Lending Club Case Study Assignment

Jatin Arora

Objective

- **Company**: Lending Club 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.

- **Context:** Lending Club wants to understand the **driving factors** behind loan default or non-default which are strong parameter of default. The company can utilize this knowledge for its portfolio and risk assessment in terms of loan issue.

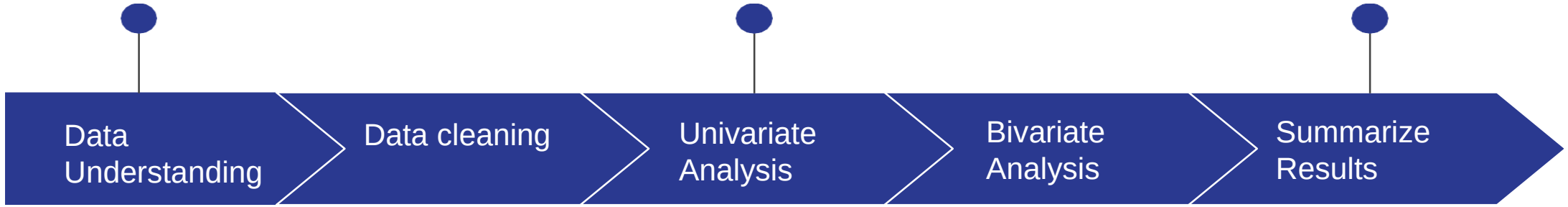
- **Problem Statement:** As a person working for Lending Club analyze the dataset containing information about past loan applicants using EDA to understand which consumer attributes and loan attributes influences the tendency to default.

Analysis Approach

- Load dataset in python
- Identify categorical and quantitative fields
- Drop rows for loans that are current

•Check data distributions of quantitative variables and default rate across categorical variables
•Create derived variables

Publish insights and observations



•Drop columns with null values, single value
•Impute data for columns with missing values
•Remove Outliers and text/symbols- '%',etc
•Create necessary columns required for analysis

•Check how two variables affect each other or a third variable
• Analyze categorical to categorical and continuous to categorical distributions analysis.

Data Understanding

- The dataset has 39717 rows and 111 columns.
- There are many columns which consists of null value and NAN values.
- The dataset has mixture of categorical and continuous data. Below are selected values of importance.

Categorical variables=['term', 'grade', 'sub_grade', 'emp_length', 'home_ownership', 'verification_status', 'issue_d', 'loan_status', 'purpose', 'addr_state', 'pub_rec_bankruptcies']

Quantitative variables =['loan_amnt', 'funded_amnt', 'int_rate', 'annual_inc', 'dti'].

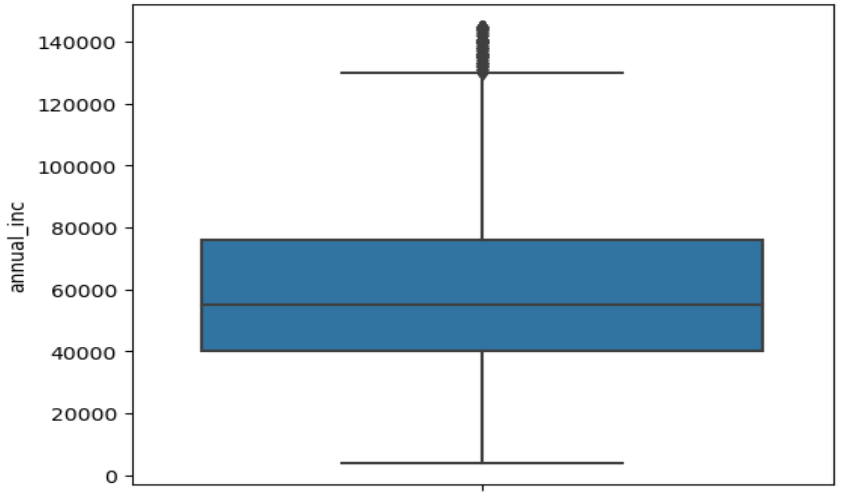
- Loan status Charged off means defaulted loans and Current means currently running.
- Employment Length shows the work experience of customer applying for loan.
- Dti is debt to income ratio.
- Our dataset needs a good amount of pre-processing and data imputation for 2 columns of importance.

Data Manipulation and Pre-Analysis

- We have loaded data and then dropped all the duplicate values in the dataset.
- We removed the columns that have single occurrence values or Null
- Then we have dropped the unnecessary columns like “id” and “url” etc.
- Interest rate column consists of percentage(%) which has been removed. Term column has ‘months’ which is also removed.
- ‘Emp_length’ and ‘Pub_rec_bankruptcies’ columns have missing values which are imputed with mode.
- We have set date related columns into proper date format like ‘2021-02-21’ and selected appropriate data type for Interest rate, month, year and Term
- Outliers are removed from annual income, interest rate and loan_amount columns using IQR.
- Now our dataset is ready for data analysis.

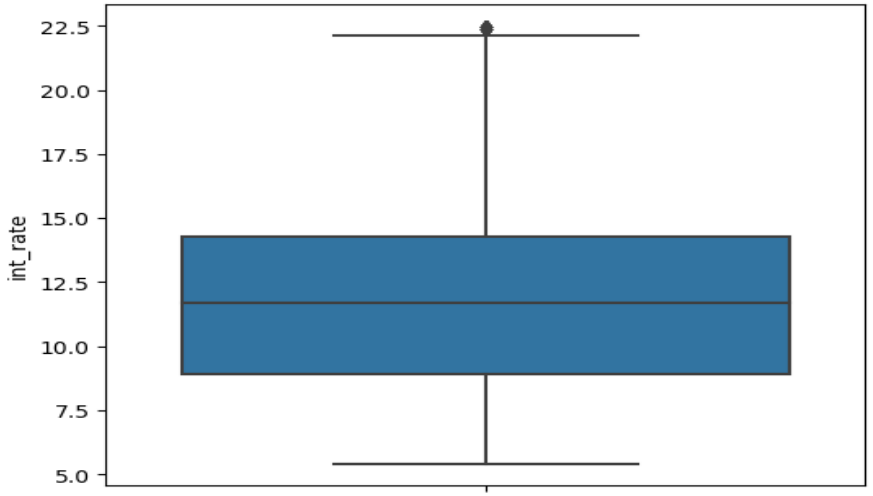
Univariate Analysis - Continuous

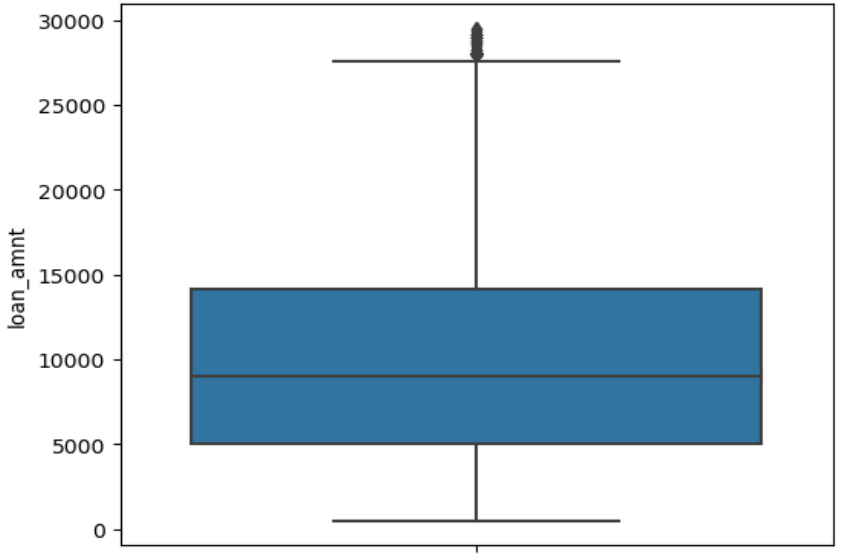
- We plotted a Histogram and Boxplot of the column ‘annual_inc’, ‘Int_rate’, ‘loan_amnt’ and ‘dti’



count	35945
mean	60451.14
min	4000
Q1-25%	40000
Q2-50%	55000
Q3-75%	76000
max	145000

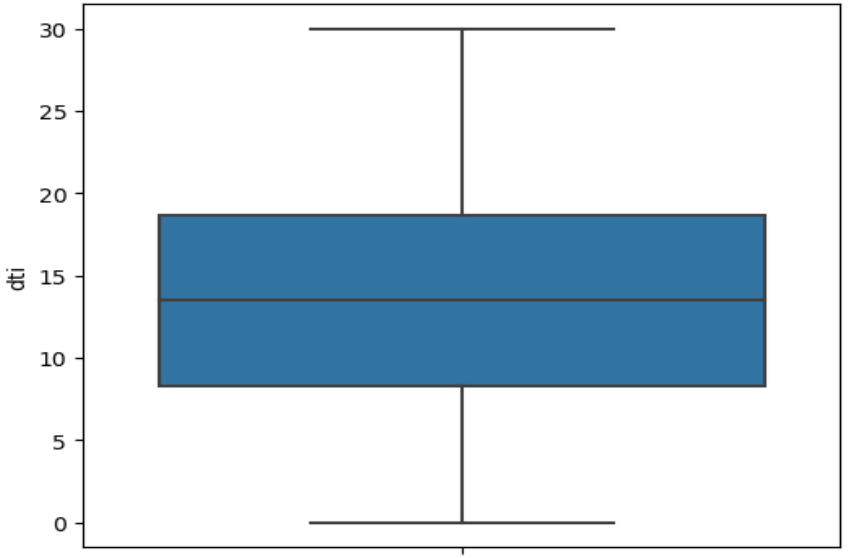
count	35945
mean	11.79
min	5.42
Q1-25%	8.9
Q2-50%	11.71
Q3-75%	14.26
max	22.48



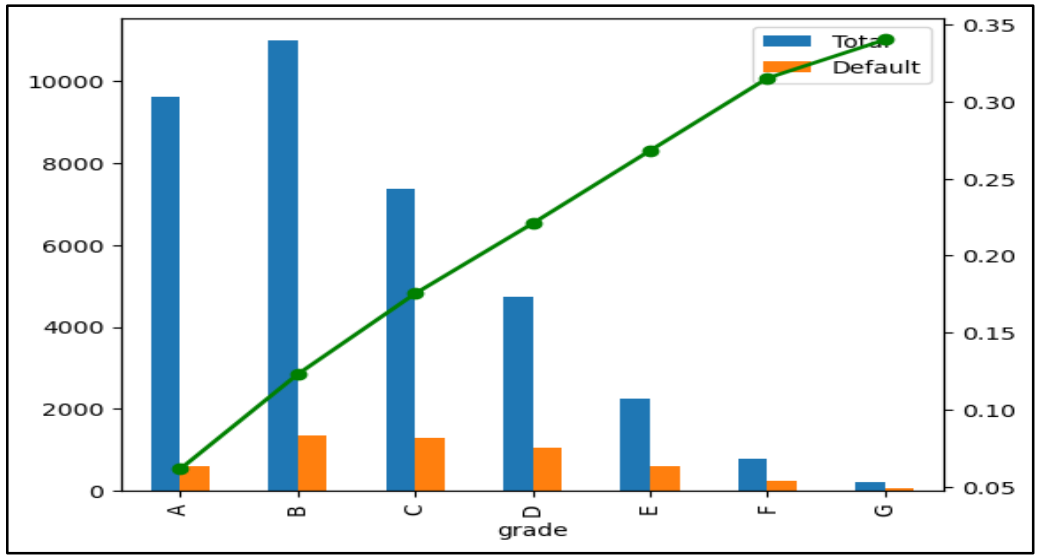


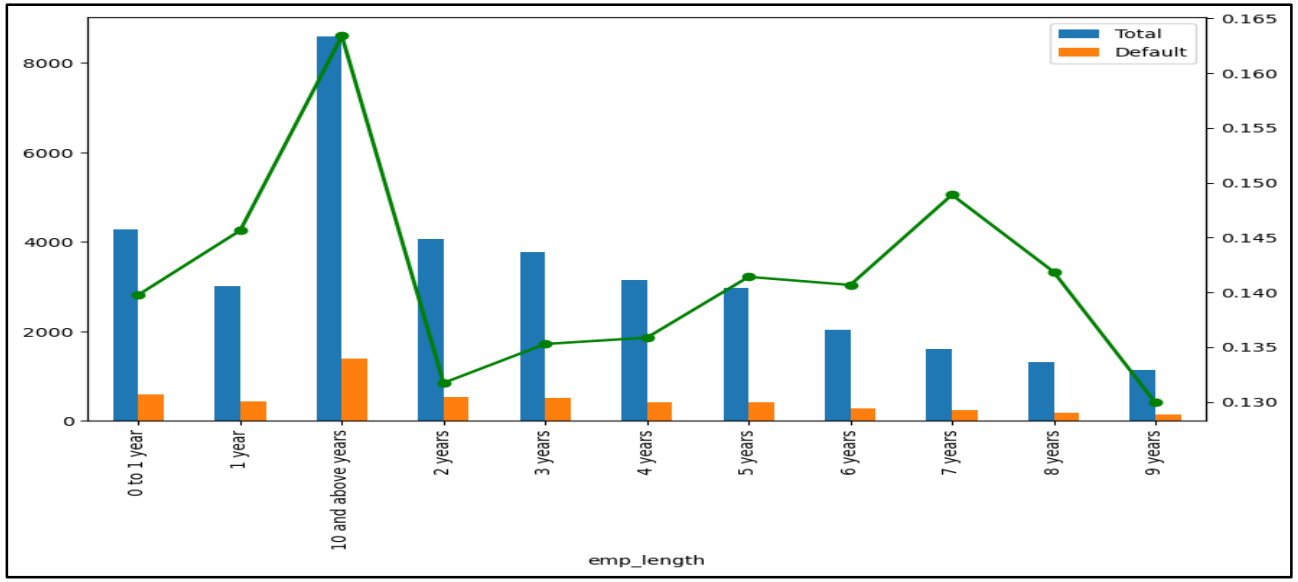
count	35945
mean	10184.59
min	500
Q1-25%	5000
Q2-50%	9000
Q3-75%	14125
max	29500

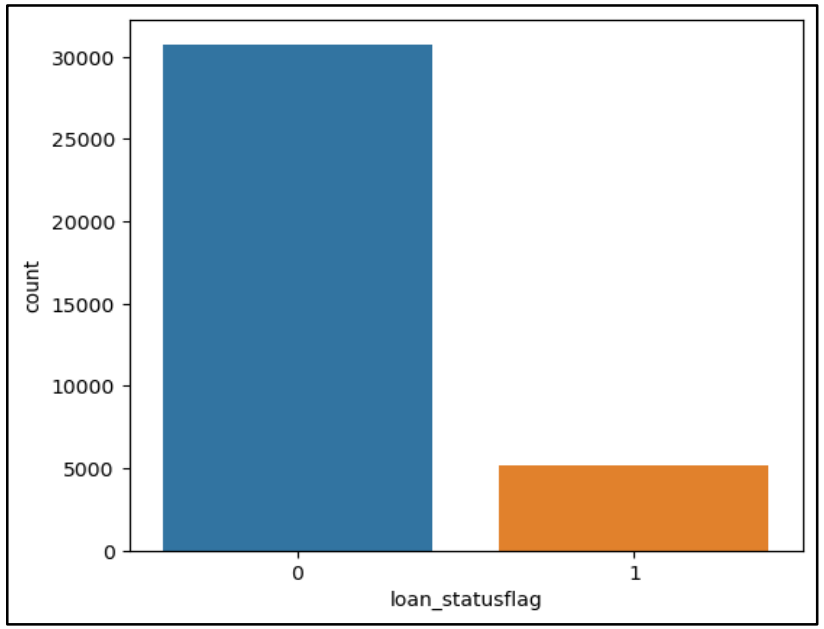
count	35945
mean	13.41
min	0
Q1-25%	8.31
Q2-50%	13.53
Q3-75%	18.68
max	29.99

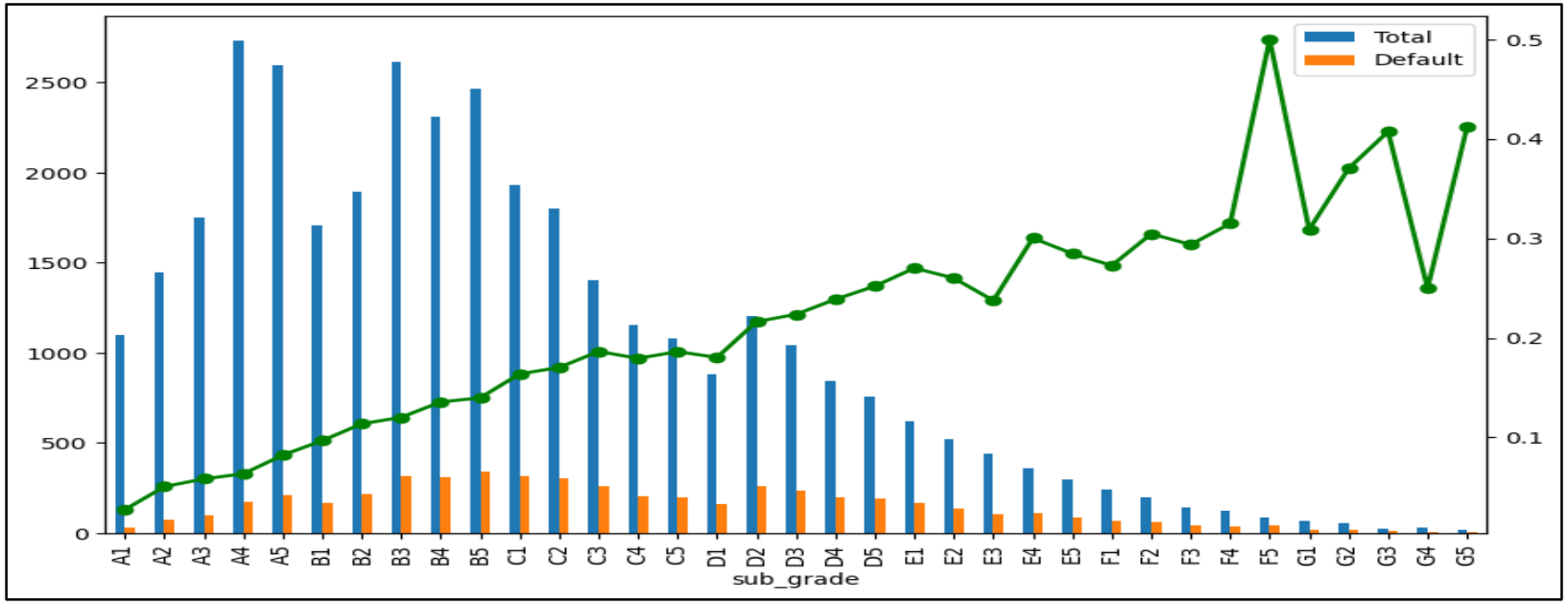


Univariate Analysis - Categorical

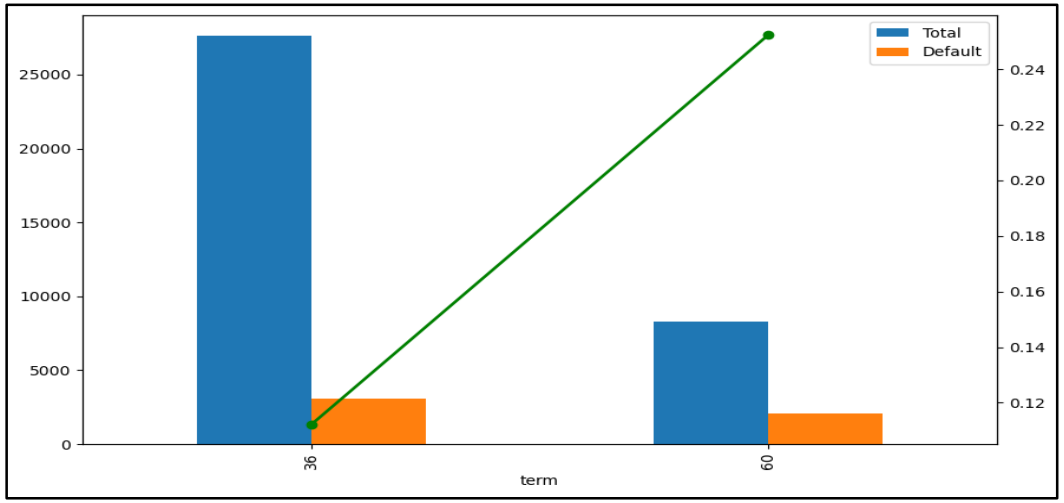


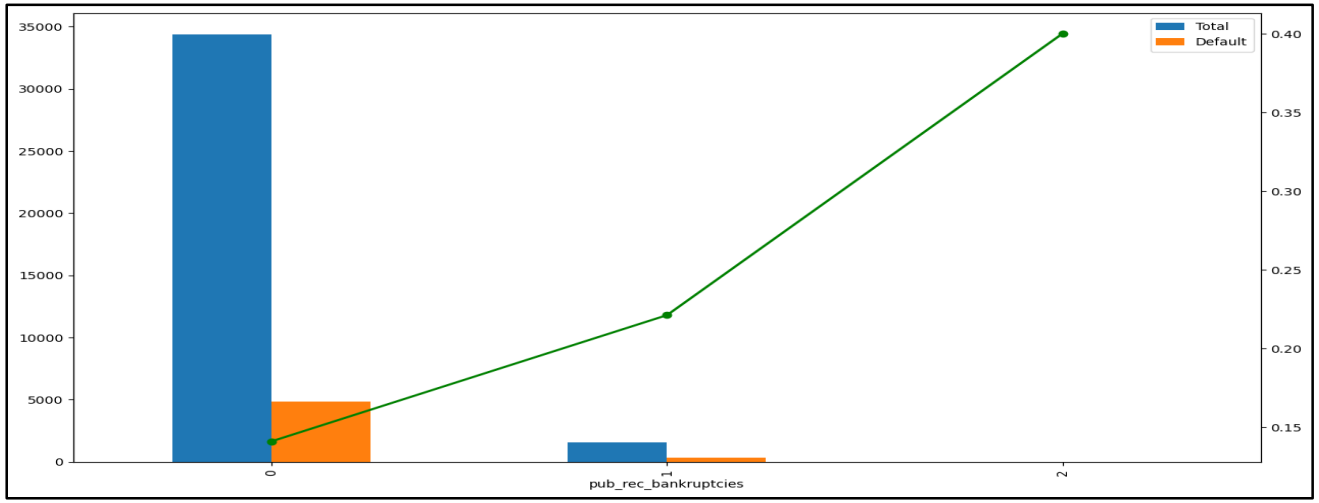


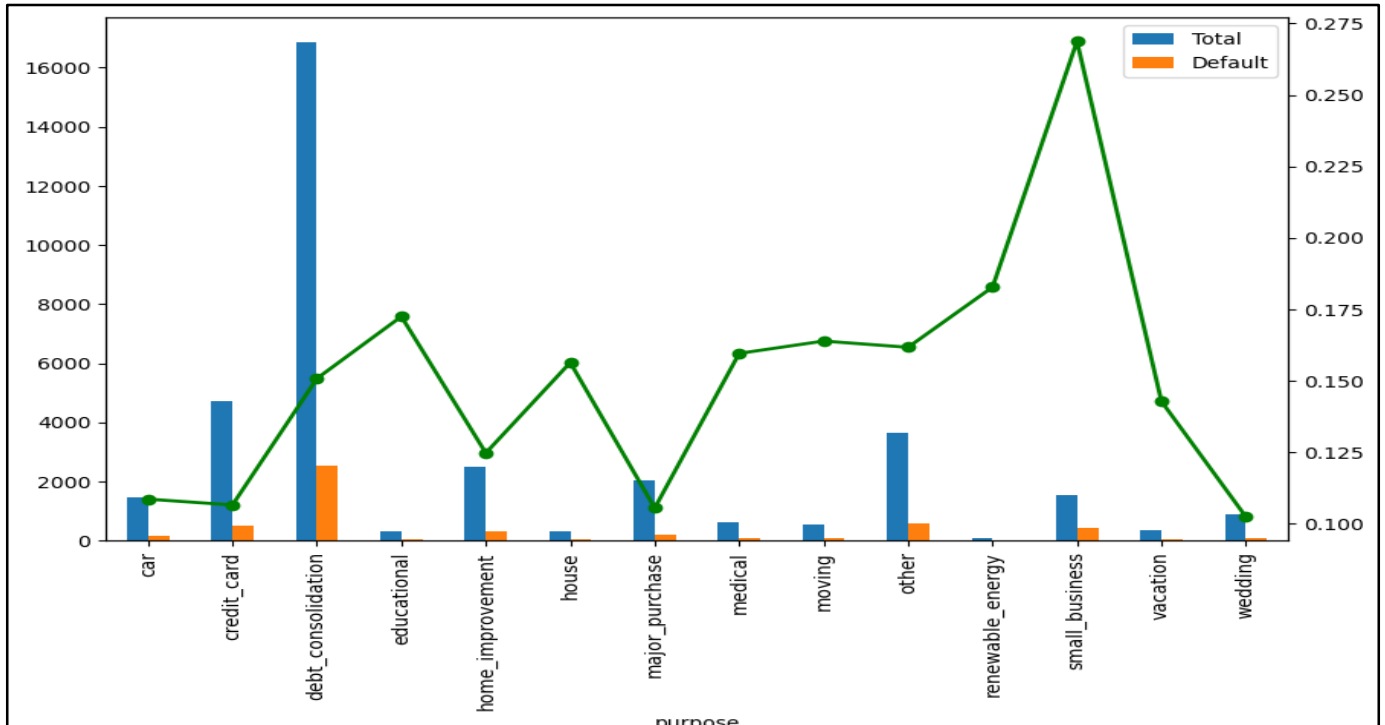




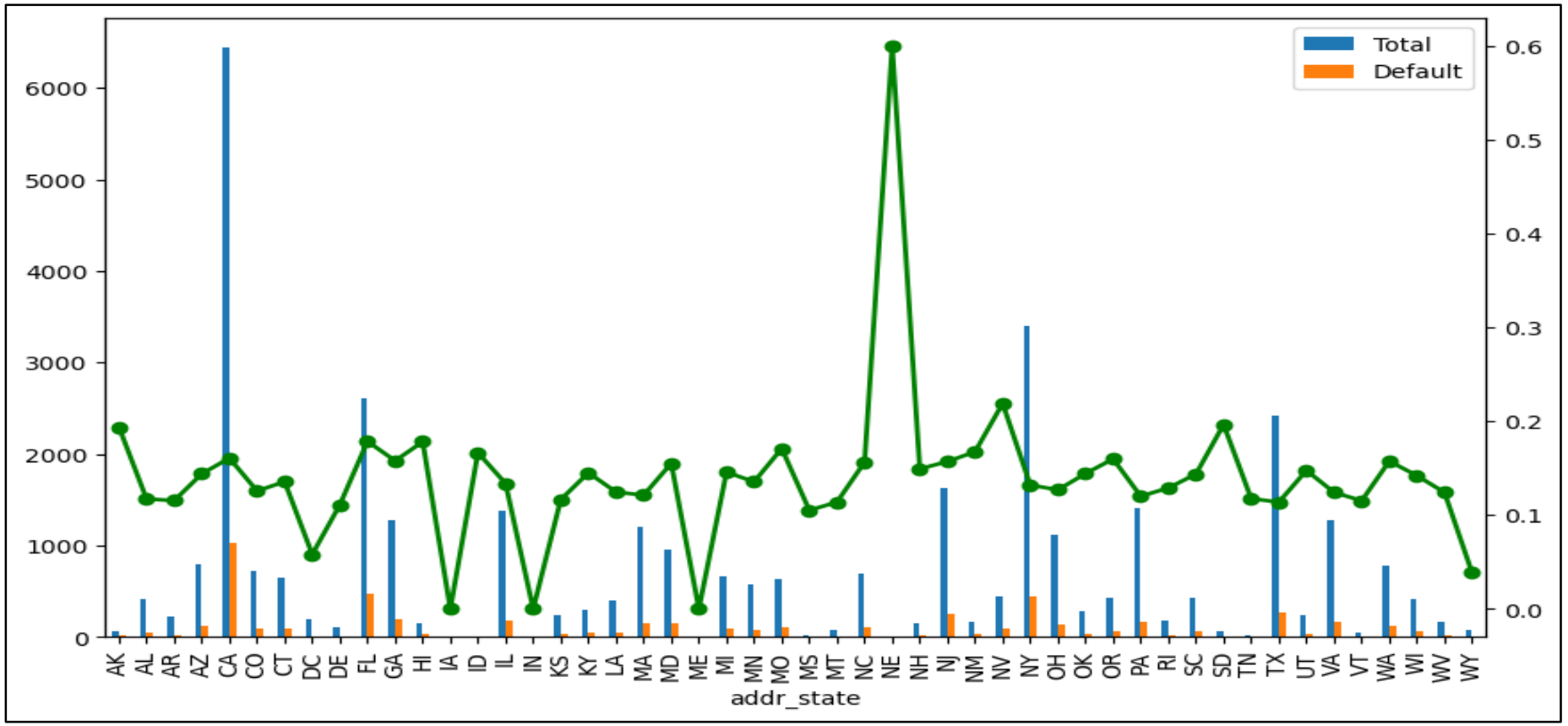
Univariate Analysis - Categorical







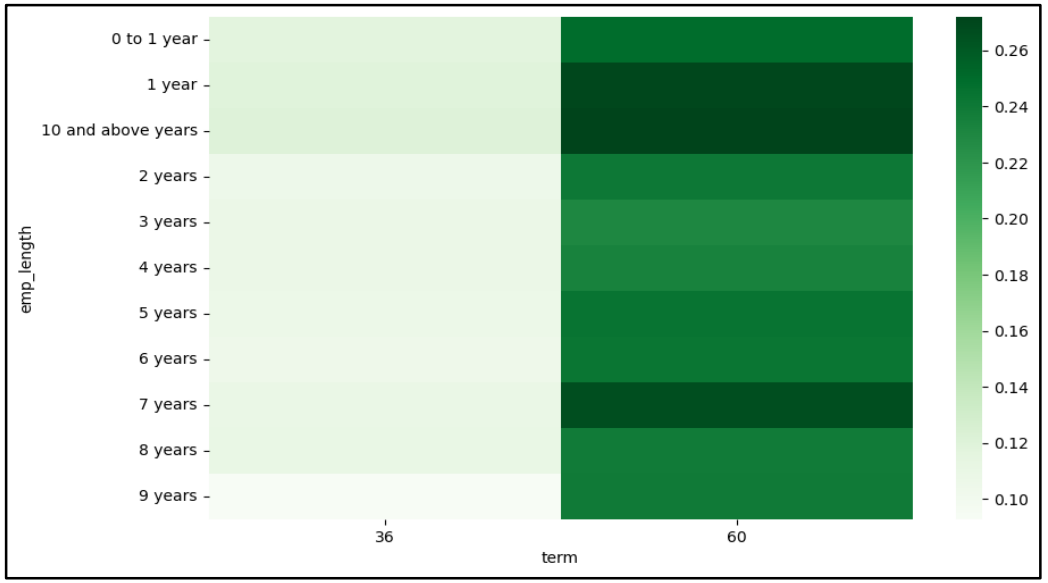
Univariate Analysis - Categorical

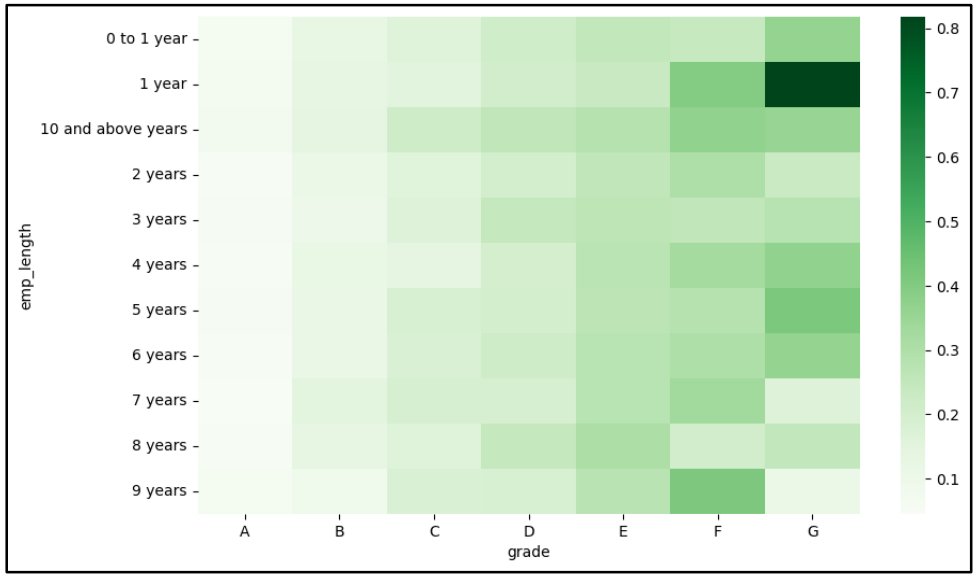


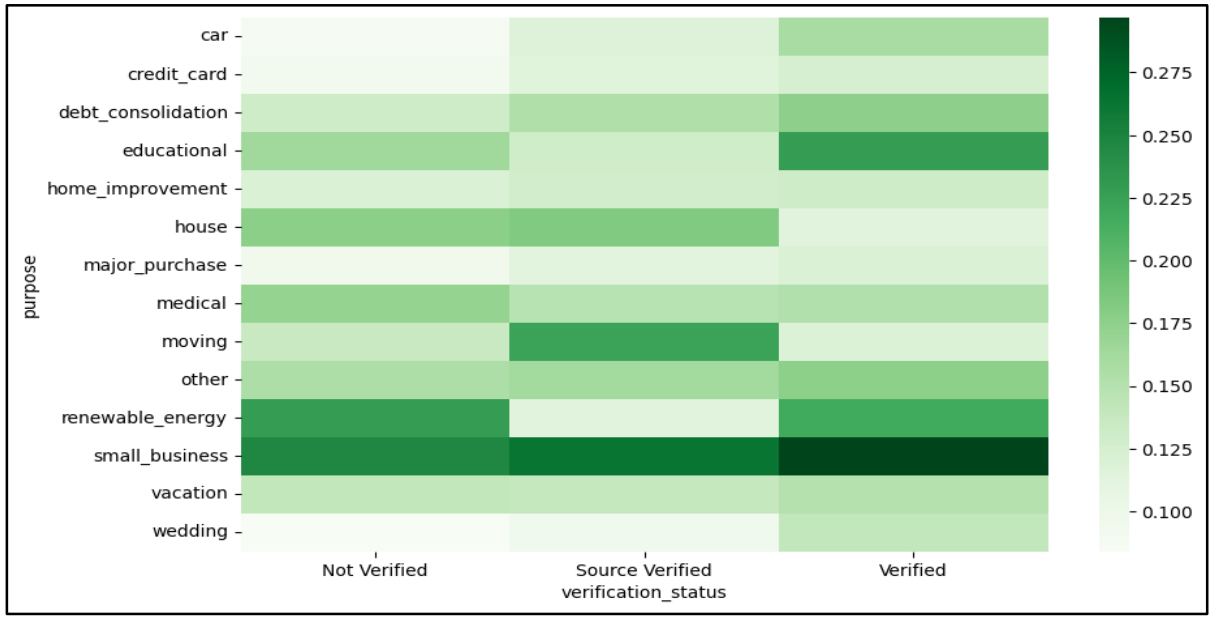
Univariate analysis– Categorical results

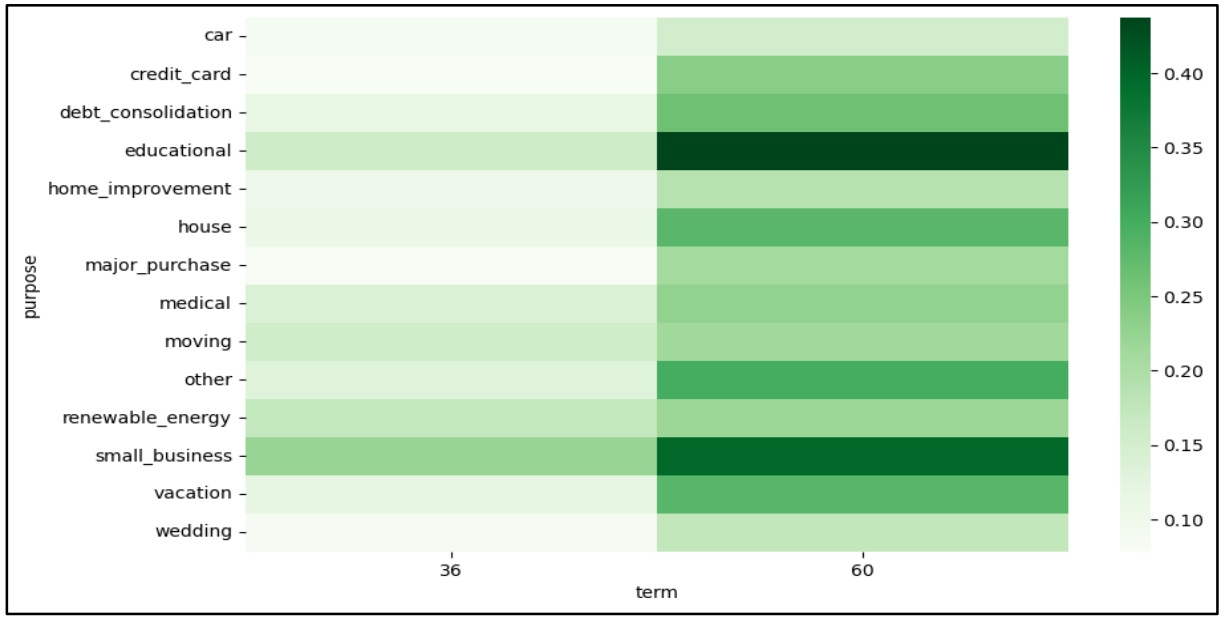
- Out of the entire data default rate is 5196/35945 = 14.4% and in the original data it is 5627/39717=14.2%
- Grades E,F,G have the highest default rates of 27 %,32% and 34%, although these categories get the lesser number of defaults as compared to Grade B which have 26% of overall defaults.
- Highest default rates are for sub grades F5 and G5.
- Experience of 10 yrs and above have the highest default rate of 16 % and contribute to 27% of total defaults across all categories.
- Purpose of small business have the highest default rate of 27%, followed by renewable energy and educational loans,
- Term of 60 months has the highest default rate of 25% and contribute to 40 % of total defaults across the data
- Public bankruptcy with just 1 record increases the chances of defaulting substantially. Pub rec bankruptcy 1 has default rate of 22% we don’t have sufficient enough data for 2 which still shows default rate of 40%.

Bivariate Analysis – categorical to categorical

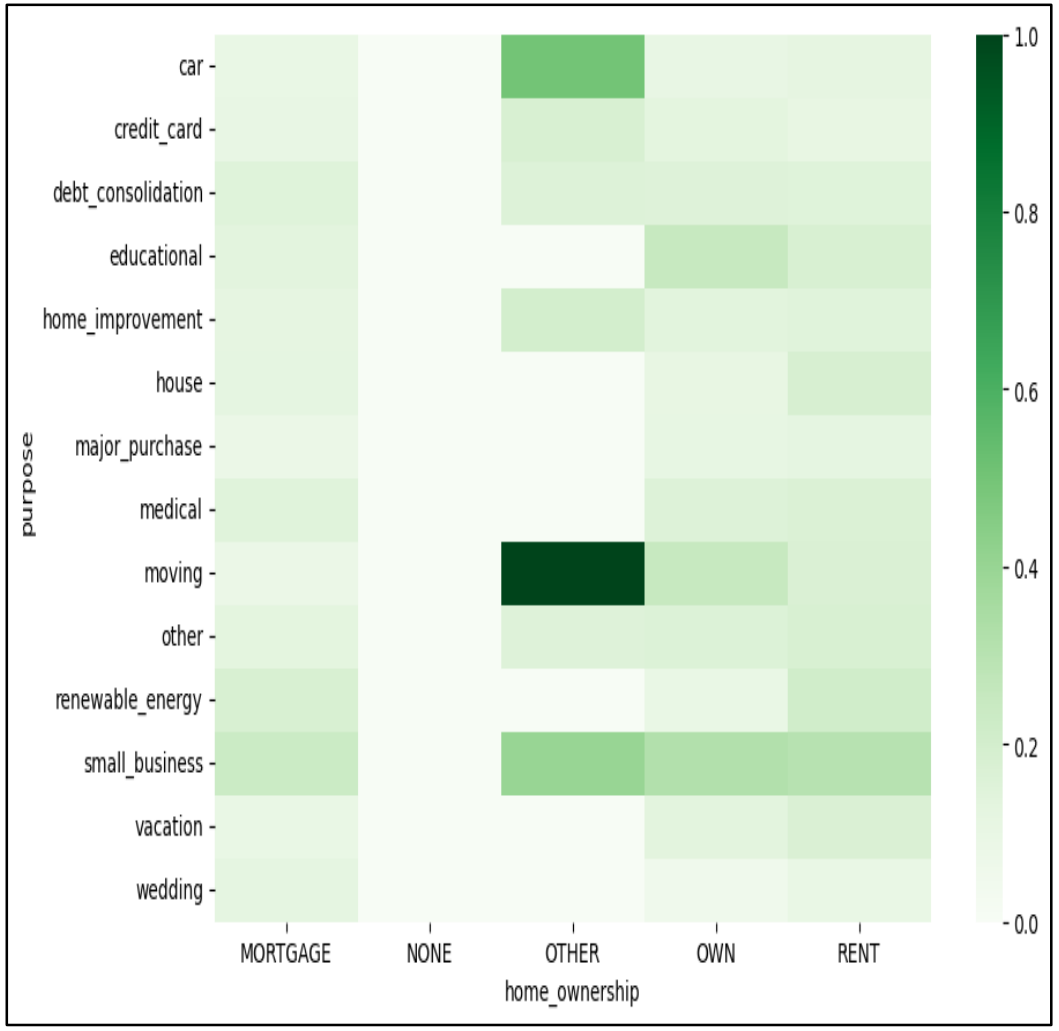


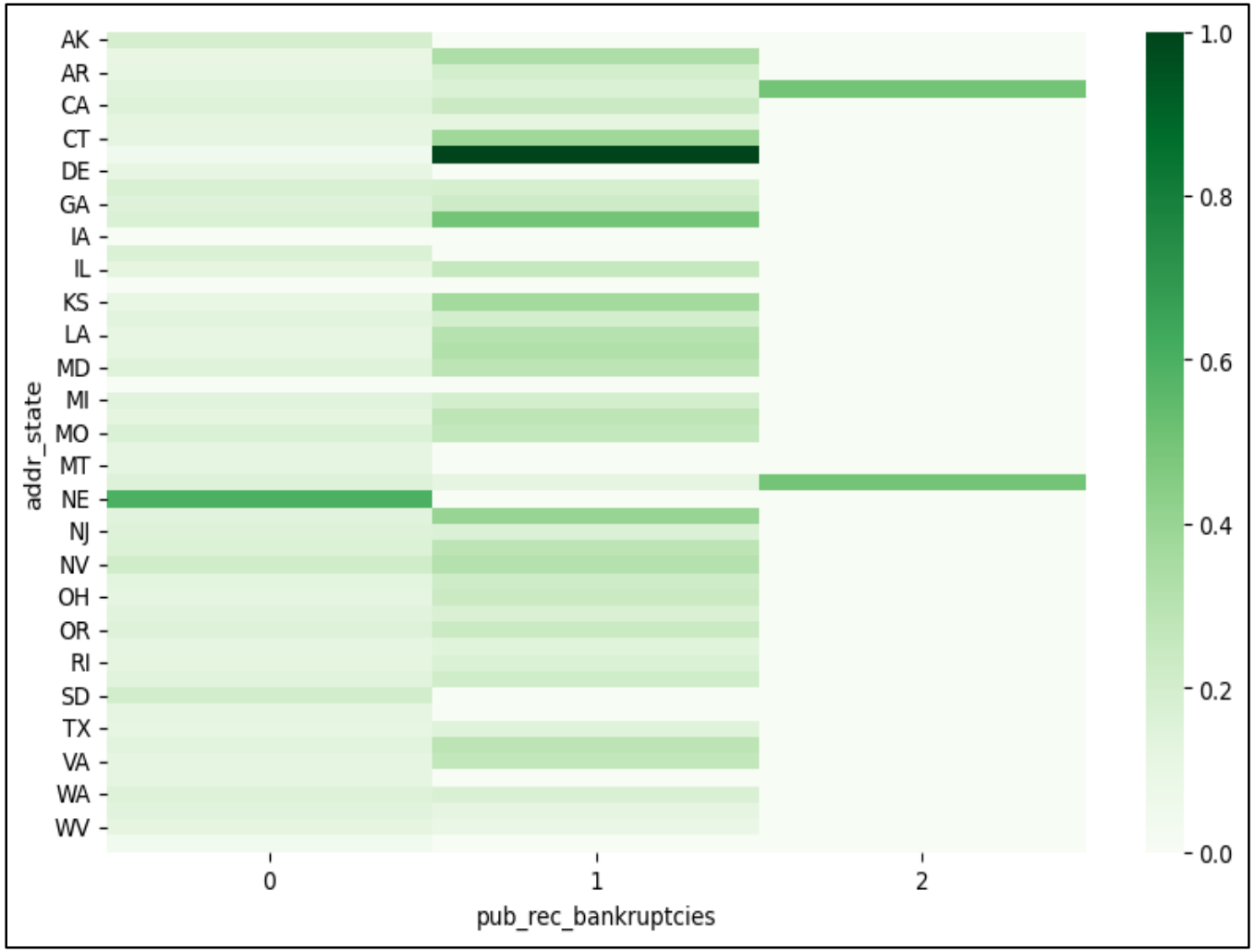




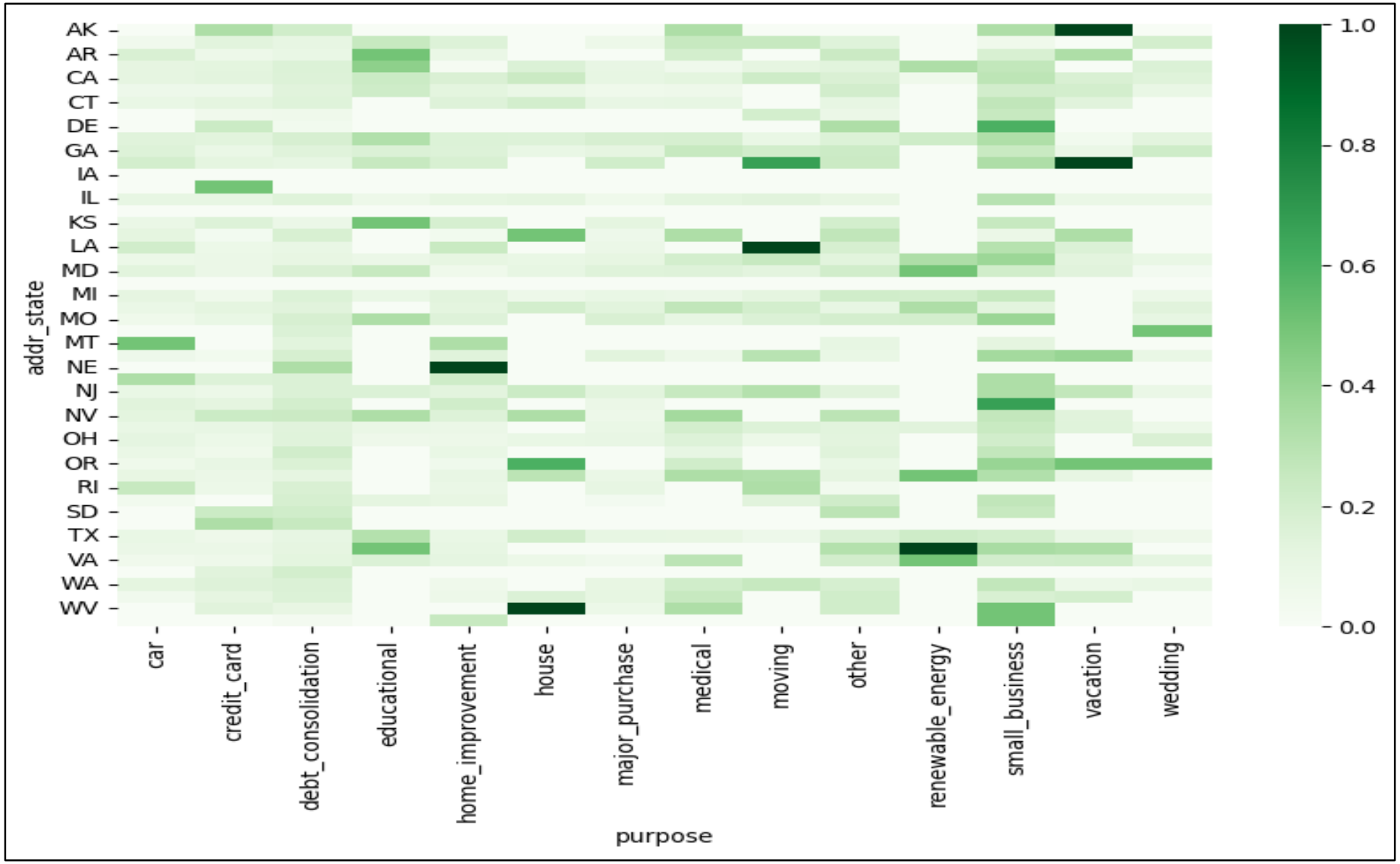


Bivariate Analysis - categorical to categorical





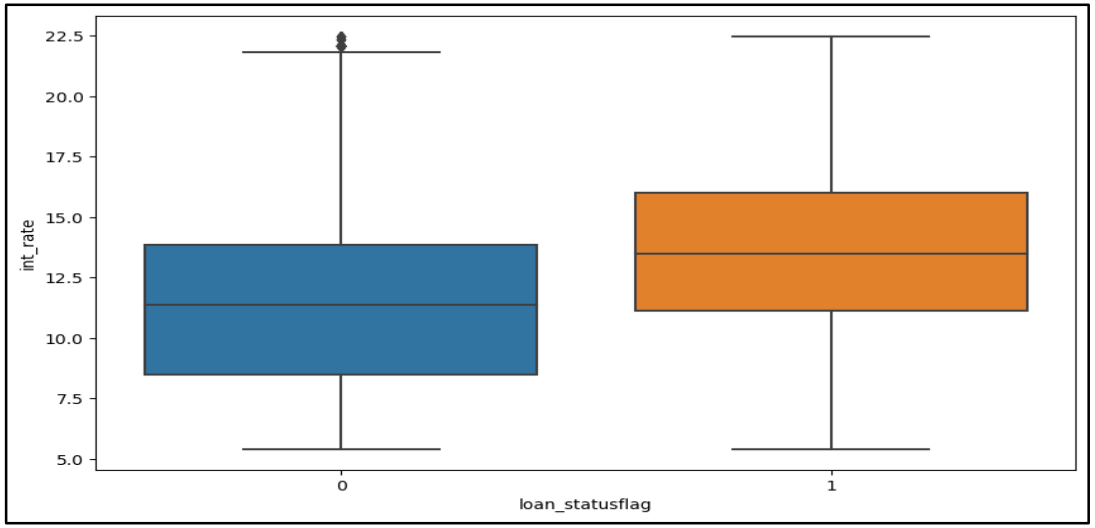
Bivariate Analysis - categorical to categorical

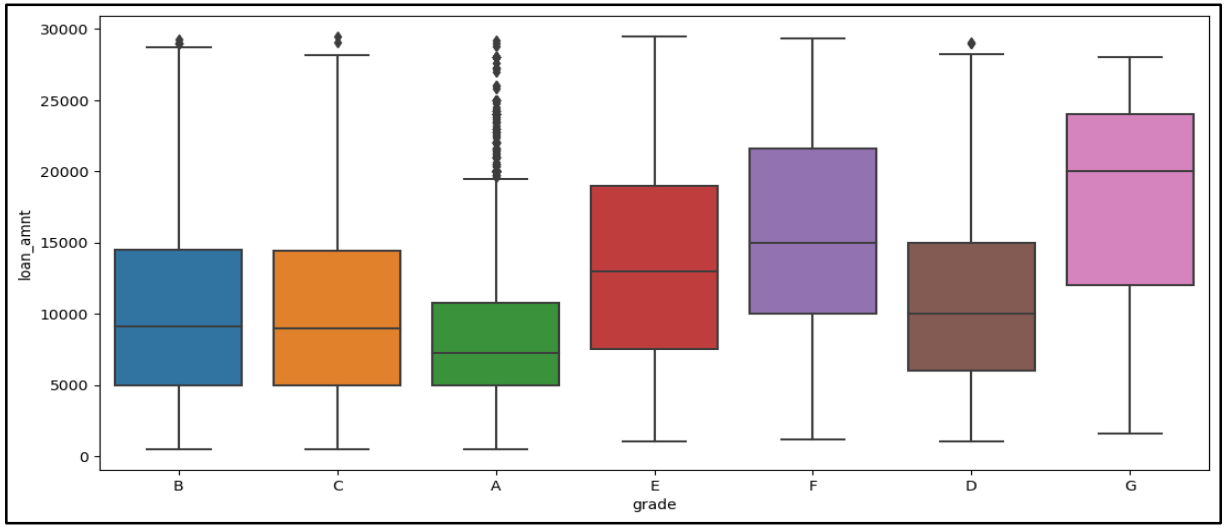


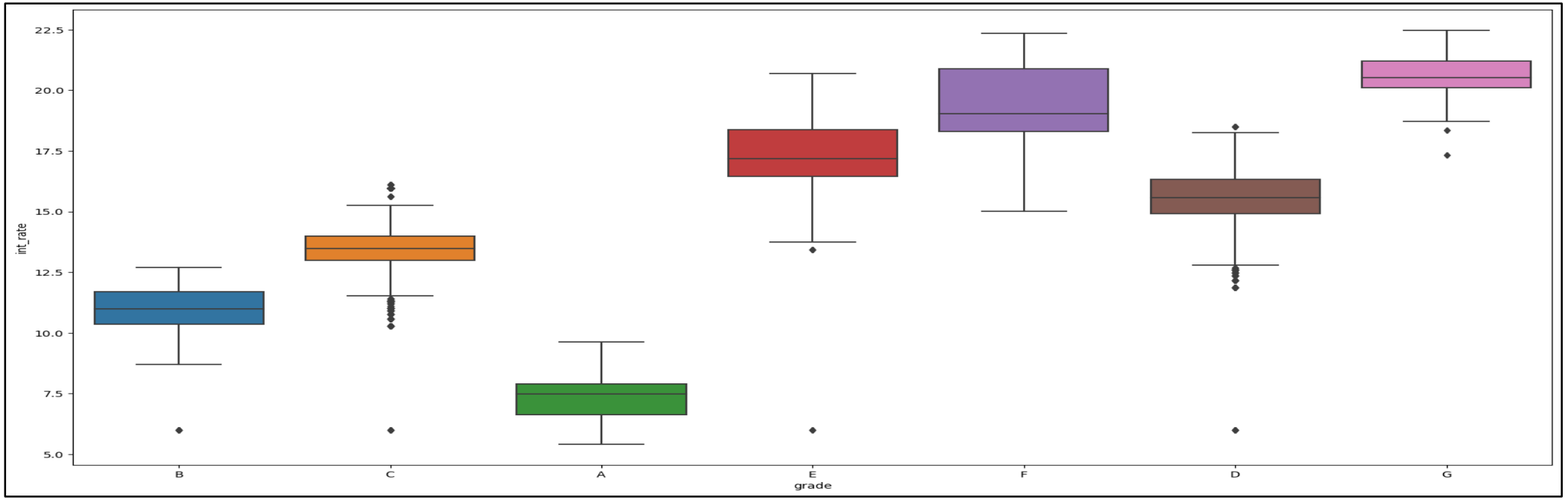
Bivariate Analysis Results – categorical to categorical

- 10 yrs and above exp customers have default rate of 27% for a term of 60 months whereas for a term of 36 it is less than half at default rate ~ 12%. So experience of 1 yr and 10 yrs with term of 60 months are at maximum default rate.
- Education and small business of term of 60 months have highest default rate of more than 40%. For a term of 36 month small business and debt consolidation default rate are at 0.22 and 0.11 which is roughly half of their values at 60 months term. Car loan of 60 month is at less default rate than renewable, educational, moving loan of any terms.
- Small business loans of any verified status is at highest default rate above 25%. Renewable energy source verified is half default rate of not verified.
- In states like CA, NY, FL,TX most of the loans are of purpose type Debt consolidation, credit and Other.
- Grade F loans to customers with exp of 10 yrs and above are at more default rate than G grade to the same type of customer.
- Grade F loans to customer with exp of 1 year are at roughly same default rate of 40% which increases to 82% for grade G.
- Pub rec of 1 bankruptcy is at roughly same default rate across NY, CA and FL which have largest contribution of loans in data. We can do a further analysis to which purpose type of loans have the highest default rate for pub rec of 1 bankruptcy.

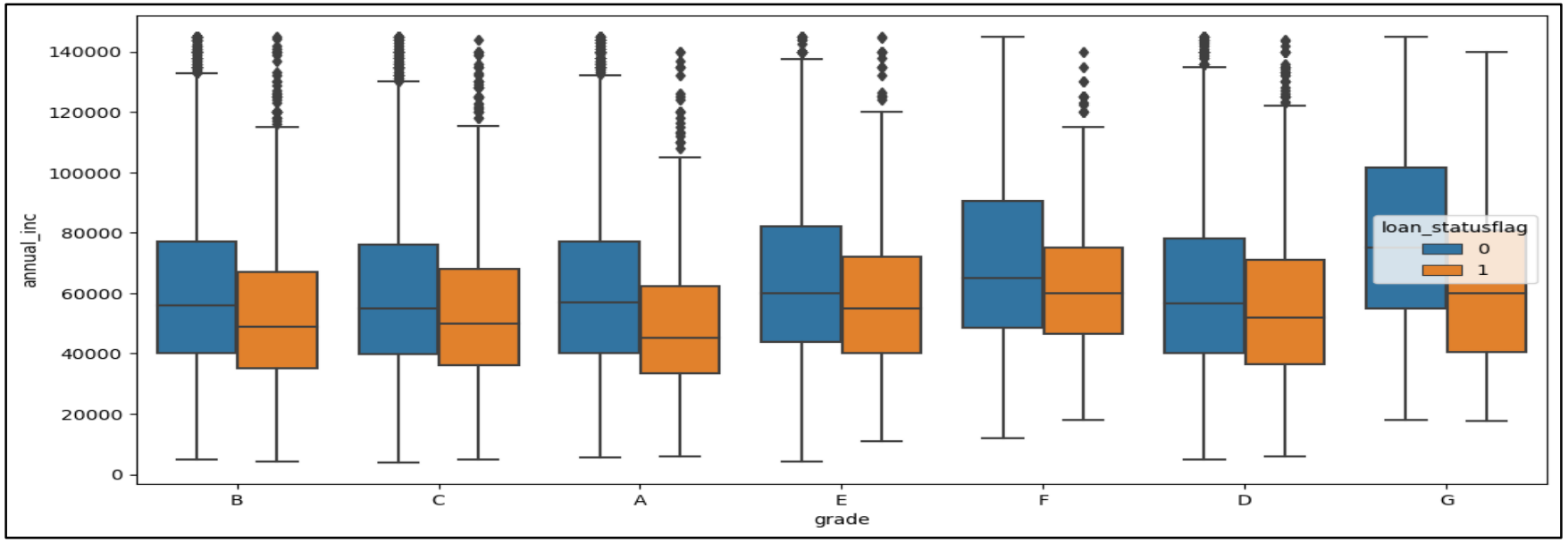
Bivariate Analysis – continuous to categorical

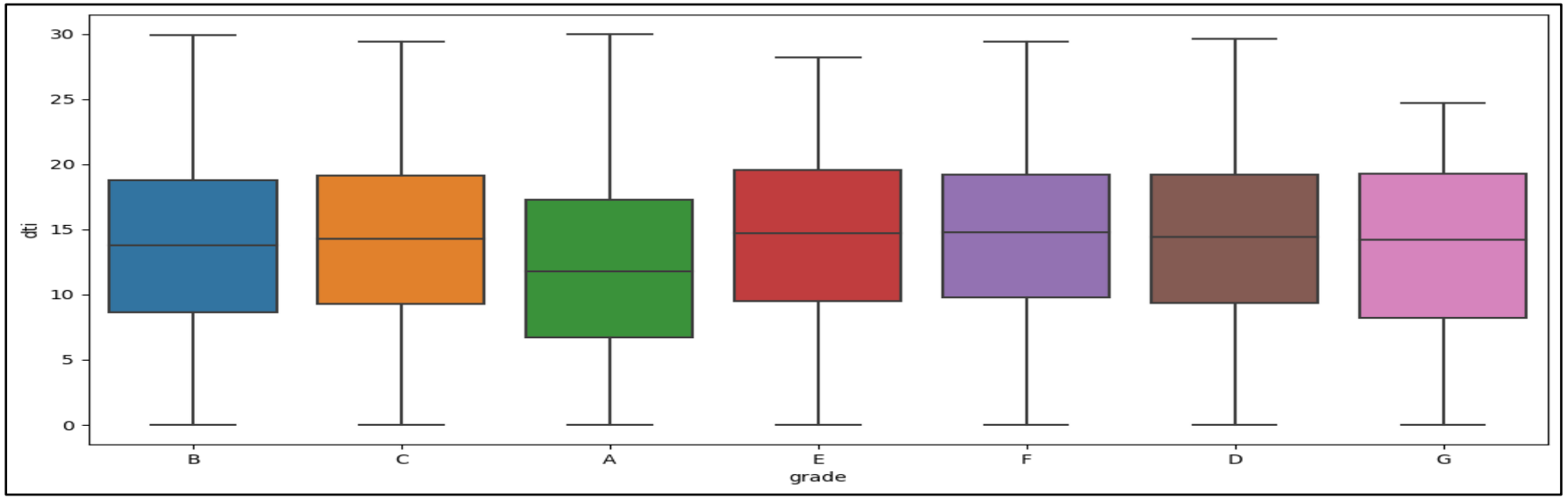




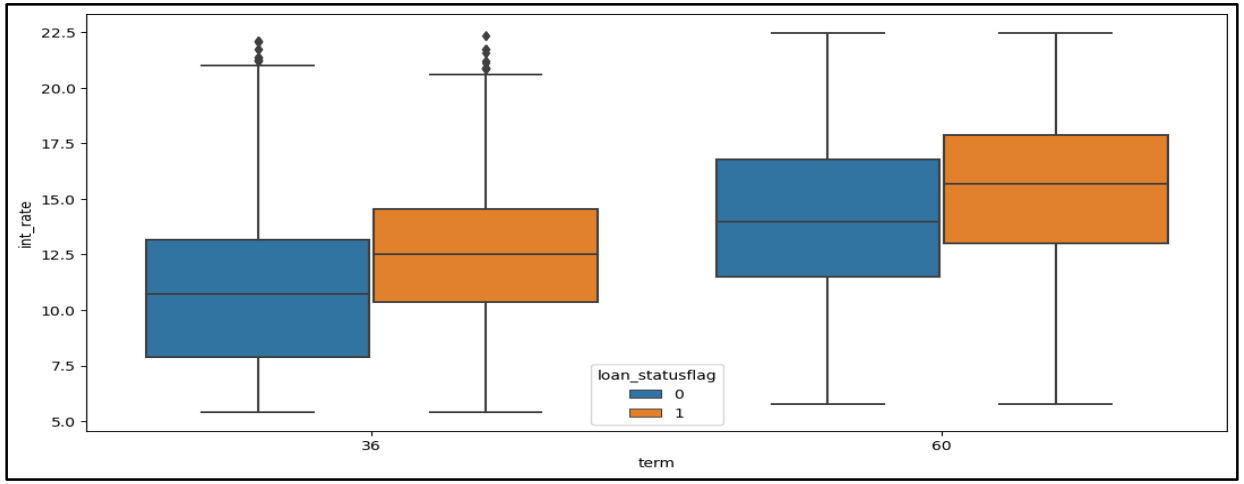


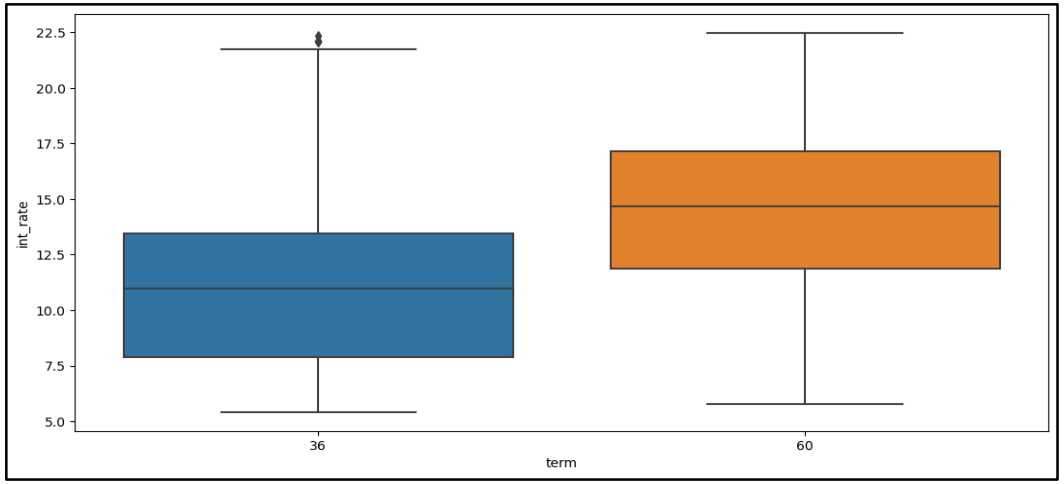
Bivariate Analysis - continuous to categorical

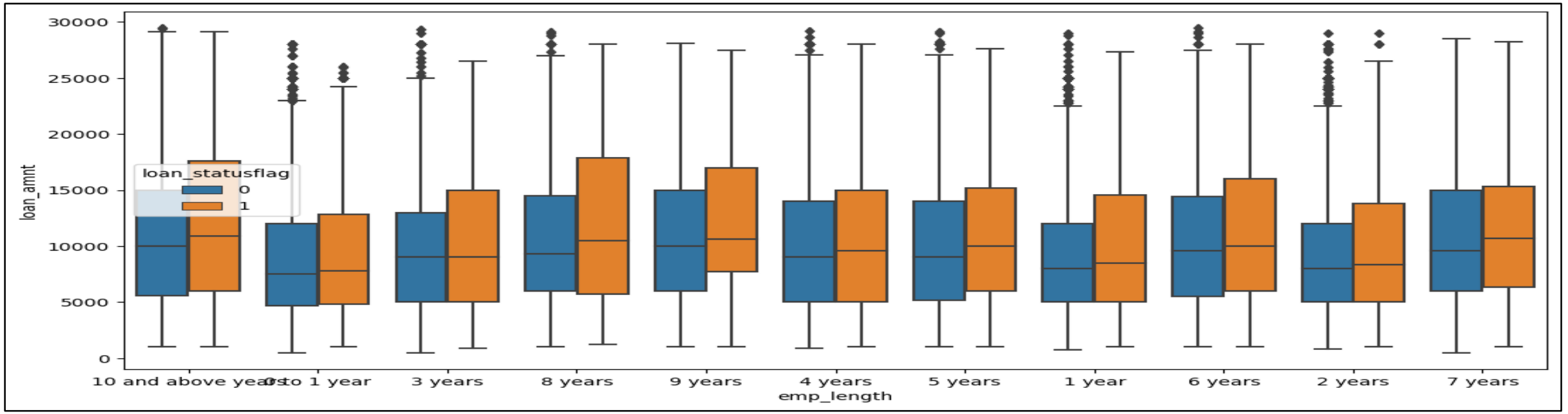




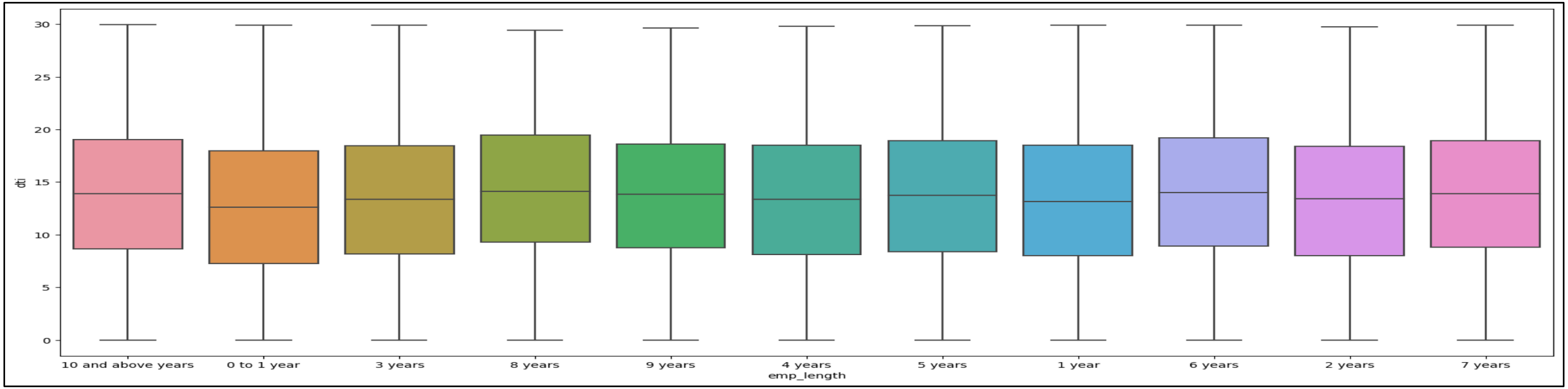
Bivariate Analysis - continuous to categorical

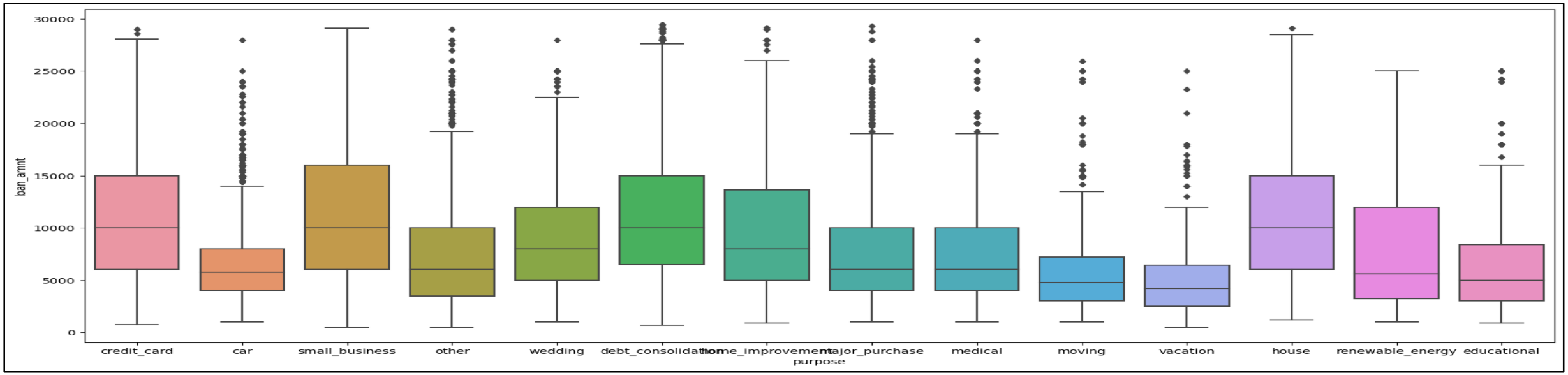




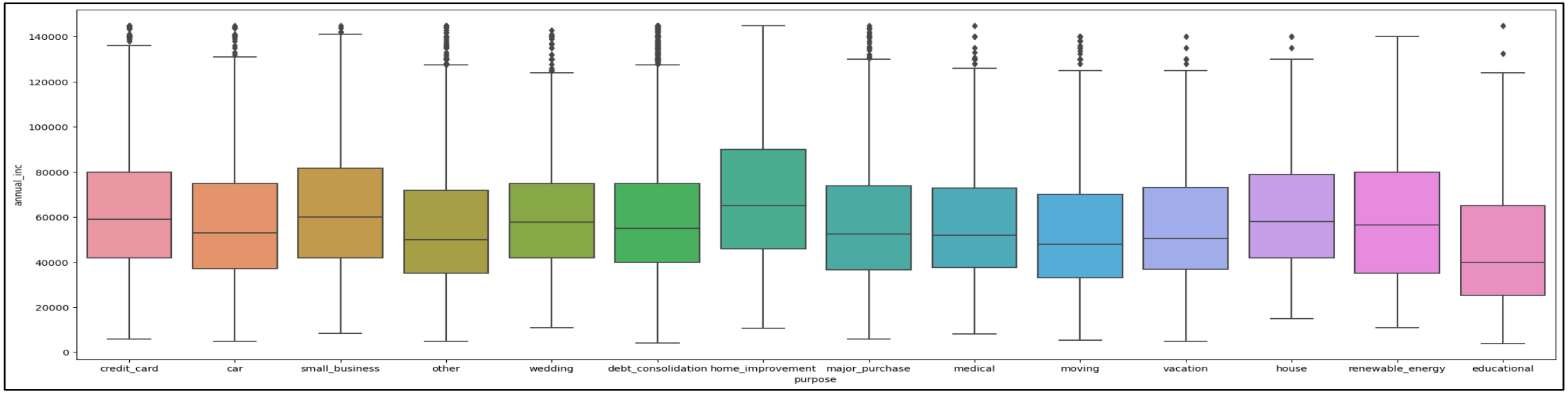


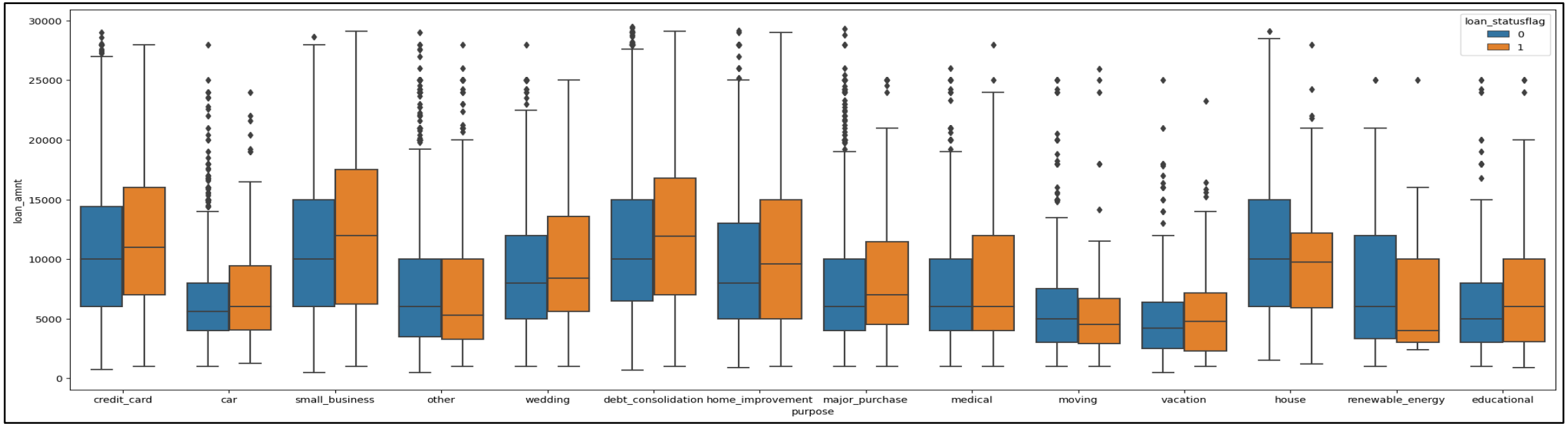
Bivariate Analysis - continuous to categorical



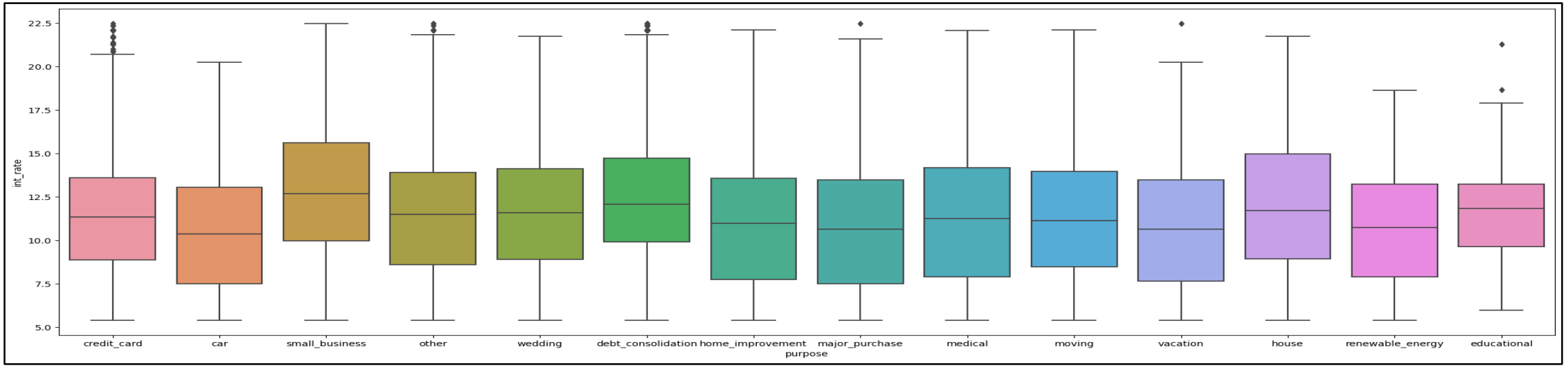


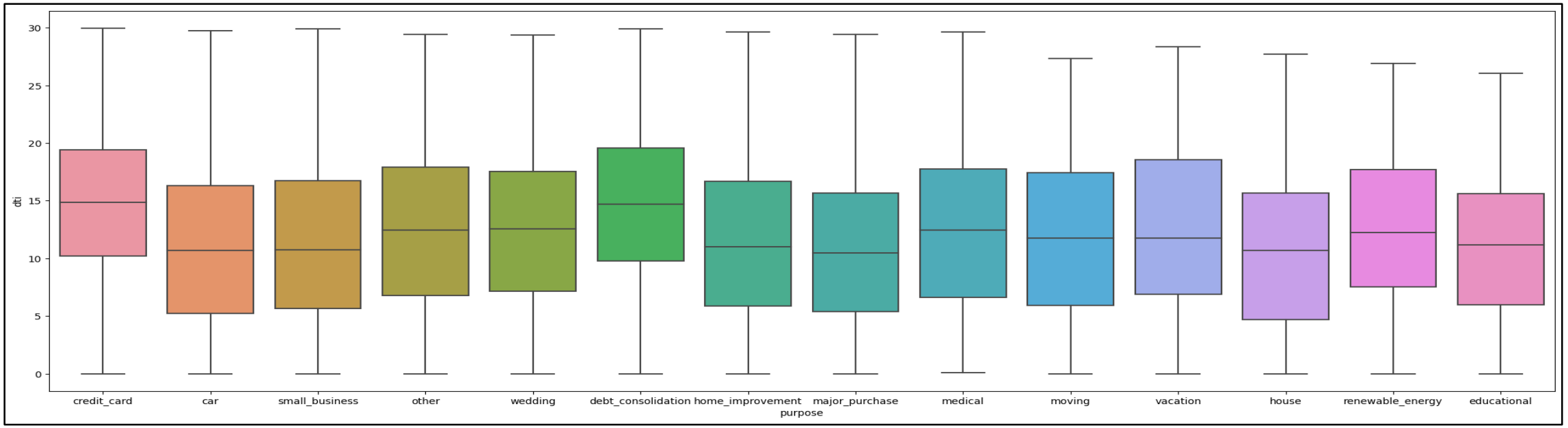
Bivariate Analysis - continuous to categorical





Bivariate Analysis - continuous to categorical





Bivariate Analysis Results- continuous to categorical

- Loans of grade E,F and G shows high interest rate value, their median loan amounts are highest as well and have similar debt to income ratio. Hence they are very likely to default.
- Loans of term 60 months carry a high interest rate and also have interest rate above 13% and are likely to default.
- For customers with experience of 7,8,9,10 years and above a high loan amount above 10K is very likely to default. For customer with experience of 7,8,9,10 years and above the median debt to income ratio are highest. For any experience employee interest rate above 13 % is likely to default.
- For credit and debt consolidation purpose loans the dti is likely to be high as seen in data as compared to small business.

However small business have roughly higher median income than the credit and debt consolidation purpose loans. Also small business loans are less likely to default if it is below 10K. Hence we can put a upper cap on small business and reduce the int_rate. Currently the median interest rate for small business is 12.5 % which makes them likely to default. If we reduce int_rate for these loans to 10% it will default lesser.

Summary

- Variables like [Grade, Employment experience, Public record of bankruptcy, Purpose, term and interest rates] are strong indicator of default.
- If public record of bankruptcy increases then chances of defaults are high, if grade is high in alphabetical order or employment experience is 10yrs or above the chances to default increases. Loan Purpose of small business is a high default rate loan type. High interest rate and term length of 60 months have large default chance.
- 10 years and above experience customers have default rate of 27% for a term of 60 months whereas for a term of 36 it is less than half at default rate ~ 12%. Which is below default rate of other 60 months loans for other employment experience.
- For a term of 36 month small business and debt consolidation default rate are at 0.22 and 0.11 which is roughly half of their values at 60 months term. Car loan of 60 month is at less default rate than renewable, educational, moving loan of any terms.
- Renewable energy source verified is at half default rate of not verified.
- In states like CA, NY, FL,TX most of the loans are of purpose type Debt consolidation, credit and Other.
- For employment experience of 7,8,9,10 years and above a high loan amount above 10K is very likely to default. For customer with experience of 7,8,9,10 years and above the median debt to income ratio are highest. For any experience employee int rate above 13 % is likely to default.
- Small business have roughly higher median income than the credit and debt consolidation purpose loans. Also small business loans are less likely to default if it is below 10K. Hence we can put a upper cap on small business and reduce the int_rate. Currently the median interest rate for small business is 12.5 % which makes them likely to default. If we reduce int_rate for these loans to 10% it will default lesser.

Recommendations

- 10 years and above experience customers have default rate of 27% for a term of 60 months whereas for a term of 36 months it is less than half at default rate ~ 12%. Which is below default rate of other 60 months loans for other employment experience.

Hence we can reduce 60 months loans to customers with experience 10 years and above and increase loans with term 36 months to these customers.

- For a term of 36 months small business and debt consolidation default rate are at 0.22 and 0.11 which is roughly half of their values at 60 months term. Car loan of 60 month is at less default rate than renewable, educational, moving loan of any terms.

Hence we can reduce 60 months loans to small businesses and debt consolidation loan types and increase 36 months loans of same purposes. Also since car loans are less defaulting than renewable and educational we can increase more of them.

- Renewable energy source verified loans is at half default rate of not verified.

Hence we should allow such loan types to be source verified only as it carries less chance of credit loss.

- Small business have roughly higher median income than the credit and debt consolidation purpose loans. Also small business loans are less likely to default if it is below 10K.

Hence we can put a upper cap on small business and reduce the int_rate. Currently the median interest rate for small business is 12.5 % which makes them likely to default. If we reduce int_rate for these loans to 10% it will have less chance of credit loss.

Conclusions

We have conducted EDA on the dataset. We have understood the columns of the dataset and perform data cleaning and manipulation. Then we have added data new column in the dataset for better analysis and performed univariate and bivariate analysis. We have successfully completed EDA on our dataset and have published results in Summary. The recommendations given are suggestions with help of which we can increase the business and decrease the chance of credit loss.

Thank you.