**Date:12/04/2015**

**DATA SCIENCE & DATA ANALYTICS|AI|ML**

**PROJECT 1 REPORT**

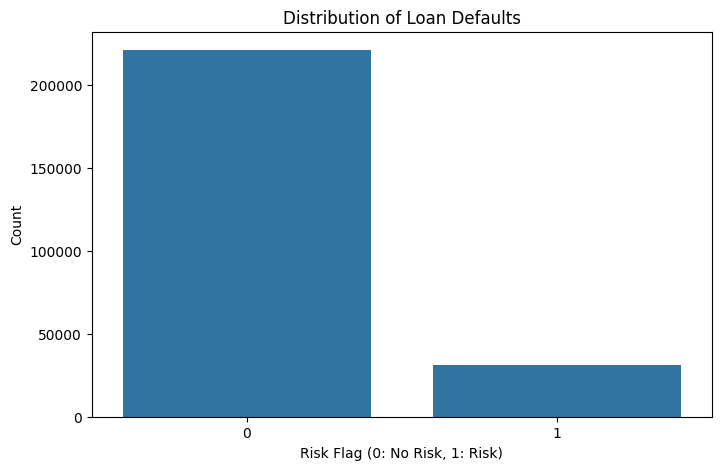
Alpha dreamers Banking consortium is a financial institution that provides business loans, savings accounts, and checking accounts for individuals and businesses. The bank has been experiencing problems with regards to a large number of clients failing to pay back the loans.This report outlines the findings of an analysis aimed at predicting loan defaults for Alpha Dreamers Banking Consortium. The objective is to develop a predictive model that can assist the loans department in identifying potential defaulters and mitigating associated risks.

The following are the research questions that were answered through the exploratory data analysis:

1. What is the overall distribution of loan defaults (Risk\_Flag)?
2. How does marital status (Married/Single) relate to loan default risk?
3. Does homeownership status (House\_Ownership) impact loan default rates?
4. Is there a relationship between car ownership (Car\_Ownership) and loan default risk?
5. How are income and age related to loan defaults?
6. What is the income distribution for defaulting vs. non-defaulting borrowers?
7. How does age distribution vary between defaulting and non-defaulting borrowers?
8. Do certain professions have significantly higher default rates?
9. Are there significant regional variations in loan default rates?
10. What are the key relationships between features in the dataset?

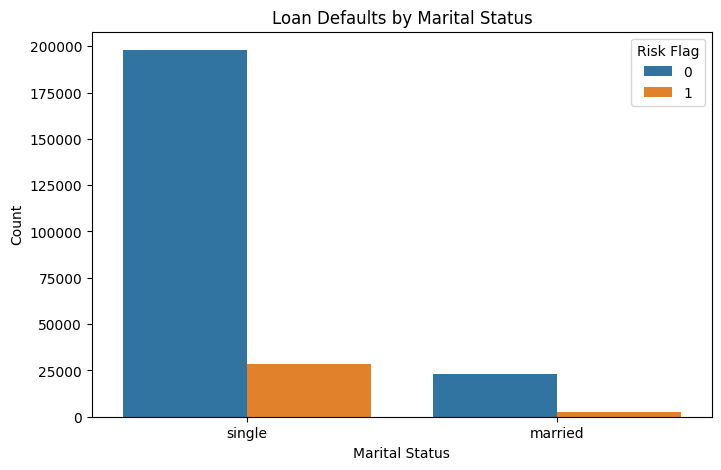
**Exploratory Data Analysis (EDA)**

The EDA phase involved a comprehensive analysis of the provided loan data. Key visualizations and their insights are summarized below:



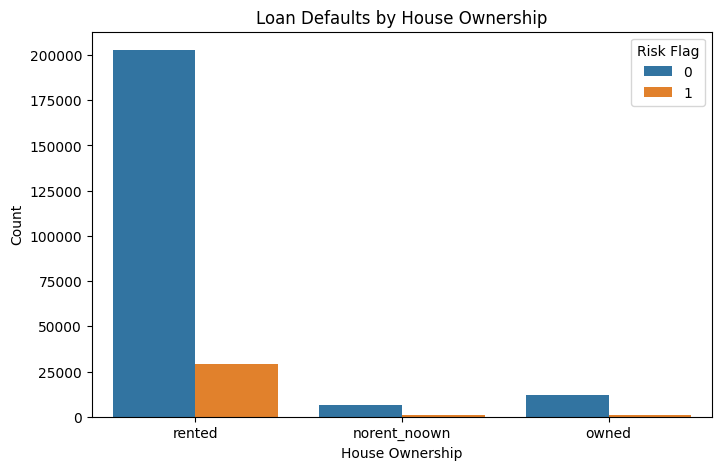
**Figure 1: Histogram of risk flag**

The analysis revealed a significant class imbalance, with a considerably higher number of loans with "No Risk" (Risk\_Flag = 0) compared to "Risk" (Risk\_Flag = 1). This imbalance will need to be addressed during model development.



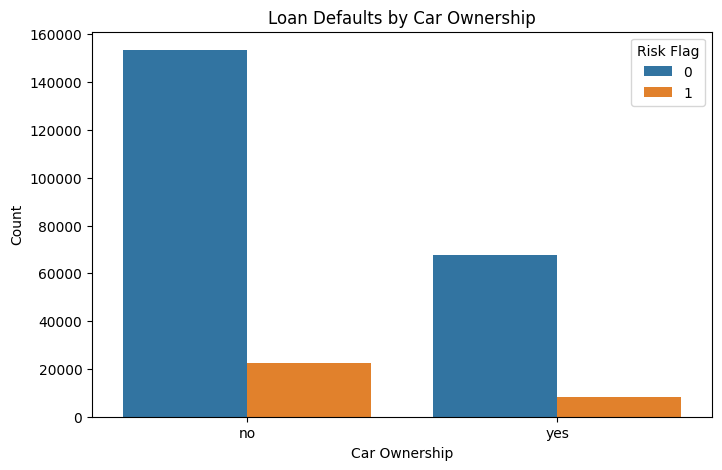
**Figure 2: Marital Status (Married/Single) and Default Risk**

The analysis showed that single individuals had a slightly lower proportion of defaults compared to married individuals.However, the overall difference in default rates between the two groups was not substantial.

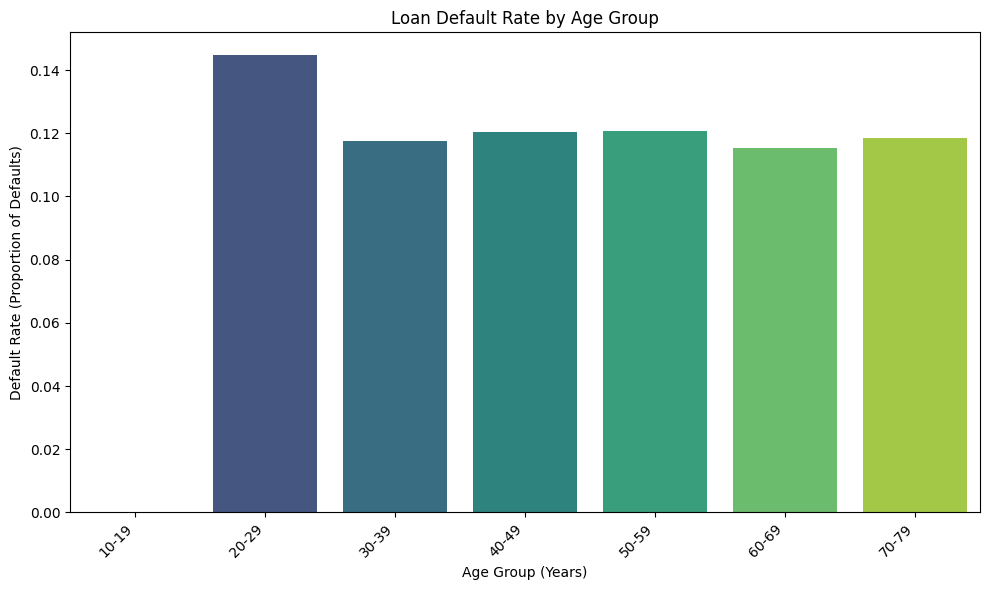


**Figure 3: House Ownership and Default Risk**

Borrowers who rented their homes exhibited a slightly higher proportion of defaults compared to homeowners or those who neither owned nor rented.However, the overall number of defaults within the "Rented" category was still relatively low.



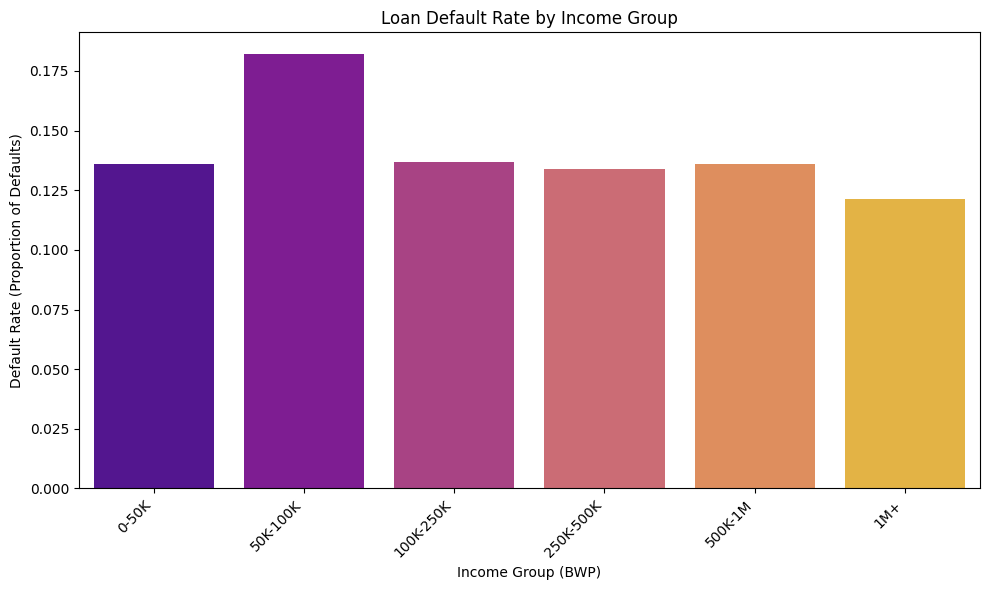
**Figure 4: Car Ownership (Car\_Ownership) and Default Risk**

Borrowers who did not own cars showed a slightly higher proportion of defaults compared to car owners.However, the overall default rate within both groups was relatively low.

**Figure 5: Relationship between age and risk flag**

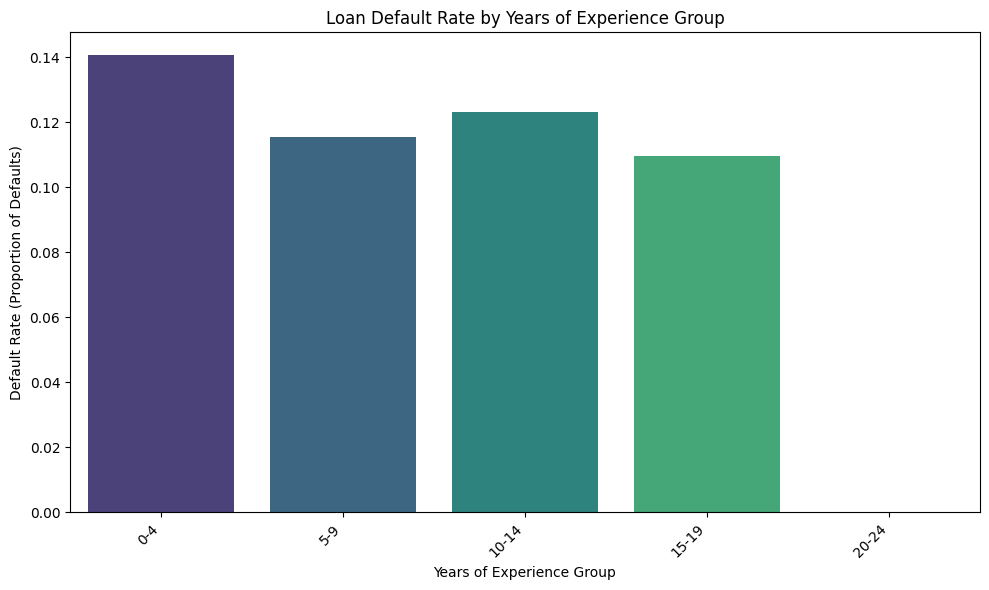
The 20-29 age group has the highest number of total loans (43,657) compared to other age groups. The 20-29 age group also has the highest number of defaults (6,318).The 70-79 age group shows a lower number of defaults (4,447), likely due to a lower number of total loans in this group compared to the 20-69 age ranges

The 20-29 age group has the highest proportion default loans rates which means that this age range is most likely to default on loans. The age group for 60-69 years has the lowest default rate.All the age groups except (and also older than 20-29yrs)20-29 the age groups have a similar default range between 0.10 and 0.12.



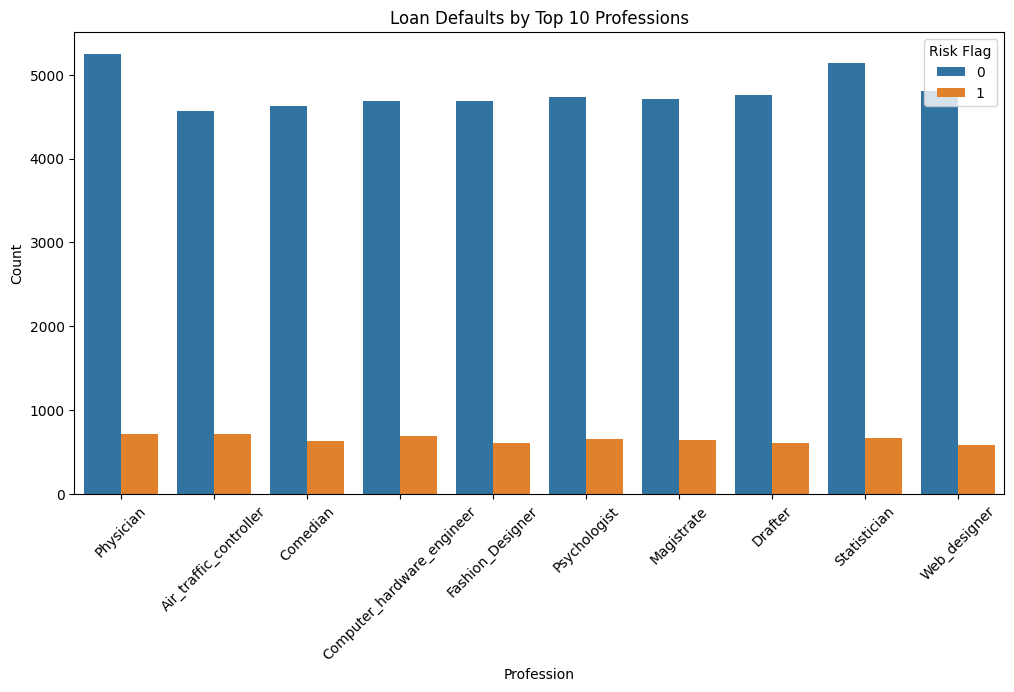
**Figure 6: Relation between income and risk flag**

The people who have an income range of 50k-100K have the highest default rate which means that people in this income group are most likely to default on a loan. The 0-50K, 100K -250K, 250K-500K and 500K-1M all fall within a default range of 0.125 to 0.150. The people with an income of more than 1M have the lowest default loan rate which means they are less likely to default on a loan compared to the other groups.



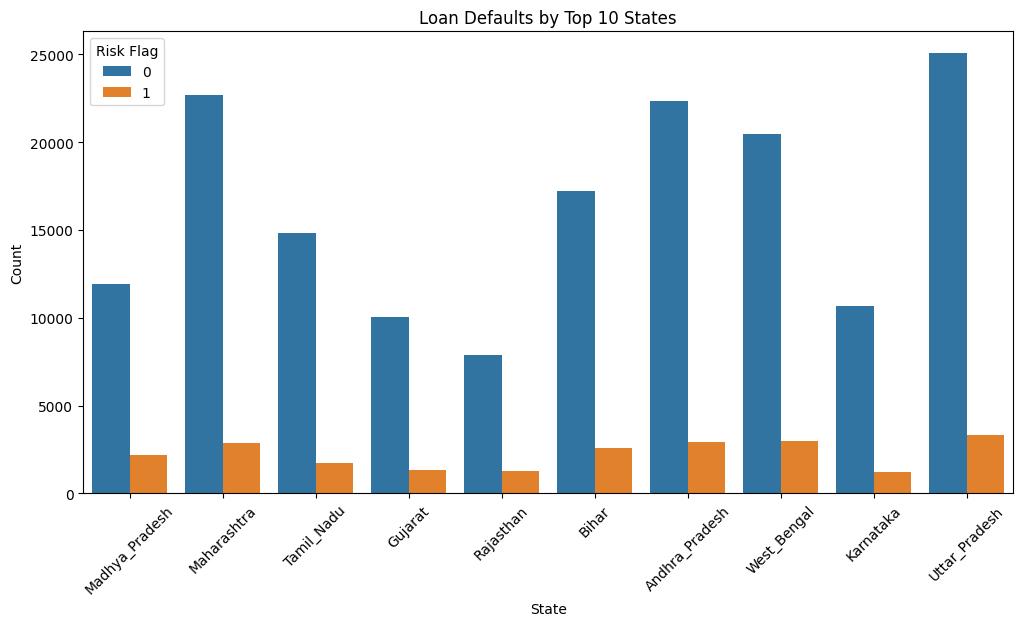
**Figure 7: Relationship between experience in years and risk flag**

The people who have experience of 0-4 years have the highest default rate which means these people are most likely to default on their loan. Those with 10-14 years of experience have the 2nd highest default rate.The 15-19 years had the lowest probability of defaulting the loan.



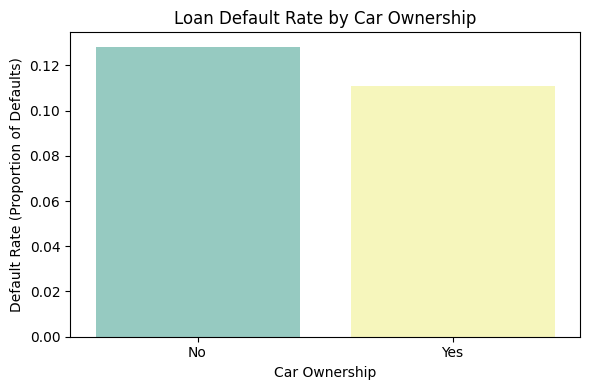
**Figure 8: Relationship between profession and default risk**

Analysis of the top 10 professions revealed slight variations in default rates across different professions.Physicians, Web Designers, and Statisticians exhibited slightly lower default rates.However, the differences in default rates across professions were relatively minor.



**Figure 9: Relationship between state and default risk**

Analysis of the top 10 states revealed some regional variations in default rates. Borrowers in Uttar Pradesh exhibited the lowest default rates among the top 10 states.Overall, most states showed a low proportion of defaults.



**Figure 10: Relationship between car ownership and default loan rates**

Those who do not own cars have the highest risk of defaulting the loan. The default rate is 0.12 for those who own cars, whereas, the default rate for those who do not own cars is 0.10.

The findings from the descriptive statistics that were done show that Income has a wide range, with a few very high earners. The mean and median are relatively close, indicating a somewhat symmetric distribution, but the large standard deviation suggests significant variability.The experience distribution is relatively symmetric, with a mean and median around 10 years.This is a binary variable, where 0 represents no risk and 1 represents high risk. The mean of 0.12 indicates that the dataset is imbalanced, with significantly more loans having no risk. The current house years variable has a very small spread.

**Machine learning model**

A logistic regression model was chosen due to its suitability for binary classification problems, such as predicting loan defaults.The data was preprocessed, including dropping irrelevant columns ('ID', 'CITY'), one-hot encoding of categorical variables (Married/Single, House\_Ownership, Car\_Ownership, Profession, STATE), splitting data into training and testing sets (80/20) and standardizing features using StandardScaler.A Logistic Regression model was instantiated and trained on the scaled training data.

The model's performance was evaluated using the following metrics:

* Accuracy: 0.88
* Precision: 0.77 (weighted avg)
* Recall: 0.88 (weighted avg)
* F1-score: 0.82 (weighted avg)
* AUC-ROC: 0.58

A precision of 0.77 suggests that 77% of the instances predicted as defaults by the model were actually defaults. A recall of 0.88 indicates that the model correctly identified 88% of the actual defaulters.An F1-score of 0.82 suggests a reasonable balance between precision and recall in the model's predictions.An AUC-ROC of 0.58 indicates that the model's ability to distinguish between defaulting and non-defaulting loans is **relatively poor**.

**Inferential statistics**

Statistical analysis test was performed on each variable to explore if any of these factors affect the probability of someone defaulting on the loan. The factors that are statistically significant on the probability of the patient defaulting on the loan. The following variables have been seen as statistically significant: some professions, some states,house ownership, age, married/single status, experience, current job years and car ownership.

**Summary**

In conclusion some factors like marital status, house ownership, and certain professions showed slight associations with default risk.Income and age did not appear to be strong predictors of loan defaults in this analysis.The model exhibits relatively good accuracy and recall.However, the low precision and AUC-ROC scores indicate that the model may be prone to making false positive predictions.