Merging Datasets Using R

Assil Noureddine (400924750)

2025-06-11

noureddine.assil@stud.hs-fresenius.de

Abstract

This handout demonstrates how to merge two datasets using the `tidyverse` package in R. We use fictional employee and salary data to explore `left_join()` and `inner_join()`, visualize results, and summarize key findings. The project showcases essential skills for working with relational data and reporting results using Quarto.

1. Introduction

Data analysis often involves working with multiple datasets, each containing pieces of relevant information. To gain meaningful insights, analysts must combine these datasets accurately. Merging datasets means integrating them based on one or more common keys (columns) (R Core Team, 2025).

Often, information about a subject (such as an employee) is spread across multiple tables. Joining these datasets using common identifiers allows analysts to create comprehensive datasets for analysis.

In this project, we simulate this situation using two datasets: - `employees.csv`: Contains basic employee information. - `salaries.csv`: Contains salary data for some employees.

We'll merge these datasets and conduct simple analysis and visualization.

2. Why Merge Datasets?

Merging is essential in many scenarios, such as:

- Combining data from different sources: For example, linking sales data with customer demographics.
- Appending information: Adding additional columns (attributes) to an existing dataset.
- Longitudinal data analysis: Joining time-series data points from different periods.
- Data cleaning: Identifying mismatches or duplicates across datasets.

Merging facilitates integrated data views and supports more complex analyses (Wickham, 2023).

3. Types of Dataset Joins

Understanding different types of joins is crucial. Each determines how rows from two datasets are matched:

Join Type	Description	Use Case Example
Inner	Keeps only rows where the key exists in both datasets	Common customers in two systems
Left	All rows from left dataset + matching from right	Add additional info, keep all left rows

Join Type	Description	Use Case Example
Right	All rows from right dataset + matching from left	Rarely used, opposite of Left
Full	All rows from both datasets, with NA if no match	Merge all records, including unmatched
Anti	Rows in one dataset but not in the other	Find missing records

4. Tools & Packages in R for Merging

Tidyverse: dplyr Joins

The dplyr package provides a set of join functions (Wickham, 2023):

Function	Description
inner_join()	Inner join
left_join()	Left join
right_join()	Right join
full_join()	Full outer join
anti_join()	Rows only in left dataset

5. Step-by-Step Guide with Examples Using CSV Files

Suppose you have two CSV files:

- employees.csv:

employee_id	name	department
1	Alice	HR
2	Bob	IT
3	Charlie	Finance
4	David	marketing

- salaries.csv:

employee_id	salary
2	60000
4	75000
5	50000

Step 1: Load the CSV files into R

```
employees <- read.csv("Data/employees.csv")
salaries <- read.csv("Data/salaries.csv")
```

Step 2: Perform Different types of joins using dplyr

Make sure to load dplyr first:

```
library(dplyr)
```

5.1 Inner Join — Employees with salaries available

```
inner_merged <- inner_join(employees, salaries, by = "employee_id")
print(inner_merged)
```

```
  employee_id  name department salary
1           2   Bob         IT  60000
2           4 David  Marketing  75000
```

5.2 Left Join — All employees, with salaries where available

```
left_merged <- left_join(employees, salaries, by = "employee_id")
print(left_merged)
```

```
  employee_id    name department salary
1           1   Alice         HR     NA
2           2     Bob         IT  60000
3           3 Charlie    Finance     NA
4           4   David  Marketing  75000
```

5.3 Full Join — All employees and all salary records combined

```r
# Full join to merge all rows from both dataframes
library(dplyr)

full_merged <- full_join(employees, salaries, by = "employee_id")
full_merged
```

```
  employee_id    name department salary
1           1   Alice         HR     NA
2           2     Bob         IT  60000
3           3 Charlie    Finance     NA
4           4   David  Marketing  75000
5           5    <NA>       <NA>  50000
```

5.4 Anti join

```r
unmatched_employees <- anti_join(employees, salaries, by = "employee_id")
print(unmatched_employees)
```

```
  employee_id    name department
1           1   Alice         HR
2           3 Charlie    Finance
```

6. Handling Common Issues

- Duplicated keys: Multiple rows with the same key cause the merged dataset to grow unpredictably. Use distinct() or aggregation beforehand.
- Missing keys: Always check for NA values after joins. Use anti_join() to find unmatched records.
- Data type mismatch: Ensure keys are of the same type (e.g., both numeric or both character).
- Performance: For very large datasets, the data.table package's merge function is optimized for speed (R Core Team, 2025).

7. Best Practices

- Clean and preprocess datasets before merging.
- Always verify results after merging (row counts, NA values).
- Document the join type used and rationale.
- Use version control (Git) to track changes in data processing scripts (Association, 2020).

8. Summary & Further Reading

Merging datasets is foundational in data science. R provides multiple methods—from base R to powerful tidyverse tools. Choosing the right join and handling edge cases correctly ensures clean, reliable data for analysis.

Recommended Reading:

- Wickham, H., & Grolemund, G. (2016). R for Data Science. O'Reilly Media.
- Wickham, H. (2023). dplyr join functions documentation. https://dplyr.tidyverse.org/reference/join.html
- R Core Team. (2025). merge function documentation. https://stat.ethz.ch/R-manual/R-devel/library/base/html/merge.html

Discussion

We observe differences in average salary across departments. This could reflect job roles, experience levels, or salary negotiation practices. The left_join() strategy helps include all employees in the analysis, even if some lack salary records.

Future Work Could:

Include more datasets (e.g., performance data) Explore mismatched joins Handle duplicate or missing IDs

Conclusion

This project showed how to combine datasets using dplyr joins in R. We practiced left_join() and inner_join(), and demonstrated data exploration and visualization techniques. These tools are powerful for preparing data in real-world projects.

Affidavit

I hereby affirm that this submitted paper was authored unaided and solely by me. Additionally, no other sources than those in the reference list were used. Parts of this paper, including tables and figures, that have been taken either verbatim or analogously from other works have in each case been properly cited with regard to their origin and authorship. This paper either in parts or in its entirety, be it in the same or similar form, has not been submitted to any other examination board and has not been published.

I acknowledge that the university may use plagiarism detection software to check my thesis. I agree to cooperate with any investigation of suspected plagiarism and to provide any additional information or evidence requested by the university.

Checklist:

[X]The handout contains 3-5 pages of text.

[X]The submission contains the Quarto file of the handout.

[X]The submission contains the Quarto file of the presentation.

[X]The submission contains the HTML file of the handout.

[X]The submission contains the HTML file of the presentation.

[X]The submission contains the PDF file of the handout.

[X]The submission contains the PDF file of the presentation.

[X]The title page of the presentation and the handout contain personal details (name, email, matriculation number).

[X]The handout contains a abstract.

[X]The presentation and the handout contain a bibliography, created using BibTeX with APA citation style.

[X]Either the handout or the presentation contains R code that proof the expertise in coding.

[X]The handout includes an introduction to guide the reader and a conclusion summarizing the work and discussing potential further investigations and readings, respectively.

[X]All significant resources used in the report and R code development.

[X]The filled out Affidavit.

[X]A concise description of the successful use of Git and GitHub, as detailed here: https://github.com/hubchev/make_a_pull_request.

[X]The link to the presentation and the handout published on GitHub.

[Assil Noureddine,] [11 june 2025,] [Kôln]

References

Association, A. P. (2020). *Publication manual of the american psychological association: The official guide to APA style* (7th ed.). American Psychological Association.

R Core Team. (2025). *Merge function — r base documentation.* https://stat.ethz.ch/R-manual/R-devel/library/base/html/merge.html

Wickham, H. (2023). *Dplyr join functions.* https://dplyr.tidyverse.org/reference/join.html