NAND Project #4 Data Wrangling Report

Introduction

This document reports on the Data Wrangling phase of NAND Project #4. The goal of this project is to gather, assess and clean specified data associated with the WeRateDogs Twitter account. This Twitter account tweets humorous comments about photos and videos submitted that typically contain one or more dogs. These tweets typically contain the following:

- The name(s) of the dog(s)
- One or more cute, descriptive terms for the dog
- A rating on a 0 to 10 scale but this rating often exceeds 10 (e.g., 12/10)
- A link to submitted photo(s) or video

An archive of tweets from this account from 11-15-2015 through 8-01-2017 was provided for this project in the form of a .csv file (twitter-archive-enhance.csv). In addition to the text of the original tweets this file also contains a number of enhancements to facilitate analysis. These enhancements include the following:

- Tweet ID the uniquely identifies the tweet
- Timestamp identifying when the tweet was sent
- Source
- “in reply to” tweet status and user IDs
- Retweeted status and user IDs and timestamp
- Expanded URLs for submitted media
- Rating numerator extracted from the tweet text
- Rating denominator extracted from the tweet text
- Dogs name extracted from the tweet text
- Cute, descriptive terms “doggo”, “floofer”, “pupper” and/or “puppo” extracted from the tweet text and put into one of four corresponding columns

In addition to the twitter-archive-enhance.csv file data from two other sources was required:

- Photos from these archived tweets were process by a neural network to predict dog breed or other animal or other objects found and the output is hosted on an internal server
- The Twitter API was used to collect retweet counts and favorite counts for each tweet in the archive

These data sources were used to gather the required data for the analysis and this code used to perform this step is available in the Jupyter Notebook, wrangle_act.ipynb. While these sources provide a wealth of data on the WeRateDogs account these data are not without issues. The

details of my assessment are documented in the wrangle_act.ipynb Jupyter Notebook but, in brief

- A number of data quality and a couple of tidiness issues were identified in the twitter-archive-enhanced data.
- Some of the tweets IDs found in the twitter archive were not found using the Twitter API
- Some of the photos in the image prediction data appear more than once and there was one significant data tidiness issue identified

The details of how these issues were cleaned can be found in the Jupyter Notebook, but the data was cleaned and data from the twitter archive was merged with the extra variables from the Twitter API and this merged data was stored in twitter_archive_master.csv. The image prediction data was retained as a separate table and stored in image_predictions_master.csv.

The results from the analysis conducted on the cleaned dataset is reported in a separate document, act_report.pdf.