

Bank Loan Case Study

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. So we are doing EDA for this dataset to check two types of risks after receiving the application i.e

1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
2. If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Approach : To begin with I imported the given datasets in Microsoft Excel. I saw the tasks to done for above datasets, then I saw tables in the dataset and initially I cleaned the dataset based on the blank cell percentage in columns and columns with more than 40% was dropped and unnecessary columns were also removed. Based on the task I did mind mapping like which tables , columns and excel functions to use for required task and later the same was used in Microsoft Excel.

Tech Stack Used : Microsoft Excel

1. Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)

Here to begin with I calculated all the columns blank cells percentage and based on the percentage I have dropped or replaced columns with mode imputation, i.e

- More than 40%, I have dropped all the columns.
- 10%-40%, I have replace the blank cells with mode imputation i.e highest frequency value.
- Less than 10%, I have trimmed the whole row since we have large dataset.

<i>column_name</i>	<i>blank_cell_in %</i>
SK_ID_CURR	0.00
TARGET	0.00
NAME_CONTRACT_TYPE	0.00
CODE_GENDER	0.00
FLAG_OWN_CAR	0.00
FLAG_OWN_REALTY	0.00
CNT_CHILDREN	0.00

<i>column_name</i>	<i>blank_cell_in %</i>
AMT_INCOME_TOTAL	0.00
AMT_CREDIT	0.00
NAME_INCOME_TYPE	0.00
NAME_EDUCATION_TYPE	0.00
NAME_FAMILY_STATUS	0.00
NAME_HOUSING_TYPE	0.00
REGION_POPULATION_RELATIVE	0.00
DAYS_BIRTH	0.00
DAYS_EMPLOYED	0.00
DAYS_REGISTRATION	0.00
DAYS_ID_PUBLISH	0.00
FLAG_MOBIL	0.00
FLAG_EMP_PHONE	0.00
FLAG_WORK_PHONE	0.00
FLAG_CONT_MOBILE	0.00
FLAG_PHONE	0.00
FLAG_EMAIL	0.00
REGION_RATING_CLIENT	0.00
REGION_RATING_CLIENT_W_CITY	0.00
WEEKDAY_APPR_PROCESS_START	0.00
HOUR_APPR_PROCESS_START	0.00
REG_REGION_NOT_LIVE_REGION	0.00
REG_REGION_NOT_WORK_REGION	0.00
LIVE_REGION_NOT_WORK_REGION	0.00
REG_CITY_NOT_LIVE_CITY	0.00
REG_CITY_NOT_WORK_CITY	0.00
LIVE_CITY_NOT_WORK_CITY	0.00
ORGANIZATION_TYPE	0.00
FLAG_DOCUMENT_2	0.00
FLAG_DOCUMENT_3	0.00
FLAG_DOCUMENT_4	0.00
FLAG_DOCUMENT_5	0.00
FLAG_DOCUMENT_6	0.00
FLAG_DOCUMENT_7	0.00
FLAG_DOCUMENT_8	0.00
FLAG_DOCUMENT_9	0.00
FLAG_DOCUMENT_10	0.00
FLAG_DOCUMENT_11	0.00
FLAG_DOCUMENT_12	0.00
FLAG_DOCUMENT_13	0.00
FLAG_DOCUMENT_14	0.00

<i>column_name</i>	<i>blank_cell_in %</i>
FLAG_DOCUMENT_15	0.00
FLAG_DOCUMENT_16	0.00
FLAG_DOCUMENT_17	0.00
FLAG_DOCUMENT_18	0.00
FLAG_DOCUMENT_19	0.00
FLAG_DOCUMENT_20	0.00
FLAG_DOCUMENT_21	0.00
DAYS_LAST_PHONE_CHANGE	0.00
CNT_FAM_MEMBERS	0.00
AMT_ANNUITY	0.00
AMT_GOODS_PRICE	0.09
EXT_SOURCE_2	0.21
OBS_30_CNT_SOCIAL_CIRCLE	0.33
DEF_30_CNT_SOCIAL_CIRCLE	0.33
OBS_60_CNT_SOCIAL_CIRCLE	0.33
DEF_60_CNT_SOCIAL_CIRCLE	0.33
NAME_TYPE_SUITE	0.42
AMT_REQ_CREDIT_BUREAU_HOUR	13.50
AMT_REQ_CREDIT_BUREAU_DAY	13.50
AMT_REQ_CREDIT_BUREAU_WEEK	13.50
AMT_REQ_CREDIT_BUREAU_MON	13.50
AMT_REQ_CREDIT_BUREAU_QRT	13.50
AMT_REQ_CREDIT_BUREAU_YEAR	13.50
EXT_SOURCE_3	19.83
OCCUPATION_TYPE	31.35
EMERGENCYSTATE_MODE	47.40
TOTALAREA_MODE	48.27
YEARS_BEGINEXPLUATATION_AVG	48.78
YEARS_BEGINEXPLUATATION_MODE	48.78
YEARS_BEGINEXPLUATATION_MEDI	48.78
FLOORSMAX_AVG	49.76
FLOORSMAX_MODE	49.76
FLOORSMAX_MEDI	49.76
HOUSETYPE_MODE	50.18
LIVINGAREA_AVG	50.19
LIVINGAREA_MODE	50.19
LIVINGAREA_MEDI	50.19
ENTRANCES_AVG	50.35
ENTRANCES_MODE	50.35
ENTRANCES_MEDI	50.35
APARTMENTS_AVG	50.75

<i>column_name</i>	<i>blank_cell_in %</i>
APARTMENTS_MODE	50.75
APARTMENTS_MEDI	50.75
WALLSMATERIAL_MODE	50.84
ELEVATORS_AVG	53.30
ELEVATORS_MODE	53.30
ELEVATORS_MEDI	53.30
NONLIVINGAREA_AVG	55.18
NONLIVINGAREA_MODE	55.18
NONLIVINGAREA_MEDI	55.18
EXT_SOURCE_1	56.38
BASEMENTAREA_AVG	58.52
BASEMENTAREA_MODE	58.52
BASEMENTAREA_MEDI	58.52
LANDAREA_AVG	59.38
LANDAREA_MODE	59.38
LANDAREA_MEDI	59.38
OWN_CAR_AGE	65.99
YEARS_BUILD_AVG	66.50
YEARS_BUILD_MODE	66.50
YEARS_BUILD_MEDI	66.50
FLOORSMIN_AVG	67.85
FLOORSMIN_MODE	67.85
FLOORSMIN_MEDI	67.85
LIVINGAPARTMENTS_AVG	68.35
LIVINGAPARTMENTS_MODE	68.35
LIVINGAPARTMENTS_MEDI	68.35
FONDKAPREMONT_MODE	68.39
NONLIVINGAPARTMENTS_AVG	69.43
NONLIVINGAPARTMENTS_MODE	69.43
NONLIVINGAPARTMENTS_MEDI	69.43
COMMONAREA_AVG	69.87
COMMONAREA_MODE	69.87
COMMONAREA_MEDI	69.87

Here we are doing mode imputation i.e all blank cells in respective columns are replaced by highest repeated cells of category .

Values	Count of AMT_REQ_CREDIT_BUREAU_YEAR	percentage
0	71801	26.99%
1	63405	23.84%
2	50192	18.87%
3	33628	12.64%
4	20714	7.79%
5	12052	4.53%
6	6967	2.62%
7	3869	1.45%
8	2127	0.80%
9	1096	0.41%
10	22	0.01%
11	31	0.01%
12	30	0.01%
13	19	0.01%
14	10	0.00%
15	6	0.00%
16	3	0.00%
17	7	0.00%
18	4	0.00%
19	4	0.00%
20	1	0.00%
21	1	0.00%
22	1	0.00%
23	1	0.00%
25	1	0.00%
(blank)		0.00%
Grand Total	265992	100.00%

Values	Count of AMT_REQ_CREDIT_BUREAU_DAY	percentage
0	264503	99.44%
1	1292	0.49%
2	106	0.04%
3	45	0.02%
4	26	0.01%
5	9	0.00%
6	8	0.00%
8	1	0.00%
9	2	0.00%
(blank)		0.00%
Grand Total	265992	100.00%

Values	Count of OCCUPATION_TYPE	percentage
Accountants	9813	4.65%
Cleaning staff	4653	2.20%
Cooking staff	5946	2.82%
Core staff	27570	13.06%
Drivers	18603	8.81%
High skill tech staff	11380	5.39%
HR staff	563	0.27%
IT staff	526	0.25%
Laborers	55186	26.14%
Low-skill Laborers	2093	0.99%
Managers	21371	10.12%
Medicine staff	8537	4.04%
Private service staff	2652	1.26%
Realty agents	751	0.36%
Sales staff	32102	15.21%
Secretaries	1305	0.62%
Security staff	6721	3.18%
Waiters/barmen staff	1348	0.64%
(blank)		0.00%
Grand Total	211120	100.00%

Values	Count of AMT_REQ_CREDIT_BUREAU_QRT	percentage
0	215417	80.99%
1	33862	12.73%
2	14412	5.42%
3	1717	0.65%
4	476	0.18%
5	64	0.02%
6	28	0.01%
7	7	0.00%
8	7	0.00%
19	1	0.00%
261	1	0.00%
(blank)		0.00%
Grand Total	265992	100.00%

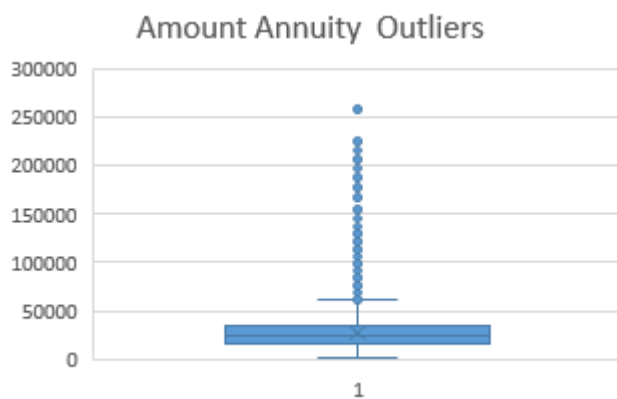
Values	Count of AMT_REQ_CREDIT_BUREAU_WEEK	percentage
0	257456	96.79%
1	8208	3.09%
2	199	0.07%
3	58	0.02%
4	34	0.01%
5	10	0.00%
6	20	0.01%
7	2	0.00%
8	5	0.00%
(blank)		0.00%
Grand Total	265992	100.00%

Values	Count of AMT_REQ_CREDIT_BUREAU_MON	percentage
0	222233	83.55%
1	33147	12.46%
2	5386	2.02%
3	1991	0.75%
4	1076	0.40%
5	602	0.23%
6	343	0.13%
7	298	0.11%
8	185	0.07%
9	206	0.08%
10	132	0.05%
11	119	0.04%
12	77	0.03%
13	72	0.03%
14	40	0.02%
15	35	0.01%
16	23	0.01%
17	14	0.01%
18	6	0.00%
19	3	0.00%
22	1	0.00%
23	1	0.00%
24	1	0.00%
27	1	0.00%
(blank)		0.00%
Grand Total	265992	100.00%

Values	percentage	Count of EXT_SOURCE_3
0.746300213	12.11%	1460
0.7136314	10.91%	1315
0.694092643	10.59%	1276
0.689479143	9.44%	1138
0.670651753	9.88%	1191
0.652896552	9.58%	1154
0.621226338	9.20%	1109
0.595456203	9.43%	1136
0.581483706	9.47%	1141
0.554946769	9.39%	1132
Grand Total	100.00%	12052

Median Imputation for annuity feature

MIN	1615.5
1st Quartile	16524
Median	24903
3rd Quartile	34596
Max	258025.5



Median Imputation for amount goods price t feature

MIN	40500
1st Quartile	238500
Median	450000
3rd Quartile	679500
Max	4050000



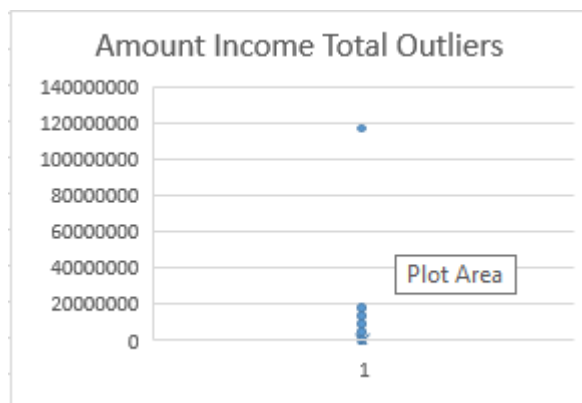
Since there are lot of outliers we are replacing all blanks cells in amount annuity column and amount goods price column with median which is 24903 and 45000 respectively.

The highlighted cells in first table is dropped since they have blank cells percentage more than 40% and all highlighted cells in remaining tables were replaced as mode imputation to blank cells in 10%-40% category in their respective columns and later all unnecessary columns which are not usefull for our analysis are dropped.

2. Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier.

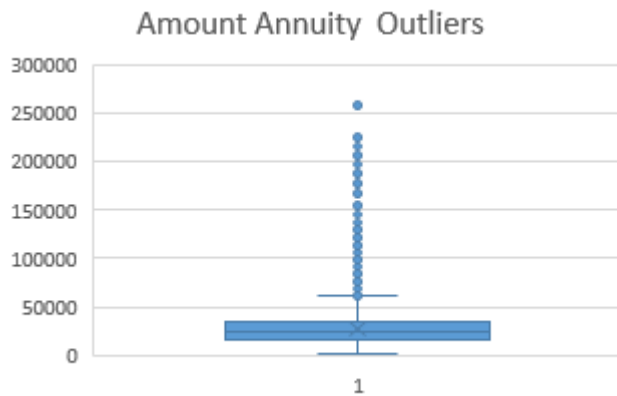
1. Outliers in amount income total features

MIN	25650
1st Quartile	112500
Median	147150
3rd Quartile	202500
Max	117000000



2. Outliers in amount annuity feature

MIN	1615.5
1st Quartile	16524
Median	24903
3rd Quartile	34596
Max	258025.5



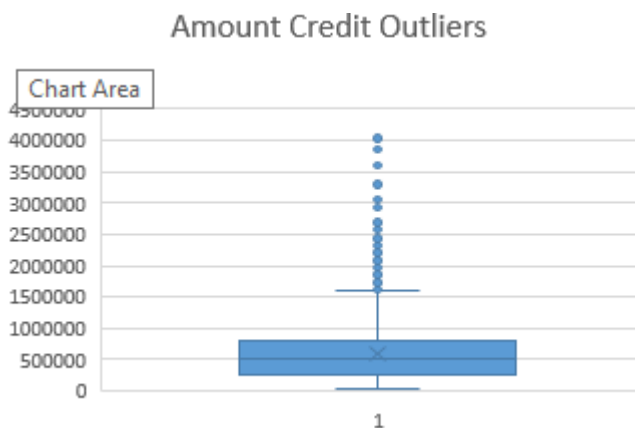
3. Outliers in amount goods price t feature

MIN	40500
1st Quartile	238500
Median	450000
3rd Quartile	679500
Max	4050000



4. Outliers in amount credit feature

MIN	45000
1st Quartile	270000
Median	513531
3rd Quartile	808650
Max	4050000

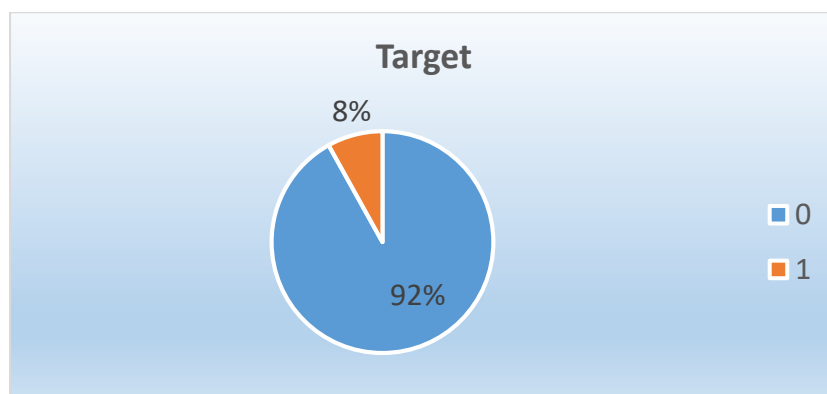


All the four columns i.e AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE have outliers has shown in box and whisker chart and its min, 1st quartile, median, 3rd quartile, max values are represented in above tables respectively. But here we are not replacing outliers with anything since it is related income which varies from person to person.

3. Univariate Analysis

Target Analysis with pie chart

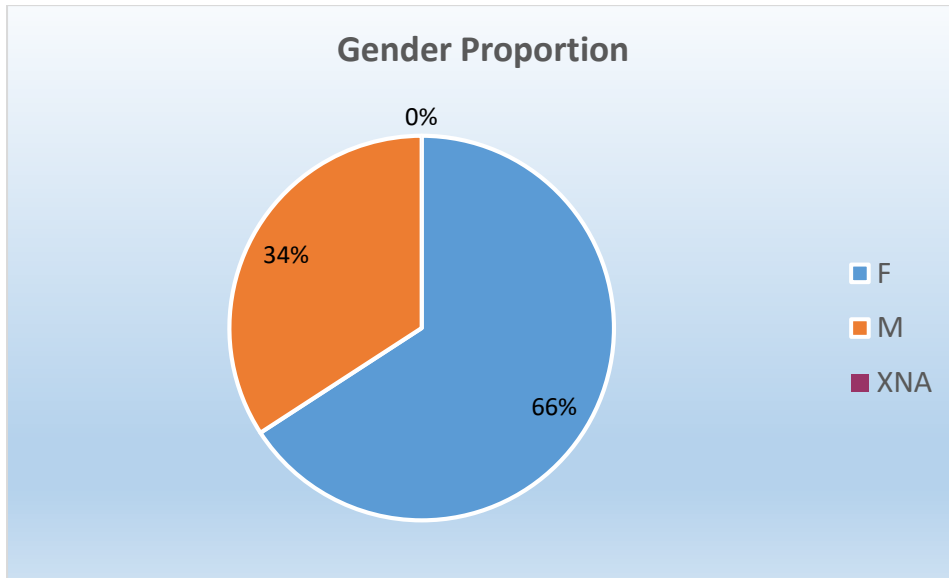
Row Labels	Count of TARGET
0	282686
1	24824
Grand Total	307510



Here 92% people does not have any difficulty in repaying loans (0) and 8% people finds difficulty in repaying loans (1) , they can become defaulters.

Gender proportion

Row Labels	Count of CODE_GENDER
F	202448
M	105058
XNA	4
Grand Total	307510

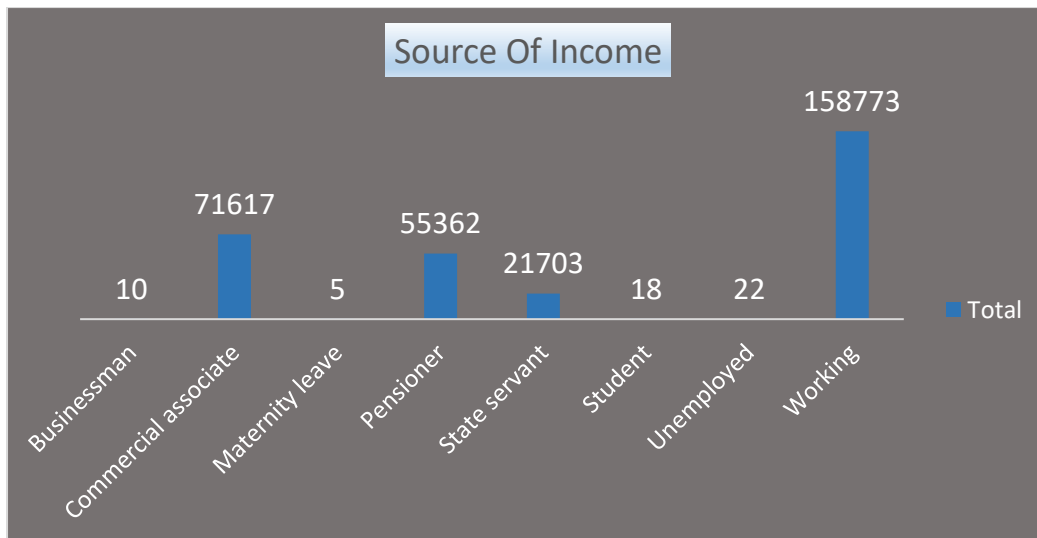


Here 66% are females and 34% are males.

XNA count is only 4 which can be neglected.

Source of Income Type

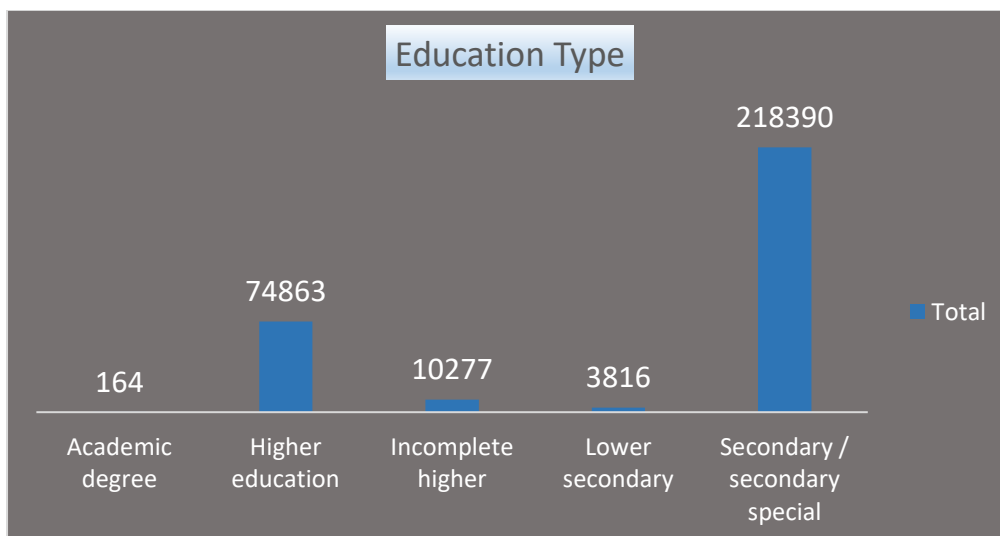
Row Labels	Count of NAME_INCOME_TYPE
Businessman	10
Commercial associate	71617
Maternity leave	5
Pensioner	55362
State servant	21703
Student	18
Unemployed	22
Working	158773
Grand Total	307510



Here **working class** people has high percentage of applying loans.

Education qualification of loan appliers

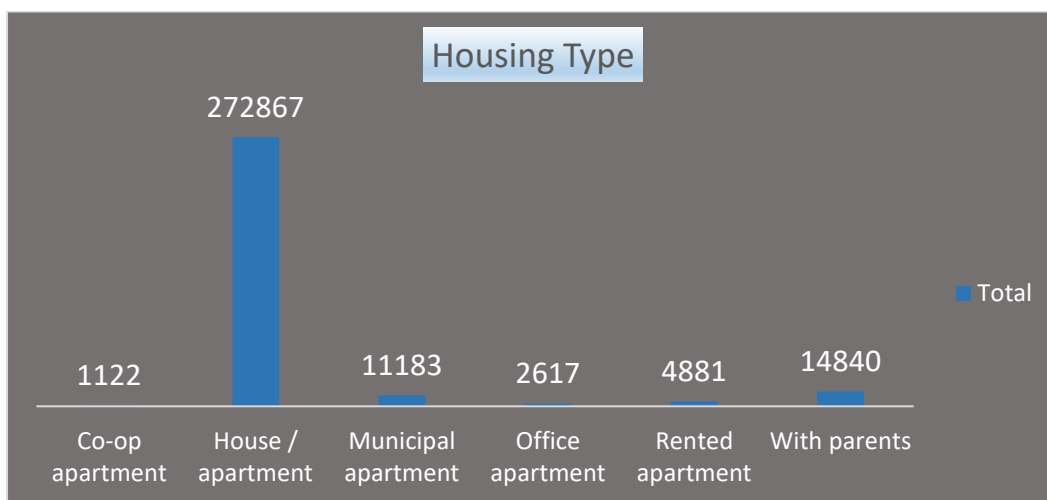
Row Labels	Count of NAME_EDUCATION_TYPE
Academic degree	164
Higher education	74863
Incomplete higher	10277
Lower secondary	3816
Secondary / secondary special	218390
Grand Total	307510



Here **Secondary/secondary special** education high percentage of applying loans.

Housing Type Analysis wrt loan application

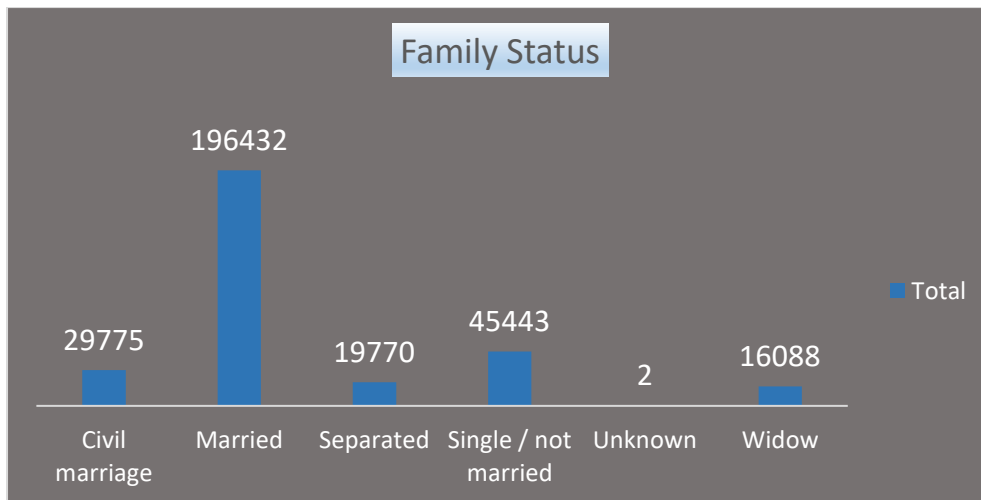
Row Labels	Count of NAME_HOUSING_TYPE
Co-op apartment	1122
House / apartment	272867
Municipal apartment	11183
Office apartment	2617
Rented apartment	4881
With parents	14840
Grand Total	307510



Here people living in **house/apartment** housing type have high percentage of applying loans.

Family Status wrt loan application

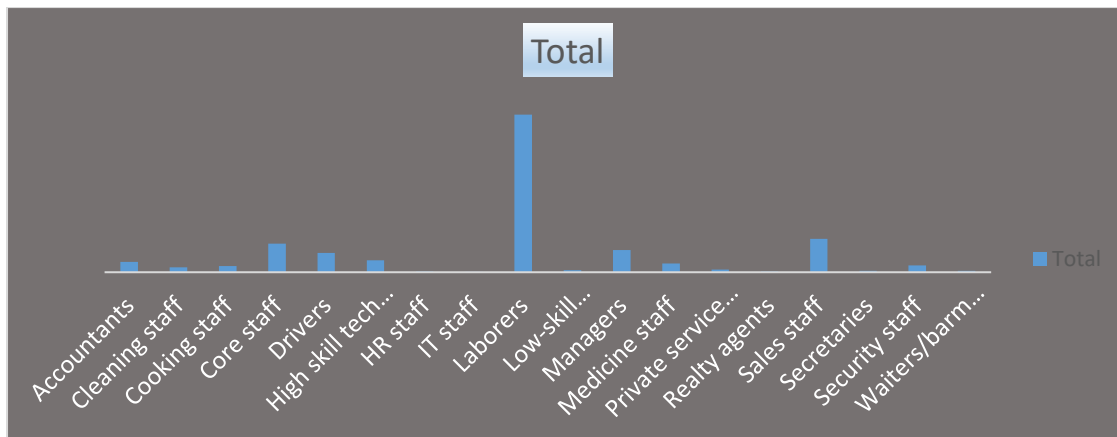
Row Labels	Count of NAME_FAMILY_STATUS
Civil marriage	29775
Married	196432
Separated	19770
Single / not married	45443
Unknown	2
Widow	16088
Grand Total	307510



People who have **married** has high percentage in applying loans.

Occupation Type wrt loan application

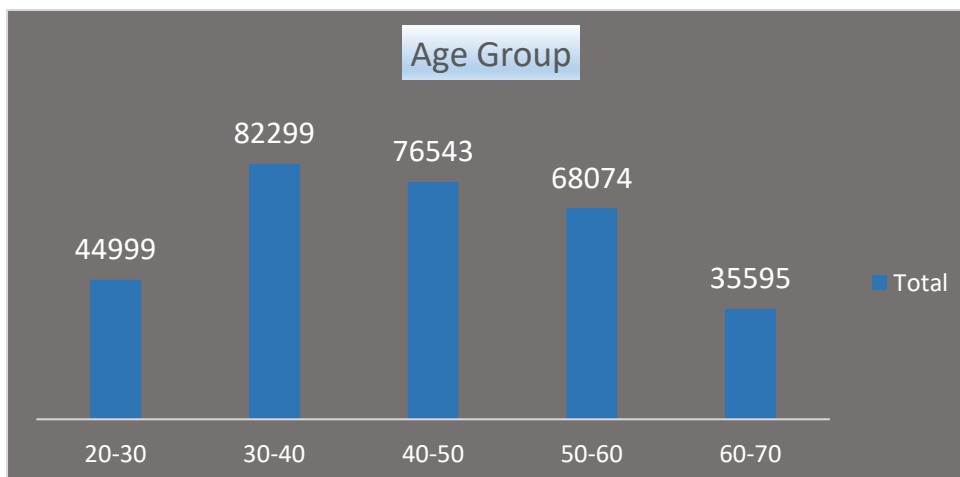
Row Labels	Count of OCCUPATION_TYPE
Accountants	9813
Cleaning staff	4653
Cooking staff	5946
Core staff	27570
Drivers	18603
High skill tech staff	11380
HR staff	563
IT staff	526
Laborers	151576
Low-skill Laborers	2093
Managers	21371
Medicine staff	8537
Private service staff	2652
Realty agents	751
Sales staff	32102
Secretaries	1305
Security staff	6721
Waiters/barmen staff	1348
Grand Total	307510



Here clearly **labourers** occupation type dominated in applying loans.

Age Group wrt loan application

Row Labels	Count of Age_in_years
20-30	44999
30-40	82299
40-50	76543
50-60	68074
60-70	35595
Grand Total	307510

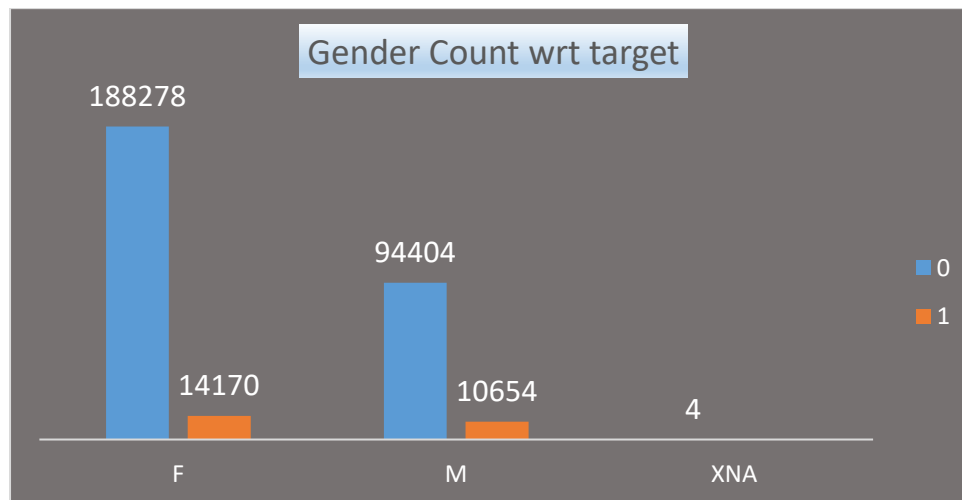


Clearly with age group **30-40** has high percentage in applying loans.

Univariate Segmented Analysis

Gender proportion wrt target

Count of CODE_GENDER	Column Labels		Grand Total
	0	1	
Row Labels			
F	188278	14170	202448
M	94404	10654	105058
XNA	4		4
Grand Total	282686	24824	307510



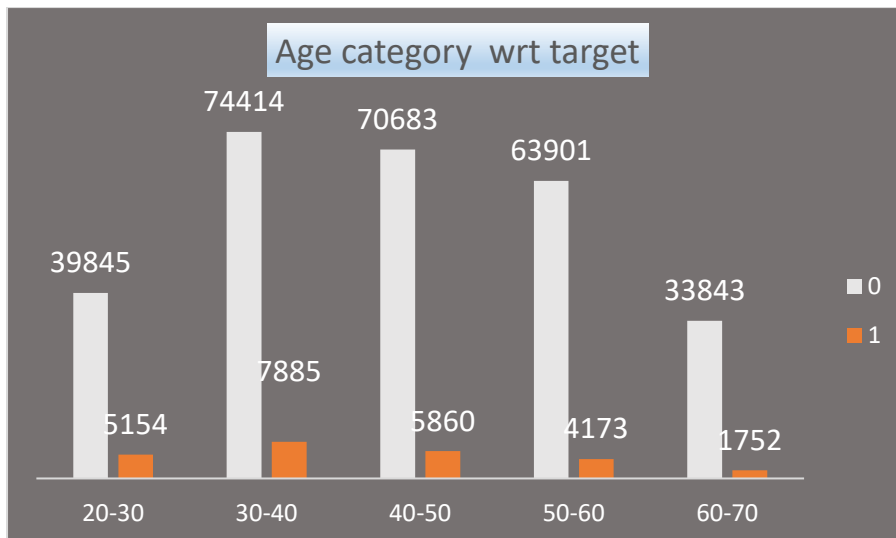
Here male become more defaulters than females comparing to their proportion in difficulty in repaying loans i.e.

Female 66% 14170 defaulters

Male 34% 10654 defaulters

Age Range wrt Target

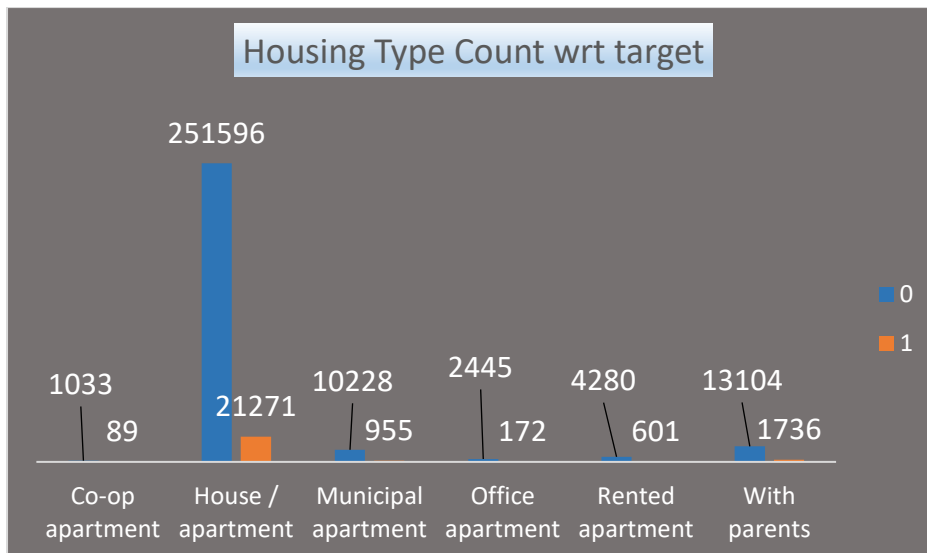
Count of Age_in_years	Column Labels		Grand Total
	0	1	
Row Labels			
20-30	39845	5154	44999
30-40	74414	7885	82299
40-50	70683	5860	76543
50-60	63901	4173	68074
60-70	33843	1752	35595
Grand Total	282686	24824	307510



Here 30-40 age group people have highest proportion in repaying loans and people with age group 20-30 have high chances becoming defaulters since 11.5% of total count has problem in paying loans.

Housing Type wrt target

Count of NAME_HOUSING_TYPE			
Column Labels			
Row Labels	0	1	Grand Total
Co-op apartment	1033	89	1122
House / apartment	251596	21271	272867
Municipal apartment	10228	955	11183
Office apartment	2445	172	2617
Rented apartment	4280	601	4881
With parents	13104	1736	14840
Grand Total	282686	24824	307510

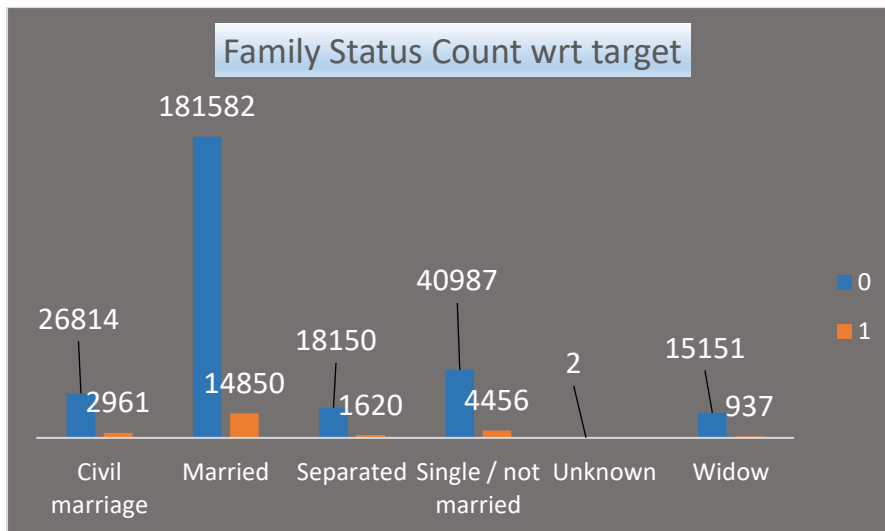


Here people living in house/apartment have highest proportion in repaying loans and people living in rented apartment have high chances becoming defaulters since 12.3% of total count has problem in paying loans.

Family Status wrt target

Count of NAME_FAMILY_STATUS	Column Labels		
Row Labels	0	1	Grand Total
Civil marriage	26814	2961	29775
Married	181582	14850	196432
Separated	18150	1620	19770
Single / not married	40987	4456	45443
Unknown	2		2
Widow	15151	937	16088
Grand Total	282686	24824	307510

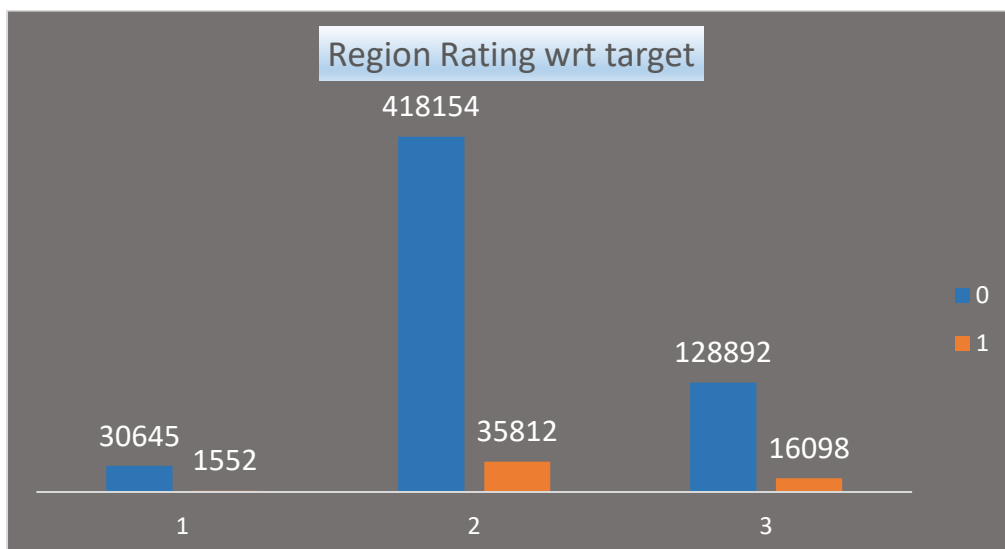
.



Here married people have highest proportion in repaying loans and single/not married have high chances becoming defaulters since 9.8% of total count has problem in paying loans.

Region Rating wrt target

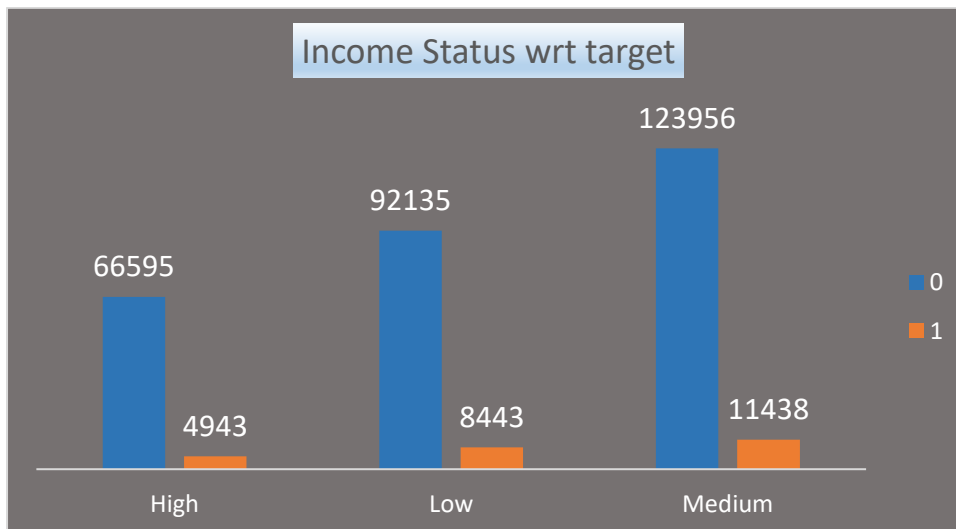
Sum of REGION_RATING_CLIENT		Column Labels	
Row Labels		0	1
		Grand Total	
1		30645	1552
2		418154	35812
3		128892	16098
Grand Total		577691	53462
		631153	



Here tier 2 region people have highest proportion in repaying loans and tier 3 region people have high chances becoming defaulters since 11.1% of total count has problem in paying loans.

Amount Income Range wrt tartget

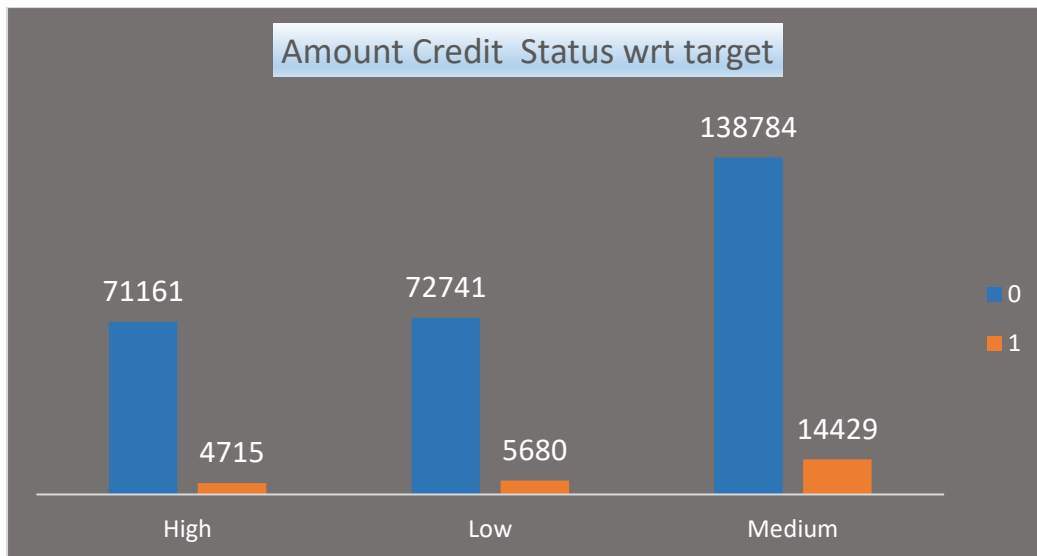
Count of Amt_income_status	Column Labels		
Row Labels	0	1	Grand Total
High	66595	4943	71538
Low	92135	8443	100578
Medium	123956	11438	135394
Grand Total	282686	24824	307510



Here medium income range people have highest proportion in repaying loans and medium income range people have high chances becoming defaulters since 8.4% of total count has problem in paying loans.

Amount Credit Range wrt target

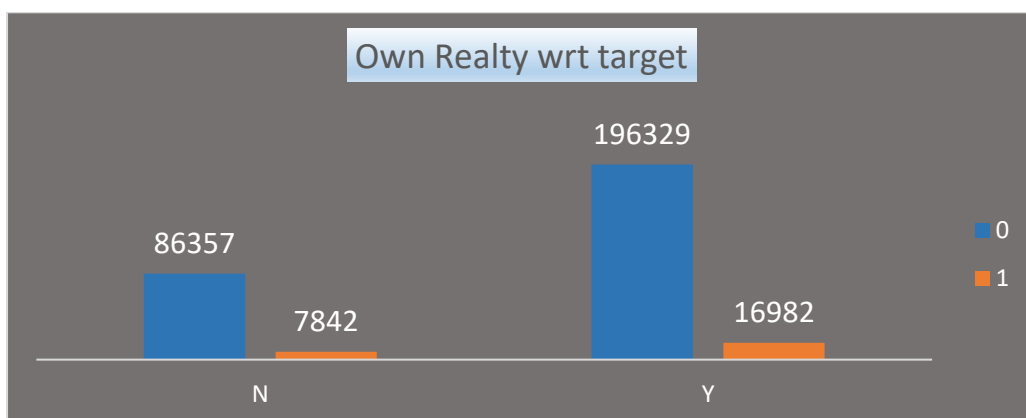
Count of Amt_credit_status	Column Labels		
Row Labels	0	1	Grand Total
High	71161	4715	75876
Low	72741	5680	78421
Medium	138784	14429	153213
Grand Total	282686	24824	307510



Here medium credit range people have highest proportion in repaying loans and medium credit range people have high chances becoming defaulters since 9.4% of total count has problem in paying loans.

People own realty wrt target

Count of FLAG_OWN_REALTY		Column Labels		Grand Total
Row Labels		0	1	
N		86357	7842	94199
Y		196329	16982	213311
Grand Total		282686	24824	307510

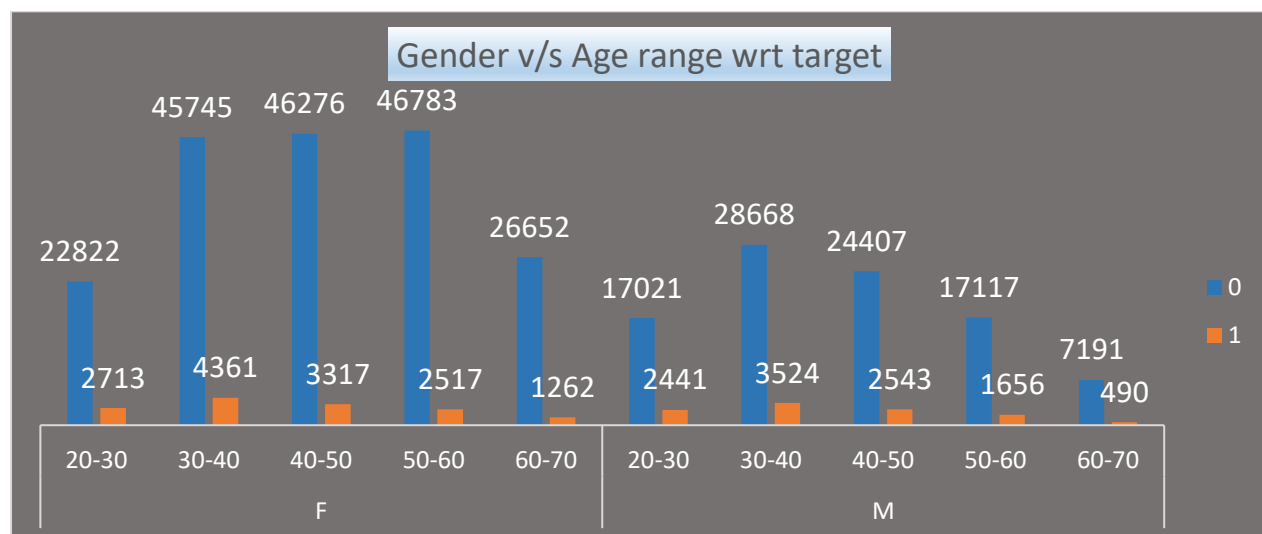


People who own realty have highest proportion in repaying loans than people without realty since 9.4% of total count has problem in paying loans.

Bivariate Analysis

Gender Proportion with age wrt target

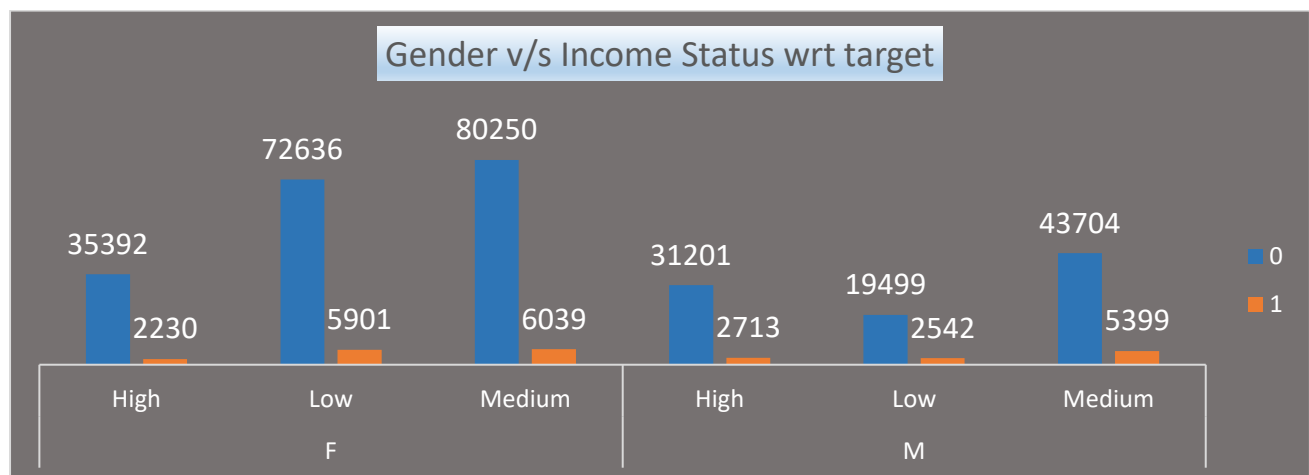
Count of Age_in_years		Column Labels		Grand Total
Row Labels		0	1	
F		188278	14170	202448
20-30		22822	2713	25535
30-40		45745	4361	50106
40-50		46276	3317	49593
50-60		46783	2517	49300
60-70		26652	1262	27914
M		94404	10654	105058
20-30		17021	2441	19462
30-40		28668	3524	32192
40-50		24407	2543	26950
50-60		17117	1656	18773
60-70		7191	490	7681
Grand Total		282682	24824	307506



Here 50-60 age group women and 30-40 age group men face no difficulty in repaying loans and both 20-30 age group women and men find difficulty in repaying loan.

Gender Proportion with income status wrt target

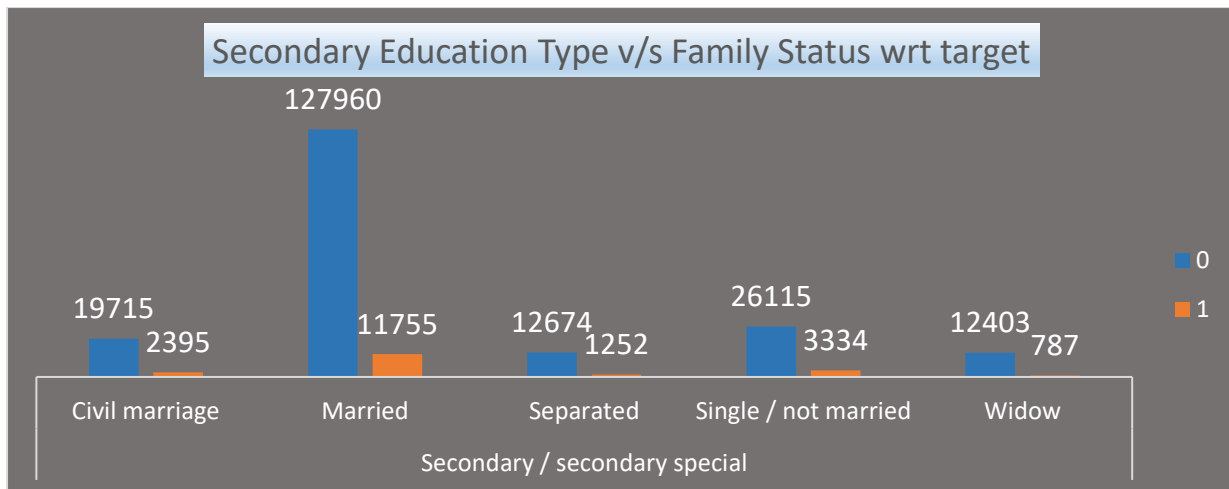
Count of Amt_income_status	Column Labels		
Row Labels	0	1	Grand Total
F	188278	14170	202448
High	35392	2230	37622
Low	72636	5901	78537
Medium	80250	6039	86289
M	94404	10654	105058
High	31201	2713	33914
Low	19499	2542	22041
Medium	43704	5399	49103
Grand Total	282682	24824	307506



Here both men and women of medium income range face no difficulty in repaying loans and both men and women of low income range find difficulty in repaying loan.

Secondary Education Type with family status wrt target

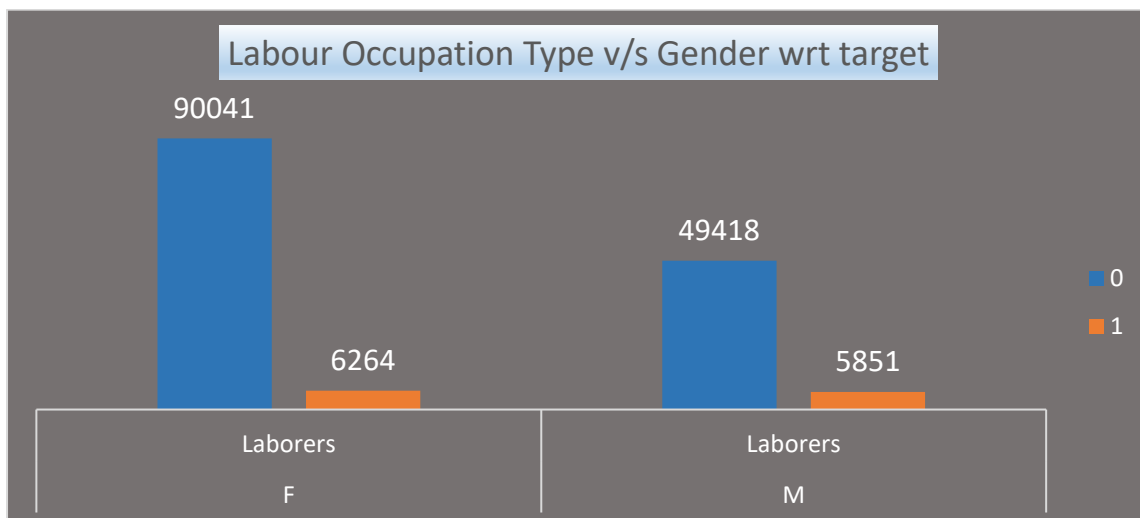
Count of NAME_FAMILY_STATUS	Column Labels		
Row Labels	0	1	Grand Total
Secondary / secondary special	198867	19523	218390
Civil marriage	19715	2395	22110
Married	127960	11755	139715
Separated	12674	1252	13926
Single / not married	26115	3334	29449
Widow	12403	787	13190
Grand Total	198867	19523	218390



Here people who are married with secondary education qualification face no difficulty in repaying loans and people who are single/not married with secondary education qualification find difficulty in repaying loan.

Labour occupation type with gender wrt target

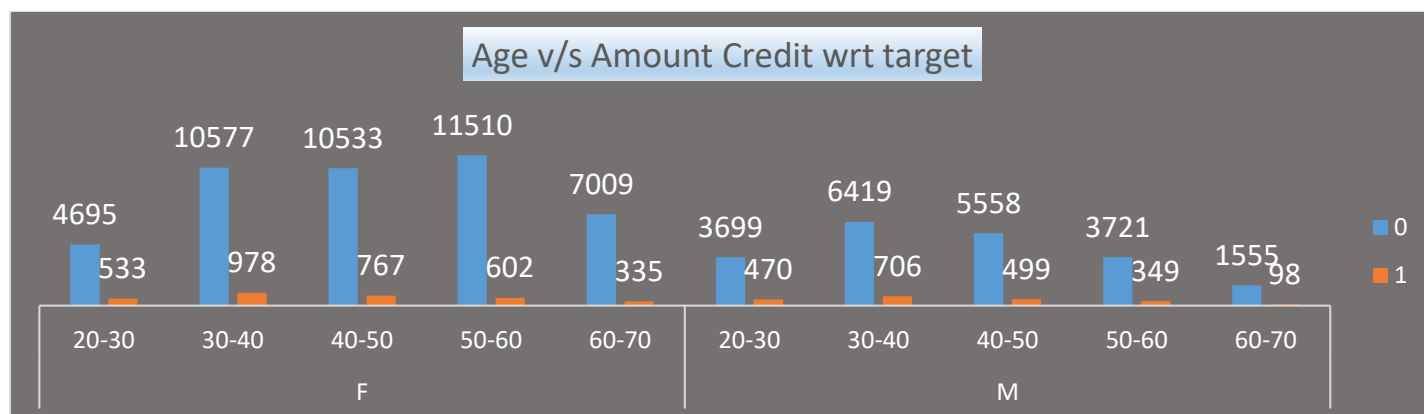
Count of OCCUPATION_TYPE		Column Labels	
Row Labels		0	1
F		90041	6264
Laborers		90041	6264
M		49418	5851
Laborers		49418	5851
Grand Total		139459	12115



Here female labourers face no difficulty in repaying loans and find difficulty in repaying loan.

Amount Credit Quarterly with Gender wrt target

Sum of AMT_REQ_CREDIT_BUREAU_QRT	Column Labels		Grand Total
Row Labels	0	1	
F	44324	3215	47539
20-30	4695	533	5228
30-40	10577	978	11555
40-50	10533	767	11300
50-60	11510	602	12112
60-70	7009	335	7344
M	20952	2122	23074
20-30	3699	470	4169
30-40	6419	706	7125
40-50	5558	499	6057
50-60	3721	349	4070
60-70	1555	98	1653
Grand Total	65276	5337	70613



Here 50-60 age group women and 30-40 age group men face no difficulty in repaying loans and both 20-30 age group women and men find difficulty in repaying loan.

Application Data Analysis Dataset Link

<https://docs.google.com/spreadsheets/d/1V1MsZl-ltZXh16ucwehmxco60ltr4N1Y/edit?usp=sharing&ouid=113123883243133768755&rt=pof=true&sd=true>

Now we are moving to second dataset i.e previous _application

Here also to begin with we have to do EDA as similar to what we have done to our first dataset and then we have to analysis regarding approved, cancelled, refused, and unused offer of previous bank loan application.

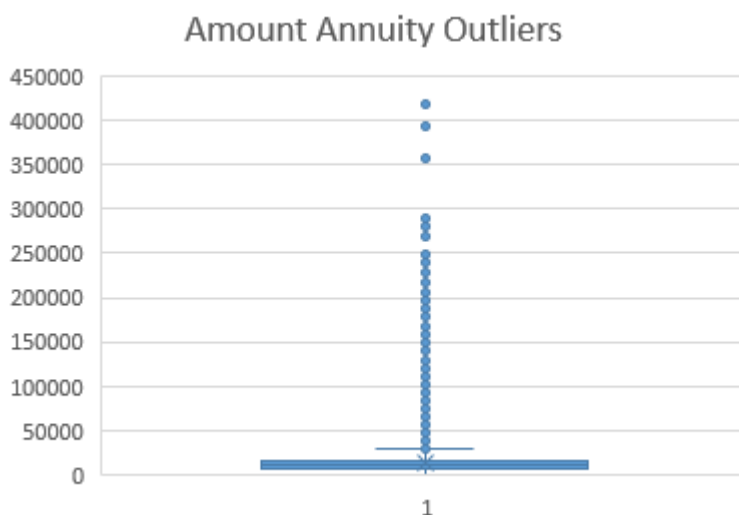
column name	blank cell in %
SK_ID_PREV	0
SK_ID_CURR	0
NAME_CONTRACT_TYPE	0
AMT_APPLICATION	0
AMT_CREDIT	0
WEEKDAY_APPR_PROCESS_START	0
HOUR_APPR_PROCESS_START	0
FLAG_LAST_APPL_PER_CONTRACT	0
NFLAG_LAST_APPL_IN_DAY	0
NAME_CASH_LOAN_PURPOSE	0
NAME_CONTRACT_STATUS	0
DAYS_DECISION	0
NAME_PAYMENT_TYPE	0
CODE_REJECT_REASON	0
NAME_CLIENT_TYPE	0
NAME_GOODS_CATEGORY	0
NAME_PORTFOLIO	0
NAME_PRODUCT_TYPE	0
CHANNEL_TYPE	0
SELLERPLACE_AREA	0
NAME_SELLER_INDUSTRY	0
NAME_YIELD_GROUP	0
PRODUCT_COMBINATION	0.021362
CNT_PAYMENT	22.22123
AMT_ANNUITY	22.22151
AMT_GOODS_PRICE	22.98026
DAYS_FIRST_DRAWING	40.12192
DAYS_FIRST_DUE	40.12192
DAYS_LAST_DUE_1ST_VERSION	40.12192
DAYS_LAST_DUE	40.12192
DAYS_TERMINATION	40.12192
NFLAG_INSURED_ON_APPROVAL	40.12192
NAME_TYPE_SUITE	49.12758
AMT_DOWN_PAYMENT	53.34826

RATE_DOWN_PAYMENT	53.34826
RATE_INTEREST_PRIMARY	99.64523
RATE_INTEREST_PRIVILEGED	99.64523

Here all the highlighted columns which have blank cell percentage more than 40% are dropped and also the irrelevant columns which are not useful for analysis were also dropped.

Amount Annuity Outliers

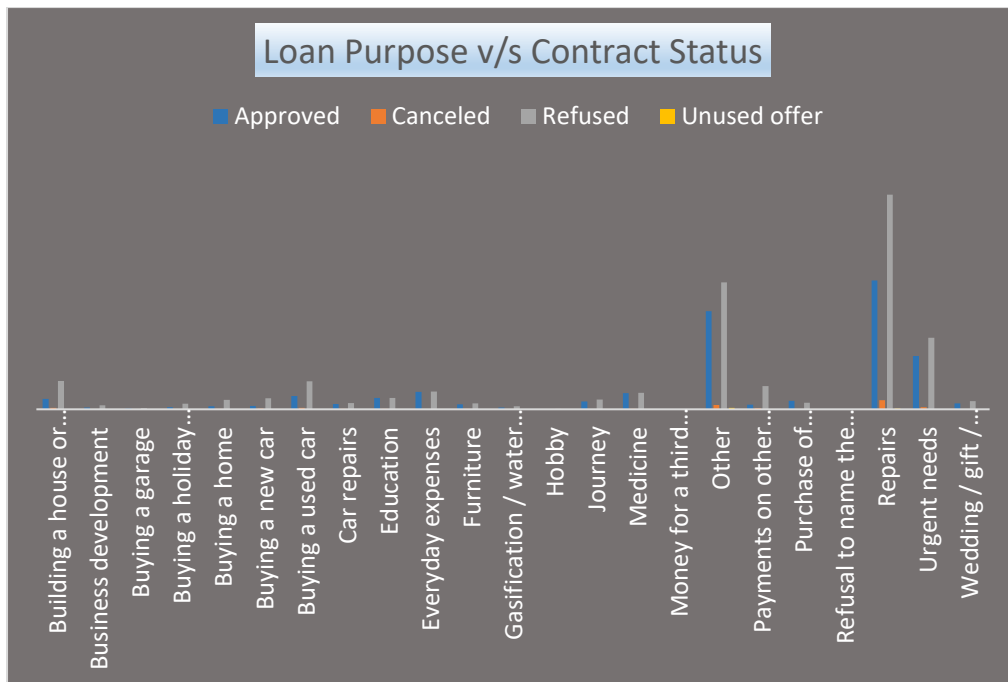
Min	0
1st Quartile	7506.788
Median	11250
3rd Quartile	16737.23
Max	418058.1



Since there were lot of outliers we are doing median imputation to all blank cells i.e replacing all blank cells in amount annuity column with median which is 11250 and we are not removing any outliers because it is related to income which varies from person to person.

Loan Purpose v/s Contract Status

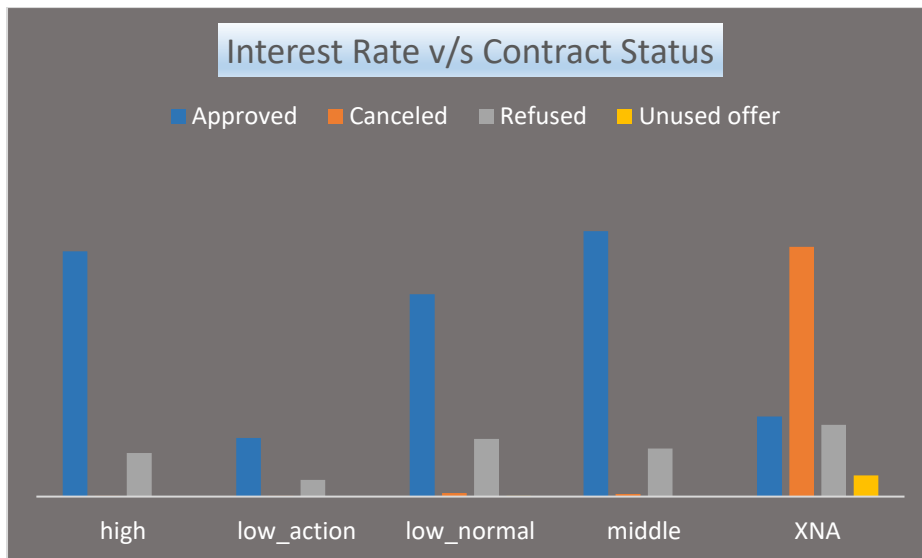
Count NAME_CONTRACT_STATUS	Column Labels				
Row Labels	Approved	Cancelled	Refused	Unused offer	Grand Total
Building a house or an annex	434	60	1188		1682
Business development	78	12	164		254
Buying a garage	28	5	51		84
Buying a holiday home / land	91	13	230		334
Buying a home	130	23	393		546
Buying a new car	139	29	465	4	637
Buying a used car	552	57	1166	9	1784
Car repairs	223	14	256		493
Education	481	14	476	4	975
Everyday expenses	732	8	740	7	1487
Furniture	210	15	250		475
Gasification / water supply	75	3	125		203
Hobby	11		20		31
Journey	329	10	404	2	745
Medicine	676	25	696	5	1402
Money for a third person	10		6		16
Other	4106	186	5310	62	9664
Payments on other loans	189	45	973	3	1210
Purchase of electronic equipment	357	4	280	3	644
Refusal to name the goal	1		7		8
Repairs	5385	381	8973	28	14767
Urgent needs	2228	83	2998		5309
Wedding / gift / holiday	248	10	336		594
Grand Total	16713	997	25507	127	43344



Here clearly we can see that loan for repairs has high percentage refused status which means either they would struggle to repay loan or there are some documentation problem.

Interest Rate v/s Contract Status

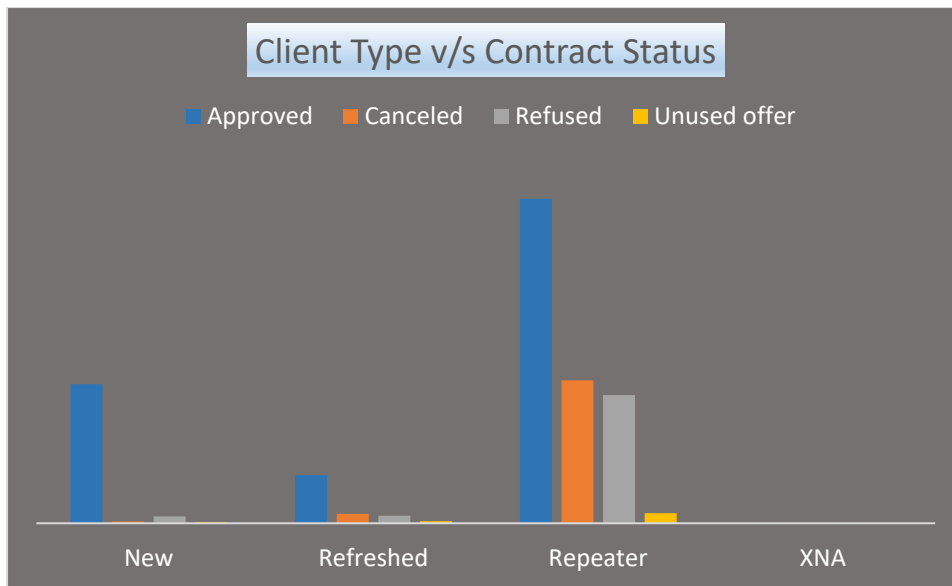
Count NAME_CONTRACT_STATUS	of Column Labels				
Row Labels	Approved	Cancelled	Refused	Unused offer	Grand Total
high	187933	576	33346	45	221900
low_action	44957	564	12679		58200
low_normal	154936	2749	44092	429	202206
middle	203411	1897	36912	107	242327
XNA	61248	191445	55054	16194	323941
Grand Total	652485	197231	182083	16775	1048574



Here we can clearly see that middle interest rate category is approved by high percentage of people and XNA interest rate category is cancelled by high percentage of people.

Client Type v/s Contract Status

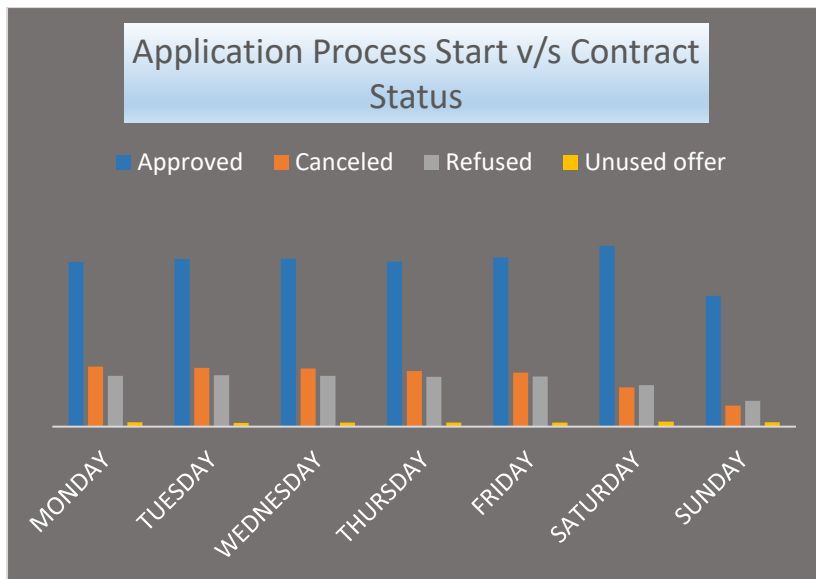
Count of NAME_CONTRACT_STATUS	Column Labels				
Row Labels	Approved	Cancelled	Refused	Unused offer	Grand Total
New	177394	2241	8972	1383	189990
Refreshed	61324	12193	9444	2450	85411
Repeater	413412	182176	163443	12925	771956
XNA	355	621	224	17	1217
Grand Total	652485	197231	182083	16775	1048574



Here we can clearly see that repeater client gets approved by high percentage .

Application Process Start v/s Contract status

Count NAME_CONTRACT_STATUS of	Column Labels				
Row Labels	Approved	Cancelled	Refused	Unused offer	Grand Total
MONDAY	93828	34111	28845	2394	159178
TUESDAY	95366	33477	29263	2184	160290
WEDNESDAY	95539	33042	28928	2258	159767
THURSDAY	93911	31592	28287	2231	156021
FRIDAY	96377	30811	28575	2384	158147
SATURDAY	102901	22278	23549	2804	151532
SUNDAY	74563	11920	14636	2520	103639
Grand Total	652485	197231	182083	16775	1048574



Here most applications approval process starts on Saturday and most applications cancelling process starts on Tuesday.

Previous Application Analysis Dataset Link

https://docs.google.com/spreadsheets/d/1j0yrKIJwY-8_8BBrhtSRkNomAqeU-FHg/edit?usp=sharing&ouid=113123883243133768755&rtpof=true&sd=true