Reporting: wrangle_report

This current notebook is organised into the following sections:

Introduction

> **This notebook takes us through the steps, methodology and processes deployed during this project, from the gathering stage to the assessment of the gathered datasets and various cleaning methods applied.**

Dependencies

> **In order to perform any wrangling process, dependencies and libraries are vital to the overal process. The various commands used to access these depndencies used for this project are within the cell below.**

In [1]:

```
#importing necessary libraries and dependencies

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import requests
import os
from PIL import Image
from io import BytesIO
import json
```

Gathering

The first stage of wrangling which is as important as every other stage involved. Without gathering data, we can't necessarily assess nor perform any form data cleaning procedure. Brief details about the sourced files used for this wrangling exercise will be looked at in this section.

Data used for this project was sourced from three different files types namely:

1. `twitter-archive-enhanced.csv` - Contains tweets from WeRateDogs twitter handle and the respective properties of each tweet such as ratings, dog name, time stamp of tweets, dog stages and source. This file was sourced by manual download.

---

1. `image-predictions.tsv` - This file is a neural network prediction of images found in tweets made by the WeRateDogs twitter handle. This file was sourced using programtic download method by using the `requests.get()` method.

---

1. `tweet-json.txt` - This is a TXT file with JSON contents extracted from Twitter(via API) and provided by Udacity. Contains tweet properties just like the first file on the list but having more details. This files helped me generate the favorite counts and retweet counts. Two needed columns from this file were extractd using the `.readlines()` , `pd.DataFrame()` and `.append()` methods.

Re-gathering

> - **Before heading to the 'Assessing' section of this notebook, it is important to note that some of the dataset had a few complications which was mentioned in the motivation section of this project. This informed the decision to re-gather some sections of the data in the `weratedogs_archive` dataframe such as the `name` , `dog_stage` , `rating_numerator` and `rating_denominator` columns.**
>   **The Re-gathering process was performed using the `.str.extract` method in combination with regular expressions on the `text` column of the `weratedogs_archive` dataframe.**

Assessing

> **After gathering necessary files, data assessment is the key next step in every wrangling project. For this project, assessment was done using both visual and programmatic methods.**

Visual Assessment

> **Visual assessment was performed by firstly reading each dataset into a pandas dataframe using the `pd.DataFrame()` method for the data extracted from the `tweet-json.txt` file, `pd.read_csv` for data contained in the `twitter-archive-enhanced.csv` as well as the `image-predictions.tsv` with an inclusive separator argument of `\t` . Furthermore, each dataframe was simply called within the Jupyter notebook for a first level visual perusal inorder to identify necessary data tidiness and quality issues that were possible using this method.**

Programmatic Assessment

> **Programmatic assessment drills down further into our dataset to help identify tidiness and quality issues that might not have been seen during the visual assessment process. Some useful methods and commands used during programmatic assessments are: `.info()` , `.duplicated()` , `.describe()` , `.unique()` and `.sum()`**

Cleaning

> **Several Data cleaning procedures were applied to the dataframes gathered and assessed in the previous sections for specific purposes**

1. `.merge()` : combining all three dataframes from generated from the gathering processes needed to be done for data tidiness reasons. There was no need to have three different dataframes.

---

1. `.drop()` method: This method was used on several occassions during the cleaning session of the project. Excluding rows with unwanted values, columns uneeded for my analysis and other duplicate rows that came as a result of the `.melt()` method used during the cleaning stage.

---

1. `.melt()` method: used to collapse(unpivot) the different dog stages found in the `weratedogs_archive` dataframe for proper data tidiness.

---

1. `.fillna()` method: was useful for filling some missing values with the average values in the dataset.

Summary/Conclusion

> **Data wrangling is an ittirative process and this project isn't exempt from that statement. Several processes were revisited during the wrangling acts for this project. For example, gathering process was revisited during the assessment section and data cleaning switched between tackling tidiness and quality issues on several occassions. All sections in this notebook covers a surmmarised breakdown of all warangling stages for this project. The more untidy a dataset is, the more ittirative wrangling processes can be.**