CASE STUDY ON BANK LOAN DEFAULT

Jyothi Punya Prasad Bale
DS C47

EDA WORKFLOW

- Business Understanding and Problem statement
- Importing necessary libraries into Jupyter notebook
- Loading dataset
- Understanding dataset
- Data Cleaning
- Data Analysis
- Insights

PROBLEM STATEMENT

A consumer finance company which specializes in lending various types of loans to urban customers, wants to analyze the patterns present in the data to 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.

IMPORTING LIBRARIES

The following libraries were used for doing EDA

- Pandas
- NumPy
- Matplotlib
- Seaborn
- Warnings

LOADING DATASET

This dataset has 3 files :

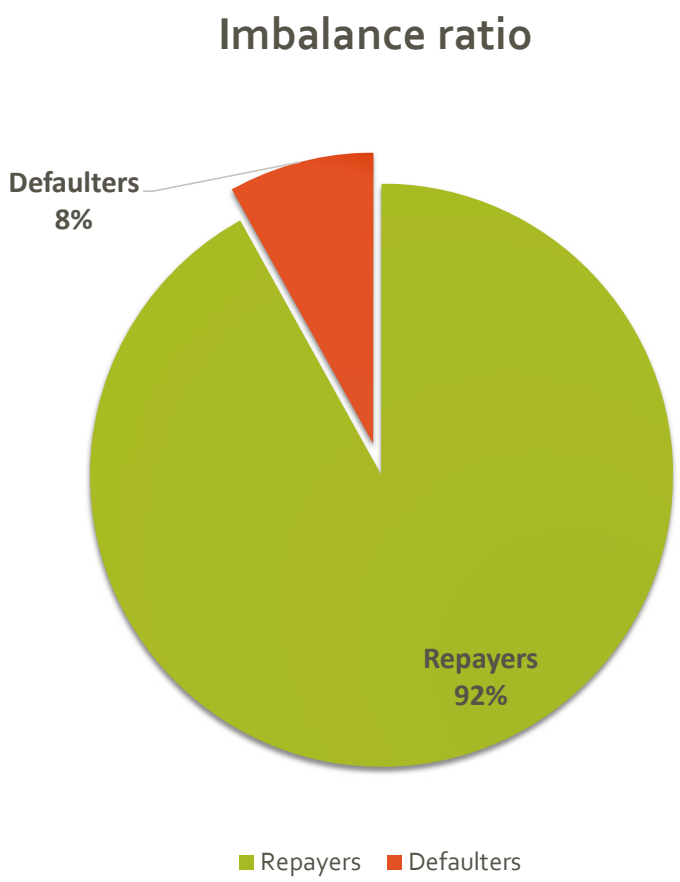
1. *application_data.csv* : contains all the information of the client at the time of application.
   The data is about whether a **client has payment difficulties.**

2. *previous_application.csv:* contains information about the client's previous loan data. It contains the data on whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

3. *columns_description.csv:* data dictionary which describes the meaning of the variables.

UNDERSTANDING DATASET

- Current Applications dataset is highly imbalanced. Imbalance ratio is ~ 8% i.e., 8 out of 100 applicants are defaulting the loans.
- Hence, divided the dataset into defaulters and repayers for analyzing separately.



DATA CLEANING

Data Cleaning involved checking for data types, handling missing values and outliers

- **Handling missing values:**
  1. Dropping columns: Current applications data has 49 columns with null values more than 40%. Hence, dropped these columns
  2. Imputing columns: Imputed categorical columns with mode of the columns. Almost all numerical columns have outliers, hence imputed numerical columns with median of the column
  3. After dropping all the unwanted columns we finally have a total 39 columns for analysis
     - Continuous columns: 20
     - Categorical columns: 19

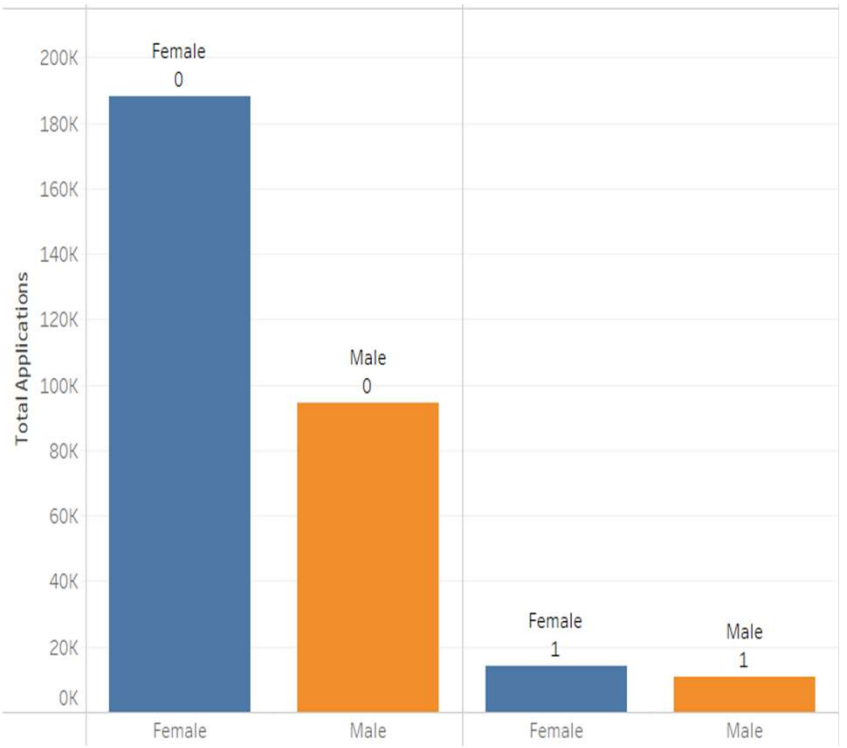
Cont....

- **Handling Outliers:**
  1. **CNT_CHILDREN** has outliers. Count of children more than 3 are oultiers in the data. Max children count is 19
  2. **AMT_INCOME_TOTAL** has outliers. Individuals Income who fall under these outliers has income much more than the rest of the individuals
  3. **AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE** has outliers
  4. **DAYS_EMPLOYED** has a lot of outliers where years of employment is shown as 1000 years
  5. **LAST_PHONE_CHANGE** has an outlier where applicant has not changed phone for more than 10 years
  6. **COUNT_FAMILY_MEMBERS** has outliers where total family members more than 5
- We can impute/remove these outliers but Since it was mentioned that for this exercise, it is not necessary to remove any data points, No changes have been made in the context of outliers

DATA ANALYSIS

GENDER:

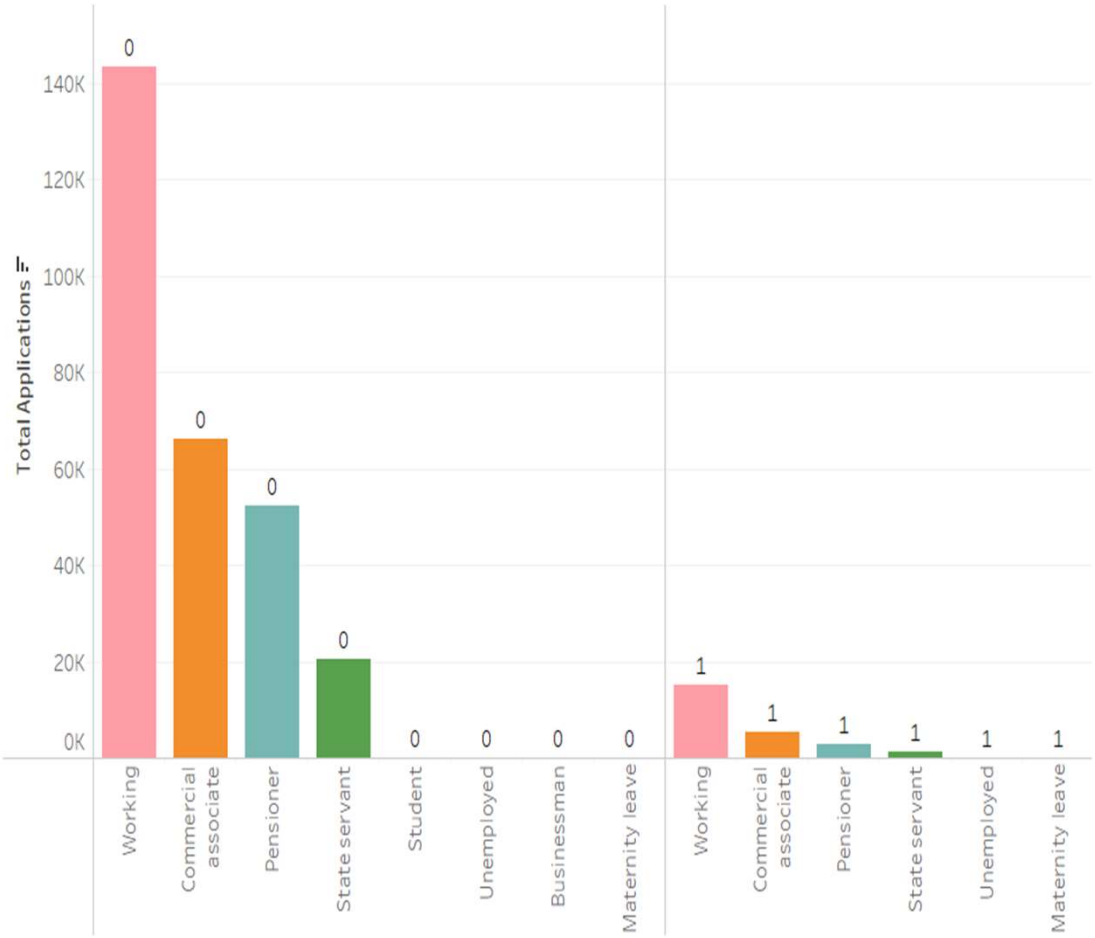
- Majority of the applicants are females with 65.8% applications
- Men are majority of the defaulters



DATA ANALYSIS

INCOME TYPE:

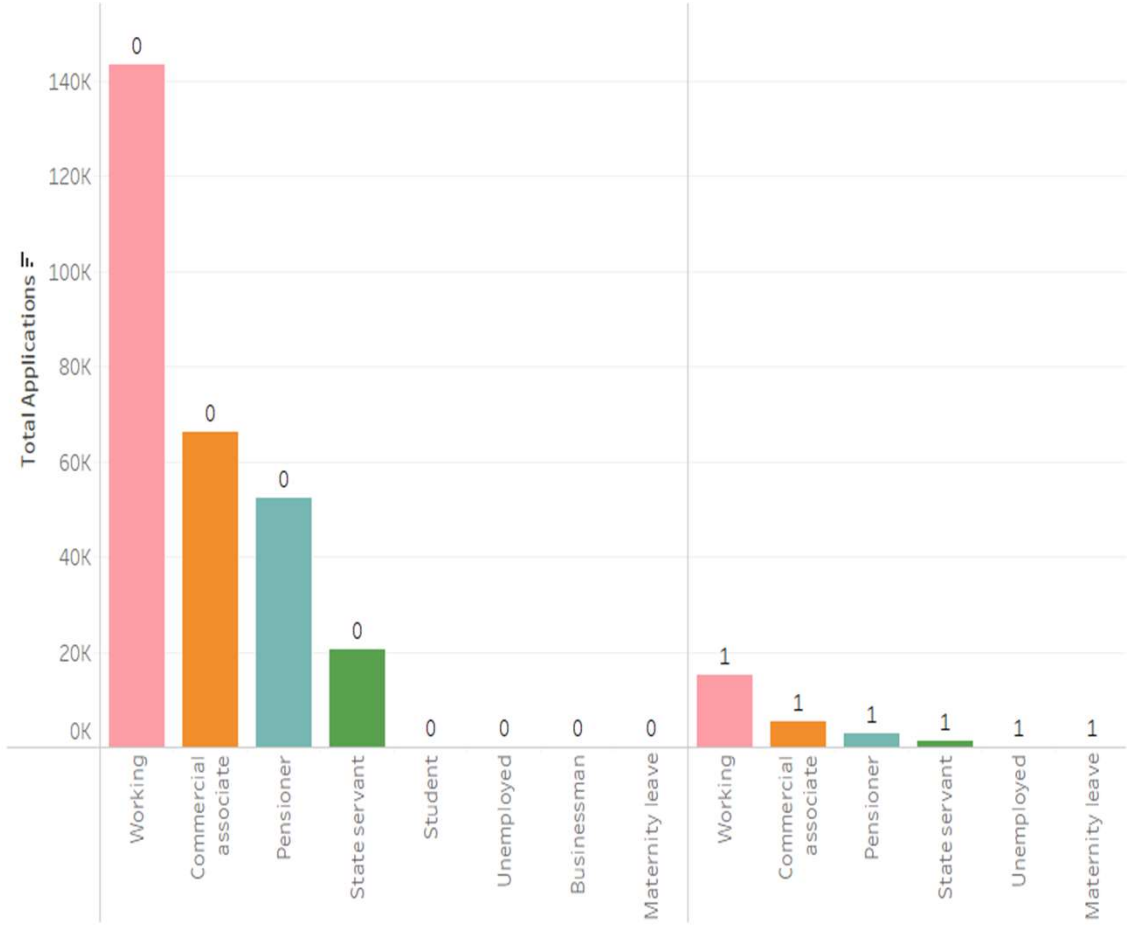
- More than 50% of the applicants are from working class followed by commercial associates and pensioners
- Maternity leave or Unemployed applicants have higher default rate



DATA ANALYSIS

EDUCATION TYPE:

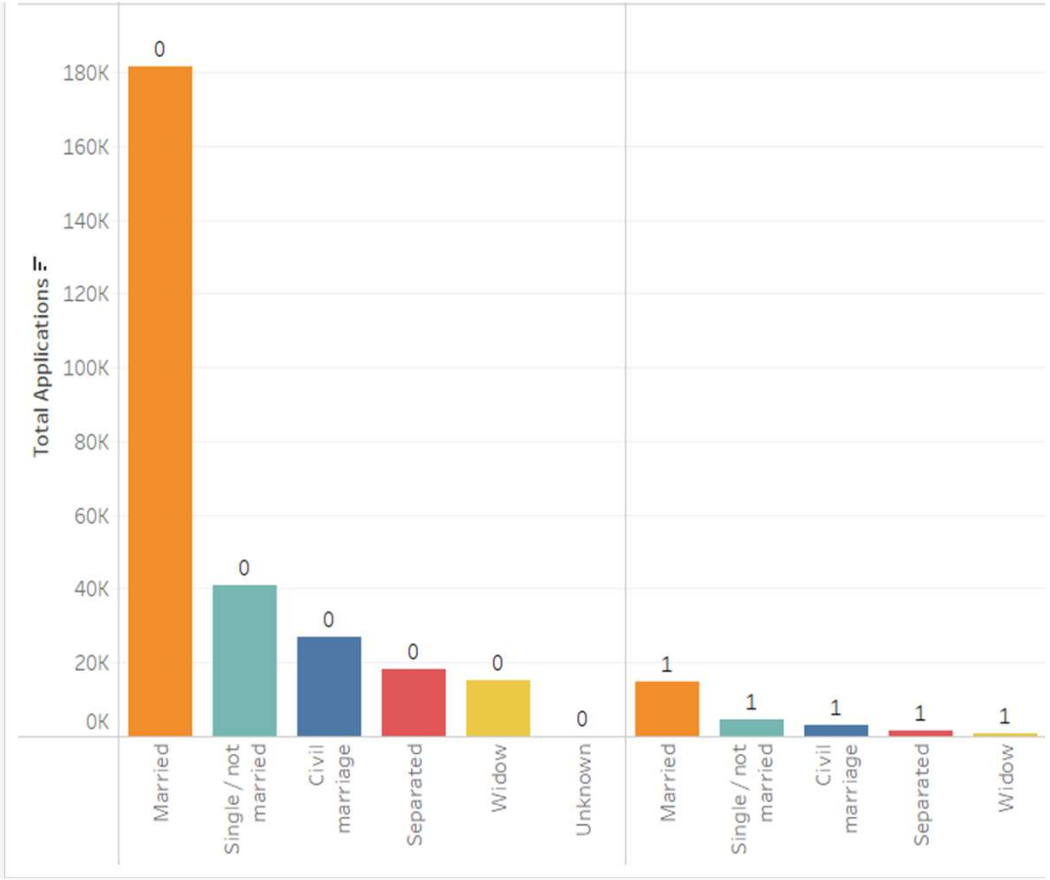
- 70% of the applicants have secondary / secondary special education
- Applicants with lower secondary, secondary or incomplete higher education are defaulting more



DATA ANALYSIS

FAMILY STATUS:

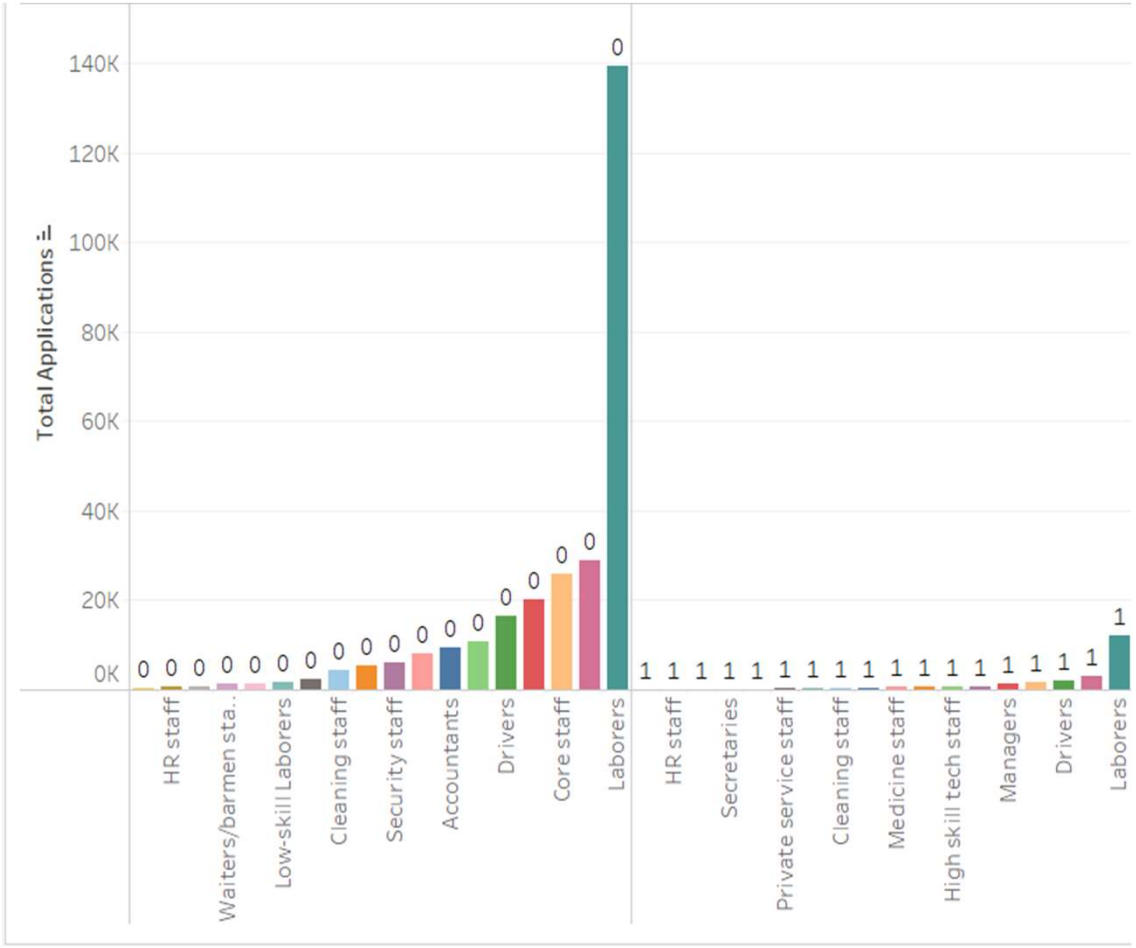
- Married people the most applicants for the Laon with 60% of loans by them
- Applicants who are civil married are single have higher default rate



DATA ANALYSIS

OCCUPATION TYPE:

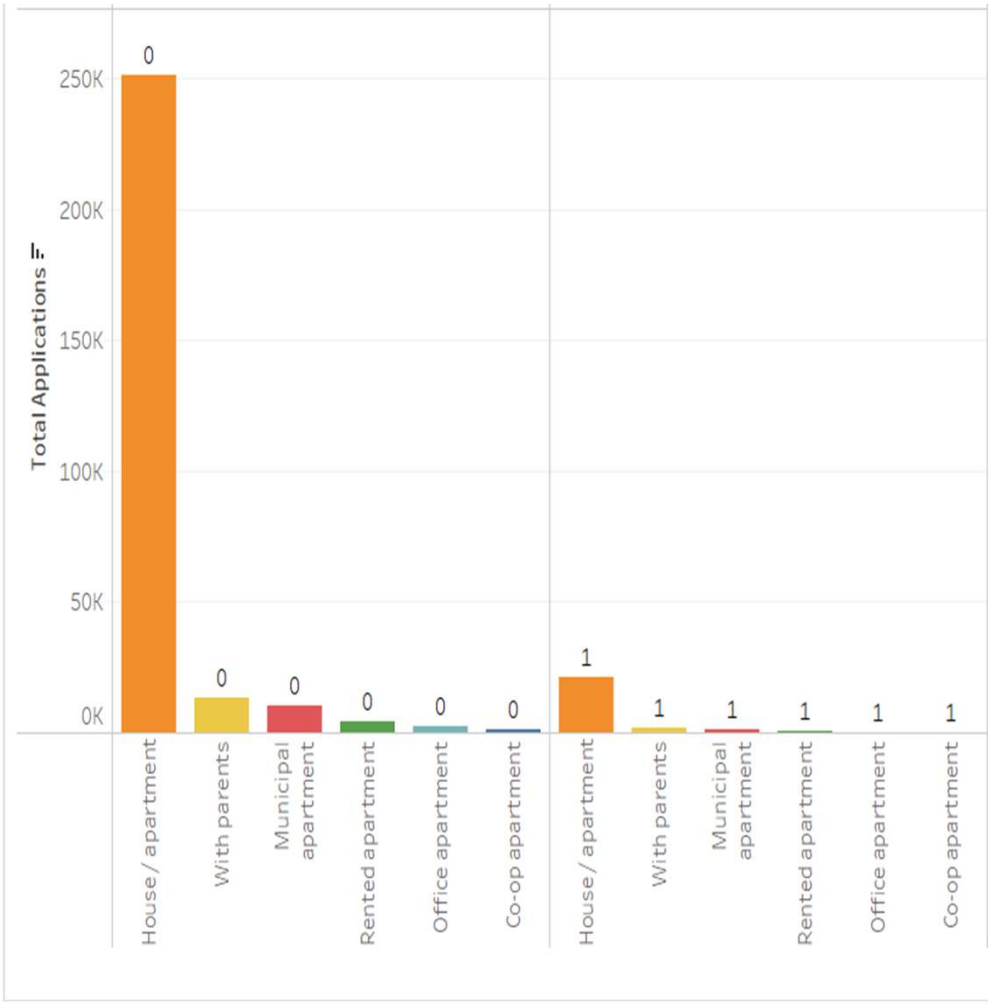
- Majority of the loan applicants are laborers with 50% of applications. Least number of applications received are from IT an HR staff
- Default rate is higer for applicants who are Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff



DATA ANALYSIS

HOUSING TYPE:

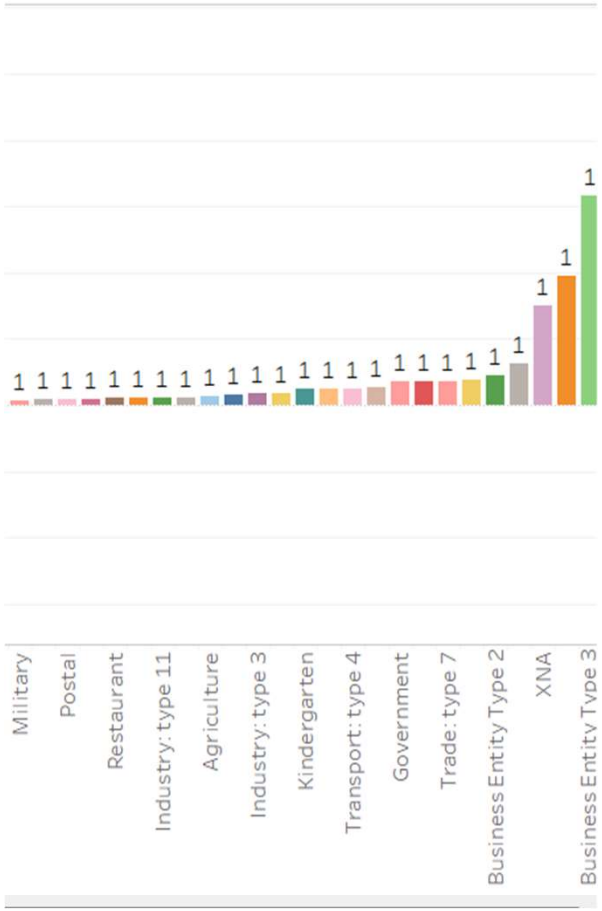
- More than 80% of applicants live in either own house or own apartment
- Applicants who are living in rented apartments, living with parents or in municipal apartment are having high default rates

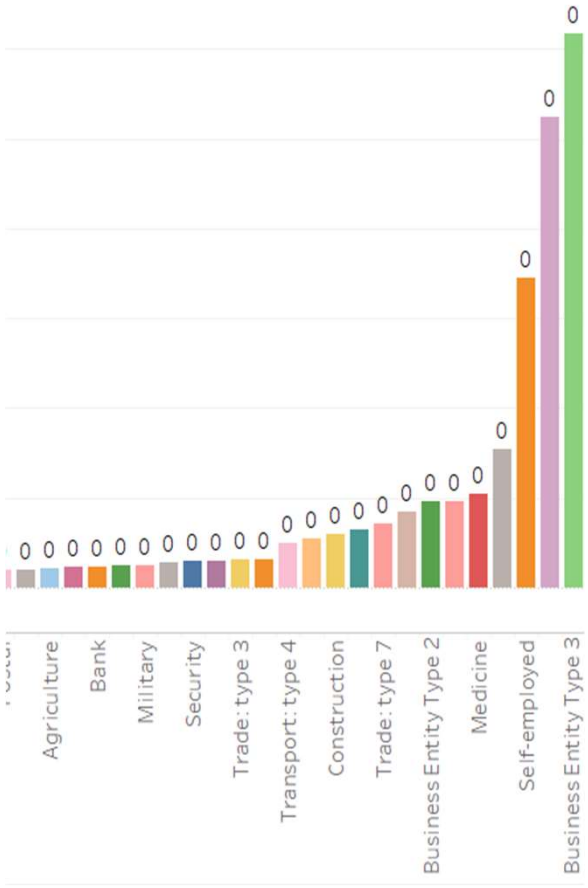


DATA ANALYSIS

ORGANIZATION TYPE:

- People who belongs to Business Entity type 3 are the top 1 applicants for loans
- Organizations with highest percent defaults are Transport: type 3, Industry: type 13, Industry: type 8 and Restaurant. Self-employed people also have relative high default rate

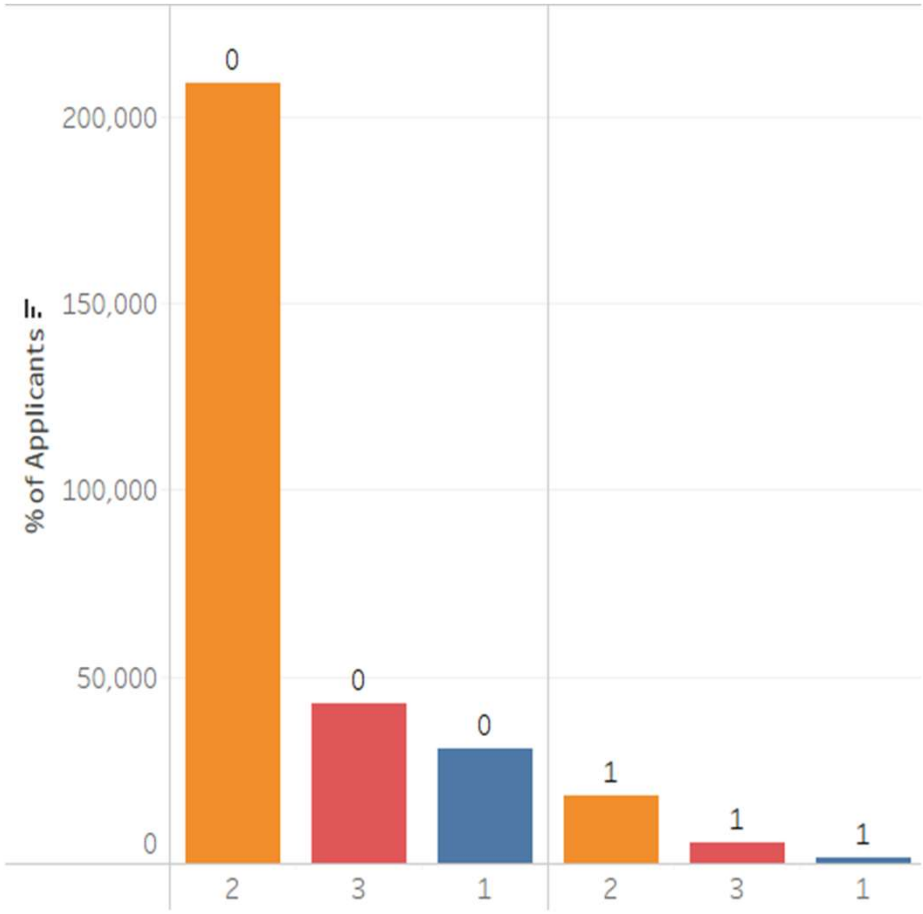




DATA ANALYSIS

REGION RATING:

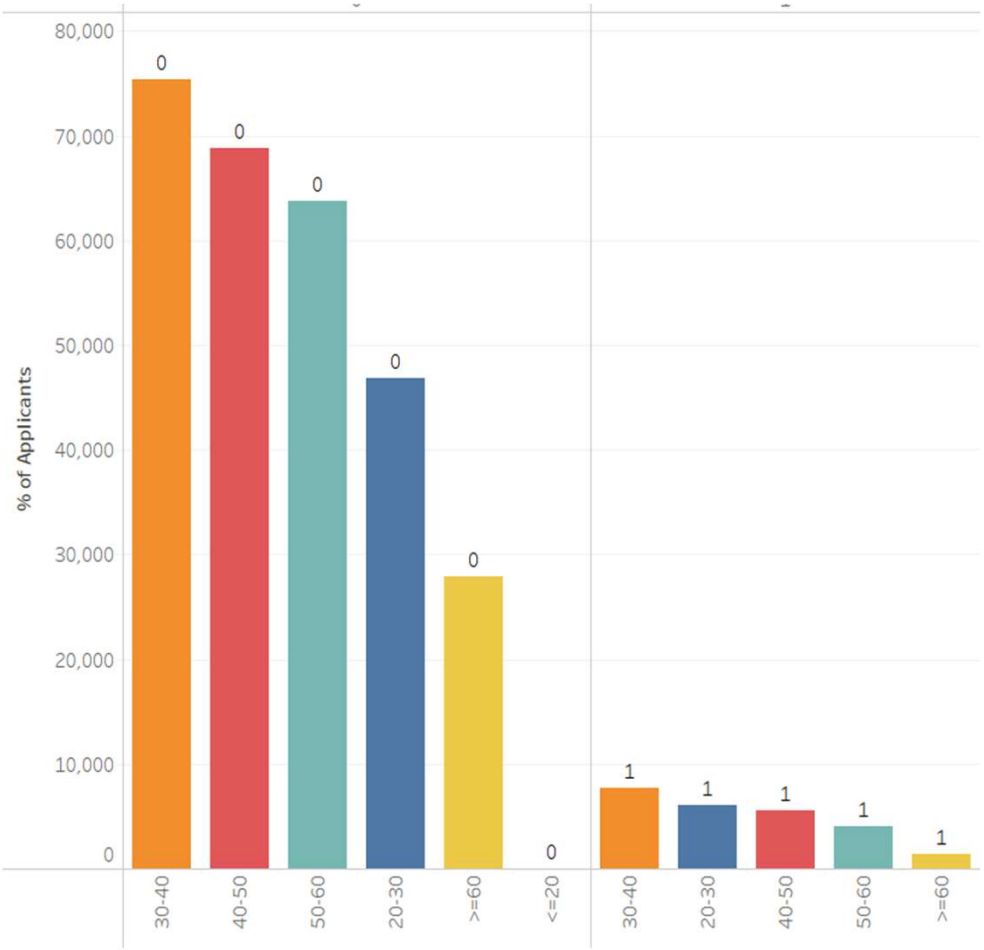
- More than 70% of applicants are from region rating 2
- Applicants coming regions of rating 3 have high default rates



DATA ANALYSIS

AGE GROUP:

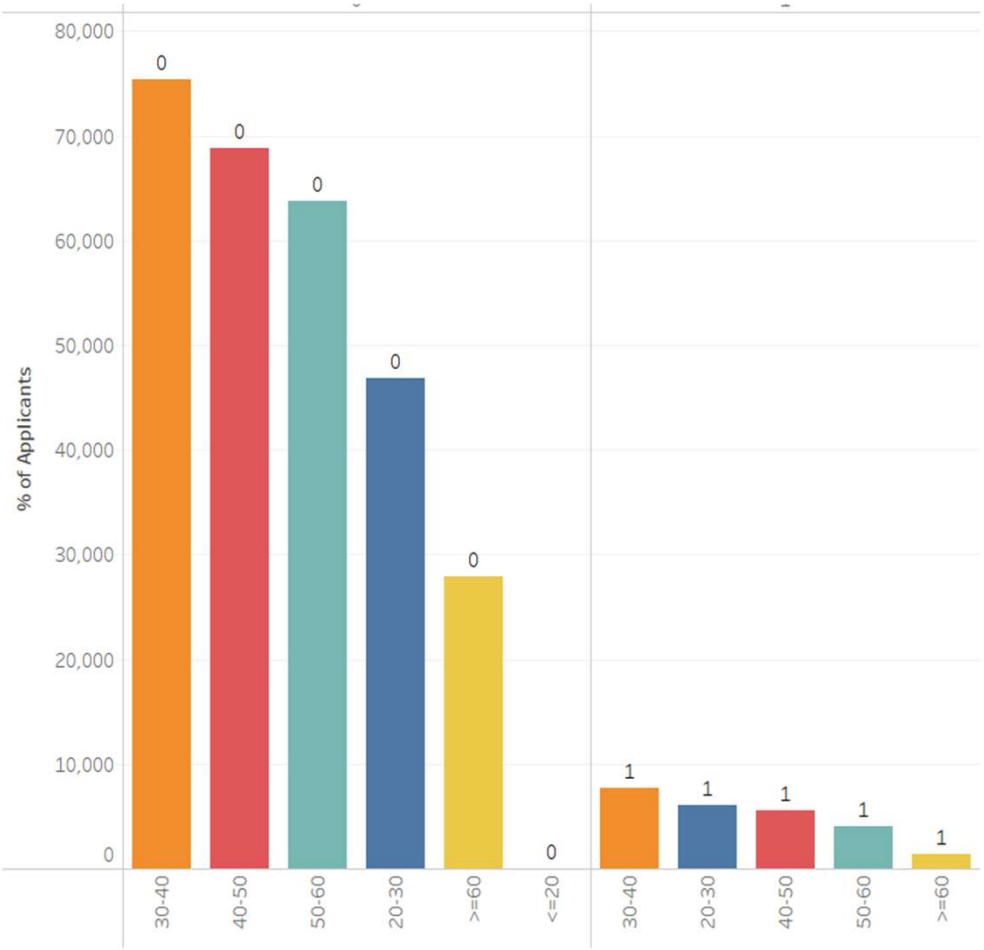
- More than 25% of applicants belongs to 30-40 age group
- People belonging to 20-30 and 30-40 age group tends to default loans more comparatively



DATA ANALYSIS

INCOME RANGE:

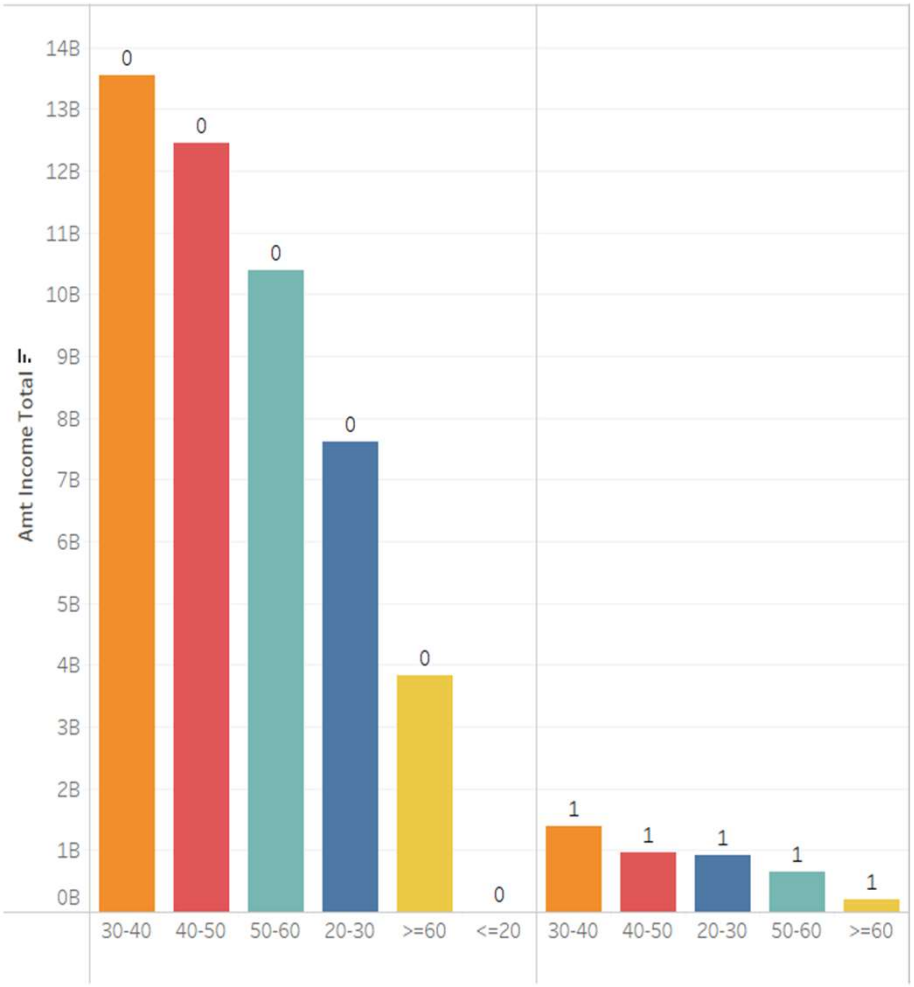
- Over 80% of applicants have an income less than 250,000
- Applicants having income less than 250,000 have more defalut rate



DATA ANALYSIS

AGE GROUP – INCOME - TARGET:

- Applicants of age group 30-40 and 40-50 have the highest income
- People of age group 30-40 and 40 -50 having high income range are defaulting more



CONCLUSION

- Factors for a potential defaulter:
  - Gender: Women are majority of the defaulters with default rate of more than 50%
  - Income type:
    - Of all the income types, Working people are defaulting more with a default rate of more than 60%
    - Students have 0% default rate
  - Education type: More than 80% of defaulters have secondary / secondary special education
  - Family status: Married people are defaulting the loans at 60% default rate
  - Occupation type: Laborers are defaulting the loans more than anyone with a 50% default rate

Cont....

- Housing type: - 80% of the defaulters have their own house / apartment
- Organization type: Business entity type 3 and self employed people default more.
- Age Group: People belonging to 30-40 age group tends to default loans more comparatively
- Region rating : Over 70% of defaulters are coming from region rating 2
- Income range : More than 80% of loan defaults are by the people having income less than 250,000

- People of age group 30-40 and 40 -50 having high income range are defaulting more
- Defaulters have highest education as academic education