Bank Loan Case Study

Summary:

In this project, I will be a data analyst at a finance company that offers various types of loans to urban customers. The company faces two major risks: losing business if it denies loans to customers capable of repayment, and incurring financial losses if it approves loans for customers who cannot repay. The primary task is to analyse a dataset using **Exploratory Data Analysis (EDA)** to uncover patterns that help identify whether a customer is likely to default on their loan.

The dataset includes customer applications that have led to one of four outcomes: **Approved, Cancelled, Refused**, or **Unused Offer**. Customers are categorized into two main groups based on their payment history:

- Customers who have had difficulty with payments (late by more than X days in at least one of the first Y instalments).
- Customers who paid their instalments on time.

The goal is to leverage EDA to discover how various customer and loan attributes affect the likelihood of default. This will help the company make informed decisions about loan approvals, such as denying loans, reducing loan amounts, or increasing interest rates for risky customers.

Approach:

1. **Data Understanding**:
    - I Began by understanding the dataset, which likely contains both **customer attributes** (age, income, credit score, etc.) and **loan attributes** (loan amount, term, interest rate, etc.).
    - Identify the different types of customers based on their payment behaviour.
2. **Data Cleaning**:
    - Checked for missing or incorrect data and applied appropriate methods like imputation or exclusion.
    - Normalize and transform data as required, especially for continuous variables (e.g., loan amount, income).
3. **Exploratory Data Analysis (EDA)**:
    - **Univariate Analysis**: I Started by analysing individual variables to get an overview of the distribution and detect any anomalies. For example, understand the distribution of credit scores, loan amounts, etc.
    - **Bivariate and Multivariate Analysis**: Examined how two or more variables interact, such as how loan default rates change with income levels, credit scores, or loan amounts.
    - Use visualizations like **histograms, scatter plots, box plots, and correlation matrices** to detect relationships and patterns in the data.
4. **Identifying Risk Factors**:
    - Identified attributes that are closely associated with loan defaults, such as low credit scores, high loan amounts, or low-income levels.
    - Investigate if certain types of loans (e.g., higher interest rates or longer terms) lead to a higher probability of default.
5. **Segmentation**:
    - Segment the customers into groups based on their likelihood of default, which could help develop targeted loan approval policies. For instance, high-risk applicants might get loans with stricter conditions or higher interest rates.
6. **Conclusion and Recommendations**:

Bank Loan Case Study

   - Based on the insights, provide actionable recommendations. Suggest changes in approval policies to mitigate the risks, such as rejecting high-risk applicants, offering smaller loans, or adjusting interest rates to reflect the level of risk.
7. **Research on Risk Analytics**:
   - Before diving into the project, it's crucial to understand the basics of **risk analytics in banking**. Focus on variables like **credit history, debt-to-income ratios, payment history**, and their significance in determining loan risk.

Key Business Objective:

The ultimate goal is to help the company make better decisions regarding loan approvals by identifying factors that predict whether a customer is likely to default, thereby minimizing financial losses and missed business opportunities.

This approach ensures that capable applicants are not rejected, and risky applicants are either rejected or subjected to more stringent loan conditions.

INSIGHTS:

A. Missing Data:

- Identifying missing data helps maintain data integrity. Missing values can be addressed by imputation (e.g., using the average for numerical data) or removal, depending on the extent of missingness. Visualizing the proportion of missing data across variables highlights key areas requiring attention.

B. Outliers:

- Outliers can distort analysis and lead to misleading conclusions. Detecting them allows the company to either adjust or remove extreme values, ensuring that results reflect the general population's behaviour accurately.

C. Data Imbalance:

- If the dataset is imbalanced (e.g., more non-defaulters than defaulters), predictive models could become biased. Recognizing imbalance enables the use of techniques to mitigate this, improving the reliability of decision-making.

D. Univariate, Segmented Univariate, and Bivariate Analysis:

- **Univariate analysis** reveals the distribution of individual variables, such as income or loan amounts.
- **Segmented univariate analysis** provides deeper insights by comparing variable distributions for different customer groups (e.g., defaulters vs. non-defaulters).
- **Bivariate analysis** highlights relationships between variables, such as income levels and default risk, allowing for more informed decision-making.

E. Top Correlations:

Bank Loan Case Study

- Identifying top correlations between variables (like income, loan amount) and loan default for different customer groups provides actionable insights, helping to predict default risk and inform loan approval policies.

  TASK1- To Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

  Approach: Excel functions like COUNT, ISBLANK, and IF were used to identify missing data. Functions like AVERAGE or MEDIAN for imputation or other appropriate methods available in Excel.

  Outcomes-

- With the use of a count functions and other arithmetic formulas, I calculated the percentage of blank values in each column.
- After that, I dropped all the columns were blank percentage were more than 40 percent, I filled the null values of the remaining columns with median or mode depending upon the data type.
- For null cells with the data of the numeric type I utilized the median values, and for data of the categorical type I preferred its mode.
- I changed the data set from days to years.

  Fallowing are the results and visualizations from the excel-

Column1	COUNT BLANK	BLANK %
SK_ID_CURR	0	0
TARGET	0	0
NAME_CONTRACT_TYPE	0	0
CODE_GENDER	0	0
FLAG_OWN_CAR	0	0
FLAG_OWN_REALTY	0	0
CNT_CHILDREN	0	0
AMT_INCOME_TOTAL	0	0
AMT_CREDIT	0	0
AMT_ANNUITY	1	0.00200004
AMT_GOODS_PRICE	38	0.07600152
NAME_TYPE_SUITE	192	0.38400768
NAME_INCOME_TYPE	0	0
NAME_EDUCATION_TYPE	0	0
NAME_FAMILY_STATUS	0	0
NAME_HOUSING_TYPE	0	0
REGION_POPULATION_RELATIVE	0	0
DAYS_BIRTH	0	0
DAYS_EMPLOYED	0	0
DAYS_REGISTRATION	0	0
DAYS_ID_PUBLISH	0	0

Bank Loan Case Study

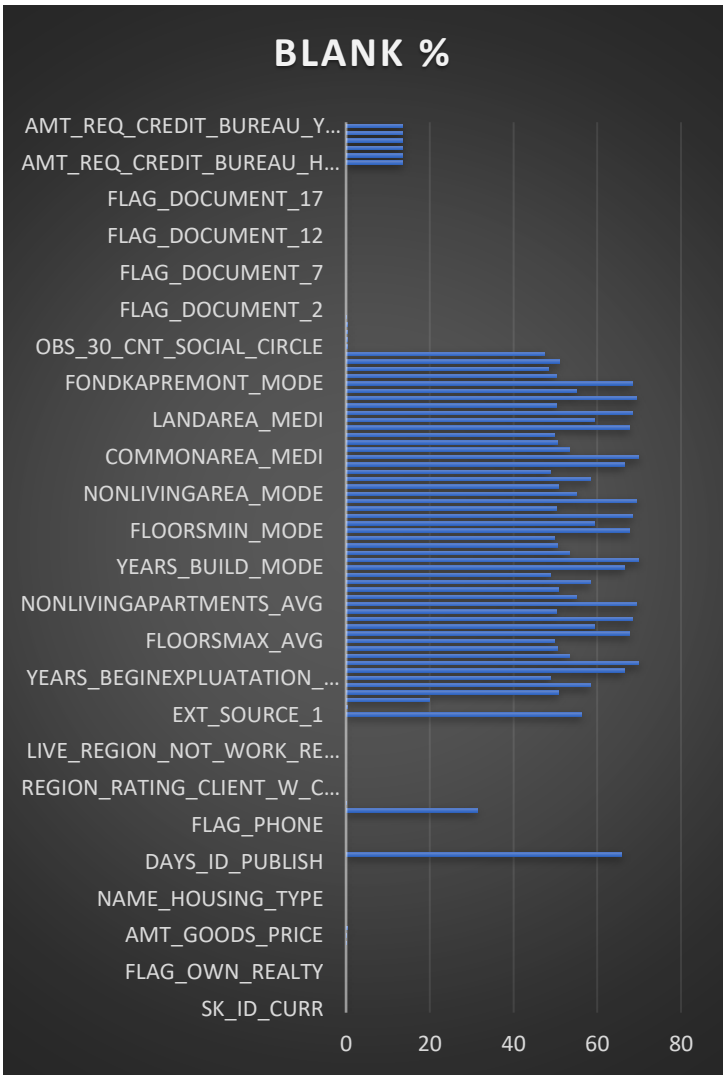
OWN_CAR_AGE	32950	65.90131803
FLAG_EMP_PHONE	0	0
FLAG_WORK_PHONE	0	0
FLAG_CONT_MOBILE	0	0
FLAG_PHONE	0	0
FLAG_EMAIL	0	0
OCCUPATION_TYPE	15654	31.30862617
CNT_FAM_MEMBERS	1	0.00200004
REGION_RATING_CLIENT	0	0
REGION_RATING_CLIENT_W_CITY	0	0
WEEKDAY_APPR_PROCESS_START	0	0
HOUR_APPR_PROCESS_START	0	0
REG_REGION_NOT_LIVE_REGION	0	0
REG_REGION_NOT_WORK_REGION	0	0
LIVE_REGION_NOT_WORK_REGION	0	0
REG_CITY_NOT_LIVE_CITY	0	0
REG_CITY_NOT_WORK_CITY	0	0
LIVE_CITY_NOT_WORK_CITY	0	0
ORGANIZATION_TYPE	0	0
EXT_SOURCE_1	28172	56.3451269
EXT_SOURCE_2	126	0.25200504
EXT_SOURCE_3	9944	19.88839777
APARTMENTS_AVG	25385	50.77101542
BASEMENTAREA_AVG	29199	58.39916798
YEARS_BEGINEXPLUATATION_AVG	24394	48.78897578
YEARS_BUILD_AVG	33239	66.47932959
COMMONAREA_AVG	34960	69.92139843
ELEVATORS_AVG	26651	53.30306606
ENTRANCES_AVG	25195	50.39100782
FLOORSMAX_AVG	24875	49.75099502
FLOORSMIN_AVG	33894	67.78935579
LANDAREA_AVG	29721	59.44318886
LIVINGAPARTMENTS_AVG	34226	68.45336907
LIVINGAREA_AVG	25137	50.2750055
NONLIVINGAPARTMENTS_AVG	34714	69.42938859
NONLIVINGAREA_AVG	27572	55.1451029
APARTMENTS_MODE	25385	50.77101542
BASEMENTAREA_MODE	29199	58.39916798
YEARS_BEGINEXPLUATATION_MODE	24394	48.78897578
YEARS_BUILD_MODE	33239	66.47932959
COMMONAREA_MODE	34960	69.92139843
ELEVATORS_MODE	26651	53.30306606
ENTRANCES_MODE	25195	50.39100782
FLOORSMAX_MODE	24875	49.75099502
FLOORSMIN_MODE	33894	67.78935579
LANDAREA_MODE	29721	59.44318886

Bank Loan Case Study

LIVINGAPARTMENTS_MODE	34226	68.45336907
LIVINGAREA_MODE	25137	50.2750055
NONLIVINGAPARTMENTS_MODE	34714	69.42938859
NONLIVINGAREA_MODE	27572	55.1451029
APARTMENTS_MEDI	25385	50.77101542
BASEMENTAREA_MEDI	29199	58.39916798
YEARS_BEGINEXPLUATATION_MEDI	24394	48.78897578
YEARS_BUILD_MEDI	33239	66.47932959
COMMONAREA_MEDI	34960	69.92139843
ELEVATORS_MEDI	26651	53.30306606
ENTRANCES_MEDI	25195	50.39100782
FLOORSMAX_MEDI	24875	49.75099502
FLOORSMIN_MEDI	33894	67.78935579
LANDAREA_MEDI	29721	59.44318886
LIVINGAPARTMENTS_MEDI	34226	68.45336907
LIVINGAREA_MEDI	25137	50.2750055
NONLIVINGAPARTMENTS_MEDI	34714	69.42938859
NONLIVINGAREA_MEDI	27572	55.1451029
FONDKAPREMONT_MODE	34191	68.38336767
HOUSETYPE_MODE	25075	50.15100302
TOTALAREA_MODE	24148	48.29696594
WALLSMATERIAL_MODE	25459	50.91901838
EMERGENCYSTATE_MODE	23698	47.39694794
OBS_30_CNT_SOCIAL_CIRCLE	168	0.33600672
DEF_30_CNT_SOCIAL_CIRCLE	168	0.33600672
OBS_60_CNT_SOCIAL_CIRCLE	168	0.33600672
DEF_60_CNT_SOCIAL_CIRCLE	168	0.33600672
DAYS_LAST_PHONE_CHANGE	1	0.00200004
FLAG_DOCUMENT_2	0	0
FLAG_DOCUMENT_3	0	0
FLAG_DOCUMENT_4	0	0
FLAG_DOCUMENT_5	0	0
FLAG_DOCUMENT_6	0	0
FLAG_DOCUMENT_7	0	0
FLAG_DOCUMENT_8	0	0
FLAG_DOCUMENT_9	0	0
FLAG_DOCUMENT_10	0	0
FLAG_DOCUMENT_11	0	0
FLAG_DOCUMENT_12	0	0
FLAG_DOCUMENT_13	0	0
FLAG_DOCUMENT_14	0	0
FLAG_DOCUMENT_15	0	0
FLAG_DOCUMENT_16	0	0
FLAG_DOCUMENT_17	0	0
FLAG_DOCUMENT_18	0	0
FLAG_DOCUMENT_19	0	0

Bank Loan Case Study

FLAG_DOCUMENT_20	0	0
FLAG_DOCUMENT_21	0	0
AMT_REQ_CREDIT_BUREAU_HOUR	6734	13.46826937
AMT_REQ_CREDIT_BUREAU_DAY	6734	13.46826937
AMT_REQ_CREDIT_BUREAU_WEEK	6734	13.46826937
AMT_REQ_CREDIT_BUREAU_MON	6734	13.46826937
AMT_REQ_CREDIT_BUREAU_QRT	6734	13.46826937
AMT_REQ_CREDIT_BUREAU_YEAR	6734	13.46826937



TASK2- To Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables**.**

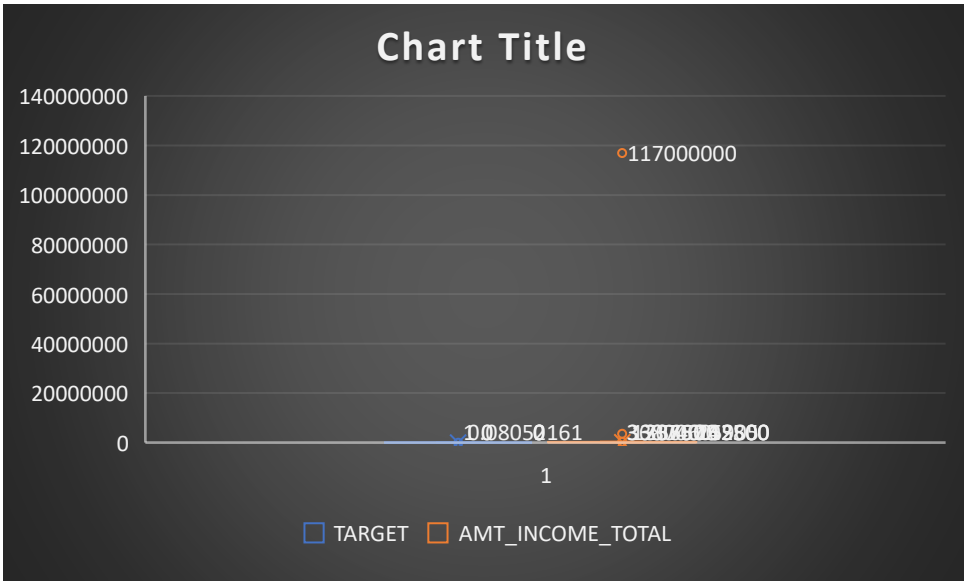
APPROACH-Functions like quartile, statistical and conditional formatting were used to identify the potential outliers.

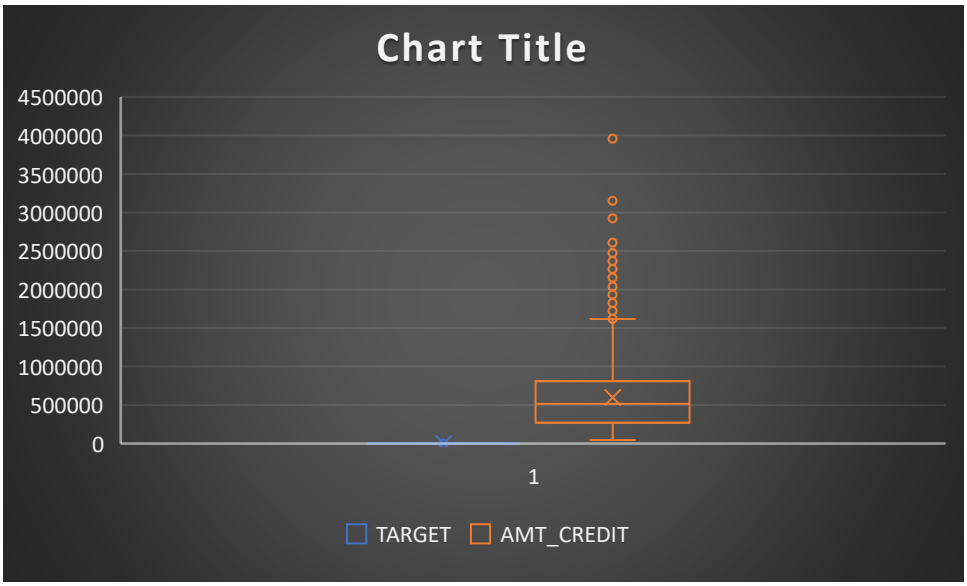
Outcomes-

Bank Loan Case Study

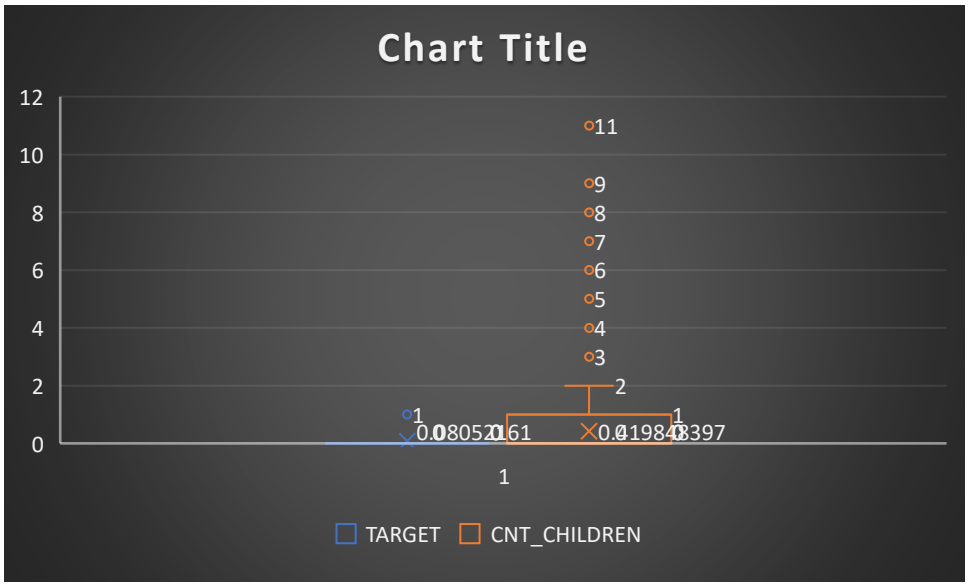
- Amount_ income_ total contains an outlier that I discovered ,one of the outlier has an extraordinary salary of 117,000,000.
- In Cnt _Children some of the outliers propose 11 children's which does not seem feasible in today's world
- Days_ employed_ yrs I discovered some outliers were I discovered people working for longer than 1000 years.
- However, in days_ birth and other columns I didn't find potential outliers.

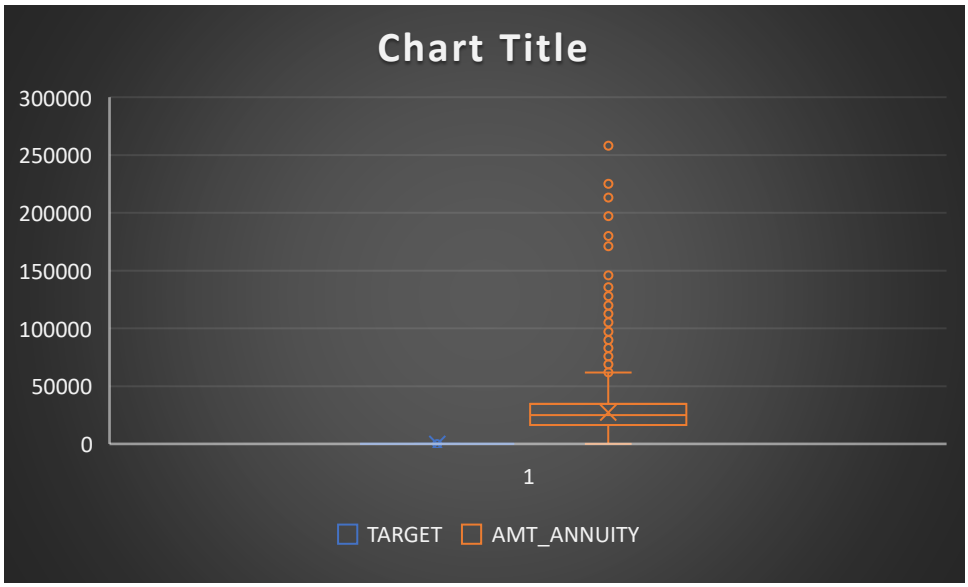
Following are the visualization table from excel.





Bank Loan Case Study

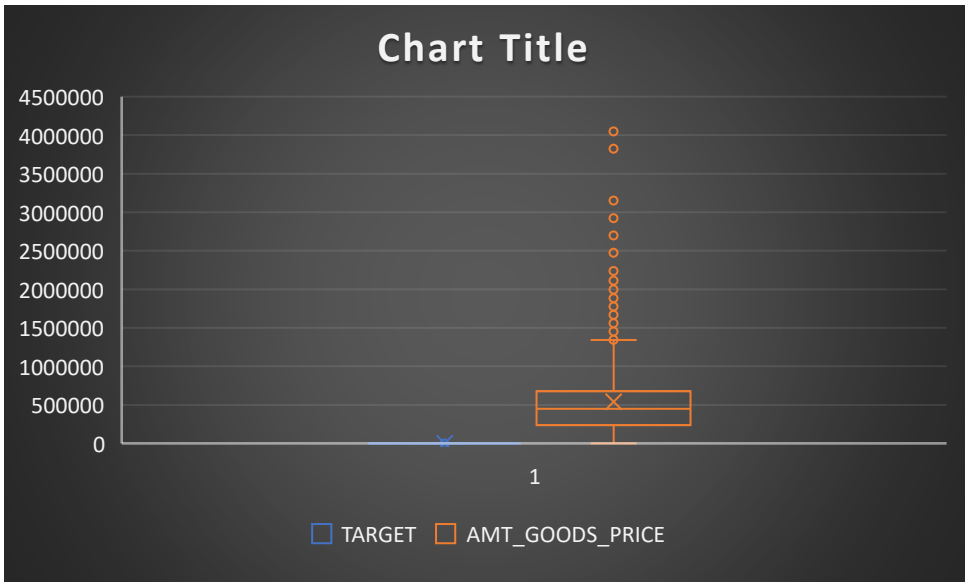


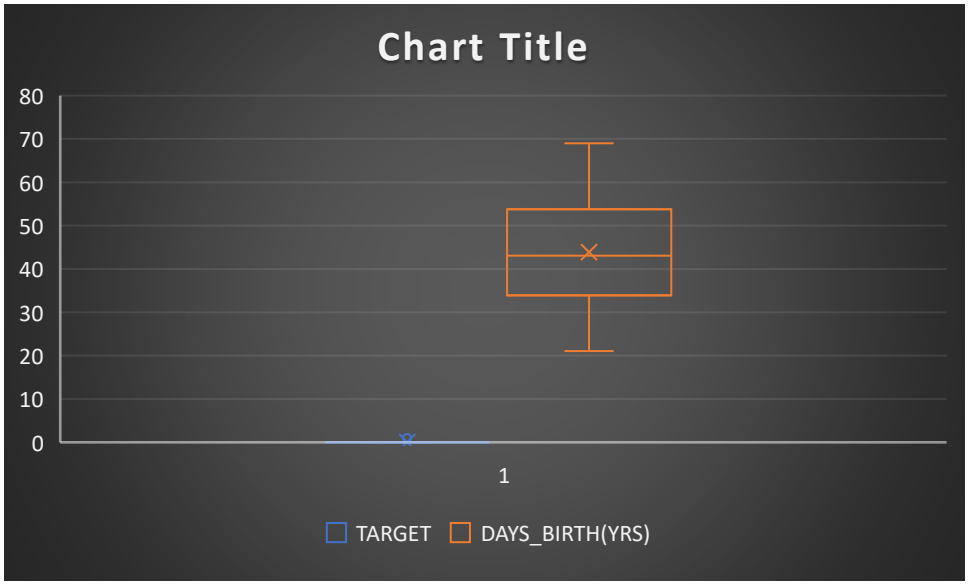


TASK3-Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

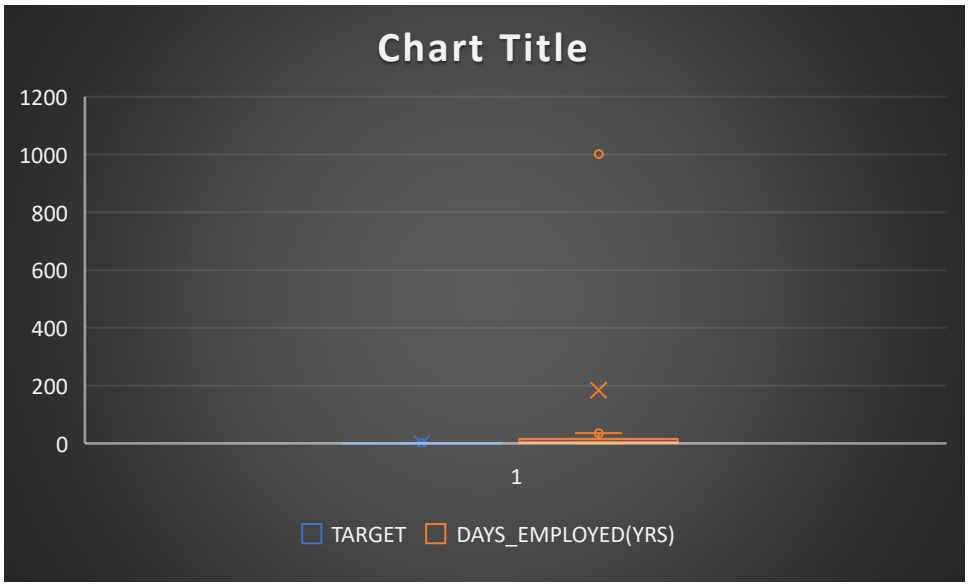
APPROACH- Utilized Excel functions like COUNTIF and SUM to calculate the proportions of each class.

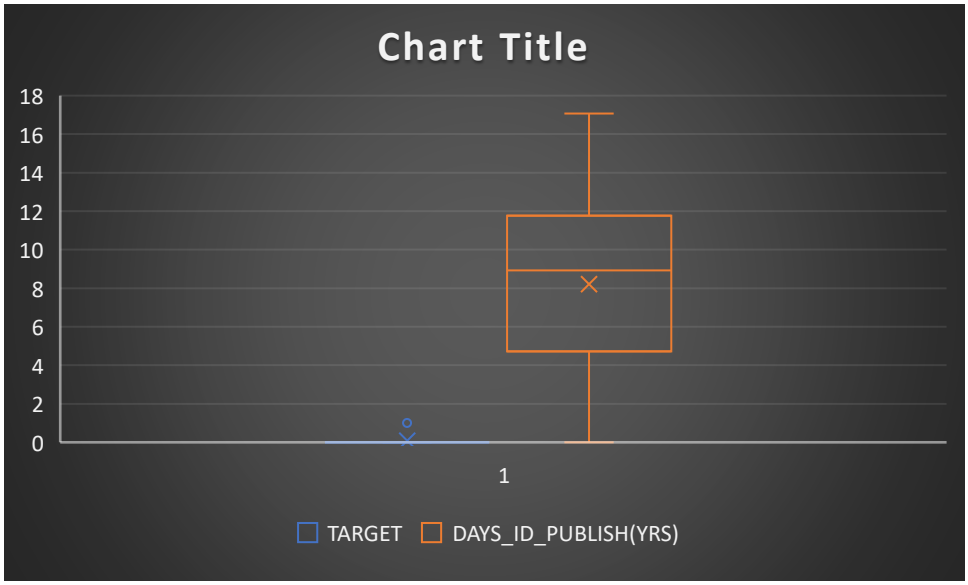
Bank Loan Case Study



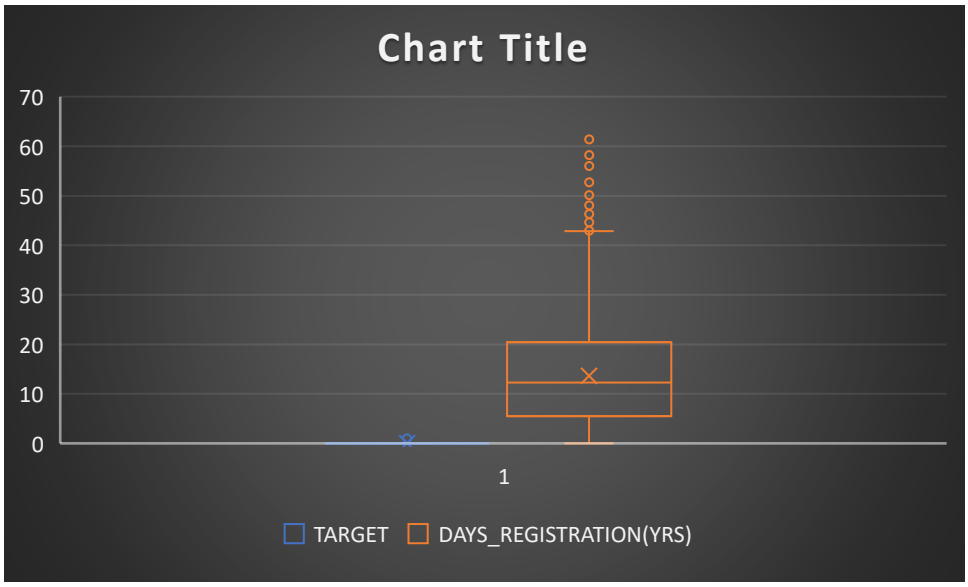


Bank Loan Case Study





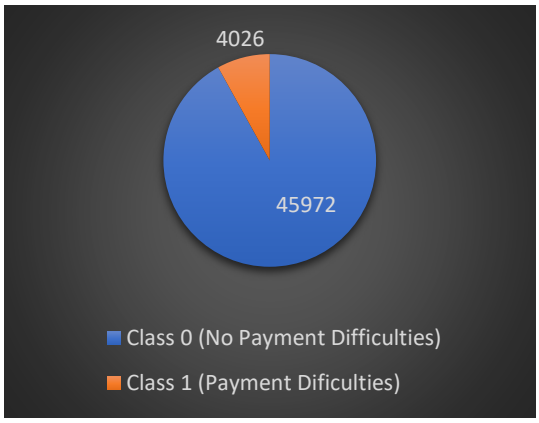
Bank Loan Case Study



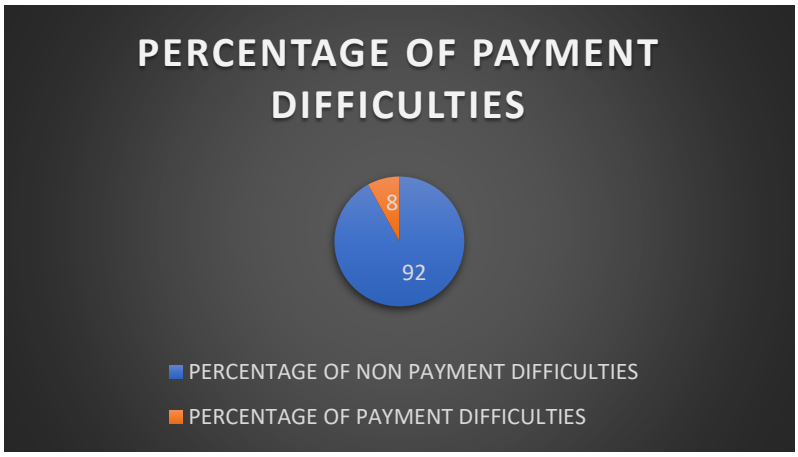
TASK3- Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

APPROACH- Utilized Excel functions like COUNTIF and SUM to calculate the proportions of each class.

OUTCOMES-
Fallowing is the graph showing the number of people with no payment difficulties and number of people with payment difficulties.
Here class 0 represents the people with no payment difficulties
Class1 represents people with payment difficulties.



Bank Loan Case Study



TASK4-Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

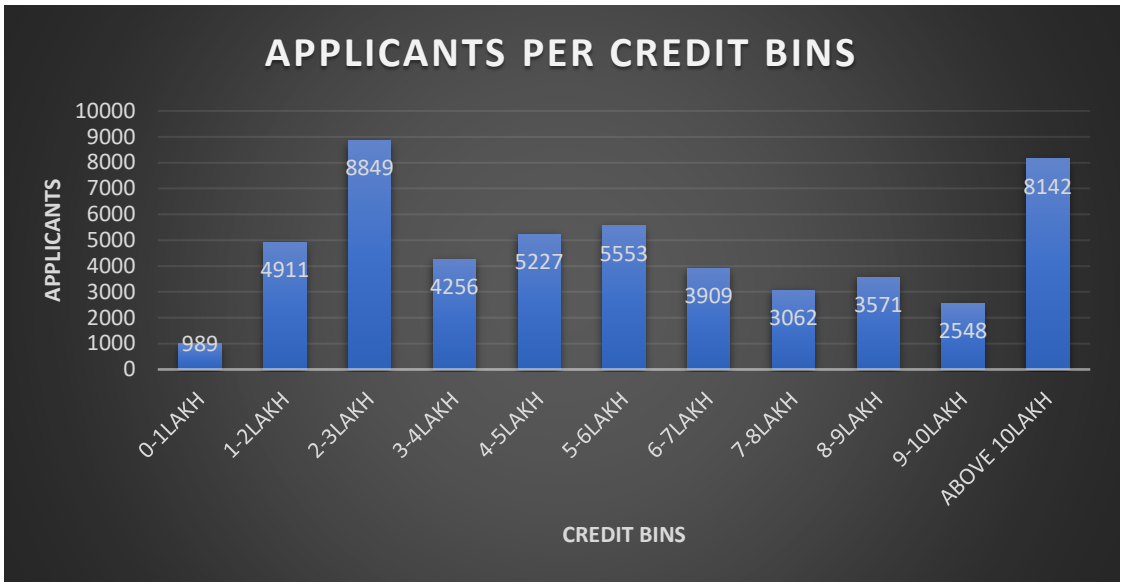
APPROACH-Utilized Excel functions like COUNT, AVERAGE, MEDIAN, and statistical functions for descriptive analysis. Utilized Excel features like filters, sorting, and pivot tables for segmented and bivariate analysis.

OUTCOMES- The fallowing are the results and visualizations from the excel.

CREDIT BINS	APPLICANTS
0-1LAKH	989
1-2LAKH	4911
2-3LAKH	8849
3-4LAKH	4256
4-5LAKH	5227
5-6LAKH	5553
6-7LAKH	3909
7-8LAKH	3062
8-9LAKH	3571
9-10LAKH	2548
ABOVE 10LAKH	8142

UNIVARIATE ANALYSIS

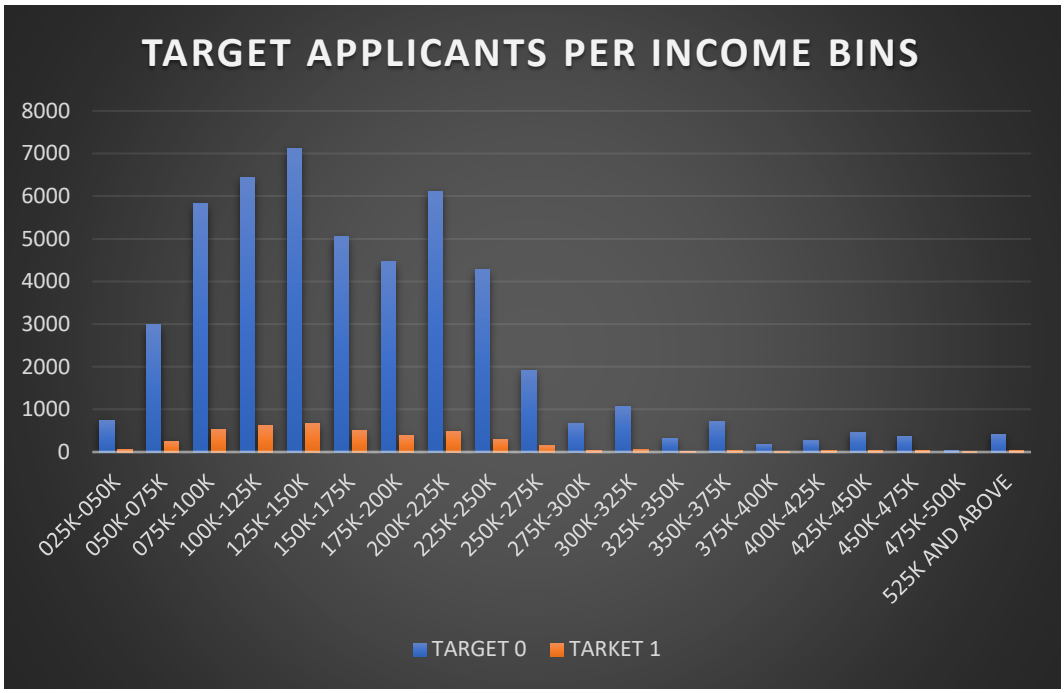
Bank Loan Case Study



INCOME BINS	TARGET 0	TARKET 1
025K-050K	741	63
050K-075K	2980	246
075K-100K	5826	536
100K-125K	6428	620
125K-150K	7126	678
150K-175K	5060	501
175K-200K	4458	389
200K-225K	6121	491
225K-250K	4279	304
250K-275K	1919	143
275K-300K	681	45
300K-325K	1076	59
325K-350K	322	24
350K-375K	723	34
375K-400K	186	14
400K-425K	263	26
425K-450K	456	36
450K-475K	375	34
475K-500K	44	3
525K AND ABOVE	423	31

SEGMENTED UNIVARATE ANALYSIS

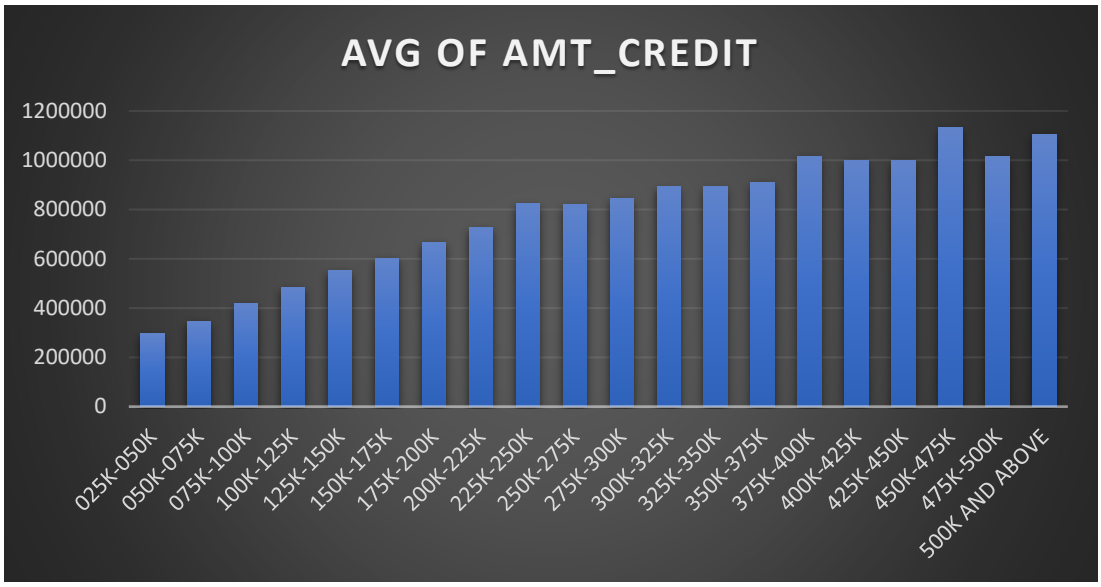
Bank Loan Case Study



INCOME BINS	AVG OF AMT_CREDIT
025K-050K	297752.0765
050K-075K	345240.3585
075K-100K	417267.8771
100K-125K	483568.8073
125K-150K	553042.1642
150K-175K	602034.4016
175K-200K	667004.421
200K-225K	727198.4449
225K-250K	822956.3582
250K-275K	820255.3451
275K-300K	842725.6488
300K-325K	892300.0718
325K-350K	892332.6503
350K-375K	910353.0482
375K-400K	1016914.375
400K-425K	999208.199
425K-450K	999153.6402
450K-475K	1132882.5
475K-500K	1015150.404
500K AND ABOVE	1105365.122

BIVARIATE ANALYSIS

Bank Loan Case Study



TASK5- Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

APPROACH- Utilize Excel functions like CORREL to calculate correlation coefficients between variables and the target variable within each segment.

OUTCOMES-

CNT_CHILDREN:

- Positively correlated with DAYS_BIRTH (YRS) (0.329) and REGION_RATING_CLIENT (0.0259), suggesting that more children may be weakly associated with older age and slightly higher regional ratings.
- Negatively correlated with DAYS_EMPLOYED (YRS) (-0.241), indicating that people with more children tend to have slightly fewer years of employment.

AMT_INCOME_TOTAL:

- Positively correlated with AMT_CREDIT (0.069), suggesting a weak positive relationship between income and credit amount.
- Weak negative correlation with DAYS_BIRTH (YRS) (-0.016), showing no strong relationship between income and age.

AMT_CREDIT:

- Positively correlated with REGION_POPULATION_RELATIVE (0.095), meaning higher credit amounts are slightly associated with higher relative population in a region.
- Negatively correlated with REGION_RATING_CLIENT (-0.100), implying that higher credit amounts may be given to clients in lower-rated regions.

REGION_POPULATION_RELATIVE:

- Positively correlated with AMT_CREDIT (0.095), as discussed.
- Negatively correlated with REGION_RATING_CLIENT (-0.125), indicating that clients from more populated regions might receive slightly lower ratings.

DAYS_BIRTH (YRS):

Bank Loan Case Study

- Negatively correlated with CNT_CHILDREN (-0.329), meaning older individuals tend to have fewer children.
- Weak negative correlations with DAYS_EMPLOYED (-0.067) and REGION_RATING_CLIENT (-0.103), showing minimal relationships between age and these variables.

DAYS_EMPLOYED (YRS):

- Strong negative correlation with REGION_RATING_CLIENT (-0.532), indicating that individuals with more years of employment tend to be from lower-rated regions.

REGION_RATING_CLIENT:

- Strongly negatively correlated with DAYS_EMPLOYED (YRS) (-0.613), as mentioned, suggesting that higher ratings are associated with fewer years of employment.
- Weak positive correlation with CNT_CHILDREN (0.0259), as discussed.

CORRELATION FOR APPLICANTS WITH PAYMENT MADE ON TIME

CNT_CHILDREN	1	0.009589	0.004972	-0.02556	0.329263	-0.23969	0.025913
AMT_INCOME_TOTAL	0.009589	1	0.069316	0.009589	0.009589	-0.03162	-0.03819
AMT_CREDIT	0.004972	0.069316	1	0.004972	0.004972	-0.07047	-0.10051
REGION_POPULATION_RELATIVE	-0.02556	0.029841	0.095111	-0.02556	-0.02556	-0.11045	-0.1258
DAYS_BIRTH(YRS)	-0.32926	-0.016	0.059343	-0.32926	-0.32926	-0.06783	-0.10368
DAYS_EMPLOYED(YRS)	-0.24154	-0.03151	-0.06774	-0.24154	-0.24154	-0.0041	-0.53267
REGION_RATING_CLIENT	0.025914	-0.03819	-0.10051	0.025914	0.025914	-0.61355	0.016779

CNT_CHILDREN	1	0.009589	0.004972	-0.02556	0.329263	-0.23969	0.025913
AMT_INCOME_TOTAL	0.009589	1	0.069316	0.009589	0.009589	-0.03162	-0.03819

Bank Loan Case Study

AMT_CREDIT	0.004972	0.069316	1	0.004972	0.004972	-0.07047	-0.10051
REGION_POPULATION_RELATIVE	-0.02556	0.029841	0.095111	-0.02556	-0.02556	-0.11045	-0.1258
DAYS_BIRTH(YRS)	-0.32926	-0.016	0.059343	-0.32926	-0.32926	-0.06783	-0.10368
DAYS_EMPLOYED(YRS)	-0.24154	-0.03151	-0.06774	-0.24154	-0.24154	-0.0041	-0.53267
REGION_RATING_CLIENT	0.025914	-0.03819	-0.10051	0.025914	0.025914	-0.61355	0.016779

CORRELATION FOR APPLICANTS WITH PAYMENT DIFFICULTIES