

Phase 3: Database creation, table population, and business questions

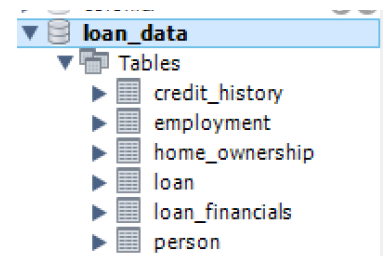
Phase 3: Database creation, table population, and business questions.....	1
Discussion of how you converted the dataset into tables.....	2
Challenges faced while importing your data and how did you overcome these data importation challenges:.....	3
A complete data dictionary for every table in your database.	4
The list of business questions.....	5

Discussion of how you converted the dataset into tables.

1. Converting the Dataset into Tables

To start with, I received a dataset in a single CSV file (loan_data.csv) that contained all the relevant information. Based on the ERD created earlier, I had split this dataset into six distinct tables to create a normalized relational structure and clearly define relationships between entities. Here's how I approached the process:

2. **Identifying Entities:** Using the ERD as a guide, I identified the main entities in the data: PERSON, EMPLOYMENT, CREDIT_HISTORY, LOAN, LOAN_FINANCIALS, and HOME_OWNERSHIP. I mapped specific fields in the dataset to each of these entities, which helped in breaking down the data into logical groupings.
3. **Data Extraction:** I used a Python script to load and parse the CSV file. Each row in the CSV represented a single record, which I then divided based on the entity it corresponded to. By organizing the data in this way, I was able to assign each attribute to the appropriate table.
4. **Data Transformation:** I applied a few transformations to prepare the data for import:
 - a. Converted text fields to appropriate data types, such as double for income and loan amount fields, and integer for ID fields.
 - b. Standardized categorical values like loan status and home ownership to ensure consistency across the database.
5. **Data Insertion into Tables:** Once the data was organized and transformed, I imported it into six separate tables in MySQL, as specified in the ERD:



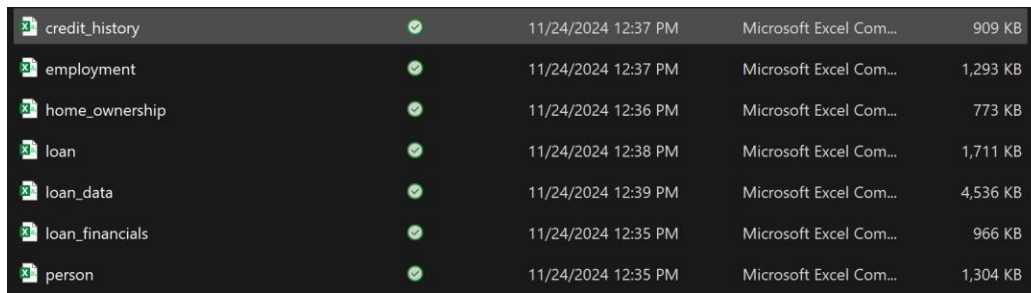
- a. **PERSON:** Stored personal information for each individual.
- b. **EMPLOYMENT:** Stored employment-related information linked to the PERSON table via person_id.
- c. **CREDIT_HISTORY:** Stored each person's credit score, credit history length, and any previous loan defaults.
- d. **LOAN:** Stored the details of individual loans, including loan amount, interest rate, and purpose.
- e. **LOAN_FINANCIALS:** Stored additional loan financial data such as the interest rate and percentage of income allocated to the loan.
- f. **HOME_OWNERSHIP:** Stored the home ownership status of each individual.

This approach helped me create a normalized database structure that minimized redundancy and maintained clear relationships between entities, making it well-suited for analysis.

Challenges faced while importing your data and how did you overcome these data importation challenges:

During the process of importing the data and splitting it into multiple tables, I encountered several challenges, which I tackled as follows:

1. **Splitting Data into Multiple Tables:** The initial dataset was in a single flat file, so I had to break it down into multiple tables according to the ERD structure. This required careful mapping of attributes to each table and ensuring that relationships were correctly represented. I used foreign keys to maintain referential integrity between tables, ensuring that each table could reference related data accurately.



credit_history	✓	11/24/2024 12:37 PM	Microsoft Excel Com...	909 KB
employment	✓	11/24/2024 12:37 PM	Microsoft Excel Com...	1,293 KB
home_ownership	✓	11/24/2024 12:36 PM	Microsoft Excel Com...	773 KB
loan	✓	11/24/2024 12:38 PM	Microsoft Excel Com...	1,711 KB
loan_data	✓	11/24/2024 12:39 PM	Microsoft Excel Com...	4,536 KB
loan_financials	✓	11/24/2024 12:35 PM	Microsoft Excel Com...	966 KB
person	✓	11/24/2024 12:35 PM	Microsoft Excel Com...	1,304 KB

- 2.
3. **Ensuring Correct Data Types:** Since the CSV format stores all data as text, I needed to ensure that each field was converted to the correct data type, such as integer, double, or string, based on the ERD's specifications. I handled this by explicitly casting each attribute to the correct data type within the Python script before inserting it into MySQL. This prevented type mismatches and improved data integrity.
4. **Handling Missing or Inconsistent Data:** Some fields in the dataset had missing or inconsistent values. For instance, credit scores or employment records were not available for every individual. I addressed this by assigning NULL values where appropriate and cleaning the data to ensure consistent formatting (e.g., standardizing categorical values).
5. **Foreign Key Constraints and Referential Integrity:** To maintain the relationships between tables, I used foreign keys. This required careful attention to avoid referential integrity issues, ensuring, for example, that person_id in the EMPLOYMENT, CREDIT_HISTORY, LOAN, and HOME_OWNERSHIP tables referenced existing records in the PERSON table. I verified data consistency and cross-referenced records during the insertion process to avoid orphan records.
6. **Data Loading Performance:** Given the large volume of data, loading it into multiple tables with relational constraints could have slowed down the import process. To optimize performance, I implemented bulk inserts in the Python script, which significantly sped up the data load. Additionally, I added indexes on primary and foreign keys to improve query performance, especially for complex joins during analysis.

By addressing these challenges, I was able to import the dataset into MySQL as a set of normalized, relational tables that adhered to the ERD structure. This process not only organized the data but also ensured it was optimized for further analysis.

(Attached Python Script for Loading Data into MySQL Workbench)

A complete data dictionary for every table in your database.

Tables	Column	Data Type	Description
PERSON	person_id (PK)	varchar	Unique identifier for each person (e.g., P53211858).
	person_age	int	Age of the person (e.g., 22).
	person_gender	string	Gender of the person (e.g., female).
	person_edu	string	Education level of the person (e.g., Master).
EMPLOYMENT	employment_id (PK)	varchar	Unique identifier for each employment record (e.g., employer ID).
	person_id (FK)	varchar	Foreign key referencing PERSON table (e.g., P53211858).
	person_income	double	Annual income of the person (e.g., 71948).
	person_emp_exp	int	Years of employment experience (e.g., 0).
CREDIT_HISTORY	person_id (FK, PK)	varchar	Foreign key and primary key referencing PERSON table.
	credit_score	int	Credit score of the person (e.g., 561).
	cb_person_cred_hist_length	int	Length of the person's credit history (e.g., 3).
	previous_loan_defaults_on_file	string	Indicates past loan defaults (Yes or No).
LOAN	loan_id (PK)	varchar	Unique identifier for each loan (e.g., L00327147).
	person_id (FK)	varchar	Foreign key referencing PERSON table (e.g., P53211858).
	loan_intent	string	Purpose or intent of the loan (e.g., PERSONAL, EDUCATION).
	loan_percent_income	double	Percentage of income allocated to the loan payment (e.g., 0.49).
	loan_status	int	Status of the loan (1 = approved, 0 = defaulted).
LOAN_FINANCIALS	loan_id (PK, FK)	varchar	Foreign key referencing LOAN table (e.g., L00327147).
	loan_int_rate	double	Interest rate on the loan (e.g., 16.02).
	loan_amnt	double	Loan amount (e.g., 35000).
HOME_OWNERSHIP	person_id (PK, FK)	varchar	Foreign key and primary key referencing PERSON table.
	person_home_ownership	string	Home ownership status (e.g., RENT, OWN).

The list of business questions.

1. What is the default rate for loans based on the applicant's credit score?
2. Is there a correlation between loan amount and loan interest rate?
3. What is the average income of applicants by employment experience level?
4. What are the key factors (such as income, employment experience, credit history length, or previous defaults) that most strongly predict loan defaults, and how do these factors vary by loan amount and purpose?
5. How does the loan default rate vary across different combinations of credit score and loan intent, and what are the average loan amounts and interest rates for each combination?
6. For applicants with the same income and similar employment experience, how does the loan interest rate vary by credit score, home ownership status, and previous loan defaults?
7. Which combination of demographic factors (such as age, gender, education) and loan characteristics (intent, amount, income percentage) is most associated with loan approvals and rejections?
8. How do credit scores and employment experience affect the probability of loan default for applicants in different income brackets, and does this relationship vary by loan intent?
9. What is the average time to loan repayment (for loans that have been fully repaid) by applicant demographic factors and credit history, and how does this vary by loan amount and purpose?
10. How does the percentage of income allocated to loan payments affect the likelihood of default, and does this relationship differ based on employment experience, education level, and home ownership status?
11. What are the patterns of default rates for applicants with specific combinations of credit score, loan amount, and previous loan defaults, and how does this influence the average interest rate offered?