GRAMENER CASE STUDY

SUBMISSION

Group Name:

1. Swati Shejwalker
2. Vijaya Lakshmi Potturu
3. Benjamin Turner
4. Somasundaram Balasubramaniyam

Abstract

Objective:

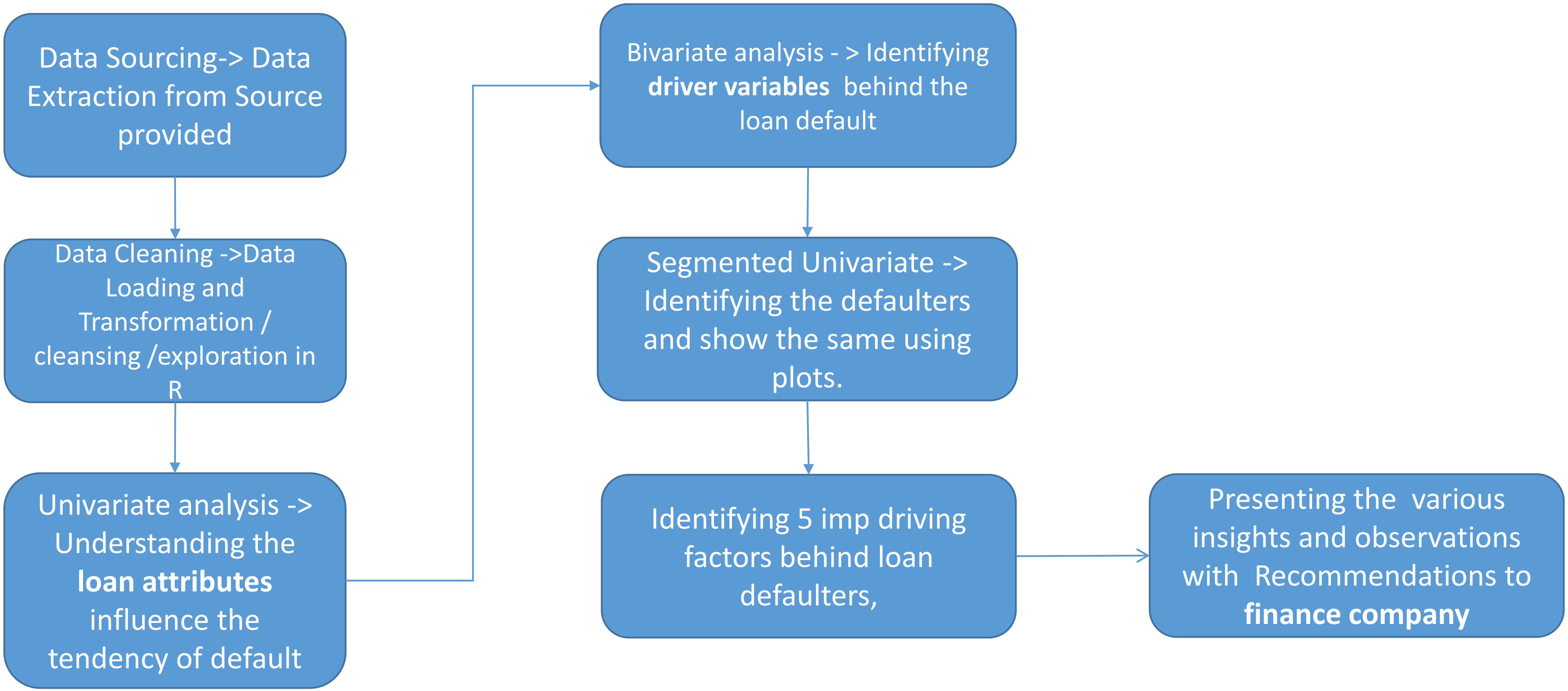
•A **consumer finance company** which specialises in lending various types of loans to urban customers .

•The aim is to **identify patterns which indicate if a person is likely to default** for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

•Use EDA to understand how **consumer attributes** and **loan attributes** influence the tendency of default.

•Understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default to utilise this information for its portfolio and risk assessment. (identify at least the 5 important driver variables)

Problem solving methodology



Data Sourcing : Assumptions

- There are three possible loan scenarios/statuses: fully paid, current, charged-off. But we are interested in identifying clients who default (charged-off) so derived additional field to simplify the three statuses into a defaulted binary (If loan status is charged off then defaulted is 1 otherwise 0).
- Annual income outliers are imputed with 95th percentile income ($142,000).
- During the data analysis to identify the driving factors behind the defaulters are based on default rate ratio.

Data cleaning : Data Quality Issues

- There are 397171 observations with 111 Variables in the Original dataset.
- Fix rows and columns & Missing Values –
  - Many columns containing only value - NA, 0, 'f', 'n' etc are removed from the source dataset.
  - There are columns which contains only 2 unique values i.e 0 and NA, Individual, so these columns are omitted.
- Manipulation of strings and dates
  - Fix Invalid Values
    - Incorrect data types- issue_d, earliest_cr_line , last_credit_pull_d, last_pymnt_d converted R Date format.
  - Standardise Text
    - Remove extra characters from values and convert to numeric (% in int_rate, revol_util) columns in all the Rows
    - Columns 'term' contains char "months" which makes it non-numerical, so remove chars and make the column numeric.
    - Columns 'int_rate' contains char "%" which makes it non-numerical, so remove chars and make the column numeric
    - emp_length' column contains chars like "years", "year", " "; Need to remove these chars to make it numeric.
    - Zip code with XX is removed.

Data cleaning : Data Quality Issues

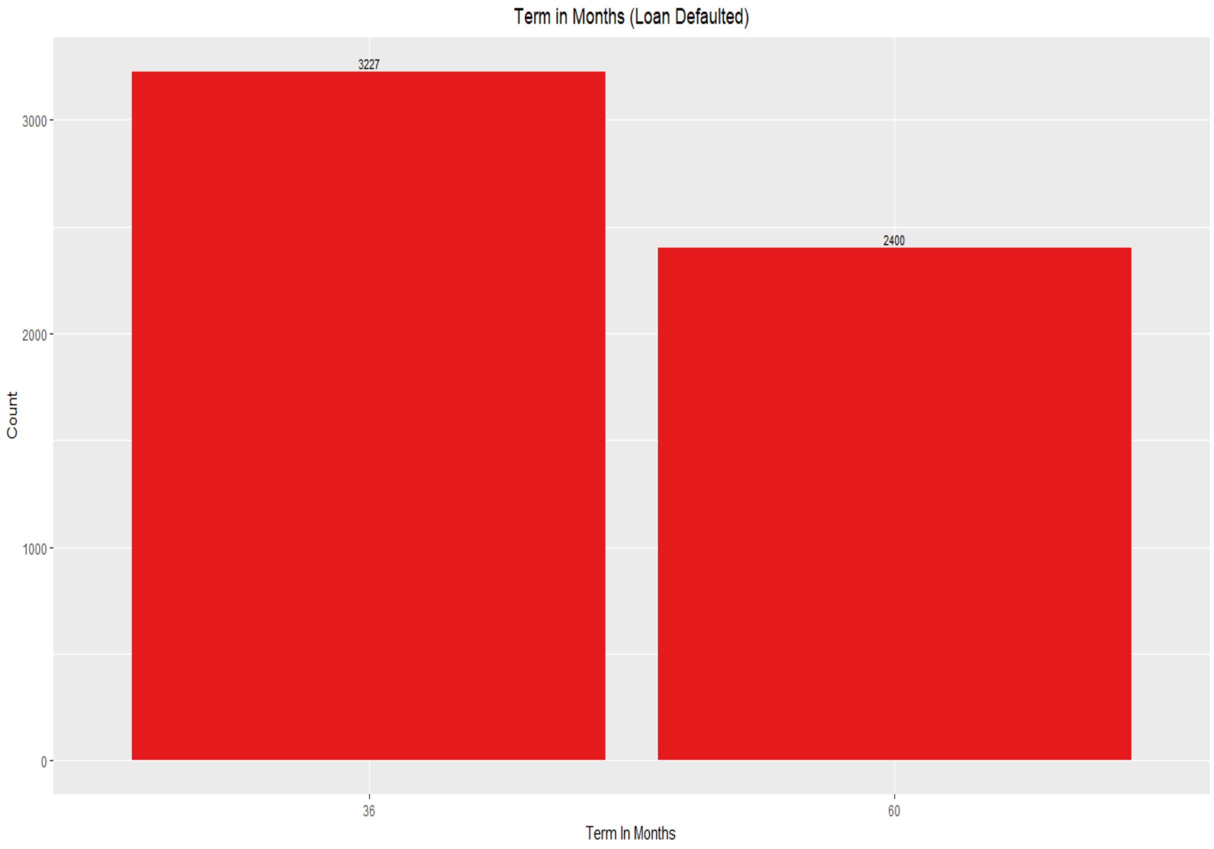
- Driven Metrics – (14 New variables derived)
  - Business-driven – annual_inc_range, dti_bucket ,income_bin, loan_amnt_bin, dti_bin, revol_util_bin for **segmented univariate analysis** are created.
  - Type – driven - Issue Month and year column - issue_dyr (Year) Issue_dm (Month) , earliest_cr_line_year ,generated a latitude and longitude,city,state for plotting data on a map.
  - data-driven metrics – defaulted
- Standardise Numbers- Over-precision in funded_amnt_inv column
- Filter Data – Columns irrelevant to analysis are removed (Desc,URL)
- Missing value imputation, outlier treatment
  - public record bankruptcies -NAs with median figure, which is 0.
  - Title for the loan entered by the borrower is empty. Set NA to empty string.
  - Outlier in annual income does not seem to make much difference, but handling of outliers is on the evaluation rubric, Cap high incomes (above 1.5 * IQR = # $145,144) at the value of the 95th percentile income ($142,000).

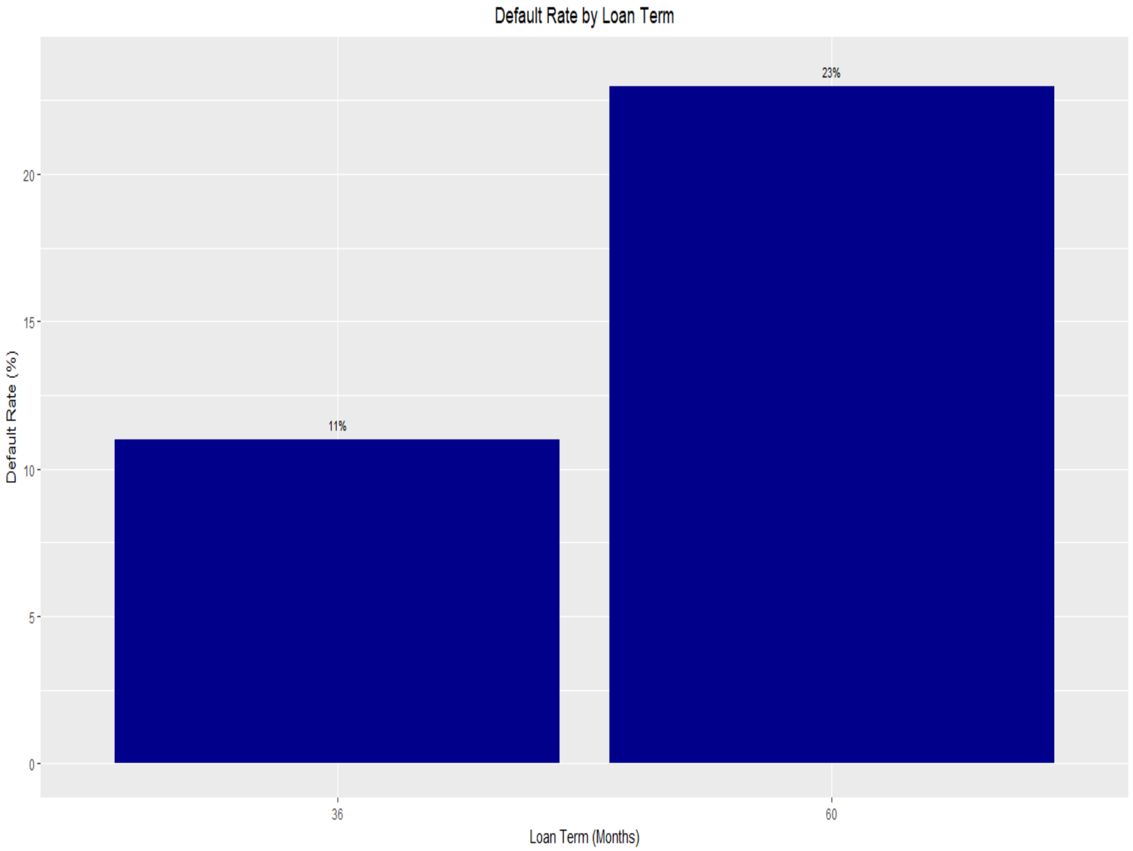
Data Analysis: univariate analysis

- **Univariate Analysis :**
  - Loans , loan amount > $17,500 have default rate >15%, those for smaller amounts all have <15% defaults.
  - **60-month term loans (23% default rate), others have 11% so it almost doubles the default rate.**
  - **Grade E (25% default rate), F and G both >30% ,with in E , subgrade E4 >28% ,F -> F4,F5 >28%, G->G2,G3 and G5 >28%.**
  - Emp length of "n/a" has default rate of ~21% (others all <15%)
  - Home ownership of "OTHER" has rate of ~18%. Others all <15%
  - **Annual income < $20,000 has default rate of 24%, others all <18%.**
  - Verification status verified (16%), compared to not verified (13%)
  - **Purpose - small business (26%), others all <18%.**
  - inq_last_6mths >= 6 (25%)
  - **Revol util NA (33%). Everything other than NA has less than 22% default rate**
  - **Pub rec bankruptcies of 1 are >(22%) compared to 14% with no previous bankruptcy.**
  - The higher the grade (more risky loan), the higher the interest rates.
  - No Clear pattern observer with dti.
- **Most Important driving factors :** Term, Grade, Annual Income, Purpose, Revolving line utilization rate and Public record bankruptcies

Data Analysis : Univariate analysis -Term

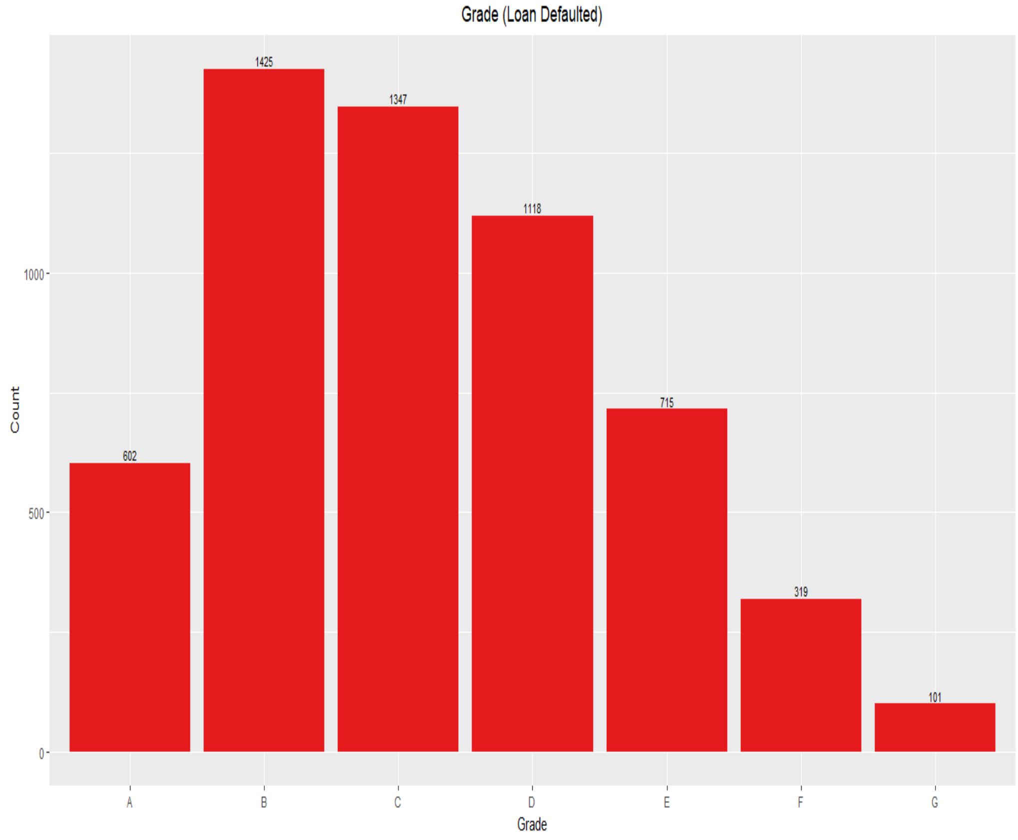
- **60-month term loans are (23% default rate), others have 11% so it almost doubles the default rate.**

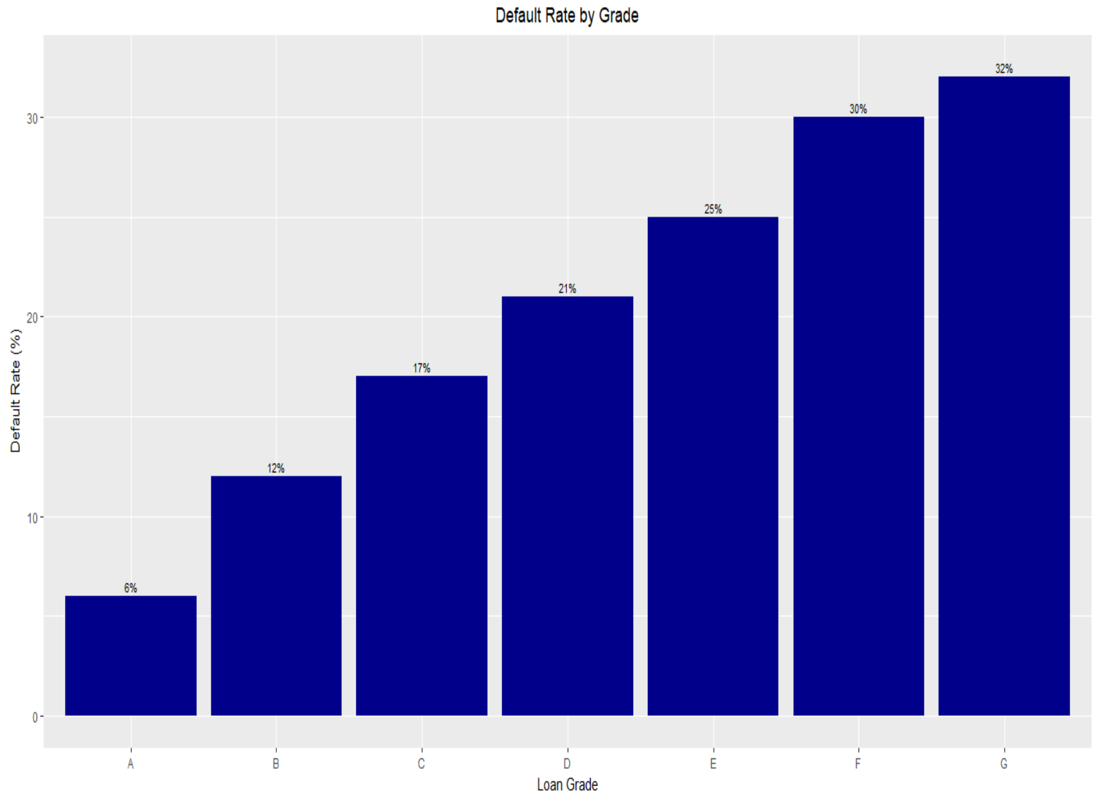




Data Analysis : Univariate analysis - Grade

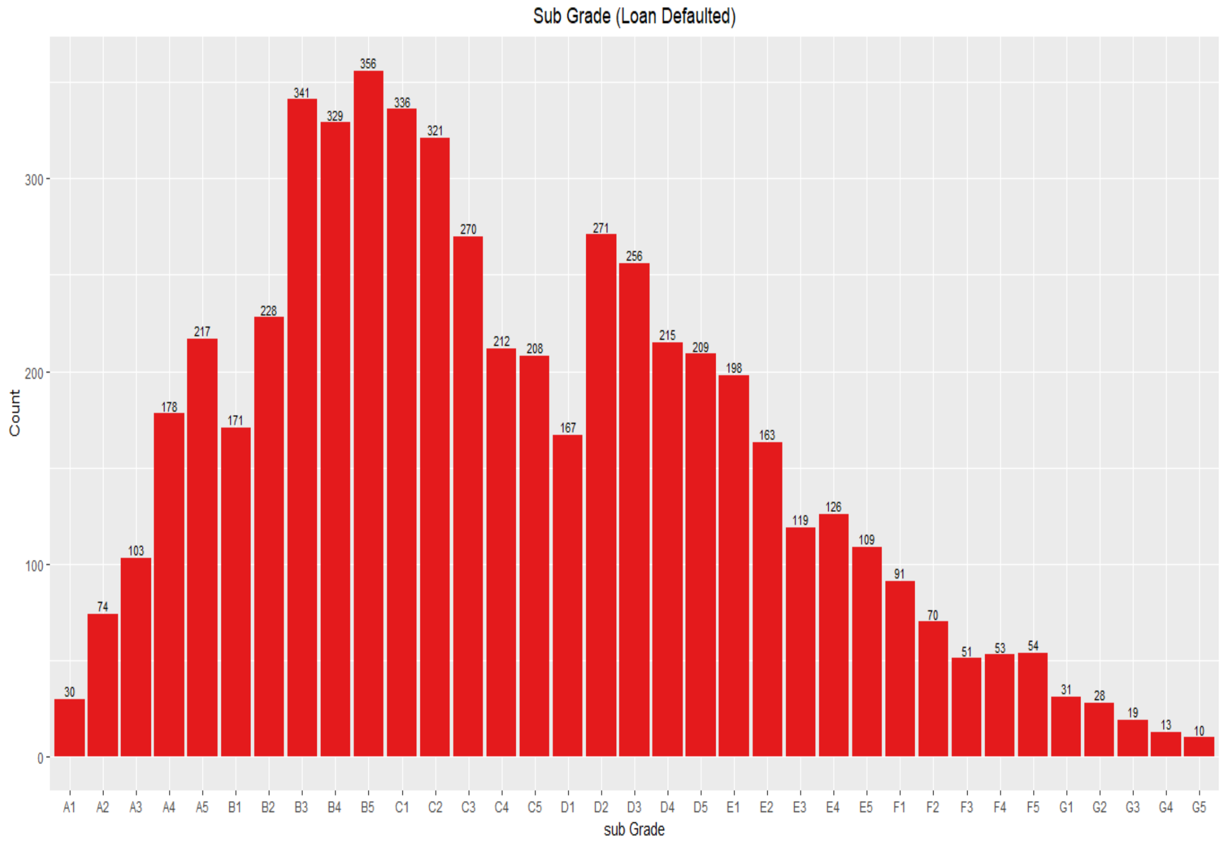
- **Grade E (25% default rate), F and G both >30%.**

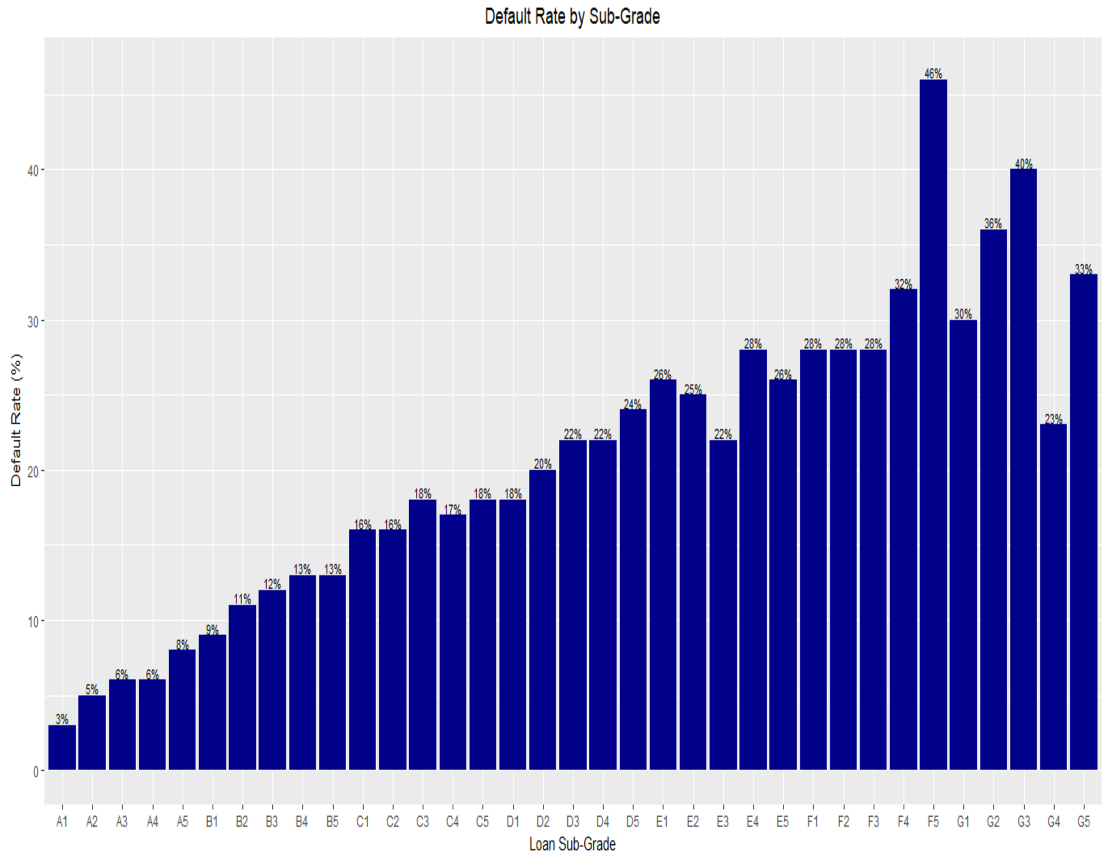




Data Analysis : Univariate analysis –Sub Grade

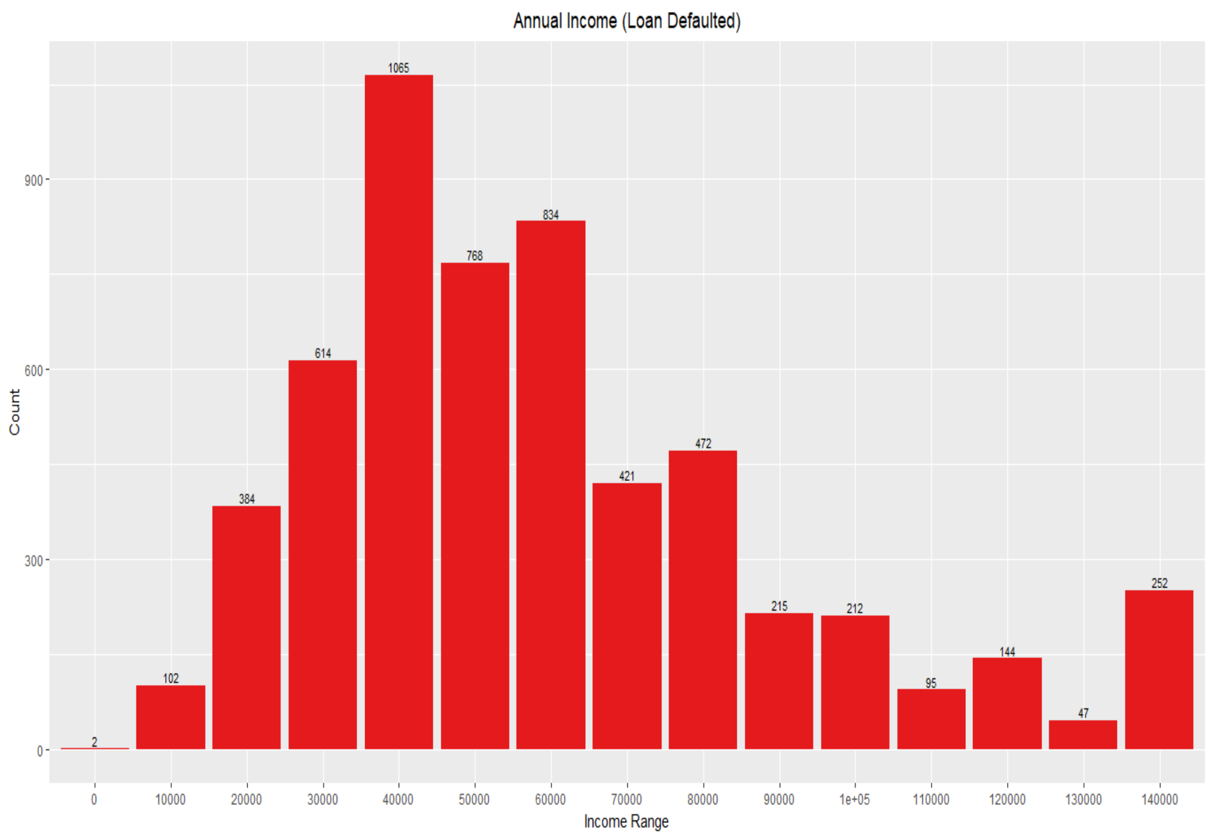
- **Within E, Subgrade E4 >28% ,F -> F4,F5 >28%, G->G2,G3 and G5 >28%.**

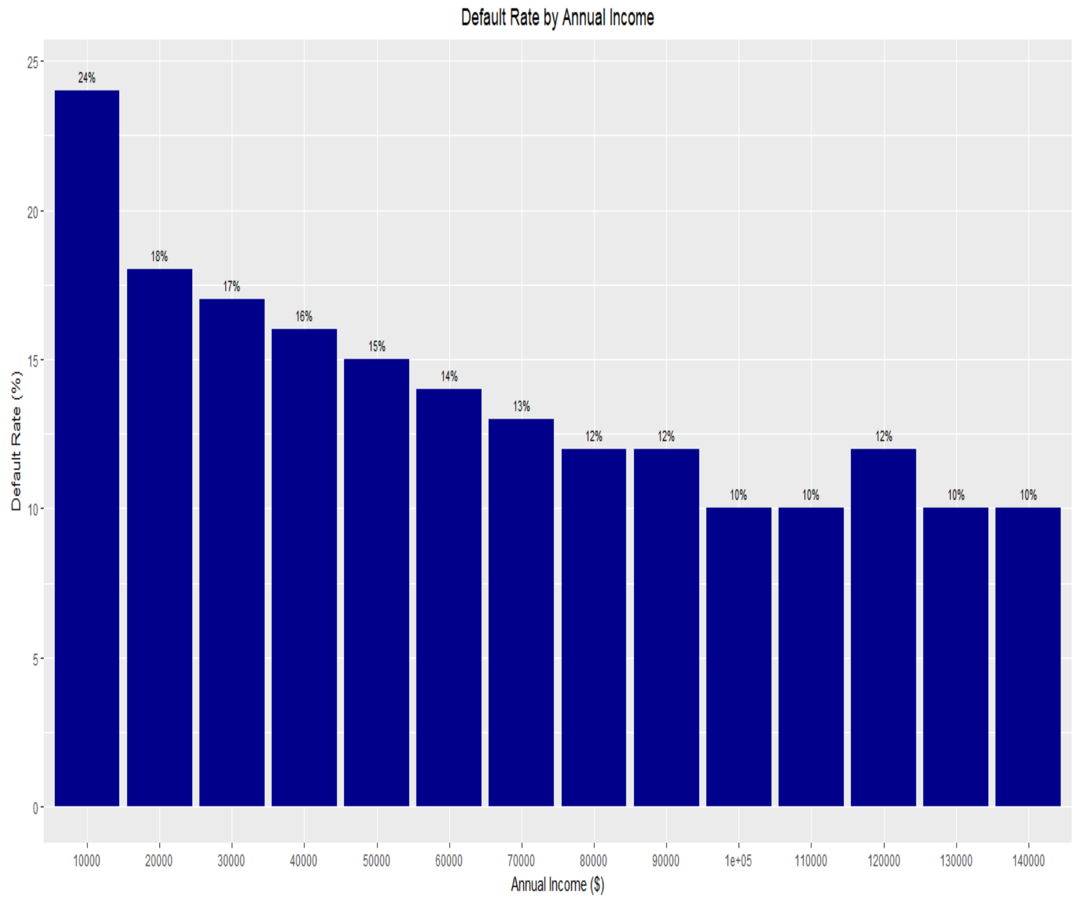




Data Analysis : Univariate analysis –Annual income

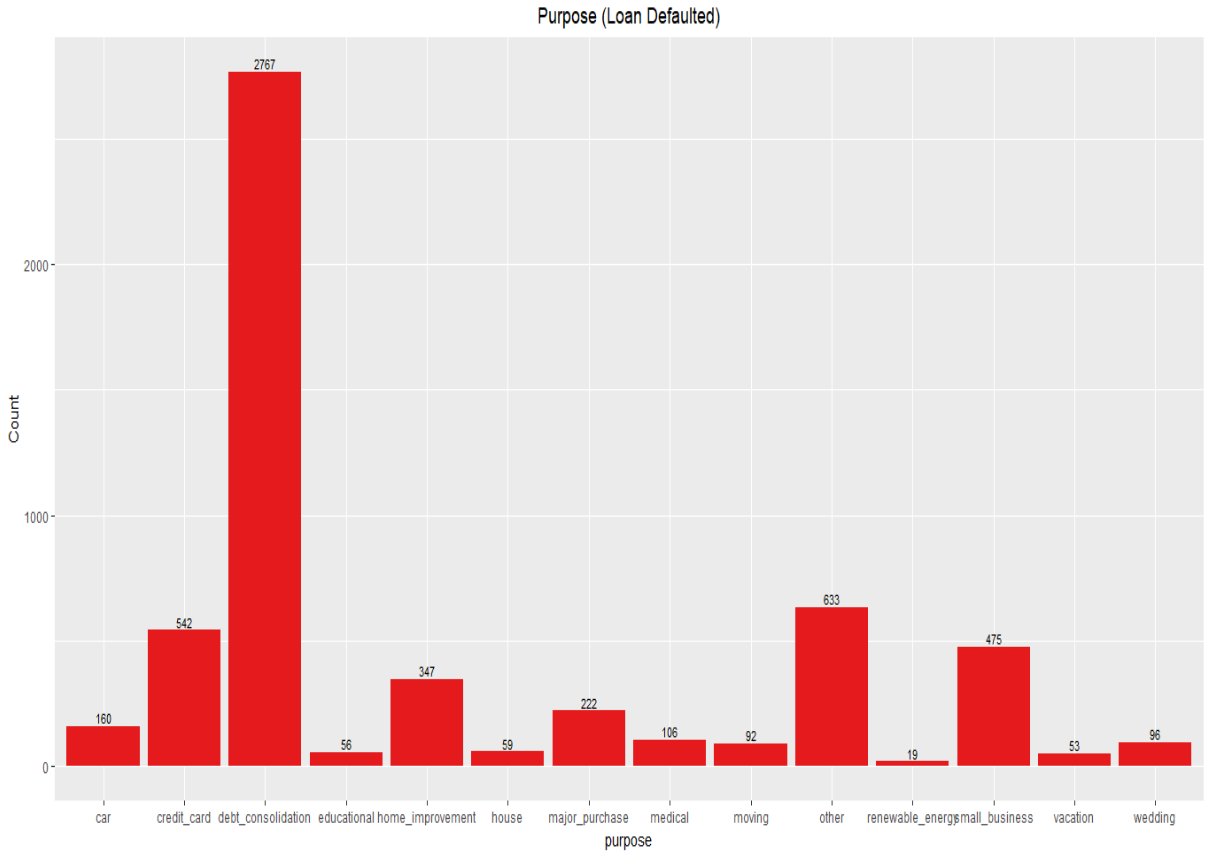
- **Annual income < $20,000 has default rate of 24%, others all <18%.**

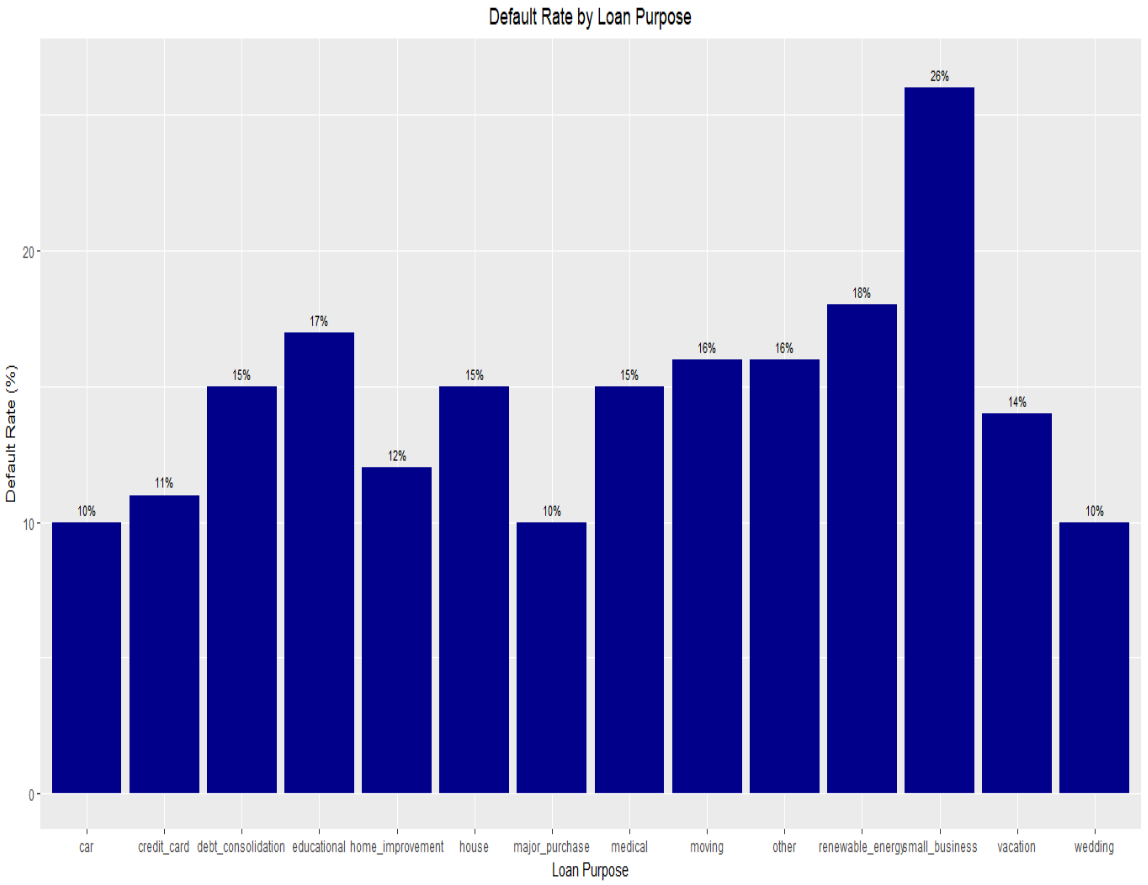




Data Analysis : Univariate analysis –Purpose income

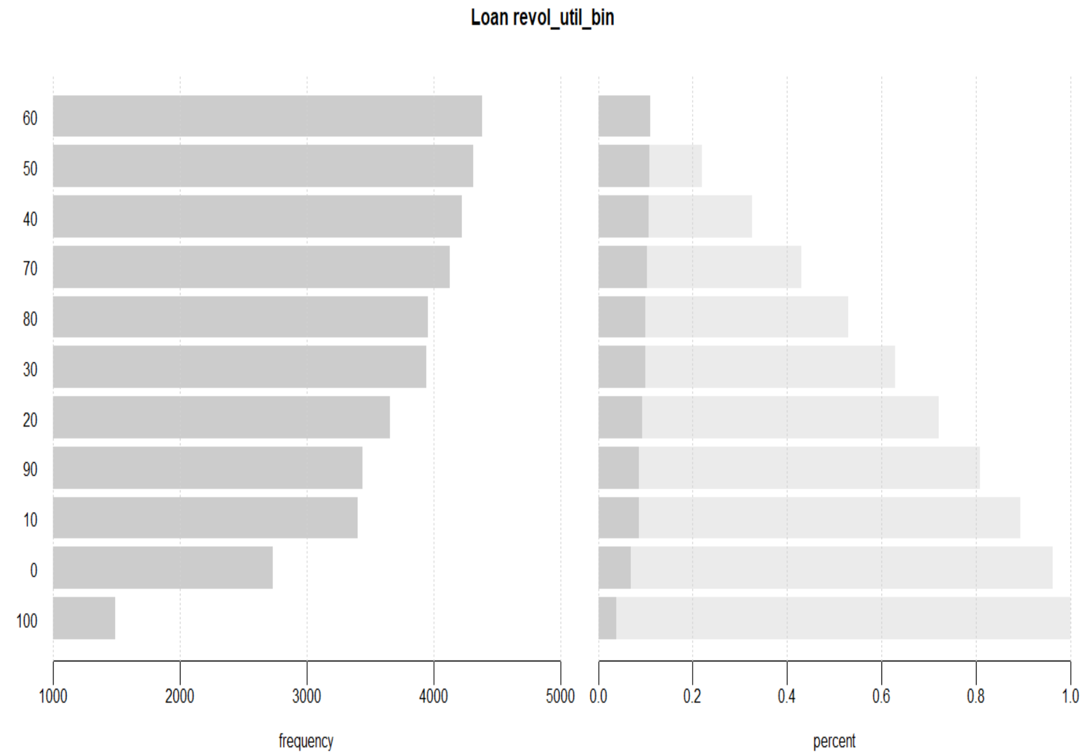
- **Purpose – small business (26%), others all <18%.**

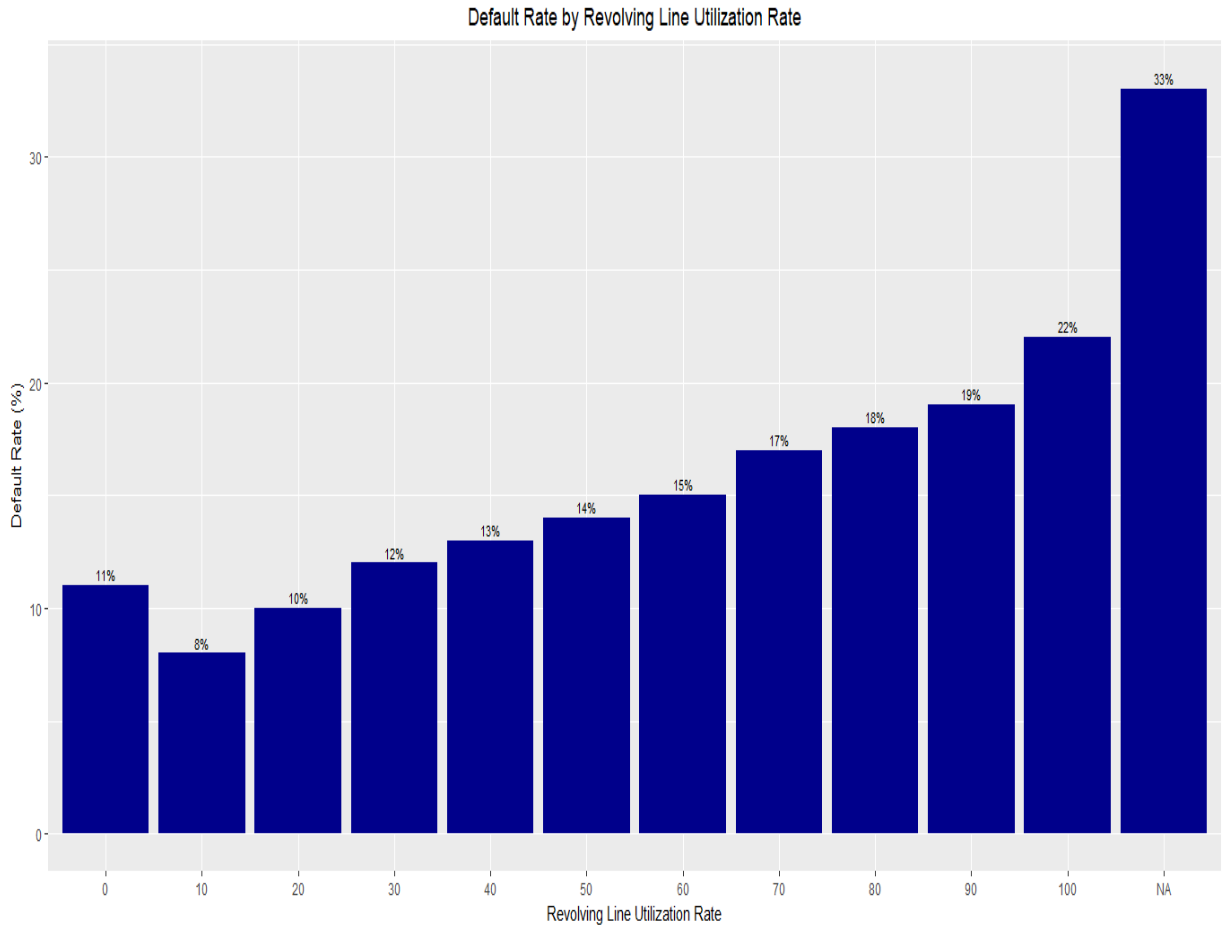




Data Analysis : Univariate analysis –**Revolving line utilization rate.**

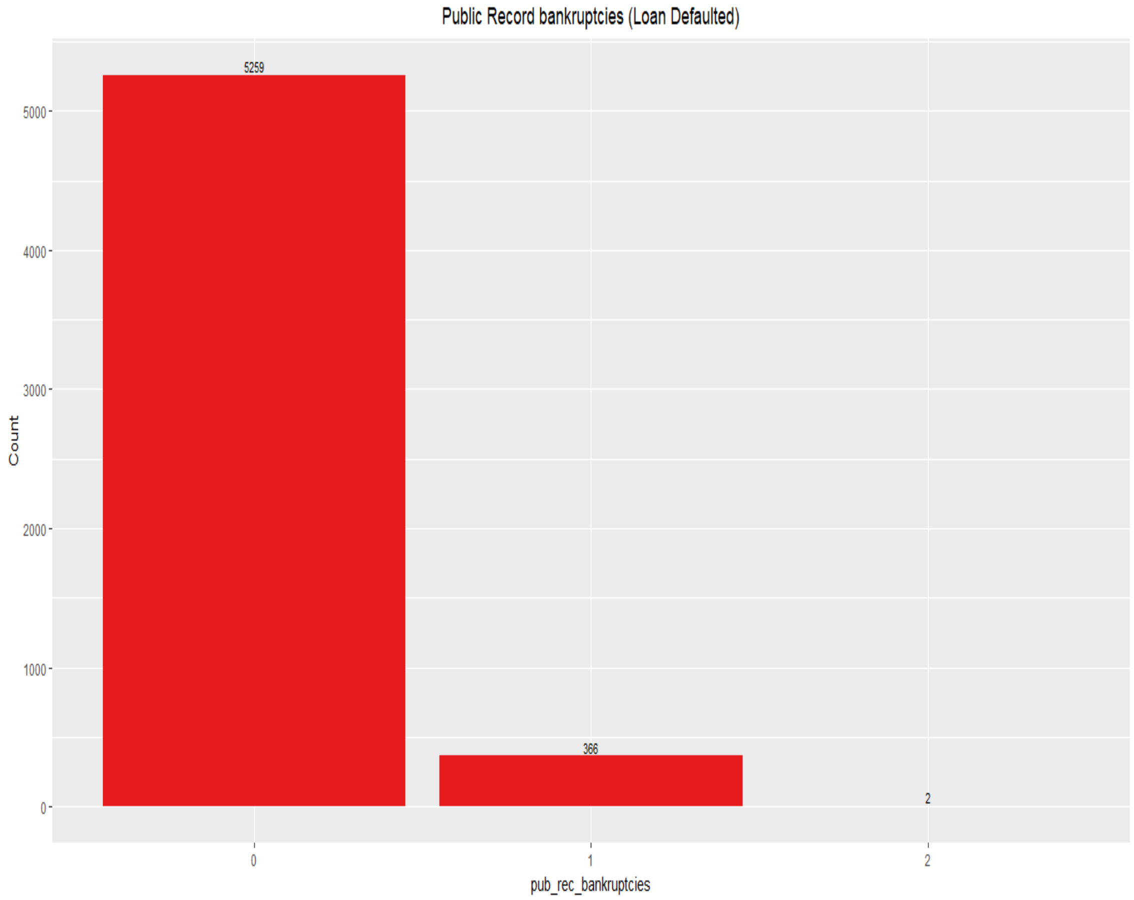
- **Revol_util NA (33%). Everything other than NA has less than 22% default rate**

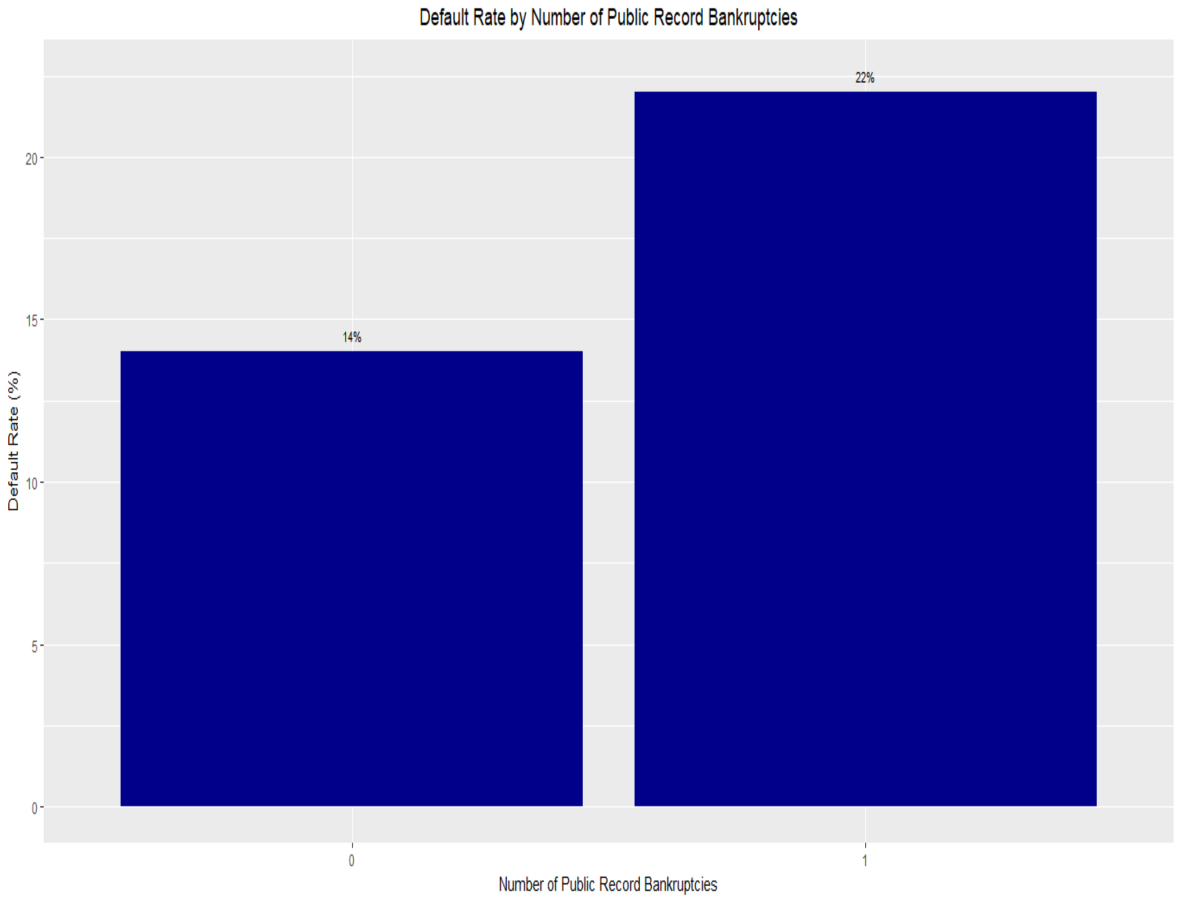




Data Analysis : Univariate analysis – public record bankruptcies

- **Pub_rec_bankruptcies of 1 are >(22%) compared to 14% with no previous bankruptcy**



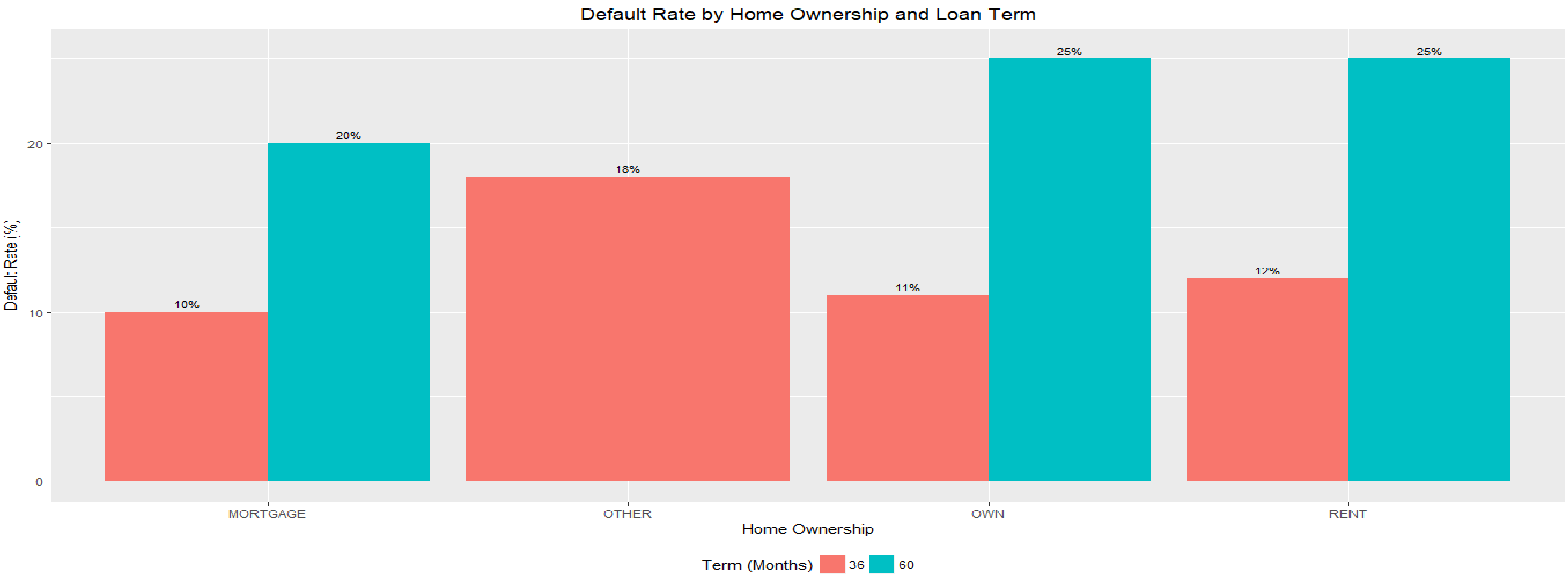


Data Analysis : Bivariate analysis

- **Bivariate Analysis:**
  - Grade G loans of term 36 months: default rate 38% (but only accounts for 56 loans in total)
  - Employment length “n/a” for 60 month loans: default rate 28% (266 loans in total)
  - **Home ownership OWN (757 loans) or RENT (4207 loans) AND 60-month loan_term (both have 25% default rate)**
  - **Annual income <$50,000 and 60-month term all income levels have default rate >25% and make up 2717 loans in total**
  - **Purpose - small business AND 60-month term: 35% default rate of 589 loans**
  - **Pub_rec_bankruptcies >0 AND 60-month term: 35% default rate of 498 loans**
  - annual income <$10,000 and home ownership RENT 26% (324 loans)
  - **Purpose small business AND home ownership OWN (32% default rate of 110 loans default) or RENT (29% default rate of 775 loans default)**
  - Pub_rec_bankruptcies > 1+ ,loan_amnt >12,500 (>25% of 523 loans)
  - Pub_rec_bankruptcies >1+ and verification status verified (24% of 527 loans) or Source Verified (25% 407 loans)
- **Most Important driving factors :** Home Ownership Vs Term, Annual Income vs Term, Purpose, Vs Term, Public record bankruptcies Vs Term , Purpose Vs Home Ownership.

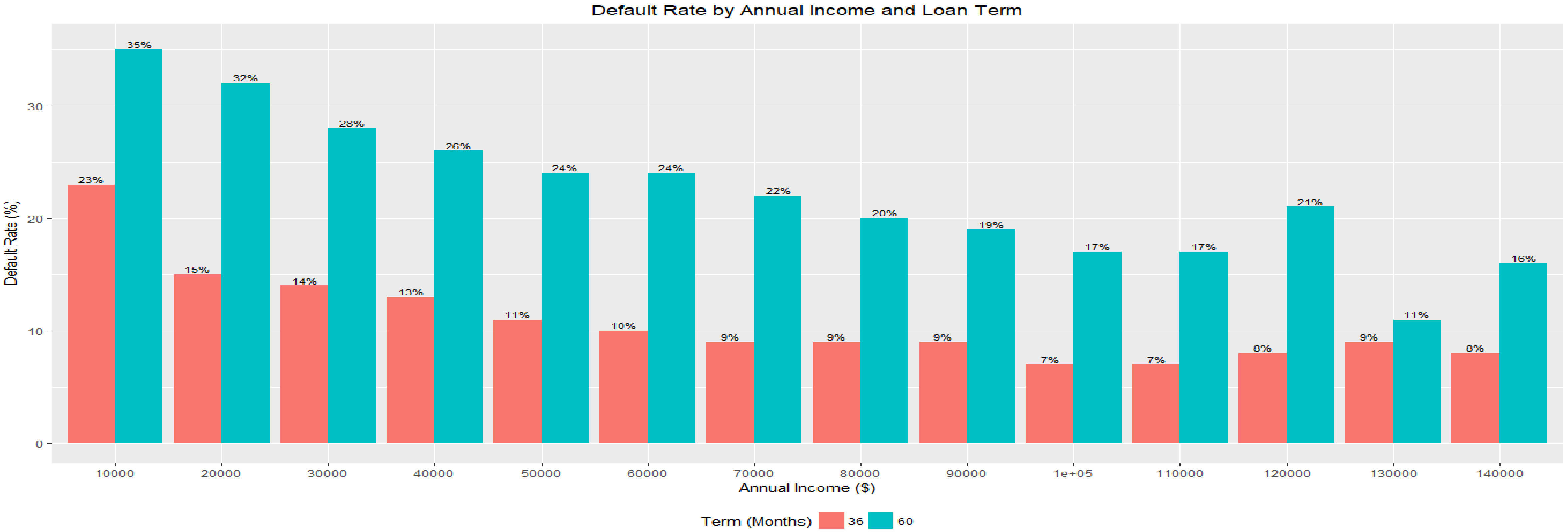
Data Analysis : Bivariate analysis – Home Ownership Vs Term

Home Ownership OWN (757 loans) or RENT (4207 loans) AND 60-month loan term (both have 25% default rate)



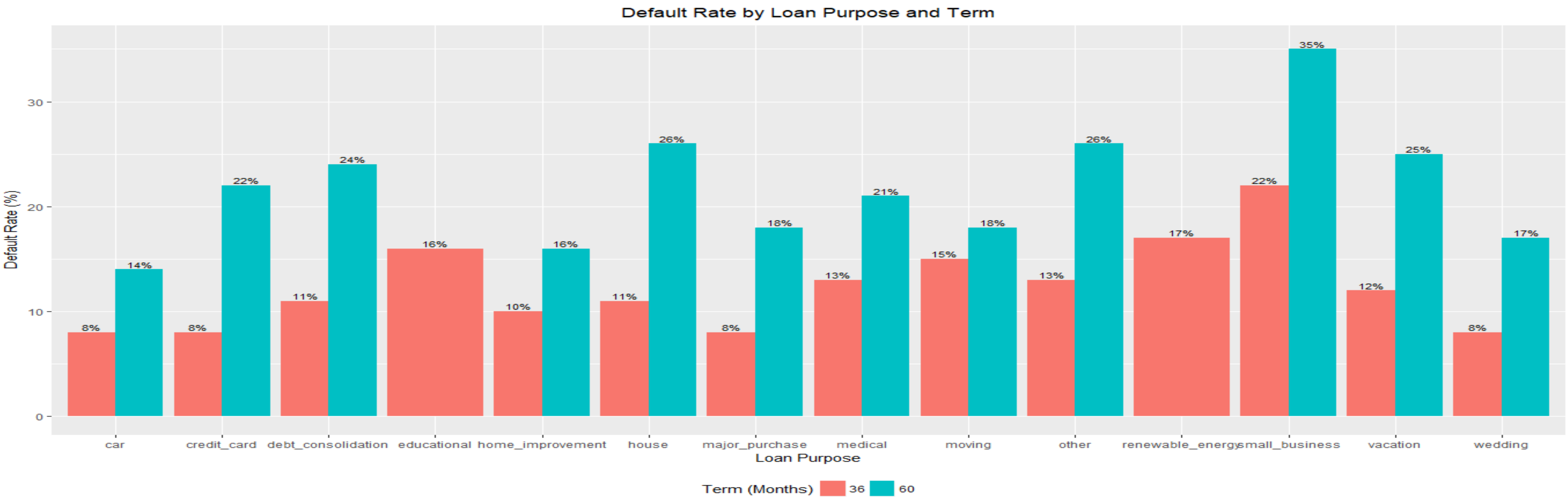
Data Analysis : Bivariate analysis – Annual income Vs Term

- **Annual income <$50,000 and 60-month term all income levels have default rate >25% and make up 2717 loans in total**



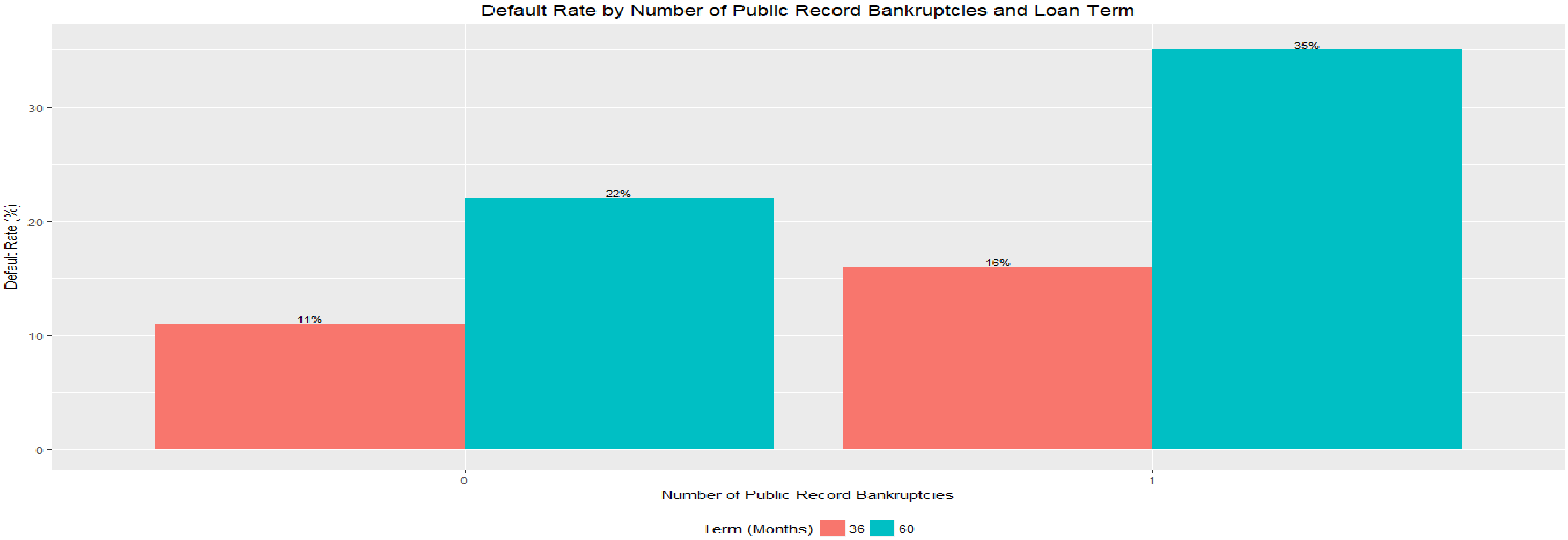
Data Analysis : Bivariate analysis- Purpose Vs Term

Purpose -Small_business AND 60-month term: 35% default rate of 589 loans



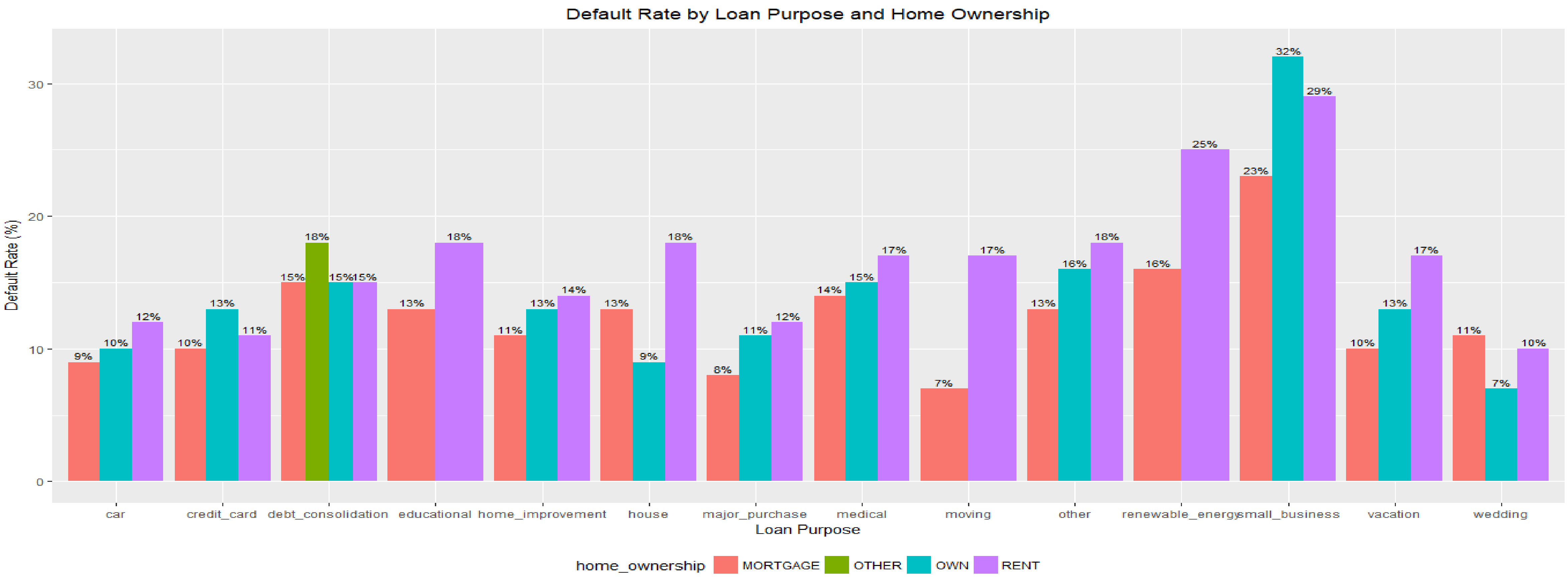
Data Analysis : Bivariate analysis- Public Record Bankruptcies Vs Term

- **Pub_rec_bankruptcies >0 AND 60-month term: 35% default rate of 498 loans**



Data Analysis : Bivariate analysis- Purpose Vs Home Ownership

Purpose - small_business AND home ownership OWN (32% default rate of 110 loans default) or RENT (29% default rate of 775 loans default)



Conclusions

- Important driver variables - Term, Annual Income, Purpose, Loan Amount, Grade and Sub Grade, Revolving line utilization rate and public record bankruptcies, Home OwnerShip are behind the Loan defaulter analysis.
- Dti and interest rate does not seem to show an strong dependency.
- Verification Status, Employment Length does not seem to show an strong dependency.

Recommendation

- Build a credit model and see if we can predict reliably defaulters with important variables identified.
- Build a more robust process for customer verification to control the defaults.
- Based on the Model from important variables, classify clearly the customers segmentations to take a decision for denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.