Wrangle Report

Udacity Data Analyst Nanodegree, Term 2
Created by: Matthias Rieke

Introduction:
This project is part of the nanodgree "Data Analyst". This project is about wrangling, assessing, cleaning and analyzing data associated with the twitter user WeRateDogs (@dog_rates).
The twitter user WeRateDogs rates pictures of dogs in a very entertaining way. What's special about the rating is, that WeRateDogs ratings almost always have a numerator greater 10, while the denominator is 10. The argument for the numerator being greater than 10 and the denominator being 10 is, that all of the dogs are good dogs, therefore deserving at least a ten, but in most cases deserve better.

The project is separated in different steps:

1. Gathering Data
The data we need for this project is gathered in different ways:

1. Tweet Image predictions are programmatically downloaded from a source provided by Udacity. Machine learning was used to predict the dog type based on the picture.
2. An enhanced twitter archive was manually downloaded and then being accesses from the local storage. This file contains different variables like tweet id, text, ratings, names, timestamps etc.
3. Additional data is downloaded via the twitter api using tweepy. Data which was gathered this way are retweet counts, favorite counts and more.

The cleaned data has then been saved into a new csv.

2. Assessing Data
After gathering all data from the various sources, data of the different datasets is assessed visually and programmatically. The following methods were used:

- .head()
- .info()
- .value_counts()
- .duplicated()
- .count()
- .str.contains
- .loc
- .str.islower
- .sample()
- .text[]
- .sum()

The following have been discovered:

Tidiness:

- Data separated into different dataframes

- Variables separated in different columns, e.g. dog stage. Same goes for columns which contain predictions

Quality:
- Data contained retweets and replies
- Incorrect datatyps, e.g. for retweets, timestamp
- name, doggo, floofer, pupper and puppo columns contained „None“ instead of NaN
- names were sometimes missing or inaccurate extraced
- rating numerators contained decimals
- rating were not standardized
- columns which are not needed were present in our datasets
- there were denominators other than 10

3. Cleaning data

After assessing the data visually and programmatically, it has been cleaned. The following techniques have been used to address the issues which have been assessed:
- .merge()
- .drop()
- .isnull()
- .str.slice()
- Pd.to_datetime()
- .loc[]
- .drop_duplicates()
- .apply()
- .append()
- Astype()
- .value_counts()
- .info()
- .head()
- .str.extract()
- Regular expressions
- Replacing values by using ==

Summary:

Before I was able to analyze the data, there was quite some work to do.
Taking a look at the number of assessment and cleaning operations which have been taken, we can see that in order to have clean data, you may need to put quite some work into gathering, investigation (assessment) and cleaining. This may differ depending on which data you are working on, but it shows that you have to be critical about the data if you gather them initially.