Lending case study

Kiran sunkari
Kunal choudhary

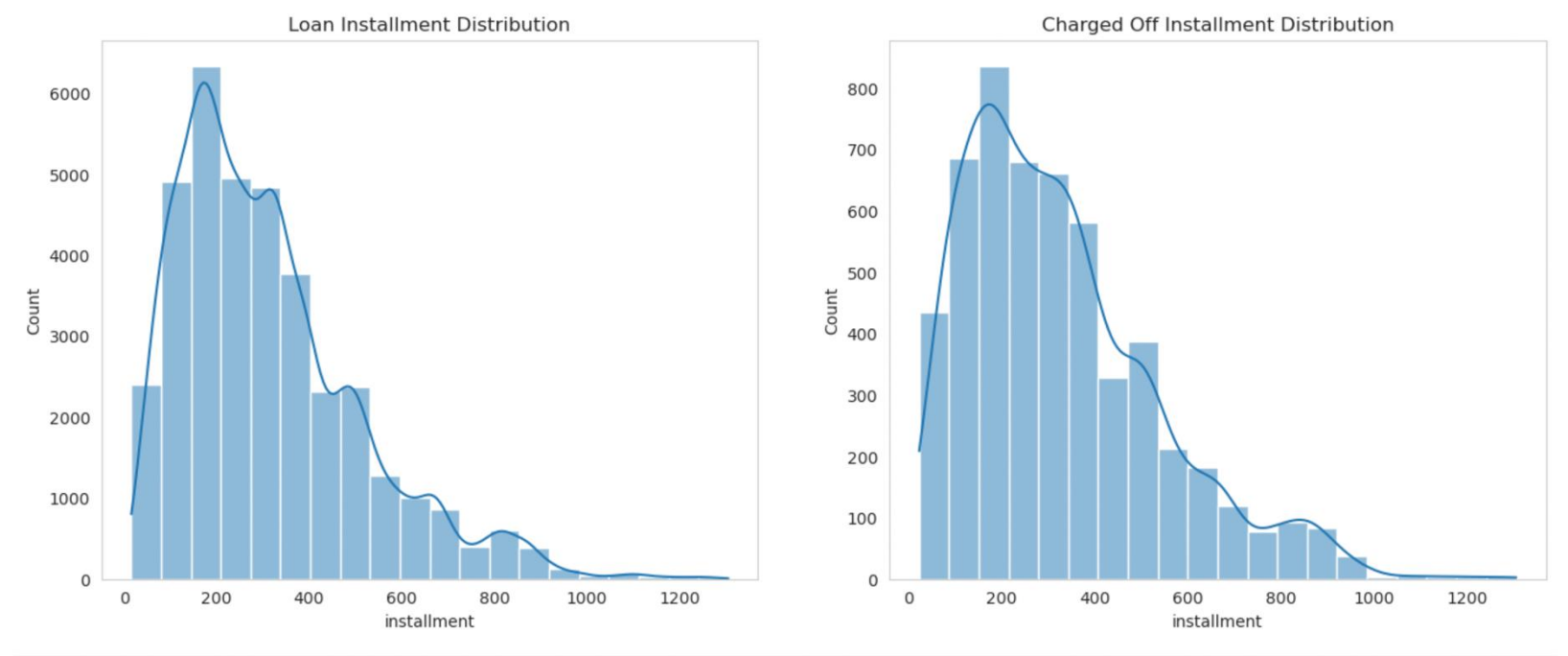
Objective

- The objective of this case study is to explore various methods for analyzing data, apply Exploratory Data Analysis (EDA) techniques to a real-world problem, and derive meaningful insights
- For any ML/AI solution, understanding the business problem is the most critical step. In this context, EDA serves as the first step to identify the key features for our ML or AI solutions.

Dataset understanding &cleaning

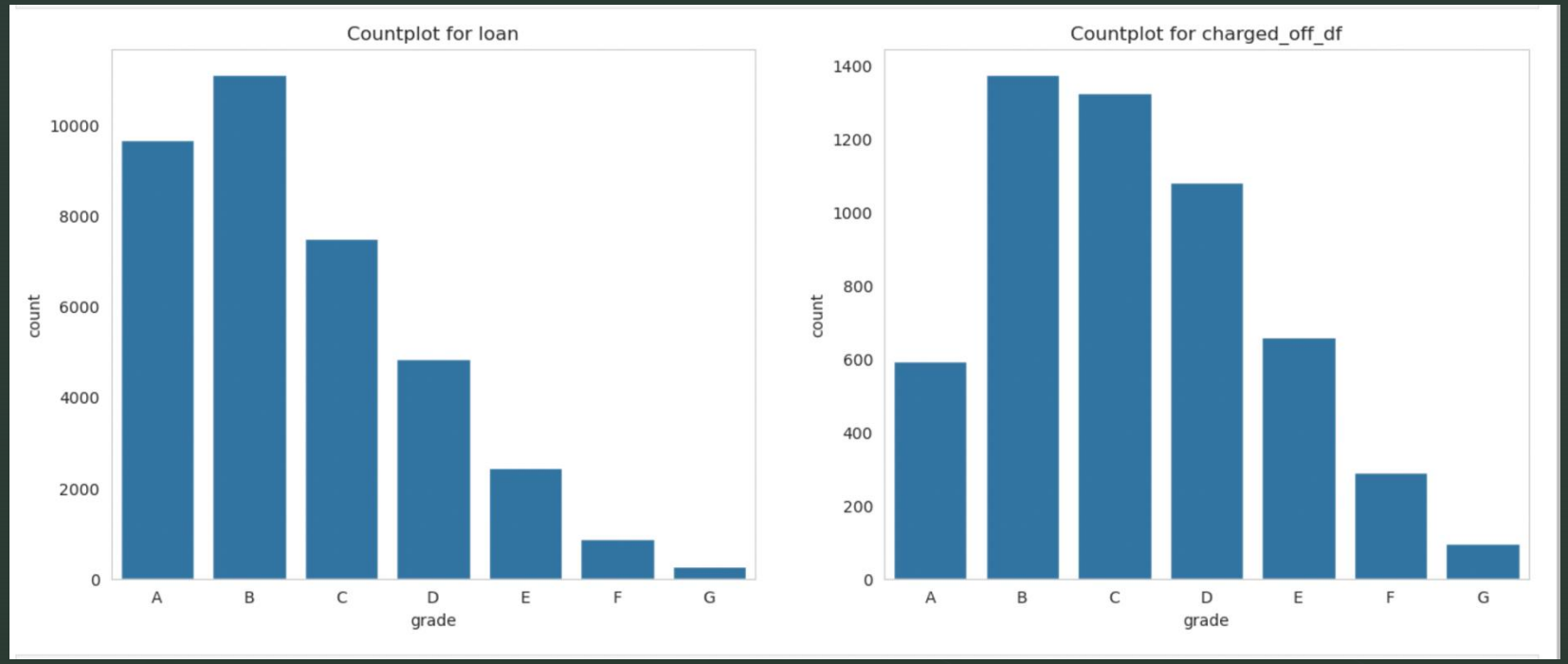
- Removed all columns and rows that contain only null values.
- Deleted the duplicate rows and columns
- Grouped the quantitative variables, such as annual_inc, int_rate, and dti
- Analyzed individual columns and removed outliers. For example, in the case of annual_inc, all records beyond the 95th percentile were excluded, retaining only the records up to the 95th percentile.

Loan Installment Distribution



Performed univariate analysis by plotting the distribution graph for loan installment property, comparing loan-defaulters with the entire dataset. No signals were observed in the loan installment data.

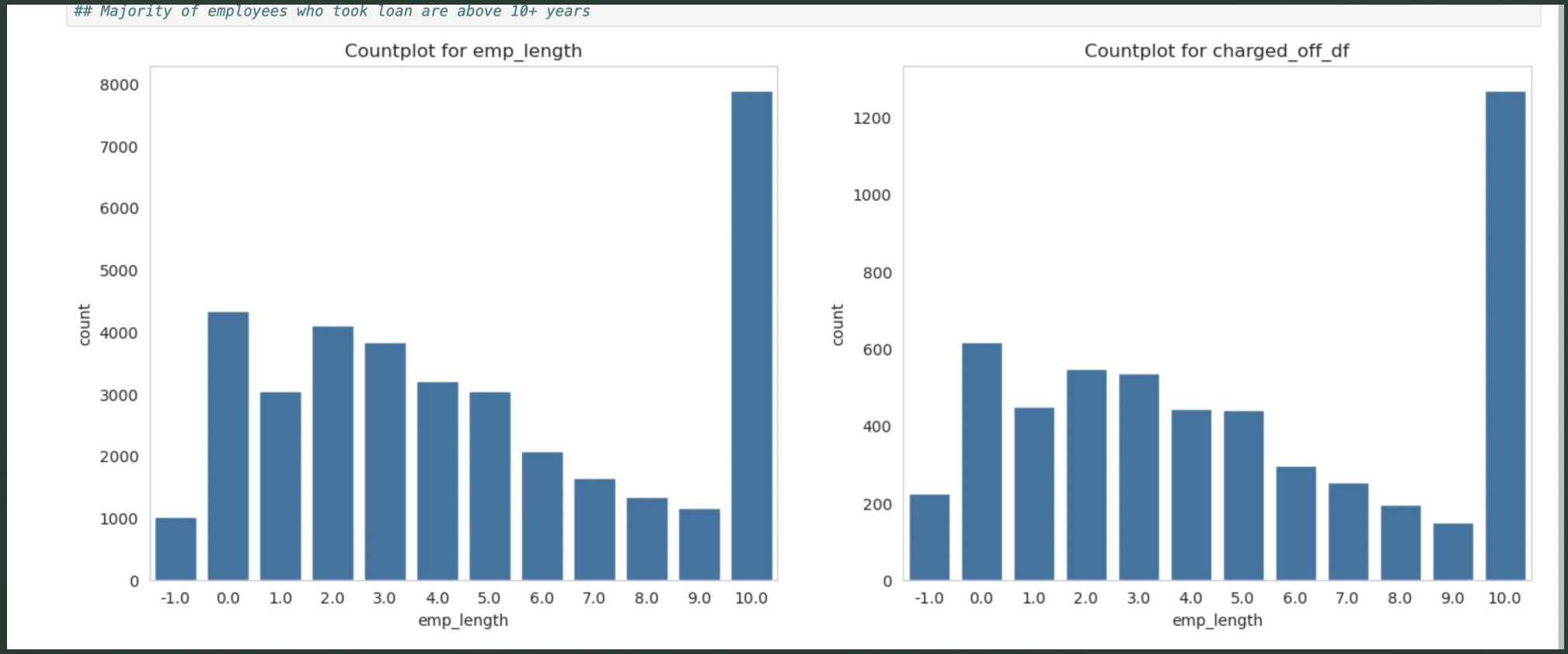
Grade property analysis



When analyzing the **Grade** property, we observe that the **Grade 'C'** category has a higher proportion of defaulters compared to other grades.

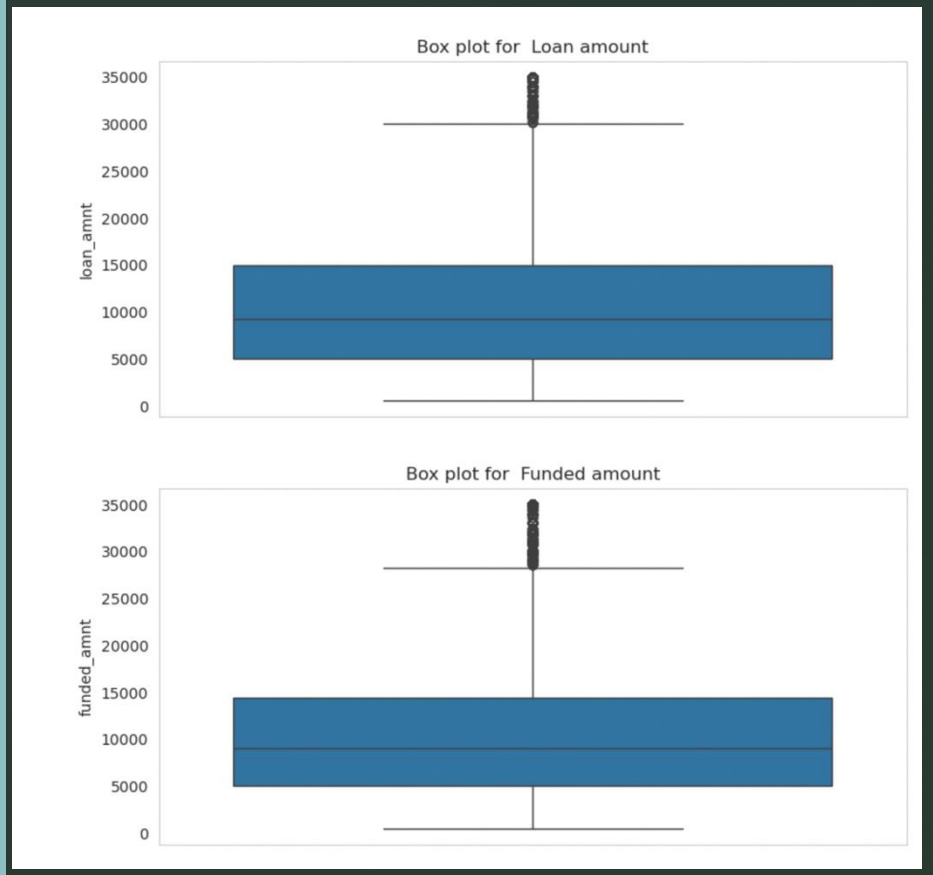
Employee length

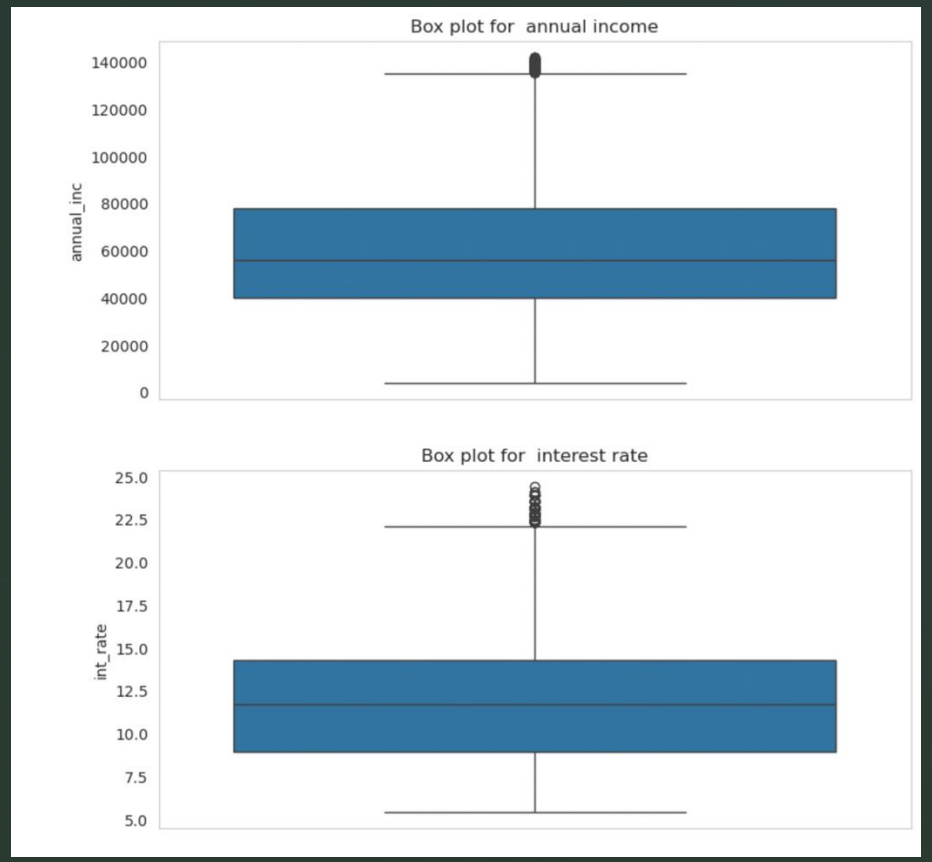
```
## Majority of employees who took loan are above 10+ years
```



By examining the distributions, the **employee length** does not provide any meaningful signal. We need to add another property to validate this.

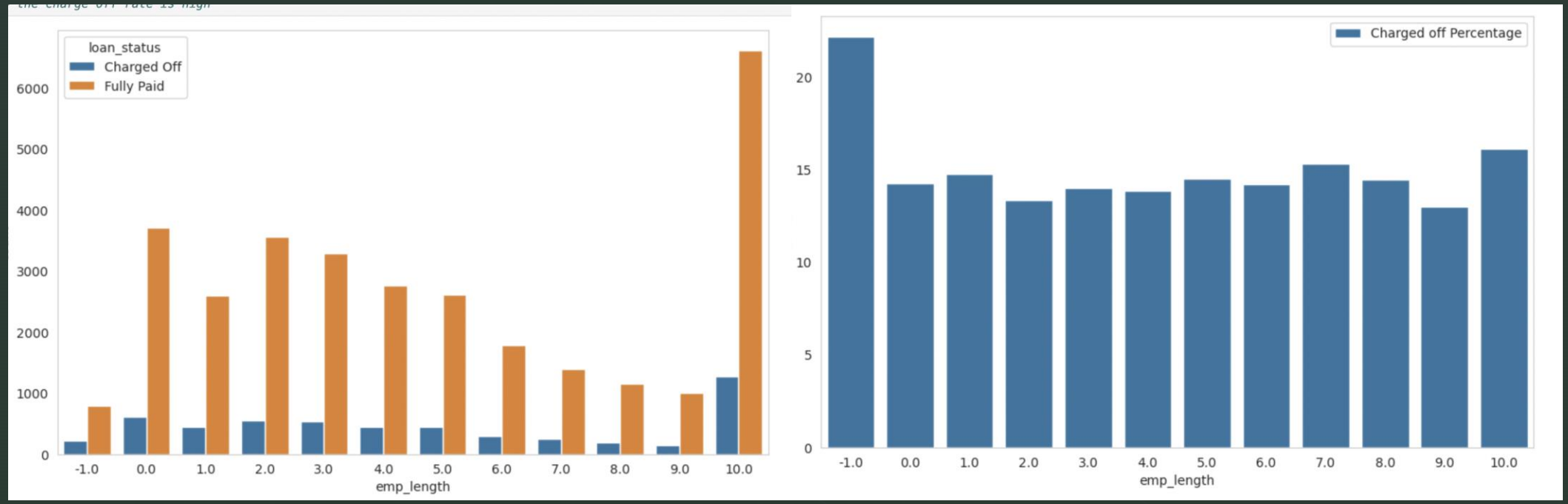
Analyzing quantitative variables





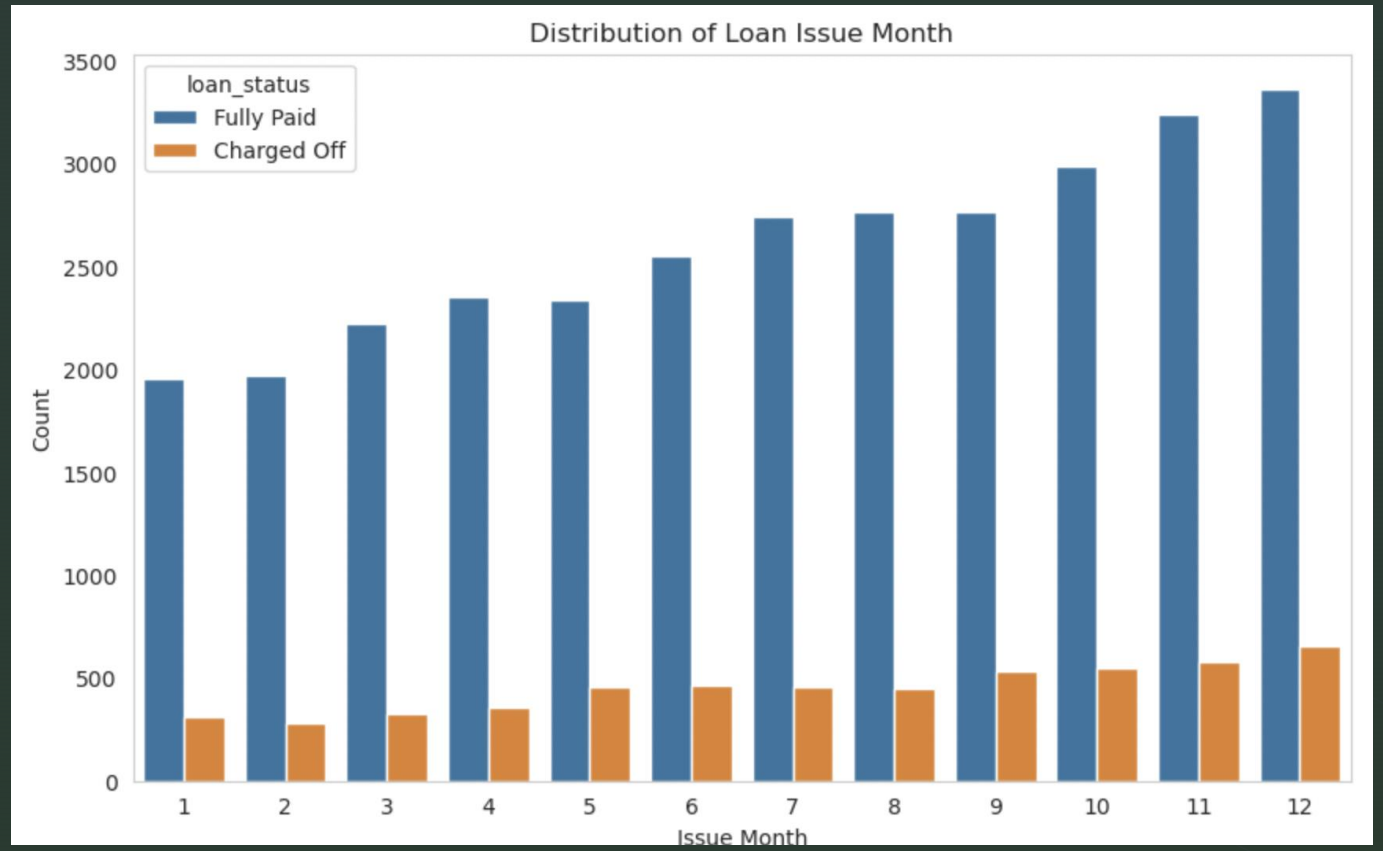
Plotted the box plots to understand the distribution of quantitative variables, remove the outliers based on the IQR, and perform data bucketing based on the IQR distribution

Employee length analysis



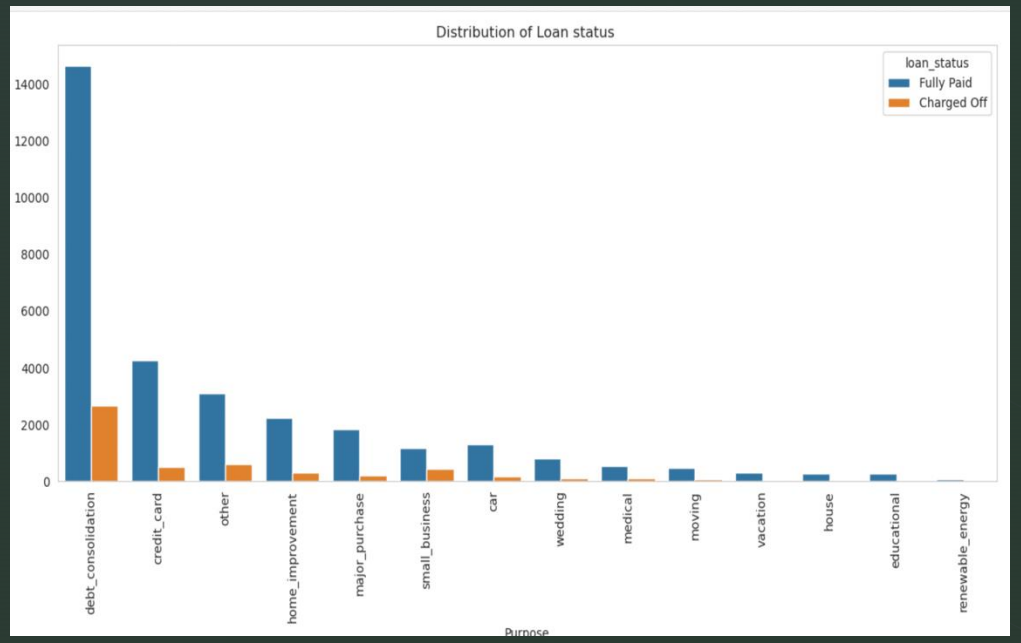
multivariate analysis: when plotting the **employee length** against **loan status**, no significant signal is observed. However, when plotting the **charged-off percentage**, it becomes apparent that records without an **employee length** (which may indicate that the applicant is not employed) have a higher default rate.

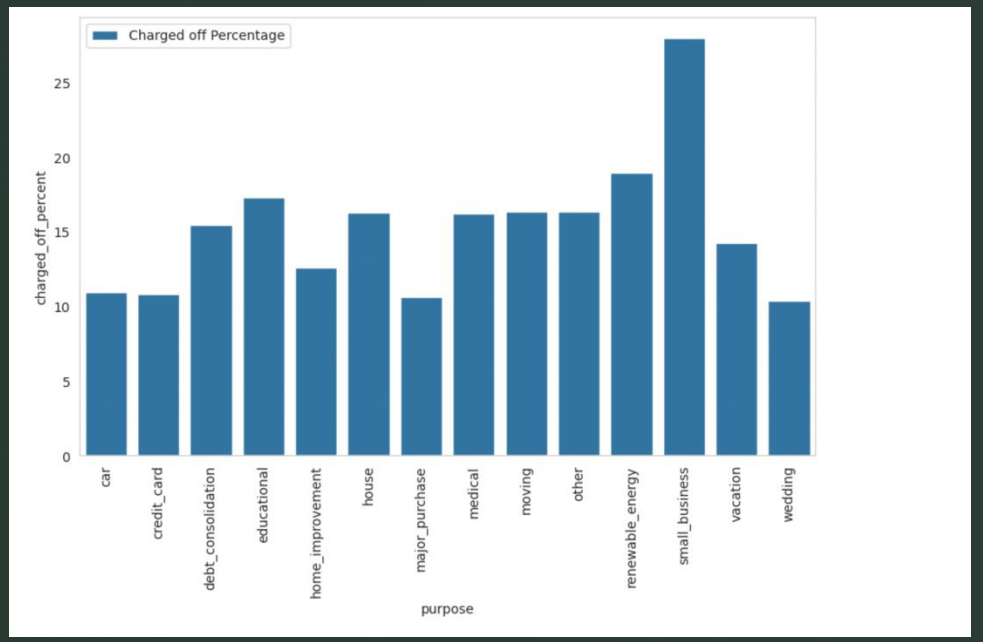
Month attribute analysis



Based on the distribution, we can observe that loans issued at the end of the year have a higher default rate

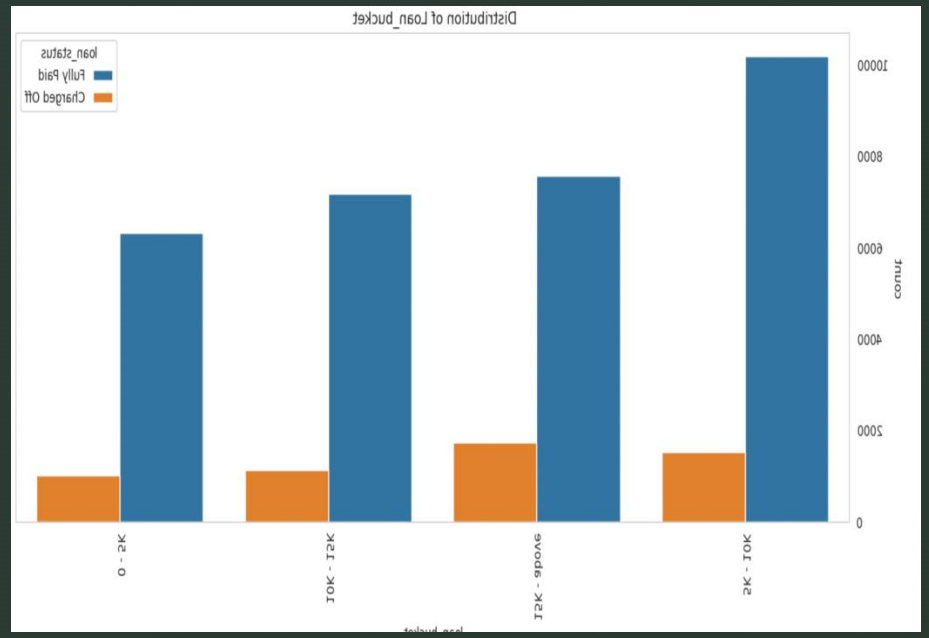
Purpose analysis

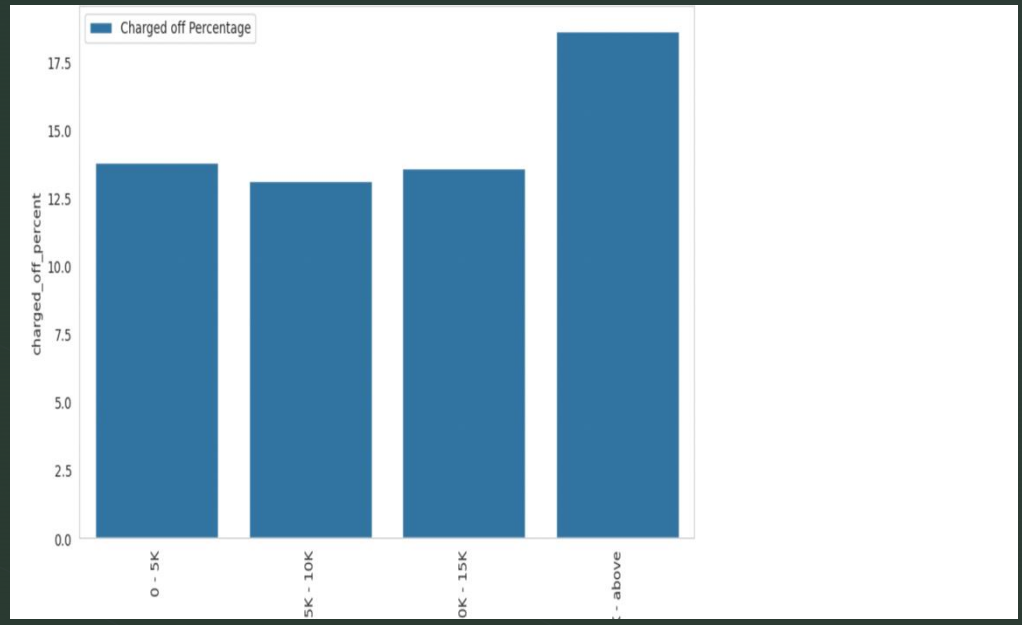




Upon analysis, we can observe that loans given for small businesses have a higher chance of default..

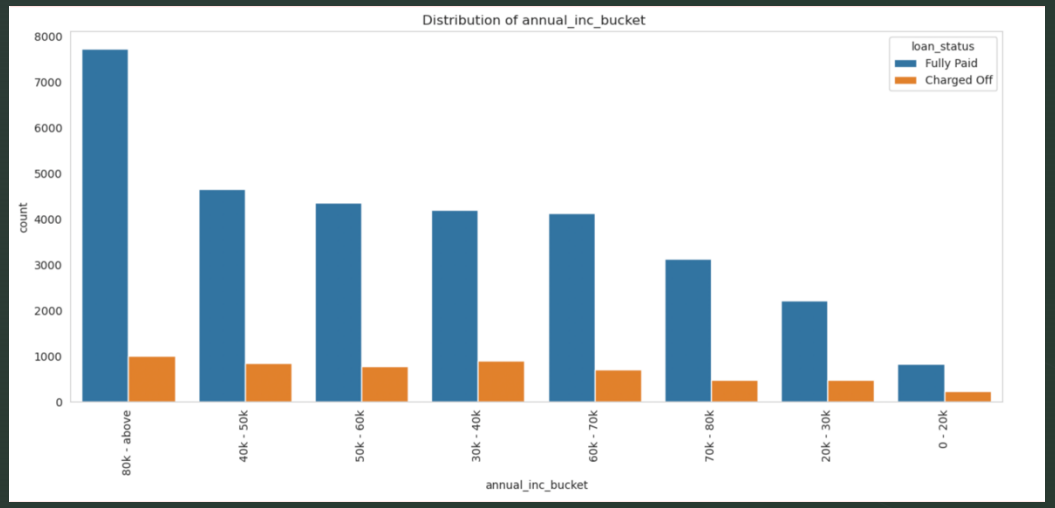
Loan amount bucket analysis

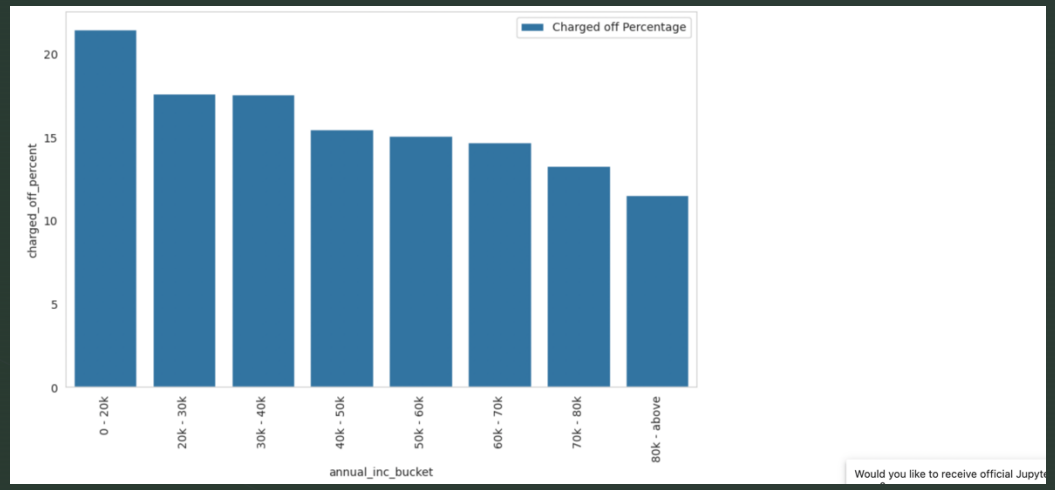




Upon analysis, we can observe that loan amount greater than 15k has a higher chance of default. But we need to check with other parameters

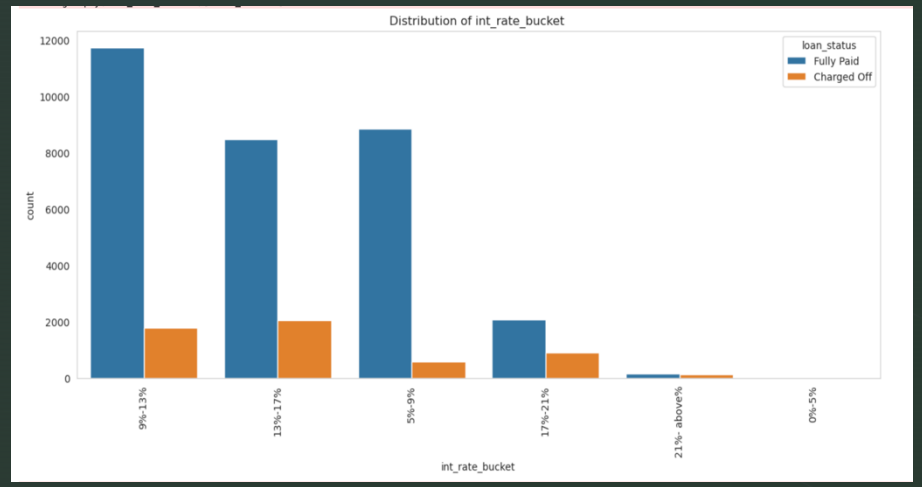
Annual income analysis

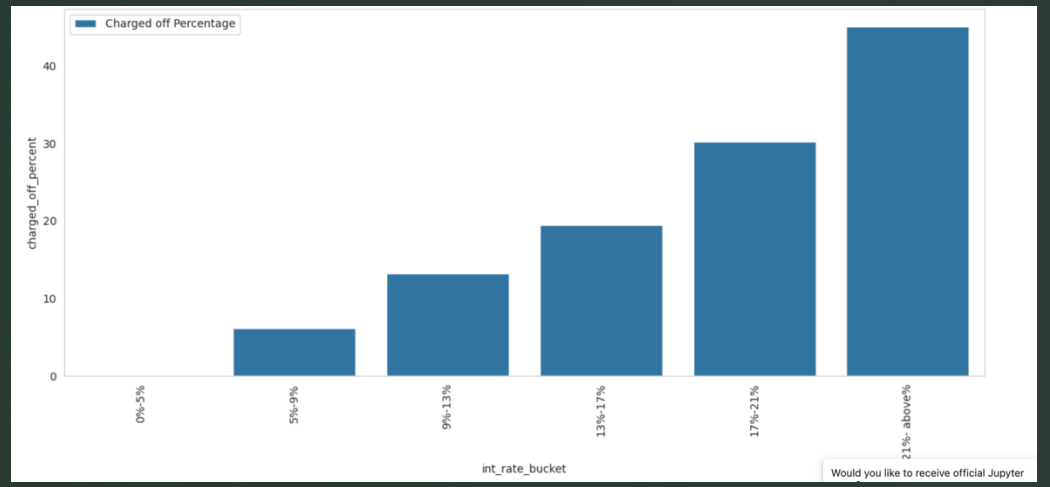




Upon analysis, we can observe that loans given for people with annual income less than 20k has a higher chance of default.

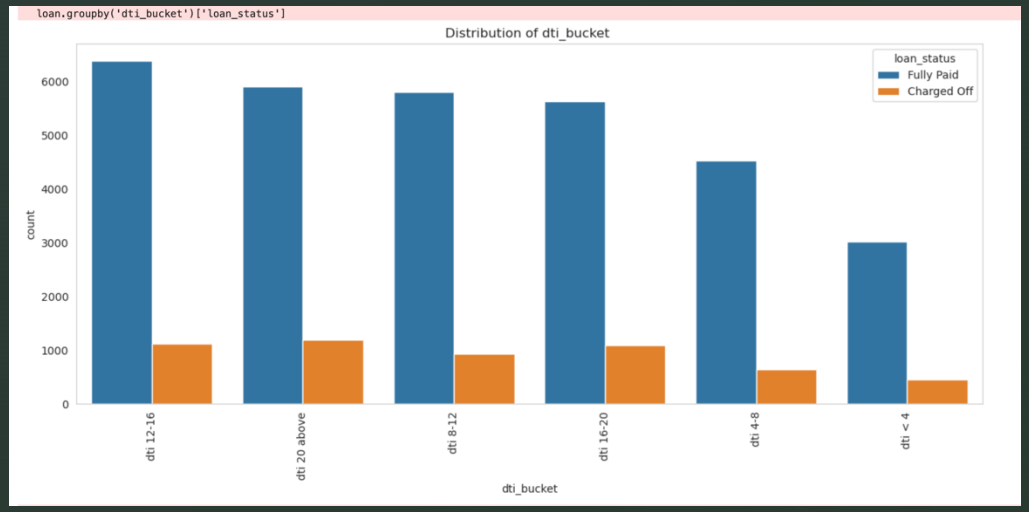
Interest rate analysis

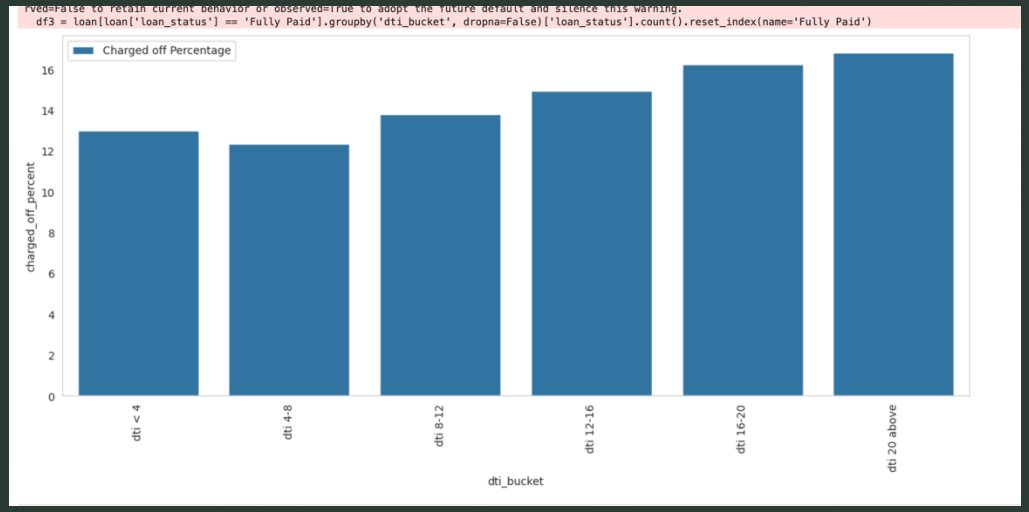




Upon analysis, we can observe that loans given with higher interest rate has higher chance of default

DTI Analysis





Upon analysis, we can observe that applicants with high dti rate has more chance of loan default