Innovation & Entrepreneurship Hub for Educated Rural Youth (SURE Trust – IERY)

Loan Approval Intelligence Suite

The domain of the Project:

Data Analytics & Data Science

Team Mentor (and their designation):

Mr. Purnangshu Nath Roy

AI CONSULTANT @CSR BOX

Presented By:

Ms. Rishika Dipak Sahu

Period of the project:

November 2025 to December 2025

Declaration

The project titled “Loan Approval Intelligence Suite” has been mentored by Mr. Purnangshu Nath Roy, organised by SURE Trust, from November 2025 to December 2025, for the benefit of the educated unemployed rural youth for gaining hands-on experience in working on industry relevant projects that would take them closer to the prospective employer. I declare that to the best of my knowledge the participant’s name mentioned below, has worked on it successfully and enhanced her practical knowledge in the domain.

Participant Name:

Ms. Rishika D. Sahu

Mentored By:

Mr. Purnangshu Nath Roy
AI Consultant
CSR BOX

Prof. Radhakumari
Executive Director & Founder
SURE Trust

Table of contents

Executive Summary

The *Loan Approval Intelligence Suite* project was developed to simulate a real-world, end-to-end analytics workflow for evaluating loan applications in a financial institution. The objective was to build a unified system that cleans raw data, performs structured analysis, applies predictive modelling, and delivers insights through an interactive business dashboard. Four tools—Excel, SQL, Python, and Power BI—were integrated to demonstrate a complete data lifecycle.

The project began with Excel, where raw loan application data was cleaned, validated, and explored using pivot tables, formulas, and dashboard elements. The processed dataset was then imported into SQL for deeper analytical operations, including aggregations, joins, risk classification, and creation of reusable views. Insights from SQL were carried into Python, where exploratory data analysis, preprocessing, and a logistic regression model were used to predict loan approval outcomes and generate applicant risk scores. These results formed the basis of the final Power BI dashboard, which presented KPIs, slicers, trends, drill-through pages, and executive-level insights for data-driven decision-making.

Key findings highlight clear patterns in approval behaviour across income groups, credit history, education levels, and property areas. The predictive model provided measurable accuracy and helped identify attributes most influential in loan decisions. The final dashboard offered a streamlined, interactive view for stakeholders to monitor approval trends, assess applicant risk, and support faster, evidence-based decisions.

Overall, the project demonstrates a practical, interconnected data analytics pipeline suitable for modern financial workflows. It is recommended that organizations adopt similar structured approaches—combining data cleaning, database analytics, machine learning, and visualization—to enhance transparency, reduce manual effort, and make more consistent loan approval decisions.

Introduction

- Background and context of the project:
  Financial institutions process thousands of loan applications daily, and decisions must be accurate, consistent, and data-driven. With increasing data volume, banks rely heavily on analytics to assess applicant eligibility, predict risks, and optimize approval workflows. This project, *Loan Approval Intelligence Suite*, replicates a real industry-level data analytics pipeline using Excel, SQL, Python, and Power BI to demonstrate how data moves across tools and transforms into actionable insights.

- Problem statement and goal:
  The primary goal of this project is to build an end-to-end system that evaluates loan applications efficiently and intelligently. Traditional manual screening leads to delays, inconsistencies, and limited visibility into risk factors. To address this, the project aims to:
  - Clean and structure loan data for accuracy and usability
  - Analyze approval patterns using SQL
  - Build a predictive model to forecast loan approval likelihood
  - Create an interactive dashboard for decision-makers
    This ensures faster, more reliable, and insight-driven loan assessment.

- Scope and limitations:
  Scope:
  - Covers data cleaning, preprocessing, EDA, risk scoring, and visualization
  - Integrates four major tools (Excel, SQL, Python, Power BI)
  - Focuses on loan approval patterns, prediction, and reporting
  - Includes model evaluation and dashboard insights

  Limitations:
  - Uses an assumed or limited dataset as real banking data is confidential
  - Predictive accuracy depends on available features and data quality
  - Does not include deployment into a real-time production environment
  - Some complex financial parameters (e.g., credit scoring systems) are simplified for project purposes

- Innovation component in the project:
  - The project brings four tools together into a single connected workflow**,** simulating real fintech pipelines.
  - Introduces risk scoring using machine learning**,** giving an additional decision layer beyond simple approval logic.
  - Builds an executive-ready Power BI dashboard with drill-through insights, KPIs, and dynamic filters.
  - Focuses on interpretability using feature importance, helping stakeholders understand *why* an applicant is approved or rejected.
  - Encourages a shift from manual evaluation to automated data-driven decision-making**,** improving speed and consistency.

Project Objectives

- Objectives and goals:
    - To build an end-to-end analytics pipeline for loan approval using Excel, SQL, Python, and Power BI.
    - To clean, structure, and analyze loan data for accuracy and deeper insights.
    - To develop a predictive model that identifies the likelihood of loan approval and applicant risk.
    - To create an interactive, decision-focused dashboard for stakeholders to monitor trends and make informed decisions.
    - To simulate a real-world financial analytics workflow that demonstrates practical industry skills.

- Expected outcomes and deliverables:
    - A fully cleaned and validated loan dataset ready for analysis.
    - SQL-based analytical reports identifying approval trends, risk categories, and applicant patterns.
    - A Python-generated machine learning model with performance metrics and applicant risk scores.
    - A Power BI dashboard containing KPIs, slicers, visual trends, and drill-through reports.
    - A unified, interlinked workflow showing how data progresses across all four tools in the project.

Methodology and Results

- Methods/Technology used:

  The project follows a complete data analytics lifecycle, moving step-by-step across tools that simulate real industry workflows:
  - **Data Cleaning & Preparation:** Raw loan data was cleaned, validated, and transformed to make it analysis-ready.
  - **Data Exploration & SQL Analytics:** Patterns were explored using SQL queries, aggregations, filtering, and risk classifications.
  - **Predictive Modelling:** A logistic regression model was built in Python to predict loan approval outcomes.
  - **Risk Scoring:** Each applicant was assigned a risk score based on model outputs and key attributes.
  - **Interactive Visualization:** Insights and predictions were visualized through a Power BI dashboard designed for executives.

- Tools/Software used:

  Microsoft Excel: Data cleaning, validation, pivots, dashboards

  SQL Server: Data storage, transformations, analytical querying

  Python (Pandas, NumPy, Scikit-learn, Matplotlib): EDA, modelling, evaluation, risk scoring

  Power BI: KPI creation, dashboards, slicers, drill-through reports

  Jupyter Notebook: Coding and experimentation

  GitHub: Version control and project storage

- Project Architecture:

  Step 1: Excel – Data Cleaning & Feature Preparation
  - Remove duplicates, fix missing values
  - Create derived fields (total income, loan ratios)
  - Prepare final clean dataset for SQL ingestion

  Step 2: SQL – Structured Analysis & Insights
  - Store data in relational tables
  - Run analytical queries: approval counts, property area distribution, credit history patterns
  - Create views and filter high-risk profiles
  - Export refined dataset to Python

Step 3: Python – Modelling & Risk Scoring

- Load SQL output
- Perform EDA (histograms, correlations, boxplots)
- Encode categorical variables
- Build Logistic Regression model
- Evaluate accuracy, precision, recall
- Generate applicant-level risk score
- Export final dataset for BI use

Step 4: Power BI – Visualization & Reporting

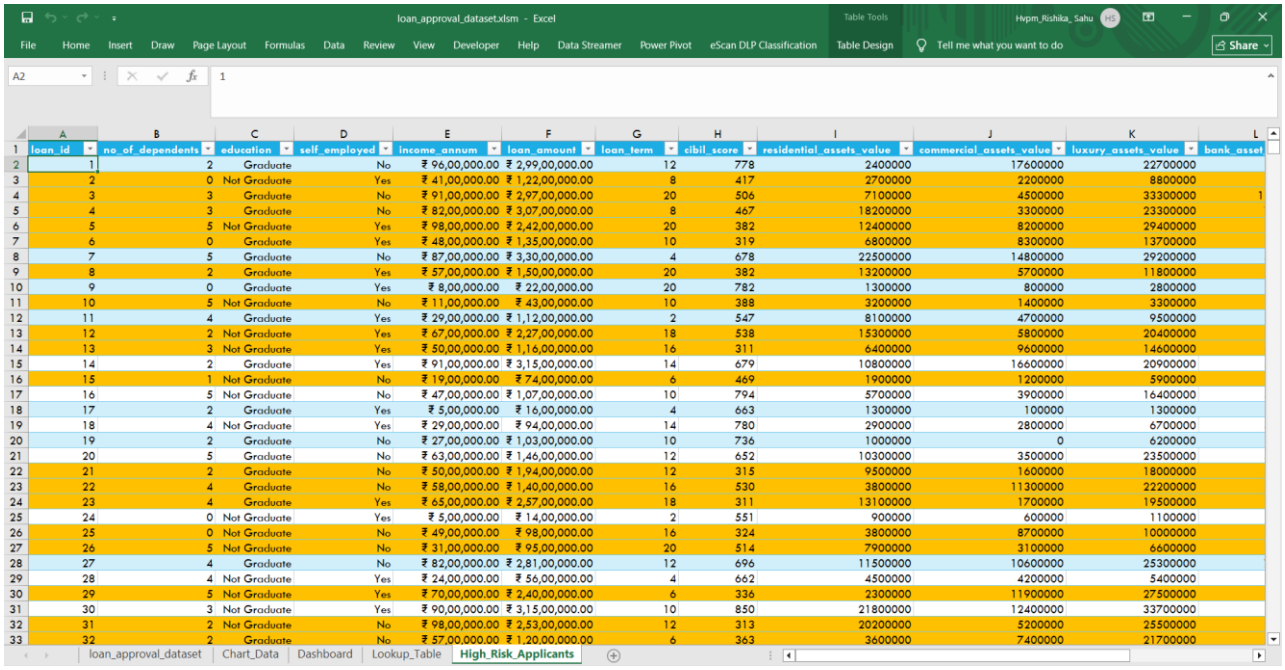
- Import Python output
- Create KPIs: approval rate, average income, high-risk count
- Use slicers for gender, education, property area
- Final dashboard summarizing predictions and insights

Architecture Flow:

Excel → SQL → Python (Model & Risk Scores) → Power BI Dashboard

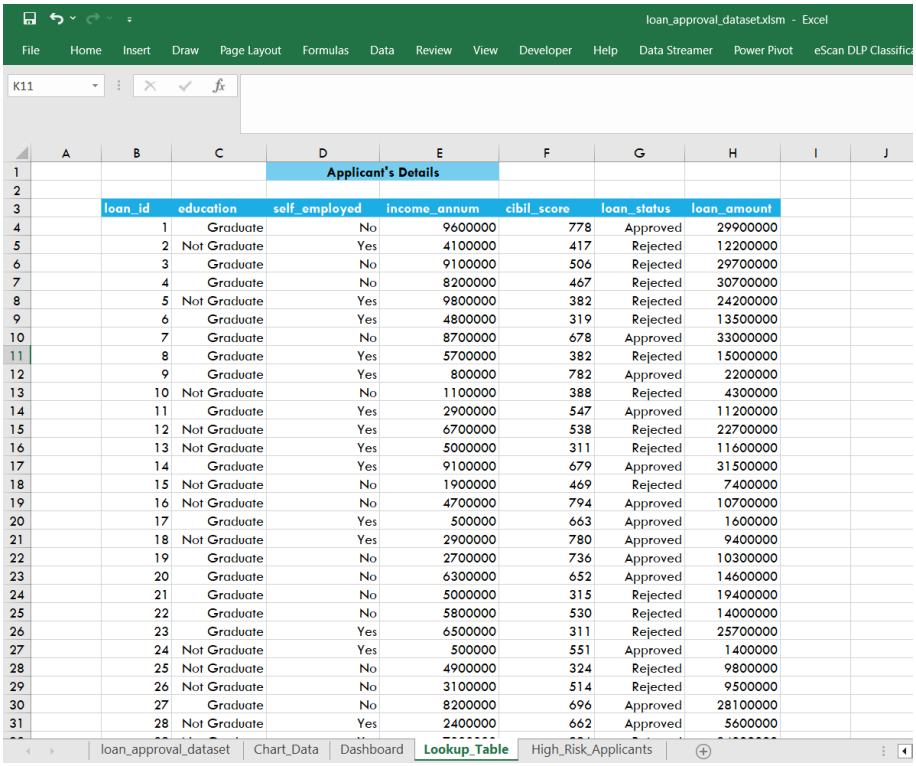
- Final project working screenshots:

Screenshot 1: Excel



loan_id	no_of_dependents	education	self_employed	income_annum	loan_amount	loan_term	cibil_score	residential_assets_value	commercial_assets_value	luxury_assets_value	bank_asset
1	2	Graduate	No	₹ 96,00,000.00	₹ 2,99,00,000.00	12	778	2400000	17600000	22700000	
2	0	Not Graduate	Yes	₹ 41,00,000.00	₹ 1,22,00,000.00	8	417	2700000	2200000	8800000	
3	3	Graduate	No	₹ 91,00,000.00	₹ 2,97,00,000.00	20	506	7100000	4500000	33300000	1
4	3	Graduate	No	₹ 82,00,000.00	₹ 3,07,00,000.00	8	467	18200000	3300000	23300000	
5	5	Not Graduate	Yes	₹ 98,00,000.00	₹ 2,42,00,000.00	20	382	12400000	8200000	29400000	
6	0	Graduate	Yes	₹ 48,00,000.00	₹ 1,35,00,000.00	10	319	6800000	8300000	13700000	
7	5	Graduate	No	₹ 87,00,000.00	₹ 3,30,00,000.00	4	678	22500000	14800000	29200000	
8	2	Graduate	Yes	₹ 57,00,000.00	₹ 1,50,00,000.00	20	382	13200000	5700000	11800000	
9	0	Graduate	Yes	₹ 8,00,000.00	₹ 22,00,000.00	20	782	1300000	800000	2800000	
10	5	Not Graduate	No	₹ 11,00,000.00	₹ 43,00,000.00	10	388	3200000	1400000	3300000	
11	4	Graduate	Yes	₹ 29,00,000.00	₹ 1,12,00,000.00	2	547	8100000	4700000	9500000	
12	2	Not Graduate	Yes	₹ 67,00,000.00	₹ 2,27,00,000.00	18	538	15300000	5800000	20400000	
13	3	Not Graduate	Yes	₹ 50,00,000.00	₹ 1,16,00,000.00	16	311	6400000	9600000	14600000	
14	2	Graduate	Yes	₹ 91,00,000.00	₹ 3,15,00,000.00	14	679	10800000	16600000	20900000	
15	1	Not Graduate	No	₹ 19,00,000.00	₹ 74,00,000.00	6	469	1900000	1200000	5900000	
16	5	Not Graduate	No	₹ 47,00,000.00	₹ 1,07,00,000.00	10	794	5700000	3900000	16400000	
17	2	Graduate	Yes	₹ 5,00,000.00	₹ 16,00,000.00	4	663	1300000	100000	1300000	
18	4	Not Graduate	Yes	₹ 29,00,000.00	₹ 94,00,000.00	14	780	2900000	2800000	6700000	
19	2	Graduate	No	₹ 27,00,000.00	₹ 1,03,00,000.00	10	736	1000000	0	6200000	
20	5	Graduate	No	₹ 63,00,000.00	₹ 1,46,00,000.00	12	652	10300000	3500000	23500000	
21	2	Graduate	No	₹ 50,00,000.00	₹ 1,94,00,000.00	12	315	9500000	1600000	18000000	
22	4	Graduate	No	₹ 58,00,000.00	₹ 1,40,00,000.00	16	530	3800000	11300000	22200000	
23	4	Graduate	Yes	₹ 65,00,000.00	₹ 2,57,00,000.00	18	311	13100000	1700000	19500000	
24	0	Not Graduate	Yes	₹ 5,00,000.00	₹ 14,00,000.00	2	551	900000	600000	1100000	
25	0	Not Graduate	No	₹ 49,00,000.00	₹ 98,00,000.00	16	324	3800000	8700000	10000000	
26	5	Not Graduate	No	₹ 31,00,000.00	₹ 95,00,000.00	20	514	7900000	3100000	6600000	
27	4	Graduate	No	₹ 82,00,000.00	₹ 2,81,00,000.00	12	696	11500000	10600000	25300000	
28	4	Not Graduate	Yes	₹ 24,00,000.00	₹ 56,00,000.00	4	662	4500000	4200000	5400000	
29	5	Not Graduate	Yes	₹ 70,00,000.00	₹ 2,40,00,000.00	6	336	2300000	11900000	27500000	
30	3	Not Graduate	Yes	₹ 90,00,000.00	₹ 3,15,00,000.00	10	850	21800000	12400000	33700000	
31	2	Not Graduate	No	₹ 98,00,000.00	₹ 2,53,00,000.00	12	313	20200000	5200000	25500000	
32	2	Graduate	No	₹ 57,00,000.00	₹ 1,20,00,000.00	6	363	3600000	7400000	21700000	

Fig 1.1: Conditional Formatting – High risk applicants



loan_approval_dataset.xlsm - Excel

File Home Insert Draw Page Layout Formulas Data Review View Developer Help Data Streamer Power Pivot eScan DLP Classifica

K11

	A	B	C	D	E	F	G	H	I	J
1				Applicant's Details						
2										
3		loan_id	education	self_employed	income_annum	cibil_score	loan_status	loan_amount		
4		1	Graduate	No	9600000	778	Approved	29900000		
5		2	Not Graduate	Yes	4100000	417	Rejected	12200000		
6		3	Graduate	No	9100000	506	Rejected	29700000		
7		4	Graduate	No	8200000	467	Rejected	30700000		
8		5	Not Graduate	Yes	9800000	382	Rejected	24200000		
9		6	Graduate	Yes	4800000	319	Rejected	13500000		
10		7	Graduate	No	8700000	678	Approved	33000000		
11		8	Graduate	Yes	5700000	382	Rejected	15000000		
12		9	Graduate	Yes	800000	782	Approved	2200000		
13		10	Not Graduate	No	1100000	388	Rejected	4300000		
14		11	Graduate	Yes	2900000	547	Approved	11200000		
15		12	Not Graduate	Yes	6700000	538	Rejected	22700000		
16		13	Not Graduate	Yes	5000000	311	Rejected	11600000		
17		14	Graduate	Yes	9100000	679	Approved	31500000		
18		15	Not Graduate	No	1900000	469	Rejected	7400000		
19		16	Not Graduate	No	4700000	794	Approved	10700000		
20		17	Graduate	Yes	500000	663	Approved	1600000		
21		18	Not Graduate	Yes	2900000	780	Approved	9400000		
22		19	Graduate	No	2700000	736	Approved	10300000		
23		20	Graduate	No	6300000	652	Approved	14600000		
24		21	Graduate	No	5000000	315	Rejected	19400000		
25		22	Graduate	No	5800000	530	Rejected	14000000		
26		23	Graduate	Yes	6500000	311	Rejected	25700000		
27		24	Not Graduate	Yes	500000	551	Approved	1400000		
28		25	Not Graduate	No	4900000	324	Rejected	9800000		
29		26	Not Graduate	No	3100000	514	Rejected	9500000		
30		27	Graduate	No	8200000	696	Approved	28100000		
31		28	Not Graduate	Yes	2400000	662	Approved	5600000		

loan_approval_dataset | Chart_Data | Dashboard | Lookup_Table | High_Risk_Applicants

Fig 1.2: VLOOKUP – Applicant's details

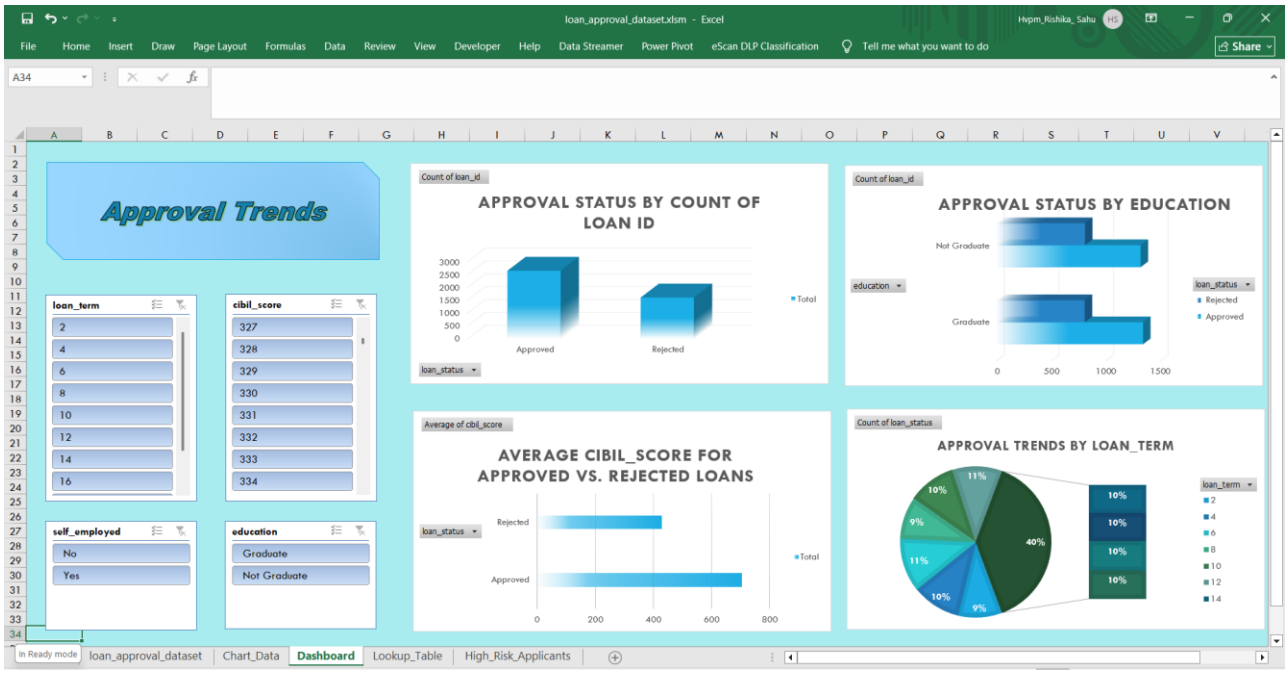


Fig 1.3: Slicer enabled Dashboard – Approval trends

Screenshot 2: SQL Query Output

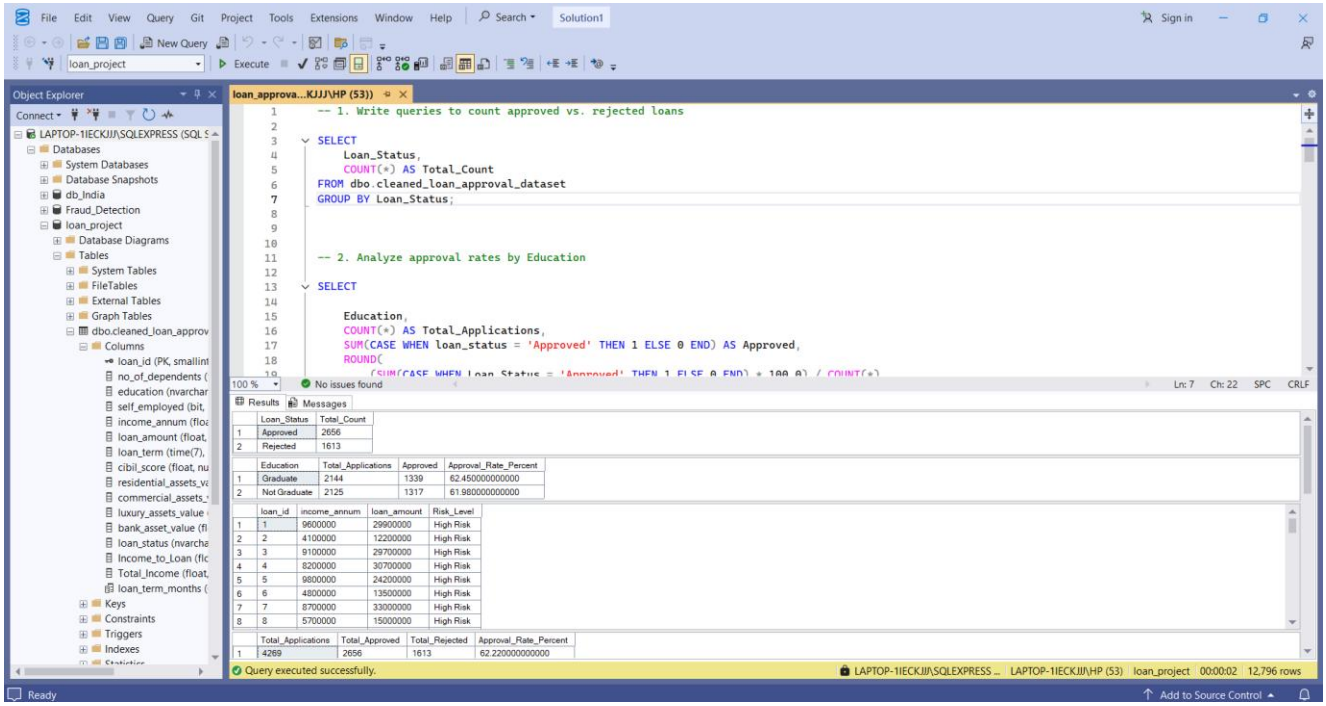


Fig 2: SQL Server Management Studio – Queries and Output

Screenshot 3: Python

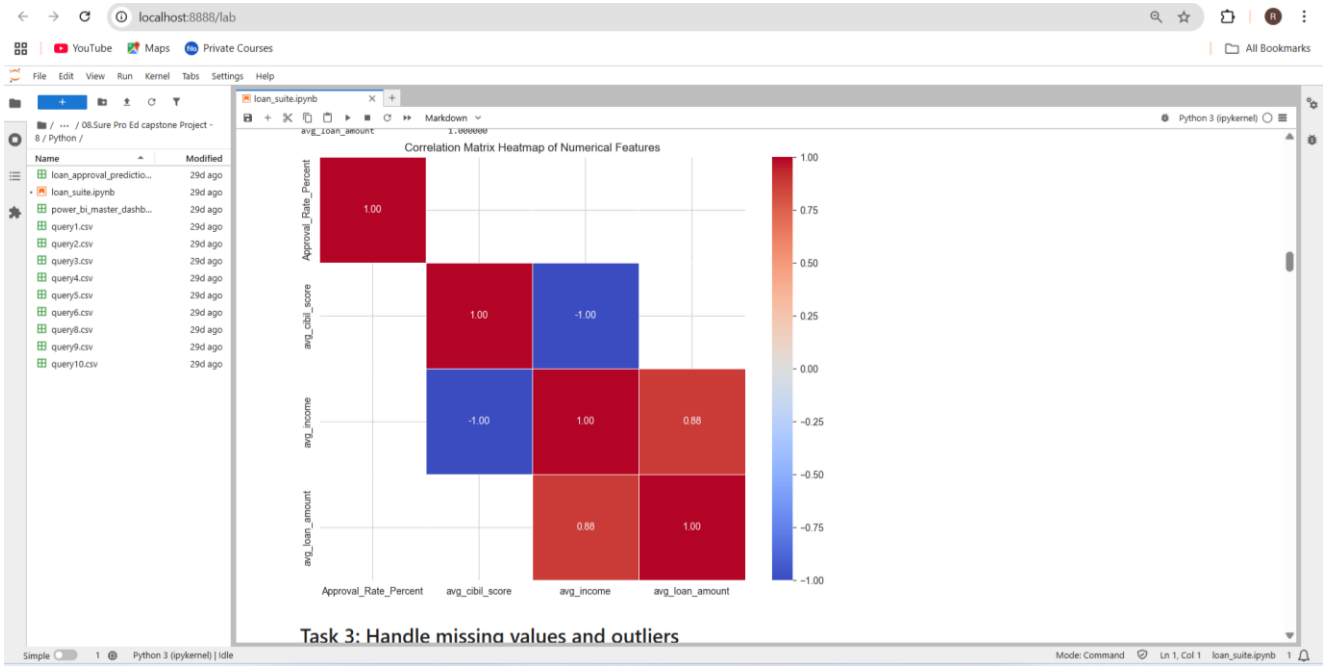


Fig 3.1 Correlation Matrix Heatmap of Numerical features

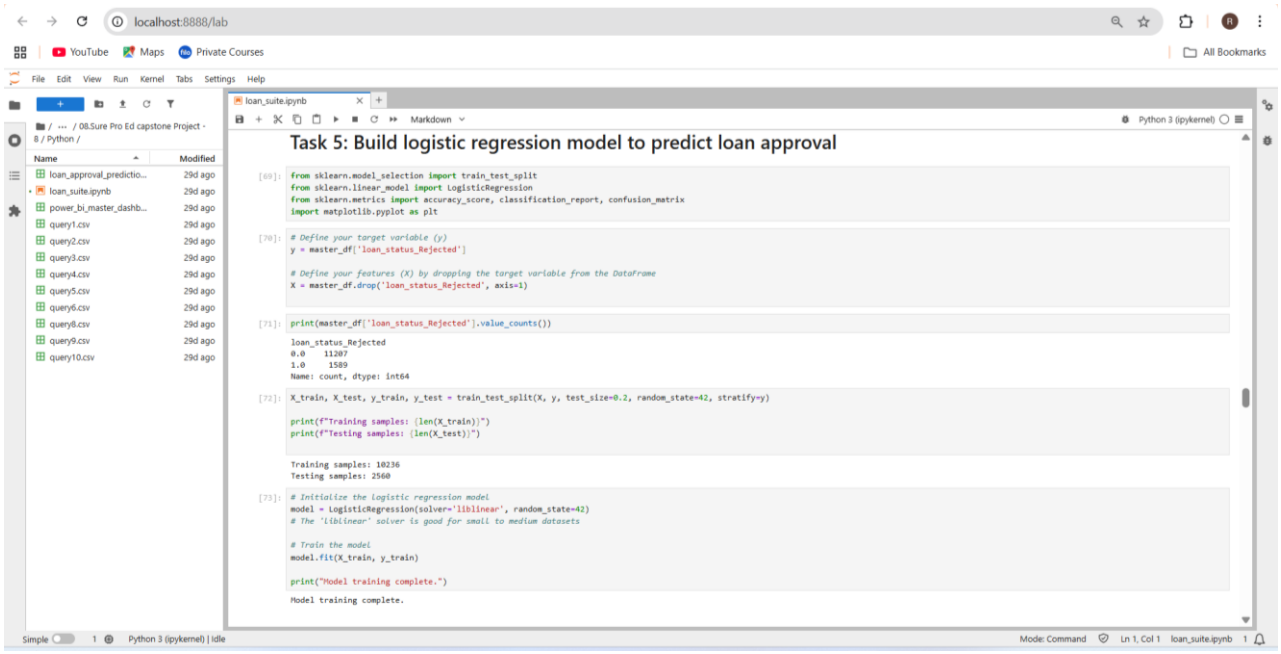


Fig 3.2: Model Building – Logistic Regression

Screenshot 4: Power BI Dashboard

Fig 4: Power BI Dashboard

- Project GitHub Link –

  https://github.com/Rishika-s11/Loan_Approval_Intelligence_Suite

Social/ Industry Relevance of the Project

The *Loan Approval Intelligence Suite* holds significant relevance in both social and industry contexts, as it addresses real challenges faced by financial institutions and the communities they serve.

Industry Relevance

- **Improved Decision-Making:**
  Banks and lending companies rely on data-driven systems to evaluate applications accurately. This project demonstrates how analytics and machine learning can streamline loan approval processes and reduce manual errors.
- **Operational Efficiency:**
  Automating data cleaning, analysis, and risk evaluation helps financial institutions process applications faster, saving time and resources while handling large volumes of data.
- **Risk Management:**
  Predictive models, like the one developed in this project, help identify high-risk applicants early, reducing chances of loan default and improving portfolio health.
- **Data Integration Workflow:**
  The project reflects modern industry practices by connecting Excel, SQL, Python, and Power BI—tools widely used in banking, fintech, and analytics roles.

Social Relevance

- **Fair and Transparent Loan Decisions:**
  Data-driven evaluation reduces human bias and promotes fairer approval processes, helping deserving applicants get access to financial support.
- **Financial Inclusion:**
  By understanding patterns in applicant behavior, institutions can design more inclusive lending policies, supporting low-income groups and first-time borrowers.
- **Empowering Individuals with Better Insights:**
  Predictive analytics helps applicants know their approval chances and financial standing, allowing them to make smarter borrowing decisions.
- **Strengthening Economic Growth:**
  Faster and more accurate loan approvals enable more people to access funds for education, housing, business, and personal development—supporting broader economic progress.

Learning and Reflection

1. New Learnings

Working on this individual capstone project allowed me to gain hands-on experience across multiple stages of the data analytics lifecycle. Key learnings include:

- **Technical Skills:**
    - Improved my ability to clean and preprocess datasets using Excel functions, pivot tables, and validation tools.
    - Strengthened SQL querying skills, including joins, aggregations, views, subqueries, and risk classification techniques.
    - Gained practical experience in Python for EDA, preprocessing, logistic regression modelling, and evaluating machine learning performance.
    - Learned how to integrate model outputs into Power BI and design professional dashboards with KPIs, slicers, and drill-through pages.
- **Analytical & Problem-Solving Skills:**
    - Understood how different data attributes influence loan approval decisions.
    - Learned how to interpret model outputs using metrics and feature importance.
- **Management & Workflow Skills:**
    - Managed the entire project independently, from planning to execution.
    - Developed better time management and task breakdown skills by completing structured modules.
    - Learned how to document work effectively.

2. Overall Experience

Completing this project individually was a highly rewarding experience. It gave me the opportunity to work across multiple technologies and simulate a real-world workflow used in financial institutions. Moving step-by-step through Excel, SQL, Python, and Power BI helped me understand how each tool contributes uniquely to decision-making.

I also gained confidence in solving challenges independently, such as handling missing data, debugging SQL queries, improving model accuracy, and designing meaningful visual dashboards. Overall, the project strengthened both my technical capabilities and my ability to manage a complete analytics pipeline from scratch. It has been a practical, insightful, and growth-oriented experience that adds strong value to my data analytics journey.

Future Scope and Conclusion

Future Scope:

- **Advanced Machine Learning Models:**
  Future versions can include Random Forest, XGBoost, or Neural Networks to improve prediction accuracy and handle complex relationships.
- **Real-Time Decision System:**
  Integrating the model into a web application or API can help banks perform instant loan eligibility checks.
- **Automated Data Pipeline:**
  Tools like Airflow or Power Automate can be used to automate data refresh from Excel → SQL → Python → Power BI.
- **Enhanced Feature Engineering:**
  Incorporating external data such as credit bureau scores, employment history, and financial behavior can improve model robustness.
- **Enterprise-Level Dashboard:**
  Adding forecasting visuals, scenario analysis, and role-based views can make the dashboard more powerful for business users.
- **Deployment in Cloud:**
  Hosting the workflow on cloud platforms like Azure or AWS can support scalability, collaboration, and secure access.

Conclusion:

The primary objective of this project was to build a complete, industry-style analytics pipeline for evaluating loan applications. This was successfully achieved by integrating Excel, SQL, Python, and Power BI into a smooth, interconnected workflow. The project cleaned and structured raw data, performed detailed SQL analysis, developed a predictive model for loan approval using logistic regression, and visualized insights through an interactive Power BI dashboard.

Through this process, the project demonstrated how data flows across tools—from cleaning to modelling to visualization—and highlighted key patterns in loan approval behaviour. The generated risk scores and performance metrics also provided deeper understanding of applicant eligibility and decision factors. Overall, the project met all intended goals and delivered a functional, data-driven solution suitable for financial decision-making scenarios.