wrangle_report

January 7, 2020

1 WeRateDogs

1.0.1 Data Wrangling by Ovidiu Anicai

As aprt of the Udacity Nanodegree in Data Analysis I built up a Data Wrangling project in order to prove and exercise the new data processing skills I aquired. Data Wrangling is a very important part when we analyze various datasets, during my career I found that this part takes a lot of time and it proves to be crucial in building reports.

The project starts with the Gathering data step, I begun with the datasets provided by Udacity. One dataset is the WeRateDogs twitter archive that provides more than 2500 tweets with dog photos rated in a funny way by the account owner. The second dataset is also provided by Udacity and includes a series of predictions made with multiple algorithms over the dog photos of the twitter account. The third piece of data I had to fetch directly from Twitter API and it includes up to dat likes and retweets counts of the previously collected tweets.

I started the process as recomended in the course with the **Gathering Data** step, I documented the datasets and downloaded each one using a specific method, then saved all results locally. While downloading tweets details from Twitter's API I encountered few errors related to deleted tweets, but most of the process went very well.

The next step in the Data Wrangling process was the **Data Assesment**, this is done having in mind that the data might have issues with the content of the data (Quality) or to the structure (Tidiness). The first approach was by **visually** getting acquainted with the datasets columns and visible values. Just by looking at the tables I started to see some issues like missing values, inaccurate data or even corrupded data. Also the structure of the datasets surely needed work in order to be ready for analysis. The next approach was using **programatic**, here functions like **.info()** or **.describe()** are very usefull in spotting wrong data types, oddly named columns, missing values and statistics about the numeric values. Further more I started investigating some columns that I wanted to see how many distinct values have, in this **.value_counts()** proved to be very usefull and I started spotting columns with few distinct values or with lots of 'None' values. This step also gave me some insight into the rating system details, here some of the results hinted me that the rating extraction from text was not properly done or there are some tweets with totally different formats. Continuing the assesment I checked for duplicated values or particular cases.

Before starting listing and addressing each issue, I built a small Data Dictionary in order to better understand what each column represents and includes. Then I started listing all **Quality** and **Tidiness** issues and hinted few possible approaches. During the cleaning process I strikedout some of the issues because they became redundant or would not help if implemented.

Now the most important step of the Wrangling process **Data Cleaning**. I started by making a copy of all already imported datasets to be able to get back anytime to the initial values, because during the cleaning lots of things break and I had to rerun from this step. I used the recommended steps **Define** -> **Code** -> **Test** methodolody and this made me understand better the importance of this approach and the clearness it provides to the reader. I started with mosto if the Titiness issues because they affect more the structure of the dataframes and continued one by one with each identified issue. From time to time I reordered the issues so the cleaning will make more sense and gain advantages of the previously steps. I struggled a lot with some new notions learnt in the course videos, especially with the **.melt()** function, but at the end I discovered a better and simpler way to solve my issues with **.wide_to_long()** During the Test step I used assertions in order to break the process if the data is not as expected.

At the end of the cleainig process I reinspected all columns and felt ok with the results and saved all resulted dataframes to different files on the disk.

Then I started the **Analisys** section and tried to answer few questions about the cleaned datasets. Having in mind the two extra documents I have to write, I drawn few charts that can help answer my questions. In the last part I made sure all Charts are properly built and the resulted images are saved in the media folder where I can include them in the two reports. I wrote the two reports and needed to document more about the WeRateDogs account and the traction it created.

I had a great experience in building this Notebooks and made me understand better the importance of Data Wrangling and the mindset of having to present your results to other people