The Art of Lending Loans: A Bank Loan Case Study

Project Description:

The dataset provided is related to Bank loan lending various types of loans to urban customers. The aim of this report is to simply use Exploratory Data Analysis (EDA) to analyse patterns in the data and ensure that capable applicants are not rejected. The impact of this analysis are significant for Banks, lenders, NBFCs who want to understand how customer attributes and loan attributes influence the likelihood of default, based on these insights informed decisions in their future disbursement is to be taken into action. Utilizing a dataset that includes records of previous clients, this project will aim to address specific questions designed to improve the decision making and effectiveness of the disbursement and its procedures. The ultimate goal is to provide data-driven recommendations that contribute to better outcomes for the bankers to generate maximum revenue for the Bank.

ExcelSheet Link Here: Click Here!

OR use the below mentioned link:
https://docs.google.com/spreadsheets/d/1y7MCwuePibLLnkHX56VYv1B7xRVOe7Mi/edit?usp=sharing&ouid=109807839321036950449&rtpof=true&sd=true

Objectives:

A. **Identify Missing Data and Deal with it Appropriately:** Identify the missing data in the dataset and decide on an appropriate method to deal with it.
B. **Identify Outliers in the Dataset:** Detect and identify outliers in the dataset
C. **Analyze Data Imbalance**: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance.
D. **Perform Univariate, Segmented Univariate, and Bivariate Analysis**: 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 variables.
E. **Identify Top Correlations for Different Scenarios:** Segment the dataset based on different scenarios and find correlations for each segmented.

Tech-Stack Used:

Microsoft Excel 2010.

Microsoft Excel 2010 is a powerful spreadsheet application that serves a wide range of purposes, from data organization and analysis to financial modelling and reporting. It offers a user-friendly interface with features such as advanced formulas, pivot tables, and charting tools that enable users to manipulate and visualize data effectively.

Excel 2010 is widely used in various fields, including business, finance, education, and research, allowing users to perform calculations, create budgets, track expenses, and analyse trends. Its ability to handle large datasets and automate repetitive tasks through macros makes it an invaluable tool for professionals seeking to enhance productivity and decision-making based on data insights.

Approach & Results:

#Gather the required data by downloading the datasheet. Open the same using MS Excel(Currerntly using MS Excel 2010)

A. Identify Missing Data and Deal with it Appropriately:

Solution:

>> There are no. of ways one can perform this task. Here we have done it using functions and formulas.

>> There were around 122 attributes in the application data. In order to find the missing values we first found the blank cells in each column(=countblank). From the total of 50000 rows we then converted those blank cells in percentage.

>> We then define the criteria where;

Analyse data	*Discard/Anaylse data*	*Discard data*
<5%	***<30% But >=5%***	***>=30***

- If the column contains < 5% of blank cells we keep the attribute and replace the missing values with descriptive statistical values(mean, median, mode)
- If it contains =>5% but <30% blank cells we keep the data and analyse it later if required further anywhere in our analysis else delete the column.
- If the column contains =>30% of blank cells we directly discard the data attribute; since even if we replace that data with synthetic values we will end up with erroneous values for our analysis.
- We later colour coded these criteria for attributes with conditional formatting.

>> With the above criteria we ended up with 64 columns (attributes) from 122 columns.

>> We even then converted some attributes from days to years with formula =abs(values_in_days/365).

(Total employment year, age of client, etc)

>> We then replaced the missing values with Descriptive Statistical values (mean, median, mode). More details kindly refer to the attached sheet.

>> Wherever necessary we have replaced those missing values based on histogram chart type either mean or median. For categorical attributes mode was used. (Details mentioned in the excel sheet)

B. Outliers in the Dataset:

Detection and identification of outliers in the dataset.

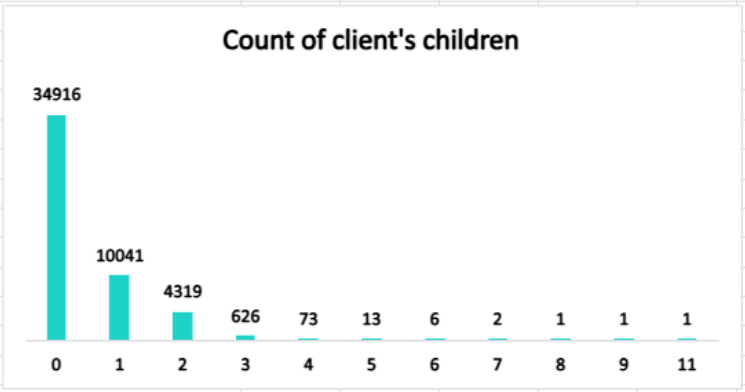
Solution:

>> Majorly for the numerical attributes there have been numerous outliers, some of the outliers and its detection have been mentioned below:

>> Count of Client's Children: With the help of pivot table and column graph we plotted these values and noticed clients having => 8 children as outliers. It's almost not possible to have => 8 children in 21st century.

>> The bar graph for the client's children have been plotted is shown below:

Count of child	Count of Client_CHILDREN
0	34916
1	10041
2	4319
3	626
4	73
5	13
6	6
7	2
8	1
9	1
11	1
Grand Total	49999

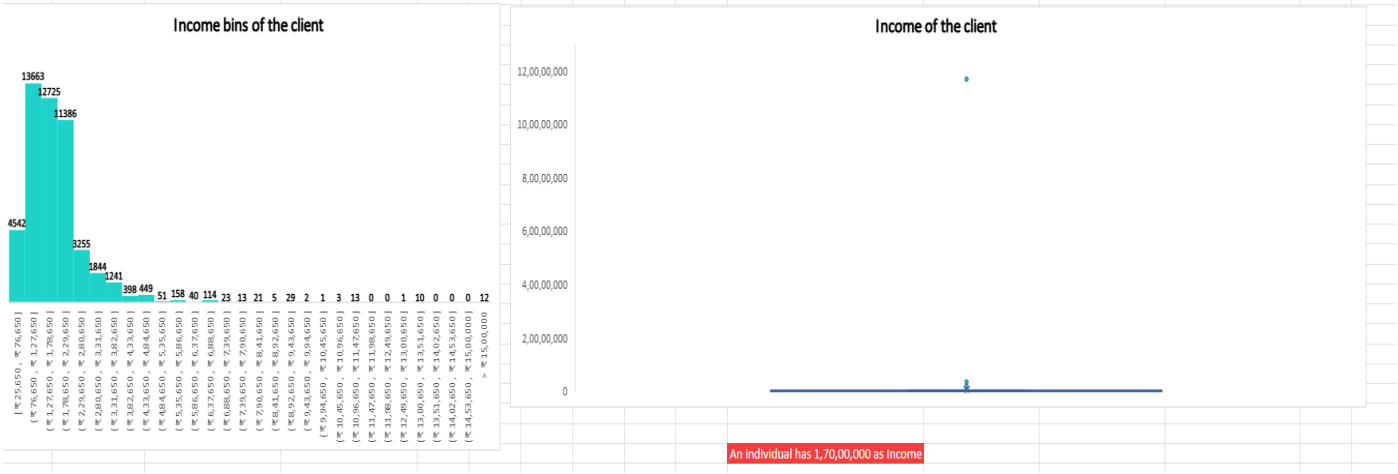


>> Income of the Client: We plotted a histogram of the income of the clients to get the breadth of the income distribution.

>> We observed the data is right skewed as we have earlier replaced the missing values for the same with median. We then plotted the whisker plot for this data.

>> We noticed a few clients having income 1,70,00,000/- as annual income, which can easily be seen on the whisker plot as an outlier beyond the max values. By judging from the rest of the client this income in this dataset was irrational. Hence we removed these rows as outliers.

>> The Income histogram and the whisker plot has been pasted below for reference.

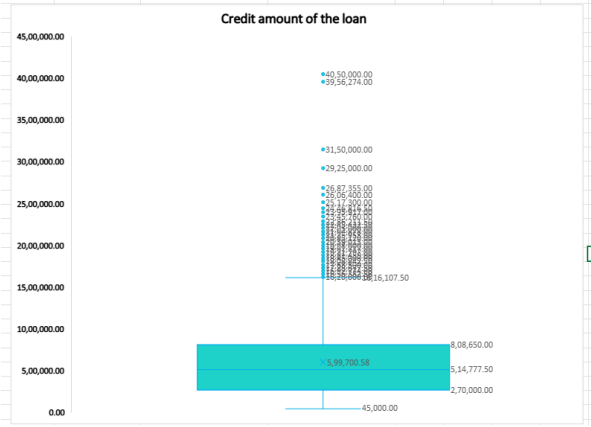


>> Credit amount of the loan: We plotted a whisker plot for the amount of the loans credited from the dataset with the help for quartile values.

>> We noticed many incomes as outliers. But to genuinely detect the outliers we used a filter where clients **income <3,00,000/- ; credit amount of the loan >25,00,000/- and amount of goods price < loan amount credited** using conditional formatting.

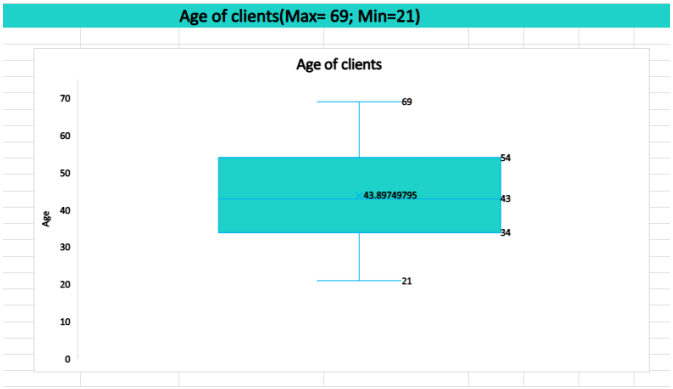
>> We detected these outliers and then removed them from the dataset

>> SS for the same has been pasted below:



Outlier			
<3,00,000	>25,00,000		Collateral <Loan amount
Income of the client	Credit amt of the loan	AMT_ANNUITY(payments	AMT_GOODS_PRICE(Colateral)
2,70,000	25,17,300.00	1,02,658.50	22,50,000.00
2,70,000	25,17,300.00	66,402.00	22,50,000.00
2,25,000	25,17,300.00	69,354.00	22,50,000.00
2,02,500	25,17,300.00	66,532.50	22,50,000.00
2,34,000	26,95,500.00	77,382.00	22,50,000.00
2,25,000	25,17,300.00	66,532.50	22,50,000.00
2,92,500	25,17,300.00	66,402.00	22,50,000.00
2,29,500	25,17,300.00	72,274.50	22,50,000.00
2,25,000	25,17,300.00	73,732.50	22,50,000.00
2,70,000	25,17,300.00	69,223.50	22,50,000.00
1,80,000	25,17,300.00	69,223.50	22,50,000.00
1,80,000	26,06,400.00	79,155.00	22,50,000.00
2,70,000	25,17,300.00	66,532.50	22,50,000.00
2,83,500	25,17,300.00	66,532.50	22,50,000.00
2,70,000	25,17,300.00	66,532.50	22,50,000.00
1,35,000	25,17,300.00	66,532.50	22,50,000.00
2,11,500	25,17,300.00	66,402.00	22,50,000.00
2,25,000	25,17,300.00	69,354.00	22,50,000.00
1,90,611	29,31,660.00	74,110.50	27,00,000.00
2,25,000	26,95,500.00	74,254.50	22,50,000.00
2,25,000	25,17,300.00	66,532.50	22,50,000.00
2,92,500	26,95,500.00	71,235.00	22,50,000.00

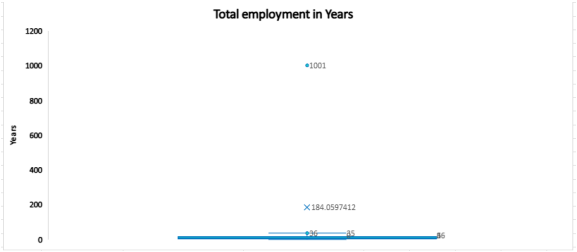
>> **Age of the client:** We plotted a whisker plot a noticed no outliers, the same have been pasted below:



>> **Total Employment in Years:** We found the Min and the MAX values for the same and noticed clients having 1001 years of employment.

>> We later then plotted a box plot for the same. Every other values were normal except for 1001 years of employment in years.

>> SS for the values of box plot have been pasted for your reference:



>> After removing these rows we were left with 41,108 Rows from 50,000 within this dataset.

C. Analyze Data Imbalance:

Determination of data imbalance in the loan application dataset and calculation of the ratio of data imbalance.

Solution:

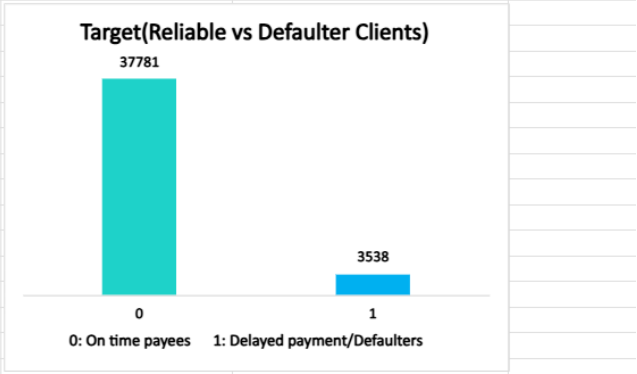
>> We plotted a bar graph for 'Target' attribute where '0' being the number of Non-Defaluter and '1' as other cases as Defaulters.

>> We noticed a clear case of bata imbalance here where the count of Defaulter stood at 3538 and Non-defaulters at 37781

>> In terms of banks repo and profitability this data show a good sign with TARGET Clients (As most of them are paying on time). However, the ratio of imbalance is 10.678632 i.e., for every 1 defaulters there were 10 Non-defaulters against it, clearly highlighting data imbalance within the dataset.

>> SS for the same has been attached below:

TARGET(1-defaulters;0-On_	Count of TARGET(1-defaulters;0-On_	Target Percent
0	37781	91%
1	3538	9%
Grand Total	**41319**	**100%**

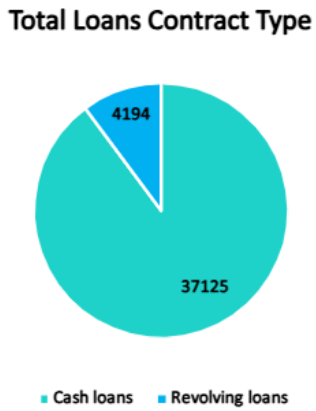


>> Later, to further analysis this data we plotted a pie chart for revolving loans and cash loans.

>> 90% of contract indicated cash loans and 10% indicated Revolving loans.

>> Ratio of imbalance stood at 8.8.

>> The SS for the same have been pasted below for your reference.



D. Perform Univariate, Segmented Univariate, and Bivariate Analysis:

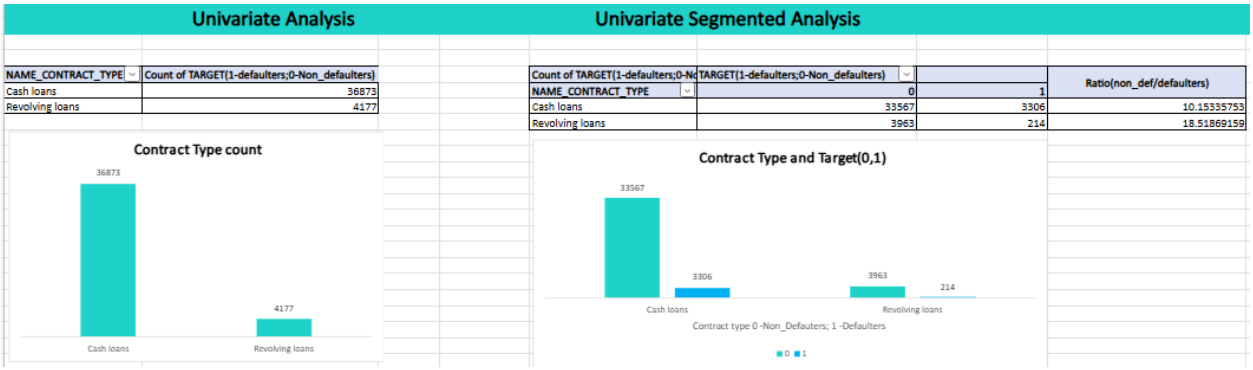
Determination of Univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variables.

Solution:

>> Univariate Analysis of Contract type and Segmented Univariate Analysis of Contract type with Target attribute.

>> We plotted a bar-chart for this dataset, and noticed; For every 1 defaulters under cash loans there are 10 Non-defaulters and Similarly, for every 1 defaulters in Revolving loans there are 18 non defaulters. Hence proving Revolving loans are more prudent and Non-defaulting in nature then the cash loan contract for the banks.

>> The SS for the same has been pasted below for the reference:



>> Univariate Analysis of Code_Gender and Segmented Univariate Analysis of Code Gender with Target attribute.

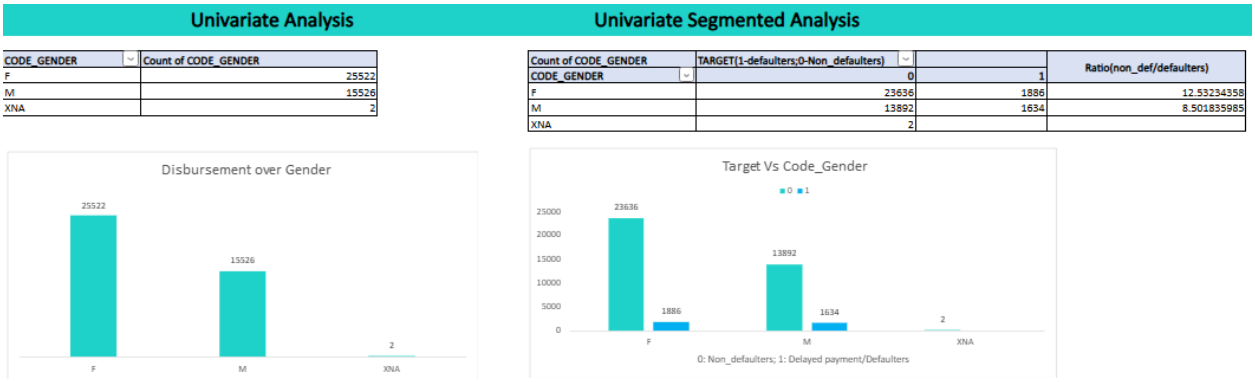
>> We plotted a bar chat for the above mentioned attributes and observed majority of the loans have been distributed to Females compared to Males

>> The data is imbalanced where females have received more loans credit compared to males; this could also imply that the counts of defaulters for Females are eventually high.

>> Additionally, after calculating the ratio of Target attribute data indicates males have higher default ration when compared to Females

>> Data indicates banks must focus more on Females for distribution of loans as they have low default ratio.

>> The SS for the same has been pasted below for the reference:



>> Univariate Analysis of Client's_Children and Segmented Univariate Analysis of Client's_Children with Target attribute.

>> The data indicates clients with lower count of children have defaulted more compared to higher count of children.

>> Therefore by calculating the ratio of Non-defaulters to defaulters, we can conclude; clients with high count of children have higher default occurrences.

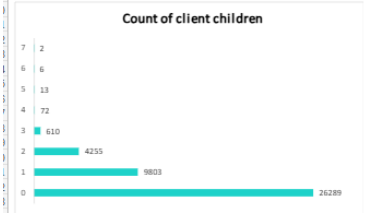
>> The SS for the same has been pasted below for the reference:

Univariate Analysis

Client_CHILDREN	Count of Client_CHILDREN
0	26289
1	9803
2	4255
3	610
4	72
5	13
6	6
7	2

Univariate Segmented Analysis

Count of Client_CHILDREN	TARGET(1-defaulters;0-Non_defaulters)		Ratio(non_def/defaulters)
Client_CHILDREN	0	1	
0	24114	2175	11.08689655
1	8906	897	9.928651059
2	3879	376	10.31648936
3	555	55	10.09090909
4	58	14	4.142857143
5	10	3	3.333333333
6	6		
7	2		



>> Univariate Analysis of Client's_Income and Segmented Univariate Analysis of Client's_Income with Target attribute.

>> We divided the client's income into 8 bins ranging from 25k to 100K (75K interval) up until 550k and Above.

>> We distributed the dataset in these bin for our client income

>> Clients from 1laks to 1.75 lakhs of income have the highest count of individual, for the same the defaulters are also high.

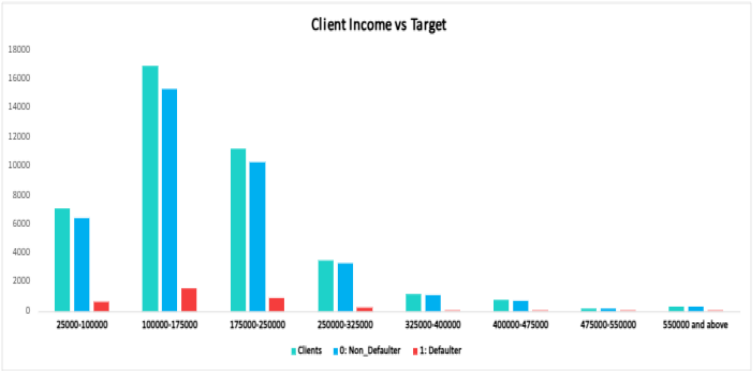
>> However, considering the target ratio for Non-Defaulters/Defaulters we see individuals from 1lakh-1.75 lakh have defaulted the most followed by the lower income group 25K to 1lakh.

>> We also observe a pattern, as the income increases the defaulters of loan decreases.

>> Banks must be careful when disbursing loan to clients having lower income level

>> The SS for the same has been pasted below for the reference:

Client Income Interval	Count of Clients	Target		Ratio of Target
		0-Non Defau	1-Defaulters	
25000-100000	7073	6419	654	9.8
100000-175000	16871	15287	1584	9.7
175000-250000	11183	10281	902	11.4
250000-325000	3510	3285	225	14.6
325000-400000	1191	1124	67	16.8
400000-475000	745	688	57	12.1
475000-550000	161	150	11	13.6
550000 and above	316	296	20	14.8



>> Univariate Analysis of Income_Type and Segmented Univariate Analysis of Income_Type with Target attribute.

>> For the income_type Univariate analysis we plotted a Pivot table and based on that we managed to graph a line chart for each category.

>> The disbursement is highest in Commercial associates and Working but the defaulters are high in working clients.

>> State servants here the least defaulters in this dataset compared to any other category, maximum defaulters are from working category.

>> Banks must disburse more loans to state servant followed by commercial associates.

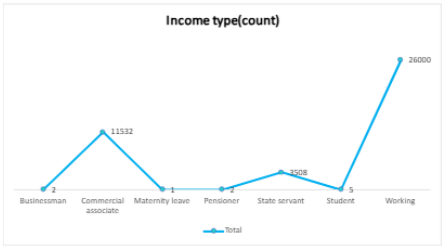
>> The SS for the same has been pasted below for the reference:

Univariate Analysis

Client_INCOME_TYPE	Count of Client_INCOME_TYPE
Businessman	2
Commercial associate	11532
Maternity leave	1
Pensioner	2
State servant	3508
Student	5
Working	26000

Univariate Segmented Analysis

Count of Client_INCOME_TYPE	TARGET(1-defaulters;0-Non_defaulters)		Ratio of Target(Non def to Defaulters)
Client_INCOME_TYPE	0	1	
Businessman	2		
Commercial associate	10669	863	12.3626883
Maternity leave	1		
Pensioner	2		
State servant	3310	198	16.71717172
Student	5		
Working	23541	2459	9.573403823



>> Univariate Analysis of Client's Education and Segmented Univariate Analysis of Client's Education with Target attribute.

>> With the Client's education data we managed to plot a bar chart and we observed that the disbursement is highest in client's having Secondary/secondary special education followed by Higher education clients.

>> But the Secondary/secondary special education clients have defaulted the most followed by Higher education from our Analysis.

>> Banks should consider education of client before disbursing the loan because clients having good education have defaulted the least.

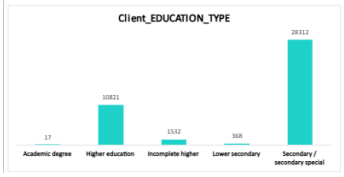
>> The SS for the same has been pasted below for the reference:

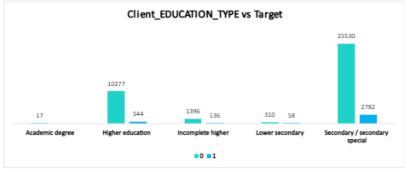
Univariate Analysis

Client_EDUCATION_TYPE	Count of Client_EDUCATION_TYPE
Academic degree	17
Higher education	10821
Incomplete higher	1532
Lower secondary	368
Secondary / secondary speci	28312

Univariate Segmented Analysis

Count of Client_EDUCATION_TYPE	TARGET(1-defaulters;0-Non_defaulters)	
Client_EDUCATION_TYPE	0	1
Academic degree	17	
Higher education	10277	544
Incomplete higher	1396	136
Lower secondary	310	58
Secondary / secondary special	25530	2782





>> Bivariate Analysis of Data(Age group and Loan amount Credited)

>> For Bivariate analysis we have considered attributes Age group of the clients and the loan amount credited.

>> We have divide the client's age in 12 Bins with 4 years are age individuals. Since 21 was the youngest and 69 being the oldest.

>> We then plotted the age group of clients against the loan amount credited. Our observation indicates the group 37-40 have the highest count of loan disbursement then any of the other age groups.

>> Followed by age group 41-44. Age group 0f 37-40 have defaulted/delayed payments higher in number when compared to age group of 41-44.

>> Another pattern is as the age increases clients with loan decreases. i.e., clients with high age don't usually apply for loans.

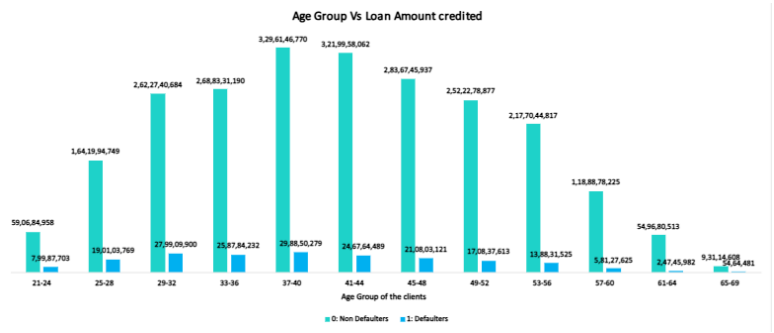
>> The age group with least defaulters are 61-64.

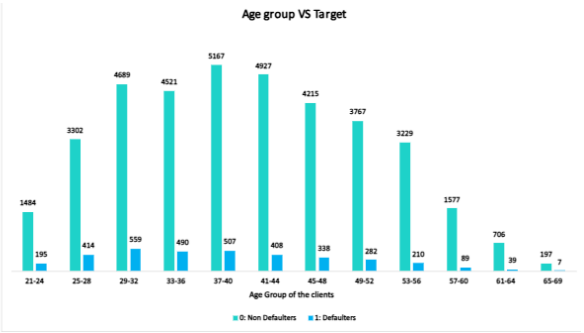
>> The bank must focus more on individuals who are having age above/beyond 37, as they are less prone to being defaulters.

>> The SS for the same has been pasted below for the reference:

Bivariate Analysis of Data(Age group and Loan amount Credited)

Age Group of Clients	Count of Age_years		Sum of Credit amt of the loan		Ratio (Non-defaulters/Defaulters)
	0	1	0	1	
21-24	1484	195	59,06,84,958	7,99,87,703	7.38
25-28	3302	414	1,64,19,94,749	19,01,03,769	8.64
29-32	4689	559	2,62,27,40,684	27,99,09,900	9.37
33-36	4521	490	2,68,83,31,190	25,87,84,232	10.39
37-40	5167	507	3,29,61,46,770	29,88,50,279	11.03
41-44	4927	408	3,21,99,58,062	24,67,64,489	13.05
45-48	4215	338	2,83,67,45,937	21,08,03,121	13.46
49-52	3767	282	2,52,22,78,877	17,08,37,613	14.76
53-56	3229	210	2,17,70,44,817	13,88,31,525	15.68
57-60	1577	89	1,18,88,78,225	5,81,27,625	20.45
61-64	706	39	54,96,80,513	2,47,45,982	22.21
65-69	197	7	9,31,14,608	54,64,481	17.04





>> Bivariate Analysis of Data (Employment in years and Loan amount Credited)

>> For Bivariate analysis we have considered attributes employment in years of the clients and the loan amount credited.

>> We have divide the client's employment time (years) in 10 Bins with 4 years as interval.

>> It can be noticed that majority of the loans are being taken by employment group of 0-4 and 5-9 years. Reasons can be many as they have recently joined the workplace and have many necessities and less liabilities along with less cash.

>> However, it can be noticed that as the years of employment increases the defaulters in loan amount decreases.

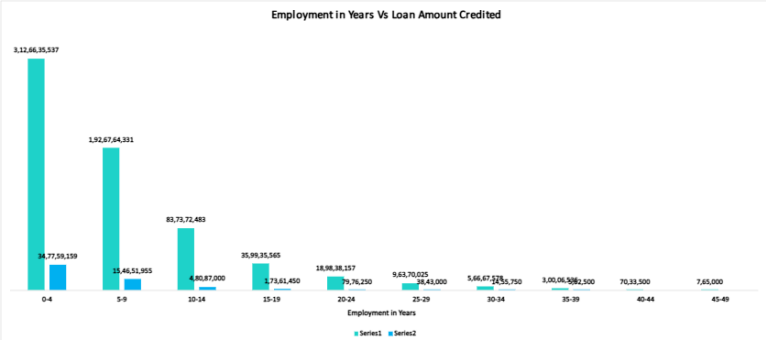
>> It is advisable to bank that it concentrates more on higher experienced individuals to disburse loan. As they tend to have least defaulters.

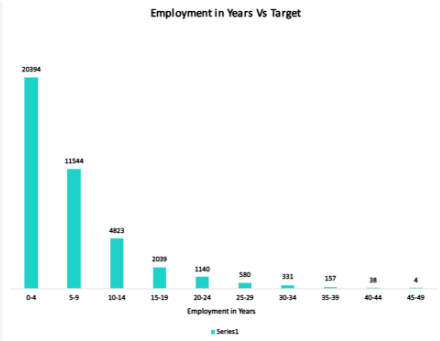
>> The SS for the same has been pasted below for the reference:

Bivariate Analysis of Data(Employment in years Vs Loan amount credited)

Employment in years	Count of TARGET
0-4	20394
5-9	11544
10-14	4823
15-19	2039
20-24	1140
25-29	580
30-34	331
35-39	157
40-44	38
45-49	4

Employment in years	loan credited(Non-Defaulters)	loan credited(Defaulters)	Ratio (Non-defaulters/Defaulters)
	0	1	
0-4	3,12,66,35,537	34,77,59,159	8.990807174
5-9	1,92,67,64,331	15,46,51,955	12.4587131
10-14	83,73,72,483	4,80,87,000	17.41369774
15-19	35,99,35,565	1,73,61,450	20.73188383
20-24	18,98,38,157	79,76,250	23.80042708
25-29	9,63,70,025	38,43,000	25.07676932
30-34	5,66,67,578	14,55,750	38.92672334
35-39	3,00,06,536	5,62,500	53.344952
40-44	70,33,500		
45-49	7,65,000		





D. Identify Top Correlations for Different Scenarios: Correlations of data based on different scenarios with Target as Attribute.

Solution:

>> This is done using the inbuilt option of Excel: Data Analysis (Correlation).

>> We first filtered data for Non-defaulter and later for defaulter.

>> We have used tributes like *income of the client, credit amount,* ***Amt_Annunity, Amt_Goods_price, region population relative, age_years, total_employment in years, client family, client's children.***

>> We simply plotted a correlation analysis with heat map(using conditional formatting)

>> **Rules of Correlation(r)**

>> -ve Correlation is inverse relation

>> +ve Correlation is proportional relation.

>> '0' is No relationship.

>> #Key highlights for Correlation of Non-Defaulters and Defaulters.

>> Credit amt of loan and Amt goods price have the highest correlation. Which simply means the higher the loan amount; equivalent or more the amount of goods is needed.

>> Followed by Credit amount of loan with Amt Annunity. I.e., Higher the loan higher the annunity amount the client has to pay.

>> Amt annunity and amt goods have also +ve strong correlation.

>> Slightly negative Correlation between Age of client along with Client_family and Client_children exists.

>> The SS for the same has been pasted below for the reference:

0(Non-defaulters)	Income of the client	Credit amt of the loan	AMT_ANNUITY(payments	AMT_GOODS_PRICE(Colateral)	REGION_POPULATION_RELATIVE	Age_years	Total_employement	Client_FAM	Client_CHILDREN
Income of the client	1.000								
Credit amt of the loan	0.360	1.000							
AMT_ANNUITY(payments at regular interval)	0.430	0.760	1.000						
AMT_GOODS_PRICE(Colateral)	0.370	0.990	0.760	1.000					
REGION_POPULATION_RELATIVE	0.180	0.100	0.110	0.100	1.000				
Age_years	0.050	0.160	0.100	0.160	0.050	1.000			
Total_employement_year	0.040	0.100	0.060	0.100	0.000	0.350	1.000		
Client_FAM_MEMBERS	0.000	0.040	0.040	0.040	-0.030	-0.170	-0.030	1.000	
Client_CHILDREN	-0.010	-0.020	0.000	-0.020	-0.030	-0.240	-0.060	0.890	1.000

1(Defaulters)	Income of the client	Credit amt of the loan	AMT_ANNUITY(payments	AMT_GOODS_PRICE(Colateral)	REGION_POPULATION_RELATIVE	Age_years	Total_employement	Client_FAM	Client_CHILDREN
Income of the client	1								
Credit amt of the loan	0.361	1							
AMT_ANNUITY(payments at regular interval)	0.430	0.760	1						
AMT_GOODS_PRICE(Colateral)	0.368	0.986	0.764	1					
REGION_POPULATION_RELATIVE	0.185	0.096	0.114	0.100	1				
Age_years	0.054	0.165	0.098	0.159	0.049	1			
Total_employement_year	0.037	0.096	0.061	0.098	-0.003	0.351	1		
Client_FAM_MEMBERS	-0.002	0.038	0.044	0.036	-0.030	-0.174	-0.033	1	
Client_CHILDREN	-0.007	-0.015	-0.002	-0.019	-0.031	-0.236	-0.064	0.894	1

==========

Insights/Results:

This project has been an invaluable source of hands-on experience, greatly boosting my confidence in using MS Excel and its functions. I've come to realize that merely watching an instructor is not enough to truly master the tool—focusing solely on theory has its limitations. The real progress happens when you put the concepts into practice. This is the key to mastering any programming language: the practical application of knowledge gained throughout the process. Moreover, this project helped me understand how MS Excel can be used to manipulate data and achieve the desired results, a crucial skill in many industries. It also strengthened my understanding of statistics and its practical use in data analysis.

==========

All Objectives outlined in our project plan have been successfully completed. We have extracted the necessary data analysis from our dataset, and the results, along with the corresponding methodologies, have been published in detail.

- Revolving loans are more prudent and non-defaulting in nature. Hence, banks must focus in distributing revolving loans among the clients.
- Male clients exhibit a higher tendency to default than female clients.
- As the age increases the defaulters of loan decreases, Banks must be careful when disbursing loan to clients having lower income level and with younger clientele.
- It is advisable to bank that it concentrates more on higher experienced (employment in years) individuals to disburse loan. As they tend to have least defaulters.
- Banks should consider education of client before disbursing the loan because clients having good education have defaulted the least whereas lower secondary/ secondary special clients have defaulted the most.

Excel sheet link (GDrive):
https://docs.google.com/spreadsheets/d/1y7MCwuePibLLnkHX56VYv1B7xRVOe7Mi/edit?usp=sharing&ouid=109807839321036950449&rtpof=true&sd=true

Project by – Sakib R. Jamadar