**BANK LOAN CASE STUDY**

**Project Description**:

* Project aims is to analyse the loan of bank. When the company receives a loan application, the company has to decide for loan approval on the basis of applicant profile.
* This case study attempts to demonstrate the application of EDA in real-World business. In addition to using the techniques learned in Eda module, it will help in gaining a basic of risk analytics in banking and financial services.
* The goal is to identify patterns that indicate whether a customer is likely to face difficulty in repaying their loan installments. By understanding the key factors behind loan default, the company aims to make informed decisions about loan approvals, mitigating risks, and ensuring capable applicants are not rejected.
* This analysis will help the company avoid financial losses from defaulting customers while optimizing business opportunities.

**APPROACH:**

This case study has two enormous datasets: the current and the previous applications. Each dataset includes several unneeded columns that would be useless for risk assessment and contains many blank data entries. Therefore, the first step is to clean the data by removing unnecessary columns and handling missing values. After cleaning the data, outliers were identified and addressed. Then, the analysis part begins, where univariate and bivariate analyses are performed using pivot tables and charts. Finally, we will find the correlations between variables to gain insights into their relationships.

**TECH-STACK Used:**

1. MS-Excel
2. Ms-Word

Data Understanding:

1.’**application\_data.csv’** contains all the information of the client at the time of application.

The data is about whether a client has payment difficulties.

2.’**previous\_application.csv’** contains information about the client’s previous loan data. It contains the data whether the previous application had been approved, Cancelled, Refused or Unused offer.

3,’**columns\_description.csv’** is data dictionary which describe the meaning of the variables.

1. **Identify Missing Data and Deal with it Appropriately:**

* I used COUNTA function to count the total rows in each column. Then found the percentage of null values in each column.
* After that, I remove the columns whose percentage of blank values are greater.
* It is essential to handle missing data effectively to ensure the accuracy of the analysis.
* we first cleaned the data by removing unnecessary columns and handling missing values. To identify missing data, we used the COUNT and IF functions in Excel.
* For handling missing values, we opted to impute them using the AVERAGE/Median function. If a cell was missing, we replaced it with the average of the available data in the same column.
* I kept only relevant variable to get the insight from it and converted the days into years.

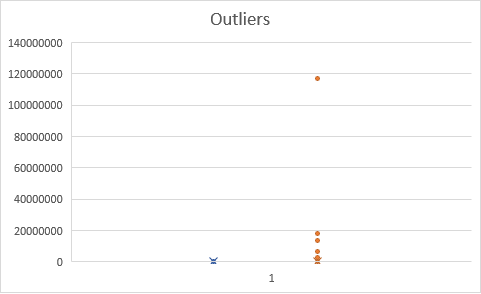
Excel functions effectively to manage missing data, allowing us to perform meaningful exploratory data analysis and draw valuable insights into loan default patterns.

1. **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.

Outliers for AMT\_INCOME\_TOTAL

In this XY plane, we can see that for the target variable 1 there are few applicants who draw an income of 11 crores whereas majority of the applicant drawing an income in lacs only.

|  |
| --- |
| Q1 |
| 112500 |
|  |
| Q3 |
| 202500 |
|  |
| IQR |
| 90000 |
|  |
| UL |
| 337500 |
|  |
| Ll |
| -22500 |

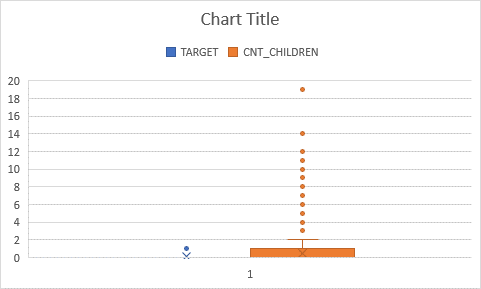


Outliers for CNT\_CHILDREN

In this XY plane, we can see that for the target variable 0 applicant having max 19 children’s which is highly unusual in these days.

Quartile 1=0, Quartile3=1, IQR=1

UL=3, Ll=-2



Outliers for DAYS\_EMPLOYED(Year)

In this plot we can see that there are few applicants in target 0 and 1 were employed for 1000 years which is impossible whereas as majority of the applicant are employed for around 80-90 years.

Q1=3 Q3=16 IQR=13 Ul=35.5 Ll=-16.5

DATA IMBALANCE:

Data imbalanced sheet show the ratio of total application with payment difficulties., target 1 to the total applicant with payment made on time and target 0 is in ratio of 11.39 times.

Out of total applicant of 307511, the percentage of applicants which makes payment on time is 92% thus makes a majority class whereas the percentage of applicant of applicant with payment difficulties is 8% thus makes a minority class.

UNIVARIATE ANALYSIS:

Univariate Analysis refers to the analysis of the data contain only one variable. It doesn’t deal with the causes or relationship and the main purpose of the analysis is to describe the data and find pattern exist in it.

* This graph is an example of univariate analysis which shows the count of every application (0 and 1) of the columns AMT\_CREDIT grouped in different income bins. Majority of the applicant got loan approval of credit range 9 lacs and above.
* Those with 0 to 7 years of work experience are the most likely to seek for loans.
* Those who are married have taken out more loans.
* More loans have been requested by working people.

SEGMENTED UNIVARIATE ANALYSIS:

Segmented Univariate analysis refers to data contains only one variable. Here segmented analysis refers to analysis in subset of variable.

This belove graph is an example of segmented analysis that shows most of the applicants(0 and 1) drawing an income between 1 lacs – 2.25 lacs.

Very few targets 1 applicants are drawing an income 5 lacs and above. Thus, which can be the reason of difficulty in payment.

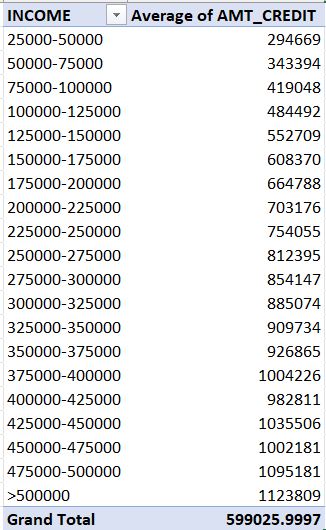
|  |  |  |  |
| --- | --- | --- | --- |
| **A** | **Target** |  |  |
| **Income Bin** | **0** | **1** | **Grand Total** |
| 25000-50000 | 4174 | 343 | 4517 |
| 50000-75000 | 17849 | 1526 | 19375 |
| 75000-100000 | 36450 | 3356 | 39806 |
| 100000-125000 | 39860 | 3841 | 43701 |
| 125000-150000 | 43837 | 4053 | 47890 |
| 150000-175000 | 31685 | 2978 | 34663 |
| 175000-200000 | 27190 | 2454 | 29644 |
| 200000-225000 | 18464 | 1635 | 20099 |
| 225000-250000 | 25945 | 2093 | 28038 |
| 250000-275000 | 11846 | 887 | 12733 |
| 275000-300000 | 4000 | 306 | 4306 |
| 300000-325000 | 6342 | 410 | 6752 |
| 325000-350000 | 1987 | 135 | 2122 |
| 350000-375000 | 4282 | 255 | 4537 |
| 375000-400000 | 1180 | 85 | 1265 |
| 400000-425000 | 1696 | 115 | 1811 |
| 425000-450000 | 640 | 38 | 678 |
| 450000-475000 | 2407 | 153 | 2560 |
| >500000 | 2852 | 162 | 3014 |
| **Grand Total** | **282686** | **24825** | **307511** |

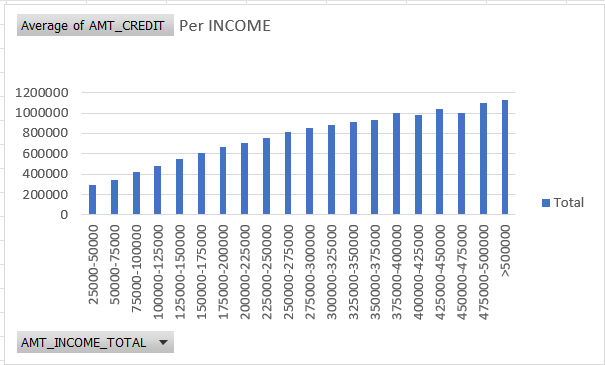
BIVARIATE ANALYSIS:

Bivariate analysis refers to the data contains two different variables. It deals with the causes and relationship among two variables.

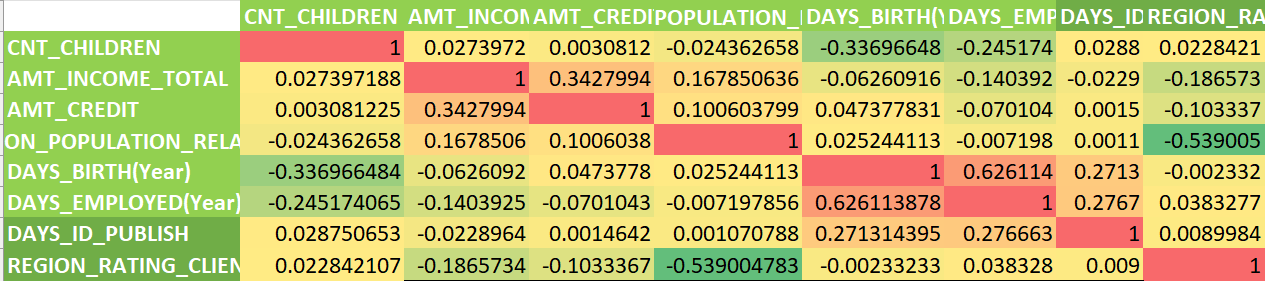
The above graph shows the relationship between applicant and different income bins which are directly proportional to each other.

Thus, income increases the amount credit will also increases.





**Identify Top Correlations for Different Scenarios:**

****

The above heat map shows the correlation between the different variables for target 0 i.e., applicant with payment made on time.

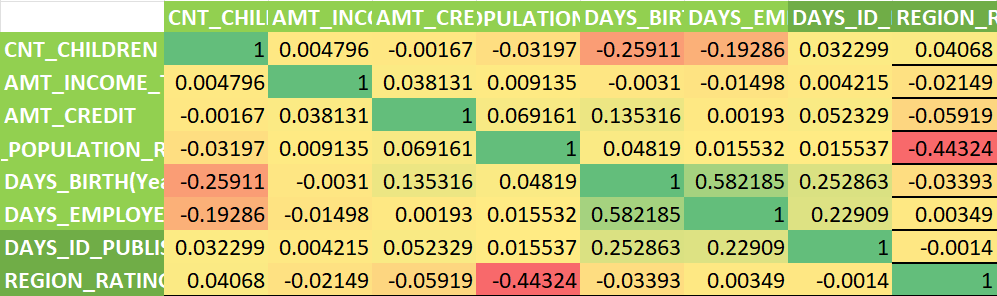
Colors used in heat are red, yellow and green where color green indicates the strongest correlation among variables.

Thus, most relevant correlations among are

DAYS\_EMPLOYED(Year) to DAYS\_BIRTH(Year)

DAYS\_ID\_PUBLISH to DAYS\_BIRTH(Year)

DAYS\_ID\_PUBLISH to DAYS\_EMPLOYED(Year)



The above heat map shows the correlation between the different variables for target 1 i.e., applicant with payment not made on time.

Colors used in heat are red, yellow and green where color green indicates the strongest correlation among variables.

Thus, most relevant correlations among are

DAYS\_EMPLOYED(Year) to DAYS\_BIRTH(Year)

DAYS\_ID\_PUBLISH to DAYS\_BIRTH(Year)

DAYS\_ID\_PUBLISH to DAYS\_EMPLOYED(Year)

**Result:**

Most Applicants drawing the higher income were offered the higher credit amount.

Majority of the applicants got loan approval of credit range 9 lacs and above.

Most of the applicants (0 and 1) drawing an income between 1 lac – 2.25 lacs.

[application\_data work.xlsx](application_data%20work.xlsx)

<https://docs.google.com/spreadsheets/d/1ILA5Bd-4LPEMOAWuy9pr_8nZyaA7gzm7/edit?usp=sharing&ouid=106154228007115112439&rtpof=true&sd=true>