

■ Data Cleaning Process

Before modifying any column, a copy of the column should be created. All transformations must be applied to this copy, and once the results are validated, the original column is deleted and replaced with the cleaned version.

1. Column name

The column contained names, surnames, and honorifics in the same field. The steps taken were:

- Identify records with more than three words to detect potential honorifics.
- The honorifics found were: Ms., Mrs., Mr., DDS, MD, Jr., Dr., III, PhD, DVM, Miss.
- A conditional column was created in Power Query to extract honorifics when present.
- Replace functions were then used to remove honorifics from the name column.
- Finally, the column was split into 'First Name' and 'Last Name' using space as a delimiter.

2. Column gender

Values showed inconsistencies like 'f', 'FEMALE', 'female'. To standardize:

- Applied 'Capitalize Each Word'.
- Replaced 'F' with 'Female' and 'M' with 'Male' (exact match option enabled).
- Verified that only 'Male' and 'Female' remained.

3. Column age

- Null values were found.
- Calculated the average using 'Average' in the 'Statistics' tab.
- The rounded average was used to replace null values.

4. Column years_of_experience

- Same process as in 'age': average calculation and replacement of nulls.

5. Column hourly_rate (USD)

- Mixed formats found: '100', 'USD 100', '\$40'.
- Replaced 'USD ' and '\$' with empty string.
- Changed data type to numeric.

6. Column rating

- Removed values equal to 0.
- Kept null values to avoid distorting analysis.

7. Column is_active

- Contained values 0, 1, N, Y.
- Applied 'Capitalize Each Word'.
- In Power Query replaced: '0' and 'N' with 'Inactive'; '1' and 'Y' with 'Active'.

- Finally converted to boolean (TRUE/FALSE).