Gramener Case Study

Group
- Rajarshi Ghoshal
- Piyush Baid

PGDDS C7
June'18
Group 2

Business Objectives

The Company for which we are working, is the **largest online loan marketplace**, facilitating **personal loans**, **business loans**, and **financing of medical procedures**. Borrowers can easily access lower interest rate loans through a fast online interface.

- ❑ Identify patterns which indicate if a person is likely to default, i.e., identifying the **High Risk Loan Applicants.**
- ❑ Use **EDA** to understand how consumer attributes and loan attributes influence the tendency of default.
- ❑ Analyze the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default.
- ❑ Reducing the amount of Credit Loss (amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed).

Metadata & Data Understanding

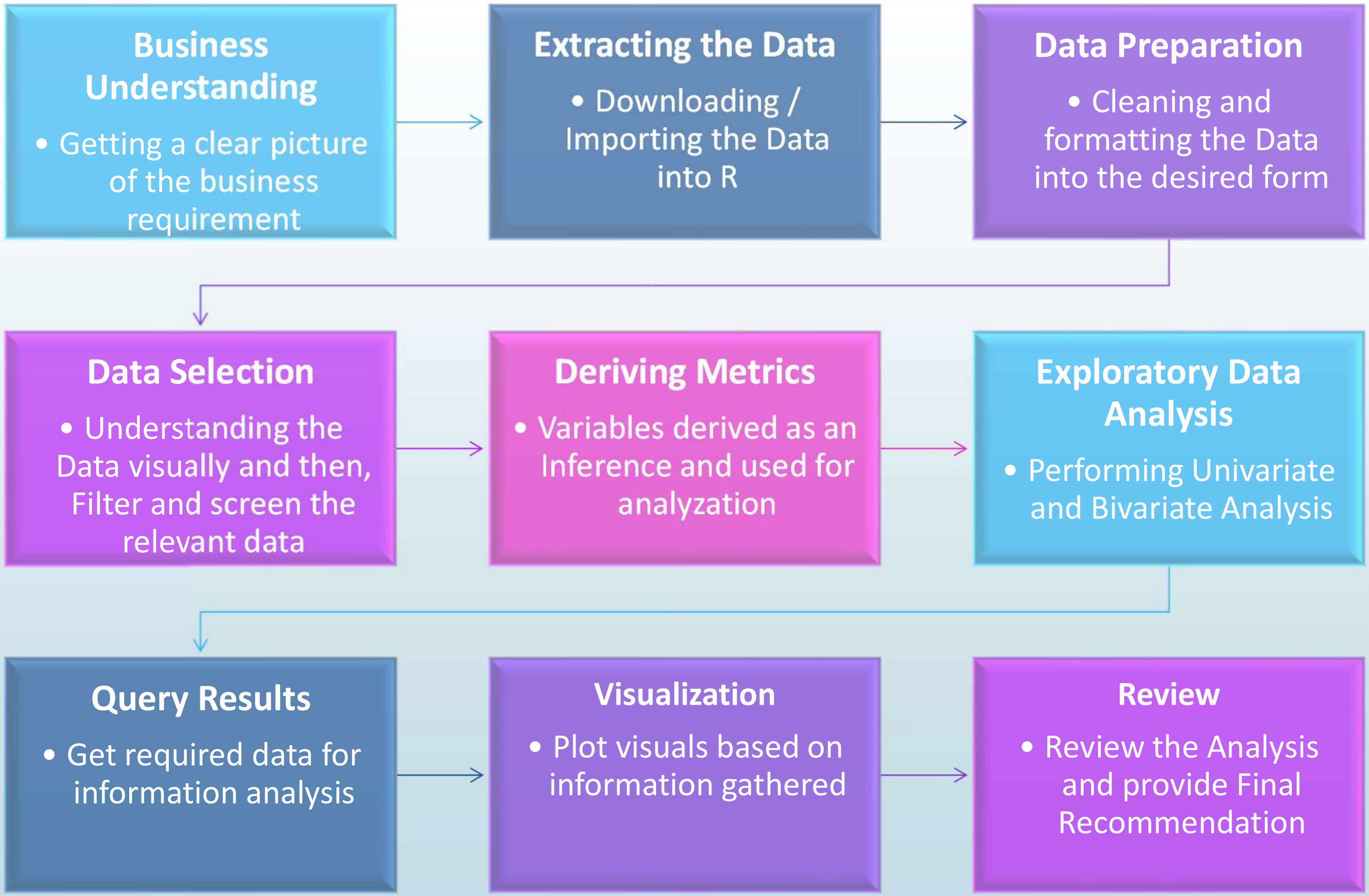
There are 111 attributes (variables) for every individual who requested for a loan:

- Data for 4 years have been listed under the 3 major categories of Loan types: **“Fully Paid”, “Charged Off” and “Current”.**
- Customer’s information, like their **“Annual Income”, “Purpose of the loan”, “Employment Length”,** etc have been provided as well.
- Loan Details, such as, **“Loan Tenure”, “Interest Rate”, “Loan Amount”, “Loan Issued Date”,** etc were given.
- Total records accounted for **39717** for the data.
- No Duplicate data was observed with the data set provided.

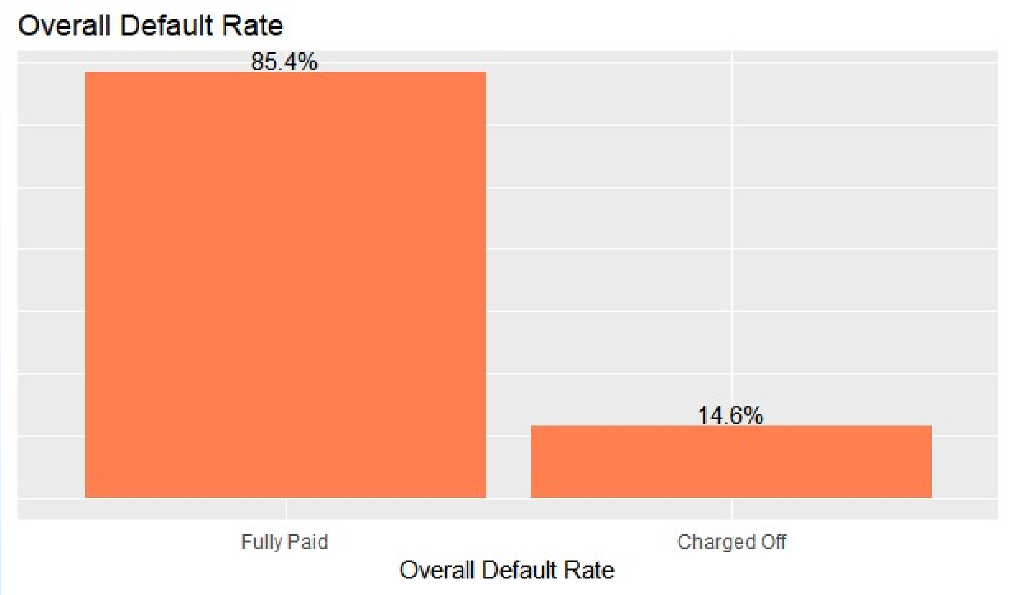
Assumptions

- Most of the analysis were performed excluding the data for loans with **“Current”** status.
- Plots for the Annual Income were **capped to around 100000** get a better visual.
- Columns with **constant/null values** were removed as the wouldn’t have been useful for data analysis.
- Most of the columns, such as- **“emp_title", "collection_12_mths", "chargeoff_within_12_mths", "tax_liens”, "addr_state", "zip_code", and "title“** were removed because it didn’t had any impact on the analysis.
- In the emp_length column of the data, the value- **“<1”** were imputed to **“0”.**
- Few numerical columns having some NA values in between had the imputation of **mean/median value for the NA.**

Problem Solving Methodology

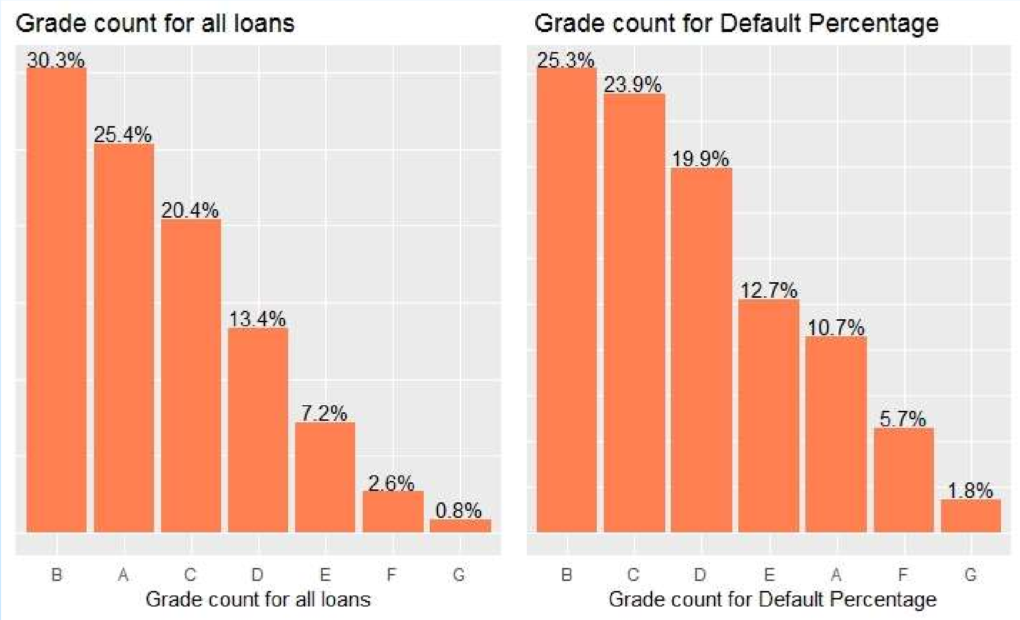


Plot 1 : Peeking at Default Rate and Grades

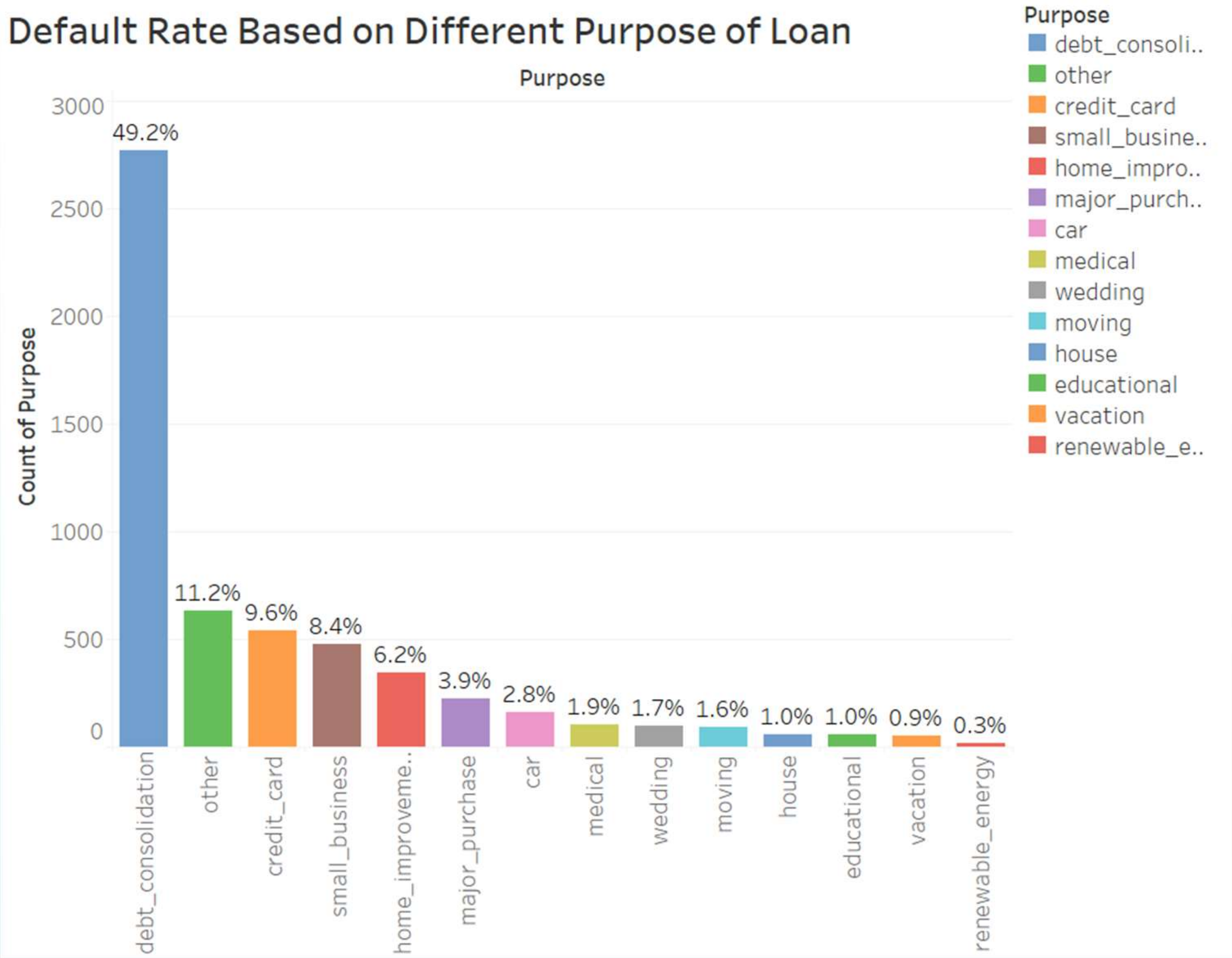


- Around **85.4%** loans are **fully paid**.
- Around **14.6%** loans get **defaulted**

- **Grades B, C, D and E** are more risky grades as compared to others.
- **Grade A, F and G** seems to be risk free.



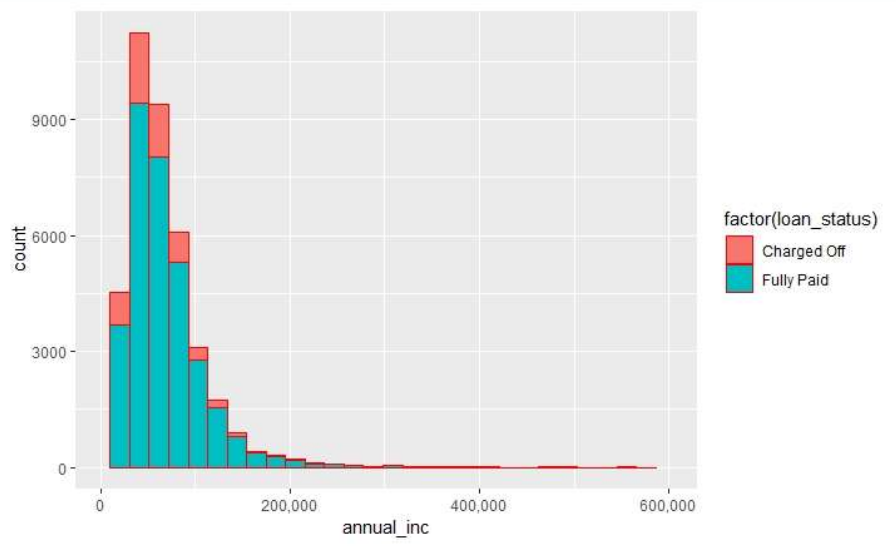
Plot 2 : Loan Purposes vs Charged Off data

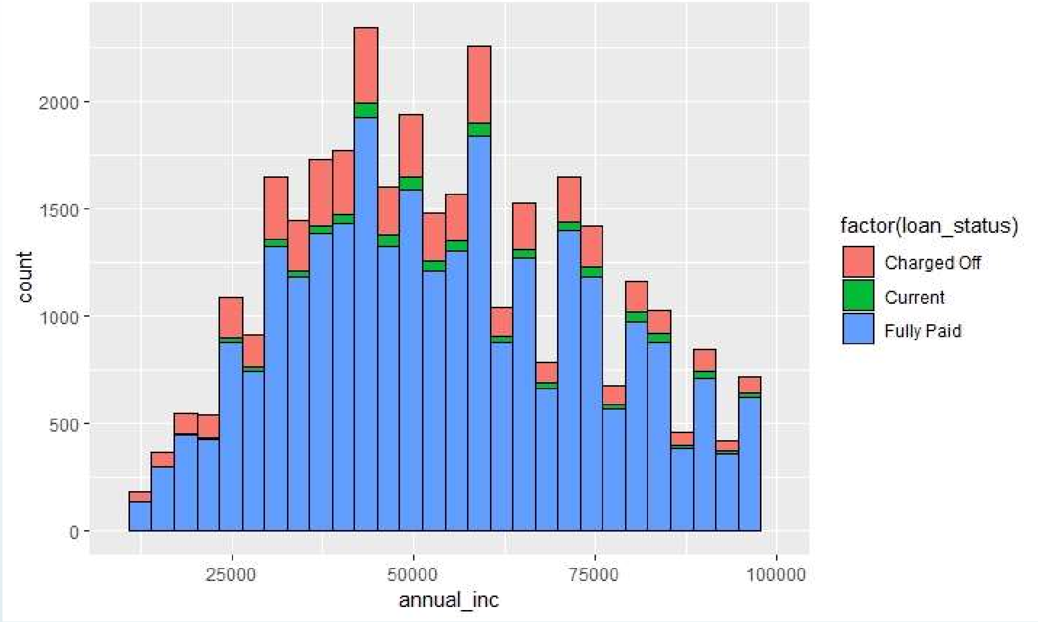


The Top 5 Loan purposes for the default% are:

I. **Debt-consolidation**
II. **Others**
III. **Credit Card**
IV. **Small Business**
V. **Home Improvement**

Plot 4 : Annual Income vs Loan Status

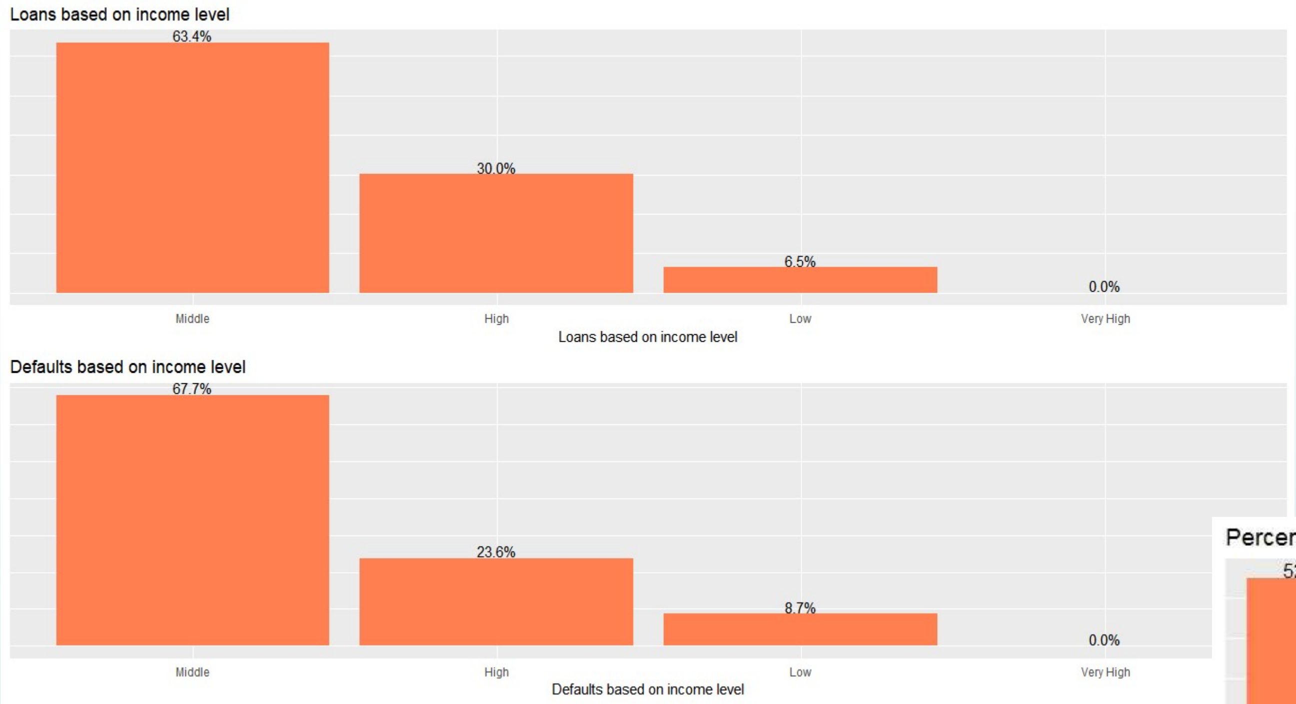




- As we don't see much Charged off data after 100000, we will cap the annual income plots below that value to get a clear visual.

- After capping the Annual Income, as mentioned before, we see that most of the loans are given to the people who have the annual income between **40k-70k approx**. Moreover, most of the defaulters lie in this same range as well (Charge-off data).

Plot 5 : Income groups & Employment Length groups Data

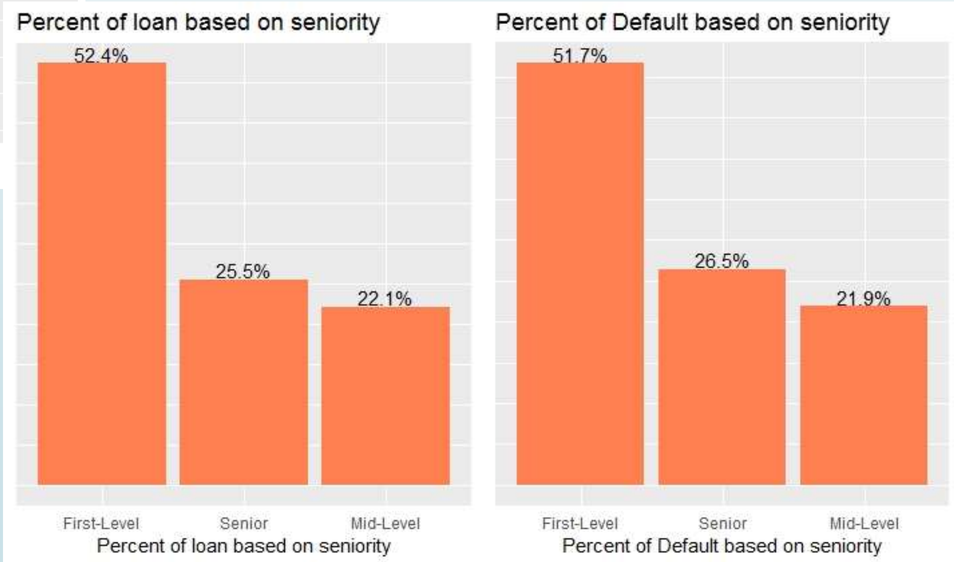


- Annual Income data was divided into 4 groups: **Low** (less than equal to 25k), **Middle** (less than equal to 75k), **High** (less than equal to 100000) and **Very High** (greater than 100000).

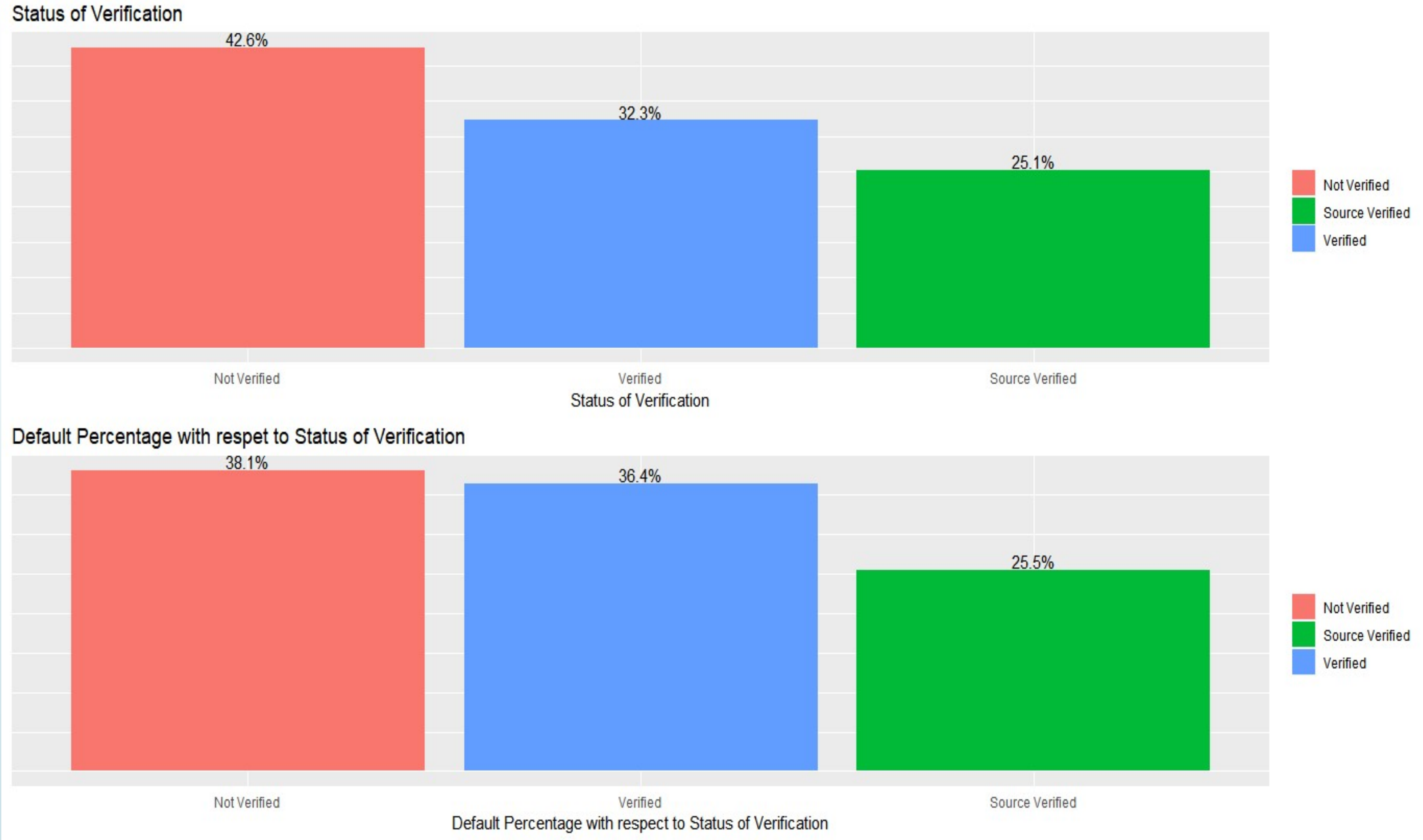
- **Middle group** is more riskier income group for defaults.

- Employee length data was divided into 3 groups: **First_Level** (less than equal to 4), **Mid- Level** (less than equal to 8) and **Senior** (greater than 8).

- Maximum loan was given to more junior people **(First_Level)** and maximum default are coming from them as well.

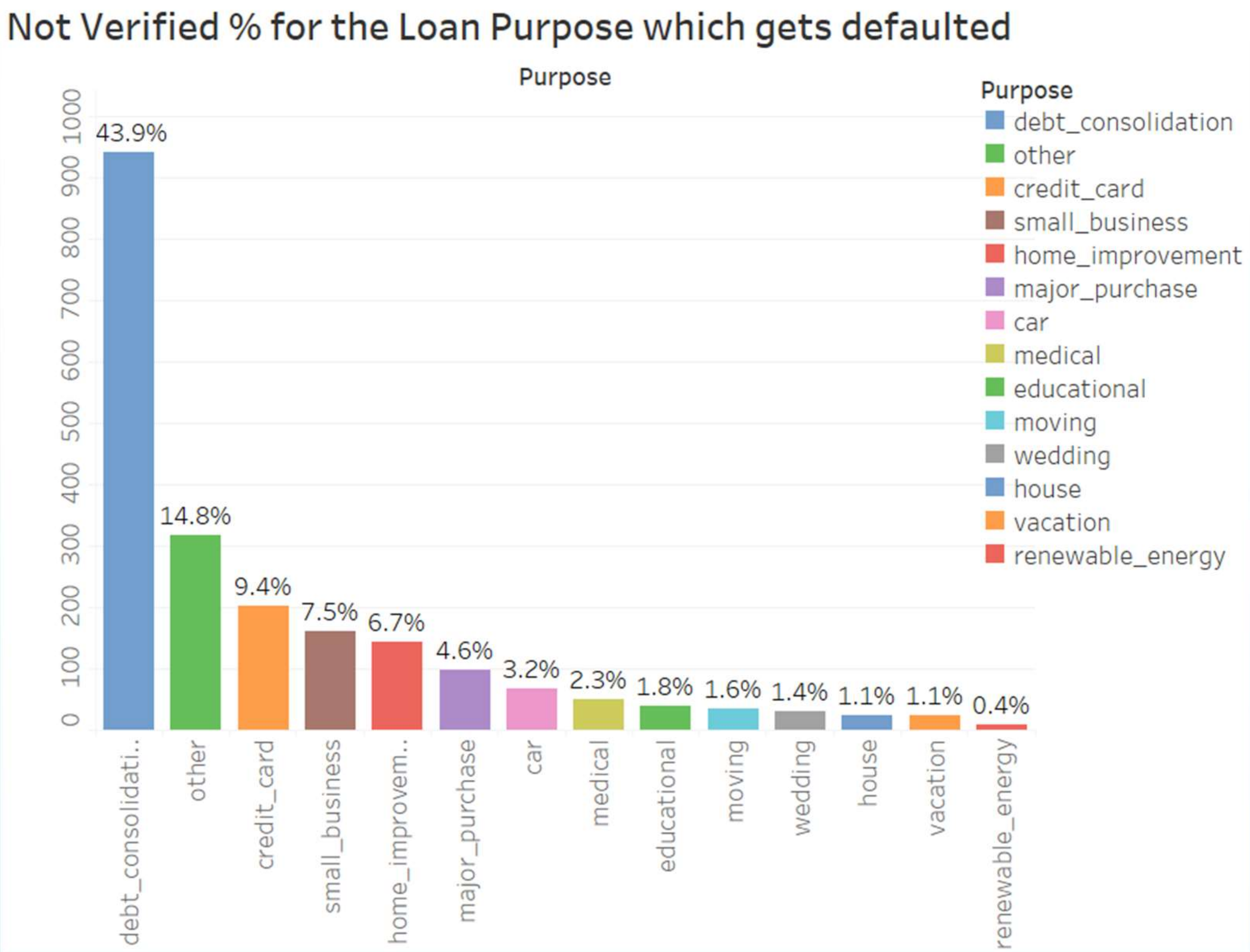


Plot 6 : Verification Status Analysis



- Majority of loans **(42.6%)** are **not verified** in any way.
- Majority of default loans **(38.1%)** are **not verified** as well.

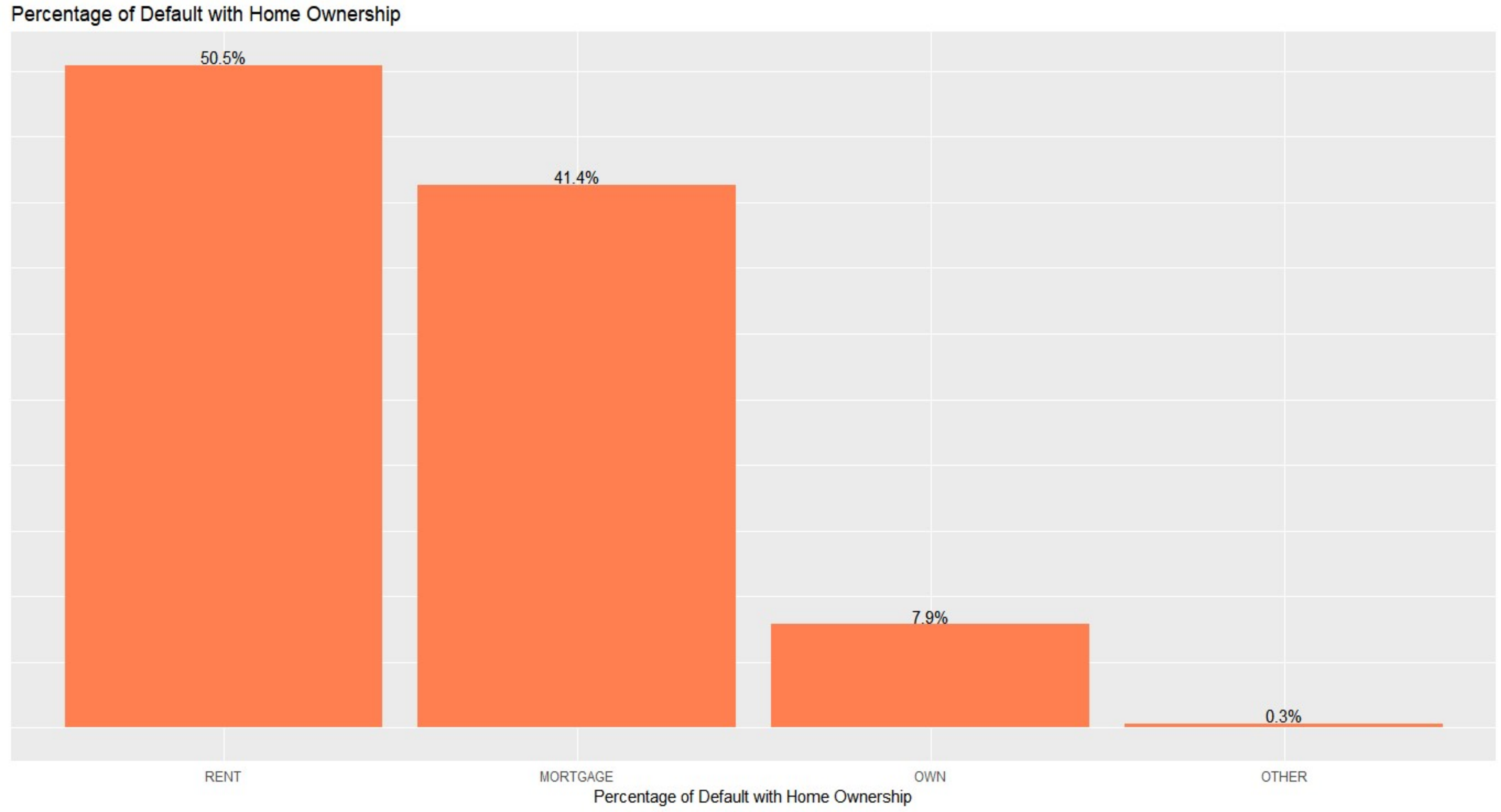
Plot 8 : Verification Status Analysis contd.



- Here we see that top 5 purposes where loan gets defaulted and are not verified are :-
  I. **Debt Consolidation**
  II. **Other**
  III. **Credit Card**
  IV. **Small Business**
  V. **Home Improvement**

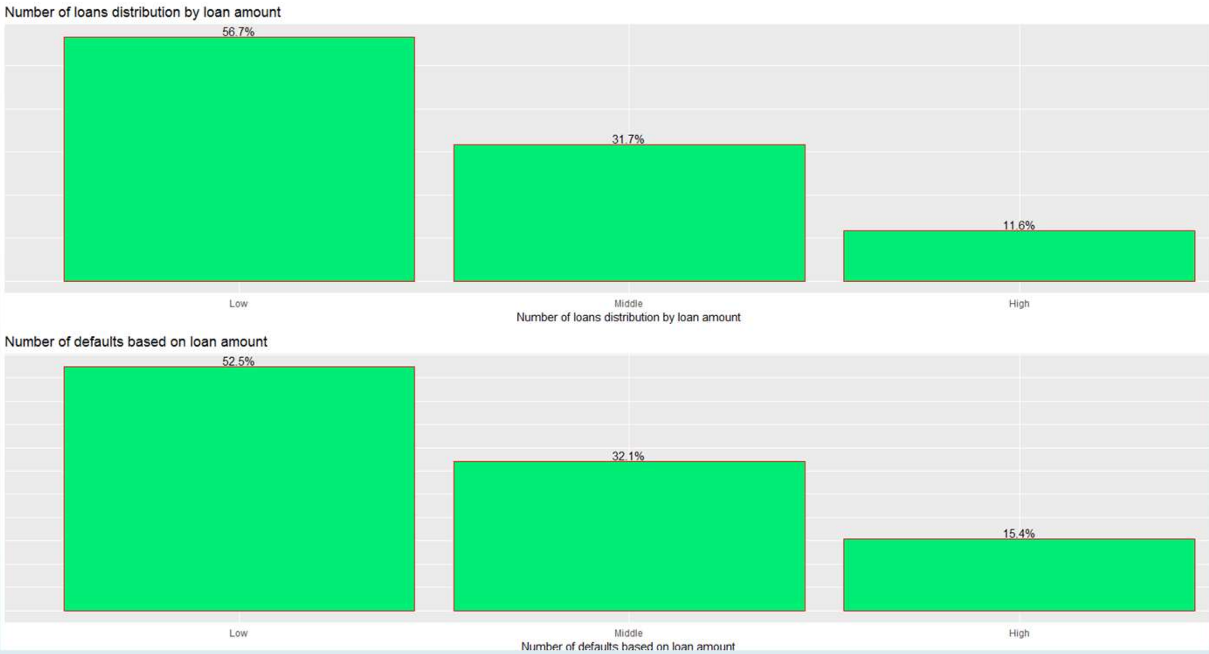
- So, these areas need more attention.

Plot 9 : Home Ownership Status



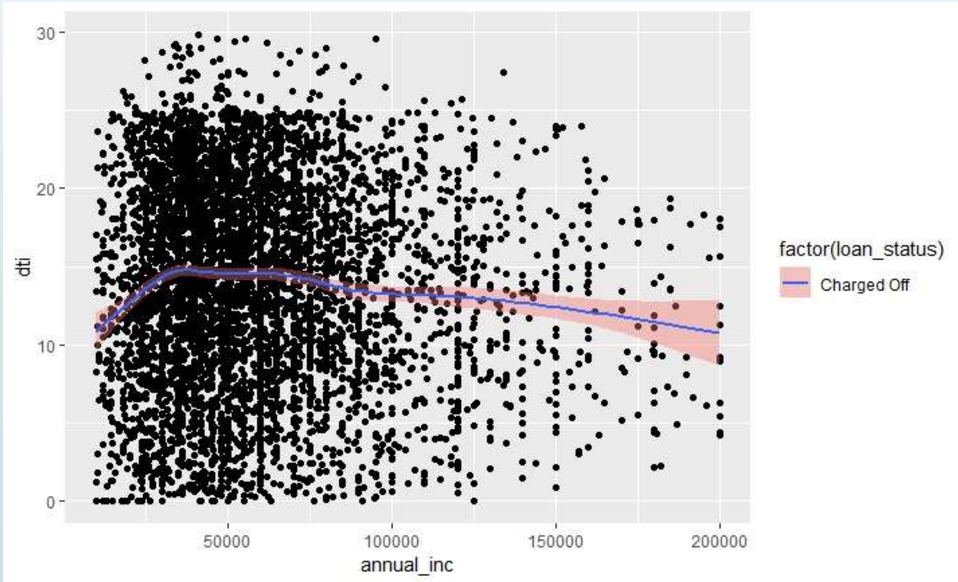
- We can clearly observe people who owns **unmortgaged home** are much **less likely to get defaulted**.
- So, people who **doesn't own home** or have **mortgaged** one are much **riskier**.

Plot 10 : Loan Amount and DTI Analysis



- **DTI means debts to income ratio**. As the income increases, DTI decreases (which is obvious).

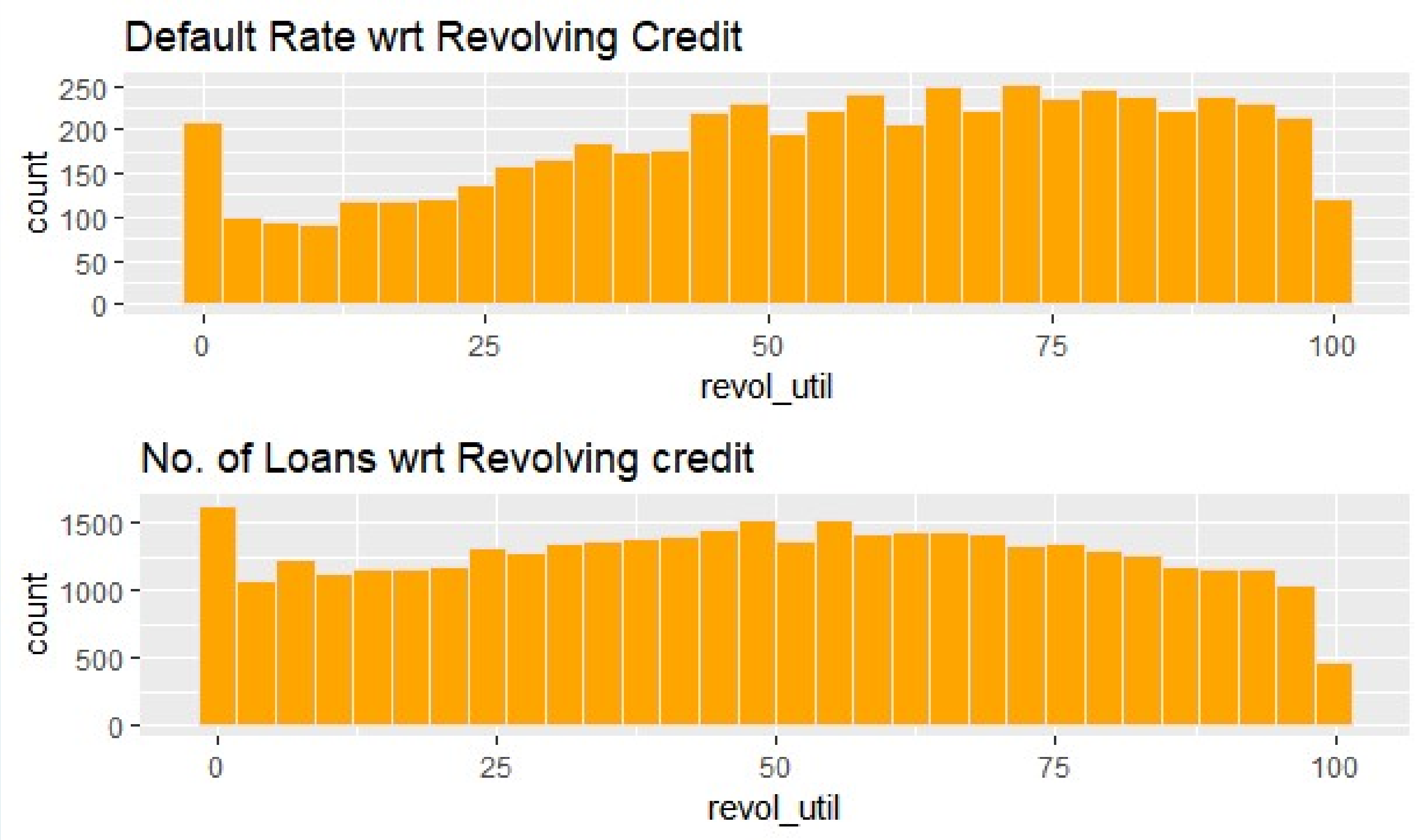
- But main thing to note is that **DTI is high** for the **Annual income: 40k-70k** approx.



- Loan Amount data was divided into 3 groups: **Low** (less than equal to 10k), **Middle** (less than equal to 20k), **High** (greater than 20k).

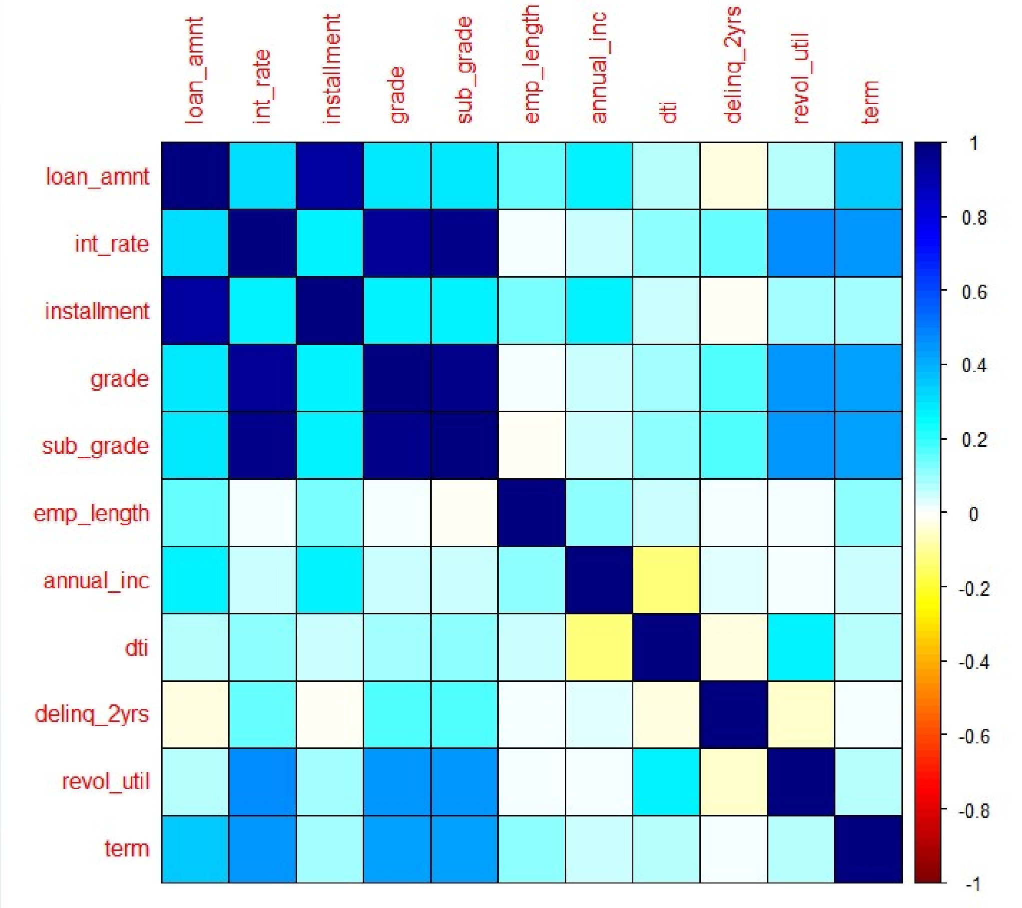
- The plot shows that lower the Loan amount, higher the default risk.

Plot 11 : Revolving Credit Analysis



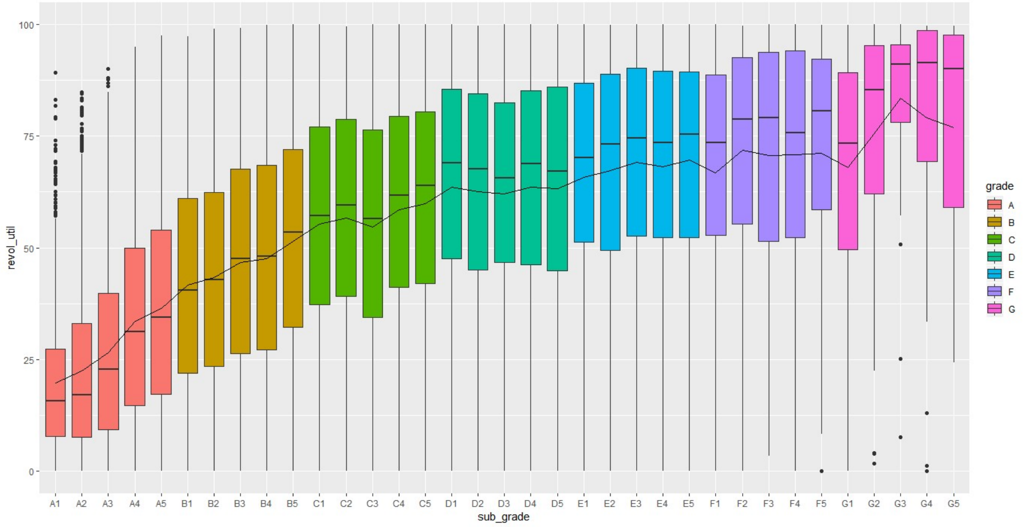
- With the increase of revolving credit chance of default increases.

Plot 12 : Correlation Plot



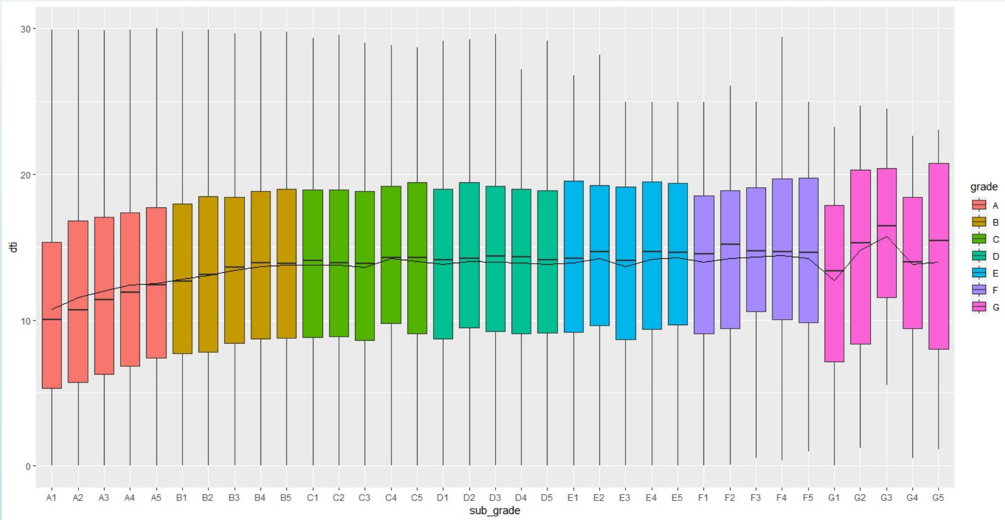
- As expected loan amount and installment are highly correlated. Interest rate and term also have strong correlation with loan amount. And interest rate is 100% correlate with grade/sub-grade.
- Term is mainly correlated with interest rate as expected.
- Revolving line utilization rate (revol_util) has correlation with interest rate and small correlation with DTI.
- Annual income and DTI has small negative correlation as expected.
- Employment length as almost no correlation with any other variables.

Plot 13 : Sub-grade/Grade v/s Others

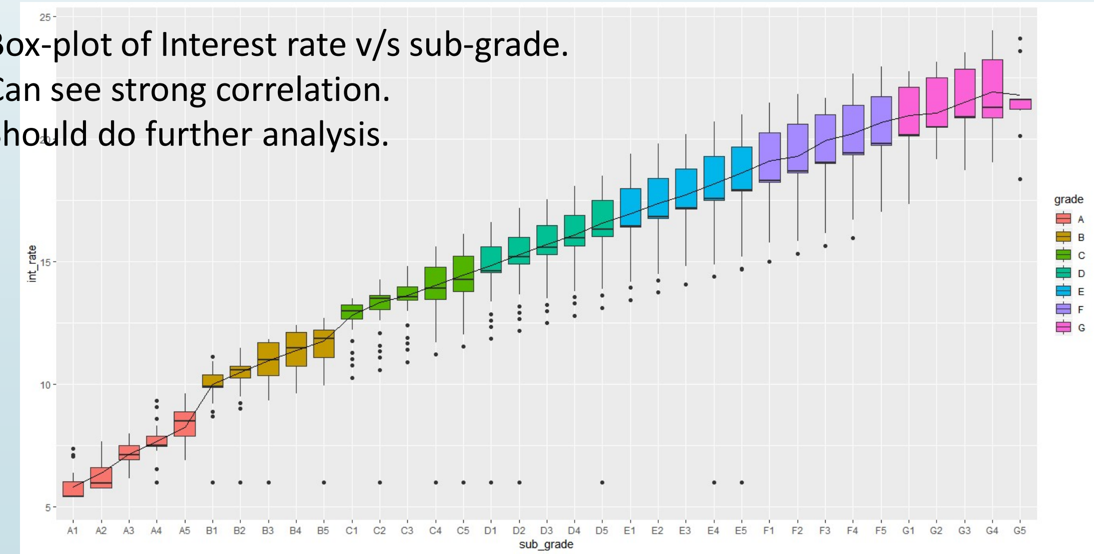


- Box-plot of revolv_util v/s sub-grade.
- Not so much correlation between revol_util and sub-grade.
- Better to keep them separate.

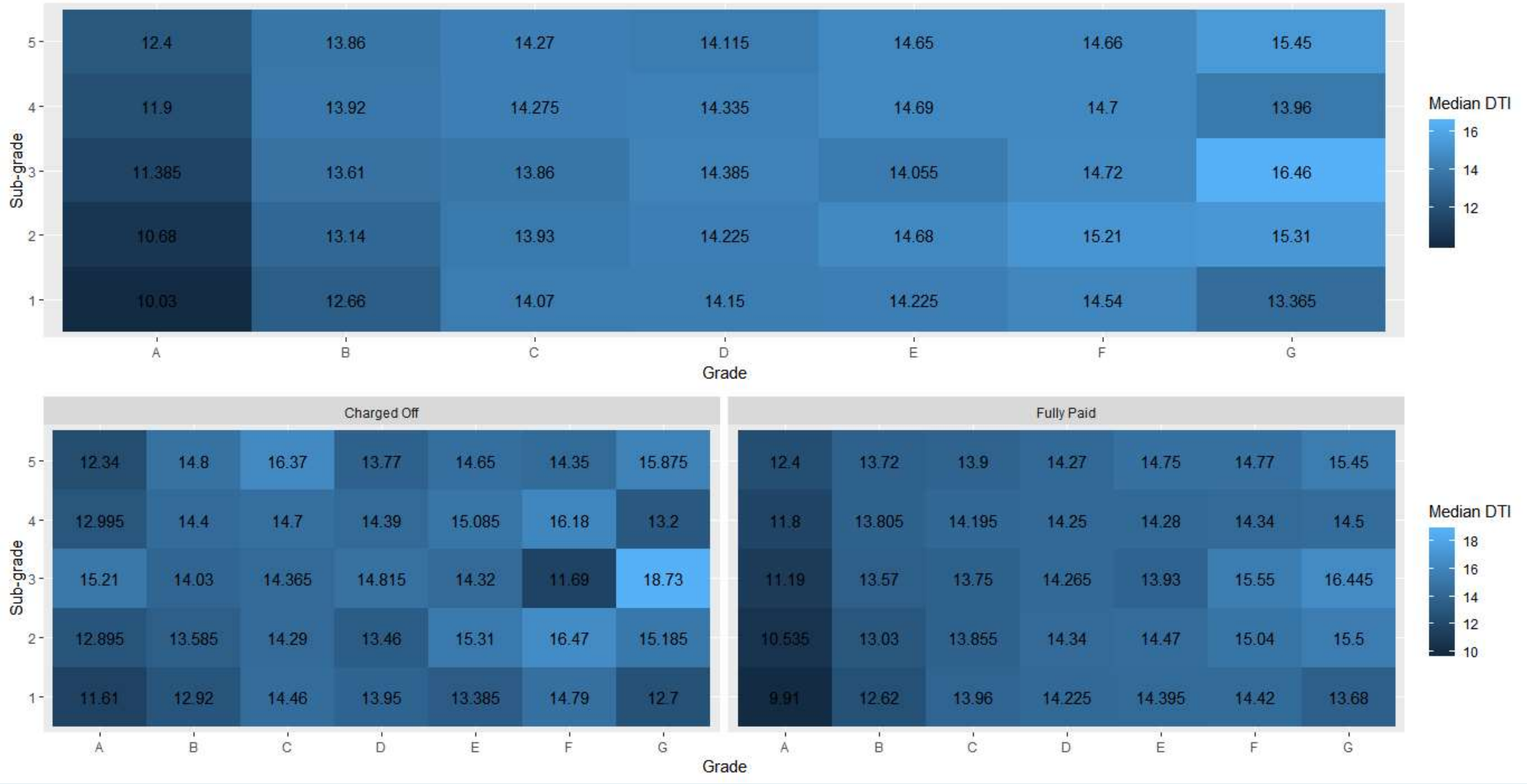
- Box-plot of dti v/s sub-grade.
- Can see strong correlation.
- Should do further Analysis.



- Box-plot of Interest rate v/s sub-grade.
- Can see strong correlation.
- Should do further analysis.

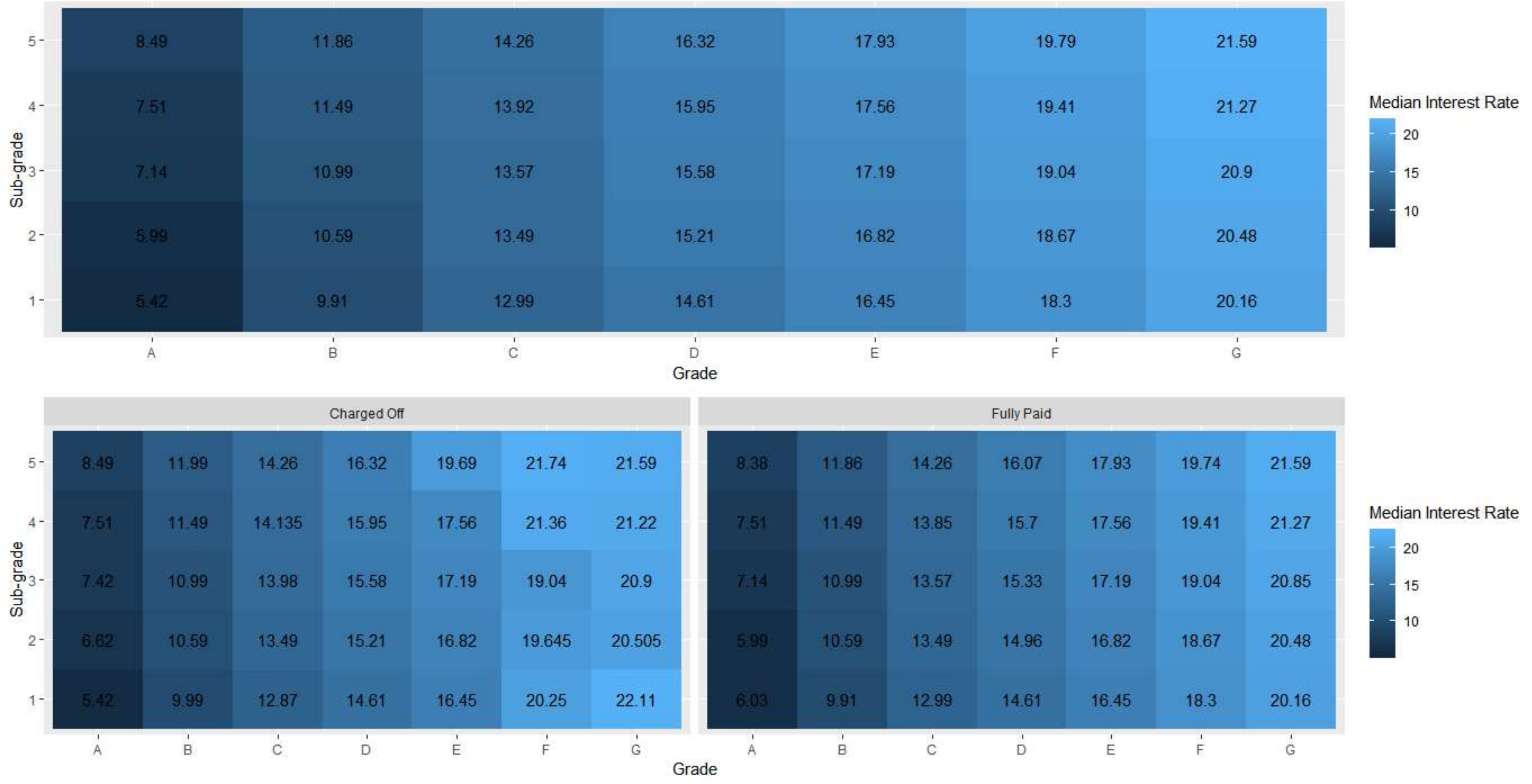


Plot 14 : Sub-grade/Grade v/s DTI



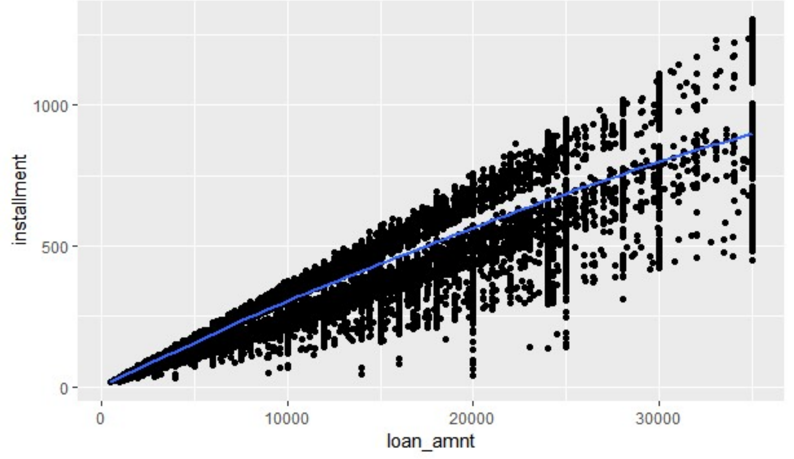
- Heat-map analysis of dti v/s grade/sub-grade.
- Shows very strong correlation.
- **G3 sub-grade is an outlier** throughout the plots.
- Grade/Sub-grade combination can be used as a perfect reflector of DTI.

Plot 15 : Sub-grade/Grade v/s Interest Rate

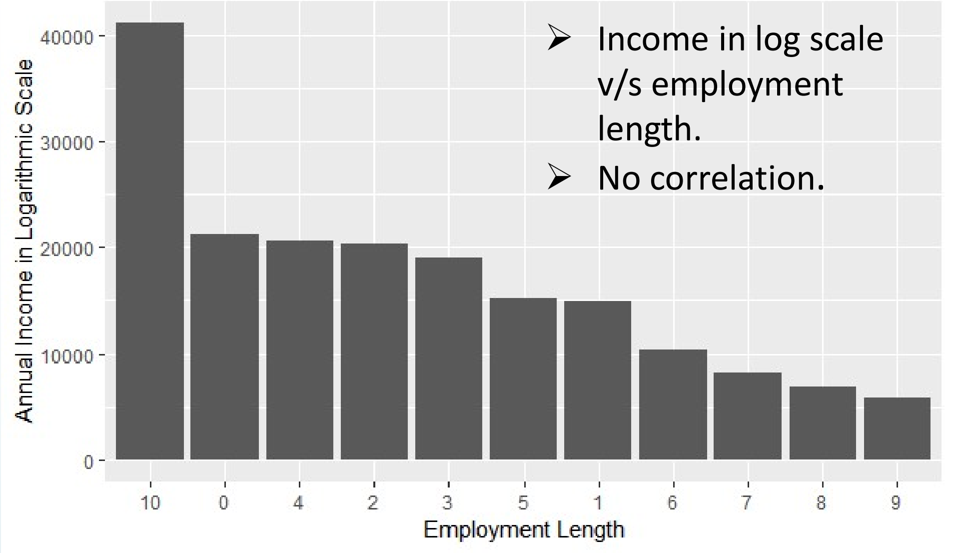


- Heat-map analysis of interest rate v/s grade/sub-grade.
- Shows 100% correlation.
- Does not have any outlier.
- Grade/Sub-grade combination can be used as a perfect reflector of interest rate.

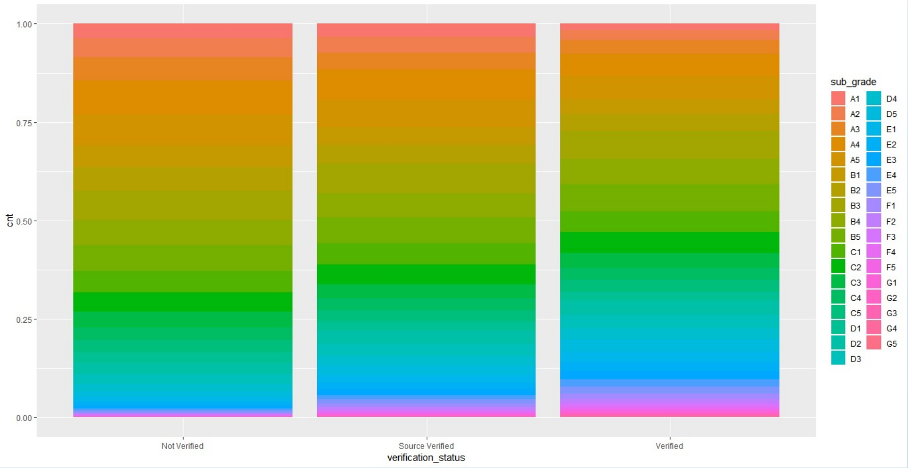
Plot 16 : Other Correlations



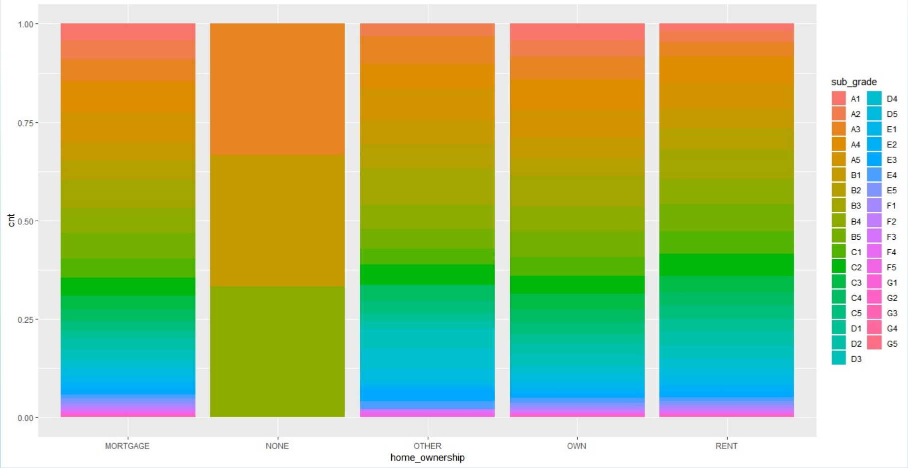
- Scatter plot between loan amount and installment.
- Loan amount can be used as an reflector of installment.



- Income in log scale v/s employment length.
- No correlation.



- Verification status v/s sub-grade.
- No correlation.



- Home ownership v/s sub-grade.
- No correlation.

Conclusions

- From the above univariate and bivariate analysis we get insights about instances affecting default rate in loans.

- **Annual income, verification status and home ownership** play important roles about whether a loan will be fully paid or charged off. People with **annual income between 25k and 75k are riskiest**; people who aren't verified are more prone to charge off; people having **no home/ mortgaged home are riskier**.

- Grade and sub-grade are very important instances; not just they strongly forecast possibility of defaulter but also they reflect number of other instances, i.e., interest rate etc. **Grades B, C, D and E are riskier than others**.

- Purpose is also an important player. So, it to needed to examine carefully for future loans. Some purpose are riskier than others.

- Loan amount and term are another two instances with importance. With **36 months** term there is **higher chance of default** than that of with 60 months. Surprisingly, **percentage of default increases with lower amounts of loan**.

- Employment length and revolving line utilization rate are the last two important instances affecting default percentage. **People with lower employment length are riskier**. And as revolving line utilization rate increase there is increase in number of default loans.

Recommendations

- **Customers should be classified based on above instances for better risk assessment.**
- **More and more loans should be given after proper verifications.**
- **To keep risks low increase of time in terms could be introduced.**
- **Some purpose are associated with more risks, they need to addressed properly.**