Untitled

January 18, 2021

1 WRANGLING AND ANALYZING THE "WeRateDogs" TWITTER ACCOUNT

```
Wrangling a dataset consists of three parts:
   - Gathering
   - Assessing
   - Cleaning
```

1.1 Gathering:

```
For this project, i gathered three datasets. The first one was the (twitter_archive_enhanced.csv
```

1.2 Assessing

```
i assessed the data in each dataset programmatically at first using pandas methods like .head()

# Issues
```

1.2.1 Quality

```
- Change Timestamp to datetime64 instead of String
- Fix numerator and denominator's wrong values
- Drop useless columns
- Change missing name values from "NaN" to None
- Change missing floofer, pupper, doggo and puppo from "NaN" to None
- Reduce dogs with multiple stage_name values to 1 stage name
- Drop invalid names
- Drop duplicated images
- Change source values to ('Iphone' , 'Vine' , 'Twitter_Web_Client' , 'TweetDeck')
```

1.2.2 Tidiness

```
- Combine the doggo, floofer, pupper and puppo columns into "dog_stage" column with type String
- Merge all the dataframes
```

1.3 Cleaning

As i didnt find any issues with the Tweet_json dataset, i created a copy of the archive and pred
- Which is the most common source used.
- What is the most favourited dog breed.
- What are the most common false predictions in the dog prediciton algorithim.

In [ ]: