DATA WRANGLING REPORT

TAYEBWA CRISPUS

6th September, 2022

1. Data gathering process

Th datasets used in this project were gathered from three sources. The 'twitter_archive_enhanced.csv' file was provided by Udacity and I manually downloaded it from the provided resources.

Then, I programmatically downloaded the 'image_predictions.tsv' file from the Udacity servers by deploying the requests library (a python library).

I then used the python library called Tweepy to query the twitter API, in order to get the extra data about the tweets, from twitter. Each of these tweets' contents were stored on a single line in a text file. Which were quintessential to the successful and comprehensive completion of our project. I later extracted the necessary parameters and stored them in a csv file.

2. Data assessment process

2.1. Manual assessment

I did some manual assessment of the datasets so as to get familiar with the provided data and ascertain the consistency in the column values of the datasets. Microsoft excel was an important and efficient tool in the accomplishment of this task.

2.2. Programmatic assessment

this an iterative process in data cleaning, where I used methods like .info(), .describe(), .head(), .tail() and many more. These methods aided me in assessment of the various segments of the datasets for tidiness and quality issues.

3. Data cleaning process

3.1. Quality issues

Firstly, I checked for the data quality issues. These are issues that are associated with the content in the table.

The table below summarizes the quality issues I highlighted as well as the proposed solutions.

Quality issue	**dataset**	**solution**
Unclean values (contain HTML)	Archive dataset	Extract the HTML from the value
Wrong data types	All datasets	Convert to suitable datatypes
Invalid dog names	Archive dataset	Replace the invalid names
Wrong data	Archive dataset	Delete the retweets
Inconsistency in cases of the names	Archive dataset	Convert all names to uppercase as per the first name letter

3.2. Tidiness issues

After assessing for quality issues, I then examined the datasets for tidiness issues. These are issues related to the structural layout of the data. The table below summarizes the few tidiness issues I highlighted as well as the proposed solutions.

Tidiness issue	dataset	Solution
Unmerged tables	All datasets	Merge all tables into archive table
Many columns for dog stages	Archive dataset	Merge them into one column
Two values in single column	Archive dataset	subdivide the column into two

4. Data storage process

In order to start my analysis process, I needed to concatenate all my data into a single clean master table. I saved the table to *twitter_archive.csv* file as instructed in the project guidelines.