LENDING CLUB CASE STUDY

SUBMISSION

Group Assignment

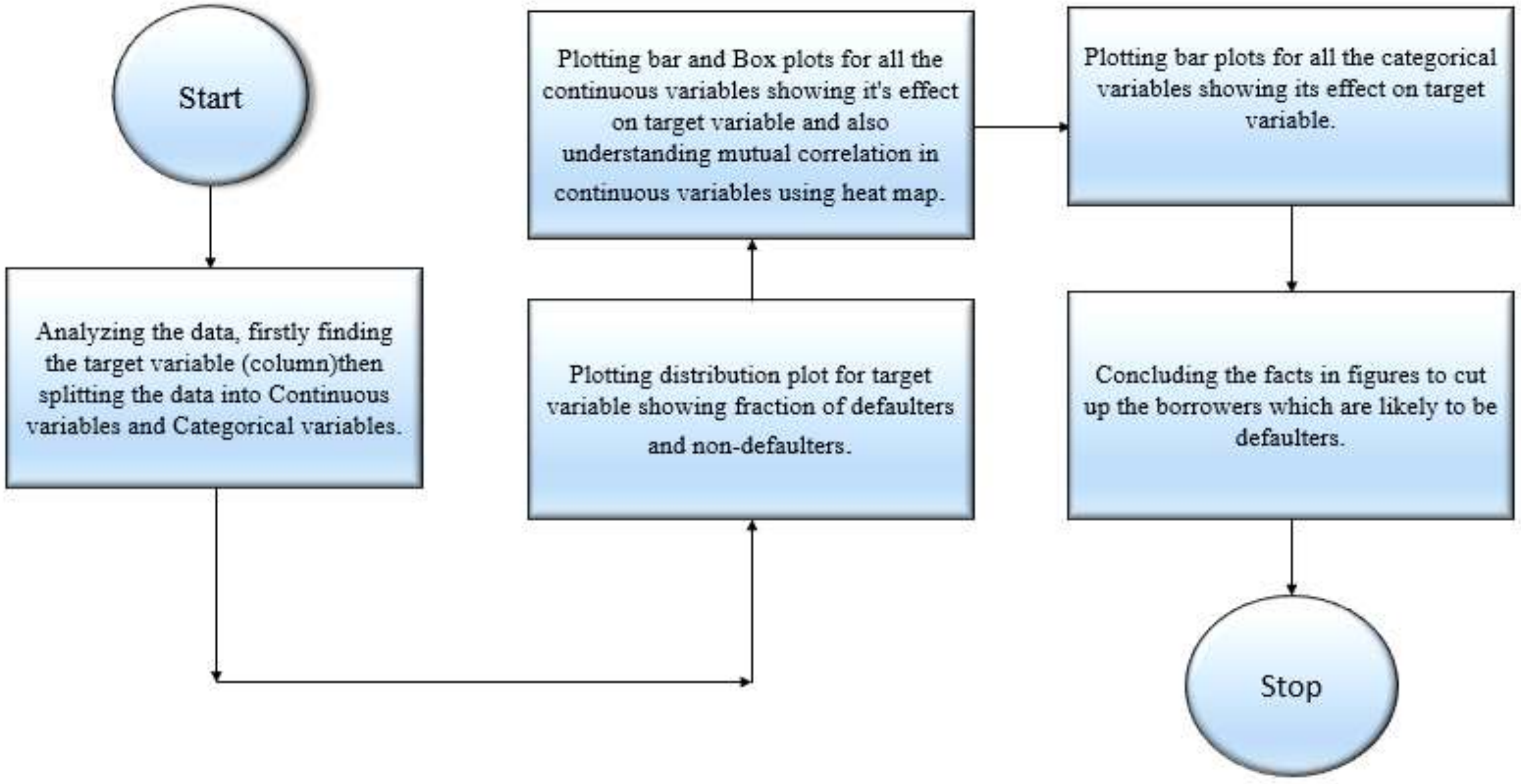
Submitted by : ***Navoneel Chakrabarty***

Ankita Chakrabarti

Abstract:

For the analysis on the past customer data provided by Lending Club, firstly target variable is found and then the impact of all other variables, on the target variable is analyzed so as to conclude whether the borrowers are likely to be defaulters or not.

Flowchart:



Analysis:

- Original data-frame is loaded, and distribution plot for Target Variable (loan_status) is constructed and shown in Fig 1.
- Continuous Variable Analysis:
  - ✓ The features containing null values are detected. Those containing medium % of NaN values (mths_since_last_delinq) are imputed with median, those containing large % of NaN values (mths_since_last_record) are removed and for those containing very small % of NaN values (revol_util), corresponding rows are dropped.
  - ✓ The required continuous features are extracted from the main data-frame and mutual correlation is calculated and presented in the form of a heatmap in Fig 2.
  - ✓ The resulting data-frame with only continuous features is further bifurcated into 2 sub data-frames, df_not_default and df_default containing 'fully paid' and 'charged off' data respectively.
  - ✓ Default Rate for each category (synthetic class intervals) based Bar Plots are constructed for analysis along with Box-Plots for both not_default and default data for outlier-range detection.
- Categorical Variable Analysis:
  - ✓ The features containing null values are detected. Those containing medium % of NaN values (emp_title, emp_length, pub_rec_bankruptcies) are imputed with median, those containing large % of NaN values (next_pymnt_d), are removed and for those containing very small % of NaN values (last_pymnt_d, last_credit_pull_d), corresponding rows are dropped.
  - ✓ The resulting data-frame with only categorical features is further bifurcated into 2 sub data-frames, df_not_default and df_default contaning 'fully paid' and 'charged off' data respectively.
  - ✓ Default Rate for each category based Bar Plots are constructed for analysis.

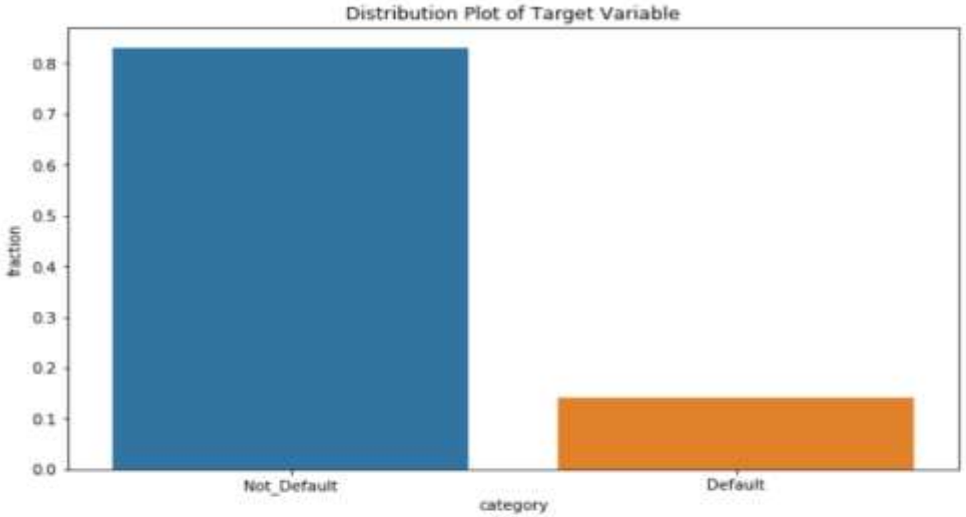


Fig 1. Distribution Plot for Target Variable

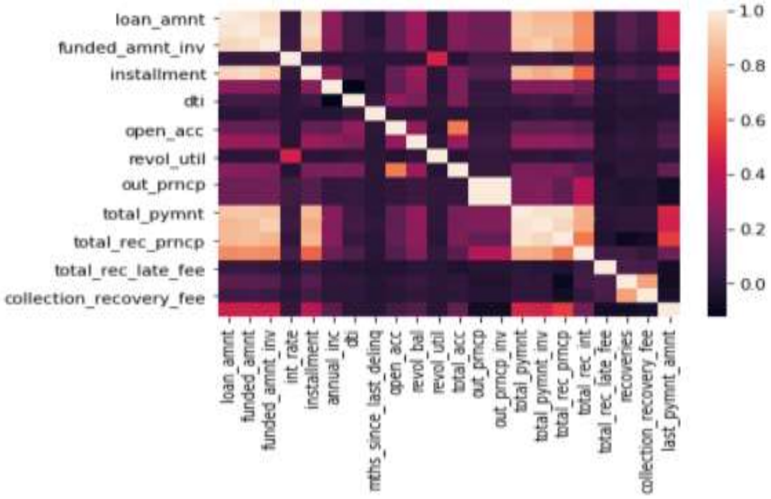


Fig 2. Heat-Map showing mutual correlation for continuous variables

Results and Conclusions: For Continuous Variables

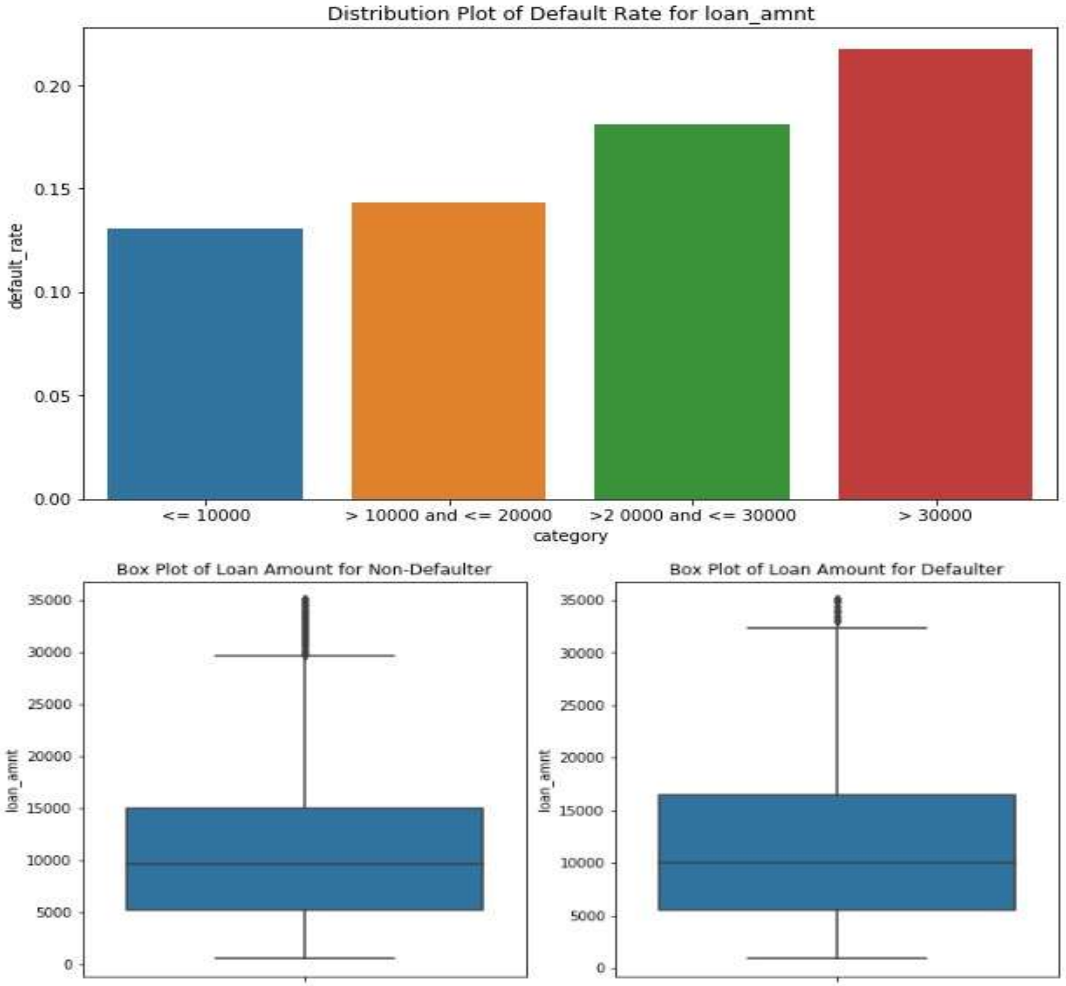


Fig 3. It can be concluded that for ***loan_amnt*** > 29700.0 there are higher chances of the borrower to be a DEFAULTER.

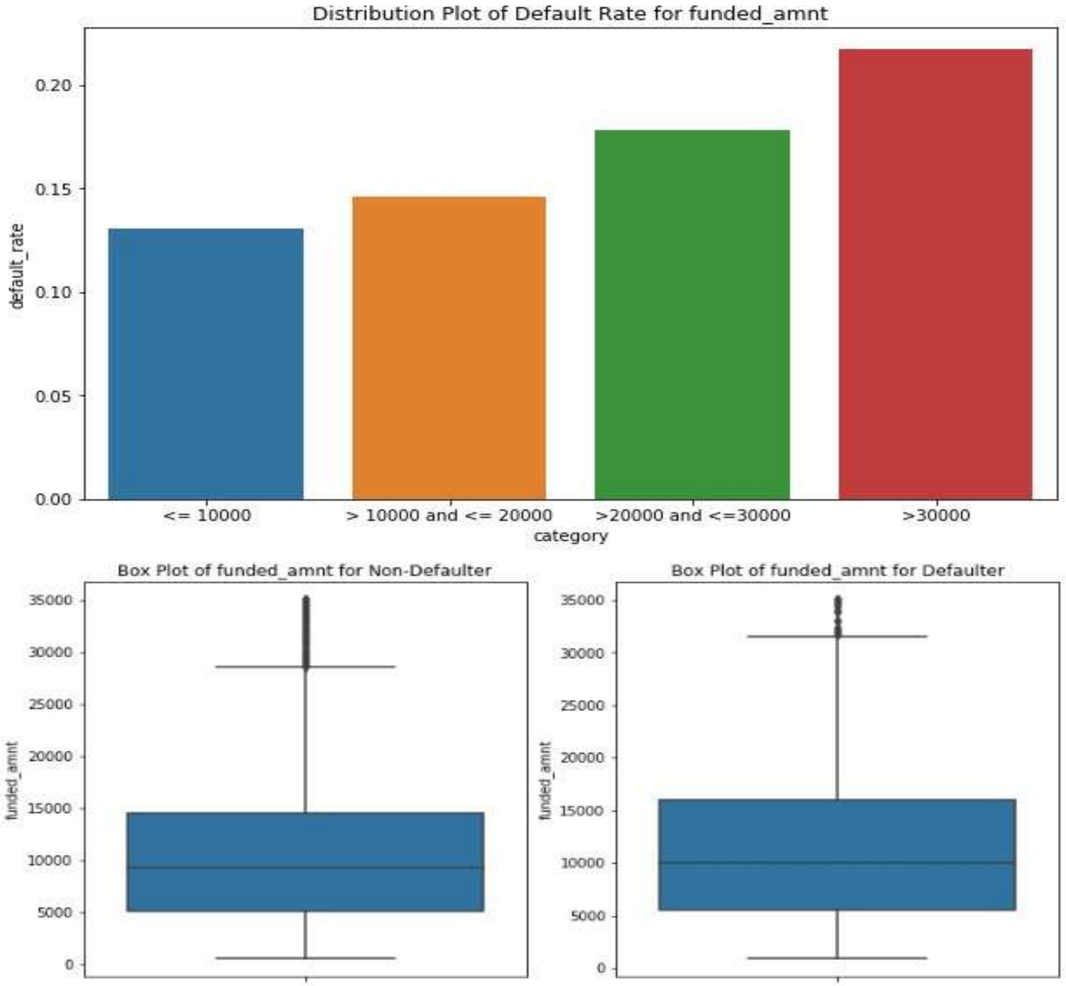


Fig 4. It can be concluded that if the ***funded_amnt*** > 28600.0, there are higher chances for the borrower to be a DEFAULTER.

Results and Conclusions: For Continuous Variables

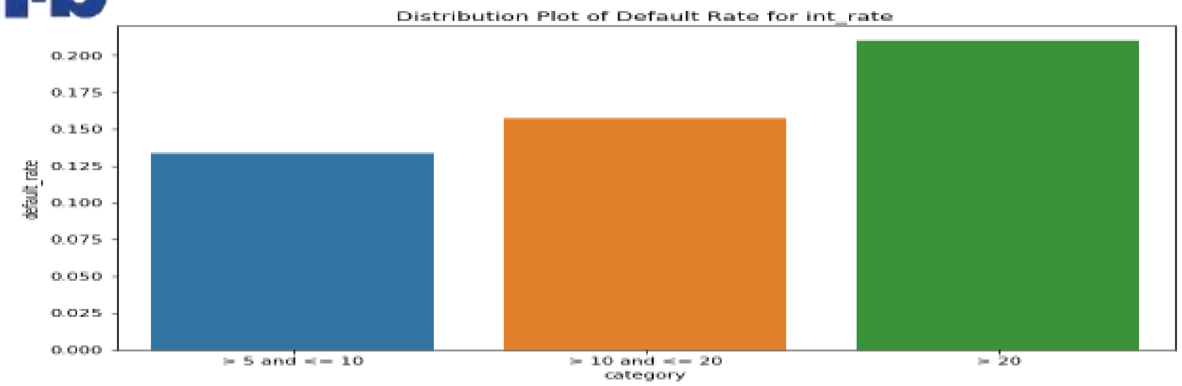
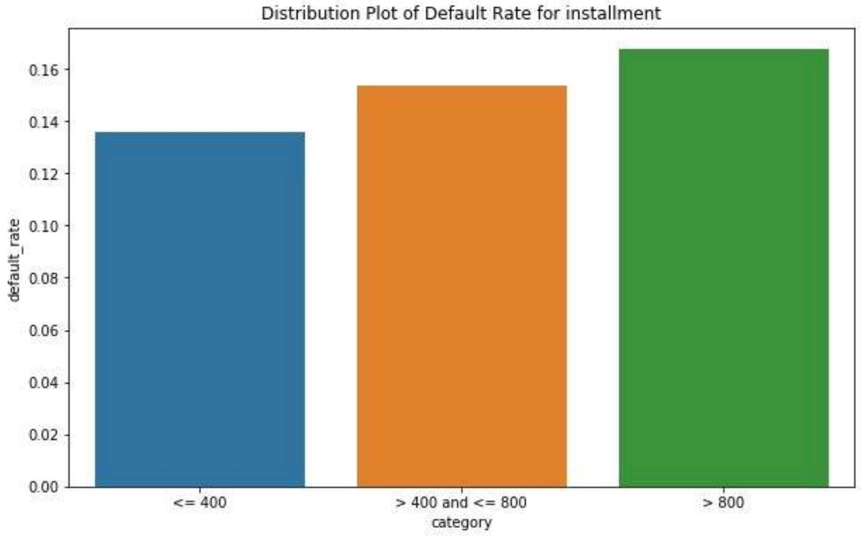


Fig 5 . It can be concluded that for ***int_rate*** > 20%, there are higher chances of the borrower to be a DEFAULTER.



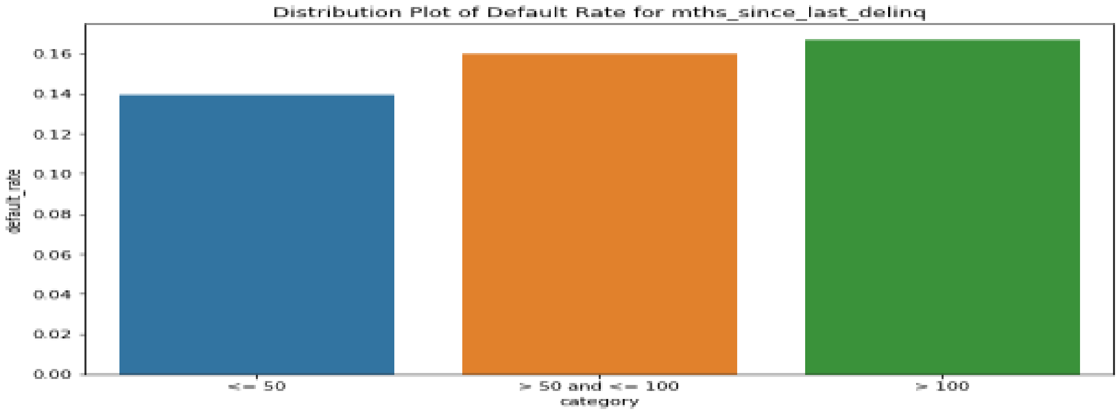


Fig 6. It can be concluded that for ***mths_since_last_delinq*** > 100, there are higher chances of the borrower to be a DEFAULTER.

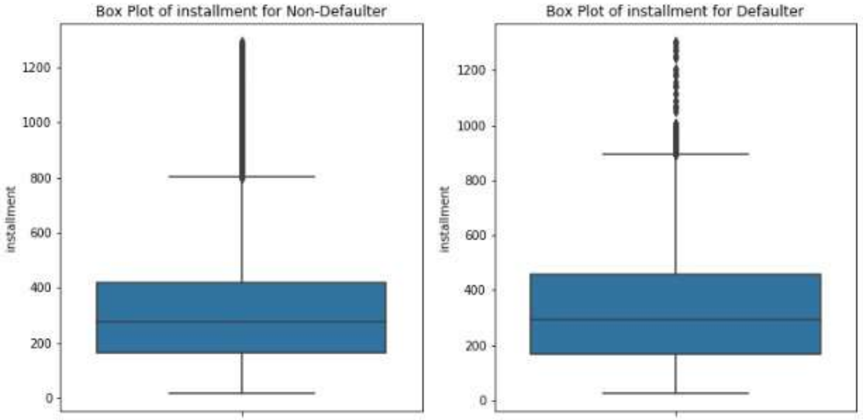


Fig 7. It can be concluded that if the ***installment*** > 800, there are higher chances for the borrower to be a DEFAULTER.

Results and Conclusions: For Continuous Variables

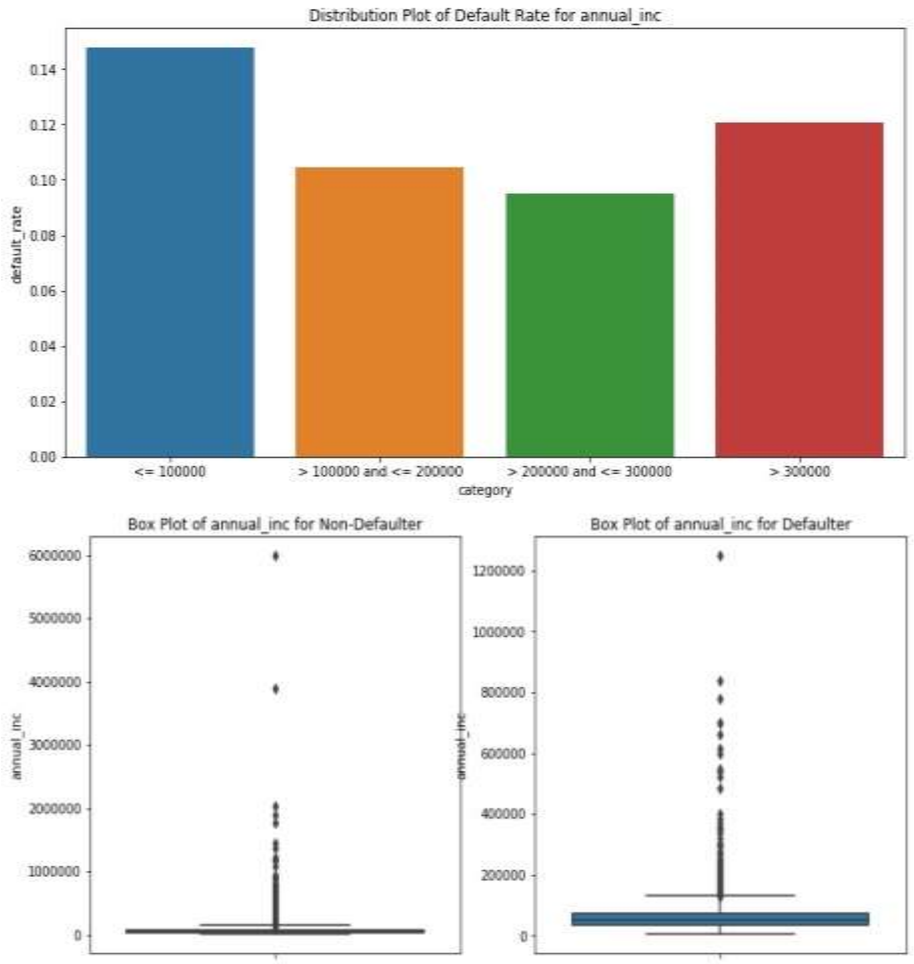


Fig 8. It can be concluded that for ***annual_inc*** <= 100000, there are higher chances of the borrower to be a DEFAULTER.

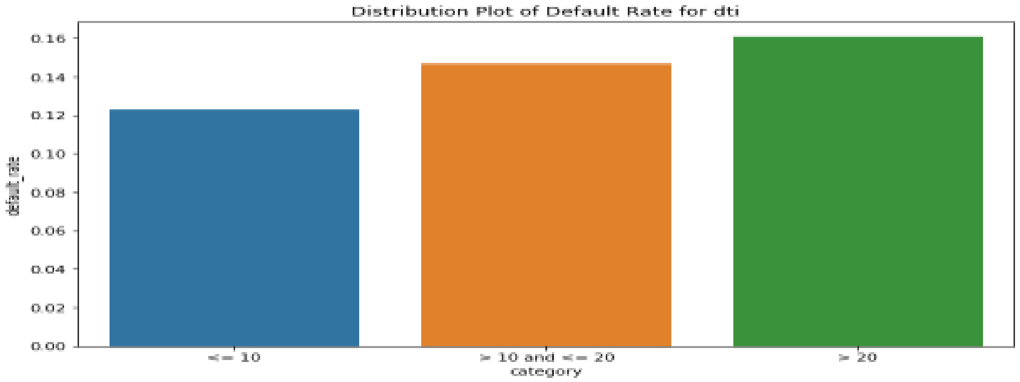


Fig 9. It can be concluded that for ***dti*** > 20, there are higher chances of the borrower to be a DEFAULTER.

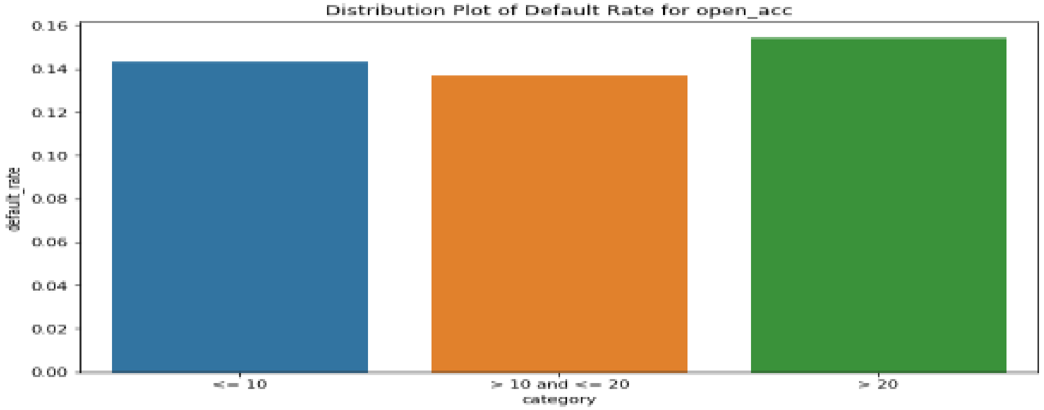


Fig 10. It can be concluded that for ***open_acc*** > 20, there are higher chances of the borrower to be a DEFAULTER.

Results and Conclusions: For Continuous Variables

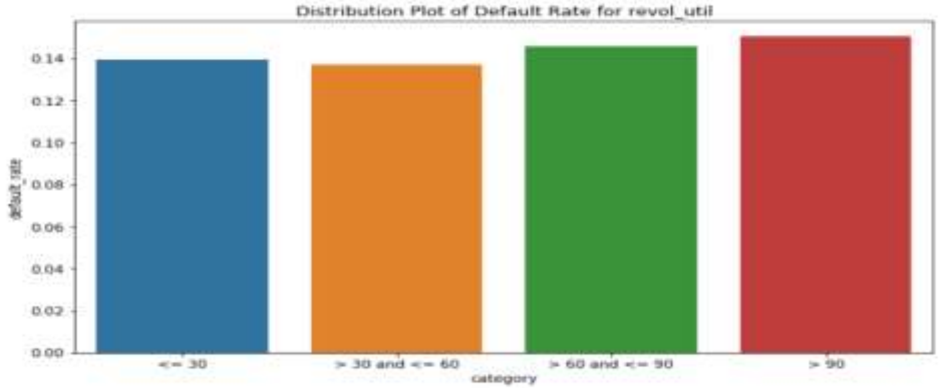


Fig 11. It can be concluded that for ***revol_util*** > 90, there are higher chances of the borrower to be a DEFAULTER.

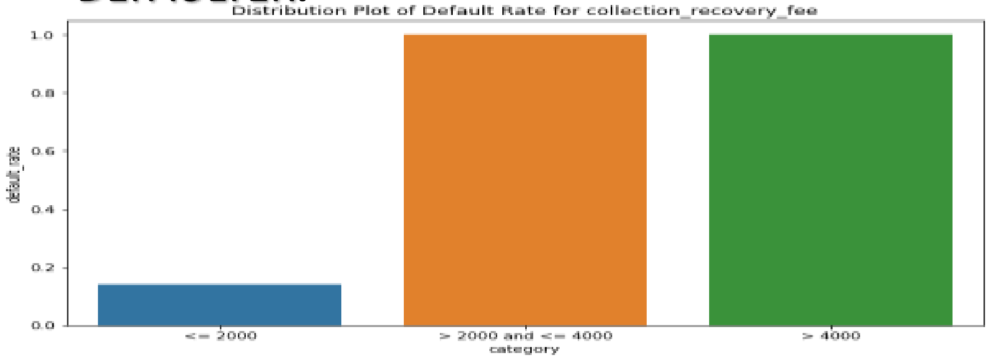


Fig 12. It can be concluded that for ***collection_recovery_fee*** > 2000, there are higher chances of the borrower to be a DEFAULTER

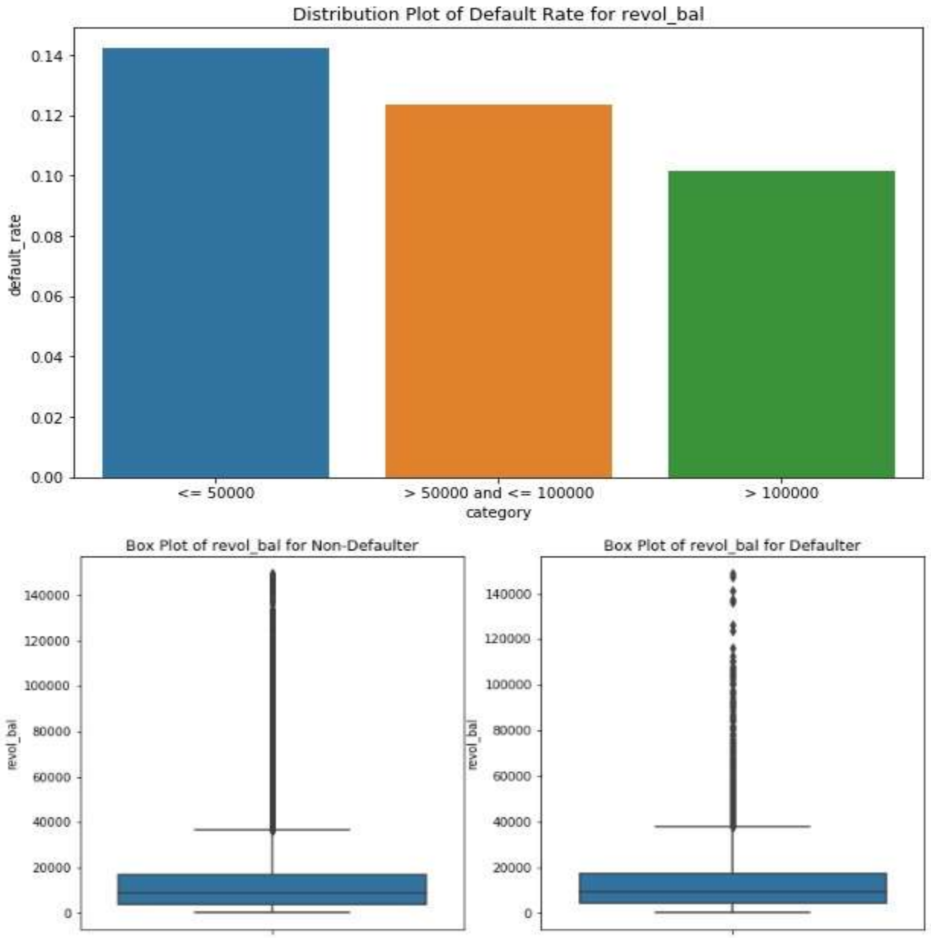


Fig 13. It can be concluded that if the ***revol_bal*** <= 50000, there are higher chances for the borrower to be a DEFAULTER.

Results and Conclusions: For Continuous Variables

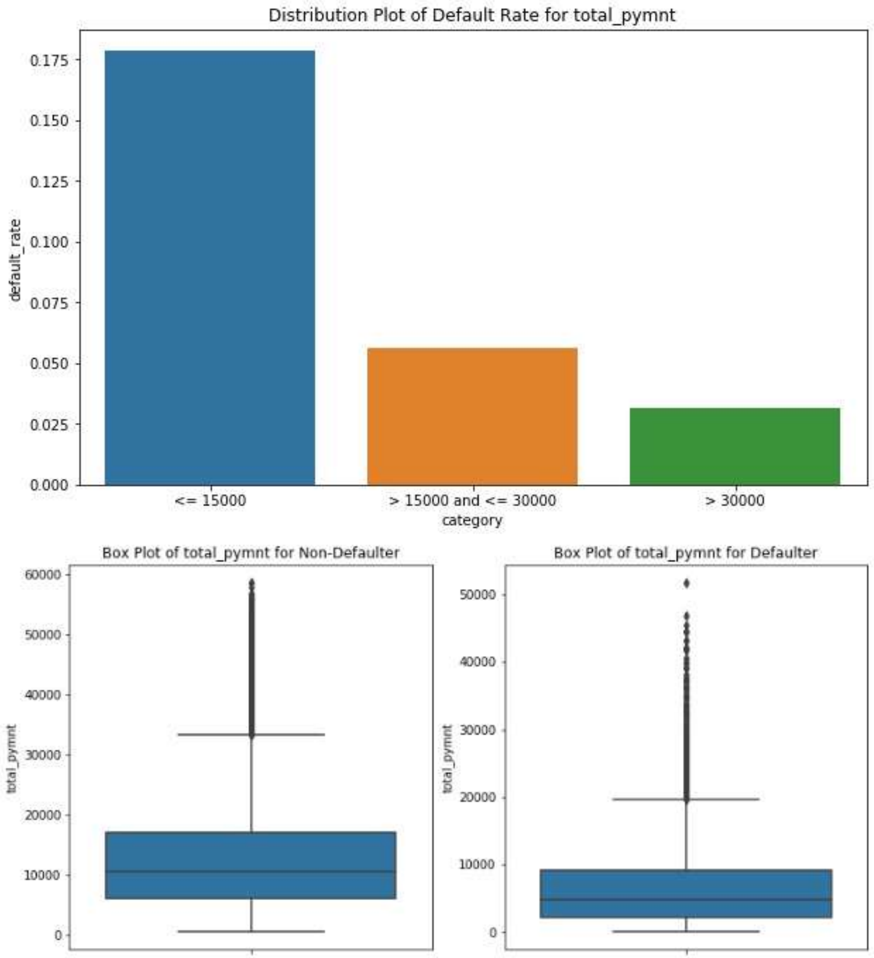


Fig 14. It can be concluded that for ***total_pymnt*** <= 15000, there are higher chances of the borrower to be a DEFAULTER.

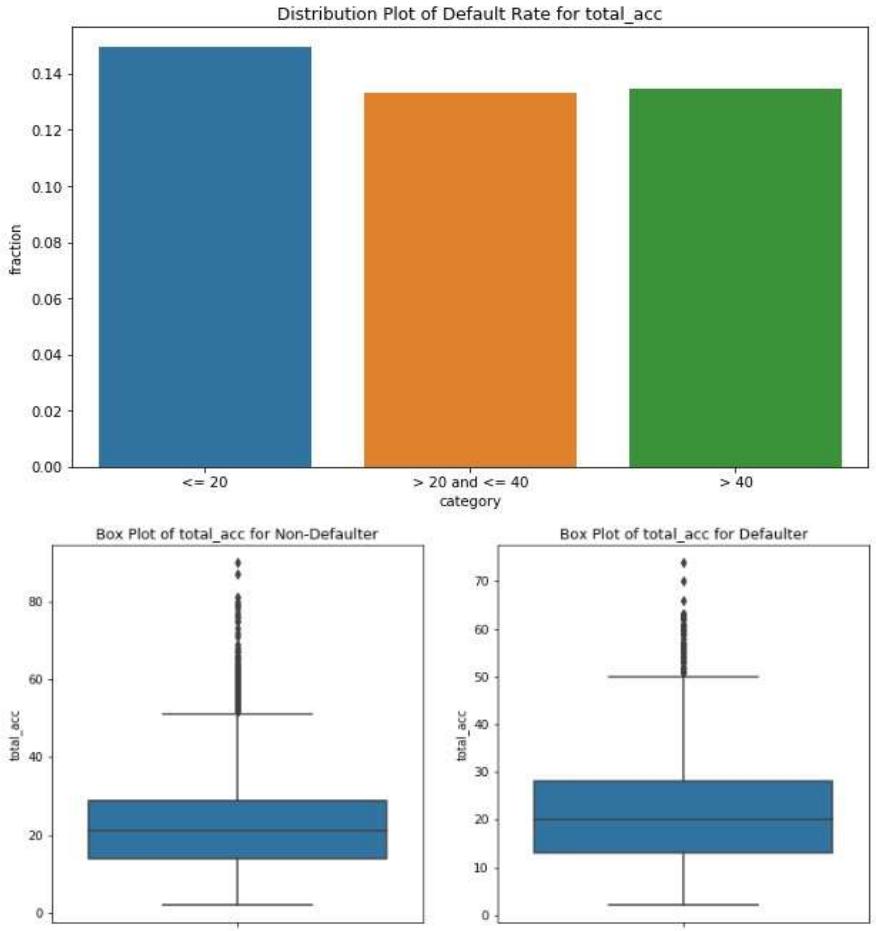


Fig 15. It can be concluded that for ***total_acc*** <= 20, there are higher chances of the borrower to be a DEFAULTER.

Results and Conclusions: For Continuous Variables

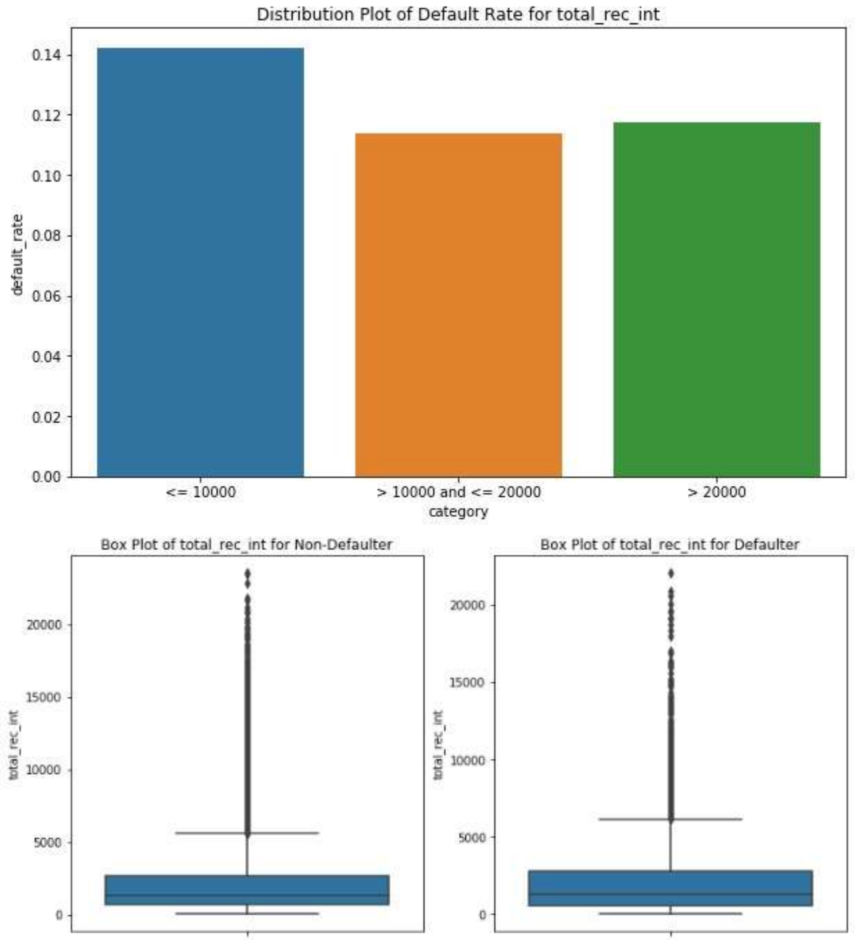


Fig 16. It can be concluded that if ***total_rec_int*** <= 10000, then there are higher chances that the borrower is a DEFAULTER.

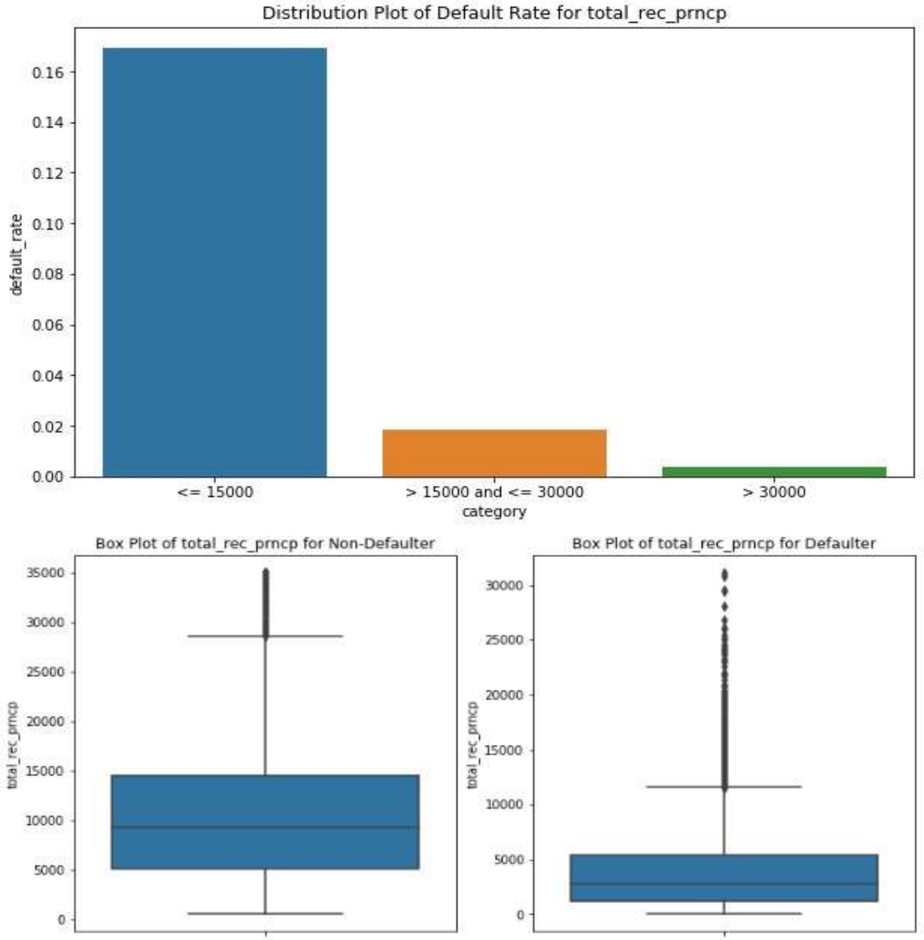


Fig 17. It can be concluded that if ***total_rec_prncp*** < 11577.76, then there are higher chances that the borrower is a DEFAULTER.

Results and Conclusions: For Continuous Variables

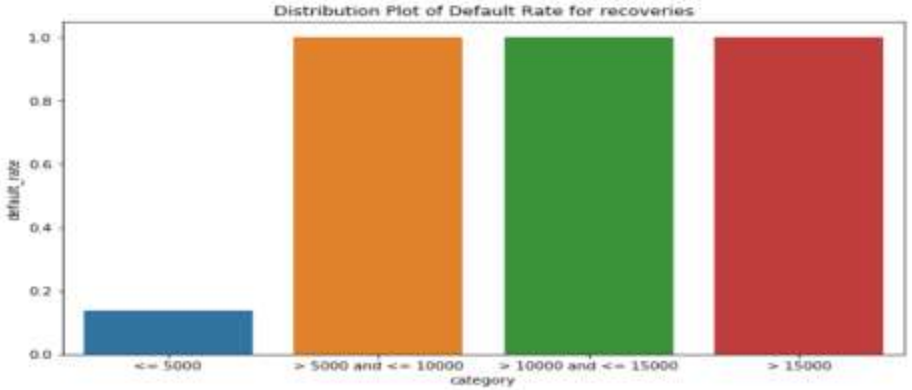


Fig 18. It can be concluded that for ***recoveries*** > 5000, there are higher chances of the borrower to be a DEFAULTER.

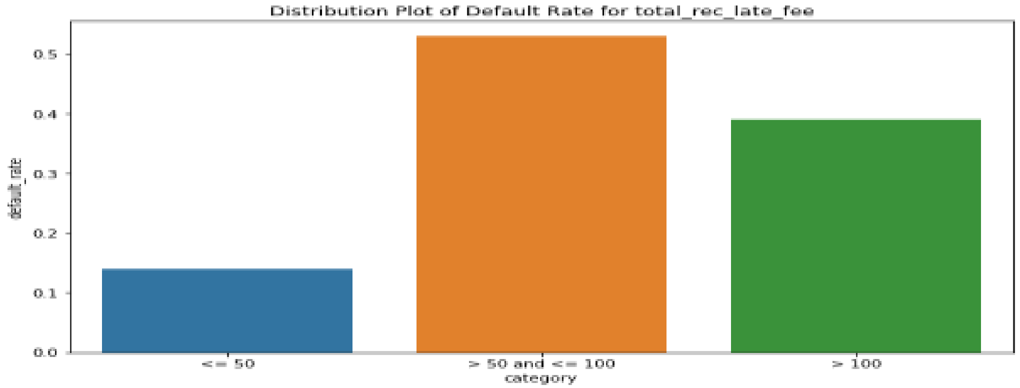
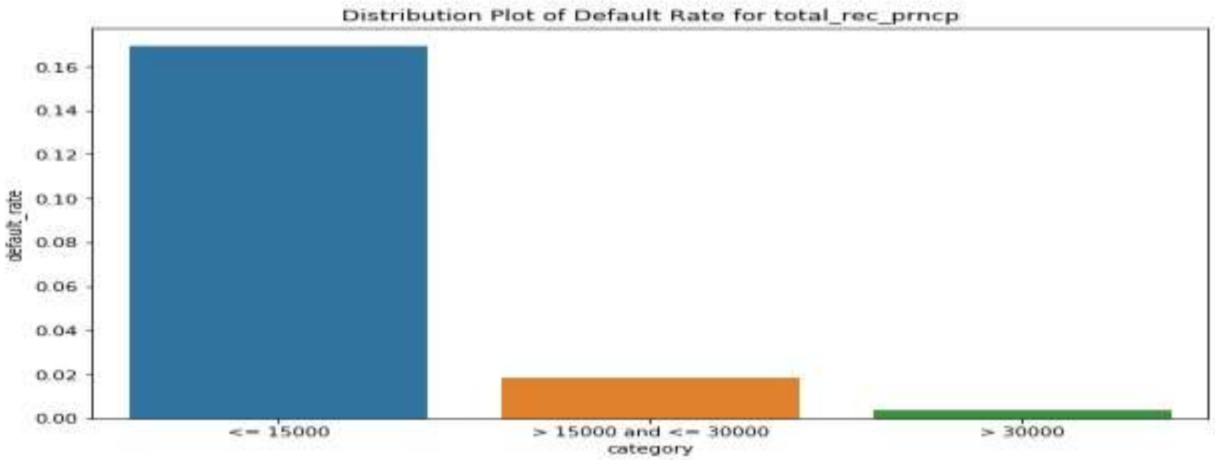
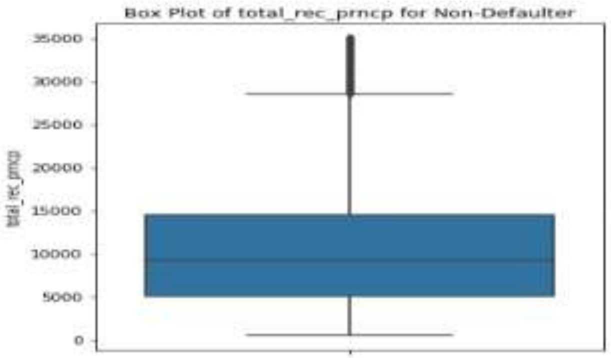


Fig 19. It can be concluded that for ***total_rec_late_fee*** in range 50-100 (100 included), there are higher chances of the borrower to be a DEFAULTER.





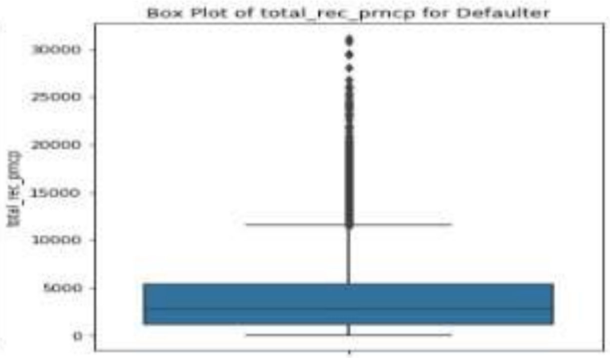
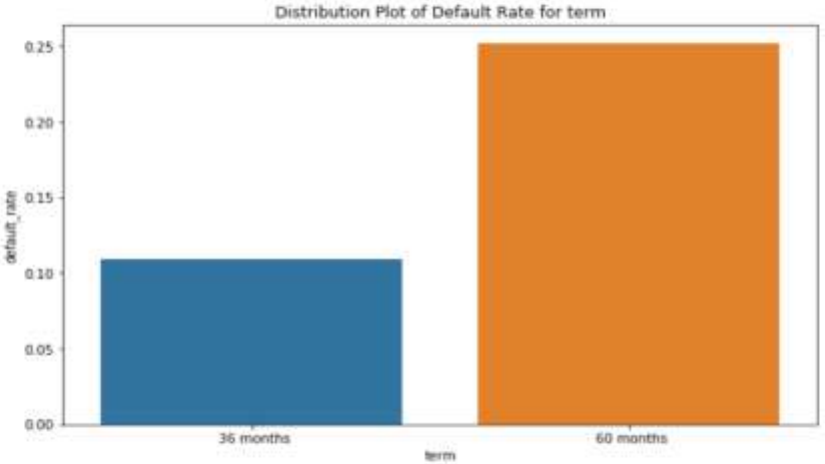


Fig 20. It can be concluded that if ***total_rec_prncp*** < 845.9 then there are higher chances that the borrower is a DEFAULTER.

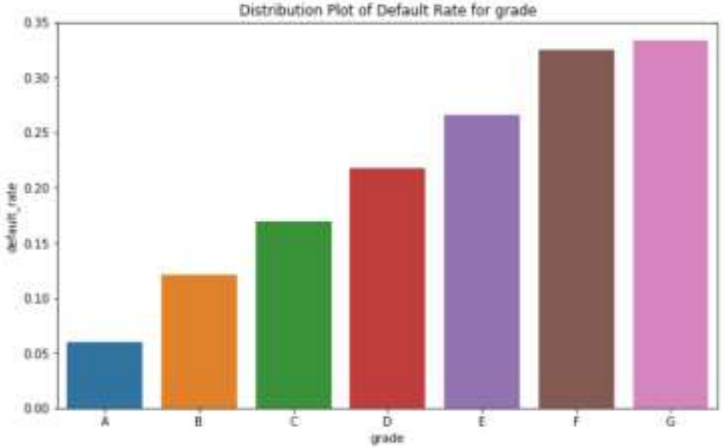
Results and Conclusions: For Categorical Variables

1. term



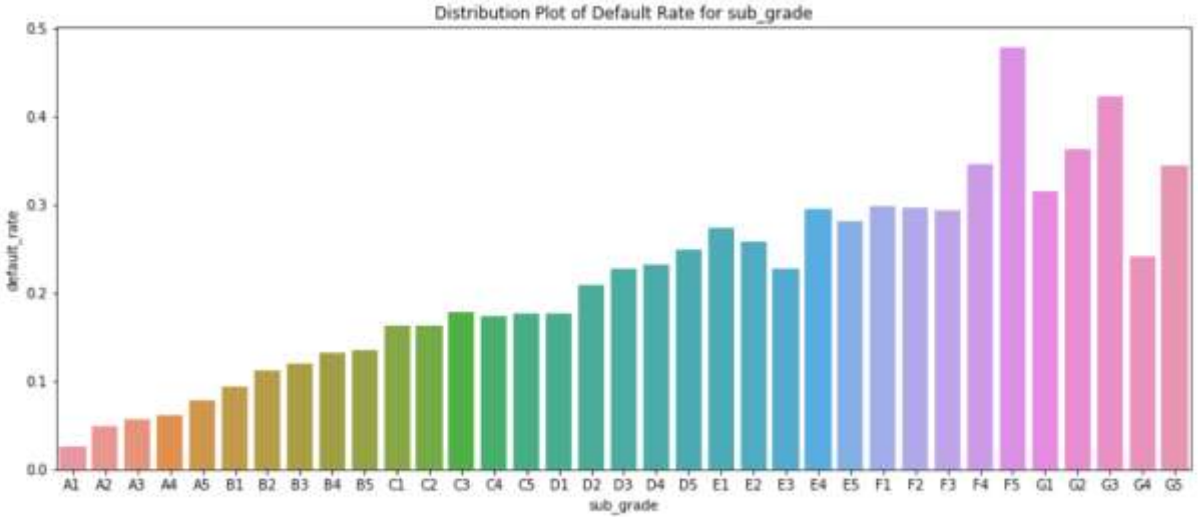
For a 60-month-term, there are higher chances of the borrower to be a Defaulter.

2. grade



For grade G, there are higher chances of the borrower to be a Defaulter.

3. sub_grade



For sub_grade F5, there are higher chances of the borrower to be a Defaulter.

4. emp_title

```
emp_title_analysis.sort_values(by = 'default_rate', ascending = False, inplace = True)
```

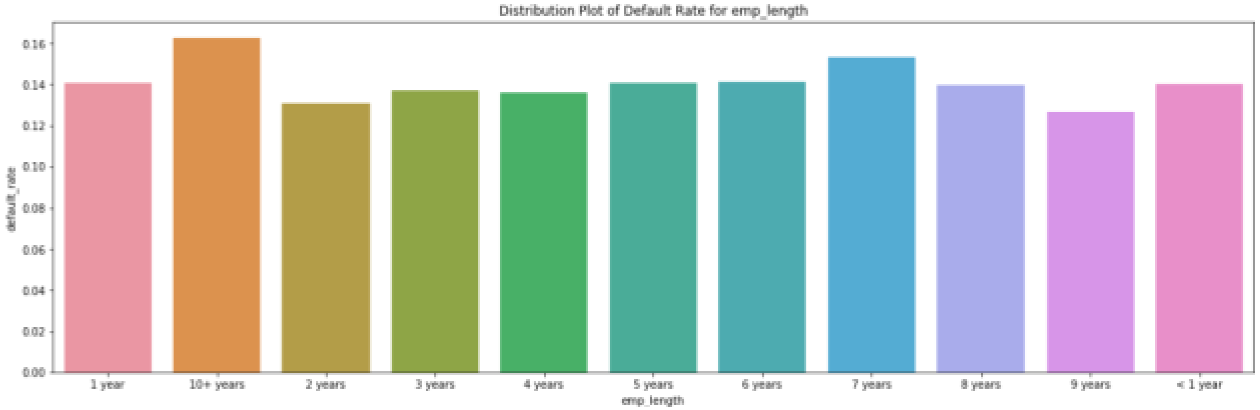
```
emp_title_analysis.head()
```

	emp_title	Not_Defaulter_Count	Defaulter_Count	default_rate
19578	UNITED STATES POSTAL SERVICE	1	3.0	0.750000
10991	Level 3 Communications	1	3.0	0.750000
13069	National Grid	1	3.0	0.750000
2414	Blockbuster	1	2.0	0.666667
23607	shaw group	1	2.0	0.666667

For UNITED STATES POSTAL SERVICE, Level 3 Communications and National Grid, there are higher chances of the borrower to be a Defaulter.

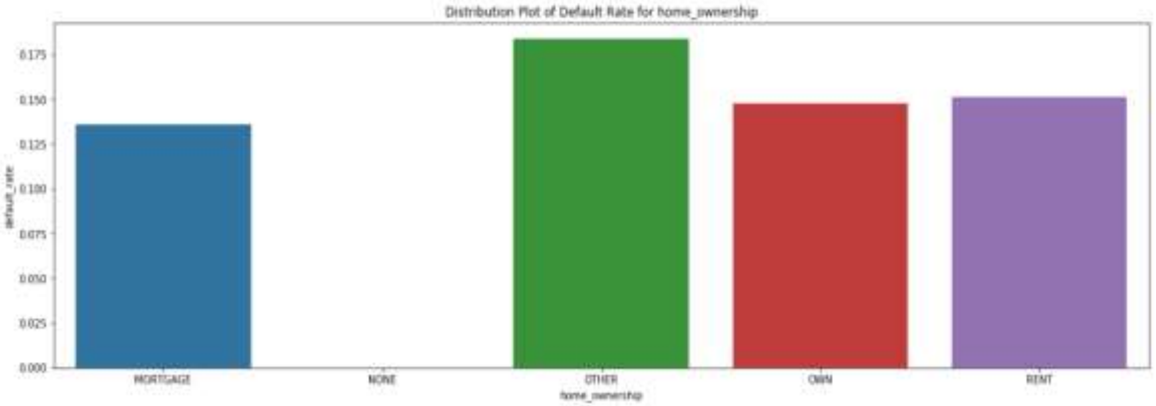
Results and Conclusions: For Categorical Variables

5. emp_length



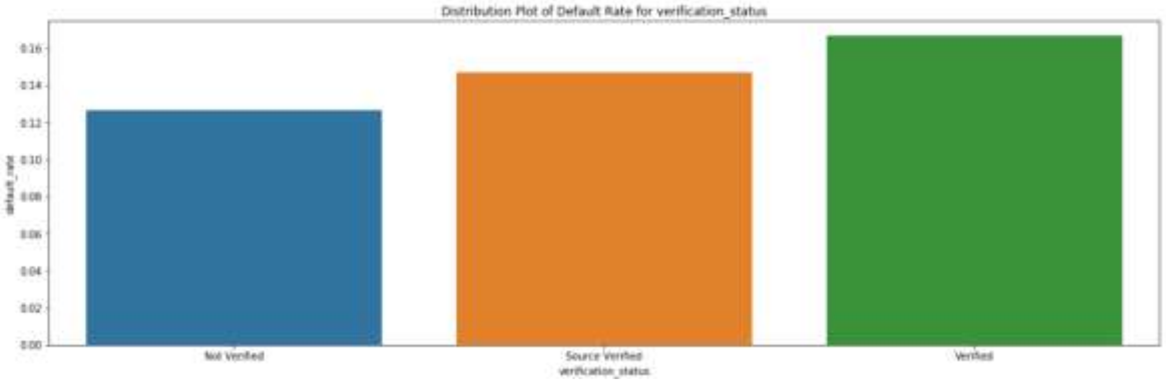
For 10+ years of service, there are higher chances of the borrower to be a Defaulter.

6. home_ownership



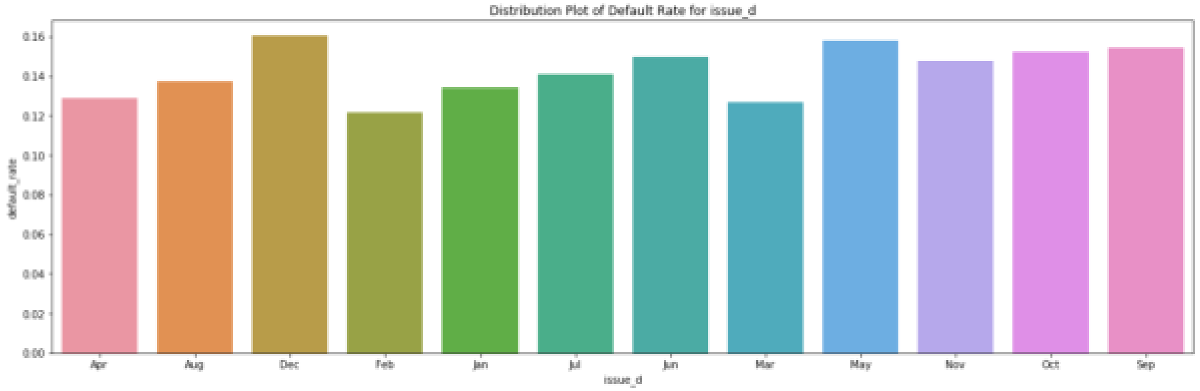
For home_ownership OTHER, there are higher chances of the borrower to be a Defaulter.

7. verification_status



For Verified status, there are higher chances of the borrower to be a Defaulter.

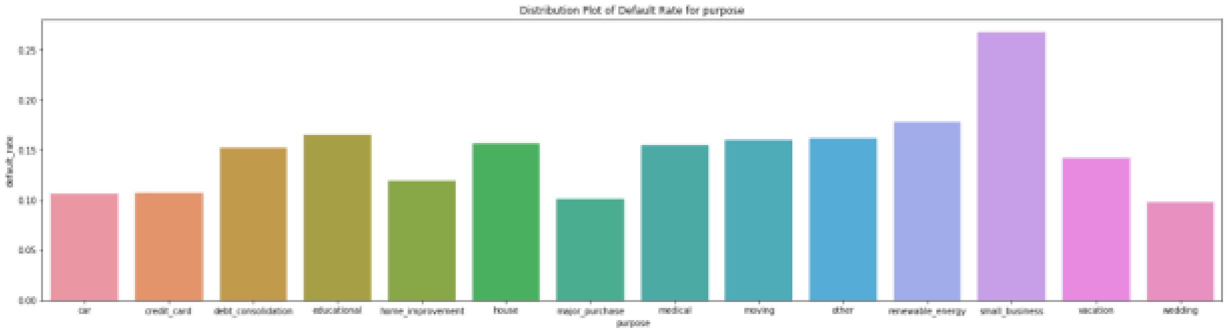
8. issue_d



For issue_month to be December, there are higher chances of the borrower to be a Defaulter.

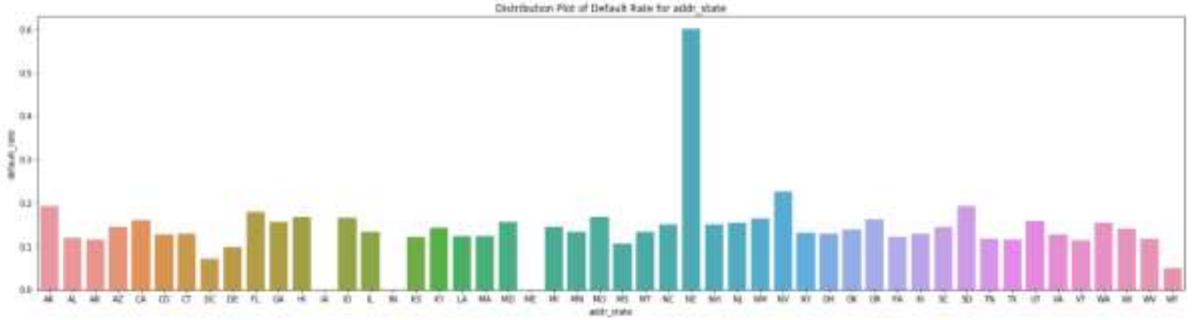
Results and Conclusions: For Categorical Variables

9. purpose



For small business, there are higher chances of the borrower to be a Defaulter.

10. addr_state



For addr_state NE, there are higher chances of the borrower to be a Defaulter.

11. zip_code

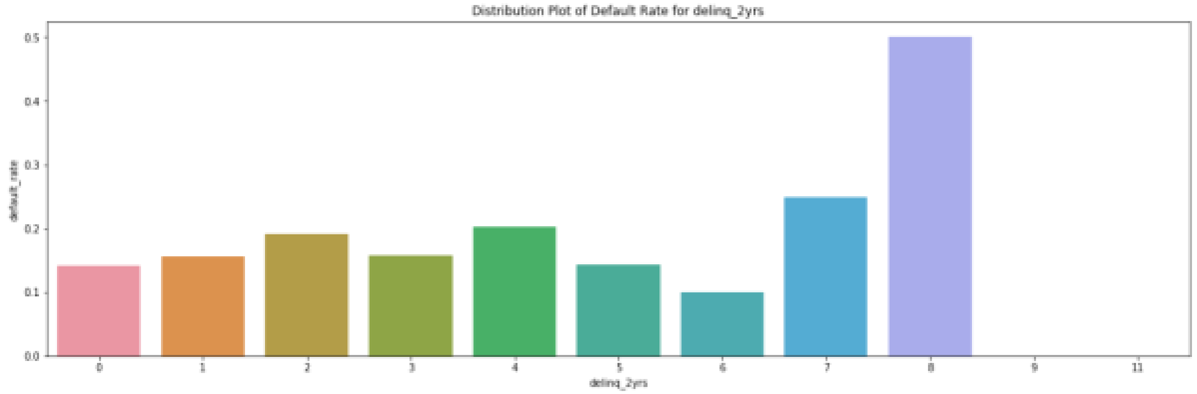
```
zip_code_analysis.sort_values(by = 'default_rate', ascending = False, inplace = True)
```

```
zip_code_analysis.head(5)
```

	zip_code	Not_Defaulter_Count	Defaulter_Count	default_rate
605	746xx	2	5.0	0.714286
468	561xx	2	5.0	0.714286
566	685xx	1	2.0	0.666667
433	496xx	3	3.0	0.500000
558	673xx	1	1.0	0.500000

For zip_code of 746xx and 561xx, there are higher chances of the borrower to be a Defaulter.

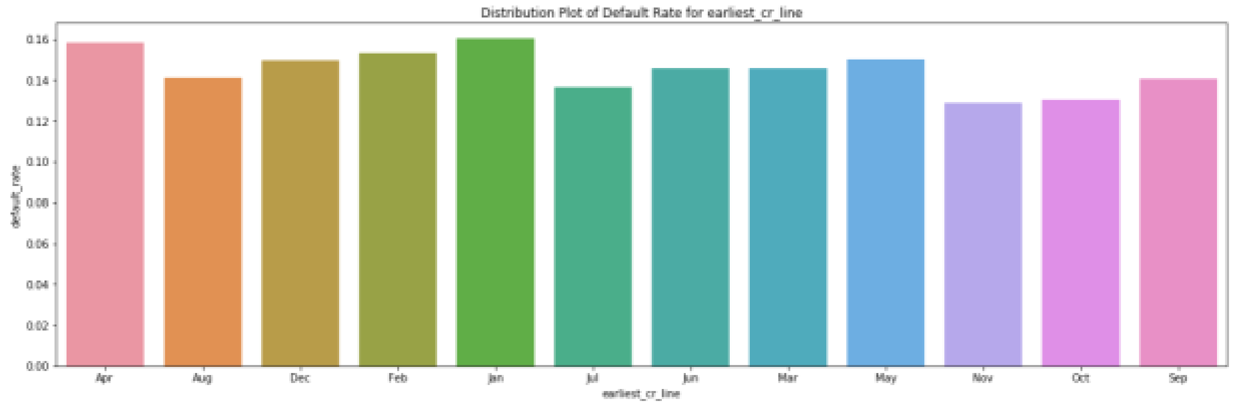
12. delinq_2yrs



For delinq_2yrs to be 8, there are higher chances of the borrower to be a Defaulter.

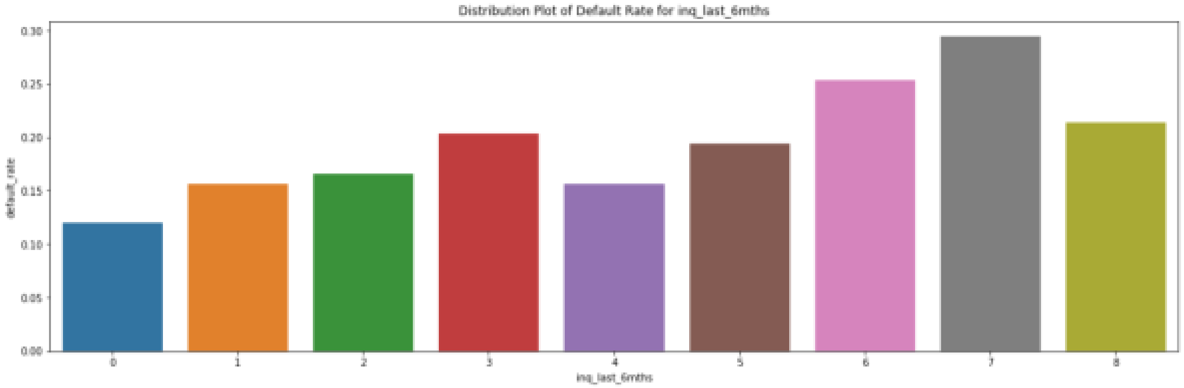
Results and Conclusions: For Categorical Variables

13. earliest_cr_line



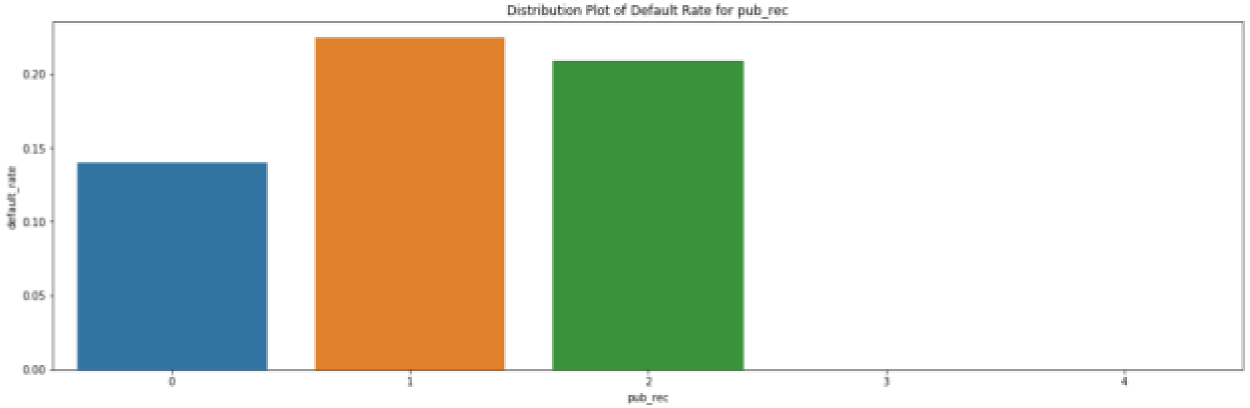
For earliest_cr_line in the month of January, there are higher chances of the borrower to be a Defaulter.

14. inq_last_6mths



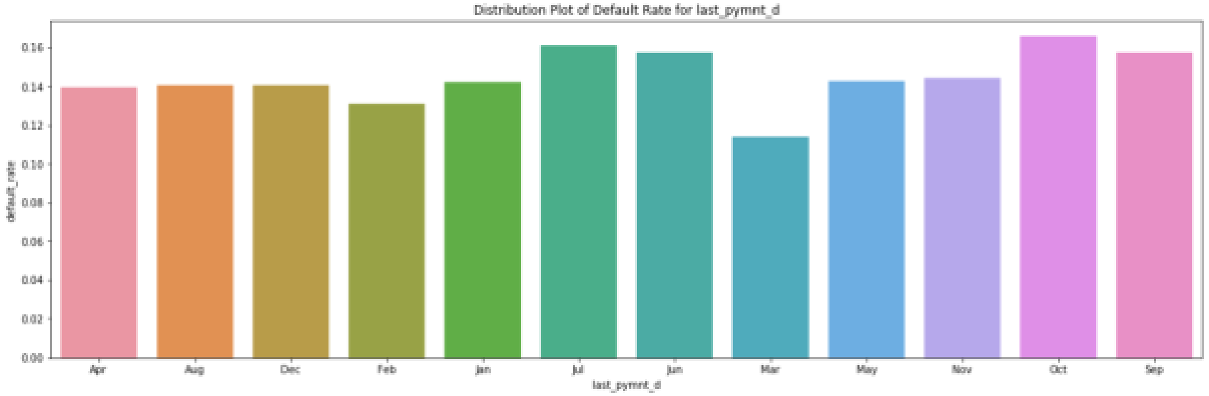
For inq_last_6mths to be 7, there are higher chances of the borrower to be a Defaulter.

15. pub_rec



For pub_rec of 1, there are higher chances of the borrower to be a Defaulter.

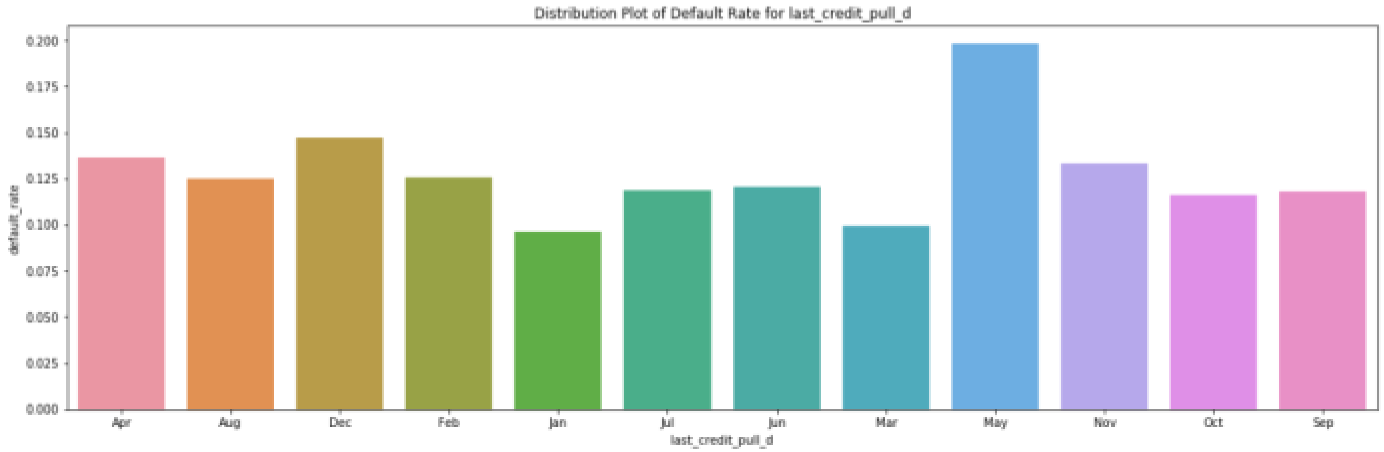
16. last_pymnt_d



For last_pymnt_d to be October, there are higher chances of the borrower to be a Defaulter.

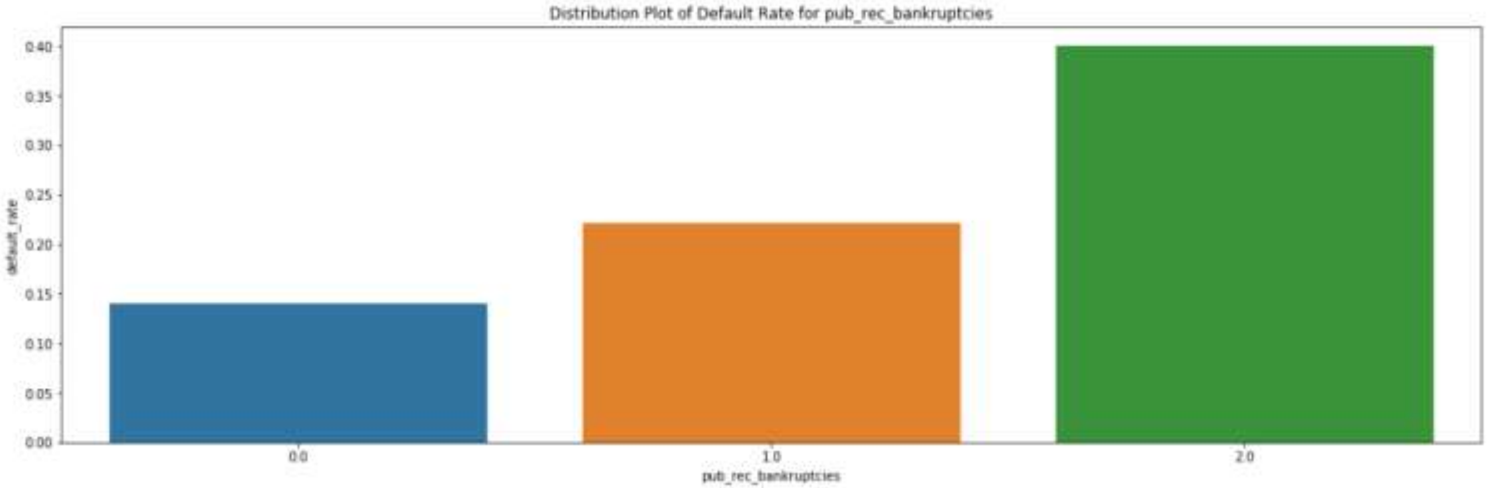
Results and Conclusions: For Categorical Variables

17. last_credit_pull_d



For last_credit_pull_d to be in the month of May, there are higher chances of the borrower to be a Defaulter.

18. pub_rec_bankruptcies



For pub_rec_bankruptcies of 2.0, there are higher chances of the borrower to be a Defaulter.