Khadeejah Alaslani

Abstract

In this project, I analyzing and visualizing is the tweet archive of Twitter user @dog\_rates, also known as WeRateDogs.

Data wrangling

**Introduction**

The aim of this project to improve my skills in wrangle and analyze data. The dataset that I will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user [@dog\_rates](https://twitter.com/dog_rates), also known as [WeRateDogs](https://en.wikipedia.org/wiki/WeRateDogs).

**Tools**

* Python
* Jupyter Notebook

**Libraries**

* Pandas
* Numpy
* Request
* Tweepy
* Joson
* Matplotlib

**Project Flow**

* **Gathering**

The data of this project is from three files:

1. **Enhanced Twitter Archive**

Reading file from twitter-archive-enhanced.csv

1. **Data via the Twitter**

 Reading the data from JSON which providing by Udacity. I cannot use twitter API because my account not approved.

1. **Image Predictions**

Reading image predictions which hosted on Udacity's servers and downloaded programmatically using the Requests library.

* **Assessing**

1. **Enhanced Twitter Archive**

**Quality issues**

1. The column (in\_reply\_to\_status\_id) representing if the tweet was original or a reply to another tweet (You only want original ratings (no retweets) that have images.). Drop all the rows in (in\_reply\_to\_user\_id) has a value that is not NaN.
2. Drop some columns like ( in\_reply\_to\_user\_id, retweeted\_status\_id, retweeted\_status\_user\_id,retweeted\_status\_timestamp,expanded\_urls) they have missing data:in\_reply\_to\_status\_id, in\_reply\_to\_user\_id : 78 instead of 2356 retweeted\_status\_id,retweeted\_status\_user\_id,retweeted\_status\_timestamp 181 instead of 2356 expanded\_urls : 2297 instead of 2356  
   and not helpful in the analysis.
3. There are invalid dog names like (a, an and the) and with bad style non-capitalized letter. Update all the name which are non-capitalized names and "None" to Null.
4. Rename the columns to readable name like: (name) to (dog\_name), (timestamp) to (tweet\_timestamp), (text) to (tweet\_text).
5. Incorrect dtype for (tweet\_id), (timestamp), (source), (rating\_numerator) and (rating\_denominator).
6. The ratings probably aren't all correct. There are (23) cases which denominator of rating (! = 10), (1) denominator of rating (==0), (2) numerator of rating (==0) and numerator of rating (>20). So, drop all of these data.
7. The tweet\_text value has the rating number with the text. So, we need to delete the rating number from text.
8. The source column extracts the important part (iPhone, Twitter, Vine, Tweet Deck)

**Tidiness Issues**

1. Add column of dog\_stages (i.e. doggo, floofer, pupper, and puppo), and drop the column (doggo, floofer, pupper, and puppo)
2. Add column of rating\_number (rating\_numerator/rating\_denominator), and drop the column (rating\_numerator, rating\_denominator).
3. **Data via the Twitter**

**Quality Issues**

1. Column id is saved as(int) datatype instead of (object) datatype & rename as tweet\_id

**Tidiness Issues**

1. Drop unneeded columns
2. **Image Predictions**

**Quality Issues**

1. Change the type of column tweet\_id to (string object)

**Tidiness Issues**

1. Change the columns (p1, p2, p3, p1\_conf, p2\_conf, p3\_conf) to readable name.
2. Drop the columns (jpg url, img\_num) no need for them

* **Cleaning**

Cleaning data, which consists of: define, code and test. It is explanation in **wrangle\_act. ipynb** all these steps.

* **Storing**

Store the clean Data Frame(s) in a CSV file with the main one named twitter\_archive\_master.csv

* **Analyzing, and visualizing wrangled data**