Wrangle report

Introduction

There are 3 main data sources: tweets with data surrounding the tweets, twitter data regarding the tweets such as dog names, and predictions of the images of dogs. Beneath are the cleaning actions which are taken to enhance and merge the data.

Tidiness

Recall that tidiness is that each variable is a column, each observation is a row, and each observational unit is a table.

- This means that we should combine these 3 data sources. Each data source has as observation base a tweet, or more specific a tweet_id. These data is merged on tweet_id. Tweet_id is set as index since it is the unique identifier.
- Some columns are in multiple tables. Therefore, redundant columns are deleted.
- Information regarding the twitter account is deleted. Since this is the same for all observations.

Quality

As quality issues the beneath mentioned problems are cleaned:

- The source column contains html code and this is changed to only the text part.
- Columns with only Nans/null values are omitted since they do not provide extra information.
- Datetime columns are changed to date columns
- The columns with the 'dog-kind' i.e. floofer or puppo are merged to one column containing that information.
- There is information with dog predictions, the respective probability and a Boolean column if the predicted object is a dog. There is a new column added where the column contains the predicted object with the highest probability if the predicted object is a dog race. If the top 3 predictions don't contain dog races, this is null.
- Delete all retweets and tweets of other twitter accounts.
- Set the values of columns as retweet count, favorite count, display text start, display text end, and img num to int values.
- Empty lists are set to nan values.
- Set entities hashtags to the hashtags for information regarding most used hashtags.
- Replace all empty lists with a nan value for clarity.
- A score is added which is the denominator divided by the numerator