TRAINITY

**PROJECT – 6**

**BANK LOAN CASE STUDY**

Submitted By

**BOJJA JAHNAVI**

Data Analytics Trainee

Submitted On:- 03-12-2023

**BANK LOAN CASE STUDY**

**Project Description:**

This Project is analyzed to find the issues occurring to sanction the loans for candidates. The main objective of the project is to draw the insights from the given data to represent whether a person has difficulty in approval of the loan.

Company gets two types of risks when a customer applies for a loan. They are

1. The candidate can repay the loan but it is not approved
2. The candidate cannot repay the loan but it is approved

When a customer applies for a loan there are total 4 possible outcomes

1. Approved – loan is approved by the company
2. Refused – loan is rejected by the company
3. Cancelled – The loan is cancelled by the applicant
4. Unused offer – Loan is approved but candidate is not using the offer

These are the 4 outcomes when a candidate applies for a loan.

The dataset provided for this project contains the applicant details. It contains payment difficulties like who cannot pay and takes extra X days in the initial Y installments.

As a data analyst I should find the data imbalances, correlation, requirements to approve the loan etc.

**Approach:**

The project contains three types of datasets namely application\_data, previous\_application, columns\_description. Download the datasets and execute the functions to derive the solutions.

**A. Identify Missing Data and Deal with it Appropriately:**The given dataset contains large data with numerous columns. The unnecessary as well as blank columns must be removed. Blank cells are found by using the functions like countblank, count. Columns with more than 30% blank cells must be deleted and those with less than 30% are remained in the anlaysis.

**Function:**

=(countblank(A:A)/count(A:A))\*100

**Output:**

These are some of the columns that are deleted which have more than 30% blank cells.

Housetype\_mode, livingarea\_avg, wallsmaterial\_mode, elevations\_avg, nonlivingarea\_avg, ext\_source\_1, floorsmin\_avg, own\_car\_age, commonarea\_avg etc.

**B. Identify Outliers in the Dataset:** There are few outliers present in the dataset. These outliers leads to impact on the results which provide wrong analysis. Therefore excel functions like quartile, conditional formatting are used to identify the outliers. I have taken the following variables that contains the numerical values.

**Funtion:**

=QUARTILE(RANGE,QUART)

**For AMT\_INCOME\_TOTAL**

=QUARTILE(H:H,1)

=QUARTILE(H:H,3)

=QUARTILE(H:H,4)

**For AMT\_CREDIT**

=QUARTILE(I:I,1)

=QUAERTILE(I:I,3)

=QUARTILE(I:I,4)

**FOR AMT\_ANNUITY**

=QUARTILE(J:J,1)

=QUARTILE(J:J,3)

=QUARTILE(J:J,4)

**FOR AMT\_GOODS\_PRICE**

=QUARTILE(K:K,1)

=QUARTILE(K:K,3)

=QUARTILE(K:K,4)

**Output:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variables** | **Quartile 1** | **Quartile 3** | **Max value** | **IQR** |
| AMT\_INCOME\_TOTAL | 112500 | 202500 | 117000000 | 90000 |
| AMT\_CREDIT | 270000 | 808650 | 4050000 | 538650 |
| AMT\_ANNUITY | 16524 | 34596 | 258025.5 | 18072 |
| AMT\_GOODS\_PRICE | 238500 | 679500 | 4050000 | 441000 |
| CNT\_FAM\_MEMBERS | 2 | 3 | 20 | 1 |

From the given table we can see the outliers for every quartile. The maximum values of the variables are considered as the outliers which are different from the remaining values.

**C. Analyze Data Imbalance:** Data Imbalance affects the accuracy of data. By this we can find the details of the imbalanced data by comparing the frequencies of every class using countif function. Here I got the information of the people who have payment difficulties and who paid the installments on time from the target variable. Likewise I found the imabalance of name\_contract\_type, cnt\_children, Year\_birth, Name\_education\_type etc.

**Function:**

For target

=countif(B:B,BX14)

=countif(B:B,BX15)

=sum(BY14+BY15)

=percentage(BY14/BY16)\*100

**Output:**

|  |  |
| --- | --- |
| **Target** |  |
| 0 | 282686 |
| 1 | 24825 |
| Total sum | 307511 |
| Percentage of installments paid on time (0) | 91.92712 |
| Percentage of payment difficulties (1) | 8.072882 |

From the table there are 92% applicants present who can pay the installments on time without delay. Whereas 8% applicants have payment difficulties.

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

In the given dataset we need to find and compare the variables to analyze the data and establish relationship between them.

**Univariate Analysis:**

Univariate Analysis refers to finding the relation of the single variable.

**Function:**

I used Pivot Table to analyze the relationship in the Amt\_Income\_type, Amt\_credit.

**Output:**

For Amt\_Income\_Type

|  |  |
| --- | --- |
| **Row Labels** | **Count of NAME\_INCOME\_TYPE** |
| 25000-275000 | 280466 |
| 275000-525000 | 24399 |
| 525000-775000 | 2062 |
| 775000-1025000 | 337 |
| 1025000-1275000 | 101 |
| 1275000-1525000 | 62 |
| 1525000-1775000 | 21 |
| 1775000-2025000 | 19 |
| 2025000-2275000 | 24 |
| 2275000-2525000 | 1 |
| 2525000-2775000 | 1 |
| 2775000-3025000 | 1 |
| 3025000-3275000 | 2 |
| 3275000-3525000 | 2 |
| 3525000-3775000 | 2 |
| 3775000-4025000 | 2 |
| 4275000-4525000 | 4 |
| 6525000-6775000 | 1 |
| 8775000-9025000 | 1 |
| 13275000-13525000 | 1 |
| 17775000-18025000 | 1 |
| 116775000-117025000 | 1 |
| **Grand Total** | **307511** |

There are more candidates whose income is between 25000 - 275000

For Amt\_credit

|  |  |
| --- | --- |
| **Row Labels** | **Count of AMT\_CREDIT** |
| 45000-1045000 | 264037 |
| 1045000-2045000 | 42028 |
| 2045000-3045000 | 1415 |
| 3045000-4045000 | 23 |
| 4045000-5045000 | 8 |
| **Grand Total** | **307511** |

Majority of the applicants are approved for the loan credit between 45000 – 1045000.

**Univariate Segmented Analysis:**

Univariate Segmented Analysis refers to finding the analysis of single varible that is having subsets.

**Function:**

For Target and Contract\_name\_type

Used pivot tables

**Output:**

|  |  |
| --- | --- |
| **Row Labels** | **Count of TARGET** |
| 0 | 282686 |
| 1 | 24825 |
| **Grand Total** | **307511** |

There are more applicants who can pay installments on time.

|  |  |
| --- | --- |
| **Row Labels** | **Count of NAME\_CONTRACT\_TYPE** |
| Cash loans | 278232 |
| Revolving loans | 29279 |
| **Grand Total** | **307511** |

Cash loans are provided more than revolving loans.

**Bivariate Analysis:**

The relationship established between two variables is called bivariate analysis.

**Function:**

Used pivot tables

**Output:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Count of TARGET** | **Column Labels** |  |  |
| **Row Labels** | **0** | **1** | **Grand Total** |
| 25000-275000 | 257300 | 23166 | 280466 |
| 275000-525000 | 22885 | 1514 | 24399 |
| 525000-775000 | 1949 | 113 | 2062 |
| 775000-1025000 | 318 | 19 | 337 |
| 1025000-1275000 | 97 | 4 | 101 |
| 1275000-1525000 | 58 | 4 | 62 |
| 1525000-1775000 | 20 | 1 | 21 |
| 1775000-2025000 | 17 | 2 | 19 |
| 2025000-2275000 | 24 |  | 24 |
| 2275000-2525000 | 1 |  | 1 |
| 2525000-2775000 | 1 |  | 1 |
| 2775000-3025000 | 1 |  | 1 |
| 3025000-3275000 | 1 | 1 | 2 |
| 3275000-3525000 | 2 |  | 2 |
| 3525000-3775000 | 2 |  | 2 |
| 3775000-4025000 | 2 |  | 2 |
| 4275000-4525000 | 4 |  | 4 |
| 6525000-6775000 | 1 |  | 1 |
| 8775000-9025000 | 1 |  | 1 |
| 13275000-13525000 | 1 |  | 1 |
| 17775000-18025000 | 1 |  | 1 |
| 116775000-117025000 |  | 1 | 1 |
| **Grand Total** | **282686** | **24825** | **307511** |

Majority of the applicants have income range of 25000 - 27500

**E. Identify Top Correlations for Different Scenarios**: correlations of different variables are found to know the relation between them about the payment difficulties and other cases.

**Function :**

Used Correlation formula in Data Analysis which is in the Data Menu.

**Output:**

Correlation for payments made on time:

Relevant correlations are seen between Amt\_total\_income to amt\_credit, Days\_employed to days\_birth.

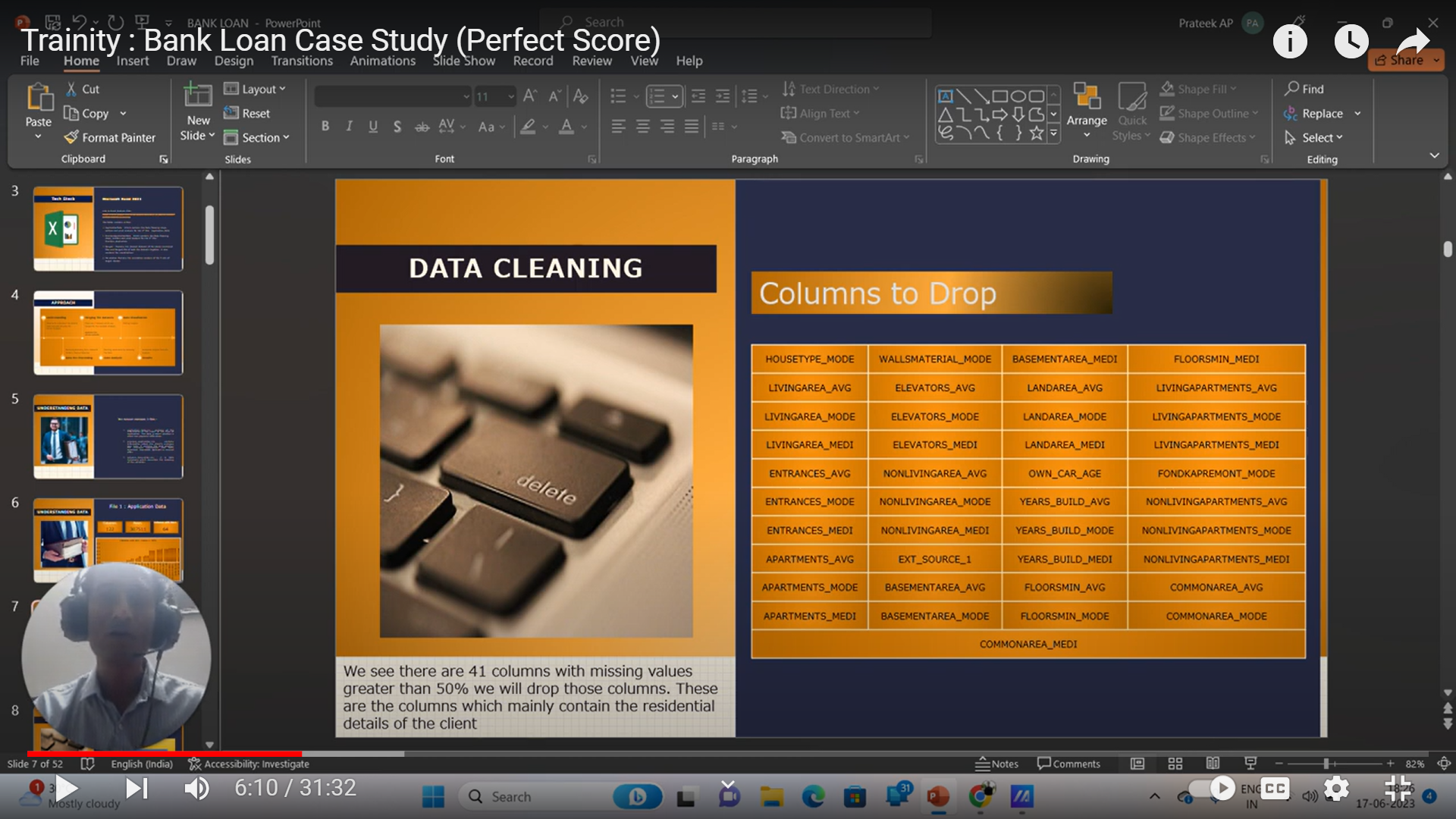
Correlation for other cases:

Relevant correlations are seen between Amt\_total\_income to amt\_credit, Days\_employed to days\_birth.

**Tech-Stack Used:** The Project “Bank Loan Case Study” is done in Microsoft Excel 2021. This is very simple to learn and handle for beginners. It has every function to solve a question and derive the answer. Therfore it is flexible to use and is free of cost.

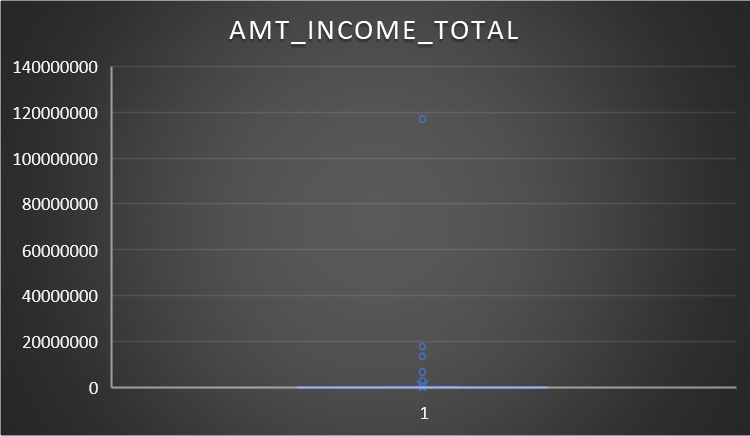
**Insights:**

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

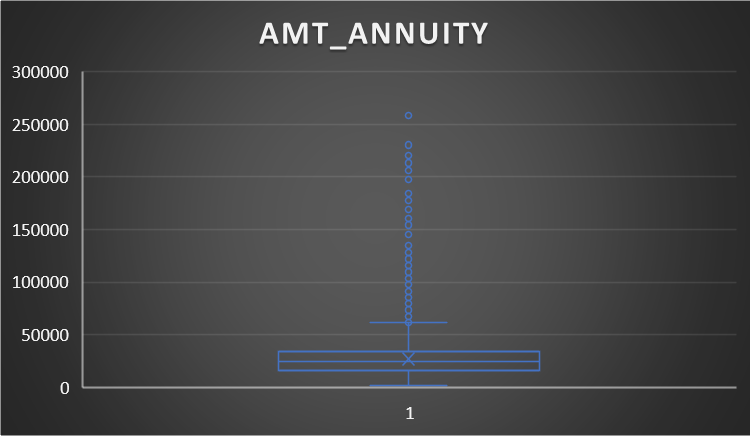
****

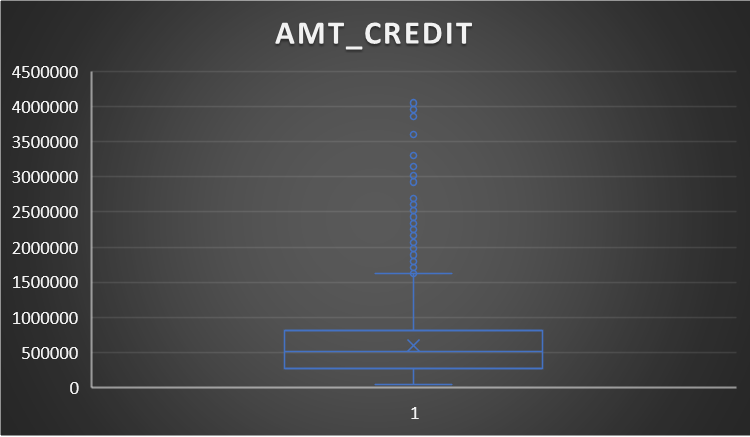
These are the columns to be dropped while cleaning the data.

1. **Identify Outliers in the Dataset:**

****

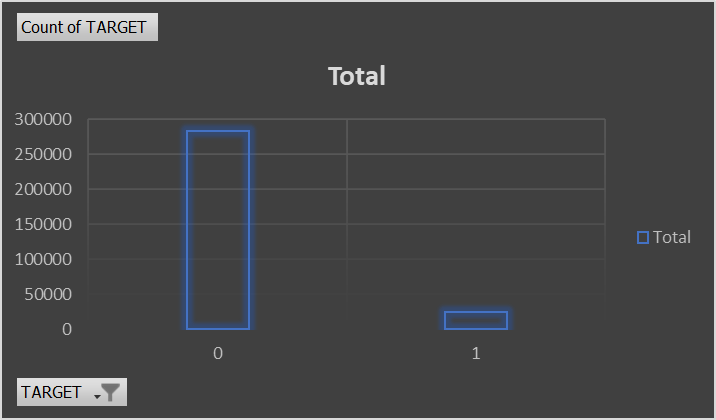
For target 1 the maximum income of the applicant is 11 crores.

****

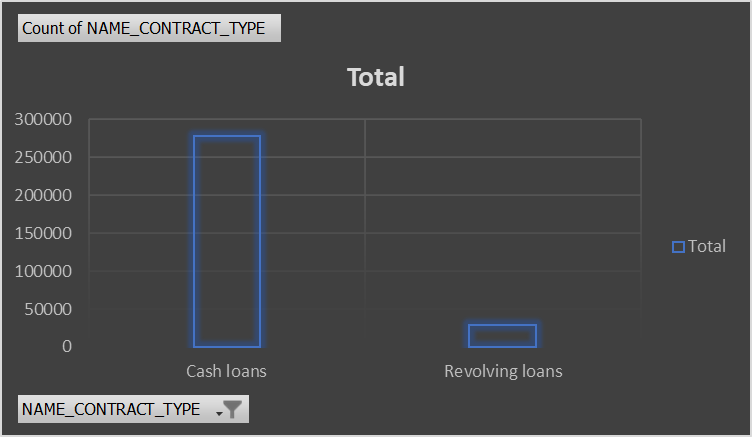
****

The amount credited by bank loan is above 4000000.

1. **Analyze Data Imbalance:**



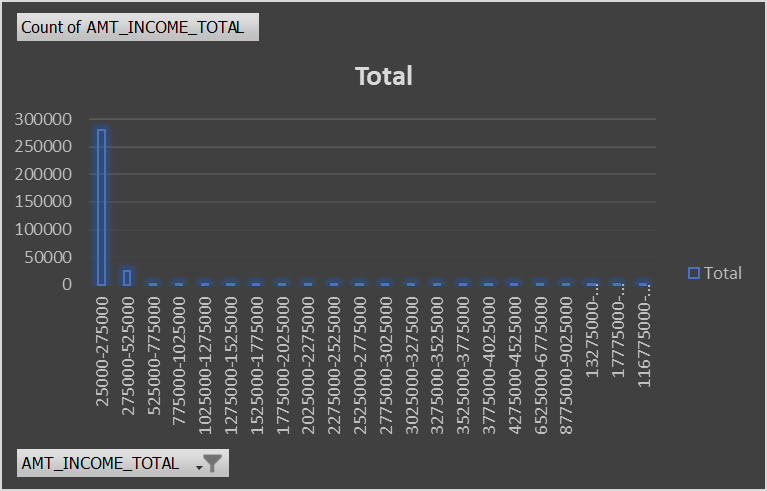
Candidates who can pay the loans on time are more than those who delay.



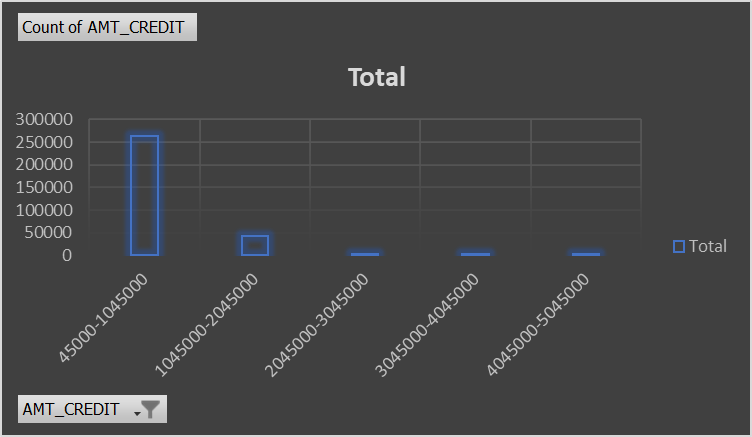
Cash loans are more than revolving loans.

1. **Perform Univariate, Segmented Univariate, and Bivariate Analysis:**

**Univariate Analysis:**

****

There are more candidates under the income ranging between 25000 – 27500



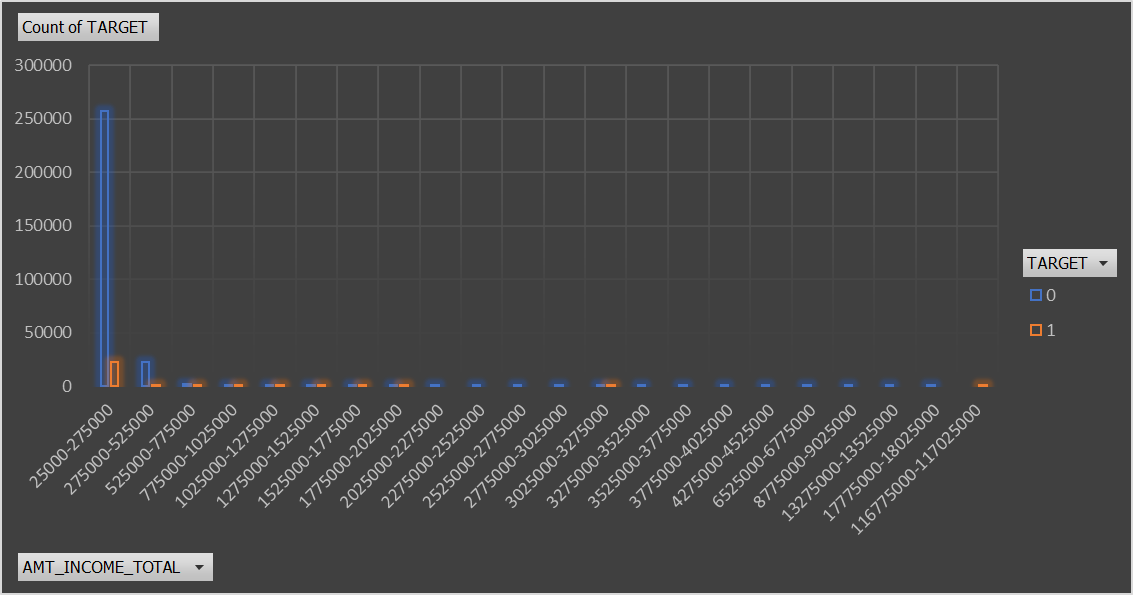
Majority of the applicants are credited with a loan of 45000 – 1045000

**Univariate Segmented Analysis:**

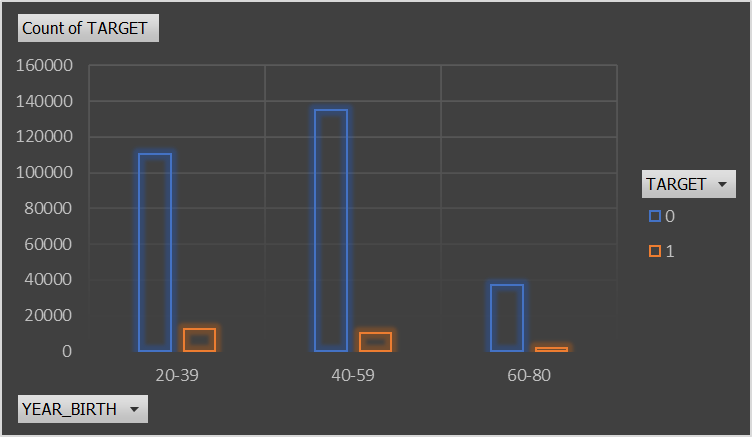
****

92% of the loans are cash loans while remaining 8% loans are revolving loans.

**Bivariate Analysis:**

****

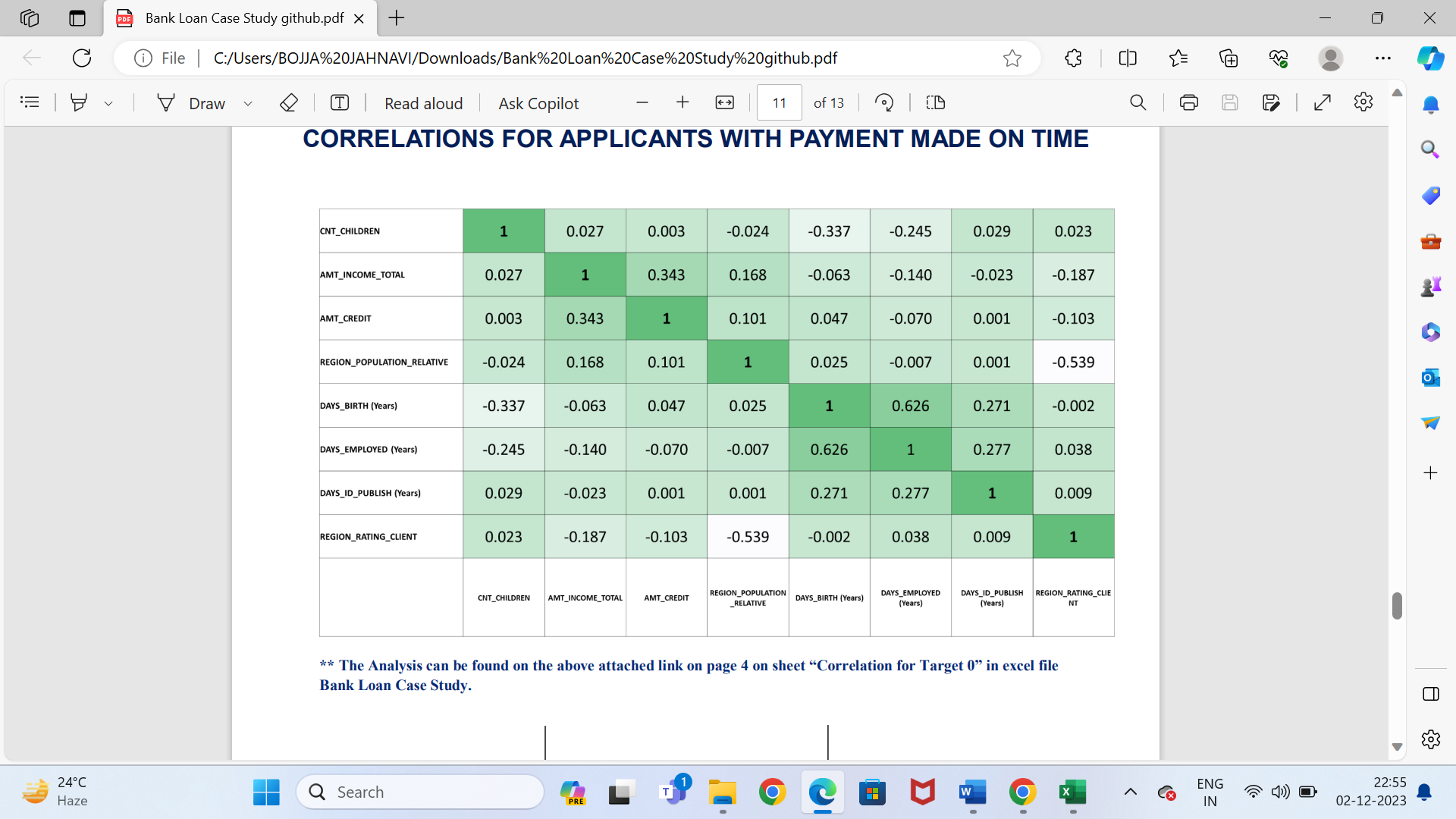
**Here we have two variables Amt\_income\_total and Target.**

****

Applicants belonging to the age group 40 – 59 are able to pay installments within time.

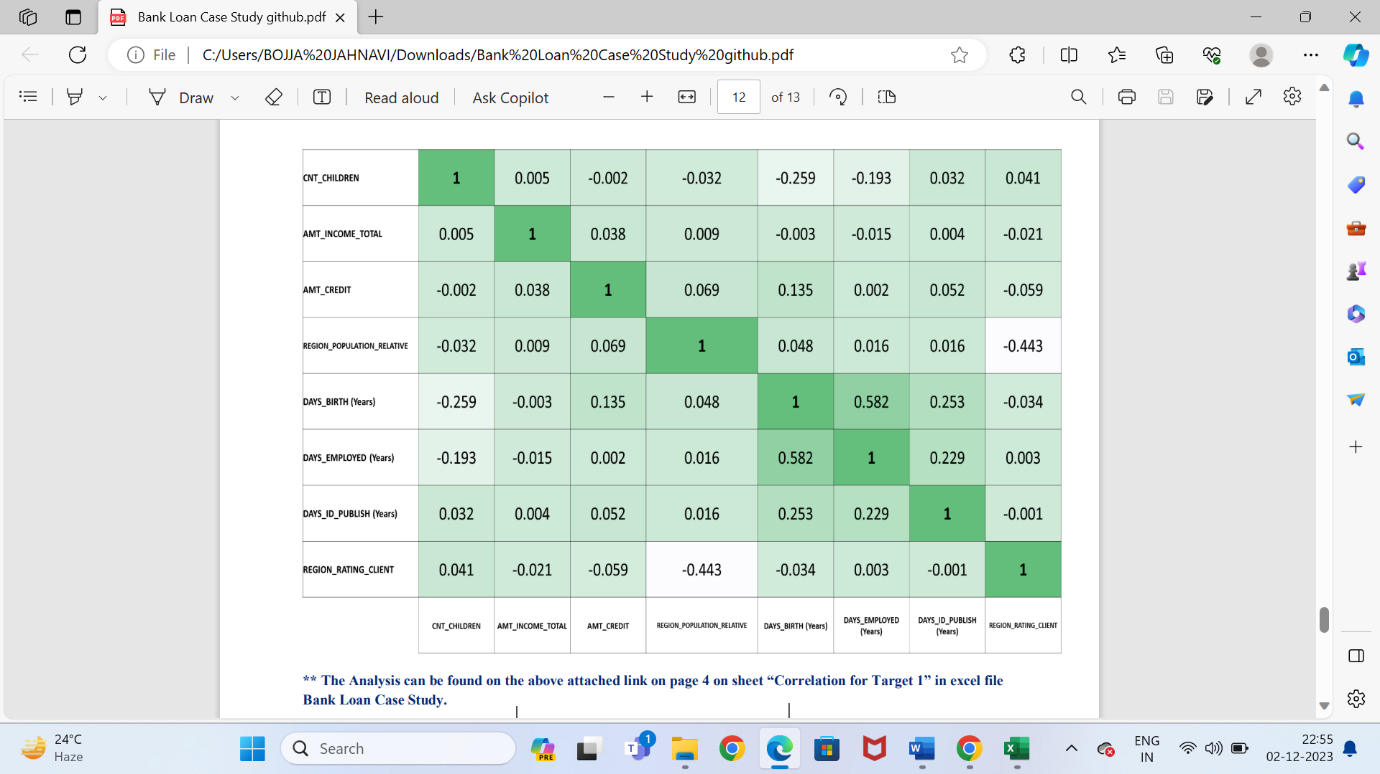
1. **Identify Top Correlations for Different Scenarios:**

For installments paid on time:

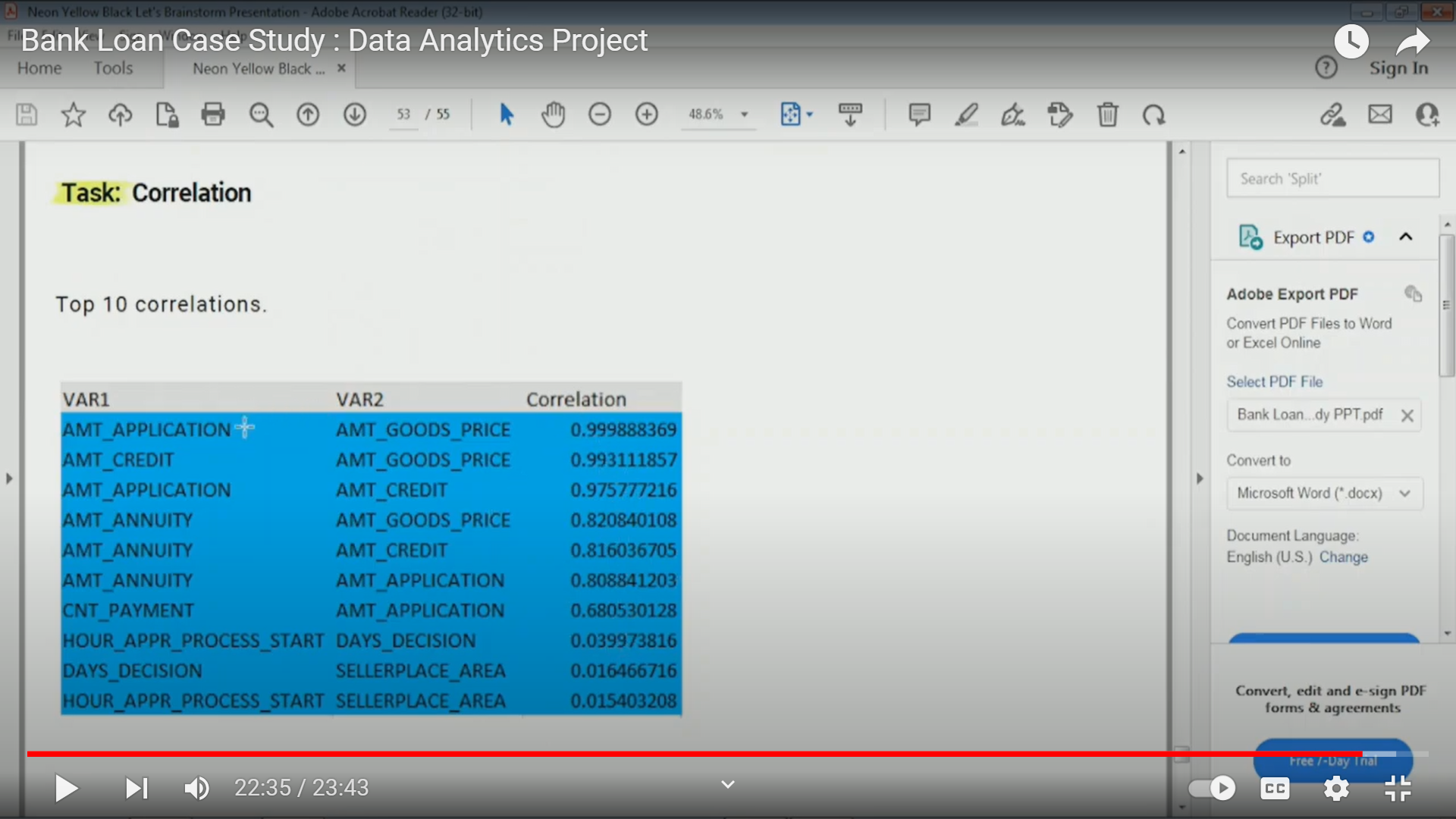
****

Relevant correlations are seen between Amt\_total\_income to amt\_credit, Days\_employed to days\_birth.

For Delayed Payments:



These are the top 10 correlations of the variables



**Result:** From the project “Bank Loan Case Study” I have learnt the advanced excel functions. Firstly there is a huge data which must be cleaned by removing the null values. There are few reasons analyzed in which the loan is approved. The project is done by using the exploratory data. Most of the applicants have the income between 125000 to 175000. Correlations made to analyze the large datasets. So, the bank issues loans based on the income, age etc.

**My Excel Work sheet:**

[**https://docs.google.com/spreadsheets/d/1rT7LhEaSAGjeM2dJnvj1iAuFChGUBnK8/edit?usp=sharing&ouid=101880124803050791429&rtpof=true&sd=true**](https://docs.google.com/spreadsheets/d/1rT7LhEaSAGjeM2dJnvj1iAuFChGUBnK8/edit?usp=sharing&ouid=101880124803050791429&rtpof=true&sd=true)

**Video Presentation:**

[**https://drive.google.com/file/d/1ai0KjXTHJFnch4yjyepM9oxohb-Xphi8/view?usp=sharing**](https://drive.google.com/file/d/1ai0KjXTHJFnch4yjyepM9oxohb-Xphi8/view?usp=sharing)