GRAMENER CASE STUDY

SUBMISSION

Group Name: Busy4

1. Member name: **Sujata Ray (Lead)**
2. Member name: **Bijayalaxmi Sahoo**
3. Member name: **Ashish Katiyar**
4. Member name: **Rajshekar Swadi**

Lending Case Study

Abstract:

A consumer finance company, is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures.

For each loan the company has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

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

Business objective (Problem statement):

The objective is to identify the risky applicants. The borrowers who refuses to pay (the defaulters – who are labelled as charged-off).

The company wants to understand the driving factors behind loan default. The company can utilize this knowledge for its portfolio and risk assessment.

Problem Analysis Approach

- Requirement Understanding
- Data sourcing and Understanding
- Data Cleanup/Manipulations
- Derived Metrics Creations
  - Type-driven metrics
  - Business-driven metrics
  - Data-driven metrics
- Data Analysis
  - Univariate Analysis
  - Segmented Univariate Analysis
  - Bivariate Analysis
- Data Visualizations for better Analysis
- Analysis Outcome / Recommendations / Conclusions

Analysis - Understanding

- **Requirement Understanding**

Company has provided loan data for its customers. As an analyst we need to understand the driving factors behind loan default based on the customer profile.

We need to identify top 5 driving factors (variables) needs to be understood to avoid risky applicants by the lending company.

- **Data Sourcing and Understanding**

The loan data is shared by the company in an csv file. The data is being sourced into R for understanding.

Following understanding (and some assumptions) could be done after loan data imported/explored:

- Has 39717 loan account details with 111 attributes for each account.
- Has following customer attributes (fields/columns) that are not useful for analysis **(REDUDANT):**
  - Has many attributes (fields/columns) that completely filled with NA (nearly 50+)
  - Also many attributes that has same values across customers (eg. Payment plan, initial list status..etc)
  - Many customer related information (such as area, zipcode, url, descriptions, reason, title etc)
  - Many loan related attributes (principal, accounts, revol account..etc)
  - Correlated fields – such as:
    - Interest rate, grade and subgrade (**only one will be used**)
    - Loan amount, Funded amount, funded amount investor (**only Funded amount will be used** as this the one that bank finally lended to the customer)

Analysis - Cleanup

- **Data Cleanup/Manipulations**
- As part of data understanding in the previous section we identified lot of **REDUNDANT** attributes for the loan applicants. All these are removed, namely – NA columns, correlated columns, non useful columns
- The DATE format is not proper – the same is being standardized
- We observe that many of the columns still have significant NA values. If any field has more than 20% of the values as NA, they may add value for analysis, better to get rid of such columns/fields.
- For loan default analysis we just need – Full Paid and Charged OFF data. the Current option that indicates ongoing customers who are still paying EMI does not help. Hence we filter and remove Current status option.
- Convert some fields from character to number – such as Interest rate and Employment length.
- We better convert all the data types to factor for better manipulations and analysis.

Analysis - Cleanup

- **Derived Metrics Creation**
- For better and narrow down analysis we need to created derived metrics from the exists data, so that analysis can done in a better way.
- **Funded Amount Categorization (Grouping)**
  - We see that this value ranges from 500 to 3500 and the mean is 1500.
  - With such a huge range it is difficult to analyze properly. So we will categorize them into 4 buckets/bins:
    - Small <= 5000
    - Medium >5000 & <=15000
    - High >15000 & <=25000
    - Very High >25000
- **Interest Rate Categorization (Grouping)**
  - We see that this value ranges from 5.42 to 24.40 and the mean is 11.93
  - So we will create bucket for the same:
    - Low <=10%
    - Medium >10% & <=15%
    - High >15%

Analysis - Cleanup

- **Derived Metrics Creation**
- **EMI or Installment Categorization (Grouping)**
  - We see that this value ranges from 15.69 to 1305.19 and the mean is 322.46.
  - So we will create bucket for the same:
    - Small <=200
    - Medium >200 & <=400
    - High >400 & <=600
    - Very High >600
- **Annual Income Categorization (Grouping)**
  - We see that this value ranges from 400 to 600000 to 24.40 and the mean is 68778
  - So we will create bucket for the same:
    - Small <=50000
    - Medium >50000 & <=100000
    - High > 100000 & <=150000
    - Very High >150000
- **Experience Categorization (Grouping)**
  - For this too we categorize into Fresher's, Junior, Senior and Expert.
- **We will create a new derived column for Year (From loan date) for yearly analysis.**

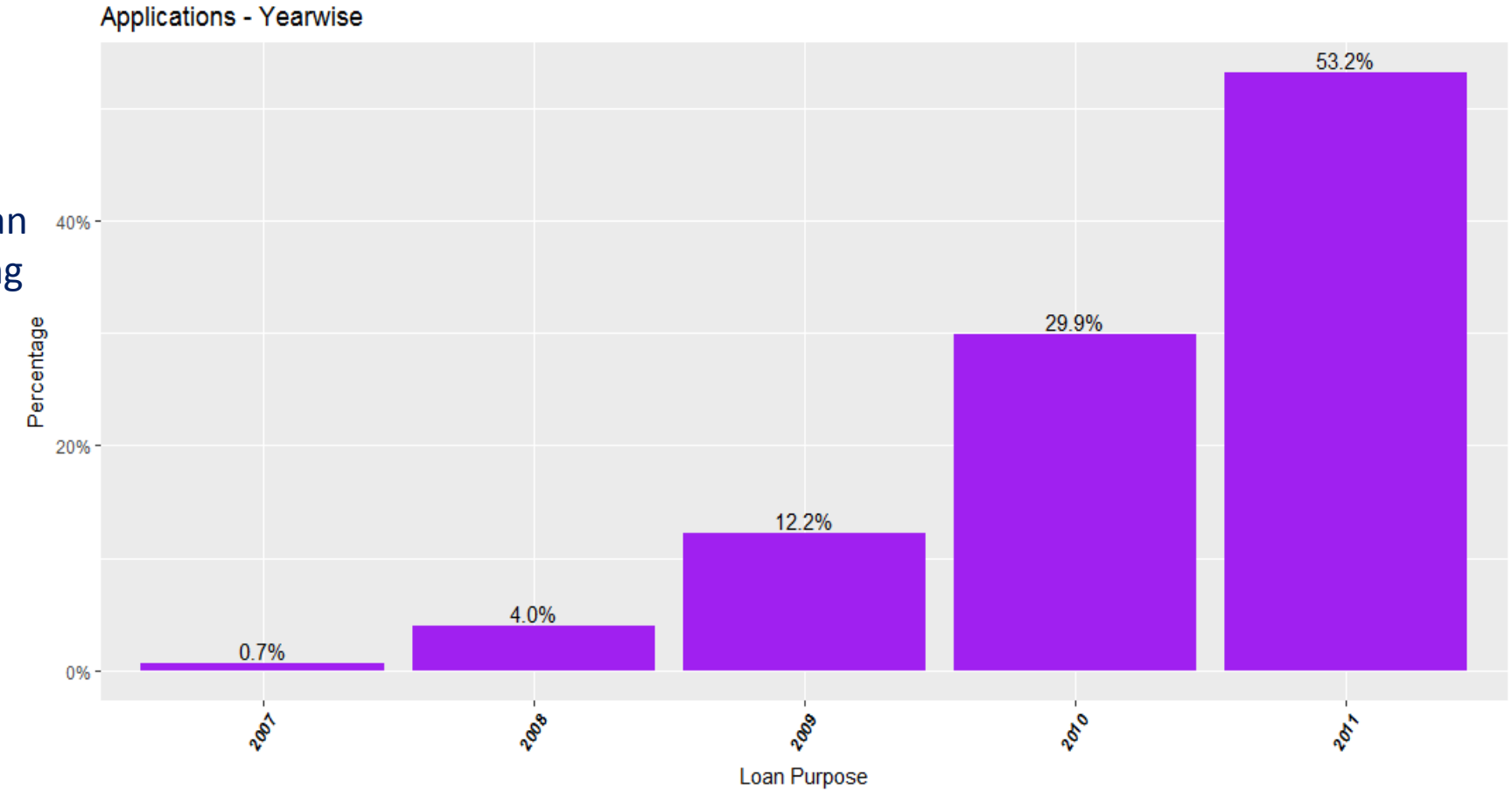
Analysis – Univariate

- **Univariate Analysis**

Analyzing one variable at a time. This gives a trend before we move to segmented and bivariate analysis. Following slides visualize the univariate analysis for the following variables against total number of loan **Applications processed:**

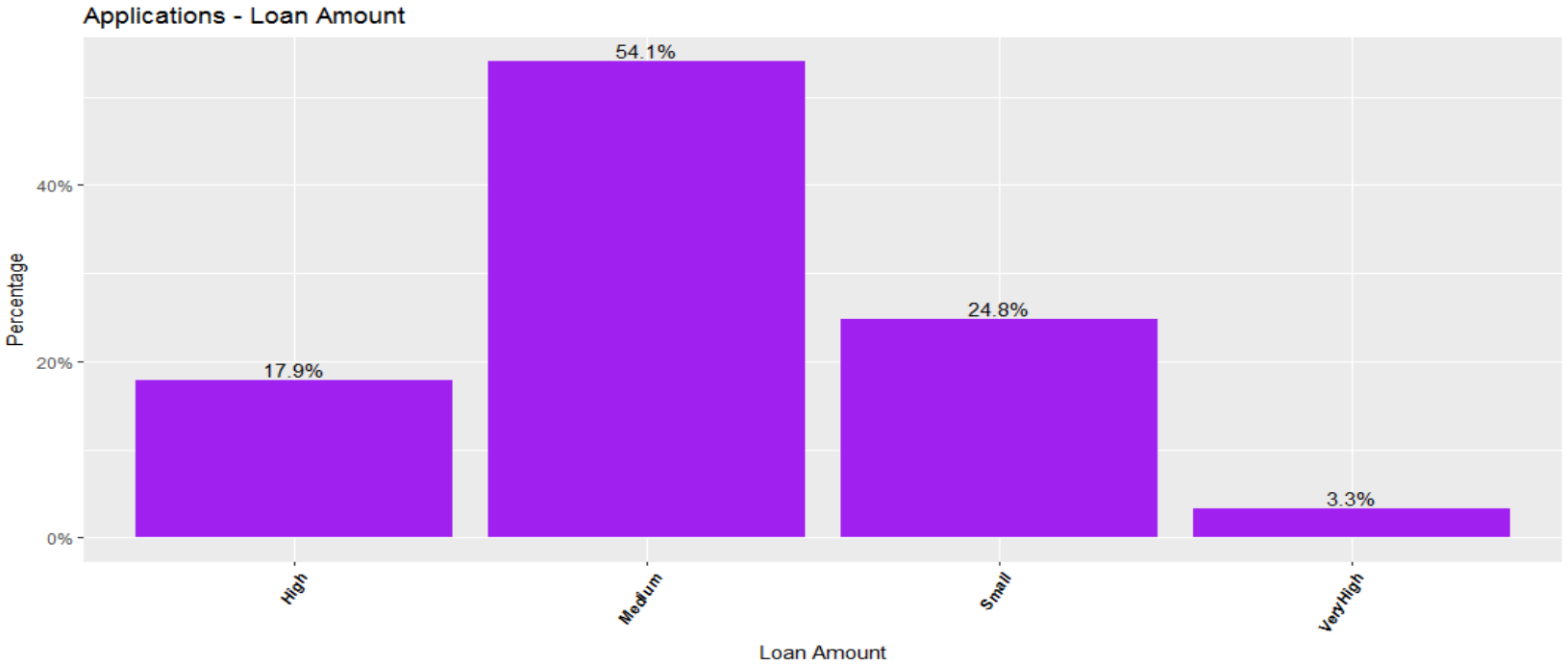
- **Year wise –**

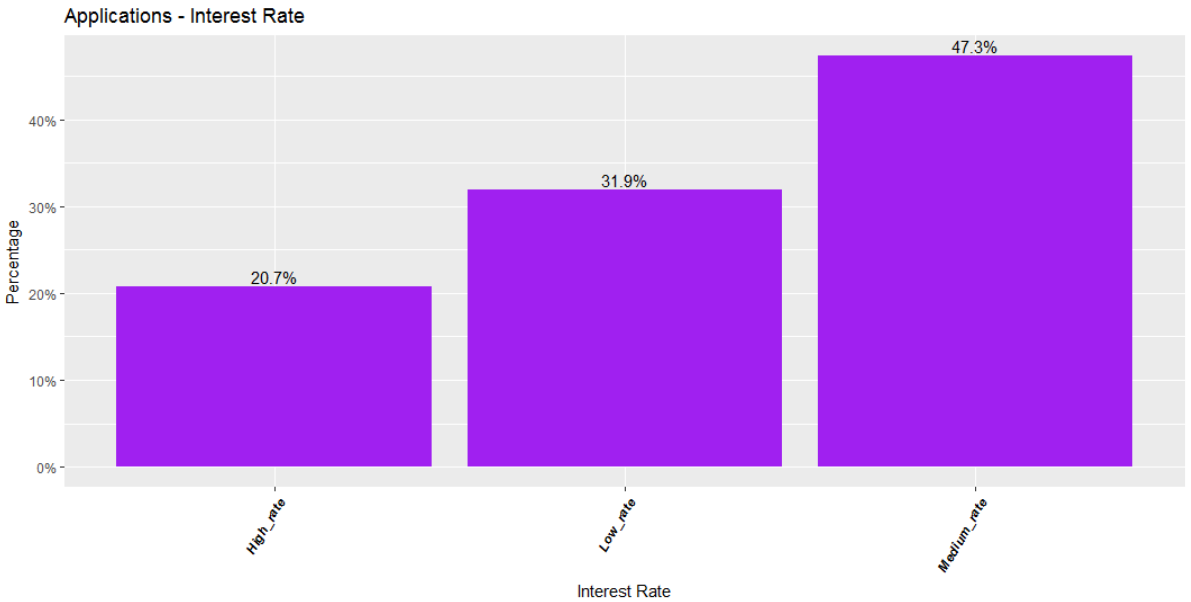
  The loan of approved loan applications are increasing significantly year wise – almost doubled for the recent year.



Analysis – Univariate

- **Loan Amount wise -**
  54% of the loan is taken given medium range (5K to 15K)

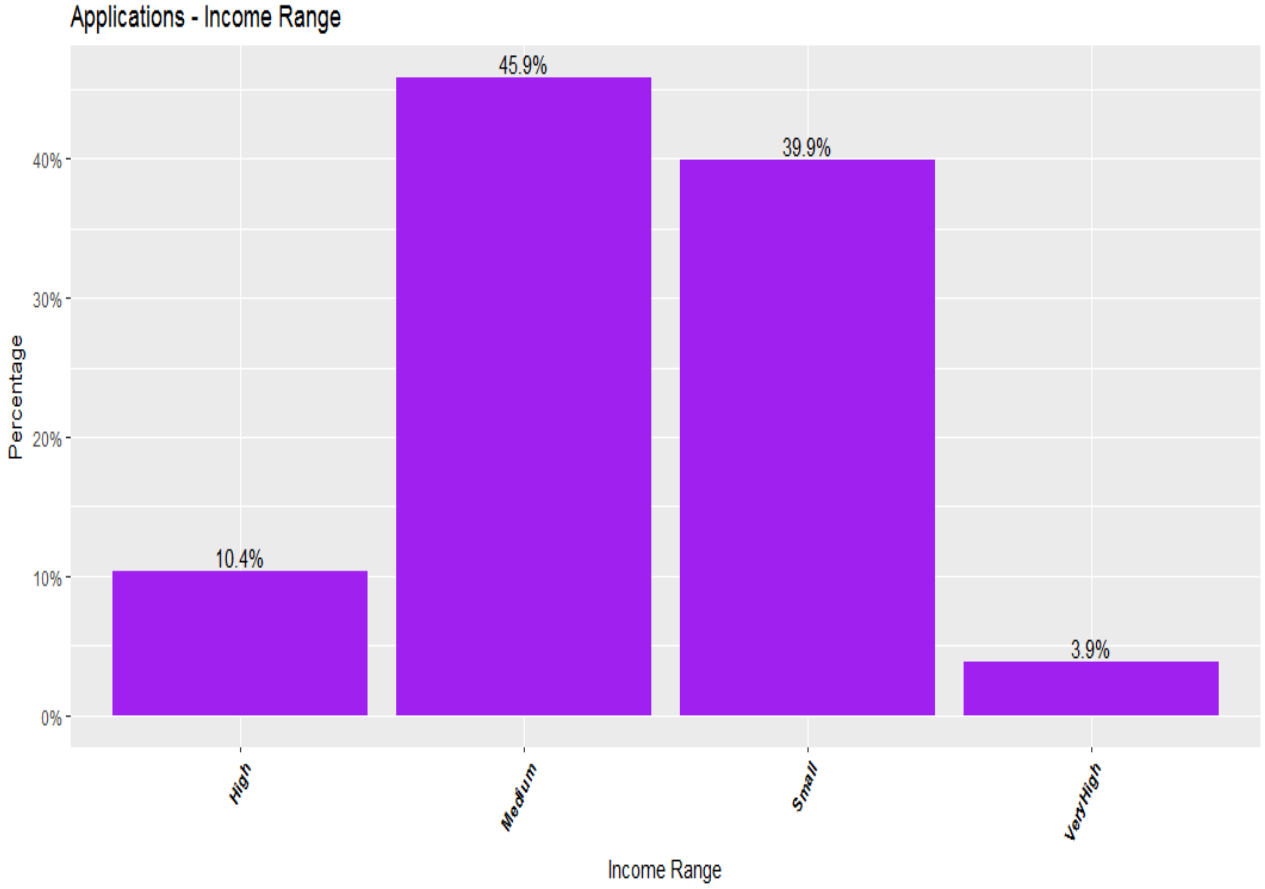




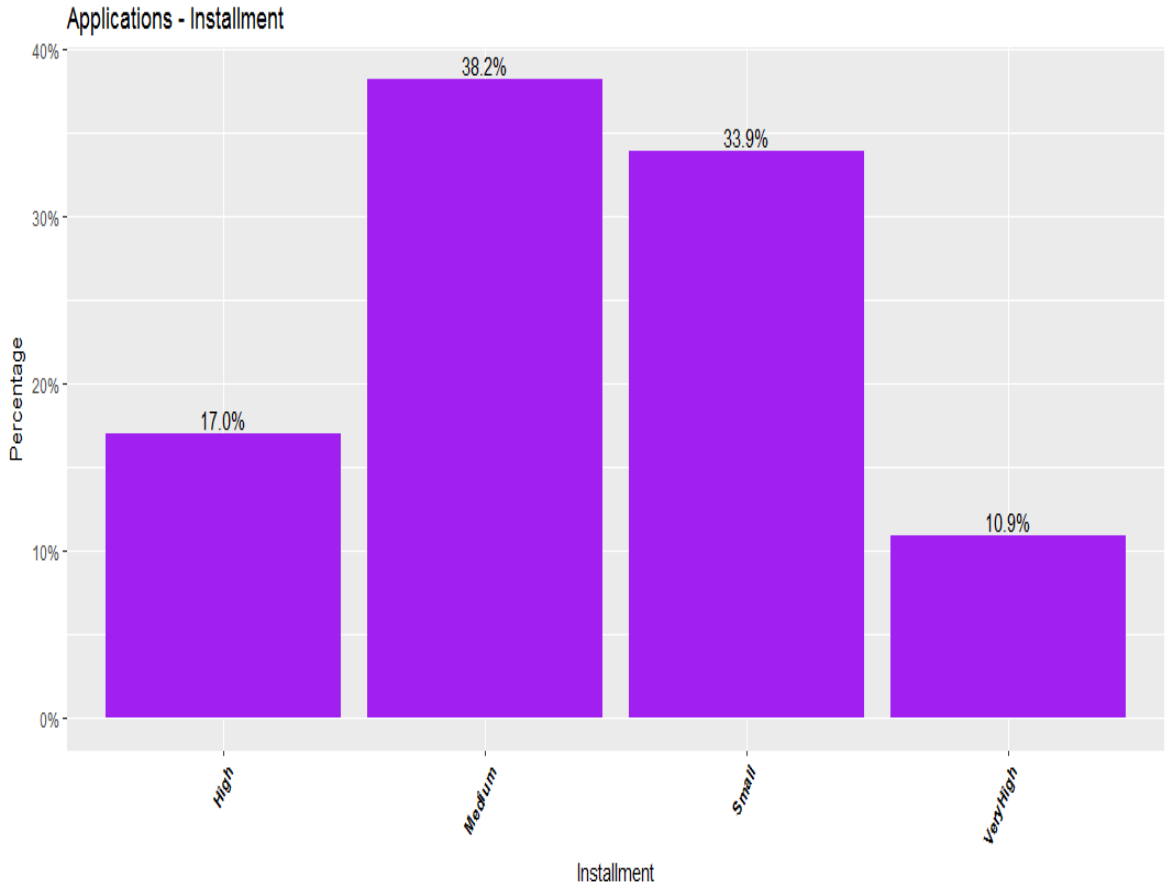
- **Interest Rate –**
  47% Applications carry medium interest rate (10 to 15%)

Analysis – Univariate - Plots

- **Income Range** –
  Nearly 90% of customer fall in Small and Medium income range



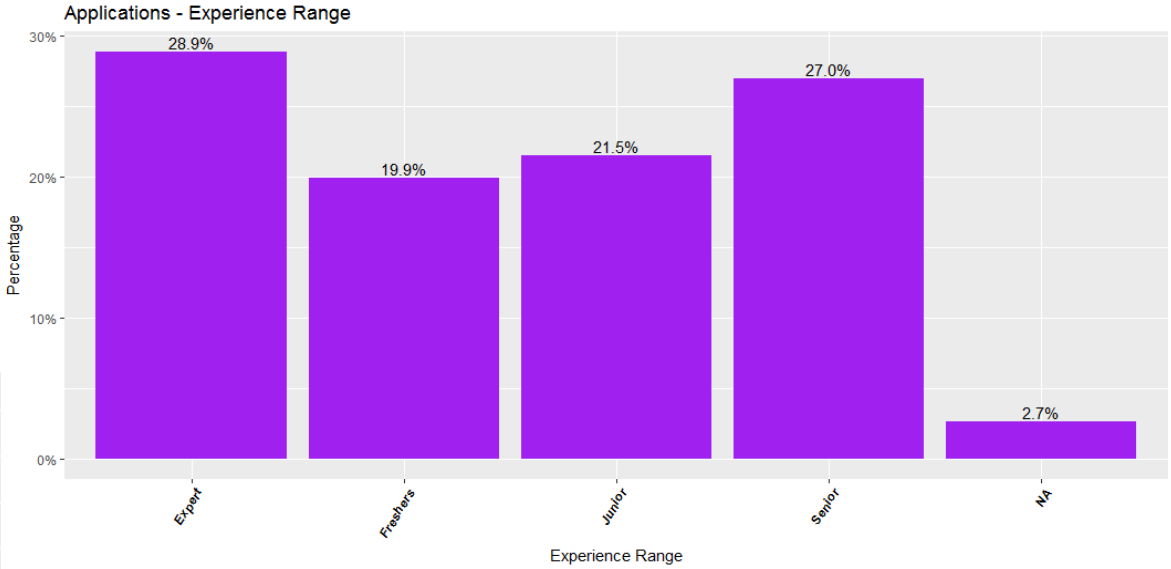
Installment (EMI) –
Nearly 70+% EMI falls into small and medium EMI

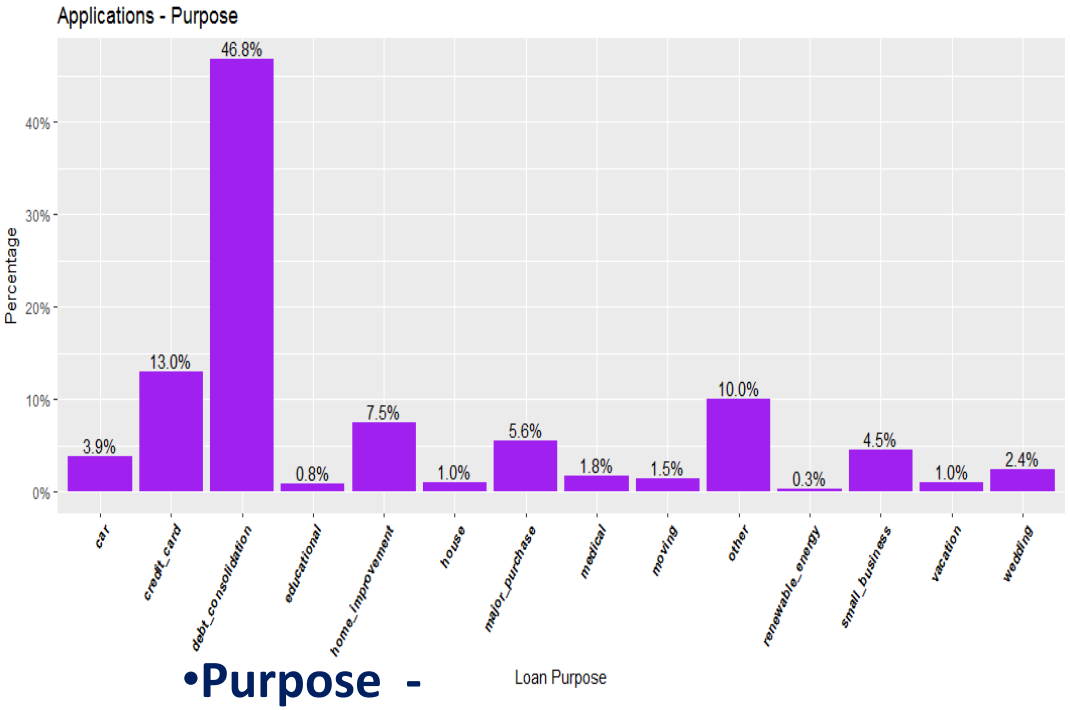


Analysis – Univariate - Plots

- **Experience Range –**

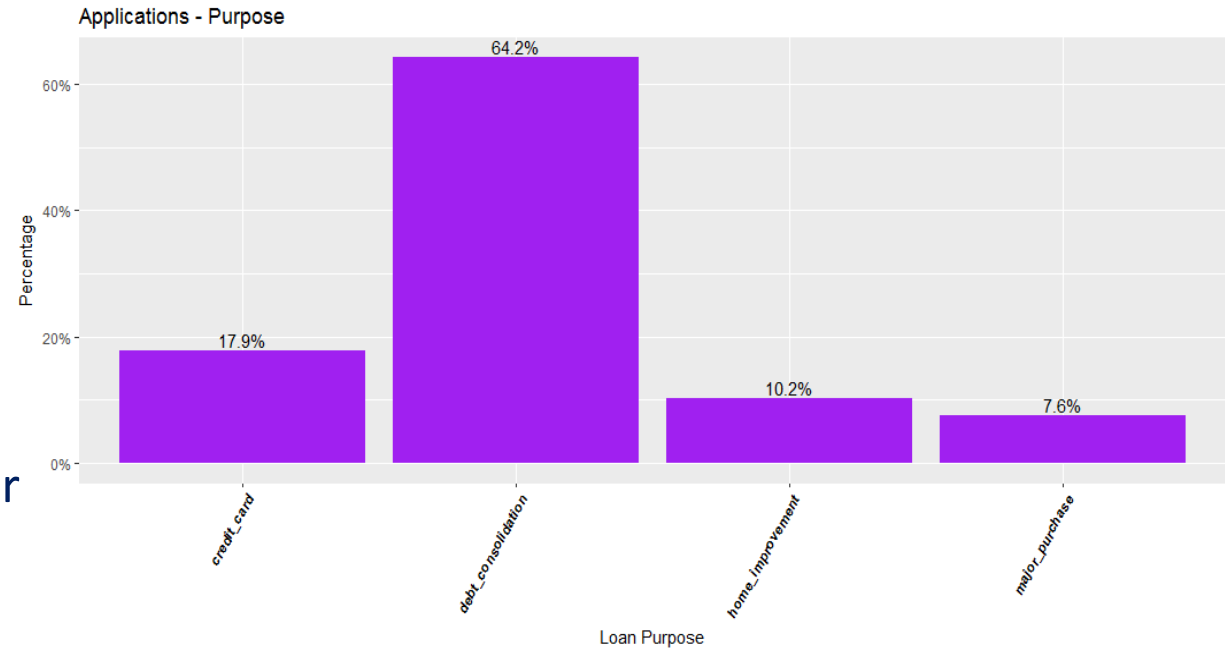
  We see that number of applications equally distributed across experience range.





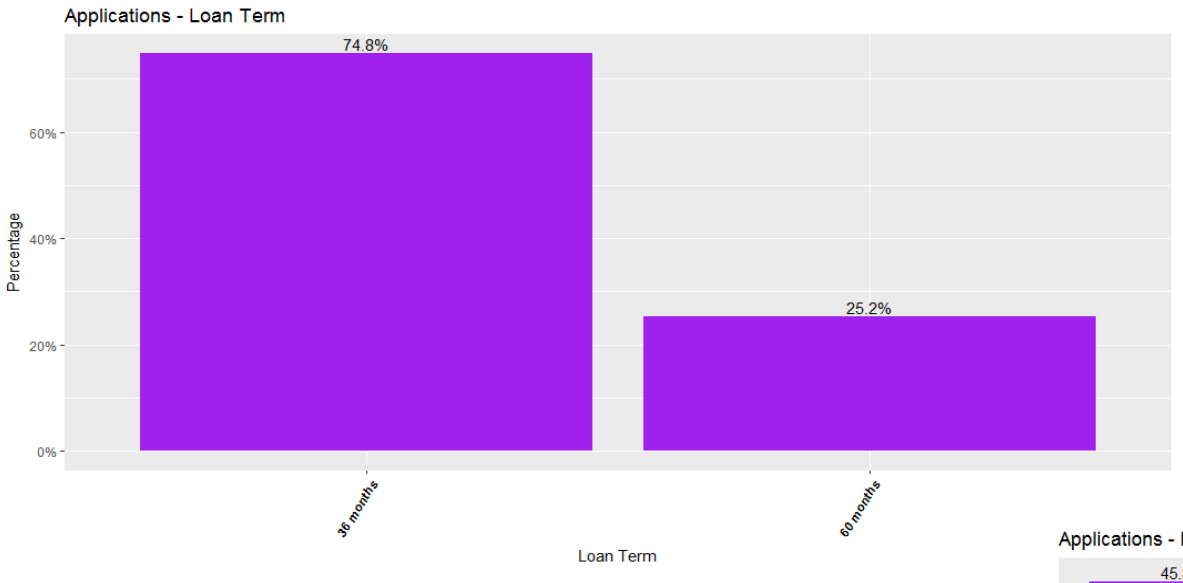
- **Purpose -**

  We see that mainly 4 categories have major contributions – Debt Cons, Credit Card, Home Improvement and Purchase

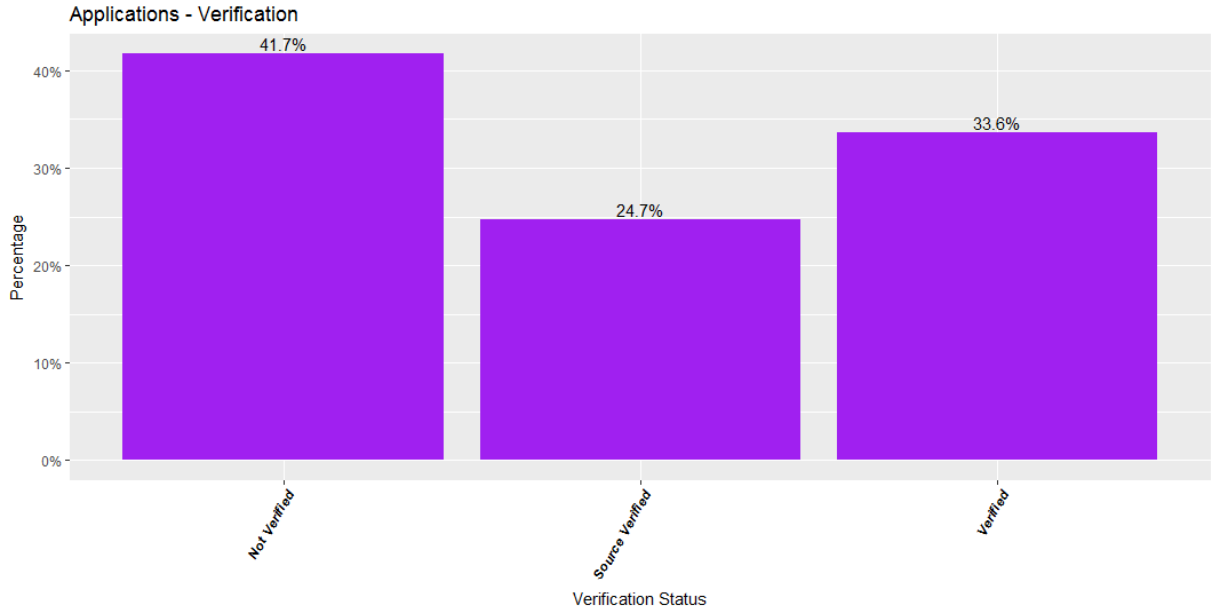


Analysis – Univariate - Plots

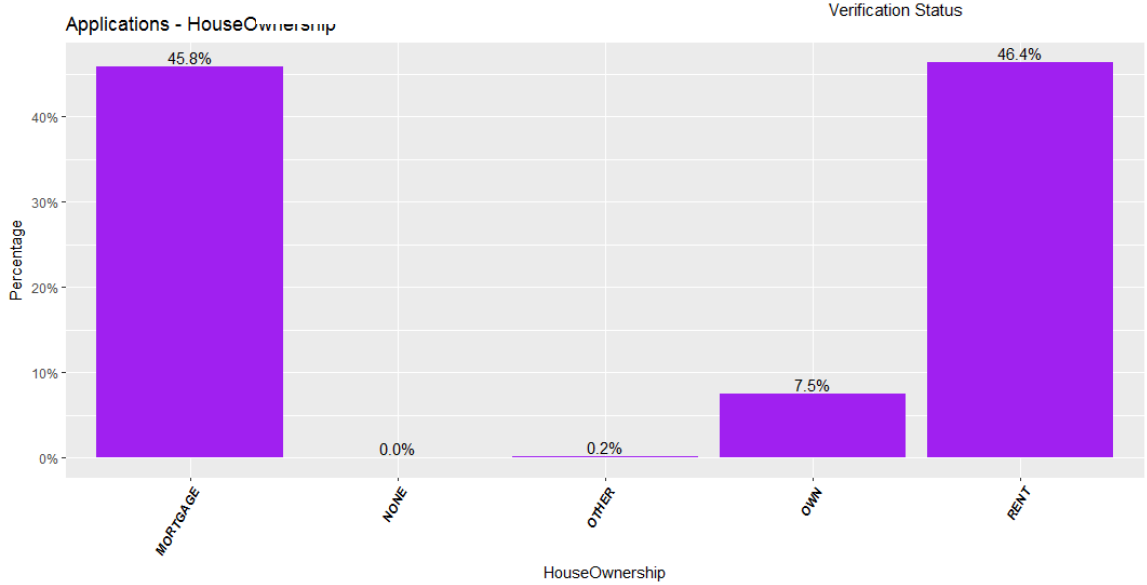
- **Term -** Around 75% loan applications are for short term (36months)



- **Verification Status –** Around 41% were not verified



- **House Ownership –**
Around 90% applications are of mortgage and rent



Analysis – Bivariate

Bivariate Analysis

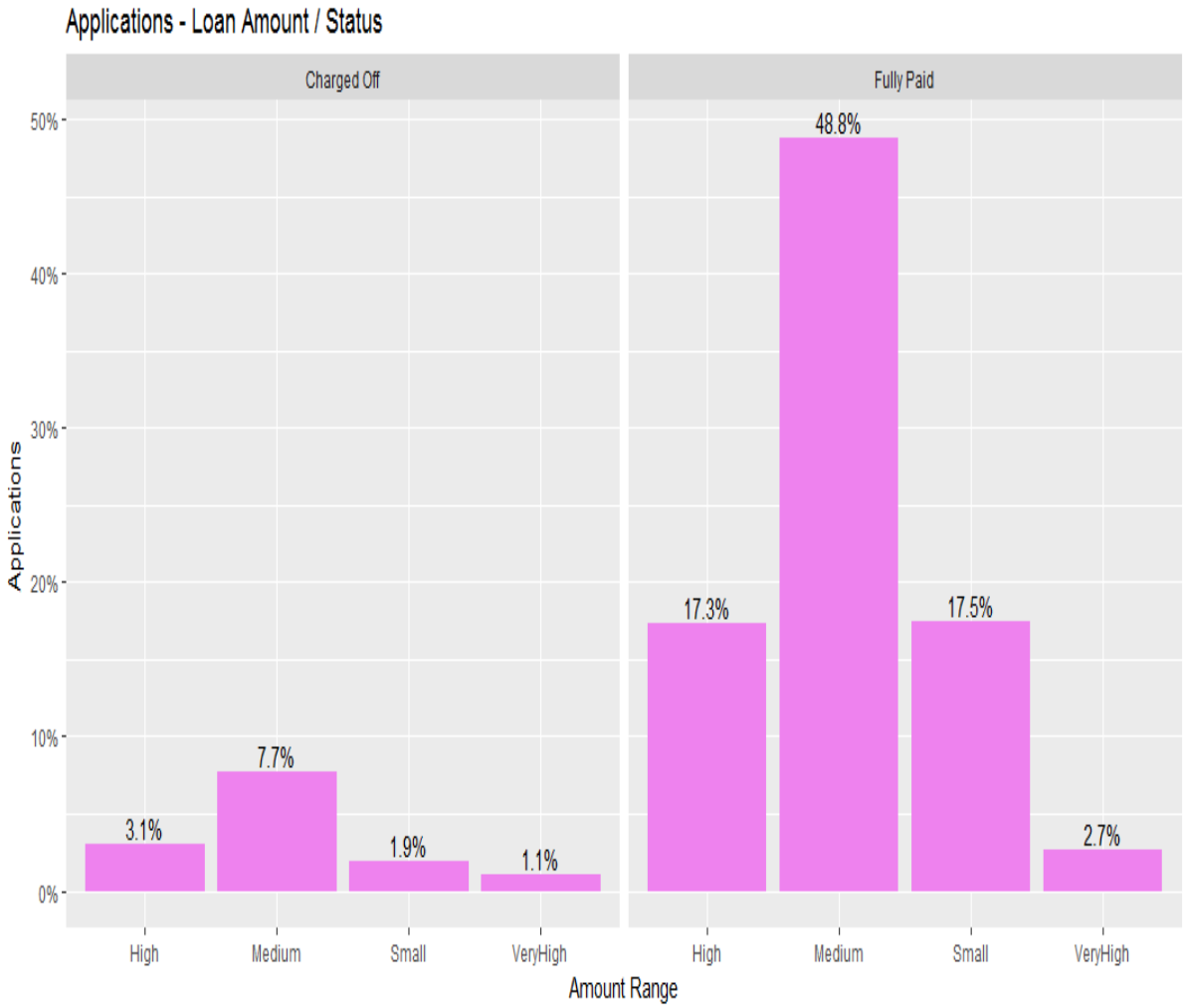
Analyzing more than one variable at a time. This gives relationships and helps to understand the impact of one variable on others.

Following slides visualize the bivariate analysis for the following variables against total number of loan Applications processed, the other main base variable kept constant is status (that provides information about fully paid or charged off).

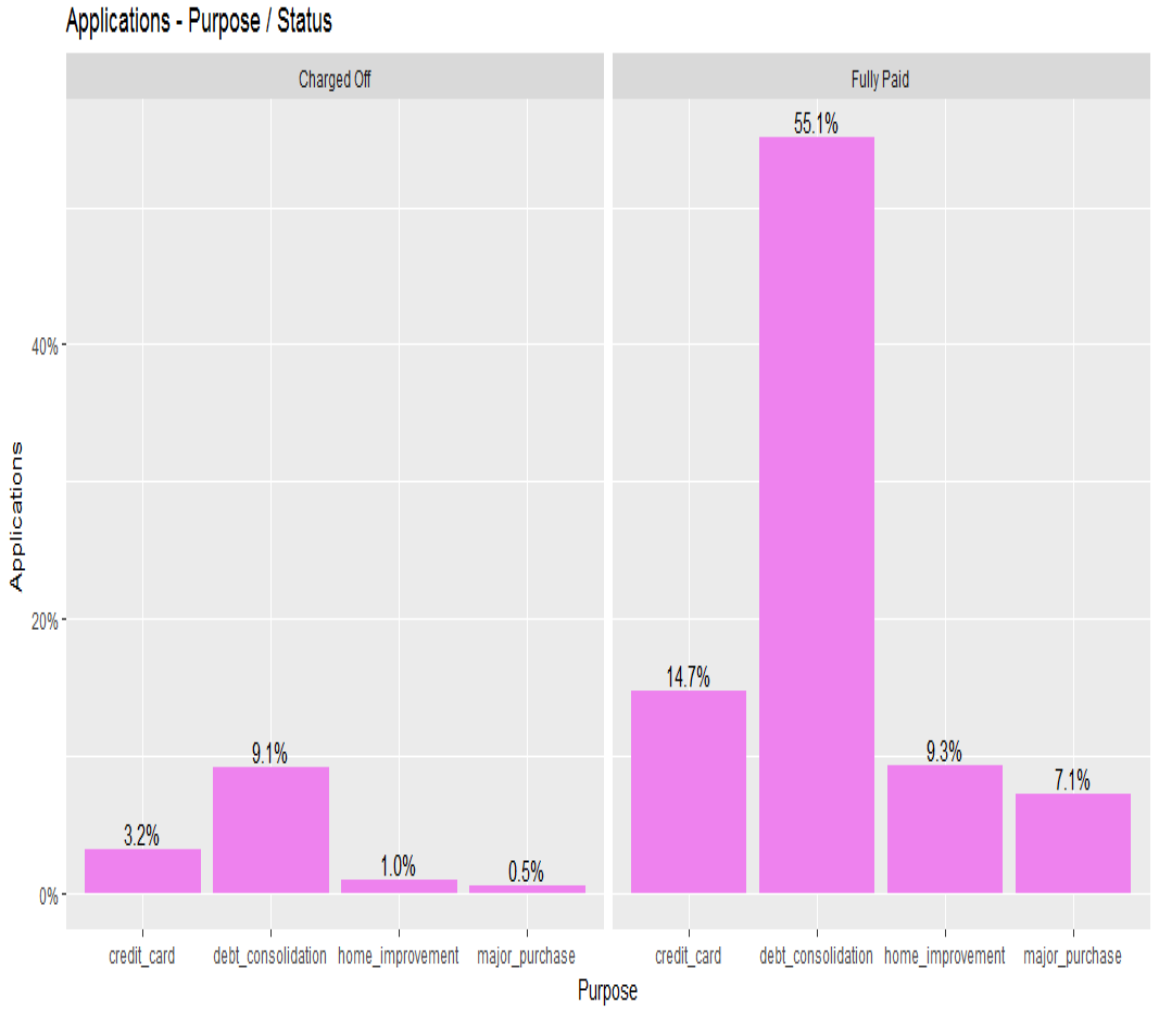
One of the objective of this analysis is understand the factors that are resulting in charged off:

- **Loan Amount / Status –** The highest category for charged off is medium income salary customers.
- **Purpose / Status –** The debt consolidation loan seekers are major defaulters then credit card people.
- **Interest rate / Status –** This is does not seem to have a major impact or differentiator.
- **Term / Status –** Higher default rate exists for short term (36 months) (nearly 3 times of 60months)
- **Experience / Status –** Higher experience people are major defaulters.
- **Verification / Status –** Does not seem to have any impact – equal % of defaulters.
- **Homeownership / Status** – Rent and Mortage are the major defaulters.

Analysis – Bivariate - Plots



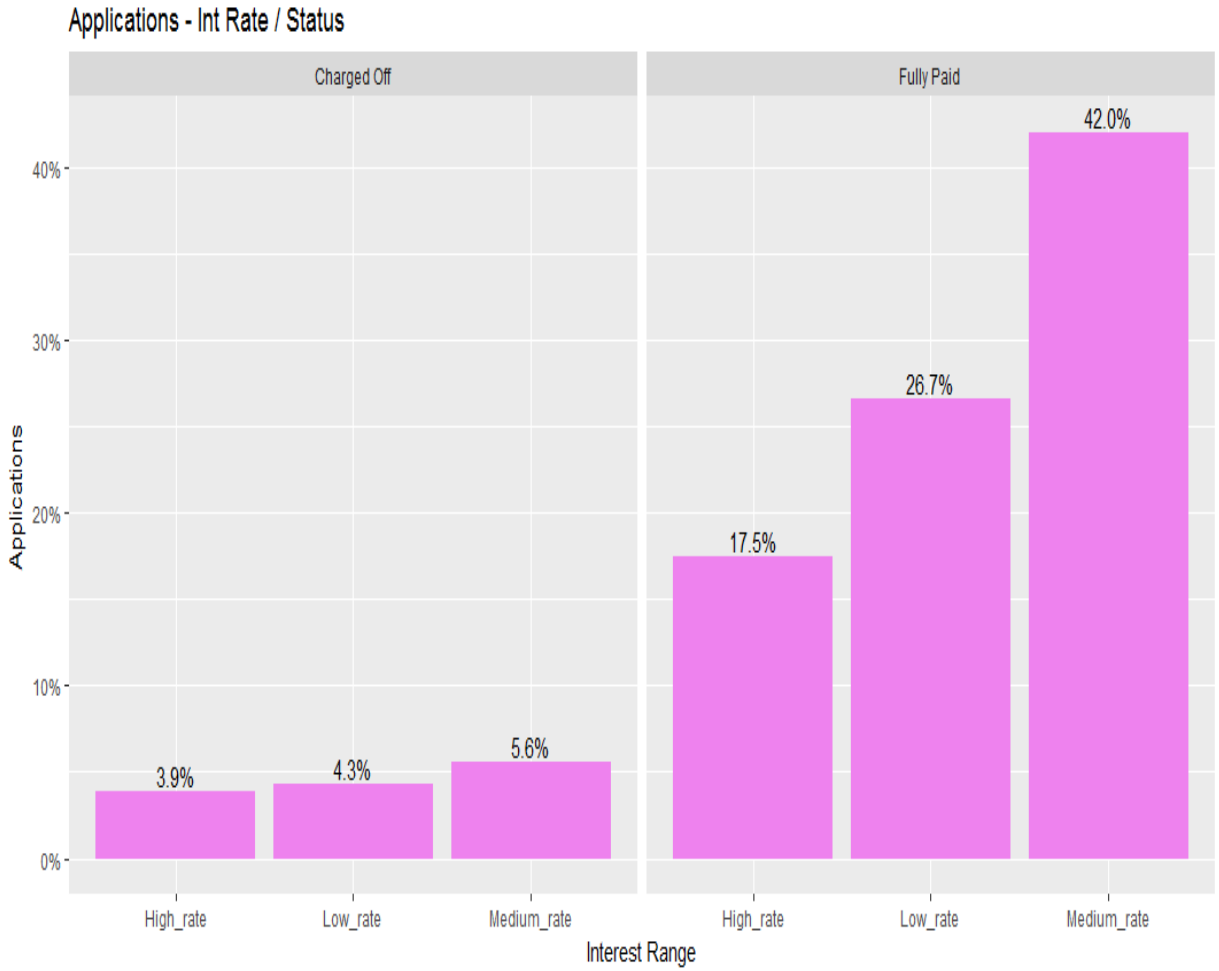
- **Loan Amount / Status –** The highest category for charged off is medium income salary customers.



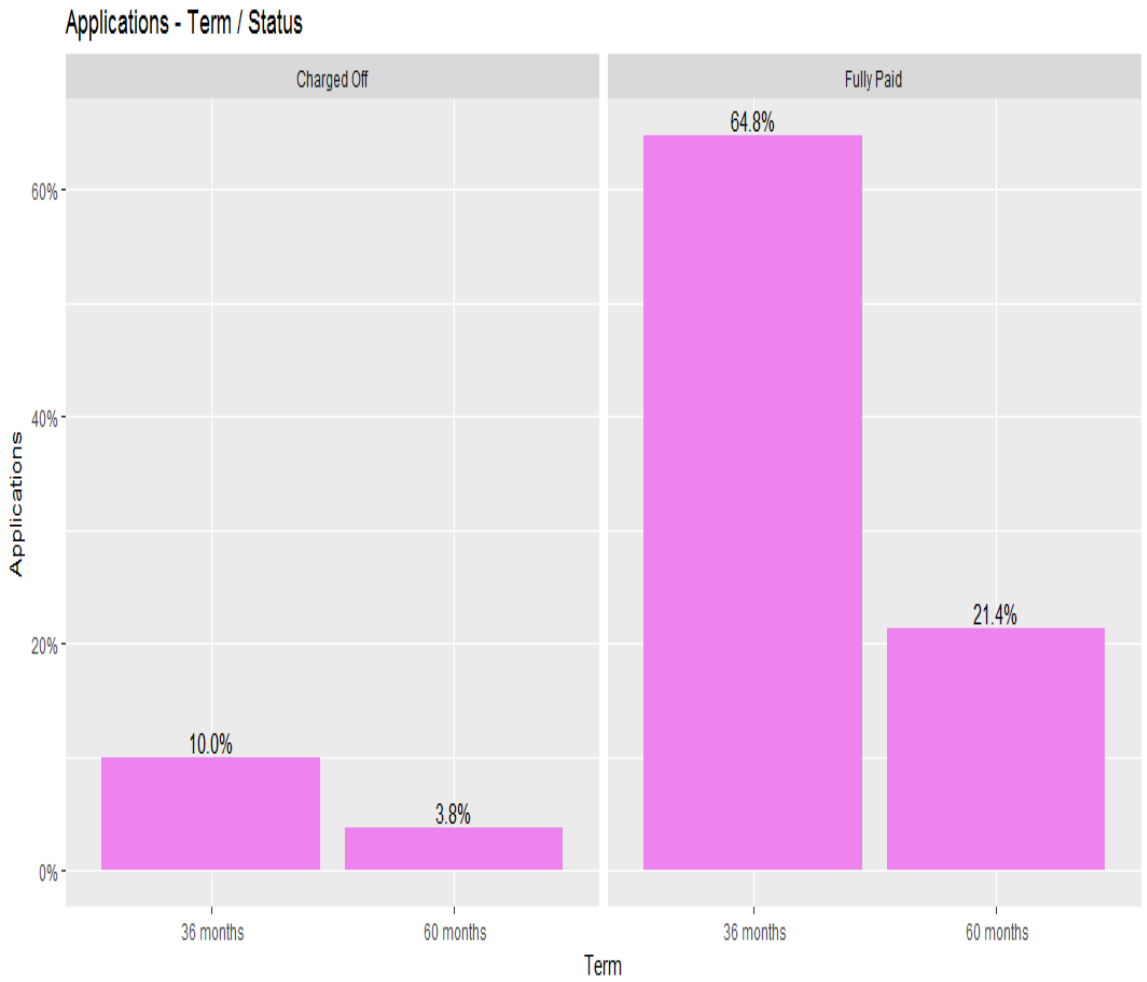
- **Purpose / Status –** The debt consolidation loan seekers are major defaulters then credit card people

Analysis – Bivariate - Plots

- **Interest rate / Status –** This is does not seem to have a major impact or differentiator.

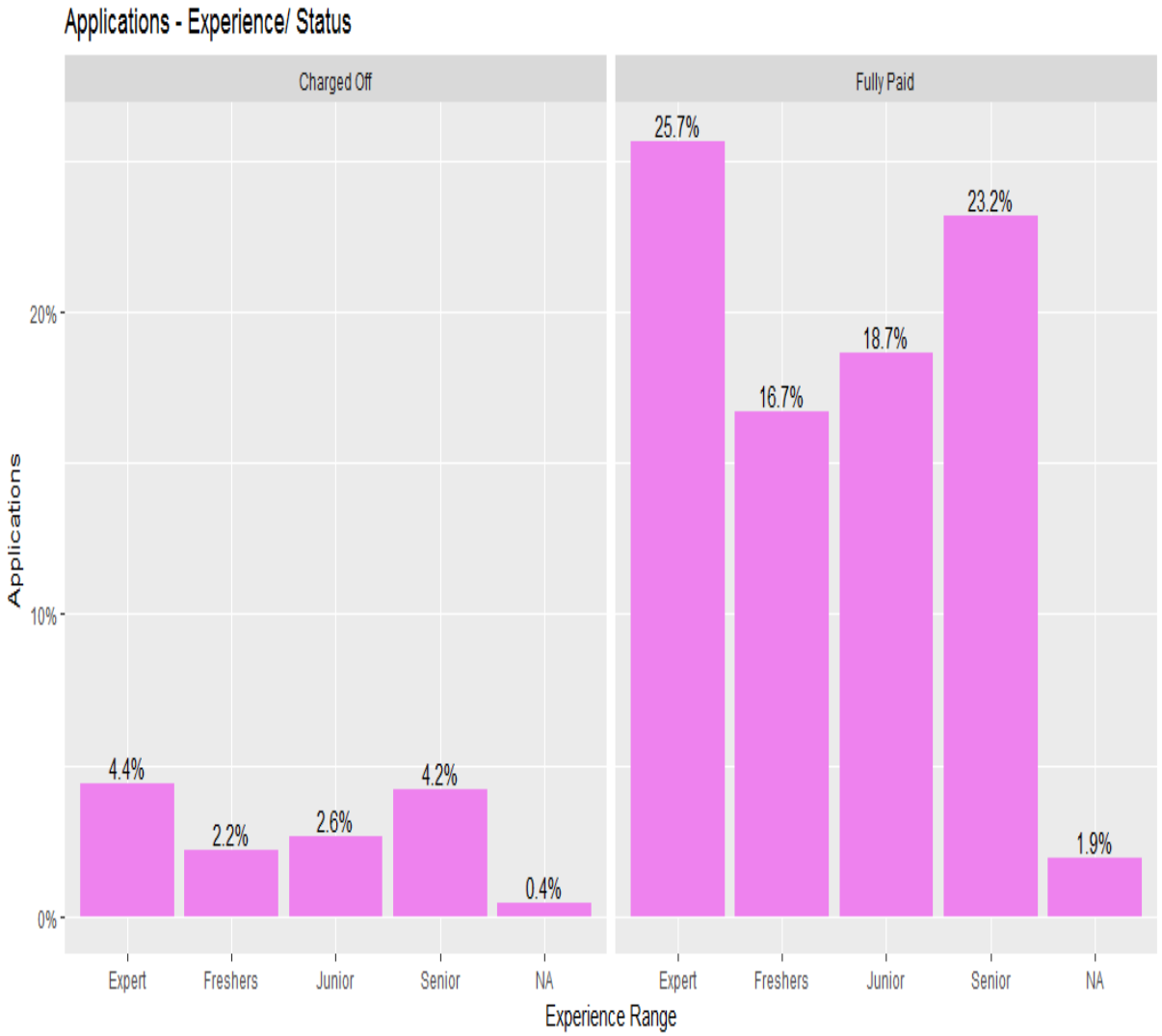


- **Term / Status –** Higher default rate exists for short term (36 months) (nearly 3 times of 60months)

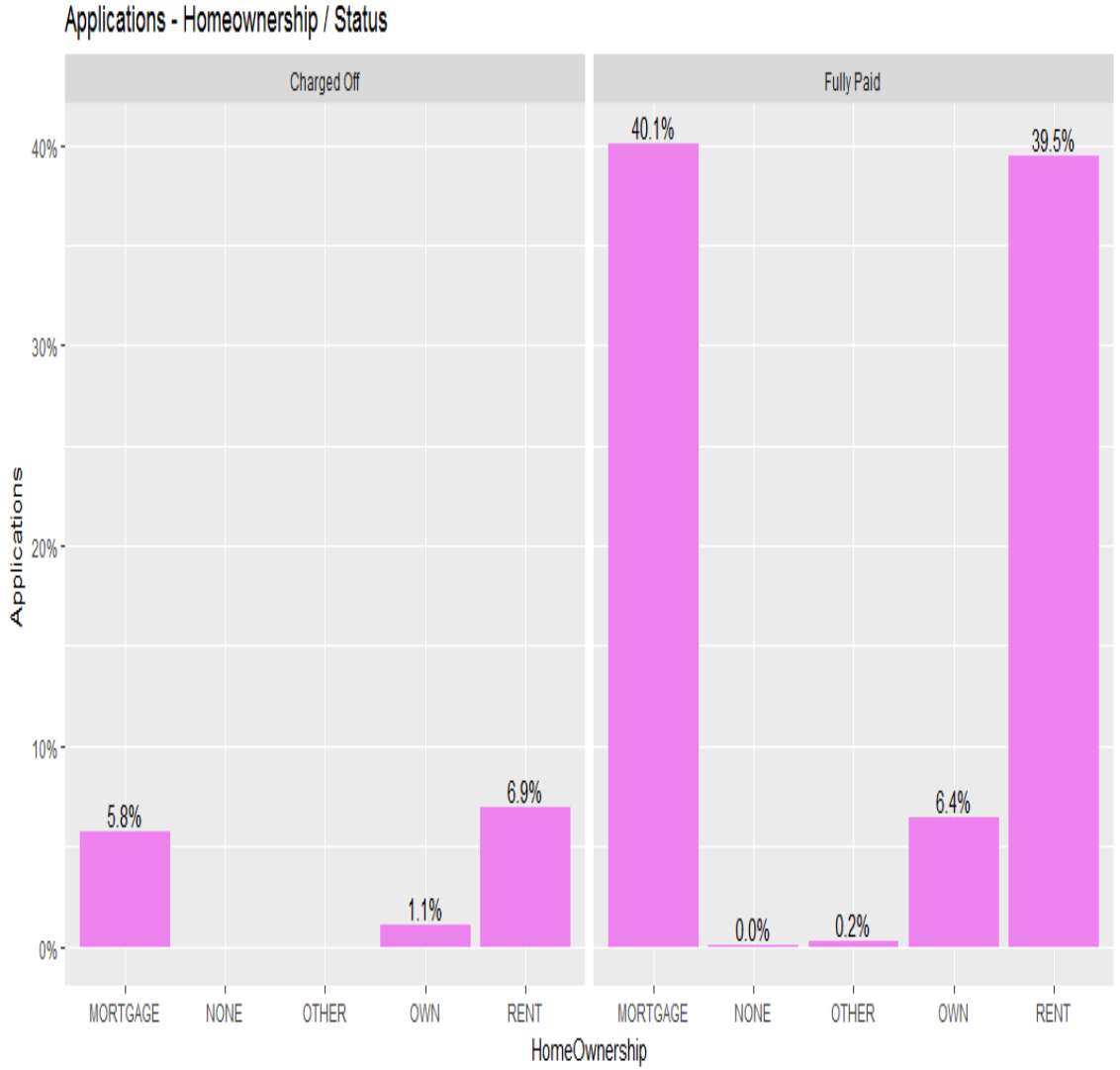


Analysis – Bivariate - Plots

- **Experience / Status –** Higher experience people are major defaulters.

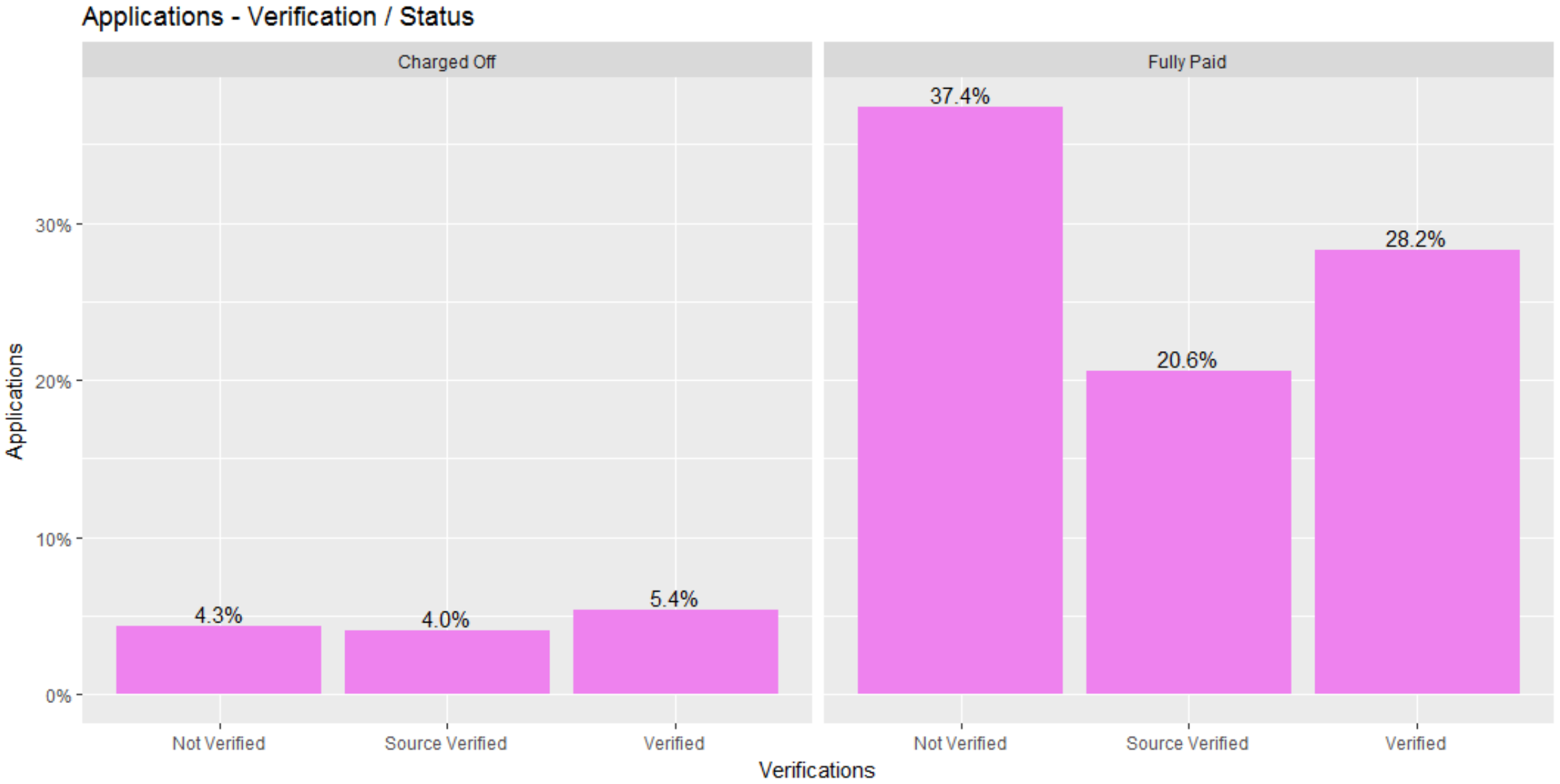


- **Homeownership / Status** – Rent and Mortgage are the major defaulters.



Analysis – Bivariate - Plots

- **Verification / Status –** Does not seem to have any impact – equal % of defaulters.



Conclusions

Based on the above analysis, we could identify following top 5 driving factors (variables) needs to be understood to avoid risky applicants by the lending company.

1. Medium Income Range
2. Debt Consolidation Purpose
3. Short Term Loans
4. Mortgage ownership type
5. Higher Experience