Project: **Wrangling and Analyze Data**
Student: **Mehrol Bazarov**

Udacity 4th project
WRANGLE REPORT
(300-600 words) - mine (804 words)

I am pleased to present my data wrangling report for the WeRateDogs project. This project involved the data gathering, assessing, cleaning, and storing processes to prepare the data for further analysis and visualization.

Data Gathering

For this project, I collected data from three different sources. The first source was the WeRateDogs Twitter archive, which was available as a downloadable file named "twitter_archive_enhanced.csv". The second source was the image predictions file, "image_predictions.tsv", which contained predictions for each tweet's images using a neural network. I obtained this file programmatically. The third source was the "tweet_json.txt" file, which provided additional information such as retweet count and favorite (like) count for each tweet. I gathered this data using the Twitter API. Actually the third source was also ready in the Udacity website.

Assessing Data

To assess the data for quality and tidiness issues, I performed both visual and programmatic assessments. For visual assessment, I used Excel to get a better understanding of the data and identify any noticeable anomalies. Programmatic assessment was conducted using the Pandas library, allowing me to programmatically analyze the data for issues such as incorrect values, missing data, inconsistent formatting, and duplicate entries.

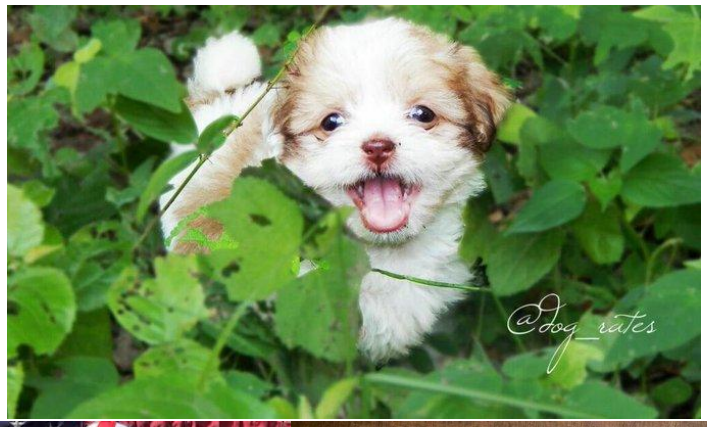
During the assessment, I identified a total of *nine* data quality issues and *three* tidiness issues. These issues encompassed various aspects of the data, including abnormal rating denominators and numerators, incorrect data types, inconsistent naming conventions, duplicates, and entries not related to dogs in the image predictions.

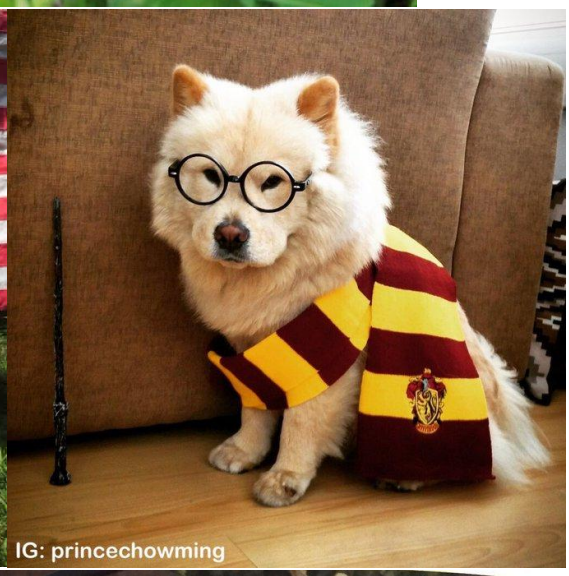
Cleaning Data

To address the identified issues, I followed a systematic data cleaning process. For each issue, I documented the problem, defined the cleaning operations required, and implemented the necessary code to rectify the issues. The cleaning process involved tasks such as converting data types, removing unnecessary columns, resolving inconsistencies, dropping duplicates, and filtering out irrelevant entries.

One interesting issue I encountered was the handling of abnormal rating numerators in the Twitter Archive. To address this, I used value counts to identify incorrect ratings and cross-checked them with the corresponding tweet text to determine if the ratings were not related to dogs. I corrected the numerator values based on the text analysis and dropped entries that were not dog-related. To make this process more engaging, I

included images of the dogs whose numerator values were changed based on the tweet text.





Abnormal Rating Numerators

In the Twitter Archive, I observed abnormal rating numerators such as 1, 2, 3, and 4, which did not seem to represent dog ratings. To resolve this, I examined the corresponding tweet text and identified that these ratings were not associated with dogs. Therefore, I decided to drop these entries from the dataset. Here are some pictures of these non-dog animals:






By removing these non-dog entries, we reduced the number of unwanted animals from this dataset. Anyway, we can not delete all of them, it is impossible to read all of the dataset.

Handling Tidiness Issues
After resolving the data quality issues, I turned my attention to addressing the tidiness issues.

Combining Dog Stages
The Twitter Archive dataset included separate columns for dog stages, namely doggo, floofer, pupper, and puppo. To adhere to the principles of tidy data, I combined these columns into a single column named "dog_stage." This consolidation improved the structure and consistency of the dataset.

Merging Tweet JSON Data
To integrate the additional information from the tweet_json file, I merged it with the Twitter Archive dataset. This allowed me to enrich the dataset with essential metrics such as retweet count and favorite count, providing valuable insights into the popularity and engagement of each tweet.

Merging Image Prediction Data
Finally, I merged the image_prediction data with the Twitter Archive dataset. This integration enabled us to include image predictions for each tweet, providing additional context and visual information. To ensure data integrity, I filtered out the rows with null values, retaining only the entries with valid image predictions.

These tidiness operations streamlined the dataset and made it more comprehensive for subsequent analysis.

Storing Data

After completing the data cleaning process, I saved the gathered, assessed, and cleaned master dataset to a CSV file named "twitter_archive_master.csv". This file contained the consolidated and cleaned data ready for analysis and visualization.

Conclusion

In conclusion, the data wrangling phase of the WeRateDogs project was an exciting and rewarding experience. Through data gathering, assessing, cleaning, and storing, I transformed raw and messy data into a clean and structured dataset suitable for further analysis. By addressing various quality and tidiness issues, I ensured the integrity and reliability of the data. This project allowed me to enhance my skills in data wrangling and provided valuable insights into the challenges and intricacies of working with real-world data.

I am eager to proceed to the next phase of the project, where I will analyze and visualize the data to gain further insights into the fascinating world of WeRateDogs.