Loan Analysis & Insights for Risk Identification

By - Sandeep Kumar

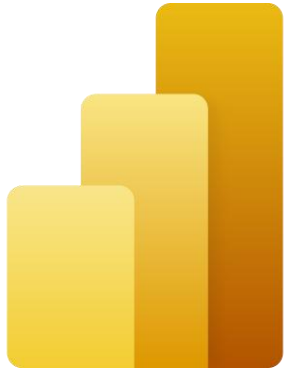
Tools Used:
Python | Pandas | NumPy | Seaborn | SQL | Power BI | Excel

End-to-End Data Analysis Workflow
(Excel → Python → SQL → Power BI)

→

→

→



Explore,understand, validate

Clean, transform, analyse

Store, query, optimise

Visualise, present, deliver insights

Raw Data → Processing & Analysis → Database → Dashboard & Insights

Problem Statement

Banks give loans to many customers, but some customers do not repay, which causes financial loss.

This project aims to:
•Analyze customer and loan-related data
•Identify key factors that influence default
•Detect high-risk borrower profiles
•Provide actionable insights that help minimize losses
•Support better credit decision-making using data

Business Impact
•Improve approval quality
•Reduce risk exposure
•Enhance credit policies using data-backed insights

Excel – Initial Data Understanding

Before moving to Python, the dataset was first explored in Excel to perform basic validation and early insights.

1. Data Import & Structure Check

- Loaded the raw dataset into Excel
- Verified total rows, columns, and header format
- Ensured no misaligned or broken columns

2. Basic Data Quality Checks

- Used **filters** to spot blank cells, inconsistent values
- Checked for **duplicate rows** using *Remove Duplicates*

Excel – Initial Data Understanding

3. Quick Summary Using Excel Tools

•Created **Pivot Tables** to review:

- Loan Status distribution (Good vs Default)
- Loan Purpose counts
- Grade-wise loan distribution
- Loan amount ranges

4. Initial Trend & Pattern Identification

•Created quick charts:

- Loan Amount histogram
- Interest Rate distribution
- Loan Purpose bar chart

5. Export for Python

•Cleaned file saved as **.xlsx / .csv**

•Handed over to Python for advanced cleaning & EDA

Data Summary

Dataset Overview

The dataset contains information about customers, their loan details, and their repayment status. It helps us understand the factors that influence whether a loan is repaid or defaulted.

Dataset Size

- **Total Rows:** 38576
- **Total Columns:** 24

Key Columns Include

- **Customer Information:** age, income, employment type
- **Loan Details:** loan amount, interest rate, loan term
- **Financial Indicators:** credit score, debt-to-income ratio
- **Loan Status:** Good loan / default

Data Cleaning & Preparation

1. Missing Values Handling
- Checked missing values in all columns
- The column **emp_title** had missing values
- Filled missing emp_title with **"Unknown“**

2. Duplicate Records
- Verified for duplicated entries
- **Total duplicate rows: 0**

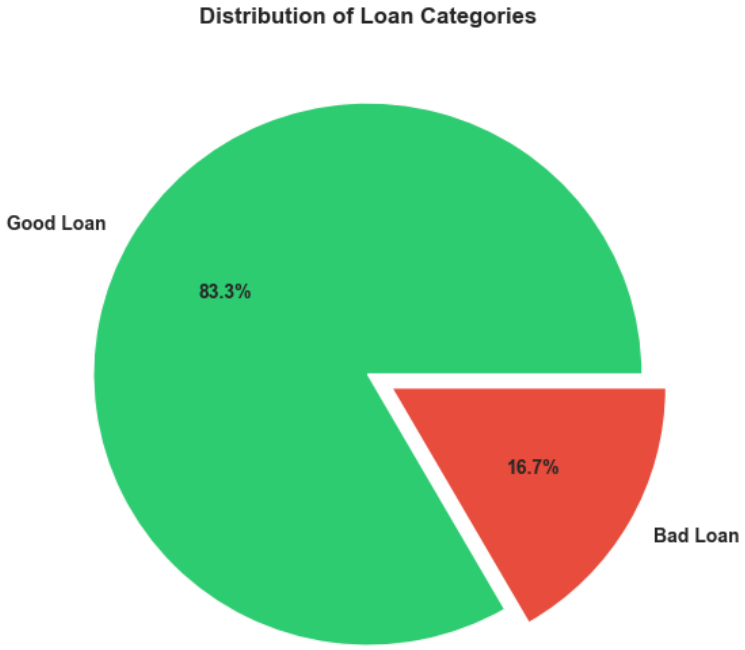
3. Outlier Detection
- Identified outliers using boxplots
- Outliers were **not removed**, kept for analysis to avoid data loss

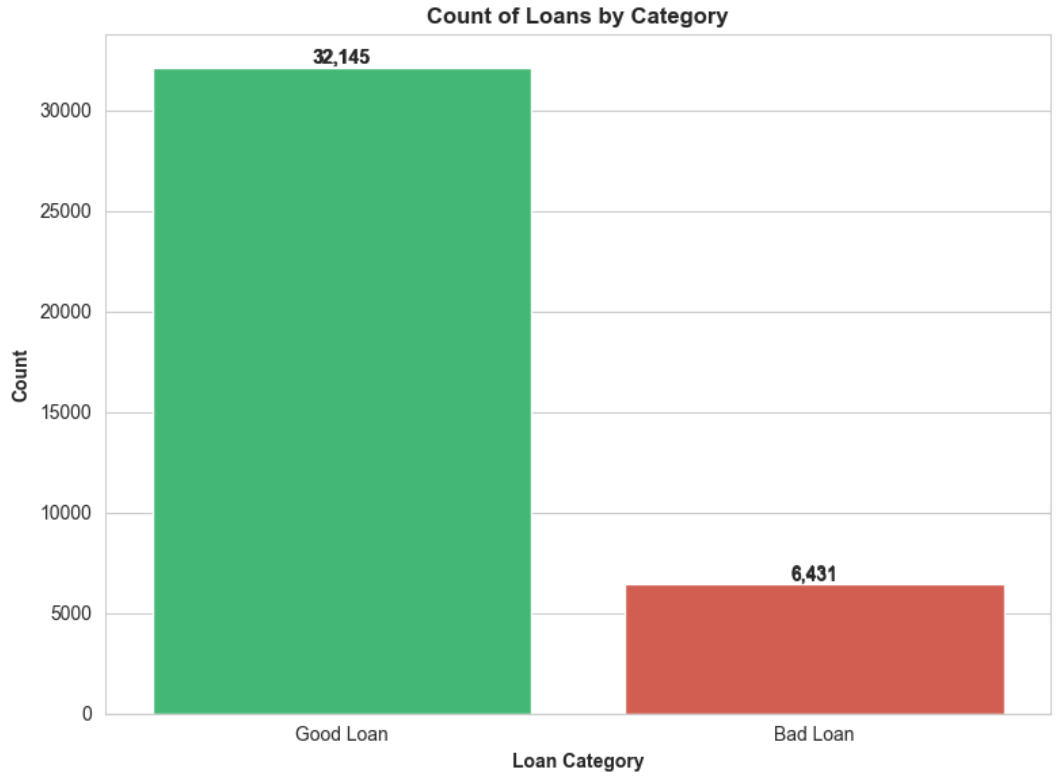
4. Fixing Data Types
- Converted date-related columns into proper datetime format:
    - issue_date
    - last_credit_pull_date
    - last_payment_date
    - next_payment_date

Dependent Variable

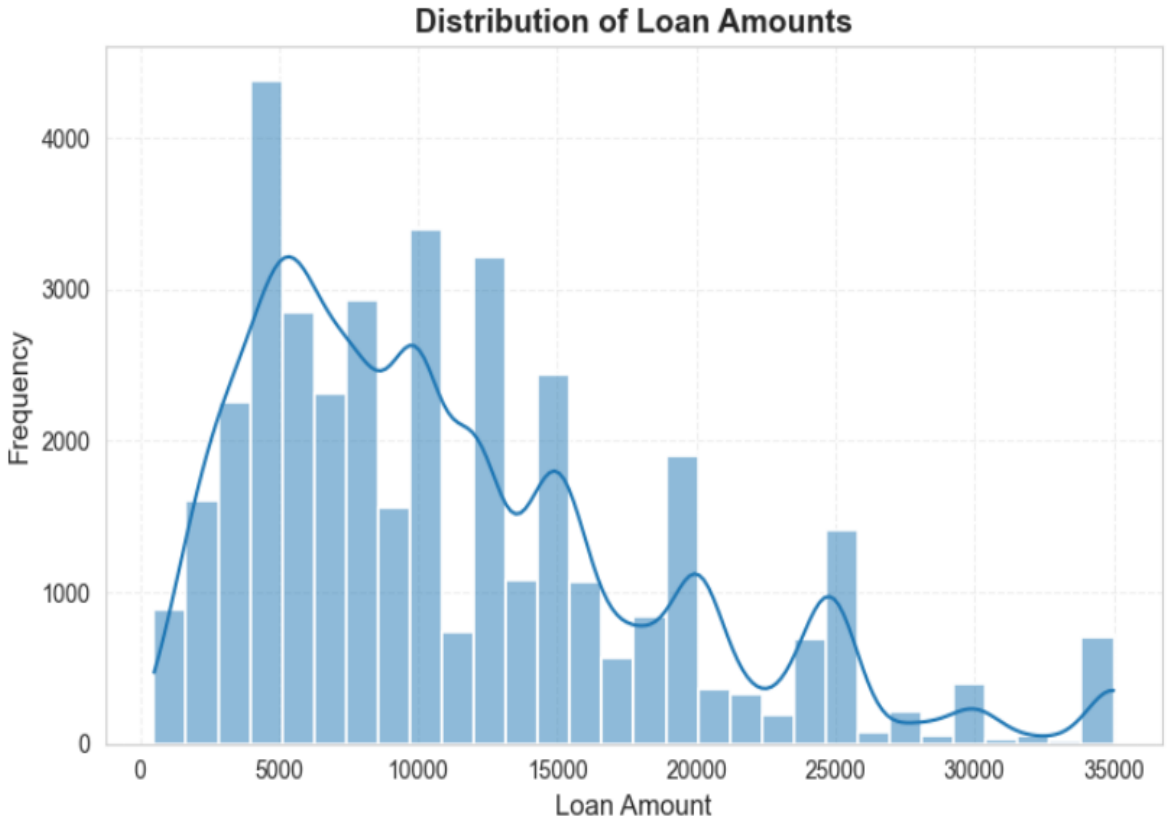
📌 loan_status (Good Loan vs Default Loan)

This column tells you whether a borrower **repaid the loan** or **defaulted**.





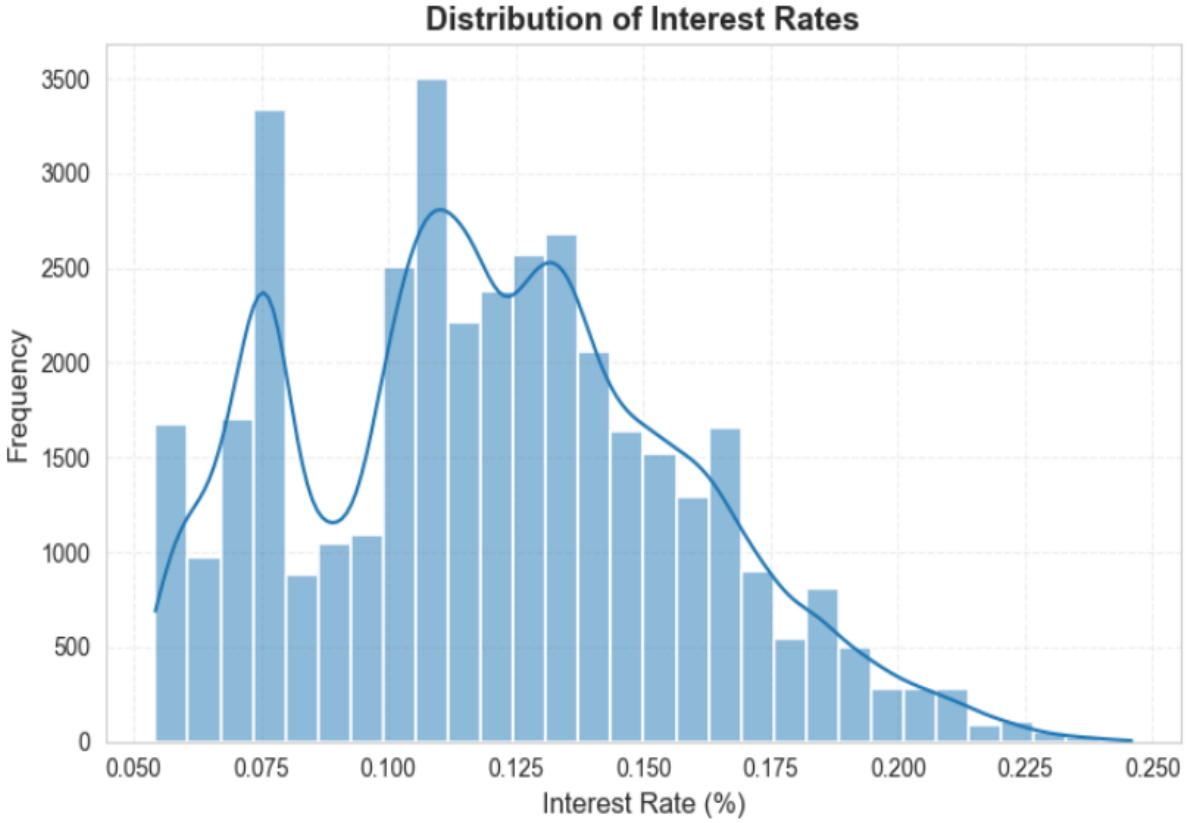
Exploratory Data Analysis (EDA)

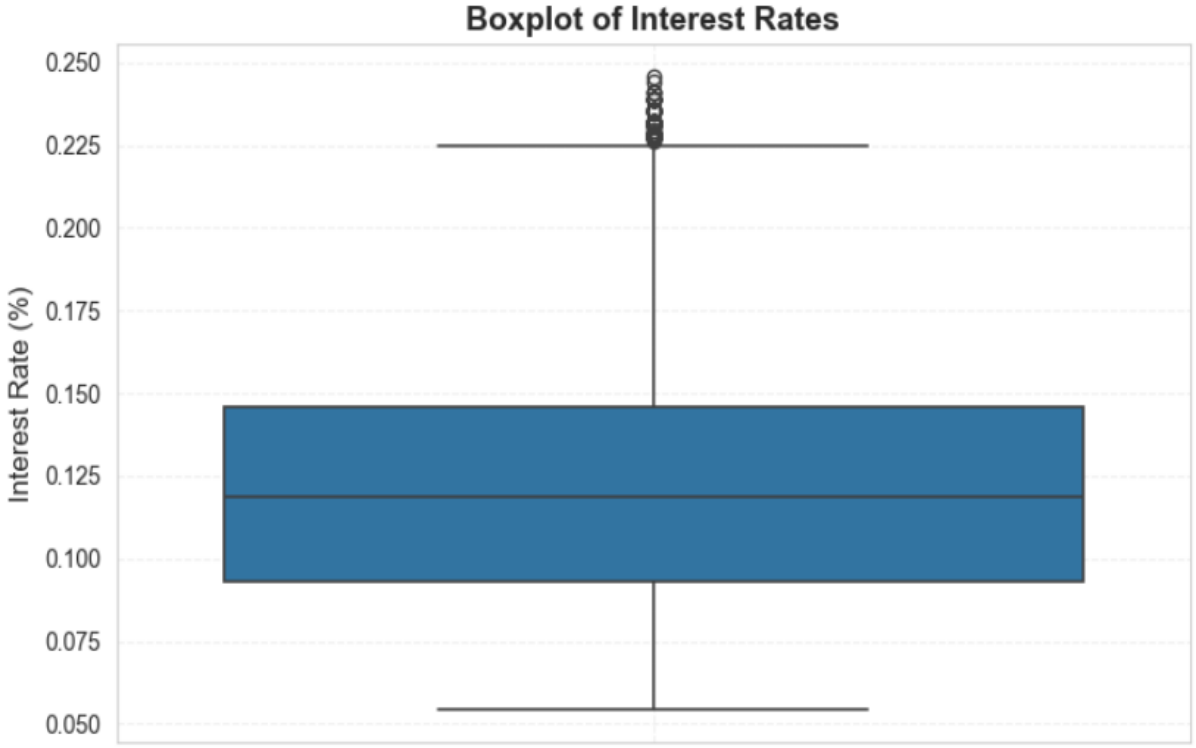


Loan Amount

- Most loan amounts lie between **$5,000 and $15,000**.
- The distribution is **right-skewed**, meaning some customers take very high loan amounts.
- Boxplot shows **clear outliers**, but we did **not remove them** because they represent real high-value loans

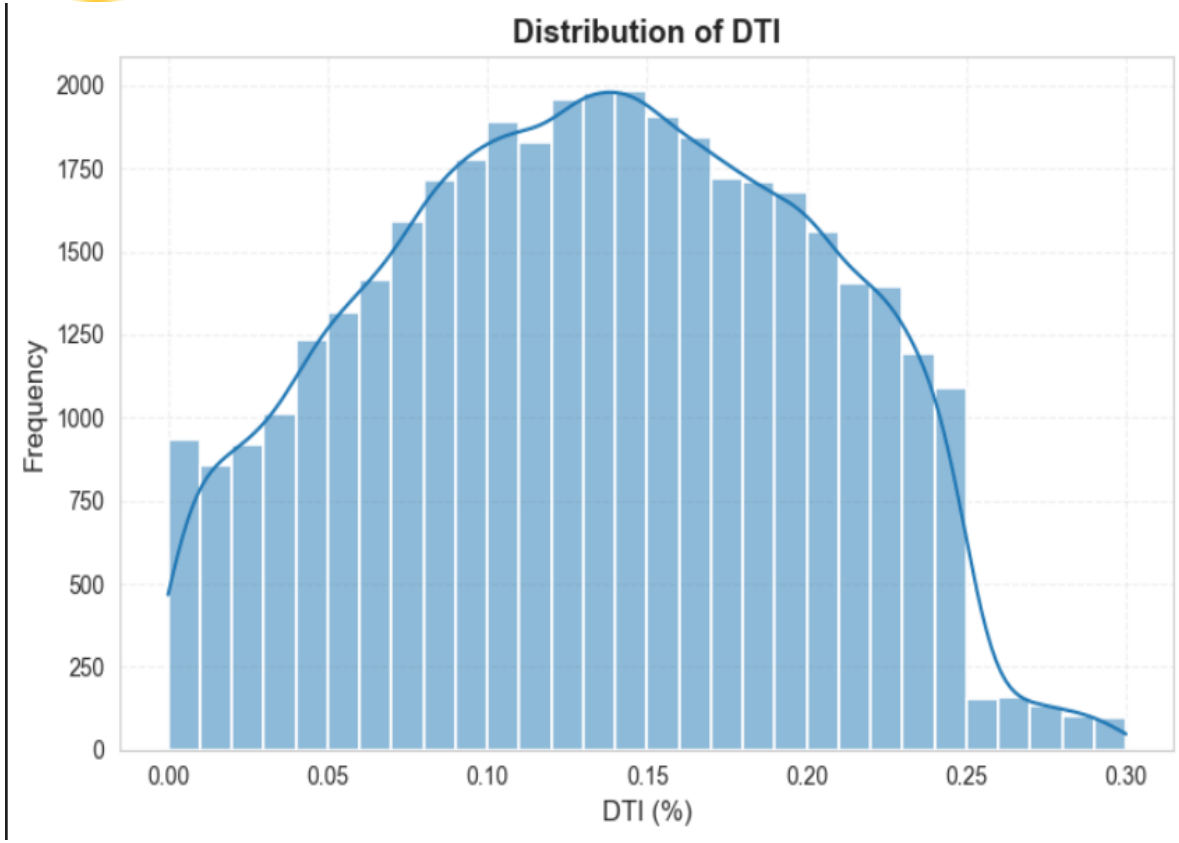
Exploratory Data Analysis (Continued)

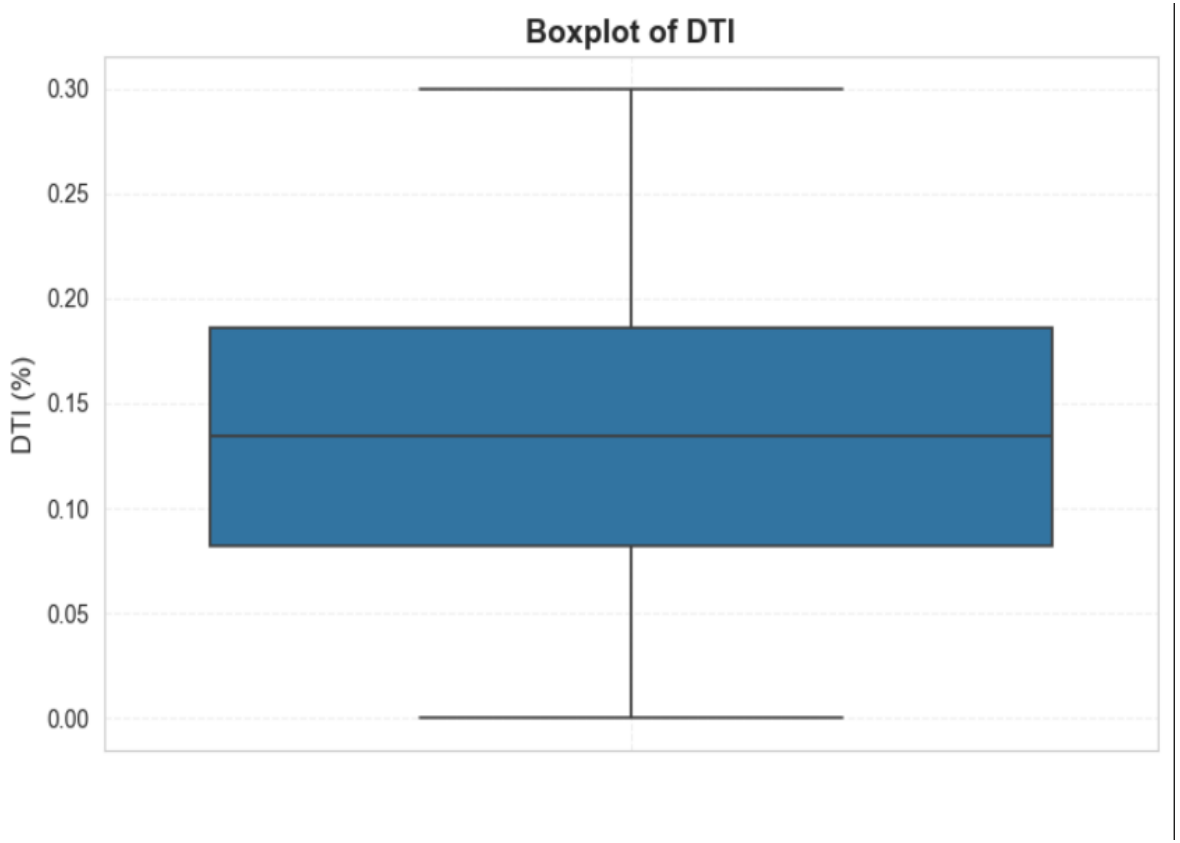




- Interest rates are mostly between **8% and 14%**.
- The distribution is **slightly right-skewed**, with a few customers getting higher interest rates.
- Outliers are present at the upper end (18%–24%).
- These outliers are expected due to risk-based pricing.

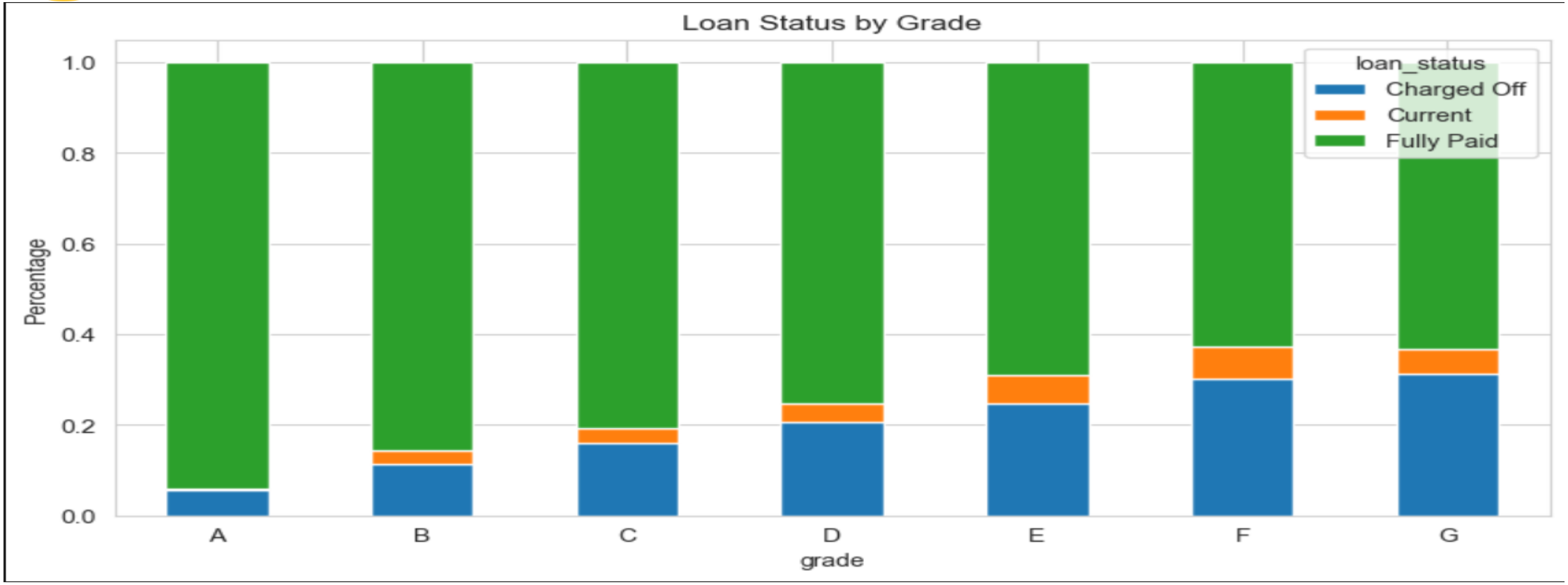
Exploratory Data Analysis (Continued)





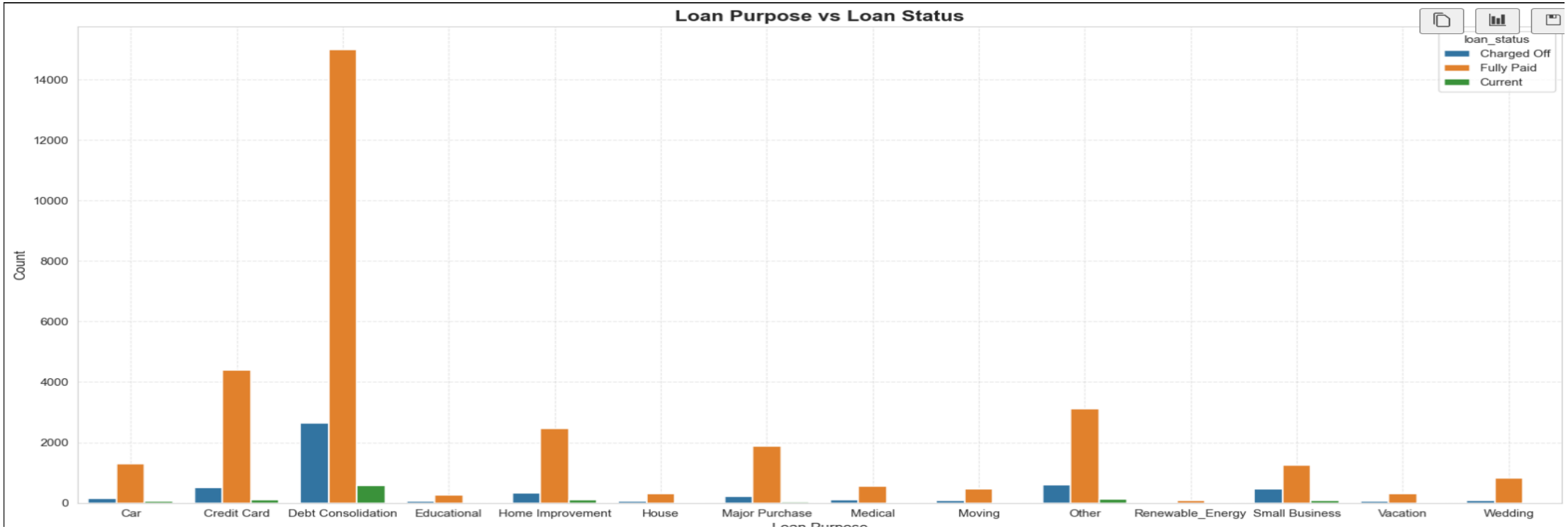
- Most DTI values lie between **8%–18%**.
- Distribution is **slightly right-skewed**.
- **Median DTI ≈ 13%**.
- Few **high DTI cases**, but **no major outliers**

Exploratory Data Analysis (Continued)



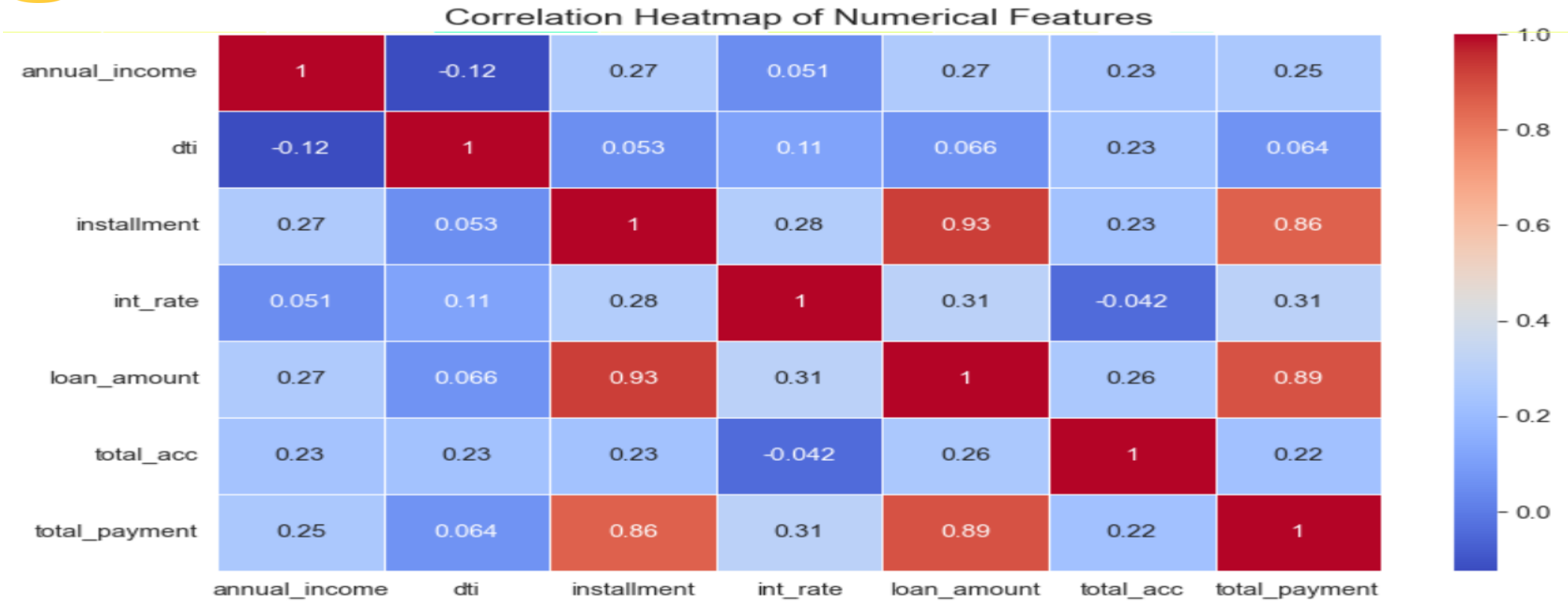
- Higher grades (A, B, C) have **very low Charged-Off rates**.
- Lower grades (D, E, F, G) show **increasing default (Charged-Off) percentages**.
- Most loans across all grades are **Fully Paid**, but risk clearly increases as grade decreases.
- This confirms that **loan grade is strongly linked to credit risk**.

Exploratory Data Analysis (Continued)



- **Debt Consolidation** is the most common loan purpose; it also has a high number of **Charged-Off** loans.
- Credit Card, Home Improvement, and Other categories also show moderate levels of default.
- Very few loans across all categories remain in **Current** status.
- Loan purpose can help identify **high-risk borrower segments**.

Exploratory Data Analysis (Continued)



- **Installment** and **Loan Amount** show **very high correlation**.
- **Total Payment** is strongly correlated with both **Loan Amount** and **Installment**.
- **Annual Income**, **DTI**, and **Interest Rate** have **weak correlations** with other features.
- No strong multicollinearity issues except the **Loan Amount ↔ Installment** pair

Key Insights from EDA

Loan Status

•Dataset is **imbalanced** → ~83% Good Loans, ~17% Bad Loans.

Loan Amount

•Most loan amounts fall between **$5K–$15K**.

•**Higher loan amounts** are more common in **Current** loans.

•Borrowers with **higher loan amounts** are more likely to be **Charged Off**.

Interest Rate

•**Charged-off loans** have **higher interest rates**.

•**Fully Paid** loans have **lower interest rates**.

•Higher interest rates increase the chance of **default**.

Key Insights from EDA

DTI (Debt-to-Income Ratio)

- **Charged-off borrowers** have **higher DTI** than Fully Paid borrowers.
- Higher DTI increases the likelihood of **loan default**.

Loan Grade

- Lower grades (**E, F, G**) show **higher default rates**.
- Higher grades (**A, B, C**) indicate **lower risk**.

Loan Purpose

- **Debt Consolidation** is the largest category and has the **most defaults**.
- Other risky purposes: **Credit Card**, **Home Improvement**, **Other**.

Correlation Insights

- **Loan Amount ↔ Installment** show **very high correlation**.
- All other numerical features show **weak correlation → no multicollinearity issues**.

Feature Engineering

1. Date Feature Extraction
•Extracted **issue_year** and **issue_month** from the loan issue date
→ helps analyze trends over time.

2. Income Bucketing
•Grouped annual income into **Low**, **Medium**, **High**, and **Very High** categories
→ simplifies analysis and improves model interpretability.

3. Income-to-Loan Ratio
•Created a new feature: **income_to_loan_ratio = annual_income / loan_amount**
→ captures borrower repayment capacity more effectively.

4. DTI Bucketing
•Converted continuous DTI values into risk buckets
→ **Low DTI**, **Medium DTI**, **High DTI** groups.

5. Interest Rate Bucketing
•Converted continuous interest rates into labeled buckets such as:
Low (≤10%), **Medium (10–15%)**, **High (>15%)**
•Helps group borrowers based on the cost of borrowing, making it easier to analyze default patterns.

SQL Data Storage & Querying

Why SQL?

After cleaning and preparing the dataset in Python, the final DataFrame was uploaded into a SQL database for:

•Efficient querying

•KPI calculations

•Portfolio segmentation

•Data validation

•Power BI connection

Database Setup:

•CREATE DATABASE Bank loan data;
•USE Bank loan data;

SQL Analysis & KPIs

Table Creation:

```
CREATE TABLE loans_cleaned AS
SELECT * FROM loans_python_clean;
```

1.Portfolio Overview:

```
SELECT
  COUNT(*) AS total_loans,
  SUM(CASE WHEN loan_status='Good' THEN 1 END) AS good_loans,
  SUM(CASE WHEN loan_status='Bad' THEN 1 END) AS bad_loans
FROM loans_cleaned;
```

2. Good vs Bad Loan Percentage:

```
SELECT
  loan_status,
  ROUND(100 * COUNT(*) / (SELECT COUNT(*) FROM loans_clean), 2) AS percent
FROM loans_cleaned
GROUP BY loan_status;
```

SQL Analysis & KPIs

3. Grade-wise Risk:

```
SELECT
  grade,
  SUM(CASE WHEN loan_status='Bad' THEN 1 END) AS bad_loans
FROM loans_cleaned
GROUP BY grade;
```

4. Interest Rate Trend Using Window Function:

```
SELECT
    issue_date,
    COUNT(*) AS loans_issued,
    SUM(COUNT(*)) OVER (ORDER BY issue_date) AS cumulative_loans
FROM loans_clean
GROUP BY issue_date;
```

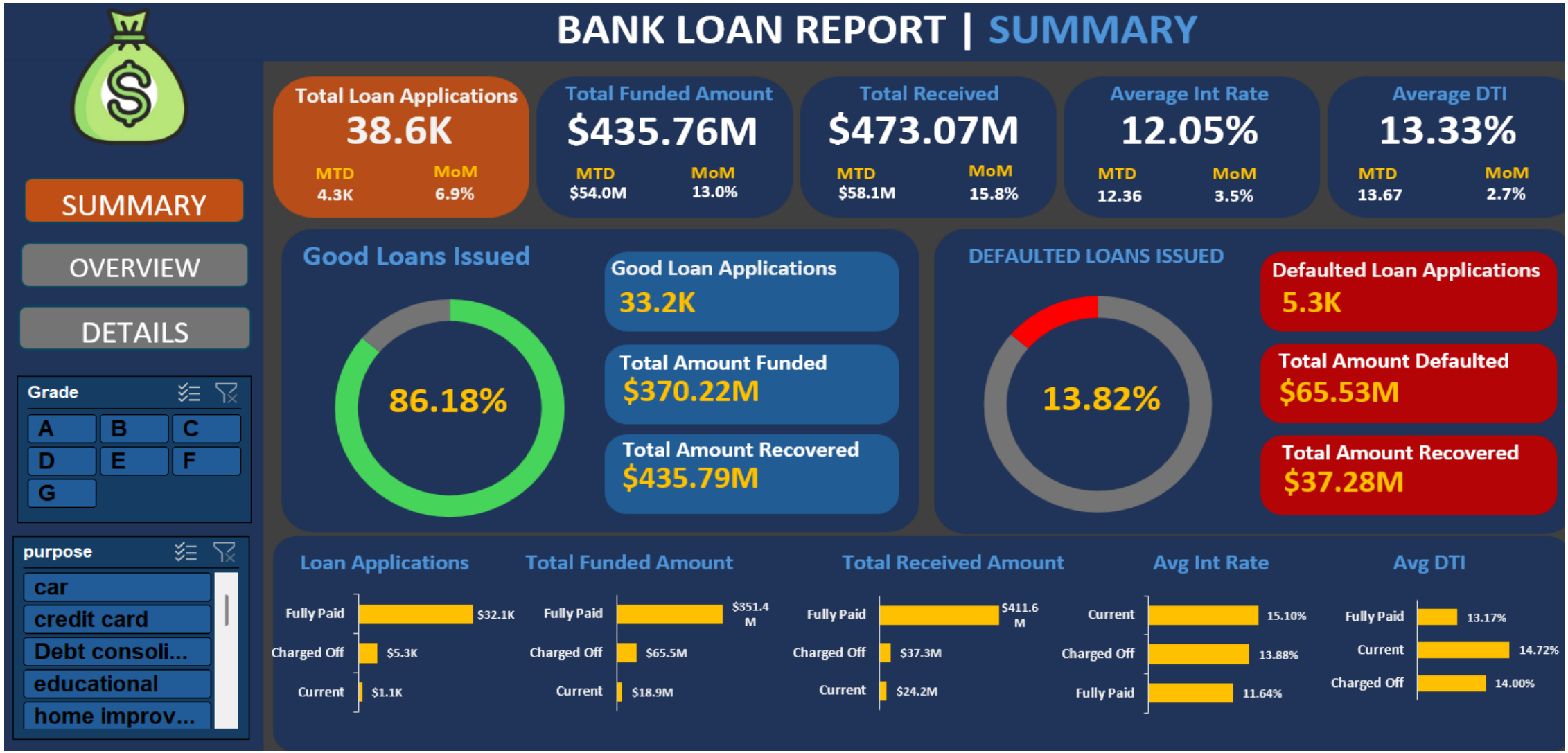
SQL Analysis & KPIs

5. CTE for High-Risk Segments:

```
WITH risky AS (
  SELECT *
  FROM loans_clean
  WHERE dti > 20 AND interest_rate > 15
)
SELECT COUNT(*) AS high_risk_borrowers
FROM risky;
```

“The cleaned SQL table served as the data source for Power BI, enabling real-time updates, dynamic filtering, and smooth KPI dashboards.”

Visual Insights from Dashboards



Visual Insights from Dashboards (Continued)

Insights from Summary Dashboard

1. High Loan Success Rate
•**86.18% of all loans issued are Good Loans**, indicating strong repayment behavior.

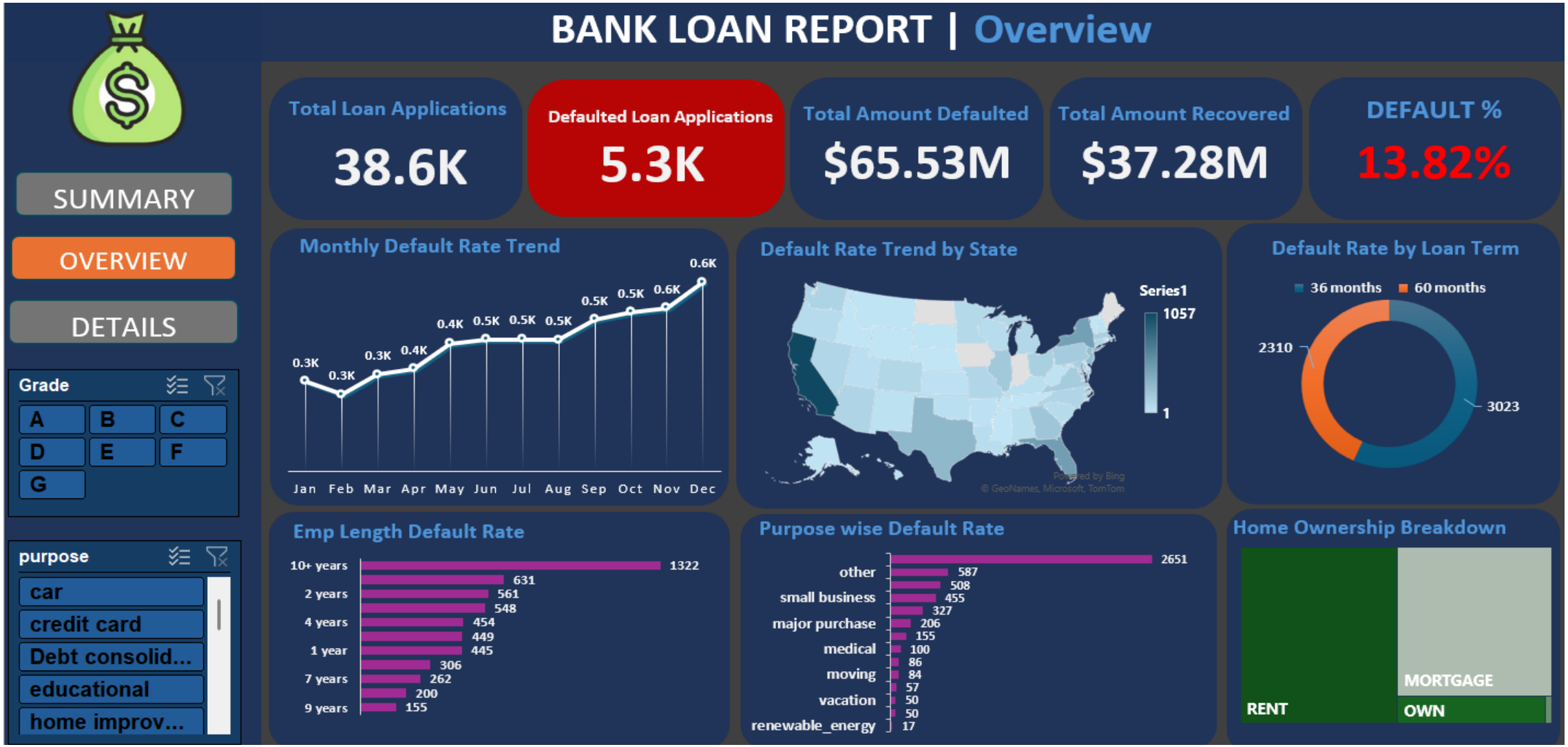
2. Default Level is Moderate but Significant
•**13.82% loans are in default**, resulting in **$65.53M** total defaulted amount.

3. Good Loans Drive Most of the Revenue
•Good loans recovered **$435.79M**, far higher than other segments.

4. Interest Rate Difference Explains Risk
•Charged-off loans have **higher interest rates** (13.88%+) compared to Fully Paid (11.64%).

5. Loan Purpose & Grade Affect Risk Profile
•Filters (Grade & Purpose) allow identifying high-risk segments like **lower grades** and **Debt Consolidation** loans.

Visual Insights from Dashboards (Continued)



Visual Insights from Dashboards (Continued)

Insights from Overview Dashboard

1. Default Rate is Gradually Increasing
•Monthly trend shows defaults rising from **0.3K in Jan to 0.6K in Dec** → consistent upward risk pattern.

2. State-wise Defaults Vary Significantly
•Some states show **very high default concentration**, while others remain low → geographic risk differences.

3. Longer Loan Terms Have Higher Default Rates
•**60-month loans** default more than **36-month loans** → longer commitments increase risk.

4. Certain Loan Purposes Are High Risk
•Highest defaults come from **Other**, **Small Business**, and **Debt Consolidation** categories.

5. Employment Length Doesn't Guarantee Low Risk
•Even borrowers with **10+ years experience** show high defaults → stability doesn't always mean safe borrowing.

Final Recommendations

- **Tighten loan approval rules** for high-risk segments (low grades: E, F, G).
- **Increase interest rates** or add stricter checks for borrowers with **high DTI** or **high loan amounts**.
- **Strengthen verification** for high-risk loan purposes (Debt Consolidation, Small Business, Other)
- **Promote shorter loan terms** (36-month) to reduce default probability.
- **Use automated alerts** for risky profiles (high DTI, high interest rate, long loan term).

Conclusion

- This project analyzed **loan applications, borrower profiles, and repayment behavior** using Python and Power BI.

- Clear patterns emerged: higher **DTI**, higher **interest rates**, lower **grades**, and certain **loan purposes** increase default risk.

- Insights help the business **improve approval policies, identify high-risk customers, and reduce default losses**.

- Overall, the analysis supports **smarter credit decisions** and better **portfolio risk management**.