

BANK LOAN CASE STUDY

Project Description:

In this project, we have performed Exploratory Data Analysis (EDA) on Bank Loan Dataset. This case study aims to give you an idea of applying EDA in a real business scenario. The loan-providing companies find it hard to give loans to people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming defaulters.



Business Understanding:

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter.

Suppose we work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

- ❖ When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
 - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
 - If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.
- ❖ When a client applies for a loan, there are **four** types of decisions that could be taken by the client/company:

1. **Approved:** The company has approved loan application
2. **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
3. **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
4. **Unused Offer:** Loan has been cancelled by the client but on different stages of the process.

In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

❖ **From this project, we have to discover some insights such as-**

- Which type of applicants are likely to pay the loan amount and which type of applicants are taking advantage of the loan amount?
- Identifying such defaulters and taking strict actions like denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected.
- Presenting approach of removing non-useful columns, and null values, identifying outliers and data imbalance in the data, explaining the results of univariate, segmented univariate, bivariate analysis, etc. in business terms, drawing out top 10 correlation for the client with payment difficulties and all other cases.
- Discovering such results by performing EDA to help the company to decide whether to approve or refuse the loan application.

Tech-Stack used:

1. **MS Excel:** It is useful for understanding what the dataset exactly means, also useful in analysis work when the data is small, and for understanding the columns' description of data.
2. **Jupyter notebook:** Jupyter notebook is an interactive environment for arranging notebooks, code, and data using many programming languages. I have used it for whole analysis work using the **python** language.
3. **Python:** Python is a simple and easy programming language used in web development and machine learning applications, etc.
4. **MS word:** It is used for making the report.

Approach:

Steps -

1. Used MS Excel to understand what the dataset is about and its columns' meaning, which is important.
2. Then imported the required python libraries (Pandas, Numpy, Seaborn, etc.) into Jupyter notebook.
3. After libraries, datasets (Application_Data and Previous_Application) are imported.
4. Identified missing values and removed columns which are not useful for analysis work.
5. Identified outliers of different columns and their relations.
6. Identified data imbalance and its ratio and presented using the graphs.
7. Performed univariate, segmented univariate, and bivariate analysis of data.
8. Found out the top 10 correlations with respect to TARGET variables.
9. Lastly presented all the senseful data (analysis work) in the form of different graphs and charts.

You are required to provide a detailed report for the below data record mentioning the answer to the questions that follows:

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

- **Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)**
 1. There are two datasets namely 'Applicaton_Data' and 'Previous_Application'.
 2. Initially calculated the null value percentage of each column and found out the columns containing null values above 40% in Application data are 64 columns and in Previous Application are 11 columns.
- ❖ In Application_Data –
 3. In application_data, we found 'AMT_GOODS_PRICE' as a useful column in data analysis work that's why using the **median()** operation we imputed the missing values in the 'AMT_GOODS_PRICE' columns.
 4. Similarly, the missing values in the column 'AMT_ANNUITY' is imputed using the same **median()** operation.
 5. In **median imputation**, the missing values are replaced with the **median value** of the entire feature column.
 6. The columns like 'FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_DOCUMENT_1' and others which are not useful for analysis work are dropped off from data frame.

❖ In Previous_Application –

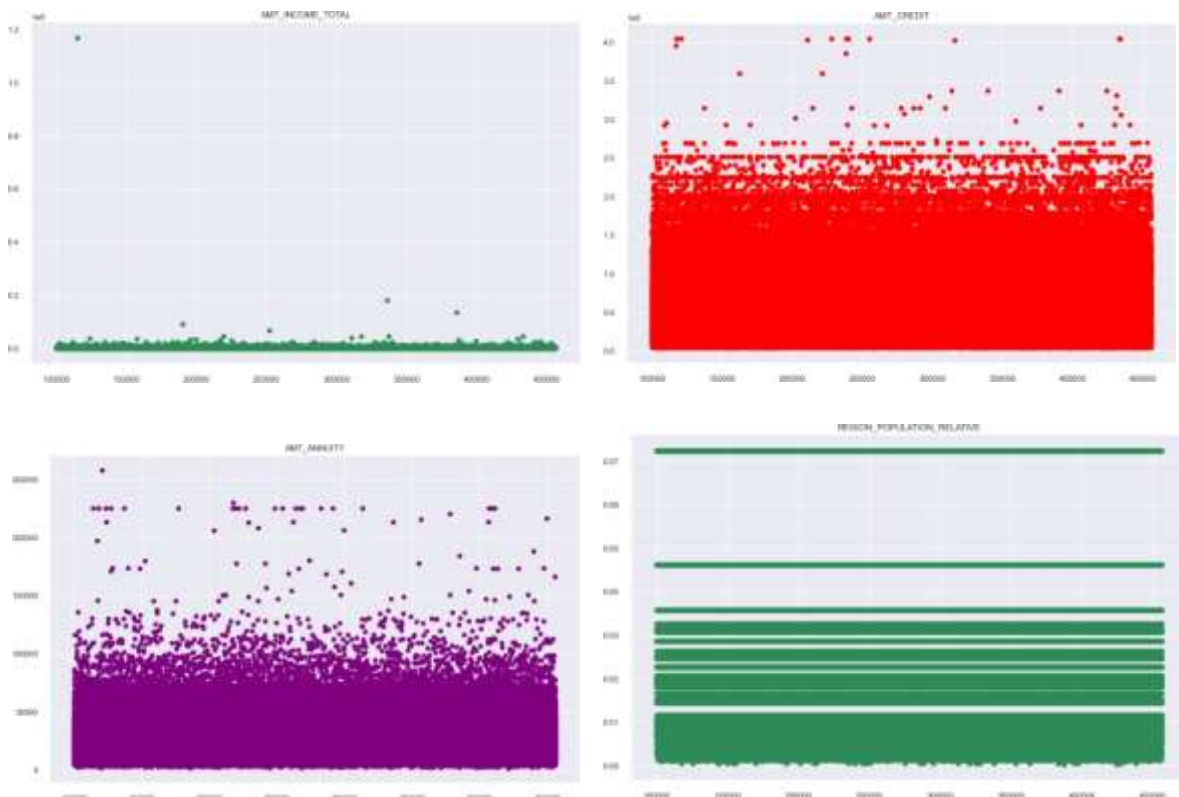
7. Imputation in Previous application.
8. Converting negative days to positive as days can't be negative.
9. Days in Columns 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH' are converted into positive values.
10. Imputing Null Values/Missing values in Previous Application-
 - a. After checking the null values percentage, we discovered columns like AMT_ANNUITY, AMT_GOODS_PRICE, AMT_DOWN_PAYMENT, and CNT_PAYMENT, have some null values/missing values.
 - b. For imputing null values, in column 'AMT_ANNUITY' nulls are filled using the **median** operation.
 - c. In 'AMT_GOODS_PRICE' and 'AMT_CREDIT', we have used **mode()**.
 - d. And imputation in 'CNT_PAYMENT' are done by filling null by '0'.

- **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.

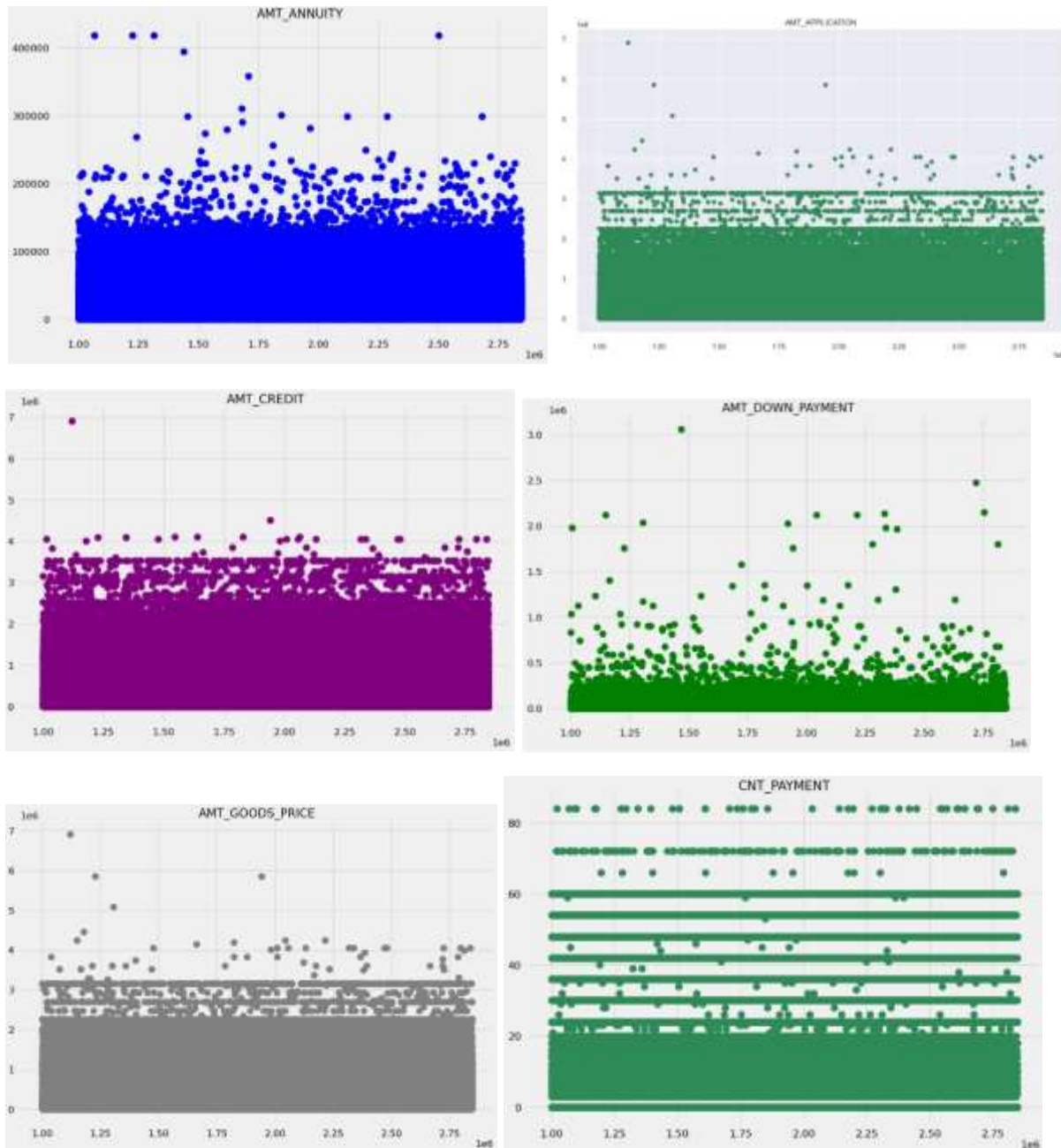
The outliers in this analysis work are identified using Scatter plots. A scatter plot is also called a scatter chart, scattergram, scatter plot, and XY graph. The scatter diagram graphs numerical data pairs, with one variable on each axis, showing their relationship.

❖ In Application_Data:



1. **AMT_ANNUITY** and **AMT_CREDIT** have some number of outliers.
2. **AMT_INCOME** has huge outliers.
3. In **REGION_POPULATION_RELATIVE**, we found no client have 0.06 region_population_relative.
4. **DAYS_EMPLOYED** have inappropriate values.

In Previous_Application:

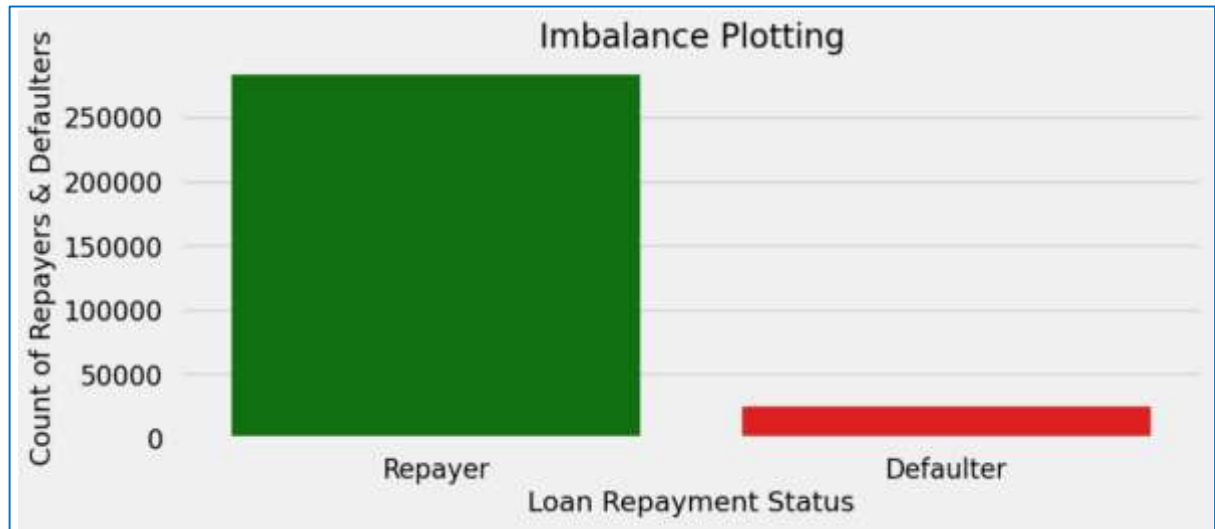


1. **AMT_ANNUITY**, **AMT_APPLICATION**, **AMT_CREDIT**, **AMT_DOWN_PAYMENT**, and **AMT_GOODS_PRICE** have big number of outliers.

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

The result of data imbalance shows-

- ✓ Ratio of data imbalance in relative with respect to Repayer and Defaulter data is **11.39 : 1**



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

❖ **For target variable 0 (other cases/defaulters)**

1. Most of the applicants are female but the twist is a smaller number of loans are taken by males but in case of not repaying the loan (defaulters), the number of males is in large numbers.
2. About **188278** females were found to be **defaulters** whereas males were **94404** in number.
3. Females have overtaken males in terms of not repaying the loan.
4. Most Females have 'Working' as their source of income, as compared to males, females are in larger numbers in terms of not repaying the loan.
5. Meanwhile, Commercial Associates (both female and male) have a great contribution to defaulters.
6. Maternity leave people have the least defaulter count.
7. Here, we can see most of the defaulters have income between 1Lakh-2Lakh whereas approx.
8. Rich clients (income above 1M) are less in the defaulters' line.
9. 20% of defaulters have income less than a Lakh.
10. More than 16% of people used loan money above 10 lakh as for their advantage.

11. Meanwhile, loans ranging from 2Lakh-3Lakh are taken with the intention of not repaying where females are in large numbers.
12. Hence as compared to revolving loans (9.21%), cash loans (90.79%) are taken in bulk.
13. Almost clients applied for cash loans later they turned into defaulters.
14. Clients having no car have more defaulter rate (69%).
15. In case of not returning loans, clients have a house or flat in large numbers (65%).
16. Business Entity type 3 people have a big bar in defaulter cases followed by Self-Employed people and others.

❖ **For target variable 1 (with payment difficulties/Repayers)**

17. 57% of females are in the repayers counts which is less than females count in defaulters.
18. Here students and businessmen are missing whereas people with 'working' as the source of income is leading.
19. In terms of repaying loans, the percentage of people got increased from 50% in defaulter rate to 61% in repayment.
20. People having income 2lakh-3lakh are in more number in case of not repaying loan where here in terms of returning loan they are in fewer numbers as compared to defaulters chart.
21. Vice versa for the people having income less than or equal to 1Lakh they are in more number for returning loans as compared to defaulter status.
22. Revolving loans returned by males are very less in numbers. Case loans in repayment status and defaulting status in quite the same.
23. Females who don't have a car have a big chance of repaying the loan and vice versa.
24. Males owning no flat or house have less chance of returning the loan.
25. Loan applicants not having a car are in better numbers in repayment status (69%).
26. Loan applicants owning a house or flat are also in good numbers (68%).

Result of bivariate analysis:

27. Clients having academic degrees who have the most income and have more counts in defaulters than repayers.
28. Clients educated from lower secondary are an equal number of counts in defaulters and repayers.
29. Working people having 2 family members are an equal number of repayers and defaulters.
30. People who get income through Maternity Leave tend to be more Defaulter when they have more Family Members.
31. Married people are high in defaulter counts as they have more children this may be a reason for default.

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

Top 10 correlations for the client with payment difficulties (Repayer) –

1. AMT_GOODS_PRICE and AMT_CREDIT are highly correlated (0.98).
2. Also, AMT_GOODS_PRICE is correlated with AMT_ANNUITY at 0.75
3. CNT_FAM_MEMBERS and CNT_CHILDREN are also correlated at 0.88
4. LIVE_REGION_NOT_WORK_REGION and REG_REGION_NOT_WORK_REGION comes in pictures in term of correlation (0.84).

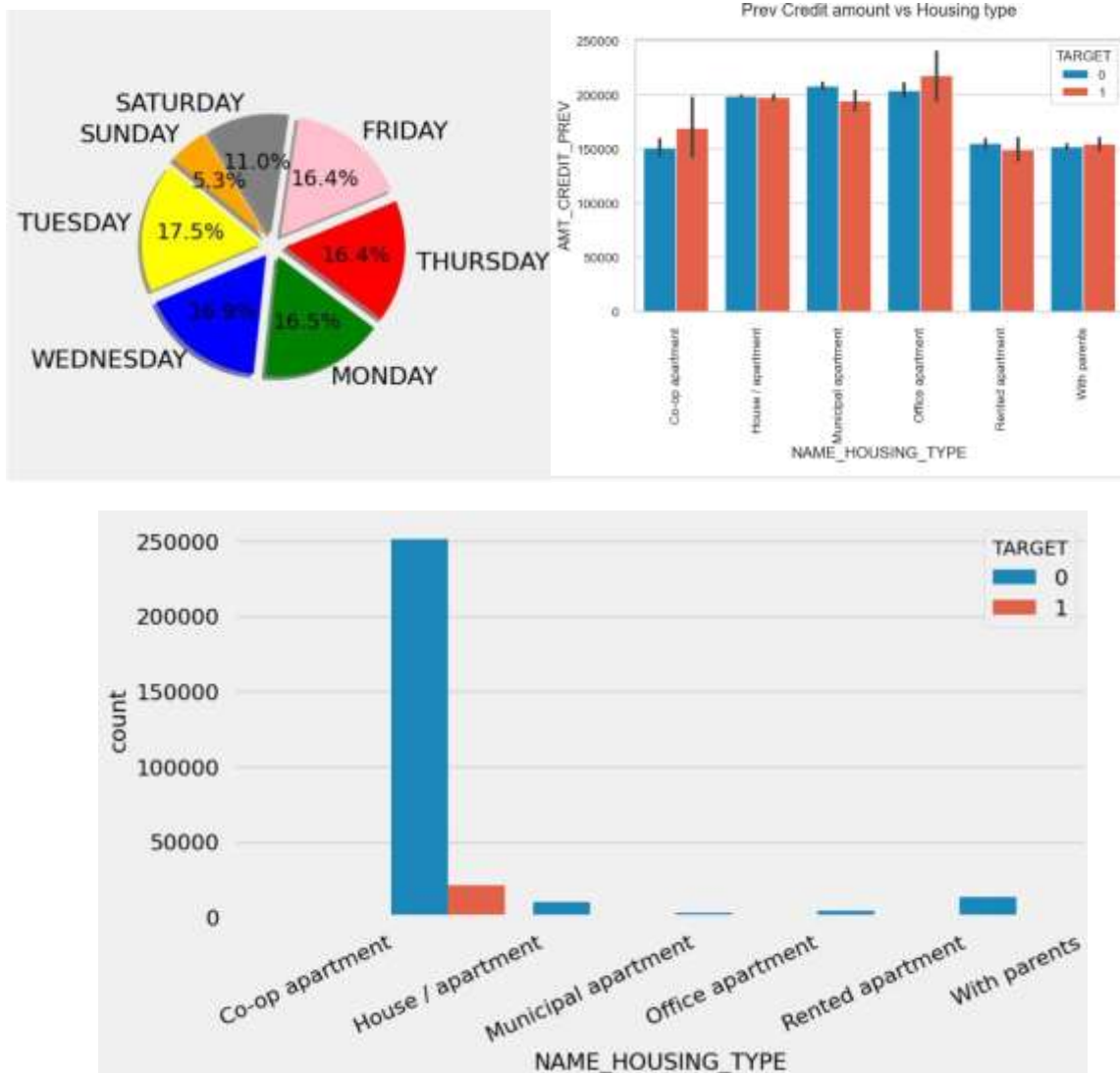
	VAR1	VAR2	Correlation
124	AMT_GOODS_PRICE	AMT_CREDIT	0.982783
242	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
335	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.847885
398	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.778540
125	AMT_GOODS_PRICE	AMT_ANNUITY	0.752295
104	AMT_ANNUITY	AMT_CREDIT	0.752195
188	DAYS_EMPLOYED	DAYS_BIRTH	0.575097
314	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.497937
377	REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.472052
354	REG_CITY_NOT_LIVE_CITY	REG_REGION_NOT_LIVE_REGION	0.322628

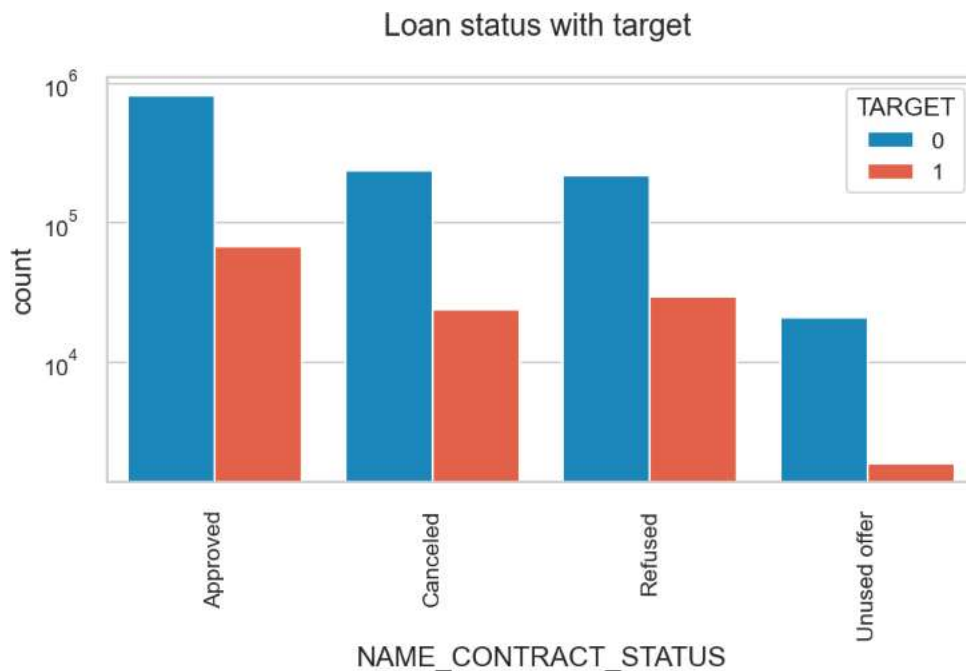
Top 10 correlations for the client with other cases (Defaulter) –

1. The correlation of AMT_GOODS_PRICE with AMT_CREDIT AND AMT_ANNUITY is quite same as in Repayment(above).
2. Whereas it's a bit increase in correlation of CNT_FAM_MEMBERS and CNT_CHILDREN (0.88) in defaulters as compared to repayment status (0.87).
3. Slightly decrease in correlation between AMT_GOODS_PRICE and AMT_ANNUITY (0.75) in the default case in comparison with repayment (0.77).
4. Same as in the case of AMT_ANNUITY and AMT_CREDIT, the correlation reduced from 0.77 to 0.75 in the defaulter case.
5. The correlation of DAYS_EMPLOYED and DAYS_BIRTH seems to be increased from 0.57 in repayment to 0.61 in defaulter.
6. Almost top 6 correlation are same but with slightly different values
7. In defaulter case, AMT_ANNUITY and AMT_INCOME_TOTAL at 0.41 come in top_10 correlation.

	VAR1	VAR2	Correlation
124	AMT_GOODS_PRICE	AMT_CREDIT	0.987022
242	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
335	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861
398	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.830381
125	AMT_GOODS_PRICE	AMT_ANNUITY	0.776421
104	AMT_ANNUITY	AMT_CREDIT	0.771297
188	DAYS_EMPLOYED	DAYS_BIRTH	0.618048
314	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.446101
377	REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.435514
103	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418948

- Include visualizations and summarize the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables.





Visualization and summary of insights are shown here – [BANK LOAN CASE STUDY](#)

Insights:

1. Females were found to be more defaulters as compared to males.
2. Out of 100 only 16% of clients have applied on weekend days(Saturday & Sunday) for current loans meanwhile on TUESDAY most of the clients about 17% have shown interest in applying.
3. Clients living in apartments and houses have the most defaulters and it's not safe side to approve such clients' loan applications.
4. 18% of clients have applied for the very first time whereas almost 73% of clients have applied for loans again.
5. After XAP, the common reason behind the rejection of a loan is HC.
6. In Previous applications, about 41% of clients have applied for POS followed by Cash with 22%.
7. Other than XAP and XNA, for repairing purposes most of the defaulters have taken advantage of the loan amount
8. Buying a used car is also a major reason for applying for a loan after urgent needs.
9. Company should think before approving loan applications for such reasons.
10. People living in office apartments is having higher credit for defaulting and people in co-op apartments have repaid the loan despite taking huge amounts as a loan.
11. Municipal apartments also have huge bars in not repaying the loan.

12. Defaulters use the unused offer for their benefit whereas only more than the half-approved loans are repaid.
13. New clients have returned their loan payments but there are still defaulters more than repayers.
14. Whereas clients applying for loans again has most counts in terms of not paying loan status.
15. Company should pay attention on client whose previous application was for POS, Cash, etc because they are also contributing to the defaulters counts.

Result:

- ✓ From the project we get to know how a company should manage risk during giving loans to clients.
- ✓ Understood visualizing techniques using python libraries (pandas, matplotlib, seaborn, etc.) and the senseful data from the graphs and charts.
- ✓ Learned how to present valuable insights and driving factors from the huge dataset.
- ✓ Learned about correlations of important variables and the idea of presenting them.
- ✓ Removal of null values, imputation of null values, data imbalance, outliers, univariate, bivariate analysis, etc. have also been studied.
- ✓ Helped me to use EDA (Exploratory Data Analysis) in real business case scenes.

Jupyter Notebook (Python file Link) – [Bank loan case study](#)