Wrangle Report

This report describes the wrangling process of WeRateDogs Data Analysis Project of Udacity's Nanodegree Program. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment and a numerical rating for the dog at the end of each tweet. The goal of this project is to analyze the Twitter account and get insights of their contents.

This wrangling process consists of:

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

Part I – Gathering Data

The project analysis involves obtaining data from three different datasets, each obtained from various sources.

The first source for the project is a Twitter tweet archive provided by Udacity as a CSV file. This data is to provide the basic information about WeRateDogs' twitter content from November 2015 to August 2017, including each post's ID, timestamp, text, name of dog mentioned in the tweet, ratings for each dog, etc. This provides the base and is the most robust dataset out of three.

Secondly, we obtain data about what breed of dog is present in each tweet by a tweet image prediction website. To obtain this file, we need to send a request to the website informing our interest in using and uploading the data to our work frame in Jupyter Notebook. As WeRateDogs post tweets about comments and ratings for dogs accompanied by a picture, we are using this system to translate the pictures of each tweet into a categorical data which is the breed of the dog.

The third data set is obtained by querying the Twitter's API and using a Python library named Tweepy to obtain further data of the account's tweets in the CSV file. This includes gathering each tweet's retweet count and favorite count. We can achieve this by using the tweet IDs from the Twitter archive and inquiring for each tweet's JSON data and store the entire set of the data in a txt file. Then, we read this file into a Pandas DataFrame.

After all the data is gathered from these three data sources, we continue to the next step of data cleaning and visualization.

Part II – Assessing Data

We assess the gathered data visually and programmatically for both quality and tidiness issues. The visual assessment and programmatic assessment are done inside Jupyter with Pandas. The two issues are documented in the wrangle_act.ipynb Jupyter Notebook for each dataset.

The quality issues are related to the content of the data. The issues found in all three datasets are not all of the data are in the most appropriate data type. Some specific issues include some invalid inputs in timeframe, ratings, and names in the first dataset, and unequal number of rows for the three DataFrames.

The tidiness issues are related to the structure of the data. Some issues found were some columns that can be combined into one. The three datasets are then merged into a single data frame with selected data containing only relevant variables in each column.

We also checked and found no duplicated data in all three datasets.

Part III – Cleaning Data

While assessing the data, we also involve the process of cleaning the data for each of the quality and tidiness issues documented. This includes merging DataFrames, changing data types, removing and adding columns, removing tweets with incomplete data, etc. which are described in detail in the ipynb file. This is done using tools such as def functions and Pandas built-in functions with some manual data cleaning. The final file generated is a clean master Pandas DataFrame.