DATA WRANGLE REPORT ON WERATEDOGS (@Dog_Rates) TWITTER HANDLE

BY AGU, JAMES IFEANYICHUKWU

INTRODUCTION

This project is designed to assess our Data Wrangling abilities, and this is part of Udacity's Data Analytics Nanodegree Program.

During the data wrangling process, a number of steps were undertaken in order to get the data ready for analysis.
The entire data wrangling process was done using the Jupyter Notebook.

The data wrangling steps are listed below:

1. Gathering Data
2. Assessing Data
3. Cleaning Data
4. Storing Data

Gathering data

There were three datasets for our analysis and three separate methods were taken to download each of them.

The first method was to directly download the **twitter_archive_enhanced.csv** dataset and manually load it using the pandas **read_csv** object. This dataset contained over 5000+ datasets of

The second method was using the **Requests** library to download the **image_prediction** dataset from the website url provided.

The third and final method using the **Tweepy** library to query the remaining dataset **(tweet_json.txt)** through the Twitter API. The returned result was in JSON format, and this result was saved to our directory for further assessment and analysis.

Assessing data

This process involved both manual and programmatic assessment of our datasets. We leveraged the Pandas library functions for the programmatic assessment. The

functions used were DataFrame.info(), DataFrame.describe(), DataFrame.head(), etc. The matplotlib library was useful for visualization of the datasets in order to draw more insights on relationships among the columns.

We also checked our datasets for both **Quality** and **Tidiness** issues.

Quality issues are those relating to the content of our datasets which include missing data, invalid data, inaccurate date or inconsistent data.

Tidiness issues on the other hand, refers to structure of our data (columns, rows or table). We were able to observe about 8 quality issues and 2 Tidiness issues on our datasets. After the assessment, the three datasets were merged into a single column for further analysis.

Cleaning data

This step was all about cleaning our dataset based on all of the issues observed during the assessment stage. The cleaning process involved changing datatypes, dropping columns and rows that would not be useful, replacing or dropping missing rows/columns, correcting spelling errors, collapsing columns of similar features into a single column.

Storing data

At the end of the cleaning process, the cleaned data was saved and stored as **twitter_archive_master.csv** for exploratory and explanatory data analysis.