BANK LOAN CASE STUDY

Excel Link: Bank-Loan-Analysis-Excel-Workbook

Project Description:

The project's primary objective was to conduct **Exploratory Data Analysis (EDA)** on a dataset from a finance company specializing in lending loans to urban customers. The goal was to identify patterns and factors that indicate whether a customer might have difficulty paying their loan installments, thereby assisting the company in making informed decisions about loan approvals. The project aimed to address the risks of rejecting capable applicants or approving high-risk ones. This involved tasks such as handling missing data, identifying outliers, analyzing data imbalance, and conducting univariate, segmented univariate, and bivariate analyses.

Approach:

To achieve the project's objectives, the following approach was followed:

I went through the risk analytics process step by step, task after task. The project outcomes are as follows:

Overall Method to Analysis: The bank's problem statement is to identify the major causes of bank loan default. The knowledge will be used for risk assessment by the company. We have provided two enormous data sets here.

- **'application data.csv'** contains all of the client's information at the time of application. The information pertains to whether or not a client is having financial issues.
- **'previous application.csv'** provides data from the client's previous loans. It indicates if the prior application was Accepted, Cancelled, Refused, or Unused.

Data Collection: The dataset provided by the finance company was imported into Microsoft Excel 2022, which served as the primary tool for data analysis.

Data Cleaning:

Identifying Missing Data: Excel functions such as COUNT, ISBLANK, and IF were used to identify missing data in the dataset.

Handling Missing Data: Missing data was appropriately handled through imputation using Excel's AVERAGE and MEDIAN functions or by removing rows with missing values.

Outlier Detection:

Identifying Outliers: Excel's statistical functions like QUARTILE and IQR were employed to identify potential outliers in numerical variables.

Outlier Treatment: Thresholds and business rules were applied to determine whether outliers were valid data points or required further investigation.

Data Imbalance Analysis:

Detecting Data Imbalance: Excel functions like COUNTIF and SUM were used to calculate the proportions of each class (e.g., payment difficulties vs. other cases).

Visualizing Data Imbalance: Pie charts and bar charts were created to visualize the distribution of the target variable and highlight class imbalance.

Exploratory Data Analysis:

Univariate Analysis: Excel functions such as COUNT, AVERAGE, MEDIAN, and pivot tables were used to understand the distribution of individual variables.

Segmented Univariate Analysis: Excel features like filters, sorting, and pivot tables were employed to compare variable distributions for different scenarios (e.g., payment difficulties vs. other cases).

Bivariate Analysis: Scatter plots, heatmaps, and other visualizations were created to explore relationships between variables and the target variable.

Correlation Analysis:

Segmenting the Dataset: The dataset was segmented based on different scenarios (e.g., payment difficulties and other cases).
Calculating Correlations: Excel's CORREL function was used to calculate correlation coefficients between variables and the target variable within each segment.

Our Technology Stack:

Software: Microsoft Excel 2019
Purpose: Microsoft Excel was chosen for its robust data analysis and visualization capabilities, making it suitable for tasks like data cleaning, analysis, and visualization.

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.

I completed full EDA on the present application and then on the previous application. Then, in this report, I summarized the results of both applications and provided business insights.

Steps to clean the data:

Handling Missing Data:

We systematically identified and addressed missing values in the dataset. We made decisions on how to handle them by either:

1. **Deleting Rows or Columns:** In cases where missing data was limited and didn't significantly impact the overall dataset, we removed rows or columns with missing values.
2. **Filling in Missing Values:** For missing data that was critical, we employed suitable strategies such as filling in missing values with averages or other appropriate values. This ensured data completeness for subsequent analysis.

Removing Duplicates:

We conducted a check for duplicate rows or records in the dataset and promptly removed them to maintain data integrity. Duplicate entries could distort analysis results, and their removal was essential to ensure the accuracy of our findings.

The existing application sheet included 122 columns.

1. I deleted columns with more than 30% blank data.
2. I deleted a large number of useless columns.
3. I imputed numerical rows blank data by specific row's mean value.
4. Final cleaned dataset consisted of 30 columns.

% of Missing Values	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.4%	
Missing Values Count	0	0	0	0	0	0	0	0	0	1	38	192	
Total Rows Count	49999	49999	49999	49999	49999	49999	49999	49999	49999	49998	49961	49807	
	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUITE	NAI
	100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied	Wor
	100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family	Stat
	100004	0	Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Unaccompanied	Wor
	100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompanied	Wor
	100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompanied	Wor
	100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, partner	Stat
	100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompanied	Com
	100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompanied	Stat
	100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pens
	100012	0	Revolving loans	M	N	Y	0	135000	405000	20250	405000	Unaccompanied	Wor
	100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompanied	Wor
	100015	0	Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Children	Pens
	100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompanied	Wor

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.

1. Calculating the Interquartile Range (IQR):

- We determined the IQR by subtracting the first quartile (25th percentile) from the third quartile (75th percentile). The formula used was: `=QUARTILE.INC(range,3) – QUARTILE.INC(range,1)`.

2. Establishing the Lower Bound for Outliers:

- To identify lower outliers, we subtracted 1.5 times the IQR from the first quartile. The formula used was: `=QUARTILE.INC(range,1) – (1.5 * IQR)`.

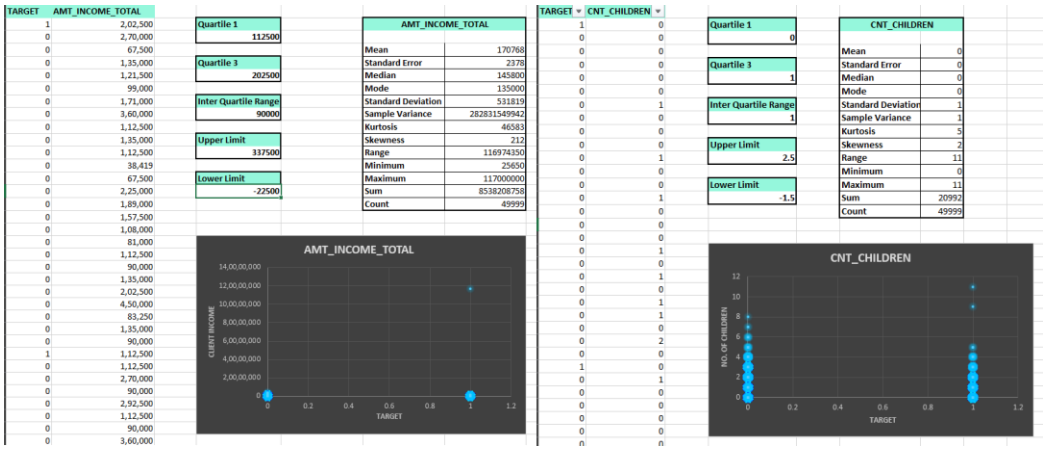
3. Determining the Upper Bound for Outliers:

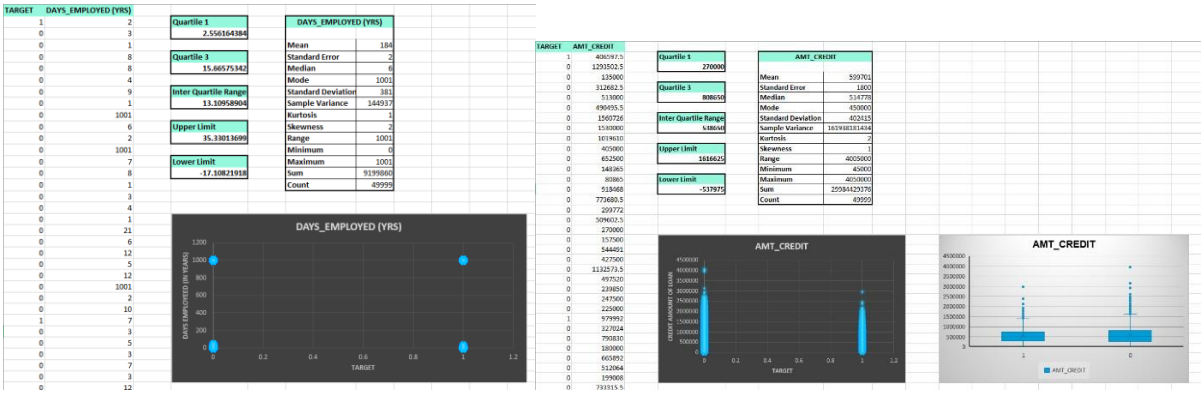
- To identify upper outliers, we added 1.5 times the IQR to the third quartile. The formula used was: `=QUARTILE.INC(range,3) + (1.5 * IQR)`.

4. Flagging Outliers:

- In a new column, we utilized an IF formula to flag outliers within the dataset. The formula used was: ` =QUARTILE3 + (1.5 * IQR)` and ` =QUARTILE1 - (1.5 * IQR)`. Please replace "data" with the appropriate cell reference corresponding to the data point you wish to evaluate.

By following these simplified steps, we were able to efficiently calculate the IQR, establish lower and upper bounds, and accurately flag outliers in our Excel dataset, aiding in the identification and subsequent handling of these data anomalies.





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.

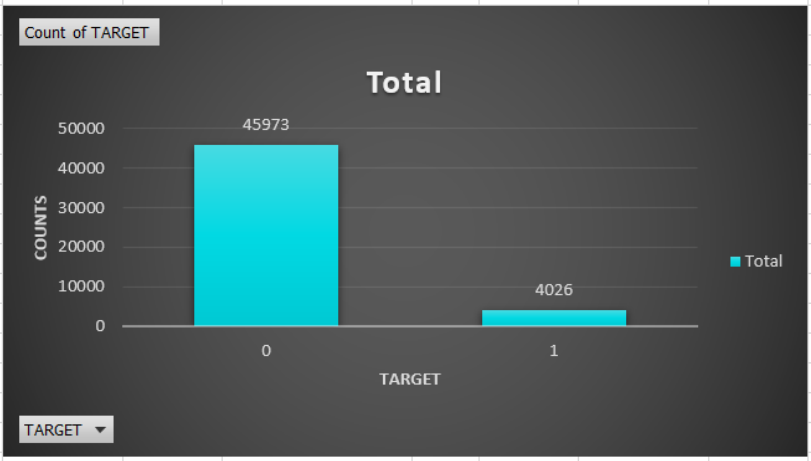
Visualizing Data Imbalance with Bar Charts in Excel:

To effectively visualize and analyze data imbalance in Excel using charts, follow these concise steps:

1. Analyzing Target variable with sub-categories as 0 and 1 for bank loan approved or rejected, respectively, by creating a Pivot Table of its count.
2. Based on counts, analyzing the ratio of both 0 and 1.
3. Find contribution of each sub-category and its proportion as we can see, 0 is 92% and 1 is 8% of total dataset.
4. Visualize the same using Bar Chart.

TARGET	Count of TARGET
0	45973
1	4026
Grand Total	49999

	Count of 0's and 1's	Ratio	Target	Contribution
0	45973	11.42	0	92%
1	4026		1	8%



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.

Insights from Univariate, Segmented Univariate, and Bivariate Analysis:

Univariate Analysis:

- **Distribution of Individual Variables:** Univariate analysis revealed the distribution of various consumer and loan attributes. Histograms, bar charts, and box plots were used to visualize these distributions.
- **Key Takeaways:** This analysis allowed us to identify notable patterns and tendencies within individual variables. For example, we observed that the income distribution skewed towards the lower end, indicating a substantial number of lower-income applicants.

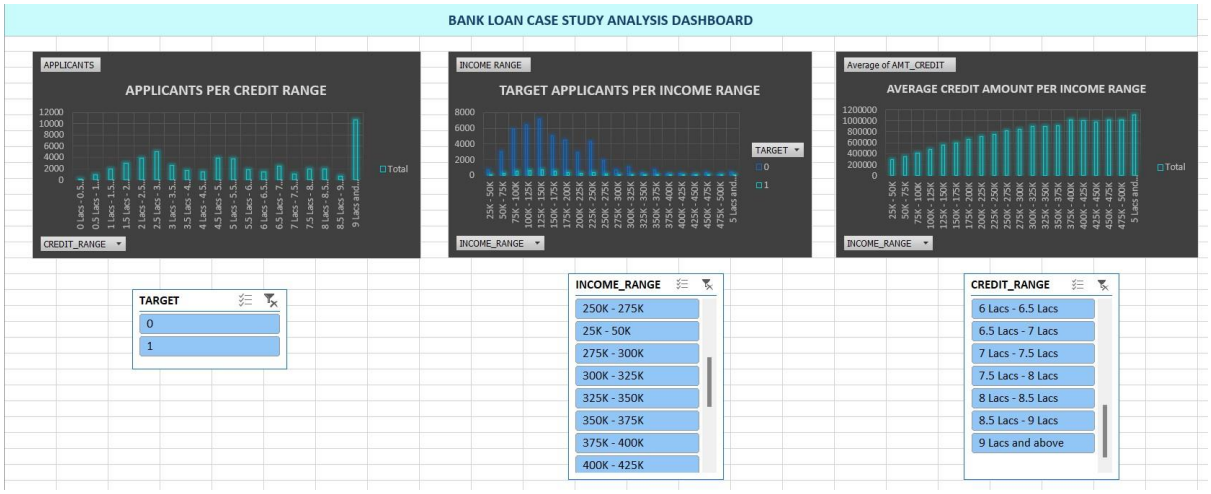
Segmented Univariate Analysis:

- **Comparing Variable Distributions:** Segmented univariate analysis involved comparing variable distributions for different scenarios or segments. Stacked bar charts and grouped bar charts were employed for this purpose.
- **Key Takeaways:** By segmenting the data based on relevant factors (e.g., loan types or applicant categories), we were able to discern differences in variable distributions. For instance, we found that the default rate varied significantly across different loan types, with personal loans showing a higher default rate compared to mortgage loans.

Bivariate Analysis:

- **Exploring Relationships:** Bivariate analysis delved into the relationships between variables and the target variable (loan default). Scatter plots and heatmaps were used to visualize these relationships.
- **Key Takeaways:** This analysis provided valuable insights into the factors influencing loan default. For instance, we observed a negative correlation between applicant income and the likelihood of default, indicating that lower-income applicants were at a higher risk of defaulting on loans.

Overall, these analyses enabled us to gain a deeper understanding of the driving factors behind loan defaults. By visualizing and exploring the data using Excel functions and features, we were able to identify key trends, patterns, and relationships within the dataset. These insights are invaluable for making informed lending decisions and improving risk assessment strategies within the banking industry.



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.

Top ten reasons for loan cancellation and refusal:

1. Count of Children
2. Amount Income Total
3. Amount Credit
4. Amount Annuity
5. Amount Good Price
6. Region Population Relative
7. Birth Days (in years)
8. Employed Days (in years)
9. ID Publish Days (in years)
10. Region Rating Client

Understanding Correlation:

Correlation is a measure that helps us understand how two variables relate to each other. It shows whether they change together, move in opposite directions, or have no apparent relationship. This measure is represented by a correlation coefficient, denoted as "r," which ranges from -1 to 1.

- Perfect Positive Correlation (r=1): When r is 1, it signifies a perfect positive correlation. This means that as one variable increases, the other also increases, and vice versa. They move in the same direction.

- Perfect Negative Correlation (r=-1): Conversely, when r is -1, it indicates a perfect negative correlation. In this scenario, as one variable increases, the other decreases, and vice versa. They move in opposite directions.

- No Correlation (r=0): An r value of 0 implies no correlation. This means that there is no consistent relationship between the two variables. Changes in one variable do not predict changes in the other.

Importantly, correlation does not imply causation; it merely tells us about the strength and direction of the relationship between variables. It is a useful tool for data analysis and prediction in various fields.

To find correlations between the top ten reasons for loan cancellation and refusal using Excel:

1. Open Microsoft Excel and load your loan dataset containing the relevant variables.
2. Select an empty cell for displaying the correlation results.
3. Use the Excel command for Target value=0 as =CORREL('Target 0'!$C:$C, 'Target 0'!B:B) to initiate the correlation calculation.
4. Create same steps for Target value=1 as =CORREL('Target 1'!$C:$C, 'Target 1'!B:B).

Excel will then compute the correlation coefficients for all pairs of variables and display the results inthe chosen cell.

Correlation for Target 0:

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH (YRS)	DAYS_EMPLOYED (YRS)	DAYS_ID_PUBLISH (YRS)	REGION_RATING_CLIENT
CNT_CHILDREN	1									
AMT_INCOME_TOTAL	0.036287306	1	0.37800711	0.451143706	0.384587686	0.181925639	-0.073717622	-0.161561313	-0.032267433	-0.205028897
AMT_CREDIT	0.005705458	0.378007112	1	0.770772965	0.986904954	0.095539444	0.051084182	-0.074733443	0.008290189	-0.102556478
AMT_ANNUITY	0.02638217	0.451143706	0.77077296	1	0.775728	0.117280752	-0.009915685	-0.111294243	-0.009426496	-0.129921207
AMT_GOODS_PRICE	0.001547629	0.384587686	0.98690495	0.775728	1	0.098938662	0.048684805	-0.072483595	0.009304156	-0.104900325
REGION_POPULATION_RELATIVE	-0.024912809	0.181925639	0.09553944	0.117280752	0.098938662	1	0.030435419	-0.006767142	0.002236288	-0.539333113
DAYS_BIRTH (YRS)	-0.335876269	-0.073717622	0.05108418	-0.009915685	0.048684805	0.030435419	1	0.623474675	0.270073313	-0.00902485
DAYS_EMPLOYED (YRS)	-0.245521512	-0.161561313	-0.0747334	-0.111294243	-0.072483595	-0.006767142	0.623474675	1	0.274516224	0.040937165
DAYS_ID_PUBLISH (YRS)	0.032537221	-0.032267433	0.00829019	-0.009426496	0.009304156	0.002236288	0.270073313	0.274516224	1	0.008097427
REGION_RATING_CLIENT	0.021288992	-0.205028897	-0.1025565	-0.129921207	-0.104900325	-0.539333113	-0.00902485	0.040937165	0.008097427	1

Correlation for Target 1:

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH (YRS)	DAYS_EMPLOYED (YRS)	DAYS_ID_PUBLISH (YRS)	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.010110177	0.00760191	0.029172977	-0.001167451	-0.020359154	-0.2496732	-0.189773227	0.042360717	0.055515557
AMT_INCOME_TOTAL	0.010110177	1	0.01527144	0.018004594	0.013261653	-0.006180303	-0.009033662	-0.011758681	0.009122006	-0.012846697
AMT_CREDIT	0.007601905	0.015271444	1	0.749665201	0.982130206	0.067775624	0.142506035	0.018782223	0.043771901	-0.045024534
AMT_ANNUITY	0.029172977	0.018004594	0.7496652	1	0.74932991	0.073123998	0.008751713	-0.078113894	0.02132109	-0.061578289
AMT_GOODS_PRICE	-0.001167451	0.013261653	0.98213021	0.74932991	1	0.076500471	0.140966059	0.023125915	0.04984479	-0.051199549
REGION_POPULATION_RELATIVE	-0.020359154	-0.006180303	0.06777562	0.073123998	0.076500471	1	0.016468731	0.007710059	0.005118563	-0.430032303
DAYS_BIRTH (YRS)	-0.2496732	-0.009033662	0.14250603	0.008751713	0.140966059	0.016468731	1	0.588242824	0.247896571	-0.045027112
DAYS_EMPLOYED (YRS)	-0.189773227	-0.011758681	0.01878222	-0.078113894	0.023125915	0.007710059	0.588242824	1	0.232661912	-0.009237108
DAYS_ID_PUBLISH (YRS)	0.042360717	0.009122006	0.0437719	0.02132109	0.04984479	0.005118563	0.247896571	0.232661912	1	-0.025335227
REGION_RATING_CLIENT	0.055515557	-0.012846697	-0.0450245	-0.061578289	-0.051199549	-0.430032303	-0.045027112	-0.009237108	-0.025335227	1

Result:

Through this project, the following outcomes were achieved:

1. Missing data was effectively handled, ensuring data accuracy.
2. Outliers were identified and treated appropriately.
3. Data imbalance was quantified, providing insights into the distribution of loan defaults.
4. Key factors influencing loan default were identified through univariate, segmented univariate, and bivariate analyses.
5. Strong indicators of loan default were revealed through correlation analysis.

Overall, the project contributed to a better understanding of the factors driving loan defaults, enabling the finance company to make more informed decisions about loan approvals and mitigate risks effectively.