Wrangle Report - Project 4 Udacity

by Bianca Friz

This project revolved around data wrangling of the twitter account “WeRateDogs”. This account tweets all types of dogs and rates them. However, it has a greater purpose, which is, beyond telling who they are, it also visualizes the problems of these dogs and gives them a platform so that more people can help them. This project requires to show abilities for data analysis, a process that includes 4 steps: Gather, Assess, Clean and Explore. Data wrangling is one important part to get data analysis of quality and make good decisions based on a good profiled dataframe. However, it is a process that requires time and attention to notice problems.

For this work, the data collected from the “WeRateDogs” account was not easy, mostly because of the different ways of gathering the data in first place. Moreover, the fact that it required 3 different dataframes: one from the twitter API, the other one from the image-prediction.csv file and finally, the twitter-archive-enhanced.csv.

In this project I used Jupyter Notebook because I found it more friendly and secured than the workspace in Udacity. One of the first difficulties faced was gathering the data from the twitter API, because twitter did not reply to my API access request. Nevertheless, I could use the code that Udacity gave so that I could progress in the project. But, besides that, the rest of the open tasks for the other archives did not presented bigger issues thanks to the methods given by pandas.

On the other side, for the Assessment part, it was divided in two steps: The visual assessment and the programmatical assessment. The first one was made via excel which could show all the rows and columns without the cut display that Jupyter notebook has for bigger dataframes. Meanwhile, the programmatical assessment were made in the Jupyter notebook with methods like .shape(), .info(), .describe(), .duplicated(), sample(), .head(), .value_counts() and .nunique(). Those methods helped to a faster visualization of the quality and tidiness problems needed to find.

After that step, the cleaning step of the dataframes was the part that took the longest time, because it required a lot of coming back to clean problems that were not noticed in first place in the assessment step. Due to that, it was a constant search for improvement. In the cleaning stage, it was required the format “Define-Code-Test”, which helped to be clearer and more orderly in the explanation of the process. Before starting to clean the dataframes, I made a copy for each one in case it was required to re-run the dataframe if there was an error made in the cleaning step.

For the last part, after cleaning and merching the dataframes, it was easier to find insights and visualization with one single dataframe with all the contents of the three dataframes, which made a more efficient way of the exploration process. Finally, thanks to the libraries such as matplotlib, the charts and plots were easier and more intuitive to make. Also, I found really helpful to make subsets of the data that I wanted to show before visualizing in on charts, so that it only focused on the main columns and for that, iloc, loc, pivot_table and sort_values were the ones that I used more. But one of the things that I rescue the most out of this project is that I learned how to display images out of a url, which I found really interesting.