DATA WRANGLING REPORT

Paula Jasper, November 12, 2018

This document has been written to support the third project of Term 2 of the Udacity Data Analysis Nanodegree course and marks the end of the Data Wrangling element of the Nanodegree.

The aim of this project was to wrangle WeRateDogs Twitter data to create interesting analyses and visualizations using python. The project supports the three stages of the data wrangling process:

1. Data Gathering.
2. Data Assessing.
3. Data Cleaning.

The work culminates with storing, analysing and visualising the cleansed data.

All work was completed using Python within Jupyter Notebook.

STEP 1: DATA GATHERING

Three pieces of data were gathered using three different methods.

1. The WeRateDogs archive (twitter-archive-enhanced.csv) contains basic tweet data for over 5000 tweets. This was downloaded manually from the link provided using the pandas read_csv() function.

2. The tweet image predictions file (image_predictions.tsv) contains predictions of dog breed in each tweet from a neural network, developed as part of Udacity's Machine Learning course. This was downloaded programmatically using python's requests library.

3. Finally, (the most challenging part of the data gathering stage) additional data from the tweets was downloaded using the Twitter API. In order to do this, I requested a developer account from Twitter and keys and tokens were provided for me to access it. I stored the keys and tokens in environment variables on my computer to keep them hidden. I then queried the Twitter API for each tweet's JSON data using python's Tweepy library and wrote it to a text file. This process took about 30 minutes. I included printouts in my code, and a timer in order to follow the progress. Finally, I read the file into a pandas DataFrame storing tweet ID, retweet count and favorite count.

STEP 2: DATA ASSESSING

The three dataframes were then assessed according to Quality (completeness, validity, accuracy and consistency) and Tidiness (one variable per column, one observation per row and one unit per table).

One dataframe at a time, I looked at samples, structure and tested for duplicates. Then I examined one column at a time in more detail and made a list of the issues I found. The full list is contained in wrangle_act.ipynb.

STEP 3: DATA CLEANING

Firstly, I made copies of the three dataframes, then cleaned many of the issues I documented in the assessment step. I did not clean all of them due to time constraints. Also, the requirement was to clean at least 8 quality and 2 tidiness issues. However, I identified and cleaned more than this.

I cleaned the data using the 'define, code and test' process and I documented this fully in wrangle_act.ipynb. I also recorded "DONE" plus a brief description against the issues to assist me.

Many of the issues, such as incorrect datatypes, dropping incomplete records, renaming columns and changing lowercase to uppercase were simple to correct. However, others such as reshaping the archive dataframe to one dog_stage column instead of 4 columns was difficult and took longer.

When I dealt with merging individual pieces of data according to the rules of tidy data, I originally planned to merge only the archive and api dataframes as the data in both is associated with the tweets. I planned to keep image_predictions as a separate dataframe as this is additional enriched data. However, ultimately, I decided to merge them all into one dataframe to facilitate the visualisations stage.

DATA STORING, ANALYSIS & VISUALISATION

I wrote the cleansed and merged dataframe to twitter_archive_master.csv.

The analysis and visualization outputs are described in act_report.pdf.