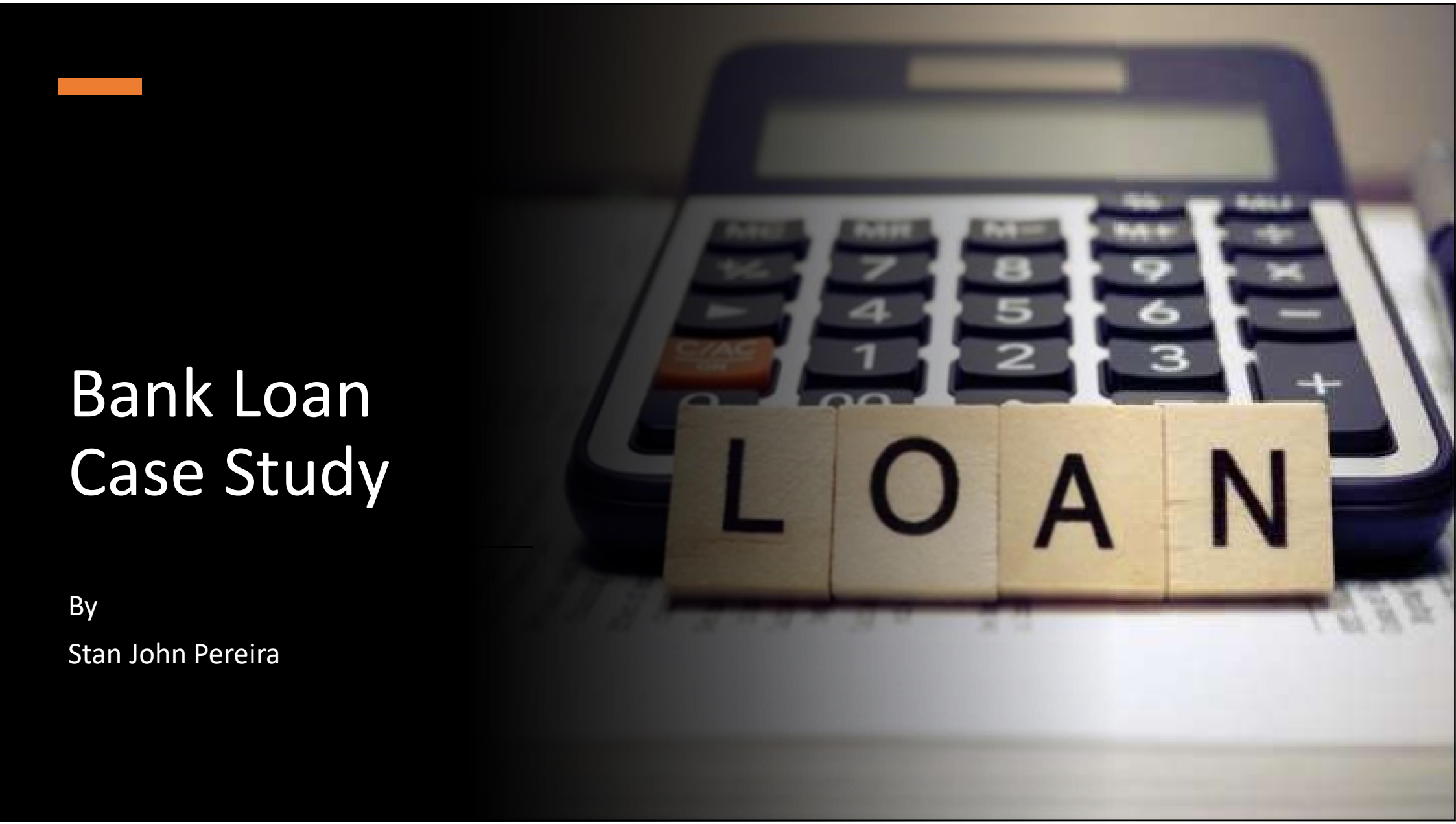
Bank Loan Case Study



By

Stan John Pereira

Project Description

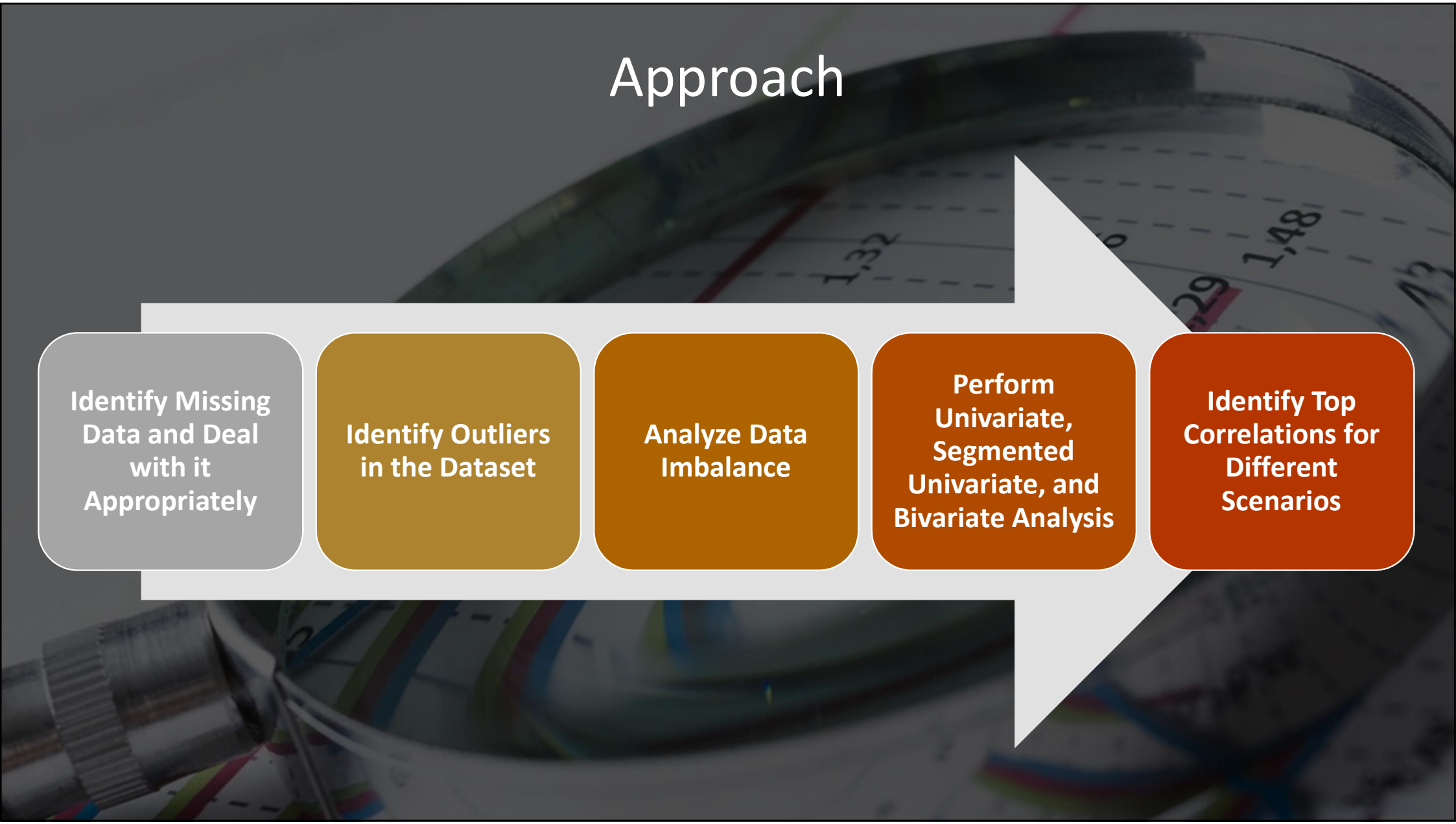
The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their installments.

This information can be used to make decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants.

The company wants to understand the key factors behind loan default so it can make better decisions about loan approval.

As a Data Analyst, my task is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.

Approach


Identify Missing Data and Deal with it Appropriately
Identify Outliers in the Dataset
Analyze Data Imbalance
Perform Univariate, Segmented Univariate, and Bivariate Analysis
Identify Top Correlations for Different Scenarios

Tech Stack Used

Ability to perform calculations, data analysis, data visualization, data transformation, and data cleaning with Excel tools and functions.

Transform and clean data with features like Power Query and Flash Fill.

Microsoft Excel 2010
Version 14.0.7628.5000

Code to automate tasks and customize functions with VBA (Visual Basic for Applications).

Availability of free templates and code to customize and automate Excel.


Insights

Task 1 : Identify Missing Data and Deal with it Appropriately

As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

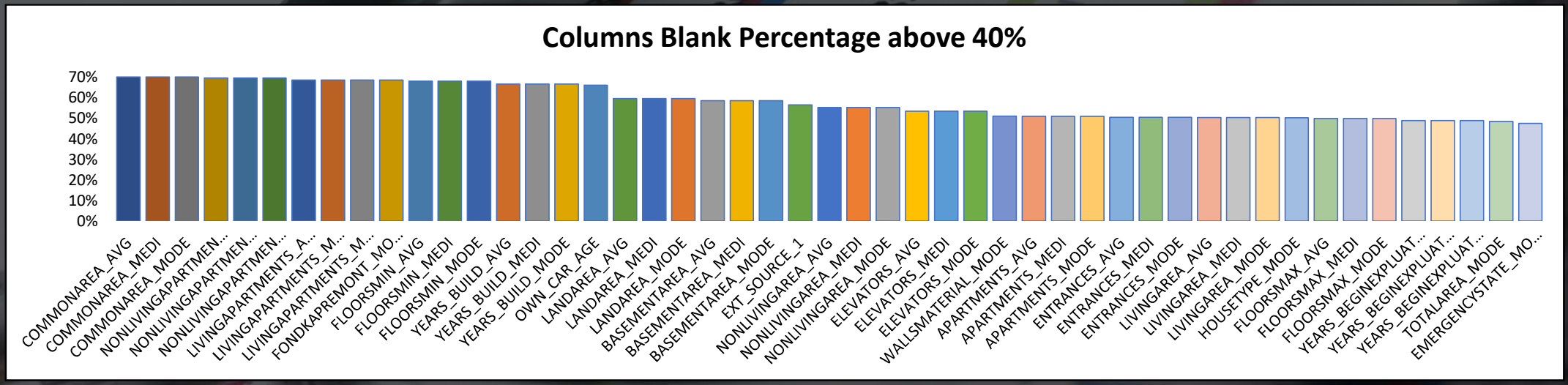
- Understand the Data
- Find and Handle Missing Values (Rows and Columns)
- Remove Irrelevant Columns
- Remove Duplicates
- Impute Missing Values with Mean/Median/Mode

Understand the Data

Application Data Table

Provides details about the current loan applications.

Application Data (before Cleaning)	
Total Rows	49999
Total Columns	122
Columns with Null %age >40%	49



Find and Handle Missing Values (Rows and Columns)

NO DUPLICATES FOUND – SK_ID_CURR

Columns Dropped due to Blank Percentage over 40%	
COMMONAREA_AVG	NONLIVINGAREA_MEDI
COMMONAREA_MEDI	NONLIVINGAREA_MODE
COMMONAREA_MODE	ELEVATORS_AVG
NONLIVINGAPARTMENTS_AVG	ELEVATORS_MEDI
NONLIVINGAPARTMENTS_MEDI	ELEVATORS_MODE
NONLIVINGAPARTMENTS_MODE	WALLSMATERIAL_MODE
LIVINGAPARTMENTS_AVG	APARTMENTS_AVG
LIVINGAPARTMENTS_MEDI	APARTMENTS_MEDI
LIVINGAPARTMENTS_MODE	APARTMENTS_MODE
FONDKAPREMONT_MODE	ENTRANCES_AVG
FLOORSMIN_AVG	ENTRANCES_MEDI
FLOORSMIN_MEDI	ENTRANCES_MODE
FLOORSMIN_MODE	LIVINGAREA_AVG
YEARS_BUILD_AVG	LIVINGAREA_MEDI
YEARS_BUILD_MEDI	LIVINGAREA_MODE
YEARS_BUILD_MODE	HOUSETYPE_MODE
OWN_CAR_AGE	FLOORSMAX_AVG
LANDAREA_AVG	FLOORSMAX_MEDI
LANDAREA_MEDI	FLOORSMAX_MODE
LANDAREA_MODE	YEARS_BEGINEXPLUATATION_AVG
BASEMENTAREA_AVG	YEARS_BEGINEXPLUATATION_MEDI
BASEMENTAREA_MEDI	YEARS_BEGINEXPLUATATION_MODE
BASEMENTAREA_MODE	TOTALAREA_MODE
EXT_SOURCE_1	EMERGENCYSTATE_MODE
NONLIVINGAREA_AVG	

Columns Dropped as not required for Analysis	
EXT_SOURCE2	EXT_SOURCE3

Columns Having Blank Percentage Data Imputed based on Mean/Median/Mode	
OCCUPATION_TYPE	Mode Used. New Type assigned as Blank had the most records - 'Unknown'
AMT_REQ_CREDIT_BUREAU_DAY	Median Used
AMT_REQ_CREDIT_BUREAU_HOUR	Median Used
AMT_REQ_CREDIT_BUREAU_MON	Median Used
AMT_REQ_CREDIT_BUREAU_QRT	Median Used
AMT_REQ_CREDIT_BUREAU_WEEK	Median Used
AMT_REQ_CREDIT_BUREAU_YEAR	Median Used
NAME_TYPE_SUITE	Mode Used. Changed the Blanks to 'Unaccompanied'
DEF_30_CNT_SOCIAL_CIRCLE	Median Used
DEF_60_CNT_SOCIAL_CIRCLE	Median Used
OBS_30_CNT_SOCIAL_CIRCLE	Median Used
OBS_60_CNT_SOCIAL_CIRCLE	Median Used
AMT_GOODS_PRICE	Median Used

Columns with Minimal Blank Percentage Rows with Blank Values dropped	
AMT_ANNUITY	01 Row Dropped
CNT_FAM_MEMBERS	01 Row Dropped
DAYS_LAST_PHONE_CHANGE	01 Row Dropped

Task 2: Identify Outliers in the Dataset

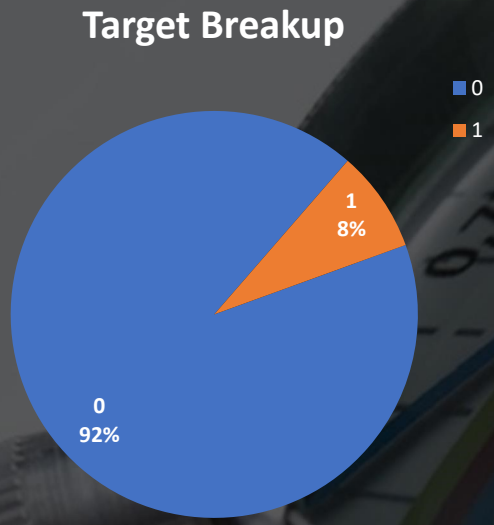
Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

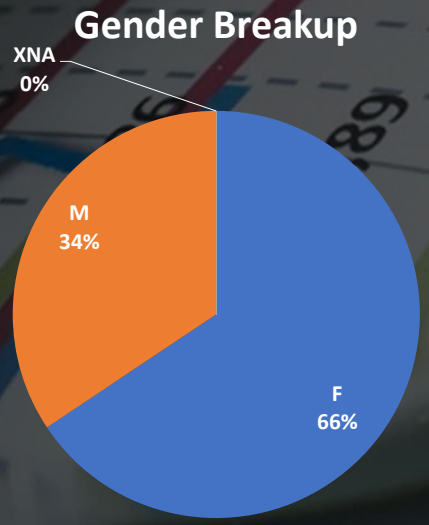
Outliers Analysis

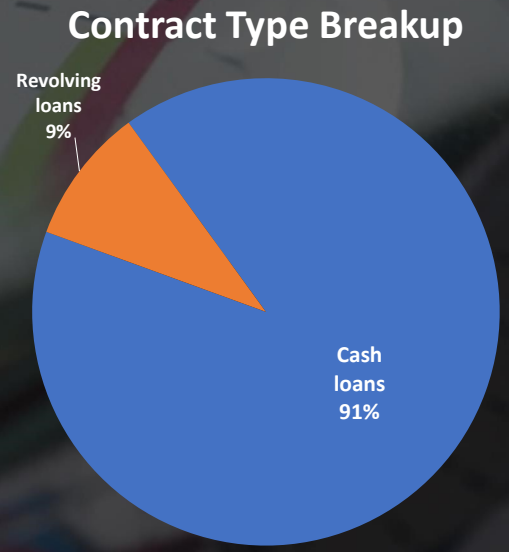
- There are few high outliers in AMT_ANNUITY which does not give a very healthy analysis and can skew the mean.
- There are a few outliers in AMT_GOODS_PRICE & AMT_CREDIT where the amount is more than normal.
- There is an outliers in AMT_INCOME_TOTAL which is extraordinarily high, which is unusual.
- The YEARS_WORKED data has quite a lot of data that has the value - 1000 years. This is impossible. On further investigating the data, it was found that the value was assigned for ‘Pensioners’
- There are a few outliers in DAYS_LAST_PHONE_CHANGE which suggests that many people are using the same phone for a long time, even over 8 years
- Most of the people in the data are over 20 years and below 75 years, which shows that the data is not skewed from the point of age.
- Lastly, there are few outliers in CNT_CHILDREN, where a few people have more than 4 children, which is impractical and can put financial pressure.

Task 3 : Analyze Data Imbalance

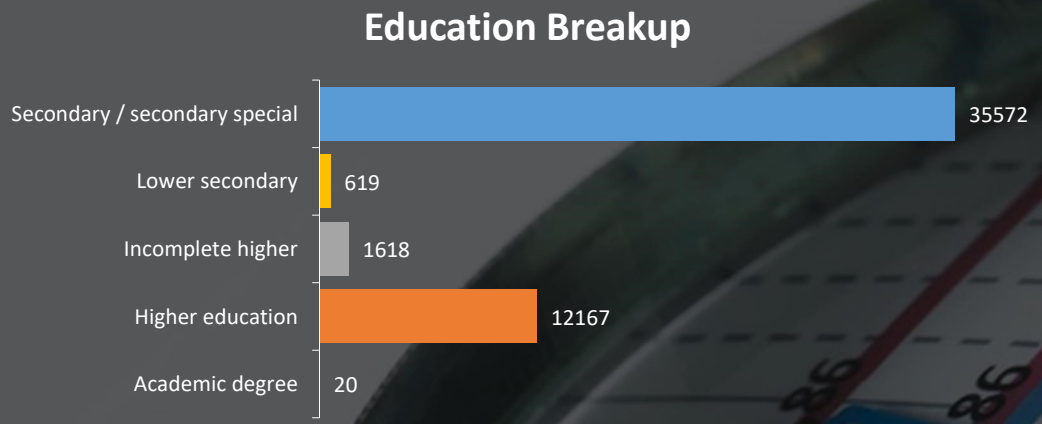
Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

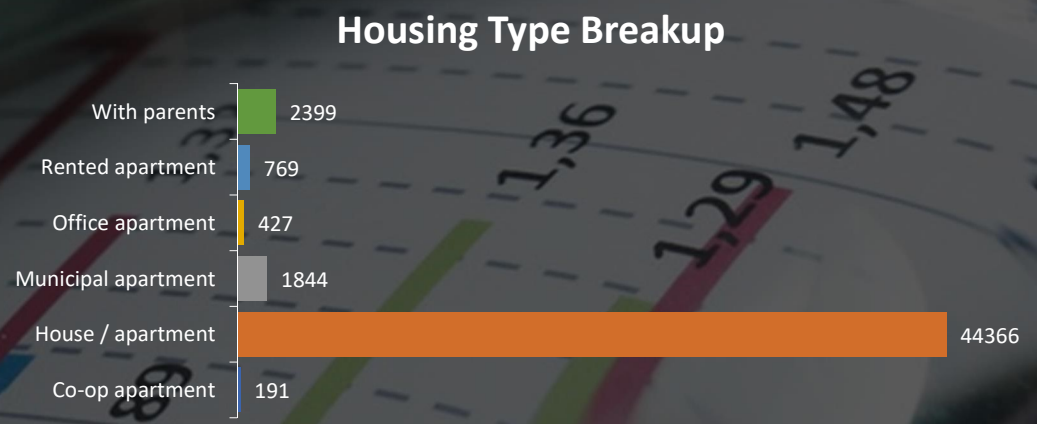


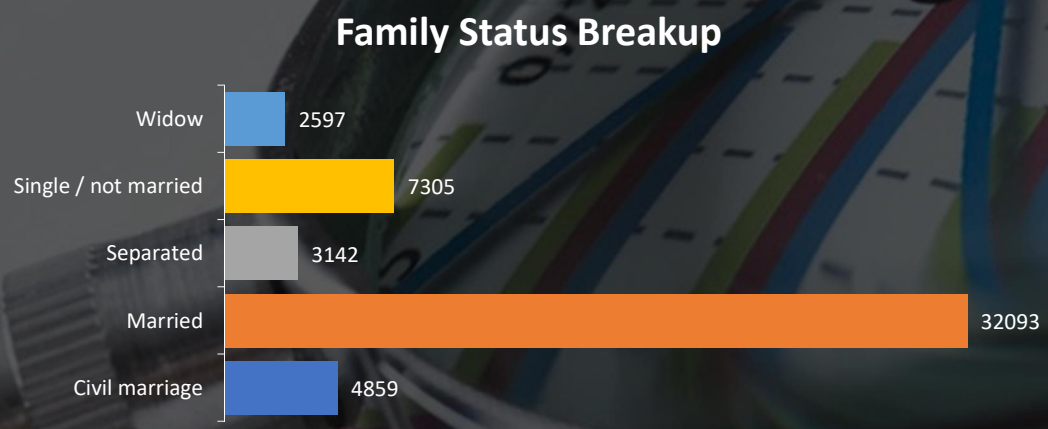




Data Imbalance







All the charts shown have Data Imbalance

The Pie Charts have RePayers, Females and Cash Loans are majority in the data.

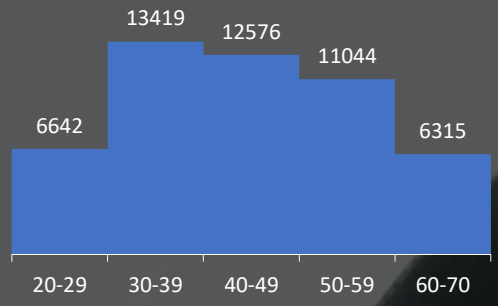
The Bar Charts have people who have studied till Secondary, Married and People staying in Apartments taking up most of the data

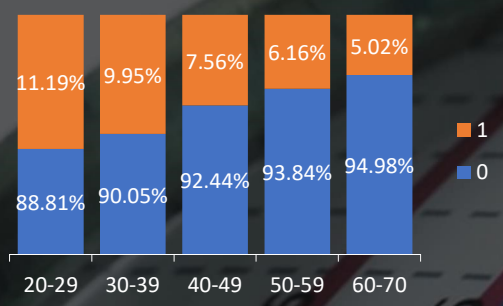
Task 4 : Perform Univariate, Segmented Univariate, and Bivariate Analysis

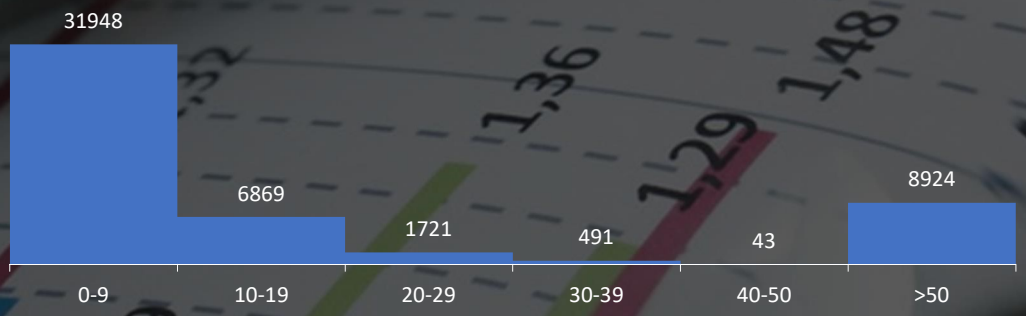
To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.
To attain an analysis of the data, we will be performing the below:

- Univariate Analysis to understand the distribution of individual variables
- Segmented Univariate Analysis to compare variable distributions for different scenarios
- Bivariate Analysis to explore relationships between variables and the target variable using Excel functions and features.

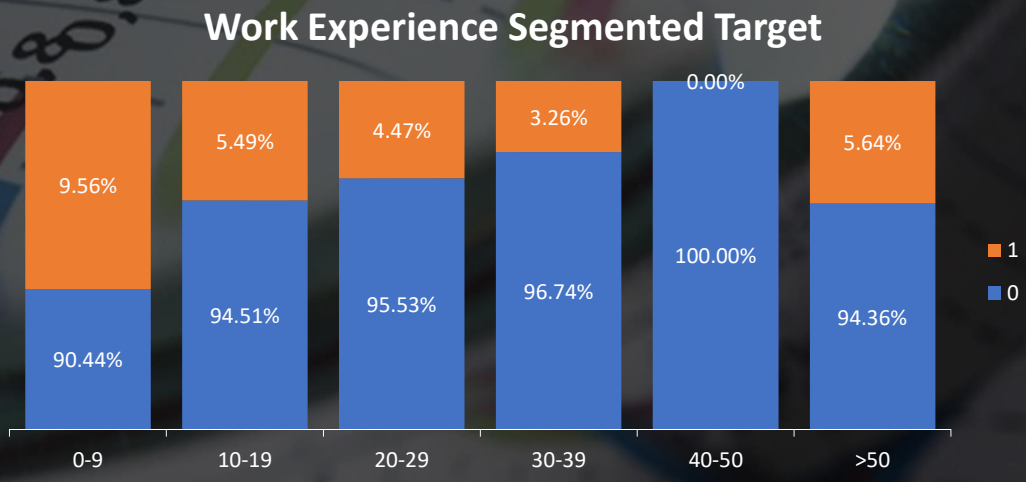
Univariate and Segmented Univariate Analysis



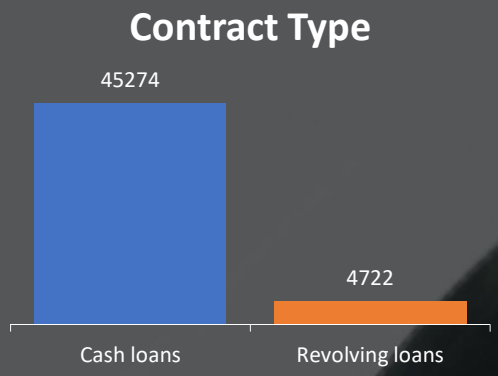


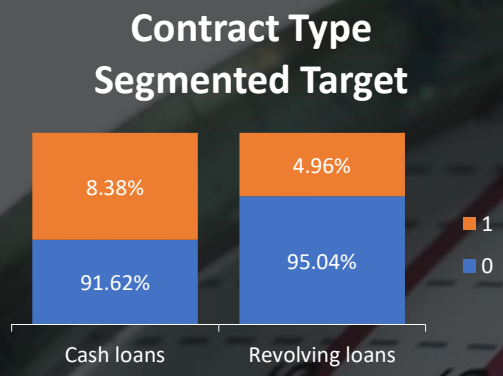


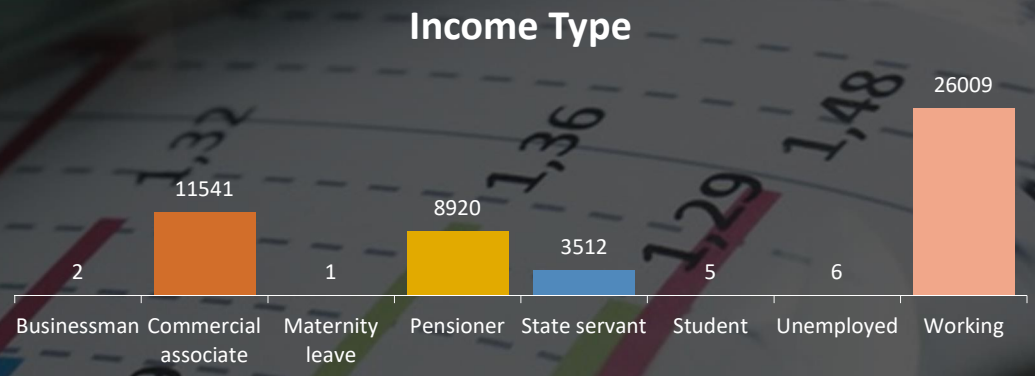
- Middle Age group have been given the most loans
- The percentage of Defaulters decrease as age increases
- Most people taking loans have less than 10 years work experience
- Defaulters decrease as work experience increase



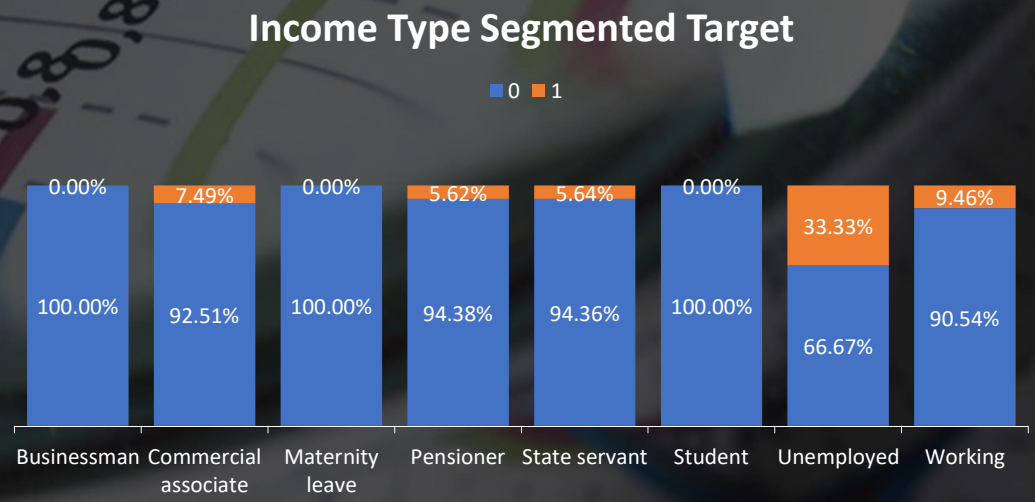
Univariate and Segmented Univariate Analysis



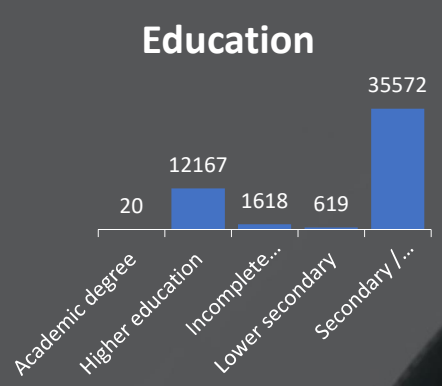


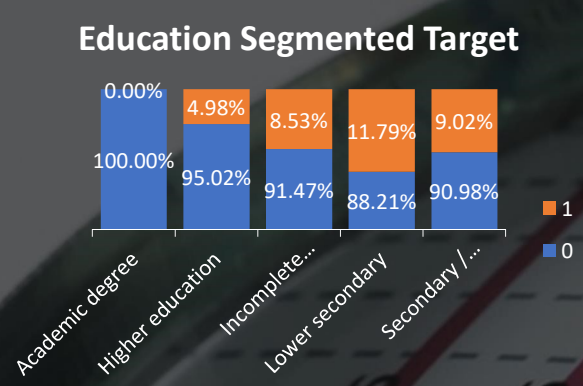


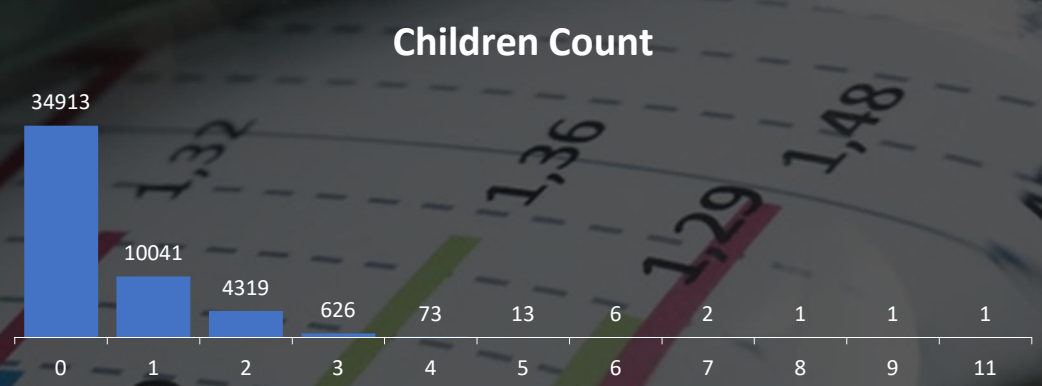
- Most people take Cash Loans
- Cash Loans has the highest Defaulter Percentage
- Majority of the Applied Loans are by Working Professionals
- Highest Defaulters are Unemployed
- Students and Businessmen are less likely to default



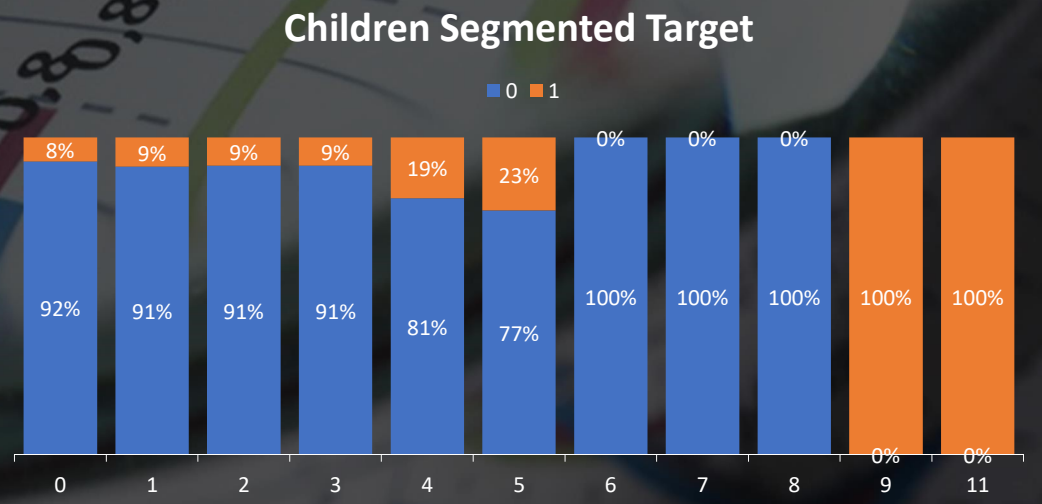
Univariate and Segmented Univariate Analysis



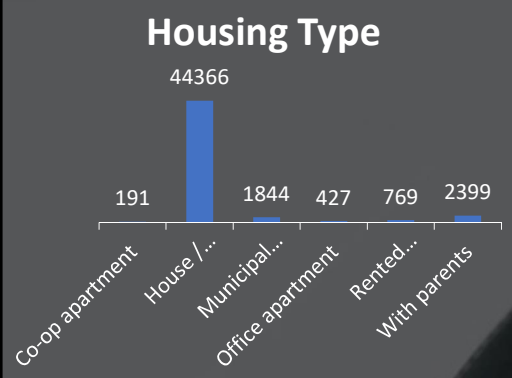


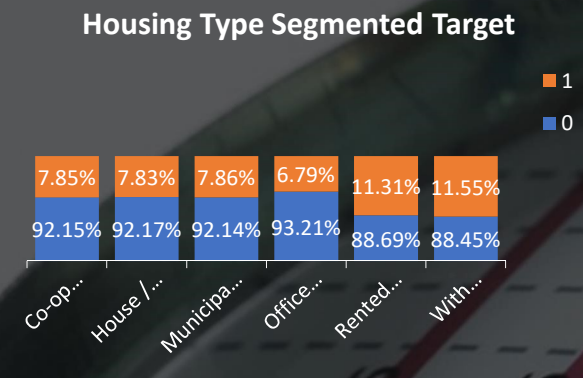


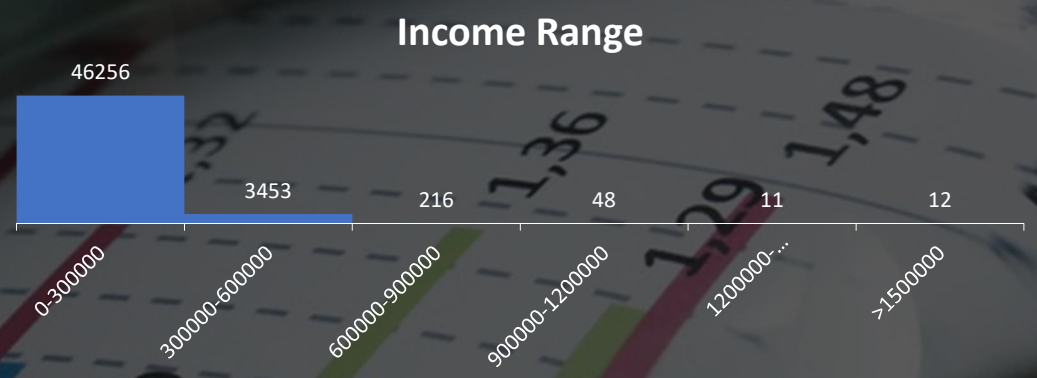
- Most people applying for loans have only had Secondary Education
- Lower Education has the highest defaulters
- Singles or People with no children are the highest loan appliers
- Outliers with over 8 kids are defaulters



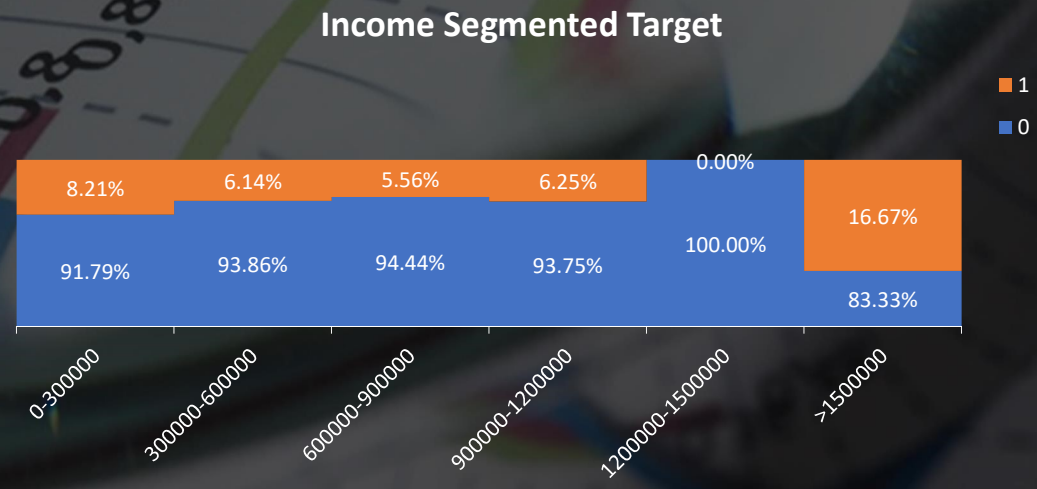
Univariate and Segmented Univariate Analysis



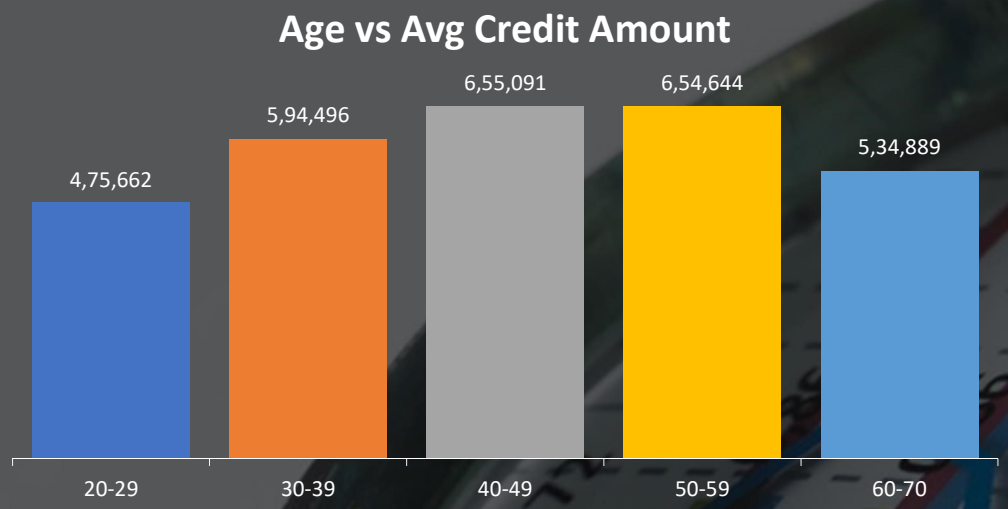


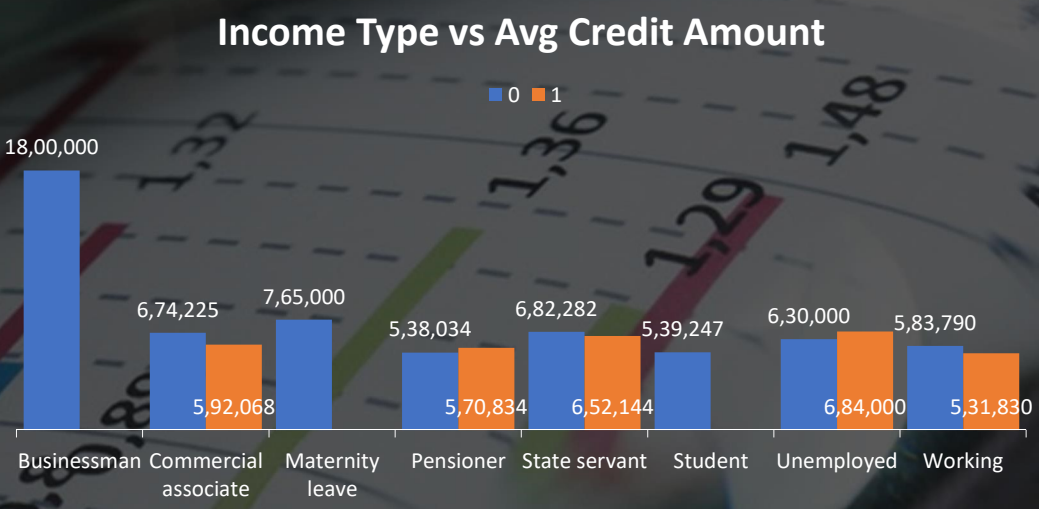


- Most people applying for loans stay in Apartments
- Most Defaulter Percentage stay with their parents or on rent
- Majority of the people applying for loans are in the low income group
- The High Income Groups are more likely to pay their loans, but the Very High Income Groups have the highest defaulter percentage



Bivariate Analysis





More credit is given to people as they grow older, but once they are reaching their retirement, the credit decreases. This is due to the fact that with age you default less

Businessmen get the highest credit amount whereas Students get the lowest

Task 5 : Identify Top Correlations for Different Scenarios

Correlation Matrix for Defaulters

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	AGE	DAYS_EMPLOYED	YEARS_WORKED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	DAYS_LAST_PHONE_CHANGE	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
CNT_CHILDREN	1																							
AMT_INCOME_TOTAL	0.01011018	1																						
AMT_CREDIT	0.00760191	0.01527144	1																					
AMT_ANNUITY	0.02917298	0.01800459	0.7496652	1																				
AMT_GOODS_PRICE	-0.00107967	0.0132695	0.98226796	0.74950403	1																			
REGION_POPULATION_RELATIVE	-0.02035915	-0.0061803	0.06777562	0.073124	0.07663549	1																		
DAYS_BIRTH	0.2496732	0.00903366	-0.14250603	-0.00875171	-0.1410059	-0.01646873	1																	
AGE	-0.24961576	-0.00844428	0.14238416	0.00886209	0.14086284	0.01653014	-0.99968736	1																
DAYS_EMPLOYED	-0.18932418	-0.01155596	0.01603957	-0.07955601	0.02023535	0.00774291	-0.58147904	0.58106649	1															
YEARS_WORKED	-0.18978832	-0.01173541	0.01877681	-0.07811983	0.02318228	0.00770574	-0.58825779	0.58785843	0.99960013	1														
DAYS_REGISTRATION	0.15211312	-0.00956115	-0.0428444	0.02158165	-0.04332022	-0.04613029	0.28843784	-0.28793306	-0.18871844	-0.19244489	1													
DAYS_ID_PUBLISH	-0.04236072	-0.00912201	-0.0437719	-0.02132109	-0.04972323	-0.00511856	0.24789657	-0.24792541	-0.23006367	-0.23266468	0.09029149	1												
CNT_FAM_MEMBERS	0.89252187	0.01312168	0.06124869	0.07583846	0.05513581	-0.01725715	0.1991414	-0.19900248	-0.18356011	-0.18337761	0.15178655	-0.04403782	1											
OBS_30_CNT_SOCIAL_CIRCLE	0.01793193	-0.01128092	0.03346617	0.01381902	0.03272397	-0.00887544	-0.01115023	0.01122317	0.00352185	0.00470808	-0.0057933	-0.02731374	0.03999054	1										
DEF_30_CNT_SOCIAL_CIRCLE	-0.01361871	-0.00797944	-0.02494668	-0.03454537	-0.01909661	0.02780592	-0.02083879	0.02099375	0.02985635	0.02977283	0.00099818	-0.02842652	-0.00645388	0.36507385	1									
OBS_60_CNT_SOCIAL_CIRCLE	0.01514587	-0.01121117	0.03443931	0.01409863	0.03387918	-0.007065	-0.01257029	0.01262327	0.00420868	0.00541281	-0.00592661	-0.02621248	0.0375377	0.99806585	0.36805994	1								
DEF_60_CNT_SOCIAL_CIRCLE	-0.0185057	-0.00672696	-0.02900724	-0.04047103	-0.02059292	0.02714232	-0.02575665	0.02585942	0.0238941	0.02378984	-0.00641263	-0.02789635	-0.00887718	0.29795102	0.89051161	0.30142085	1							
DAYS_LAST_PHONE_CHANGE	0.01133933	0.01245711	-0.12453934	-0.10047094	-0.12883245	-0.06710568	0.12460949	-0.12387487	-0.01573254	-0.01936405	0.07860465	0.13808778	-0.00573115	-0.0219161	0.00415783	-0.0230033	0.015271	1						
AMT_REQ_CREDIT_BUREAU_HOUR	-0.0002876	-0.00110418	0.01780636	0.03739749	0.01526195	0.00935622	0.02489871	-0.0249157	-0.00304646	-0.00355628	0.00638373	0.01407583	0.00486301	-0.01408506	0.00272831	-0.01358072	-0.01317789	0.0124725						
AMT_REQ_CREDIT_BUREAU_DAY	-0.03060525	-0.00144685	-0.0085184	-0.01868834	-0.00631921	-0.00383354	-0.02267042	0.02261487	0.04947762	0.04939249	-0.00147508	-0.00643299	-0.03314229	-0.01702919	0.01223625	-0.01744512	-0.01032405	-0.00679158	0.3511998	1				
AMT_REQ_CREDIT_BUREAU_WEEK	-0.03060405	-0.00221861	0.00012537	0.03472145	0.00011449	0.01206424	-0.00966098	0.00993477	0.0203878	0.020222	-0.01817817	0.01953762	-0.0280826	0.0058366	-0.01162117	0.00555602	-0.00393597	-0.00103167	0.01931666	0.06198856	1			
AMT_REQ_CREDIT_BUREAU_MON	0.008161	-0.00086402	0.0834082	0.07129522	0.0789087	0.0753956	-0.0072774	0.00748929	-0.03306561	-0.03208909	-0.001526	-0.03791731	0.01616533	0.01607779	0.00808821	0.01698449	0.01303479	-0.05674853	-0.00102354	-0.01629214	-0.00095331	1		
AMT_REQ_CREDIT_BUREAU_QRT	-0.0115206	-0.00374923	-0.01936131	-0.00163066	-0.02036764	0.01531017	-0.00878324	0.0087175	0.01787588	0.01760253	-0.00629042	-0.03267147	0.00246556	0.03483581	0.0201013	0.03640041	0.02534777	0.00042674	0.03109885	0.02553544	0.01169736	0.0199464	1	
AMT_REQ_CREDIT_BUREAU_YEAR	-0.03080113	-0.00510098	-0.01645997	0.00156927	-0.02347544	0.02402393	-0.09012732	0.08962933	0.01769246	0.01811595	-0.02509419	-0.08164306	-0.00506845	0.05051753	0.02101665	0.05070851	0.02062616	-0.10067865	0.00615969	0.02760701	0.03091697	0.0387895	0.10363174	1

Task 5 : Identify Top Correlations for Different Scenarios

Correlation Matrix for Non-Defaulters

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	AGE	DAYS_EMPLOYED	YEARS_WORKED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	DAYS_LAST_PHONE_CHANGE	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
CNT_CHILDREN	1																							
AMT_INCOME_TOTAL	0.0363546	1																						
AMT_CREDIT	0.005693	0.3779851	1																					
AMT_ANNUITY	0.026387	0.4511483	0.7707771	1																				
AMT_GOODS_PRICE	0.0015021	0.3846215	0.9870017	0.7758435	1																			
REGION_POPULATION_RELATIVE	-0.024923	0.1819764	0.0955331	0.1172846	0.0989572	1																		
DAYS_BIRTH	0.3359474	0.0737412	-0.051051	0.009911	-0.048736	-0.030412	1																	
AGE	-0.335761	-0.073614	0.0512097	-0.009707	0.0489161	0.0303618	-0.999707	1																
DAYS_EMPLOYED	-0.243613	-0.162693	-0.077379	-0.113004	-0.07512	-0.006618	-0.615291	0.6150742	1															
YEARS_WORKED	-0.245549	-0.161668	-0.074743	-0.111287	-0.07246	-0.006781	-0.623467	0.6232501	0.999533	1														
DAYS_REGISTRATION	0.1830823	0.0689103	0.0080533	0.0346046	0.0112642	-0.058497	0.335036	-0.334776	-0.204367	-0.208838	1													
DAYS_ID_PUBLISH	-0.032536	0.0323623	-0.008266	0.0094387	-0.00937	-0.002223	0.2700702	-0.270217	-0.272233	-0.274523	0.1035609	1												
CNT_FAM_MEMBERS	0.8792434	0.0416157	0.0648659	0.0778914	0.0628827	-0.022999	0.2844454	-0.284258	-0.233756	-0.234792	0.171489	-0.025058	1											
OBS_30_CNT_SOCIAL_CIRCLE	0.0161794	-0.033097	0.0008616	-0.010001	0.0004953	-0.019072	0.0123028	-0.012352	0.0056494	0.0055721	0.0109714	-0.011823	0.0242932	1										
DEF_30_CNT_SOCIAL_CIRCLE	-0.002833	-0.031999	-0.013516	-0.019745	-0.015219	0.0089004	0.0007102	-0.000764	0.0170239	0.0166534	0.003453	0.0023139	-0.002824	0.3061583	1									
OBS_60_CNT_SOCIAL_CIRCLE	0.0163343	-0.033069	0.0011701	-0.009685	0.0007178	-0.018015	0.0123099	-0.012352	0.0055107	0.0054417	0.0112891	-0.012125	0.0245776	0.9983575	0.3085654	1								
DEF_60_CNT_SOCIAL_CIRCLE	-0.003341	-0.032523	-0.018573	-0.023009	-0.019743	0.0032491	0.0022298	-0.002301	0.016508	0.0161195	0.0062859	0.0026435	-0.004596	0.2291725	0.850995	0.2312825	1							
DAYS_LAST_PHONE_CHANGE	-0.004803	-0.049511	-0.071191	-0.064449	-0.074233	-0.044133	0.0725007	-0.072429	0.0329686	0.0291783	0.0477788	0.0850658	-0.025007	-0.014342	0.0025038	-0.015119	0.0022878	1						
AMT_REQ_CREDIT_BUREAU_HOUR	0.0026143	0.0081277	3.49E-05	0.0101412	0.000809	-0.003133	0.0014926	-0.001468	-0.004296	-0.004401	-0.003688	0.0028246	0.0036848	0.0023638	-0.004399	0.0025849	-0.0032	-0.00128	1					
AMT_REQ_CREDIT_BUREAU_DAY	0.0011966	0.0094819	0.0134851	0.0091571	0.0136393	-0.00034	0.0019892	-0.001957	0.0016163	0.0015184	-0.003383	0.0035151	0.0006473	0.0009729	0.0036862	0.0008657	0.0027768	-0.000453	0.2307631	1				
AMT_REQ_CREDIT_BUREAU_WEEK	0.0043217	0.009497	0.0053717	0.0189105	0.0058066	0.0026421	-0.002388	0.0022825	-0.006486	-0.006243	0.0006618	-0.004665	0.0061138	-0.004288	-0.005038	-0.004878	-0.005732	-0.005992	0.012125	0.2491225	1			
AMT_REQ_CREDIT_BUREAU_MON	-0.011618	0.074876	0.0639709	0.0379869	0.0657025	0.070733	-0.002431	0.0023181	-0.032963	-0.032245	-0.010721	-0.013232	-0.00451	0.0081697	0.0076822	0.0081263	0.0039669	-0.047332	0.0095461	-0.000655	-0.010603	1		
AMT_REQ_CREDIT_BUREAU_QRT	-0.004724	0.0157967	0.0267999	0.010067	0.027519	-0.009716	-0.021539	0.0216167	0.0145838	0.0146865	0.0031317	-0.024582	-0.004241	0.0088515	0.0053543	0.0086808	0.0083102	-0.012883	0.003519	-0.007869	-0.014597	0.011892	1	
AMT_REQ_CREDIT_BUREAU_YEAR	-0.035752	0.0313498	-0.031578	-0.004172	-0.034428	0.0046453	-0.070236	0.0702236	0.0441713	0.0443584	-0.022957	-0.044693	-0.022929	0.0341607	0.014498	0.0345735	0.0151975	-0.11761	0.0040933	-0.000859	0.024733	0.0193038	0.1217536	1


Results

Results

This project involved an extensive use of Excel and Statistics. It was an enormous challenge trying to work in Excel and manipulate such massive data.

Nonetheless, the project was very insightful and informative. It allowed me to learn to work with such datasets.

Throughout the project, I got a deeper understanding on EDA, right from analysing the data columns, to handling missing data and creating new analysis.

I was also able to learn about new add-ins in Excel such as Data Analyzer.

Link to Excel File


THANK YOU