Wrangle Report

Introduction

This project is about wrangling data from twitter 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. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. WeRateDogs has over 4 million followers and has received international media coverage.

In this project, we want to collect data from various sources, assess and clean efficiently to use it in our analyses.

1- Data Gathering

In this part, data was gathered using three different sources:

1. Twitter archive enhanced (csv file).

First data, a tweet archive from WeRateDogs Twitter account was gathered manually by using the file provided. (twitter_archive_enhanced.csv)

Download twitter archive from this link:

https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv

2. Images predictions for dogs (tsv file).

The tweet image predictions, I downloaded it using the Requests library and the following URL:

Download image predictions from this link:

https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3. Tweet Data from twitter API (txt file).

Third data, in which you can find retweet and favorite counts of each tweet, was gathered using Twitter API and Tweepy library. This file was stored in JSON format. I used the twitter API using my developer account this time but we can download twitter data using links that udacity provide.

Download twitter data from this link:

https://video.udacity-data.com/topher/2018/November/5be5fb7d_tweet-json/tweet-json.txt

Or download twitter data from this link:

https://video.udacity-data.com/topher/2018/November/5bf60fc9_tweet-json/tweet-json.zip

2- Assessing data

Both visual assessment and programmatic assesment were performed. In visual assessment, data was scrolled from start to end. In programmatic assessment, various functions were used such as .info(), .describe() and value_counts(). After the assesment, the issues were listed below.

Tidiness:

- 1- We have 4 seperate columns of dog categories, we shoulld merge them in one column.
- 2- We have 3 datasets, twitter_archive, image_predictions, and tweets_json dataset, we should merge it in one dataframe.

Quality:

Twitter-archive table:

- 1- There are 181 retweets indicated by retweeted_status_id.
- 2- Some dog names are invalid (None, an, &).
- 3- Invalid tweet_id data type.
- 4- invalid timestamp data type.
- 5- Sources difficult to read.
- 6- Some rows have less than 10 rating_denominator.

image_predictions table:

- 1- Missing photos for some ids.
- 2- Underscores are used in names in columns p1, p2, & p3.
- 3- Some P names start with uppercase letter, other starts with lowercase.

tweets_info table:

- 1- Missing entries.

3- Cleaning Data

3.1- Creating copies from dataframes

3.2- Cleaning tidiness issues

Dog stage data is separated into 4 columns.

Define:

Merge 4 columns into 1 called dog_stage.

Data is separated into 3 dataframes.

Define:

Merge 3 dataframes into one dataset based on tweet_id.

3.3- Cleaning quality issues

Not quality issues will be cleaned.

There are 181 retweets indicated by retweeted_status_id

Define:

Delete rows that represent retweets and all related columns.

Some dog names are invalid (None, an, &).

Define:

Convert invalid names to NaN and extract the correct names from text columns after named.

Invalid tweet_id data type.

Define:

Correct invalid data type from integer to string.

Invalid timestamp to datetime.

Define:

Correct invalid by converting timestamp to datetime.

Sources difficult to read.

Define:

Change sources to more readable categories.

Some rows have less than 10 rating_denominator.

Define:

Remove rating_denominator which less than 10.

Missing photos for some ids.

Define:

Delete rows with missing photos.

Underscores are used in names in columns p1, p2, & p3.

Define:

Changing underscore to spaces.

Some P names start with uppercase letter, other starts with lowercase.

Define:

Convert lowercase letters to uppercase letters.

Missing entries.

Define:

Delete rows without retweet_count entries

4- Storing data

The last version of the merged data explained in Section 3 was saved and stored as a csv file called 'twitter_archive_master.csv'.