Report

On

Data Wrangling

The Waterdogs Twitter project goals included:

1) Wrangling the twitter data through the following processes:

   - Gathering Data
   - Assessing Data
   - Cleaning Data

2) Storing, analysing and visualizing your wrangled data
3) Reporting on the data wrangling efforts and data analyse and visualization

Gathering Data

The wrangling methods used data from different sources that includes

twitter_archive_enhanced.csv

tweet image predictions

Twitter API that uses Tweepy library to gather] tweet

Assessing Data

Quality Issue:

twitter-archive-enhanced-2.csv

- Completeness:
  missing data in the following columns:
  in_reply_to_status_id,
  in_reply_to_user_id,
  retweeted_status_id,
  retweeted_status_user_id,
  retweeted_status_timestamp,
  expanded_urls.

tweet_id is an int (applies to all tables)

- Validity:

dog names: some dogs have 'None' as a name, or 'a', or 'an.'
This data-set includes retweets, which means there is duplicated data

Accuracy:
retweeted_status_timestamp inconstance data type
ime-stamp (wrong data type)
Consistency:
The Source column still has the HTML tags
rating_denominator should be a standard 10, but there are a multitude of other values

image_predictions.tsv
Validity:
p1, p2 and p3 columns have invalid data
Consistency:
the dog breed listed is all lowercase
some columns there is an underscore for multi-word dog breeds.
tweet_json
Completeness:
Missing Some Data

Tidiness Issue:
twitter-archive-enhanced-2.csv
The last four columns all relate to the same variable (dogoo, floofer, pupper, puppo).

image_predictions.tsv
This data set is part of the same observational unit as the data in the 'twitter-archive-enhanced-2.csv' - one table with all basic information about the dog ratings.

tweet_json
This data set is also part of the same observational unit - one table with all basic information about the dog ratings.

Cleaning Data

After the assessment, cleaning the data through the following means:

Define, Code and Test:

Merge the clean versions of archive, images, and twitter_counts_df data frames
Correct the dog types
Remove columns no longer needed.
Change tweet_id from an integer to a string.

Change the timestamp to correct datetime format.
Correct naming issues and Standardize dog ratings.