Lending Club Case Study

Team Members	
Dhawal Dethe	Swarajkumar Thummapudi

The Problem

Company-Consumer finance company which specializes in lending various types of loans to urban customers

Context-When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile

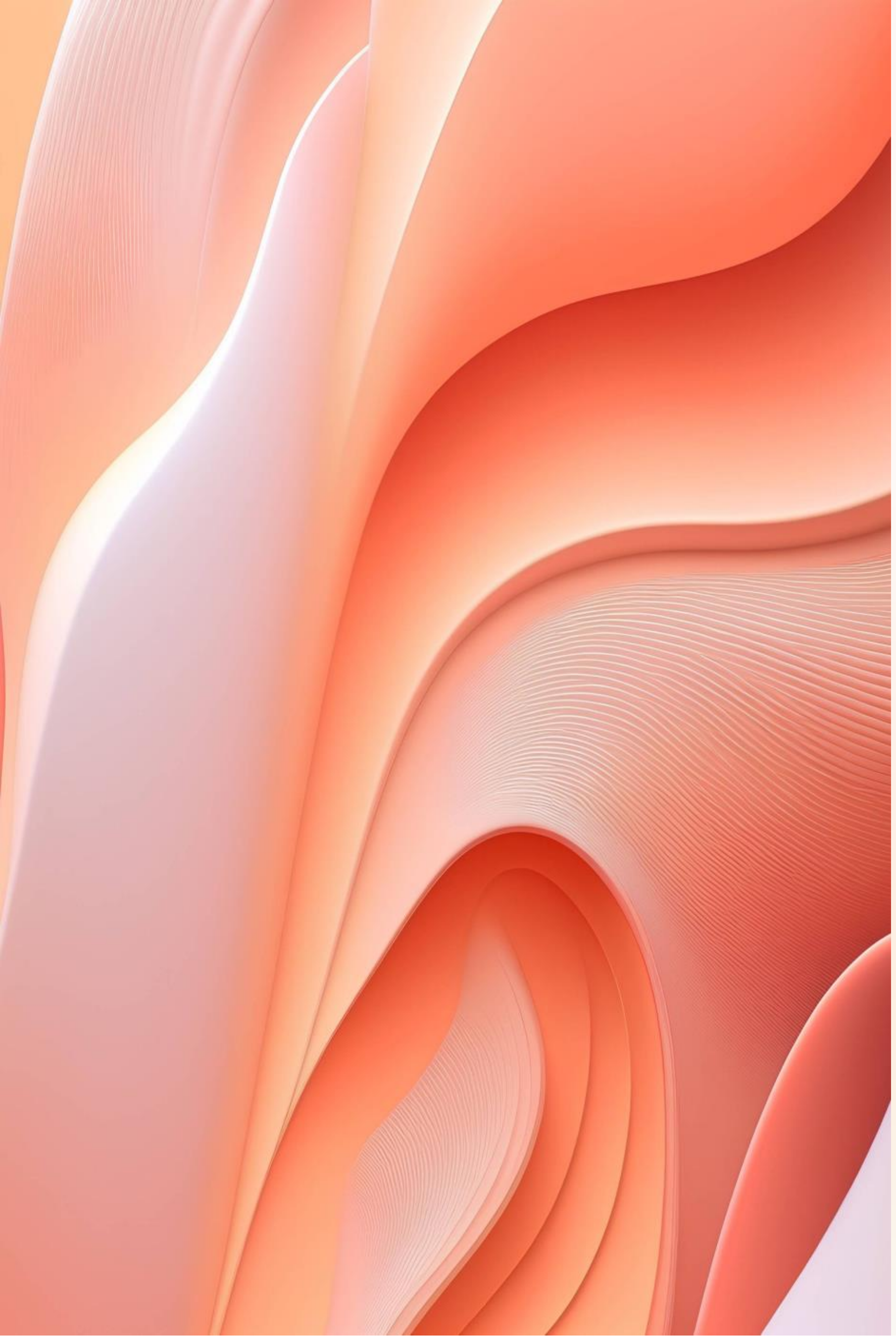
Problem Statement-The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default

Solution

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

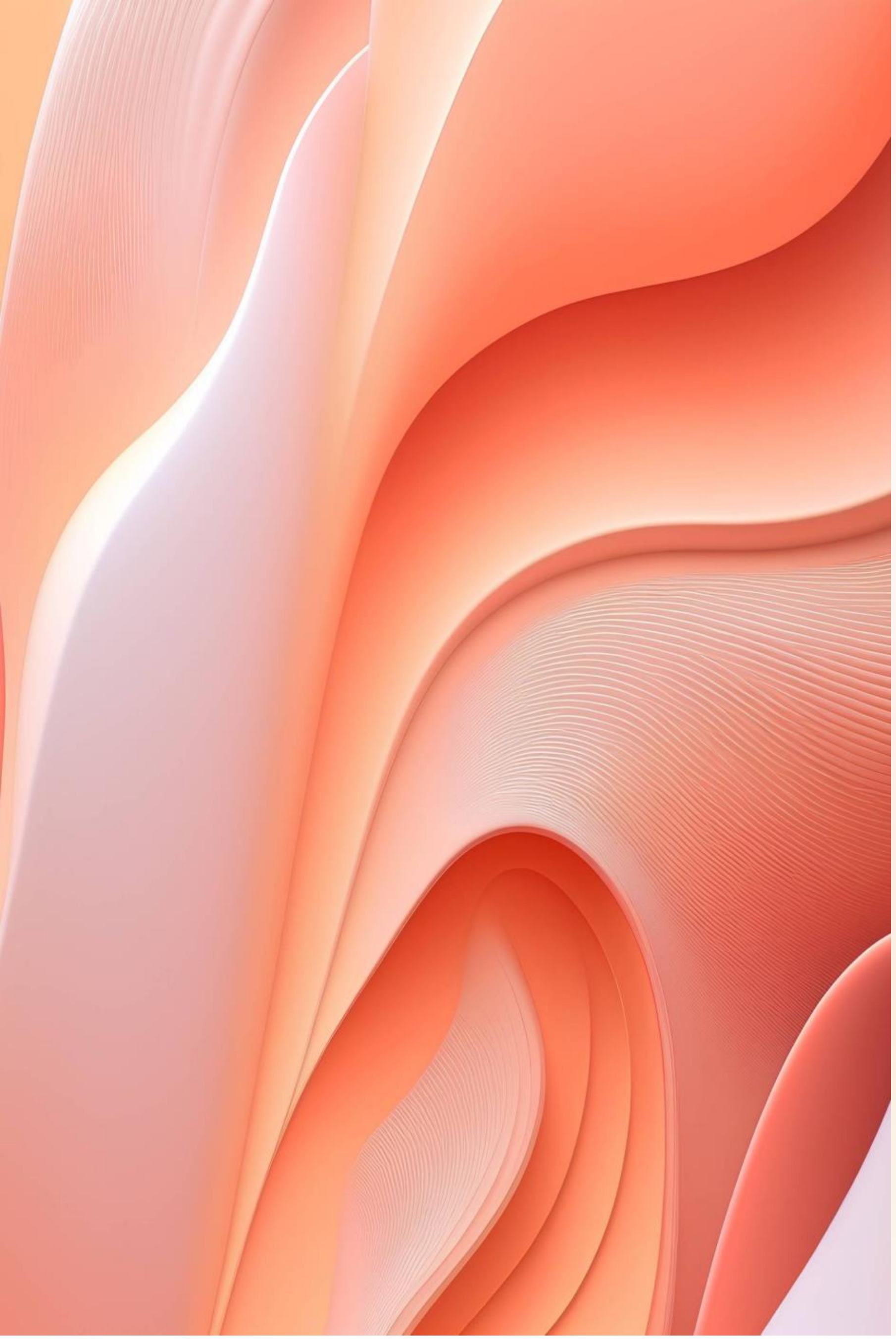
EDA Process

1. Data Understanding
2. Data Cleaning
3. Univariate Analysis
4. Segmented Univariate Analysis
5. Bivariate Analysis
6. Summary of Findings
7. Recommendations



Data Understanding

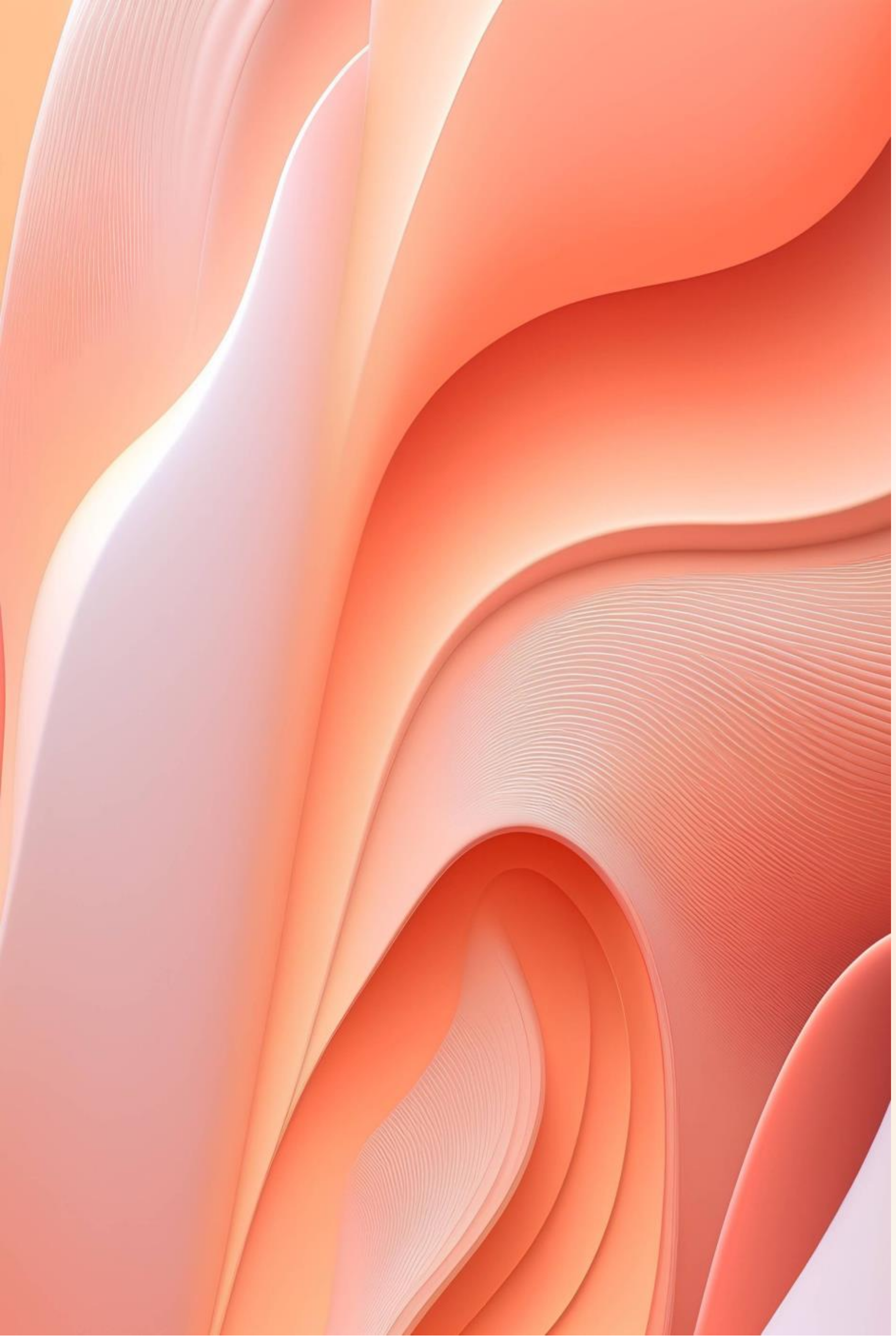
- Total records - 39717
- Attributes - 111
- Float64 Type of attributes - 74
- Int64 Type of attributes - 13
- object type of attributes - 24



Data Cleaning

Data cleaning is the process of identifying and correcting errors, inconsistencies, and missing values in a dataset to ensure data quality and reliability for analysis.Key steps are used for better readability

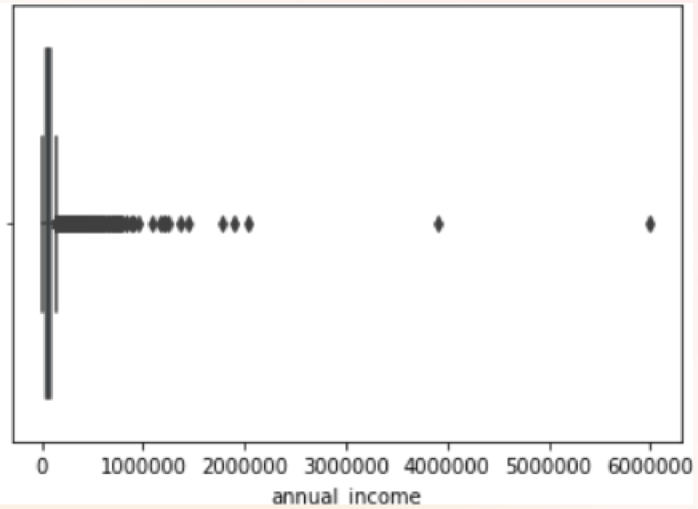
1. Handling Missing Data
2. Duplicate Data
3. Data Consistency
4. Data Types
5. Outliers
6. Data Transformation
7. Handling Categorical/Numerical Data

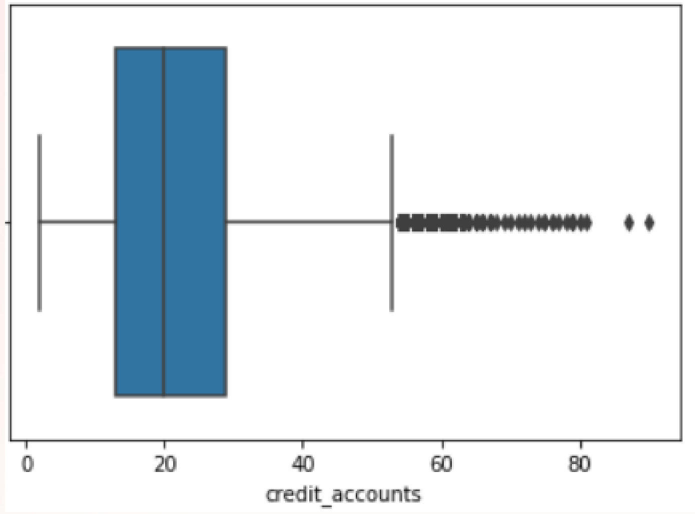


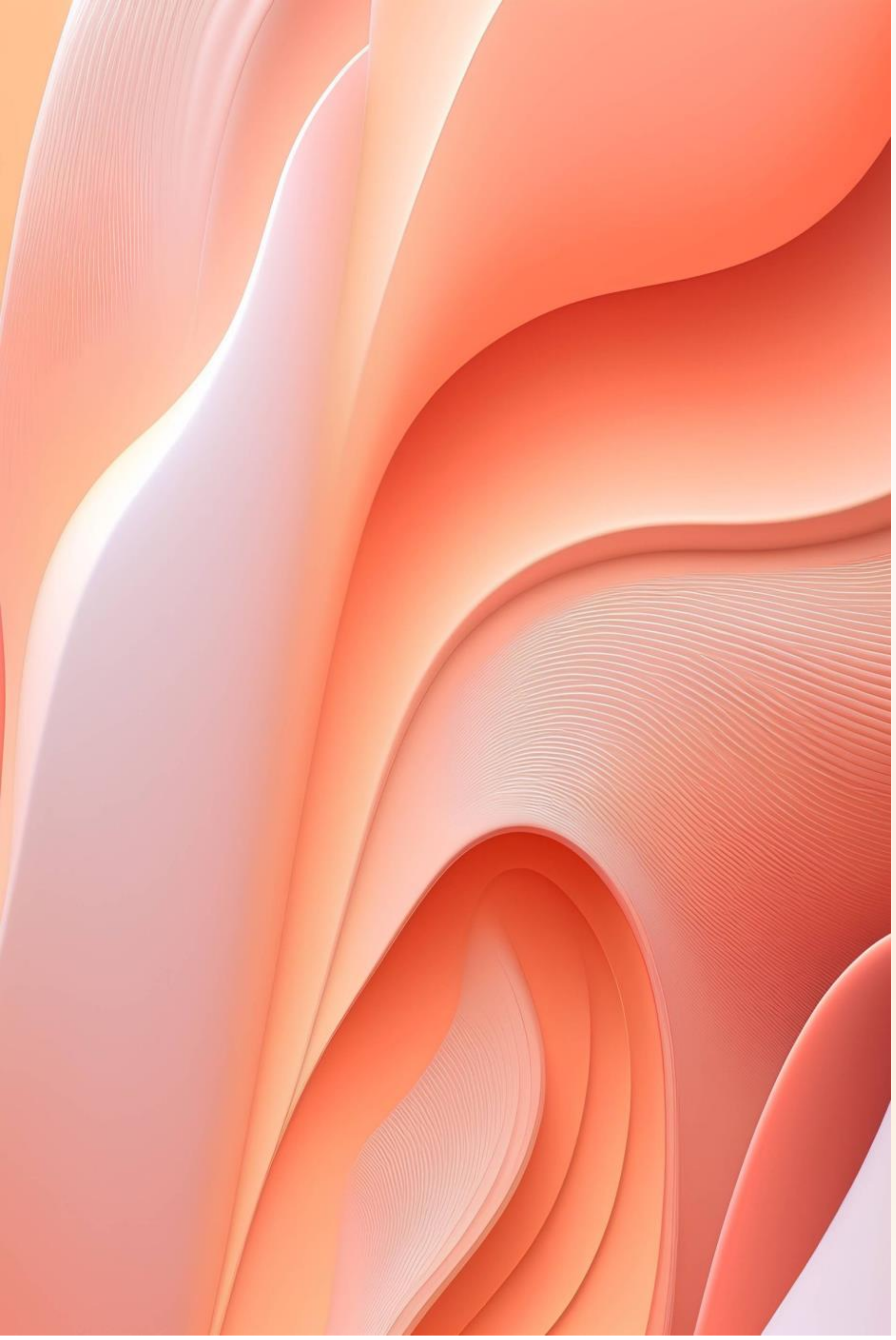
Fix Invalid Values

Fixed outliers using box plot for:

- *annual_income*
- *credit_accounts*







Summary-Data Cleaning

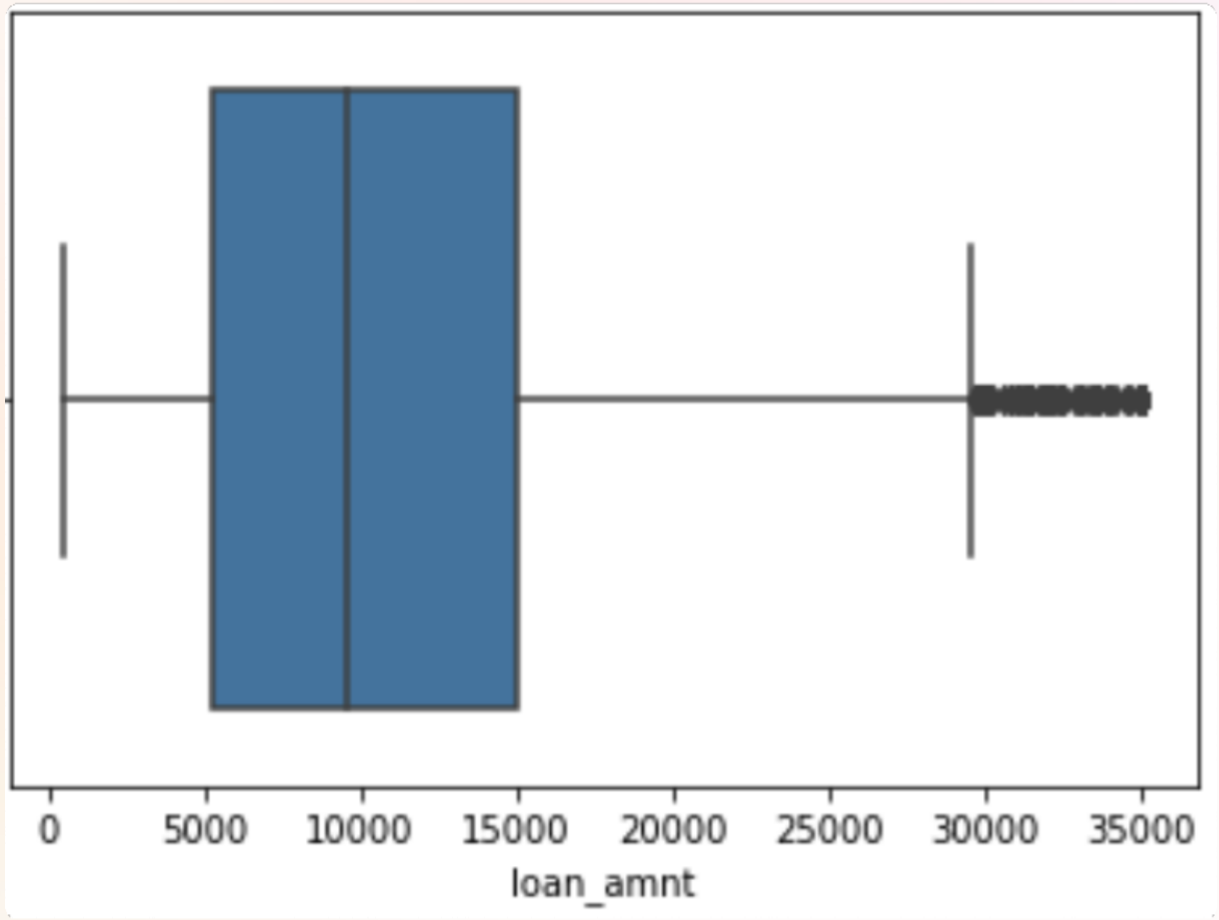
Following is the summary after performing data cleaning :

1. 54 Columns with All values as null are dropped
2. 4 Columns with 25% values as null are dropped
3. 9 Columns having same value for all records are dropped
4. 16 Irrelevant columns based on observation of rows are dropped
5. Renamed few columns for better understanding.
6. NAN and Null values are replaced by appropriate value.
7. Dropping rows where loan status is current
8. Removed outliers

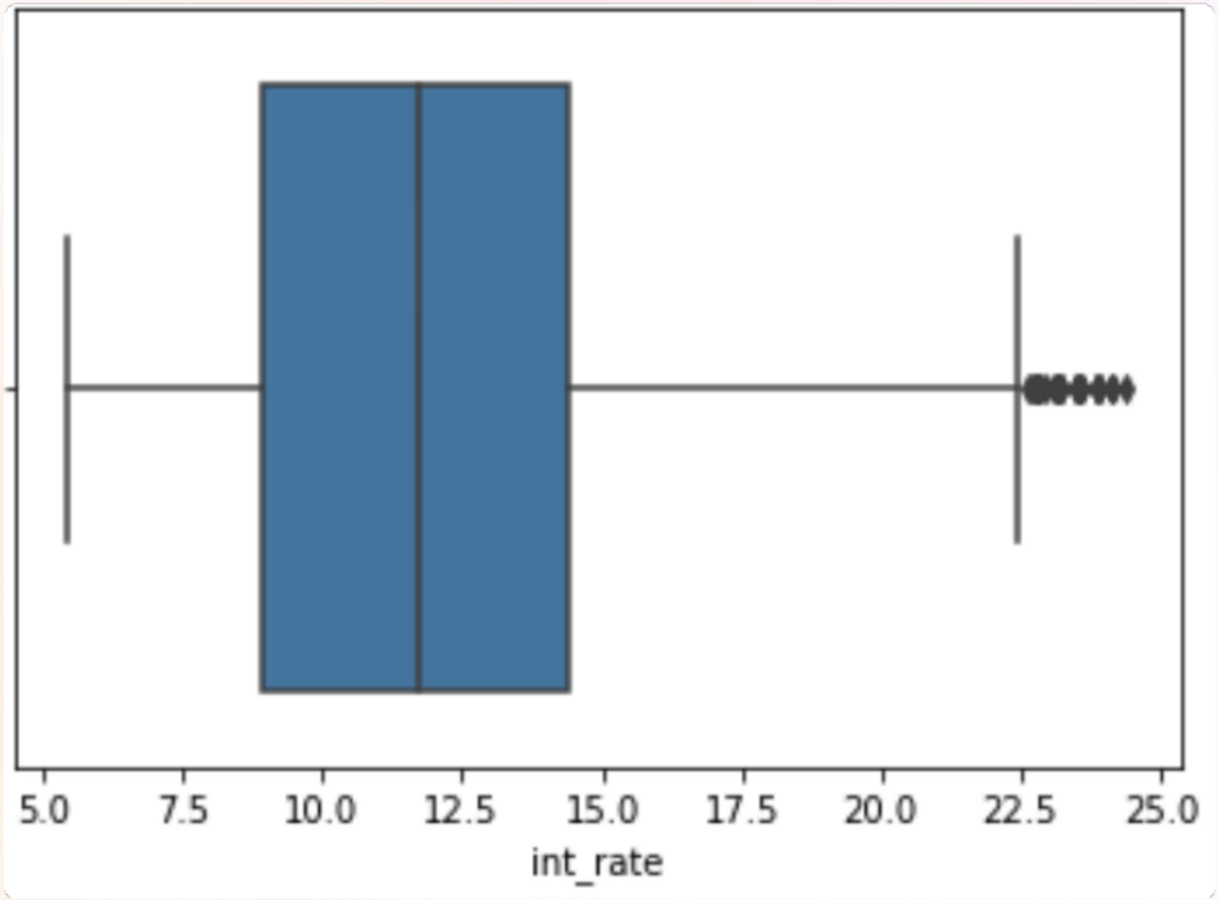
Univariate Analysis

Univariate analysis is the examination of a single variable to understand its characteristics and distribution within a dataset.

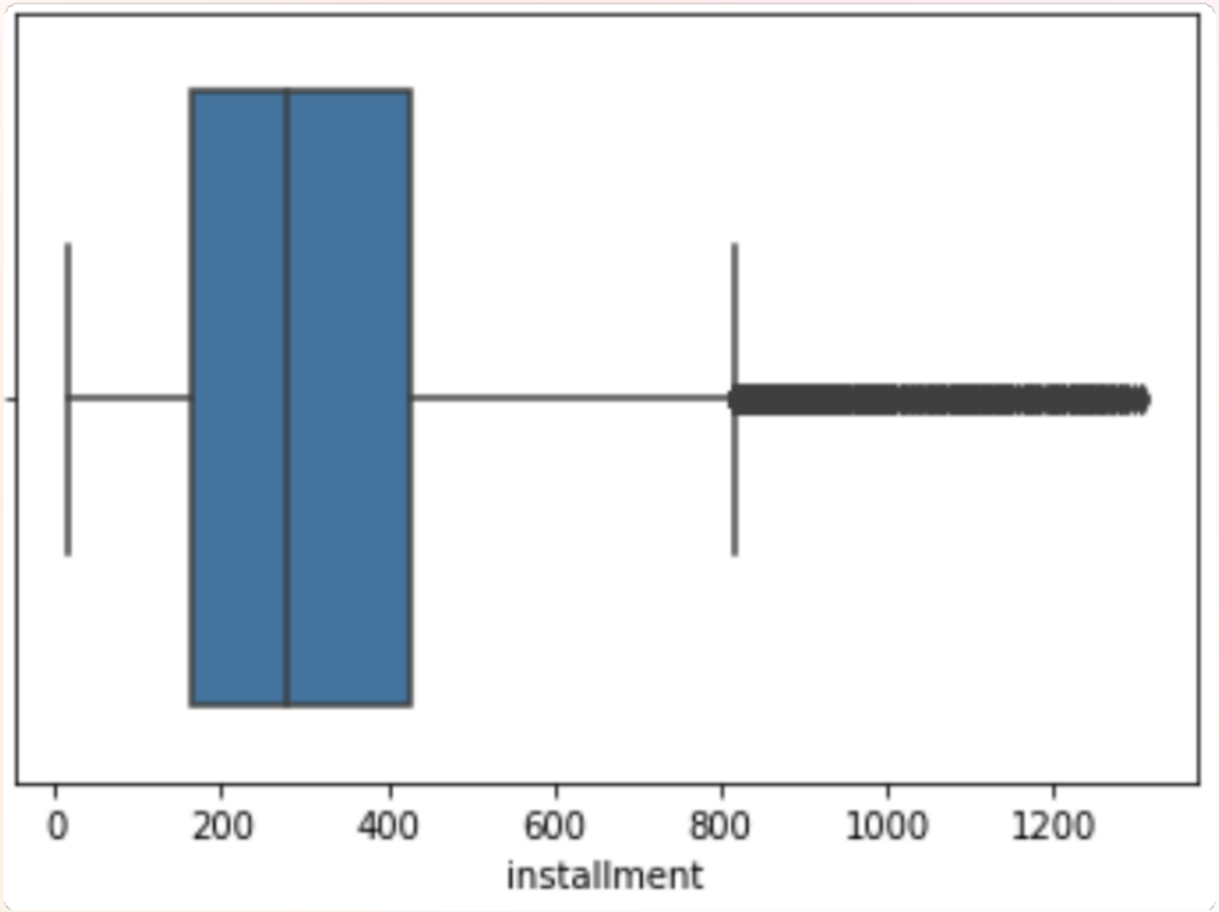
1) In the **loan amount** box plot, it is evident that the majority of loans are within the $5,000 to $15,000 range, and the median loan amount is $10,000. This indicates a prevailing preference among applicants for loans falling within this range



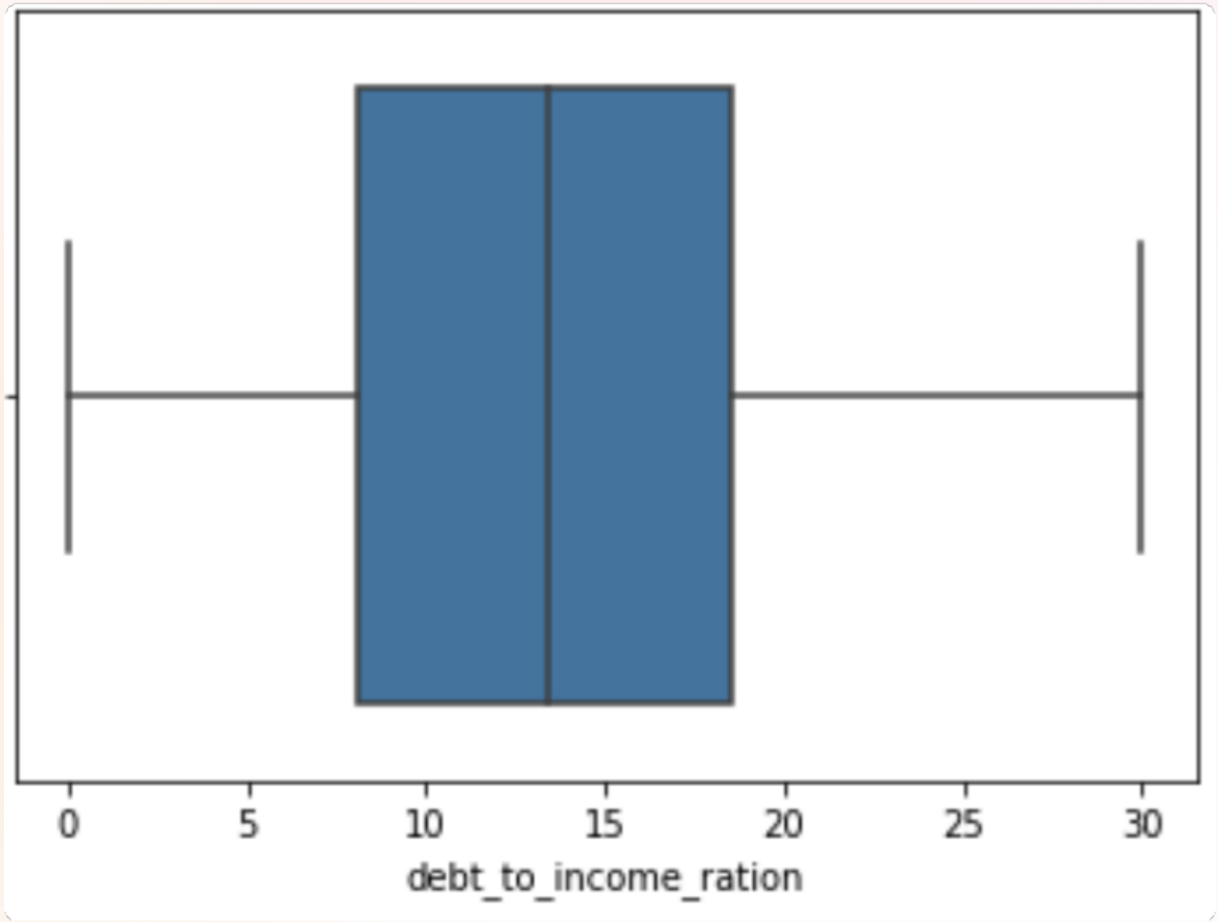
2) In the **int_rate box plot**, the central tendency of **interest rates** is evident, with a median rate of 15.0%. The majority of loans exhibit rates between **10.0% and 12.5%**, while some outliers extend to a maximum of 22.5%, reflecting variations in borrower risk profiles.

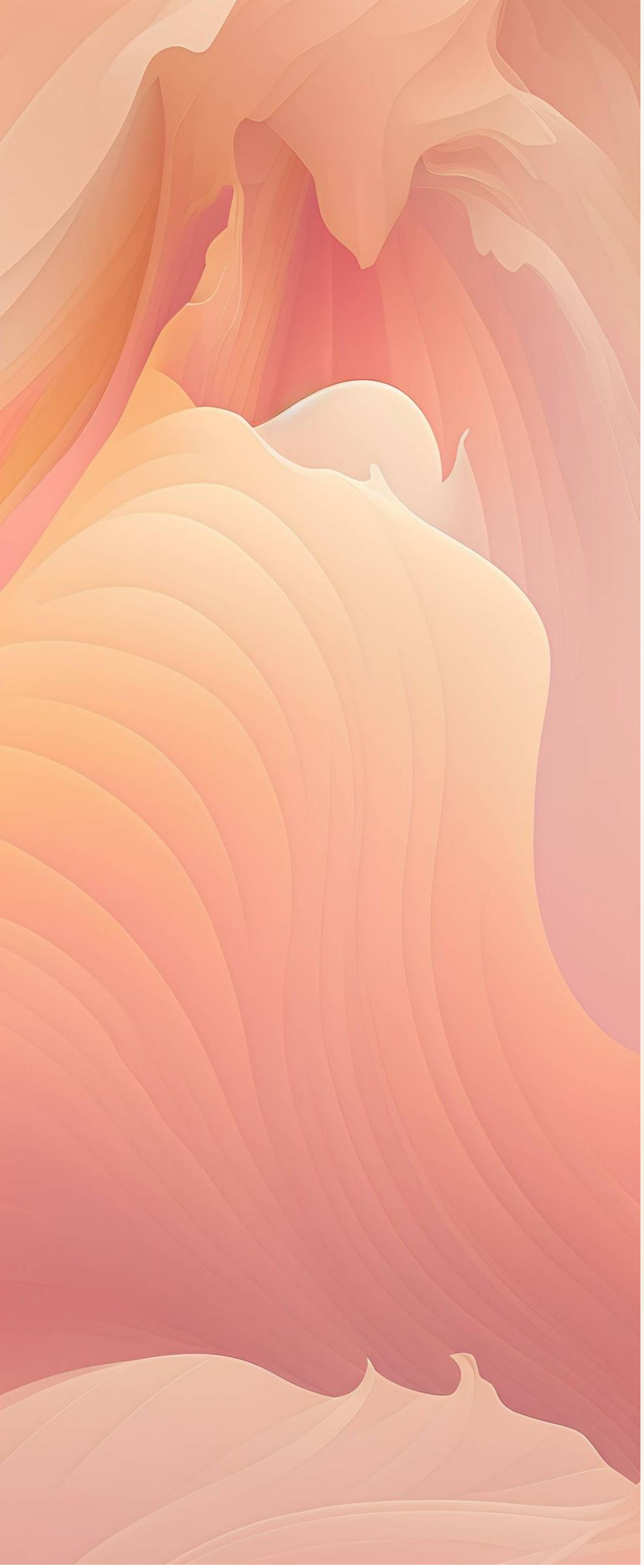


3) In the **installments box plot**, we observe that the majority of borrowers have monthly installments falling within the range of 200 to 400 months (between the 25th and 75th percentiles). However, it's noteworthy that there are instances where monthly installments exceed 800. This distribution suggests that most borrowers opt for manageable monthly payments, but a segment of borrowers may have chosen longer-term loans with higher installments, possibly to accommodate larger loan amounts or specific financial preferences.

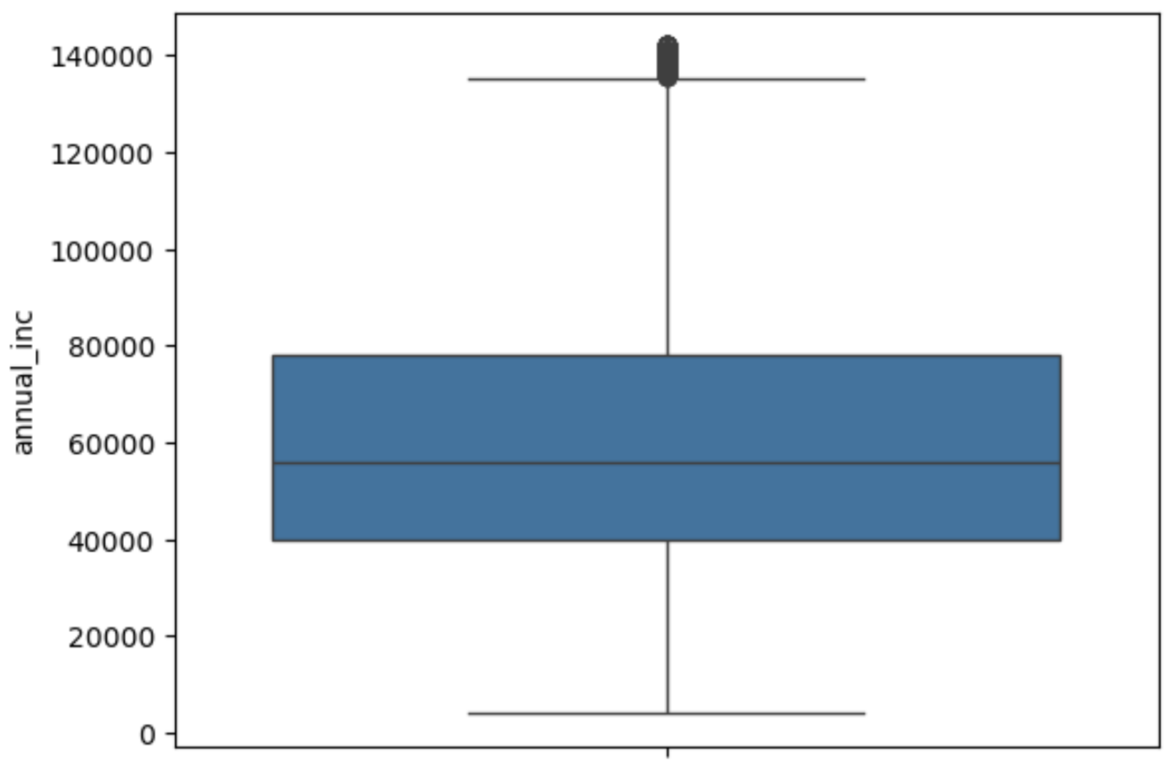


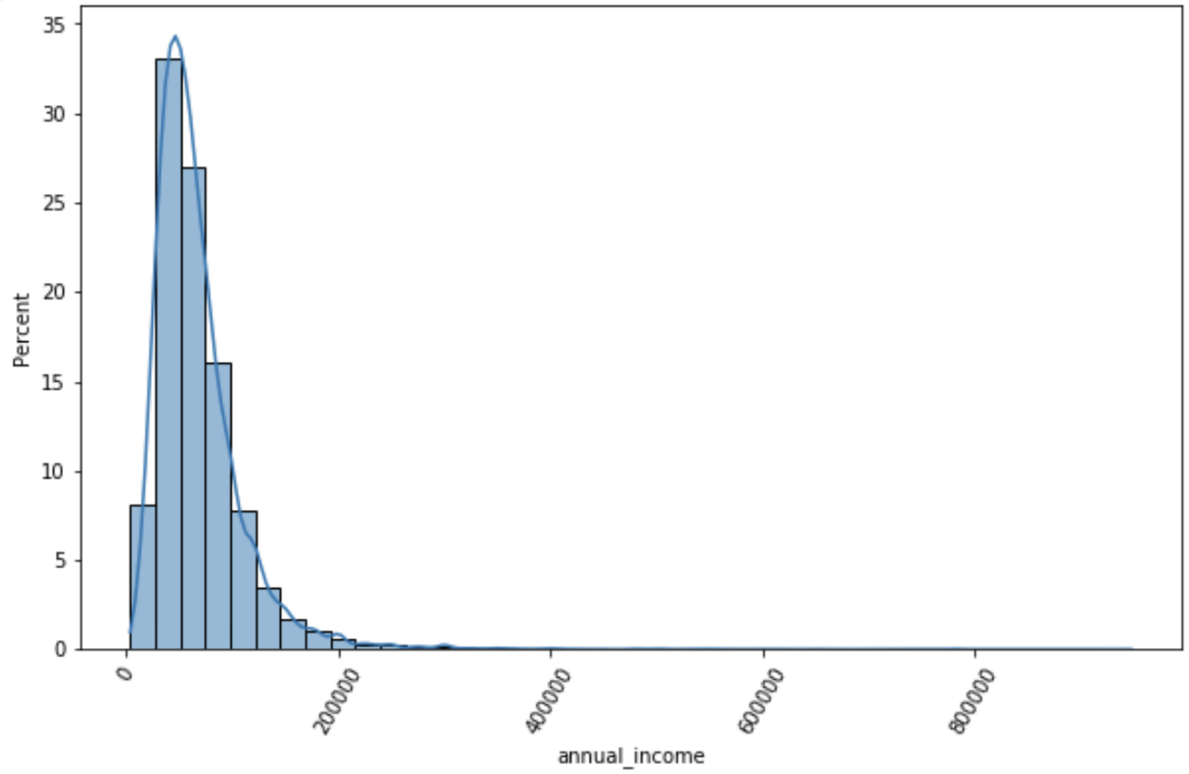
4) In the **debt-to-income** ratio **box plot**, the data shows a relatively tight distribution with minimal outliers. This indicates that the majority of applicants have debt-to-income ratios within a healthy range of 8 to 18 (between the 25th and 75th percentiles). This is a positive sign for lending, as it suggests that most borrowers maintain a sustainable balance between debt and income, reducing the risk of defaults. Lending institutions can focus on approving loans within this range to promote responsible borrowing behavior.

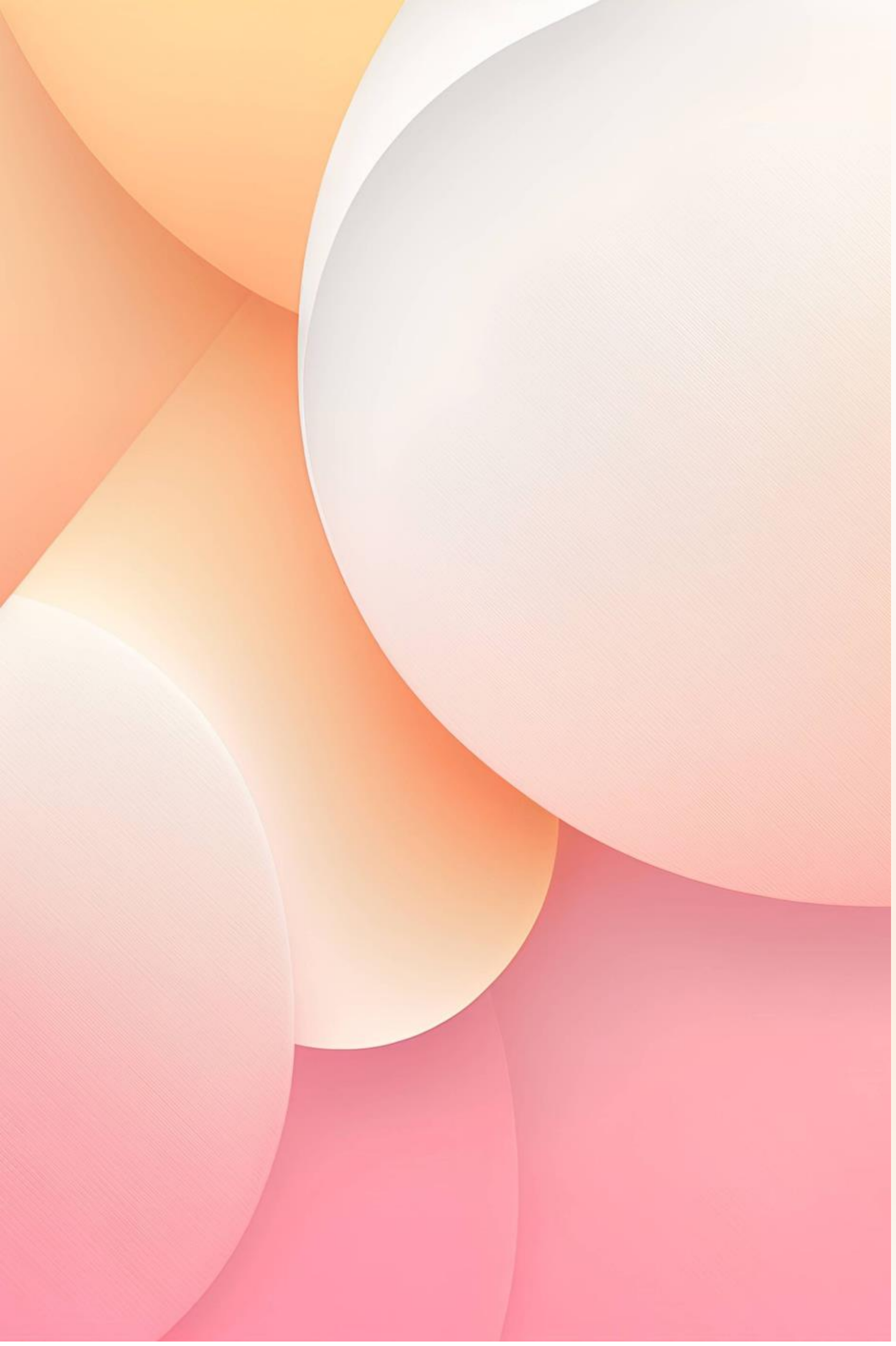




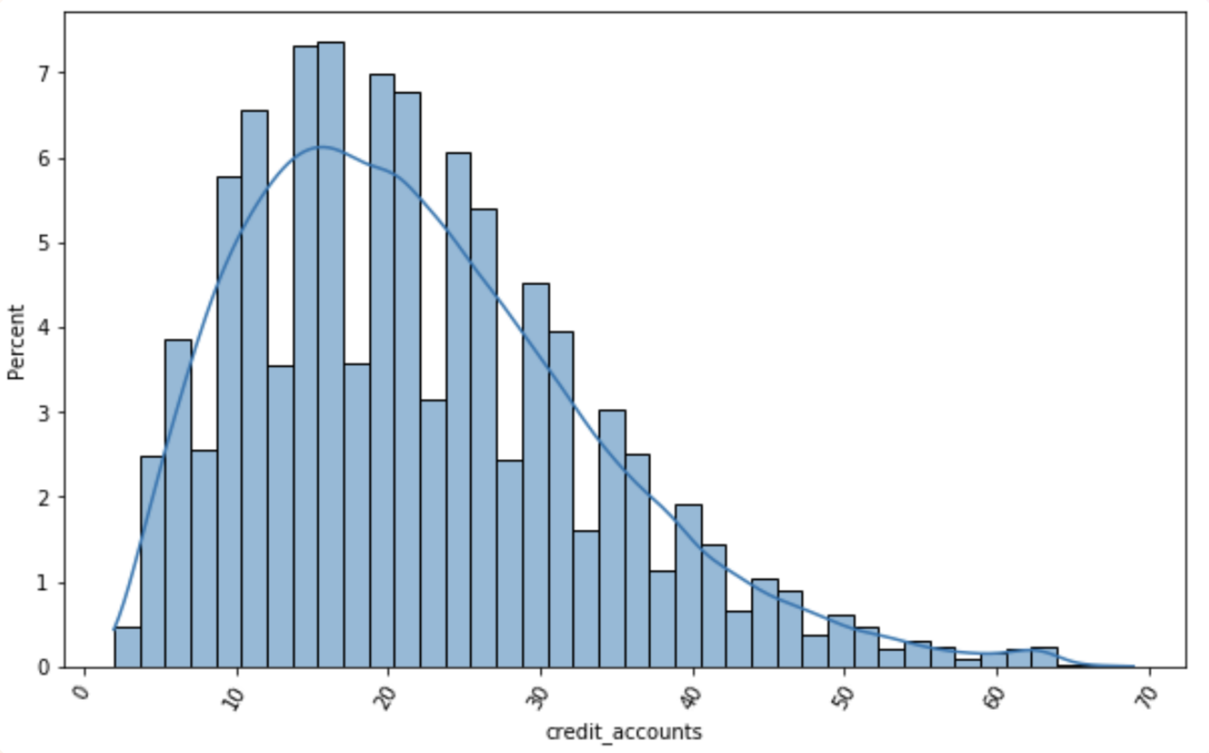
5) Most individuals indicate **annual incomes** in the range of 40K - 80K ₹. Yet, a substantial segment reports incomes exceeding 1.5L ₹. This diversity in income levels presents lending institutions with the potential to serve a broad spectrum of applicants, each with distinct financial profiles and requirements, thereby enhancing the diversity of their loan portfolio.

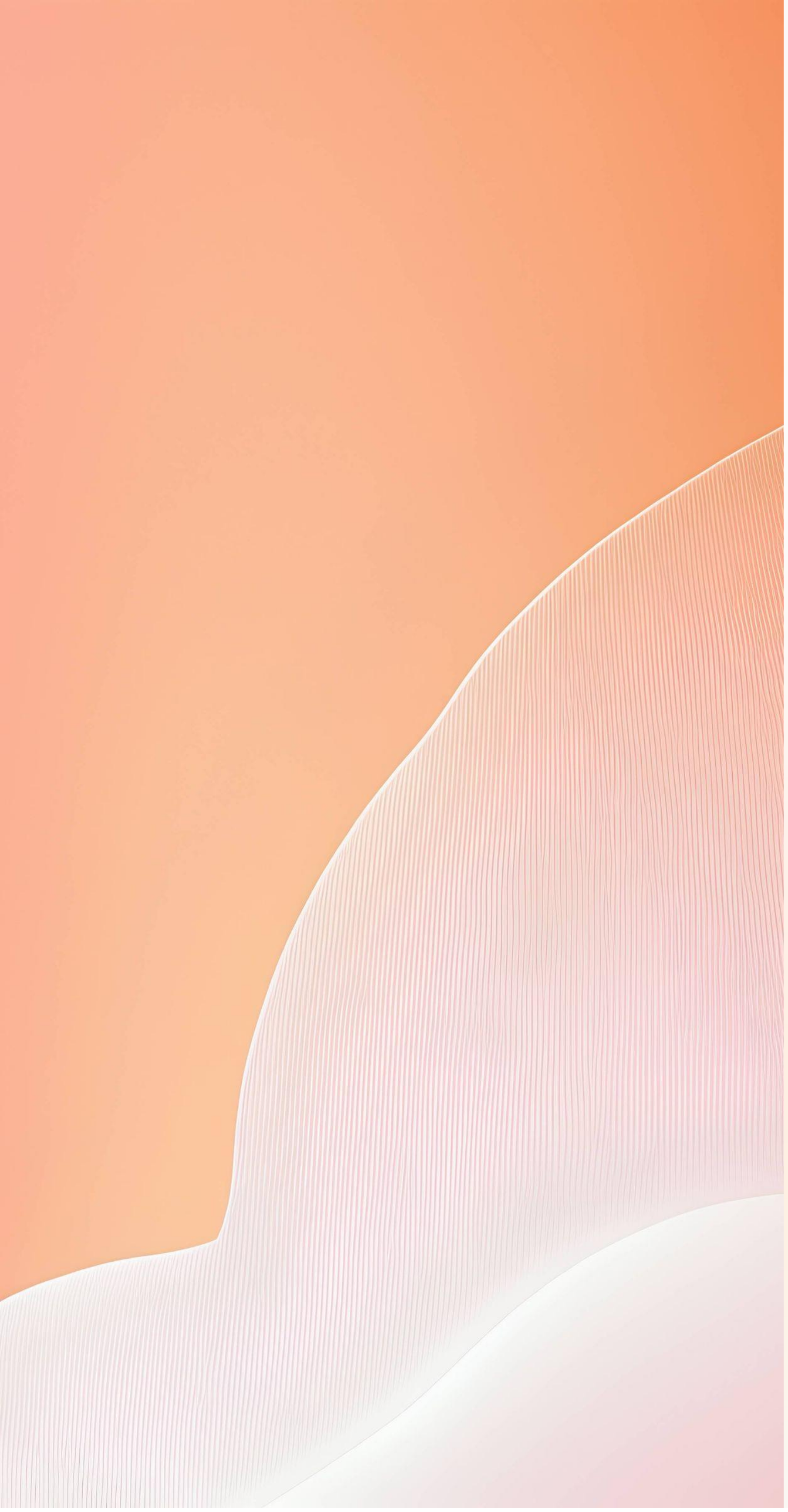






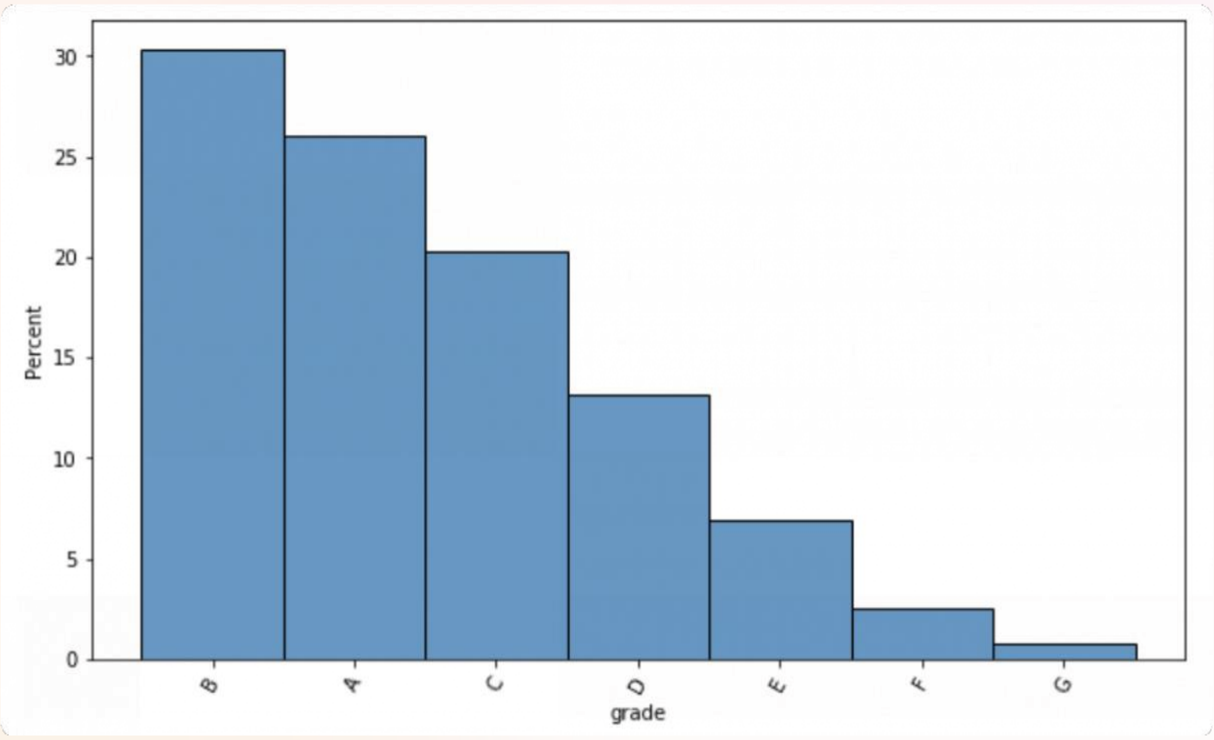
6) Higher **credit account** percentages in the **20-24 age group** indicate greater **credit adoption** among **young adults**, Conversely, **declining percentages** with age suggest **older adults** tend to close accounts upon **debt repayment and retirement**. The rising trend in credit account usage over time may stem from factors like convenience, card availability, and favorable interest rates.





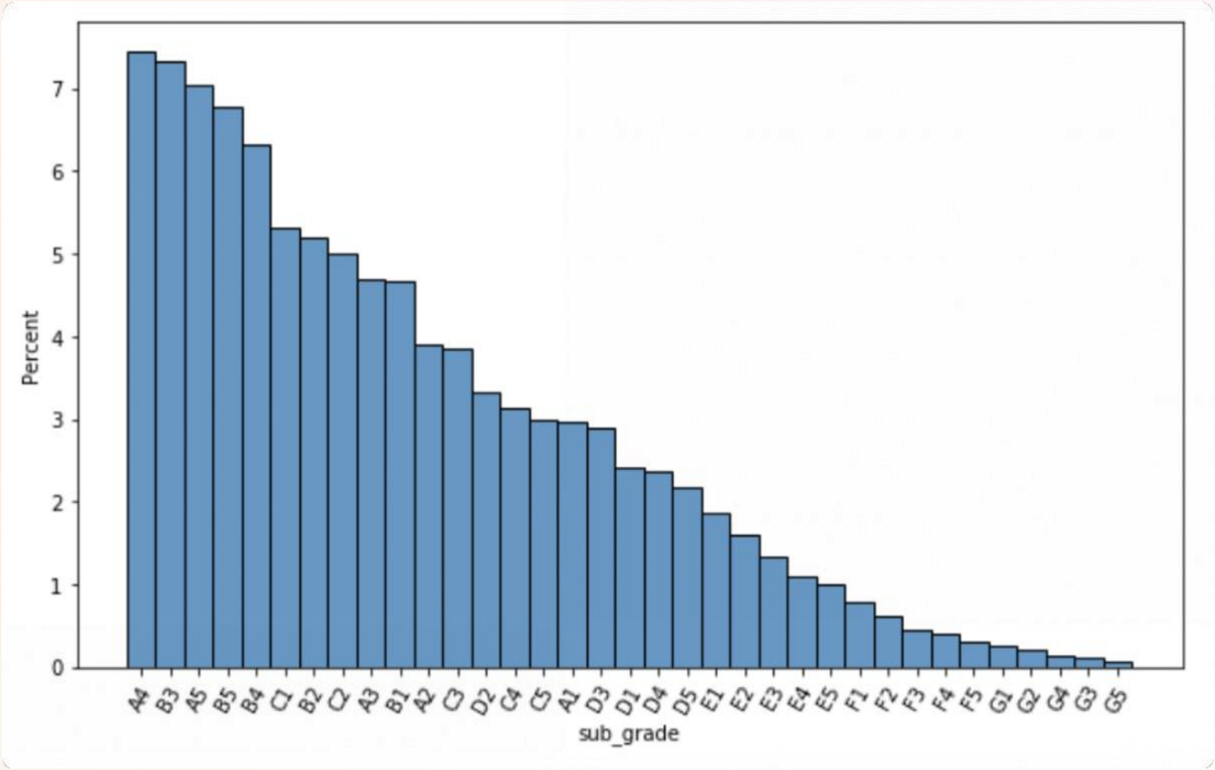
7) **Grades**

- **Insight 1:** The majority of loans are in the B and A grade categories. This suggests that borrowers with these credit grades are the most likely to be approved for loans.
- **Insight 2:** The number of loans in the C and D grade categories is significantly lower than the number of loans in the B and A grade categories. This suggests that borrowers with these credit grades are less likely to be approved for loans, or they may be approved for loans with less favorable terms.



8) **Sub-grades**

- The concentration of loans in the lower B, upper C, and upper D sub-grades suggests that there may be a need for lenders to develop products and services that are specifically tailored to the needs of borrowers in these sub-grades.
- Lenders can also use this information to identify borrowers who may be at risk of defaulting on their loans. This information can be used to develop early intervention strategies to help these borrowers avoid default.



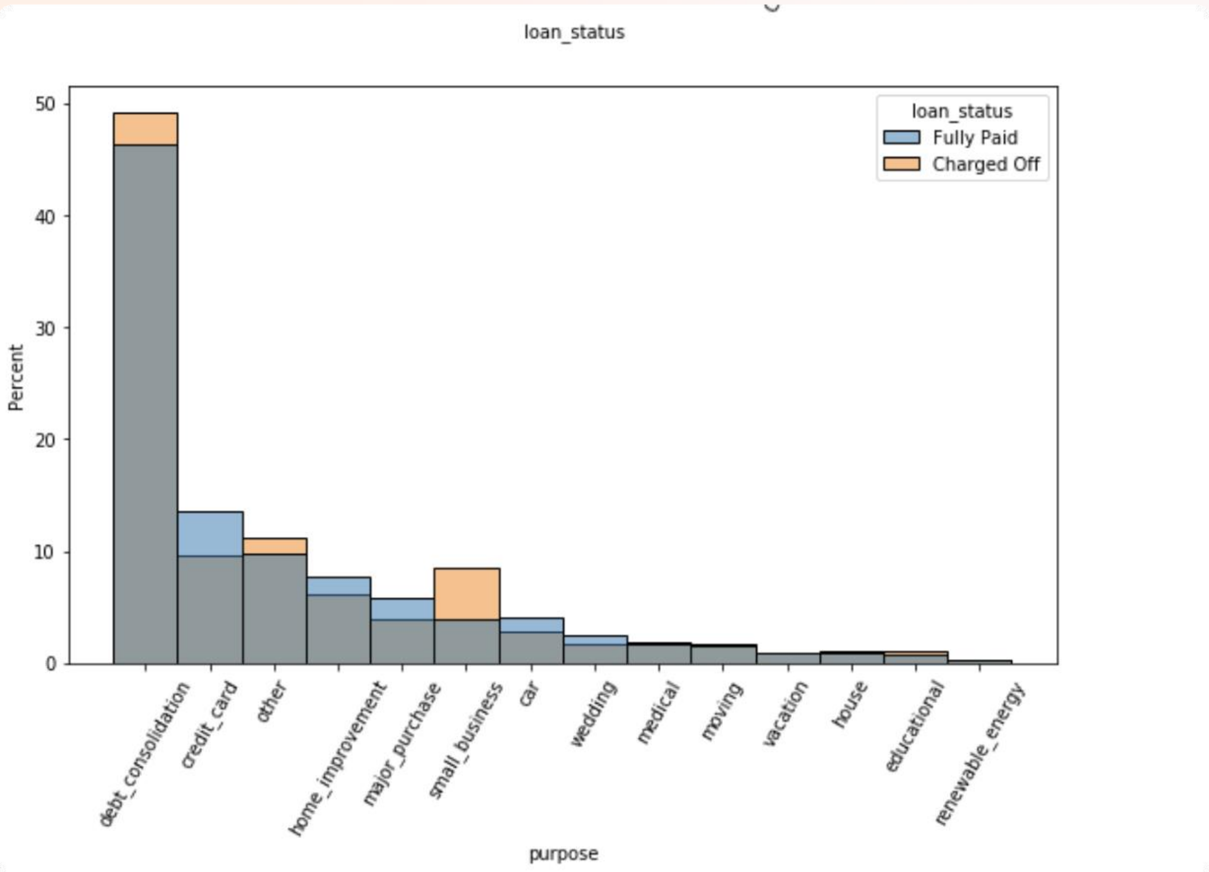
Segmented Univariate Analysis

is the study of a single variable within distinct subgroups or segments of a dataset to identify patterns or differences specific to each segment, helping to uncover hidden insights. It allows for a more detailed examination of individual groups within the data.

9) **Loan Status** The percentage of charged off loans is highest for personal loans and credit cards. This suggests that borrowers are more likely to default on these types of loans, compared to other types of loans.

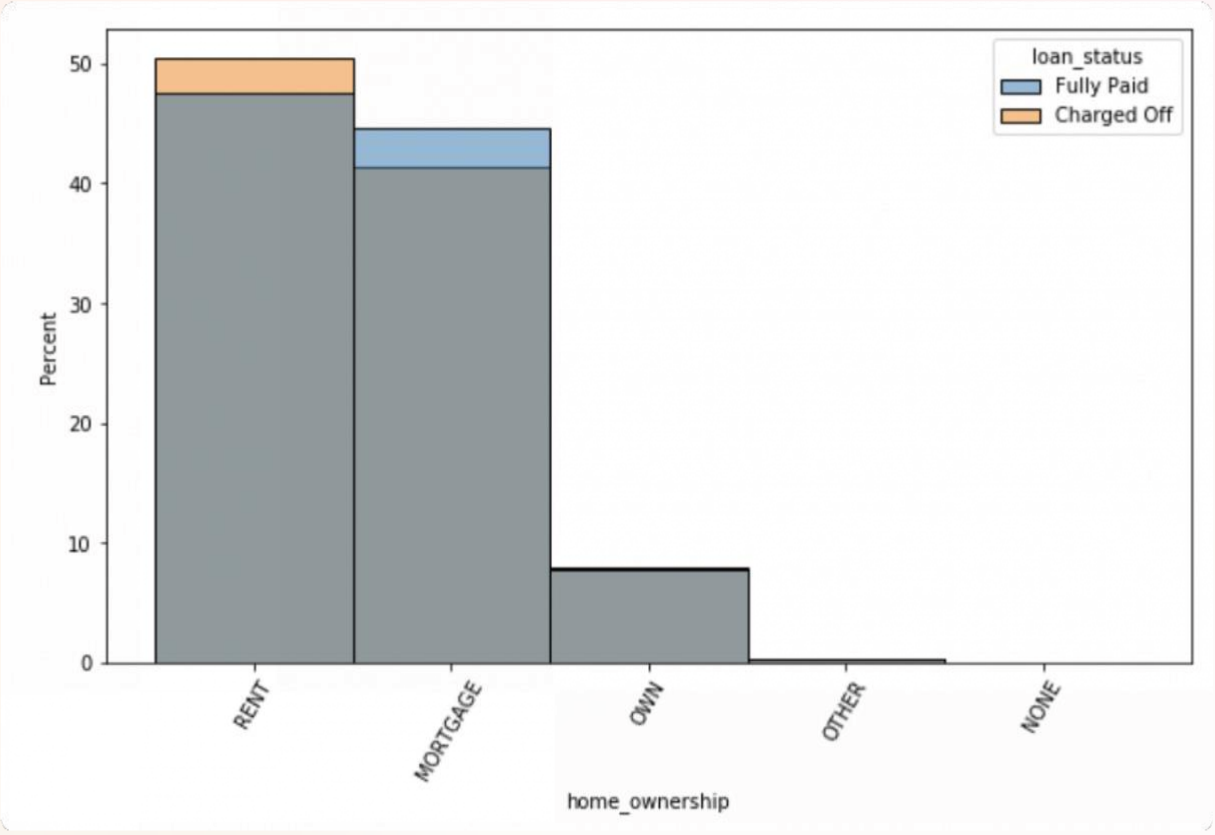
The percentage of charged off loans is also relatively high for business loans and education loans. This suggests that businesses and students are also at risk of defaulting on their loans.

The percentage of charged off loans is lowest for home loans and auto loans. This suggests that borrowers are less likely to default on these types of loans, compared to other types of loans.



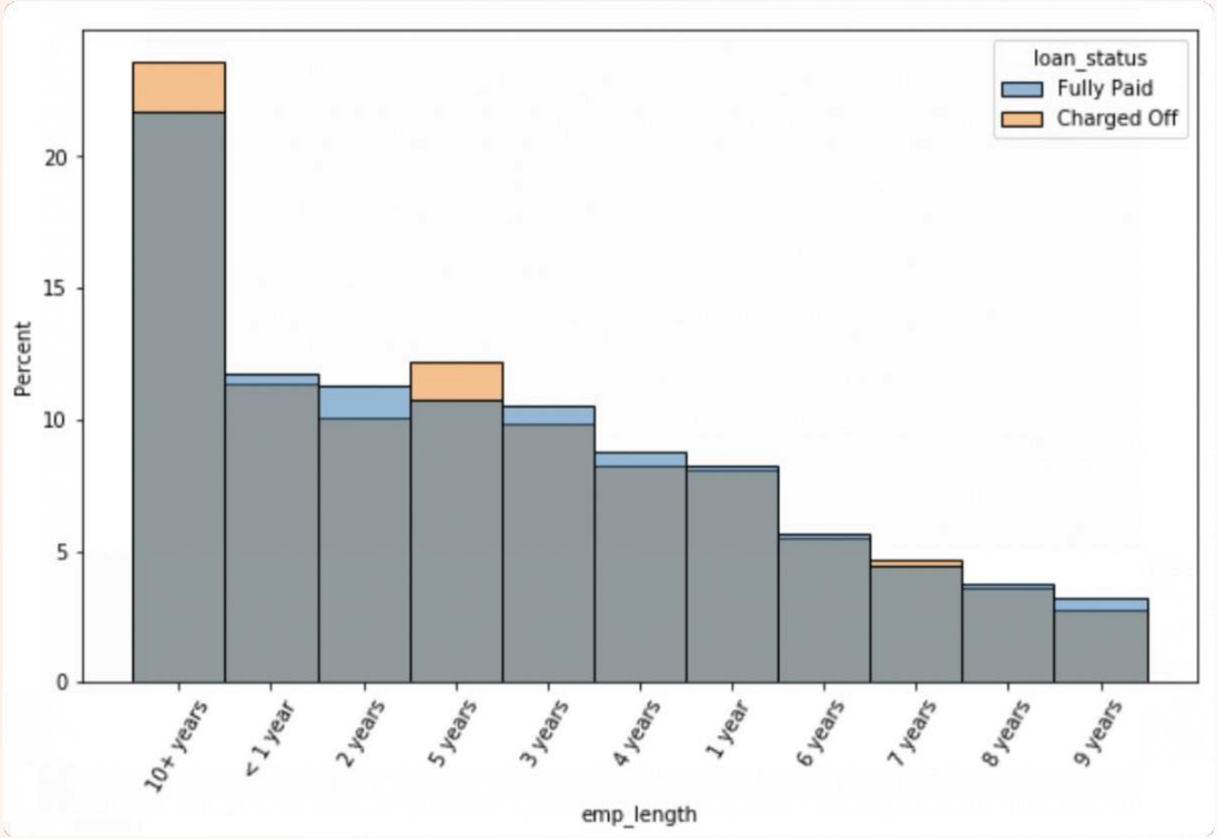
10) **Home_Ownership** Overall, the graph provides valuable insights into the housing status of borrowers in India. Lenders can use this information to develop more targeted marketing and underwriting strategies. Lenders can also use this information to set pricing for loans in a way that reflects the risk of default for borrowers with different housing statuses.

- **Offer affordable housing loans to borrowers with lower incomes.**
- **Work with borrowers who are renting to develop a plan for homeownership.**
- **Provide financial counseling and other support services to borrowers who are struggling to afford their housing costs.**



11) Emp_length Here are some additional insights that can be drawn from the graph

- The percentage of loans that are fully paid decreases as the number of years since origination increases. This is likely due to the fact that borrowers with longer-term loans are more likely to default on their loans.
- The percentage of loans that are charged off increases as the number of years since origination increases. This is likely due to the fact that lenders are more likely to charge off loans that have been delinquent for a longer period of time.
- The percentage of loans that are in other categories (such as delinquent or in forbearance) remains relatively constant over time. This suggests that there is a steady stream of loans that are in various stages of distress.

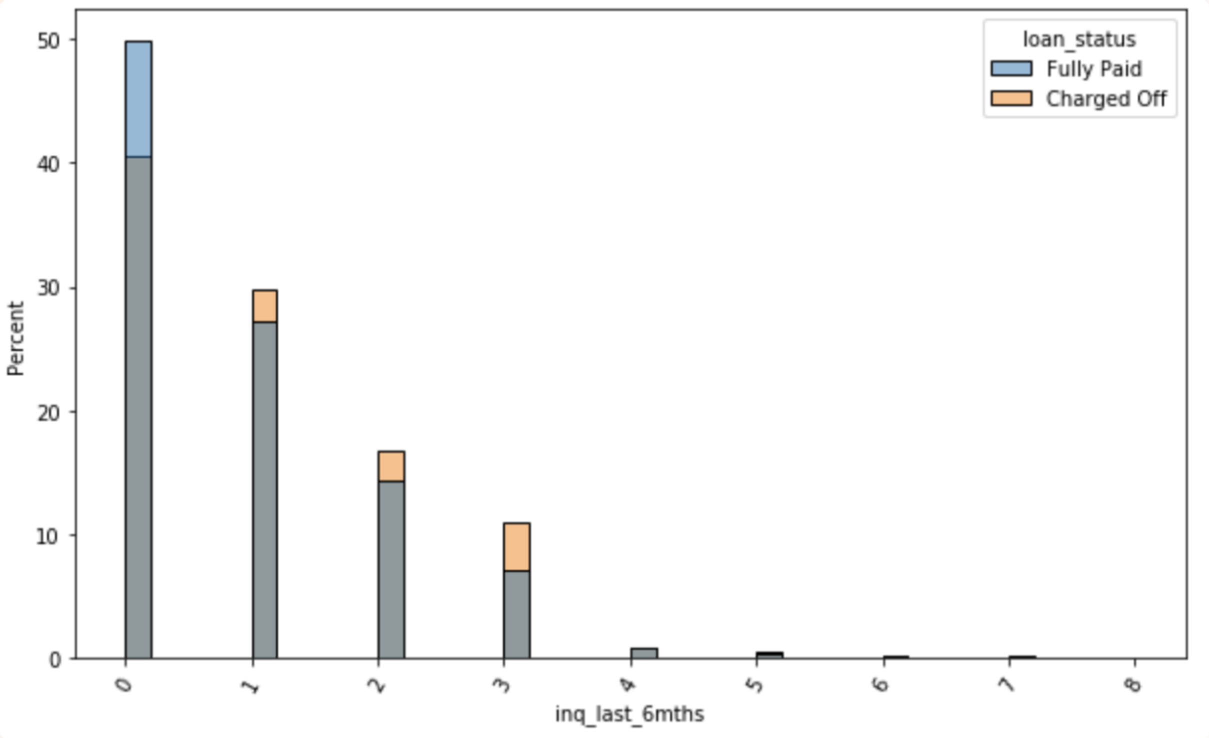


12) Here are some observations for the below **inq_last_6 months** graph for the lending company:

- The majority of loans (50%) have 0-2 inquiries in the last 6 months.
- There is a significant decrease in the percentage of loans as the number of inquiries increases.
- The percentage of loans with 7+ inquiries is relatively small, but it is important to note that these borrowers may be at a higher risk of default.

Here are some **additional insights** that can be drawn from the graph:

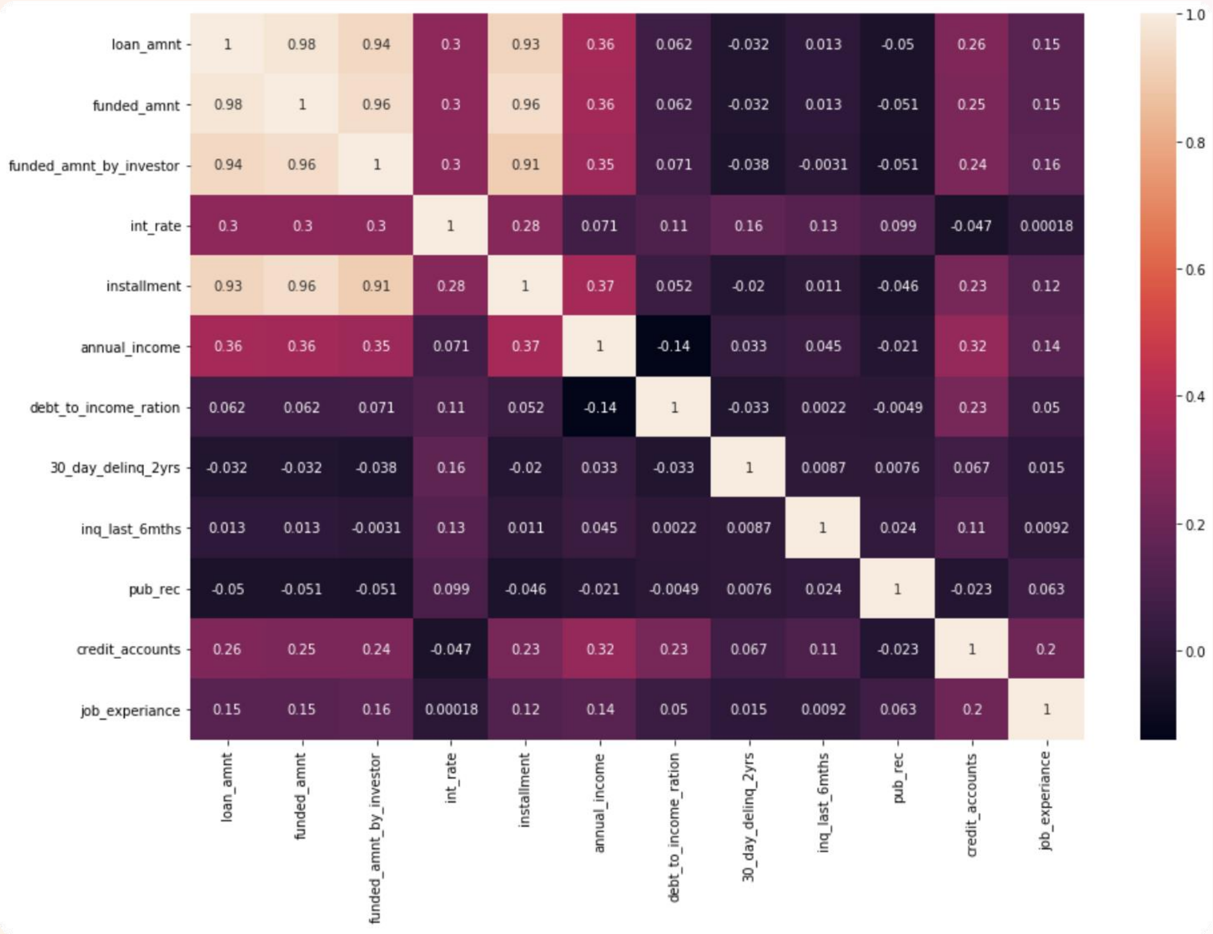
- The graph suggests that borrowers with fewer inquiries are more likely to be approved for loans.
- Borrowers with more inquiries may be seen as a higher risk by lenders, as they may be struggling to manage their debt.



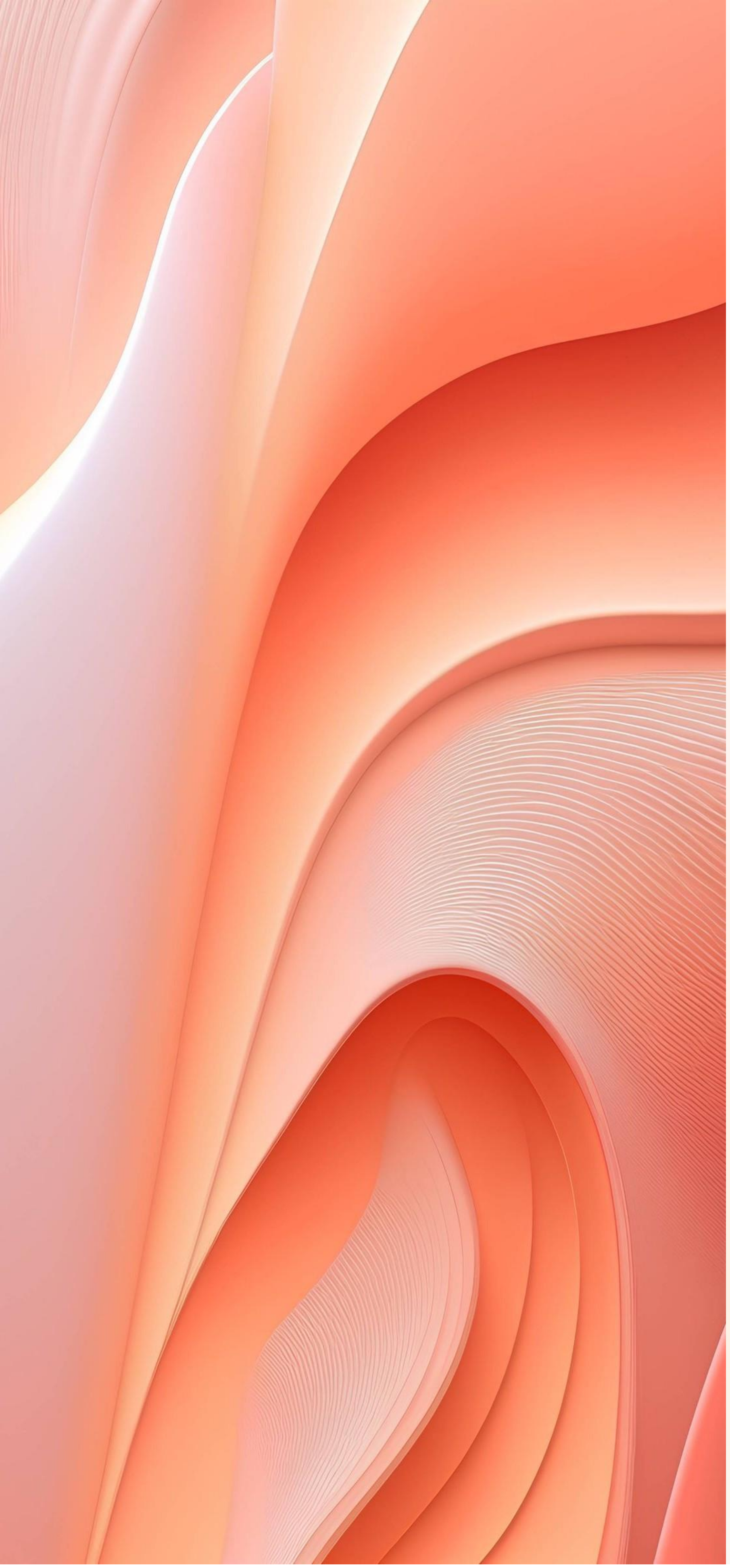
Bivariate Analysis

13) It is the examination of two variables in a dataset to understand how they are related or interact with each other, helping to uncover connections and associations between the two variables.

- **The majority of loans are given to borrowers with a debt-to-income ratio of less than 50%.** This suggests that lenders are more likely to give loans to borrowers who have a lower debt burden.
- **There is a significant decrease in the percentage of loans given to borrowers as the debt-to-income ratio increases.** This suggests that lenders are less likely to give loans to borrowers with a higher debt burden.
- **The percentage of loans given to borrowers with a debt-to-income ratio of 75% or more is relatively small, but it is important to note that these borrowers may be at a higher risk of default.**

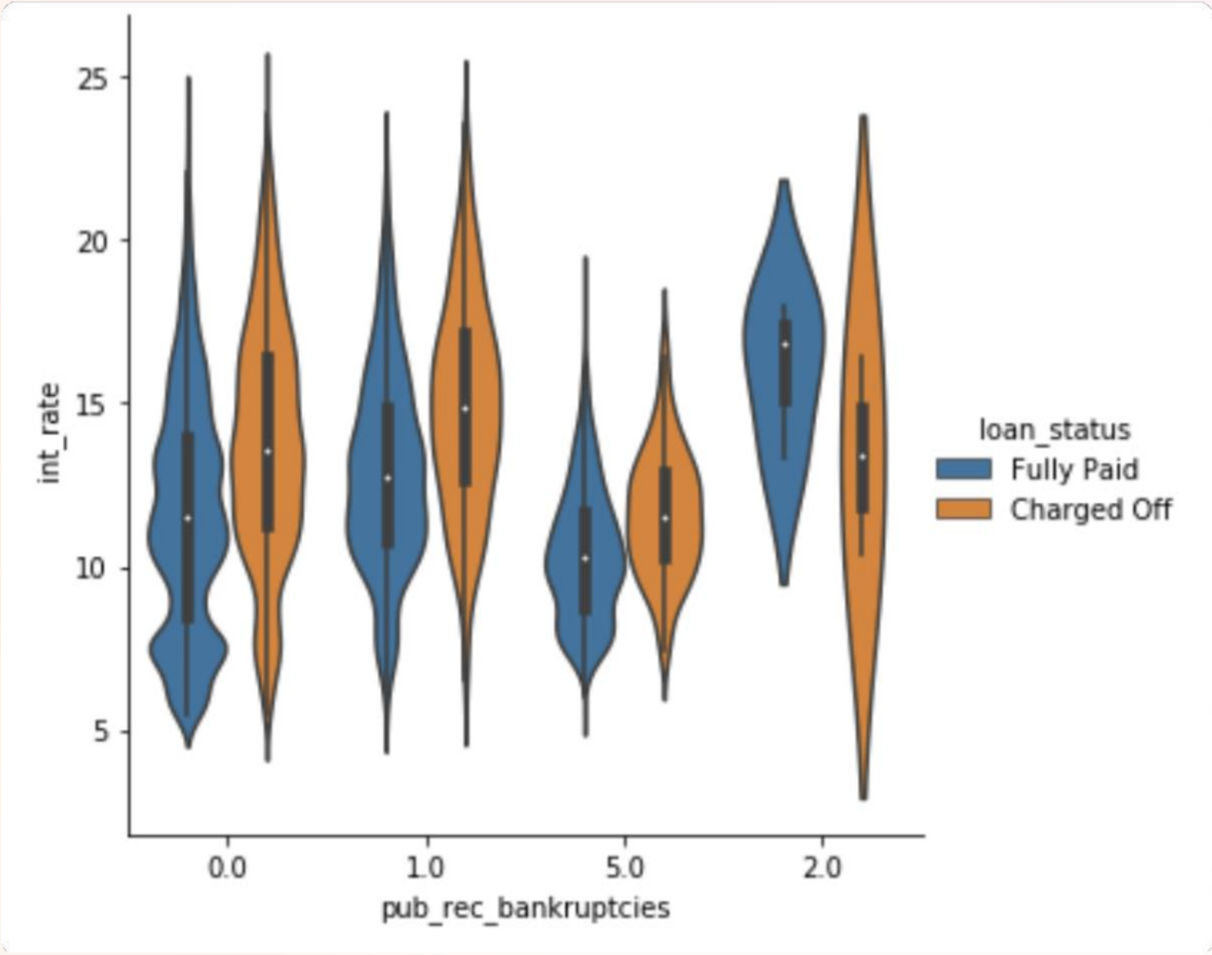


	loan_amnt	funded_amnt	funded_amnt_by_investor	int_rate	installment	annual_income	debt_to_income_ration	30_day_delinq_2yrs	inq_last_6mths	pub_rec	credit_accounts	job_experiance
loan_amnt	1	0.98	0.94	0.3	0.93	0.36	0.062	-0.032	0.013	-0.05	0.26	0.15
funded_amnt	0.98	1	0.96	0.3	0.96	0.36	0.062	-0.032	0.013	-0.051	0.25	0.15
funded_amnt_by_investor	0.94	0.96	1	0.3	0.91	0.35	0.071	-0.038	-0.0031	-0.051	0.24	0.16
int_rate	0.3	0.3	0.3	1	0.28	0.071	0.11	0.16	0.13	0.099	-0.047	0.00018
installment	0.93	0.96	0.91	0.28	1	0.37	0.052	-0.02	0.011	-0.046	0.23	0.12
annual_income	0.36	0.36	0.35	0.071	0.37	1	-0.14	0.033	0.045	-0.021	0.32	0.14
debt_to_income_ration	0.062	0.062	0.071	0.11	0.052	-0.14	1	-0.033	0.0022	-0.0049	0.23	0.05
30_day_delinq_2yrs	-0.032	-0.032	-0.038	0.16	-0.02	0.033	-0.033	1	0.0087	0.0076	0.067	0.015
inq_last_6mths	0.013	0.013	-0.0031	0.13	0.011	0.045	0.0022	0.0087	1	0.024	0.11	0.0092
pub_rec	-0.05	-0.051	-0.051	0.099	-0.046	-0.021	-0.0049	0.0076	0.024	1	-0.023	0.063
credit_accounts	0.26	0.25	0.24	-0.047	0.23	0.32	0.23	0.067	0.11	-0.023	1	0.2
job_experiance	0.15	0.15	0.16	0.00018	0.12	0.14	0.05	0.015	0.0092	0.063	0.2	1



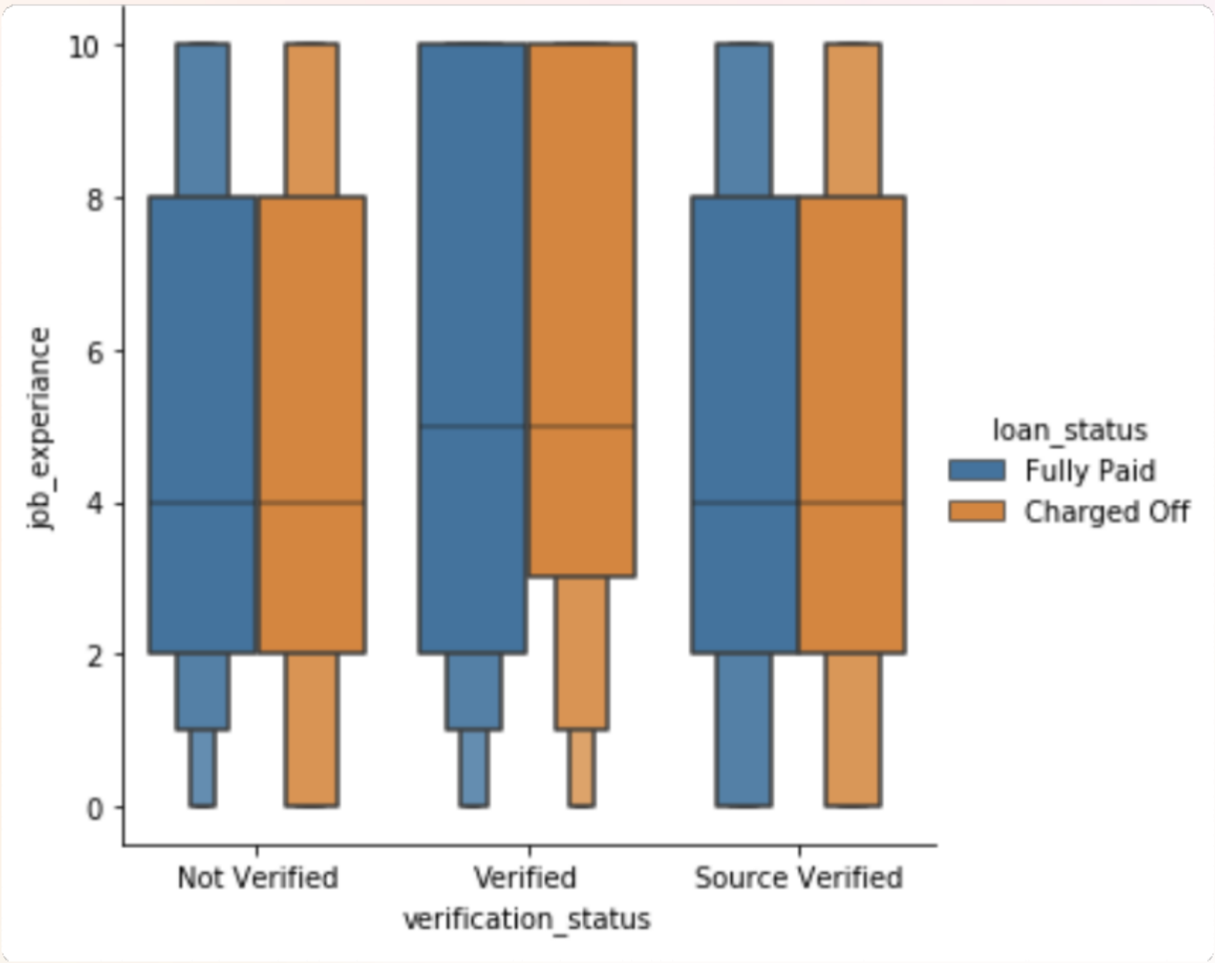
14) **pub_rec_bankruptcies**

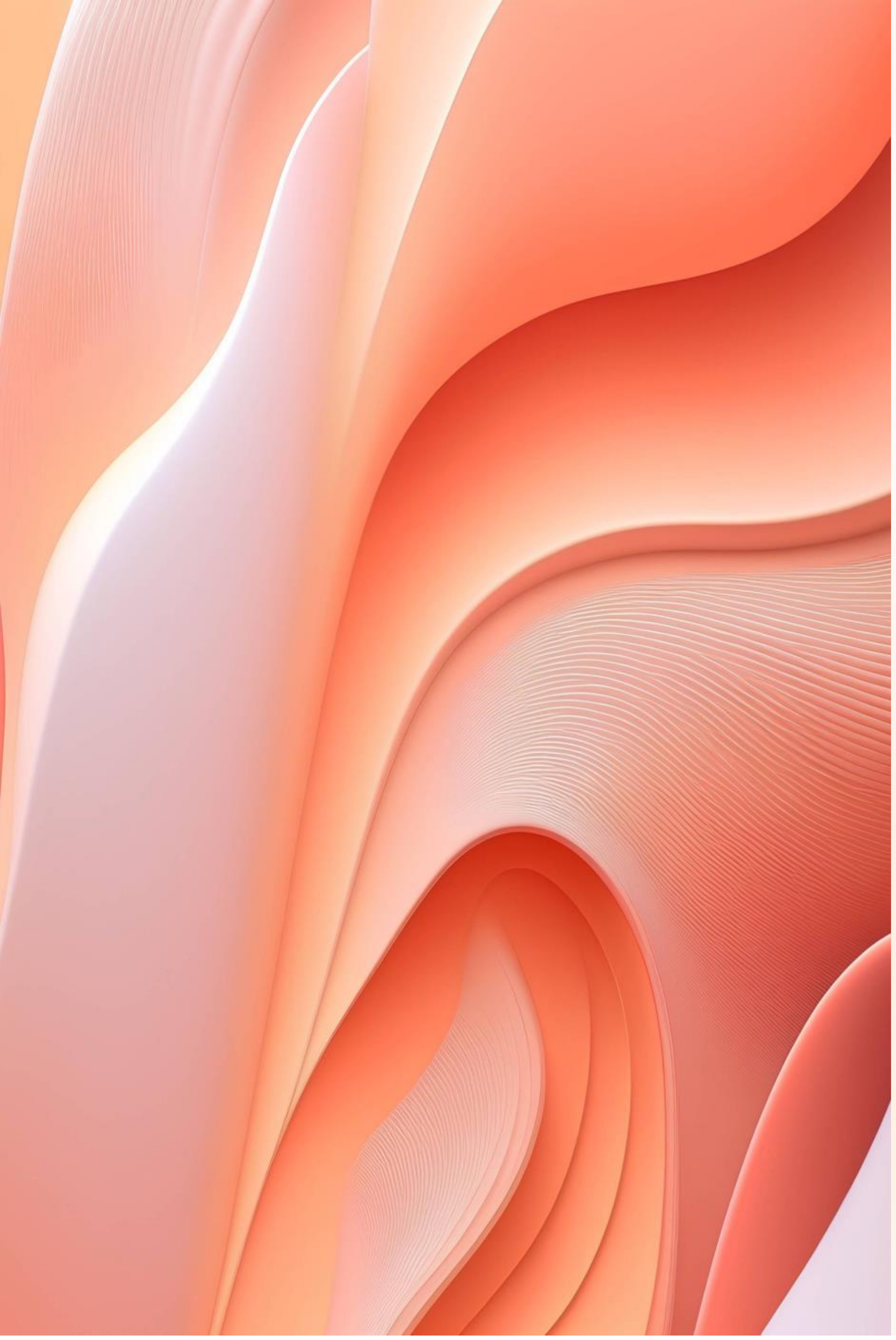
- The majority of loans are given to borrowers with no public record of bankruptcy.
- There is a significant decrease in the percentage of loans given to borrowers as the number of public record of bankruptcies increases.
- The percentage of loans given to borrowers with 2+ public record of bankruptcies is relatively small, but it is important to note that these borrowers may be at a higher risk of default.
- Lenders may charge higher interest rates to borrowers with a higher number of public record of bankruptcies.



15) **Verification_status**

- This is evidenced by the fact that the percentage of loans given to people with verified incomes increases as job experience increases.
- People with verified incomes are more likely to get loans than people with unverified incomes
- Lenders may use verification status as a factor in determining a borrower's credit score.
- Lenders may charge higher interest rates to borrowers with unverified incomes.

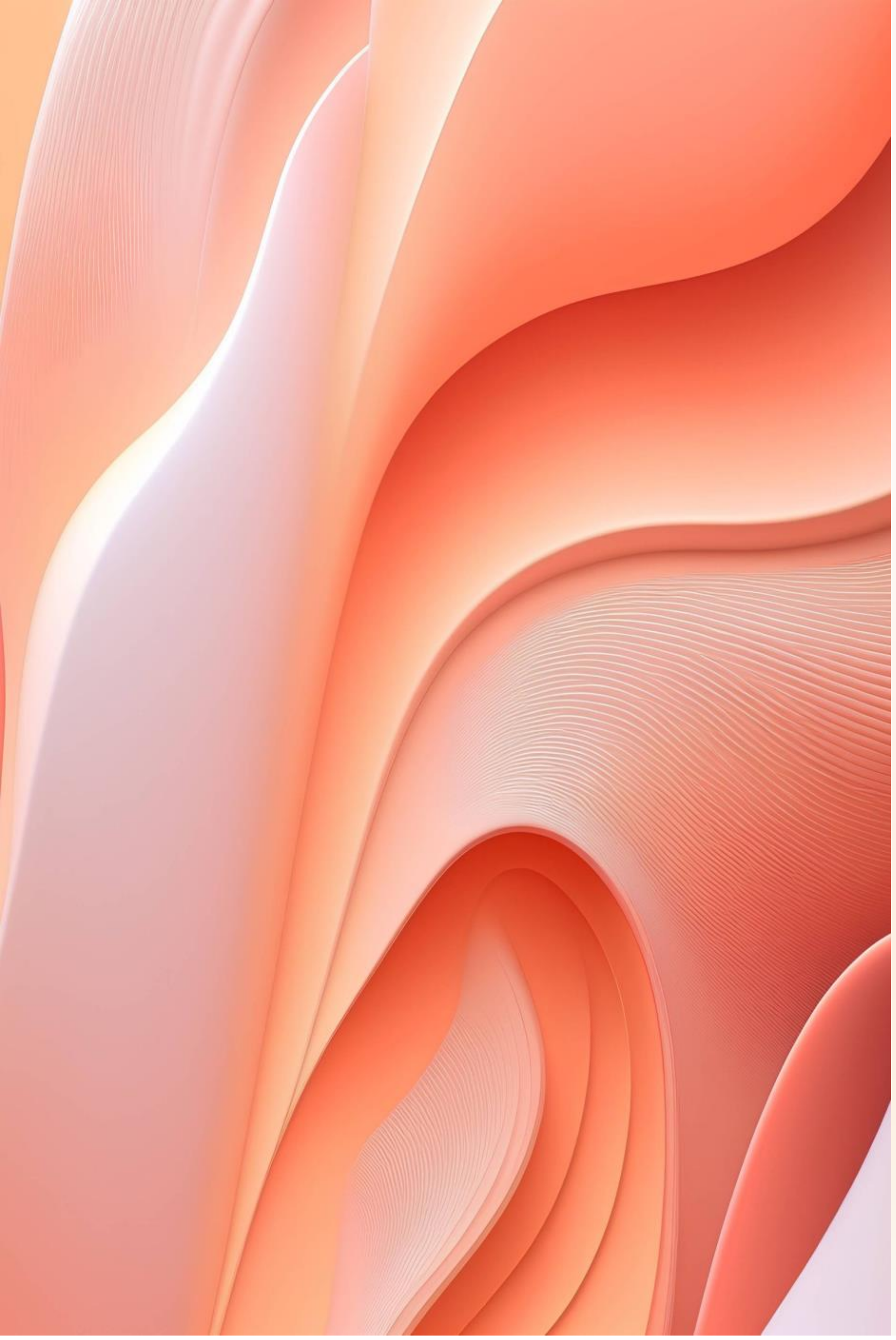




Recommendations

- Identified 14 parameters which needs to be considered while giving the loan.
- If more than 5 parameters matches in applicant’s profile then we should avoid giving loan for such applicants
- Following are the parameters :

1. **grade** - Grades given by LC are C,D,E,F or G
2. **term** - Loan tenure selected is 60 months
3. **home**_ownership - Home ownership is Rent
4. **verification**_status - Verification status is Verified
5. **purpose** - Purpose of the loan is Small Business or Debt Consolidation
6. **add_state** - States are one of the following CA, FL or NV
7. **pub_rec_bankruptcies** - Any public record of bankruptcies
8. **loan_amnt** - Loan amount is 15K or more
9. **int_rate** - Interest rate is 12.5 or more
10. **annual_income** - Annual income is 10K or less than 10K
11. **debt_to_income_ration** - DTI is 12 or more
12. **inq_last_6mths** - 2 or more inquiries in last 6 months
13. **credit_accounts** - Credit accounts are 15 or less
14. **job_experiance** - Job experience is 5 or more



Thank You

Technologies used and versions

- pandas 1.3.5
- numpy 1.21.6
- matplotlib 3.1.1
- seaborn 0.12.2

Reference:

- https://stackoverflow.com/questions/51070985/find-out-the-percentage-of- missing-values-in-each-column-in-the-given-dataset
- https://www.geeksforgeeks.org/select-columns-with-specific-data-types-in- pandas-dataframe/
- https://www.geeksforgeeks.org/how-to-set-a-seaborn-chart-figure-size/
- https://stackoverflow.com/questions/63373194/how-to-plot-percentage-wit h-seaborn-distplot-histplot-displot
- https://seaborn.pydata.org/generated/seaborn.histplot.html
- https://seaborn.pydata.org/generated/seaborn.pairplot.html
- https://stackoverflow.com/questions/56942670/first-and-last-row-cut-in-hal f-of-heatmap-plot
- https://www.geeksforgeeks.org/multi-plot-grid-in-seaborn/