

LOAN DATASET ANALYSIS

Indhumathi S
COURSE – DATA ANALYTICS
Batch – OCT-2025

DATA PREPARATION

The loan dataset was generated using Mockaroo for analytical practice. It consists of 2000 records and 10 attributes representing customer demographic, financial, and loan-related information. The dataset is well-structured and designed to simulate real-world loan data scenarios. It is suitable for analysis using Excel, SQL, and Power BI tools. The primary objective of data preparation was to ensure the dataset could support data cleaning, statistical analysis, SQL processing, and visualization. Proper preparation helped establish a strong foundation for accurate and meaningful insights.

Field Name	Type	Options
Customer_id	Row Number	blank: 0 % Σ X
Customer_name	First Name	blank: 0 % Σ X
Gender	Last Name	blank: 0 % Σ X
Age	Number	min: 18 max: 69 decimals: 0 blank: 0 % Σ X
Martial_Status	Custom List	Single, Married, Divorced  random  blank: 0 % Σ X
City	City	blank: 0 % Σ X
Country	Country	restrict countries...  blank: 0 % Σ X
Occupation	Job Title	blank: 0 % Σ X
Annual_Income	Number	min: 300000 max: 2000000 decimals: 0 blank: 0 % Σ X

DATA CLEANING

The dataset was thoroughly reviewed to identify missing, inconsistent, and duplicate values. Irrelevant records were removed to maintain data quality. Data types were standardized across numeric, text, and date fields to ensure uniformity. Logical consistency checks were applied to key financial columns such as income, loan amount, and credit score. Derived columns were created using SQL to support advanced analysis. These steps ensured that the dataset was reliable and ready for accurate analytical processing.

This screenshot shows a Microsoft Excel spreadsheet with data in columns A through I. Column A contains customer IDs and names, while columns B through I contain various demographic and professional details. A data validation dialog box is open in the center, displaying the message "No duplicate values found." with an "OK" button.

Customer_id	Full_Name	Gender	Age	Marital_Status	City	Country	Occupation	Annual_Income_Lo
1	Pyotr Maskrey	Male	35	Divorced	Yugdan	China	Environmental Specialist	496079 Pe
2	Zachary Lang	Female	40	Married	Almamah	Kuwait	Food Changer	716156 Ed
3	Astin Lewin	Male	60	Single	Ciaco de Vila	Portugal	Desktop Support Technician	1710354 Ed
4	Alida Goosdie	Female	64	Single	Waigete	Indonesia	Marketing Assistant	1789186 Ed
5	Haywood Baptist	Male	27	Married	Reinaldes		Account Coordinator	854601 Ho
6	Yard Stuer	Male	56	Single	Läupai		Accountant II	1149985 Pe
7	Lilli MacHostie	Female	25	Divorced	Asaita		Senior Developer	381180 Au
8	Kallia Norrington	Female	18	Single	Sheksna		Recruiter	1920330 Pe
9	Troy Tech	Female	41	Single	Minstanteng		Human Resources Assistant I	1009500 Pe
10	Vivienne Broschek	Female	65	Divorced	Korikova		Assistant Professor	157639 Pe
11	Jed Garell	Male	33	Single	Ostiek	Poland	Sales Associate	1463060 Pe
12	Penelope Woodrooffe	Female	68	Single	Anxi	China	Administrative Assistant I	1869904 Ho
13	Etheline Spinio	Female	38	Divorced	Sakai	Japan	Registered Nurse	1721836 Pe
14	Dorothea Whaplington	Female	43	Single	Ayna	Peru	Quality Engineer	1383237 Ed
15	Ethelred Yushkov	Male	29	Divorced	Mazamet	France	Financial Advisor	1686578 Au
16	Kerri Sappy	Female	27	Single	Haukivuori	Finland	Account Executive	1838160 Pe
17	Vikki Barnwell	Female	56	Divorced	Al Masa'ah ash Sharqiyah	Palestinian Territory	VP Accounting	626263 Au
18	Elspeth Dohmen	Male	44	Divorced	Adam	Mali	Geologist II	96400 Pe
19	Aly Jersch	Female	47	Divorced	Rangah	Indonesia	Geologist I	927562 Ed
20	Cart Hechlin	Male	52	Married	Mindresti	Moldova	Senior Editor	1701340 Ed
21	Rycca Pettman	Female	19	Single	Tha Yang	Thailand	Sales Representative	1552231 Pe

This screenshot shows a Microsoft Excel spreadsheet with data in columns J through T. The columns represent different loan types and their associated metrics like amount, interest rate, term, and status. A color-coded bar chart is overlaid on the data, where each row's background color corresponds to its 'Credit_Score_Band'. The chart highlights segments such as Personal, Education, Home, Auto, and others, with colors ranging from light blue for 'Excellent' to red for 'Poor'.

J	K	L	M	N	O	P	Q	R	S	T
1	Loan_Type	Loan_Amount	Interest_Rate	Loan_Term_Year	Loan_Status	Credit_Score	Income_Ratio	Loan_Eligibility_Status	Income_Catagory	Credit_Score_Band
2	Personal	133267	14	2	Ongoing	678	0.26864068	Not Eligible	Low Income	Average
3	Education	152751	11	9	Ongoing	644	0.213298275	Not Eligible	Medium Income	Average
4	Education	197723	10	29	Closed	419	0.115603553	Not Eligible	High Income	Poor
5	Education	51599	11	4	Default	470	0.028839372	Not Eligible	High Income	Poor
6	Home	74810	7	30	Closed	481	0.087537927	Not Eligible	Medium Income	Poor
7	Personal	162256	13	17	Default	362	0.141094014	Not Eligible	Medium Income	Poor
8	Auto	190982	15	30	Closed	598	0.501028386	Not Eligible	Low Income	Poor
9	Personal	58291	12	29	Default	793	0.030354679	Eligible	High Income	Good
10	Personal	131597	8	1	Closed	376	0.130353816	Not Eligible	Medium Income	Poor
11	Personal	87363	5	9	Default	564	0.055446076	Not Eligible	High Income	Poor
12	Personal	137662	6	1	Ongoing	678	0.094091835	Not Eligible	Medium Income	Average
13	Home	116242	10	16	Default	469	0.062164689	Not Eligible	High Income	Poor
14	Personal	84278	6	25	Closed	494	0.04894659	Not Eligible	High Income	Poor
15	Education	165313	17	5	Closed	423	0.119511696	Not Eligible	Medium Income	Poor
16	Auto	145486	12	9	Default	657	0.085727432	Not Eligible	High Income	Average
17	Personal	126331	13	1	Default	593	0.068726879	Not Eligible	High Income	Poor
18	Auto	143114	14	5	Approved	324	0.228520606	Not Eligible	Low Income	Poor
19	Personal	139920	10	2	Closed	767	0.145084275	Eligible	Medium Income	Good
20	Education	127195	8	5	Default	712	0.1371283	Eligible	Medium Income	Good
21	Education	158050	12	12	Ongoing	765	0.092897363	Eligible	High Income	Good
22	Personal	86227	14	20	Approved	707	0.055550366	Eligible	High Income	Good

DERIVED COLUMNS IN EXCEL

Several derived columns were created in Excel to enhance analytical depth. Loan Eligibility Status was calculated using customer income and credit score to classify applicants as Eligible or Not Eligible. Income Category segmented customers into High, Medium, and Low-income groups. Credit Score Band classified borrowers into Excellent, Good, Average, and Poor categories. Additionally, the Income Ratio was calculated as the ratio of loan amount to annual income. These derived columns supported eligibility assessment and risk evaluation.

➤ LOAN ELIGIBILITY STATUS

```
=IF(AND(O2>=700 ,I2>=400000),"Eligible", "Not Eligible")
```

➤ INCOME CATEGORY

=IF(I2>=1500000,"High Income", IF(I2>=700000,"Medium Income", "Low Income"))

➤ CREDIT SCORE BAND

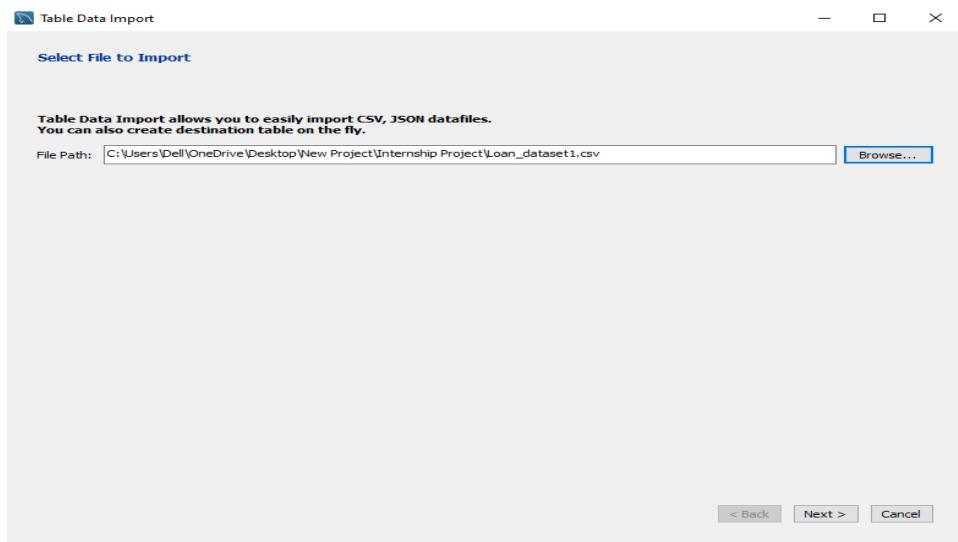
=IF(O2>=800,"Excellent",IF(O2>=700,"Good",IF(O2>=600,"Average","Poor")))

➤ INCOME RATIO

=LOAN AMOUNT/ANNUAL INCOME

SQL PROCESSING

The cleaned Excel-based loan dataset was imported into a SQL database for structured processing. SQL queries were used to filter relevant records and aggregate numerical data. GROUP BY clauses helped summarize loan amounts and customer income across categories. WHERE conditions were applied to segment loans based on eligibility and status. SQL processing improved efficiency, accuracy, and scalability of analysis. The final SQL tables were prepared for reporting and visualization purposes.



DATABASE CREATION IN SQL

A dedicated SQL database was created to manage and analyze the loan data efficiently. Aggregate functions such as SUM, AVG, and COUNT were used to summarize loan amounts and applicant income. Grouping operations enabled category-wise and status-wise analysis. Filtering techniques helped isolate active, closed, and eligible loan records. This structured database approach

ensured reliable data storage and supported advanced analytical queries required for business insights.

The screenshot shows four separate query windows in MySQL Workbench:

- Query 1 (top left):** A series of SELECT statements on the loan_dataset1 table. The results are shown in a Result Grid, with the last row being max_loan_amount = 199968.
- Query 2 (top right):** A series of SELECT statements on the loan_dataset1 table. The results are shown in a Result Grid, with the last row being total_loan = 362336.
- Query 3 (middle left):** A series of SELECT statements on the loan_dataset1 table. The results are shown in a Result Grid, with the last row being total_customers = 2000.
- Query 4 (middle right):** A series of SELECT statements on the loan_dataset1 table. The results are shown in a Result Grid, with the last row being avg_credit_score = 571.9229.
- Query 5 (bottom left):** A series of SELECT statements on the loan_dataset1 table. The results are shown in a Result Grid, with the last row being a LIMIT 5 result.
- Query 6 (bottom right):** A series of SELECT statements on the loan_dataset1 table. The results are shown in a Result Grid, with the last row being a LIMIT 5 result.

STATISTICAL ANALYSIS

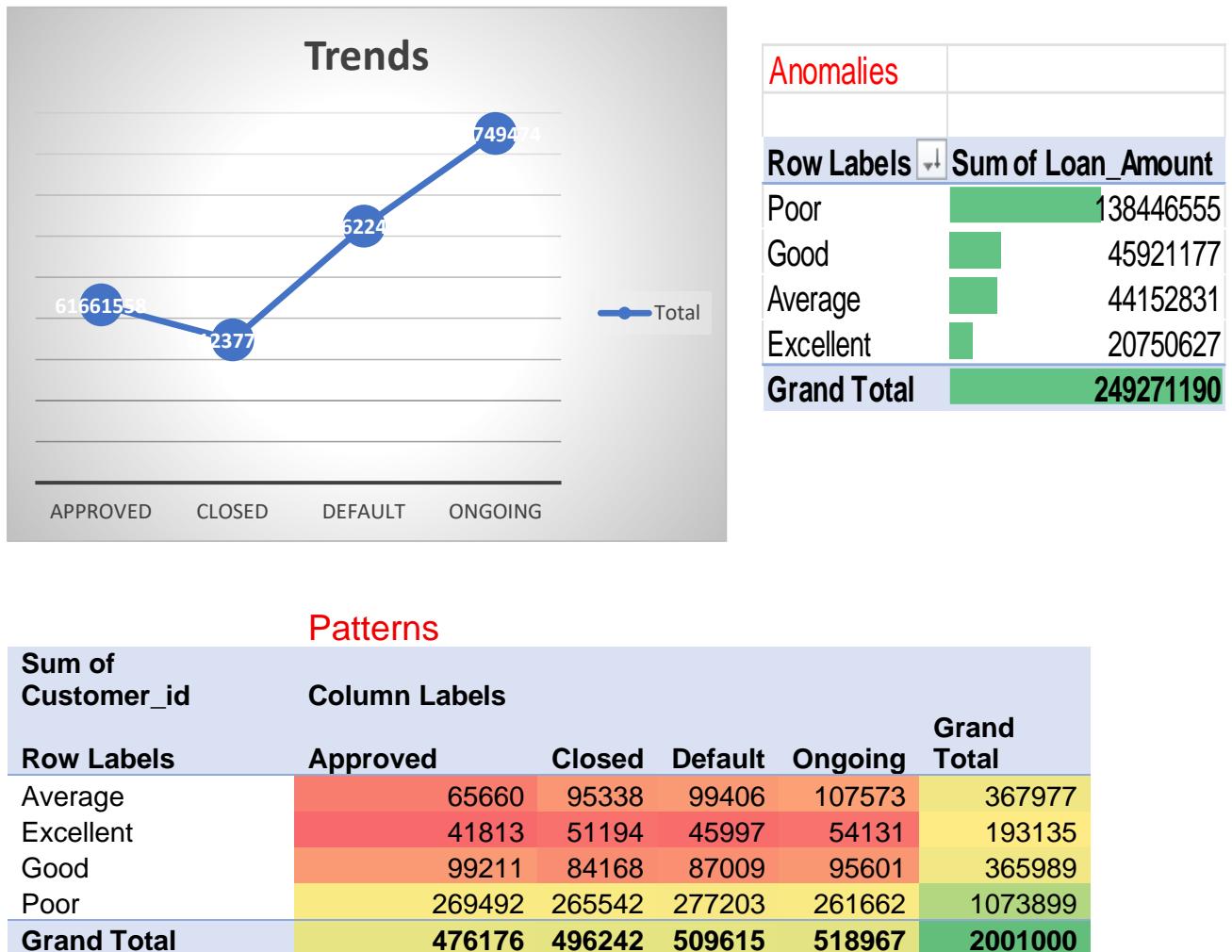
Statistical analysis was conducted to understand customer credit behavior and risk distribution. The mean credit score was calculated as 573.09, while the median score was 573, indicating a balanced distribution. Credit scores ranged from a minimum of 300 to a maximum of 850, highlighting wide variability in borrower risk profiles. These statistics helped identify overall creditworthiness trends and supported data-driven decision-making in loan evaluation.

<i>Statistical Analysis</i>	
Mean	573.088044
Standard Error	3.560118672
Median	573
Mode	847
Standard Deviation	159.1735388
Sample Variance	25336.21547
Kurtosis	-1.20193171
Skewness	0.02292431
Range	550
Minimum	300
Maximum	850
Sum	1145603
Count	1999
Confidence Level(95.0%)	6.981933906

Trend, Pattern, and Anomaly Analysis Using Pivot Tables

The pivot analysis clearly explains trends, patterns, and anomalies in the loan dataset. The trend shows that ongoing loans account for the highest loan amount, indicating maximum active exposure in the portfolio. Patterns reveal that customers with good and excellent credit scores consistently receive higher loan approvals and maintain healthier loan statuses. This confirms that credit score plays a major role in loan decision-making. At the same time, poor credit score customers show a higher tendency toward defaults. An anomaly is observed where poor credit score customers hold a disproportionately high total loan amount. This unexpected exposure indicates potential risk or relaxed

lending criteria. Overall, this combined analysis supports better risk management and data-driven lending decisions.



POWER QUERY – CONDITIONAL & DERIVED COLUMNS IN POWER BI

Power Query was used to create conditional and derived columns for advanced analysis. Risk Level was categorized as High, Medium, or Low based on credit score thresholds. Interest Rate Category classified loans into Low, Medium, and High interest groups. Additional calculations such as Total Amount, Interest Amount, and EMI Amount were derived using loan amount, interest rate, and loan tenure. These transformations automated data preparation and enhanced the analytical capability of Power BI dashboards.

Conditional Column Formula

- Risk Level = (if [Credit Score] < 600 then "High Risk" else if [Credit Score] < 750 then "Medium Risk" else "Low Risk")
- Interest Rate Category = (if [Interest Rate] < 8 then "Low Interest" else if [Interest Rate] <= 14 then "Medium Interest" else "High Interest")

Column Formula

- Total Amount = [Loan Amount]*[Interest Rate])

- Interest Amount = ([Loan Amount]*[Interest Rate])/100)

EMI Amount = (([Loan Amount]*[Interest Rate])/100)/[Loan term year])



Data	
<input type="checkbox"/>	Full_Name
<input type="checkbox"/>	Gender
<input type="checkbox"/>	Interest_Amou...
<input type="checkbox"/>	Interest_Rate ...
<input type="checkbox"/>	\sum Interest_Rate
<input type="checkbox"/>	Loan_Term_Cat...
<input type="checkbox"/>	\sum Loan_Amount
<input type="checkbox"/>	Loan_Status
<input type="checkbox"/>	\sum Loan_term_year
<input type="checkbox"/>	Loan_Type
<input type="checkbox"/>	Loans_Issued
<input type="checkbox"/>	Marital_Status
<input type="checkbox"/>	Occupation
<input type="checkbox"/>	Ongoing_Loan...
<input type="checkbox"/>	Risk_Level
<input type="checkbox"/>	Total_Loan_Am...
<input type="checkbox"/>	Total_Amount
<input type="checkbox"/>	Total_Customers
<input type="checkbox"/>	Total_Interest ...

DAX MEASURES

DAX measures were created to calculate key performance indicators for loan analysis. KPI cards displayed total loans issued, total customers, average interest rate, and counts of active and closed loans. These measures enabled real-time tracking of loan performance and customer activity. DAX calculations improved dashboard interactivity and provided a quick summary of portfolio health. They played a crucial role in monitoring business performance and lending efficiency.

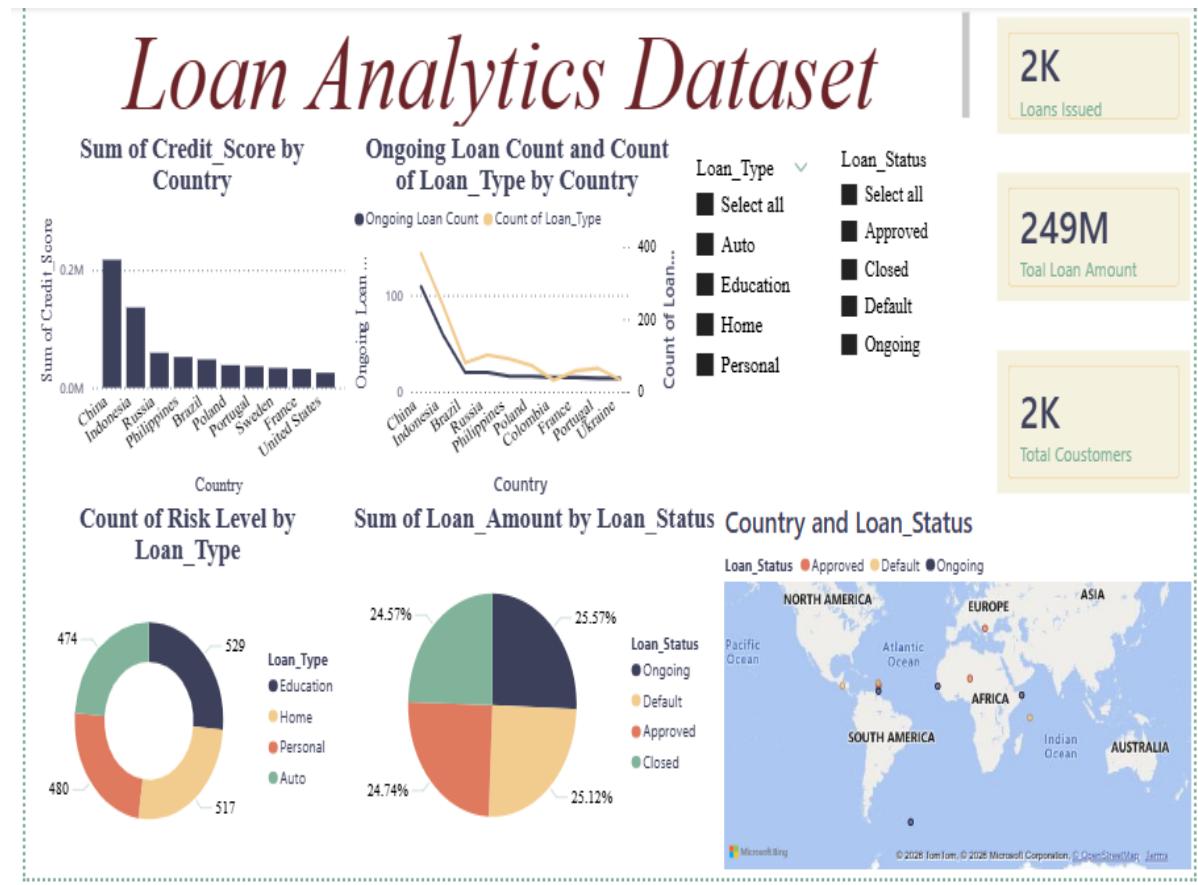
Columns Formula

- Loan Term Category = IF(data[Loan term year] > 5, "Long Term", "Short Term")
- Customer Profile = IF(data[Annual Income] > 600000 && data[Credit Score] >= 700, "Premium Customer", "Regular Customer")
- EMI Category = IF(data[Loan Amount] / (data[Loan term year] * 12) > 20000, "High EMI", "Low EMI")

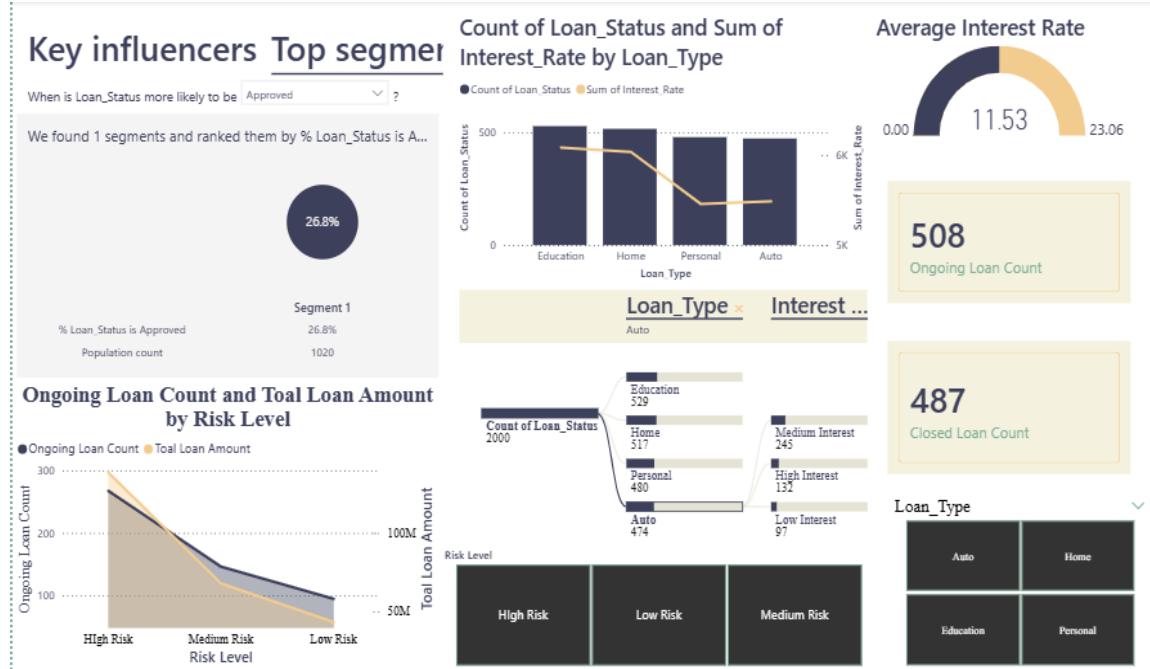
Measures Formula

- Ongoing Loan Count = CALCULATE(COUNT(data[Customer id]), data[Loan Status] = "Ongoing")
- Loans Issued = COUNT(data[Loan Amount])
- Total Customer's = DISTINCTCOUNT(data[Customer id])
- Average Interest Rate = AVERAGE(data[Interest Rate])
- Closed Loan Count = CALCULATE(COUNT(data[Customer id]), data[Loan Status] = "Closed")

First Dashboard view



Second Dashboard View



PROBLEM UNDERSTANDING & OBJECTIVE CLARITY

- The main objective of the project was to evaluate loan performance by analyzing customer profiles, loan characteristics, and repayment behaviour.
- The analysis aimed to identify key factors influencing loan approval, risk levels, and portfolio health.
- Credit score, income, interest rate, and loan tenure were considered critical variables.
- The project supports data-driven decision-making for risk management and loan optimization.
- Clear objectives ensured focused and meaningful analysis.

INSIGHT RELEVANCE & BUSINESS ALIGNMENT

- The analysis revealed strong relationships between credit score, interest rate, and loan performance.
- Customers with higher credit scores demonstrated better repayment behaviour, aligning with profitability goals.
- High-risk and high-interest loans showed increased default probability, supporting the need for stricter monitoring.

- Customer segmentation enabled targeted lending strategies and improved portfolio management.
- Overall, the insights aligned well with business objectives of reducing risk and maximizing returns.

Business Insights & Impact

- Loan approval decisions are primarily driven by income stability, credit history, and loan amount.
- Improving data quality and strengthening validation checks can significantly enhance decision accuracy.
- Targeted loan products can be designed for low-risk customer segments to improve approval rates while controlling defaults.

RECOMMENDATIONS

- Based on the analysis, it is recommended to strengthen credit evaluation for low credit score customers.
- High-interest and long-tenure loans should be closely monitored to reduce default risk.
- Risk-based pricing strategies can improve portfolio performance.
- Customer segmentation should be used to tailor loan products effectively. Continuous monitoring through dashboards will support proactive decision-making and enhance loan management efficiency.

CONCLUSION

- The loan dataset analysis provided valuable insights into customer behaviour and risk patterns.
- Data cleaning and preparation ensured accuracy and reliability. Statistical analysis highlighted key credit trends, while SQL enabled structured data processing.
- Power BI dashboards offered clear visualization and effective interpretation of complex data.
- Implementing these insights can improve decision-making, reduce defaults, and enhance overall business outcomes.

THANK YOU