

ANALYSIS REPORT

EMPLOYEES DATASET

Generated: 2026-02-03 13:46

SUMMARY

Total records analyzed:	300
Records after cleaning:	300
Records with valid data:	300
Instances created:	300

1. Exploratory Data Analysis

1.1 General Information of the Dataset

The original dataset contains **300** records. After the cleaning process, **300** records were retained, representing a retention rate of **100%**.

The data includes personal, contact, and employment information of employees, including hire dates, termination dates, department, and salary.

1.2 Missing Values

Table 1: Missing Values by Column

Column	Missing	Percentage
phone	13	4.33%
address	19	6.33%
termination_date	145	48.33%

2. Gender Consistency Analysis

2.1 Gender vs Name Validation

The **gender-guesser** library was used to validate the consistency between the registered gender and the employee's name.

- Genders validated using name inference
- Employees with inconsistent gender: **0**
- Method: Comparison of registered gender vs inferred from name

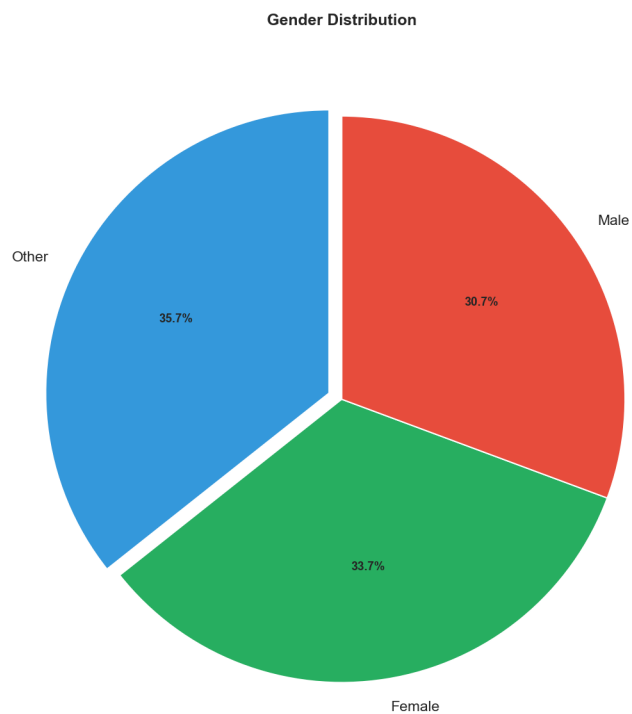


Figure 1: Gender Distribution

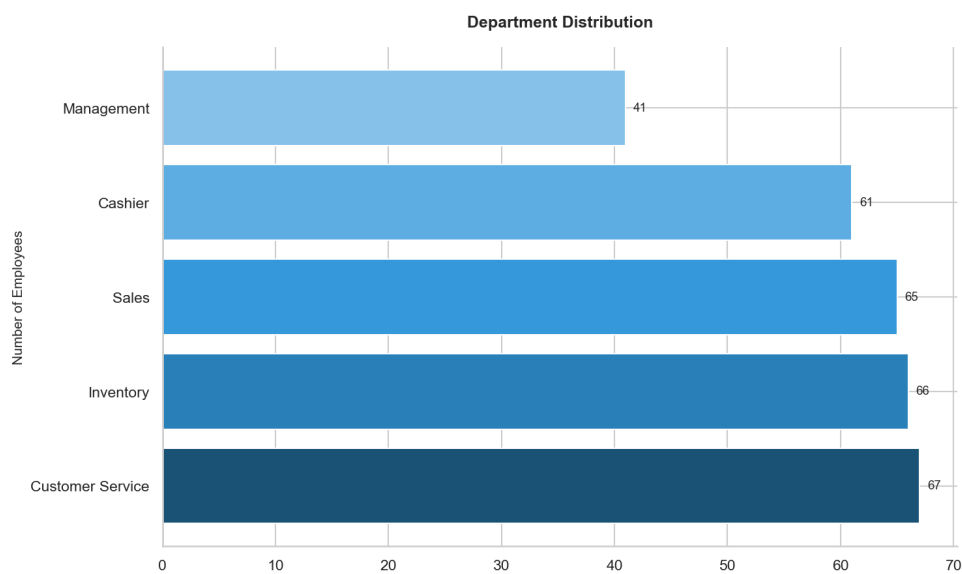


Figure 2: Department Distribution

3. Date and Consistency Validation

3.1 Age at Hiring

- Employees hired as adults: **83**
- Employees hired as minors: **217**

3.2 Termination vs Hiring

- Employees terminated before hiring: 1

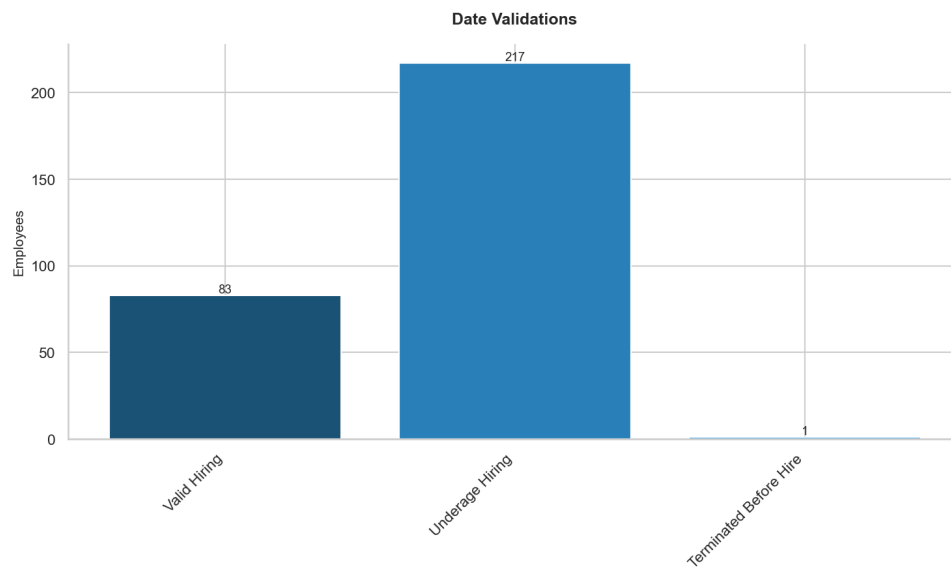


Figure 3: Date Validations

4. Applied Cleaning Process

4.1 transformations performed

- Email validation and cleaning (standardized format)
- Phone number cleaning (international format, add 10 zeros to invalid numbers)
- Gender inference based on name (gender-guesser)
- Removal of records without employee_id or name
- Removal of Age column (redundant/inconsistent data)
- Conversión de fechas a formato datetime

4.2 Cleaning Statistics

Table 2: Cleaning Process Summary

Metric	Value
Original records	300
Records after cleaning	300
Records removed	0
Emails processed	300
Phones cleaned	300
Genders inferred	300

5. Relations Between Datasets

5.1 Sales Activity Analysis

- Employees with associated sales: **229**
- Employees without associated sales: **71**

5.2 Relationship Inconsistencies

- Sales with invalid Employee ID: **0**

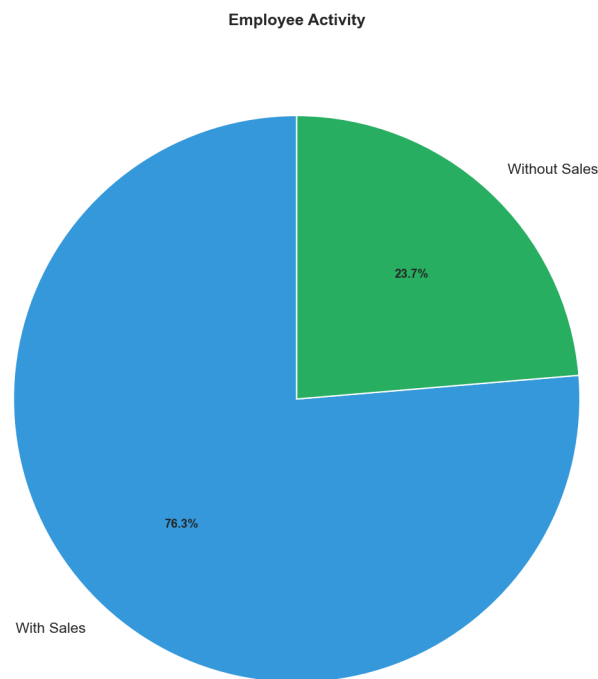


Figure 4: Employee activity

6. Pipeline Validation Results

6.1 Applied Validation Rules

- **termination_is_after_hire**: Termination date major that hire date
- **termination_is_after_birthdate**: Employee major or igual 18 years old at termination
- **valid_email**: Email with valid format
- **valid_phone**: Phone with valid format

6.2 Validation Results

Table 3: Validation Results by Rule

Validation	VValid	Invalid	N/A	% Success
termination_is_after_h	154	146	145	51.3%
termination_is_after_b	56	244	145	18.7%



Figure 5: Validation Details