ND111 - Project 02 - Data Science II

Wrangle and Analyze Data

Anderson Hitoshi Uyekita

30 December 2018

Wrangle Report

Synopsis

Along the Data Wrangling process, in the `twitter_archive_enhanced.csv` file, I have found several problems in the dog's name column, probably the regex used to gather/find it (from the Twitter user `@dog_rates` also known as WeRateDogs™ (https://twitter.com/dog_rates)) was not well calibrated, and in many cases has gathered articles, nouns, etc. or any other ordinary word. I have fixed it assuming these problematic dog's names as `None` .

I have also found problems in `rating_numerator` and `rating_denominator` columns, both from `image_predictions.tsv` file, which has required a new process of "scrapping" these values from the `text` column.

Finally, I have combined the files `twitter_archive_enhanced.csv` and `image_predictions.tsv` into a new data frame called `twitter_archive_master.csv` , which I have aggregated some new features:

- `retweet_count` , and;
- `favorite_count` .

Both features, are gathered from the WeRateDogs™ tweets using the `tweepy` package.

1. Introduction

This Wrangle Report is a part of a Data Science Course Project offered by Udacity (ND111 - Data Science II (https://br.udacity.com/course/fundamentos-data-science-ii--nd111)). The project aims to gather data from Twitter and combine it with a third party data frame to create analysis about the tweets and the predicted dog's breed.

2. Data Gathering

I have gathered the files `image_predictions.tsv` and `twitter_archive_enhanced.csv` using the `requests` package. Although the `image_predictions.tsv` file has almost all the information from the WeRateDogs™ user, there is some missed variable, which I have gathered using the `tweepy` package.

3. Data Assessing

The Data Assessing process have found several issues, which I have detailed in Table 1:

Table 1 - Summary of Issues Identified.

Issue ID	Table	Issue Type	Dimension	Method	Column	Description
1	df_ach	Quality	Validity	Visual	name	Invalid names or non-standard names.
2	df_ach	Tidiness	-	Visual	source	HTML tags, URL, and content in a single column.
3	df_ach	Quality	Validity	Programmatic	rating_numerator	Invalid ratings. Value varies from 1776 to 0. Data Structure must be converted from `int` to `float` .
4	df_ach	Quality	Validity	Programmatic	rating_denominator	Invalid denominator, I expected a fixed base. Data Structure must be converted from `int` to `float` .
5	df_ach	Tidiness	-	Programmatic	doggo, floofer, pupper, and puppo	This is a categorical variable, and I can combine these columns into one column.
6	df_ach	Tidiness	-	Programmatic	text	There is two information in a single column. Split the text from the URL.
7	df_ach	Quality	Validity	Programmatic	timestamp	Convert to date.
8	df_ach	Quality	Validity	Programmatic	tweet_id	Following the example of zip code, it must be a string.
9	df_ach	Quality	Accuracy	Programmatic	retweeted_status_id	The same dog could be recorded twice or more in cases of retweets.
10	df_ach	Quality	Accuracy	Programmatic	in_reply_to_status_id	The same dog could be recorded twice or more in cases of reply.
11	df_img	Quality	Consistency	Visual	p1, p2, and p3	Dog's breed has no standard. Capital letter or lowercase names.
12	df_img	Quality	Validity	Programmatic	tweet_id	Convert to string.
13	df_img	Quality	Validity	Programmatic	jpg_url	It has duplicated images and consequently double entry.
14	twt_ach_mstr	Tidiness	-	Programmatic	-	Merging these two tables (`df_ach` and `df_img`) into one.
15	df_img	Quality	Completeness	Programmatic	"retweet count"	Gather additional info in `tweet_json.txt` file.
16	df_img	Quality	Completeness	Programmatic	"favorite count"	Gather additional info in `tweet_json.txt` file.

Issue ID	Table	Issue Type	Dimension	Method	Column	Description
17	twt_ach_mstr	Quality	Validity	Programmatic	“many columns”	Remove `in_reply_to_status_id` , `in_reply_to_user_id` , `retweeted_status_timestamp` , `retweeted_status_id` , and `retweeted_status_user_id` .

Legend:

- `df_ach` : Loaded data frame from `twitter_archive_enhanced.csv` ;
- `df_img` : Loaded data frame from `image_predictions.tsv` , and;
- `twt_ach_mstr` : Loaded data frame from `twitter_archive_master.csv` .

4. Data Cleaning

The dog’s names issue was solved evaluating if it starts with a capital letter it was a name if not it was an ordinary word and I have converted to “None”. Most of the issues involving non-usual values to rating_numerator and rating_denominator were solved using a new tailored regular expression to gather the ratings from `text` column.

In respect to the data type problems in `timestamp` and `tweet_id columns` , were fixed using the `.astype()` method and `.loc[]` .

In regard to the duplicated information, I decided to remove all retweets and reply to avoid double entries of the same dog.

Finally, I have solved the tidiness issues combining the tables `twitter_archive_enhanced.csv` and `image_predictions.tsv` in one called `twitter_archive_master.csv` . I have also merged 4 columns (doggo, pupper, puppo, and floofer) into one, which I have bundled and named as dogtionary.

5. Conclusions

I have documented 17 issues but this final file version is not totally free of issues, because I faced the Data Wrangle as an iterative process, what I did so far was the first iteration.

For this reason, the `twitter_archive_master.csv` file is the final file version with a minored number of issues, and ready for a Data Analysis. This file has 1968 observations and 24 features.

Caveats.: Bear in mind, there are some `tweet_id` that do not have `retweet_count` and `favorite_count` , which means there are observations with NaN.

Additional Info

For further information about Project 02 from Data Science II, you can access the following link:

- ND111 - Project 02 - Repository (https://github.com/AndersonUyekita/ND111_data_science_foundations_02/tree/master/03-Chapter03/00-Project_02) (Github Repository)
- ND111 - Project 02 - Wrangle Act (https://github.com/AndersonUyekita/ND111_data_science_foundations_02/blob/master/03-Chapter03/00-Project_02/wrangle_act.ipynb) (Jupyter Notebook File)
- ND111 - Project 02 - Act Report (http://rpubs.com/AndersonUyekita/nd111_project_02_act_report) (Markdown File)
- ND111 - Data Science II - Nanodegree Repository (https://github.com/AndersonUyekita/ND111_data_science_foundations_02) (Github Repository)

A work by AH Uyekita (https://andersonuyekita.github.io/site/cv.html)

anderson.uyekita[at]gmail.com

(https://linkedin.com/in/andersonuyekita/) (https://github.com/AndersonUyekita/)