WRANGLING REPORT

Author: Thais Ruiz

January 2019

The data wrangling process consists basically of three steps: Gathering, Assessing and Cleaning.

Either they are well defined, it is often necessary to iterate through them or intertwist them, or also steps might be combined in one operation.

DATA GATHERING

In the gathering phase, the data came from different sources on varied file formats.

The 'WeRateDogs' Twitter archive (in csv format) was provided by Twitter to Udacity, and available on hand; therefore gathering manually this data and then reformatted in a DataFrame using Pandas built-in function pd.read_csv, made it a breeze.

But usually, in the real-world, data analysts have to retrieve data from external sources and in different formats, which require doing it programmatically, as the case for the following two datasets.

The tweet image predictions file in tsv format, hosted on Udacity's servers, was downloaded programmatically using the Requests library and the provided URL. With few lines of code, this task was still easy, and again the function pd.read_csv was used (with the 'tab' separator in the parameters).

The challenging data gathering came when retrieving additional data from Twitter's API using the Tweepy library. To be able to retrieve this data, this API required authentication. Additionally, some tweets on this data might have been deleted, so it was needed handling this exception when querying the additional data from the API through the tweet ids on the archive. Another encountered limitation downloading this data, was that this API has a rate limit, then it was necessary to set the wait_on_rate_limit and

wait_on_rate_limit_notify parameters to True in the tweepy.api class. Not impossible, but definitely the data gathering from Twitter's API was not a trivial task.

DATA ASSESSING

As in the real-world, data rarely comes clean.

Therefore, the major assessing task was focused in the 'WeRateDogs' archive file.

Some assessing was done visually and other programmatically, for quality and tidiness issues.

Per the visual limitation in Jupyter notebooks, some visual assessment was done opening the csv file in another editor (PyCharm and Excel) and by applying some filters in the columns was easy to identify some issues. Once more familiarized with the data, the same filters were done programmatically in the notebook to be shown as issues to tackle. Excel as an editor of the archive file, misread the tweet ids data!

For the quality issues I found missing data, inaccurate names and rating, miscategorized dogs, (re)tweets with no images, … Commonly, when you find an issue, it leads to another, and this changes the strategy to assess the data.

For the tidiness issues I found that columns from the images and JSON data could be included in the tweet archive file.

DATA CLEANING

When tackling an issue, it is usually necessary to prioritize the tasks. For example, tidiness cleaning (like dropping extraneous data: retweets with no images) should be done before cleaning quality issues (like inaccurate names). Otherwise, it would result working twice and drawing back in the cleaning process.

But experience will come with practice!

The order should be addressing the missing data, then the tidiness issues and finally the quality issues cleaning.

In the cleaning process, new detected issues made me return to the assessment stage of the wrangling work.

A different problem I encountered when coding, that make me realize the importance of checking the version of the installed packages in the workspace, was that some functions threw errors even though they should work per their documentation.

Overall, this project has been a hard but worthy journey in becoming a Data Analyst!