Project: Wrangle and analyse data

Table of Contents

Introduction

This academic project aims to wrangle data from @dog_rates Twitter, also known as WeRateDogs, create interesting and trustworthy analyses and visualizations. Additional gathering, then assessing and cleaning is required for "Wow!"-worthy analyses and visualizations.

Find below main tasks:

1. Data wrangling, which consists of:

    1.1. Gathering data (downloadable files and querying Twitter API).

    1.2. Assessing data;

    1.3. Cleaning data;

2. Storing, analyzing, and visualizing the wrangled data;

3. Communication and Reporting on my data wrangling efforts, data analysys and visualizations, as well.

Gathering Data

In this project I'm going to work with the data below:

1. The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets: **twitter_archive_enhanced.csv**

    1.1. DogoDictionary:

```
doggo - a big pupper, usually older;

pupper - a small doggo, usually younger;

puppo - a transitional phase between doggo to pupper
;

blep - one's tongue protuding ever so slightly from
the mouth as an extremely subtle act that occurs;
```

```
snoot - the dog's nose;

floof - any dog really.
```

2. The tweet image predictions: **image-predictions.tsv**

3. Using the tweet IDs in the WeRateDogs Twitter archive, I'm going to query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

Assessing Data

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues. In this session, I am going to detect and document at least:

- eight (8) quality issues and
- two (2) tidiness issues in your wrangle_act.ipynb Jupyter Notebook.

To meet specifications, the issues need to satisfy the key points:

- I only want original ratings (no retweets) that have images.
- Though there are 5000+ tweets in the dataset, not all are dog ratings and some are retweets.

Cleaning Data

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. This data is usually not necessary or helpful when it comes to analyzing data because it may hinder the process or provide inaccurate results.

Data cleaning is not simply about erasing information to make space for new data, but rather finding a way to maximize a data set’s accuracy without necessarily deleting information.

We can consider two key points for Data Cleaning:
- Data Quality > Accuracy and Consistency
- Data Tydiness > Structure, Standard and Organise Data

1. FINDINDS:

1.1. Archive Dataset

- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be integer instead of float and there are only 78 rows non-null

- retweeted_status_timestamp, timestamp should be datetime instead of object
- Some of rating_numerator values are relatively high (24 rows contains value > "20")
- Some of rating_denominator values seems to be inconsistent (23 rows contains value <> "10")
- Several columns contains "None" values (None to NaN)
- Invalid column Name: i.e "None" & "a", "the", "an", etc
- Dog's Stage are organized in 4 columns (doggo, floofer, pupper, puppo)
- We want to keep only original ratings (no retweets) & having images (if in_reply_to_status_id <> "" or retweeted_status_id <> "")

1.2. Predictions Image Dataset

- Missing values from images dataset (2075 rows instead of 2356)
- Some tweet_ids have the same jpg_url (66 rows)
- There are " _ " and " - " in the column name

1.3 Json Tweets Dataset

- created_at should be datetime instead of object (format='%Y-%m-%d %H:%M:%S')

2. TASKS: In this project I'm focusing in Quality & Tidyness as assessing and cleaning the entire dataset completely would require a lot of time and is not necessary to practice and demonstrate your skills in data wrangling.

2.1. DEFINE Quality Issues

2.1.1. Replace "a", "the", "an", etc values in the column name to NAN value

2.1.2. Delete retweeted rows and drop the columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id as there are only 78 rows non-null

2.1.3. Change the data type of the columns: timestamp (FROM object TO datetime)

2.1.4. Fixing rating_numerator relatively high

2.1.5. Fixing rating_denominator <> 10

2.1.6. Replace " _ " and " - " in the column name

2.1.7. Replace "None" value to NAN value

2.1.8. Delete Missing values from archive where there is no rows in the images dataset (2075 rows instead of 2356)

2.2. DEFINE Tidyness Issues

2.2.1. Create a DogStage column, storing "doggo", "floofer", "pupper", "puppo"

2.2.2. Rename columns name: p1, p1_conf, p1_dog, p2, p2_conf, p2_dog to understable names

Conclusion of the learning

Amazing job! In this project, I've learned about the 3 main process of Data Cleaning, such as Gathering, Assessing and Cleaning. My main difficult was about accessing Twitter API and understand rating definition.