## Project Description

To find trends that can be used to determine whether a customer has trouble making their payments, which may be used to decide whether to refuse the loan, reduce the loan's size, or lend to a riskier applicant. More expensive interest rates, etc. By doing this, it will be ensured that only borrowers who can repay the loan will be accepted.

There are two sections to the analysis, or should I say two datasets:

1. Application data
2. Previous application data

## Approach

Prior to analysis, data must be understood and cleaned. Finding null values, outliers, and identifying each column in our dataset to determine how many of them are irrelevant for analysis are all part of the cleaning process.

After data cleaning, we must analyses the data using univariate and bivariate methods, which aid in data analysis and provide insightful information about the relationship between two variables, or the interdependence of one and more factors.

## Tech-Stack Used

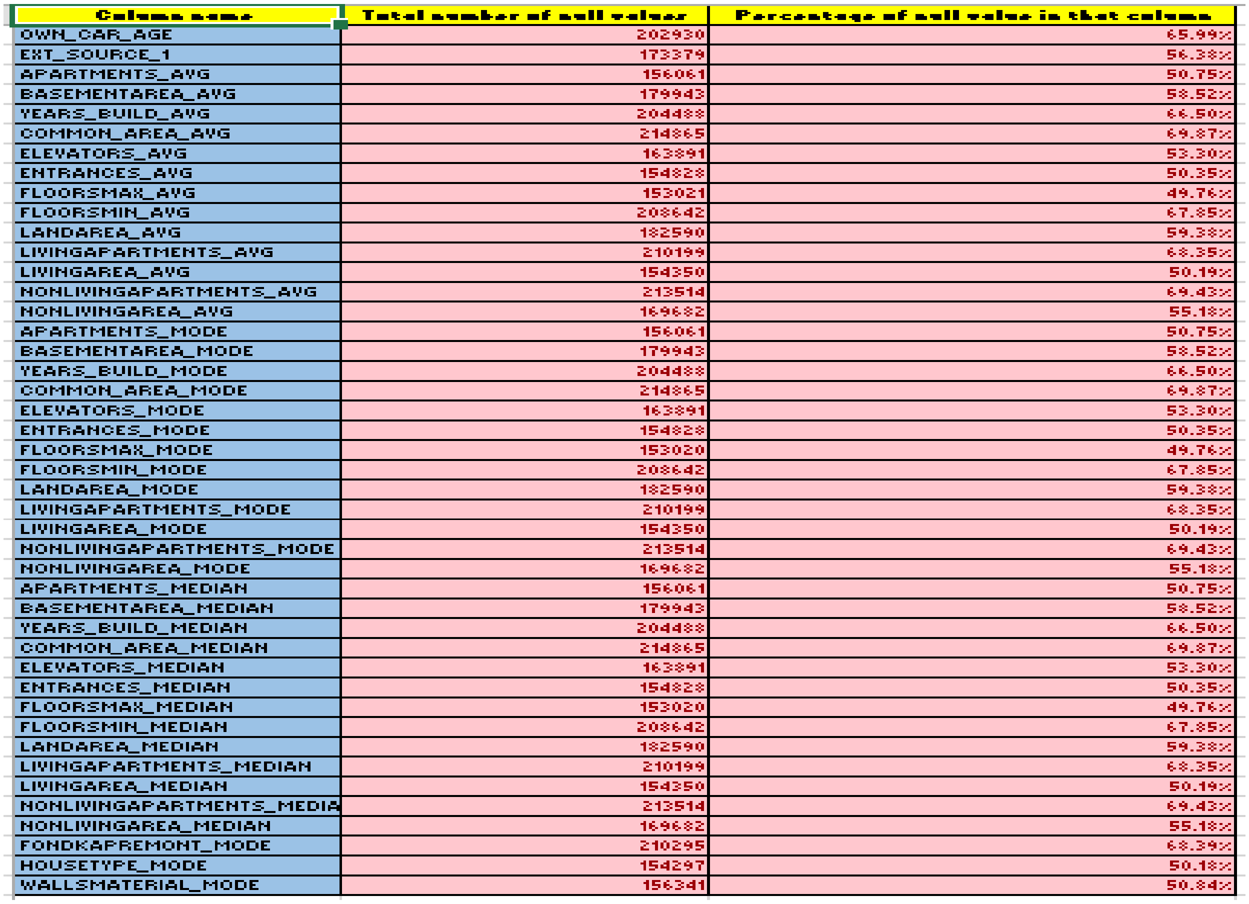
Here I am using Microsoft Excel 2016, I will be able to clean the data and develop a pivot table, which is useful for data analysis. We can visualize data using graphs in Excel as well. Due to the size of the dataset, I'm also using a Jupyter notebook to help me analyses the data.

## Insights

Application Dataset – NULL values

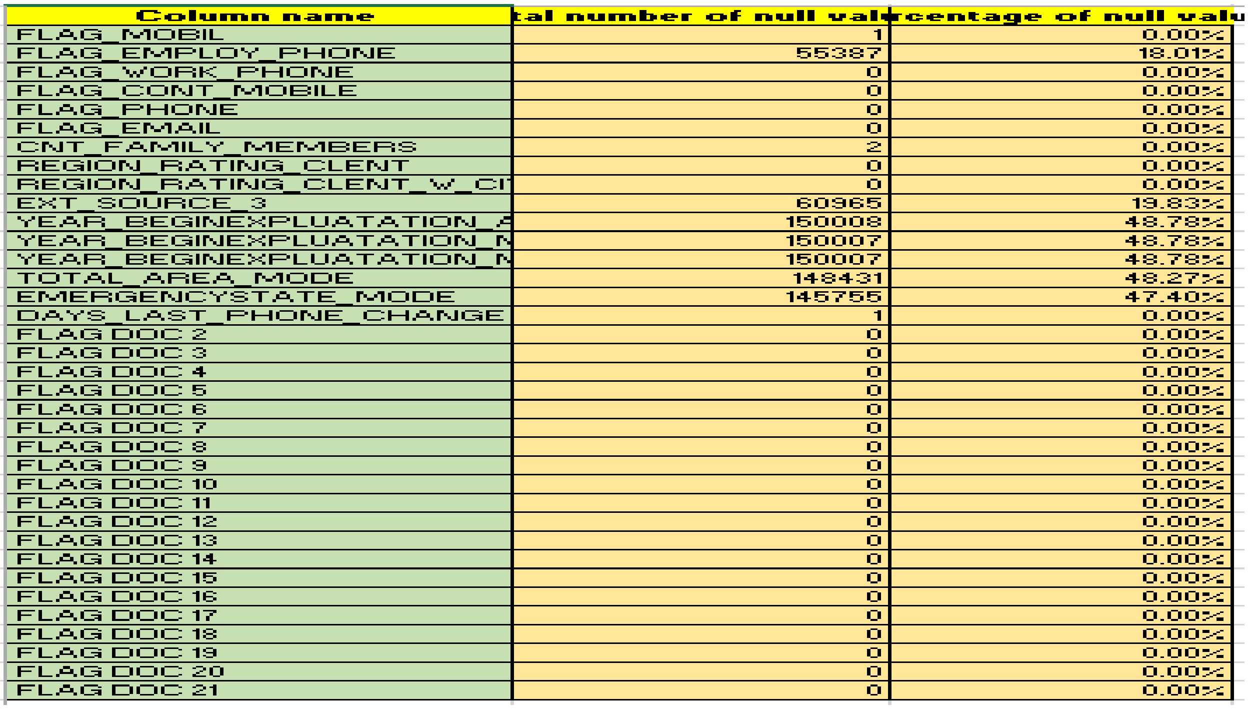
First, the proportion of null values must be examined, and any columns with more than 50% of the data being null must be removed.Additionally, any columns with less than 50% of null data must be replaced with the mean, median, or the category variable with the highest frequency.

As all of the below column names have null values greater than or equal to 50%, they can all be dropped down.



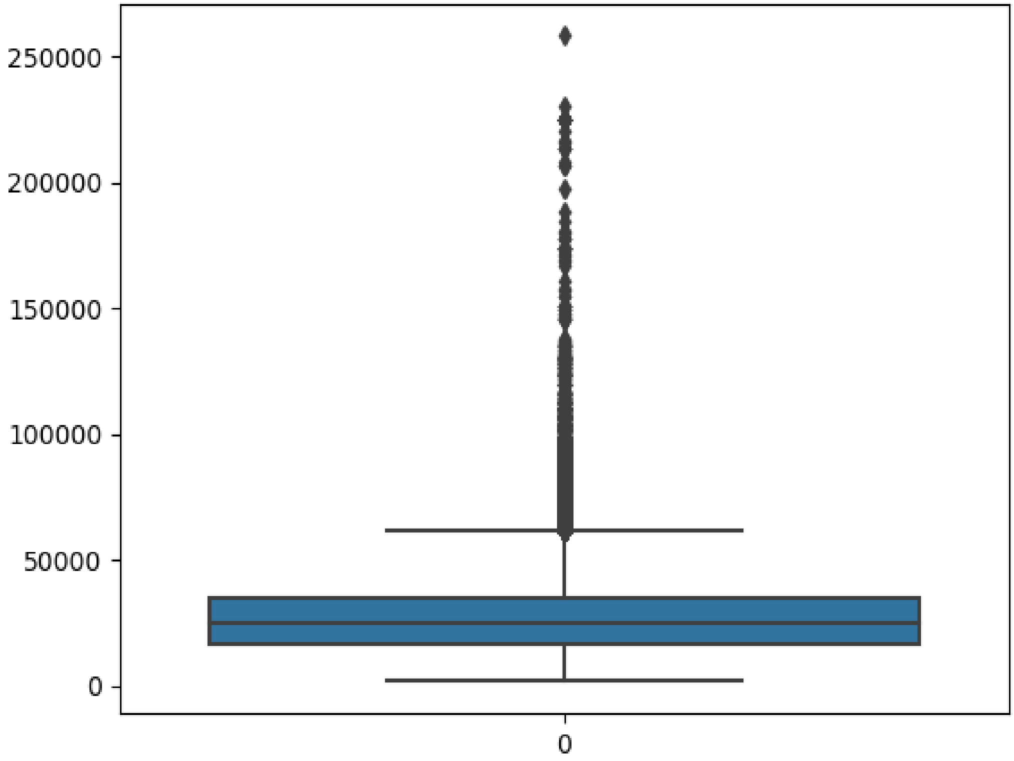
| Column name | Total number of null values | Percentage of null value in that column |
|---|---|---|
| OWN_CAR_AGE | 202930 | 65.99% |
| EXT_SOURCE_1 | 173379 | 56.38% |
| APARTMENTS_AVG | 156061 | 50.75% |
| BASEMENTAREA_AVG | 179943 | 58.52% |
| YEARS_BUILD_AVG | 204488 | 66.50% |
| COMMON_AREA_AVG | 214865 | 69.87% |
| ELEVATORS_AVG | 163891 | 53.30% |
| ENTRANCES_AVG | 154828 | 50.35% |
| FLOORSMAX_AVG | 153021 | 49.76% |
| FLOORSMIN_AVG | 208642 | 67.85% |
| LANDAREA_AVG | 182590 | 59.38% |
| LIVINGAPARTMENTS_AVG | 210199 | 68.35% |
| LIVINGAREA_AVG | 154350 | 50.19% |
| NONLIVINGAPARTMENTS_AVG | 213514 | 69.43% |
| NONLIVINGAREA_AVG | 169682 | 55.18% |
| APARTMENTS_MODE | 156061 | 50.75% |
| BASEMENTAREA_MODE | 179943 | 58.52% |
| YEARS_BUILD_MODE | 204488 | 66.50% |
| COMMON_AREA_MODE | 214865 | 69.87% |
| ELEVATORS_MODE | 163891 | 53.30% |
| ENTRANCES_MODE | 154828 | 50.35% |
| FLOORSMAX_MODE | 153020 | 49.76% |
| FLOORSMIN_MODE | 208642 | 67.85% |
| LANDAREA_MODE | 182590 | 59.38% |
| LIVINGAPARTMENTS_MODE | 210199 | 68.35% |
| LIVINGAREA_MODE | 154350 | 50.19% |
| NONLIVINGAPARTMENTS_MODE | 213514 | 69.43% |
| NONLIVINGAREA_MODE | 169682 | 55.18% |
| APARTMENTS_MEDIAN | 156061 | 50.75% |
| BASEMENTAREA_MEDIAN | 179943 | 58.52% |
| YEARS_BUILD_MEDIAN | 204488 | 66.50% |
| COMMON_AREA_MEDIAN | 214865 | 69.87% |
| ELEVATORS_MEDIAN | 163891 | 53.30% |
| ENTRANCES_MEDIAN | 154828 | 50.35% |
| FLOORSMAX_MEDIAN | 153020 | 49.76% |
| FLOORSMIN_MEDIAN | 208642 | 67.85% |
| LANDAREA_MEDIAN | 182590 | 59.38% |
| LIVINGAPARTMENTS_MEDIAN | 210199 | 68.35% |
| LIVINGAREA_MEDIAN | 154350 | 50.19% |
| NONLIVINGAPARTMENTS_MEDIA | 213514 | 69.43% |
| NONLIVINGAREA_MEDIAN | 169682 | 55.18% |
| FONDKAPREMONT_MODE | 210295 | 68.39% |
| HOUSETYPE_MODE | 154297 | 50.18% |
| WALLSMATERIAL_MODE | 156341 | 50.84% |

As they are irrelevant columns for doing our analysis, ALL OF THE FOLLOWING COLUMN NAMES NEED TO BE DROPPED DOWN.



| Column name | tal number of null val | rcentage of null valu |
|---|---|---|
| FLAG_MOBIL | 1 | 0.00% |
| FLAG_EMPLOY_PHONE | 55387 | 18.01% |
| FLAG_WORK_PHONE | 0 | 0.00% |
| FLAG_CONT_MOBILE | 0 | 0.00% |
| FLAG_PHONE | 0 | 0.00% |
| FLAG_EMAIL | 0 | 0.00% |
| CNT_FAMILY_MEMBERS | 2 | 0.00% |
| REGION_RATING_CLENT | 0 | 0.00% |
| REGION_RATING_CLENT_W_CI | 0 | 0.00% |
| EXT_SOURCE_3 | 60965 | 19.83% |
| YEAR_BEGINEXPLUATATION_A | 150008 | 48.78% |
| YEAR_BEGINEXPLUATATION_M | 150007 | 48.78% |
| YEAR_BEGINEXPLUATATION_M | 150007 | 48.78% |
| TOTAL_AREA_MODE | 148431 | 48.27% |
| EMERGENCYSTATE_MODE | 145755 | 47.40% |
| DAYS_LAST_PHONE_CHANGE | 1 | 0.00% |
| FLAG DOC 2 | 0 | 0.00% |
| FLAG DOC 3 | 0 | 0.00% |
| FLAG DOC 4 | 0 | 0.00% |
| FLAG DOC 5 | 0 | 0.00% |
| FLAG DOC 6 | 0 | 0.00% |
| FLAG DOC 7 | 0 | 0.00% |
| FLAG DOC 8 | 0 | 0.00% |
| FLAG DOC 9 | 0 | 0.00% |
| FLAG DOC 10 | 0 | 0.00% |
| FLAG DOC 11 | 0 | 0.00% |
| FLAG DOC 12 | 0 | 0.00% |
| FLAG DOC 13 | 0 | 0.00% |
| FLAG DOC 14 | 0 | 0.00% |
| FLAG DOC 15 | 0 | 0.00% |
| FLAG DOC 16 | 0 | 0.00% |
| FLAG DOC 17 | 0 | 0.00% |
| FLAG DOC 18 | 0 | 0.00% |
| FLAG DOC 19 | 0 | 0.00% |
| FLAG DOC 20 | 0 | 0.00% |
| FLAG DOC 21 | 0 | 0.00% |

Replacing blanks in the Application Dataset's AMT_ANNUTIY column with the median value of AMT_ANNUITY since the column contains outliers.

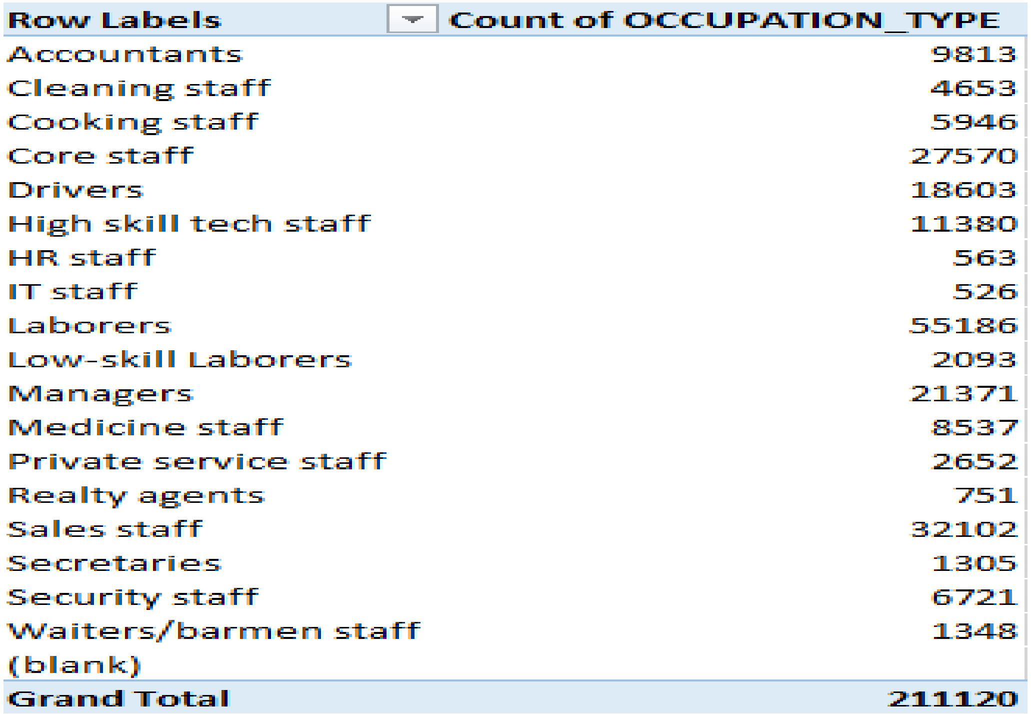


AMT_ANNUITY

| | |
|---|---|
| Median | 24903 |

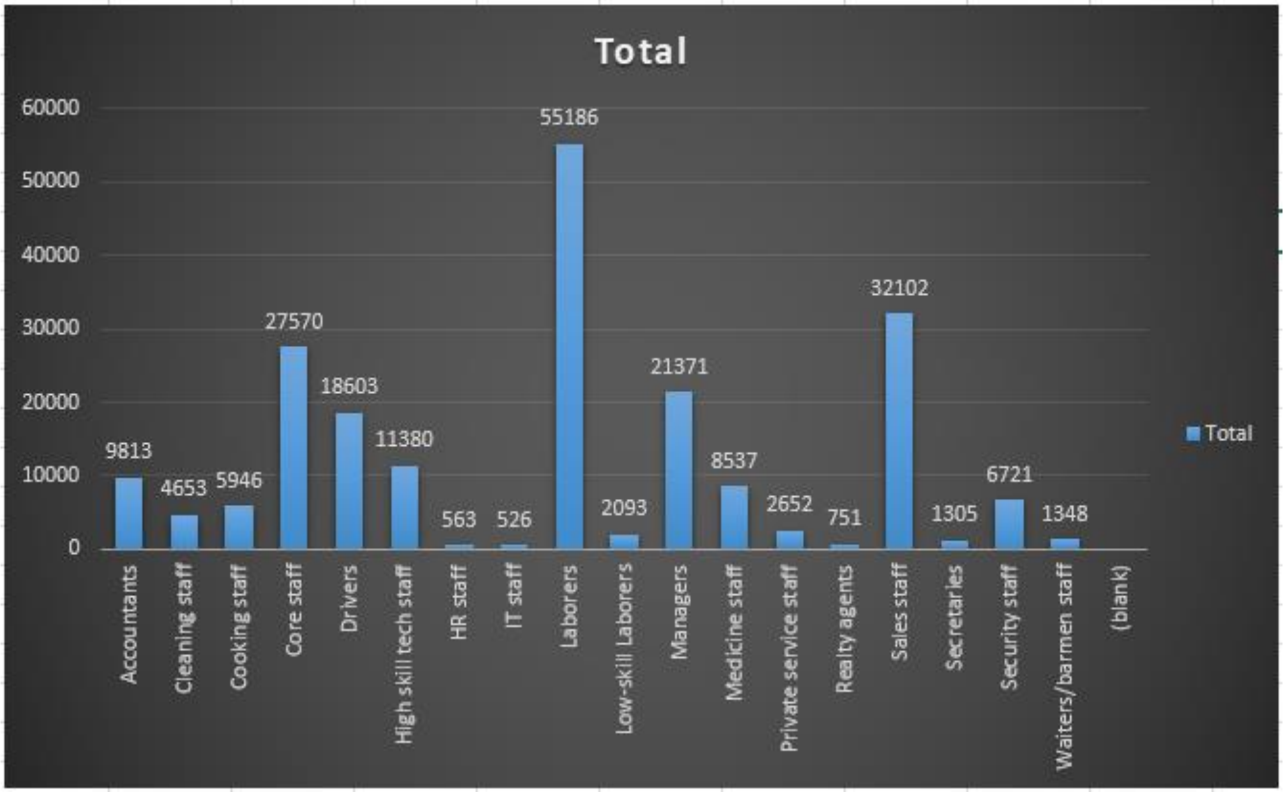
Replacing Blanks with Median

Filling in blanks in the Application Dataset's Occupation_Type column with the categorical variable with the highest frequency

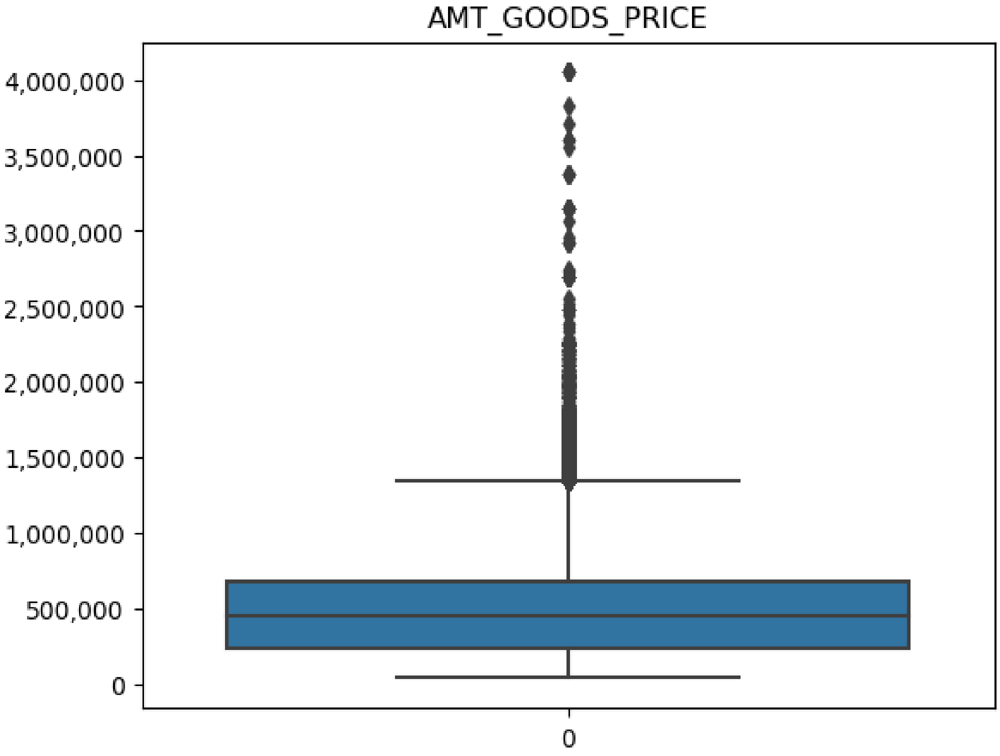


| Row Labels | Count of OCCUPATION_TYPE |
|---|---|
| Accountants | 9813 |
| Cleaning staff | 4653 |
| Cooking staff | 5946 |
| Core staff | 27570 |
| Drivers | 18603 |
| High skill tech staff | 11380 |
| HR staff | 563 |
| IT staff | 526 |
| Laborers | 55186 |
| Low-skill Laborers | 2093 |
| Managers | 21371 |
| Medicine staff | 8537 |
| Private service staff | 2652 |
| Realty agents | 751 |
| Sales staff | 32102 |
| Secretaries | 1305 |
| Security staff | 6721 |
| Waiters/barmen staff | 1348 |
| (blank) | |
| **Grand Total** | **211120** |

*Highest occurring categorical variable is 'Laborers'*



Blanks in the Application Dataset's AMT_GOODS_PRICE column should be replaced with the median of AMT_GOODS_PRICE since the column contains outliers.

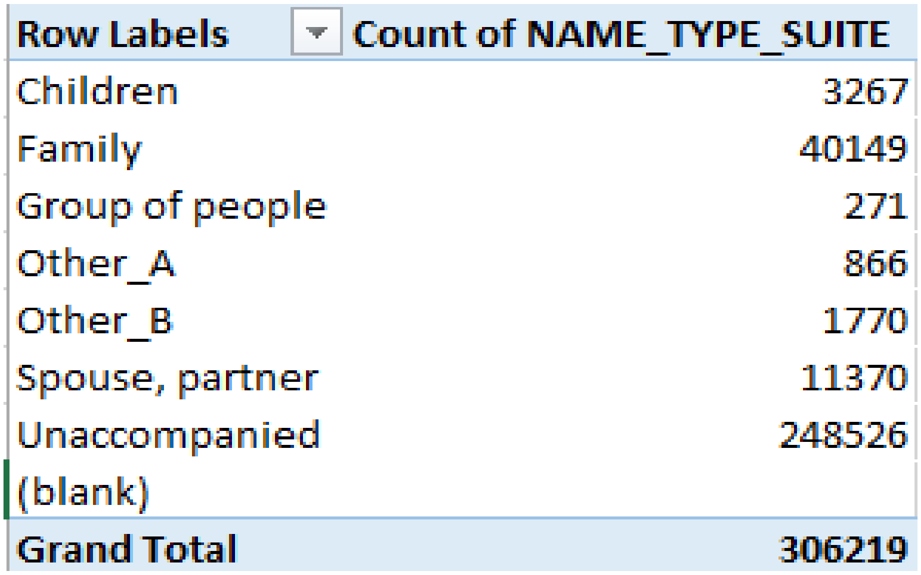


AMT_GOODS_PRICE

| | |
|---|---|
| Median | 450000 |

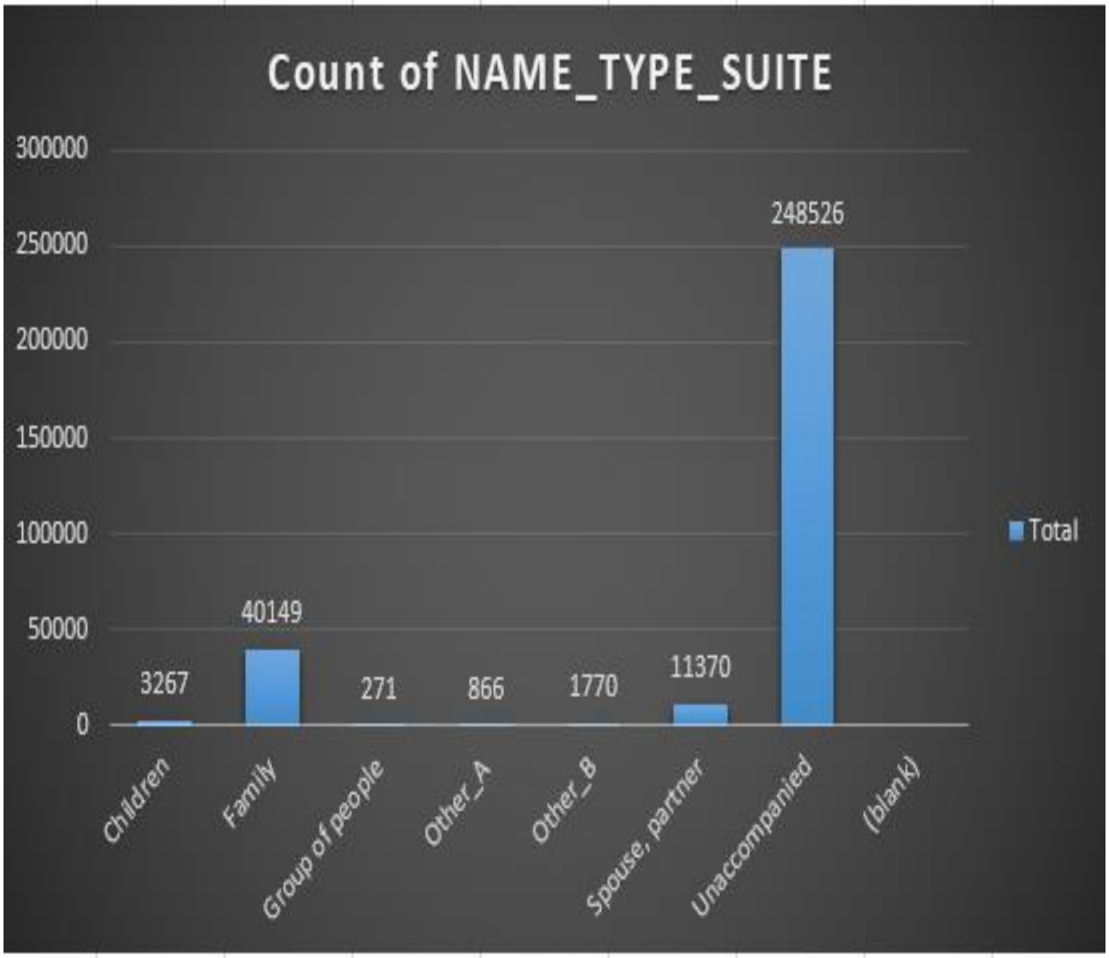
Replacing Blanks with Median

Filling in blanks in the Name_Type_Suite column of the Application Dataset with the categorical variable with the highest frequency

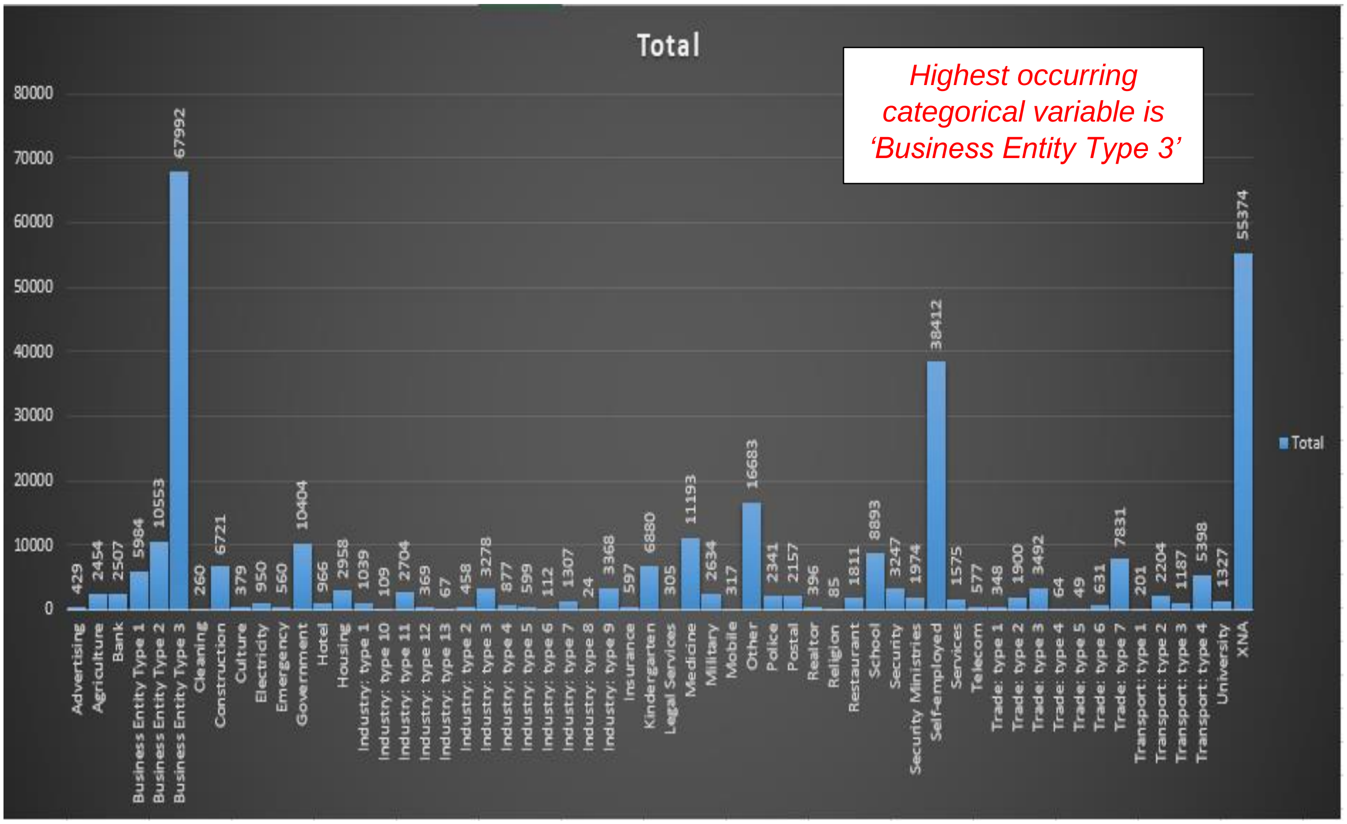


| Row Labels | Count of NAME_TYPE_SUITE |
|---|---|
| Children | 3267 |
| Family | 40149 |
| Group of people | 271 |
| Other_A | 866 |
| Other_B | 1770 |
| Spouse, partner | 11370 |
| Unaccompanied | 248526 |
| (blank) | |
| **Grand Total** | **306219** |

*Highest occurring categorical variable is 'Unaccompanied'*

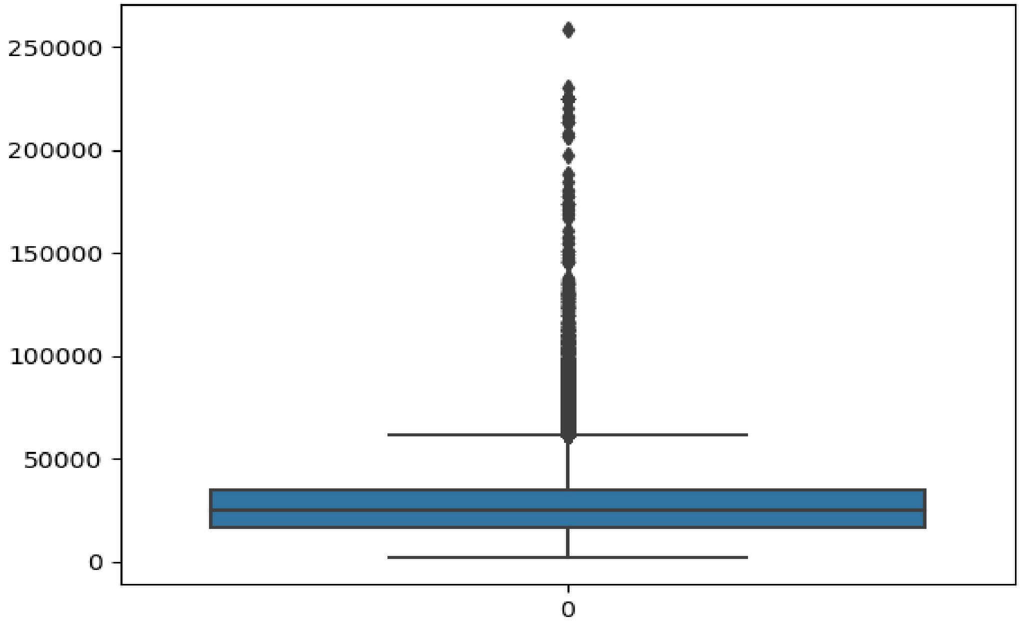


## Filling in blanks in the Application Dataset's Organization_type column with the most frequent categorical variable

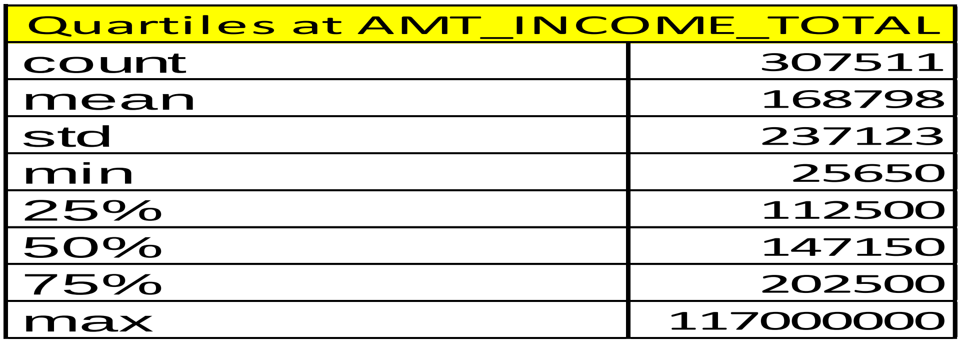


*Highest occurring categorical variable is ‘Business Entity Type 3’*

## Application Dataset – Outliers

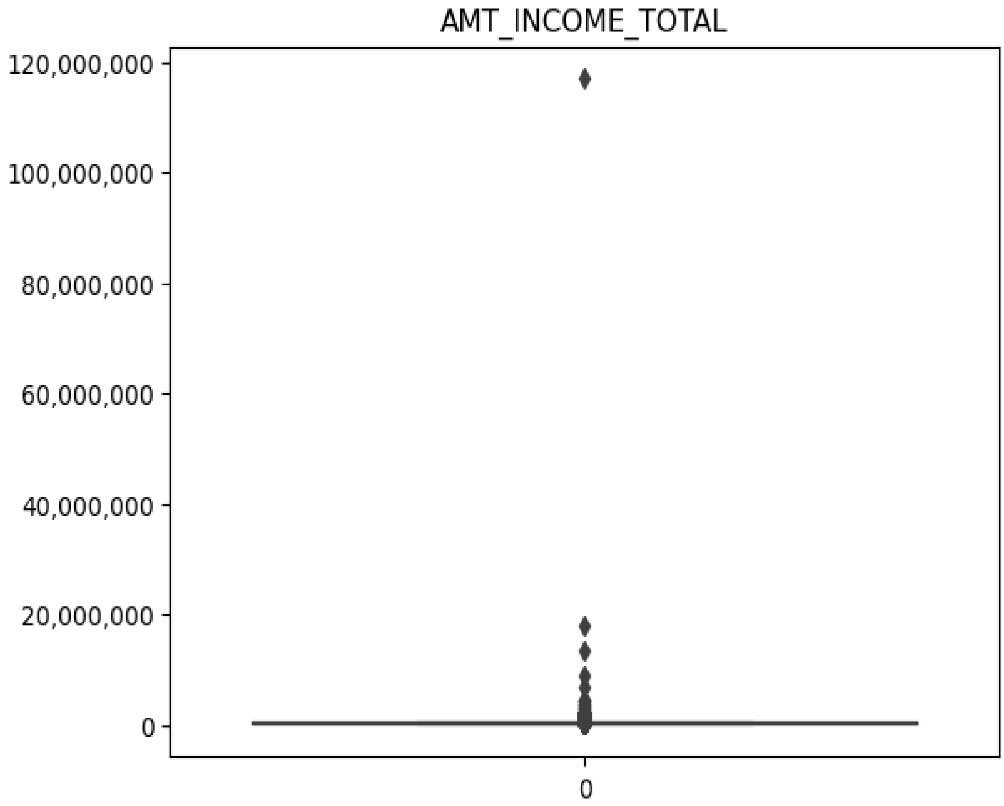


*AMT_ANNUITY, which is more than 250000, is the first outlier.24903 is used in place of this anomaly.*

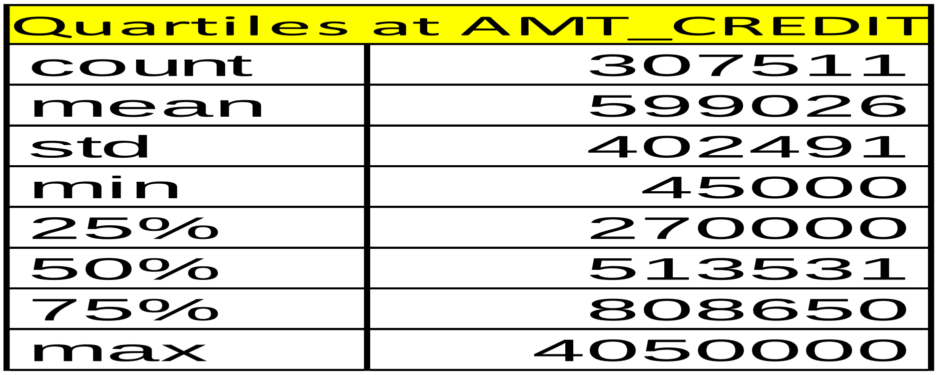


| Quartiles at AMT_INCOME_TOTAL | |
|---|---|
| count | 307511 |
| mean | 168798 |
| std | 237123 |
| min | 25650 |
| 25% | 112500 |
| 50% | 147150 |
| 75% | 202500 |
| max | 117000000 |

Here, we can see that the existence of outliers causes a significant difference between the 25%, 50%, and 75% quartiles.We won't get rid of the outliers though because overall income differs from person to person.



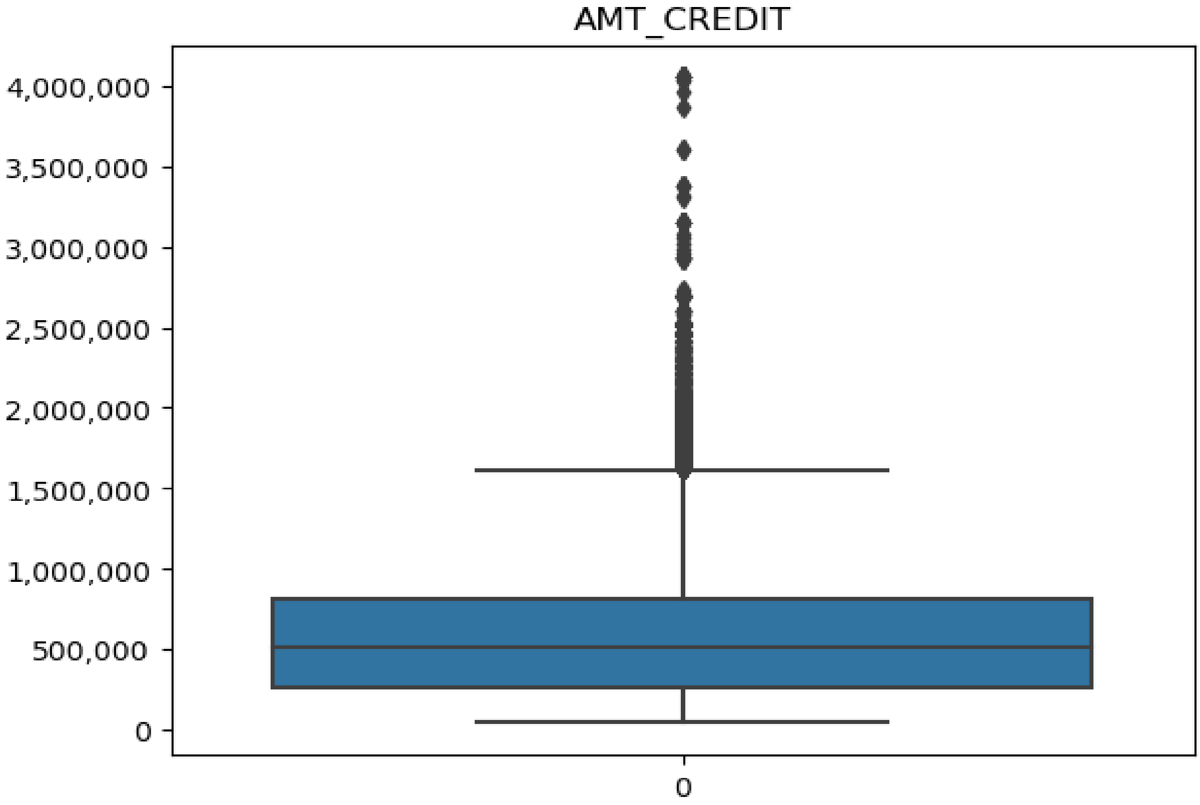
*outliers at extreme points i.e. max 117000000*

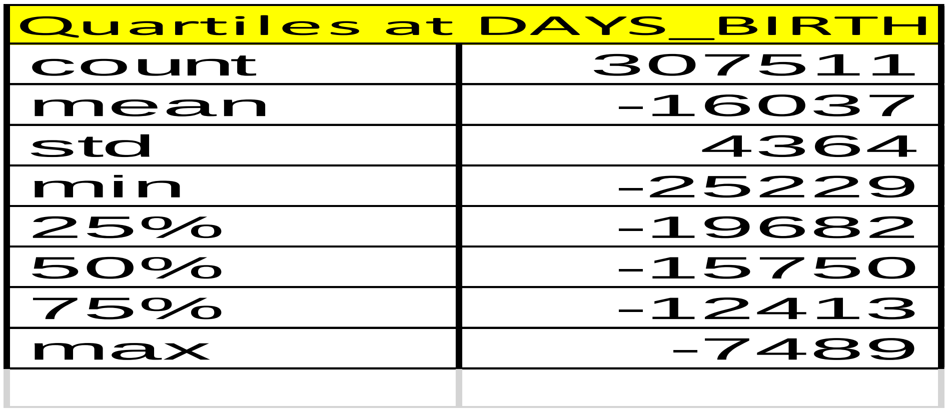


| Quartiles at AMT_CREDIT | |
|---|---|
| count | 307511 |
| mean | 599026 |
| std | 402491 |
| min | 45000 |
| 25% | 270000 |
| 50% | 513531 |
| 75% | 808650 |
| max | 4050000 |

It is evident from the figure that outliers are located on the 98% and close to the maximum side of the box plot.Additionally, there is a large discrepancy between the 75% quartile and the maximum value, which is caused by the existence of outliers.

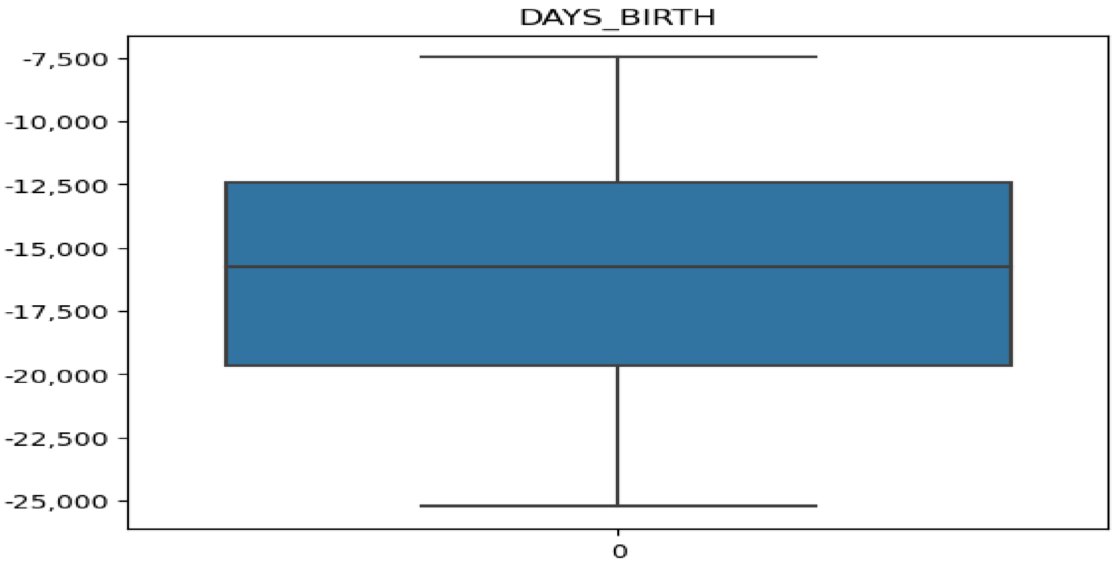
But because each person receives a different amount of credit, we won't eliminate the outliers.

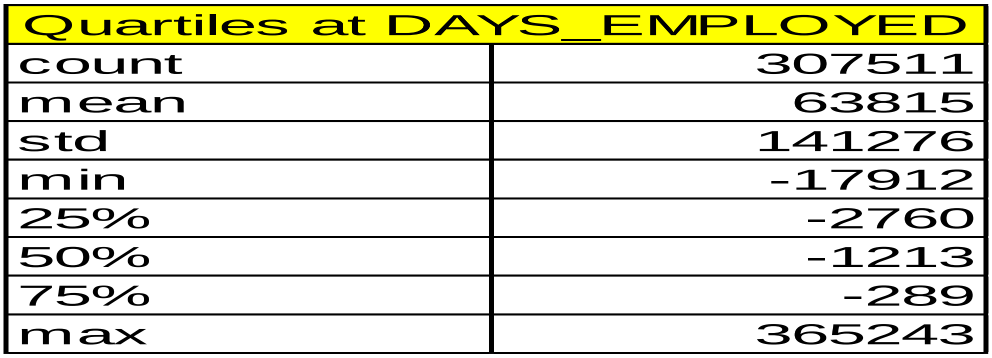




| Quartiles at DAYS_BIRTH | |
|---|---|
| count | 307511 |
| mean | -16037 |
| std | 4364 |
| min | -25229 |
| 25% | -19682 |
| 50% | -15750 |
| 75% | -12413 |
| max | -7489 |

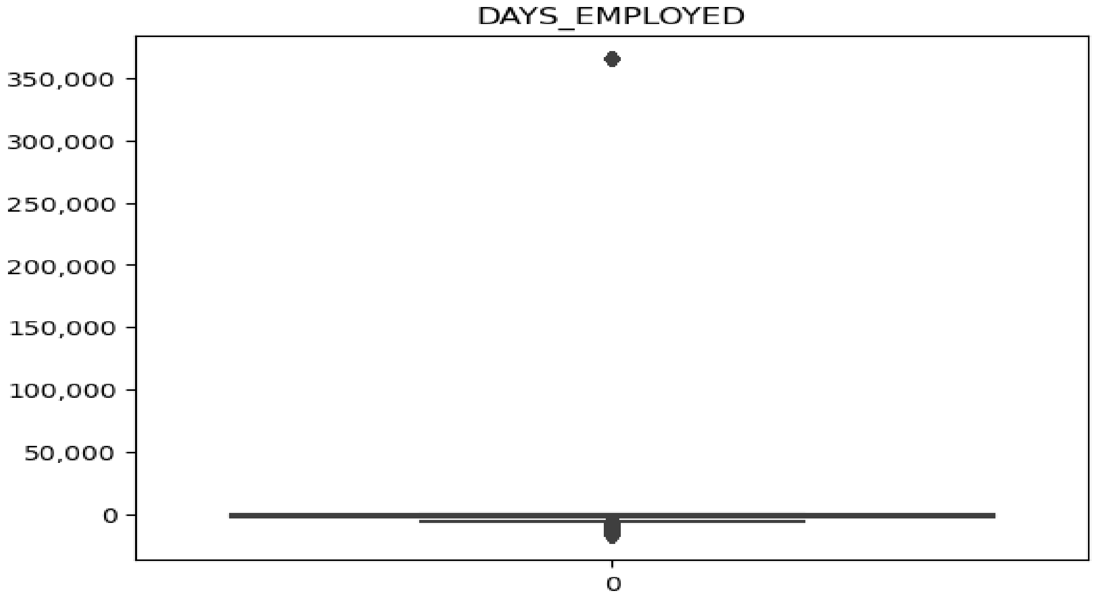
As seen from the boxplot it is clear that there are no outliers The data of DAYS_BIRTH is well distributed.





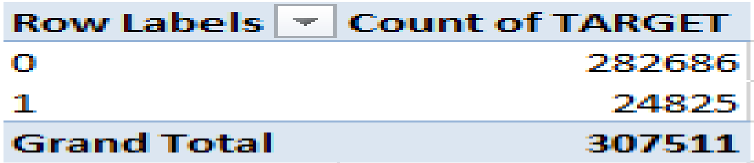
| Quartiles at DAYS_EMPLOYED | |
|---|---|
| count | 307511 |
| mean | 63815 |
| std | 141276 |
| min | -17912 |
| 25% | -2760 |
| 50% | -1213 |
| 75% | -289 |
| max | 365243 |

There is only one outlier, which is + or 365243; the median value is -1213.00.

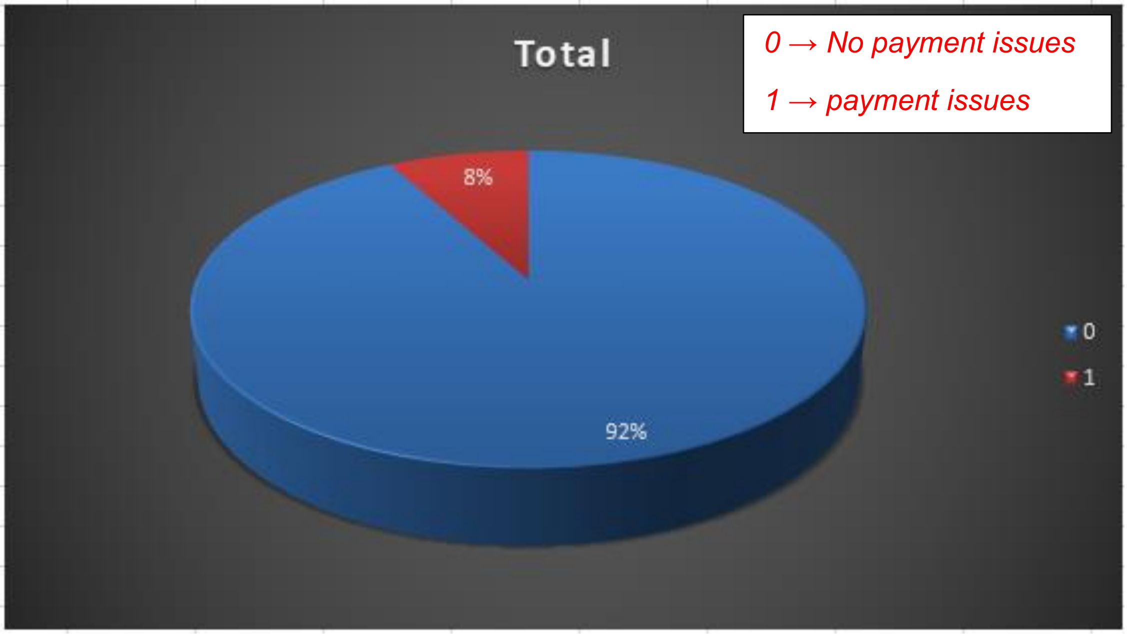


## Application Dataset –Univariate Analysis

### TARGET VARIABLE



| Row Labels | Count of TARGET |
|---|---|
| 0 | 282686 |
| 1 | 24825 |
| Grand Total | 307511 |

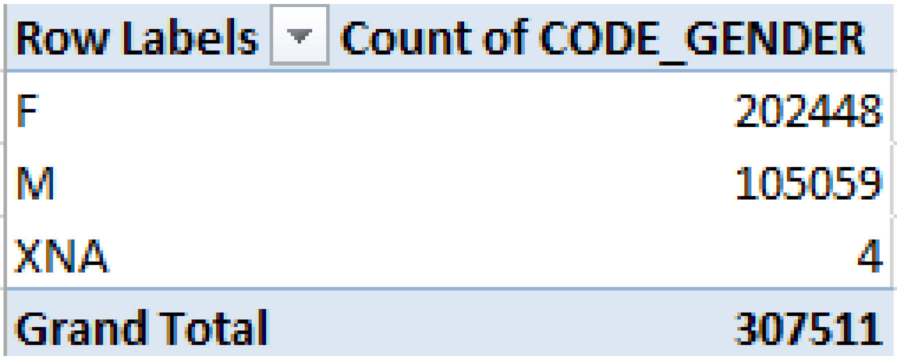


*0 → No payment issues*

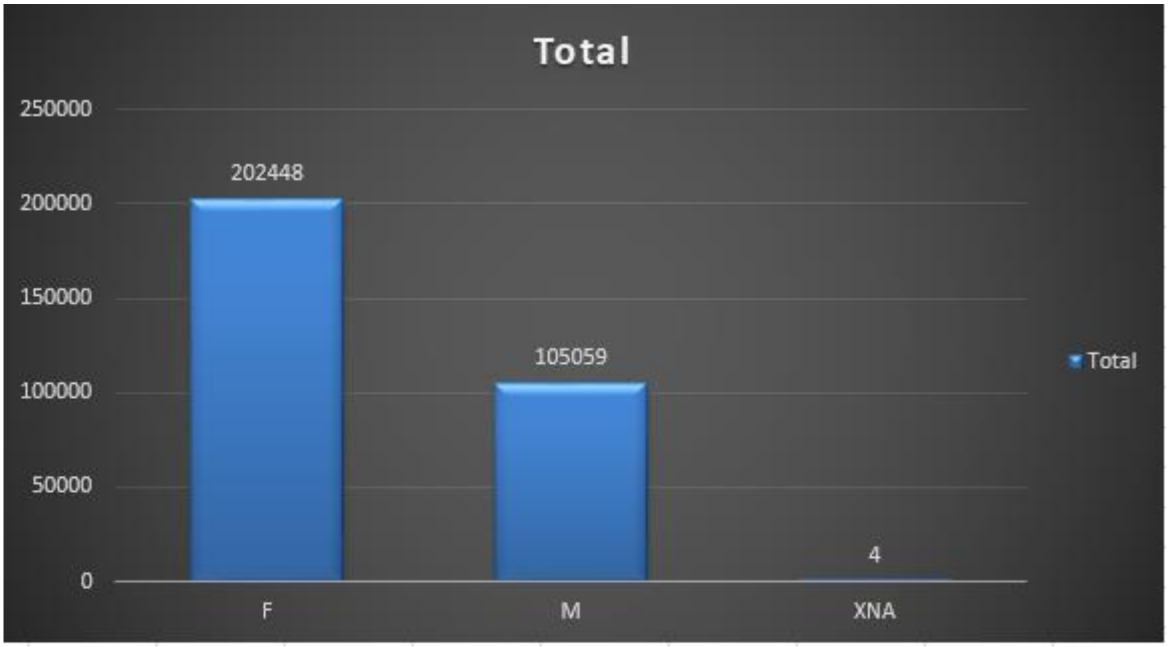
*1 → payment issues*

Nearly 92% of all clients had the target variable, according to the Target Variable Pie Chart.Whereas 8% of clients had some sort of issue upon payment, there were no problems.

### GENDER VARIABLE

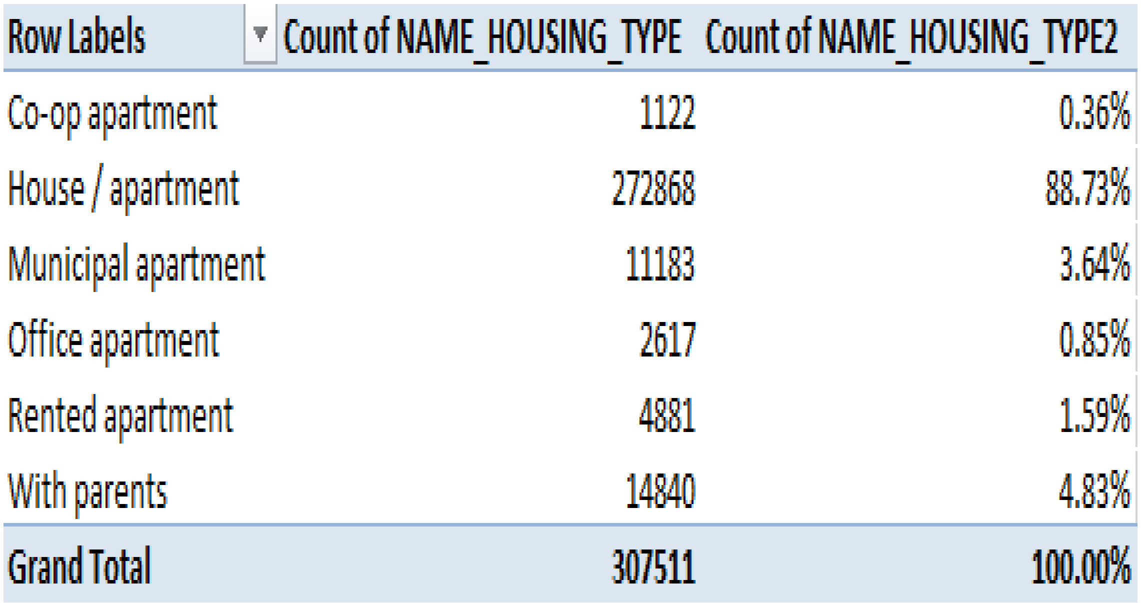


| Row Labels | Count of CODE_GENDER |
|---|---|
| F | 202448 |
| M | 105059 |
| XNA | 4 |
| Grand Total | 307511 |

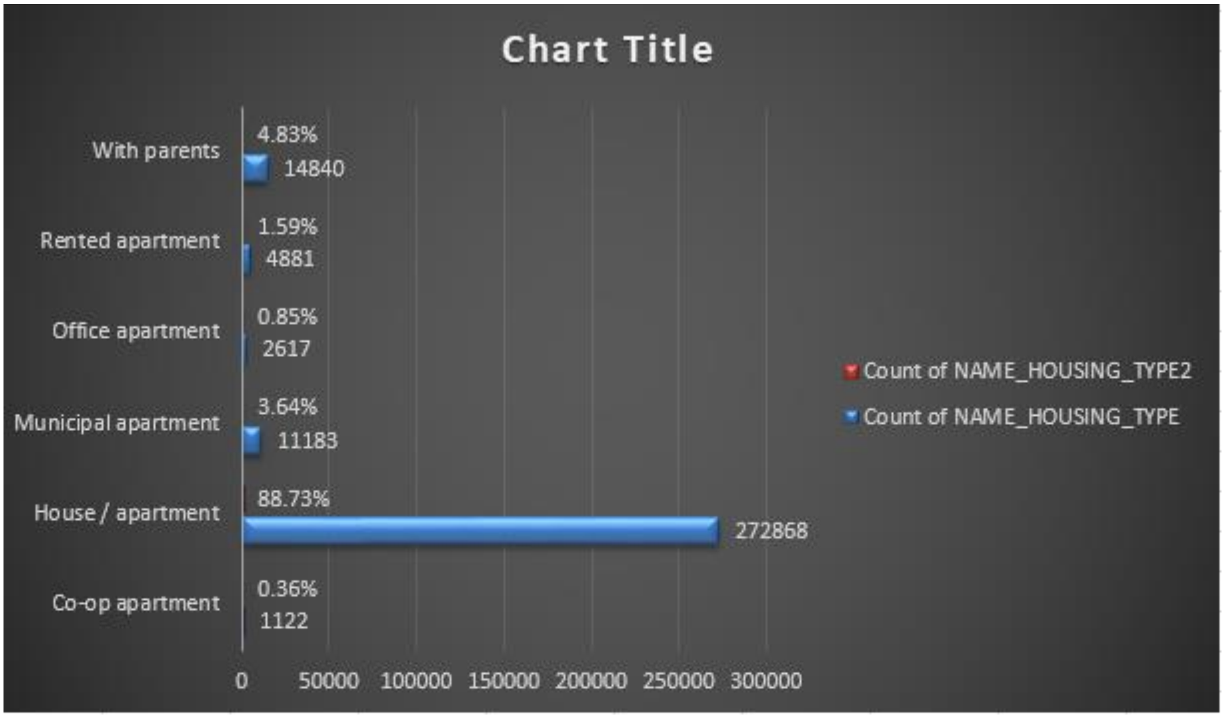


We may deduce that around 66% of clients are female and 34% are male based on the GENDER_VARIABLE pie chart.The four applicants' XNA gender designations can be disregarded.

NAME_HOUSING_TYPE

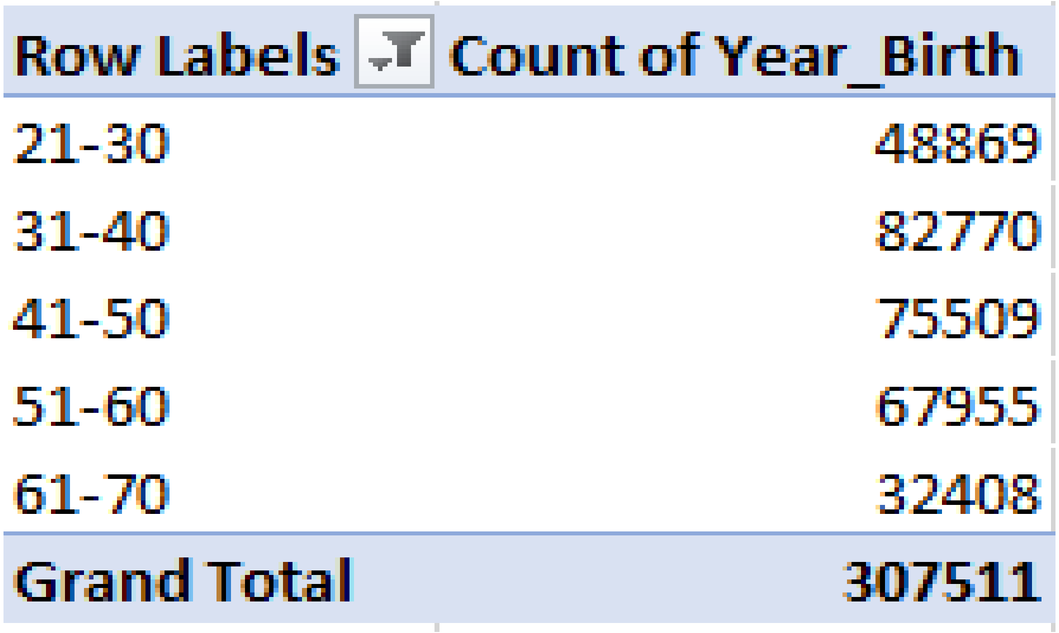


| Row Labels | Count of NAME_HOUSING_TYPE | Count of NAME_HOUSING_TYPE2 |
|---|---|---|
| Co-op apartment | 1122 | 0.36% |
| House / apartment | 272868 | 88.73% |
| Municipal apartment | 11183 | 3.64% |
| Office apartment | 2617 | 0.85% |
| Rented apartment | 4881 | 1.59% |
| With parents | 14840 | 4.83% |
| **Grand Total** | **307511** | **100.00%** |

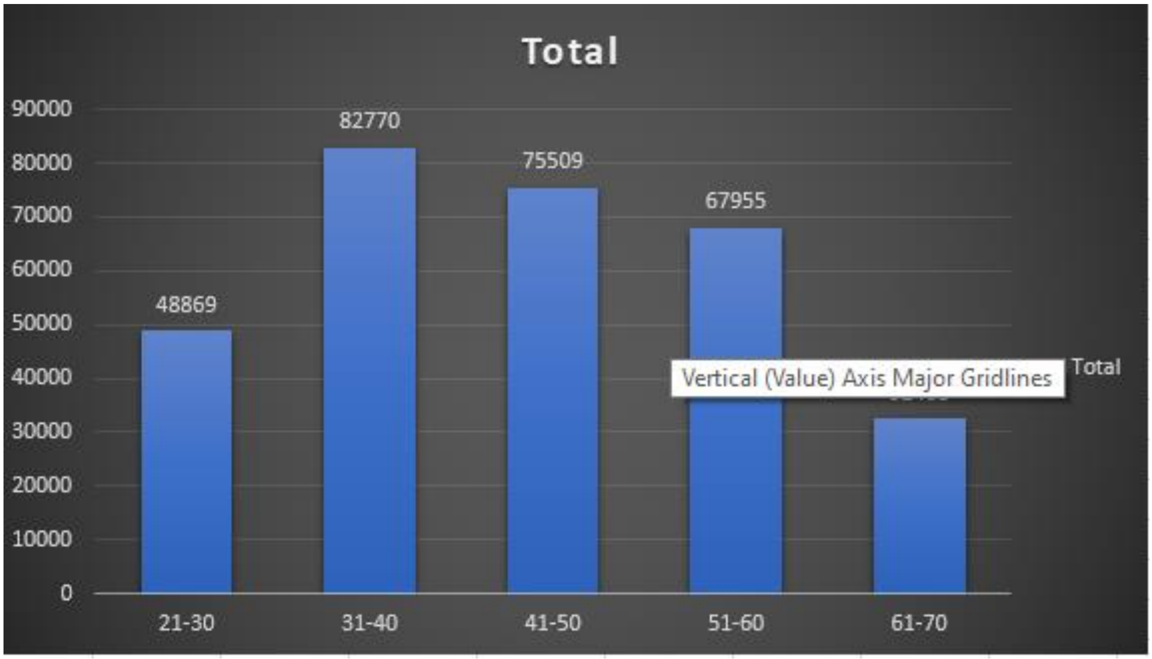


Based on the count and percentage bar graphs.The bank can focus on those populations that do not have their own apartments, such as those who live in co-ops, municipal apartments, rented apartments, and those who live with their parents.

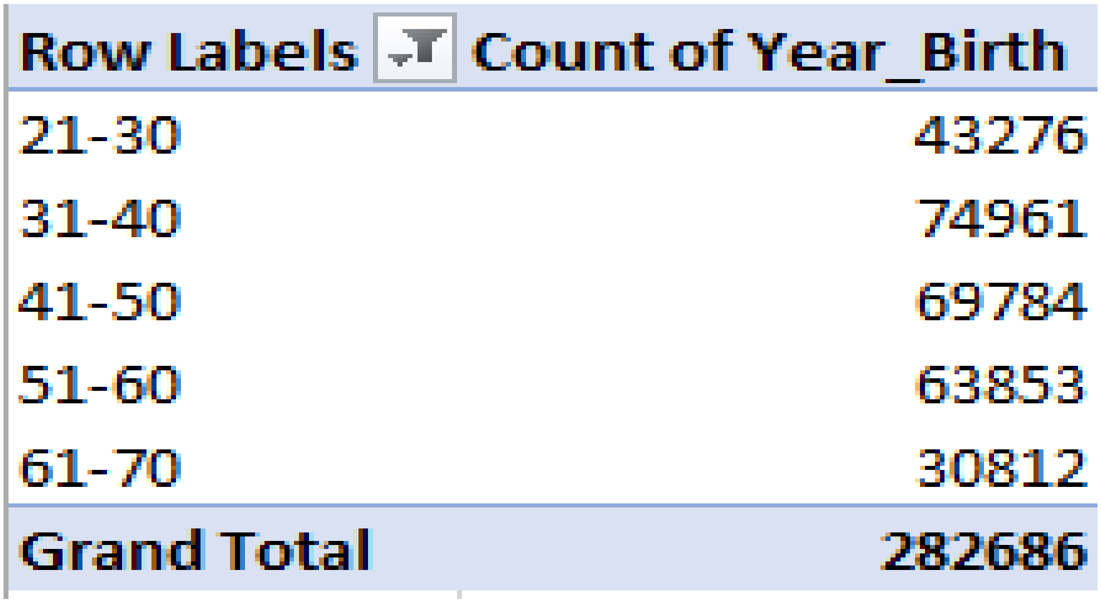
AGE GROUP



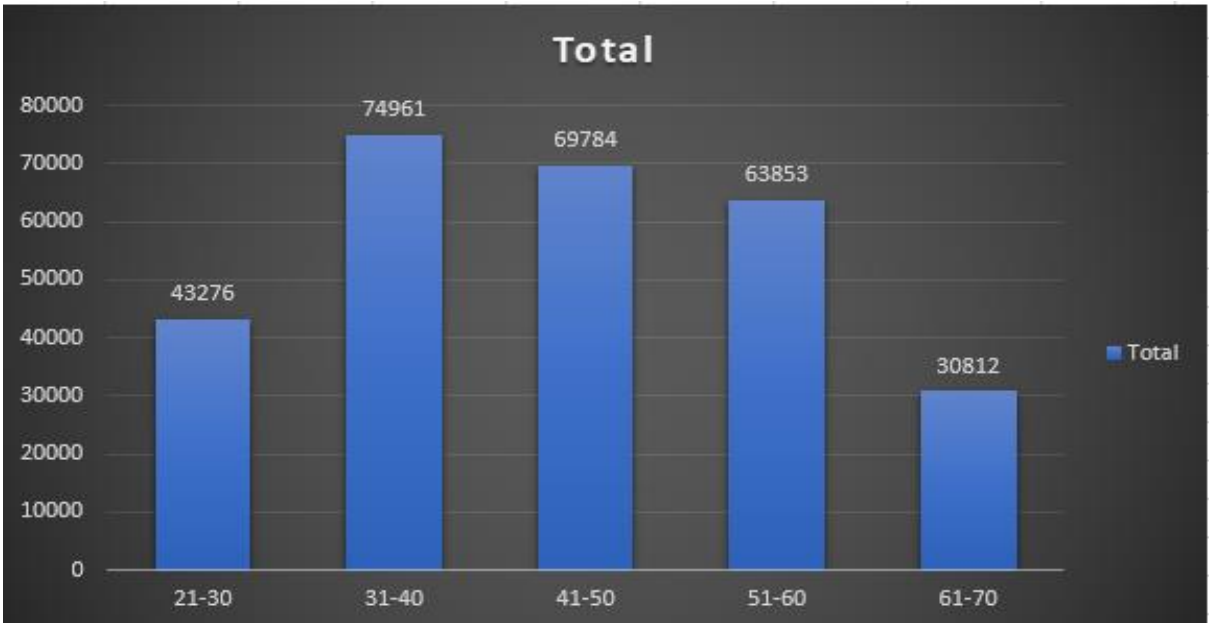
| Row Labels | Count of Year_Birth |
|---|---|
| 21-30 | 48869 |
| 31-40 | 82770 |
| 41-50 | 75509 |
| 51-60 | 67955 |
| 61-70 | 32408 |
| **Grand Total** | **307511** |



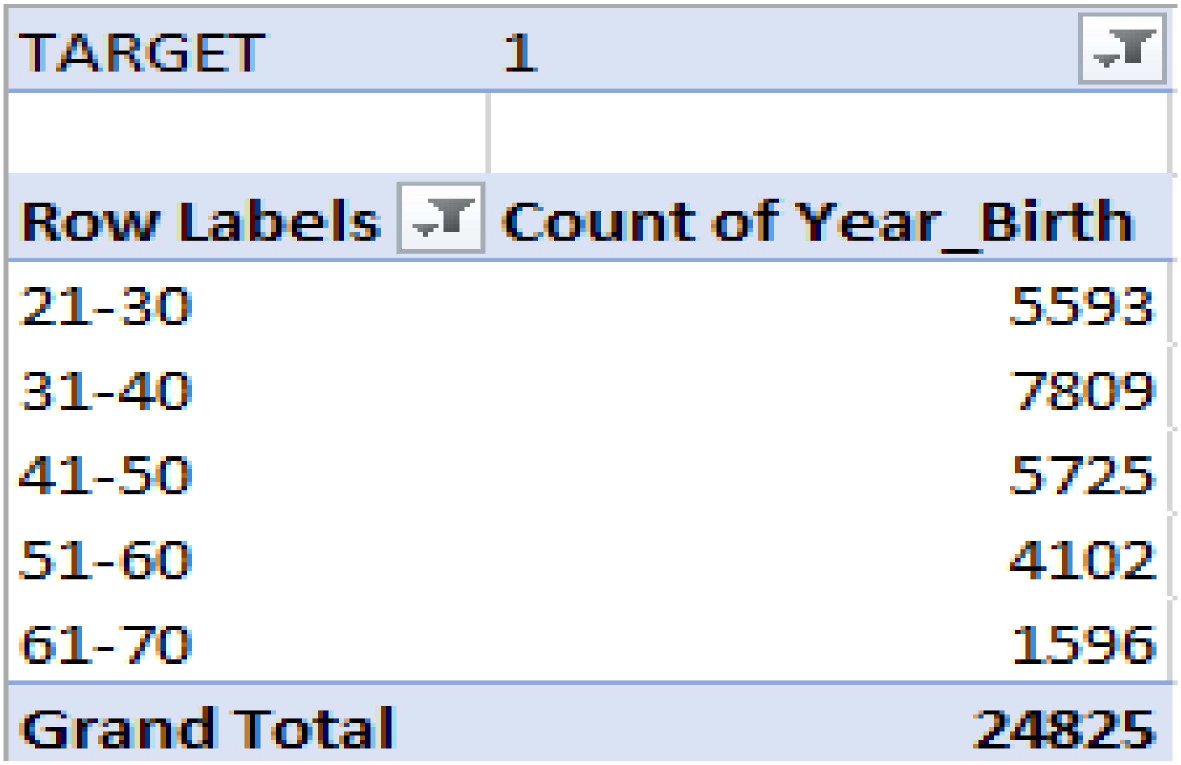
We may deduce that the majority of applicants fall into the Age Group "31-40" from the nearby pub layout.



| Row Labels | Count of Year_Birth |
|---|---|
| 21-30 | 43276 |
| 31-40 | 74961 |
| 41-50 | 69784 |
| 51-60 | 63853 |
| 61-70 | 30812 |
| **Grand Total** | **282686** |

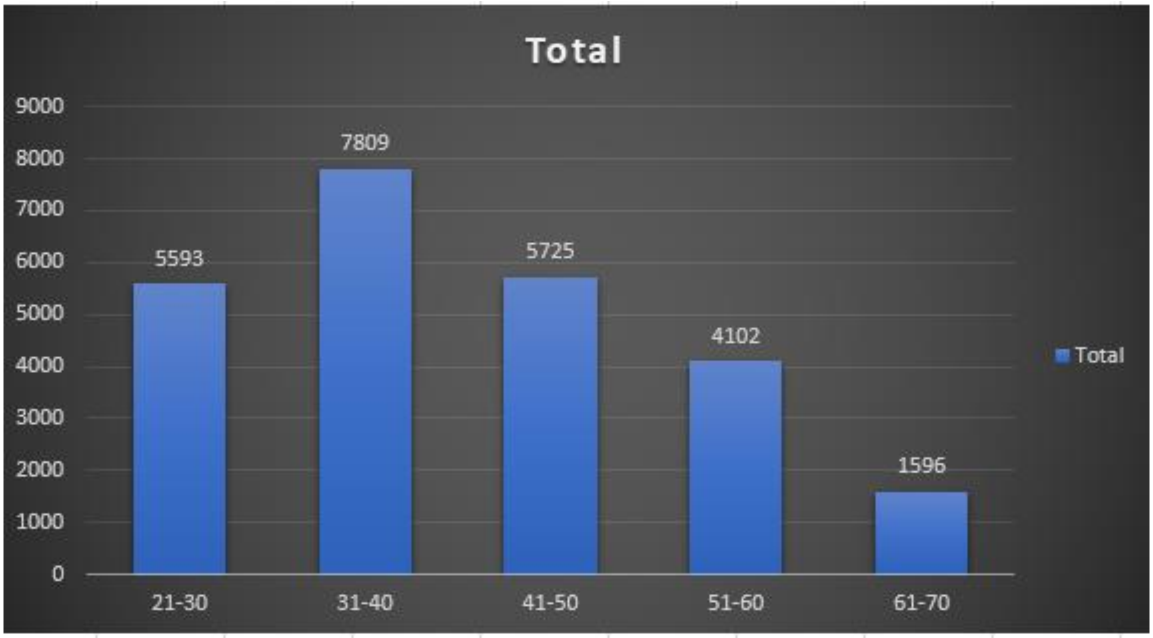


We may deduce from the adjacent bar plot that customers/applicants in the age group "31-40" have the biggest number when it comes to making or returning payments to banks.



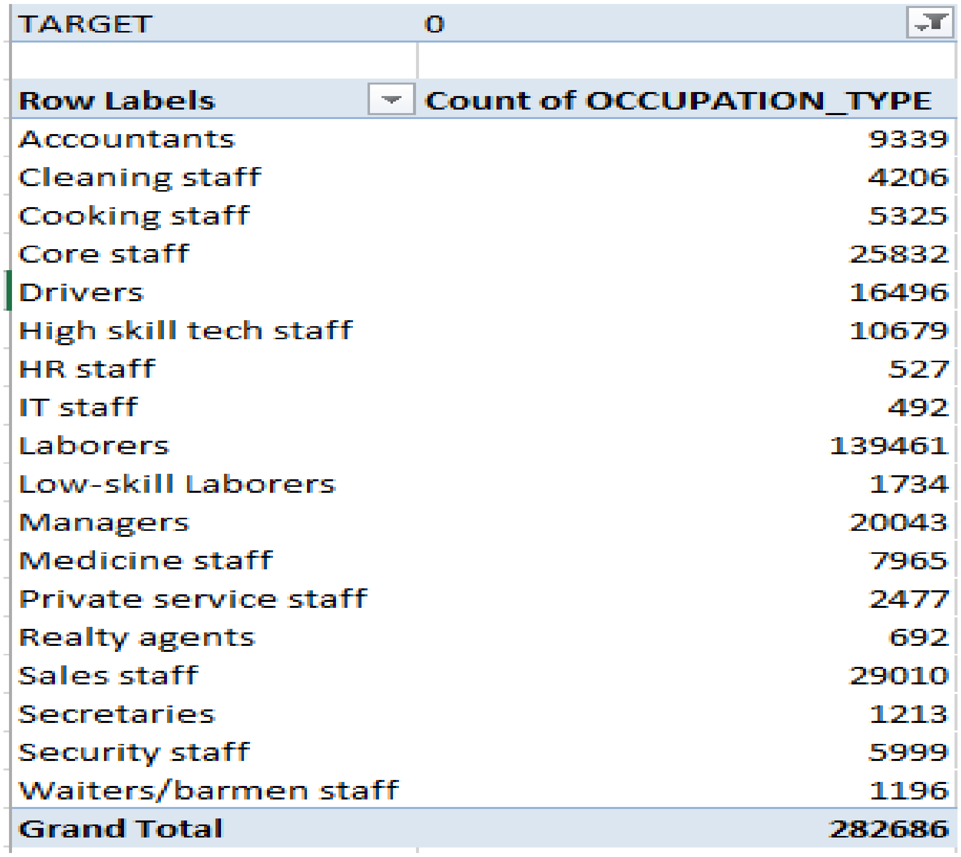
TARGET 1

| Row Labels | Count of Year_Birth |
|---|---|
| 21-30 | 5593 |
| 31-40 | 7809 |
| 41-50 | 5725 |
| 51-60 | 4102 |
| 61-70 | 1596 |
| Grand Total | 24825 |

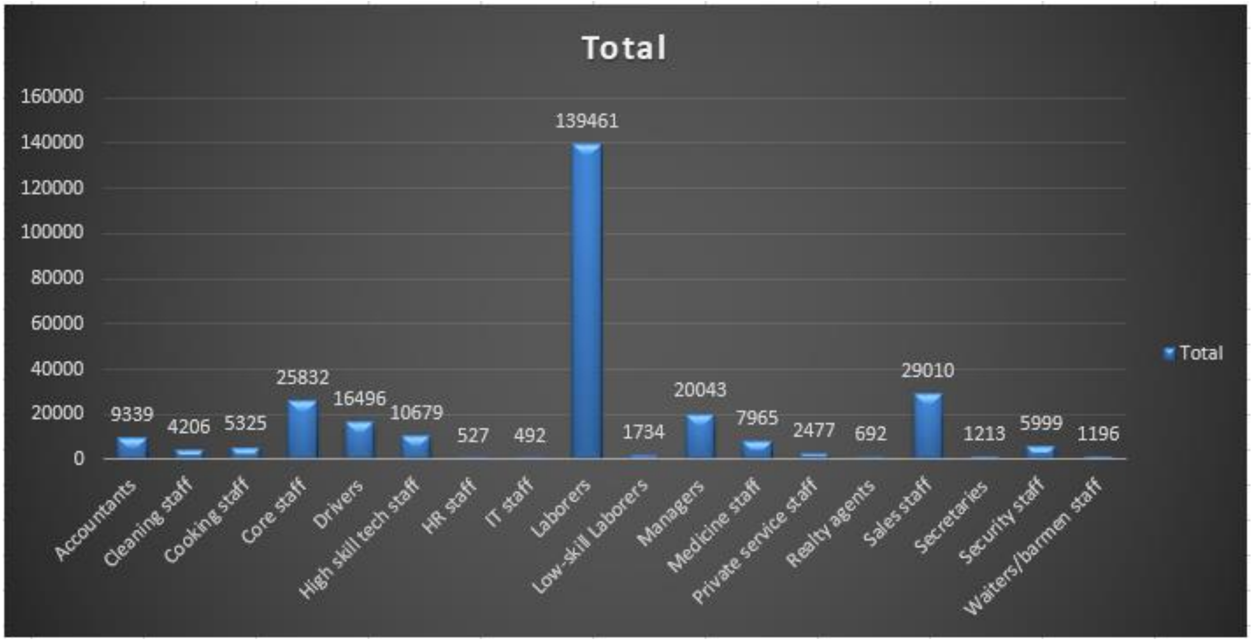


According to the adjacent bar plot, customers/applicants in the age group "31-40" experience the most payment problems while making or returning payments to banks.

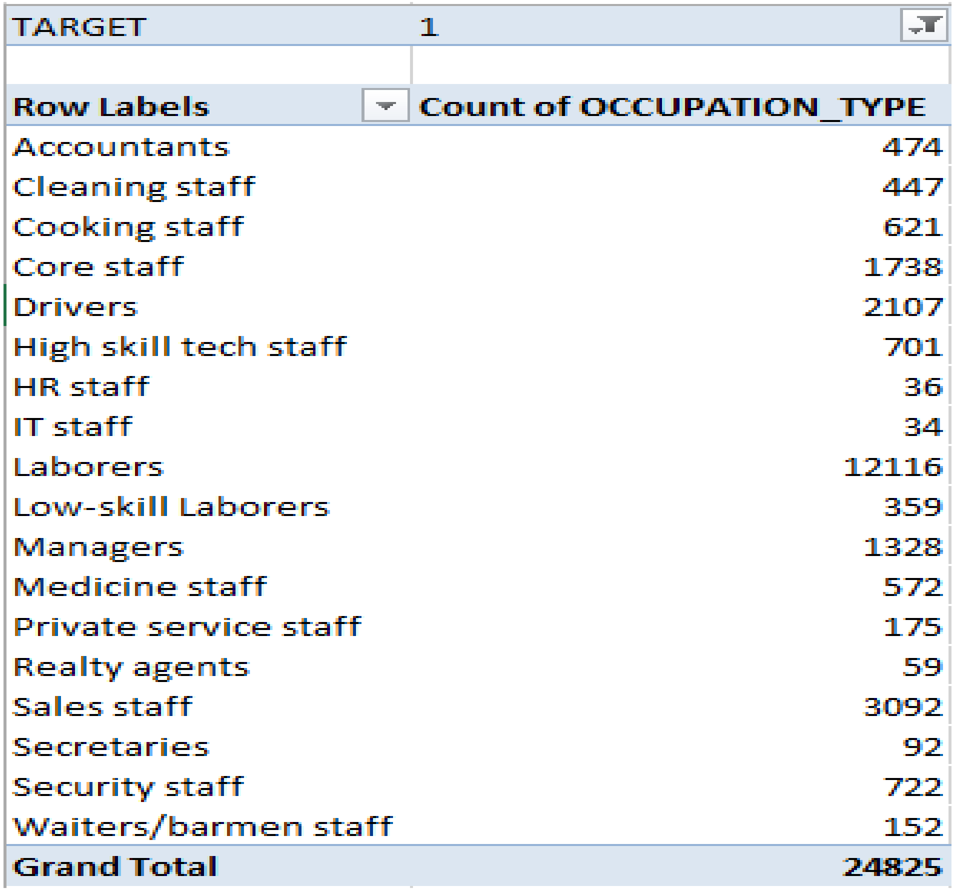
OCCUPATION_TYPE



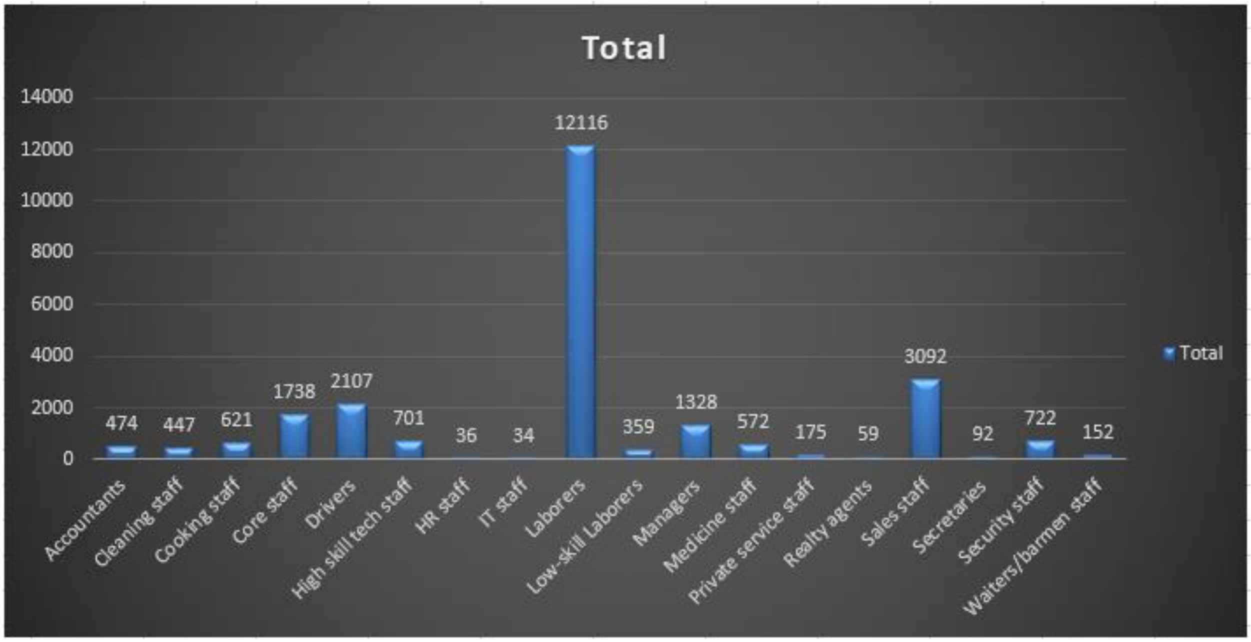
| TARGET | 0 |
|---|---|
| | |
| **Row Labels** | **Count of OCCUPATION_TYPE** |
| Accountants | 9339 |
| Cleaning staff | 4206 |
| Cooking staff | 5325 |
| Core staff | 25832 |
| Drivers | 16496 |
| High skill tech staff | 10679 |
| HR staff | 527 |
| IT staff | 492 |
| Laborers | 139461 |
| Low-skill Laborers | 1734 |
| Managers | 20043 |
| Medicine staff | 7965 |
| Private service staff | 2477 |
| Realty agents | 692 |
| Sales staff | 29010 |
| Secretaries | 1213 |
| Security staff | 5999 |
| Waiters/barmen staff | 1196 |
| **Grand Total** | **282686** |



'Labourers' occupation_type clients have the largest count when it comes to clients with no payment concerns, according to the above bar plot, it can be deduced.

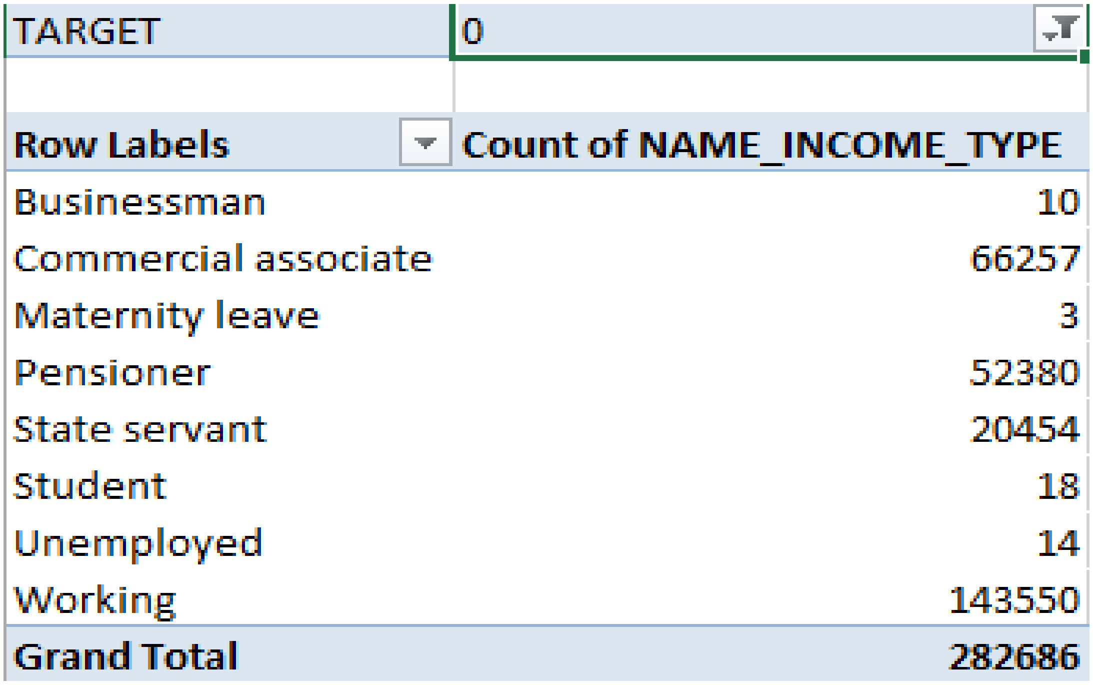


| TARGET | 1 |
|---|---|
| | |
| **Row Labels** | **Count of OCCUPATION_TYPE** |
| Accountants | 474 |
| Cleaning staff | 447 |
| Cooking staff | 621 |
| Core staff | 1738 |
| Drivers | 2107 |
| High skill tech staff | 701 |
| HR staff | 36 |
| IT staff | 34 |
| Laborers | 12116 |
| Low-skill Laborers | 359 |
| Managers | 1328 |
| Medicine staff | 572 |
| Private service staff | 175 |
| Realty agents | 59 |
| Sales staff | 3092 |
| Secretaries | 92 |
| Security staff | 722 |
| Waiters/barmen staff | 152 |
| **Grand Total** | **24825** |

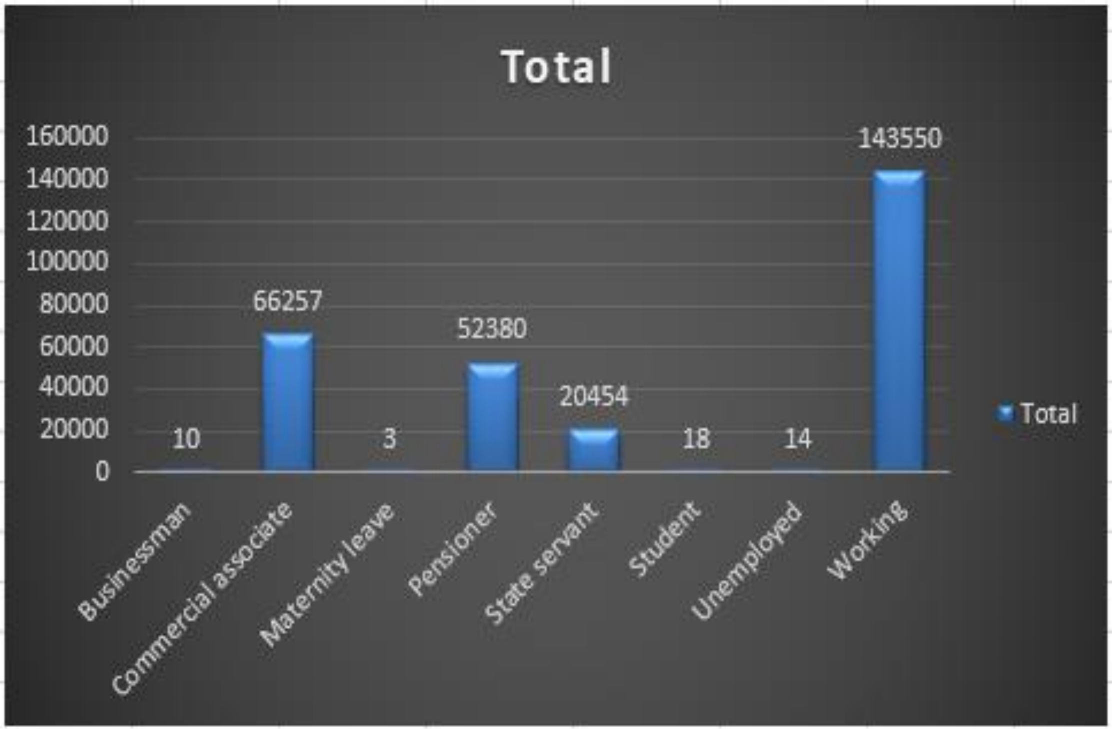


From the above bar plot we can infer that clients with occupation_type 'Laborers' have the highest number of count when it comes to clients with payment issues.

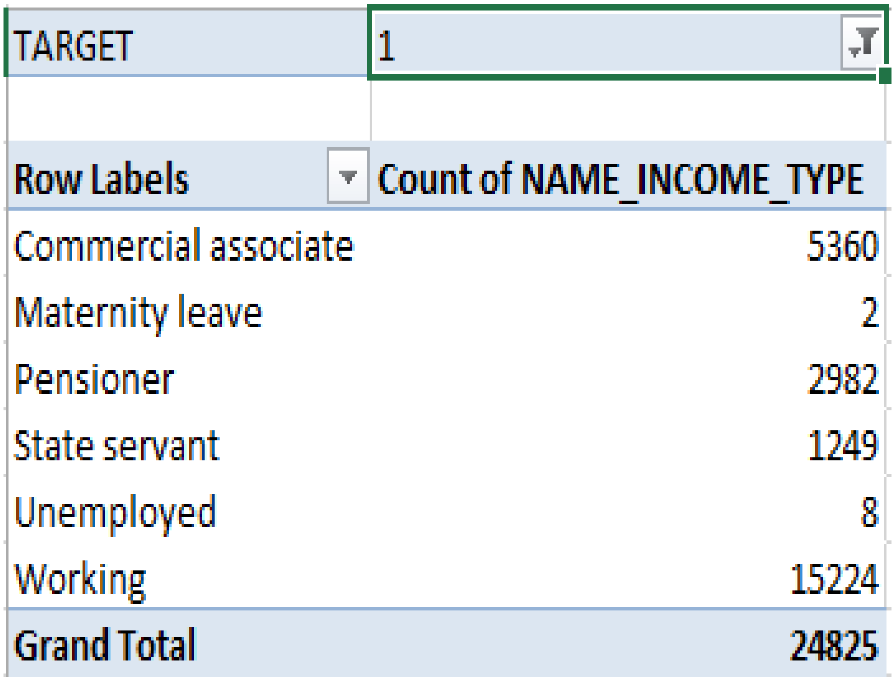
NAME_INCOME_TYPE



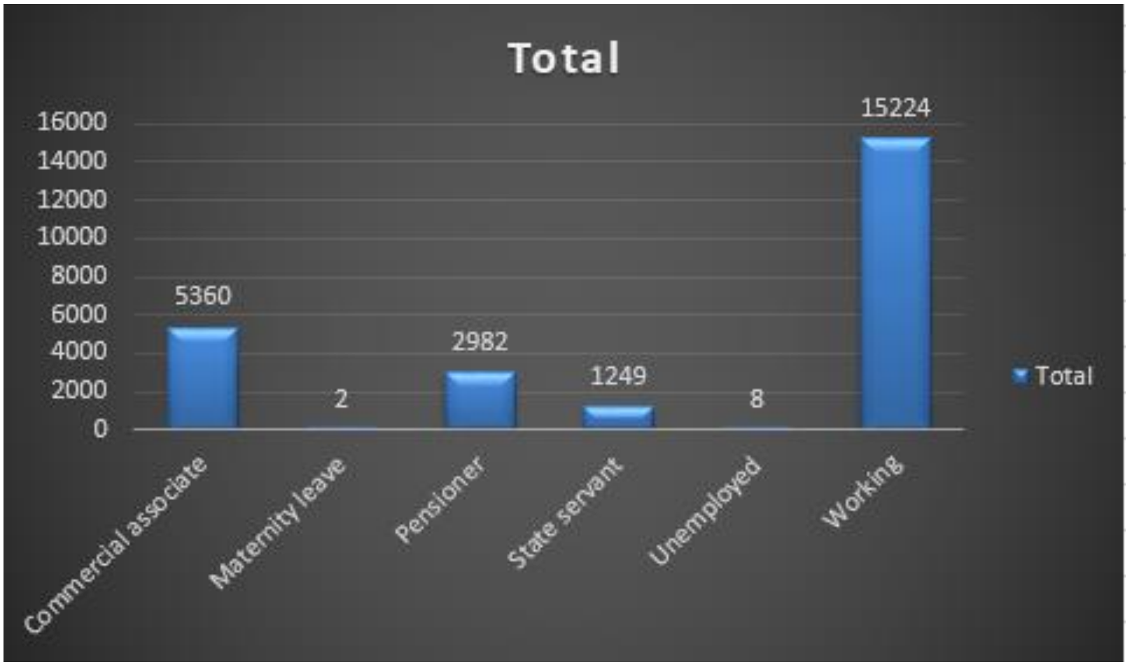
| TARGET | 0 |
|---|---|
| | |
| **Row Labels** | **Count of NAME_INCOME_TYPE** |
| Businessman | 10 |
| Commercial associate | 66257 |
| Maternity leave | 3 |
| Pensioner | 52380 |
| State servant | 20454 |
| Student | 18 |
| Unemployed | 14 |
| Working | 143550 |
| **Grand Total** | **282686** |



The 'WORKING' income_type customers have the largest count of those who have no payment concerns, according to the aforementioned Bar plot.



| TARGET | 1 |
|---|---|
| | |
| **Row Labels** | **Count of NAME_INCOME_TYPE** |
| Commercial associate | 5360 |
| Maternity leave | 2 |
| Pensioner | 2982 |
| State servant | 1249 |
| Unemployed | 8 |
| Working | 15224 |
| **Grand Total** | **24825** |

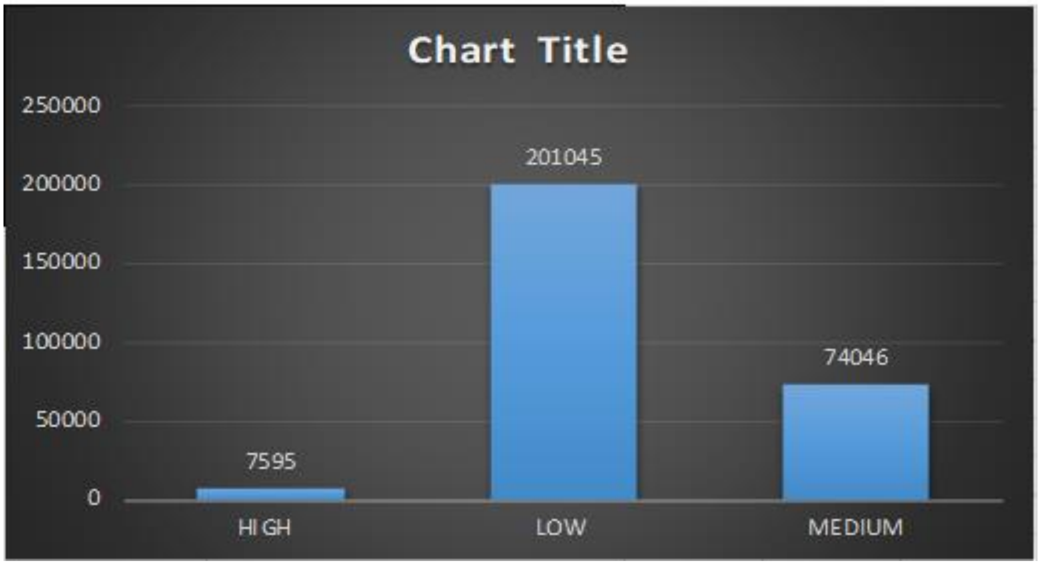


The 'WORKING' income_type customers have the largest count of clients experiencing payment concerns, according to the aforementioned Bar plot.

AMT_TOTAL INCOME

| Row Labels | Count of AMT_TOTAL INCOME |
|---|---|
| HIGH | 7595 |
| LOW | 201045 |
| MEDIUM | 74046 |
| Grand Total | 282686 |

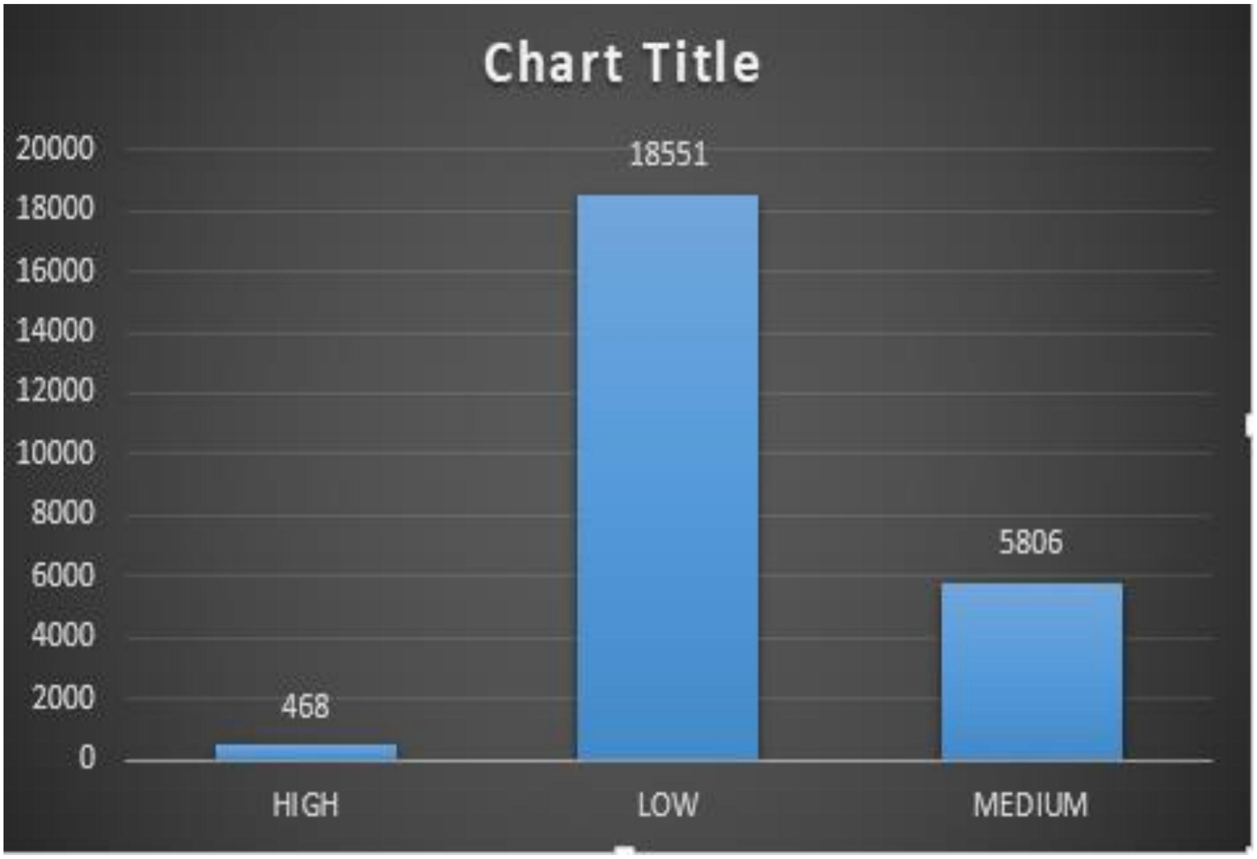
AMT_TOTAL_INCOME with no payment issues



The customer having the whole income range as 'LOW' has the largest count when it comes to clients having no payment concerns, according to the aforementioned bar plot.

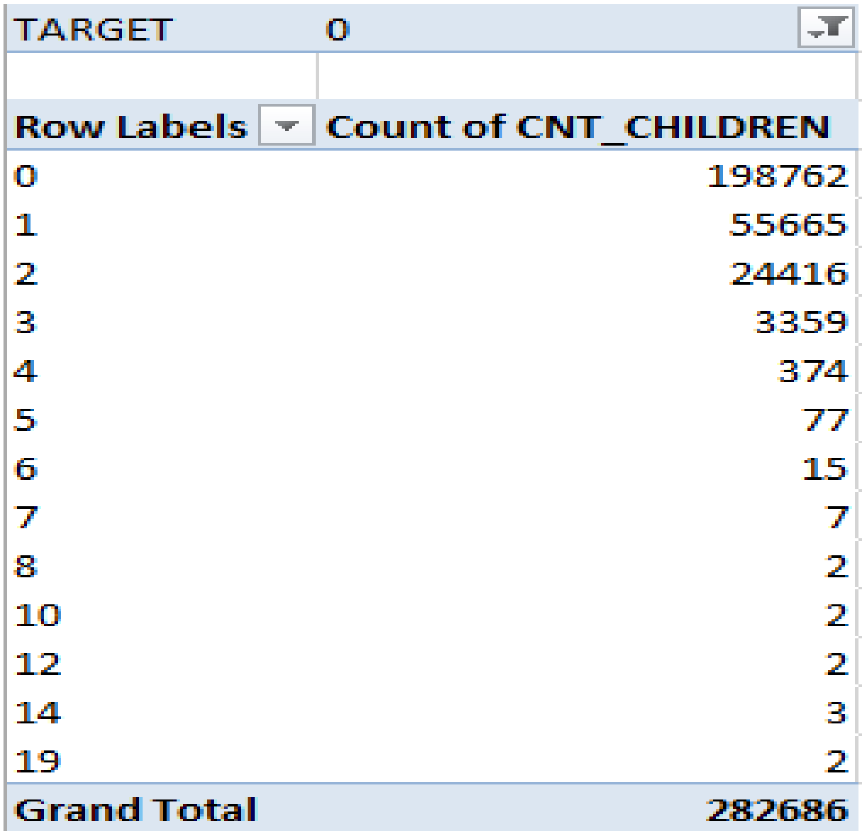
AMT_TOTAL_INCOME with payment issues

| Count of TARGET | |
|---|---|
| | |
| **Row Labels** | **1** |
| HIGH | 468 |
| LOW | 18551 |
| MEDIUM | 5806 |
| Grand Total | 24825 |



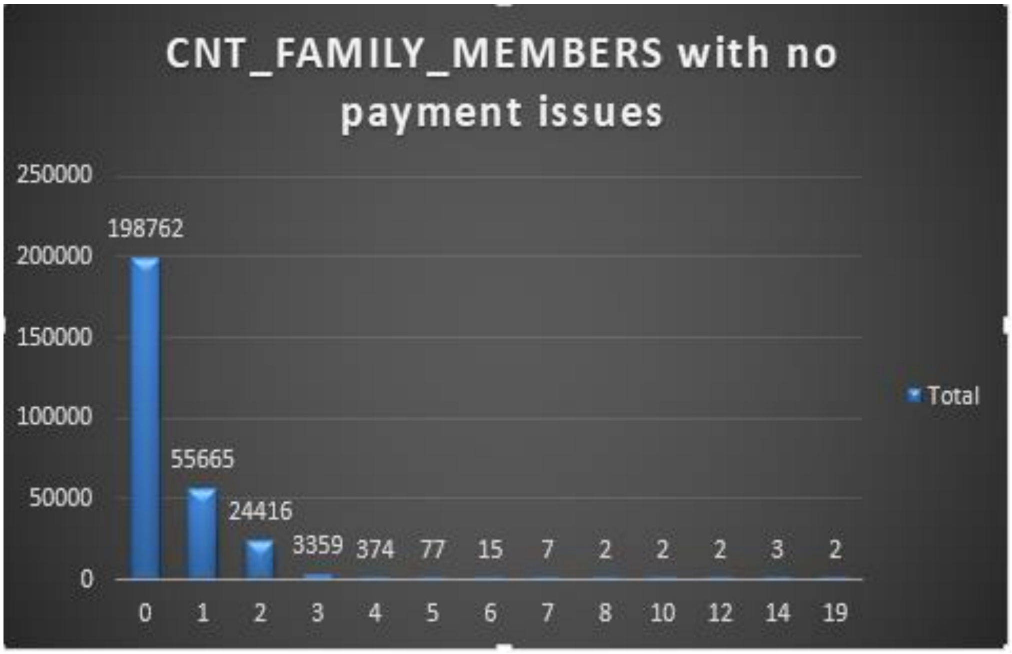
The accompanying bar plot indicates that clients with total income ranges that are "LOW" have the largest percentage of clients with payment difficulties.

CNT_FAMILY

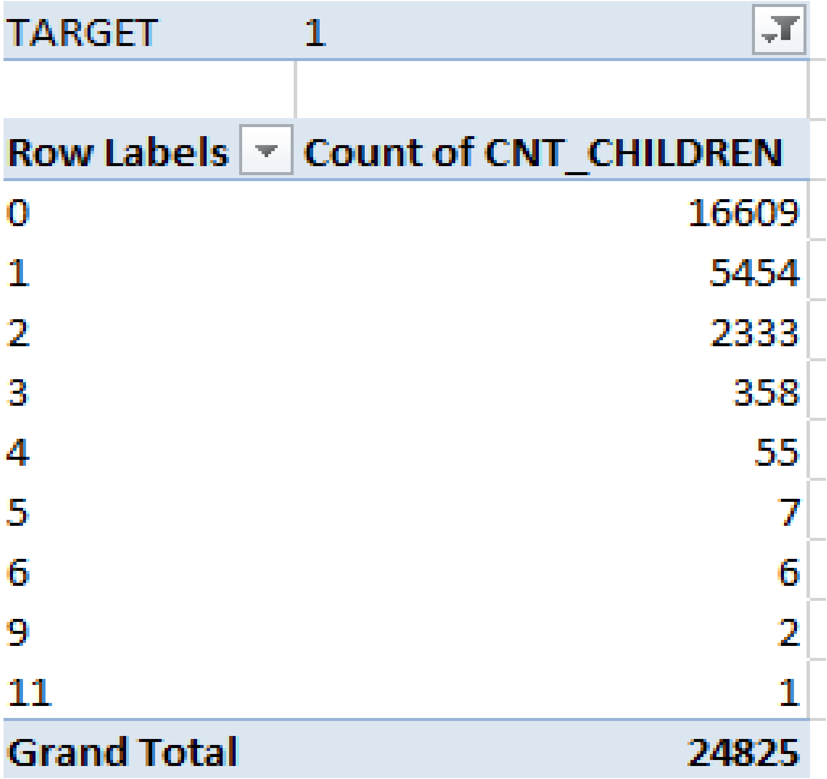


TARGET 0

| Row Labels | Count of CNT_CHILDREN |
|---|---|
| 0 | 198762 |
| 1 | 55665 |
| 2 | 24416 |
| 3 | 3359 |
| 4 | 374 |
| 5 | 77 |
| 6 | 15 |
| 7 | 7 |
| 8 | 2 |
| 10 | 2 |
| 12 | 2 |
| 14 | 3 |
| 19 | 2 |
| **Grand Total** | **282686** |

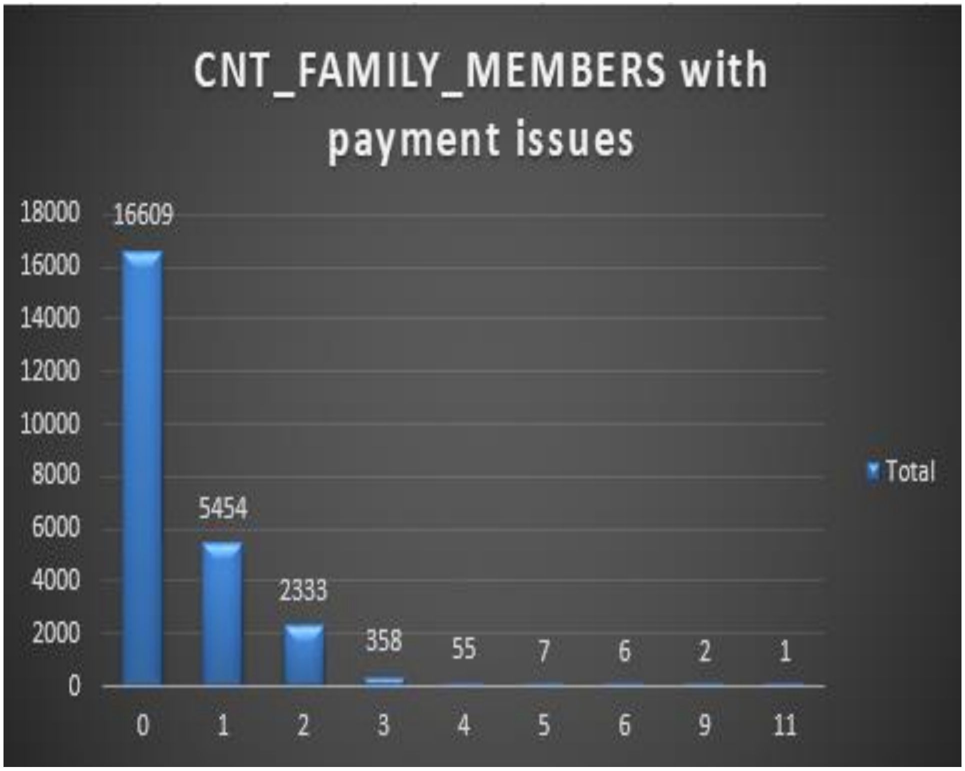


According to the above Bar Plot, clients with no family members have the highest percentage of clients with no payment concerns.



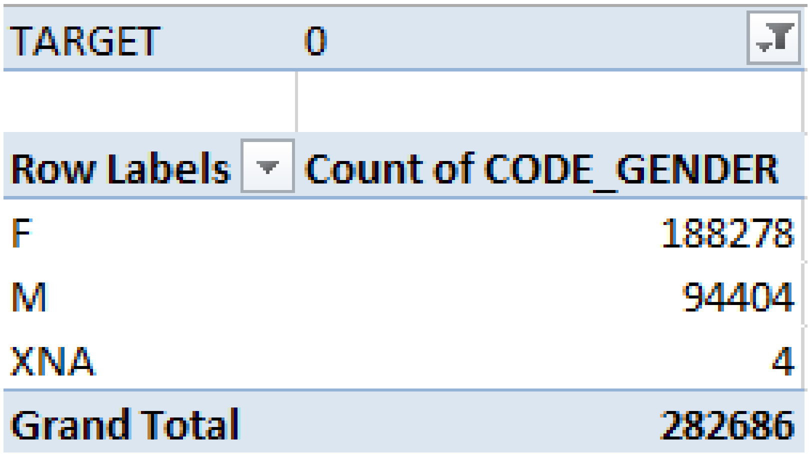
TARGET 1

| Row Labels | Count of CNT_CHILDREN |
|---|---|
| 0 | 16609 |
| 1 | 5454 |
| 2 | 2333 |
| 3 | 358 |
| 4 | 55 |
| 5 | 7 |
| 6 | 6 |
| 9 | 2 |
| 11 | 1 |
| **Grand Total** | **24825** |



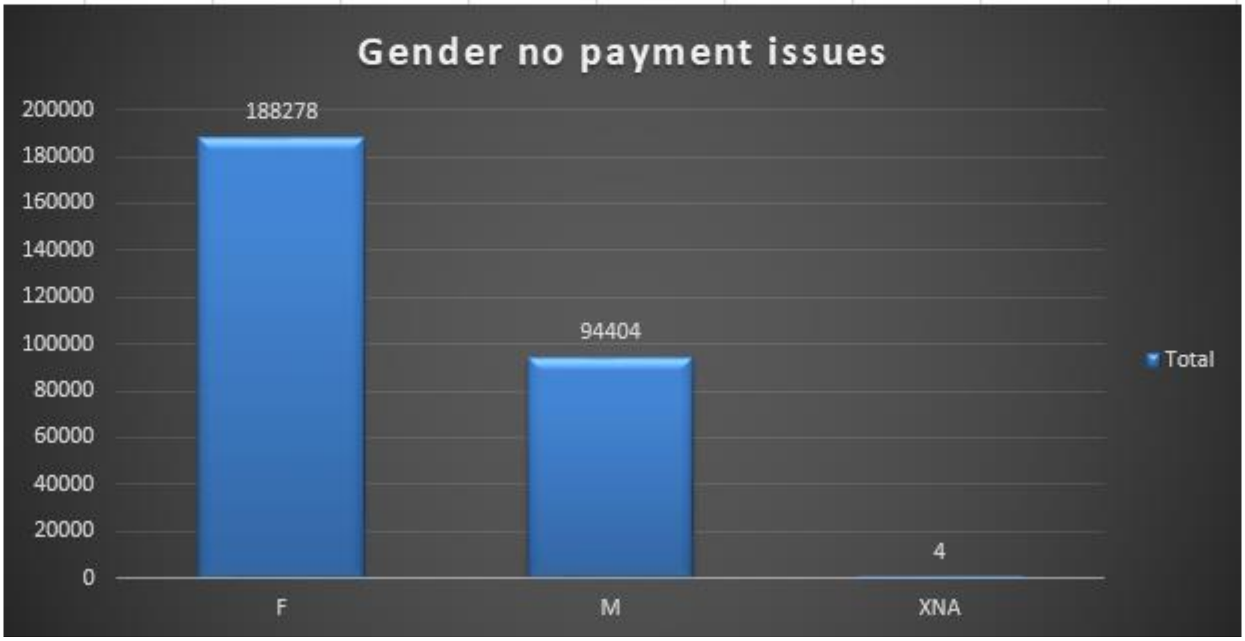
According to the aforementioned bar plot, customers with no family members are the ones who have the most number of payment problems.

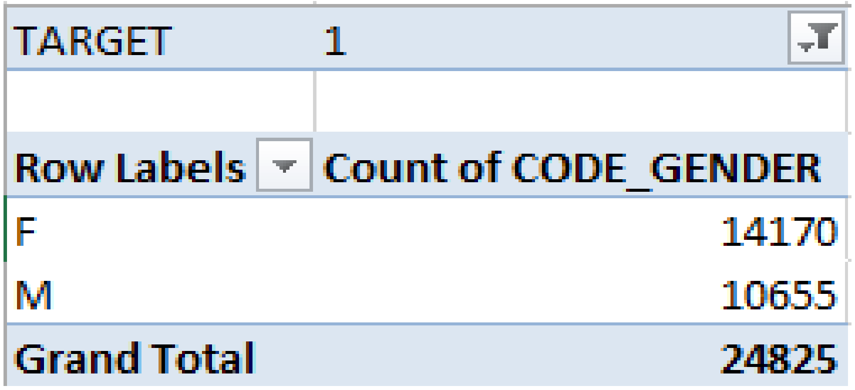
CODE_GENDER



| TARGET | 0 |
|---|---|

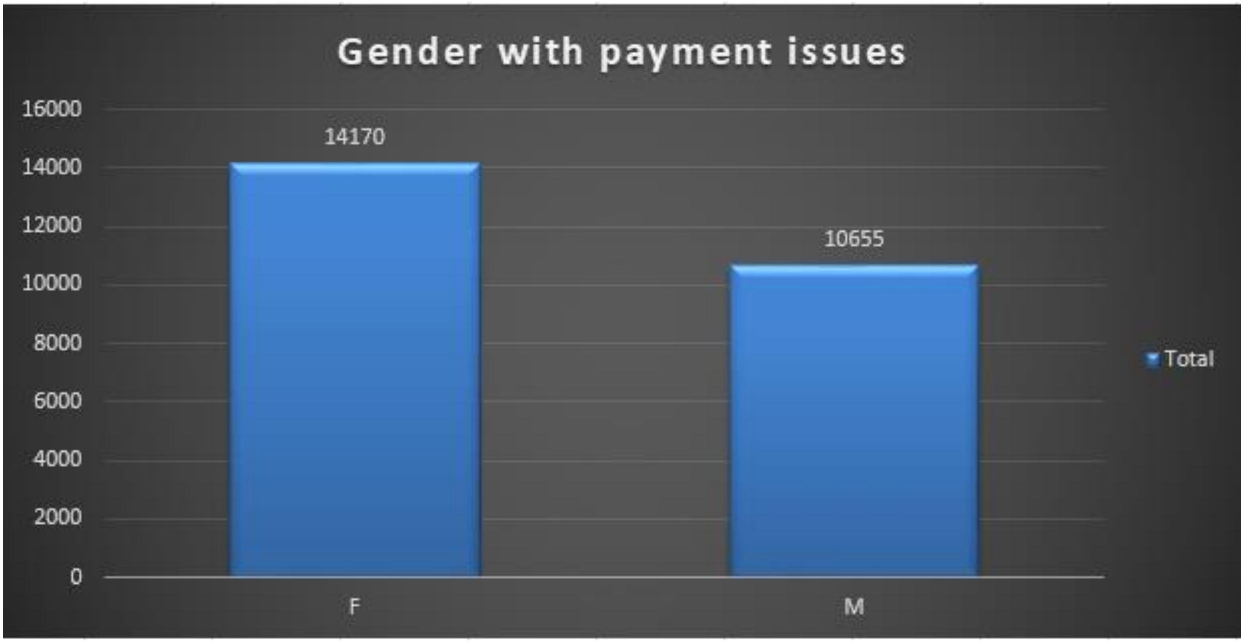
| Row Labels | Count of CODE_GENDER |
|---|---|
| F | 188278 |
| M | 94404 |
| XNA | 4 |
| **Grand Total** | **282686** |





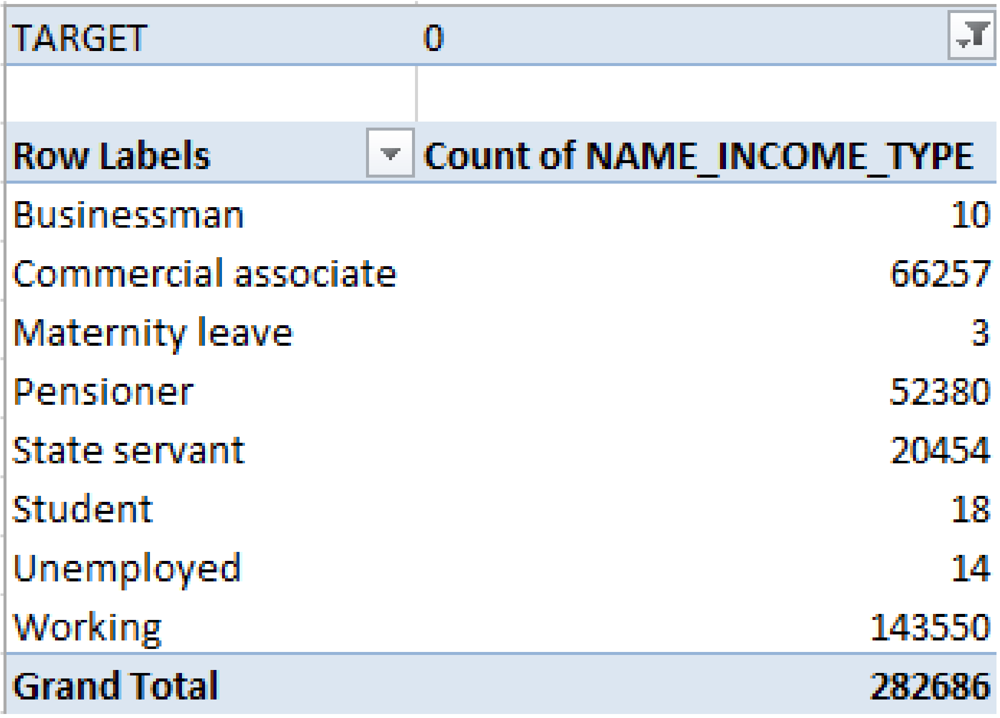
| TARGET | 1 |
|---|---|

| Row Labels | Count of CODE_GENDER |
|---|---|
| F | 14170 |
| M | 10655 |
| **Grand Total** | **24825** |

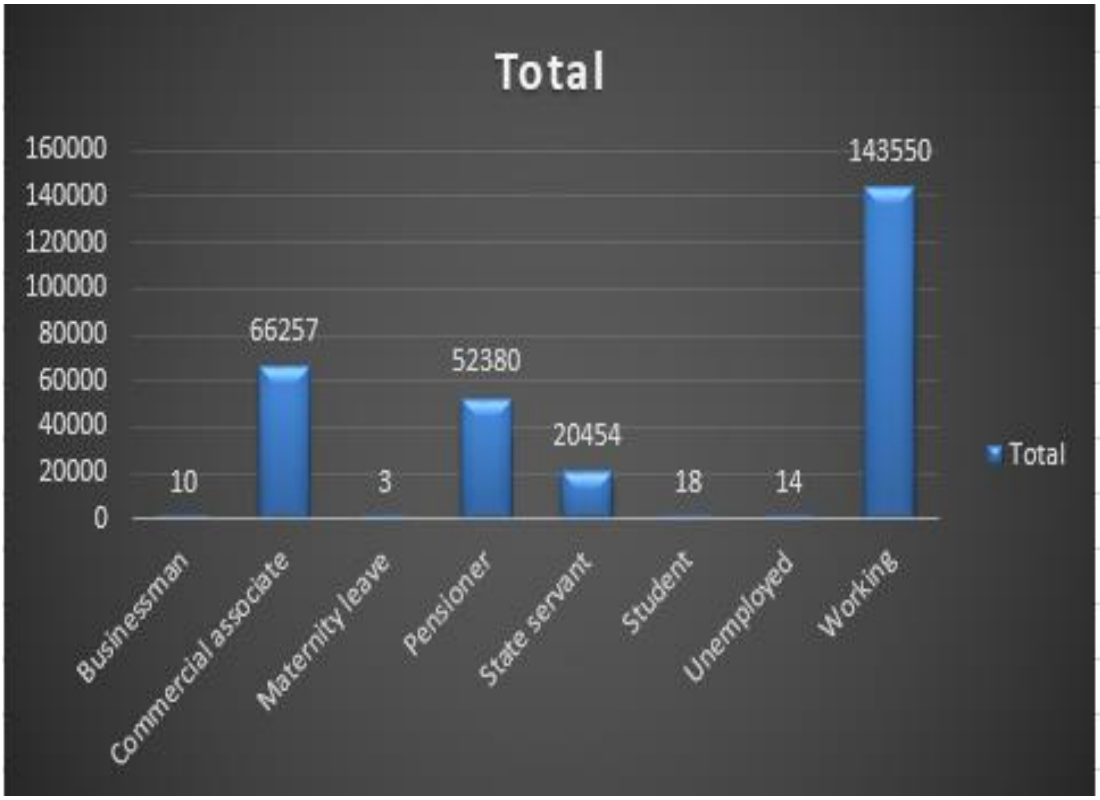


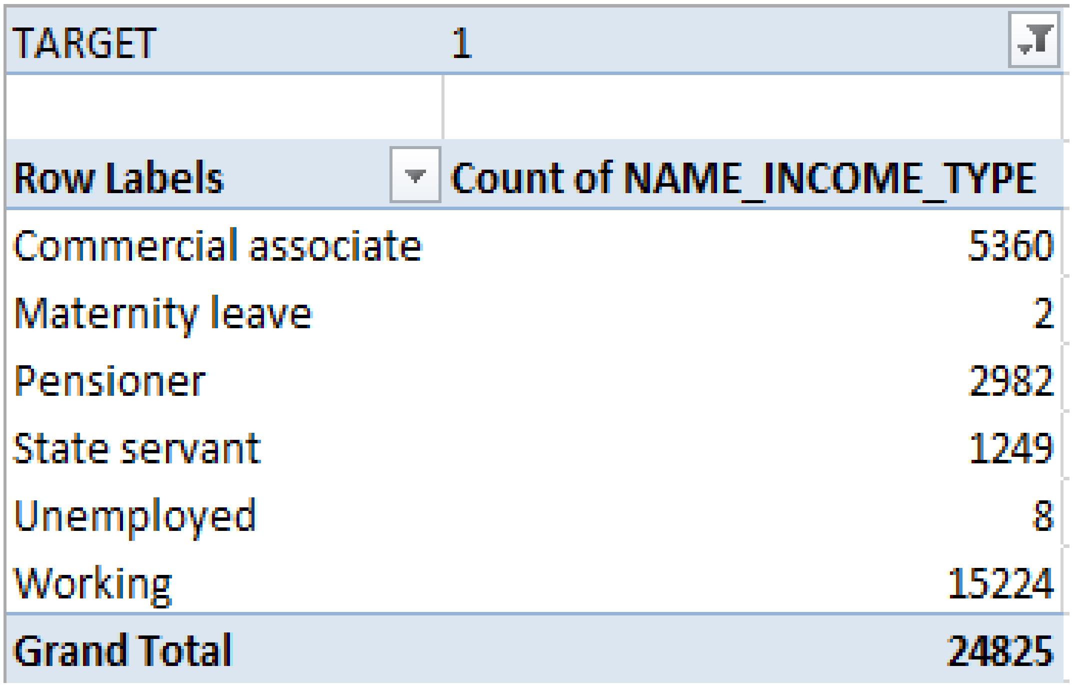
According to the above bar plot, clients with CODE_GENDER = 'F' had the most non-defaulters (188278-14170 = 174108).

## NAME_INCOME_TYPE

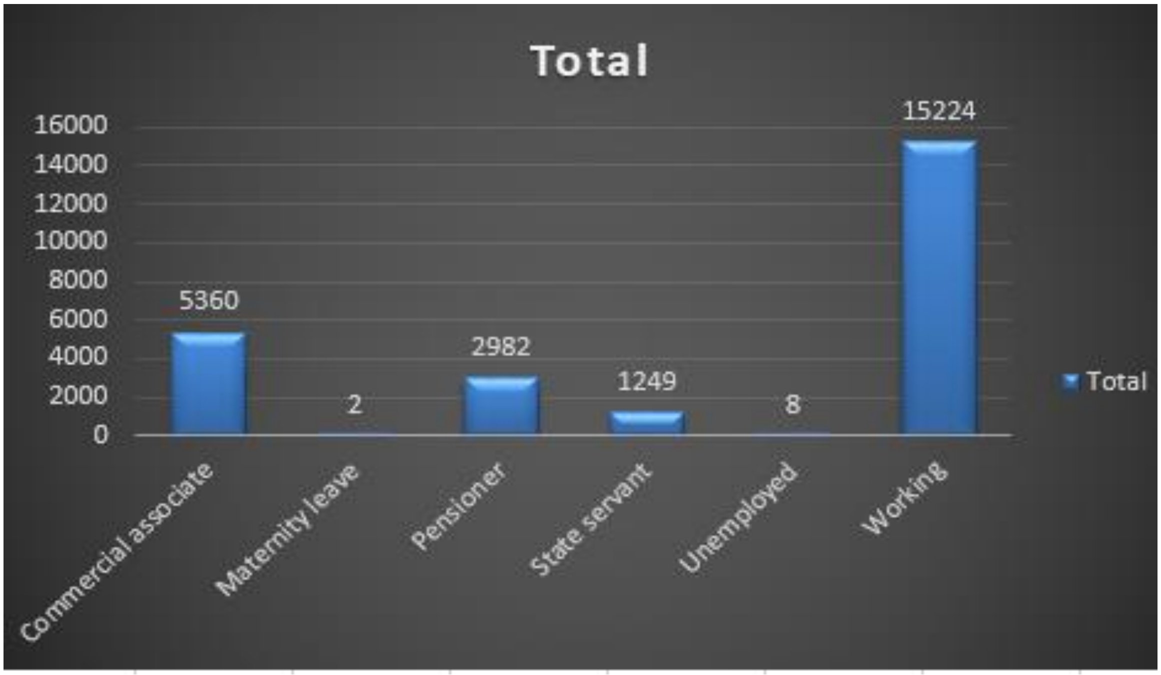


| | |
|---|---|
| TARGET | 0 |
| | |
| **Row Labels** | **Count of NAME_INCOME_TYPE** |
| Businessman | 10 |
| Commercial associate | 66257 |
| Maternity leave | 3 |
| Pensioner | 52380 |
| State servant | 20454 |
| Student | 18 |
| Unemployed | 14 |
| Working | 143550 |
| **Grand Total** | **282686** |



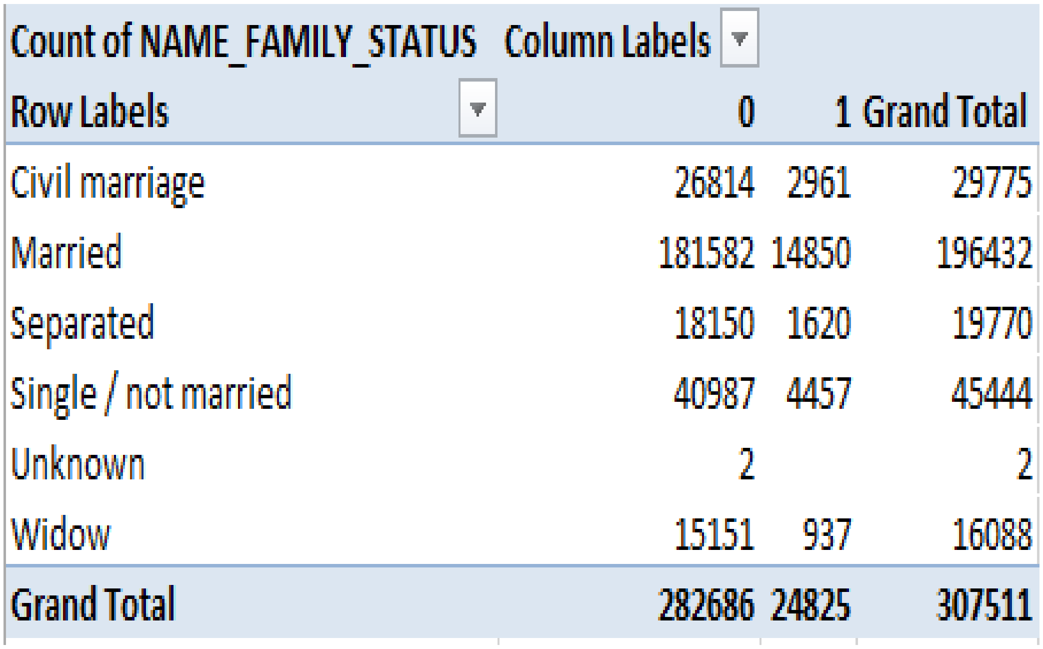


| | |
|---|---|
| TARGET | 1 |
| | |
| **Row Labels** | **Count of NAME_INCOME_TYPE** |
| Commercial associate | 5360 |
| Maternity leave | 2 |
| Pensioner | 2982 |
| State servant | 1249 |
| Unemployed | 8 |
| Working | 15224 |
| **Grand Total** | **24825** |

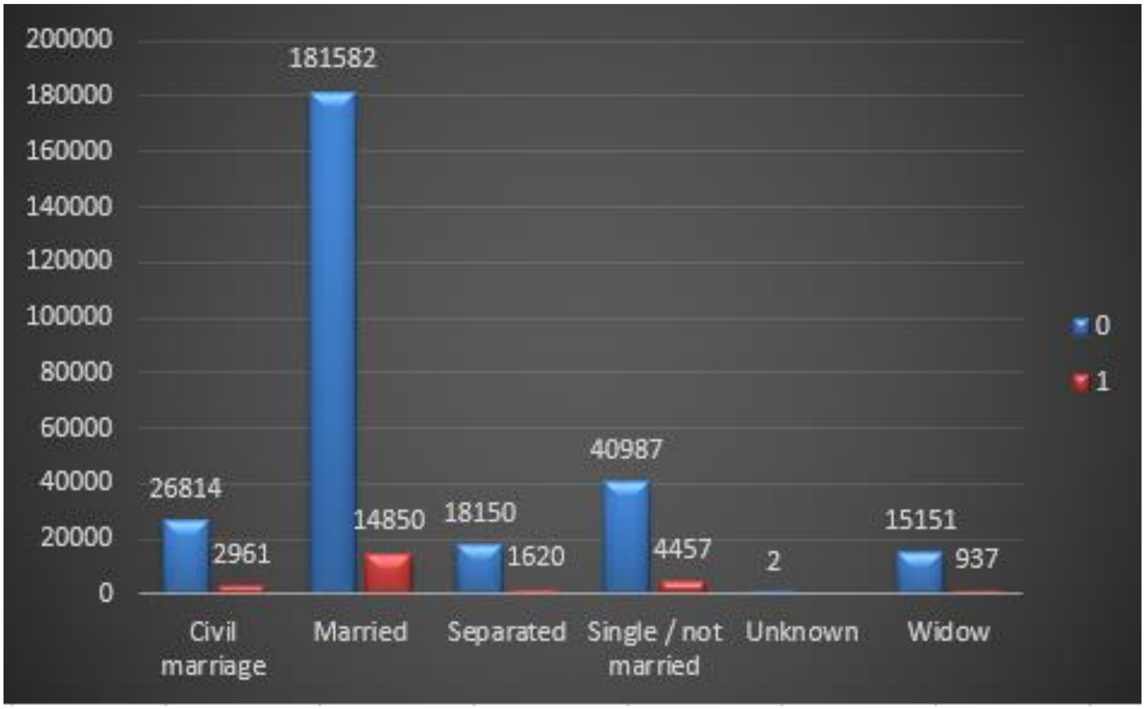


According to the adjacent bar plot, customers with NAME_INCOME_TYPE = "WORKING" have the largest number of non-defaulters, or 143550-15224 = 128326.

## NAME_FAMILY_STATUS

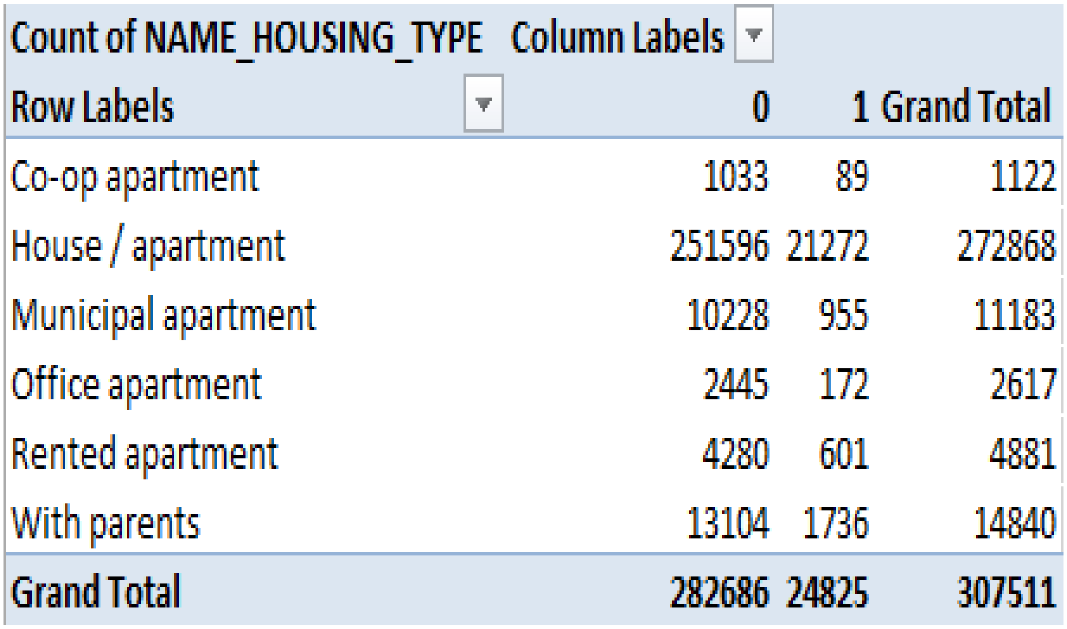


| Count of NAME_FAMILY_STATUS | Column Labels ▾ | | |
|---|---|---|---|
| Row Labels ▾ | 0 | 1 | Grand Total |
| Civil marriage | 26814 | 2961 | 29775 |
| Married | 181582 | 14850 | 196432 |
| Separated | 18150 | 1620 | 19770 |
| Single / not married | 40987 | 4457 | 45444 |
| Unknown | 2 | | 2 |
| Widow | 15151 | 937 | 16088 |
| Grand Total | 282686 | 24825 | 307511 |

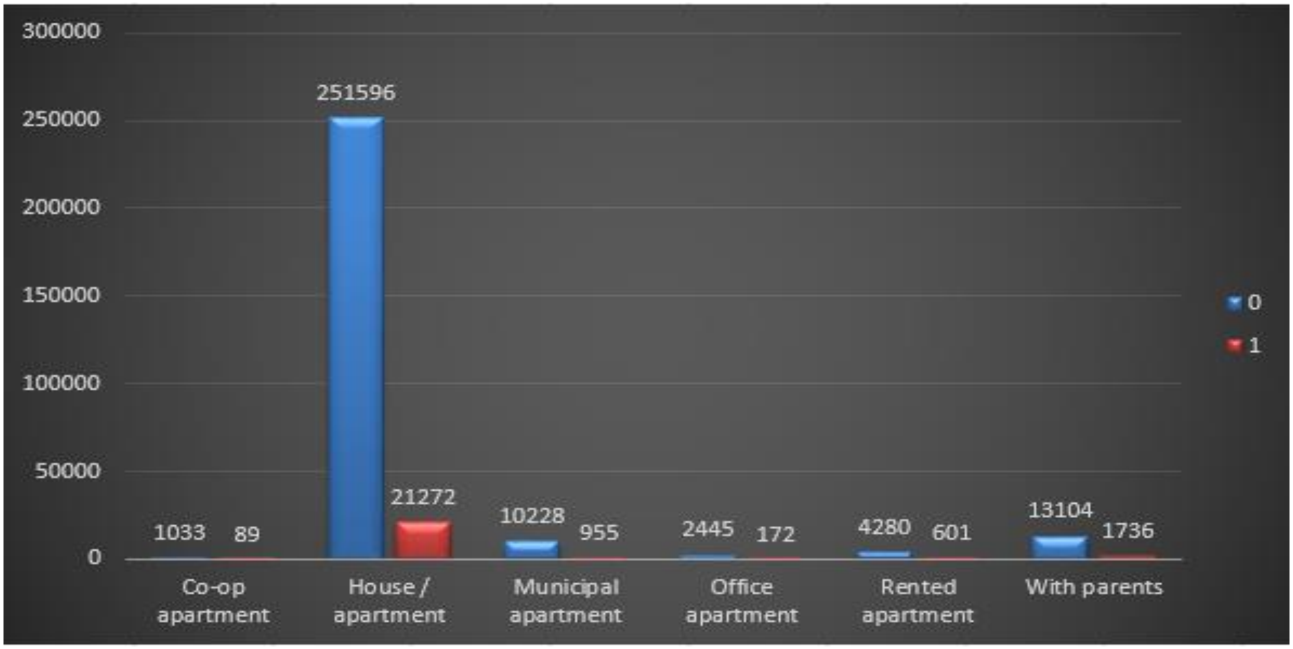


customers with NAME_FAMILY_STATUS = 'MARRIED' are, according to the adjacent Bar Plot, customers had the most nondefaulters, with a total of 166732 (181582 - 14850).

NAME_HOUSING_TYPE

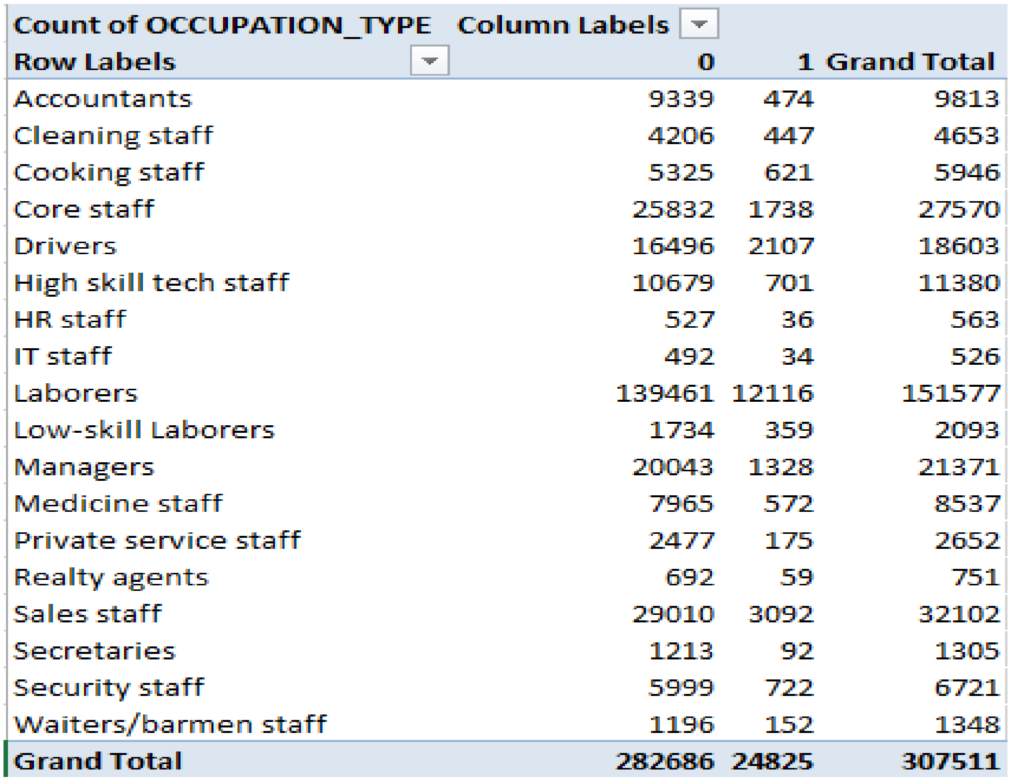


| Count of NAME_HOUSING_TYPE | Column Labels ▾ | | |
|---|---|---|---|
| Row Labels ▾ | 0 | 1 | Grand Total |
| Co-op apartment | 1033 | 89 | 1122 |
| House / apartment | 251596 | 21272 | 272868 |
| Municipal apartment | 10228 | 955 | 11183 |
| Office apartment | 2445 | 172 | 2617 |
| Rented apartment | 4280 | 601 | 4881 |
| With parents | 13104 | 1736 | 14840 |
| Grand Total | 282686 | 24825 | 307511 |

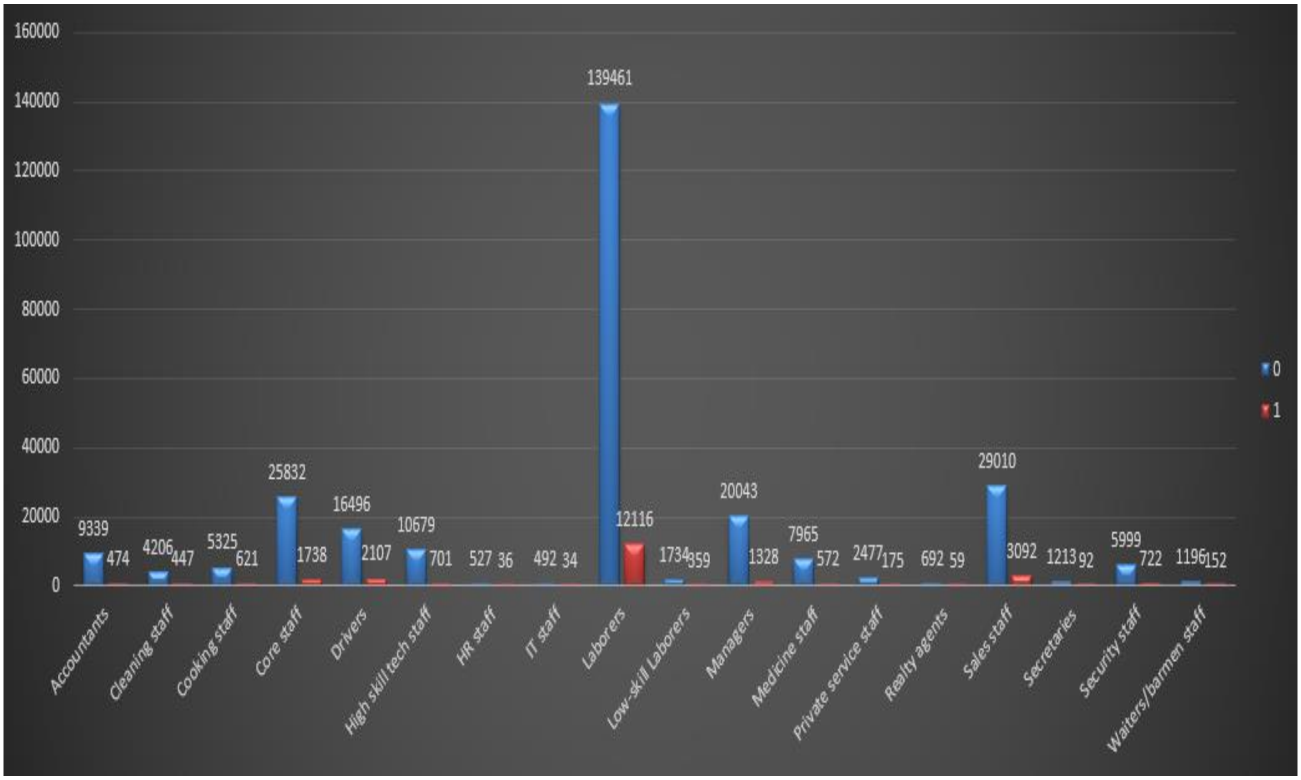


According to the above bar plot, clients with NAME_HOUSING_TYPE = "House/Apartment" had the largest number of non-defaulters, with a total of 251596-21272 = 230324.

## OCCUPATION_TYPE



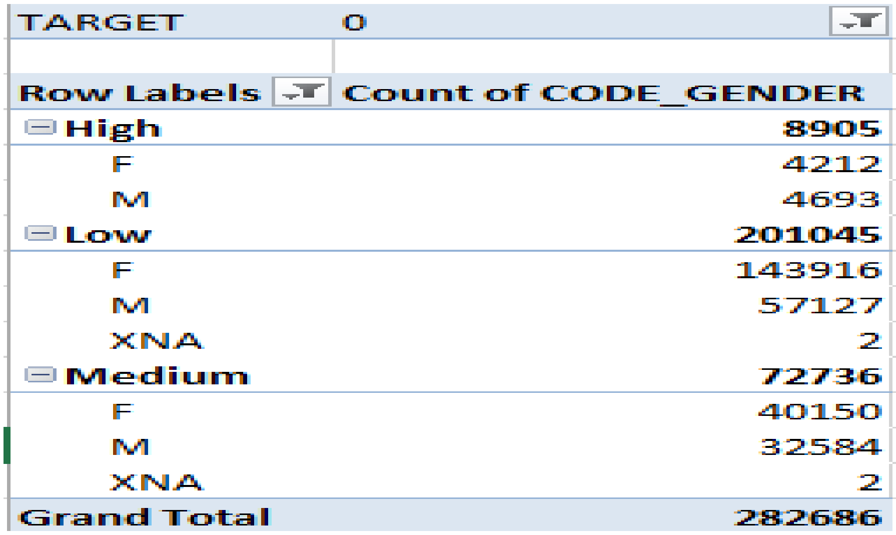
| Count of OCCUPATION_TYPE | Column Labels | | |
|---|---|---|---|
| Row Labels | 0 | 1 | Grand Total |
| Accountants | 9339 | 474 | 9813 |
| Cleaning staff | 4206 | 447 | 4653 |
| Cooking staff | 5325 | 621 | 5946 |
| Core staff | 25832 | 1738 | 27570 |
| Drivers | 16496 | 2107 | 18603 |
| High skill tech staff | 10679 | 701 | 11380 |
| HR staff | 527 | 36 | 563 |
| IT staff | 492 | 34 | 526 |
| Laborers | 139461 | 12116 | 151577 |
| Low-skill Laborers | 1734 | 359 | 2093 |
| Managers | 20043 | 1328 | 21371 |
| Medicine staff | 7965 | 572 | 8537 |
| Private service staff | 2477 | 175 | 2652 |
| Realty agents | 692 | 59 | 751 |
| Sales staff | 29010 | 3092 | 32102 |
| Secretaries | 1213 | 92 | 1305 |
| Security staff | 5999 | 722 | 6721 |
| Waiters/barmen staff | 1196 | 152 | 1348 |
| **Grand Total** | **282686** | **24825** | **307511** |



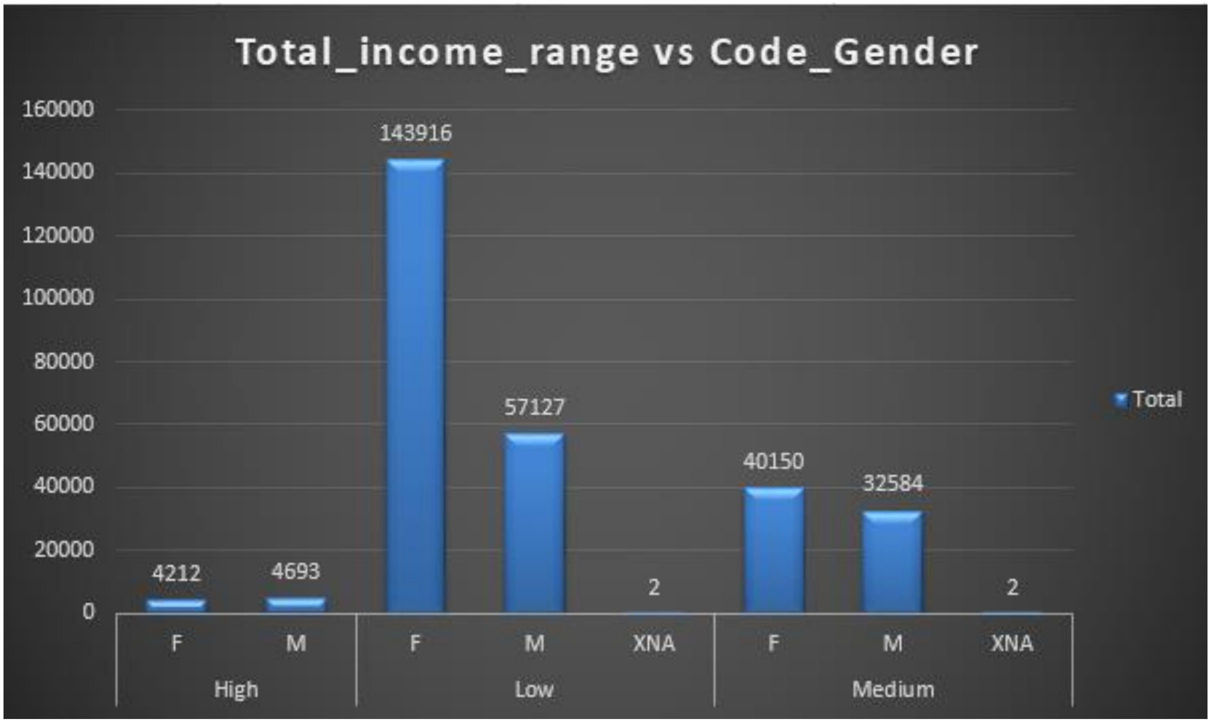
The customers with occupation_type = "Labourers" have the greatest count for non-defaulters, which is 139461-12116 = 127345, according to the adjacent Bar plot.

## Bivariate Analysis for TARGET variable

Target 0: Total_income_range vs Code_gender

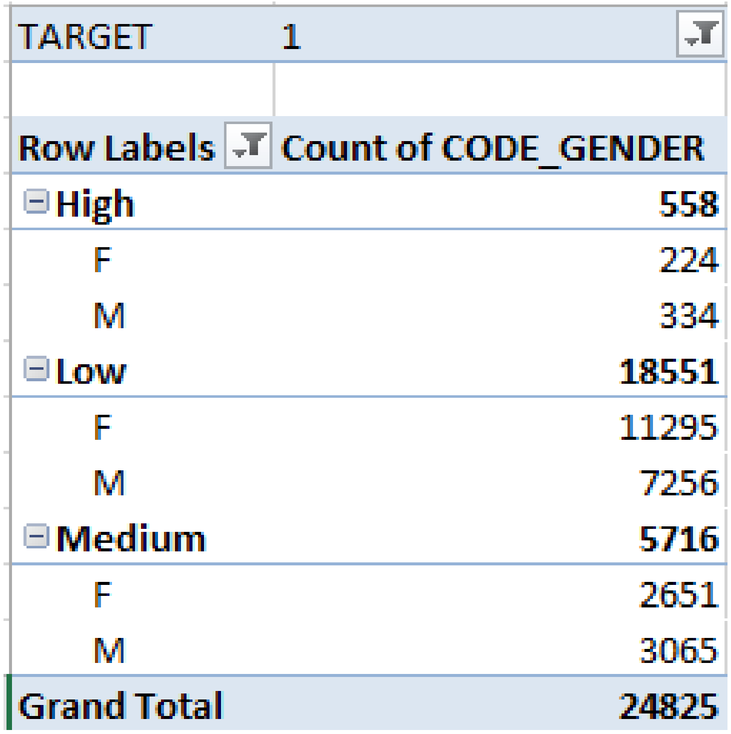


| TARGET | 0 |
|---|---|
| | |
| Row Labels | Count of CODE_GENDER |
| ⊟ High | 8905 |
| F | 4212 |
| M | 4693 |
| ⊟ Low | 201045 |
| F | 143916 |
| M | 57127 |
| XNA | 2 |
| ⊟ Medium | 72736 |
| F | 40150 |
| M | 32584 |
| XNA | 2 |
| Grand Total | 282686 |

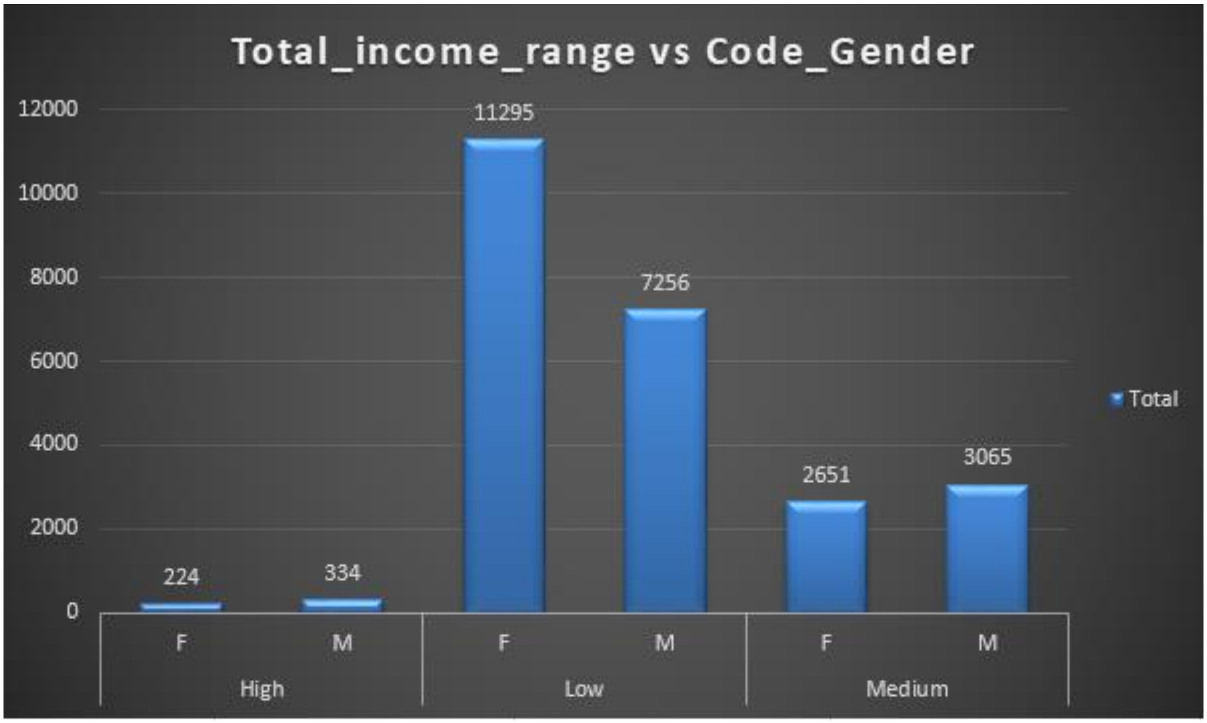


The majority of consumers without payment concerns are women who are in the low income bracket, according to the aforementioned bar map.

Target 1: Total_income_range vs Code_gender

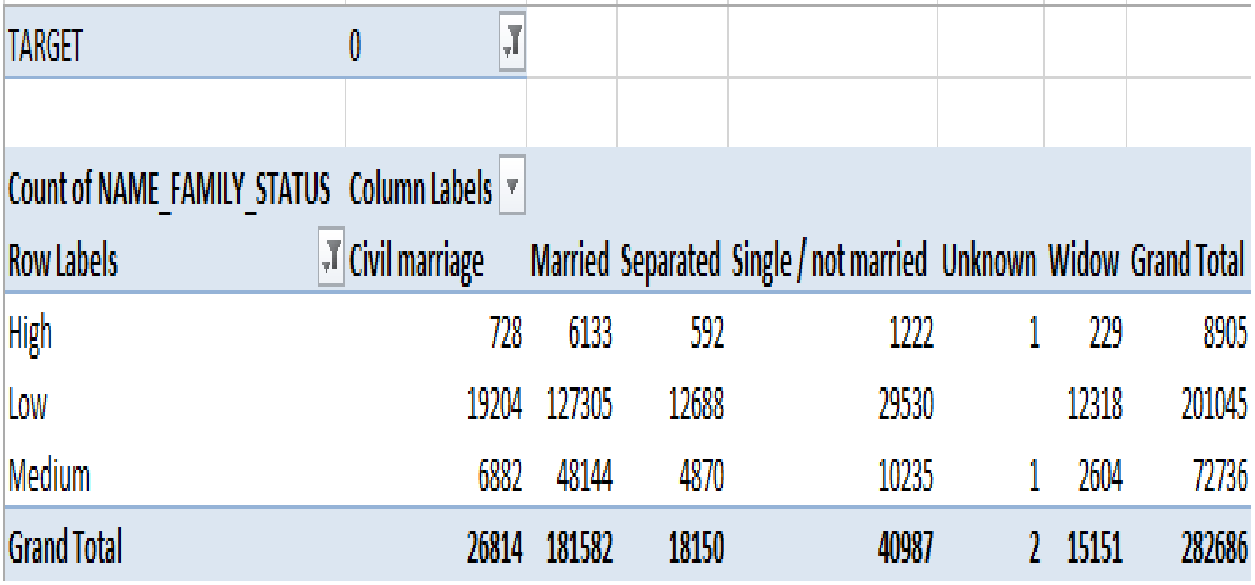


| TARGET | 1 |
|---|---|
| | |
| Row Labels | Count of CODE_GENDER |
| ⊟ **High** | **558** |
| F | 224 |
| M | 334 |
| ⊟ **Low** | **18551** |
| F | 11295 |
| M | 7256 |
| ⊟ **Medium** | **5716** |
| F | 2651 |
| M | 3065 |
| **Grand Total** | **24825** |

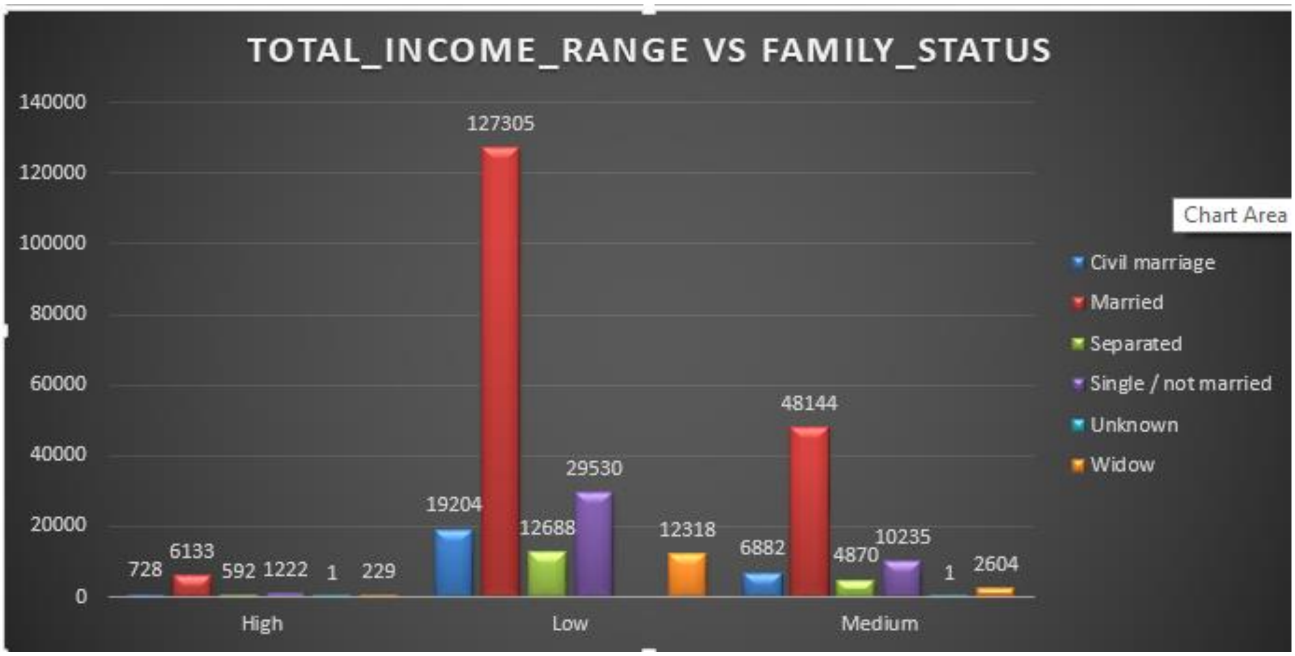


The accompanying bar plot indicates that the majority of clients that have payment concerns are women who fall into the low income group.

Target 0: Total Income vs Family status

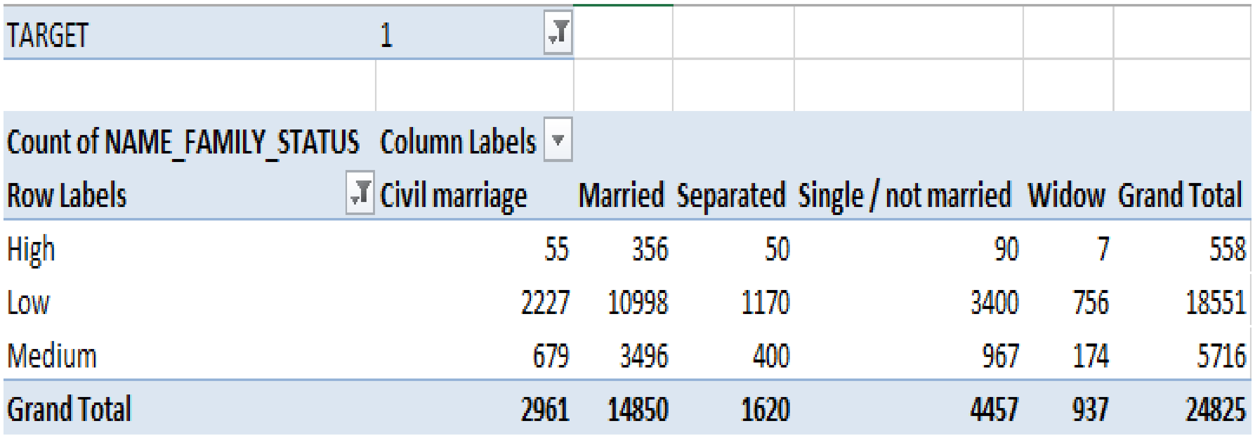


| | | | | | | | |
|---|---|---|---|---|---|---|---|
| TARGET | 0 | | | | | | |
| | | | | | | | |
| Count of NAME_FAMILY_STATUS | Column Labels | | | | | | |
| Row Labels | Civil marriage | Married | Separated | Single / not married | Unknown | Widow | Grand Total |
| High | 728 | 6133 | 592 | 1222 | 1 | 229 | 8905 |
| Low | 19204 | 127305 | 12688 | 29530 | | 12318 | 201045 |
| Medium | 6882 | 48144 | 4870 | 10235 | 1 | 2604 | 72736 |
| Grand Total | 26814 | 181582 | 18150 | 40987 | 2 | 15151 | 282686 |

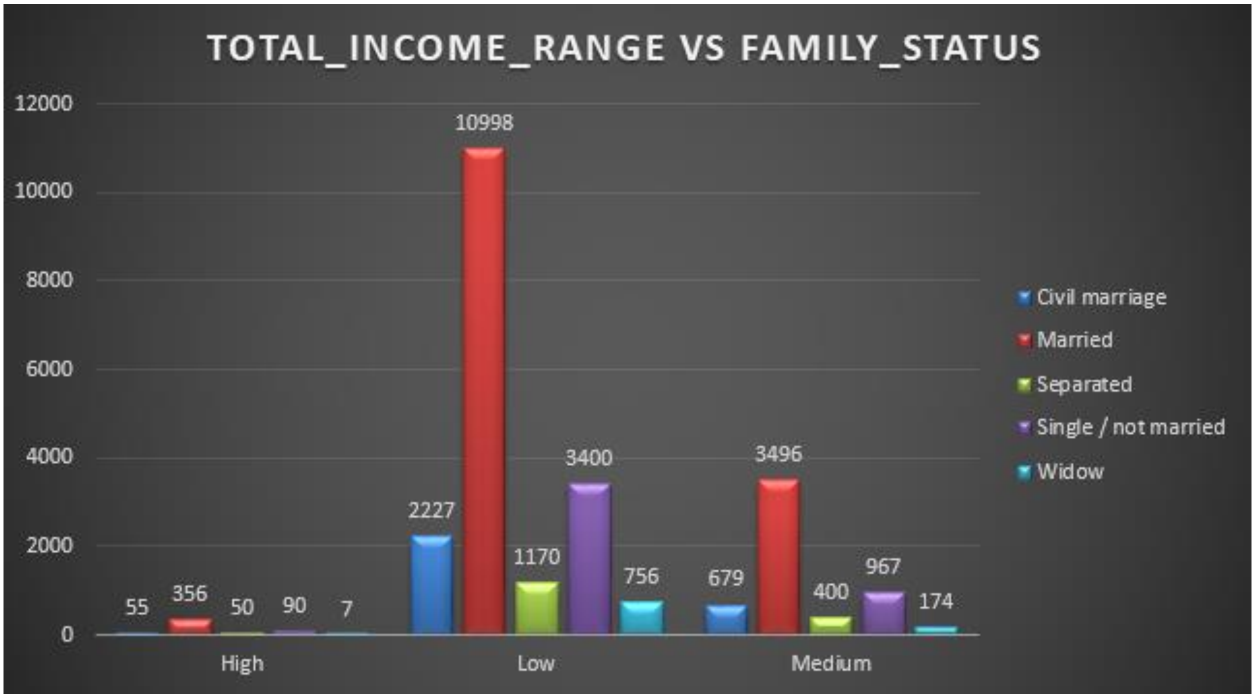


Clients with a family status of "Married" and a total income range of "Low" are those most likely to have no payment troubles, according to the adjacent Bar plot.

Target 1: Total Income vs Family status



| TARGET | 1 | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | |
| Count of NAME_FAMILY_STATUS | Column Labels | | | | | |
| Row Labels | Civil marriage | Married | Separated | Single / not married | Widow | Grand Total |
| High | 55 | 356 | 50 | 90 | 7 | 558 |
| Low | 2227 | 10998 | 1170 | 3400 | 756 | 18551 |
| Medium | 679 | 3496 | 400 | 967 | 174 | 5716 |
| **Grand Total** | **2961** | **14850** | **1620** | **4457** | **937** | **24825** |



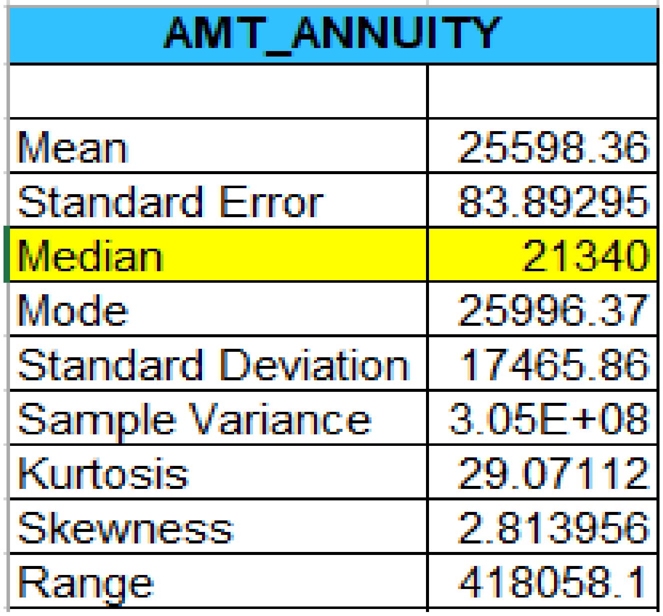
Clients with a total income range of "Low" and a family status of "Married" are those most likely to experience payment troubles, according to the adjacent Bar plot.

## Previous Application Dataset – Dropping, Imputing and analyzing Null values

The following columns from the preceding datasets for applications must be removed since they are unnecessary for doing data analysis.

1. HOUR_APPR_PROCESS_START
2. WEEKDAY_APPR_PROCESS_START_PREV
3. FLAG_LAST_APPL_PER_CONTRACT
4. NFLAG_LAST_APPL_IN_DAY
5. SK_ID_CURR
6. WEEKDAY_APPR_PROCESS_START
7. Removing the rows with the values 'XNA' &'XAP' for the column: NAME_TYPE_SUITE

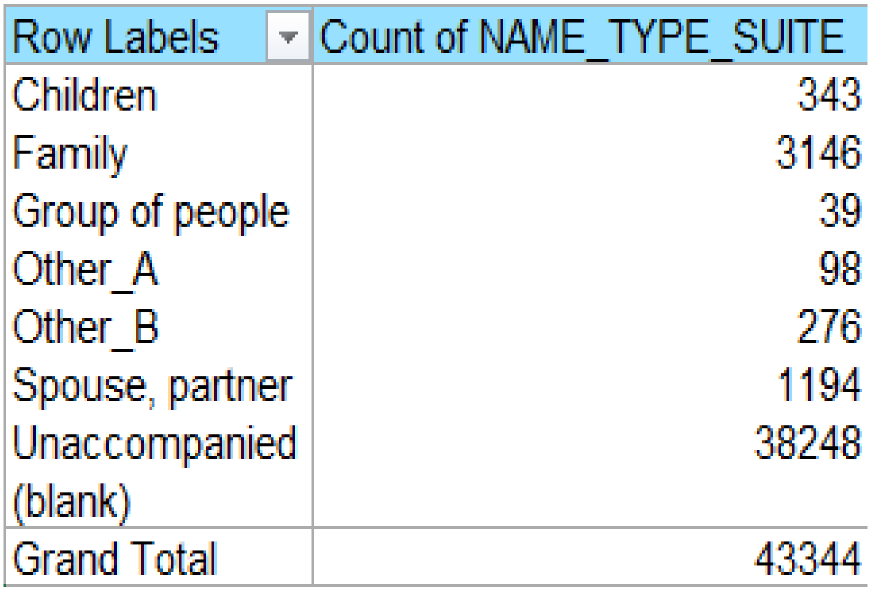
AMT_ANNUITY



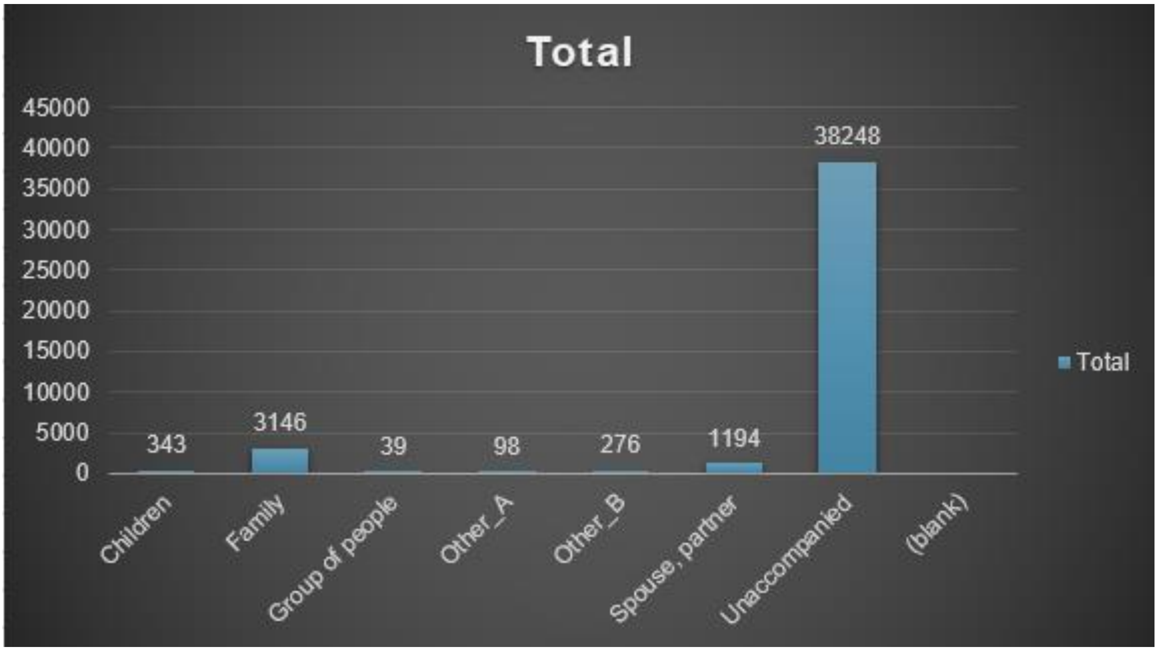
| AMT_ANNUITY | |
|---|---|
| | |
| Mean | 25598.36 |
| Standard Error | 83.89295 |
| Median | 21340 |
| Mode | 25996.37 |
| Standard Deviation | 17465.86 |
| Sample Variance | 3.05E+08 |
| Kurtosis | 29.07112 |
| Skewness | 2.813956 |
| Range | 418058.1 |

Replace Blanks with 21340

NAME_TYPE_SUITE

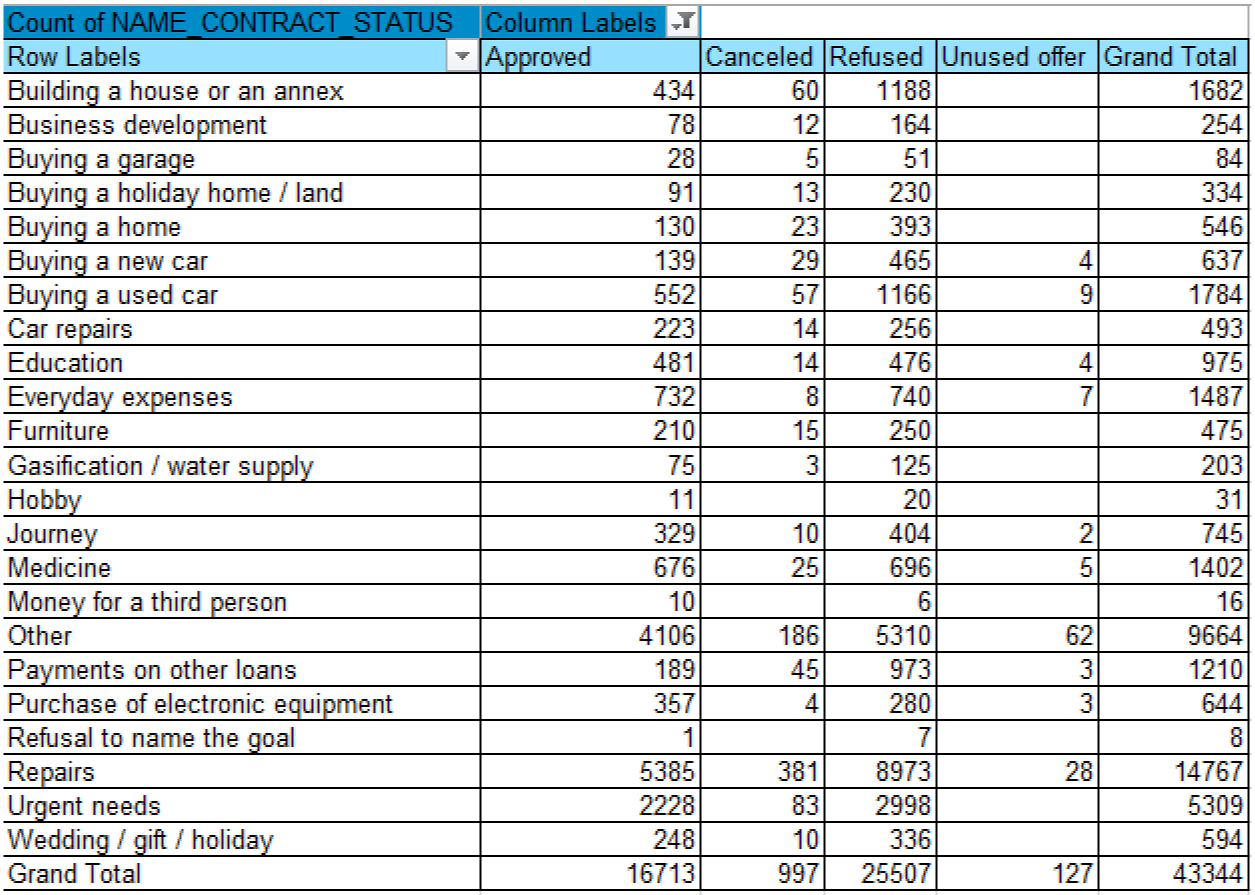


| Row Labels | Count of NAME_TYPE_SUITE |
|---|---|
| Children | 343 |
| Family | 3146 |
| Group of people | 39 |
| Other_A | 98 |
| Other_B | 276 |
| Spouse, partner | 1194 |
| Unaccompanied | 38248 |
| (blank) | |
| Grand Total | 43344 |

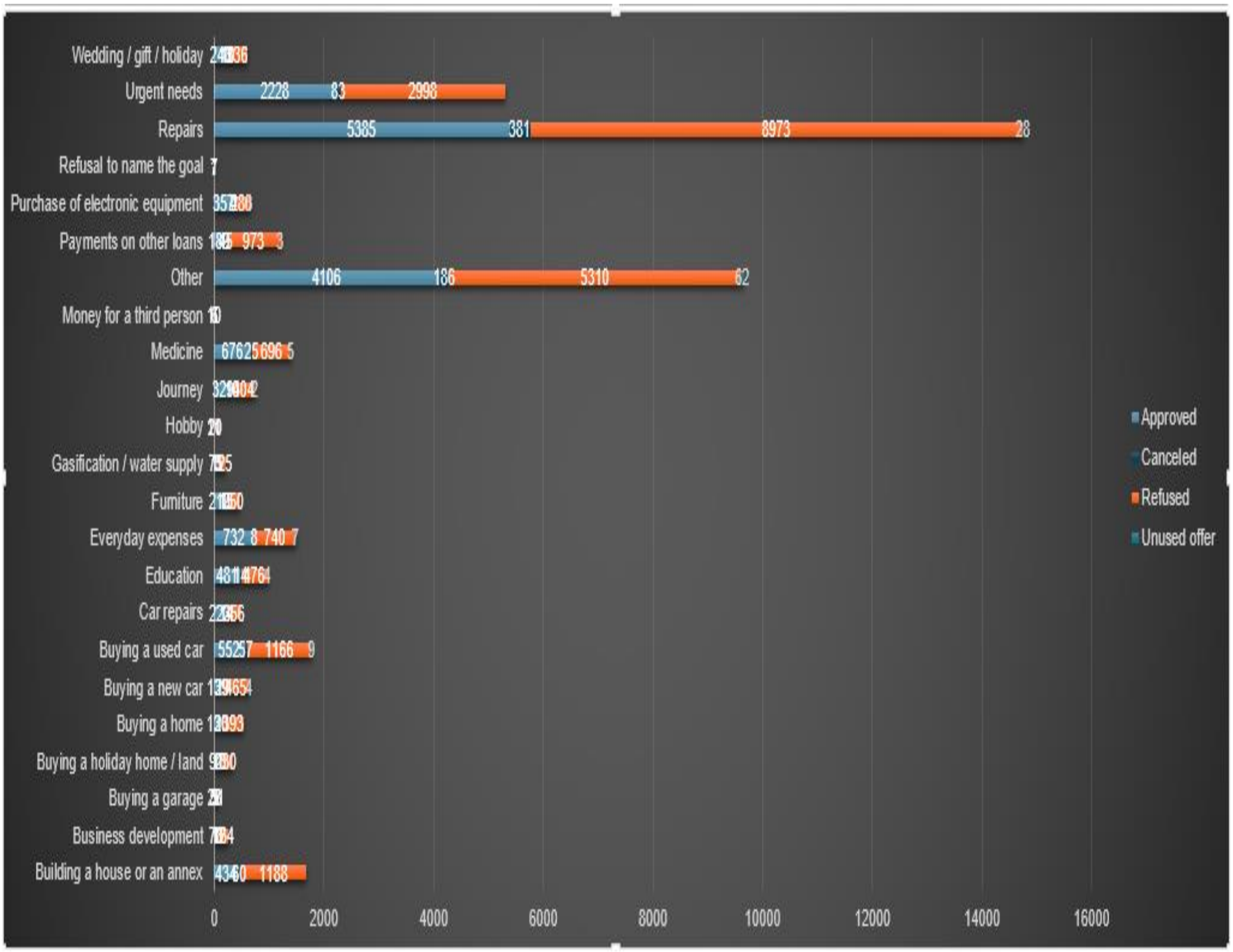


**Replace Blanks with Unaccompained**

## Distribution of Name Contract Status



| Count of NAME_CONTRACT_STATUS | Column Labels | | | | |
|---|---|---|---|---|---|
| Row Labels | Approved | Canceled | Refused | Unused offer | Grand Total |
| Building a house or an annex | 434 | 60 | 1188 | | 1682 |
| Business development | 78 | 12 | 164 | | 254 |
| Buying a garage | 28 | 5 | 51 | | 84 |
| Buying a holiday home / land | 91 | 13 | 230 | | 334 |
| Buying a home | 130 | 23 | 393 | | 546 |
| Buying a new car | 139 | 29 | 465 | 4 | 637 |
| Buying a used car | 552 | 57 | 1166 | 9 | 1784 |
| Car repairs | 223 | 14 | 256 | | 493 |
| Education | 481 | 14 | 476 | 4 | 975 |
| Everyday expenses | 732 | 8 | 740 | 7 | 1487 |
| Furniture | 210 | 15 | 250 | | 475 |
| Gasification / water supply | 75 | 3 | 125 | | 203 |
| Hobby | 11 | | 20 | | 31 |
| Journey | 329 | 10 | 404 | 2 | 745 |
| Medicine | 676 | 25 | 696 | 5 | 1402 |
| Money for a third person | 10 | | 6 | | 16 |
| Other | 4106 | 186 | 5310 | 62 | 9664 |
| Payments on other loans | 189 | 45 | 973 | 3 | 1210 |
| Purchase of electronic equipment | 357 | 4 | 280 | 3 | 644 |
| Refusal to name the goal | 1 | | 7 | | 8 |
| Repairs | 5385 | 381 | 8973 | 28 | 14767 |
| Urgent needs | 2228 | 83 | 2998 | | 5309 |
| Wedding / gift / holiday | 248 | 10 | 336 | | 594 |
| Grand Total | 16713 | 997 | 25507 | 127 | 43344 |



# Result

## The analysis that was done led to the following conclusions:

1. The Name of Contract status, i.e., Repairs work, has the largest number of Loans that have been approved, according to the above Bar Plot.
2. So, both the Applications Dataset and the Precious Applications Dataset are being used for the analysis.
3. The percentage of defaulters (target = 1) is around 8%, whereas the percentage of non-defaulters (target = 0) is approximately 92%.
4. The Bank often loans more money to female customers than to male customers since there are fewer female customers on the list of defaulters. If the credit amount is met, the bank may still hunt for additional male customers.
5. Additionally, customers from the Working Class are more likely than those from the Commercial Associate category to make their loan payments on time.

6. Consumers with education levels of secondary or higher secondary or above have a tendency to repay loans on schedule, allowing banks to prioritise lending to those consumers.
7. Clients with LOW credit amounts tend to pay off their loans on time as opposed to HIGH and MEDIUM credit amounts. Clients in the Age Groups 31–40 have the greatest rate of timely loan repayment, followed by clients in the Age Groups 41–60.
8. Compared to other housing types, customers who live with their parents often pay off their debts rapidly. Therefore, a bank may extend credit to customers who live with their parents.
9. consumers who are taking out loans to buy a new home, or who are taking out loans to buy a new car, or who have an income type like "State Servant," have a tendency to repay their debts on time, thus banks should favour consumers with this background.
10. The Bank should exercise greater caution when providing loans to customers for repairs since they have a high number of defaulters in addition to a high number of defaulters.