Bank Loan Case Study

By: Rajat Panwan

Project Description:

The project aims to use exploratory data analysis (EDA) to analyze loan application data and identify patterns that indicate if a client has difficulty paying their installments. The project aims to understand the driving factors behind loan default, i.e. the variables which are strong indicators of default, and utilize this knowledge for the company's portfolio and risk assessment.

Tech-Stack Used:

The project was performed using Excel version 2021 Excel was chosen for its powerful data analysis and visualization capabilities It is also widely used in the industry and provides a familiar environment for users.

Insights:

1.The mean income for the income is around 160000.
2.There is something wrong in the data the mean of the days_employeed in not acceptable.
3. 8 percent people have difficulties in payments.
4. It show that Amt_Goods_Price and Amt_Credit are the most correlated variable followed by Amt_Annuity & Amt_Goods_price and Amt_Credit & Amt_Annuity.
5. Although Females are more difficulties with payment but percentage wise it is male who face difficulties in payments.
6. Here are more females than males. The ratio of male and female in around 1:2.
7.The average of the income for male is more as compared to female.
8.Academic degree holder apply for the loan most.
9.Academic degree holder ask for more credit as compare to other educational background.
10.Male ask for greater money for loan as compare to the female.

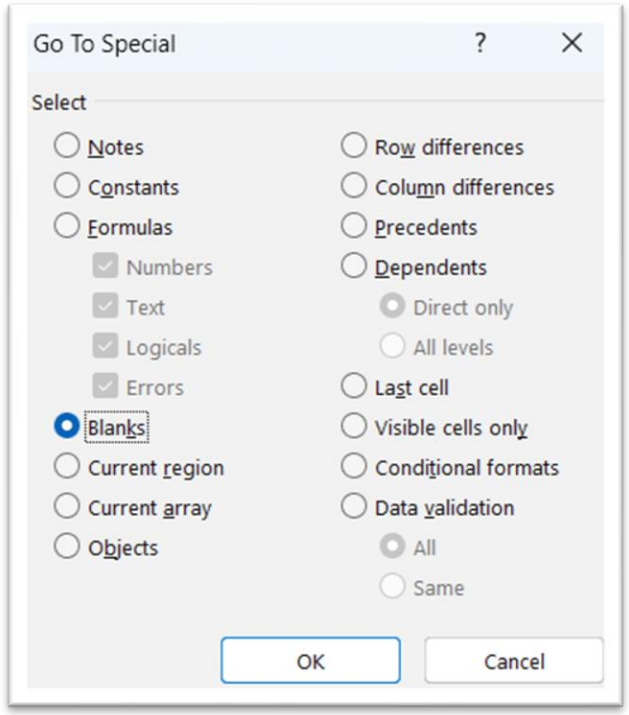
Result:

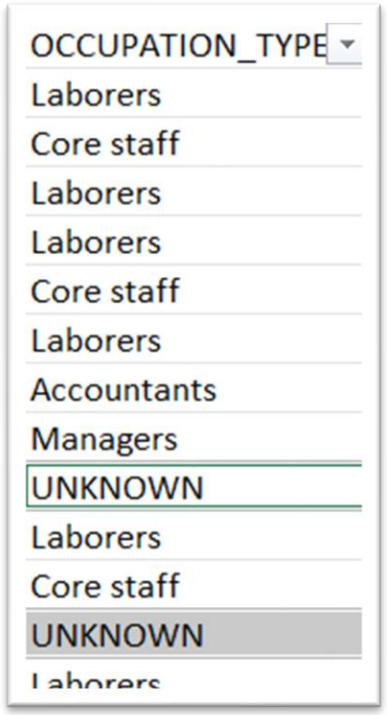
Through this project, I was able to gain valuable insights into the factors that influence loan default and develop a basic understanding of risk analytics in the banking and financial services sector. What I find is on next pages.

Present the overall approach of the analysis. Mention the problem statement and the analysis approach briefly

The project was executed by first performing data cleaning and preprocessing, followed by exploratory data analysis (EDA). During EDA, various statistical and visualization techniques were used to gain insights into the data and identify patterns related to loan default. The data was also analyzed in conjunction with the data dictionary and domain research on risk analytics.

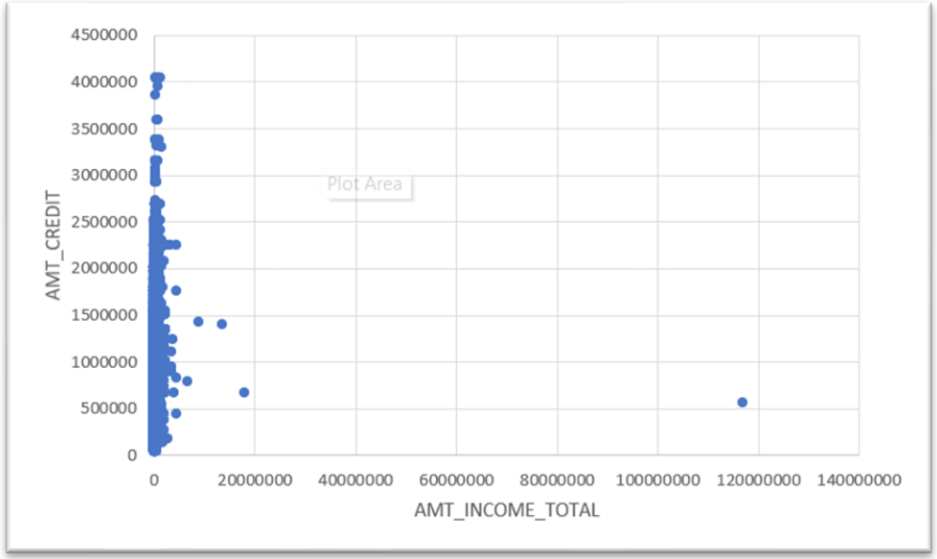
Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value):

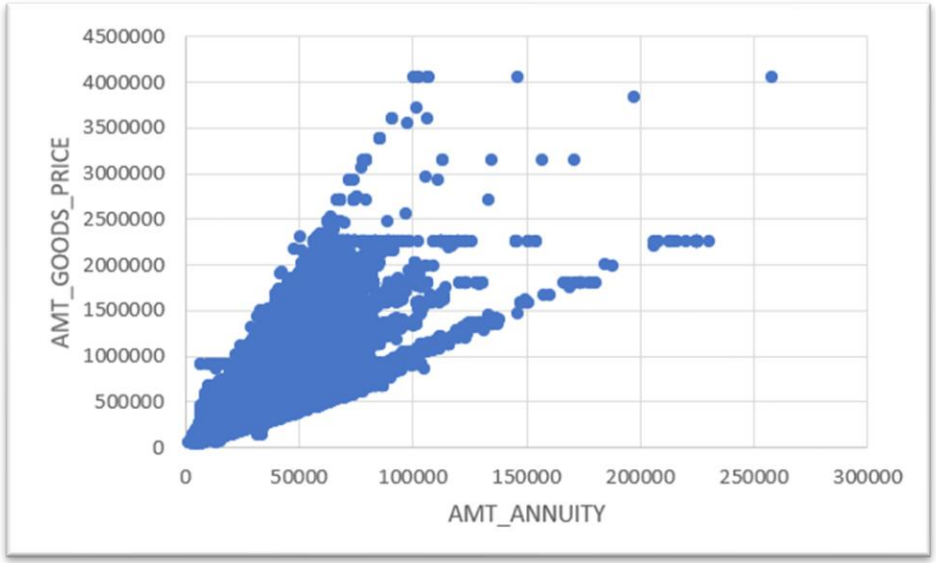




- Delete the row which has blank vlues.
- In some column such as occupation_type with blank values in replaced with unkown.
- Delete some columns which is not required in the analysis.

Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier:





- SK_ID_CURR 114967 is the outlier because his credit is very less as compare to his income.
- In second graph SK_ID_CURR 120926 is the outlier with respect to amount goods price and amount annuity.
- Scatter Plot is used for getting the outliers.

Identify if there is data imbalance in the data. Find the ratio of data imbalance:

Row Labels	Count of NAME_INCOME_TYPE	Ratio
Working	158774	31754.8
Commercial associate	71617	14323.4
Pensioner	55362	11072.4
State servant	21703	4340.6
Unemployed	22	4.4
Student	18	3.6
Businessman	10	2.0
Maternity leave	5	1.0
Grand Total	**307511**	

Row Labels	Count of HOUSETYPE_MODE	Ratio
block of flats	150503	124.2
specific housing	1499	1.2
terraced house	1212	1.0
Grand Total	**153214**	

Row Labels	Count of NAME_EDUCATION_TYPE	Ratio
Secondary / secondary special	218391	1332
Higher education	74863	456
Incomplete higher	10277	63
Lower secondary	3816	23
Academic degree	164	1
Grand Total	**307511**	

Row Labels	Count of CODE_GENDER
F	202448
M	105059
XNA	4
Grand Total	**307511**

- Pivot table is used for checking the data imbalance. Here data is imbalanced as most of the data is from the working Income Type where as very less for business. The ration for working and business income type is 31000:2
- Here data is imbalanced because most of the House type is block of flats and others are very less. The ratio of block of flats and other house type is 124:2
- Here data is imbalanced because most of the data is has education type and very less for academic degree their ratio is 1332:1
- Here are more females than males. The ratio of male and female in around 1:2.

Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms:

Univariate:

AMT_INCOME_TOTAL	
Mean	168798
Median	147150
Mode	135000

AMT_CREDIT	
Mean	599025.9997
Median	513531
Mode	450000

DAYS_EMPLOYED		In Years
Mean	63815	175
Median	-1213	-3
Mode	365243	1001

AMT_GOODS_PRICE	
Mean	538396
Median	450000
Mode	450000

REGION_POPULATION_RELATIVE	
Mean	0.02087
Median	0.01885
Mode	0.03579

DAYS_BIRTH		Age in Years
Mean	-16037	44
Median	-15750	43
Mode	-13749	38

AMT_ANNUITY	
Mean	27109
Median	24903
Mode	9000

DAYS_REGISTRATION	
Mean	-4986
Median	-4504
Mode	-1

DAYS_ID_PUBLISH		in years
Mean	-2994	8
Median	-3254	9
Mode	-4053	11

- In first table AMT_INCOME_TATAL mean is amount 168798 , median is 147150 where as mode is 135000
- In second table AMT_CREDIT mean is amount 599025, median is 513531 and mode is 450000.
- In DAYS_EMPLOYED mean is 175 years and mode is 1001 years which not acceptable as data.

```
=AVERAGE(application_data!H:H)
```

```
=MEDIAN(application_data!H:H)
```

```
=MODE.MULT(application_data!H:H)
```

- These three formulae are used for the univariate Analysis.
- In Days_Employed Column there is some mistake in data because mean of the is 174 years which is wrong and mode is 1001 years when days are converted into years.

Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms:

Segmented Univariate:

Average Amount of Income on Gender Basis	
Row Labels	Average of AMT_INCOME_TOTAL
F	156032
M	193396
XNA	186750
Grand Total	**168798**

Average Amount of Credit on Education Basis	
Row Labels	Average of AMT_CREDIT
Academic degree	723516
Higher education	689950
Incomplete higher	566731
Lower secondary	489749
Secondary / secondary special	571193
Grand Total	**599026**

Average Amount of Credit on Education Basis	
Row Labels	Average of AMT_INCOME_TOTAL
Academic degree	240009
Higher education	208652
Incomplete higher	181564
Lower secondary	130079
Secondary / secondary special	155159
Grand Total	**168798**

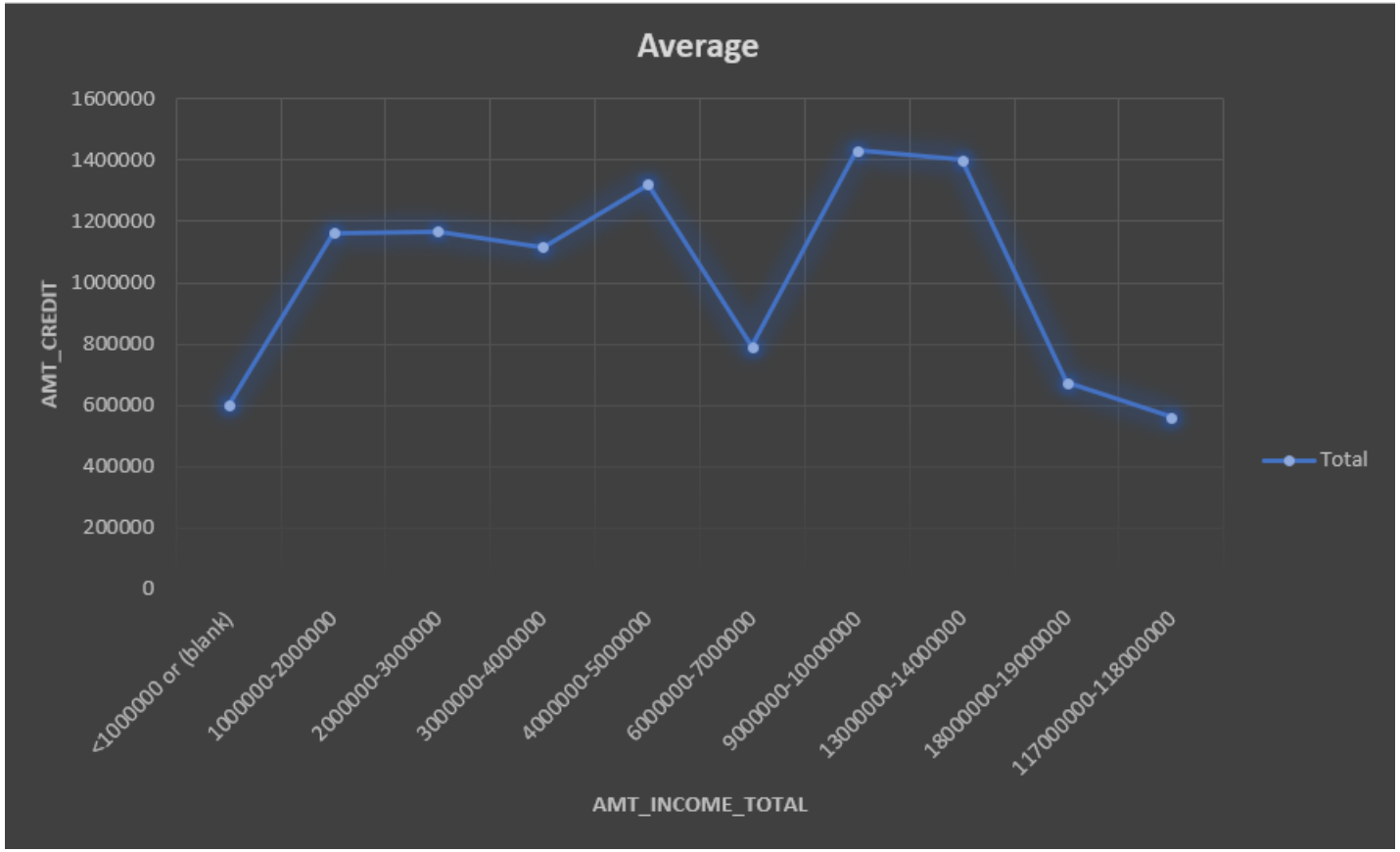
Average Amount of Credit on Gender Basis	
Row Labels	Average of AMT_CREDIT
F	592767
M	611095
XNA	399375
Grand Total	**599026**

- The average of the income for male is more as compared to female.
- Academic degree holder apply for the loan most.
- Academic degree holder ask for more credit as compare to other educational background.
- Male ask for greater money for loan as compare to the female.

Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms:

Bivariate Analysis (Part-1):

Row Labels	Average of AMT_CREDIT
<1000000 or (blank)	598568
1000000-2000000	1163950
2000000-3000000	1167367
3000000-4000000	1116665
4000000-5000000	1322595
6000000-7000000	790830
9000000-10000000	1431531
13000000-14000000	1400504
18000000-19000000	675000
117000000-118000000	562491
Grand Total	**599026**

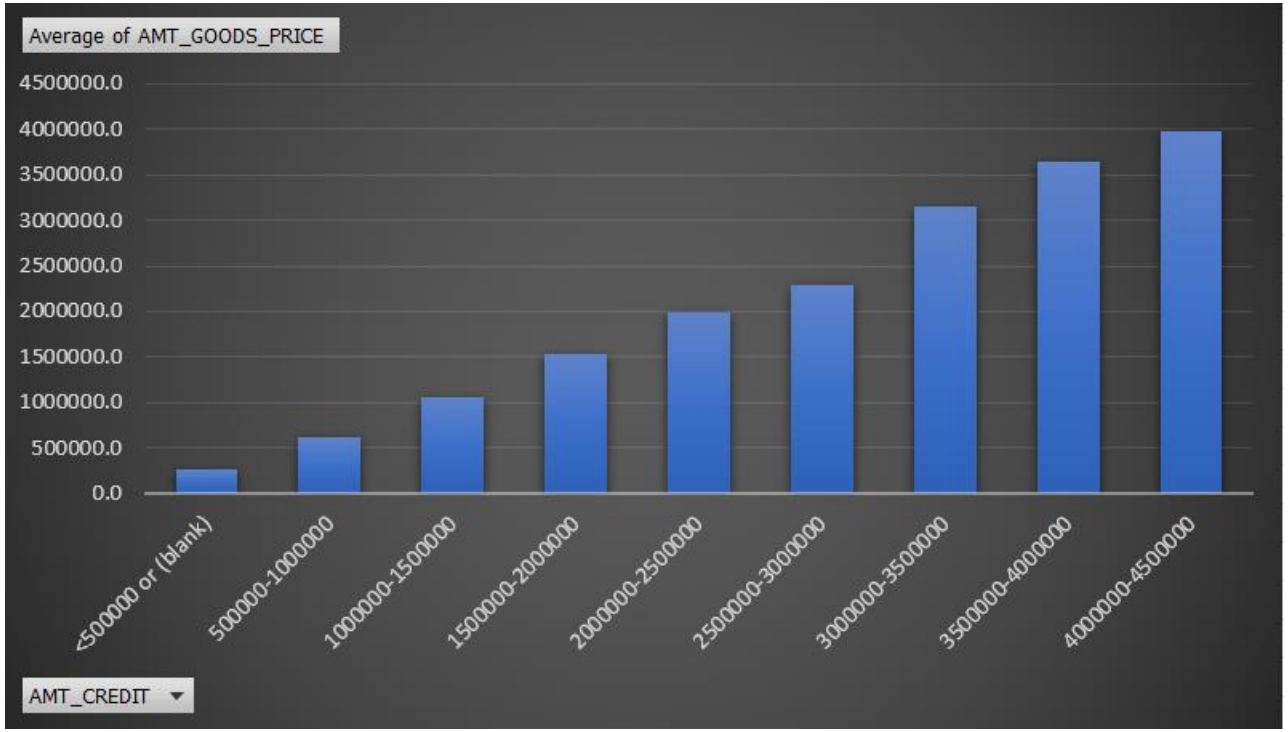


- This is Bivariate Analysis because it show how the one parameter is behave with the change in other.
- The adjacent line graph show that people with income 9000000 to 14000000 is ask for more loan.
- Where as income between 6000000 to 7000000 is lesser than the nearest income class.

Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms:

Bivariate Analysis (Part-2):

Row Labels	Average of AMT_GOODS_PRICE
<500000 or (blank)	259932.3
500000-1000000	622558.8
1000000-1500000	1064841.0
1500000-2000000	1531780.4
2000000-2500000	1989482.6
2500000-3000000	2285808.5
3000000-3500000	3157500.0
3500000-4000000	3645000.0
4000000-4500000	3971250.0
Grand Total	**538396**



- The adjacent bar graph show that Credit amount is increases with the increase of Good Price.

Find the top 10 correlation for the Clients other than payment difficulties:

	AMT_INCOM	AMT_CRE	AMT_ANN	AMT_GOO	REGION_PO	DAYS_BIRT	DAYS_EMPL	DAYS_REGIS	DAYS_ID_PU
AMT_INCOME_TOTAL	1.000								
AMT_CREDIT	0.343	1.000							
AMT_ANNUITY	0.419	0.771	1.000						
AMT_GOODS_PRICE	0.349	0.987	0.777	1.000					
REGION_POPULATION_RELATIVE	0.168	0.100	0.121	0.104	1.000				
DAYS_BIRTH	0.063	-0.047	0.013	-0.045	-0.025	1.000			
DAYS_EMPLOYED	-0.141	-0.073	-0.107	-0.071	-0.007	-0.618	1.000		
DAYS_REGISTRATION	0.065	0.014	0.039	0.016	-0.052	0.333	-0.210	1.000	
DAYS_ID_PUBLISH	0.023	-0.001	0.014	-0.004	-0.001	0.271	-0.274	0.100	1.000

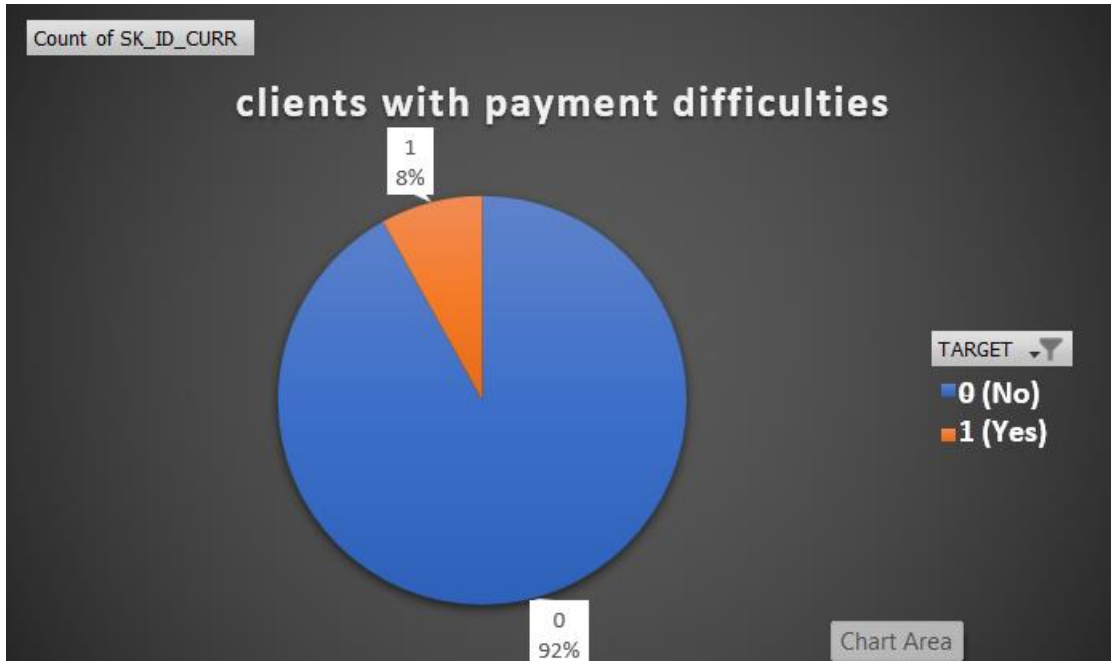
- This is the correlation matrix for clients other than payment difficulties.
- It show that Amt_Goods_Price and Amt_Credit are the most correlated variable followed by Amt_Annuity & Amt_Goods_price and Amt_Credit & Amt_Annuity.

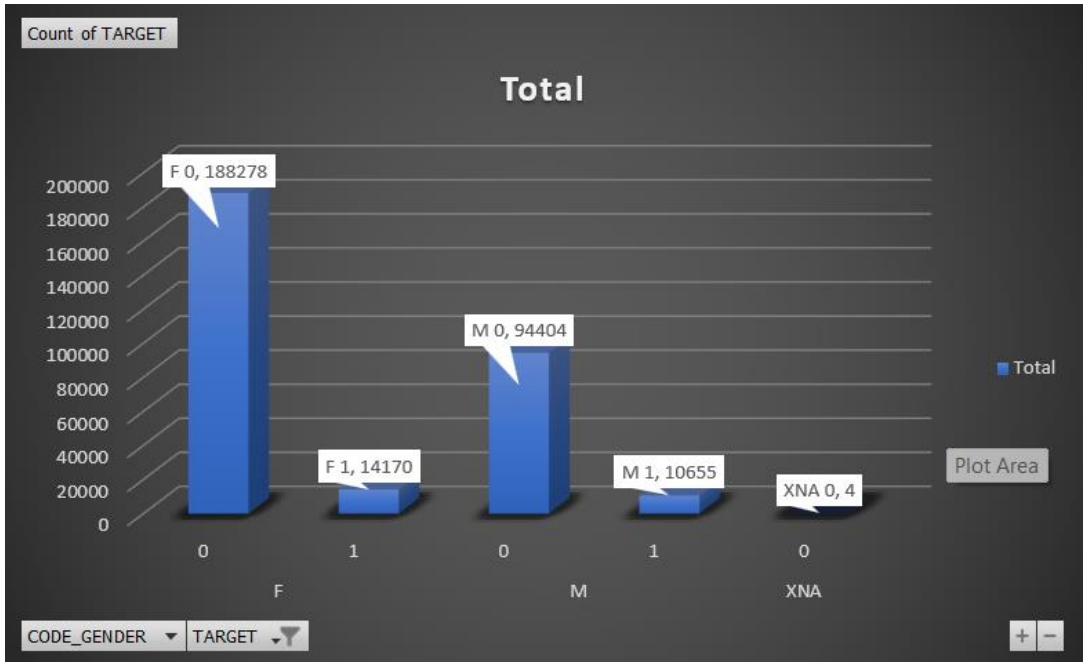
Find the top 10 correlation for the Client with payment difficulties:

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRI	REGION_POPULAT	DAYS_BIRTH	DAYS_EMPLOYE	DAYS_REGISTRATI	DAYS_ID_PUBLIS
AMT_INCOME_TOTAL	1.000								
AMT_CREDIT	0.038	1.000							
AMT_ANNUITY	0.046	0.752	1.000						
AMT_GOODS_PRICE	0.038	0.983	0.753	1.000					
REGION_POPULATION_RELATIVE	0.009	0.069	0.072	0.076	1.000				
DAYS_BIRTH	0.003	-0.135	-0.014	-0.136	-0.048	1.000			
DAYS_EMPLOYED	-0.015	-0.001	-0.083	0.004	0.015	-0.575	1.000		
DAYS_REGISTRATION	0.000	-0.026	0.034	-0.026	-0.056	0.289	-0.189	1.000	
DAYS_ID_PUBLISH	-0.004	-0.052	-0.017	-0.056	-0.016	0.253	-0.226	0.097	1.000

- This is the correlation matrix for clients with payment difficulties.
- It show that Amt_Goods_Price and Amt_Credit are the most correlated variable followed by Amt_Annuity & Amt_Goods_price and Amt_Credit & Amt_Annuity.

Include visualizations and summarize the most important results in the presentation:

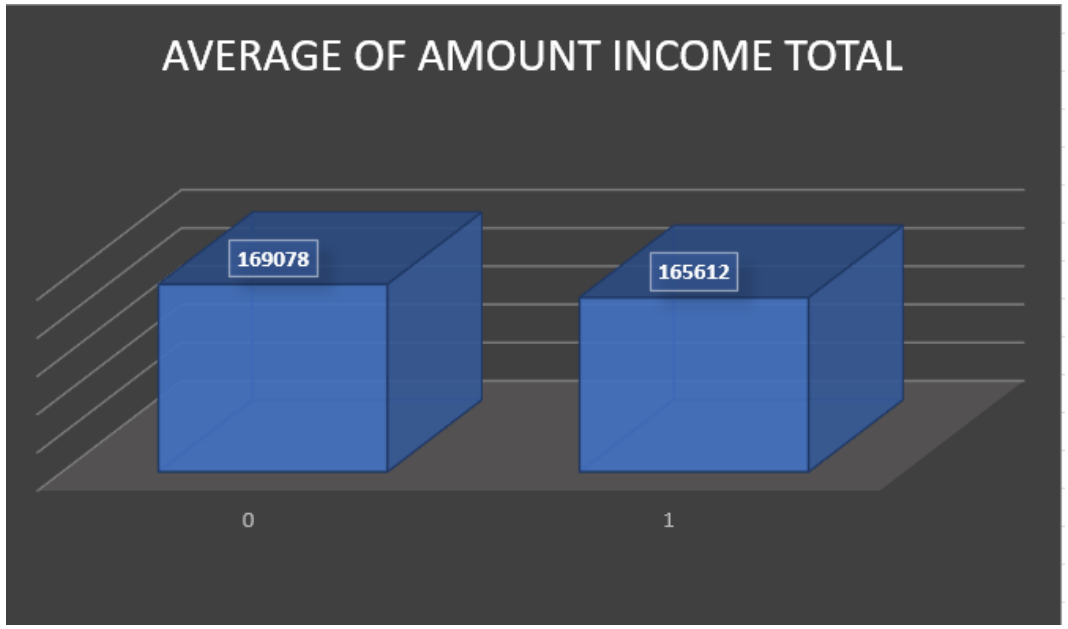


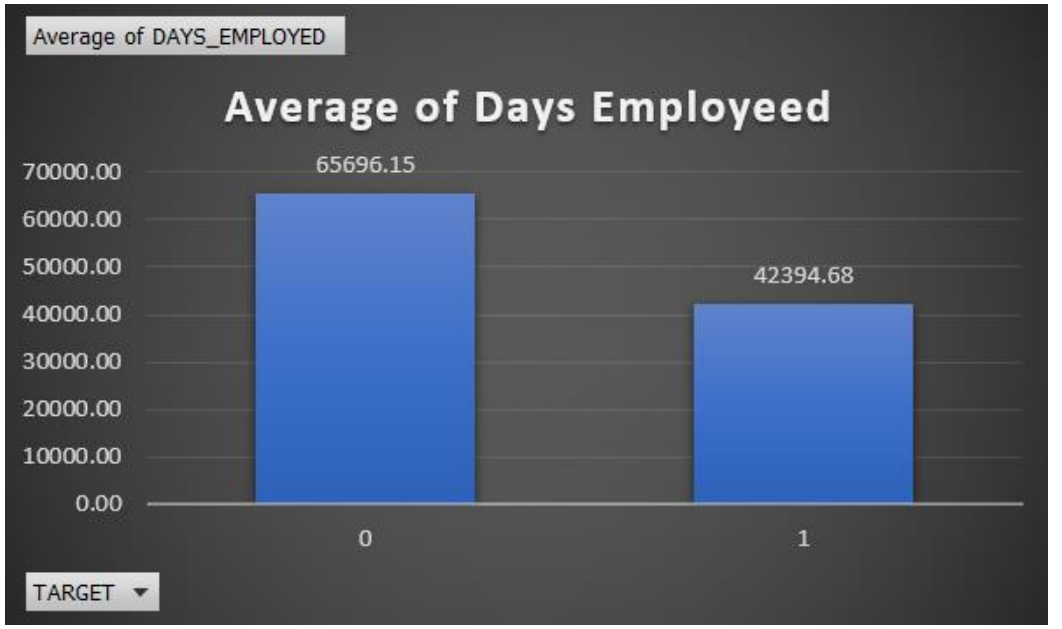


- In above pie chart we can see that on 8% of clients have payments difficulties.

- According to the above graph it is shown that although Females are more difficulties with payment but percentage wise it is male who face difficulties in payments.

Include visualizations and summarize the most important results in the presentation





- In above bar chart it can be seen that client with payment difficulties and other clients have almost same income amount.

- In above bar chart it can be seen that client with payment difficulties has lesser number of days employeed.

Rajat Panwan

rajatpawan@gmail.com

9457941019