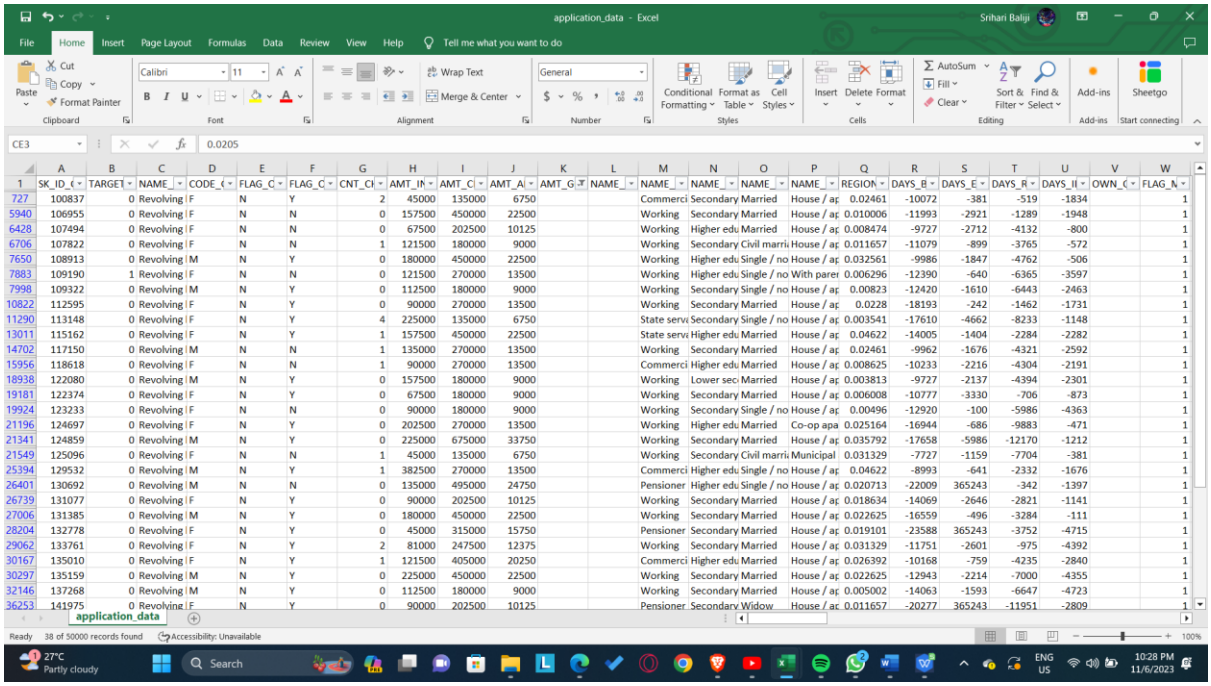
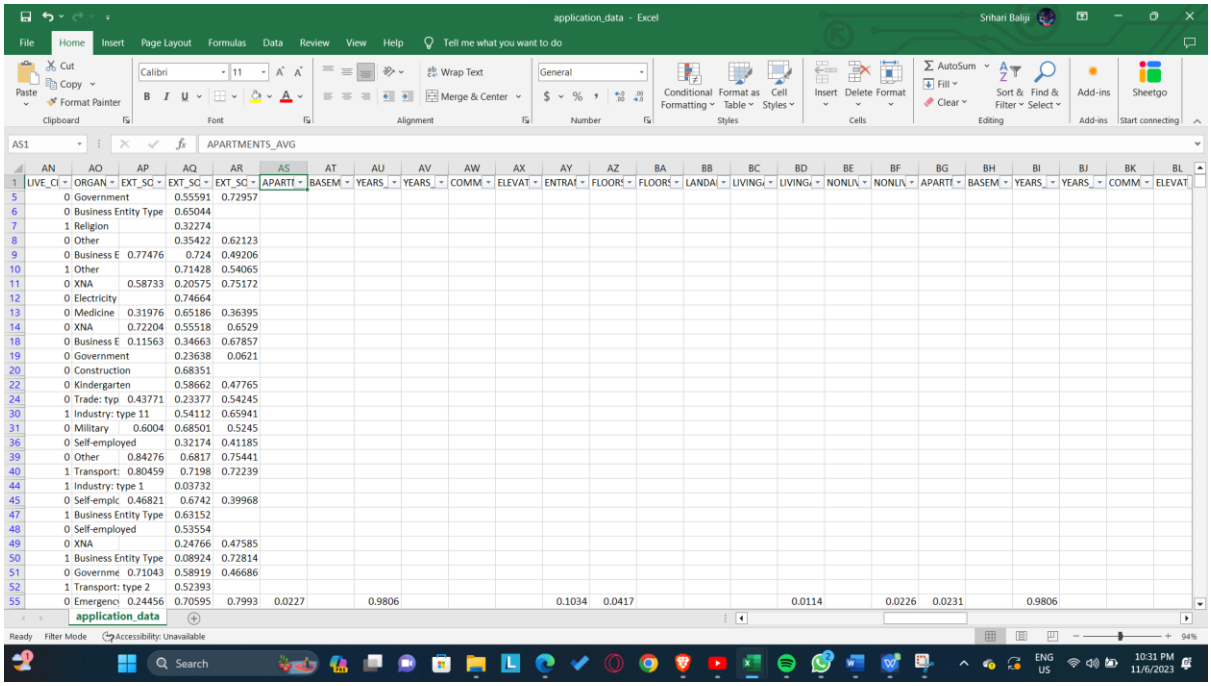
# Bank Loan Case Study

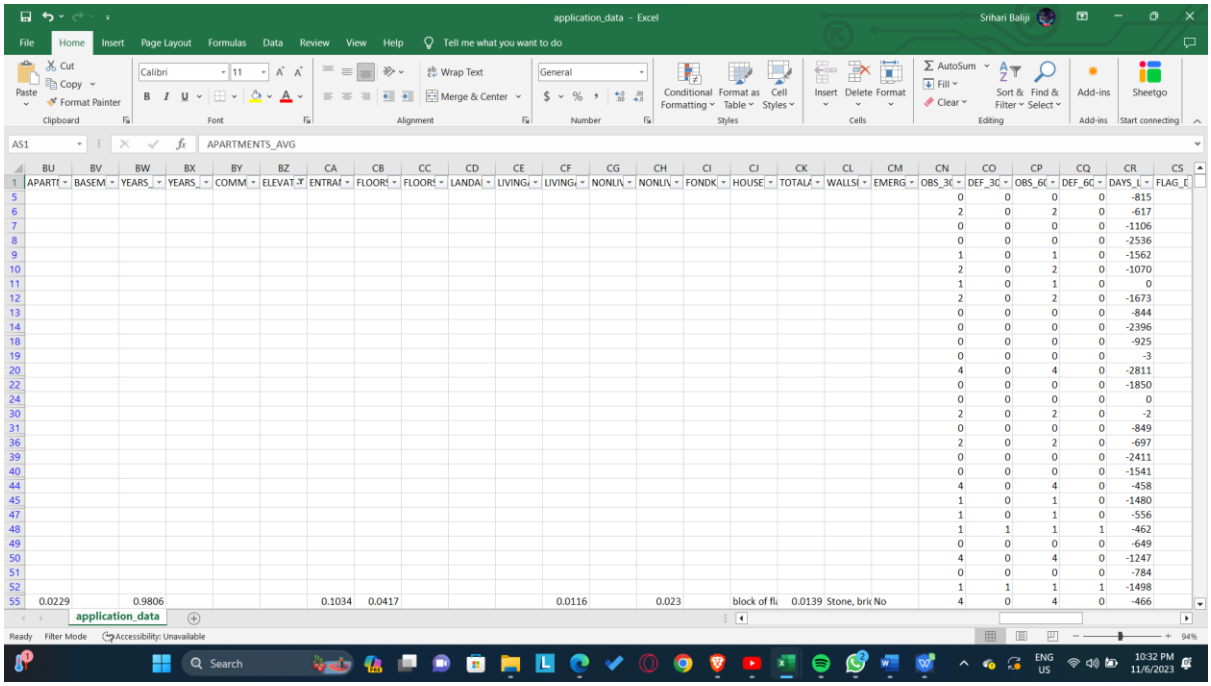
A bank loan case study analyzes a bank's loan portfolio to identify trends and patterns. This information can be used to improve the bank's lending practices and reduce risk.

## Identification of Missing Data and Dealing with it Appropriately:

Here we are using Excel formulas to see the missing data in the given data set which is “=COUNTBLANK(A:A)” We will be applying this to all the columns in the data set to find the number of missing cells in the column.





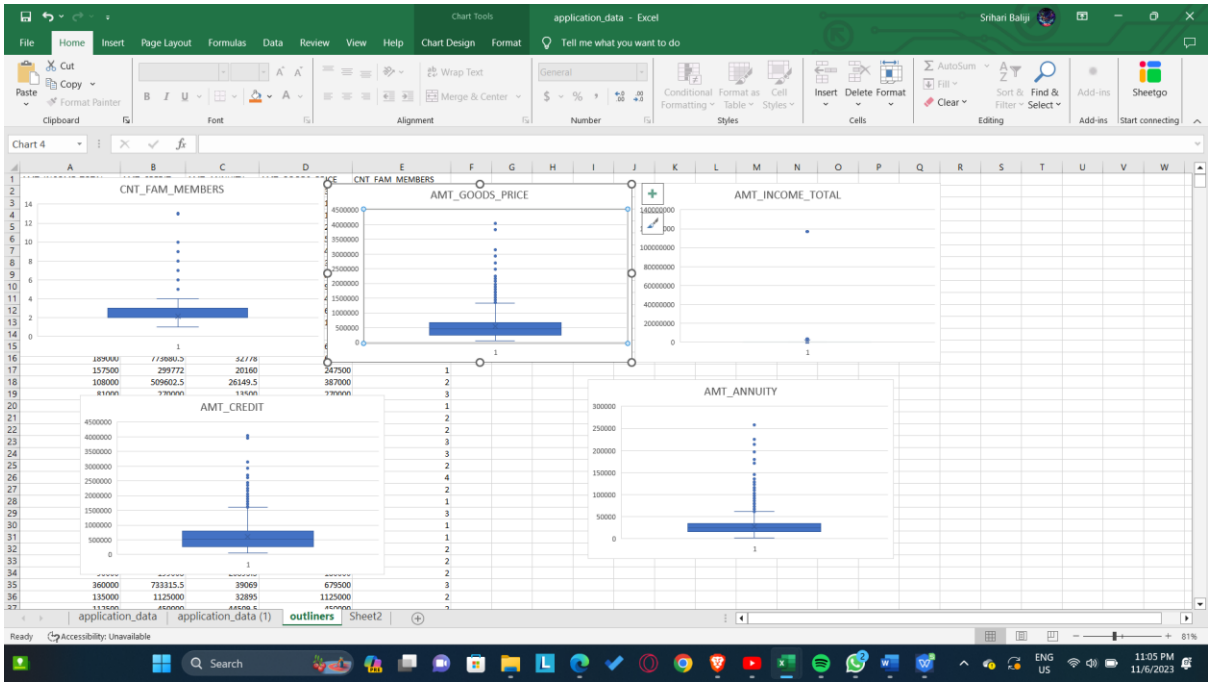

From the above images, we can say that there are moves and more missing cells in the data from

APARTMENTS_AVG to EMERGENCYSTATE_MODE

More than 35% of the cells in this columns are empty

## Identify Outliers in the Dataset:

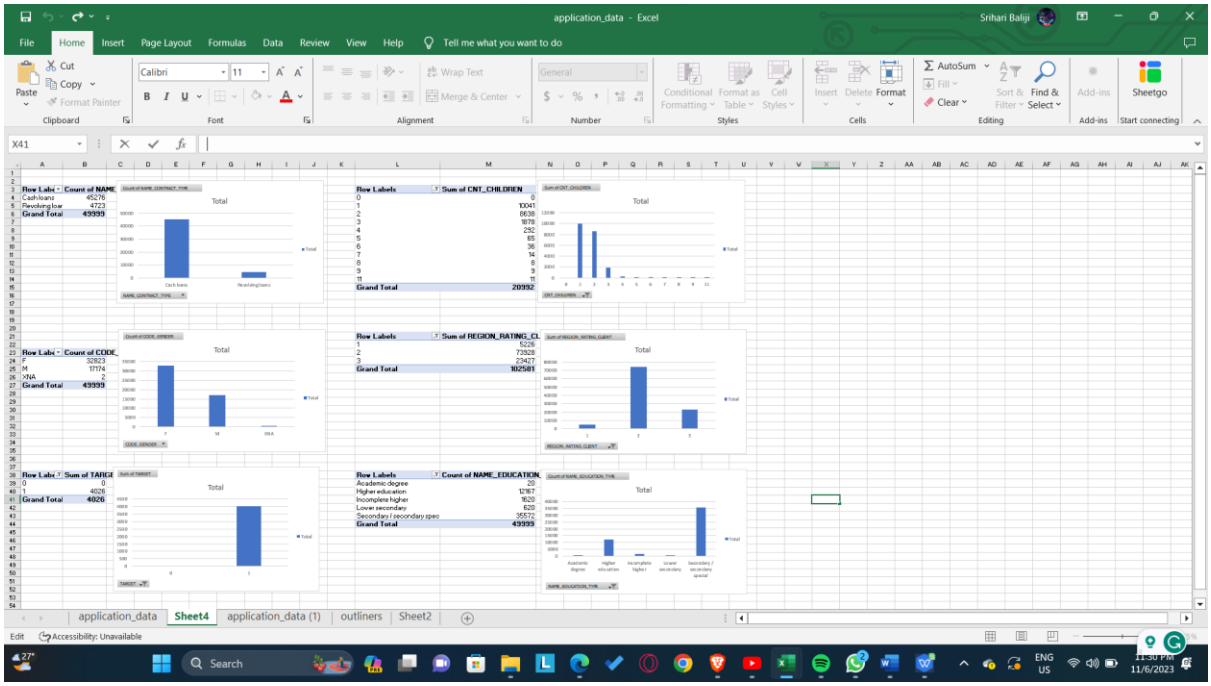
Here  Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset. By using the created plot box for identification of the outliners




## Analyzing of Data Imbalance:

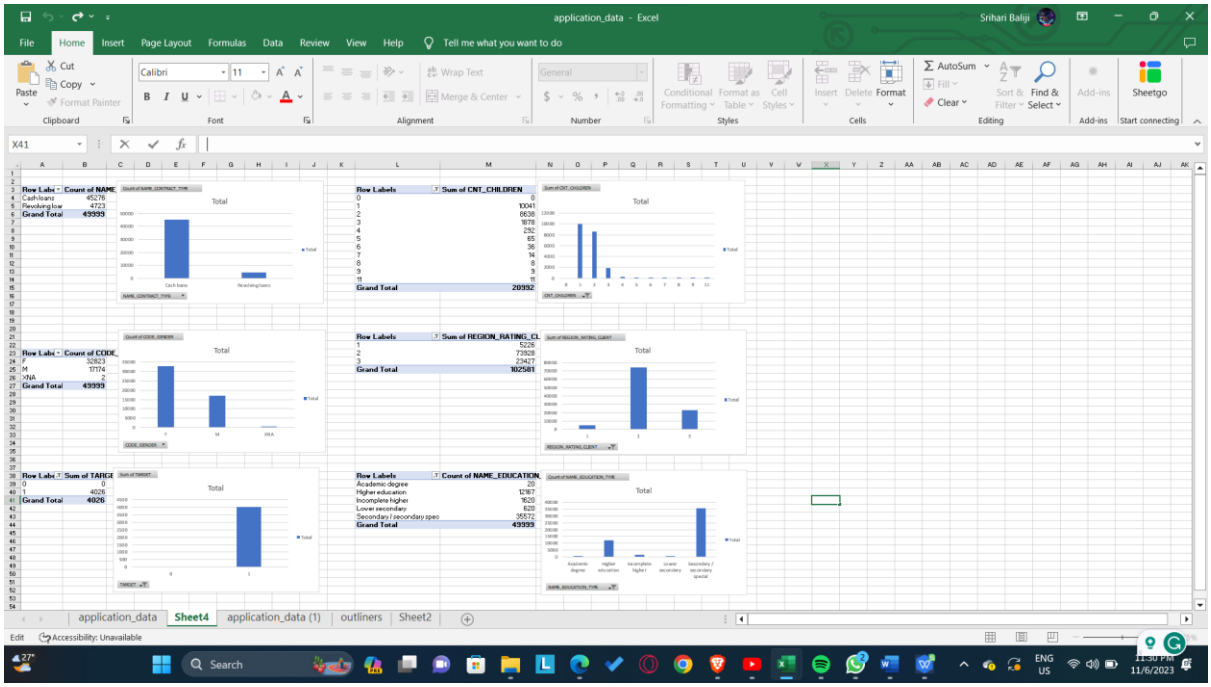
Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

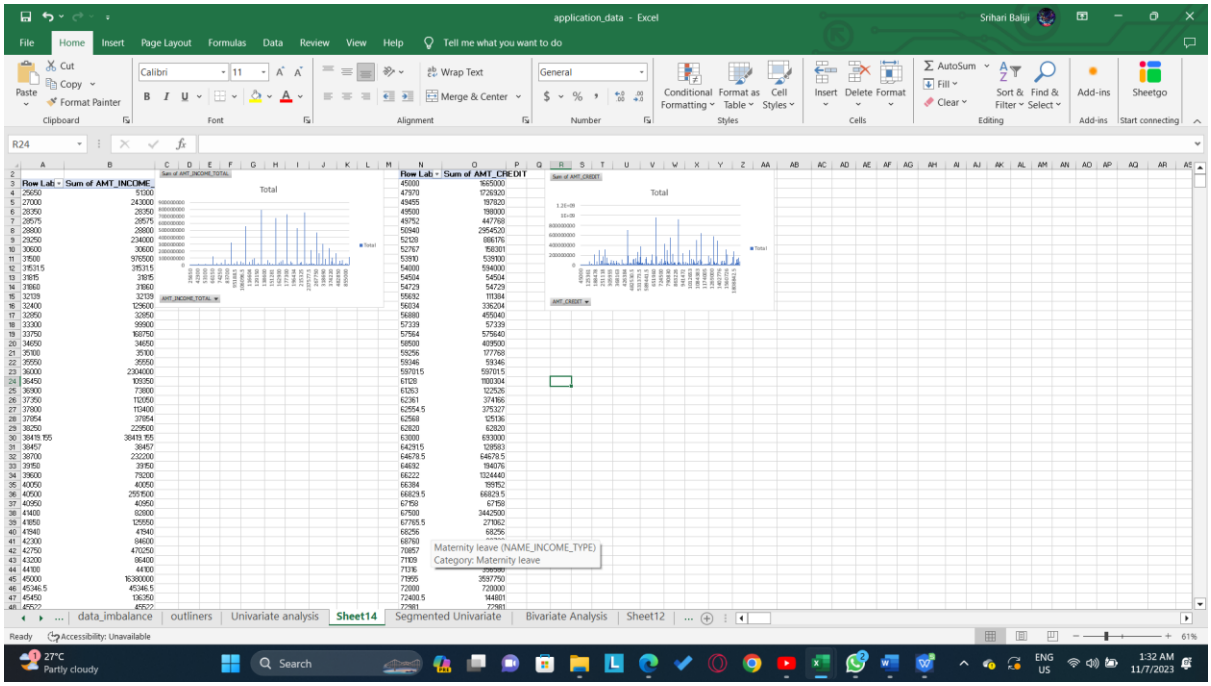
Here we use a Pivort table to show the imbalance of data



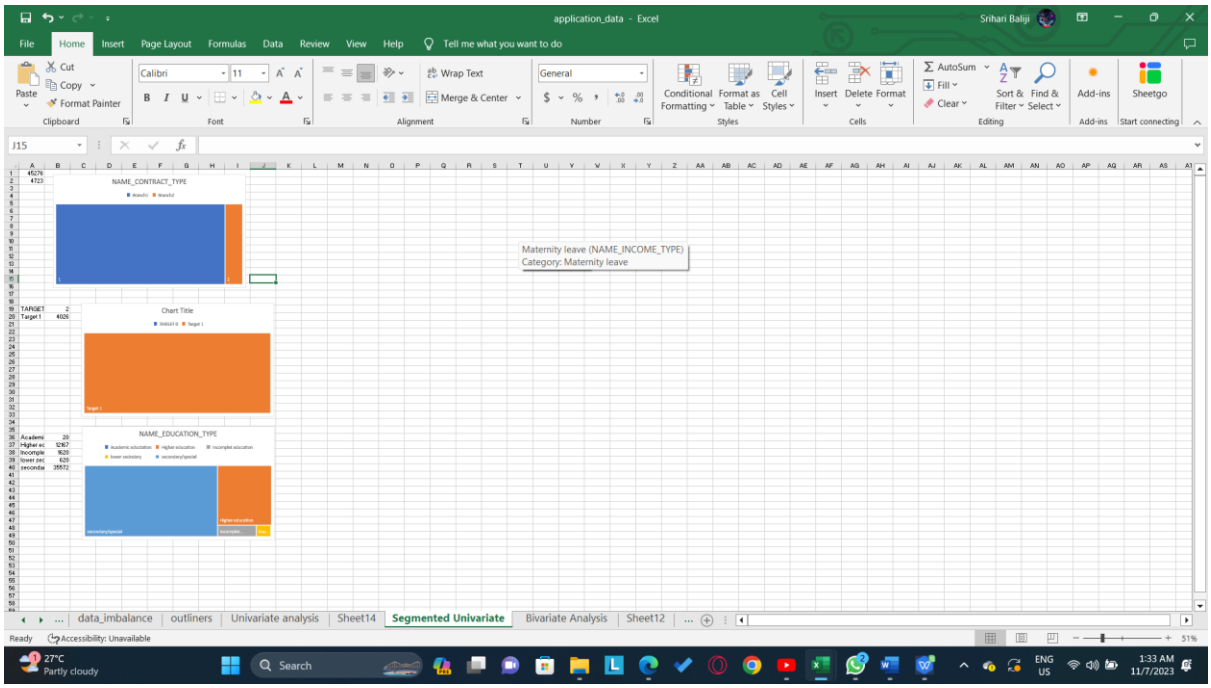

**Perform Univariate, Segmented Univariate, and Bivariate Analysis:** To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

## Univariate:

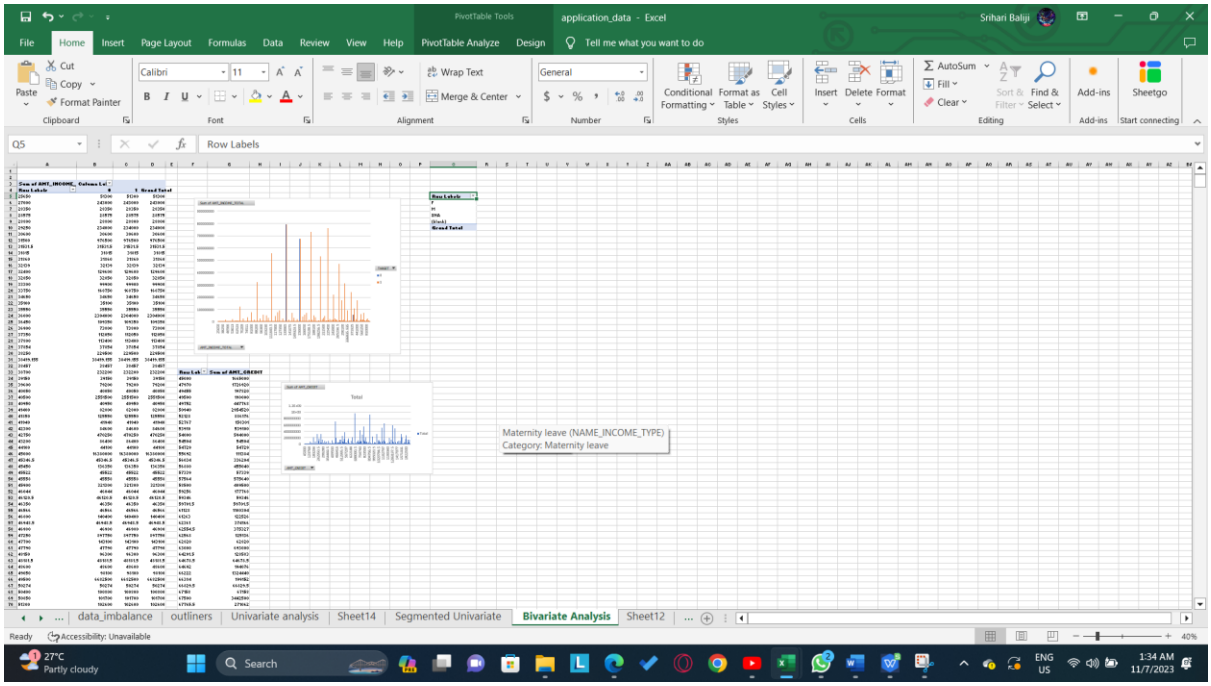




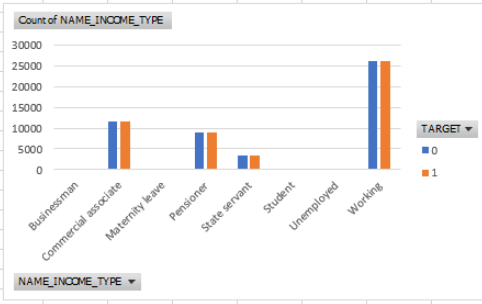

# Segmented Univariate:



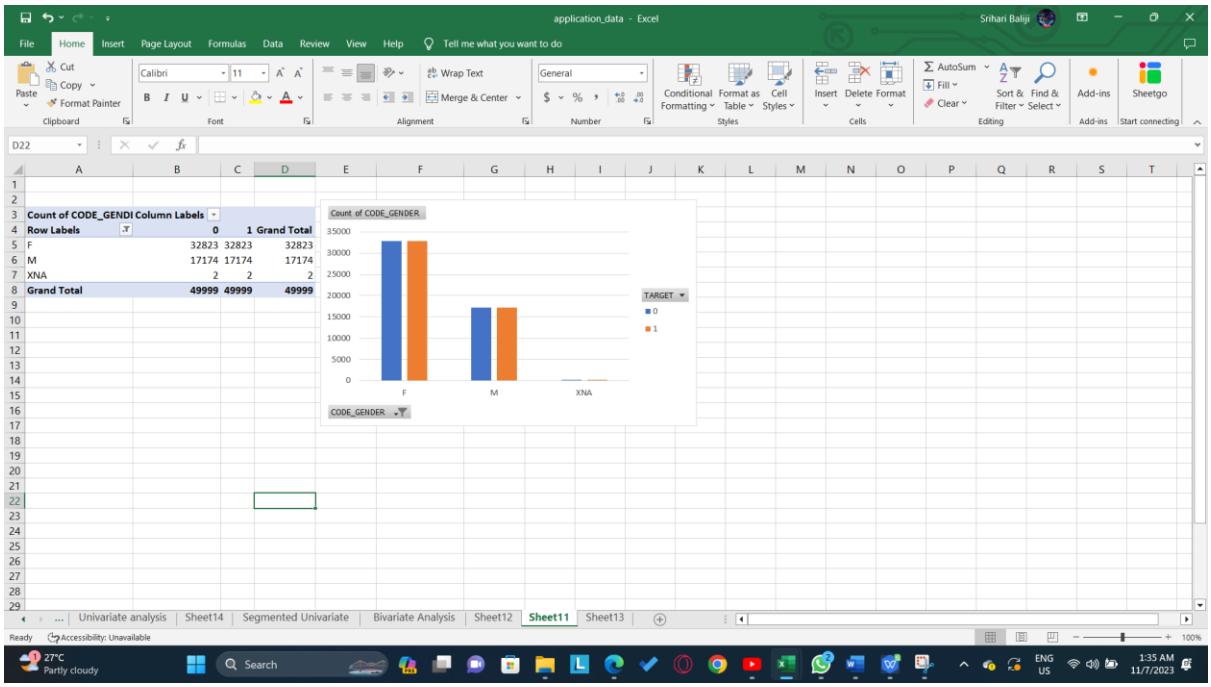

# Bivariate Analysis:



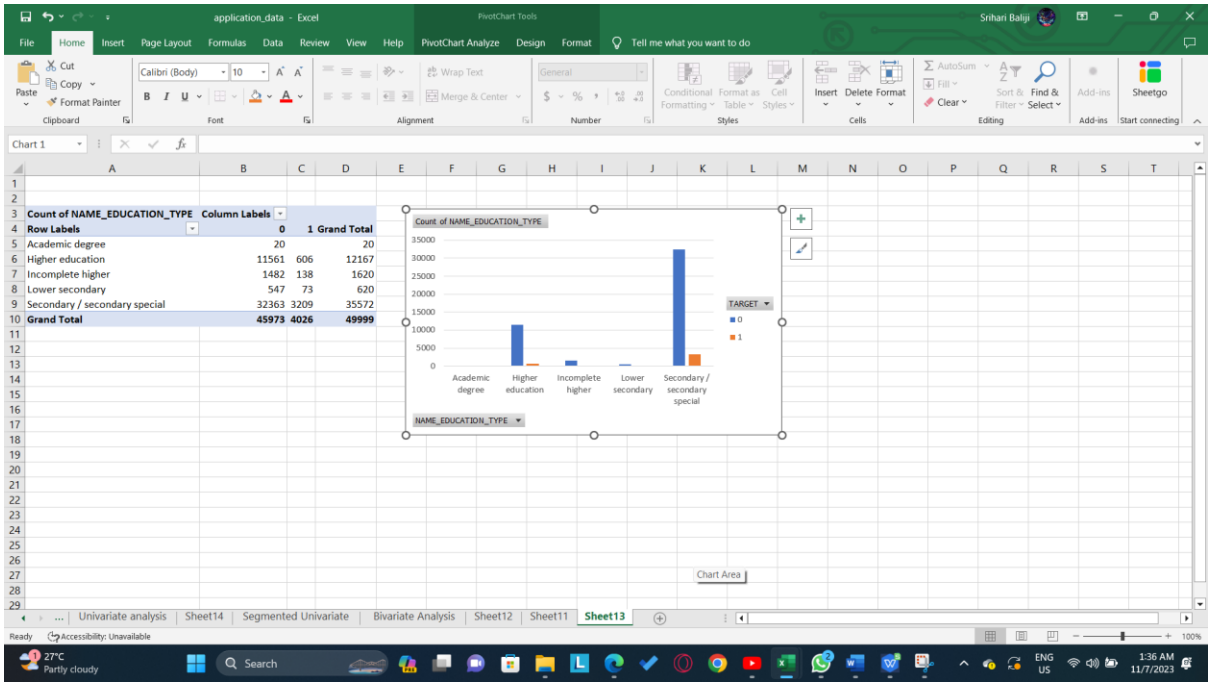

| Count of NAME_INCOME_TYPE | Column Labels | | |
|---|---|---|---|
| Row Labels | 0 | 1 | Grand Total |
| Businessman | 2 | 2 | 2 |
| Commercial associate | 11543 | 11543 | 11543 |
| Maternity leave | 1 | 1 | 1 |
| Pensioner | 8920 | 8920 | 8920 |
| State servant | 3512 | 3512 | 3512 |
| Student | 5 | 5 | 5 |
| Unemployed | 6 | 6 | 6 |
| Working | 26010 | 26010 | 26010 |
| Grand Total | 49999 | 49999 | 49999 |




Manual Filters




| Count of CODE_GENDI | Column Labels | | |
|---|---|---|---|
| Row Labels | 0 | 1 | Grand Total |
| F | 32823 | 32823 | 32823 |
| M | 17174 | 17174 | 17174 |
| XNA | 2 | 2 | 2 |
| Grand Total | 49999 | 49999 | 49999 |




## Identify Top Correlations for Different

**Scenarios:** Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

| | CNT_CHILDREN | AMT_INCOME_TOTAL | AMT_CREDIT | AMT_ANNUITY | AMT_GOODS_PRICE | REGION_POPULATION_RELATIVE | Age | Years Employed | CNT_FAM_MEMBERS | REGION_RATING_CLIENT | EXT_SOURCE_2 | EXT_SOURCE_3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CNT_CHILDREN | 1 | 0.028 | 0.003 | 0.021 | -0.001 | -0.024 | -0.337 | -0.041 | 0.879 | 0.023 | -0.015 | -0.037 |
| AMT_INCOME_TOTAL | 0.028 | 1 | 0.343 | 0.419 | 0.349 | 0.168 | -0.063 | 0.035 | 0.034 | -0.187 | 0.140 | -0.059 |
| AMT_CREDIT | 0.003 | 0.343 | 1 | 0.771 | 0.987 | 0.100 | 0.047 | 0.084 | 0.065 | -0.103 | 0.129 | 0.033 |
| AMT_ANNUITY | 0.021 | 0.419 | 0.771 | 1 | 0.777 | 0.121 | -0.013 | 0.053 | 0.076 | -0.132 | 0.127 | 0.025 |
| AMT_GOODS_PRICE | -0.001 | 0.349 | 0.987 | 0.777 | 1 | 0.104 | 0.045 | 0.085 | 0.063 | -0.104 | 0.136 | 0.036 |
| REGION_POPULATION_RELATIVE | -0.024 | 0.168 | 0.100 | 0.121 | 0.104 | 1 | 0.025 | -0.009 | -0.023 | -0.539 | 0.198 | -0.011 |
| Age | -0.337 | -0.063 | 0.047 | -0.013 | 0.045 | 0.025 | 1 | 0.226 | -0.286 | -0.002 | 0.078 | 0.175 |
| Years Employed | -0.041 | 0.035 | 0.084 | 0.053 | 0.085 | -0.009 | 0.226 | 1 | -0.010 | 0.016 | 0.074 | 0.097 |
| CNT_FAM_MEMBERS | 0.879 | 0.034 | 0.065 | 0.076 | 0.063 | -0.023 | -0.286 | -0.010 | 1 | 0.028 | -0.001 | -0.023 |
| REGION_RATING_CLIENT | 0.023 | -0.187 | -0.103 | -0.132 | -0.104 | -0.539 | -0.002 | 0.016 | 0.028 | 1 | -0.291 | -0.004 |
| EXT_SOURCE_2 | -0.015 | 0.140 | 0.129 | 0.127 | 0.136 | 0.198 | 0.078 | 0.074 | -0.001 | -0.291 | 1 | 0.076 |
| EXT_SOURCE_3 | -0.037 | -0.059 | 0.033 | 0.025 | 0.036 | -0.011 | 0.175 | 0.097 | -0.023 | -0.004 | 0.076 | 1 |

| | CNT_CHILDREN | AMT_INCOME_TOTAL | AMT_CREDIT | AMT_ANNUITY | AMT_GOODS_PRICE | REGION_POPULATION_RELATIVE | Age | Years Employed | CNT_FAM_MEMBERS | REGION_RATING_CLIENT | EXT_SOURCE_2 | EXT_SOURCE_3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CNT_CHILDREN | 1 | 0.005 | -0.002 | 0.031 | -0.008 | -0.032 | -0.259 | -0.044 | 0.886 | 0.041 | -0.012 | -0.024 |
| AMT_INCOME_TOTAL | 0.005 | 1 | 0.038 | 0.046 | 0.038 | 0.009 | -0.003 | -0.001 | 0.007 | -0.021 | 0.007 | -0.018 |
| AMT_CREDIT | -0.002 | 0.0381 | 1 | 0.752 | 0.983 | 0.069 | 0.135 | 0.098 | 0.051 | -0.059 | 0.121 | 0.052 |
| AMT_ANNUITY | 0.031 | 0.0464 | 0.752 | 1 | 0.753 | 0.072 | 0.014 | 0.041 | 0.076 | -0.074 | 0.116 | 0.032 |
| AMT_GOODS_PRICE | -0.008 | 0.0376 | 0.983 | 0.753 | 1 | 0.076 | 0.136 | 0.104 | 0.047 | -0.066 | 0.131 | 0.053 |
| REGION_POPULATION_RELATIVE | -0.032 | 0.0091 | 0.069 | 0.072 | 0.076 | 1 | 0.048 | 0.016 | -0.030 | -0.443 | 0.170 | -0.010 |
| Age | -0.259 | -0.0031 | 0.135 | 0.014 | 0.136 | 0.048 | 1 | 0.281 | -0.203 | -0.034 | 0.108 | 0.134 |
| Years Employed | -0.044 | -0.0010 | 0.098 | 0.041 | 0.104 | 0.016 | 0.281 | 1 | -0.010 | -0.005 | 0.089 | 0.056 |
| CNT_FAM_MEMBERS | 0.886 | 0.0067 | 0.051 | 0.076 | 0.047 | -0.030 | -0.203 | -0.010 | 1 | 0.044 | 0.002 | -0.029 |
| REGION_RATING_CLIENT | 0.041 | -0.0215 | -0.059 | -0.074 | -0.066 | -0.443 | -0.034 | -0.005 | 0.044 | 1 | -0.250 | 0.014 |
| EXT_SOURCE_2 | -0.012 | 0.0071 | 0.121 | 0.116 | 0.131 | 0.170 | 0.108 | 0.089 | 0.002 | -0.250 | 1 | 0.049 |
| EXT_SOURCE_3 | -0.024 | -0.0182 | 0.052 | 0.032 | 0.053 | -0.010 | 0.134 | 0.056 | -0.029 | 0.014 | 0.049 | 1 |

The correlation is calculated between two coloumns using the formula correl(array1,array2) the highest is markes using green markings