WRANGLE REPORT

1.0 INTRODUCTION

This is a data wrangling project from Udacity which is part of the requirement's for completing the Data Analyst Nanodegree Program. Data wrangling involves gathering data from different data sources (web pages, APIs, databases etc.), assessing it for quality (content) and tidiness (structural) issues, then cleaning it.

The dataset that I will be wrangling is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. This report briefly describes how I wrangled the dataset for this project.

2.0 PROJECT GOALS

The goals of this project include:

- Data Wrangling
- Analyzing and Visualizing Data

Data Wrangling

There are three steps in the data wrangling process (gather, assess, clean). The activities carried out in each of these steps are described below:

- **Data Gathering**: the data for this project involved three datasets, they were gathered as follows:
    - **The WeRateDogs Twitter archive ('twitter-archive-enhanced.csv' file):** this dataset was provided by Udacity and was downloaded manually via the provided download link.
    - **The tweet image predictions ('image_predictions.tsv' file):** this dataset contains the results obtained from running every image in the WeRateDogs twitter archive through a neural network for the purpose of classifying different breeds of dogs.

This file was downloaded programmatically using python's request library and the provided URL from Udacity.

- **Additional data from the Twitter API**: for this file, using the 'tweet_id' in the WeRateDogs Twitter archive dataset, I successfully queried and downloaded the 'favorite_count' and 'retweet_count' for each 'tweet_id' using Python's tweepy library from Twitter's API. Before accessing the Twitter API, I created a Twitter developer account, submitted a request to Twitter to grant me access to their API and created an app with which to access their API after permission was granted. Each tweets json data was written to a text file called 'tweet_json.txt' line by line. This text file was then read line by line with the 'tweet_id', 'favorite_count', and 'retweet_count' extracted and loaded into a pandas dataframe.

- **Assessing Data**: the two types of assessment used include visual and programmatic assessment.

  - **Visual assessment**: I visually assessed the data by printing the three datasets in the jupyter notebook and also using an external application like excel.

  - **Programmatic assessment**: the dataset was programmatically assessed using different pandas functions and/or methods such as .info(), .head(), .sample(), .tail(), .value_counts(), .unique(), .duplicated(), .isna() etc.

  Through these assessments, I was able to identify several quality and tidiness issues while being conscious of specified issues in the project instructions. Some of the identified issues include:

  - **Quality issues:**
    - Contains columns with retweet data and rows with retweet related data.
    - source column values represented using HTML link tags.
    - contains ratings that are not for dogs.
  - **Tidiness issues:**

    - Four variables (doggo, floofer, pupper, puppo) in 4 columns in the twitter archive dataset.
    - image_predictions table should be part of the archive table.

- **Cleaning Data**: this step involved cleaning all the quality and tidiness issues documented in the assessment step.

  The first thing I did was to make a copy of the three datasets so that the original datasets will remain available without being edited.

  Afterwards, using the define-code-test framework on each of the documented quality and tidiness issues, the three datasets were cleaned with each cleaning operation defined before hand and verified afterwards with tests to ensure the operation was successful.

  Some of the basic cleaning operations carried out include changing column data types, dropping unnecessary rows and columns etc.
  While some of the more advanced cleaning operations include, extracting values from one column and replacing the values in that column with the extracted values, extracting values from one column and storing it in another column, using np.select() to select values based on a condition and choice list and merging three tables into one.

Storing Data

The cleaned master dataset was stored in a CSV file called 'twitter_archive_master.csv'

3.0 CONCLUSION

Real-world data is messy and is rarely available in a ready to use format. This is why data wrangling is an important skill to have as a data analyst/scientist. In this project, I have used Python and its libraries to gather data from different sources (web, API) and in different formats (csv, tsv, txt, json), assessed it visually and programmatically using Pandas (A versatile and powerful python library for data manipulation and analysis), cleaned it programmatically and then storing the cleaned master dataset in a CSV file making it ready for analysis.