Data Wrangling Report

By Oladayo Isaac Oladipupo
September 2022

In this report, I outline the wrangling efforts to assemble and clean the data required for analysis of the WeRateDogs Twitter Archive.

The work was divided into
- Data Gathering
- Data Assessment
- Data Cleaning
- Data Storing
- Analysis and Visualization

Data Gathering:

Gathering Data for this Project involved obtaining three different datasets from three different sources. Each one tests a different way of obtaining a dataset.
The data was gathered from three different resources in three different formats.

a. **"twitter_archive_enhanced.csv" file.** This file was downloaded manually and then imported into our working environment using the Pandas library.
b. **"image_prediction.tsv" file.** This file was downloaded programmatically using the Requests library from a provided URL. This file has image prediction results for the dogs' breeds obtained through a neural network on most of the tweets in the archive file.
c. **"api_json.txt" file.** This file was gathered from twitter's API via the Tweepy library by querying the API to obtain extra information pertinent to the tweets' ids in the archive file, e.g. retweets count and favorite count.

Data Assessment:

In this step, the imported files were investigated both visually and Programmatically on the Jupiter notebook using code to examine the data provided. We inspected our datasets to make sure it is eligible for our next analysis stage. The datasets were assessed under two criteria, quality and tidiness. When an issue was detected it was documented under one of these two criteria.

Quality refers to issues related to the content of the data, sometimes called dirty data. The standard criteria of completeness, validity, accuracy, and consistency of the data were used to

identify quality issues. These issues were varied and are listed in the assessment section of the “wrangle_act.ipynb” Jupiter notebook.

Tidiness refers to issues related to the structure of the data, sometimes called messy data. The basis for assessment is that each variable forms a column, each observation forms a row and each type of observational unit forms a table. After assessing, an assessment summary was created to list all discovered issues for the next stage.

Data Cleaning:

The final step in the wrangling process is cleaning the data for quality and tidiness issues. The cleaning followed the standard process of defining, coding, and testing for each of the issues, and they were tackled in a logical order, which is reflected in the numbering order in the “wrangle_act.ipynb”. Most of the cleaning was performed programmatically, such as defining functions, developing regular expressions to capture the right records, or using the panda's built-in functions (merge, melt, extract, etc.).

Data Storing:

As requested the final Twitter archive was saved in the csv format under the name twitter_archive_master.csv

Analysis and Visualization:

The insight from the dataset were analyzed and visualized in the wrangle_act.ipynb file as requested.