Consumer Finance Loan Risk Analysis

EDA case study - Identifying drivers of loan default (Fully Paid vs Charged Off)

Prepared by:Anshuman Mohapatra
Date:13-02-2026

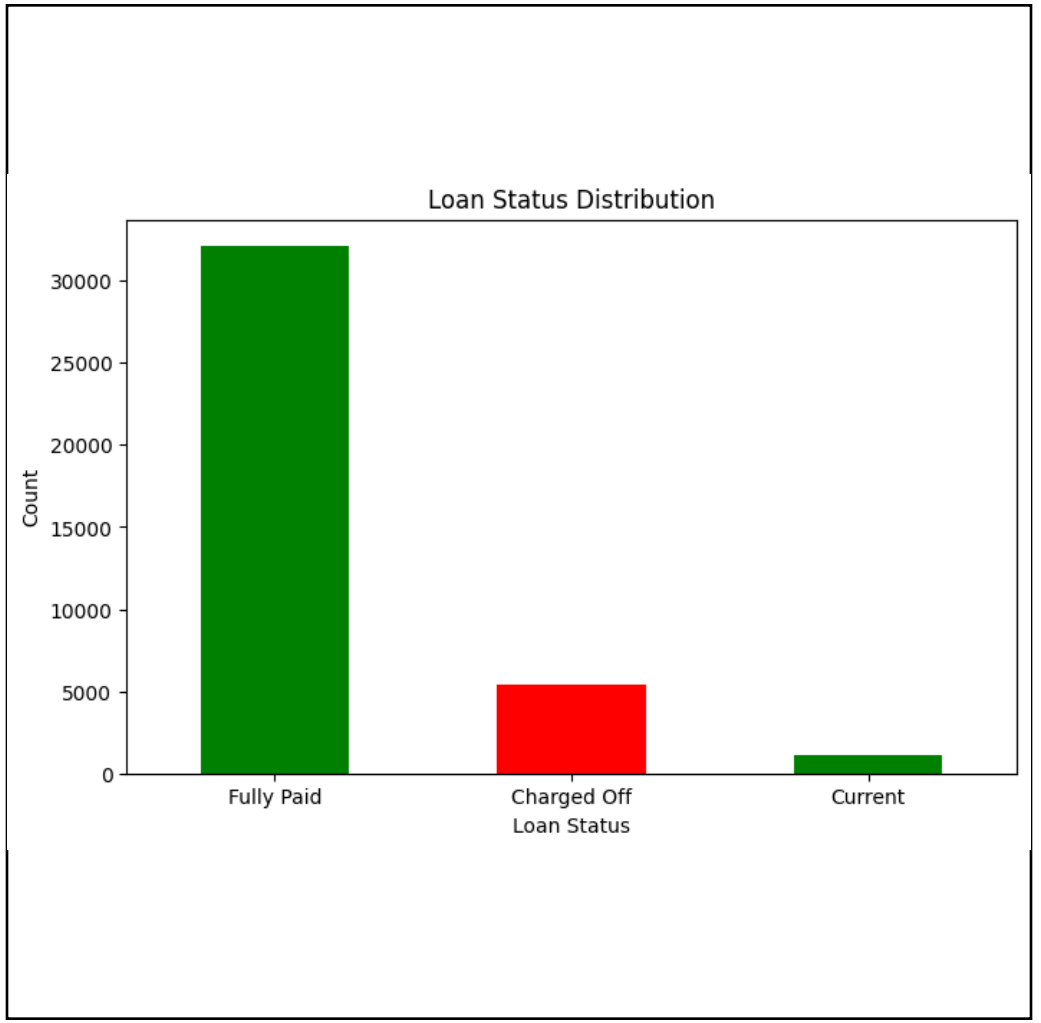
Notebook: Gramener Consumer Finance Loan Risk Case Study - FINAL.ipynb

Problem Statement & Objective

- Goal: understand the key driver variables behind loan default.
- Business need: reduce credit loss by screening high-risk applicants early.
- Target variable: loan_status (focus on Fully Paid vs Charged Off).
- Approach: structured EDA - data cleaning, univariate, bivariate, and multivariate analysis.

Data Overview & Target Definition

- Dataset contains borrower profile, loan details, and credit history variables.
- Target created: is_default = 1 for Charged Off, 0 for Fully Paid.
- Current loans were excluded to avoid unknown outcomes.
- Observation: dataset is imbalanced (defaults are minority).



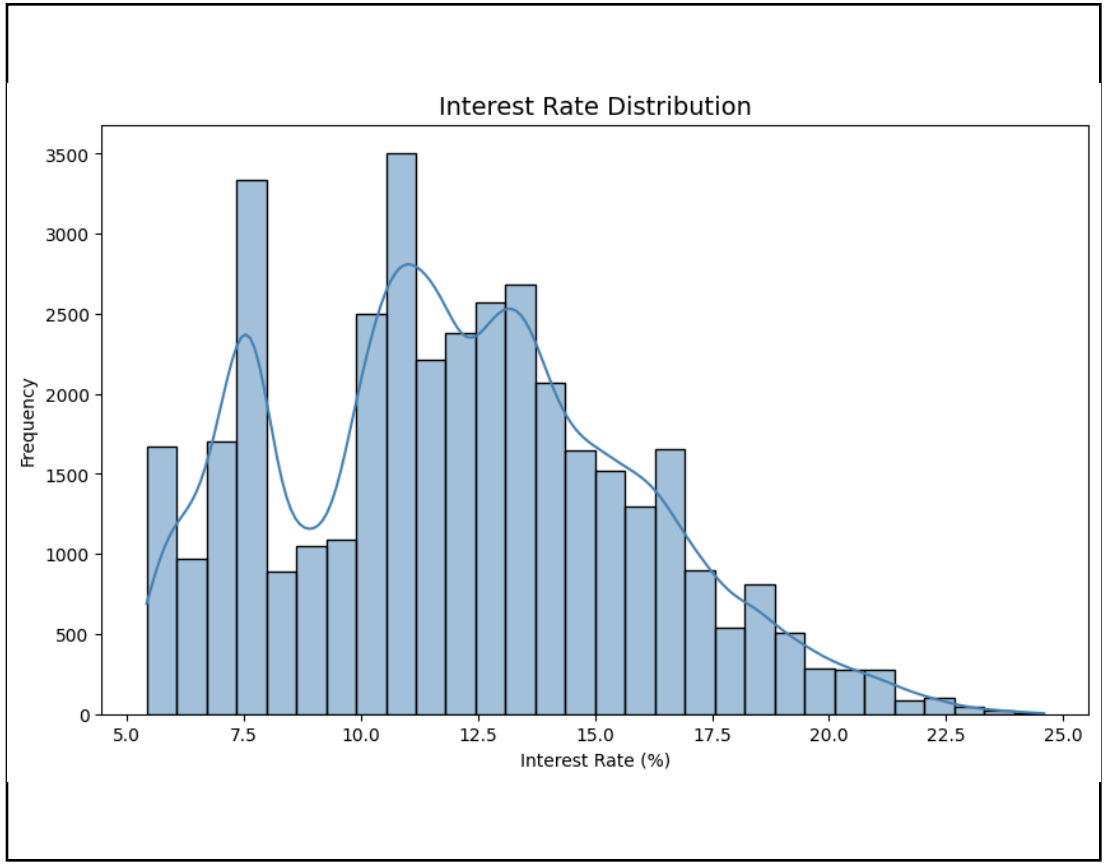
Loan Status Distribution (Notebook cell 92)

Data Cleaning & Preparation

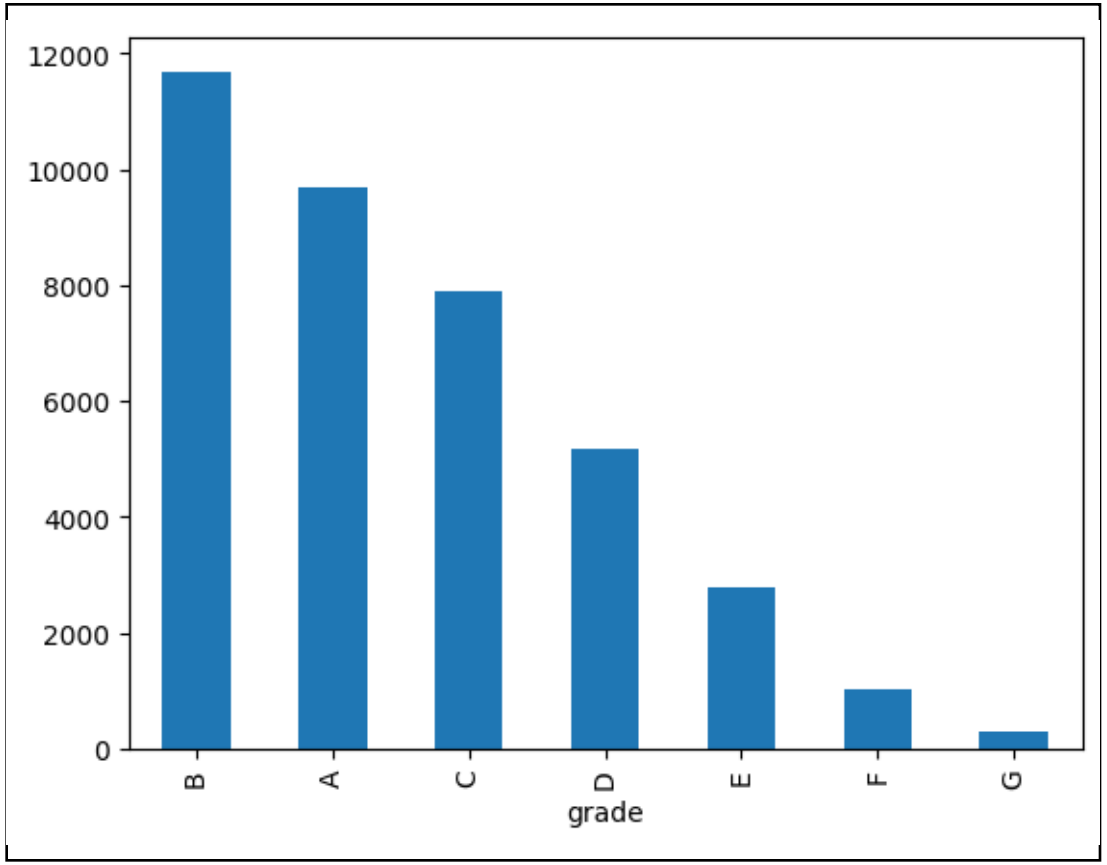
- Removed fully-null columns and reviewed partially-missing columns (drop/keep based on missing %).
- Dropped redundant/irrelevant columns for default analysis.
- Standardized data types:
- int_rate: removed '%' and converted to float
  - term: extracted numeric months (36/60)
  - emp_length: cleaned text and converted to numeric
- Created derived metrics for analysis: default flag + binned bands for numerical features (income, loan amount, DTI, utilization, interest rate).
- Outlier handling for visuals: capped extreme values (e.g., annual income and revolving balance) at high percentiles (95th/99th) instead of removing real records.

Univariate Insights - Pricing & Internal Risk Grade

- Interest rate ranges ~5% to ~25% with right-skew (many loans in 8-16%).
- Grades are concentrated in A-C; F/G are rare but represent highest risk tiers.



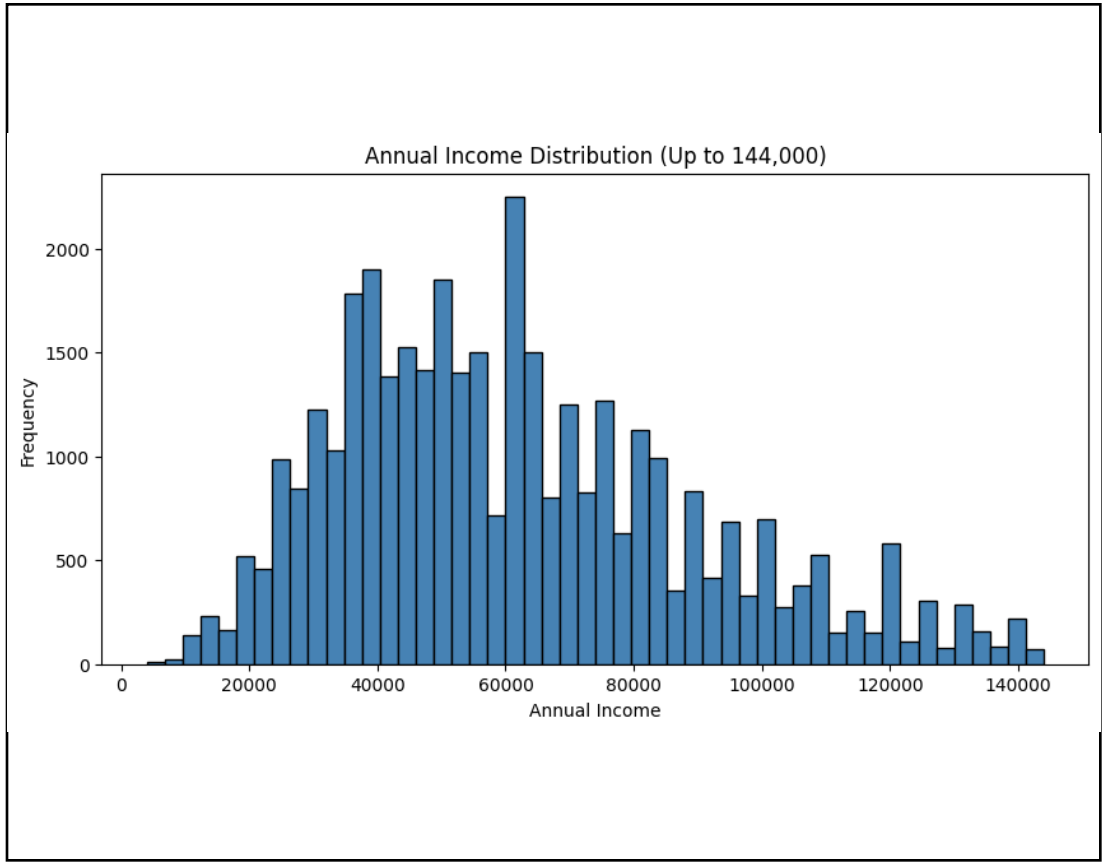
Interest Rate Distribution (Notebook cell 99)



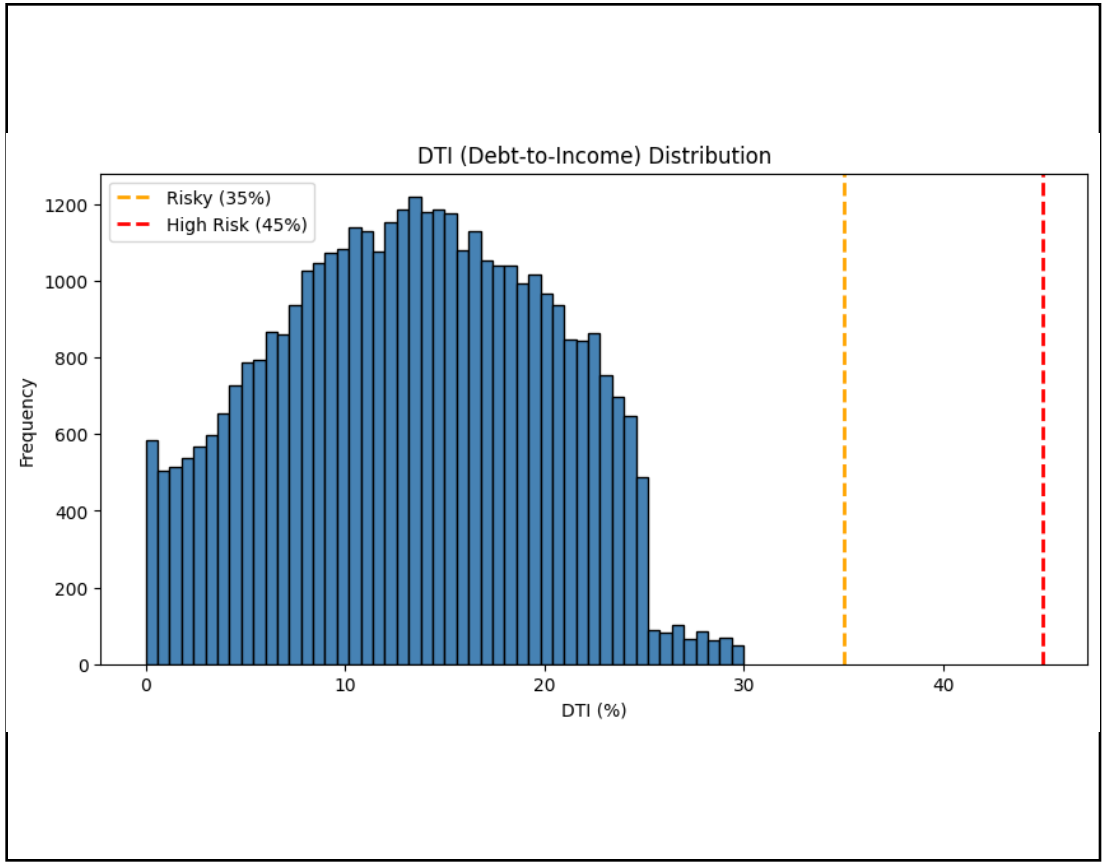
Loan Grade Distribution (Notebook cell 110)

Univariate Insights - Affordability (Income & DTI)

- Annual income is heavily right-skewed; capping helps interpretation (most borrowers $30K-$100K).
- DTI is relatively clean; higher values indicate repayment stress.



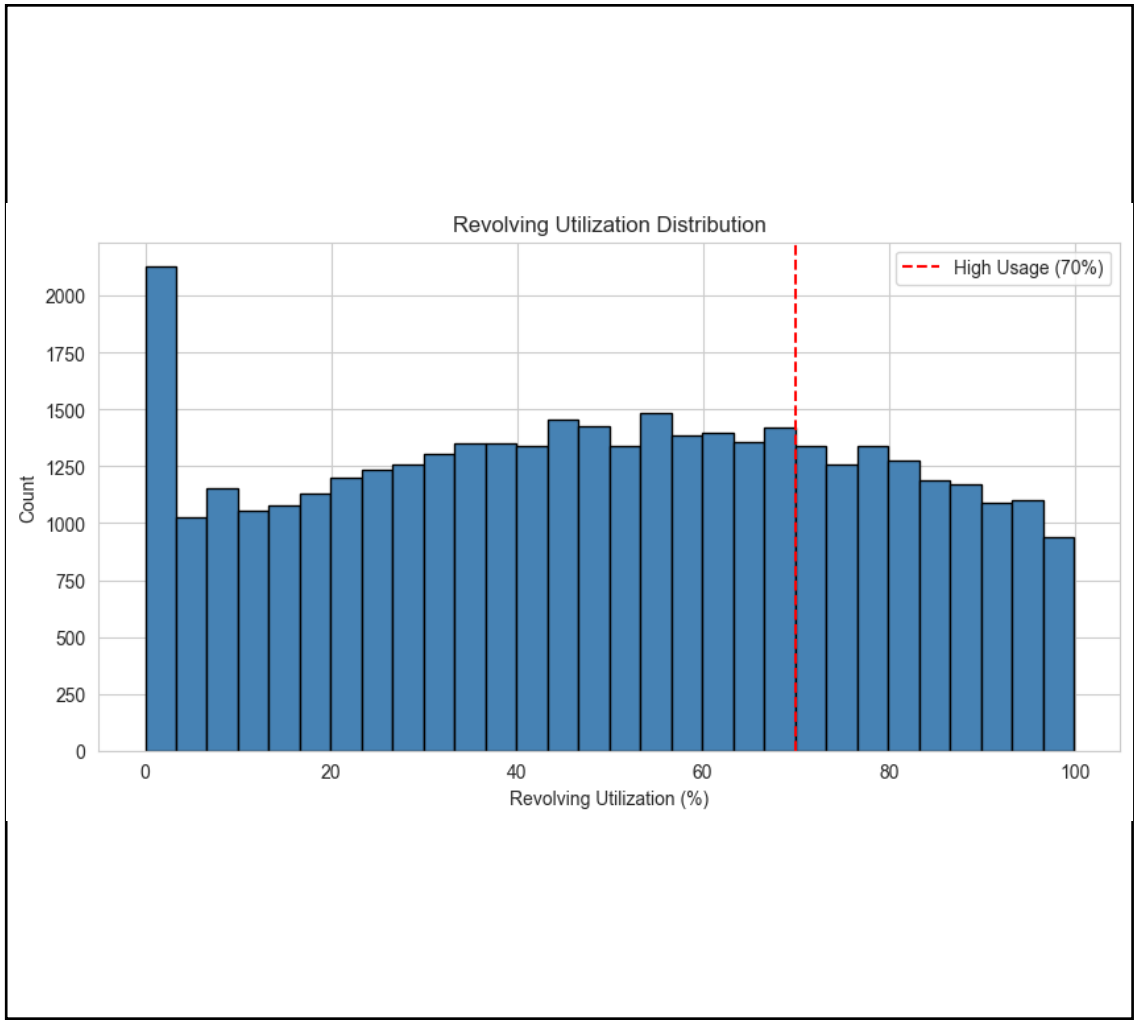
Annual Income Distribution (capped) (Notebook cell 118)



DTI Distribution (Notebook cell 121)

Univariate Insights - Credit Utilization

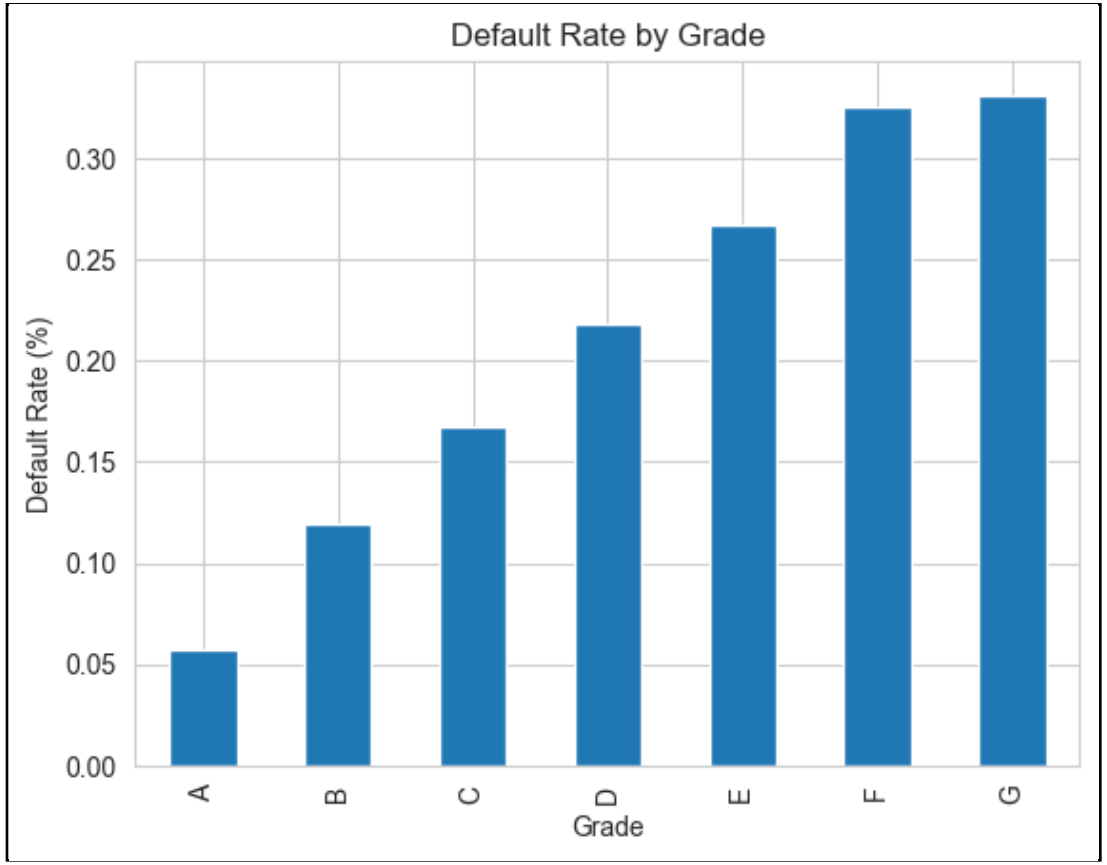
- Utilization shows a wide spread; higher utilization generally signals financial stress.
- Used later as a strong driver in bivariate/multivariate analysis.



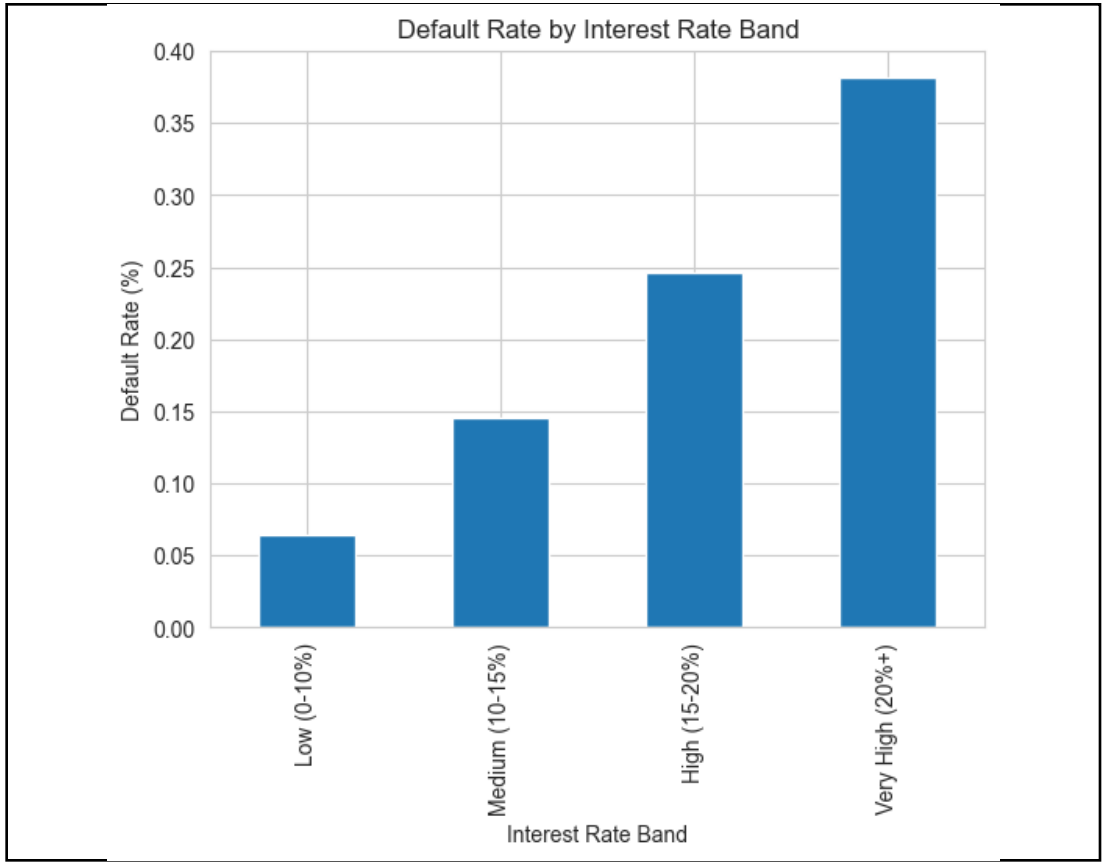
Revolving Utilization Distribution (Notebook cell 140)

Bivariate Drivers - Grade & Interest Rate

- Default rate increases monotonically from Grade A (~5.5%) to Grade G (~33%) - grading is strongly predictive.
- Interest rate is a top indicator: defaults rise from ~6% (0-10%) to ~38% (20%+).



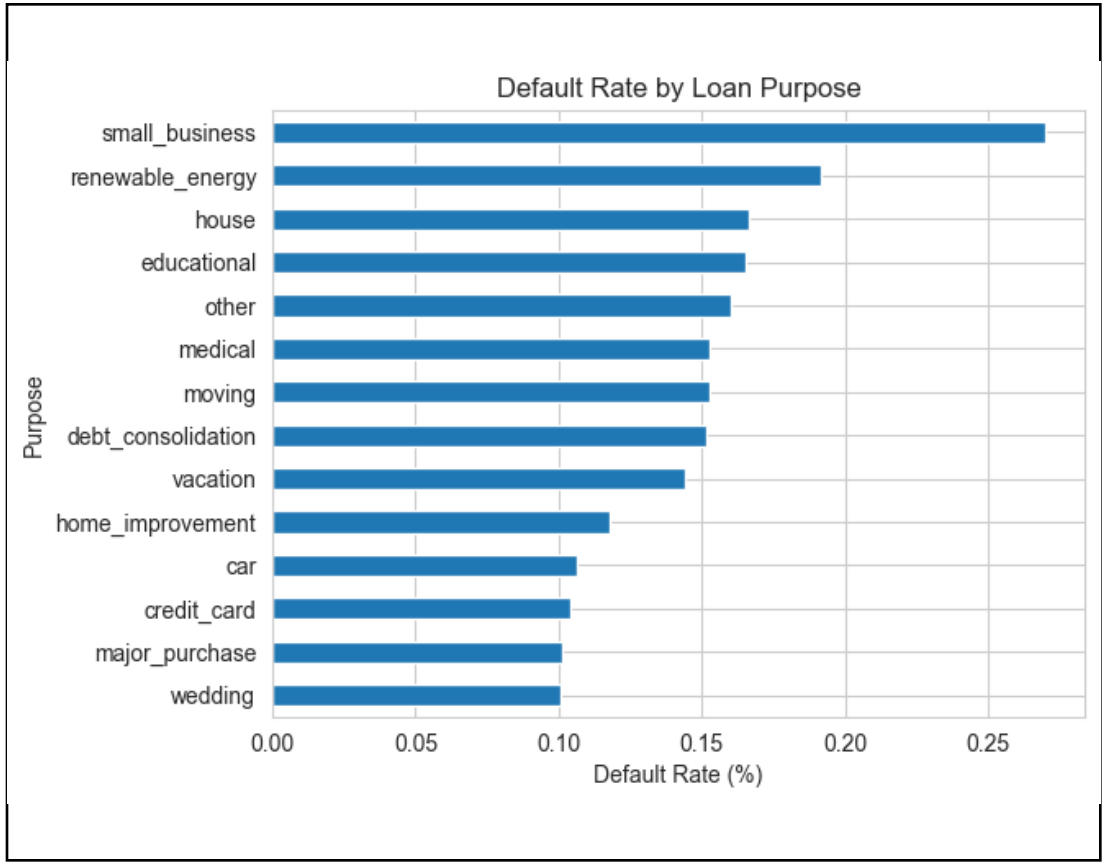
Default Rate by Grade (Notebook cell 154)



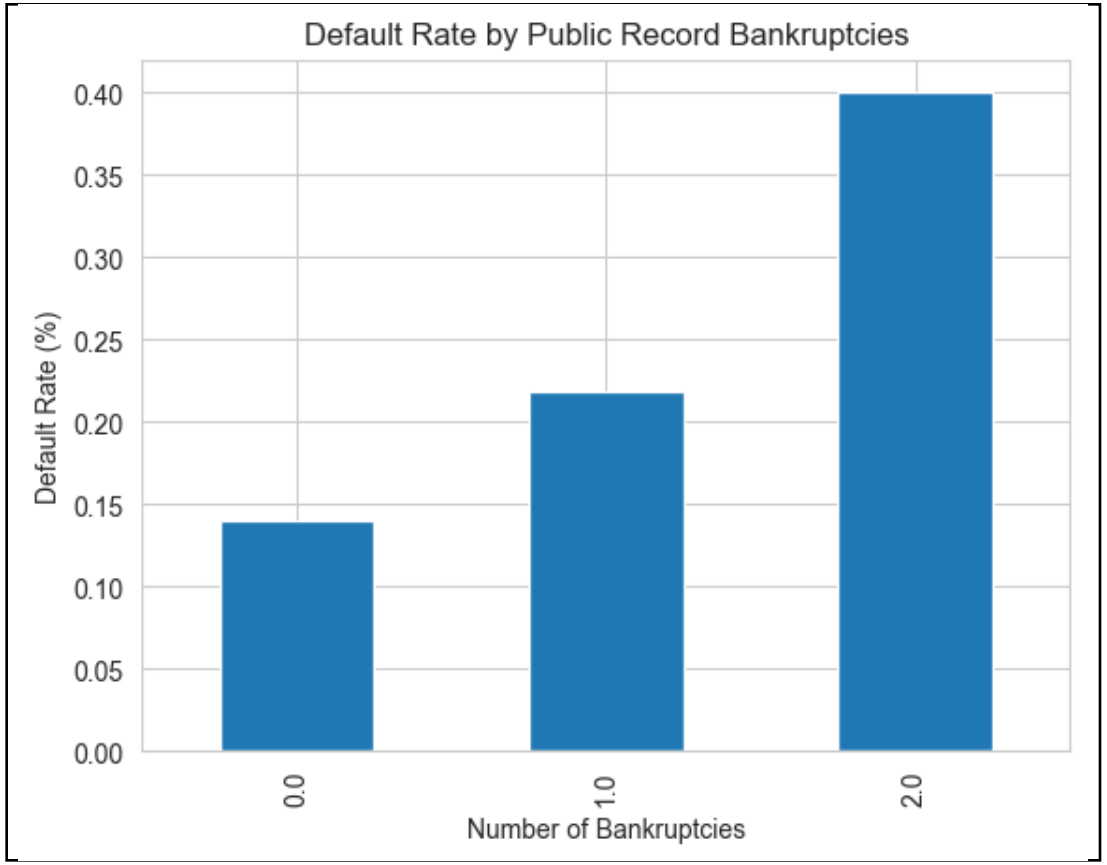
Default Rate by Interest Rate Band (Notebook cell 170)

Bivariate Drivers - Purpose & Public Records

- Loan purpose matters: small_business has the highest default (~27%), while wedding/major_purchase are among the lowest (~10%).
- Past bankruptcies are highly predictive: default rises from ~14% (0) to ~22% (1) to ~40% (2).



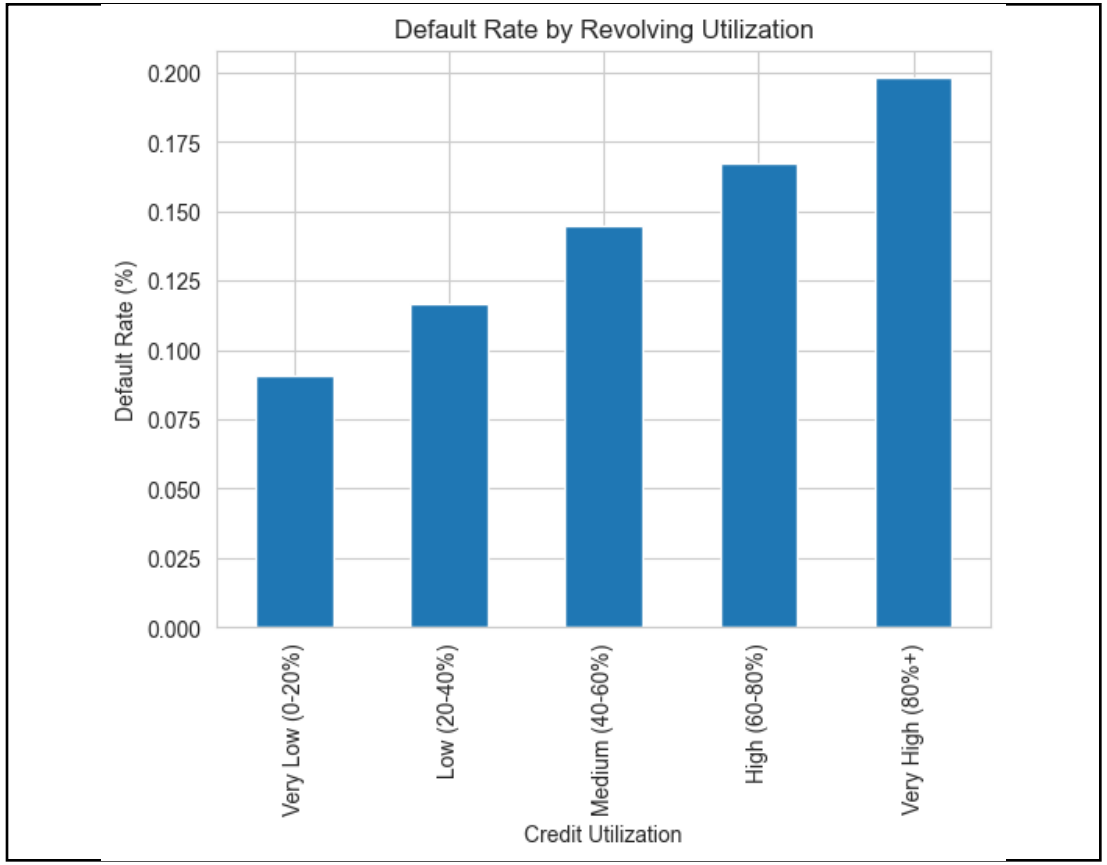
Default Rate by Loan Purpose (Notebook cell 157)



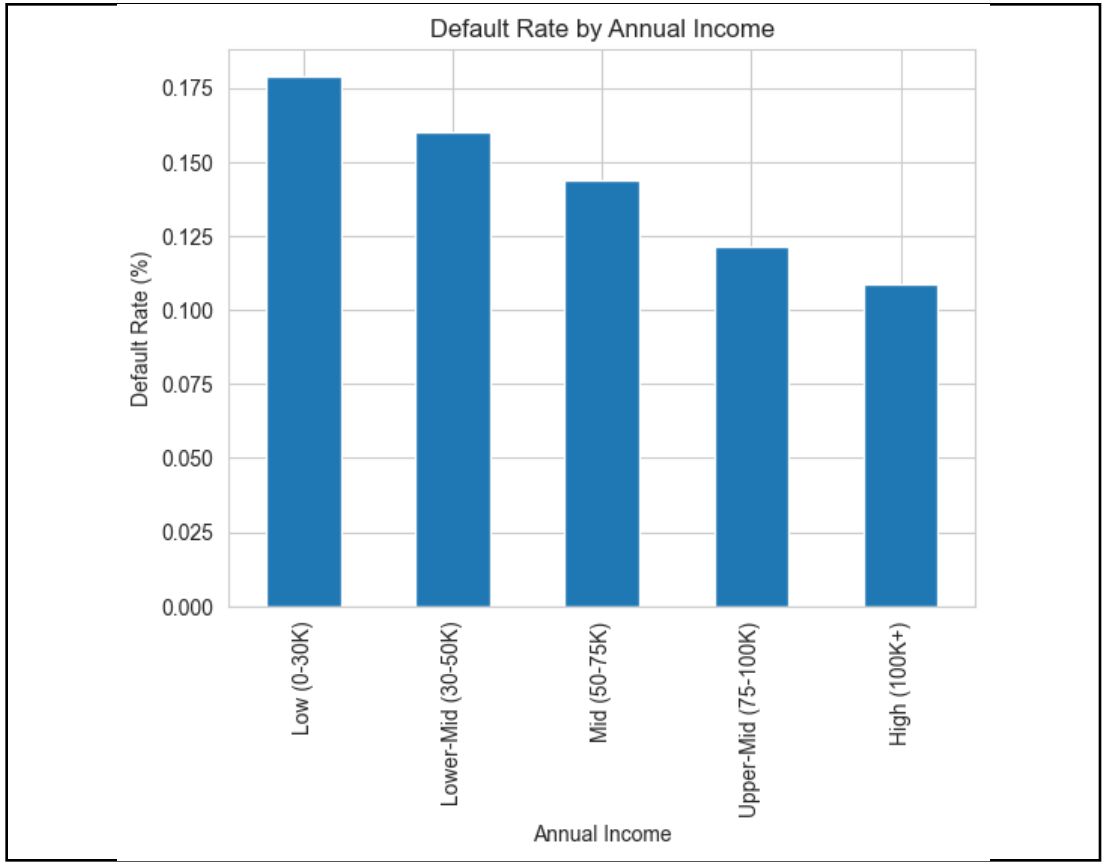
Default Rate by Public Record Bankruptcies (Notebook cell 189)

Bivariate Drivers - Utilization & Income

- Credit utilization is strongly linked to default: default rate rises from ~9% (0-20%) to ~20% (80%+).
- Income shows an inverse trend: low income (0-30K) ~18% default vs 100K+ ~11%.



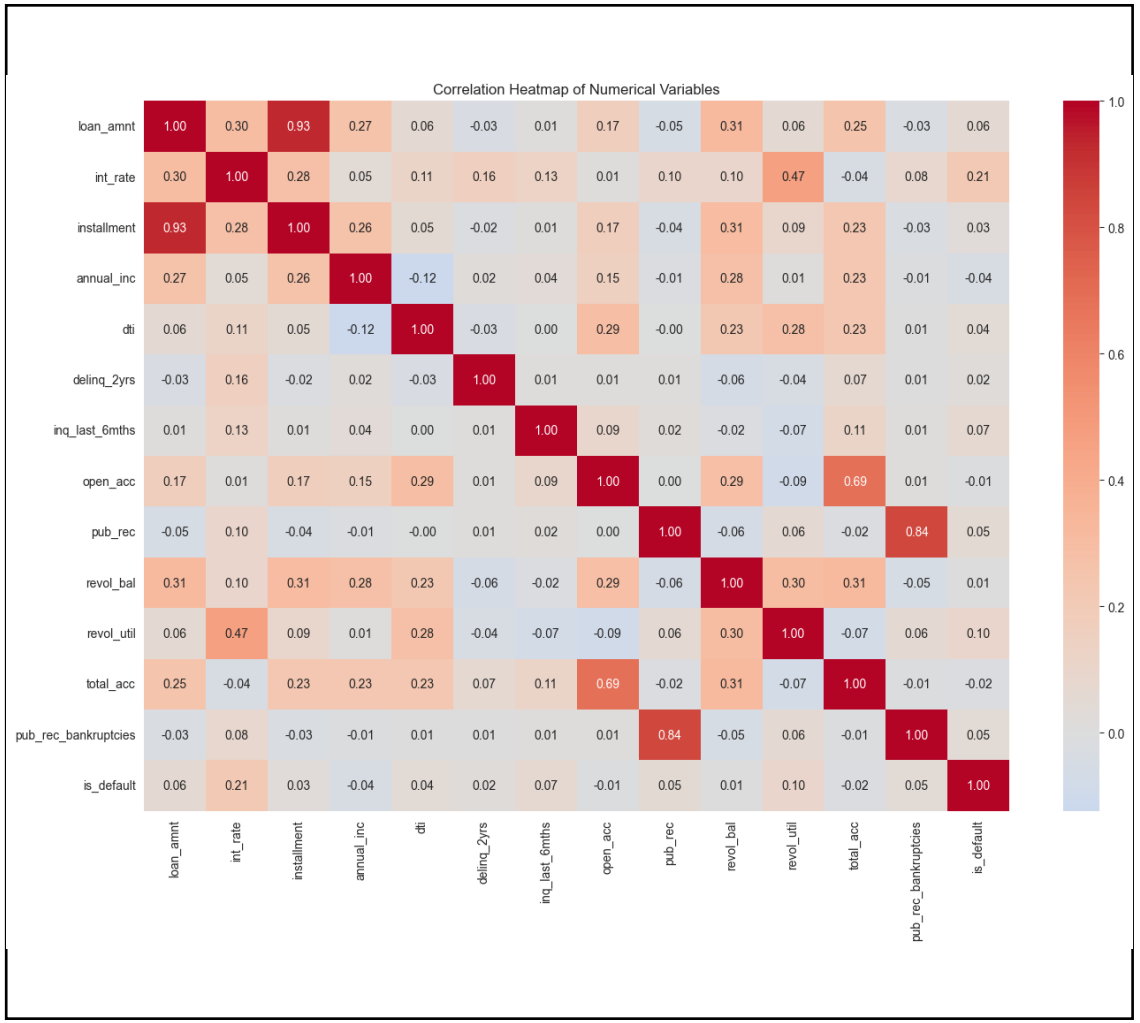
Default Rate by Revolving Utilization (Notebook cell 185)



Default Rate by Annual Income (binned) (Notebook cell 178)

Correlation Summary & Redundancy Check

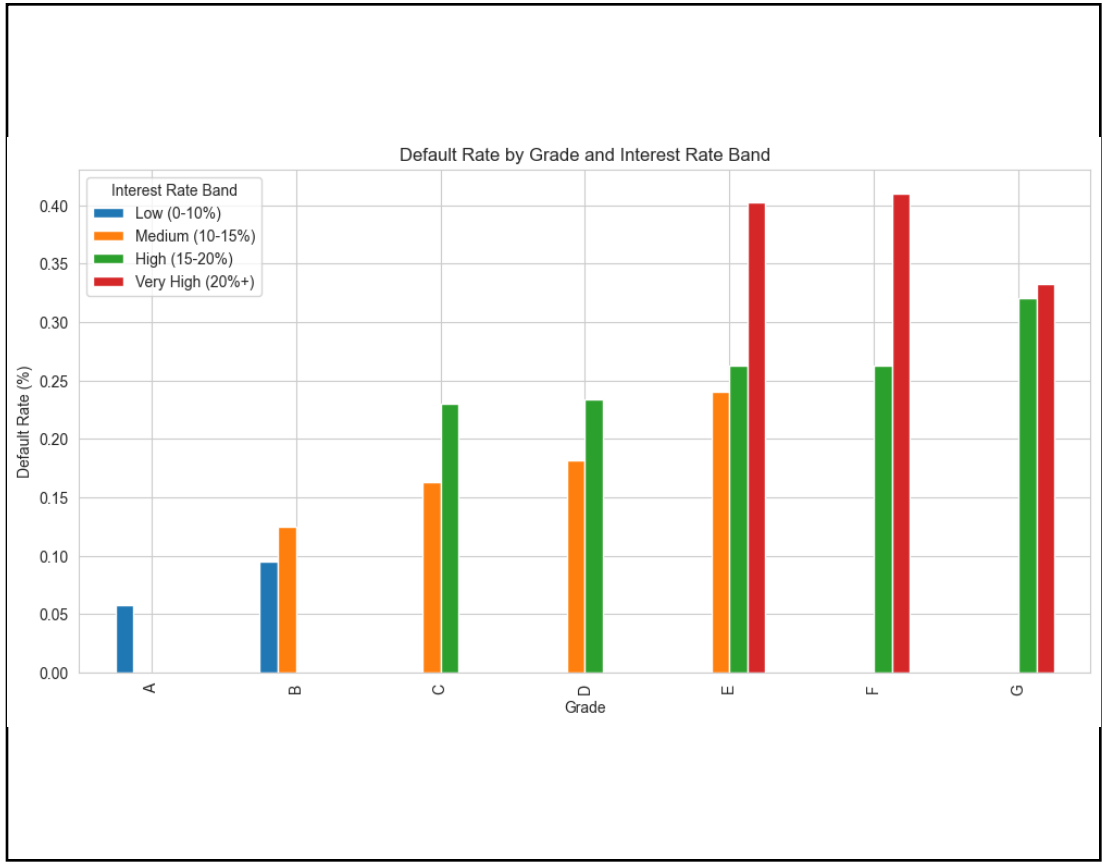
- Strongest correlations with default: int_rate (~0.21), revol_util (~0.10), pub_rec_bankruptcies (~0.05).
- Key multicollinearity: loan_amnt vs installment (~0.93), pub_rec vs pub_rec_bankruptcies (~0.84) - consider dropping redundant fields in modeling.
- Income has a small negative correlation (higher income tends to lower default).



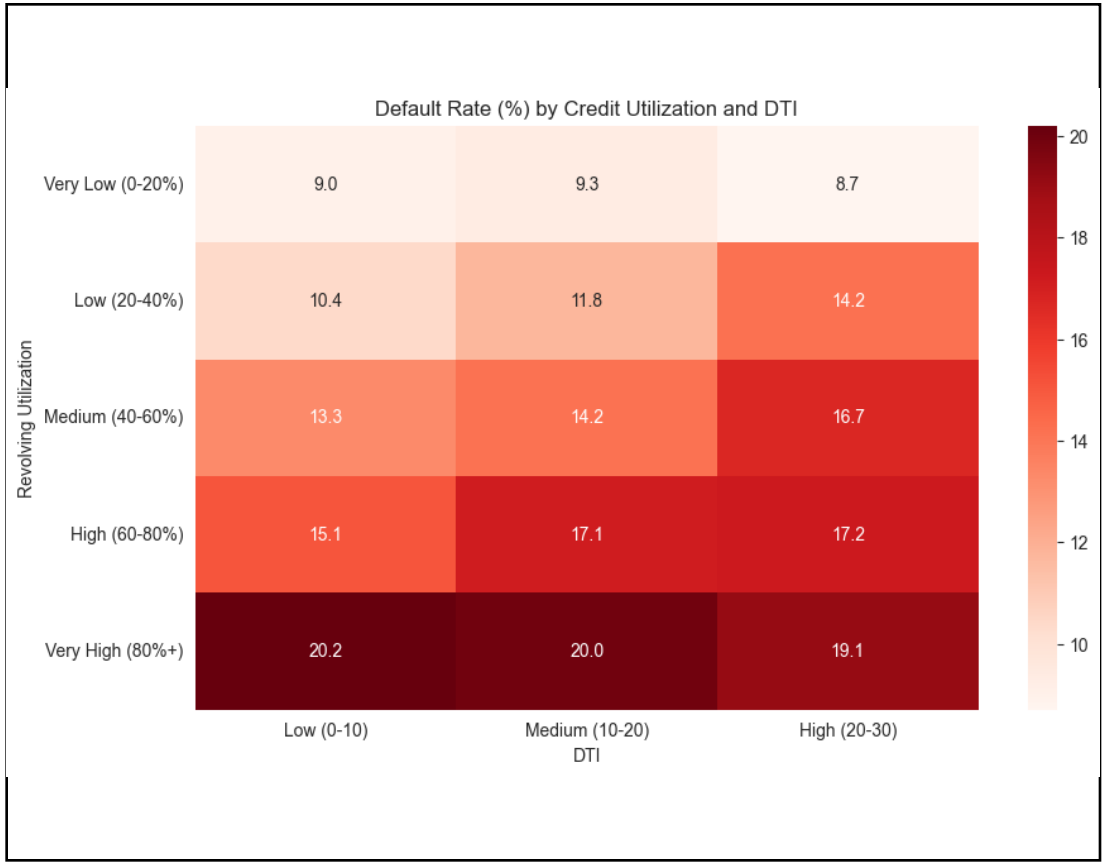
Correlation Heatmap of Numerical Variables (Notebook cell 192)

Multivariate Profiles - When Risk Factors Combine

- Grade and interest rate together amplify risk (e.g., Grade F with very high rate shows extremely high defaults).
- Utilization is dominant: very high revol_util (~80%+) shows ~20% default regardless of DTI band.



Default Rate by Grade & Interest Rate Band (Notebook cell 196)

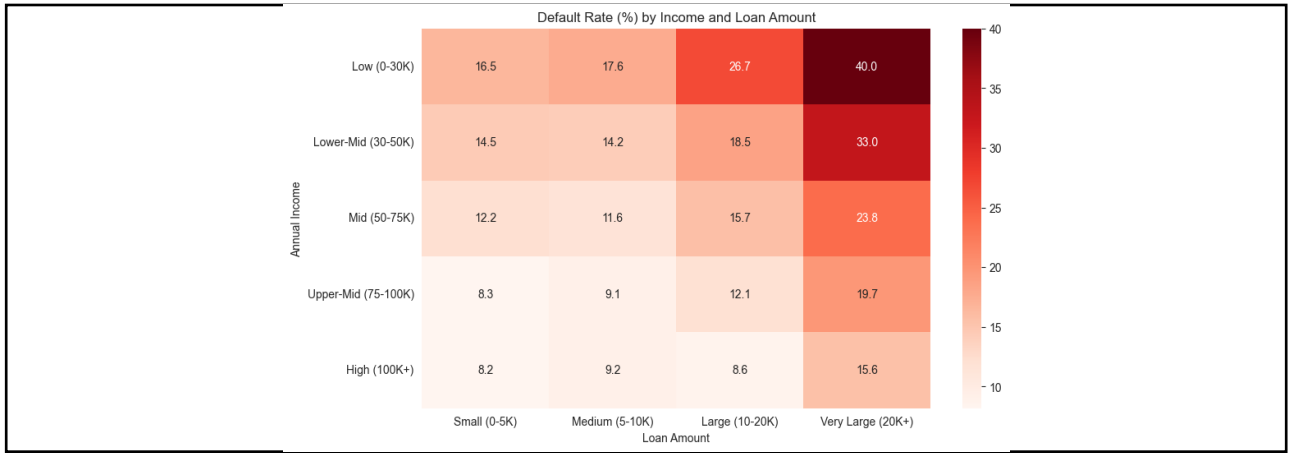


Default Rate by Credit Utilization & DTI (Notebook cell 208)

Multivariate Profiles - Income vs Loan Amount

- High-risk zone: low income combined with large loan amount (defaults cluster there).
- Heatmap highlights the worst combo: income 0-30K + loan 20K+ ~40% default.
- Actionable rule: set loan limits relative to income to avoid over-burdening borrowers.

Annual Income vs Loan Amount (colored by default) (Notebook cell 204)



Default Rate by Income & Loan Amount (Notebook cell 206)

Key Risk Drivers & Recommendations

- Top drivers of default (from EDA): Grade/Sub-grade, Interest Rate, Loan Purpose, Public Records/Bankruptcies, Revolving Utilization, DTI, and Income.
- Risk hotspots:
- Grades F/G (often >30% default)
  - Interest rate 20%+ (very high default)
  - Small Business purpose (high default)
  - 1+ bankruptcies (sharp increase in default)
- Recommendations:
- Cap/limit exposure for Grades F & G or require collateral/guarantees
  - Apply specialized underwriting for Small Business loans
  - Re-evaluate pricing: even 20%+ loans show very high default; consider declining some segments instead of only pricing higher
  - Tighten DTI thresholds for lower income applicants and enforce income-based loan limits

Assumptions: Only Fully Paid vs Charged Off used; Current loans excluded; percentile capping used for visualization.

Appendix - Plot Index (What to paste where)

Slide	Notebook cell	Plot / graph to use
3	92	Loan Status Distribution
5	99	Interest Rate Distribution
5	110	Loan Grade Distribution
6	118	Annual Income Distribution (capped)
6	121	DTI Distribution
7	140	Revolving Utilization Distribution
8	154	Default Rate by Grade
8	170	Default Rate by Interest Rate Band
9	157	Default Rate by Loan Purpose
9	189	Default Rate by Public Record Bankruptcies
10	185	Default Rate by Revolving Utilization
10	178	Default Rate by Annual Income (binned)
11	192	Correlation Heatmap of Numerical Variables
12	196	Default Rate by Grade & Interest Rate Band
12	208	Default Rate by Credit Utilization & DTI
13	204	Annual Income vs Loan Amount (colored by default)
13	206	Default Rate by Income & Loan Amount

Tip: If you want to replace images, export the corresponding figure from the notebook cell and paste over this area.