download it then I used pandas to read it.
- The third dataset : I use tweeter API to gather the data and save it to [tweet_json.txt] file.

Assessing

In this section I tried to explore each dataset to find the issues that needs to be fixed. And I found these :

A. Quality

- Missing value in dog name and stages
- duplicated tweet_id
- missing dogs' names
- Incorrect dogs' names
- Some tweets does not contain a rating
- Incorrect datatypes
- Tweets with no images
- **Extra characters after the symbol &**
- Some columns like (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id) are missing too many values and might not be needed in this analysis
- Null value define as None
- Some of rating denominator is greater than 10

B. Tidiness

- There are 4 columns (doggo, floofer, pupper, puppo) for dogs' stages
- 3 Data frame

Cleaning

I choose some of the issues which I found while assessing the datasets and I fixed them as follows:

A. In quality part

- I drop missing value in the dogs' names and stages, duplicated tweet_id, tweets that does not contain ratings, tweets with no images, and columns that I do not need in analysis.
- I deleted extra characters after the symbol & by using (str.replace) function.
- I corrected null values that define as none to Nan .

B. In tidiness part

- I created new columns [dog_stage] and merge (puppo, pupper, floofer, doggo) with it using join() function then I deleted unnecessary columns.
- I merged the 3 datasets using merge() function.

Conclusion

After finishing cleaning I stored the cleaned file as an csv file and started the analysis part.