Introduction

This report provides a detailed guide/feedback on the Data Analysis wrangling process (Gather, Assess & Clean) performed on **WeRateDogs** Twitter archive.

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The Data Wrangling process used in this analysis include:

1. Gather

- At the onset, essential packages (*pandas, numpy, requests, os, io*) were imported into the Jupyter Notebook. Subsequently, **WeRateDogs** twitter archive holding basic tweet data, *twitter-archive-enhanced.csv*, was loaded into the project workspace and stored in a *pandas* dataframe called *df1*.
- Along the same line, a dataset containing image predictions for the rated dogs was downloaded programmatically using the *requests library* and stored in a *pandas* dataframe called *df2*.
- In order to ascertain additional data (favorite_count & retweet_count) on the dogs, Twitter API was queried using the Tweepy library and stored in a text file, *tweet_json.txt*. Afterwards, this text file was read into a *pandas* dataframe called *df3*.

2. Assess

- Each of the gathered data (*df1, df2 & df3*) was accessed both programmatically and visually.
- Visual Assessment: For visual assessment of each dataframe, a Microsoft Excel Workbook & Notepad were pivotal. Jupyter Notebook was not used in the visual assessment because it didn't display a plethora of the rows in each dataframe; hence, the need for a Microsoft Excel & Notepad.
- Programmatic Assessment: Here, Jupyter Notebook was very essential. Methods like *.info(), .describe(), .nunique()* were used in programmatically assessing each dataframe.
- Documentation of quality and tidiness issues was executed in tandem with the programmatic & visual assessment.

3. Clean

- On completion of data assessment, a copy of the 3 dataframes was created.
- The first cleaning operation tackled a tidiness issue, unwanted data. These unwanted data includes data that do not represent original tweets. While assessing the data, it was gathered that some of the data included were either retweets or replies. These data were removed accordingly
- Following the eradication of unwanted data, the tidiness issues were addressed. Using pd.melt, the doggo, floofer, pupper & puppo column was commingled into one column called dog_stage.
- Afterwards, all three columns were joined accordingly.
- The key next step was to address the remaining quality issues such as wrong column data type, wrong dog ratings, multiple dog_stage etc.

Conclusion

After the gamut of the data wrangling process was conducted, the clean data was stored as a csv file called *twitter_archive_master.csv.*