Trainity Assignment-6

Muthiah Sivavelan S
Ph: 8525021258

Excel Document: Bank Loan Case Study

Project Description:

Imagine you're a data analyst at a finance company that specializes in lending various types of loans to urban customers. Your company faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans. Your task is to use Exploratory Data Analysis (EDA) to analyse patterns in the data and ensure that capable applicants are not rejected.

The goal in this project is to use EDA to understand how customer attributes and loan attributes influence the likelihood of default.

Through this project, we also try to identify patterns that indicate if a customer will have difficulty paying their instalments. 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.

Tech-Stack Used:

I have used Microsoft Excel 2019 for this project, since that's the version that came preinstalled with my laptop and it has most of functionalities similar to Microsoft Excel 2022.

Tasks:

A. **Identify Missing Data and Deal with it Appropriately**: As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

Approach:

First, I have checked the relevant columns used for this project. Then, I imported the necessary files required (application_data.csv and previous_application.csv). I, then deleted the irrelevant columns that are not required for this project. For the remaining columns, I imputed the missing values with the median value/ mode value/ mean value/ logical imputing/ linear regression, etc.

- Occupation_Type column:

I removed the spaces surrounding the words in each row and removed any invisible spaces available. Then, I imputed the blanks with value Unknown. I then, found out the mode of this column(excluding Unknown) based on their counts and imputed with this mode value.

- External sources columns:

I imputed the columns with the median value for each column.

- AMT_GOODS_PRICE column:

I did mean imputation for this column.

- AMT_ANNUITY column:

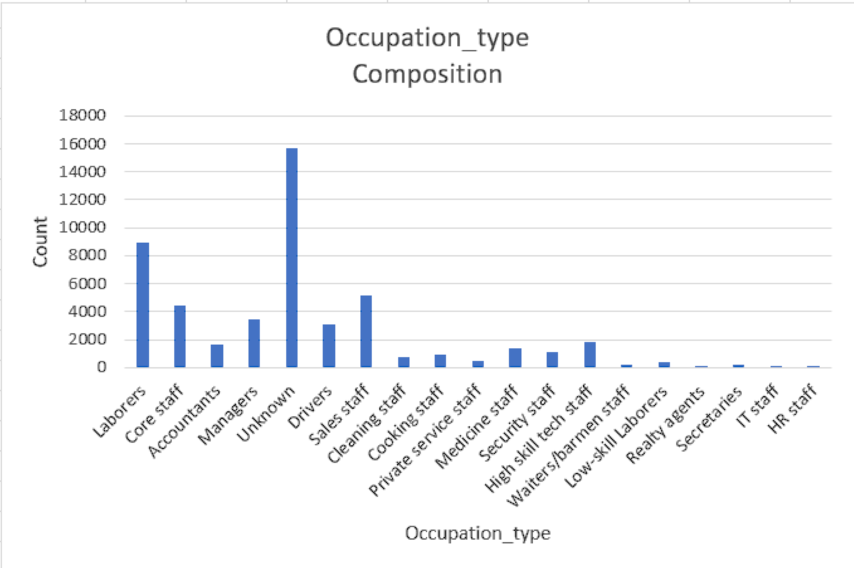
For this, column, I first found out the correlation with AMT_INCOME_TOTAL column, AMT_CREDIT column, AMT_GOODS_PRICE column and I found out that the correlation with AMT_CREDIT and AMT_GOODS_PRICE column is very high and I decided to use Linear Regression for finding the coefficients and imputed the value using linear regression.

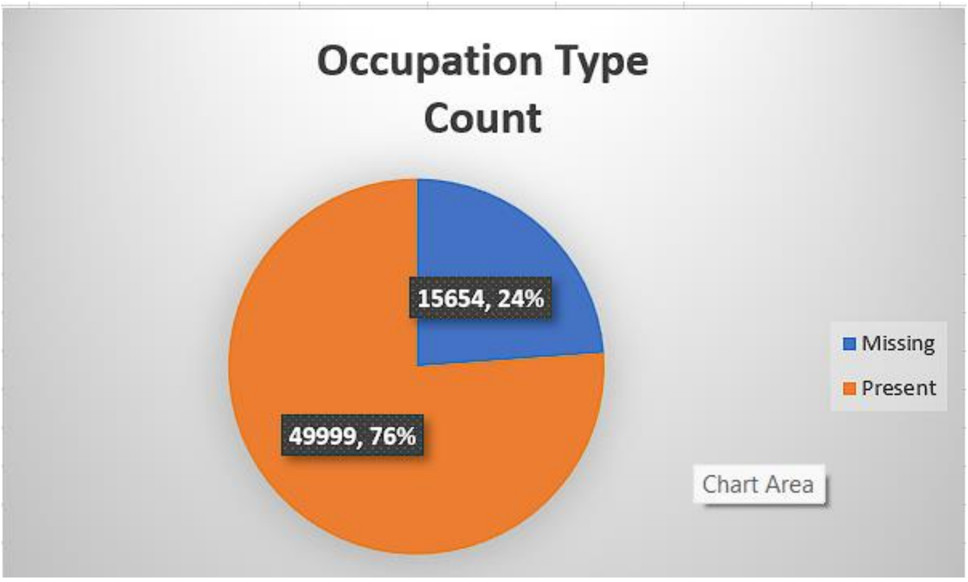
- NAME_TYPE_SUITE column:

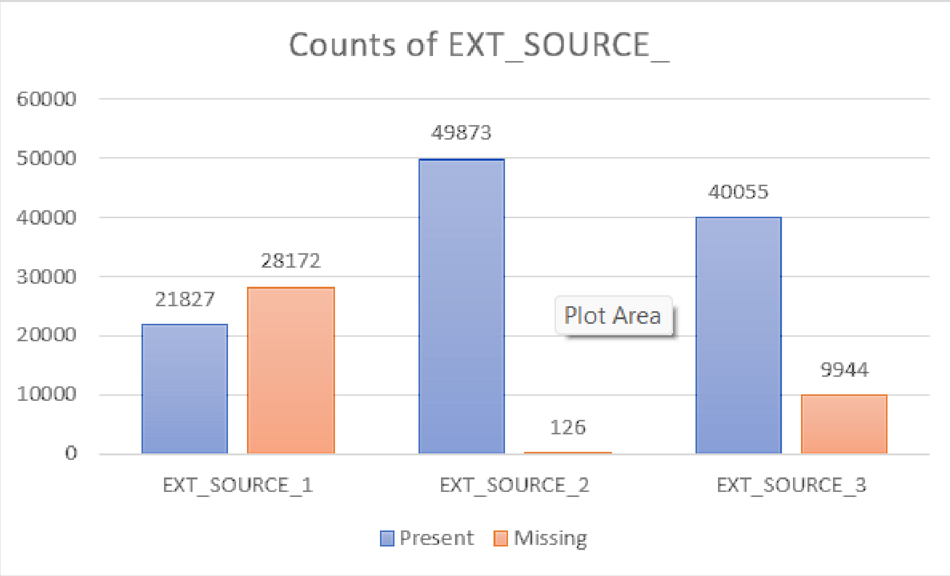
For this column, I used NAME_FAMILY_STATUS as a helper column for imputation. My logic of imputation was if family status is single/not married, widow or separated, the NAME_TYPE_SUITE column should be Unaccompanied while if the family status is married or civil marriage, the NAME_TYPE_SUITE column is

Spouse, partner. If the NAME_FAMILY_STATUS is Unknown then the NAME_TYPE_SUITE column will also be unknown. In the second step of imputation, I imputed the unknown values with the maximum occurrence of the NAME_TYPE_SUITE column(mode value, i.e., Unaccompanied).

Distinct_Occupation	Count
Laborers	8952
Core staff	4434
Accountants	1621
Managers	3489
Unknown	15654
Drivers	3044
Sales staff	5160
Cleaning staff	739
Cooking staff	963
Private service staff	447
Medicine staff	1403
Security staff	1140
High skill tech staff	1852
Waiters/barmen staff	228
Low-skill Laborers	357
Realty agents	123
Secretaries	212
IT staff	80
HR staff	101





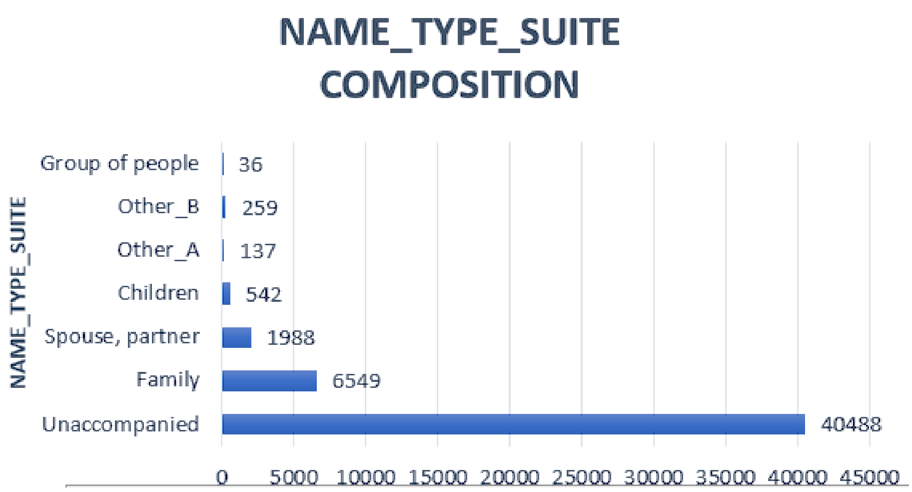


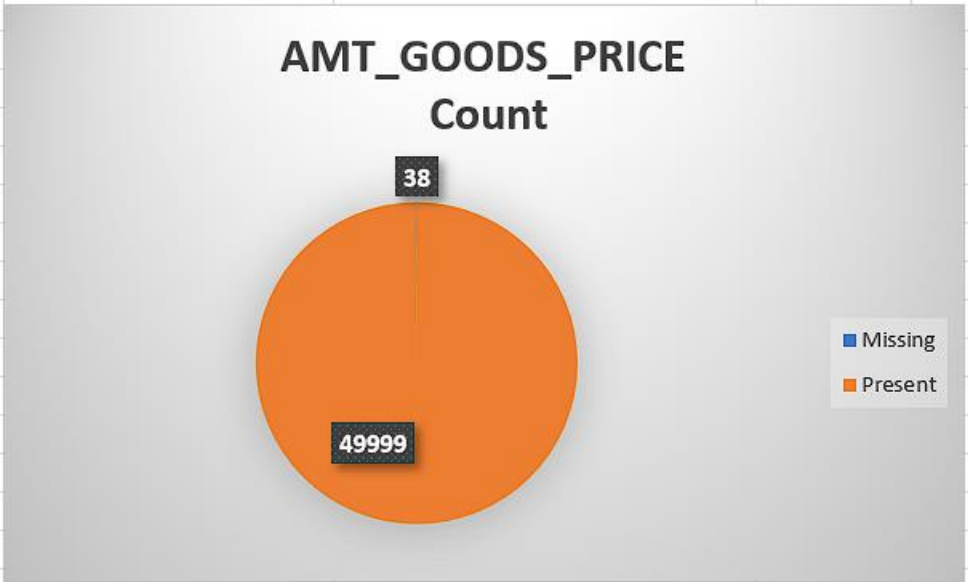
Name_Family_Status	NAME_TYPE_SUITE_MAPPING
Single / not married	Unaccompanied
Married	Spouse, partner
Civil marriage	Spouse, partner
Widow	Unaccompanied
Separated	Unaccompanied
Unknown	Unknown

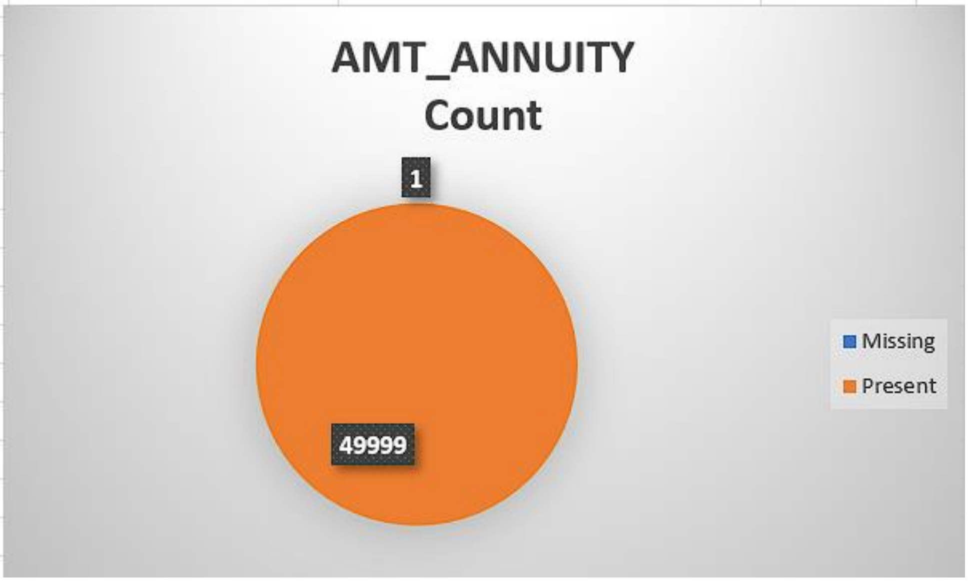
Number of unknown in imputed column 1:

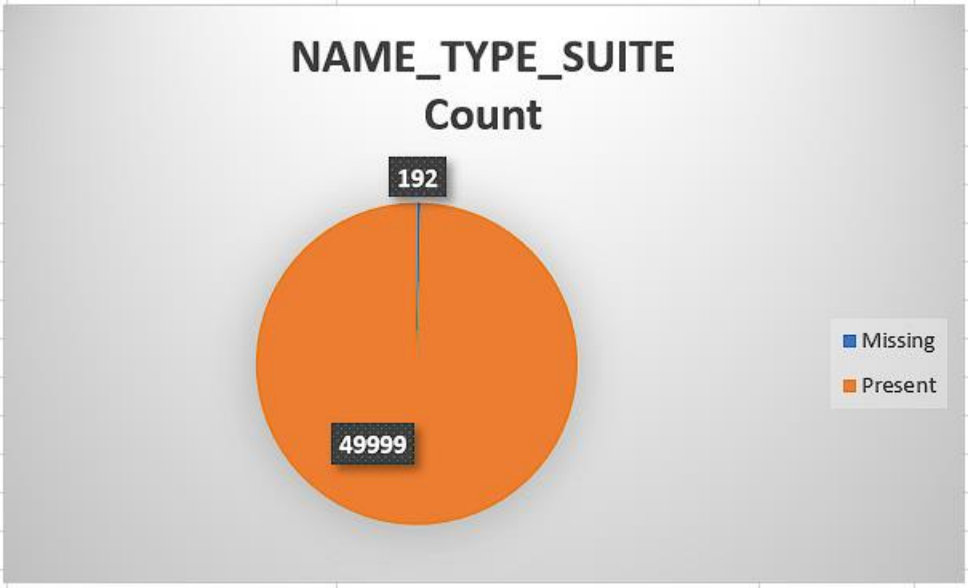
1

We will use the maximum occurrence to impute the rows with Unknown value




AMT_GOODS_PRICE Count
38
49999
Missing
Present


AMT_ANNUITY Count
1
49999
Missing
Present


NAME_TYPE_SUITE Count
192
49999
Missing
Present

B. Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

- **Task:** Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

Approach:

I copied the required numberical variables in a separate sheet after imputing the missing values. I, then created a table containing the required columns as the rows and the columns as 25th Quartile, 75th Quartile, IQR(Inter Quartile Range), Lower_Bound value, Upper_Bound value.

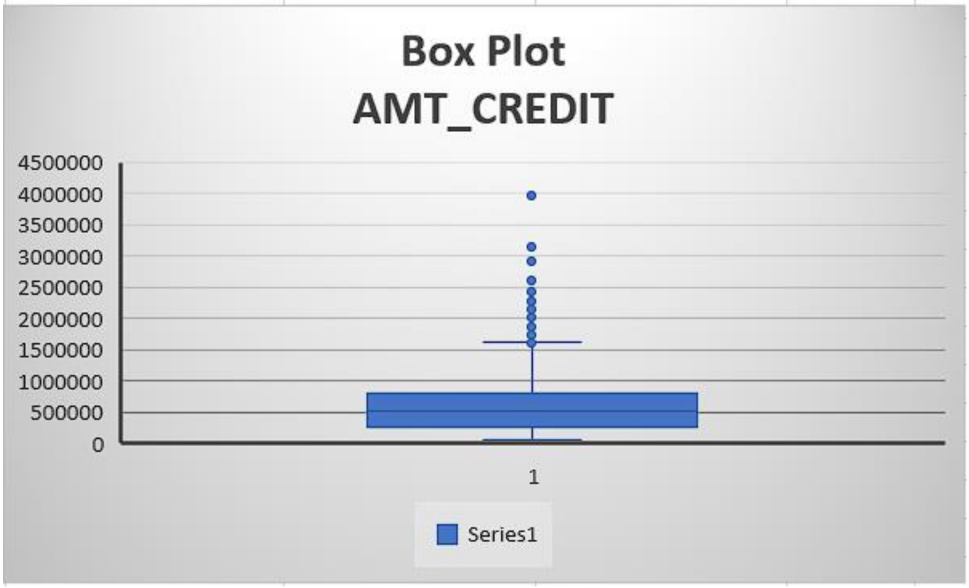
I used the following formulas for the respective columns to fill up the table.

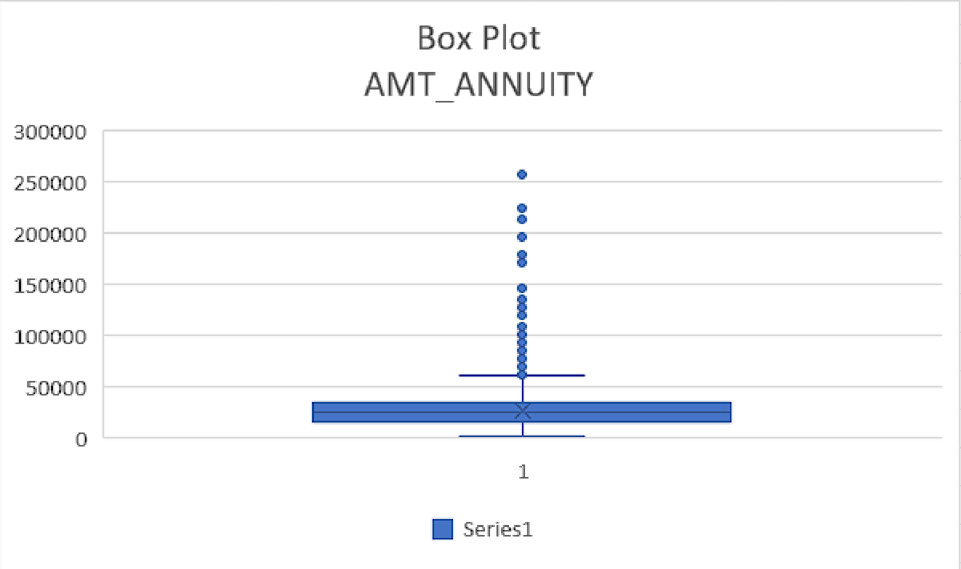
25th Quartile:
=QUARTILE.EXC(Table15[AMT_INCOME_TOTAL],1)
75th Quartile:
=QUARTILE.EXC(Table15[AMT_INCOME_TOTAL],3)
IQR:
=[@[75th Quartile(Q3)]]-[@[25th Quartile(Q1)]]
Lower_Bound:
=[@[25th Quartile(Q1)]]-1.5*[@IQR]
Upper_Bound:
=[@[75th Quartile(Q3)]]+1.5*[@IQR]

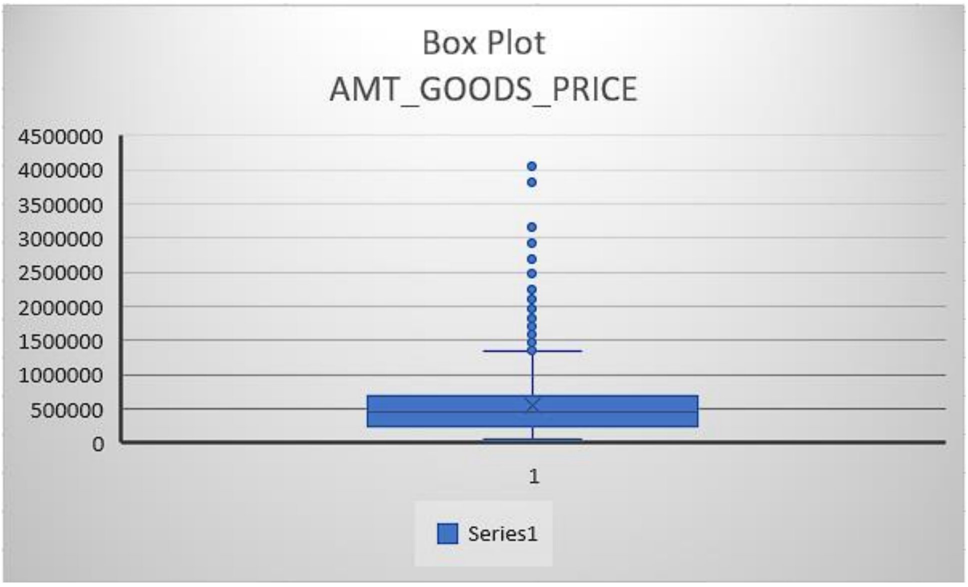
The values which are not in the range of Lower Bound and Upper Bound are the outliers. I, then verified this using the box plots for the required columns. I used conditional formatting to highlight the values that are not in the range of Lower_Bound and Upper_Bound.

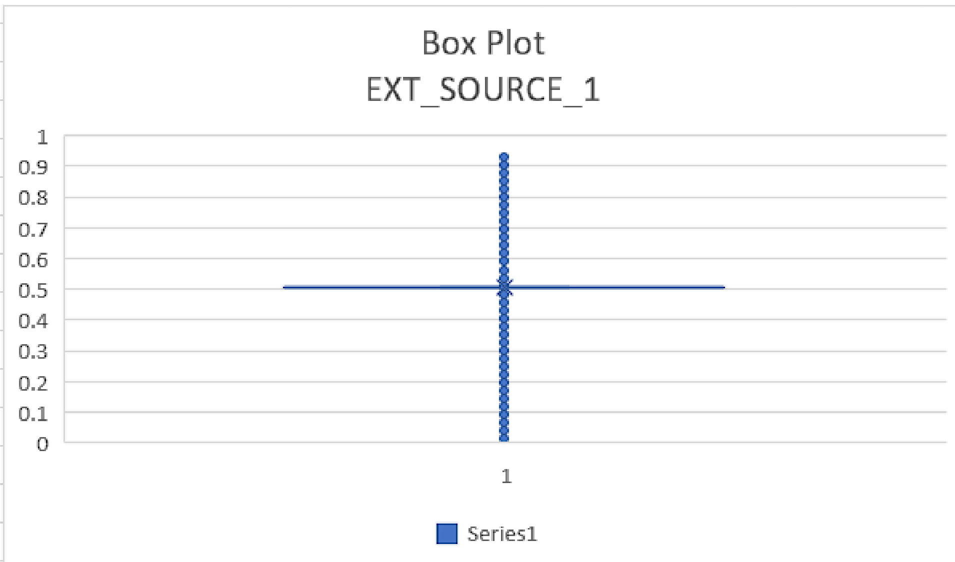
AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	EXT_SOURCE_1	EXT_SOURCE_2	EXT_SOURCE_3
202500	406597.5	24700.5	351000	0.083036967	0.262948593	0.13937578
270000	1293502.5	35698.5	1129500	0.311267311	0.622245775	0.53527625
67500	135000	6750	135000	0.506883944	0.555912083	0.729566691
135000	312682.5	29686.5	297000	0.506883944	0.65044169	0.53527625
121500	513000	21865.5	513000	0.506883944	0.322738287	0.53527625
99000	490495.5	27517.5	454500	0.506883944	0.354224732	0.621226338
171000	1560726	41301	1395000	0.774761413	0.723999852	0.492060094
360000	1530000	42075	1530000	0.506883944	0.714279286	0.54065445
112500	1019610	33826.5	913500	0.587334047	0.205747288	0.751723715
135000	405000	20250	405000	0.506883944	0.746643629	0.53527625
112500	652500	21177	652500	0.319760172	0.651862333	0.363945239
38419.155	148365	10678.5	135000	0.72204445	0.555183162	0.652896552
67500	80865	5881.5	67500	0.464831117	0.715041819	0.176652579
225000	918468	28966.5	697500	0.506883944	0.566906613	0.77008707
189000	773680.5	32778	679500	0.721939769	0.642656205	0.53527625
157500	299772	20160	247500	0.115634337	0.346633981	0.678567689
108000	509602.5	26149.5	387000	0.506883944	0.23637784	0.062103038
81000	270000	13500	270000	0.506883944	0.683513346	0.53527625

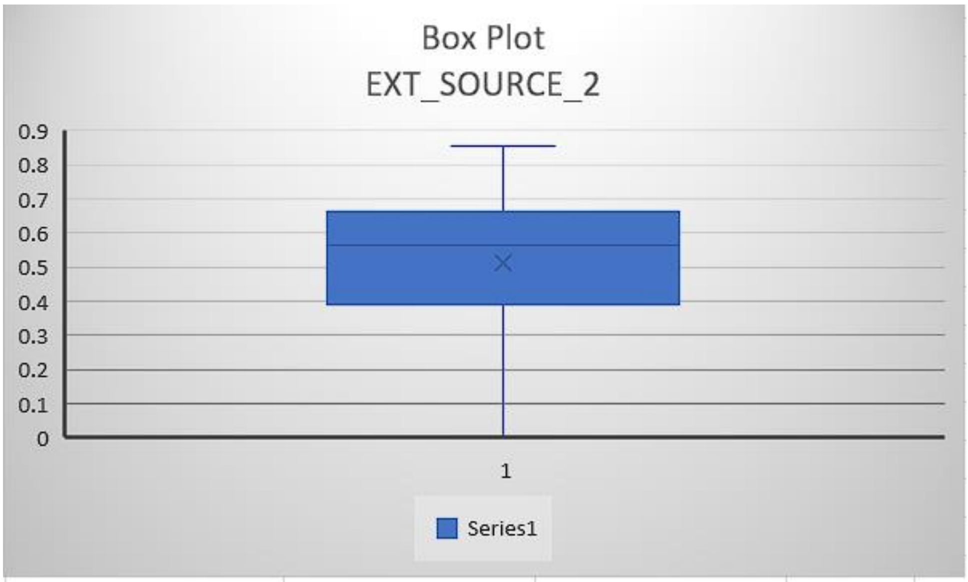
Name	25th Quartile(Q1)	75th Quartile(Q3)	IQR	Lower_Bound	Upper_Bound
AMT_INCOME_TOTAL	112500	202500	90000	-22500	337500
AMT_CREDIT	270000	808650	538650	-537975	1616625
AMT_ANNUITY	16456.5	34596	18139.5	-10752.75	61805.25
AMT_GOODS_PRICE	238500	679500	441000	-423000	1341000
EXT_SOURCE_1	0.506883944	0.506883944	0	0.506883944	0.506883944
EXT_SOURCE_2	0.392217599	0.66316296	0.270945	-0.014200443	1.069581002
EXT_SOURCE_3	0.417099668	0.638043528	0.220944	0.085683878	0.969459318

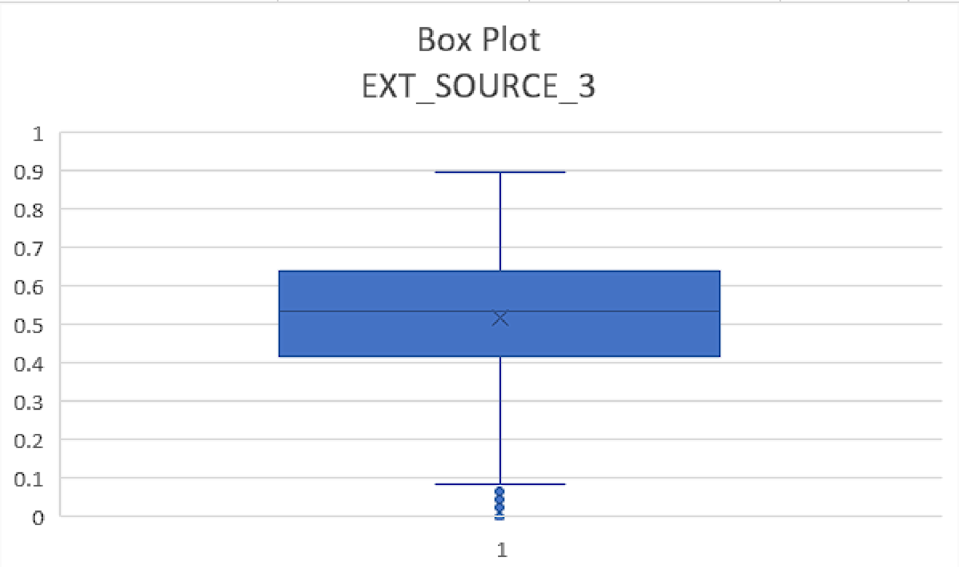



Box Plot
AMT_ANNUITY
300000
250000
200000
150000
100000
50000
0
1
Series1


Box Plot
AMT_GOODS_PRICE
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1
Series1


Box Plot
EXT_SOURCE_1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
Series1





C. Analyze 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.

- **Task:** Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

Approach:

I copied the target column into a new sheet and removed the duplicates. I created a new table with this column and added the columns, Frequency and Percentage. I used the countif function to find the frequency and for the percentage, I divided the frequency with total number of rows and multiplied by 100.

I found that 91.94 % of the target column, contained the value 0 and only 8.05 % of the target column contained the value 1. This shows there is imbalance in the data.

I found the ratio of data imbalance by calculating the ratio between minimum percentage to maximum percentage, i.e, 8.05/91.94

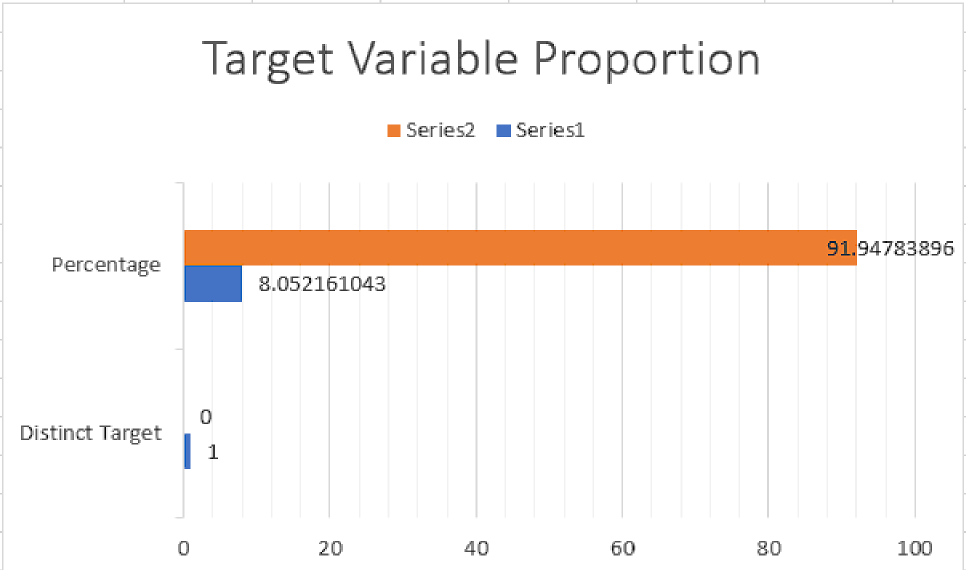
I visualised the results using pie chart, horizontal bar chart.

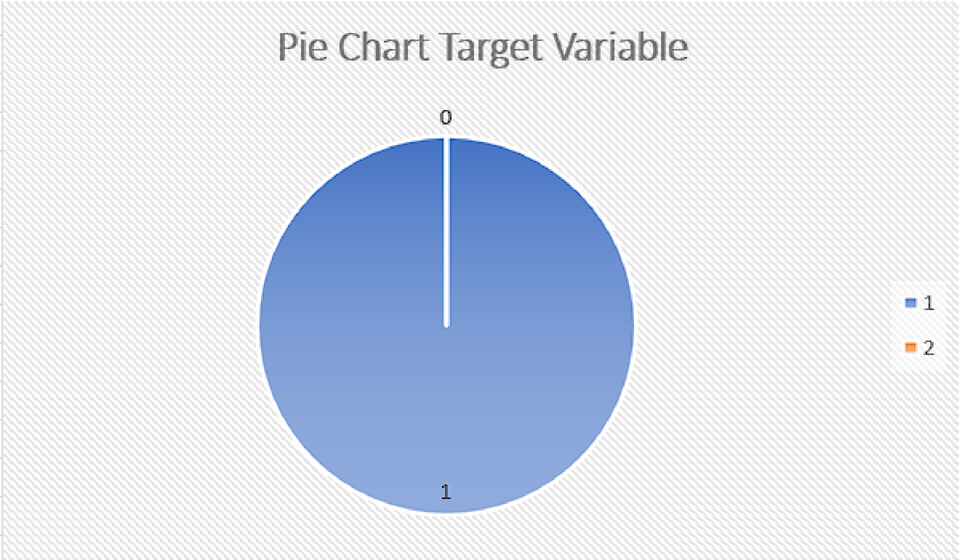
Distinct Target	Frequency	Percentage
1	4026	8.052161043
0	45973	91.94783896
	Total	49999

This definitely indicates the data imbalance.

Frequency of 0 class in target variable is much higher than frequency of 1 class in target variable.

Ratio of data imbalance	0.087573141





D. 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.

- **Task:** Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

Approach:
I have done the univariate analysis first itself, while imputing the various columns (finding the mean, median, mode of numerical columns and stuff). My main focus here, was to find the relation of each variable to the target variable.

Demographic factors:

- Gender & payment difficulty:

  I created a pivot table from the gender column and target column where the rows contained the distinct values of gender column, columns contained the distinct values of target column and the values themselves are the count of target variable. After this, I found the proportion facing difficulty and I found out that 6.898 % of the females are facing payment difficulties while 10.26 % of the men are facing payment difficulties.

From this, I arrived at a conclusion that men face more difficulties in payment compared to women.

Count of TARGET	Column Label			Proportion facing difficutly
Row Labels	0	1	Grand Total	
F	30559	2264	32823	6.897602291
M	15412	1762	17174	10.25969489
XNA	2		2	
Grand Total	45973	4026	49999	Men face more difficulties in payment compared to women.

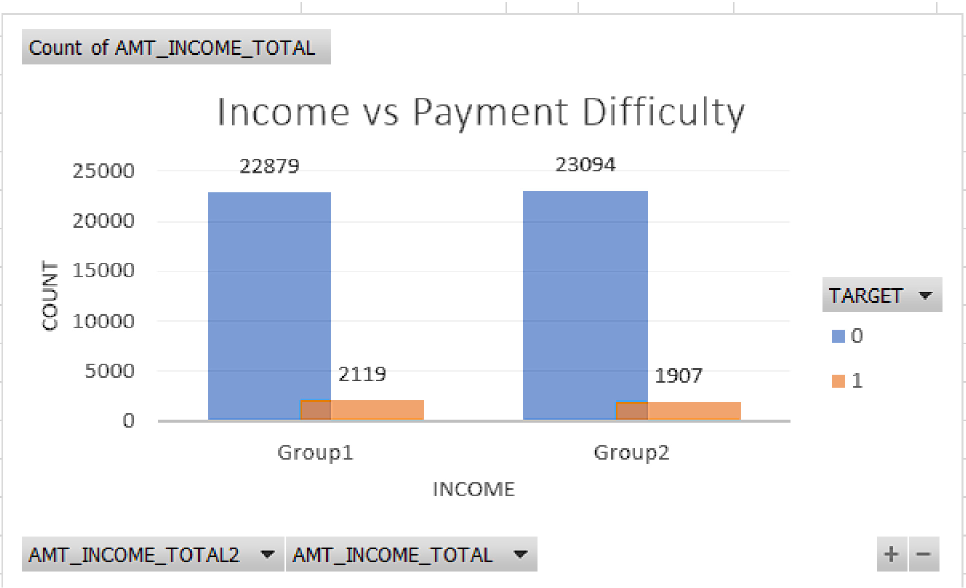
- Income & Payment Difficulty:

First, I found the median of AMT_INCOME_TOTAL column. Then, I created a pivot table containing the target as columns and income as the rows, count of target variable as the values. I grouped the rows into two groups based on the median value (Group1- lower income group and Group2- higher income group).

I found the proportion of the population facing payment difficulties in both the groups and I came to a conclusion that, people with lower income face more difficulties towards payment when compared to people with higher income.

Count of AMT_INCOME_TOTAL	Target Labels		
Group based on income	0	1	Grand Total
Group1	22879	2119	24998
Group2	23094	1907	25001
Grand Total	45973	4026	49999

Median	Proportion
145800	8.476678
	7.627695
	Yes, people with lower income face more difficulties towards payment when compared to people with higher income



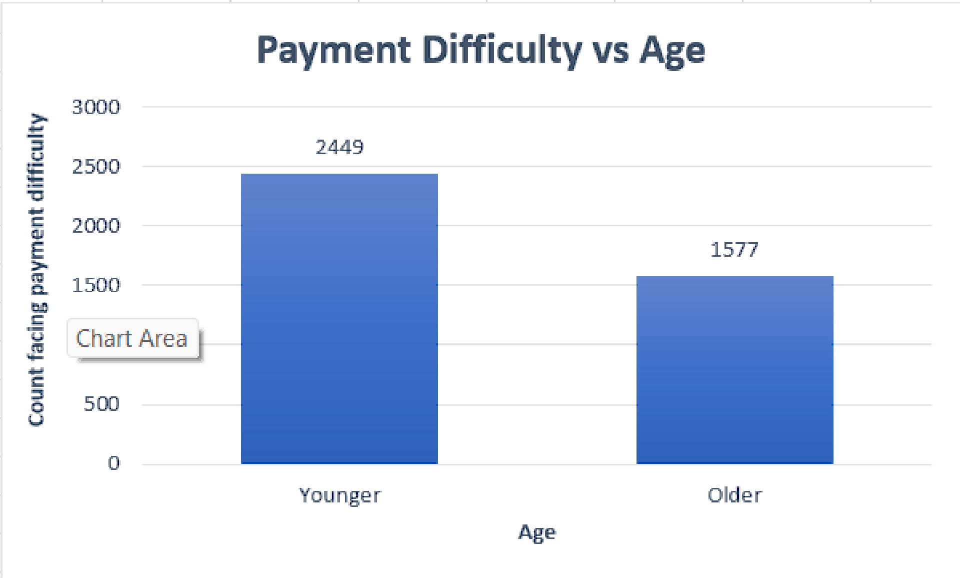
- Age & Payment Difficulty:

I created a pivot table containing the rows as the DAYS_BIRTH column, target as the columns, count of target variable as the values. I found out the median, minimum and the maximum values of the DAYS_BIRTH column.

I created a new table with DAYS_BIRTH column from the pivot table. I used the logic to sum the column where target=1 if value of DAYS_BIRTH column in the table is greater than the median value and less than the maximum value for finding the count of people who are younger and are facing payment difficulties. Similarly, I found the count of the older people who are facing payment difficulties (sum if DAYS_BIRTH is less than or equal to median value and greater than or equal to the minimum value). It is important to note that, DAYS_BIRTH column is negative.

I found out that the people who are younger face more difficulties in payment when compared to the older people.

Median	Min	Max
-15731	-25184	-7680
Group	Count facing difficulty	
Younger	2449	
Older	1577	
Answer:		
Yes, younger audience face more difficulty in payments.		



Loan Characteristics:

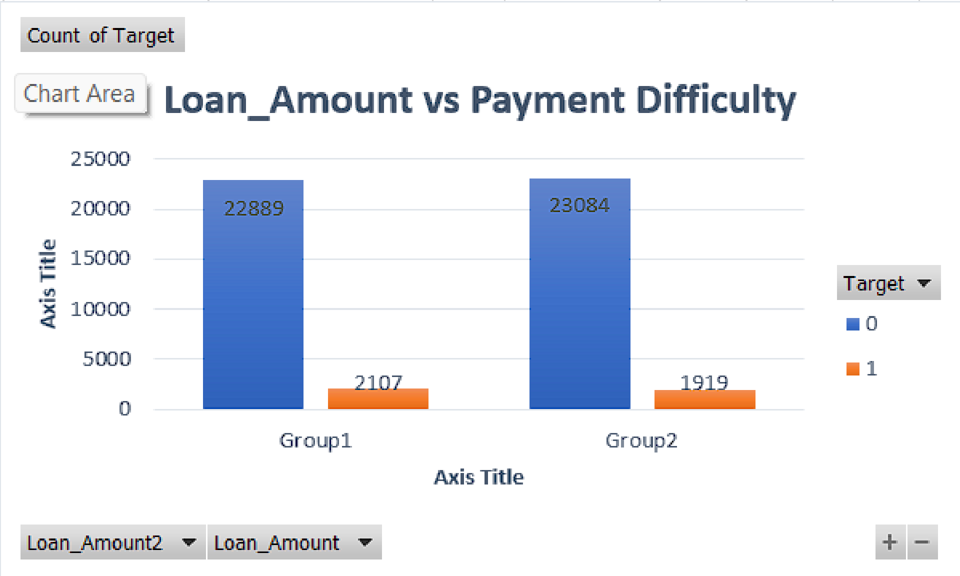
- Loan Amount & Payment Difficulty:
  First, I copied the AMT_CREDIT and Target columns to a new sheet. I checked the correlation between the two columns and I found that they were slightly negative. This shows that lower loan amount is associated with increased payment difficulty. To verify this, I first found the median of the AMT_CREDIT column. Then, I created a pivot table from the two columns with the AMT_CREDIT being the rows and Target being the columns and the count of target variable being the values. I then, grouped the rows based on the median value, if row less than median value then it belongs to Group1(Lower loan amount) otherwise Group2(Higher loan amount). I found out the proportion of the

population facing payment difficulty in both the groups. I verified that the lower loan amount group had the higher payment difficulty proportion when compared to the higher loan amount group.

Correlation	
-0.032428347	
Median	
Loan_Amount	514777.5
Total Count	49999
# Facing difficulty	4026
High Loan_Amount:	
Total Count	25003
# Facing difficulty	1919
Proportion (in %)	7.675079
Low Loan_Amount:	
Total Count	24996
# Facing difficulty	2107
Proportion (in %)	8.429349
Answer:	
No, lower loan amounts is associated with increased payment difficulty	

Group 1 represents lower loan amount
Group 2 represents higher loan amount

Count of Target	Column Labels		
Row Labels	0	1	Grand Total
Group1	22889	2107	24996
Group2	23084	1919	25003
Grand Total	45973	4026	49999



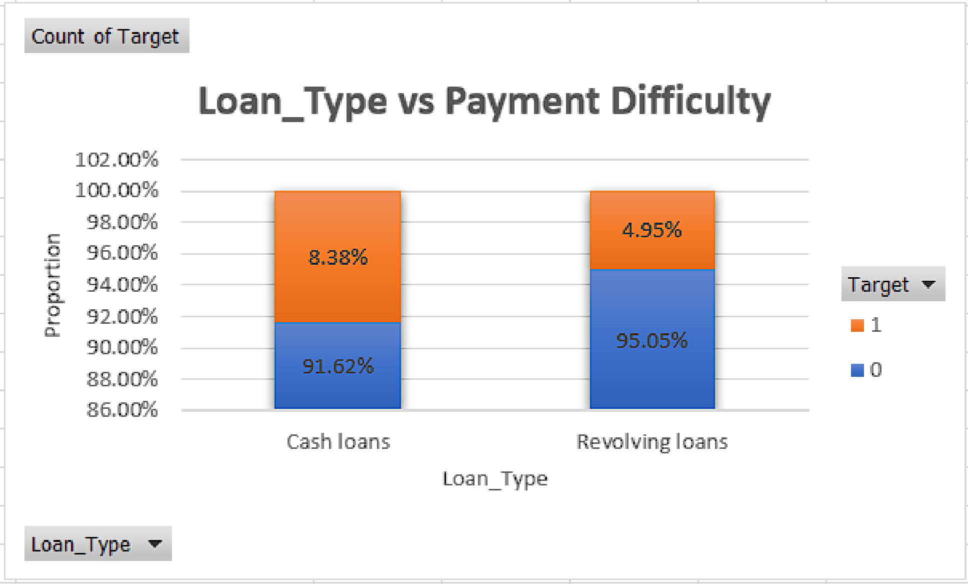
- Loan Type & Payment Difficulty:
  Similar to the approach I used for the Loan Amount & Payment Difficulty part, here also I did the same with NAME_CONTRACT_TYPE and Target columns.

Distinct Loan_Type
Cash loans
Revolving loans

Total Count	49999
# Facing difficulty	4026
Loan_Type = Cash loans	
Count	45276
# Facing difficulty	3792
Proportion	8.375298
Loan_Type = Revolving Loans	
Count	4723
# Facing difficulty	234
Proportion	4.954478

Count of Target	Column Labels		
Row Labels	0	1	Grand Total
Cash loans	91.62%	8.38%	100.00%
Revolving loans	95.05%	4.95%	100.00%
Grand Total	91.95%	8.05%	100.00%

Answer:
Yes, Cash loans are associated with more payment difficulties than Revolving loans.



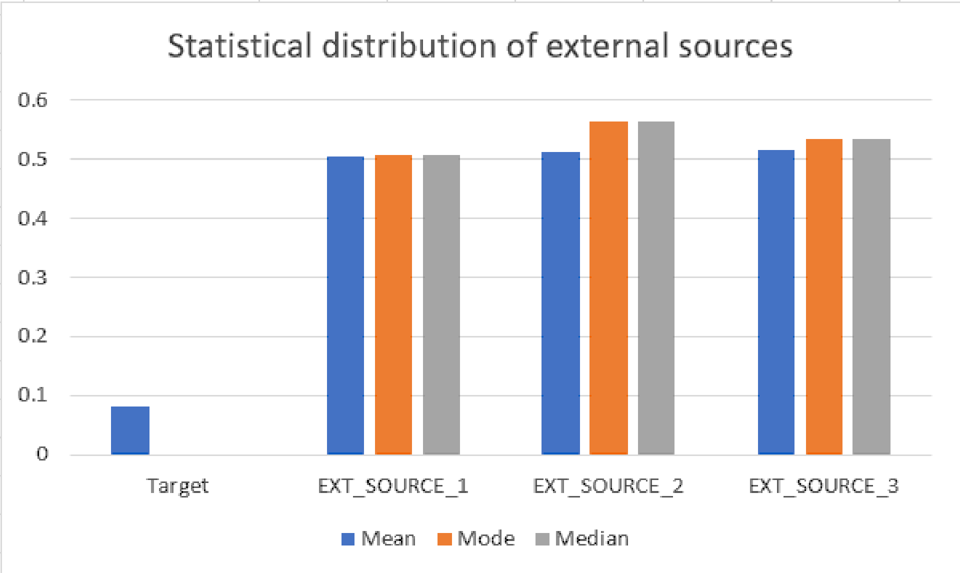
External Factors and Creditworthiness:

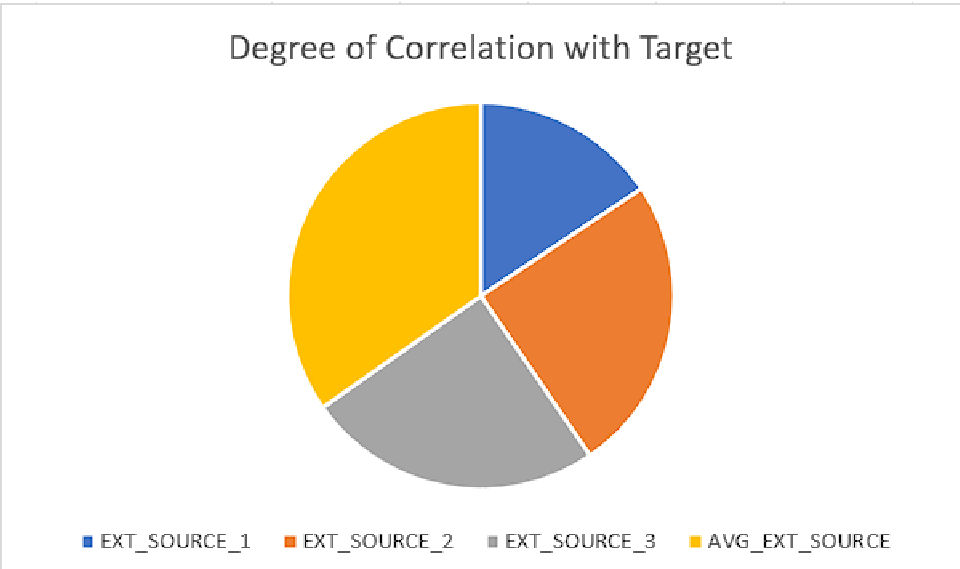
- External Ratings & Payment Difficulty:
  I copied the target and external source columns to a new sheet. I found out the mean, mode, median of these columns first. I, then found out the correlation of each of the columns with the target column. I created a new column containing the average of all the external source ratings. Surprisingly, the correlation of this column with the target variable is more than that of each of the columns. The correlation was negative which shows that lower the external scores, the higher the target is.
  Therefore, Lower scores from external sources correlates with increased payment difficulties.

Name	Mean	Mode	Median
Target	0.080522	0	0
EXT_SOURCE_1	0.504864	0.506884	0.506884
EXT_SOURCE_2	0.513954	0.565585	0.565585
EXT_SOURCE_3	0.516534	0.535276	0.535276

Correlation	Target
EXT_SOURCE_1	-0.09974
EXT_SOURCE_2	-0.15829
EXT_SOURCE_3	-0.15818
AVG_EXT_SOURCE	-0.22185

Negative correlation which implies lower external scores, the higher the target is.
Therefore, Lower scores from external sources correlates with increased payment difficulties.





Employment & Occupation:

- Employment Duration & Payment Difficulty:
  I first copied the Target column, DAYS_EMPLOYED column into a new sheet. Then, I created a table with these columns, I added a column named ADJ_DAYS_EMPLOYED. This column would contain -1*(DAYS_EMPLOYED) column, if the result is negative, then it is imputed with the abs(median(DAYS_EMPLOYED)) column. Then, I found out

the median of the DAYS_EMPLOYED column. It is important to note that DAYS_EMPLOYED column contains negative values.

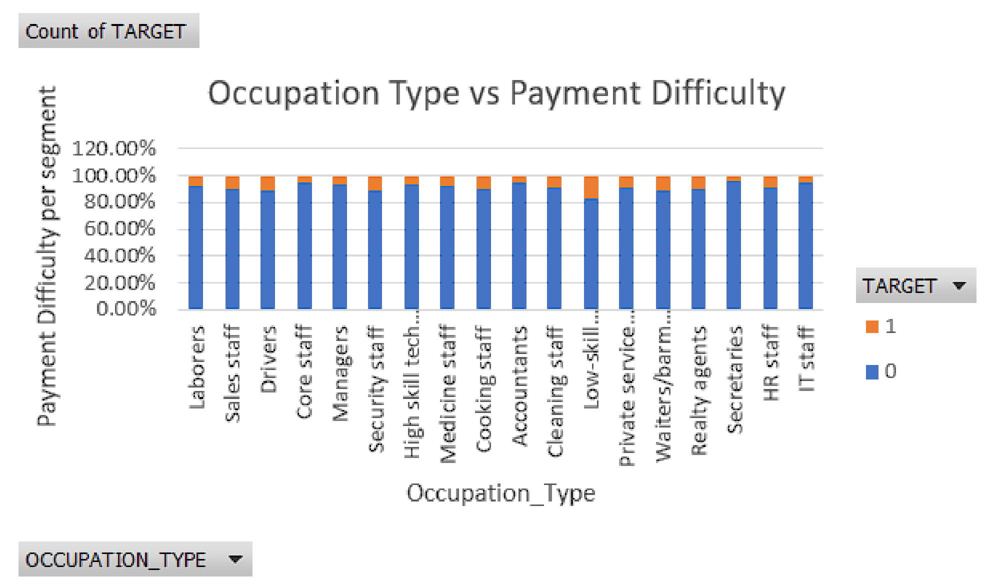
I found out the correlation between Target column and ADJ_DAYS_EMPLOYED column. It came out to be slightly negative which shows that people who had shorter employment duration has more difficulties in payment. To verify this, I found the count of people who were employed less than the median value of DAYS_EMPLOYED and the count of people who were employed more than the median value of DAYS_EMPLOYED. For these two groups, I found out the count of people who are having payment difficulties. Then, I found out the proportion of people who are facing payment difficulties in each group separately. Therefore, I verified that clients with shorter employment duration has more difficulties in payment.

Median	
-1221	
Correlation between days employed and target	
-0.061291159	
Yes, clients with shorter employment duration has more difficulties in payment.	
This statement is not completely true since, the correlation is almost neutral but very slightly negative(true upto some extent)	

Totat Count	49999
# Facing difficulty	4026
Less # of days employed:	
Total Count	Chart Area 4985
# Facing difficulty	2253
Proportion (in %)	9.017410446
More # of days employed:	
Total Count	25014
# Facing difficulty	1773
Proportion (in %)	7.088030703
Therefore, clients with shorter employment duration has more difficulties in payment.	

- Occupation Type & Payment Difficulty:
  I created a pivot table with OCCUPATION_TYPE as rows, Target as columns and count of target variable as the values. I then, sorted the values according to the count of target from largest to smallest. I, then changed the values to show them as the percentage of the row total. I used conditional formatting to highlight the top 40 % of the percentage where the target is 1. This would give us the OCCUPATION_TYPE who generally have more difficulty in payment. Therefore, it was clear that people with certain Occupation Type had more Payment Difficulty when compared to others.

Count of TARGET	**Column Labels**		
Row Labels	**0**	**1**	**Grand Total**
Laborers	92.09%	7.91%	100.00%
Sales staff	90.47%	9.53%	100.00%
Drivers	88.90%	11.10%	100.00%
Core staff	94.36%	5.64%	100.00%
Managers	93.04%	6.96%	100.00%
Security staff	89.04%	10.96%	100.00%
High skill tech staff	93.63%	6.37%	100.00%
Medicine staff	92.44%	7.56%	100.00%
Cooking staff	89.51%	10.49%	100.00%
Accountants	95.00%	5.00%	100.00%
Cleaning staff	90.80%	9.20%	100.00%
Low-skill Laborers	82.91%	17.09%	100.00%
Private service staff	91.72%	8.28%	100.00%
Waiters/barmen staff	89.04%	10.96%	100.00%
Realty agents	89.43%	10.57%	100.00%
Secretaries	95.75%	4.25%	100.00%
HR staff	91.09%	8.91%	100.00%
IT staff	95.00%	5.00%	100.00%
Grand Total	**91.95%**	**8.05%**	**100.00%**



Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
Laborers	22660	1946	24606
Sales staff	4668	492	5160
Drivers	2706	338	3044
Core staff	4184	250	4434
Managers	3246	243	3489
Security staff	1015	125	1140
High skill tech staff	1734	118	1852
Medicine staff	1297	106	1403
Cooking staff	862	101	963
Accountants	1540	81	1621
Cleaning staff	671	68	739
Low-skill Laborers	296	61	357
Private service staff	410	37	447
Waiters/barmen staff	203	25	228
Realty agents	110	13	123
Secretaries	203	9	212
HR staff	92	9	101
IT staff	76	4	80
Grand Total	**45973**	**4026**	**49999**

In the given dataset, Laborers, Sales staff, Drivers, core staff have more difficulty in payment
Taking into account of the proportions in the total population,
it seems Low-skill Laborers, Drivers, Security Staff, Waiters, Realty agents and cooking staff have more difficulty in payment

Application Details:

- Application Timing & Payment Difficulty:

I, first copied the Target column, WEEKDAY_APPR_PROCESS_START column, HOUR_APPR_PROCESS_START column into a new sheet and created a table with the column names, Target, WeekDay, Hour.

I wrote separately in the sheet, what I mean by a weekend (Saturday, Sunday while other days are working days) and late hours (from 9 pm to morning 6 am while left out timing is working hours).
I added two more columns named IS_LATE_HOUR and IS_WEEKEND. These columns act as helper column to find out if it is a late hour and if it is a weekend.

	H	I	J
1			
2		From	To
3	Weekends:	SATURDAY	SUNDAY
4	Late Hours:	21	6

I used the following formulas for filling out the IS_LATE_HOUR and IS_WEEKEND column:

IS_LATE_HOUR:
=IF(OR([@Hour]>$I$4,[@Hour]<$J$4),1,0)
IS_WEEKEND:
=IF(OR([@WeekDay]="SATURDAY",[@WeekDay]="SUNDAY"),1,0)

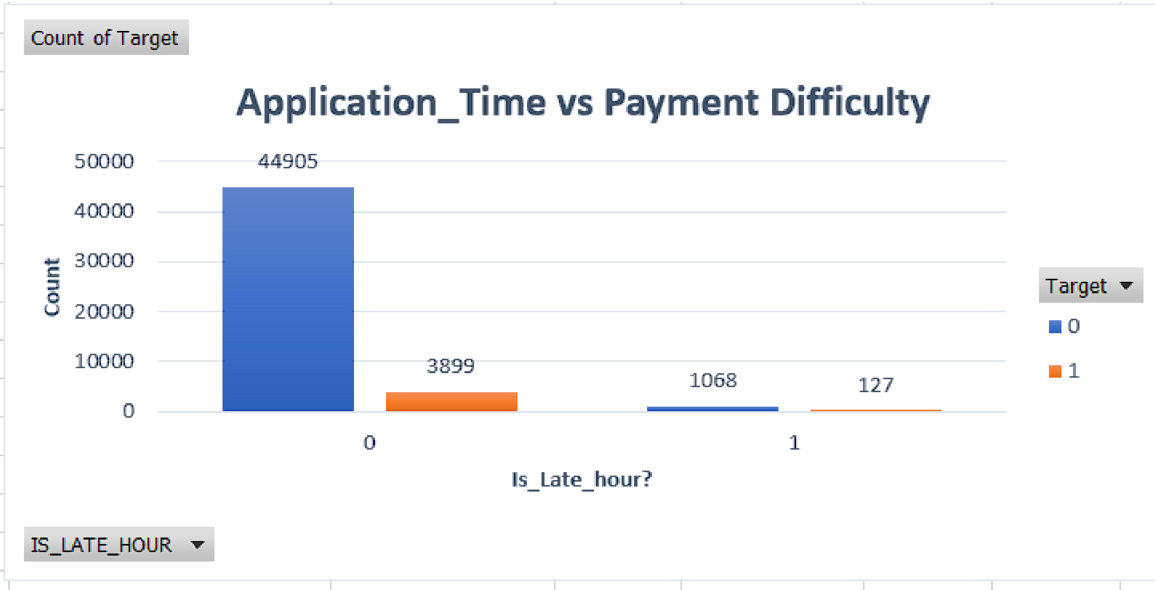
I calculated the total number of people who are facing the payment difficulty using the countif function. Then, I calculated the total number of people who are facing payment difficulty and have done the payment in late hours (IS_LATE_HOUR=1). Similarly, I calculated the number of people who are facing payment difficulty

and have done the payment in weekend. I calculated the total number of late hours and total number of weekends. Using this, I calculated the proportion of people who are facing payment difficulty in late hours and in weekends separately.
After this, I calculated the total number of weekdays, total number of working hours. Then, I calculated the number of people who were facing difficulties in payment during the workdays and the number of people who were facing difficulties in payment during the normal hours. From this, I calculated the proportion of people who were facing difficulties in normal hours and in work days.
Comparing the above proportions with the proportions which were calculated earlier, I came to a conclusion that payment difficulty during late hours(after 9pm and before 6am) is more when compared to normal hours and payment difficulty during weekends is less when compared to work days.

Total Difficulty:	4026		
Total Late Hours:	1195		
Total Weekends:	8083		Proportion(%)
Difficulty in Late Hours:	127		10.62761506
Difficulty in Weekends:	634		7.843622417
Difficulty in Weekdays:	3392		8.092375227
Difficulty in Normal hours:	3899		7.989099254
Total weekdays:	41916		
Total normal hours:	48804		

Payment difficulty during late hours(after 9pm and before 6am) is more when compared to non late hours
Payment difficulty during weekends is less when compared to non weekend days.

Count of Target	**Column Labels**		
Row Labels	**0**	**1**	**Grand Total**
0	44905	3899	48804
1	1068	127	1195
Grand Total	**45973**	**4026**	**49999**



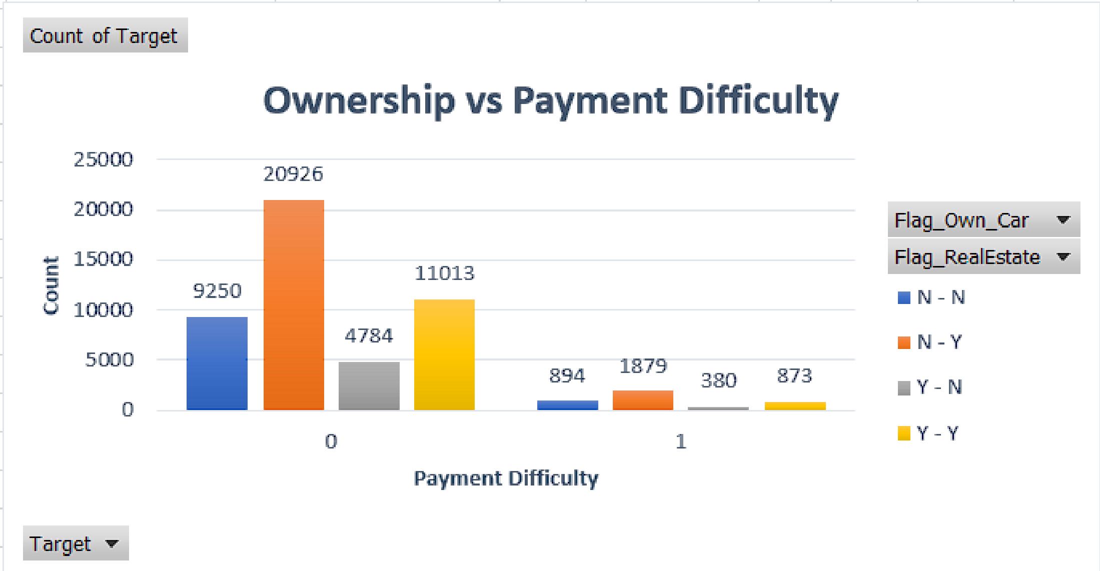
Property and Reality:

- Property Ownership & Payment Difficulty:
  First, I copied the columns Target, FLAG_OWN_CAR and FLAG_REALTY to new sheet. I added a new column named OWN_EITHER. This column would contain TRUE if any of the other two columns contains Yes otherwise FALSE. I used OR function for this. Using this column, I calculated the number of people who owns either of the property and then I calculated the number of people who owns neither of the property. I, then calculated the number of people who has payment difficulties and owns some property. I also calculated the number of people who has payment difficulties and doesn’t own any property. I calculated the proportion for the both and I found out that, people who dont own any property has more payment issues when compared to those who own some property.

People who own			
Either of property	39855		
Neither of property	10144		
Has Payment issues			Proportion
Own Either property	3132		7.858487015
Own Neither property	894		8.813091483

People who don't own any property has more payment issues when compared to those who own some property

Count of Target	Column Labels						
	N		N Total	Y		Y Total	Grand Total
Row Labels	N	Y		N	Y		
0	9250	20926	30176	4784	11013	15797	45973
1	894	1879	2773	380	873	1253	4026
Grand Total	**10144**	**22805**	**32949**	**5164**	**11886**	**17050**	**49999**

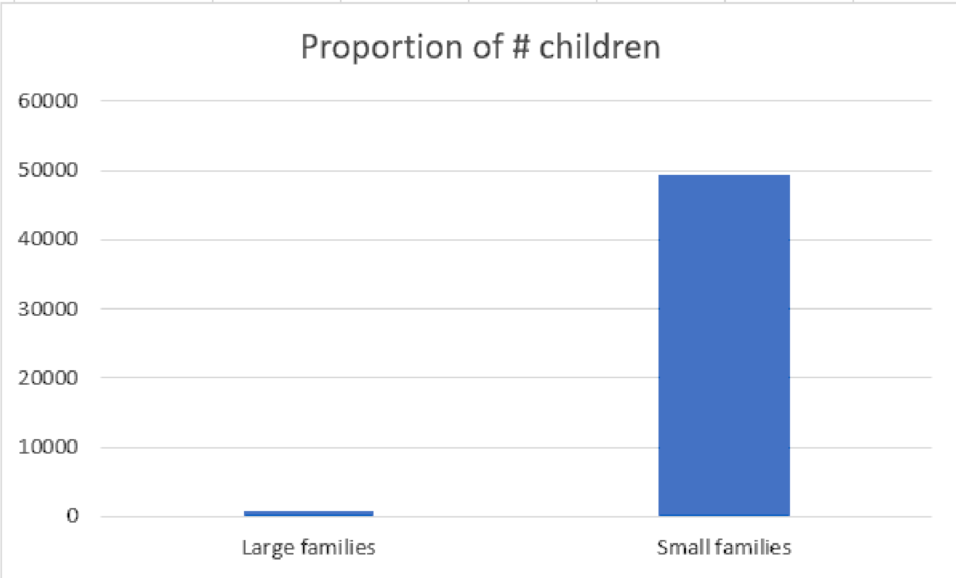


- Family Size & Payment Difficulty:
  Here, I considered that the people who has more than 2 children as a large family.
  I, first copied the columns Target, CNT_CHILDREN to a new sheet. I added a new column named IS_LARGE_FAMILY, this would contain 1 if it is a large family (CNT_CHILDREN>2).

  I counted the total number of large families and small families using the helper column. Then, I calculated the number of people who were facing difficulties and is a large family. Similarly, I calculated for the small family as well. After this, I calculated the proportion of people who were facing difficulties for each group.

From this, I came to a conclusion, that large families(more than 2 children) get payment difficulty more often when compared to small families.

Let's consider people who has children more than 2 as large family			
Total			
Large families	723		
Small families	49276		
Having difficulty			Proportion
Large families	75		10.37344
Small families	3951		8.018102
Yes, large families(more than 2 children) get payment difficulty more often when compared to small families			



- Client's Region & Payment Difficulty:
  I, first copied the columns, Target, REGION_RATING, REGION_RATING_W_CITY into a new sheet and created a table. I created a pivot table with REGION_RATING as rows and Target as columns and count of target as values. I created another pivot table with REGION_RATING_W_CITY as rows and Target as columns and count of target as values.

From the pivot tables, I found out the proportion of the population facing difficulty for each rating in the REGION_RATING and then REGION_RATING_W_CITY.

From this, I found out that,
as the REGION_RATING increases, the Target increases (proportion of people facing payment difficulties increases) and as the REGION_RATING_W_CITY increases, the Target increases (proportion of people facing payment difficulties increases).

I verified this relation using correlation function as well.

Count of Target	Column Labels		
Row Labels	**0**	**1**	**Grand Total**
1	5006	220	5226
2	34043	2921	36964
3	6924	885	7809
Grand Total	**45973**	**4026**	**49999**

Region Rating	Proportion(in %)
1	4.209720628
2	7.902283303
3	11.33307722

Count of Target	Column Labels		
Row Labels	**0**	**1**	**Grand Total**
1	5320	241	5561
2	34379	2962	37341
3	6274	823	7097
Grand Total	**45973**	**4026**	**49999**

Region_Rating_w_City	Proportion(in %)
1	4.333752922
2	7.932299617
3	11.5964492

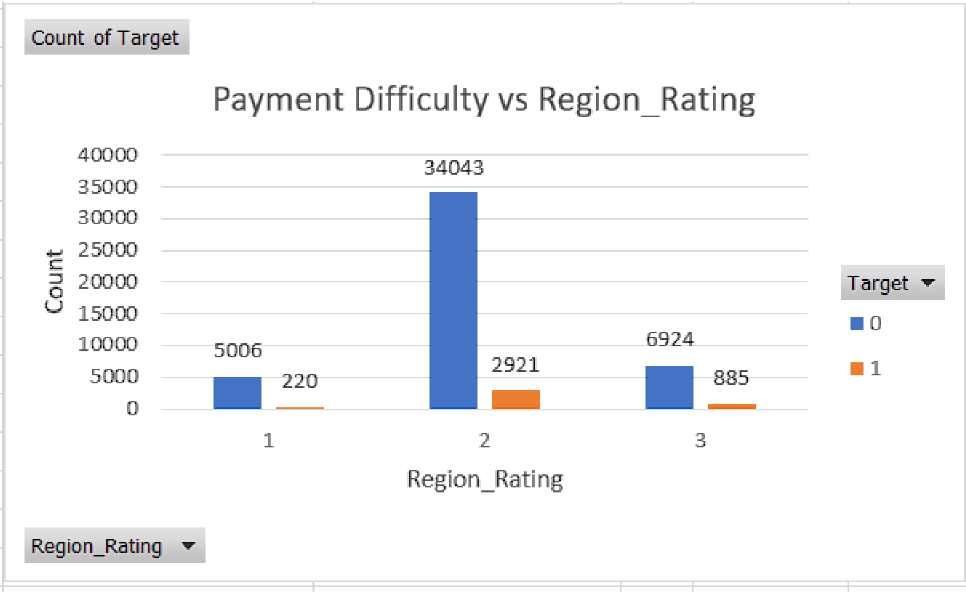
Region Rating	Proportion(in %)	Region_Rating_w_City	Proportion(in %)
1	4.209720628	1	4.333752922
2	7.902283303	2	7.932299617
3	11.33307722	3	11.5964492

As Region_rating increases target increases(proportion of people facing payment difficulties increases)

As Region_rating_w_city increases target increases(proportion of people facing payment difficulties increases)

Correlation with target	
0.066130148	-> Region_Rating
0.067079294	-> Region_Rating_w_City

Chart Area



E. **Identify Top Correlations for Different Scenarios:** Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

- **Task:** Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

Approach:

First, I copied all the numerical variables, Target separately into a new sheet. Then, I filtered the target column to have only 1. I copied the resultant rows to a new sheet and formed a new table.

Then, I changed the filter so that the target only has the value 0. I copied the resultant rows to a new sheet and formed a new table. This is how I segmented the data based on target variable. Correlation of each variable with the target variable in each segment gives division by zero error since, the target variable is constant.

Then, I installed the data analysis toolpak from the add-ins menu present inside the options menu. For each of the segmented data, I selected the segment and clicked on the Data Analysis icon present in Data menu. I then clicked on Correlation matrix. I deleted the main diagonal, since it has 1 (we already know that correlation of a column to itself is 1, and I don't want to include those cells while ranking the correlation).

I selected the correlation matrix and I used 3-color scale conditional formatting on it for colouring negative, positive and neutral correlation differently. I then, copied this correlation matrix and then filled out the abs of the values (so that I can rank positive correlation and negative correlation based on intensity and not based on sign). I then, created a format similar to the correlation matrix but for the values I used the following formula:
=RANK.EQ(C23,$C$22:$P$35,0)
Here, 0 denotes the descending order ranking, that is I want the strongest correlation to have the first rank. I then filled out the table. I used conditional formatting on the ranks table to highlight the bottom 10 % and I summarised the results for each of the segments separately.

	Target	Cnt_Children	Amt_Income_Total	Amt_Credit	Annuity	Goods_Price	POPULATION	AYS_BIRTH	S_EMPLOY	_RATING_	TING_CLIE	PR_PROCE	T_SOURCE	T_SOURCE	T_SOURCE_3
Target															
Cnt_Children	#DIV/0!														
Amt_Income_	#DIV/0!	0.010110177													
Amt_Credit	#DIV/0!	0.007601905	0.01527												
Annuity	#DIV/0!	0.029172977	0.01800	0.74967											
Goods_Price	#DIV/0!	-0.001167452	0.01326	0.98213	0.74933										
REGION_POPU	#DIV/0!	-0.020359154	-0.00618	0.06778	0.07312	0.07650									
DAYS_BIRTH	#DIV/0!	0.2496732	0.00903	-0.14251	-0.00875	-0.14097	-0.01647								
DAYS_EMPLOY	#DIV/0!	-0.189324184	-0.01156	0.01604	-0.07956	0.02018	0.00774	-0.58148							
REGION_RATI	#DIV/0!	0.055515557	-0.01285	-0.04502	-0.06158	-0.05120	-0.43003	0.04503	-0.00915						
REGION_RATI	#DIV/0!	0.054802235	-0.01267	-0.05295	-0.07942	-0.05659	-0.43168	0.03809	-0.00414	0.95077					
HOUR_APPR_	#DIV/0!	-0.006884357	0.01448	0.04540	0.04489	0.05752	0.15605	0.05789	-0.05207	-0.27888	-0.25307				
EXT_SOURCE_	#DIV/0!	-0.118205866	0.00128	0.11220	0.03190	0.11413	0.06113	-0.39635	0.18218	-0.03681	-0.04141	0.01450			
EXT_SOURCE_	#DIV/0!	-0.015410537	-0.01623	0.11919	0.11370	0.13334	0.15922	-0.11141	-0.01880	-0.23905	-0.23918	0.13390	0.06602		
EXT_SOURCE_	#DIV/0!	-0.014958105	-0.02646	0.04584	0.01773	0.04757	-0.02159	-0.13959	0.08319	0.02150	0.01782	-0.02579	0.08087	0.04734	

	Cnt_Children	mt_Income_Tot	Amt_Credit	Annuity	Goods_Price	POPULATION	AYS_BIRTH	S_EMPLOY	_RATING_	TING_CLIE	PR_PROCE	T_SOURCE	T_SOURCE	T_SOURCE_3
Cnt_Children														
Amt_Income_	81													
Amt_Credit	86	73												
Annuity	58	66	3											
Goods_Price	91	77	1	4										
REGION_POPU	63	88	35	34	33									
DAYS_BIRTH	11	83	18	84	19	69								
DAYS_EMPLOY	14	80	71	31	64	85	5							
REGION_RATI	42	78	52	37	46	7	51	82						
REGION_RATI	43	79	44	32	41	6	55	89	2					
HOUR_APPR_	87	76	50	53	40	17	39	45	9	10				
EXT_SOURCE_	24	90	27	57	25	38	8	15	56	54	75			
EXT_SOURCE_	72	70	23	26	22	16	28	65	13	12	21	36		
EXT_SOURCE_	74	59	49	68	47	61	20	29	62	67	60	30	48	

The ranks are done in descending order and excluding the correlation of a variable with itself

Rank	G1	G2	Correlation
1	Goods_Price	Amt_Credit	0.98213
2	REGION_RATI	GION_RATING_CLIEN	0.95077
3	Annuity	Amt_Credit	0.74967
4	Goods_Price	Annuity	0.74933
5	DAYS_EMPLOY	DAYS_BIRTH	-0.58148

	Target	Cnt_Children	Amt_Income_Total	Amt_Credit	Annuity	Goods_Price	POPULATION	AYS_BIRTH	S_EMPLOY	_RATING_	TING_CLIE	PR_PROCE	T_SOURCE	T_SOURCE	T_SOURCE_3
Target	1														
Cnt_Childr	#DIV/0!	1													
Amt_Incor	#DIV/0!	0.036319722	1												
Amt_Credi	#DIV/0!	0.005705458	0.377965752	1											
Annuity	#DIV/0!	0.026384785	0.451134889	0.770773157	1										
Goods_Pri	#DIV/0!	0.001547629	0.384548363	0.986904954	0.775728	1									
REGION_P	#DIV/0!	-0.024912809	0.181941261	0.095539444	0.117279	0.098939	1								
DAYS_BIRT	#DIV/0!	0.335876269	0.073769425	-0.051084182	0.00991	-0.04868	-0.03044	1							
DAYS_EMF	#DIV/0!	-0.243591518	-0.162702675	-0.077367219	-0.113	-0.07514	-0.00661	-0.61529	1						
REGION_R	#DIV/0!	0.021288992	-0.205031899	-0.102556478	-0.12992	-0.1049	-0.53933	0.009025	0.040506	1					
REGION_R	#DIV/0!	0.017873365	-0.220044862	-0.111639948	-0.1432	-0.11319	-0.53686	0.007084	0.042899	0.950468	1				
HOUR_API	#DIV/0!	-0.005272551	0.08543156	0.056524809	0.053565	0.065358	0.167612	0.096389	-0.09236	-0.28282	-0.26176	1			
EXT_SOUR	#DIV/0!	-0.100014976	0.060504292	0.113361371	0.081712	0.116968	0.065702	-0.35455	0.136695	-0.07785	-0.07677	0.018594	1		
EXT_SOUR	#DIV/0!	-0.013466448	0.156172404	0.136259888	0.130023	0.143251	0.201089	-0.08033	-0.03409	-0.29442	-0.28906	0.155796	0.126386	1	
EXT_SOUR	#DIV/0!	-0.039263601	-0.073475401	0.028989838	0.018671	0.031215	-0.01388	-0.17908	0.098573	0.002023	0.003246	-0.04745	0.090395	0.068882	1

Ranks:

	Cnt_Children	Amt_Income_Total	Amt_Credit	Annuity	Goods_Pric	PULATION	AYS_BIRTH	'S_EMPLOY	_RATING	TING_CLIE	PR_PROCE	T_SOURCE	T_SOURCE	T_SOURCE_3
Cnt_Children														
Amt_Incor	70													
Amt_Credi	87	10												
Annuity	75	8	4											
Goods_Pri	91	9	1	3										
REGION_P	76	21	46	34	43									
DAYS_BIR	12	56	64	83	65	73								
DAYS_EMF	17	24	53	38	55	86	5							
REGION_R	77	19	41	32	40	6	84	68						
REGION_R	80	18	39	28	37	7	85	67	2					
HOUR_API	88	49	62	63	60	23	45	47	15	16				
EXT_SOUR	42	61	36	50	35	59	11	29	52	54	79			
EXT_SOUR	82	25	30	31	27	20	51	71	13	14	26	33		
EXT_SOUR	69	57	74	78	72	81	22	44	90	89	66	48	58	

Rank	G1	G2	Correlation
1	Goods_Price	Amt_Credit	0.986904954
2	REGION_RATING	GION_RATING_CLIE	0.950468157
3	Goods_Price	Annuity	0.775728255
4	Annuity	Amt_Credit	0.770773157
5	DAYS_EMPLOYE	DAYS_BIRTH	-0.615289978

The ranks are done in descending order and excluding the correlation of a variable with itself

Insights:

The banks should be more careful in providing the loans to the people who has payment difficulties more often. The following are the factors that affect the payment difficulties generally:

- Most people have the OCCUPATION_TYPE as Laborers.
- Most of the people are Unaccompanied.
- EXT_SOURCE_1 has the most outliers and it should be removed while EXT_SOURCE_2 has the least outliers.
- There is a huge data imbalance in the target class, 0 value constitutes over 91.75% while value 1 constitutes just 8.05%.
- Men face more difficulties in payment compared to women.
- People with lower income face more difficulties towards payment when compared to people with higher income.
- People who are younger face more difficulty in payments when compared to people who are older.
- Lower loan amounts is associated with increased payment difficulty.
- Cash loans are associated with more payment difficulties than Revolving loans.
- Lower scores from external sources correlates with increased payment difficulties.

- Clients with shorter employment duration has more difficulties in payment.
- Low-skill Laborers, Drivers, Security Staff, Waiters, Realty agents and cooking staff have more difficulty in payment when compared to other occupations.
- Payment difficulty during late hours (after 9pm and before 6am) is more when compared to normal hours.
- Payment difficulty during weekends is less when compared to week days.
- People who don't own any property has more payment issues when compared to those who own some property.
- Large families(more than 2 children) get payment difficulty more often when compared to small families.
- As the REGION_RATING increases, the proportion of people facing payment difficulties increases.
- As the REGION_RATING_W_CITY increases, the proportion of people facing payment difficulties increases.
- AMT_GOODS_PRICE and AMT_CREDIT has the strongest correlation in both the segments when the people were facing payment difficulties and when the people weren't facing payment difficulties.

Result:
From this project, I have learnt how banks use EDA (Exploratory Data Analysis) to prevent business loss (loss in customers) and financial loss. I felt like, this project really simulated the real-world scenario and I understood the complete practical applications of Excel.
I understood about various loan attributes, customer attributes and their influence in the likelihood of default. I understood the patterns that indicate if a customer will have difficulty paying their instalments.

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 companies can make better decisions about loan approval from this project.