Wrangle report

4. February. 2019

Table of Contents

Introduction

The goal of this project is to wrangle the WeRateDogs Twitter data to create interesting and trustworthy analysis and visualizations. The challenge is that the Twitter archive is great, but it only contains very basic tweet information that comes in JSON format. For a successful project, I needed to gather, asses and clean the Twitter data for a worthy analysis and visualization.

The dataset that we will be wrangling, analyzing and visualizing is the tweet archive of Twitter user [@dog_rates](https://twitter.com/dog_rates)

As described above the wrangling process contain three steps:

1. Gathering data.
2. Assessing data.
3. Cleaning data.

Gathering data

In this project we will gather Data from three sources as described below:

- The WeRateDogs Twitter archive. We will download this file manually from: [twitter_archive_enhanced.csv](https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv)
- The tweet image predictions, this file (image_predictions.tsv) hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
- Tweet's retweet count and favorite (tweet.json.txt) I downloaded from udacity because I couldn't gain twitter developer account- as described in project details (Twitter API) section if I cannot get this file from twitter I can proceed with the "Gathering Data", Then read this tweet_json.txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count."

Gathering summary

Gathering data is the **first step in the data wrangling process and we gathered the data from various resources:**

- Getting data from an existing file (twitter-archive-enhanced.csv) Reading from csv file using pandas.
- Downloading a file from the internet (image-predictions.tsv) Downloading file using requests.
- Read tweet_json.txt file line by line into a pandas dataframe then convert to csv file.

Assessing data

After gathering the data, assess this data will be the next step to be sure that the quality of this data is suitable to move to last step which is the cleaning step.

Quality

Content issues

archive dataset

- Timestamp, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be we will change the data type of all columns.
- In several columns null objects are non-null (None to NaN).
- Some entries have invalid values i.e. 'None', 'a', 'an'.
- The ratings are not extracted correctly some have error.

images dataset

- Some tweet_ids have the same jpg_url.
- Some tweets are having 2 different tweet_id one redirect to the other.

tweet dataset

- The tweet_id column should be named same in all the DataFrames and its datatype should be same in all the tables (id_str column name to tweet_id so I can merge the file with other files [to be a key column in all dataset].

In all dataset

- The name column has many invalid values like, a, an, the.

Tidiness

structural issues

- No need to all the informations in images dataset, (tweet_id and jpg_url what matters)
- Merge 4 columns into 1 column (Merge the 'doggo', 'floofer', 'pupper' and 'puppo' columns into one column named 'dog_stage') in archives dataset.
- All tables should be part of one dataset.

Additional

- We may want to add a gender column from the text columns in archives dataset.

Cleaning data

Cleaning our data is the third step in data wrangling. It is where we will fix the quality and tidiness issues that we identified in the assess step.

We used the two types of cleaning, the manual and programmatic even the manual not recommended but the issues were one-off occurrences. Our process was Define, Code and Test. We didn't spot all the quality and tidiness assessments at the assessing data section, so we have been iterating and revisiting assessing to add these assessments to our notes.

Conclusion

Data wrangling indeed is a core skill that everyone who works with data should be familiar with since so much of the world's data isn't clean. If we analyze, visualize, or model our data before we wrangle it, our consequences could be making mistakes, missing out on cool insights, and wasting time. We couldn't be able to make some of the visualization without wrangling (i.e. dog gender partition) So always best practices is wrangling data.