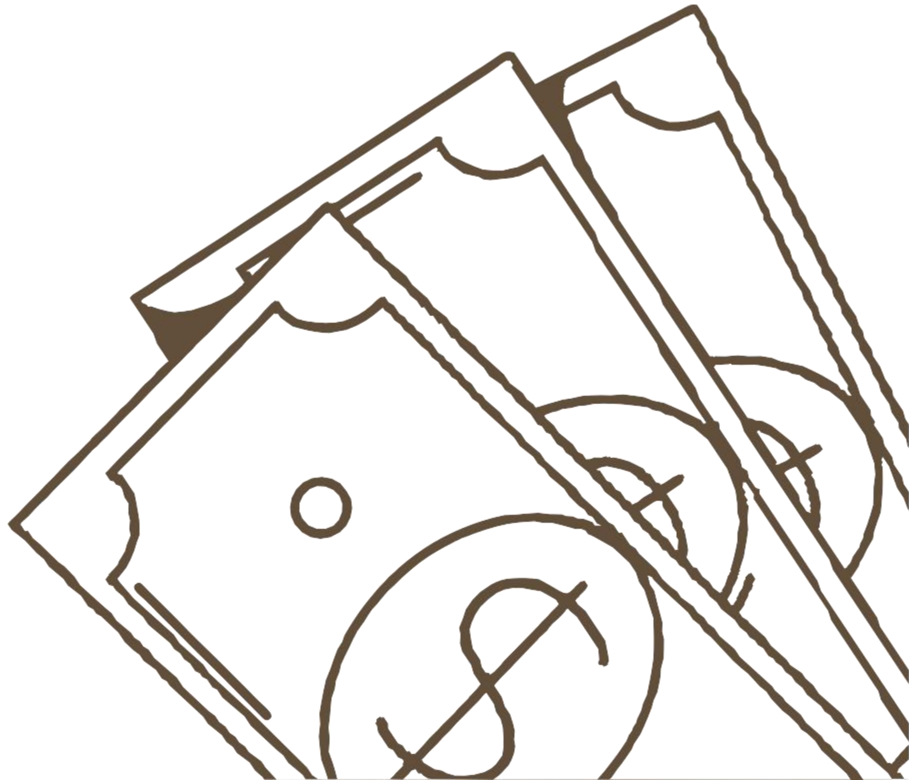
Bank Loan Case Study

Project Report

By

Shubhangi Chaudhary



Contents

01 Project Description

02 Approach

03 Tech stack

04 Insights

05 Results

06 Thank you

Project Description

A finance company that specializes in lending various types of loans to urban customers faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans. The task is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.

Approach

For the given data the process of analysis the trends are as follows:

- **Identification of Missing Data and Dealing with it Appropriately.** It is essential to handle missing data effectively to ensure the accuracy of the analysis.
- **Identify Outliers in the Dataset,** as Outliers can significantly impact the analysis and distort the results. We identify outliers in the loan application dataset.
- **Analyze Data Imbalance,** since it can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

- **Perform Univariate, Segmented Univariate, and Bivariate Analysis,** to gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.
- **Identify Top Correlations for Different Scenarios.** Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

Tech stack

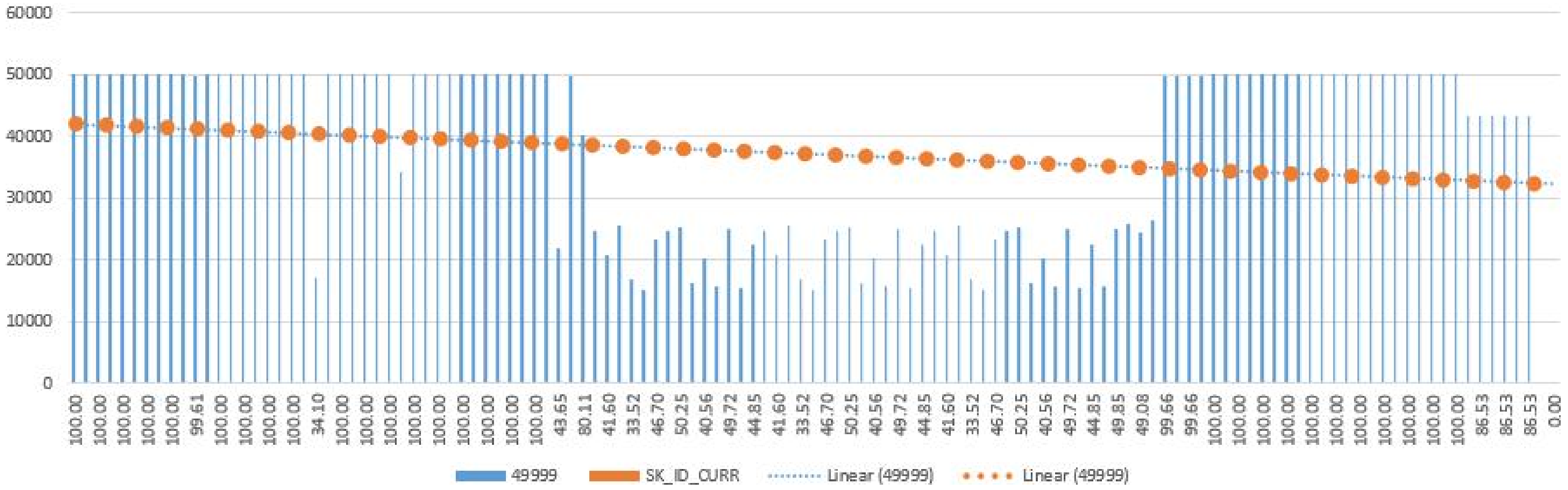
The project required me to work extensively with MS-Excel (2022), allowing me to gain a better understanding related to its various features and how I could seek relevant insights with the same when dealing with a huge quantity of data.

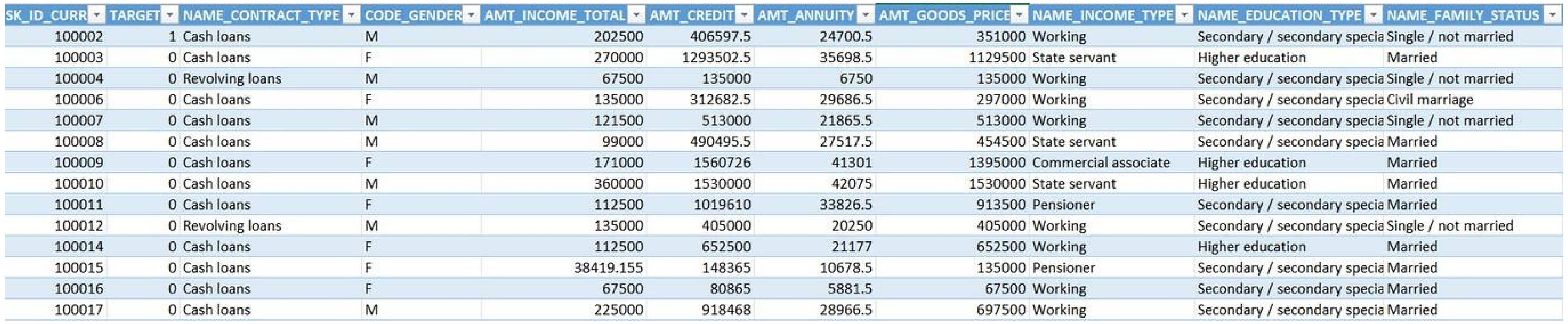
Identifying Missing Data and Dealing with it Appropriately

Insights

The amount of missing values was estimated and columns with nulls of about a certain percentage or close, were dropped. The data with numbers having blanks can be substituted with an average value and others can make use of median and mode values. The columns irrelevant to the analysis were also dropped.

Graph showing trend of null values in the raw data



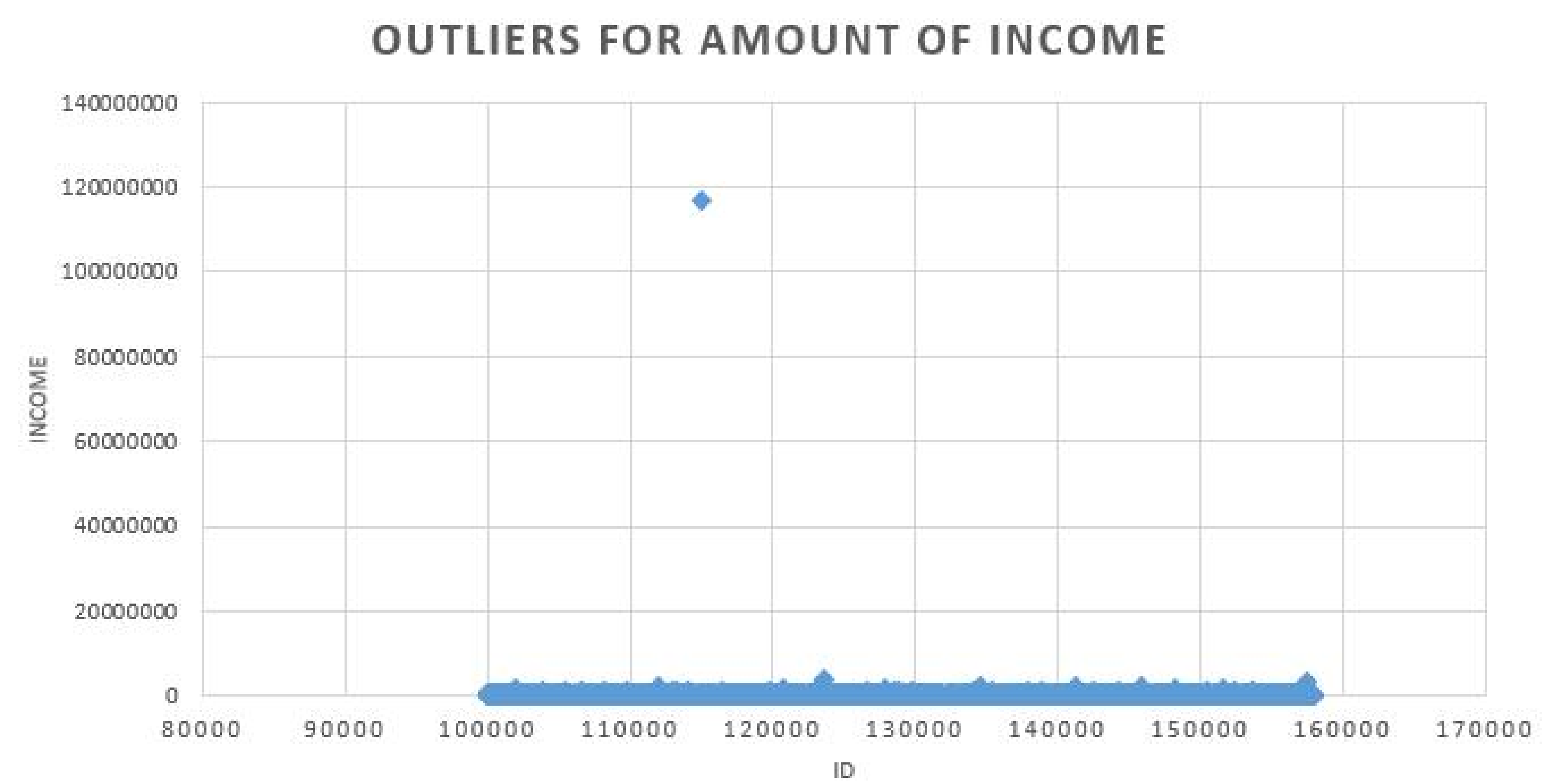


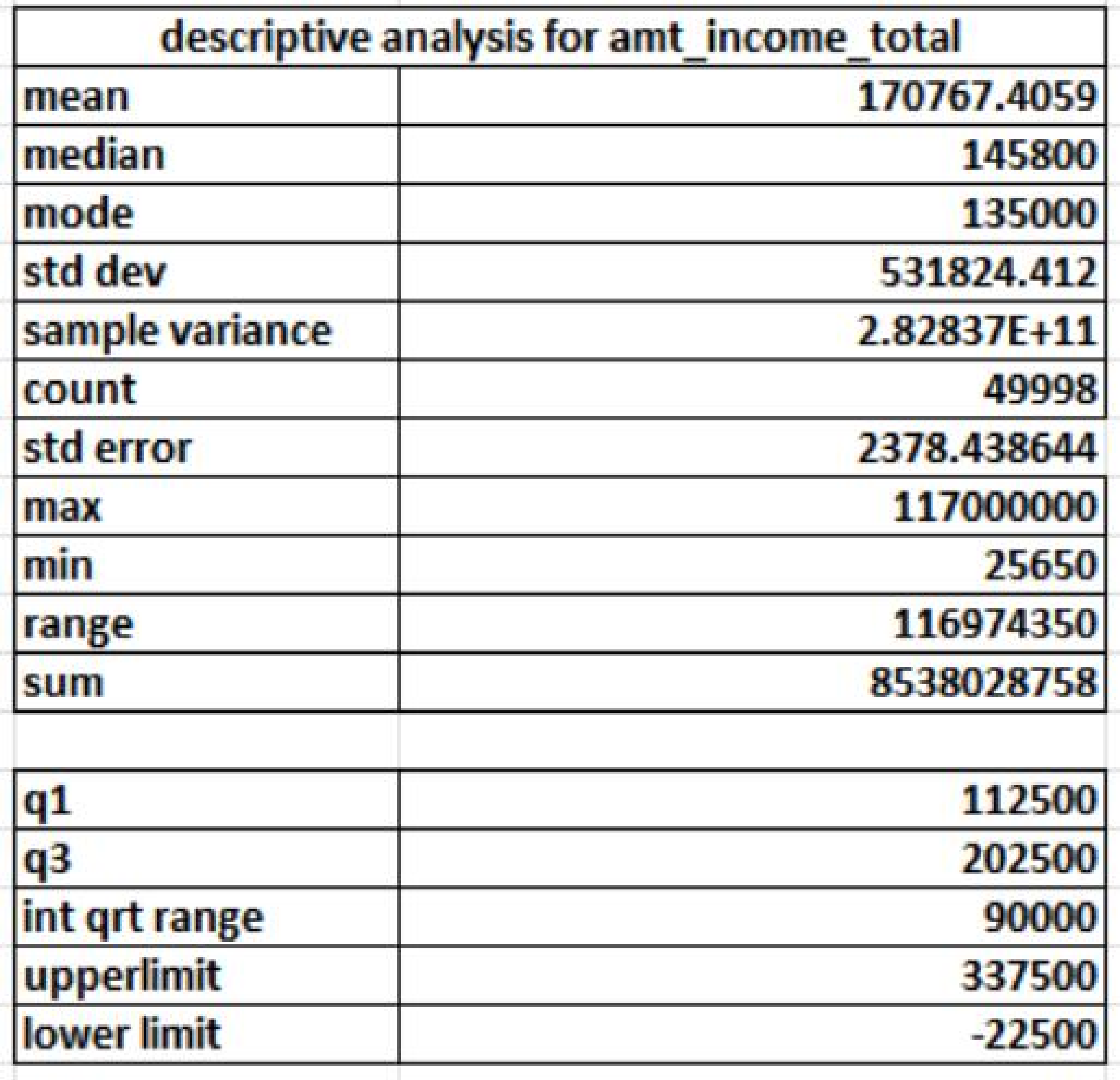
SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS
100002	1	Cash loans	M	202500	406597.5	24700.5	351000	Working	Secondary / secondary specia	Single / not married
100003	0	Cash loans	F	270000	1293502.5	35698.5	1129500	State servant	Higher education	Married
100004	0	Revolving loans	M	67500	135000	6750	135000	Working	Secondary / secondary specia	Single / not married
100006	0	Cash loans	F	135000	312682.5	29686.5	297000	Working	Secondary / secondary specia	Civil marriage
100007	0	Cash loans	M	121500	513000	21865.5	513000	Working	Secondary / secondary specia	Single / not married
100008	0	Cash loans	M	99000	490495.5	27517.5	454500	State servant	Secondary / secondary specia	Married
100009	0	Cash loans	F	171000	1560726	41301	1395000	Commercial associate	Higher education	Married
100010	0	Cash loans	M	360000	1530000	42075	1530000	State servant	Higher education	Married
100011	0	Cash loans	F	112500	1019610	33826.5	913500	Pensioner	Secondary / secondary specia	Married
100012	0	Revolving loans	M	135000	405000	20250	405000	Working	Secondary / secondary specia	Single / not married
100014	0	Cash loans	F	112500	652500	21177	652500	Working	Higher education	Married
100015	0	Cash loans	F	38419.155	148365	10678.5	135000	Pensioner	Secondary / secondary specia	Married
100016	0	Cash loans	F	67500	80865	5881.5	67500	Working	Secondary / secondary specia	Married
100017	0	Cash loans	M	225000	918468	28966.5	697500	Working	Secondary / secondary specia	Married

Identify Outliers in the Dataset

Insights

Oultiers were the lone points in the graphs that can greatly impact the average leading to misinterpretation of data. The following outliers were found for the total amount of income.
The descriptive analysis was also carried out.





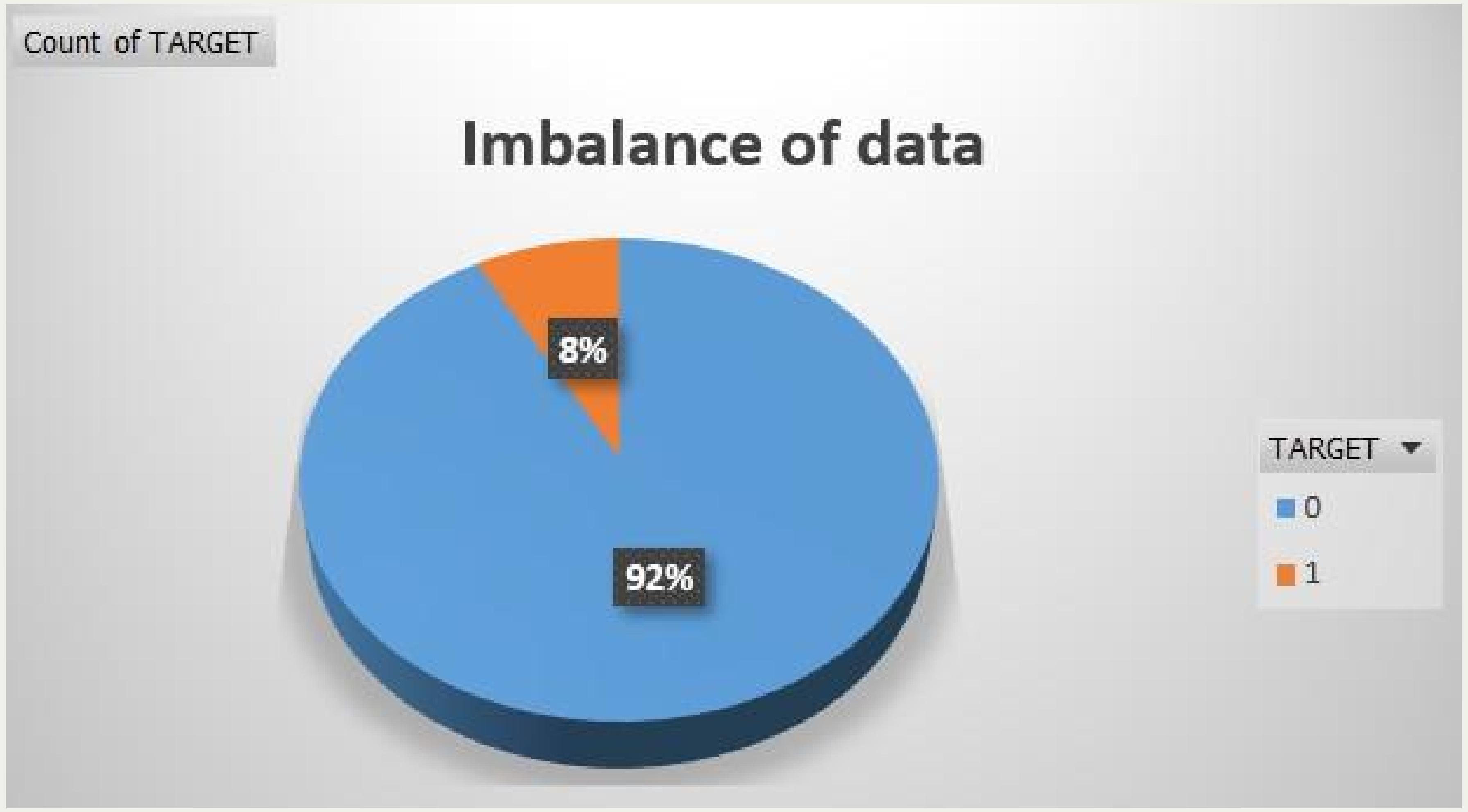
descriptive analysis for amt_income_total	
mean	170767.4059
median	145800
mode	135000
std dev	531824.412
sample variance	2.82837E+11
count	49998
std error	2378.438644
max	117000000
min	25650
range	116974350
sum	8538028758

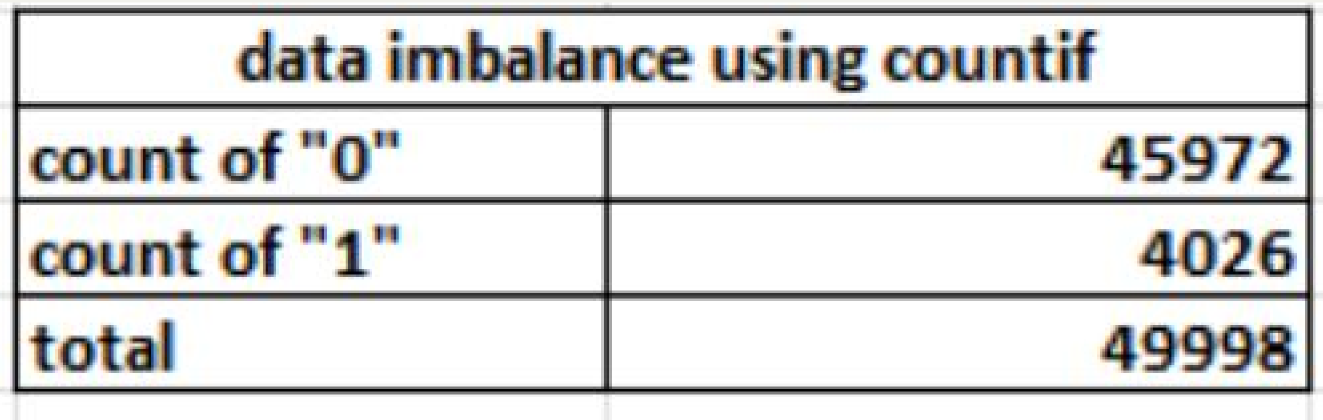
q1	112500
q3	202500
int qrt range	90000
upperlimit	337500
lower limit	-22500

Analyze Data Imbalance

Insights

The imbalance was quite prominent, only a small percentage of people(8%) face any difficulties .



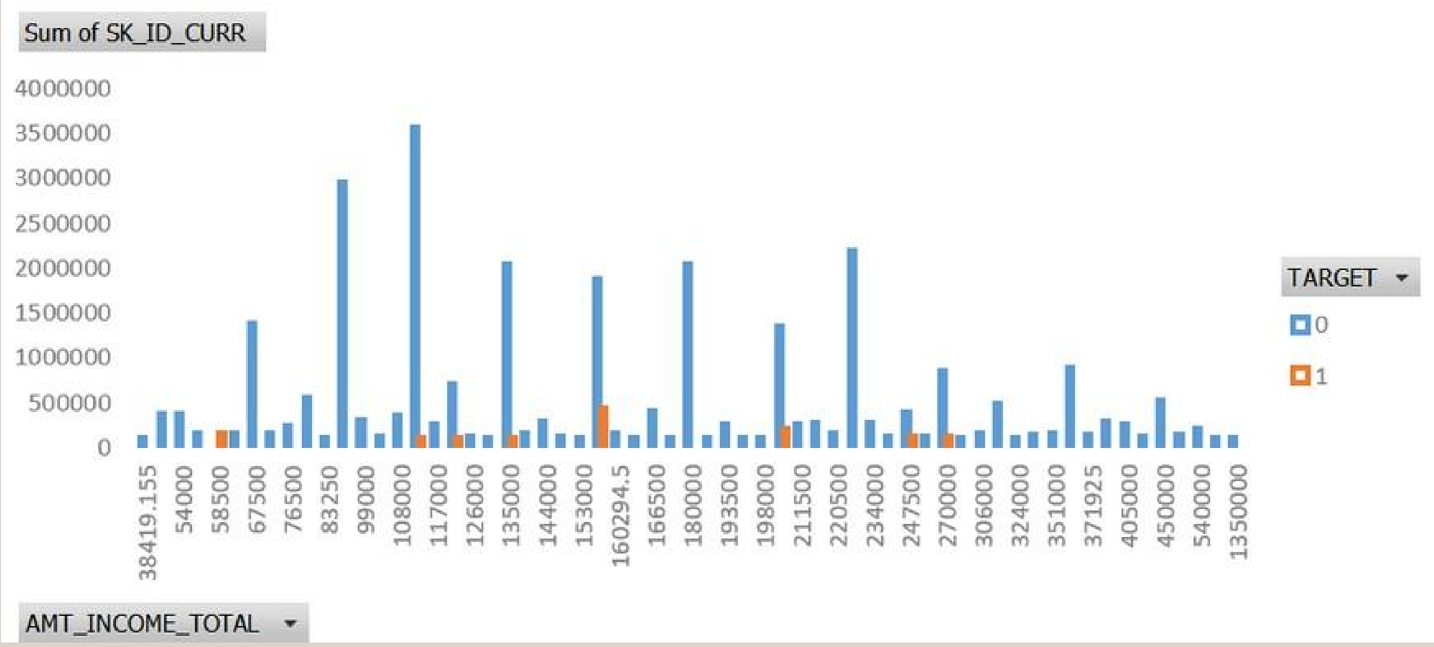


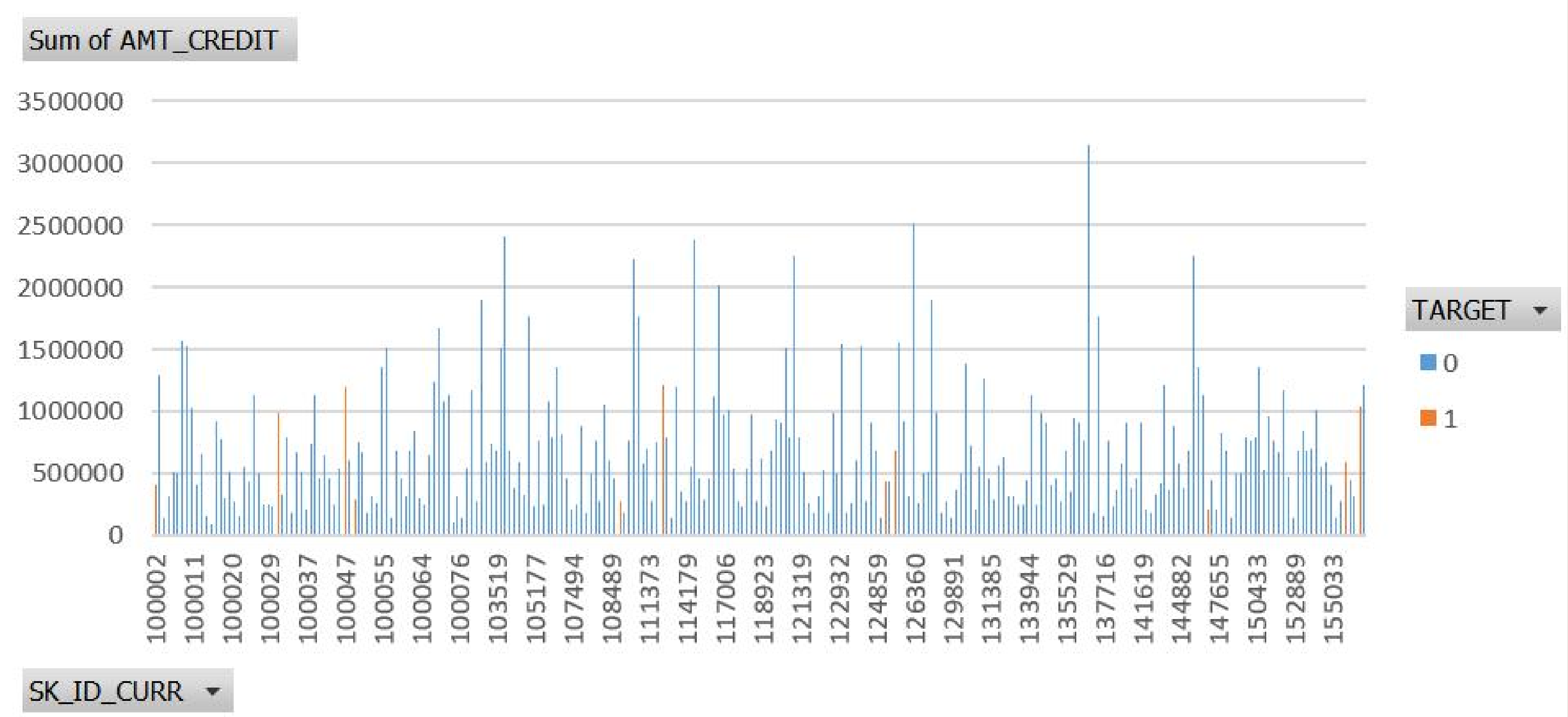
data imbalance using countif	
count of "0"	45972
count of "1"	4026
total	49998

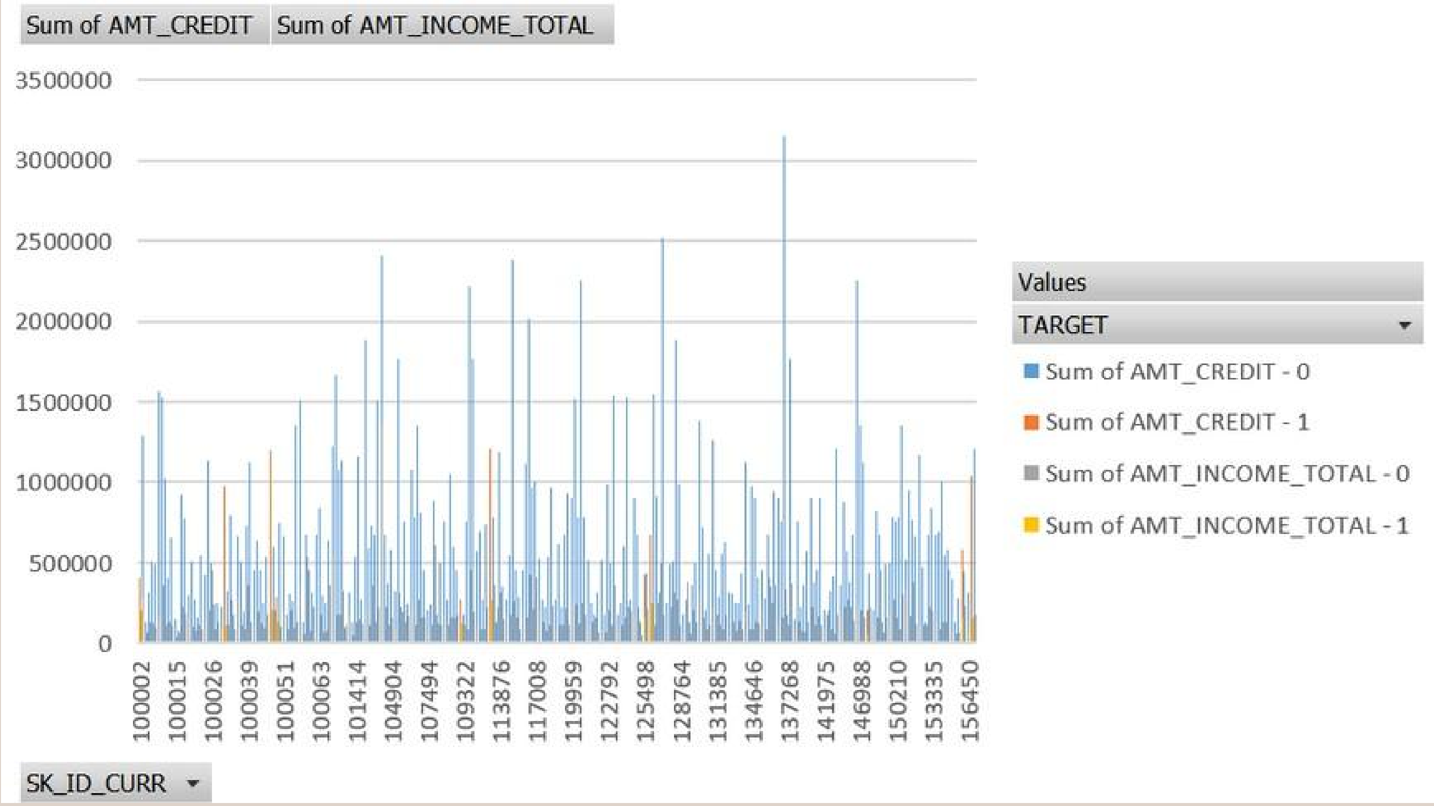
Perform Univariate, Segmented Univariate, and Bivariate Analysis

Insights

The visualizations obtained were quite distinct.

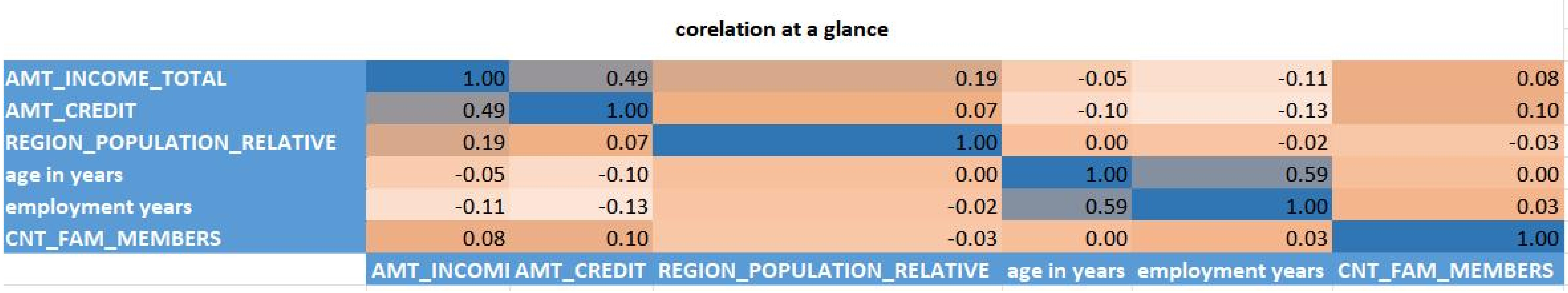






Identify Top Correlations for Different Scenarios

Insights



corelation at a glance

AMT_INCOME_TOTAL	1.00	0.49	0.19	-0.05	-0.11	0.08
AMT_CREDIT	0.49	1.00	0.07	-0.10	-0.13	0.10
REGION_POPULATION_RELATIVE	0.19	0.07	1.00	0.00	-0.02	-0.03
age in years	-0.05	-0.10	0.00	1.00	0.59	0.00
employment years	-0.11	-0.13	-0.02	0.59	1.00	0.03
CNT_FAM_MEMBERS	0.08	0.10	-0.03	0.00	0.03	1.00
	AMT_INCOMI	AMT_CREDIT	REGION_POPULATION_RELATIVE	age in years	employment years	CNT_FAM_MEMBERS

The correlations were quite prominent between factors as income and credit amount and ages of individuals. It shows that people with higher income will have better chances of facing no trouble and people of higher age are more likely to face no issues .

Results

Through this project, I got to better understand the many ways in which we can gather meaningful insights from a large amount of data. MS Excel can indeed be quite a powerful tool that can be helpful in answering a number of questions extensively to better address various doubts and make informed decisions by observing trends over time.

link to excel workbook & video

Thank you