**Assignment-1**

1. How would you segment customers based on their risk (of default)?

From the comprehensive analysis, we observe the following risky groups based on the given dataset:

* + Younger individuals
  + Females (slightly higher chance than males)
  + Divorced/Remarried females and divorced males
  + Unskilled residents and non-residents
  + Individuals with multiple pending loans
  + Individuals living on a free basis (slightly)
  + Applicants for education loans and vehicle loans (slightly)
  + Applicants owning real estate and having some agreement or savings (slightly)
  + The absence of a coapplicant is risky, but the presence of a coapplicant does not guarantee less risk.
  + Whether a guarantor is present or not, there is associated risk in both cases.

Elaborating on these points:

* + Males have almost equal probabilities of being low or highly risky. However, young females (between the ages of 25-40 years) are more likely to be highly risky.
  + Single and married individuals are less likely to be risky, while divorced and remarried individuals are more likely to be risky.
  + Unskilled residents, unemployed unskilled non-residents are highly risky, while individuals in employment and high positions have an equal likelihood of being both.
  + Individuals with very high savings balances are less likely to be highly risky, but others have an equal likelihood.
  + Having no previous loans or fully paid loans and delayed loan payments are very unlikely.
  + Surprisingly, existing loan holders and loans at this bank are equally likely to be highly risky or less risky. However, those with multiple pending loans are more skewed towards being highly risky.
  + Applicants who own their house and freeloaders are slightly more likely to be highly risky than renters. However, this may be due to a lack of more data points.
  + Education loans are highly risky, while loans for used vehicles, business, and domestic appliances are less risky. Career development loans have zero associated risk. However, this may also be due to a lack of more data points. For other categories, the risk levels are equal.
  + People who own real estate properties and have agreements/savings are slightly risky.
  + Having no coapplicant is risky, but the presence of a coapplicant does not guarantee less risk. Similarly, whether a guarantor is present or not, there is associated risk in both cases.
  + The number of loans at this bank does not reveal any significant features.
  + Overall, drawing conclusive insights about the groups is difficult due to the limited number of data points (around 1000).

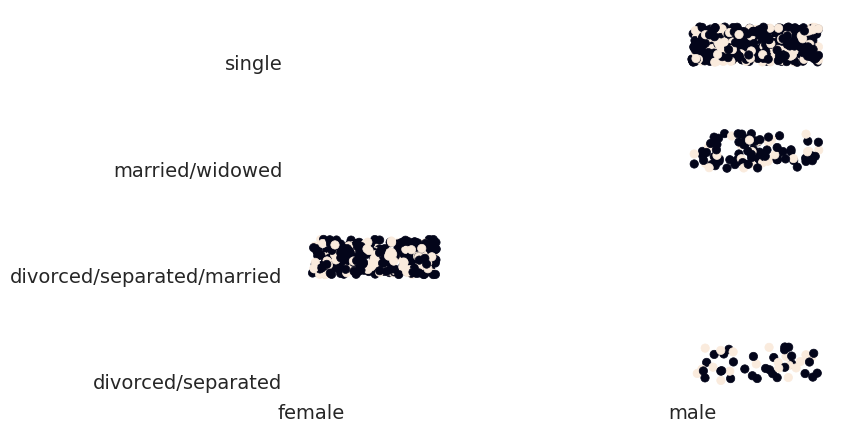
1. Which of these segments / sub-segments would you propose be approved?

* Married males
* Management/Highly qualified employees
* Very high saving account balance
* No loans taken/All loan paid back fully and previous loans duly paid back
* People having their own house
* Domestic appliances, Business and Career development loan applicants
* Has coapplicant and guarantor with fewer loans at current bank

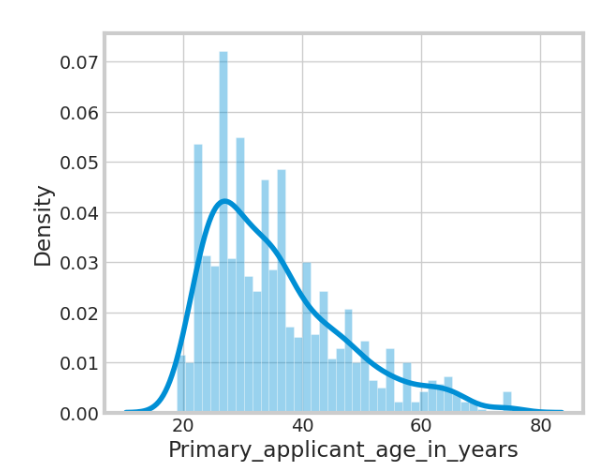
1. What other insights in general can you share about these segments and observations on the data itself (completeness, skews) and how you would treat any anomalies (for eg - missing data)?

Perhaps due to lack of datapoints, there was dominance of some groups over others. Since, it was observed that riskier groups came from those datapoints who have maximum number of observations. There rarely happened to be any outlier.

Eg-1: Only Divorced/Remarried females are present in the dataset

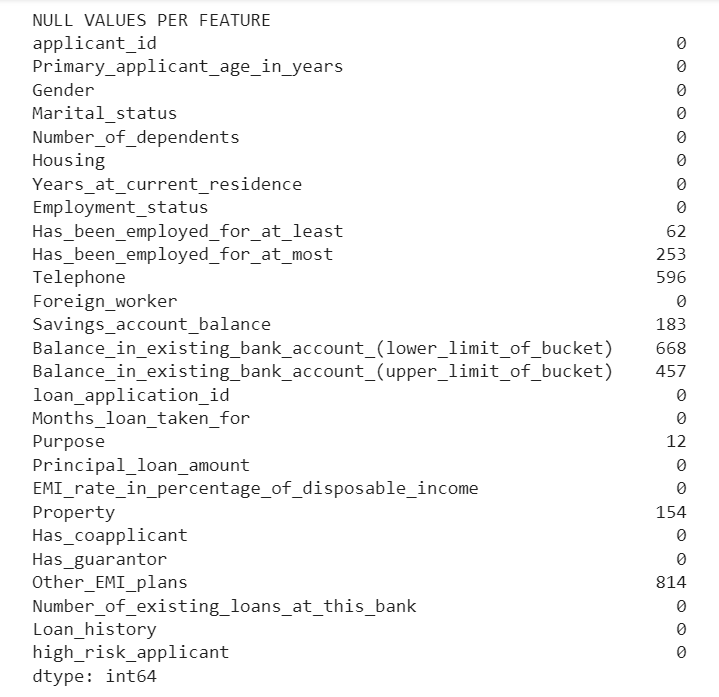


Eg-2: Few number of old people were present



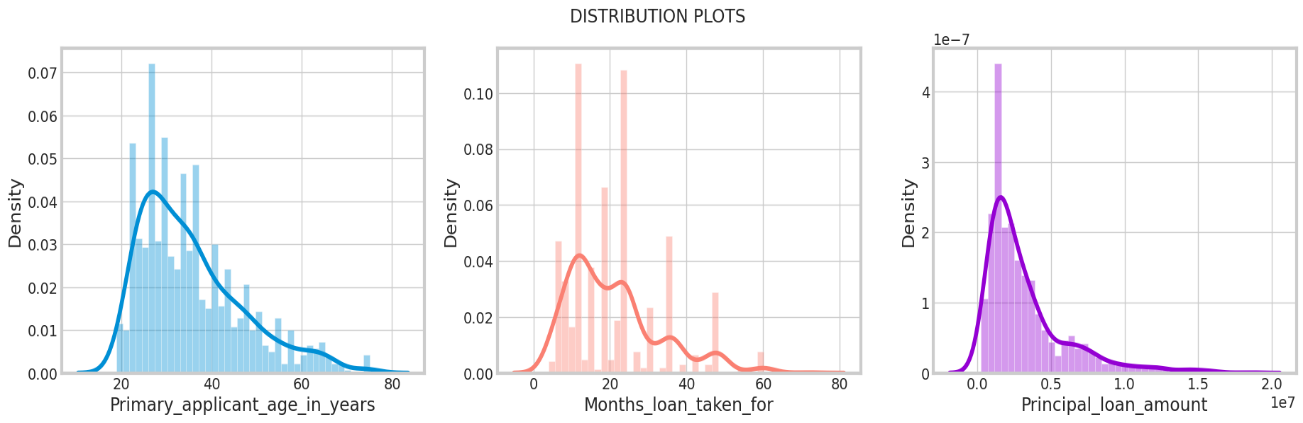
In the given dataset,

* No duplicate values were found.
* Dataset is balanced with respect to high\_risk variable (only 30% are highly risky).
* However, several features contained null values, including Telephone, Range of bank balance, Property, Other EMI plans, Saving account balance (categorical), and maximum working period (in years).
* The features Telephone, Range of bank balance, and Other EMI plans were dropped from the analysis as they were either redundant or not useful.
* For the categorical feature Saving account balance, the most\_frequent technique from Scikit-learn's Imputer was used to replace the null values. The same process was applied to the Purpose feature.
* For the maximum working period (in years), NaN values can be replaced with the difference between the current age and 23 (assuming the minimum working age). This assumption is based on the observation that most cases with missing values belonged to senior workers and older individuals.



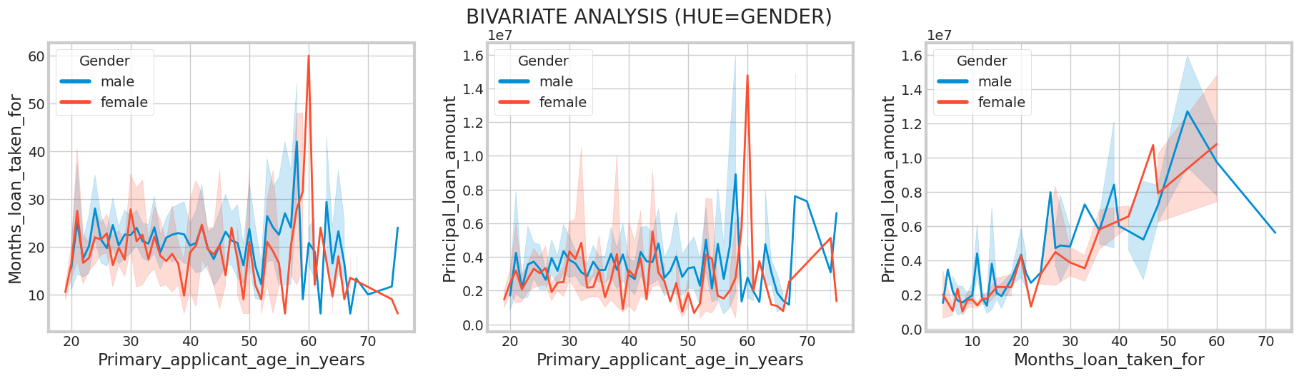
* For less frequent NaN columns like Least employment period (in years), we can replace the missing values with 0 for individuals with a maximum working period of 0. Otherwise, we can replace them with the difference between the current age and 23. However, the latter case occurred very rarely. Therefore, the primary approach was to impute the missing values with 0.
* Age and Principal loan exhibited positive skewness, indicating that the majority of the values were concentrated towards the lower end with a few extreme values in the higher range. On the other hand, the loan month period showed multiple peaks, such as at 12 months, 24 months, 36 months, 42 months, and so on.

Image:

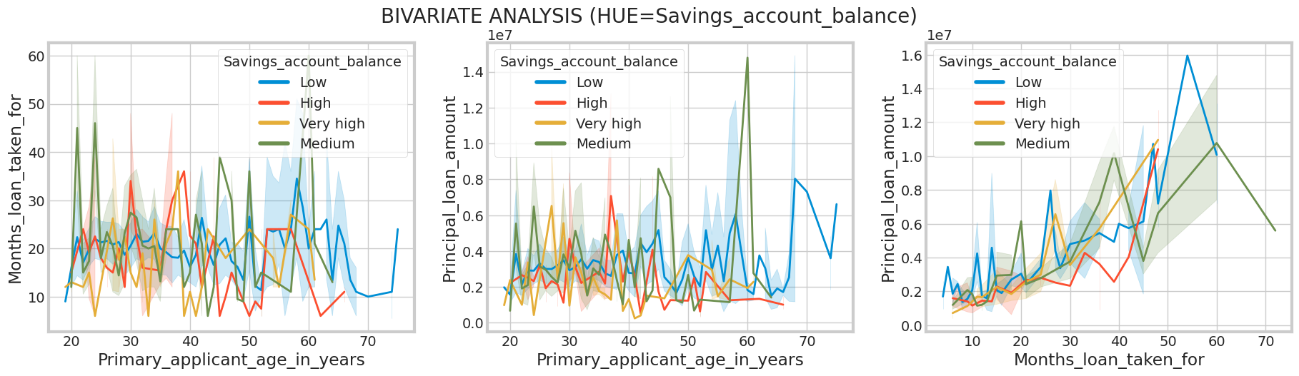


More insights:

* The majority of loans were taken by individuals in the age group of 25-40 years.
* The most common loan amount range chosen by applicants was between 20 Lac Rs and 45 Lac Rs, with the most popular choice being 25 Lac Rs.
* A significant number of individuals opted to clear their loans within a period of 1-2 years, with the majority selecting a 1.5-year loan period.
* The majority of loan applicants consist of single males and remarried males.
* Approximately 63% of loan applicants are skilled professionals, indicating a higher level of creditworthiness. However, a significant portion of unskilled individuals and people in high-level positions have also applied for loans.
* Many house owners have opted for a loan, particularly for the purpose of acquiring electronic equipment, new and used vehicles, and furniture, fixtures, and equipment (FF&E) for business purposes.
* Around 75% of loan applicants have low cash in their savings accounts, highlighting the need for a loan to meet their financial requirements.
* More than 50% of the applicants have a history of being loyal and paying back their loans on time. However, approximately 30% of applicants have not done so, posing a high risk of default.
* Overall, around 30% of the applicants are considered highly risky, while the remaining 70% are considered suitable for loan approval. This indicates an imbalance in the dataset, with a higher proportion of risky applicants.
* Interestingly, the majority of applicants do not have a co-applicant or a guarantor, which are factors that typically help convince lenders of lower risk.
* Lastly, most applicants have a low number of pending loans at the specific bank being analyzed.



* Based on the above plots, it is evident that individuals in the younger and middle age group (below 55 years) tend to take principal loans in the most frequent range, as described earlier, with a loan period of 1-2 years.
* In contrast, senior or retired citizens tend to take larger loan amounts and opt for longer loan periods.



* Among senior citizens, those with low to medium savings account balances tend to opt for retirement age loans more frequently, while those with high balances rarely choose this option.
* In contrast, among the younger population, applicants with high savings account balances are more likely to apply for loans.

1. Real-world data features to see in the dataset and why these features might enrich the analysis?

* Including additional features such as Area (in the form of State or City/Town), Company/Startup the applicant is working for, Job Profile, and Educational qualification would have significantly enhanced the analysis.
* Area information would provide context for house loans, allowing us to consider regional factors that could impact creditworthiness.
* Company/Startup information would offer insights into the job security and growth potential of the applicant. Working for a reputable company/startup would indicate a higher level of creditworthiness.
* Job Profile and Educational qualifications would provide similar insights into the applicant's professional background and competence.

By incorporating these features into the analysis, we would have a more comprehensive understanding of the applicant's creditworthiness and the broader context surrounding their financial profile.