Lending club case study

Problem statement: Help the business to understand the driving factors behind the loan default i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk management.

Process adopted

- Data understanding
- Data cleaning
- Data analysis
- Recommendation

Data Understanding :

Analyzed the column and selected the feature which is required.

Divided the columns as per the data into below division

Numerical columns : loan_amnt,funded_amnt,int_rate,installment,emp_length,annual_inc,dti

Interval data (time duration): issue_d, term

Nominal : home ownership, verification, status, loan_status, purpose,addr_state

Ordinal : grade, subgrade, pub_rec_bankruptcies

Cleaning of the data

```
steps adopted for cleaning
```

1 Identifying the columns whose null values are 60% and dropping them
2 Identifying duplicate records, the dataset didn't contain any duplicate records
3 Dropping the columns which has features that did not add value to the problem statement
4 Filling the nan values of emp_length with 0 years and the other categorical with unknown values
5 Removing the outliers from annual_income, loan_amnt
annual income we took values of till 99 percentile removing the last percentile data

Created derived columns

Two derived columns were created with issue_d column which is of type date
issue_d_years and issue_d_months

For the numerical variable :
loan_amount_grp columns was derived which is the range of loan_amnt column
int_rate_grp columns was derived which is the range of int_rate column
dti_grp columns was derived which is the range of dti column
annual_inc_grp columns was derived which is the range of annual_inc column

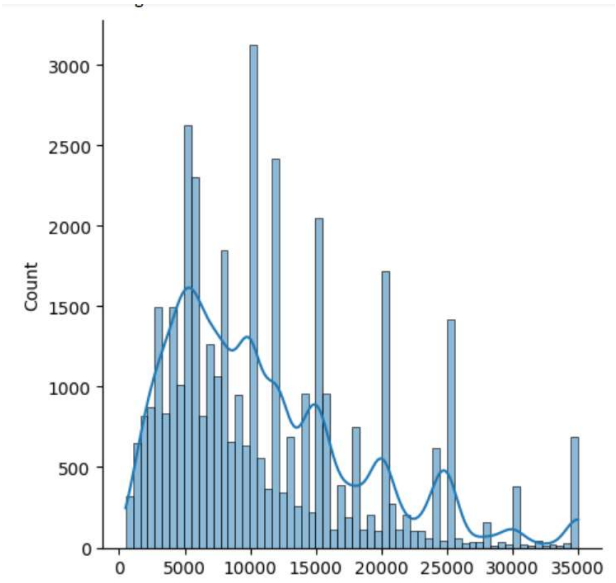
Data analysis:

Univariate analysis

Below are the plots and summarization of the result

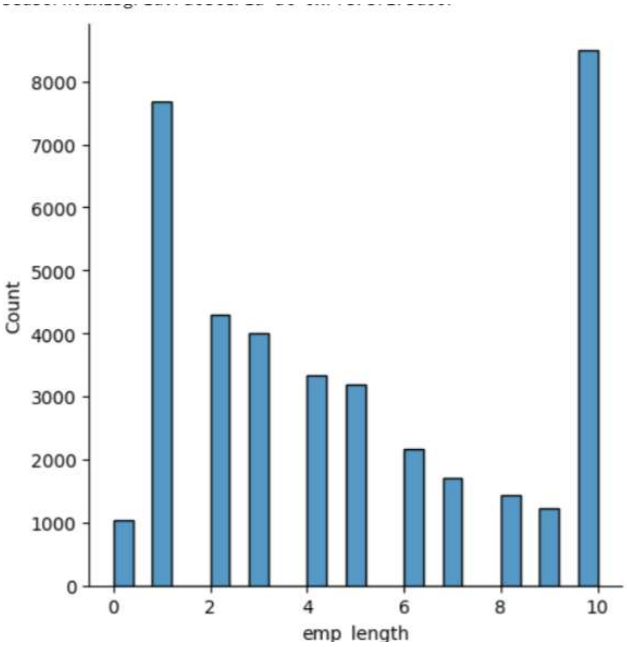
Loan amount vs loan count

Inference : Maximum loan application is in the range 5000- 15000



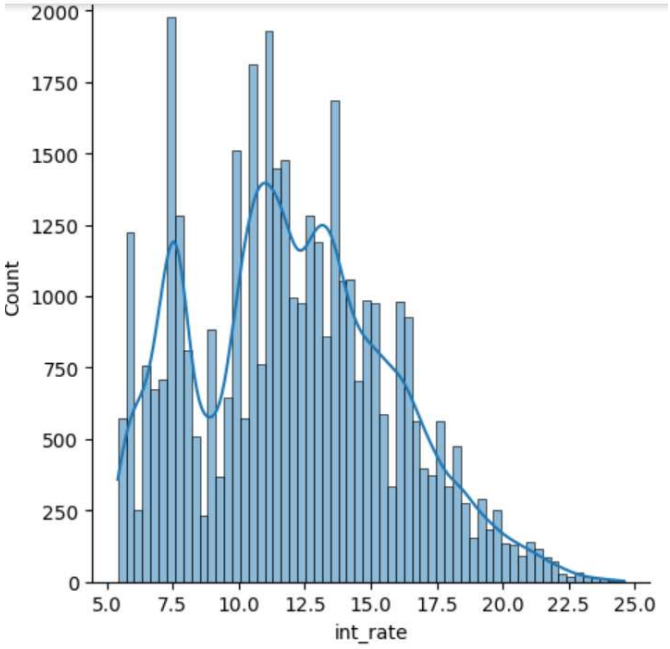
Emp_length vs loan count

Inference: more loan application is from more experience and less of 1 year experience



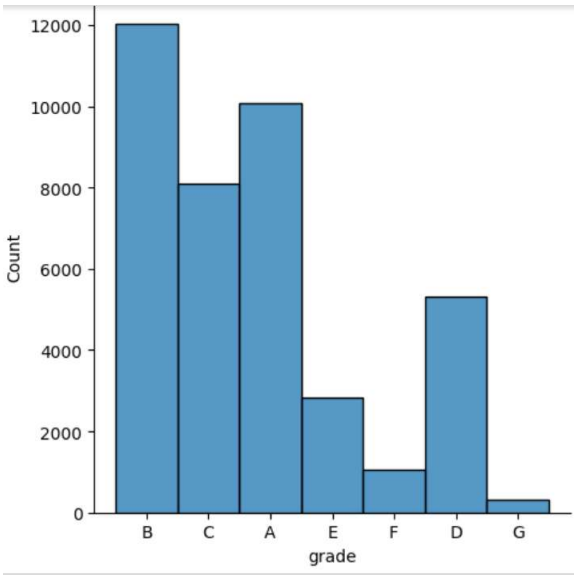
Int_rate vs loan count

Inference: As the interest rate increases loan count decreases



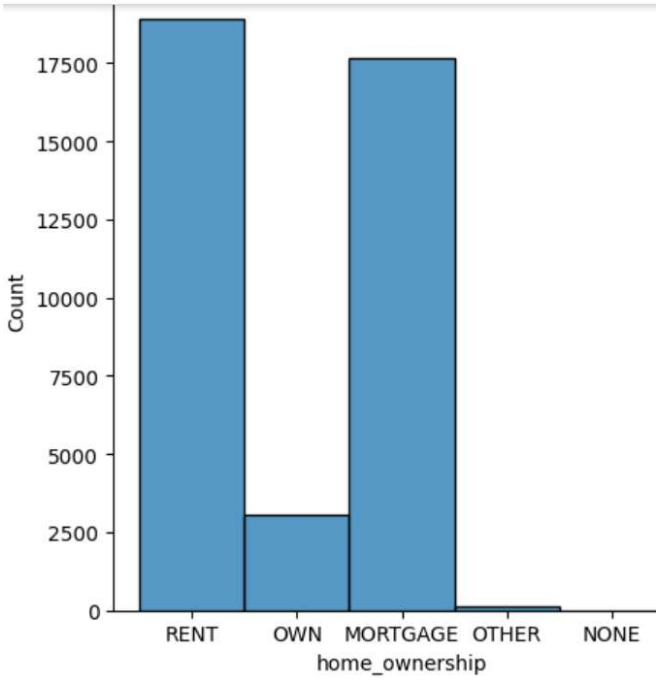
Grade vs loan count

Inference:As the credit grade moves from A to G loan application decreases



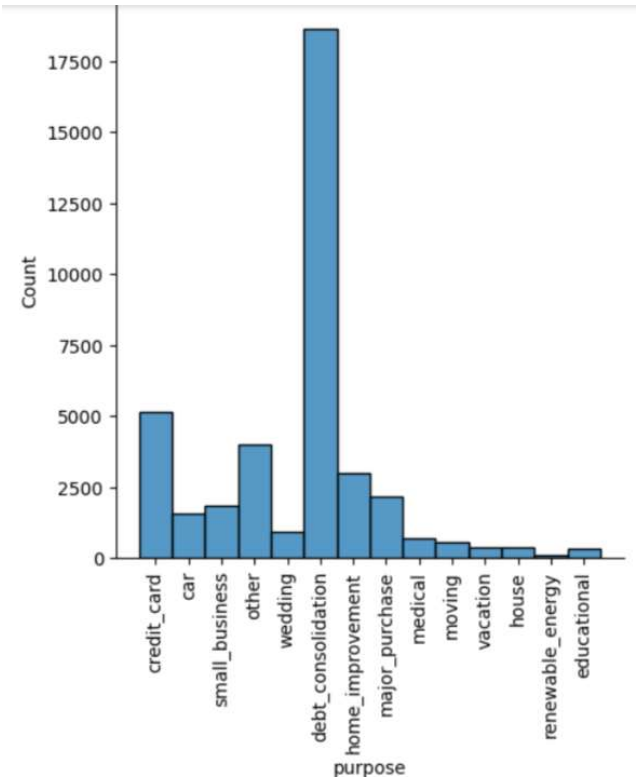
Home ownership vs the loan count

Inference: Rented and Mortgage people take more loan



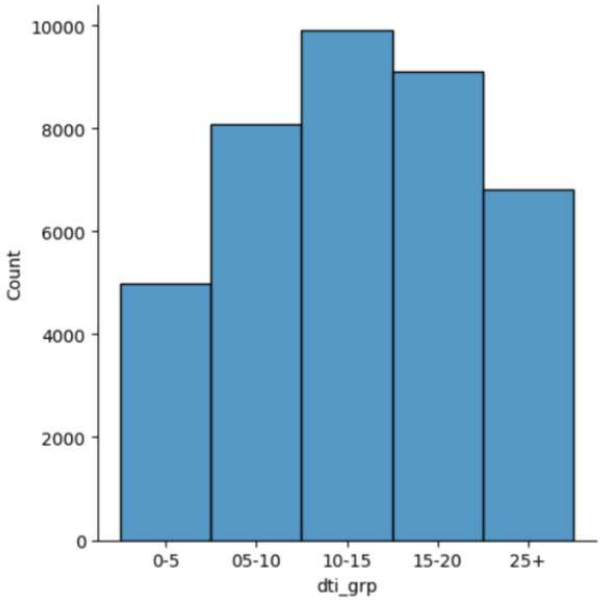
Purpose vs loan application count

Inference : Loan taken is highest for debt_consolidation



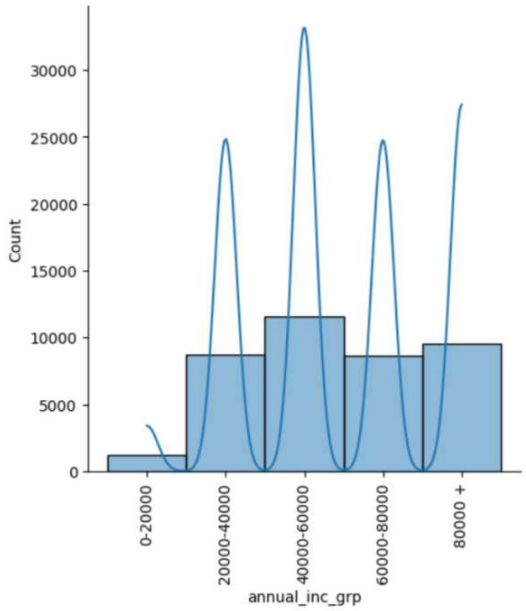
Dti vs loan application count

Inference : dti between 10-15 are the highest loan applicant



Annual_inc_grp vs loan count

Inference : maximum loan taken is between 40000-60000



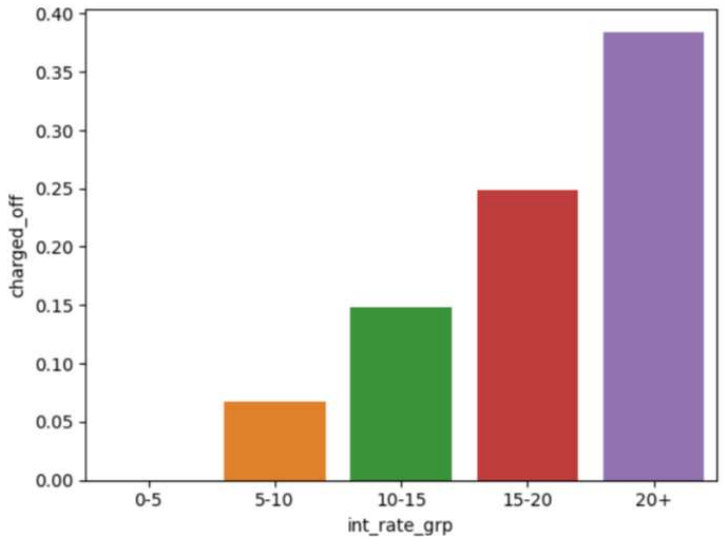
Observation from univariate analysis

- Maximum people have taken loan between 5000 to 15000 and 10000. People have taken loan from 500 to max 35000 loan
- Loan application is more for 36 months than longer duration 60 months
- Number of loans is decreasing as the interest rate increases , maximum loan is in between 10% to 15 %
- Grade B has taken maximum loan
- Maximum loan is from the annual income group 40k - 60 k , 0-20k people have lesser loan application
- Maximum loan is taken for debt_consolidation
- 14% of total loan is charged off
- Maximum loan taken is in the range of 5000 to 15000.
- Loan applicant are more from Rented and Mortgaged category
- 14% loans were charged off out of total loan issued

Bivariate analysis

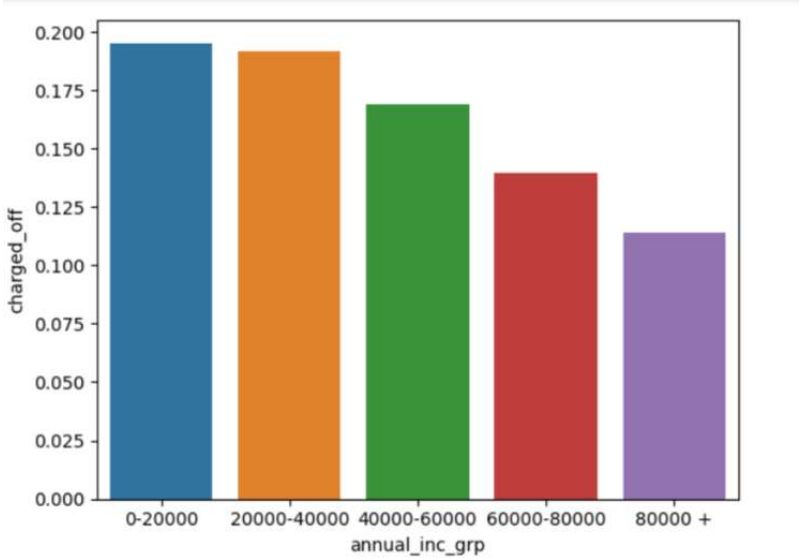
Int_rate_grp vs charged off proportion

Inference: Higher interest rate cause more default



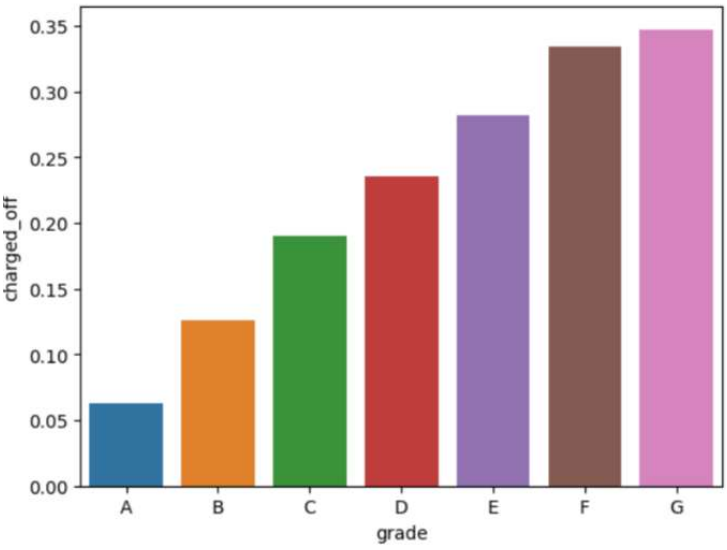
Annual_inc_grp vs charged off

Inference:Lesser the annual income more the default



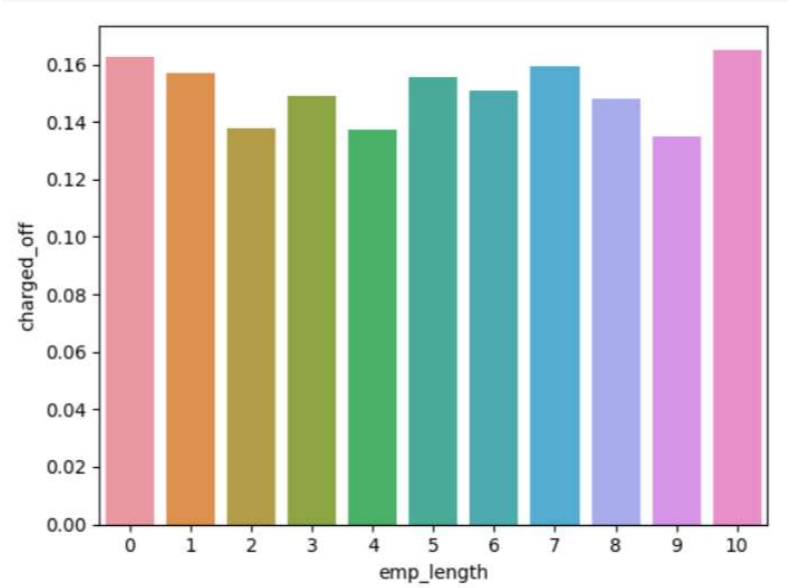
Grade vs Charged off

Inference:As the credit grade move from A to G default increases



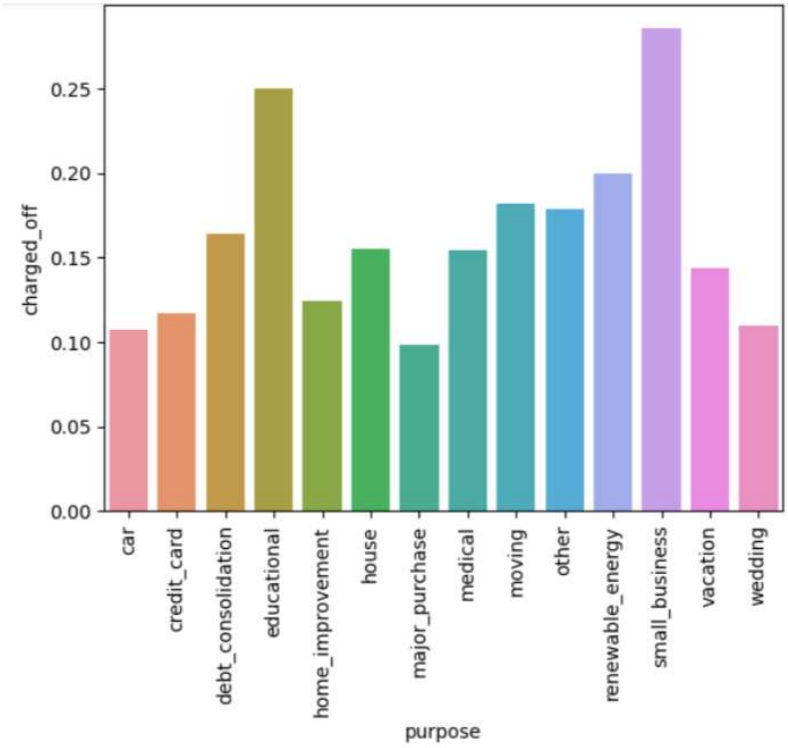
Emp_length vs charged off

Inference: People with 1 and less year are defaulting more and then the 10+ years



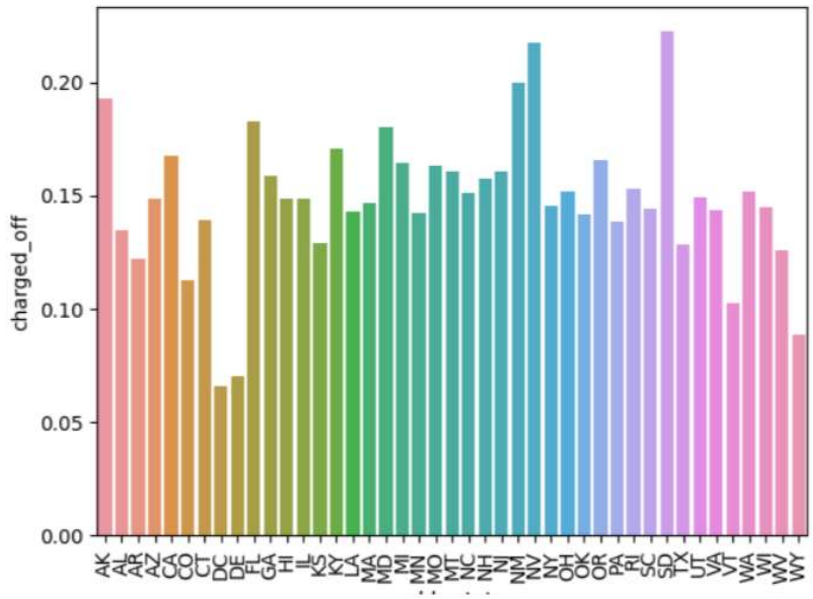
Purpose vs Charged off

Inference: Loan taken for small business are defaulting more



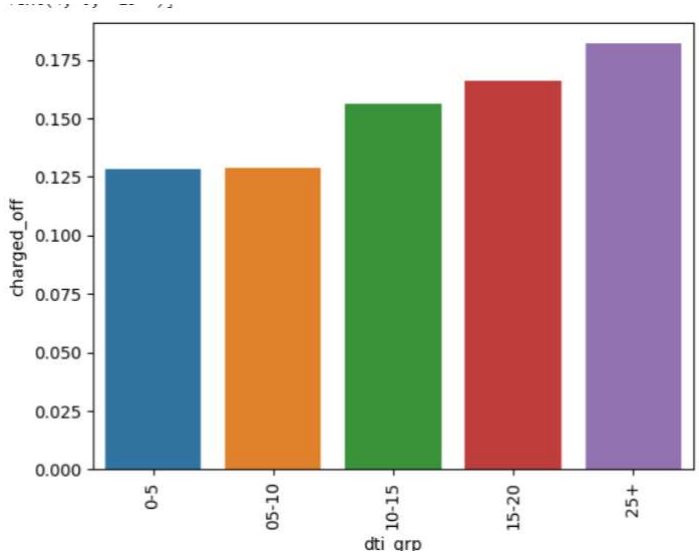
State vs charged off

Inference: Loan from the AK ,NV and SD state are more default



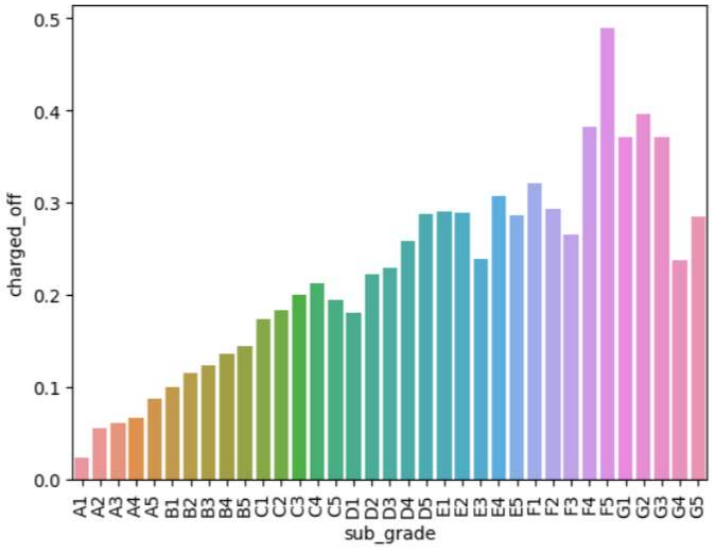
Dti_grp vs Charged Off

As the range of dti increases default increases



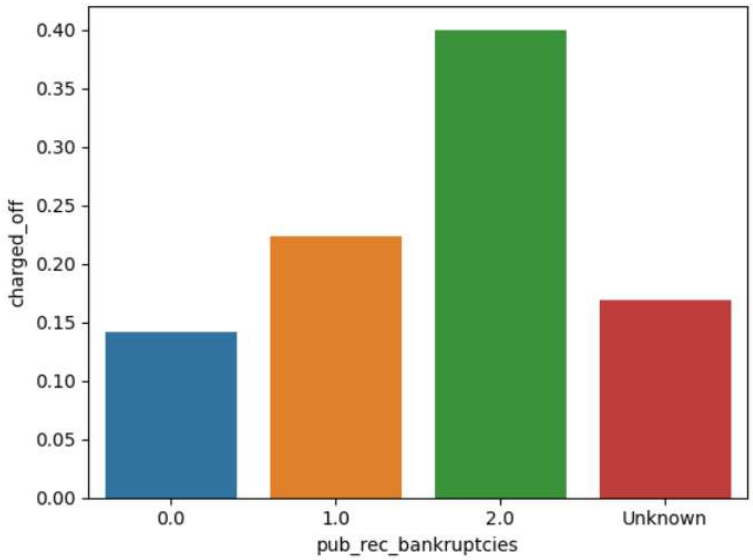
Sub_grade vs Charged Off

Inference : subgrade in F and G are more likely to default



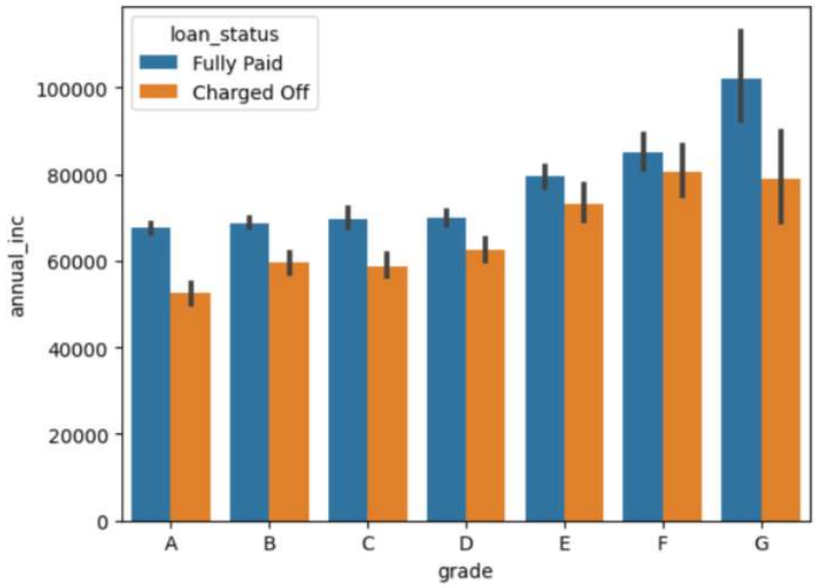
Pub_rec_bankruptcies vs charged off

Customer who have defaulted once will default again



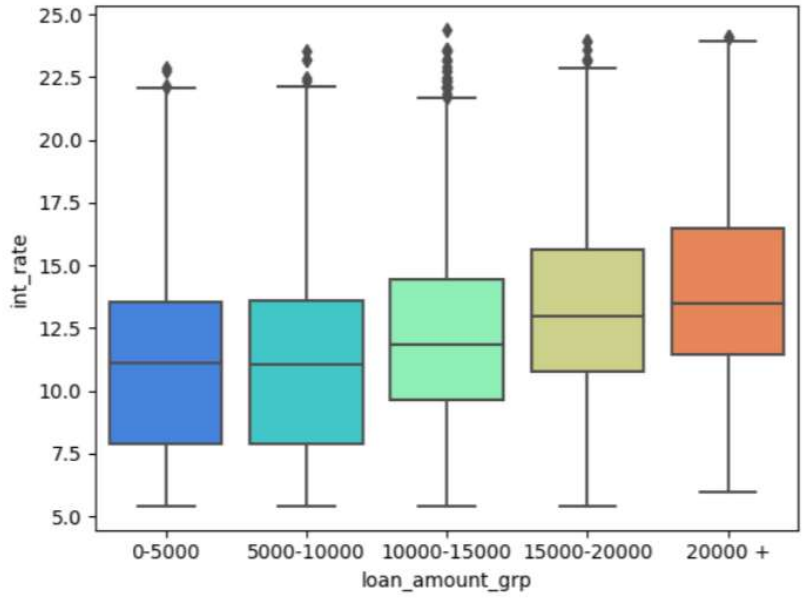
Grade vs annual income

Inference: Annual income of default group is less than the fully paid one



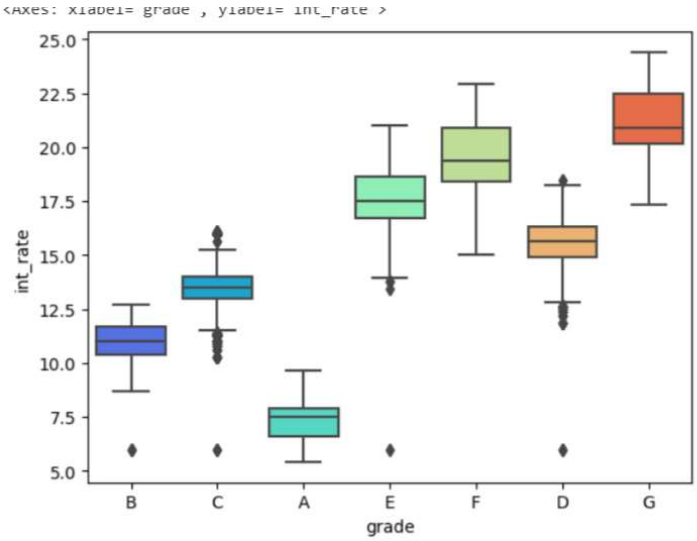
Loan_amount_grp vs int_rate

Inference: Interest rate increases with the loan amount



Grade vs interest rate

Inference: Interest rate increases with grade



Observation of the Bivariate analysis

- Higher interest rate cause more default
- Lesser the annual income more the default
- Loan taken for small business are defaulting more
- Lesser experience default more as the income is less
- Lesser income people more money goes for debt so debt to income is high
- Annual income of default group is less than the fully paid one
- Grade "A" has very less chances of charged off.
- Grade "F" and "G" have very high chances of charged off.
- Interest rate increase with loan amount
- Interest rate increase with debt

Recommendation:

- More the loan taken more is the interest rate and that leads to default

- Default given to lesser in come group have more chances to default despite the less interest offer to them.
- Loan defaulters have high chances of default again
- Analysis done on the segment where the income is above 80000+ with loan taken is with low interest rate shows high percentage of such loans was taken for debt consolidation suggestion that loan given for debt consolidation is risky
- Grade "F" and "G" have very high chances of default.
- Loans given for small business are having more chances to default