WeRateDogs

Data Wrangling

Andreina Tirado, 2019

The overall project consisted in the data collection (3 different sources) a csv file, an url endpoint and twitter api. Getting each of those three sources had different levels of difficulty and determining how to complete the data collection was one of the first challenges. In this case, by reading the documentation (eg Tweepy) and using request library it was easy to get the data to start the analysis. During this step one of the most important remarks are:

- Delete credentials and secrets otherwise there will be a vulnerability in the code that could increase the likelihood of a security incident.
- Use tweepy library and its documentation to collect the information required per tweet taking into consideration rate limits
- Use the request library to download the images-prediction files.

At this point the data wrangling process can start, there are 2 methods:

- Visually

  Better to use with small data sets, otherwise it can be a bit cumbersome almost impossible to identify data errors only by looking at the data directly. Some errors that can be spotted are missing values, multiple data points in one columns, incorrect or misleading column names.

- Programatically

  The preferred method, especially with larger data sets depending on how complex are the different data sources it can turn more complex than necessary, nonetheless there are a couple of factors that must be checked to verify data quality: duplicated data, incorrect data types, missing data and unnecessary columns.

This particular step was one of the most challenging and complex up until this point during the course, mostly because it required a lot of detail and comprehensive documentation. There were at least 2 specific cases where incorrect information was parse to the tweet file (original):

1. Incorrect numerators: for this case using a regular expression to capture the entire number proved to be the best approach
2. Multiple stages assigned to one dog: in this case the approach was to combine the value of all columns with stages to later analyze them in detail.

In addition to the mentioned above, one formula that can be use to decrease the error rate during this step, as suggested in the course, is:

- Document finding
- Define how to fix and code it
- Test the changes

This 3 steps process proves value and helps to achieve the final goal, which is a clean data set to work with. It's also important to make sure that the data is tidy in that way it will be easy to conduct the analysis to verify among others, why people love WeRateDogs, what is the most popular dog (based on retweets)? And the most popular dog breed.