Microsoft 365

Lending Case Study

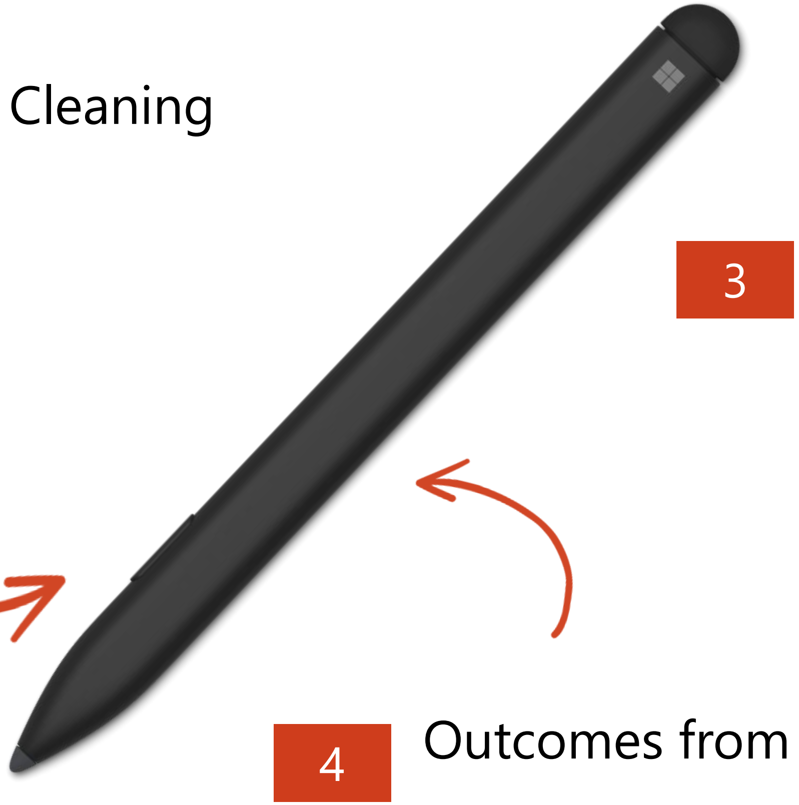
Agenda

1 Understand the Problem Statement

2 Data Understanding

&

Data Cleaning



3 Univariate and Bivariate Analysis

4 Outcomes from case study

Problem Statement

Lending Club was founded on the idea of disrupting the traditional banking industry by connecting borrowers directly with investors.

Lending Club is a peer-to-peer lending platform that connects borrowers with investors, essentially bypassing traditional banks.

Analyze the given dataset to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment on the basis of univariate and bivariate analysis results.

Analysis Steps:

- **Data Cleaning:** Check for missing or corrupt data and remove it if necessary. Also, remove outliers or data that doesn't belong.
- **Data Exploration:** Examine the distribution of the variable by plotting it as a histogram, box plot, or density plot. Calculate basic statistics such as mean, median, and standard deviation.
- **Univariate Visualization:** Choose the best visualization that represents the data well. Common univariate visualizations include histogram, density plot, box plot, and bar chart.
- **Bivariate Visualization:** Choose the best visualization that represents the relationship between the two variables well. Common bivariate visualizations include scatter plot, line plot, and heat map.
- **Interpret the Visualization:** Interpret the visualization by examining the relationship between the two variables. Look for patterns, trends, and correlations.

Data Cleaning:

- Given dataset containing 111 columns.
- We dropped the columns which were having more than 50% nulls and were left with 54 columns.
- Still, we had 11 columns left, which contain null values.
- For non- numeric columns: nulls were replaced with dummy placeholder ‘Missing value’
- For numeric columns: We tried using KNN imputer.

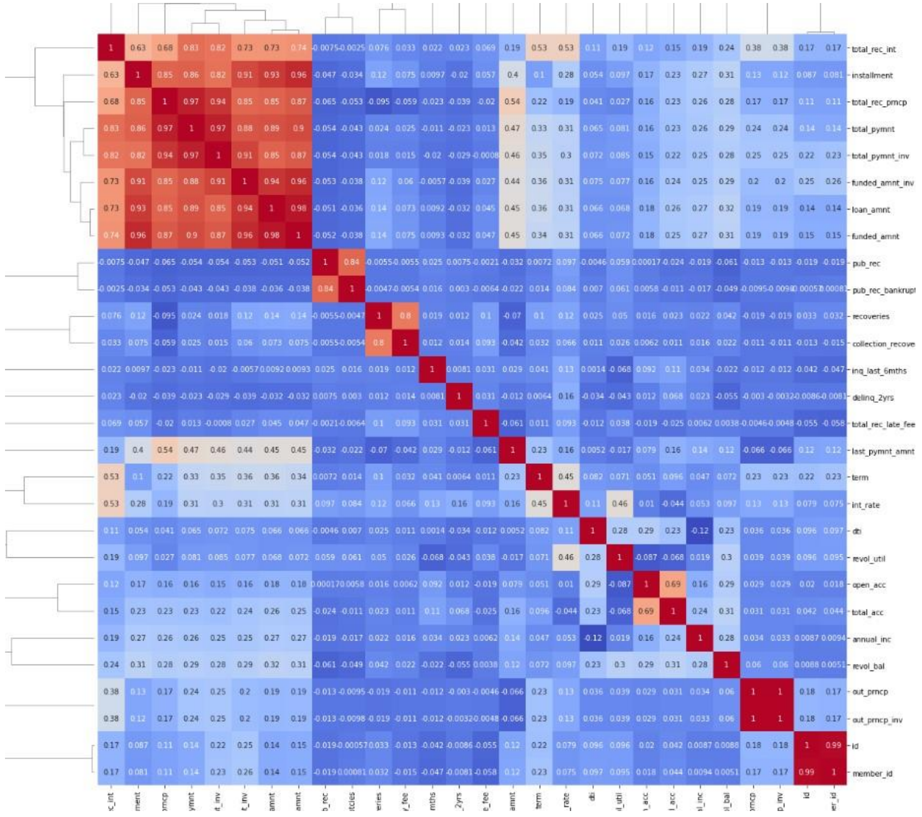
But we found that for column ‘pub_rec_bankruptcies’, most people are getting assigned a value of 1, which we don't think should be correct as most people would not have a single bankruptcy. So, we believe that we are losing a lot of data when we select only numeric features to compute using KNN imputer. Then, we settled with filling missing values with median only.

Data Cleaning:

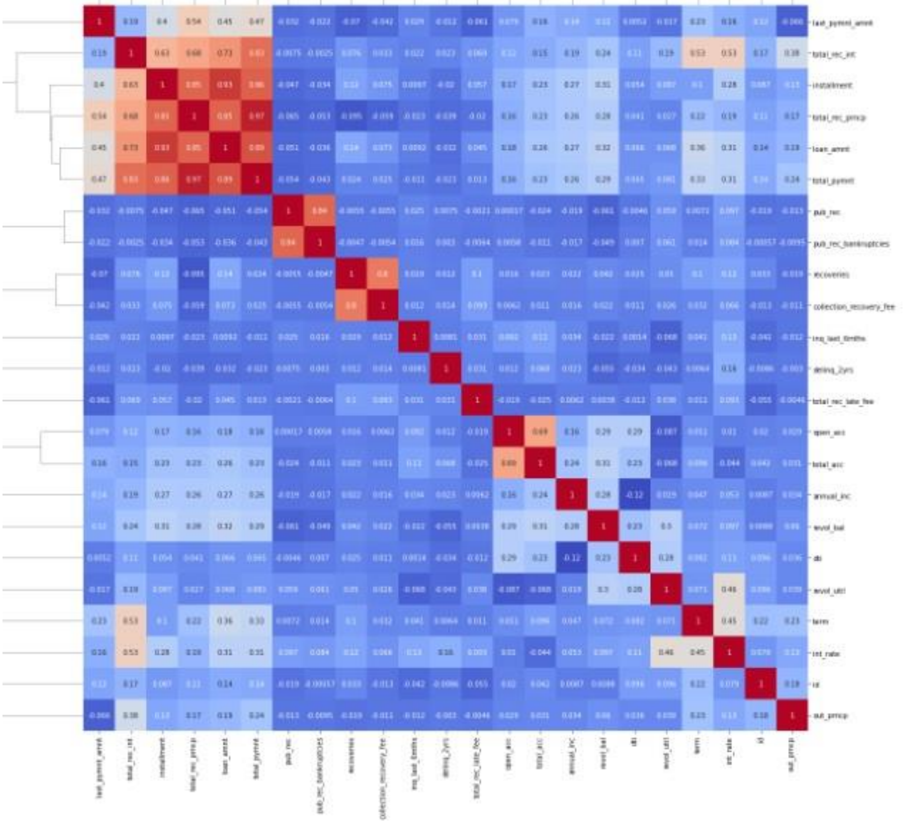
- Then we dropped the columns having only one value
- Fixed the data type for interest rate col(removing %) and term(removing ‘months’).
- We analyzed that there is a huge inconsistency in data contained in df2(filter dataset where Principal Remaining should be 0 in case of loan_status Fully Paid). When we subtract the loan amount with total principal paid, the difference remains vary large in above shown cases. We might be reading something wrong. So, for the sake of keeping the analysis consistent, we added new column 'prncp_rem' and assigned prncp_rem as 0 when the loan is marked as fully paid.
- We removed url, desc zip_code column as well, which seems not having relevant data to be analyzed.
- Finally, we left with 38 columns for analyzing the data.
- We filtered on loan_status and kept only ‘Fully Paid’ and ‘Charged off’ rows.

Bivariate Analysis for dropping highly correlated columns

- Tried finding and dropping highly corelated columns using clustermap.

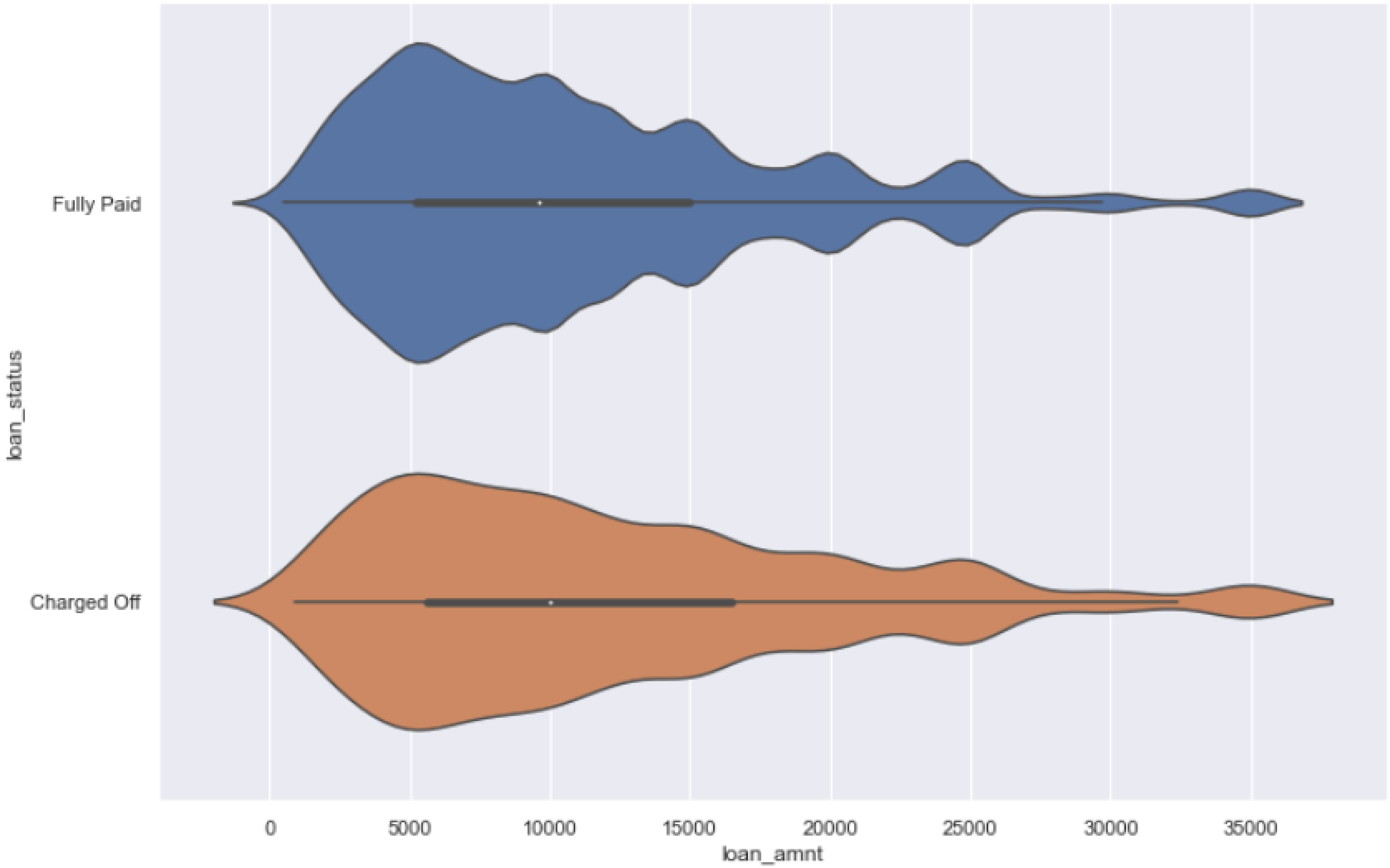


Before



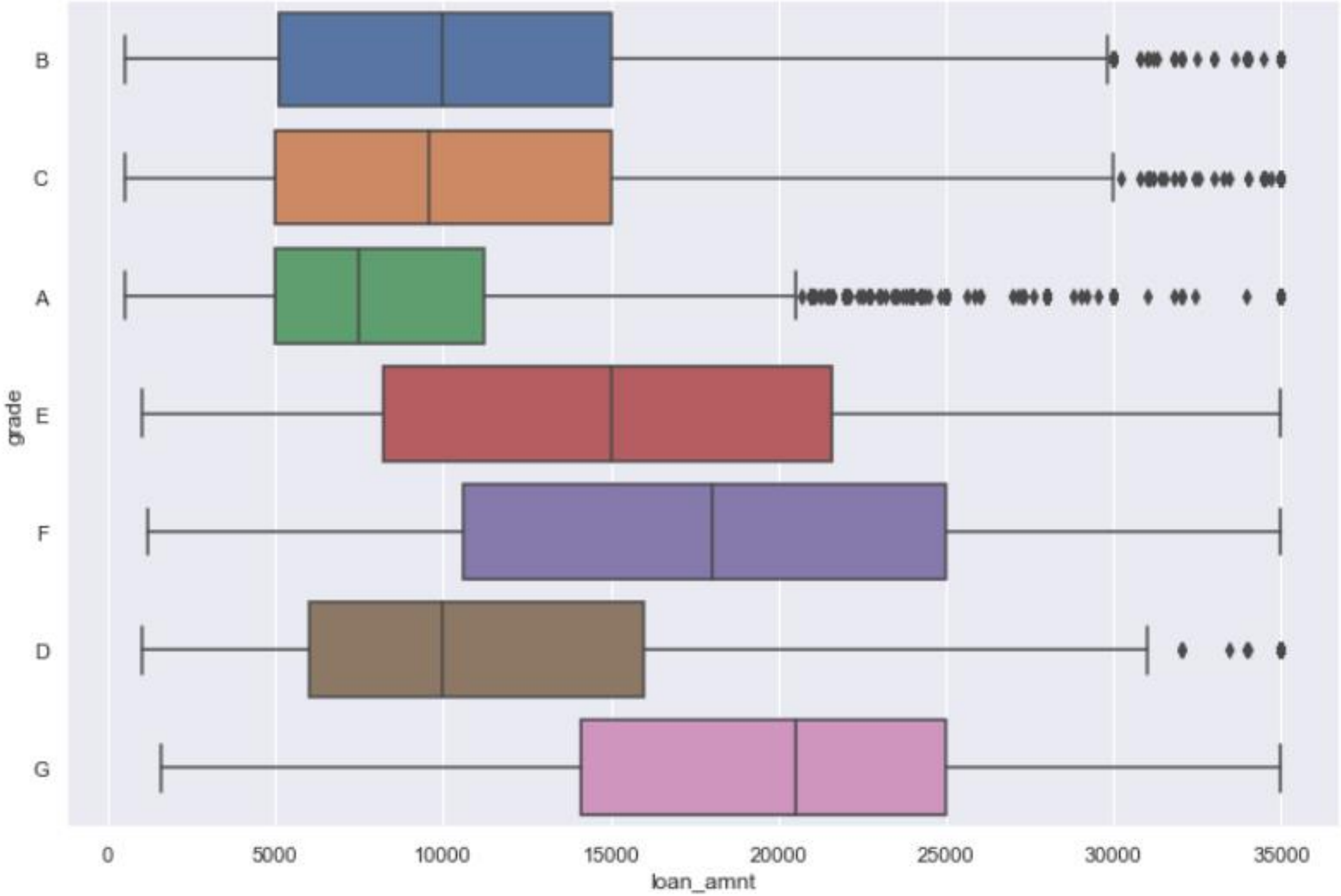
After dropping 5 more highly correlated columns

Univariate Analysis



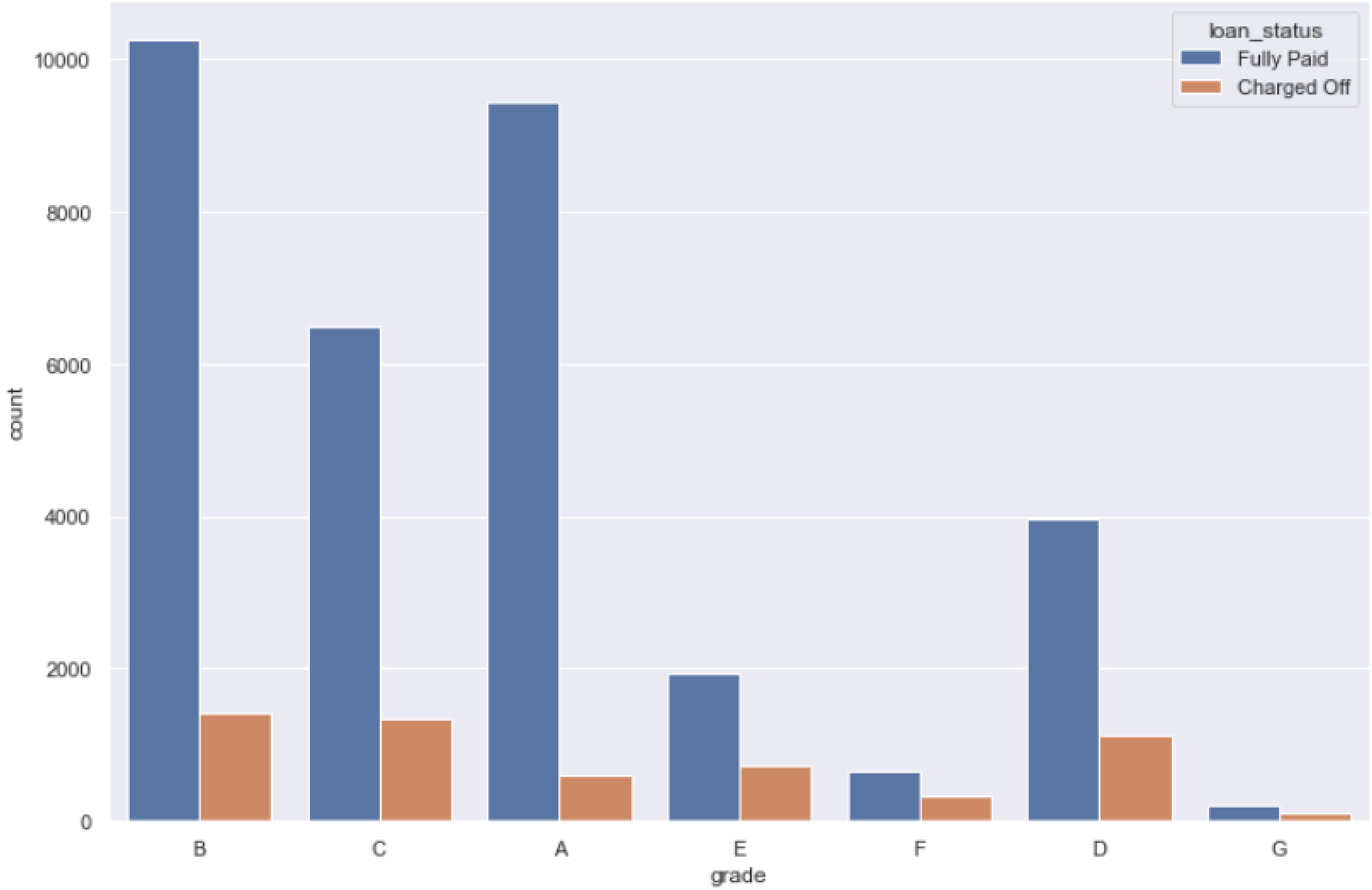
Observation *Loan amount does not seem to paly any role in the people defaulting the loan, with histograms and quantiles nearly overlapping*

Univariate Analysis



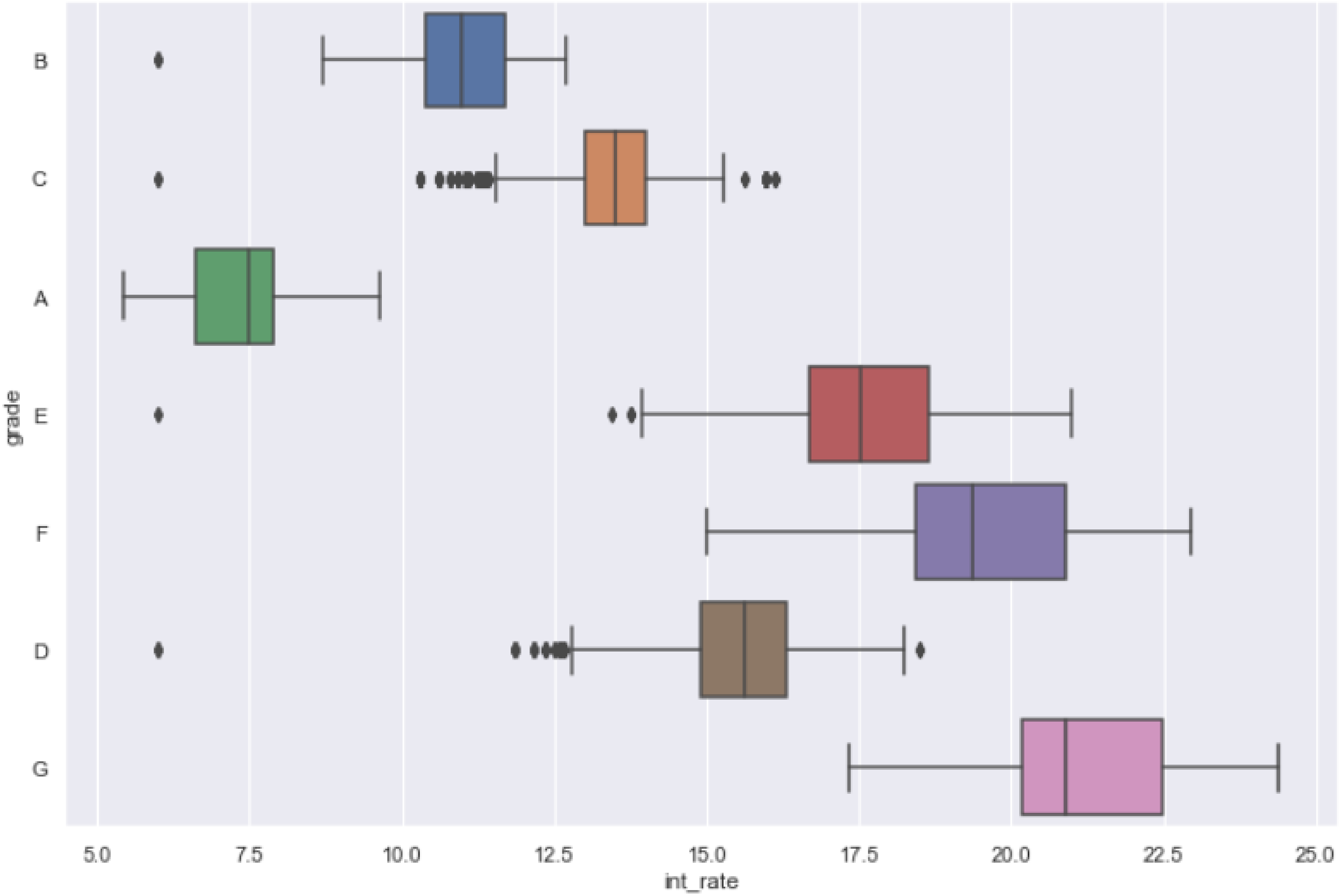
Observation *Loan grade is not assigned based on the total loan amount*

Univariate Analysis



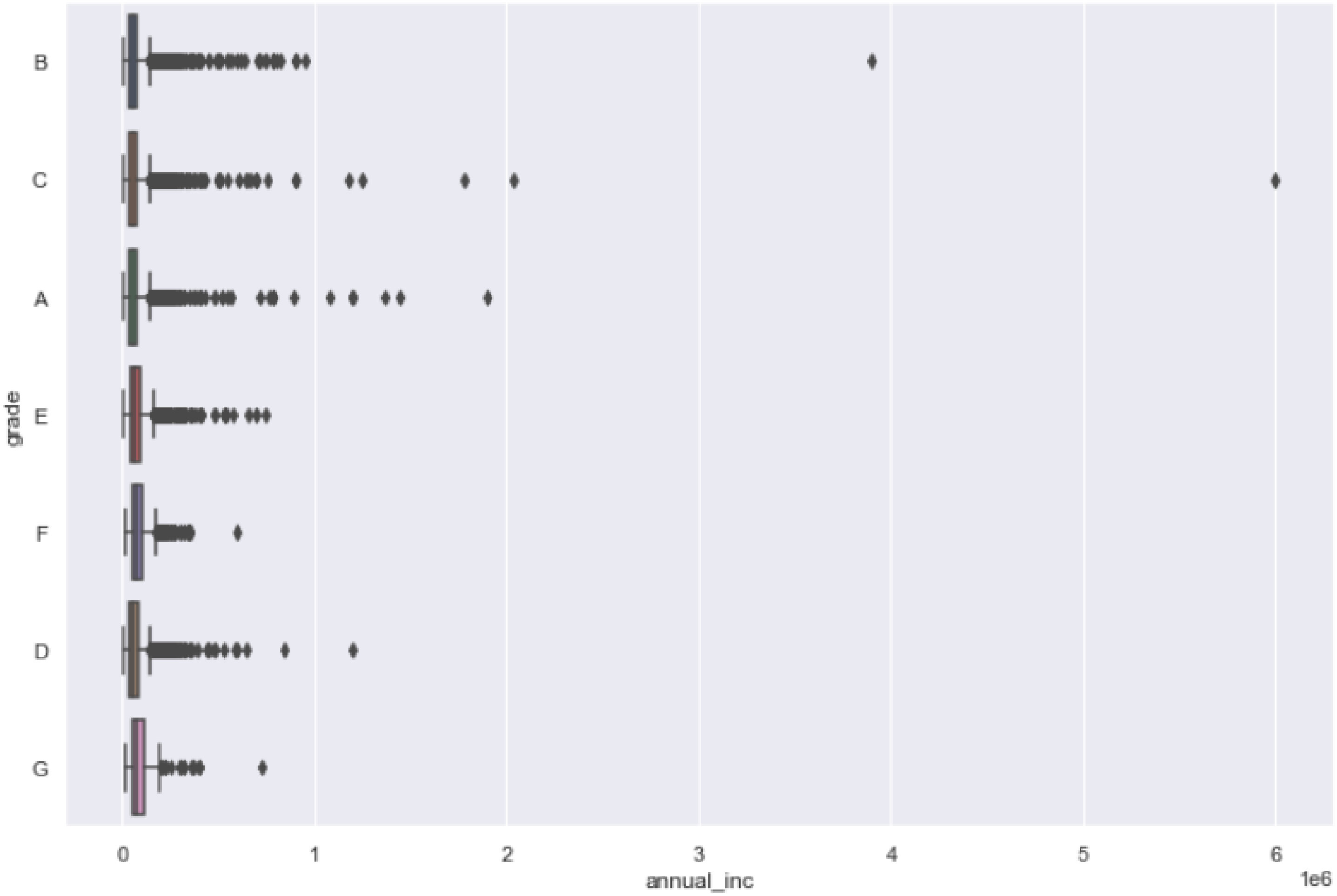
Observation *Higher graded loans have more chance of default in %age terms.*

Univariate Analysis



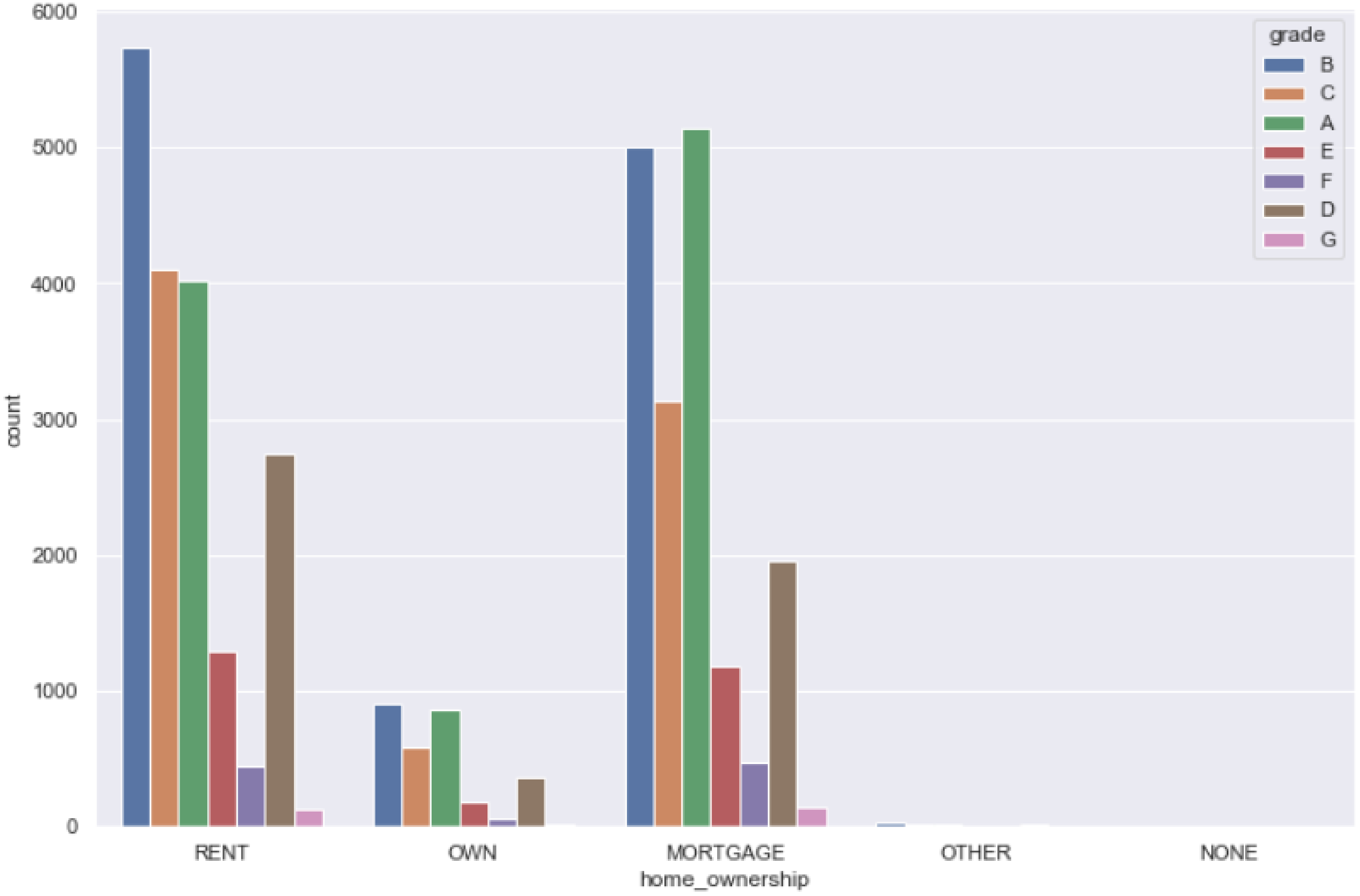
Observation *Yes, our assumption was correct the lender might be doing risk assessment and charging higher interest on the higher graded loans. This also explains the previous observation of (higher graded loans have more chance of default in %age terms)*

Univariate Analysis



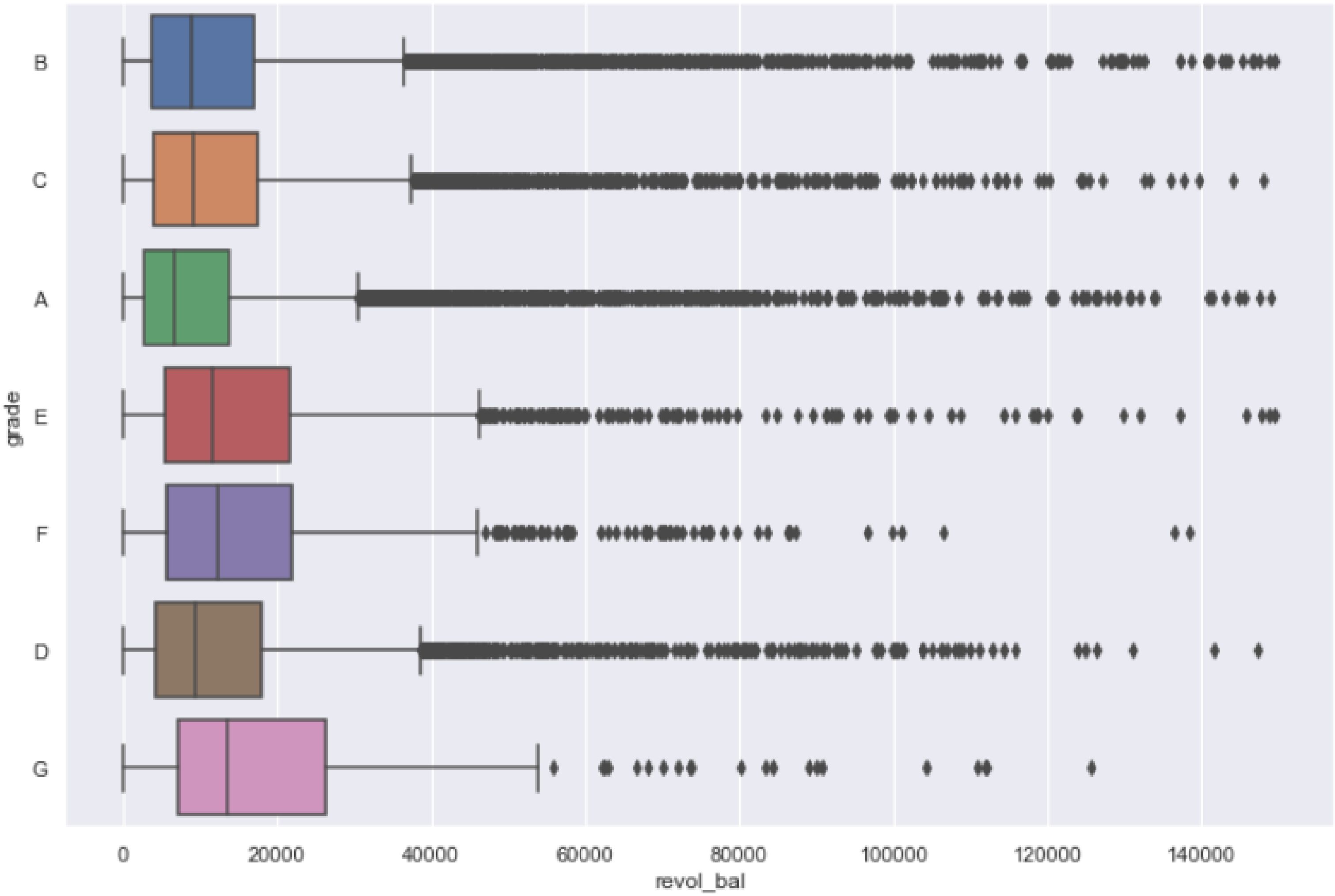
Observation *The quantiles and median seems to be very similarly distributed in each of the grades with longer tail of outliers in A, B, C grades, meaning it plays good role to be getting assigned lower grade but might not be that significant to cause huge shift in quantiles or median towards right.*

Univariate Analysis



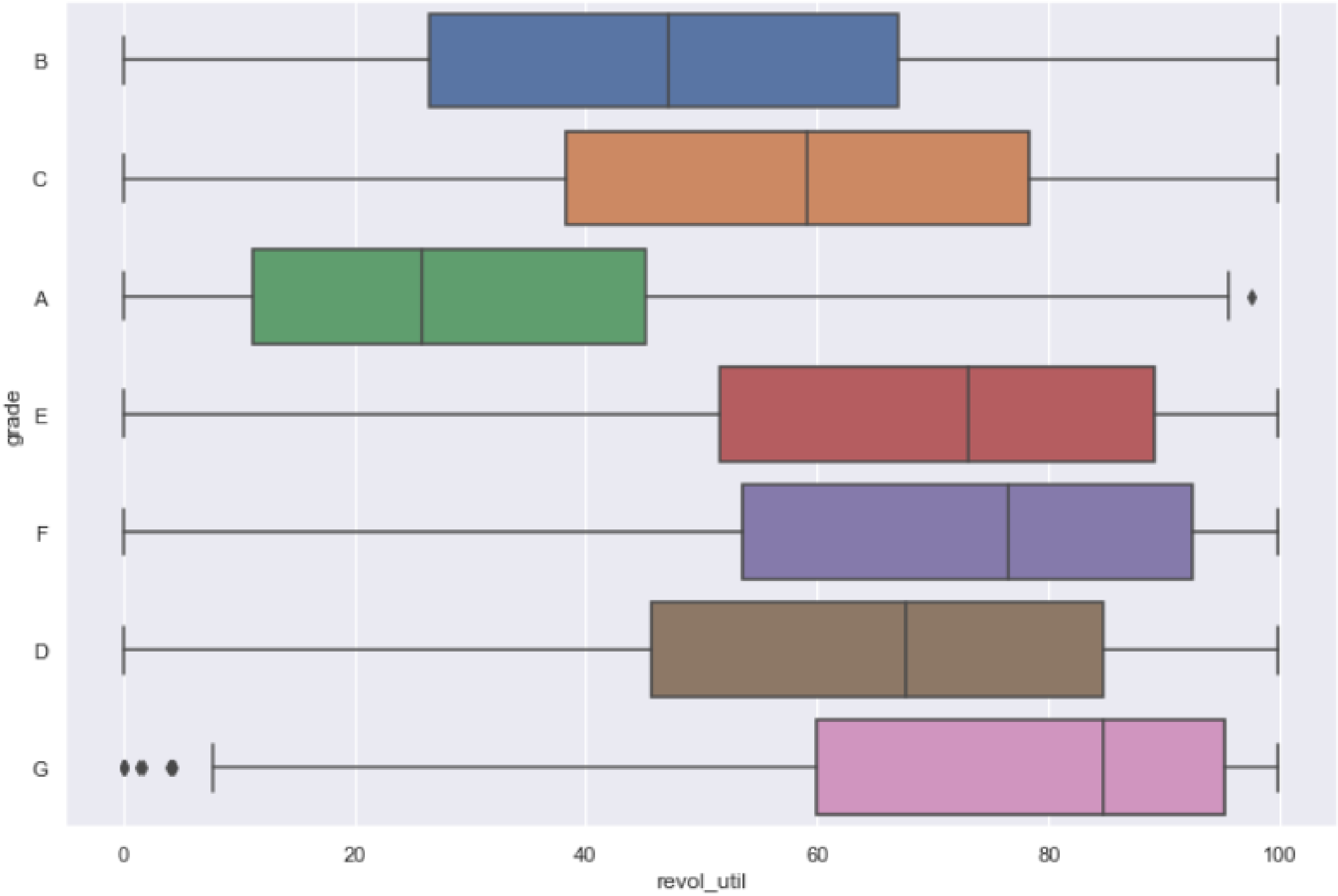
Observation *We can draw two observations here, one is the lender has granted lesser loans to homeowners when compared to other categories. Also, we can clearly see that homeowners and mortgagers get significantly higher-grade A (less risky, low interest) loans. Infact, grade A loan numbers in both these two categories are similar to grade B loan numbers*

Univariate Analysis



Observation *Revolving balance is the average amount of credit taken in absolute terms. We can clearly see that the people with lower grades utilize less amount of credit on average. It is a major factor in calculating the credit score, hence the people with better money management habits get assigned lower graded loans that charge lower interest. But there are many outliers in all the grades.*

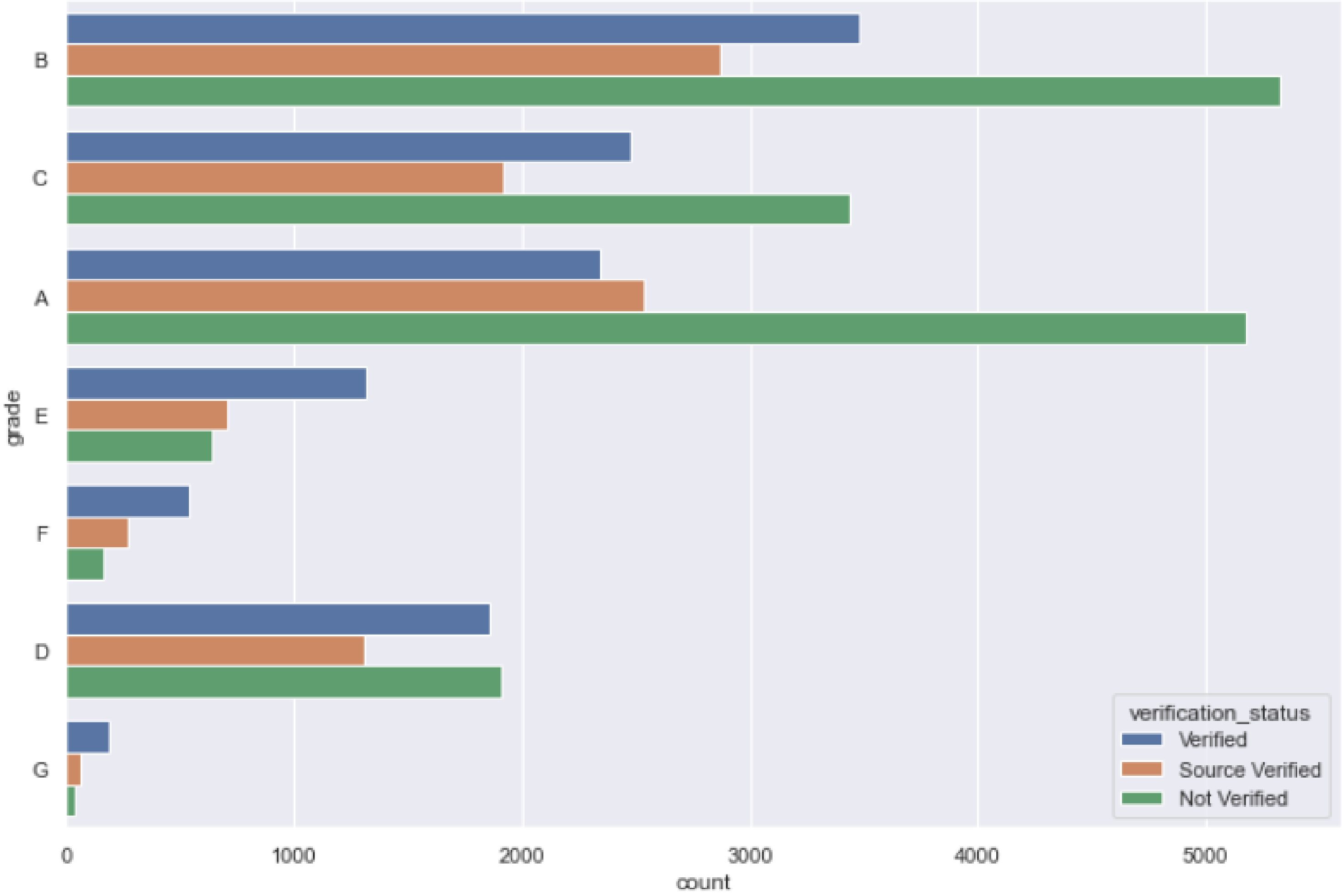
Univariate Analysis



Observation *Revolving Utilization is the percentage of assigned credit a person is utilizing. Here in percentage terms, it becomes very obvious how people with lower grades utilize their credit properly and might be having better credit scores to get lower graded loans.*

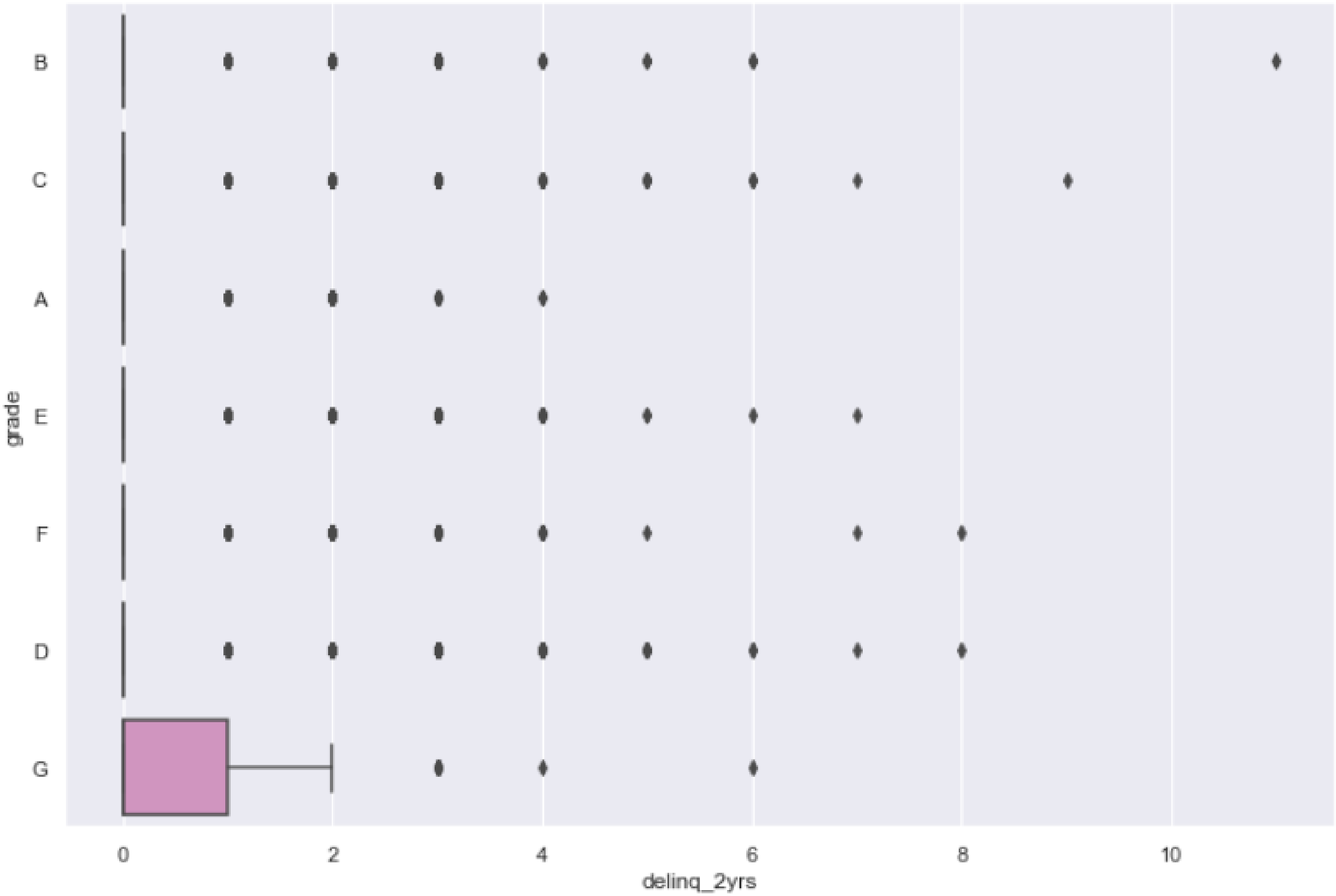
In [77]:

Univariate Analysis



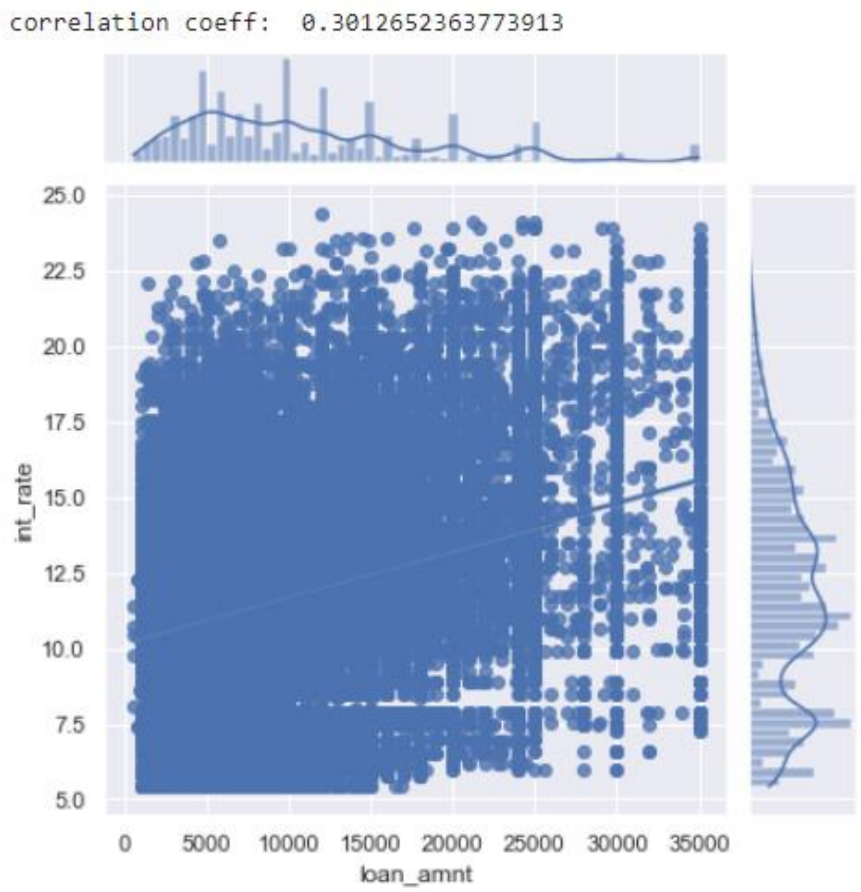
Observation *Not very obvious at every grade but bank does more verification (verified or source verified) in cases of higher graded loans. Maybe as the people might have worse credit scores, making banks do more verification.*

Univariate Analysis

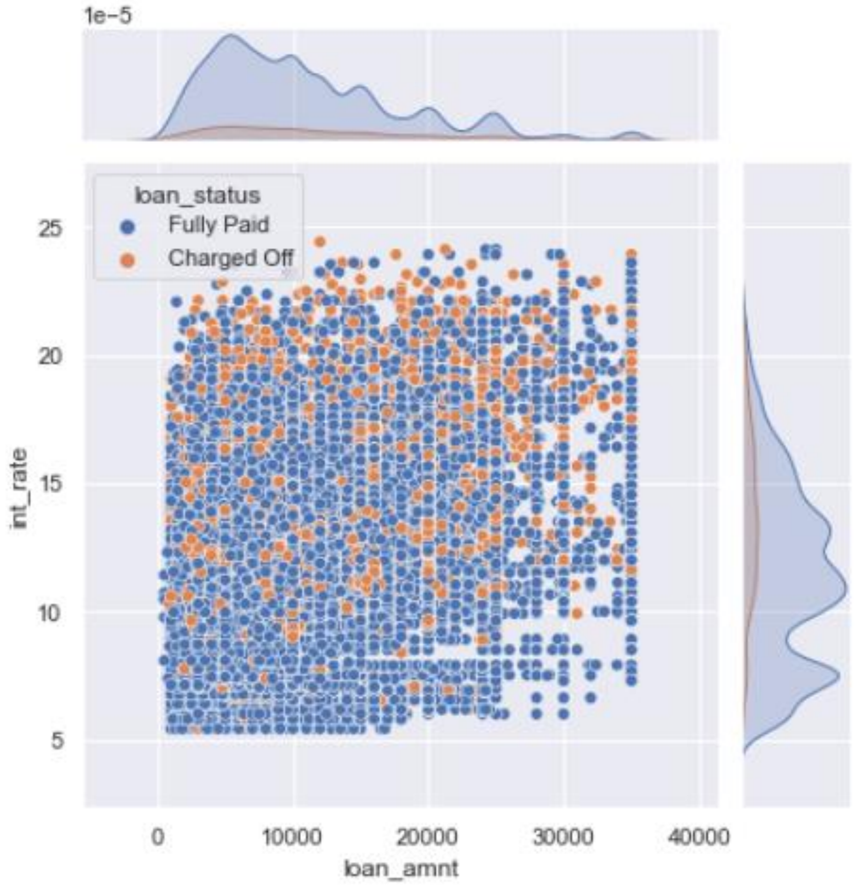


Observation *Majority of people have not missed payments by more than 30 days in last 2 years.(Many outliers also present in each category) Only in G grade loans at least 25% of people have missed at least 1 payment.

Bivariate Analysis



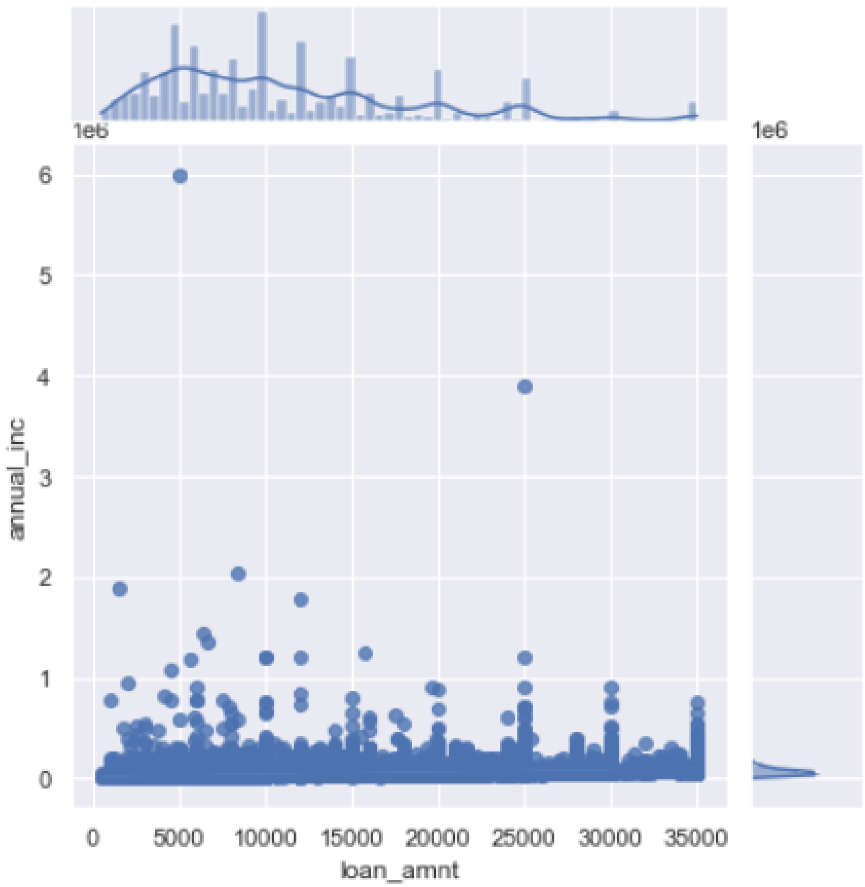
```
correlation coeff:  0.3012652363773913
```

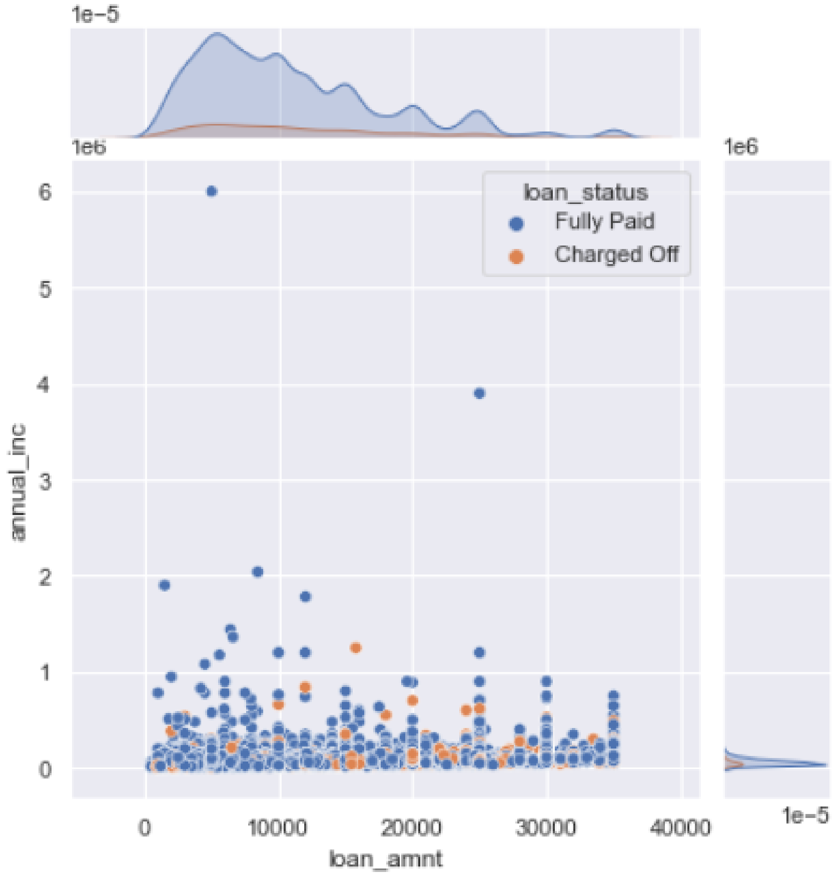


Observation *There is a minor correlation of 30%.*

Bivariate Analysis

```
correlation coeff:  0.2689993990756643
```

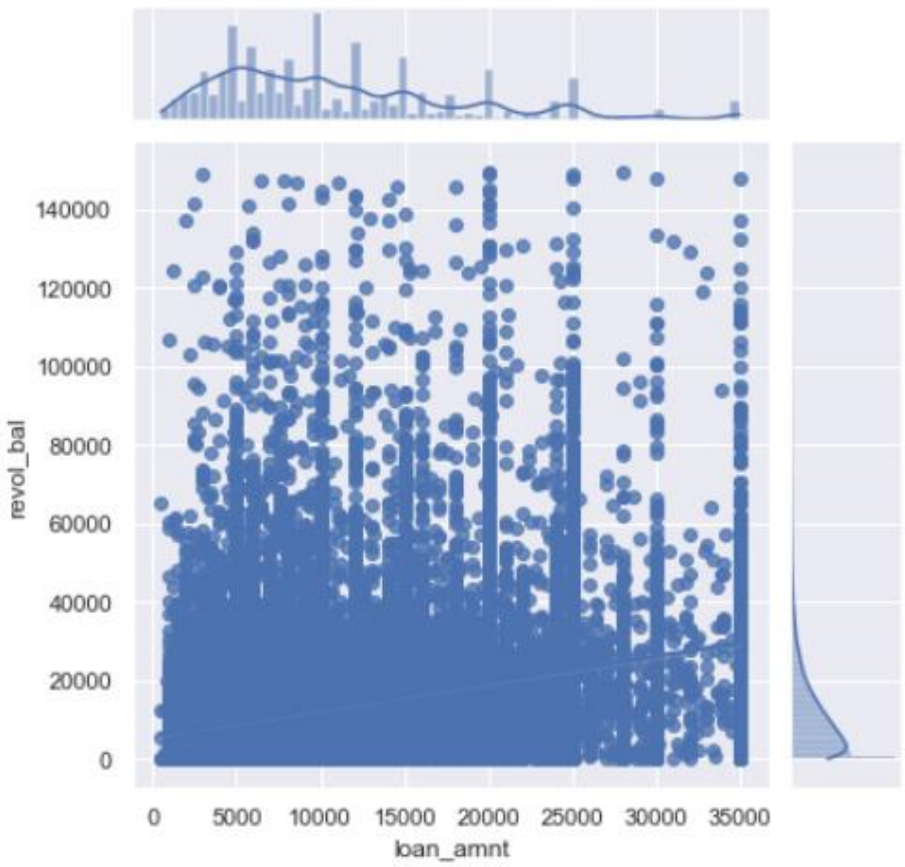


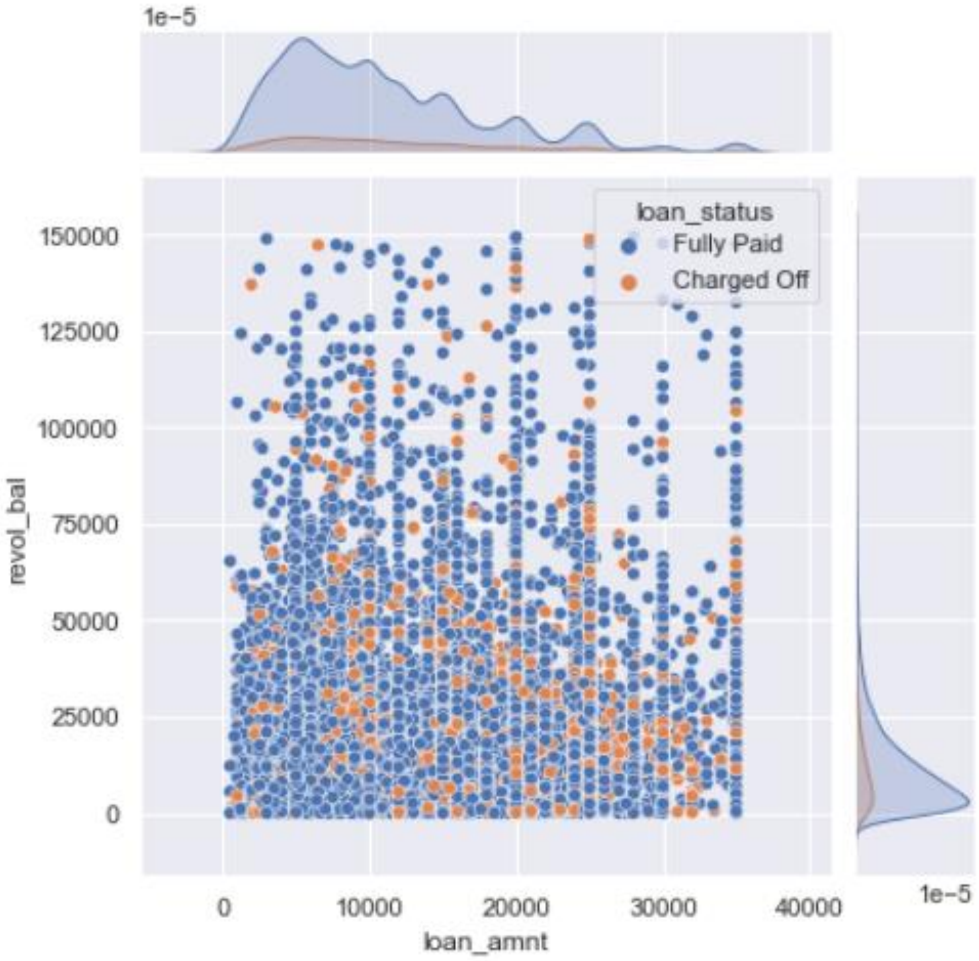


Observation *We can see that there is minor correlation.*

Bivariate Analysis

```
correlation coeff:  0.3140222066227738
```

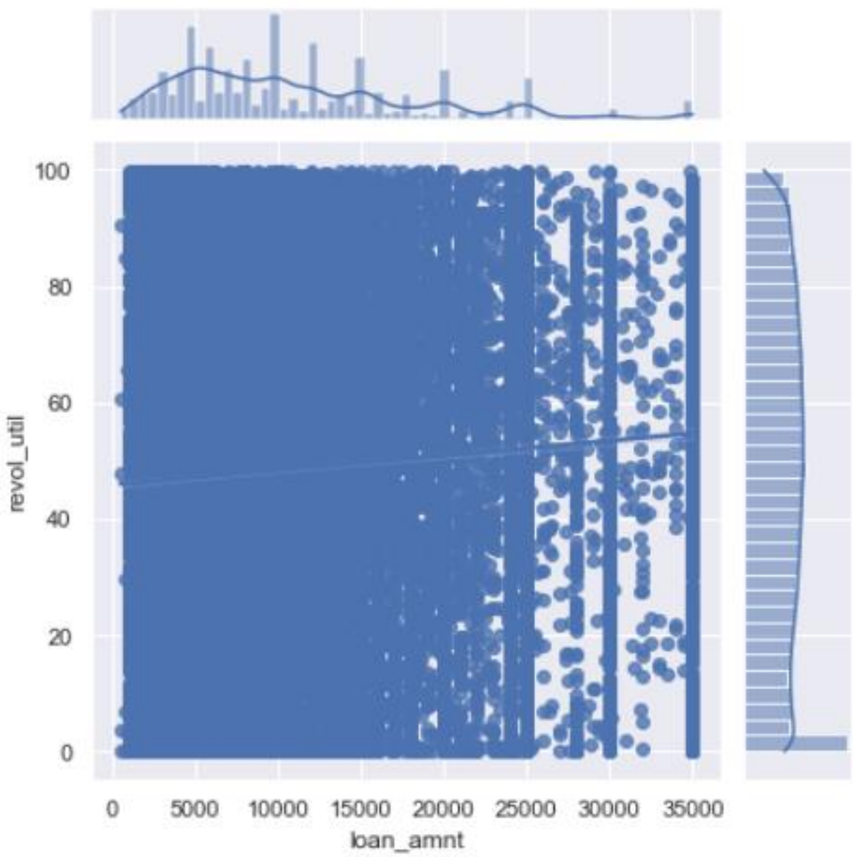


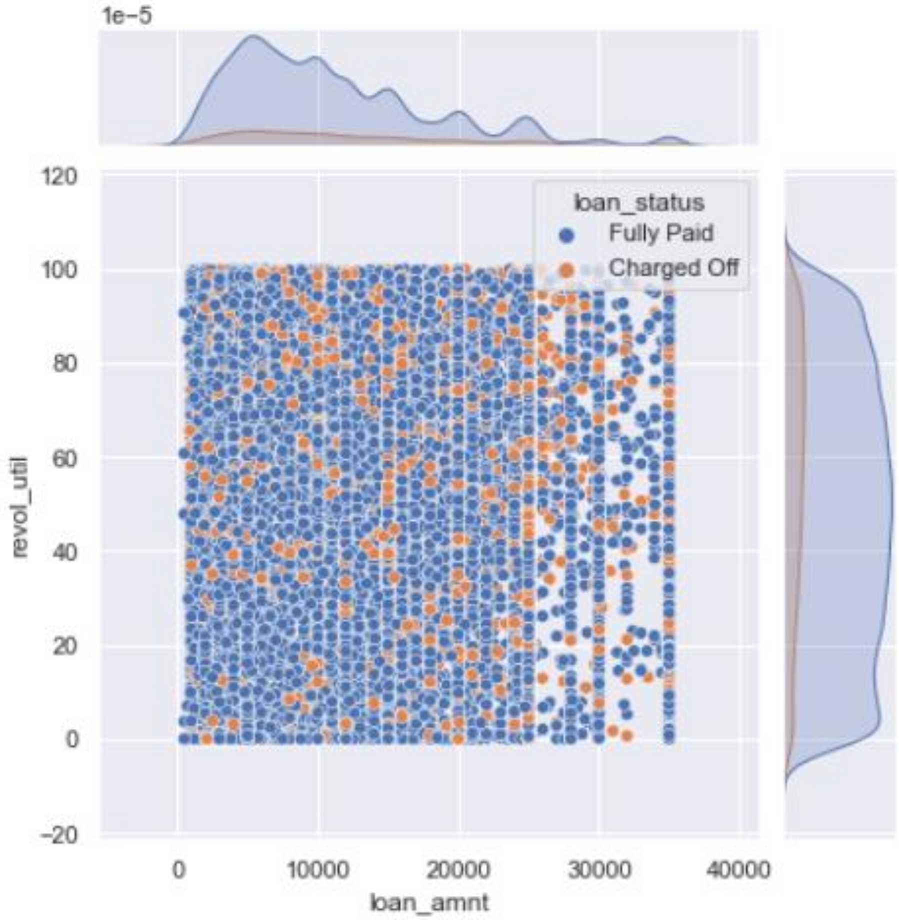


Observation *We can see that there is a good 30+% correlation. Concluding that people who use more absolute credit are most likely to have higher loan sanctioned by the lender.*

Bivariate Analysis

```
correlation coeff:  0.0652470091155389
```

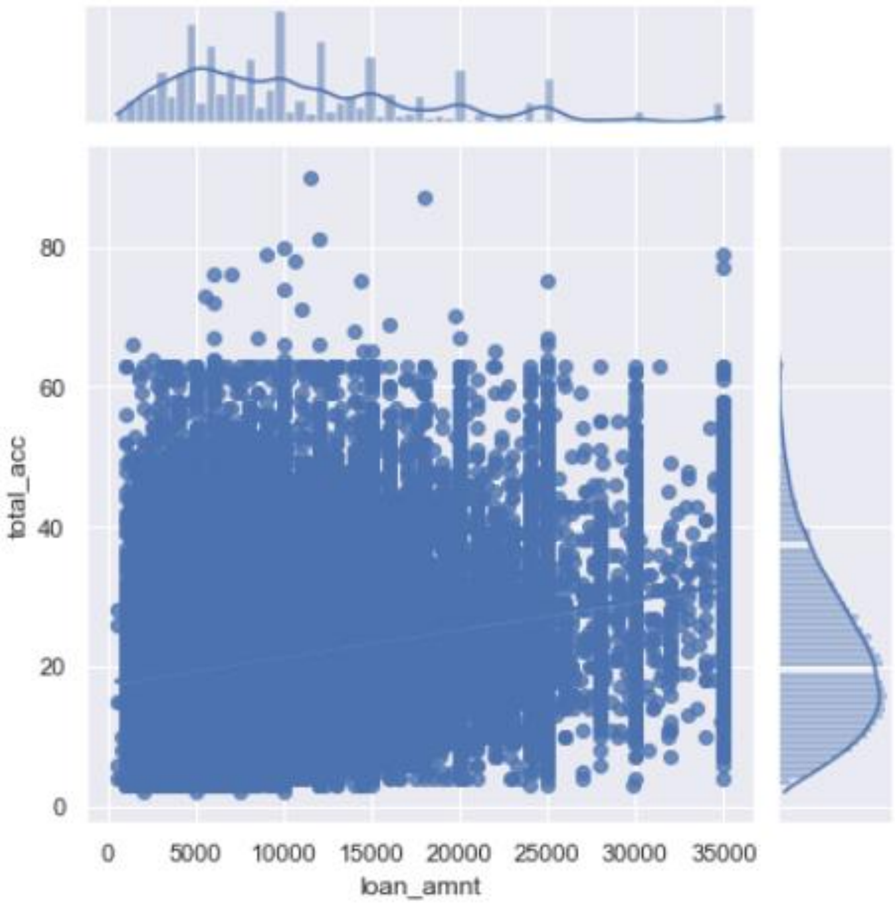


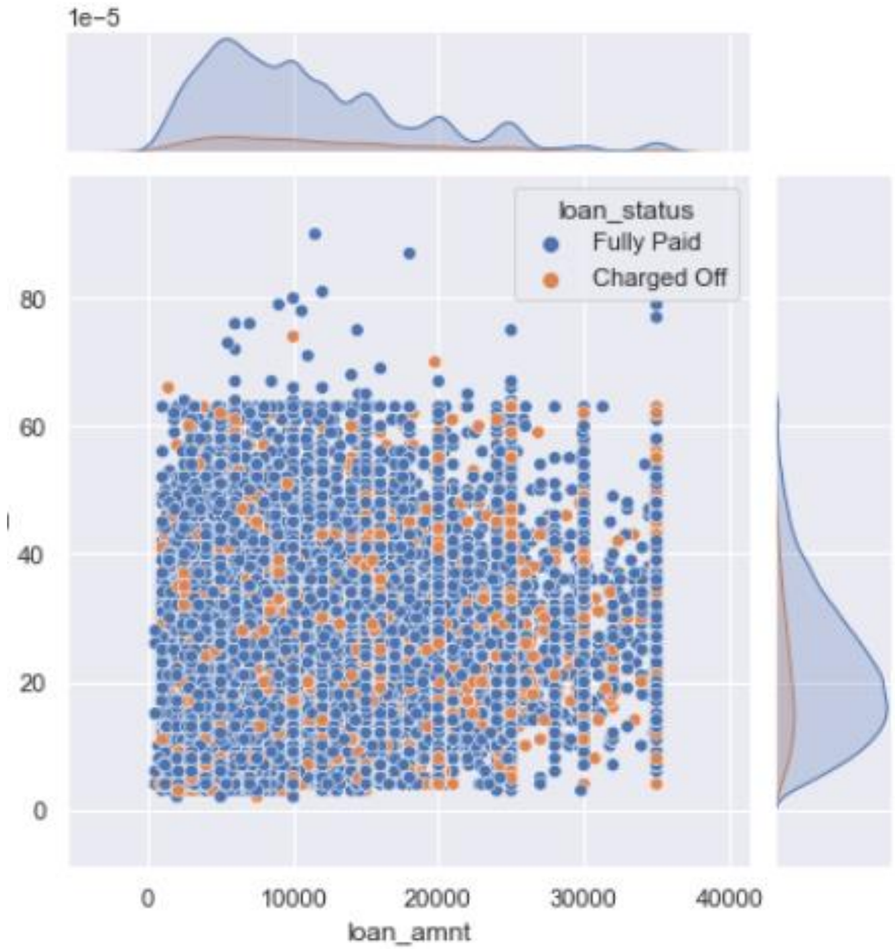


Observation *There barely any correlation between the two variables*

Bivariate Analysis

```
correlation coeff:  0.25617929816607554
```

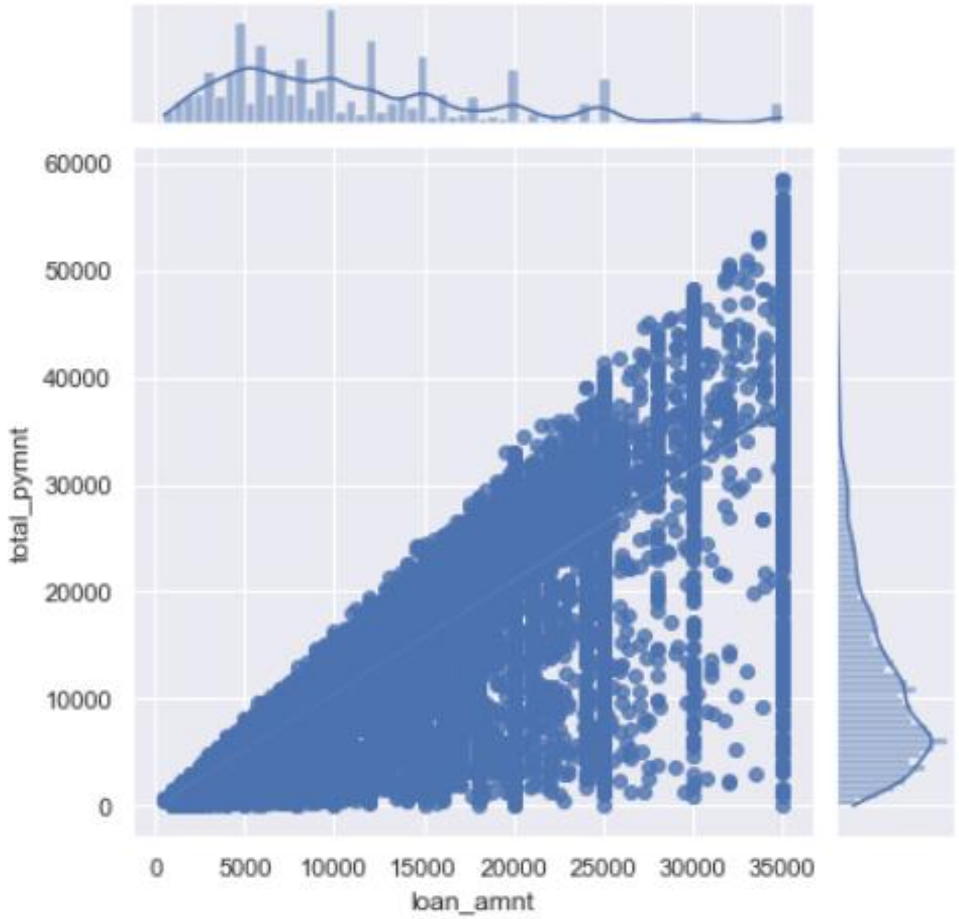


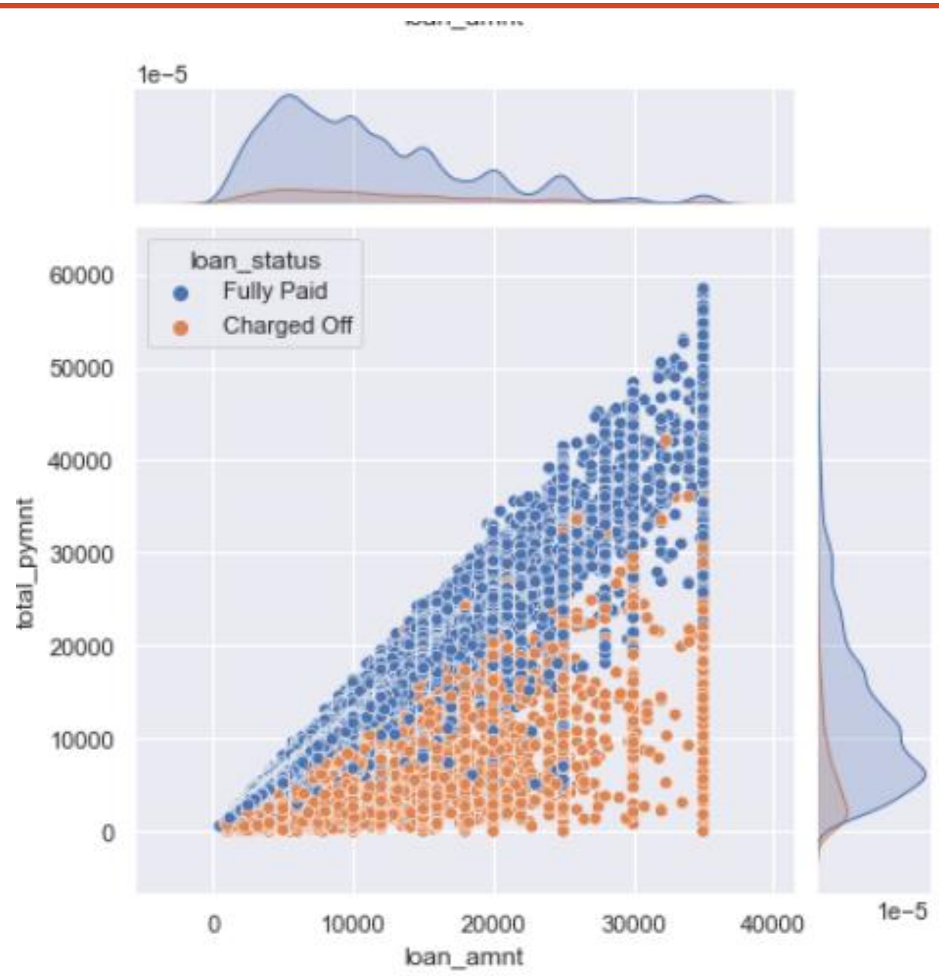


Observation *There is about minor 25% correlation between the two variables.*

Bivariate Analysis

```
correlation coeff:  0.8819103258047376
```

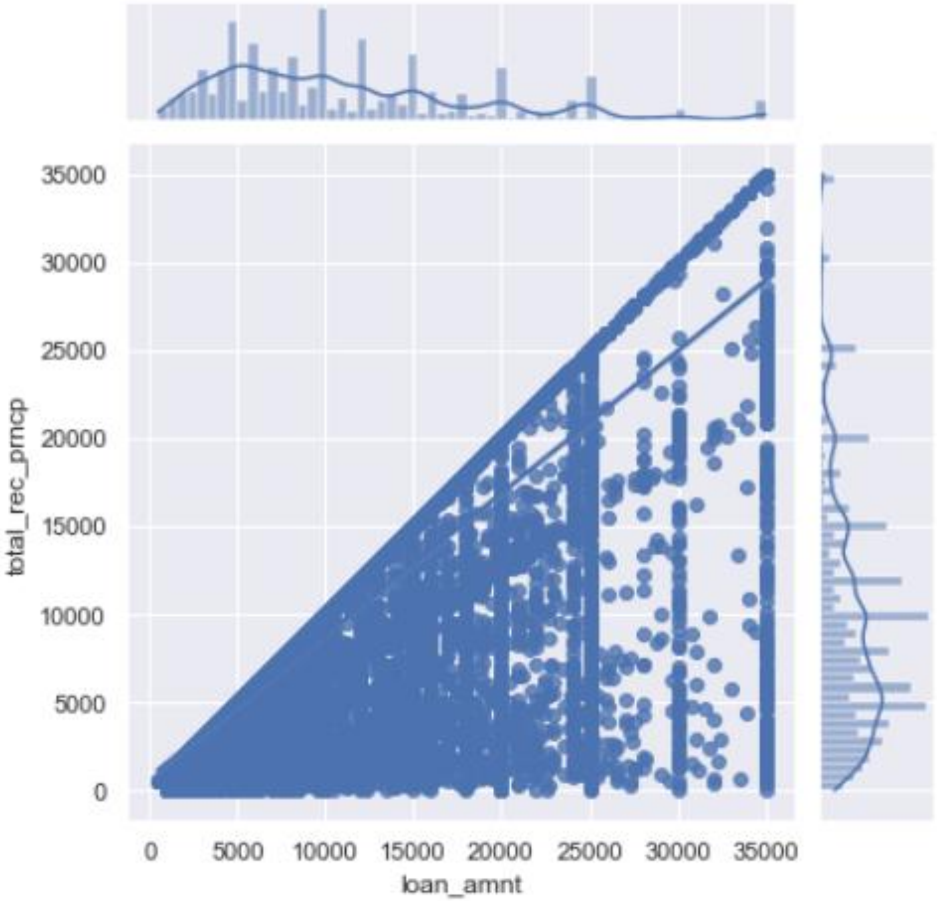


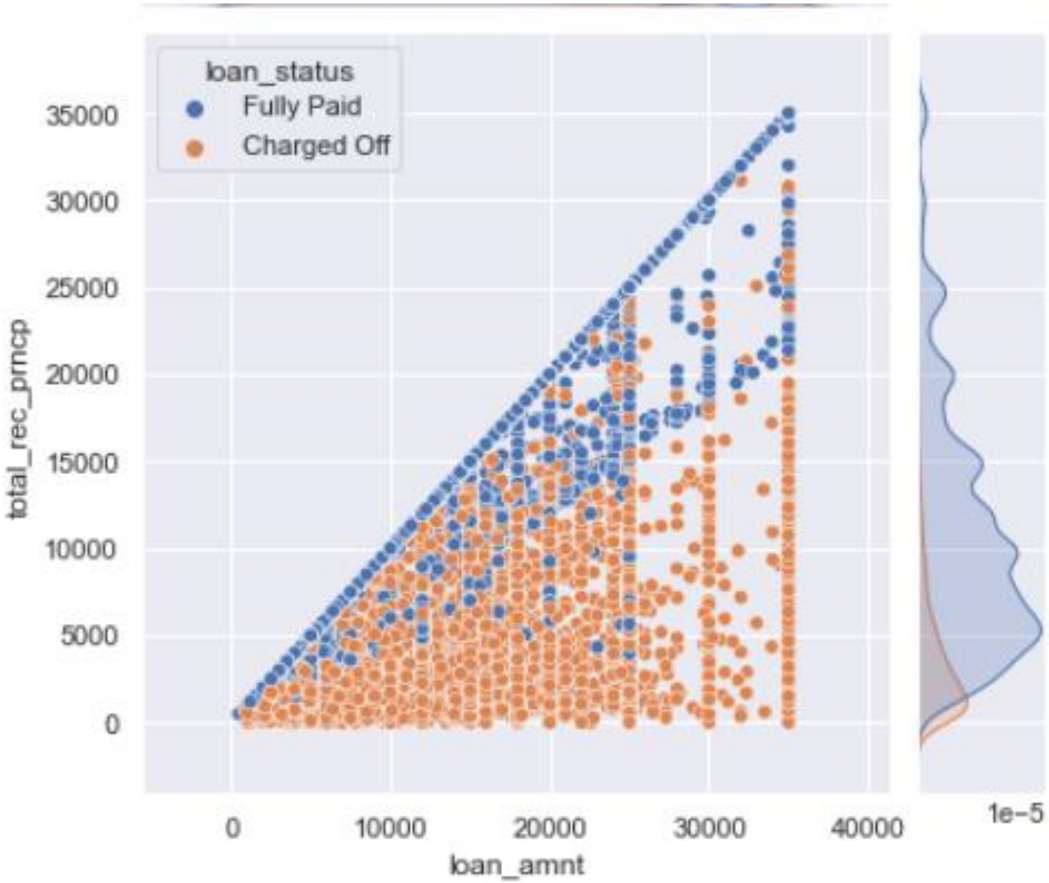


Observation *There is very high 88% correlation between the two variables. This is mainly due to the fact that most of the would have paid the loan amount and the interest amount and hence are very correlated. Also, we can clearly see that charged off (defaulters) and fully paid users are segregated, which is understandable. Also, we can see few fully paid points in between defaulters, we had identified this before as well. There could be some errors in data set or there could be some other factor into play that is not getting captured in the dataset, something like government subsidy (the borrower would*

Bivariate Analysis

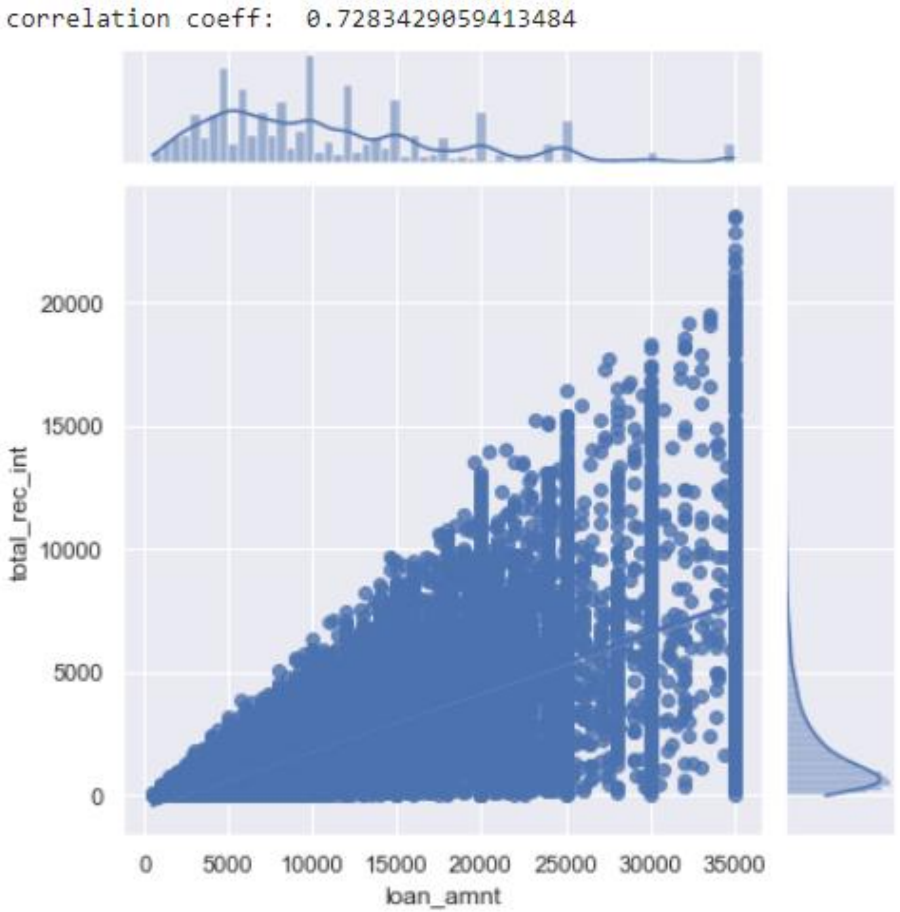
```
correlation coeff:  0.84586979587196
```



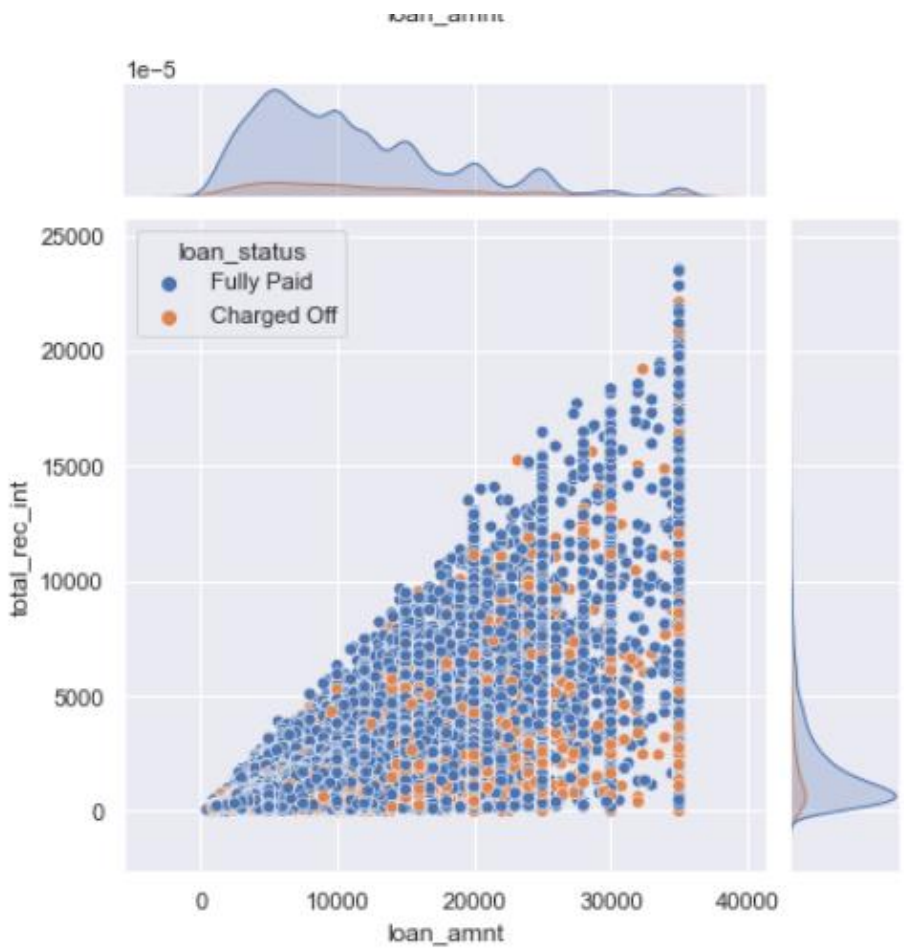


Observation *Again there is a very high 84% correlation for the same reasons as previous two variables. Same issues also getting highlighted as well.*

Bivariate Analysis



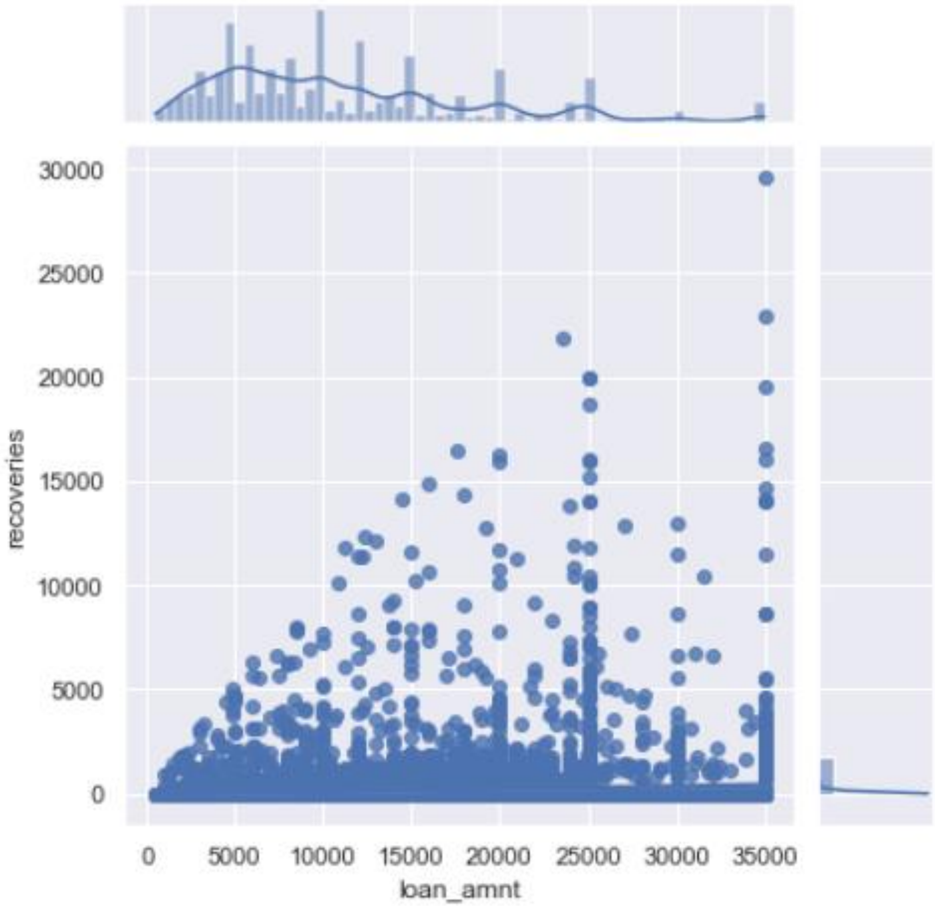
```
correlation coeff:  0.7283429059413484
```

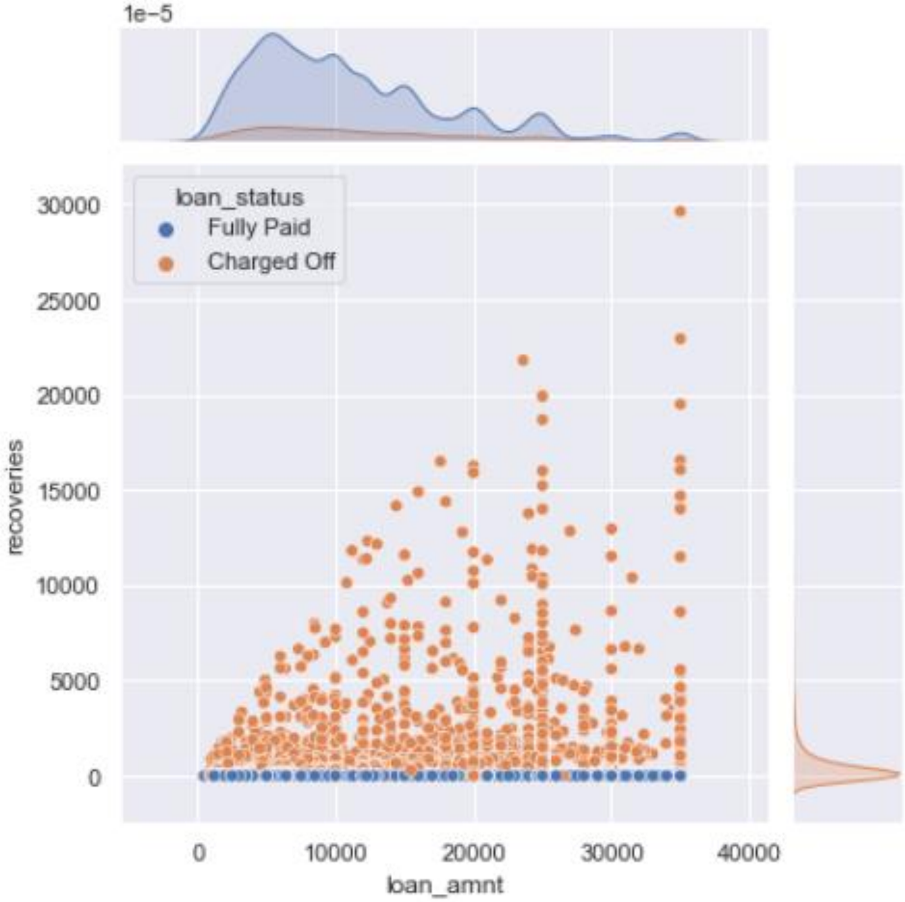


Observation *Again there is a good 73% correlation for the same reasons as previous variables.*

Bivariate Analysis

```
correlation coeff:  0.1427892610412201
```

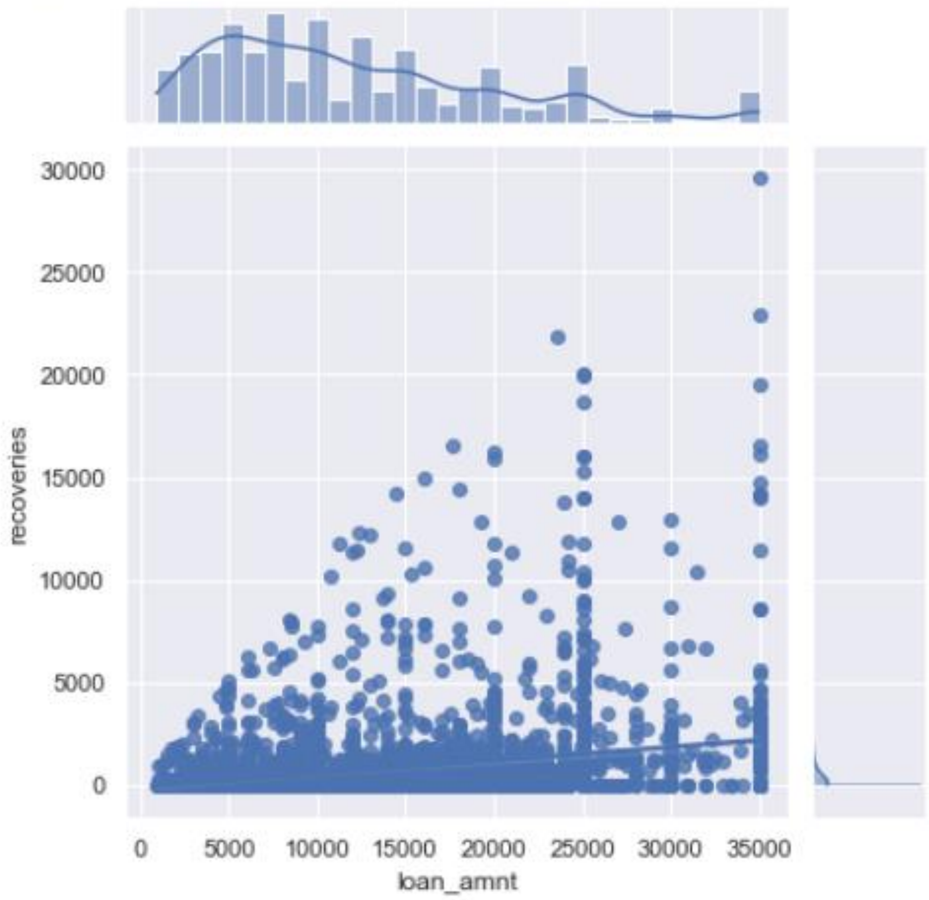


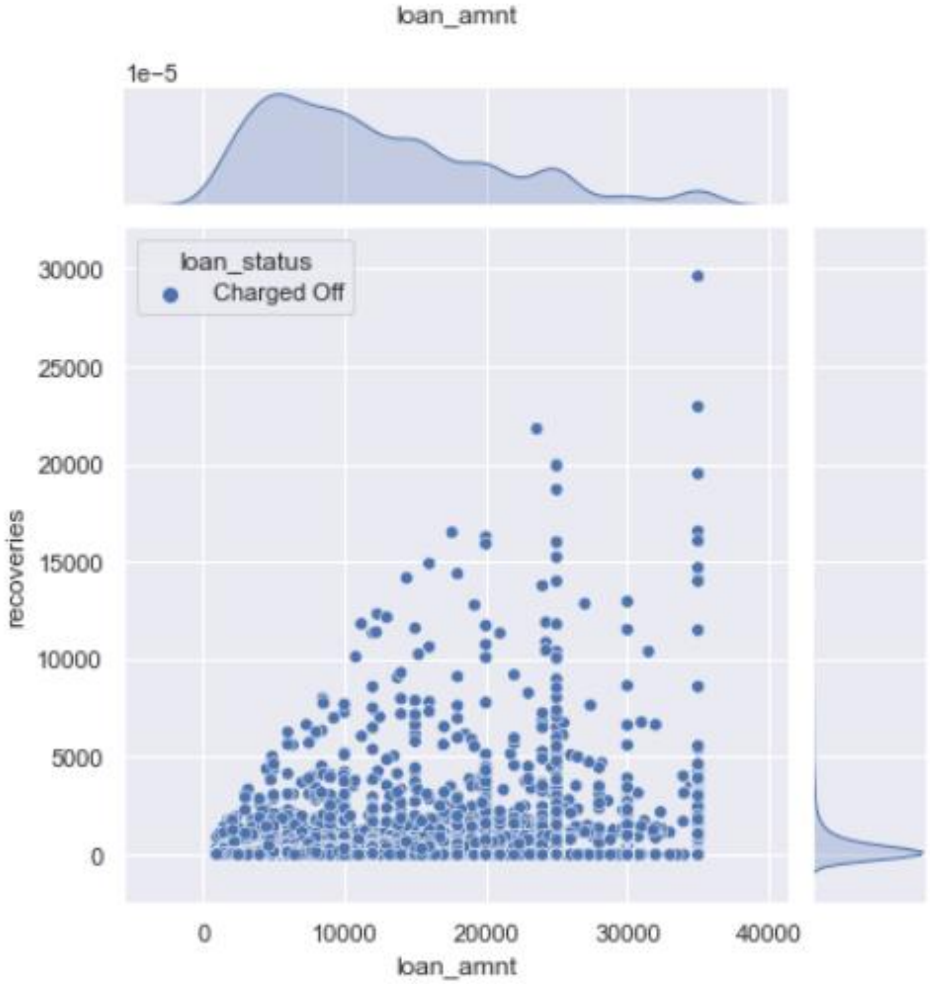


Observation *There is a very low correlation here. But we can clearly see that this could be due to majority of the people how have paid the whole loan and didn't need to pay via recovery. We can try to check this for the defaulters only to see the correlation for better picture*

Bivariate Analysis

```
correlation coeff:  0.31018492071809645
```





Observation *We can clearly see that there is a good 30+% correlation between the two variables when we use the sebset of people who have defaulted.*

Key Findings & Outcomes:

1. Loan amount doesn’t seem to play any significant role in people defaulting.

2. The lender has graded the loan in the cat of A-G.

3. Higher graded loans are having the high probability of being defaulted.

4. Hence, the lender is charging high interest rate on higher graded loans.

5. Following factors seem to play significant role on grading loans:
    1. Home ownership
    2. Revol_balance
    3. Revol_util
    4. Delinquite_2yrs

6. Also, annual_inc, purpose are not the driving factors for loan grading.

7. There are entries where loan principal amount had not been fully paid but those entries were marked as ‘Fully paid’, this leads to huge data discrepancy.

Learnings:

1. We learnt how to use KNN imputer and during null values imputation found out that KNN imputers only worked for numeric values.

2. Before during any analysis, we should check for highly correlated columns and drop them to avoid duplicity.

3. Segment Univariate analysis helps us to identify trends using different quartiles, which are very helpful for decision making.

4. Learnt how to see the dependency of one variable on another using correlation coefficient.

ThankYou