DATA WRANGLING

Abstract

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

Khadeejah Alaslani

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, also known as 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

2- **Data via the Twitter**

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

3- 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).

2- 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

3- 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**