DATA ANALYST – PROJECT: DATA WRANGLING

1. GATHERING DATA

The first step was to gather all data of the three sources. Getting the archive data of a manually downloaded file via pd.read_csv. This data contains information about several dog tweets of WeRateDog. The second task was to programmatically download a tsv-file with the prediction data by using the requests library and get the response of the given url in the project details. The third task was to gather directly from Twitter by using the Twitter API Tweepy. Unfortunately, I had some problems to creating a twitter account and used the provided json file. But in meantime I created an account and used the given code with my access token etc.. I probably didn't import the json file in the best way, I didin't notice that there is and read_json method in pandas.

2. ASSESSING DATA

I took two assessing steps. For each dataframe I displayed the first rows and used the info() and describe method() to get a better understanding of the data. In the first place I looked for obvious mistakes like missing values or the datatype of each column. In the dataframe df_archive I found the most issues. The main Issues are mentioned in the jupyter notebook. After merging the three dataframe I reiterate to Asses Step und looked for other issues which I didn't see in the first asses step. For example after merging all dataframes there were some rows or tweets without images. (The task was to analyze only tweets with images)

3. CLEANING DATA

Before the cleaning part I copied all dataframes. In the first cleaning step I mostly cleaned quality issues like more descriptive column names, changing datatypes of some columns (e.g. timestamp from string to datetime, tweet_id from integer to string) or dropping rows or columns of the dataframe which weren't necessary for the analysis like all retweet entries. After that I cleaned tidiness issues like combining all doggolingo columns to one column because this only one variable. After merging all dataframes I reassess the data and looked for other issues.

In the second cleaning part I cleaned some quality issues like dropping all rows/tweets without images and tweets without favorite count (This tweets aren't available anymore). The last step was to make a second dataframe with the tweet text, url, jpg_url cause these columns aren't necessary for the analysis (and drop these in the main dataframe).