**Project – 6**

**Bank Loan Case Study**

**Project Description**

The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their installments. This information can be used to make decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants. The company wants to understand the key factors behind loan default so it can make better decisions about loan approval.

**Approach**

* Dataset was downloaded and imported into Microsoft excel
* Then clean the dataset finding the blanks and missing values ,imputing the missing values with the appropriate method(mean, median ,mode).
* Then I found the outliers in the dataset.
* After all these I used pivot tables and basic charts to visualise the data.
* Then I found top 10 correlation between the columns of repayer and defaulters

**Tech-Stack Used**

Microsoft® Word 2019 MSO (Version 2501 Build 16.0.18429.20132) 64-bit for data analysis, pivot table creation and chart generation.

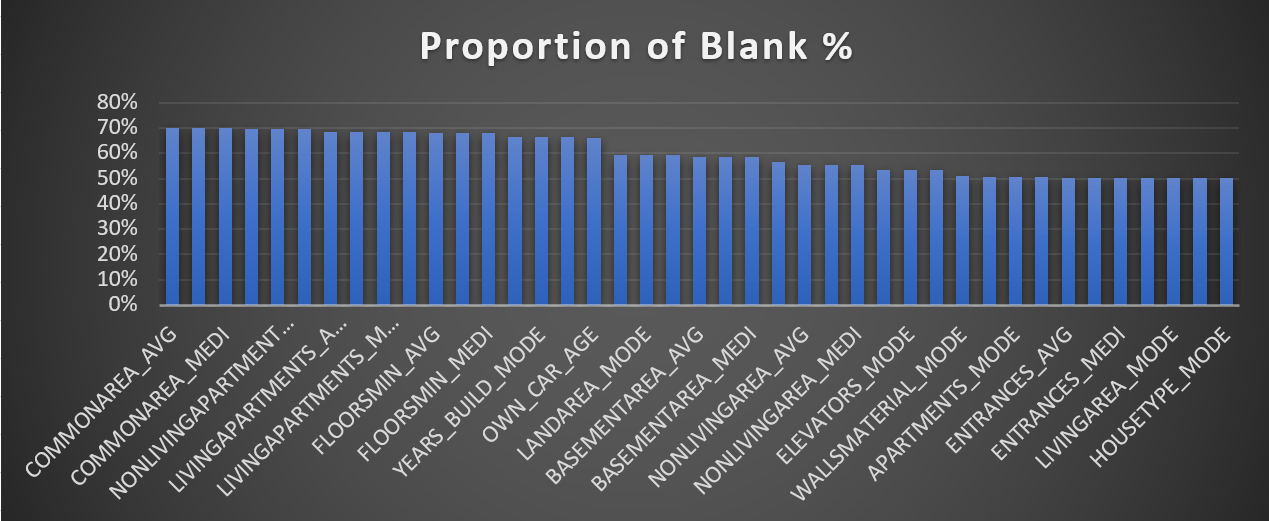
**Insights**

**FOR CURRENT APPLICATION**

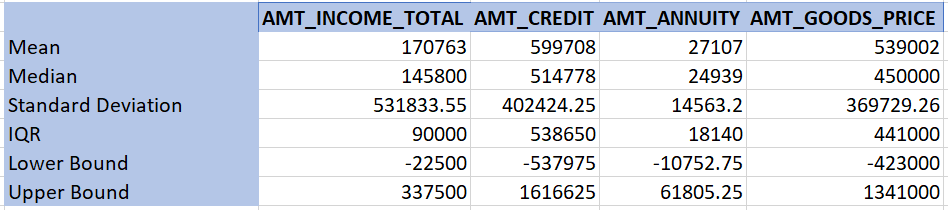
1. **Identify Missing Data and Deal with it Appropriately:**

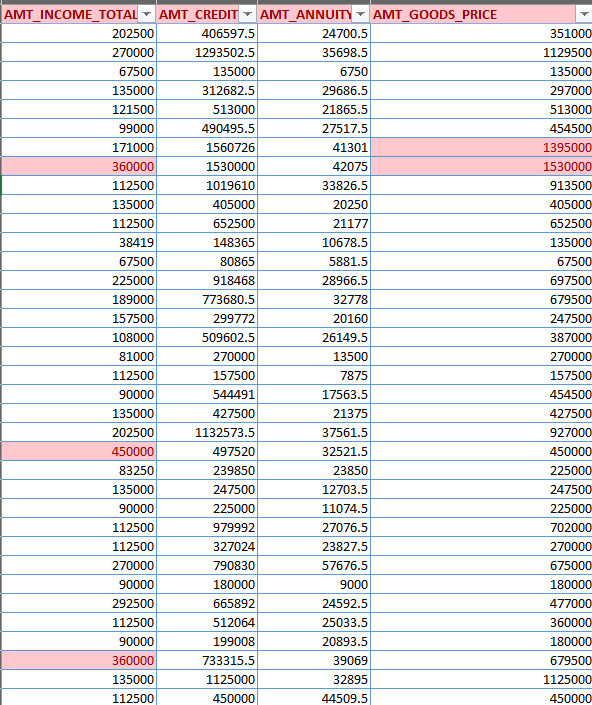
* Initially there were 122 columns and 50000 rows including header row in loan application dataset.
* First we used count blank formula to count all the blank cells in column
* Then we converted it into percent to plot the bar graph by transposing it into new sheet.
* Then we filled null values of numerical column with median as data contains outliers therefore mean is inappropriate.
* For non numerical data we deleted whole column
* Columns removed are:

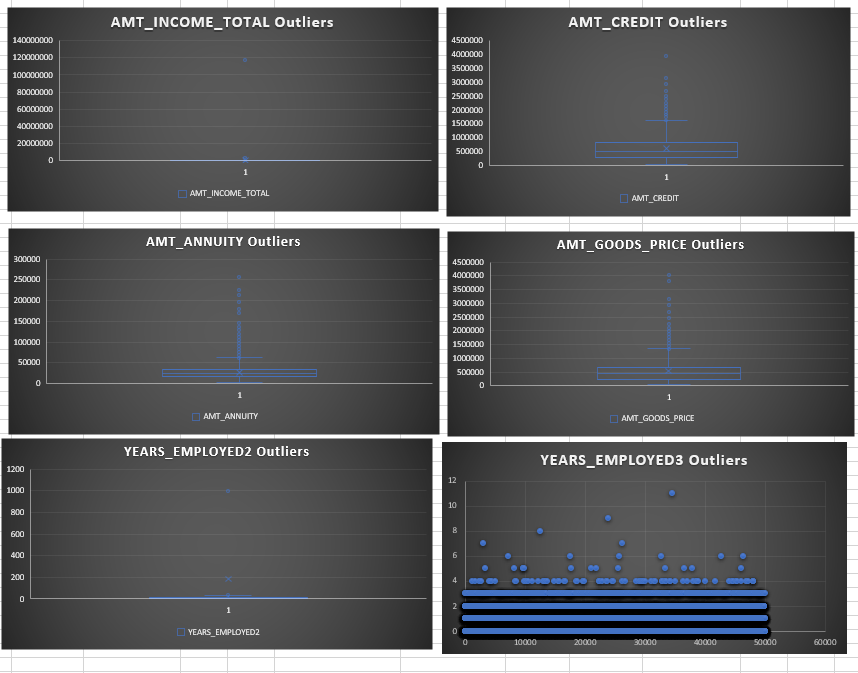
|  |
| --- |
| **COMMONAREA\_AVG** |
| **COMMONAREA\_MODE** |
| **COMMONAREA\_MEDI** |
| **NONLIVINGAPARTMENTS\_AVG** |
| **NONLIVINGAPARTMENTS\_MODE** |
| **NONLIVINGAPARTMENTS\_MEDI** |
| **LIVINGAPARTMENTS\_AVG** |
| **LIVINGAPARTMENTS\_MODE** |
| **LIVINGAPARTMENTS\_MEDI** |
| **FONDKAPREMONT\_MODE** |
| **FLOORSMIN\_AVG** |
| **FLOORSMIN\_MODE** |
| **FLOORSMIN\_MEDI** |
| **YEARS\_BUILD\_AVG** |
| **YEARS\_BUILD\_MODE** |
| **YEARS\_BUILD\_MEDI** |
| **OWN\_CAR\_AGE** |
| **LANDAREA\_AVG** |
| **LANDAREA\_MODE** |
| **LANDAREA\_MEDI** |
| **BASEMENTAREA\_AVG** |
| **BASEMENTAREA\_MODE** |
| **BASEMENTAREA\_MEDI** |
| **EXT\_SOURCE\_1** |
| **NONLIVINGAREA\_AVG** |
| **NONLIVINGAREA\_MODE** |
| **NONLIVINGAREA\_MEDI** |
| **ELEVATORS\_AVG** |
| **ELEVATORS\_MODE** |
| **ELEVATORS\_MEDI** |
| **WALLSMATERIAL\_MODE** |
| **APARTMENTS\_AVG** |
| **APARTMENTS\_MODE** |
| **APARTMENTS\_MEDI** |
| **ENTRANCES\_AVG** |
| **ENTRANCES\_MODE** |
| **ENTRANCES\_MEDI** |
| **LIVINGAREA\_AVG** |
| **LIVINGAREA\_MODE** |
| **LIVINGAREA\_MEDI** |
| **HOUSETYPE\_MODE** |



1. **Identify Outliers in the Dataset:**

* Outliers are the datapoints which deviate signifiacantly from the rest of the datapoints.
* So to analyze the datapoints we have used box plot and scatter plot to find out the outliers.
* We also used statistical formula to find upper bound manually to highlight the values greater than the upper bound value.





**Insights**

* AMT\_TOTAL\_INCOME

By the box plot we can understand that most of the values are below the 40,00,000.

The only value which seems unrealistic is 11,70,00,000. As it is outlier it affects the measure of central tendency.

* CNT\_CHILDREN

In few cases there are errors in data entry as some have count upto 11 childrens which seems unrealistic in today’s world.

* YEARS\_EMPLOYED2

By observing the scatter plot we can analyse that some values are above 1000 years which is clearly not possible and is an outlier.

* AMT\_GOOD\_PRICES

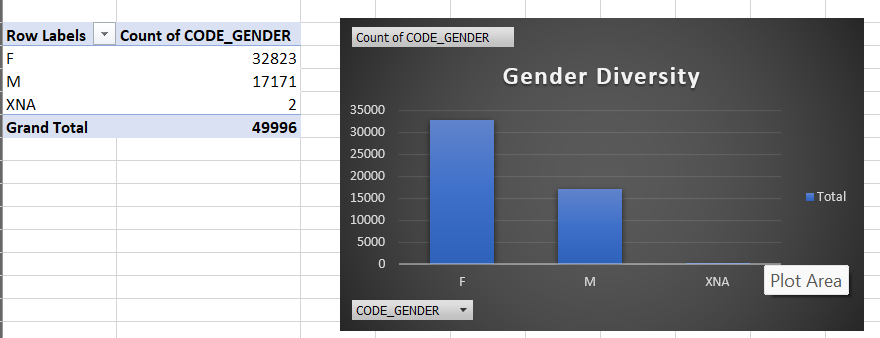
Some the client are applying for loan with extremely high prices and not realistic. Therefore treated as outliers.

* AMT\_ANNUITY

As there is very few chances of having annuity more than 200000 the points observed in box plot above the upper bound are treated as outliers.

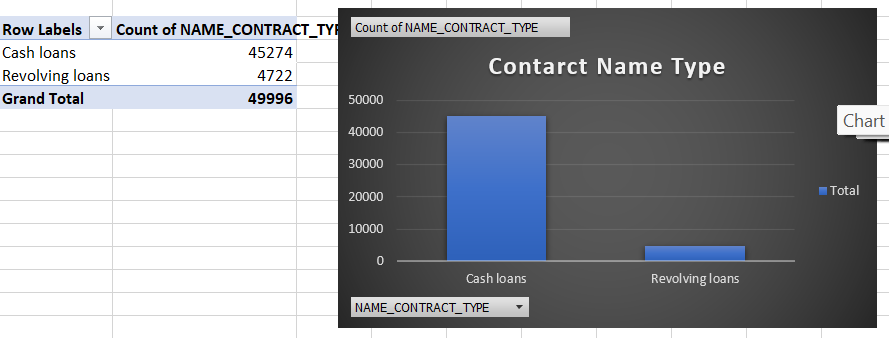
* AMT\_CREDIT

As there is very few chances that a client has this much credit more than 200000 the points above the upper bound are treated as outliers.

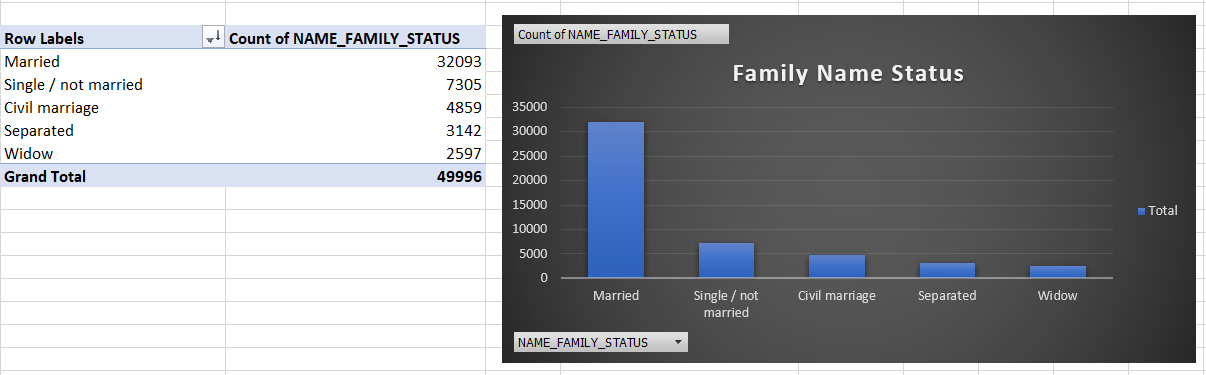
1. **Analyze Data Imbalance:**

**Gender Diversity:**

The dataset displays that number of females is much higher than males and also the imbalanced distribution in number of xna.

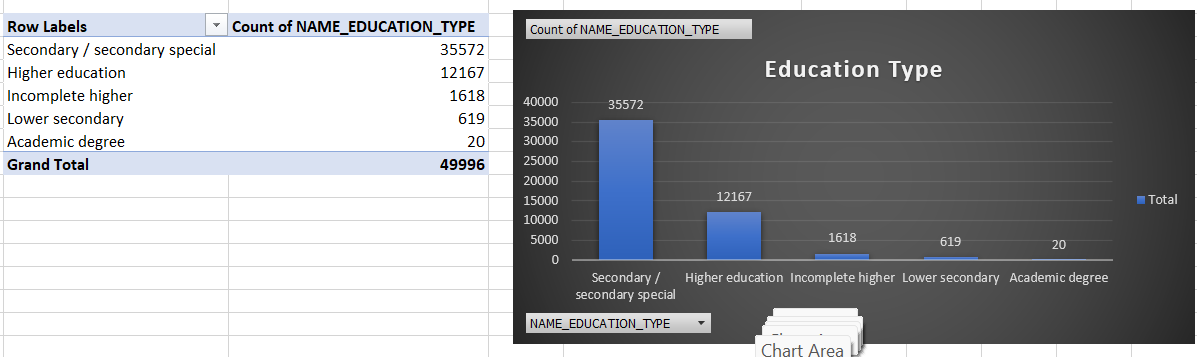
****

**Contract name type:**

The frequency of cash loans is much greater than the revolving loans. The revolving loan has small fraction of part out of the total number which is 10.4%.

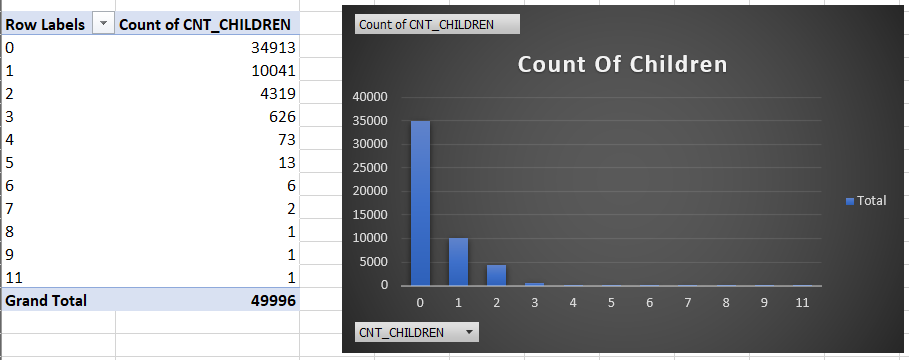
**Family Name status:**

Married people have taken most of the loans followed by Single/not married and civil marriage.

****

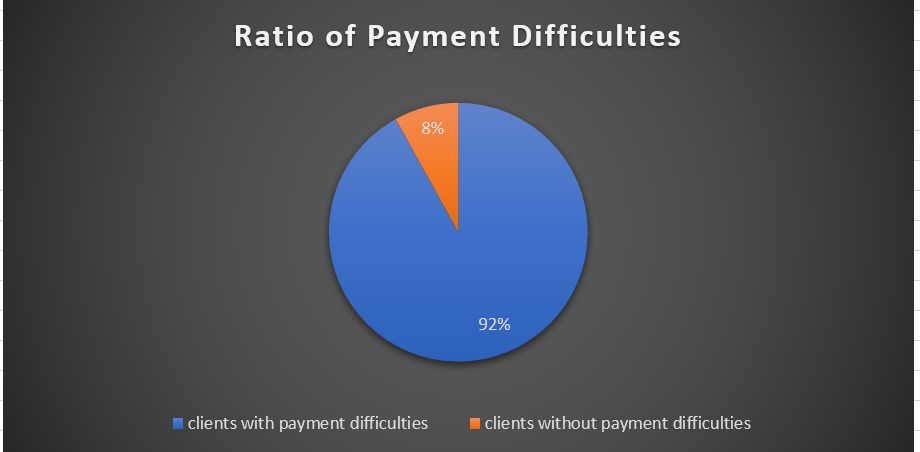
**Education Type:**

Secondary/secondary special education is of majority clients then followed by Higher education. Clients having academic degree has very small number.

****

**Count of Children:**

Clients with no children have higher frequency while other are less. It shows the imbalanced distribution in number of children feature.

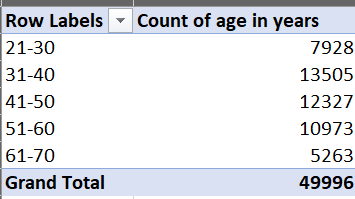
****

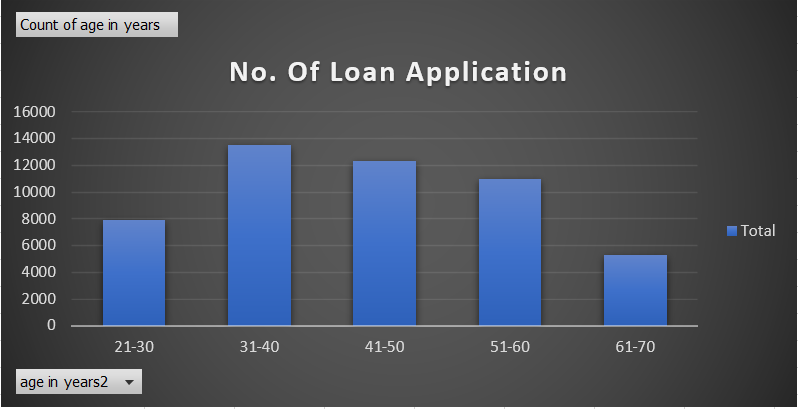
**Ratio of Payment Difficulties:**

We observe that most of the clients are facing difficulties in payment which 92% and only 8% are able to do payment without any difficulties. This shows that the data is imbalanced as the one class is under represented.

1. **Perform Univariate, Segmented Univariate, and Bivariate Analysis:**

**Univariate:**

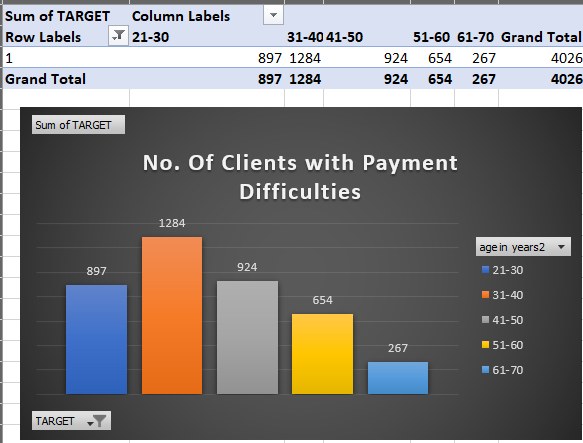




**Insight:**

Data shows that majority of loan applicants fall within 31-60 age group.

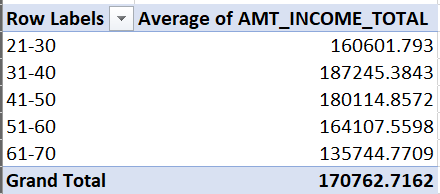
**Bivariate:**

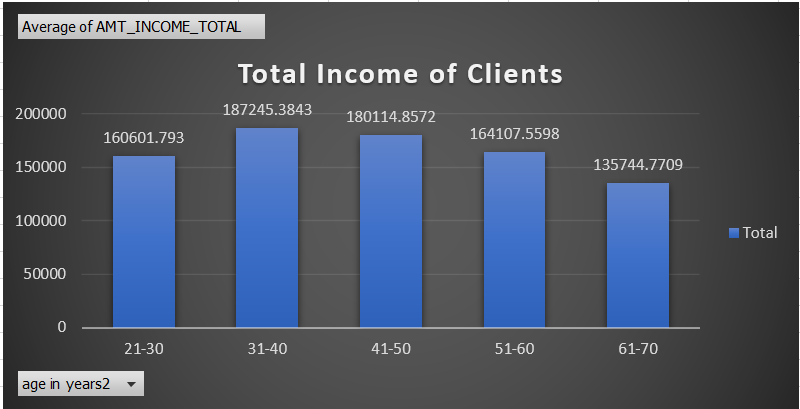


**Insight:**

The data shows that after age of 31 the number of clients with payment difficulties decreases as age increases. This gives us a insight that older clients may have more financial stability and experience in managing finances so they have less difficulties in payment of loan.

**Univariate:**

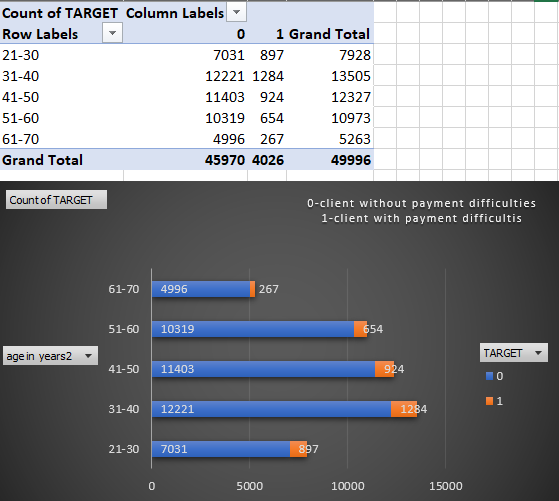




**Insight:**

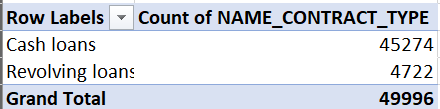
It indicates that clients within 31-60 age range tend to have higher incomes.

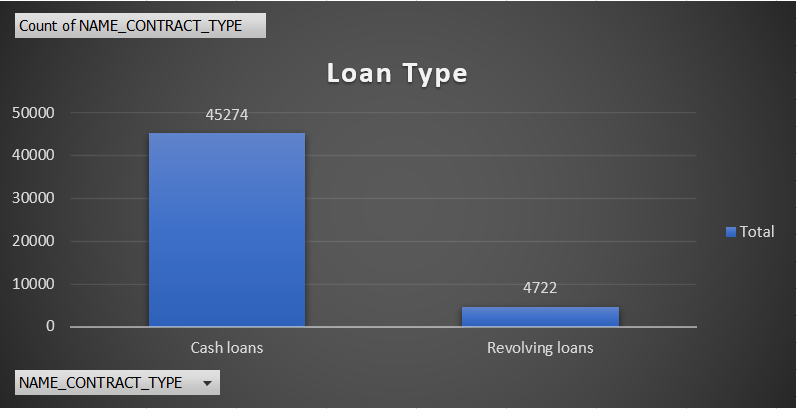
**Bivariate:**



**Insight:**

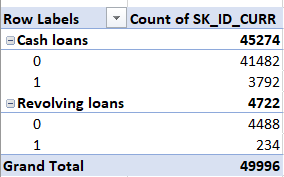
The data shows that clients with higher income generally do not experience payment difficulties. That shows the strong correlation between higher income levels and better financial stability.

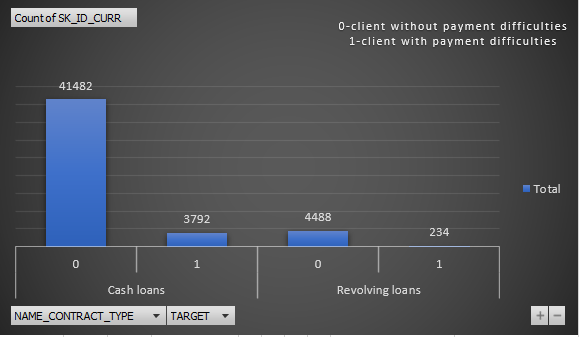
**Univariate:**



**Insight:**

The data shows that significant number of clients opt for cash loans compared to revolving loans.

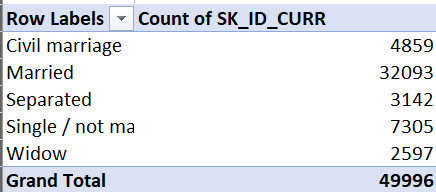
**Bivariate:**

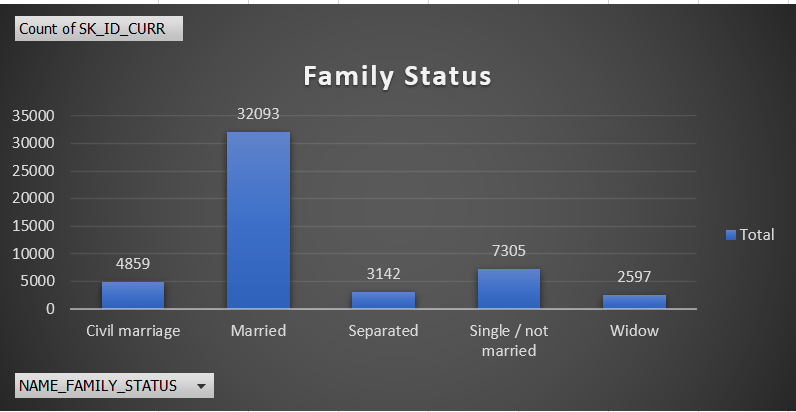


**Insight:**

We can observe that clients with payment difficulties are more relevant among who take cash loans as compared to revolving loans. This suggests the cash loans might pose greater risk for payment challenges, which might be due to higher loan amounts or less flexible repayment terms.

**Univariate:**

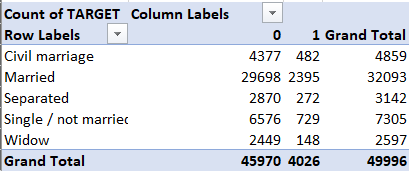


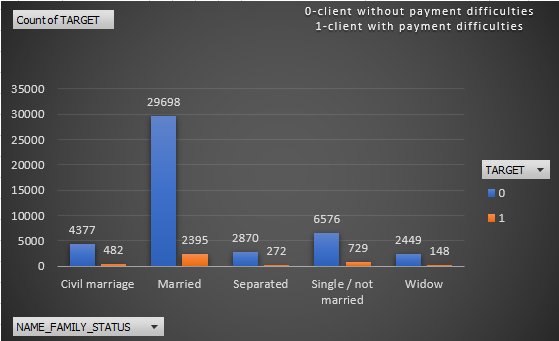


**Insight:**

We can observe that married individuals are more likely to take loans compared to others.

**Bivariate:**

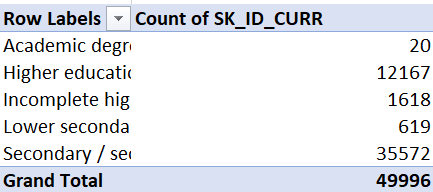


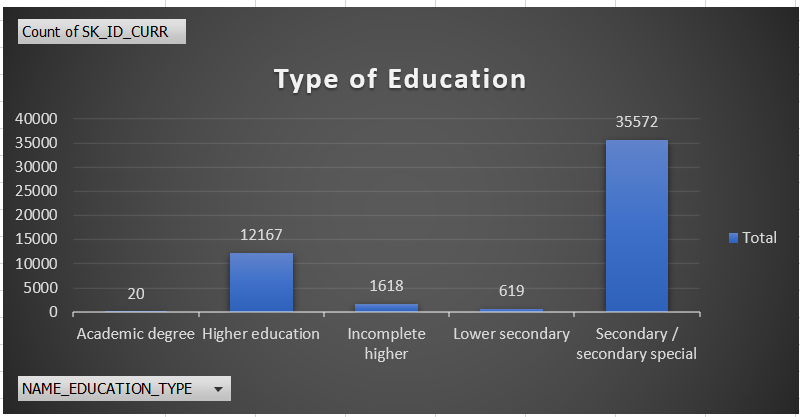


**Insight:**

We can observe that number of married clients experience lower payment difficulties as compared to who do not face payment difficulties. This gives us a insight that married individuals may have better payment behaviour.

**Univariate:**

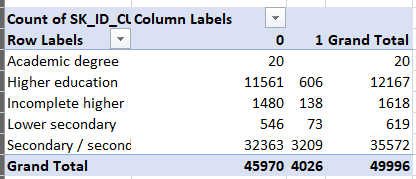


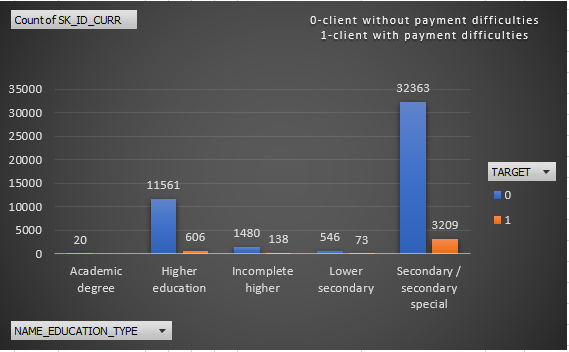


**Insight:**

The data shows that high number of loan applicants have secondary or secondary special education. Which gives us insight that many of this may be seeking financial assistance for pursuing studies abroad.

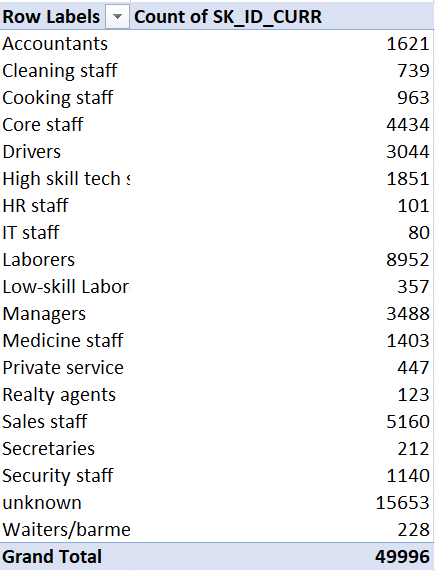
**Bivariate:**

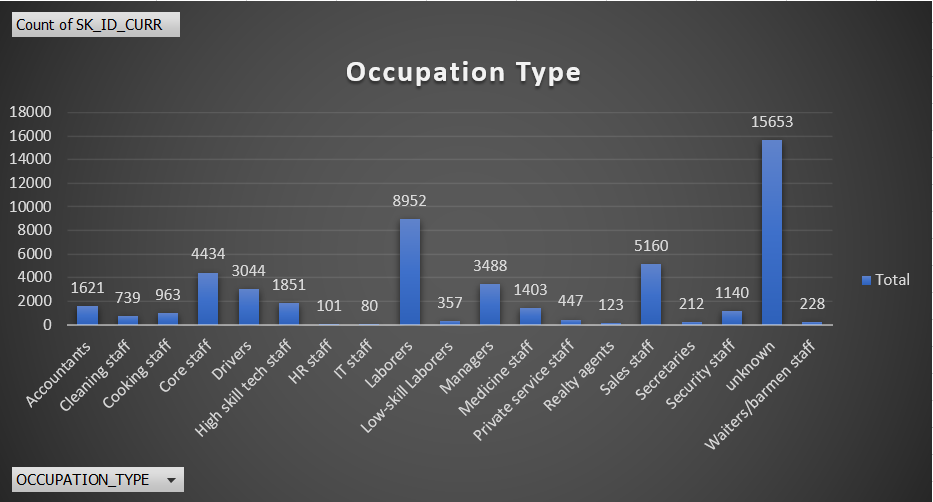




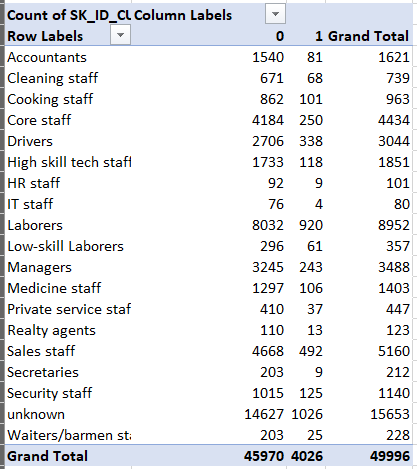
**Insight:**

It show that most of the individuals in secondary/secondary special education do not have payment difficulties, which suggests that they have higher incomes.

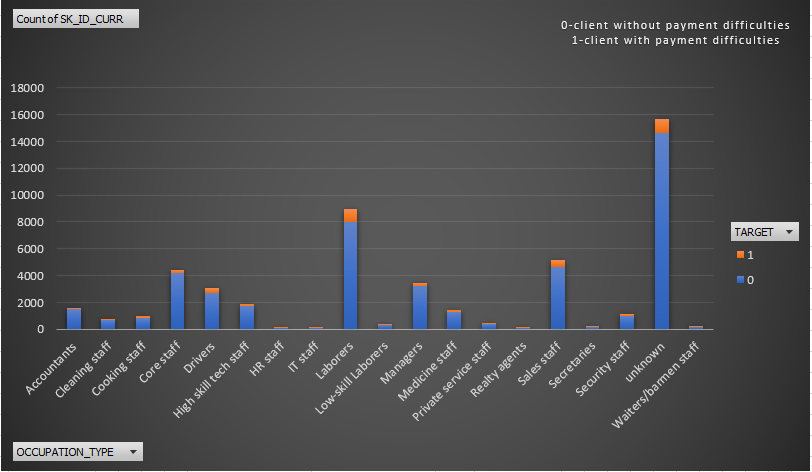
**Univariate:**



**Insight:**

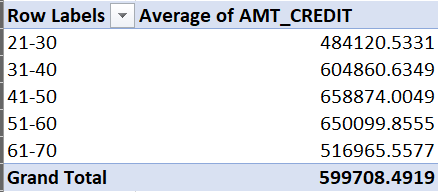
We can observe that applicants who are labourers are the highest which indicates lower financial situation among this group.

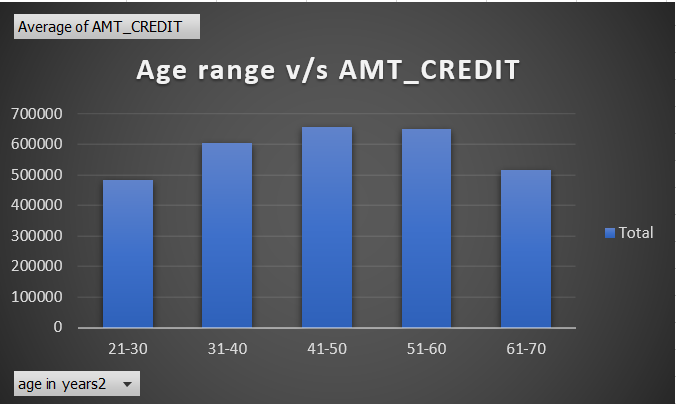
**Bivariate:**



**Insight:**

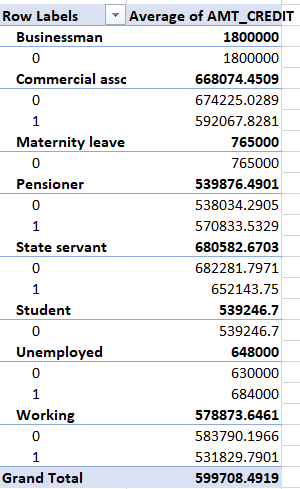
We can observe clearly that applicants from labourers are highest in both categories of clients with payment difficulties and those without. This indicates labourers are significantly represented in loan applications across different financial situations. This reflects their broader need for financial assistance.

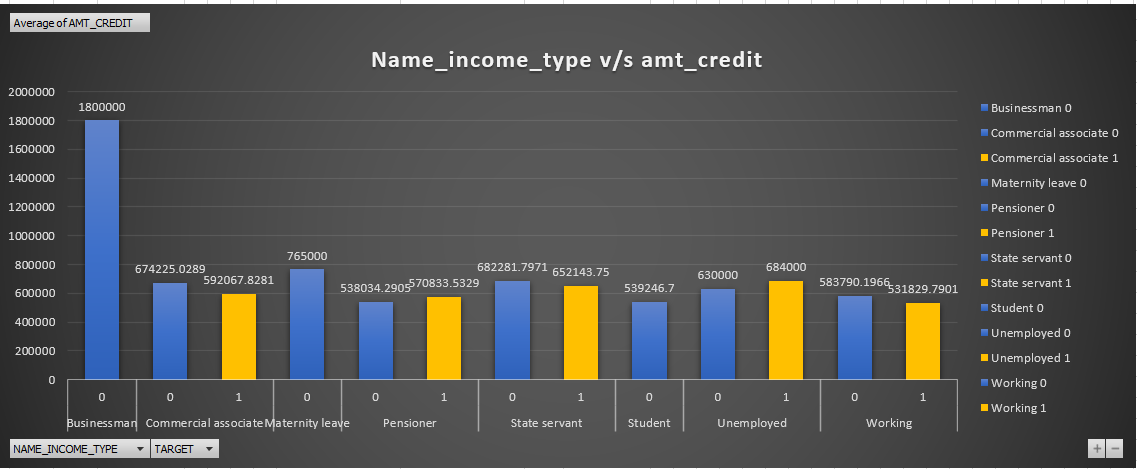
**Univariate:**



**Insights:**

We can observe that individuals within 31-60 age range have higher credit scores. This gives us insight that clients in this age group generally maintain better credit, which my be due to more financial stability and credit history.

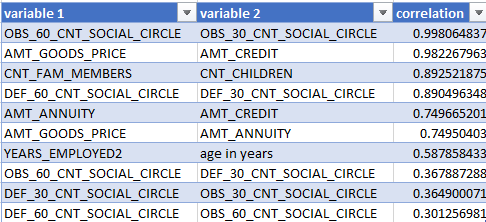
**Bivariate:**

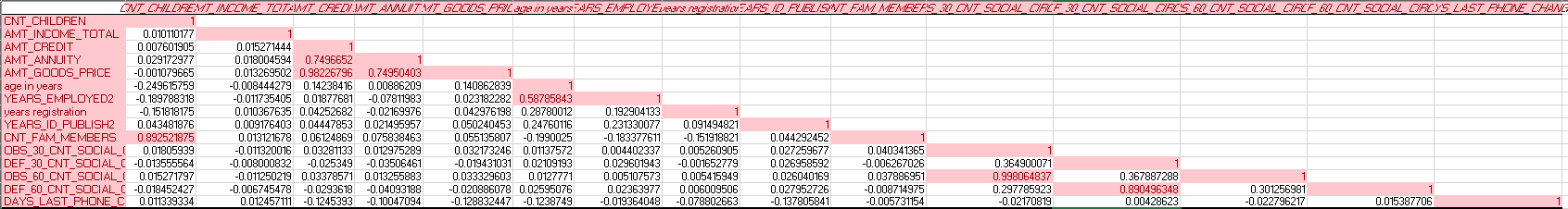


**Insight:**

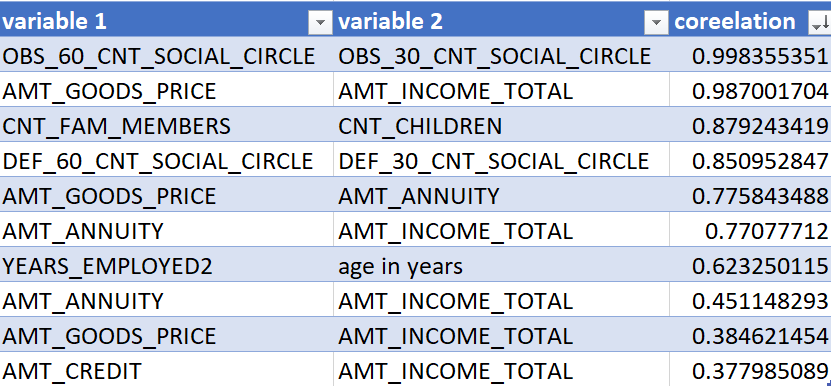
We can observe that clients with category of businessman, maternity leaves, state servant and commercial associate do not experience payment difficulties. The clients who are business man have higher credit score.

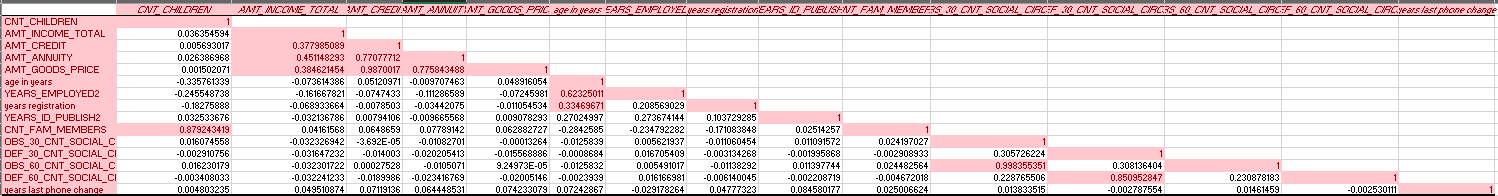
1. **Identify Top Correlations for Different Scenarios:**

**The top 10 correlation between different variables and client who are defaulter is:**

****

**The top 10 correlation between different variable and clients who are repayer**

****

****

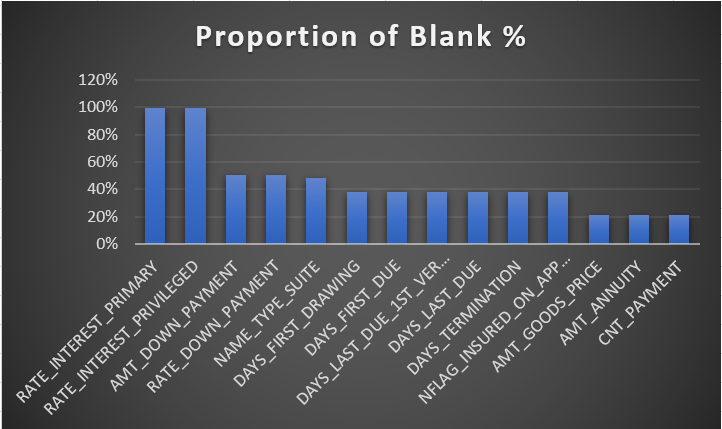
Drive link:

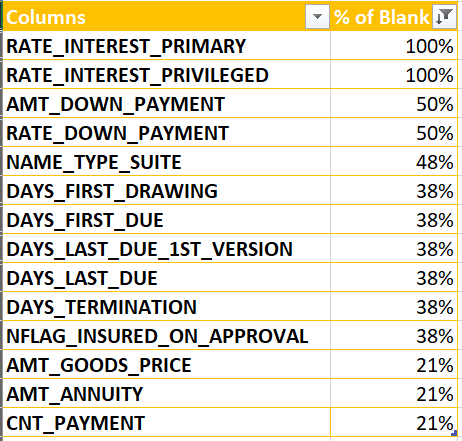
[**https://docs.google.com/spreadsheets/d/14lAwEXxC09Dxz\_GIuOaeyzTIn3rLk5YT/edit?usp=sharing&ouid=109001208060904860088&rtpof=true&sd=true**](https://docs.google.com/spreadsheets/d/14lAwEXxC09Dxz_GIuOaeyzTIn3rLk5YT/edit?usp=sharing&ouid=109001208060904860088&rtpof=true&sd=true)

**FOR PREVIOUS APPLICATION**

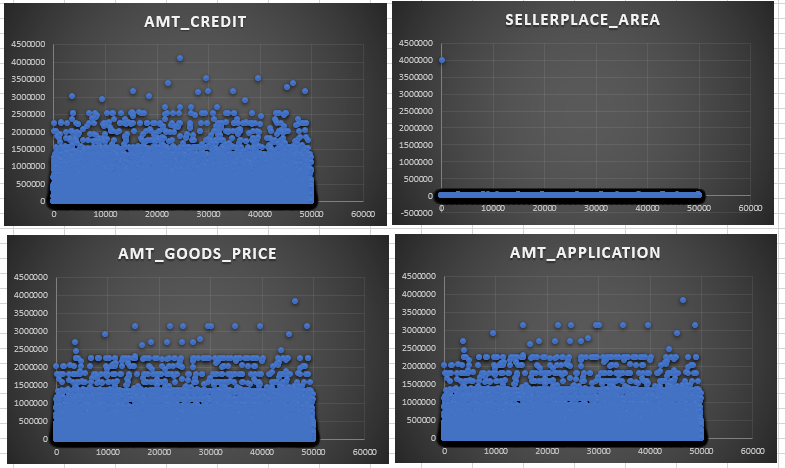
Similarly performing all the tasks for previous application dataset

1. **Identify Missing Data and Deal with it Appropriately:**

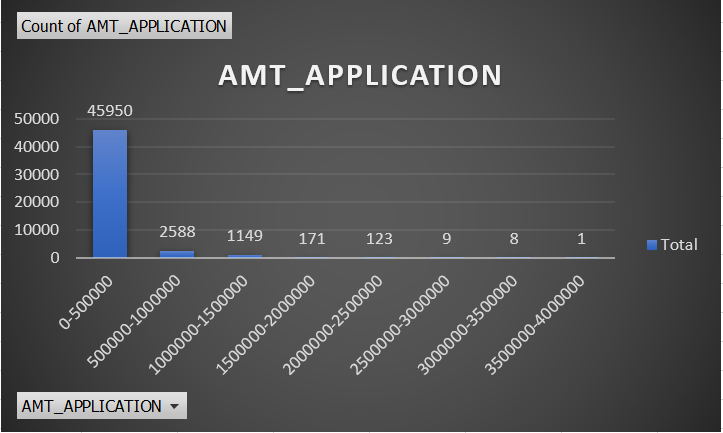




1. **Identify Outliers in the Dataset:**

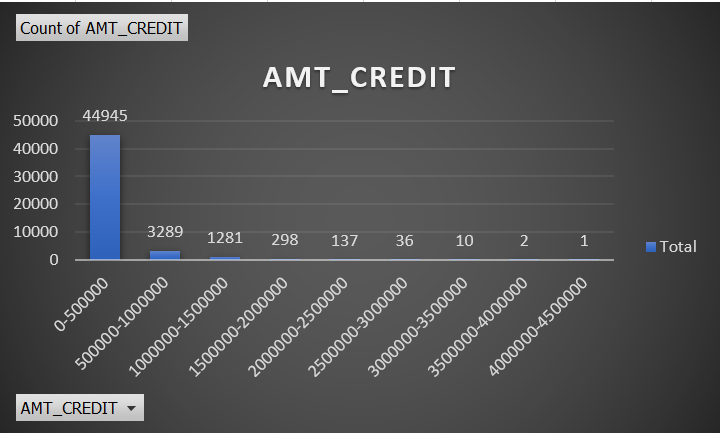


1. **Analyze Data Imbalance:**

****

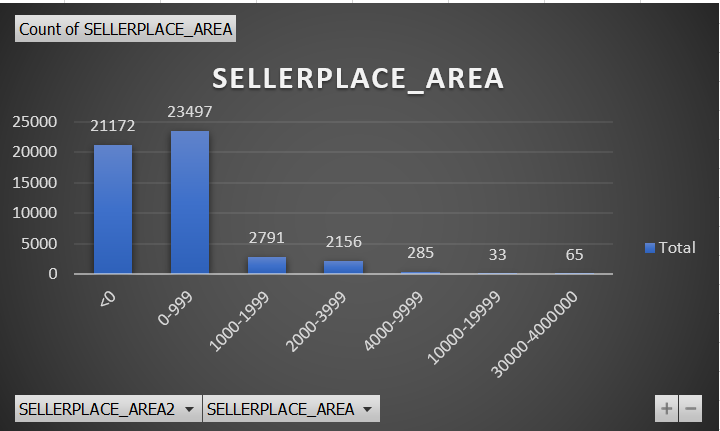
**Insights:**

It contains highly imbalanced data. As 0-500000 contains 90% of the data.



**Insights:**

Majority clients belong to credit group of 0-500000. It also contains 90% of loan applications.

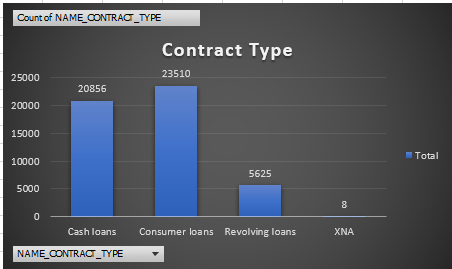


**Insights:**

It contains outliers and also highly imbalanced data as area between 0-9999 contains maximum of the data.

1. **Perform Univariate, Segmented Univariate, and Bivariate Analysis:**

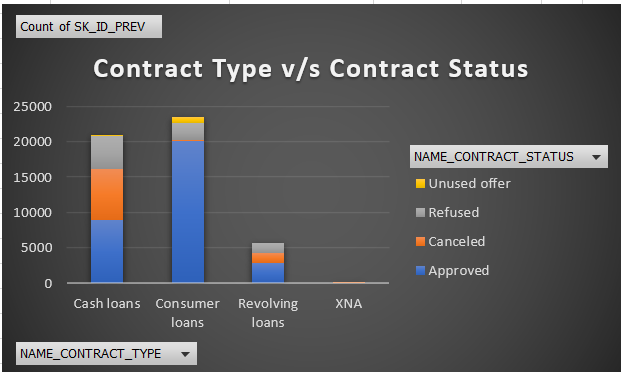
**Univariate:**



**Insights**:

It is observed that consumer of the bank are taking the most amount of the loans.

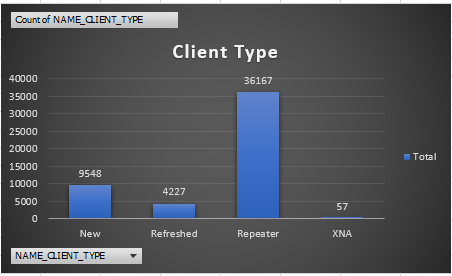
**Bivariate**:



**Insights**:

We can observe that most of the consumer loans were approved and none where cancelled.

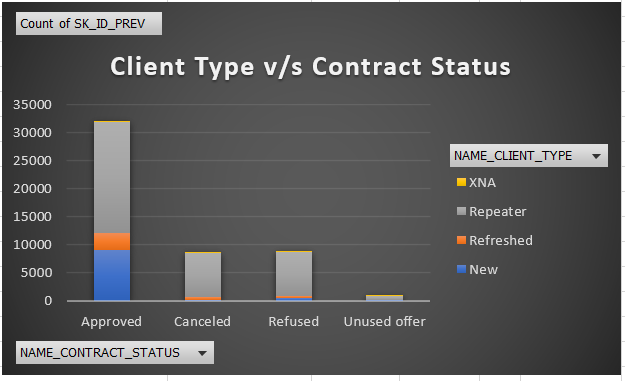
**Univariate**:



**Insights:**

It is observed that most of the clients are repeaters and it will make the process easy as they are aware about the process.

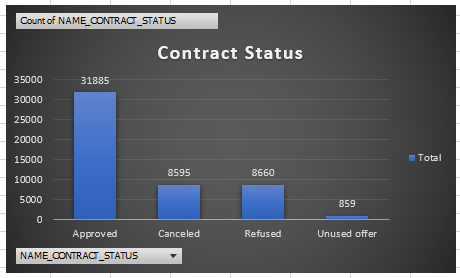
**Bivariate**:



**Insights**:

We can observe that repeaters are highly approved and new where the second. And the maximum cancelled were also repeaters.

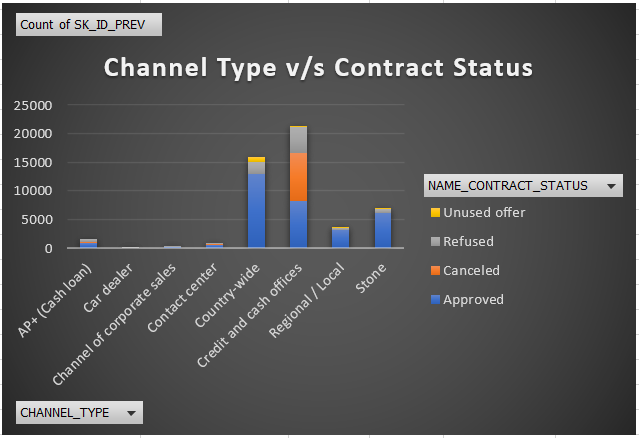
**Univariate**:



**Insights**:

It is seen that most of the loans are approved. This tells us that most of the applicants are bonified.

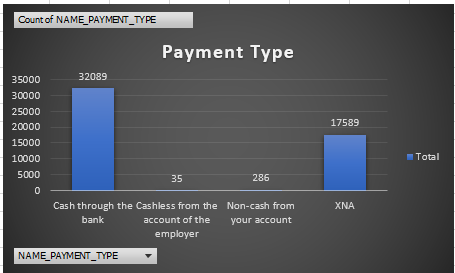
**Bivariate**:



**Insights**:

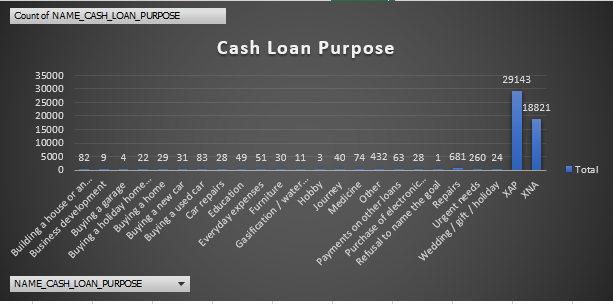
We can say that majority of loans approved were of country-wide channel followed by credit and cash offices and the stone. Also cash and credit offices also have maximum rejections.

**Univariate**:



We can observe that most of the payments are done through cash by the clients. This show people are promoting banking lines of payment.

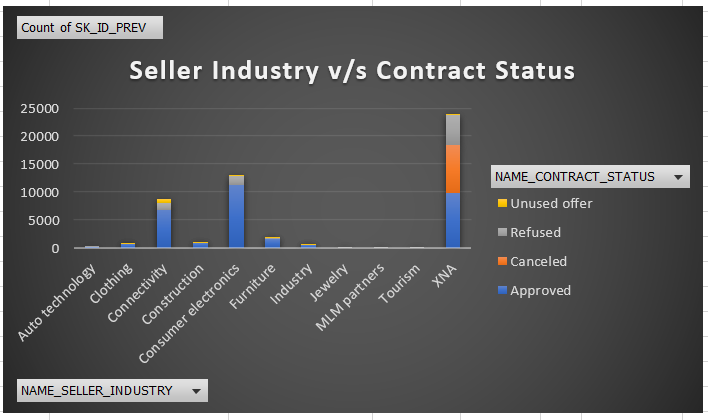
**Univariate**:



**Insights**:

We can observe that most of the clients have purpose of taking loan as XAP followed by XNA. Both in combine constitute 95% of loan application purpose.

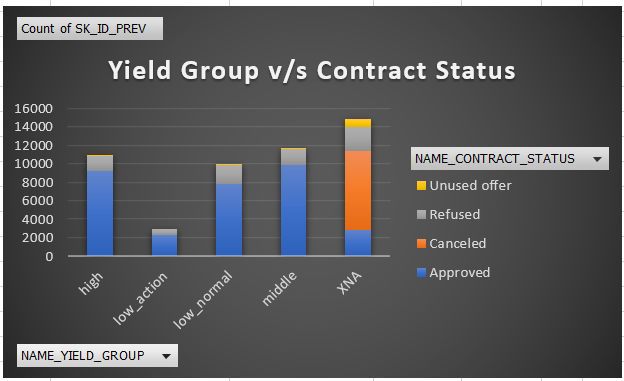
**Bivariate**:



**Insights**:

We can see that consumer electronic industry gets most of the loan approved followed by XNA and then connectivity.

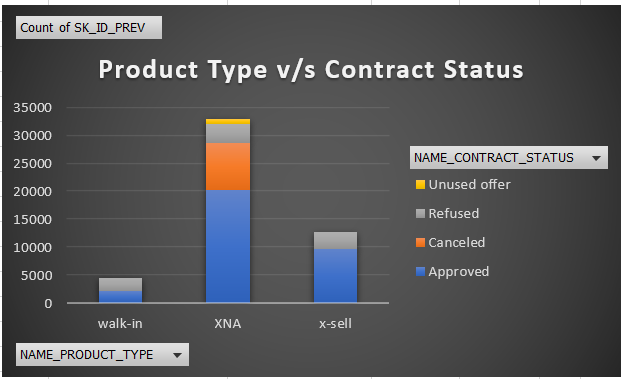
**Bivariate**:



**Insights**:

By observing the graph we can tell that group interest rate into small medium and high shows middle rate is highest approved followed by high and others are more or less same.

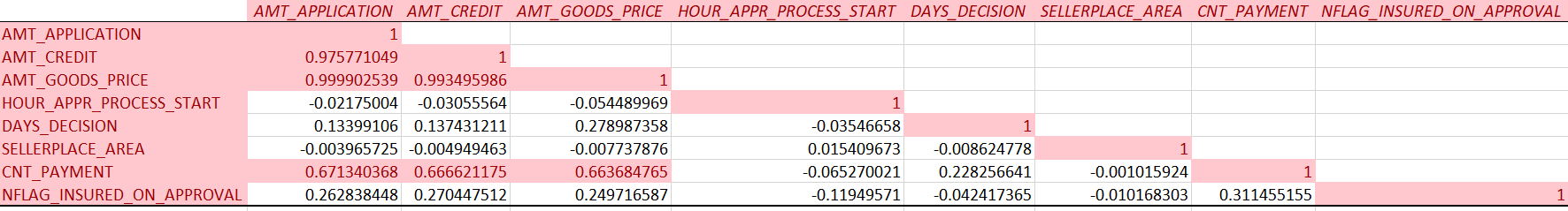
**Bivariate**:

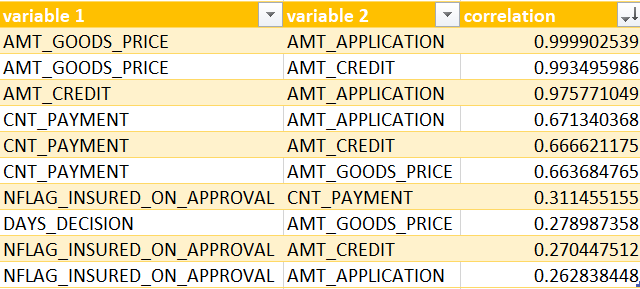


**Insights**:

It is observed that XNA is highly approved and followed by X-sell. Also XNA were also most cancelled.

1. **Identify Top Correlations for Different Scenarios:**





**Drive Link:**

[**https://docs.google.com/spreadsheets/d/1uSWZS87SSBVZpexxIr2B6UEDpoiA7Iu\_/edit?usp=sharing&ouid=109001208060904860088&rtpof=true&sd=true**](https://docs.google.com/spreadsheets/d/1uSWZS87SSBVZpexxIr2B6UEDpoiA7Iu_/edit?usp=sharing&ouid=109001208060904860088&rtpof=true&sd=true)

**Conclusion**

* The dataset was observed to be imbalanced with one class as being underrepresented.
* The major demographic age group for seeking loan were observed to be between 31-60.
* The age group of 31-60 where observed to have higher incomes which suggests the prime stage of earning in life.
* The clients having higher incomes generally did not face difficulties in payment of loan.
* It was observed cash loans were the most common loan type which gave us insight of liquidity needs among the clients.
* Clients with secondary/secondary special education were more likely to take loans which showed us correlation between education level and loan dependency.
* Married individuals are more likely to apply for loan which suggest greater financial obligations or goals.
* Largest occupation group of loan applicants is observed of labourers which tells financial need in this sector.
* There large amount of repeat clients suggesting customers returning more often for additional loans.
* Age group between 31-60 tend to have more credit score, further establishing this age group as financially stable and reliable.

**Loom video link**

[**https://www.loom.com/share/7cec464278554ca9977d04f4a2b17372?sid=d7abcf55-1141-47fe-aacf-f5e2300c3548**](https://www.loom.com/share/7cec464278554ca9977d04f4a2b17372?sid=d7abcf55-1141-47fe-aacf-f5e2300c3548)