Report: wrangle_report

Introduction to the project.

For this project, I wrangled data that's associated with tweets of Twitter user "WeRateDogs". The dataset is based on three individual sets being provided by different sources:

a. A twitter archive dataset containing tweet data directly from the user. This dataset is being provided as a download.

b. An image prediction dataset containing information about the dog breeds. This dataset is being hosted on a server and should be downloaded programmatically.

c. A dataset containing retweet and favorite counts of tweets. This dataset needs to be created based on a request to the Twitter API.

The goal is to wrangle all datasets to create interesting and trustworthy analyses and visualizations.

Gathering data

First, I start with importing all needed modules for this project.

The first, twitter archive enhanced, available as a CSV, can be downloaded and uploaded into the notebook.

The second, image prediction dataset, required me to request for access oa a given url leading to the file on the server.

The third, queried twitter API dataset, I used the downloaded version because of some difficulties.

Assessing data

I visually and programmatically access all three datasets one by one. I am looking at the content and structure given in each table and column and then have a closer look via programatically checking for duplicates, counting values, checking datatypes and the distribution of the data.

There are many things that can be fixed and cleaned, but I focus on only eight quality and two tidiness issues.

Cleaning data

Before cleaning the data, I made copies of all three datasets and continued working with the copies. Most of the quality issues in the datasets are based on missing values, wrong content, or wrong formatting. All issues were fixed with the help of various codes to replace, query, drop etc. For the tidiness issues, I melted various columns of the stages and merged two tables together.

Storing data

I merged all the tables into a master table and dropped the columns that are not needed for our analysis.

Analyzing and visualizing data

For the analyzing, I check the newly created master dataset for findings about the dog breeds and which has the highest likes. I also looked at the likes and retweets distributions.

For the visualization, I plotted the breed with the most highest likes in the tweets . Secondly, I ploted the likes and retweets distribution , which showed to be a positive.