Wrangel Report DAND Project 4

In this report all data wrangling steps (gather, assess, clean) of the project 4 will be explained.

Gather

In this project we used three data sources. A manually downloaded file (*archive-enhanced.csv*) from the online classroom, which contains the basic tweet data. A programmatically downloaded file (*image_predictions.tsv*) using requests library, which contains predictions on image labels for most of the tweets. And a file (*archive-enhanced.csv*) which contains tweet details which were downloaded by using twitter api (tweepy) and store it in tweet_json.txt.

Asses

To assess the previously downloaded data DataFrames from the Pandas library were used.

For the image_redictions.tsv the data was read in by using the read_csv method with the tab as separator.

The built-in Pandas functions came in handy also for getting first general overview of the data, **visual assessment**. So head() and sample() gave a quick general insight into the structure and concept of the data. Info(), describe() and duplicated() helped to identify missing values, wrong data formats, detect outliers and duplicated entries as well as finding tidiness issues, **programmatic assessment**.

Clean

In cleaning two main block were approached:

1. Cleaning issues (renaming columns, aligning columns, handling not existent values, handling extreme values, transform types)

2. tidiness issues (according to Hadley Wickham's definition of tidy[1])

1. Detected and approached cleaning issues by:

Cleaning the structure:

- rename column **id** to **tweet_id**
- make all labels same format lower case

Get rid of bad values:

- remove records with rating_denominator unequal 10 and remove rating_denominator column
- remove rows that are retweeted or replaying and remove empty cols afterwards

[1] Hadley Wickham's definition of tidy in (https://vita.had.co.nz/papers/tidy-data.pdf):

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

- replace all not set names and to short names with a default name.

Extract information:

- extract @recipient from text into column recipient
- extract weekday from created at and store as category

Cast wrong types:

- transform created at to datetime type
- retweet status timestamp to date time type
- timestamp to date time type
- convert in_reply_to_status_id/ in_reply_to_user_id/retweeted_status_id/retweeted_user_id to int
- convert source to categorical type (iPhone, web, vine, TwitterDeck)

2. Detected and approached tidiness issue were detected and approached:

- p1 p2 p3 is one variable (melt to a new table with new vars p, breed, conf, isdog)
- doggo/floofer/pupper/puppo melt to one var of dtype="category"
- extract retweeted /replay columns to its own table

Please read the attached act_report.pdf for documentation of analysis and insights into final data