Data Wrangling - Project Two

Ashinze Emmanuel Chidi

Udacity Nanodegree Program

WRANGLE REPORT

Project Overview¶

The project is divided into three sections:

- Gathering Data
- Assessing Data
- Cleaning Data

Gathering Data

The first step to data wrangling is gathering data which could be really tasking. I gathered data in the following steps:

- I started by importing the required libraries; NumPy, Pandas, seaborn, matplotlib.
- I manually downloaded the files provided by Udacity which are the twitter-archive-enhanced.csv, image-predictions.tsv and tweet-json.txt files.
- The twitter-archive-enhanced.csv and image-predictions.tsv were read using the pd.read_csv() functions and the tweet-json.txt was Looped through to generate a more structured data set.
- The twitter API would have been an excellent choice, but the application was not approved.

Assessing and Cleaning Data

After gathering data, it was time to show my detective skills. Using the pandas profiling function and Pandas functions, I was able to find the following issues and solve the issue with the solutions on the table below:

Quality Issues	Solutions
Too many columns. Columns that are not needed will be dropped.	Drop Columns in the twitter archive dataset.
Inconsistent Datatypes across various Dataset. (tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp, retweeted_status_id, source, retweeted_status_user_id, retweeted_status_timestamp, doggo, floofer, pupper, and puppo)	Datatype conversion to the proper datatype
Errors in the rating_numerator and rating_denominator. There are some entries here with decimals that needs to be fixed.	Using Regex to find the affected entries and replacing them.

Missing values in the Dataset and incorrect representation of missing values in the name and URL column	Dropping the missing values
Errors in Data entries in the dog names column	Replacing the incorrect names with None
Improper Formatting of the Source Column	Using Regexes to extract and fix the improper Formatting
In the Text Column, some of the contexts are not ratings in twitter archive data	Using Regexes to find the affected entries and rectifying it.
Missing images in the image predictions dataset	Dropping the missing images
Tidiness Issue	**Solution**
Multiple Dog Stages- doggo, floofer, pupper and puppo columns in twitter_archive table should be merged into one column named "dog_stage"	Merge the Columns into one and drop the columns not needed
Many datasets trying to show the same thing (Archive data, Image data and Api data)	Merge the Datasets

After the iterative process, the dataset was stored as twitter_archive_master.csv for analysis.