Lending Club Case Study

Course 1: Statistics, Module 8 (MLC 64)

Presented by:-

Akshaay Vijay HS & Vidya Siddaramappa

Context

Problem Statement
Business Understanding and Objective
Data Understanding (Data Sourced)
Data Cleaning
Data Extraction
Univariate Analysis
Bivariate Analysis
Correlations
Conclusions

Problem Statement, Business Understanding and Objective

You work for a consumer finance company specializing in providing various loans to urban customers. When the company receives a loan application, it must decide whether to approve the loan based on the applicant's profile. There are two types of risks associated with this decision:

1. If the applicant is likely to repay the loan, not approving the loan results in a loss of business for the company.

2. If the applicant is unlikely to repay the loan and is likely to default, approving the loan may lead to a financial loss for the company.

Consider a financial company which is a largest online loan marketplace, offering personal loans, business loans, and financing for medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

Like many other lending companies, lending to ‘risky’ applicants is the largest source of financial loss, known as credit loss. Credit loss occurs when the borrower refuses to pay or absconds with the owed money. In other words, borrowers who default cause the most significant losses for lenders. In this context, customers labeled as 'charged-off' are the 'defaulters'.

Identifying these risky loan applicants can reduce the number of such loans, thereby minimizing credit loss. The goal of this case study is to use Exploratory Data Analysis (EDA) to identify these risky applicants.

In other words, the company aims to understand the driving factors (or driver variables) behind loan default, i.e., the variables that strongly indicate default. The company can use this knowledge for better portfolio and risk assessment.

Data Understanding

Loan.csv file which contains ~39000 rows and ~110 columns containing the attributes from Banking sector. The attributes seen are

1. Loan attributes
2. Customer attributes
3. Banking Attributes

Data Cleaning

1. There are in total 39717 rows and 111 columns in the data set Loan.csv.
2. On detailed analysis of the data set, we found no headers, footers, summary rows, etc,. Also, no duplicate rows were found.
3. 1140 rows with **`loan_status ='current'`** were deleted as they do not participate in the analysis.
4. 55 columns with all values as null or blank were removed as they do not contribute to the analysis.
5. The **`url`** and **`member_id`** columns, being unique in nature, were deleted. The **`id`** column was retained for future analysis.
6. The **`desc`** and **`title`** columns, containing text/description values that do not participate in the analysis, were dropped.
7. Analysis is limited to the **'Group'** level, so sub-group data has been dropped.
8. 21 columns of behavioral data, which are captured post-loan approval and do not participate in the analysis, were deleted.
9. Eight columns with values consistently equal to 1, indicating uniqueness, were dropped from the analysis.
10. Two columns with more than 50% of their data as NA were removed.
11. After the data cleaning process, we are left with 38,577 rows and 20 columns.

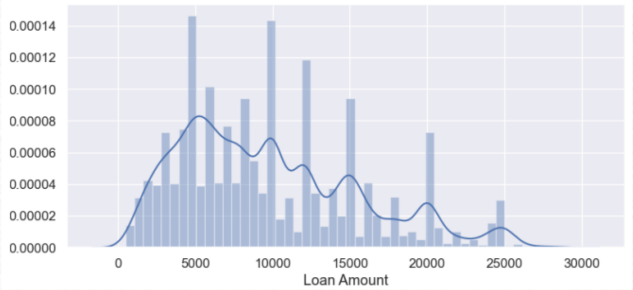
Data Extraction

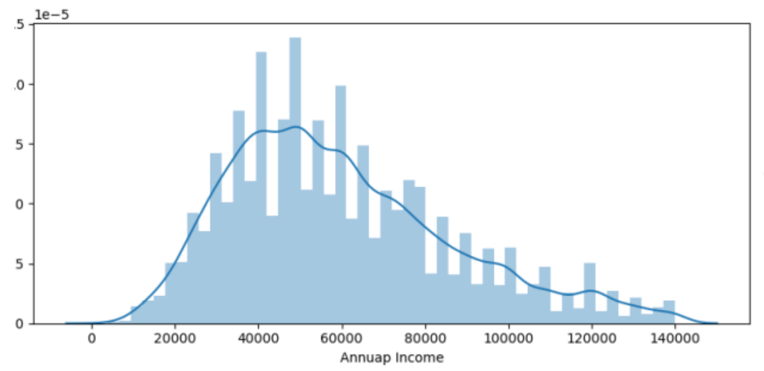
1. Truncated additional string values from the 'term' column and converted them to integer data type.
2. Converted **'int_rate'** from string to integer, removing additional '%’.
3. Converted **'loan_funded_amnt'** and **'funded_amnt'** columns to float.
4. Rounded values in **'loan_amnt'**, **'funded_amnt'**, **'funded_amnt_inv'**, **'int_rate'**, and **'dti'** columns to two decimal points.
5. Converted **'issue_d'** column to the appropriate data type.
6. Derived **'issue_year'** and **'issue_month'** columns from **'issue_d'** for further analysis purposes.
7. Created derived columns **'loan_amnt_b'**, **'annual_inc_b'**, **'int_rate_b'**, and **'dti_b'** (binned continuous data) to facilitate better analysis.

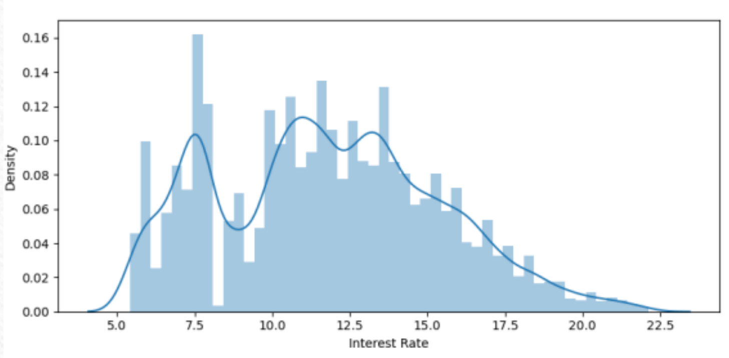
UNIVARIATE ANALYSIS

Annual Income, Loan Amount & Interest Rate

1. Many loan applications fell within the range of 5,000 to 14,000.
2. The highest loan amount applied for was approximately 27,000.
3. The annual income of most applicants ranges from 40,000 to 75,000.
4. The average annual income is 59,883.0.
5. Many applicants have an interest rate ranging between 8% and 14%.
6. The average interest rate is 11.7%.

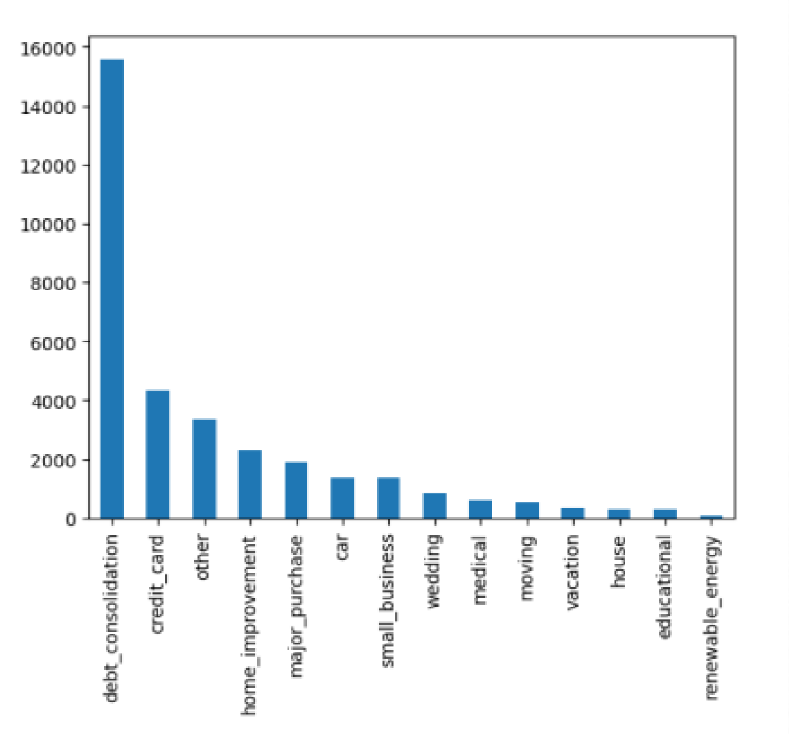


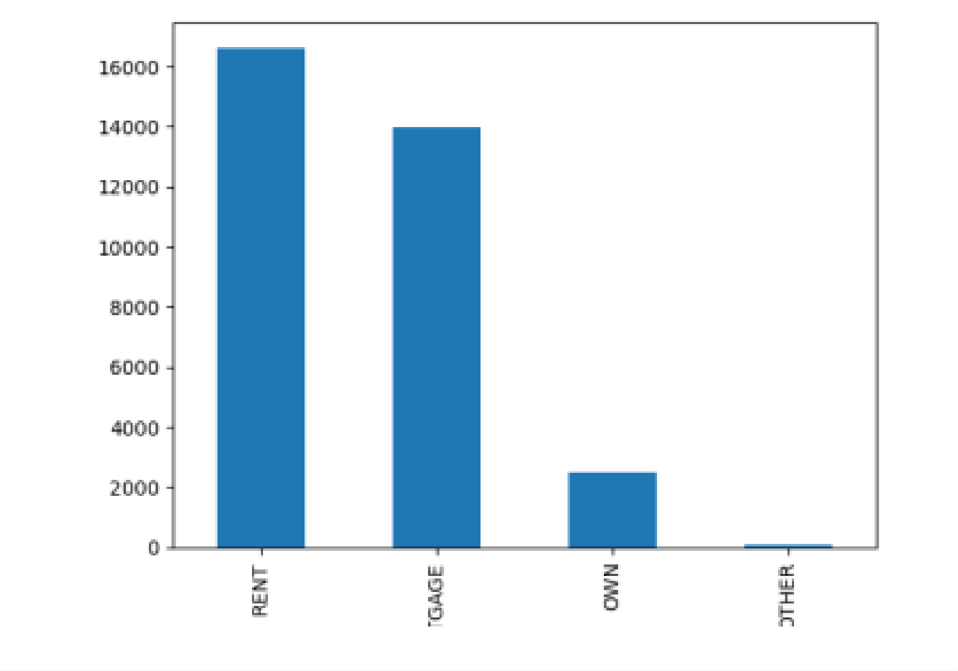




Ordered and Unordered Categorical Variables

1. Most applications indicate 10 or more years of experience.
2. Most loan applicants are seeking debt consolidations.
3. Most loan applicants either rent or have a mortgage.
4. Many loan applicants are from California (CA).

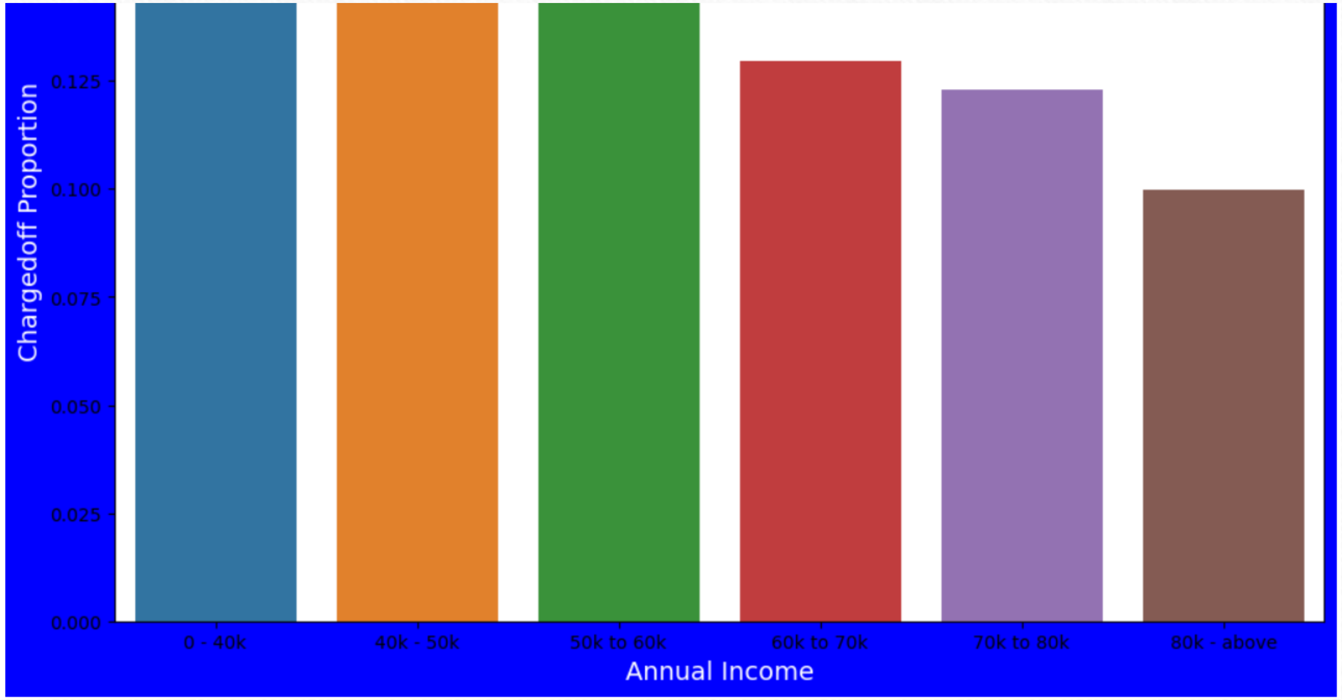




BIVARIATE ANALYSIS

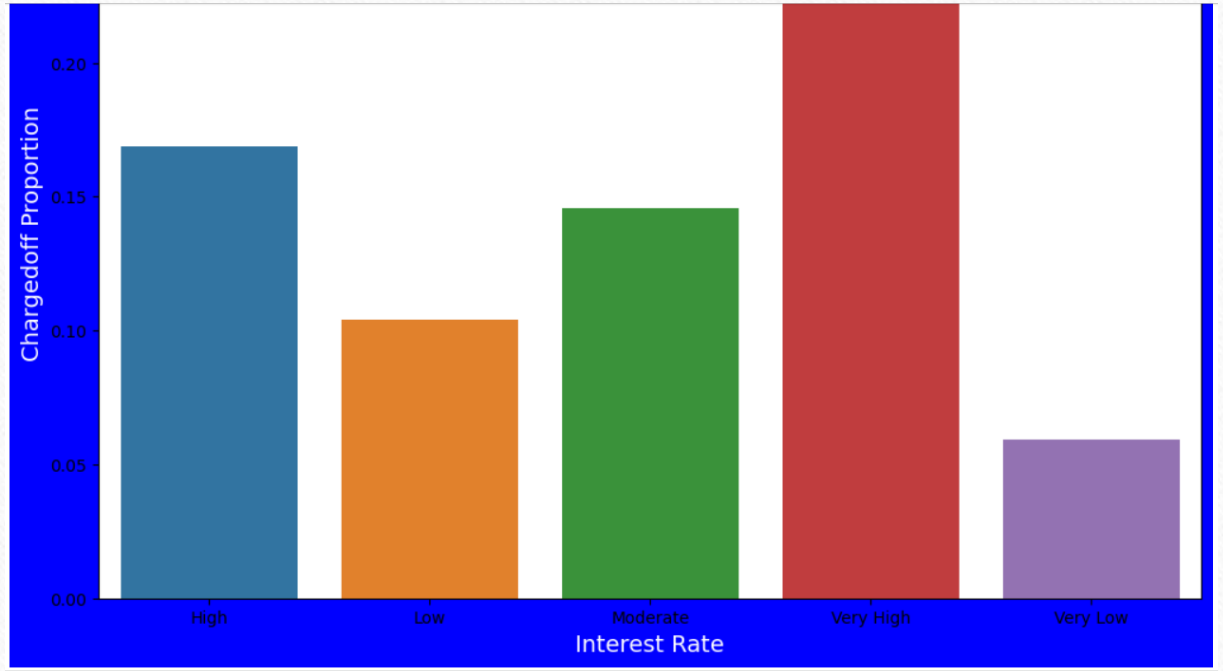
Annual Income vs Charged Off

1. Applicants with an income above 80,000 are less likely to experience charge-offs.
2. Applicants with an income between 0 and 20,000 are more likely to experience charge-offs.
3. It is observed that as annual income increases, the proportion of charge-offs decreases.



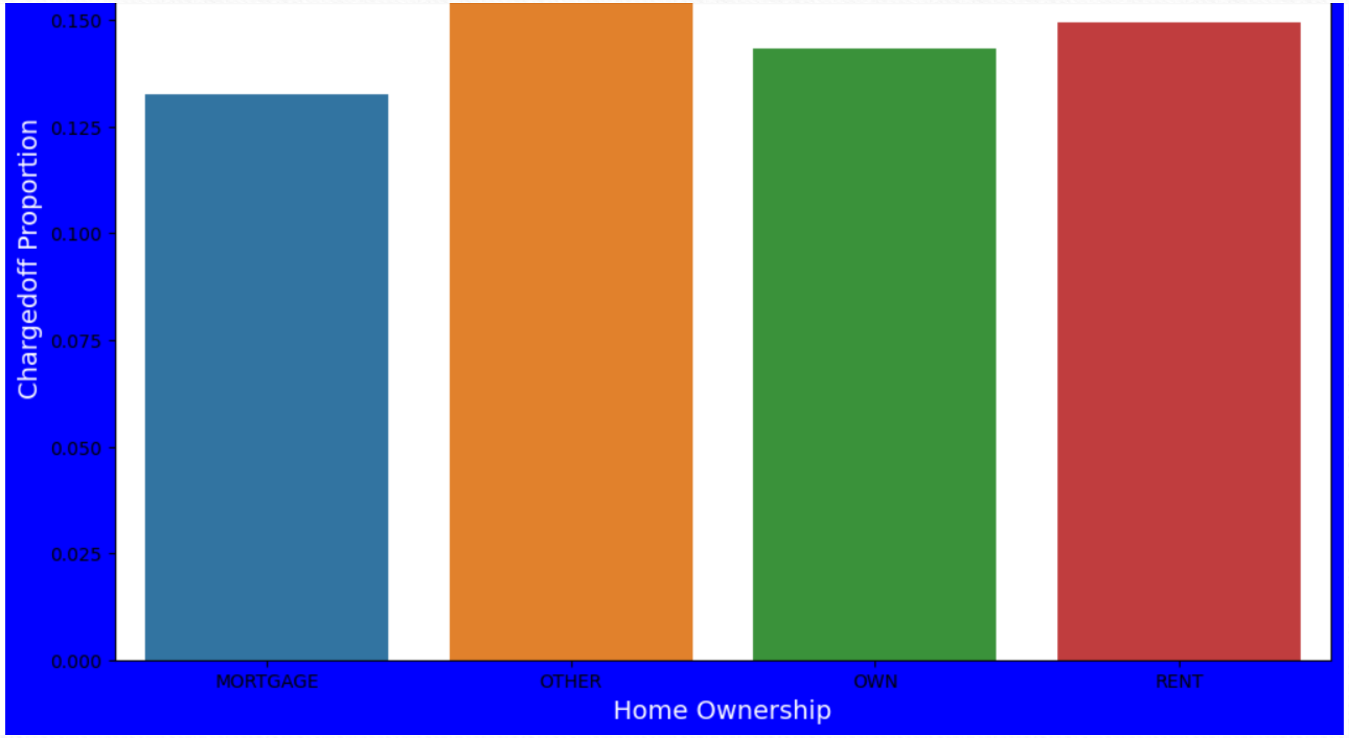
Interest Rate vs Charged Off

1. Loans with interest rates below 10%, particularly those starting from a minimum of 5%, have a very low likelihood of being charged off.
2. Loans with interest rates above 16%, categorized as very high, are more likely to experience charge-offs compared to other interest rate categories.
3. The proportion of charged-offs tends to increase with higher interest rates.



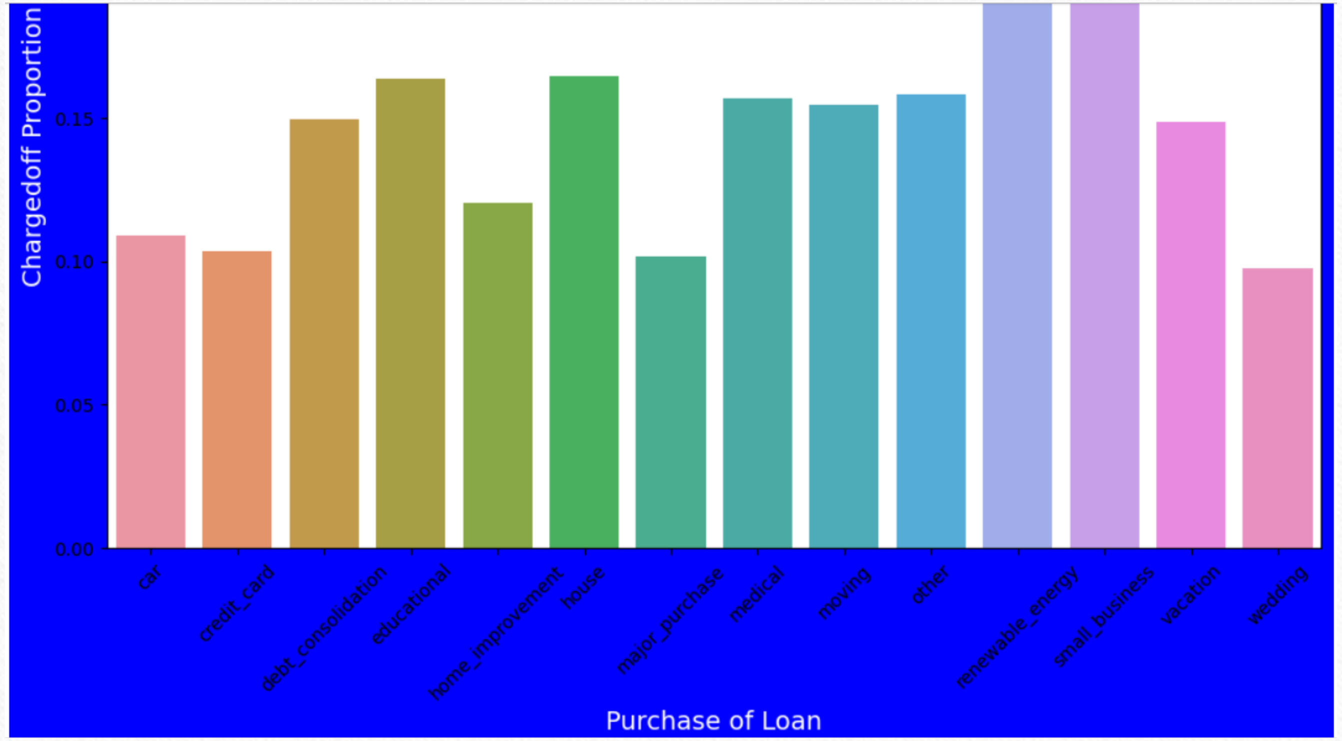
Home Ownership vs Charged Off

1. Individuals who do not own a home have a higher likelihood of being loan defaulters.
2. The graph also indicates a higher proportion of charge-offs, although the available data for this category is limited compared to others.



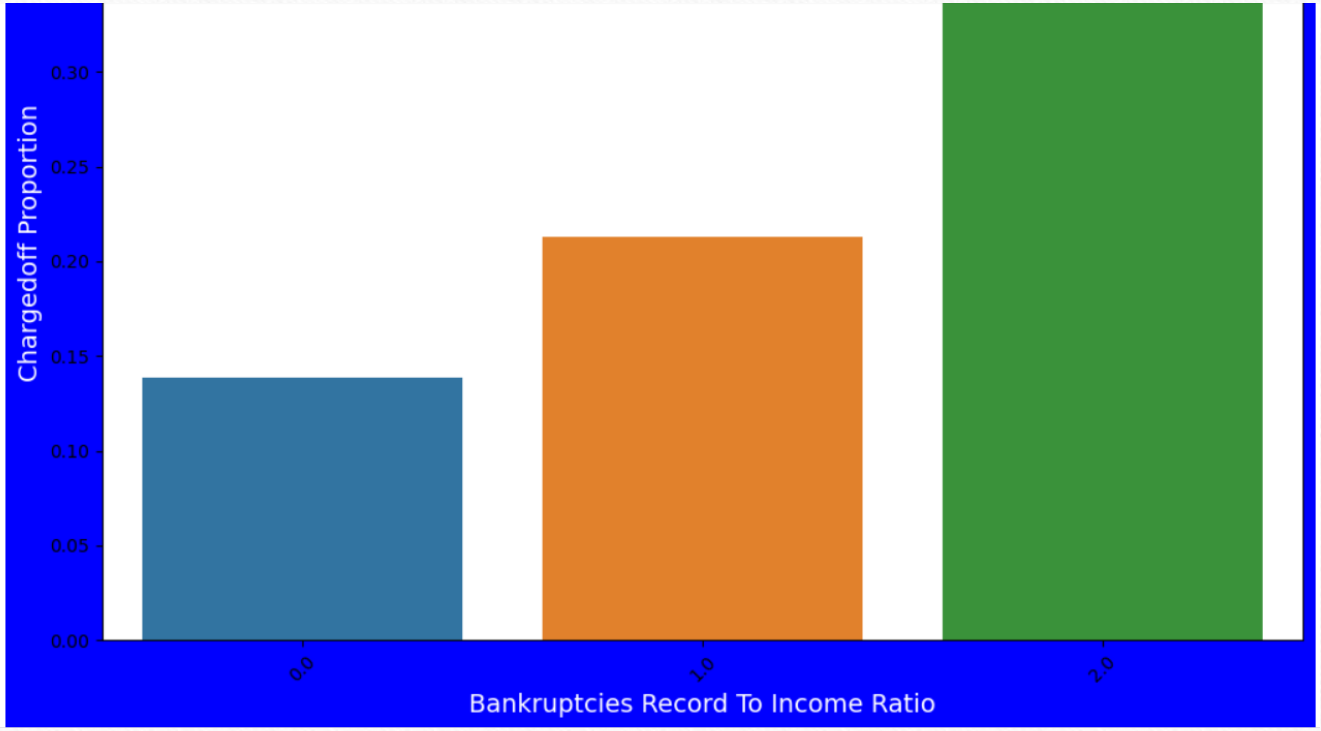
Purpose vs Charged Off

1. Applicants with a home loan have a lower likelihood of loan defaults.
2. Applicants with loans for small businesses have a higher likelihood of loan defaults.



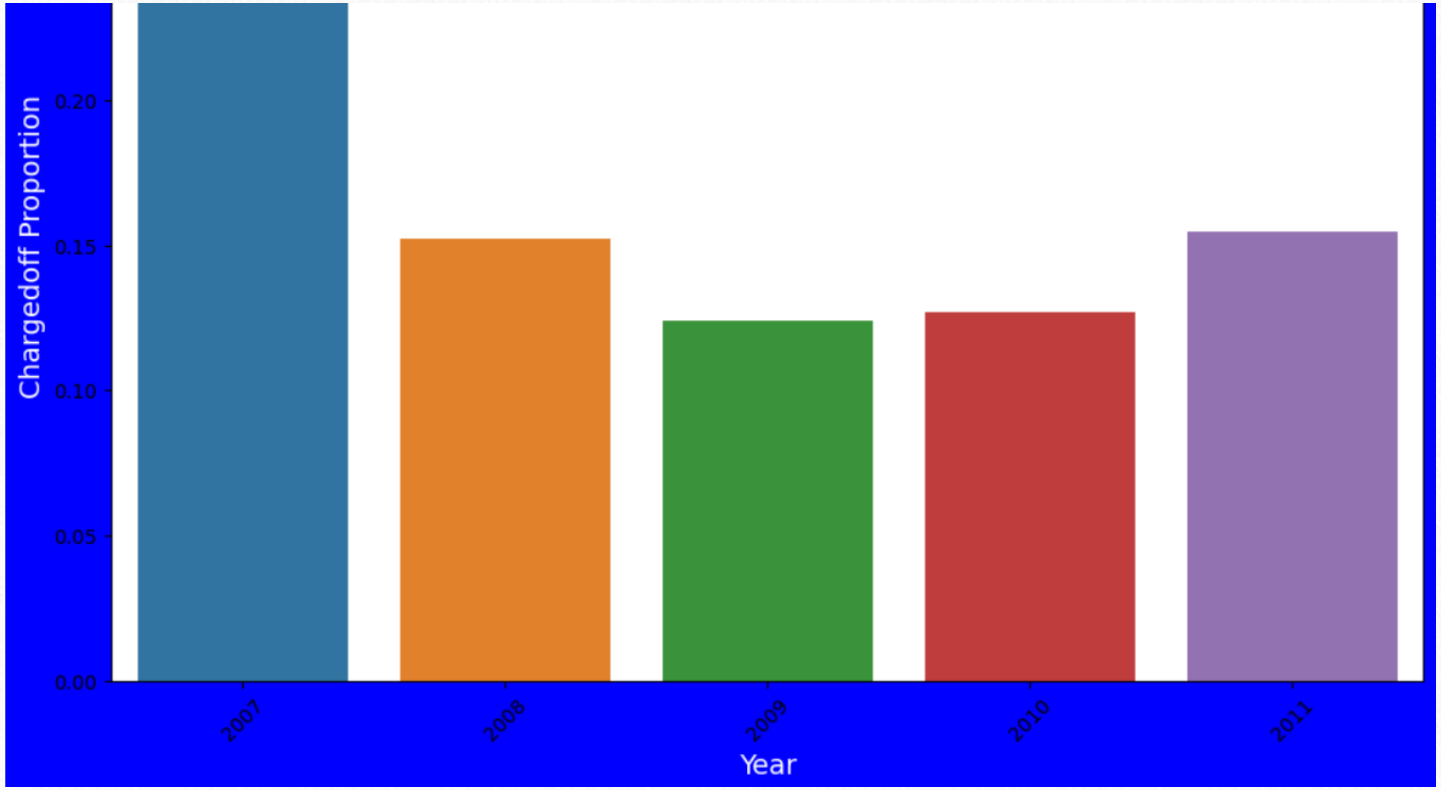
Bankruptcies Record vs Charged off

1. A bankruptcy record with 2 has a significant impact on loan defaults.
2. A bankruptcy record with 0 has minimal impact on loan defaults.
3. Lower bankruptcy records correspond to lower risk levels.



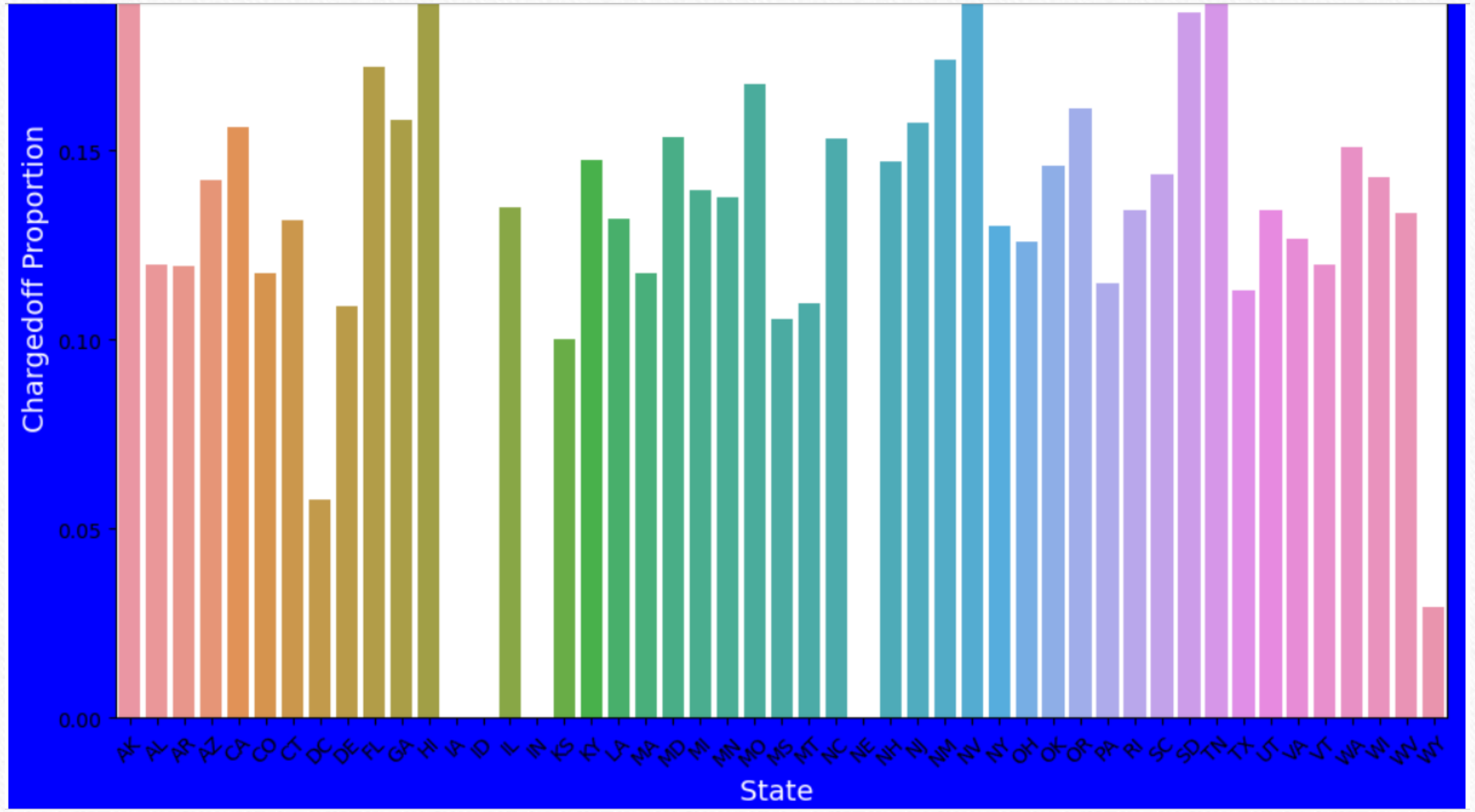
Issue Year vs Charged off

1. The year 2007 experienced the highest number of loan defaults.
2. The year 2009 saw the lowest number of loan defaults.



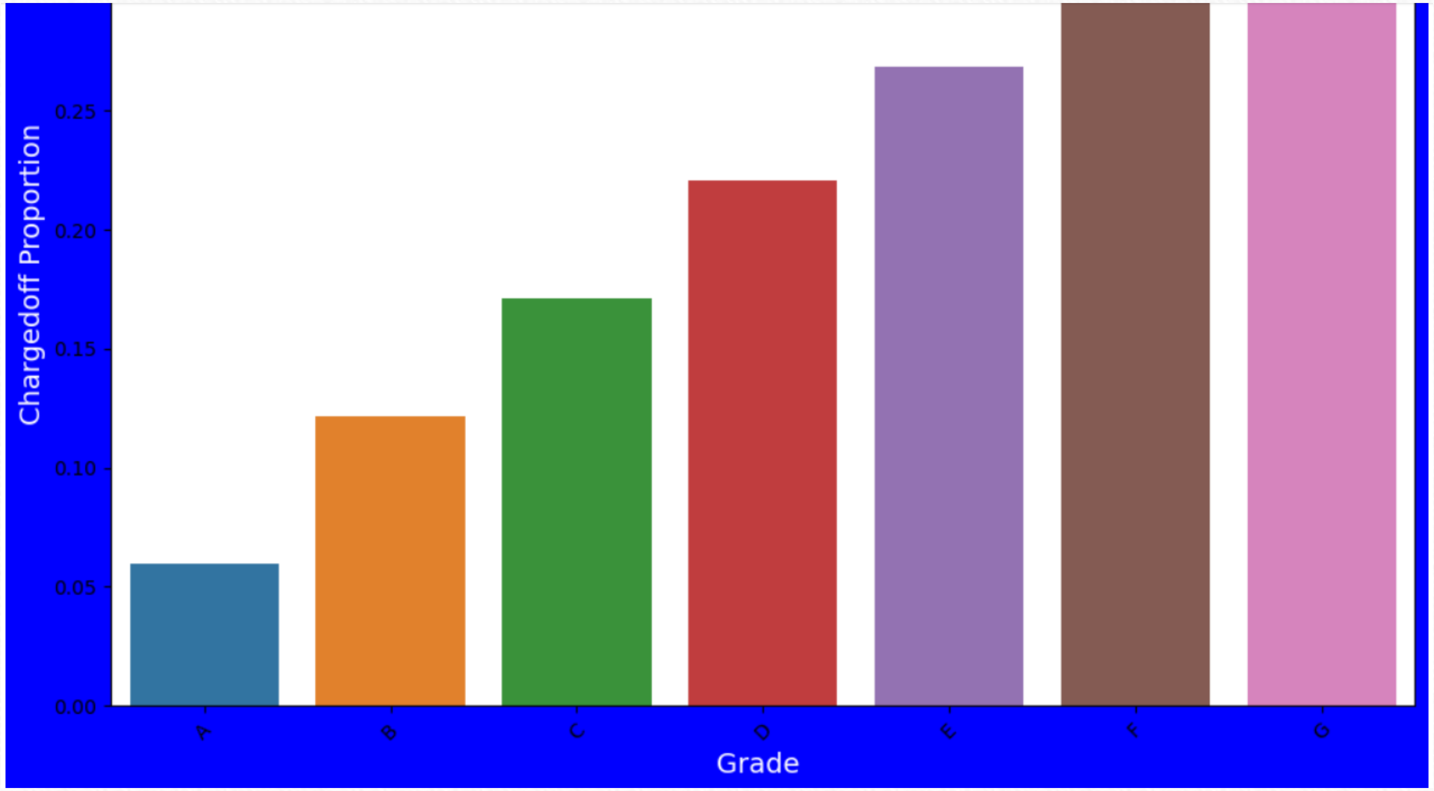
State vs Charged off

1. Delaware (DE) has the highest number of loan defaults.
2. California (CA) has a low number of loan defaults.



Grade vs Charged off

1. Loan applicants with loan grade G have the highest loan defaults.
2. Loan applicants with loan grade A have the lowest loan defaults.



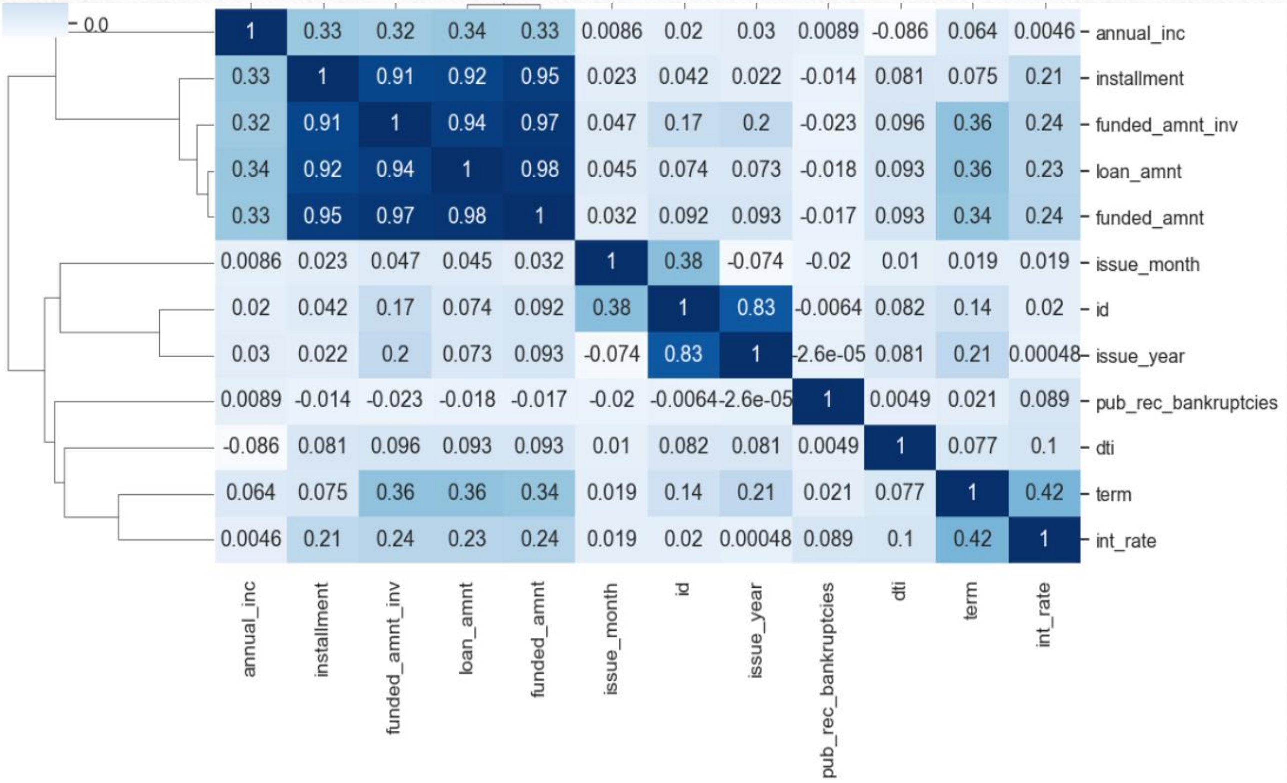
Correlations

Negative Correlation:

1. Loan amount has a negative correlation with pub_rec_bankruptcies.
2. Annual income has a negative correlation with dti.

Strong Correlation:

1. Term has a strong correlation with loan amount.
2. Term has a strong correlation with interest rate.
3. Annual income has a strong correlation with loan amount.



	annual_inc	installment	funded_amnt_inv	loan_amnt	funded_amnt	issue_month	id	issue_year	pub_rec_bankruptcies	dti	term	int_rate
annual_inc	1	0.33	0.32	0.34	0.33	0.0086	0.02	0.03	0.0089	-0.086	0.064	0.0046
installment	0.33	1	0.91	0.92	0.95	0.023	0.042	0.022	-0.014	0.081	0.075	0.21
funded_amnt_inv	0.32	0.91	1	0.94	0.97	0.047	0.17	0.2	-0.023	0.096	0.36	0.24
loan_amnt	0.34	0.92	0.94	1	0.98	0.045	0.074	0.073	-0.018	0.093	0.36	0.23
funded_amnt	0.33	0.95	0.97	0.98	1	0.032	0.092	0.093	-0.017	0.093	0.34	0.24
issue_month	0.0086	0.023	0.047	0.045	0.032	1	0.38	-0.074	-0.02	0.01	0.019	0.019
id	0.02	0.042	0.17	0.074	0.092	0.38	1	0.83	-0.0064	0.082	0.14	0.02
issue_year	0.03	0.022	0.2	0.073	0.093	-0.074	0.83	1	-2.6e-05	0.081	0.21	0.00048
pub_rec_bankruptcies	0.0089	-0.014	-0.023	-0.018	-0.017	-0.02	-0.0064	-2.6e-05	1	0.0049	0.021	0.089
dti	-0.086	0.081	0.096	0.093	0.093	0.01	0.082	0.081	0.0049	1	0.077	0.1
term	0.064	0.075	0.36	0.36	0.34	0.019	0.14	0.21	0.021	0.077	1	0.42
int_rate	0.0046	0.21	0.24	0.23	0.24	0.019	0.02	0.00048	0.089	0.1	0.42	1

Conclusion

1. Income range between 0 and 20,000 has a high likelihood of being charged off.
2. Interest rates above 16% have a higher chance of being charged off compared to other interest rate categories.
3. Non-homeowners have a higher chance of loan default.
4. Applicants with loans for small businesses have a high likelihood of loan defaults.
5. High DTI values are associated with a high risk of defaults.
6. The higher the bankruptcy record, the greater the chance of loan defaults.
7. Delaware (DE) has the highest number of loan defaults.
8. Loan applicants with grade G have the highest loan defaults.

The above factors will cause an applicant to default the loan.