Data Wrangling Process Report

By Amr Mohamed ElHelly

March 2021

A report detailing steps used in Udacity Data Wrangling project as part of Data Analysis Professional Track Nanodegree. It contains a step-by-step description of the analysis on WeRateDogs twitter account.

First Step: Data Gathering

Three sources were used to gather the needed data for the project:

1. twitter-archive-enhanced-2.csv file. This file was downloaded from Project page on Udacity under Supporting Materials section. The file was imported as a Pandas DataFrame using "pandas.read_csv" function. The file contains tweet data of WeRateDogs twitter account such as tweet time, text and ID.

2. image_predictions.tsv file. This file was downloaded from an online page using Python "requests" library and read as a Pandas DataFrame also. It contains predictions on images included in WeRateDogs tweets whether they contain a dog or not with a specific confidence level.

3. tweet-json copy.json file. This file was obtained through Udacity Supporting Material section and processed using Python "json" library line by line to extract favorite count, retweet count and user count of WeRateDogs. This file is queried from Twitter API but it also requires Twitter Developer account, so I only reviewed the code.

Second Stage: Data Assessing and Data Cleaning

Data Assessment

Both Visual and Programmatic assessments were conducted on all data sets in order to assess their Quality and Tidiness:

Visual Assessment

- Visual assessment was carried out first by printing out the entire DataFrames into Jupyter Notebook and scrolling through them to detect issues.
- However due to columns and rows grouping done automatically by Pandas, not all columns were visible so all DataFrames were exported as CSV files and viewed using Microsoft Excel.

Programmatic Assessment

- Programmatic assessment is more important when dealing with large data sets and has been done using Pandas DataFrame statistics functions mainly "info()" function.

Data Cleaning

Ten Quality issues and two Tidiness issues were detected and dealt with across the three data sets, and below is a detailed breakdown of them:

Data Sets	Issue	Solution
	Quality Issues	
Twitter Archive	There are tweets that are not original tweets instead retweets and replies	Removed some of the retweets and replies guided by in_reply_to_status_id column and image_predictions data set
	Twitter archive file has two columns about retweets that have many empty cells.	Removed mentioned columns using Pandas DataFrame manipulation functions.
	Not found dog names are represented as None not NaN.	Replaced None values with NaN to not interfere with data analysis.
	Dog types in twitter archive file contain None instead of Nan	Replaced None values with NaN to not interfere with data analysis.
	some dog names are erroneous such as "a","an" and "the".	Replaced the incorrectly extracted dog names with NaN.
	Timestamp column needs to be converted to datetime in order to facilitate analysis.	Converted its type into datetime.
	Timestamp column needs to be split into month and year columns for clarity.	Used Pandas DataFrame manipulation tool to extract month and year data into their own columns.
Twitter API Data	Twitter API data file should be merged with twitter archive to represent one observational unit about each tweet.	Merged both data sets together based on their tweet_ids.
Image Predictions	Image predictions file, p1_conf and p1_dog are not descriptive column names.	Changed their names into p1_confidence and p1_IsDog?.
	p2 and p3 are low confidence and low-quality indicators that do not add value.	Removed mentioned columns to reduce distractions.
	Tidiness Issues	
Twitter Archive	Dog names should be merged into one column named dog type.	Merged the four columns into one column that represents the dog type according to WeRateDogs.
Image Predictions	Image predictions file does not have separate columns, instead all data are clustered into one column.	Separated each data variable into a column based on the white space between them.

Third Stage: Data Sorting, Data Analysis and Visualization

Data Sorting

A master data set name Twitter_archive_master.csv was generated based on merging the three data sets and the data cleaning that was done on them. Also, all original data sets were saved in CSV format.

Data Analysis and Visualization

A report containing all insights and visualizations created from the data set was prepared. Insights were focused on interactions with WeRateDogs twitter account. their tweeting frequency and their dogs vs no dogs posts.