Wrangle Report

A REPORT COVERING ALL THE STEPS DONE IN THE WRANGLING PROCESS

DATA WRANGLING

The data wrangling process was performed following three steps:

1- Gathering the Data
2- Assessing the Data
3- Cleaning the Data

Data Gathering

In this project the data was gathered from different sources and by different methods.

The We Rate Dogs archive “Twitter_archive_enhanced.csv” was downloaded manually.

The second file “the tweet image predictions (image predictions.csv) was downloaded programmatically using the Requests library.

Finally, a JSON file called tweet_json.txt was created by gathering the data from WeRate Dogs twitter page using twitter API by Pyhton’s Tweepy library.

Data Assessing

Assessing is the second step in data wrangling. Assessing data is searching for unclean data. The searching procedure can be done in two methods: visual, and programmatic. While assessing the data the following issues were raised.

Quality Issues

From the archive table:

- Wrong column datatypes: tweet_id, retweeted_status_id, and retweeted_status_user_id are integers and floats rather than strings. Timestamp, and retweeted_status_timestamp are strings rather that time format.
- In accurate dog names: some dogs are named ‘a’ and ‘such’ instead of their correct names.
- In accurate rating_numerators some numerators were extremely high, others were decimals.
- In accurate rating_denominators the denominators should all be ‘10’ while some of them were multiples of 10 that was due to the presence of more than one dog in the picture.

From the image_pred table:

- Wrong column data type: tweet_id should be a string instead of an integer.

From the retweet table:

- 179 retweets are not usefull and thus will not be used.

Tidiness Issues

The columns doggo, floofer, pupper, and puppo should be merged into 1 column instead of four.

All the data frames can merged into one master data frame.

Cleaning Data

Cleaning data will be done in three steps: define, code, and test.

First the steps to clean the data should be defined. Second, a code Is written to convert those definitions for example the melt and merge commands were used to fix the tidiness issues.

After gathering, assessing, and cleaning the data the final data frame is saved as a csv.