Wrangling Report

By Afnan Alabdan

Introduction

Real-world data rarely comes clean. Using Python and its libraries, I have gathered data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it.

The dataset that wrangled 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. These ratings almost always have a denominator of 10.

The data Wrangling process:

- Gathering data.

- Assessing data

- Cleaning data.

Gathering data

In this section I have gathered the data from the following three sources:

- The WeRateDogs Twitter archive, provided to Udacity students to download it manually, upload it and read the data into a pandas DataFrame.

- The tweet image predictions, download it programmatically using the requests library and the URL.

- Additional data from the Twitter API, while I wasn’t able to extract the data through API since my developer accounts in twitter wasn’t approved, so I have used the option of getting the data from ‘json-text.txt’.

Assessing data

In this section, I have assessed the data in both two ways: visual assessment and programmatic assessment. Also, I have documented each quality and tiredness issue. I founded the following 8 Quality issues and 2 Tiredness issues:

Quality issues

1. Not all the rows are original tweet, there are 181 retweet.

2. tweet_counts_df has inconsistent id column with the other tables.

3. archeive_df has a timestamp column with a string data type.

4. There are columns that wont be used in the analysis.

5. there are a naming issues with 'a, an , and ' dog names.

6. There are duplicated images urls.

7. There are 23 rating denominator does not equal 10

8 .The datatype of tweet_id is integer.

Tidiness issues

1. Four columns that specify the stage of each dog.

2. all tables should be in one dataset.

Cleaning data

In this section, I started cleaning the documented issues that was found in the ‘Assessment part’ by Defining the issue and write the code to clean it, after that test our cleaning code and I have repeated it to each issue need to be solved.

Storing the data

After wrangling the project successfully, I have stored it to twitter_archive_master as csv file.