DATA WRANGLING REPORT

@WeRateDogs
Twitter

DATA GATHER & ASSESS

Objective to Question

- Rating distribution
- Dog breed with highest rating
- Dog breed with highest retweet rate
- Dog stage with higest favorite rate

Data

The basis for the analysis consists of three different types of data:

- WeRateDogs Twitter accound data
- Twitter interactiondata from tweepy, Twitter API
- Image prediction data from Udacity server

Data Gathering

WeRateDogs twitter account data was gathered from a (.csv)-file by pd.read_csv(). Twitter interaction data was gathered by the tweepy API (extended mode), creation of a local (.txt)-file with the content of interest (tweet_id.values). For image predictions, a request with the url was sent to Udacity servers, the content was written to a local (.tsv)-file and read by pd.read_csv. All data were read into a Pandas data frame for further use.

Data Assessment

The data assessment aims to identify tidiness (each variable forms a column, observation forms a row, each type of observational unit forms a table) and quality (completeness, validity, accuracy, consistency) issues in the data and consists of two parts: Visual (Numbers) and programmatic assessment (Jupiter Notebook). Of all findings, 8 quality and 2 tidiness issues have been addressed.

DATA WRANGLING

Assesment Findings

Twitter Account Data

Quality

- Validity issue, we are only interested in original tweet data
- Validity issue, the rating columns are not correct extracted from the text
- Validity issue, the dog stages are not correct extracted from the text
- Accuracy issue, change A to None in name column
- Consistency issue with name starting with lower letter in name column; such,quite (looks like overwriting of data)
- Consistency issue with with datatype in id-columns: in_reply_to_status,in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id
- Consistency issue with datatype for timestamp

Tidyness

- The variable dog stage can be found in different columns

Twitter Interaction Data

Tidyness

- Tidyness issue, all tweet is not data in the same table

Image Prediction Data

Quality

- Consistency issue in p1, p2 and p3 columns, first letter
- Consistency issue in p1, p2 and p3 columns, replace ‘_’ with ‘ ‘
- Consistency issue in p1_conf, p2_conf and p3_conf columns, len - round()

Tidyness

- The table also contains data from the same observational unit (the tweet) and should therefore be in the same table as the account_data

DATA CLEANING

Data Cleaning

Before cleaning, data copies was created by df.copy(). The cleaning started by addressing missing data and tidiness issues. Thereafter, addressing quality issues in the order of validity, accuracy and consistency.

First, I did not work with the missing data, I don't have any information of how it could be retrieved, treated or recovered. Later, I had to address mismatch of data after table merges, keep it in mind for future purposes.

Every step of cleaning was done following the process define, code and test.

The data wrangling source code can be found on GitHub through the link below.

Karin Wiberg

DATA SCIENTIST

KarinWiberg