Topic: Wrangle and Analyze Data

Prepared by Ifeoma Onyiuke

Date: 26th March 2021

INTRODUCTION:

This report is prepared as part of the Udacity data wrangling project carried out for the @WeRateDogs twitter data. It briefly describes the wrangling efforts (stpes and methods applied to make this data ready for analysis and creating visualisations.The aim of this project is to wrangle @WeRateDogs Twitter data in order to create interesting and trustworthy analyses and visualisations.

This project is made up of three phases;

1. Data wrangling which is made up of three steps:
   - ✓ Data gathering
   - ✓ Data assessment
   - ✓ Data cleaning
2. Data Storage
3. Analysing and visualising wrangled data

DATA GATHERING

Data gathering for this project was carried out in three phases.

Phase1: Manual data download:

Weratedogs twitter archive data, which contains some tweets, dog ratings was provided by Udacity and I downloaded this data from the resources section for the course and saved on my local storage as a csv file(**twitter_archive_enhanced.csv).**

Phase2:

The second data needed for this project, **the tweet image projection data**. This data was downloaded programmatically using a **get request** from the Udacity's server and saved on my local storage as a tsv file**(image_predictions.tsv)**

URL for assessing image prediction data: (**https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv**).

Kindly go here to see detailed codes and steps.

Phase 3:

The data dataset required for this project was downloaded using twitters' API Tweepy. Using the Tweet_ID for each tweet contained in the archived data from **Phase1.** I downloaded each tweets **retweet count and favorite("Like") count** for this twitter handle.

- Extract tweet IDs from the twitter archive data downloaded in pahse 1
- Query the twitter API for each tweet's JSON data using Python's Tweepy library

- Store each tweet's entire set of JSON data in a file called twwet_json.txt
- Each tweet's json should be written on its own line
- Read the .txt file line by line into a pandas frame with these columns (tweet_id, retweet_count, favorite count).
- Other issues: ratings are not probably correct/dog name/dog stage
- Assess amd clean these columns to be used for analysis and visualisation

ASSESSING DATA

After gathering the data, the next step was to assess the gathered data. I did this programmatically, and some of them manually. I used pandas functions to assess the tidiness and the quality of the data.

Quality: Low quality data is commonly referred to as dirty data. Dirty data has issues with its content. The Data Quality Dimensions are **Completeness, Validity, Accuracy and Consistency.**

Tidiness: Untidy data is commonly referred to as "messy" data. Messy data has issues with its structure. Tidy data is where:

1. Each variable forms a column.

2. Each observation forms a row.

3. Each type of observational unit forms a table.

Some of the Tidyness and Quality issues discovered during this phase are as follows:

ASSESSING THE TWITTER_ARCHIVED_DATA

QUALITY ISSUES

- missing data: in_reply_to_status_id and in_reply_to_user_id has only 78 rows available out of the 2356.
- not original tweet, columns retweeted_status_timestamp, retweeted status_id and user_id means that rows are retweet and not original tweet. (drop rows)
- dog name is none for 745 rows - check if the same rows with retweet
- dog names recorded as a or an should be None
- rows with a also have dog style
- does not contain retweet and favorite counts
- inconsistent rating_denominator -value should be 10, values greater than 10 should be removed
- very high rating_numerator as much 1776, not an issue but keep in mind
- time stamp has object data type, change to datetime
- some rows have several identical values in the expanded_url column concatenated by a comma.

TIDINESS ISSUES

- create one column for dog stage, collapse multiple colunmns and rows. convert the rows (name, doggo, floofer,pupper,puppo into two columns, one with dog name and one with dog level specifying either from the list)
- expanded url has multiple url on one row
- the three datasets can be one,all have tweet_id

ASSESSING IMAGE PREDICTIONS DATA

Quality issues

- inconsistent naming for dog breeds p1, remove underscore
- convert prediction number to type int and remove letter p
- inconsistent labelling for dog breeds, convert all breed name to lower case
- some breed predictions are false

Tidiness issues

- Predictions are spread in three columns.
- Confidence intervals are spread in three columns.
- Dog tests are spread in three columns.
- Melt all three into two columns (breed and confidence)

ASSESSING RETWEET AND FAVORITE COUNT DATA

Quality issues

- first column(unamed 0) not needed - drop

CLEANING

After assesing the various datasets, the issues highlighted were then corrected to make the data clean and ready for visualisation. I followed three steps to complete each cleaning phase.

- ✓ Define: Define the issue
- ✓ Code: The code to correct the issue define.
- ✓ Test: To ensure method applied was functional and error/issue corrected

Detailed code and information for each step can be found in this Jupyter notebook.

DATA STORAGE

Cleaned datasets were merged using Pandas merge and stored as a .csv called twitter_archive_master.

VISUALISATION AND DATA ANALYSIS

I used the cleaned data to create some visuals and analysis to give more insight to the data and if possible provide an understanding of the rating system applied by @WeRateDogs before rating dogs. See results and insights in act_report.

CONCLUSION

Completing this project gave me the opportunity to try out all the data wrangling steps and processes learned in the classroom. It was also an opportunity for further research. Assessing more data from twitter was an invaluable experience and I am glad to have completed it. I had the impression that it would be the easiest part of this nanodegree but so far it has been the toughtest. I have a deeper understanding of Data wrangling processes, developed more skills in use of Python libraries pandas, seaborn, matplotlib, Numpy etc.

In conclusion, real-world data rarely comes clean. Using Python and its libraries, I gathered the necessary data from a variety of sources and in a variety of formats, assessed the quality and tidiness and then clean it. A new dataset was created which contains the the three various datasets assessed and cleaned. This was used to create visuals and analysed for more insights. Details and information are contained in the second part of this report called act_report.

RESOURCES

Stack overflow

https://www.earthdatascience.org/courses/use-data-open-source-python/intro-to-apis/twitter-data-in-python/

https://towardsdatascience.com/twitter-analytics-weratedogs-a441be7d4a85