Bank Loan Case Study

By Divyasri Jegan

Introduction:

Welcome to the comprehensive case study on bank loans, a critical aspect of financial services that significantly impacts both individuals and businesses. This study delves into the mechanisms, challenges, and outcomes associated with bank loans, providing an in-depth analysis of the loan process from application to repayment.This project is about the analysis of Bank Loan case study.

Project Description :

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

The case study is given with three datasets:

- Previous_data previous_dataset
- Application_data application_dataset
- Column_data column data

Company’s perspective :

When a customer applies for a loan, your company faces two risks:

1. If the applicant can repay the loan but is not approved, the company loses business.
2. If the applicant cannot repay the loan and is approved, the company faces a financial loss.

Two types of scenarios:

1. Customers with payment difficulties: These are customers who had a late payment of more than X days on at least one of the first Y installments of the loan.
2. All other cases: These are cases where the payment was made on time.

When a customer applies for a loan, there are four possible outcomes:

1. Approved: The company has approved the loan application.
2. Cancelled: The customer cancelled the application during the approval process.
3. Refused: The company rejected the loan.
4. Unused Offer: The loan was approved but the customer did not use it.

Approach:

01 Introduction

02 Project Description

03 Tech Stack Used

04 Insights

05 Analysis

06 Result

07 Conclusion

Tech Stack Used :

Microsoft Excel

Used for data cleaning and EDA analysis to make interactive graphs

PowerPoint

To make ppts and also represent my projects in a interactive way.

Business Objective :

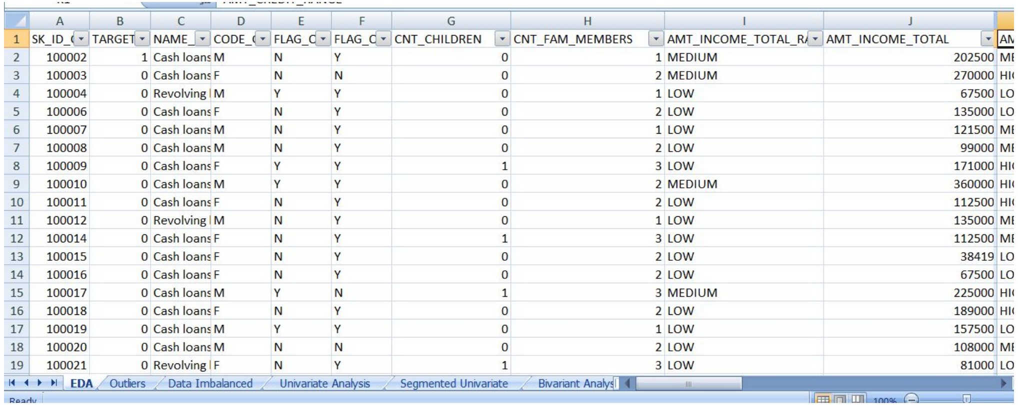
The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their installments. This information can be used to make decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants. The company wants to understand the key factors behind loan default so it can make better decisions about loan approval.

Note: To better understand this project, you might want to research a bit about risk analytics in banking and financial services. Understanding the types of variables and their significance should be enough.

Insights:

1) **Identify Missing Data and Deal with it Appropriately:** As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

- **Task:** Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

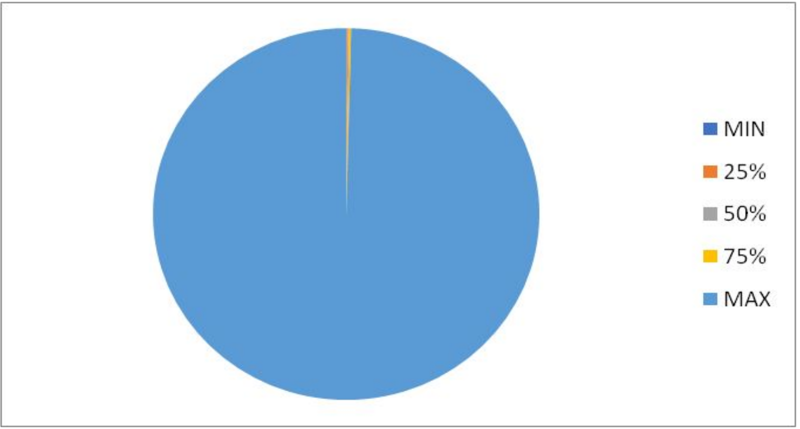


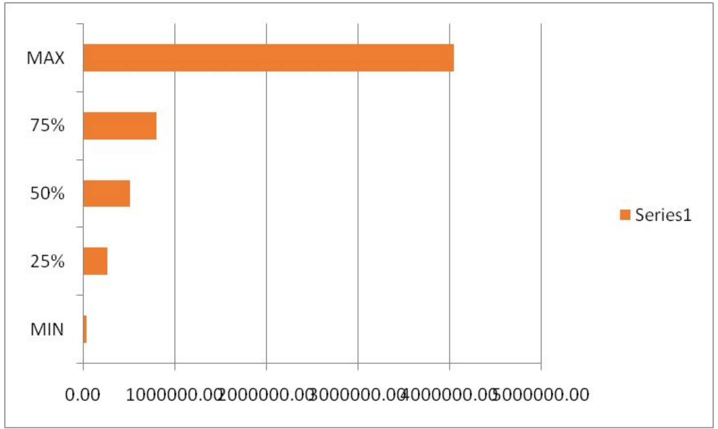
	A	B	C	D	E	F	G	H	I	J	
1	SK_ID_(	TARGET	NAME_	CODE_(	FLAG_C	FLAG_C	CNT_CHILDREN	CNT_FAM_MEMBERS	AMT_INCOME_TOTAL_RA	AMT_INCOME_TOTAL	AM
2	100002	1	Cash loans	M	N	Y	0	1	MEDIUM	202500	ME
3	100003	0	Cash loans	F	N	N	0	2	MEDIUM	270000	HI(
4	100004	0	Revolving I	M	Y	Y	0	1	LOW	67500	LO
5	100006	0	Cash loans	F	N	Y	0	2	LOW	135000	LO
6	100007	0	Cash loans	M	N	Y	0	1	LOW	121500	ME
7	100008	0	Cash loans	M	N	Y	0	2	LOW	99000	ME
8	100009	0	Cash loans	F	Y	Y	1	3	LOW	171000	HI(
9	100010	0	Cash loans	M	Y	Y	0	2	MEDIUM	360000	HI(
10	100011	0	Cash loans	F	N	Y	0	2	LOW	112500	HI(
11	100012	0	Revolving I	M	N	Y	0	1	LOW	135000	ME
12	100014	0	Cash loans	F	N	Y	1	3	LOW	112500	ME
13	100015	0	Cash loans	F	N	Y	0	2	LOW	38419	LO
14	100016	0	Cash loans	F	N	Y	0	2	LOW	67500	LO
15	100017	0	Cash loans	M	Y	N	1	3	MEDIUM	225000	HI(
16	100018	0	Cash loans	F	N	Y	0	2	LOW	189000	HI(
17	100019	0	Cash loans	M	Y	Y	0	1	LOW	157500	LO
18	100020	0	Cash loans	M	N	N	0	2	LOW	108000	ME
19	100021	0	Revolving I	F	N	Y	1	3	LOW	81000	LO

- Using EDA analysis and COUNTA and NULL function I have handled the missing values of given dataset.

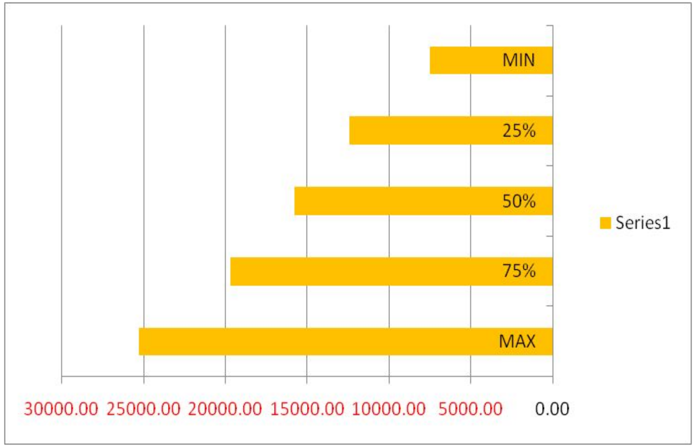
2). Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

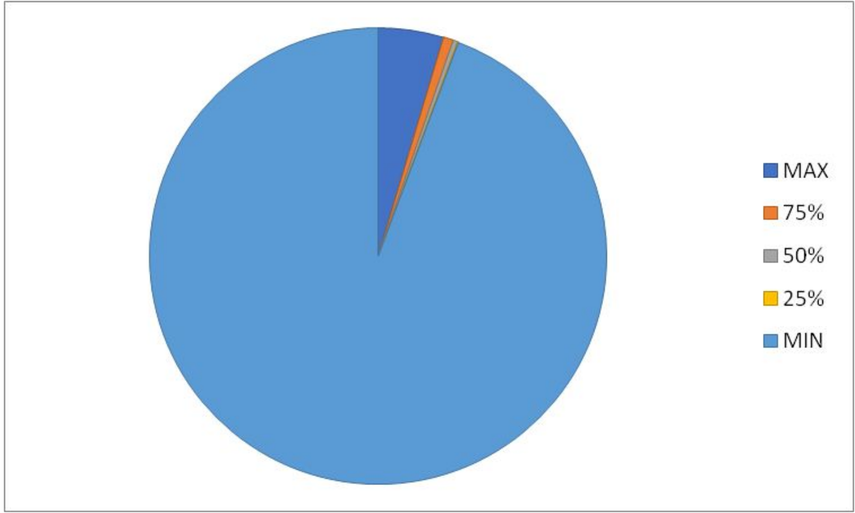
- **Task:** Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.





Outliers in AMT_ANNUITY
AMT_CREDIT

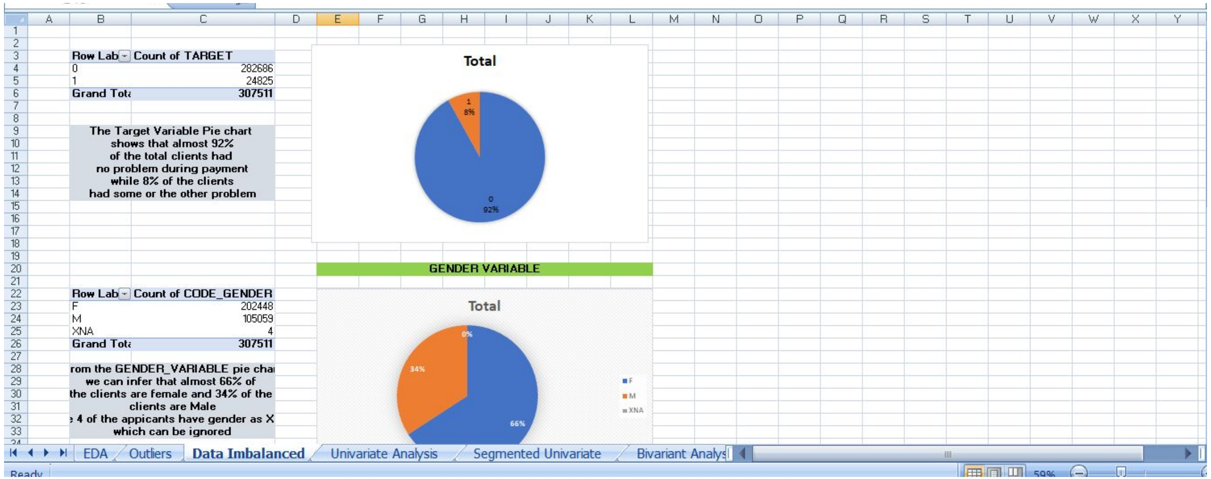




Outliers in DAYS_OF_BIRTH and DAYS_EMPLOYED

3) Analyze Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

- **Task:** Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.



Row Lab	Count of TARGET
0	282686
1	24825
Grand Tota	307511

The Target Variable Pie chart shows that almost 92% of the total clients had no problem during payment while 8% of the clients had some or the other problem

GENDER VARIABLE

Row Lab	Count of CODE_GENDER
F	202448
M	105059
XNA	4
Grand Tota	307511

rom the GENDER_VARIABLE pie char we can infer that almost 66% of the clients are female and 34% of the clients are Male
: 4 of the appicants have gender as X which can be ignored

- The missing data is handled and also the data with outlier is made and imbalance data have me cleared to make a accurate insights providing dataset was made.

4) 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.

- **Task:** Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

UNIVARIATE ANALYSIS

Univariate analysis is the simplest form of analyzing data, where the data has only one variable. It involves analyzing each variable separately, without considering the relationships between variables.

SEGMENTED UNIVARIATE ANALYSIS

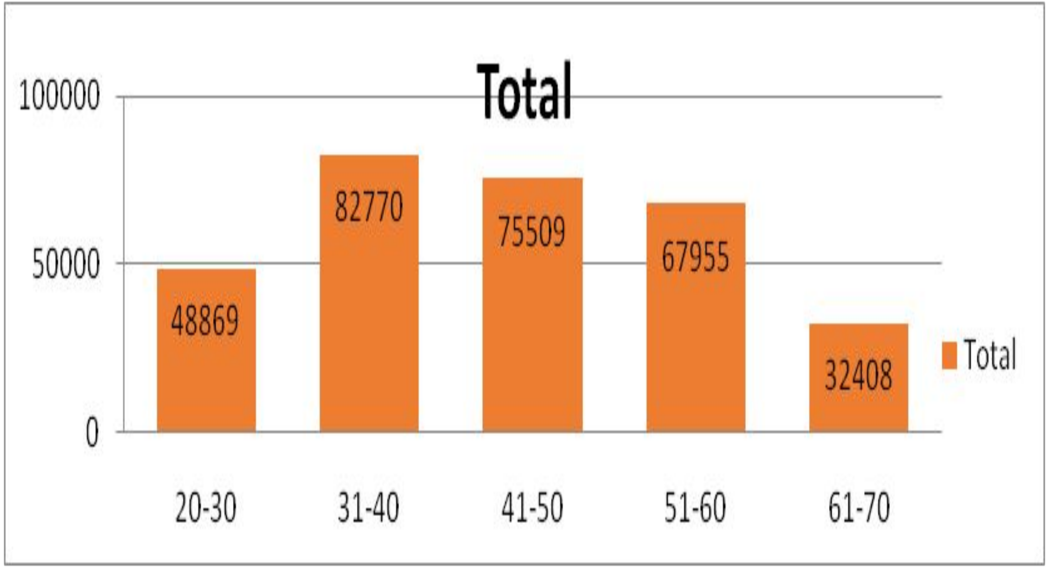
Segmented univariate analysis can be used to find summary of a single data variable in the form of segments. It also used to detect the central tendencies such as mean,median,mode; variance and standard deviation.

BIVARIATE ANALYSIS

Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them. Bivariate analysis can be helpful in testing simple hypotheses of x and y.

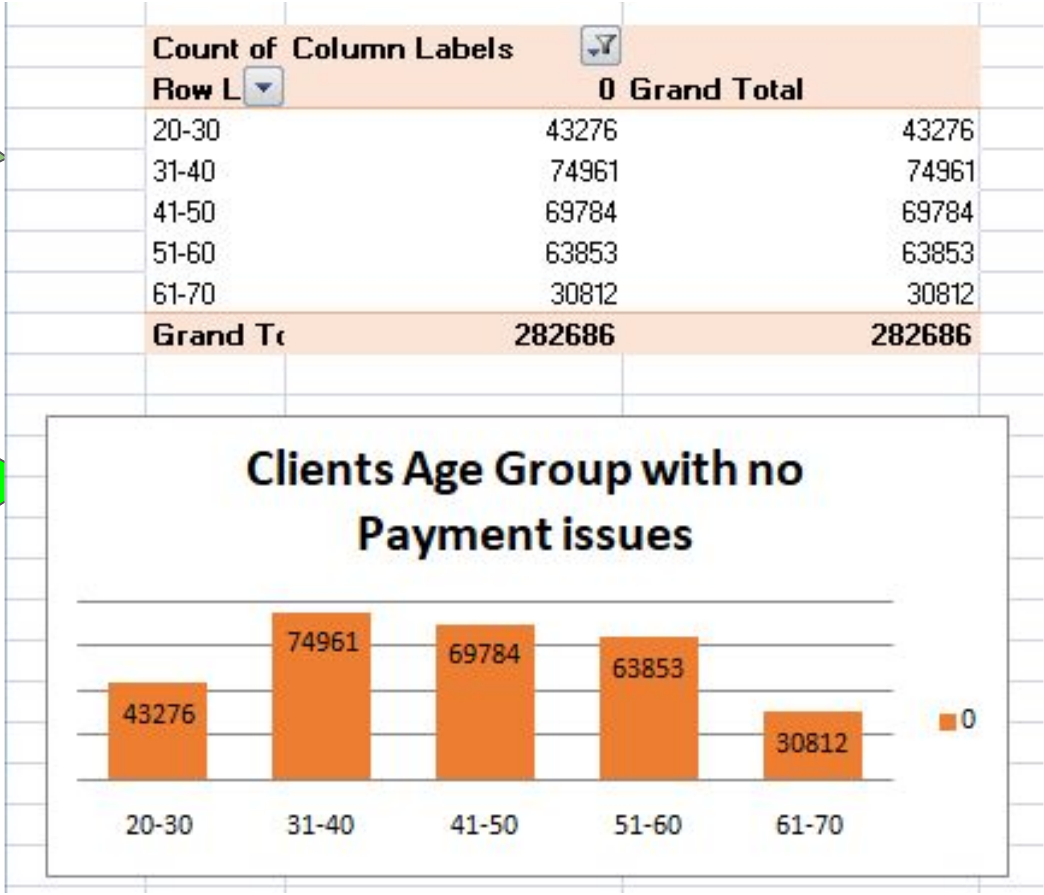
Univariate Analysis

Row Labels	Count of YEARS_BIRTH_RANGE
20-30	48869
31-40	82770
41-50	75509
51-60	67955
61-70	32408
Grand Total	307511



From the adjacent bar plot we can infer that most of the applicants belong to the Age Group ’31-40’.

AGE GROUP

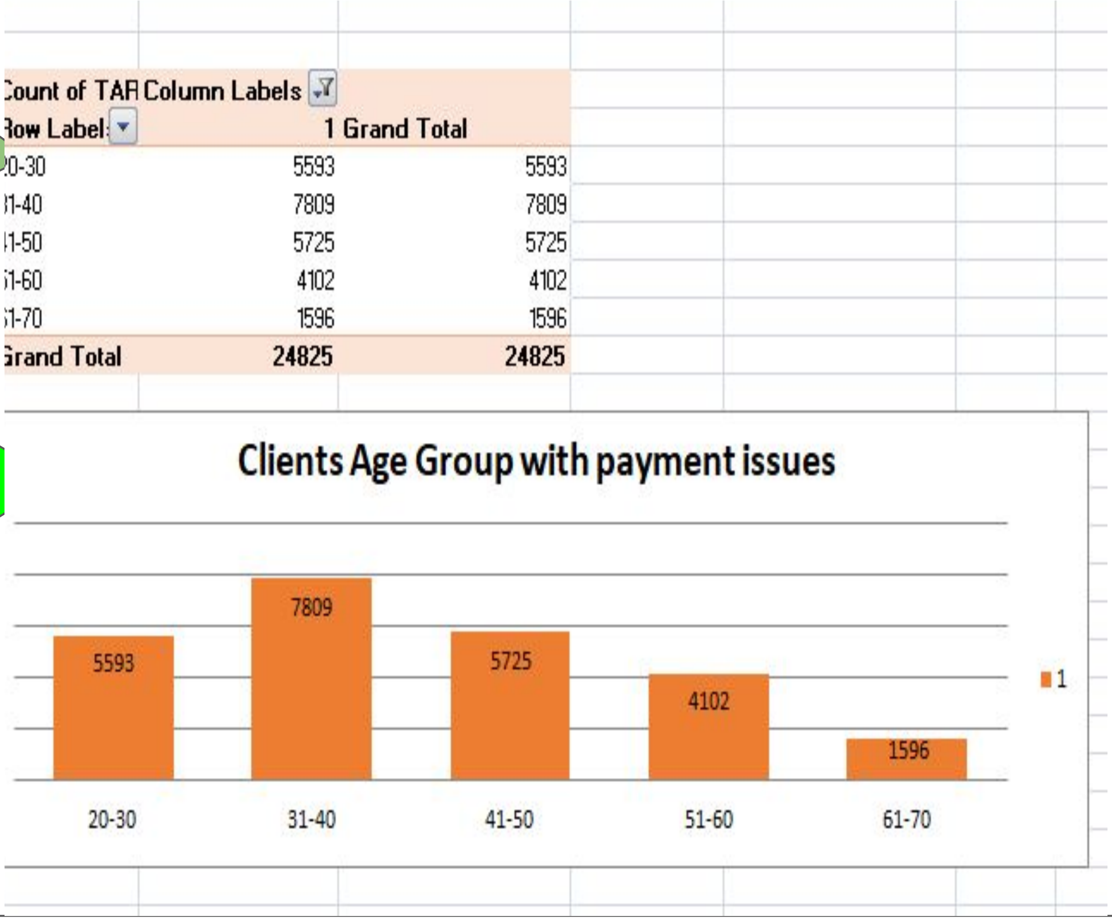


Count of	Column Labels	
Row L	0	Grand Total
20-30	43276	43276
31-40	74961	74961
41-50	69784	69784
51-60	63853	63853
61-70	30812	30812
Grand Tc	**282686**	**282686**

Clients Age Group with no Payment issues

From the adjacent Bar plot we can infer that clients/applicants in the Age Group ’31-40’ are having the highest number when it comes to doing/returning Payment to Banks

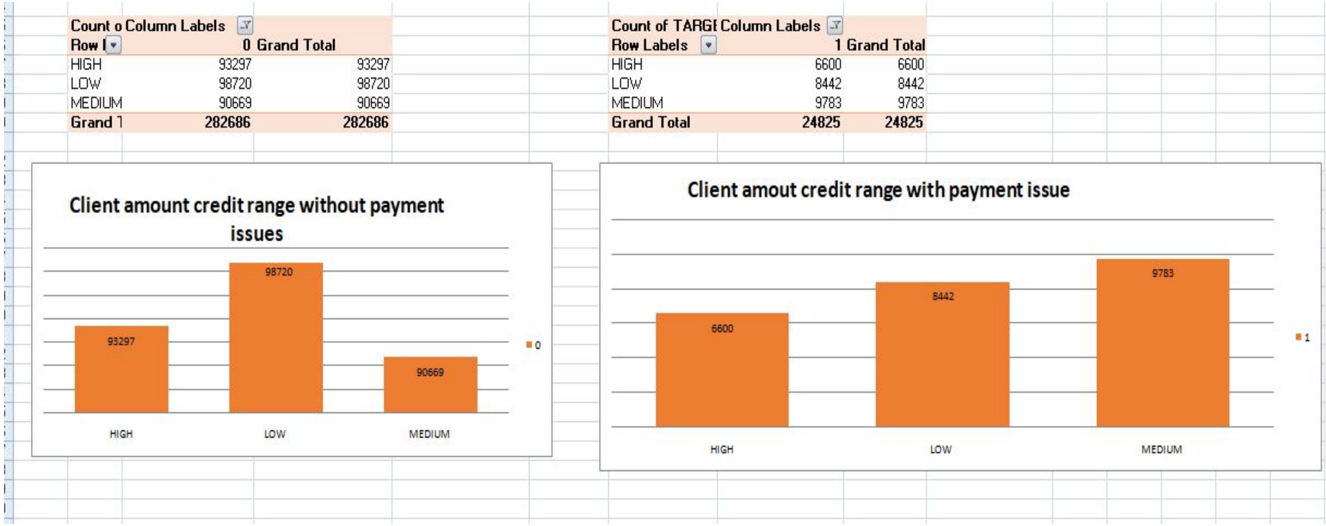
AGE GROUP



ount of TAR	Column Labels	
ow Label	1	Grand Total
0-30	5593	5593
1-40	7809	7809
1-50	5725	5725
1-60	4102	4102
1-70	1596	1596
rand Total	**24825**	**24825**

From the adjacent Bar plot we can infer that clients/applicants in the Age Group ’31-40’ are having the highest number of payment issues when it comes to doing/returning Payment to Banks

Client amount credit range



Count o	Column Labels	
Row l	0	Grand Total
HIGH	93297	93297
LOW	98720	98720
MEDIUM	90669	90669
Grand T	282686	282686

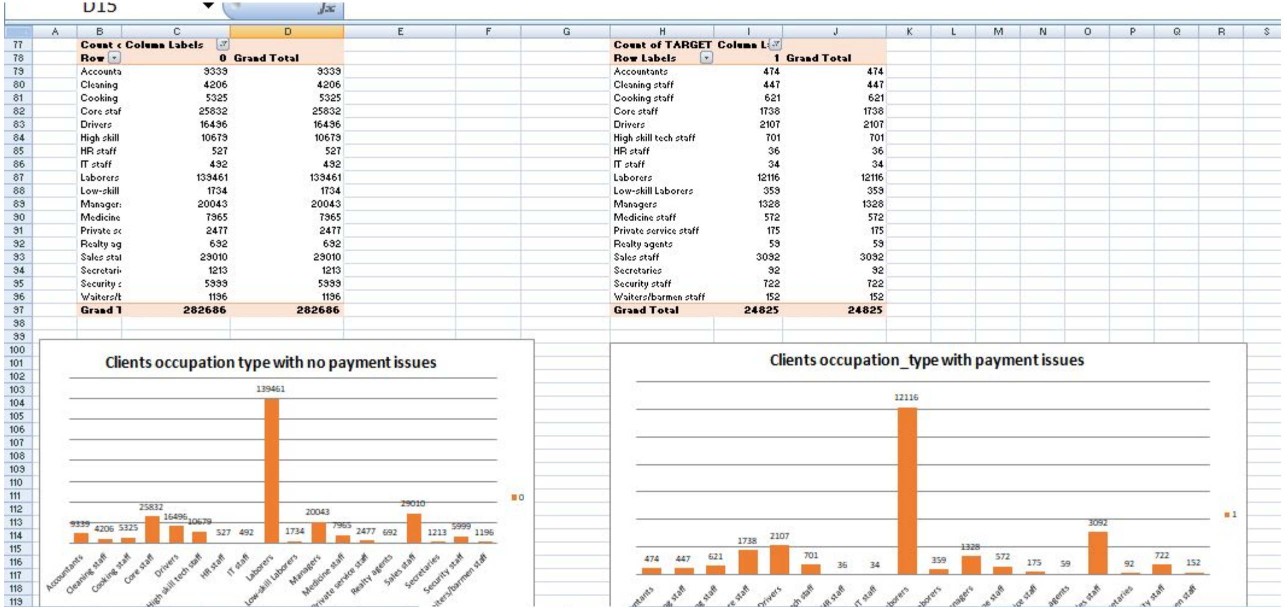
Count of TARGI	Column Labels	
Row Labels	1	Grand Total
HIGH	6600	6600
LOW	8442	8442
MEDIUM	9783	9783
Grand Total	24825	24825

Client amount credit range without payment issues
93297
98720
90669
HIGH
LOW
MEDIUM
0
Client amout credit range with payment issue
6600
8442
9783
HIGH
LOW
MEDIUM
1

From the adjacent Bar plot we can infer that clients belonging to 'Low' income range have the highest count when it comes to clients with no payment issues

From the adjacent Bar plot we can infer that clients belonging to 'Medium' income range have the highest count when it comes to clients with payment issues

OCCUPATION_TYPE



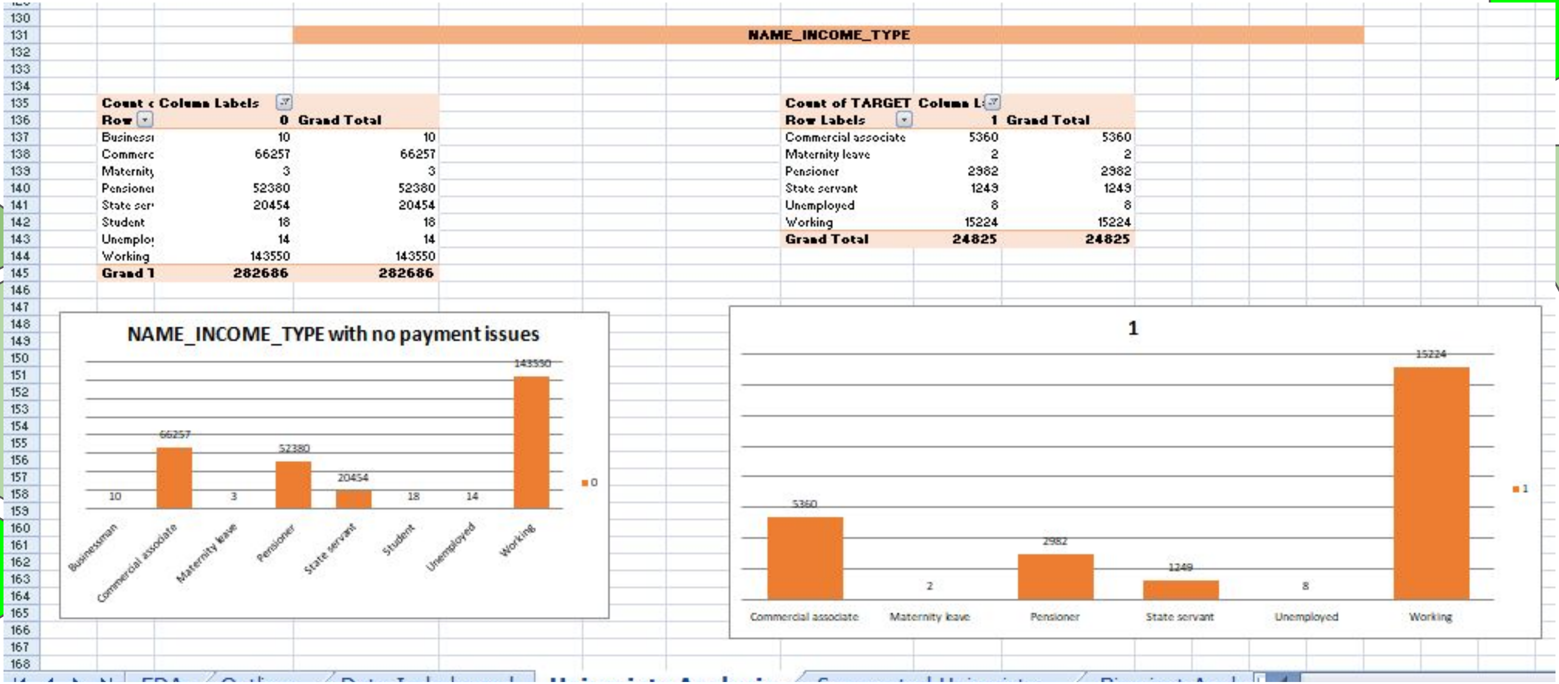
Count c	Column Labels	
Row	0	Grand Total
Accounta	9339	9339
Cleaning	4206	4206
Cooking	5325	5325
Core staf	25832	25832
Drivers	16496	16496
High skill	10679	10679
HR staff	527	527
IT staff	492	492
Laborers	139461	139461
Low-skill	1734	1734
Manager:	20043	20043
Medicine	7965	7965
Private sc	2477	2477
Realty ag	692	692
Sales stal	29010	29010
Secretari	1213	1213
Security :	5999	5999
Waiters/t	1196	1196
Grand T	**282686**	**282686**

Count of TARGET	Column L	
Row Labels	1	Grand Total
Accountants	474	474
Cleaning staff	447	447
Cooking staff	621	621
Core staff	1738	1738
Drivers	2107	2107
High skill tech staff	701	701
HR staff	36	36
IT staff	34	34
Laborers	12116	12116
Low-skill Laborers	359	359
Managers	1328	1328
Medicine staff	572	572
Private service staff	175	175
Realty agents	59	59
Sales staff	3092	3092
Secretaries	92	92
Security staff	722	722
Waiters/barmen staff	152	152
Grand Total	**24825**	**24825**

Clients occupation type with no payment issues

Clients occupation_type with payment issues

- From the above bar plot we can infer that clients with occupation_type 'Laborers' have the highest number of count when it comes to clients with no payment issues
- From the above bar plot we can infer that clients with occupation_type 'Laborers' have the highest number of count when it comes to clients with payment issues



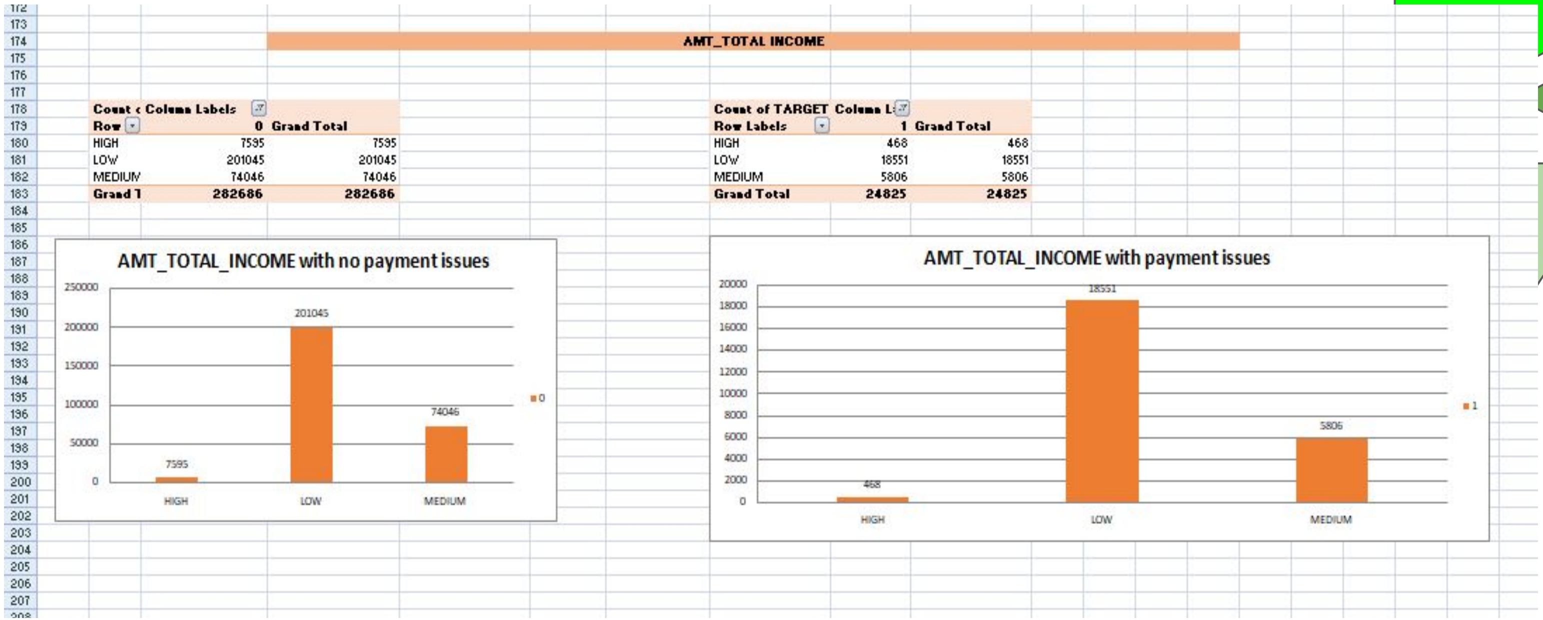
NAME_INCOME_TYPE

Count c	Column Labels	
Row	0	Grand Total
Businessı	10	10
Commerc	66257	66257
Maternity	3	3
Pensionei	52380	52380
State ser	20454	20454
Student	18	18
Unemplo	14	14
Working	143550	143550
Grand T	**282686**	**282686**

Count of TARGET	Column L	
Row Labels	1	Grand Total
Commercial associate	5360	5360
Maternity leave	2	2
Pensioner	2982	2982
State servant	1249	1249
Unemployed	8	8
Working	15224	15224
Grand Total	**24825**	**24825**

– From the above Bar plot we can infer that clients having income_type as 'WORKING' have the highest count when it comes to clients with no payment issues

– From the above Bar plot we can infer that clients having income_type as 'WORKING' have the highest count when it comes to clients with payment issues



AMT_TOTAL INCOME

Count c	Column Labels	
Row	0	Grand Total
HIGH	7595	7595
LOW	201045	201045
MEDIUM	74046	74046
Grand T	**282686**	**282686**

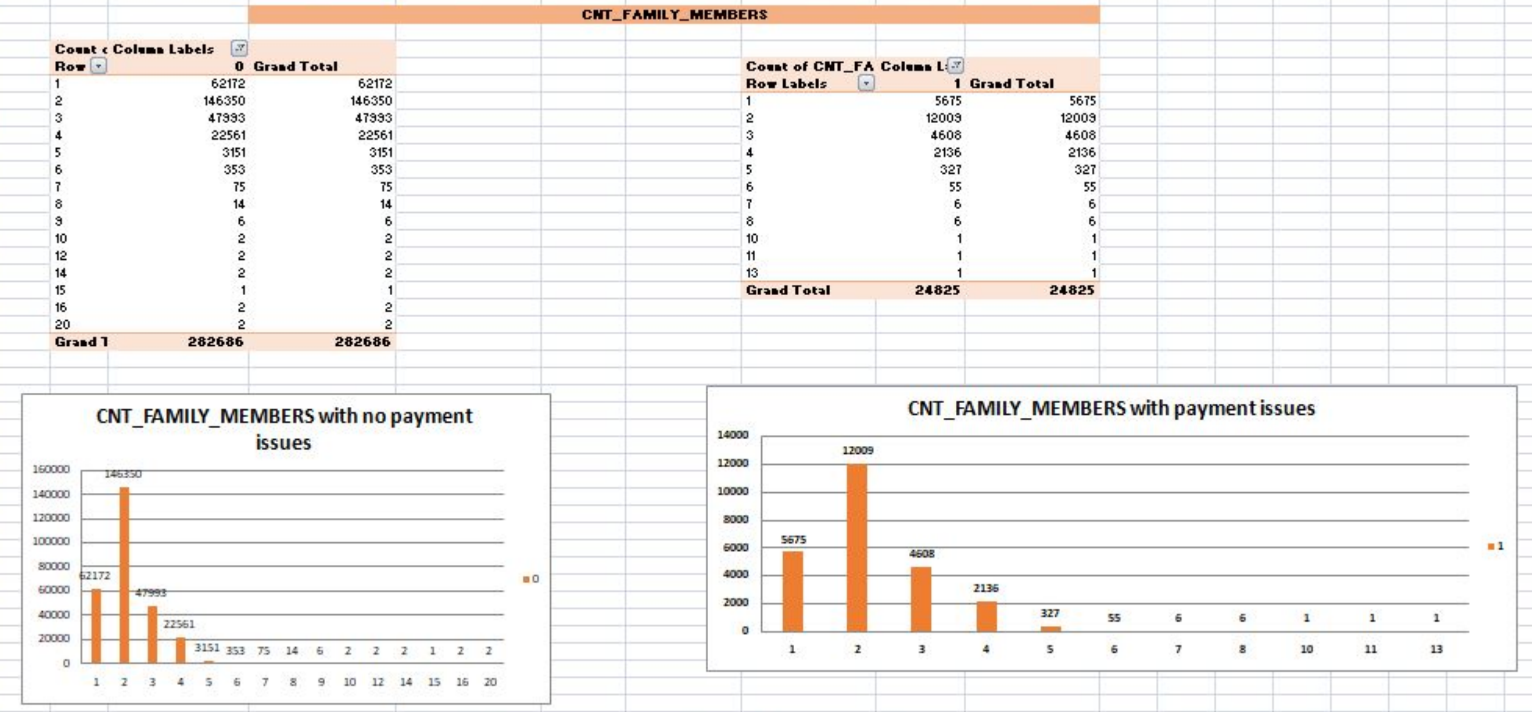
Count of TARGET	Column L	
Row Labels	1	Grand Total
HIGH	468	468
LOW	18551	18551
MEDIUM	5806	5806
Grand Total	**24825**	**24825**

AMT_TOTAL_INCOME with no payment issues

AMT_TOTAL_INCOME with payment issues

– **From the above Bar plot we can infer that client having the total income range as ‘LOW’ have the highest count when it comes to clients having no payment issues**

– **From the above Bar plot we can infer that client having the total income range as ‘LOW’ have the highest count when it comes to clients having payment issues**



CNT_FAMILY_MEMBERS

Count c	Column Labels	
Row ▾	0	Grand Total
1	62172	62172
2	146350	146350
3	47993	47993
4	22561	22561
5	3151	3151
6	353	353
7	75	75
8	14	14
9	6	6
10	2	2
12	2	2
14	2	2
15	1	1
16	2	2
20	2	2
Grand T	282686	282686

Count of CNT_FA	Column L	
Row Labels ▾	1	Grand Total
1	5675	5675
2	12009	12009
3	4608	4608
4	2136	2136
5	327	327
6	55	55
7	6	6
8	6	6
10	1	1
11	1	1
13	1	1
Grand Total	24825	24825

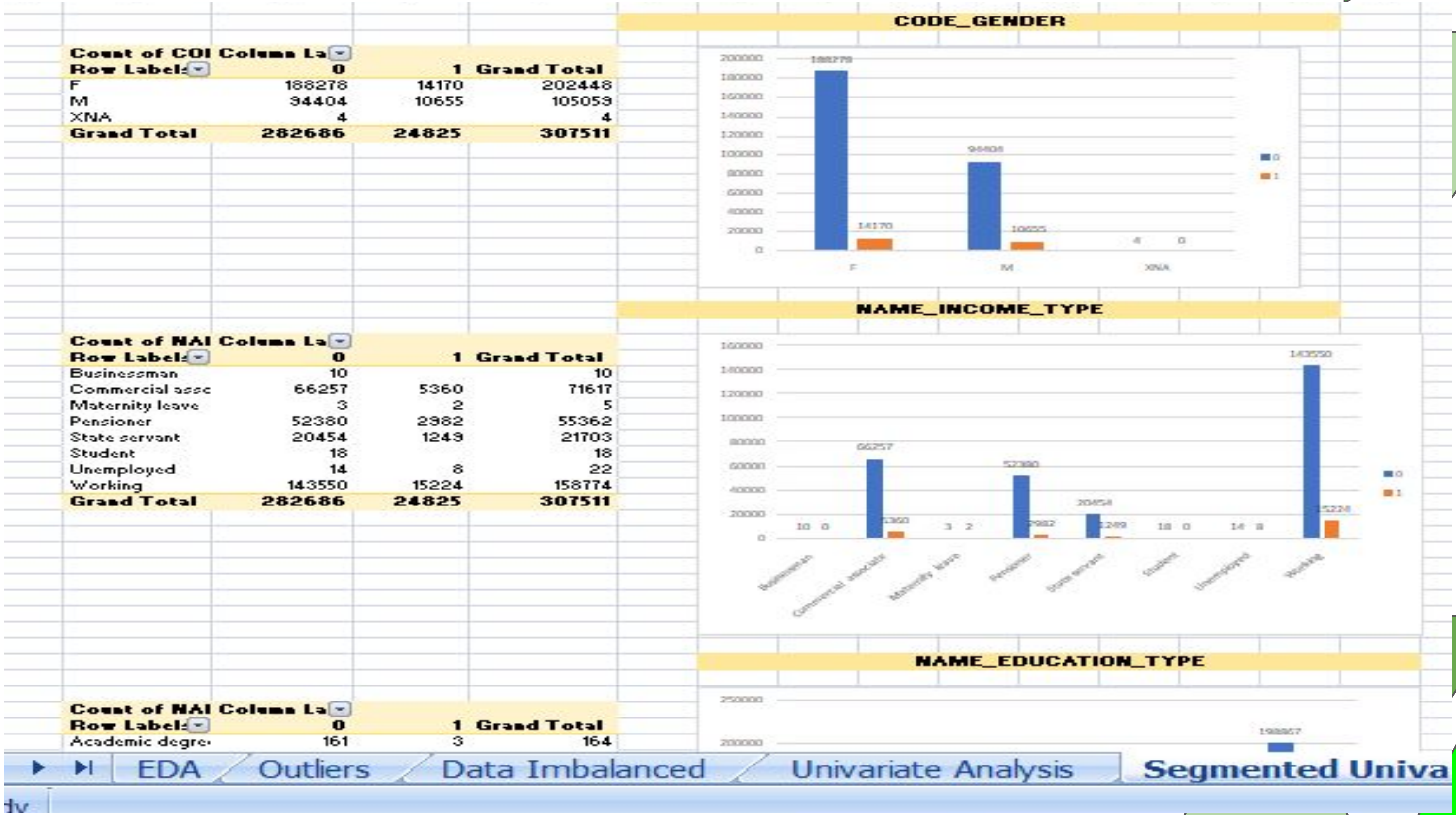
CNT_FAMILY_MEMBERS with no payment issues

CNT_FAMILY_MEMBERS with payment issues

– From the above Bar plot we can infer that clients having total count of family members as 2 have the highest count when it comes to clients having no payment issues

– From the above Bar plot we can infer that clients having total count of family members as 2 have the highest count when it comes to clients having payment issues

Segmented Univariate Analysis:

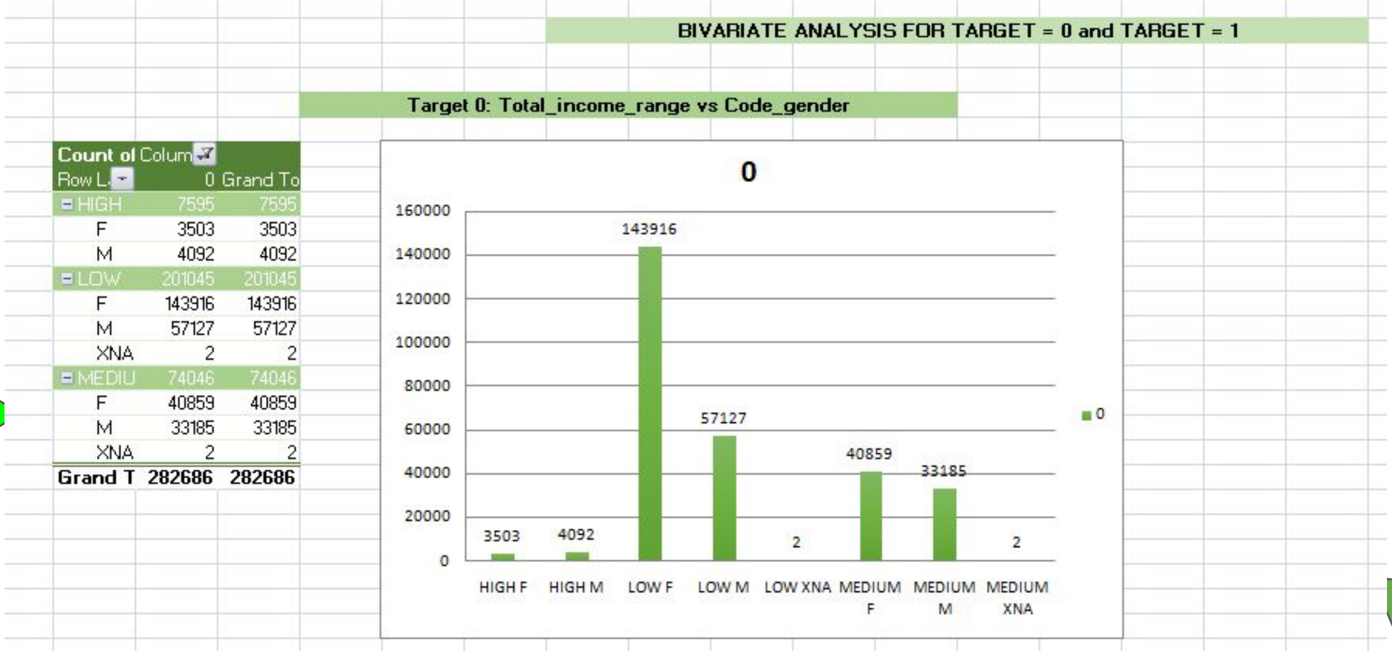


Count of COI Column La			
Row Label:	0	1	Grand Total
F	188278	14170	202448
M	94404	10655	105059
XNA	4		4
Grand Total	**282686**	**24825**	**307511**

Count of NAI Column La			
Row Label:	0	1	Grand Total
Businessman	10		10
Commercial assc	66257	5360	71617
Maternity leave	3	2	5
Pensioner	52380	2982	55362
State servant	20454	1249	21703
Student	18		18
Unemployed	14	8	22
Working	143550	15224	158774
Grand Total	**282686**	**24825**	**307511**

Count of NAI Column La			
Row Label:	0	1	Grand Total
Academic degre	161	3	164

Bivariate Analysis :

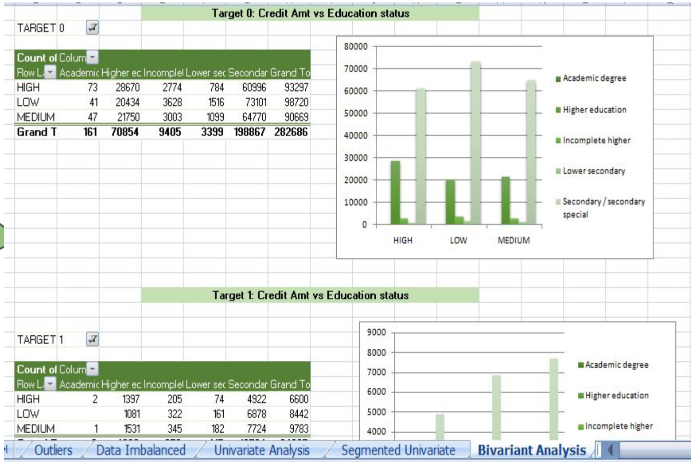


BIVARIATE ANALYSIS FOR TARGET = 0 and TARGET = 1

Target 0: Total_income_range vs Code_gender

Count of Colum		
Row L.	0	Grand To
⊟ HIGH	7595	7595
F	3503	3503
M	4092	4092
⊟ LOW	201045	201045
F	143916	143916
M	57127	57127
XNA	2	2
⊟ MEDIU	74046	74046
F	40859	40859
M	33185	33185
XNA	2	2
Grand T	**282686**	**282686**

From the above Bar plot we can infer that Females belonging to Low income group are the highest number of clients with no payment issues



Target 0: Credit Amt vs Education status

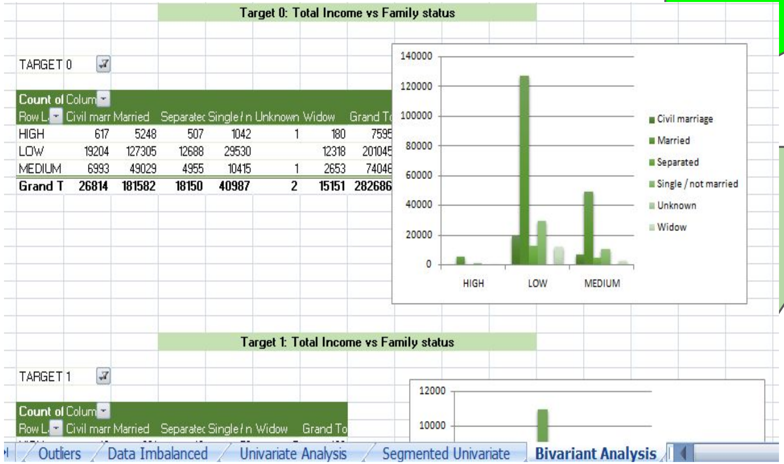
TARGET 0

Count of Colum						
Row L	Academic	Higher ec	Incomplet	Lower sec	Secondar	Grand To
HIGH	73	28670	2774	784	60996	93297
LOW	41	20434	3628	1516	73101	98720
MEDIUM	47	21750	3003	1099	64770	90669
Grand T	161	70854	9405	3399	198867	282686

Target 1: Credit Amt vs Education status

TARGET 1

Count of Colum						
Row L	Academic	Higher ec	Incomplet	Lower sec	Secondar	Grand To
HIGH	2	1397	205	74	4922	6600
LOW		1081	322	161	6878	8442
MEDIUM	1	1531	345	182	7724	9783



Target 0: Total Income vs Family status

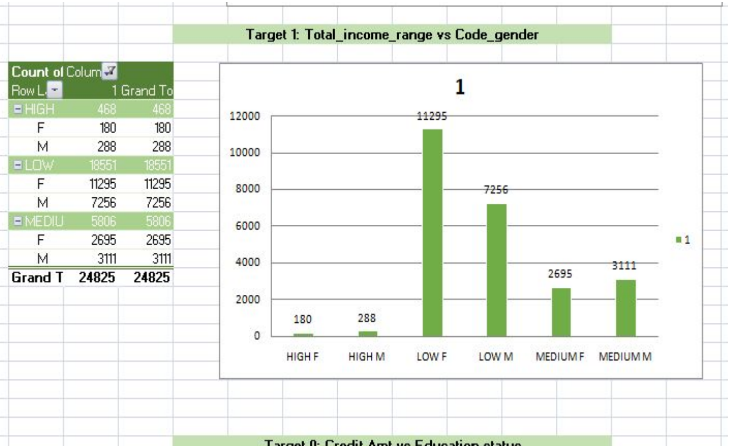
TARGET 0

Count of Colum							
Row L	Civil marr	Married	Separatec	Single / n	Unknown	Widow	Grand To
HIGH	617	5248	507	1042	1	180	7595
LOW	19204	127305	12688	29530		12318	201045
MEDIUM	6993	49029	4955	10415	1	2653	74046
Grand T	26814	181582	18150	40987	2	15151	282686

Target 1: Total Income vs Family status

TARGET 1

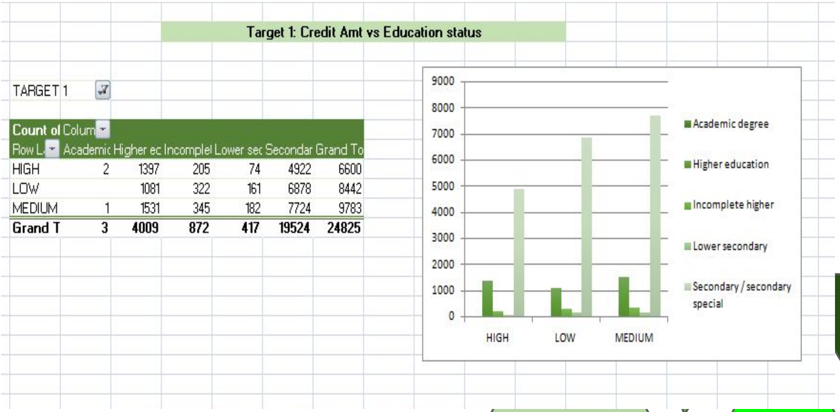
Count of Colum						
Row L	Civil marr	Married	Separatec	Single / n	Widow	Grand To



Target 1: Total_income_range vs Code_gender

Count of Colum		
Row L	1	Grand To
⊟ HIGH	468	468
F	180	180
M	288	288
⊟ LOW	18551	18551
F	11295	11295
M	7256	7256
⊟ MEDIU	5806	5806
F	2695	2695
M	3111	3111
Grand T	24825	24825

Target 0: Credit Amt vs Education status



Target 1: Credit Amt vs Education status

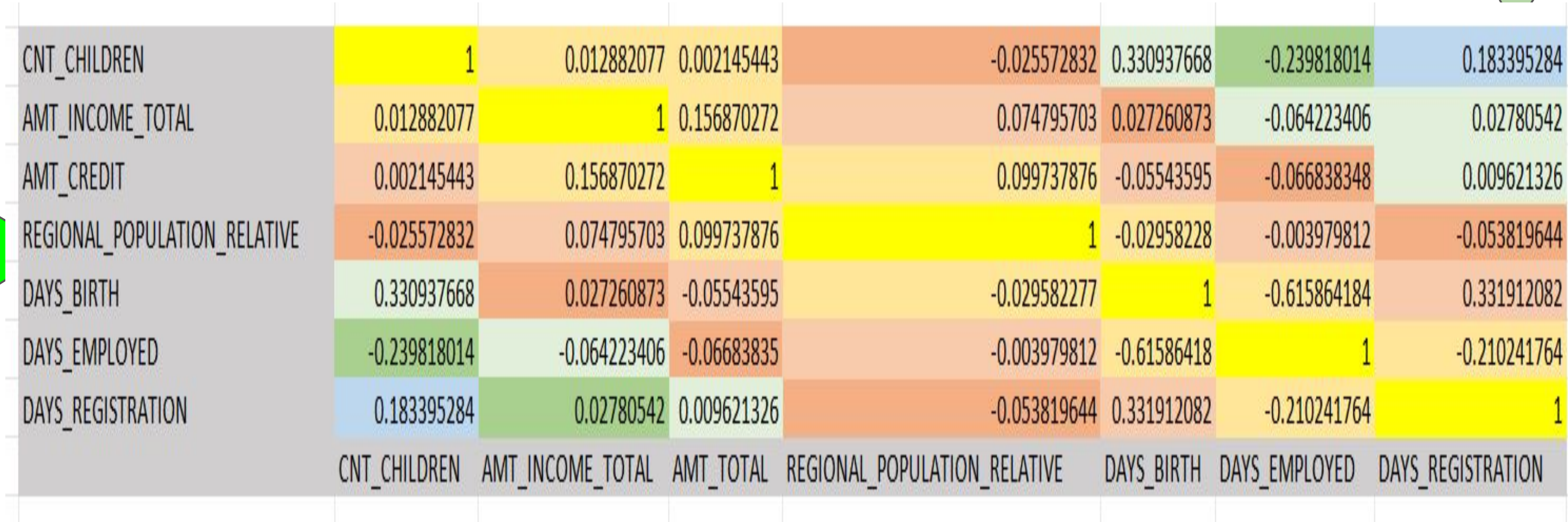
TARGET 1

Count of Colum						
Row L	Academic	Higher ec	Incomplet	Lower sec	Secondar	Grand To
HIGH	2	1397	205	74	4922	6600
LOW		1081	322	161	6878	8442
MEDIUM	1	1531	345	182	7724	9783
Grand T	3	4009	872	417	19524	24825

5) Identify Top Correlations for Different Scenarios: Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

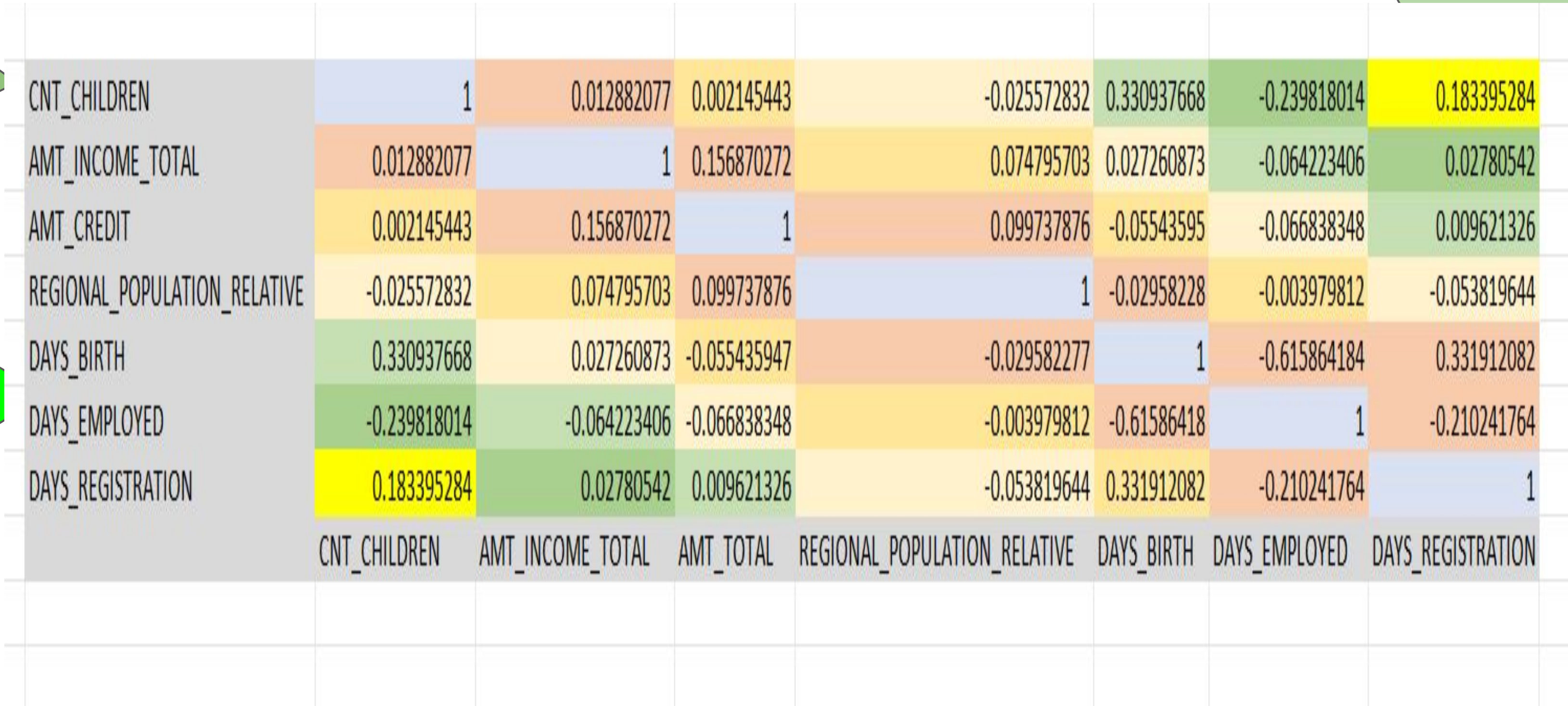
- **Task:** Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

Correlation for target 0



CNT_CHILDREN	1	0.012882077	0.002145443	-0.025572832	0.330937668	-0.239818014	0.183395284
AMT_INCOME_TOTAL	0.012882077	1	0.156870272	0.074795703	0.027260873	-0.064223406	0.02780542
AMT_CREDIT	0.002145443	0.156870272	1	0.099737876	-0.05543595	-0.066838348	0.009621326
REGIONAL_POPULATION_RELATIVE	-0.025572832	0.074795703	0.099737876	1	-0.02958228	-0.003979812	-0.053819644
DAYS_BIRTH	0.330937668	0.027260873	-0.05543595	-0.029582277	1	-0.615864184	0.331912082
DAYS_EMPLOYED	-0.239818014	-0.064223406	-0.06683835	-0.003979812	-0.61586418	1	-0.210241764
DAYS_REGISTRATION	0.183395284	0.02780542	0.009621326	-0.053819644	0.331912082	-0.210241764	1
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_TOTAL	REGIONAL_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION

Correlation for Target 1



CNT_CHILDREN	1	0.012882077	0.002145443	-0.025572832	0.330937668	-0.239818014	0.183395284
AMT_INCOME_TOTAL	0.012882077	1	0.156870272	0.074795703	0.027260873	-0.064223406	0.02780542
AMT_CREDIT	0.002145443	0.156870272	1	0.099737876	-0.05543595	-0.066838348	0.009621326
REGIONAL_POPULATION_RELATIVE	-0.025572832	0.074795703	0.099737876	1	-0.02958228	-0.003979812	-0.053819644
DAYS_BIRTH	0.330937668	0.027260873	-0.055435947	-0.029582277	1	-0.615864184	0.331912082
DAYS_EMPLOYED	-0.239818014	-0.064223406	-0.066838348	-0.003979812	-0.61586418	1	-0.210241764
DAYS_REGISTRATION	0.183395284	0.02780542	0.009621326	-0.053819644	0.331912082	-0.210241764	1
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_TOTAL	REGIONAL_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION

Result:

Hence the analysis are being done on both datasets Applications Dataset and Precious Applications Dataset
The following conclusions were drawn from the analysis done

- The proportion/percentage of the defaulters(target = 1) is around 8% and that of non-defaulters(target = 0) is around 92%

- The Bank generally lends more loan to Female clients as compared to Males clients as the count of Female clients in the defaulter's list is less than that of Males. Still Bank can look for more Male clients if their credit amount is satisfied

- Also the clients who belong to Working class tend to pay their loans on time followed by the clients who fall under Commercial Associate

- **Clients having Education status like Secondary/ Higher Secondary or more tend to pay loan on time so bank can prefer lending loans to clients having such Education Status**
- **Clients who fall in the Age Group 31-40 have the highest count for paying off their loans on time followed by the clients who fall in the Age Groups 41-60**
- **Clients having LOW credit amount range tend to pay off their loans on time than compared to HIGH and MEDIUM credit range**
- **Clients living with their Parents tend to pay off their loans quickly as compared to other housing type. So Bank can lend loan to clients having housing type** à **Living with Parents**

- **Clients taking loan for purchasing New Home i.e. clients taking Home Loans or purchasing New Car i.e. Car Loans and clients who have a income type as State Servant tend to pay their loans on time and hence Bank should prefer clients having such background**

- **The Bank should be more cautious when lending money to clients with Repairs purpose because they have high count of Defaulters along with High count of Defaulters**

Task 6 Workbook: application data.xlsx

Task 6 presentation : Presentation

Conclusion:

- Thus the solution to the given case study was found.
- I have able to create decision and do analysis on a large dataset.
- It involved EDA method and also you data visualization technique to provide a understandable graphs that makes it easy for understandable and every step made me to get strong knowledge in MS Excel.