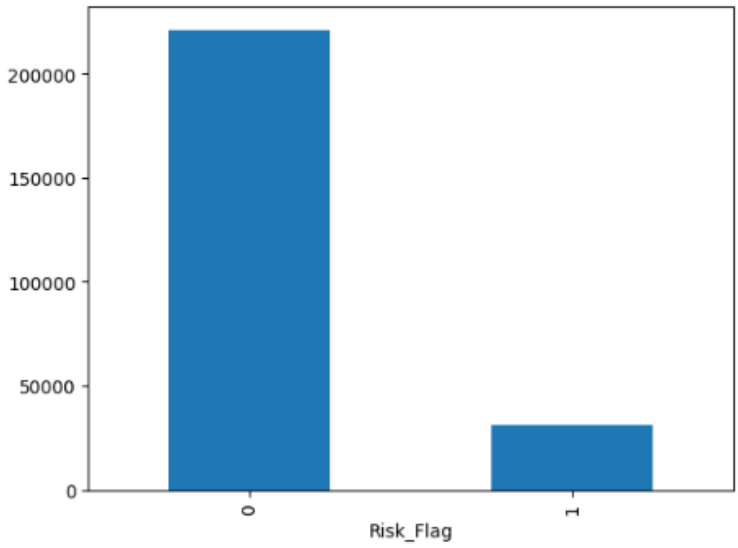
Data visualizations:

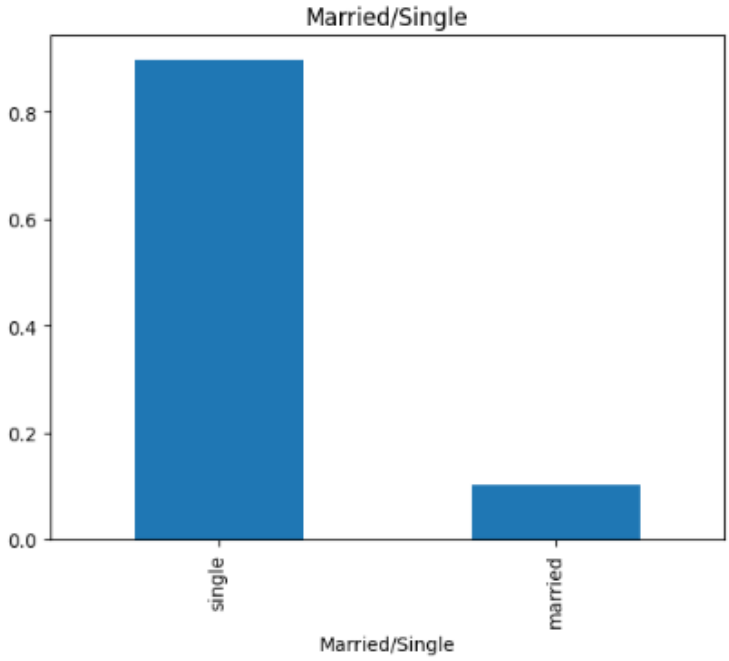
First I create some visualizations based on some features of the dataset to the dataset well.

Below are some visualizations to show how the data is distributed.

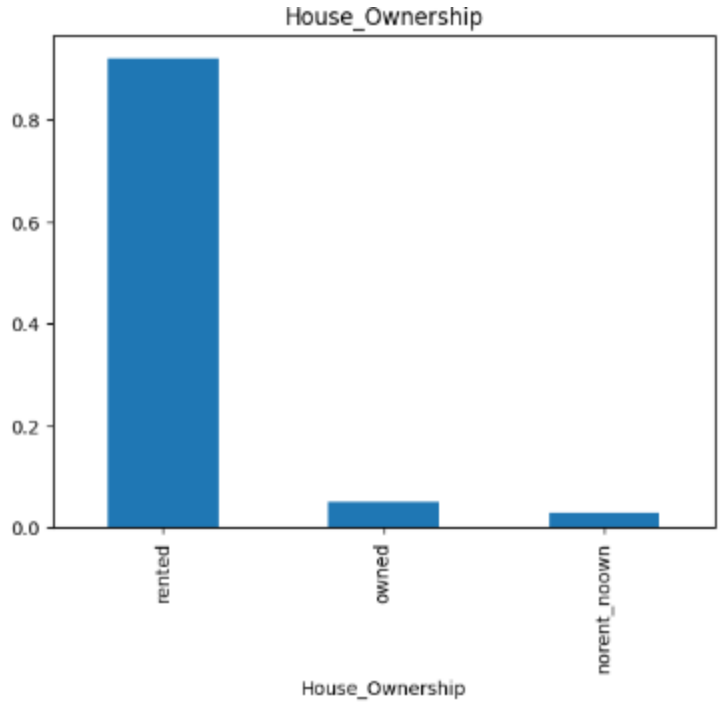
From the below picture we can see that most of the values of Risk_Flag is 0.



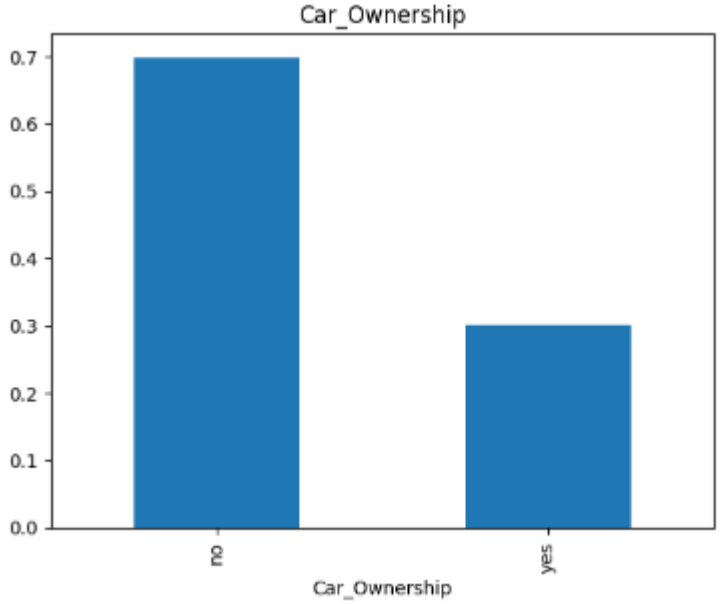
Plot to see distribution of married and single people in the dataset.



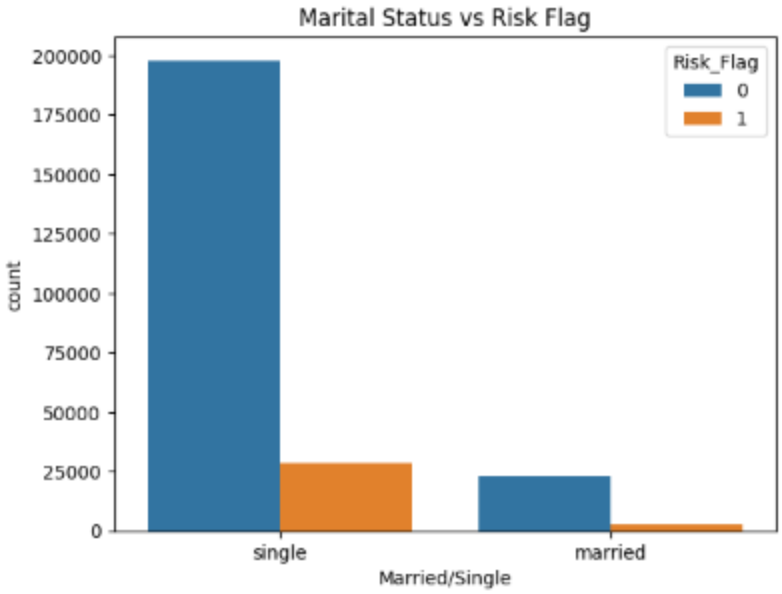
From the below pictue we can see that most of the people has rented house



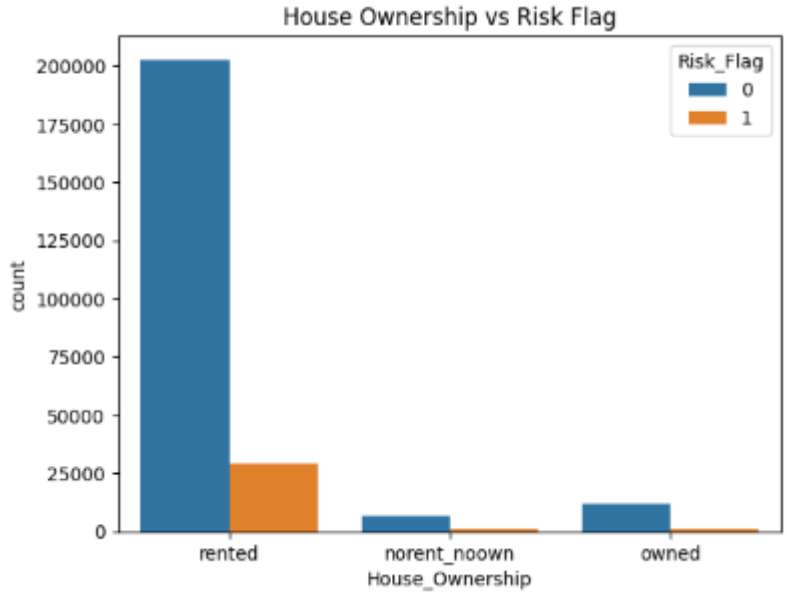
Also most of the people don’t have car.



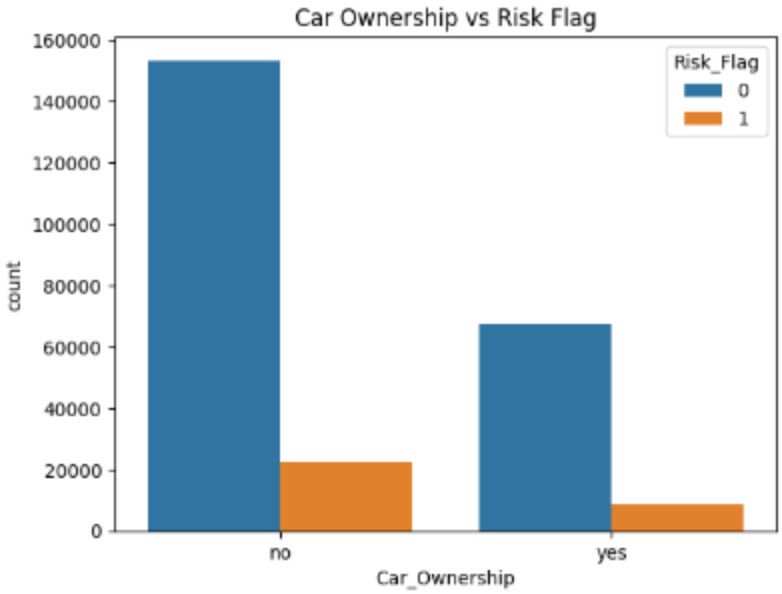
Below is the visualization of Marital status vs risk flag.



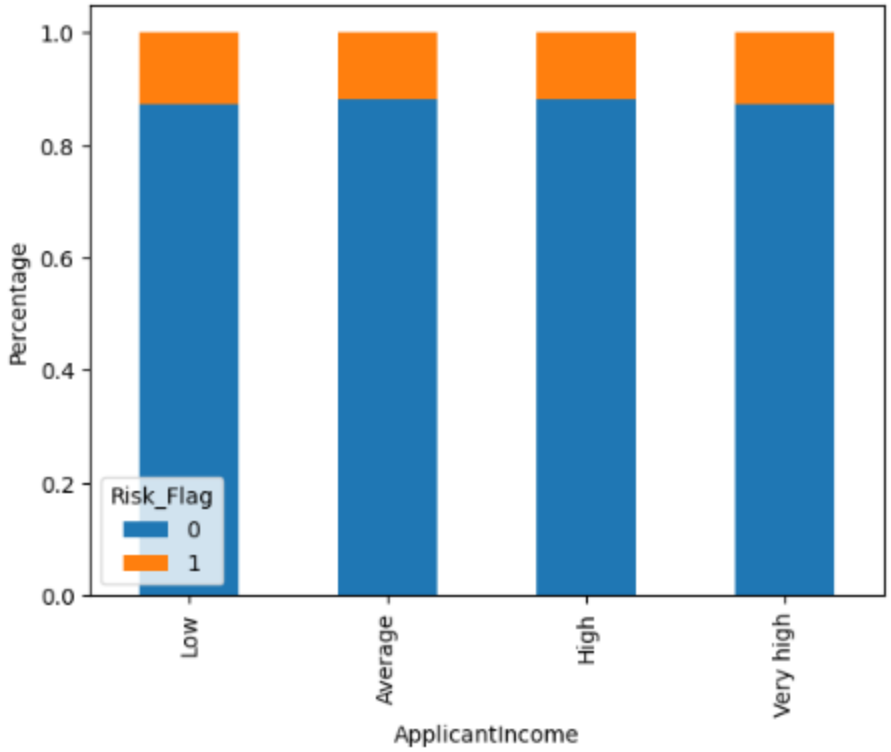
Below is the visualization of House ownership vs risk flag



From the below diagram of car ownership vs Risk flag we can see that proportion of having a car and Risk Flag is high



From the Below visualization we can see proportion of risk flag=0 and 1 are almost same for all the category of applicant income.



Model Performance Analysis

Logistic Regression:

- **Accuracy:** 0.88
- **ROC AUC:** 0.56
- **Cross-validation AUC Scores:** [0.555, 0.557, 0.552, 0.559, 0.558]

With an accuracy of 0.88, the Logistic Regression model accurately identifies a significant amount of the data. Its ROC AUC score (0.56) is comparatively low, indicating that it has little capacity to distinguish between clients who pose a high risk and those who do not. The classification report, which demonstrates flawless memory for class '0' but no recall for class '1', lends more credence to this. This suggests a bias in the model's prediction of low-risk (class '0') clients.

Decision Tree:

- **Accuracy:** 0.86
- **ROC AUC:** 0.69
- **Cross-validation AUC Scores:** [0.661, 0.686, 0.691, 0.675, 0.660]

Compared to Logistic Regression, the Decision Tree model performs much better in ROC AUC (0.69) but somewhat lower in accuracy (0.86). Although a sizable percentage of class '1' clients are still misclassified, the classification report shows more recall and precision for this class as compared to Logistic Regression, indicating that the Decision Tree model is more balanced and capable of identifying high-risk clients.

Random Forest:

- **Accuracy:** 0.91
- **ROC AUC:** 0.92
- **Cross-validation AUC Scores:** [0.921, 0.920, 0.919, 0.920, 0.919]

The Random Forest model performs better than all other models in terms of accuracy (0.91) and ROC AUC (0.92), demonstrating good predictive performance and the capacity to discriminate between clients who pose a significant risk and those who do not. The categorization report exhibits some bias towards class '0', but it also demonstrates a reasonable balance between precision and recall for both classes. It can be seen from the confusion matrix that the model misclassifies comparatively less frequently.

Gradient Boosting:

- **Accuracy:** 0.88
- **ROC AUC:** 0.73
- **Cross-validation AUC Scores:** [0.729, 0.727, 0.729, 0.727, 0.727]

While the Gradient Boosting model performs better in terms of ROC AUC (0.73), it has accuracy comparable to that of Logistic Regression (0.88). Although there has been progress, the categorization report reveals that it has low recall for class '1', meaning that although it can recognise low-risk clients with ease, it finds it difficult to recognise high-risk clients.

Main Deciding Factors Associated with Risk

The feature importance from the Random Forest model highlights the main factors influencing the risk prediction:

1. **Id:** 0.319
2. **CITY:** 0.128
3. **Age:** 0.114
4. **Profession:** 0.112
5. **STATE:** 0.079
6. **Experience:** 0.069
7. **CURRENT_JOB_YRS:** 0.060
8. **CURRENT_HOUSE_YRS:** 0.050
9. **Income:** 0.041
10. **Car_Ownership:** 0.012
11. **Married/Single:** 0.008
12. **House_Ownership:** 0.008

Key Insights:

• id: This characteristic, which is usually a unique identity, unexpectedly demonstrates a high importance, suggesting that there may have been data leakage or correlation with other significant characteristics.
• CITY and STATE: Risk is significantly influenced by geographic location, maybe as a result of local market dynamics and economic conditions.
• Age and Experience: These two characteristics are significant since they suggest that people with greater age and experience may be viewed as carrying a different level of danger.
• Profession: Because different vocations have distinct default probabilities and income stability, professions have a substantial impact on risk assessment.
• CURRENT_JOB_YRS and CURRENT_HOUSE_YRS: Risk assessment is influenced by the stability of employment and housing conditions, with longer periods typically suggesting lower risk.
• Income: Although significant, income is not the only determining factor, indicating that the model takes a comprehensive approach to a client's profile.