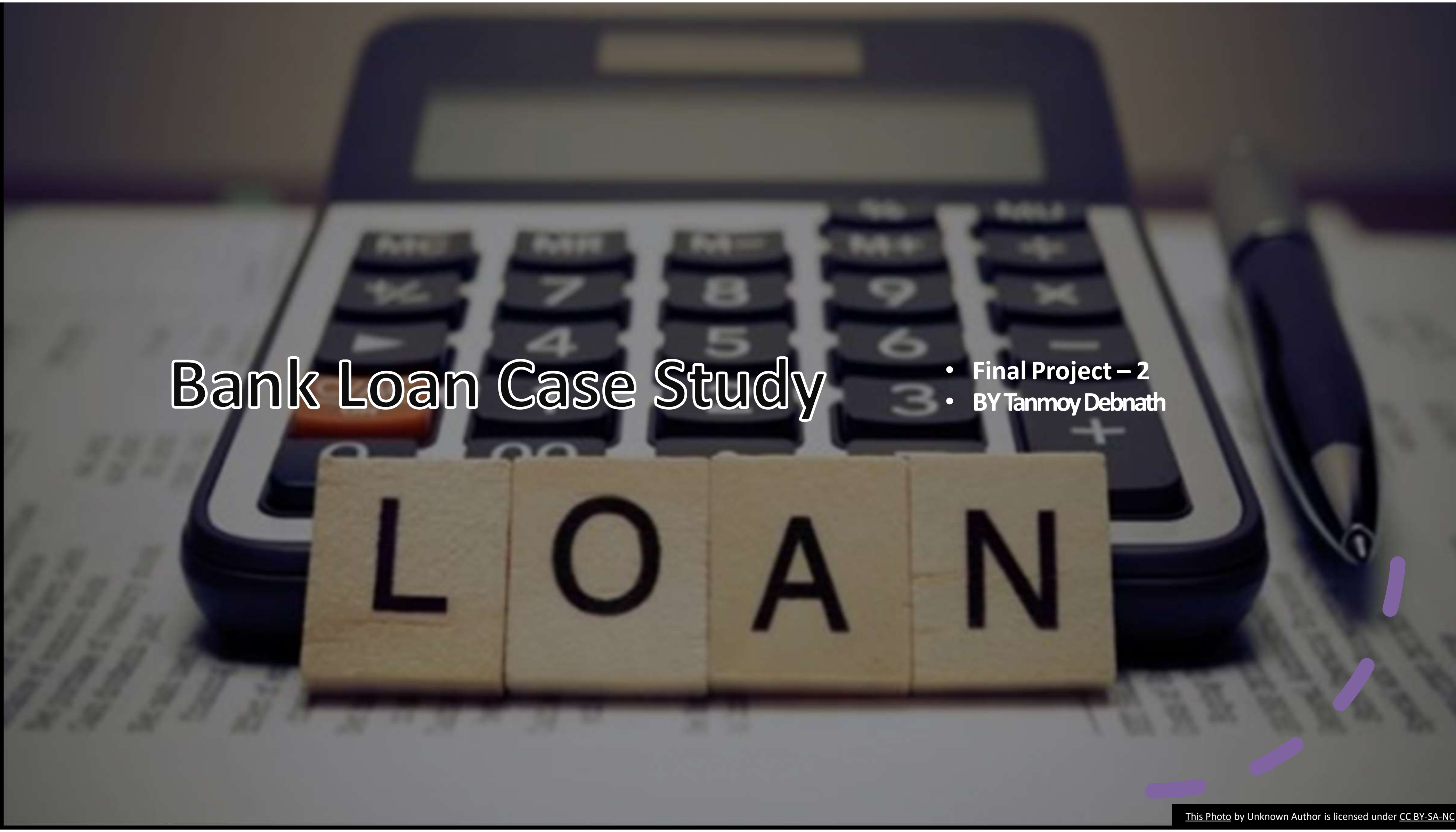



# Bank Loan Case Study

- Final Project – 2
- BY Tanmoy Debnath

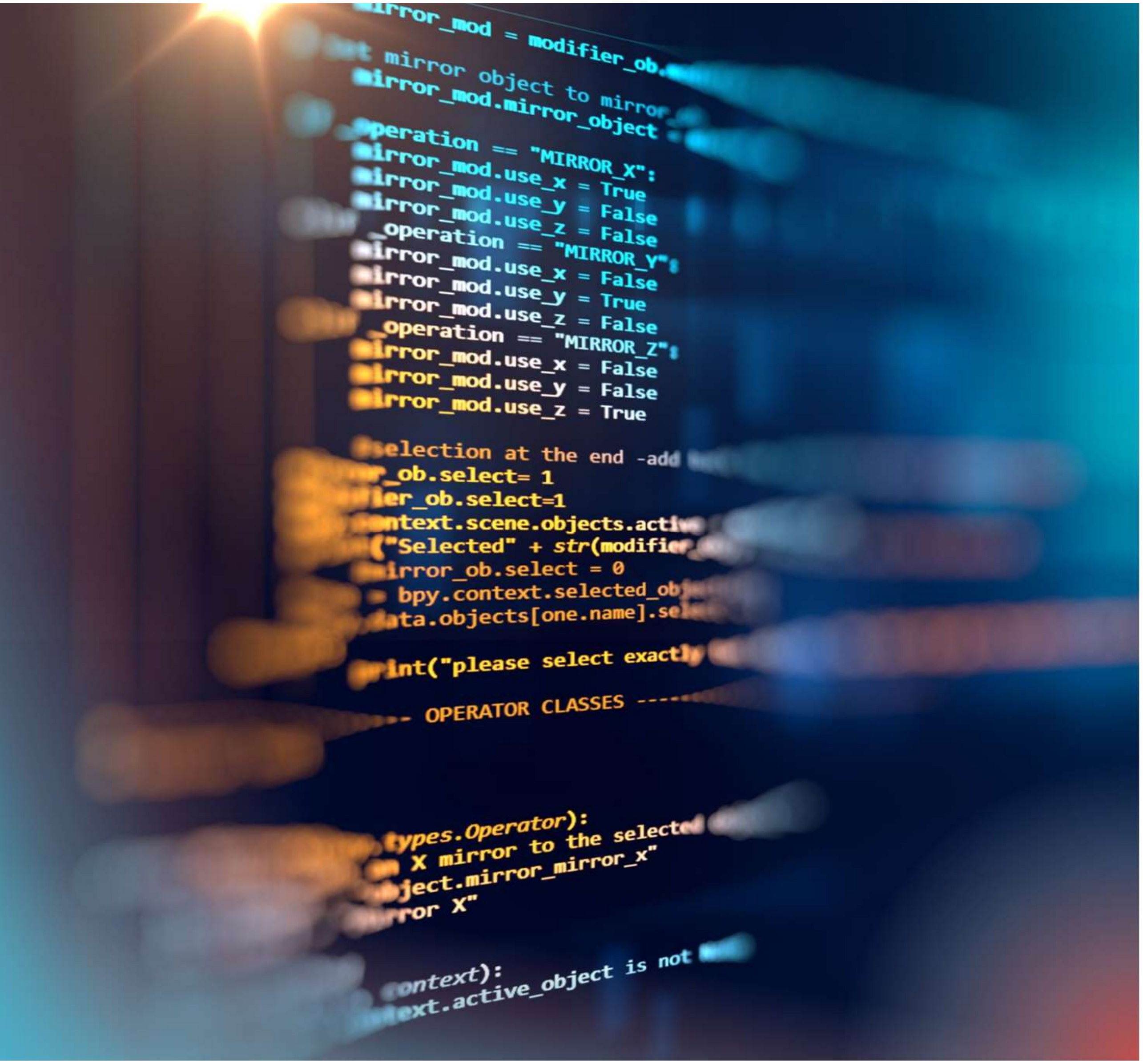
# Project Description:

- This case study attempts to demonstrate the application of EDA in a real-world business environment. In this case study, in addition to using the techniques learned in the EDA module, it will help in gaining a basic grasp of risk analytics in banking and financial services, as well as how data is utilized to reduce the risk of losing money when lending to consumers
- The company wants to understand the driving factors (or driver variables) behind
- loan default, i.e. the variables which are strong indicators of default.

# APPROACH

This case study has two enormous data sets: the current application and the previous application. Each included several unneeded columns that would be useless for risk assessments, as well as many blank data. So, first step is cleaning the data.

To evaluate his enormous set of data, I first cleaned the data, located some outliers and deleted them, and then began performing univariate and bivariate analysis using pivot tables and charts.




# TECH-STACK USED

- Software And The Version Used While Making The Project :
  - **MS Excel** (For working, analysing and reporting insights)
  - **Microsoft Power Point** (For presenting the detailed analysis)

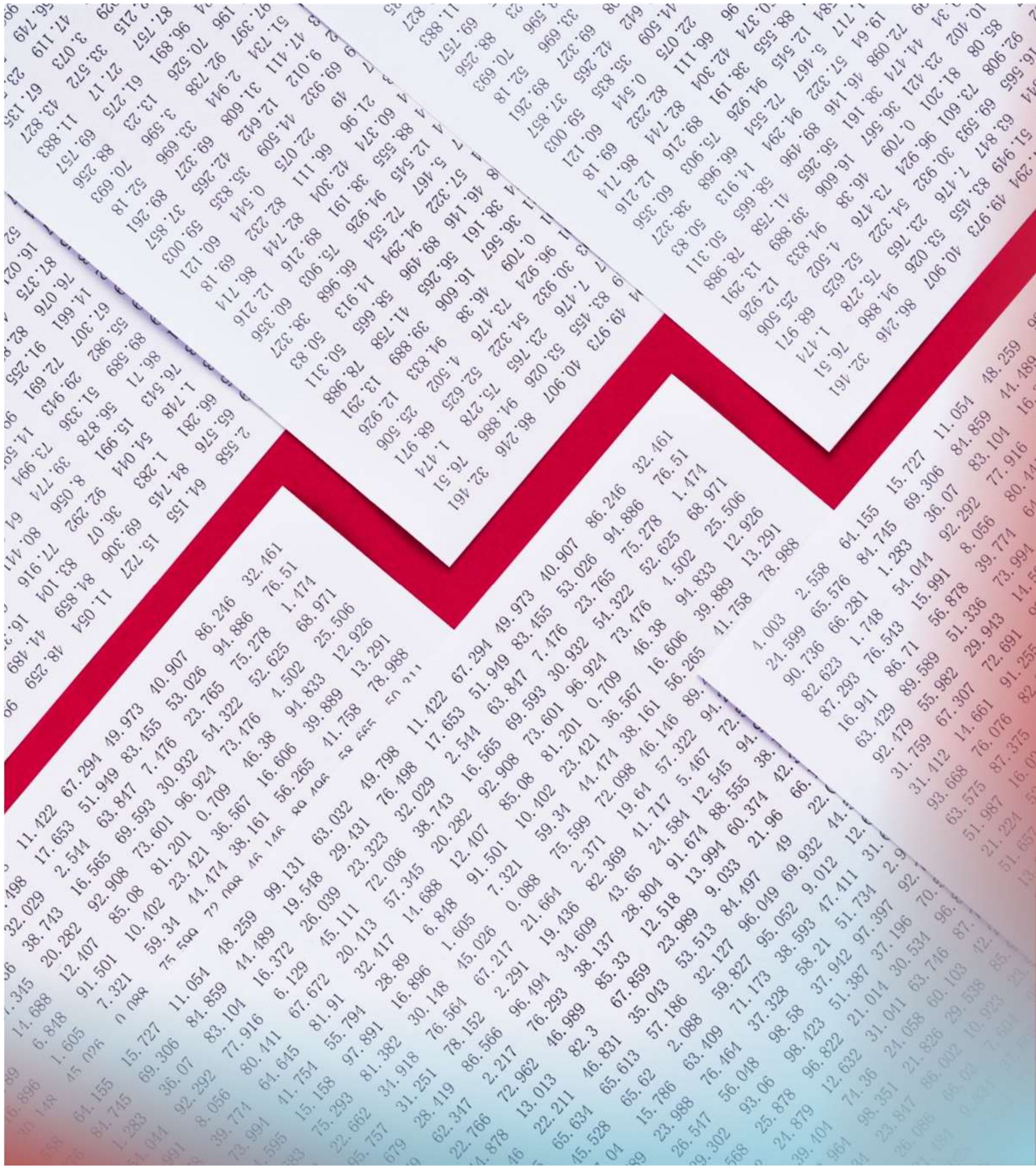
# Data Understanding:

- **`application_data.csv`** contains all the information of the client at the time of application.
- The data is about wheather a client has payment difficulties.
- **`previous_application.csv`** contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.
- **`columns_descrption.csv`** is data dictionary which describes the meaning of the variables.

**Task:**

Present the overall approach of the analysis. Mention the problem statement and the analysis approach briefly

- Both the CSV files will be checked for any unnecessary data and unwanted columns/rows and will be cleaned/removed if necessary. Then they will be checked for outliers, if any, to find if there is skewness in the given columns which would affect the final visualization and insight. Data Imbalance will be checked. Different types of analysis will be done to understand the relationships between different variable to find the Driving Factors. Different visualizations will be observed to understand the relationships



# AFTER CLEANING THE TABLES

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | | Table | Row | Description | Special | | | |
| 2 | 1 | application_data | SK_ID_CURR | ID of loan in our sample | | | | |
| 3 | 2 | application_data | TARGET | Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases | | | | |
| 4 | 5 | application_data | NAME_CONTRACT_TYPE | Identification if loan is cash or revolving | | | | |
| 5 | 6 | application_data | CODE_GENDER | Gender of the client | | | | |
| 6 | 7 | application_data | FLAG_OWN_CAR | Flag if the client owns a car | | | | |
| 7 | 8 | application_data | FLAG_OWN_REALTY | Flag if client owns a house or flat | | | | |
| 8 | 9 | application_data | CNT_CHILDREN | Number of children the client has | | | | |
| 9 | 10 | application_data | AMT_INCOME_TOTAL | Income of the client | | | | |
| 10 | 11 | application_data | AMT_CREDIT | Credit amount of the loan | | | | |
| 11 | 12 | application_data | AMT_ANNUITY | Loan annuity | | | | |
| 12 | 13 | application_data | AMT_GOODS_PRICE | For consumer loans it is the price of the goods for which the loan is given | | | | |
| 13 | 14 | application_data | NAME_TYPE_SUITE | Who was accompanying client when he was applying for the loan | | | | |
| 14 | 15 | application_data | NAME_INCOME_TYPE | Clients income type (businessman, working, maternity leave,...) | | | | |
| 15 | 16 | application_data | NAME_EDUCATION_TYPE | Level of highest education the client achieved | | | | |
| 16 | 17 | application_data | NAME_FAMILY_STATUS | Family status of the client | | | | |
| 17 | 18 | application_data | NAME_HOUSING_TYPE | What is the housing situation of the client (renting, living with parents, ...) | | | | |
| 18 | 19 | application_data | REGION_POPULATION_RELATIVE | Normalized population of region where client lives (higher number means the client lives in more populated region | normalized | | | |
| 19 | 20 | application_data | DAYS_BIRTH | Client's age in days at the time of application | time only relative to the application | | | |
| 20 | 21 | application_data | DAYS_EMPLOYED | How many days before the application the person started current employment | time only relative to the application | | | |
| 21 | 22 | application_data | DAYS_REGISTRATION | How many days before the application did client change his registration | time only relative to the application | | | |
| 22 | 23 | application_data | DAYS_ID_PUBLISH | How many days before the application did client change the identity document with which he applied for the loan | time only relative to the application | | | |
| 23 | 24 | application_data | OWN_CAR_AGE | Age of client's car | | | | |
| 24 | 25 | application_data | FLAG_MOBIL | Did client provide mobile phone (1=YES, 0=NO) | | | | |
| 25 | 26 | application_data | FLAG_EMP_PHONE | Did client provide work phone (1=YES, 0=NO) | | | | |
| 26 | 27 | application_data | FLAG_WORK_PHONE | Did client provide home phone (1=YES, 0=NO) | | | | |
| 27 | 28 | application_data | FLAG_CONT_MOBILE | Was mobile phone reachable (1=YES, 0=NO) | | | | |

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | SK_ID_PREV | SK_ID_CURR | NAME_CONTRACT_TYPE | AMT_ANNUITY | AMT_APPLICATION | AMT_CREDIT | AMT_DOWN_PAYMENT | AMT_GOODS_PRICE | WEEKDAY_APPR_PROCESS_START | HOUR_APPR_PROCESS_START | FLAG_LAST_APPL_PER_CON |
| 2 | 2030495 | 271877 | Consumer loans | 1730.43 | 17145 | 17145 | 0 | 17145 | SATURDAY | 15 | Y |
| 3 | 2802425 | 108129 | Cash loans | 25188.615 | 607500 | 679671 | | 607500 | THURSDAY | 11 | Y |
| 4 | 2523466 | 122040 | Cash loans | 15060.735 | 112500 | 136444.5 | | 112500 | TUESDAY | 11 | Y |
| 5 | 2819243 | 176158 | Cash loans | 47041.335 | 450000 | 470790 | | 450000 | MONDAY | 7 | Y |
| 6 | 1784265 | 202054 | Cash loans | 31924.395 | 337500 | 404055 | | 337500 | THURSDAY | 9 | Y |
| 7 | 1383531 | 199383 | Cash loans | 23703.93 | 315000 | 340573.5 | | 315000 | SATURDAY | 8 | Y |
| 8 | 2315218 | 175704 | Cash loans | | 0 | 0 | | | TUESDAY | 11 | Y |
| 9 | 1656711 | 296299 | Cash loans | | 0 | 0 | | | MONDAY | 7 | Y |
| 10 | 2367563 | 342292 | Cash loans | | 0 | 0 | | | MONDAY | 15 | Y |
| 11 | 2579447 | 334349 | Cash loans | | 0 | 0 | | | SATURDAY | 15 | Y |
| 12 | 1715995 | 447712 | Cash loans | 11368.62 | 270000 | 335754 | | 270000 | FRIDAY | 7 | Y |
| 13 | 2257824 | 161140 | Cash loans | 13832.775 | 211500 | 246397.5 | | 211500 | FRIDAY | 10 | Y |
| 14 | 2330894 | 258628 | Cash loans | 12165.21 | 148500 | 174361.5 | | 148500 | TUESDAY | 15 | Y |
| 15 | 1397919 | 321676 | Consumer loans | 7654.86 | 53779.5 | 57564 | 0 | 53779.5 | SUNDAY | 15 | Y |
| 16 | 2273188 | 270658 | Consumer loans | 9644.22 | 26550 | 27252 | 0 | 26550 | SATURDAY | 10 | Y |
| 17 | 1232483 | 151612 | Consumer loans | 21307.455 | 126490.5 | 119853 | 12649.5 | 126490.5 | TUESDAY | 7 | Y |
| 18 | 2163253 | 154602 | Consumer loans | 4187.34 | 26955 | 27297 | 1350 | 26955 | SATURDAY | 12 | Y |
| 19 | 1285768 | 142748 | Revolving loans | 9000 | 180000 | 180000 | | 180000 | FRIDAY | 13 | Y |
| 20 | 2393109 | 396305 | Cash loans | 10181.7 | 180000 | 180000 | | 180000 | THURSDAY | 14 | Y |
| 21 | 1173070 | 199178 | Cash loans | 4666.5 | 45000 | 49455 | | 45000 | SATURDAY | 16 | Y |
| 22 | 1506815 | 166490 | Cash loans | 25454.025 | 450000 | 491580 | | 450000 | MONDAY | 6 | Y |
| 23 | 1182516 | 267782 | Cash loans | 20361.6 | 405000 | 451777.5 | | 405000 | SATURDAY | 4 | Y |
| 24 | 1172842 | 302212 | Cash loans | | 0 | 0 | | | TUESDAY | 9 | Y |
| 25 | 1172937 | 302212 | Cash loans | 39475.305 | 1129500 | 1277104.5 | | 1129500 | THURSDAY | 5 | Y |
| 26 | 1555330 | 199353 | Cash loans | | 0 | 0 | | | SATURDAY | 6 | Y |
| 27 | 1543131 | 275707 | Cash loans | 22619.52 | 229500 | 241920 | | 229500 | THURSDAY | 8 | Y |

previous_application

| | SK_ID_CURR | TARGET | NAME_CONTRACT_TYPE | CODE_GENDER | FLAG_OWN_CAR | FLAG_OW | CNT_CHILDREN | AMT_INCOME_TOTAL | AMT_CREDIT | AMT_ANNUITY | AMT_GOODS_PRICE | NAME_TYPE_SUITE | NAME_INCOME_TYPE | NAME_ED | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 100002 | 1 | Cash loans | M | N | Y | 0 | 202500 | 406597.5 | 24700.5 | 351000 | Unaccompanied | Working | Secondary | S |
| 3 | 100003 | 0 | Cash loans | F | N | N | 0 | 270000 | 1293502.5 | 35698.5 | 1129500 | Family | State servant | Higher edu | M |
| 4 | 100004 | 0 | Revolving loans | M | Y | Y | 0 | 67500 | 135000 | 6750 | 135000 | Unaccompanied | Working | Secondary | S |
| 5 | 100006 | 0 | Cash loans | F | N | Y | 0 | 135000 | 312682.5 | 29686.5 | 297000 | Unaccompanied | Working | Secondary | C |
| 6 | 100007 | 0 | Cash loans | M | N | Y | 0 | 121500 | 513000 | 21865.5 | 513000 | Unaccompanied | Working | Secondary | S |
| 7 | 100008 | 0 | Cash loans | M | N | Y | 0 | 99000 | 490495.5 | 27517.5 | 454500 | Spouse, partner | State servant | Secondary | M |
| 8 | 100009 | 0 | Cash loans | F | Y | Y | 1 | 171000 | 1560726 | 41301 | 1395000 | Unaccompanied | Commercial associate | Higher edu | M |
| 9 | 100010 | 0 | Cash loans | M | Y | Y | 0 | 360000 | 1530000 | 42075 | 1530000 | Unaccompanied | State servant | Higher edu | M |
| 10 | 100011 | 0 | Cash loans | F | N | Y | 0 | 112500 | 1019610 | 33826.5 | 913500 | Children | Pensioner | Secondary | M |
| 11 | 100012 | 0 | Revolving loans | M | N | Y | 0 | 135000 | 405000 | 20250 | 405000 | Unaccompanied | Working | Secondary | S |
| 12 | 100014 | 0 | Cash loans | F | N | Y | 1 | 112500 | 652500 | 21177 | 652500 | Unaccompanied | Working | Higher edu | M |
| 13 | 100015 | 0 | Cash loans | F | N | Y | 0 | 38419.155 | 148365 | 10678.5 | 135000 | Children | Pensioner | Secondary | M |
| 14 | 100016 | 0 | Cash loans | F | N | Y | 0 | 67500 | 80865 | 5881.5 | 67500 | Unaccompanied | Working | Secondary | M |
| 15 | 100017 | 0 | Cash loans | M | Y | N | 1 | 225000 | 918468 | 28966.5 | 697500 | Unaccompanied | Working | Secondary | M |
| 16 | 100018 | 0 | Cash loans | F | N | Y | 0 | 189000 | 773680.5 | 32778 | 679500 | Unaccompanied | Working | Secondary | M |
| 17 | 100019 | 0 | Cash loans | M | Y | Y | 0 | 157500 | 299772 | 20160 | 247500 | Family | Working | Secondary | S |
| 18 | 100020 | 0 | Cash loans | M | N | N | 0 | 108000 | 509602.5 | 26149.5 | 387000 | Unaccompanied | Working | Secondary | M |
| 19 | 100021 | 0 | Revolving loans | F | N | Y | 1 | 81000 | 270000 | 13500 | 270000 | Unaccompanied | Working | Secondary | M |
| 20 | 100022 | 0 | Revolving loans | F | N | Y | 0 | 112500 | 157500 | 7875 | 157500 | Other_A | Working | Secondary | W |
| 21 | 100023 | 0 | Cash loans | F | N | Y | 1 | 90000 | 544491 | 17563.5 | 454500 | Unaccompanied | State servant | Higher edu | S |
| 22 | 100024 | 0 | Revolving loans | M | Y | Y | 0 | 135000 | 427500 | 21375 | 427500 | Unaccompanied | Working | Secondary | M |
| 23 | 100025 | 0 | Cash loans | F | Y | Y | 1 | 202500 | 1132573.5 | 37561.5 | 927000 | Unaccompanied | Commercial associate | Secondary | M |
| 24 | 100026 | 0 | Cash loans | F | N | N | 1 | 450000 | 497520 | 32521.5 | 450000 | Unaccompanied | Working | Secondary | M |
| 25 | 100027 | 0 | Cash loans | F | N | Y | 0 | 83250 | 239850 | 23850 | 225000 | Unaccompanied | Pensioner | Secondary | M |
| 26 | 100029 | 0 | Cash loans | M | Y | N | 2 | 135000 | 247500 | 12703.5 | 247500 | Unaccompanied | Working | Secondary | M |
| 27 | 100030 | 0 | Cash loans | F | N | Y | 0 | 90000 | 225000 | 11074.5 | 225000 | Unaccompanied | Working | Secondary | M |

application_data

# Identify Missing Data and Deal with it Appropriately

**In Applicant_data.csv**
Before Cleaning, the number of Columns and rows are 122 and 3075124 respectively.

-

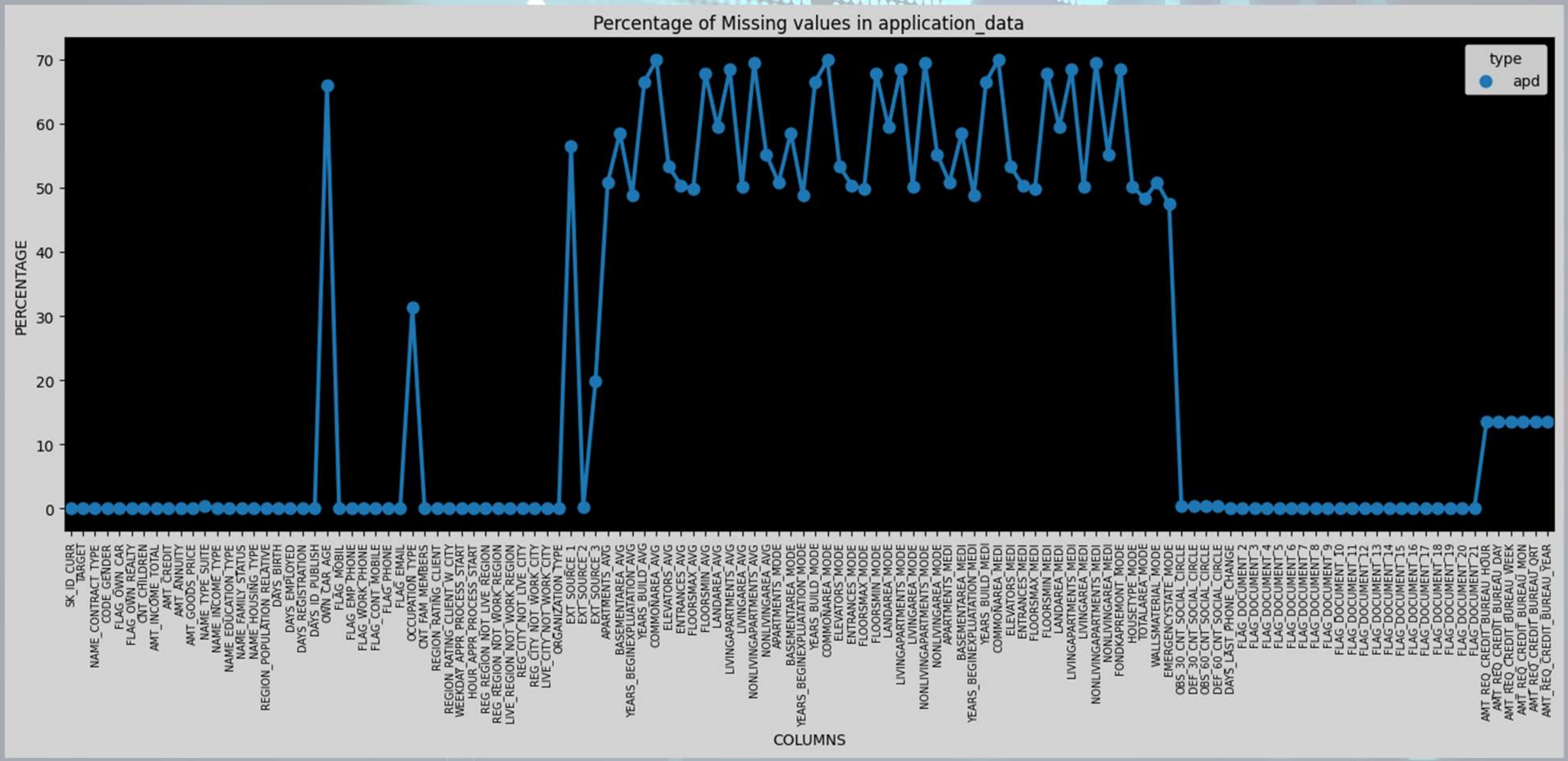
Items removed from the original dataset are :

* There are columns having more than 40% null data.
* There are more than 50 unwanted columns or columns not desirable for our analysis.

(Hint: Note that in EDA, since it is not necessary to replace the missing value, but if you have to replace the missing value, what should be the approach. Clearly mention the approach.)

*There are columns with null values less than 40%. They can be treated in 2 ways. I can delete those columns but then I might lose some important information required for my analysis. I can retain it but then I will have to do treatment. If I impute them, I will introduce bias. The decision to delete or retain basically depends on the Understanding of the problem statement, the usefulness of the variable, total size of available data. Here it seems that those columns can be removed So, I have removed them. There are still some columns will very little missing values which will be treated if necessary or left as it is.

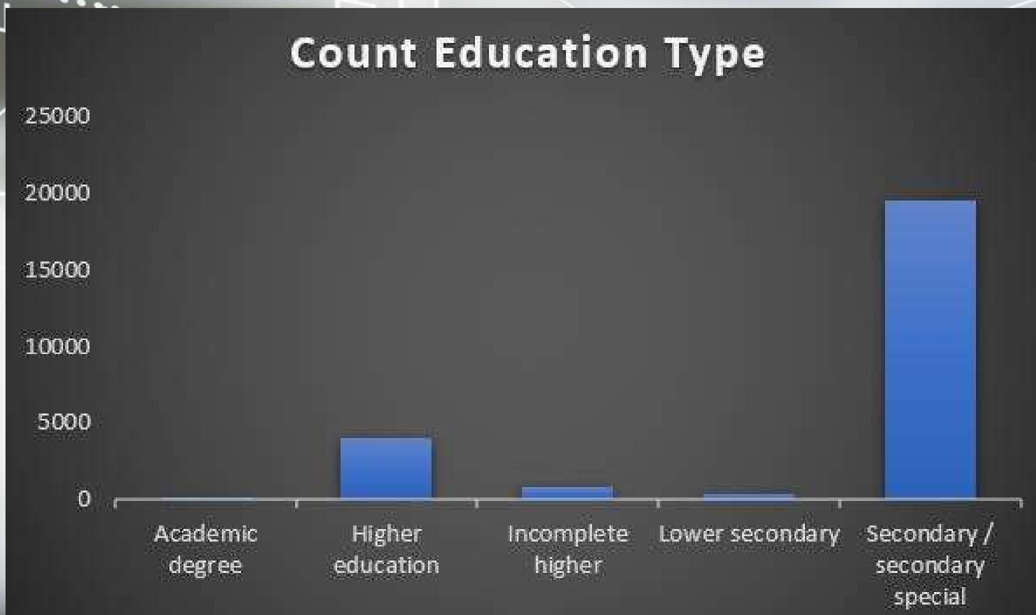
# Identify Missing Data and Deal with it Appropriately

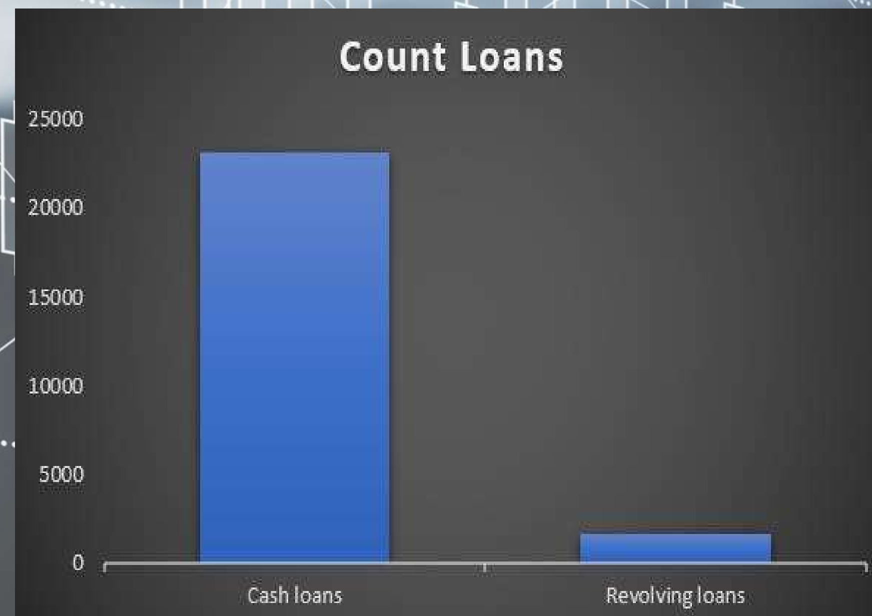


# Identify Outliers in the Dataset

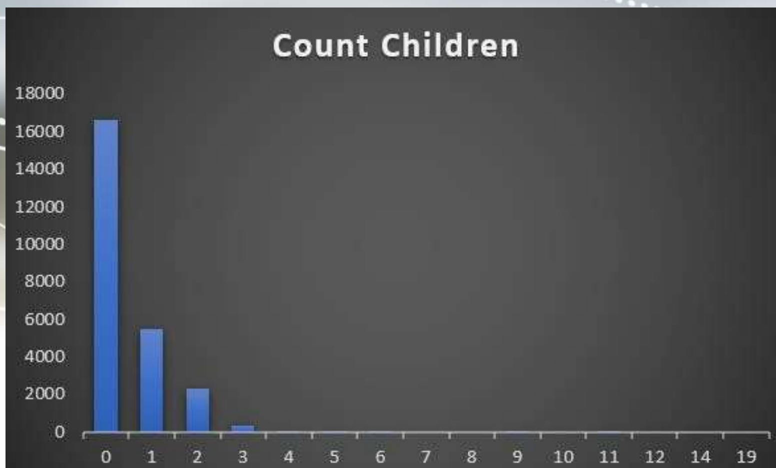


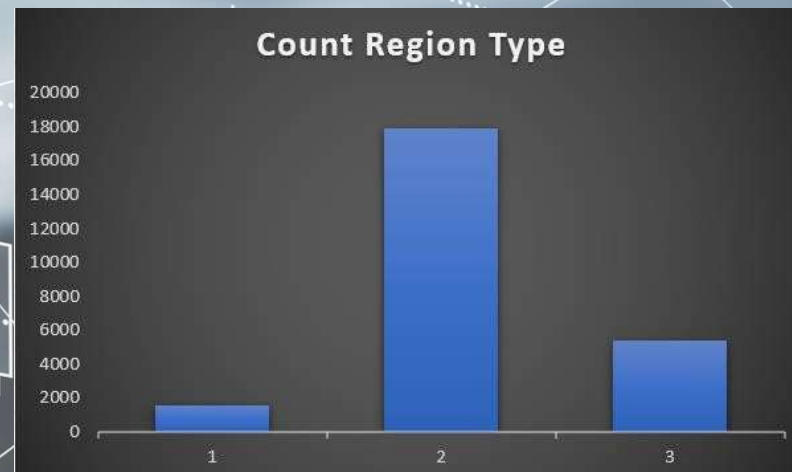
# Analyze Data Imbalance

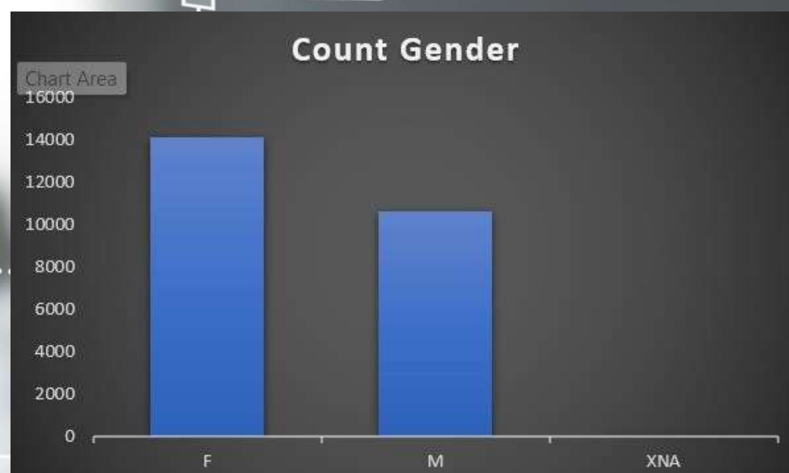


# Analyze Data Imbalance



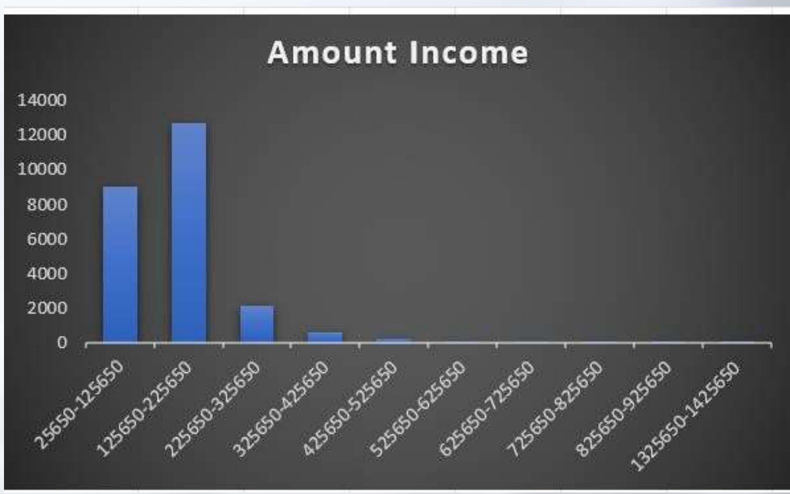



# Perform Univariate, Segmented Univariate, and Bivariate Analysis

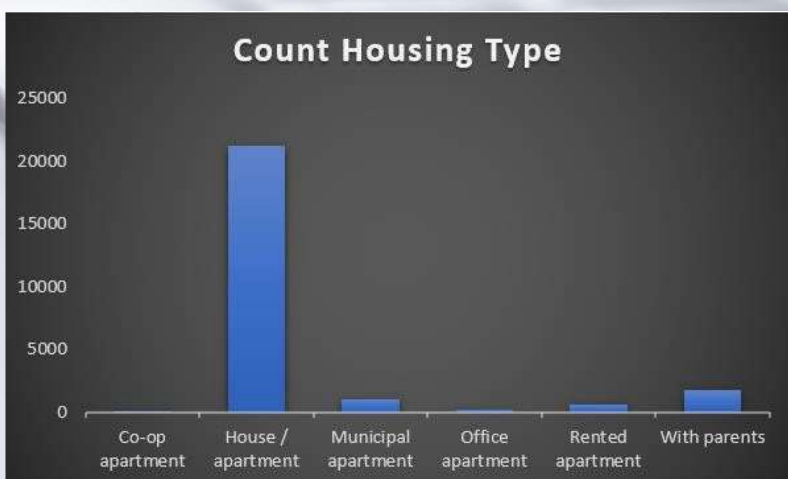
- ## UNIVARIATE ANALYSIS :

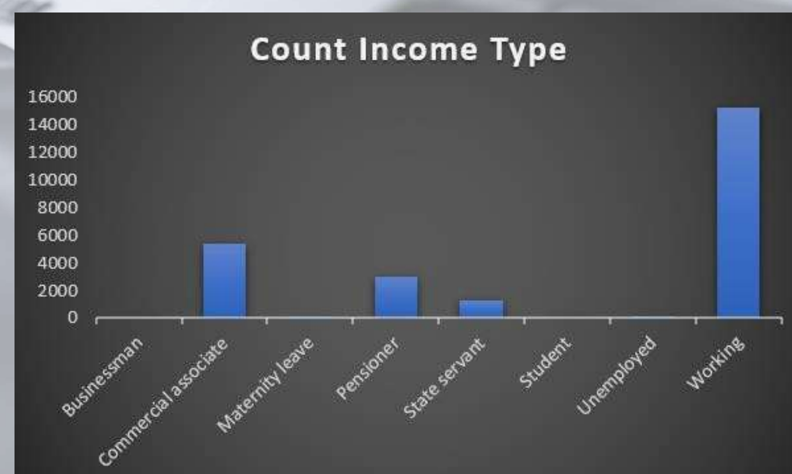
  - **Individuals with higher incomes are less likely to apply for loans.**
  - **The credit amount of a bank loan is typically in the range of 45000 to 1045000.**
  - **The majority of loan applications have come from people between the ages of 35 and 50.**
  - **Those with 0 to 8 years of work experience are the most likely to seek for loans.**
  - **Individuals who own homes are more likely to apply for loans than others.**
  - **Those who are married have taken out more loans.**
  - **More loans have been requested by working people.**
  - **Unaccompanied minors have requested for extra loans.**

- **UNIVARIATE ANALYSIS :**

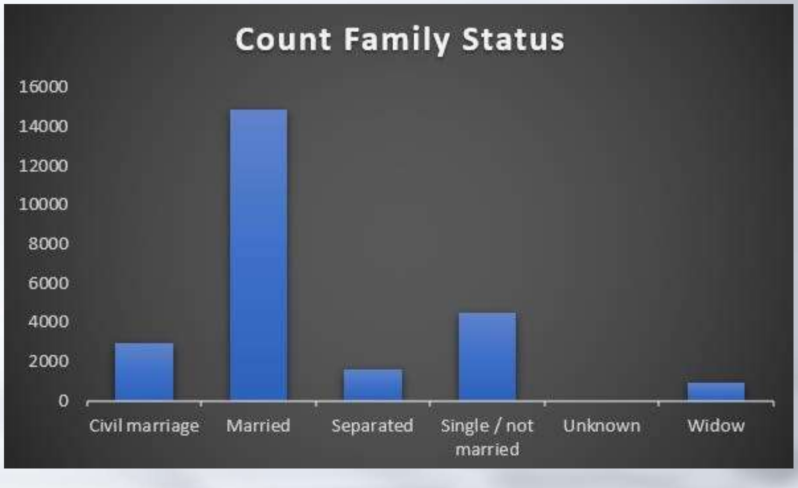


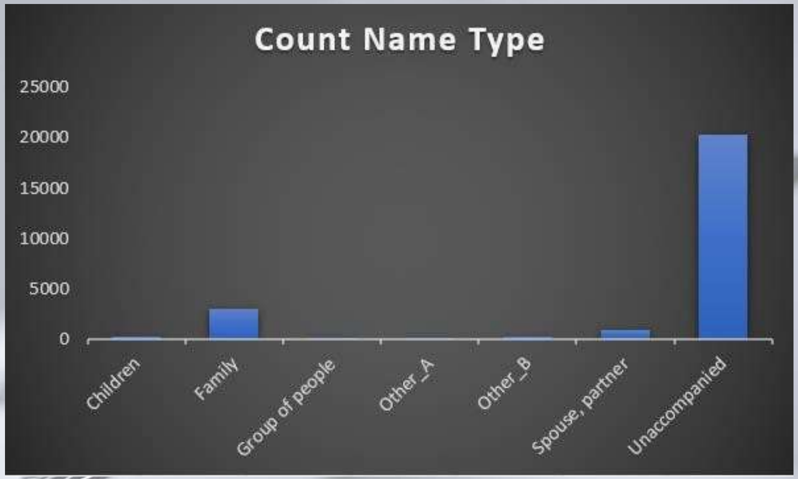




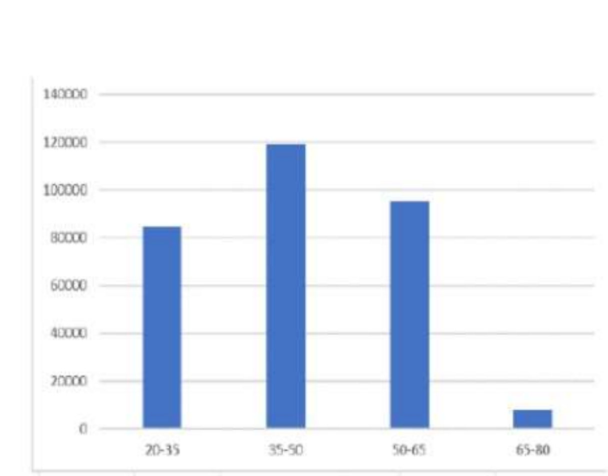
- **UNIVARIATE ANALYSIS :**



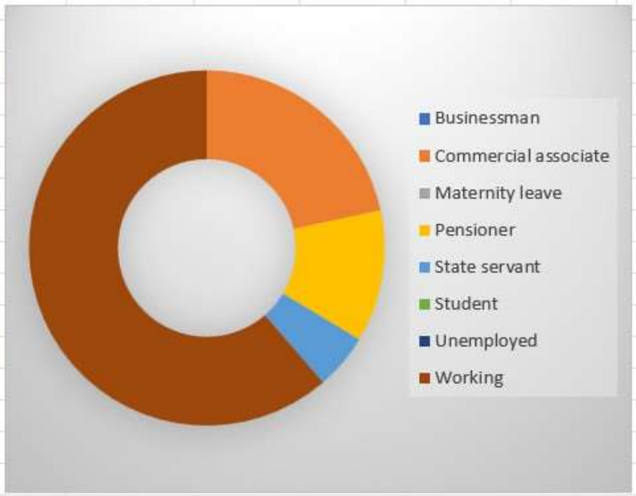


AGE

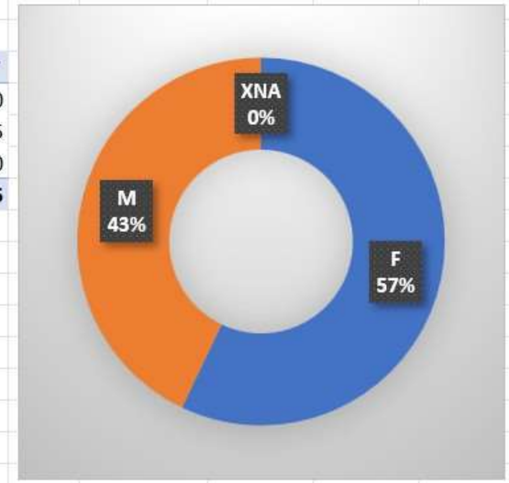


- # SEGMENTED UNIVARIATE ANALYSIS:

| Row Labels | Sum of TARGET |
|---|---|
| Businessman | 0 |
| Commercial associate | 5360 |
| Maternity leave | 2 |
| Pensioner | 2982 |
| State servant | 1249 |
| Student | 0 |
| Unemployed | 8 |
| Working | 15224 |
| **Grand Total** | **24825** |

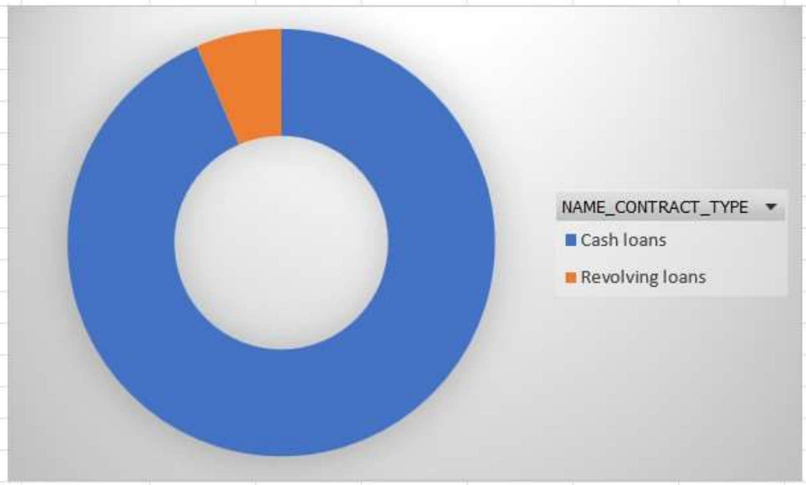



| Row Labels | Sum of TARGET |
|---|---|
| F | 14170 |
| M | 10655 |
| XNA | 0 |
| **Grand Total** | **24825** |



- **SEGMENTED UNIVARIATE ANALYSIS:**

| Row Labels | Sum of TARGET |
|---|---|
| Cash loans | 23221 |
| Revolving loans | 1604 |
| **Grand Total** | **24825** |

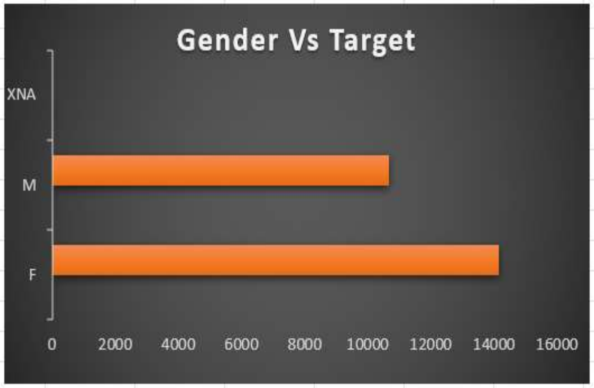


# BIVARIATE ANALYSIS :

- Customers who live in low-rating areas will have higher defaults.
- Individuals with lower incomes are more likely to default.
- Young people are more likely to default, and the trend of defaulters declines with age.
- Ladies are less inclined than males to have defaults.
- More defaults are predicted due to maternity leave and unemployment.
- Customers with more than five family members are more likely to default on their bank loan.
- Customers with fewer educational qualifications are more likely to fail on a bank loan.
- Customers with hardly work experience are more likely to have defaults.

# BIVARIATE ANALYSIS :

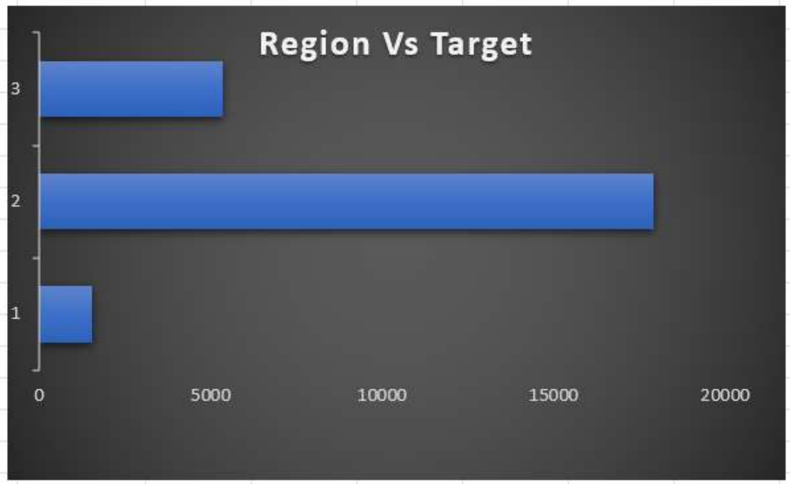
| Sum of TARGET | Column Labels | | |
|---|---|---|---|
| Row Labels | 0 | 1 | Grand Total |
| F | 0 | 14170 | 14170 |
| M | 0 | 10655 | 10655 |
| XNA | 0 | | 0 |
| **Grand Total** | **0** | **24825** | **24825** |




| | Sum of TARGET | Column Labels | | |
|---|---|---|---|---|
| 4 | Row Labels | 0 | 1 | Grand Total |
| 5 | 25650-125650 | 0 | 9067 | 9067 |
| 6 | 125650-225650 | 0 | 12686 | 12686 |
| 7 | 225650-325650 | 0 | 2129 | 2129 |
| 8 | 325650-425650 | 0 | 590 | 590 |
| 9 | 425650-525650 | 0 | 208 | 208 |
| 10 | 525650-625650 | 0 | 68 | 68 |
| 11 | 625650-725650 | 0 | 44 | 44 |
| 12 | 725650-825650 | 0 | 7 | 7 |
| 13 | 825650-925650 | 0 | 11 | 11 |
| 14 | 1325650-1425650 | 0 | 3 | 3 |
| 15 | **Grand Total** | **0** | **24813** | **24813** |
| 16 | | | | |

- # BIVARIATE ANALYSIS :

| Row Labels | Sum of TARGET |
|---|---|
| 1 | 1552 |
| 2 | 17907 |
| 3 | 5366 |
| **Grand Total** | **24825** |



# Identify Top Correlations for Different Scenarios

Top 10 driving factors in current application.csv

1. Income type
2. Count of Family Members
3. Children count
4. External source
5. Region rating of client
6. Age
7. Months Employed
8. Amount credit
9. Amount Goods Price
10. Amount total income


Defaulter By Occupation Type
7000
6000
5000
4000
3000
2000
1000
0
Accountants
Cleaning staff
Cooking staff
Core staff
Drivers
High skill tech staff
HR staff
IT staff
Laborers
Low-skill Laborers
Managers
Medicine staff
Private service staff
Realty agents
Sales staff
Secretaries
Security staff
Waiters/barmen staff
(blank)
Total
Linear (Total)
Defaulter By Education Type
25000
20000
15000
10000
5000
0
-5000
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special
Total
Linear (Total)

| Loan- Highly Recommended Groups | Loan- High Risk Groups |
|---|---|
| 1. Previous application approved clients | 1. Clients that are unemployed |
| 2. Married clients | 2. Youth clients 3. Customers whose prior applications were denied |
| 3. Senior clients | 4. Low-income clientele |
| 4. More educated clients | 5. Clients with insufficient external sources |
| 5. Customers with a High Income | 6.Customers with little work experience |
| 6.Clients with a greater external source | 7.Customers on Maternity Leave |
| 7. Females | 8. Clients with a larger number of family members |
| 8.Customers with strong work experience | |

# Result

- After performing the analysis, we can rectify whether a client will repay the loan or not.
- The people who are likely to face problem in loan repayment are labourers.
- People with Secondary /secondary special education might face problem in loan repayment.
- Moreover, those who are living in house/apartment are facing difficulty in loan repayment (may be because of extra home loan, EMIs and so on).