We Rate Dogs Data Wrangling Report

London, 28th of May of 2019
Author: Vivian Sacon

Content:

1. Overview

   This report has the objective of describing my work performing data wrangling with data from the twitter webpage WeRateDogs, using the data provided for this project.

   The project's objective is evaluate the data and treat it using data wrangling concepts and tools learned during the data wrangling module of the Udacity Data Science Fundamentals course.

   As I have learned during this module, the data wrangling consists in 3 steps: gather, asses and clean the data, I have followed the three steps in that order and documented every step in the jupyter notebook document called wrangle_act.jpynb.

2. Gather

The work started with the creation of the three dataframes that should be used during the wrangling. Each dataframe was derivate from a different source, so I could practice different ways of gathering data.

The first dataframe called WeRateDogs was created from a .csv file provided by Udacity; this file contains the majority of the information related to the tweets of the page as tweet_id, text, datetime, dog stage, etc.

The second dataframe called ImagePredictions, was created using a file available in the Udacity servers, for that one I needed to download the file programmatically and then turn it into a dataframe. This dataframe contains predictions about the dogs breeds created with artificial intelligence based in the pictures from the tweets.

The third dataframe was the most complicated one, the project specifications requested information about count of favorites and retweets and this information needed to be extracted using a twitter api.

This three steps concluded the gather phase of the project.

3. Assess

To asses the data I have used some code assess tools as info(), head(), count_values(), however, the visual analysis was very important resource to detect the number of problems requested to meet the requirements.

At the end I could identify 8 quality issues and 2 tidiness issues, as listed below.

Quality:

1. All three files have different row count.

WeRateDogs:
2. Timestamp should be datetype datetime.
3. The project specification says that just original tweets must be taking into account, therefore, the lines with values <> '' in the field retweeted_status_id must be removed.
4. The project specification also says that tweets without images shouldn't be taking into account, therefore expanded_urls without value must be removed.
5. Remove the 48 tweets not related with dogs by using the "We only rate dogs" filter.
6. Replace the 49 values 'a'in the column Name to 'None', to keep the pattern when we dont have the dog name.
7. There are 22 rows with denominator different from 10, this is another pattern problem that should be fixed to better evaluate the data later.

ImagePredictions:
8. There are predictions that are not related to dog breeds and therafore should be removed from the final analysis.

Tidiness:

WeRateDogs:
1. One variable spread in 4 columns (variable dog_stage: doggo, floofer, pupper, puppo).

ImagePredictions:
2. Dog prediction should be converted in one variable and therafore be presented in one column.

4. Clean

   The cleaning part consisted in use code to clean all the issues that I have found in the asses phase of the project.
   For this phase I kept the original dataframes and created new ones with the cleaned data.

   After cleaning the necessary issues individually in each dataframe, I have created a final dataframe called wrangle_act.jpynb joining all the tweets that have information in the 3 dataframes, for this final dataframe I have kept just the important columns, dropping some unnecessary columns.

5. Analyze

   After the three data wrangling steps, I have took some time to analyze the data produced and create some insights and visualizations, this step will be used to create the act_report requested in the project specifications.