# Bank Loan Case Study

:Priyanshu Kamal
:callmepriyanshu4@gmail.com

# Project Description

This project is a case study of the process through which banks give loan to its customers. What are the different methods used for minimizing company's loss which includes Risk Analytics also and to understand the driving factor behind loan default. The project is majorly to understand how Exploratory Data Analysis (EDA) is used in real life scenario by different institutions, in this case by Banks which involves doing EDA of very large datasets.

# Approach

The approach of this project is to first clean the data by finding null values and dropping columns and replaced the missing data making the data in relevant form to make it ready for analysis, and then use Exploratory Data Analysis (EDA) steps to analyze the pattern present in the data.

I also Changed the duration from days to years by dividing it by 365 for easier understanding and analysis.

Used Quartiles to find out the outliers in the data.

# Tech-Stack Used

I have used Microsoft Excel 365 to do the analysis of the given data.

I have used Microsoft Excel because it is a spreadsheet developed for various platforms. It has calculation or computation capabilities with graphing tools, pivot tables, and  also has a macro programming language called Visual Basic for Applications which would make doing the analysis and visualisation effortless.

Hyperlink of excel sheet:
https://docs.google.com/spreadsheets/d/1mo0pI6IlzeGJP95vr1mnw4V6G_S-5Qe9/edit?usp=sharing&ouid=115516897268360068412&rtpof=true&sd=true

# Data cleaning

The very first step for data cleaning is to check the percentage of blank cells in each column which is done by COUNTA function which counts the number of nonblank cells and the finding out its percentage. If the blank cell percentage of a column is equal to or greater than 30% then we drop the column.

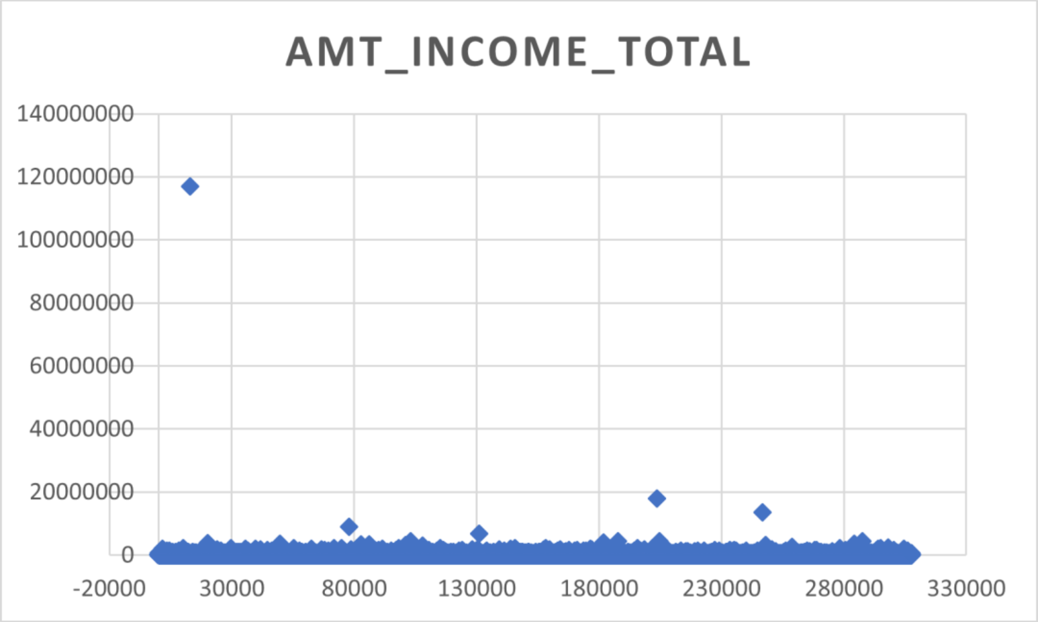
For those cells which has Blank cells percentage less than 30% I filled the missing data using distribution Statistics i.e., mean, median, mode to the relevant cells.

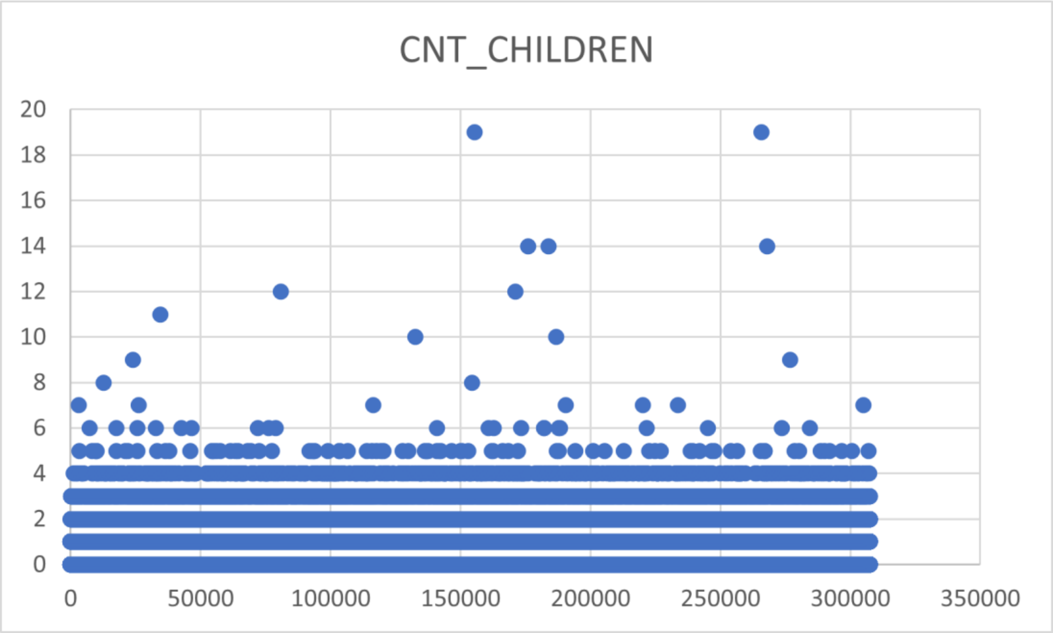
# Outliers

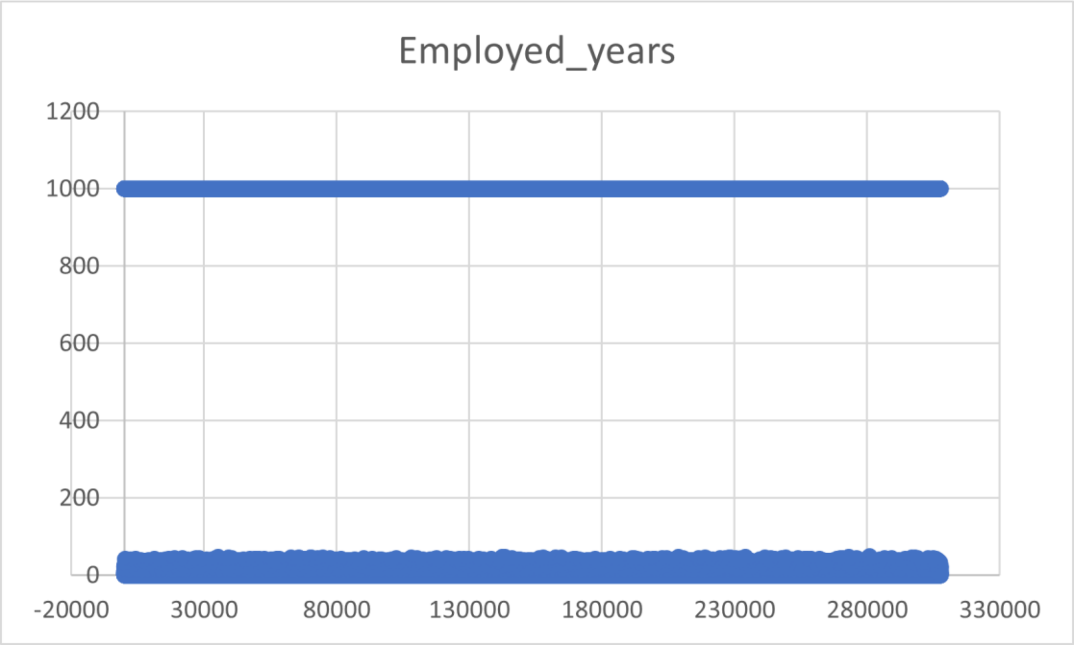
Outlier is a data point that significantly deviates from the overall pattern or distribution of the rest of the data. It is an observation that lies an abnormal distance away from other values in a dataset. It can affect the insights that we derive from data analysis.

Outliers can be done through various ways; I’ve used Quartile method visual method.

This data had a lot of Outliers but I’ve included the highly deviated ones in this presentation.


AMT_INCOME_TOTAL
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
-20000
30000
80000
130000
180000
230000
280000
330000

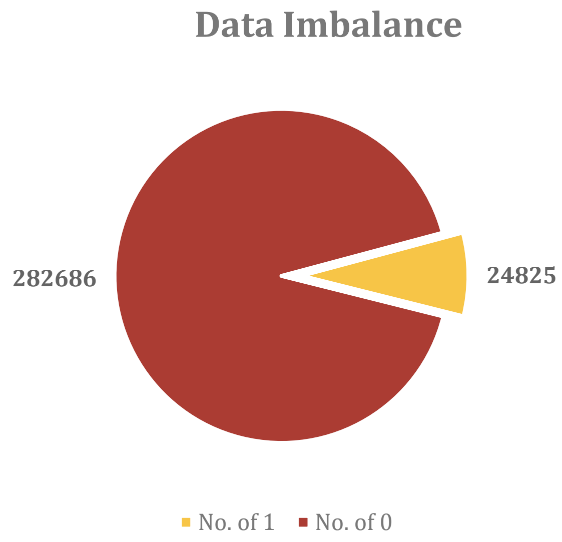

CNT_CHILDREN
20
18
16
14
12
10
8
6
4
2
0
0
50000
100000
150000
200000
250000
300000
350000


Employed_years
1200
1000
800
600
400
200
0
-20000
30000
80000
130000
180000
230000
280000
330000

# Data Imbalance

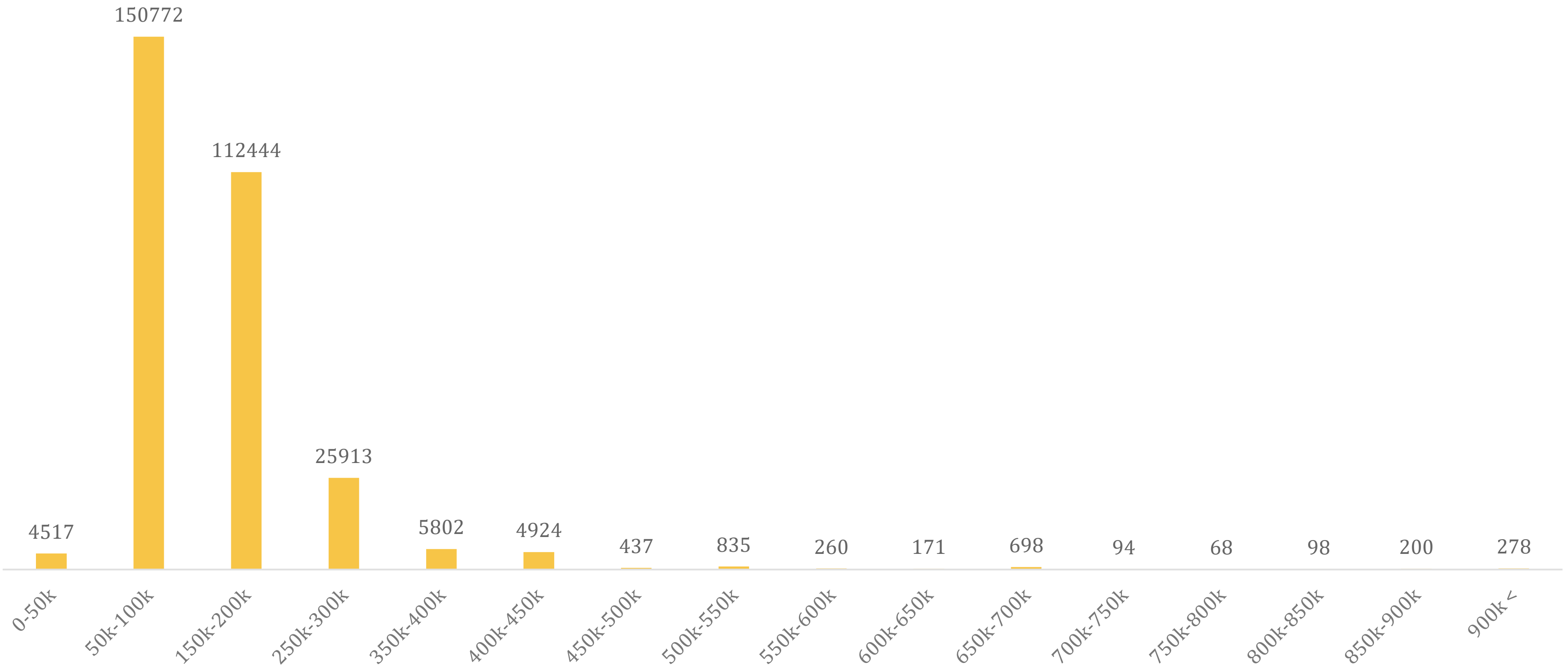
Data Imbalance is the situation where one or more class has significantly more numbers of instances compared to other class.

**This data is highly imbalanced** considering the number of applicants with difficulty and no difficulty having a **ratio of 11.39**



# Univariate Analysis

Univariate Analysis is the analysis of data in which only one variable is used for deriving insights and patterns from the data.

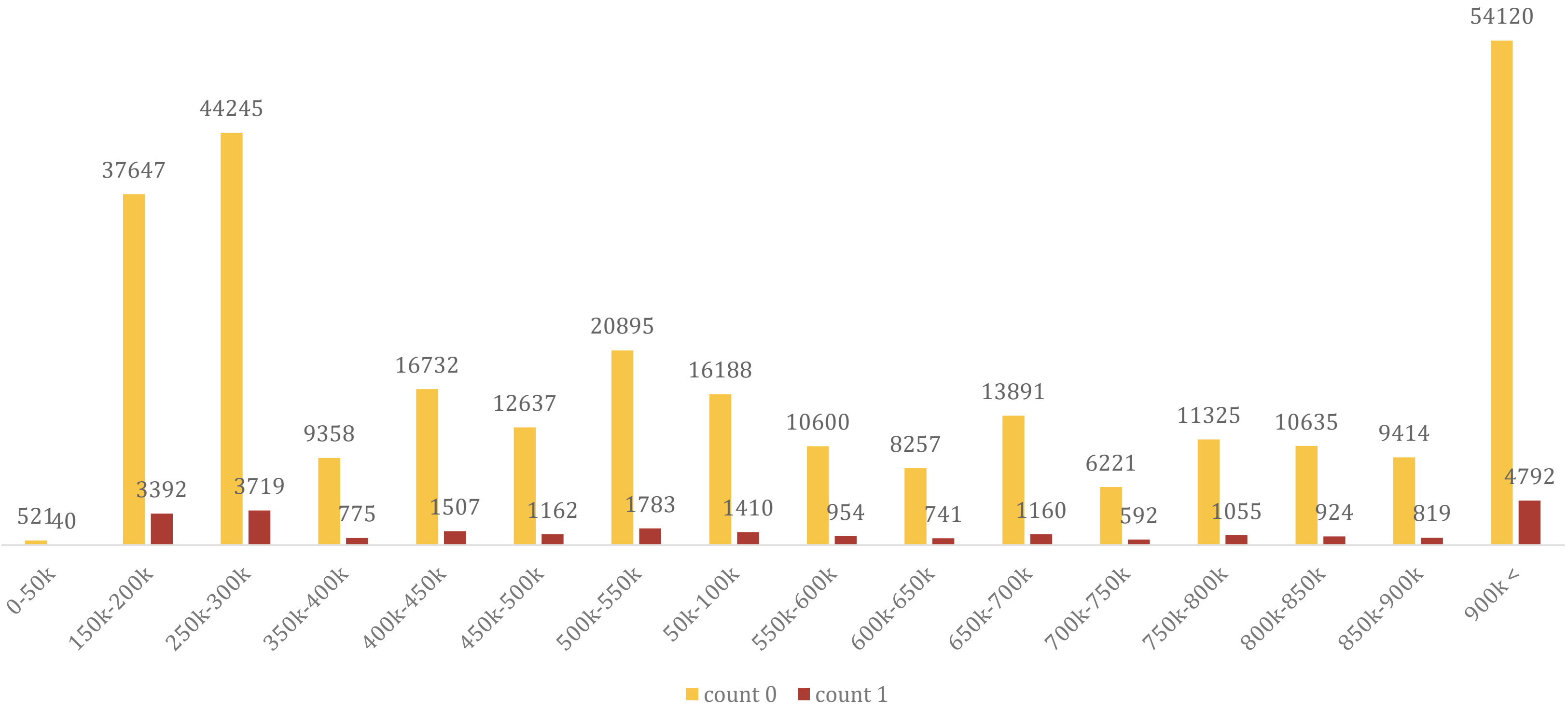


No of application in every income range

# Segmented Univariate Analysis

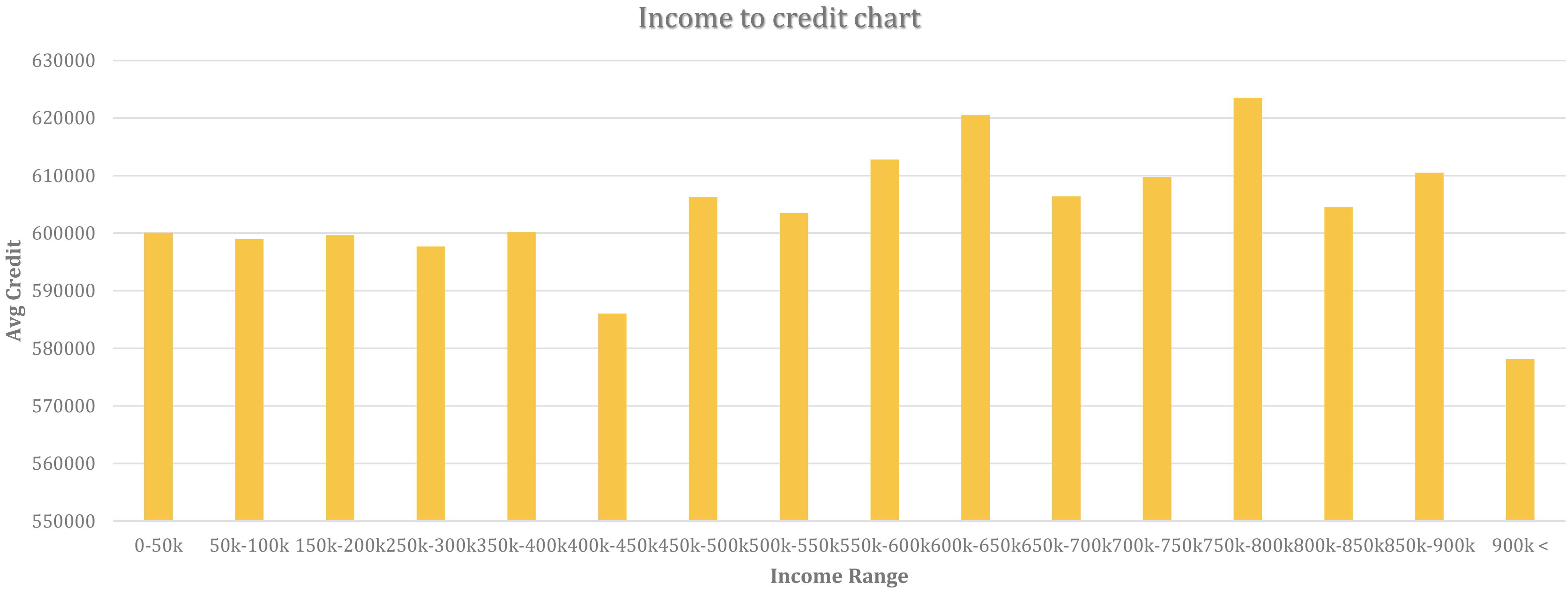
Segmented Univariate Analysis is a type of Univariate Analysis in which only one variable is used to draw the insights but they are of different class or groups.

credit range with respective difficulty



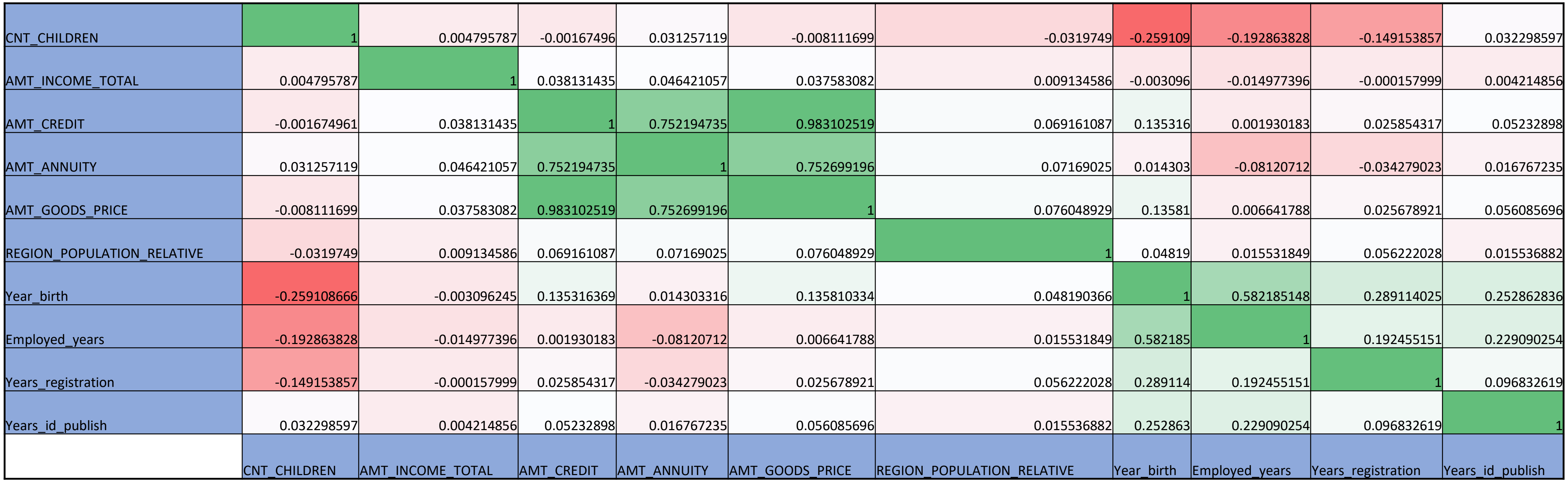
# Bivariate Analysis

Bivariate Analysis is the analysis of data in which two variable is used to analyze or derive the insights from it.



# Correlation

For Applicant with **payment difficulty**. I have used CORREL function and then conditional formatting.

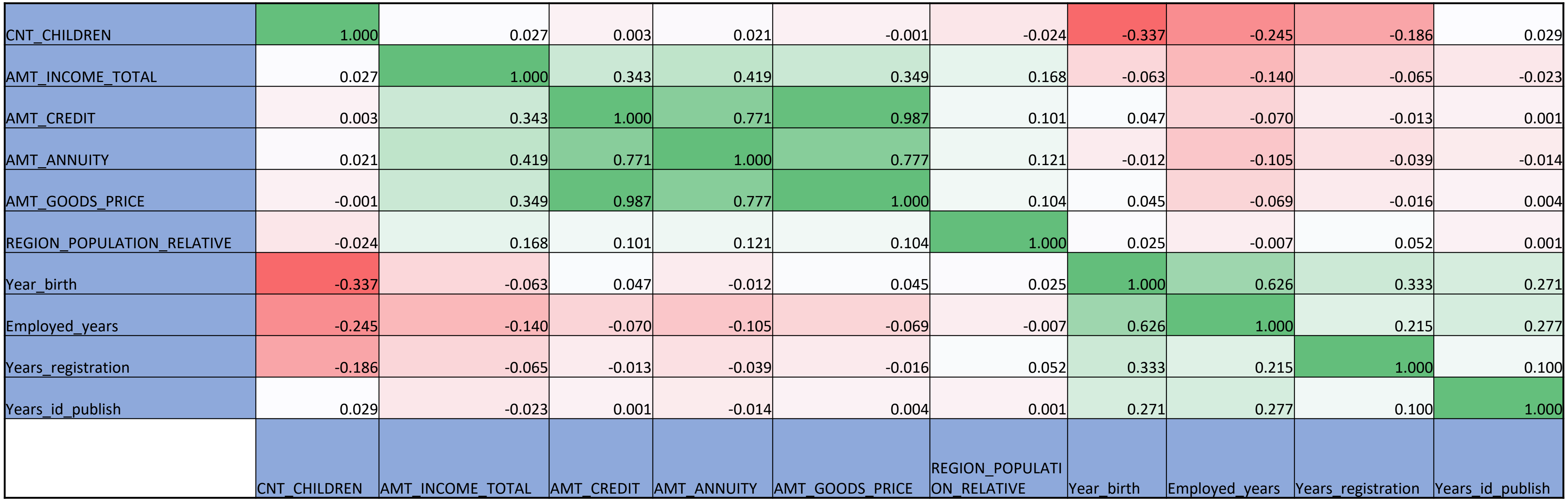


| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| CNT_CHILDREN | 1 | 0.004795787 | -0.00167496 | 0.031257119 | -0.008111699 | -0.0319749 | -0.259109 | -0.192863828 | -0.149153857 | 0.032298597 |
| AMT_INCOME_TOTAL | 0.004795787 | 1 | 0.038131435 | 0.046421057 | 0.037583082 | 0.009134586 | -0.003096 | -0.014977396 | -0.000157999 | 0.004214856 |
| AMT_CREDIT | -0.001674961 | 0.038131435 | 1 | 0.752194735 | 0.983102519 | 0.069161087 | 0.135316 | 0.001930183 | 0.025854317 | 0.05232898 |
| AMT_ANNUITY | 0.031257119 | 0.046421057 | 0.752194735 | 1 | 0.752699196 | 0.07169025 | 0.014303 | -0.08120712 | -0.034279023 | 0.016767235 |
| AMT_GOODS_PRICE | -0.008111699 | 0.037583082 | 0.983102519 | 0.752699196 | 1 | 0.076048929 | 0.13581 | 0.006641788 | 0.025678921 | 0.056085696 |
| REGION_POPULATION_RELATIVE | -0.0319749 | 0.009134586 | 0.069161087 | 0.07169025 | 0.076048929 | 1 | 0.04819 | 0.015531849 | 0.056222028 | 0.015536882 |
| Year_birth | -0.259108666 | -0.003096245 | 0.135316369 | 0.014303316 | 0.135810334 | 0.048190366 | 1 | 0.582185148 | 0.289114025 | 0.252862836 |
| Employed_years | -0.192863828 | -0.014977396 | 0.001930183 | -0.08120712 | 0.006641788 | 0.015531849 | 0.582185 | 1 | 0.192455151 | 0.229090254 |
| Years_registration | -0.149153857 | -0.000157999 | 0.025854317 | -0.034279023 | 0.025678921 | 0.056222028 | 0.289114 | 0.192455151 | 1 | 0.096832619 |
| Years_id_publish | 0.032298597 | 0.004214856 | 0.05232898 | 0.016767235 | 0.056085696 | 0.015536882 | 0.252863 | 0.229090254 | 0.096832619 | 1 |
| | CNT_CHILDREN | AMT_INCOME_TOTAL | AMT_CREDIT | AMT_ANNUITY | AMT_GOODS_PRICE | REGION_POPULATION_RELATIVE | Year_birth | Employed_years | Years_registration | Years_id_publish |

Amount credited has the best correlation with goods price and cnt_children has worst correlation with Year_birth (age).

# Correlation

For Applicant with no payment difficulty. I have used CORREL function and then conditional formatting.

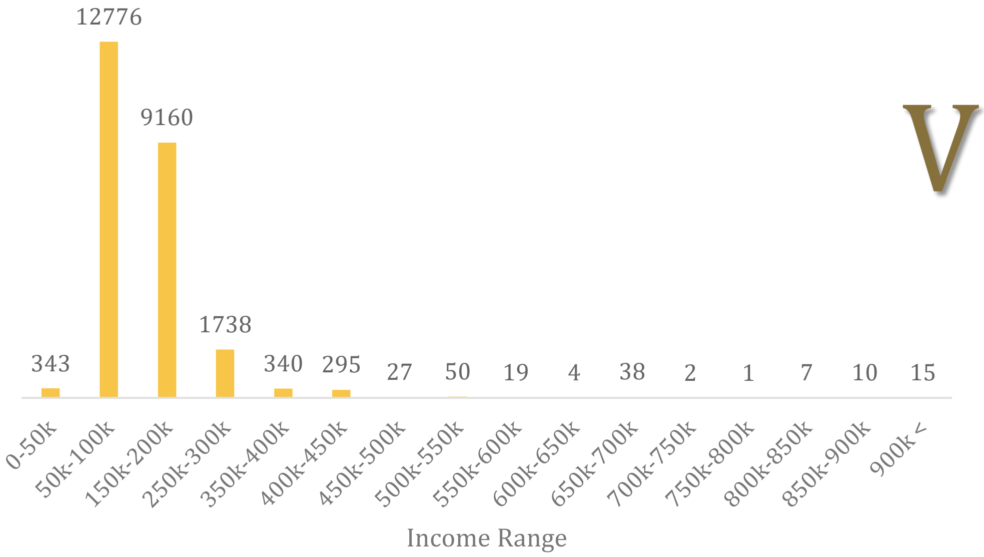


| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| CNT_CHILDREN | 1.000 | 0.027 | 0.003 | 0.021 | -0.001 | -0.024 | -0.337 | -0.245 | -0.186 | 0.029 |
| AMT_INCOME_TOTAL | 0.027 | 1.000 | 0.343 | 0.419 | 0.349 | 0.168 | -0.063 | -0.140 | -0.065 | -0.023 |
| AMT_CREDIT | 0.003 | 0.343 | 1.000 | 0.771 | 0.987 | 0.101 | 0.047 | -0.070 | -0.013 | 0.001 |
| AMT_ANNUITY | 0.021 | 0.419 | 0.771 | 1.000 | 0.777 | 0.121 | -0.012 | -0.105 | -0.039 | -0.014 |
| AMT_GOODS_PRICE | -0.001 | 0.349 | 0.987 | 0.777 | 1.000 | 0.104 | 0.045 | -0.069 | -0.016 | 0.004 |
| REGION_POPULATION_RELATIVE | -0.024 | 0.168 | 0.101 | 0.121 | 0.104 | 1.000 | 0.025 | -0.007 | 0.052 | 0.001 |
| Year_birth | -0.337 | -0.063 | 0.047 | -0.012 | 0.045 | 0.025 | 1.000 | 0.626 | 0.333 | 0.271 |
| Employed_years | -0.245 | -0.140 | -0.070 | -0.105 | -0.069 | -0.007 | 0.626 | 1.000 | 0.215 | 0.277 |
| Years_registration | -0.186 | -0.065 | -0.013 | -0.039 | -0.016 | 0.052 | 0.333 | 0.215 | 1.000 | 0.100 |
| Years_id_publish | 0.029 | -0.023 | 0.001 | -0.014 | 0.004 | 0.001 | 0.271 | 0.277 | 0.100 | 1.000 |
| | CNT_CHILDREN | AMT_INCOME_TOTAL | AMT_CREDIT | AMT_ANNUITY | AMT_GOODS_PRICE | REGION_POPULATION_RELATIVE | Year_birth | Employed_years | Years_registration | Years_id_publish |

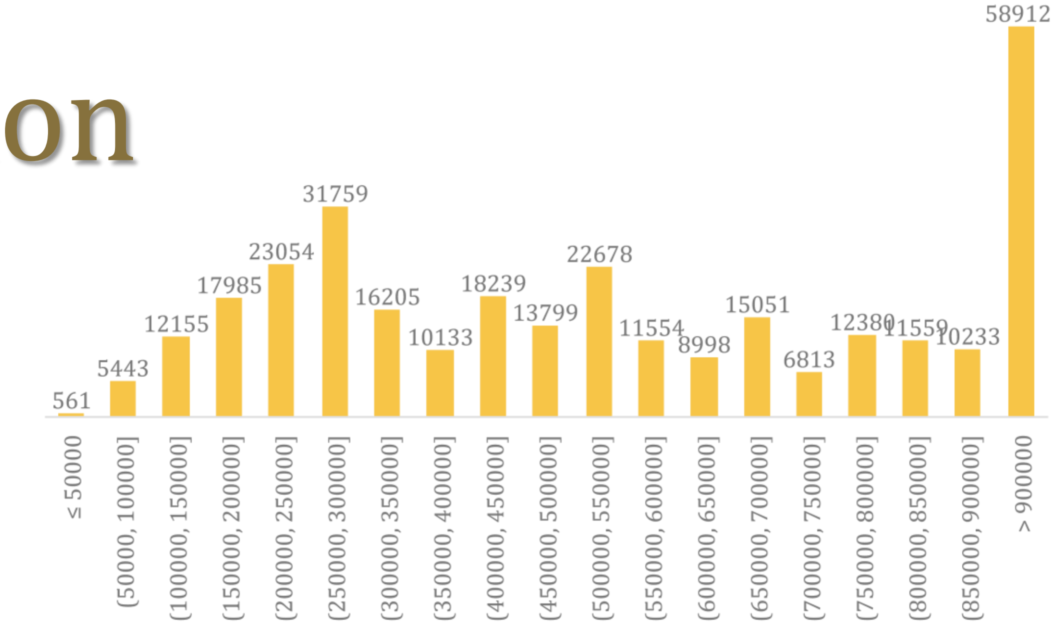
Amount credited has the best correlation with goods price and cnt_children has worst correlation with Year_birth (age).

# Visualization

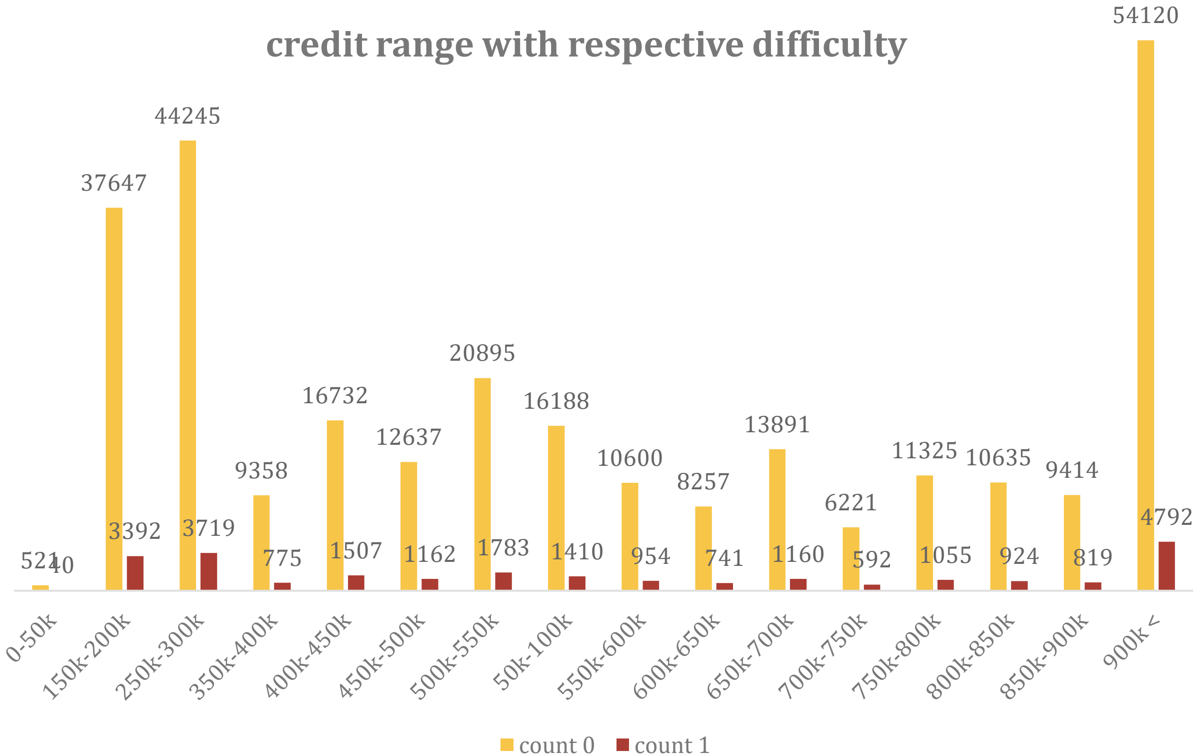
## No of application for clients with difficulty



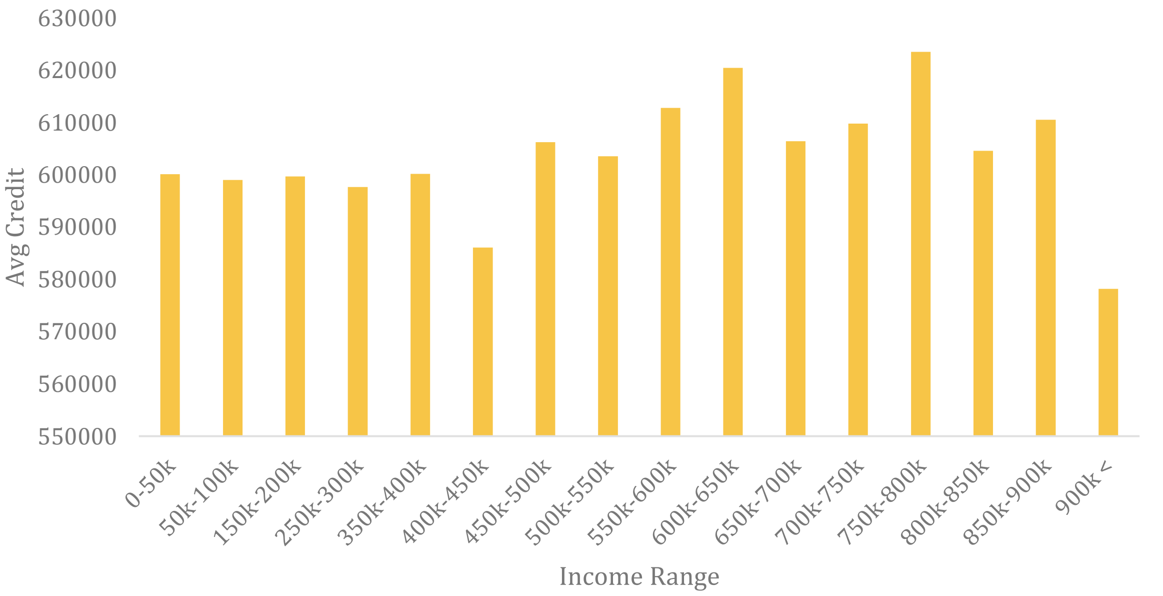
## No of application per credit range.



## credit range with respective difficulty



## Income to credit chart



# Insights

As we can see from the analysis and the charts that it is important to differentiate clients based on their payment difficulties or in other case. As most of the clients having payment difficulties belong to relatively lower income group. Clients having high income range got high credit allotted to them. Most of the loan allotted was for the credit range of 9 lakh and above.

# Result

By doing this project I got to understand how to do Exploratory Data Analysis (EDA) in real life scenario including big datasets. I got to learn a little about Risk Analytics too which banks use in making decision for giving out loans. I learnt how outliers can affect the data analysis. I also  learnt about the various types of analysis and how it is useful to draw out insights.

# THANK YOU

:Priyanshu Kamal
:callmepriyanshu4@gmail.com