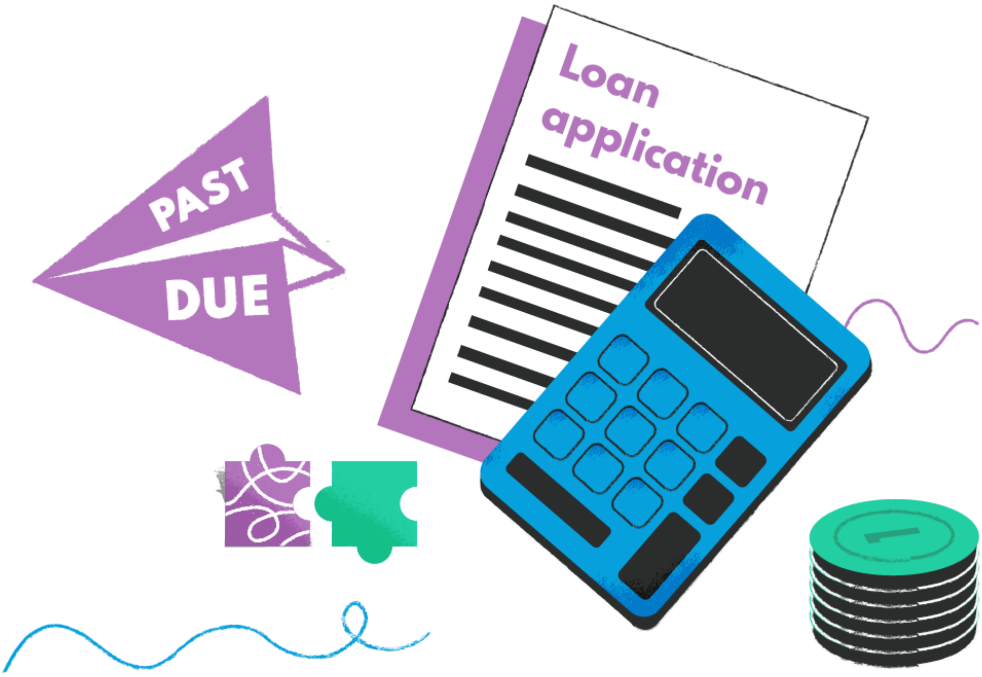
Bank Loan Case Study



- **Samruddhi Pawar**

Project Description:

Imagine you're a data analyst at a finance company that specializes in lending various types of loans to urban customers. The company faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans. As a data analyst, you are tasked to use Data Analytics tools and skills to analyse patterns in the data and ensure that capable applicants are not rejected.

The main aim of this project is 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 the loan, or lending at a higher interest rate to risky applicants. The company wants to understand the key factors behind loan default to make better decisions about loan approval.

Business Understanding:

The loan-providing companies find it hard to give loans to people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming defaulter.

How are we going to handle the things?

Suppose we work for a consumer finance company that specializes in lending various types of loans to urban customers. We will use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
  - If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

- When a client applies for a loan, there are four types of decisions that could be taken by the client/company:
  - Approved: The company has approved loan application
  - Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
  - Refused: The company had rejected the loan (because the client does not meet their requirements etc.).
  - Unused Offer: Loan has been cancelled by the client but on different stages of the process.

In this case study, we will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

- Following are the things that we are going to find out through this case study:
  - Our aim is to identify the patterns that indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of the loan, lending (too risky applicants) at a higher interest rate, etc.
  - The driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.
  - Presenting the overall approach of the data analysis, cleaning the dataset, finding outliers, data imbalance, univariate, segmented univariate, bivariate analysis, etc.
  - The top 10 correlations for the Client with payment difficulties and all other cases (Target variable).

Tech-Stack Used:

Microsoft Excel: All the analysis has been performed in Excel. This tool is also used to create a graphical representation of the results and to understand the result set better.

Microsoft Word: It is used to make a report (PDF) to be presented to the leadership team.

Data Understanding:

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

Source of Data:
https://docs.google.com/spreadsheets/d/1VoLQFMauM1TngYFY7dKx8xLrzj6HzUSzhp-1fbbhCMo/edit?usp=sharing

https://docs.google.com/spreadsheets/d/1ma-6d17Llnrlu9RrG46F_eJ6h68WIHkPsqnc_cSJxek/edit?usp=sharing

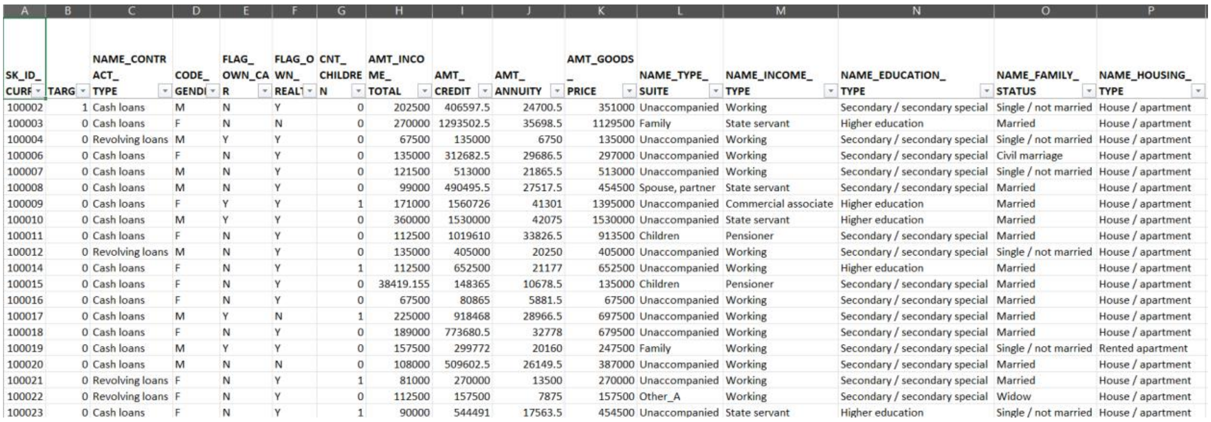
Insights:

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.

Task: Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

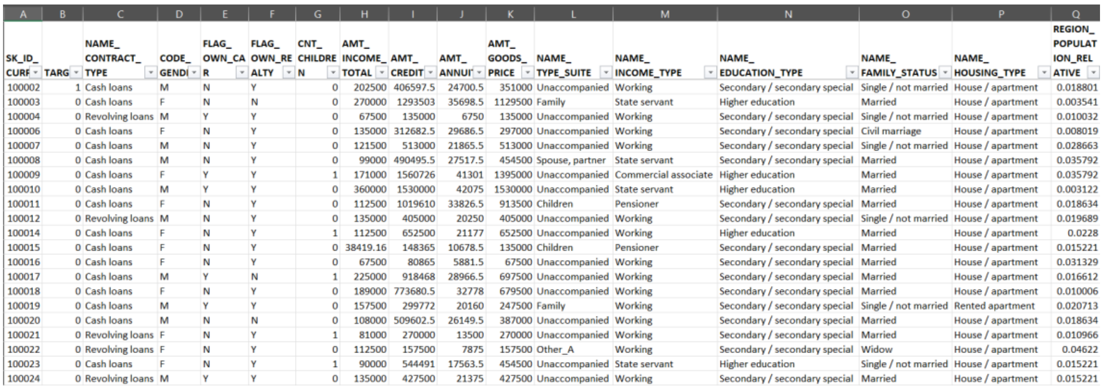
Result:

- I have used the COUNTA function to count the total rows in each column.
- After that, I have found the percentage of null values in each column using the formula 1- (Total Row Counts for each column / Total Row Counts).
- After that, I have removed all the columns having null value percentages of more than 30%. For columns having less than 30% null value percentages I have done mean imputation for the missing values for columns having null value percentages less than 30%.
- I have also found the outliers using the interquartile range method considering relevant columns. After going through each column description, I have kept only relevant columns to bring out the insights.
- The columns having days are converted into years by simply dividing the days by 365.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
SK_ID_ CURF	TARG	NAME_CONTR ACT_ TYPE	CODE_ GENDI	FLAG_ OWN_CA R	FLAG_O WN_ REALT	CNT_ CHILDRE N	AMT_INCO ME_ TOTAL	AMT_ CREDIT	AMT_ ANNUITY	AMT_GOODS _ PRICE	NAME_TYPE_ SUITE	NAME_INCOME_ TYPE	NAME_EDUCATION_ TYPE	NAME_FAMILY_ STATUS	NAME_HOUSING_ TYPE
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment
100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family	State servant	Higher education	Married	House / apartment
100004	0	Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompanied	Working	Secondary / secondary special	Civil marriage	House / apartment
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, partner	State servant	Secondary / secondary special	Married	House / apartment
100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompanied	Commercial associate	Higher education	Married	House / apartment
100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompanied	State servant	Higher education	Married	House / apartment
100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pensioner	Secondary / secondary special	Married	House / apartment
100012	0	Revolving loans	M	N	Y	0	135000	405000	20250	405000	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment
100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompanied	Working	Higher education	Married	House / apartment
100015	0	Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Children	Pensioner	Secondary / secondary special	Married	House / apartment
100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment
100017	0	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment
100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment
100019	0	Cash loans	M	Y	Y	0	157500	299772	20160	247500	Family	Working	Secondary / secondary special	Single / not married	Rented apartment
100020	0	Cash loans	M	N	N	0	108000	509602.5	26149.5	387000	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment
100021	0	Revolving loans	F	N	Y	1	81000	270000	13500	270000	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment
100022	0	Revolving loans	F	N	Y	0	112500	157500	7875	157500	Other_A	Working	Secondary / secondary special	Widow	House / apartment
100023	0	Cash loans	F	N	Y	1	90000	544491	17563.5	454500	Unaccompanied	State servant	Higher education	Single / not married	House / apartment

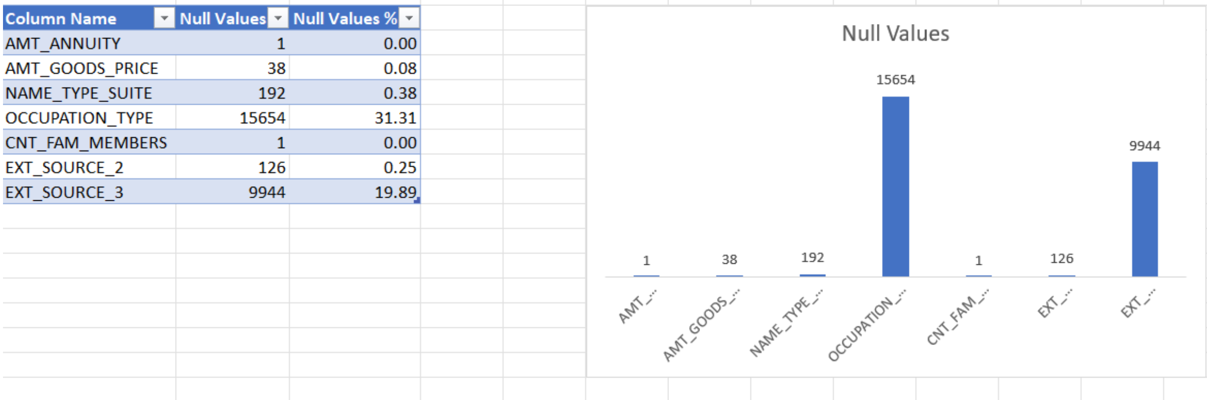
Before Cleaning



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
SK_ID_ CURF	TARG	NAME_ CONTRACT_ TYPE	CODE_ GENDI	FLAG_ OWN_CA R	FLAG_ OWN_RE ALTY	CNT_ CHILDRE N	AMT_ INCOME_ TOTAL	AMT_ CREDIT	AMT_ ANNUI	AMT_ GOODS_ PRICE	NAME_ TYPE_SUITE	NAME_ INCOME_TYPE	NAME_ EDUCATION_TYPE	NAME_ FAMILY_STATUS	NAME_ HOUSING_TYPE	REGION_ POPULAT ION_REL ATIVE
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment	0.018801
100003	0	Cash loans	F	N	N	0	270000	1293503	35698.5	1129500	Family	State servant	Higher education	Married	House / apartment	0.003541
100004	0	Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment	0.010032
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompanied	Working	Secondary / secondary special	Civil marriage	House / apartment	0.008019
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment	0.028663
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, partner	State servant	Secondary / secondary special	Married	House / apartment	0.035792
100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompanied	Commercial associate	Higher education	Married	House / apartment	0.035792
100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompanied	State servant	Higher education	Married	House / apartment	0.003122
100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pensioner	Secondary / secondary special	Married	House / apartment	0.018634
100012	0	Revolving loans	M	N	Y	0	135000	405000	20250	405000	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment	0.019689
100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompanied	Working	Higher education	Married	House / apartment	0.0228
100015	0	Cash loans	F	N	Y	0	38419.16	148365	10678.5	135000	Children	Pensioner	Secondary / secondary special	Married	House / apartment	0.015221
100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	0.031329
100017	0	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	0.016612
100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	0.010006
100019	0	Cash loans	M	Y	Y	0	157500	299772	20160	247500	Family	Working	Secondary / secondary special	Single / not married	Rented apartment	0.020713
100020	0	Cash loans	M	N	N	0	108000	509602.5	26149.5	387000	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	0.018634
100021	0	Revolving loans	F	N	Y	1	81000	270000	13500	270000	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	0.010966
100022	0	Revolving loans	F	N	Y	0	112500	157500	7875	157500	Other_A	Working	Secondary / secondary special	Widow	House / apartment	0.04622
100023	0	Cash loans	F	N	Y	1	90000	544491	17563.5	454500	Unaccompanied	State servant	Higher education	Single / not married	House / apartment	0.015221
100024	0	Revolving loans	M	Y	Y	0	135000	427500	21375	427500	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	0.015221

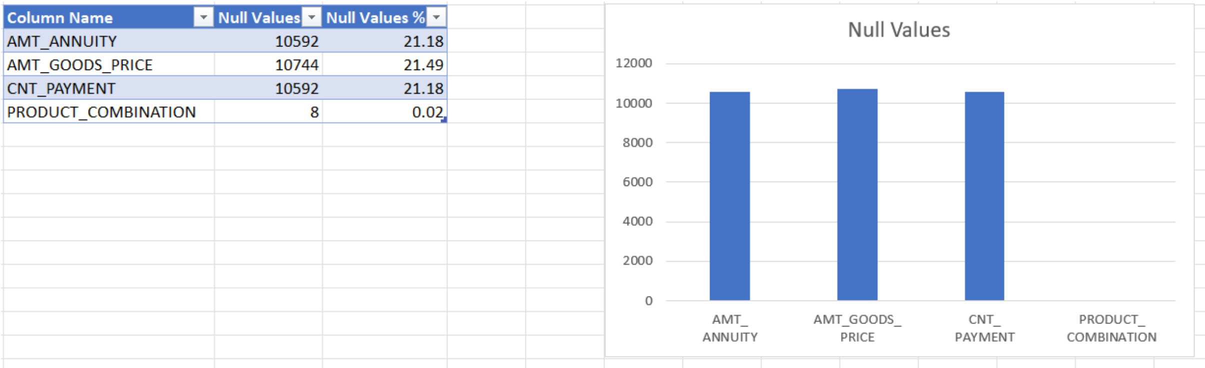
After Cleaning

- ***Application_data:***



Column Name	Null Values	Null Values %
AMT_ANNUITY	1	0.00
AMT_GOODS_PRICE	38	0.08
NAME_TYPE_SUITE	192	0.38
OCCUPATION_TYPE	15654	31.31
CNT_FAM_MEMBERS	1	0.00
EXT_SOURCE_2	126	0.25
EXT_SOURCE_3	9944	19.89

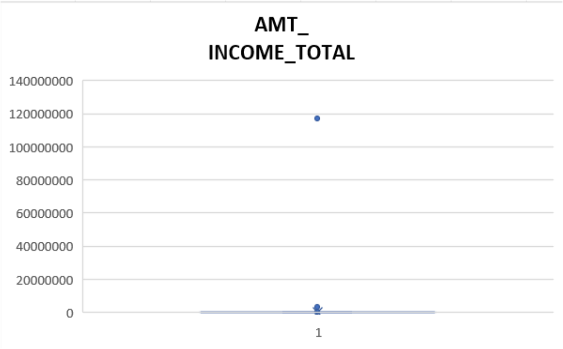
- ***Previous_application_data:***

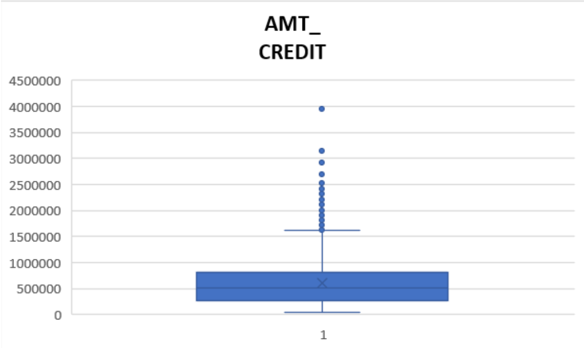


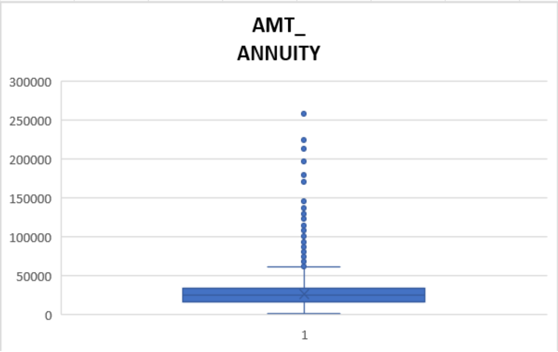
Column Name	Null Values	Null Values %
AMT_ANNUITY	10592	21.18
AMT_GOODS_PRICE	10744	21.49
CNT_PAYMENT	10592	21.18
PRODUCT_COMBINATION	8	0.02

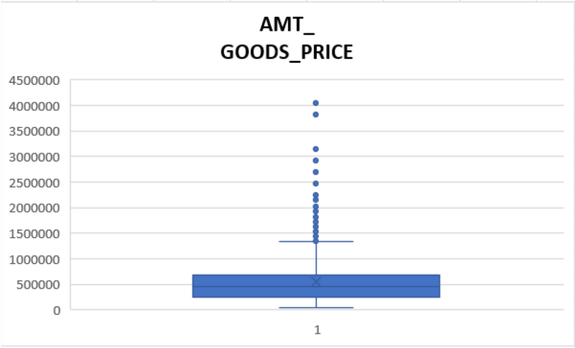
B. **Identify Outliers in the Dataset:** Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.
Task: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.
Result:
An outlier is an observation that lies an abnormal distance from other values in a random
sample from a population. An outlier can be identified from a box plot graph. If the value lies above maximum and below minimum, they are considered outliers.

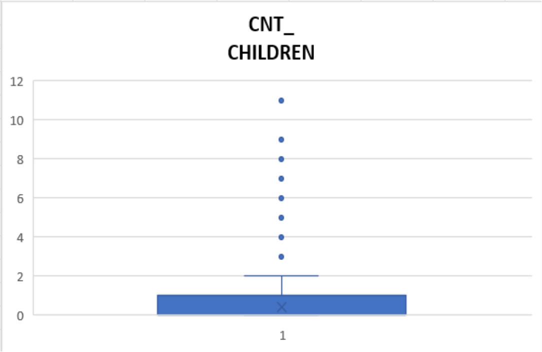
- ***Application_Data:***
 - AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE, CNT_CHILDREN and REGION_POPULATION_RELATIVE have some number of outliers.
 - AMT_INCOME_TOTAL has a huge number of outliers which indicates that few of the loan applicants have high income compared to the others.
 - AGE has no outliers which means the data available is reliable.
 - YEARS_EMPLOYED has outlier value which is around 958 years which is impossible and hence this has to be an incorrect entry.

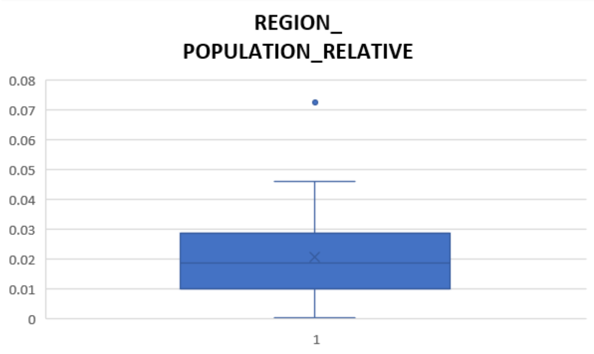

AMT_
INCOME_TOTAL
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
1

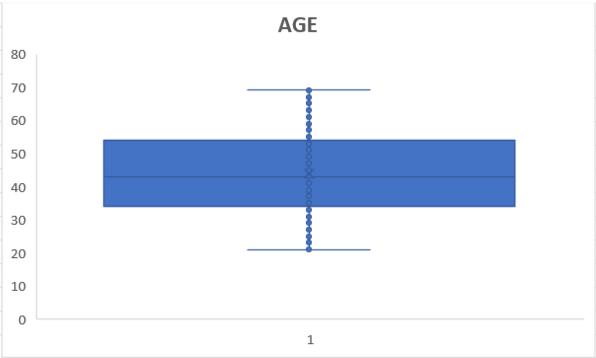

AMT_
CREDIT
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1

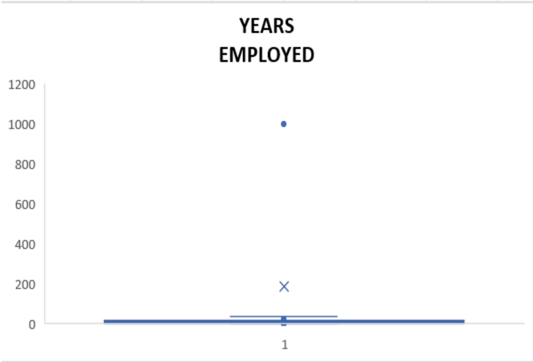

AMT_
ANNUITY
300000
250000
200000
150000
100000
50000
0
1


AMT_
GOODS_PRICE
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1

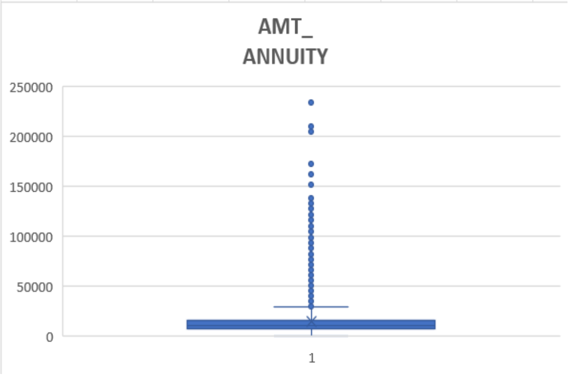

CNT_
CHILDREN
12
10
8
6
4
2
0
1

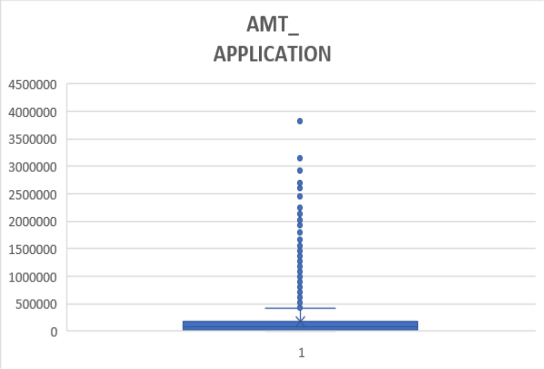

REGION_
POPULATION_RELATIVE
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
1

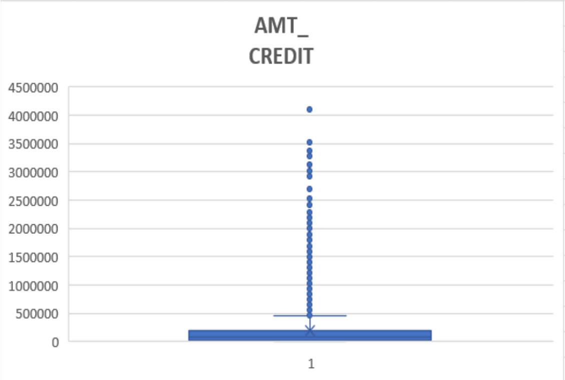

AGE
80
70
60
50
40
30
20
10
0
1

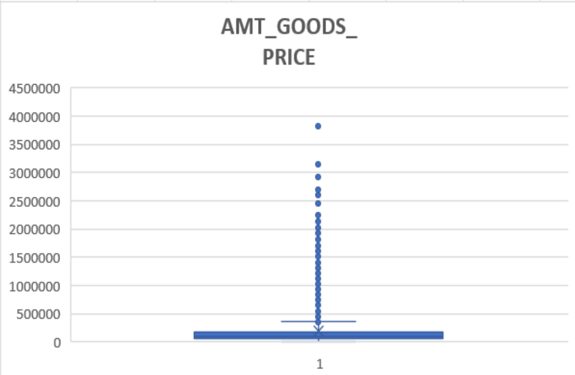

YEARS
EMPLOYED
1200
1000
800
600
400
200
0
1

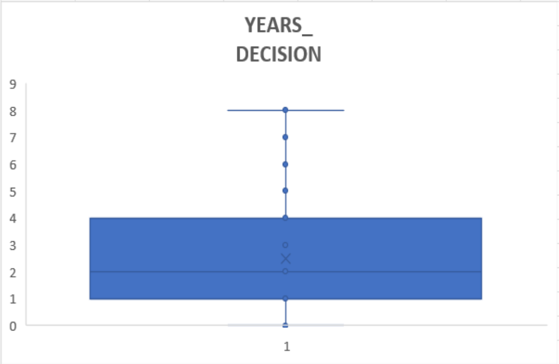
- ***Previous_Application_data:***
  - AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.
  - CNT PAYMENT has few outlier values.
  - YEARS_DECISION has a small number of outliers indicating that these previous application decisions were taken long back.

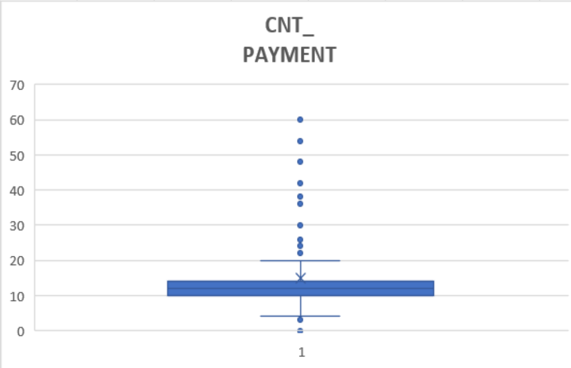












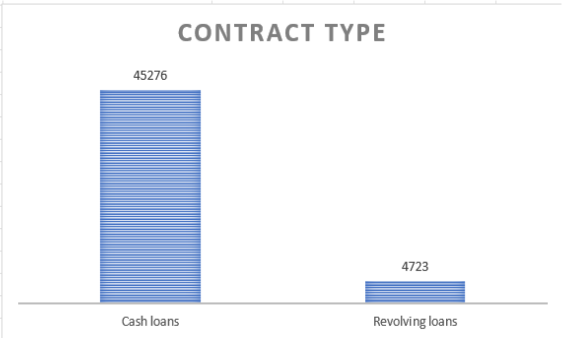
C. **Analyse Data Imbalance:** Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

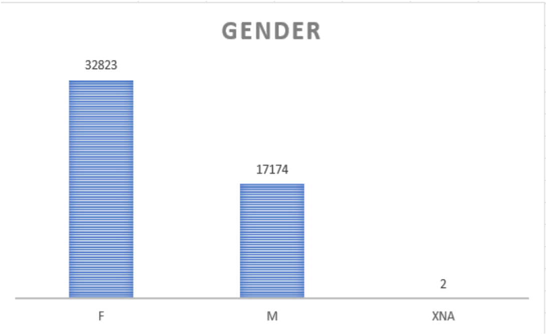
Task: Determine if there is a data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

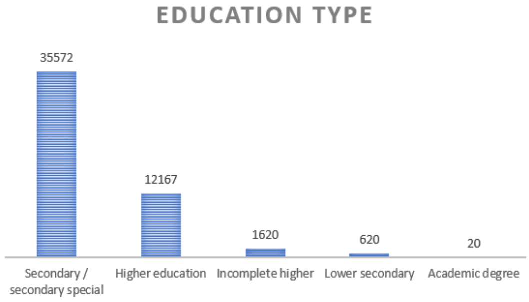
Result:

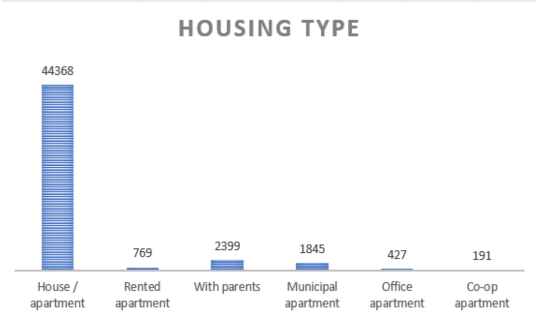
Data imbalance can be observed in many analyses of the following dataset.

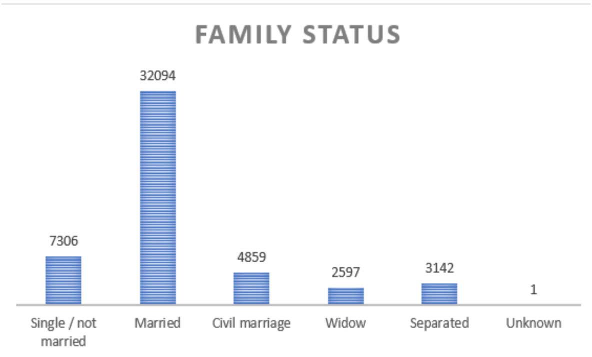
- ***Application_data:***
  - In contract type, as we can see loans have been disbursed in two ways: Cash loans and Revolving loans. There is a huge disparity between how loans are being disbursed. 45276 are given as cash loans and only 4723 are given as revolving loans.
  - As from the graph we can see that female applicants are more than male applicants. There are 32823 female applicants and 17174 male ones.
  - Of the given applicants, the majority of them have Secondary as their education. 35572 have secondary education, 12167 as higher education, 1620 as incomplete higher, 620 as lower secondary, and only 20 as academic degrees.
  - Most of the applicants are applying for loans for houses/apartments which is 44368. Followed by 1845 for municipal apartments, 769 for rented apartments, and 427 for office apartments.
  - Out of applicants, 32094 are married, 7306 are single, and 3142 are separated.
  - Out of the applicants, 34916 of them have no children. 10041 have 1 child and 4319 applicants have 2 children.

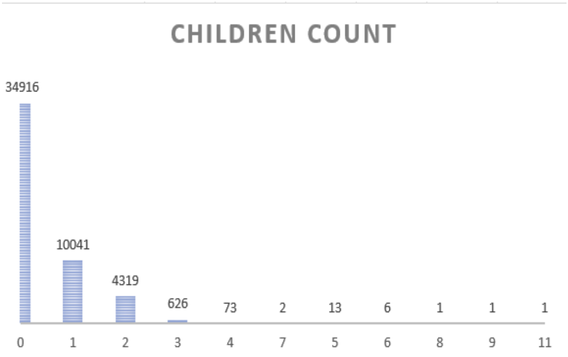




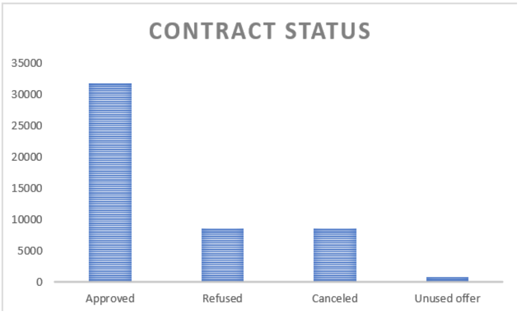


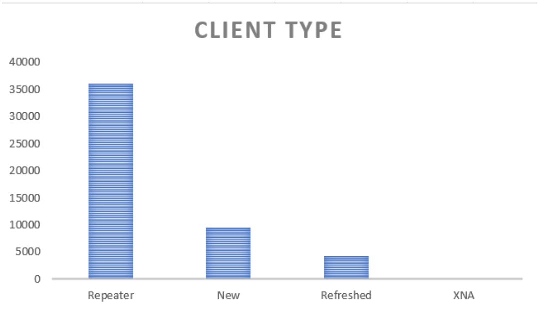






- ***Previous_application_data:***
  - From the application received, 31885 have been approved, 8660 have been refused, and 8595 have been cancelled.
  - Out of the approved clients, 36167 are repeaters, 9548 are fresh applicants, and 4227 are refreshed applicants.

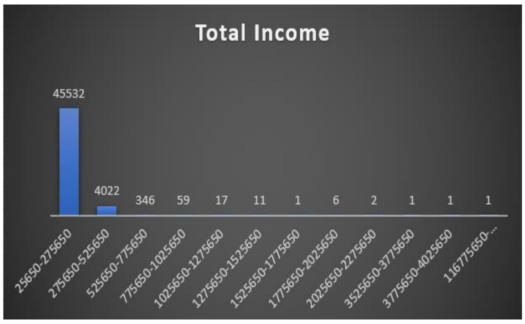


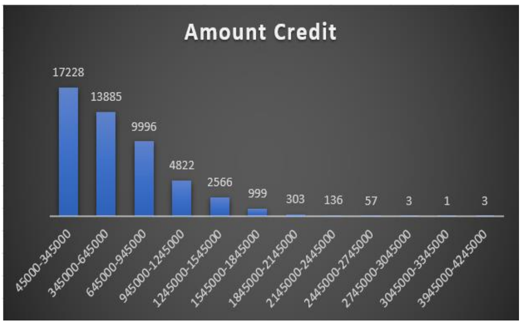


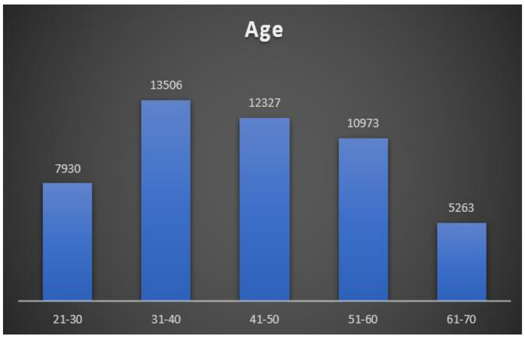
D. **Perform Univariate, Segmented Univariate, and Bivariate Analysis:** To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.
Task: Perform univariate analysis to understand the distribution of individual variables, segmented 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.
Result:

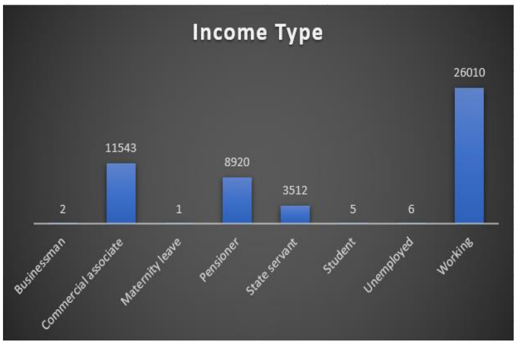
- ***Univariate Analysis:***
- ***Application_data***
  - Most of the applicant's total income lies in the range between $25650-$275650, followed by an income range of $275650-$525650.
  - Most of the applicants applied for the amount of credit in the range of $45000-$345000 with a total number of 17228. 13885 people have applied for amount credit in the range of $345000-$645000.
  - Out of the applications received, the majority of them were in the age group of 31-40 with a total number of 13506. Followed by 41-50 and then 51-60.

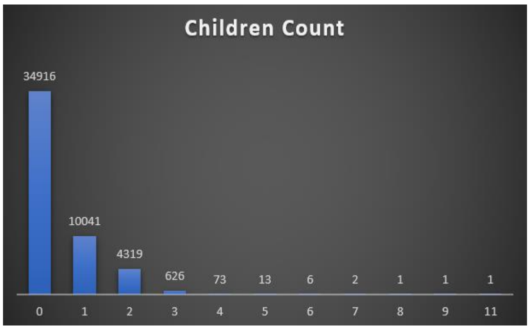
- Working professionals were the most received loan applications with a total number of 26010, followed by commercial associates and then pensioners.
- Of the total applicants, 34916 of them have no children, 10041 have 1 child, and 4319 applicants have 2 children.
- Married people prefer to apply for a loan as it's easy to pay it off, followed by 7306 applicants who are unmarried.
- 44368 people have applied for houses/apartments, 1845 for municipal apartments.
- Most of the applicants have work experience of 0-9 years with a total number of 31951 applicants, followed by 6869 applicants with work exp of 10-19 years.

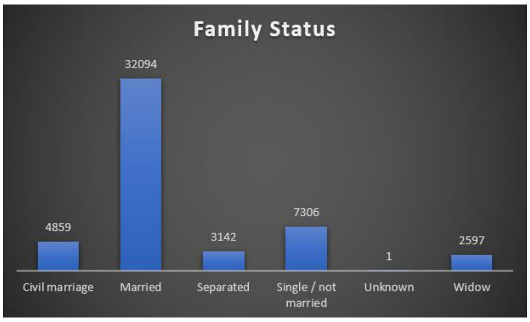


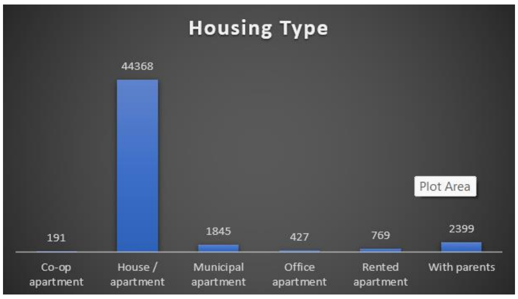


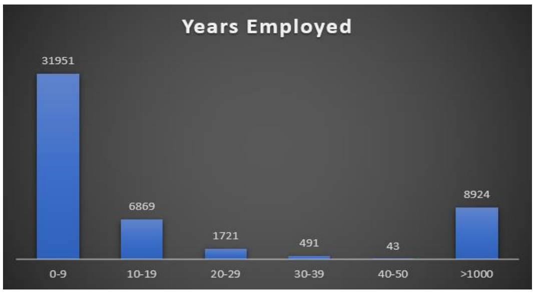




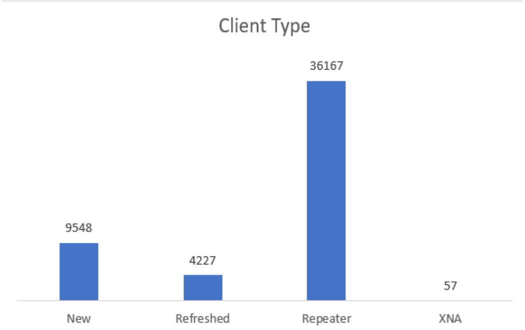


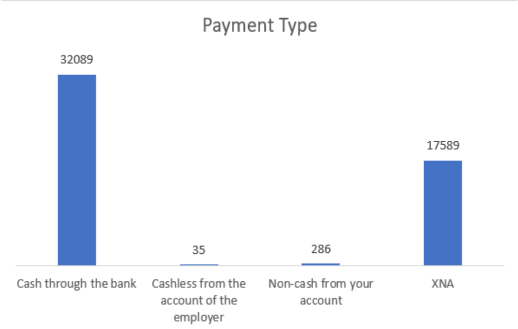


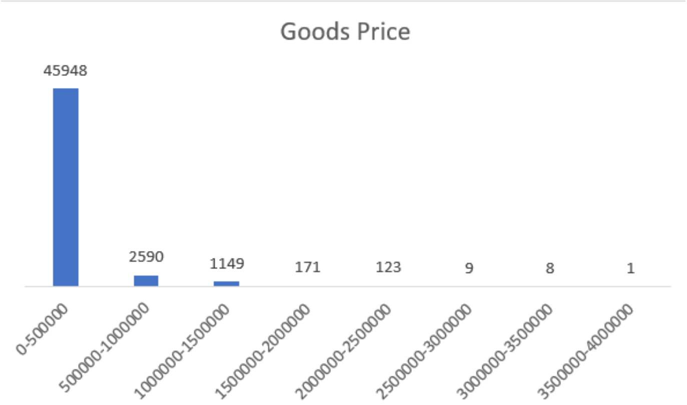




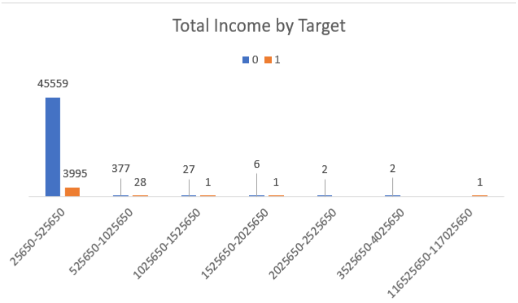
- ***Previous_application_data:***
  - 45948 applicants have goods price in the range of 0-500000.
  - 36167 are repeated applicants, 9548 are new ones, and 4227 are refreshed applicants.
  - Most of the loan payments have been done in cash through the bank.

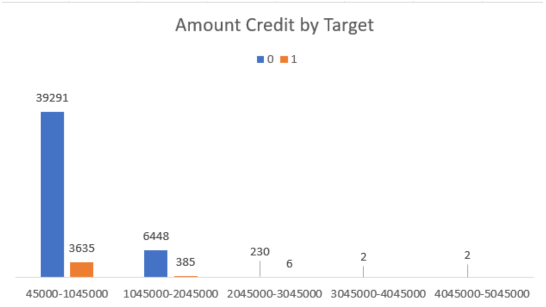


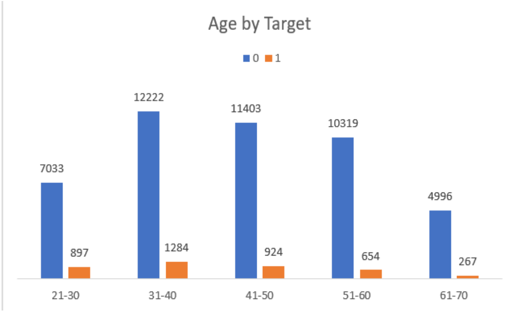


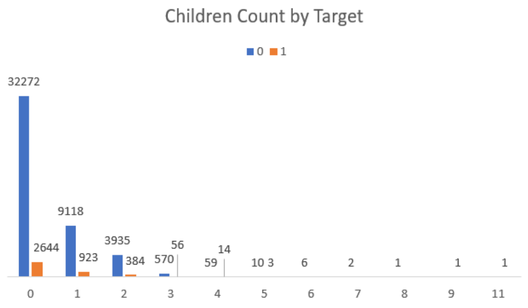


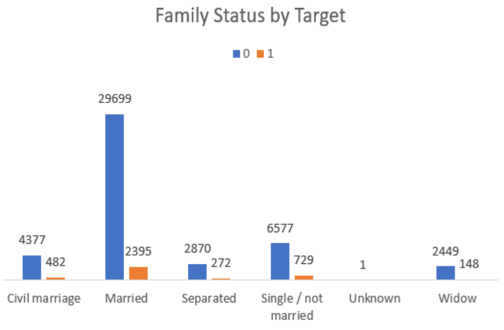
- ***Bivariate Analysis:***
- ***Application_data***
  - In total income by target, 1 denoted late payment and 0 for all other cases. Out of all the applicants, a very handful of people have made late payments. 3995 applicants have made late payments and 45559 applicants have cleared payments on time in the income slab of $25650-$525650.
  - For the amount of credit, 39291 have made the payment on time and 3635 applicants have delayed the payment in the range of $45000-$1045000.
  - Most of the defaulters are in the age group of 31-40 with 1284 applicants and 12222 applicants have made payment on time.
  - Out of the applicants with 0 children, 322272 have paid EMIs on time and 2644 haven’t made it on time
  - Out of the applicants who are married, 29699 have made payments on time, and 2395 are defaulters.
  - Most of the applicants are working professionals, out of which 23549 have paid EMIs on time and 2461 are defaulters.

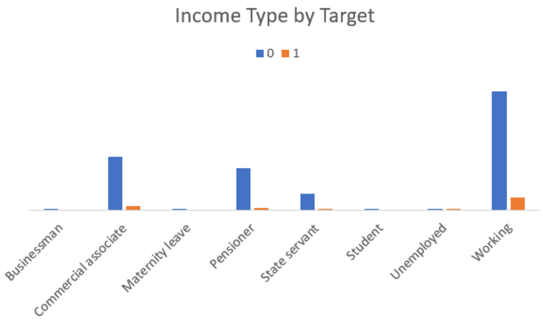






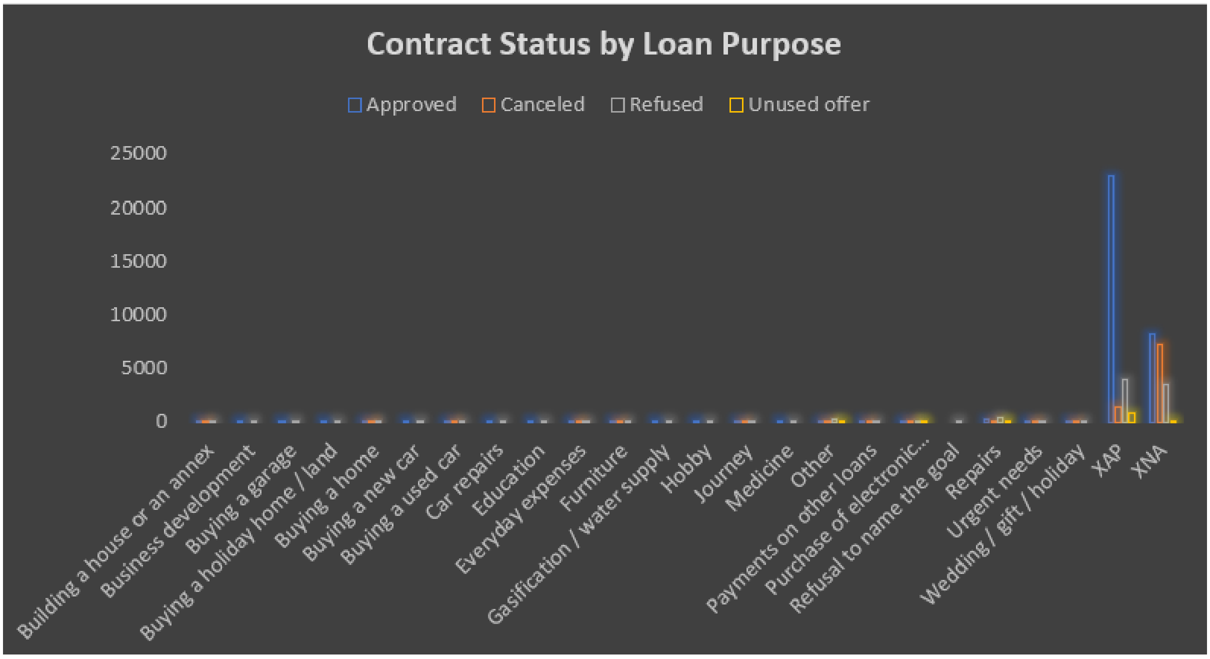


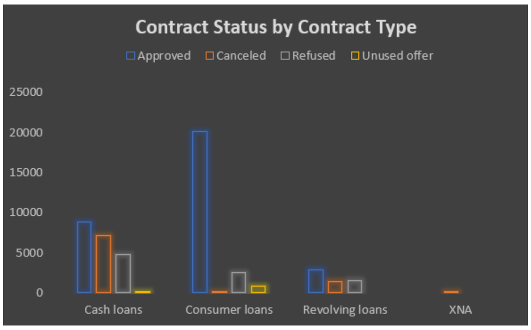


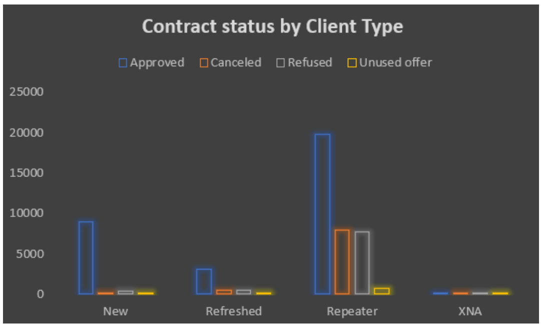


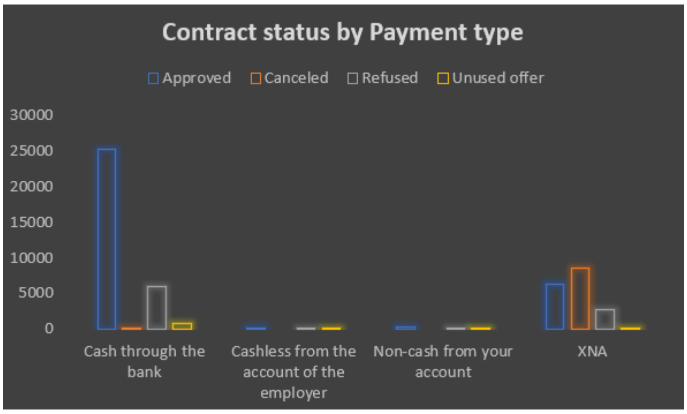
➢ ***Previous_application_data:***

- Out of the total applications received, 22986 have been approved, 1396 have been cancelled, and 3919 have been refused.
- Out of the total applications received, the majority of them were consumer loans. In this, 20149 applications have been approved, 51 are cancelled, and 2468 are refused ones.
- Most of the applicants are repeaters, in which, 19814 are approved, 7924 are cancelled, and 7761 are refused.
- In contract status by payment type, most of them are done by cash through the bank. 25296 are approved, 66 are cancelled, and 5884 are refused.







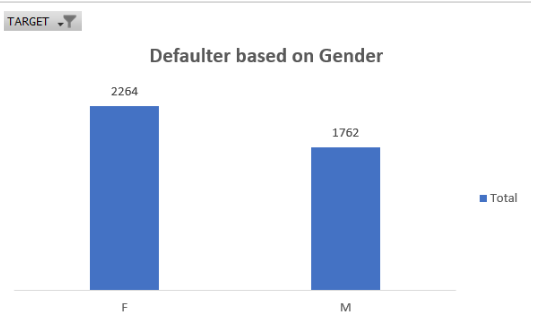


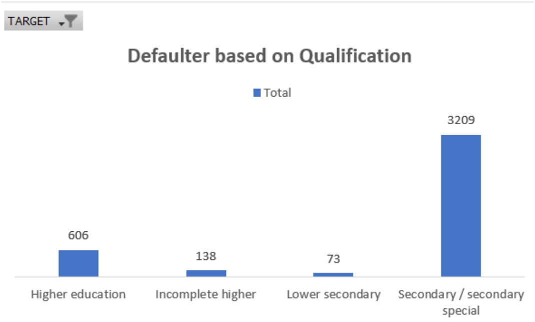
- ***Segmented Univariate Analysis:***
- ***Application_data:***

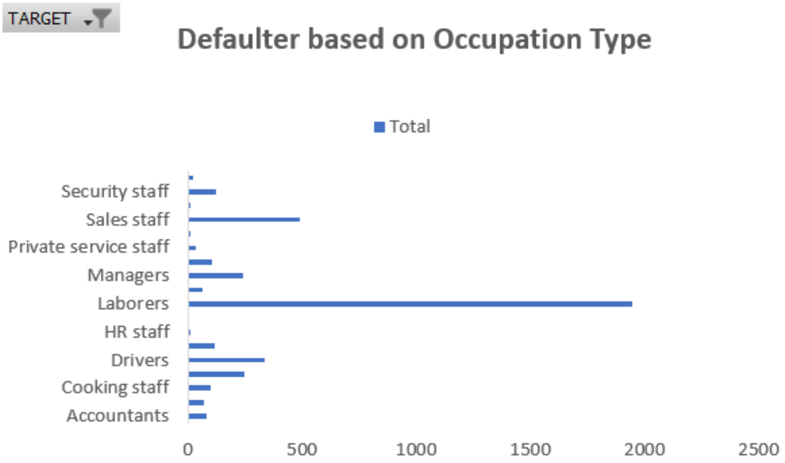
For Segmented Univariate Analysis, I have drawn a survey of loan applicants who are defaulters

- There are more female defaulters than men. 2264 are females and 1762 are males.

- Majority of the defaulters have secondary qualifications with a total of 3209. Followed by Higher education of 606 defaulters, and 138 defaulters of Incomplete higher as a qualification.
- Laborers are the most number of defaulters of all the other occupations. There are 1946 labourers who are defaulters, followed by 492 sales staff and 250 core staff as defaulters.



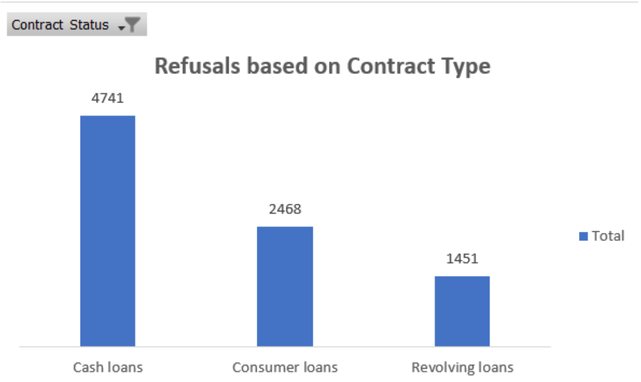


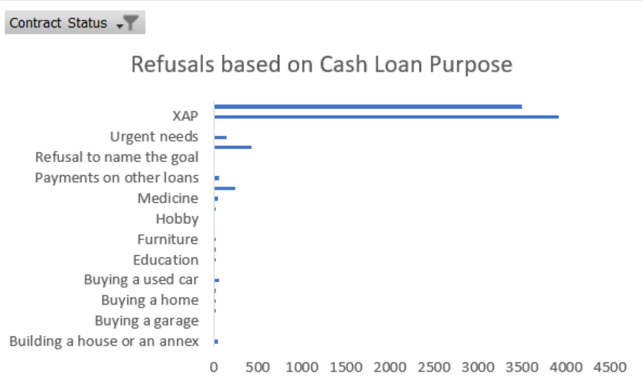


➢ ***Previous_Application_data:***

For Segmented Univariate Analysis, I have drawn a survey of loan applicants who were refused to grant the loan.

- Out of the loans that were refused, the majority of them applied for cash loans with a count of 4741, followed by 2468 consumer loans and then 1451 revolving loan applicants.
- For cash loans that were refused, most of the applicants didn’t mention the reason for taking that loan with a total number of applicants of 3919.



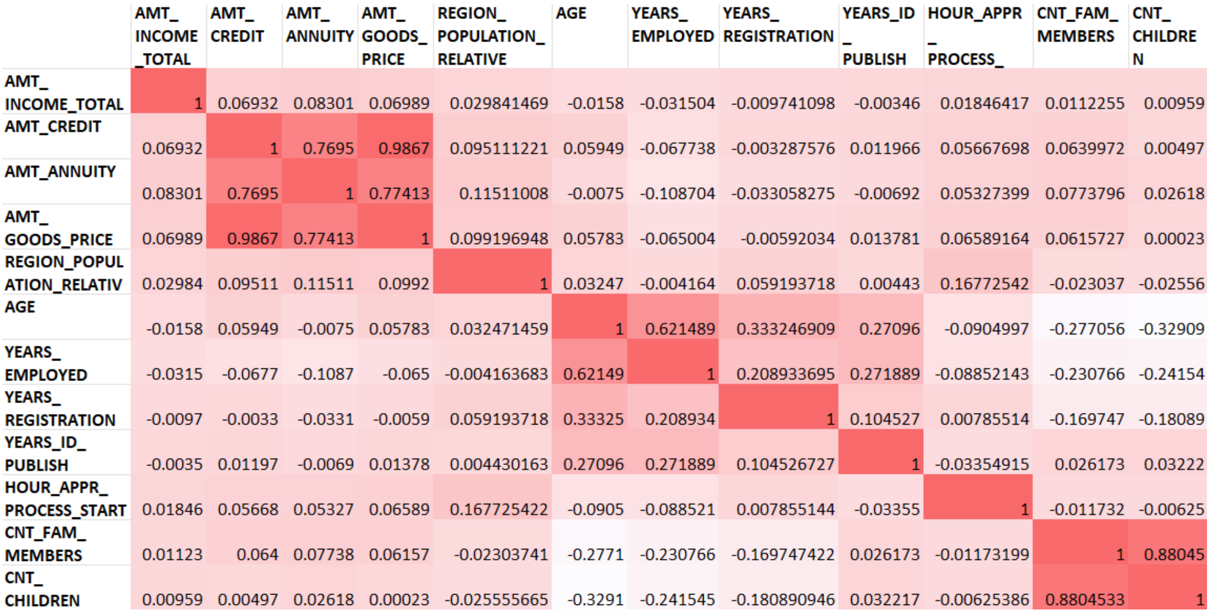


E. **Identify Top Correlations for Different Scenarios:** Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

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.

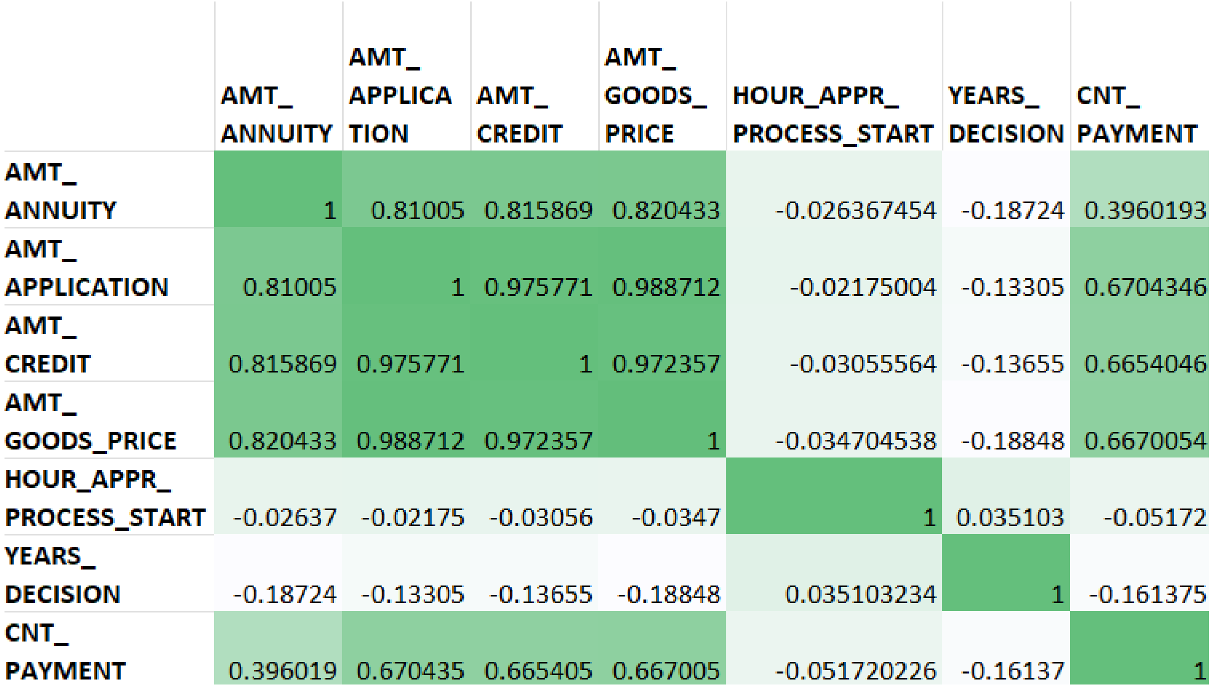
Result:

- ***Application_data:***
  - AMT_CREDIT and AMT_GOODS_PRICE are highly and positively correlated as the Credit amount request is for the Goods whose price is in the AMT_GOODS_PRICE column.
  - CNT_FAM_MEMBERS and CNT_CHILDREN are highly and positively correlated as we observed before that all applicants were either Single Parents or had Nuclear Families.
  - AMT_CREDIT and AMT_ANNUITY are highly and positively correlated and have almost the same Correlation values.
  - Credit amount is highly correlated with amount of goods price which is same as repayments.
  - But the loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to repayment(0.77).



	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	AGE	YEARS_EMPLOYED	YEARS_REGISTRATION	YEARS_ID_PUBLISH	HOUR_APPR_PROCESS_	CNT_FAM_MEMBERS	CNT_CHILDREN
AMT_INCOME_TOTAL	1	0.06932	0.08301	0.06989	0.029841469	-0.0158	-0.031504	-0.009741098	-0.00346	0.01846417	0.0112255	0.00959
AMT_CREDIT	0.06932	1	0.7695	0.9867	0.095111221	0.05949	-0.067738	-0.003287576	0.011966	0.05667698	0.0639972	0.00497
AMT_ANNUITY	0.08301	0.7695	1	0.77413	0.11511008	-0.0075	-0.108704	-0.033058275	-0.00692	0.05327399	0.0773796	0.02618
AMT_GOODS_PRICE	0.06989	0.9867	0.77413	1	0.099196948	0.05783	-0.065004	-0.00592034	0.013781	0.06589164	0.0615727	0.00023
REGION_POPULATION_RELATIV	0.02984	0.09511	0.11511	0.0992	1	0.03247	-0.004164	0.059193718	0.00443	0.16772542	-0.023037	-0.02556
AGE	-0.0158	0.05949	-0.0075	0.05783	0.032471459	1	0.621489	0.333246909	0.27096	-0.0904997	-0.277056	-0.32909
YEARS_EMPLOYED	-0.0315	-0.0677	-0.1087	-0.065	-0.004163683	0.62149	1	0.208933695	0.271889	-0.08852143	-0.230766	-0.24154
YEARS_REGISTRATION	-0.0097	-0.0033	-0.0331	-0.0059	0.059193718	0.33325	0.208934	1	0.104527	0.00785514	-0.169747	-0.18089
YEARS_ID_PUBLISH	-0.0035	0.01197	-0.0069	0.01378	0.004430163	0.27096	0.271889	0.104526727	1	-0.03354915	0.026173	0.03222
HOUR_APPR_PROCESS_START	0.01846	0.05668	0.05327	0.06589	0.167725422	-0.0905	-0.088521	0.007855144	-0.03355	1	-0.011732	-0.00625
CNT_FAM_MEMBERS	0.01123	0.064	0.07738	0.06157	-0.02303741	-0.2771	-0.230766	-0.169747422	0.026173	-0.01173199	1	0.88045
CNT_CHILDREN	0.00959	0.00497	0.02618	0.00023	-0.025555665	-0.3291	-0.241545	-0.180890946	0.032217	-0.00625386	0.8804533	1

- ***Previous_application_data:***
  - AMT_CREDIT and AMT_GOODS_PRICE are highly and positively correlated as the Credit amount request is for the Goods whose price is in the AMT_GOODS_PRICE column.
  - CNT_FAM_MEMBERS and CNT_CHILDREN are highly and positively correlated as we observed before that all applicants were either Single Parents or had Nuclear Families.
  - AMT_CREDIT, AMT_GOODS_PRICE are highly and negatively correlated as ANNUITY%CREDIT is a derived feature which is inversely proportional to AMT_CREDIT and AMT_CREDIT and AMT_GOODS_PRICE are highly and positively correlated as in point 1.
  - There is a severe drop in the correlation between total income of the client and the credit amount(0.038) amongst defaulters whereas it is 0.342 among repayment.
  - Days_birth and number of children correlation has reduced to 0.259 in defaulters when compared to 0.337 in repayment.
  - There is a slight increase in defaulted to observed count in social circle among
  - defaulters(0.264) when compared to repayment (0.254).



	AMT_ ANNUITY	AMT_ APPLICA TION	AMT_ CREDIT	AMT_ GOODS_ PRICE	HOUR_APPR_ PROCESS_START	YEARS_ DECISION	CNT_ PAYMENT
AMT_ ANNUITY	1	0.81005	0.815869	0.820433	-0.026367454	-0.18724	0.3960193
AMT_ APPLICATION	0.81005	1	0.975771	0.988712	-0.02175004	-0.13305	0.6704346
AMT_ CREDIT	0.815869	0.975771	1	0.972357	-0.03055564	-0.13655	0.6654046
AMT_ GOODS_PRICE	0.820433	0.988712	0.972357	1	-0.034704538	-0.18848	0.6670054
HOUR_APPR_ PROCESS_START	-0.02637	-0.02175	-0.03056	-0.0347	1	0.035103	-0.05172
YEARS_ DECISION	-0.18724	-0.13305	-0.13655	-0.18848	0.035103234	1	-0.161375
CNT_ PAYMENT	0.396019	0.670435	0.665405	0.667005	-0.051720226	-0.16137	1

Summary:

Decisive Factors whether an applicant will Repay:

- NAME_EDUCATION_TYPE: Academic degree has fewer defaults.
- NAME_INCOME_TYPE: Students and Businessmen have no defaults.
- REGION_RATING_CLIENT: RATING 1 is safer.
- ORGANIZATION_TYPE: Clients with Trade Type 4 and 5 and Industry Type 8 have defaulted less than 3%
- DAYS_BIRTH: People above the age of 50 have a low probability of defaulting
- DAYS_EMPLOYED: Clients with 40+ years of experience having less than 1% default rate
- AMT_INCOME_TOTAL: Applicants with Income more than 700,000 are less likely to default
- NAME_CASH_LOAN_PURPOSE: Loans bought for Hobby, buying garage are being repaid
  mostly.
- CNT_CHILDREN: People with zero to two children tend to repay the loans.

Decisive Factors on whether an applicant will Default:

- GENDER: Men are at the relatively higher default rate
- NAME_FAMILY_STATUS: People who have civil marriages or who are single default a lot.
- NAME_EDUCATION_TYPE: People with Lower Secondary & Secondary education
- OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as the default rate is huge.

- ORGANIZATION_TYPE: Organizations with the highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self-employed people have a relatively high defaulting rate and thus should be avoided to be approved for loans or provided loans with higher interest rates to mitigate the risk of defaulting.
- DAYS_BIRTH: Avoid young people who are in the age group of 20-40 as they have a higher probability of defaulting
- DAYS_EMPLOYED: People who have less than 5 years of employment have high default rate.
- CNT_CHILDREN & CNT_FAM_MEMBERS: Clients who have children equal to or more than 9 default 100% and hence their applications are to be rejected.

Conclusion:

In this case study, I applied the EDA in the real business case scenario.

- I learned the basics of risk analytics in banking and financial services and understood how
  data is used to minimize the risk of losing money while lending to customers.
- This case study helped me in learning how to summarize a huge dataset to gain valuable insights.
- This project was very challenging. I implemented the study of correlation between different variables to extract the necessary insights for the clients.
- I learned about data imbalance, outliers, and driving factors for the datasets.
- It helped me visualize the huge dataset and summarize the most important results helpful to the client.