EDA ASSIGNMENT

Roy Mathew Benjamin

Introduction

This assignment aims to give you an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Business Understanding

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicant’s profile. Two types of risks are associated with the bank’s decision:

If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

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.

STEPS TAKEN

- Check for missing values and handle them
- Check for outliers and fix them if needed
- Check for correlation for clients with payment difficulties and other variables within the files
- Identifying and analysing relevant correlationships

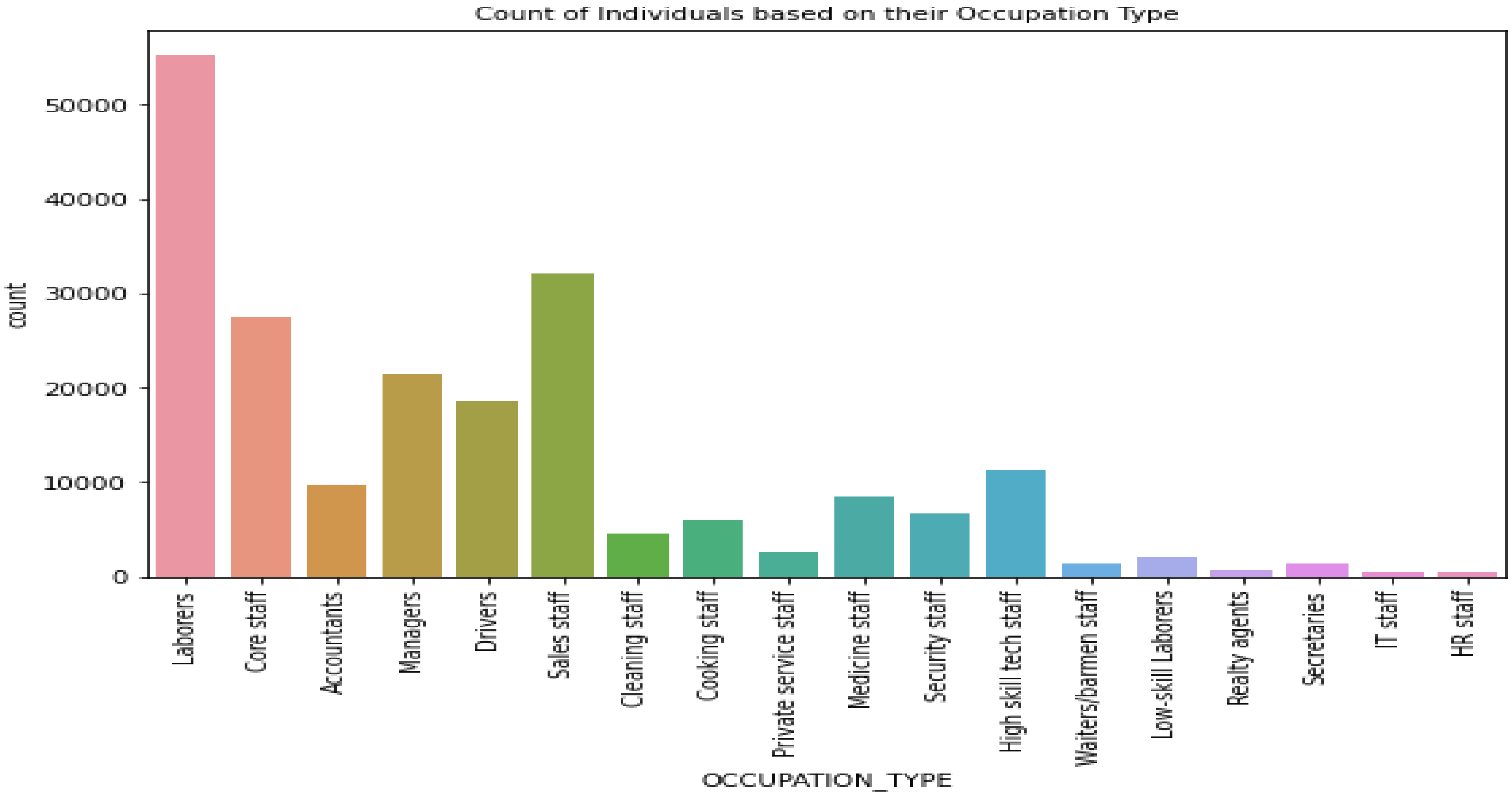
MISSING VALUES

- Find out columns with more than 50% null values and remove them
- Find out columns with 13-15% null values and fill them with valid/meaningful values

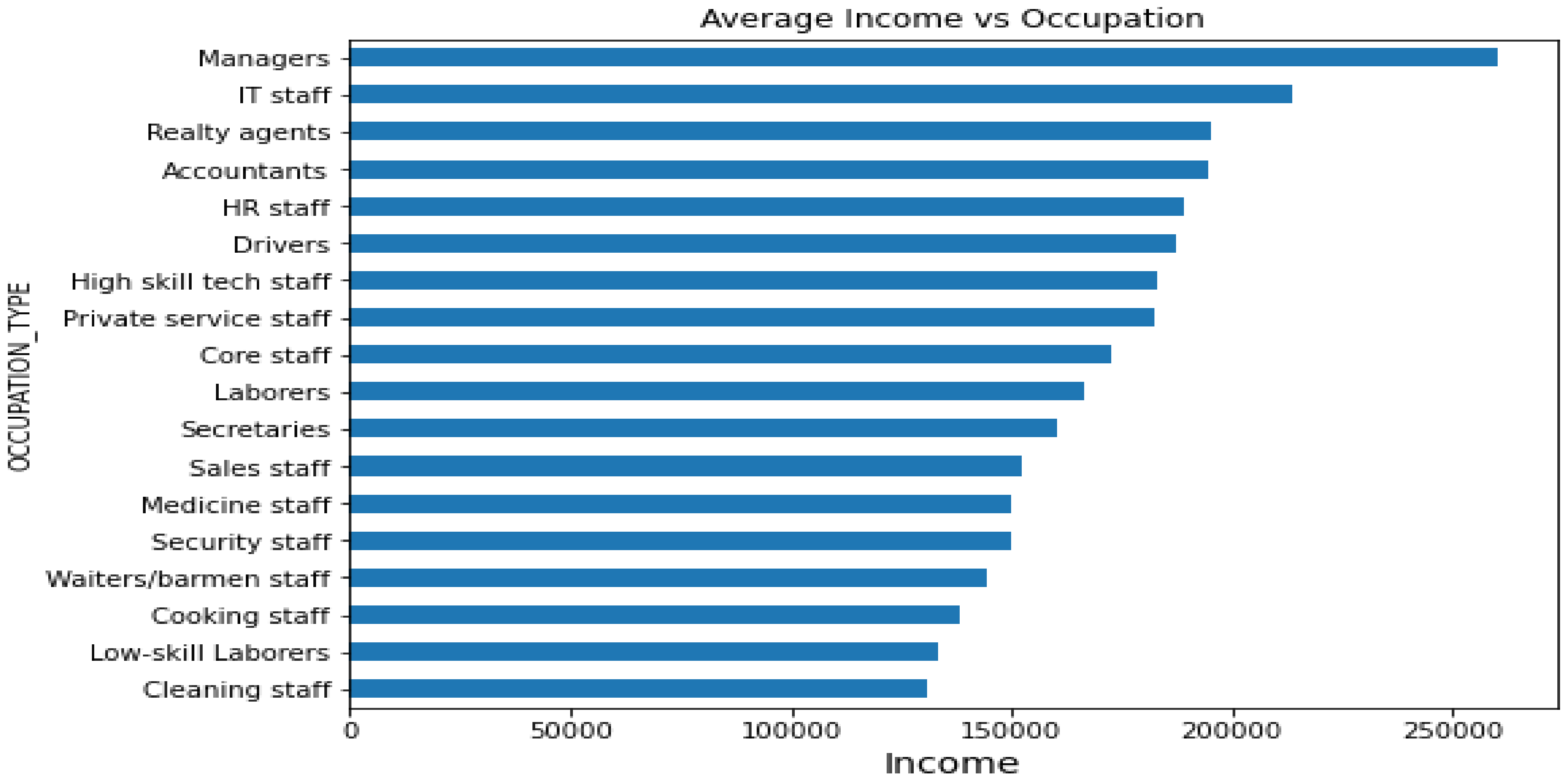
OUTLIERS

- An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.
- Usually an outlier is fixed with mean/median/mode

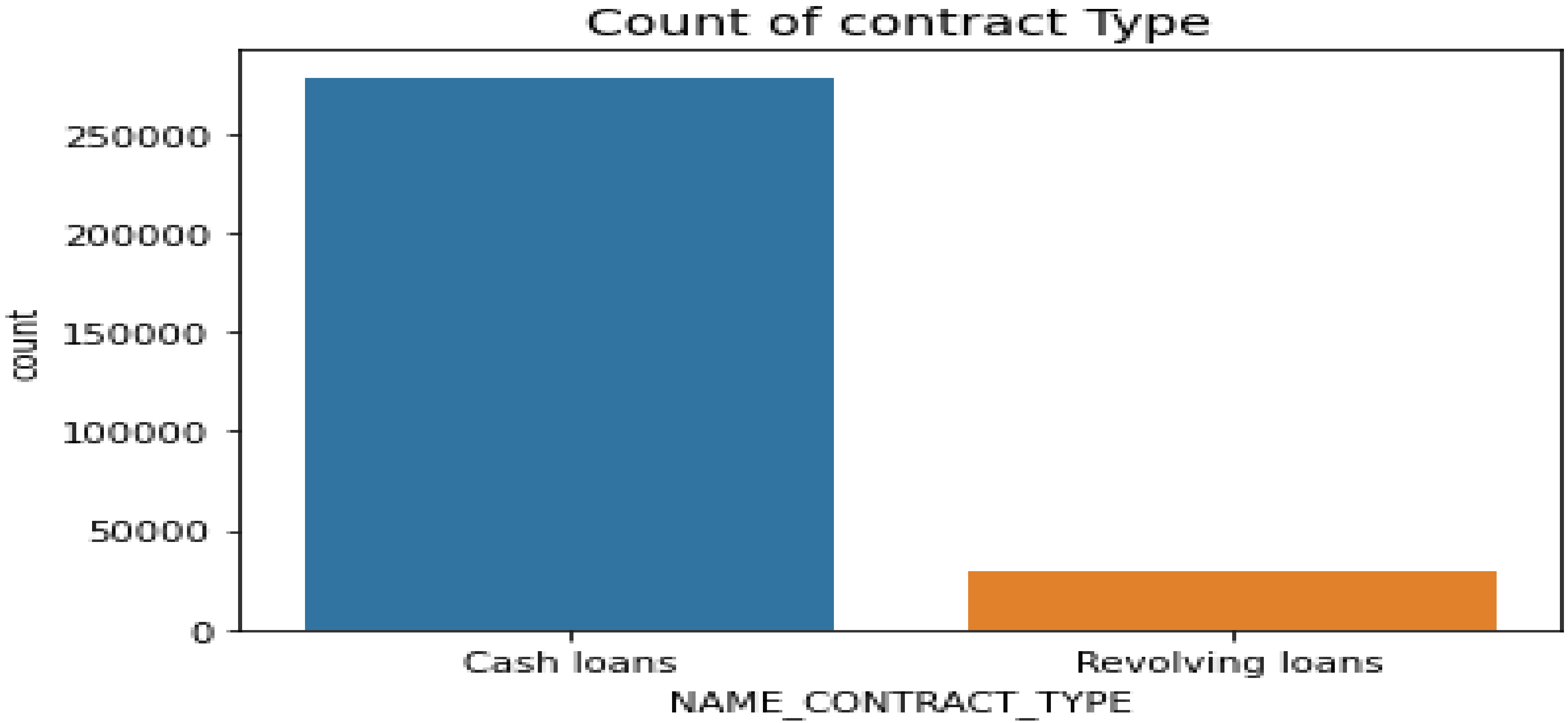
Number of applicants as per occupation



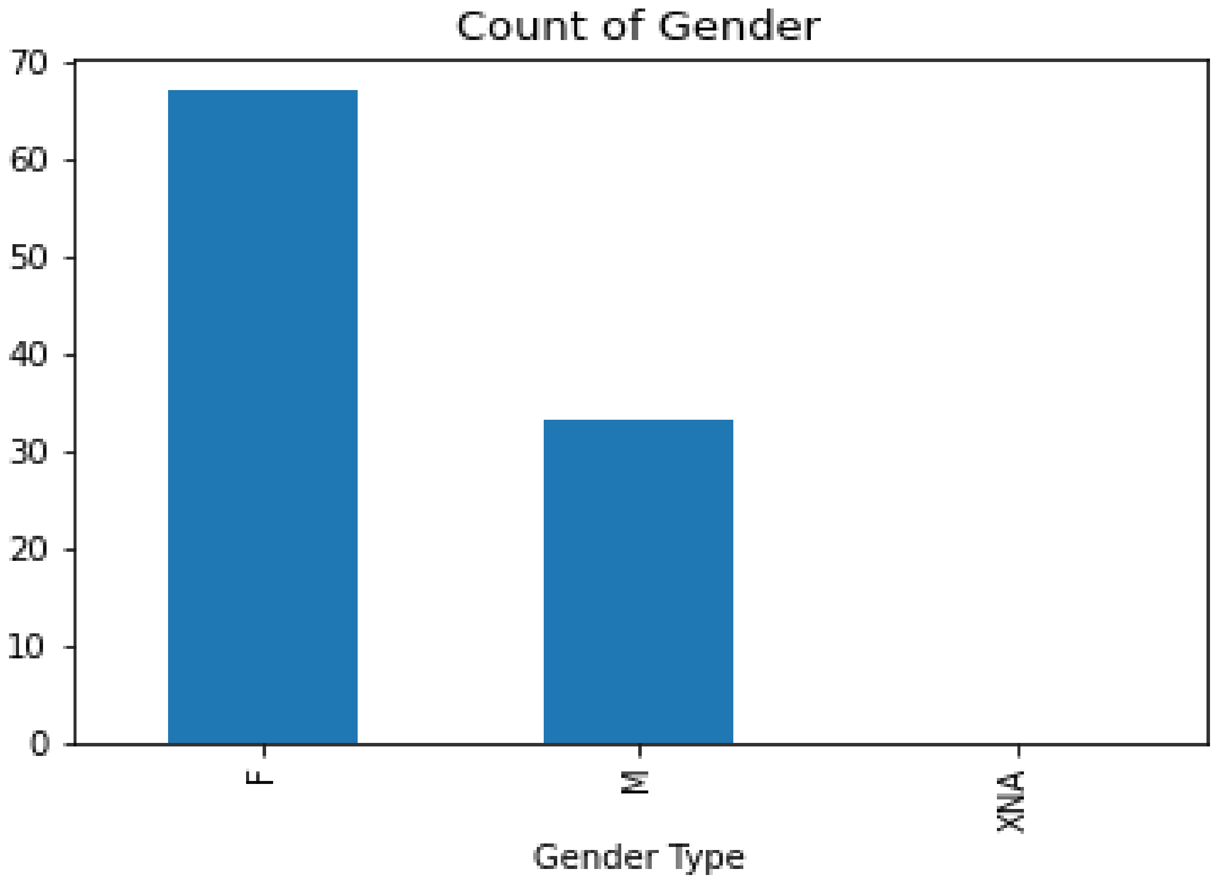
Average income as per occupation



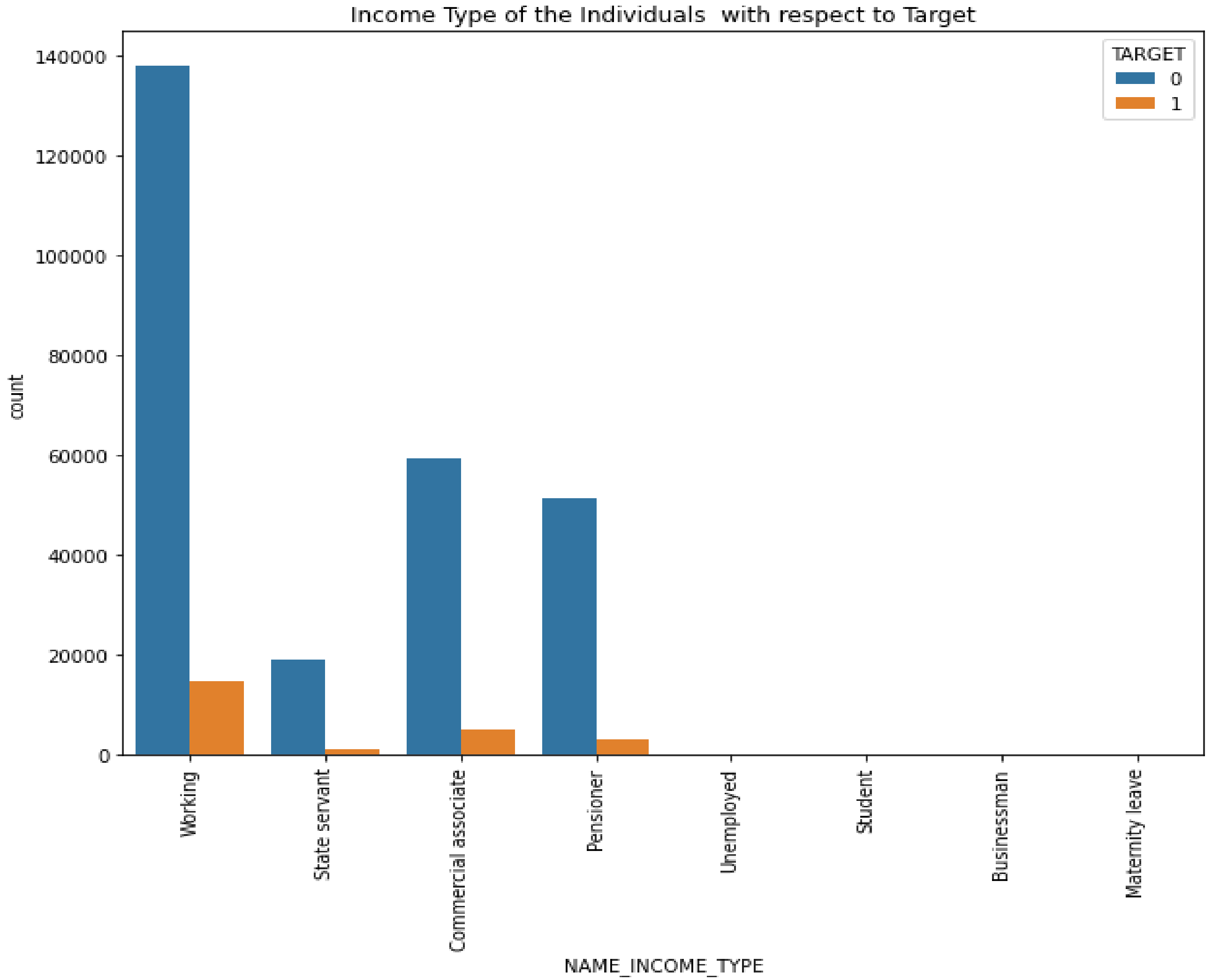
Different types of loans given



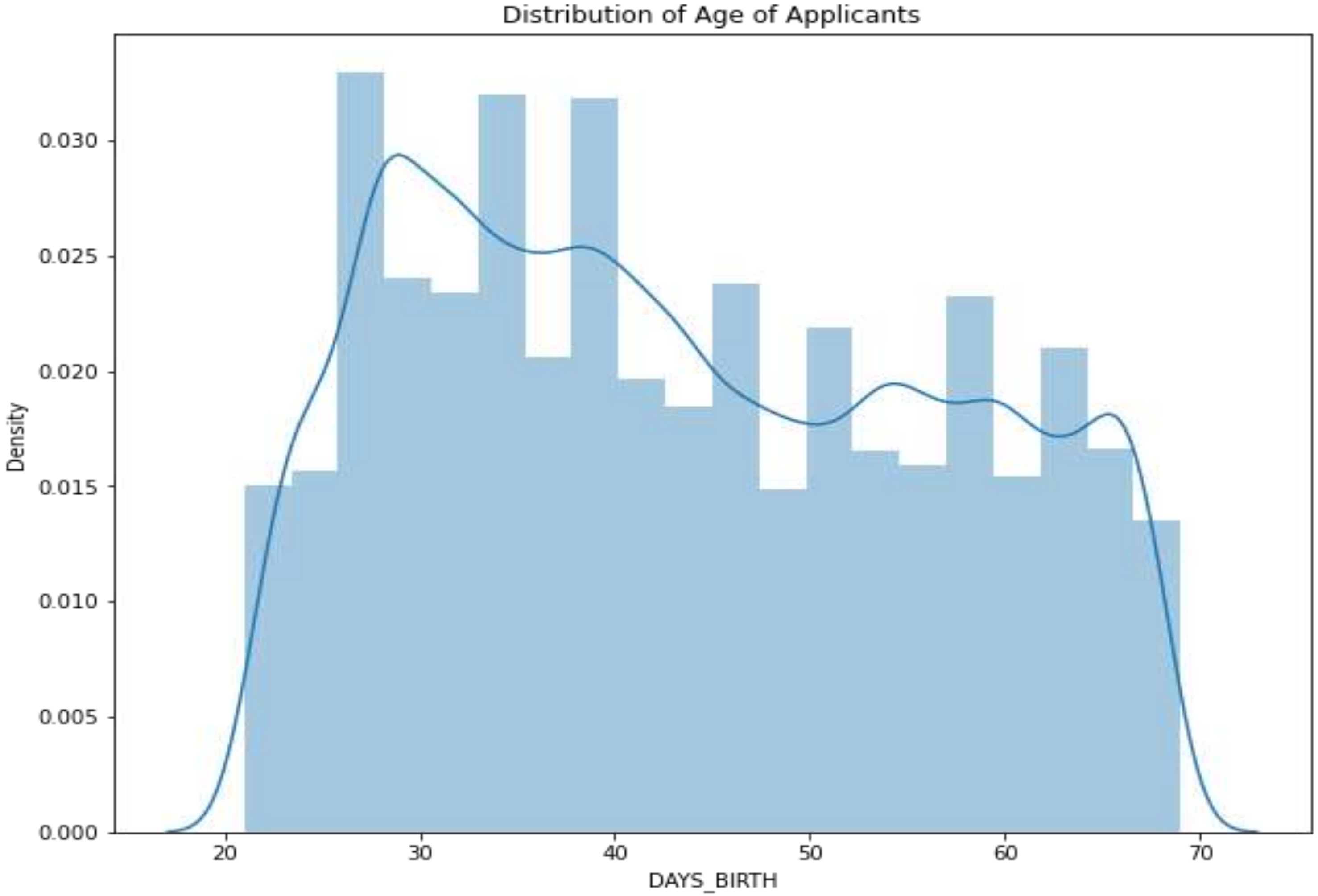
Loan applied as per gender



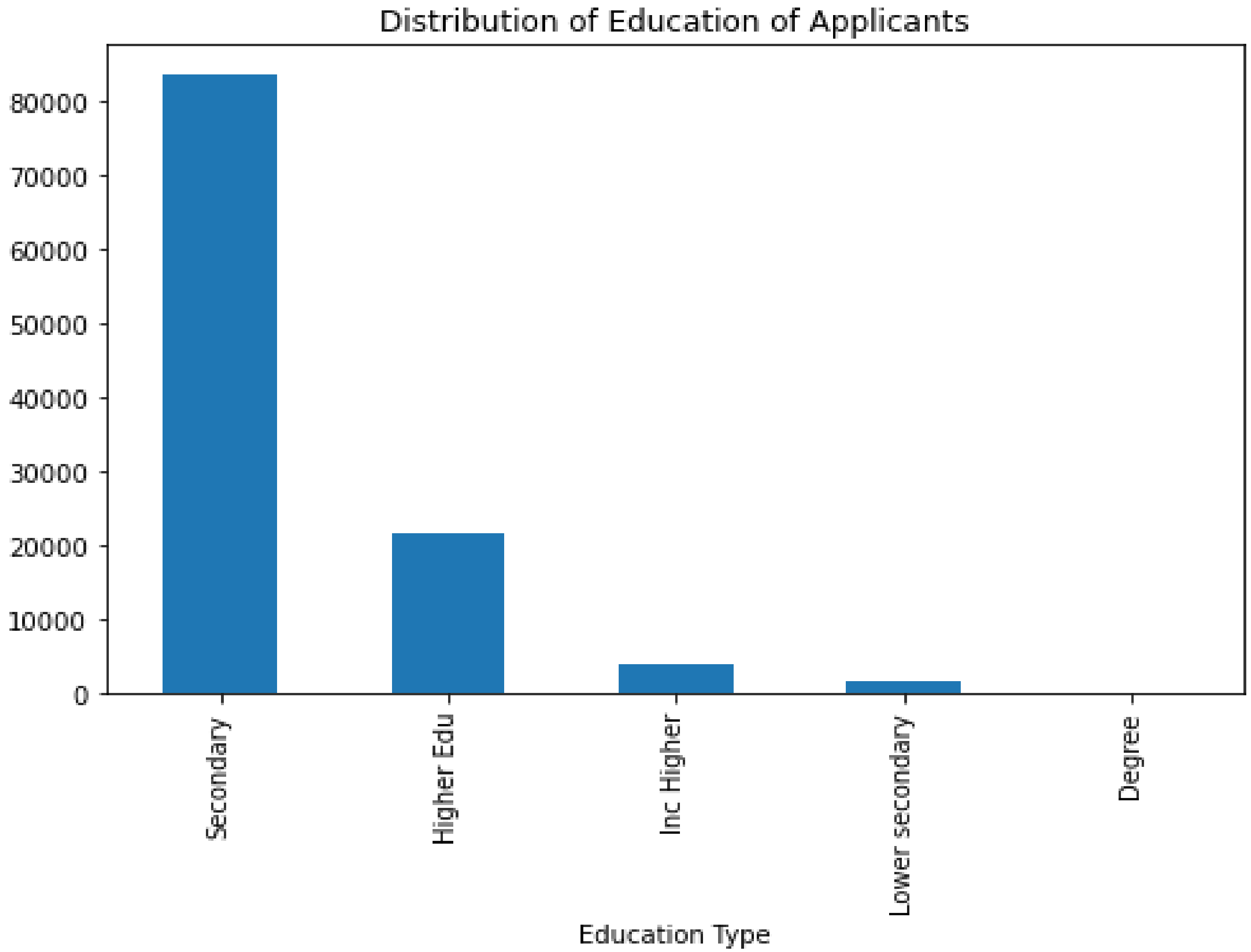
Loan given VS Occupation



Loan given distribution as per age



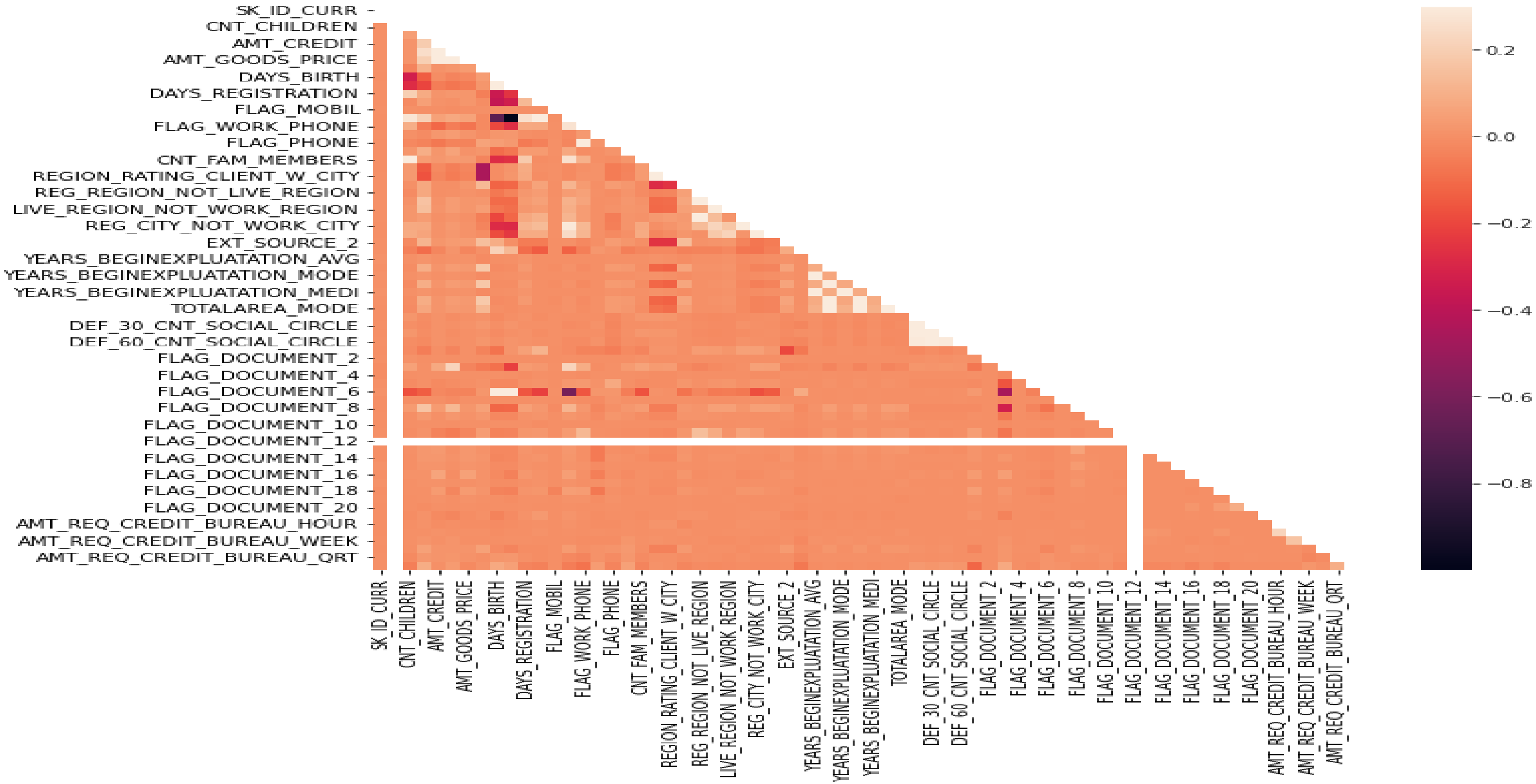
Loan given distribution as per level of education



ANALYSIS

- Number of clients paying on time – 101820
- Number of client with payment difficulties – 8966
- Ratio – 11.35

Correlation between variables



MERGING OF FILES

- File 1 – application_data
- File 2 – previous_application
- SK_ID_CURR is a unique identifier which can be used to merge the 2 files
- After merging both the files, the new file will also have duplicate number of SK_ID_PREV which can be used to figure out if any patter is present by including the cases if a customer has previously taken loan more than once

Conclusion

- A wide variety of people of different class/age/occupation/gender will apply for loan for various reasons
- It is very important to assess the previous data for common applicants to check out whether they lie under the category of good customer or defaulter
- These data and records can help maintain stability and help banks to filter out users whether to give them loan or not
- With the help of such data banks can grow economically by maintaining a good record of users and filtering out defaulters