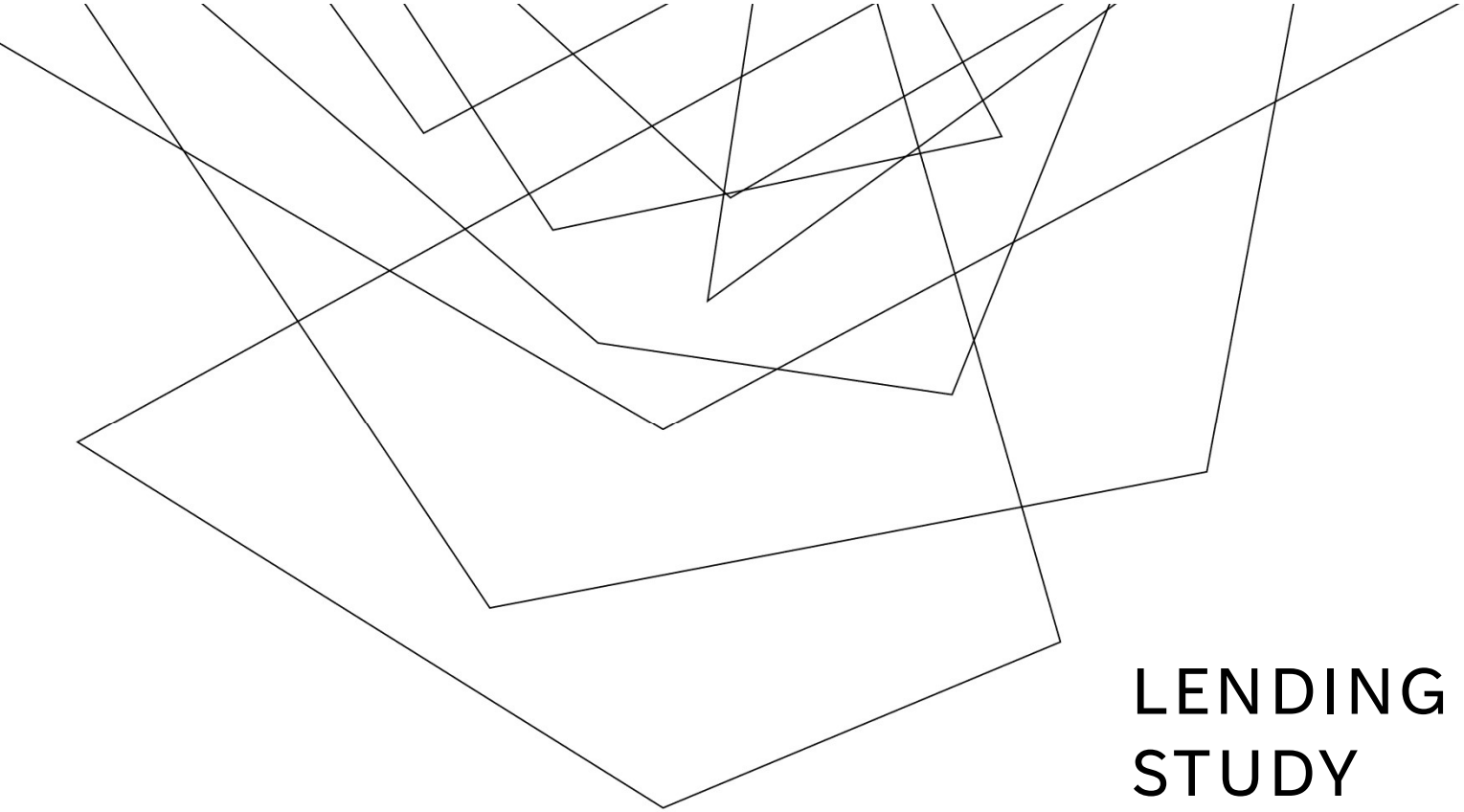
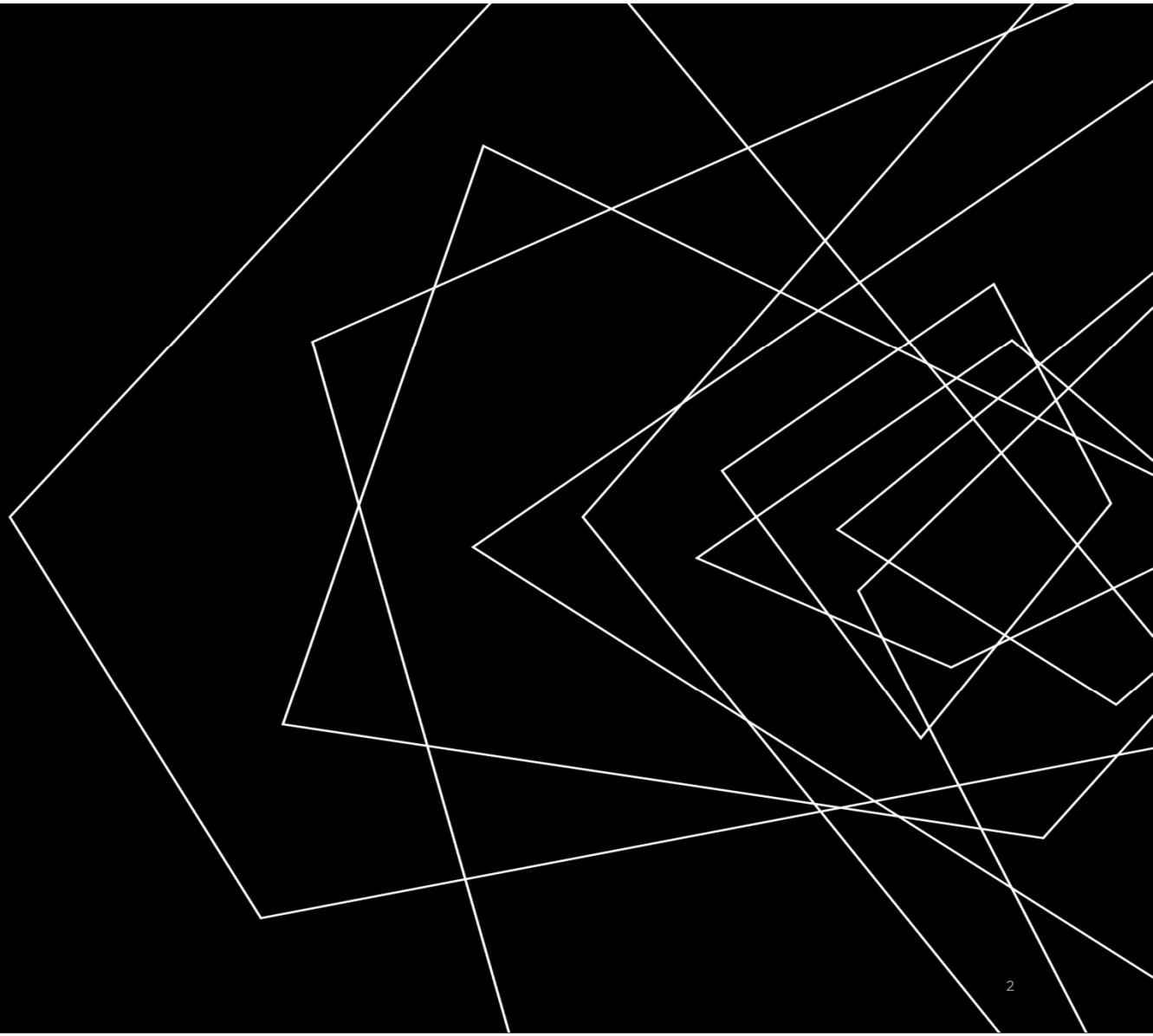


LENDING CLUB CASE STUDY

By: Milind Awade & Vishal Ganguly
ML-C67

TOPICS

- Problem Statement
- Objective
- Constraints
- Data Used
- Preparing Data
    - Cleaning
    - Imputation
- Data Conversion
- Treating Outliers
- Functions
- Analysis:
    - Univariate
    - Bivariate
    - Correlations
- Conclusion



PROBLEM STATEMENT

You work for a consumer finance company that specializes in offering various types of loans to urban customers. When a loan application is submitted, the company must decide whether to approve it based on the applicant's profile.

This decision involves two potential risks:

- If the applicant is likely to repay the loan, rejecting the application leads to a missed business opportunity for the company.
- If the applicant is likely to default, approving the loan could result in a financial loss for the company.

OBJECTIVE

Perform Exploratory Data Analysis (EDA) to investigate how consumer characteristics and loan features impact the likelihood of loan default.

CONSTRAINTS:

When a person applies for a loan, there are **two types of decisions** that could be taken by the company:

1. Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:

1. **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
2. **Current**: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
3. **Charged-off**: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has **defaulted** on the loan

2. Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

DATA USED

Loan.csv file is provided for conducting the EDA.

This dataset contains 39717 rows and 111 columns.

In this dataset, there are two types of Attributes : Customer and Loan

Data Requires cleaning and imputation before analysis.

PREPARING DATA

DATA CLEANING:

1. No duplicates rows found
2. Dropped 1140 rows with loan status as 'CURRENT'; as the loan is already granted and will not be useful in conducting analysis for the customer who may defaults in future.
3. Removing 55 columns with Null/Missing value as it wont be needed in analysis.
4. Deleting columns with unique value or value as 1, as it does not contribute in EDA, total 11 columns found. *(pymnt_plan, initial_list_status, out_prncp, out_prncp_inv, collections_12_mths_ex_med, policy_code, application_type, acc_now_delinq, chargeoff_within_12_mths, delinq_amnt, tax_liens)*
5. Removed Non contributing textual field *'url', 'title', 'emp_title', 'desc',* as it won't be helpful in analysis.
6. For the business purpose, behavioural data is captured which doesn't participate in analysis. Hence dropping the 21 behavioural columns.
7. Additionally there are 2 columns having more than 50% of data as NA has removed.
8. Finally after the cleaning process, we are left with 18 columns and 38577 rows for analysis.

PREPARING DATA

DATA CONVERSION / IMPUTATION:

1. Converting ‘term’ column to int data type by stripping text/characters .
2. Similarly converting ‘int_rate’ of string to float by striping ‘%’.
3. Converted ‘loan_funded_amnt’ and ‘funded_amnt’ converted to float.
4. Following columns ‘loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'int_rate', ‘dti' values are rounded off to decimal points.
5. ‘issue_d’ has been converted to date.
6. Created bins (range) as derived columns : ‘loan_amnt_b’, ‘annual_inc_b’, ‘int_rate_b, and ‘dti_b’ (for multiple group range of data) for better analysis.

Treating Outliers and Imputing Rows:

1. Columns - ‘emp_length’ and ‘pub_rec_bankruptcies’ contains a small % of data as null - 2.67% and 1.80% respectively, hence dropped them.
2. Overall % of rows deleted with null values : 4.48%,
3. As Outliers exits for data columns 'loan_amnt’, ‘annual_inc’, 'int_rate’, ‘dti’and in order to treat this outliers, **quantile method** has been used to treat them.

FUNCTIONS AND MORE..

Lambda functions used for creating bin groups

```
# Grouping Loan Amount for meaningful ranges
df['loan_amnt_b'] = df['loan_amnt'].apply(
    lambda x: '0 - 5K' if x <= 5000 else
              '5K - 10K' if 5000 < x <= 10000 else
              '10K - 15K' if 10000 < x <= 15000 else
              '15K - above'
)
```

```
# Grouping Annual Income for meaningful ranges
df['annual_inc_b'] = df['annual_inc'].apply(
    lambda x: '0 - 40k' if x <= 40000 else
              '40k - 50k' if 40000 < x <= 50000 else
              '50k to 60k' if 50000 < x <= 60000 else
              '60k to 70k' if 60000 < x <= 70000 else
              '70k to 80k' if 70000 < x <= 80000 else
              '80k - above'
)
```

```
# Grouping Interest Rate for meaningful ranges

df['int_rate_b'] = df['int_rate'].apply(
    lambda x: 'Very Low (< 9)' if x <= 9 else
              'Low (9 - 11)' if 9 < x <= 11 else
              'Moderate (11 - 13)' if 11 < x <= 13 else
              'High (13 - 15)' if 13 < x <= 15 else
              'Very High  (> 15)'
)
```

For Plotting the graph a generic function is created, it is a generic function to help in creating plots for univariate analysis called "**plt_graph**"

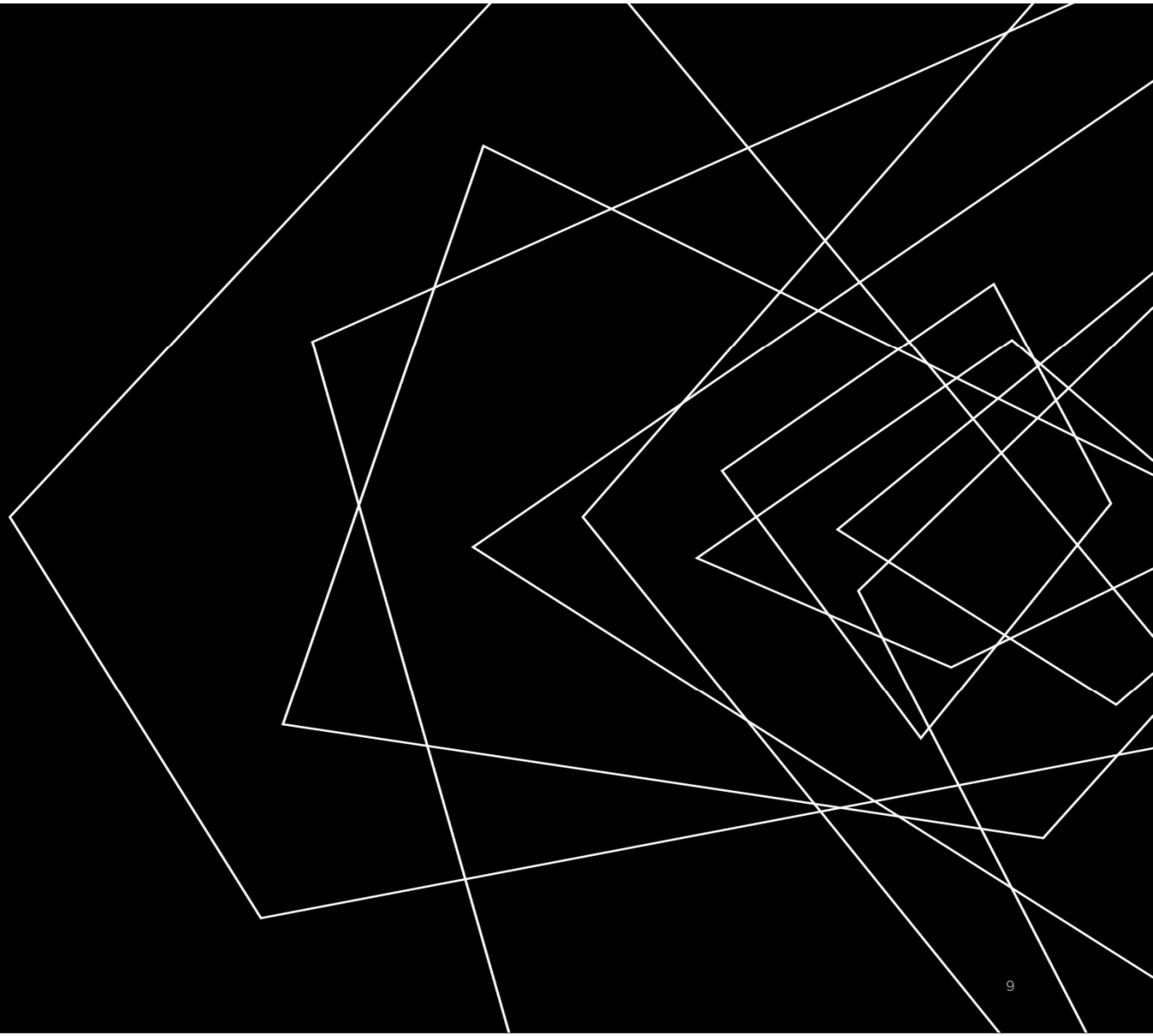
```
# Creating a function to generate graph
def plt_graph(dfrm, column):
    fig, p = plt.subplots(1,2, figsize=(16, 4))
    sea.distplot(dfrm.loc[dfrm[column].notnull(), column], kde=True, hist=True, ax=p[0])
    sea.boxplot(x=column, data=dfrm, ax=p[1])
    p[0].set_xlabel(col_Titles[column])
    p[1].set_xlabel(col_Titles[column])
    plt.tight_layout()
    plt.show()
```

Column Labels are added for Axis Titles / Legends

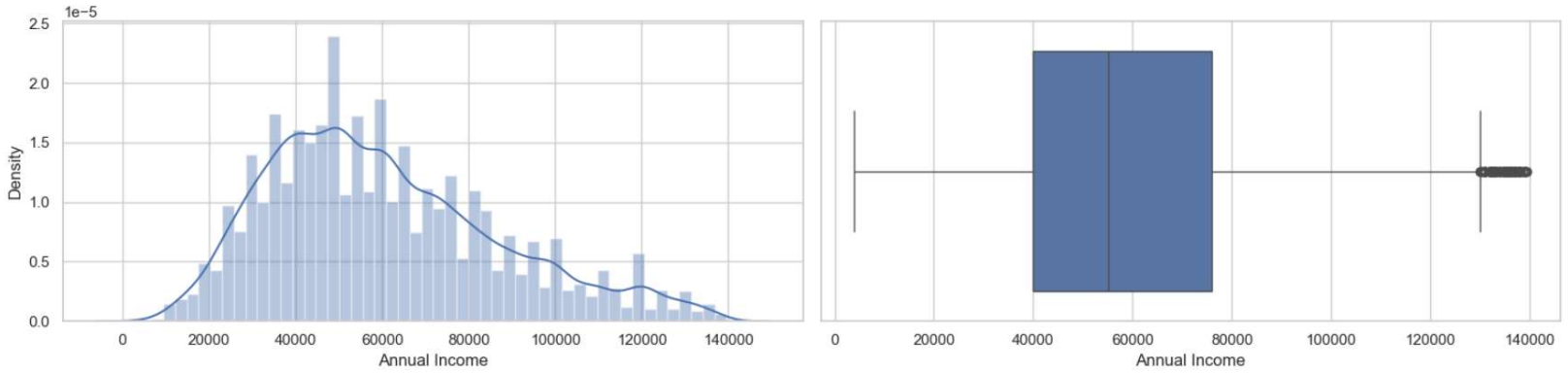
```
# Refering Metadata Definition for using labels/Legends on graph plots

col_Titles = {
    'grade': 'Grade',
    'emp_length': 'Employment Length',
    'loan_amnt': 'Loan Amount',
    'term': 'Loan Term',
    'int_rate': 'Interest Rate',
    'home_ownership': 'Home Owner Status',
    'annual_inc': 'Annual Income',
    'annual_inc_b': 'Annual Income',
    'verification_status': 'Verification Status',
    'issue_d': 'Issue Date',
    'loan_status': 'Loan Status',
    'purpose': 'Purpose of Loan',
    'addr_state': 'State',
    'dti': 'Debt To Income Ratio',
    'pub_rec_bankruptcies': 'Bankruptcies Record',
    'loan_amnt_b': 'Loan Amount Bin',
    'int_rate_b': 'Interest Rate Bin',
    'dti_b': 'DTI Bin'
}
```

UNIVARIATE ANALYSIS



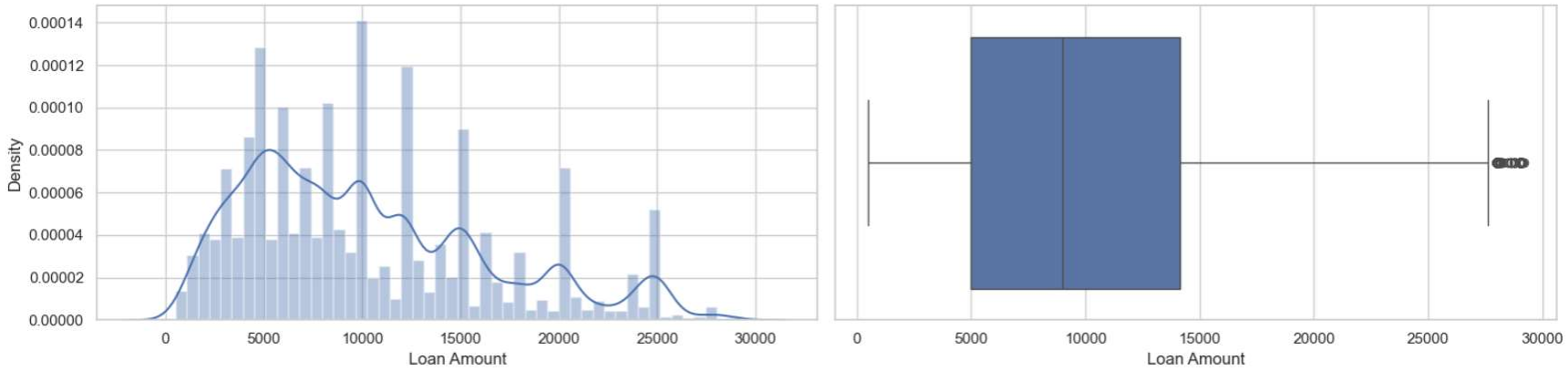
ANNUAL INCOME



Observation:

1. Annual Average Income is : 60314
2. Annual income of most applicants are in between the range 40000 – 75000

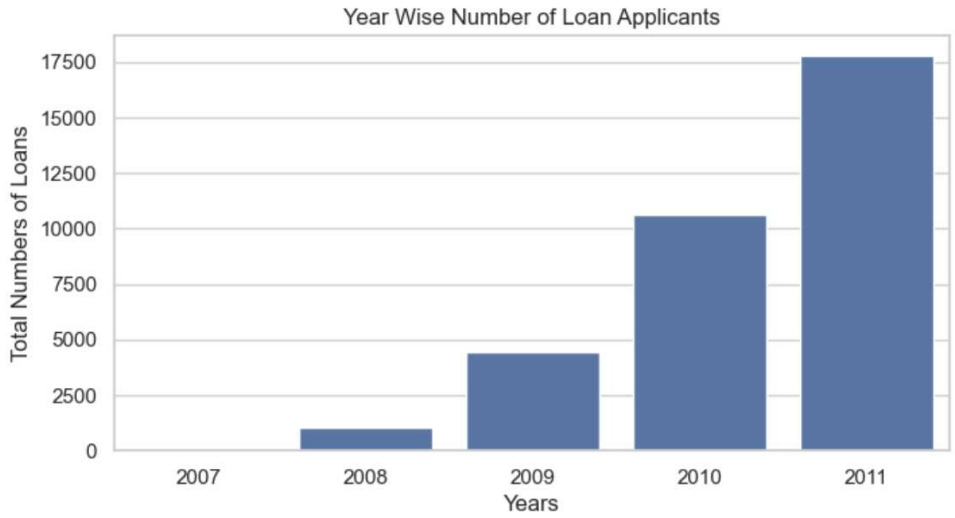
FUNCTIONS



Observation:

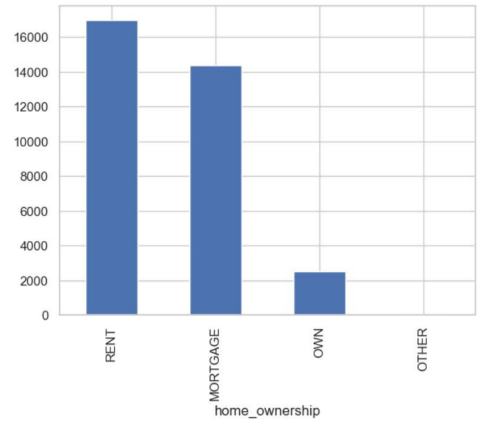
- Max Loan amount applied was ~ 29000
- Most of the loan amount applied was in the range of 5000 - 140000

YEAR WISE APPLICANTS

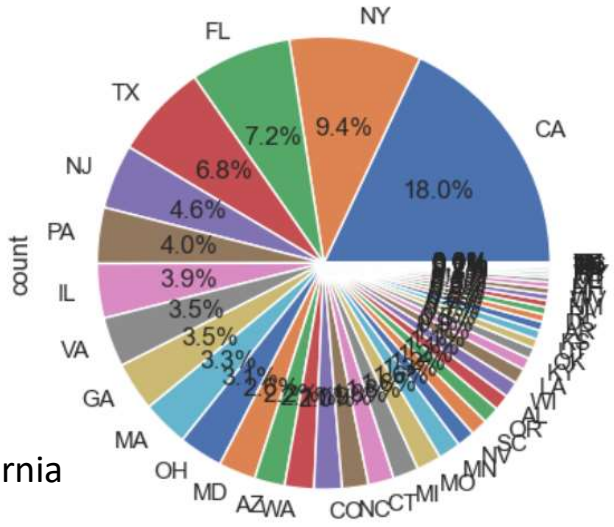


Observation: We can see there is a Substantial growth took place between 2008-2011 in the number of loan applicants

CATEGORICAL VARIABLE ANALYSIS

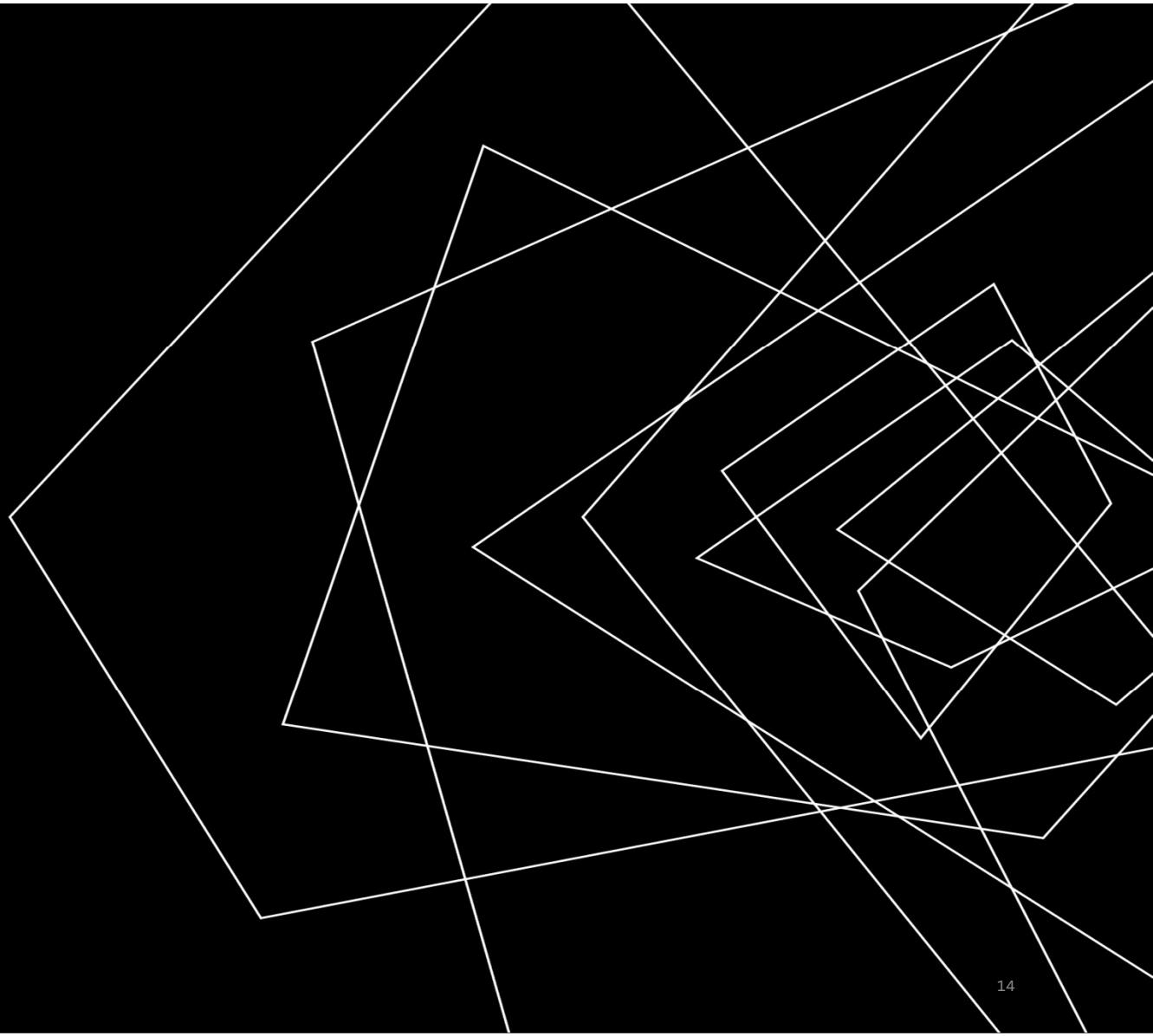


Observation: Majority of applicants are either living on Rent or on Mortgage

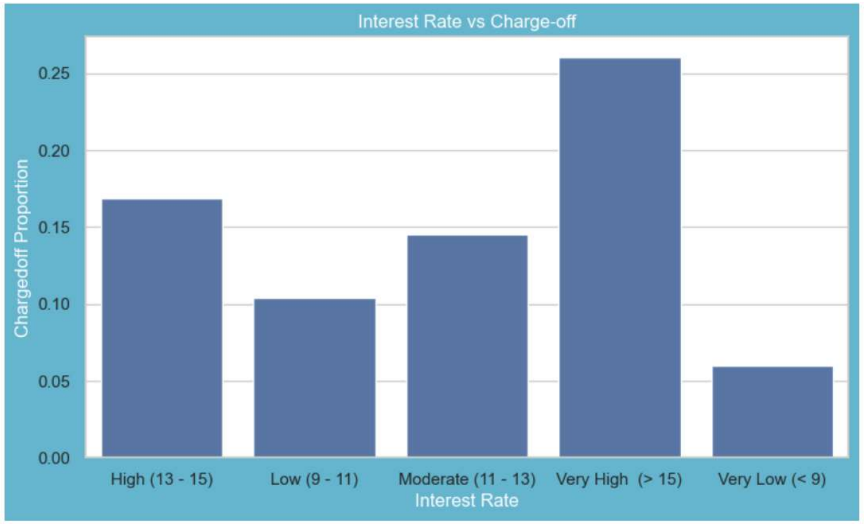


Observation: Most of the Loan applicants are from California State

BIVARIATE ANALYSIS



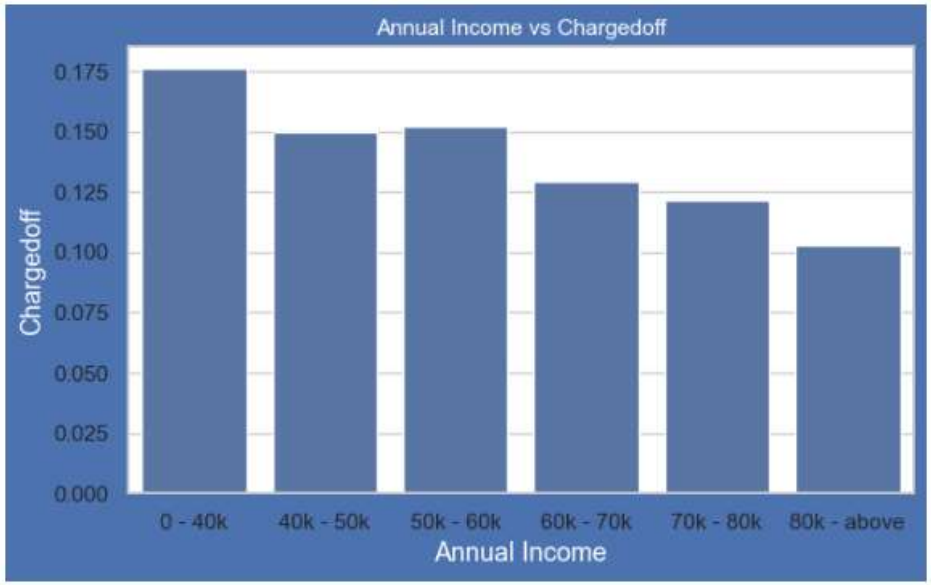
INTEREST RATE Vs CHARGE-OFF



Observations:

- Interest rates having less than 9% have very low chances of charged off.
- Interest rate with more than 15% have good chances of charged off as compared to other category interest rates.
- Charged off increases with higher interest rates.

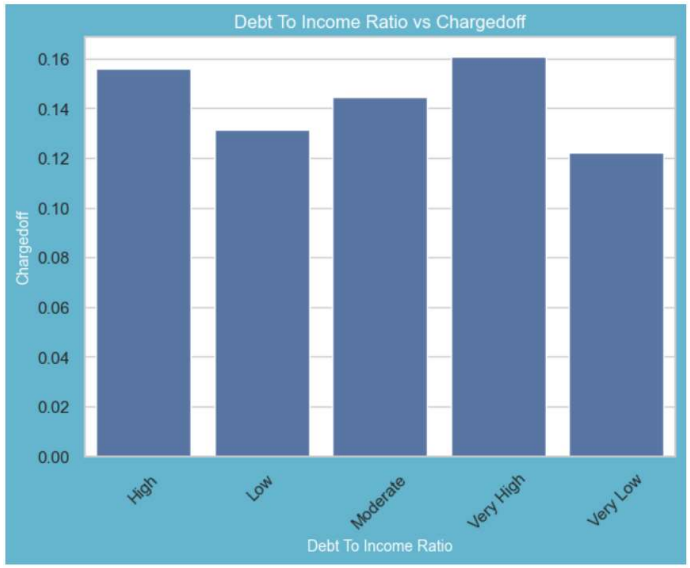
ANNUAL INCOME Vs CHARGE-OFF



Observations:

- Income range 0-20000 has higher chances of charged off.
- Income range of 80000 & above has less chances of charged off.
- With the Increase in annual income decreases the charged off.

DEBT TO INCOME RATIO Vs CHARGED-OFF

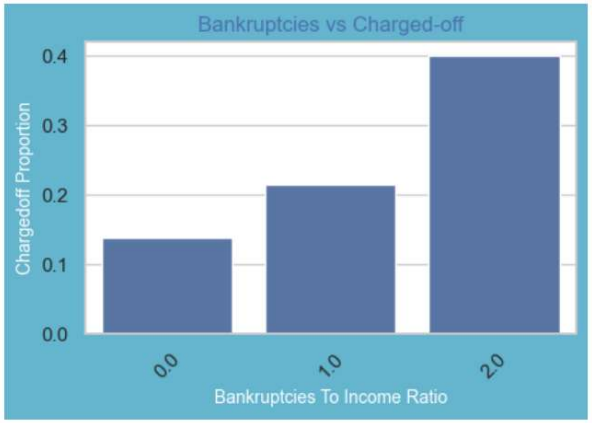


Observation:

- Higher the DTI value having high risk of defaults

- Lower the DTI having low chances of loan defaults

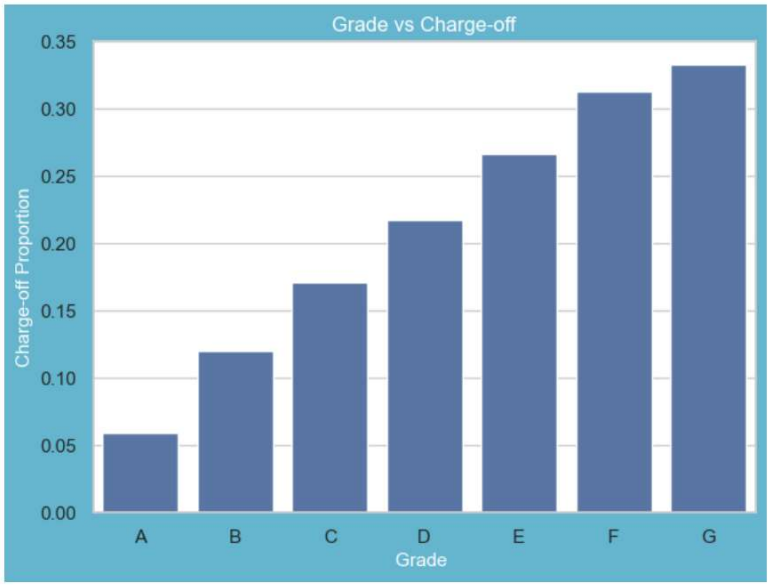
BANKRUPTCIES Vs CHARGED-OFF



Observations:

- Bankruptcies with 2 is having high impact on loan defaults
- Lower the Bankruptcies lower the risk

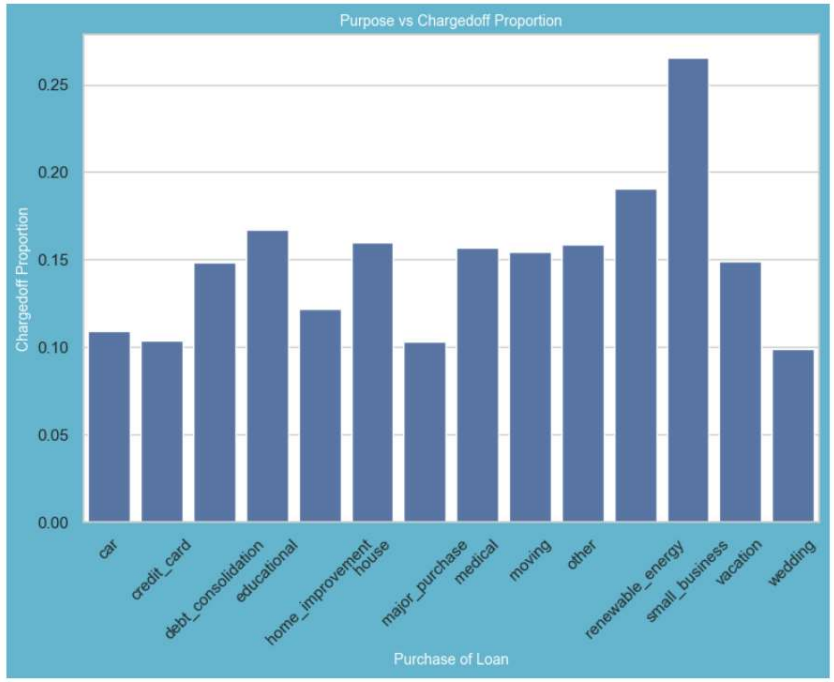
GRADE Vs CHARGE-OFF



Observations :

- The Loan applicants with loan Grade G is having highest Loan Defaults.
- The Loan applicants with loan A is having lowest Loan Defaults.

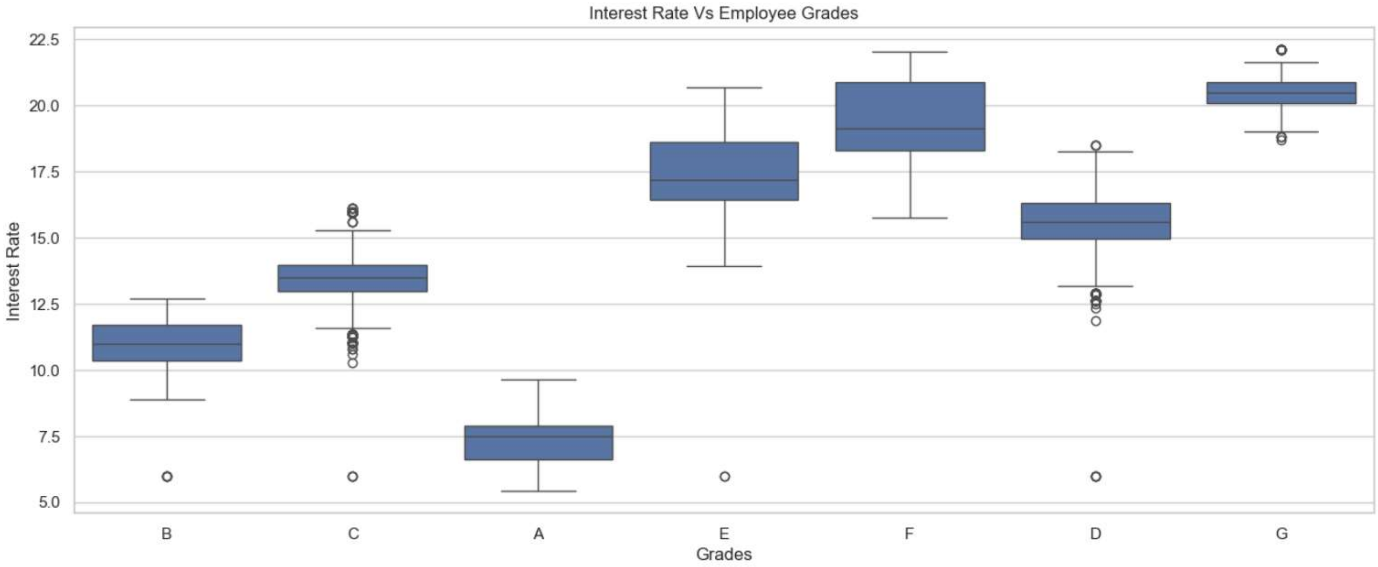
GRADE Vs CHARGE-OFF



Observation:

- Applicants having loan for small business is having high chances for loan defaults.
- Applicants having home loan is having low chances of loan defaults.

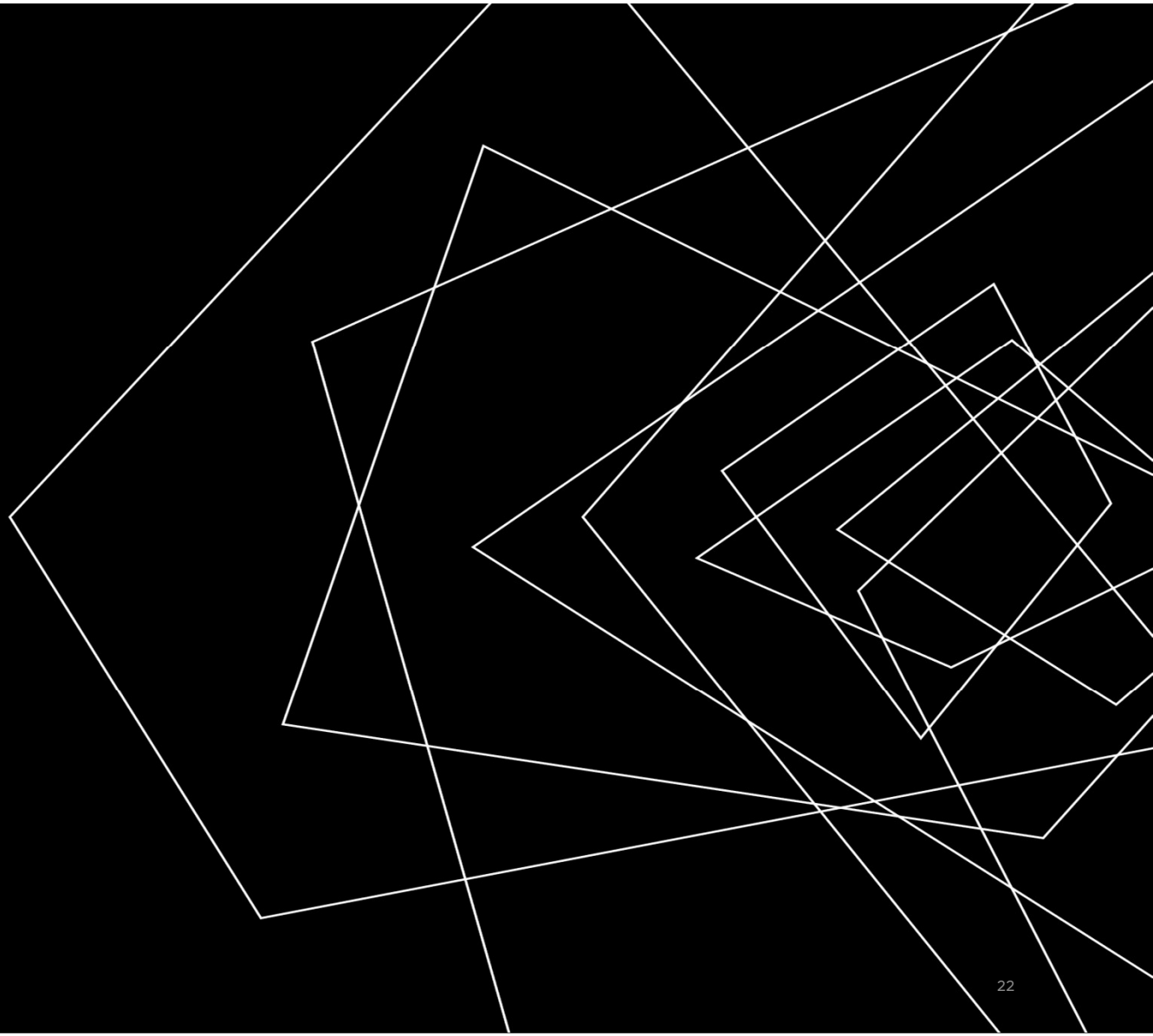
INTEREST RATE Vs EMPLOYEE GRADES



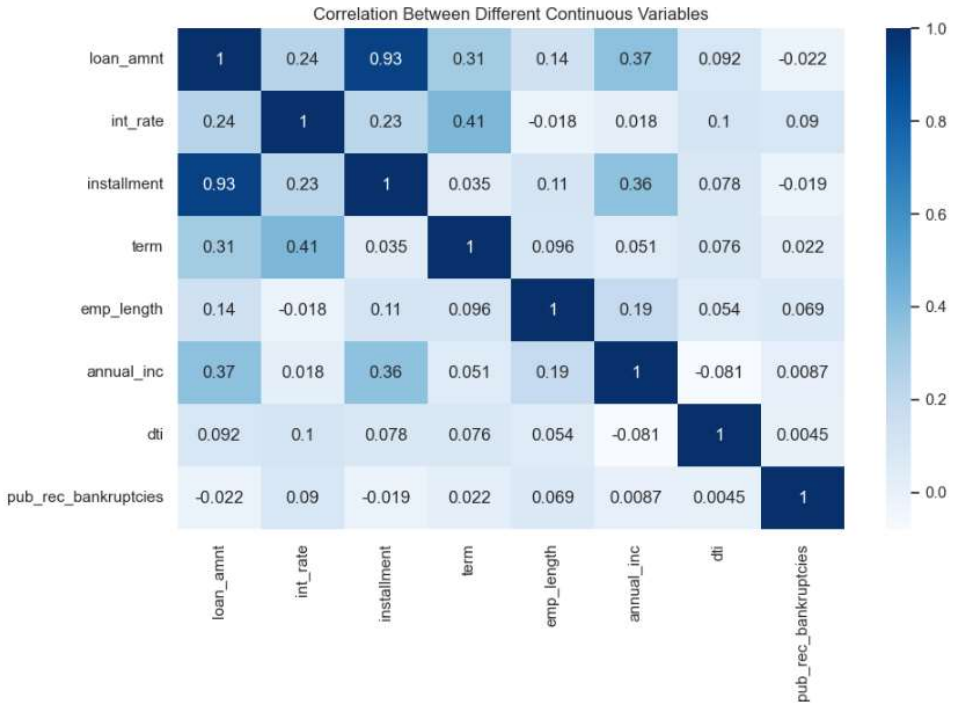
Observation:

- Higher the grade higher the interest rate
- ‘A’ Grade is having minimum interest rate, whereas ‘G’ is having highest interest rate

CORRELATION ANALYSIS



Correlation Between Different Continuous Variables



Positive Correlation:

•Term has a positive correlation with "Interest Rate" and "Loan Amount“

•Annual Income has a strong correlation with "Loan Amount" and "Installments"

Negative Correlation:

•"Annual income" has a negative correlation with "DTI“

•"Loan Amount" has negative correlation with "Bankruptcies"

Conclusion

- Interest rate more than 15% has good chances of charged off as compared to other category interest rates.
- Income range 0-20000 has higher chances of charged off.
- Higher the DTI value having high risk of defaults
- Applicants having loan for small business is having high chances for loan defaults.
- Higher the Bankruptcies record higher the chance of loan defaults.
- The Loan applicants with loan Grade G is having highest Loan Defaults

THANK YOU

Milind Awade

Vishal Ganguly

https://github.com/MilindAwade/LendingClubCaseStudy

ML-C67, Date: 24th Sep 2024