Wrangle Report

WeRateDogs Analysis Project

Introduction

This report aim to summarise how I wrangle the data in the project of WeRateDogs Analysis and the followings will be divided into three parts: gathering data, assessing data and cleaning data.

Gathering data

Data is collected in three methods:

(1) **WeRateDogs Twitter archive**: Downloaded the .csv file provided by Udacity - twitter_archive_enhanced.csv. Read the .csv file into DataFrame "twitter_archive".

(2) **Tweet image predictions**: Downloaded image_predictions.tsv file from the Udacity's Servers programmatically using the Requests library. Read the .tsv file into DataFrame "image_predictions".

(3) **Tweet's retweet count and favourite count**: Queried the Twitter API for each tweet's JSON data using Python's Tweepy library and store them in tweet_json.txt file., which was then read line by line into a DataFrame "df_api".

Takeaway: Be careful when processing JSON data. When dealing with the JSON data, I failed to queried the targeted data, failed to write them into different lines and had to queried several times from Twitter API which took me lots of time.

Assessing data

Three datasets were assessed visually and programmatically for quality and tidiness issues. The issues identified are as follows:

(1) Quality Issues:

twitter_archive

- re-tweeted, and replied tweets should be removed as we only want original ratings.
- Variables 'rating_denominator' and 'rating_numerator' have wrongly extracted.
- Some tweets are not dog rating tweets.
- Some dog names are wrongly extracted.
- Some dogs have more than one dog stage (wrongly extracted & two dogs in a pic).
- Source should be truncated.

image_predictions

- There are some duplicated data in pic prediction.
- Some dog breeds starts with capital alphabet and some don't.

df_api

- Some tweets are not available, resulting in missing retweet count or favourite count.
- There are duplicated data that should be removed.

(2) Tidiness

- Three dataset should be integrated into one.
- In twitter_archive, dog stage should be integrated into one variable.

(3) Takeaways:

- Always remember to check duplicated data. Particularly, in this dataset, need to check duplicated tweet_id.
- Don't overthink.

Cleaning data

Datasets were first copied and saved as twitter_archive_clean, image_predictions_clean, df_api_clean. They were then cleaned according to the issues mentioned in the previous section, presented below.

(1) Quality issues

- Delete data that are retweets or replies and drop useless columns.
- Deal with variable 'rating_denominator' with number other than 10: wrongly extraction; more than one dogs - averaged; not dog rating tweets - deleted. Then drop it as all data have rating_denominator of 10.
- Deal with variable 'rating_numerator' that has extreme values: A tweet that is not dog rating, need deleted; Some tweets are wrongly extraction and need adjusted. Then rename the variable as “rating”.
- Deal with dog names that are wrongly extracted. They are wrongly extracted and should be change to None.
- Check all data with at least two stages. Adjust the data have wrongly extracted dog stages.
- Remove tweets did not have pictures (not in image_predictions_clean).
- Truncated variable source to 'Twitter for (platform)'.
- Change timestamp into data type datetime.
- Change all the types predictions to words with the first letter capitalised in dataset image_predictions_clean.
- Remove duplicated data and useless columns in dataset df_api_clean.

(2) Tidiness¶

- Integrate three datasets into one. Drop data with incomplete information
- Integrate four variables regarding Dog stage into one variable.

(3) Takeaways:

- Save the data as copy before cleaning.
- The order of cleaning does matter. A proper order might save much time.

Conclusion

The final cleaned dataset includes 1962 entries and 20 columns, and was saved as twitter_archive_master.csv file.