Gathering, Assessing and Cleaning
some of the @WeRateDogs tweets

This document details my efforts to gather, assess and clean data regarding the weratedogs twitter channel. The present paper is part of the Udacity's Data Analyst Nanodegree. In the following lines I will briefly explain my actions regarding each part of the process.

Gathering Data

There are three sources of data used in this project: a .csv WeRateDogs Twitter archive, a .tsv tweet image predictions file and the actual twitter account weratedogs tweets. Below I will describe the steps undertaken to "gather" the data. In this instance, gathering means importing data into pandas DataFrames

WeRateDogs Twitter archive

To read the file into a DataFrame the function "pd.read_csv" was used.

Tweet image predictions file

This file is located on an external resource, given this the "requests" library was used to "get" the file, write it on the local disc and reading it into a DataFrame. Since this is a tsv file the '\t' separator was used as a separator.

Actual twitter account weratedogs tweets

In importing this data into DataFrame, first the tweets were read from twitter.com and saved as json objects into a local file, then, this json file was read and a dictionary was created based on these json objects, afterwards the dictionary data was loaded into a DataFrame.

Assessing Data

Data was assessed using two approaches: Visual Inspection and Programmatic one.

Visual Inspection

In order to visualize data *head(), sample(),* and *info()* functions were used.

Programmatic Inspection

In this type of approach, the data types and the format of the data was tested. In checking the data type *type()* function was used. In checking the format the RegEx expressions were used.

Cleaning Data

Cleaning the data was done in two stages: cleaning quality issues and cleaning tidiness issues. In cleaning data the define -> code -> test approach was used.

Quality issues

In order to clean the data, replacements, changing of data type and deletion of rows were performed. Briefly, the changes that were done are:

1) The text from the column “source” was extracted from under HTML tags
2) The timestamps were changed to the timedate format
3) Rows with the inappropriate dog names were deleted
4) The retweets were also deleted
5) The rows with missing expanded urls were also purged
6) The predicted breeds of dogs were lowercased
7) The “-“ separator was changed into “_” separator
8) The predicted probability type was changed from string to float
9) The rows that contained the column titles were deleted

Tidiness issues

Two tidiness issued were addressed:

1) The “stage” of the dogs where in separate columns. This issue was cleaned by melting all four columns into one
2) The tables contain information that are pertinent to a single observation, in this case the tweet id, therefore all tables were linked together using this ID and the *join()* function.

Summary

As you have seen, I gathered data from multiple sources, I assessed the data and cleaned the data according to my assessment.