WERATEDOGS Twitter Archive: Wrangle Report

Introduction

The main purpose of this report is to summarize the effort deployed to wrangle the data of **the WeRateDogs Twitter**, which is a Twitter account that rates people's dogs with a humorous comment about the dog.

Data wrangling refers to the process of gathering, assessing and cleaning the raw data available into a more usable format, so as to create interesting and trustworthy analyses and visualizations.

1. Gathering data

The first step of wrangling data is gathering it. For this project, I have gathered 3 sources of data:

- **The twitter archive file,** that I downloaded it manually from Udacity servers.
- **The tweet image predictions**, I downloaded it programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
- **Twitter API & JSON:** I downloaded it from the Udacity servers and read this text file line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count.

I loaded the 3 raw data files into separate tables: ***twitter_archive, image_predictions, tweets_df.***

2. Assessing and cleaning data

After gathering the data, the second step is to **assess data** of the three datasets, visually and programmatically. The main objective of this section is to look for uncleaned data in all the three DataFrames, in particular for quality issues (completeness, validity, accuracy and consistency) and tidiness. The later is where each variable forms a column, each observation forms a row and each type of observational unit forms a table.

Visually, I used two ways, the first one by printing the three dataframes in jupyter and the second wat by viewing it in excel. Programmatically, by using different functions, such as, info, value_counts, describe, duplicated, etc…

The third step is the **cleaning** task. It doesn’t mean that I have changed the data to make it say something different, I have just cleaned data when inaccurate, removed when irrelevant and replaced when missing.

First of all, I have created a copy of the three original dataframes, which is a very practical and helpful method.

Concerning the quality issues, as an example[1], I have dropped all rows containing non null values in the retweets and replies columns, because we are only interested by original tweets, and after that dropped these columns. There were also some erroneous datatypes, especially for timestamp, tweet_id, rating numerator and denominator, I converted them to the right datatype. In parallel there were many challenging tasks, such as, cleaning some incorrect data for numerators and denominator by extracting them from the text. Moreover, I had to deal with some missing data (urls and names) and mislabelled data.

Concerning the tidiness issues[2], I have first joined the twitter_archive and image_predictions, then the new one and tweets_df. Also, I have melted the 4 dog stage columns to the dog_stage column and the tweets without stages were set to 'NaN'….

3. Storing data

At the end of the wrangling data process, I have stored the cleaned data in a csv file "twitter_archive_master".

[1] These are just some examples. You will find all the quality issues in the jupyter notebook' wrangle_act '

[2] These are just some examples. You will find all the tidiness issues in the jupyter notebook' wrangle_act '