Data Engineering Assignment Report

Student Name: Jeevana Sree B
Registration No: 12220183
Role: Data Engineering Intern
Date: 19-04-2025

Introduction

In this assignment, I worked with four different but related datasets to explore how data engineering concepts like cleaning, transformation, and merging help solve real-world business problems. Using Python, I walked through each dataset step-by-step to draw meaningful insights and answer four specific questions.

Tools Used: Python (pandas, matplotlib, seaborn) in a Jupyter Notebook environment

Datasets Used

1. **industry_client_details.csv** – Information about each client's ID, company size, industry, and location.
2. **subscription_information.csv** – Subscription details like type, start and end dates, and renewal status.
3. **payment_information.csv** – Records of how much each client paid, when, and using what payment method.
4. **financial_information.csv** – Economic indicators such as inflation and GDP growth, given by date range.

Implementation and Analysis

This section outlines how each question was approached and implemented using Python. I explain the steps I took to clean, manipulate, and analyse the data to reach meaningful answers.

Step 1: Loading the Data

I started by loading each dataset using pd.read_csv(). I checked the structure using .head(), .info(), .isnull().sum() to get a feel for the data and identify missing or unusual values and .duplicated().sum() to find the duplicates

Step 2: Cleaning the Data

I cleaned each dataset individually before combining them. Here's how:

industry_client_details:

- Fixed text formatting issues (extra spaces, inconsistent casing)
- Converted client_id to integers to ensure consistency
- Verified no duplicates or null values were present

subscription_information:

- Converted subscription dates to datetime format
- Ensured renewed was stored as a boolean (True/False)
- Made sure client_id values matched other datasets

payment_information:

- Cleaned payment_method names (fixed spacing and capitalization)
- Converted payment dates and extracted the year for time-based analysis
- Converted amount_paid to numeric in case of text errors

financial_information:

- Removed unnecessary index columns
- Made sure the financial period dates were in datetime format
- Checked that inflation and GDP data were numerical

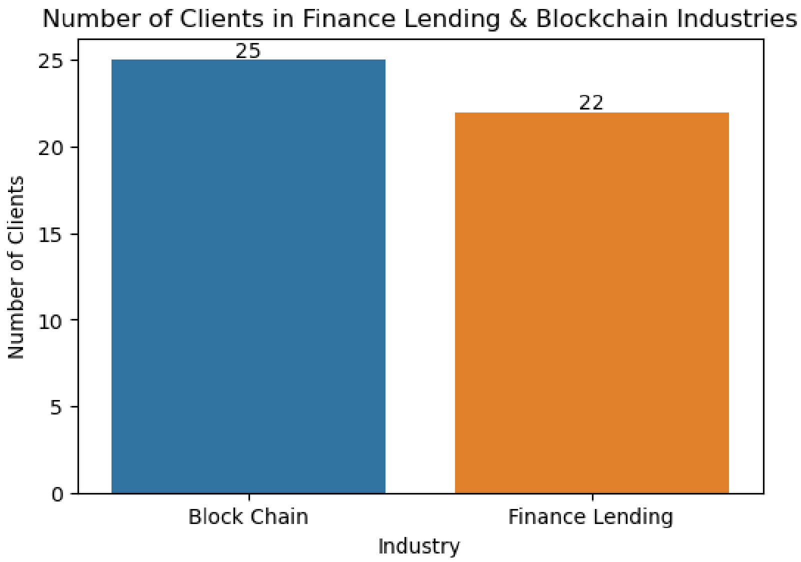
Step 3: Exploring the Data (Data Inspection)

Before jumping into analysis, I did some basic exploration using value_counts():

- Counted values in key columns like industry, company size, and location
- Looked at how many subscriptions were renewed or not
- Observed trends in payment methods across years
- Understood inflation and GDP values across time periods

Question-Wise Approach & Answers

Question 1: How many Finance Lending and Blockchain clients?



Approach:

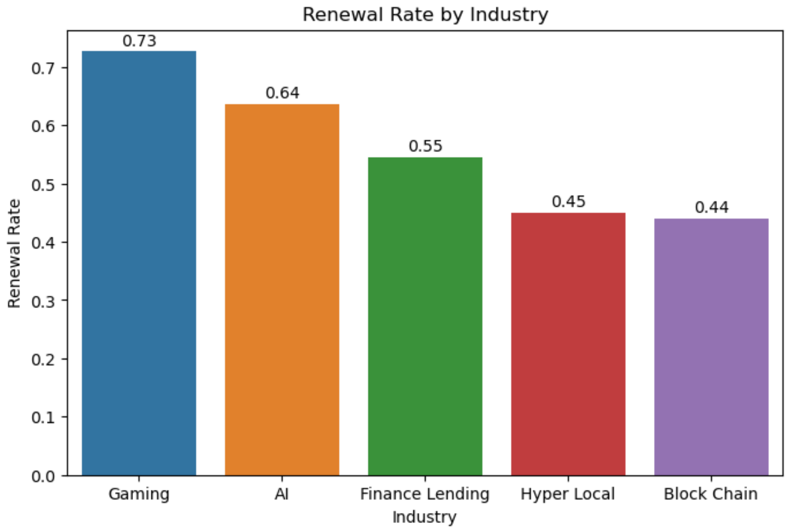
- Filtered the industry-client dataset to select only "Finance Lending" and "Block Chain"
- Counted how many clients were in each category
- Created a simple bar plot to show the numbers visually

Result: There are 47 clients in total — 22 from Finance Lending and 25 from Blockchain industries.

Question 2: Which industry has the highest renewal rate?

Approach:

- Combined (Merged) subscription data with industry data
- Grouped by industry and calculated the average renewal rate
- Used a bar chart to display the result



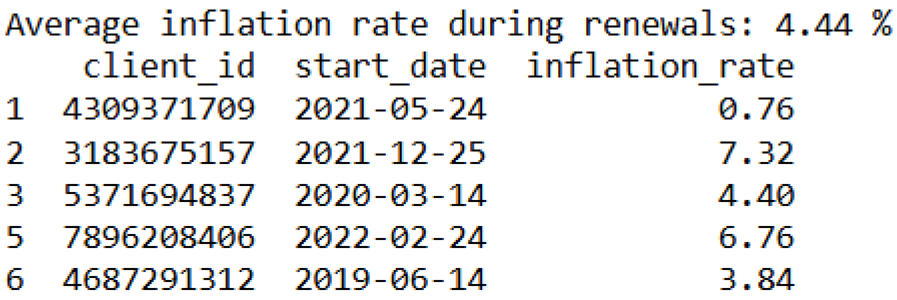
Result: The Gaming industry had the highest renewal rate, with around 73% of its clients renewing.

✅ Question 3: What was the average inflation rate during renewals?

What I did:

- Focused only on subscriptions that were renewed
- Checked which financial period each subscription start date fell into
- Collected the matching inflation rate for each and averaged them

Result: The average inflation rate during renewals was about **4.44%**.



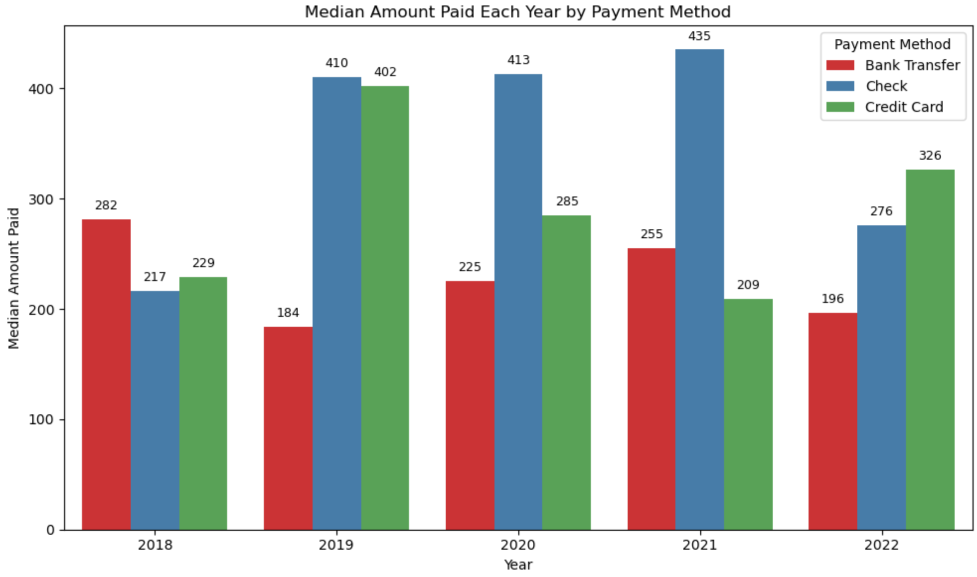
```
Average inflation rate during renewals: 4.44 %
    client_id  start_date  inflation_rate
1  4309371709  2021-05-24            0.76
2  3183675157  2021-12-25            7.32
3  5371694837  2020-03-14            4.40
5  7896208406  2022-02-24            6.76
6  4687291312  2019-06-14            3.84
```

Question 4: What is the median amount paid each year for all payment methods?

Approach:

- Extracted the year from each payment date
- Grouped the data by year and payment method
- Calculated the median payment in each group
- Visualized it with a grouped bar chart showing labels for better readability

Result: Median payments varied year to year. In most years, payments made via Check had the highest medians — ₹435 in 2021, for example.



Declaration

I hereby declare that this assignment has been completed by me to the best of my knowledge and ability. No unfair means have been used, and all the work presented here is my own.