

DATA CLEANING AND LEARNING REPORT

Intern Tasks – February 2026

Introduction

Data cleaning is an important step in data analysis. In this internship, I worked on different datasets and I did many cleaning operations. Real datasets have many problems like missing values, wrong data, and incorrect formats. I did these tasks to understand real-world data and improve my practical skills.

Objectives

- I learned basic data cleaning methods.
- I found errors in datasets and fixed them.
- I improved data quality before analysis.
- I worked with different tools and file types.
- I gained practical experience by doing real tasks.

Manual Checking and Missing Data

Manual Audit and Logic Check

First, I checked a large Excel file manually to understand the data. I did this operation to find logical errors. I found problems like negative age values and future dates, which are not possible in real life.

Missing Data Cleaning

For this task I used titanic dataset having garbage values , I checked missing values in the dataset. I did this operation to make the data consistent. I converted all missing values into one standard format.

Logic and Language Tasks

Logic Detection

I worked on a housing dataset and checked conditions that looked incorrect. I did this operation to detect unrealistic data. For example, bathrooms were greater than bedrooms and some housing prices were negative.

R Language Task

I cleaned an iris flower dataset using only R language. I did this task to improve my R skills and understand how cleaning works without using Python.

Text, Date and Time Cleaning

Regex Task

For this task I have used Employees dataset. I cleaned text data using regular expressions. I did this operation to separate combined information into different columns and make the dataset more structured.

Encoding and Time Zones

For this task I have used Global Sales dataset. I fixed character encoding issues and converted different date and time formats into one standard format. I did this operation to maintain consistency in the dataset.

Data Visualization

Visual Data Quality Audit

For this task I have used Superstore dataset. I created a Power BI dashboard. I did this operation to visually check missing values and understand data quality using charts.

Advanced Data Tasks

Uncleanable JSON

For this task I have used Employees dataset. I converted nested JSON data into CSV format. I did this operation to make complex data easier to analyze. I also handled missing and list values during this process.

Data Pipeline

I created, merged, and cleaned multiple 10 raw CSV files. Also, created raw files using python script. Then done operations to create a final clean dataset and then created a Power BI report from it.