BANK LOAN CASE STUDY

Project Description:

This case study aims to showcase the practical implementation of Exploratory Data Analysis (EDA) within a real-world business scenario. By applying EDA techniques acquired through the module, we will delve into the realm of risk analytics within the banking and financial services industry. Moreover, we will explore how data is effectively leveraged to mitigate the risk of financial losses associated with consumer lending.

Approach:

In this bank loan case study, we encountered two extensive data sets: the current application data and the previous application data. These datasets contained numerous columns that were deemed irrelevant for risk assessments, as well as a significant amount of missing data. Therefore, our initial step involved data cleaning and preprocessing.

After cleaning the data, we proceeded to evaluate the datasets. We identified and addressed any outliers present in the data, ensuring their removal to avoid skewing our analysis. To gain insights from the data, we conducted both univariate and bivariate analyses.

For the univariate analysis, we employed pivot tables and charts to examine individual variables and their distributions. This allowed us to understand the characteristics and patterns within each variable independently.

Moving on to the bivariate analysis, we investigated the relationships and dependencies between pairs of variables. By utilizing pivot tables and charts, we were able to identify any correlations or associations between different variables, which aided in understanding the interactions and potential influences among them.

Overall, by adopting this approach, we effectively cleaned the data, identified outliers, and conducted comprehensive univariate and bivariate analyses using pivot tables and charts. These steps provided us with valuable insights for further risk assessment in the bank loan case study.

Tech-Stack Used:

MS Excel – I used this tool because this tool is used to create graphical representation of the result and understand the result set better.

Jupyter Notebook -

Insights:

Task1:

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

- **Problem Statement**
  In this case study, we want to learn certain things. First, we want to find patterns that show if a client has trouble paying their loan installments. This information can help us make decisions like denying a loan, giving a smaller loan amount, or lending to risky applicants at a higher interest rate.

  Second, we want to identify the main factors that contribute to loan defaults. These factors are strong indicators that someone might not be able to repay their loan.

- **Analysis Approach**
  These are the following steps we have to do for analysis:
  - Import the datasets (Application_Data & Previous_Application) in jupyter notebook.
  - Analyze the data to determine the appropriate approach, locate any missing datasets, and work on them to achieve the desired results.
  - Detect outliers and demonstrate their impact on our dataset.
  - Understand the ratio of imbalance in our data.
  - Explore the correlation between the variables and the target variables, and identify the top three correlations.
  - Present data visually using charts and graphs.

Task2:

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

To check the %age of missing value plotting the column vs percentage of missing value of previous application data.

Sorted percentage of null values in each column

Out[40]:

	column_name	Percentage
14	RATE_INTEREST_PRIVILEGED	99.643698
13	RATE_INTEREST_PRIMARY	99.643698
6	AMT_DOWN_PAYMENT	53.636480
12	RATE_DOWN_PAYMENT	53.636480
20	NAME_TYPE_SUITE	49.119754
36	NFLAG_INSURED_ON_APPROVAL	40.298129
35	DAYS_TERMINATION	40.298129
34	DAYS_LAST_DUE	40.298129
33	DAYS_LAST_DUE_1ST_VERSION	40.298129
32	DAYS_FIRST_DUE	40.298129
31	DAYS_FIRST_DRAWING	40.298129
7	AMT_GOODS_PRICE	23.081773
3	AMT_ANNUITY	22.286665
28	CNT_PAYMENT	22.286366
30	PRODUCT_COMBINATION	0.020716

We find that, Rate_INTEREST_PRIMARY and RATE_INTEREST_PRIVILEGED have huge amount of null values that's why drop these column.

```
df1 = prev_app.drop('RATE_INTEREST_PRIMARY',axis=1)
```

```
df2 = df1.drop('RATE_INTEREST_PRIVILEGED',axis=1)
```

Then drop only the cells having null values of these columns

```
df3 = df2.dropna(subset=['AMT_DOWN_PAYMENT','RATE_DOWN_PAYMENT','NAME_TYPE_SUITE',
                         'NFLAG_INSURED_ON_APPROVAL','DAYS_TERMINATION','DAYS_LAST_DUE',
                         'DAYS_FIRST_DRAWING','AMT_GOODS_PRICE','CNT_PAYMENT'])
```

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE
13	1397919	321676	Consumer loans	7654.860	53779.5	57564.0	0.0	53779.5
15	1232483	151612	Consumer loans	21307.455	126490.5	119853.0	12649.5	126490.5
16	2163253	154602	Consumer loans	4187.340	26955.0	27297.0	1350.0	26955.0
28	2075578	418383	Consumer loans	7656.705	74610.0	65610.0	9000.0	74610.0
35	2027074	208000	Consumer loans	12065.535	56655.0	66411.0	0.0	56655.0

5 rows × 35 columns

These are the sorted null values percentage data of application_data we find it contains large amount of missing values.

	column_name	Percentage
76	COMMONAREA_MEDI	69.872297
48	COMMONAREA_AVG	69.872297
62	COMMONAREA_MODE	69.872297
70	NONLIVINGAPARTMENTS_MODE	69.432963
56	NONLIVINGAPARTMENTS_AVG	69.432963
...	...	...
15	NAME_HOUSING_TYPE	0.000000
14	NAME_FAMILY_STATUS	0.000000
13	NAME_EDUCATION_TYPE	0.000000
12	NAME_INCOME_TYPE	0.000000
0	SK_ID_CURR	0.000000

122 rows × 2 columns

I find the column having greater than 40% null values and I got there are 49 column which have null values greater than 40%.

	Column Name	Percentage of null values
21	OWN_CAR_AGE	65.990810
41	EXT_SOURCE_1	56.381073
44	APARTMENTS_AVG	50.749729
45	BASEMENTAREA_AVG	58.515956
46	YEARS_BEGINEXPLUATATION_AVG	48.781019
47	YEARS_BUILD_AVG	66.497784
48	COMMONAREA_AVG	69.872297
49	ELEVATORS_AVG	53.295980
50	ENTRANCES_AVG	50.348768
51	FLOORSMAX_AVG	49.760822
52	FLOORSMIN_AVG	67.848630
53	LANDAREA_AVG	59.376738
54	LIVINGAPARTMENTS_AVG	68.354953
55	LIVINGAREA_AVG	50.193326
56	NONLIVINGAPARTMENTS_AVG	69.432963
57	NONLIVINGAREA_AVG	55.179164
58	APARTMENTS_MODE	50.749729
59	BASEMENTAREA_MODE	58.515956
60	YEARS_BEGINEXPLUATATION_MODE	48.781019
61	YEARS_BUILD_MODE	66.497784
62	COMMONAREA_MODE	69.872297
63	ELEVATORS_MODE	53.295980
64	ENTRANCES_MODE	50.348768
65	FLOORSMAX_MODE	49.760822

```
len(blank2)
```

```
49
```

Task3:

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

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. An outlier can be identified from a box-

plot graph. If the value lies above the maximum and below the minimum, they are considered as outliers.

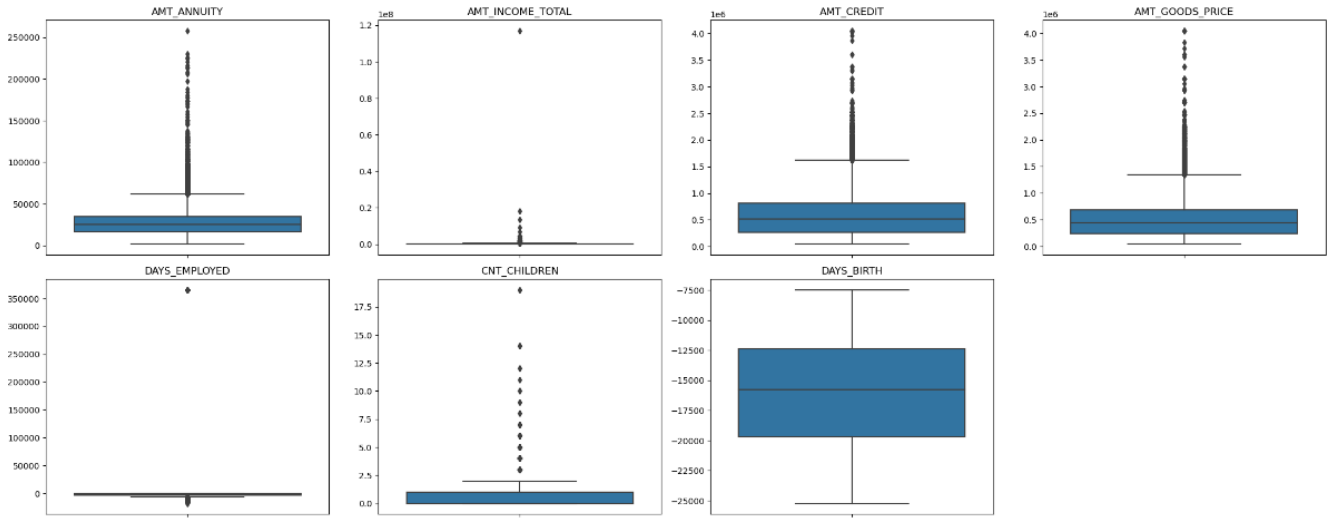
Application_data:

1. AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE, and CNT_CHILDREN exhibit some number of outliers, suggesting the presence of extreme values in these variables.

2. AMT_INCOME_TOTAL shows a significant number of outliers, indicating that a few loan applicants have considerably higher incomes compared to others.

3. DAYS_BIRTH does not contain any outliers, indicating that the available data for age is reliable and does not contain extreme values.

4. DAYS_EMPLOYED contains outlier values around 350,000 days, equivalent to approximately 958 years, which is impossible and likely an incorrect entry that needs to be addressed.



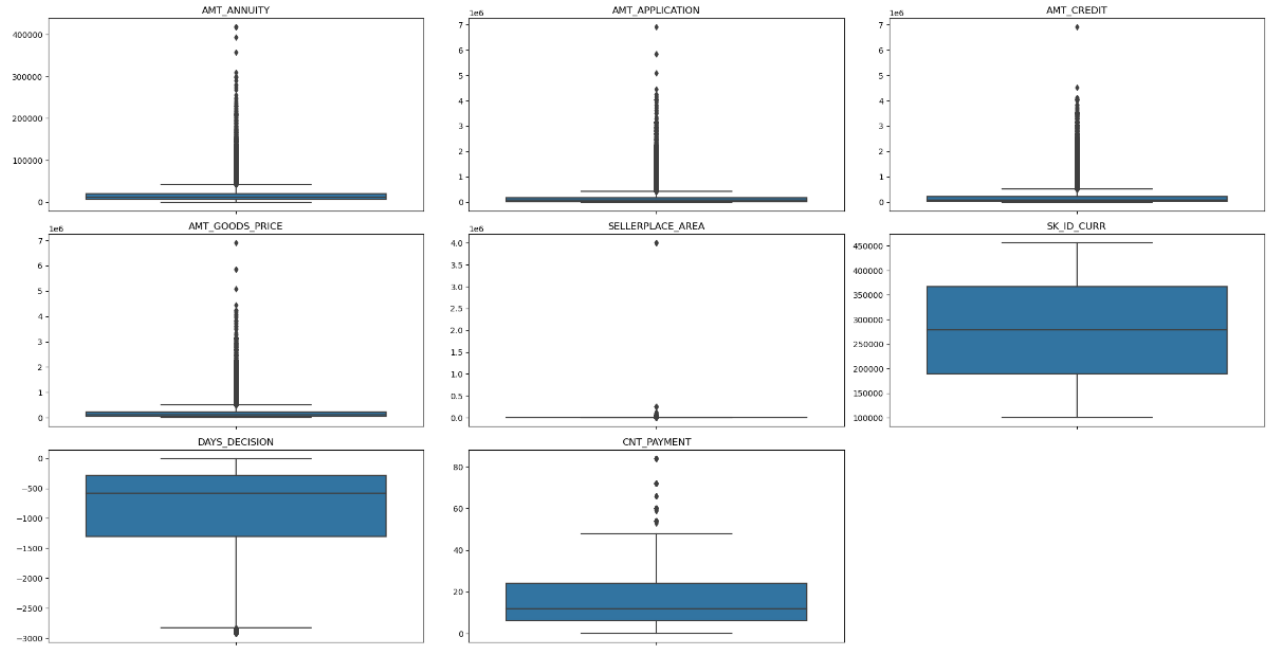
Privious_application:

1. Variables such as AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, and SELLERPLACE_AREA exhibit a significant number of outliers, indicating the presence of extreme values in these variables. These outliers suggest the existence of exceptional cases or unusual data points in these categories.

2. The variable CNT_PAYMENT has a few outlier values, suggesting that there are a limited number of instances where the payment count differs significantly from the majority of cases.

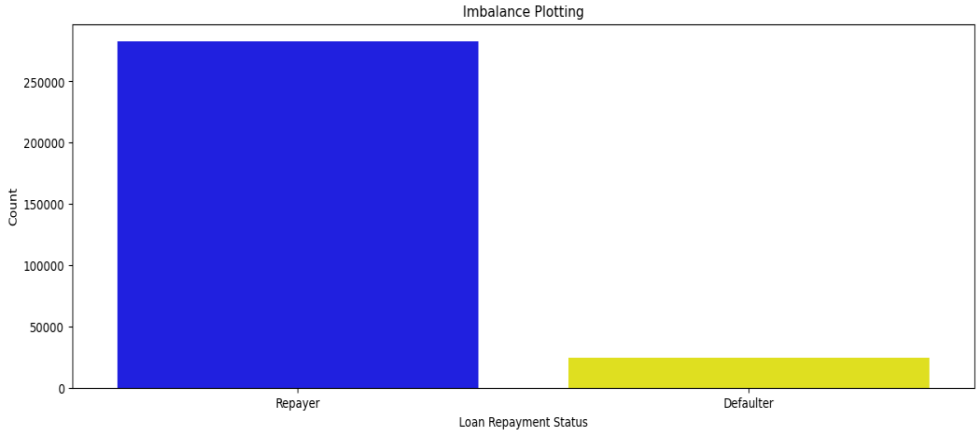
3. SK_ID_CURR represents an ID column and typically does not have outliers. It serves as an identifier for each individual and doesn't contain numerical values that can be considered outliers.

4. The variable DAYS_DECISION shows a relatively small number of outliers, indicating that the decisions for these previous applications were made a long time ago. These outliers may represent exceptional cases with significantly longer or shorter decision times compared to the majority of applications.



Task4:

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



This data is highly imbalanced as number of defaulter is very less in total population.

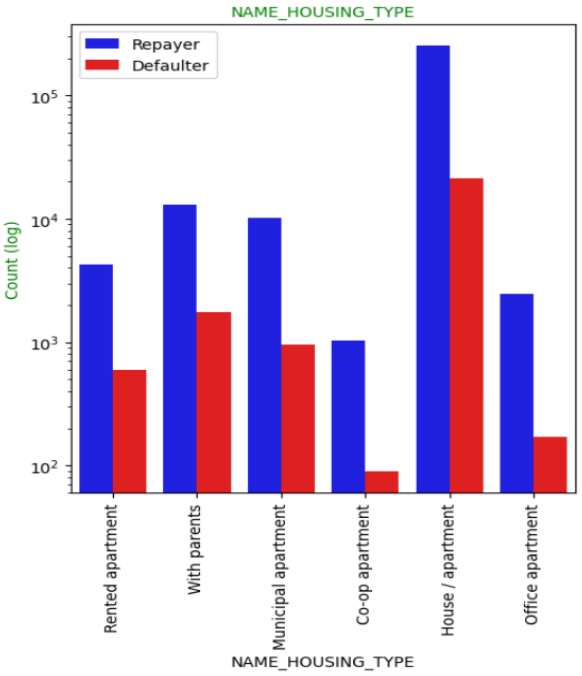
```
Imbalance ratio in percentage with respect to Repayer and Defaulter datas are: 91.93 and 8.07
```

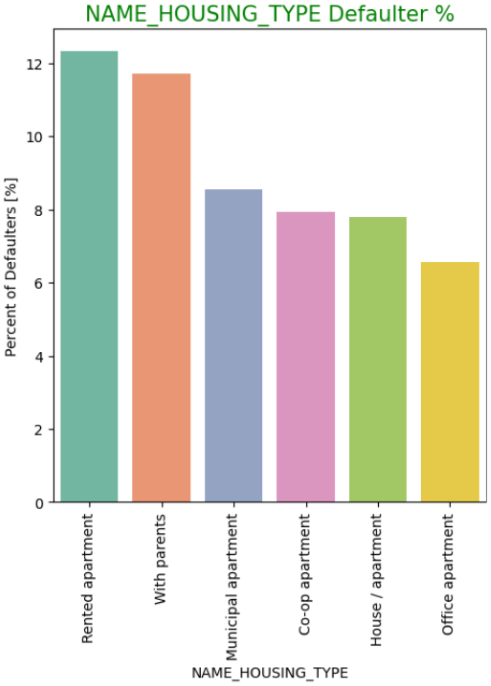
Task5:

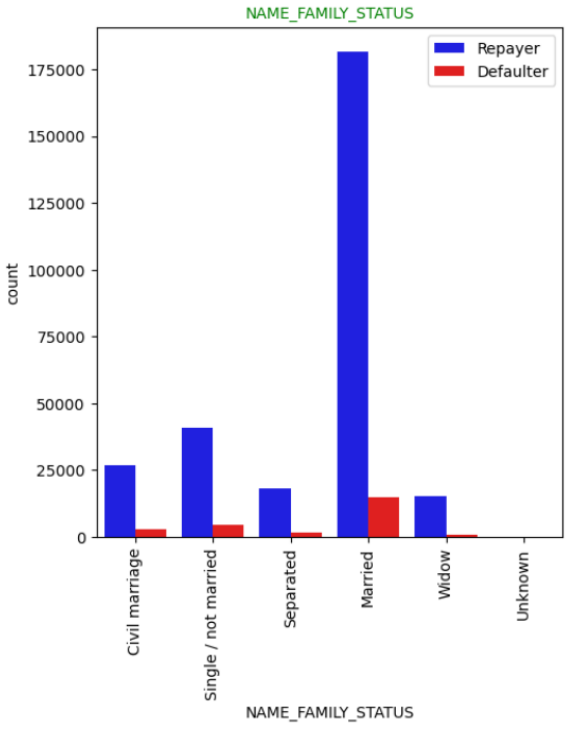
Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.

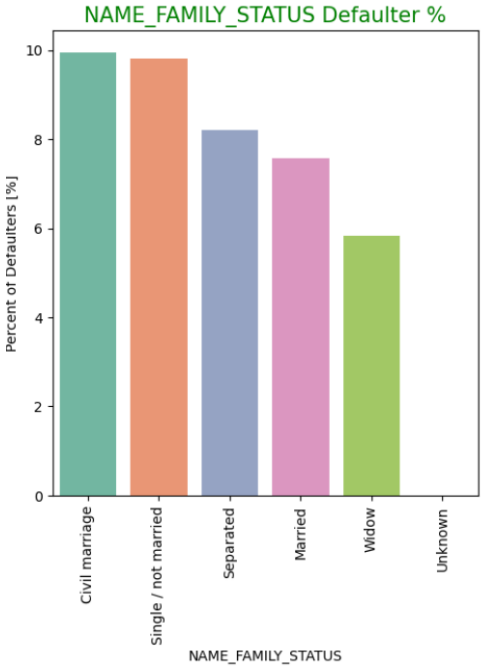
Univariate Analysis:

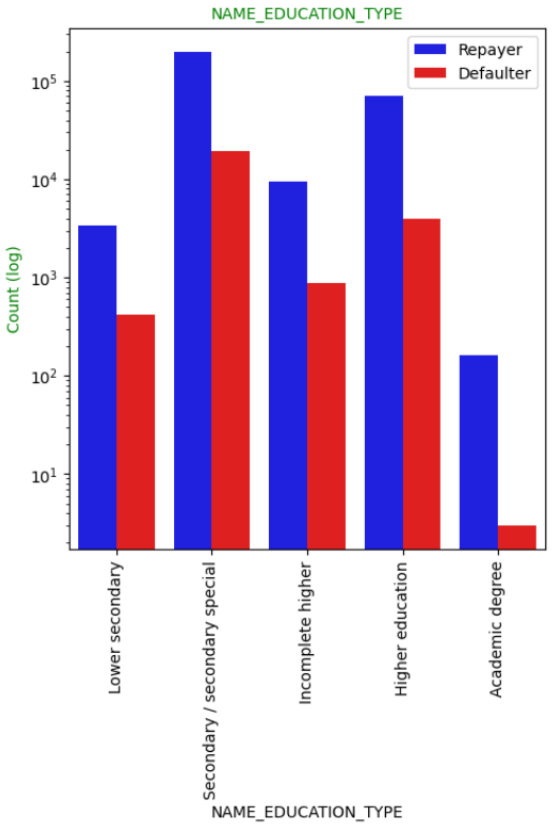
Univariate analysis is a method used to analyze data by focusing on a single variable at a time. It aims to describe the data and identify patterns within that particular variable. Instead of examining relationships between variables, univariate analysis concentrates on summarizing and understanding the distribution, central tendency (typical value), and variability of the data. It uses statistical measures and graphical representations to achieve this. By studying one variable in isolation, univariate analysis helps us identify trends and patterns that can be useful for further analysis and decision-making.

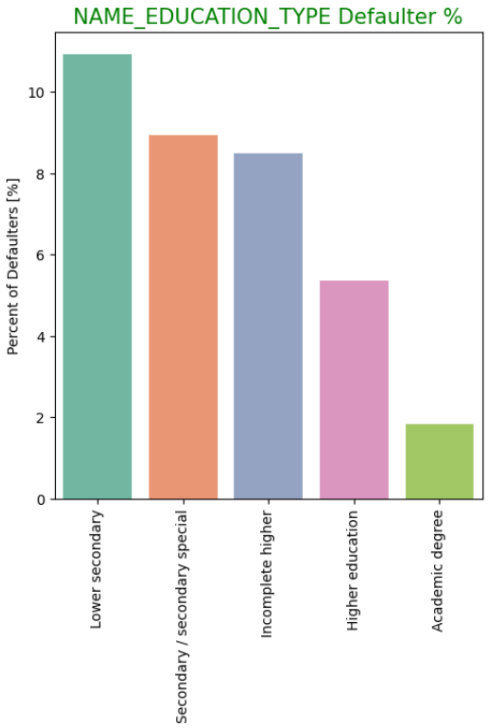

NAME_HOUSING_TYPE
Repayer
Defaulter
Count (log)
10^5
10^4
10^3
10^2
Rented apartment
With parents
Municipal apartment
Co-op apartment
House / apartment
Office apartment
NAME_HOUSING_TYPE

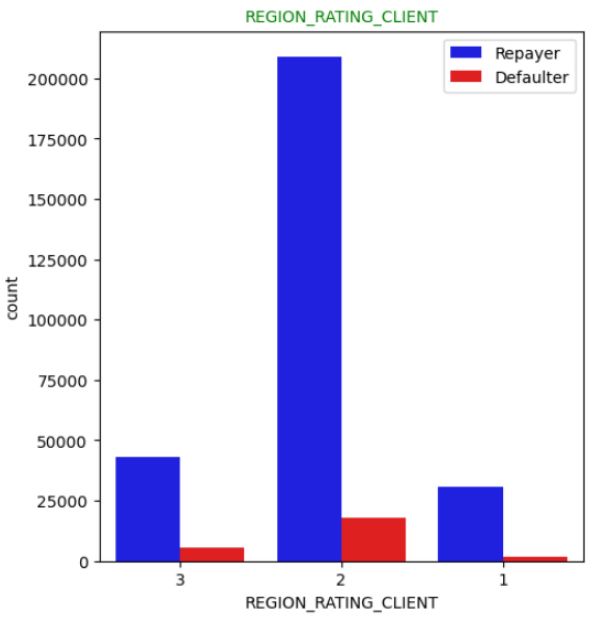

NAME_HOUSING_TYPE Defaulter %
Percent of Defaulters [%]
0
2
4
6
8
10
12
Rented apartment
With parents
Municipal apartment
Co-op apartment
House / apartment
Office apartment
NAME_HOUSING_TYPE

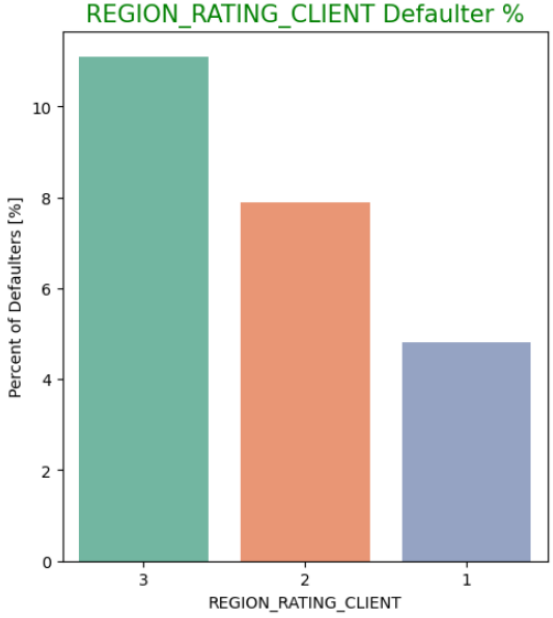

NAME_FAMILY_STATUS
Repayer
Defaulter
count
0
25000
50000
75000
100000
125000
150000
175000
Civil marriage
Single / not married
Separated
Married
Widow
Unknown
NAME_FAMILY_STATUS

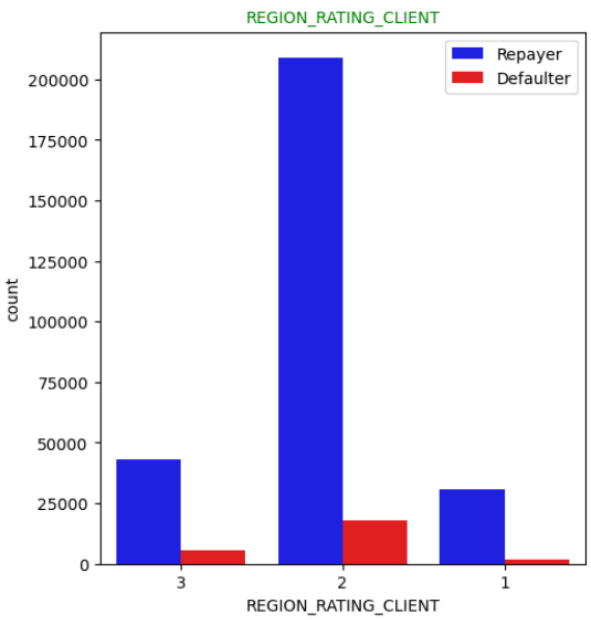

NAME_FAMILY_STATUS Defaulter %
Percent of Defaulters [%]
0
2
4
6
8
10
Civil marriage
Single / not married
Separated
Married
Widow
Unknown
NAME_FAMILY_STATUS

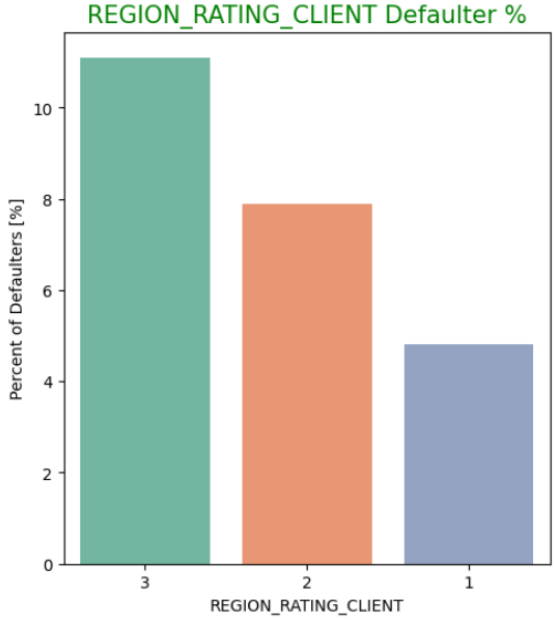

NAME_EDUCATION_TYPE
Repayer
Defaulter
Count (log)
10^5
10^4
10^3
10^2
10^1
Lower secondary
Secondary / secondary special
Incomplete higher
Higher education
Academic degree
NAME_EDUCATION_TYPE

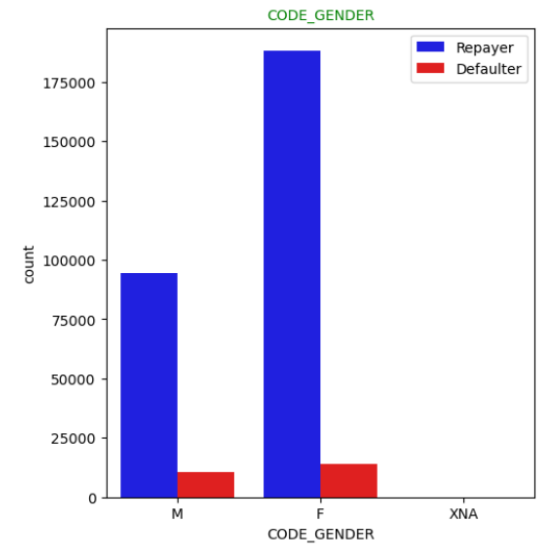

NAME_EDUCATION_TYPE Defaulter %
Percent of Defaulters [%]
0
2
4
6
8
10
Lower secondary
Secondary / secondary special
Incomplete higher
Higher education
Academic degree

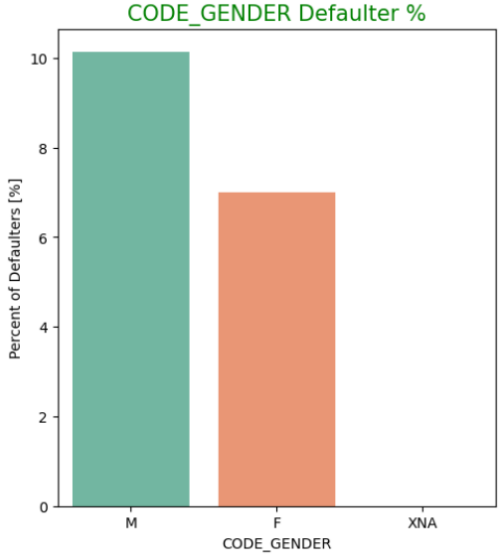

REGION_RATING_CLIENT
Repayer
Defaulter
count
0
25000
50000
75000
100000
125000
150000
175000
200000
3
2
1
REGION_RATING_CLIENT

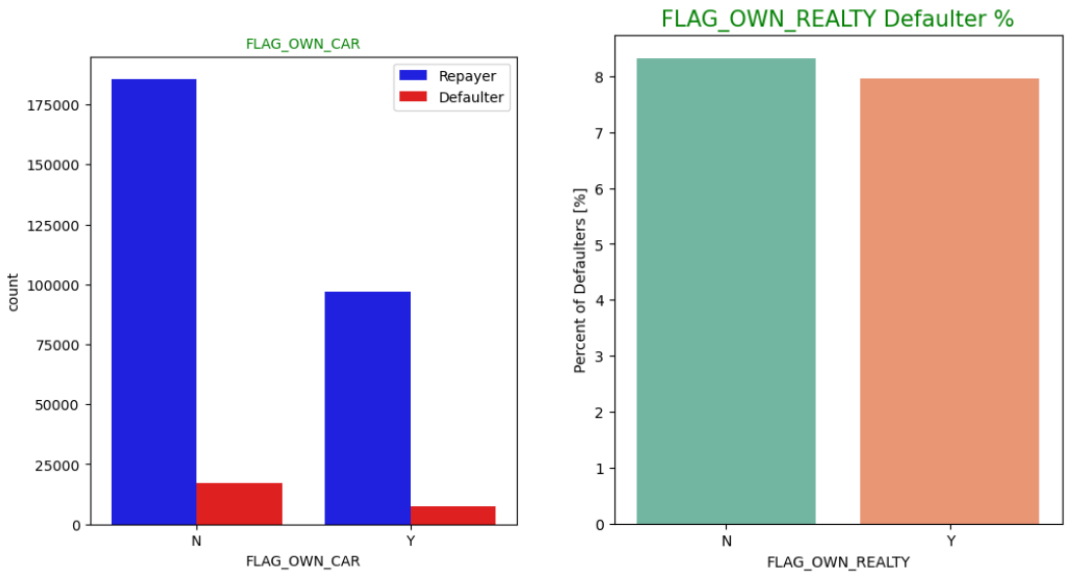

REGION_RATING_CLIENT Defaulter %
Percent of Defaulters [%]
0
2
4
6
8
10
3
2
1
REGION_RATING_CLIENT


REGION_RATING_CLIENT
Repayer
Defaulter
count
200000
175000
150000
125000
100000
75000
50000
25000
0
3
2
1
REGION_RATING_CLIENT


REGION_RATING_CLIENT Defaulter %
Percent of Defaulters [%]
10
8
6
4
2
0
3
2
1
REGION_RATING_CLIENT

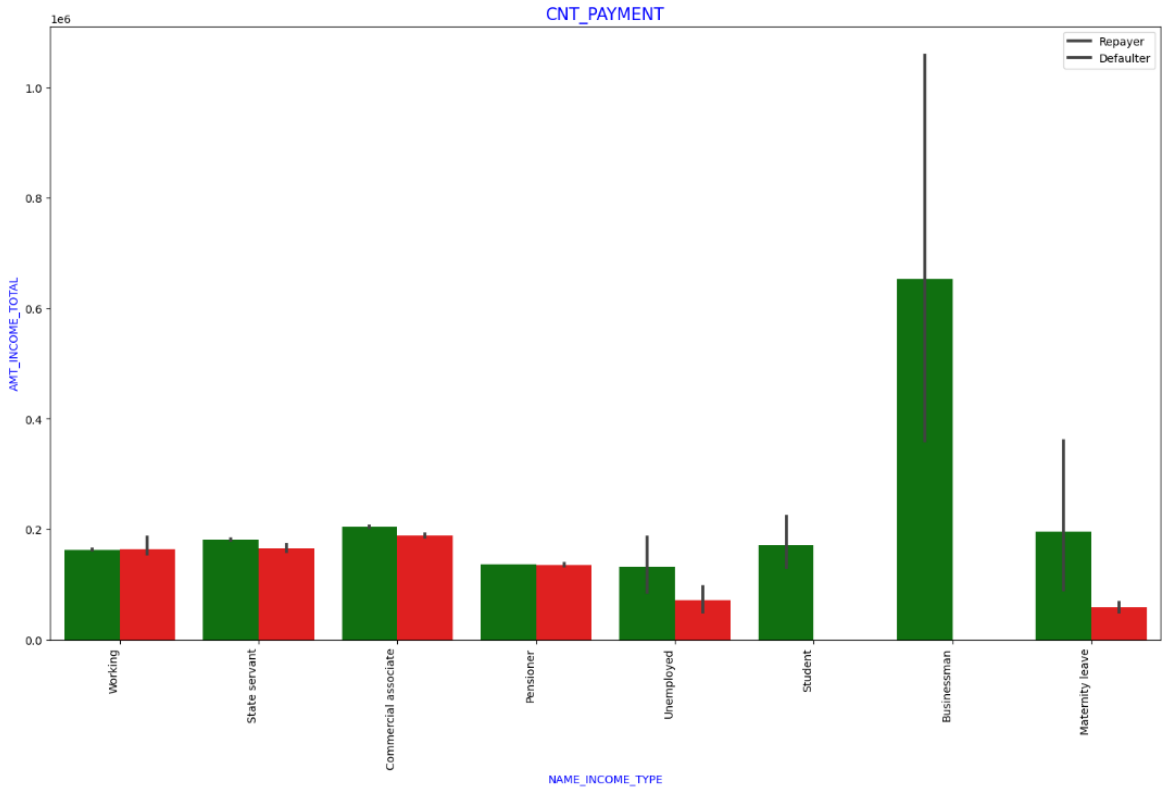

CODE_GENDER
Repayer
Defaulter
count
175000
150000
125000
100000
75000
50000
25000
0
M
F
XNA
CODE_GENDER


CODE_GENDER Defaulter %
Percent of Defaulters [%]
10
8
6
4
2
0
M
F
XNA
CODE_GENDER



Bivariate Analysis:

- Bivariate analysis is a statistical technique used to explore the relationship between two variables. It involves analyzing and studying the data from two different variables simultaneously. The primary objective of bivariate analysis is to understand how the values of one variable are related to the values of another variable.
  I have taken NAME_INCOME_TYPE and AMT_INCOME_TOTAL columns in application data.



Task6:

Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable).

The top 10 correlation for the Client with repayment:

	VAR1	VAR2	Correlation
122	AMT_GOODS_PRICE	AMT_CREDIT	0.987250
371	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950149
300	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
495	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861
588	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.830381
123	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686
92	AMT_ANNUITY	AMT_CREDIT	0.771309
216	DAYS_EMPLOYED	DAYS_BIRTH	0.618048
335	REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	0.539005
365	REGION_RATING_CLIENT_W_CITY	REGION_POPULATION_RELATIVE	0.537301

Client with default:

	VAR1	VAR2	Correlation
122	AMT_GOODS_PRICE	AMT_CREDIT	0.983103
371	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637
300	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
495	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.847885
588	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.778540
123	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699
92	AMT_ANNUITY	AMT_CREDIT	0.752195
216	DAYS_EMPLOYED	DAYS_BIRTH	0.575097
464	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.497937
557	REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.472052

.

1. . Clients who make repayments:

- Strong correlation between credit amount and goods price
- Correlation between credit amount and repayments
- Strong correlation between loan annuity and credit amount
- Strong correlation between total income and credit amount
- Strong correlation between age and the number of children

- Slightly weaker correlation between defaulted and observed counts in the social circle variable compared to defaulters

2. Default clients:

- Strong correlation between credit amount and goods price

- Correlation between credit amount and repayments

- Weaker correlation between loan annuity and credit amount compared to repayments

- Weaker correlation between total income and credit amount compared to clients who make repayments

- Weaker correlation between age and the number of children compared to clients who make repayments

- Slightly stronger correlation between defaulted and observed counts in the social circle variable compared to repayments

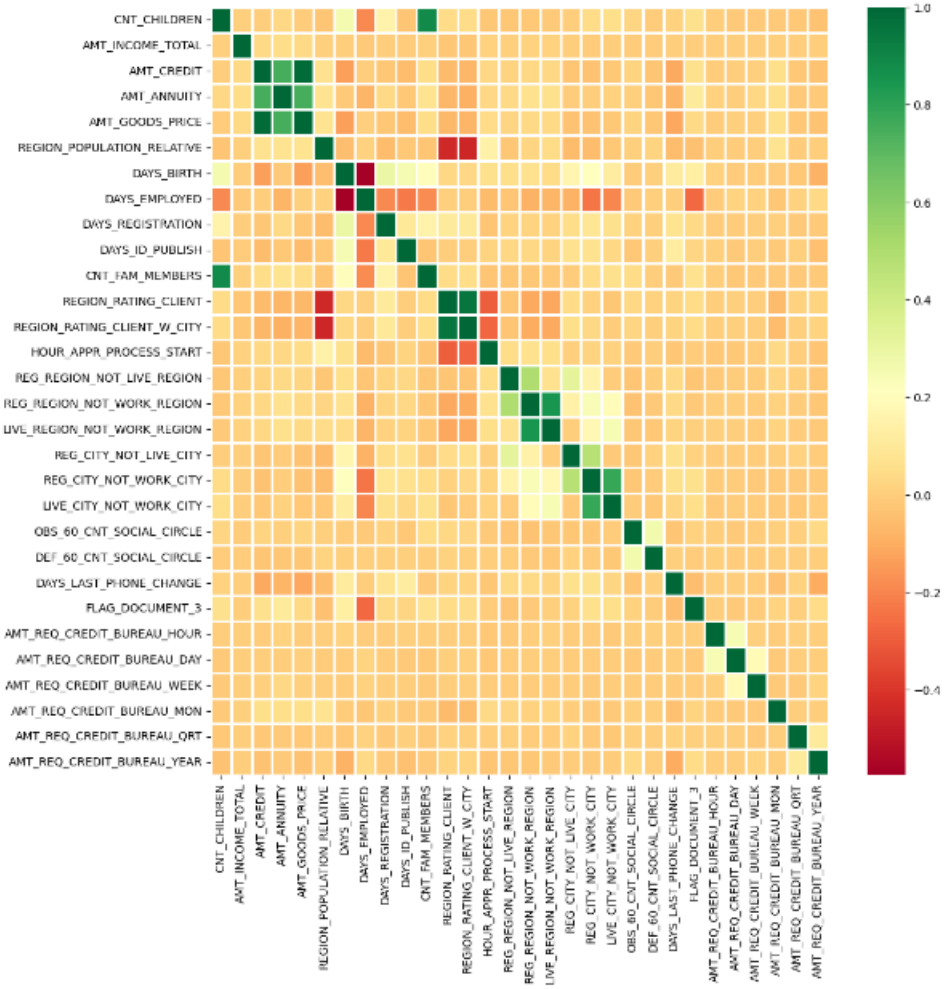
Task7:

Include visualizations and summarize the most important results in the presentation

Repairs:

1. Loan repayers demonstrate a strong positive correlation between the credit amount and the following factors: the amount of goods price, loan annuity, and total income. This indicates that repayers tend to have higher credit amounts associated with larger goods prices, higher loan annuities, and greater total income levels.
2. Additionally, there is a notable correlation between the number of days employed and loan repayers. This suggests that individuals who have been employed for a longer duration are more likely to be successful in repaying their loans.

Defaulters:



1. Credit amount is highly correlated with the amount of goods price for both repayers and defaulters.

2. The link between the loan installment amount and the credit amount is slightly weaker for defaulters compared to those who repay their loans successfully.

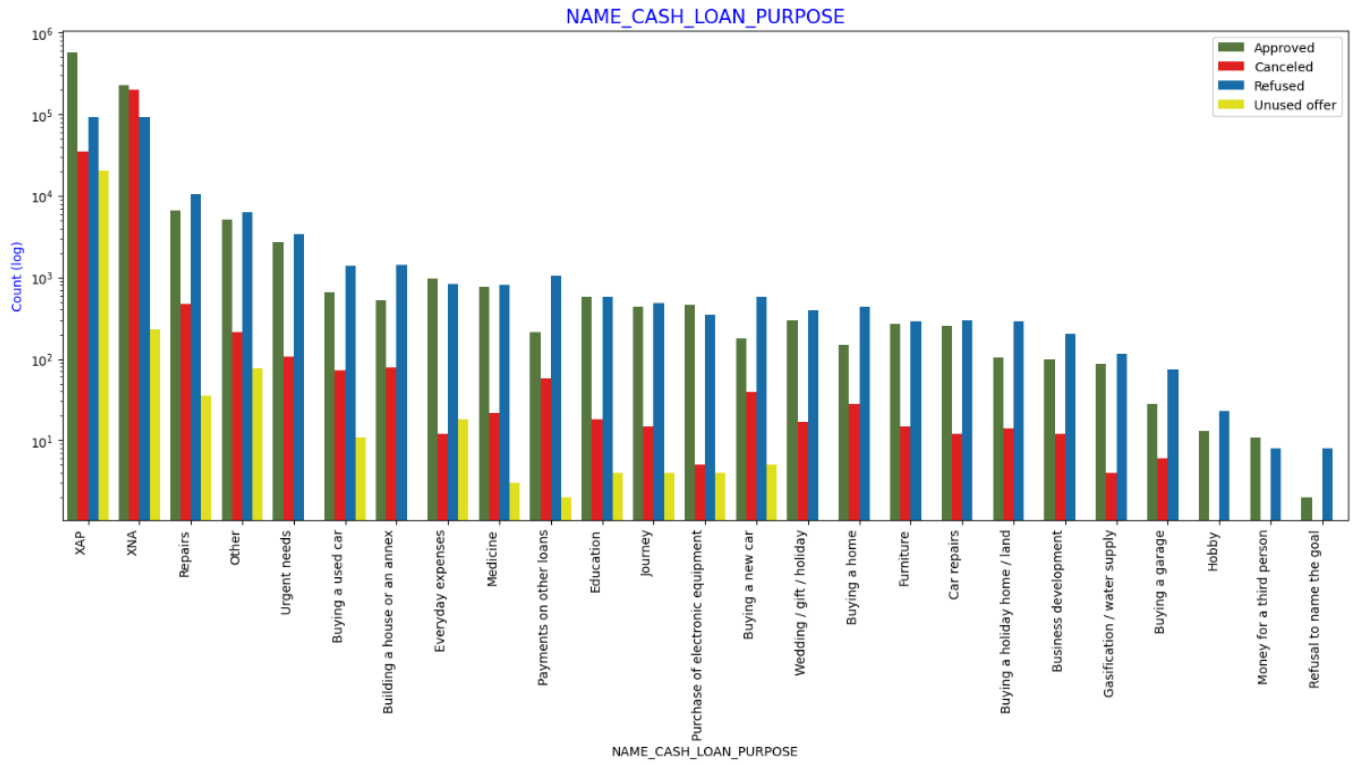
3. Repayers have a higher correlation between the number of days employed and loan repayment than defaulters.

4. The correlation between total income and credit amount is significantly lower for defaulters than repayers.

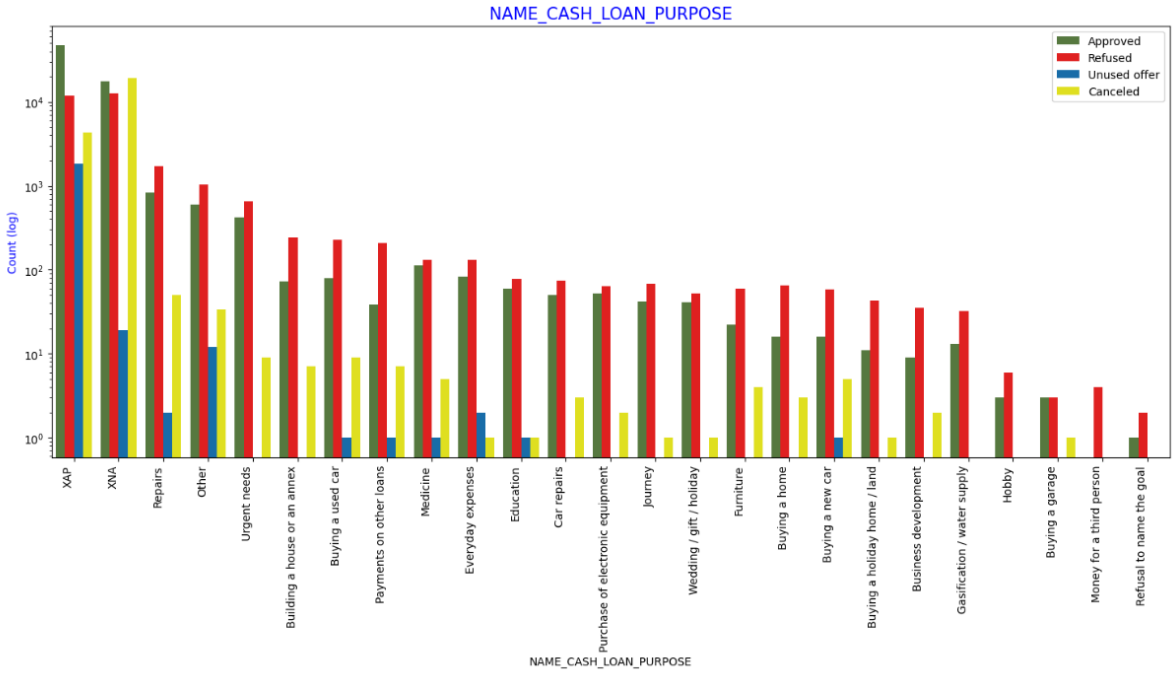
5. The correlation between age (days_birth) and the number of children is weaker for defaulters compared to repayers.

6. There is a slight increase in the correlation between the count of defaulted loans and the count of observed individuals in the social circle for defaulters compared to repayers.

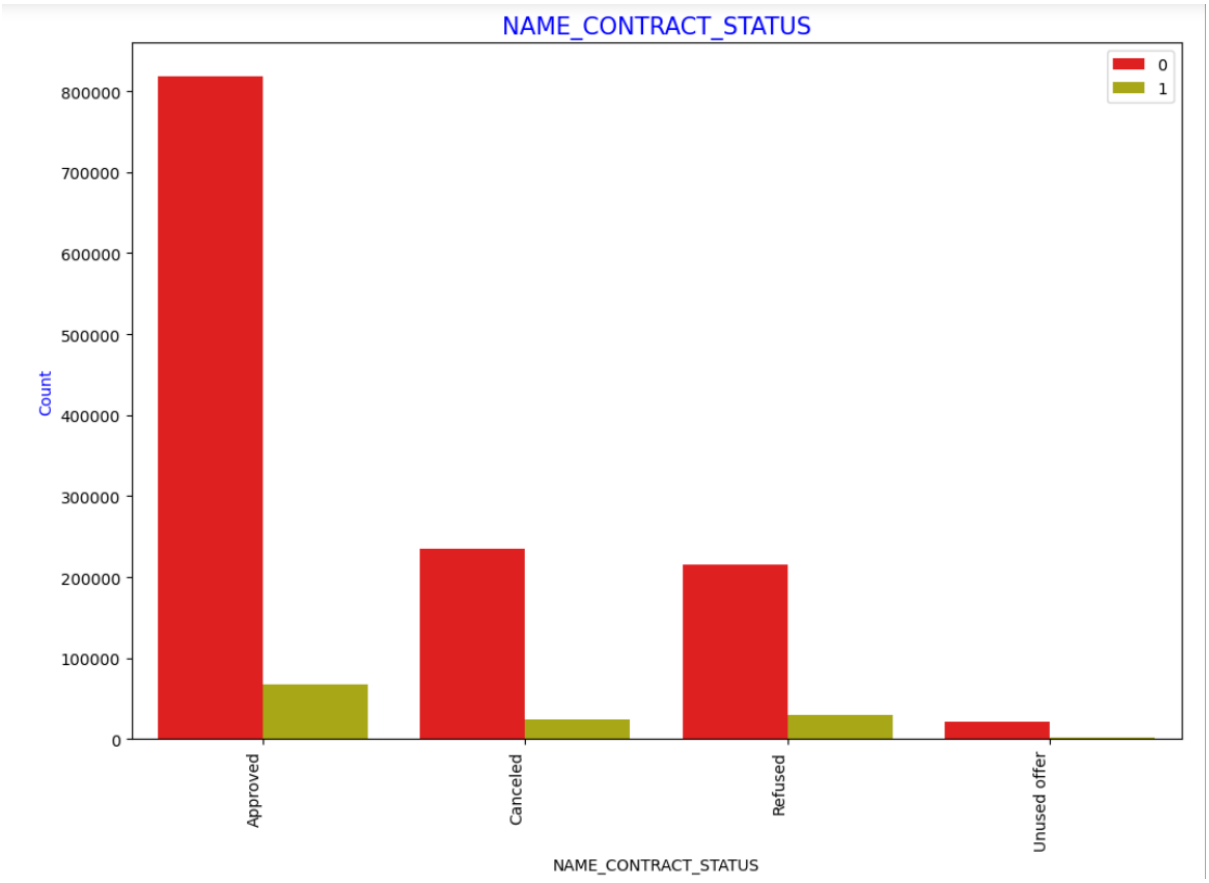
Repairs:



Defaulters:



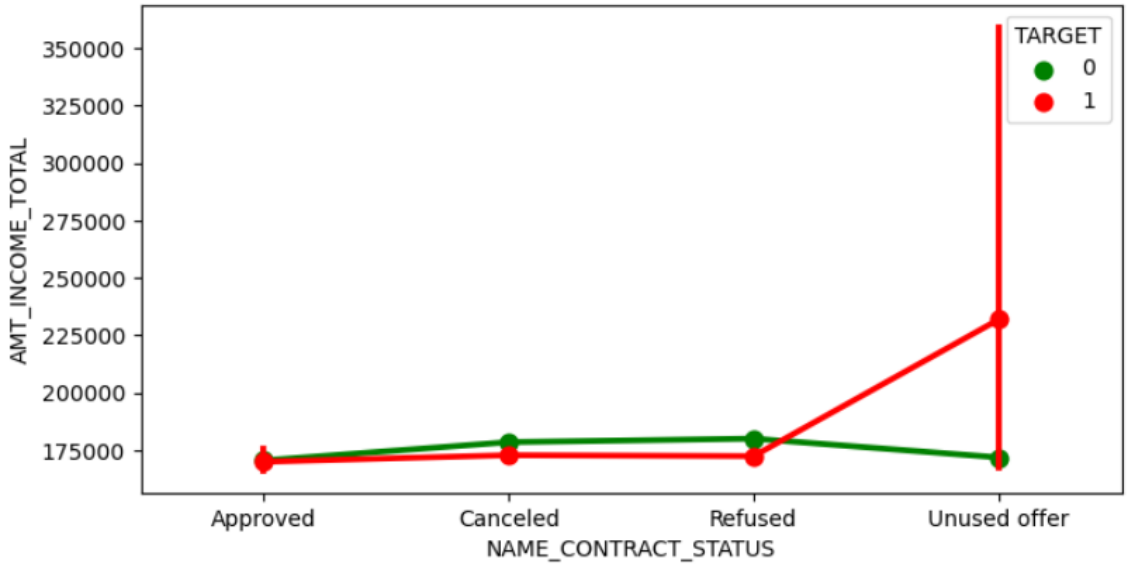
A significant number of loan applications have been rejected by the bank or refused by the clients when the purpose of the loan is stated as "repair" or other similar purposes. This suggests that the bank considers loan applications for repair purposes to be high-risk. As a result, these applications are either rejected outright or the bank offers loan interest rates that are perceived as unfeasible by the clients, leading them to refuse the loan. This indicates that the purpose of repair is viewed as a high-risk category by the bank, resulting in a higher likelihood of loan rejections or unaffordable loan terms for clients.



NAME_CONTRACT_STATUS	TARGET	Counts	Percentage
Approved	0	818856	92.41%
	1	67243	7.59%
Canceled	0	235641	90.83%
	1	23800	9.17%
Refused	0	215952	88.0%
	1	29438	12.0%
Unused offer	0	20892	91.75%
	1	1879	8.25%

1. Among the clients who had their loan applications previously cancelled, approximately 90% of them have successfully repaid their loans in the current case. This indicates a high likelihood of repayment for these clients. Therefore, revisiting the interest rates offered to these clients could potentially lead to increased business opportunities and a higher chance of loan repayment.
2. For clients who were previously refused a loan, around 88% of them have successfully repaid the loan in the current case. This suggests that despite being declined in the past, these clients have demonstrated their ability to repay the loan. Recording the reasons for previous loan refusals would be beneficial for further analysis. These clients have the

potential to become reliable customers who are more likely to repay their loans.



The point plot show that the people who have not used offer earlier have defaulted even when there average income is higher than others

Results:

In this case study, I conducted an Exploratory Data Analysis (EDA) in a real business scenario related to bank loans. Through this project, I gained valuable insights and learned several important concepts in risk analytics within the banking and financial services sector.

The key points that I have learned from this project:

1. Understanding Risk Analytics: I gained a fundamental understanding of how data analysis is employed in the banking and financial services industry to mitigate the risk of financial losses when granting loans to customers.

2. Data Summarization: This case study provided me with the opportunity to learn how to effectively summarize large datasets to extract meaningful insights. By employing various techniques, I was able to condense the data and derive valuable conclusions.

3. Dealing with Challenges: The project presented several challenges that I successfully addressed. One of the challenges involved studying the correlation between different variables in the dataset to uncover crucial insights for the clients.

4. Data Imbalance and Outliers: I encountered issues related to data imbalance and outliers, which allowed me to understand their impact on the analysis. I learned strategies to handle these challenges and ensure the accuracy of the results.

6. Data Visualization: I developed expertise in visualizing large datasets, which proved vital in summarizing and presenting the most important results to the client. Through visual representations, I effectively communicated the insights gained from the analysis.

Overall, this bank loan case study equipped me with a solid foundation in risk analytics, data summarization, correlation analysis, dealing with data imbalances and outliers, identification of driving factors, and data visualization. These skills will enable me to make informed decisions and provide valuable insights in similar business scenarios in the future.

Here I attached my jupyter notebook file: jupyter notebook file

THANK YOU