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 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.

- ## Approach:

This case study aims to identify patterns which 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 loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected.

Analysis Approach:

1. Imported the datasets (Application_Data & Previous_Application)
2. Identification of both Datasets
3. Outliers: Identified outliers
4. Imbalance: Understanding the ratio of imbalance in our data.
5. Correlation Analysis: Finding the correlation between the variables with respect to the
target variables and find the top three correlations.

- ## Tech-Stack Used:

• Microsoft Excel 365: It enables users to format, organise and calculate data in a spreadsheet. It organises data in an easy-to-navigate way. It has been used to have an overall.

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

1. There are a total of 122 columns in Application_Data and 37 columns in Previous_Application which have missing values greater than 40%. Also removed columns which have more than 30% missing values and the remaining column fill with Median or Mode imputation if required.
2. On further analysis, we found that "EXT_SOURCE_2","EXT_SOURCE_3" has no correlation with the "TARGET" column.

4. There is almost no correlation of ‘FLAG_MOBIL’, ‘FLAG_EMP_PHONE’, ‘FLAG_WORK_PHONE’, ‘FLAG_CONT_MOBILE’, ‘FLAG_PHONE’, ‘FLAG_EMAIL’ with the “TARGET“ column.
5. ‘WEEKDAY_APPR_PROCESS_START’, ‘HOUR_APPR_PROCESS_START’, ‘FLAG_LAST_APPL_PER_CONTRACT’, ‘NFLAG_LAST_APPL_IN_DAY’ are the column in the Previous_Application which are not needed for the analysis.
6. Dropping all the above mentioned columns which will total 26 in Application_Data and 21 in Previous_Application.
7. Converting the negative days column into positive days.
8. Imputing the remaining null values columns needed for data analysis with mean, median
(numerical data) and mode (categorical data).
9. Imputed categorical variable 'NAME_TYPE_SUITE' using mode, 'OCCUPATION_TYPE' by adding an ‘Unknown’ category, numerical variables

'10. Imputed AMT_ANNUITY with median, AMT_GOODS_PRICE with mode, CNT_PAYMENT with 0 as the NAME_CONTRACT_STATUS for these indicate that most of these loans.

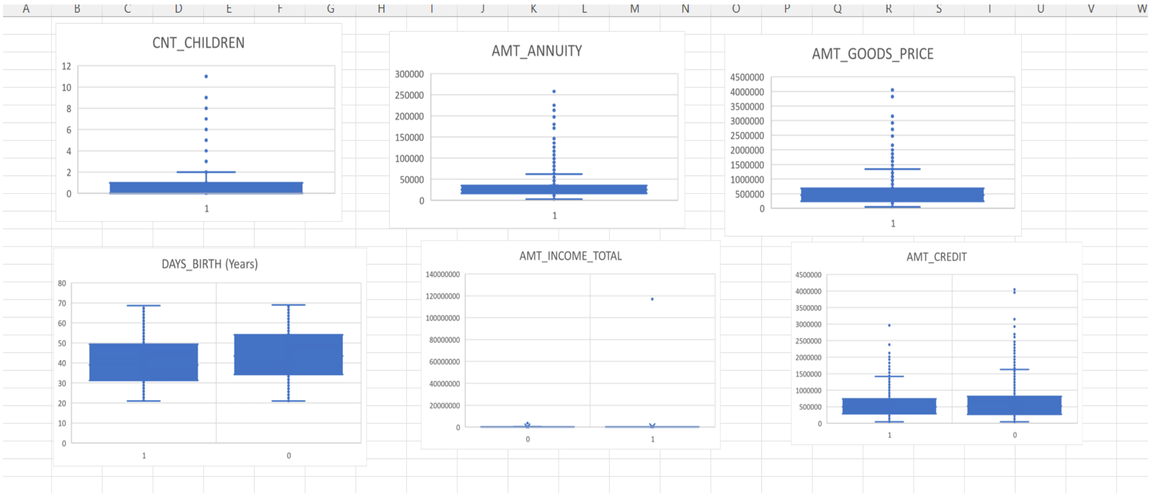
Excel Sheet Link -
https://docs.google.com/spreadsheets/d/1dc0GxW7HY-0C0E3ziw_UYZ_fJHUKLbrh/edit?usp=sharing&ouid=115986816887265464875&rtpof=true&sd=true

https://docs.google.com/spreadsheets/d/1kST86vRb857WeVWLm3pjzDq2A0HTOoiD/edit?usp=share_link&ouid=115986816887265464875&rtpof=true&sd=true

2. 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.

- Application_Data:

1. AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have some number of outliers.
2. AMT_INCOME_TOTAL has a huge number of outliers which indicate that few of the loan applicants have high income compared to the others.
3. DAYS_BIRTH has no outliers which means the data available is reliable.
4. DAYS_EMPLOYED has outlier values around 365243 (days) which is around 1001 years which is impossible and hence this has to be an incorrect entry.



Previous_Application:

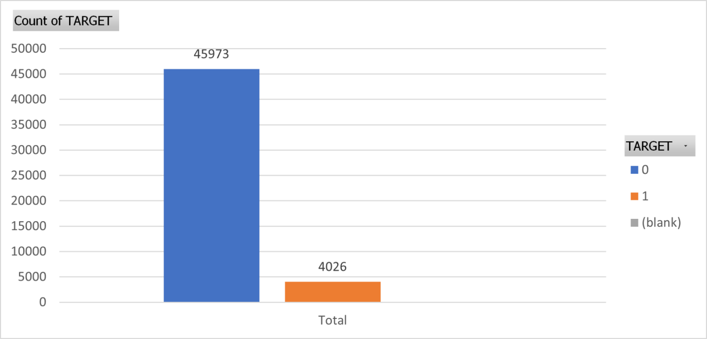
1. AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA has a huge number of outliers.
2. CNT_PAYMENT has few outlier values.
3. SK_ID_CURR is an ID column and hence no outliers.
4. DAYS_DECISION has little number of outliers indicating that these previous applications decisions were taken long back.

Excel Sheet Link - https://docs.google.com/spreadsheets/d/1dc0GxW7HY-0C0E3ziw_UYZ_fJHUKLbrh/edit?usp=sharing&ouid=115986816887265464875&rtpof=true&sd=true

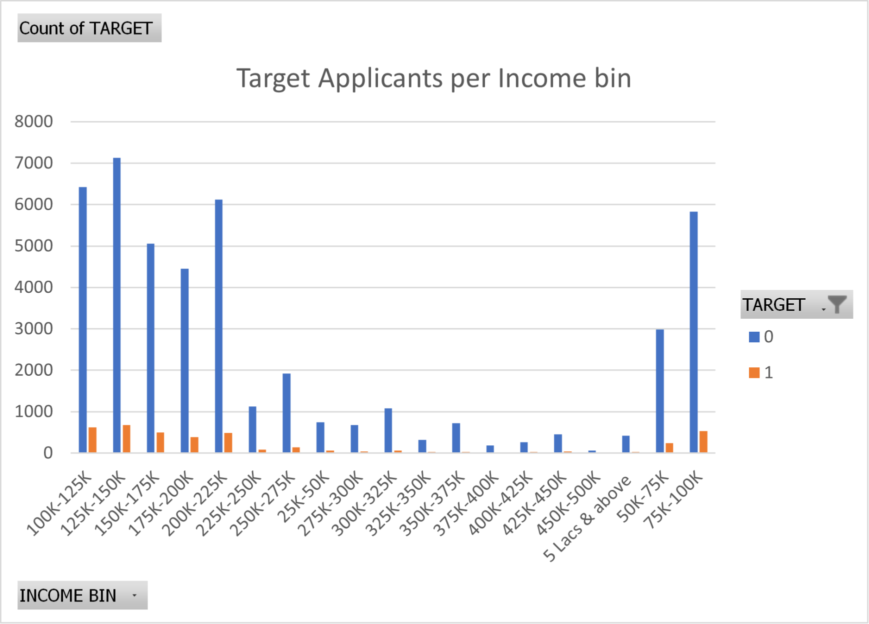
https://docs.google.com/spreadsheets/d/1kST86vRb857WeVWLm3pjzDq2A0HTOoiD/edit?usp=share_link&ouid=115986816887265464875&rtpof=true&sd=true

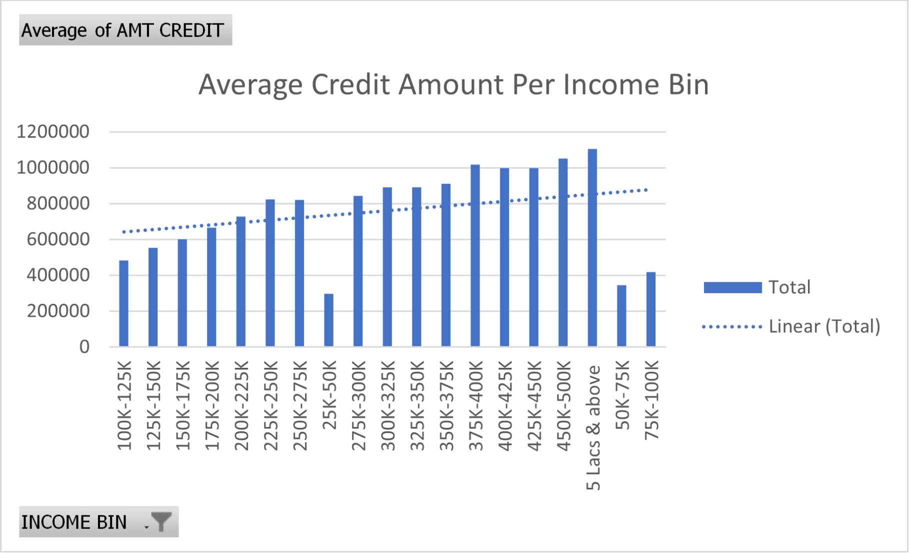
3. Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

This data is highly imbalanced as the Count of 1 's is very less than the Count of 0 's. Data Imbalance Ratio with respect to 0 & and 1 is 11. : 42



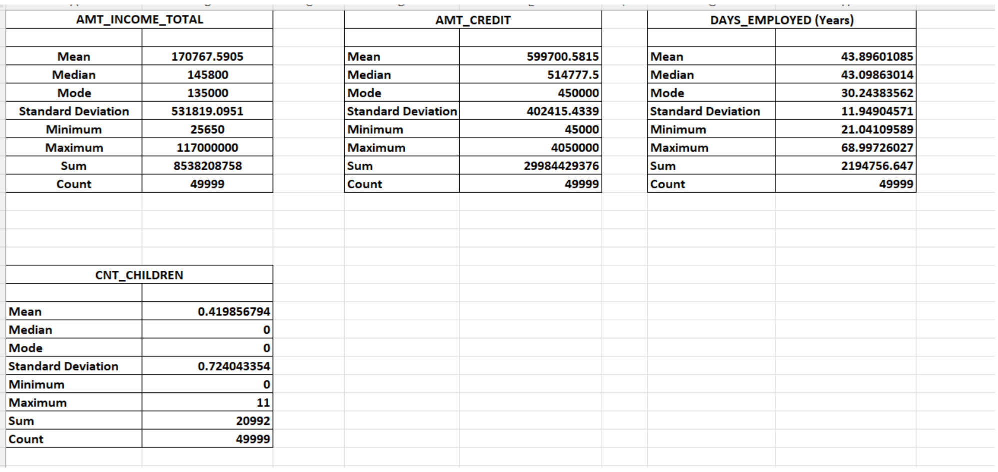
4. Perform Univariate, Segmented Univariate, and Bivariate Analysis: To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.





Excel Sheet Link - https://docs.google.com/spreadsheets/d/1dc0GxW7HY-0C0E3ziw_UYZ_fJHUKLbrh/edit?usp=sharing&ouid=115986816887265464875&rtpof=true&sd=true

https://docs.google.com/spreadsheets/d/1kST86vRb857WeVWLm3pjzDq2A0HTOoiD/edit?usp=share_link&ouid=115986816887265464875&rtpof=true&sd=true



AMT_INCOME_TOTAL	
Mean	170767.5905
Median	145800
Mode	135000
Standard Deviation	531819.0951
Minimum	25650
Maximum	117000000
Sum	8538208758
Count	49999

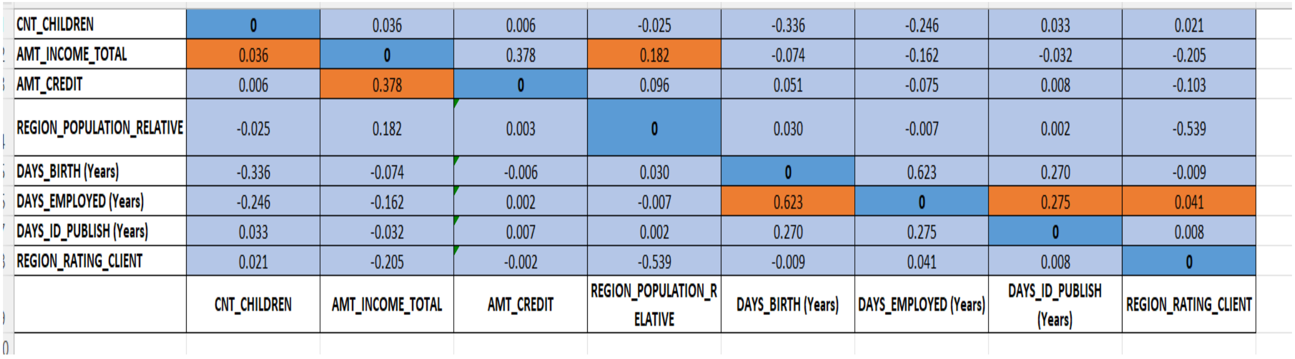
AMT_CREDIT	
Mean	599700.5815
Median	514777.5
Mode	450000
Standard Deviation	402415.4339
Minimum	45000
Maximum	4050000
Sum	29984429376
Count	49999

DAYS_EMPLOYED (Years)	
Mean	43.89601085
Median	43.09863014
Mode	30.24383562
Standard Deviation	11.94904571
Minimum	21.04109589
Maximum	68.99726027
Sum	2194756.647
Count	49999

CNT_CHILDREN	
Mean	0.419856794
Median	0
Mode	0
Standard Deviation	0.724043354
Minimum	0
Maximum	11
Sum	20992
Count	49999

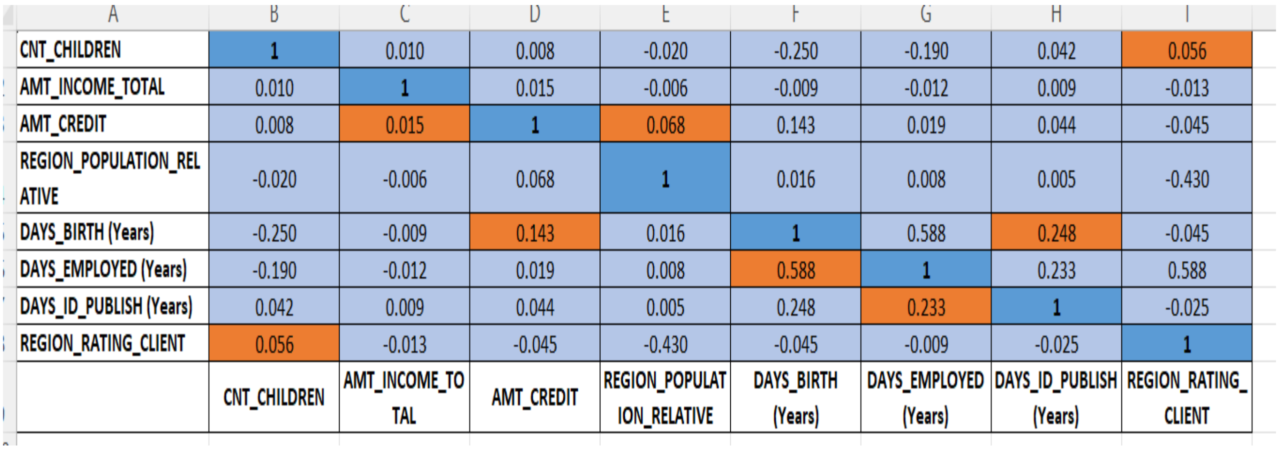
5. Identify Top Correlations for Different Scenarios: Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

- Correlation For Target 0



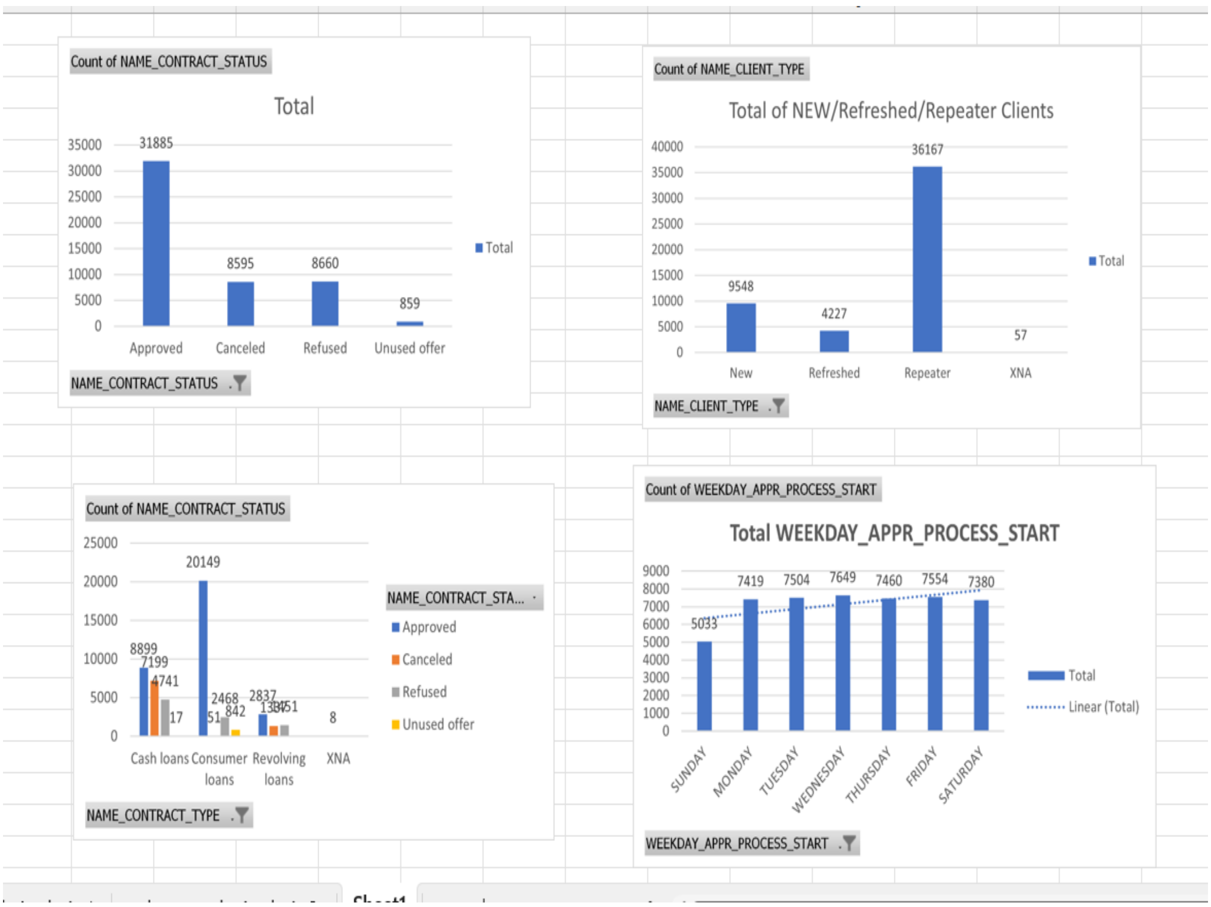
CNT_CHILDREN	0	0.036	0.006	-0.025	-0.336	-0.246	0.033	0.021
AMT_INCOME_TOTAL	0.036	0	0.378	0.182	-0.074	-0.162	-0.032	-0.205
AMT_CREDIT	0.006	0.378	0	0.096	0.051	-0.075	0.008	-0.103
REGION_POPULATION_RELATIVE	-0.025	0.182	0.003	0	0.030	-0.007	0.002	-0.539
DAYS_BIRTH (Years)	-0.336	-0.074	-0.006	0.030	0	0.623	0.270	-0.009
DAYS_EMPLOYED (Years)	-0.246	-0.162	0.002	-0.007	0.623	0	0.275	0.041
DAYS_ID_PUBLISH (Years)	0.033	-0.032	0.007	0.002	0.270	0.275	0	0.008
REGION_RATING_CLIENT	0.021	-0.205	-0.002	-0.539	-0.009	0.041	0.008	0
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH (Years)	DAYS_EMPLOYED (Years)	DAYS_ID_PUBLISH (Years)	REGION_RATING_CLIENT

- Correlation For Target 1



A	B	C	D	E	F	G	H	I
CNT_CHILDREN	1	0.010	0.008	-0.020	-0.250	-0.190	0.042	0.056
AMT_INCOME_TOTAL	0.010	1	0.015	-0.006	-0.009	-0.012	0.009	-0.013
AMT_CREDIT	0.008	0.015	1	0.068	0.143	0.019	0.044	-0.045
REGION_POPULATION_RELATIVE	-0.020	-0.006	0.068	1	0.016	0.008	0.005	-0.430
DAYS_BIRTH (Years)	-0.250	-0.009	0.143	0.016	1	0.588	0.248	-0.045
DAYS_EMPLOYED (Years)	-0.190	-0.012	0.019	0.008	0.588	1	0.233	0.588
DAYS_ID_PUBLISH (Years)	0.042	0.009	0.044	0.005	0.248	0.233	1	-0.025
REGION_RATING_CLIENT	0.056	-0.013	-0.045	-0.430	-0.045	-0.009	-0.025	1
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH (Years)	DAYS_EMPLOYED (Years)	DAYS_ID_PUBLISH (Years)	REGION_RATING_CLIENT

- **Include visualizations and summarize the others results in the presentation Results:**



- **Results :**

• learned basic of risk analytics in banking and financial services and understood how
data is used to minimise the risk of losing money while lending to customers.
• Helped me in learning how to summarise a huge dataset to gain the valuable insights.
• Implemented the study of correlation between different variables to extract the necessary insights for the clients.
• Learned about data imbalance, outliers, driving factors for the datasets.
• Helped me in visualising the huge dataset and summarising the most important results helpful to the client.