Internal document
Author: Stanislava Stachova

Documentation for data wrangling steps: gather, assess, and clean, covering brief description of wrangling efforts.

The document *act_wrangling* is divided into following parts:

- Part I - Gathering data
- Part II - Assessing data
- Part III - Cleaning data
- Part IV - Analyse and Visualize data

Within Part I, I have gathered three dataset files in a following way:

- I obtained WeRateDogs Twitter archive. As it was downloaded to pre-defined folder Downloads, I used shutil to move it programmaticaly to target folder.
- The tweet image predictions file (image_predictions.tsv) was hosted on Udacity's servers and was downloaded and saved programmatically.
- I was not able to obtain access to developer account, therefore I downloaded file tweet_json.txt manually and read it

All datasets were read to python and displayed through .head() to show first rows of dataset.

Firstly, I have assessed all three datasets visually. I used mainly jupyter notebook for this analysis. I looked to each table separately to get acquainted with datasets, understand what was the table about. I checked columns, column names, rows and scammed observations in order to identify possible quality and tidiness issues seen at first sight.

Secondly, I used built in functions to assess tables programmatically. I used primarily following functions:

- .**info()** to check values in columns, rows that I planned to analyze. It gave me info regarding data types, null and non-null rows, number of rows, columns and appropriately visible list of column names;
- **.head(), .tail(), .sample()** to get overview of observations at the beginning of table, at the end of table and randomly selected sample from whole dataset;
- **.value_counts()** to get acquainted with the count per each group of observation in selected columns;

- filter data using **.loc/ . query()** to look closely on selected observations and its possible later analysis;
- .**duplicated()** to understand if table had or had not duplicates which would be required to clean.

When I identified quality or tidiness issue, I made a note under the code cell. At the end of Part 2 Assessing data, I gathered issues identified within assessing process and summarized them in the table for further cleaning process.

The Part III deals with cleaning process. This part is divided to Qualitative issue cleaning and Tidiness issue cleaning. Issues were cleaned in line with standard cleaning process documentation and so Define, Code and Test. In this part I worked with the copies of read datasets.

I decided to clean following issues:

Qualitative issues:

- Change data type (object to datetime)
- Removal of retweets
- Drop selected columns not used for the purposes of analysis later on
- Cleaning of denominator rating and nominator rating
- Remove duplicates
- Change column names (id to tweet_id to merge datasets, p1 to prediction_1 to give more descriptive information)

Tidiness issues:

- New column gathering dog stage created
- Merge of all three datasets

At the end of cleaning process, I merged three cleaned datasets to one master dataset called twitter_archive_master and save it to ‘csv’ file called twitter_archive_master.csv.