BANK LOAN CASE STUDY

-ANDRI

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.

- **When a customer applies for a loan, a company faces two risks:**

1. If the applicant can repay the loan but is not approved, the company loses business.
2. If the applicant cannot repay the loan and is approved, the company faces a financial loss.

Approach and tech stack used

- Understanding the data.
- Cleaning / pre-processing the data and handling all NULL values and outliers.
- Merging the datasets/csvs to gain deeper insights.
- Data analysis and visualization.
- Gaining insights/hypothesis from the analysis.

I have used MS-Office 2019 for analysis and powerpoint presentation.

Understanding the dataset

- Our dataset consists of three .csv files providing information of various aspects of the loan application mentioned below:

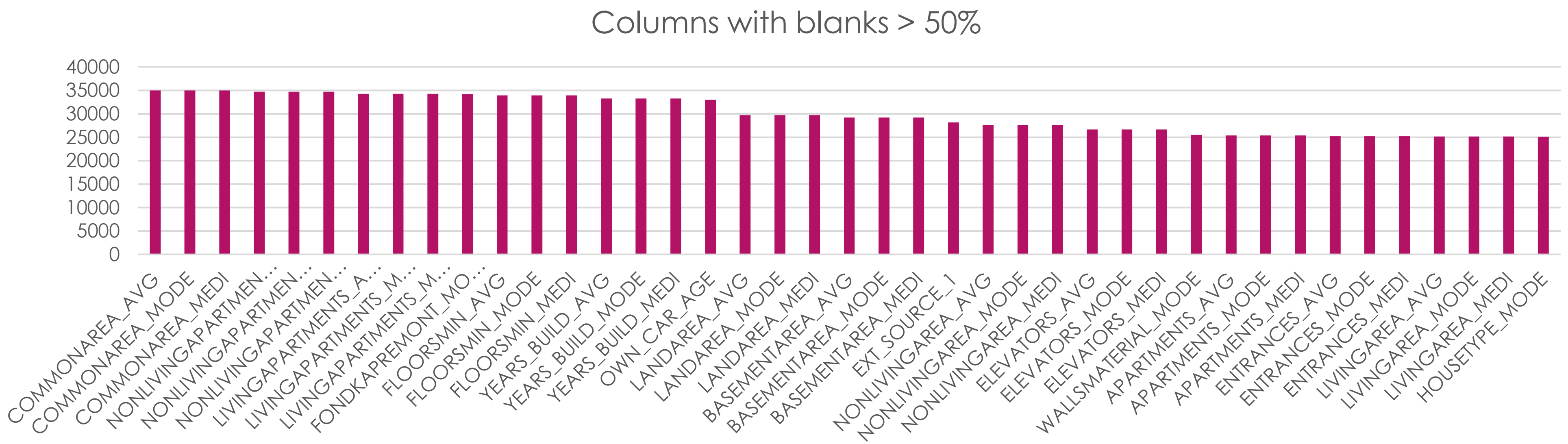
1. application_data.csv: gives information about current loan applications.
2. previous_application.csv: has information about client’s previous loan data.
3. columns_description.csv: contains information about columns present in the above two datas.

Application data file

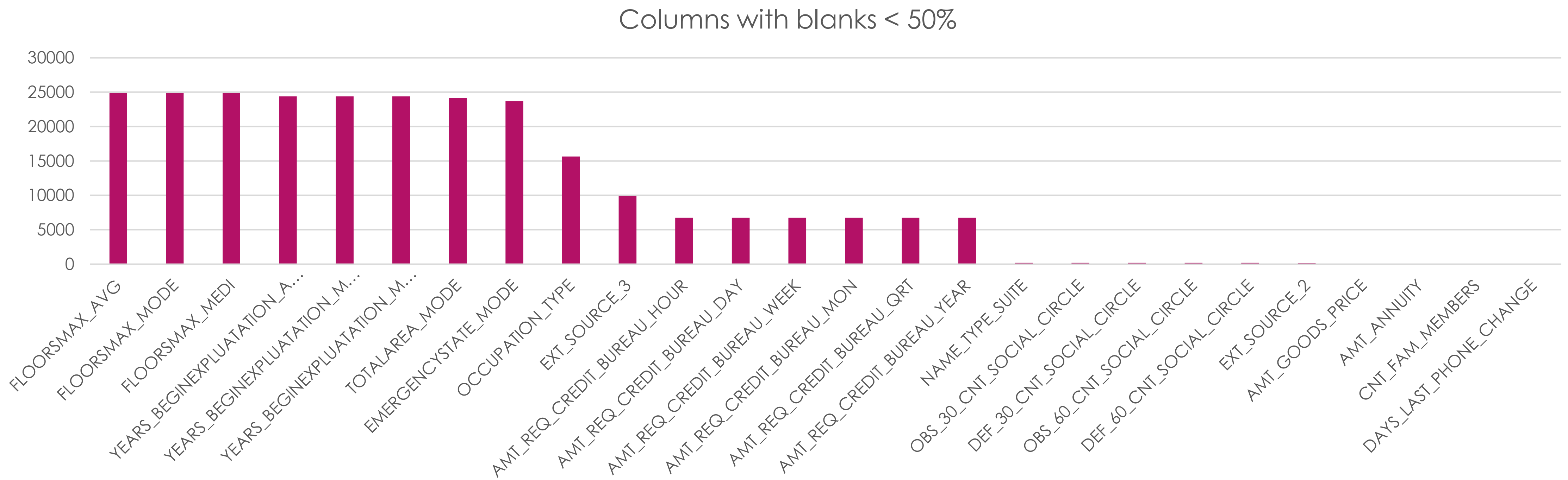
- Rows: 50000
- Columns: 122
- Columns with blank cells: 67
- Columns with high number of blank cells (>50% of total rows): 41
- Columns with less number of blank cells (<50% of total rows): 26s

Data cleaning

- Identifying the columns (41 in number) with a high number of blank cells will be **dropped** as imputation would not work on it.



Data cleaning



Data cleaning

- Out of the remaining 26 columns, certain columns are not required for analysis such as:

FLOORSMAX_AVG
FLOORSMAX_MODE
FLOORSMAX_MEDI
YEARS_BEGINEXPLUATATION_AVG
YEARS_BEGINEXPLUATATION_MODE
YEARS_BEGINEXPLUATATION_MEDI
TOTALAREA_MODE
EMERGENCYSTATE_MODE
EXT_SOURCE_3
EXT_SOURCE_2

- Dropping these 10 columns, we are left with 16 columns that need to be imputed with either **mean, median or mode for numerical** and **mode for categorical** columns.

Data Cleaning

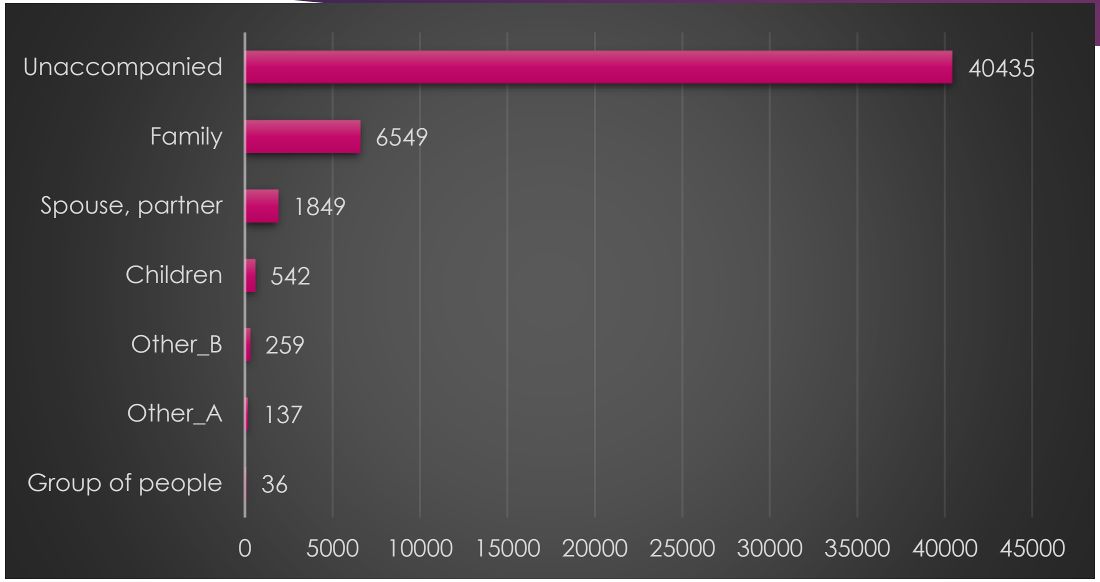
- Dropping the blank rows from the below columns as there is just one blank cell.

AMT_ANNUITY
CNT_FAM_MEMBERS
DAYS_LAST_PHONE_CHANGE

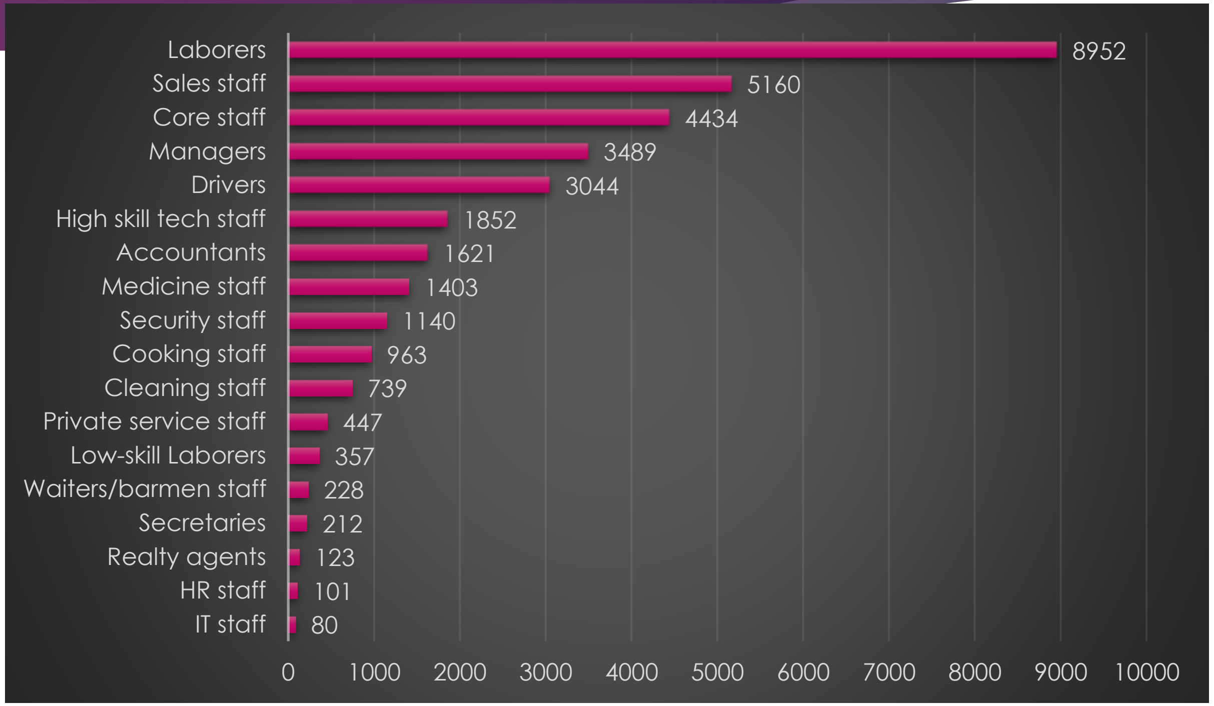
- Features to impute finally 13 columns:

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
NAME_TYPE_SUITE
OBS_30_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
OCCUPATION_TYPE
AMT_GOODS_PRICE

Mode Imputation:



NAME_TYPE_SUITE blank cells imputed with "Unaccompanied"

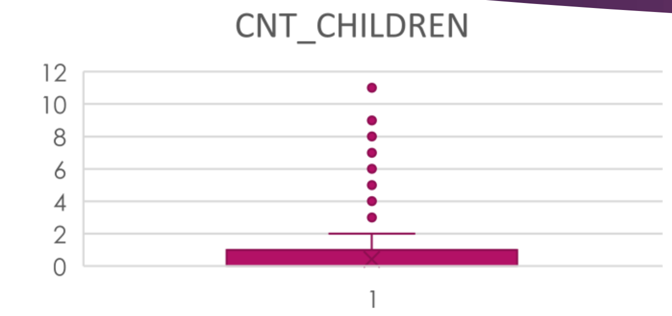


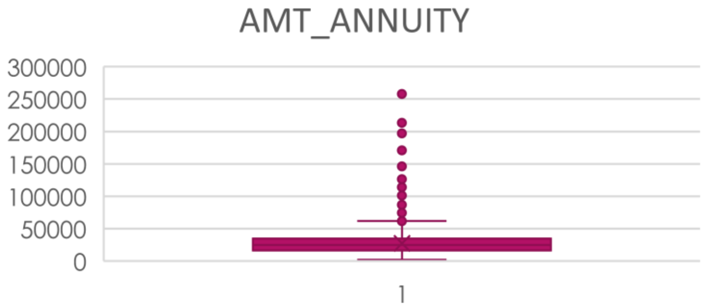
OCCUPATION_TYPE column has high number of blank cells (~15k) which is more than the highest occupation category hence imputed with "Unknown"

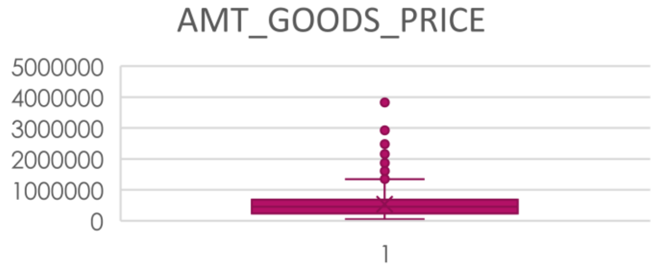
Median imputation

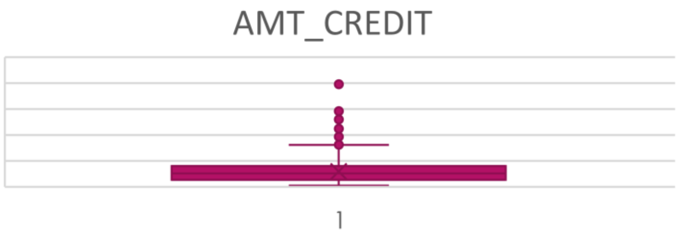
- AMT_REQ_CREDIT_BUREAU_HOUR : 0
- AMT_REQ_CREDIT_BUREAU_DAY : 0
- AMT_REQ_CREDIT_BUREAU_WEEK : 0
- AMT_REQ_CREDIT_BUREAU_MON : 0
- AMT_REQ_CREDIT_BUREAU_QRT : 0
- AMT_REQ_CREDIT_BUREAU_YEAR : 1
- OBS_30_CNT_SOCIAL_CIRCLE : 0
- DEF_30_CNT_SOCIAL_CIRCLE : 0
- OBS_60_CNT_SOCIAL_CIRCLE : 0
- DEF_60_CNT_SOCIAL_CIRCLE : 0
- AMT_GOODS_PRICE : 450000

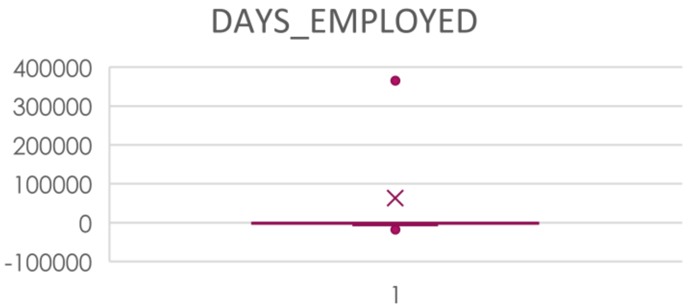
Outliers removal

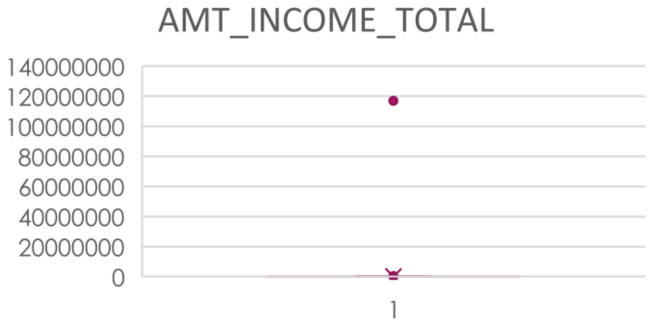








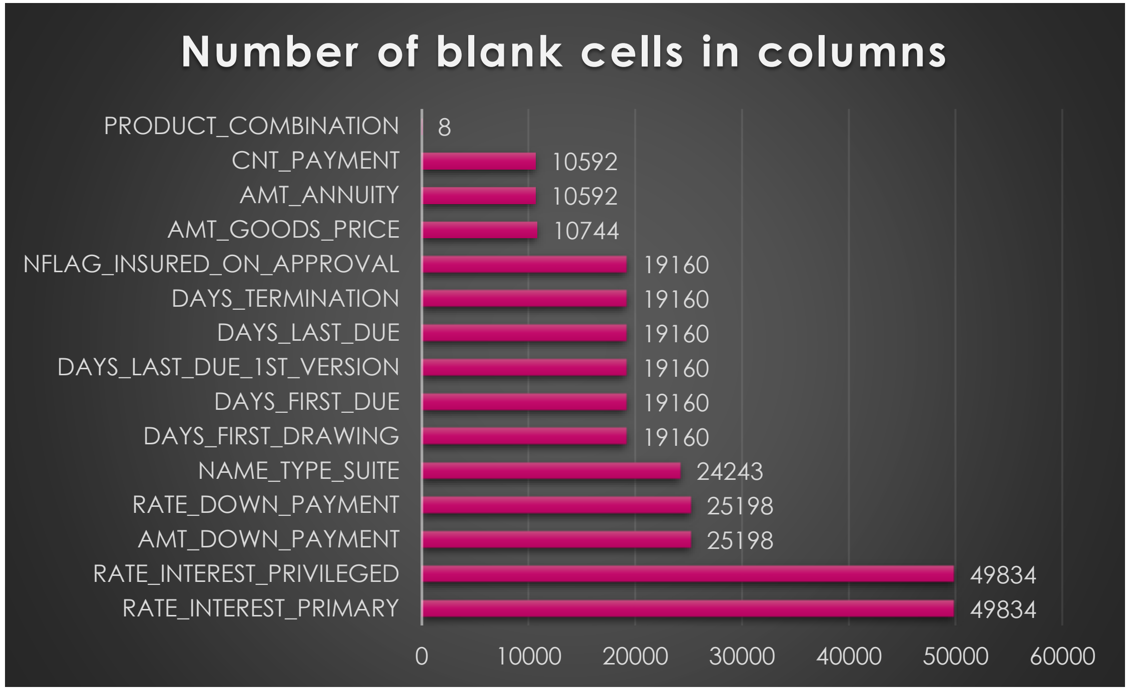




- Removing the outliers from the numerical columns
- CNT_CHILDREN in today's age can't be more than 3-4, in the dataset it goes up to 11, I have taken max 5 children.
- Removing vague values from DAYS_EMPLOYED **(~8k)** and AMT_INCOME_TOTAL that are too high, 1000 years for DAYS_EMPLOYED and 117000000 for AMT_INCOME_TOTAL.
- Removing the outliers from other numerical columns such as AMT_ANNUITY, AMT_GOODS_PRICE, AMT_CREDIT, DAYS_LAST_PHONE_CHANGE.

Previous_application data file

- No. of columns: 37
- No. of rows: 50000
- No. of columns with blank: 15



Data cleaning

- Dropping columns with more than 50% of blank cells:

RATE_INTEREST_PRIMARY
RATE_INTEREST_PRIVILEGED
AMT_DOWN_PAYMENT
RATE_DOWN_PAYMENT

- Dropping these 4 columns, we are left with 33 columns, out of which we can impute the remaining columns:

NAME_TYPE_SUITE
DAYS_FIRST_DRAWING
DAYS_FIRST_DUE
DAYS_LAST_DUE_1ST_VERSION
DAYS_LAST_DUE
DAYS_TERMINATION
NFLAG_INSURED_ON_APPROVAL
AMT_GOODS_PRICE
AMT_ANNUITY
CNT_PAYMENT
PRODUCT_COMBINATION

Data cleaning

- Dropping unnecessary columns which don't provide any information:

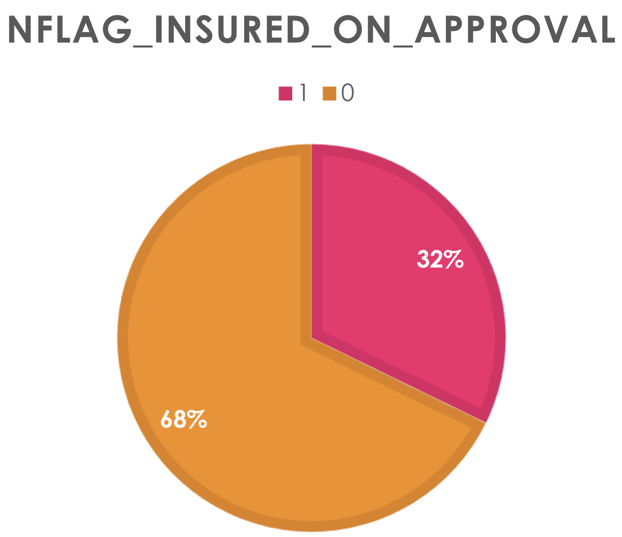
1. NAME_TYPE_SUITE
2. WEEKDAY_APPR_PROCESS_START
3. HOUR_APPR_PROCESS_START
4. FLAG_LAST_APPL_PER_CONTRACT
5. NFLAG_LAST_APPL_IN_DAY

Median imputation

1. AMT_ANNUITY
2. AMT_GOODS_PRICE
3. DAYS_FIRST_DRAWING
4. DAYS_FIRST_DUE
5. DAYS_LAST_DUE_1ST_VERSION
6. DAYS_LAST_DUE DAYS_TERMINATION

MODE IMPUTATION

- NFLAG_INSURED_ON_APPROVAL

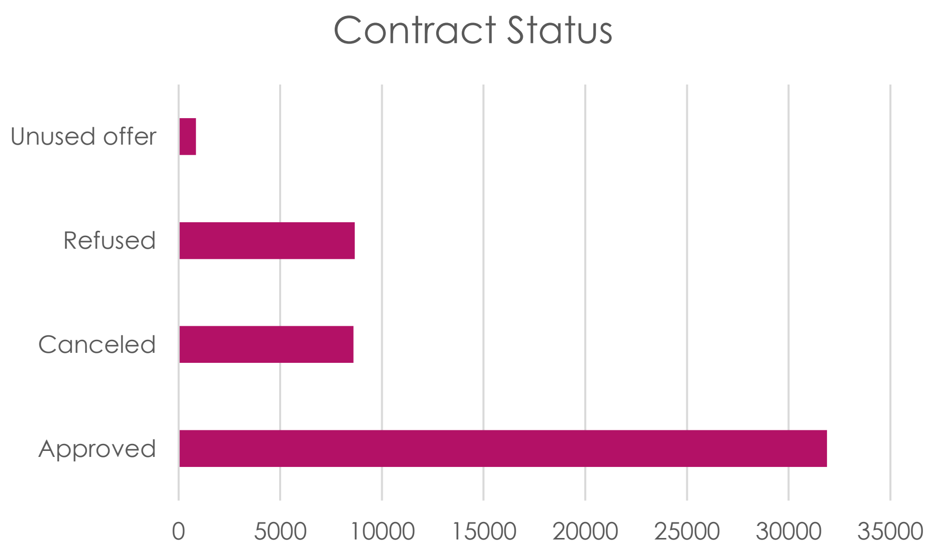


Data cleaning: Dropping

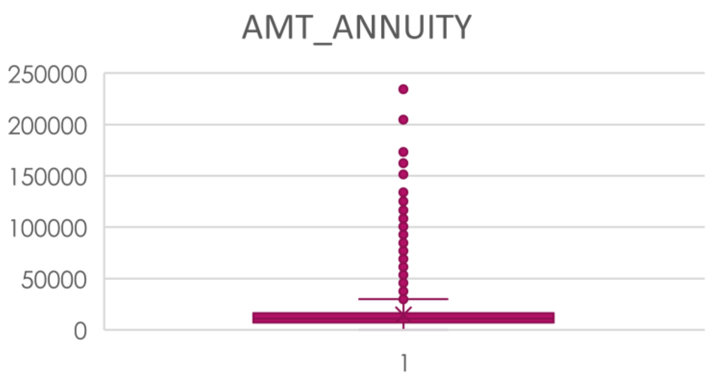
- Dropping the blank cells from “PRODUCT_COMBINATION” feature as the blanks were just 8 in number and dropping would be better than imputation.

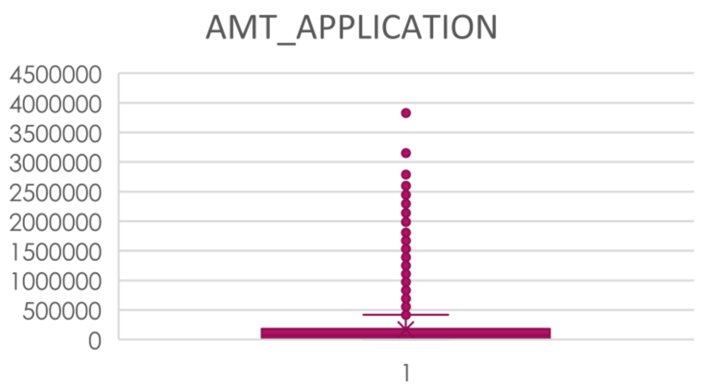
Data cleaning: Custom imputation

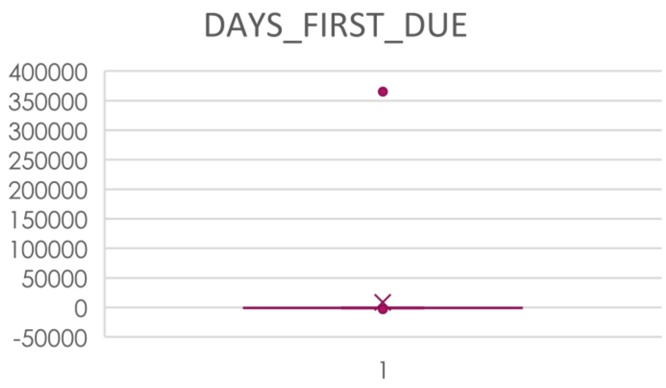
- On analyzing the contract status for the CNT_PAYMENT, most of the contracts were approved, hence imputing CNT_PAYMENT with median would be better.

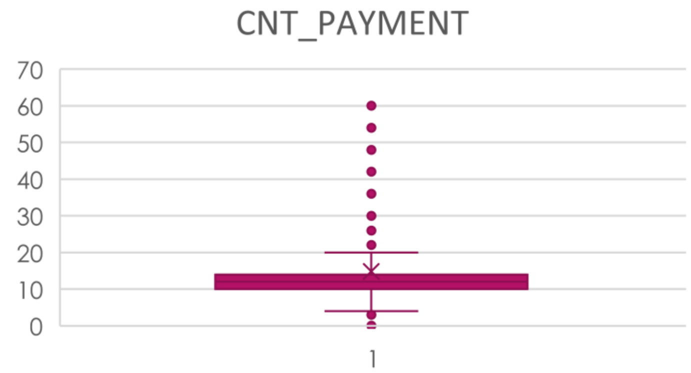


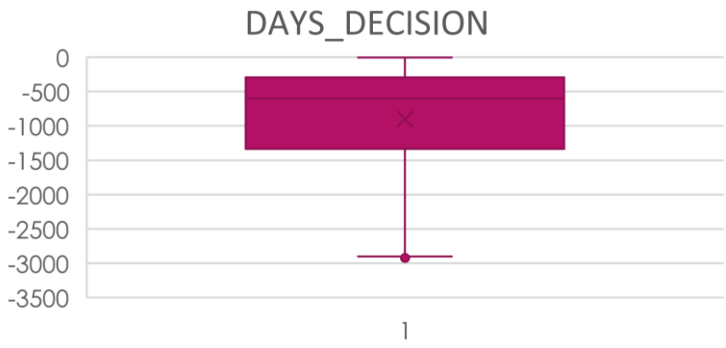
Outliers







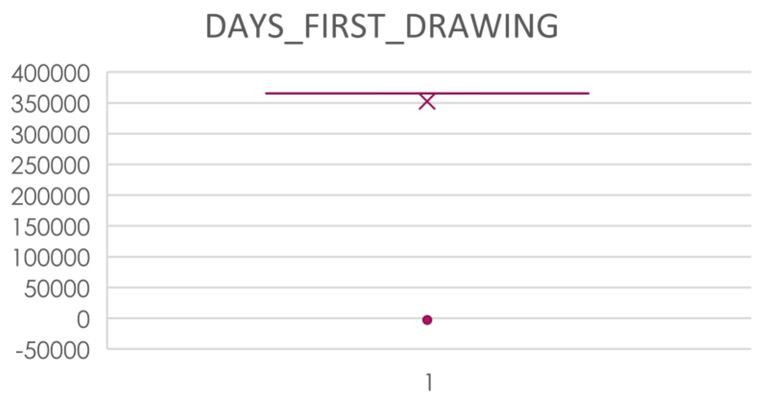


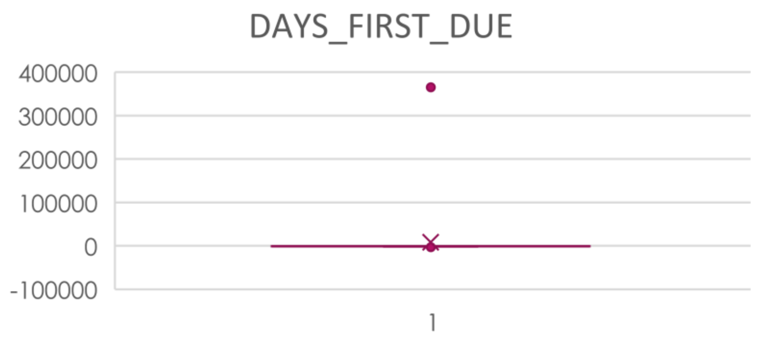


- There are many outliers in the columns like AMT_ANNUITY, AMT_APPLICATION etc.
- Few outliers in CNT_PAYMENT.
- Few outliers in DAYS_DECISION are such that it indicates that the decision time taken is high, which is not a good practice.

Removing outliers

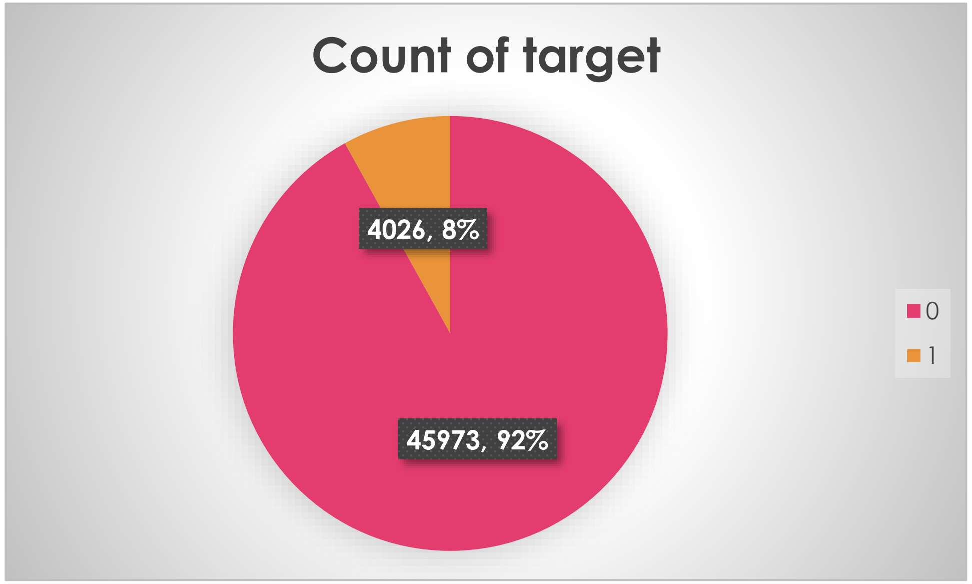
- I have cleaned the outliers from the columns such as DAYS_FIRST_DRAWING, DAYS_FIRST_DUE, DAYS_LAST_DUE_1ST_VERSION, DAYS_LAST_DUE, DAYS_TERMINATION as the outliers were unrealistic values.





Data imbalance

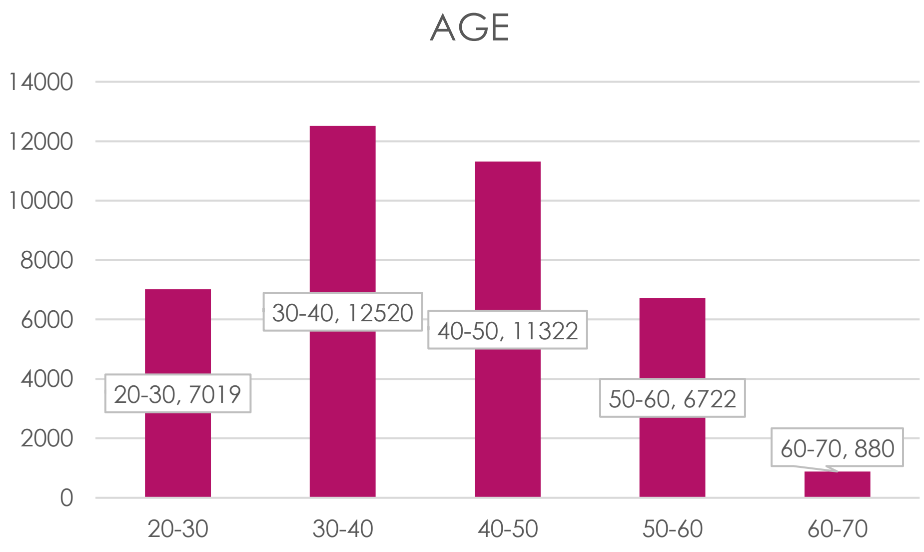
As we can see there is data imbalance and the number of defaulters is way less than the number of re-payers in this dataset provided.



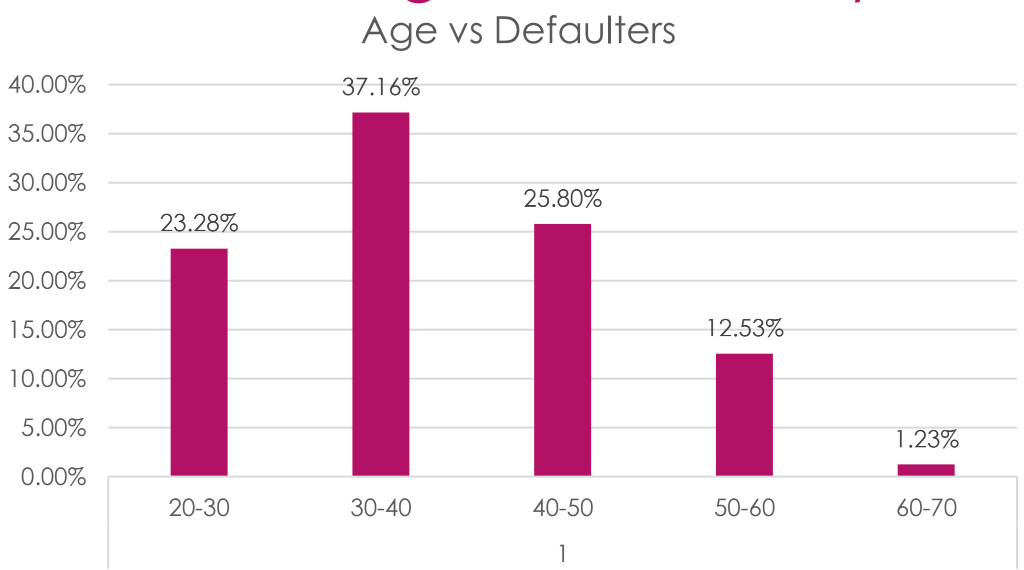
UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS: Age

Most of the loan applicants are from age group 30-40 years. As the age increases the rate of defaulting decreases.

Univariate analysis



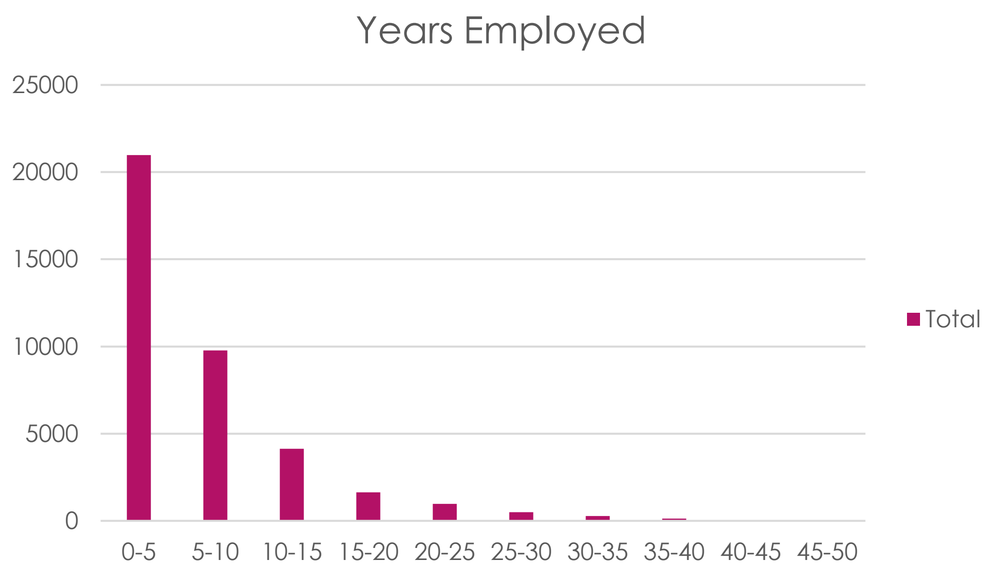
Univariate segmented analysis



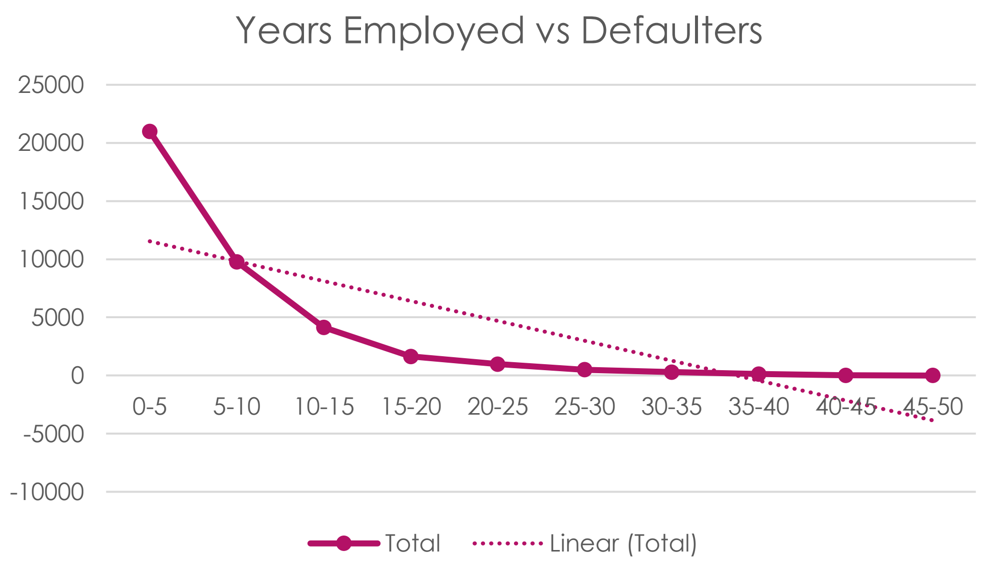
UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS: Years Employed

Most of the loan applicants are from 0-5 years of experience. As the experience increases the rate of defaulting decreases

Univariate analysis



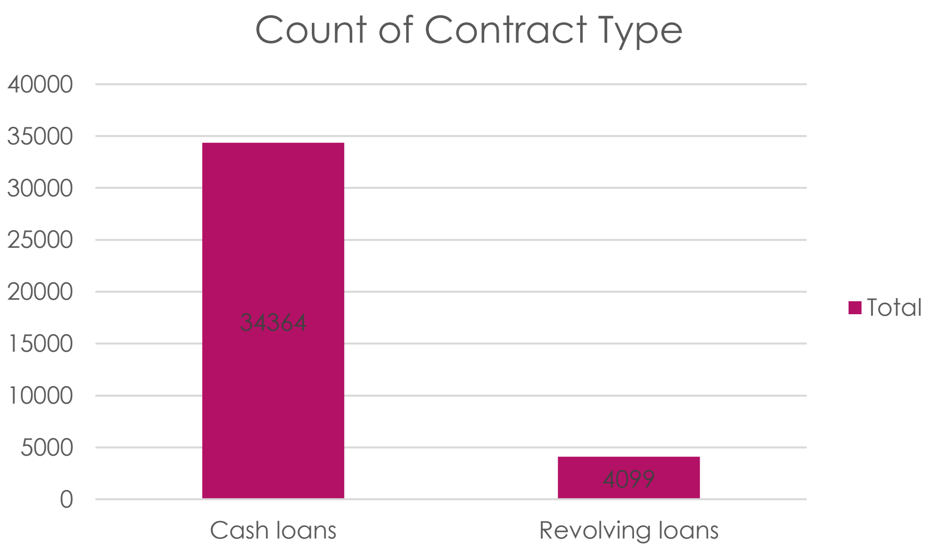
Univariate segmented analysis



UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS: Contract type

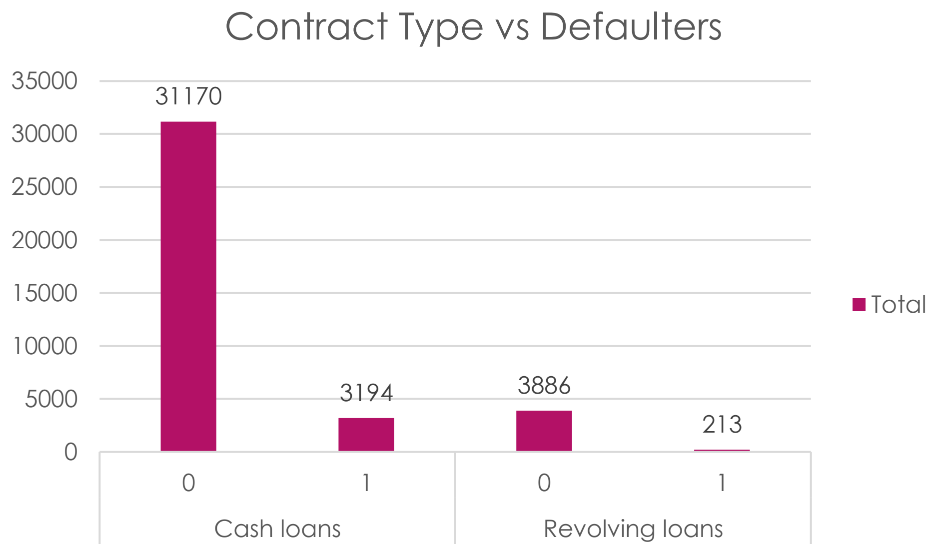
Univariate analysis

- Highest are cash loans.



Univariate segmented analysis

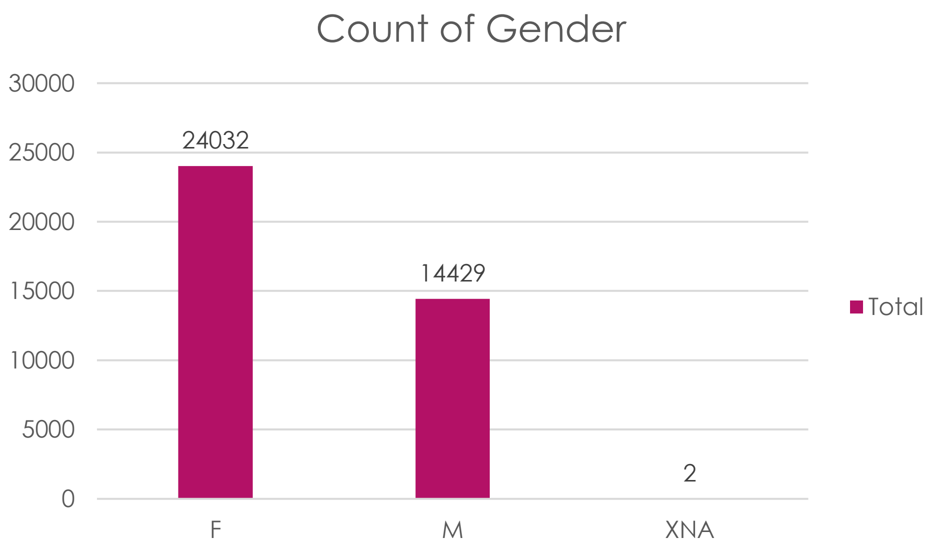
- High defaulters are from Cash loans
- Cash loans: 9.2%
- Revolving loans: 5.1%



UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS: Gender

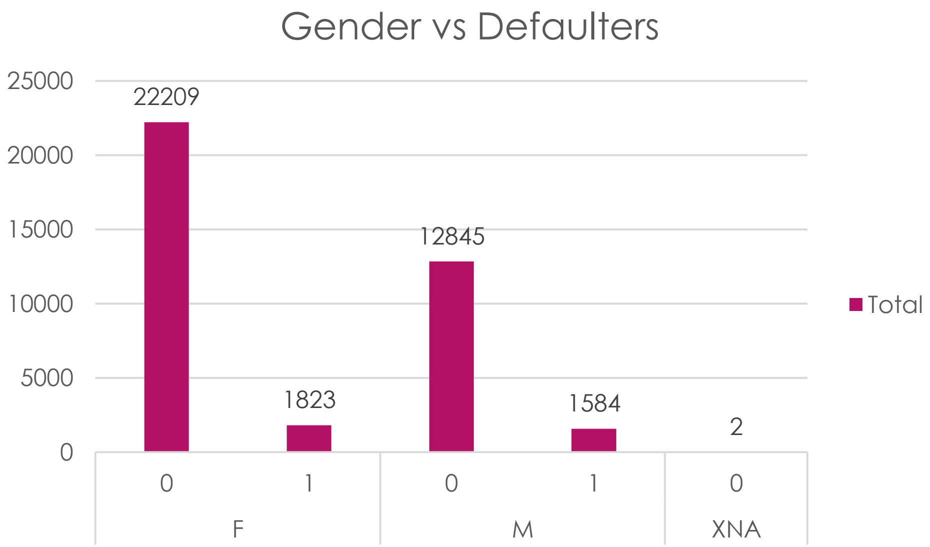
Univariate analysis

- Most of the females have taken loan.



Univariate segmented analysis

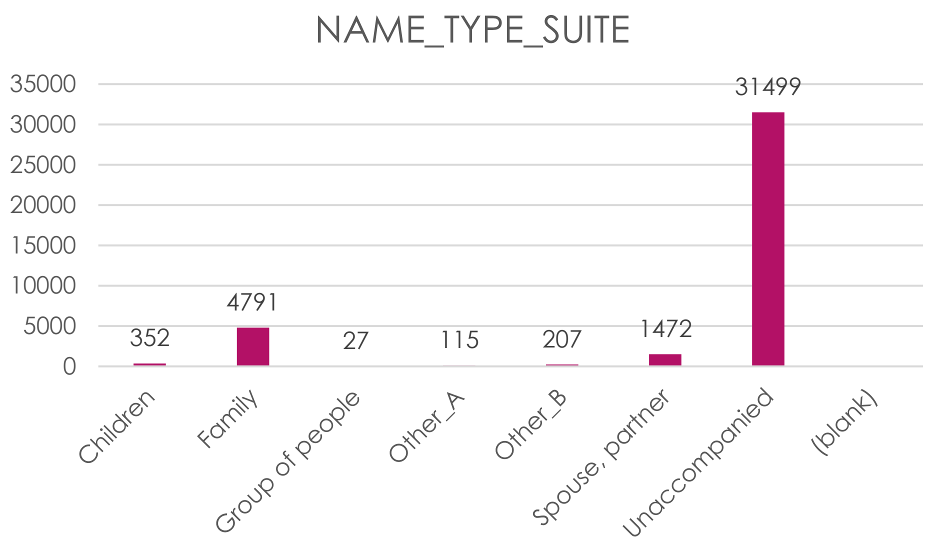
- Female defaulters: 7.5%
- Male defaulters: ~11%



UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS: NAME_TYPE_SUITE

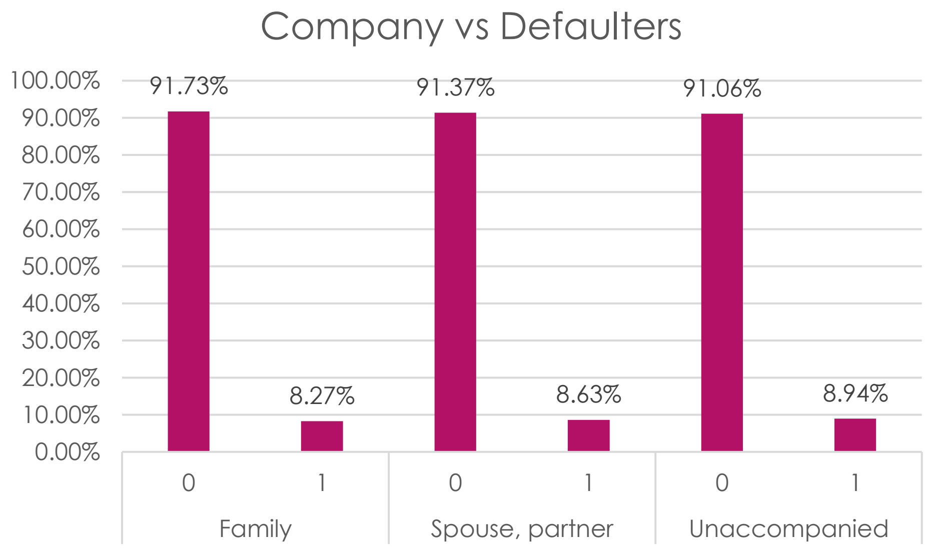
Univariate analysis

- Most of the people were unaccompanied while taking loan.



Univariate segmented analysis

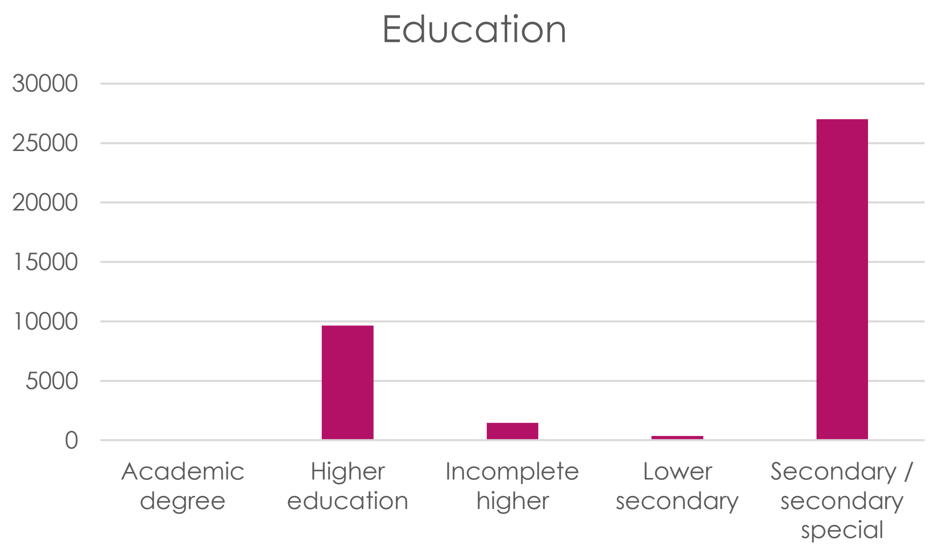
- Here I have shown the categories with significant numbers.



UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS: Education

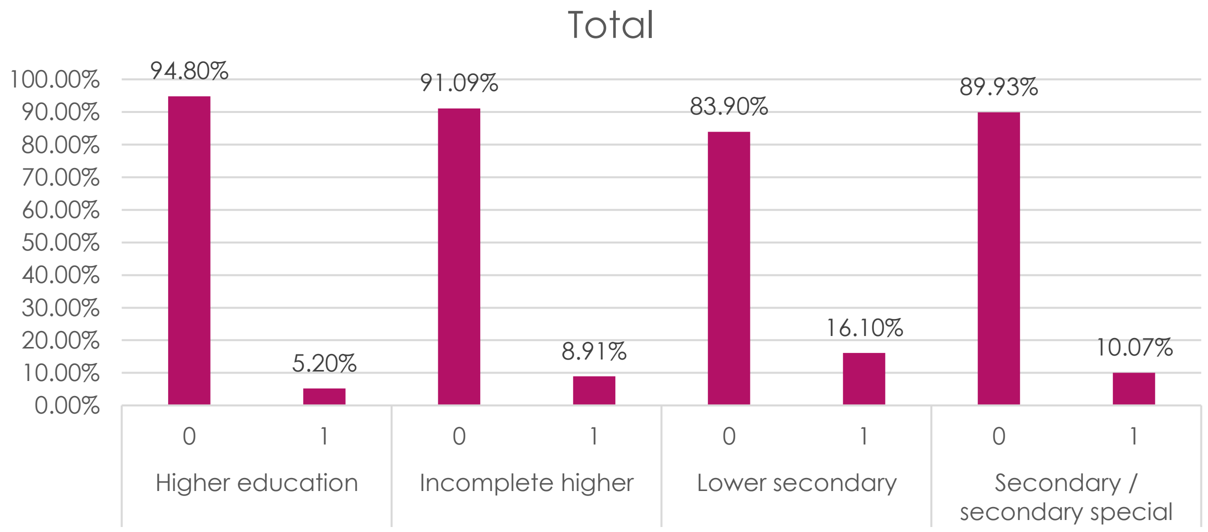
Univariate analysis

- Most of the customers have secondary and higher education.



Univariate segmented analysis

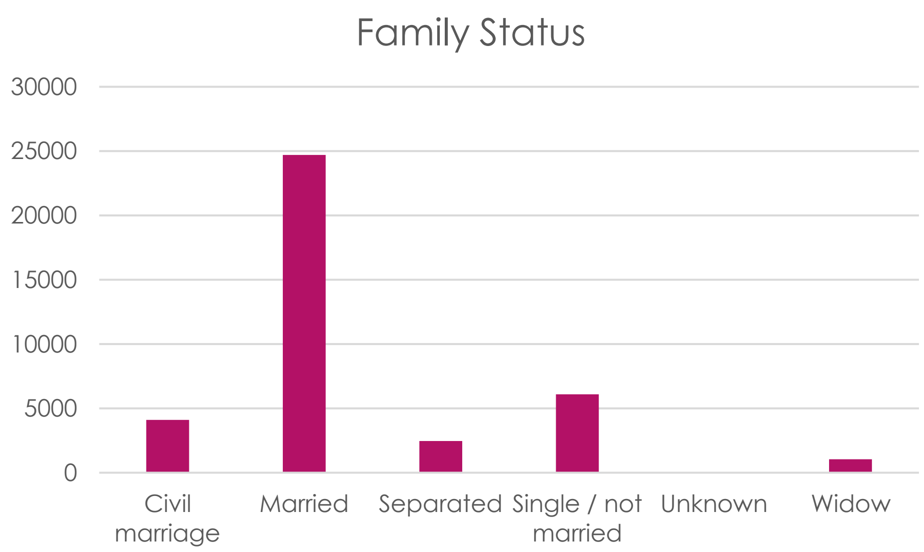
- Highest and lowest defaulters have lower secondary (16%) and academic degree(0%) respectively.



UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS: Family Status

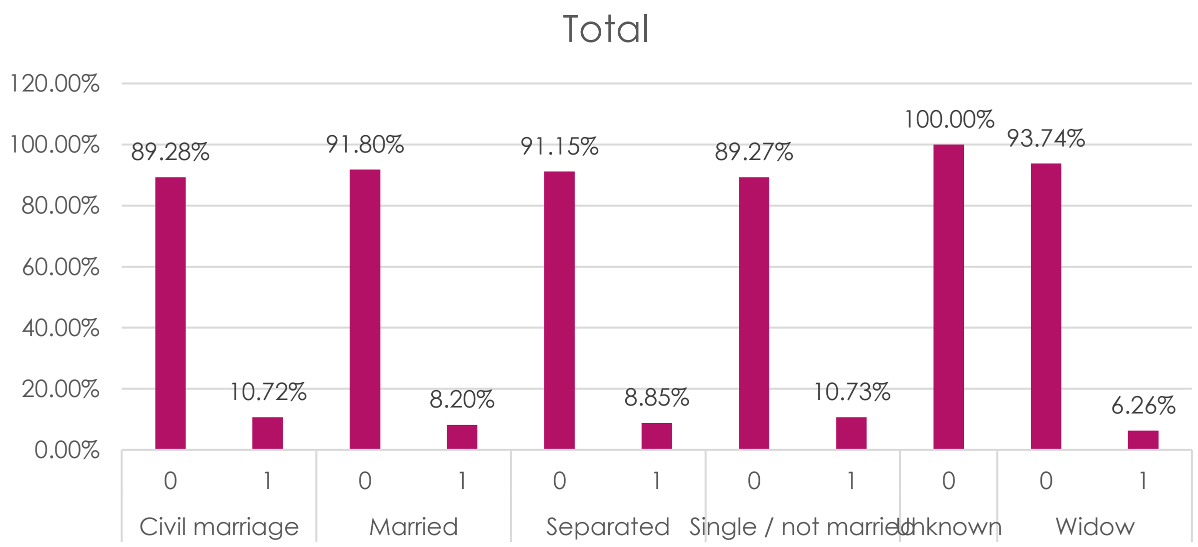
Univariate analysis

- Highest number of people who took loans are married.



Univariate segmented analysis

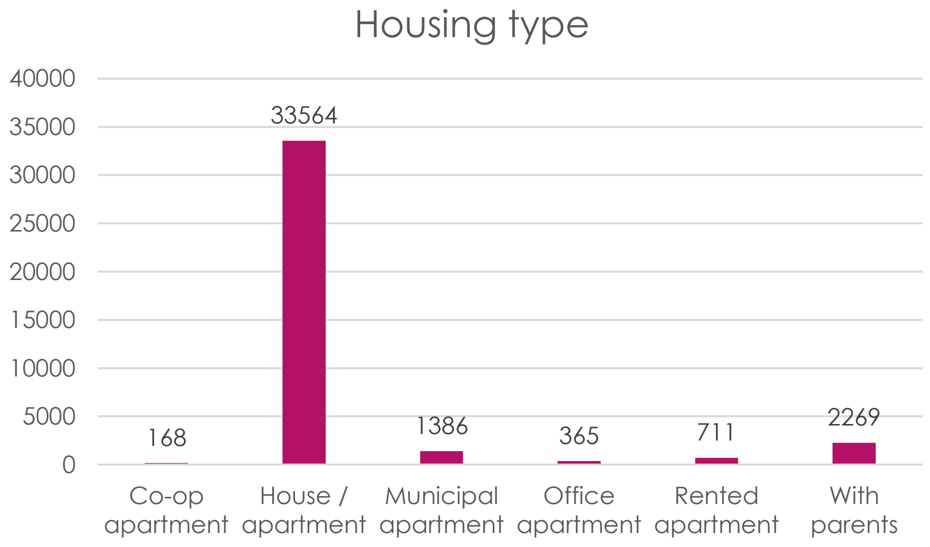
- Highest default percent is in civil marriage and single people ~10%



UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS: Housing Type

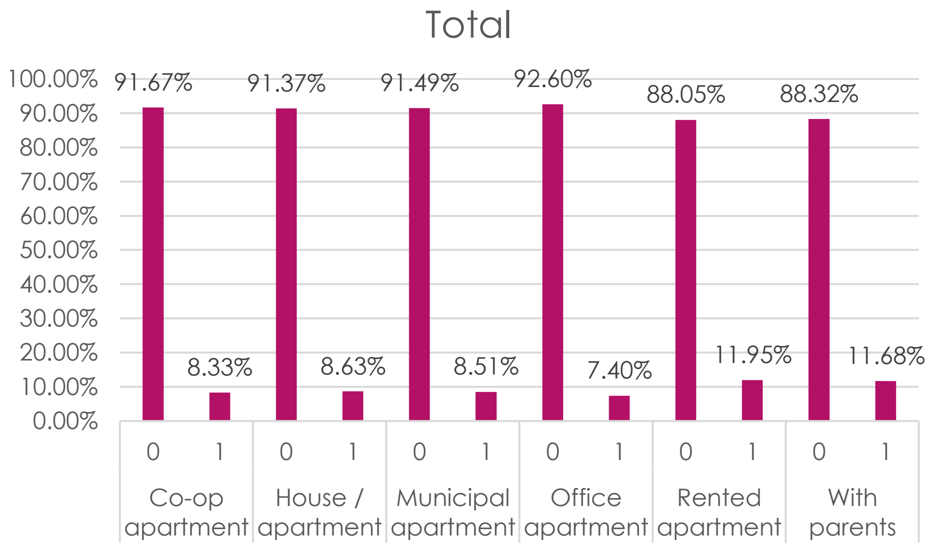
Univariate analysis

- People who took loans had houses/apartments.



Univariate segmented analysis

- Highest defaulters are from rented apartments and with parents ~11%



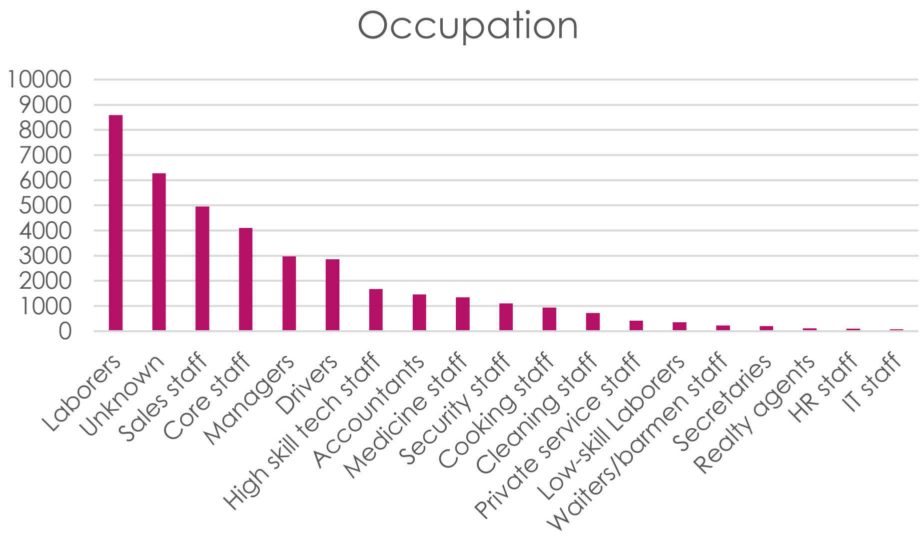
UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS: Occupation

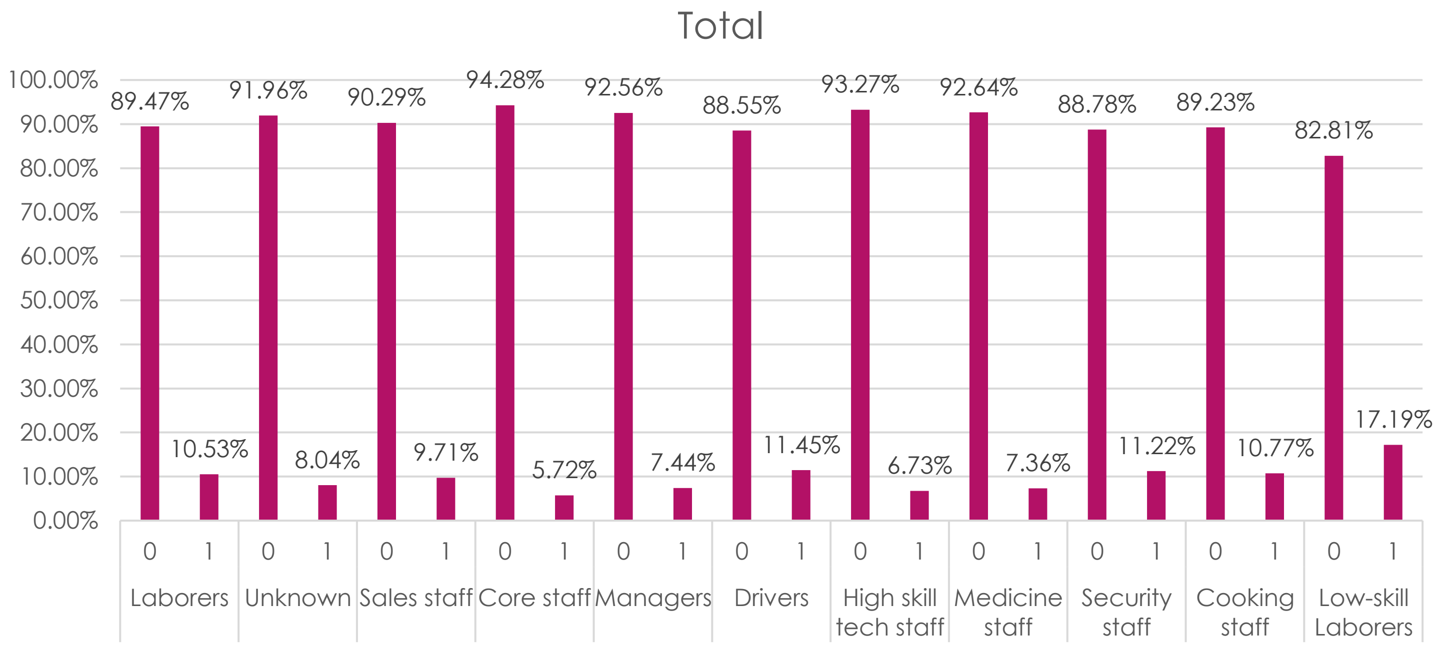
Univariate analysis

- Laborers are the highest loan taking occupation.

Univariate segmented analysis

- Highest defaulter percent is for low skilled labourers ~17%.

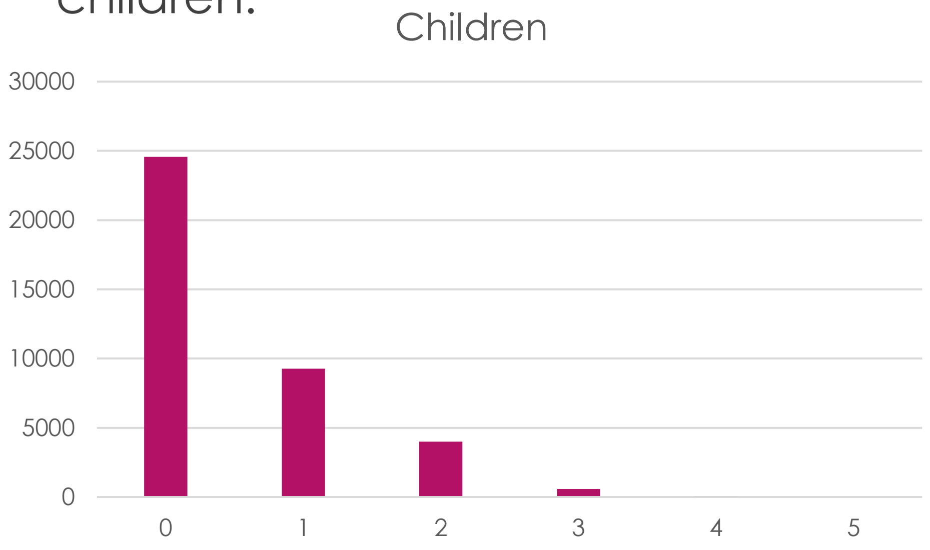




UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS: Children

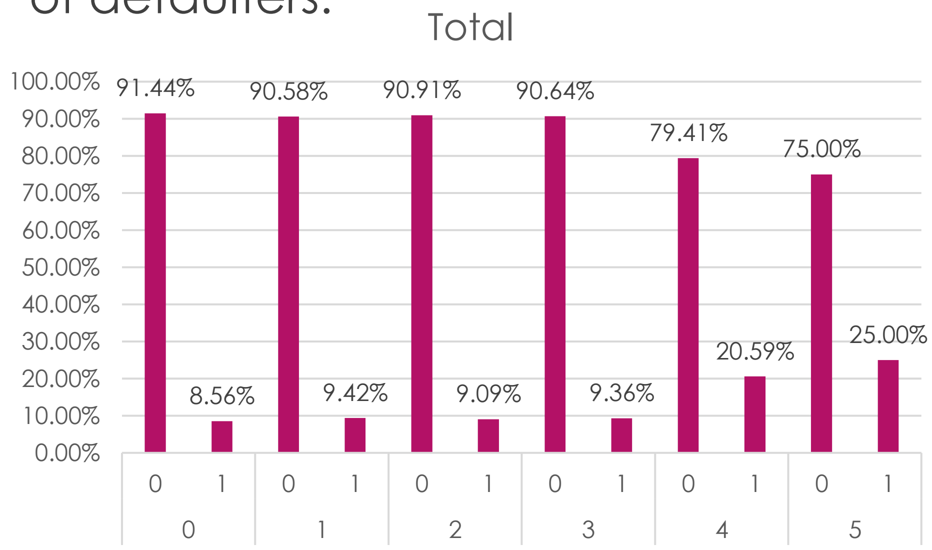
Univariate analysis

- Most of the people taking loans are either having no child or one/two children.



Univariate segmented analysis

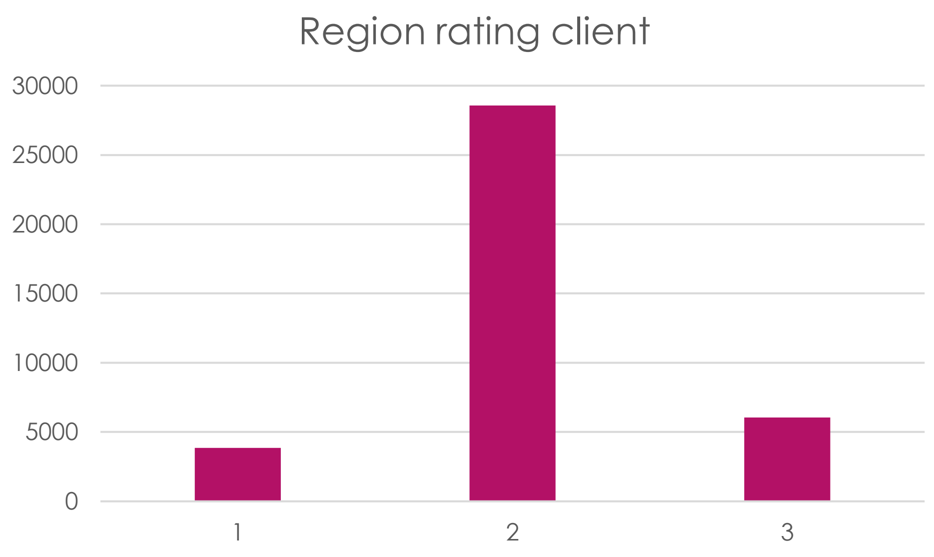
- The highest defaulters are people having more number of children 25% of defaulters.



UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS: Region_Rating_Client

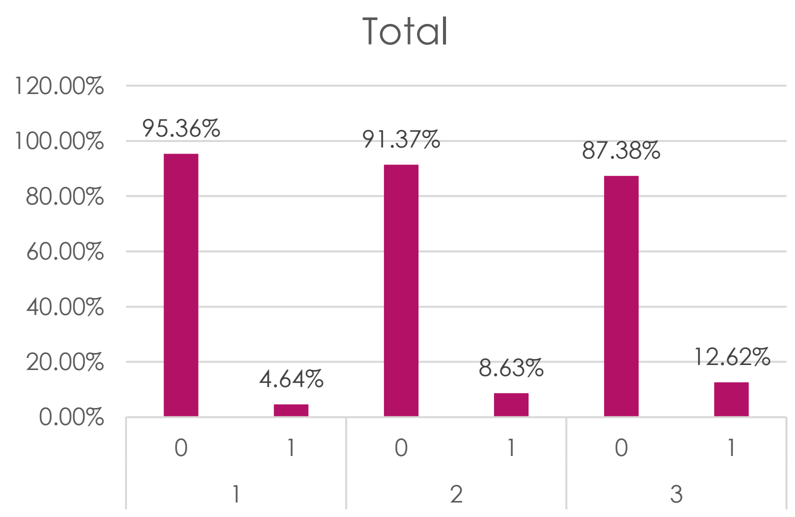
Univariate analysis

- Most of the people taking loans are from region 2.



Univariate segmented analysis

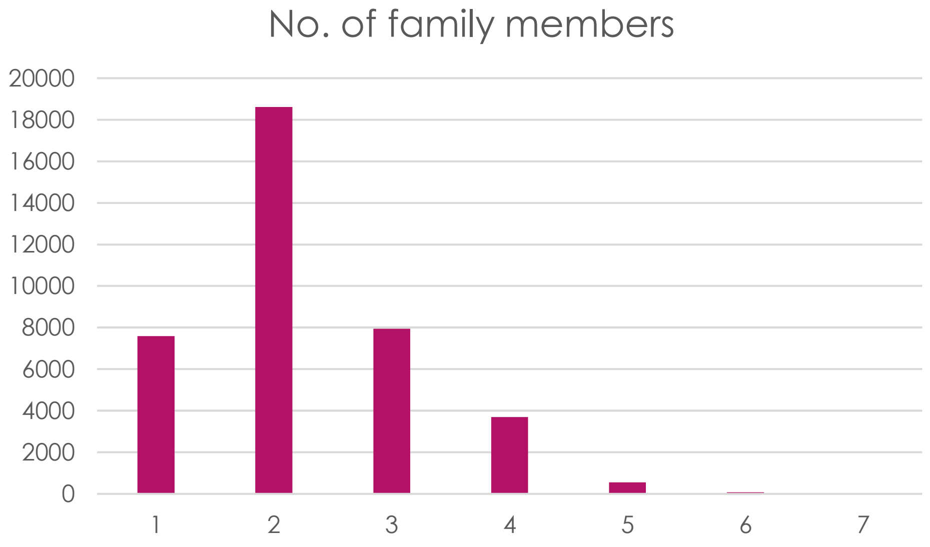
- Region rating 3 have the highest defaulters.



UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS:No. of family members

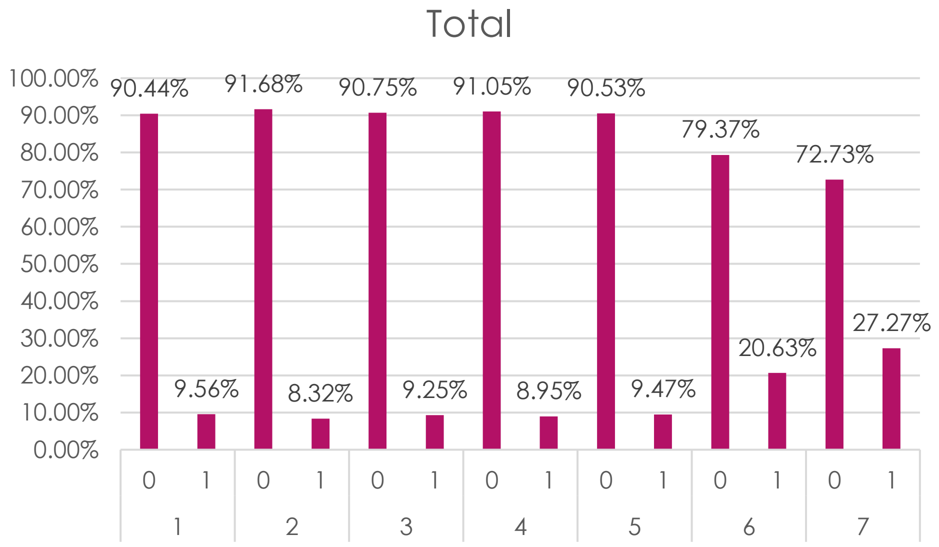
Univariate analysis

- Most of the people taking loans have 2 family members.



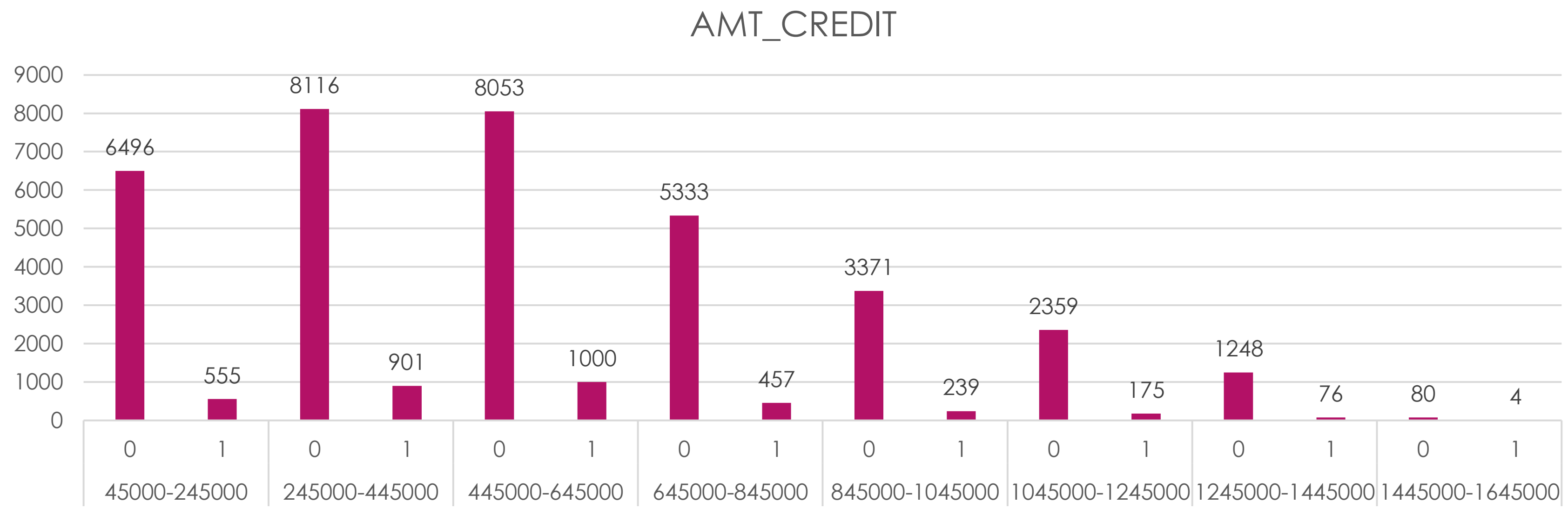
Univariate segmented analysis

- Highest percent of defaulters are from people having more family members.

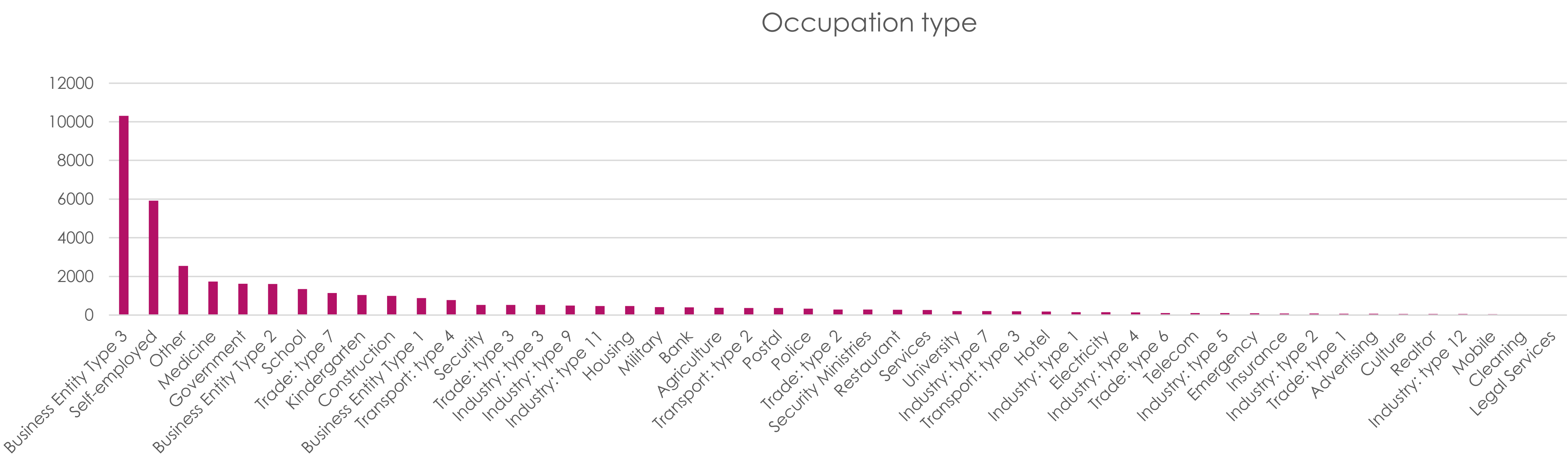


UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS

- The most of the loans are taken between 245000 to 645000.
- Highest defaulters are from 445000 – 645000.



UNIVARIATE ANALYSIS:



Business entity Type 3 takes the highest amount of loans.

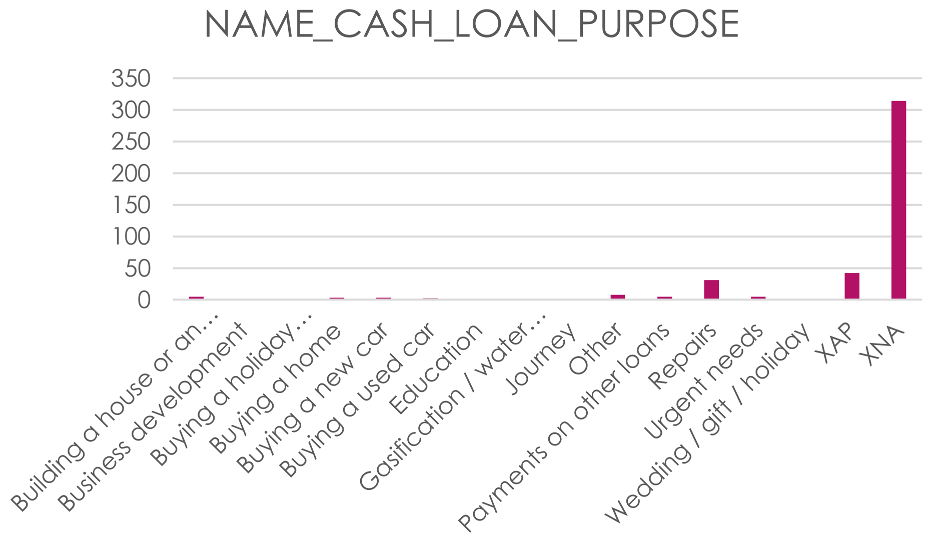
UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS

- Most of the defaulters are from the income range of 25650 – 1025650.

UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS: previous_application

Univariate analysis

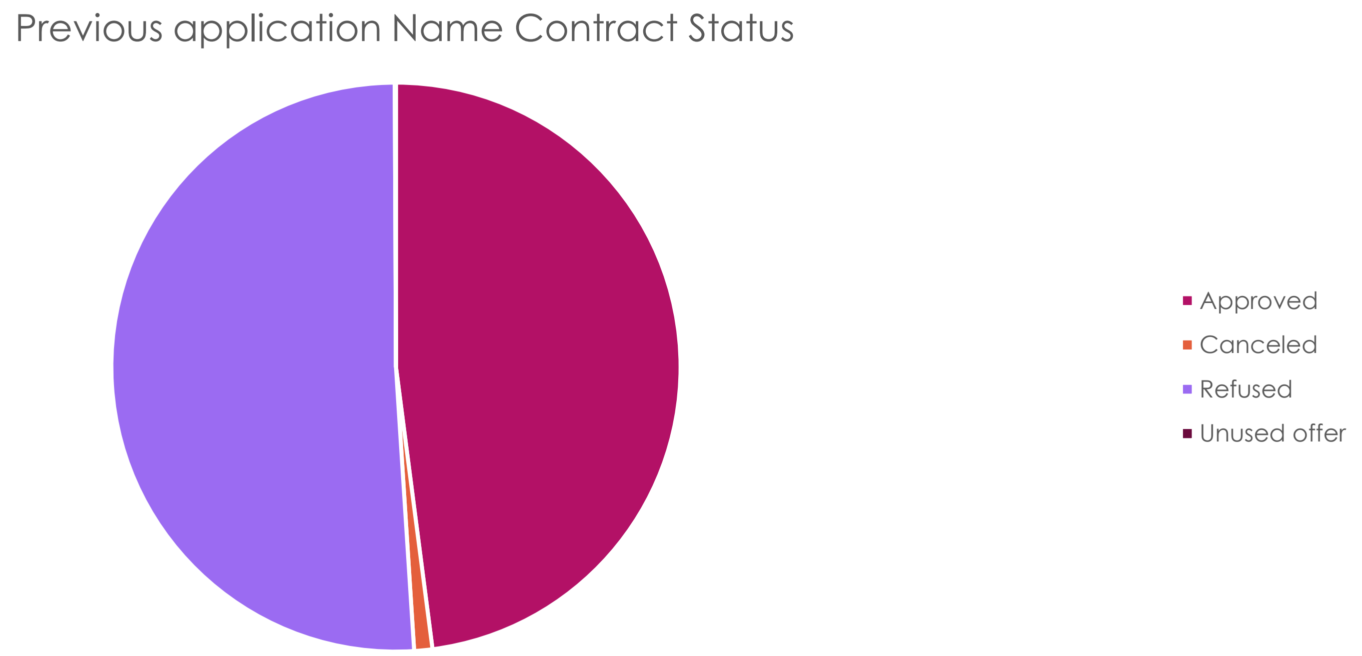
- Most of the previous loan takers are for XNA category.



Univariate segmented analysis

- Most of the loan defaulters are from XNA category.

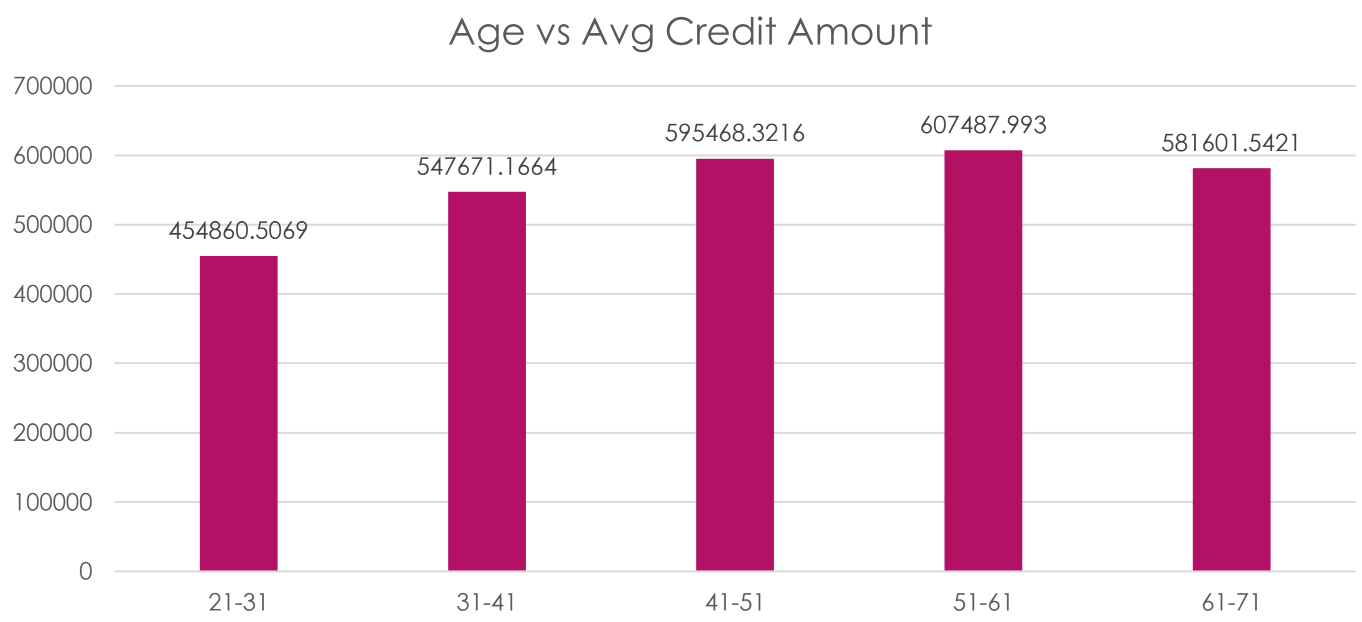
UNIVARIATE ANALYSIS: previous_application



Approximately equal number of previous loans were either approved or refused.

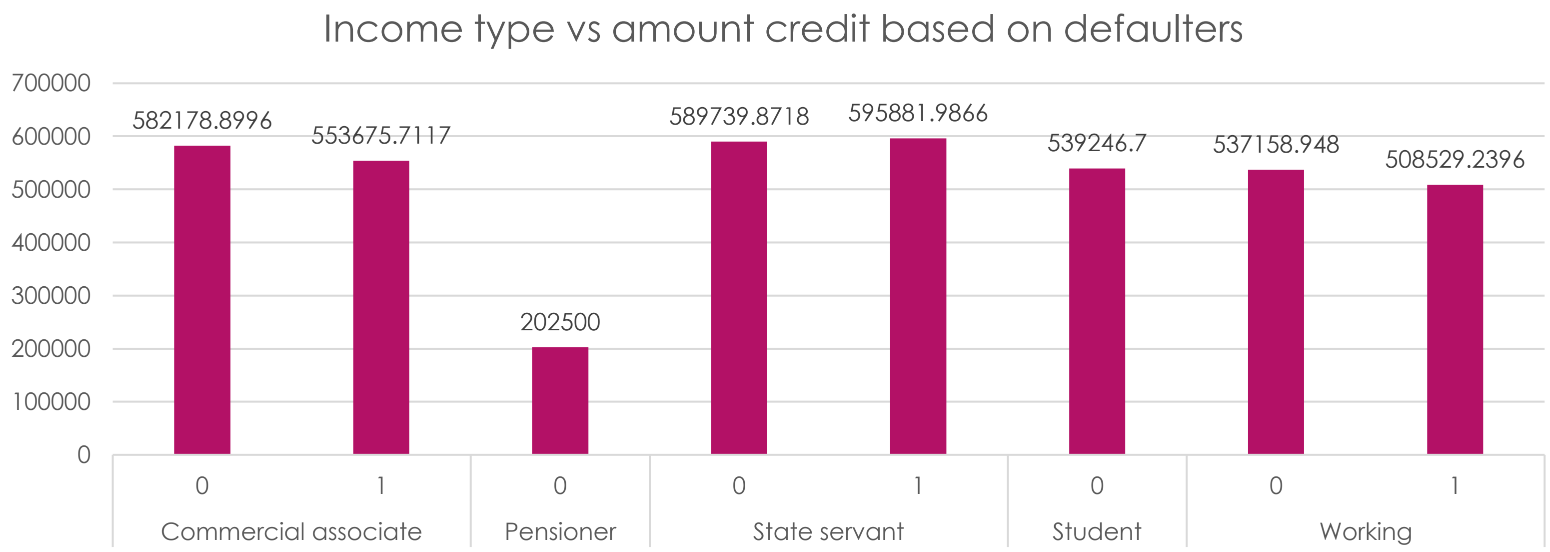
Bivariate/Segmented Bivariate Analysis

- We can see the credit amount is highest for people in age bracket 51-61 years so the insight is that the 51-61 year people default less than the others.



Bivariate/Segmented Bivariate Analysis

- We see pensioners having the least amount credited but also that they don't default.



Correlation among the features: Defaulters

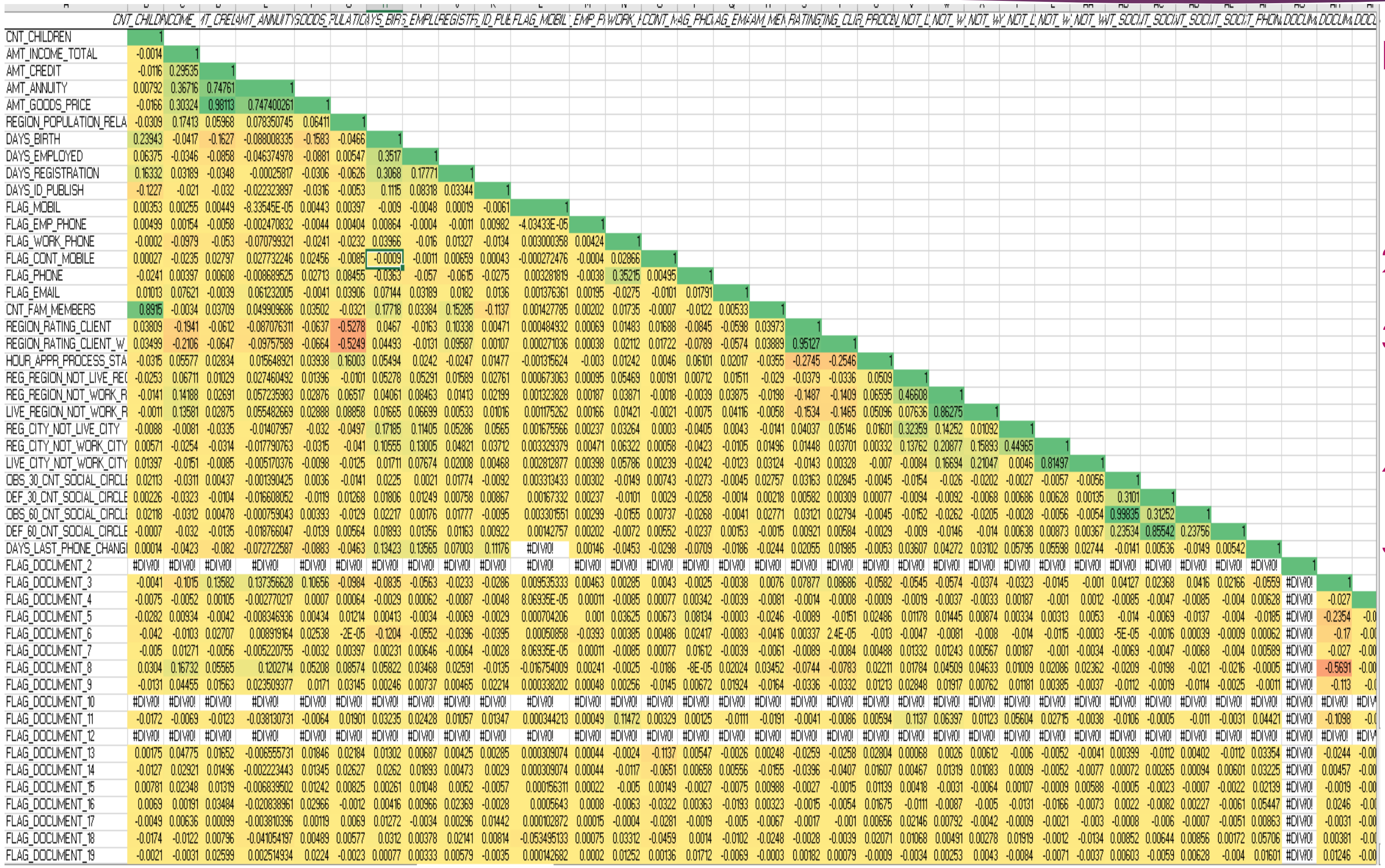
Column1	CNT_CHILDRE	AMT_INCOME_TOTA	AMT_CREDI	AMT_ANNUIT	AMT_GOODS_PRIC	REGION_POPULATION_RELATIV	DAYS_BIRT
CNT_CHILDREN	1						
AMT_INCOME_TOTAL	-0.037075868	1					
AMT_CREDIT	0.018296664	0.292192232	1				
AMT_ANNUITY	0.01388093	0.341453384	0.735256487	1			
AMT_GOODS_PRICE	0.011626404	0.297662806	0.977513624	0.737409349	1		
REGION_POPULATION_RELATIVE	-0.015678121	0.093467914	0.058928665	0.043860922	0.066538303	1	
DAYS_BIRTH	0.166087259	-0.095223734	-0.186047072	-0.07111406	-0.17958462	-0.01651809	1
DAYS_EMPLOYED	0.0165888	-0.025464352	-0.10513514	-0.050884178	-0.115233121	-0.000352962	0.30953554
DAYS_REGISTRATION	0.136017706	-0.00229044	-0.048651406	0.02029698	-0.047981067	-0.046921166	0.244588236
DAYS_ID_PUBLISH	-0.112118914	-0.034664139	-0.051491211	-0.046669977	-0.057941283	-0.006804691	0.122633931
FLAG_MOBIL	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
FLAG_EMP_PHONE	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
FLAG_WORK_PHONE	0.02588478	-0.101133349	-0.057642707	-0.05304025	-0.025691342	-0.027161174	0.060156231
FLAG_CONT_MOBILE	0.00767177	-0.062204615	0.032779119	0.042170336	0.029337791	0.001781923	-0.025894248
FLAG_PHONE	-0.005504041	0.012045091	0.031441771	-0.002549976	0.048539918	0.049959335	-0.030412544
FLAG_EMAIL	0.004980182	0.054238619	-0.007055289	0.083846921	-0.005548285	0.038533337	0.063309792
CNT_FAM_MEMBERS	0.895691634	-0.028726217	0.061753496	0.052342935	0.058969404	-0.020172553	0.104124224
REGION_RATING_CLIENT	0.065580661	-0.151028768	-0.042849823	-0.052080745	-0.050244418	-0.429147258	0.056656144
REGION_RATING_CLIENT_W_CITY	0.063995992	-0.160022222	-0.052137806	-0.070952814	-0.057257597	-0.4323543	0.051376248
HOUR_APPR_PROCESS_START	-0.032561449	0.064606205	0.045755966	0.042429111	0.060700469	0.151543664	0.034687502
REG_REGION_NOT_LIVE_REGION	-0.022582349	0.056980061	0.00856628	0.028687162	0.009031181	-0.006618057	0.021924859
REG_REGION_NOT_WORK_REGION	-0.020433771	0.113982496	0.026913656	0.068951093	0.031652187	0.020285719	0.031885153
LIVE_REGION_NOT_WORK_REGION	-0.013506401	0.112801758	0.035888602	0.072064077	0.040020128	0.058085723	0.018073234
REG_CITY_NOT_LIVE_CITY	-0.014396765	-0.00438027	-0.033058386	0.000205211	-0.032989312	-0.030069437	0.123689628
REG_CITY_NOT_WORK_CITY	-0.000415753	-0.018121666	-0.01335855	0.011897842	-0.017211947	-0.040261818	0.105683563
LIVE_CITY_NOT_WORK_CITY	0.01734573	-0.007206205	0.016944903	0.018939156	0.011093291	-0.022210273	0.030893167
OBS_30_CNT_SOCIAL_CIRCLE	0.020779203	-0.019498609	0.018695746	0.009804005	0.014530655	-0.007863017	-0.001486731
DEF_30_CNT_SOCIAL_CIRCLE	-0.011114197	-0.043559602	-0.037214427	-0.035929123	-0.031649123	0.016575356	-0.006362859
OBS_60_CNT_SOCIAL_CIRCLE	0.017533428	-0.018045785	0.019152095	0.009638339	0.015188141	-0.00621451	-0.002805411
DEF_60_CNT_SOCIAL_CIRCLE	-0.017135641	-0.029938718	-0.042909455	-0.041739693	-0.03415881	0.016475415	-0.013936387
DAYS_LAST_PHONE_CHANGE	0.001782097	-0.096004637	-0.123513993	-0.094114831	-0.127299795	-0.067230297	0.14573565
FLAG_DOCUMENT_2	-0.011660111	-0.000919044	0.046955865	0.018222831	0.05766707	-0.018488398	-0.036410495
FLAG_DOCUMENT_3	-0.009364854	-0.073455765	0.093929892	0.111281542	0.066931859	-0.028812242	-0.040046658
FLAG_DOCUMENT_4	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
FLAG_DOCUMENT_5	-0.027119984	-0.007192003	-0.029603612	-0.023318453	-0.017080056	-0.002863054	0.003250626
FLAG_DOCUMENT_6	-0.01774014	-0.004520921	0.034720314	0.019156506	0.031560365	0.003691754	-0.080529289
FLAG_DOCUMENT_7	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
FLAG_DOCUMENT_8	0.002085501	0.159504527	0.051360744	0.106202667	0.050151881	0.035096628	0.004318011
FLAG_DOCUMENT_9	0.014385246	0.057933305	0.008468483	0.009341416	0.00655654	0.008422027	0.030213197
FLAG_DOCUMENT_10	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
FLAG_DOCUMENT_11	0.005379829	-0.022126113	-0.045752982	-0.053272955	-0.0456109	0.031329308	0.026713436
FLAG_DOCUMENT_12	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
FLAG_DOCUMENT_13	-0.016492308	0.021486096	0.004410645	0.003751927	0.009950104	0.017981294	-0.011826919

Column1	DAYS_EMPLOYE	DAYS_REGISTRATIO	DAYS_ID_PUBLIS	FLAG_MOB	FLAG_EMP_PHON	FLAG_WORK_PHON	FLAG_CONT_MOBIL	FLAG_PHON	FLAG_EM
CNT_CHILDREN									
AMT_INCOME_TOTAL									
AMT_CREDIT									
AMT_ANNUITY									
AMT_GOODS_PRICE									
REGION_POPULATION_RELATIVE									
DAYS_BIRTH									
DAYS_EMPLOYED	1								
DAYS_REGISTRATION	0.154476676	1							
DAYS_ID_PUBLISH	0.102770224	0.044523103	1						
FLAG_MOBIL	#DIV/0!	#DIV/0!	#DIV/0!	1					
FLAG_EMP_PHONE	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	1				
FLAG_WORK_PHONE	-0.013346664	0.023232697	-0.018477429	#DIV/0!	#DIV/0!	1			
FLAG_CONT_MOBILE	-0.00883226	-0.001124023	-0.006169565	#DIV/0!	#DIV/0!	0.017679879	1		
FLAG_PHONE	-0.064110856	-0.058031107	-0.040011889	#DIV/0!	#DIV/0!	0.336640732	0.01579947	1	
FLAG_EMAIL	0.030156605	-0.018074243	0.019793754	#DIV/0!	#DIV/0!	0.009790362	0.007354386	0.031891025	
CNT_FAM_MEMBERS	-0.007270723	0.13490667	-0.11184541	#DIV/0!	#DIV/0!	0.031706513	-0.002166857	0.008154656	-0.00833
REGION_RATING_CLIENT	0.006685917	0.133465594	0.028895115	#DIV/0!	#DIV/0!	0.021984903	0.030137831	-0.053666996	-0.01445
REGION_RATING_CLIENT_W_CITY	0.003856483	0.123998038	0.021256022	#DIV/0!	#DIV/0!	0.028390782	0.029139451	-0.047415299	-0.01175
HOUR_APPR_PROCESS_START	-0.010889432	-0.063362677	-0.001137479	#DIV/0!	#DIV/0!	0.042081951	-0.020212515	0.05273293	-0.00900
REG_REGION_NOT_LIVE_REGION	0.040164089	0.005285146	0.013477124	#DIV/0!	#DIV/0!	0.064351434	0.004299686	0.017707846	0.01730
REG_REGION_NOT_WORK_REGION	0.069680257	8.6325E-05	0.021416908	#DIV/0!	#DIV/0!	0.087051002	0.007234766	0.036432133	0.00522
LIVE_REGION_NOT_WORK_REGION	0.04664693	-0.001623883	0.009731827	#DIV/0!	#DIV/0!	0.060518825	0.006099552	0.037838971	-0.00624
REG_CITY_NOT_LIVE_CITY	0.097544216	0.034377759	0.046127266	#DIV/0!	#DIV/0!	-0.006881511	0.011447208	-0.03655406	-0.0018
REG_CITY_NOT_WORK_CITY	0.134009059	0.065976643	0.026541346	#DIV/0!	#DIV/0!	0.053834921	0.021749861	0.000505865	-0.01268
LIVE_CITY_NOT_WORK_CITY	0.071911384	0.038869291	-0.012740353	#DIV/0!	#DIV/0!	0.066855588	0.017411457	0.034916877	-0.01850
OBS_30_CNT_SOCIAL_CIRCLE	-0.04170605	0.008798129	-0.023732187	#DIV/0!	#DIV/0!	-0.073398801	0.010756952	-0.064638221	-0.01893
DEF_30_CNT_SOCIAL_CIRCLE	0.002363023	0.001131154	-0.025507001	#DIV/0!	#DIV/0!	-0.028875709	-0.025656514	-0.025786976	-0.01847
OBS_60_CNT_SOCIAL_CIRCLE	-0.042357902	0.008463905	-0.022288819	#DIV/0!	#DIV/0!	-0.073220435	0.010615919	-0.065487326	-0.01747
DEF_60_CNT_SOCIAL_CIRCLE	-6.62931E-05	-0.007801486	-0.030941367	#DIV/0!	#DIV/0!	-0.028428071	-0.0333164	-0.022989055	-0.01926
DAYS_LAST_PHONE_CHANGE	0.140168213	0.087776161	0.135165988	#DIV/0!	#DIV/0!	-0.008496777	-0.017548191	-0.06617992	0.03267
FLAG_DOCUMENT_2	0.00632065	-0.022252701	-0.022878733	#DIV/0!	#DIV/0!	-0.010204485	0.000508679	-0.00911915	-0.0042
FLAG_DOCUMENT_3	-0.048849062	-0.01606743	-0.04439463	#DIV/0!	#DIV/0!	-0.002128998	-0.01339082	-0.03184293	0.02115
FLAG_DOCUMENT_4	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0
FLAG_DOCUMENT_5	-0.05853412	-0.024587368	-0.014760612	#DIV/0!	#DIV/0!	0.062005701	0.003695909	0.072289039	-0.03084
FLAG_DOCUMENT_6	0.016660395	-0.002630308	-0.019787088	#DIV/0!	#DIV/0!	0.027007702	0.002393304	0.036616099	0.01142
FLAG_DOCUMENT_7	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0
FLAG_DOCUMENT_8	0.030596457	0.016092522	-0.005367299	#DIV/0!	#DIV/0!	-0.014667961	0.008656786	0.018610531	0.00730
FLAG_DOCUMENT_9	0.013851524	0.01214234	0.014115252	#DIV/0!	#DIV/0!	-0.017381288	0.001906944	-0.012059929	0.0037
FLAG_DOCUMENT_10	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0
FLAG_DOCUMENT_11	0.03287023	0.020857316	0.041869441	#DIV/0!	#DIV/0!	0.081462171	0.001440241	0.003424847	-0.01201
FLAG_DOCUMENT_12	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0
FLAG_DOCUMENT_13	0.013689839	0.009088912	0.011840019	#DIV/0!	#DIV/0!	-0.014433441	0.000719486	-0.012898319	-0.0060

- Top 5 correlations are from:

1. OBS_30_CNT_SOCIAL_CIRCLE - OBS_60_CNT_SOCIAL_CIRCLE
2. AMT_CREDIT - AMT_GOODS_PRICE
3. REGION_RATING_CLIENT - REGION_RATING_CLIENT_W_CITY
4. CNT_CHILDREN - CNT_FAM_MEMBERS
5. REG_REGION_NOT_WORK_REGION - LIVE_REGION_NOT_WORK_REGION

Correlation among the features: Repayers



- Top 5 correlations are from:

1. OBS_30_CNT_SOCIAL_CIRCLE - OBS_60_CNT_SOCIAL_CIRCLE
2. AMT_CREDIT - AMT_GOODS_PRICE
3. REGION_RATING_CLIENT - REGION_RATING_CLIENT_W_CITY
4. CNT_CHILDREN - CNT_FAM_MEMBERS
5. REG_REGION_NOT_WORK_REGION - LIVE_REGION_NOT_WORK_REGION

Insights and hypothesis:

- Elderly people with good experience years are less likely to default on a loan, hence can be considered good customers for loans.
- Most people take cash loans and hence there is a high defaulter rate for cash loans.
- Females take more loans than males but the default rate for males is higher, hence the banks can consider giving out loans to females more than men.
- People with less educational qualifications tend to default more than the ones having academic degrees.
- Most of the single or civil marriage people default more than any other family status.
- Taking a high amount as the loan has a high default rate, this could be considered while giving out loans.
- People with region rating 3 default more than any other region rating.
- People with less income default more.
- Pensioners have less amount of credit but they don't default much, hence can be considered for providing loans in the future.

Results:

- This project was of a great help to work with huge dataset with data cleaning and handling outliers using the domain knowledge.
- This project helped me understand how to merge two excel sheets and use add-ins like data analyse for correlation.

Drive Links:

- Application_data: https://docs.google.com/spreadsheets/d/1f4LFl0Z9NhbGHCyuxTpe6rzyQKI_BKbr/edit?usp=sharing&ouid=115109770037321084146&rtpof=true&sd=true
- Previous_application: https://docs.google.com/spreadsheets/d/1PdOBodjo_yOuVJycu9P0xddODKUKUSB8/edit?usp=sharing&ouid=115109770037321084146&rtpof=true&sd=true
- Merged sheets: https://docs.google.com/spreadsheets/d/1aAJemTTQC8RxI_jPQmlz1PSiB3x9Jom-/edit?usp=sharing&ouid=115109770037321084146&rtpof=true&sd=true
- Correlation: https://docs.google.com/spreadsheets/d/1SboxQXWvU-F6tPMJZDoCIlje-Hpk6-Ck/edit?usp=sharing&ouid=115109770037321084146&rtpof=true&sd=true
- Analysis: https://docs.google.com/spreadsheets/d/1al0q0--yOniq_62NPtfyDc9WAu2xjJsR/edit?usp=sharing&ouid=115109770037321084146&rtpof=true&sd=true

ThankYou