Mehdi Iddar
2019

WRANGLE REPORT

In this project, I conducted a data wrangling process through gathering data from a variety of sources and in a variety of formats:

- First is downloaded manually a .csv file named 'twitter_archive_enhanced.csv' and stored it in 'archive' table
- Then, I used the Requests python library to download programmatically a '.tsv' file named 'tweet-image-predictions.tsv' and I stored it in 'images' table. This file contains the results of a neural network's analysis which predicts a dog's breed based on images.
- After this, I created an API object that I used to programmatically download a Json file stored as 'twitter_counts' table, which contains additional Twitter data.

In the second section of the project, which is devoted to data assessing, I first, looked for quality issues that pertain to the content of data I identified ten quality issues:

For the Archive Table :

- Erroneous datatype : timestamp and retweeted_status_timestamp should be datetime
- Erroneous datatype: tweet_id should be string
- Doggo, floofer, pupper, and puppo should be categories
- Denominator is not 10 for 23 tweets
- unnecessary html tags in source column in place of utility name

For the Images Table :

- Erroneous datatype : tweet_id should be string
- ('p1', 'p1_conf', 'p1_dog', 'p2','p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog') Columns names are not informative. Names should be changed to become more informative.
- Erroneous datatype : p1, p2 and p3 should be categorical

For the Twitter_counts Table:

- Erroneous datatype: tweet_id should be string

Then I examined tidiness issues, Tidiness issues pertain to the structure of data. The requirements for tidy data are:

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational unit forms a table.

I identified two tidiness issues:

- For the archive table : doggo, floofer, pupper, puppo should be categories of a single variable named "dog_stage"

- Archive and twitter_counts can be consolidated into a single table for which the observational units are tweets. Images can be left as-is, because the observational units are images.

In the last section of the wrangling process I structured and cleaned dirty data into the desired format for better analysis and visualizations using Python and its libraries. For each identified issue, I defined the actions to undertake before translating those actions to lines of code. I also tested every code to check the result of the cleaning.