BANK LOAN CASE STUDY

Trainity Project No. 6

Prepared By:
Nilesh Kulkarni

AGENDA

- Project Description
- Project Approach
- Tech Stack Used
- Insights/Tasks
- Results

Prepared By:
Nilesh Kulkarni

Project Description

With given bank Loan applicants metadata, in this project, the aim to analyze a loan application data to gain insights into factors influencing loan defaults. The project involves analyzing loan applicant's data and find out relations and factors influencing loan defaults.

The project involves below mentioned data analytics tasks:

A. **Identify Missing Data and Deal with it Appropriately** – Identify blank/nulls, chart them in visuals and come up with strategy to handle those scenarios.
B. **Identify Outliers in the Dataset** – Identify outliers using Quartile/IQR and Excel functions as well as relevance to business domain
C. **Analyse Data Imbalance** – based on distribution of target variable
D. **Perform Univariate, Segmented Univariate, and Bivariate Analysis** – Useful to gain insights into loan defaults, analyse various loan attributes
E. **Identify Top Correlations for Different Scenarios** – identify top relationship attributes among given loan data

For datasets, Please refer to drive link:
Application-Data
Previous-Application-Data

Prepared By:
Nilesh Kulkarni

Project Approach

High level steps for the Project approach are as outlined below:

- **Data cleaning/preparation:** Work on identifying missing data and strategy to handle missing data (imputation/drop rows or columns) as appropriate
- **Data Analysis:** Understand the given data set and its attributes, identify the outliers based on Quartile/Box-Plot as well as relevance to business domain.
- **Insights Analysis**: Analyze each insights requirement in detail and prepare MS Excel formula or functions to extract insights. Select optimal and efficient approach.
- **Extract insights:** Use MS Excel as a tool to extract new insights as required including visuals/charts.
- **Review:** Review and cross check output to verify it matches with the requirements/insights required
- **Document:** Document the insights and results to be shared across business teams

Prepared By:
Nilesh Kulkarni

Tech Stack used

Data Analytics tool: Microsoft Excel (Office 365)
Bank Loan data has been provided in Excel format and excel is further used for data cleaning, analysis and creating visuals/charts to demonstrate insights. Excel is user friendly and functionally rich tool to analyze, visualize and report the data insights.

Operating System: Microsoft Windows 11 Version 22H2

Documentation: Microsoft office 365 (Power Point) & Acrobat PDF

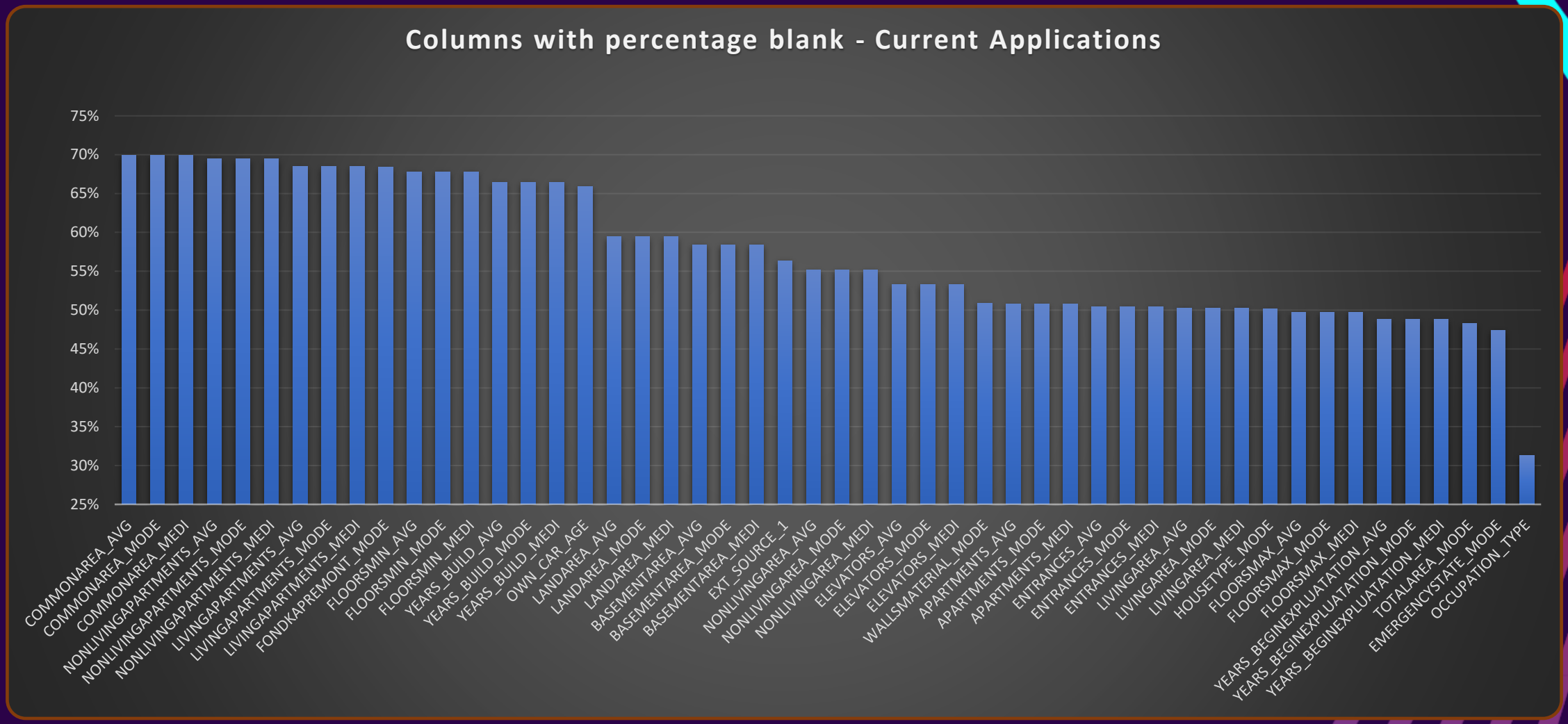
Prepared By:
Nilesh Kulkarni

Insights/Tasks

Prepared By:
Nilesh Kulkarni

Task A: Identify Missing Data and Deal with it Appropriately-Identify the missing data in the dataset & decide on an appropriate method to deal with it using Excel built-in functions and features.

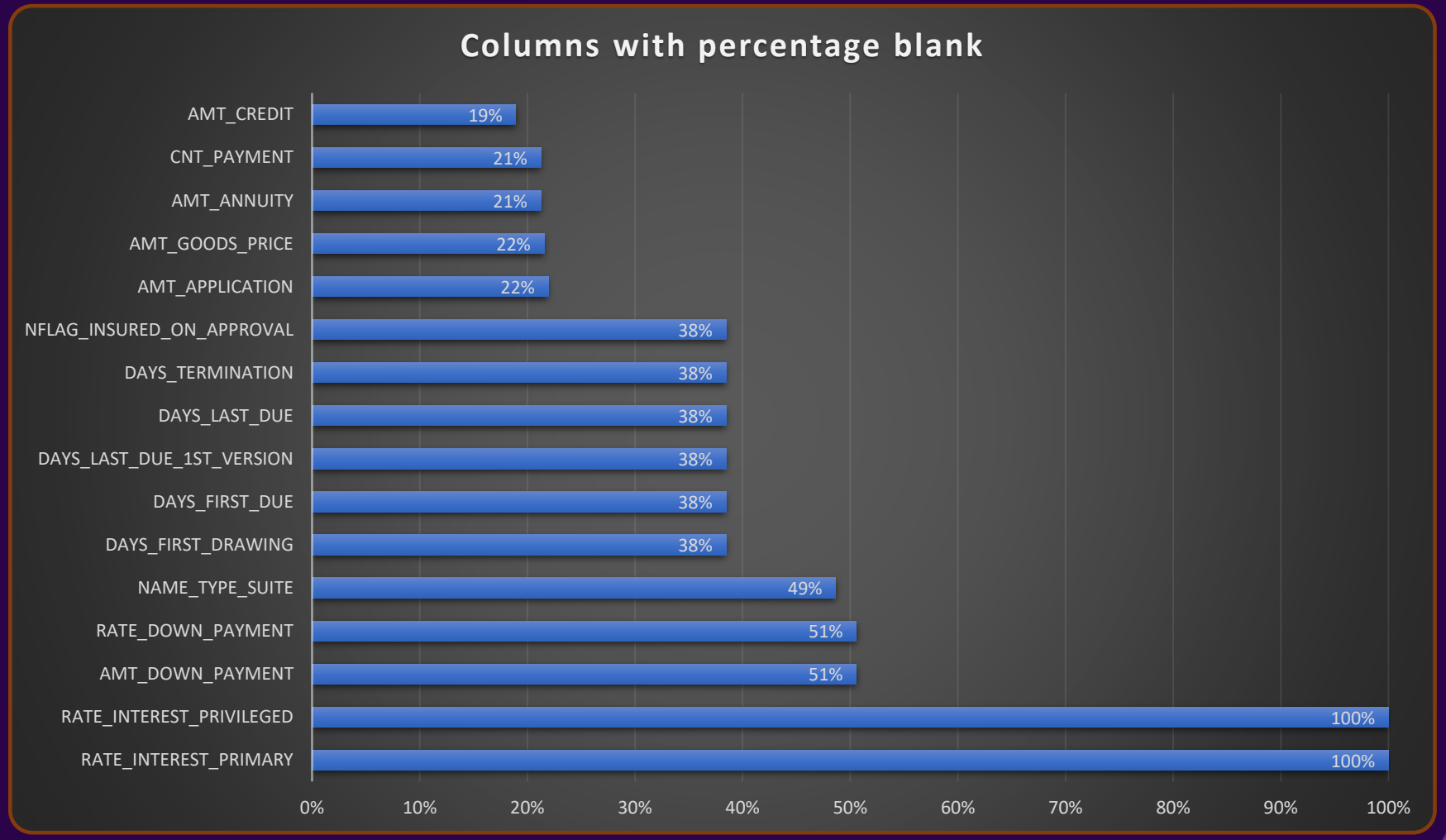
Application data has 50 columns having more than 30% blank rows



Prepared By:
Nilesh Kulkarni

Task A: Identify Missing Data and Deal with it Appropriately-Identify the missing data in the dataset & decide on an appropriate method to deal with it using Excel built-in functions and features.

Previous Application data has 16 columns having more than 15% blank rows



Prepared By:
Nilesh Kulkarni

Task A: Identify Missing Data and Deal with it Appropriately-Identify the missing data in the dataset & decide on an appropriate method to deal with it using Excel built-in functions and features.

Following 50 columns are deleted from current applications data since those having more than 30% rows as blank, potential to impact analysis due to blank rows:

1. COMMONAREA_AVG
2. COMMONAREA_MODE
3. COMMONAREA_MEDI
4. NONLIVINGAPARTMENTS_AVG
5. NONLIVINGAPARTMENTS_MODE
6. NONLIVINGAPARTMENTS_MEDI
7. LIVINGAPARTMENTS_AVG
8. LIVINGAPARTMENTS_MODE
9. LIVINGAPARTMENTS_MEDI
10. FONDKAPREMONT_MODE
11. FLOORSMIN_AVG
12. FLOORSMIN_MODE
13. FLOORSMIN_MEDI
14. YEARS_BUILD_AVG
15. YEARS_BUILD_MODE
16. YEARS_BUILD_MEDI
17. OWN_CAR_AGE
18. LANDAREA_AVG
19. LANDAREA_MODE
20. LANDAREA_MEDI
21. BASEMENTAREA_AVG
22. BASEMENTAREA_MODE
23. BASEMENTAREA_MEDI
24. EXT_SOURCE_1
25. NONLIVINGAREA_AVG
26. NONLIVINGAREA_MODE
27. NONLIVINGAREA_MEDI
28. ELEVATORS_AVG
29. ELEVATORS_MODE
30. ELEVATORS_MEDI
31. WALLSMATERIAL_MODE
32. APARTMENTS_AVG
33. APARTMENTS_MODE
34. APARTMENTS_MEDI
35. ENTRANCES_AVG
36. ENTRANCES_MODE
37. ENTRANCES_MEDI
38. LIVINGAREA_AVG
39. LIVINGAREA_MODE
40. LIVINGAREA_MEDI
41. HOUSETYPE_MODE
42. FLOORSMAX_AVG
43. FLOORSMAX_MODE
44. FLOORSMAX_MEDI
45. YEARS_BEGINEXPLUATATION_AVG
46. YEARS_BEGINEXPLUATATION_MODE
47. YEARS_BEGINEXPLUATATION_MEDI
48. TOTALAREA_MODE
49. EMERGENCYSTATE_MODE
50. OCCUPATION_TYPE

Prepared By:
Nilesh Kulkarni

Task A: Identify Missing Data and Deal with it Appropriately-Identify the missing data in the dataset & decide on an appropriate method to deal with it using Excel built-in functions and features.

Following 36 columns are deleted because those are irrelevant for the analysis from application data worksheet

1. FLAG_WORK_PHONE
2. FLAG_CONT_MOBILE
3. FLAG_PHONE
4. FLAG_DOCUMENT_2
5. FLAG_DOCUMENT_3
6. FLAG_DOCUMENT_4
7. FLAG_DOCUMENT_5
8. FLAG_DOCUMENT_6
9. FLAG_DOCUMENT_7
10. FLAG_DOCUMENT_8
11. FLAG_DOCUMENT_9
12. FLAG_DOCUMENT_10
13. FLAG_DOCUMENT_11
14. FLAG_DOCUMENT_12
15. FLAG_DOCUMENT_13
16. FLAG_DOCUMENT_14
17. FLAG_DOCUMENT_15
18. FLAG_DOCUMENT_16
19. FLAG_DOCUMENT_17
20. FLAG_DOCUMENT_18
21. FLAG_DOCUMENT_19
22. FLAG_DOCUMENT_20
23. FLAG_DOCUMENT_21
24. EXT_SOURCE_1
25. EXT_SOURCE_2
26. AMT_REQ_CREDIT_BUREAU_HOUR
27. AMT_REQ_CREDIT_BUREAU_DAY
28. AMT_REQ_CREDIT_BUREAU_WEEK
29. AMT_REQ_CREDIT_BUREAU_MON
30. AMT_REQ_CREDIT_BUREAU_QRT
31. AMT_REQ_CREDIT_BUREAU_YEAR
32. OBS_30_CNT_SOCIAL_CIRCLE
33. DEF_30_CNT_SOCIAL_CIRCLE
34. OBS_60_CNT_SOCIAL_CIRCLE
35. DEF_60_CNT_SOCIAL_CIRCLE
36. DAYS_LAST_PHONE_CHANGE

Prepared By:
Nilesh Kulkarni

Task A: Identify Missing Data and Deal with it Appropriately-Identify the missing data in the dataset & decide on an appropriate method to deal with it using Excel built-in functions and features.

Following 11 columns are deleted from previous applications data since those having more than 30% rows as blank, potential to impact analysis due to blank rows:

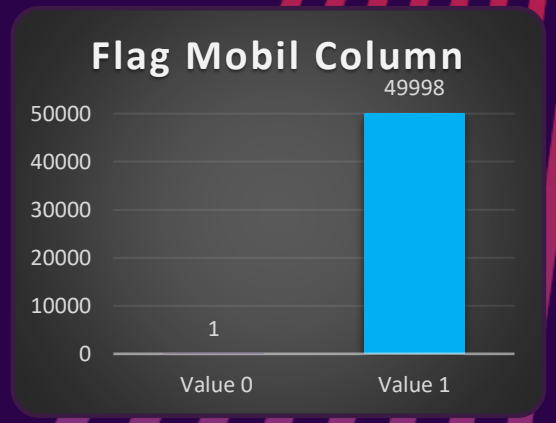
1. RATE_INTEREST_PRIMARY
2. RATE_INTEREST_PRIVILEGED
3. AMT_DOWN_PAYMENT
4. RATE_DOWN_PAYMENT
5. NAME_TYPE_SUITE
6. DAYS_FIRST_DRAWING
7. DAYS_FIRST_DUE
8. DAYS_LAST_DUE_1ST_VERSION
9. DAYS_LAST_DUE
10. DAYS_TERMINATION
11. NFLAG_INSURED_ON_APPROVAL

Prepared By:
Nilesh Kulkarni

Task A: Identify Missing Data and Deal with it Appropriately-Identify the missing data in the dataset & decide on an appropriate method to deal with it using Excel built-in functions and features.

Features enhancement/Data Preparation – Current Application Data

- ✓ Numeric columns set to median Values of rest of the data:
  - AMT_ANNUITY column has 1 row blank so set to median value 24939
  - AMT_GOODS_PRICE column has 38 blank rows so set to median value 450000
  - CNT_FAM_MEMBERS column has 1 row blank so set to median value 2
- ✓ Text/categorical column NAME_TYPE_SUITE has 192 rows blank so set to most occurring value ‘unaccompanied’ based on column data
- ✓ Introduced 3 new derived columns from data:
  - DAYS_BIRTH(Yrs) - calculated as number of years from days in column DAYS_BIRTH
  - DAYS_EMPLOYED(Yrs) - calculated as number of years from days of DAYS_EMPLOYED
  - DAYS_REGISTRATION(Yrs) – calculated as number of years from days of column DAYS_REGISTRATION
  - Also changed negative sign of above columns to no sign
- ✓ Column FLAG_MOBIL has all values set to 1 except one row with zero so deleted



Prepared By:
Nilesh Kulkarni

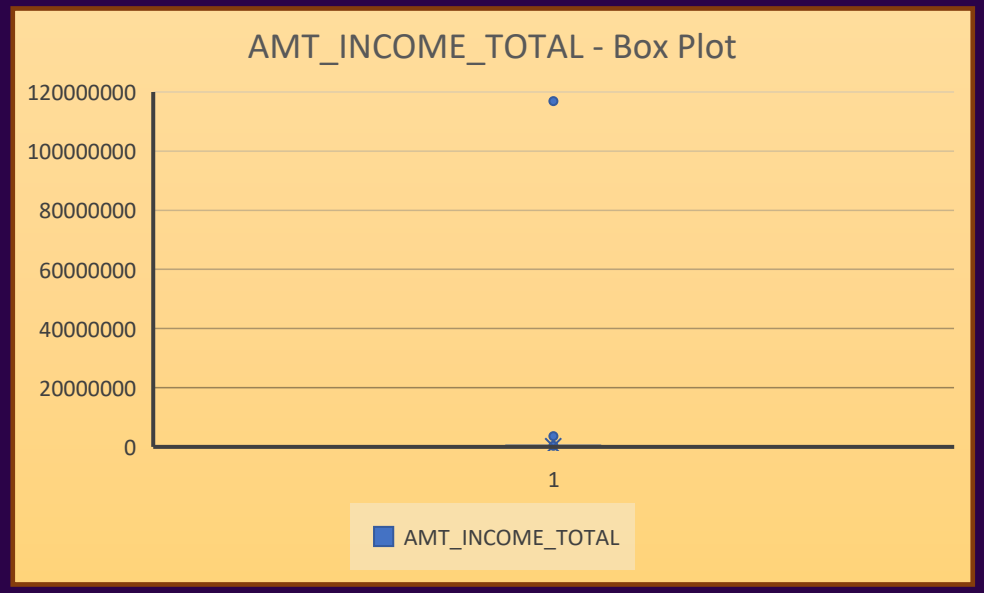
Task A: Identify Missing Data and Deal with it Appropriately-Identify the missing data in the dataset & decide on an appropriate method to deal with it using Excel built-in functions and features.

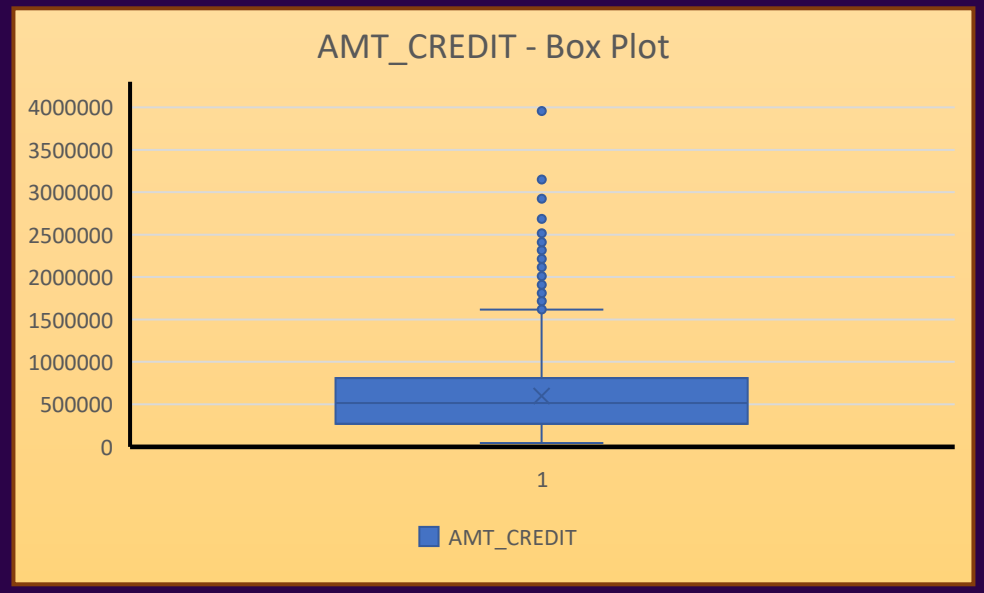
Features enhancement/Data Preparation – Previous Application Data

- ✓ Numeric columns set to median Values of rest of the data:
  - AMT_ANNUITY column has 10592 rows blank so set to median value 10879.92
  - AMT_GOODS_PRICE column has 10744 blank rows so set to median value 104017.5
  - CNT_PAYMENT columns has 10592 blank rows so set to median value 24
- ✓ AMT_APPLICATION and AMT_CREDIT has many columns with value as zero, since zero is valid numeric value – those have not been amended.
- ✓ Below listed columns were deleted because of their irrelevance to the data analysis:
  - WEEKDAY_APPR_PROCESS_START
  - HOUR_APPR_PROCESS_START
  - HOUR_APPR_PROCESS_START
- ✓ Previous application dataset kept separate since the combined dataset would be slow and performance intensive

Prepared By:
Nilesh Kulkarni

Task B: Identify Outliers in the Dataset: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.



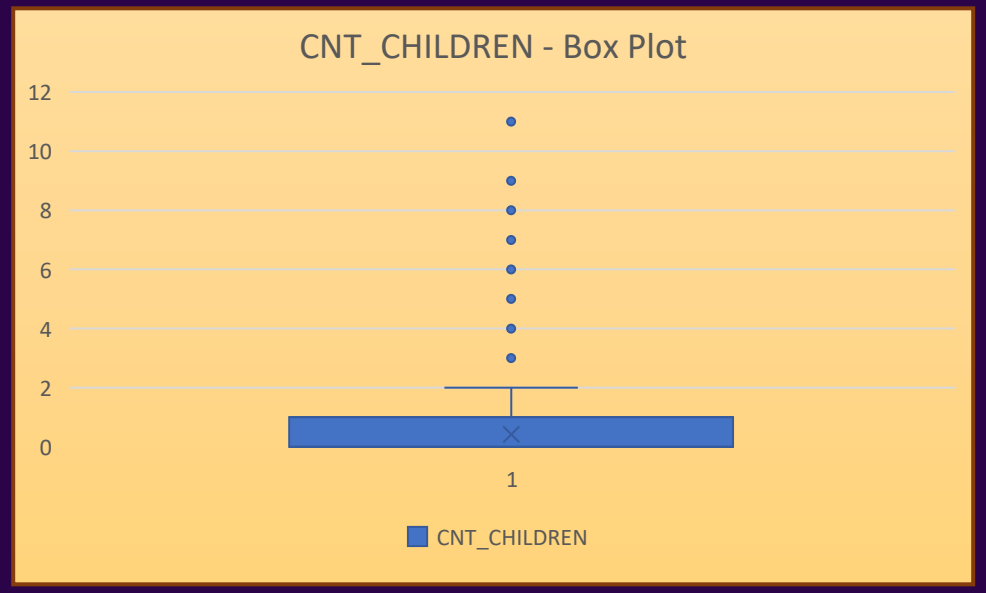


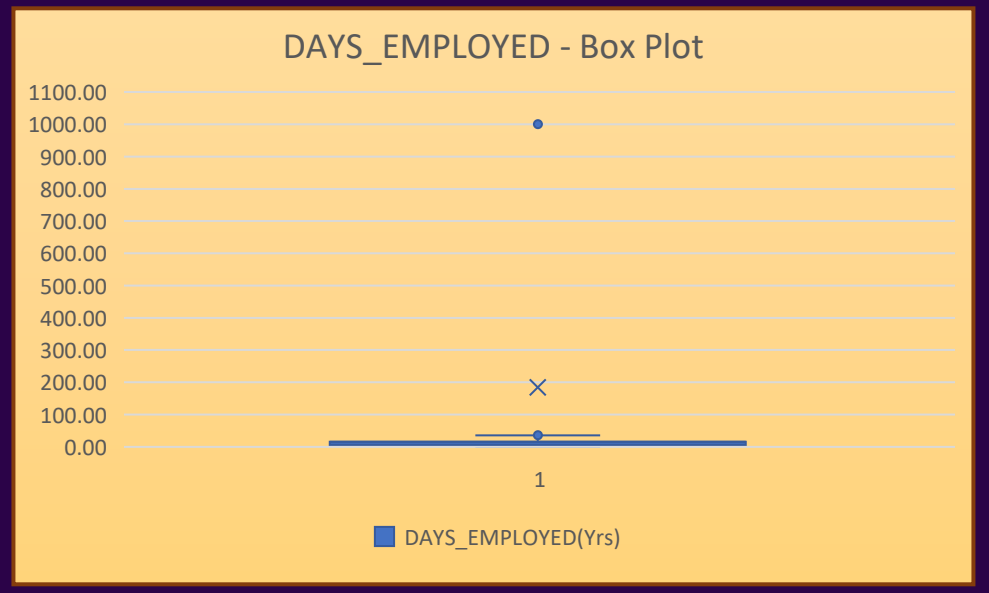
Income total amount has an outlier with income of 11.7 crore however this can be valid business data

Amount credit has few outliers up to data of 40 lakh as shown, however this can be valid business data

Prepared By:
Nilesh Kulkarni

Task B: Identify Outliers in the Dataset: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.





Count of children has few outliers as shown But those can be genuine data points

DAYS_EMPLOYED (years employed) clearly has an outlier of 1000+ years – which can be potential data error – 8924 such rows were updated with average value

Prepared By:
Nilesh Kulkarni

Task B: **Identify Outliers in the Dataset**: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

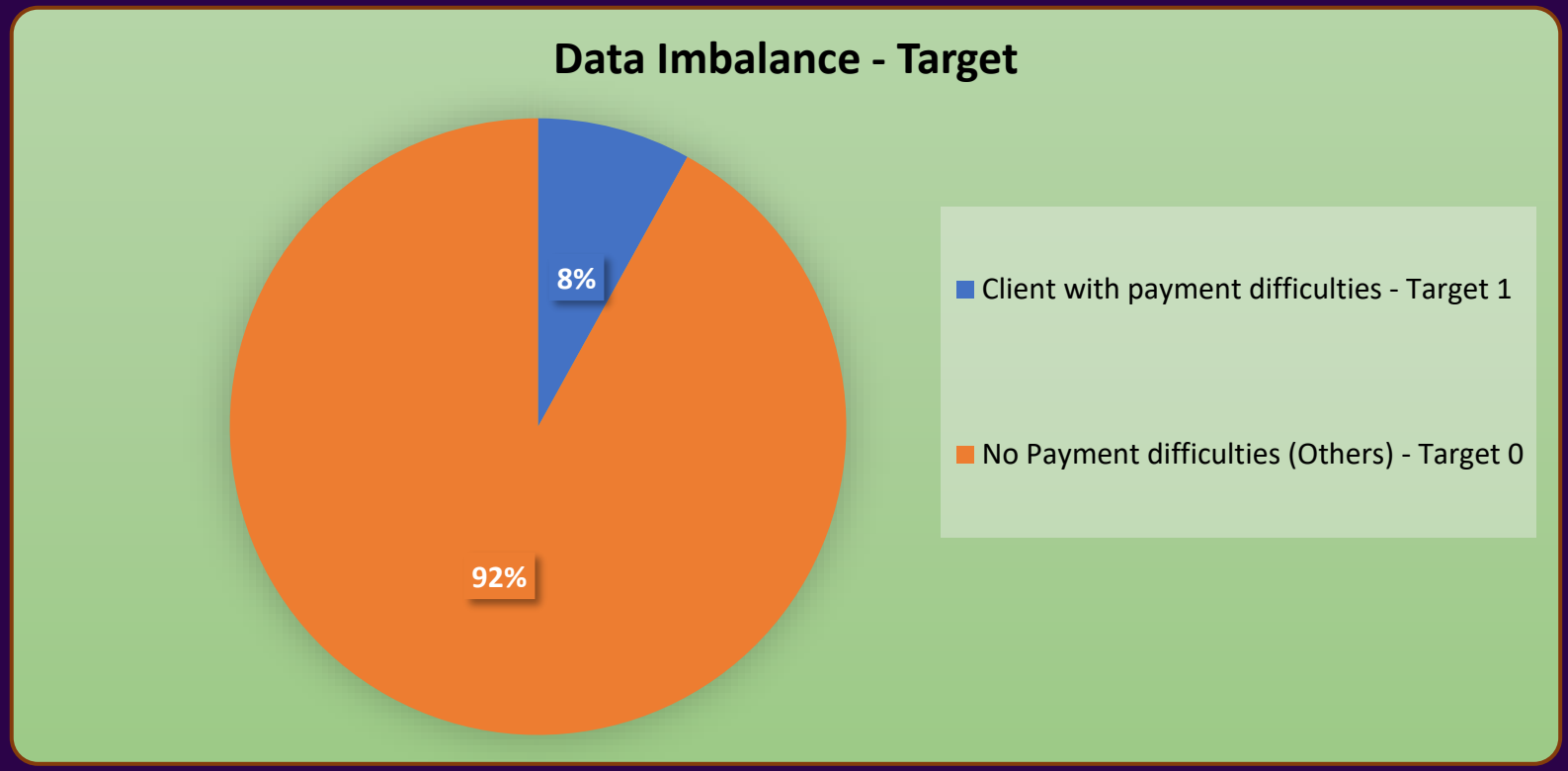
Calculations for finding outliers based on Quartile & IQR as shown below for income amount (column AMT_INCOME), there are more than 2295 outliers based on IQR method

Q1 (Quartile-1)	112500	Quartile has ben calculated including median values
Q3 (Quartile-3)	202500	
IQR (Inter Quartile Range) = Q3-Q1	90000	
High Limit = Q3+(1.5*IQR)	337500	
Low Limit = Q1-(IQR*1.5)	0	Negative value -22500 set to zero since it is amount
Count of Outliers	2295	
Since low limit is zero, we just need to find amount income more than high limit to find outliers in amount income		

Prepared By:
Nilesh Kulkarni

Task C: Analyse Data Imbalance: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

Data Imbalance shown using Target as element/parameter
Data is highly imbalanced (8% defaulters Vs 92% normal customers) as shown and skewed towards target 0
Data modelling can be misleading based on this data

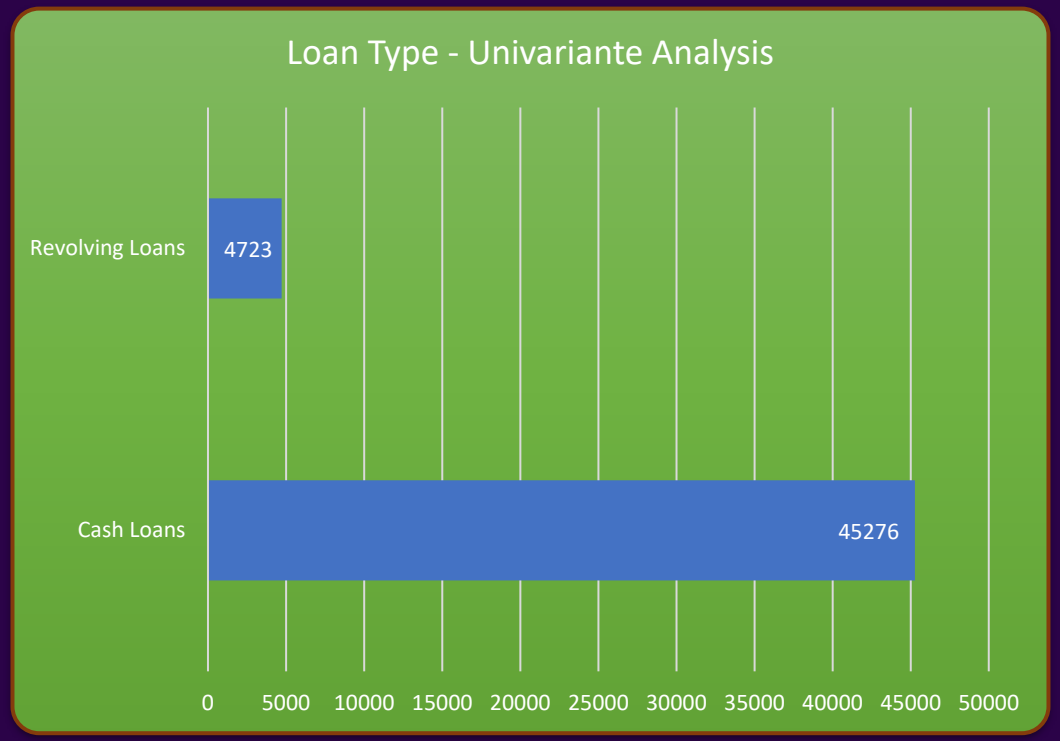


Prepared By:
Nilesh Kulkarni

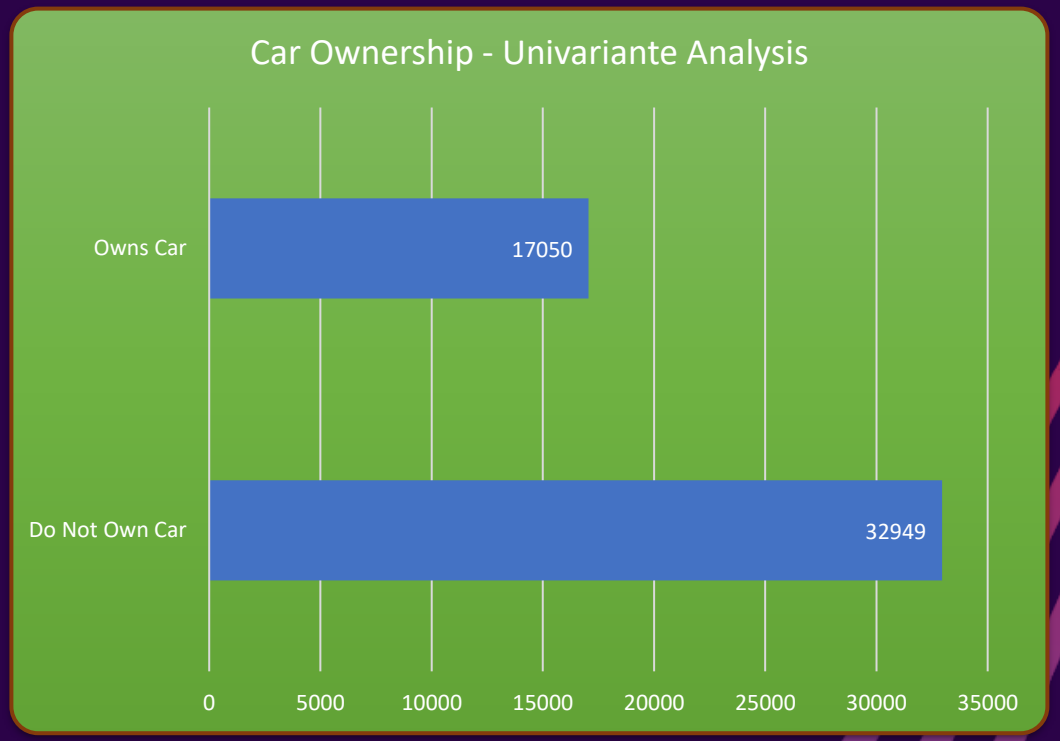
Task D: Perform Univariate, Segmented Univariate, and Bivariate Analysis: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

Univariate Analysis

Loan type - Cash loans are clearly more favourite at 91% of total loans



Car Ownership - 66% applicants do not own cars almost 2/3rd of total

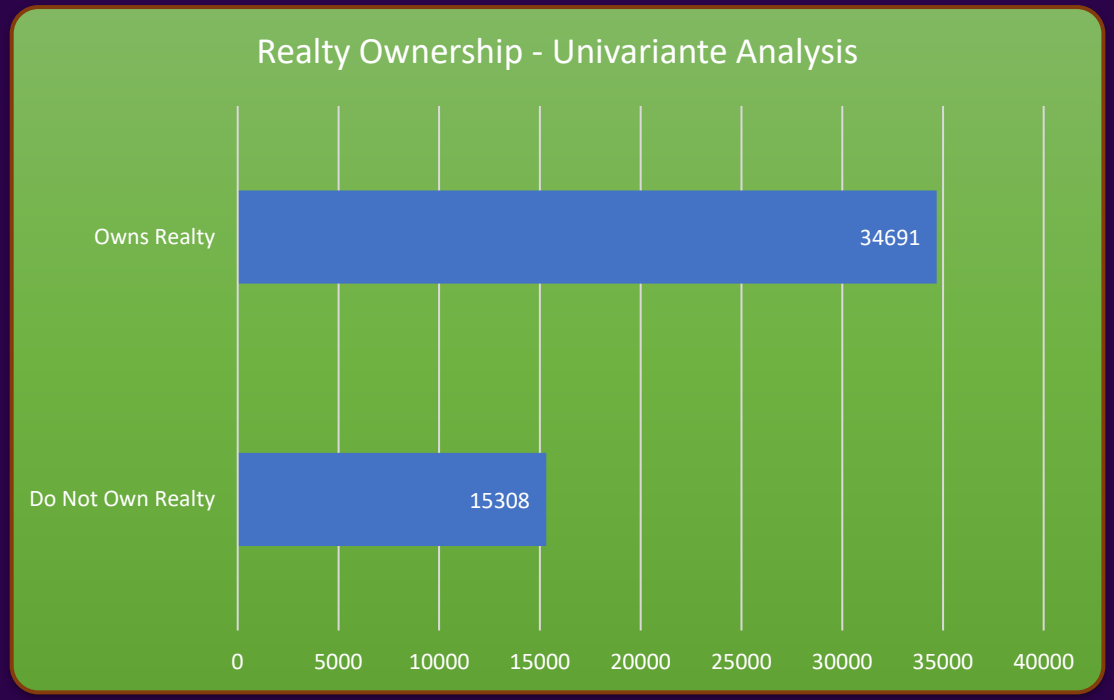


Prepared By:
Nilesh Kulkarni

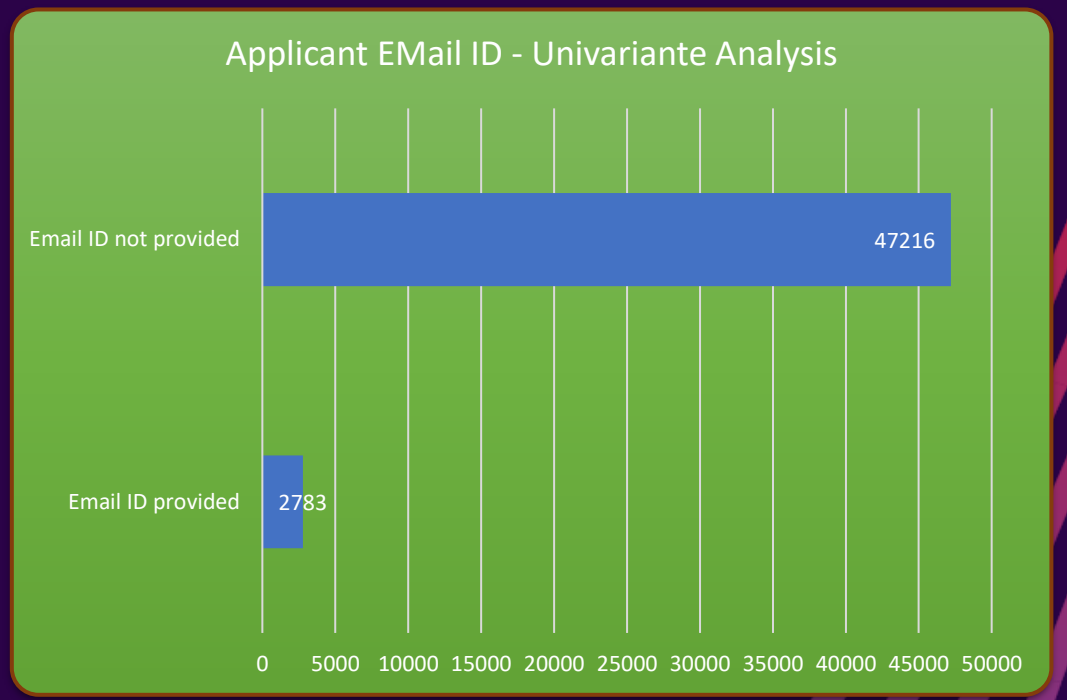
Task D: Perform Univariate, Segmented Univariate, and Bivariate Analysis: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

Univariate Analysis

Realty Ownership – 69% applicants own property (more than 2/3rd of total)



Email Id provided – Majority of applicants (94%) have not provided Email address

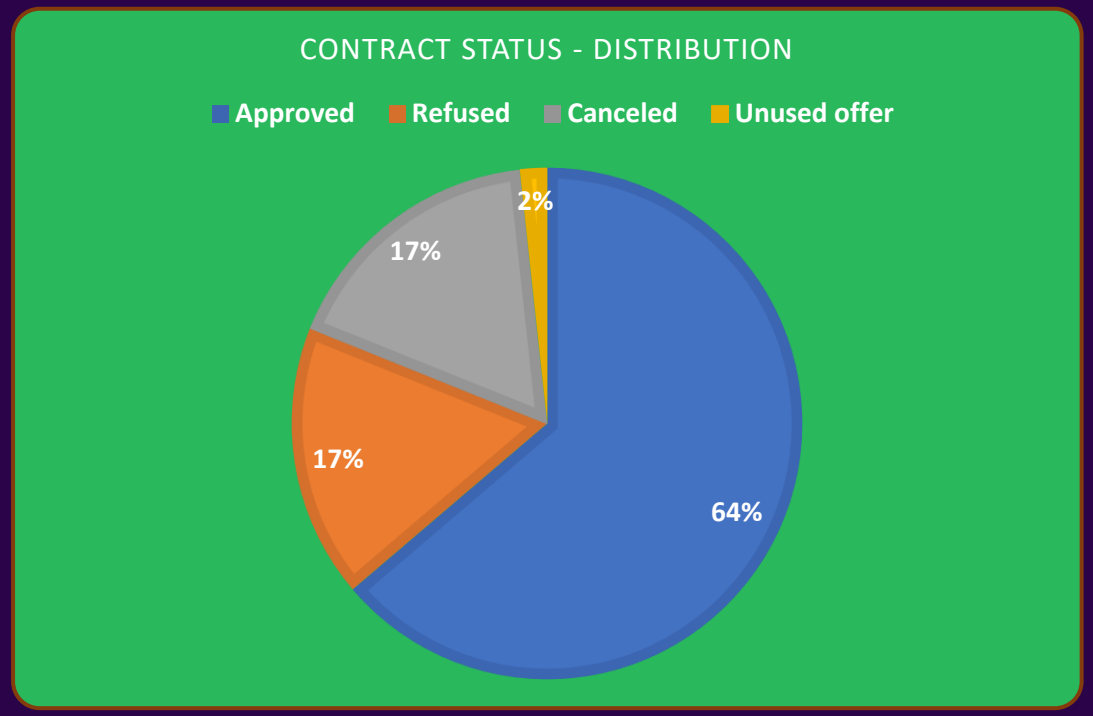


Prepared By:
Nilesh Kulkarni

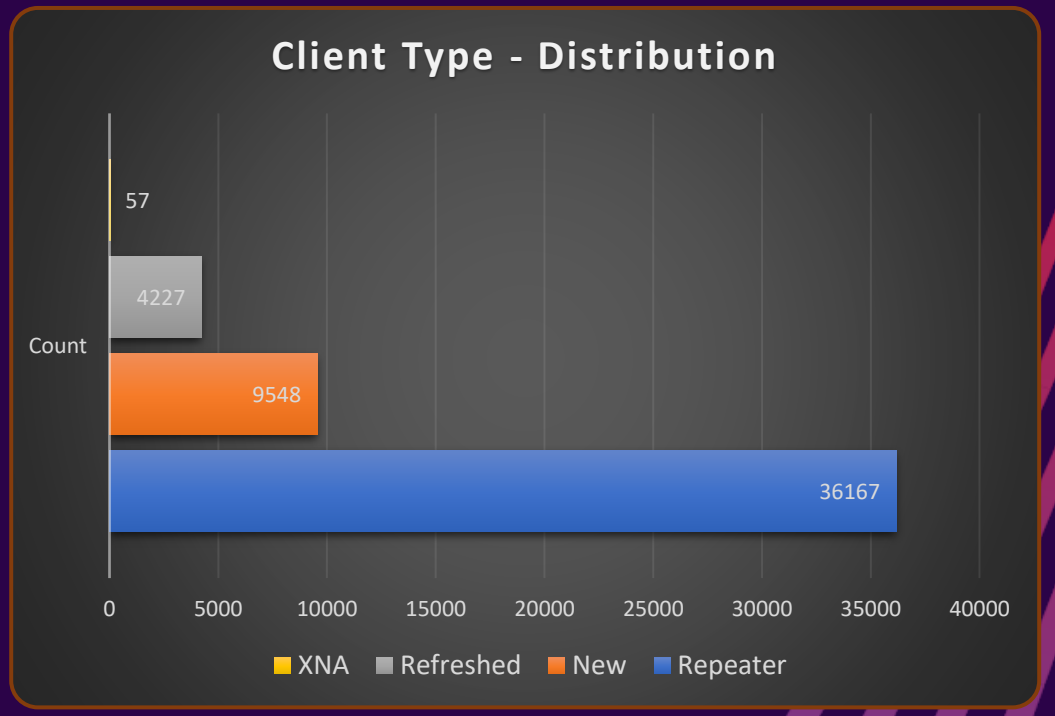
Task D: Perform Univariate, Segmented Univariate, and Bivariate Analysis: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

Univariate Analysis – Previous application data

64% loan applications approved & only 2% were unused offers



More than 70% previous applications were Repeaters & new were less than 20%



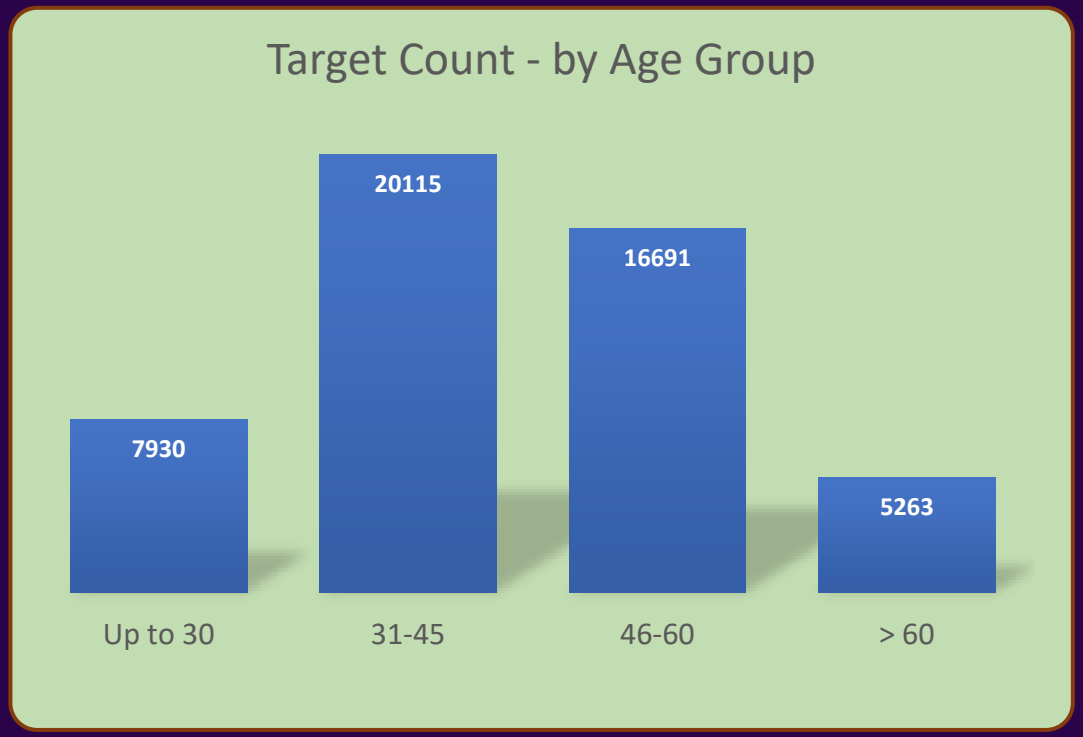
Prepared By:
Nilesh Kulkarni

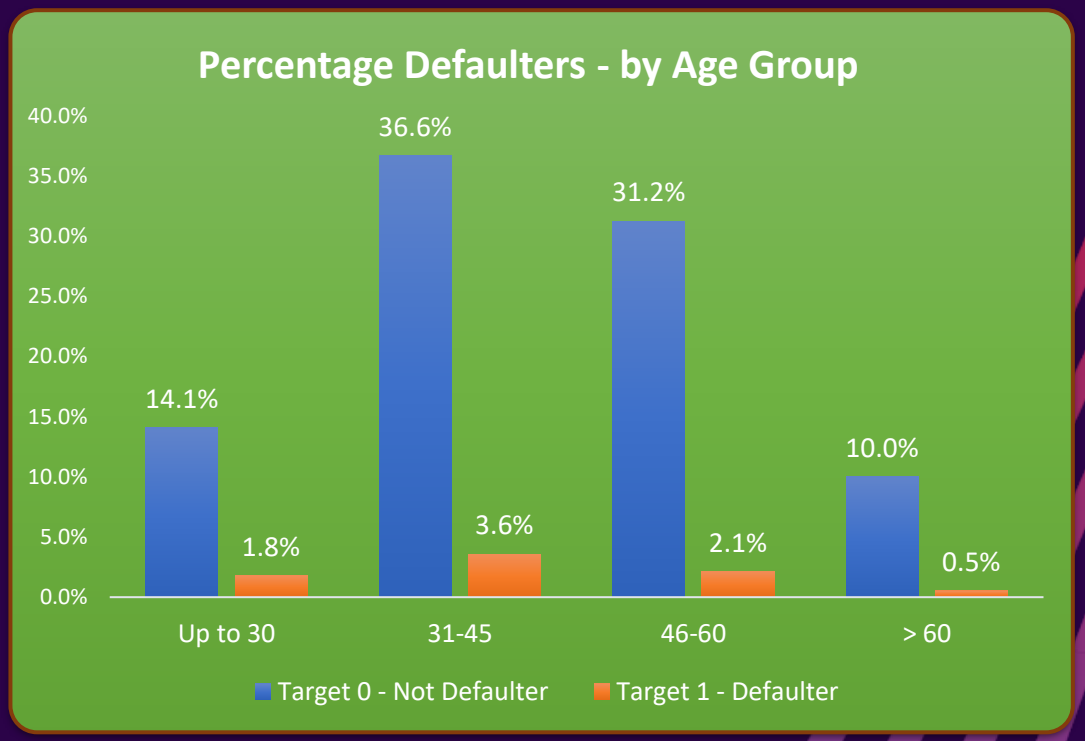
Task D: Perform Univariate, Segmented Univariate, and Bivariate Analysis: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

Segmented Univariate Analysis

Majority of loan applicants are in age group 31-45 and 46-60

60+ age group has least percent of defaulters and 31-45 age group has most percent of defaulters



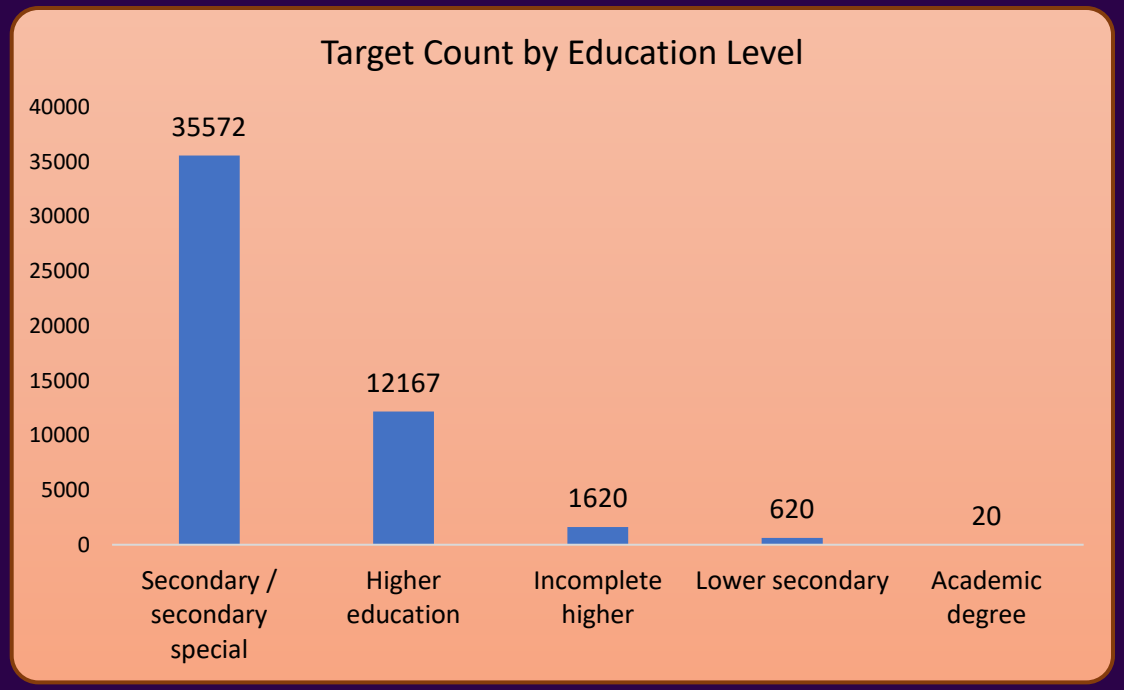


Prepared By:
Nilesh Kulkarni

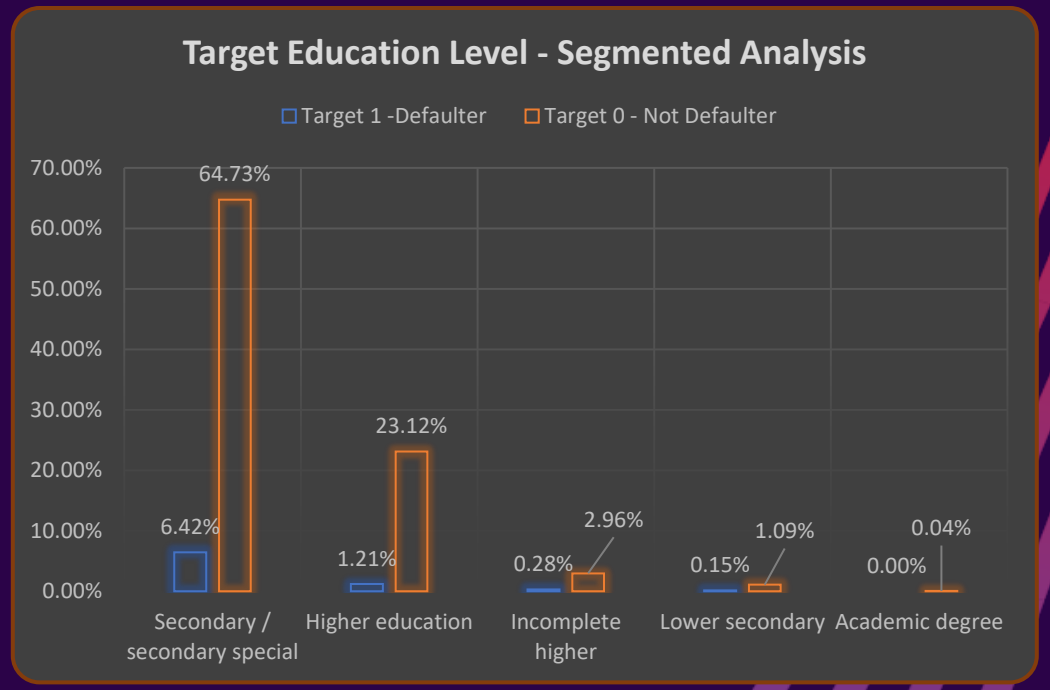
Task D: Perform Univariate, Segmented Univariate, and Bivariate Analysis: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

Segmented Univariate Analysis

Least applicants with academic degree & most applicants are secondary education qualification



No defaulters with academic degree and most defaulter percent with secondary education

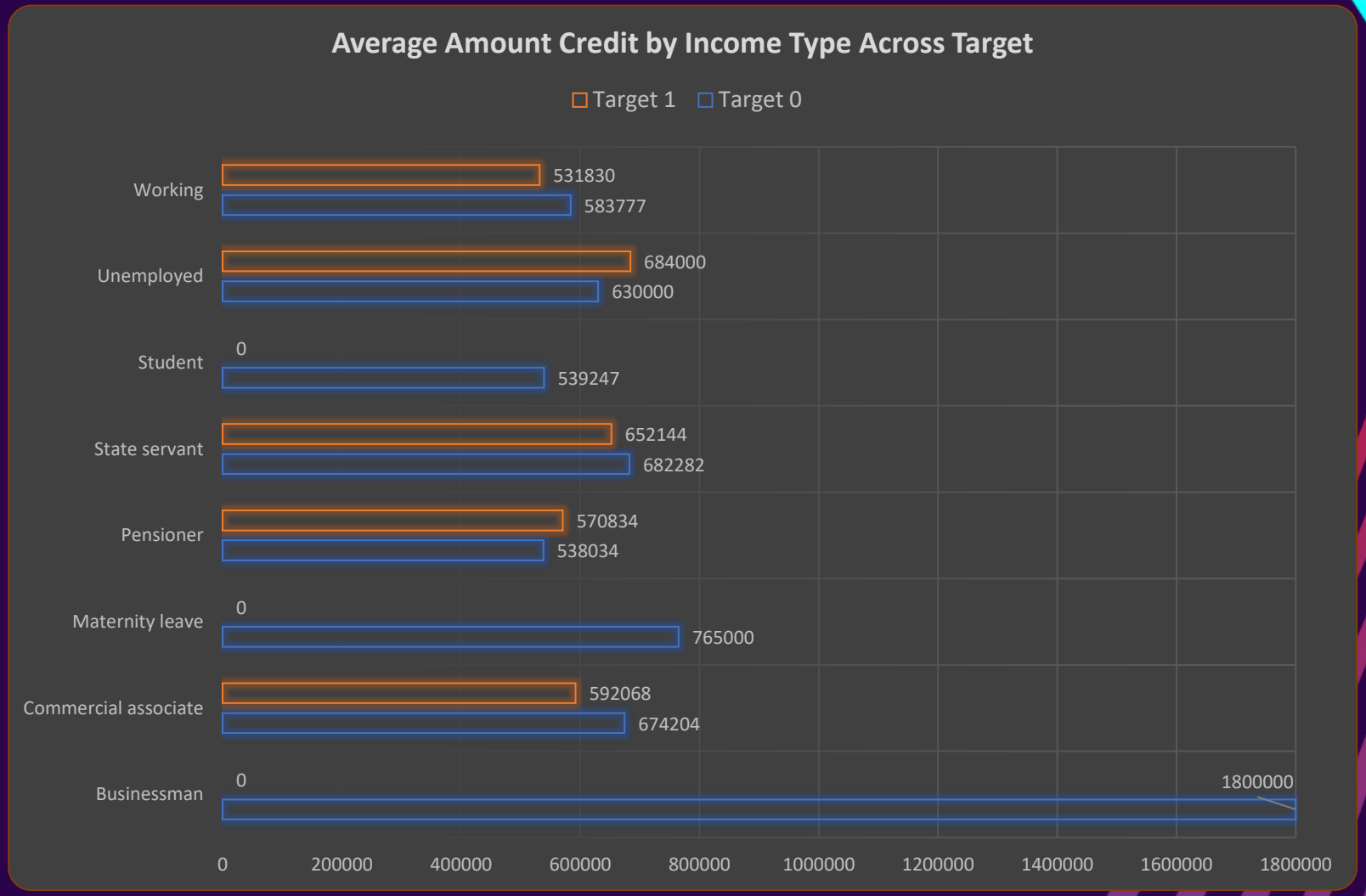


Prepared By:
Nilesh Kulkarni

Task D: Perform Univariate, Segmented Univariate, and Bivariate Analysis: 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.

Bivariate Analysis

Zero defaulters with income-type student, Maternity leave and businessman. Maximum defaulter rate with 'unemployed'. Businessman category has highest average amount credit.

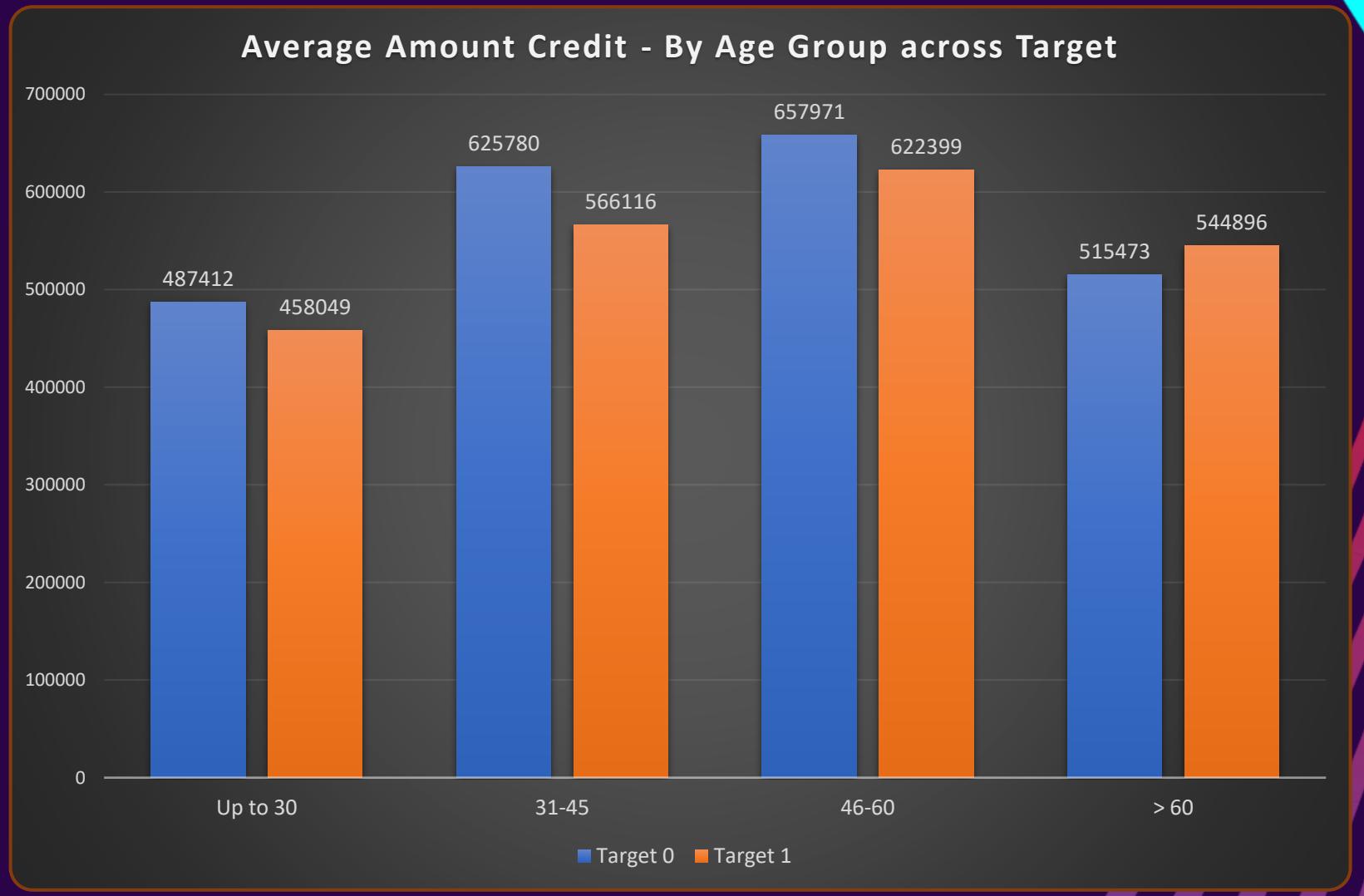


Prepared By:
Nilesh Kulkarni

Task D: Perform Univariate, Segmented Univariate, and Bivariate Analysis: 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.

Bivariate Analysis

Age group 46-60 has highest credit amount across defaulters as well as normal applicants. Age group Upto-30 has least average amount credit across both target types.



Prepared By:
Nilesh Kulkarni

Task E: Identify Top Correlations for Different Scenarios: Task: Segment the dataset based on different scenarios (e.g., clients with payment difficulties & all other cases) & identify the top correlations for each segmented data using Excel functions.

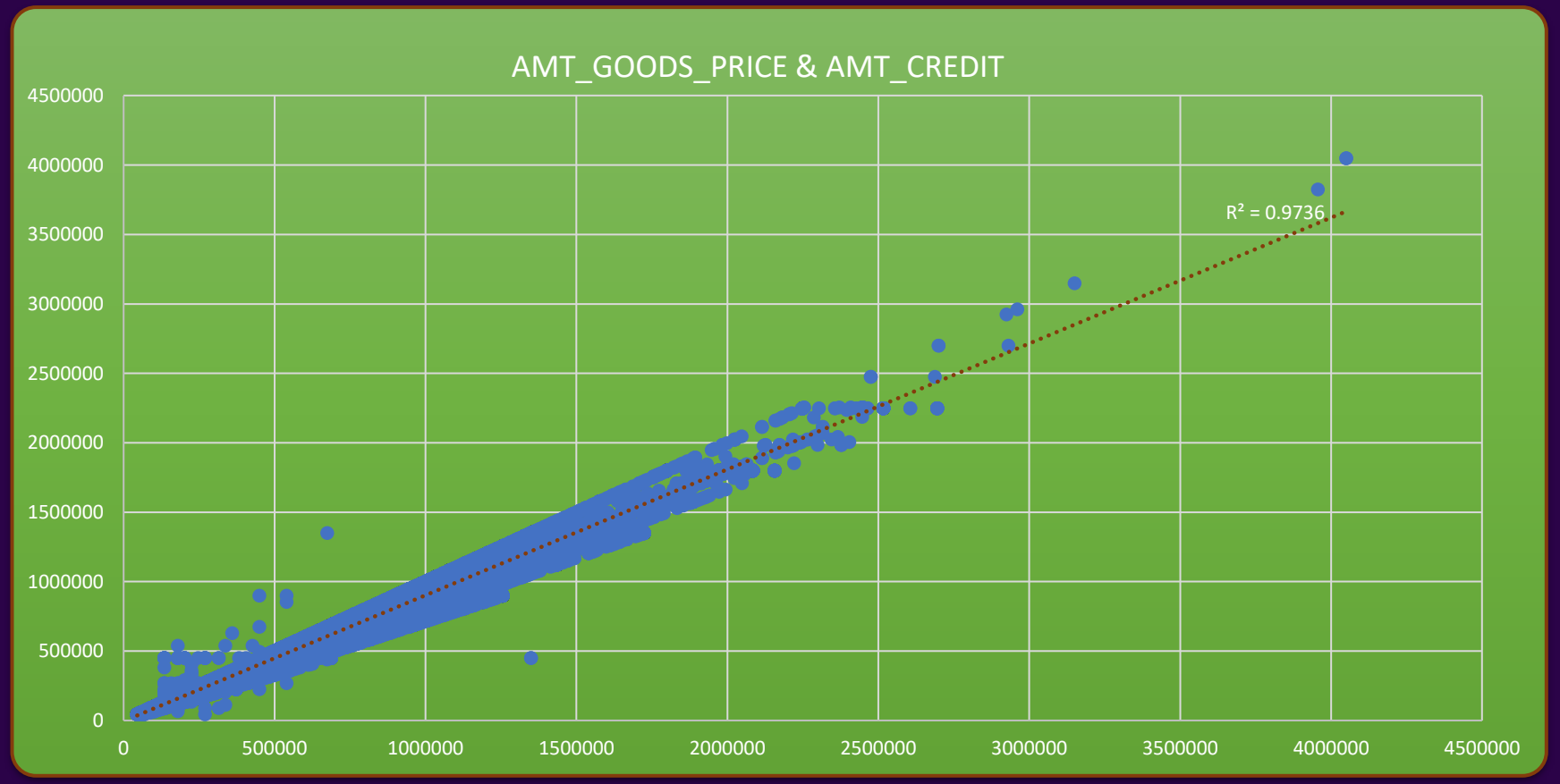
Relationship between numeric variables along with relationship co-efficient has been shown below – It includes heatmap on relationship co-efficient as well as type of relationship (positive or negative)

Feature 1	Feature 2	Relationship Co-efficient	Type - Relationship
CNT_CHILDREN	CNT_FAM_MEMBERS	0.880453292	Strongly Positive
AMT_GOODS_PRICE	AMOUNT_CREDIT	0.986704386	Strongly Positive
FLAG_EMP_PHONE	AGE	-0.617469697	Strongly Negative
AMT_CREDIT	AMT_ANNUITY	0.769498787	Strongly Positive
AMT_INCOME_TOTAL	AMT_CREDIT	0.069315897	Mild Positive
Age	FLAG_EMAIL	-0.09217863	Mild Negative

Prepared By:
Nilesh Kulkarni

Task E: Identify Top Correlations for Different Scenarios: Task: Segment the dataset based on different scenarios (e.g., clients with payment difficulties & all other cases) & identify the top correlations for each segmented data using Excel functions.

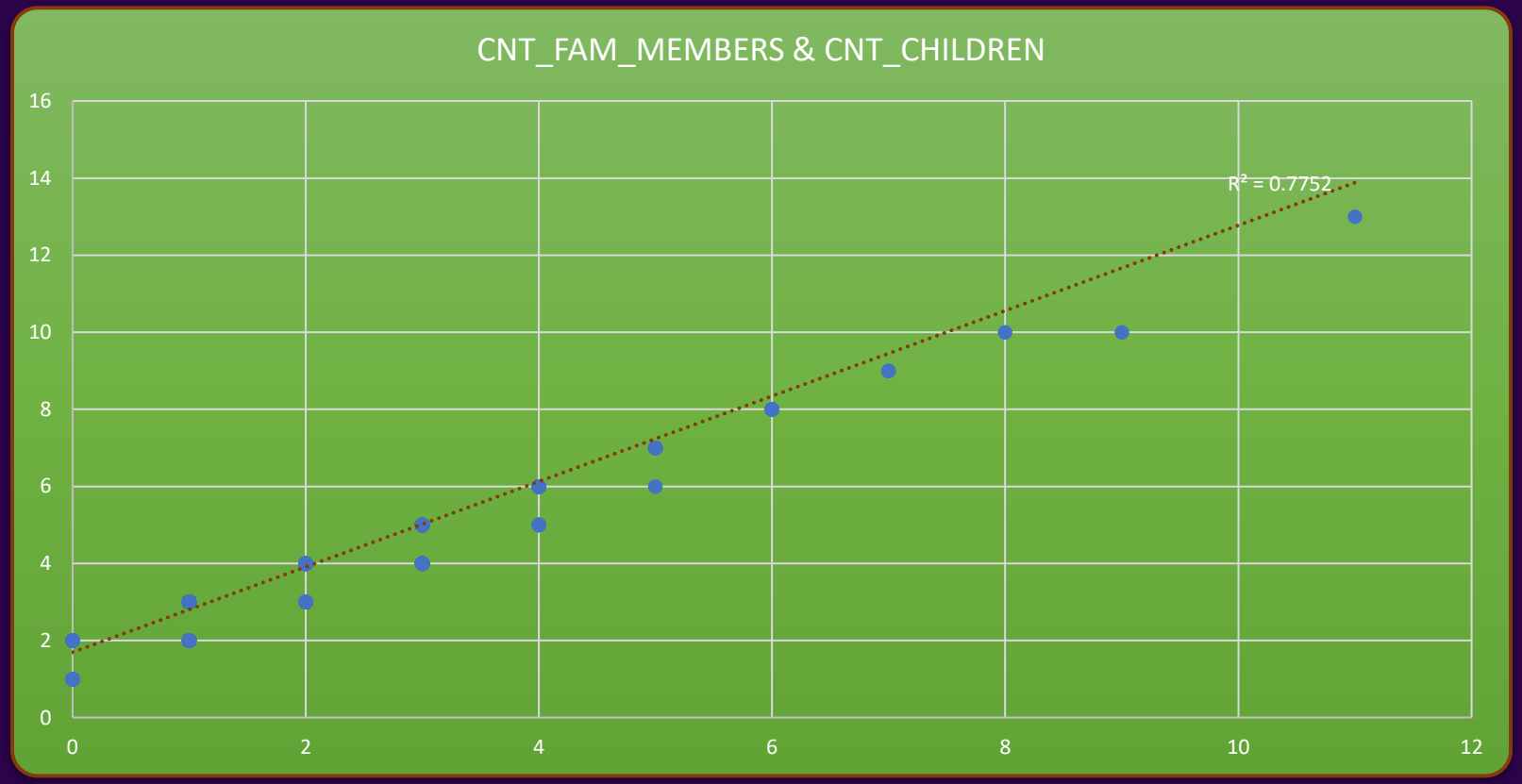
Scatter diagram demonstrating relationship between Amount credit and Amount goods price – the relation ship is positive i.e. amount credit increases with increase in amount goods price



Prepared By:
Nilesh Kulkarni

Task E: Identify Top Correlations for Different Scenarios: Task: Segment the dataset based on different scenarios (e.g., clients with payment difficulties & all other cases) & identify the top correlations for each segmented data using Excel functions.

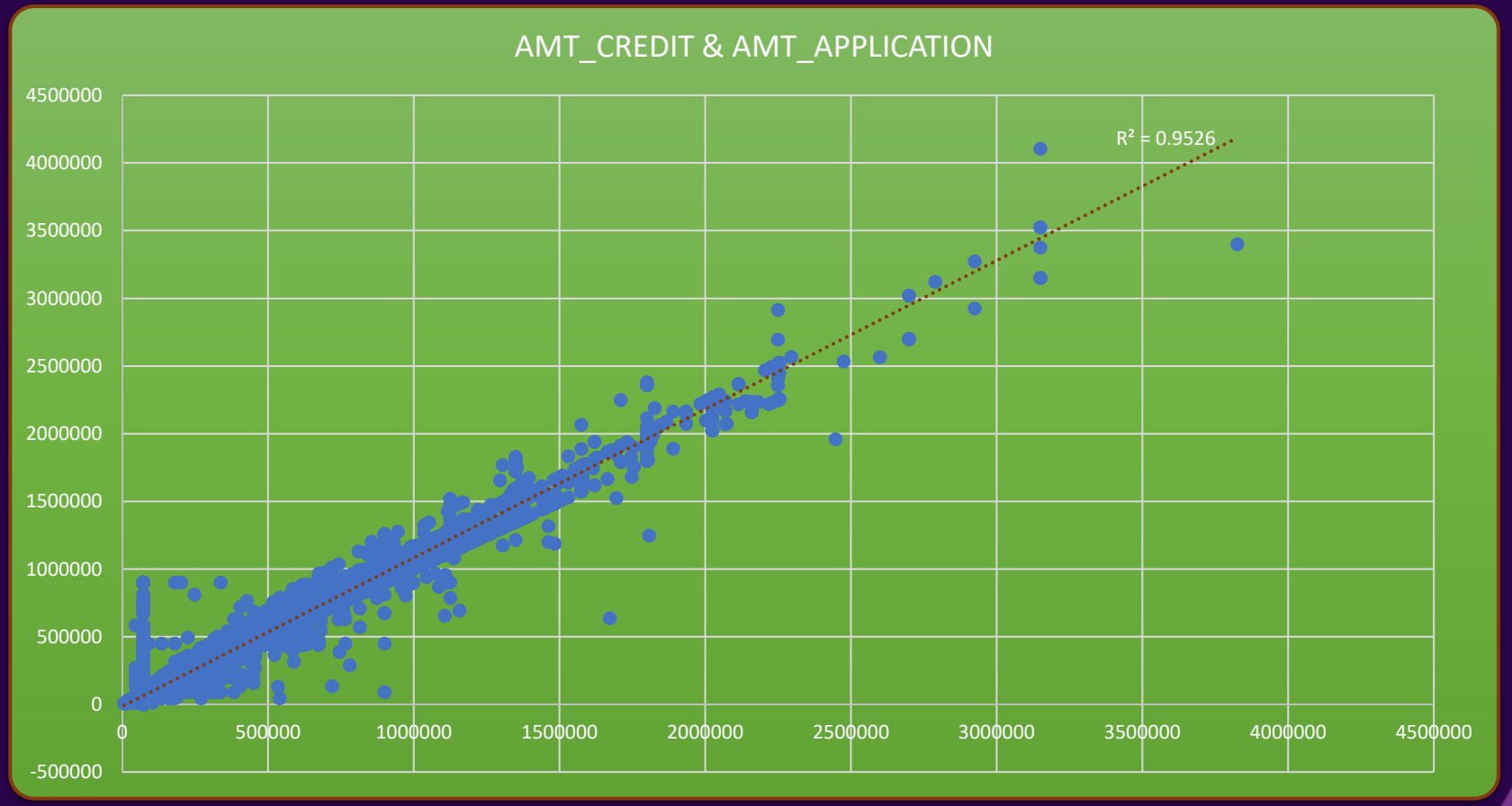
Scatter diagram demonstrating relationship between children count and family members count– the relation ship is positive i.e. family members count increases with increase in children count



Prepared By:
Nilesh Kulkarni

Task E: Identify Top Correlations for Different Scenarios: Task: Segment the dataset based on different scenarios (e.g., clients with payment difficulties & all other cases) & identify the top correlations for each segmented data using Excel functions.

Previous Applications Data – Strong positive relationship between amount application & amount credit as shown on scatter plot below



Prepared By:
Nilesh Kulkarni

Insights - Summary

- The data has many issues like missing values (more than 30% rows blank) and outliers so there is scope for improving data quality at source.
- Median based imputation for numeric data and Mode (most occurring value) based imputation for categorical data are quick and reliable ways to handle missing data or data errors.
- The data has few outliers in features like income amount, credit amount, children count - however practical bank domain relevance need to be checked while handling outliers.
- The provided data is largely imbalanced and data modelling might not be correct on this data.
- Univariate Analysis shows that cash loans mostly preferred, most applicant do not own car, majority own realty and email id usage is low among applicants.
- Segmented univariate analysis shows that – age group 60+ & applicants with academic degree has least defaulters whereas age group 31-45 & applicants with secondary education has the most.
- Bivariate analysis shows that no defaulters for students, with maternity leave and businessman whereas most defaulters for 'unemployed' category of income type. Also Age group 46-60 has highest average amount credit and age group Upto-30 has least average credit amount.
- There are certain variables in data which shows strong negative or positive co-relation between those which can be traced by relationship co-efficient/scatter charts.

Prepared By:
Nilesh Kulkarni

Results – Up-Skilling

The project helped for individual upskilling and gain experience in Data Analytics. The project helped me to gain knowhow and experience in listed areas:

- ✓ Retail Banking related domain knowhow and overview of risk analytics
- ✓ Data preparation and transformation skills using MS-Excel
- ✓ Preparation of visuals and various charts in MS-Excel with huge data
- ✓ Relating domain knowhow with statistics and MS-Excel function to transform data including handling missing values and outliers
- ✓ Data analysis and insights extraction using statistics basics and domain knowhow
- ✓ Various charts/functions in MS-Excel like Box-Plot & Quartile
- ✓ Preparation of comprehensive project report using MS-PowerPoint and MS-Excel

Overall, the project has been great learning and professionally enriching experience.

Prepared By:
Nilesh Kulkarni

THANK YOU

Nilesh Kulkarni

inileshkulkarni@gmail.com