BANK LOAN CASE STUDY

Data Analytics Project by Soumya Singhal

PROJECT DESCRIPTION

- In this project, the goal is to analyze loan application data to identify patterns that predict the likelihood of loan default.
- The project involves Exploratory Data Analysis (EDA) to explore how different customer attributes and loan characteristics influence the probability of defaulting on a loan.
- The key business objectives are to reduce risk by identifying potentially problematic applicants and improving the loan approval process.

APPROACH

To achieve the project objectives, a systematic and structured approach was followed, leveraging Excel's advanced features and statistical techniques to derive meaningful insights.

1. **Data Preparation**:
   - Analyzed missing data using functions like COUNTBLANK and IF.
   - Filled missing values with median or mode where applicable, and removed columns with over 40% missing data.

2. **Outlier Detection**:
   - Used statistical measures like IQR to identify and visualize outliers.
   - Created box plots to display the data distribution and pinpoint anomalies.

3.Data Imbalance Analysis:

- •Evaluated the distribution of the target variable using COUNTIF.
- •Visualized the class imbalance through pie charts and bar charts to highlight potential issues.

4.Exploratory Data Analysis:

- •Performed **Univariate Analysis** to examine individual variable distributions using histograms and bar charts.
- •Conducted **Segmented Univariate Analysis** by comparing variable distributions for defaulters and non-defaulters.
- •Executed **Bivariate Analysis** to study relationships between variables, employing scatter plots and heatmaps for insights.

5.Correlation Analysis:

- •Generated correlation matrices for the overall dataset, defaulters, and non-defaulters separately.
- •Highlighted strong correlations (>0.5) using color coding for easy interpretation.
- •Identified key correlated variables and summarized them in a table for further analysis.

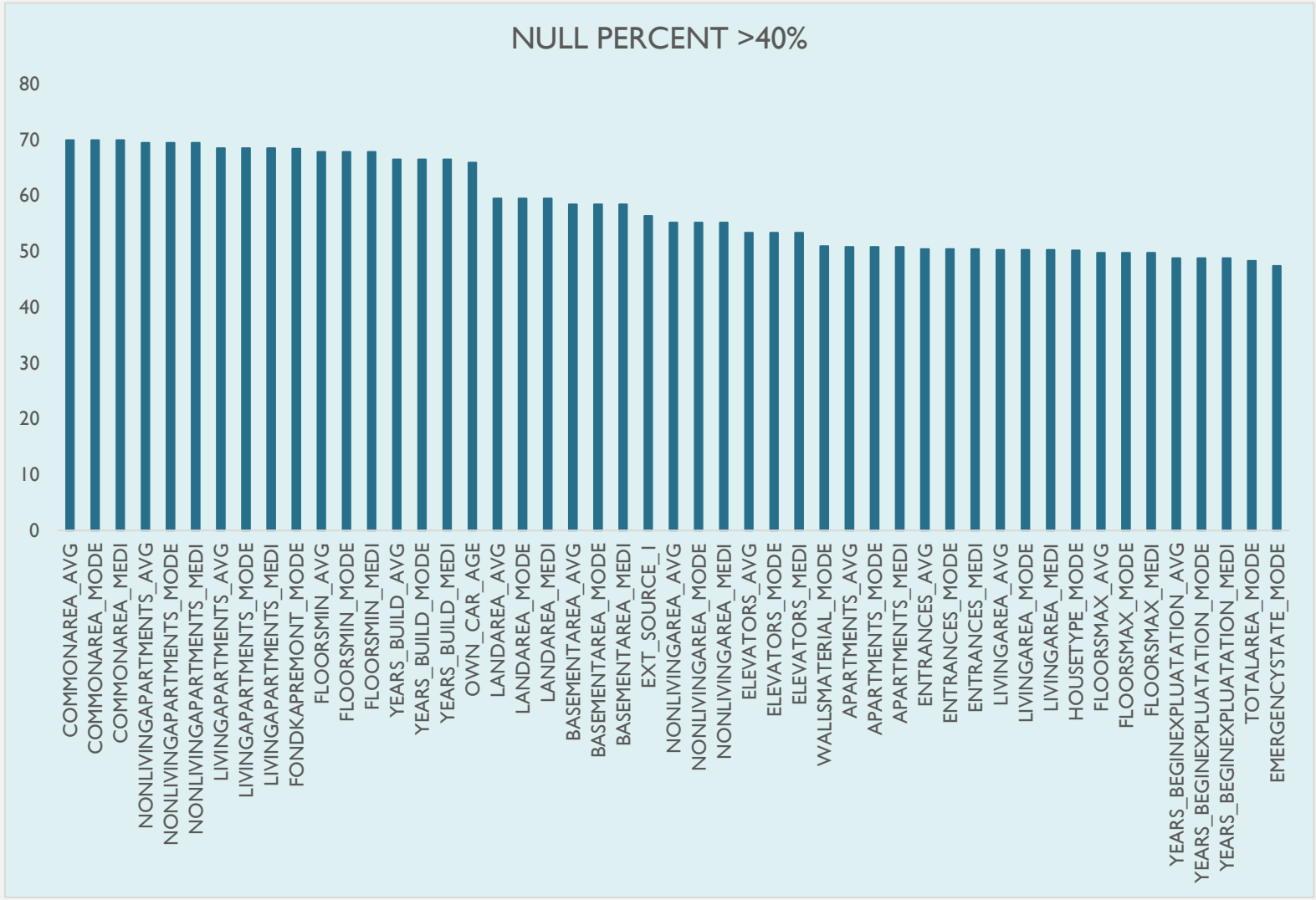
This comprehensive approach ensured a deep understanding of the dataset and provided actionable insights into loan default patterns.

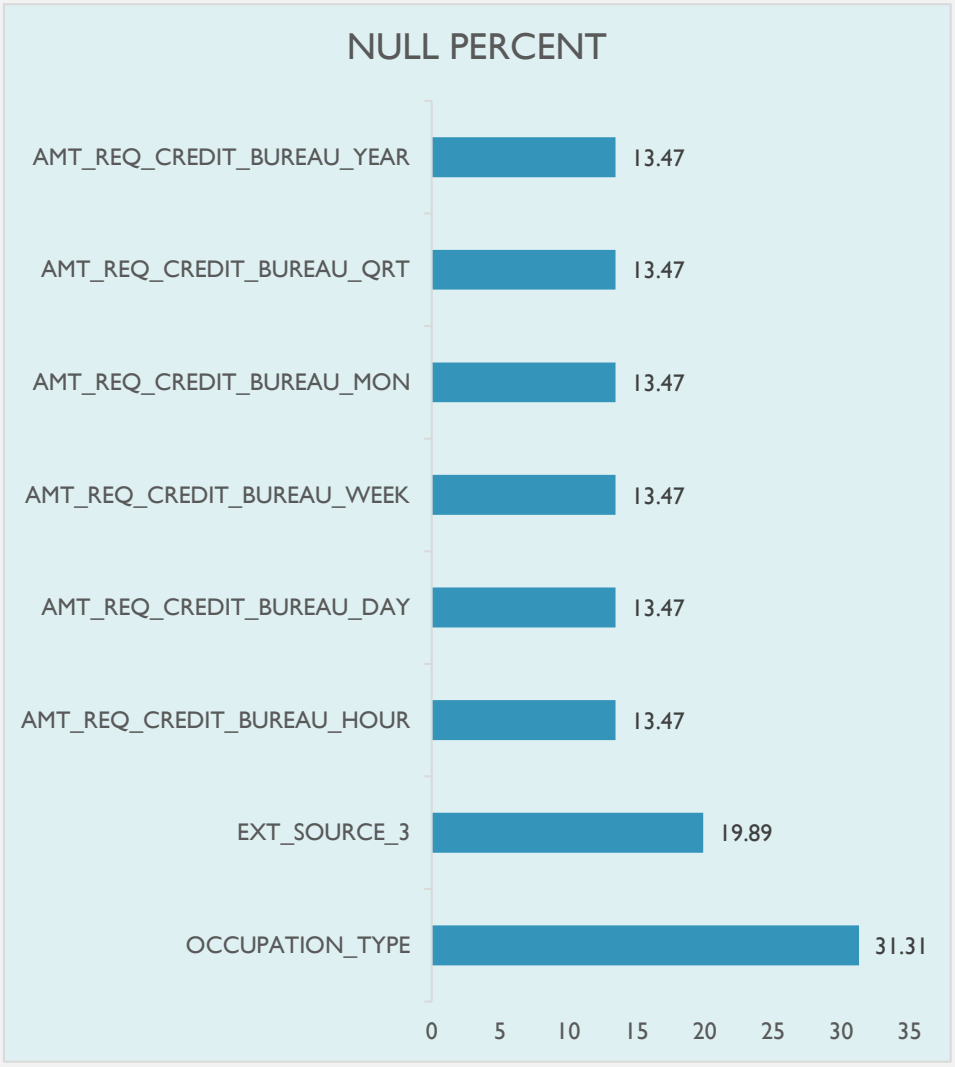
TECH-STACK USED

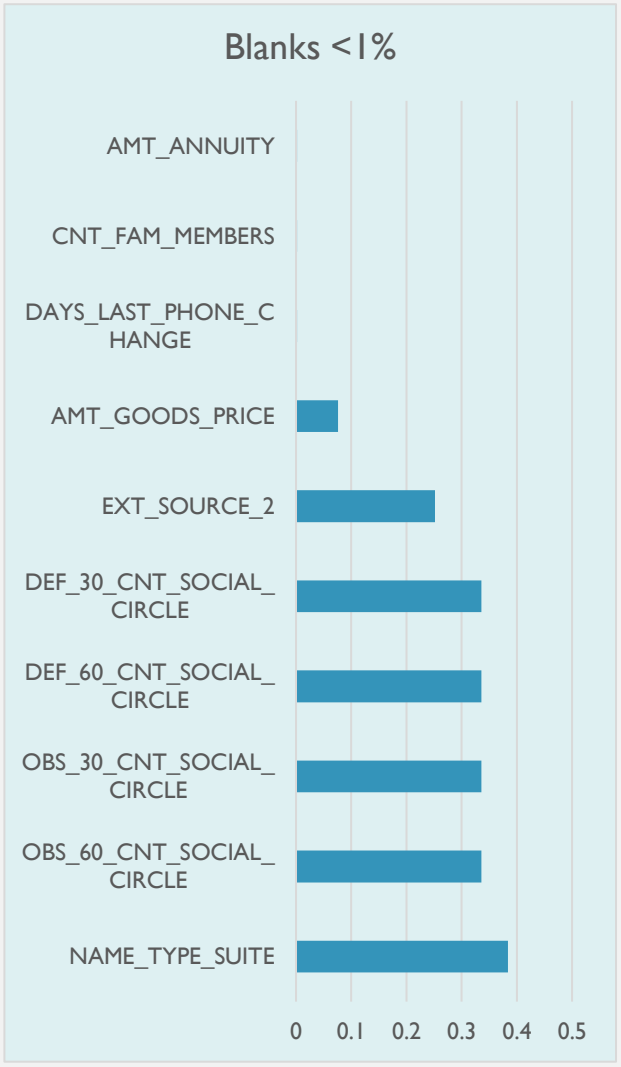
- MS Excel
- MS PowerPoint

TASK A: IDENTIFY AND HANDLE MISSING DATA

- Counted the number of blank rows in each column using the COUNTBLANK function and calculated the percentage of missing values.
- Created a bar chart and column chart to visualise the proportion of missing values for each variable.
- Removed columns(49) that contained more than 40% blank cells.
- Filled missing values in the `NAME_TYPE_SUITE` column using the mode i.e. Unaccompanied.
- Filled missing values in other columns, where less than 40% of the data was missing, using the median or mode.

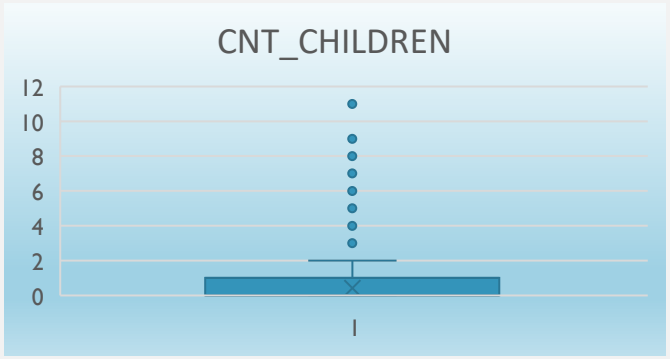

NULL PERCENT >40%
80
70
60
50
40
30
20
10
0
COMMONAREA_AVG
COMMONAREA_MODE
COMMONAREA_MEDI
NONLIVINGAPARTMENTS_AVG
NONLIVINGAPARTMENTS_MODE
NONLIVINGAPARTMENTS_MEDI
LIVINGAPARTMENTS_AVG
LIVINGAPARTMENTS_MODE
LIVINGAPARTMENTS_MEDI
FONDKAPREMONT_MODE
FLOORSMIN_AVG
FLOORSMIN_MODE
FLOORSMIN_MEDI
YEARS_BUILD_AVG
YEARS_BUILD_MODE
YEARS_BUILD_MEDI
OWN_CAR_AGE
LANDAREA_AVG
LANDAREA_MODE
LANDAREA_MEDI
BASEMENTAREA_AVG
BASEMENTAREA_MODE
BASEMENTAREA_MEDI
EXT_SOURCE_1
NONLIVINGAREA_AVG
NONLIVINGAREA_MODE
NONLIVINGAREA_MEDI
ELEVATORS_AVG
ELEVATORS_MODE
ELEVATORS_MEDI
WALLSMATERIAL_MODE
APARTMENTS_AVG
APARTMENTS_MODE
APARTMENTS_MEDI
ENTRANCES_AVG
ENTRANCES_MODE
ENTRANCES_MEDI
LIVINGAREA_AVG
LIVINGAREA_MODE
LIVINGAREA_MEDI
HOUSETYPE_MODE
FLOORSMAX_AVG
FLOORSMAX_MODE
FLOORSMAX_MEDI
YEARS_BEGINEXPLUATATION_AVG
YEARS_BEGINEXPLUATATION_MODE
YEARS_BEGINEXPLUATATION_MEDI
TOTALAREA_MODE
EMERGENCYSTATE_MODE

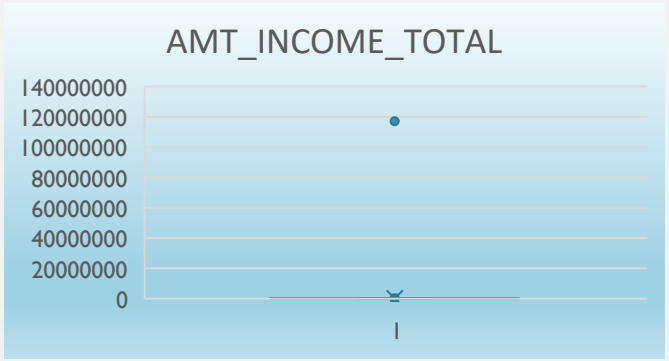

NULL PERCENT
AMT_REQ_CREDIT_BUREAU_YEAR
13.47
AMT_REQ_CREDIT_BUREAU_QRT
13.47
AMT_REQ_CREDIT_BUREAU_MON
13.47
AMT_REQ_CREDIT_BUREAU_WEEK
13.47
AMT_REQ_CREDIT_BUREAU_DAY
13.47
AMT_REQ_CREDIT_BUREAU_HOUR
13.47
EXT_SOURCE_3
19.89
OCCUPATION_TYPE
31.31
0
5
10
15
20
25
30
35

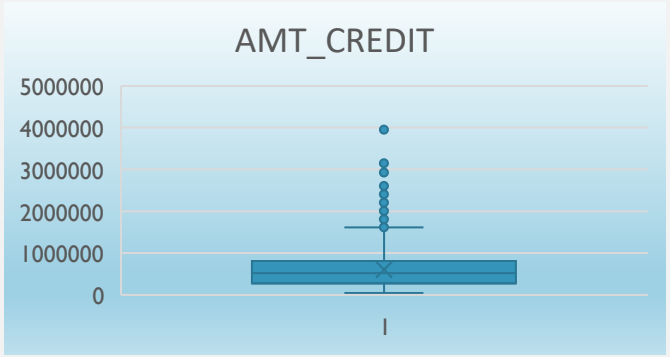

Blanks <1%
AMT_ANNUITY
CNT_FAM_MEMBERS
DAYS_LAST_PHONE_CHANGE
AMT_GOODS_PRICE
EXT_SOURCE_2
DEF_30_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
OBS_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
NAME_TYPE_SUITE
0
0.1
0.2
0.3
0.4
0.5

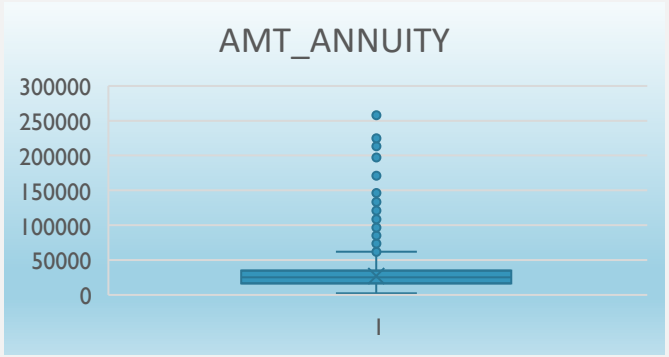
TASK B: IDENTIFY OUTLIERS

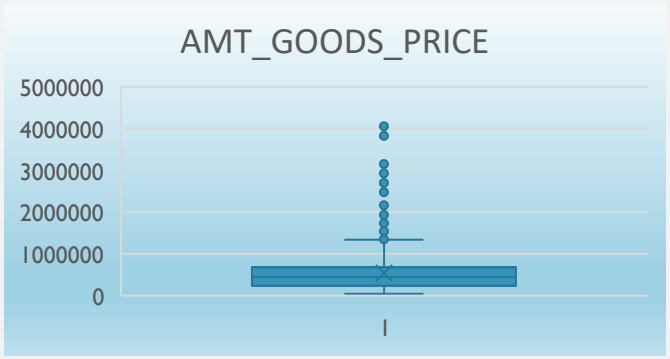
- Detect outliers using statistical methods such as QUARTILE.EXC and IQR in Excel.
- Visualize outliers using box plots or scatter plots.

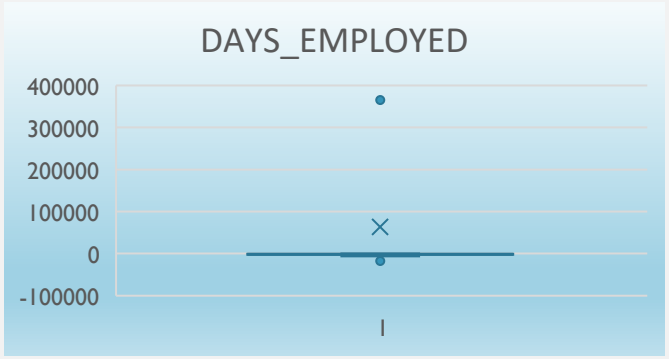

CNT_CHILDREN
12
10
8
6
4
2
0
1

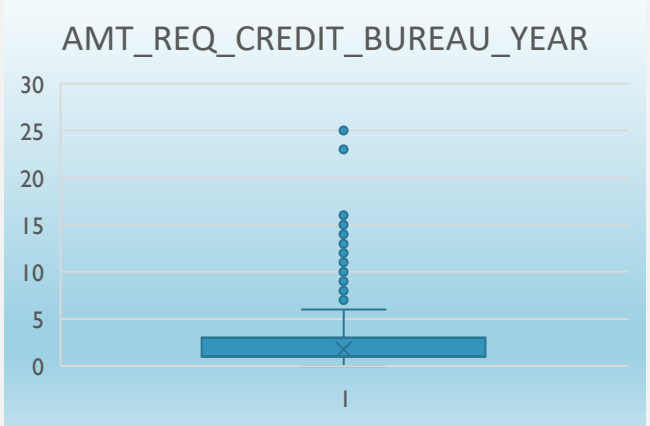

AMT_INCOME_TOTAL
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
1

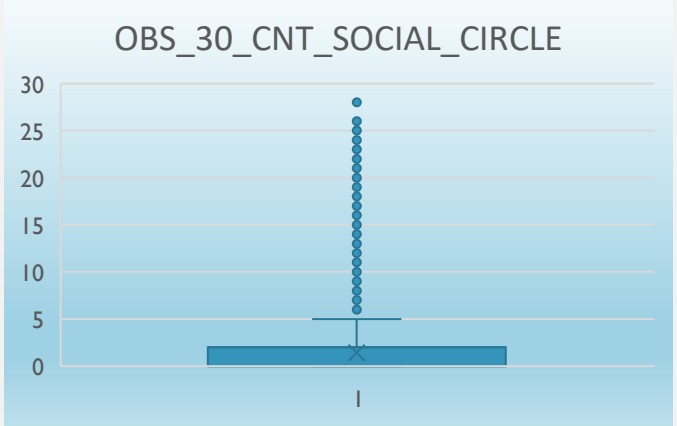

AMT_CREDIT
5000000
4000000
3000000
2000000
1000000
0
1

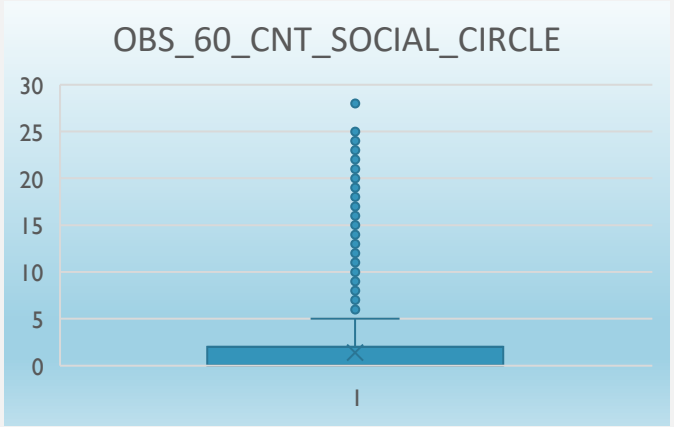

AMT_ANNUITY
300000
250000
200000
150000
100000
50000
0
1


AMT_GOODS_PRICE
5000000
4000000
3000000
2000000
1000000
0
1


DAYS_EMPLOYED
400000
300000
200000
100000
0
-100000
1


AMT_REQ_CREDIT_BUREAU_YEAR
30
25
20
15
10
5
0
1

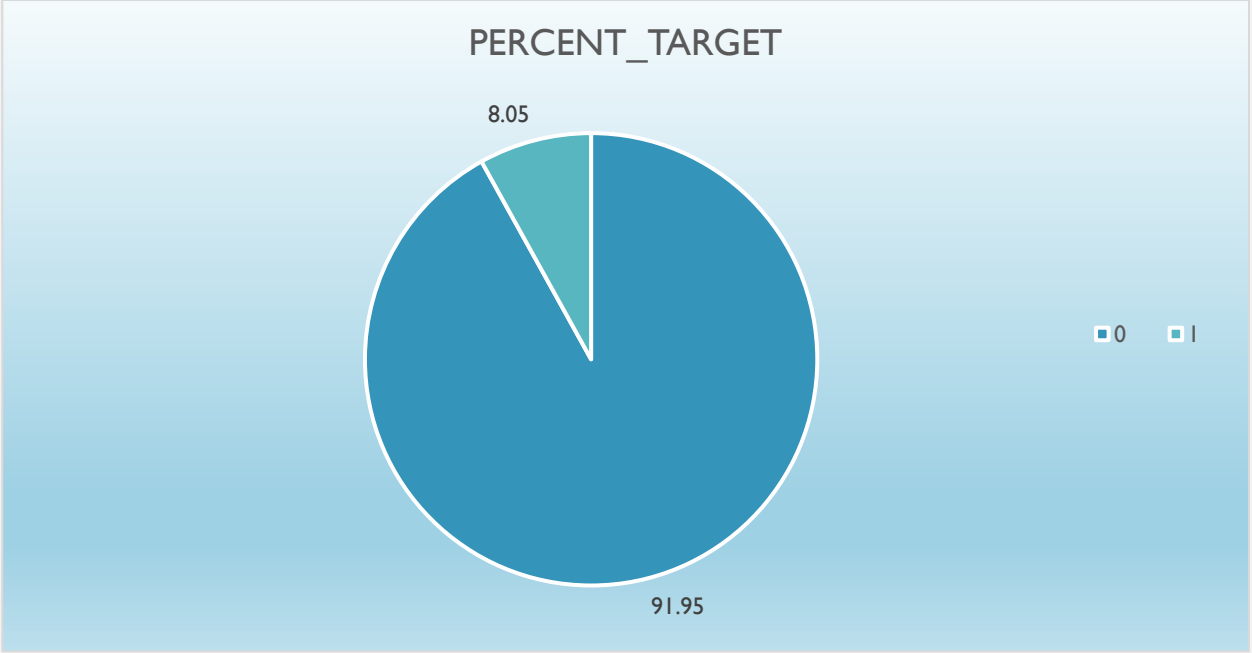

OBS_30_CNT_SOCIAL_CIRCLE
30
25
20
15
10
5
0
1


OBS_60_CNT_SOCIAL_CIRCLE
30
25
20
15
10
5
0
1

TASK C: ANALYZE DATA IMBALANCE

- Determine if data imbalance exists using COUNTIF and SUM functions in Excel.
- Visualize the distribution of the target variable with pie charts or bar charts.
- The data is significantly imbalanced for non-defaulters(Target = 0).
- I have created a bar chart to clearly illustrate this imbalance.

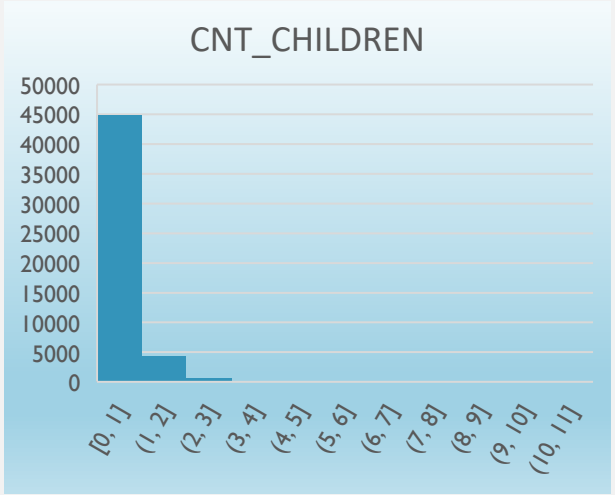
TARGET	COUNT_TARGET	PERCENT_TARGET
0	45973	91.94783896
1	4026	8.052161043



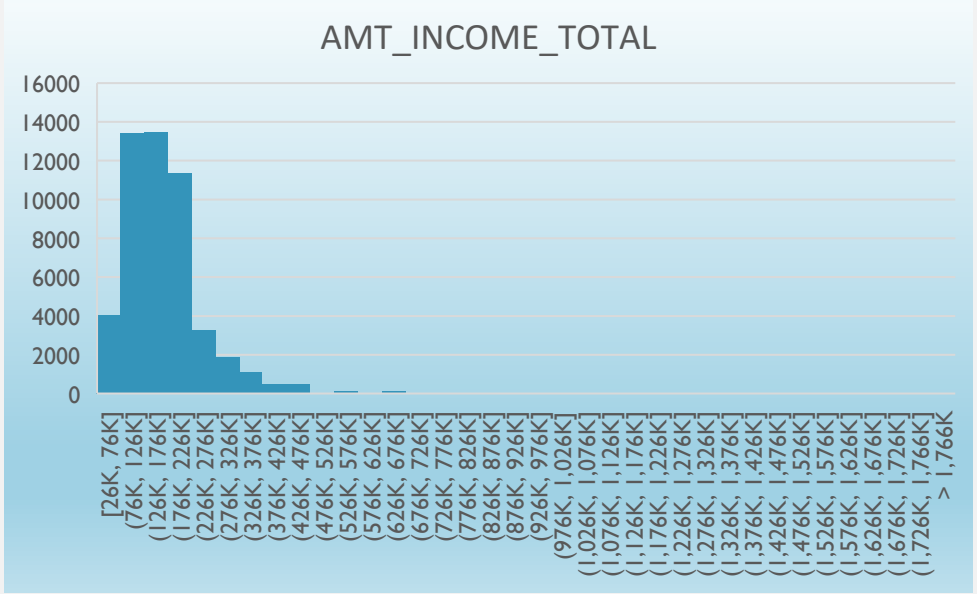
TASK D: UNIVARIATE AND BIVARIATE ANALYSIS

- Perform univariate analysis to understand individual variable distributions.
- Conduct bivariate analysis to explore relationships between variables and the target variable.
- Visualize the relationships using histograms, bar charts, scatter plots, and heatmaps.

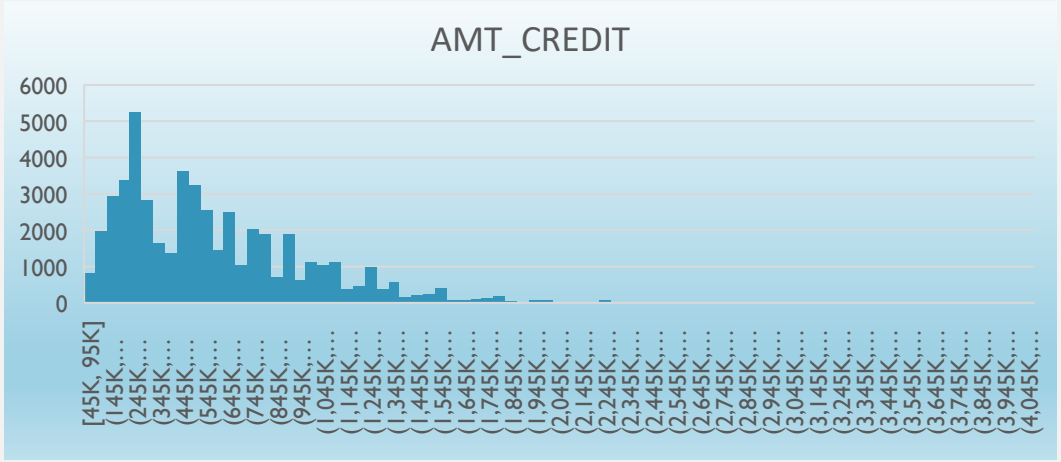
Coloumns	AVERAGE	MEDIAN	STD DV	MINIMUM	MAXIMUM
CNT_CHILDREN	0.419848397	0	0.724031307	0	11
AMT_INCOME_TOTAL	170767.5905	145800	531813.7768	25650	117000000
AMT_CREDIT	599700.5815	514777.5	402411.4096	45000	4050000
AMT_ANNUITY	27107.33399	24939	14562.6564	2052	258025.5
AMT_GOODS_PRICE	538992.3491	450000	369717.1252	45000	4050000
REGION_RATING_CLIENT	2.051661033	2	0.507972786	1	3
YEAR_BIRTH	43.8960057	43.09863014	11.94892234	21.04109589	68.99726027



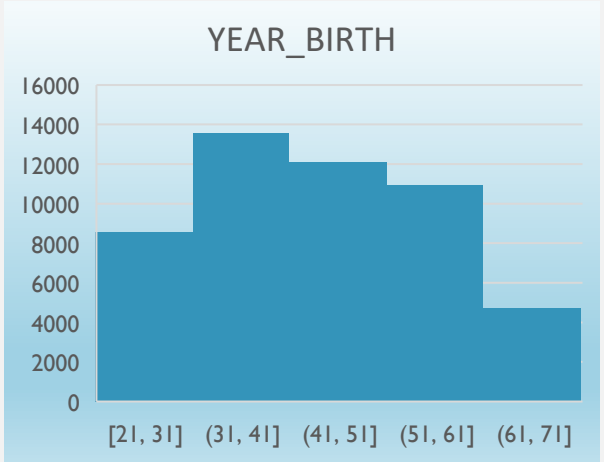
Most people who take out loans have either no children or just one child.



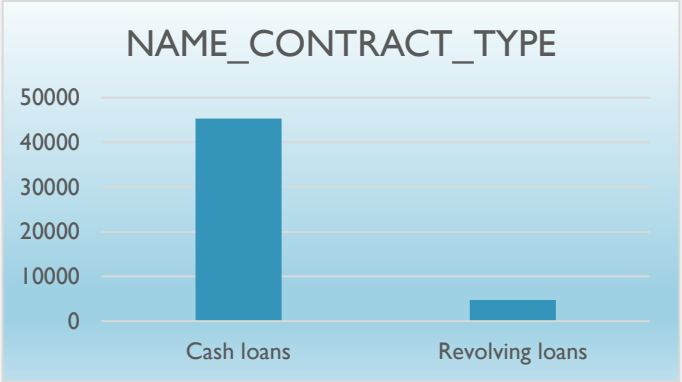
Individuals with incomes between 70k and 2.5 lakh tend to take out more loans, whereas those with higher incomes may require fewer loans.

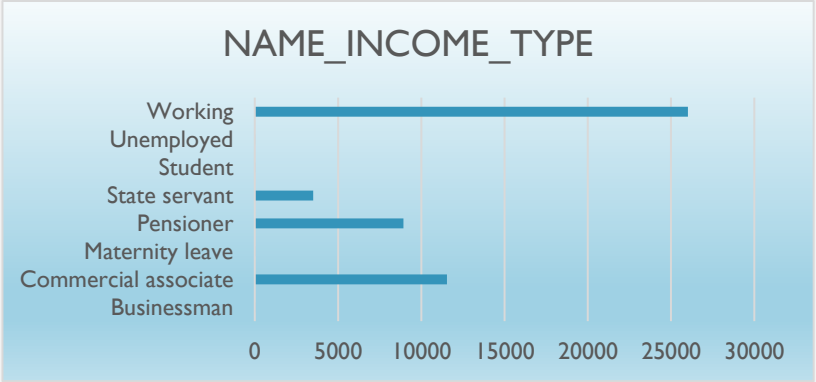


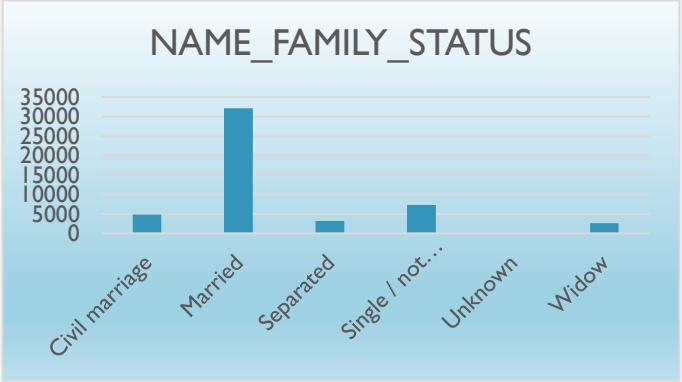
Most people require loans ranging from 250,000 to 300,000 rupees.

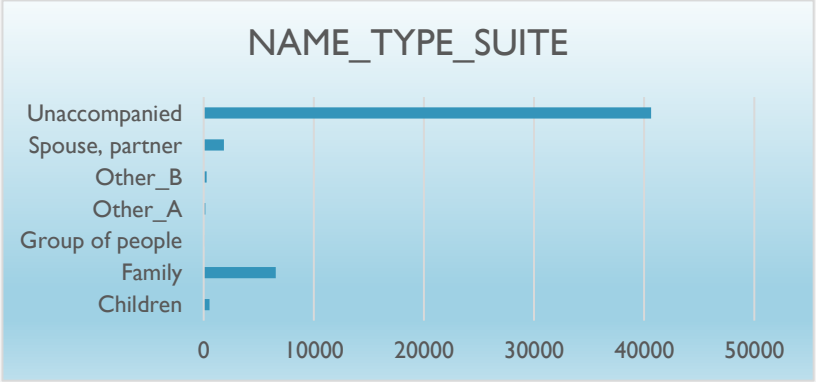


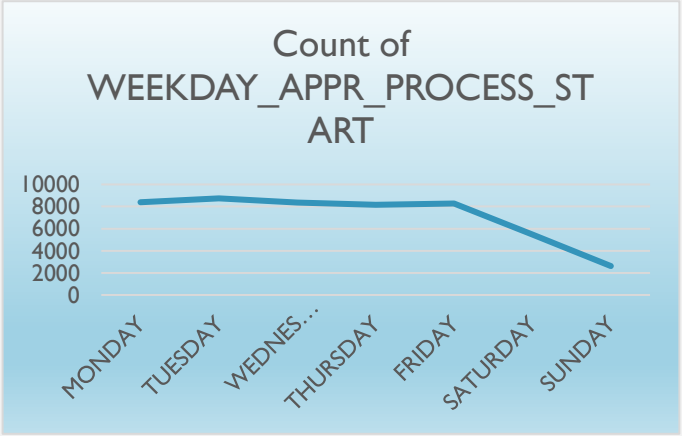
People aged 31 to 41, along with those in middle age, are more likely to take out loans.

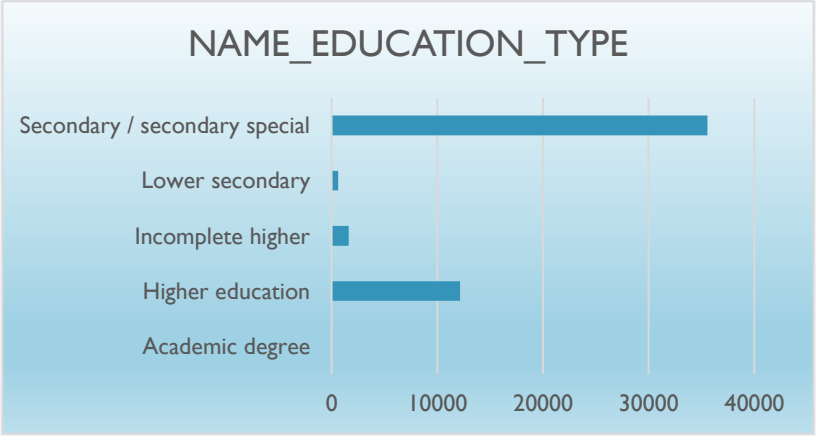

NAME_CONTRACT_TYPE
50000
40000
30000
20000
10000
0
Cash loans
Revolving loans

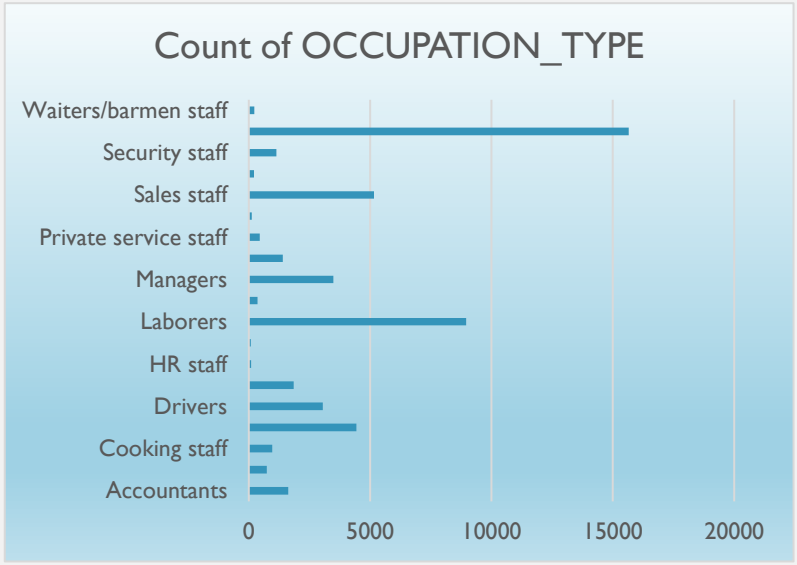

NAME_INCOME_TYPE
Working
Unemployed
Student
State servant
Pensioner
Maternity leave
Commercial associate
Businessman
0
5000
10000
15000
20000
25000
30000

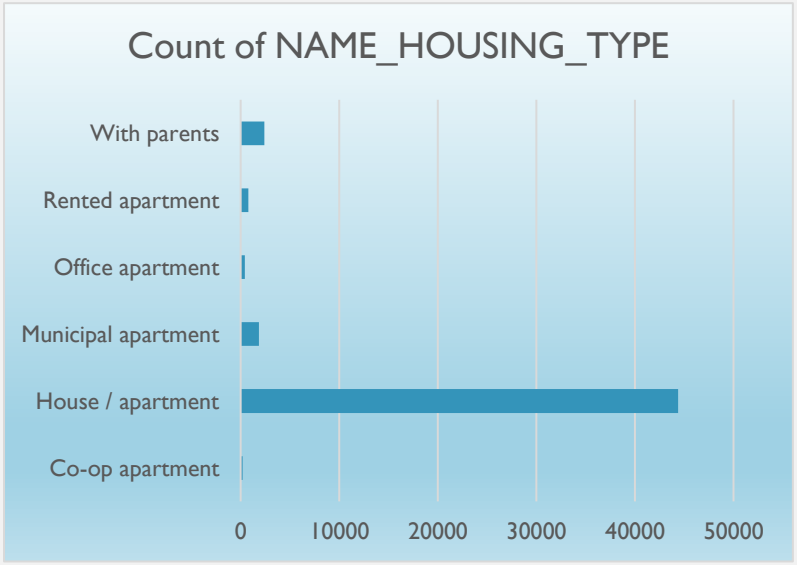

NAME_FAMILY_STATUS
35000
30000
25000
20000
15000
10000
5000
0
Civil marriage
Married
Separated
Single / not...
Unknown
Widow

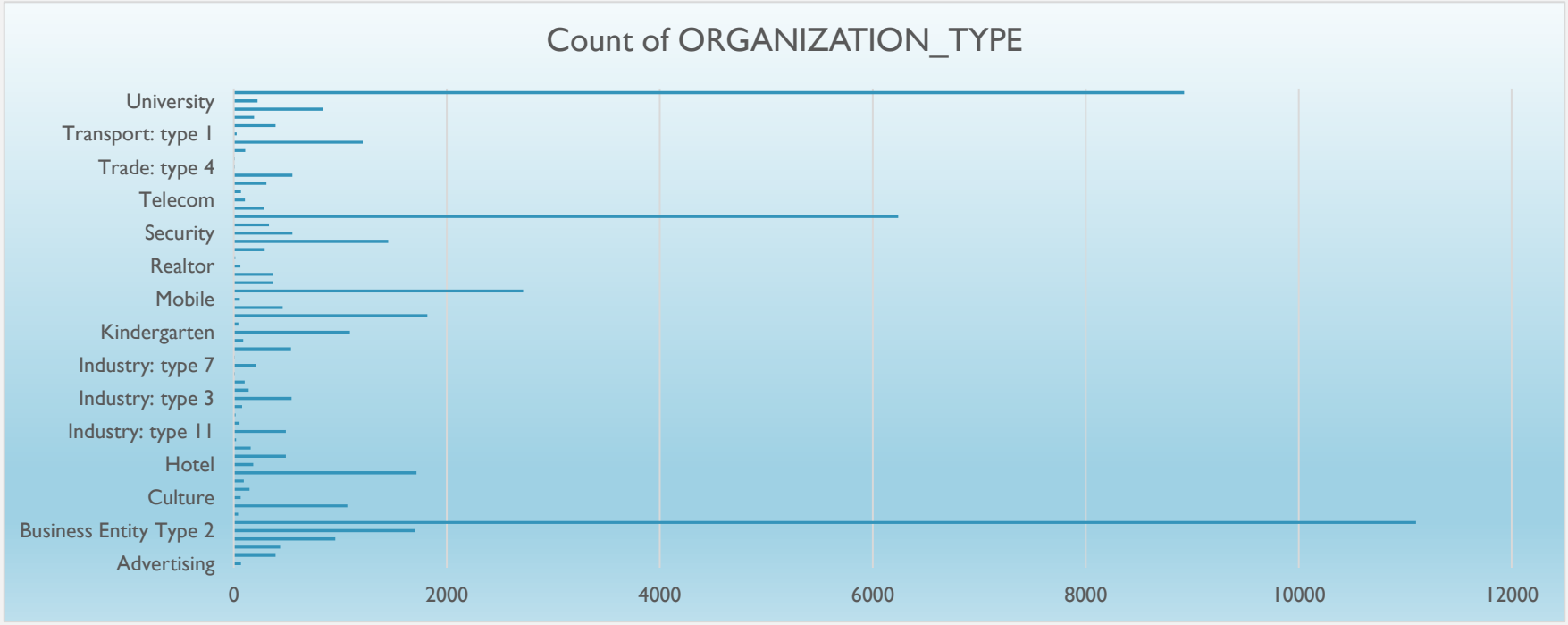

NAME_TYPE_SUITE
Unaccompanied
Spouse, partner
Other_B
Other_A
Group of people
Family
Children
0
10000
20000
30000
40000
50000

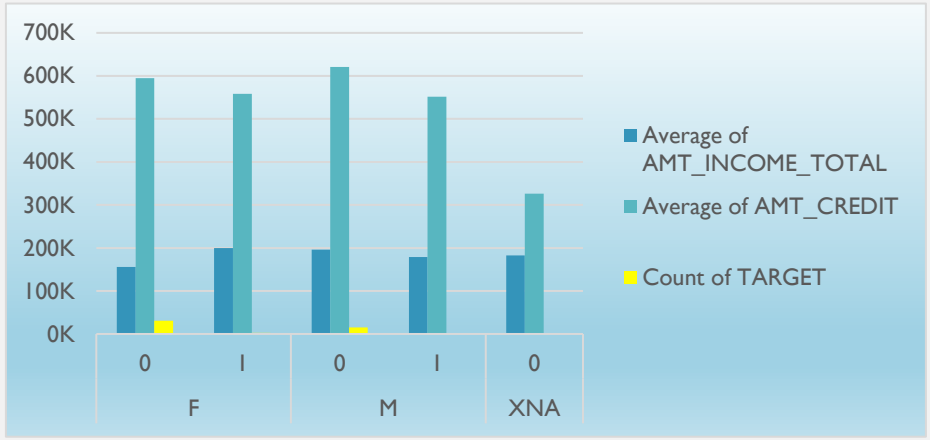

Count of WEEKDAY_APPR_PROCESS_START
10000
8000
6000
4000
2000
0
MONDAY
TUESDAY
WEDNES...
THURSDAY
FRIDAY
SATURDAY
SUNDAY

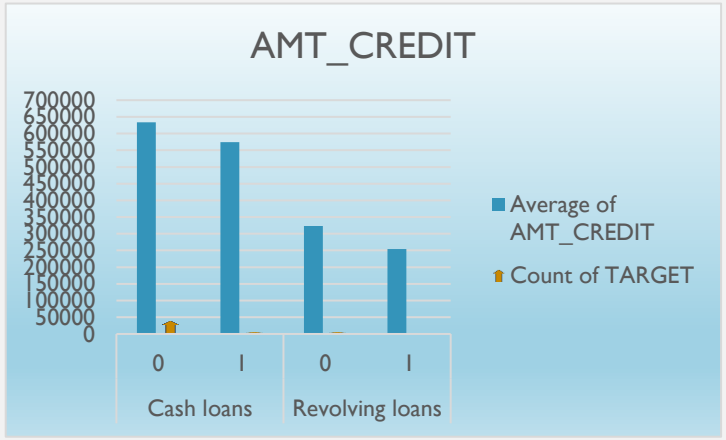

NAME_EDUCATION_TYPE
Secondary / secondary special
Lower secondary
Incomplete higher
Higher education
Academic degree
0
10000
20000
30000
40000

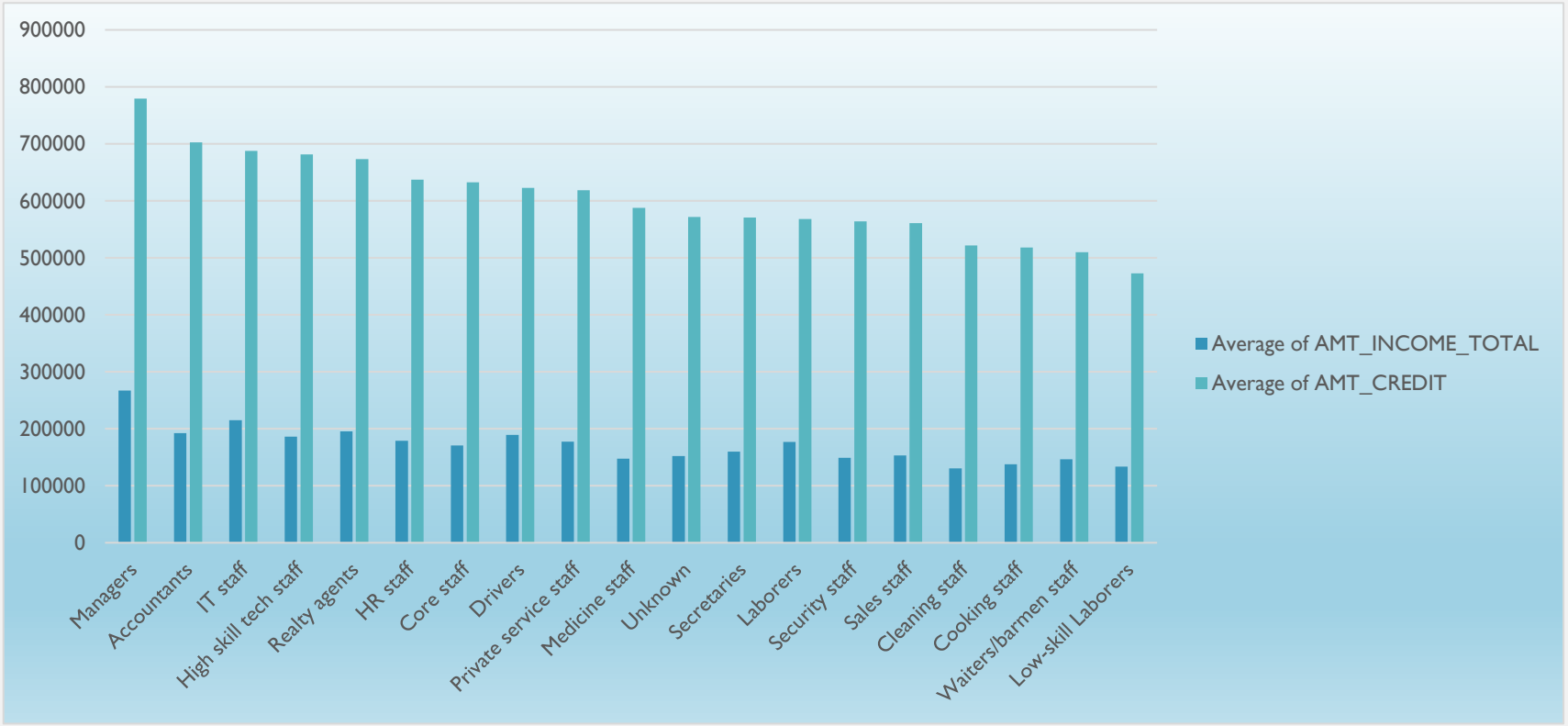

Count of OCCUPATION_TYPE
Waiters/barmen staff
Security staff
Sales staff
Private service staff
Managers
Laborers
HR staff
Drivers
Cooking staff
Accountants
0
5000
10000
15000
20000

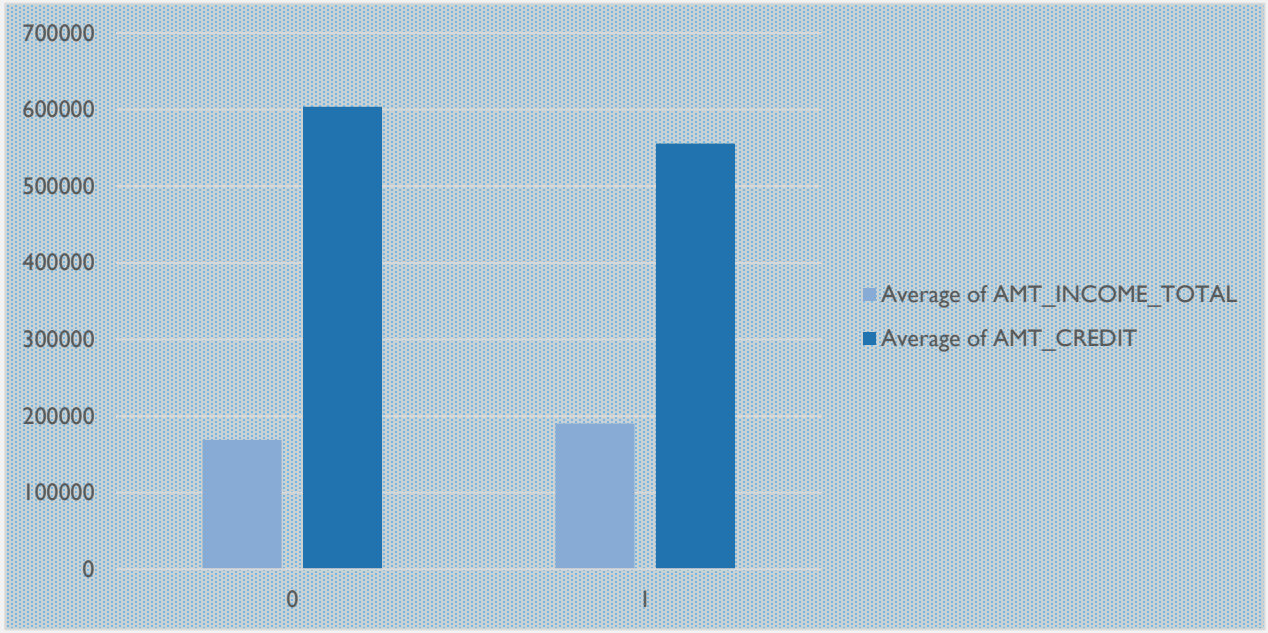

Count of NAME_HOUSING_TYPE
With parents
Rented apartment
Office apartment
Municipal apartment
House / apartment
Co-op apartment
0
10000
20000
30000
40000
50000

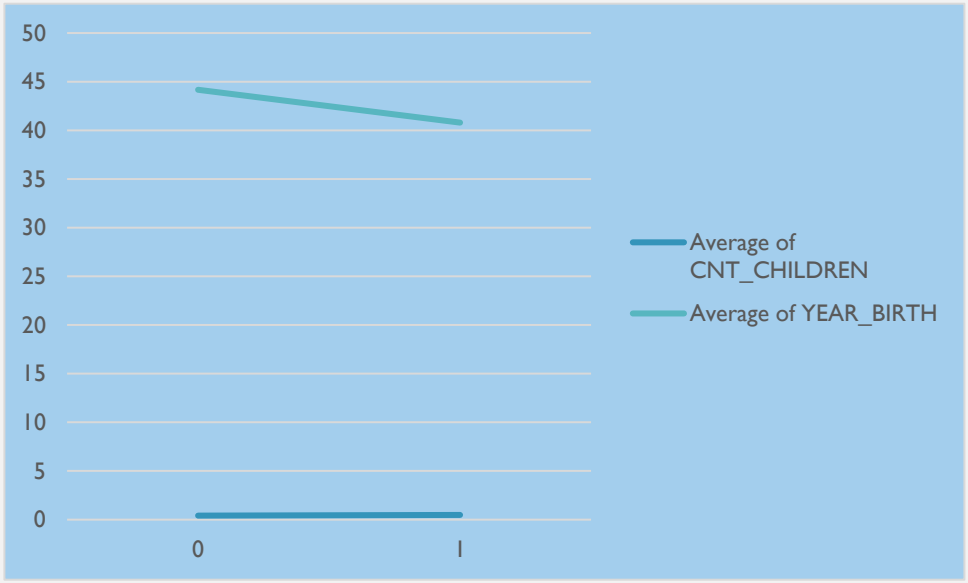

Count of ORGANIZATION_TYPE
University
Transport: type 1
Trade: type 4
Telecom
Security
Realtor
Mobile
Kindergarten
Industry: type 7
Industry: type 3
Industry: type 11
Hotel
Culture
Business Entity Type 2
Advertising
0
2000
4000
6000
8000
10000
12000

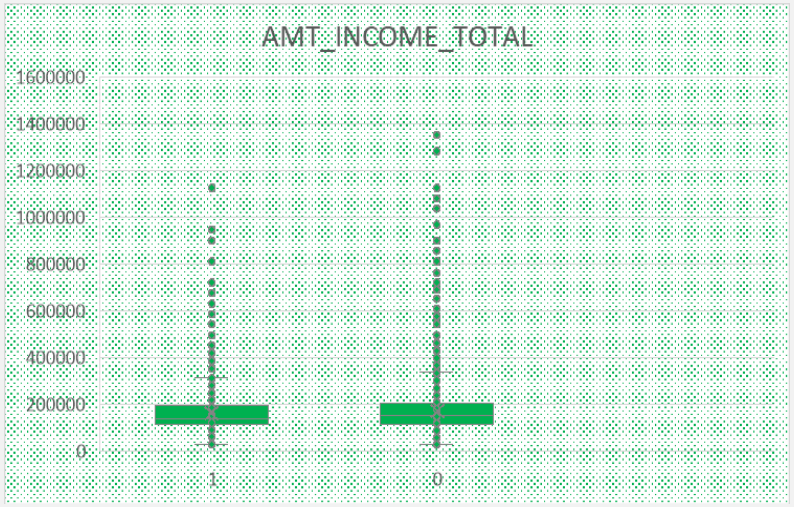

700K
600K
500K
400K
300K
200K
100K
0K
0
1
0
1
0
F
M
XNA
Average of AMT_INCOME_TOTAL
Average of AMT_CREDIT
Count of TARGET

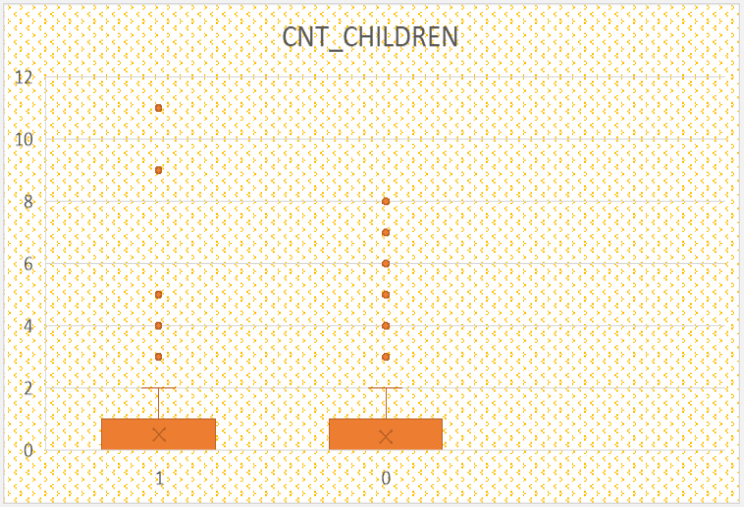

AMT_CREDIT
700000
650000
600000
550000
500000
450000
400000
350000
300000
250000
200000
150000
100000
50000
0
0
1
0
1
Cash loans
Revolving loans
Average of AMT_CREDIT
Count of TARGET

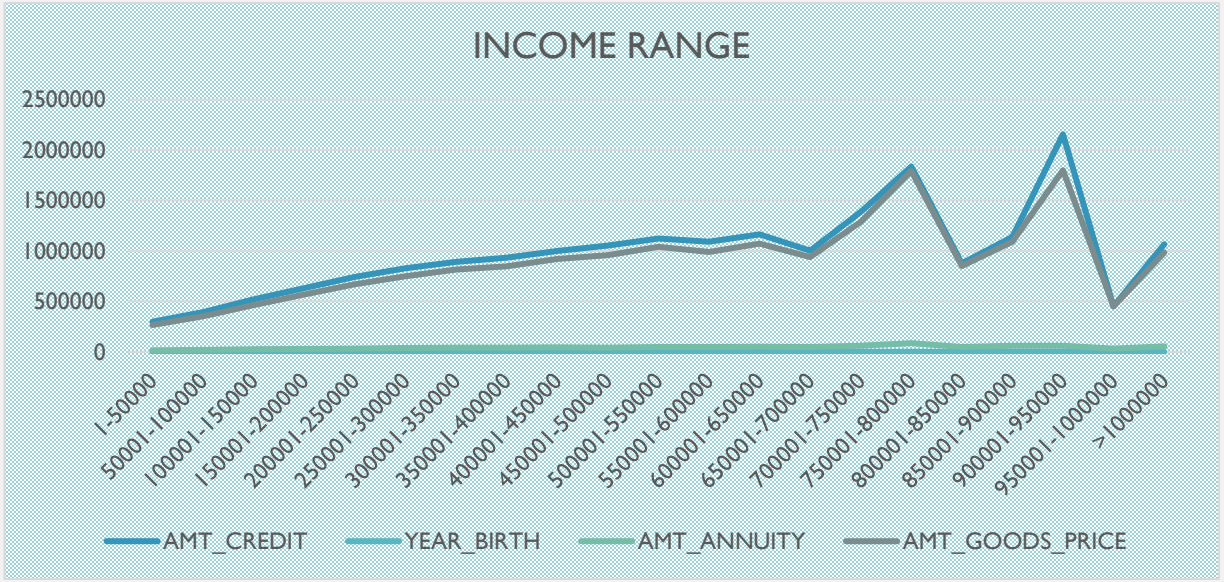

900000
800000
700000
600000
500000
400000
300000
200000
100000
0
Managers
Accountants
IT staff
High skill tech staff
Realty agents
HR staff
Core staff
Drivers
Private service staff
Medicine staff
Unknown
Secretaries
Laborers
Security staff
Sales staff
Cleaning staff
Cooking staff
Waiters/barmen staff
Low-skill Laborers
Average of AMT_INCOME_TOTAL
Average of AMT_CREDIT


700000
600000
500000
400000
300000
200000
100000
0
0
1
Average of AMT_INCOME_TOTAL
Average of AMT_CREDIT


50
45
40
35
30
25
20
15
10
5
0
0
1
Average of CNT_CHILDREN
Average of YEAR_BIRTH


AMT_INCOME_TOTAL
1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
1
0


CNT_CHILDREN
12
10
8
6
4
2
0
1
0


INCOME RANGE
2500000
2000000
1500000
1000000
500000
0
1-50000
50001-100000
100001-150000
150001-200000
200001-250000
250001-300000
300001-350000
350001-400000
400001-450000
450001-500000
500001-550000
550001-600000
600001-650000
650001-700000
700001-750000
750001-800000
800001-850000
850001-900000
900001-950000
950001-1000000
>1000000
AMT_CREDIT
YEAR_BIRTH
AMT_ANNUITY
AMT_GOODS_PRICE

TASK E: IDENTIFY TOP CORRELATIONS

- Created correlation matrices for the entire dataset, as well as for defaulters and non-defaulters.
- Organized the data into tables for improved clarity.
- Used deep red to highlight correlations greater than 0.5 and light red for correlations between 0 and 0.5.
- Summarized the highest correlations in a table of key insights.
- Separated the analysis for defaulters and non-defaulters to identify specific patterns within each group.

COMBINED CORRELATION

Variable 1	Variable 2	CORRELATION
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998332863
AMT_GOODS_PRICE	AMT_CREDIT	0.986704386
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950710179
CNT_FAM_MEMBERS	CNT_CHILDREN	0.880453292
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.857141677
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.856262392
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.821583789
AMT_GOODS_PRICE	AMT_ANNUITY	0.774134041
FLAG_EMP_PHONE	DAYS_BIRTH	0.617702887
DAYS_EMPLOYED	FLAG_DOCUMENT_6	0.591672763

INSIGHTS AND RESULTS

- The dataset needed to be cleaned for empty cells, but it was free of duplicate rows.
- Many important columns contained outliers, which significantly affected the average.
- The data was highly imbalanced with respect to non-defaulters.
- During the univariate and bivariate analysis, we created numerous graphs that provided valuable insights into how individuals' loan-seeking behavior depended on age, gender, occupation, income, and education.
- We also discovered the correlation between many variables.

LINKS

- Excel
- Video Presentaion