BANK LOAN CASE STUDY

Excel Hyperlink

PROJECT DESCRIPTION

In this project, we're working as a Data Analyst in a finance company that specializes in lending various types of loans to urban customers. Our company faces challenge some customers who don't have a sufficient credit history take advantage of this and default their loans. Our task to do EDA to analyze the patterns in the data and ensure that capable applicants are not rejected.

When a customer applies for a loan, your company faces two risks:

1. If the applicant can repay the loan but is not approved, the company loses business.
2. If the applicant cannot repay the loan and is approved, the company faces a financial loss.

We're working with the dataset contains information about loan applications. It include two types of scenarios.

1. Customers with payment difficulties: These are customers who had a late payment of more than X days on at least one of the first Y installments of the loan.
2. All other cases: These are cases where the payment was made on time.

When a customer applies for a loan, there are four possible outcomes:

1. Approved: The company has approved the loan application.
2. Cancelled: The customer cancelled the application during the approval process.
3. Refused: The company rejected the loan.
4. Unused Offer: The loan was approved but the customer did not use it.

APPROACHES

Our main main of this project to identify patterns that indicate if a customer will have difficulty paying their installments. 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.

Firstly, we downloaded our dataset and identifying the numerical and categorical values in our dataset, after this we handled the missing values and null values in our dataset and also outliers.

After this we done the EDA and feature engineering by using many excel functions, graphs, pivot table, correlation and heatmap.

We also extracted the valuable insights according to the questions.

TECH-STACK USED

I have used the MS Excel 2019 version. MS word to make report and convert into PDF by using online word to pdf converter.

INSIGHTS

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

In this question, we have to identify the missing data in the dataset and decide a suitable method to deal with it.

In our dataset, there are 50001 rows and 124 columns are present. But there are so many null values are present

between them so we have to handle to do feature engineering and EDA without any issue.

So we're using the isblank function to check the null value in our dataset and also used the count function to count the null values of our dataset we made the new rows to count the null values for each column.

After this, we are using the column_description table to know the columns and their description. After this we made a new sheet and added the columns of application_data and made a new column count of missing values and added their missing values with respect to their columns.

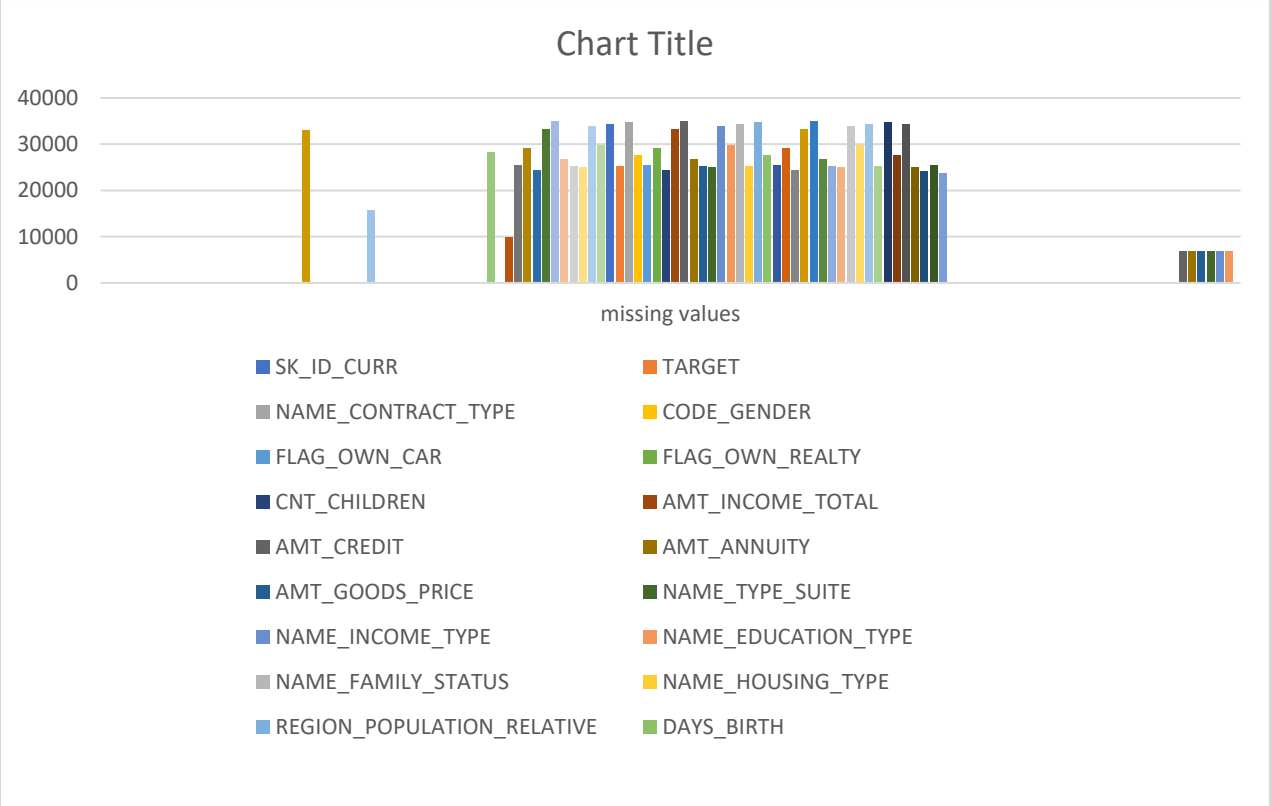
After this we made a bar graph to visualize and easy understanding the missing values with respect to their columns.

After this, we find the mean, median, mode and other values of data to analyze and know our data.

The mean is 12299.27 of our dataset, which shows the average null values of our dataset, it means averagely 12299.27 null values are present in our dataset.

The median is 168 of our dataset, which the 168 null values are present in columns.

The mode is 0 which means the majority of columns are no missing or null values in our dataset.



Descriptive Statistics	
Mean	12299.27
Standard Error	1294.54
Median	168
Mode	0
Standard Deviation	14239.95
Sample Variance	2.03E+08
Kurtosis	-1.64784
Skewness	0.439983
Range	34960
Minimum	0
Maximum	34960
Sum	1488212
Count	121
Largest(1)	34960
Smallest(1)	0

B. **Detect and identify outliers in the dataset using Excel**

statistical functions and features, focusing on numerical variables.

In this question, we have to identify the outliers in the dataset also we have to use only numerical variables because system are denied to detect the outliers of categorical columns.

Firstly, we handled all missing values and remove the all null values and also we remove the columns more that the 5% of null values because it can create problems while doing EDA and other functions. We made the new application_data sheet after removing the all null values.

We are using the new application_data to handling the outliers of our dataset.

Firstly, we chosen the numerical columns like amt_credit, amt_total_income and amt_annunity to handled the outliers.

Firstly, we are checking the outliers of amt_credit columns with respect to Target column.

We are finding the values of Q1,Q3, and Inter-Quartile to find the outliers and after this we find the upper limit and lower limit to detect the outliers.

We made the new column of outliers and using the excel function we are able to detect the outliers as True or False and shows the all outliers.

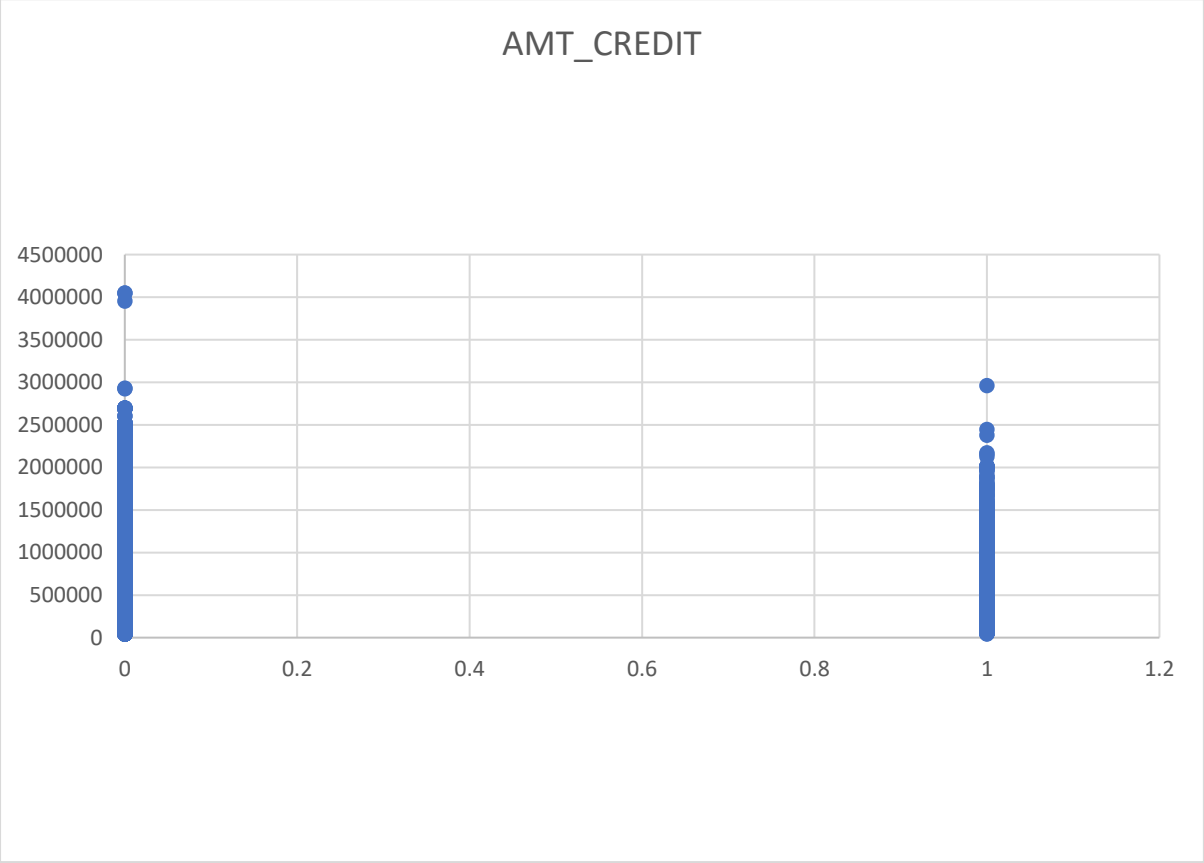
We also used the conditional formatting to make rules to easy functioning and applied the business rules to determine the outliers at valid data points.

After this, we made a scatter plot to visualize the outliers of dataset and for better understanding.

The mean values shows the averagely outliers in our dataset and median shows the middle values of outliers and mode shows the majority of outliers are present in our dataset.

We also followed the same process with another numerical columns.

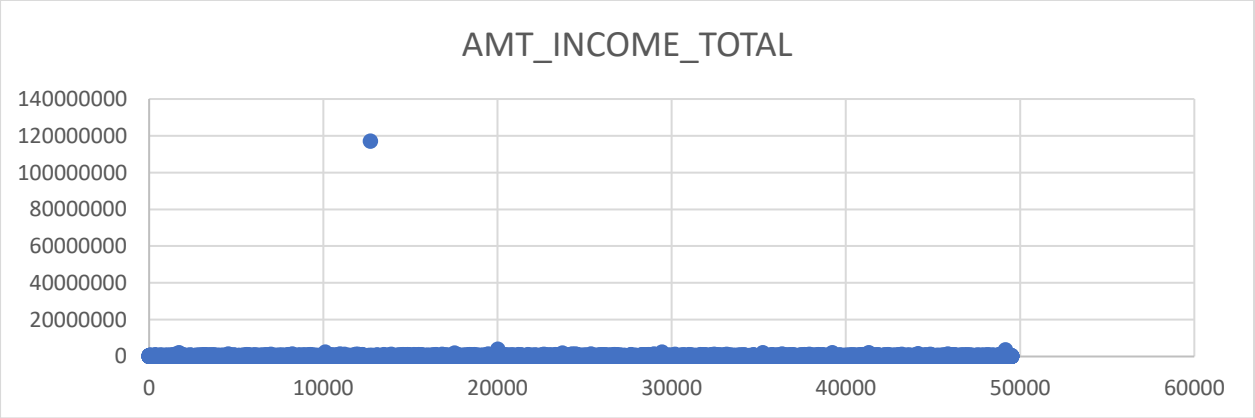
Here’s our output :



Descriptive Statistics	
Mean	600158.7578
Standard Error	1807.234198
Median	517500
Mode	450000
Standard Deviation	402132.9604
Sample Variance	1.61711E+11
Kurtosis	1.873899798
Skewness	1.214742771
Range	4005000
Minimum	45000
Maximum	4050000
Sum	29715060416
Count	49512
Largest(1)	4050000
Smallest(1)	45000

Q1	270000
Q3	808650
Inter_Quartile	538650
lower limit	-537975
upper limit	1616625

For Amt_Income :

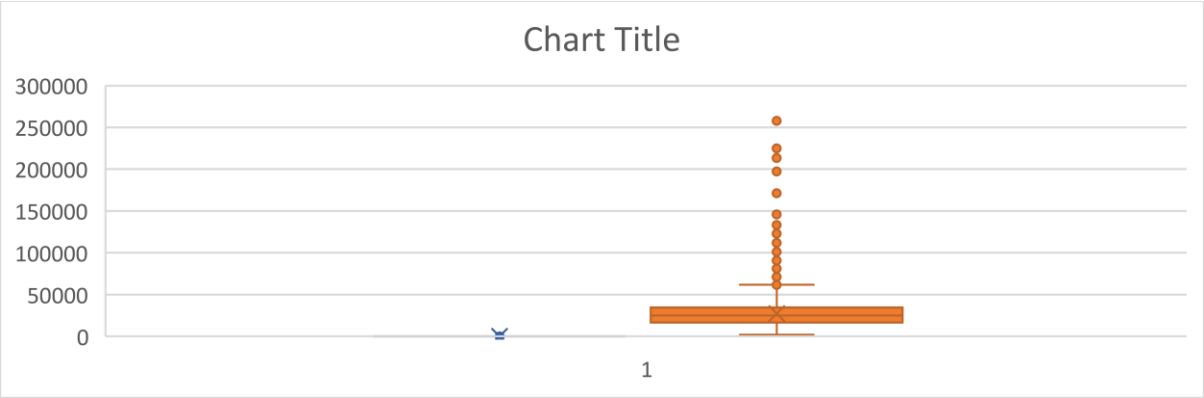


Descriptive Statistics

Mean	170638.018
Standard Error	2401.163959
Median	144000
Mode	135000
Standard Deviation	534290.0063
Sample Variance	2.85466E+11
Kurtosis	46176.78137
Skewness	211.2066458
Range	116974350
Minimum	25650
Maximum	117000000
Sum	8448629549
Count	49512
Largest(1)	117000000
Smallest(1)	25650

Q1	112500
Q3	202500
Inter_Quartile	90000
lower bound	-22500
upper bound	337500

For amt_annuity :

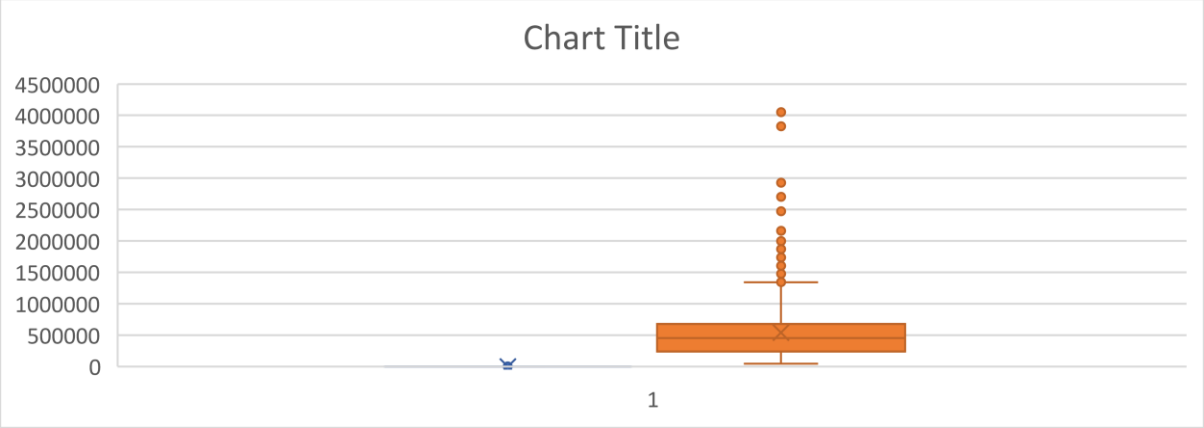


Descriptive Statistics

Mean	27139.88754
Standard Error	65.39439266
Median	24961.5
Mode	9000
Standard Deviation	14551.09733
Sample Variance	211734433.4
Kurtosis	9.497913998
Skewness	1.691483678
Range	255973.5
Minimum	2052
Maximum	258025.5
Sum	1343750112
Count	49512
Largest(1)	258025.5
Smallest(1)	2052

Q1	16521.75
Q3	34605
Inter_Quartile	18083.25
lower limit	-10603.1
upper limit	61729.88

For amt_Good_price :



Descriptive Statistics	
Mean	539151.1762
Standard Error	1660.625137
Median	450000
Mode	450000

Standard Deviation	369510.5498
Sample Variance	1.36538E+11
Kurtosis	2.431318023
Skewness	1.338238309
Range	4005000
Minimum	45000
Maximum	4050000
Sum	26694453038
Count	49512
Largest(1)	4050000
Smallest(1)	45000

Q1	238500
Q3	679500
Inter_Quartile	441000
lower limit	-423000
upper limit	1341000

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

In this question, we have to determine the data imbalance in the loan application dataset and also calculate the ratio of data imbalance because data imbalance can affect the accuracy of the analysis, especially the binary classification problem.

Firstly, we have to make a new sheet to do analysis with the target column the target has 0 and 1 value which means TARGET: Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases).

Then, we make the class and frequency table and extract the unique values from the target value which are 0 and 1.

The we used the countif function to count the number of target values with respect to class. After this we used the sum function to calculate the total number of targeted values.

			Ratio	Contribution
Target	0	45507	11.3597104343485:1	91.90919556
Frequencncy	1	4006	0.0880304129035094:1	8.090804435
	Total	**49513**		

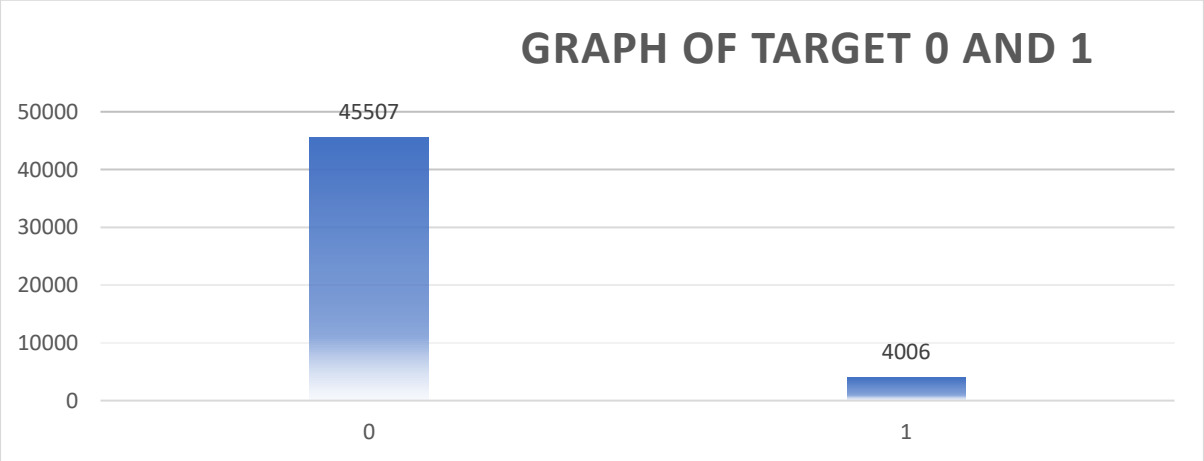
The target 0 has 45507 values and target 1 has 4006 number of values.

Then, we also find the ratio of our data imbalance using the excel function. Also we find the contribution of both values in the data imbalance.

Target 0 contribute 92% and Target 1 has 8%.

From the above values help us to know the data imbalance and we use those analysis in loan application dataset.

The 92% shows the loan repayers and 8% are defaulters.



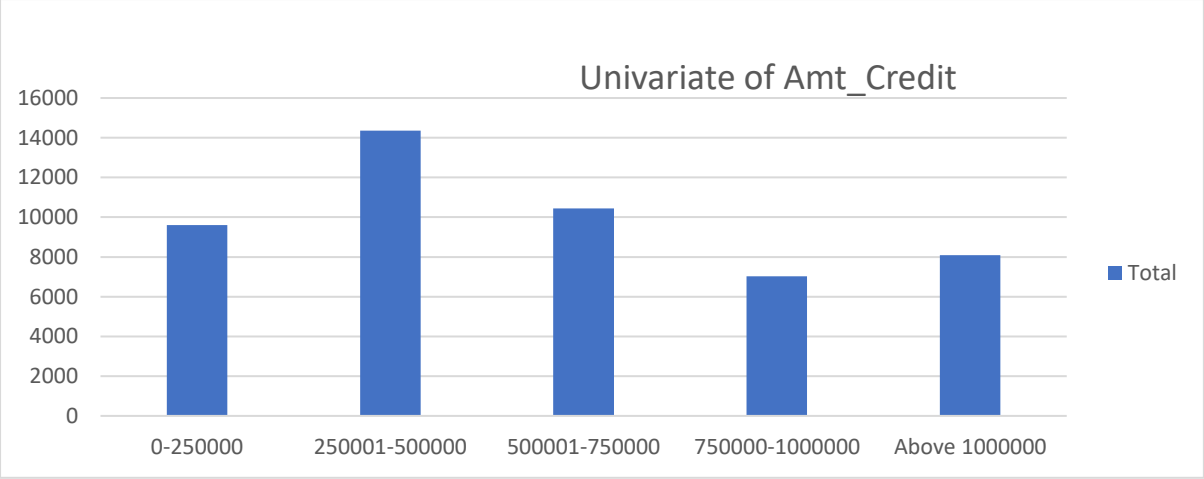
D. Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

In this question we have to perform the univariate analysis to understand the distribution of individual variables and segmented univariate also we also bivariate analysis to explore relationships between the variables and the target variables.

Firstly, we are starting from univariate analysis to explore each variable in a dataset separately. It looks at the range of values, as well as the central tendency of the values. It describes the pattern of response to the variable. It describes each variables of its own.

So, we are doing the univariate analysis of AMT_CREDIT column. We make the class and frequency by using the Pivot table. Also we make the graph to visualize the output and also for easy understanding.

The graph shows the highest number of people are take amount as credit between 250000 to 500000. And very least number of peoples take above 10 lakhs of amout as a credit.



Now, we're doing the segmented univariate with respect to amt_total_income column.

The segmented univariate analysis can be used to find summary of a single data variable in form of segments.

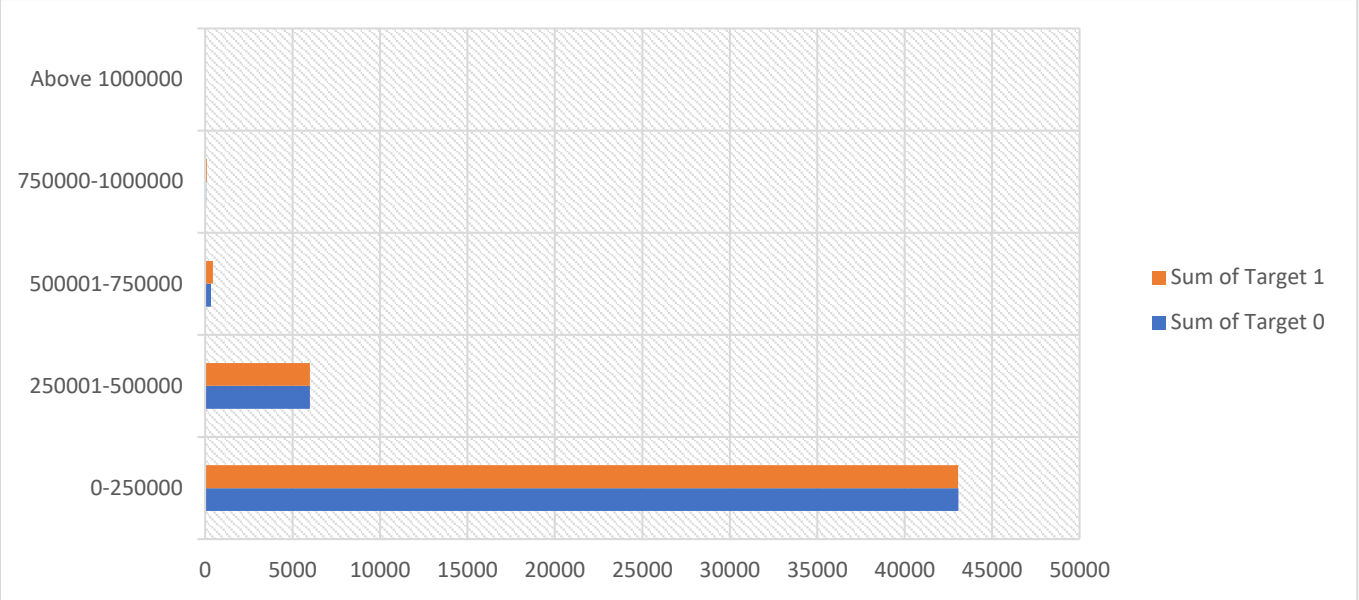
We take the amt_income column and make the class of between them and also we make the two new column sum of target 0 and target 1.

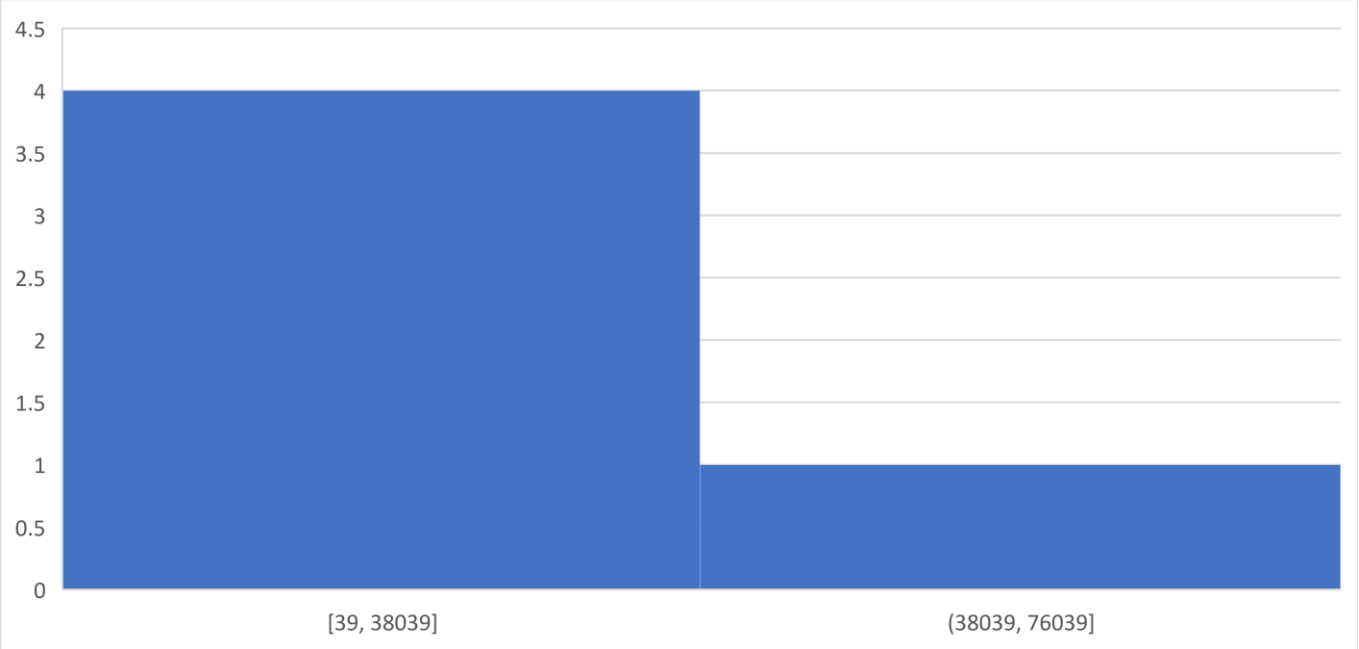
We used the pivot table and sum function to calculate the values as per required.

The graph shows the majority of peoples having income between 0 to 2.5 lakhs and very few peoples having income more than 10 lakh.

This condition is applied for both target 1 and target 2.

The histograms and other graphs shows the visualization to better understanding.





Now, we are using the bivariate analysis with respect to amt_credit column.

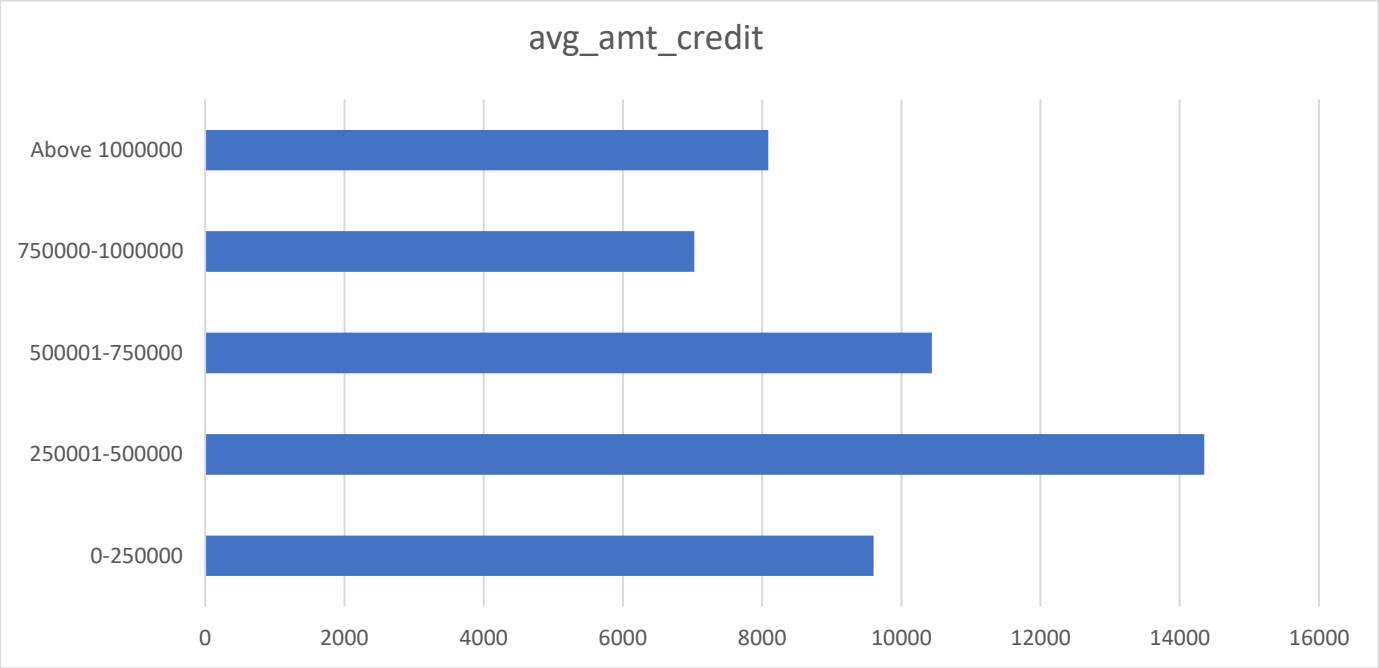
The bivariate analysis is a statistical method to examining how two different things are related. The bivariate analysis aims to determine if there is a statistical link between the two

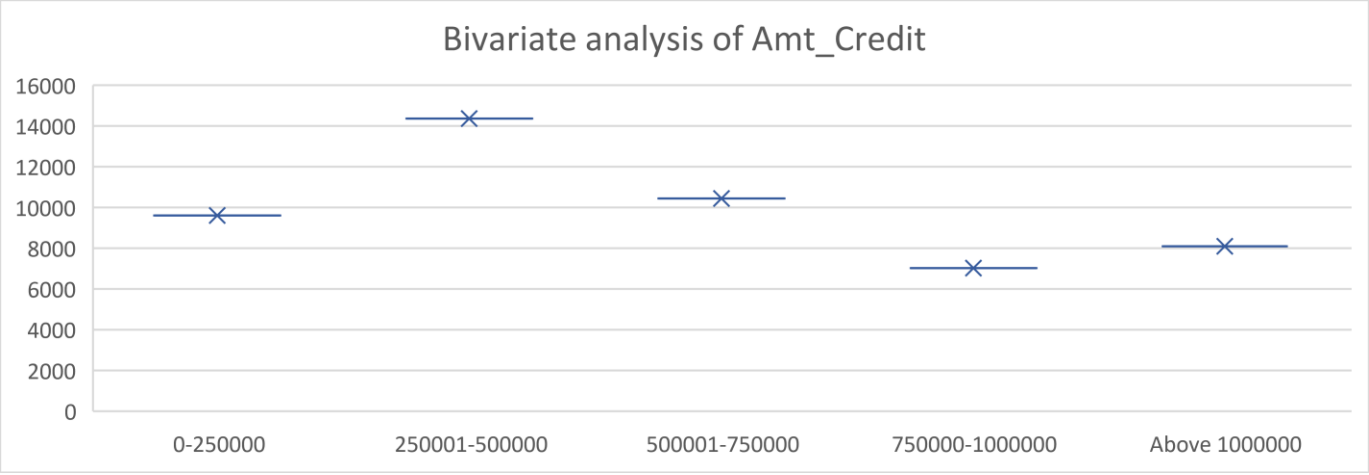
variables and if, so how strong and in which direction that link is.

We make the class and frequency table of amt_credit column and we are taking the average of their frequencies.

These average shows the peoples take average amout of those class.

Majority of peoples take credit amount between 2.5 lakh to 5 lakh and very fews peoples take credit of above 10 lakhs.





E. 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.

In this question, we have to find the segments the dataset based on different scenarios and also we have to identify the top correlations for each segmented data using the various excel functions.

Firstly, we take the various columns which we have to find the correlations between them.

The main benefits of correlation analysis are that it helps companies to determine which variables they want to investigate further, and it allows for rapid hypothesis testing.

	TARGET	*AMT_INCOME_TOTAL*	*AMT_CREDIT*	*AMT_ANNUITY*	*AMT_GOODS_PRICE*	*DAYS*
TARGET	1					
AMT_INCOME_TOTAL	0.011087	1				
AMT_CREDIT	-0.03238	0.068598	1			
AMT_ANNUITY	-0.01229	0.0822	0.768724	1		
AMT_GOODS_PRICE	-0.04116	0.069142	0.986937	0.773787	1	
DAYS_BIRTH	0.077168	0.015921	-0.06059	0.007768	-0.05909	
DAYS_EMPLOYED	-0.04043	-0.0316	-0.07013	-0.1107	-0.06748	
DAYS_REGISTRATION	0.042869	0.009786	0.002499	0.032777	0.005279	
CNT_CHILDREN	0.026497	0.009658	0.004899	0.026248	3.42E-05	

Also we used the heatmaps for those columns and Analysts use heatmaps to analyze the magnitude of an event using visual crues.

A data visualization technique, heatmaps are utilize to derive quick interpretations of the intensity of an event and do course corrections accordingly.

CONCLUSION

- From the above analysis, we could determine some key factors which helps us to analyze if the client would be defaulters or non-defaulters.
- As the age and years of experience increases, the chance of default increase. From this analysis the bank should priotize older and experience clients.
- Also we determine the amt of credit could take as per their annual income. Majority of peoples are taking credit amount between 2.5 lakh to 5 lakh.
- Very fews peoples takes credit amount more than 10 lakhs.
- We also learnt, how to identified the patterns that indicate if a customers will have difficulty paying their installments. Also this kind of 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.
- From this project, we learnt how the Data Analyst doing their work in finance company and lending the loans to the urban customers.
- I learnt how to deal with null or missing values and how to clean the data it can helps to enhance the better feature engineering.
- Also, I gained skills how to use pivot table and how to visualize by using graph, chart, scatter plot to the given output to better understanding the data.

- From this project, I gained my Advanced Excel skills and also learnt about how Data Analyst extract the valuable insights from the data.
- We learnt, how to do EDA, Feature Engineering , Excel functions as per given questions and to extract the valuable insights from them.

Linkedin Profile : https://www.linkedin.com/in/pawan-kashyap-832515230

Instagram : https://www.instagram.com/pawankkkashyap

THANK YOU