Wrangling Report
7th October 2022

This Data Analytics project was intended to test the student's ability to gather, assess and clean data in order to extract meaningful findings.

This project was derived from the @WeRateDogs Twitter profile, a popular internet sensation that rates images of dogs submitted by the public while briefly describing the dog post in a humorous way. To test our gathering skills, we were required to obtain the data from 3 sources;

1. The first dataset was graciously given to us and it contained archived tweets from the @WeRateDogs Twitter account. This data was provided in csv format and once downloaded, was imported into python Jupyter notebook using Pandas read_csv tool. This dataset contained 2356 rows and 17 columns.

2. The second dataset which included images of the dogs had to be downloaded programmatically using the Pandas requests tool from a given URL. This data was then imported into Jupyter notebook using the read_csv function of Pandas. This data set contained 2075 rows and 11 columns.

3. The third dataset included key data which was missing from dataset 1 and was crucial for the analysis. Particularly important were the retweet count and favorites count information. Gathering this data proved most challenging as I was required to obtain this from a Twitter API via a Twitter Developer account. After Twitter approved the account, I received api credentials unique to me which I was to use to obtain this data. After obtaining the required information, I converted the list of tweets into a dataframe using the pd.DataFrame tool, to align with the other datasets already gathered.

Due to the large dataset, manually/visually assessing it proved challenging. I resorted to programmatically assessing the data using the below functions:

1. .info() - This provides summary information of the dataset including number of null values, datatypes, number of columns and rows as well as column heading names.

2. .describe() - This provides statistical information on the dataset including row count of each column, mean value for each numerical related column, standard deviation, minimum and maximum values for each numerical related column.

3. .duplicated() - This checks if any row has been duplicated in the dataset.

I noted all issues observed during the 'assessing' stage and this was going to be resolved during the 'cleaning' stage. The cleaning stage is where unstructured and untidy data is transformed into structured and tidy data. In the cleaning stage, columns which were not meaningful to my end goal were dropped. I also dropped outliers which would skew the findings.

For meaningful analysis to be conducted, I merged the 3 datasets into one file using the .merge() function and I saved this into a datafame.