Bank Loan Case Study

Project Description

This Project involves analyzing of Lending history data of a Finance company that specializes in lending various types of loans to urban customers and to identify patterns and ensure that capable applicants are not rejected and identify customers who don't have a sufficient credit history take advantage of this and default on their loans

As a data analyst, first I will understand the project requirements and purpose and then I will understand the data which is provided in the project attachment and will perform analysis to get the meaningful insights.

Analysis: involves understanding the relationships between different variables, use of EDU(Exploratory Data Analysis) and statistical analysis to look for impact of factors like Annual Income, Age, Job Experience on Loan Credit Target, and correlation between Target and above factors

For example, Analysis of impact of Income variable over Target variable can reveal that applicants with lower income levels might have a higher likelihood of default

Approach

1. Data Cleaning: Checking for missing or inconsistent data. Handle outliers if necessary. This step ensures that the data is accurate and will not lead to misleading analysis results.
2. Descriptive Statistics: Generating basic statistics such as mean, median, mode, standard deviation etc. This will give a general understanding of the dataset's distribution.
3. Correlation Analysis: Identifying relationships between different variables in the dataset. This can help in understanding how different attributes relate to the likelihood of default.
4. Visual Exploratory Data Analysis: Using plots and charts to visualize the data and identify patterns, trends and outliers. This could include:
    - Bar plots or pie charts for categorical variables like loan status (Approved, Cancelled, Refused, Unused Offer).
    - Histograms or box plots for numerical variables to understand their distribution.
    - Scatter plots to understand the relationship between two numerical variables.
5. Segmentation: Divide the dataset into different groups based on certain criteria (like customers with payment difficulties and all other cases) and analyze these groups separately.

Throughout this process, my goal was to provide clear, accurate, and helpful insights and answers to questions.

Tech-Stack Used

The software and versions you used for the project:

Microsoft Excel
Pivot Tables to classify the data and to calculate Mean, StdDev etc.
Power Query to filter and sort the data
Data analysis tool pack for descriptive statistics

Insights:

I will provide my insights in the form of answers for the questions posed by management team

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.

Output:

1. **Missing Data:** We have two excel files to analyze i.e., “Application Data” and “Previous Application”. In both we have data related to the lending history with multiple columns related to customer and loan details.

   In Application Data excel document we have total 124 columns out of which 60 columns have blank cells.

   In Previous_Application excel document total 38 columns are there out of which 15 Columns have blank cells

   Formula: Countblank() excel function is used to count the number of blank cells in each of above mentioned number of columns and listed them in separate excel sheet “Question A” along with graphical representation.

This Figure shows sample graph that shows count of blank cells in certain columns in application data.
Please see “Question A Graph” in application data excel for complete graph and details

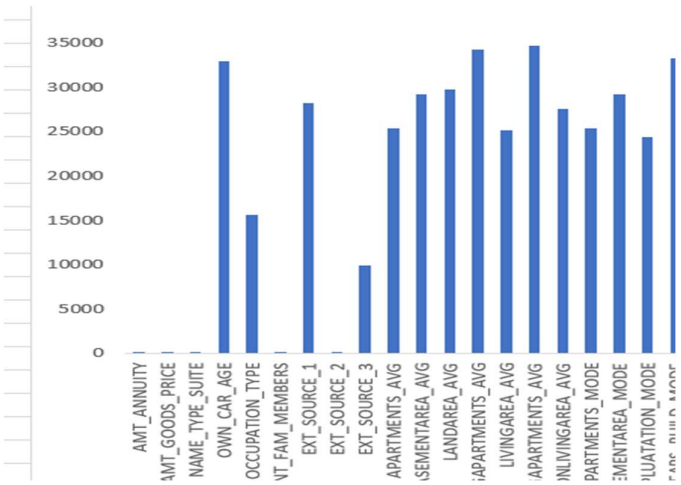
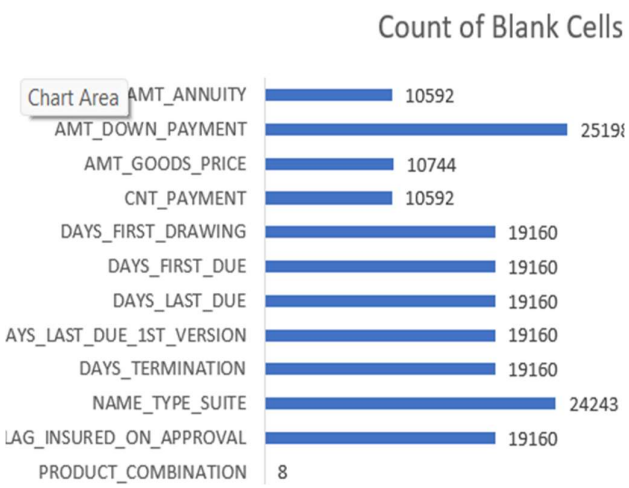


Figure shows sample graph that shows count of blank cells in certain columns in application data.
Please see “Question A Graph” in application data excel for complete graph and details

Figure shows sample graph that shows count of blank cells in certain columns in Precvious_application data.
Please see "Question A Graph" in Previous_application data excel for complete graph and details



2. **Replaced Values:** Based on the data type and data in the columns calculated the values to replace blank cells. For example Mean value for a numerical column, most repeated value for a text or numerical column and fill the blank cells with value in the above cell for text columns etc.

Column Name	Count of Blank Ce	Column Description	Method to replace
AMT_ANNUITY	1	Loan annuity	Loan amount in amt_credit against blank cell is 450000. So, Mean value of the column cells for which Amt_Credit column value is 450000
AMT_GOODS_PRICE	38	For consumer loans it is the price of the goods for which the loan is given	Mean Value of the column
NAME_TYPE_SUITE	192	Who was accompanying client when he was applying for the loan	Most repeated value in the column
OWN_CAR_AGE	32950	Age of client's car	Median of column
OCCUPATION_TYPE	15654	What kind of occupation does the client have	Most repeated value in the column
CNT_FAM_MEMBERS	1	How many family members does client have	Mode of the column
		Normalized score from external	

Column Names	Count of Blank Ce	Method to replace	Replaced Valu
AMT_ANNUITY	10592	Most of the corresponding Amt_Credi	0
AMT_DOWN_PAYMENT	25198	Mean value of the column	6557.57
AMT_GOODS_PRICE	10744	Most of the corresponding Amt_Appli	0.00
RATE_DOWN_PAYMENT	25198	Most of the values are between 0 and	0.08
RATE_INTEREST_PRIMARY	49834	Most of the values are between 0 and	0.19
RATE_INTEREST_ Plot Area D	49834	Most of the values are between 0 and	0.79
NAME_TYPE_SUITE	24243	most repeated value of the column	Unaccompanied
CNT_PAYMENT	10592	Mean value of the column	15.55589109
PRODUCT_COMBINATION	8	most repeated value of the column	POS household with interest
DAYS_FIRST_DRAWING	19160	Fill Down with value in above cell	
DAYS_FIRST_DUE	19160	Fill Down with value in above cell	

Above screen shots shows the methods to replace blank cells. For more details please see "Question A" sheet in both application data and Previous_Application excel files.

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

OUPUT:

Method 1:

Calculated 25th and 75th percentiles and IQR using excel functions on numerical columns to find the upper limit and lower limit of Interquartile range and considered the values outside the range as outliers

	Quartile1(25th Percentile)	Quartile2(Median)	Quartile3(75th Percentile)	IQR	Upper Limit	Lower Limit
5905	112500	145800	202500	90000	337500	-22500
5815	270000	514777.5	808650	538650	1616625	-537975
3457	16456.5	24939	34596	18139.5	61805.25	-10752.75
0361	238500	450000	679500	441000	1341000	-423000
4208	-19644	-15731	-12378.5	7265.5	-1480.25	-30542.25
2449	-2786	-1221	-292	2494	3449	-6527
2666	-7463.5	-4490	-1998	5465.5	6200.25	-15661.75

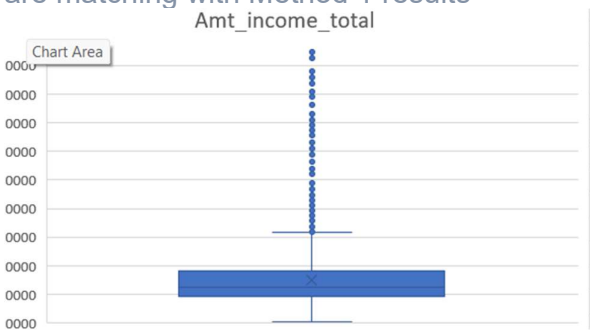
Above picture shows sample of outliers calculation. Please see “Question B” sheet in both application data and previous_application excel files.

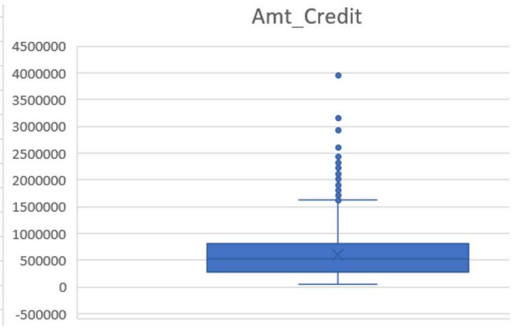
Method 2:

Tried to use Z-Score method to find outliers but because of huge amount of data excel file is getting hanged and getting error while saving. So avoided this method.

Method 3:

Used boxplot graphs on all numerical columns to find the outliers. The results are matching with Method 1 results





Based on the results highlighted the outliers using conditional formatting.

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.

OUTPUT:

Count the Instances of Each Class:
Used Pivot tables to count the number of instances for each class(0,1) in target column.

Imbalance Ratio Analysis

Target	Frequency of TARGET
0	45973
1	4026
Grand Total	**49999**

Calculate the Ratio of Data Imbalance: Divide the number of instances in the minority class by the number of instances in the majority class. This will give you the imbalance ratio.

Imbalance Ratio	11
Proportion of 0	92%
Proportion of 1	8%

Imbalance Ratio	0.087573141

From the above calculation the imbalance ratio of class 1 (which is the class of customers defaulted in loan payments) is significantly less than 1 i.e., 0.08. this indicates a high level of data imbalance.

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.

OUTPUT:

Univariate Analysis: This analysis is used to understand the distribution of individual variables.

Performed statistical analysis like Mean, Median, StdDeviation and Variance on individual columns of application data and previous_application excel files to find their impact.

Univariate Analysis

Column Name	Min	Max	Mean	Median	Mode	StdDeviation	Variance
AMT_INCOME_TOTAL	25650	117000000	170767.591	145800	135000	531813.7768	2.82826E+11
AMT_CREDIT	45000	4050000	599700.582	514777.5	450000	402411.4096	1.61935E+11
AMT_ANNUITY	2052	258025.5	27107.3457	24939	9000	14562.65489	212070917.4
AMT_GOODS_PRICE	45000	4050000	539060.036	450000	450000	369708.9789	1.36685E+11
DAYS_BIRTH	-25184	-7680	-16022.042	-15731	-11653	4361.356655	19021431.87
DAYS_EMPLOYED	-17531	365243	63219.4245	-1221	365243	140793.1977	19822724515

One point found that Loan credit is comparatively more than the customers Income, which can be the reason for Loan defaults.

From the standard deviation results in your data:

1. **AMT_INCOME_TOTAL**: The standard deviation is quite high, indicating a significant variation in the income of the loan applicants. This suggests that the company services a diverse range of customers in terms of income levels.
2. **AMT_CREDIT**: The standard deviation is also high, indicating a wide range of loan amounts that the company has approved. This could mean that the company offers a variety of loan products, from small to large loans.

Segmented Univariate Analysis: This analysis is used to understand and compare the distribution of variables across different segments or groups.

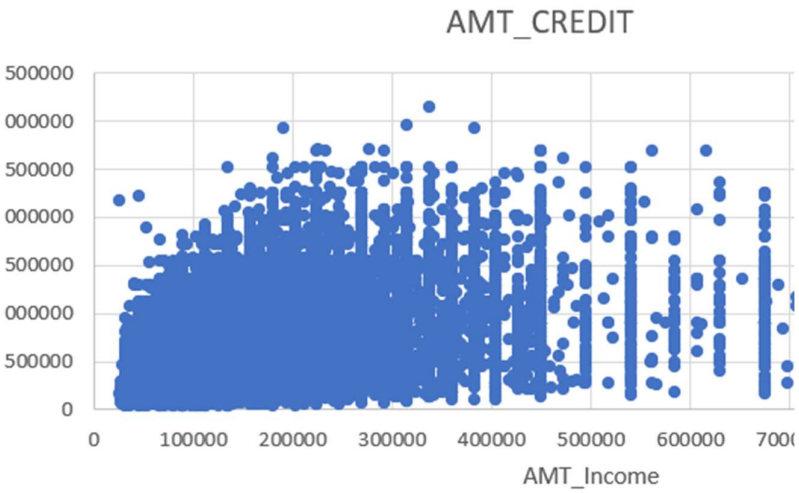
Created Pivot tables, charts to segment the numerical columns under different classes like Target class, age, job experience, loan type to find the patterns and impact

Target	Average of AMT_INC	Average of A	Average of .
0	169053.1507	603562.3	27160.7767
1	190344.8238	555603.52	26497.2161
Grand Total	**170767.5905**	**599700.58**	**27107.3457**

2	**Gender Impact**	**Column Labels**		
3	**Target Class**	**F**	**M**	**X**
4	0	30559	15412	
5	1	2264	1762	
6	**Grand Total**	**32823**	**17174**	
7				
8	**Count of TARGET**	**Column Labels**		
9	**Loan Type**	**0**	**1**	**G**
0	Cash loans	41484	3792	
1	Revolving loans	4489	234	
2	**Grand Total**	**45973**	**4026**	

Bivariate Analysis: This analysis is used to find out if there is a relationship between two sets of values.

Used Scatterplot to find the relation between numerical columns like AMT_Income vs AMT_Credit



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.

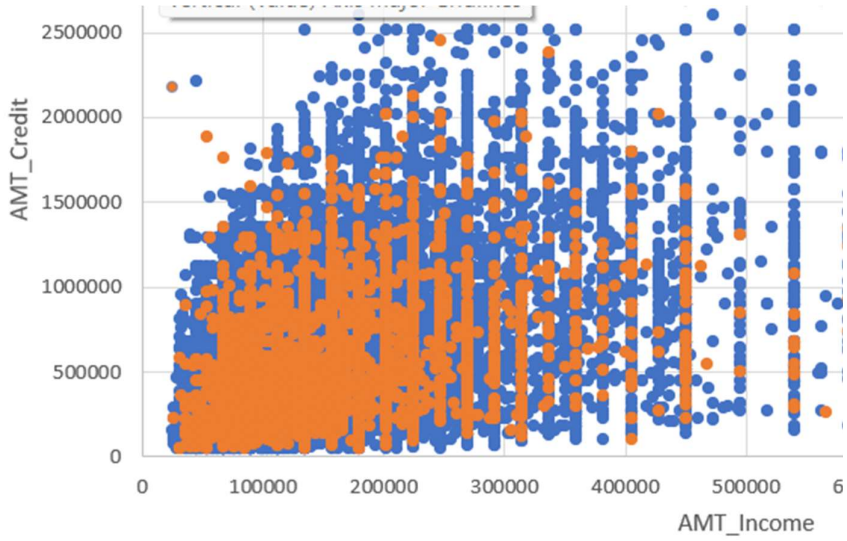
- **Task:** Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

OUTPUT:

Performed correlations calcuations on various numerical columns under various segments to find the indicators of loan default.

1. **Income and Loan**: There is a positive correlation of 0.069 between income and loan amount. This suggests that as income increases, the loan amount also tends to increase, but the relationship is not very strong.
2. **Age and Loan**: For the Target 1 segment (clients with payment difficulties), there is a stronger positive correlation (0.142) between age and loan amount compared to the Target 0 segment (all other cases), which has a correlation of 0.051. This could suggest that older clients in the Target 1 segment tend to take larger loans.
3. **Experience (Exp) and Loan**: For the Target 1 segment, there is a weak negative correlation (-0.016) between experience and loan amount, suggesting that more experienced clients in this segment tend to take slightly smaller loans. However, for the Target 0 segment, there is a positive correlation (0.077), indicating that more experienced clients tend to take slightly larger loans.
4. **Experience (Exp) and Income**: For both segments, there is a positive correlation between experience and income, but it's stronger for the Target 0 segment (0.162) compared to the Target 1 segment (0.011). This could suggest that more experienced clients generally have higher incomes, especially in the Target 0 segment.
5. **Age and Income**: For both segments, there is a weak negative correlation between age and income. This could suggest that income tends to decrease slightly with age in both segments

Used Scatterplot to find the correlation between Customer Income and Loan amount credited



Result:

1. Through Segmented and Correlation analysis and through various charts it is found that there is week correlation between customers income and loan credited. So Finance company should decide, limit and fix the loan amounts based on customer income. In the given data it is found that higher loan amounts given to the customers with lower income.
2. The customers with older age are mostly tend to default the loan payments as per analysis on Target class 1 segment. As per data Finance company is giving higher loans to elder customers which is leading to loan default. Company should change its decision on the age segment.
3. Analysis on Customer's Job experience it is found that income is higher for the more experienced customers but finance company is giving slightly smaller loans to most experienced clients. So company has to change its decision on this segment also.

Drive Link:

For Excel files:

https://drive.google.com/file/d/1VAKt3-pgODR2FmRGFQjGaQME3kk6vTua/view?usp=sharing