Name: Heshika Pokala

Gmail: heshikapokala@gmail.com

Data Visualizations

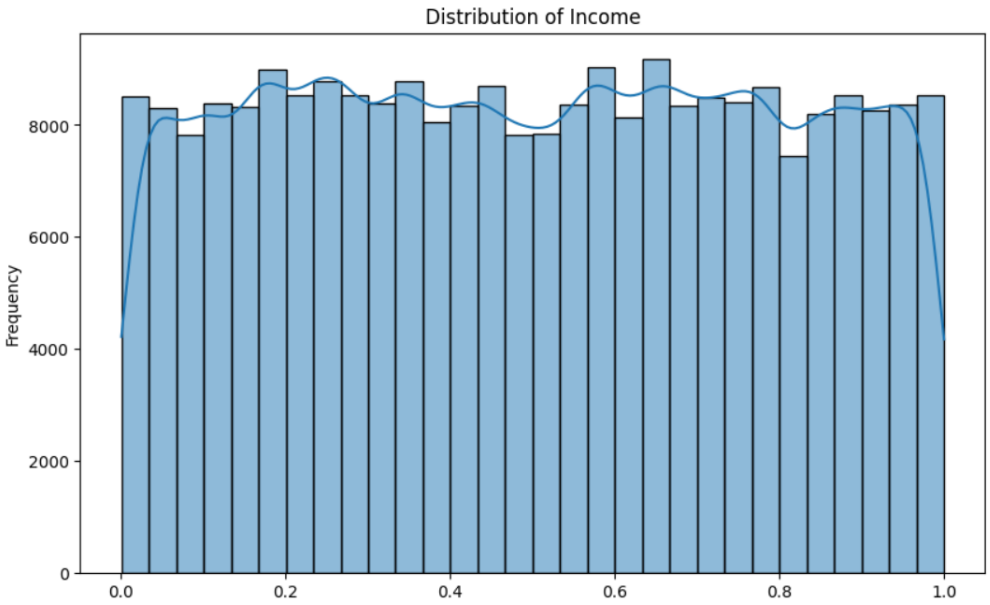
```
import pandas as pd
data = pd.read_json("loan_approval_dataset.json")
data.head()
```

	Id	Income	Age	Experience	Married/Single	House_Ownership	Car_Ownership	Profession	CITY	STATE	CURRENT_JOB_YRS	CURREN
0	1	1303834	23	3	single	rented	no	Mechanical_engineer	Rewa	Madhya_Pradesh	3	
1	2	7574516	40	10	single	rented	no	Software_Developer	Parbhani	Maharashtra	9	
2	3	3991815	66	4	married	rented	no	Technical_writer	Alappuzha	Kerala	4	
3	4	6256451	41	2	single	rented	yes	Software_Developer	Bhubaneswar	Odisha	2	
4	5	5768871	47	11	single	rented	no	Civil_servant	Tiruchirappalli[10]	Tamil_Nadu	3	

This is how our data can be seen as soon as we load our json file.

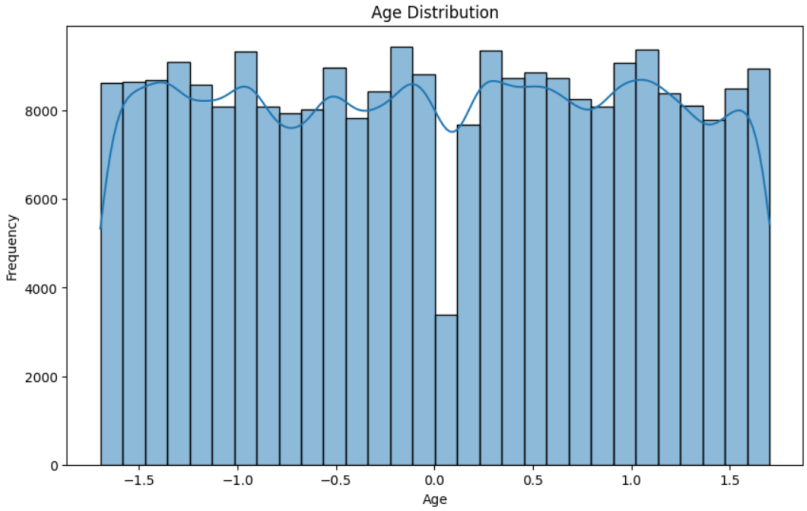
These graphs are created by **matplotlib** and **seaborn** libraries in Python.

Distribution of Income:



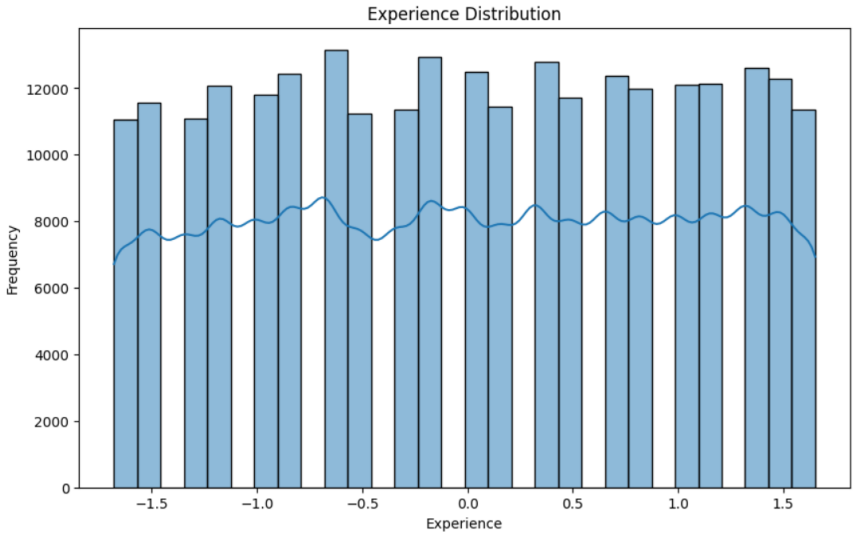
The income distribution plot shows the range and frequency of income levels among the clients. We can observe if there is a wide variation in income levels or if most clients fall within a certain income bracket, which can help in understanding the financial stability.

Age Distribution:



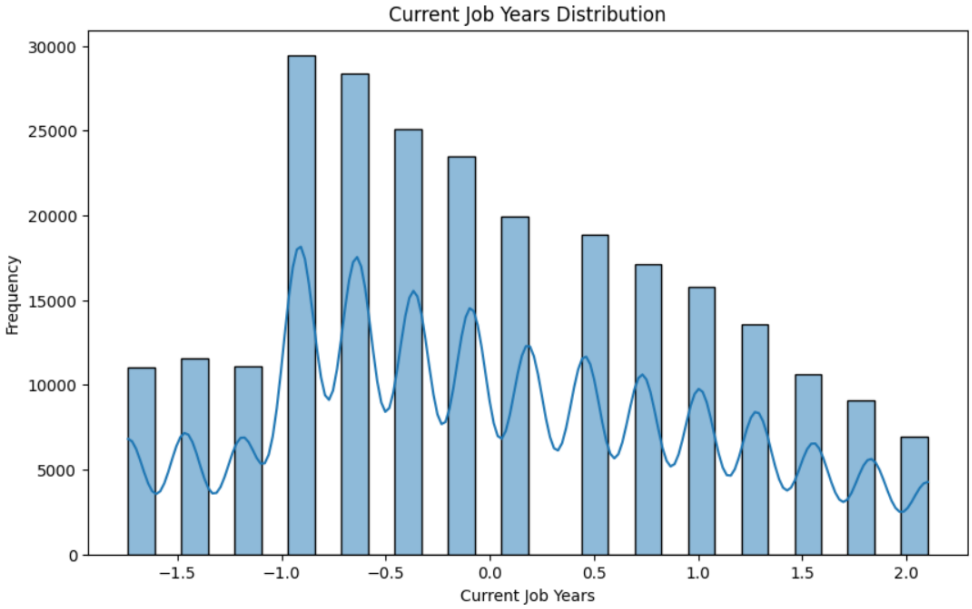
The age distribution plot reveals the age demographics of the clients. It helps identify if there is a concentration of clients in certain age groups, which can be useful for tailoring marketing strategies and understanding risk associated with different age brackets.

Experience Distribution:



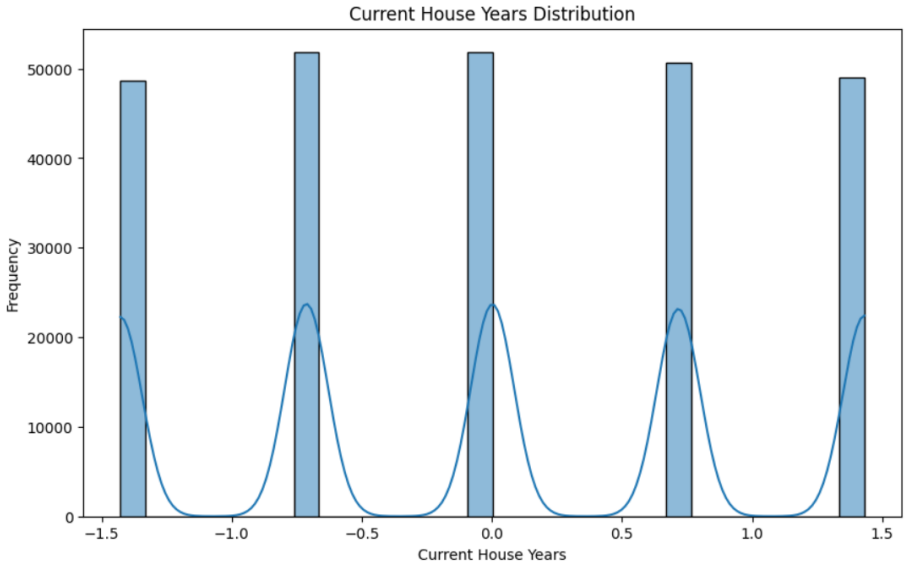
The experience distribution plot shows the number of years of professional experience among clients. This can indicate the potential job stability and earning capacity of the clients, which are important factors in assessing their creditworthiness.

Current Job Years Distribution:



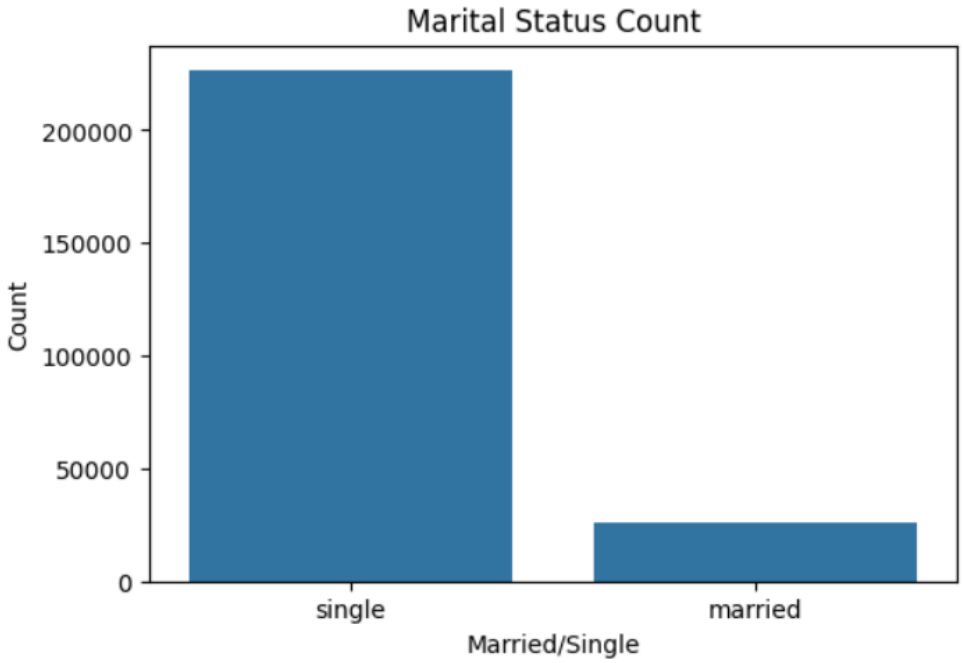
The current job years distribution highlights how long clients have been in their current job. Longer job tenure can suggest job stability, which is an important factor in assessing the risk of loan default.

Current House Years Distribution:



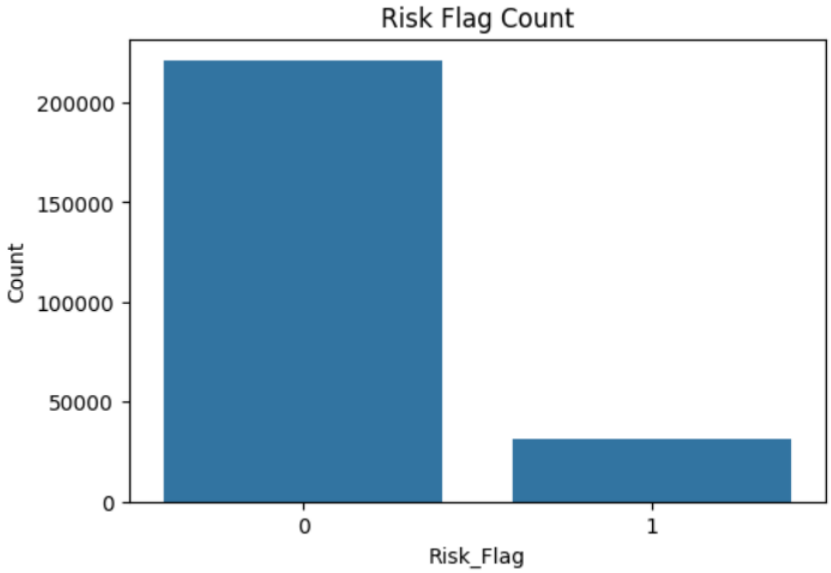
The current house years distribution indicates how long clients have been living in their current residence. Longer residency can suggest stability and a lower risk of default.

Marital Status Count:



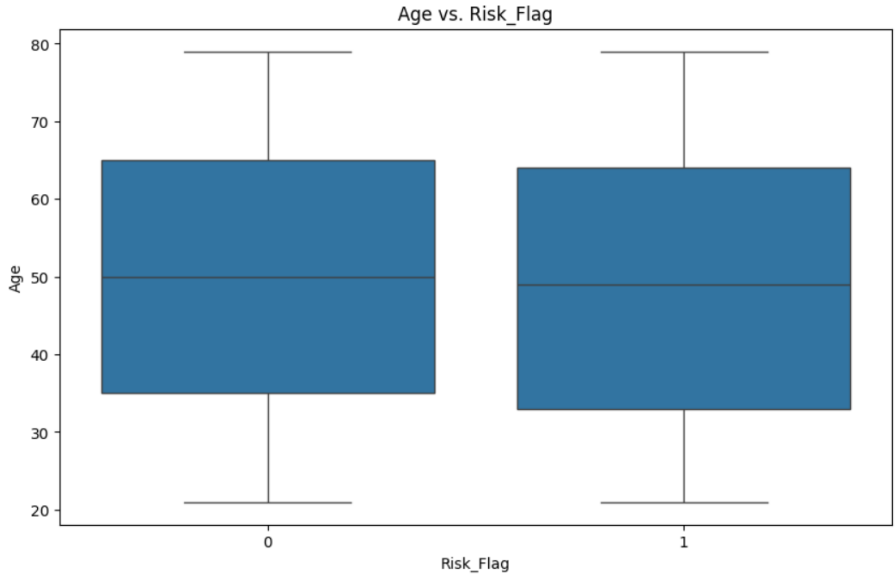
The marital status distribution shows the proportion of married versus single clients. This can be relevant as marital status may impact financial stability and spending behavior, influencing loan repayment ability.

Risk Flag Count:



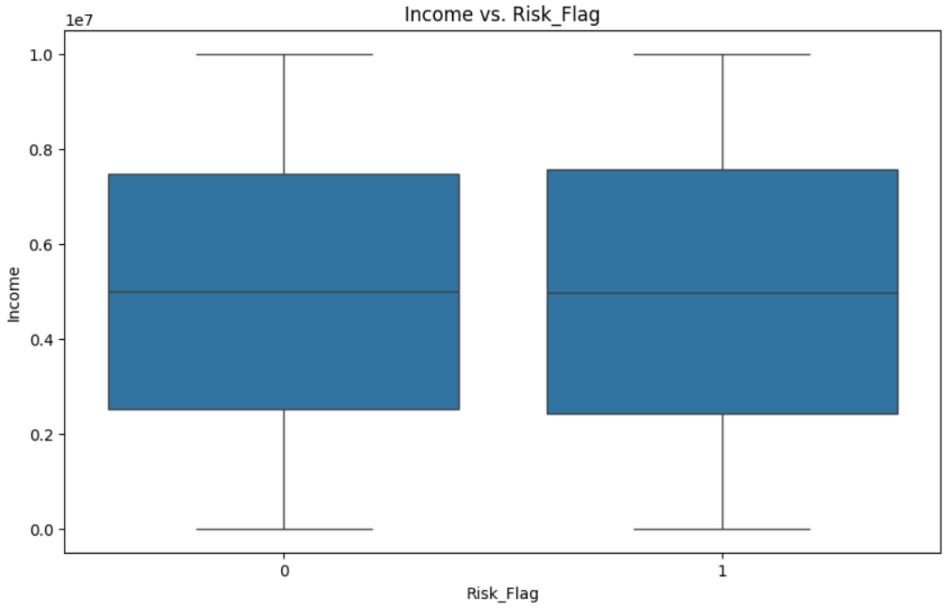
The risk flag distribution highlights the proportion of clients classified as high risk versus low risk. This provides a clear picture of the overall risk profile of the client base, which is crucial for financial planning and risk management

Age vs Risk Flag:



This box plot will show how the age distribution varies between low-risk and high-risk clients. It helps identify if there are specific age groups that are more likely to be classified as high risk.

Income vs Risk Flag:



This box plot will illustrate the distribution of income for low-risk and high-risk clients. It allows us to see if there are significant differences in income levels between the two risk categories, which can inform risk assessment criteria.

Data Exploration Insights

- **Income**: The income distribution is right-skewed, indicating most clients have lower incomes, with a few clients having very high incomes.
- **Age**: The age distribution is fairly uniform across different age groups, with some peaks at certain age ranges.
- **Experience**: Experience distribution shows a higher frequency for lower years of experience.
- **Current Job Years**: Most clients have been in their current job for fewer years.
- **Current House Years**: The distribution is fairly uniform with a slight right skew.
- **Marital Status**: There are more single clients than married ones.
- **House Ownership**: Most clients are renting their houses.
- **Car Ownership**: A significant portion of clients do not own a car.
- **Profession**: Various professions are represented, with some having higher counts.

Model Performance

Why Random Forest Model?

Random Forest is often considered a strong choice for loan risk prediction due to its high accuracy and robustness. It mitigates over fitting by averaging multiple decision trees and effectively handles complex non-linear relationships. It also provides insights into feature importance, aiding interpretability. Its versatility in handling various types of data, including those with missing values, makes it particularly suitable for this problem.

Random Forest Model Evaluation:

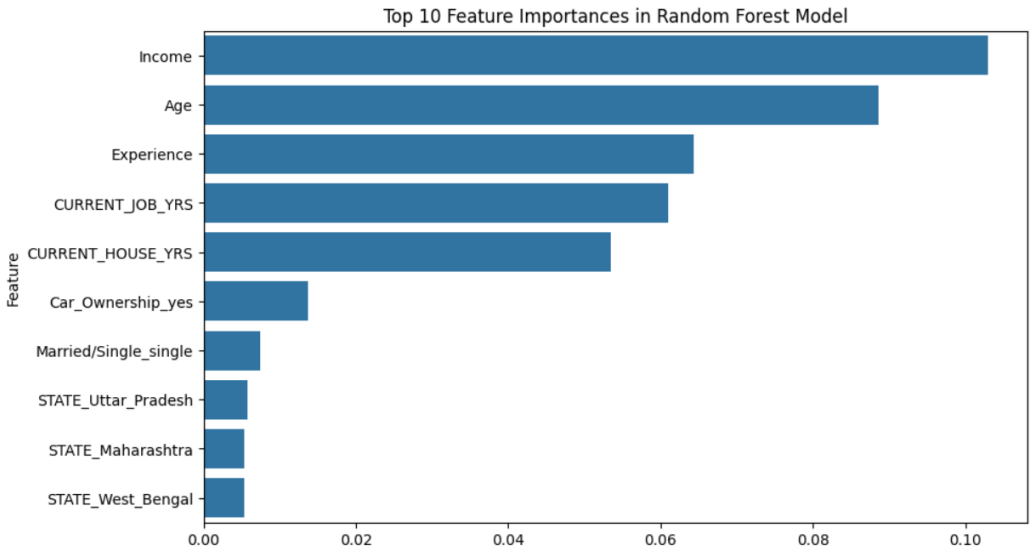
Accuracy: 0.8992

Precision: 0.6072

Recall: 0.5314

F1 Score: 0.5668

Understanding Main Deciding Factors Associated with Risk



- **Income**: Higher or lower income can significantly affect the risk assessment.
- **Age**: Older clients may have different risk profiles compared to younger ones.
- **Experience**: More experienced individuals might be considered lower risk.
- **Current Job Years**: Stability in current employment can be a strong indicator of lower risk.
- **Current House Years**: Longer duration in the current house may indicate stability.

By analysing these feature importances, lenders can gain insights into which factors are most critical when assessing loan risk and adjust their decision-making processes accordingly.