Wrangle Report

By: Guilherme Landim Frota Leitão

Date: July, 2018

The Data Wrangling project was difficult and challenging, due to that I've learned a lot because I had to search for many techniques to solve the issues I'd encountered.

Data was gathered from three sources for this data analysis.

- csv file WeRateDogs gave exclusive access for Udacity to use on their class. Containing some basic information.

This file I read with pandas library

- tsv file of neural network predictions of the dog's breed in the pics were downloaded from the Udacity servers.

To download this file I had to use requests library, but while trying new things I dicouvered that pandas.read_csv function could read the file while on the servers.

- json file we created by data mining with the TWEEPY API to get more information about the tweets from the first file.

Using tweepy to get the additional data for the tweets was my biggest challenge since I had to learn how to use tweepy from my own looking for information online in sites like stackoverflow, videos on Youtube and in tweepy documentation. I don't know how many dificulties I had and how many sites and videos I've search to solve them.

After I gather the data from all sources I initiate a cycle of Asses the data (8 quality problems and 2 tidiness problems needed) and clean it, several times while I was cleaning one of the issues I've found another one to be clean.

Clean the issues was not so hard, some of them were wrong columns data types, wrong information in rating columns, repeated and unused columns,duplicated information, retweets and replies in the tweets dataframe and columns that should be unified in just one column.

To conclude this project was a really important learning experience since I've to learn about several techniques to solve the issues encountered, and by this, the project was really successful.