Data Wrangling report on WeRateDogs Twitter archive Data

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. This project aims to create exciting and trustworthy analyses and visualizations from the data obtained from the Twitter archives of the platform. Before the visualizations, the data must be subjected to the Data Wrangling process to ensure that the visualizations created are accurate and trustworthy.

The Data wrangling process is divided into four stages; Data Gathering, Assessing Data, Cleaning Data and Exploratory Analysis.

1. **The Data Gathering Process**

Datasets for the project were not readily available and are located in three different sources. The methods for gathering them were also different. The first dataset was a "csv." File from the Twitter archive of WeRateDogs. I got this by downloading and loading the file into the Jupyter Notebook. The second dataset contained the retweet and favorite count for the dog posts, which I assessed using the file URL and the request. get library. I saved it as image_predictions_df. The third dataset involved querying the Twitter API to get the JSON file. The request for access to Twitter API was denied, so I used the tweet-json.txt file provided in Udacity for the project. After loading the three datasets into the Notebook, I moved on to start assessing the data visually and programmatically.

2. **Assessing Data**

The datasets gathered were assessed for Quality (Dirty) and Tidiness (Messy) issues. I carried a visual assessment and programmatic assessment using codes such as; .info(), .isnull(), .describe etc. The Quality and Tidiness issues discovered include;

Quality Issues

Twitter archive table

1. in_reply_to_status_id has missing and invalid rows.
2. inreply_to_user_id has missing/ invalid rows etc.
3. Dogs' names are missing, and some appear as "a", 'an,' O.
4. Rating numerator less than 10
5. Rating denominator less and more than 10.

Image Prediction table

i. Name of columns (p1, p1_conf, p1_dog etc) is not clear.

ii. Img_num column is not conveying any information.

iii. False p1_dog is usually not a dog.

Tidiness Issues

1. Image_prediction table and retweet_count and favourite_count table(df3) should be merged with the Twitter archive.

2. The rating numerator and denominator should be in one column.

3. The different classifications of the dogs (doggo,floofer,pupper,puppo) should be in one column.

After identifying the above-listed issues in the data set, I created copies of the data sets to clean.

3. Cleaning Data

I used the define, code, and test framework to do the cleaning task. Defining the task, followed by the code for the task, and finally, testing for completion.

The first cleaning process was to merge the three datasets into one based on their tweet_id. This process solves the tidiness issues where three columns have the tweet_id column. Then the next is to melt the different dog classes or stages (doggo, floofer) into one column named “dog_class”. The issues with the rating numerator and denominator were resolved. In addition to this, both ratings were merged into one column named “ratings”. The issues with dogs’ names were resolved by dropping some false dog names. The columns containing invalid/ missing data were also removed from the data set. Other issues found when assessing the data were also resolved. Finally, the clean dataset was saved and ready for analysis.

4. Exploratory Data Analysis

The clean data is then analyzed for trends and patterns. One of the critical results from the analysis is that majority of the dogs rated are in the pupper stage