Financial Risk Analytics: Loan Default Prediction

Project Description

This project aimed to analyze customer application data to identify key factors influencing loan default risk. The objective was to apply exploratory data analysis (EDA) techniques to uncover trends, detect data quality issues, and generate insights that can inform risk-based decision-making in the banking sector.

Approach

The project followed a structured approach involving data cleaning, outlier treatment, univariate and bivariate analysis, and correlation analysis across customer segments.

Tech Stack Used

- **Microsoft Excel 2019**
  Utilized for all data cleaning, outlier handling, exploratory data analysis, visualizations (histograms, box plots), and correlation matrix calculations.

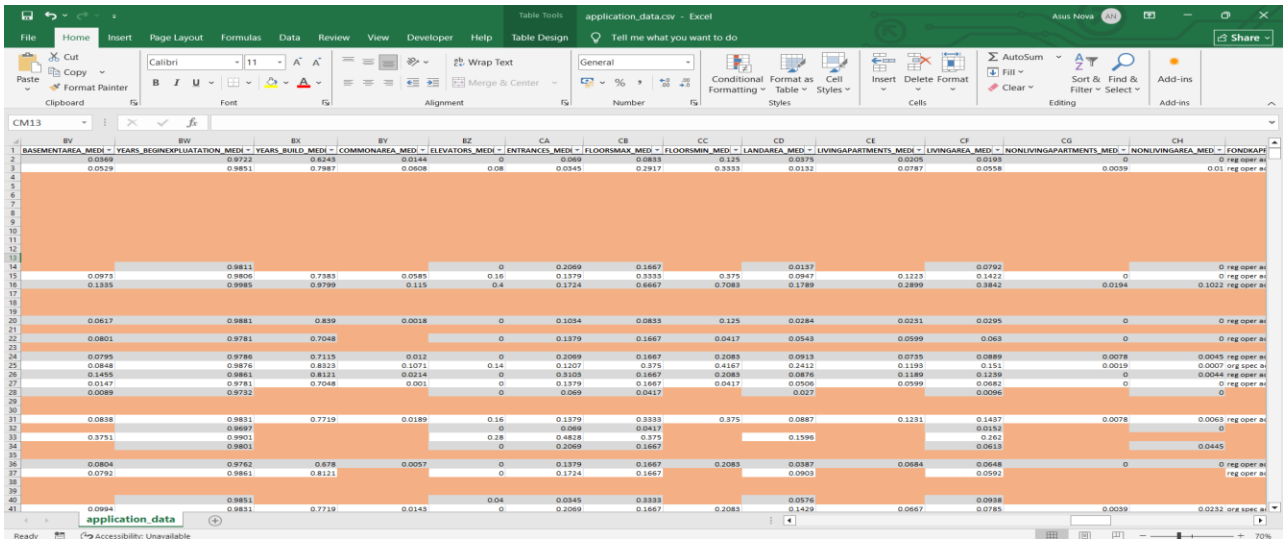
Key Insights

- Higher income levels are associated with lower default rates.
- Employment stability (longer duration) correlates with reduced default risk.
- Younger applicants, particularly those aged 27–35, have higher default rates.
- Manual labour occupations show higher default rates than white-collar jobs.
- Higher education levels correspond to lower default risk.
- Most previous loans had moderate monthly payments and short durations.
- Strong positive correlations exist between credit amount and goods price, as well as between credit amount and annuity.

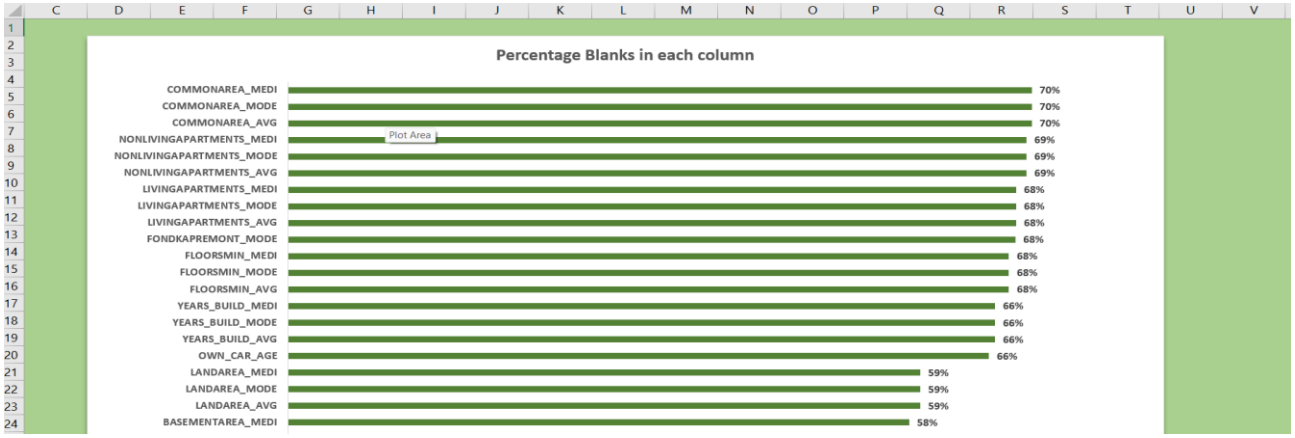
Tasks

A. Handling Missing Values

- Highlighted blanks with conditional formatting



a. Column Removal



- **Action:** Removed columns with more than 40% missing values.
- **Purpose:** Improved dataset reliability by focusing on complete records.

b. Handling Missing Values

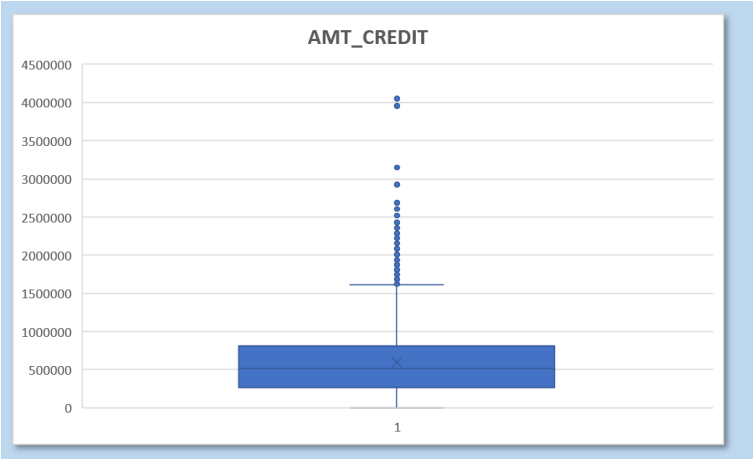
Blanks	0	0	0	0	0
Blanks %	0.0%	0.0%	0.0%	0.0%	0.0%
median	514777.5	514777.5	24939	24939	
mode					Unaccompanied

- **Numerical Columns:** Replaced missing values with the median to reduce distortion.
- **Categorical Columns:** Filled missing values with the mode to preserve distribution.

B. Outlier Detection and Treatment

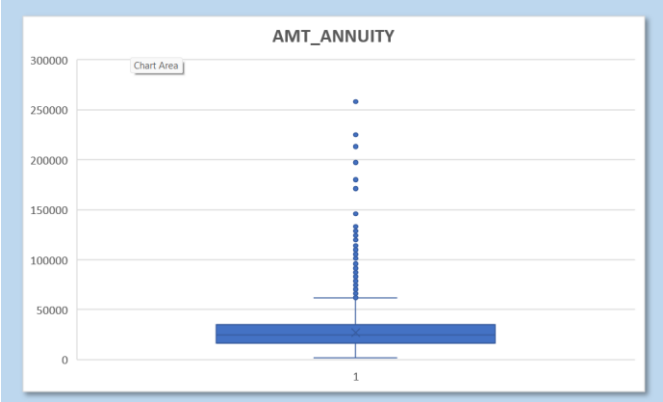
Outliers were capped using the Interquartile Range (IQR) method:

- **AMT_CREDIT:**



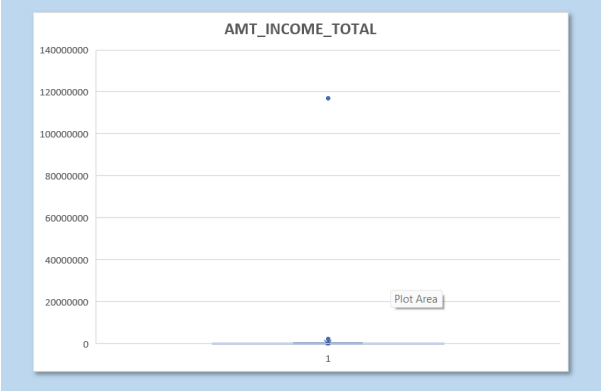
Controlled large loan values to maintain data consistency.

- **AMT_ANNUITY:**



Reduced influence of unusually high annuities.

- **AMT_INCOME_TOTAL:**



Capped to reduce extreme upper values.

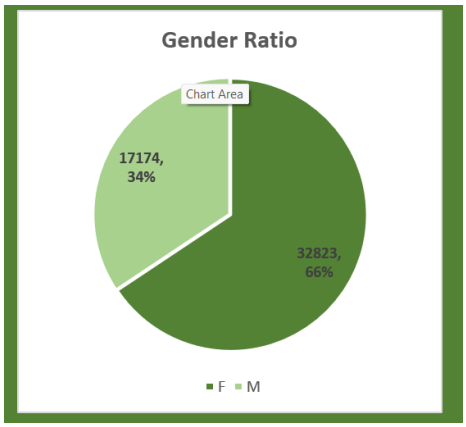
Purpose: Prevent skewed results while preserving overall data structure.

C. Data Imbalance Analysis

a. TARGET (Loan Default Status):

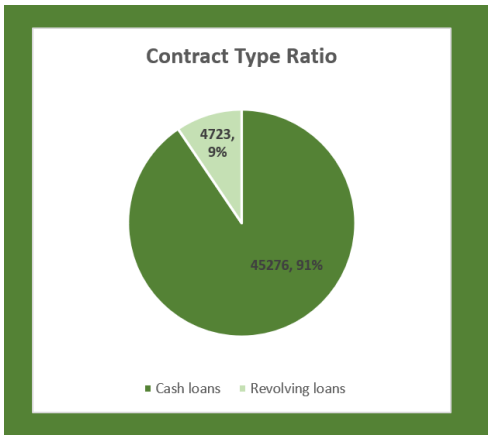
- **Non-default (0):** 91.95% (45,973 cases)
- **Default (1):** 8.05% (4,026 cases)
- **Insight:** Highly imbalanced (~11.4:1), which may affect model performance.

b. CODE_GENDER:



- **Female:** 65.65%
- **Male:** 34.35%
- **Insight:** Female applicants dominate the dataset (1.91:1 ratio).

c. NAME_CONTRACT_TYPE:



- **Cash Loans:** 90.55%
- **Revolving Loans:** 9.45%
- **Insight:** Cash loans are significantly more common.

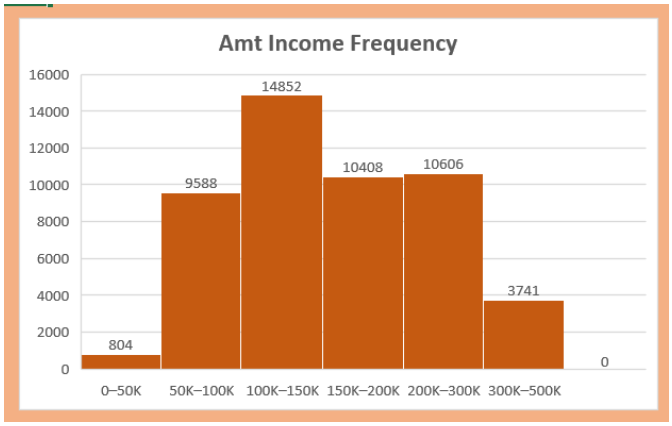
D. Univariate, Segmented Univariate, and Bivariate Analysis

Univariate Analysis

Analyzed key numeric and categorical variables:

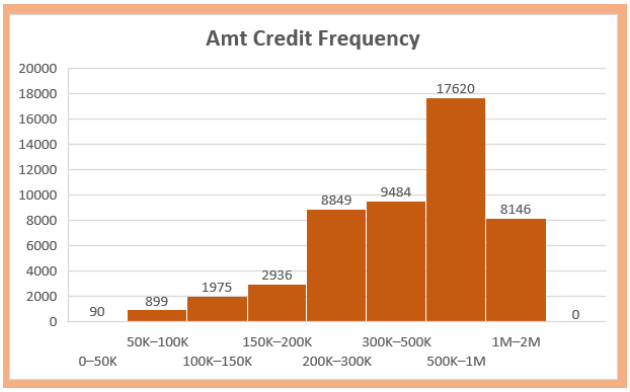
- **AMT_INCOME_TOTAL (Capped):**
  Right-skewed. Common income ≈ 135,000. Capped at 337,500.

Class	Frequency
0–50K	804
50K–100K	9588
100K–150K	14852
150K–200K	10408
200K–300K	10606
300K–500K	3741



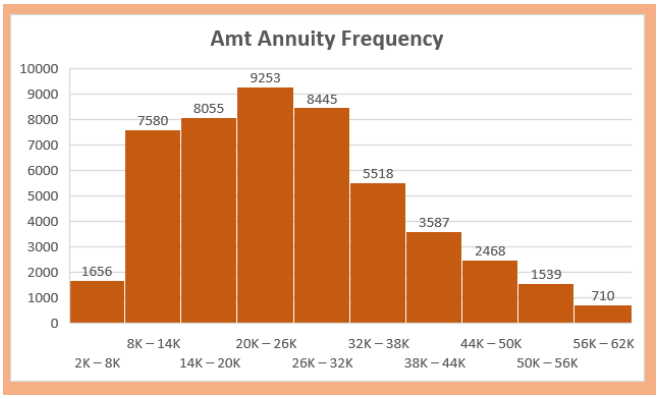
- **AMT_CREDIT (Capped):**
  Broad range. Common value ≈ 450,000. Capped at 2.7 million.

Class	Frequency
0–50K	90
50K–100K	899
100K–150K	1975
150K–200K	2936
200K–300K	8849
300K–500K	9484
500K–1M	17620
1M–2M	8146



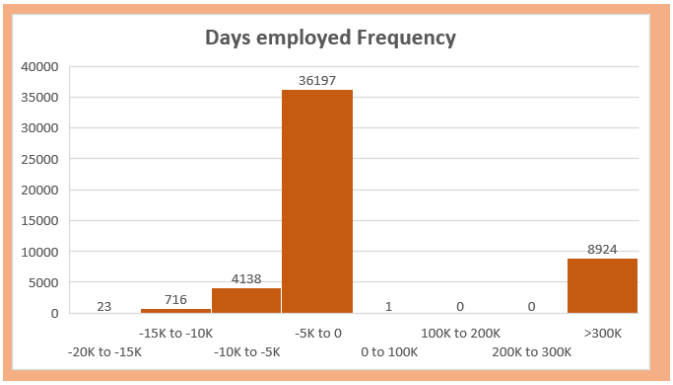
- **AMT_ANNUITY (Capped):**
  Concentrated around 28,000–35,000. Outliers capped.

Class Interval (₹)	Frequency
2K – 8K	1656
8K – 14K	7580
14K – 20K	8055
20K – 26K	9253
26K – 32K	8445
32K – 38K	5518
38K – 44K	3587
44K – 50K	2468
50K – 56K	1539
56K – 62K	710



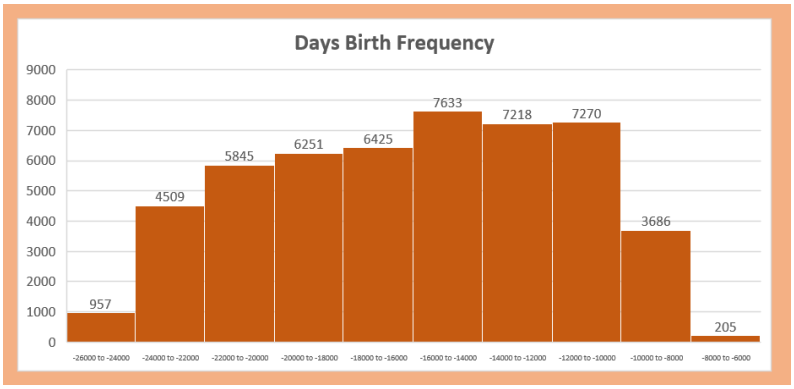
- **DAYS_EMPLOYED (Capped):**
  Most values between 3–5 years. Extreme value (365,243) treated separately (likely retirees).

Statistics	Values
Count	49999
Mean	63219
Median	-1221
Mode	365243
Minimum	-17531
Maximum	365243
Standard Deviation	140793



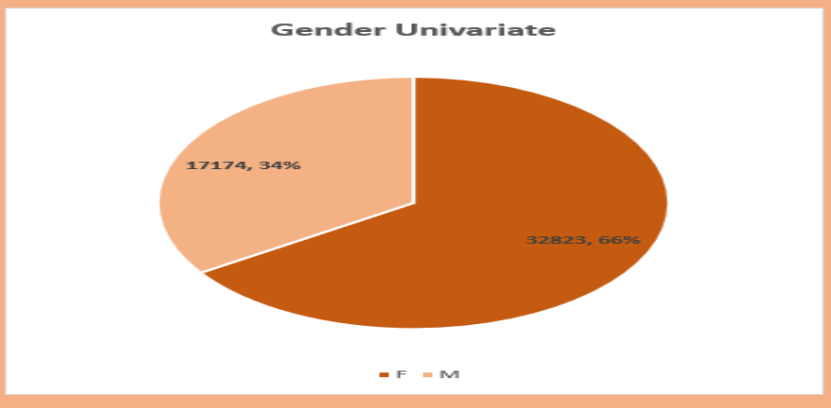
- **DAYS_BIRTH:**
  Average age ≈ 44(16022days) years. Majority between 20–60 years.

Statistics	Values
Count	49999
Mean	-16022
Median	-15731
Mode	-11039
Minimum	-25184
Maximum	-7680
Standard Deviation	4361



- **CODE_GENDER:**
  65.6% female. Non-binary or anomalous values are negligible (0.004%).

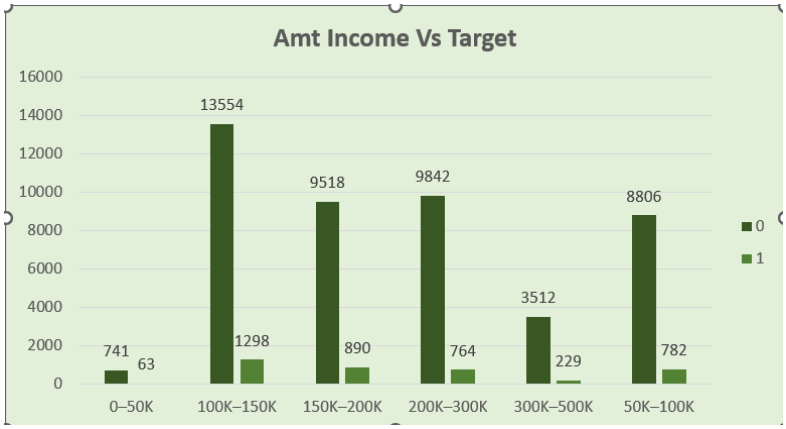
Gender	Count of CODE_GENDER	Percentage
F	32823	65.65%
M	17174	34.35%
XNA	2	0.00%



Segmented Univariate Analysis

a. Income vs. Default Rate

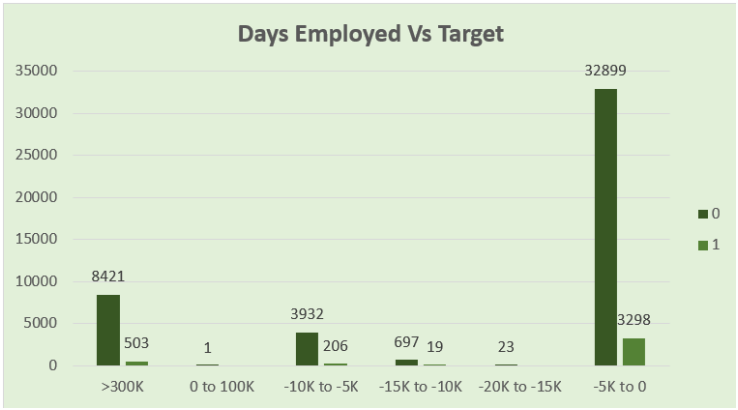
Class	0	1	Default rate
0–50K	741	63	7.84%
100K–150K	13554	1298	8.16%
150K–200K	9518	890	8.74%
200K–300K	9842	764	8.55%
300K–500K	3512	229	7.20%
50K–100K	8806	782	6.12%



- **Insight:** Higher income reduces risk.

b. Employment Duration vs. Default

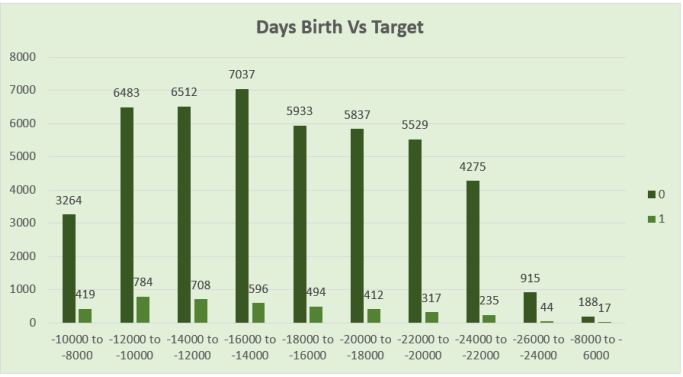
Class	0	1	Default Rate
>300K	8421	503	6%
0 to 100K	1		0%
-10K to -5K	3932	206	5%
-15K to -10K	697	19	3%
-20K to -15K	23		0%
-5K to 0	32899	3298	9%



- **Insight:** Longer employment history indicates lower default risk.

c. Age vs. Default Rate

Class	0	1	Default Rate
-10000 to -8000	3264	419	11%
-12000 to -10000	6483	784	11%
-14000 to -12000	6512	708	10%
-16000 to -14000	7037	596	8%
-18000 to -16000	5933	494	8%
-20000 to -18000	5837	412	7%
-22000 to -20000	5529	317	5%
-24000 to -22000	4275	235	5%
-26000 to -24000	915	44	5%
-8000 to -6000	188	17	8%

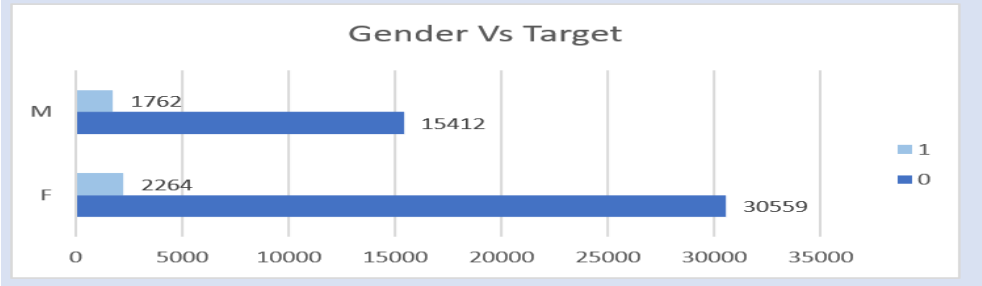


- **Insight:** Younger age correlates with higher default risk.

Bivariate Analysis

a. Gender vs. Default

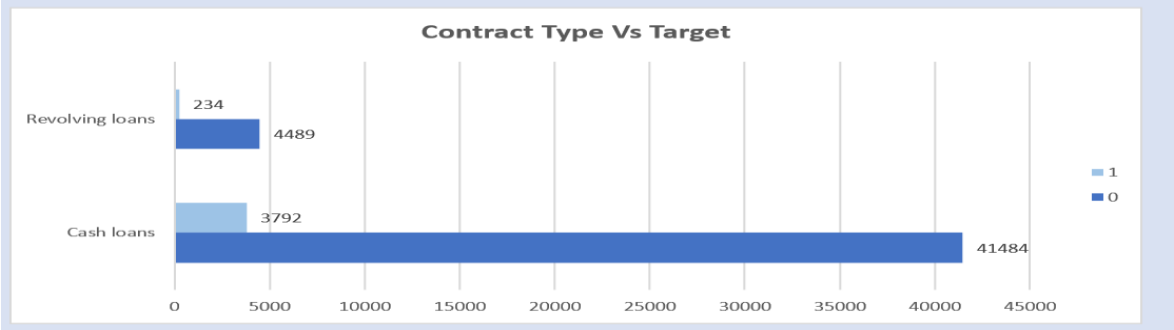
Genders	0	1	Default Rate
F	30559	2264	7%
M	15412	1762	10%



- Slightly higher default rate among males.
- **Insight:** Gender differences exist but are modest.

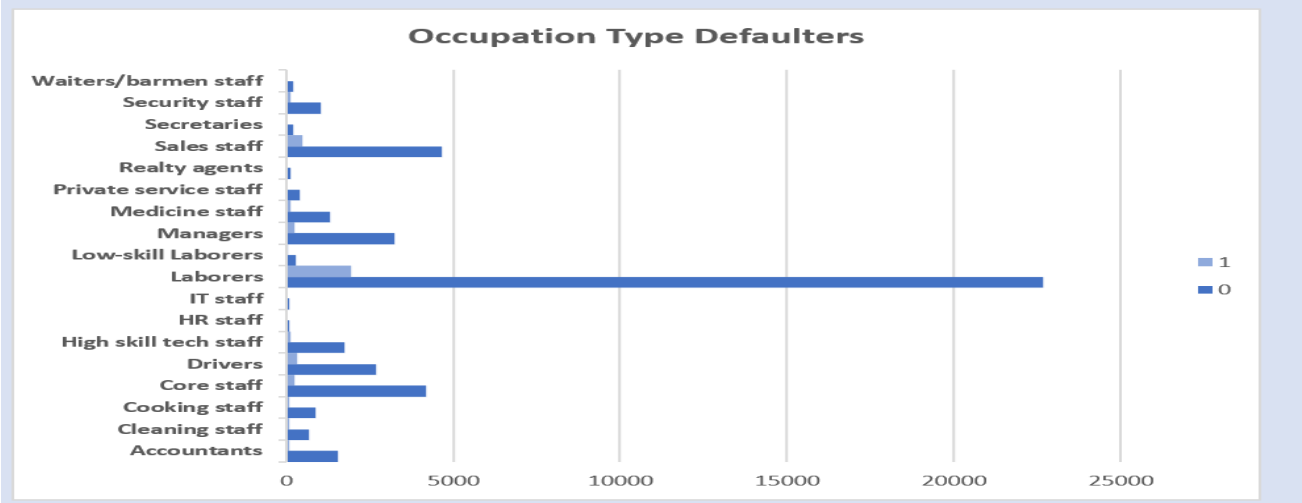
b. Contract Type vs. Default

Contract Type	0	1	Default Rate
Cash loans	41484	3792	8%
Revolving loans	4489	234	5%



- Cash loans show significantly higher default risk compared to revolving loans.

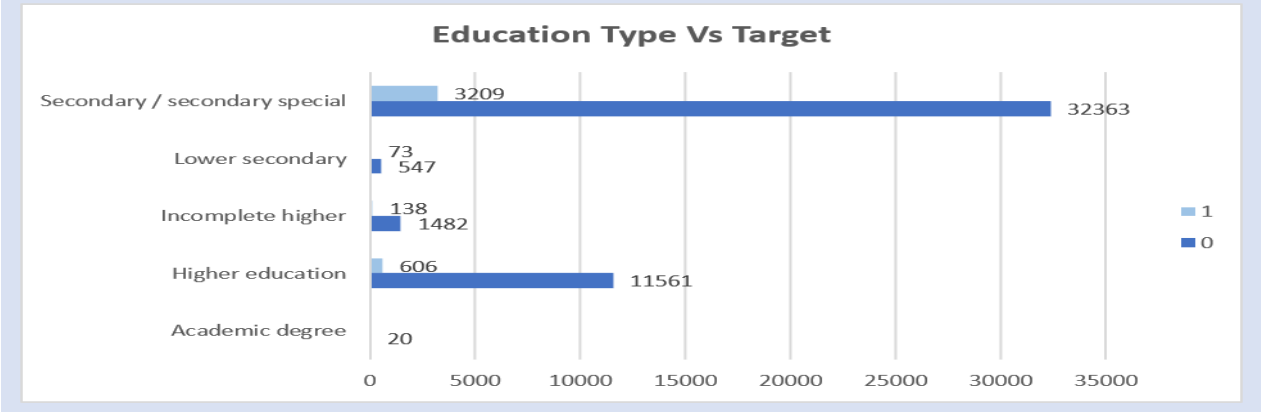
c. Occupation vs. Default



- Manual roles (e.g., Laborers, Sales) have higher default rates.
- White-collar roles (e.g., Managers, Accountants) show lower risk.

d. Education vs. Default

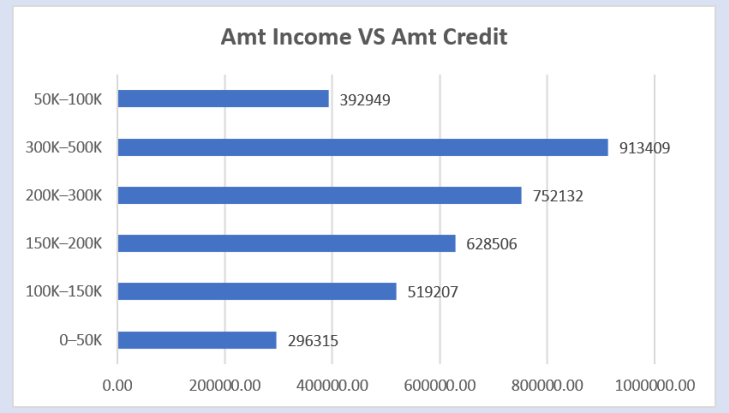
Education Type	0	1	Default Rate
Academic degree	20		0%
Higher education	11561	606	5%
Incomplete higher	1482	138	9%
Lower secondary	547	73	12%
Secondary / secondary special	32363	3209	9%



- Lower education (e.g., Secondary) linked to higher defaults.
- Higher education corresponds to lower risk.

e. Income vs. Credit

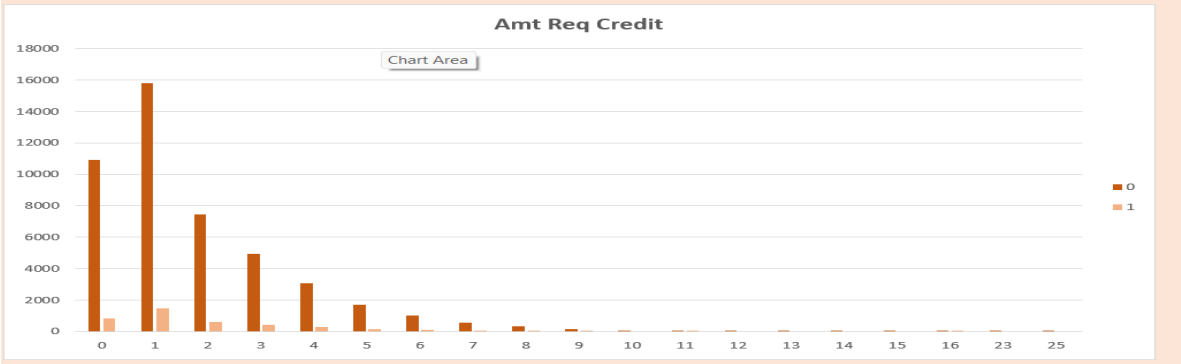
Income Range	Average Credit
0–50K	296314.92
100K–150K	519206.77
150K–200K	628505.72
200K–300K	752132.41
300K–500K	913408.65
50K–100K	392948.77



- Higher income generally results in higher credit amounts.
- Some low-income individuals take large loans, indicating possible overextension.

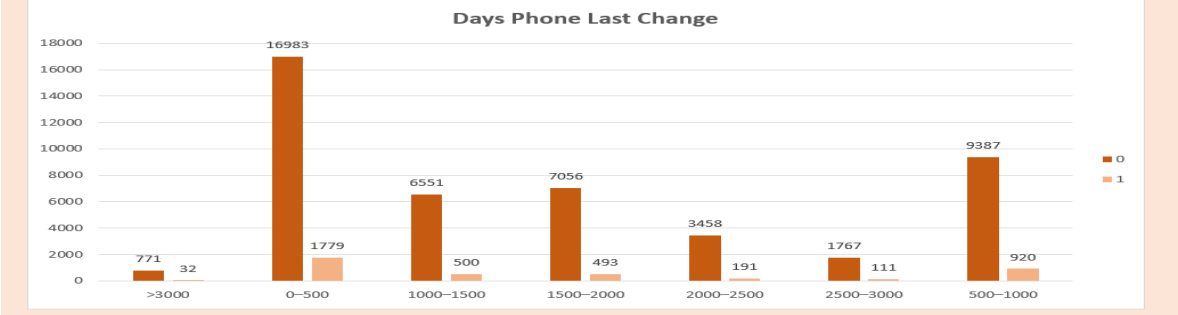
Payment Difficulty Analysis

a. Credit Bureau Inquiries & Phone Records



- Frequent credit inquiries and outdated contact information are linked to higher default rates.

b. Behavioural Patterns



- Stable phone contact (updated within 1,000 days) correlates with lower default probability.

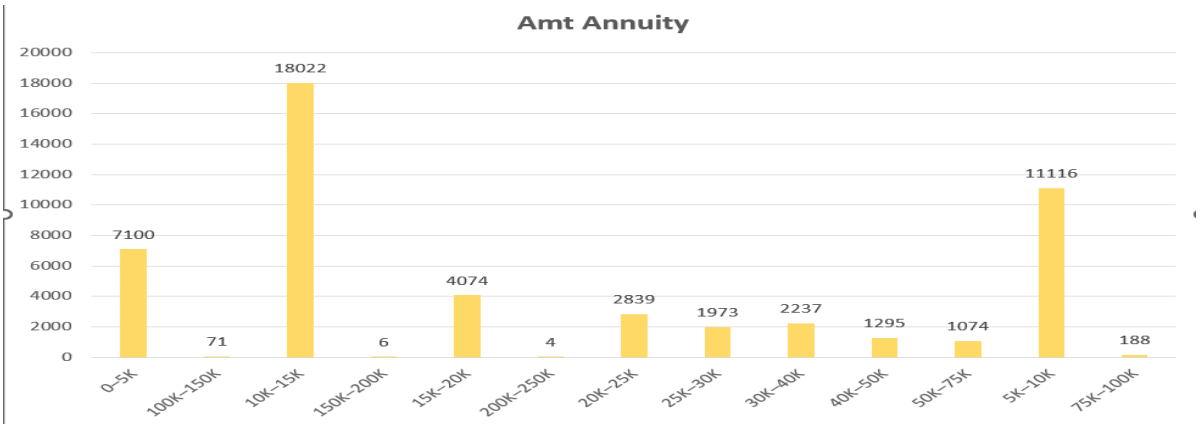
Insight: Consistent contact information and fewer credit checks suggest responsible behavior.

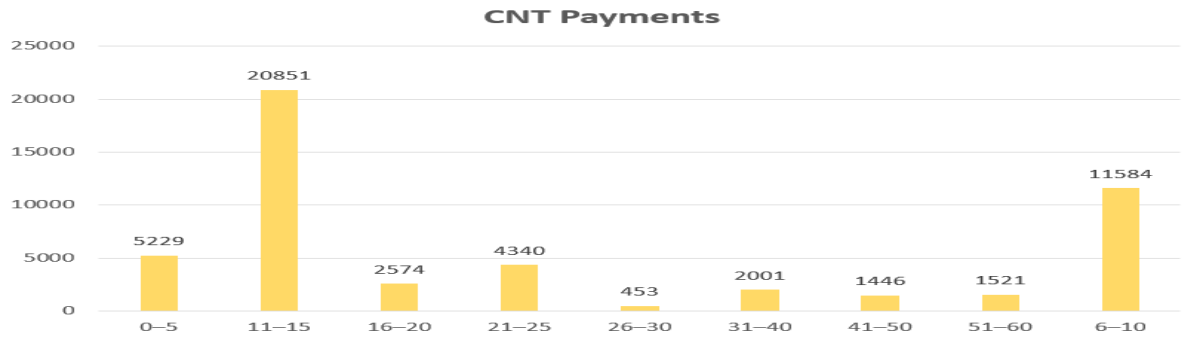
Previous Application Analysis

- ❖ **Univariate Analysis**

Statistic	Values Annuity
Minimum	0
Maximum	234478.395
Mean	14507.54628
Median	10879.92
Standard Dev	13036.5375
Count	49999

Statistic	Values Term
Minimum	0
Maximum	60
Mean	14.80259605
Median	12
Standard Dev	12.5003519
Count	49999





- Median annuity: ~10,900
- Median term: ~12 months
- **Insight:** Most previous loans were short-term and of moderate value.

❖ **Segmented Analysis**

Contract Status	Average Of Amt Annuity	Count Of Amt Annuity
Approved	14402.58025	31885
Canceled	11032.65919	8595
Refused	18704.30367	8660
Unused offer	10863.28268	859

Contract Status	Average of CNT_PAYMENT	Count of CNT_PAYMENT
Approved	13.8911714	31885
Canceled	12.18219895	8595
Refused	21.04399538	8660
Unused offer	11.93015134	859

Product Combination	Approved	Canceled	Refused	Unused offer	Grand Total
Card Street	1189	1331	803		3323
Card X-Sell	1648	6	648		2302
Cash		7135	787	17	7939
Cash Street: high	1139	1	612		1752
Cash Street: low	280	29	747		1056
Cash Street: middle	490	10	460		960
Cash X-Sell: high	1332	1	324		1657
Cash X-Sell: low	2432	15	1092		3539
Cash X-Sell: middle	3226	8	719		3953
POS household with interest	7478	16	910	114	8518
POS household without interest	2450		349		2799
POS industry with interest	2946		72	10	3231
POS industry without interest	364		26		390
POS mobile with interest	5530	13	792	694	7029
POS mobile without interest	661		47	23	731
POS other with interest	633	27	67	1	728
POS others without interest	87		5		92

Chart Area

- Most previous loans were approved.
- Common loan types: Cash loans, Mobile loans.
- **Insight:** Behavior in previous applications mirrors current trends.

E. Correlation Analysis

Correlations for nine key numerical variables were analyzed separately for defaulters and non-defaulters.

a. Non-Defaulters (TARGET = 0)

Correlations (Non Defaulters)	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPU	DAYS_BIRTH	DAYS_EMPLOYED	REGION_RATING_CLIENT
CNT_CHILDREN	1.000	0.038	0.006	0.030	0.002	-0.025	0.336	-0.244	0.021
AMT_INCOME_TOTAL	0.038	1.000	0.419	0.494	0.424	0.185	0.090	-0.193	-0.223
AMT_CREDIT	0.006	0.419	1.000	0.795	0.975	0.092	-0.053	-0.079	-0.099
AMT_ANNUITY	0.030	0.494	0.795	1.000	0.791	0.112	0.013	-0.120	-0.128
AMT_GOODS_PRICE	0.002	0.424	0.975	0.791	1.000	0.099	-0.049	-0.075	-0.105
REGION_POPULATION_RELATIVE	-0.025	0.185	0.092	0.112	0.099	1.000	-0.030	-0.007	-0.539
DAYS_BIRTH	0.336	0.090	-0.053	0.013	-0.049	-0.030	1.000	-0.615	0.009
DAYS_EMPLOYED	-0.244	-0.193	-0.079	-0.120	-0.075	-0.007	-0.615	1.000	0.041
REGION_RATING_CLIENT	0.021	-0.223	-0.099	-0.128	-0.105	-0.539	0.009	0.041	1.000

- **Strongest Positive:**
  - AMT_CREDIT vs. AMT_GOODS_PRICE: +0.97
  - AMT_CREDIT vs. AMT_ANNUITY: +0.79
- **Strongest Negative:**
  - DAYS_BIRTH vs. DAYS_EMPLOYED: –0.62

b. Defaulters (TARGET = 1)

Correlations (Defaulters)	CNT_CHILD	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPU	DAYS_BIRTH	DAYS_EMPLOYED	REGION_RATING_CLIENT
CNT_CHILDREN	1.000	-0.008	0.010	0.030	-0.001	-0.020	0.250	-0.189	0.056
AMT_INCOME_TOTAL	-0.008	1.000	0.349	0.425	0.348	0.108	0.016	-0.141	-0.164
AMT_CREDIT	0.010	0.349	1.000	0.755	0.974	0.069	-0.142	0.016	-0.046
AMT_ANNUITY	0.030	0.425	0.755	1.000	0.750	0.073	-0.007	-0.080	-0.065
AMT_GOODS_PRICE	-0.001	0.348	0.974	0.750	1.000	0.077	-0.141	0.020	-0.051
REGION_POPULATION_RELATIVE	-0.020	0.108	0.069	0.073	0.077	1.000	-0.016	0.008	-0.430
DAYS_BIRTH	0.250	0.016	-0.142	-0.007	-0.141	-0.016	1.000	-0.581	0.045
DAYS_EMPLOYED	-0.189	-0.141	0.016	-0.080	0.020	0.008	-0.581	1.000	-0.009
REGION_RATING_CLIENT	0.056	-0.164	-0.046	-0.065	-0.051	-0.430	0.045	-0.009	1.000

- **Strongest Positive:**
    - AMT_CREDIT vs. AMT_GOODS_PRICE: +0.97
    - AMT_CREDIT vs. AMT_ANNUITY: +0.76
- **Strongest Negative:**
    - DAYS_BIRTH vs. DAYS_EMPLOYED: –0.58

Insight: Correlation structures are similar between groups, though slightly weaker among defaulters, indicating mild behavioural variation.

Result

The project successfully conducted exploratory data analysis on both current and previous loan application datasets. Key risk indicators—such as income, age, employment history, loan types, and education—were identified. These findings enhance the understanding of loan default risk and can inform future development of risk-based lending models and predictive analytics solutions.

Drive Link

Main File (Application Data)

https://docs.google.com/spreadsheets/d/1_uW8t7NLwP9kHStcY3uJiq5HwVtDB8go/edit?usp=sharing&ouid=115912486417524737249&rtpof=true&sd=true

Secondary File for In-depth Analysis (Previous Data)

https://docs.google.com/spreadsheets/d/1mtig2AMKr0aO-vOJuIC2Qgxb8VKtnzwb/edit?usp=sharing&ouid=115912486417524737249&rtpof=true&sd=true