EDA Analysis for Loan Data

Apurva Ukande

Analysis context and Objective

We are provided with loan data for various customers, where each customer is assigned an unique **loan_id.** Our objective is to carry out exploratory data analysis (EDA) to bring out major patterns in loaning behaviour.

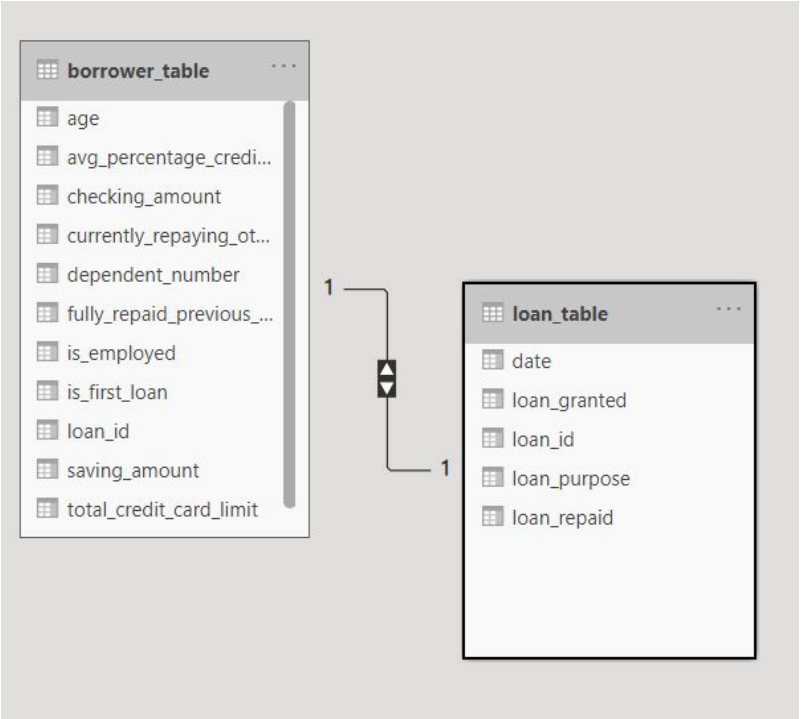
This is intended to guide us in devising a loaning strategy to help bank reduce number of people defaulting the laon.

Provided Data:
We have 2 files, **borrower_table.csv** and **loan_table.csv**.There are 101098 instances of borrower data and 101010 of loan data. Total number of attributes equals 17 including primary key of borrower and loan data file. Borrower and loan table are connected on the basis of loan Id which acts as primary key.

Data Exploration and Modelling

Factors consider-

1. Age of Customer
2. Credit Card Spending
3. Per Person Salary Spent
4. Employment
5. Purpose for Loan
6. Saving vs Checking Amount Relation

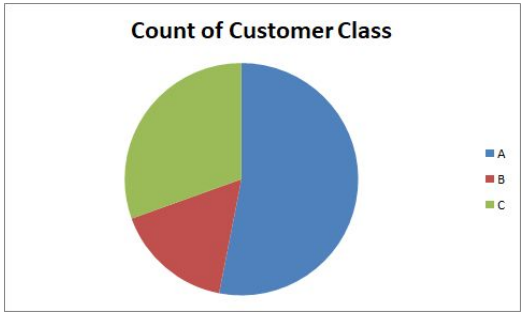


Category 1:

These are group of customers that have no previous loan history and are applying for loan the first time.

Analyzing relative distribution of customer classes:

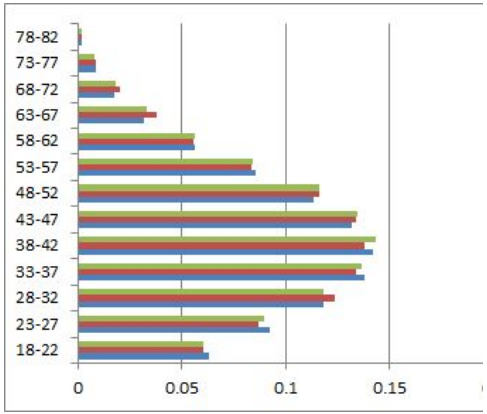
Class	Count of Customer Class	Percentage
A	29158	53.06568147
B	9026	16.42673849
C	16763	30.50758003
Grand Total	54947	



People in category-1 are those who have applied for loan for the first time, and it's been observed from data that **bank reject almost 50%** of loan. But among those whose loan was approved most of them **(66% of remaining) returned the loan**.

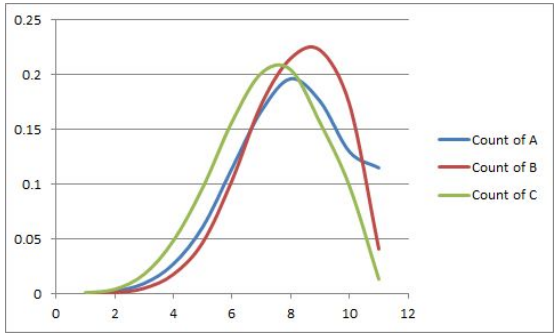
Analyzing the normalized age distribution among various customer categories

Age	Count of age A	Count of age B	Count of age C
18-22	0.063036	0.060603	0.060431
23-27	0.092153	0.086971	0.089423
28-32	0.117875	0.123643	0.118117
33-37	0.137698	0.133725	0.13673
38-42	0.141916	0.137935	0.143053
43-47	0.131799	0.133836	0.134463
48-52	0.113245	0.115998	0.116089
53-57	0.085328	0.083315	0.083875
58-62	0.056314	0.055506	0.056076
63-67	0.032135	0.037669	0.033526
68-72	0.017765	0.020053	0.018076
73-77	0.008814	0.008752	0.008232
78-82	0.001921	0.001994	0.001909



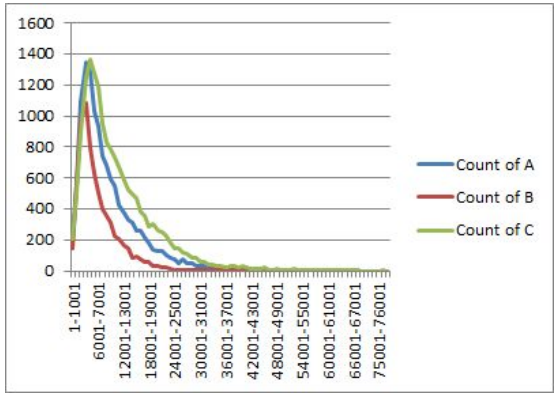
We can see that that **age factor is not very useful** to differentiate among customers that apply for first loan. Normalized distribution for each class remains roughly same

Analyzing Normalized Avg of % Credit Card limit used against customer classes



We can observed from this visual that among the people who grated loan, the people belonging to **class B use more percentage of credit card limit**, indicating possible financially instability.

Analyzing effective salary available per dependent of client



Per person salary expenditure for class C is more as compared to other two classes as median for class C lie in 8K to 9K salary per person range, and also graph of class is is more skewed than other classes.

Analyzing Saving vs Checking Amount of customer against all classes

Saving Amount	Avg Checking Amt of A	Avg Checking Amt of B	Avg Checking Amt of C
0-999	2439.579054	2070.841997	4313.406379
1000-1999	2544.339877	2075.349989	4351.385031
2000-2999	4011.696146	2060.846939	4289.352606
3000-3999	4300.016087	0	4262.865606
4000-4999	4301.089897	0	4255.888889
5000-5999	4271.269949	0	4317.776084
6000-6999	4436.883408	0	4178.268293
7000-7999	4319.5	0	4419.232323
8000-8999	4146.230769	0	3902.307692
9000-10000	7383.5	0	4003.2
10000-10999	0	0	1290

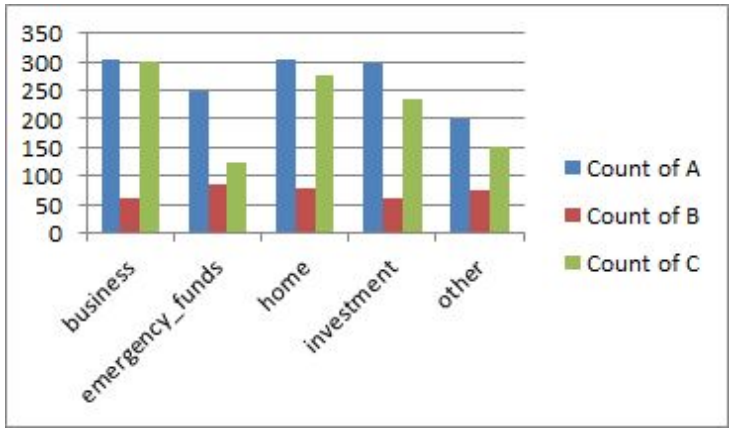
We observe from this data that among all people that **didn't return** the loan, they have **savings <3k** .While those who return have higher savings. Also we note that all the people **who return** typically have **checking amounts greater than 4K** but those who don't have checking amounts only around 2K.

Analyzing Employment status of different customer classes

Employed	Count for A	Count for B	Count for C	% rejected	%defaulter among accepted
No	16450	1906	422	87.6	81.8
Yes	12708	7120	16341	35.1	30.3

Out of all unemployed people that applied for loan, **87.6% of applications are rejected** and also from the rest that are somehow granted the loan, **81.8% did not return the loan**.

Analyzing Purpose of loan against customer classes



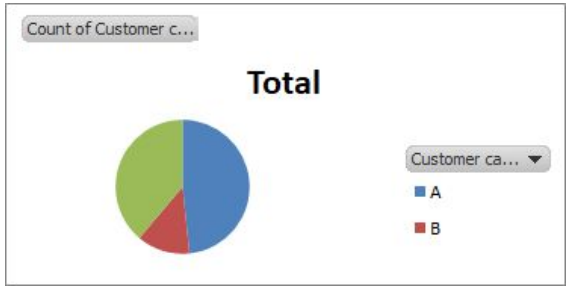
Among all who are granted loan, there is a **high rate of loan defaulters** in case the purpose for loan was to cover **emergency funds and other miscellaneous reasons**. This relate to the fact that business needs , loan for home, or for investing somewhere are usually planned activities so the borrower already has means to return the loan.

Category 2:

These are group of customers that have previous loan history and have neither paid their last loan nor currently paying it

Analyzing relative distribution of customer classes:

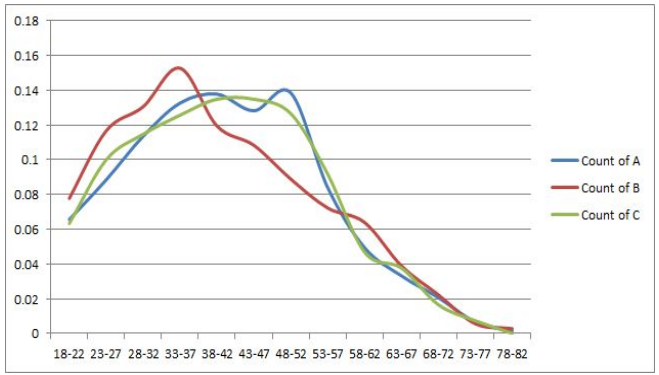
Customer Category	Count of Customer category	Percentage
A	1357	48.34342715
B	360	9.975062344
C	1090	38.8314927
Grand Total	**2807**	



People in category-2 are those who have previous loan(not applied 1st time) and currently not repaying previous loan, and it's been observed from data that **bank reject almost 50%** of loan. But among those whose loan was approved most of them **(75% of remaining) returned the loan**. Indicating stringent financial analysis by bank

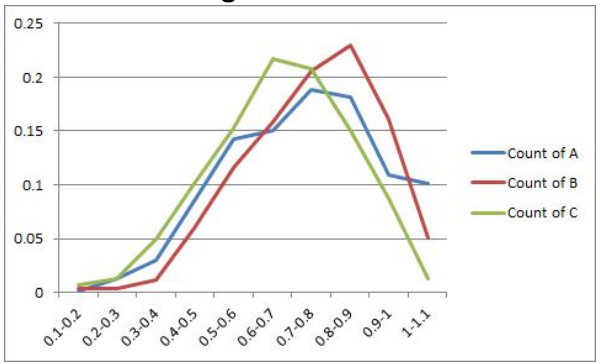
Analyzing the normalized age distribution among various customer categories

Age	Count of A	Count of B	Count of C
18-22	0.065585851	0.077777778	0.063302752
23-27	0.088430361	0.116666667	0.1
28-32	0.11348563	0.130555556	0.114678899
33-37	0.132645542	0.152777778	0.125688073
38-42	0.137803979	0.119444444	0.134862385
43-47	0.128224024	0.108333333	0.134862385
48-52	0.138540899	0.088888889	0.126605505
53-57	0.084008843	0.072222222	0.091743119
58-62	0.049373618	0.063888889	0.046788991
63-67	0.033161385	0.038888889	0.037614679
68-72	0.020633751	0.022222222	0.016513761
73-77	0.006632277	0.005555556	0.00733945
78-82	0.001473839	0.002777778	0



We observe that people who had **not paid their previous loan are younger(red curve)** than those who returned (relatively older age people , aged around 38 to 52 year old).

Analyzing Normalized Avg of % Credit Card limit used against customer classes

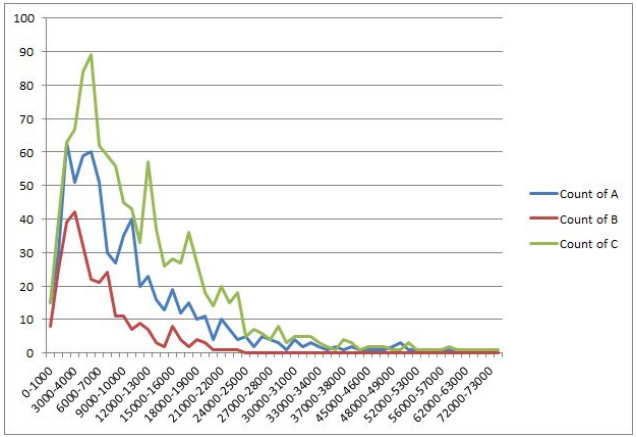


We can observed from this visual that among the people **who returned loan, use 60 to 80% of credit card limit, while class B use 80 to 90% limit.** This implies financially stable people rely less on credit card

Analyzing Saving vs Checking Amount of customer against all classes

Saving Amount	Avg of Checking Amt of A	Avg of Checking Amt of B	Avg of checking Amt of C
7-1006	2793.377119	2062.218182	4281.958974
1007-2006	2736.174098	2065.240437	4365.961702
2007-3006	3977.357955	1551.583333	4093.854545
3007-4006	4536.221311	0	4456.73545
4007-5006	4113.857143	0	3931.847619
5007-6006	4858.583333	0	4485.464286
6007-7006	5008.166667	0	4179.793103
7007-8006	4026	0	3691.666667
8007-9006	855	0	0
9007-10006	8037	0	0

Analyzing effective salary available per dependent of client



From data we can observed that bank thoroughly revised every application while granting loan as people with per person less expenditure, have not granted loan. Also those who granted loan, we can see from median value that class C graph is more skewed compare to class B, indicating financial stability of class A.

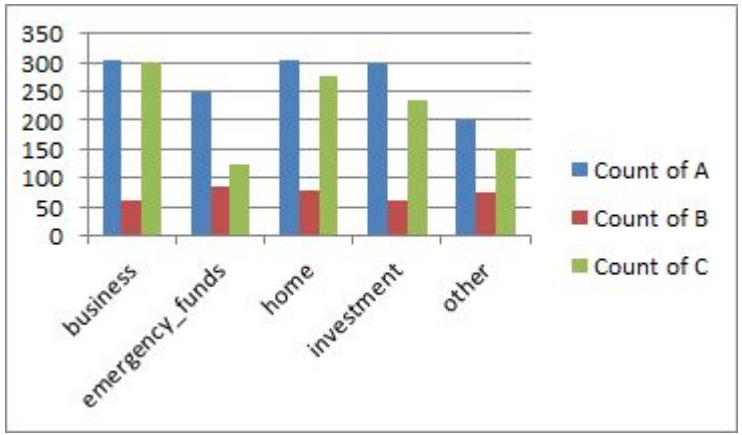
We observe from this data that among all people that **didn't return** the loan, they have **savings <3k** .While those who return have higher savings. Also we note that all the people **who return** typically have **checking amounts greater than 4K** but those who don't have checking amounts only around 2K.

Analyzing Employment status of different customer classes

Employed	Count for A	Count for B	Count for C	% rejected	%defaulter among accepted
No	680	72	26	87.4	73.5
Yes	677	288	1064	33.4	21.3

Out of all unemployed people that applied for loan, **87.% of applications are rejected** and also from the rest that are somehow granted the loan, **73.5% did not return the loan**.

Analyzing Purpose of loan against customer classes



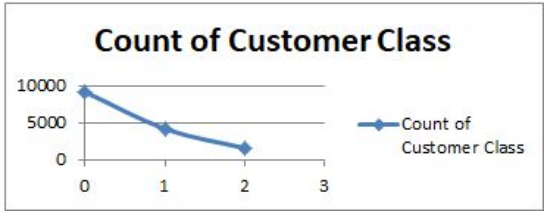
Among all who are granted loan, there is a **high rate of loan defaulters** in case the purpose for loan was to cover **emergency funds and other miscellaneous reasons**. This relate to the fact that business needs , loan for home, or for investing somewhere are usually planned activities so the borrower already has means in place to return the loan.

Category 3:

These are group of customers that have previous loan history and failed to repay their last loan. But they are currently paying their loans. This could be due to improved financial status. These could be potential good customers for bank.

Analyzing relative distribution of customer classes:

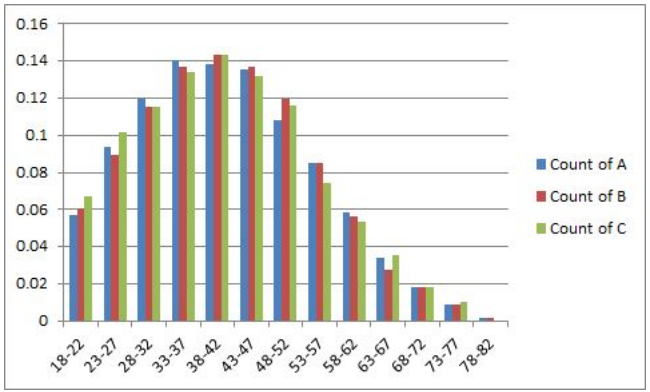
Class	Count of Customer Class	Percentage
A	1159	62.95491581
B	533	28.95165671
C	149	8.093427485
Grand Total	**1841**	



By looking at data we observed that bank reject around **63% of applications as they failed to pay their last loan,** but they currently paying loan. Only few percent people granted loan, **among which only 8% returned.**

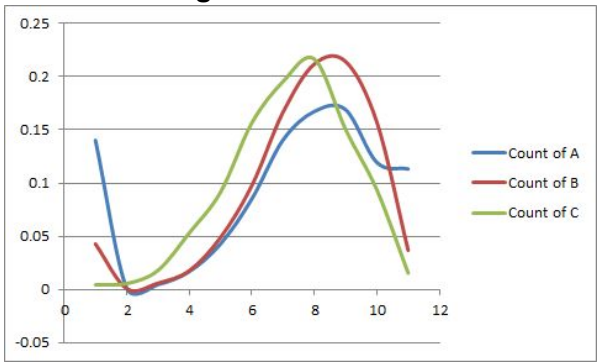
Analyzing the normalized age distribution among various customer categories

Age	Count of A	Count of B	Count of C
18-22	0.04659189	0.069418386	0.053691275
23-27	0.088869715	0.095684803	0.053691275
28-32	0.130284728	0.112570356	0.053691275
33-37	0.146678171	0.144465291	0.053691275
38-42	0.129421915	0.1369606	0.053691275
43-47	0.133735979	0.127579737	0.053691275
48-52	0.120793788	0.12945591	0.053691275
53-57	0.079378775	0.084427767	0.053691275
58-62	0.062985332	0.048780488	0.053691275
63-67	0.036238136	0.022514071	0.053691275
68-72	0.018119068	0.016885553	0.053691275
73-78	0.006902502	0.011257036	0.053691275



Here also, we can see that that **age factor is roughly same among all three classes.** So age factor is not helping us to draw insight among these customer in this category.

Analyzing Normalized Avg of % Credit Card limit used against customer classes

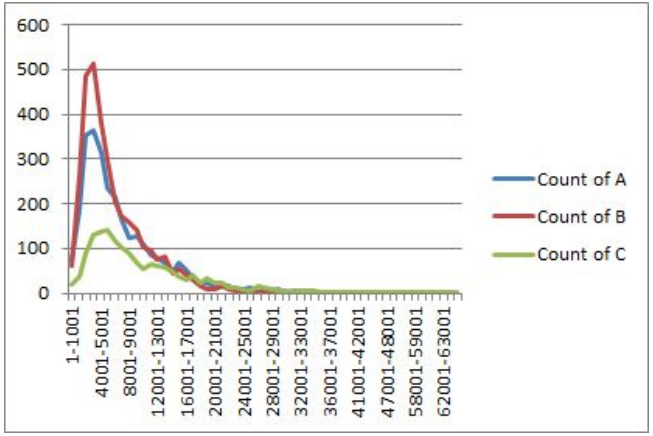


Here also we can deduced that people who don't have any financial crunch depend less on credit card as compare to those who are financially unstable.

Analyzing Saving vs Checking Amount of customer against all classes

We observe from this data also that among all people that **didn't return** the loan, they have **savings <3k** .While those who return have higher savings. Also we note that all the people who return typically have checking amounts greater than 4K but those who don't have checking amounts only around 2K.

Analyzing effective salary available per dependent of client



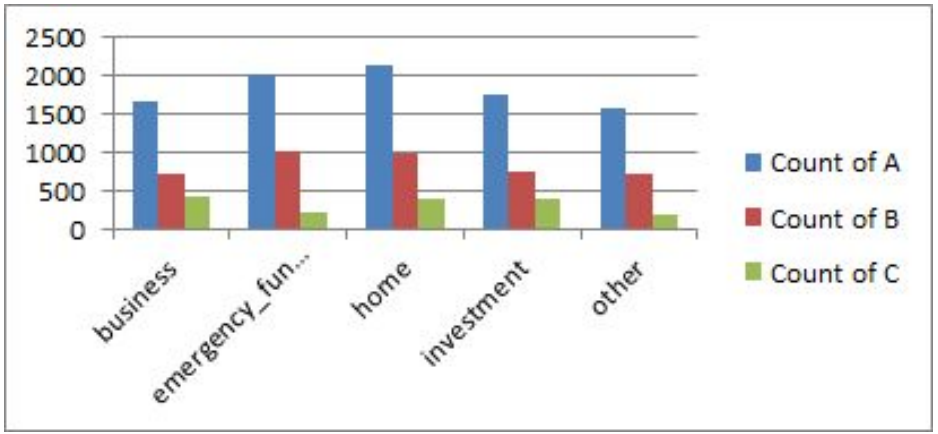
Median of class A and B lie in range 4K to 6K, whereas median of class C lie in range 7K to 8K range. This implies data of C is more skewed as compare to other, indicating more per person spent in case of class C. More spent in general implies more liquidity.

Analyzing Employment status of different customer classes

Employed	Count for A	Count for B	Count for C	% rejected	%defaulter among accepted
No	780	97	2	88.7	97.9
Yes	379	436	147	39.4	74.7

Out of all unemployed people that applied for loan, **88.7% of applications are rejected** and also from the rest that are somehow granted the loan, **97.9% did not return the loan**.

Analyzing Purpose of loan against customer classes



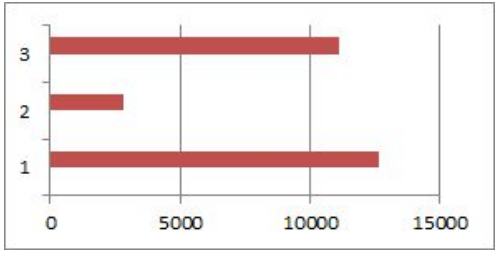
Among all who are granted loan, there is a **high rate of loan defaulters** in case the purpose for loan was to cover **emergency funds and other miscellaneous reasons**. This relate to the fact that business needs , loan for home, or for investing somewhere are usually planned activities so the borrower already has means to return the loan.

Category 4:

These are group of customers that have previous loan history and have repaid their previous loan and not currently paying any other loans

Analyzing relative distribution of customer classes:

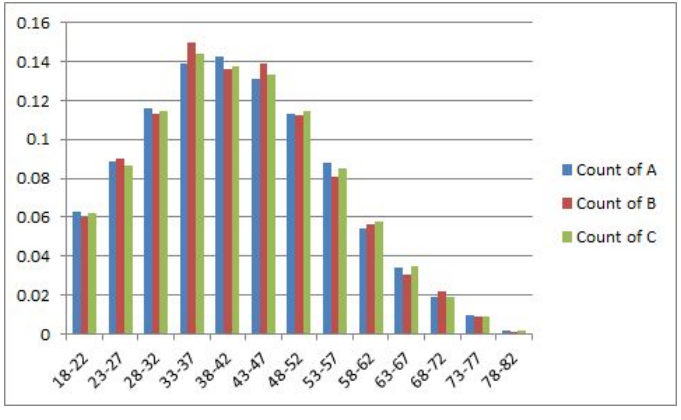
Class	Count of Customer Class	Percentage
A	12626	0.475896122
B	2818	0.106215371
C	11087	0.417888508



Bank approved application of **53% people among all who applied for loan.** These people have previous loan history and paid their last loan, but currently not paying, so **almost half of the loan applications are rejected**.

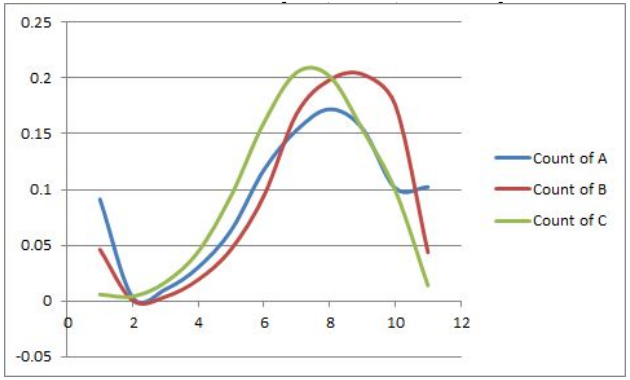
Analyzing the normalized age distribution among various customer categories

Age	Count of A	Count of B	Count of C
18-22	0.062649	0.059972	0.062145
23-27	0.088943	0.090135	0.086408
28-32	0.116189	0.112846	0.114549
33-37	0.139078	0.150106	0.143952
38-42	0.142563	0.136267	0.137729
43-47	0.130762	0.139106	0.133039
48-52	0.113179	0.112491	0.114639
53-57	0.087676	0.080908	0.085145
58-62	0.054491	0.056423	0.057815
63-67	0.034136	0.030163	0.034725
68-72	0.01885	0.021647	0.018761
73-77	0.009663	0.008872	0.0092
78-82	0.001822	0.001065	0.001894



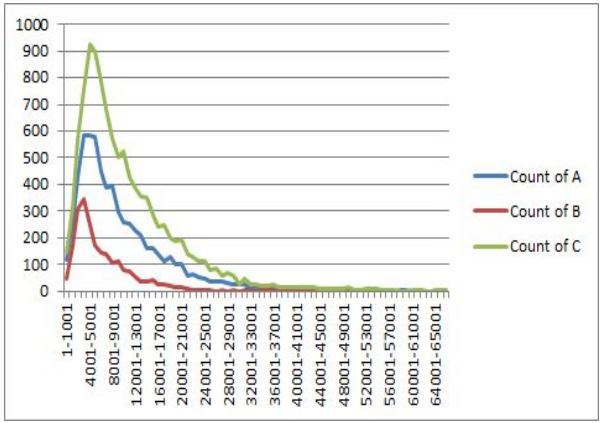
Similar trend is observed among all classes so age data is not helping us to draw meaningful insights to segregate people.

Analyzing Normalized Avg of % Credit Card limit used



Similar trend is observed here too, **people who uses 80 to 100% of their credit card limit, didn't return their loan** indicating financial problem.

Analyzing effective salary available per dependent of client



Graph of class C is more skewed as compare to other classes as value of median for class C lie in 8K to 9 K range whereas for class A it's 7K to 8K and for class B it's 4K to 5 K. As 4k to 5K is low amount, which implies class B financial status is not good as people belonging to this class is unable to return loan and also per person spent is less means cannot afford more expenditure.

Analyzing Saving vs Checking Amount of customer against all classes

Saving Amount	Avg of checking Amt of A	Avg of Checking Amt of B	Avg of Checking Amt of C
0-999	2778.365306	2032.441493	4251.383651
1000-1999	2885.372163	2080.473299	4377.586332
2000-2999	4113.593162	2094.293103	4252.927301
3000-3999	4339.628458	0	4298.786727
4000-4999	4407.015815	0	4254.639104
5000-5999	4267.539894	0	4269.487223
6000-6999	4271.4	0	4042.468619
7000-7999	4212.065217	0	4245.785714
8000-8999	3745.25	0	4880.733333
9000-9999	3075	0	4996.5

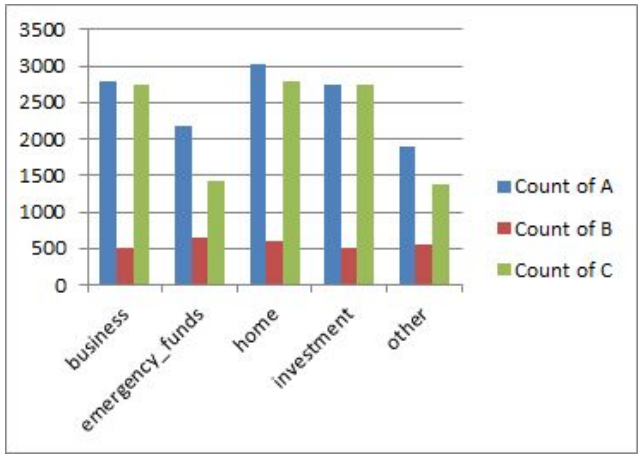
Similar observation as category 1, category 2 and category 3.

Analyzing Employment status of different customer classes

Employed	Count for A	Count for B	Count for C	% rejected	%defaulter among accepted
No	6075	566	277	87.8	67.2
Yes	6551	2252	1080	33.4	17.2

Out of all unemployed people that applied for loan, **87.8% of applications are rejected** and also from the rest that are somehow granted the loan, **67.2% did not return the loan**.

Analyzing Purpose of loan against customer classes



Among all who are granted loan, there is a **high rate of loan defaulters** in case the purpose for loan was to cover **emergency funds and other miscellaneous reasons**. This relate to the fact that business needs , loan for home, or for investing somewhere are usually planned activities so the borrower already has means to return the loan.

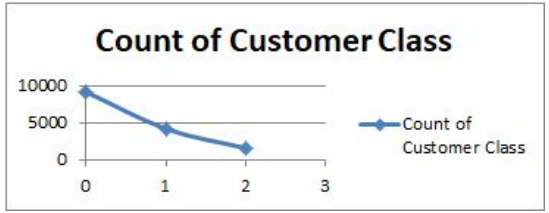
Category 5:

These are group of customers that have previous loan history and have repaid their loans and also currently paying others

61% loan are rejected by bank of people belonging to this category.

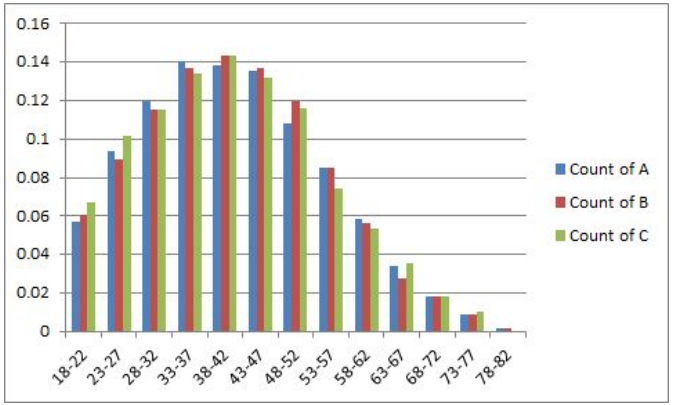
Analyzing relative distribution of customer classes:

Class	Class	Percentage
0	9146	0.61079204
1	4211	0.281220783
2	1617	0.107987178
Grand Total	14974	



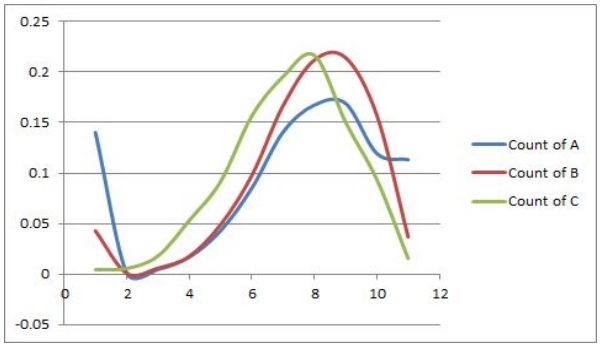
Analyzing the normalized age distribution among various customer categories

Age dist	Count of A	Count of B	Count of C
18-22	0.057293	0.060793	0.06679
23-27	0.093374	0.08929	0.101422
28-32	0.119615	0.115175	0.115028
33-37	0.140389	0.136785	0.134199
38-42	0.13864	0.143671	0.143476
43-47	0.135469	0.137022	0.131725
48-52	0.108135	0.119687	0.115646
53-57	0.085283	0.085253	0.074212
58-62	0.058605	0.056044	0.053803
63-67	0.034223	0.027784	0.03525
68-72	0.018259	0.018285	0.017934
73-77	0.008747	0.008787	0.010513
78-82	0.001968	0.001425	0



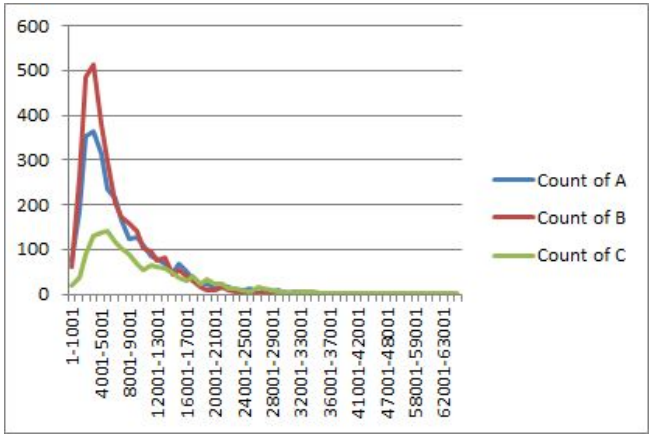
We can see that that **age factor is not very useful** to differentiate among customers that apply for first loan. Normalized distribution for each class remains roughly same

Analyzing Normalized Avg of % Credit Card limit used against customer classes



We can observed from this visual that among the people who grated loan, people who returned their sanction loan depend relatively less on credit card as compare to those who didn't return.

Analyzing effective salary available per dependent of client



Similar trend is observe in this category as other category.

Analyzing Saving vs Checking Amount of customer against all classes

Row Labels	Avg of Checking Amt of A	Avg of Checking Amt of B	Avg of Checking Amt of C
0-999	2144.581898	2047.221841	4251.784983
1000-1999	2180.360301	2070.073842	4180.489209
2000-2999	3466.303704	2042.051546	4338.705882
3000-3999	4439.874346	0	4236.925566
4000-4999	3942.340206	0	4252.05
5000-5999	3887.12069	0	4399.849315
6000-6999	4201.625	0	4413.3125
7000-7999	4208.6	0	3645
9000-9999	1109	0	0

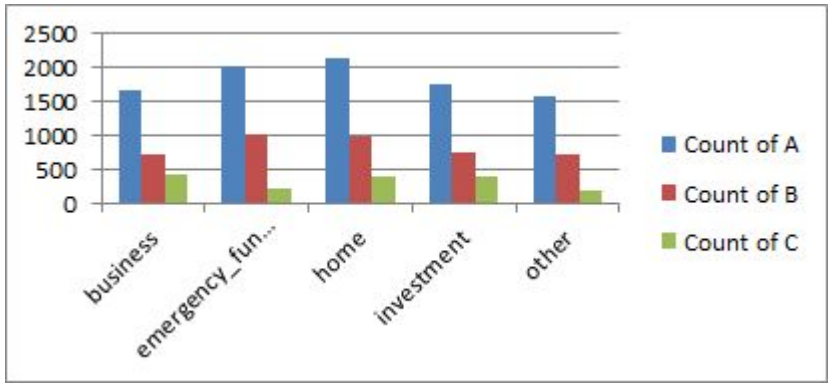
Here also, we observe from this data that among all people that didn't return the loan, they have savings <3k .While those who return have higher savings. Also we note that all the people who return typically have checking amounts greater than 4K but those who don't have checking amounts only around 2K.

Analyzing Employment status of different customer classes

Employed	Count for A	Count for B	Count for C	% rejected	%defaulter among accepted
No	6226	885	44	87.0	95.3
Yes	2920	3326	1573	37.3	67.8

Out of all unemployed people that applied for loan, **87% of applications are rejected** and also from the rest that are somehow granted the loan, **95.3% did not return the loan**.

Analyzing Purpose of loan against customer classes



Among all who are granted loan, there is a **high rate of loan defaulters** in case the purpose for loan was to cover **emergency funds and other miscellaneous reasons**. This relate to the fact that business needs , loan for home, or for investing somewhere are usually planned activities so the borrower already has means to return the loan.

Strategy

Key factor Bank should consider while considering loan applications:

Age - Bank should prefer to give loan to age group 30 to 50 years old as they are considered financially stable. People in this age group have worked for a years and still have years left to reply the loan easily. People above 60 years old find it challenging to repay back.

Employment - Employment is very crucial while considering loan application. Unemployed people serve negative impression as they find difficult to pay back. So bank should prefer employed person as self-employed people usually undergo more scrutiny than salaried people with stable monthly income.

Amount of Loan - Banks should look into the amount of credit that the borrower has applied for. A higher loan amount will lead to greater scrutiny by the bank. On the other hand, a smaller loan application could be approved more quickly based on your relationship with the bank.

Purpose of Loan - Bank should ask to disclose purpose of loan. In case of high amount application there is chance of high risk like business with no experience then bank should reject the application, or charge higher interest. In case the loan amount is for low-risk purposes, like renovations and repair to your home or construction of a house, then its relatively fine to approve it easily.

Strategy

Other factors bank give a check while sanctioning applications are **Repayment History**, **Credit card limit of previous loan, Character, Bonds etc.**

Categorization and Modelling of data rough sheet link(my work work):

EDA Work - https://drive.google.com/file/d/1GNK9tRqqBgrv6KpLptcqH_g0d7ll_rCl/view?usp=sharing

Categorization work - https://drive.google.com/file/d/16kehDDslr_abkoZcNcP_HPdfc9KbTluZ/view?usp=sharing

Thank You