BANK LOAN CASE STUDY FINAL PROJECT-2

EXPLORATORY DATA ANALYSIS (EDA) AND INSIGHTS

Presented By :
Dhruv Singh

Presented By :
Dhruv Singh

INTRODUCTION

Objective:

- To analyze patterns in loan application data to identify factors influencing loan defaults.

Key Risks:

- Approving loans for customers who cannot repay.
- Rejecting loans for customers who can repay.
- Goal: Enable data-driven decisions for loan approval based on customer and loan attributes.

APPROACH

Data Preparation:

- Downloaded the dataset and inspected its structure.
- Identified missing values and handled them using appropriate imputation techniques.
- Detected and addressed outliers to ensure the integrity of the data.

Analysis:

- Conducted univariate, segmented univariate, and bivariate analyses to explore data distributions and relationships.
- Evaluated data imbalance and calculated class proportions.
- Computed correlations between variables and the target variable within segmented data.

Visualization:

- Used Excel charts and graphs, such as histograms, bar charts, pie charts, box plots, and scatter plots, to visualize findings.

Reporting:

- Documented insights and key findings in this report.
- Hyperlinked Excel sheets with the dataset analysis for easy reference.

TECH STACK USED

Microsoft Excel 2022:
Used for data preparation, analysis, and visualization.

Google Drive:
Hosted Excel files for sharing and integration.

Canva:
Used for Creating Report

HANDLING MISSING VALUES

Missing Data:

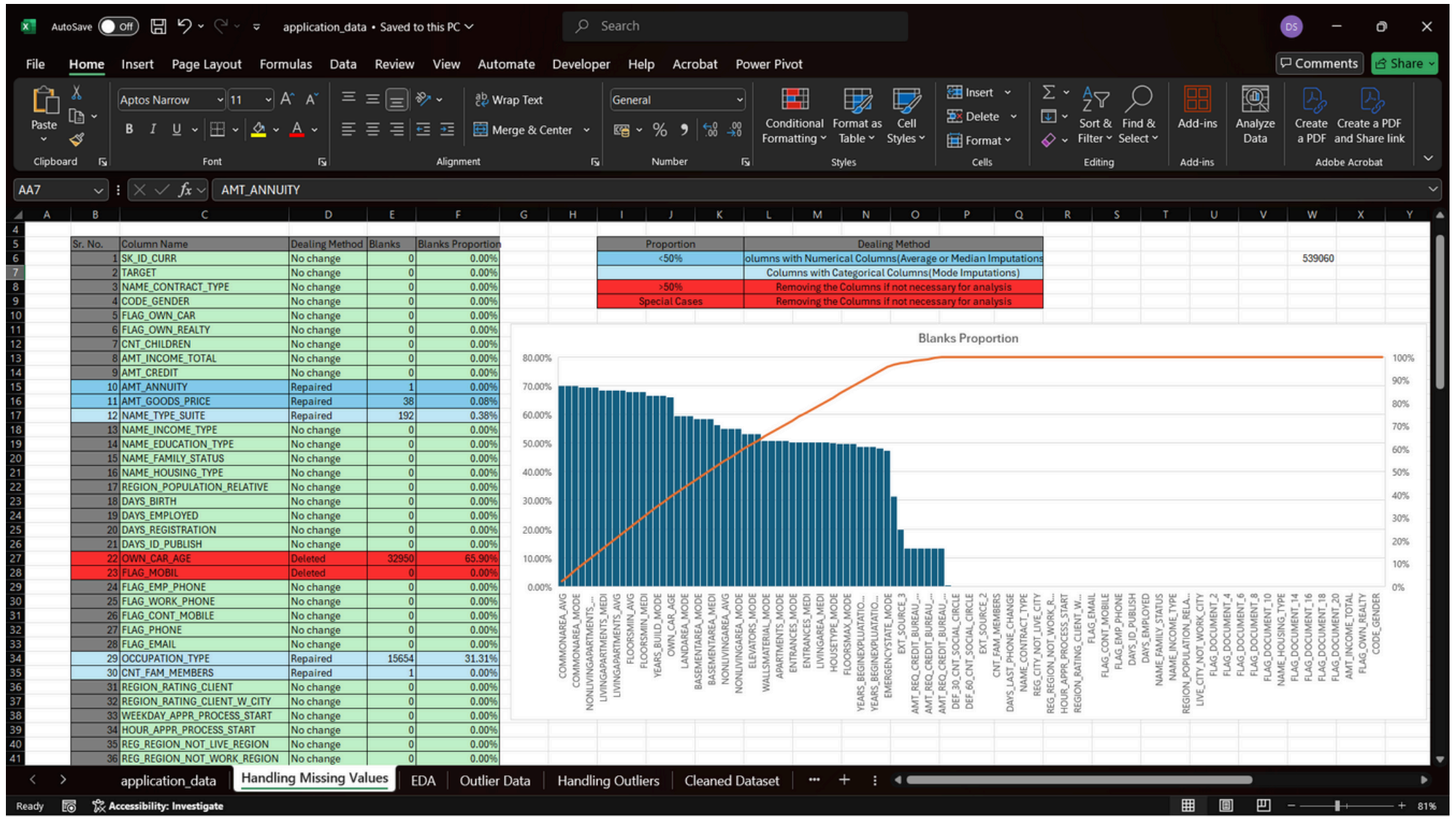
- Identified variables with missing values using Excel.
- Imputed missing values using mean/median for numerical variables.
- For categorical variables, used mode or created 'Unknown' category.

Visualization:

- Bar chart showing proportion of missing values (see Excel).

A	B	C	D	E	F
Percentage Value					
Blanks	=countb				
	S COUNTBLANK		Counts the number of empty cells in a specified range of cells		_CAR
	100002	1	Cash loans	M	N
	100003	0	Cash loans	F	N
	100004	0	Revolving loans	M	Y

Counted all the blank values using COUNTBLANK() Function



Sr. No.	Column Name	Dealing Method	Blanks	Blanks Proportion
1	SK_ID_CURR	No change	0	0.00%
2	TARGET	No change	0	0.00%
3	NAME_CONTRACT_TYPE	No change	0	0.00%
4	CODE_GENDER	No change	0	0.00%
5	FLAG_OWN_CAR	No change	0	0.00%
6	FLAG_OWN_REALTY	No change	0	0.00%
7	CNT_CHILDREN	No change	0	0.00%
8	AMT_INCOME_TOTAL	No change	0	0.00%
9	AMT_CREDIT	No change	0	0.00%
10	AMT_ANNUITY	Repaired	1	0.00%
11	AMT_GOODS_PRICE	Repaired	38	0.08%
12	NAME_TYPE_SUITE	Repaired	192	0.38%
13	NAME_INCOME_TYPE	No change	0	0.00%
14	NAME_EDUCATION_TYPE	No change	0	0.00%
15	NAME_FAMILY_STATUS	No change	0	0.00%
16	NAME_HOUSING_TYPE	No change	0	0.00%
17	REGION_POPULATION_RELATIVE	No change	0	0.00%
18	DAYS_BIRTH	No change	0	0.00%
19	DAYS_EMPLOYED	No change	0	0.00%
20	DAYS_REGISTRATION	No change	0	0.00%
21	DAYS_ID_PUBLISH	No change	0	0.00%
22	OWN_CAR_AGE	Deleted	32950	65.90%
23	FLAG_MOBIL	Deleted	0	0.00%
24	FLAG_EMP_PHONE	No change	0	0.00%
25	FLAG_WORK_PHONE	No change	0	0.00%
26	FLAG_CONT_MOBILE	No change	0	0.00%
27	FLAG_PHONE	No change	0	0.00%
28	FLAG_EMAIL	No change	0	0.00%
29	OCCUPATION_TYPE	Repaired	15654	31.31%
30	CNT_FAM_MEMBERS	Repaired	1	0.00%
31	REGION_RATING_CLIENT	No change	0	0.00%
32	REGION_RATING_CLIENT_W_CITY	No change	0	0.00%
33	WEEKDAY_APPR_PROCESS_START	No change	0	0.00%
34	HOUR_APPR_PROCESS_START	No change	0	0.00%
35	REG_REGION_NOT_LIVE_REGION	No change	0	0.00%
36	REG_REGION_NOT_WORK_REGION	No change	0	0.00%

- **Created a Separate sheet “Handling Missing Values” to solve the task in hand.**
- **Columns named Header name, Dealing Method and the Blank values with percentage are created and added in a table.**
- **The dealing method will be explained in the next slide.**
- **The visualization is done using a pareto chart to show the analysis of all the table headers with their Blank Percentage.**

The dealing method of columns were decided with the following criteria

Proportion	Dealing Method
<50%	olumns with Numerical Columns(Average or Median Imputations
	Columns with Categorical Columns(Mode Imputations)
>50%	Removing the Columns if not necessary for analysis
Special Cases	Removing the Columns if not necessary for analysis

Your paragraph text

- **All the columns with more than 50% of the blanks were deleted and the rest were repaired.**
- **The following table stores the data of all the columns repaired and the descriptive analysis used to imputate the data.**
- **The numerical columns were dealt through either mean or median while the categorical were dealth with the help of mode.**

Dealing with Missing Values

S No.	Columns Replacing Missing Value	Mean	Median	Mode
1	AMT_ANNUITY	27107.15	24939	
2	AMT_GOODS_PRICE	539052.3	450000	
3	NAME_TYPE_SUITE			Unaccompanied
4	OCCUPATION_TYPE			Laborers
5	CNT_FAM_MEMBERS	2.158726	2	
6	OBS_30_CNT_SOCIAL_CIRCLE	1.419397	0	
7	DEF_30_CNT_SOCIAL_CIRCLE	0.141346	0	
8	OBS_60_CNT_SOCIAL_CIRCLE	1.402336	0	
9	DEF_60_CNT_SOCIAL_CIRCLE	0.098004	0	
10	DAYS_LAST_PHONE_CHANGE	-964.315	-755	
11	AMT_REQ_CREDIT_BUREAU_HOUR	0.00614	0	
12	AMT_REQ_CREDIT_BUREAU_DAY	0.0065	0	
13	AMT_REQ_CREDIT_BUREAU_WEEK	0.028021	0	
14	AMT_REQ_CREDIT_BUREAU_MON	0.233889	0	
15	AMT_REQ_CREDIT_BUREAU_QRT	0.225829	0	
16	AMT_REQ_CREDIT_BUREAU_YEAR	1.627725	1	

3) Mode of Name Type suite

Category	Count
Unaccompanied	40626
Family	6549
Spouse, partner	1849
Children	542
Other_A	137
Other_B	259
Group of people	36
Blanks	192

4) OCCUPATION_TYPE

Category	Count
Laborers	24606
Core staff	4434
Accountants	1621
Managers	3488
Drivers	3044
Sales staff	5160
Cleaning staff	739
Cooking staff	963
Private service staff	447
Medicine staff	1403
Security staff	1140
High skill tech staff	1852
Waiters/barmen sta	228

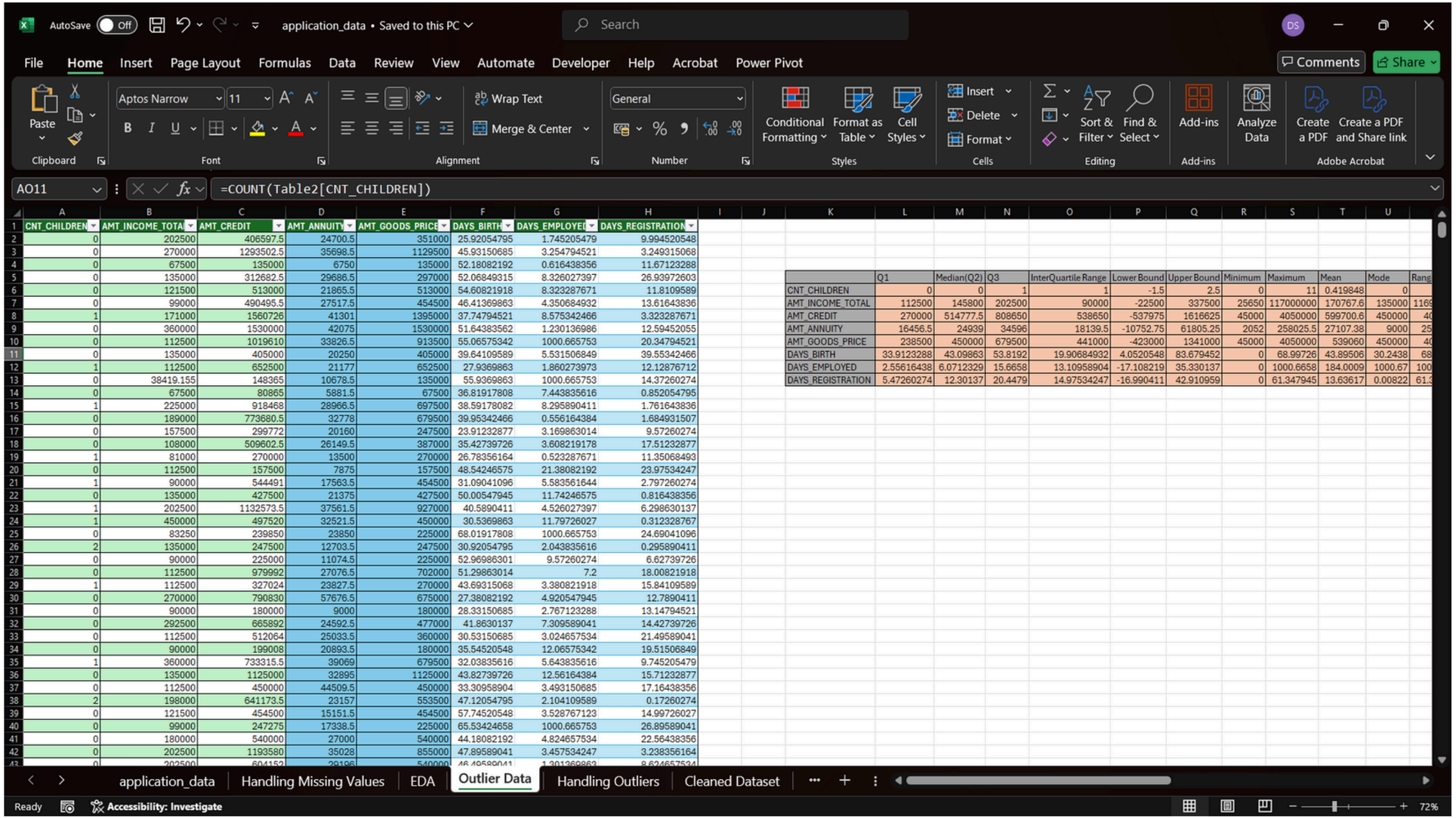
OUTLIER ANALYSIS

Method:

- Identified outliers using Interquartile Range (IQR).
- Calculated Q1, Q3, and IQR.
- Defined lower and upper bounds: Q1 - 1.5*IQR, Q3 + 1.5*IQR.
- Treated outliers by capping/extending within bounds.

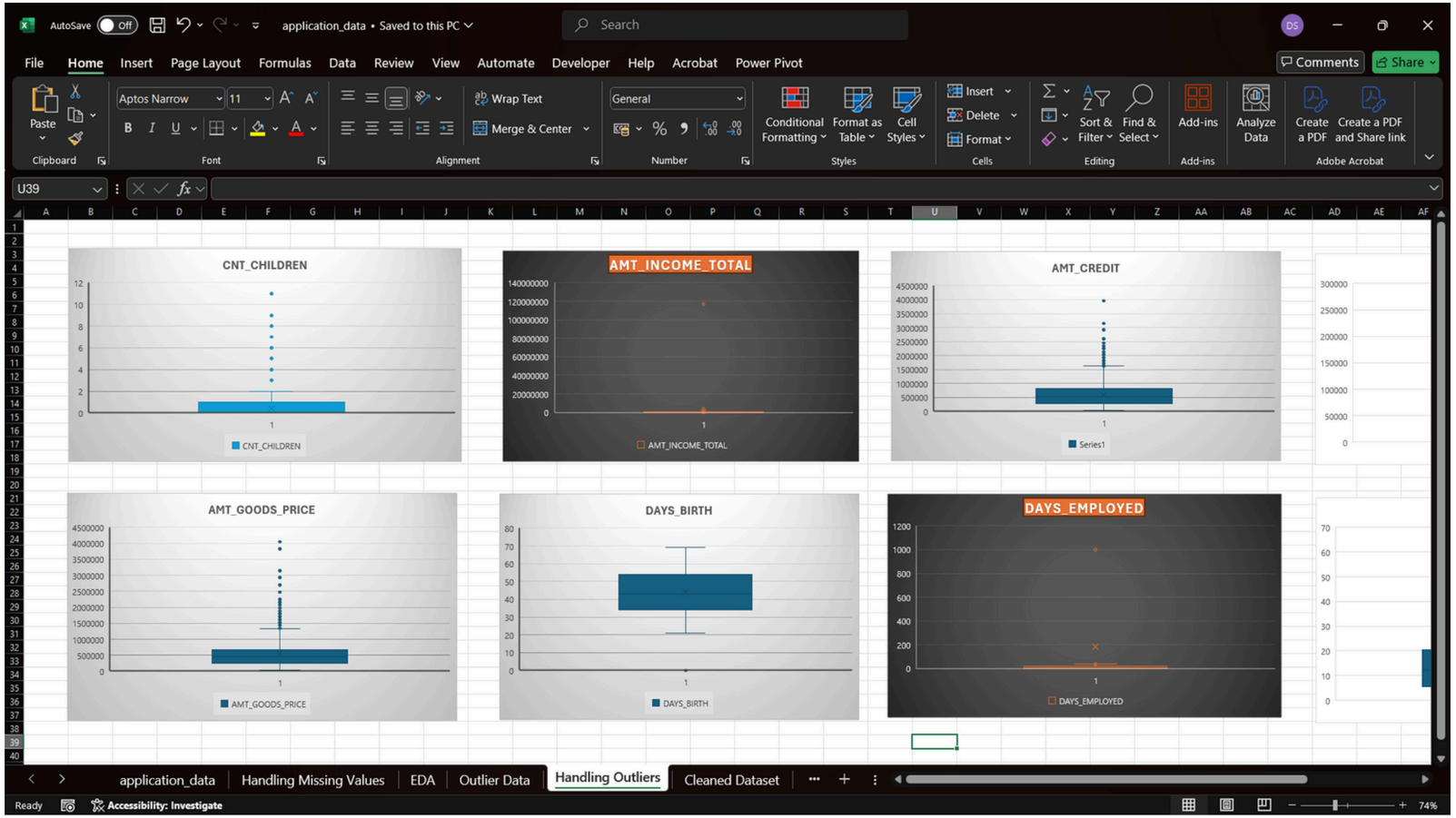
Visualization:

- Box plots for numerical variables (see Excel).



	Q1	Median(Q2)	Q3	InterQuartile Range	Lower Bound	Upper Bound	Minimum	Maximum	Mean	Mode	Rang
CNT_CHILDREN	0	0	1	1	-1.5	2.5	0	11	0.419848	0	
AMT_INCOME_TOTAL	112500	145800	202500	90000	-22500	337500	25650	117000000	170767.6	135000	116
AMT_CREDIT	270000	514777.5	808650	538650	-537975	1616625	45000	4050000	599700.6	450000	4(
AMT_ANNUITY	16456.5	24939	34596	18139.5	-10752.75	61805.25	2052	258025.5	27107.38	9000	25
AMT_GOODS_PRICE	238500	450000	679500	441000	-423000	1341000	45000	4050000	539060	450000	4(
DAYS_BIRTH	33.9123288	43.09863	53.8192	19.90684932	4.0520548	83.679452	0	68.99726	43.89506	30.2438	68
DAYS_EMPLOYED	2.55616438	6.0712329	15.6658	13.10958904	-17.108219	35.330137	0	1000.6658	184.0009	1000.67	100
DAYS_REGISTRATION	5.47260274	12.30137	20.4479	14.97534247	-16.990411	42.910959	0	61.347945	13.63617	0.00822	61.3

- **The above tables depict two different set of data.**
- **The first are the numerical rows suspected of having outliers found through filtering the data.**
- **The second contains the quartile values and IQR(Inter Quartile Range) as well as the bounds required for the Box Plots and some other descriptive analysis such as mean, median, mode, standard deviation etc.**



- Contains all the box plots of the previous table.
- The box plot helps to find the outliers.
- It is clearly visible some charts have a lot of them.
- One of the outlier to be removed can be found in the days employed(years) which has the value of 1000 years. Now it is impossible for someone to be working for a 1000 years and thus it is clearly a case of a mistype or an error.
- Thus, it was removed from the table.

DATA IMBALANCE

Analysis:

- **Assessed class distribution of target variable using COUNTIF.**
- **Imbalance found in loan outcomes, i.e., the male to female criteria has a big gap.**

Visualization:

- **Pie chart representing class distribution (see Excel).**

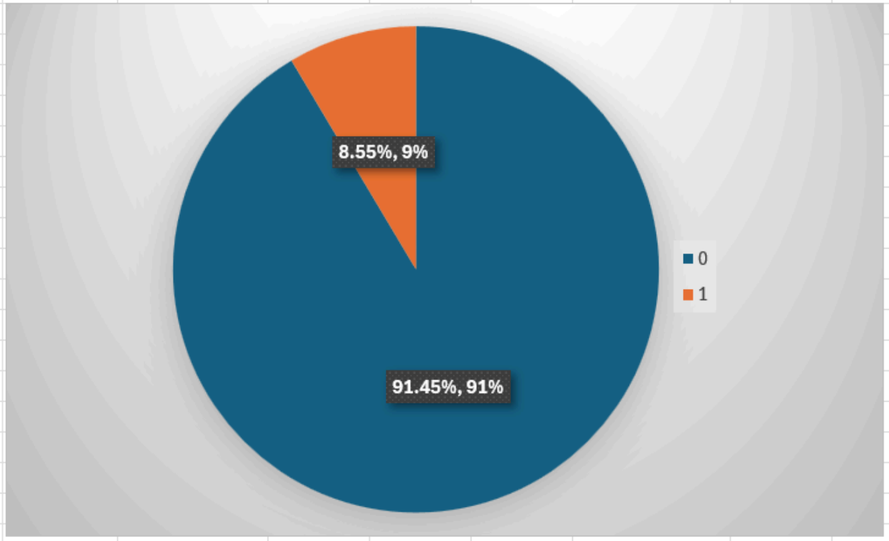
Data Imbalance

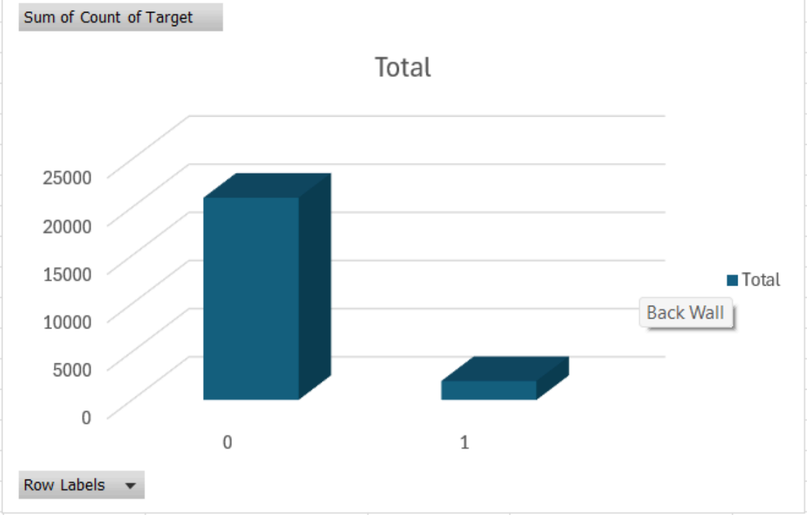
Row Labels	Count of Target
0	20994
1	1964
Total	22958

	Count of 0 and 1	Ratio	Target	Contribution
0	20994	10.6894	0	91.45%
1	1964		1	8.55%

Row Labels	Sum of Count of Target
0	20994
1	1964
Grand Total	22958

Note	
1	Clients With payment difficulties(Defaulters)
0	All other cases(Non-Defaulters)





Handling Missing Values | EDA | Outlier Data | Handling Outliers | Cleaned Dataset | Data Imbalance

UNIVARIATE AND BIVARIATE ANALYSIS

Univariate Analysis:

- Examined distributions of individual variables using histograms and bar charts.

Segmented Univariate Analysis:

- Compared distributions for scenarios (e.g., defaulters vs non-defaulters).

Bivariate Analysis:

- Analyzed relationships between variables using scatter plots and pivot tables.

Visualizations:

- Various charts (see Excel).

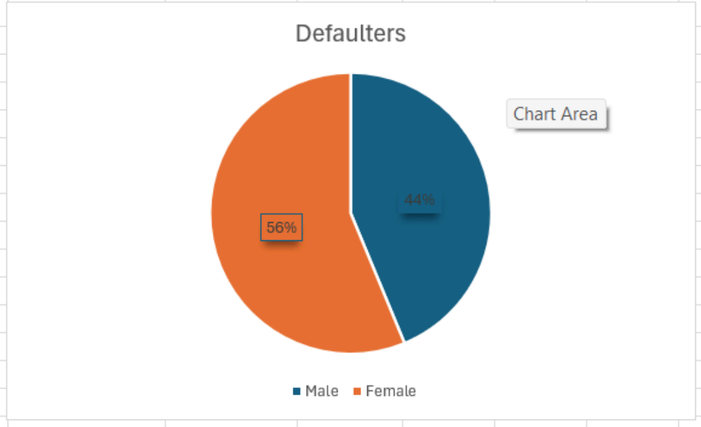
Univariate Analysis

Task D	Non-Defaulters	0
	Defaulters	1

CODE_GENDER	TARGET
M	1
F	0
M	0
F	0
M	0
M	0
F	0
M	0
F	0
M	0
F	0
F	0
F	0
M	0
F	0
M	0
F	0
F	0
M	0
F	0
F	0
M	0
F	0
F	0
M	0
F	0
F	1

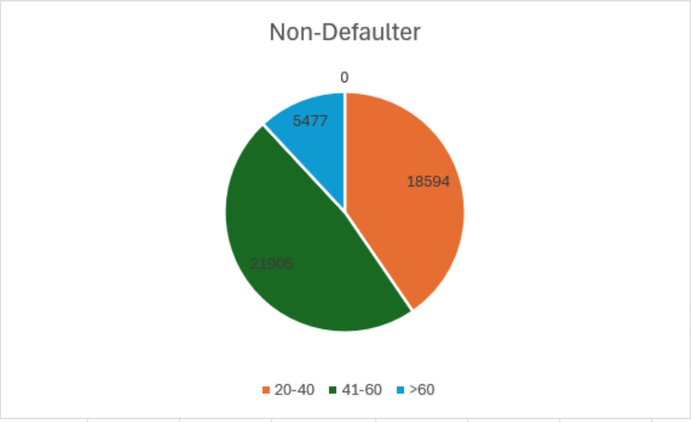
	Male	Female
Defaulters	1761	2264
Non-Defaulters	15412	30559
Total	17173	32823

	Defaulters	Percentage	Total
Male	1761	2264	4025
Female	44%	56%	100%

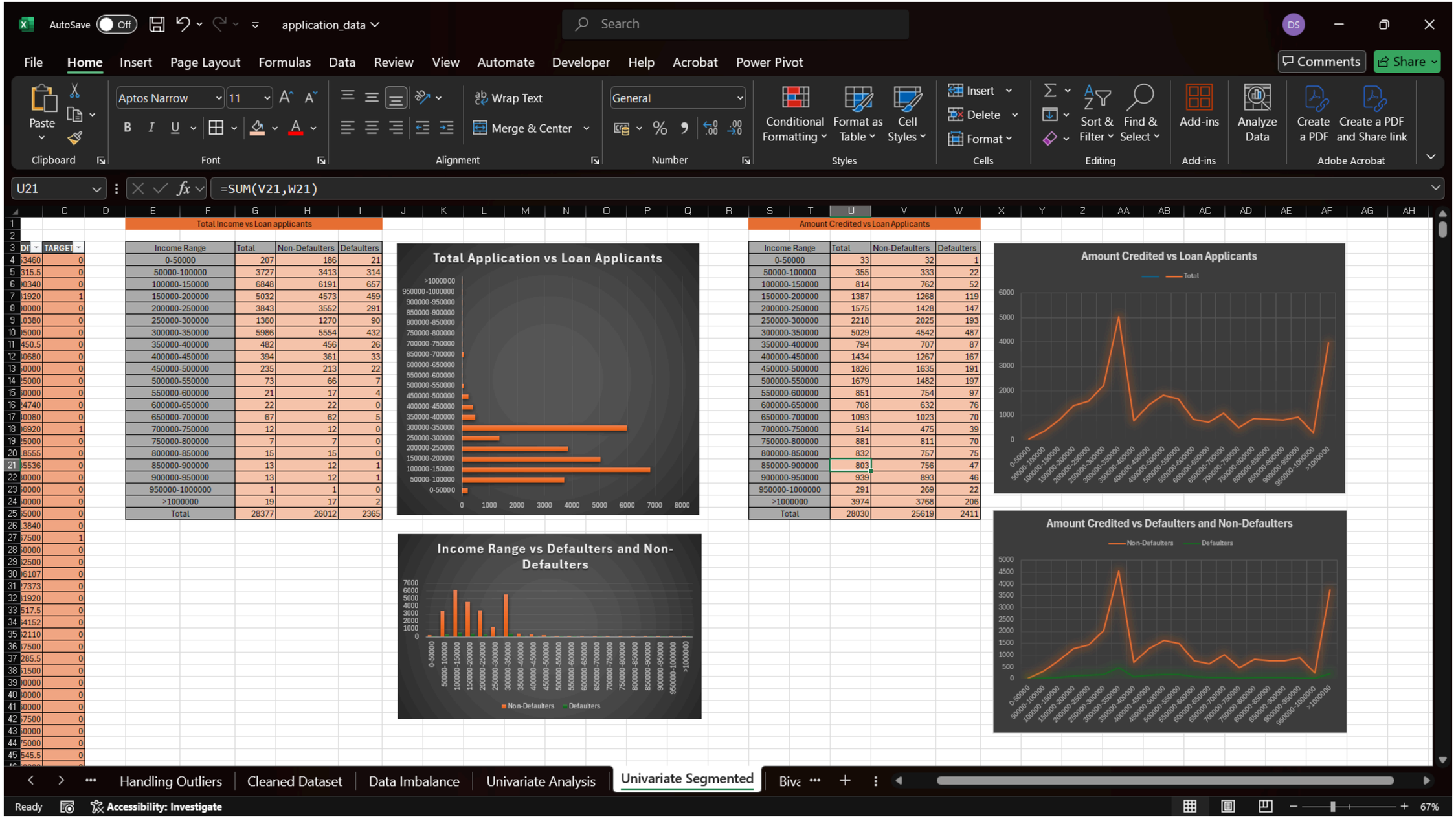


Client Age	TARGET
25.920548	1
45.931507	0
52.180822	0
52.068493	0
54.608219	0
46.413699	0
37.747945	0
51.643836	0
55.065753	0
39.641096	0
27.936986	0
55.936986	0
36.819178	0
38.591781	0
39.953425	0
23.912329	0
35.427397	0
26.783562	0
48.542466	0
31.090411	0
50.005479	0
40.589041	0
30.536986	0
68.019178	0
30.920548	0
52.969863	0
51.29863	1

Age	0	1	Loan Taken
20-40	18594	2130	20724
41-60	21906	1607	23513
>60	5477	289	5766



Univariate Segmented Analysis



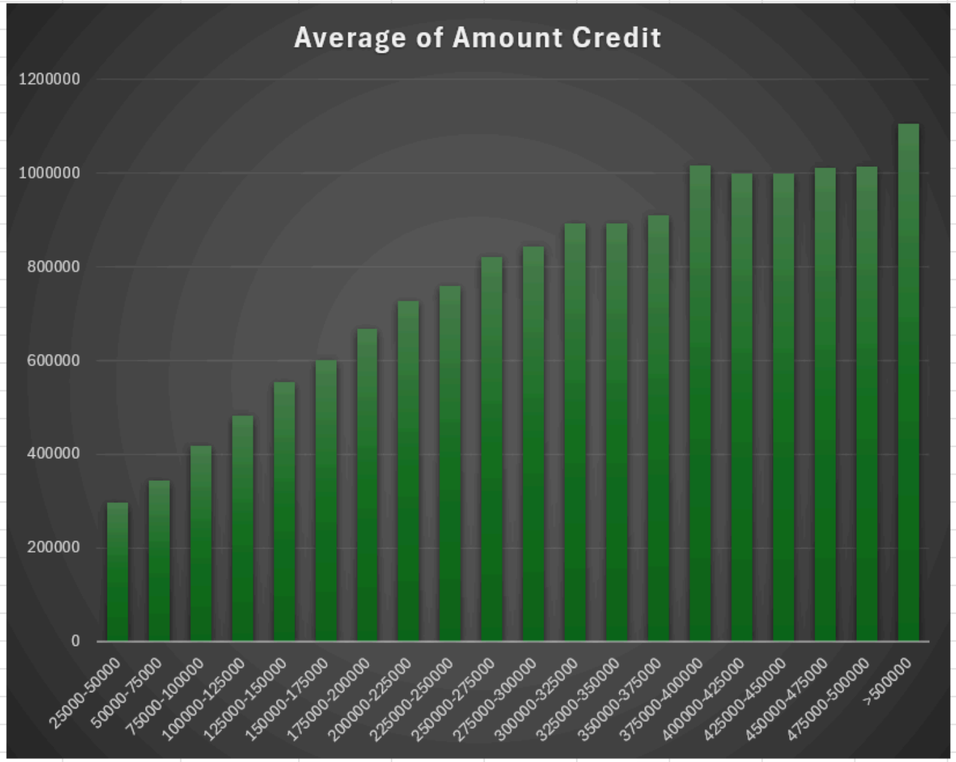
Income Range	Total	Non-Defaulters	Defaulters
0-50000	207	186	21
50000-100000	3727	3413	314
100000-150000	6848	6191	657
150000-200000	5032	4573	459
200000-250000	3843	3552	291
250000-300000	1360	1270	90
300000-350000	5986	5554	432
350000-400000	482	456	26
400000-450000	394	361	33
450000-500000	235	213	22
500000-550000	73	66	7
550000-600000	21	17	4
600000-650000	22	22	0
650000-700000	67	62	5
700000-750000	12	12	0
750000-800000	7	7	0
800000-850000	15	15	0
850000-900000	13	12	1
900000-950000	13	12	1
950000-1000000	1	1	0
>1000000	19	17	2
Total	28377	26012	2365

Income Range	Total	Non-Defaulters	Defaulters
0-50000	33	32	1
50000-100000	355	333	22
100000-150000	814	762	52
150000-200000	1387	1268	119
200000-250000	1575	1428	147
250000-300000	2218	2025	193
300000-350000	5029	4542	487
350000-400000	794	707	87
400000-450000	1434	1267	167
450000-500000	1826	1635	191
500000-550000	1679	1482	197
550000-600000	851	754	97
600000-650000	708	632	76
650000-700000	1093	1023	70
700000-750000	514	475	39
750000-800000	881	811	70
800000-850000	832	757	75
850000-900000	803	756	47
900000-950000	939	893	46
950000-1000000	291	269	22
>1000000	3974	3768	206
Total	28030	25619	2411

Bivariate Analysis

NCOME_TOTAL	AMT_CREDIT
202500	406597.5
270000	1293502.5
67500	135000
135000	312682.5
121500	513000
99000	490495.5
171000	1560726
360000	1530000
112500	1019610
135000	405000
112500	652500
38419.155	148365
67500	80865
225000	918468
189000	773680.5
157500	299772
108000	509602.5
81000	270000
112500	157500
90000	544491
135000	427500
202500	1132573.5
450000	497520
83250	239850
135000	247500
90000	225000
112500	979992
112500	327024
270000	790830

Lower Bound	Upper Bound	Income Bin	Average of Amount Credit
25000	50000	25000-50000	297752.0765
50000	75000	50000-75000	345240.3585
75000	100000	75000-100000	417267.8771
100000	125000	100000-125000	483568.8073
125000	150000	125000-150000	553042.1642
150000	175000	150000-175000	602034.4016
175000	200000	175000-200000	667004.421
200000	225000	200000-225000	727198.4449
225000	250000	225000-250000	759541.3782
250000	275000	250000-275000	820255.3451
275000	300000	275000-300000	842725.6488
300000	325000	300000-325000	892198.254
325000	350000	325000-350000	892332.6503
350000	375000	350000-375000	910363.0482
375000	400000	375000-400000	1016814.375
400000	425000	400000-425000	999208.199
425000	450000	425000-450000	999153.6402
450000	475000	450000-475000	1011521.839
475000	500000	475000-500000	1015150.404
500000		>500000	1105365.122



CORRELATION ANALYSIS

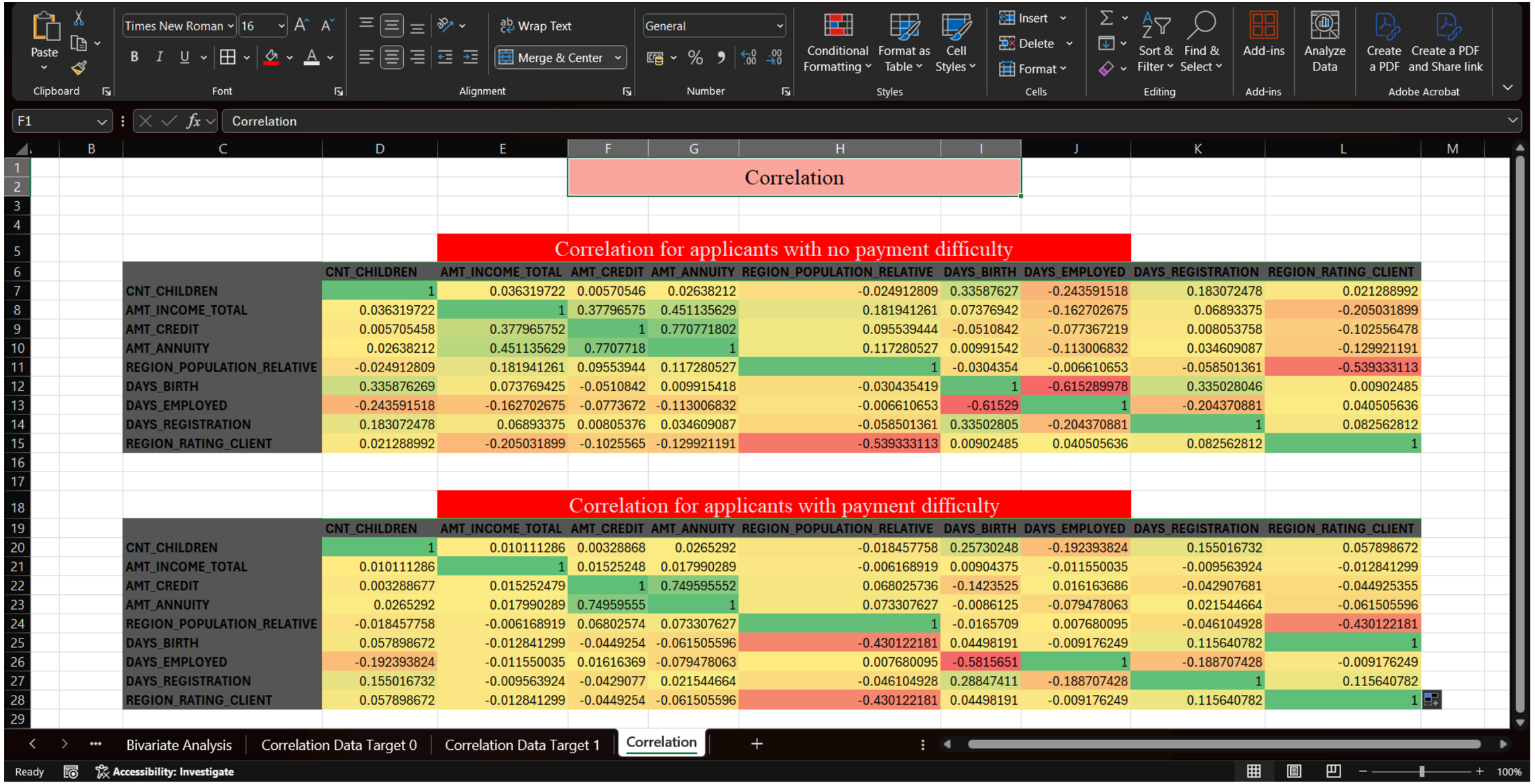
Segmented Correlation:

- Identified correlations within segmented data (e.g., defaulters).
- Used CORREL function to calculate correlation coefficients.
- Highlighted top predictors influencing loan default.

Visualization:

- Correlation heatmaps (see Excel).

Correlation Analysis



Correlation

Correlation for applicants with no payment difficulty

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.036319722	0.00570546	0.02638212	-0.024912809	0.33587627	-0.243591518	0.183072478	0.021288992
AMT_INCOME_TOTAL	0.036319722	1	0.37796575	0.451135629	0.181941261	0.07376942	-0.162702675	0.06893375	-0.205031899
AMT_CREDIT	0.005705458	0.377965752	1	0.770771802	0.095539444	-0.0510842	-0.077367219	0.008053758	-0.102556478
AMT_ANNUITY	0.02638212	0.451135629	0.7707718	1	0.117280527	0.00991542	-0.113006832	0.034609087	-0.129921191
REGION_POPULATION_RELATIVE	-0.024912809	0.181941261	0.09553944	0.117280527	1	-0.0304354	-0.006610653	-0.058501361	-0.539333113
DAYS_BIRTH	0.335876269	0.073769425	-0.0510842	0.009915418	-0.030435419	1	-0.615289978	0.335028046	0.00902485
DAYS_EMPLOYED	-0.243591518	-0.162702675	-0.0773672	-0.113006832	-0.006610653	-0.61529	1	-0.204370881	0.040505636
DAYS_REGISTRATION	0.183072478	0.06893375	0.00805376	0.034609087	-0.058501361	0.33502805	-0.204370881	1	0.082562812
REGION_RATING_CLIENT	0.021288992	-0.205031899	-0.1025565	-0.129921191	-0.539333113	0.00902485	0.040505636	0.082562812	1

Correlation for applicants with payment difficulty

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.010111286	0.00328868	0.0265292	-0.018457758	0.25730248	-0.192393824	0.155016732	0.057898672
AMT_INCOME_TOTAL	0.010111286	1	0.01525248	0.017990289	-0.006168919	0.00904375	-0.011550035	-0.009563924	-0.012841299
AMT_CREDIT	0.003288677	0.015252479	1	0.749595552	0.068025736	-0.1423525	0.016163686	-0.042907681	-0.044925355
AMT_ANNUITY	0.0265292	0.017990289	0.74959555	1	0.073307627	-0.0086125	-0.079478063	0.021544664	-0.061505596
REGION_POPULATION_RELATIVE	-0.018457758	-0.006168919	0.06802574	0.073307627	1	-0.0165709	0.007680095	-0.046104928	-0.430122181
DAYS_BIRTH	0.057898672	-0.012841299	-0.0449254	-0.061505596	-0.430122181	0.04498191	-0.009176249	0.115640782	1
DAYS_EMPLOYED	-0.192393824	-0.011550035	0.01616369	-0.079478063	0.007680095	-0.5815651	1	-0.188707428	-0.009176249
DAYS_REGISTRATION	0.155016732	-0.009563924	-0.0429077	0.021544664	-0.046104928	0.28847411	-0.188707428	1	0.115640782
REGION_RATING_CLIENT	0.057898672	-0.012841299	-0.0449254	-0.061505596	-0.430122181	0.04498191	-0.009176249	0.115640782	1

INSIGHTS AND RECOMMENDATIONS

Key Insights:

- High-income customers less likely to default.
- Loan amounts exceeding a threshold increase default risk.
- Data imbalance in loan approvals impacts analysis.

Recommendations:

- Implement stricter checks for high-risk applicants.
- Offer loans with adjusted terms to risky applicants (e.g., higher interest).
- Improve data collection to reduce missing values.

CONCLUSION

The analysis revealed actionable insights for loan approval strategies. By focusing on high-risk attributes such as low income, high loan amounts, and unfavorable debt-to-income ratios, the company can:

- Deny loans to high-risk applicants.
- Adjust loan amounts or interest rates based on risk.
- Develop policies for shorter loan tenures for specific customer segments.

PROJECT DELIVERABLES

Google Drive Link (Full Report & Resources)

Excel File Link (Detailed Analysis)

Video Presentation Link

THANK YOU