Project: WeRateDogs Insights from Twitter API

Ahmed Zidane

1. Data Gathering

- I used 3 separate sources of data, the archive data, image data and tweets data.
- I couldn't query twitter APIs, I tried everything and googled my way out, but it always didn't work during the extraction of tweets info using the tweet id, so I used tweets_json.txt as recommended in the project page
- Later I will merge these 3 data sets into 1 master dataset

2. Quality Issues

- First, I checked for obvious things like datatypes, null values and duplicates.
- Then I was checking for wrong naming, outliers or data that didn't make sense

Here are the issues I found:

Quality Issues:

Archive Dataset

```
1. Some datatypes are wrong (timestamp,retweeted_status_timesta
mp)
2. Some records have denominator not equal to 10
3. 745 records have dog name = None
4. Some records are retweets and replies
5. There are some outliers in rating numerator (Very high numbe
rs)
6. Extended URL have some null values
7. Doggo, Puppo, Pupper and Floofer columns have None values, s
hould be Nan
```

Images Dataset

```
8. P1, P2, P3 has no real dog names sometimes
9. The dataset doesn't have records for all 2354 ids. it only h
ave 2075 ids
```

Tideness:

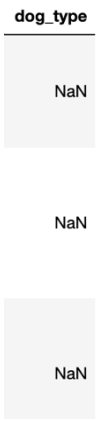
```
. All three dataset should be joined together
. Doggo, Puppo, Pupper, Floofer should be merged together in on
e column stating the type
. All retweets and replies should be deleted
```

3. Data Cleaning

in order to fix those issues, I first created a copy of each dataset, then:

- Corrected the wrong data types (`timestamp, retweeted_status_timestamp`)
- Removed records that was a retweet or a reply
- Merged Doggo, Floofer, Pupper and Puppo columns into one column to look like this

  Where if a dog type is doggo, we will have a doggo record. And if there's no type for tweet it will be Nan



- Then I removed some unnecessary columns that i am not going to use
- Then I removed 23 records that have denominator Not equal to 10

- Then I classified my numerators into 3 types
  - **Normal** (10 to 15) since most of the ratings have numerators in this range
  - **Low** (Less than 10) I found some ratings less than 10, but I needed to know if this is a typo or this is a real rating
  - **Outlier** (More than 15) some of the records have very high numbers more than 1000

  I compared a random sample of my low numerators to the text and I found that it's low on purpose, I decided to keep it as part of the analysis.
  For the Outlier High numbers, I decided to remove then because even if they are high on purpose, they will distort my analysis and skew some averages.

- Then I changed the missing dog names from None to NaN
- Last but not least, I merged the 3 datasets together