Wrangle Report

Introduction

In this project, I will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Gathering

In this step we're asked to gather 3 different datasets

- twitter-archive-enhanced.csv: This file is obtained by downloading it manually from the Udacity site
- image-predictions.tsv: This file is obtained by downloading it programmatically using the python's requests library
- tweets.txt: this file is obtained by calling the twitter API for each tweet, using the tweet id as parameter, which we obtained from the twitter-archive-enhanced.csv data

Assessing

In this step, we try to gather some quality and tidiness issue from the data we have collected in the gathering step.

Quality & Tidiness

After assessing, we found some quality & tidiness issues. The issues are listed below

twitter-archive-enhanced.csv

- Column timestamp should be datetime instead of object
- Column source is wrapped in html
- Column rating_numerator contains false data; all the data should be greater than 10
- Column rating_denominator contains false data; all the data should equal to 10
- Column name contains false data, some dogs' name is None and a

tweets.txt

- Contains retweet data
- Column entities contains redundent data
- Column extended_entities contains redundent data

image-predictions.tsv

- Some jpg_url are duplicated

Tidiness

- tweets.txt and image_predictions should be part of twitter_archive table
- Entities data seems to contain image information which are already contained in the twitter archive data, like the image_url and extended url
- Extended entities column contains duplicate information of the entity’s column

Cleaning

In this step we are trying to clean the data, hence we will clean some of the issues we mentioned above.

twitter-archive-enhanced.csv

- Problem: Column timestamp should be datetime instead of object
  Solution: Change the timestamp to datatime
- Problem: Column rating_numerator contains false data; all the data should be greater than 10
  Solution: Extract the rows that the rating_numerator greater than 10
- Problem: Column rating_denominator contains false data; all the data should equal to 10
  Solution: Extract the rows that the rating_denominator equal to 10
- Problem: Column name contains false data, some dogs' name is None and a
  Solution: Drop the rows that the name contains None or a

tweets.txt

- Problem: Contains retweet data
  Solution: Remove retweet data based on retweet status
- Problem: Column entities contains redundent data
- Solution: Drop the column
- Problem: Column extended_entities contains redundent data
- Solution: Drop the column

image-predictions.tsv

- Problem: Some jpg_url are duplicated
  Solution: Drop duplicated data (keep the first one)