Introduction:

In this project, I undertook comprehensive data wrangling efforts to clean and prepare two datasets for analysis: `image_prediction.csv` and `twitter_archive_enhanced.csv`. The goal was to address quality and tidiness issues to ensure accurate insights and robust data analysis.

Quality Issues and Solutions:

Image Prediction Dataset:

1. **Prediction Values Exceeding Limits:** Row 106 contained predictions over 1, which was resolved by dropping the row entirely. Ensuring predictions are within the expected range maintains the integrity of our analysis.
2. **Inconsistent Float Precision:** The `p1_conf`, `p2_conf`, and `p3_conf` columns contained floats of varying lengths. These were standardized by rounding each value to six decimal places, ensuring consistency across the dataset.
3. **Duplicate URLs:** The `jpg_url` column had duplicate entries, which were removed to ensure each image was unique. This step was crucial to avoid redundant data and potential bias in image analysis.

Twitter Archive Enhanced Dataset:

4. **Missing Expanded URLs:** Some entries in `expanded_urls` were missing, indicating tweets without images. These rows were dropped to maintain consistency, as image presence is critical for our analysis.
5. **HTML in Source:** The `source` column contained full HTML anchor tags, which were cleaned to retain only the text within the `href` attribute. This made the source data more readable and easier to work with.
6. **Non-Datetime Timestamps:** The `timestamp` field was not a proper datetime object. This was corrected by converting it to the appropriate datatype, facilitating accurate time-based analysis.
7. **Rating Outliers:** Outliers in `rating_numerator` and `rating_denominator` were addressed using:
    - **IQR Method:** Replaced outliers in `rating_numerator` with the mean of the normal range, ensuring extreme values didn't skew the data.
    - **Z-Score Method:** Replaced outliers in `rating_denominator` with the mean of the normal range, maintaining the integrity of our rating analysis.
8. **Invalid Names:** Entries like 'a', 'an', and 'None' in the `name` column were removed to ensure all names were valid. This step was essential for accurate categorization and analysis of the dataset.

Tidiness Issues and Solutions:

Image prediction Dataset:

The columns `p1`, `p2`, `p3` were repetitive with different attributes. These were combined into a single-column structure: `prediction_level`, `prediction`, `confidence`, and `is_dog_breed`. This consolidation made the dataset more streamlined and easier to interpret.

Twitter Archive Enhanced Dataset:

The dog stage columns (`doggo`, `floofer`, `pupper`, `Puppo`) were consolidated into one `dog_stage` column. If multiple stages were applied, the entry was marked as "multiple." This tidied up the dataset, making it more logical and reducing redundancy.

Conclusion

Through these cleaning efforts, the datasets are now well-prepared for analysis. The data's integrity and usability have been significantly improved by addressing quality issues and enhancing tidiness. These transformations ensure accurate, reliable insights in any subsequent analysis. The wrangling process involved meticulous attention to detail, ensuring the final datasets were clean, consistent, and ready for in-depth exploration.