Lending Club Case Study

Study Group Members:

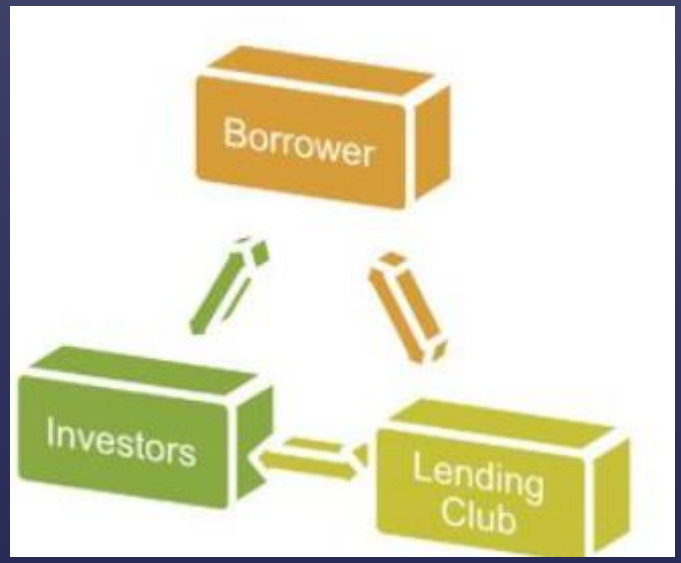
Karan Handa
Sonia Raina

Agenda

- Business Objectives
- Approach
- Analysis
- Key findings
- Recommendations

Business Objectives

- **Minimizing Credit Loss**: The primary business objective is to minimize credit loss for the company. As a leading online loan marketplace, the company faces financial losses due to loan defaults, particularly from borrowers categorized as 'charged-off.' The primary goal is to identify and reduce the risk associated with these applicants to lower credit loss, which is a significant financial concern for the company**.**
- **Risk Assessment and Portfolio Management:** By identifying the driving factors or key variables that are strong indicators of loan default, the company aims to improve its risk assessment processes. This knowledge can be applied to make informed decisions about which loans to extend and to optimize the composition of its loan portfolio.
- **Enhancing Lending Efficiency:** The company's fast online interface is a key feature, and by identifying risky loan applicants more effectively through exploratory data analysis (EDA), the company can streamline its lending process. This not only reduces credit loss but also ensures a smoother and more reliable lending experience for borrowers.



Approach

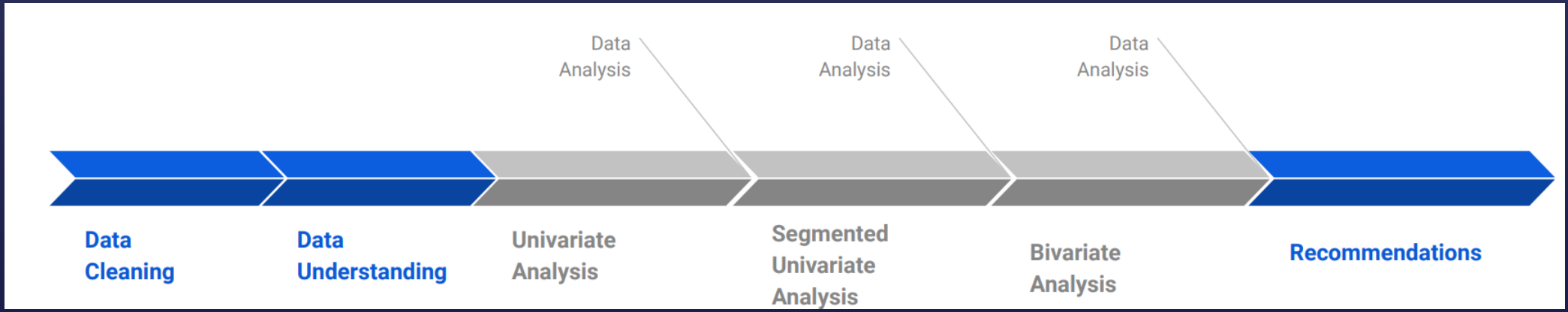
Our approach focused on establishing meaningful relationships between loan status and key parameters. We employed the following steps:

1. **Univariate Analysis:** We conducted univariate analysis both on the entire dataset and specifically for 'Charged Off' and 'Fully Paid' loans to identify relevant patterns.
2. **Shortlisting Key Columns:** After identifying important columns, we proceeded to bivariate analysis tailored to the type of columns.
3. **Numerical Columns:** For numerical columns, we employed box plots to assess differences in median values, seeking insights into their impact on loan status.
4. **Categorical Columns:** We utilized pivot tables to understand how parameter variations affect the percentages of 'Charged Off' and 'Fully Paid' loans.
5. **Bar Charts:** Lastly, for categorical columns, we employed bar charts to isolate and highlight significant categories that influence loan status.

Analysis

Data cleaning

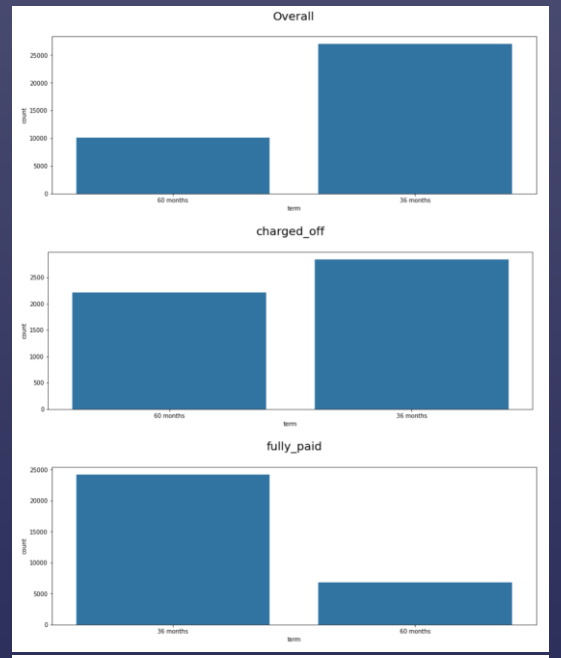
- **Initial Data Assessment:** Upon data import, our initial step involved a thorough examination of the data, its columns, and overall structure. We also referred to the data dictionary to gain a comprehensive understanding of column relationships.
- **Handling Null Values:** To ensure data quality, we took measures to address missing values. We began by dropping columns with 100% null values, as well as those with only a single unique value, which offered no analytical relevance.
- **Column Selection:** Following this, we further refined the dataset by removing columns that were either irrelevant to our analysis or not conducive to our objectives. Examples include 'desc' and 'url' columns.
- **Row Deletion:** As part of our data cleaning approach, we opted to remove rows with null values to enhance dataset cleanliness.
- **Data Standardization:** We converted interest rate columns to float data type and meticulously handled leading/trailing spaces, ensuring data uniformity and consistency.

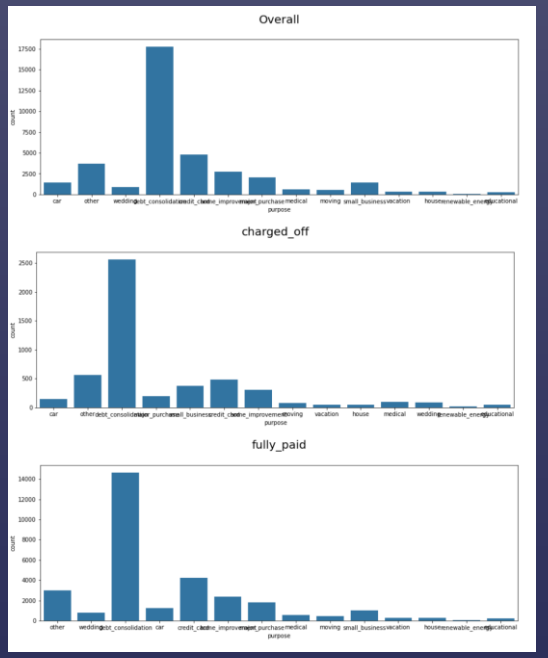


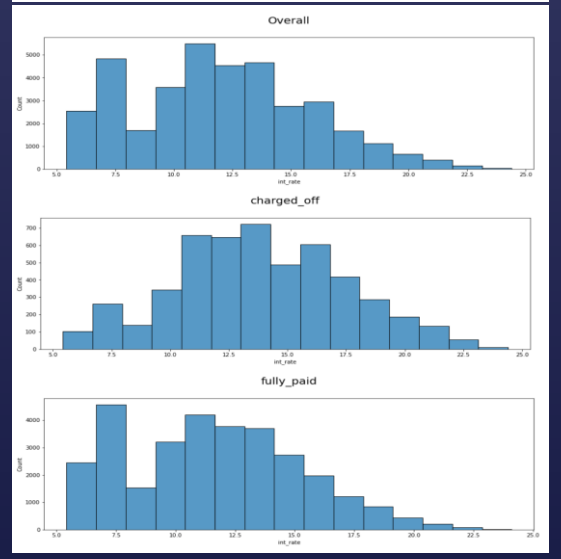
Analysis

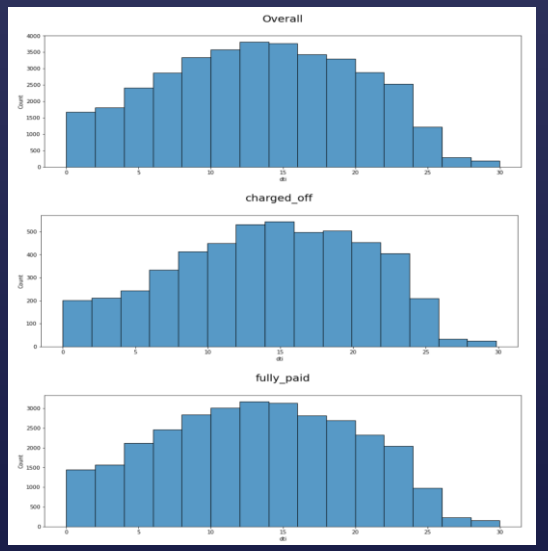
Univariate analysis

- **Data Segmentation:** Our initial step involved categorizing target columns into three distinct categories: categorical, continuous, and extra columns (e.g., 'loan_id') that were deemed irrelevant for our analysis.
- **DataFrame Divisions:** We divided the dataset into three separate dataframes: 'Overall,' 'Fully Paid,' and 'Charged Off,' allowing us to assess each category independently.
- **Categorical Columns:** For categorical columns, we employed count plots to visualize the frequency distribution of each attribute, gaining valuable insights into their patterns and importance.
- **Continuous Columns:** We created histograms to analyze the distribution of continuous columns within each dataframe. This visualization aided in evaluating the data's statistical characteristics.
- **Informed Decision-Making:** By applying these analytical steps to all three dataframes simultaneously, we efficiently garnered directional insights, guiding our decisions on whether specific columns warranted further bivariate analysis.





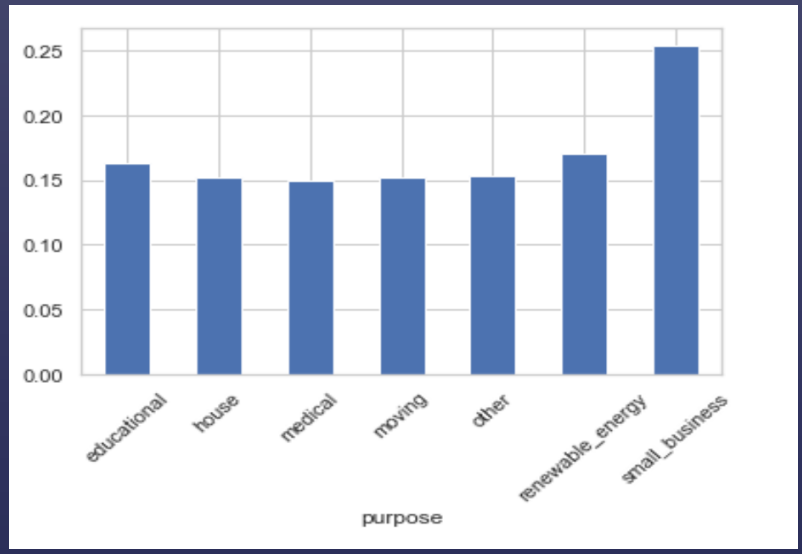


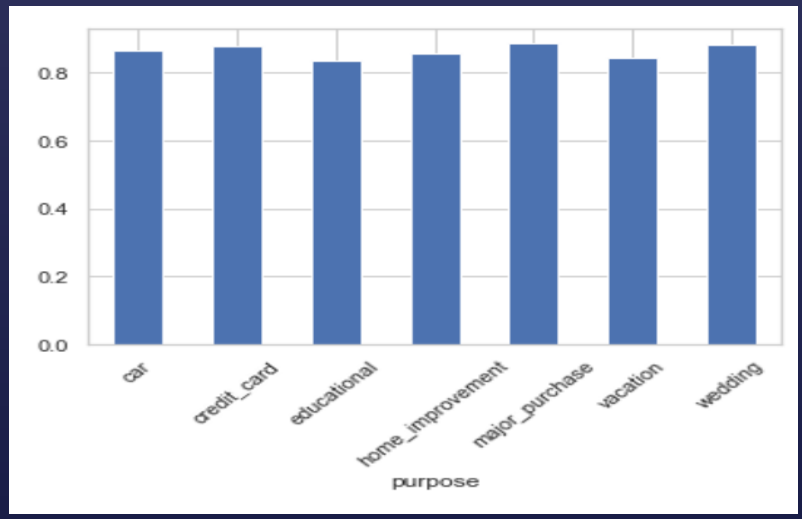


Analysis

Bivariate analysis – 1) Loan Status vs Purpose

- **Pivot Analysis:** To examine the relationship between loan status and the purpose of the loan, we constructed a pivot table. This pivot table provided the relative percentages of 'Charged Off' and 'Fully Paid' loans for each unique purpose in the dataset.
- **Shortlisting High 'Charged Off' Purposes:** Our next step involved comparing the relative 'Charged Off' percentages with the median 'Charged Off' percentage for all purposes. We identified and shortlisted purpose categories with 'Charged Off' percentages exceeding the median value. This selection was based on the logic of using the median 'Charged Off' percentage as a threshold to identify high 'Charged Off' purposes.
- **Purpose Overlap Analysis:** Similarly, we shortlisted purpose categories with 'Fully Paid' percentages exceeding the median 'Fully Paid' percentage. This step allowed us to explore potential overlaps and select the most relevant purpose categories for further in-depth analysis and recommendations.



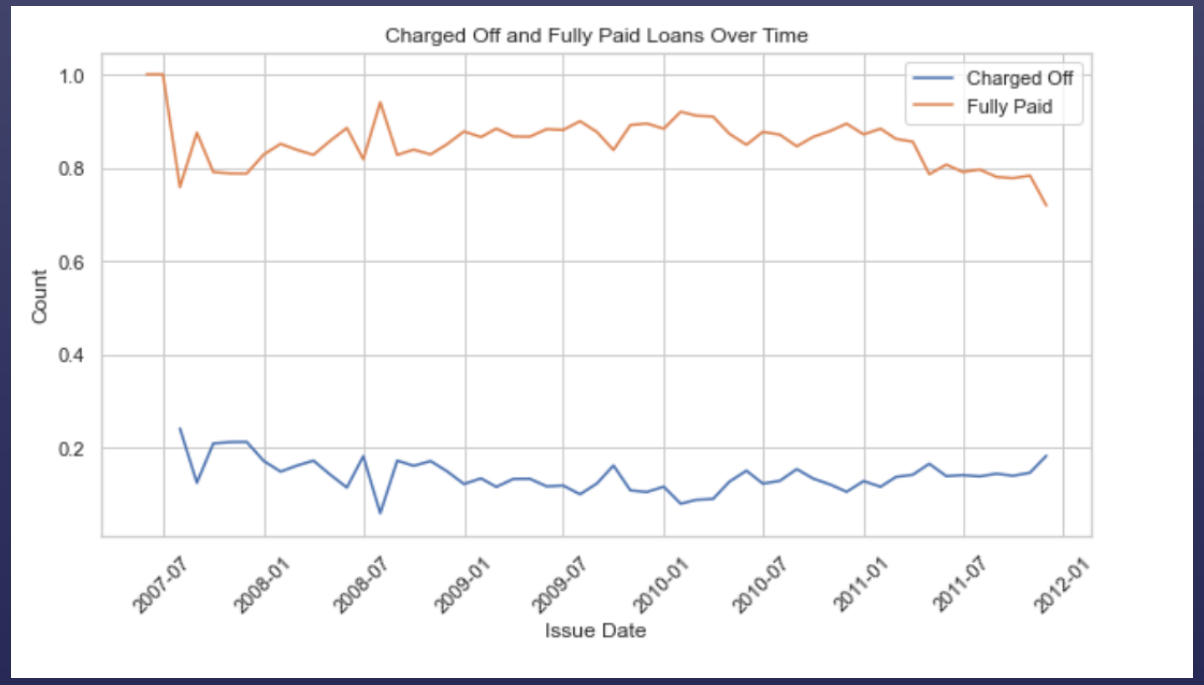


Loans categorized under 'small_business' and 'renewable_energy' purposes are associated with higher risk, with percentages of 'Charged Off' loans exceeding the median threshold.

Analysis

Bivariate analysis – 2) Loan Status – Year-Month analysis

- **Pivot Analysis:** Our analysis delved into the dynamic relationship between loan status and the temporal dimension, specifically the year and month of loan origination. We commenced by constructing a pivot table, a fundamental analytical tool. This table enabled us to assess the relative percentages of 'Charged Off' and 'Fully Paid' loans, providing valuable insights into how loan status varies across different time periods.
- **Visual Representation:** Building on the insights gleaned from the pivot analysis, our subsequent step involved visualizing the data. We employed a line chart that plotted the year-month on the x-axis against the loan status on the y-axis. This dynamic visualization effectively illustrated the fluctuations in 'Charged Off' and 'Fully Paid' loan categories over time, enhancing our understanding of temporal trends.



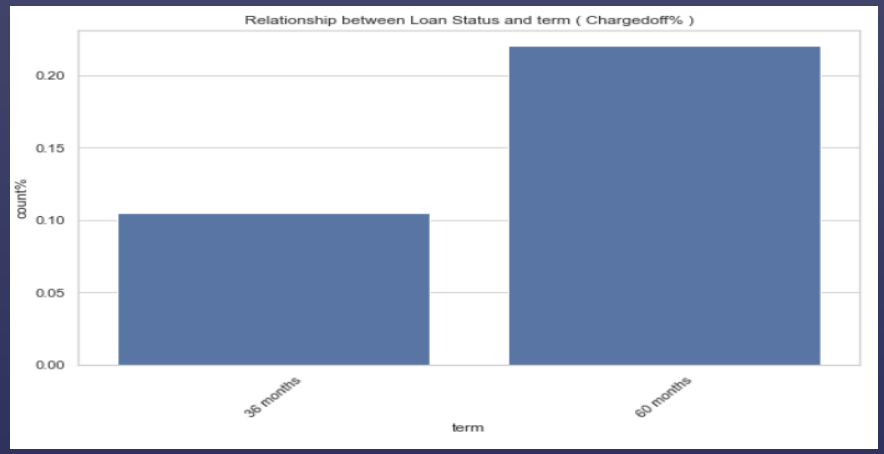
The Lending Club has effectively maintained 'Charged Off' percentages below 20%, demonstrating a remarkable achievement. A noticeable decline in 'Charged Off' percentages commences in January 2008, with a particularly steep drop in July 2008. Investigating the best practices employed during this period could provide valuable insights for the company's risk management strategies.

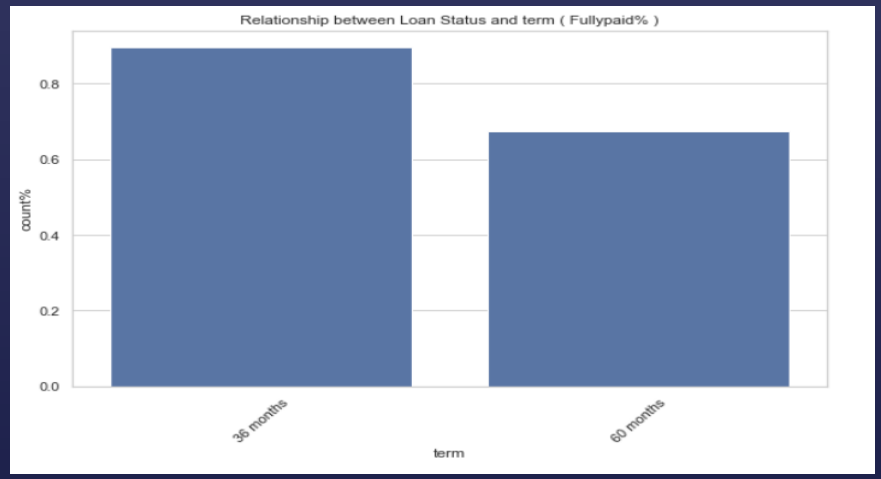
However, as we approach the beginning of 2011, the 'Charged Off' curve appears to stabilize and even increase towards the end of the observed period. A deeper analysis is recommended to uncover the factors contributing to this trend, as it may necessitate adjustments to the company's lending practices or risk assessment protocols.

Analysis

Bivariate analysis – 3) Loan Status - term analysis

- **Pivot Analysis:** Our analysis explored the dynamic relationship between loan status and the loan term. We initiated this investigation by constructing a pivotal pivot table, a fundamental tool in our analytical toolkit. This pivot table allowed us to gain a comprehensive understanding of the relative percentages of 'Charged Off' and 'Fully Paid' loans, shedding light on the variations in loan status across different term categories.
- **Visual Representation:** Building upon the insights extracted from our pivot analysis, our subsequent step involved the visual presentation of data. We created informative charts that depicted the 'Charged Off' and 'Fully Paid' percentages from the pivot table, enabling a direct comparison to determine which term category is more susceptible to loan default.



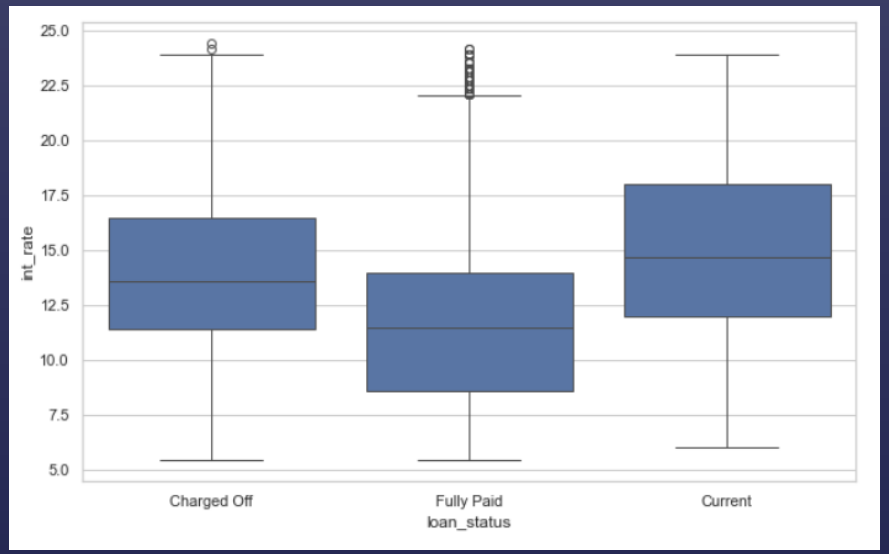


The percentage of 'Charged Off' loans rises notably from 10% to 22% as the term of the loan increases from 36 to 60 months. This indicates that loans with a 60-month term are more likely to default, suggesting a correlation between longer-term loans and higher default rates.

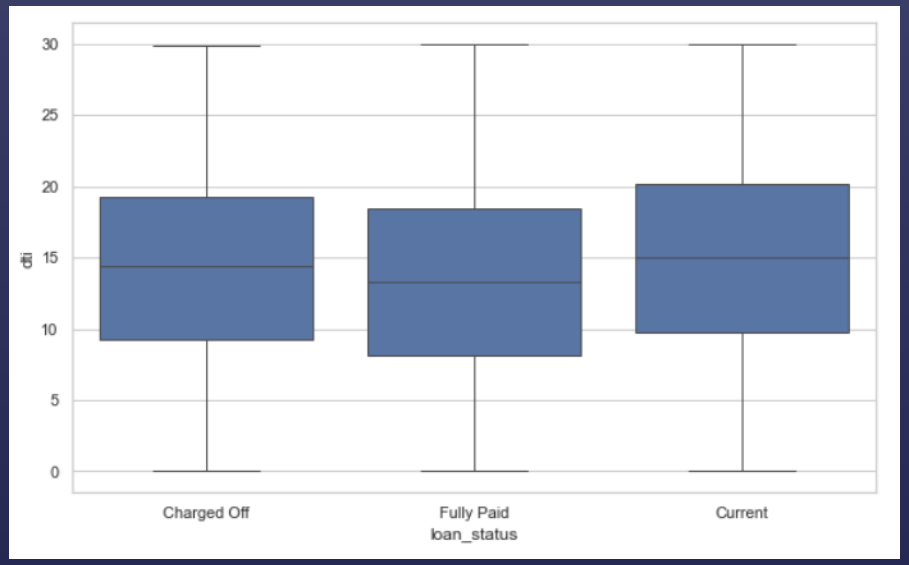
Analysis

Bivariate analysis – 4) Loan status and int_rate & 5) Loan status and dti

We utilized box plots to compare 'Loan Status' against 'Interest Rate' (int_rate) and 'Debt-to-Income Ratio' (dti), assessing the significant differences in how these parameters vary across different loan statuses.



Evidently, loans with higher interest rates (int_rate) exhibit a higher likelihood of default. This observation is supported by a significant difference in the median interest rates between 'Charged Off' and 'Fully Paid' loans.

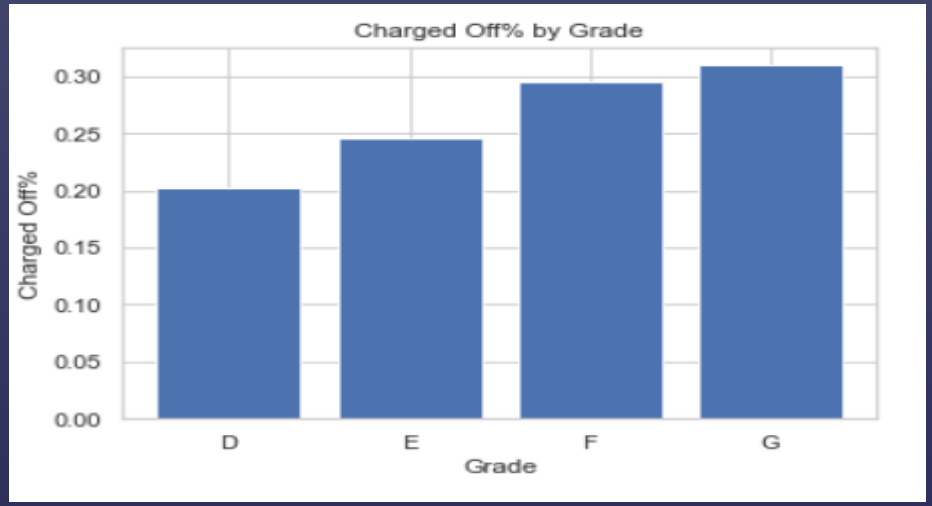


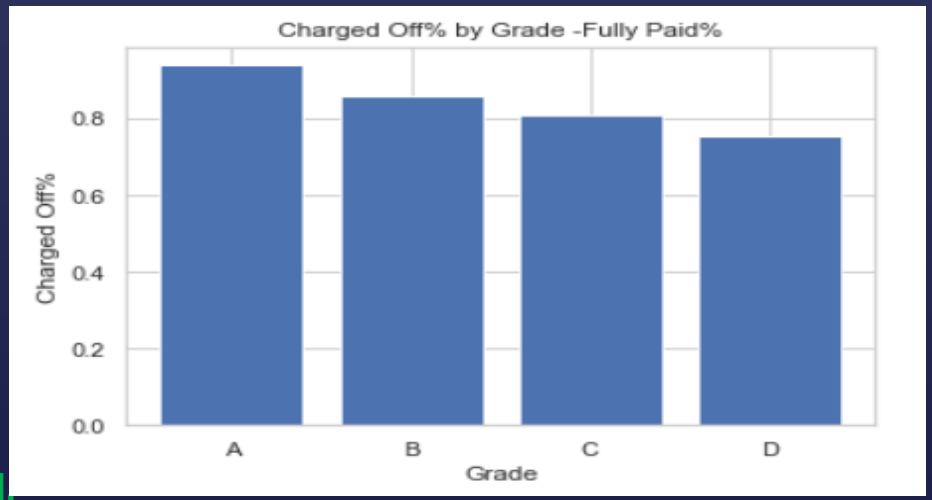
While the debt-to-income ratio (dti) for 'Charged Off' loans is slightly higher, the difference is significantly marginal. Therefore, it's inconclusive to assert a definitive relationship based on this parameter alone

Analysis

Bivariate analysis – 6) Loan status and grade

- **Pivot Analysis:** To examine the relationship between loan status and the grade of the loan, we constructed a pivot table. This pivot table provided the relative percentages of 'Charged Off' and 'Fully Paid' loans for each unique grade in the dataset.
- **Shortlisting High 'Charged Off' grades:** Our next step involved comparing the relative 'Charged Off' percentages with the median 'Charged Off' percentage for all the grade categories. We identified and shortlisted 'grade' categories with 'Charged Off' percentages exceeding the median value. This selection was based on the logic of using the median 'Charged Off' percentage as a threshold to identify high 'Charged Off' grade categories.
- **Grade Overlap Analysis:** Similarly, we shortlisted grade categories with 'Fully Paid' percentages exceeding the median 'Fully Paid' percentage. This step allowed us to explore potential overlaps and select the most relevant grade categories for further in-depth analysis and recommendations.



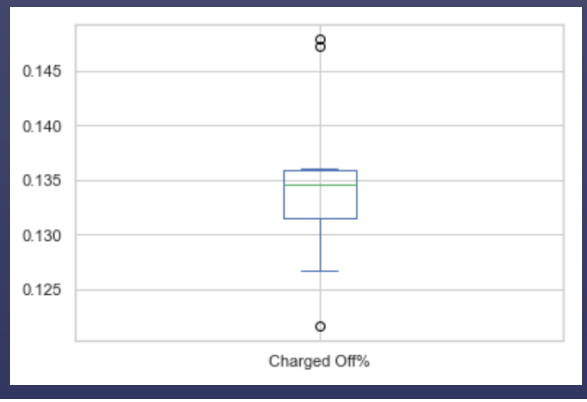


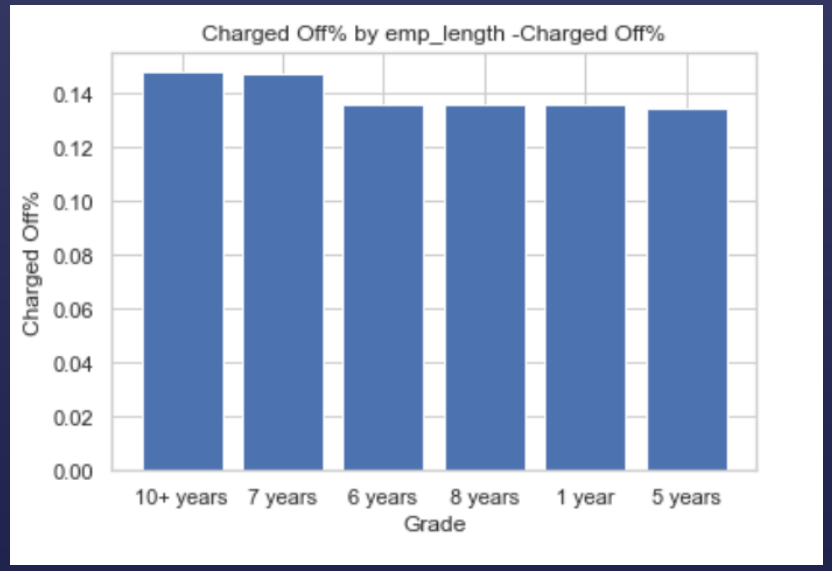
Categories D, E, F, and G loans exhibit a higher susceptibility to default, whereas categories A, B, and C are more likely to be fully paid.

Analysis

Bivariate analysis – 7) Loan status and emp_length

- **Pivot Analysis:** Our analysis explored the relationship between loan status and the borrower's employment length in years. We initiated this exploration by constructing a pivotal pivot table. This table provided us with relative percentages of 'Charged Off' and 'Fully Paid' loans, offering insights into how loan status varies across different employment lengths.
- **Shortlisting High 'Charged Off' Categories:** Following our pivot analysis, we proceeded to assess the 'Charged Off' percentages in relation to the median 'Charged Off' percentage for all employment length categories. We identified and shortlisted employment length categories where 'Charged Off' percentages exceeded the median value, using the median as a threshold to highlight high 'Charged Off' employment lengths.
- **Category Overlap Analysis:** In parallel, we also shortlisted employment length categories where 'Fully Paid' percentages exceeded the median 'Fully Paid' percentage. This step enabled us to investigate potential overlaps and pinpoint the most pertinent employment length categories for a more comprehensive analysis and recommendations.



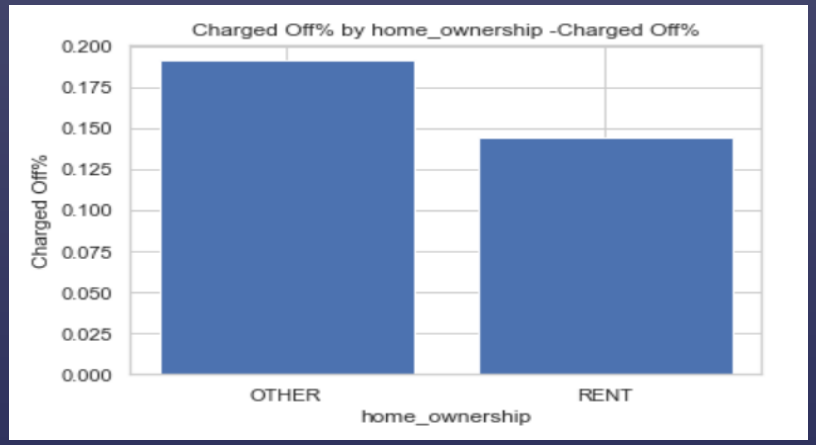


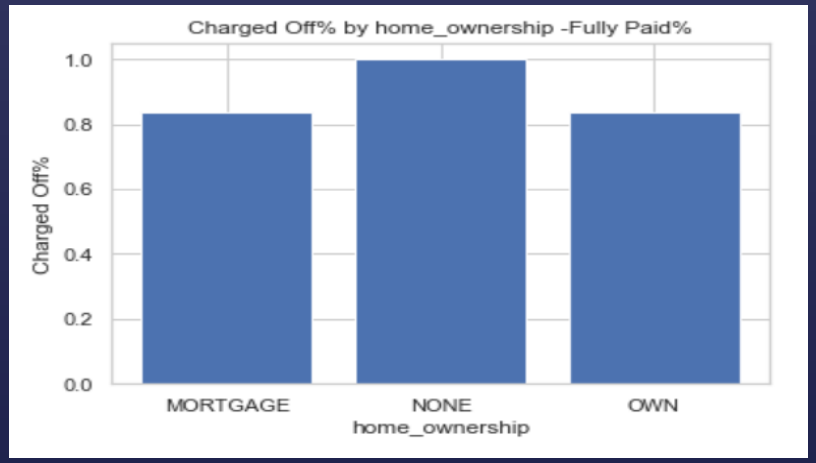
Borrowers with employment lengths of 10 years or more and 7 years are more susceptible to loan default. One hypothesis could attribute this to their later career stage, potentially resulting in limited time for loan repayment. Additionally, these individuals may have increased personal liabilities, contributing to a higher rate of defaults in the later stages of their careers.

Analysis

Bivariate analysis – 8) Loan status and Home Ownership

- **Pivot Analysis:** Our analysis explored the relationship between loan status and the borrower's home ownership status. We initiated this exploration by constructing a pivotal pivot table. This table provided us with relative percentages of 'Charged Off' and 'Fully Paid' loans, offering insights into how loan status varies across different home ownership status.
- **Shortlisting High 'Charged Off' Categories:** Following our pivot analysis, we proceeded to assess the 'Charged Off' percentages in relation to the median 'Charged Off' percentage for all home ownership status categories. We identified and shortlisted home ownership status categories where 'Charged Off' percentages exceeded the median value, using the median as a threshold to highlight high 'Charged Off' home ownership status.
- **Category Overlap Analysis:** In parallel, we also shortlisted home ownership status categories where 'Fully Paid' percentages exceeded the median 'Fully Paid' percentage. This step enabled us to investigate potential overlaps and pinpoint the most pertinent employment length categories for a more comprehensive analysis and recommendations.



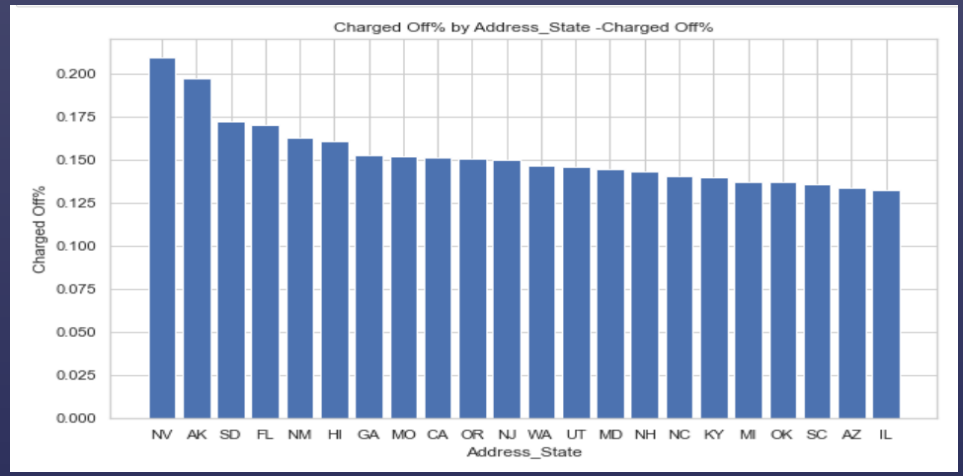


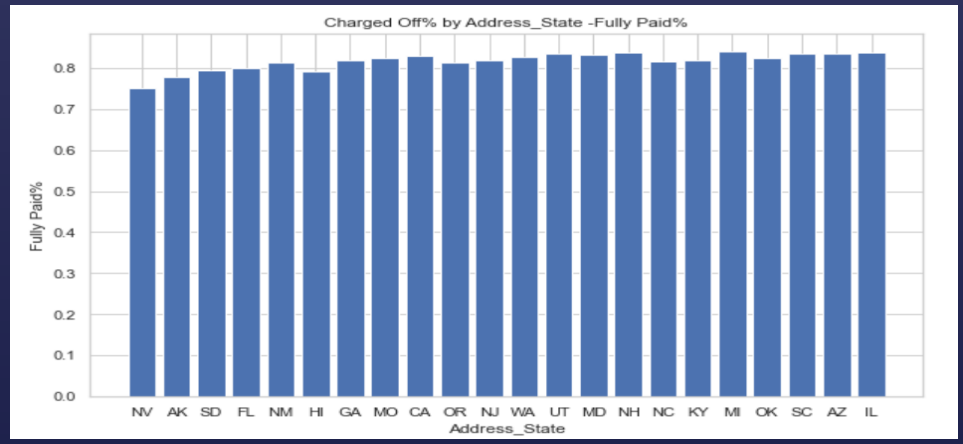
No significant relationship has been observed between homeownership categories and the 'Charged Off%' percentage.

Analysis

Bivariate analysis – 9) Loan status and addr_state

- **Pivot Analysis:** Our analysis delved into the relationship between loan status and the borrower's geographical location, specifically their state of residence (addr_state). We initiated this exploration by constructing a pivotal pivot table. This table provided us with relative percentages of 'Charged Off' and 'Fully Paid' loans, yielding valuable insights into how loan status varies across different states.
- **Shortlisting High 'Charged Off' States:** Following our pivot analysis, we proceeded to assess the 'Charged Off' percentages concerning the median 'Charged Off' percentage for all states. We identified and shortlisted states where 'Charged Off' percentages exceeded the median value, using the median as a threshold to highlight high 'Charged Off' states.
- **State Overlap Analysis:** In parallel, we also shortlisted states where 'Fully Paid' percentages exceeded the median 'Fully Paid' percentage. This step allowed us to investigate potential overlaps and pinpoint the most pertinent state categories for a more comprehensive analysis and recommendations.



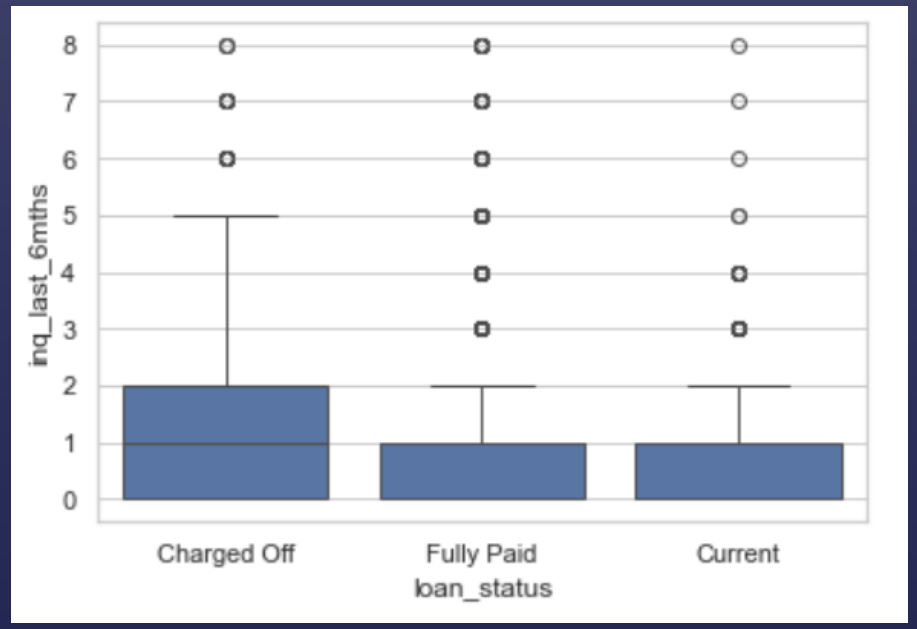


States such as NV, AK, SD, FL, NM, HI, GA, MO, and CA, where the 'Charged Off%' exceeds the median 'Charged Off%' percentage, should be subject to further in-depth analysis. These states demonstrate a higher susceptibility to default loans.

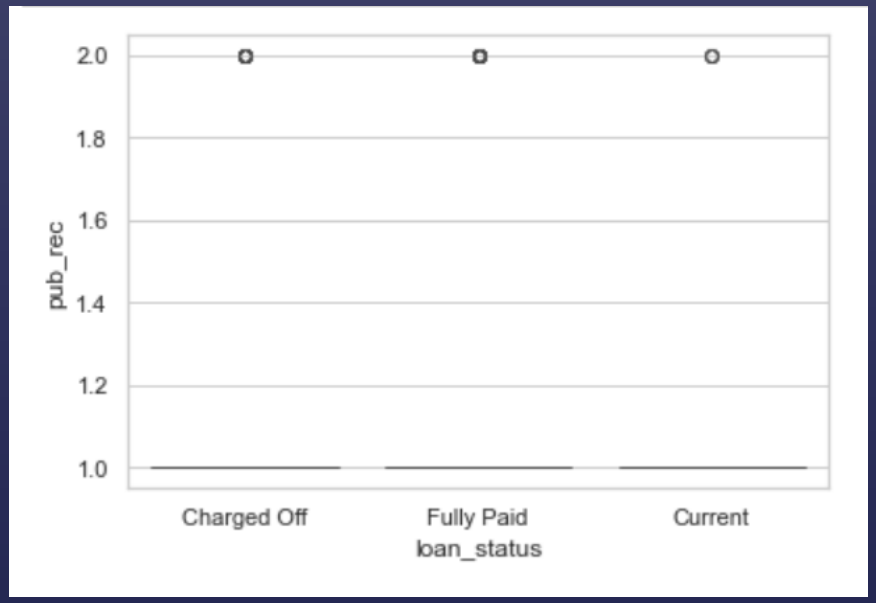
Analysis

Bivariate analysis – 10) loan_status vs inq_last_6mths &11) loan_status vs pub_rec

We utilized box plots to compare 'Loan Status' against '# inquiries in past 6 months ' (inq_last_6mths) and '# derogatory public records ' (pub_rec), assessing the significant differences in how these parameters vary across different loan statuses.



Charged-off loans exhibit a substantially higher number of inquiries in the past 6 months compared to fully paid loans. This disparity suggests that the frequency of inquiries is a noteworthy factor in loan default.

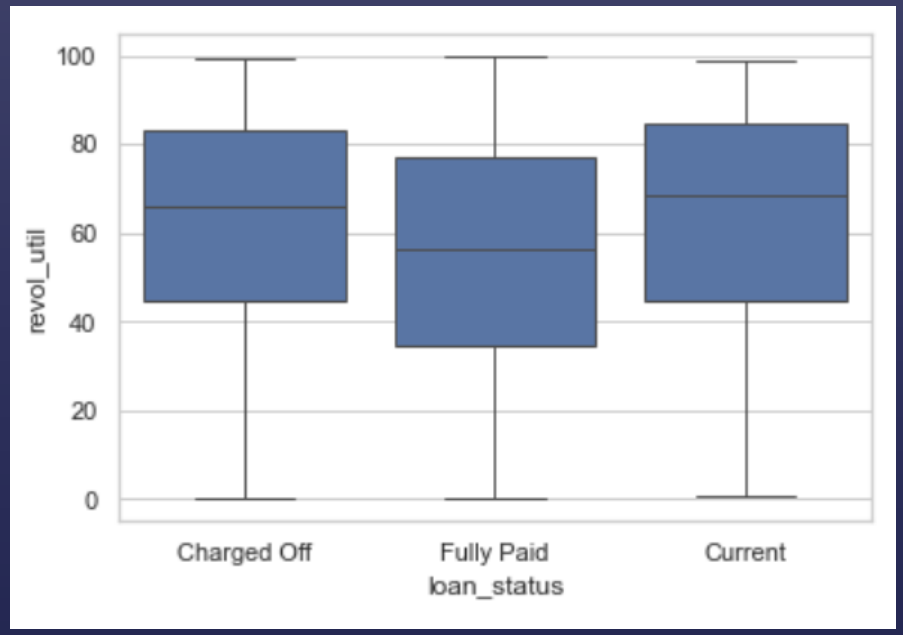


There is not clear relationship between pub_rec and loan status inferred from the above chart

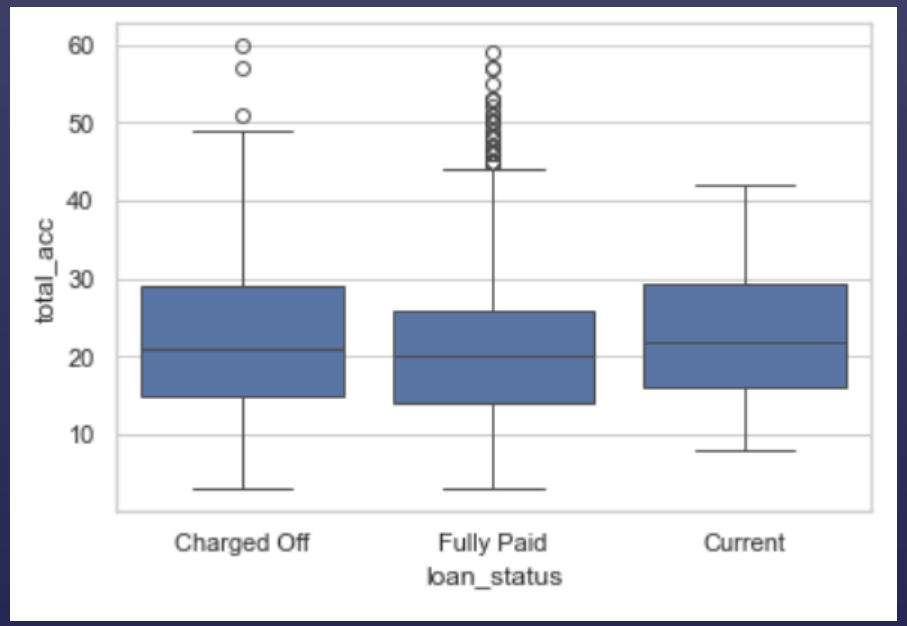
Analysis

Bivariate analysis – 12) loan_status vs revol_util & 13) loan_status vs total_acc

We utilized box plots to compare 'Loan Status' against 'Revolving line utilization rate' (revol_util) and '# credit lines currently in the borrower's credit file ' (total_acc), assessing the significant differences in how these parameters vary across different loan statuses.



It can be inferred that a higher revolving line utilization rate is a correlated indicator of default loans. This observation underscores the importance of monitoring and managing this financial metric when assessing the risk of loan defaults.



We can infer that a higher number of credit lines currently in the borrower's credit file can be a correlated indicator of default loans, as evidenced by the higher median for charged-off loans. This underscores the significance of this factor as a potential predictor of loan defaults.

Key findings (1/2)

- Loans categorized under 'small_business' and 'renewable_energy' purposes are associated with higher risk, with percentages of 'Charged Off' loans exceeding the median threshold.
- The Lending Club has effectively maintained 'Charged Off' percentages below 20%, demonstrating a remarkable achievement. A noticeable decline in 'Charged Off' percentages commences in January 2008, with a particularly steep drop in July 2008. Investigating the best practices employed during this period could provide valuable insights for the company's risk management strategies.
- However, as we approach the beginning of 2011, the 'Charged Off' curve appears to stabilize and even increase towards the end of the observed period. A deeper analysis is recommended to uncover the factors contributing to this trend, as it may necessitate adjustments to the company's lending practices or risk assessment protocols.
- The percentage of 'Charged Off' loans rises notably from 10% to 22% as the term of the loan increases from 36 to 60 months. This indicates that loans with a 60-month term are more likely to default, suggesting a correlation between longer-term loans and higher default rates.
- Evidently, loans with higher interest rates (int_rate) exhibit a higher likelihood of default. This observation is supported by a significant difference in the median interest rates between 'Charged Off' and 'Fully Paid' loans.
- While the debt-to-income ratio (dti) for 'Charged Off' loans is slightly higher, the difference is significantly marginal. Therefore, it's inconclusive to assert a definitive relationship based on this parameter alone
- Categories D, E, F, and G loans exhibit a higher susceptibility to default, whereas categories A, B, and C are more likely to be fully paid.

Key findings (2/2)

- Borrowers with employment lengths of 10 years or more and 7 years are more susceptible to loan default. One hypothesis could attribute this to their later career stage, potentially resulting in limited time for loan repayment. Additionally, these individuals may have increased personal liabilities, contributing to a higher rate of defaults in the later stages of their careers.
- No significant relationship has been observed between homeownership categories and the 'Charged Off%' percentage.
- States such as NV, AK, SD, FL, NM, HI, GA, MO, and CA, where the 'Charged Off%' exceeds the median 'Charged Off%' percentage, should be subject to further in-depth analysis. These states demonstrate a higher susceptibility to default loans.
- Charged-off loans exhibit a substantially higher number of inquiries in the past 6 months compared to fully paid loans. This disparity suggests that the frequency of inquiries is a noteworthy factor in loan default.
- There is not clear relationship between pub_rec and loan status inferred from the above chart
- It can be inferred that a higher revolving line utilization rate is a correlated indicator of default loans. This observation underscores the importance of monitoring and managing this financial metric when assessing the risk of loan defaults.
- We can infer that a higher number of credit lines currently in the borrower's credit file can be a correlated indicator of default loans, as evidenced by the higher median for charged-off loans. This underscores the significance of this factor as a potential predictor of loan defaults.

Recommendations

- Scrutinize borrowers with 10+ or 7 years of employment for stricter approval.
- Prioritize stronger indicators over homeownership in risk assessment.
- Tailor risk strategies for high-risk states: NV, AK, SD, FL, NM, HI, GA, MO, CA.
- Include recent inquiries in risk assessment.
- Combine public records with other factors for better insights.
- Manage high revolving line utilization for risk mitigation.
- Utilize the number of credit lines as a risk predictor.
- Apply stricter criteria to 'small_business' and 'renewable_energy' loans.
- Implement successful risk management practices from 2008.
- Use caution with 60-month loans; consider stricter criteria.
- Adjust rates for higher interest loans.
- Consider debt-to-income ratio, but focus on other vital indicators.
- Tailor loan terms for D, E, F, and G grades, and favor A, B, C grades.

Thank you