BANK LOAN CASE STUDY

DRIVE LINK

https://drive.google.com/drive/folders/1Mdb-ERKqcVMQJz7I9Iua_CU8QELULfhe

DATASET LINK:

https://docs.google.com/spreadsheets/d/1zPp9fBw-3hex291G4xu0GLM6zRJF1HiA/edit?usp=drivesdk&ouid=102183820204754509514&rtpof=true&sd=true

PROJECT DESCRIPTION:

As a data analyst in a finance company specializing in lending, the primary objective is to mitigate the risks associated with loan default. The company faces two critical risks: losing business by denying capable applicants and facing financial losses from approving applicants who cannot repay their loans. To address these challenges, we have to conduct Exploratory Data Analysis (EDA) on a dataset containing information about loan applications. The dataset encompasses two main scenarios: customers with payment difficulties and all other cases. The former refers to customers who have experienced late payments exceeding a certain threshold on their loan installments. The ultimate goal of this project is to discern patterns within the data to understand how customer attributes and loan attributes influence the likelihood of default. This analysis will inform decision-making processes such as loan approval, loan amount adjustments, and interest rate adjustments for risky applicants.

APPROACH:

- Identifying Missing Data:
  - Use Excel functions (COUNT, ISBLANK, IF) to identify missing data.
  - Employ appropriate methods for handling missing data, such as imputation (AVERAGE, MEDIAN).
  - Visualize missing values proportion using bar charts or column charts.
- Identifying Outliers:
  - Utilize Excel statistical functions (QUARTILE, IQR) and conditional formatting to detect outliers.
  - Determine the validity of outliers by applying thresholds or business rules.
  - Visualize outliers using box plots or scatter plots.
- Analyzing Data Imbalance:
  - Assess data distribution and calculate the imbalance ratio using Excel functions (COUNTIF, SUM).
  - Visualize class imbalance using pie charts or bar charts.
- Conducting Various Analyses:
  - Perform univariate analysis to understand variable distributions.
  - Conduct segmented univariate analysis to compare distributions across different scenarios.
  - Explore relationships through bivariate analysis using Excel functions and features.
  - Visualize distributions and relationships using histograms, bar charts, box plots, stacked bar charts, grouped bar charts, scatter plots, or heatmaps.
- Identifying Top Correlations:
  - Segment the dataset based on different scenarios (e.g., payment difficulties vs. other cases).
  - Calculate correlation coefficients within each segment using Excel functions like CORREL.
  - Rank correlations to identify top indicators of loan default and visualize them using correlation matrices or heatmaps.

By following this approach, the project aims to uncover critical patterns influencing loan default, empowering the company to make informed decisions and mitigate risks effectively.

TECH STACK USED:

Software used: Microsoft Excel 2010

Operating System used: Windows 10

Purpose: Exploratory Data Analysis

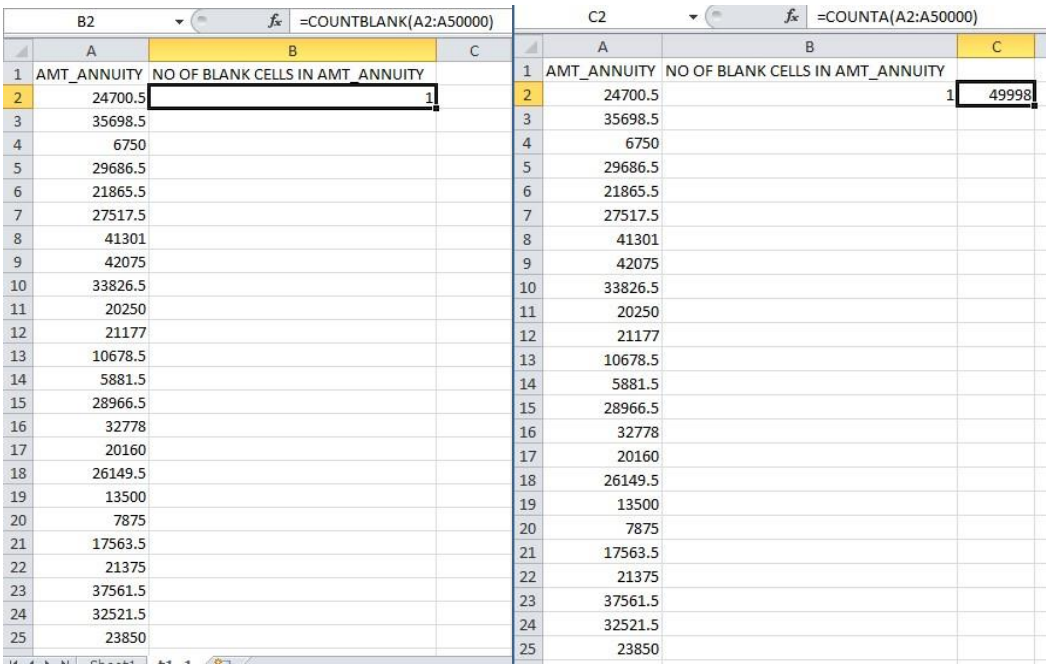
DATA ANALYTICS TASK:

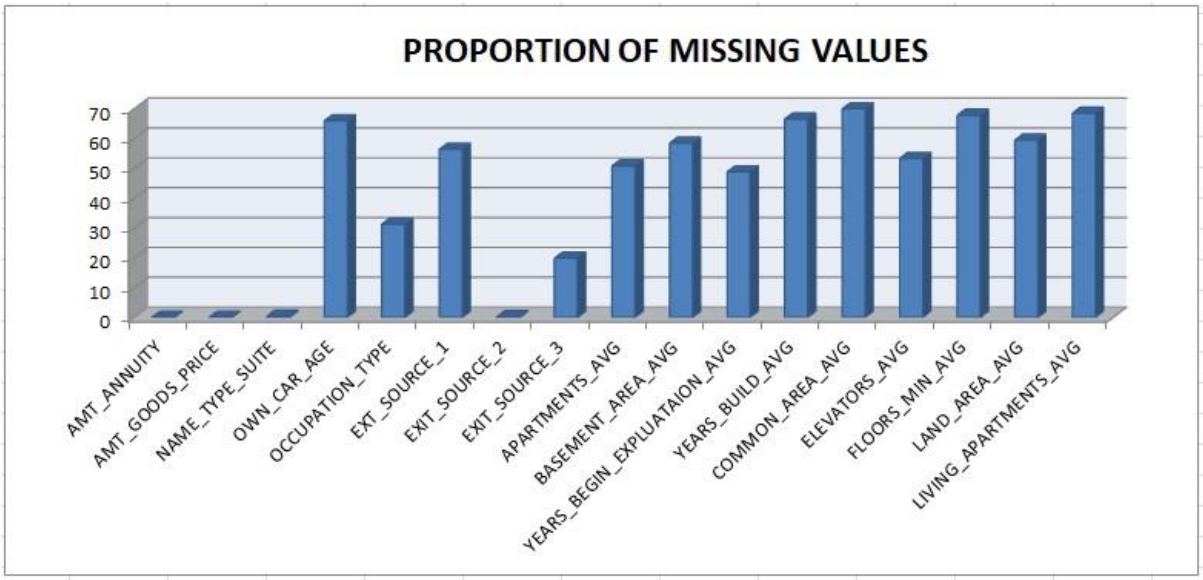
A. **Identify Missing Data and Deal with it Appropriately:** As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

- **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.

INSIGHTS:

I have used **COUNTBLANK** function to get the number of empty cells and **COUNTA** function to get the number of non-empty cells in column **AMT_ANNUITY**.





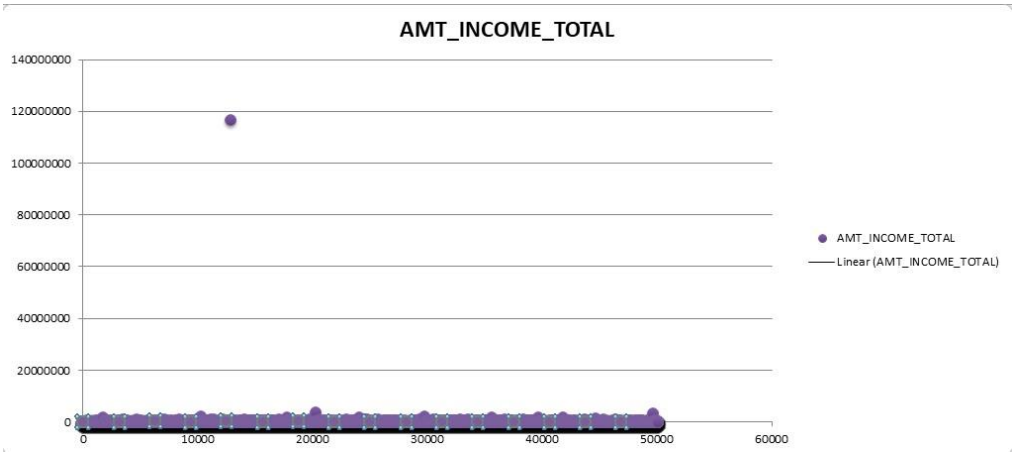
To get the proportion of missing values we will first find number of empty cells in columns using the **COUNTBLANK** function and then will divide the no. of balnk cells in a column with total no. of cells.

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.

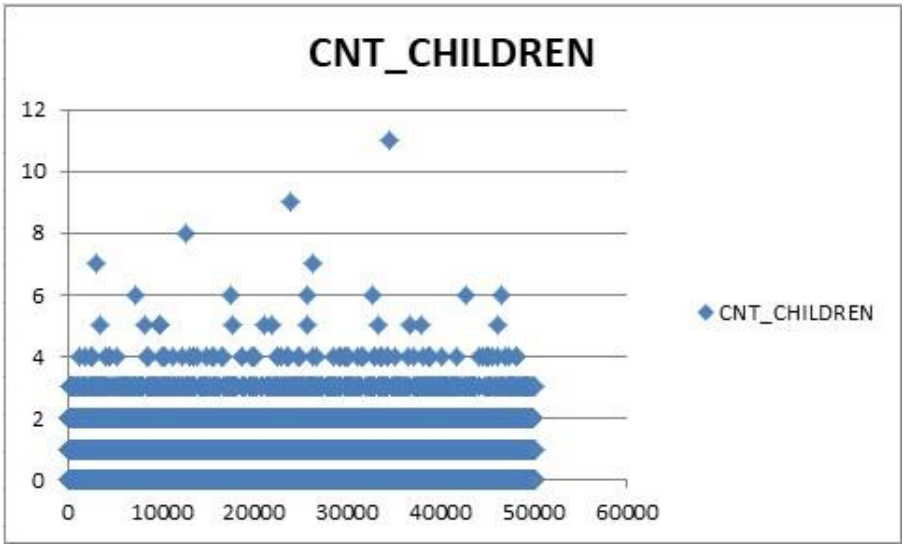
- **Task:** Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.
- **Hint:** Utilize Excel functions like QUARTILE, IQR, and conditional formatting to identify potential outliers. Consider applying thresholds or business rules to 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.

INSIGHTS:

In the below scatter plot we can see that one person's income is around 11 crores whereas majority of people have income in lacs only.



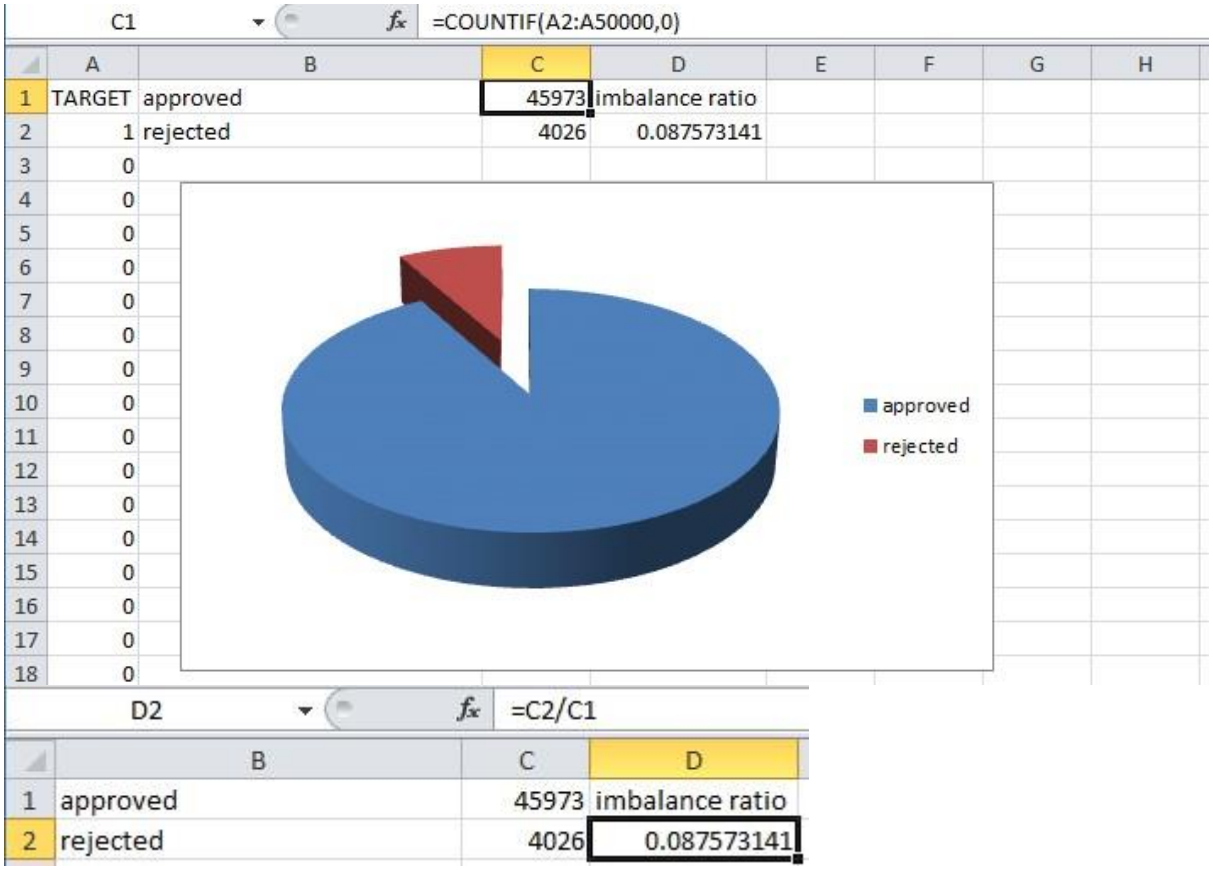
The scatter plot shows the **CNT_CHLDREN** column. One client has 11 number of children whereas majority of people have 0 to 4 number of children.



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.

INSIGHTS:



To identify data imbalance, I have taken target variable(0 – approved and 1 – rejected). I have used **COUNTIF** function to find the number of approved and rejected loans and the imbalance ratio is calculated by using formula **(rejected loan / approved loan**).

Hence, the pie chart shows data imbalance as **the approved value dominates the rejected value.**

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.

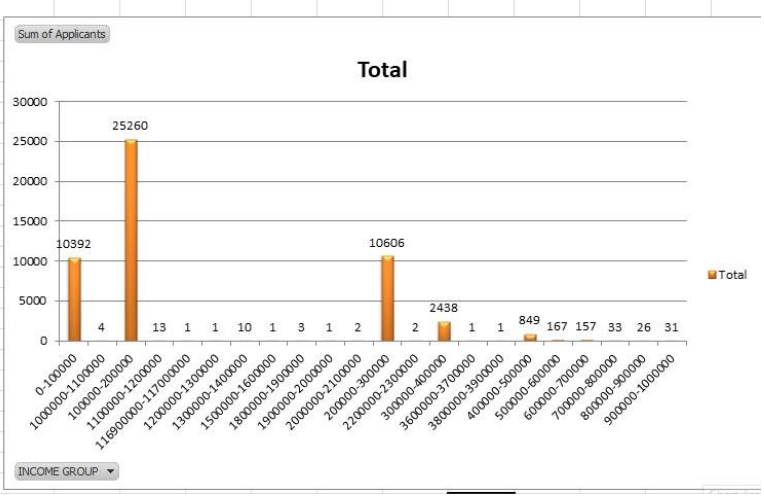
- **Task:** Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and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

INSIGHTS:

UNIVARIATE ANALYSIS

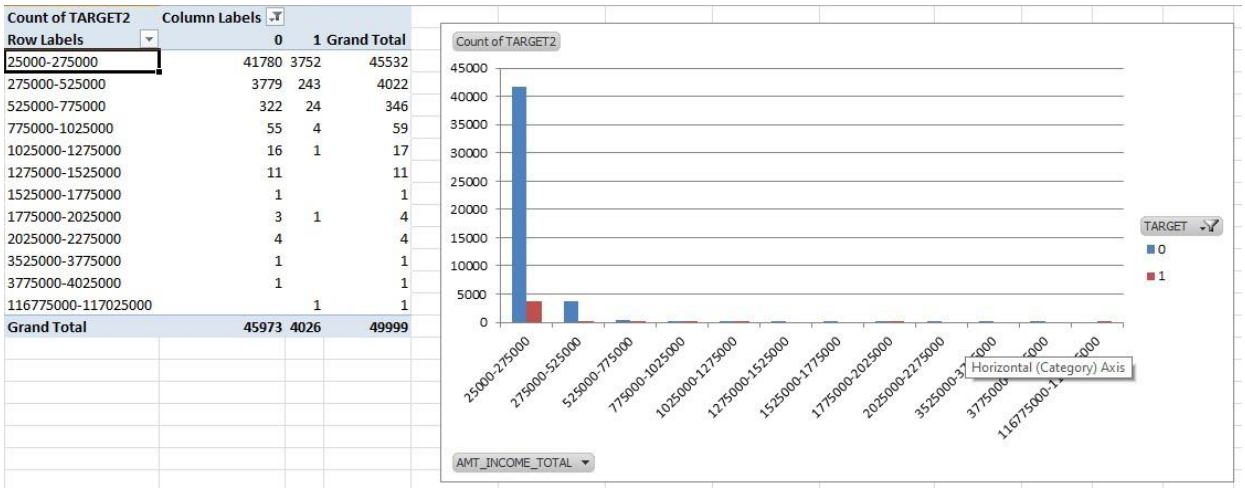
Here I have used **Pivot Table** for Univariate Analysis. I have used group function in pivot table to group the income and count function to get the number of applicants approved in that particular range. So the most number of applicants approved are having an income in the range 1000000 – 2000000.

Row Labels	Sum of Applicants
0-100000	10392
1000000-1100000	4
100000-200000	25260
1100000-1200000	13
116900000-117000000	1
1200000-1300000	1
1300000-1400000	10
1500000-1600000	1
1800000-1900000	3
1900000-2000000	1
2000000-2100000	2
200000-300000	10606
2200000-2300000	2
300000-400000	2438
3600000-3700000	1
3800000-3900000	1
400000-500000	849
500000-600000	167
600000-700000	157
700000-800000	33
800000-900000	26
900000-1000000	31
Grand Total	**49999**



SEGMENTED UNIVARIATE ANSLYSIS

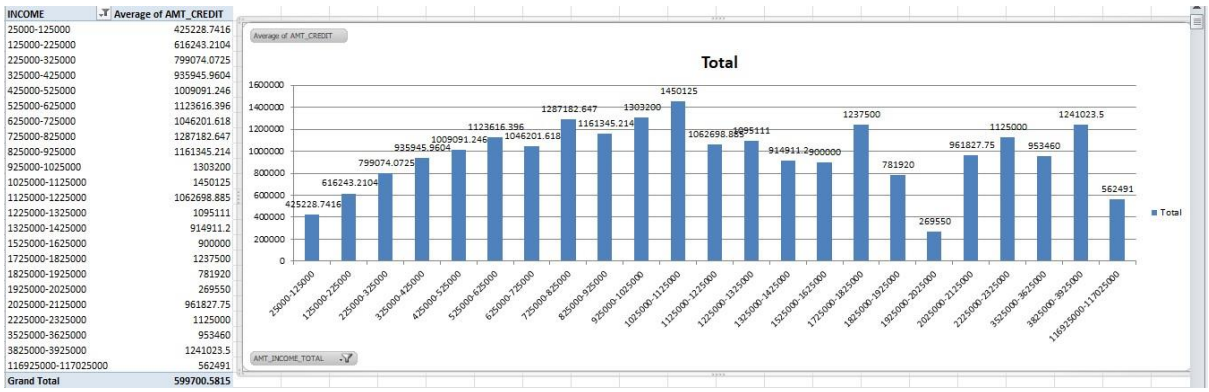
Here I have used **Pivot Table** for Segmented Univariate Analysis. I have used group function in pivot table to group the income and count function to get the number of (Approved/Rejected) applicants in that particular range. So the most number of applicants approved/rejected are having an income in the range 25000 – 275000.



Count of TARGET2	Column Labels		
Row Labels	0	1	Grand Total
25000-275000	41780	3752	45532
275000-525000	3779	243	4022
525000-775000	322	24	346
775000-1025000	55	4	59
1025000-1275000	16	1	17
1275000-1525000	11		11
1525000-1775000	1		1
1775000-2025000	3	1	4
2025000-2275000	4		4
3525000-3775000	1		1
3775000-4025000	1		1
116775000-117025000		1	1
Grand Total	45973	4026	49999

BIVARIATE ANALYSIS

Here I have used **Pivot Table** for Bivariate Analysis. I have used group function in pivot table to group the income and **AVERAGE** function to get the average of **CNT_CREDIT** in that particular range. So the maximum is 1450125.



INCOME	Average of AMT_CREDIT
25000-125000	425228.7416
125000-225000	616243.2104
225000-325000	799074.0725
325000-425000	935945.9604
425000-525000	1009091.246
525000-625000	1123616.396
625000-725000	1046201.618
725000-825000	1287182.647
825000-925000	1161345.214
925000-1025000	1303200
1025000-1125000	1450125
1125000-1225000	1062698.885
1225000-1325000	1095111
1325000-1425000	914911.2
1525000-1625000	900000
1725000-1825000	1237500
1825000-1925000	781920
1925000-2025000	269550
2025000-2125000	961827.75
2225000-2325000	1125000
3525000-3625000	953460
3825000-3925000	1241023.5
116925000-117025000	562491
Grand Total	599700.5815

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.

INSIGHTS:

The heat map shows the correlations between the different variables. I have used the colour combination of **light red** to show weakest correlations and **dark red** to show strongest correlations.

CORRELATION	CNT_CHILDR	AMT_INCOME_TO	AMT_CRED	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_IDPUBLISHED	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.009588558	0.00497	-0.025555665	0.329263754	-0.239693041	0.181217183	-0.032115773	0.025913889
AMT_INCOME_TOTAL	0.0095886	1	0.06932	0.029841469	0.016002774	-0.031615555	0.009952379	0.003506646	-0.038188511
AMT_CREDIT	0.0049716	0.069315897	1	0.095111221	-0.059342658	-0.070471393	0.003448569	-0.012228765	-0.100507425
REGION_POPULLATION_REL	-0.025556	0.029841469	0.09511	1	-0.032513748	-0.004101686	-0.059322344	-0.004345136	-0.532667302
DAYS_BIRTH	0.3292638	0.016002774	-0.05934	-0.032513748	1	-0.613553972	0.333632509	0.270825141	0.016779196
DAYS_EMPLOYED	-0.239693	-0.031615555	-0.07047	-0.004101686	-0.613553972	1	-0.204680611	-0.270382022	0.034321673
DAYS_REGISTRATION	0.1812172	0.009952379	0.00345	-0.059322344	0.333632509	-0.204680611	1	0.104298561	0.087517643
DAYS_ID_PUBLISHED	-0.032116	0.003506646	-0.01223	-0.004345136	0.270825141	-0.270382022	0.104298561	1	-0.002307011
REGION_RATING_CLIENT	0.0259139	-0.038188511	-0.10051	-0.532667302	0.016779196	0.034321673	0.087517643	-0.002307011	1