Wrangle Report

1) Data Gathering:

Goal: Gather data from 3 different sources
Action: Data were gathered from a server hosted by Udacity, a csv file provided by Udacity and via the Twitter API. The files retrieved from the Udacity server and via the Twitter API were saved to a .tsv file and a .json file respectively. 3 pandas dataframes were created with the data from the 3 different sources.
Insights: Gathering the data from the .csv file and the Udacity server via the Requests library was straightforward. Gathering the data via Twitter API with the help of tweepy was tricky insofar as there are rate limits for requests. Although this issue was adressed in the Udacity documentation the notification that „twitter api rate limit reached sleeping for 733“ caused confusion up to the point when it was clear that despite that message and an observed stand still all the data were gathered in the end and saved to the appropriate file.

2) Data Assessment:

Goal: Assessing data in order to find the quality and tidiness issues relevant for the questions asked to get insights into the data.
Action: Define the questions I want to answer in order to gain some insights into the data. Perform a visual and a programmatic assessment to define the quality and the tidiness issues. For the programmatic assessment for these data frames and my questions the most helpfull functions were .info() and .value_counts().
Insights: In the beginning I asked a few wrong questions because I did not spend enough time on getting to know the topic and the meaning of variables and the processes in WeRateDogs. Visual and programmatic assessment both are valuable for finding quality issues. With programmatic assessment you can dive deeper into data and uncover deeper hidden problems. Assessment is not a single phase in the process but work in progress all the time and also in later phases like cleaning and even during analysis new issues might be discovered.

3) Data Cleaning:

Goal: Address the quality and tidiness issues programmatically and merge the 3 dataframes into one as the master template to perform the analysis on.
Action: Start with making a copy of the dataframes and perform cleaning actions on the copies of the dataframes. Most of the cleaning effort was merging variables splitted into several columns into one column, cleaning text and extracting text from messy data. This was overall straightforward. I strictly followed the define, clean(=program), and test subdivision.
Insights: The define, clean, test order is very helpfull in structuring the cleaning effort. Text extraction can be challenging especially for a newbe to regex. This part for sure takes the most time but it is great to have a cleaned and merged master dataframe in the end.

4) Data Storage:

Goal: Store the cleaned and merged dataframe as a master data frame for future usage.
Action: Write the master data frame to a .csv file.
Insights: Once you have stored the cleaned and merged file you can start right away by reading in this dataframe without the necessity to go through all the cells with the data cleaning steps.