Wrangle Report - WeRateDogs Twitter Page

Gabriel Medeiros das Neves
May 2020

This report aims to briefly describe the wrangling efforts to prepare the necessary data for the WeRateDogs Project analyses, as a part of Udacity's Data Analyst Nanodegree.

1. Gathering

Gathering the data from this project involved collecting three different datasets from three different sources:

1. **WeRateDogs Twitter Enhanced Archive**, provided by Udacity in a *.csv* file and downloaded manually, it contains information about the page tweets, such as the tweet ID, text and URL.
2. **Tweets Image Predictions**, programmatically downloaded from Udacity servers using Python *requests* library, it consists in a *.tsv* file with dog breeds predictions calculated by a neural network that was created by Udacity.
3. **Additional Tweets Data,** *.json* file collected by querying the Twitter API with Python *Tweepy* library in order to get the tweets number of favorites and retweets. It requires a Twitter Developer Account to be downloaded.

2. Assessing

All of three datasets created with the gathered data were assessed both visually and programmatically (e.g., *Pandas* library functions, *Matplotlib* library histogram plot) in a Jupyter Notebook, aiming to recognize possible quality or tidiness issues.

Quality issues are related to the content of the data, such as missing, invalid, incorrect or inconsistent values. Usually data that has one or more of these problems is called dirty data.

Tidiness data issues, on other hand, refer to structural data problems. Essentially, an untidy data set is one that does not follow the succeeding statements:

1. Every variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

All observed issues were documented in subsections related to one of these two subjects (quality and tidiness) in the Assess section of the study.

3. Coding and Testing

I decided to organize all the cleaning code and tests together considering that would be easier to read and understand the study this way. However, each solution has its issue highlighted so that, whenever the reader wants to understand what that specific code is doing, not only he has the code commentaries, but also the issue documentation.

This final section of the wrangling process focuses on creating and cleaning up two datasets: **tweets_df** and **dogs_df**, one containing all the information related to the tweets and the other to the dogs that appear in the tweets.

After all cleaning is complete, **tweets_df**, **dogs_df** and a **combined version of these two datasets** are stored as *.csv* files (tweets_clean.csv, dogs_clean.csv and twitter_archive_master.csv, respectively) in the data folder. This way, the next time anyone need to use this data, it will already be available in its clean version in a friendly format that can be easily shared with others and opened in a spreadsheets software.