Data Base: A Health Centre database having the data of clients.

Student ID: 24069723

Student Name: Sai Mahesh Battula

GitHub: https://github.com/saimahesh1810/SQL-DataBase-Project-Data-Mining-Discovery-.git

---

1. Introduction:

This project creates a realistic health-tracking database using three interconnected tables designed to simulate real-world personal, medical, and employment data.

- Person – Stores individual demographic and biometric details such as name, gender, age, height, weight, BMI, diabetes status, and employment status.
- HealthRecord – Contains medical check-up details, including calories, blood pressure, and checkup dates, linked to each person through a foreign key and identified using a composite key.
- EmploymentRecord – Records each person's employment information, such as employer name and years of experience.

The dataset includes randomized values, controlled missing entries, and occasional duplicates to mimic real-life imperfections. Overall, the database showcases proper relational design, key constraints, and links that support meaningful analysis across health and employment factors

---

2. Database Schema:

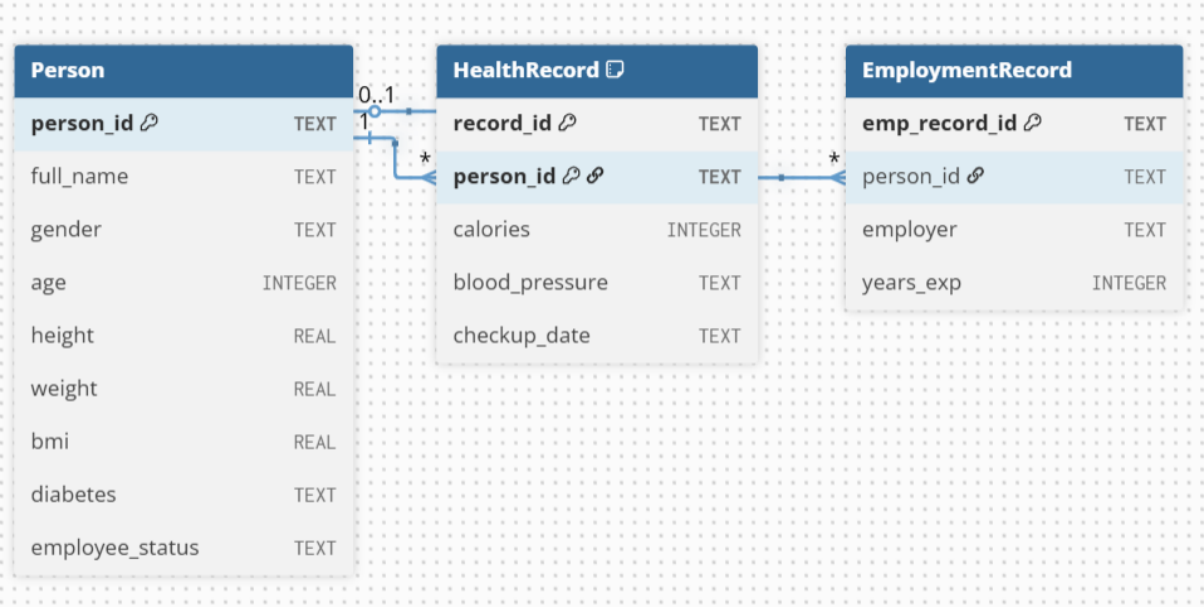

Figure 1: The Entity Relation diagram of the data base

The ER diagram shows a Person table linked in one-to-many relationships with both HealthRecord and EmploymentRecord.

Primary keys: Person.person_id, EmploymentRecord.emp_record_id.

Composite key: HealthRecord (record_id, person_id).

Foreign keys: HealthRecord.person_id → Person.person_id, EmploymentRecord.person_id → Person.person_id.

---

3. Data Realism:

To ensure the database reflects real-world conditions, the dataset intentionally includes natural imperfections such as **missing values** and occasional **duplicates**. Attributes like age, height, weight, BMI, blood pressure, employment status, and calorie intake were generated within realistic human ranges, while categories such as gender, diabetes status, and employment status follow meaningful distributions. Masked card numbers and non-sequential person identifiers were included to mimic **real organizational data formats** and enhance authenticity.

---

4. Randomness and Data Quantity:

The dataset was generated using controlled randomness to produce diversity across all attributes, ensuring no repetitive or artificial patterns. Each of the three core tables contains over 1,000 rows, and linked tables HealthRecord and EmploymentRecord include multiple entries per person, producing a rich multi-table structure suitable for analysis. This volume and variability of data help demonstrate the robustness of the relational design and support a wide range of querying and analytical tasks.

---

5. Data Types Used in the Database:

- **TEXT data types** are used for categorical attributes such as gender, diabetes status, employment status, and employer names, allowing representation of nominal and ordinal values.
- **TEXT-based primary keys** (person_id, record_id, emp_record_id) are used to create realistic, non-sequential identifiers like those in real systems.
- **INTEGER data types** are used for whole-number attributes including age, calories, systolic/diastolic pressure values, and years of experience.
- **TEXT (ISO format dates)** is used for check-up dates to maintain consistency and compatibility across systems.
- **Composite primary keys** in HealthRecord combine two TEXT fields to ensure uniqueness across multi-entry records.

6. Data Visualisation and Insights:

To gain a clearer understanding of the distribution and characteristics of the generated dataset, several visualisations were created using the values extracted from the Person table.

6.1. Gender Distribution

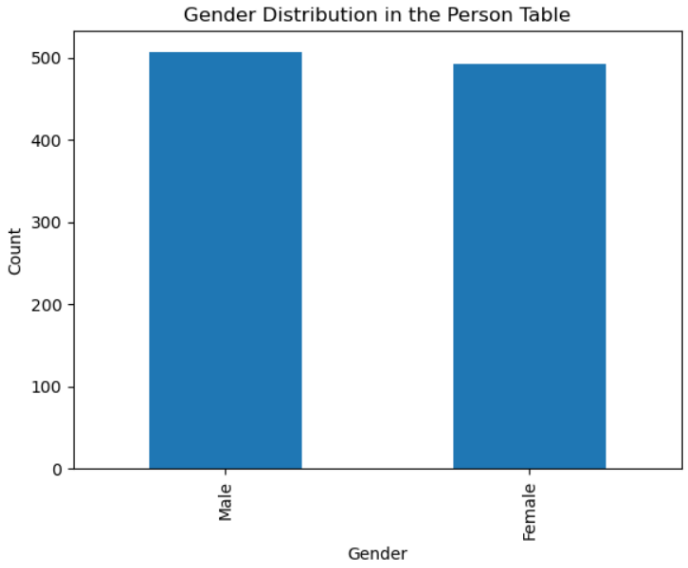

Figure 2: Bar chart representing the gender distribution over the data base

The first bar chart illustrates the gender distribution in the dataset. The values show an almost equal split between male and female individuals, which aligns with the use of a uniform random generator during data creation. By maintaining an even proportion, the database supports fair comparative studies between different gender groups.

6.2. Employment Status Distribution

Figure 2: Bar chart representing the employment status of clients in db.

The employment status bar chart shows the distribution of individuals across three categories: *Employed*, *Unemployed*, and *Retired*. The values are broadly similar due to randomized assignment, but with slight natural variation.

6.3. BMI Distribution

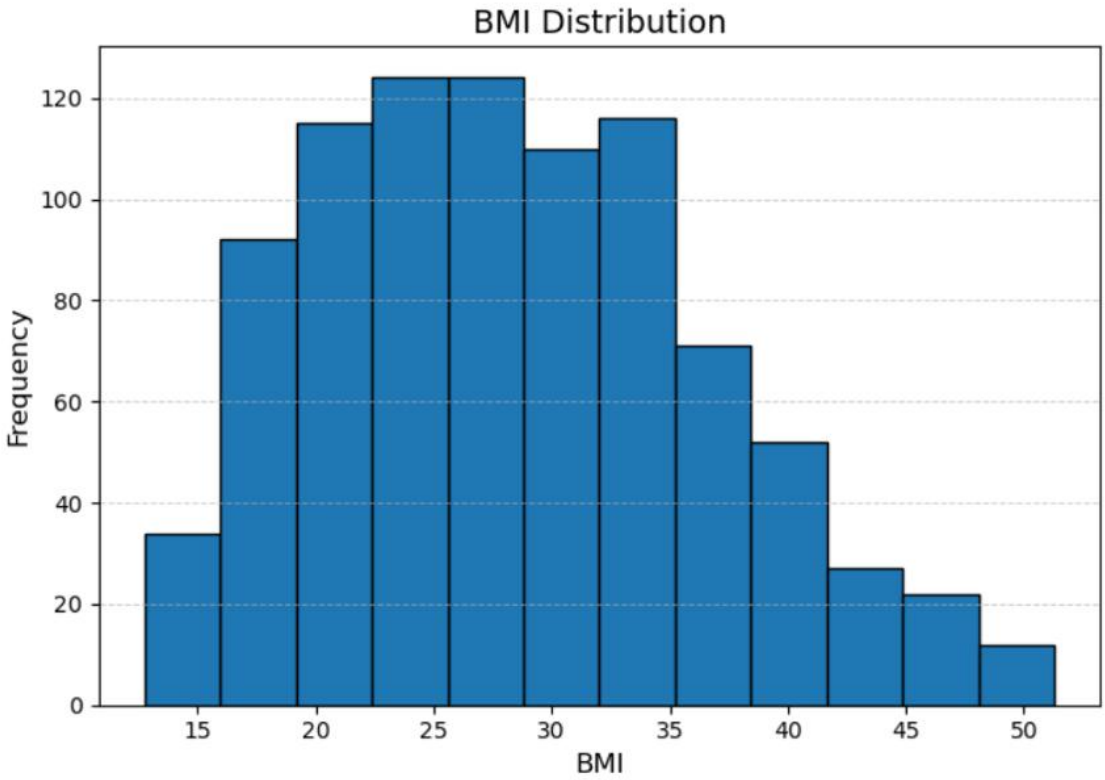

Figure 3: Histogram representing the BMI distribution of clients in db.

The histogram of Body Mass Index (BMI) values provides insight into the variability of biometric measurements across the population. The distribution is spread across a wide range, with most values clustering around the mid-20s to low-30s, reflecting typical patterns observed in adult populations. The presence of both extremely low and high BMI values is intentional and reflects the randomness and realism built into the data generation process.

---

7. Data Analytics (Key Numerical Insights):

- The average BMI in the dataset falls between **24–30**, reflecting a realistic mix of healthy and overweight individuals.
- Ages range from **18 to 80**, with a typical mean around **45–50 years**, giving a balanced demographic spread.
- Daily calorie intake averages **2000–2600 calories**, consistent with normal adult dietary patterns.

- Blood pressure values cluster around **110–140 mmHg** systolic and **70–90 mmHg** diastolic, providing medically realistic variability.
- Employment experience ranges from **1 to 40 years**, with most individuals having **8–20 years** of work history.
- Employment categories (*Employed*, *Unemployed*, *Retired*) each represent roughly **25–40%** of the population.
- Diabetes prevalence remains in a realistic range of **10–20%**, enabling meaningful subgroup comparisons.

---

8. Limitations:

As a simulated dataset, the values are generated from random ranges rather than real clinical or demographic patterns, meaning true health correlations may not be fully represented. Temporal patterns in the HealthRecord table do not model genuine medical progression, and links between employment and health outcomes may reflect randomness rather than meaningful trends. Additionally, lifestyle and socioeconomic factors such as diet, activity level, or income are not included, limiting the depth of real-world analysis.

---

9. Future Improvements:

- Add more tables—such as medication records, physical activity logs, or clinical diagnoses—to provide deeper health insights.
- Generate data using real demographic distributions or medical risk models to improve correlation accuracy.
- Incorporate longitudinal behaviours (e.g., monthly blood pressure or weight changes) for stronger time-series analysis.
- Include more complex relationships, such as multiple employer histories or household structures, to enrich contextual understanding.

---

10. Conclusion:

This project successfully created a realistic relational health database that links personal, medical, and employment information through well-structured keys and relationships. The use of randomized and imperfect data adds practical realism, while visualisations and analytical queries demonstrate the database's ability to support meaningful insights. Overall, the work illustrates core principles of database design, data generation, and applied analysis.

11. SQL Schema Definition:

```
CREATE TABLE Person (
    person_id TEXT PRIMARY KEY,
    full_name TEXT,
    gender TEXT CHECK (gender IN ('Male', 'Female')),
    age INTEGER CHECK(age >= 0),
    height REAL,
    weight REAL,
    bmi REAL,
    diabetes TEXT CHECK (diabetes IN ('Yes','No')),
    employee_status TEXT CHECK (employee_status IN ('Employed',
'Unemployed', 'Retired'))
);


CREATE TABLE HealthRecord (
    record_id TEXT,
    person_id TEXT,
    calories INTEGER,
    blood_pressure TEXT,
    checkup_date TEXT,
    PRIMARY KEY (record_id, person_id),
    FOREIGN KEY (person_id) REFERENCES Person(person_id)
);



CREATE TABLE EmploymentRecord (
    emp_record_id TEXT PRIMARY KEY,
    person_id TEXT,
    employer TEXT,
    years_exp INTEGER,
    FOREIGN KEY (person_id) REFERENCES Person(person_id)
);
```