EDA CASE STUDY

Presented By:
Deepak Goyal
Sushil Singh

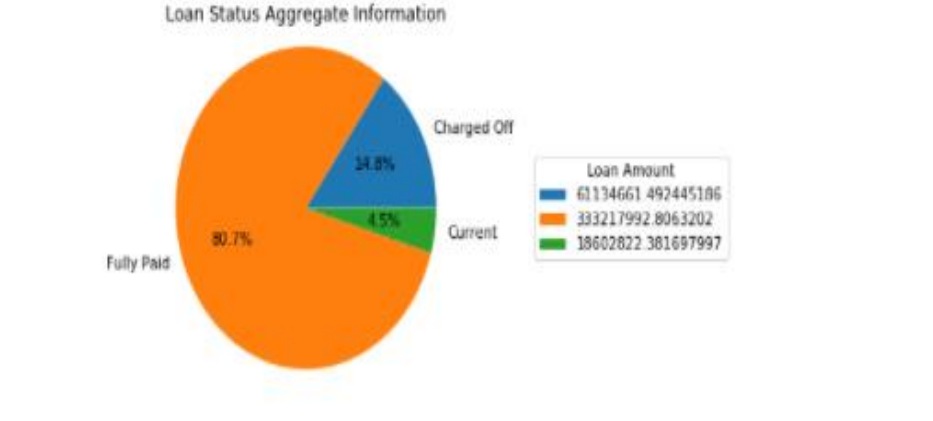
Business Requirements

- **Consumer attributes** and **loan attributes** influence the tendency of default.
- Company wants to use **driving factors (or driver variables)** behind loan default for performing portfolio and risk assessment.
- With this application one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

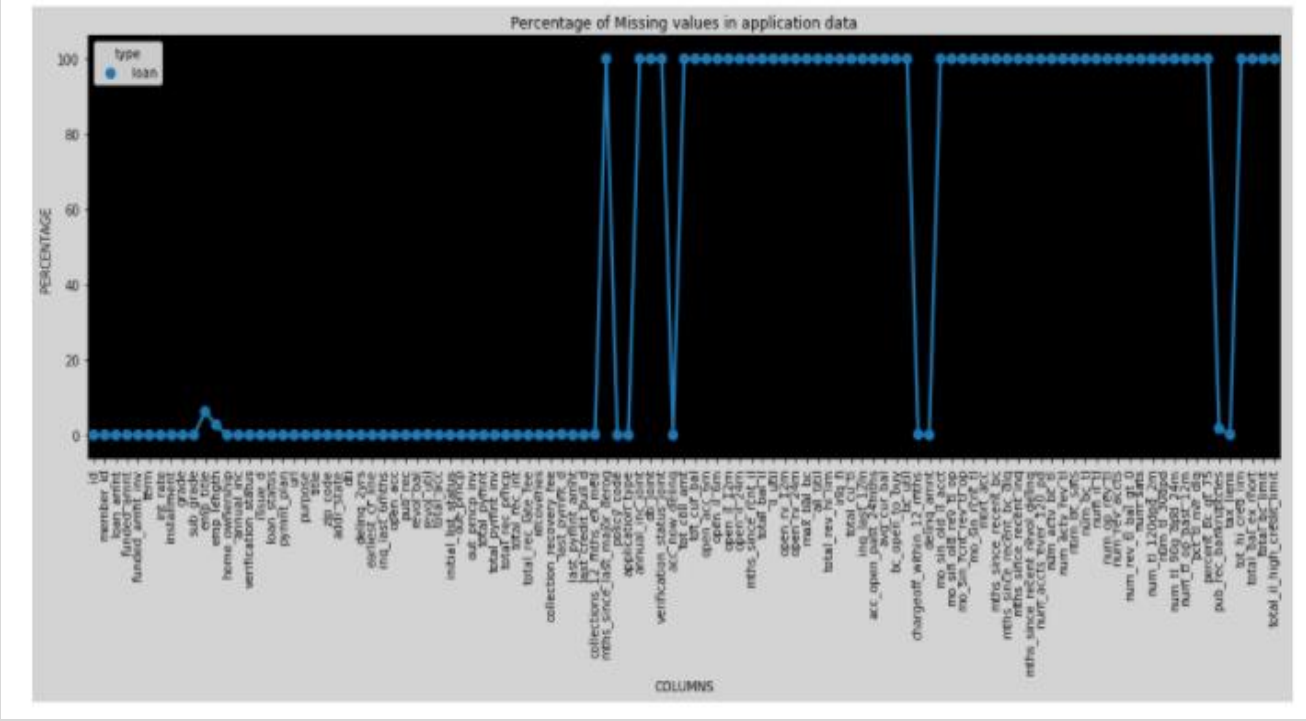
1. Data Understanding

- Data in raw share has 39717 rows and 111 columns
- Just by browsing the data we understand that there are many columns that are completely null and there many columns that are irrelevant so will be removed after analysing.
- Some of the key columns that are relevant for this analysis are:
  - Loan_amnt
  - Term
  - Int_rate
  - Grade
  - Emp_length
  - Annual_inc
  - Loan status
  - Addr_state
  - Home_ownership
  - Verification_Status

Loan Status Comparison



Identify Nulls



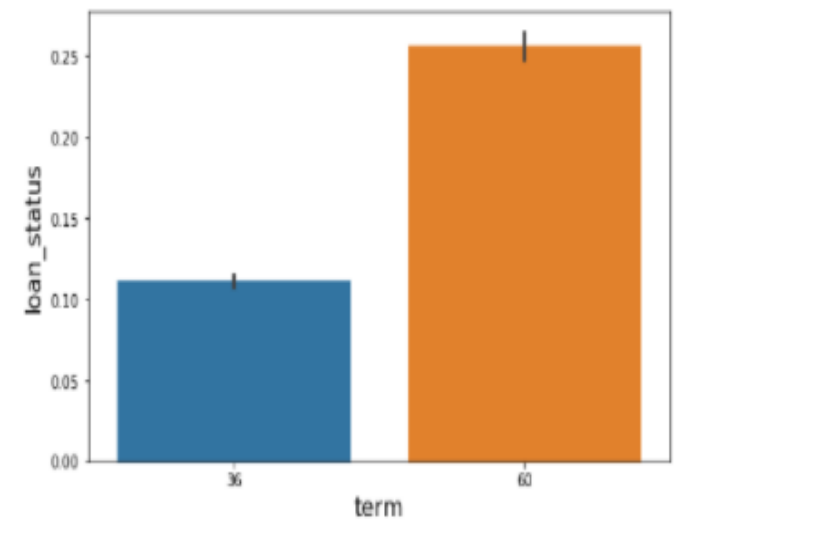
Observation: Charged-off % is very high i.e. 14.8%, so we will check why it is so high.
Identify Nulls: There are 54 columns that are completely null
Identify Outliers: Some of the columns like loan amount and income had outliers and were treated appropriately.

2. Data Cleaning and manipulation

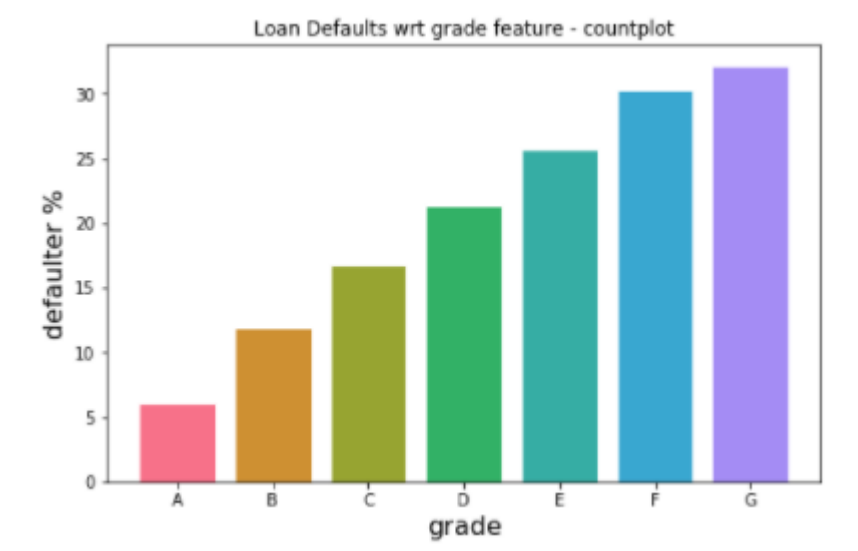
- **Remove columns that are completely null**
  - Dropped 54 columns that are completely null.
- **Remove columns having nulls greated than 30 % threshold**
  - 4 columns identified and dropped 'desc', 'mths_since_last_delinq', 'mths_since_last_record', 'next_pymnt_d'
- **Remove columns that won't help in this analysis**
  - Dropped columns "url","zip_code","id"
- **Drop the columns that are highly co-related**
  - For this we have to draw co-relation matrix and drop the variables like total_pymnt_inv, etc that are highly co-related.
- **Standardized the values** e.g. converted loan_status to numeric values for easier analysis and data is binned properly for loan_amnt, annual_inc, interst rate
- Some of the **derived metrices** like loan to income ratio was added.

3. Data analysis

Impact of term on loan status

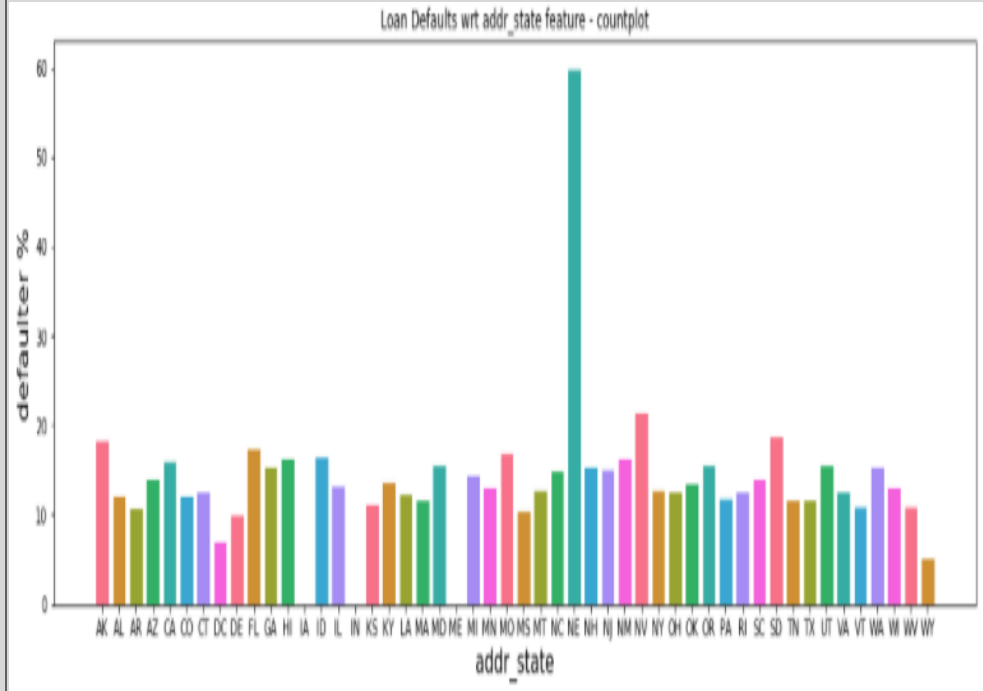


Impact of grade on loan status

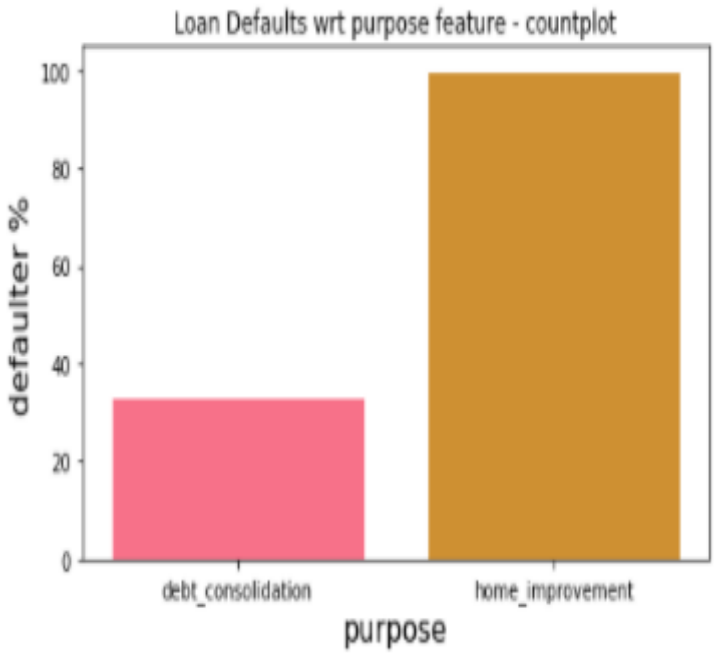


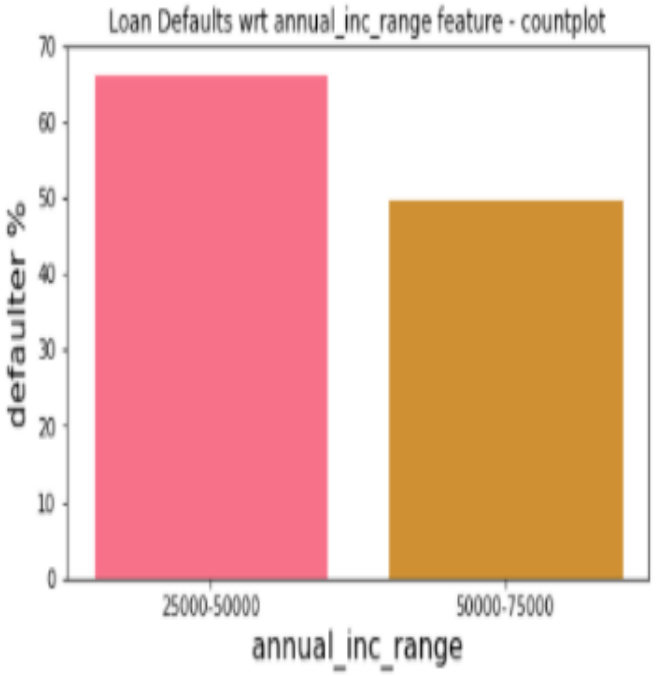
Observations: Loan has more number of defaults as term of loan increases.
% of default increases significantly as the grade of the person decreases.

Default % per state



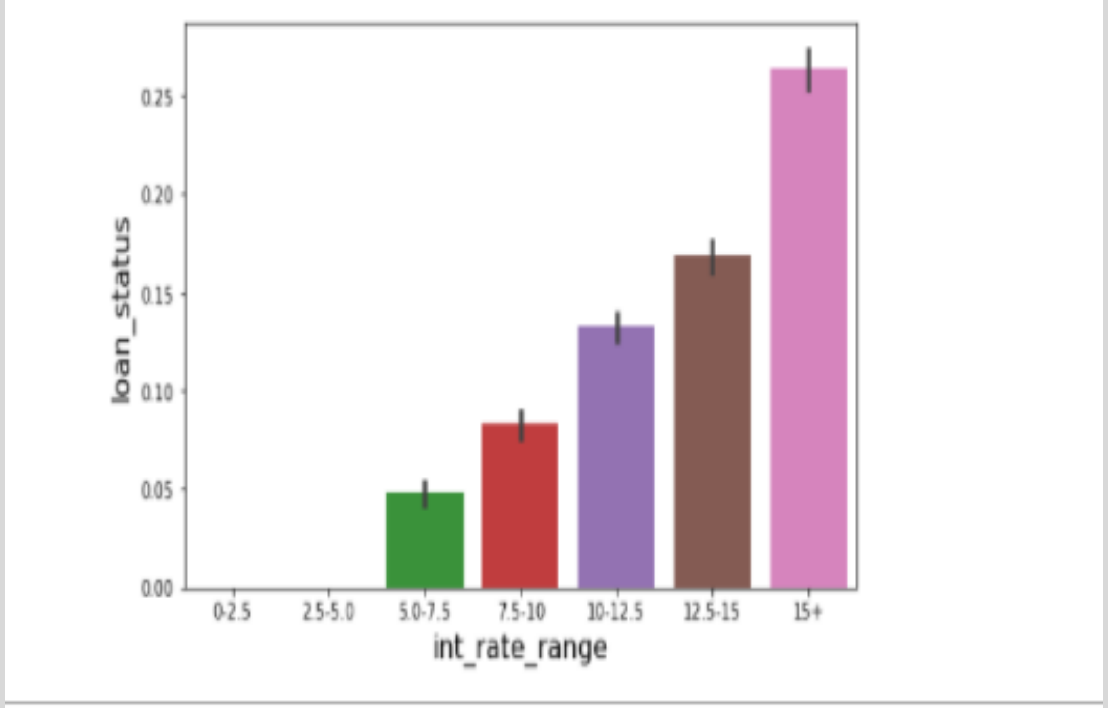
Impact of purpose and annual income in NE state



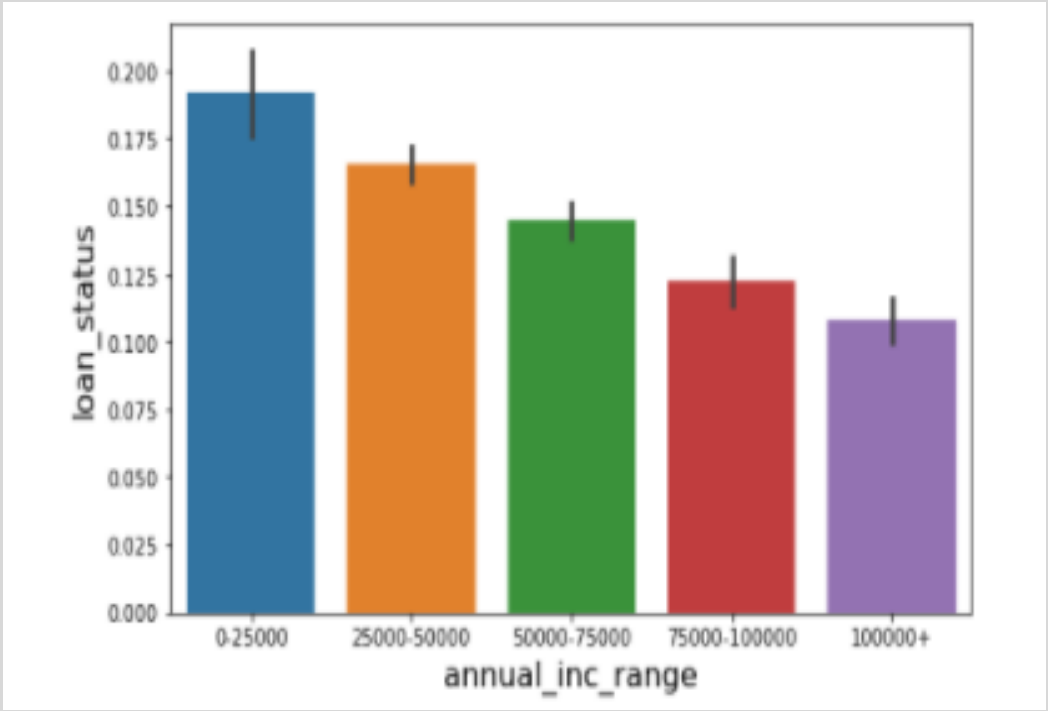


Observations: Loan default ratio is significantly high in state of "NE" as compared to other states. In State of NE, loan default is way higher if taken for home_improvement and if annual income is < 50000. Hence, we should be more vigilant in giving home_improvement loans in NE state where income < 50000.

Impact of interest rate on loan status

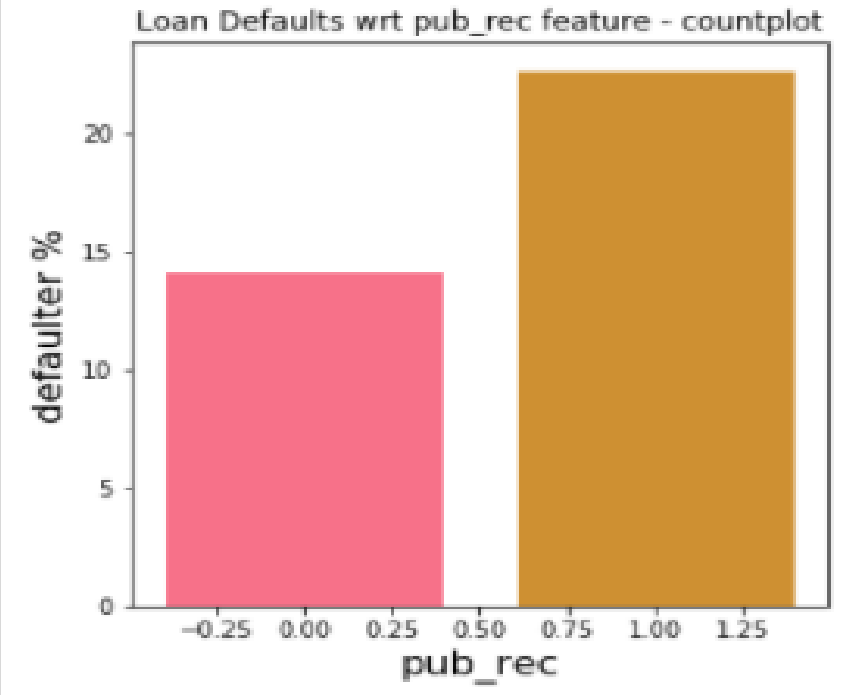


Impact of annual income on loan status

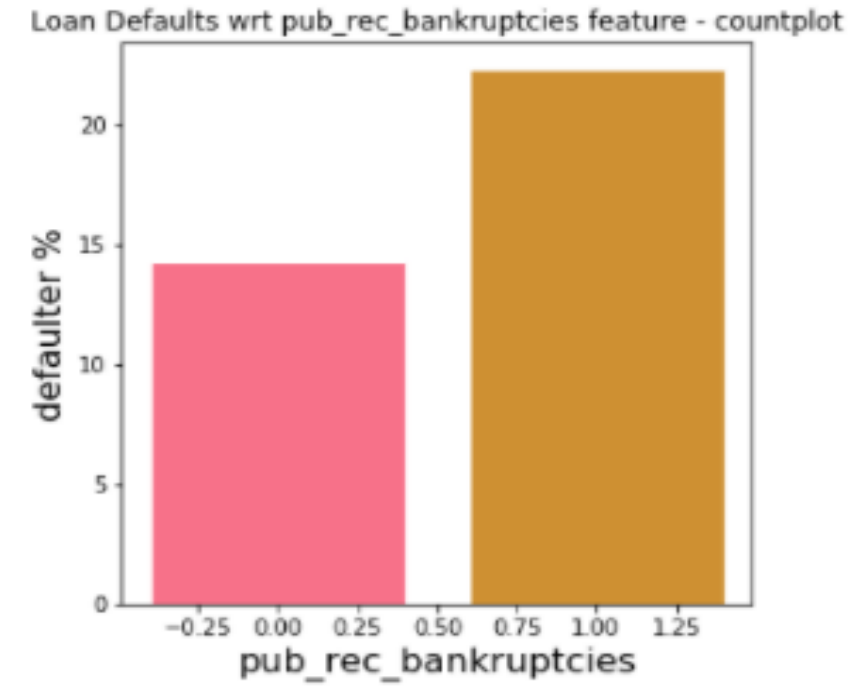


Observations: As interest rate increases loan default increases but if it is > 15 then there is significant rise in loan default.
As annual income increases loan default decreases.

Impact of public record on default %



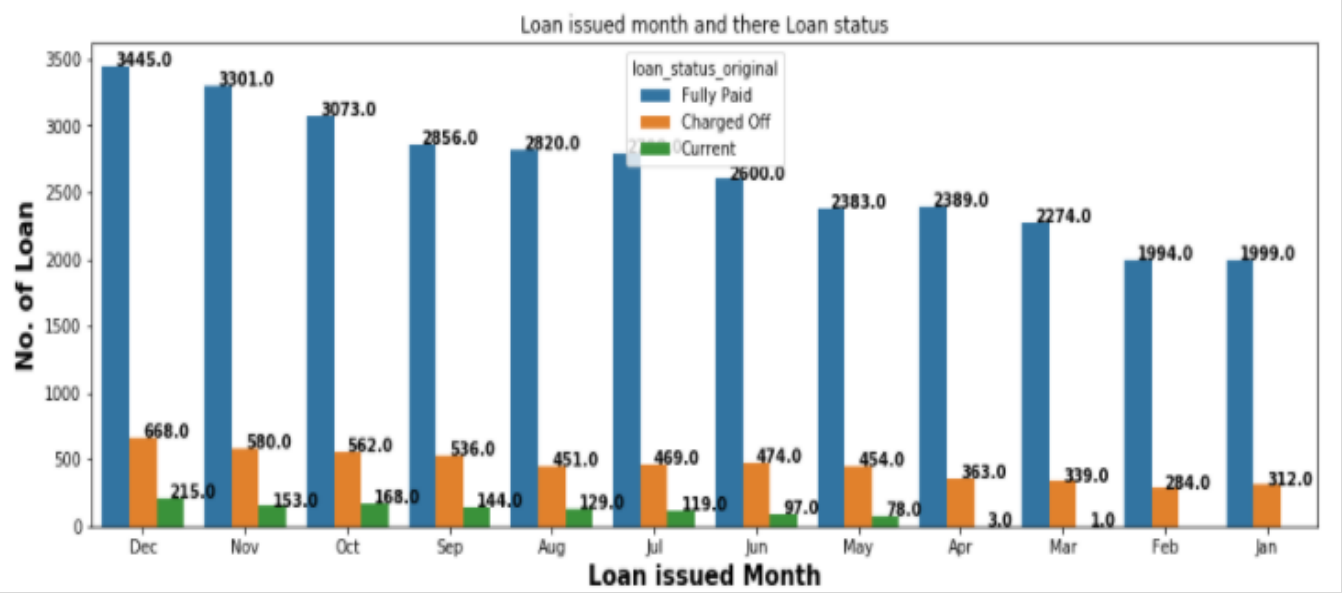
Impact of bankruptcies on default %



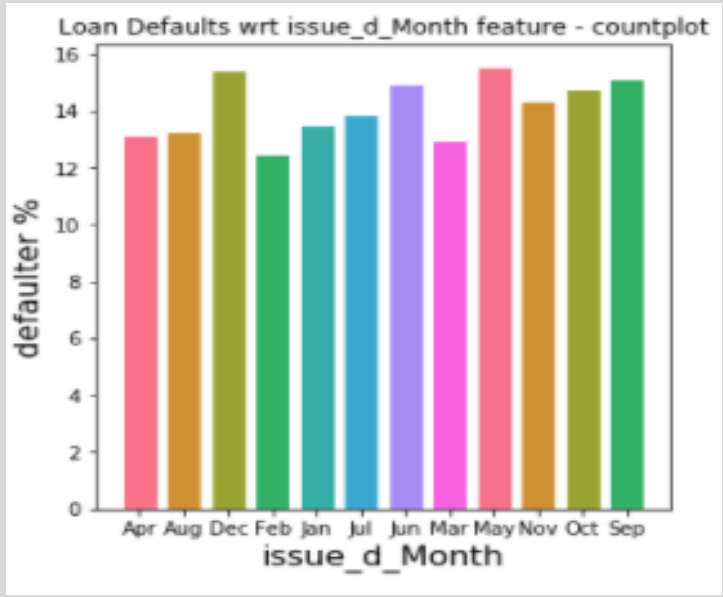
Observations: Default ratio is high if person has derogatory public record.

Borrower with bankruptcies record have high percentage of charged off.

No. of loans done month wise



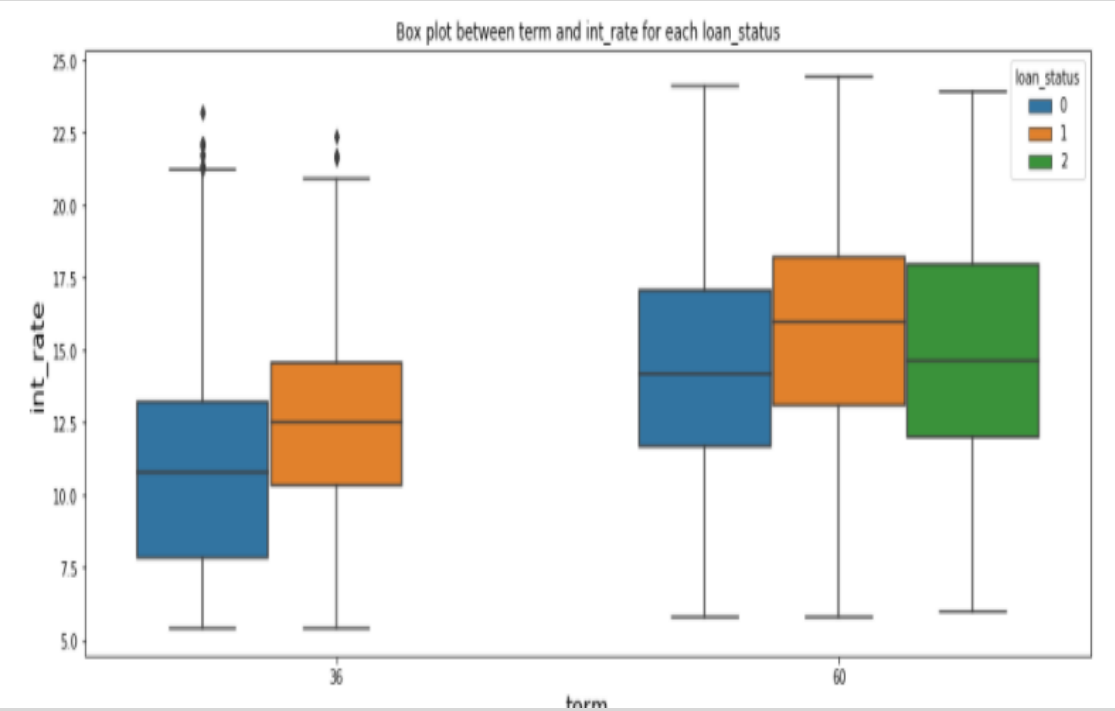
Default % of loans month wise

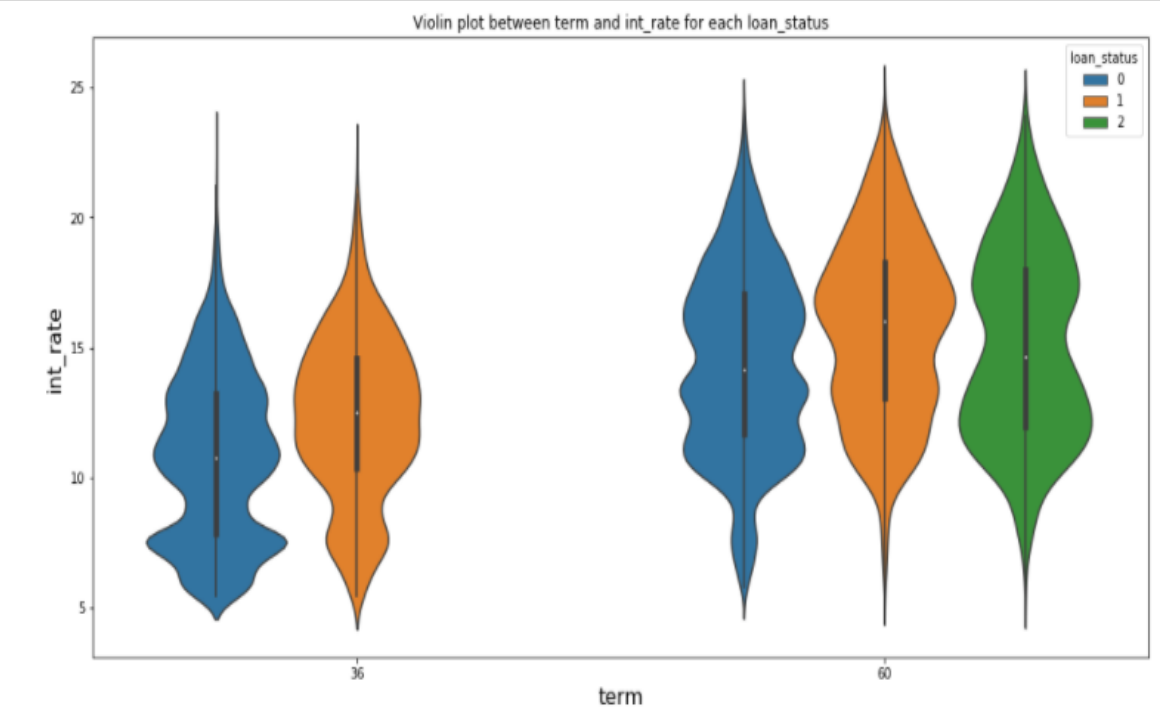


Observations: High % of loans granted in Q4 (Dec,Nov, and Oct).

Most of the loan taken in Q4 (Dec,Nov, and Oct) and default ratio is also high for these months.
Probably Sales person are giving more loans in last quarter for meeting sales target.

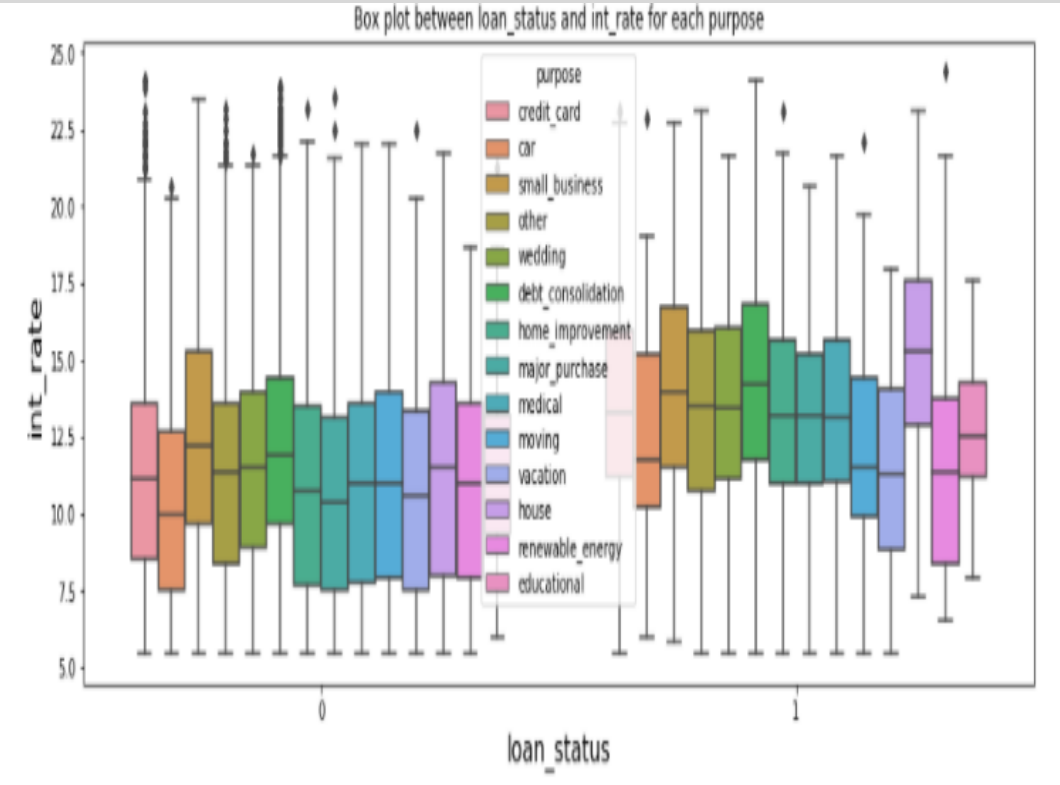
Impact of interest rate and term on loan status via box and violin plots

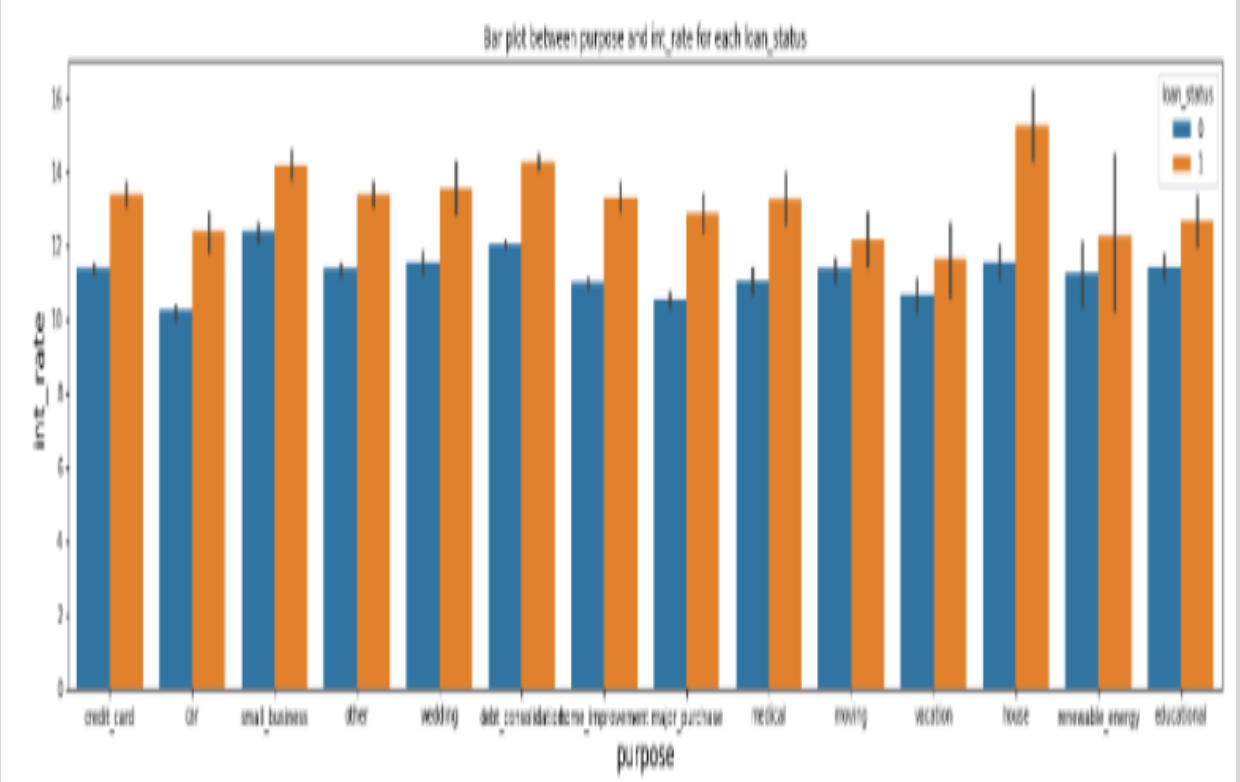




Observations: Looking at box plot, higher the term , higher is the interest rate. But within same term, chances of loan default is higher with higher interest rate.
In violin chart, for 36 months term there is high density below the median for loans that are fully paid while in charged off loans high density is above median.
Similarly in 60 months as term there is proportionate density around median for loans that are fully paid while in charged off loans high density is above median

Impact of Purpose and interest rate on loan status





Observations: In case of default home loans are charged at higher interest rate as compared to other types of loans**.**

Conclusion

1. Loan has more number of defaults as term of loan increases, so prefer to give short term loans.
2. % of default increases by 25%(Grade A vs Grade G) as the grade of the person decreases.
3. % default in NE state is 40% more than other states, Hence, we should be more vigilant in giving home_improvement loans in NE state where income < 50000.
4. Default ratio is 8% high if person has derogatory public record.
5. Borrower with bankruptcies record have 8% high percentage of charged off.
6. Most of the loan is granted in Q4 (Dec,Nov, and Oct) and default ratio is also high for these months. So don't put sales pressure to disburse loans