WRANGLING REPORT

Author: Thais Ruiz

December 2018

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) file was provided by Twitter to Udacity, and was available on hand; therefore gathering this data manually 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, even from external sources in different formats, and therefore, it requires doing it programmatically.

That was the case for the following two datasets.

The tweet image predictions file in tsv format, is hosted on Udacity's servers and was downloaded programmatically using the Requests library and the provided URL. Still with few lines of code, this task was still easy, and again the function pd.read_csv was used (this time specifying 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. Other aspect to consider when downloading this data was that some tweets in the archive file might have been deleted, so it was necessary to handle this exception when querying the additional data from the API using the tweet ids on the archive. An encountered limitation downloading this data, was that this API has a rate limit, so 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.

It is necessary, if not mandatory, to develop search skills for documentation about the data, and how to tackle the encountered issues along the process. For this project, many useful links were provided, but data analysts have to become masters looking up for the needed information. Sometimes, it might become overwhelming and time-consuming.

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, miscategorization of the dogs, some of the tweets were retweets with no images, … Sometimes when you find an issue, it leads to another, and it 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

Sometimes when tackling an issue, it is necessary to prioritize the tasks. For example, tidiness cleaning (like dropping extraneous data like 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 cleaning of quality issues.

In the cleaning process, also new issues were detected which make me to return to the assessment stage of the wrangling work.

One issue 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.

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