Project: Wrangle and Analyze Data – Wrangle Report

For this project data needs to be wrangled and presented in a report suitable for a blog post or magazine.

Wrangling the data is described in this report. It is performed in three steps:

- Gathering Data
- Assessing Data
- Cleaning Data

Gathering Data

For this project data from three different sources needs to be gathered. The sources and the method are summarized in the table below.

Source	Description
twitter-archive-enhanced.csv	This file is already provided locally and can be loaded directly into a dataframe.
image-predictions.tsv	This file is provided at a cloud space. We need to download it first and then generate a dataframe.
tweet_json.txt	We need to use the twitter API to download the tweets from twitter and create a json file out of it.

Assessing Data

Having a close look at the gathered data reveals some quality and tidiness issues:

Quality issues:

- Lots of NaN
- Some unneded Data
- Timestamp is not a data element
- Bad Formating of data, i.e. source is an url
- Lots of diplicates
- Multiple urls in dataset
- some errors in dog names
- dog breeds with odd spelling and "_"

Tidiness issues:

- Dog stats are in multiple colums
- 3 prediction of dog breeds

Cleaning Data

Before any cleaning is done all the gathered data is merged in one big dataset and a copy is made of this dataset. All further work is done on that copy.

As the assessing revealed some issues, we need to clean them up in the next step.

The following cleaning steps will be performed:

- Copy of the data is made
- Remove empty rows
- Remove unneeded columns
- Fix timestamp
- Cleanup source
- fix expanded urls
- Calculate fraction
- fix dog names
- Combine Dog stages
- Combine Dog prediction
- Fix dog breed writing

After each step we will have a look at the table head to test the results.

The final merged and cleaned dataset is stored in ‘twitter_archive_master.csv’.