Credit Risk Modelling: Predicting Loan Defaults

Credit Risk Modelling is a critical tool in the financial sector, used to predict the likelihood of borrowers defaulting on their loans. By analyzing past data, these models can assist banks in making informed decisions, minimizing risk, and ensuring that lending processes are both fair and profitable.

Objective of the Project

The primary goal of this project is to build a robust machine learning model that can accurately predict whether a loan will default based on borrower characteristics such as income, employment length, credit history, and loan details.

Steps Involved

1. **Data Cleaning**: Removing inconsistencies and handling missing data.
2. **Exploratory Data Analysis (EDA)**: Understanding relationships between variables and identifying key risk factors.
3. **Feature Engineering**: Creating new features to improve model performance.
4. **Model Training**: Testing multiple algorithms such as XGBoost, Random Forest, and Logistic Regression.
5. **Model Evaluation**: Comparing performance metrics like Accuracy, Precision, and Recall to choose the best model.

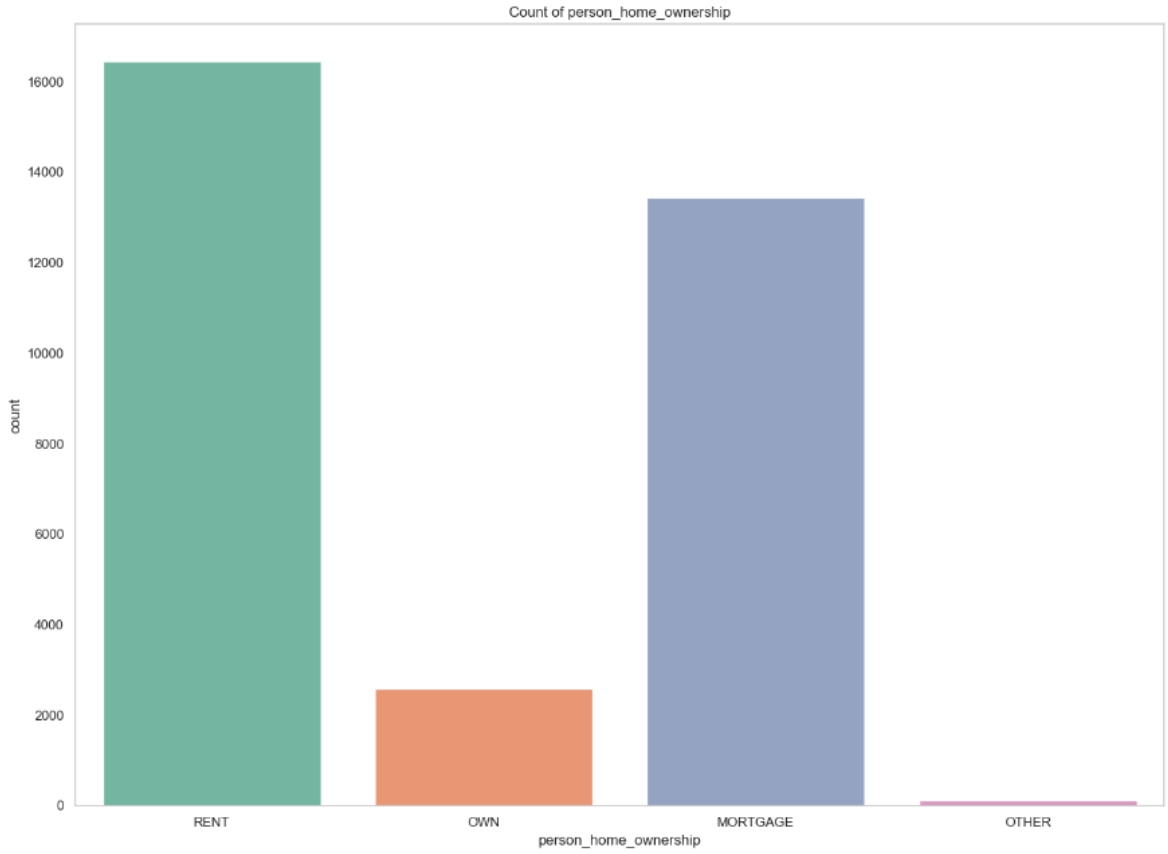
Challenges Faced

One of the significant challenges encountered was the **class imbalance** in the dataset, with a much larger proportion of non-defaulted loans. This required careful selection of metrics like Precision and Recall to avoid misleading results from a skewed dataset.

Source Code: -

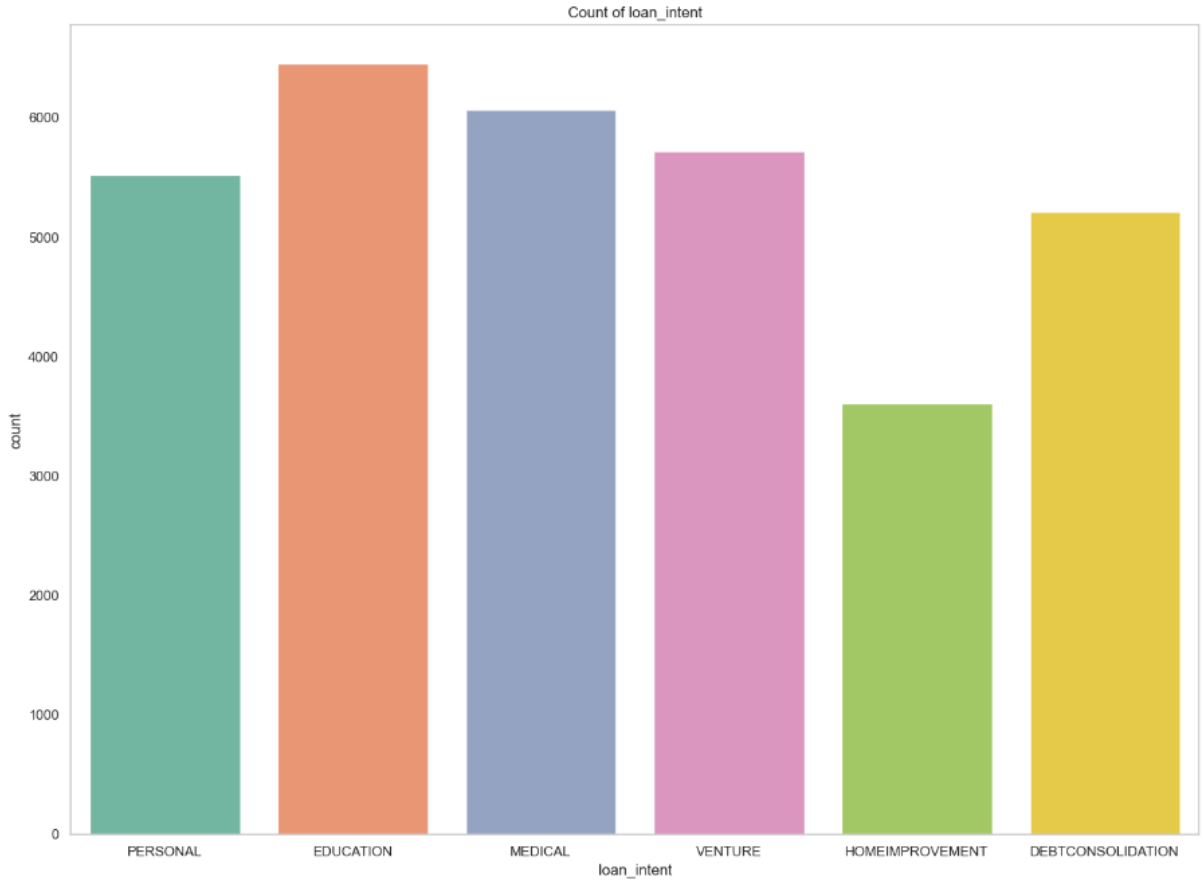
Click here for Source Code

Results from the graphs



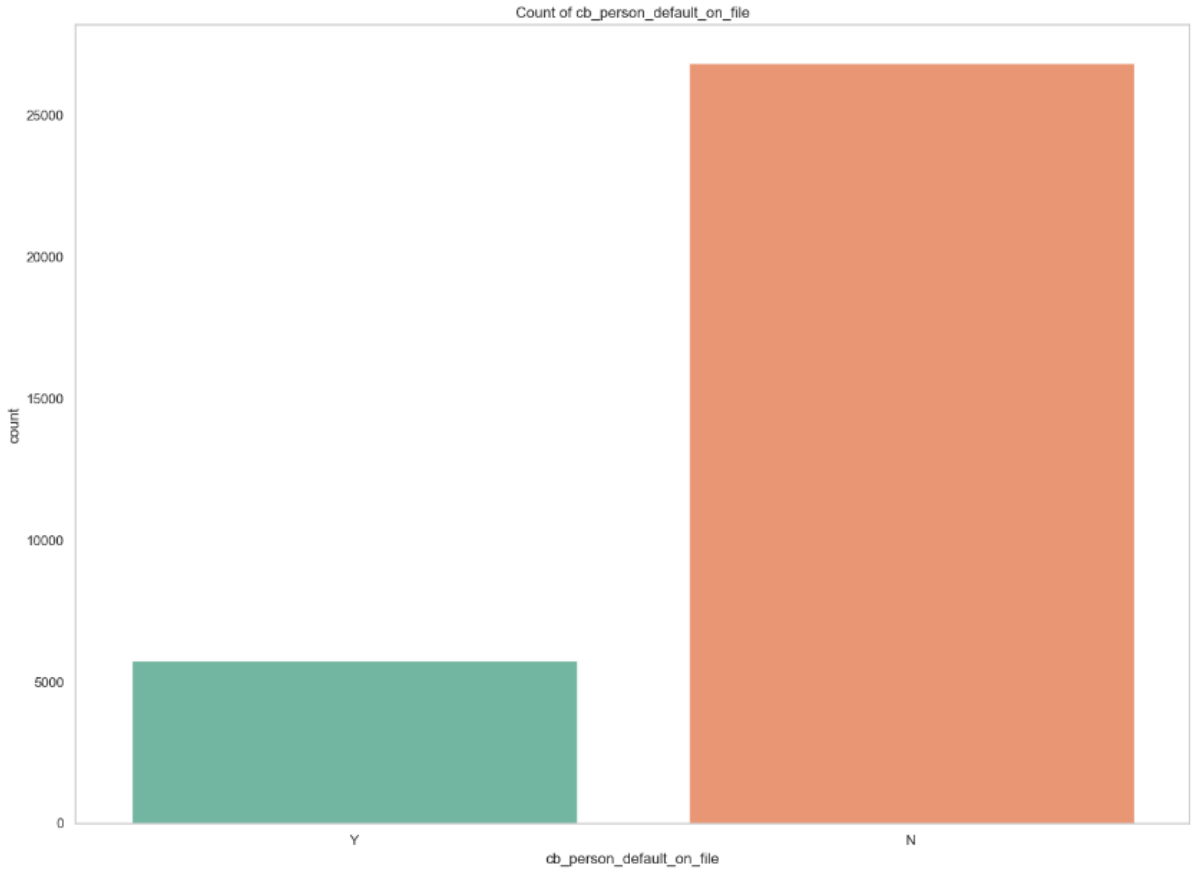
a) Count of person_home_ownership :

- Categories: This bar chart shows the distribution of different home ownership statuses: Rent, Own, Mortgage, and Other.
- Results:
    - The majority of individuals in this dataset are renting homes, followed by those with mortgages.
    - Fewer individuals fall under the Own category (those who own their homes outright), and very few are classified as Other.
    - This indicates that a significant portion of loan applicants in the dataset are renting or mortgaging homes, potentially giving insights into their financial stability.
- Implications: Borrowers with mortgages or renters may have higher loan demands, while the "Own" category may indicate borrowers with potentially lower financial risk.



b) Count of loan_intent :

- Categories: This bar chart represents different loan purposes: Personal, Education, Medical, Venture, Home Improvement, and Debt Consolidation.
- Results:
    - Education and Medical loans make up the largest portions of the dataset, followed closely by Personal loans.
    - Venture and Debt Consolidation loans are also common, but Home Improvement loans represent the smallest portion.
- Implications: This plot provides an overview of the reasons why people are applying for loans. It can be useful to study whether certain loan intents are more risky in terms of defaults. For example, loans for ventures or debt consolidation could carry more risk than those for education or medical purposes.



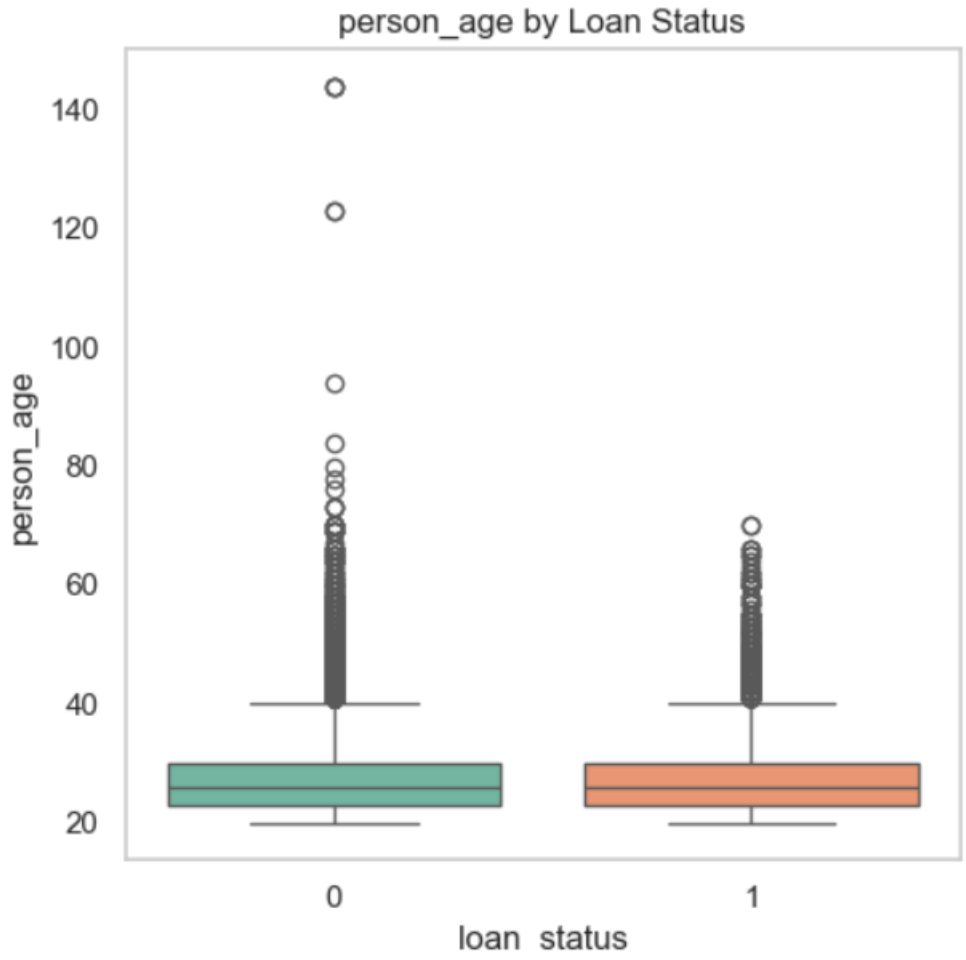
c) Count of cb_person_default_on_file :

- Categories: This bar chart shows whether a person has a history of default in their credit file: Y (Yes) or N (No).
- Results:
    - A large majority of individuals in the dataset do not have a history of defaults (category N), while a smaller portion has previously defaulted on a loan (Y).
- Implications: This graph highlights the imbalance in the dataset between defaulters and non-defaulters. Since credit risk modeling aims to predict loan defaults, the imbalance between categories can impact model performance. You might need to handle this imbalance by using techniques like SMOTE (Synthetic Minority Oversampling Technique) or adjusting class weights in the model to ensure it does not overly favor predicting non-defaults.

d) The box plots provided below show the distribution of various features against loan_status. The loan_status variable is binary, where:

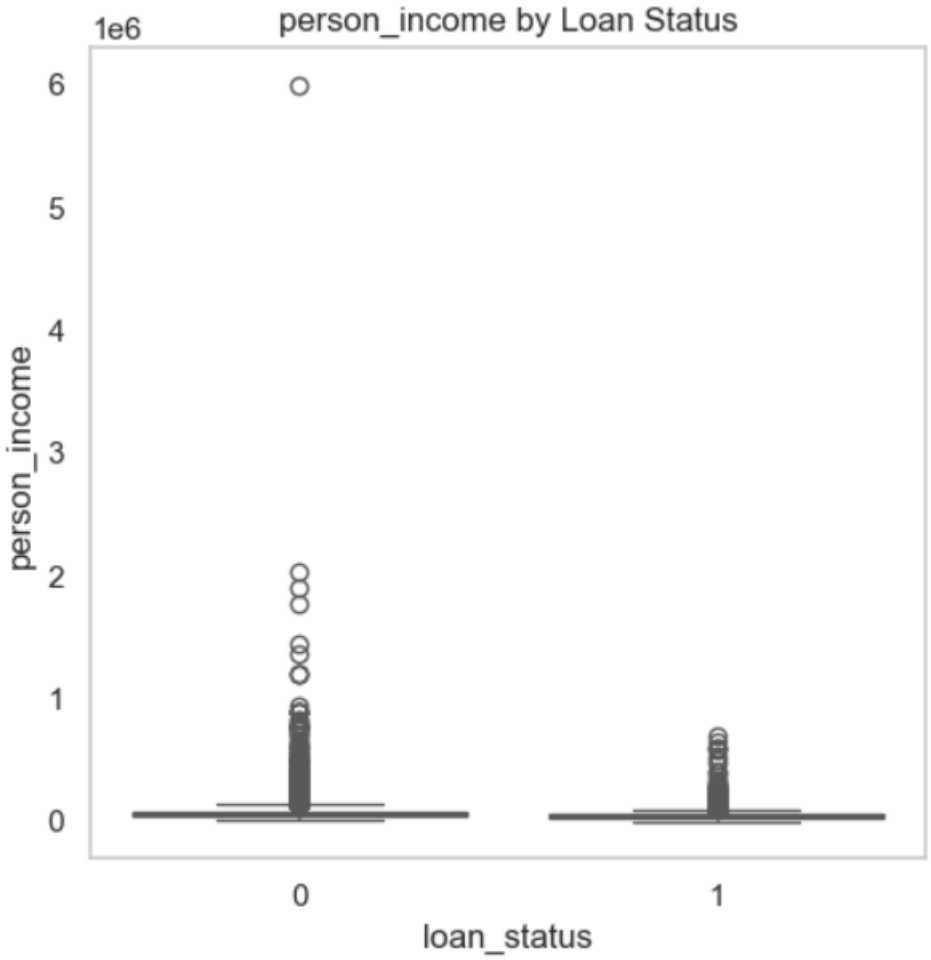
- 0: Indicates the loan was not defaulted.
- 1: Indicates the loan was defaulted.

Each plot provides insights into how the distribution of a feature differs between defaulted and non-defaulted loans.



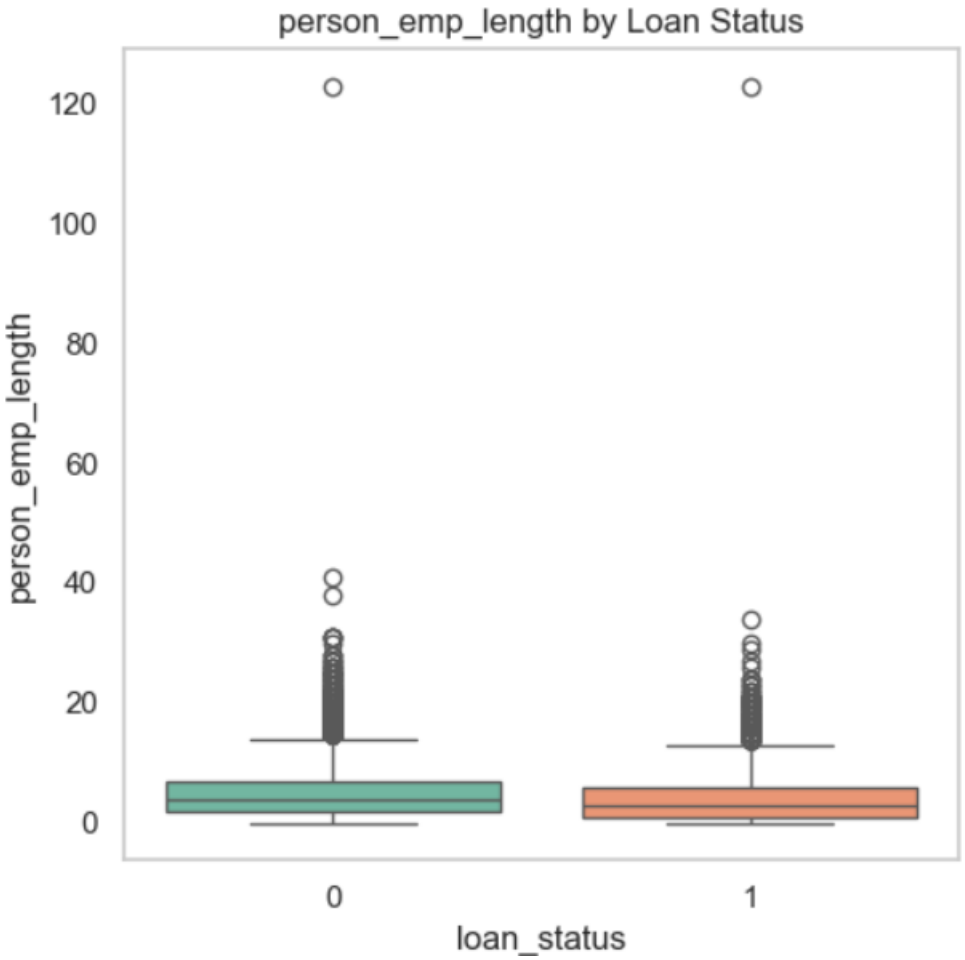
1. person_age by Loan Status

- Observation: The ages of borrowers do not seem to have a large impact on whether the loan defaults. Both defaulted and non-defaulted loans have similar age distributions, with the median age being around the same.
- Conclusion: Age might not be a strong predictor of loan defaults based on this dataset.



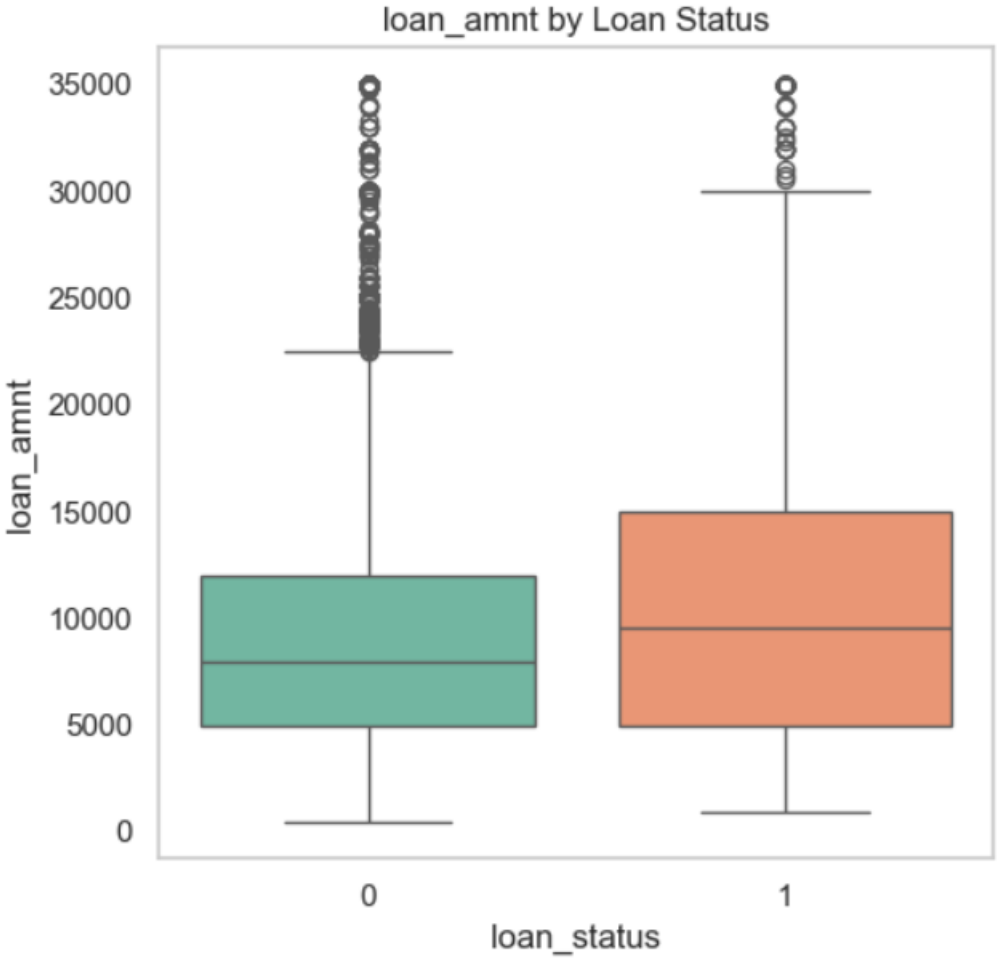
2. person_income by Loan Status

- Observation: There is a wide range of incomes, with many outliers in both defaulted and non-defaulted categories. However, both defaulted and non-defaulted loans seem to have a similar distribution in terms of income.
- Conclusion: Income might not be a strong differentiator between defaults and non-defaults, but further analysis could be done with normalization or transformation.



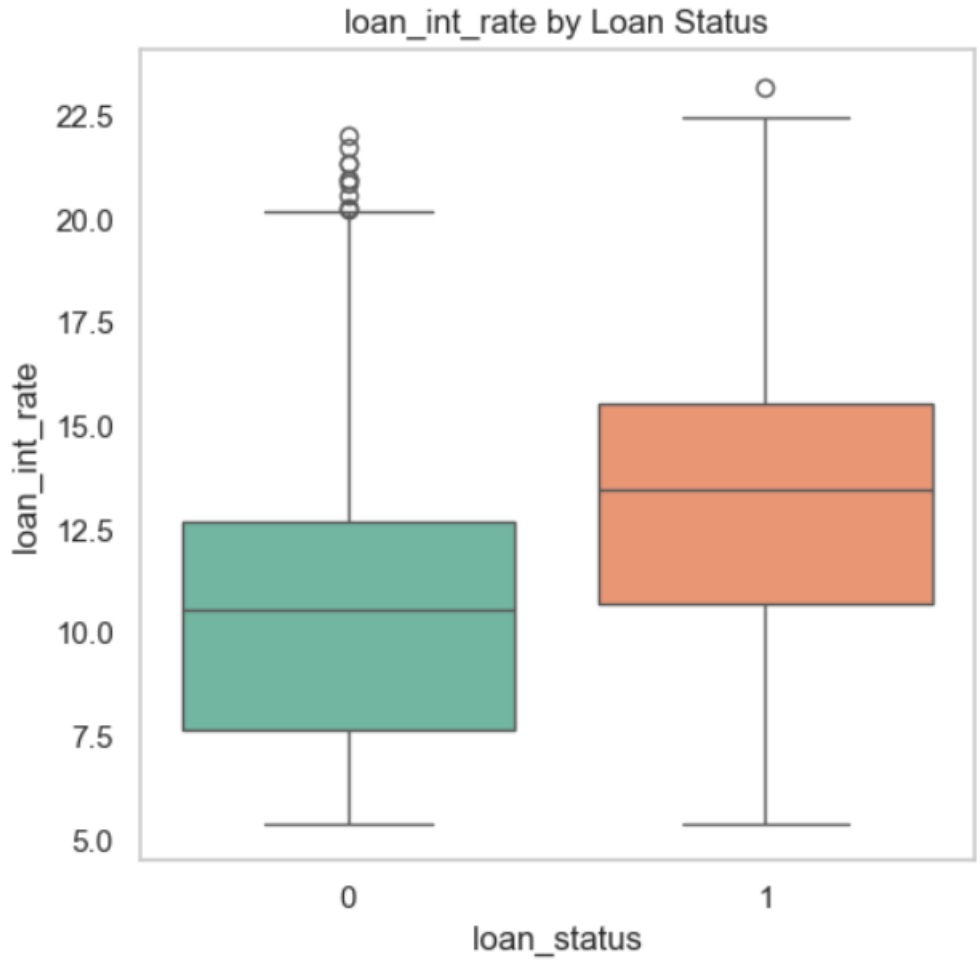
3. person_emp_length by Loan Status

- Observation: Employment length appears to have a similar distribution for both defaulted and non-defaulted loans, with the bulk of values being between 0 and 20 years. Some outliers extend much beyond this.
- Conclusion: Employment length may not be a significant factor in determining loan defaults.



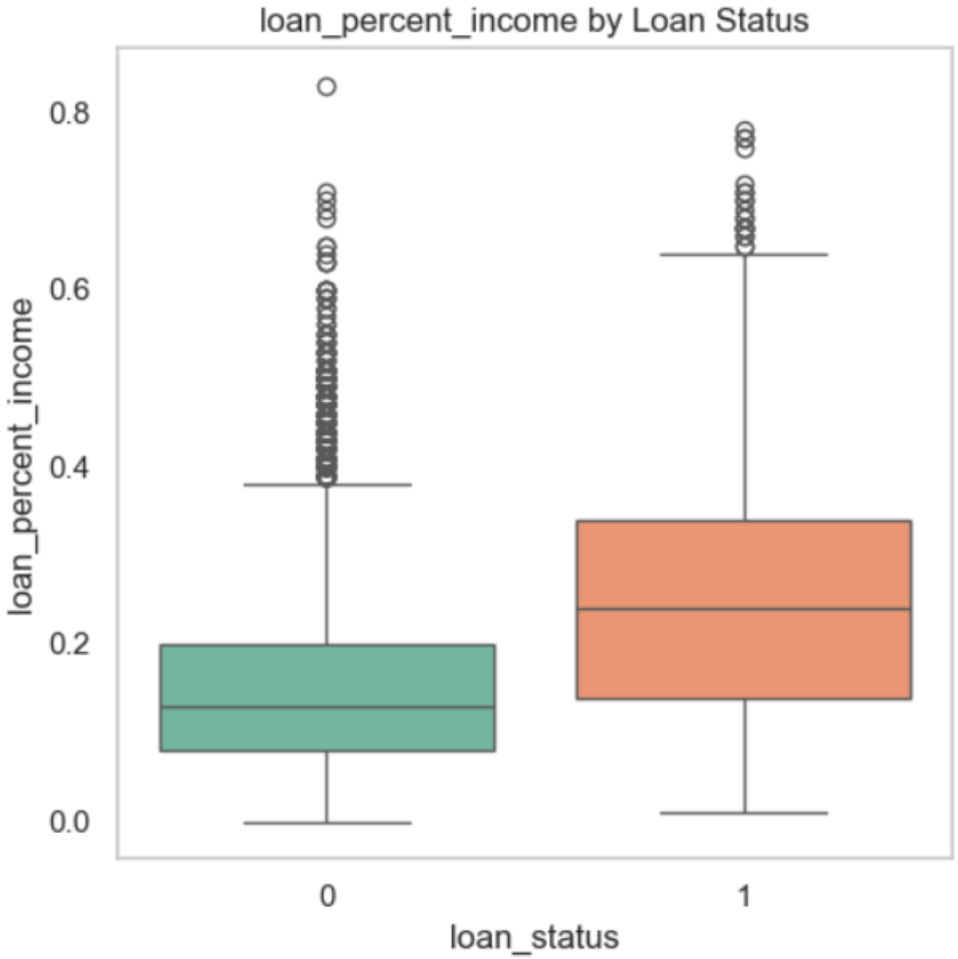
4. loan_amnt by Loan Status

- Observation: The loan amounts tend to be slightly higher for defaulted loans compared to non-defaulted loans, with the median loan amount for defaults being a bit higher.
- Conclusion: Larger loan amounts may be correlated with a higher likelihood of default, which could be useful in modeling.



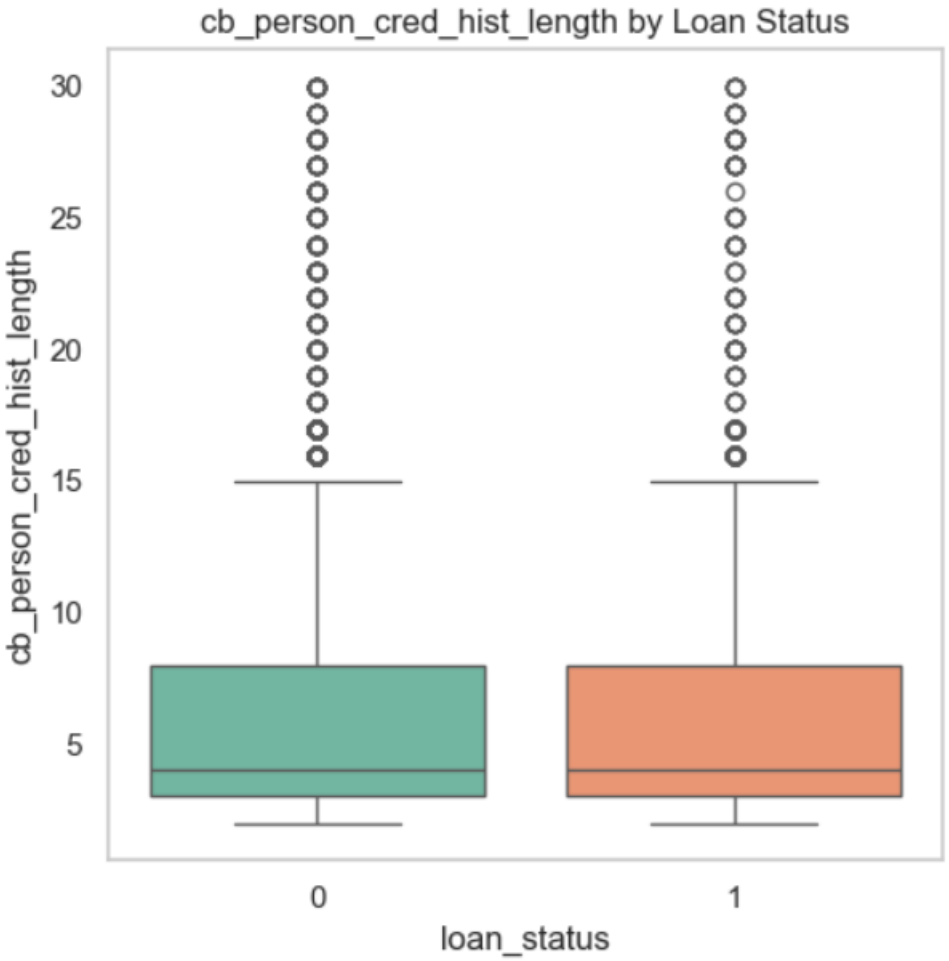
5. loan_int_rate by Loan Status

- Observation: Interest rates for defaulted loans are generally higher compared to non-defaulted loans. This is a significant insight because higher interest rates might be a risk factor for loan defaults.
- Conclusion: Loan interest rate appears to be a strong indicator of default risk. Borrowers with higher interest rates are more likely to default.



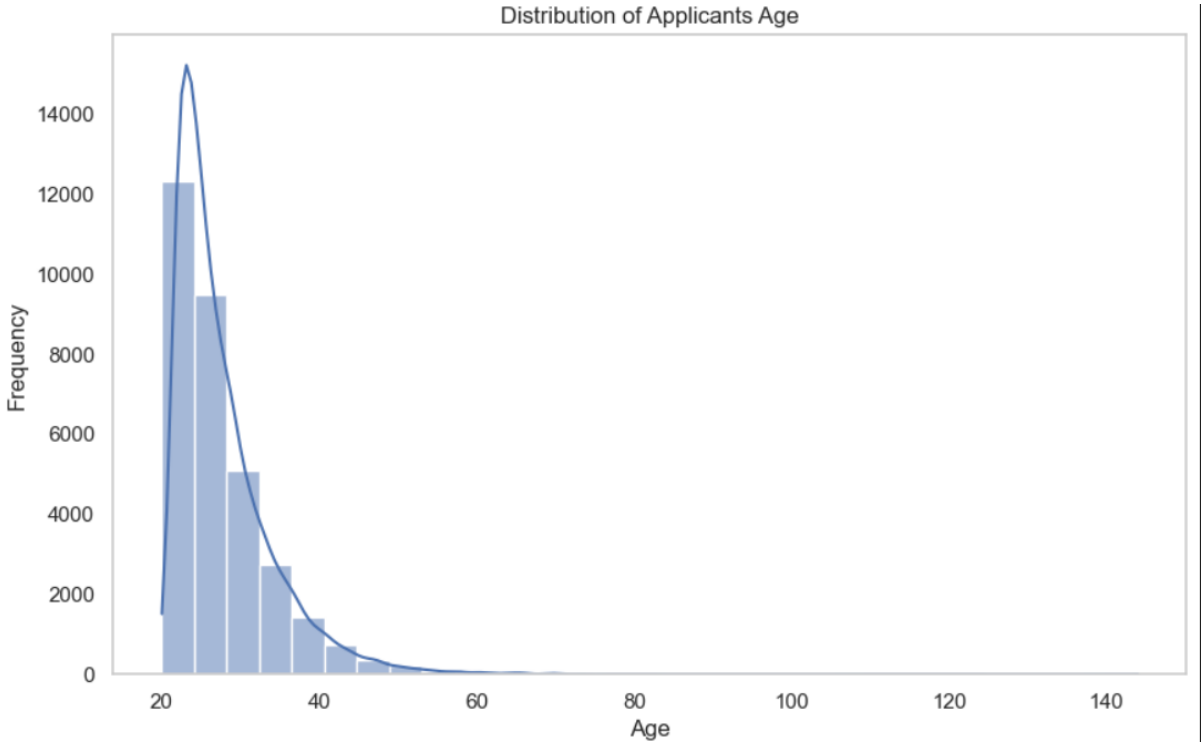
6. loan_percent_income by Loan Status

- Observation: Defaulted loans tend to have a higher percentage of income dedicated to repaying the loan. This is an important feature, as it shows that when a higher proportion of a borrower's income goes toward loan repayment, the likelihood of default increases.
- Conclusion: The percentage of income used for loan repayments is a critical factor in predicting loan defaults.



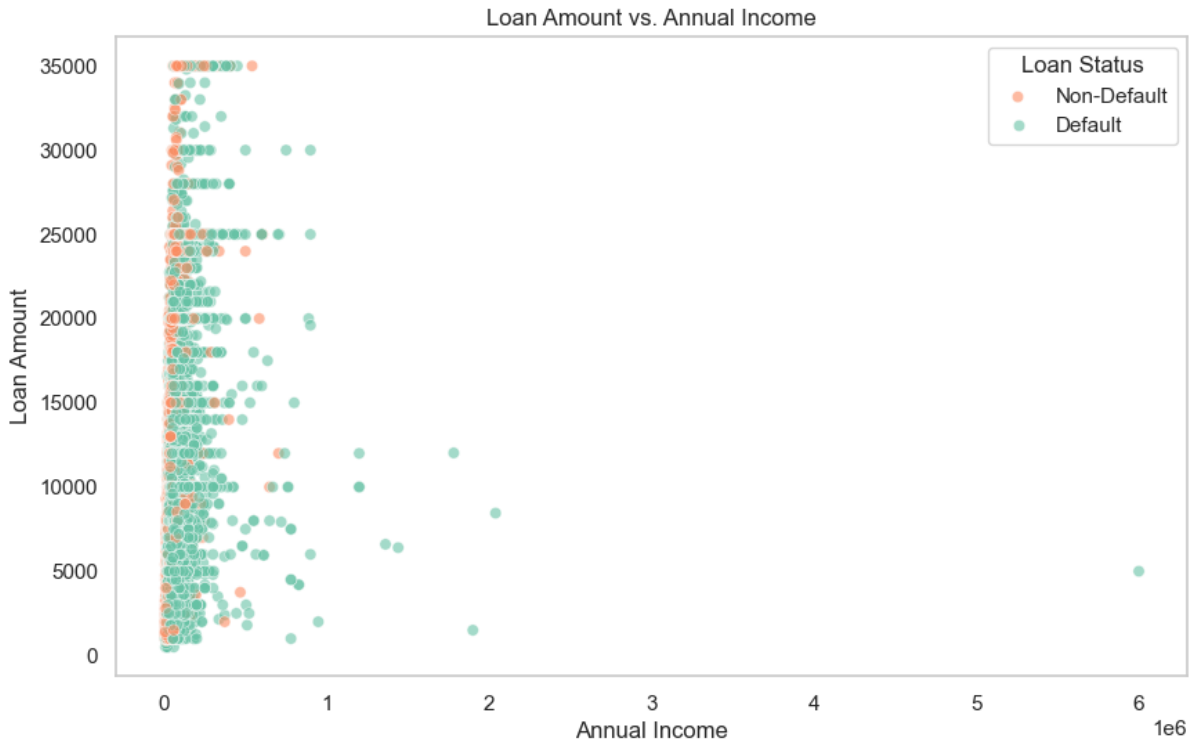
7. cb_person_cred_hist_length by Loan Status (Bottom Right Plot)

- Observation: The length of credit history is somewhat similar for defaulted and non-defaulted loans, but there may be some differences in the tails of the distribution.
- Conclusion: Credit history length alone may not be a strong predictor of loan defaults, but combined with other factors, it could be useful.



e) **Explanation of the Distribution of Applicants' Age Graph:**

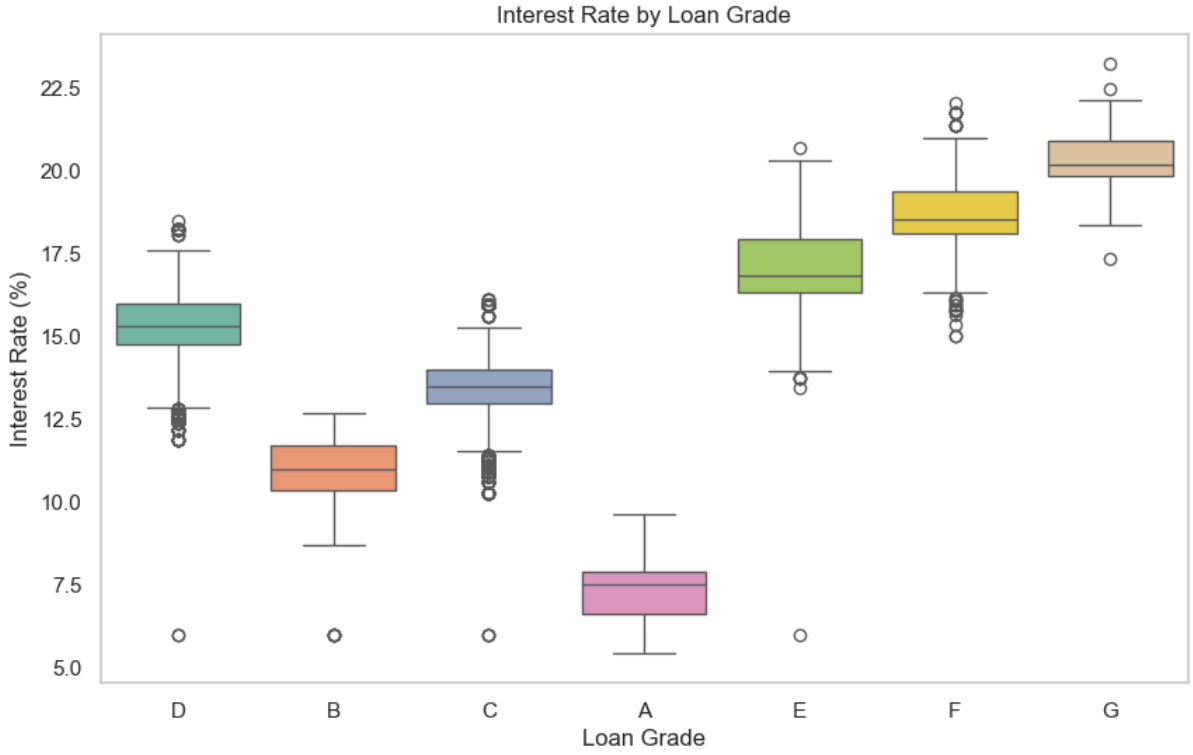
- **Graph Type**: This is a **histogram** with a **Kernel Density Estimate (KDE) curve**, which shows the distribution of ages for loan applicants.
- **Results**:
    - The **majority** of loan applicants are between the ages of **20 and 40**.
    - There is a **sharp decline** in the number of applicants as the age increases beyond 40, with very few applicants above 60.
    - There are a few outliers who are over 100 years old, which may indicate erroneous data entries or very rare cases.
- **Skewness**:
    - The distribution is **right-skewed** (positively skewed), meaning most data points are concentrated on the left side (younger applicants), with a long tail extending toward the right (older applicants).
    - This skewness suggests that younger individuals are more likely to apply for loans in this dataset.
- **Implications**:
    - **Outliers**: The very old applicants (over 100) might need further investigation or removal since they could distort the modeling results.
    - The age distribution may suggest that **younger people** are more likely to apply for loans, which could have implications for credit risk. Younger borrowers might be more likely to default, but further analysis is needed to verify that.



f) **Explanation of the Loan Amount vs. Annual Income Graph:**

- **Graph Type**: This is a **scatter plot** comparing the **loan amount** to **annual income**, with points color-coded by **loan status** (non-default and default).
- **Results**:
    - **Concentration of Data**: Most of the points are concentrated toward the lower end of the income scale (between 0 and 1 million in annual income). This indicates that most applicants have a relatively lower annual income.
    - **Loan Amount**: The loan amounts seem to range between **$0 and $35,000**, with no clear trend of higher loan amounts for higher incomes.
    - **Defaults vs. Non-Defaults**:
        - The points for **defaulted loans** (green) are scattered throughout, with no strong visible separation from **non-defaulted loans** (orange).
        - Both default and non-default loans occur frequently at lower income levels, but there are a few defaults even at higher income levels, though they are less common.
- **Key Observations**:
    - Applicants with higher incomes generally apply for similar loan amounts to those with lower incomes, indicating that income does not directly correlate with loan amount in this dataset.
    - However, the **concentration of defaults** seems to occur more at the lower income levels, suggesting that income might play some role in determining the likelihood of default, although this relationship is not clearly linear.

- **Outliers**:
    - There are a few **outliers** where applicants have very high incomes (over 2 million), and some of these have defaulted on their loans. These outliers might warrant further investigation, as they could affect the model's performance.
- **Implications**:
    - **Loan-to-income ratio** might be a more informative feature than looking at income and loan amount separately. Combining these features could improve the model's ability to predict defaults.
    - This scatter plot suggests that **lower income applicants** might be at a slightly higher risk of default, but further analysis would be needed to confirm this hypothesis.

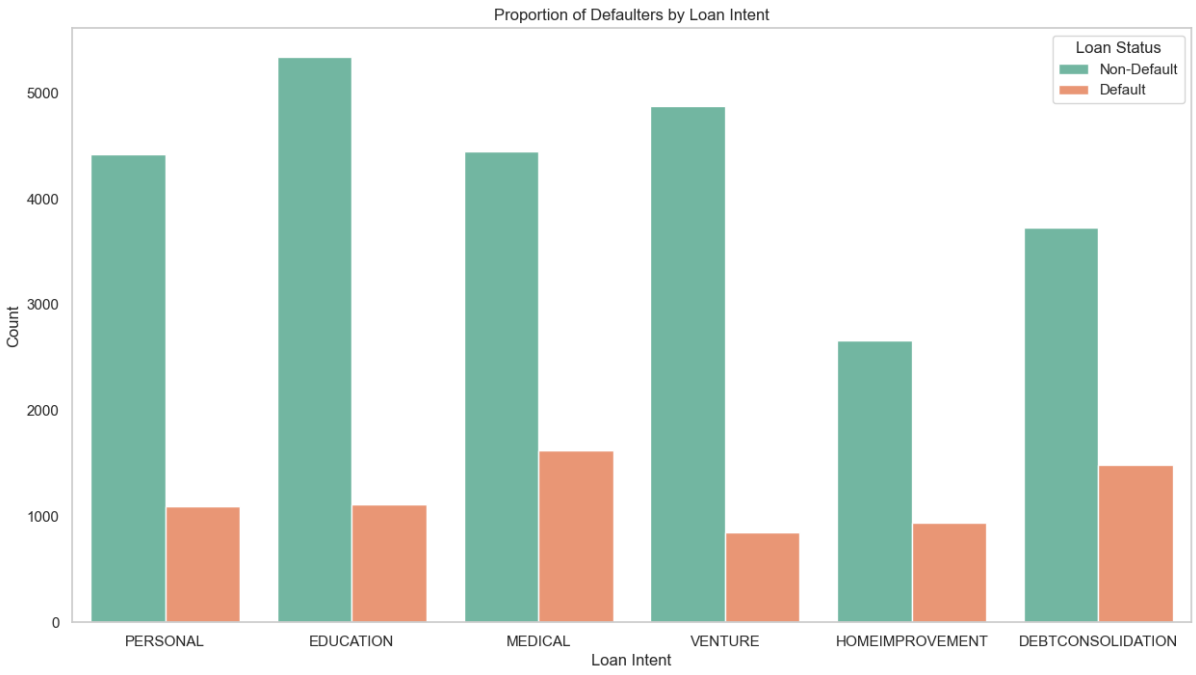


g) **Explanation of the Interest Rate by Loan Grade Box Plot:**

- **Graph Type**: This is a **box plot** comparing **interest rates** across different **loan grades** (A to G). The box plot shows the distribution of interest rates for each loan grade, with the central box representing the interquartile range (IQR), the line inside the box showing the median, and the "whiskers" extending to the minimum and maximum values (excluding outliers).
- **Results**:
    - Loan grades range from **A** (the highest quality or least risky loans) to **G** (the lowest quality or most risky loans).
    - **Grade A loans** have the **lowest interest rates**, with most rates between **6% and 9%**. This reflects the low risk associated with Grade A loans.

- As you move from **Grade A to Grade G**, the interest rates steadily **increase**, with **Grade G** loans having the highest interest rates, with medians around **20%**.
- **Grades E, F, and G** have particularly high interest rates, indicating they are considered the most risky loans, hence lenders charge higher interest to compensate for the increased risk.
- The spread of interest rates is relatively **narrow for Grade A** but **widens** for higher-risk loans, especially in **Grades D, E, F, and G**, which show more variability in interest rates.

- **Key Observations**:
  - The **positive correlation** between **loan grade** and **interest rate** is evident in this plot: higher-risk loans (Grades F and G) come with significantly higher interest rates, while lower-risk loans (Grade A) are offered at lower interest rates.
  - Outliers in the plot represent loans that were either offered at unusually high or low interest rates for their grade, which could be worth investigating.
- **Implications**:
  - Loan grade is a strong determinant of interest rate, and in credit risk modeling, this feature could serve as a powerful predictor of default probability.
  - The higher interest rates for low-grade loans (E, F, G) suggest that these loans carry a higher risk of default, and this should be reflected in any credit risk prediction model.

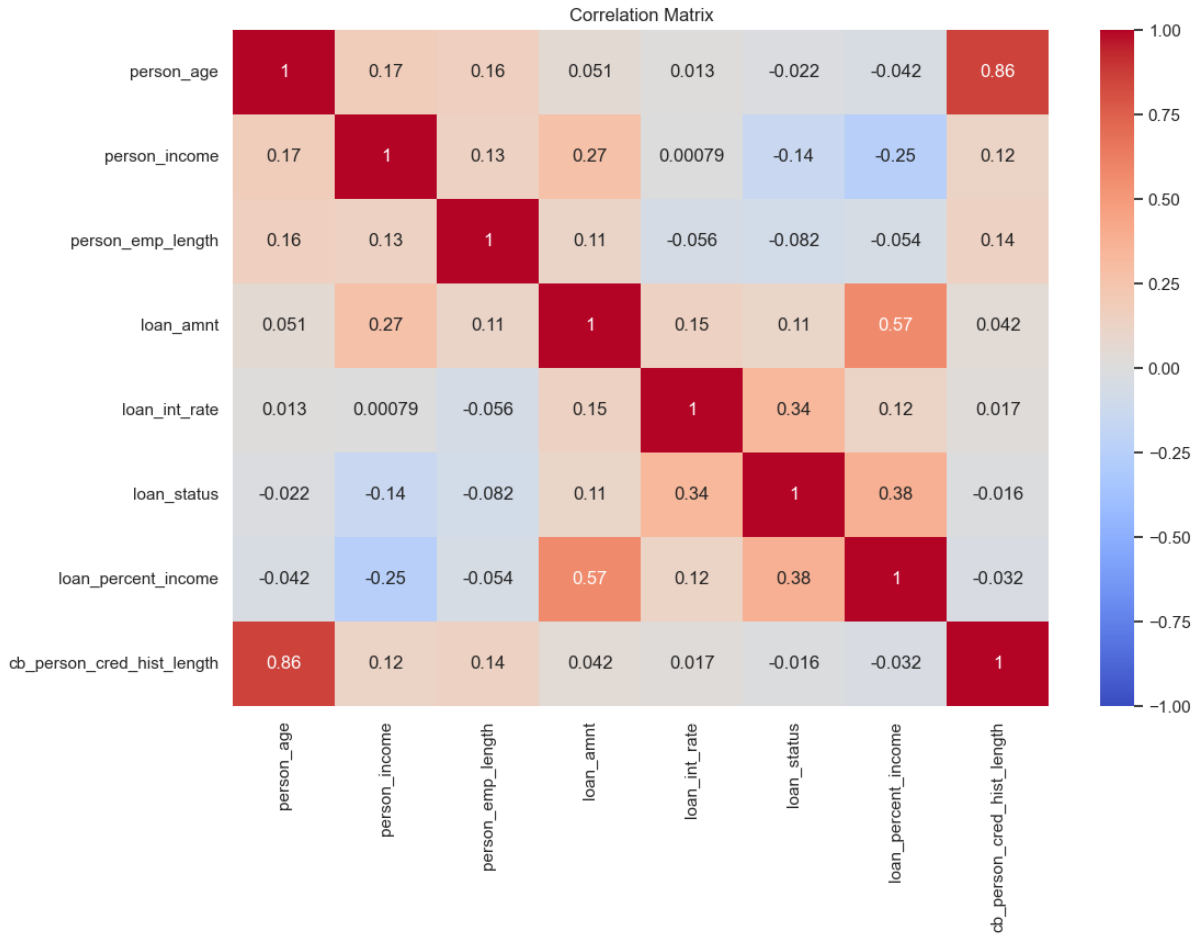


h) **Explanation of the Proportion of Defaulters by Loan Intent Bar Chart:**

- **Graph Type**: This is a **stacked bar chart** showing the proportion of **non-defaults** (green) and **defaults** (orange) for different types of loan intents. Each category of loan intent

(e.g., **Personal**, **Education**, **Medical**, etc.) shows the count of applicants who defaulted versus those who did not default.

- **Results**:
    - **Education** and **Venture** loans have the highest number of applicants, both for defaults and non-defaults.
    - **Medical** loans seem to have a relatively higher proportion of defaulters compared to other categories.
    - **Debt Consolidation** loans also have a notable proportion of defaults, which may suggest that individuals taking these loans are more financially unstable.
    - **Home Improvement** loans have the **smallest number** of total applicants, but the proportion of defaulters is still significant.
    - **Venture loans** seem to have fewer defaults compared to the total number of applicants, possibly indicating that these loans are less risky in this dataset.
- **Key Observations**:
    - **Medical Loans**: A large portion of applicants default on medical loans, which could indicate that individuals taking out medical loans are under significant financial pressure, increasing their risk of default.
    - **Debt Consolidation Loans**: The relatively high proportion of defaulters suggests that individuals using these loans are already struggling with debt, which leads to a higher likelihood of defaulting.
    - **Education and Personal Loans**: These categories have fewer defaults proportionally, meaning they may be lower risk compared to medical or debt consolidation loans.
- **Implications**:
    - **Loan Intent** is an important factor when determining the likelihood of default. Loans for medical purposes or debt consolidation are more prone to defaults, while personal and educational loans tend to have lower default rates.
    - This graph provides useful insights for credit risk models. Lenders could charge higher interest rates for high-risk loan intents (like medical or debt consolidation) to mitigate their risk.



i) Explanation of the **Correlation Matrix** Heatmap:

Key Correlations:

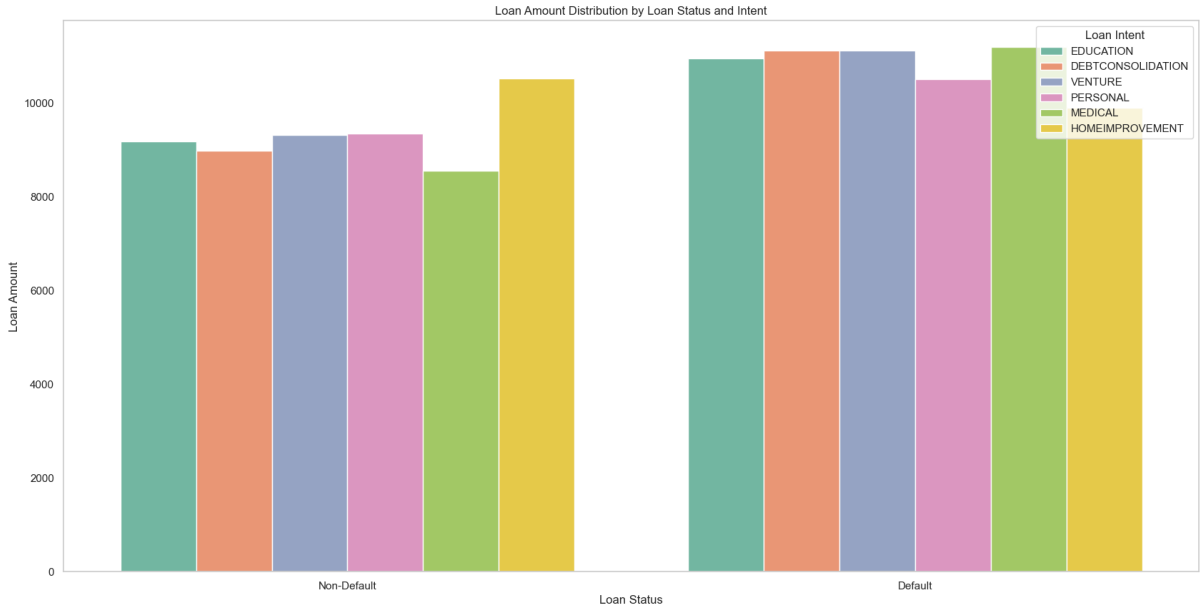
1. **person_age and cb_person_cred_hist_length (0.86)**:
    - There's a **strong positive correlation** between a person's age and their credit history length, which is expected. Older individuals tend to have longer credit histories.
2. **loan_percent_income and loan_amnt (0.57)**:
    - There's a **moderate positive correlation** between the percentage of a person's income used to repay the loan and the loan amount. This suggests that larger loans often take up a greater percentage of the borrower's income, which can be a risk factor for default.
3. **loan_status and loan_int_rate (0.34)**:
    - There's a **positive correlation** between **loan status** (default vs. non-default) and **loan interest rate**. This means that loans with higher interest rates are more likely to be associated with defaults, which aligns with expectations that riskier loans (with higher interest rates) have a higher chance of default.
4. **loan_amnt and loan_percent_income (0.57)**:

- Higher loan amounts are generally associated with a larger percentage of income being dedicated to loan payments. This might indicate that individuals taking out larger loans are under more financial stress.

5. **loan_int_rate and loan_amnt (0.15)**:
    - A weak positive correlation between loan amount and interest rate suggests that higher loan amounts tend to have slightly higher interest rates, though this relationship is not very strong.
6. **Weak or Negligible Correlations**:
    - There is **little to no correlation** between features like **income** and **loan status**, which might imply that income alone is not a strong predictor of default.
    - Other weak correlations include those between **employment length** and **loan status** or **loan amount**, meaning that employment length doesn't seem to play a significant role in loan default likelihood or loan size in this dataset.

Implications:

- **Credit History Length and Age**: Given the strong correlation between **age** and **credit history length**, these two variables may contain redundant information, and it may be worth considering using one over the other in the model.
- **Interest Rate and Loan Status**: Since there is a correlation between **interest rate** and **loan status**, this feature might play a key role in predicting defaults. Higher interest rates can be used as an indicator of higher risk.
- **Loan Amount and Income Percentage**: A higher percentage of income dedicated to loan payments correlates with larger loan amounts. Borrowers with larger loans relative to their income might be at higher risk of defaulting, making **loan_percent_income** a valuable feature for the model.



j) **Explanation of the Loan Amount Distribution by Loan Status and Intent Bar Chart:**

- **Graph Type**: This is a **grouped bar chart** showing the distribution of **loan amounts** for different loan intents, separated by **loan status** (non-default and default). The loan intents are color-coded (Education, Debt Consolidation, Venture, etc.), and the loan amount is displayed on the y-axis.

Results:

1. **Non-Default Loans (Left Group)**:
    - Across all loan intents, non-default loans have relatively similar loan amounts, with **Home Improvement** loans showing a slightly higher loan amount than the other categories.
    - **Debt Consolidation** loans and **Education** loans tend to have similar loan amounts for non-default loans.
2. **Default Loans (Right Group)**:
    - The distribution of loan amounts for defaulted loans is quite similar across different loan intents as well.
    - **Medical loans** and **Debt Consolidation loans** show higher amounts for defaulted loans compared to the other categories, which might indicate that borrowers who default on these loan types tend to have higher loan amounts.
    - **Home Improvement loans** also have a noticeable amount for defaults, but it is relatively lower compared to other categories like **Education** and **Debt Consolidation**.

Key Observations:

- **Loan Amount Consistency**: There doesn't seem to be a dramatic difference in loan amounts between non-default and default loans within each loan intent. The amounts are fairly consistent across both defaulted and non-defaulted loans.

- **Debt Consolidation** and **Medical Loans**: Both categories tend to show higher amounts for defaulted loans compared to other intents, which may suggest these loan types carry a higher financial risk, especially when borrowers take larger loans for these purposes.
- **Home Improvement Loans**: While **Home Improvement** loans appear in both defaulted and non-defaulted categories, the loan amounts for defaulted loans are lower compared to non-defaulted ones.

Implications:

- **Loan intent** plays a crucial role in understanding default risk, and certain loan types (e.g., Medical, Debt Consolidation) tend to have higher loan amounts associated with defaults, suggesting they could be more risky.
- This analysis provides useful insights for building a credit risk model, indicating that **loan amount** combined with **loan intent** can be an important feature in predicting defaults.

In summary, this chart provides a clear comparison of loan amounts across different intents and shows that, while loan amounts don't vary significantly between defaults and non-defaults, certain loan types like **Medical** and **Debt Consolidation** may involve higher risks.

	Model	Accuracy	Precision	Recall	F1-score	Support
0	GaussianNB	0.797299	0.815497	0.797299	0.804261	6517.000000
1	DecisionTreeClassifier	0.885684	0.887949	0.885684	0.886678	6517.000000
2	KNeighborsClassifier	0.877551	0.872523	0.877551	0.870176	6517.000000
3	RandomForestClassifier	0.929569	0.931929	0.929569	0.925496	6517.000000
4	LogisticRegression	0.837195	0.825224	0.837195	0.821997	6517.000000
5	AdaBoostClassifier	0.885990	0.881568	0.885990	0.881500	6517.000000
6	XGBClassifier	0.934632	0.935716	0.934632	0.931562	6517.000000
7	LGBMClassifier	0.933712	0.936186	0.933712	0.930036	6517.000000
8	NeuralNetwork	0.918060	0.922726	0.918060	0.911752	6517.000000

k) **Explanation of the Model Performance Table:**

This table summarizes the performance of several machine learning models based on key metrics: **Accuracy**, **Precision**, **Recall**, and **F1-score**.

- The **XGBoost** model is the best-performing model based on **all metrics**. It has the highest accuracy, precision, recall, and F1-score. This indicates that XGBoost is the most balanced model in terms of predicting both defaults and non-defaults.
- LightGBM also performs exceptionally well, coming in just behind XGBoost. It has the best **precision**, meaning it is the most accurate in predicting positive classes (defaults), though its recall is slightly lower than XGBoost.
- The neural network performs well but falls short compared to XGBoost and LightGBM. It has balanced performance, but its accuracy, precision, and recall are slightly lower than the top two models.
- **Tree-based models** like Random Forest and Decision Tree also perform well, but XGBoost and LightGBM outperform them in every metric.

Conclusion

In this project, we applied a variety of machine learning techniques to tackle the problem of predicting loan defaults, a crucial task in the financial industry. By using a dataset containing borrower characteristics and loan details, we explored several models, including decision trees, logistic regression, random forests, and gradient boosting models like XGBoost and LightGBM.

Through comprehensive data preprocessing, feature engineering, and model evaluation, we were able to compare different algorithms. The **XGBoost model** emerged as the best-performing algorithm, achieving the highest accuracy, precision, and F1-score. This indicates its effectiveness in correctly classifying both defaulting and non-defaulting borrowers, making it a strong candidate for practical use in credit risk modeling.

Key Insights:

- **Interest rate** and **loan-to-income ratio** were identified as critical factors in predicting defaults. Higher interest rates and larger loan amounts relative to income were strongly correlated with default risk.
- Certain loan intents, such as **debt consolidation** and **medical loans**, exhibited higher default rates, suggesting that they are riskier categories.
- **Imbalanced data** was a significant challenge, as there were far more non-defaults than defaults. This was addressed through careful metric selection and by leveraging models that can handle imbalanced data.

Future Considerations:

- **Hyperparameter tuning**: While XGBoost performed exceptionally well, further hyperparameter tuning could potentially improve its performance even more.
- **Feature engineering**: Additional features, such as borrower behavior over time or macroeconomic variables, could further enhance the model's predictive power.
- **Deployment**: With the strong performance of the XGBoost model, this approach is ready to be deployed in a real-world environment, allowing financial institutions to minimize risks and optimize lending strategies.

In conclusion, the results from this project underscore the power of machine learning in predicting credit risk. By using the best-performing model and insights gained from feature analysis, financial institutions can significantly reduce default risks while maintaining a balanced portfolio.