Description:

Imagine you're a data analyst at a finance company that specializes in lending various types of loans to urban customers. Your company faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans. Your task is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.

When a customer applies for a loan, your company faces two risks:

1. If the applicant can repay the loan but is not approved, the company loses business.
2. If the applicant cannot repay the loan and is approved, the company faces a financial loss.

The dataset you'll be working with contains information about loan applications. It includes two types of scenarios:

1. Customers with payment difficulties: These are customers who had a late payment of more than X days on at least one of the first Y installments of the loan.
2. All other cases: These are cases where the payment was made on time.

When a customer applies for a loan, there are four possible outcomes:

1. Approved: The company has approved the loan application.
2. Cancelled: The customer cancelled the application during the approval process.
3. Refused: The company rejected the loan.
4. Unused Offer: The loan was approved but the customer did not use it.

our goal in this project is to use EDA to understand how customer attributes and loan attributes influence the likelihood of default.

Data Analytics Tasks:

A. **Identify Missing Data and Deal with it Appropriately:** As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

- **Task:** Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.
- **Hint:** Utilize Excel functions like COUNT, ISBLANK, and IF to identify missing data. Consider using functions like AVERAGE or MEDIAN for imputation or other appropriate methods available in Excel.
- **Graph suggestion:** Create a bar chart or column chart to visualize the proportion of missing values for each variable.

B. Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

- **Task:** Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.
- **Hint:** Utilize Excel functions like QUARTILE, IQR, and conditional formatting to identify potential outliers. Consider applying thresholds or business rules to determine if the outliers are valid data points or require further investigation.
- **Graph suggestion:** Create box plots or scatter plots to visualize the distribution of numerical variables and highlight the outliers.

C. Analyze Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

- **Task:** Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.
- **Hint:** Utilize Excel functions like COUNTIF and SUM to calculate the proportions of each class. Compare the class frequencies to assess data imbalance.
- **Graph suggestion:** Create a pie chart or bar chart to visualize the distribution of the target variable and highlight the class imbalance.

D. Perform Univariate, Segmented Univariate, and Bivariate Analysis: To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

- **Task:** 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.
- **Hint:** Utilize Excel functions like COUNT, AVERAGE, MEDIAN, and statistical functions for descriptive analysis. Utilize Excel features like filters, sorting, and pivot tables for segmented and bivariate analysis.
- **Graph suggestion:** Create histograms, bar charts, or box plots to visualize the distributions of variables. Create stacked bar charts or grouped bar charts to compare variable distributions across different scenarios. Create scatter plots or heatmaps to visualize the relationships between variables and the target variable.

E. **Identify Top Correlations for Different Scenarios:** Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

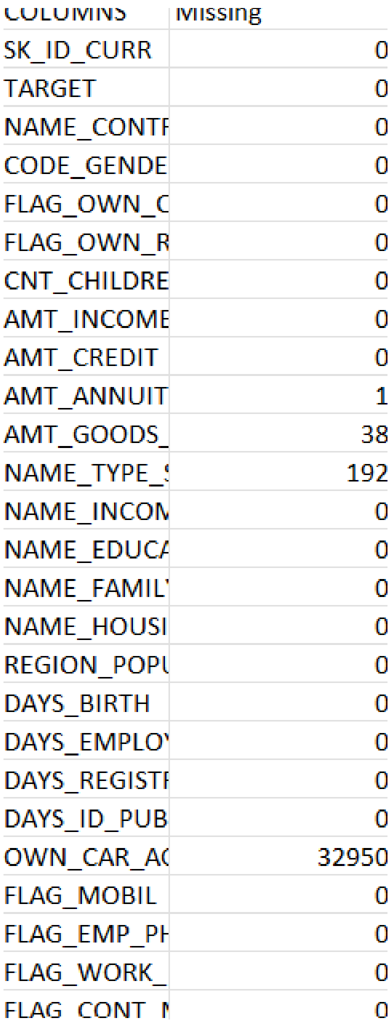
- **Task:** Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.
- **Hint:** Utilize Excel functions like CORREL to calculate correlation coefficients between variables and the target variable within each segment. Rank the correlations to identify the top indicators of loan default for each scenario.

- **Graph suggestion:** Create correlation matrices or heatmaps to visualize the correlations between variables within each segment. Highlight the top correlated variables for each scenario using different colors or shading.

APPROACH:

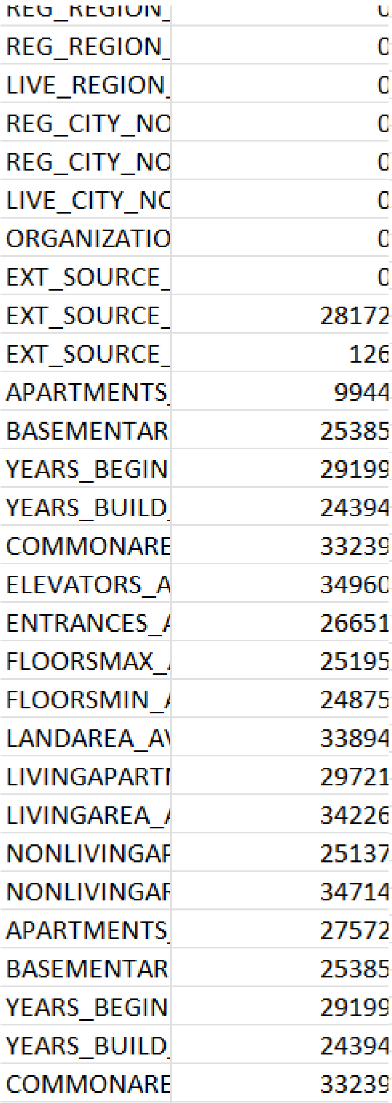
A. Identify Missing Data and Deal with it Appropriately: As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis**.**

Below are the missing values in the dataset:

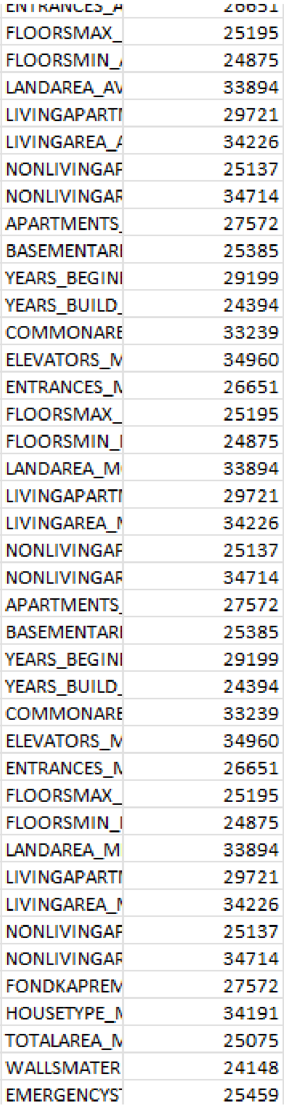


COLUMNS	Missing
SK_ID_CURR	0
TARGET	0
NAME_CONTI	0
CODE_GENDE	0
FLAG_OWN_C	0
FLAG_OWN_R	0
CNT_CHILDRE	0
AMT_INCOME	0
AMT_CREDIT	0
AMT_ANNUIT	1
AMT_GOODS_	38
NAME_TYPE_S	192
NAME_INCON	0
NAME_EDUCA	0
NAME_FAMIL'	0
NAME_HOUSI	0
REGION_POPU	0
DAYS_BIRTH	0
DAYS_EMPLO'	0
DAYS_REGISTI	0
DAYS_ID_PUB	0
OWN_CAR_A(	32950
FLAG_MOBIL	0
FLAG_EMP_PH	0
FLAG_WORK_	0
FLAG_CONT_I	0

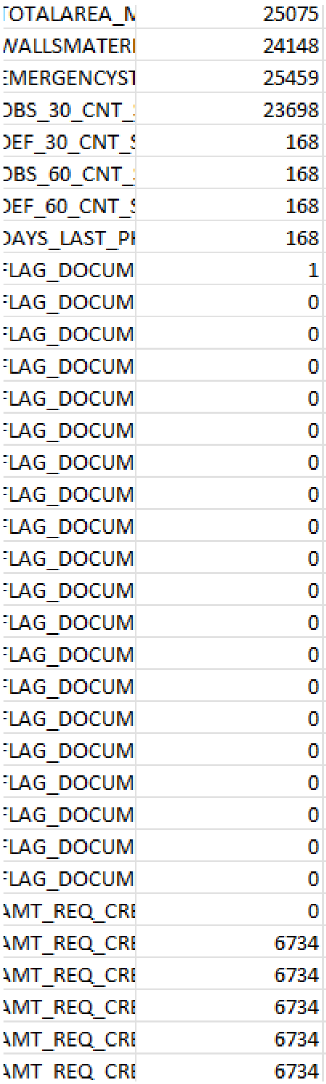
FLAG_CONT_I	0
FLAG_PHONE	0
FLAG_EMAIL	0
OCCUPATION_	15654
CNT_FAM_ME	1
REGION_RATI	0
REGION_RATI	0
WEEKDAY_API	0
HOUR_APPR_I	0
REG_REGION_	0
REG_REGION_	0
LIVE_REGION_	0
REG_CITY_NO	0
REG_CITY_NO	0
LIVE_CITY_NC	0
ORGANIZATIO	0
EXT_SOURCE_	0
EXT_SOURCE_	28172
EXT_SOURCE_	126
APARTMENTS_	9944
BASEMENTAR	25385
YEARS_BEGIN	29199
YEARS_BUILD_	24394
COMMONARE	33239
ELEVATORS_A	34960
ENTRANCES_/	26651
FLOORSMAX_/	25195
FLOORSMIN_/	24875
LANDAREA_AV	33894



REG_REGION_	0
REG_REGION_	0
LIVE_REGION_	0
REG_CITY_NO	0
REG_CITY_NO	0
LIVE_CITY_NC	0
ORGANIZATIO	0
EXT_SOURCE_	0
EXT_SOURCE_	28172
EXT_SOURCE_	126
APARTMENTS_	9944
BASEMENTAR	25385
YEARS_BEGIN	29199
YEARS_BUILD_	24394
COMMONARE	33239
ELEVATORS_A	34960
ENTRANCES_/	26651
FLOORSMAX_/	25195
FLOORSMIN_/	24875
LANDAREA_AV	33894
LIVINGAPARTI	29721
LIVINGAREA_/	34226
NONLIVINGAF	25137
NONLIVINGAF	34714
APARTMENTS_	27572
BASEMENTAR	25385
YEARS_BEGIN	29199
YEARS_BUILD_	24394
COMMONARE	33239

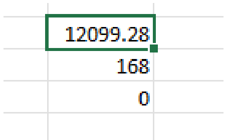


ENTRANCES_A	26651
FLOORSMAX_	25195
FLOORSMIN_/	24875
LANDAREA_AV	33894
LIVINGAPARTI	29721
LIVINGAREA_/	34226
NONLIVINGAF	25137
NONLIVINGAF	34714
APARTMENTS_	27572
BASEMENTARI	25385
YEARS_BEGINI	29199
YEARS_BUILD_	24394
COMMONARE	33239
ELEVATORS_M	34960
ENTRANCES_M	26651
FLOORSMAX_	25195
FLOORSMIN_I	24875
LANDAREA_M	33894
LIVINGAPARTI	29721
LIVINGAREA_I	34226
NONLIVINGAF	25137
NONLIVINGAF	34714
APARTMENTS_	27572
BASEMENTARI	25385
YEARS_BEGINI	29199
YEARS_BUILD_	24394
COMMONARE	33239
ELEVATORS_M	34960
ENTRANCES_M	26651
FLOORSMAX_	25195
FLOORSMIN_I	24875
LANDAREA_M	33894
LIVINGAPARTI	29721
LIVINGAREA_I	34226
NONLIVINGAF	25137
NONLIVINGAF	34714
FONDKAPREM	27572
HOUSETYPE_M	34191
TOTALAREA_M	25075
WALLSMATER	24148
EMERGENCYS	25459



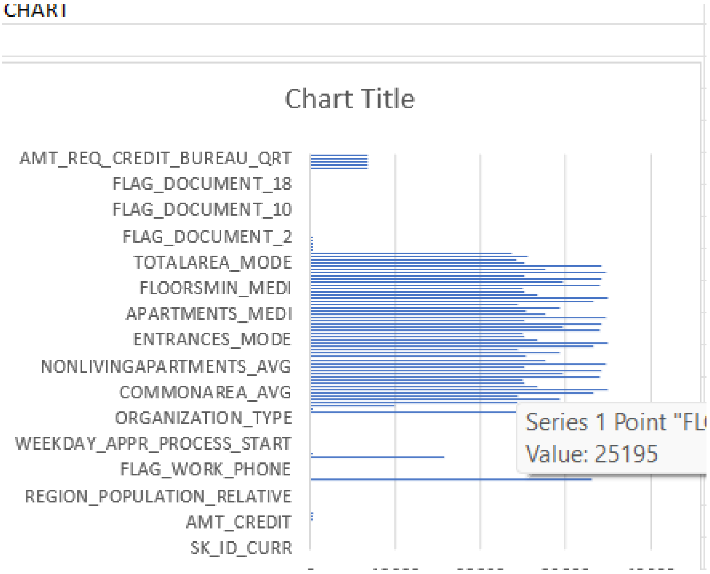
TOTALAREA_M	25075
WALLSMATERI	24148
EMERGENCYST	25459
OBS_30_CNT_	23698
DEF_30_CNT_S	168
OBS_60_CNT_	168
DEF_60_CNT_S	168
DAYS_LAST_PH	168
FLAG_DOCUM	1
FLAG_DOCUM	0
FLAG_DOCUM	0
FLAG_DOCUM	0
FLAG_DOCUM	0
FLAG_DOCUM	0
FLAG_DOCUM	0
FLAG_DOCUM	0
FLAG_DOCUM	0
FLAG_DOCUM	0
FLAG_DOCUM	0
FLAG_DOCUM	0
FLAG_DOCUM	0
FLAG_DOCUM	0
FLAG_DOCUM	0
FLAG_DOCUM	0
FLAG_DOCUM	0
FLAG_DOCUM	0
FLAG_DOCUM	0
FLAG_DOCUM	0
AMT_REQ_CRE	0
AMT_REQ_CRE	6734
AMT_REQ_CRE	6734
AMT_REQ_CRE	6734
AMT_REQ_CRE	6734
AMT_REQ_CRE	6734

MEDIAN:



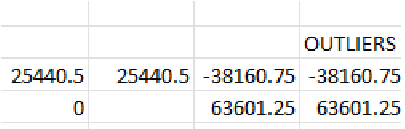
	12099.28
	168
	0

Graph:



B. Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

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



			OUTLIERS
25440.5	25440.5	-38160.75	-38160.75
0		63601.25	63601.25

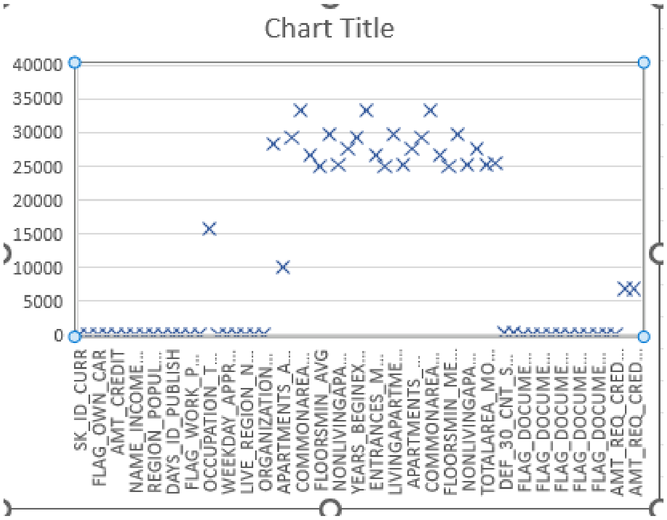
Upper Bound=

IQ*1.5+IQR

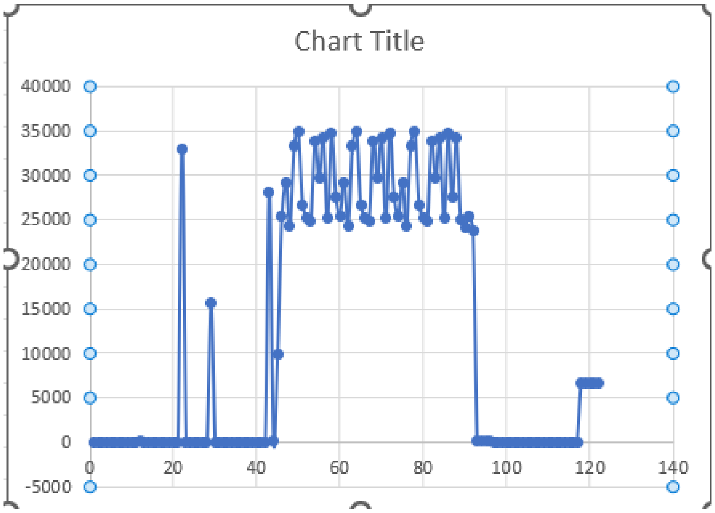
Lower Bound=

IQ*1.5-IQR

Box plot to visualize the outliers:

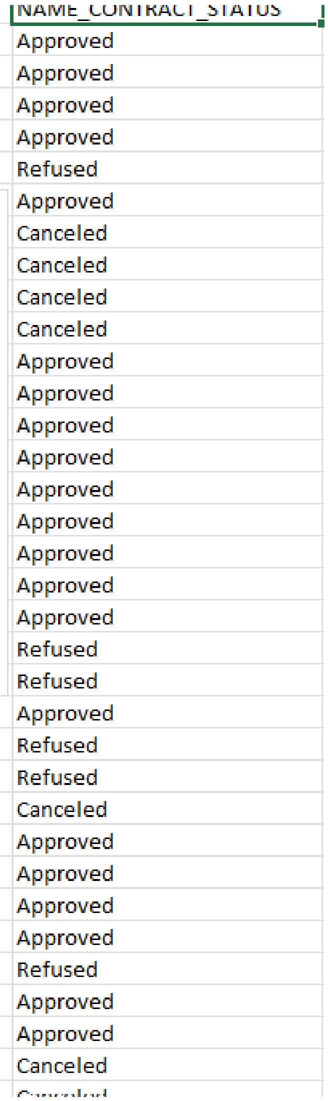


Scatter Plot To visualize outliers:



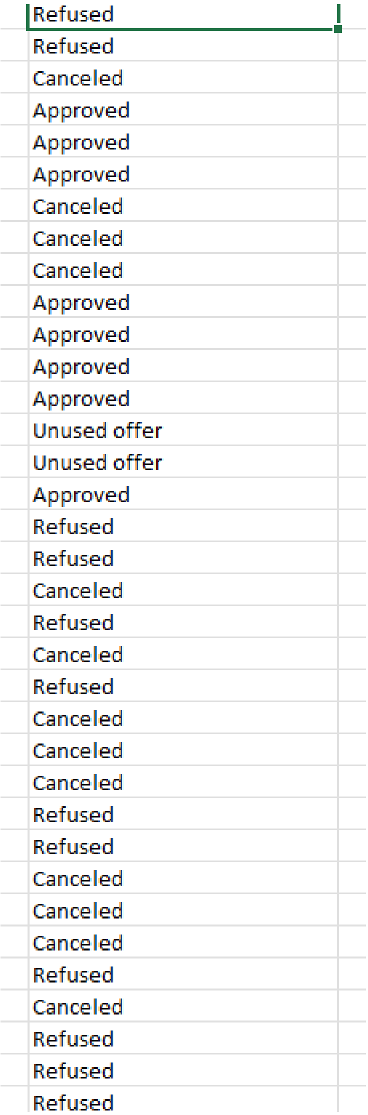
C. Analyze Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

- **Task:** Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.
- **Hint:** Utilize Excel functions like COUNTIF and SUM to calculate the proportions of each class. Compare the class frequencies to assess data imbalance.

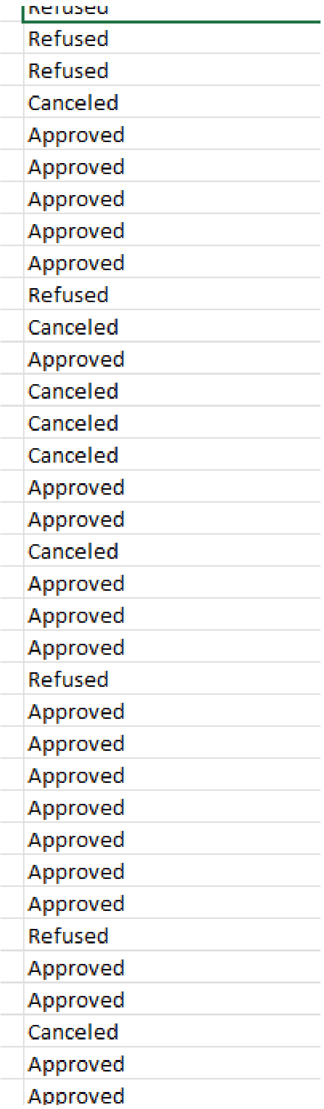


NAME_CONTRACT_STATUS
Approved
Approved
Approved
Approved
Refused
Approved
Canceled
Canceled
Canceled
Canceled
Approved
Approved
Approved
Approved
Approved
Approved
Approved
Approved
Approved
Refused
Refused
Approved
Refused
Refused
Canceled
Approved
Approved
Approved
Approved
Refused
Approved
Approved
Canceled
Canceled

Approved
Refused
Approved
Approved
Canceled
Approved
Canceled
Approved
Refused
Approved
Refused
Approved
Approved
Canceled
Approved
Approved
Canceled
Canceled
Canceled
Canceled
Canceled
Canceled
Canceled
Canceled

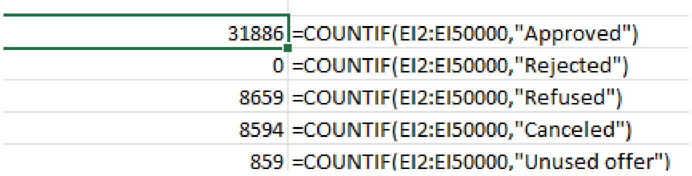


Refused
Refused
Canceled
Approved
Approved
Approved
Canceled
Canceled
Canceled
Approved
Approved
Approved
Approved
Unused offer
Unused offer
Approved
Refused
Refused
Canceled
Refused
Canceled
Refused
Canceled
Canceled
Canceled
Refused
Refused
Canceled
Canceled
Canceled
Refused
Canceled
Refused
Refused
Refused



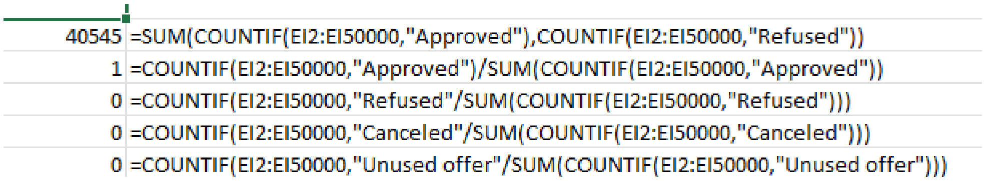
Refused
Refused
Refused
Canceled
Approved
Approved
Approved
Approved
Approved
Refused
Canceled
Approved
Canceled
Canceled
Canceled
Approved
Approved
Canceled
Approved
Approved
Approved
Refused
Approved
Approved
Approved
Approved
Approved
Approved
Approved
Refused
Approved
Approved
Canceled
Approved
Approved

Count the imbalances in the dataset:



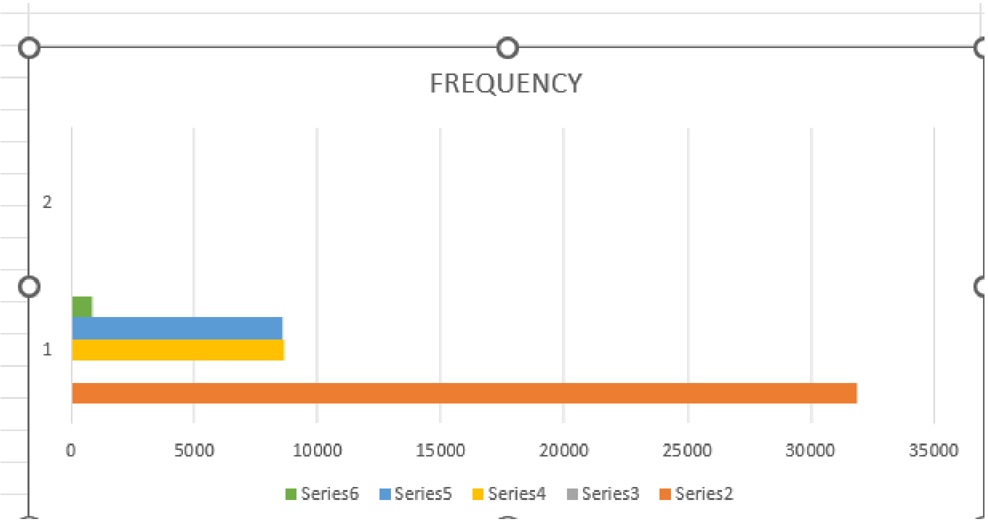
31886	=COUNTIF(EI2:EI50000,"Approved")
0	=COUNTIF(EI2:EI50000,"Rejected")
8659	=COUNTIF(EI2:EI50000,"Refused")
8594	=COUNTIF(EI2:EI50000,"Canceled")
859	=COUNTIF(EI2:EI50000,"Unused offer")

Calculate the proportion of imbalances:



40545	=SUM(COUNTIF(EI2:EI50000,"Approved"),COUNTIF(EI2:EI50000,"Refused"))
1	=COUNTIF(EI2:EI50000,"Approved")/SUM(COUNTIF(EI2:EI50000,"Approved"))
0	=COUNTIF(EI2:EI50000,"Refused"/SUM(COUNTIF(EI2:EI50000,"Refused")))
0	=COUNTIF(EI2:EI50000,"Canceled"/SUM(COUNTIF(EI2:EI50000,"Canceled")))
0	=COUNTIF(EI2:EI50000,"Unused offer"/SUM(COUNTIF(EI2:EI50000,"Unused offer")))

PLOT THE IMBALANCES USING BAR GRAPH:



D. Perform Univariate, Segmented Univariate, and Bivariate Analysis: To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

- **Task:** 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.

Univariate Analysis:

Average:

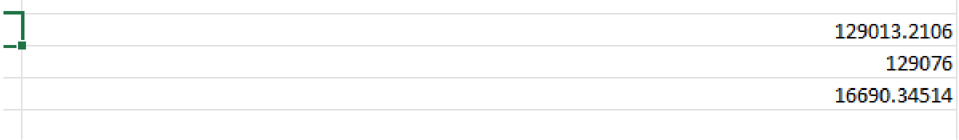
=AVG(A2:A50000)

MEDIAN

=MEDIAN(A2:A50000)

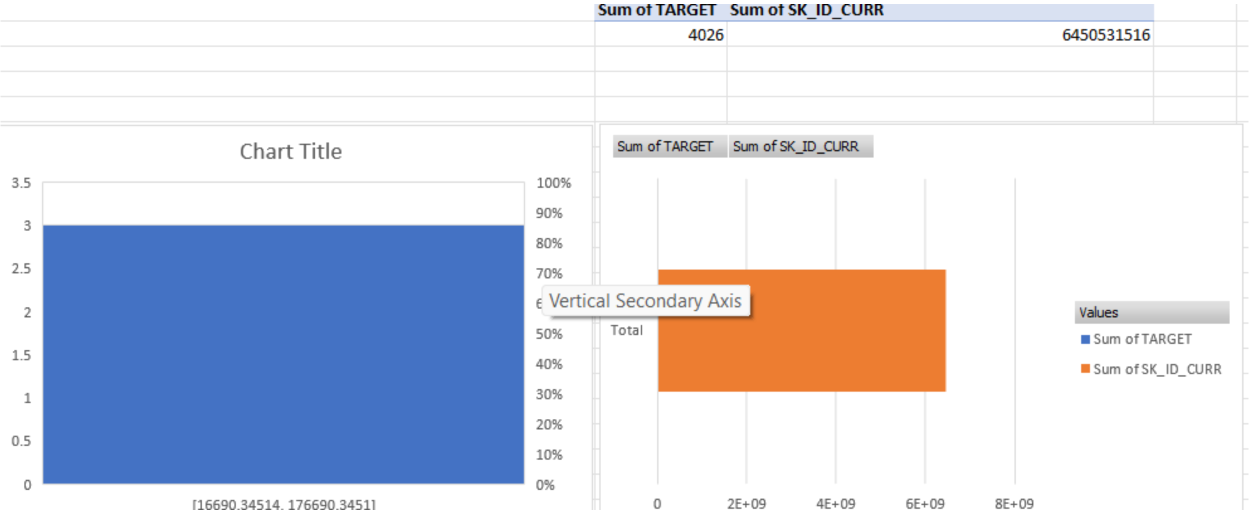
STANDARD DEVIATION:

=STDEV.P(A2:A50000)



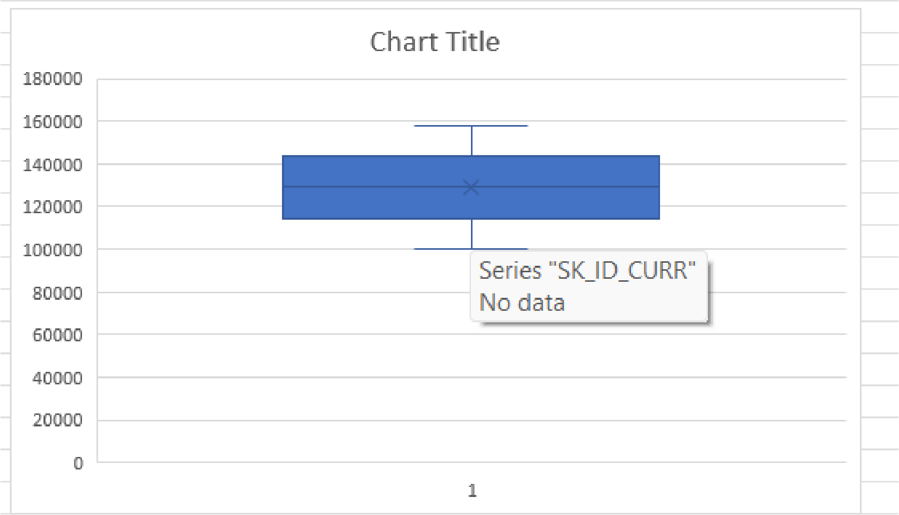
```
129013.2106
     129076
16690.34514
```

HISTOGRAM TO VISUALISE DISTRUBUTION VARIABLES :



Sum of TARGET	Sum of SK_ID_CURR
4026	6450531516

BOX PLOT :

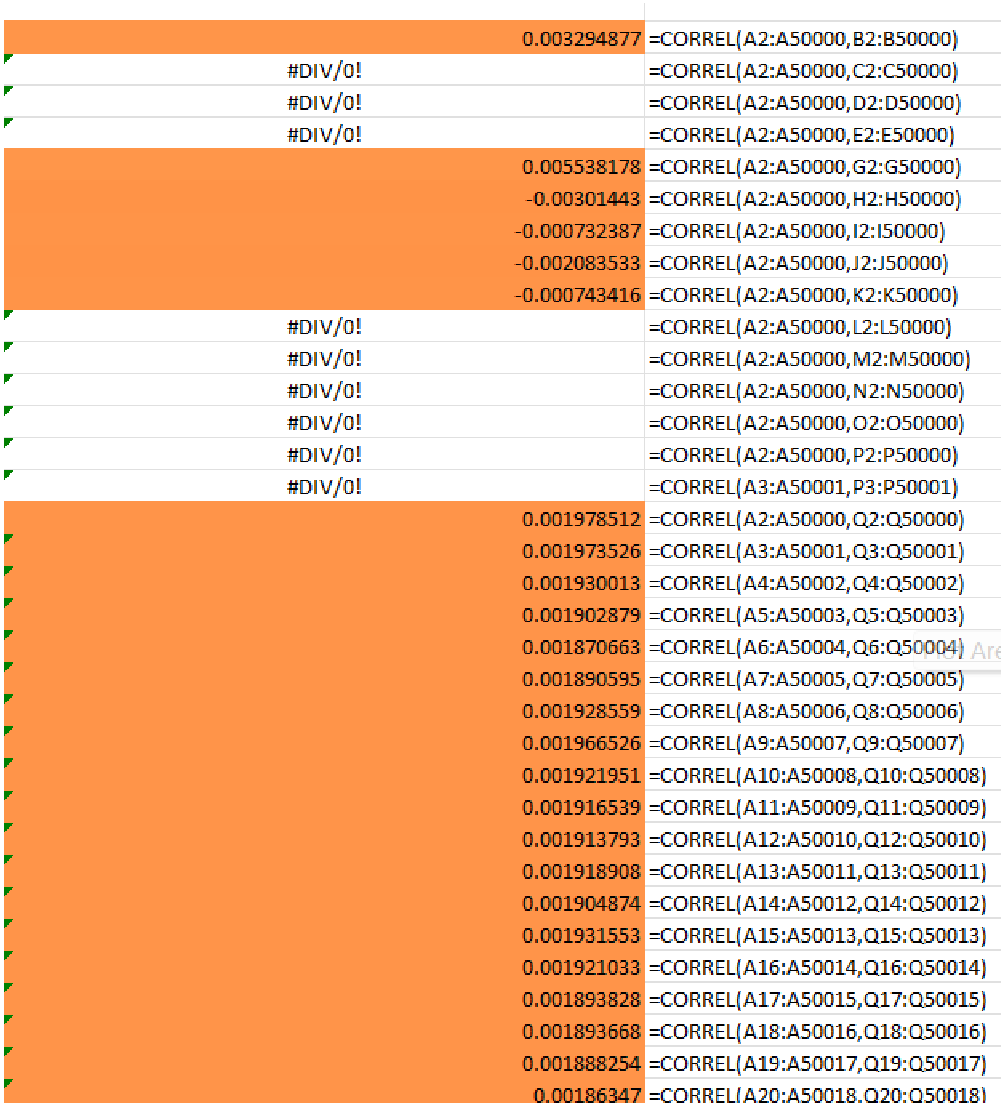


E. Identify Top Correlations for Different Scenarios: Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

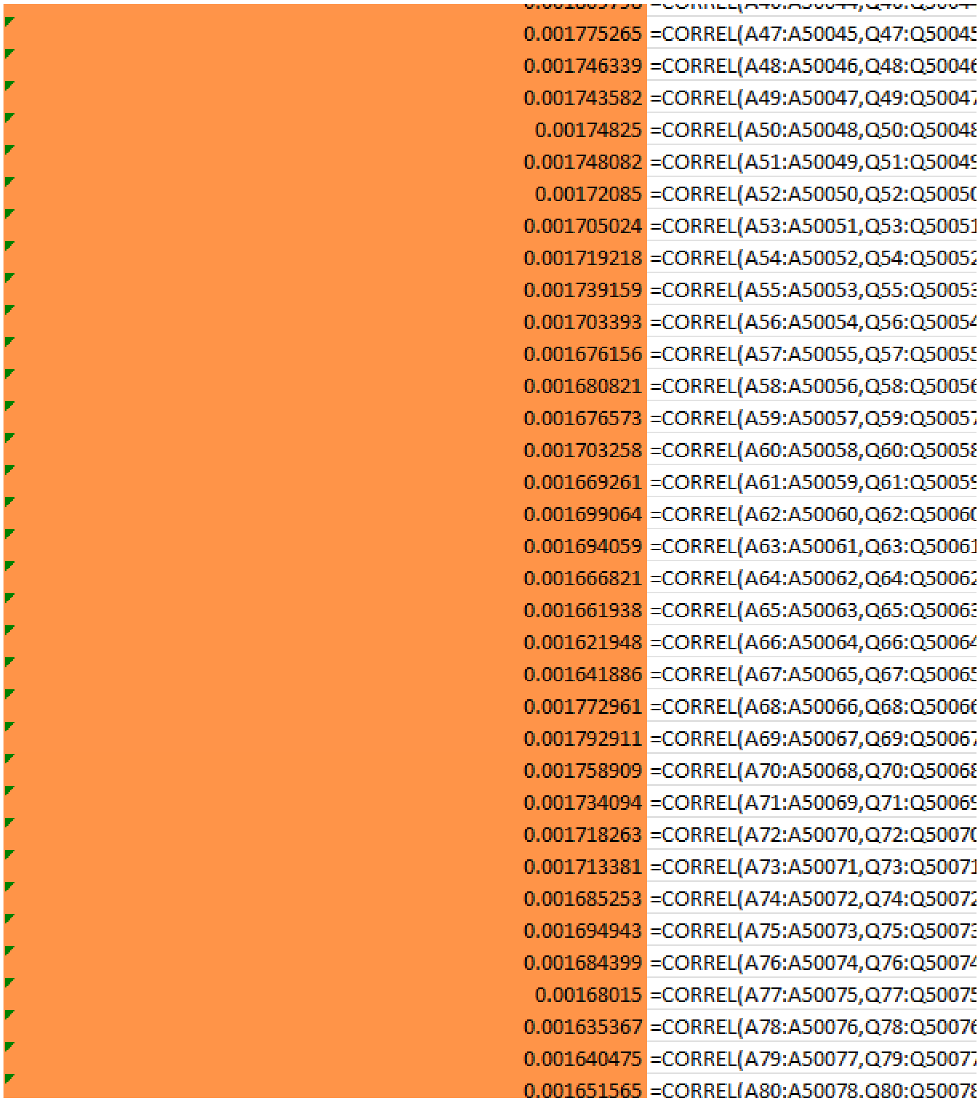
- **Task:** Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

Use correl function to determine correlation**:**

=CORREL(A2:A50000,B2:B50000)

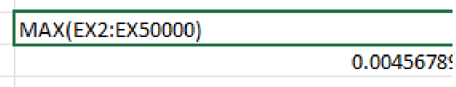


0.003294877	=CORREL(A2:A50000,B2:B50000)
#DIV/0!	=CORREL(A2:A50000,C2:C50000)
#DIV/0!	=CORREL(A2:A50000,D2:D50000)
#DIV/0!	=CORREL(A2:A50000,E2:E50000)
0.005538178	=CORREL(A2:A50000,G2:G50000)
-0.00301443	=CORREL(A2:A50000,H2:H50000)
-0.000732387	=CORREL(A2:A50000,I2:I50000)
-0.002083533	=CORREL(A2:A50000,J2:J50000)
-0.000743416	=CORREL(A2:A50000,K2:K50000)
#DIV/0!	=CORREL(A2:A50000,L2:L50000)
#DIV/0!	=CORREL(A2:A50000,M2:M50000)
#DIV/0!	=CORREL(A2:A50000,N2:N50000)
#DIV/0!	=CORREL(A2:A50000,O2:O50000)
#DIV/0!	=CORREL(A2:A50000,P2:P50000)
#DIV/0!	=CORREL(A3:A50001,P3:P50001)
0.001978512	=CORREL(A2:A50000,Q2:Q50000)
0.001973526	=CORREL(A3:A50001,Q3:Q50001)
0.001930013	=CORREL(A4:A50002,Q4:Q50002)
0.001902879	=CORREL(A5:A50003,Q5:Q50003)
0.001870663	=CORREL(A6:A50004,Q6:Q50004)
0.001890595	=CORREL(A7:A50005,Q7:Q50005)
0.001928559	=CORREL(A8:A50006,Q8:Q50006)
0.001966526	=CORREL(A9:A50007,Q9:Q50007)
0.001921951	=CORREL(A10:A50008,Q10:Q50008)
0.001916539	=CORREL(A11:A50009,Q11:Q50009)
0.001913793	=CORREL(A12:A50010,Q12:Q50010)
0.001918908	=CORREL(A13:A50011,Q13:Q50011)
0.001904874	=CORREL(A14:A50012,Q14:Q50012)
0.001931553	=CORREL(A15:A50013,Q15:Q50013)
0.001921033	=CORREL(A16:A50014,Q16:Q50014)
0.001893828	=CORREL(A17:A50015,Q17:Q50015)
0.001893668	=CORREL(A18:A50016,Q18:Q50016)
0.001888254	=CORREL(A19:A50017,Q19:Q50017)
0.00186347	=CORREL(A20:A50018,Q20:Q50018)



0.001775265	=CORREL(A47:A50045,Q47:Q50045
0.001746339	=CORREL(A48:A50046,Q48:Q50046
0.001743582	=CORREL(A49:A50047,Q49:Q50047
0.00174825	=CORREL(A50:A50048,Q50:Q50048
0.001748082	=CORREL(A51:A50049,Q51:Q50049
0.00172085	=CORREL(A52:A50050,Q52:Q50050
0.001705024	=CORREL(A53:A50051,Q53:Q50051
0.001719218	=CORREL(A54:A50052,Q54:Q50052
0.001739159	=CORREL(A55:A50053,Q55:Q50053
0.001703393	=CORREL(A56:A50054,Q56:Q50054
0.001676156	=CORREL(A57:A50055,Q57:Q50055
0.001680821	=CORREL(A58:A50056,Q58:Q50056
0.001676573	=CORREL(A59:A50057,Q59:Q50057
0.001703258	=CORREL(A60:A50058,Q60:Q50058
0.001669261	=CORREL(A61:A50059,Q61:Q50059
0.001699064	=CORREL(A62:A50060,Q62:Q50060
0.001694059	=CORREL(A63:A50061,Q63:Q50061
0.001666821	=CORREL(A64:A50062,Q64:Q50062
0.001661938	=CORREL(A65:A50063,Q65:Q50063
0.001621948	=CORREL(A66:A50064,Q66:Q50064
0.001641886	=CORREL(A67:A50065,Q67:Q50065
0.001772961	=CORREL(A68:A50066,Q68:Q50066
0.001792911	=CORREL(A69:A50067,Q69:Q50067
0.001758909	=CORREL(A70:A50068,Q70:Q50068
0.001734094	=CORREL(A71:A50069,Q71:Q50069
0.001718263	=CORREL(A72:A50070,Q72:Q50070
0.001713381	=CORREL(A73:A50071,Q73:Q50071
0.001685253	=CORREL(A74:A50072,Q74:Q50072
0.001694943	=CORREL(A75:A50073,Q75:Q50073
0.001684399	=CORREL(A76:A50074,Q76:Q50074
0.00168015	=CORREL(A77:A50075,Q77:Q50075
0.001635367	=CORREL(A78:A50076,Q78:Q50076
0.001640475	=CORREL(A79:A50077,Q79:Q50077
0.001651565	=CORREL(A80:A50078,Q80:Q50078

We can find the largest correl value using the max function on correl column :



MAX(EX2:EX50000)

0.00456789

TECH- STACK USED: I used MS Excel to apply the function keys and get the needed output. In MS Excel I could make pivot tables, graphs, relations from the given dataset.

INSIGHTS: Certainly! Let's dive into each task and provide the needed insights:

A. Identify Missing Data and Deal with it Appropriately

1. Identifying Missing Data:

- **COUNT Function:** Use =COUNTBLANK(range) to count missing (blank) cells within a specific range.

- **ISBLANK Function:** Create a new column with the formula =IF(ISBLANK(cell), "Missing", "Present") to label missing values in your dataset.

2. Dealing with Missing Data:

- **Imputation:**
  - For numerical data, calculate the mean or median of the column using =AVERAGE(range) or =MEDIAN(range) and replace missing values with this statistic.
  - For categorical data, consider using the mode or most frequent value.
- **Drop Missing Data**: If missing data is substantial and imputation is not appropriate, consider filtering out rows with missing values.

3. Visualize Missing Data:

- **Bar Chart/Column Chart:** Create a bar or column chart showing the count of missing values per variable to visualize the extent of missing data across the dataset**.**

B. Identify Outliers in the Dataset

1. Detecting Outliers:

- **Quartiles and IQR:**
  - Use =QUARTILE(range, 1) and =QUARTILE(range, 3) to calculate the first and third quartiles.
  - Compute the Interquartile Range (IQR) with =QUARTILE(range, 3) - QUARTILE(range, 1).
  - Define outlier thresholds: =QUARTILE(range, 1) - 1.5*IQR and =QUARTILE(range, 3) + 1.5*IQR.
- **Conditional Formatting:** Apply conditional formatting to highlight values that fall outside these thresholds.

2. Visualize Outliers:

- **Box Plot:** Create a box plot to visualize the distribution of numerical variables and highlight outliers.
- **Scatter Plot:** Use scatter plots to identify outliers visually by plotting data points and observing any that deviate significantly from the rest.

C. Analyze Data Imbalance

1. Determine Data Imbalance:

- **Class Distribution**: Use =COUNTIF(range, criteria) to count occurrences of each class in the target variable.

- **Calculate Ratios**: Compute the proportion of each class relative to the total using =COUNTIF(range, criteria)/TOTAL_COUNT.

2. Visualize Data Imbalance:

- **Pie Chart:** Create a pie chart to display the proportion of each class in the target variable**.**
- **Bar Chart**: Use a bar chart to show the frequency of each class and highlight any imbalances**.**

D. Perform Univariate, Segmented Univariate, and Bivariate Analysis

1. Univariate Analysis:

- **Descriptive Statistics:** Use =AVERAGE(range), =MEDIAN(range), =STDEV.P(range) for numerical variables, and =COUNTIF(range, criteria) for categorical variables**.**
- **Histograms:** Create histograms to visualize the distribution of individual numerical variables.

2. Segmented Univariate Analysis:

- **Filtering/Sorting**: Use Excel filters or sort functions to segment data (e.g., by loan approval status) and perform descriptive statistics on each segment**.**
- **Pivot Tables:** Use pivot tables to compare variable distributions across different segments

Bivariate Analysis:

- **Correlation:** Use =CORREL(range1, range2) to calculate correlations between two variables**.**
- **Scatter Plots:** Create scatter plots to explore relationships between variables and the target variable.

E. Identify Top Correlations for Different Scenarios

1. Calculate Correlations:

- **Segmented Data:** Filter or segment the dataset based on scenarios (e.g., clients with payment difficulties).
- **Correlation Coefficients:** Use =CORREL(range1, range2) to compute correlation coefficients between variables and the target variable within each segment**.**

2. Visualize Correlations:

- **Correlation Matrix/Heatmap:** Create a correlation matrix or heatmap to visualize the strength of relationships between variables in each segment**.**
- **Highlight Top Correlations:** Use different colors or shading to highlight the top correlations**.**

RESULT: This is the most interesting task I've done so far. I've learned how to combine theoretical knowledge with practical knowledge. I've learned to translate complex data insights into understandable and actionable information.