Wrangling Report

This report adequately captures my effort to wrangle the data from different data sets. The report is structured to display my efforts in gathering, assessing, and cleaning data.

Gathering Data

Data was gathered from three different sources.

1. Twitter archive data from the flat file provided.
2. The image prediction data set obtained from a url.
3. Twitter data from the api.
   I didn't use the api for sourcing the Twitter data as I couldn't get access. I used read_json to gather the Twitter data.

The gathering process was performed seamlessly with a few lines of code. The knowledge gotten from the Udacity wrangling course proved to be more than enough in handling this section. I used read_csv, the request library, and read_json to gather the Twitter archive, image prediction, and Twitter data dataframes , respectively.

Assessing Data

I assessed the data for all the dataframes visually and programmatically.

Visually, I assessed the Twitter archive dataframe on Microsoft Excel as I am quite comfortable with its use. Some quality issues were quite noticeable like an unstructured data source, some tweets were replies or retweets, and the dog life stages were in three columns.

Programmatically, I found the dot describe, info, shape, head and tail, duplicated methods very relevant in assessing the dataframes. Alongside these, I used value_count and query method to examine the data regularly.

During the process of assessing the data, I often had to switch between visual and programmatic assessments. This I found very helpful as I would have been unable to notice some quality issues across the dataset. Overall, I observed and documented twelve quality and two tidiness issues.

I encountered a few challenges during the assessment process. The most challenging was the breakdown of my personal computer which I was unable to fix for a couple of weeks. I also underwent dental surgery which left me incapacitated for several days. Fortunately, I got over these challenges and I am working on completing not only the wrangling but the entire program before the deadline.

Cleaning Data

I performed a number of activities to ensure my dataframes were cleaned and ready to be analyzed.

Firstly, I made copies of the original dataframes to avoid any distortions. This was followed by a thorough cleaning of all tidiness issues and about 95% of the quality issues identified in the assessment stage.

The tactics I implemented in cleaning the dataframes include;

- regex patterns
- changing data types
- splitting the entry into columns
- basic and advanced indexing
- dropping columns and
- merging dataframes

The tactics were very effective in cleaning the dataframes giving me a master dataframe that was used for my analysis.

Finally, I saved the dataframes to a csv file and analysed it resulting in the creation of eight insights and three visuals.

I am immensely grateful to Udacity and ALX for providing a great platform that allows individuals like me to learn data science and perform such projects as this. As an analyst, I look forward to gaining more knowledge as the course proceeds and applying these skills to my everyday activities in my workplace.