Master Project: Risk Analysis & Predictive Modeling

# Project Overview

This project focuses on the domain of Risk Analysis & Predictive Modeling using synthetic data from an insurance and lending company.   
It is divided into three major phases: Python (Data Engineering & Exploration), SQL (Data Manipulation & Analysis), and Visualization (Data Storytelling & Dashboards).  
The dataset consists of 3 interconnected CSV files with over 20,000 records each, simulating real-world financial behavior and insurance risk factors.

# Phase 1: Python

In this phase, you'll use Python for data cleaning, preprocessing, and initial analysis.   
Tasks include handling missing data, detecting outliers, correcting data inconsistencies, and feature engineering.  
Key Activities:  
- Cleaning inconsistent values (e.g., negative incomes, invalid credit scores)  
- Deriving new features like 'risk\_category', 'debt\_to\_income\_ratio'  
- Data validation and merging datasets  
- Exploratory data analysis using pandas, matplotlib, seaborn

# Phase 2: SQL

Use SQL to query the cleaned and merged data for business insights. Load the cleaned CSVs into a database.  
Key Queries:  
- Risk segmentation across regions and employment types  
- Average loan/claim amount per credit score band  
- Default rate by claim type and employment status  
- Claim status outcomes for high-risk individuals

# Phase 3: Visualization

This phase involves creating dashboards to visually communicate KPIs and insights.  
Suggested Tools: Power BI, Tableau, or Python (Plotly, Dash, Streamlit)  
Visualizations:  
- Regional risk heatmap  
- Time-series of claim types  
- Risk score vs default/claim status  
- KPI dashboard: average credit score, default rate, claim approval rate, risk distribution

# Dataset Metadata

customers.csv

- customer\_id: Unique customer identifier  
- name: Full name  
- age: Customer's age  
- income: Annual income (some negative values added intentionally)  
- employment\_status: Employment category  
- region: Geographic area  
- signup\_date: Customer sign-up timestamp

financials.csv

- customer\_id: Foreign key to customers  
- loan\_amount: Loan requested or issued  
- credit\_score: Credit score (some values >850 added)  
- late\_payments: Number of late payments  
- default\_history: Binary value if customer defaulted  
- risk\_score: Calculated score based on financial behavior  
- risk\_level: Categorical risk tier derived from risk\_score

claims.csv

- claim\_id: Unique claim identifier  
- customer\_id: Foreign key to customers  
- incident\_date: Date of insurance incident  
- claim\_amount: Amount filed for claim (some negative values added)  
- claim\_status: Result of the claim  
- claim\_type: Nature of the claim

# Known Data Issues (Intentional)

- Row 101 in customers.csv has negative income (-10000)  
- Row 201 in financials.csv has invalid credit\_score (999)  
- Row 301 in claims.csv has negative claim\_amount (-5000)  
These simulate common real-world data entry or system errors and must be cleaned during Python phase.