DECI Project: WeRateDogs Data Wrangling Report

Data Gathering

In the data gathering process, I started by importing the necessary libraries: pandas, numpy, matplotlib.pyplot, seaborn, and json. Then, I read the given datasets to start working with the data. The datasets provided were:

1. twitter-archive-enhanced.csv
2. Image-predictions.tsv
3. Tweet-json

After loading the data, I merged all the datasets into a single CSV file named twitter-archive-master.csv.

Data Assessing

In the data assessing process, I used both visual and programmatic assessment methods. I checked for duplicates, missing values, and incorrect data types.

This assessment revealed 8 quality issues and 2 tidiness issues.

Data Quality Issues

1. **Useless Columns**: Certain columns in twitter_archive and tweet_json were unnecessary and needed to be removed.
2. **Missing Data**: A significant amount of missing data was present in the features of twitter_archive and tweet_json.
3. **Representation of Missing Values**: Missing values should be represented as None in twitter_archive and tweet_json.
4. **Expanded URLs**: The expanded_url column contained more than one URL.
5. **Incorrect Data Types**: Some columns in twitter_archive had incorrect data types.
6. **P2_dog Column Type**: The type of P2_dog in image_predictions was a boolean instead of an integer.
7. **P1, P2, and P3 Formatting**: The P1, P2, and P3 columns in image_predictions needed proper formatting.
8. **Column Renaming**: The created_at column in tweet_json should be renamed to timestamp and changed its formatte.

Data Tidiness Issues

1. **Lowercase Inconsistencies**: The columns P1, P2, and P3 in image_predictions sometimes used lowercase.
2. **HTML Tags**: The source column in twitter_archive contained HTML tags that needed to be removed.
3. **Dog Stages**: The dog stages in twitter_archive (doggo, floofer, pupper, puppo) should be merged into a single column.

Data Cleaning

In the cleaning process, I started by making a copy of all datasets. Then, I tackled each issue by defining it, solving it, and testing the solution.

1. **Issue 1: Useless Columns**
   I used the drop() method to remove columns with a lot of missing data. Initially, I faced some problems, but I managed to handle them effectively.
2. **Issues 2 and 3: Missing Values**
   I replaced missing values with None using the fillna() method.
3. **Issues 4 and 5: Data Types and Expanded URLs**
   I fixed the data type of the timestamp column and formatted the expanded_url column correctly.
4. **Issues 6, 7, and 1 (Tidiness): P1, P2, P2_dogs, and P3 Columns**
   I addressed the problems associated with the P1, P2, P2_dog, and P3 columns, ensuring they were correctly formatted and consistent.
5. **Issues 8 and 2 (Tidiness): Column Renaming and HTML Tags**
   I renamed the created_at column to timestamp and removed HTML tags from the source column and changed its formatte.
6. **Issue 3 (Tidiness): Dog Stages**
   I merged the dog stages columns (doggo, floofer, pupper, puppo) into a single column.

After addressing these issues, I ensured the cleaned data was correctly formatted, consistent, and ready for analysis. The cleaned copies of the datasets were stored in twitter-archive-master.csv.

Conclusion

This data wrangling project involved gathering, assessing, and cleaning the WeRateDogs datasets to prepare them for analysis. The final twitter-archive- master.csv file is now comprehensive and ready for further analysis, providing a solid foundation for deriving meaningful insights from the WeRateDogs data.