Project by – Alokk Joshi

Bank Loan Analytics

Project by – Alokk Joshi

Bank Loan Case Study

As a data analyst at a finance company specializing in urban loans, my role is to analyze customer loan applications to minimize financial risk.

The company faces two challenges simultaneously: rejecting creditworthy applicants leads to lost business, while approving high-risk applicants results in defaults and financial losses.

Using Exploratory Data Analysis (EDA), this project aims to identify patterns in customer attributes and loan characteristics that influence loan repayment behavior. **The goal is to help the company make informed decisions on loan approvals, reducing risks while ensuring business growth.**

Key Data Analytics Tasks

- **Identify and Handle Missing Data** – Detect missing values and decide on the best approach (imputation or removal) to maintain data integrity.
- **Identify Outliers** – Find extreme values in numerical variables using statistical techniques to prevent data distortion.
- **Analyze Data Imbalance** – Examine the distribution of loan default cases to assess if the dataset is skewed towards a particular class.
- **Perform Univariate, Segmented Univariate, and Bivariate Analysis** – Explore how individual and combined attributes impact loan repayment behavior.
- **Identify Top Correlations** – Determine the strongest relationships between customer attributes and loan default risk for better decision-making.

Project 6 – Bank Loan Case Study

Project by – Alokk Joshi

Task 1

Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

Project 6 – Bank Loan Case Study

Project by – Alokk Joshi

Application_data file

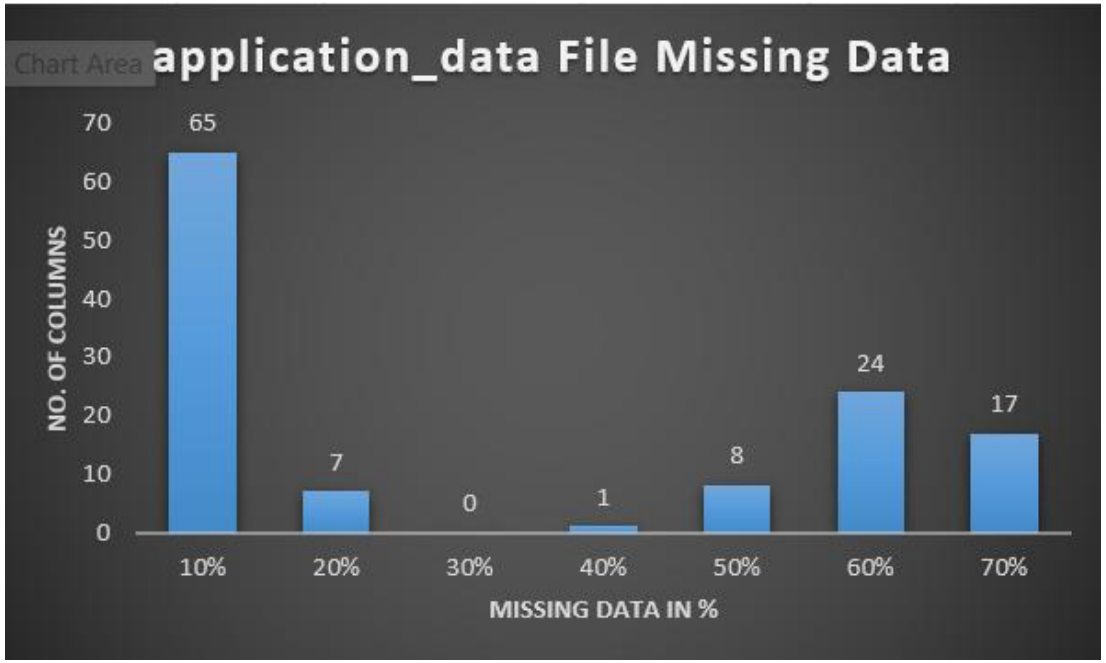
Total Columns	122
Total Rows	49999
Total Blank Cells	1488212
Columns with > 50% Blank Cells	41
% of cells having >50% missing data	34%

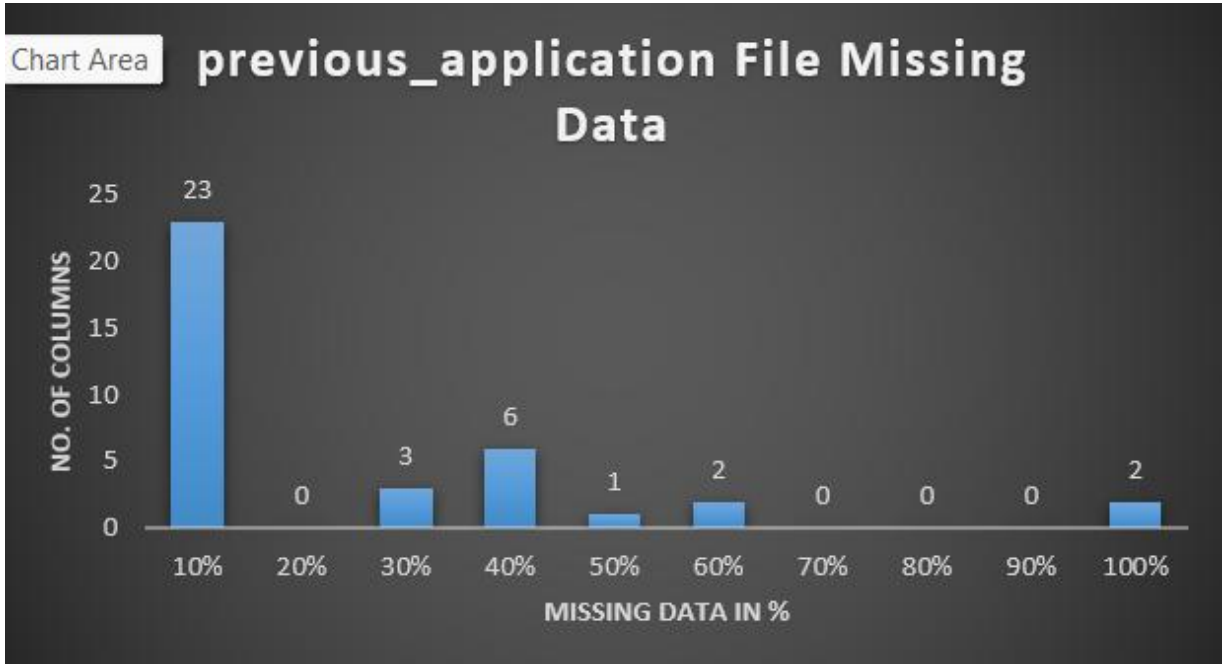
Previous_application file

Total Columns	37
Total Rows	49999
Total Blank Cells	321203
Columns with > 50% Blank Cells	4
% of cells having >50% missing data	11%

Project 6 – Bank Loan Case Study

Project by – Alokk Joshi





Project 6 – Bank Loan Case Study

Project by – Alokk Joshi

Column Name	Missing Values	% of Missing Values
COMMONAREA_AVG	34960	70%
COMMONAREA_MODE	34960	70%
COMMONAREA_MEDI	34960	70%
NONLIVINGAPARTMENTS_AVG	34714	69%
NONLIVINGAPARTMENTS_MODE	34714	69%
NONLIVINGAPARTMENTS_MEDI	34714	69%
LIVINGAPARTMENTS_AVG	34226	68%
LIVINGAPARTMENTS_MODE	34226	68%
LIVINGAPARTMENTS_MEDI	34226	68%
FONDKAPREMONT_MODE	34191	68%
FLOORSMIN_AVG	33894	68%
FLOORSMIN_MODE	33894	68%
FLOORSMIN_MEDI	33894	68%
YEARS_BUILD_AVG	33239	66%
YEARS_BUILD_MODE	33239	66%
YEARS_BUILD_MEDI	33239	66%
OWN_CAR_AGE	32950	66%
LANDAREA_AVG	29721	59%
LANDAREA_MODE	29721	59%
LANDAREA_MEDI	29721	59%
BASEMENTAREA_AVG	29199	58%
BASEMENTAREA_MODE	29199	58%
BASEMENTAREA_MEDI	29199	58%
EXT_SOURCE_1	28172	56%
NONLIVINGAREA_AVG	27572	55%
NONLIVINGAREA_MODE	27572	55%
NONLIVINGAREA_MEDI	27572	55%
ELEVATORS_AVG	26651	53%
ELEVATORS_MODE	26651	53%
ELEVATORS_MEDI	26651	53%
WALLSMATERIAL_MODE	25459	51%
APARTMENTS_AVG	25385	51%
APARTMENTS_MODE	25385	51%
APARTMENTS_MEDI	25385	51%
ENTRANCES_AVG	25195	50%
ENTRANCES_MODE	25195	50%
ENTRANCES_MEDI	25195	50%
LIVINGAREA_AVG	25137	50%
LIVINGAREA_MODE	25137	50%
LIVINGAREA_MEDI	25137	50%
HOUSETYPE_MODE	25075	50%

Application_data file

Column Name	Missing Values	% of Missing Values
RATE_INTEREST_PRIMARY	49834	99.7%
RATE_INTEREST_PRIVILEGED	49834	99.7%
AMT_DOWN_PAYMENT	25198	50.4%
RATE_DOWN_PAYMENT	25198	50.4%

Previous_application file

We will choose to delete these columns as they lack more than 50% data inside them and it will prevent us from getting meaningful insights

Application_data file missing data replacement strategy

AMT_ANNUITY	
Mean	27107.37736
Standard Error	65.12877001
Median	24939
Mode	9000
Standard Deviation	14562.94444
Sample Variance	212079350.6
Kurtosis	9.412028546
Skewness	1.688525905
Range	255973.5
Minimum	2052
Maximum	258025.5
Sum	1355314653
Count	49998

CNT_FAM_MEMBERS	
Mean	2.158946
Standard Error	0.004076
Median	2
Mode	2
Standard Deviation	0.911332
Sample Variance	0.830527
Kurtosis	1.715436
Skewness	0.949618
Range	12
Minimum	1
Maximum	13
Sum	107943
Count	49998

DAYS_LAST_PHONE_CHANGE	
Mean	-964.296
Standard Error	3.709646
Median	-755
Mode	0
Standard Deviation	829.4856
Sample Variance	688046.3
Kurtosis	-0.32418
Skewness	-0.71092
Range	4002
Minimum	-4002
Maximum	0
Sum	-4.8E+07
Count	49998

AMT_GOODS_PRICE	
Mean	539060.0361
Standard Error	1654.67948
Median	450000
Mode	450000
Standard Deviation	369853.2527
Sample Variance	1.36791E+11
Kurtosis	2.486844961
Skewness	1.347815809
Range	4005000
Minimum	45000
Maximum	4050000
Sum	26931978465
Count	49961

Will replace the null values with **Median as skewness is either positive or negative**

DS	AMT_REQ_CREDIT_BUREAU_HOUR	AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_WEEK	AMT_REQ_CREDIT_BUREAU_MON	AMT_REQ_CREDIT_BUREAU_QRT	AMT_REQ_CREDIT_BUREAU_YEAR
Mean	0.007095805	0.00751185	0.03238183	0.27028776	0.26097307	1.88103548
Mode	0	0	0	0	0	0
Median	0	0	0	0	0	1
Min	0	0	0	0	0	0
Max	3	6	6	24	8	25
SD	0.087708647	0.10799223	0.19408035	0.92856012	0.60699573	1.8650543
VAR	0.007692807	0.01166232	0.03766718	0.8622239	0.36844381	3.47842755
Skewness	13.56285992	22.2738602	7.92758702	7.97367358	2.70297074	1.29628222

Will replace the null values with **Median as skewness are highly positive**

DS	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE
Mean	1.420782244	0.141819349	1.403664386	0.098332363
Mode	0	0	0	0
Median	0	0	0	0
Min	0	0	0	0
Max	28	6	28	5
SD	2.302085879	0.440539565	2.281781763	0.357263762
VAR	5.299599396	0.194075109	5.206528013	0.127637396
Skewness	2.525749911	3.865176954	2.530120243	4.460229402

Will replace the null values with **Median as skewness are highly positive**

Project 6 – Bank Loan Case Study

Project by – Alokk Joshi

Row Labels	Count of OCCUPATION_TYPE
	15654
Laborers	8952
Sales staff	5160
Core staff	4434
Managers	3489
Drivers	3044
High skill tech staff	1852
Accountants	1621
Medicine staff	1403
Security staff	1140
Cooking staff	963
Cleaning staff	739
Private service staff	447
Low-skill Laborers	357
Waiters/barmen staff	228
Secretaries	212
Realty agents	123
HR staff	101
IT staff	80
(blank)	
Grand Total	**49999**

We will choose **MODE** imputation as they are text columns

Imputation values

Laborers
Unaccompanied
No

Row Labels	Count of EMERGENCYSTATE_MODE
No	25944
	23698
Yes	357
(blank)	
Grand Total	**49999**

Row Labels	Count of NAME_TYPE_SUITE
Unaccompanied	40435
Family	6549
Spouse, partner	1849
Children	542
Other_B	259
	192
Other_A	137
Group of people	36
(blank)	
Grand Total	**49999**

DS	FLOORSMIN_AVG	FLOORSMAX_MODE	FLOORSMAX_MEDI	YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_MEDI	TOTALAREA_MODE
Mean	0.23165	0.221489	0.225081	0.978036	0.977404	0.978031	0.102690027
Mode	0.2083	0.1667	0.1667	0.9871	0.9871	0.9871	0
Median	0.2083	0.1667	0.1667	0.9816	0.9816	0.9816	0.0685
Min	0	0	0	0	0	0	0
Max	1	1	1	1	1	1	1
SD	0.161545	0.144289	0.145574	0.056486	0.061657	0.057363	0.107950724
VAR	0.026097	0.020819	0.021192	0.003191	0.003802	0.00329	0.011653359
Skewness	0.964059	1.273558	1.265787	-16.21588	-15.42783	-16.23244	2.776809393

Will replace the null values with **Median as skewness are either positive or negative**

Previous_application file missing data replacement strategy

Row Labels	Count of NAME_TYPE_SUITE
	24243
Unaccompanied	15195
Family	6581
Spouse, partner	2098
Children	993
Other_B	551
Other_A	262
Group of people	76
(blank)	
Grand Total	**49999**

Row Labels	Count of PRODUCT_COMBINATION
POS household with interest	8510
Cash	7939
POS mobile with interest	7029
Cash X-Sell: middle	3953
Cash X-Sell: low	3539
Card Street	3323
POS industry with interest	3231
POS household without interest	2799
Card X-Sell	2302
Cash Street: high	1752
Cash X-Sell: high	1657
Cash Street: low	1056
Cash Street: middle	960
POS mobile without interest	731
POS other with interest	728
POS industry without interest	390
POS others without interest	92
	8
(blank)	
Grand Total	**49999**

MODE imputation due to text values

DS	DAYS_FIRST_DRAWING	DAYS_FIRST_DUE	DAYS_LAST_DUE_1ST_VERSION	DAYS_LAST_DUE_1ST_VERSION	DAYS_TERMINATION	AMT_GOODS_PRICE	AMT_ANNUITY	CNT_PAYMENT	NFLAG_INSURED_ON_APPROVAL
Mean	344485.1428	14217.24015	31528.14861	31528.14861	81666.16259	215141.4173	15482.59685	15.55589109	0.322351568
Mode	365243	365243	365243	365243	365243	45000	2250	12	0
Median	365243	-822	-366	-366	-500	104017.5	10879.92	12	0
Min	-2910	-2891	-2800	-2800	-2844	0	0	0	0
Max	365243	365243	365243	365243	365243	3826372.5	234478.395	60	1
SD	84683.65063	73348.98438	103691.8812	103691.8812	153101.1598	302499.2745	14530.97185	13.98517447	0.467384337
VAR	7171320684	5380073510	10752006224	10752006224	23439965135	91505811087	211149143	195.5851051	0.218448118
Skewness	-3.834616977	4.576474044	2.907553093	2.907553093	1.312382743	3.176092575	2.702398019	1.624122285	0.760230671

Will replace the null values with **Median as skewness are either positive or negative**

Task 2

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

Setting the layout for outlier detection

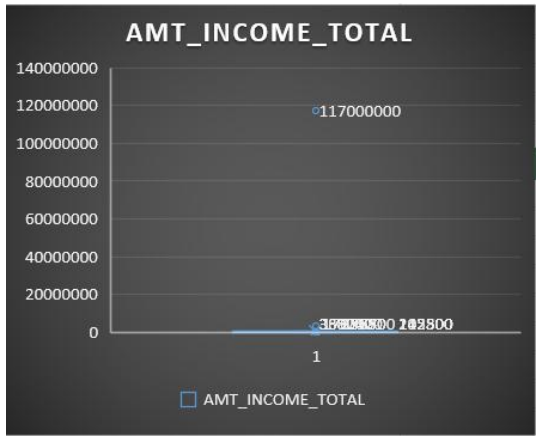
Application_data file

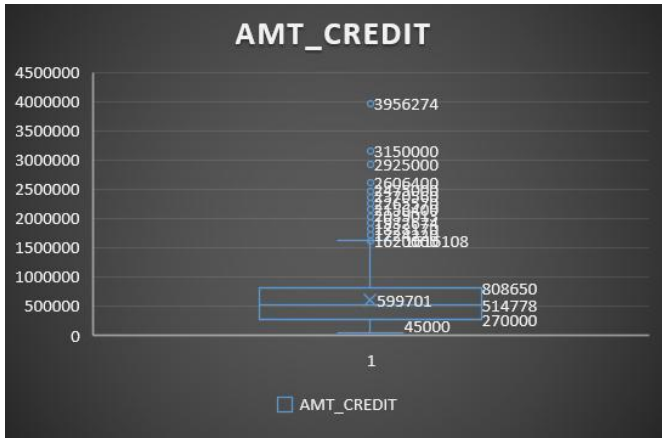
		AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	CNT_CHILDREN	YEARS_EMPLOYED
Quartile 1 = QUARTILE(B2:B50000,1)	Q1	112500	270000	16457	238500	0	3
Quartile 3 = QUARTILE(B2:B50000,3)	Q3	202500	808650	34596	679500	1	16
IQR = Quartile 3 -Quartile 1	IQR	90000	538650	18140	441000	1	13
Upper Limit = Quartile 3 + (1.5*IQR)	Upper Limit	337500	1616625	61805	1341000	3	36
Lower Limit = Quartile 1 – (1.5*IQR)	Lower Limit	-22500	-537975	-10753	-423000	-2	-17

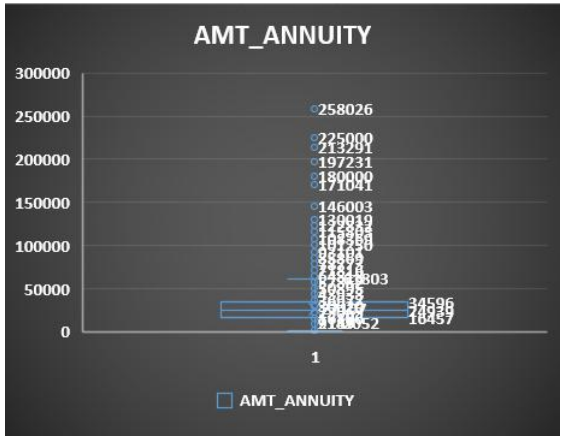
Project 6 – Bank Loan Case Study

Project by – Alokk Joshi

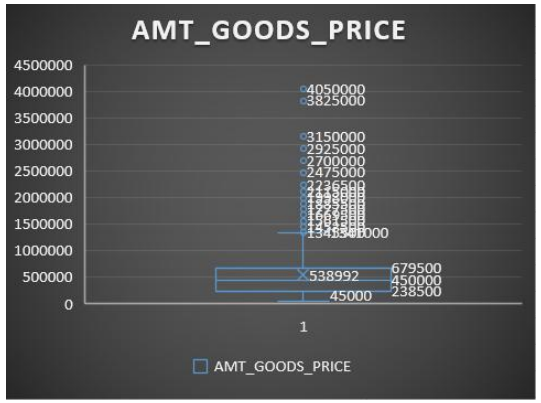
Outlier Detection

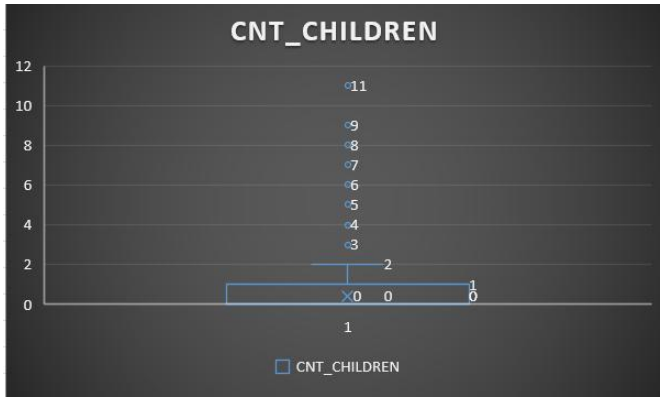


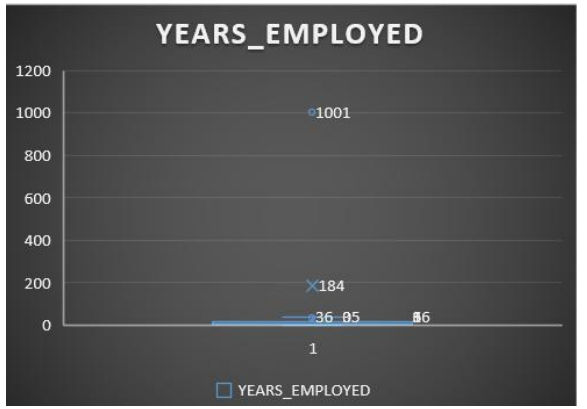




This column was additionally inserted








```
=COUNT(FILTER($B$2:$B$50000,($B$2:$B$50000>K8)+($B$2:$B$50000<K9)))
```

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	CNT_CHILDREN	YEARS_EMPLOYED
Outlier Count	2295	1063	1188	2387	723	9076

Strategy for outlier handling

Column YEARS_EMPLOYED, we can see people being employed for 1000 yrs which is beyond human capacity. We will need to correct it.

Column CNT_CHILDREN shows people are having 11 children which is not impractical but rare in normal situations. This will require data validation again.

AMT_INCOME_TOTAL one of the extreme outlier is 117000000 but we will not remove it because income of people may have different figures and it could be a real case too.

AMT_CREDIT and AMT_INCOME_TOTAL where amount is higher than the entire usual trend. We need to verify it again.

We will not remove outlier from AMT_CREDIT too as it may be one of the actual cases.

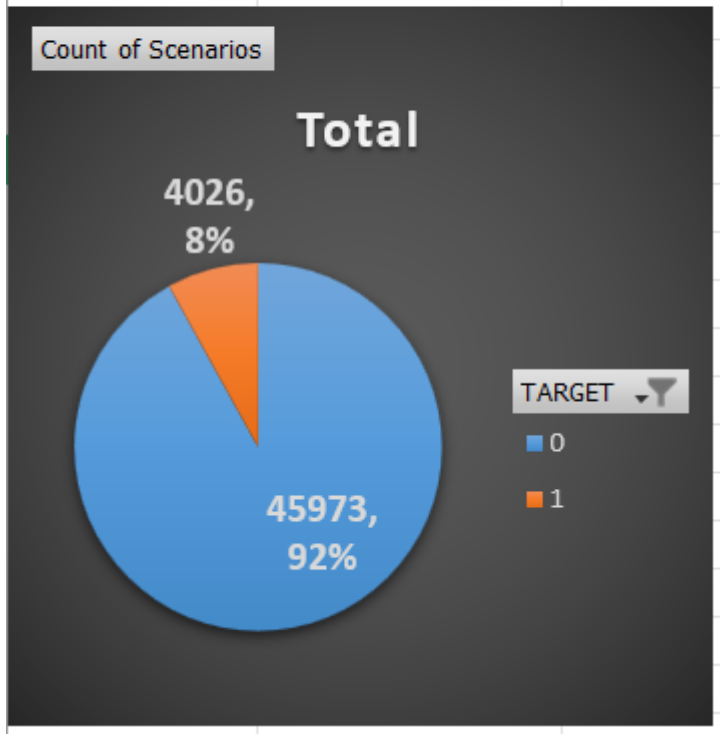
AMT_GOODS_PRICE also shows outliers but we will first understand it from the clients and then decide on it.

Task 3

Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

Class imbalance based on ‘Target’ variable

Target	Count of Scenarios
0	45973
1	4026
Grand Total	**49999**



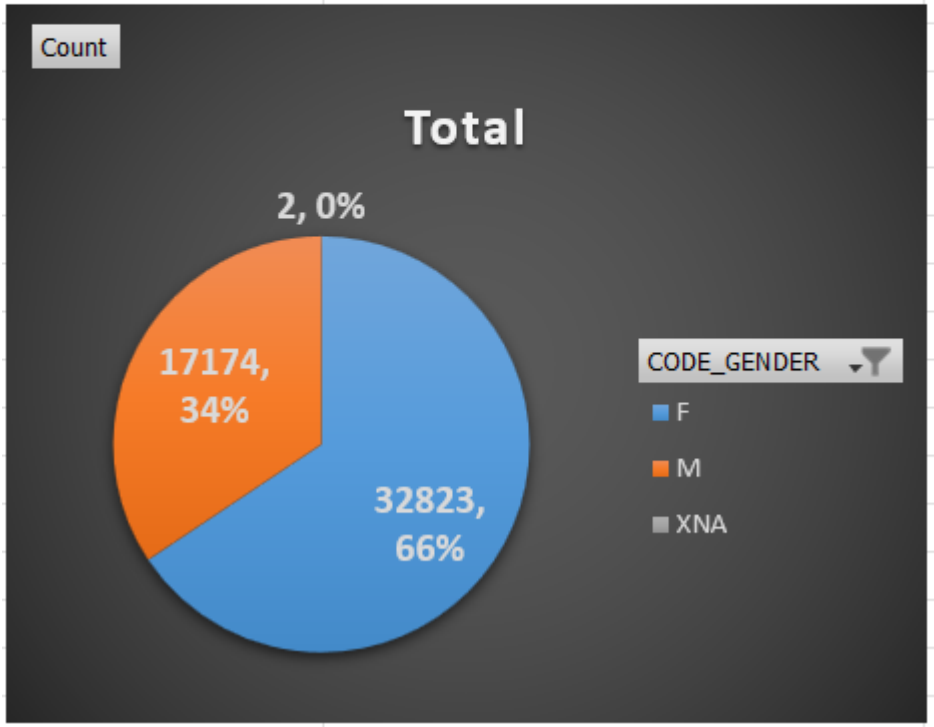
Interpretation

Almost 92% don’t fall under defaulters

8% are defaulters

Gender based observation

Gender	Count
F	32823
M	17174
XNA	2
Grand Total	**49999**



Interpretation

Females are the major debtors up to 66%
(Maybe the schemes are more liked by females)
Males account for 34%
A negligible count is of XNA

Project by – Alokk Joshi

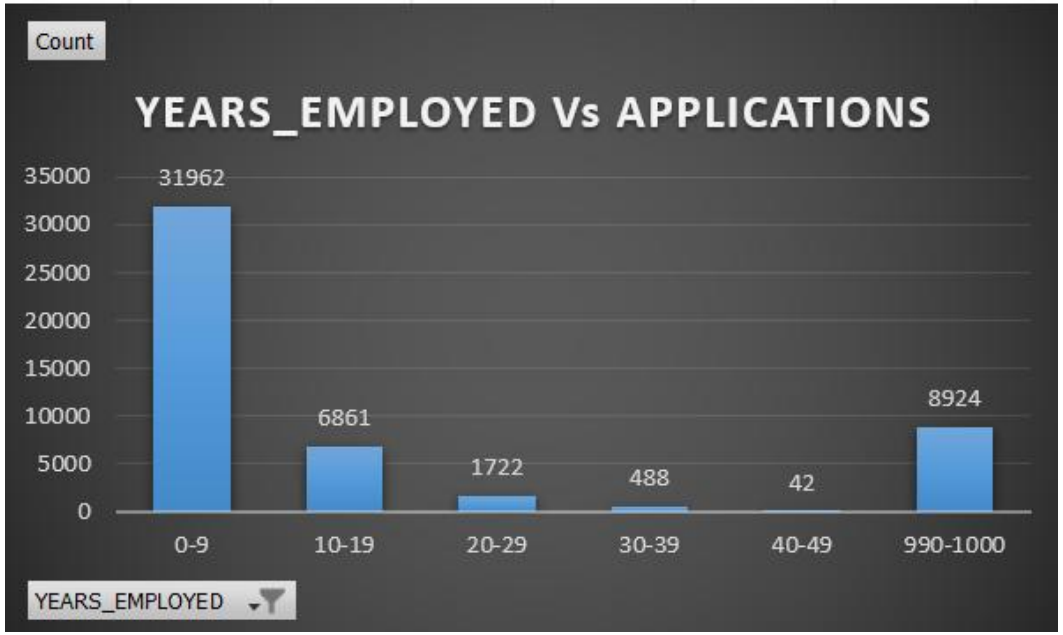
Task 4

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.

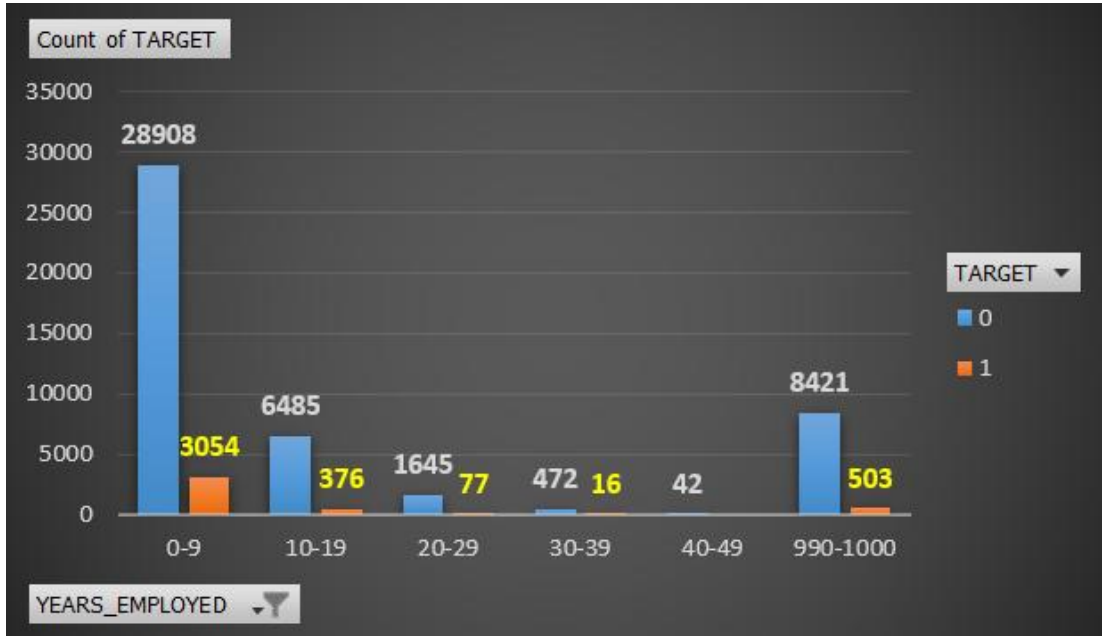
Project 6 – Bank Loan Case Study

Project by – Alokk Joshi

Duration of Employment of applicants Vs no. of applicants



Duration of Employment of applicants Vs no. of applicants & Repayment trend

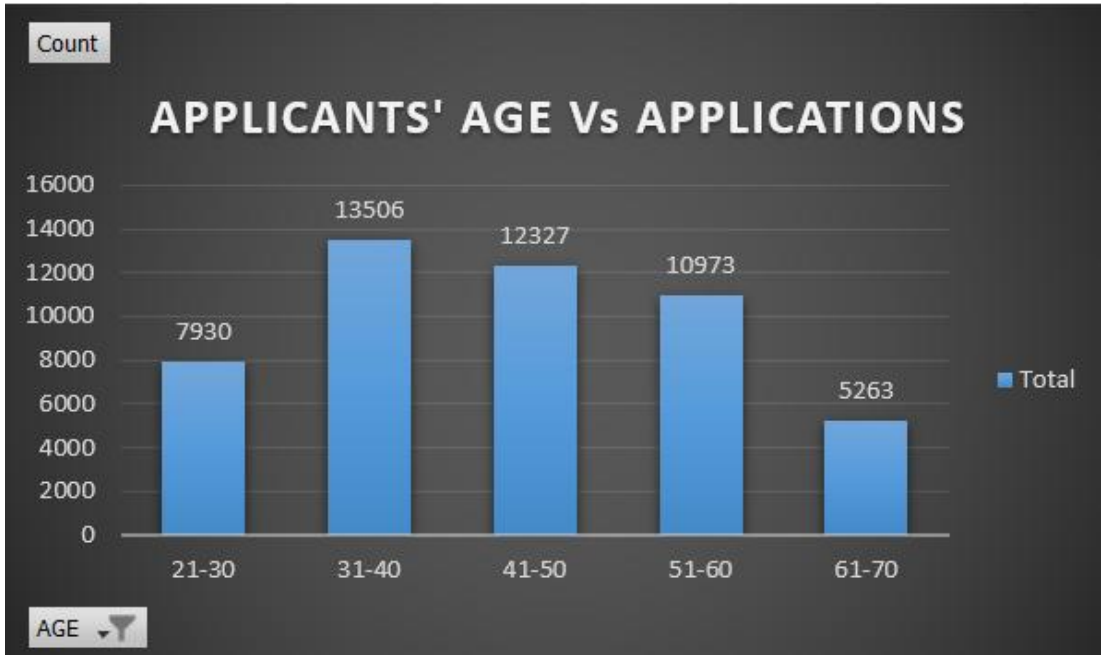


- Maximum applicants are from 0-9 years of employment range (brush off outliers here)
- As the employment age progresses, applicants have lower rate of defaulting on repayment

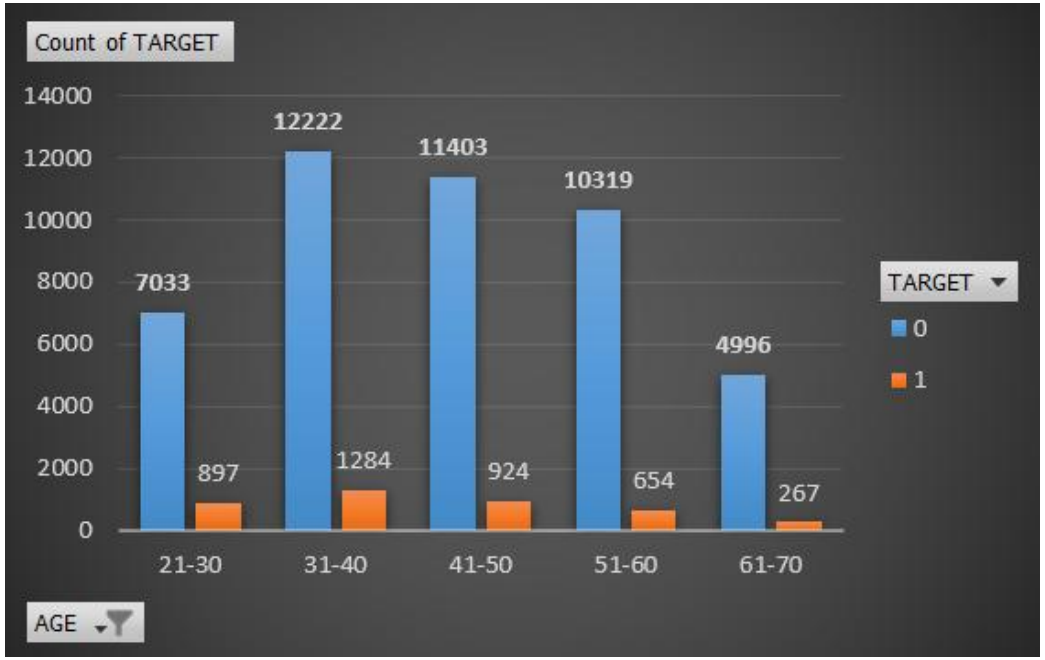
Project 6 – Bank Loan Case Study

Project by – Alokk Joshi

Applicant's age data Vs count



Duration of Employment of applicants Vs no. of applicants & Repayment trend



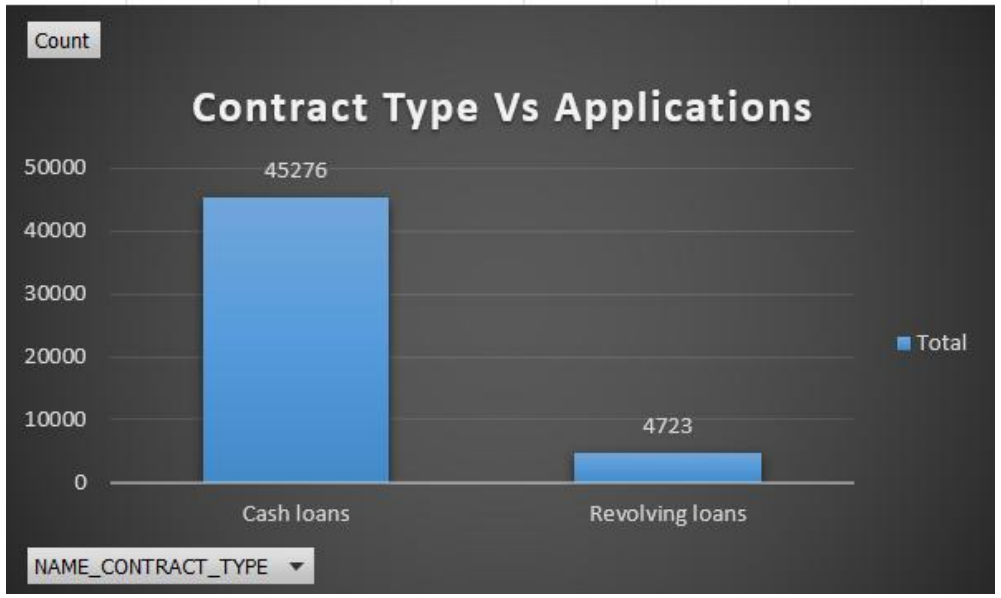
Majority of applicant's age is between 31-40
Applications decrease with age after 40
Repayment difficulty decreases with age

AGE Vs Repayment Trend				
AGE	0	1	0	1
21-30	7033	897	89%	11%
31-40	12222	1284	90%	10%
41-50	11403	924	93%	7%
51-60	10319	654	94%	6%
61-70	4996	267	95%	5%

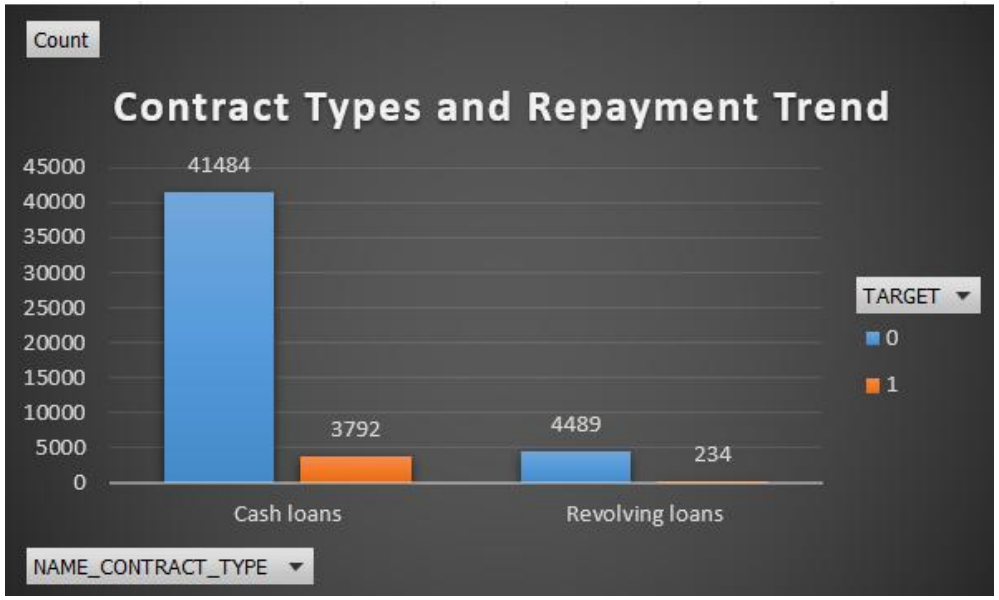
Project 6 – Bank Loan Case Study

Project by – Alokk Joshi

Types of loans and their frequency



Types of loans and repayment trend



Contract Type	0	1	0	1
Cash loans	41484	3792	92%	8%
Revolving loans	4489	234	95%	5%

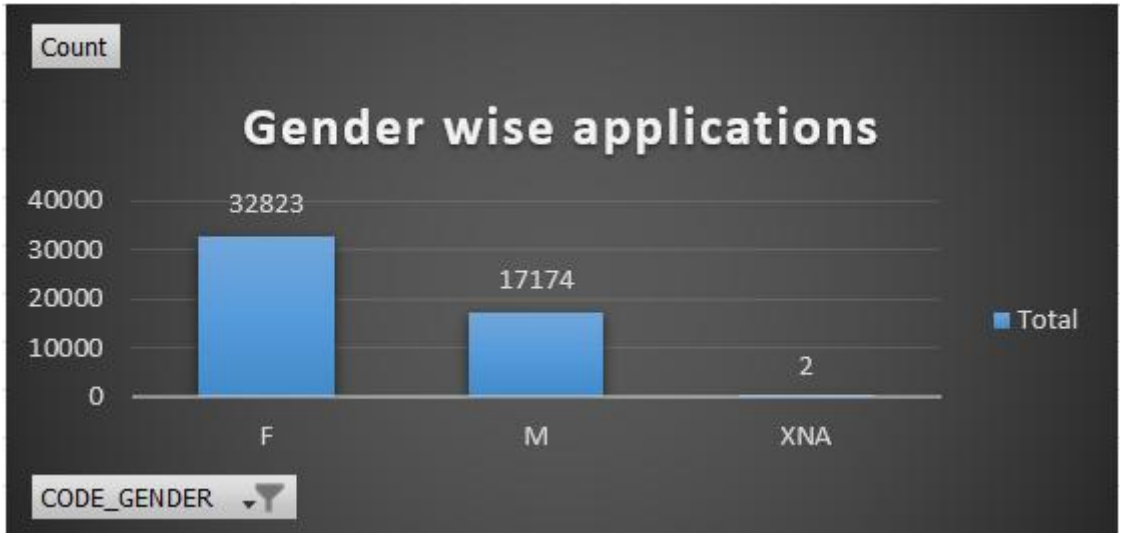
Cash loans are more in trend

Revolving loans have lesser default tendency, although the difference is huge between both

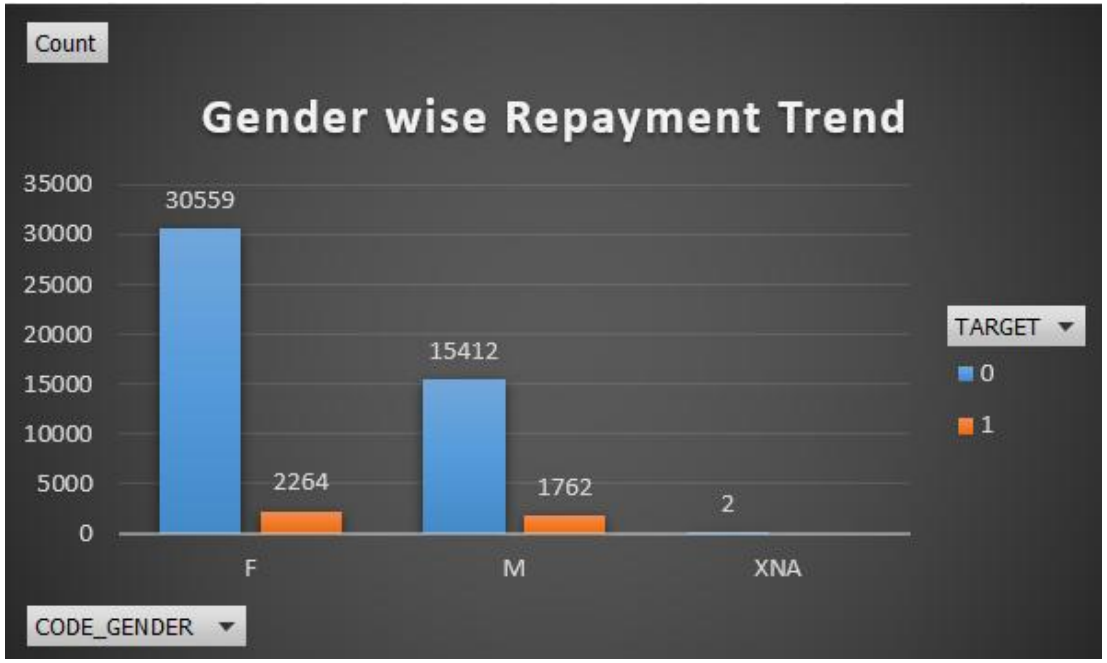
Project 6 – Bank Loan Case Study

Project by – Alokk Joshi

Gender wise application data



Gender wise application and repayment trend



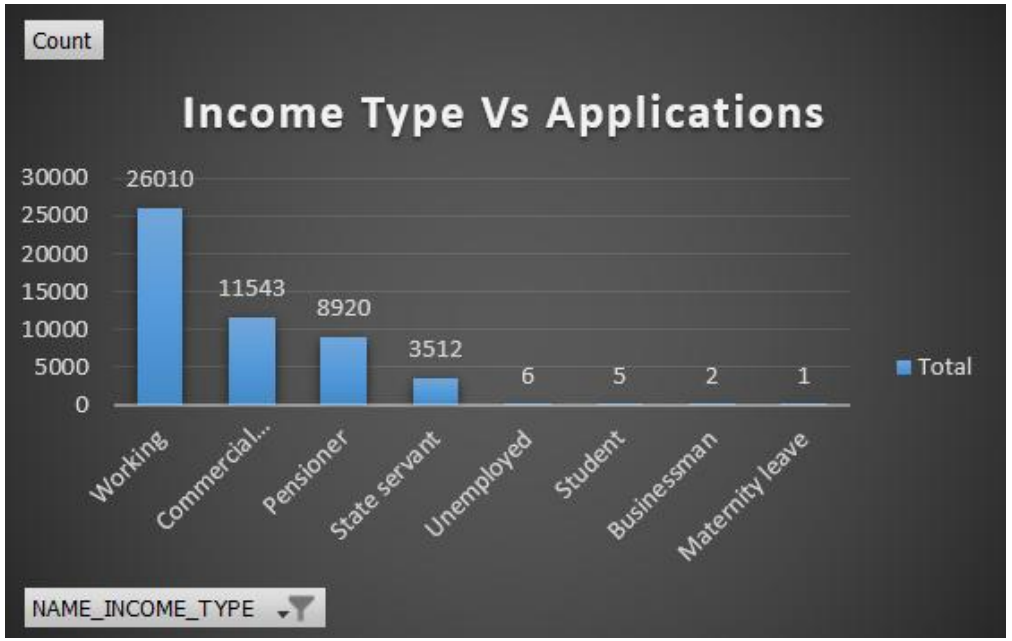
Females are major applicants
Females have slightly better repayment record than males

	Repayment Trend			
Gender	0	1	0	1
F	30559	2264	93%	7%
M	15412	1762	90%	10%
XNA	2			

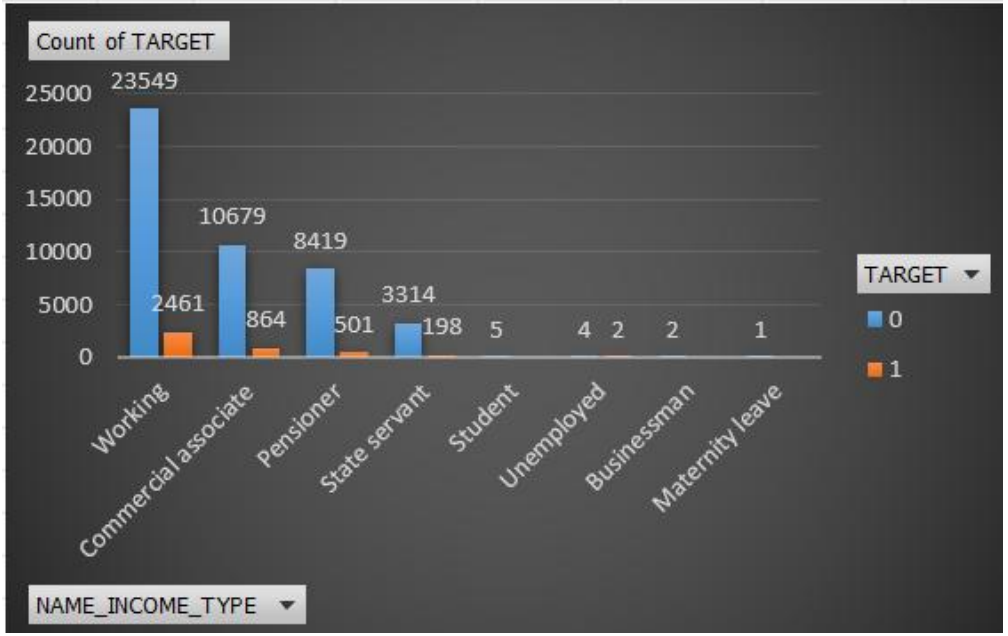
Project 6 – Bank Loan Case Study

Project by – Alokk Joshi

Income Type Vs Applications



Income Type application and repayment trend



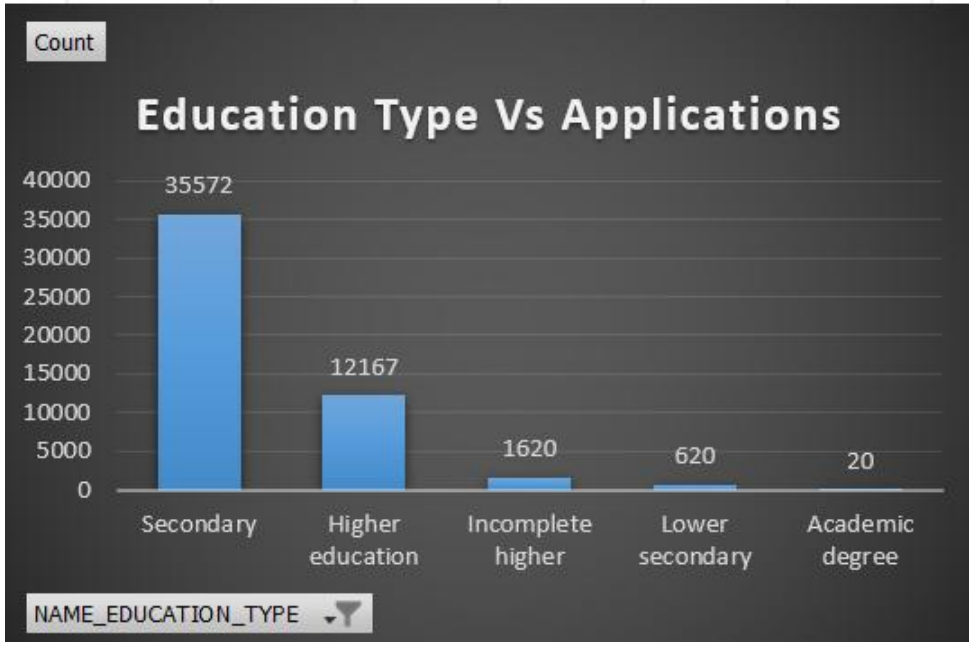
- The highest number of loan applications come from the working category followed by commercial associate and pensioner
- Students, businessmen, and maternity leave applicants have a 100% repayment rate
- Higher defaults among working applicants

Repayment Trend				
Income Type	0	1	0	1
Working	23549	2461	91%	9%
Commercial associate	10679	864	93%	7%
Pensioner	8419	501	94%	6%
State servant	3314	198	94%	6%
Student	5		100%	0%
Unemployed	4	2	67%	33%
Businessman	2		100%	0%
Maternity leave	1		100%	0%

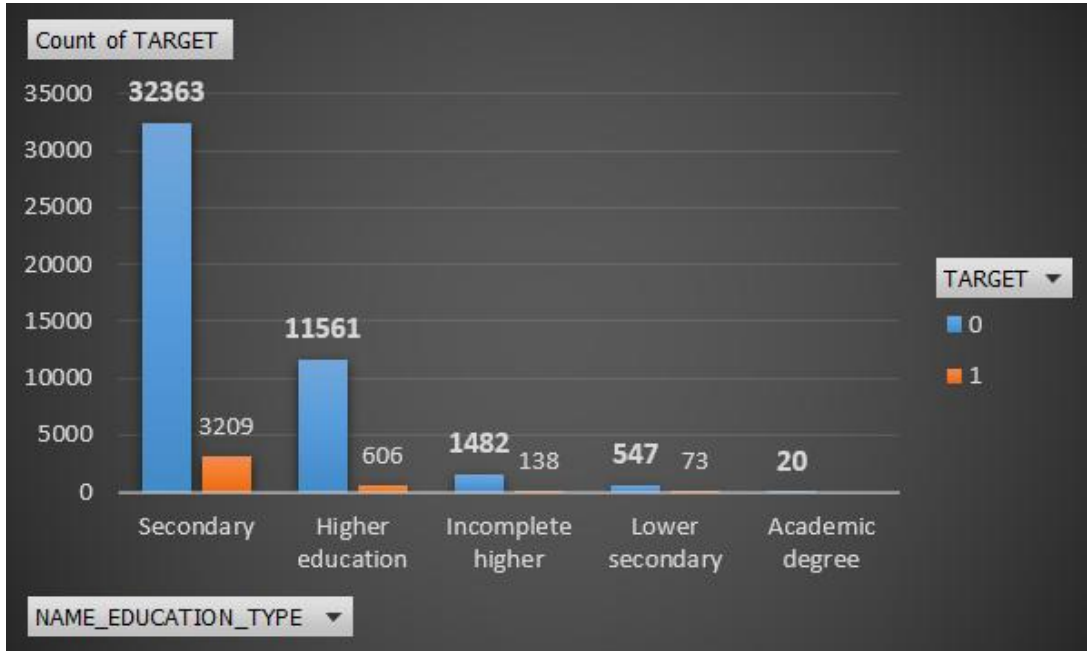
Project 6 – Bank Loan Case Study

Project by – Alokk Joshi

Education Type Vs Applications



Education Type application and repayment trend



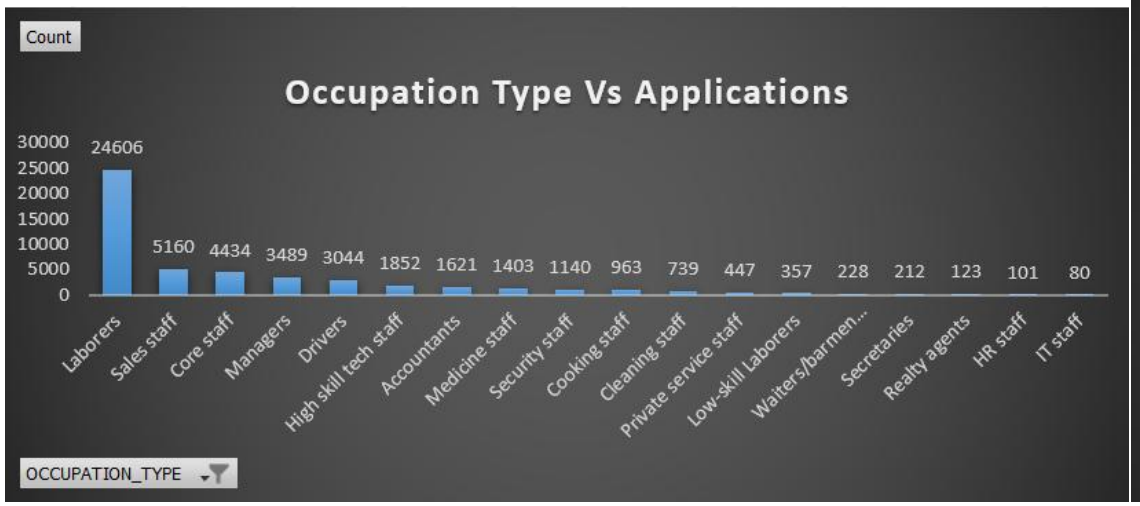
Repayment Trend				
Education Type	0	1	0	1
Secondary	32363	3209	266%	26%
Higher education	11561	606	95%	5%
Incomplete higher	1482	138	91%	9%
Lower secondary	547	73	88%	12%
Academic degree	20		100%	0%

- Secondary education dominates loan applications
- Higher education borrowers show better repayment behavior, with a 95% repayment rate
- Those with an academic degree have a 100% repayment rate, but their numbers are extremely low (only 20 applicants), making it less impactful overall.

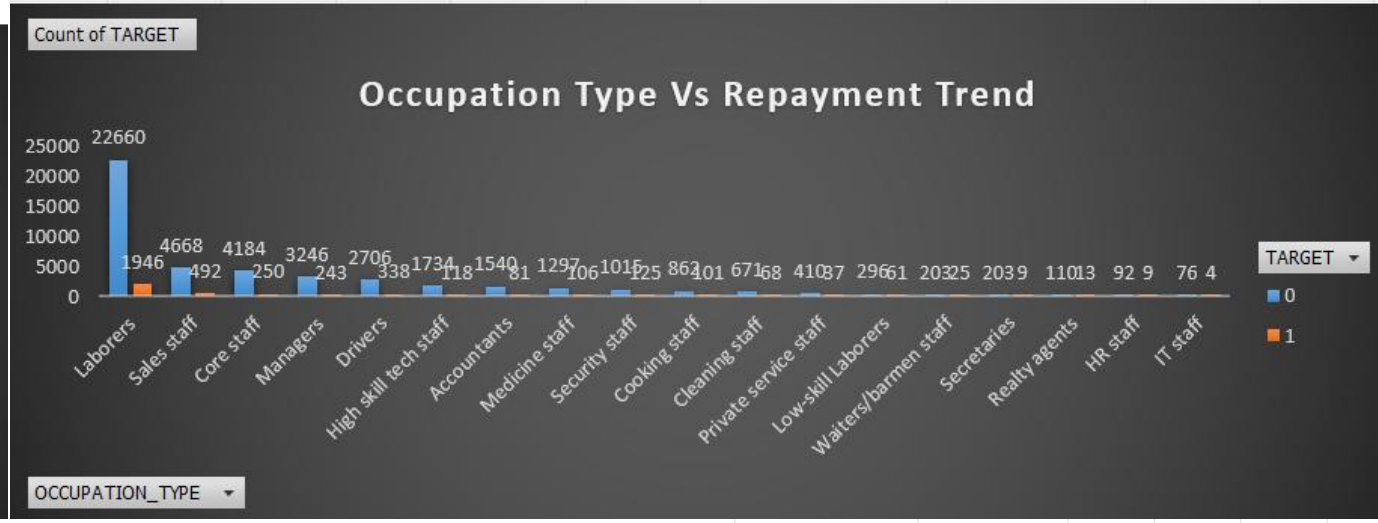
Project 6 – Bank Loan Case Study

Project by – Alokk Joshi

Occupation Type Vs Applications



Occupation Type application and repayment trend



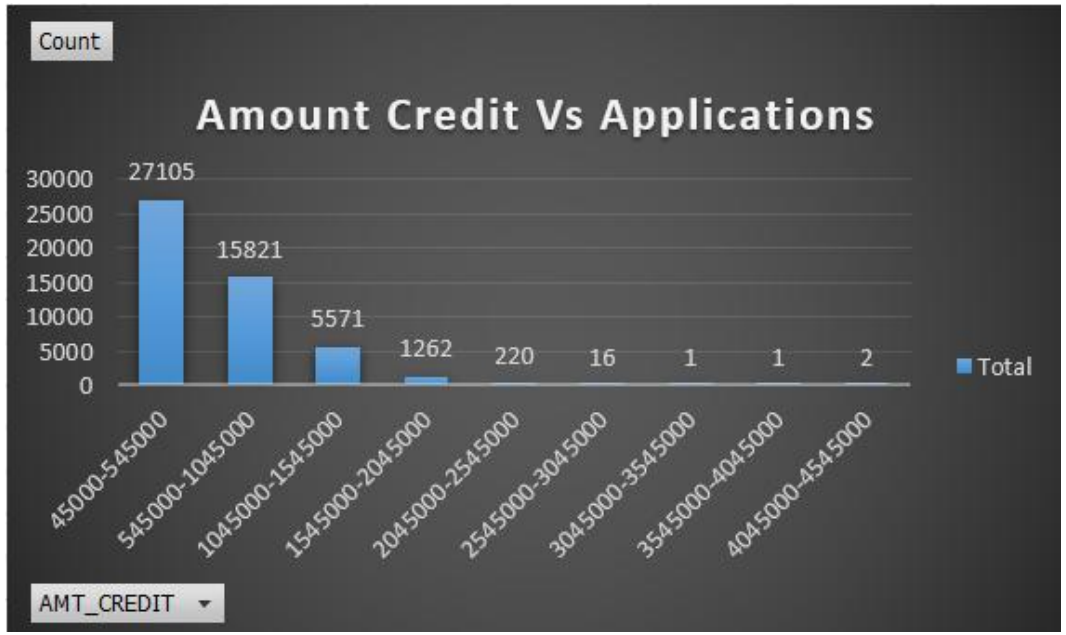
- Laborers form the largest group of borrowers (22,660) but have a higher default rate (8%), indicating financial instability in this segment.
- IT staff and accountants show the best repayment behavior, with 95% and 91% repayment rates, respectively, suggesting stable and well-paying jobs.
- Waiters/barmen and low-skill laborers have the highest default rates (11% and 17%), making them riskier borrower groups compared to others.

	Repayment Trend			
Occupation Type	0	1	0	1
Laborers	22660	1946	92%	8%
Sales staff	4668	492	90%	10%
Core staff	4184	250	94%	6%
Managers	3246	243	93%	7%
Drivers	2706	338	89%	11%
High skill tech staff	1734	118	94%	6%
Accountants	1540	81	95%	5%
Medicine staff	1297	106	92%	8%
Security staff	1015	125	89%	11%
Cooking staff	862	101	90%	10%
Cleaning staff	671	68	91%	9%
Private service staff	410	37	92%	8%
Low-skill Laborers	296	61	83%	17%
Waiters/barmen staff	203	25	89%	11%
Secretaries	203	9	96%	4%
Realty agents	110	13	89%	11%
HR staff	92	9	91%	9%
IT staff	76	4	95%	5%

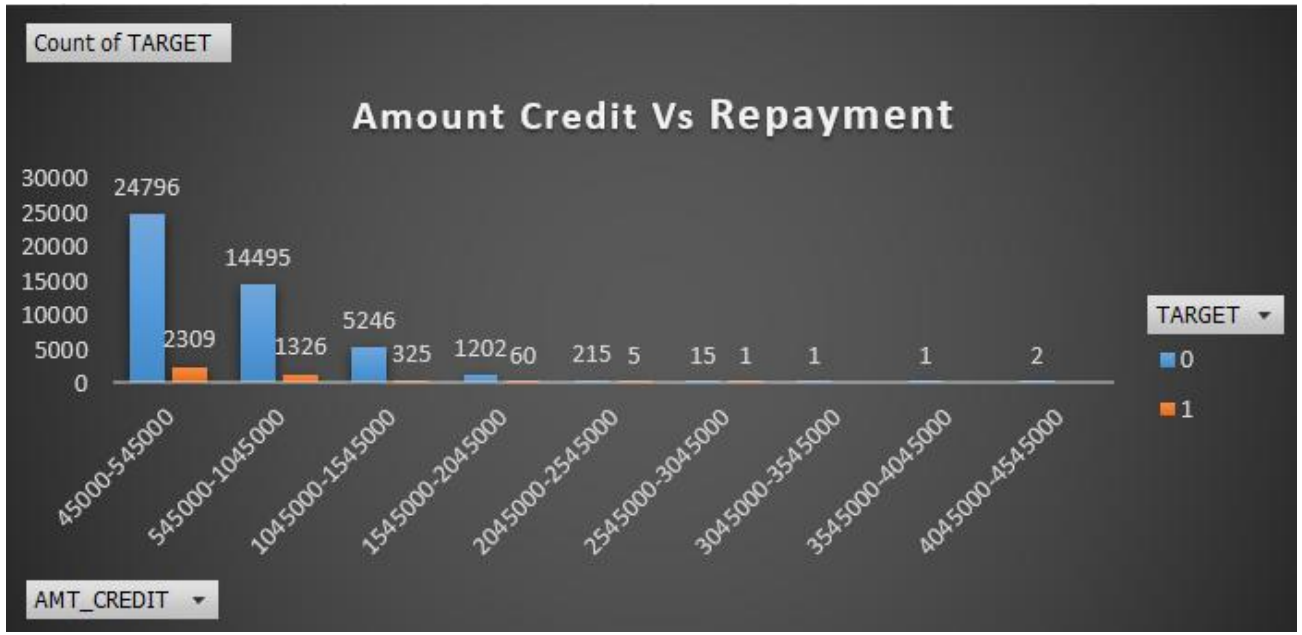
Project 6 – Bank Loan Case Study

Project by – Alokk Joshi

Amount Credit Vs Applications



Amount Credit Vs Applications



Insights in the next slide

Project 6 – Bank Loan Case Study

Project by – Alokk Joshi

• Lower credit amounts (45,000 - 5,45,000) have the highest number of borrowers but also show a higher default rate (9%), indicating financial strain on lower-credit borrowers.

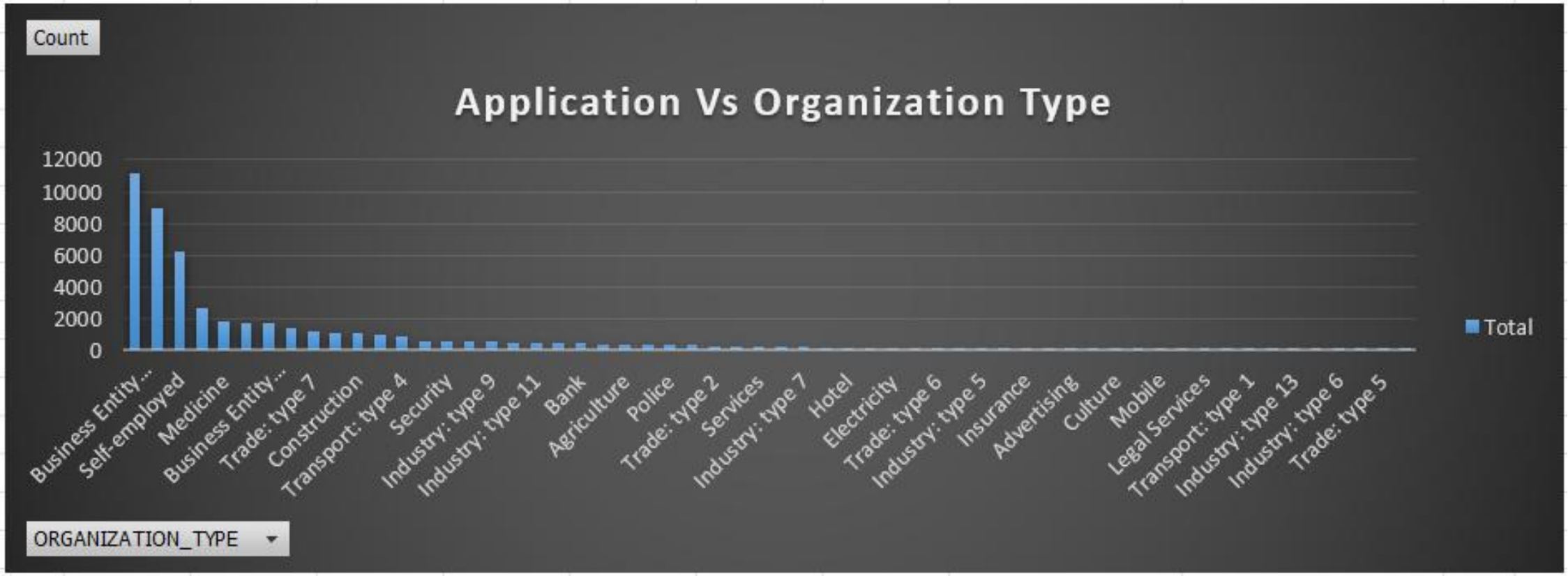
• As the credit amount increases, the repayment rate improves, with loans between 10,45,000 - 20,45,000 showing a default rate of only 5-6%, suggesting that higher-credit borrowers are more financially stable.

• Loans above 30,45,000 have a 100% repayment rate, showing that high-credit borrowers are highly reliable in loan repayments.

	Repayment			
Amount Credit	0	1	0	1
45000-545000	24796	2309	91%	9%
545000-1045000	14495	1326	92%	8%
1045000-1545000	5246	325	94%	6%
1545000-2045000	1202	60	95%	5%
2045000-2545000	215	5	98%	2%
2545000-3045000	15	1	94%	6%
3045000-3545000	1		100%	0%
3545000-4045000	1		100%	0%
4045000-4545000	2		100%	0%

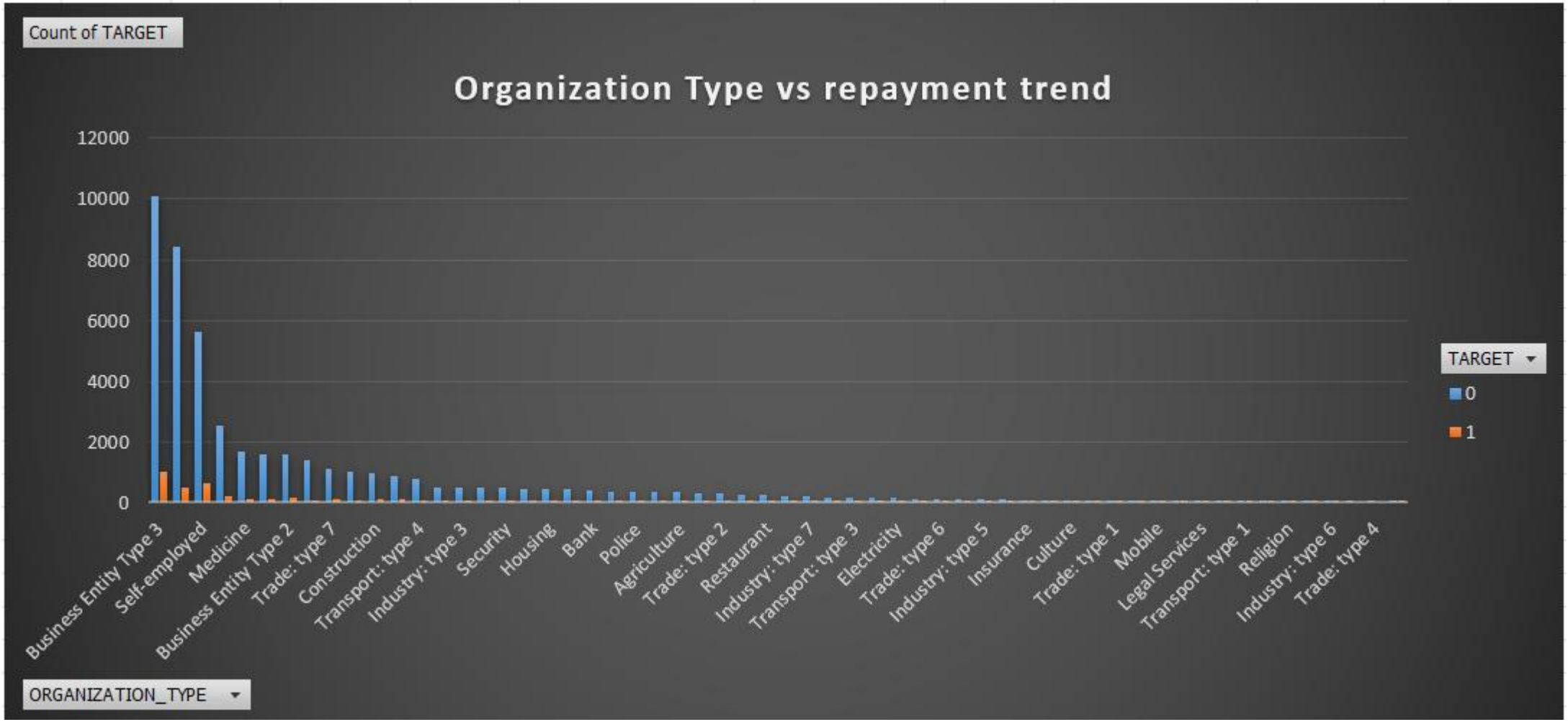
Project by – Alokk Joshi

Organization Type Vs Applications



Project by – Alokk Joshi

Organization Type Vs Repayment Trend



Project by – Alokk Joshi

Major insights

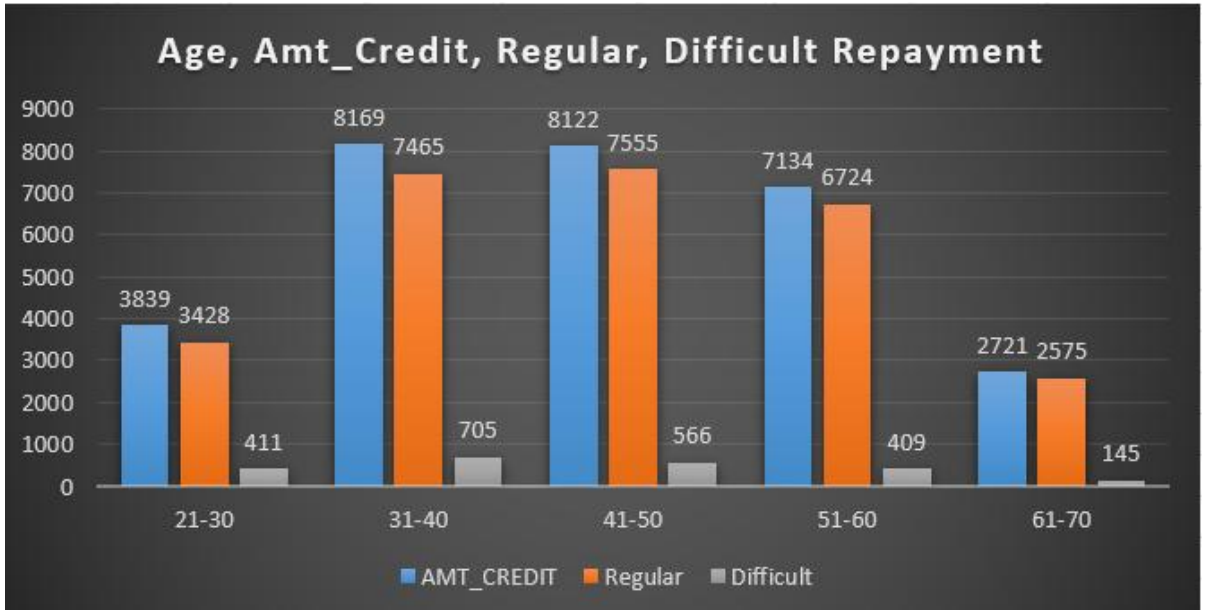
• Government and stable institutions (like Schools, Police, Universities) show the highest repayment rates (95%+), indicating that employees in these sectors have more financial stability.

• Self-employed individuals, construction workers, and trade professionals show lower repayment rates (around 90%) with higher default risks (10%+), highlighting financial unpredictability in these sectors.

• Certain industries like Agriculture (87%), Transport (87%), and Industry type 13 (73%) have the highest default rates, signaling financial instability and potential risk for lenders.

Project 6 – Bank Loan Case Study

Project by – Alokk Joshi



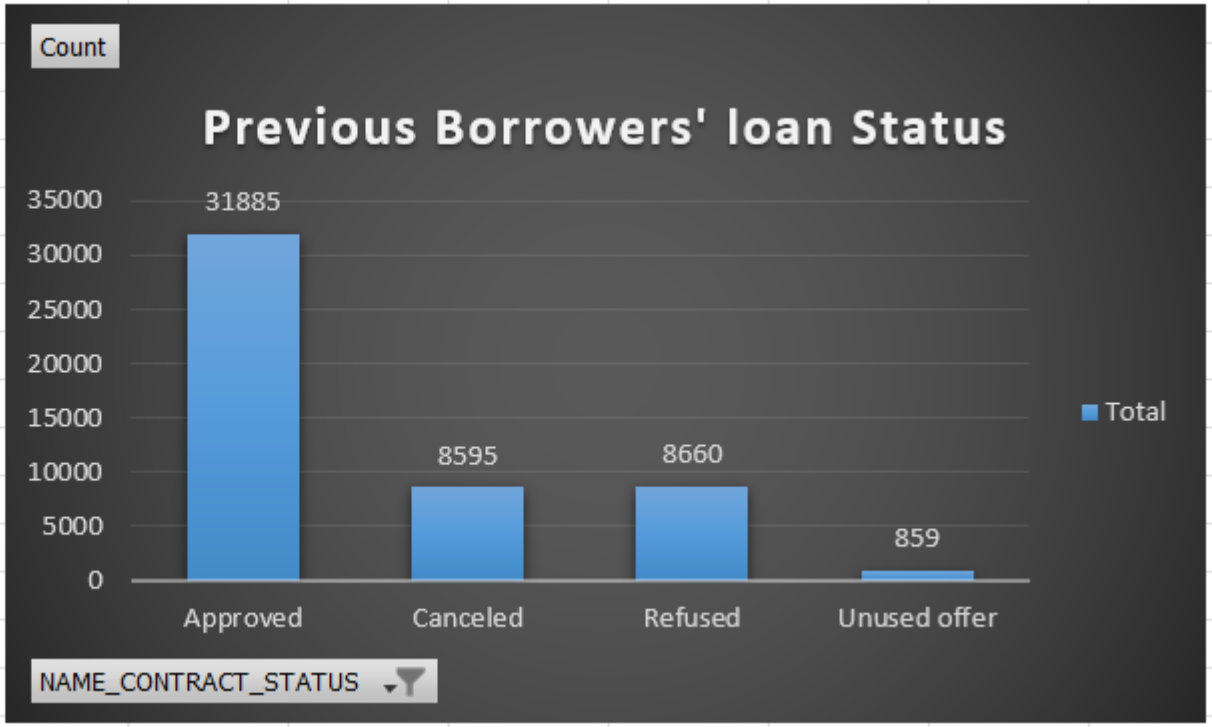
Key Insights:

- Credit amounts peak in the 31-50 age range, aligning with prime working years.
- Regular repayment trends follow a similar pattern, meaning most loans are repaid on time.
- Difficult repayment cases are higher among younger(31-50) borrowers, possibly due to unstable income sources.
- Older borrowers (61-70) take lower credit and have fewer repayment issues, likely because they are more financially stable or take conservative loans.

From Previous Application File

Previous Applications' loan status

Previous Borrowers	Count
Approved	31885
Canceled	8595
Refused	8660
Unused offer	859
Grand Total	**49999**



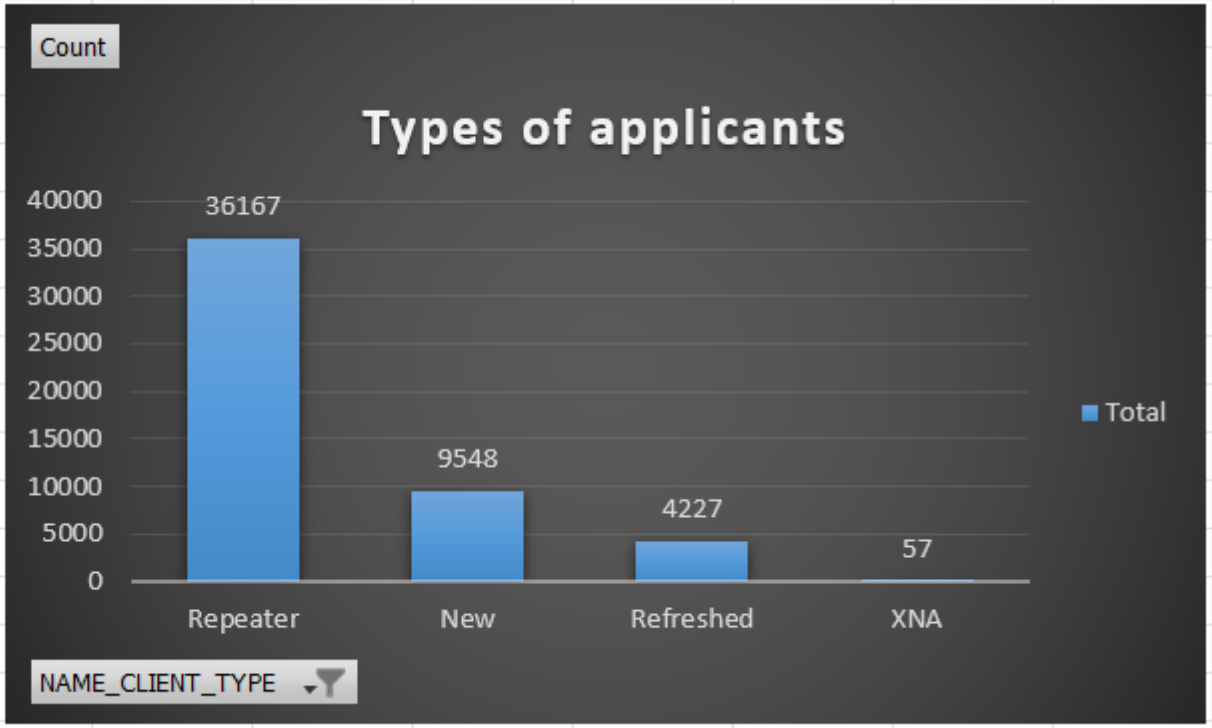
Higher number of previous applicants' loans have been approved

859 have not used their loans

8660 have refused to borrow

Previous Applications' loan status

Client Type	Count
Repeater	36167
New	9548
Refreshed	4227
XNA	57
Grand Total	**49999**



Majority of the borrowers are repeaters. Focusing on old borrowers should be focused in future business.

Here is how the application amounts were processed in previous applications

Applied Amount - Credit Amount	Count
Approved more than applied	19546
Approved less than applied	11418
Refused	9600
Not Applied	9435
Total	49999

It shows that majority the borrowers got more credit than they applied for.

Task 5

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.

Regular Repayment Correlation - 0		
AMT_CREDIT	AMT_ANNUITY	**0.770772818**
AMT_CREDIT	AMT_GOODS_PRICE	**0.986999774**
NAME_INCOME_TYPE	YEARS_EMPLOYED	**0.797293628**
CNT_CHILDREN	CNT_FAM_MEMBERS	**0.879238049**
AMT_ANNUITY	AMT_GOODS_PRICE	**0.775835204**
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	**0.950468157**
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	**0.861374946**
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	**0.825358079**
YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_MODE	**0.973531781**
YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_MEDI	**0.994674497**
YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_MEDI	**0.966738323**
FLOORSMAX_MODE	FLOORSMAX_MEDI	**0.98958882**
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	**0.998357563**
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	**0.850995792**

Project 6 – Bank Loan Case Study

Project by – Alokk Joshi

•**Employment Stability Matters** – "Years Employed" has a strong correlation (**0.7973**) with repayment, indicating that stable employment leads to better repayment behavior.

•**Higher Credit Amounts Align with Higher Annuities** – "AMT_CREDIT" and "AMT_ANNUITY" have a high correlation (**0.7708**), suggesting that larger loan amounts correspond with higher installment payments.

•**Social Circle and Default Monitoring Are Strongly Linked** – "OBS_30_CNT_SOCIAL_CIRCLE" and "OBS_60_CNT_SOCIAL_CIRCLE" show extremely high correlation (**0.9984**), meaning people who default in one short-term period are highly likely to default in another.

•**Property Age and Loan Repayment Connection** – The strong correlation of "YEARS_BEGINEXPLUATATION_AVG" (**0.9947**) with other exploitation-related variables suggests that property age or ownership history significantly impacts loan repayment behavior.

Difficulty Repayment Correlation - 1		
AMT_CREDIT	AMT_ANNUITY	**0.749665201**
AMT_CREDIT	AMT_GOODS_PRICE	**0.982267963**
CNT_CHILDREN	CNT_FAM_MEMBERS	**0.892521875**
AMT_ANNUITY	AMT_GOODS_PRICE	**0.74950403**
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	**0.950768899**
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	**0.806743886**
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	**0.783754676**
YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_MODE	**0.969745206**
YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_MEDI	**0.983626828**
YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_MEDI	**0.979592562**
FLOORSMAX_MODE	FLOORSMAX_MEDI	**0.989772825**
FLOORSMAX_AVG	FLOORSMAX_MODE2	**0.987677718**
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	**0.998065853**
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	**0.89051161**

Project 6 – Bank Loan Case Study

Project by – Alokk Joshi

•**Loan Amount and Goods Price Are Highly Correlated** – "AMT_CREDIT" and "AMT_GOODS_PRICE" have an **extremely high correlation (0.9823)**, suggesting that borrowers struggling with repayment often take loans close to the exact value of their purchased goods, leaving little financial flexibility.

•**Larger Families Struggle More** – "CNT_CHILDREN" and "CNT_FAM_MEMBERS" show a strong correlation (**0.8925**), indicating that borrowers with larger families might face financial strain, making repayment more difficult.

•**Living and Working in Different Regions Affects Repayment** – The correlation between "REG_REGION_NOT_WORK_REGION" and "LIVE_REGION_NOT_WORK_REGION" (**0.8067**) suggests that people who live far from their workplace may have increased financial burdens, impacting their ability to repay loans.

•**Property Characteristics Strongly Influence Repayment** – The correlation of "YEARS_BEGINEXPLUATATION_AVG" with different property condition measures (**ranging from 0.9697 to 0.9836**) indicates that property age and condition significantly impact repayment behavior, possibly because older or lower-value properties are linked to financial instability.

•**Social Default Risk Is Highly Predictable** – "OBS_30_CNT_SOCIAL_CIRCLE" and "OBS_60_CNT_SOCIAL_CIRCLE" have an almost perfect correlation (**0.9980**), meaning that borrowers who struggle with repayment in the short term are almost certain to continue struggling in the longer term. Additionally, "DEF_30_CNT_SOCIAL_CIRCLE" and "DEF_60_CNT_SOCIAL_CIRCLE" (**0.8905**) reinforce this pattern, showing that social circles with defaulters are a strong indicator of financial distress.

Project by – Alokk Joshi

Major Insights from the entire analytics

1. The majority of clients take cash loans.
2. Most clients are loan repayers rather than defaulters.
3. The bank lends more to women than men, but women have a lower default rate compared to men.
4. Clients with higher education levels are less likely to default compared to those with lower education, such as those with only secondary special education.
5. The bank should prioritize lending to clients with higher educational qualifications.
6. As age and experience increase, the likelihood of default decreases.
7. Older clients tend to borrow larger loan amounts, but their default rate is lower, making them less risky and more profitable for the bank.
8. As the number of children increases, the number of clients taking loans decreases.
9. The bank should exercise extra caution when lending to unemployed clients, as they have the highest default rate and take larger loan amounts.

My Learnings from the project

1. Honed Data Cleaning Techniques – Learned how to handle missing values, remove duplicates, and standardize data for accurate analysis in Excel.
2. Applied Correlation Analysis – Understood how to identify relationships between variables to draw meaningful insights from loan repayment patterns.
3. Enhanced Data Visualization Skills – Used pivot tables, conditional formatting, and charts to present complex data in a simplified and actionable manner.
4. Developed Logical Problem-Solving Approach – Improved my ability to structure data-driven questions and derive conclusions using analytical thinking.
5. Strengthened Statistical Analysis in Excel – Gained hands-on experience in using formulas like COUNTIF, UNIQUE, FILTER and VISUALIZATION tools to analyze trends.