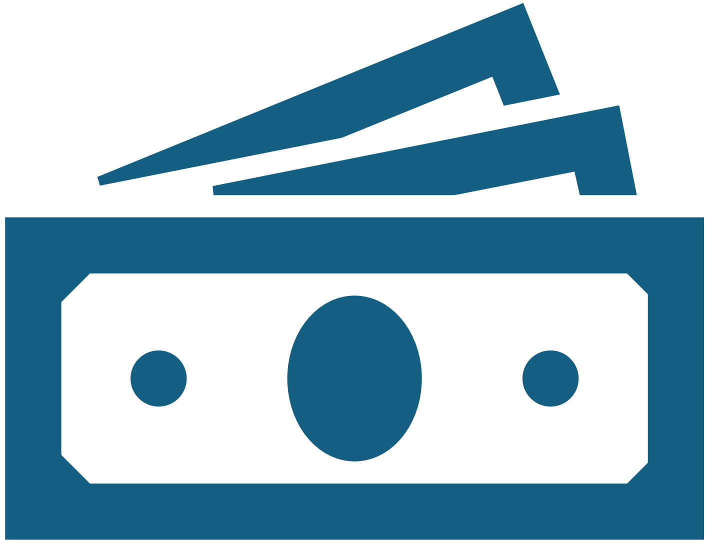
Lending Club case study



- Sathiyanathan Subramanian
- Ushasis saha

Problem Description

Our consumer finance company faces the challenge of credit loss, where loan approvals to unlikely repayers may result in credit losses.

The dataset given contains information on past loan applicants and their default status.

Our objective is to understand how consumer and loan attributes influence loan default tendencies. Our goal is to mitigate credit loss by identifying the driver variables behind loan default

Approach

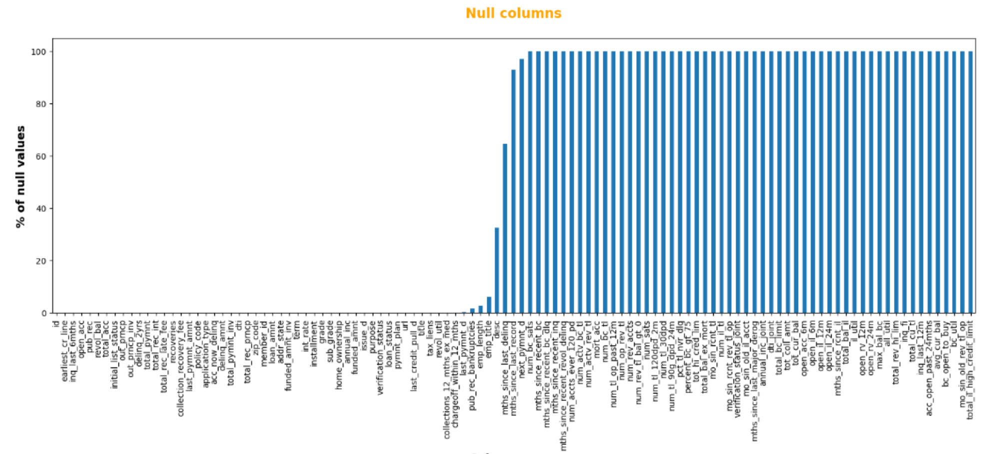
Ignored all null columns – There are 55 columns that are completely null

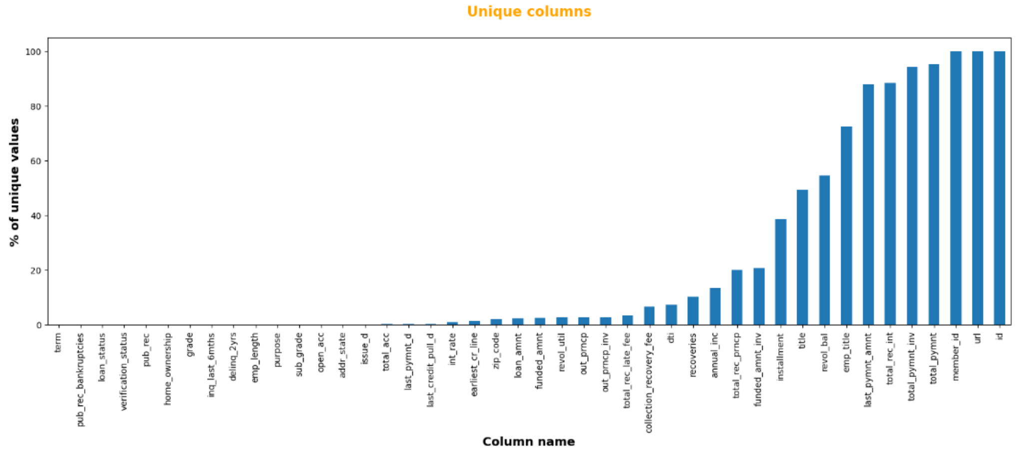
Ensured target variable integrity: No incorrect/null values identified

Removed columns with only one unique value across all records.

Columns with Categorical variables and high values for all records were removed
Ex: ‘member_id ‘ this would have unique value for all rows unique

Removed variables related to post-loan recovery (e.g., 'recoveries', 'collection_recovery_fee') as they don't influence loan approval. Confirmed unique member IDs, ensuring each user has only one loan with the company.





Approach

Manipulation of Datatypes:

Handling Date and String Variables Appropriately

Data Quality Checks:

Identification and Treatment of Missing Values

Removal of Rows with 'current' Loan Status

Treat outliners

Data Segmentation:

Splitting the Data into Two Parts:

- Defaulters
- Fully Paid Cases

Approach

Classified Variables as Numeric and Categorical (Ordered, Unordered)

Univariate Analysis:

- Numeric Variables: Distribution Plot comparing Paid vs. Defaulters
- Ordered Categorical Variables: Segmented by Bucketing, Analyzed with Bar and Pie Charts
- Unordered Categorical Variables: Analyzed with Grouped Line Chart, Bar Chart, Pie Chart, and Grouped Columns

Identified 10 Driver Variables Based on Count Analysis

Approach

Bivariate Analysis:

Numeric Variables:

- Pair Plot: Visualizes correlations between numeric variables.

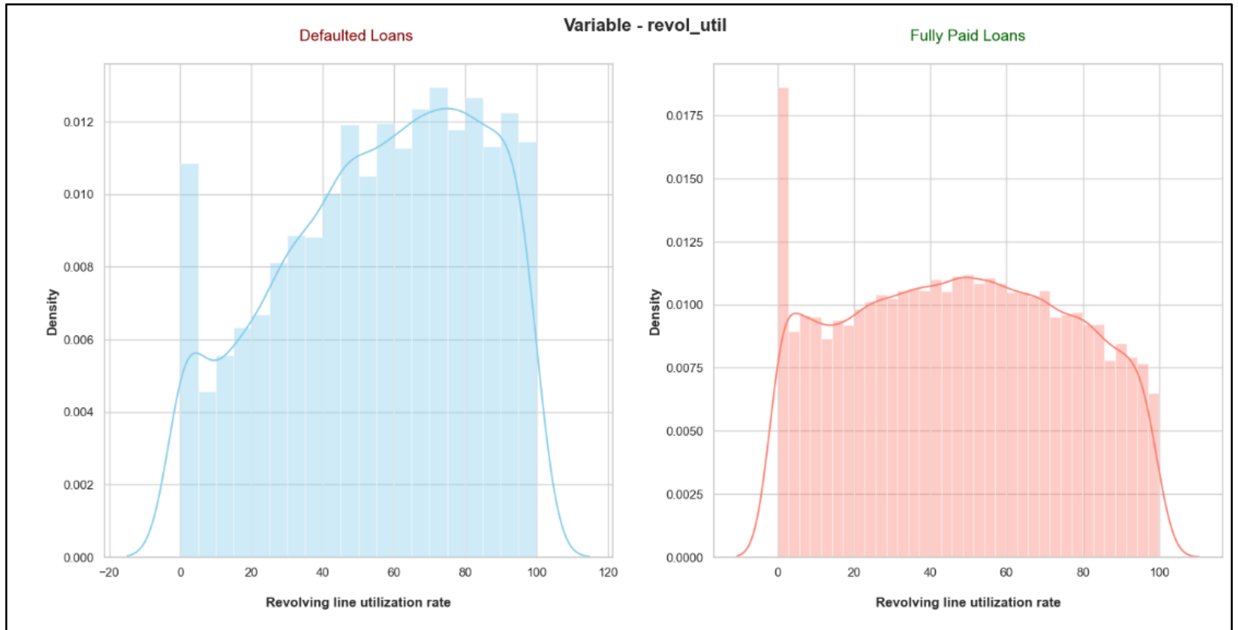
Categorical Variables:

- Pivot Table + Heatmap: Illustrates correlations between categorical variables.

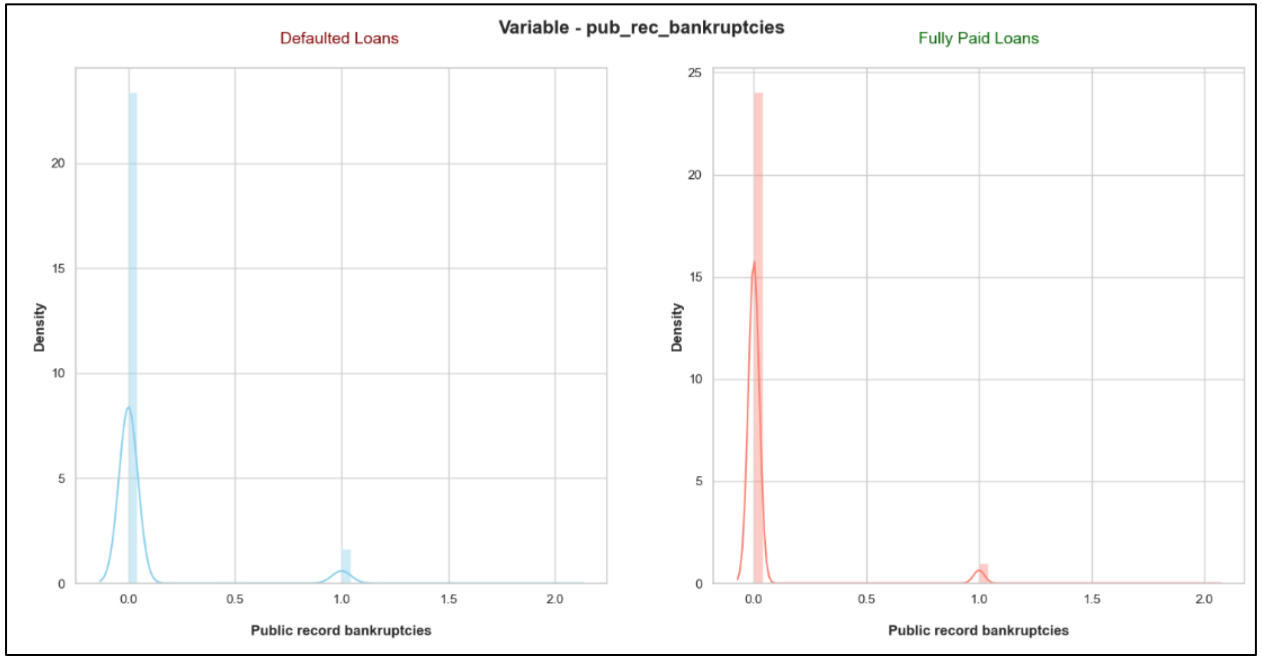
Identified Relationships:

Discovered 2 related categorical variables, 2 related numerical variables; reducing total driver variables to 9.

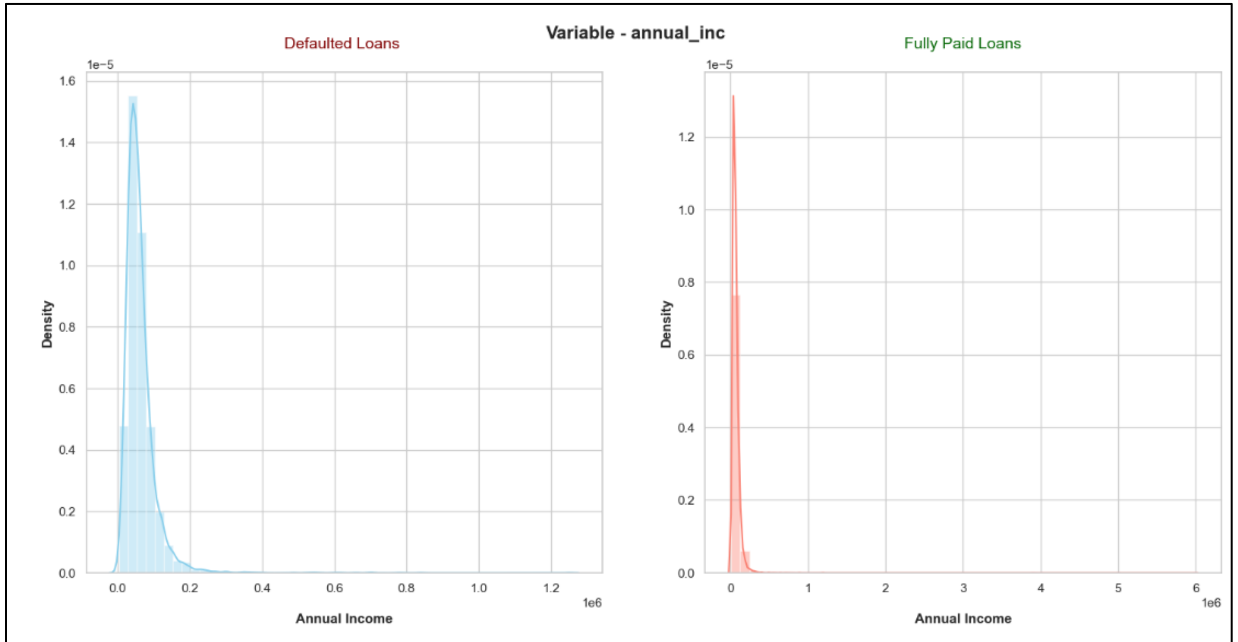
Numeric Variables



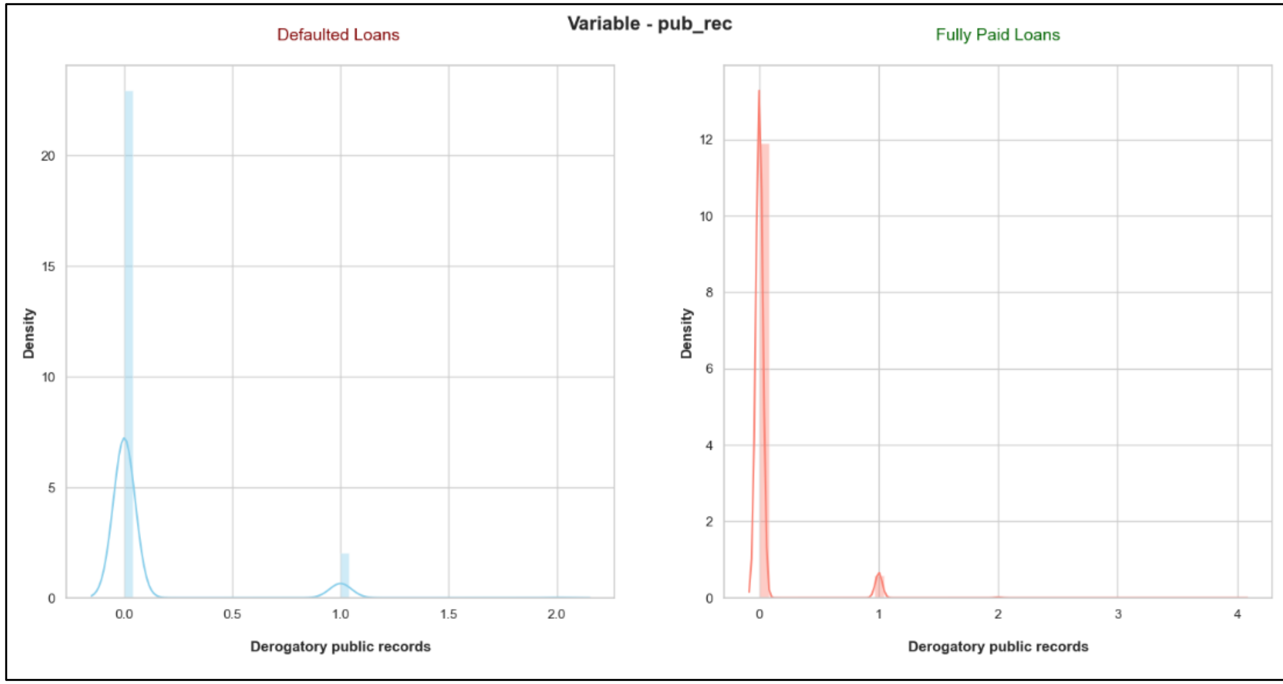
Observation: We could see that PDC is different. Desnity peaks especially around 60-80%.Therefore ***this column will be considered***



Numeric Variables



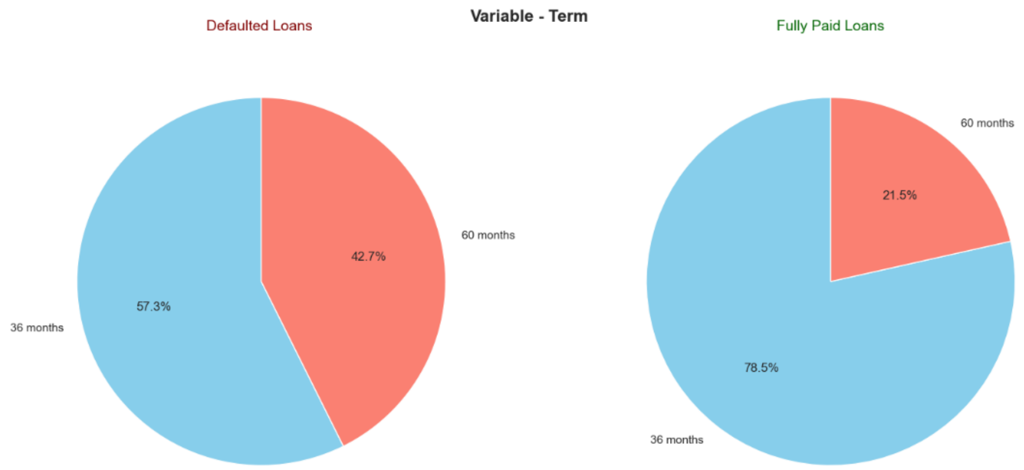
Observation: Unable to comment on annual_inc at this moment, the spike seems to be similar; but the range of annual_inc is different. Could be due to the outliers, we need to revisit to check the outliers. ***We wil consider this column***

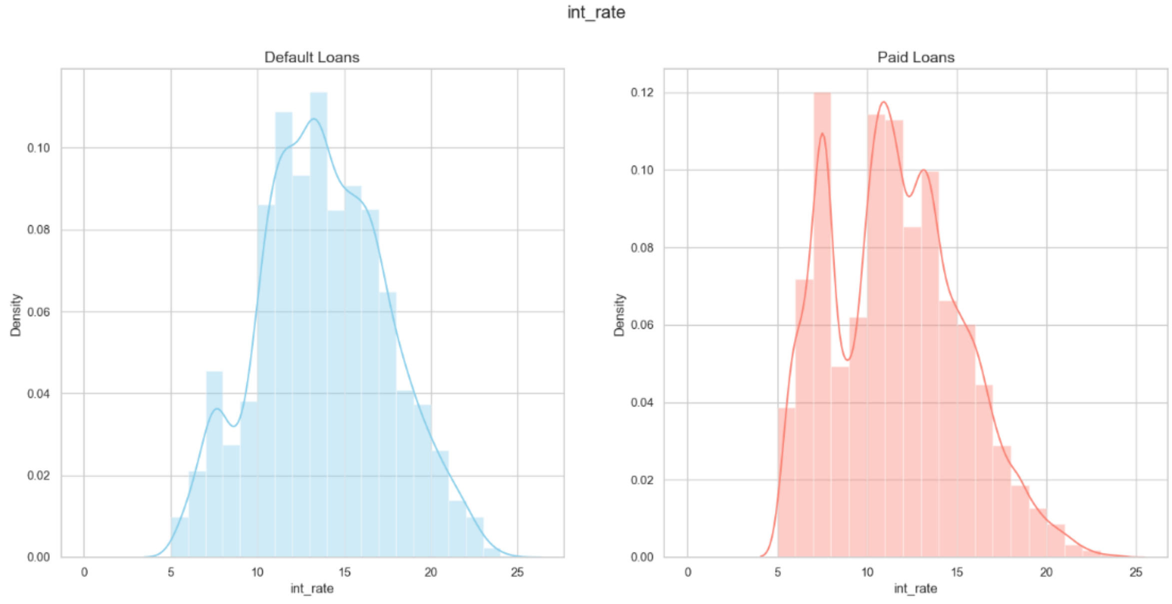


Ordered categorical Driver variables

*Observation: We could see that there is difference in distribution for 60 months segment, let's drill further into this. Plotting a pie chart, we could see that in case of Defaulted Loans, 60 months term contributes a lot compared to Fully paid. This **variable is considered for driver variables***

```
[71]: pie_chart(df_default.term, df_paid.term, 'Term')
```



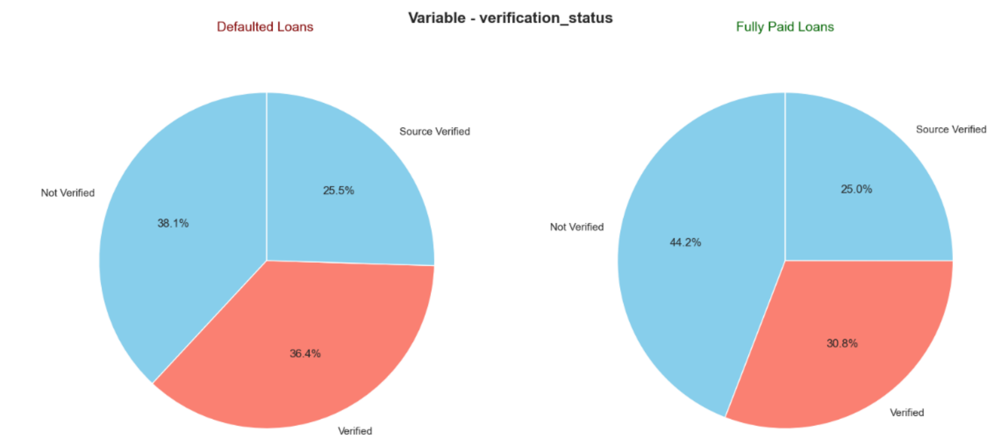


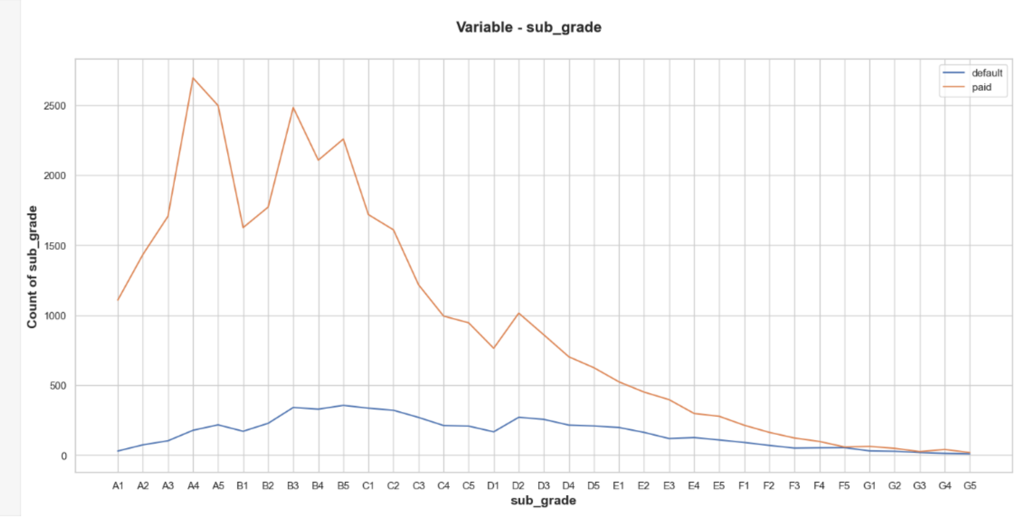
*Observation: We could see that there is difference in PDC, IN fully paid loans, there is a dip at 8, 9% but in default it a steady spike. Lets **consider this variable as driver variable***

Ordered categorical Driver variables

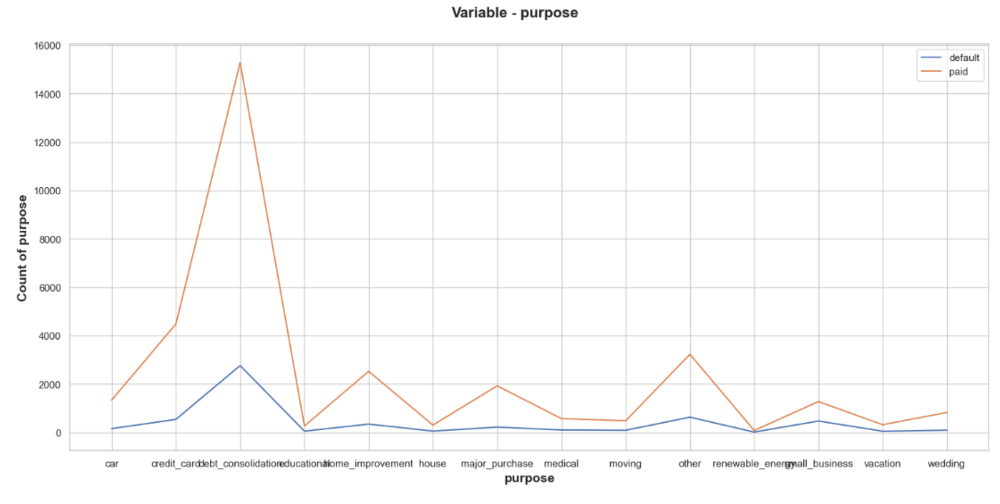
Observation: We could see a differene in pattern especially for verified, lets dig further with Pipe chart. The disribution might look similar,but there is a 6% share difference b/w verified and non-verified. Therefore lets consider this variable for driver variable.

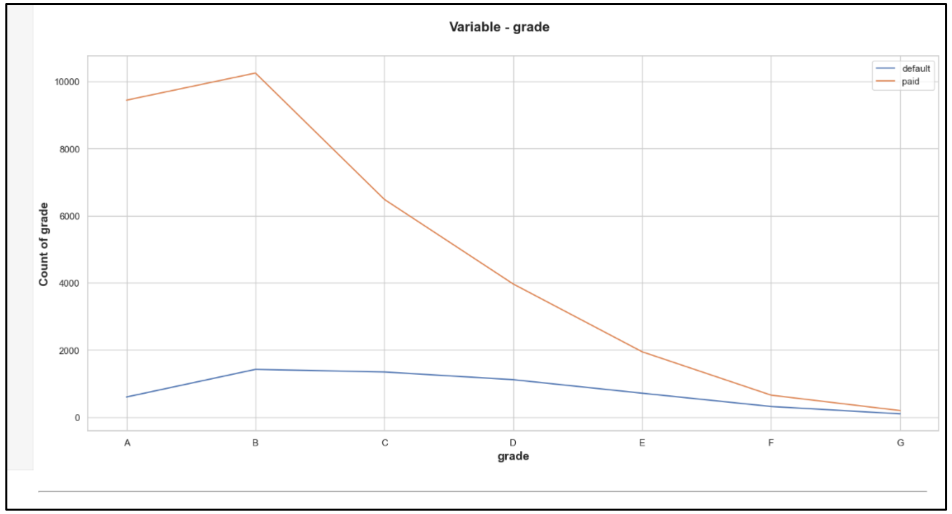
```
[85]: pie_chart(df_default.verification_status, df_paid.verification_status, 'verification_status')
```





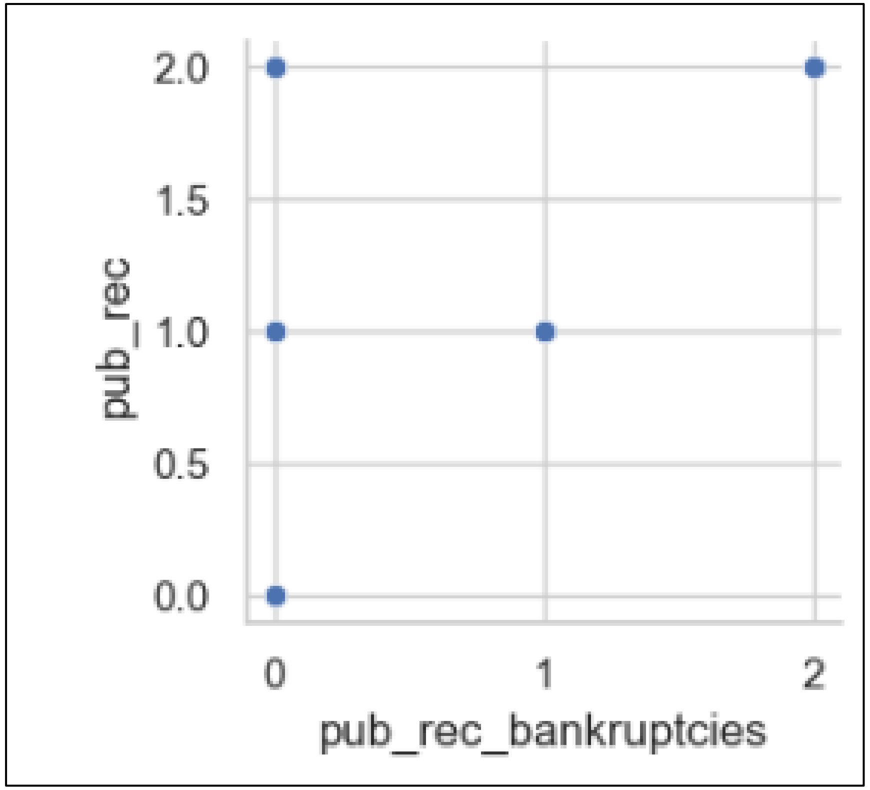
Observation: similar pattern like observed in grade. Therefore this column will be considered as a driver variable

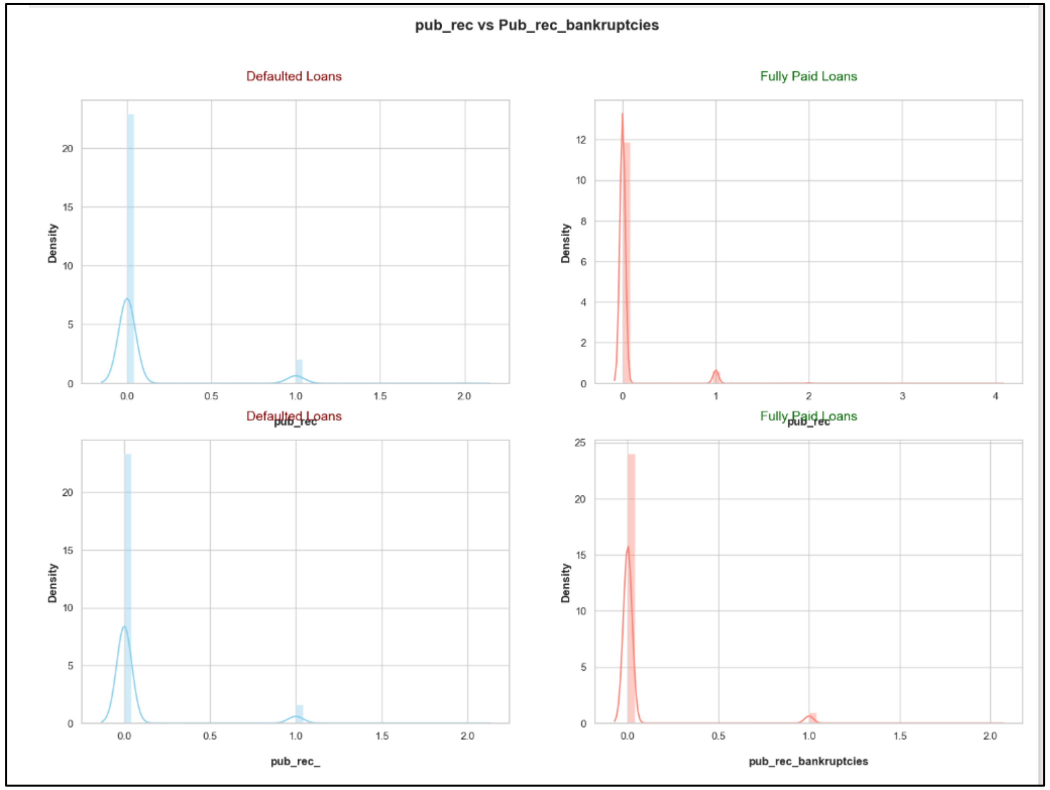




Bivariate – Numeric correlation

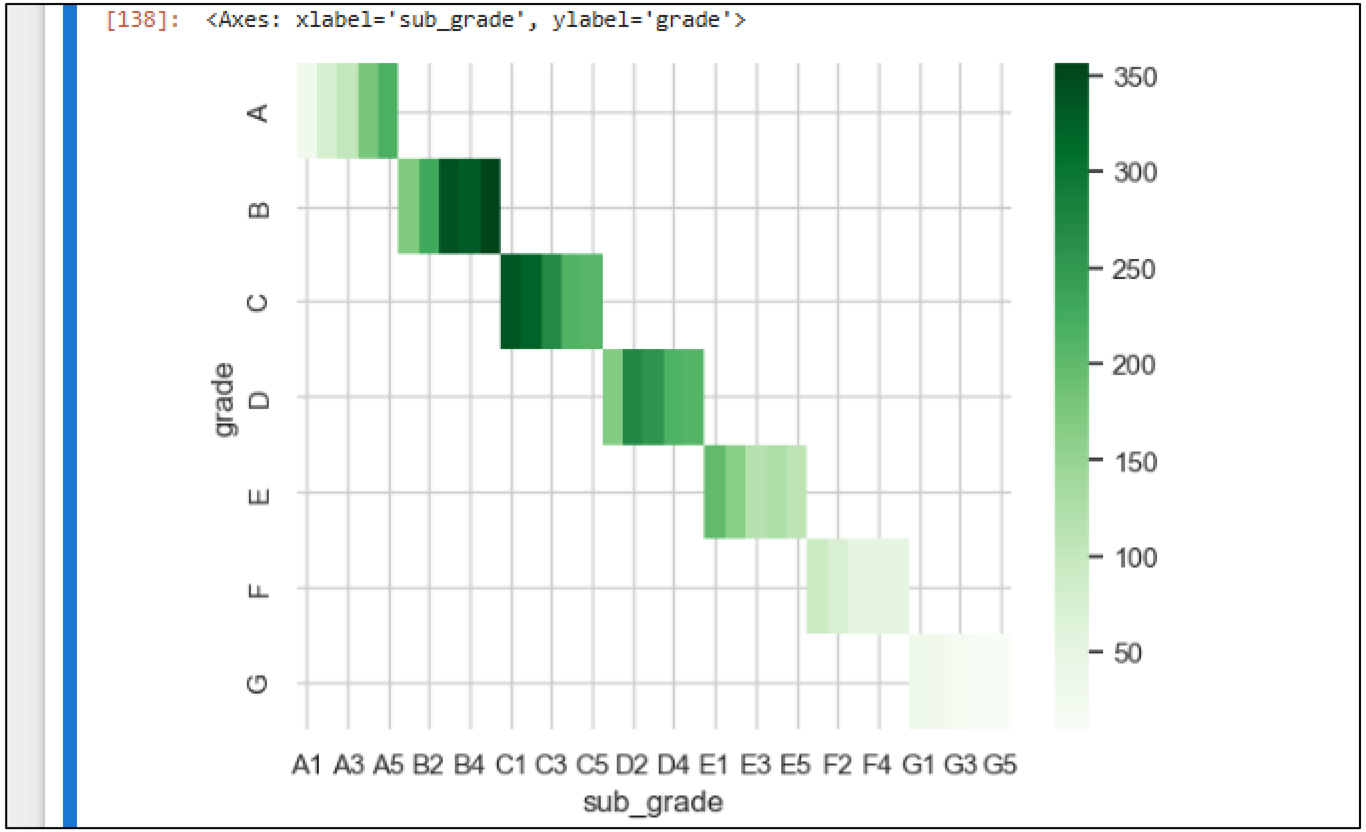
Pub_rec vs Pub_rec_bankruptcies, has linear progression relationship with each other, we shall ignore Pub_rec_bankruptcies from driver variables





Bivariate - categorical

- Grade, sub_grade – grade can be dropped
- Output is logical, because sub_grade is component of grade



Inferences/Relationships from other Bivariate analysis

Relationships:

- Pub_rec and pub_rec_bankruptices are related to each other, one can be dropped
- Similarly, Grade – Sub_grade is a component of Grade; grade can be dropped

Few Inferences:

- int_rate vs term:Loans of 36 months with interest 10-14% seems to be risky
- grade vs term: Loans of 36 months with sub_Grade B3,B5 seems to be high risky
- verification_status vs term: Loans of 36 months not verified is risky
- purpose vs term: Debt Consolidation seems to be risky - highest for 36 months and next for 60 months

Recommendations & Conclusions

- We have identified 8 variables that drive the target variable (loan_status)
- Also, inferential relationships can be identified, which can be used in machine learning model later

Variable	**Variable type**
term	Ordered Categorical
int_rate	Ordered Categorical
sub_grade	Unordered Categorical
annual_inc	Numerical
verification_status	Unordered Categorical
purpose	Unordered Categorical
pub_rec	Numerical
revol_util	Numerical

Thank you