WeRateDogs Twitter - Data Wrangling project

Alina Grigorescu
Bucharest, 2018

Introduction…

In the following project I gathered data from a variety of sources and in a variety of formats from the WeRateDogs Twitter account, assess its quality and tidiness, then clean it. Thus, the main steps are presented below:
- Gathering Data
- Assessing Data
- Cleaning Data
- Export Data to CSV file
- Insights presentation

About the data…

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. It was started in 2015 by college student Matt Nelson, and has received international media coverage both for its popularity and for the attention drawn to social media copyright law when it was suspended by Twitter.

About the data gathering….

- The WeRateDogs Twitter archive was imported and read as a given csv file
- A file with predictions regarding the dog status was read from an URL address
- A file with tweets was imported using the tweeter API (application programming interface)

About the data assessment…

2 approches to assessment:

- Visual assessment: each piece of gathered data was displayed in the Jupyter Notebook for visual assessment purposes.
- Programmatic assessment: functions and/or methods were used to assess the data.

About the Quality & Tidiness issues…

Visual assessment was carried out by opening csv files in Microsoft Excel.
Quality issues (Completeness, Validity, Accuracy, Consistency):
- Data types quality issues:
-- "tweet_id" column is integer type, but should be string type as no operations are performed on it.
-- "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", "retweeted_status_user_id" columns should be strings as well.
-- "timestamp", "retweeted_status_timestamp" columns should be date type
-- "rating_numerator", "rating_denominator" should be floating type

- Other quality issues:
-- Some "tweet_id" in the archive file (df_basic) do not appear in the prediction file (df_predict) (there are 2356 IDs in the first file, but only 2075 in the second)
-- The "text" column in df_basic always contains the URL at the end
-- The "names" in df_basic column sometimes contains invalid names (such as "a","o","all","just" etc)
-- There are missing values in the "expanded_urls" column of df_basic
-- Retweets should be excluded from the tweets
-- Some entries have missing values for all of the "doggo", "floofer", "pupper", "puppo" columns
-- Invalid ratings for where 9/11 and 7/11 is mentioned in the "text" column
-- Columns "p1","p2","p3" in the df_predict should be all capitalised

Tidiness issues:
- Columns "doggo", "floofer", "pupper", "puppo" should be only one column (named "class") containing one of the 4 values
- Column "id" in the API file should be renamed "tweet_id"
- All 3 data sources should be merged into one table

About the data cleaning…

The data quality and tidiness issues mentioned above were treated one by one, using various python libraries. Each issue was Defined, then the Code was produced and then it was Tested. Some of the issues were treated through functions defined to automate the task, some others were case by case manually solved. Cleaned data was stored in separate dataframes from the original messy ones. Also the sources were merged together into one master datasource.

Conclusions…

After cleaning the data, some basic analysis was perfomed, asking questions such as what are the most popular dog breeds, names, are there any correlations between the favorite counts and the retweet counts.