Exploratory Data Analysis (EDA)

The **Exploratory Data Analysis (EDA)** phase focuses on understanding the key characteristics of the loan dataset, detecting any patterns, and identifying potential correlations between the variables. The objective is to provide a solid foundation for further analysis by summarizing key statistics, visualizing trends, and uncovering underlying relationships.

1. Loan portfolio analysis

a. Distribution of Loan Status

The distribution of loan status is crucial for understanding the proportion of loans that are approved or rejected. By analysing loan status, we can gain insights into the performance of the loan portfolio and assess any potential biases in loan approval.

SQL query:

```
SELECT loan_status, COUNT(*) as number
FROM loan_dataset
GROUP BY loan_status;
```

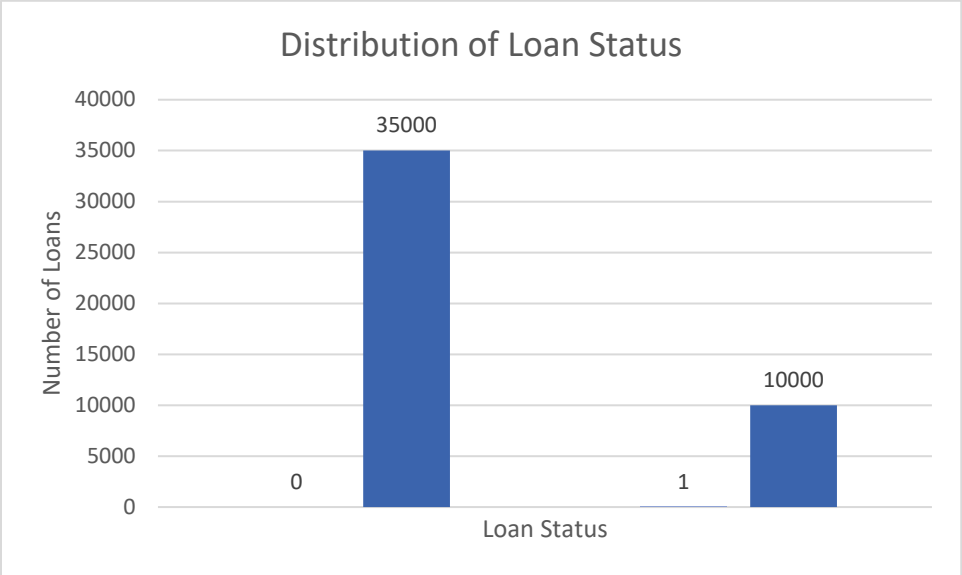
Figure 1 : Distribution of Loan Status

Result:

loan_status	number
0	35000
1	10000

Figure 2 : Distribution of Loan Status Result

Illustration:



Interpretation:

We now have the number of loans in each status category providing a clear picture of loan performance.

b. Distribution of Loan Amounts

The distribution of loan amounts is essential to understand how loans are structured in terms of their size. Understanding the distribution of loan amounts through statistical measures like **mean, median and standard deviation** provides deeper insights into how HSBC structures its loans.

SQL query:

```
WITH LoanStats AS (
    SELECT
        loan_amnt,
        PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY loan_amnt)
            OVER () AS median_loan_amount
    FROM loan_dataset
)
SELECT
    AVG(loan_amnt) AS mean_loan_amount,
    MAX(median_loan_amount) AS median_loan_amount,
    STDEV(loan_amnt) AS stdev_loan_amount
FROM LoanStats;
```

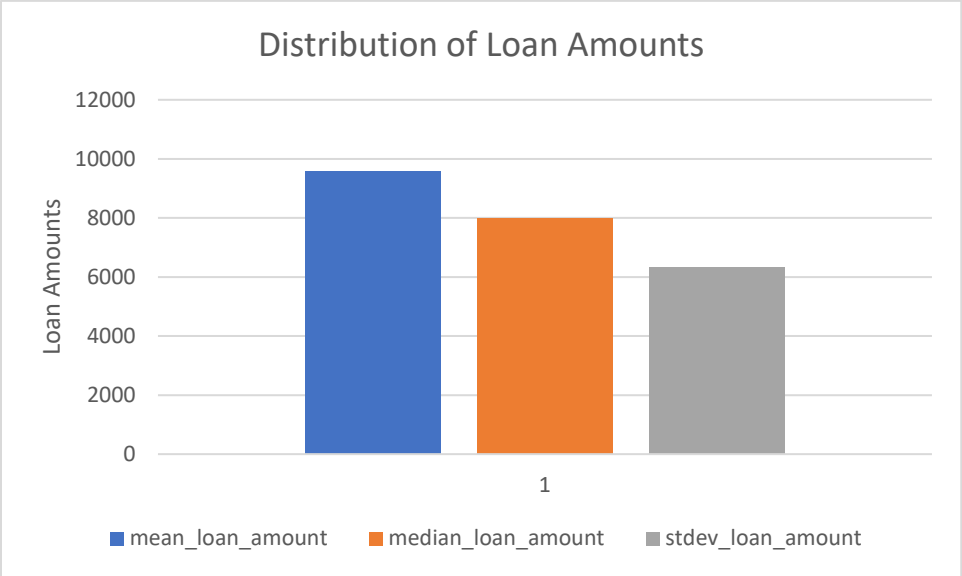
Figure 3 : Distribution of Loan Amounts

Result:

mean_loan_amount	median_loan_amount	stdev_loan_amount
9583	8000	6314,886691

Figure 4 : Distribution of Loan Amounts Result

Illustration:



Interpretation:

We now have the **mean, median, and standard deviation** of loan amounts, providing a statistical overview of loan distribution. This helps us understand typical loan sizes, detect potential outliers, and optimize lending strategies to mitigate risk.

c. Average Interest Rate Applied

The **average interest rate** is crucial for HSBC, as it helps in:

- Evaluating the **profitability of the loan portfolio**.
- Identifying trends in **interest rate policies**.
- Assessing the **impact of interest rates on defaults and loan demand**.

SQL query:

```
SELECT
    AVG(loan_int_rate) AS avg_interest_rate
FROM loan_dataset;
```

Figure 5 : Average Interest Rate Applied

Result:

avg_interest_rate
11,00660578

Figure 6 : Average Interest Rate Applied Result

Interpretation:

This result indicates that the **average interest rate applied across all loans is 11%**. HSBC can use this information to compare its rates with competitors, adjust pricing strategies, and optimize risk management policies.

2. Borrower Profile

a. Loan Distribution by Borrower Education

Understanding the **distribution of loans by borrower education** helps us:

- Identify the **most common borrower profiles**.
- Assess **risk levels** associated with different borrower types.
- Optimize **loan policies** and **targeted marketing strategies** for different customer segments.

SQL query:

```
SELECT
    person_education,
    COUNT(*) AS loan_number
FROM loan_dataset
GROUP BY person_education
ORDER BY loan_number DESC;
```

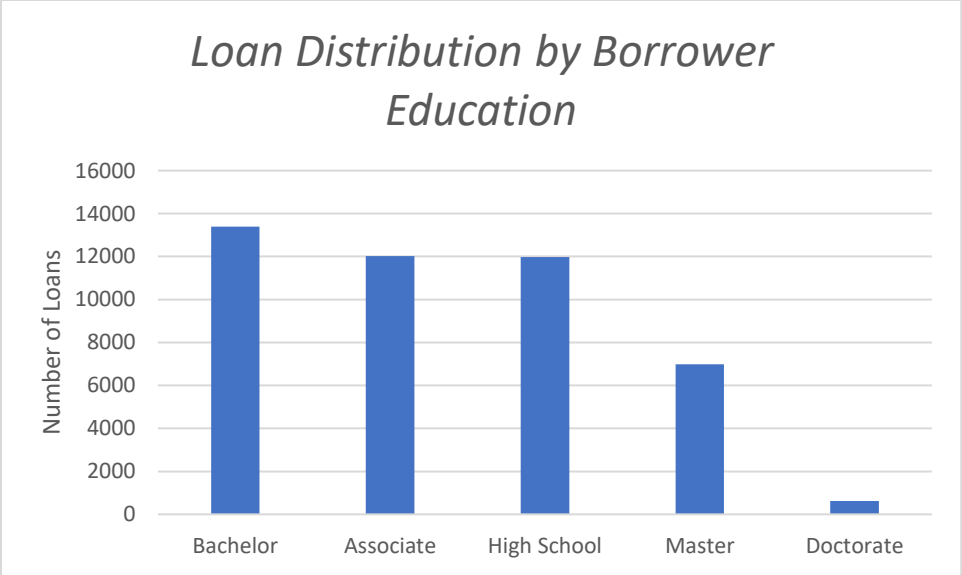
Figure 7 : Loan Distribution by Borrower Education

Result :

person_education	loan_number
Bachelor	13399
Associate	12028
High School	11972
Master	6980
Doctorate	621

Figure 8 : Loan Distribution by Borrower Education Result

Illustration:



Interpretation:

This insight helps HSBC **understand which borrower segments are most active**, assess risk associated with different education levels, and **fine-tune lending criteria** accordingly.

b. Average Income of Borrowers

Understanding the **average income of borrowers** is crucial for us to:

- Assess **borrowers' repayment capacity**.
- Identify **high-income vs. low-income segments** and their loan behaviours.
- Optimize **loan approval policies** and risk assessment models.

SQL query:

```sql
SELECT
    person_education,
    AVG(person_income) AS avg_borrower_income
FROM loan_dataset
GROUP BY person_education
ORDER BY avg_borrower_income DESC;
```

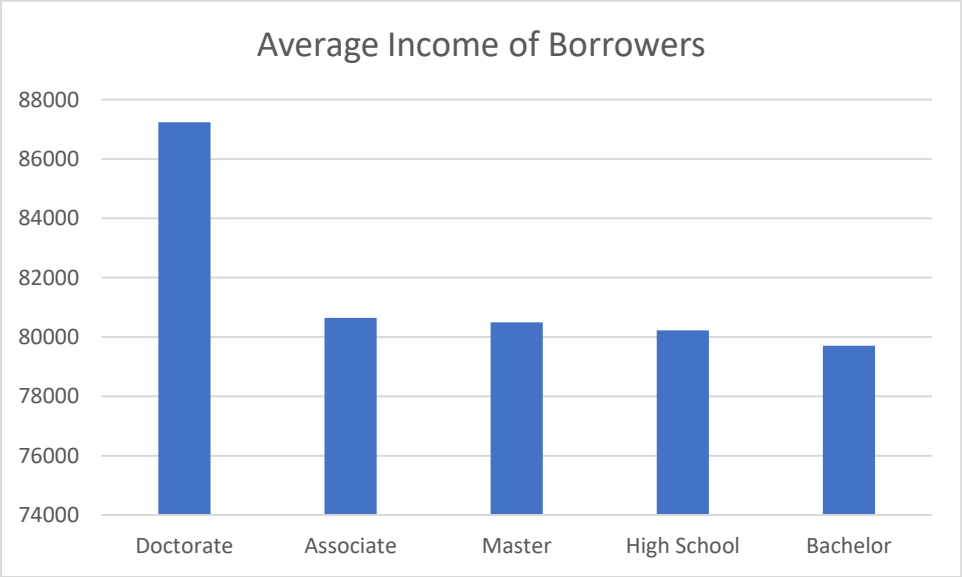
Figure 9 : Average Income of Borrowers

Result:

person_education	avg_borrower_income
Doctorate	87234
Associate	80641
Master	80491
High School	80224
Bachelor	79703

Figure 10 : Average Income of Borrowers Result

Illustration:



Interpretation:

This result indicates that the **average borrower income is approximatively £80,000**. HSBC can use this data to develop **income-based loan offerings** and ensure that loans are granted to borrowers with sufficient repayment capacity.

c. Correlation Between Income and Loan Amount

Understanding the **relationship between borrower income and loan amount** helps us to:

- Determine if higher-income individuals tend to borrow more.
- Assess whether income is a strong predictor of loan size.
- Improve credit risk models and loan approval strategies.

SQL query:

```
SELECT
    (SUM(CAST(person_income AS DECIMAL(18,2)) * CAST(loan_amnt AS DECIMAL(18,2))) -
     COUNT(*) * AVG(CAST(person_income AS DECIMAL(18,2))) * AVG(CAST(loan_amnt AS DECIMAL(18,2)))) /
    (SQRT((SUM(CAST(person_income AS DECIMAL(18,2)) * CAST(person_income AS DECIMAL(18,2))) -
           COUNT(*) * POWER(AVG(CAST(person_income AS DECIMAL(18,2))), 2)) *
          (SUM(CAST(loan_amnt AS DECIMAL(18,2)) * CAST(loan_amnt AS DECIMAL(18,2))) -
           COUNT(*) * POWER(AVG(CAST(loan_amnt AS DECIMAL(18,2))), 2))))
    AS correlation_income_loan
FROM loan_dataset;
```

Figure 11 : Correlation Between Income and Loan Amount

Result:

correlation_income_loan
0,242290131

Figure 12 : Correlation Between Income and Loan Amount Result

Interpretation:

The correlation coefficient is 0.242 so there is a weak positive correlation between income and loan amount. This indicates that there is a slight tendency for higher-income individuals to borrow more, but the relationship is **not strong**.

d. Distribution of Socio-Economic Profiles Among Borrowers

The distribution of socio-economic profiles based on age, profession, and location is crucial for understanding the key characteristics of borrowers in HSBC loan portfolio. By analysing these profiles, we can gain insights into the most common borrower groups and potentially target high-performing segments for future loan offerings.

SQL query:

```sql
SELECT
    CASE
        WHEN person_age BETWEEN 20 AND 35 THEN '20-35'
        WHEN person_age BETWEEN 36 AND 49 THEN '36-49'
        WHEN person_age >= 50 THEN '50+'
        ELSE 'Unknown'
    END AS age_group,
    person_home_ownership, AVG(person_income) AS avg_income, COUNT(*) AS borrower_number
FROM loan_dataset
GROUP BY
    CASE
        WHEN person_age BETWEEN 20 AND 35 THEN '20-35'
        WHEN person_age BETWEEN 36 AND 49 THEN '36-49'
        WHEN person_age >= 50 THEN '50+'
        ELSE 'Unknown'
    END,
    person_home_ownership
ORDER BY borrower_number DESC;
```

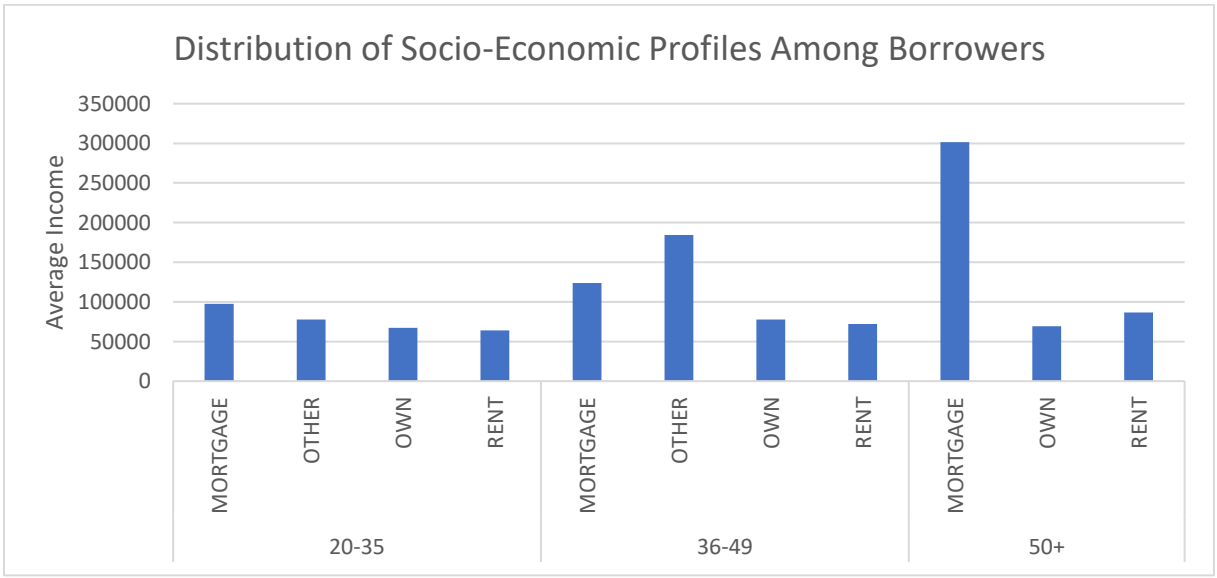
Figure 13 : Distribution of Socio-Economic Profiles Among Borrowers

Result:

age_group	person_home_ownership	avg_income	borrower_number
20-35	RENT	64103	21219
20-35	MORTGAGE	97565	16570
20-35	OWN	67315	2623
36-49	RENT	72061	1997
36-49	MORTGAGE	123823	1786
36-49	OWN	77909	304
50+	RENT	86754	227
50+	MORTGAGE	301514	133
20-35	OTHER	77851	102
50+	OWN	69373	24
36-49	OTHER	184156	15

Figure 14 : Distribution of Socio-Economic Profiles Among Borrowers Result

Illustration:



Interpretation:

- **Young Professionals** have moderate income across all property categories, but their mortgage holders have the highest average income, which is expected as mortgages often signify more financial stability.
- **Middle-Aged Professionals** tend to have higher incomes overall, especially among those with mortgages. They have a good balance of homeownership and mortgage holders, showing financial stability and career maturity.
- **Experienced Borrowers** have the highest incomes, especially those with mortgages. This could reflect long career experience, accumulated wealth, or investments in higher-value properties. Renters in this group also show relatively high-income levels, suggesting financial flexibility.

Understanding these profiles allows HSBC to tailor their loan offerings, risk assessments, and marketing strategies to the most common borrower segments. It also provides insight into the socio-economic diversity of HSBC loan portfolio.

3. Loan performance

a. Types of Loans with the Best Interest Rates

Understanding which types of loans carry the best interest rates is crucial for HSBC to evaluate its pricing strategy. By identifying the loan types with the lowest interest rates, the bank can ensure that its pricing is competitive, while also optimizing profitability for different types of borrowers.

SQL query:

```
SELECT
    loan_intent,
    AVG(loan_int_rate) AS average_interest_rate
FROM loan_dataset
GROUP BY loan_intent
ORDER BY average_interest_rate ASC;
```

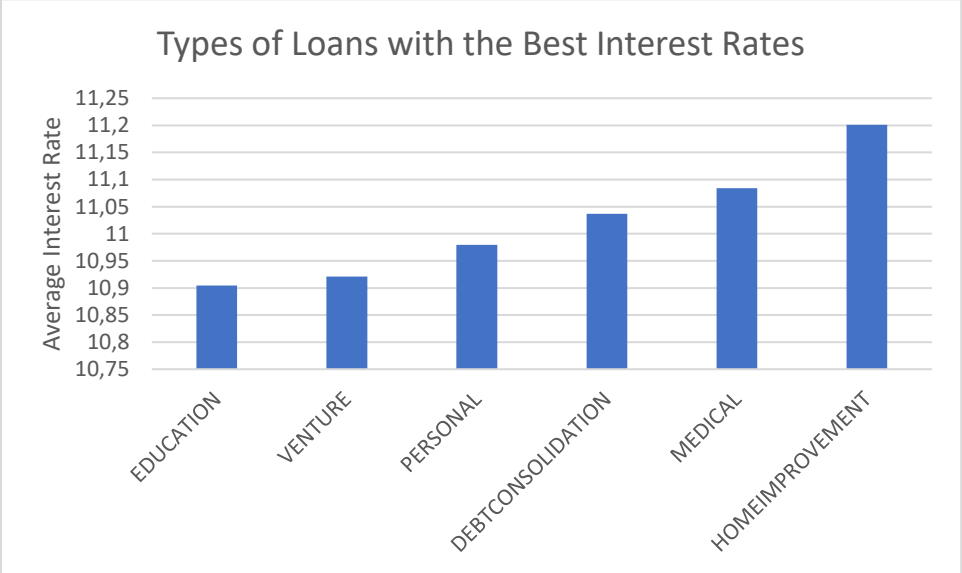
Figure 15 : Types of Loans with the Best Interest Rates

Result:

loan_intent	average_interest_rate
EDUCATION	10,90443461
VENTURE	10,9210142
PERSONAL	10,97973252
DEBTCONSOLIDATION	11,03663961
MEDICAL	11,0841694
HOMEIMPROVEMENT	11,2009931

Figure 16 : Types of Loans with the Best Interest Rates Result

Illustration:



Interpretation:

By analysing the results, we can see that the average interest rate is approximatively the same for the different types of loans.

b. Evolution of Interest Rates Based on Loan Amounts

Understanding how interest rates evolve based on the amount borrowed can help HSBC optimize its lending strategy. By analysing the relationship between loan amounts and interest rates, the bank can better assess whether larger loans are associated with higher or lower interest rates, and adjust its pricing strategies accordingly to attract specific borrower profiles.

SQL query:

```sql
SELECT
    CASE
        WHEN loan_amnt <= 5000 THEN 'Up to 5000'
        WHEN loan_amnt > 5000 AND loan_amnt <= 10000 THEN '5001 to 10000'
        WHEN loan_amnt > 10000 AND loan_amnt <= 20000 THEN '10001 to 20000'
        WHEN loan_amnt > 20000 AND loan_amnt <= 50000 THEN '20001 to 50000'
        WHEN loan_amnt > 50000 THEN 'Above 50000'
    END AS loan_amount_range,
    AVG(loan_int_rate) AS average_interest_rate
FROM loan_dataset
GROUP BY
    CASE
        WHEN loan_amnt <= 5000 THEN 'Up to 5000'
        WHEN loan_amnt > 5000 AND loan_amnt <= 10000 THEN '5001 to 10000'
        WHEN loan_amnt > 10000 AND loan_amnt <= 20000 THEN '10001 to 20000'
        WHEN loan_amnt > 20000 AND loan_amnt <= 50000 THEN '20001 to 50000'
        WHEN loan_amnt > 50000 THEN 'Above 50000'
    END
ORDER BY loan_amount_range;
```

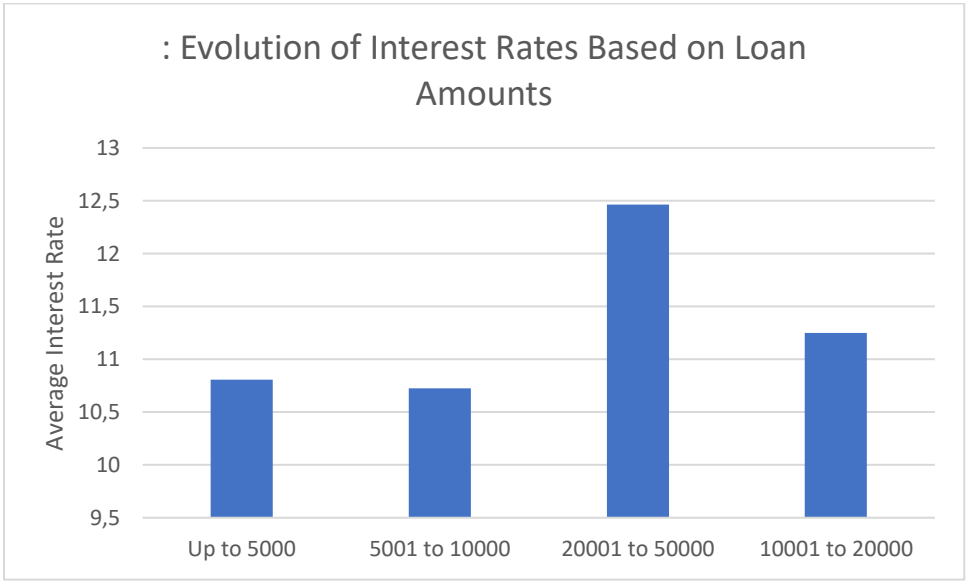
Figure 17 : Evolution of Interest Rates Based on Loan Amounts

Result :

loan_amount_range	average_interest_rate
10001 to 20000	11,24872096
20001 to 50000	12,46299131
5001 to 10000	10,72370319
Up to 5000	10,80642906

Figure 18 : Evolution of Interest Rates Based on Loan Amounts Result

Illustration:



Interpretation:

From the results, we can observe:

- Relatively stable rates for small loans: Interest rates are fairly similar for loans up to $10,000.
- Significant increase for larger loans: There is a notable increase in the interest rate for loans over $10,000, and even more so for those over $20,000.

The analysis indicates that **larger loans** tend to have **higher interest rates**, which may be due to the perceived **higher risk** associated with lending larger amounts. Borrowers requesting larger loans may be seen as riskier, or the bank may charge higher rates to offset the potential risk of non-repayment.

c. Most Profitable Customer Segments

Identifying the most profitable customer segments is essential for HSBC to refine its marketing and lending strategies. By understanding which borrower profiles bring in the most revenue, the bank can tailor its offerings and optimize its efforts to target high-value customers, while maintaining a balanced risk profile.

SQL query:

```
SELECT
    CASE
        WHEN person_age BETWEEN 20 AND 35 THEN '20-35'
        WHEN person_age BETWEEN 36 AND 49 THEN '36-49'
        WHEN person_age >= 50 THEN '50+'
        ELSE 'Unknown'
    END AS age_group,
    person_income,
    person_home_ownership,
    SUM(loan_amnt * loan_int_rate / 100) AS total_interest_revenue
FROM loan_dataset
GROUP BY CASE
        WHEN person_age BETWEEN 20 AND 35 THEN '20-35'
        WHEN person_age BETWEEN 36 AND 49 THEN '36-49'
        WHEN person_age >= 50 THEN '50+'
        ELSE 'Unknown'
    END, person_income, person_home_ownership
ORDER BY total_interest_revenue DESC;
```

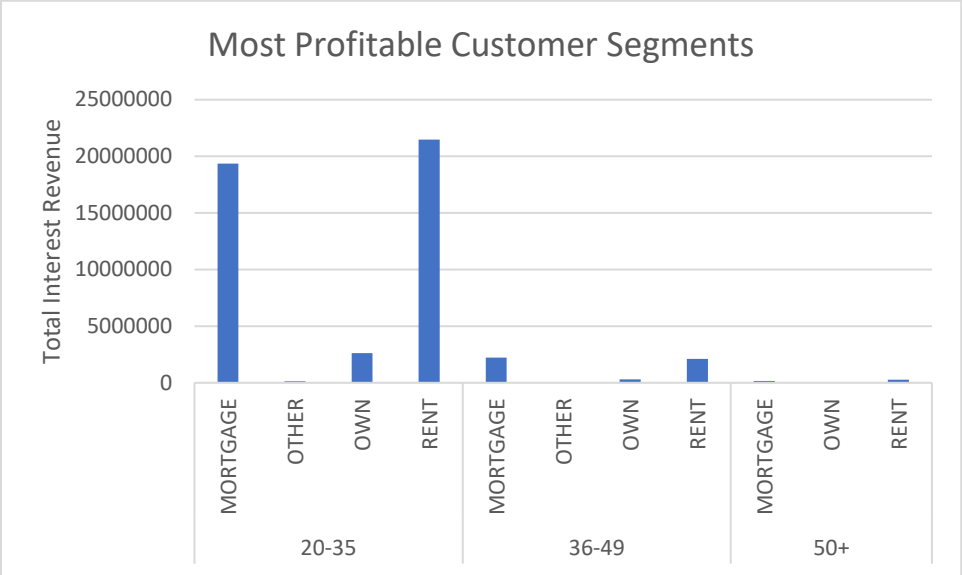
Figure 19 : Most Profitable Customer Segments

Result :

age_group	person_income	person_home_ownership	total_interest_revenue
20-35	90938	RENT	9419,15
20-35	79255	RENT	8189,56
20-35	166588	MORTGAGE	8188
20-35	97019	MORTGAGE	8013,025
20-35	85072	MORTGAGE	7974,54
20-35	47692	RENT	7955,9
20-35	73008	RENT	7824,4475
20-35	48829	RENT	7793,735
20-35	61014	RENT	7733,96

Figure 20 : Most Profitable Customer Segments Result

Illustration:



Interpretation:

From the results, we can see that the most profitable customer segments are typically **young professionals (20-35 years old)** customers with **rent and mortgage ownership**.

For HSBC, targeting young professionals' borrowers with rent or mortgage ownership could lead to higher profitability. Offering competitive rates and targeted marketing to these segments can be a strategic approach to maximize revenue while maintaining a balanced risk profile.

CONCLUSION

The study highlights key trends in borrower behaviour and loan structuring at HSBC. The main challenge lies in optimizing credit criteria and revising pricing policies to maximize profitability while maintaining controlled risk. A more detailed analysis of risk factors would make banking offerings more competitive.