BANK LOAN CASE STUDY

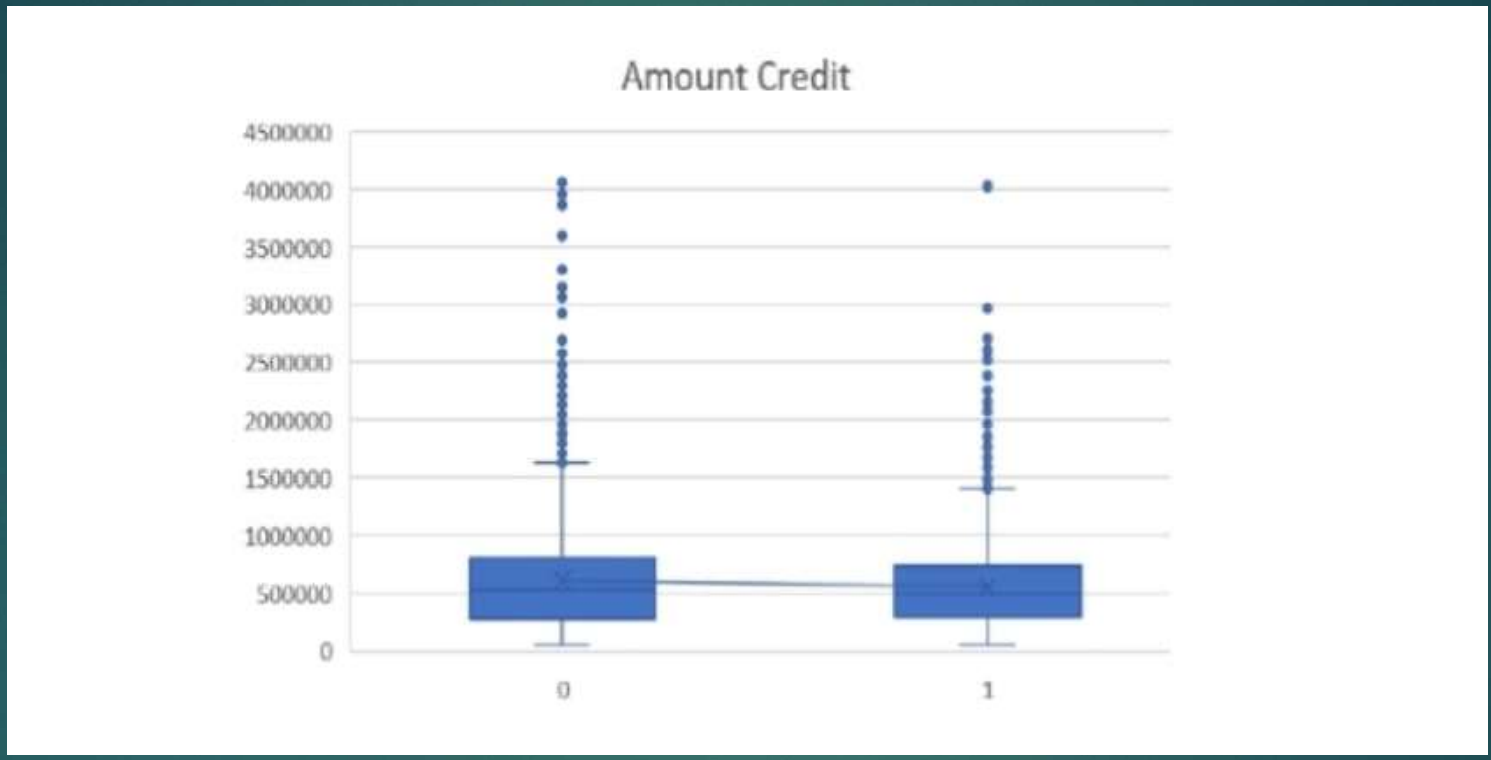
BY:-MD ANJAR

- A. **Identify Missing Data and Deal with it Appropriately:-** the Existing applications sheet included 161 columns
- I deleted columns with more than 5% blank data.
- I deleted a large number of useless columns.
- To eliminates blanks values , I used the COUNTBLANK function.

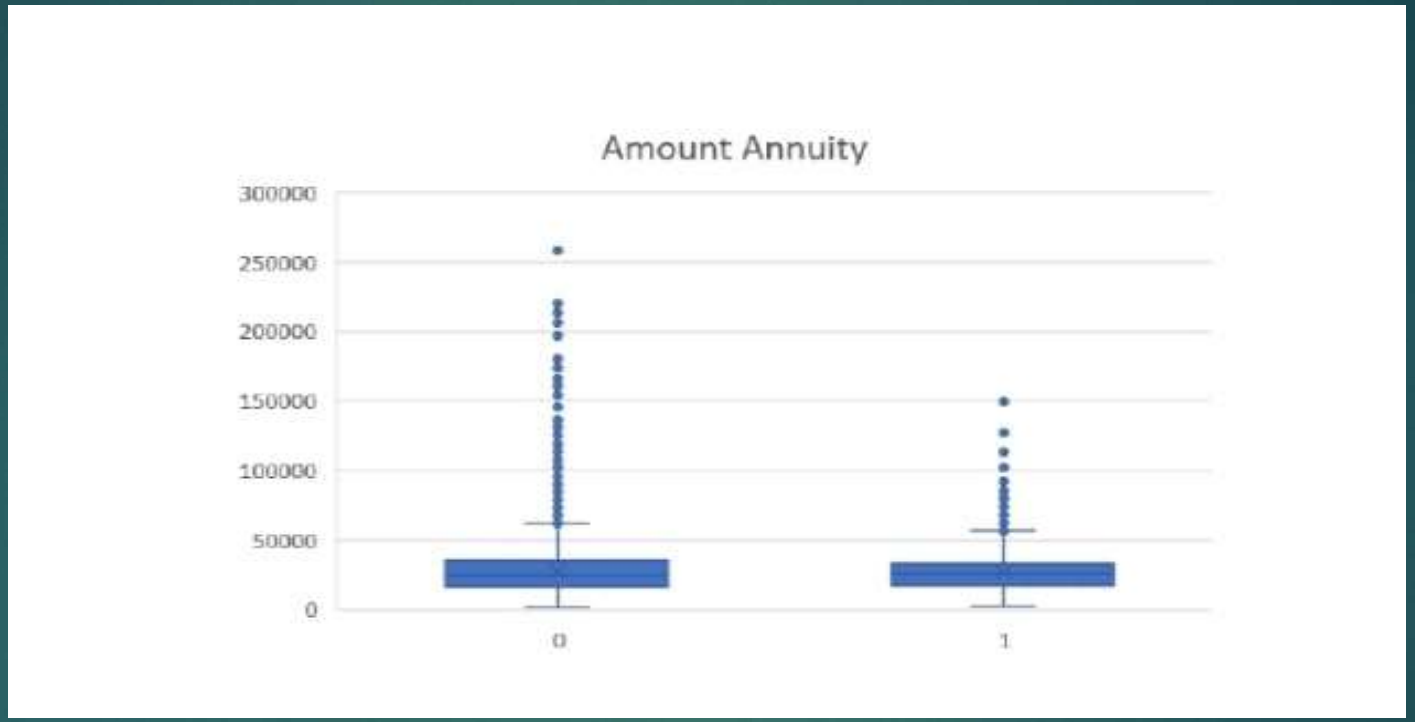
B. Identify Outliers in the Dataset:-outliers can only be identified on Numeric variables.

Box plotted Target Column Vs

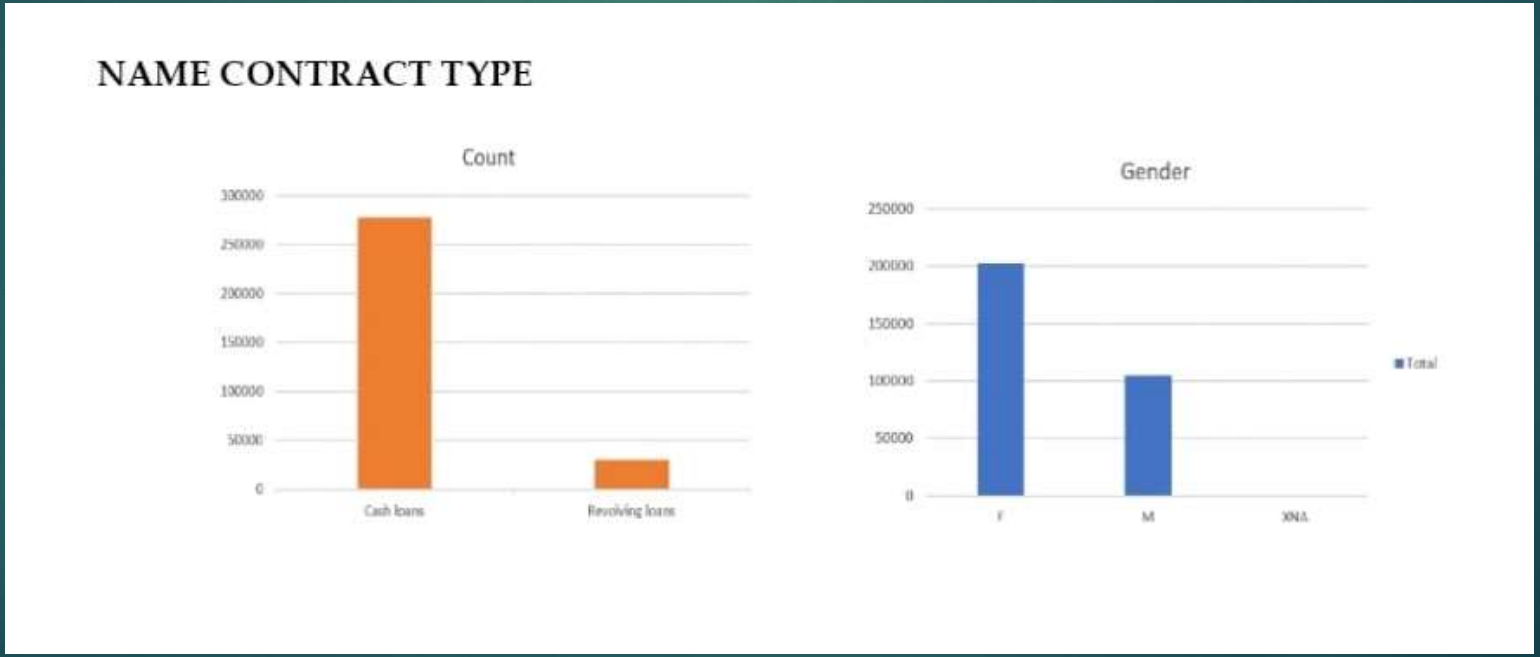
1) Amount credit
2) Amount Income
3) Amount Annuity

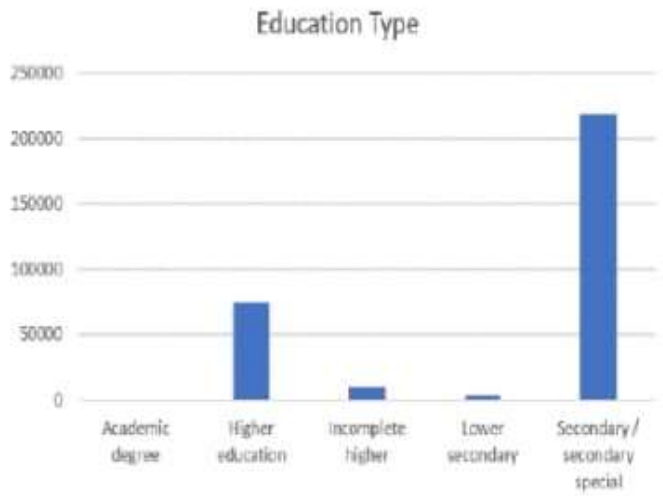

Amount Credit
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
1

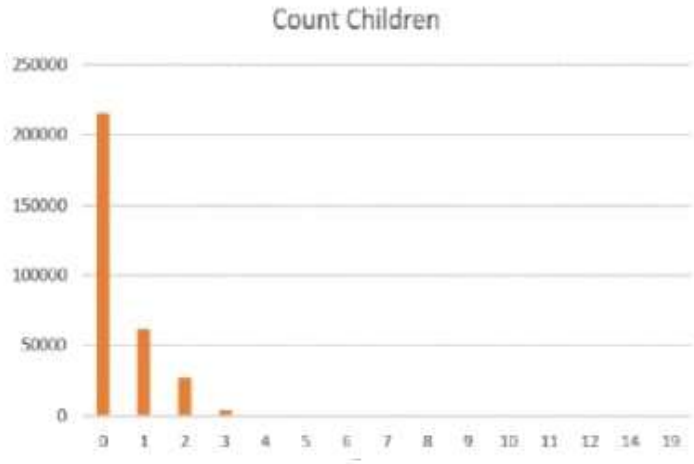

Amount Income
1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
0
1

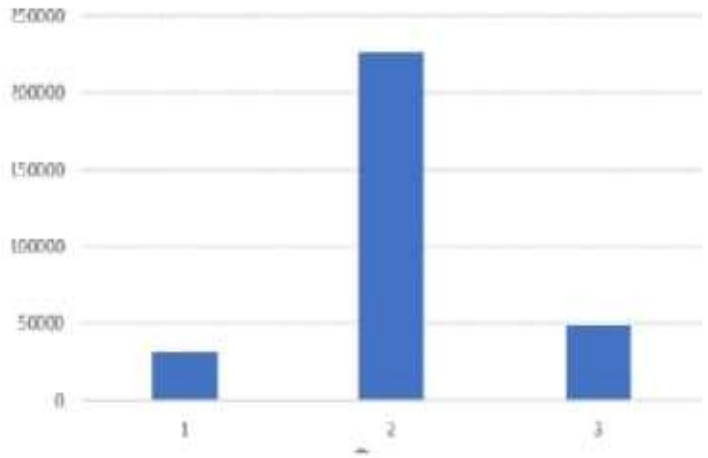

Amount Annuity
300000
250000
200000
150000
100000
50000
0
0
1

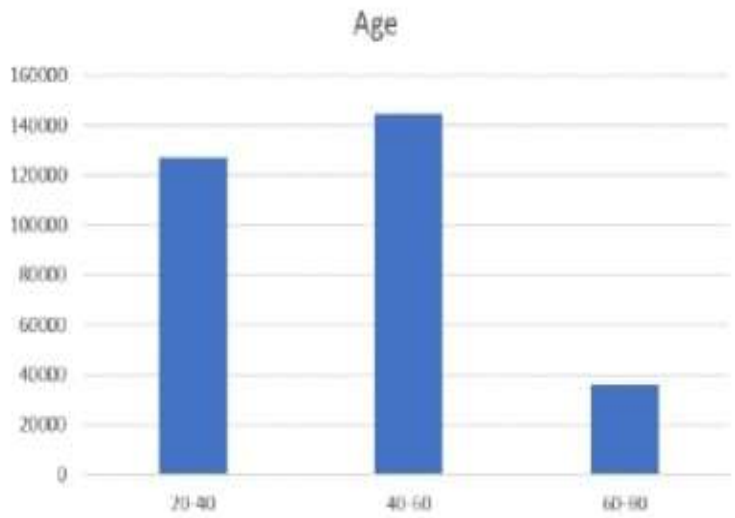
- **C. Analyze Data Imbalance:-** Data imbalance occurs when data is disseminated in an unequal manner. I plotted data imbalance using pivot charts.
- I used the Excel Function like COUNTIF and SUM To calculate the proportions of each class in the given datasets.




Education Type
250000
200000
150000
100000
50000
0
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special


Count Children
250000
200000
150000
100000
50000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
14
19


Region Rating
150000
100000
50000
0
1
2
3


Age
160000
140000
120000
100000
80000
60000
40000
20000
0
20-40
40-60
60-80

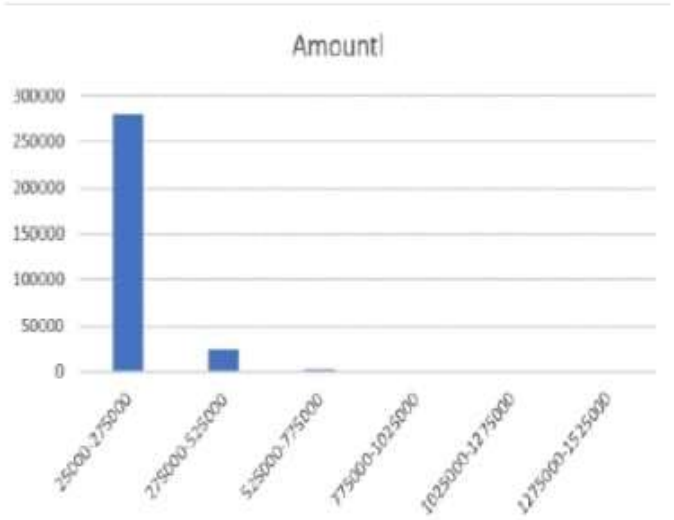
EDA

- **Perform Univariate Analysis:-**

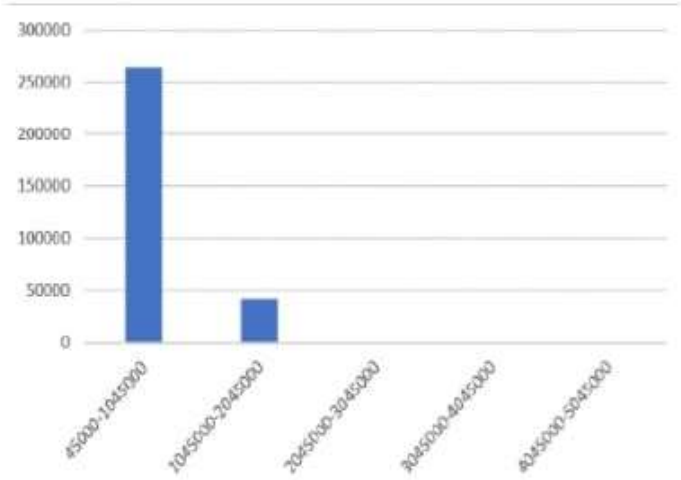
interference

Individuals with higher income are less likely to apply for loans. The credits amount of a bank loan is typically in the range of 45000 to 1045000. the majority of loan applications have come from people between the ages of 35 and 50. those with 0 to 8 years of work experiences are the most likely to seek for loans. Individuals who own homes are more likely to apply for loans than others. Those who are married have taken out more loans. More loans have been requested by working people. Unaccompanied minors have requested for extra loans.

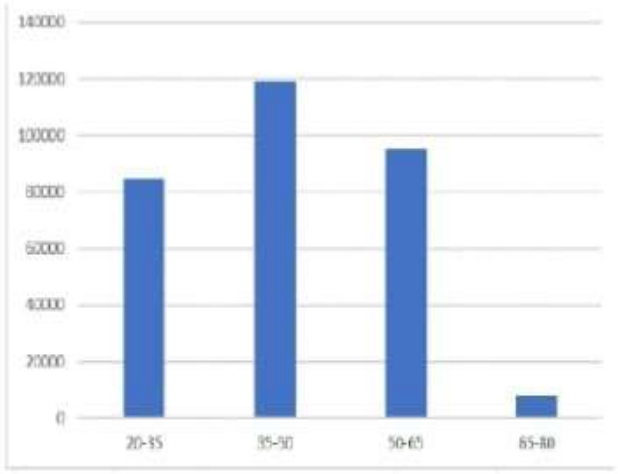
Amount Income

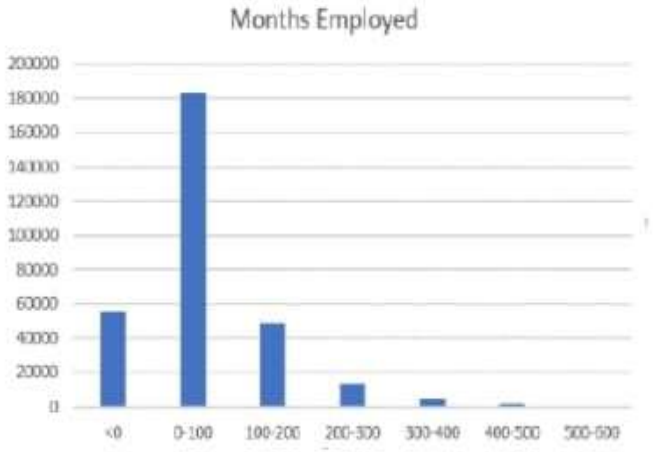


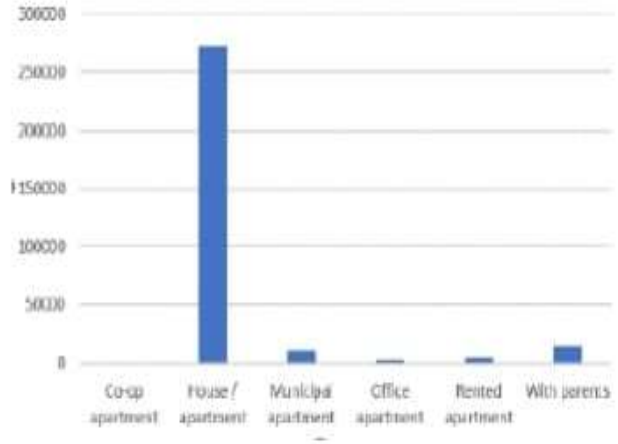
Amount Credit

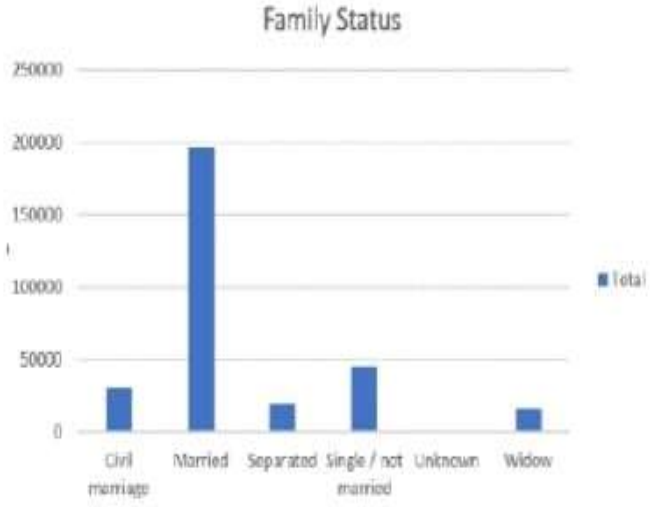


AGE

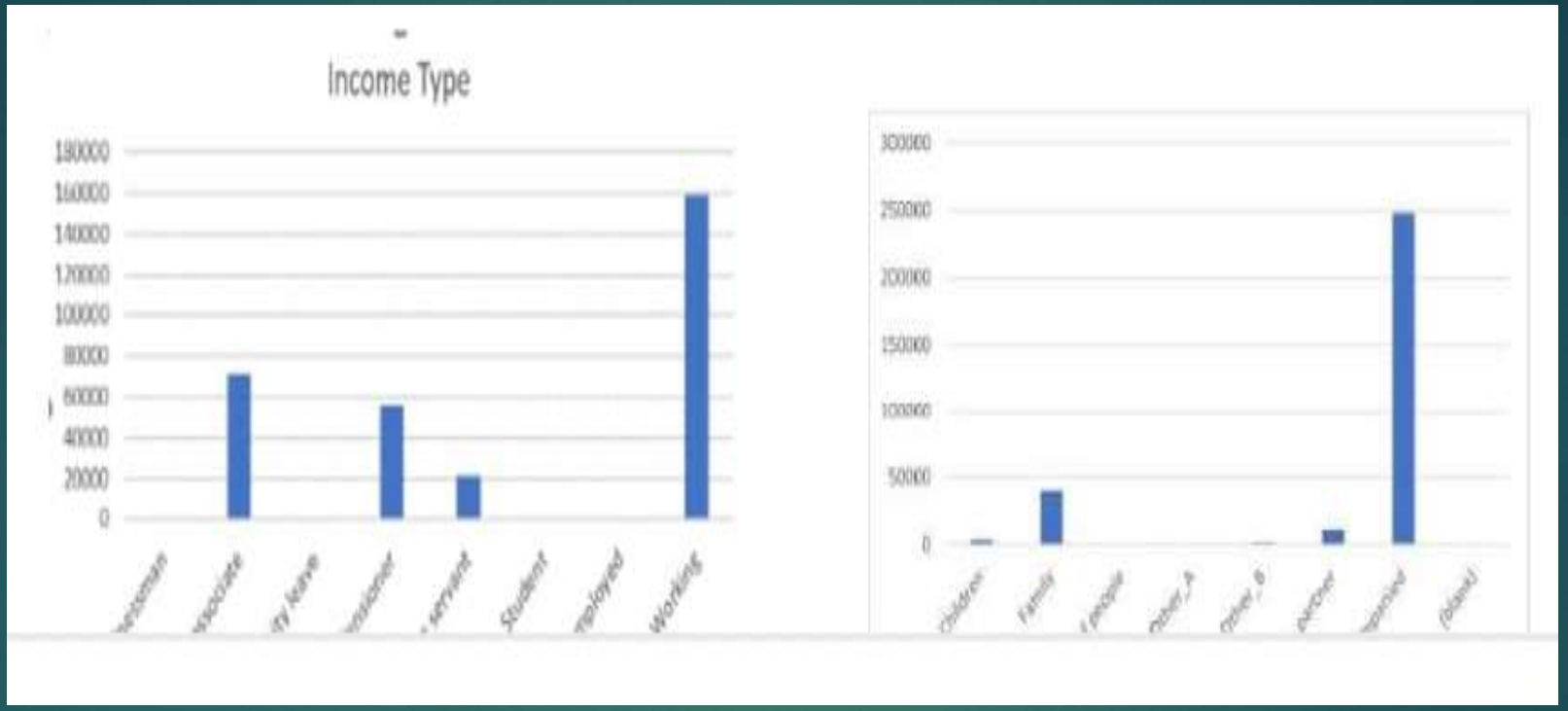

140000
120000
100000
80000
60000
40000
20000
0
20-35
35-50
50-65
65-80


Months Employed
200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
<0
0-100
100-200
200-300
300-400
400-500
500-600





Name suite type


Income Type
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
Student
Working
300000
250000
200000
150000
100000
50000
0

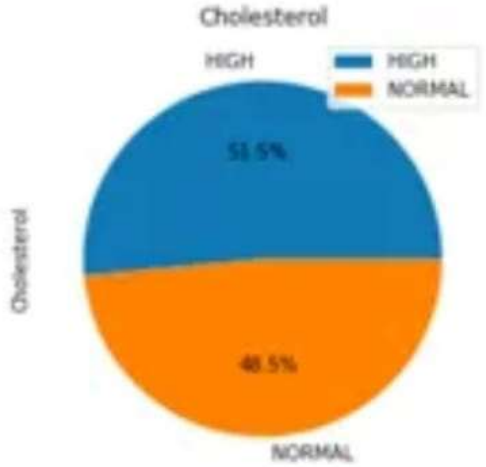
- **Segmented Univariate Analysis:-**

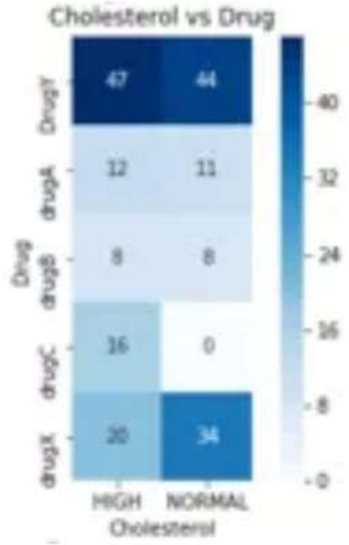
Interference

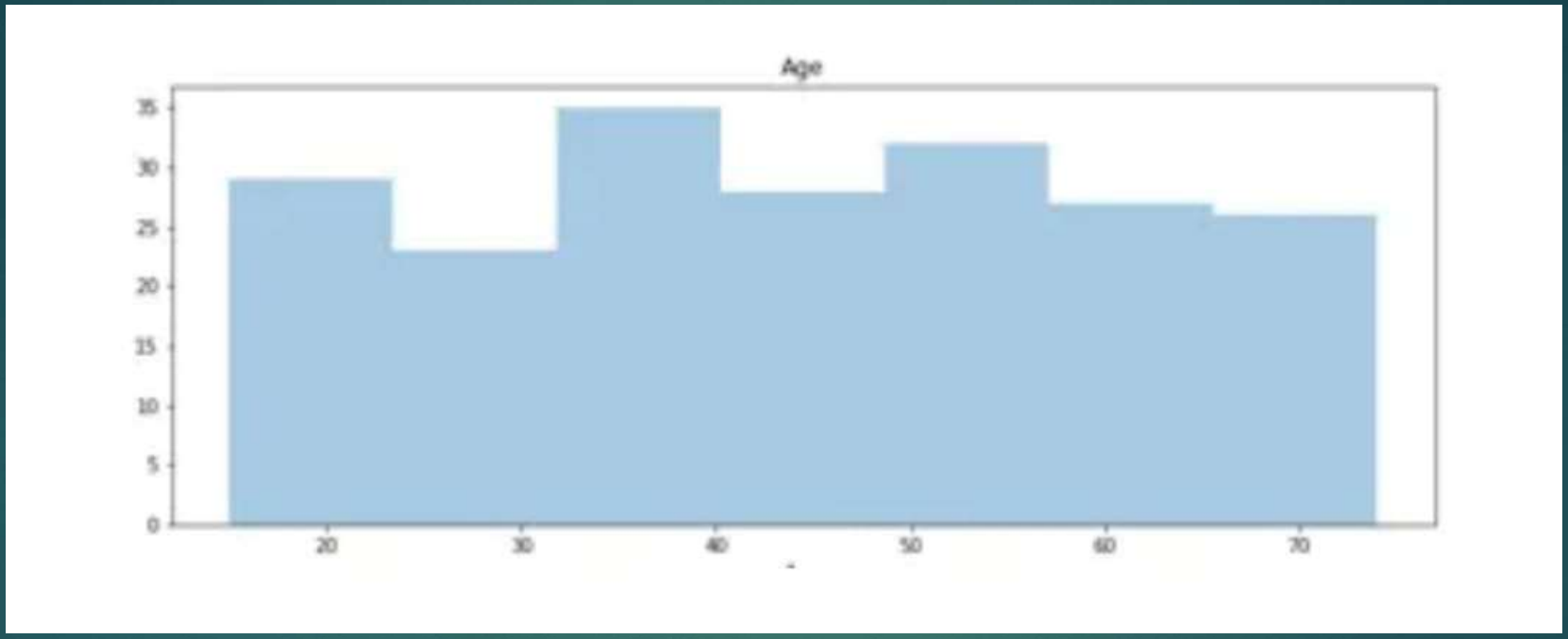
Segmented univariate analysis is one of the simplest form of visualizations to analyse data . In its name 'Uni' means one which itself describes that it considers only a single data variable for analysis.

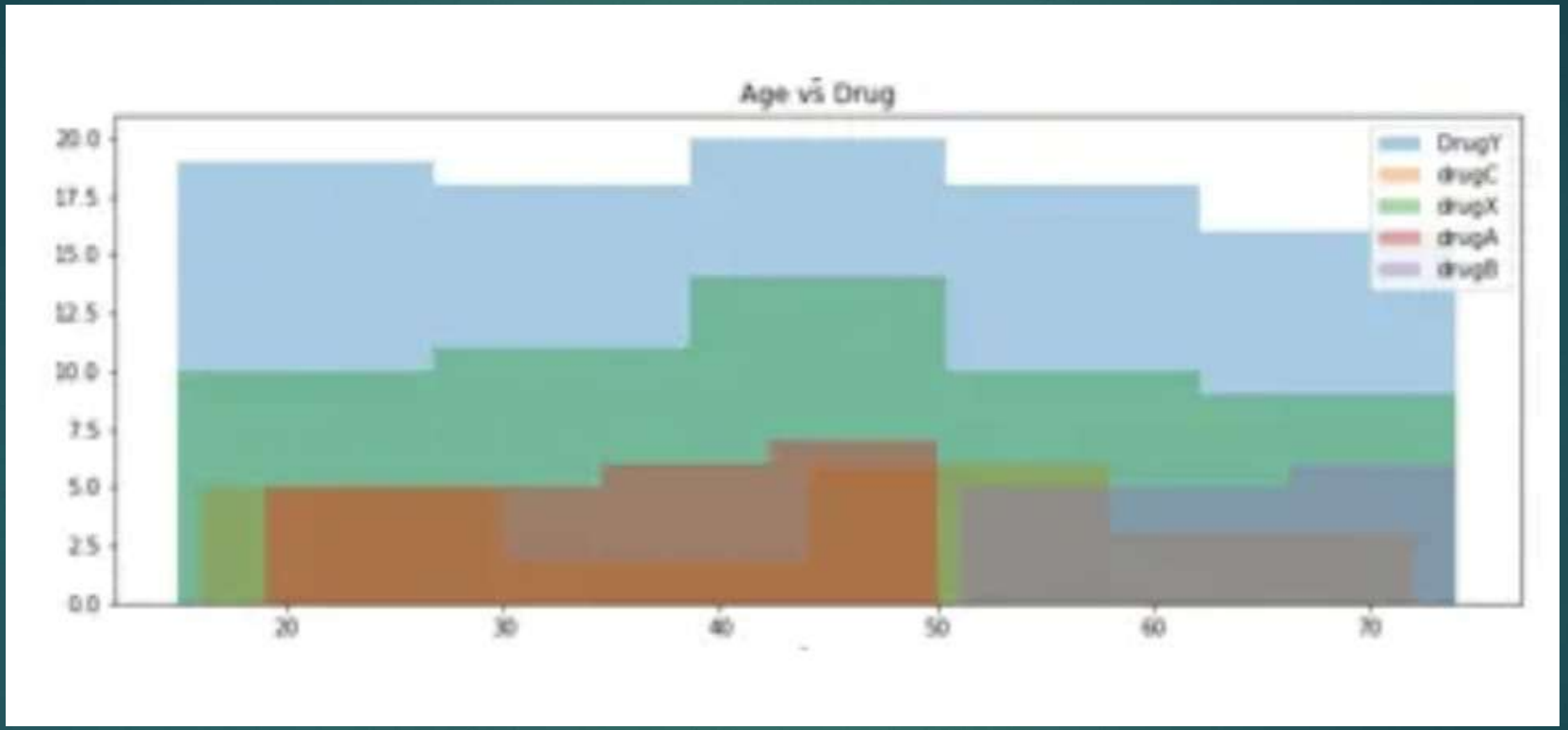
Analysis. Segmented analysis here means that the data variables is analysed in subsets.

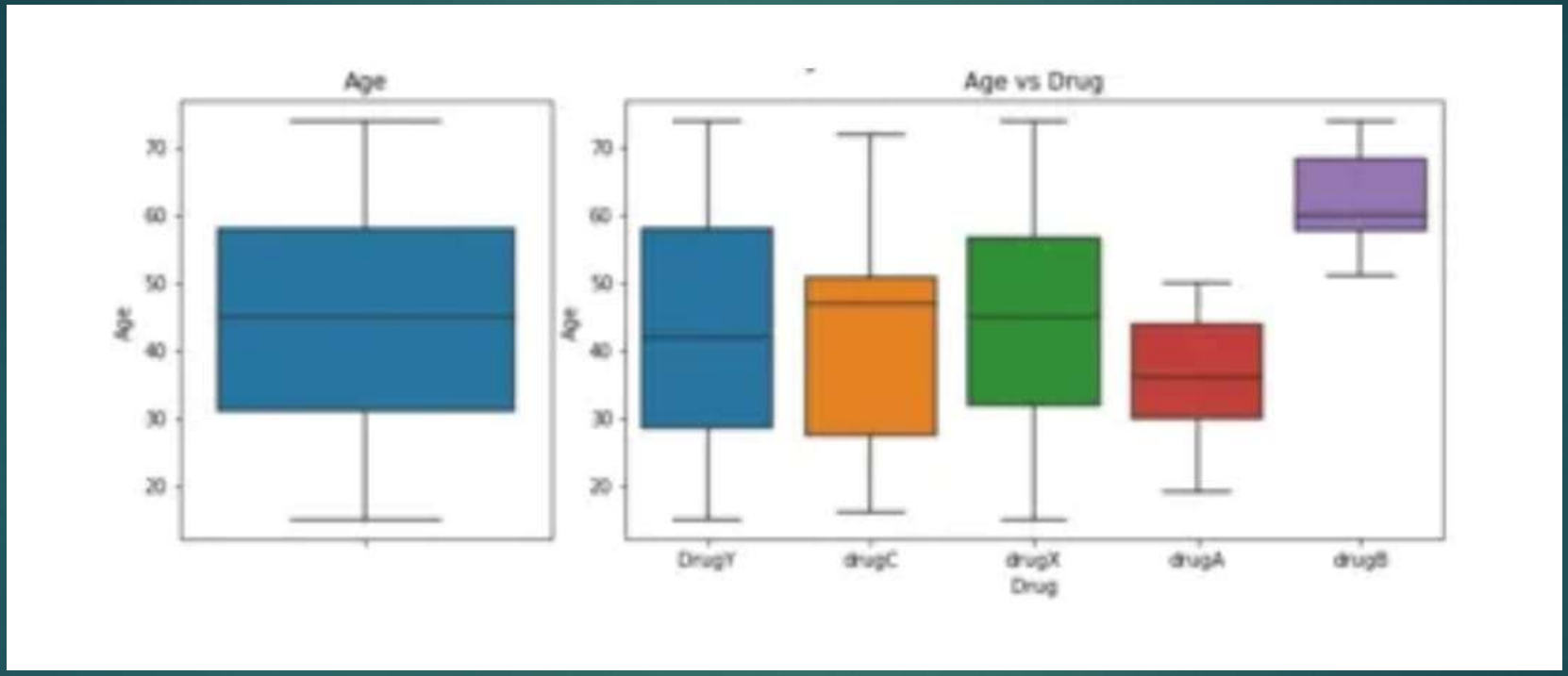
Segmented Univariate analysis can be used to find summary of a single data variable in the form of segments. It also used to detect the central tendencies such as a mean, median, and mode ; variance and standard deviation.

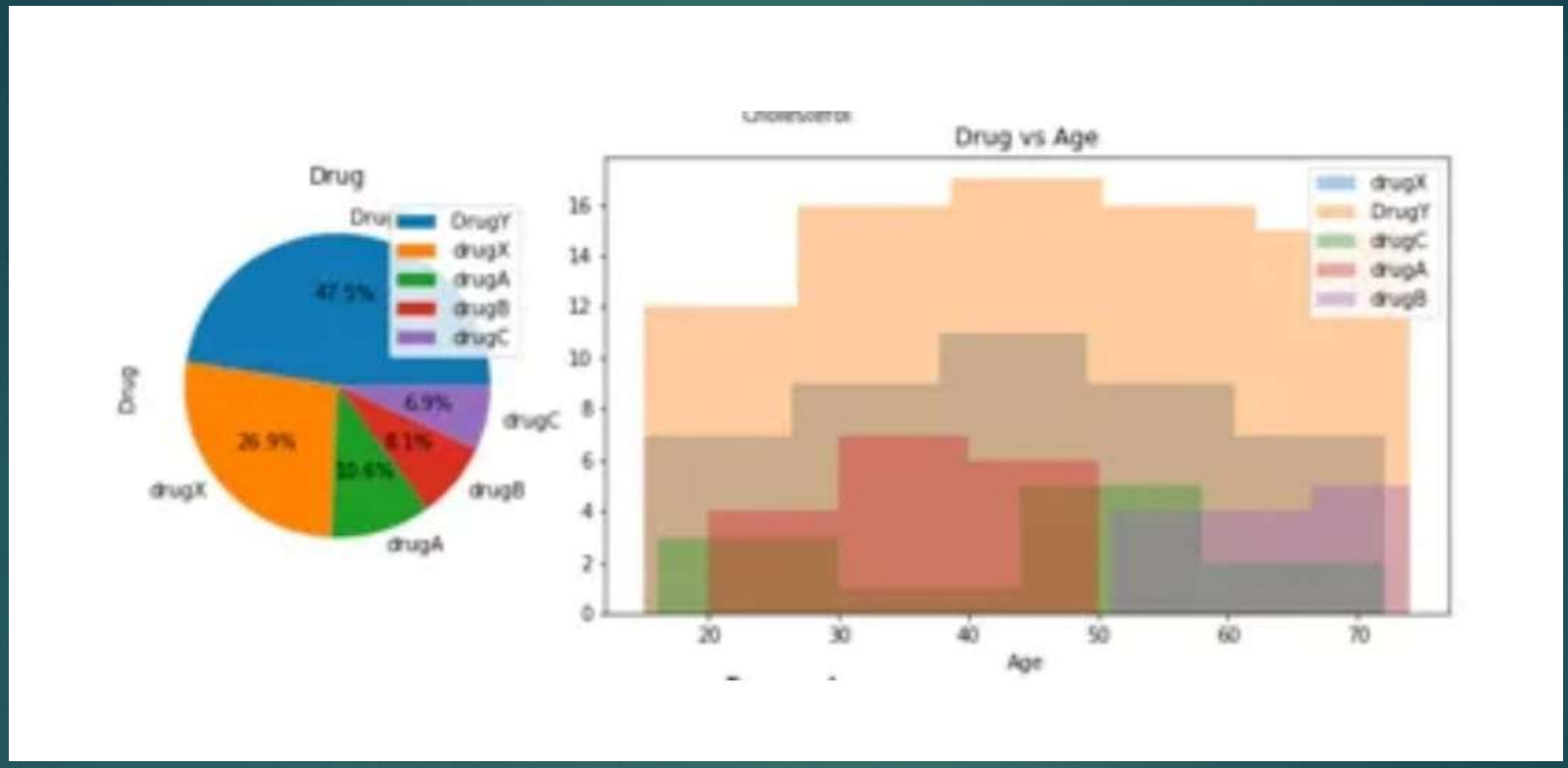

Cholesterol
HIGH
HIGH
NORMAL
51.5%
48.5%
Cholesterol
NORMAL

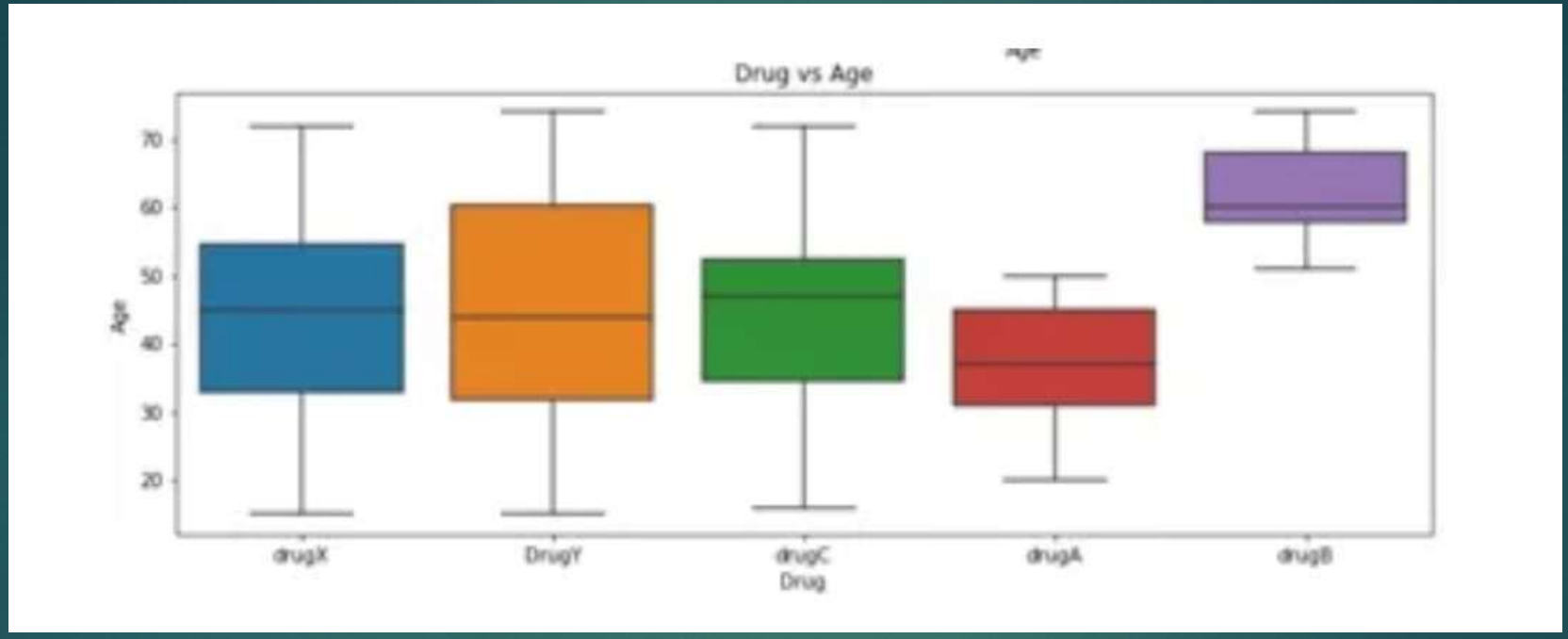

Cholesterol vs Drug
47
44
12
11
8
8
16
0
20
34
DrugY
drugA
drugB
drugC
drugX
Drug
HIGH
NORMAL
Cholesterol
40
32
24
16
8
0

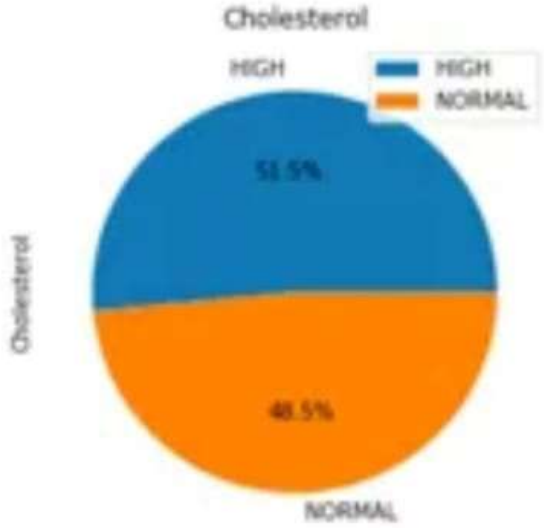

Age
35
30
25
20
15
10
5
0
20
30
40
50
60
70

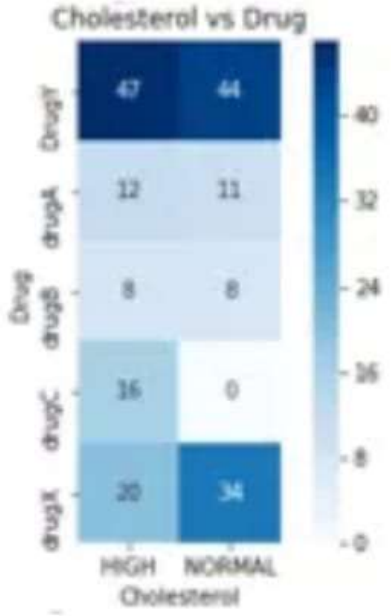

Age vs Drug
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
20
30
40
50
60
70
DrugY
drugC
drugX
drugA
drugB


Age
Age
70
60
50
40
30
20
Age vs Drug
Age
70
60
50
40
30
20
DrugY
drugC
drugX
drugA
drugB
Drug


Drug
DrugY
drugX
drugA
drugB
drugC
47.5%
26.9%
10.6%
8.1%
6.9%
drugC
drugB
drugA
drugX
Drug
Drug vs Age
drugX
DrugY
drugC
drugA
drugB
16
14
12
10
8
6
4
2
0
20
30
40
50
60
70
Age


Drug vs Age
Age
70
60
50
40
30
20
drugX
DrugY
drugC
drugA
drugB
Drug


Cholesterol
HIGH
HIGH
NORMAL
51.5%
48.5%
NORMAL
Cholesterol

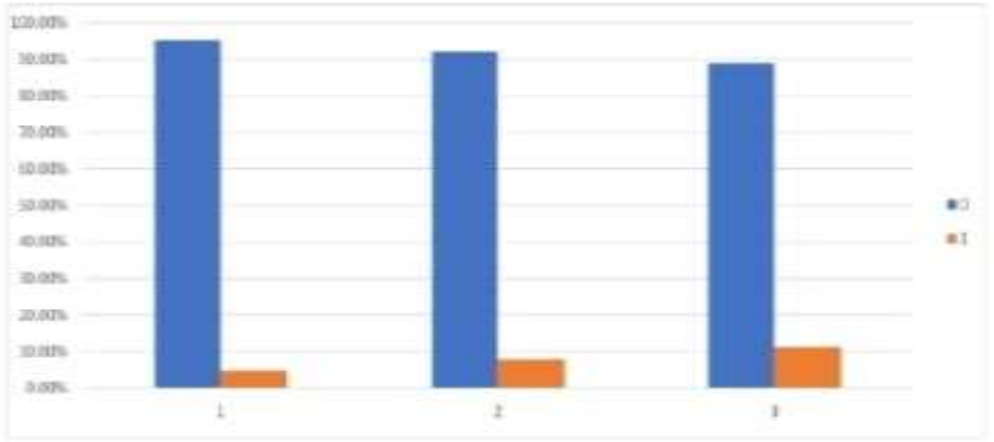

Cholesterol vs Drug
Drug
drugY
drugX
drugC
drugB
drugA
47
44
12
11
8
8
16
0
20
34
40
32
24
16
8
0
HIGH
NORMAL
Cholesterol

- ***Bivariate Analysis:-***

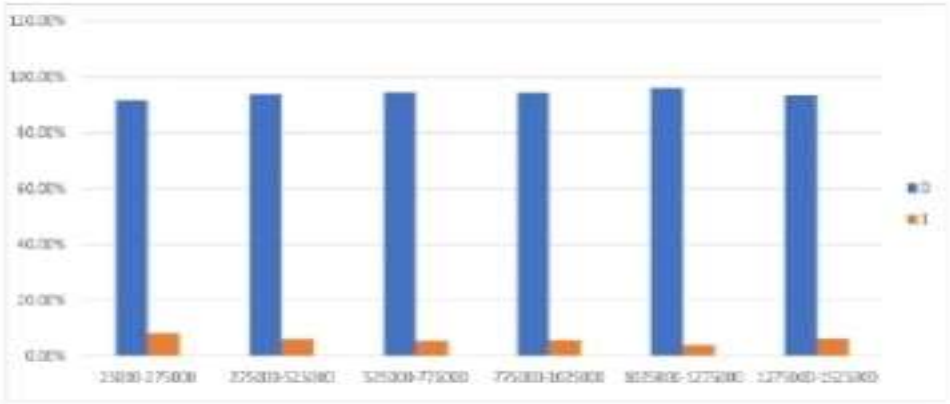
Interference

Customer who live in low-rating areas will have higher defaults. Individuals with lower income are more likely to default, and the trend of defaults declines with age. Ladies are less inclined than males to have defaults,. More defaults are predicted due to maternity leave and unemployment. Customers with more than five family members are more likely to defaults on their bank loan. Customers with fewer educational qualifications are more likely to fail on a bank loan. Customers with hardly work-experience are more likely to have defaults.

Region Rating Client vs Target



Amount Income vs Target




Age vs Target
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
20-35
35-50
50-65
65-80
0
1


Gender vs Target
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
1
F
M
XNA

Identify Top Correlations for Different Scenarios:

Considering 0.5 (absolute value) as threshold for high correlations, we can observe that:

. AMT_CREDIT and AMT_GOODS_PRICES are highly and positively correlated as thr credit amount request is for the Goods whose prices is in AMT_GOODS_PRICES columns.

.CNT_FAM_MEMBERS and CNT_CHILDREN are highly and positively correlated as we observed before that all applicants were either single parent or had nuclear Family.

.CREDIT%INCOME and ANNUITY%INCOME, AMT_ANNUITY and AMT_GOODS_PRICES,

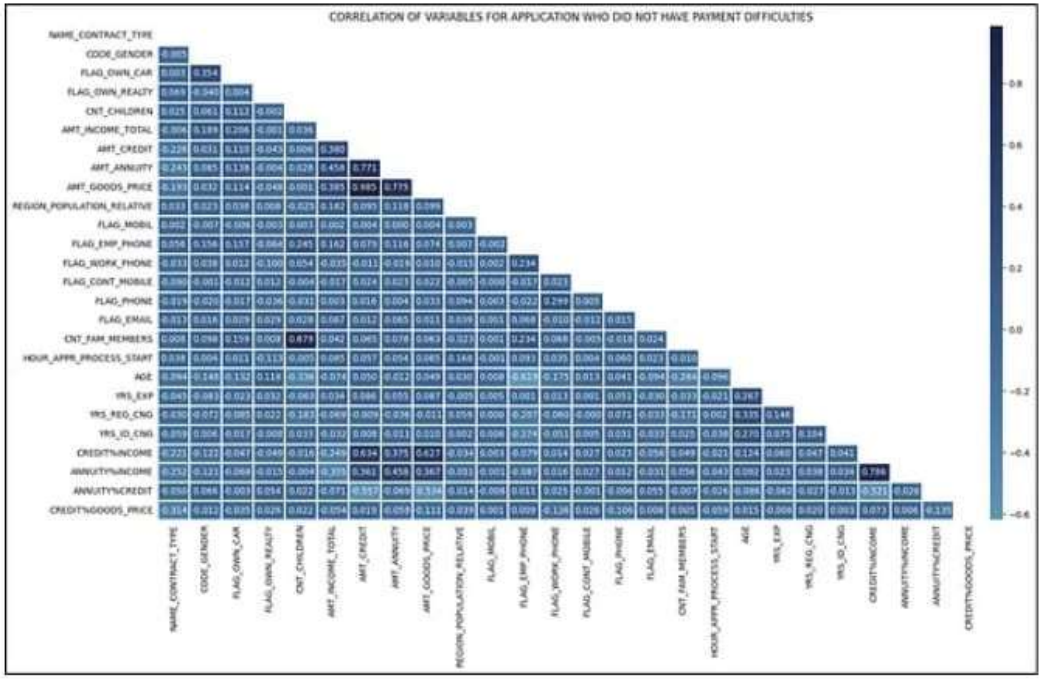
AMT_CREDIT and AMT_ANNUITY are highly and positively correlated and have some what same Correlation values.

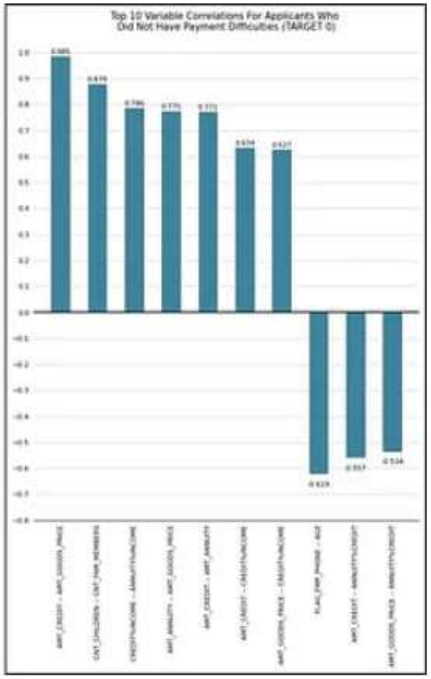
contd..

. AMT_CREDIT and CREDIT%INCOME, AMT_GOODS_PRICES and CREDIT%INCOME are highly and positively correlated as CREDIT%INCOME is a derived features which is directly proportional to AMT_CREDIT and AMT_CREDIT and AMT_GOODS_ORICES are highly and positively correlated as in point 1.

. FLAG_EMP_PHONE and AGE are highly and negatively correlated possibly because older generation people are less used to phones.

. AMT_CREDIT and ANNUIT%CREDIT,AMT_GOODS_PRICES and ANNUITY%CREDIT are highly and negatively correlated as ANNUITY%CREDIT is a derived features which is inversely proportional to AMT_CREDIT and AMT_CREDIT and AMT_GOODS_PRICES are highly and positively correlated as in point 1.


CORRELATION OF VARIABLES FOR APPLICATION WHO DID NOT HAVE PAYMENT DIFFICULTIES
NAME_CONTRACT_TYPE
CODE_GENDER
FLAG_OWN_CAR
FLAG_OWN_REALTY
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
REGION_POPULATION_RELATIVE
FLAG_MOBIL
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
CNT_FAM_MEMBERS
HOUR_APPR_PROCESS_START
AGE
YRS_EXP
YRS_REG_CNG
YRS_ID_CNG
CREDIT%INCOME
ANNUITY%INCOME
ANNUITY%CREDIT
CREDIT%GOODS_PRICE


Top 10 Variable Correlations For Applicants Who
Did Not Have Payment Difficulties (TARGET 0)

Project Description

This case study attempts to demonstrate the applications of EDA in a real-world business environment . In this case study, in addition to using the techniques learned in the EDA module, you will gain a basic grasp of risk analytics in banking and financial services, as how data is utilized to reduced the risk of losing money when lending to consumes.

- **Approach:**

this case study has two enormous data seta: the current application and the previous applications. Each included several unneeded columns that would be useless for risk assessments, as well as many blank data. I started by cleaning.

To evaluate this enormous set of data, I first cleaned the data, located some outliers and deleted them, and then began performing univariate and bivariate analysis using pivot table and charts.

Technique stack was used:

MySQL Workbench 8.0 CE, Microsoft Excel 2010

Results:

I went through the risk analytics process step by step, task after task. The project outcomes are follow:

Overall Method to analysis :

the bank's problem statement is to identify the major cause of bank loan default. The knowledge will be used for risk assessment by the company. We have provided two enormous datasets here.

1. 'applications data.csv' contains all the client's information at the time of applications. The information pertains to whether or not a client is having financial issues.

2. 'previous application.csv' provide data from the client's previous loans . It indicate if the prior application was accepted, cancelled, Refused, or not Unused.

Both sets of data contained many cleaning procedures, I splits columns in the databased on two categories of variables.

contd..

Following the data cleaning procedures, I splits columns in the dataset based on two categories of variables.

1. Categorical variables
2. Numerical Variables

Categorical variables (non-numerical variables)- person's occupation, educations status.

Numerical variables – income, credit etc.,

The following are some of the categorical and numerical variables from the provided data set.

Categorical variables	Numeric variables
Gender	Age
Name contract type	Days employed
Income type	Amount Income
Education	Amount Annuity
Housing type	Amount credit

- I completed full EDA on the present application and then on the previous application. Then in this report, I summarised the results of both applications and provided business insights.