Udacity Data Analyst Nanodegree Project 7: Wrangle Report

Oliver Kröning

October 24, 2018

1 Introduction

The purpose of this report is to describe the work, which is done within the framework of Udacity's project *Data Wrangling*. The content which I learned during the lessons is put into practice. Therefor, the tweet archive of Twitter user *WeRateDogs* (@dog_rates) as well as other sources are used to gather, assess and clean dog rating data for further analyses and visualization.

2 Data Wrangling

2.1 Gathering Data

In this section, the data required for the WeRateDogs dataset is gathered. Therefor, the Enhanced Twitter Archive is read and converted into a Panda DataFrame. Additionally, the image prediction file hosted on Udacity's server is downloaded programmatically.
At least, the tweet IDs in the WeRateDogs Twitter archive are used to query the Twitter API for each tweet's JSON data, Afterwards, the read data is saved into the tweet_json.txt file and read into a pandas DataFrame.

2.2 Assessing Data

After we obtained the datasets, we assess the data visually and programmatically.
The visual assessment is performed by using Jupyter Notebook and Microsoft Excel.
Within the programmatical approach, we used Jupyter Notebook with Python Pandas. This allows us to perform filtering and analyzing functions on the data.
We identified several data quality and tidiness issues, we have to clean in the next step:

- Data Quality (content of data):
  - The dataset contains retweets, which do not have to be considered.
  - The dataset contains tweets without any images, which do not have to be considered.

  - Some dog names are not correct or have missing values.
  - The source column contains HTML code and is hard to read.
  - Rating denominators differ from 10.
  - Rating numerators containing floats are displayed incorrectly.
  - Special characters, like '&' are not well encoded and displayed.
  - The columns' datatypes are not appropriate.

- Data Tidiness (structure of data):
  - The dog stage is given in four columns: doggo, floofer, pupper, puppo.
  - To save all datasets into one CSV-file, a dataframe combining all datasets has to be created.

2.3 Cleaning Data

Within the cleaning phase, we tackle the issues, we have identified in the assessment section. For each issue three different tasks have to be performed: defining, coding and testing.
To retain the original dataframes, we create copies of the datasets and perform the cleaning operations on them.
At first, we define the cleaning tasks, which are used to solve the issues we assessed. In the following phase, the actual coding of the cleaning tasks takes place. Comments within the source code supports the definition of the task as well as its readibility and comprehensibility.
Finally, the code is tested to check for errors or incorrect data.

3 Conclusion

In this report, we described the overall process of data wrangling within Udacity's Data Analyst Nanodegree project. The coding of the process performed by using Python and some required libraries as well as Jupyter Notebook.
In the gathering step, we collected the data given by Udacity. Additionally, we used the Twitter API to obtain more data, like the numbers of retweets and likes.
The data assessment was also supported by the usage of the Jupyter Notebook, since we could describe the problems within the dataset in markdown cells next to code cells. This enables a good readibility and transparency of the code.
The assessed issues are then tackled in the cleaning phase of the process. Python, escpecially with Pandas, also enables the manipulation of dataframes very easily.
Finally, our dataset contains 2116 observations and 24 columns.