



HR Data Cleaning Utilities (v 1 . 0)

This notebook demonstrates how to **simulate messy HR data** and then build a cleaning pipeline to make it analysis-ready.

Data cleaning is a critical step in People Analytics — poor quality data = misleading insights.

1 Create a Messy HR Dataset

We start from the processed dataset and intentionally add issues:

- Duplicates
- Missing values
- Inconsistent casing
- Invalid dates

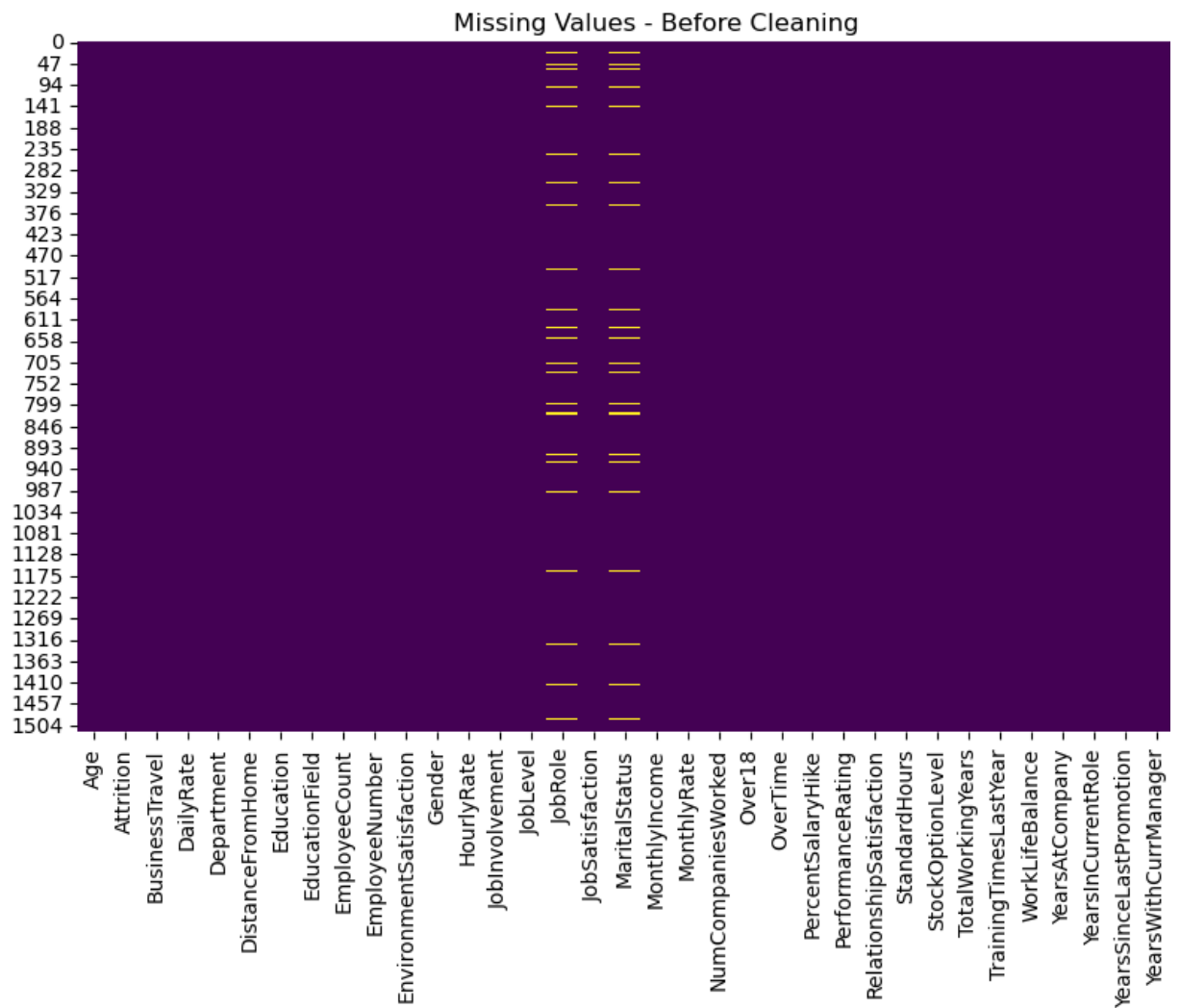
Messy dataset created !

2 Explore the Messy Data

Before cleaning, let's check shape, null values, and visualize missing data.

Initial shape: (1520, 37)

Age	0
Attrition	0
BusinessTravel	0
DailyRate	0
Department	0
dtype: int64	



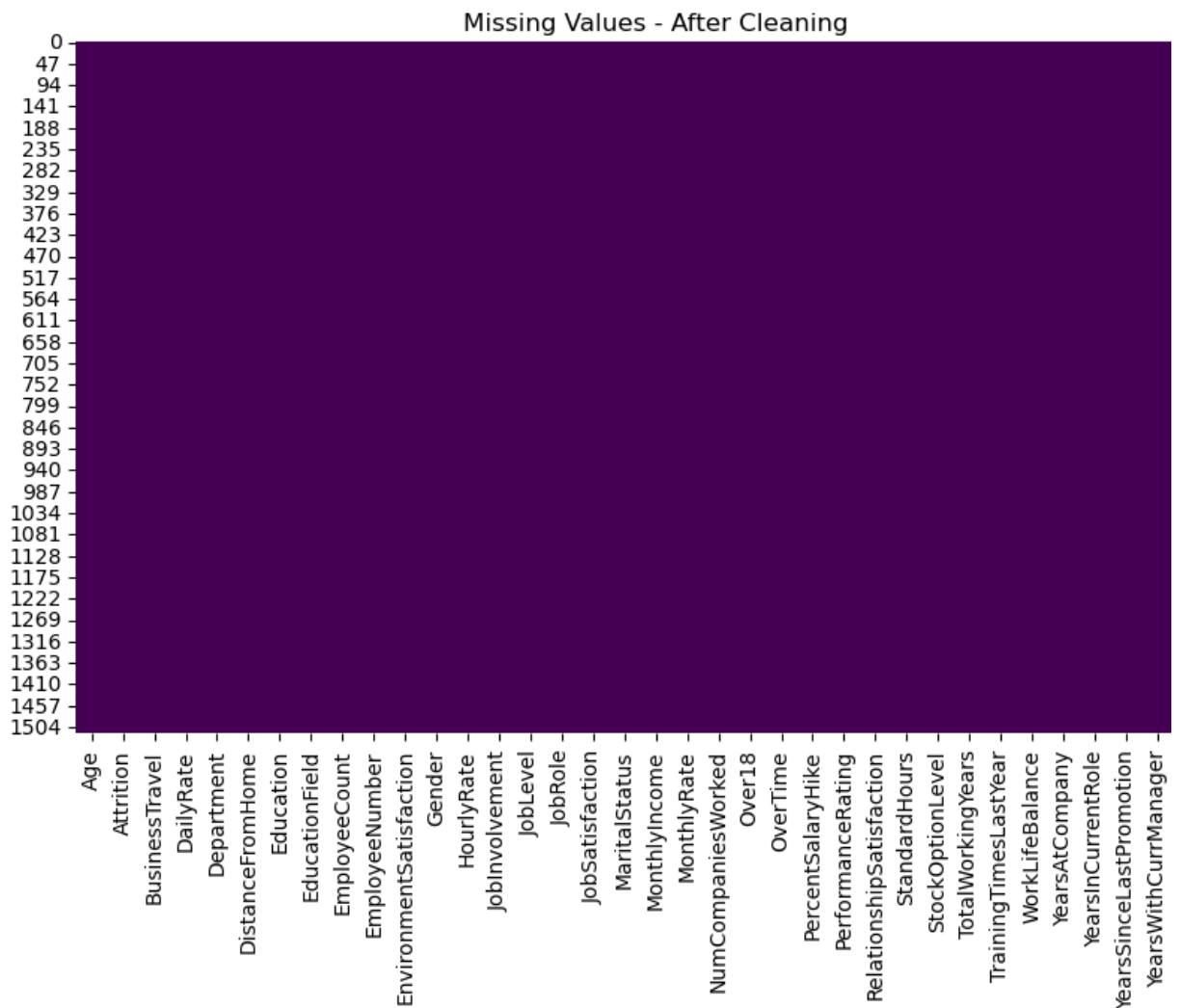
3 Apply Cleaning Steps

Now we:

- 1 . Remove duplicates
- 2 . Fill missing values
- 3 . Normalize categorical values
- 4 . Convert date columns into proper format

4 Verify the Cleaning

Check if missing values reduced and visualize again.



Final shape: (1519, 37)

5 Export the Cleaned Dataset

The cleaned dataset can now be used for further People Analytics projects.

Cleaning complete! Cleaned dataset saved as `cleaned_hr_data.csv`

Create a Before -After Collage

Collage saved at `images/missing_values_collage.png`



Conclusions

- Automated pipeline successfully cleaned the dataset.
- Issues fixed: duplicates, missing values, inconsistent casing, invalid dates.
- Before-After Collage created