Wrangle Report – Felipe Gibert

In the process to clean the “We_rate_dogs” data I followed three steps:

- Data Gathering

- Data Assessing

- Data Cleaning.

1.- Data Gathering

The data for our project was split into three different sources:

- The first one was in a csv called “twitter-archive-enhanced.csv” that was provided by Udacity to perform this task. In this CSV we can find information about: “tweet_id”,” reply status”, “tweet timestamp”, source of tweets, retweet information, tweet URLs, dogs rating and dogs stage. In this case the information has some problems with quality and tidiness.

This information was obtained by reading the file with pandas package in Python.

- The second data frame was obtained by using request library in python, and getting the information from a webpage, then saving it in a TSV file that later will be read by python pandas package.

- Finally, the third data frame was obtained by calling the twitter API and download the information using the tweet ID provided in the first data frame. To use the twitter API we need to create a developer account and configure or credentials.

2.- Data Assessing

In the different data frames, we found many issues that are detailed in the next list. I defined the dataframes like: First data = Dogs, Second Data = Images and third Data = Twitter.

Quality Issues

- Dogs

- Name "a" in dogs dataframe
- So much "None" in the field name
- There is one "unacceptable" as name
- There is one "this" as name
- There is one "actually" as name
- There are some ranking numerator that are over 14. Even when it is not a "serious" ranking it would be nice to have an accurate maximum value related to where most of the data are.
- There are some denominator that are not 10
- The source is with a link format. IT should only show the source
- Timestamp is and object and should be a date

- Images

- There are some breeds with and underscore, it would look better with and space
- There are some breeds that start with mayus and others do not. Would be good to standardize this.

- Twitter

  - "ID " column is written with and space.
  - 'Favourite_Count ' column is written with and space.

3.- Data Cleaning

After finding out some issues in the data, we proceed to clean it. We perform the following task to clean the data:

- Dogs data frame – Quality analysis

- Fixing “name” field
- Fixing ranking numerator
- Fixing ranking denominator
- Fixing “source” field
- Fixing “timestamp” field

- Dogs data frame – Tidiness analysis

- Removing unnecessary columns
- Creating a raking column
- Creating a stage column

- Image data frame – Quality analysis

- Formatting the breed columns

- Image data frame – Tidiness analysis

- Removing unnecessary columns

- Twitter data frame – Quality analysis

- Changing columns names

Finally, we merge all our clean dataset into one master dataset called “twitter_archive_master” that later we are going to use to analyze or data and get some insights.