Bank loan Case study

Project Description:

This case study aims to give an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers

Approach:

This case study has two sets of large data current application and previous application. Both had many unwanted columns that will not useful for the risk analytics and also had many blank values. First, I cleaned the data, found some outliers and removed them as well and started performing univariate and bivariate analysis using pivot tables and charts to interpret this large set of data

Tech Stack used:

Microsoft Excel 2021 , MySQL Workbench 8.0 CE

Insights:

I really had a question why bank loans get refused. In this case study I got to understand about how bank loans get refused or rejected when we apply for it. Understood the business concepts of bank loans and how risk analysis from bank standpoint is performed. After dealing with this large set of data by doing Multivariate analysis using Pivot tables, I could read a story from the pivot charts

Results:

I did task by task and performed step by step process of risk analytics. Following are the results of the project

1. Overall Approach of Analysis:

Problem statement of the bank is to find the strong driving factors behind bank loan default. The company will utilise the knowledge for risk assessment. Here we have provided with 2 large sets of data

1. **`application_data.csv`** contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
2. **`previous_application.csv`** contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer

Both sets of data had many unwanted columns which will not be used for the risk analytics and had many blanks. So I cleaned the data and inserted a new **age column** based on days birth column. **Days employed** column had days with minus sign so I inserted a new column Months Employed and also changed minus sign

After doing data cleaning process, I separated columns based on two types of variables in the dataset.

1) Categorical variables
2) Numerical variables

Categorical variables are non-numeric variables like Occupation type, Education status of a person etc. Numerical variables are variables like Amount income, Amount credit etc

These are some of the categorical and numeric variables from the given data set

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

When I did boxplot for amount income, I found so many outliers so I had to remove them for a better analysis, after removing them I Started doing Exploratory data analysis (EDA) on the data sets using Pivot tables. Using Pivot tables and charts I could find the top driving factors behind bank loan default

First, I performed complete EDA on current application and then on previous application Then I summarised the results of both applications and delivered the business insights in this report

Current application.csv

Task 2 (Identify Missing data):

In the current application sheet, it had **161 columns**

1) I removed columns which had more than 5% of blank values
2) I removed so many unwanted columns which are of no use to analysis
3) I created a new column AGE
4) I inserted DAYS EMP column with changing the minus sign of DAYS EMPLOYED column

Finally, I left with only **25 columns**

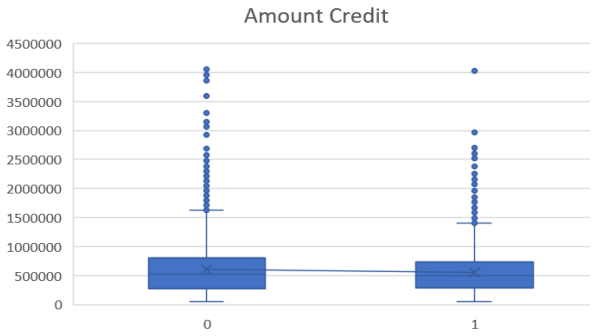
To remove blank values, I used **COUNTBLANK** function to find out the no of blank values and if the column has more than 5% blanks, I removed it.

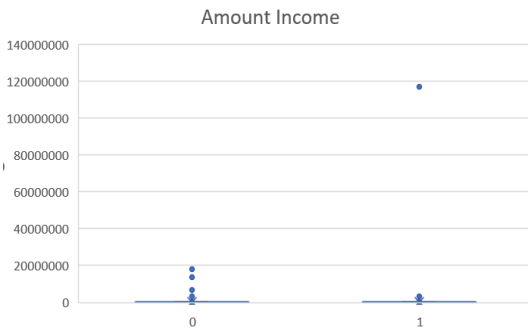
Task 3 (Outliers) :

Outliers can be found only on Numeric variables. I found outliers for 3 numeric variables from this data set

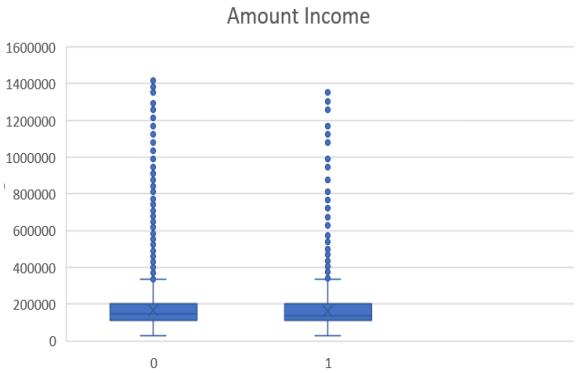
Box plotted Target column vs

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



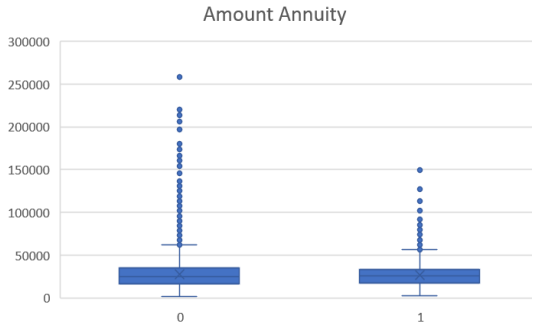


Before removal of outliers



Before removing Outliers in Amount income Median, Q1 and Q2 are not clearly visible so I had to remove those outliers

Amount Annuity

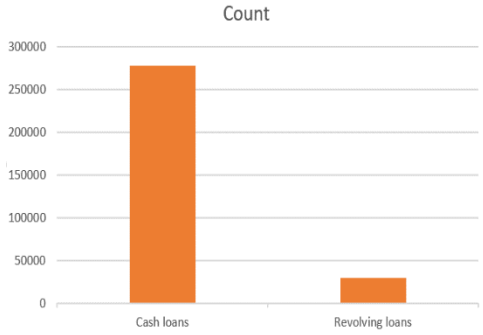


Outliers for Target 0 and 1 are found using this boxplot of Amount annuity column

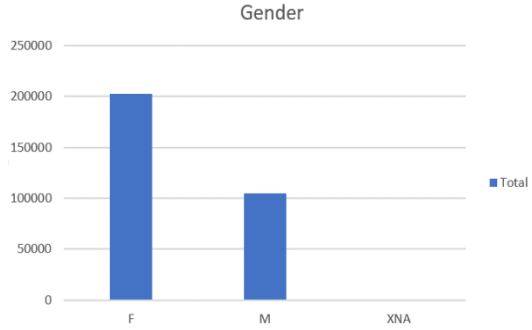
Task 4 (Data imbalance):

Data imbalance is where data is distributed so unequally. Here I found some columns where there is data imbalance. I used Pivot charts for plotting data imbalance

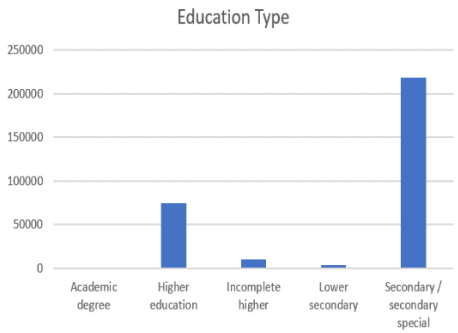
NAME CONTRACT TYPE



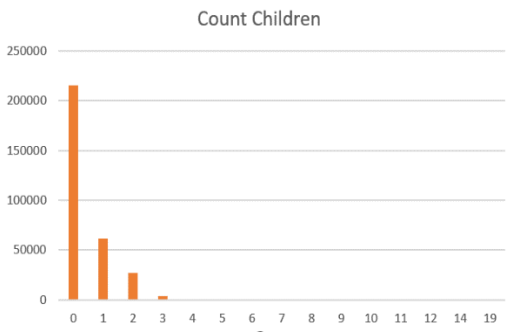
GENDER



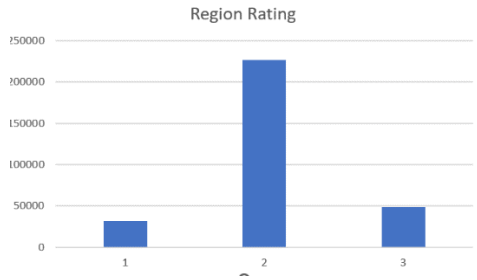
EDUCATION TYPE



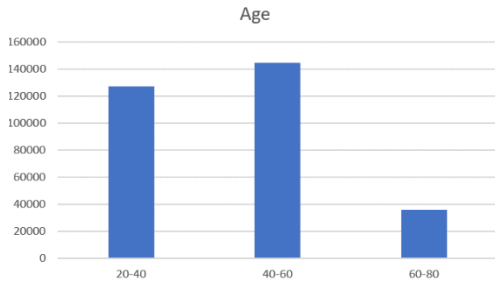
CNT CHILDREN



REGION RATING CLIENT



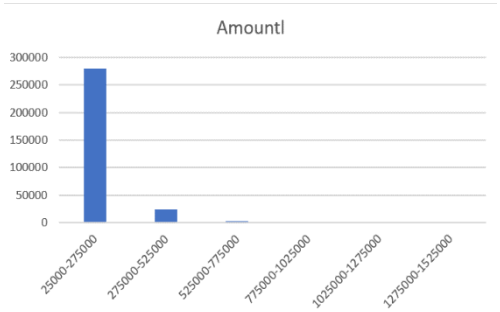
AGE



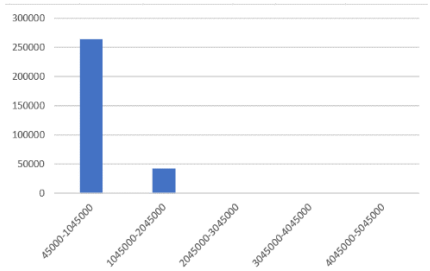
Task 5 (EDA):

Univariate Analysis:

Amount Income



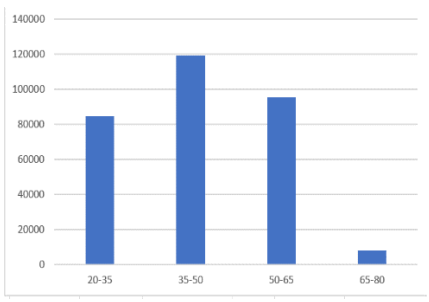
Amount Credit



Observation:

People with more income not likely to apply for loans. Credit amount of the bank loan mostly lies in the range of 45000 to 1045000

Age

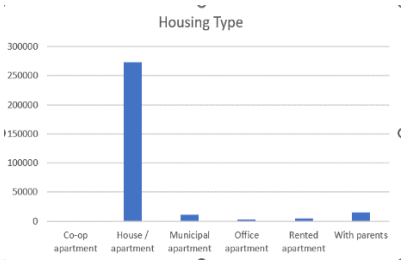


Months Employed

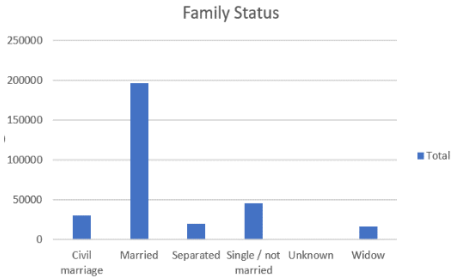
Observation:

People in the Mid age group (35 to 50) have mostly applied for loans. People with 0 to 8 years of Work experience will most likely to apply for loans

Housing type



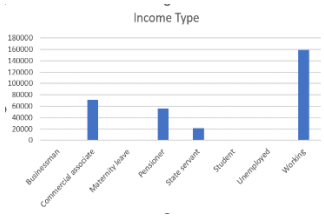
Family Status



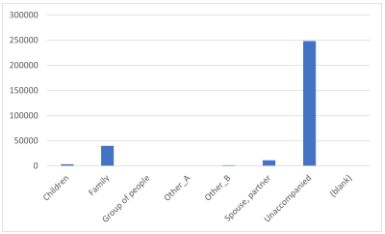
Observation:

People who have houses have applied for loans is higher than others. Married people have taken more loans

Income type



Name type Suite

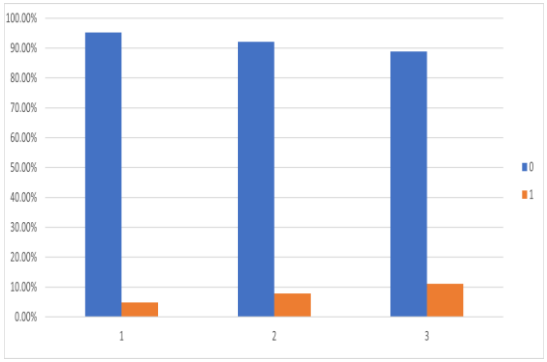


Observation:

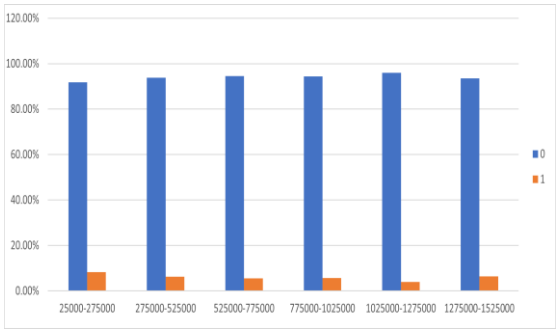
Working People have applied for more loans. People who were unaccompanied have applied for more loans

Bivariate Analysis:

Region Rating Client vs Target



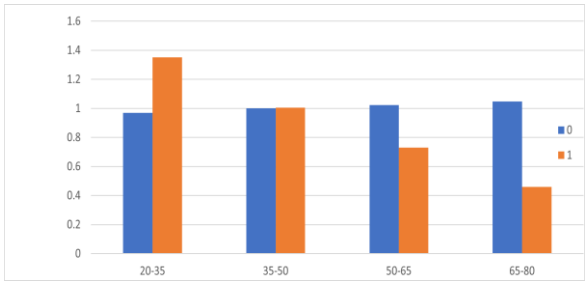
Amount Income vs Target



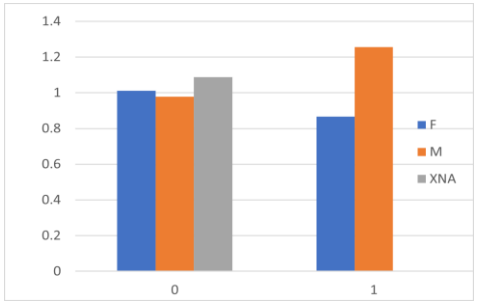
Observation:

Clients who are living in poor rating regions will have more defaults. People with less income will likely to have more defaults

Age vs Target



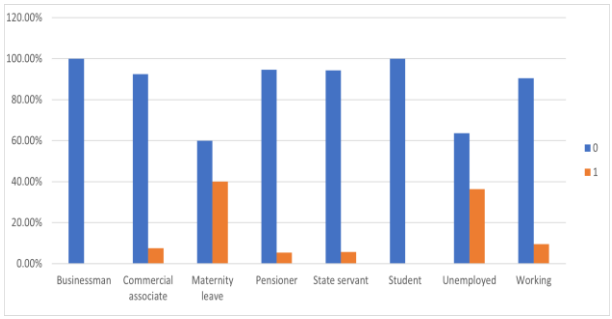
Gender vs Target



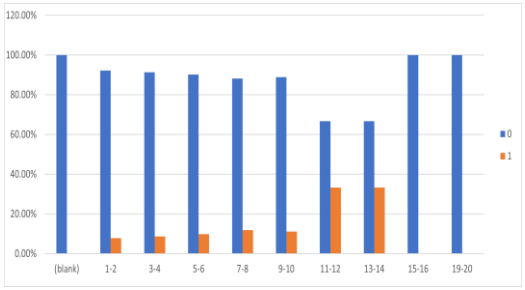
Observation:

Young People are likely to be defaulters and the trend of defaulters decreases along the age. Females are less likely to have defaults

Income type vs Target



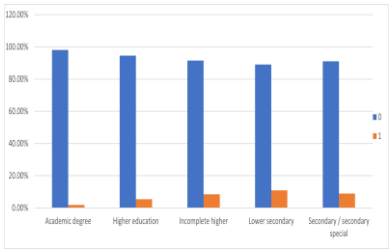
Family members vs Target



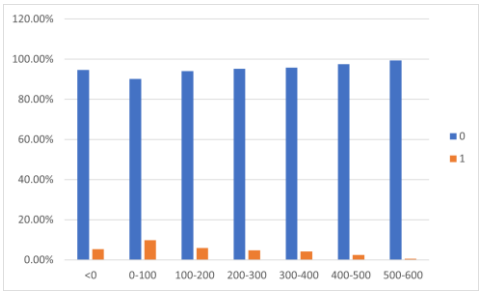
Observation:

Maternity leave and Unemployed will most likely to have more defaults. Clients whose family members are more than 5 will likely to default bank loan

Education type vs Target



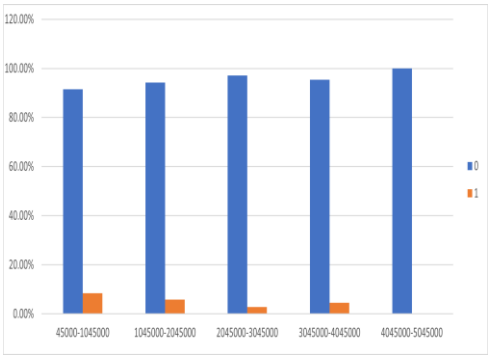
Months Employed vs Target



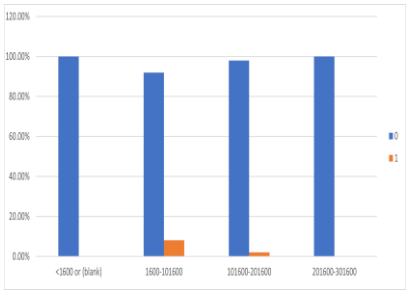
Observation:

Clients who have are educationally lower qualified will likely to default bank loan. Clients with least work experience will likely to have more defaults

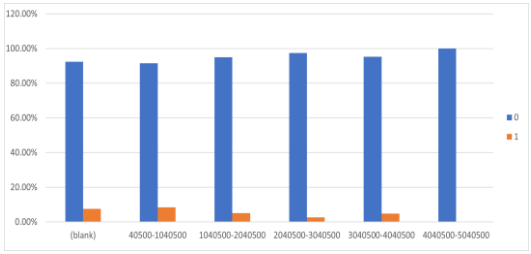
Amount Credit vs Target



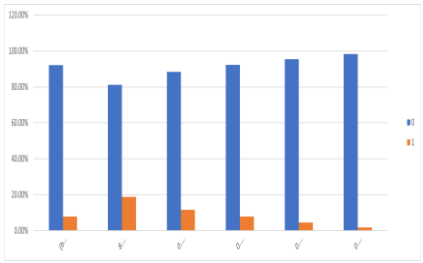
Amount Annuity vs Target



Amount Goods Price vs Target



External source vs Target



Task 6 (Finding top 10 correlations):

Top 10 driving factors in current application.csv

1. Income type
2. Count of Family Members
3. Children count
4. External source
5. Region rating of client
6. Age
7. Months Employed
8. Amount credit
9. Amount Goods Price
10. Amount total income

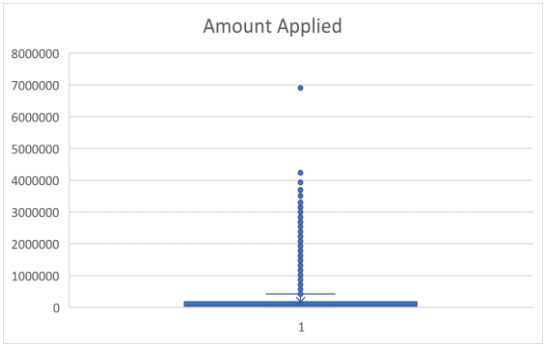
Previous Application.csv

Task 2 (Data Cleaning):

Removing columns: Used COUNTBLANK function to find the number of blanks and if it exceeds more than 5% in a column I deleted it. Also deleted few columns which are of no use to the Analysis

Inserted a new column Months decision based on Days decision column for my convenience of analysis

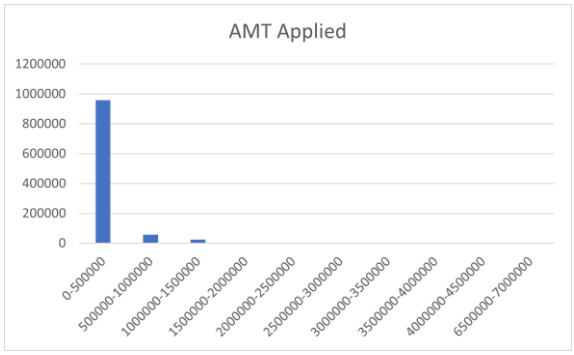
Task 3 (Finding Outliers):

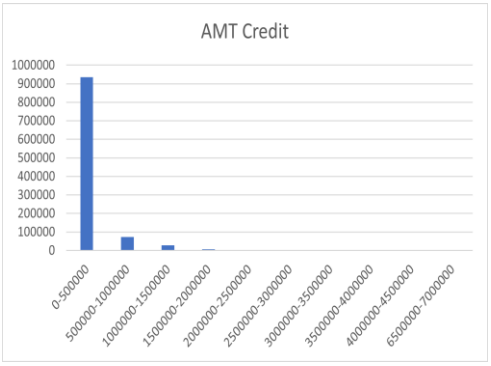


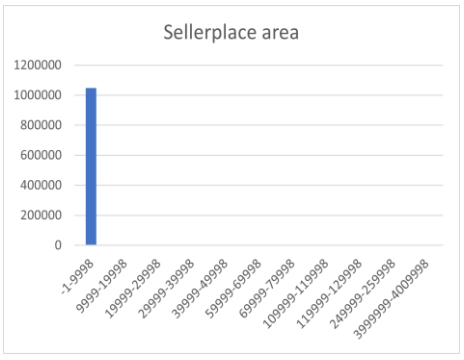
I used Boxplot for the Amount application column and outliers are found out

Task 4(Data Imbalance):

Found some columns where data is unevenly distributed

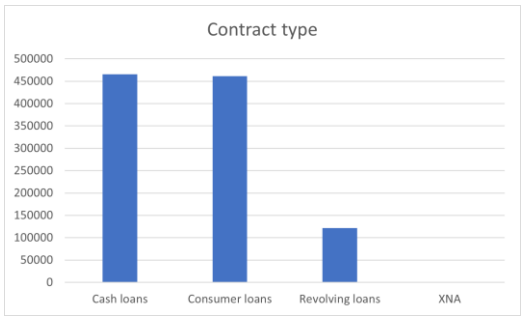


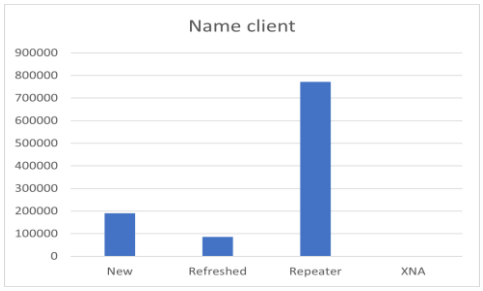




Task 5 (EDA):

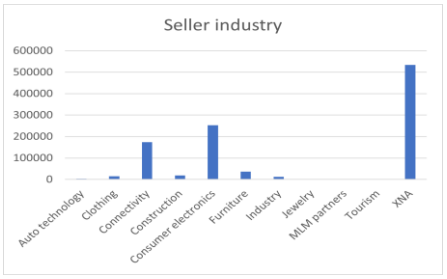
Univariate Analysis:





Observation:

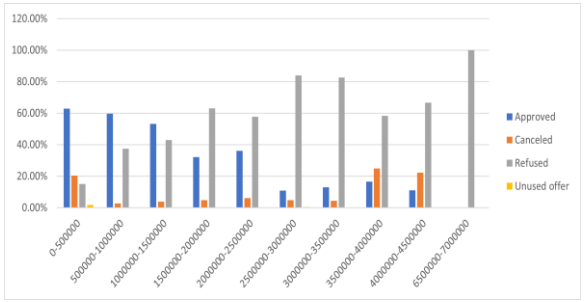
Clients have opted mostly for Cash and Consumer loans. Most of the Clients are repeaters

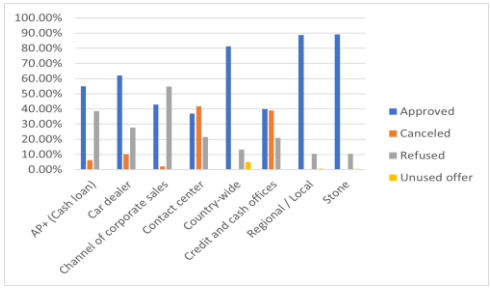


Observation:

Most of the Clients who have applied for current loan are those who have applied for loans just 10 months back. Consumer electronics have applied for more loans

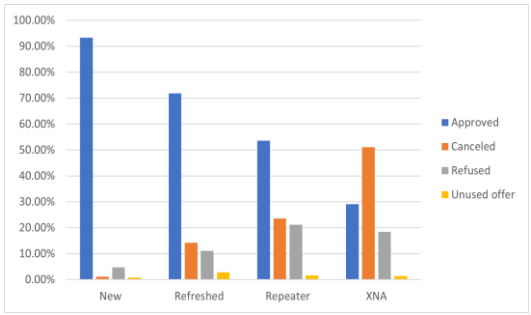
Bivariate Analysis:

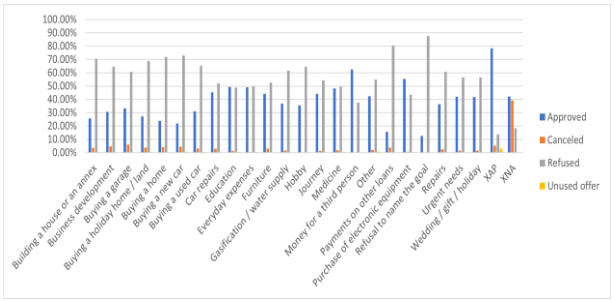


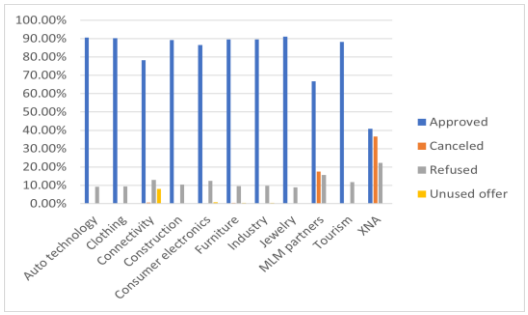


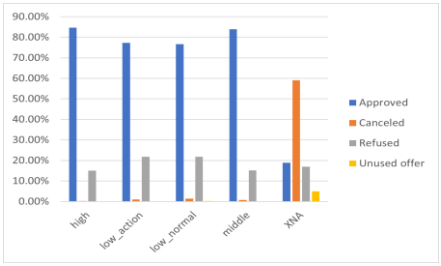
Observation:

Those Clients who applied for more than 35000000 amount will likely to get refuse. Most of the loans applied via Credit and Cash offices channel are cancelled



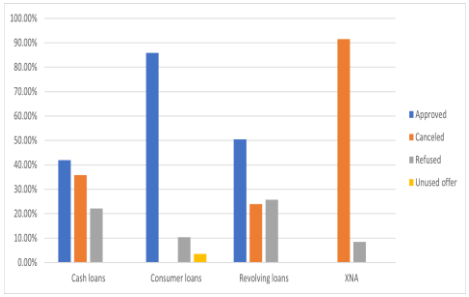


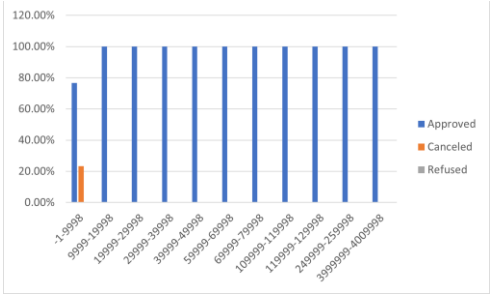




Observation:

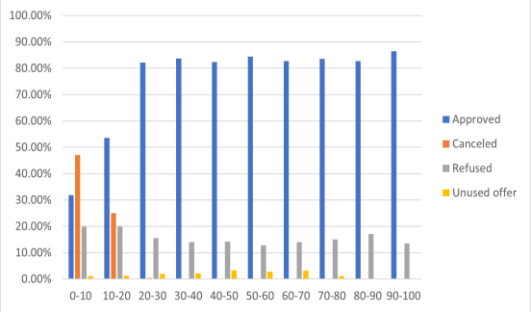
New Clients are happy as most of their loans got approved. Car loans have been refused so man. MLM partner client's loans are likely to get cancelled Almost all of the loans got approved and there's a constant number of refusals

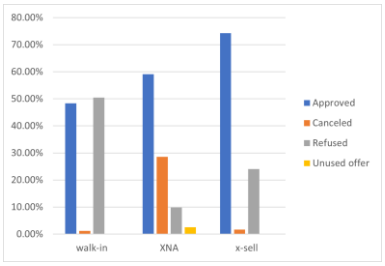




Observation:

Consumer loans have almost zero cancels and it has the highest approval percentage. Some loans were cancelled for the first Seller place area group





Observation:

It is seen that clients who have applied for another loan within 10 months of their previous loan will likely to get cancelled. Walk in loans have more no of refusal percentage

Task 6 (Finding Correlations):

Top 10 driving factors for the cancel and refusal of loans

1. Amount Application
2. Cash loan Purpose
3. Goods Category
4. Product Combination
5. Product type
6. Channel type
7. Months Decision
8. Contract type
9. Client type
10. Payment type

Task 7 (Combining two sheets):

Then I performed analysis for the common set of data by Joining Target column with previous application table. It is found that 8115 clients who applied currently for loan have already applied for loan in the same bank

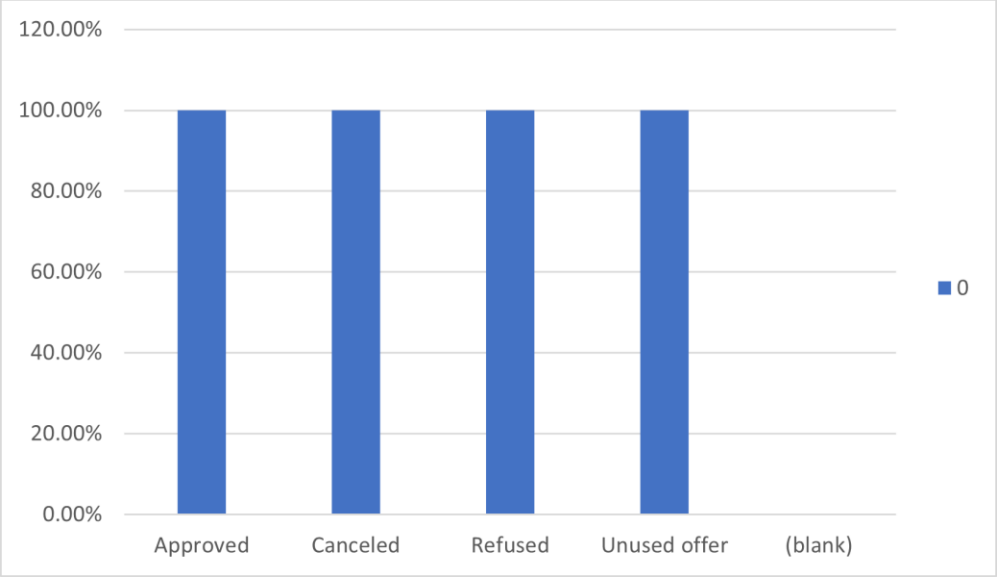
I combined them by using MySQL. I imported the data in workbench and performed the following query

Query:

```
USE bankloan;

SELECT m.TARGET,

n.SK_ID_CURR,

n.NAME_CONTRACT_TYPE,

n.AMT_APPLICATION,

n.NAME_CASH_LOAN_PURPOSE,

n.NAME_CONTRACT_STATUS,
```

```
n.NAME_CLIENT_TYPE,
n.DAYS_DECISION,
n.CODE_REJECT_REASON,
n.NAME_SELLER_INDUSTRY,
n.NAME_PORTFOLIO,
n.NAME_PRODUCT_TYPE,
n.CHANNEL_TYPE,
n.SELLERPLACE_AREA,
n.NAME_YIELD_GROUP,
n.PRODUCT_COMBINATION
FROM application_data m
JOIN prev_app n ON m.SK_ID_CURR=n.SK_ID_CURR;
```

Then I exported the results in Microsoft excel and did pivot table analysis



Clients who have applied for previous loans have no defaults in current loans

Summary:

Highly Recommended Groups	High Risk Groups
1. Approved Clients in previous applications 2. Married Clients 3. Senior Clients 4. Higher educated Clients 5. High Income Clients 6. Clients with Higher external source 7. Females 8. Clients with high work experience	1. Unemployed clients 2. Clients who are Young 3. Clients whose previous applications were rejected 4. Low Income clients 5. Clients with poor external source 6. Clients with least work experience 7. Clients who are in Maternity leave 8. Clients with more no of family members

All the Clients who re applied for the loans have no default