Probleam Statement:

- To harness Data Science to refine credit underwriting process.
- Focus on Personnel Loan Segement : Borrow Behaviour and Creditworthiness
  1. Financial Behaviour
  2. Spending Habits
  3. Potentials Risks for each borrower
- Objective is to :
  1. Optimize loan Disbursal
  2. Balancing Customer outreach with risk management.

Questionnaire

1. What percentage of customers have fully paid their Loan Amount?

Percentage would be (Fully Paid / Charged Off) = 318357 / 396030 = ~ 80.4%

```
loan_status
Fully Paid     318357
Charged Off     77673
Name: count, dtype: int64
```

```
df.shape
✓ 0.0s
(396030, 27)
```

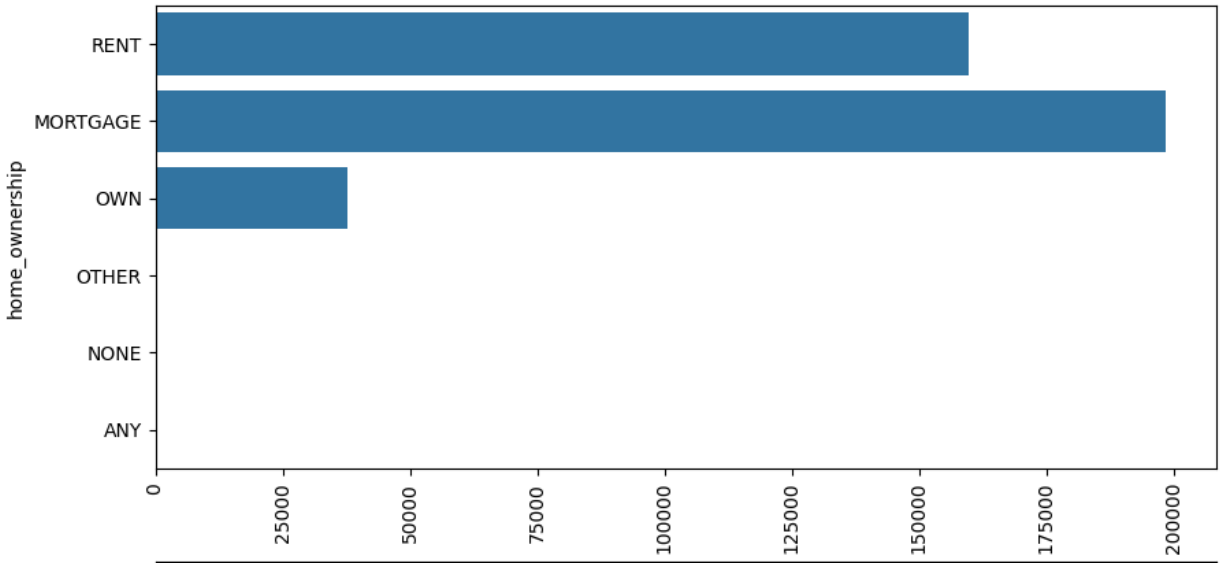
2. Comment about the correlation between Loan Amount and Installment features ?

The correlation coefficients we've obtained indicate a strong positive relationship between the two variables, both according to Pearson and Spearman correlation methods

```
Pearson Correlation: 0.953928908261621
Spearman Correlation: 0.9683337077962264
```

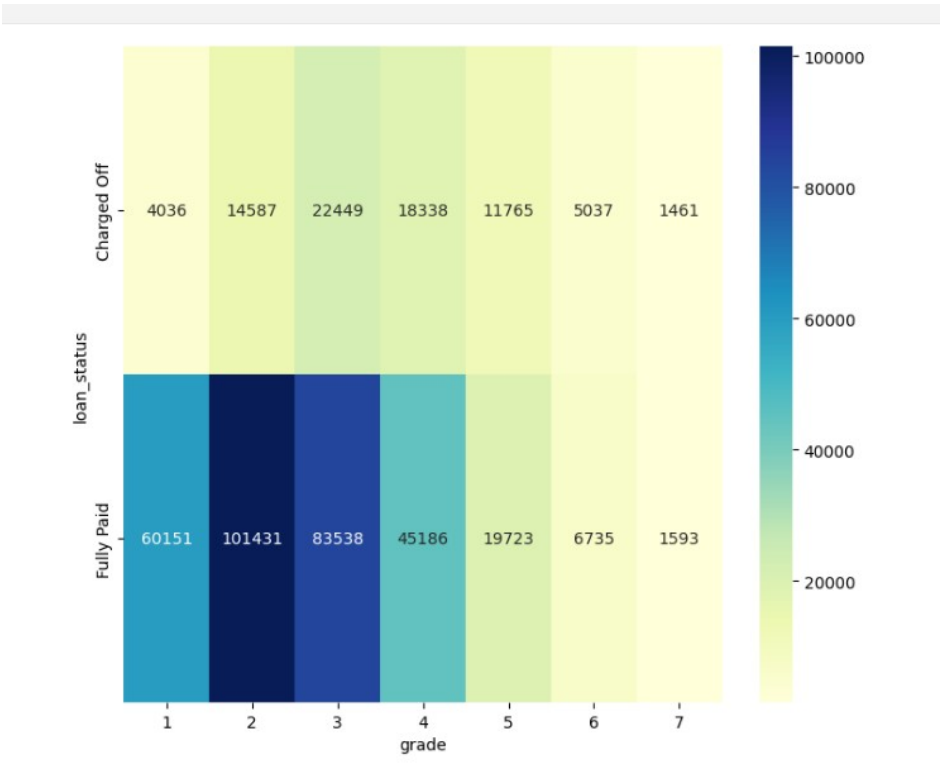
3. The majority of people have home ownership as ?

Looks like MORTAGE is majority in terms of home ownership.



4. People with grades ‘A’ are more likely to fully pay their loan. (T/F) ?

A is represented by 1 in below graph, by below analysis YES ‘A’ is most likely to fully pay the loan when we calculate the probability.



5. Name the top 2 afforded job titles.

Teacher and Manager have are the maximum count in terms of Fully Paid loan status

```
emp_title
Teacher                          3532
Manager                          3321
Registered Nurse                 1476
RN                               1467
Supervisor                       1425
                                 ...
Becton and Dickenson                1
EPI USE Labs                        1
Capital Sourcing Leader             1
lupient collision                   1
Seattle United Football Club        1
Name: count, Length: 145235, dtype: int64
```

6. Thinking from a bank's perspective, which metric should our primary focus be on..

- ROC AUC : ROC AUC measures the ability of the model to discriminate between repaid and defaulted loans across various thresholds. But as we've significant class imbalance (e.g., more repaid loans than defaulted ones), precision, recall, and F1 score might be more informative than ROC AUC.

- Precision : Precision is the proportion of approved loans that are genuinely repaid. It is relevant when the cost of approving a loan that might not be repaid (false positive) is high. Hence it is important

- Recall : Recall is the proportion of actual repaid loans that are correctly identified. It is relevant when the cost of missing a genuinely repaid loan (false negative) is high. Hence it is important.

- F1 Score : F1 Score provides a balance between precision and recall. It is suitable when there is a need to balance the trade-off between false positives and false negatives. So yes it is important matrics as well.

7. How does the gap in precision and recall affect the bank?

- High Precision, Low Recall: The bank is conservative in approving loans. The loans approved are likely to be repaid, but there's a risk of rejecting genuinely creditworthy applicants (false negatives). This approach minimizes the risk of approving risky loans but may lead to missed opportunities.

- Low Precision, high Recall: The bank is more lenient in approving loans. While capturing a larger portion of genuinely creditworthy applicants, there's a risk of approving loans that may not be repaid (false positives). This approach increases the chances of approving loans but may lead to higher default rates.

- Balanced Precision and Recall : The bank strikes a balance between minimizing the risk of approving risky loans and maximizing the approval of genuinely creditworthy applicants. The trade-off is optimized to meet the bank's risk tolerance and business objectives.

- Financial Impact :
  Low Precision : May lead to financial losses due to defaulted loans that were incorrectly approved.
  Low Recall : May result in missed opportunities and reduced revenue from potentially creditworthy applicants.

8. Which were the features that heavily affected the outcome?

Top 5 features include:
1. Pincode
2. Grade
3. Term
4. Annual Income
5. Debt-to-Income ratio (DTI)

9. Will the results be affected by geographical location? (Yes/No)

Yes, it seems that Geographical location impacts the outcome of the model.

Importance of Features:

	Feature	Absolute_Coefficient
17	pincode	0.831633
3	grade	0.495296
0	term	0.264602
19	annual_inc	0.182345
7	dti	0.159250
21	mort_acc	0.122457
4	sub_grade	0.086989
18	loan_amnt	0.070384
25	verification_status_Not Verified	0.055686
24	revol_util	0.051215
13	issue_year	0.050802
23	total_acc	0.042139
8	open_acc	0.042000
22	revol_bal	0.012679
20	emp_length	0.007921
26	verification_status_Source Verified	0.004513
14	earliest_cr_month	0.000000
11	pub_rec_bankruptcies	0.000000
10	initial_list_status	0.000000
9	pub_rec	0.000000
1	int_rate	0.000000
5	home_ownership	0.000000
...		
27	verification_status_Verified	0.000000
28	emp_duration_type_equalsto	0.000000
29	emp_duration_type_greaterthan	0.000000
30	emp_duration_type_lessthan	0.000000

END