Exploratory Analysis of Borrower Characteristics and Loan Conditions

EDA & SQL Based Project

Ironhack Bootcamp: Jan 2026
Harmandeep Singh, Isis Hassan

Table of Contents

Project Overview

Objective

Explore relationships between **borrower characteristics** *(credit score, annual income, and employment status)* and **loan attributes** *(interest rate, loan amount, and loan purpose)* using exploratory data analysis **(EDA)** and **SQL-based** queries.

Project Scope

This project focuses on exploratory data analysis (EDA) and SQL-based descriptive insights to identify patterns and relationships in loan data. The analysis is observational and descriptive in nature. It does not include predictive modeling, classification, causal inference, or risk scoring.

Research Questions:

- RQ1: How does the **credit score** have influence on deciding **Loan attributes**?
  - *How does credit score relate to interest rate?*
  - *How does credit score relate to loan amount?*
  - *How does loan purpose vary across different credit score ranges?*
- RQ2: Annual Income and Loan Size
  - How does a borrower's **annual income** relate to the **loan amount** issued?
- RQ3: Employment Status and Loan Characteristics
  - How do loan amounts differ by employment status?
  - How do interest rates vary across employment statuses?
  - How does loan purpose distribution differ by employment status?

Dataset Overview

- The dataset was collected from the **Kaggle** platform and is titled **"Predicting Loan Payback."**
- It was originally created for a **machine learning competition** focused on loan payback prediction.
- The dataset is **synthetic**, generated using **deep learning techniques**, and ideal for predictive loan payback risk analysis.
- It contains **593,994 observations (rows)** and **13 variables (columns)**.

categorical variables	*numerical variables*
• id, • gender • marital_status • education_level • employment_status • loan_purpose • grade_subgrade	• annual_income • debt_to_income_ratio • credit_score • loan_amount • interest_rate • loan_payback *(boolean, target for ML)*

Data Assumption: The dataset, although **synthetic**, closely **reflects real-world loan and borrower behavior**.

	id	annual_income	debt_to_income_ratio	credit_score	loan_amount	interest_rate	gender	marital_status	education_level	employment_status	loan_purpose	grade_subgrade	loan_paid_back
0	0	29367.99	0.084	736	2528.42	13.67	Female	Single	High School	Self-employed	Other	C3	1.0
1	1	22108.02	0.166	636	4593.10	12.92	Male	Married	Master's	Employed	Debt consolidation	D3	0.0
2	2	49566.20	0.097	694	17005.15	9.76	Male	Single	High School	Employed	Debt consolidation	C5	1.0
3	3	46858.25	0.065	533	4682.48	16.10	Female	Single	High School	Employed	Debt consolidation	F1	1.0
4	4	25496.70	0.053	665	12184.43	10.21	Male	Married	High School	Employed	Other	D1	1.0

Data Cleaning

Data Cleaning Steps

- Duplicate Rows
- Missing values & Null Values
- Drop Columns
- Handling Unique Column

Drop Columns

categorical variables	*numerical variables*
id (--> index)	annual_income
~~gender~~	~~debt_to_income_ratio~~
~~marital_status~~	credit_score
~~education_level~~	loan_amount
employment_status	interest_rate
loan_purpose	~~loan_payback *(boolean, target for ML)*~~
~~grade_subgrade~~	

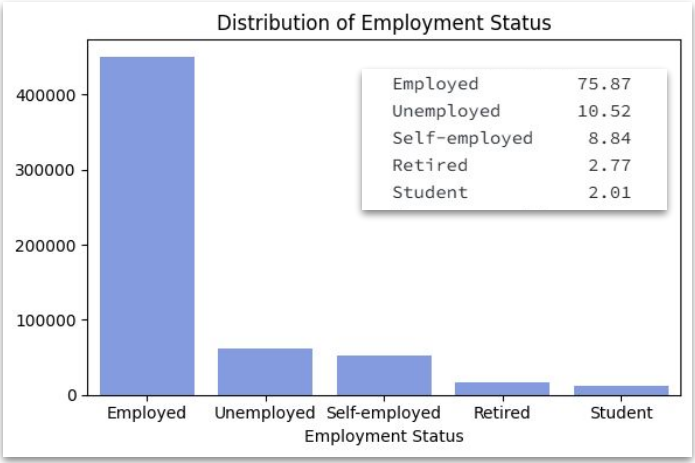
loan characteristic	~~*dropped column*~~
borrower characteristic	*Set it it as Index*

Handle `Id`Column

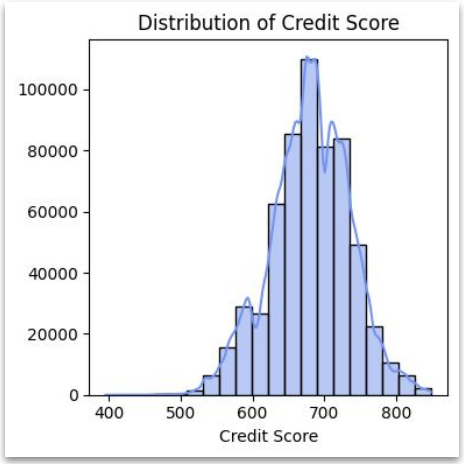
	id	annual_income
0	0	29367.99
1	1	22108.02
2	2	49566.20
3	3	46858.25
4	4	25496.70

id	annual_income
0	29367.99
1	22108.02
2	49566.20
3	46858.25
4	25496.70

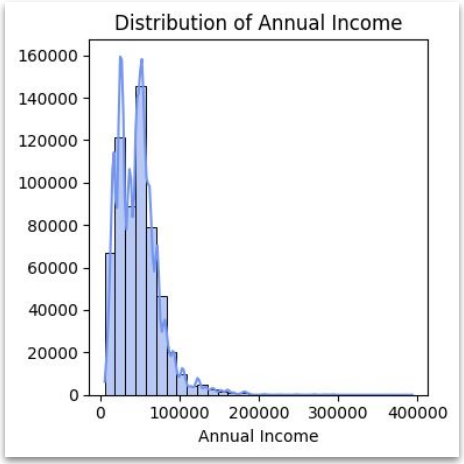
Univariate Analysis: Borrower Characteristics



- Employment status is **highly imbalanced**.
- Most loans are issued to **employed customers around 76%**.
- **Unemployed** customer received more loans as compared to other categories.

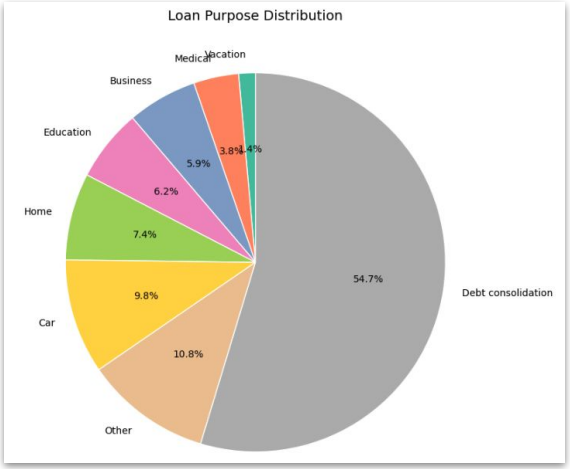


- **Credit Score Range:** Approximately **650–720.**
- **Distribution Shape: Symmetric**
- **Skewness:** Near zero (**−0.17**), supporting symmetry.

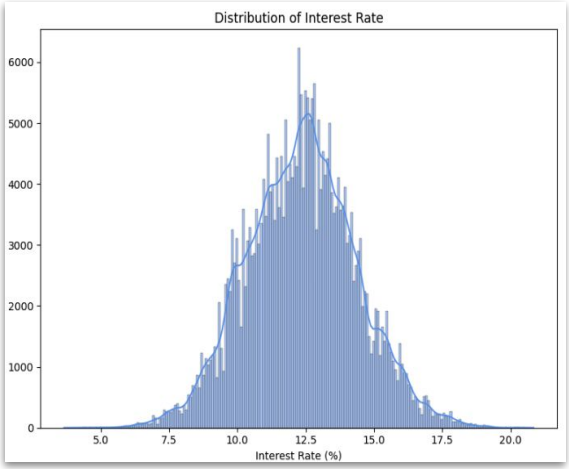


- **Mean annual income:** ~$48.2K.
- **Strong right-skewed** distribution.
- Supported by **skewness = 1.72** and **kurtosis = 7.09**
- Most borrowers fall in the **low-to-middle income range.**

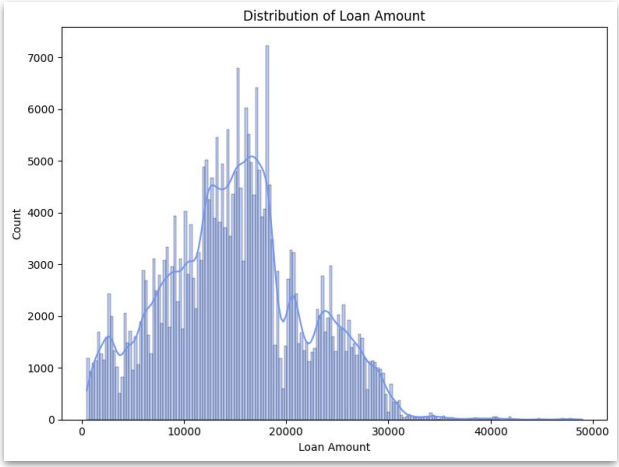
Univariate Analysis: Loan Characteristics



- Loan purpose is highly concentrated in **Debt Consolidation** (54.7%).
- Other loan purposes each represent a **much smaller share.**
- **Vacation, medical, and business** loans are the least common.

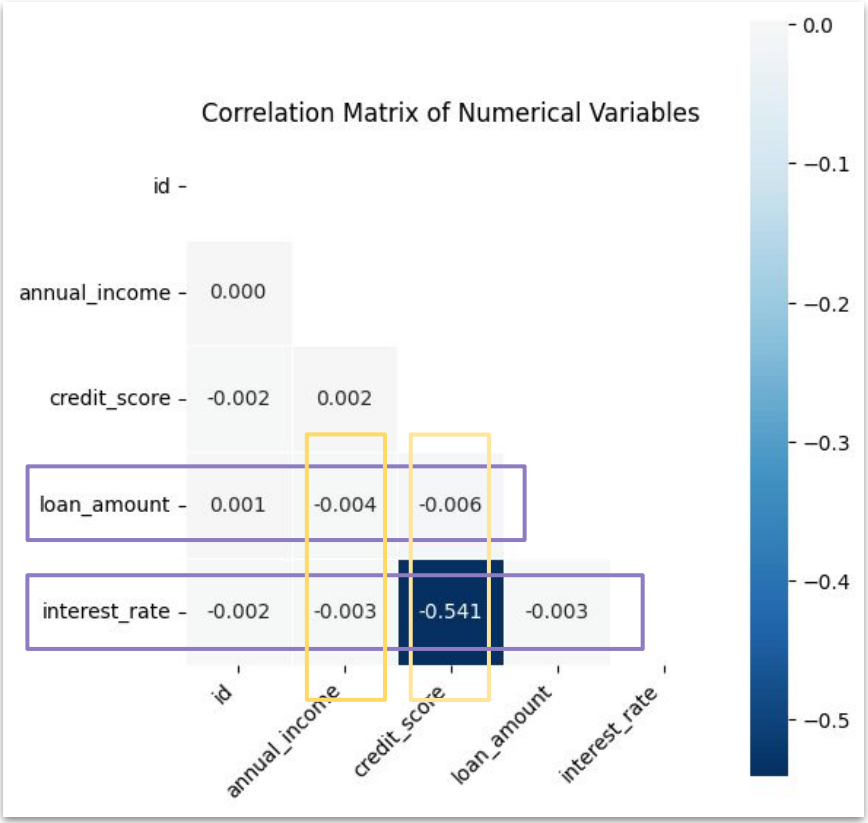


- **Median interest rate:** ~12.4%
- Most rates fall between **11%–14%.**
- **Nearly symmetric** distribution (skewness = 0.05).
- **Stable spread**, no extreme anomalies.



- **Typical loan amount:** ~$15,000 (mean ≈ median)
- Most loans fall between **$10k–$20k**
- **Nearly symmetric** distribution (skewness = 0.21)
- **No extreme tails** (kurtosis = −0.15)

Bivariate Analysis: Numerical Correlation



- Most numerical variables show **weak or no linear relationships**.

RQ1: Credit Score and Loan Attributes

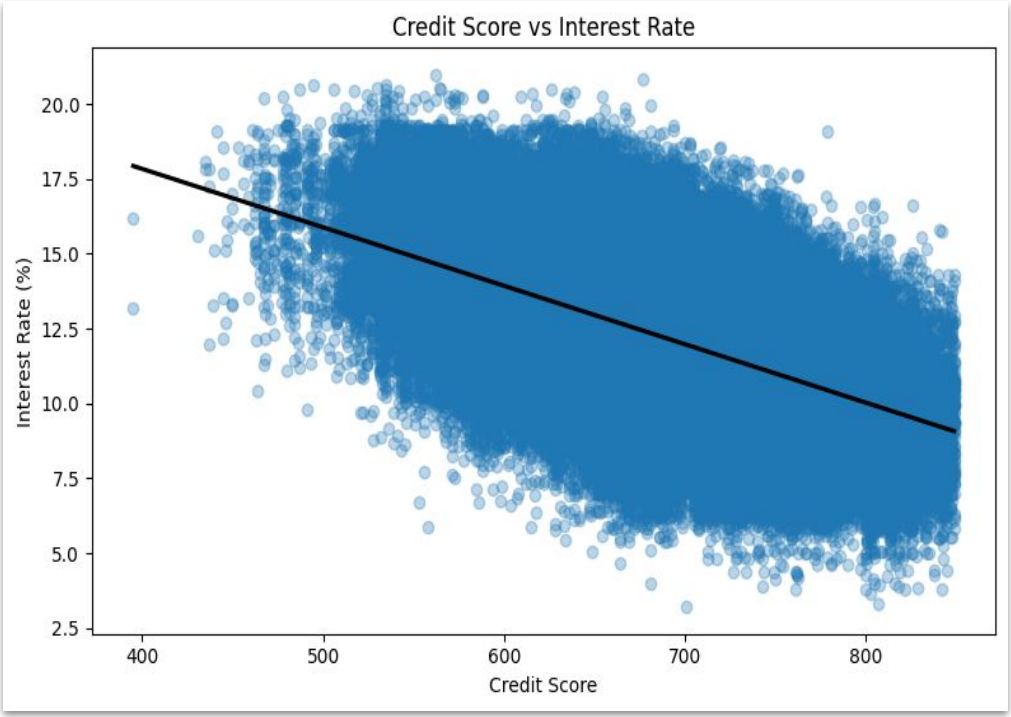
- **Credit score vs interest rate:**
  Negative correlation ($\rho \approx -0.54$), indicating higher credit scores are associated with lower interest rates.
- **Credit score vs loan amount:**
  Correlation is **near zero**, suggesting no meaningful linear relationship.
- **Credit score vs loan purpose:**
  This we see further.

RQ2: Annual Income and Loan Amount

- Annual income shows **negligible correlation** with loan amount.
- Additional Analysis with bins in slide 11.

Key observation:
Credit Score and **Interest rate** exhibited negative relationship.

Bivariate: Credit Score vs Interest Rate



- The scatter plot with regression line **confirms** that **credit score** and **interest rate** are **negatively** related.

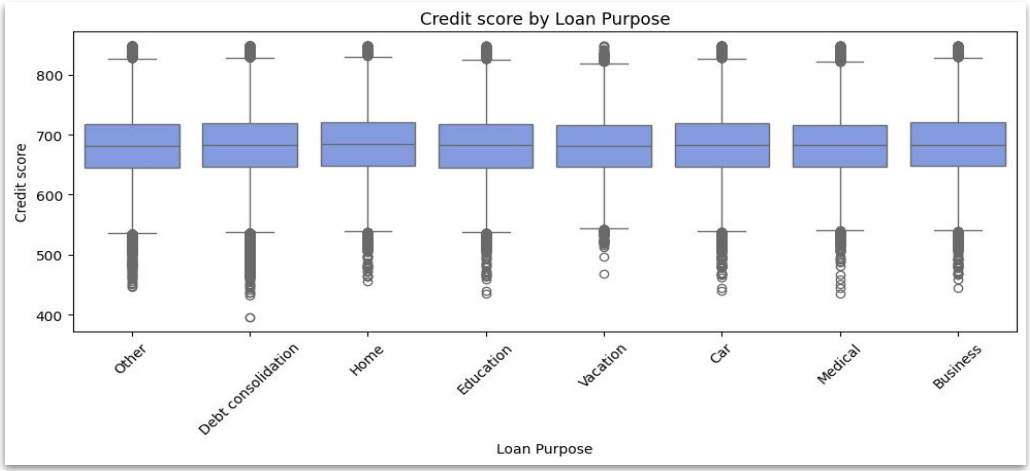
What negatively related means?

- It means higher credit scores are associated with lower interest rates, while lower credit scores correspond to higher rates.

Conclusion: Credit score is **inversely** related with **interest rate**. It is the first borrower's characteristics which affects loan characteristics.

Bivariate: Credit Score vs Loan Purpose

loan_purpose	mean	std	min	25%	50%	75%	max
Business	682.19	55.17	445.0	648.0	683.0	720.0	849.0
Car	680.87	55.67	440.0	647.0	683.0	719.0	849.0
Debt consolidation	680.94	55.47	395.0	646.0	682.0	719.0	849.0
Education	679.68	54.80	435.0	645.0	682.0	717.0	849.0
Home	682.46	55.48	456.0	648.0	684.0	721.0	849.0
Medical	679.87	54.87	435.0	646.0	682.0	716.0	849.0
Other	680.25	55.70	446.0	645.0	681.0	718.0	849.0
Vacation	680.23	54.75	468.0	647.0	681.0	716.0	849.0

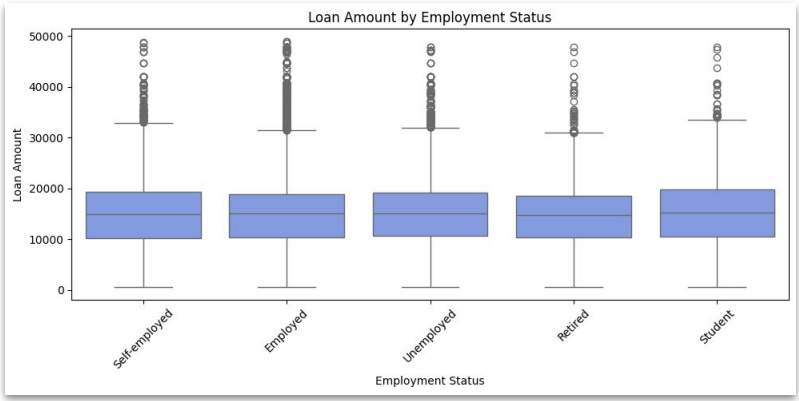


- Credit score distributions are **very similar across all loan purposes**.
- Median credit scores are clustered around **~680–684** for every category.
- IQR ranges largely **overlap**, indicating no meaningful separation by purpose.
- Minor differences in means are **not practically significant.**

Conclusion: Loan purpose does **not vary meaningfully** across different credit score ranges. So it does not strongly influence loan purpose selection.

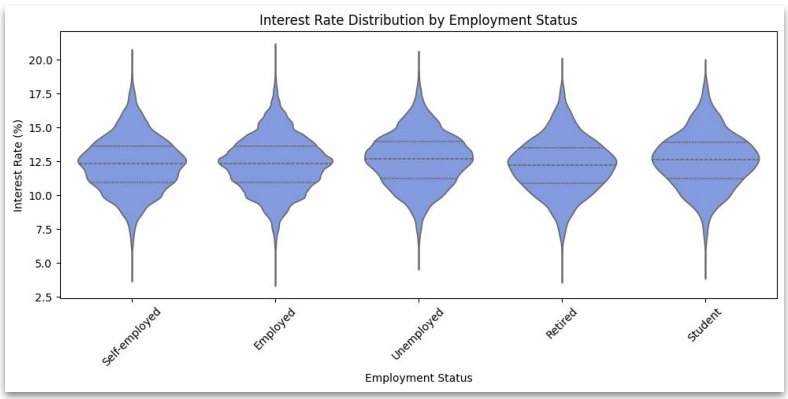
Bivariate: Employment Status Vs Loan Characteristics

employment_status	count	mean	std	min	25%	50%	75%	max
Employed	450645.0	15005.0	6915.0	500.0	10270.0	15002.0	18759.0	48960.0
Retired	16453.0	14818.0	6828.0	501.0	10198.0	14652.0	18389.0	47851.0
Self-employed	52480.0	15048.0	6998.0	509.0	10150.0	14864.0	19243.0	48832.0
Student	11931.0	15290.0	7085.0	517.0	10364.0	15207.0	19490.0	47871.0
Unemployed	62485.0	15110.0	6943.0	514.0	10559.0	15011.0	19130.0	47865.0



Employment Status vs Loan Amount

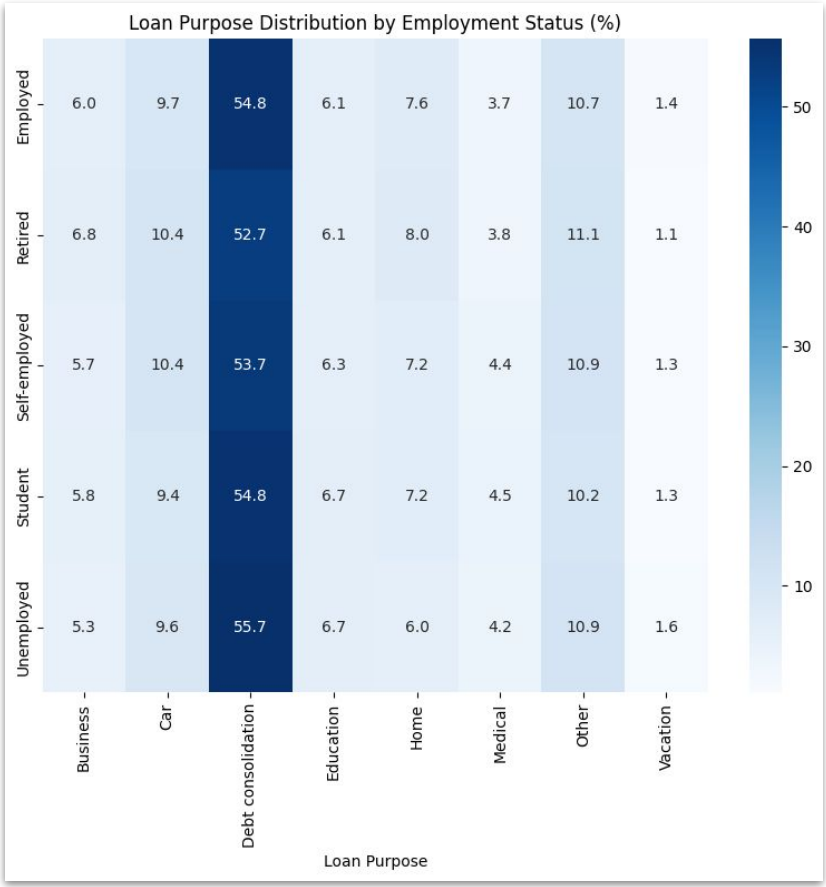
employment_status	count	mean	std	min	25%	50%	75%	max
Employed	450645.0	12.33	2.01	3.32	10.97	12.35	13.64	20.99
Retired	16453.0	12.23	2.00	4.12	10.89	12.24	13.54	20.19
Self-employed	52480.0	12.30	2.03	3.20	10.93	12.31	13.63	20.28
Student	11931.0	12.59	2.00	4.57	11.24	12.61	13.89	19.30
Unemployed	62485.0	12.62	1.98	3.89	11.24	12.65	13.93	20.24



Employment Status vs Interest Rate

Conclusion: The boxplots and violin chart show that loan amounts and interest rate are evenly distributed across employment status respectively. So it has **no impact** on loan conditions.

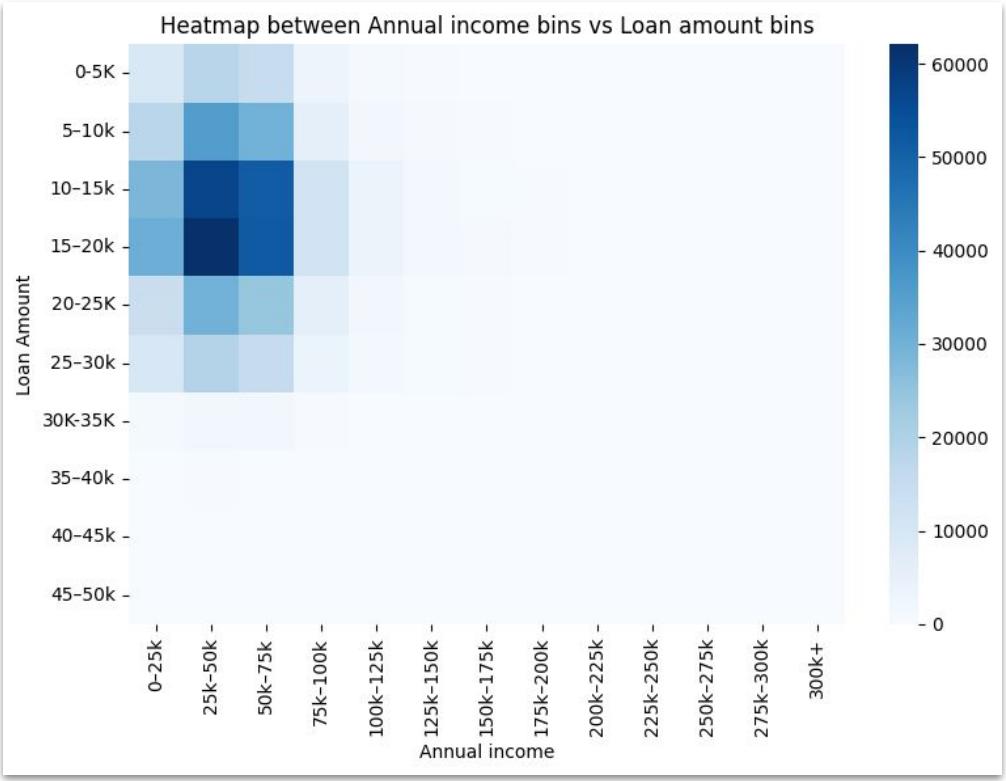
Bivariate: Employment Status Vs Loan Purpose



Chi-Square Test & Cramer Effect Size

- The chi-square test indicates a statistically significant association between employment status and loan purpose ($p < 0.05$).
- However, the effect size measured by Cramér's V (0.015) is extremely small, indicating a very **weak association**.
- This suggests that while differences in loan purpose distribution across employment status groups are detectable statistically, they are minimal in practical terms.
- Overall, **employment status** explains very little variation in **loan purpose.**

Additional: Annual Income & Loan Amount Bins



- We created bins first for both annual income and loan amount and converted them to categorical features.
- Then we created crosstab between these variables and we visualised heatmap and conducted chi-test of independence.
- The chi-square test indicates a **statistically significant association** between employment status and loan purpose. However, the e**ffect size measured by Cramér's V (0.011)** is extremely small, indicating a v**ery weak association**.
- Now we can finally conclude that there is **weak association** relationship between Loan amount & Annual income.

Key Findings

- The dataset includes a **large and diverse group of borrowers**, with most loans issued to **employed individuals** and for **debt consolidation**.
- **Credit score affects interest rates**, showing a clear negative relationship: higher credit scores lead to lower interest rates.
- Credit score does **not meaningfully influence other loan characteristics** such as loan amount, loan purpose**.**
- **Annual income** shows **no linear relationship** with loan amount and does not independently determine loan size.
- **Employment status** has **no significant impact** on loan characteristics.
- **Loan purpose differs by employment status**, but the relationship is **weak**, as indicated by a low **Cramér's V** value.

Conclusion

- This exploratory analysis examined the relationship between **borrower characteristics** and **loan conditions**.
- For an MVP product, **credit score** is the only relevant variable required to draw conclusions on **interest rate.**
- A **clear negative relationship** was observed between credit score and interest rate, confirming that borrowers with higher credit scores receive lower interest rates.
- Other factors, including **annual income, and employment status** showed **negligible or no meaningful influence** on loan conditions.

Future Research:

- In SQL, we found indicators that “grade subgrade” and “debt to income ratio” are correlated with interest rate.
- Explore other dropped borrower characteristics can be studied such as education level.

Key Learnings:

- Split the work and discuss challenges and outcomes regularly.
- Not every dataset is useful. Do not assume synthetic data is fully realistic.
- In real-world projects, **start with SQL exploration**.
- **EDA is essential part of analysis. It gives you detail insights about your dataset.**

Thanks!

Questions?

Github Repo

Github Profile Harmandeep
Github Profile Isis

Appendix

SQL Questions

1. What is the no of loans, average loan amount, avg credit score and avg interest rate per employment status?
2. What is the no of loans, average loan amount, avg credit score and avg interest rate per loan purpose?
3. What is the no. of loans per loan type/employment status?
4. What is the count and sum of loans (in millions) per employee status?
5. What is the count and sum of loans per loan purpose?
6. What is the avg. loan to annual income ratio per employment status?
7. From which three loan purposes is the bank earning the most interest per year?
8. What is the no. of loans and sum of loans given per credit score category?
9. What is the avg. debt to income ratio per credit score category?
10. What is the min/max/avg interest rate per grade_subgrade?

SQL Solutions

Q1: What is the no of loans, average loan amount, avg credit score and avg interest rate per employment status?

	employment_status	no_of_loans	avg_loan	avg_credit_score	avg_interest_rate
1	Employed	450645	15004.98	682.48	12.33
2	Retired	16453	14817.76	687.58	12.23
3	Self-employed	52480	15047.6	685.36	12.3
4	Student	11931	15290.33	670.02	12.59
5	Unemployed	62485	15109.62	666.26	12.62

Q2: What is the no of loans, average loan amount, avg credit score and avg interest rate per loan purpose?

	loan_purpose	no_of_loans	avg_loan	avg_credit_score	avg_interest_rate
1	Business	35303	14928.27	682.19	12.33
2	Car	58108	14886.04	680.87	12.36
3	Debt consolidation	324695	15058.31	680.94	12.35
4	Education	36641	14916.96	679.68	12.4
5	Home	44118	14933.44	682.46	12.32
6	Medical	22806	15010.54	679.87	12.38
7	Other	63874	15129.14	680.25	12.38
8	Vacation	8449	14972.7	680.23	12.39

Q3: What is the no. of loans per loan type/employment status?

EmploymentStatus	Other	DebtConsolidation	Home	Education	Vacation	Car	Medical	Business
Employed	48351	246853	34279	27222	6493	43548	16673	27226
Retired	1844	8628	1328	1017	180	1744	643	1069
Self-employed	5740	28180	3728	3332	670	5529	2315	2986
Student	1187	6418	941	817	144	1214	507	703
Unemployed	6752	34616	3842	4253	962	6073	2668	3319

What is the count and sum of loans (in millions) per employee status?

	employment_status	count_loans	sum_loan_millions
1	Employed	450645	6761.9
2	Unemployed	62485	944.1
3	Self-employed	52480	789.7
4	Retired	16453	243.8
5	Student	11931	182.4

SQL Solutions

Q5: What is the count and sum of loans per loan purpose?

loan_purpose	count_loans	sum_loan_millions
Debt consolidation	324695	4889.4
Other	63874	966.4
Car	58108	865.0
Home	44118	658.8
Education	36641	546.6
Business	35303	527.0
Medical	22806	342.3
Vacation	8449	126.5

Q6: What is the avg loan to annual income ratio per employment status?

loan_purpose	count_loans	sum_loan_millions
Debt consolidation	324695	4889.4
Other	63874	966.4
Car	58108	865.0
Home	44118	658.8
Education	36641	546.6
Business	35303	527.0
Medical	22806	342.3
Vacation	8449	126.5

Q7: From which three loan purposes is the bank earning the most interest per year?

loan_purpose	avg_interest_rate	avg_loan_amount	avg_interest_paid_this_year_thousands
Other	12.38	15129.0	187.3K
Debt consolidation	12.35	15058.0	186.0K
Medical	12.38	15011.0	185.9K

Q8: What is the no. of loans and sum of loans given per credit score category?

credit_score_category	no_of_loans	sum_loan_millions
1 - No Risk	9852	149.1
2 - Low Risk	208890	3135.3
3 - Medium Risk	322196	4838.7
4 - High Risk	52650	792.8
5 - Extreme Risk	406	6.2

SQL Solutions: using additional variables

Q9: What is the avg debt to income ratio per credit score category?

	credit_score_category	avg_debt_to_income_ratio
1	1 -No Risk	11.9%
2	2 - Low Risk	11.6%
3	3 - Medium Risk	12.2%
4	4 - High Risk	12.9%
5	5 - Too Risky	13.5%

Q10: What is the min/max/avg interest rate per grade_subgrade?

	grade_subgrade	min_interest_rate	avg_interest_rate	max_interest_rate
1	A1	3.32	9.7	16.24
2	A2	3.66	9.55	16.61
3	A3	3.92	9.7	16.65
4	A4	3.81	9.6	16.56
5	A5	3.79	9.64	16.12
6	B1	4.26	10.87	19.07
7	B2	4.12	10.84	17.04
8	B3	3.81	10.87	17.8
9	B4	4.48	10.85	16.89
10	B5	3.89	10.83	17.69
11	C1	4.8	11.94	19.28
12	C2	4.36	11.93	20.84
13	C3	3.98	11.98	18.72
14	C4	3.2	11.94	19.98
15	C5	4.38	11.95	19.16
16	D1	5.76	13.01	20.48
17	D2	4.68	13.07	19.86