HR Employee Data Cleaning & Preparation

A walkthrough from raw data to clean dataset | By Saayan Chowdhury

Project Objective

- Clean and prepare raw HR employee data for analysis.
- Ensure no missing values, correct data types and standardized formats.
- Prepare dataset for further use in analysis or machine learning.

Initial Dataset Columns

- EmpID, Age, DepartmentType, Attrition, Salary, Experience
- Education, GenderCode, StartDate, ExitDate, DOB
- EmployeeStatus, EmployeeType, EmployeeClassificationType, JobFunctionDescription

Problems in Raw Dataset

- Incorrect date formatting (e.g., '2020+AC0-01+AC0-17')
- Garbled text encoding artifacts like '+AC0-' and '+ACO-'
- Missing values in critical fields (DOB, StartDate, Salary)
- Mixed or incorrect data types (string where numeric expected)

Cleaning Workflow in Python

- 1. Load CSV and inspect columns
- 2. Remove junk encoding using regex replacements
- 3. Convert date fields using smart parsing
- 4. Convert numerics with coercion for safety
- 5. Drop rows missing essential values
- 6. Add derived column: TenureYears = Today - StartDate

Cleaned Output Ready for Use

- No missing values in required columns
- Date columns correctly parsed to datetime format
- Numeric columns coerced and cleaned
- TenureYears feature added for analytical use
- Exported as cleaned_hr_data.csv

Tools and Libraries Used

- Python (Pandas, Regex, Datetime)
- Jupyter Notebook

Conclusion

- Raw HR data was transformed into a clean, analysis-ready format.
- Issues like encoding noise and date parsing were systematically handled.
- Final dataset can be used for Power BI dashboards or ML modeling.
- Fully reproducible using Python scripts.