Trainity Project-6

Project-6: BANK LOAN CASE STUDY

Description:

This case study attempts to demonstrate the application 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 well as how data is utilized to reduce the risk of losing money when lending to consumers

Approach:

For this project, approach was to analyse the dataset, clean the dataset finding the blanks and missing values, imputing the missing values with the appropriate method (mean, median, mode). Then I tried to find the outliers in the dataset, there are some anomalies such as negative values which need either to be deleted or standardized. After all these I used pivot tables and basic charts to visualise the data. Moreover, insights were drawn based on my understandings

Tech-Stack:

- MS Excel (2023)
- Dataset provided (Bank Dataset)

Insights:

The dataset contains 3 files:

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.
3. columns_descrption.csv: It is data dictionary which describes the meaning of the variables

Both sets of data contained many undesired columns that will not be used for risk analytics, as well as many blanks. So, I cleaned up the data.

Following the data cleaning procedure, I split columns in the dataset based on two categories of variables. 1) Categorical variables

2) Numerical variables

Categorical variables (non-numerical variables)- person's occupation, education 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 summarized the results of both applications and provided business insights.

Current application.csv

Task 2 (Find Missing Data):

Importing the dataset in excel :

A	B	C	D	E	F	G	H	I	J	K	L
SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUIT
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied
100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family
100004	0	Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Unaccompanied
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompanied
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompanied
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, partner
100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompanied
100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompanied
100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children
100012	0	Revolving loans	M	N	Y	0	135000	405000	20250	405000	Unaccompanied
100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompanied
100015	0	Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Children
100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompanied
100017	0	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccompanied
100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccompanied
100019	0	Cash loans	M	Y	Y	0	157500	299772	20160	247500	Family
100020	0	Cash loans	M	N	N	0	108000	509602.5	26149.5	387000	Unaccompanied
100021	0	Revolving loans	F	N	Y	1	81000	270000	13500	270000	Unaccompanied
100022	0	Revolving loans	F	N	Y	0	112500	157500	7875	157500	Other_A
100023	0	Cash loans	F	N	Y	1	90000	544491	17563.5	454500	Unaccompanied
100024	0	Revolving loans	M	Y	Y	0	135000	427500	21375	427500	Unaccompanied
100025	0	Cash loans	F	Y	Y	1	202500	1132573.5	37561.5	927000	Unaccompanied
100026	0	Cash loans	F	N	N	1	450000	497520	32521.5	450000	Unaccompanied
100027	0	Cash loans	F	N	Y	0	83250	239850	23850	225000	Unaccompanied
100029	0	Cash loans	M	Y	N	2	135000	247500	12703.5	247500	Unaccompanied
100030	0	Cash loans	F	N	Y	0	90000	225000	11074.5	225000	Unaccompanied

A	B	C	D	E	F	G	H	I	J
SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_APPR_PROCESS_START	HOUR_APPR_PROCESS_START
2030495	271877	Consumer loans	1730.43	17145	17145	0	17145	SATURDAY	1
2802425	108129	Cash loans	25188.615	607500	679671		607500	THURSDAY	1
2523466	122040	Cash loans	15060.735	112500	136444.5		112500	TUESDAY	1
2819243	176158	Cash loans	47041.335	450000	470790		450000	MONDAY	
1784265	202054	Cash loans	31924.395	337500	404055		337500	THURSDAY	
1383531	199383	Cash loans	23703.93	315000	340573.5		315000	SATURDAY	
2315218	175704	Cash loans		0	0			TUESDAY	1
1656711	296299	Cash loans		0	0			MONDAY	
2367563	342292	Cash loans		0	0			MONDAY	1
2579447	334349	Cash loans		0	0			SATURDAY	1
1715995	447712	Cash loans	11368.62	270000	335754		270000	FRIDAY	
2257824	161140	Cash loans	13832.775	211500	246397.5		211500	FRIDAY	1
2330894	258628	Cash loans	12165.21	148500	174361.5		148500	TUESDAY	1
1397919	321676	Consumer loans	7654.86	53779.5	57564	0	53779.5	SUNDAY	1
2273188	270658	Consumer loans	9644.22	26550	27252	0	26550	SATURDAY	1
1232483	151612	Consumer loans	21307.455	126490.5	119853	12649.5	126490.5	TUESDAY	
2163253	154602	Consumer loans	4187.34	26955	27297	1350	26955	SATURDAY	1
1285768	142748	Revolving loans	9000	180000	180000		180000	FRIDAY	1
2393109	396305	Cash loans	10181.7	180000	180000		180000	THURSDAY	1
1173070	199178	Cash loans	4666.5	45000	49455		45000	SATURDAY	1
1506815	166490	Cash loans	25454.025	450000	491580		450000	MONDAY	
1182516	267782	Cash loans	20361.6	405000	451777.5		405000	SATURDAY	
1172842	302212	Cash loans		0	0			TUESDAY	
1172937	302212	Cash loans	39475.305	1129500	1277104.5		1129500	THURSDAY	
1555330	199353	Cash loans		0	0			SATURDAY	
1543131	275707	Cash loans	22619.52	229500	241920		229500	THURSDAY	

columns_description | application_data | previous_application | previous_application (2) | application_data (2) | Cle ...

A	B	C	D	E
Column1	Table	Row	Description	Special
1	application_data	SK_ID_CURR	ID of loan in our sample	
2	application_data	TARGET	Target variable (1 - client with payment difficulties: he/she had	
5	application_data	NAME_CONTRACT_TYPE	Identification if loan is cash or revolving	
6	application_data	CODE_GENDER	Gender of the client	
7	application_data	FLAG_OWN_CAR	Flag if the client owns a car	
8	application_data	FLAG_OWN_REALTY	Flag if client owns a house or flat	
9	application_data	CNT_CHILDREN	Number of children the client has	
10	application_data	AMT_INCOME_TOTAL	Income of the client	
11	application_data	AMT_CREDIT	Credit amount of the loan	
12	application_data	AMT_ANNUITY	Loan annuity	
13	application_data	AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the lo	
14	application_data	NAME_TYPE_SUITE	Who was accompanying client when he was applying for the lo	
15	application_data	NAME_INCOME_TYPE	Clients income type (businessman, working, maternity leave,...	
16	application_data	NAME_EDUCATION_TYPE	Level of highest education the client achieved	
17	application_data	NAME_FAMILY_STATUS	Family status of the client	
18	application_data	NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with	
19	application_data	REGION_POPULATION_RELATIVE	Normalized population of region where client lives (higher num	normalized
20	application_data	DAYS_BIRTH	Client's age in days at the time of application	time only relative to the application
21	application_data	DAYS_EMPLOYED	How many days before the application the person started curr	time only relative to the application
22	application_data	DAYS_REGISTRATION	How many days before the application did client change his reg	time only relative to the application
23	application_data	DAYS_ID_PUBLISH	How many days before the application did client change the id	time only relative to the application
24	application_data	OWN_CAR_AGE	Age of client's car	
25	application_data	FLAG_MOBIL	Did client provide mobile phone (1=YES, 0=NO)	
26	application_data	FLAG_EMP_PHONE	Did client provide work phone (1=YES, 0=NO)	
27	application_data	FLAG_WORK_PHONE	Did client provide home phone (1=YES, 0=NO)	
28	application_data	FLAG_CONT_MOBILE	Was mobile phone reachable (1=YES, 0=NO)	
29	application_data	FLAG_PHONE	Did client provide home phone (1=YES, 0=NO)	
30	application_data	FLAG_EMAIL	Did client provide email (1=YES, 0=NO)	

columns_description | application_data | previous_application | previous_application (2) | application_data (2) | Cle ...

A	B	C	D	E	F	G	H	I	J
SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_APPR_PROCESS_START	HOUR_APPR_PROCESS_ST
2030495	271877	Consumer loans	1730.43	17145	17145	0	17145	SATURDAY	
2802425	108129	Cash loans	25188.615	607500	679671		607500	THURSDAY	
2523466	122040	Cash loans	15060.735	112500	136444.5		112500	TUESDAY	
2819243	176158	Cash loans	47041.335	450000	470790		450000	MONDAY	
1784265	202054	Cash loans	31924.395	337500	404055		337500	THURSDAY	
1383531	199383	Cash loans	23703.93	315000	340573.5		315000	SATURDAY	
2315218	175704	Cash loans		0	0			TUESDAY	
1656711	296299	Cash loans		0	0			MONDAY	
2367563	342292	Cash loans		0	0			MONDAY	
2579447	334349	Cash loans		0	0			SATURDAY	
1715995	447712	Cash loans	11368.62	270000	335754		270000	FRIDAY	
2257824	161140	Cash loans	13832.775	211500	246397.5		211500	FRIDAY	
2330894	258628	Cash loans	12165.21	148500	174361.5		148500	TUESDAY	
1397919	321676	Consumer loans	7654.86	53779.5	57564	0	53779.5	SUNDAY	
2273188	270658	Consumer loans	9644.22	26550	27252	0	26550	SATURDAY	
1232483	151612	Consumer loans	21307.455	126490.5	119853	12649.5	126490.5	TUESDAY	
2163253	154602	Consumer loans	4187.34	26955	27297	1350	26955	SATURDAY	
1285768	142748	Revolving loans	9000	180000	180000		180000	FRIDAY	
2393109	396305	Cash loans	10181.7	180000	180000		180000	THURSDAY	
1173070	199178	Cash loans	4666.5	45000	49455		45000	SATURDAY	
1506815	166490	Cash loans	25454.025	450000	491580		450000	MONDAY	
1182516	267782	Cash loans	20361.6	405000	451777.5		405000	SATURDAY	
1172842	302212	Cash loans		0	0			TUESDAY	
1172937	302212	Cash loans	39475.305	1129500	1277104.5		1129500	THURSDAY	
1555330	199353	Cash loans		0	0			SATURDAY	
1543131	275707	Cash loans	22619.52	229500	241920		229500	THURSDAY	

columns description | application_data | previous_application | previous application (2) | application_data (2) | Cle ...

Imputing the missing values using mean , median and mode

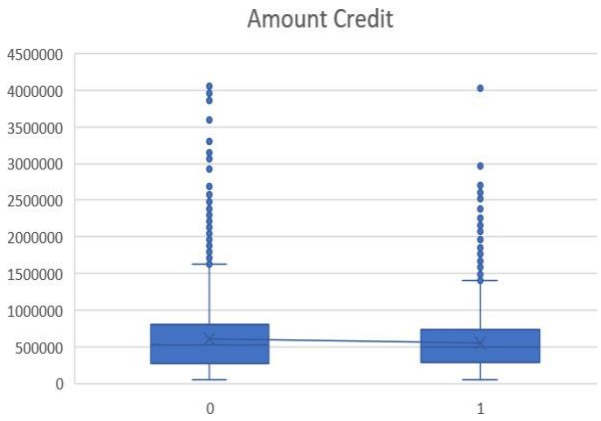
A	B	C	D	E	F	G	H	I	J	K	L
SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUI
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied
100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family
100004	0	Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Unaccompanied
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompanied
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompanied
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, partner
100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompanied
100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompanied
100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children
100012	0	Revolving loans	M	N	Y	0	135000	405000	20250	405000	Unaccompanied
100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompanied
100015	0	Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Children
100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompanied
100017	0	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccompanied
100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccompanied
100019	0	Cash loans	M	Y	Y	0	157500	299772	20160	247500	Family
100020	0	Cash loans	M	N	N	0	108000	509602.5	26149.5	387000	Unaccompanied
100021	0	Revolving loans	F	N	Y	1	81000	270000	13500	270000	Unaccompanied
100022	0	Revolving loans	F	N	Y	0	112500	157500	7875	157500	Other_A
100023	0	Cash loans	F	N	Y	1	90000	544491	17563.5	454500	Unaccompanied
100024	0	Revolving loans	M	Y	Y	0	135000	427500	21375	427500	Unaccompanied
100025	0	Cash loans	F	Y	Y	1	202500	1132573.5	37561.5	927000	Unaccompanied
100026	0	Cash loans	F	N	N	1	450000	497520	32521.5	450000	Unaccompanied
100027	0	Cash loans	F	N	Y	0	83250	239850	23850	225000	Unaccompanied
100029	0	Cash loans	M	Y	N	2	135000	247500	12703.5	247500	Unaccompanied
100030	0	Cash loans	F	N	Y	0	90000	225000	11074.5	225000	Unaccompanied

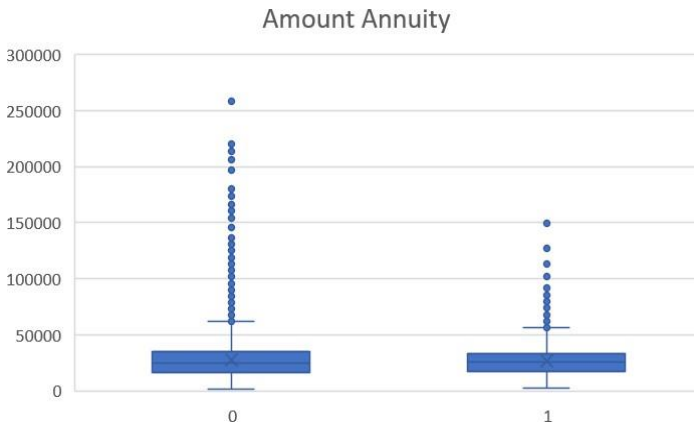
... | previous_application (2) | application_data (2) | Cleaning_application_data | Sheet1 | Targets and amt annunity | Sh ...

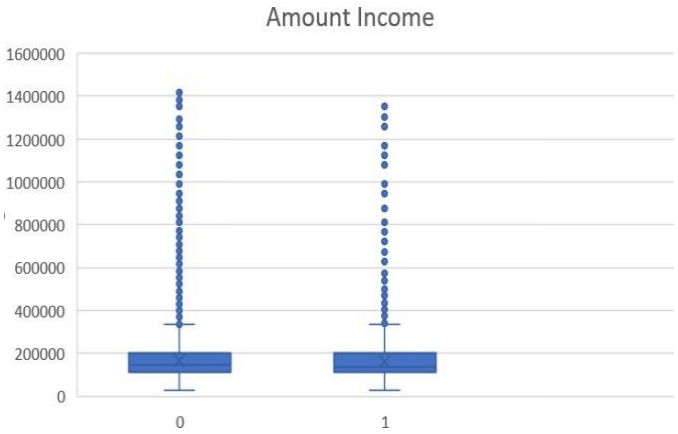
3.Outliers can only be identified on Numeric

Box plotted Target column vs

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

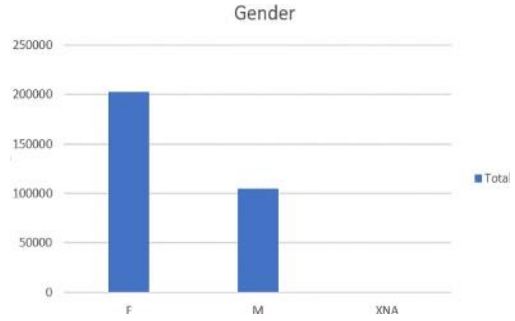


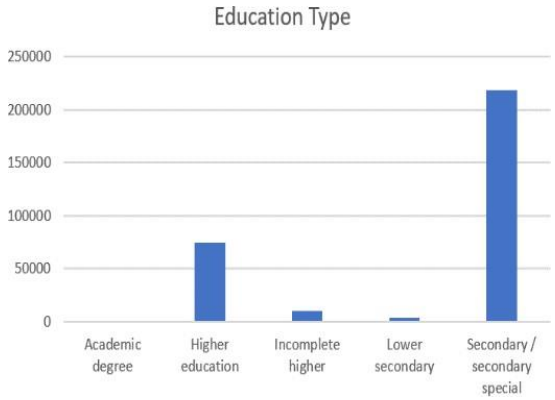


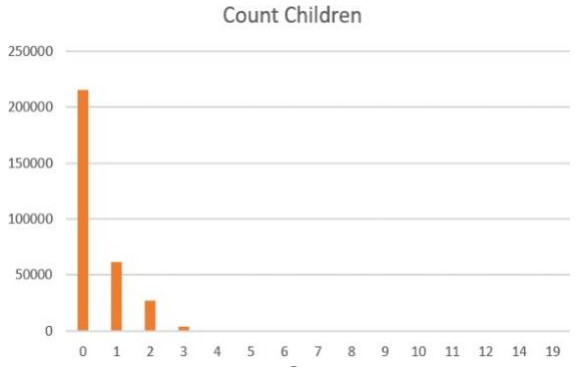


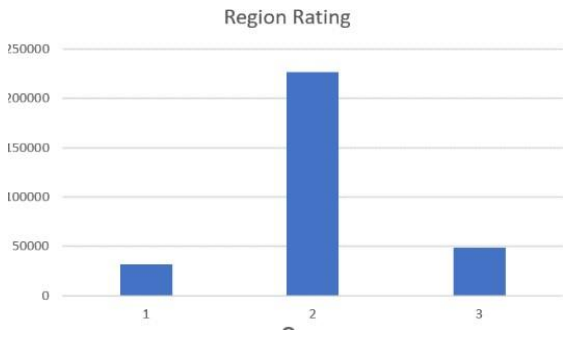
Data imbalance: Data imbalance occurs when data is disseminated in an unequal manner. I plotted data imbalance using Pivot charts.

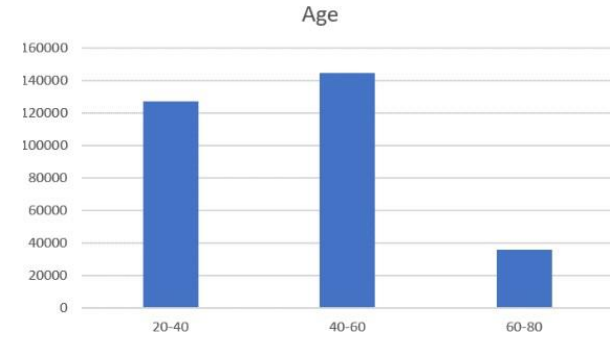
NAME CONTRACT TYPE











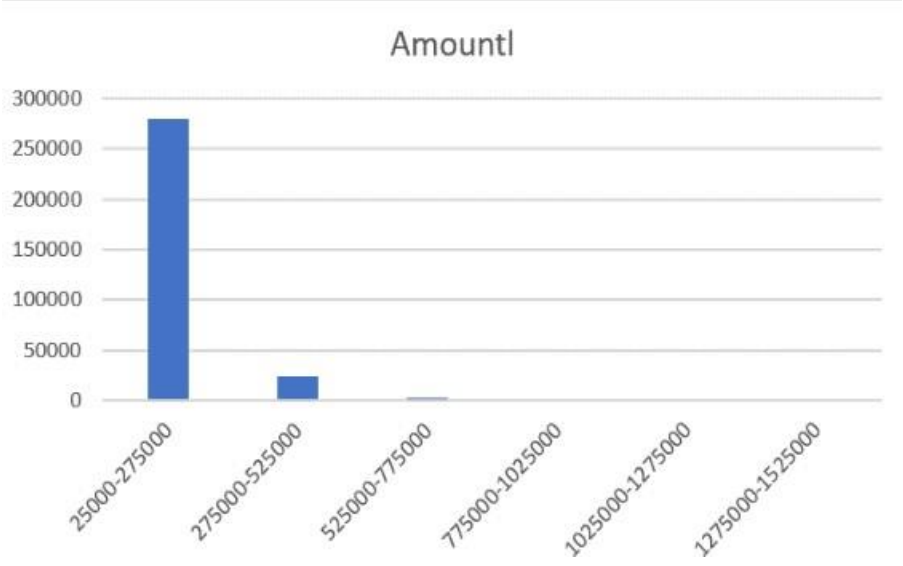
Task 5 (EDA):

Univariate Analysis:

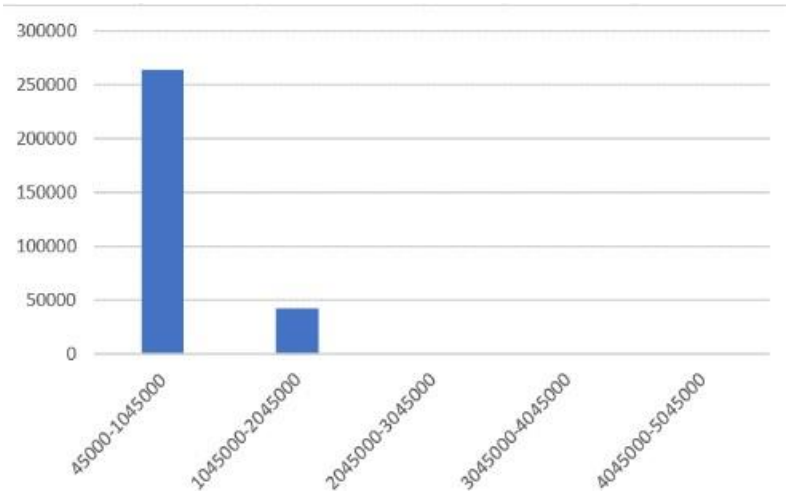
INFERENCE

Individuals with higher incomes are less likely to apply for loans. The credit 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 experience 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.

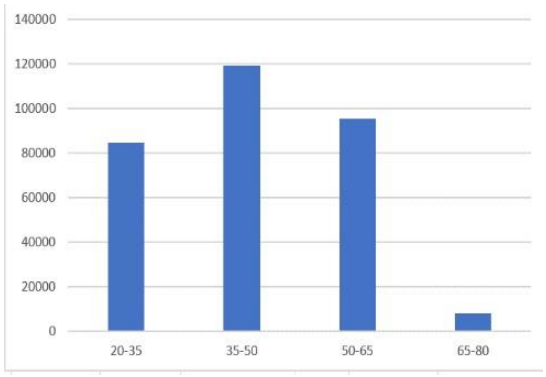
Amount Income

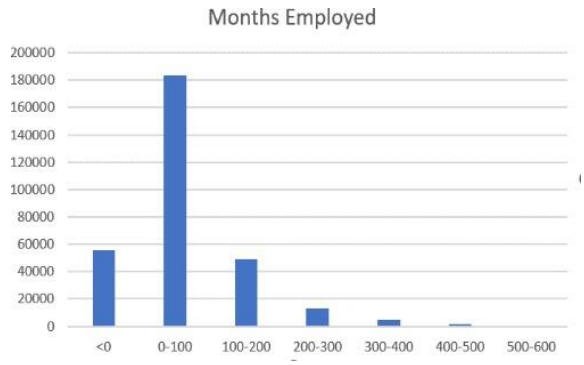


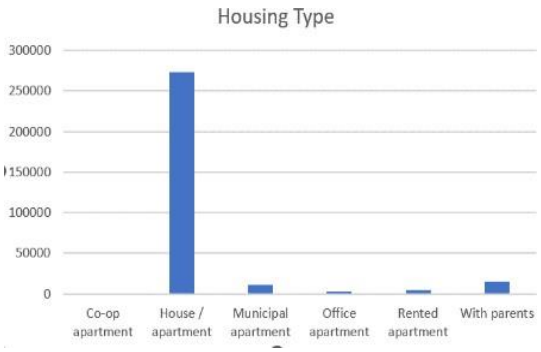
Amount Credit

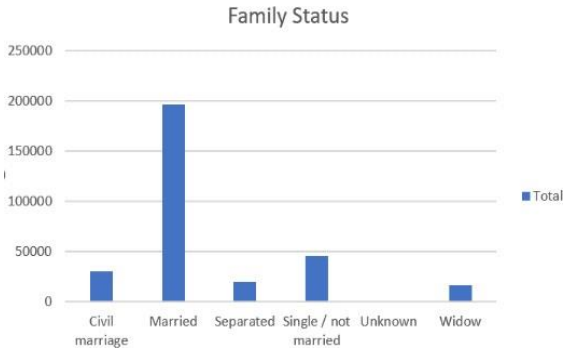


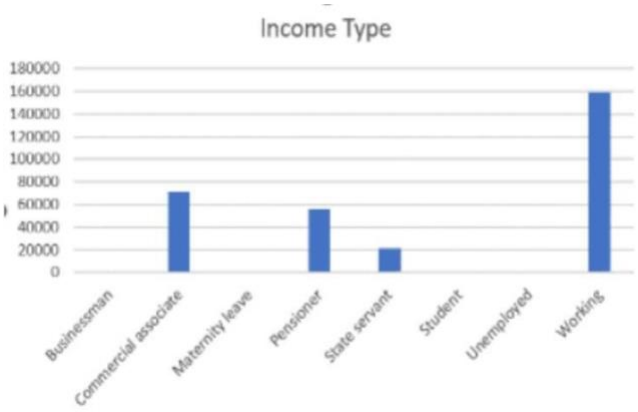
AGE



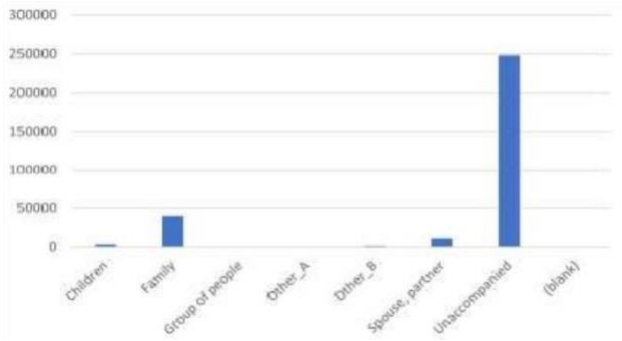








Name suite type



Bivariate Analysis:

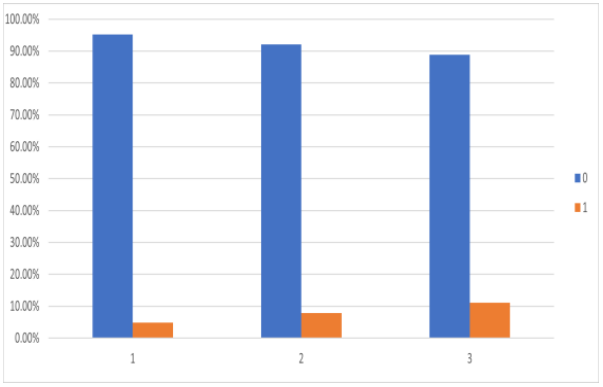
INFERENCE

Customers who live in low-rating areas will have higher defaults.

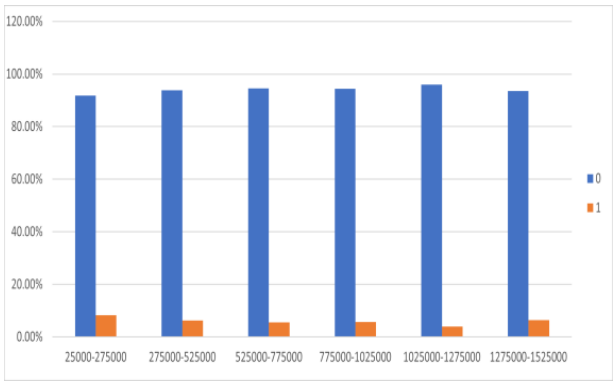
Individuals with lower incomes are more likely to default. Young people are more likely to default, and the trend of defaulters declines with age. Ladies are less inclined than males to have defaults. More defaults are predicted due to maternity leave and unemployment. Customers with more than five family members are more likely to default 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

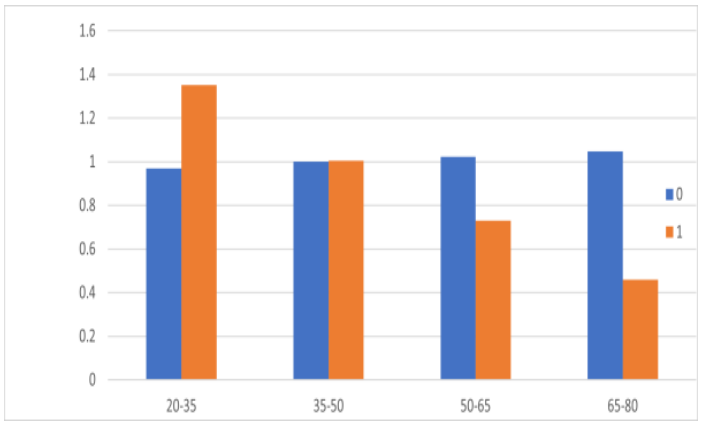
Amount Income vs Target



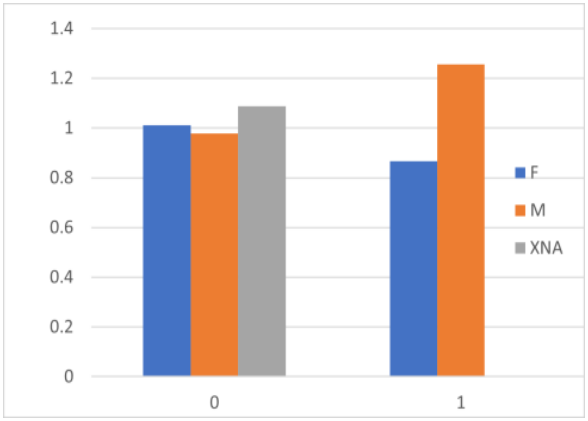
Age vs Target



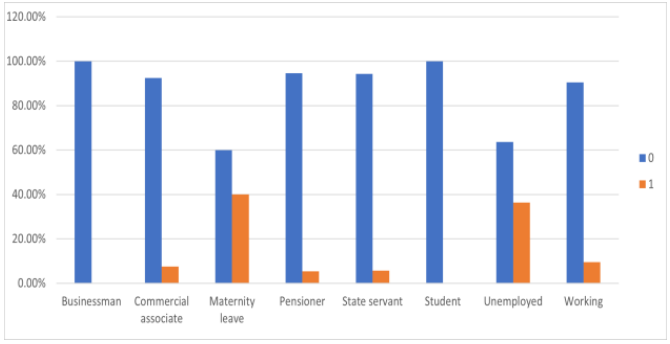
Gender vs Target



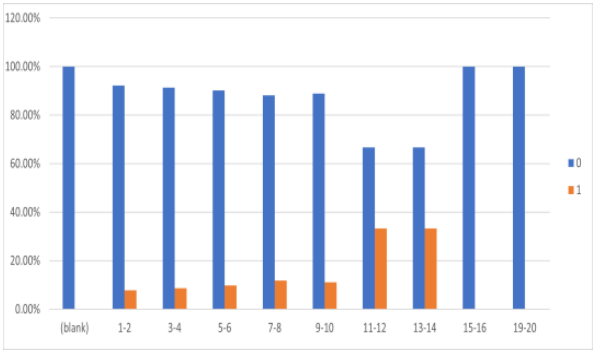
INCOME TYPE VS TARGET



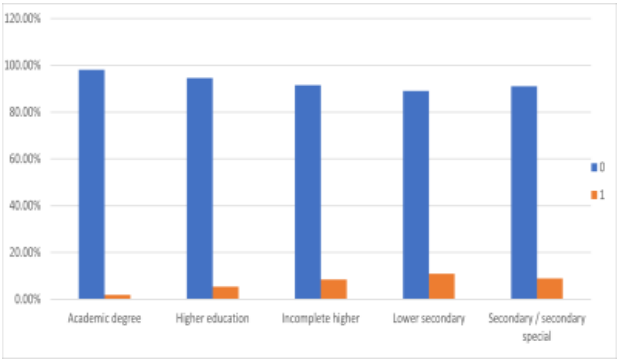
FAMILY MEMBER VS TARGET

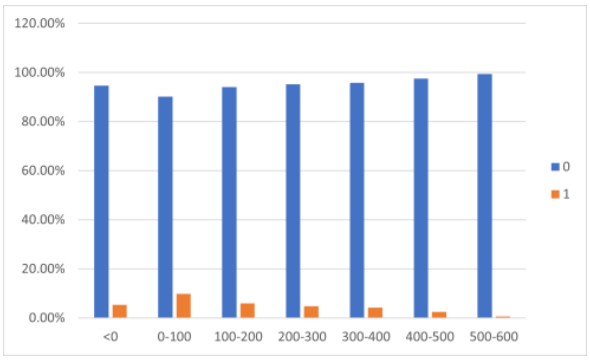


EDUCATION TYPE VS TARGET



MONTHS EMPLOYED VSTARGET

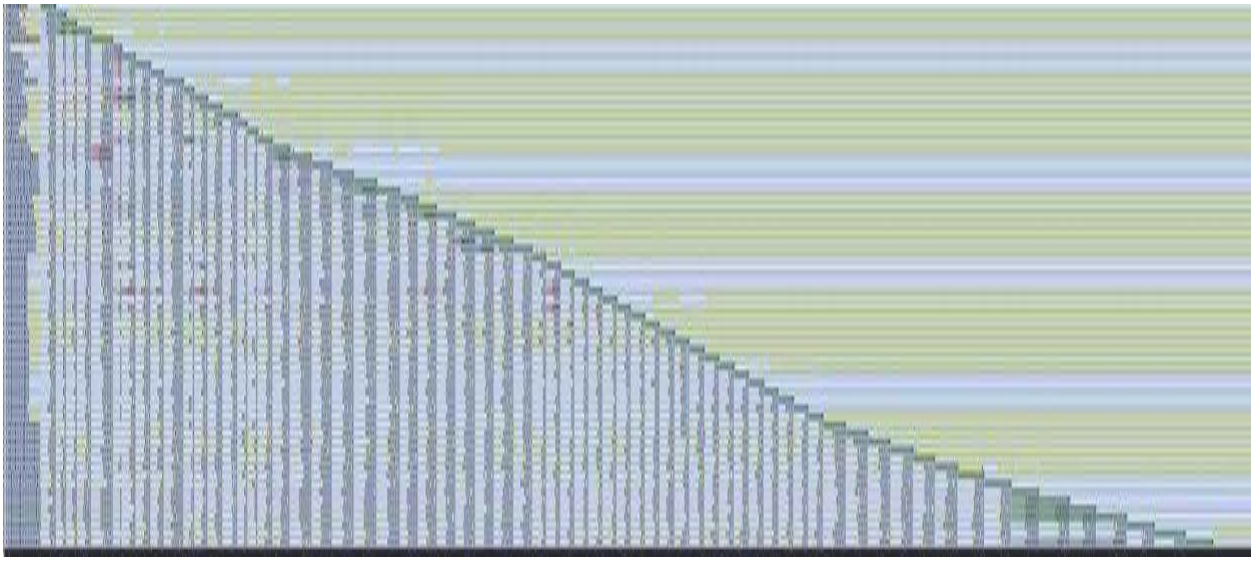




Task 6 (Finding top 10 correlations):

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



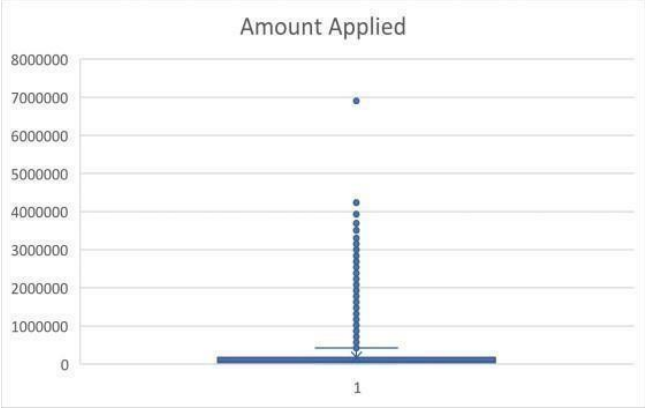
Earlier Application.csv

Task 2 (Data Cleaning):

Column removal: I used the COUNTBLANK function to determine the number of blanks in a column, and if it exceeded 5%, I eliminated it. I removed a couple columns that were of no use to the analysis

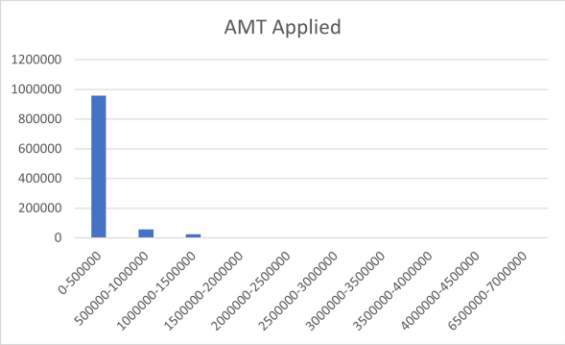
There are 1670214 rows in the dataset where as Excel has a Max limit of 1048576 rows and as per the project requirement we are supposed to use only Excel for Analysis. Hence we'd be limited to the use of 1048576 rows

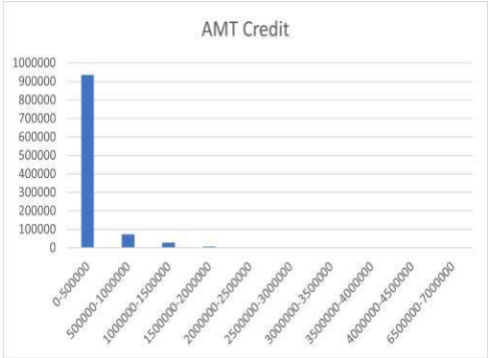
Task 3 (Finding Outliers):



Task 4(Data Imbalance):

Below are the columns where data is unevenly distributed





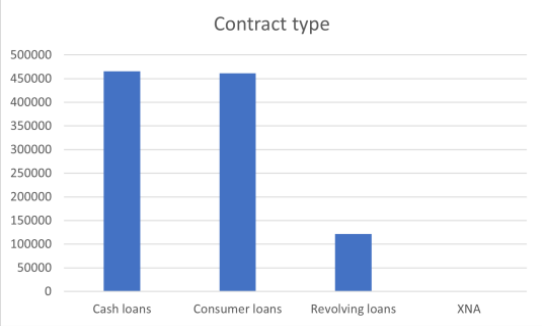
Task 5 (EDA):

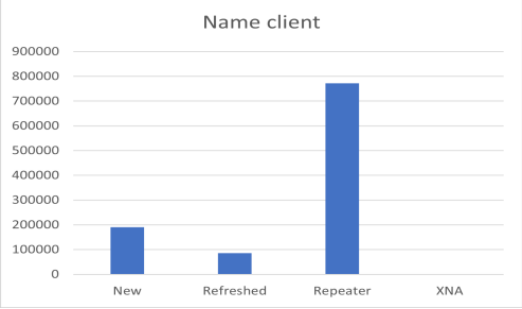
Univariate Analysis:

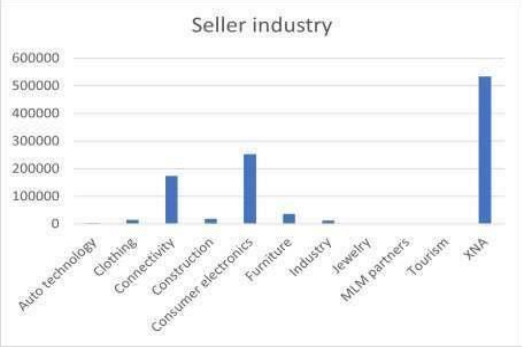
Inference

Customers have largely chosen cash and consumer loans. The majority of our clients are repeat customers.

The majority of current loan applicants are individuals who applied for loans less than ten months ago. More loans have been requested for consumer gadgets.



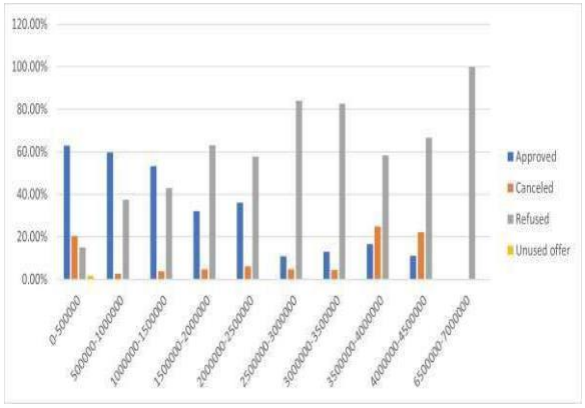


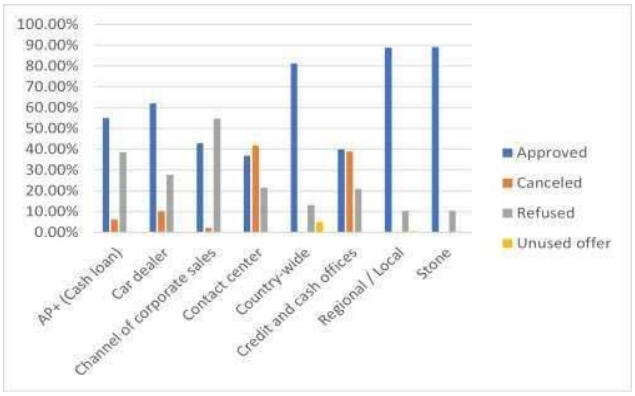


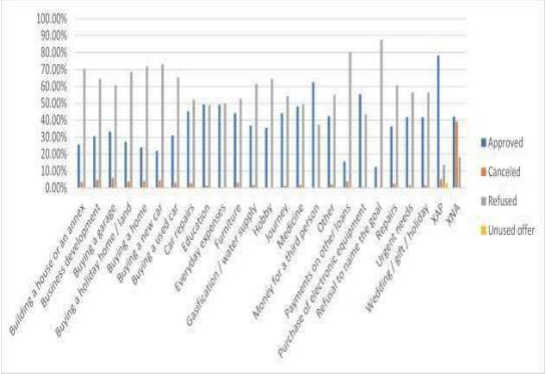
Bivariate Analysis:

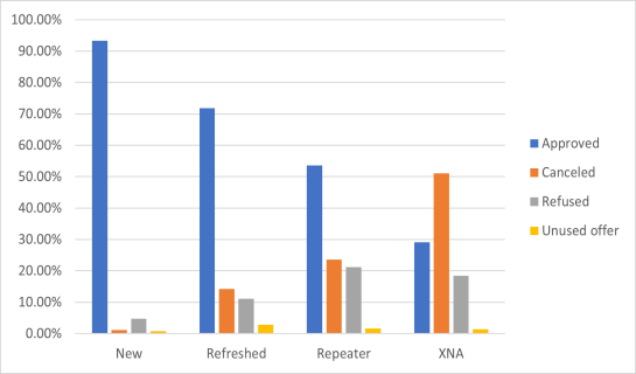
Inference

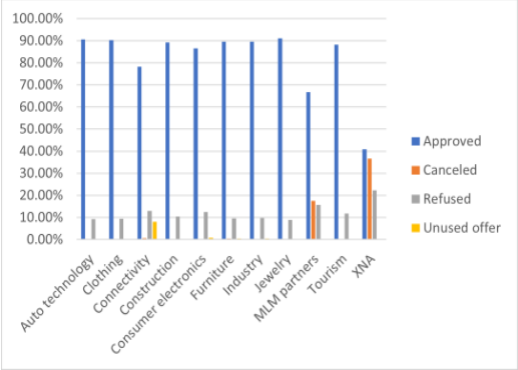
Customers who applied for more than Rs. 350,000 will most likely be denied. The majority of loans sought for through Credit and Cash agencies are cancelled.New clients are overjoyed because the majority of their loans were approved. Thus far, car loans have been denied. Loans made to MLM partner clients are likely to be cancelled. Virtually 80% of the loans were authorised, with a steady stream of rejections.Consumer loans have nearly no cancellations and the greatest approval rate. Several loans for the first Selling place area group were cancelled.Clients who apply for another loan within 10 months of their previous loan are more likely to have it cancelled. Walk-in loans have a higher refusal rate

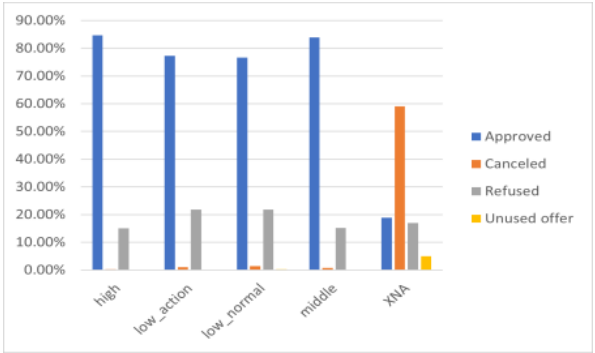


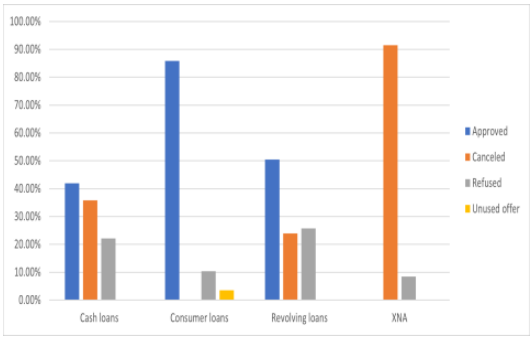


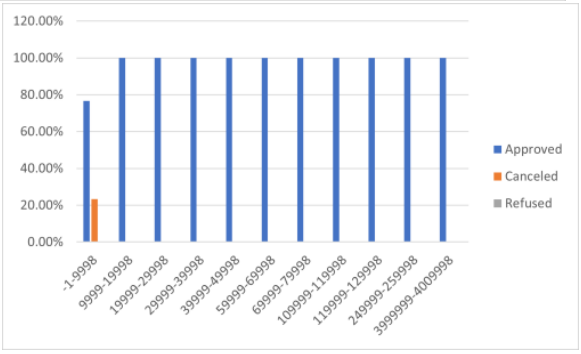


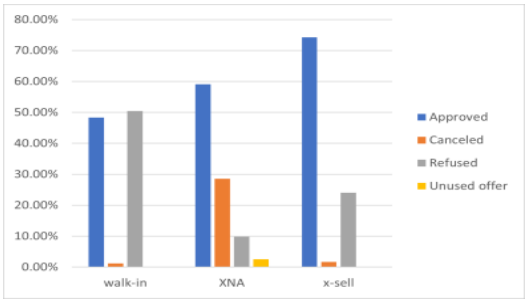


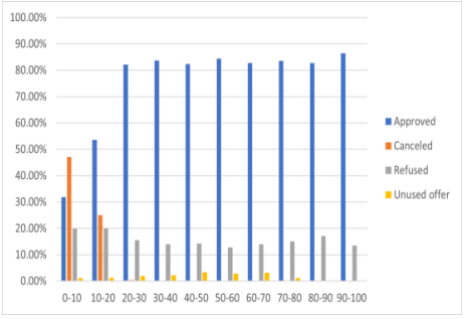






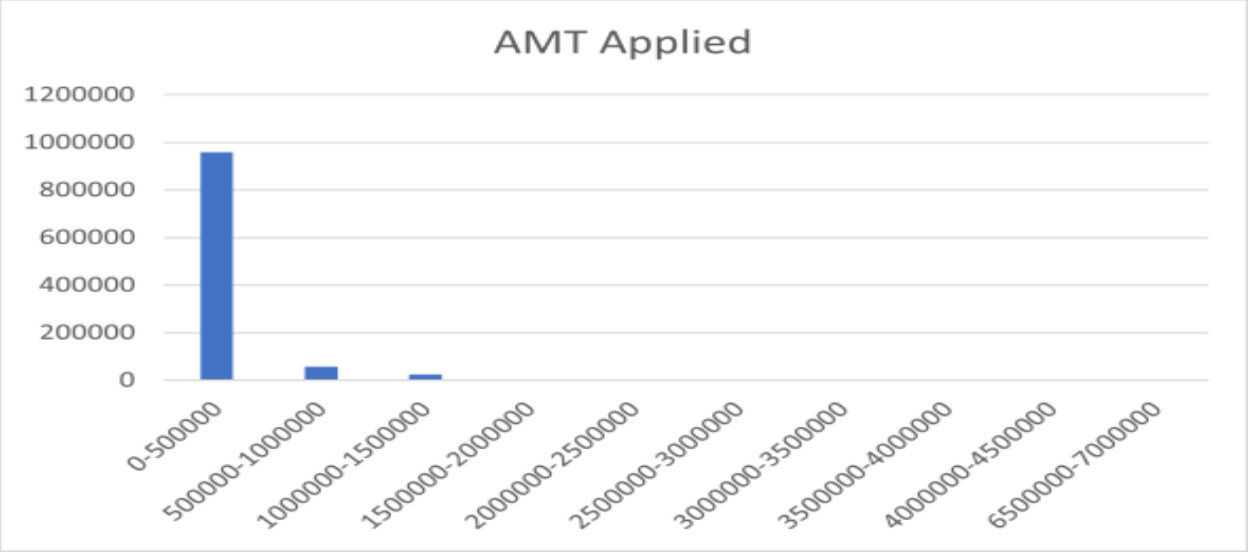






Task 6 (Finding Correlations): Top ten reasons for loan cancellation and refusal

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

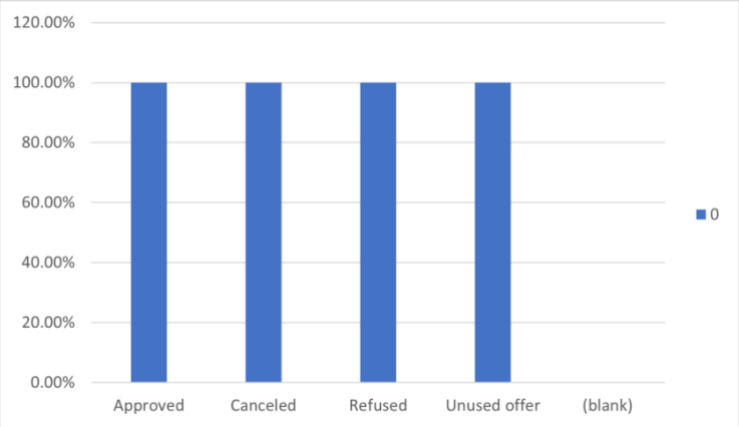


Task 7 (Combining two sheets): I then ran analysis on the common set of data by joining the Target column with the previous application table. I used MySQL to join them. I loaded the data into workbench and ran the following query.

Query:

```
SELECT TARGET,
SK_ID_CURR,
NAME_CONTRACT_TYPE,
AMT_APPLICATION,
NAME_CASH_LOAN_PURPOSE,
NAME_CONTRACT_STATUS,
NAME_CLIENT_TYPE, DAYS_DECISION,
CODE_REJECT_REASON,
NAME_SELLER_INDUSTRY,
NAME_PORTFOLIO,
NAME_PRODUCT_TYPE,
CHANNEL_TYPE, SELLERPLACE_AREA,
NAME_YIELD_GROUP,
PRODUCT_COMBINATION
FROM application_data
JOIN previous_application ON SK_ID_CURR;
```

pivot table analysis



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

Excel file:
https://docs.google.com/spreadsheets/d/1HqSNv0NdM7yg0Q0uC0IL7x_DsXt6JgLD/edit?usp=drive_link&ouid=115404029938861642621&rtpof=true&sd=true

RESULT:

This project involved extensive use of Excel. The major challenge was working with such huge data. This project helped me understand how to work with huge datasets. This helped me understand how 2 datasets are merged to analyze the details. The dataset involved a lot of missing data and outliers, handling them was a task and this project helped me understand what to how and why of handling the outliers and Null values. The project also helped me discover new add-ins such as data analyze.