wrangle_report

August 14, 2019

1 Wrangle Report We Rate dogs

1.1 Introduction

In this project we gathered, analyzed and visualized the data of the Twitter @rate_dogs known as WeRateDogs. We will document the steps for the wrangling part. The wrangling process has 3 parts : Gather data, Assess Data and Clean Data.

1.2 Gather Data

The data is being gathered from 3 differents sources : The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. This twitter archive is being given to us and downloaded manually from the Udacity website.
The Image Predictions File a table full of image predictions of breeds of dogs (the top three only) alongside each tweet ID, image URL and the image number that correspond to the most confident prediction.
This file has been downloaded programmatically using The Requests library.
The Twitter API has been used to gathered additional data. We queried the Twitter API using the Tweepy Access library, for that we registered an account on the Twitter Developer Website, the account allow us to have access tokens to use with tweepy. For security reasons these access tokens are not shown in the jupyter notebook. After querying all the tweets which ID were present in the archive we saved it locally and loaded it with keeping just the *Tweet_id*, *favorite_count* and *retweet_count* columns.

1.3 Assess Data

We assessed the data visually and programmatically.
- visually
We check for quality and tidiness issues visually detectable as invalid names of dogs, unconsistent columns and so on.
- programmatically
Using attributes like pd.info(), pd.describe(), we can check another quality and tidiness issues as wrong type of the date column and so on.
Here are the quality issues found :

- The data type of the tweet_id is not the same in all three files
- The data type of favorite_count and retweet_count in df_tweets is object
- The number of rows in image_predictions and twitter_archive are differents

- p1,p2,p3 variables in image_predictions do not start all with Capitalize letter
- The value of the rating_denominator is not always 10
- There are inconsistent values of the rating_numerator(0, 1776)
- We just need the column with the high probability to have a dog so p1_conf the other 2 are useless
- The timestamp and retweeted_status_timestamp are string data type but actually must be datetime
- Not valid names for many dogs ("a", "the", "an")
- We only want original ratings that have images.
- Remove useless columns from the final dataframe
- Rename columns of the final file for better understanding

and the tidiness issues :

- The doggo, flooter, pupper and puppo columns must be in one column.
- All 3 dataframes must be merge toghether in one dataframe

- Remove duplicates values from the final dataframe

There are others quality and tidiness issues but we had not enough time to consider them all, these are the most evidents.

1.4 Cleaning Data

The cleaning data process follow the same 3 phases for each issue :

- Define
  Here we explain with words the process to resolve the issue

- Code
  Here we programmatically resolve the issue using python code

- Test
  We make a test or assertion to verify if the issue is risolved

1.5 Store the Data

After cleaning all the data and merge all the 3 files together, we store it in a master file that we will use for the analysis. The analysis and the insights found are in the `act_report.pdf`.