Credit EDA Assignment

DS C44 - Sai Sharan Paspunuri

19th July, 2022.

Problem Statement

Business Objectives

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

Process followed for the EDA.

Approach for the Application_data dataset.

- Understand the dataset by loading & perform various actions.
- Identify the missing values in the column.
- Drop the columns where the values are missing in significant % (Note - 40% is my threshold value)
- Drop the columns which do not add any value to the analysis. (this will make my dataset lighter)
- I will find out the data imbalance ratio
- Identify the outliers and should be mindful about them.
- Perform the analysis. Pick 5 variables/attributes and perform the univariate, bivariate & segmented univariate analysis.
- I will do this analysis for 20-25 columns in loop fashion by selecting 5 variable/attribites at a time. viz., Analysis 1, Analysis 2, Analysis 3, Analysis 4, Analysis 5.
- I will find out the correlation between the top 5 driving variable/attributes along with the Target variable.

Step by step process on the previous_data dataset analysis.

- Understand the dataset by loading & perform various action.
- Identify the missing values in the columns.
- Drop the columns where the values are missing more than 40%.(40% is my threshold value)
- Drop the columns which do not add any value to the analysis. (this will make my dataset lighter)
- Identify the outliers and should be mindful about them.
- Perform the analysis. Pick 5 variables/attributes and perform the univariate, bivariate & segmented univariate analysis.

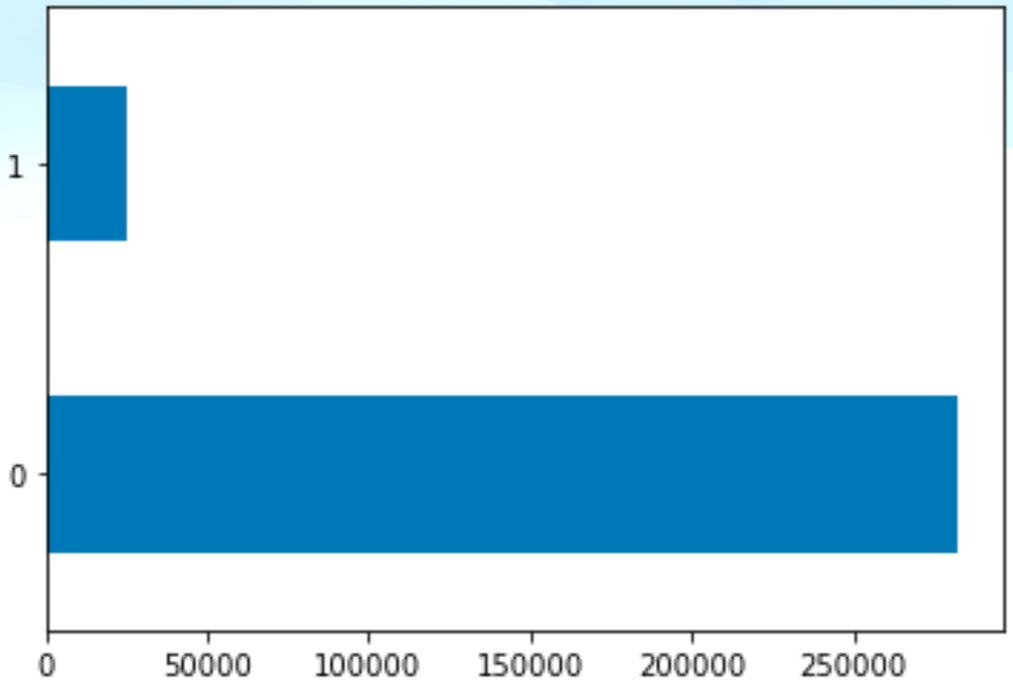
Step by step process on creating the Merged data set and performing analysis.

- Merge the data frames using the Inner join.
- Perform the various actions to understand the dataset.
- Perform the Bivariate analysis using the target variable with the previous dataset attributes.

Analysis findings

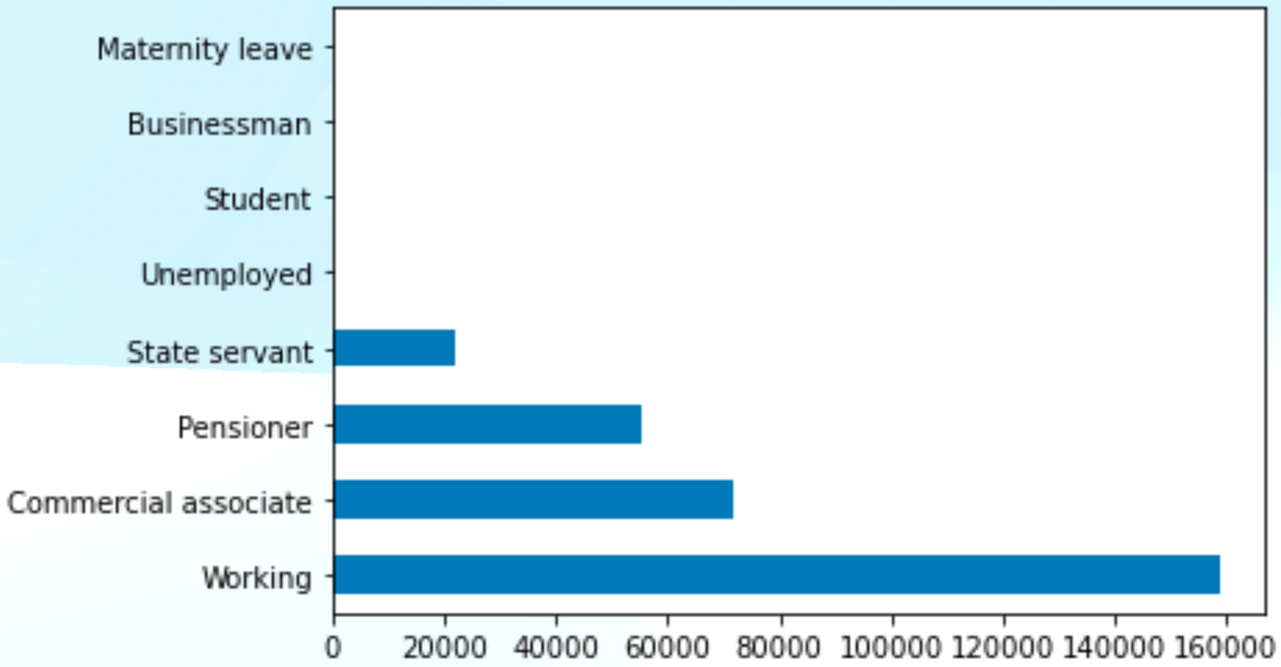
Data Imbalance Ratio.

- The graph shows that how the data is majorly one sided with all the non-defaulters.
- Around 92% of the data has no defaulters.
- The remaining 8% data is of defaulters.
- The data Imbalance ratio is **11.38 : 1**

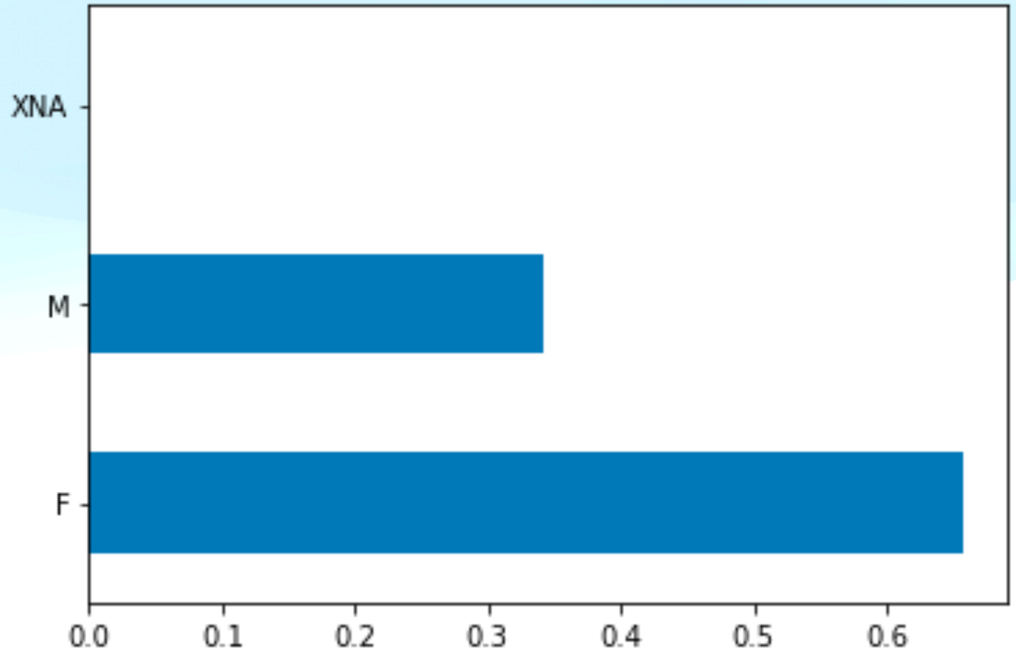


Analysis findings

Univariate Analysis



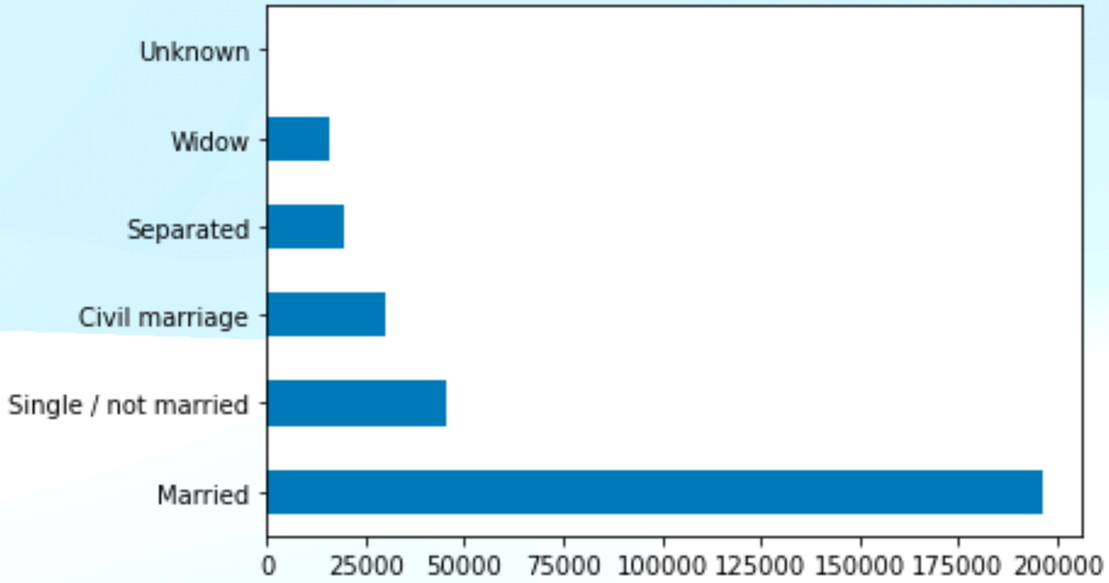
The above gives us an insight there are mostly 4 income types.
Majority are working class followed by commercial associate.



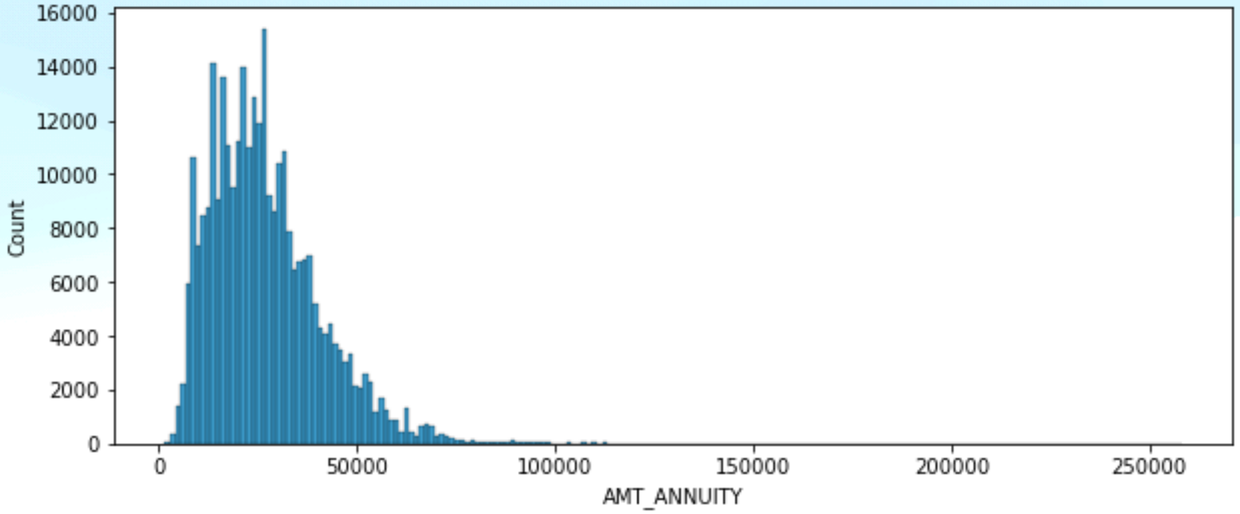
The female applications are more in number than male.
This is an interesting insight.

Analysis findings

Univariate Analysis



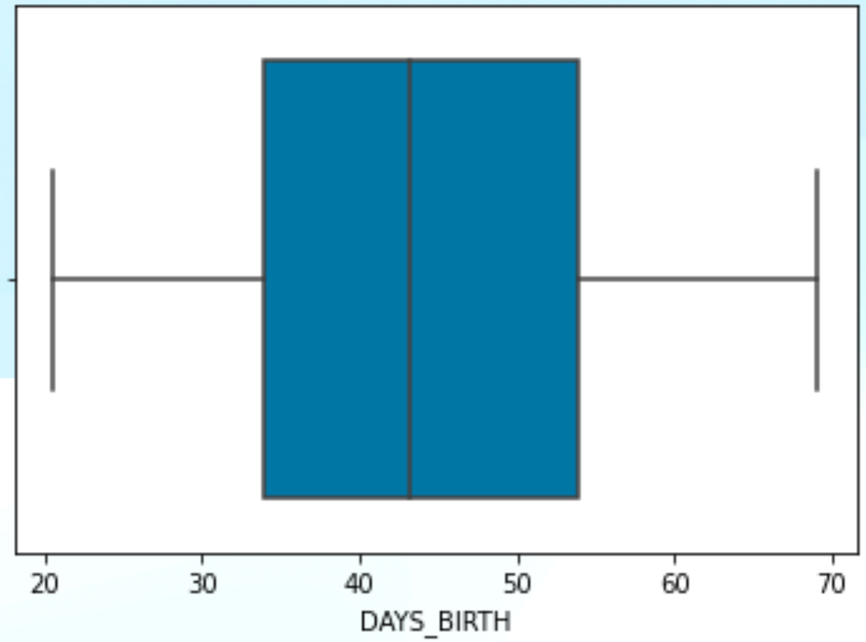
The above clearly shows that majority of the applicants **family status** is 'Married'or 'Single'



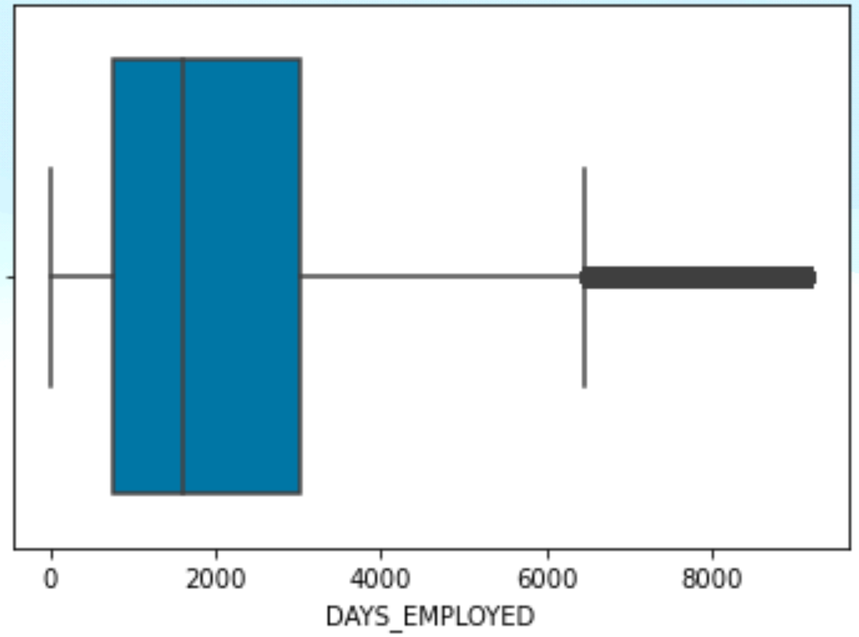
The majority of the **annuity** amount is less than 50,000.

Analysis findings

Univariate Analysis



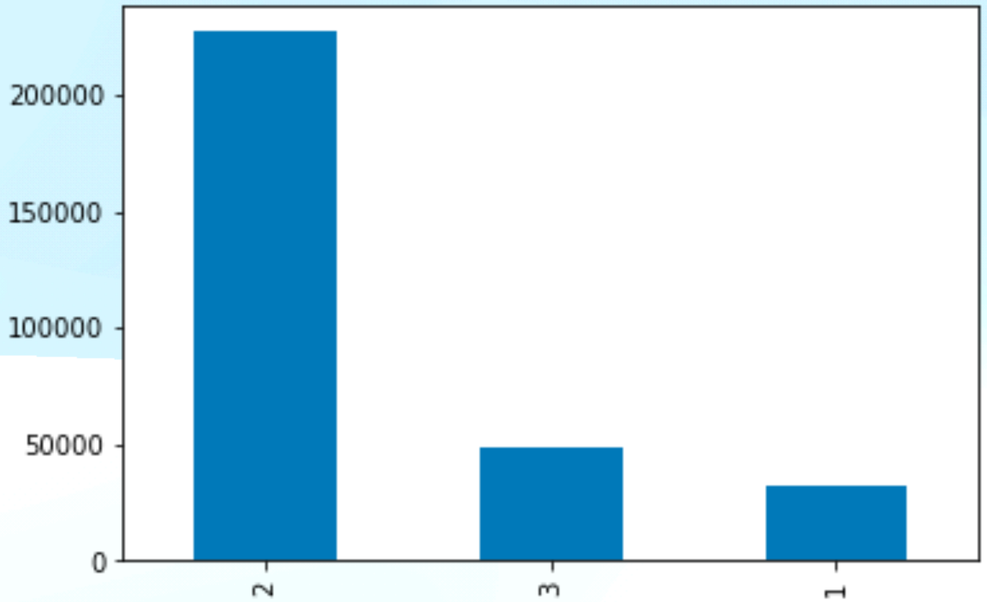
- The **median age** of the applicants is 42-43 years.



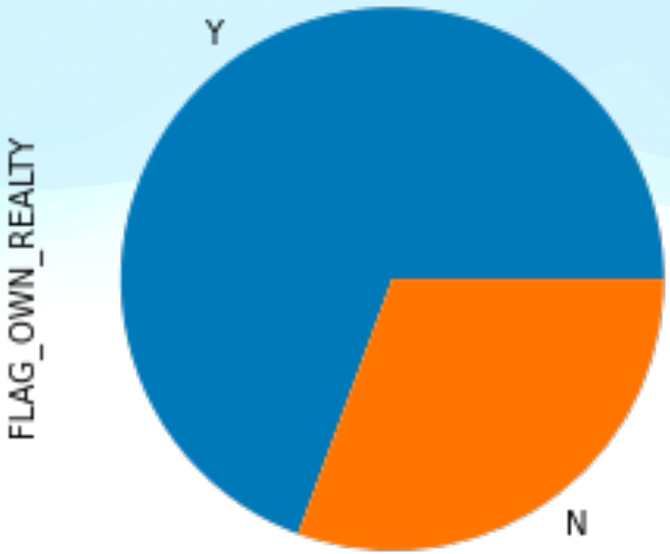
The above plot infers that more number of applicants are above the **median work exp.** It is safer to say people with more work exp tend to seek loans.

Analysis findings

Univariate Analysis



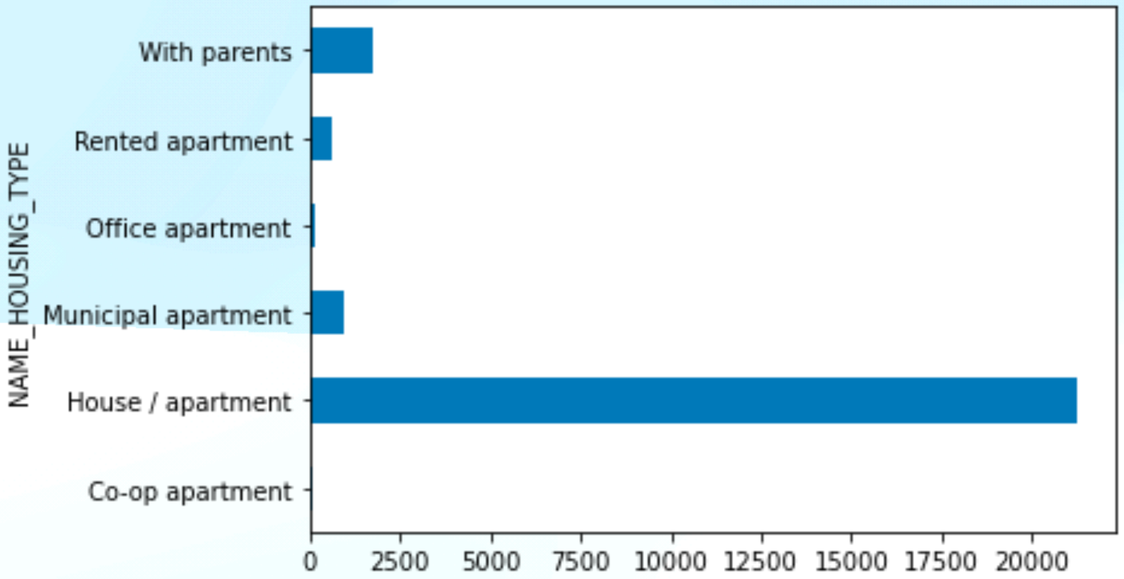
The majority of the applicants are from **region which has a rating of 2.**



The majority of the applicants **own the house**

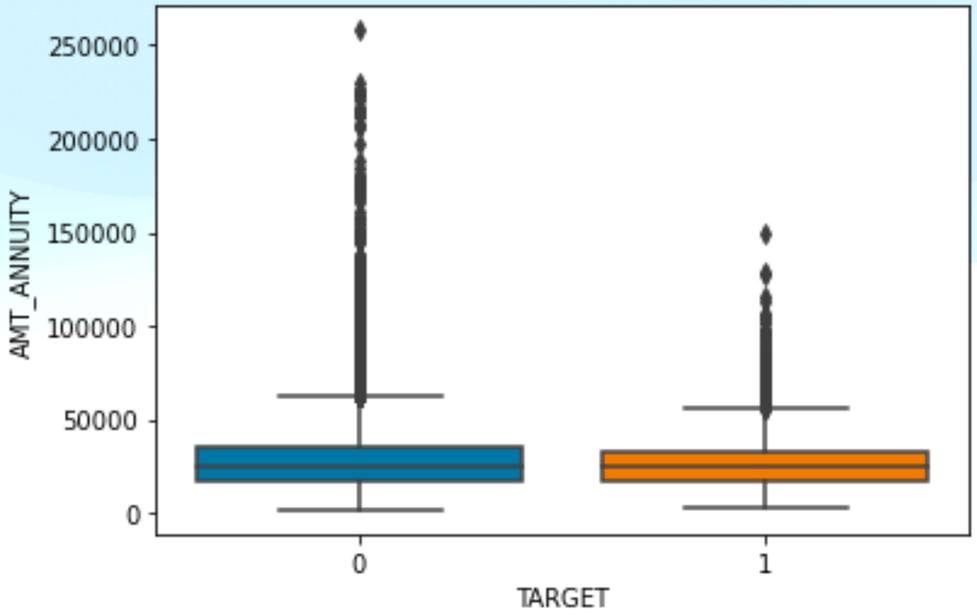
Analysis findings

Bivariate Analysis



From the above it can be inferred that,
With parents, municipal and Rented apartment - housing type has the highest defaulters.
This is by observing the value counts & the graph.

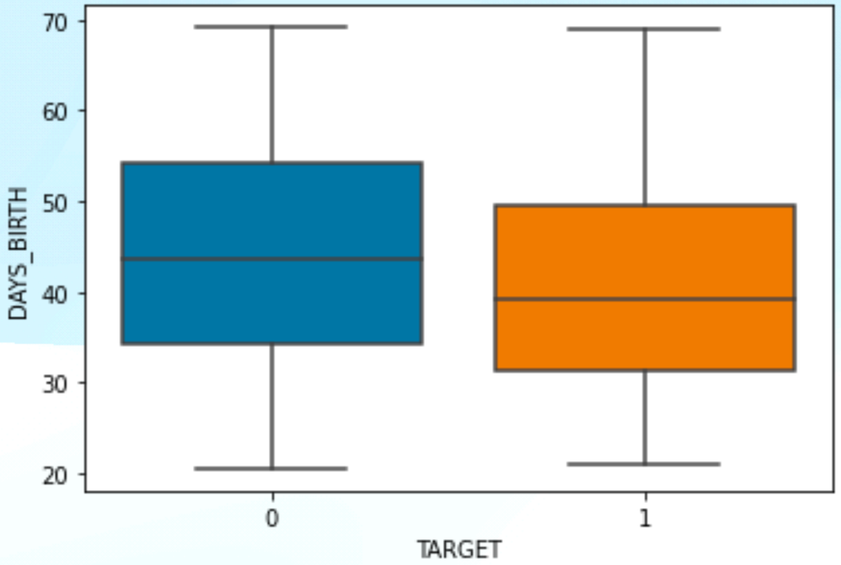
the annuity spread is higher above the median
and continuous after whisker for non-defaulters.



Inspite the defaulters annuity spread is low,
they might tend to default because of their low income.

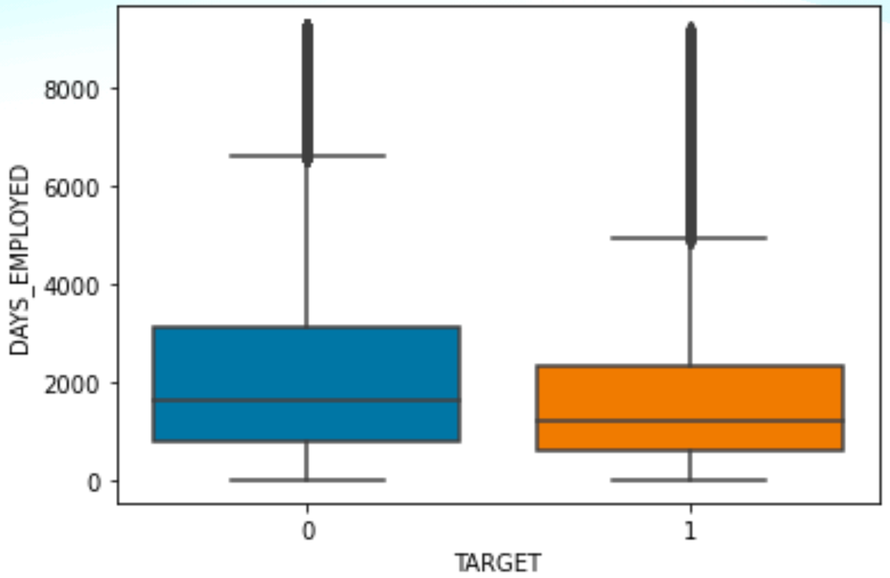
Analysis findings

Bivariate Analysis



From the above we can infer that,
the defaulters median age and the 75%ile of data is less than the non-defaulters.
It us safer to assume that
older applicants with certain criteria are a safer option than the younger ones.

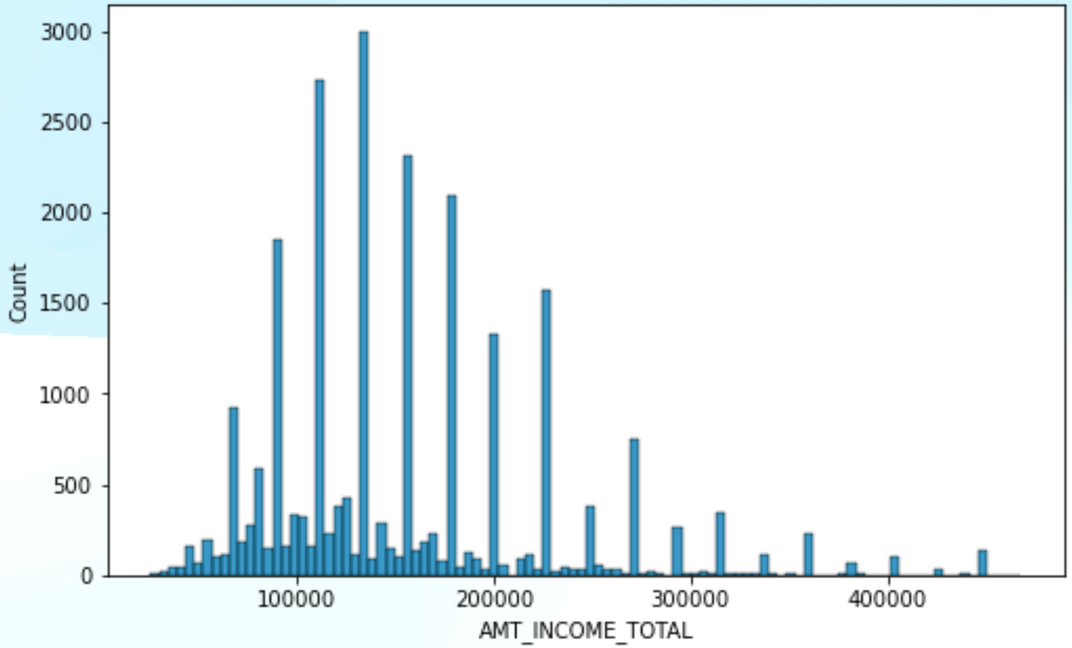
From the below we can infer that,
the defaulters median work experience and the 75%ile of data is less than the non-defaulter.
It us safer to assume that
applicants with more work exp & certain criteria are a safer option than the younger ones.



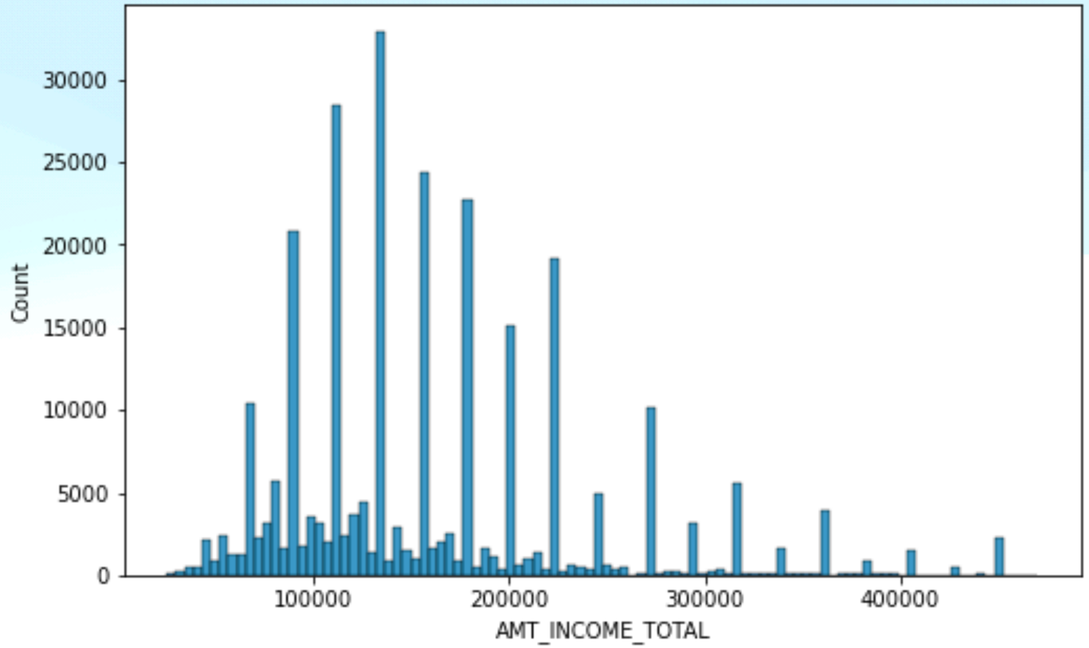
Analysis findings

Segmented Data Analysis

Defaulters



Non-Defaulters

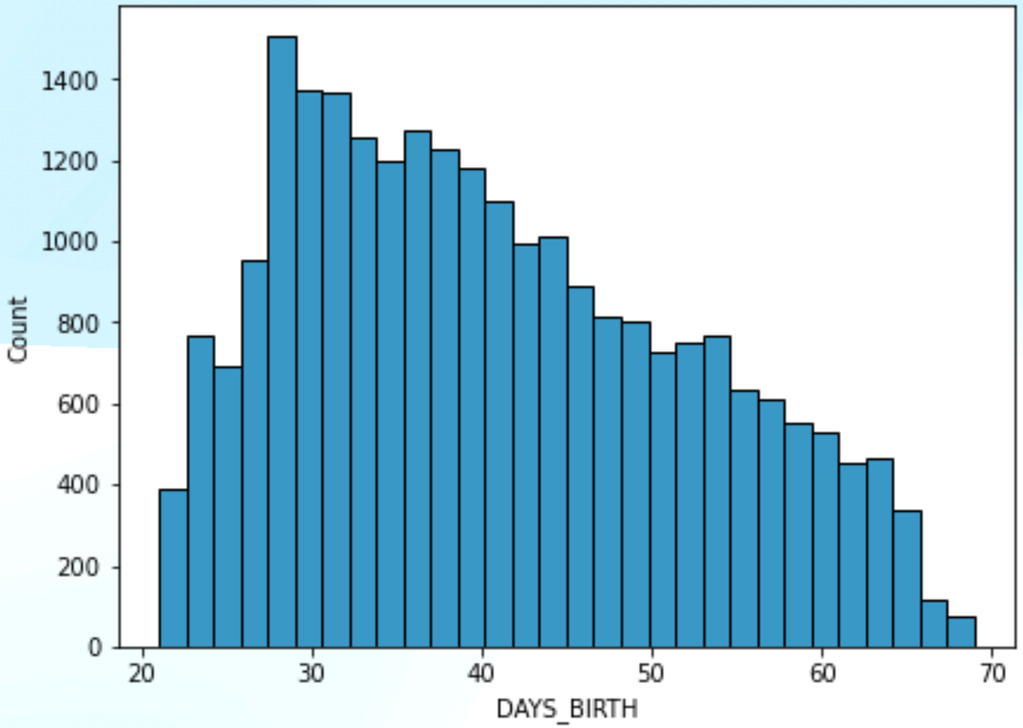


The above clearly shows that, most of the defaulters(left side graph) income levels are less than 2,50,000.

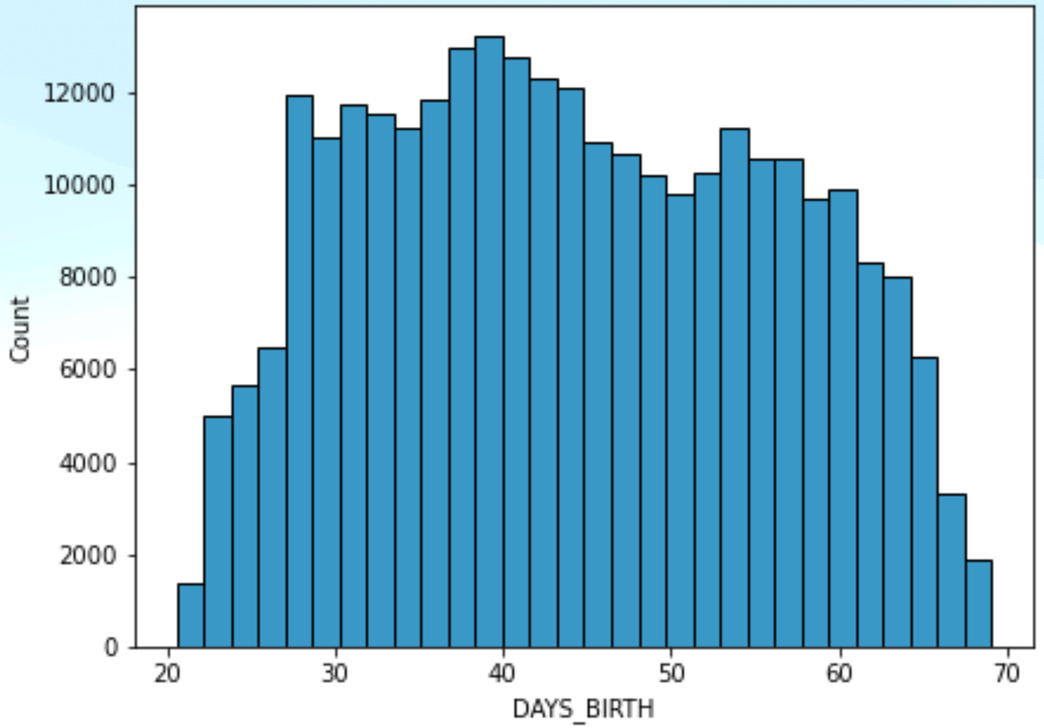
Analysis findings

Segmented Data Analysis

Defaulters



Non-Defaulters

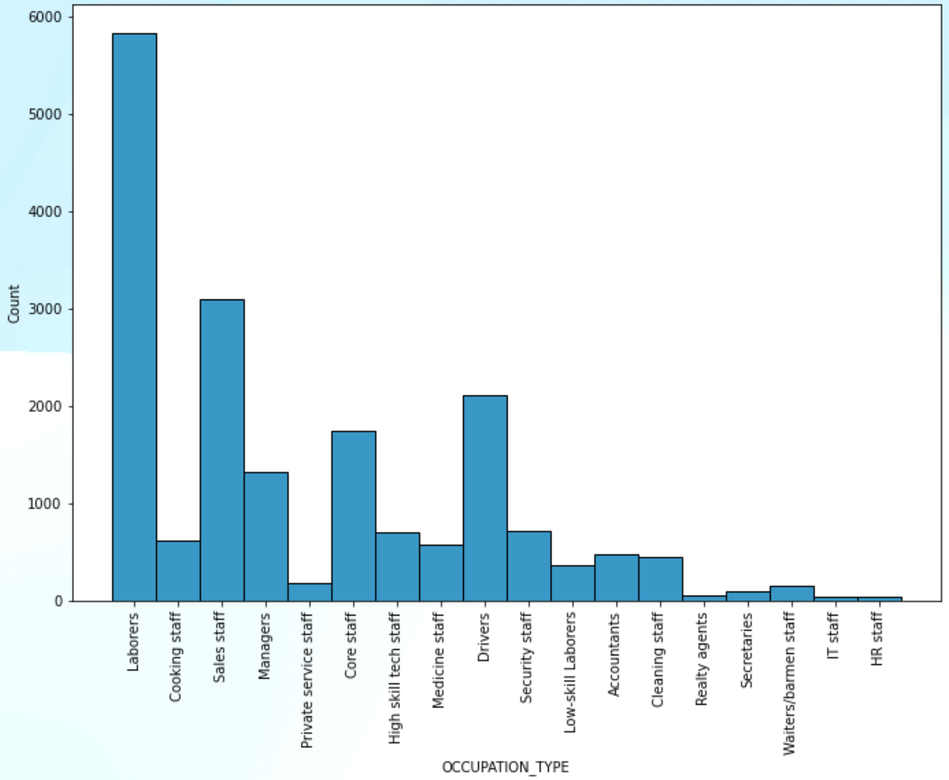


As inferred from the bivariate analysis, Applicants with less age tend to default. Here you can notice that default rate is hight among applicants with age less 30 years.

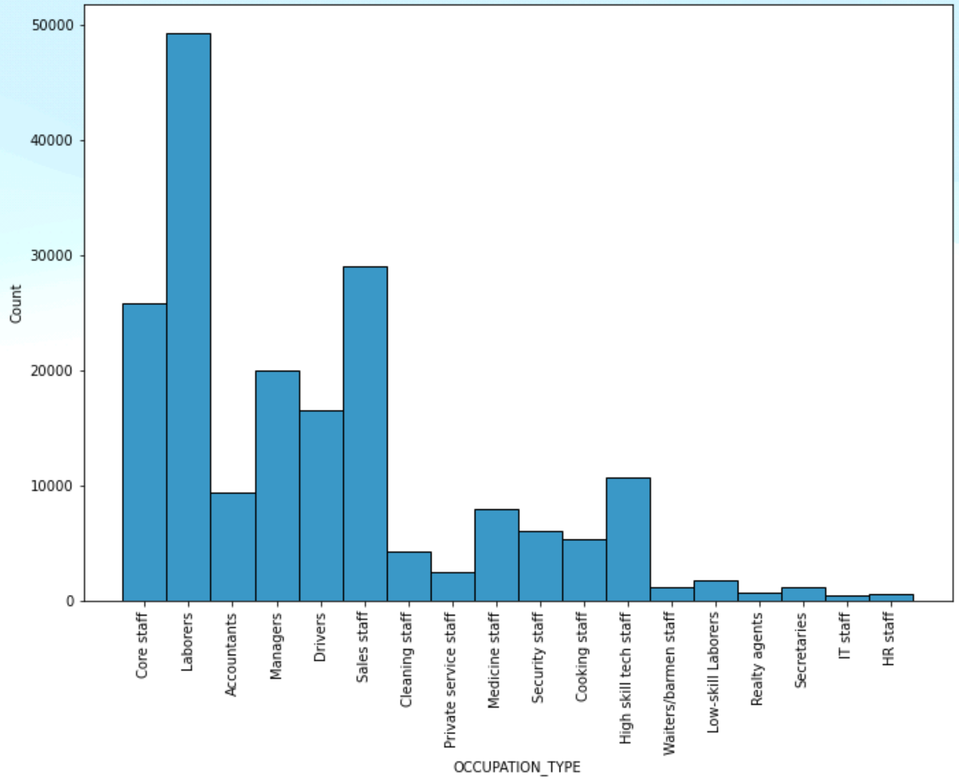
Analysis findings

Segmented Data Analysis

Defaulters



Non-Defaulters



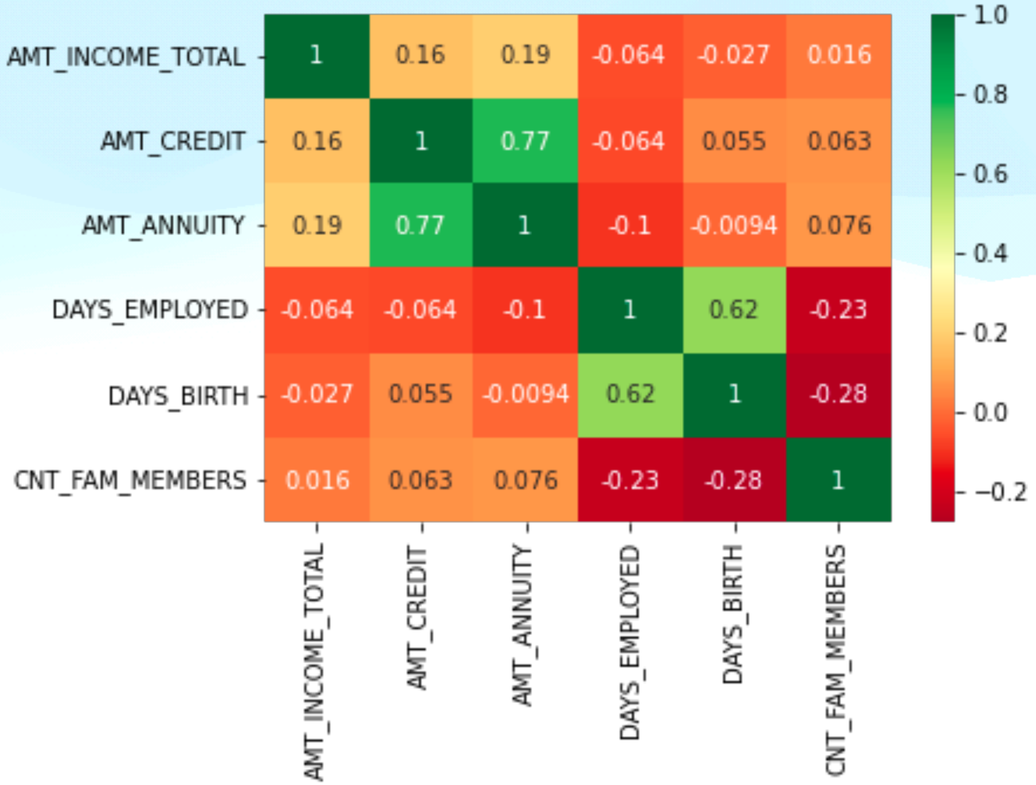
The above graph clearly states that, Labourers are the highest in applicants but, they are also highest in the % defaulters. Labourers - occupation has more than 10% defaulters where as less than 10% defaulters in other types.

Analysis findings

Data Correlation

From the above we can find the obvious correlation between loan amount & the loan annuity, and age & work experience.

There is also a slight correlation between Income & the amount credited and also with Income & the amount annuity.



Analysis findings

- The primary defaulters are the people with income less than 2,50,000/-
- There are more female applications than the male applications.
- The driving factors are the Income, Age, Work Experience, Owning a reality, & Owning a car.
- The loan can be avoided to Singles/Unmarried. The default rates are high comparatively.
- The applicants working with ‘Business Entity 3’ Organisation type can be preferred more. They are likely not to default.

Analysis findings

- Applicants with the median age of 40-45 Years and good work experience can be the top priority.
- The most type of clients are the Repeaters. A top preference can be given to the repeater applications.
- The applications coming from ‘Credit and cash offices’ are high in number. Please use all the criteria’s mentioned above to make faster decisions.

Thank you!