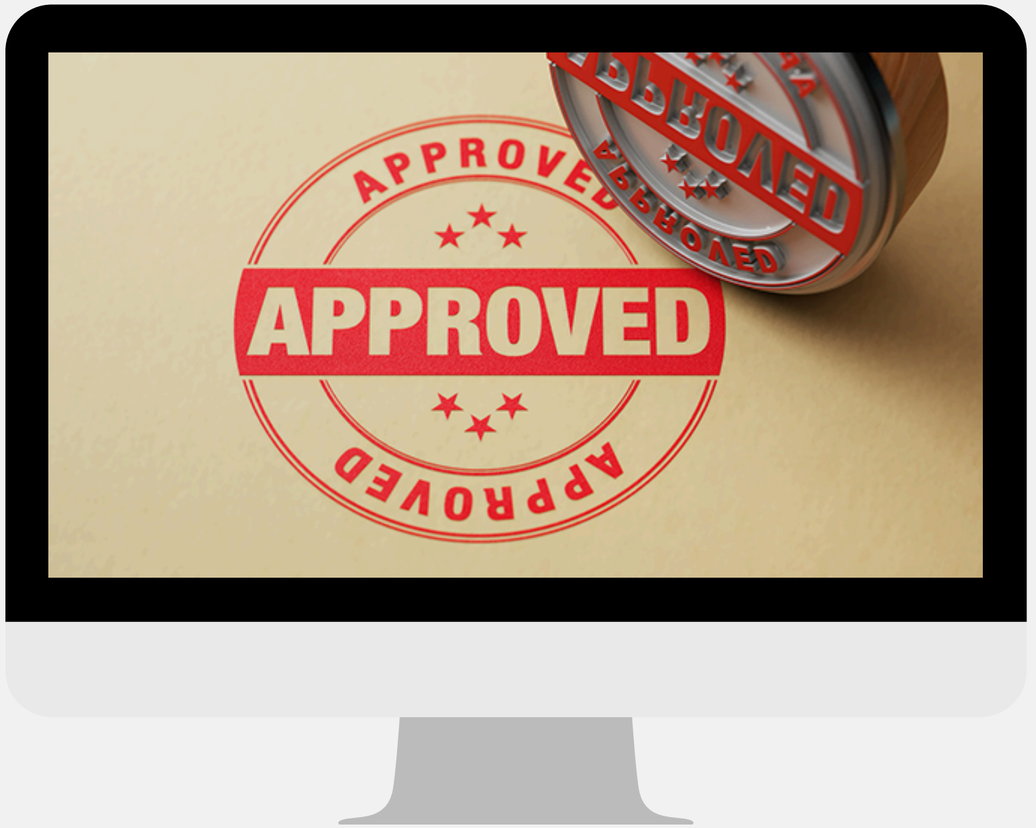
Data Analytics Capstone Project

Credit Default Risk Analysis

Murugappan Sivabalan

Problem Statment

- A financial institution aims to streamline its loan approval process by accurately predicting the likelihood of loan approval based on applicant characteristics. The dataset includes features such as income, credit history, loan amount, and more, which will be analyzed to develop a predictive model.



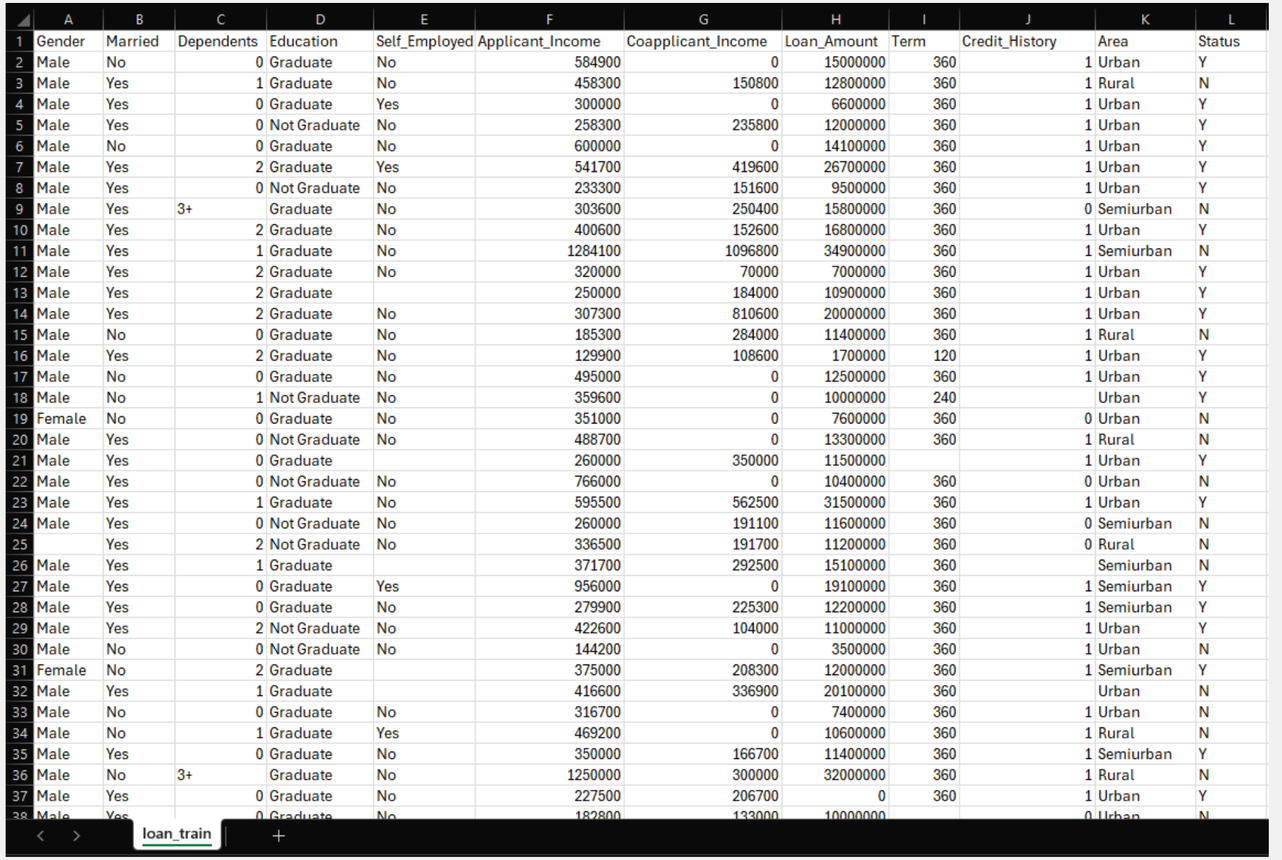
How can we enhance the accuracy of loan approval predictions and reduce the rate of defaults, while ensuring fair and efficient decision-making?

Dataset

Kaggle Dataset (loan_train excel): https://www.kaggle.com/datasets/mirzahasnine/loan-data-set_

Datasheet columns provided (12 rows):

- Gender
- Marriage status
- Number of Dependents
- Education
- Self-Employed(Y/N)
- Applicant's Income
- Co-Applicant's income (if there is one)
- Loan amount
- Loan Term (months)
- Credit history (of being able to repay)
- Area of Residence
- Status of Loan Application



	A	B	C	D	E	F	G	H	I	J	K	L
1	Gender	Married	Dependents	Education	Self_Employed	Applicant_Income	Coapplicant_Income	Loan_Amount	Term	Credit_History	Area	Status
2	Male	No	0	Graduate	No	584900	0	15000000	360	1	Urban	Y
3	Male	Yes	1	Graduate	No	458300	150800	12800000	360	1	Rural	N
4	Male	Yes	0	Graduate	Yes	300000	0	6600000	360	1	Urban	Y
5	Male	Yes	0	Not Graduate	No	258300	235800	12000000	360	1	Urban	Y
6	Male	No	0	Graduate	No	600000	0	14100000	360	1	Urban	Y
7	Male	Yes	2	Graduate	Yes	541700	419600	26700000	360	1	Urban	Y
8	Male	Yes	0	Not Graduate	No	233300	151600	9500000	360	1	Urban	Y
9	Male	Yes	3+	Graduate	No	303600	250400	15800000	360	0	Semiurban	N
10	Male	Yes	2	Graduate	No	400600	152600	16800000	360	1	Urban	Y
11	Male	Yes	1	Graduate	No	1284100	1096800	34900000	360	1	Semiurban	N
12	Male	Yes	2	Graduate	No	320000	70000	7000000	360	1	Urban	Y
13	Male	Yes	2	Graduate		250000	184000	10900000	360	1	Urban	Y
14	Male	Yes	2	Graduate	No	307300	810600	20000000	360	1	Urban	Y
15	Male	No	0	Graduate	No	185300	284000	11400000	360	1	Rural	N
16	Male	Yes	2	Graduate	No	129900	108600	1700000	120	1	Urban	Y
17	Male	No	0	Graduate	No	495000	0	12500000	360	1	Urban	Y
18	Male	No	1	Not Graduate	No	359600	0	10000000	240		Urban	Y
19	Female	No	0	Graduate	No	351000	0	7600000	360	0	Urban	N
20	Male	Yes	0	Not Graduate	No	488700	0	13300000	360	1	Rural	N
21	Male	Yes	0	Graduate		260000	350000	11500000		1	Urban	Y
22	Male	Yes	0	Not Graduate	No	766000	0	10400000	360	0	Urban	N
23	Male	Yes	1	Graduate	No	595500	562500	31500000	360	1	Urban	Y
24	Male	Yes	0	Not Graduate	No	260000	191100	11600000	360	0	Semiurban	N
25		Yes	2	Not Graduate	No	336500	191700	11200000	360	0	Rural	N
26	Male	Yes	1	Graduate		371700	292500	15100000	360		Semiurban	N
27	Male	Yes	0	Graduate	Yes	956000	0	19100000	360	1	Semiurban	Y
28	Male	Yes	0	Graduate	No	279900	225300	12200000	360	1	Semiurban	Y
29	Male	Yes	2	Not Graduate	No	422600	104000	11000000	360	1	Urban	Y
30	Male	No	0	Not Graduate	No	144200	0	3500000	360	1	Urban	N
31	Female	No	2	Graduate		375000	208300	12000000	360	1	Semiurban	Y
32	Male	Yes	1	Graduate		416600	336900	20100000	360		Urban	N
33	Male	No	0	Graduate	No	316700	0	7400000	360	1	Urban	N
34	Male	No	1	Graduate	Yes	469200	0	10600000	360	1	Rural	N
35	Male	Yes	0	Graduate	No	350000	166700	11400000	360	1	Semiurban	Y
36	Male	No	3+	Graduate	No	1250000	300000	32000000	360	1	Rural	N
37	Male	Yes	0	Graduate	No	227500	206700	0	360	1	Urban	Y
38	Male	Yes	0	Graduate	No	182800	133000	10000000		0	Urban	N

loan_train

Dataset Preparation

Data Preprocessing

For the missing values in the Term (loan duration) column, I calculated a debt-to-income ratio (DeptIncomeRatio) for each row. I then used this ratio to estimate the missing loan terms by comparing it with the available loan terms.

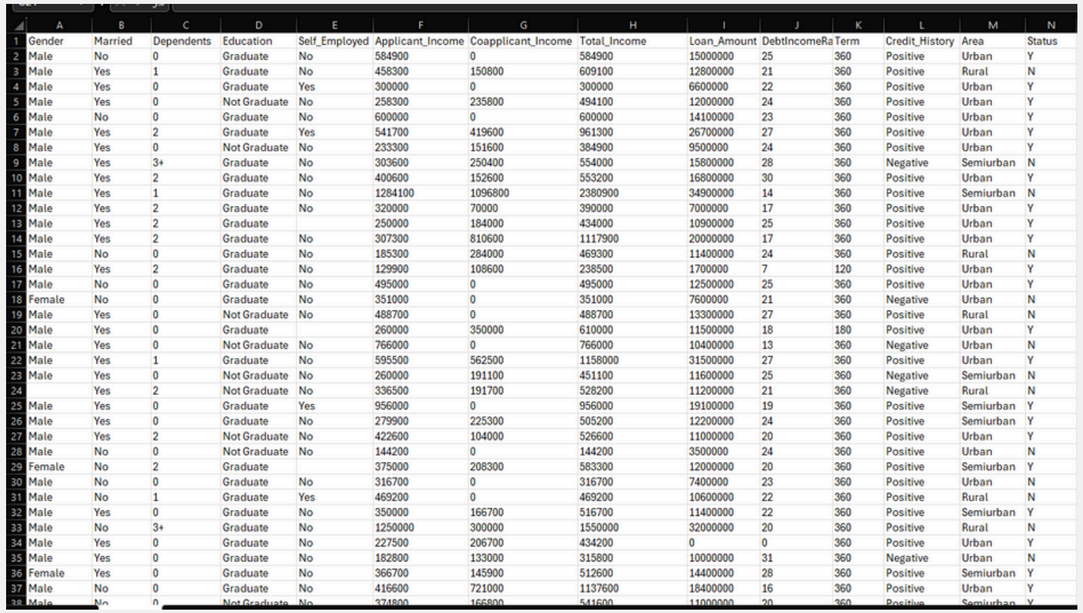
Data Integration

To make calculations easier, I added both the Applicant's Income and the Co-Applicant's income and formed a new column called Total Income

Data Cleaning

I believe Credit History and Loan Status Column are the two most important indicators and is rather hard to guess and hence removed rows that had missing values in this column

Refined Dataset



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Gender	Married	Dependents	Education	Self_Employed	Applicant_Income	Coapplicant_Income	Total_Income	Loan_Amount	DebtIncomeRa	Term	Credit_History	Area	Status
2	Male	No	0	Graduate	No	584900	0	584900	15000000	25	360	Positive	Urban	Y
3	Male	Yes	1	Graduate	No	458300	150800	609100	12800000	21	360	Positive	Rural	N
4	Male	Yes	0	Graduate	Yes	300000	0	300000	6600000	22	360	Positive	Urban	Y
5	Male	Yes	0	Not Graduate	No	258300	235800	494100	12000000	24	360	Positive	Urban	Y
6	Male	No	0	Graduate	No	600000	0	600000	14100000	23	360	Positive	Urban	Y
7	Male	Yes	2	Graduate	Yes	541700	419600	961300	26700000	27	360	Positive	Urban	Y
8	Male	Yes	0	Not Graduate	No	233300	151600	384900	9500000	24	360	Positive	Urban	Y
9	Male	Yes	3+	Graduate	No	303600	250400	554000	15800000	28	360	Negative	Semiurban	N
10	Male	Yes	2	Graduate	No	400600	152600	553200	16800000	30	360	Positive	Urban	Y
11	Male	Yes	1	Graduate	No	1284100	1096800	2380900	34900000	14	360	Positive	Semiurban	N
12	Male	Yes	2	Graduate	No	320000	70000	390000	7000000	17	360	Positive	Urban	Y
13	Male	Yes	2	Graduate		250000	184000	434000	10900000	25	360	Positive	Urban	Y
14	Male	Yes	2	Graduate	No	307300	810600	1117900	20000000	17	360	Positive	Urban	Y
15	Male	No	0	Graduate	No	185300	284000	469300	11400000	24	360	Positive	Rural	N
16	Male	Yes	2	Graduate	No	129900	108600	238500	1700000	7	120	Positive	Urban	Y
17	Male	No	0	Graduate	No	495000	0	495000	12500000	25	360	Positive	Urban	Y
18	Female	No	0	Graduate	No	351000	0	351000	7600000	21	360	Negative	Urban	N
19	Male	Yes	0	Not Graduate	No	488700	0	488700	13300000	27	360	Positive	Rural	N
20	Male	Yes	0	Graduate		260000	350000	610000	11500000	18	180	Positive	Urban	Y
21	Male	Yes	0	Not Graduate	No	766000	0	766000	10400000	13	360	Negative	Urban	N
22	Male	Yes	1	Graduate	No	595500	562500	1158000	31500000	27	360	Positive	Urban	Y
23	Male	Yes	0	Not Graduate	No	260000	191100	451100	11600000	25	360	Negative	Semiurban	N
24		Yes	2	Not Graduate	No	336500	191700	528200	11200000	21	360	Negative	Rural	N
25	Male	Yes	0	Graduate	Yes	956000	0	956000	19100000	19	360	Positive	Semiurban	Y
26	Male	Yes	0	Graduate	No	279900	225300	505200	12200000	24	360	Positive	Semiurban	Y
27	Male	Yes	2	Not Graduate	No	422600	104000	526600	11000000	20	360	Positive	Urban	Y
28	Male	No	0	Not Graduate	No	144200	0	144200	3500000	24	360	Positive	Urban	N
29	Female	No	2	Graduate		375000	208300	583300	12000000	20	360	Positive	Semiurban	Y
30	Male	No	0	Graduate	No	316700	0	316700	7400000	23	360	Positive	Urban	N
31	Male	No	1	Graduate	Yes	469200	0	469200	10600000	22	360	Positive	Rural	N
32	Male	Yes	0	Graduate	No	350000	166700	516700	11400000	22	360	Positive	Semiurban	Y
33	Male	No	3+	Graduate	No	1250000	300000	1550000	32000000	20	360	Positive	Rural	N
34	Male	Yes	0	Graduate	No	227500	206700	434200	0	0	360	Positive	Urban	Y
35	Male	Yes	0	Graduate	No	182800	133000	315800	10000000	31	360	Negative	Urban	N
36	Female	Yes	0	Graduate	No	366700	145900	512600	14400000	28	360	Positive	Semiurban	Y
37	Male	No	0	Graduate	No	416600	721000	1137600	18400000	16	360	Positive	Urban	Y
38	Male	No	0	Not Graduate	No	374800	166800	541600	11000000	20	360	Positive	Semiurban	Y

Gender	**Male/Female**
Married	Yes if married, No if single
Dependents	No.of people who rely financially on the applicant
Education	If College Degree is attained
Self-Employed	Yes if self-employed
Applicant Income	Yearly income of Applicant

Co-Applicant Income	**Yearly income of Applicant (If any)**
Total Income	Combined income of Applicant and Co-Applicant
Loan Amount	Amount of Loan applied for
DebtIncome Ratio	$\frac{\text{Loan Amount}}{\text{Total Income}}$
Term	Amount of months to repay loan
Credit History	Positive - Previously applied for loan and successfully repaid
Area	Living area (Urban, Semiurban, Rural)
Status	Y if Loan is approved

Applicants’ Background

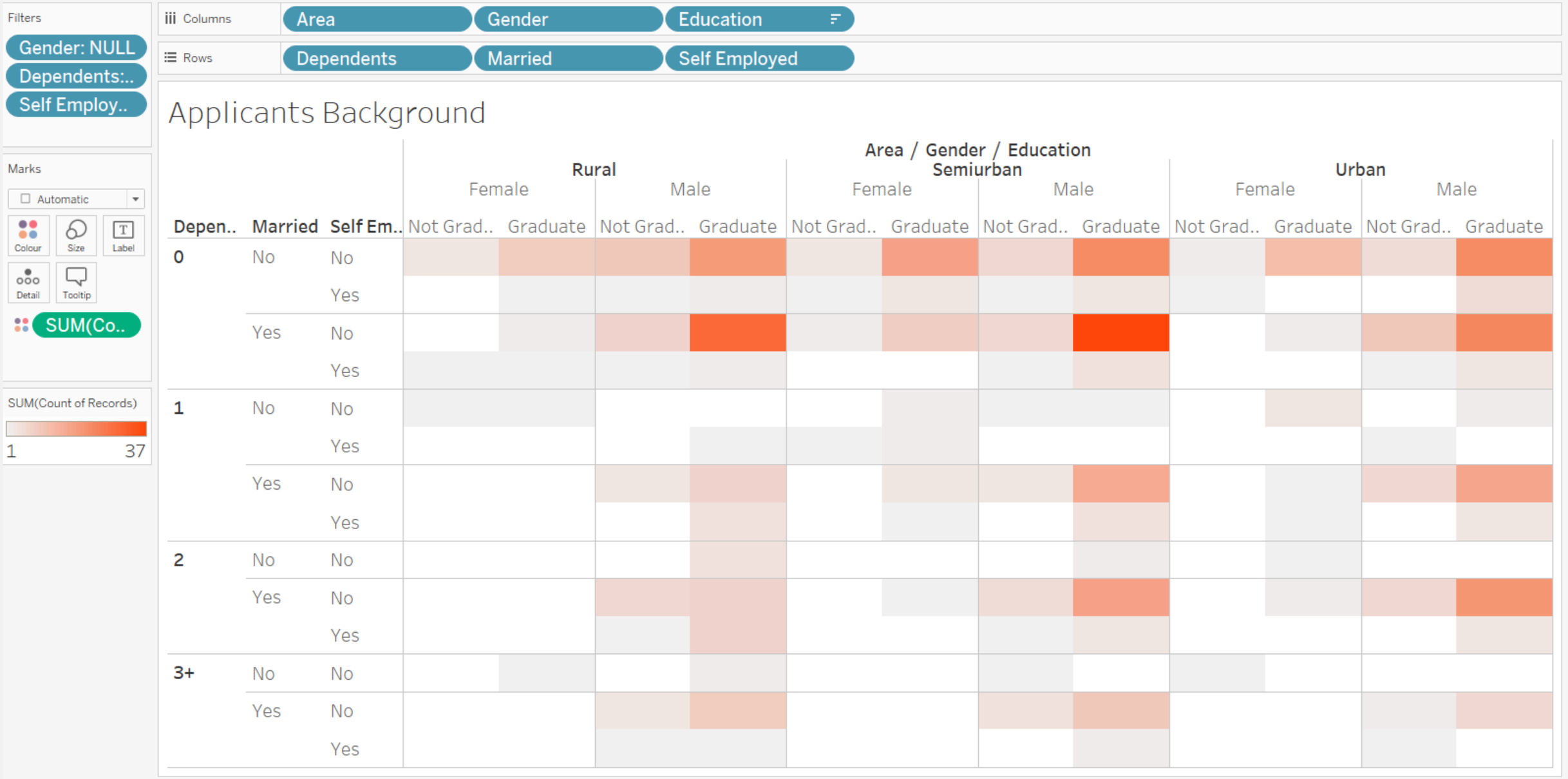


Figure 1. HeatMap of Applicant’s background against various factors

Applicants’ Background Analysis

1. From Fig 1 (previous page), the highest occurrence in the heat map reflects **Graduate Males living in the semiurban area that are married, not self employed and have no dependents** indicating that this group of people have the highest say in the selected database

2. Upon further observation, heatmap suggests that on a zoomed out focus, there is generally a high response from graduated males, so I have a second heatmap with lesser indicators to have a closer look

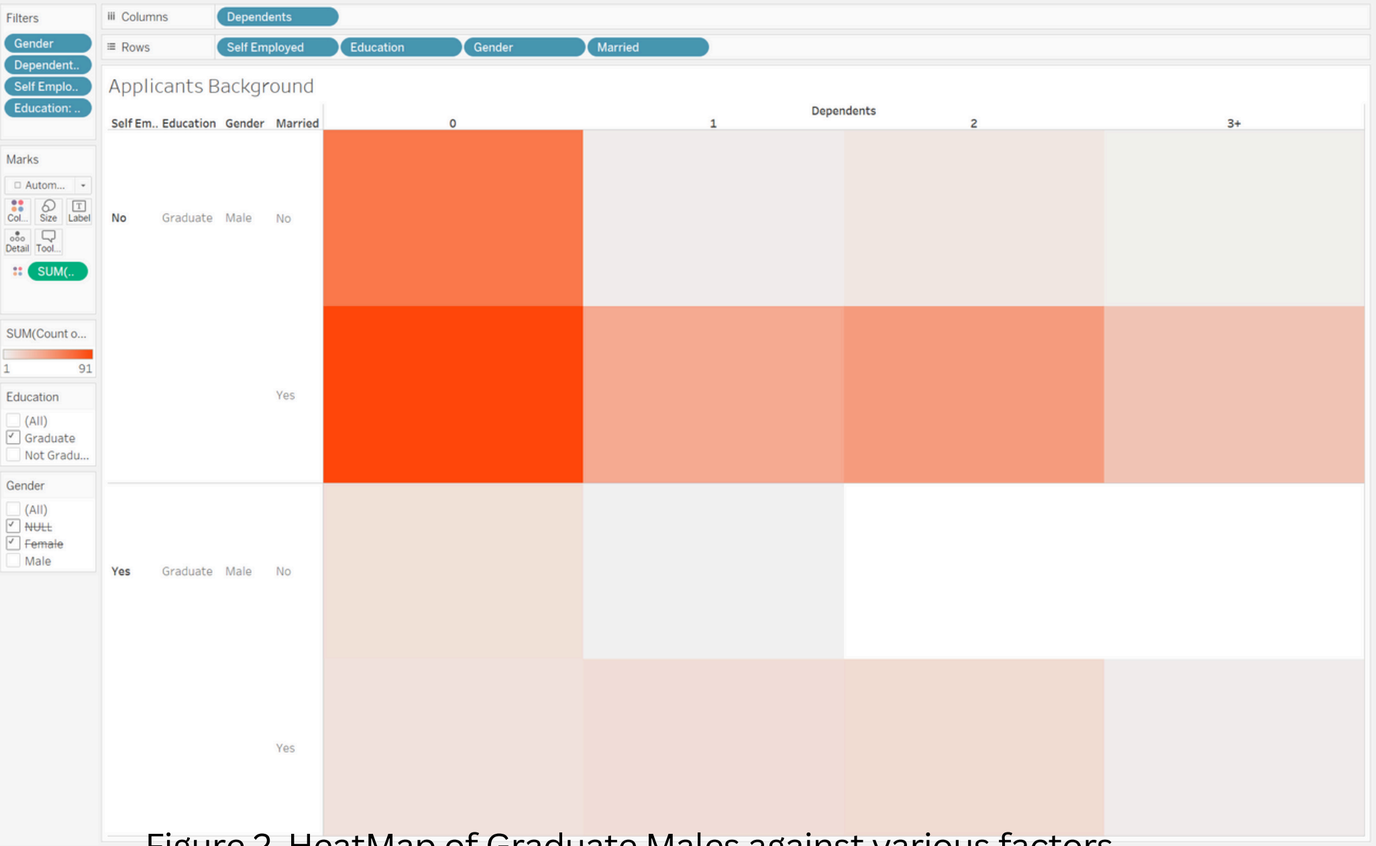


Figure 2. HeatMap of Graduate Males against various factors

3. Now looking at Fig 2, this indicates much clearly that our database selected has a large influence from **Graduate Men that are not self employed and have no dependents.**

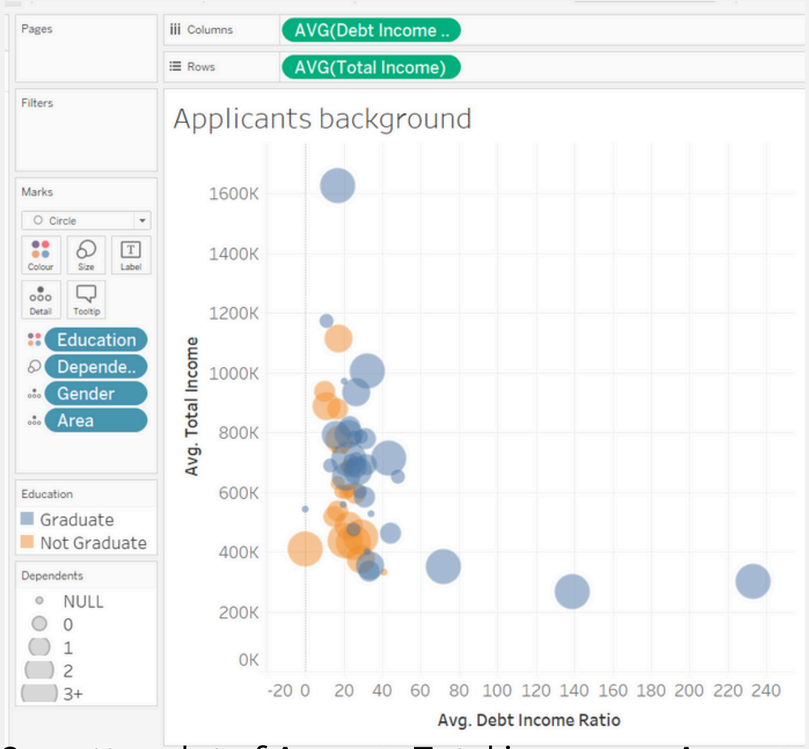


Fig 3. scatter plot of Average Total income vs Average Debt income ratio

4. Looking at Fig 3, most respondents also indicate that graduates tend to have an average higher total income and average higher debt income ratio

Factors affecting Credit Score

1. Education Vs Credit Score
2. Marriage Vs Credit Score
3. Dependables Vs Credit History
4. Total Income Vs Loan Amount
5. Area Vs Loan Approval (Status)

Education Vs Credit History

1. Positive Credit History (A good record of consistently making timely payments) from our dataset is at about 84%. Fig4 shows the educational distribution of the applicants according to gender

2. Fig 5 shows that amongst the 84%, a bulk of the proportion (67%) comes from an educated background (college degree) meaning that an educated applicant has a high chance of making timely payments. But eitherways, for both educated and non-educated backgrounds, majority of the applicants diligently do make timely payments mostly citing that education might not be a huge factor affecting credit defaulting

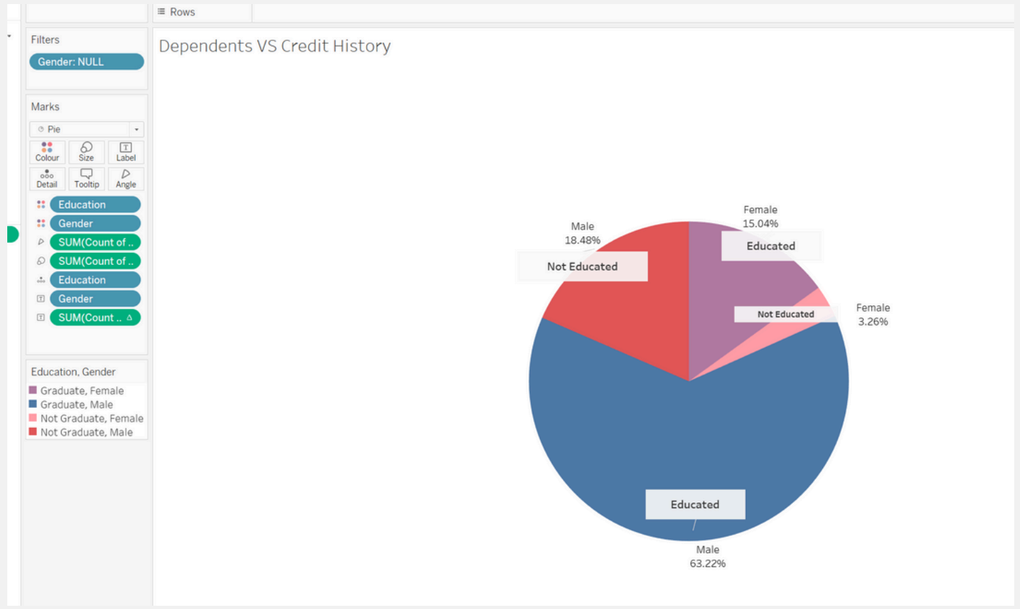


Figure 4. Eduction Distribution from database

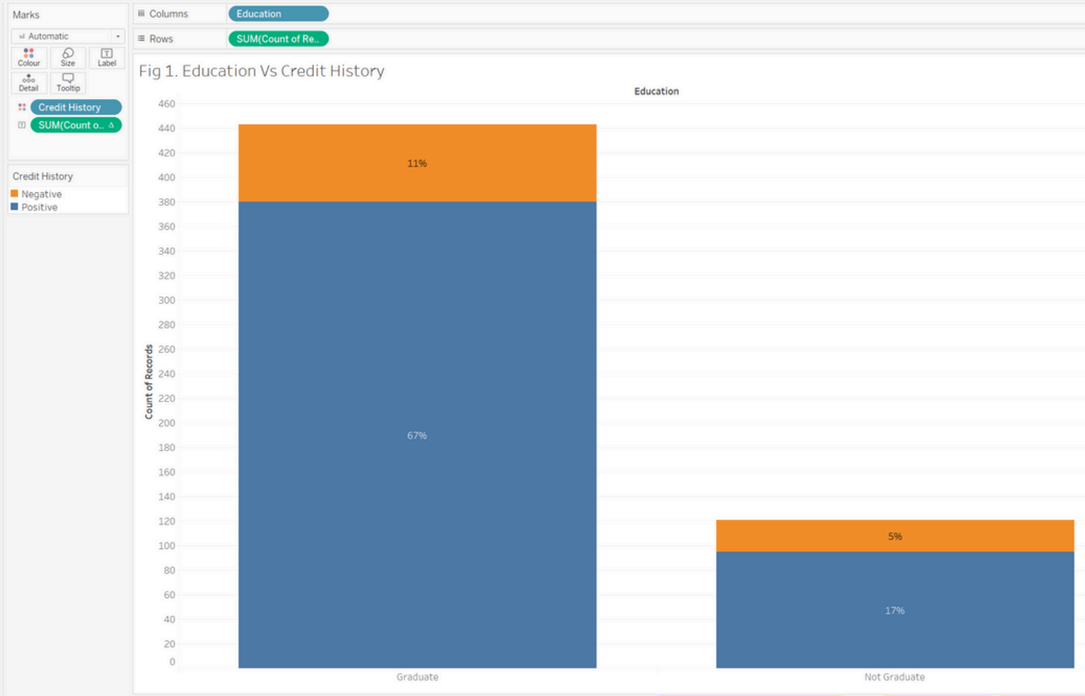


Figure 5. Bar Graph Education Vs Credit Score

Marriage Vs Credit History

Married applicants generally show a higher percentage of positive credit history compared to single applicants

Insight: Marriage may bring financial stability, dual incomes, and better financial planning, contributing to positive credit histories.

Further Insight: But the increase in positive credit history is only by a mere 20%. Does other factors inside marriage like number of dependables play a factor as to how likely applicants are able to make timely payments? (next page)

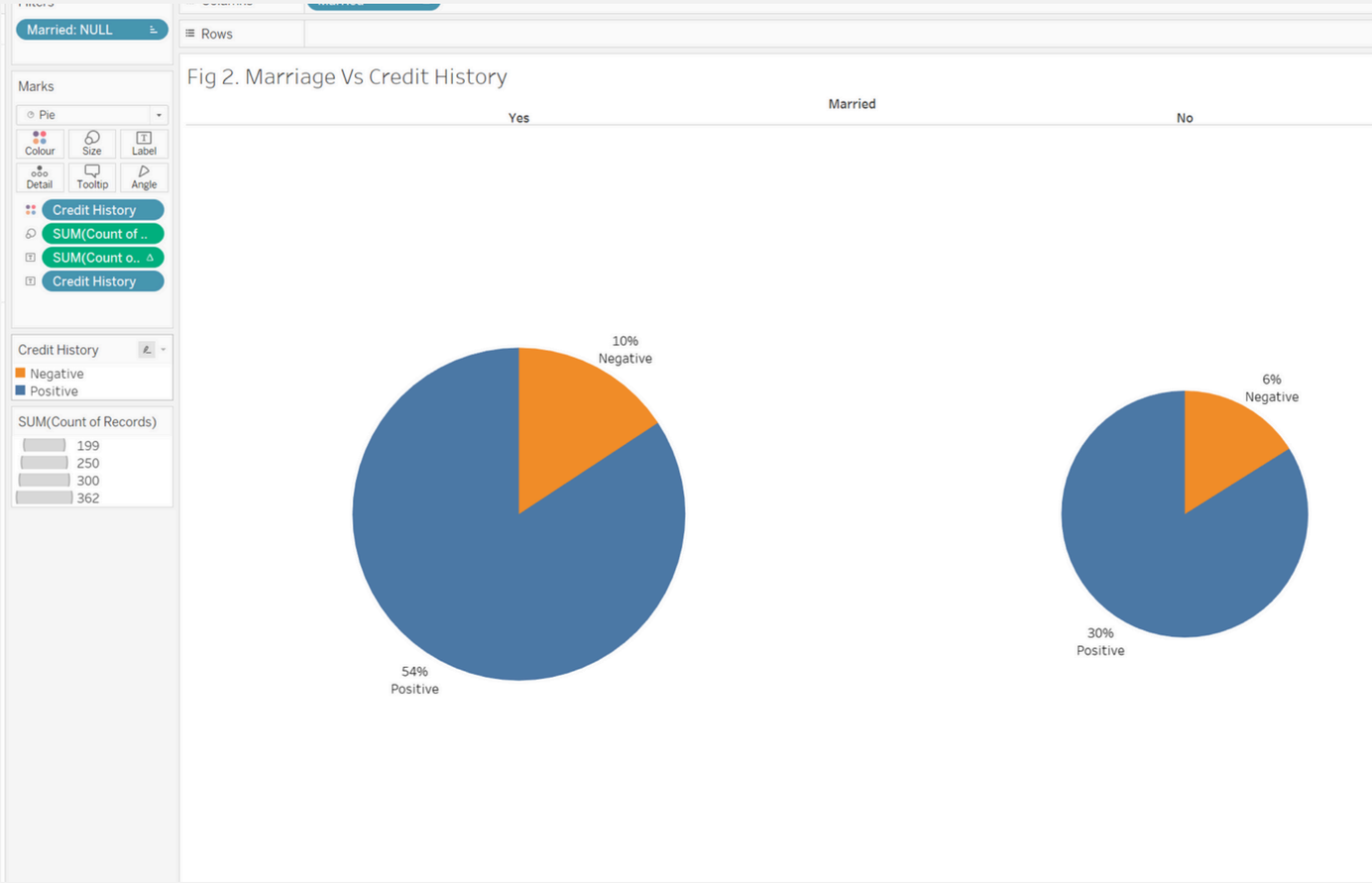


Fig 6. Marriage Vs Credit History

Dependables Vs Credit History

Figure 7 shows that married people are likely to have more people dependant on them financially. This extra increase in dependable may highly be because of children from marriage

Figure 8 shows that the higher the number of dependables, the lower the likelihood of having a positive credit history

This goes to show that while marriage increases the probability of a positive credit history, one should delve further and see the amount of dependables an applicant has as well.

Insight: Having dependents can increase financial responsibilities, which may impact loan repayment capabilities.

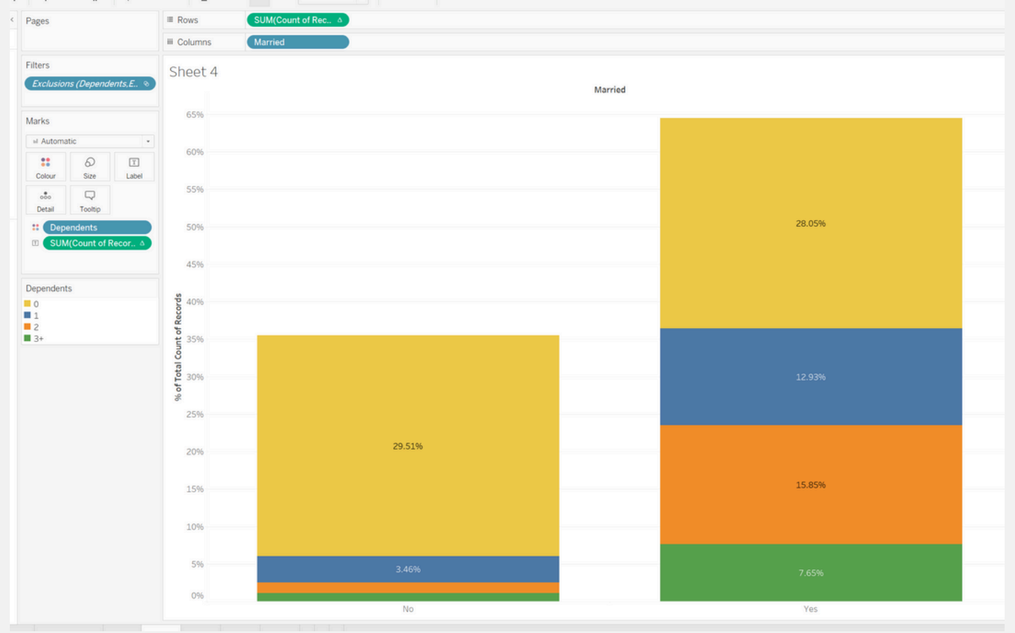


Fig 7. Marriage VS Dependables

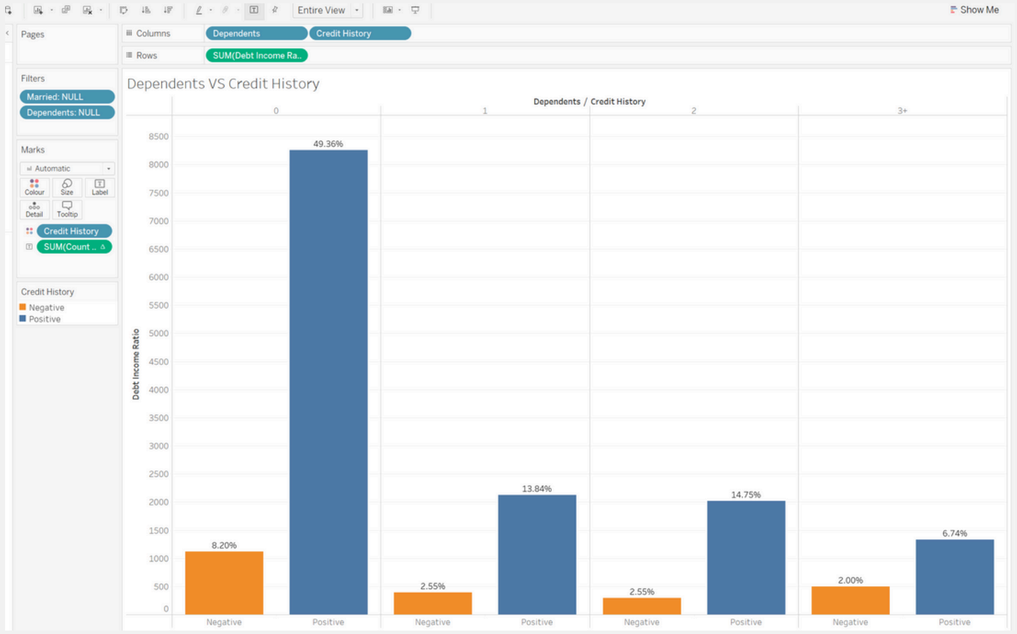


Fig 8. Dependables Vs Credit History

Income Vs Loan Amount

Finding: From Figure 9, Applicants with higher total incomes tend to have higher loan approval rates (status).

Insight: Total income is a significant factor in determining loan approval. Higher incomes provide more assurance of repayment capability.

Recommendation: Consider setting minimum total income thresholds for loan approvals to reduce the risk of default.

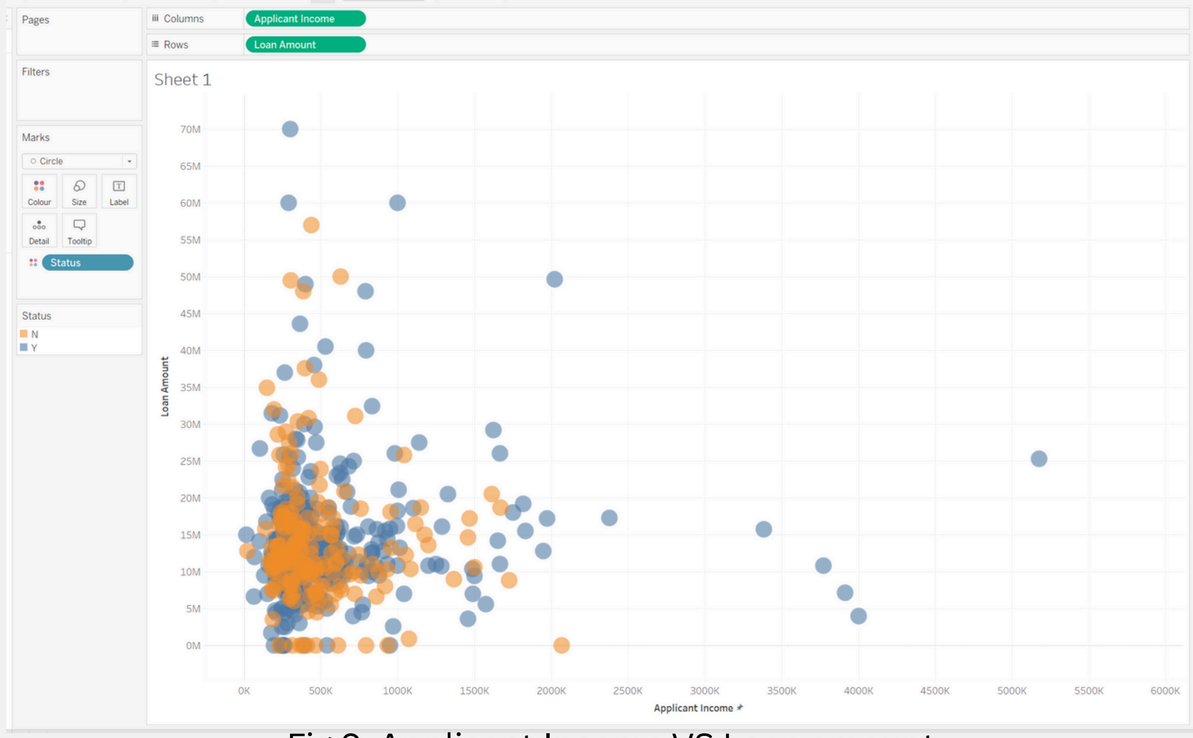


Fig 9. Applicant Income VS Loan amount

We are also able to tell from fig 10 that higher applicant incomes tend to be associated with lower debt-to-income ratios.

Insight: There appears to be a threshold for debt-to-income ratios, above which the likelihood of loan approval decreases significantly.

Recommendation: Applicants with debt-to-income ratios above a certain level should be carefully evaluated or required to provide additional collateral or guarantees.

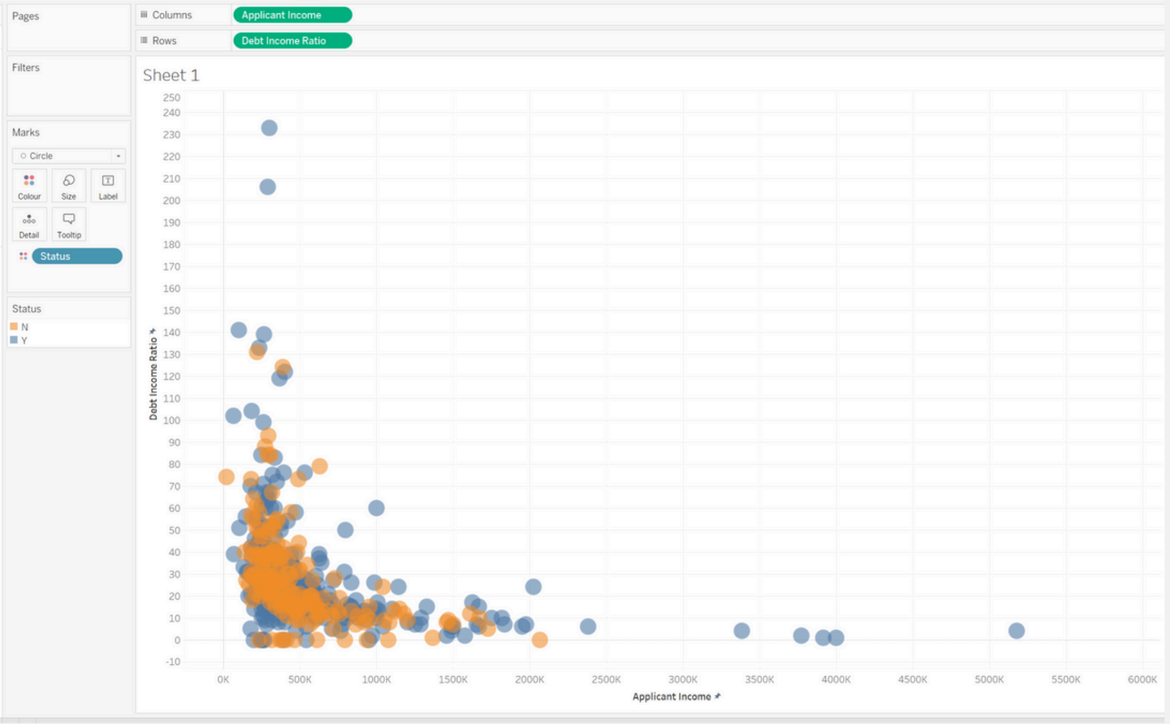


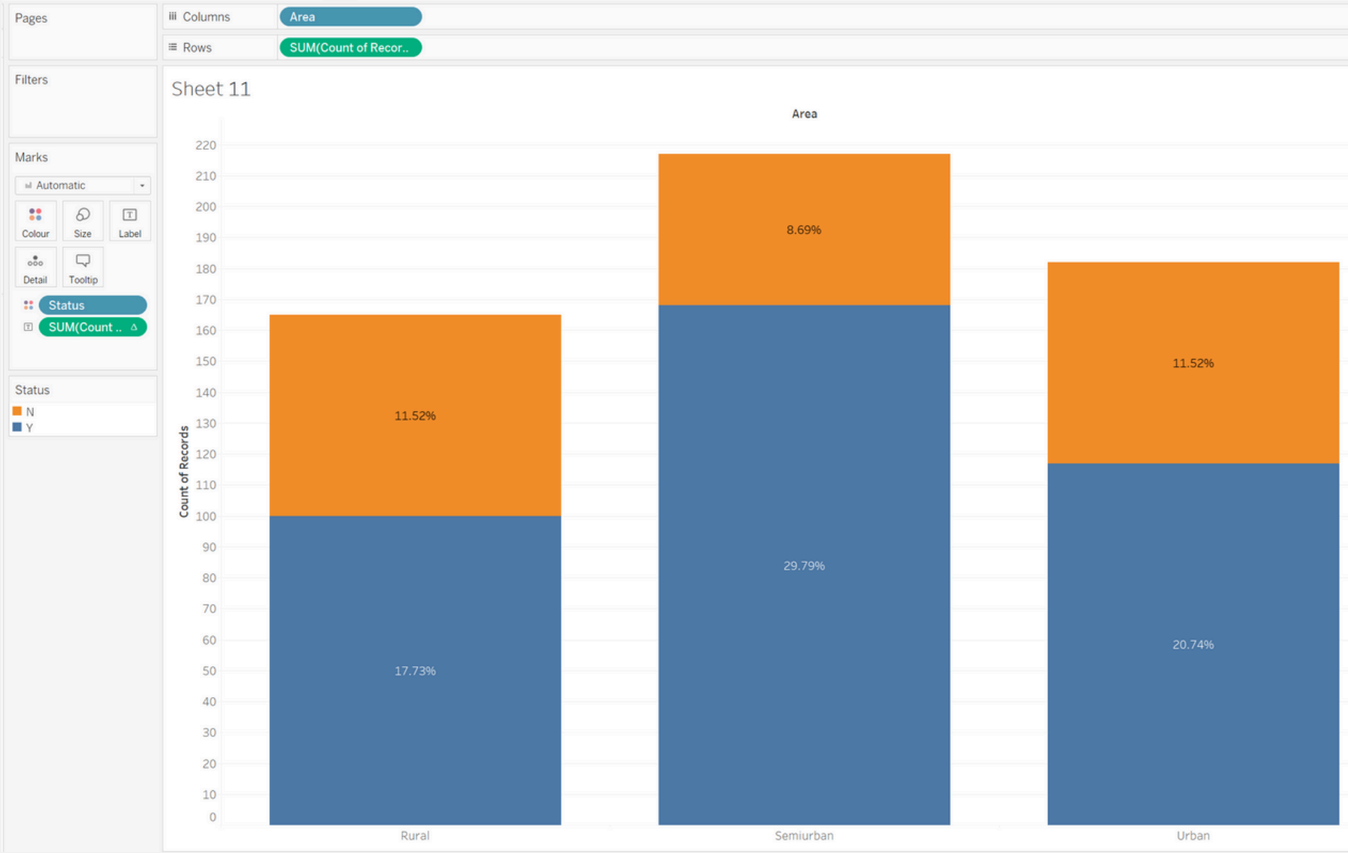
Fig 10.Applicant Income VS Debt Income Ratio

Area VS Loan Approval

Finding: SemiUrban and Urban applicants have higher loan approval rates compared to rural applicants.

Insight: Geographic location impacts loan approval, potentially due to differences in economic opportunities and financial stability.

Recommendation: Develop specialized loan products and approval criteria for rural applicants to address their unique financial situations and increase approval rates.



Recommendations

1. Prioritize Graduate Males in Semiurban Areas for Loan Approvals
Finding: Graduate males in semiurban areas who are married, not self-employed, and have no dependents show the highest loan approval rates.

Steps: Develop specific loan products and marketing strategies aimed at this demographic to leverage their high approval rates, Implement expedited approval processes for these applicants to enhance efficiency and customer satisfaction.

2. Develop Loan offers with Flexible Terms for Applicants with Dependents
Finding: Higher numbers of dependents are associated with lower positive credit history rates.

Steps: Design loan products with adjustable interest rates and repayment terms for applicants with multiple dependents to mitigate financial strain.

3. Implement a Minimum Income and Debt-Income Ratio Threshold
Finding: Higher total incomes are associated with higher loan approval rates, and managing the debt-to-income ratio is crucial for assessing an applicant's repayment capability.

Steps: Establish minimum income and DTI ratio thresholds to ensure borrowers have sufficient financial capacity to repay loans.

Possible Limitations of Project

- Credit History only indicates past behaviour of applicants when payment is due. This goes to say that first time applicants may not be considered in the positive in the dataset hence making the dataset not entirely indicative

- Credit History does not give exception to delays in payments due to any valid reason (Emergency, not in town) and hence might

Thank you!!