Data Analyst Internship - Task 1: Data Cleaning and Preprocessing

This document combines all the outputs and details from the Task 1 data cleaning and preprocessing assignment. It includes an overview, summary of transformations, dataset samples, and generated charts.

Summary of Cleaning Operations

Task 1 — Data Cleaning and Preprocessing
Files included:
- task1_raw_dataset.csv : Original synthetic raw data with deliberate dirty entries.
- task1_cleaned_dataset.csv : Cleaned dataset ready for analysis.
- cleaning_script.py : Python script which implements the cleaning pipeline.
- README.md : Instructions and repo summary.
- SUMMARY.md : Short summary of changes made.
- images/ : Charts generated showing before/after snapshots.
- repo-ready zip: data-analyst-task1.zip

Summary of changes performed:
1. Trimmed leading/trailing whitespace from string fields (name, gender, country, email).
2. Standardized name casing (Title Case).
3. Standardized gender values to consistent categories: Male, Female, Other; unknowns set to null and then filled where applicable.
4. Converted age to integer; non-numeric ages coerced and filled with median age (29).
5. Parsed and standardized signup_date to ISO datetime (YYYY-MM-DD); unparseable dates filled with the mode date.
6. Standardized country names (e.g., 'usa', 'u.s.a.' -> 'United States'; 'uk' -> 'United Kingdom').
7. Validated emails; invalid or missing emails set to null.
8. Cleaned salary to numeric (removed currency symbols and separators); missing salaries filled with median salary (4500.0).
9. Removed exact duplicate rows.
10. Renamed columns to lowercase with underscores.

Counts:
- Original rows: 14
- After dropping duplicates: 13
- Duplicates removed: 1

Note: This synthetic dataset is supplied as an example. Replace task1_raw_dataset.csv with your own dataset (for example a Kaggle CSV) and run cleaning_script.py to reproduce the same pipeline on real data.

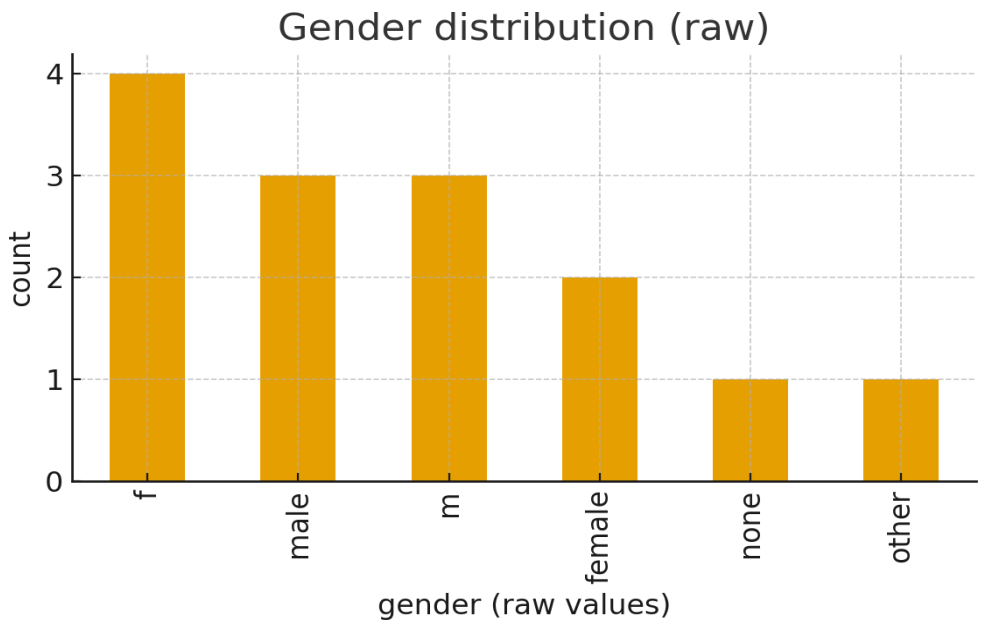
Raw Dataset (Sample)

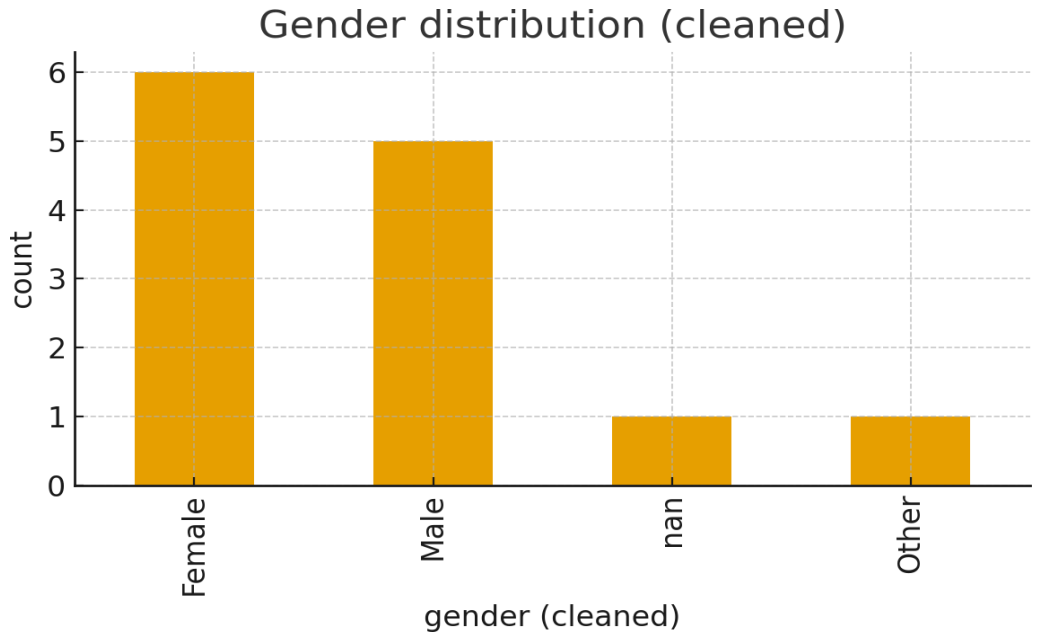
id	name	gender	age	signup_date	country	email	salary
1	Alice	F	29	2021-01-05	USA	alice@example.com	$5,000
2	bob	Male	34	05-02-2020	United States	bob@example.com	4500
3	Charlie	M	nan	2020/03/15	usa	charlie@example.com	■3000
4	diana	female	27	15-04-2021	India	diana@example.com	3500
id	name	gender	age	signup_date	country	email	salary
5	Eve	F	twenty-five	2021.06.01	U.S.A.	eve@@example.com	$4,200.00

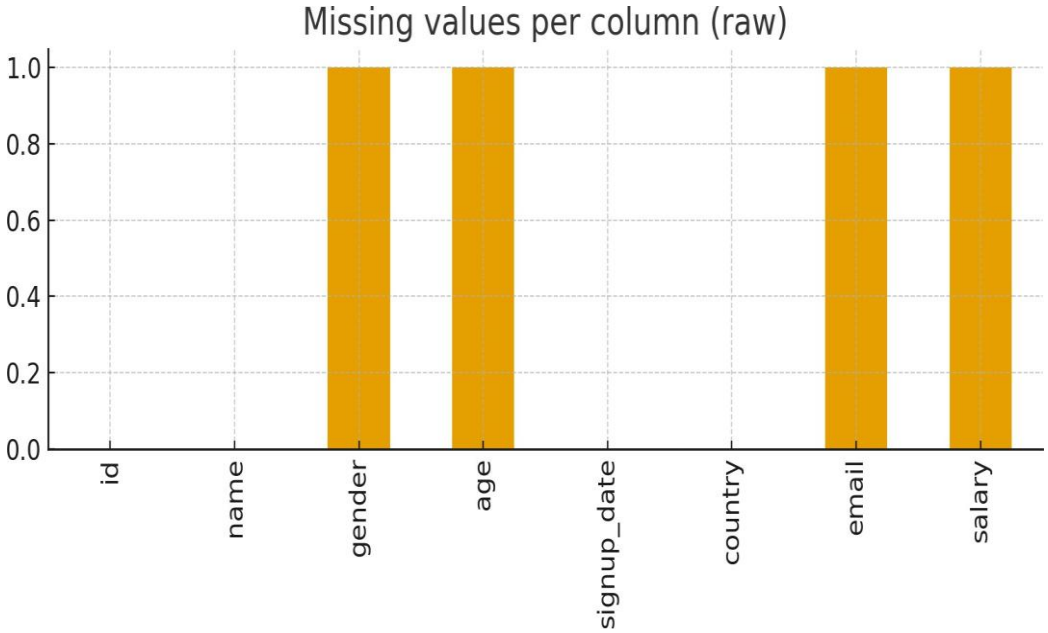
Cleaned Dataset (Sample)

id	name	gender	age	signup_date	country	email	salary
1	Alice	Female	29	2021-01-05	United States	alice@example.com	5000.0
2	Bob	Male	34	2020-02-05	United States	bob@example.com	4500.0
3	Charlie	Male	29	2020-03-15	United States	charlie@example.com	3000.0
4	Diana	Female	27	2021-04-15	India	diana@example.com	3500.0
5	Eve	Female	29	2021-06-01	United States	nan	4200.0

Generated Charts




Gender distribution (cleaned)
count
0
1
2
3
4
5
6
Female
Male
nan
Other
gender (cleaned)


Missing values per column (raw)
0.0
0.2
0.4
0.6
0.8
1.0
id
name
gender
age
signup_date
country
email
salary