Loan Approval Data Analysis

—

Done by-

- Shantaprasad Kamat

Index

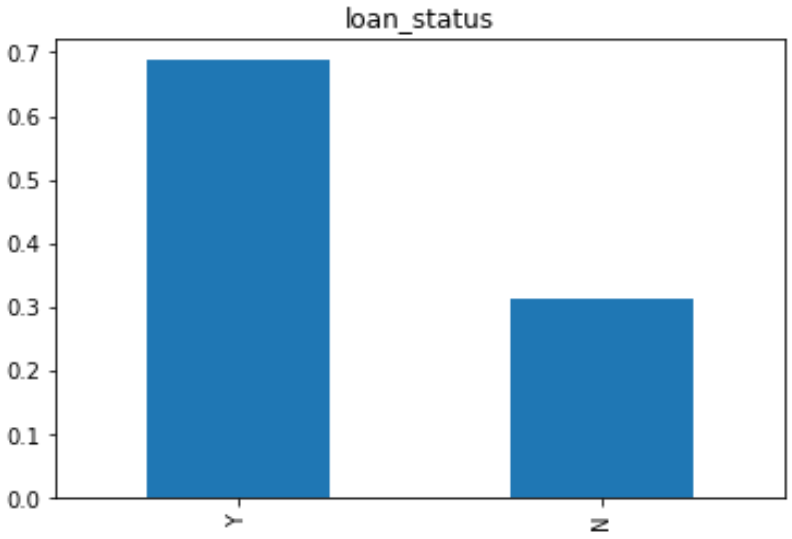
Problem Statement-

Automation has become the latest industry buzzword. There are many loan-providing companies and they receive many applications from which they have to decide whose loan they have to approve on the basis of their information. So by using this data, we tried to perform an exploratory data analysis on the data and automate this process of loan approval using a machine learning algorithm.

Data:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Loan_ID	Gender	Married	Dependen	Education	Self_Empl	ApplicantI	Coapplicar	LoanAmou	Loan_Amc	Credit_His	Property_	Loan_Status
2	LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
3	LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
4	LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
5	LP001006	Male	Yes	0	Not Gradu	No	2583	2358	120	360	1	Urban	Y
6	LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
7	LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
8	LP001013	Male	Yes	0	Not Gradu	No	2333	1516	95	360	1	Urban	Y
9	LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurbar	N
10	LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
11	LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurbar	N
12	LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
13	LP001027	Male	Yes	2	Graduate		2500	1840	109	360	1	Urban	Y
14	LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
15	LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
16	LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y

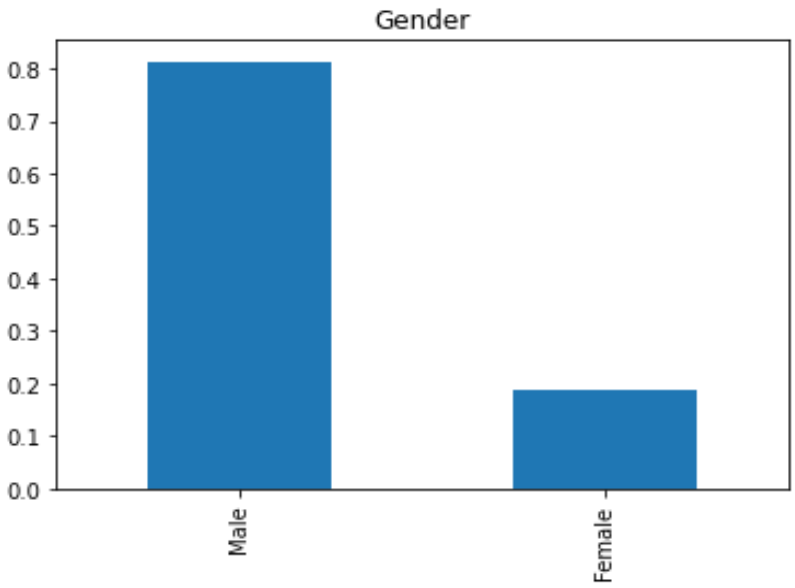
Graph 1- Loan Application Status:



The above data has been normalized and interprets that 68.73% of loans have been approved and 31.27% have not been approved out of the total 614 loan applications.

i.e Out of the recorded data (614 in total), 422 loan applications have already been approved and 192 haven't.

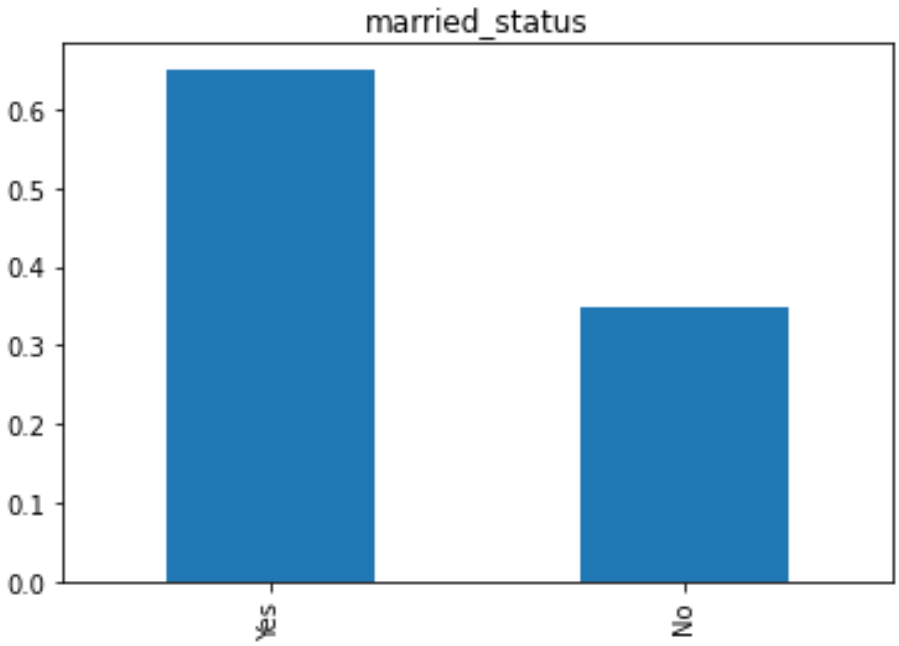
Graph 2- Ratio of male and female applicants:



The above data has been normalized and it interprets that the number of male applicants for a loan is greater than that of females, 81.36% of loan applicants are male and 18.64% are female.

i.e Out of the recorded data (601 in total), 489 are Male and 112 are Female.

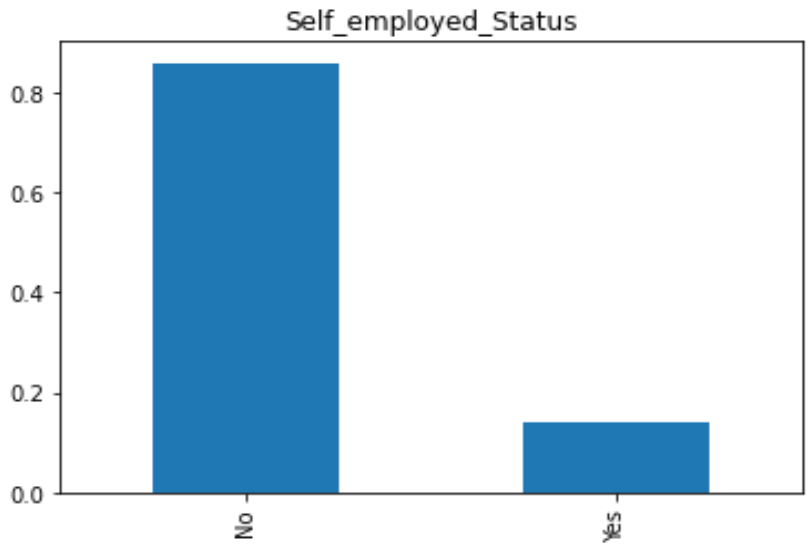
Graph 3- Ratio of married applicants:



The above data has been normalized and interprets that 65.14% of the loan applicants are married and 34.86% are not married out of the total 611 loan applications.

i.e Out of the recorded data (611 in total), 398 loan applications are married and 213 aren't.

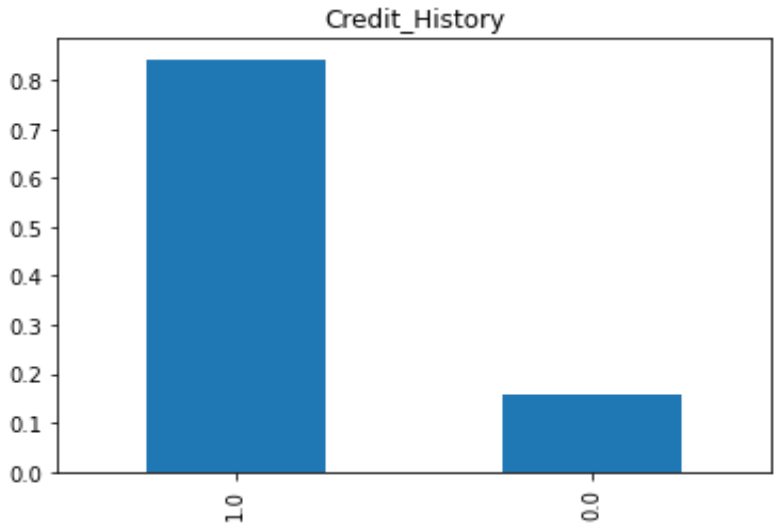
Graph 4- Ratio of self-employment status of loan applicants:



The above data has been normalized and interprets that 14.09% of the loan applicants are self-employed and 85.91% are not self-employed out of the total 582 loan applications.

i.e Out of the recorded data (582 in total), 82 loan applications are self-employed and 500 aren't.

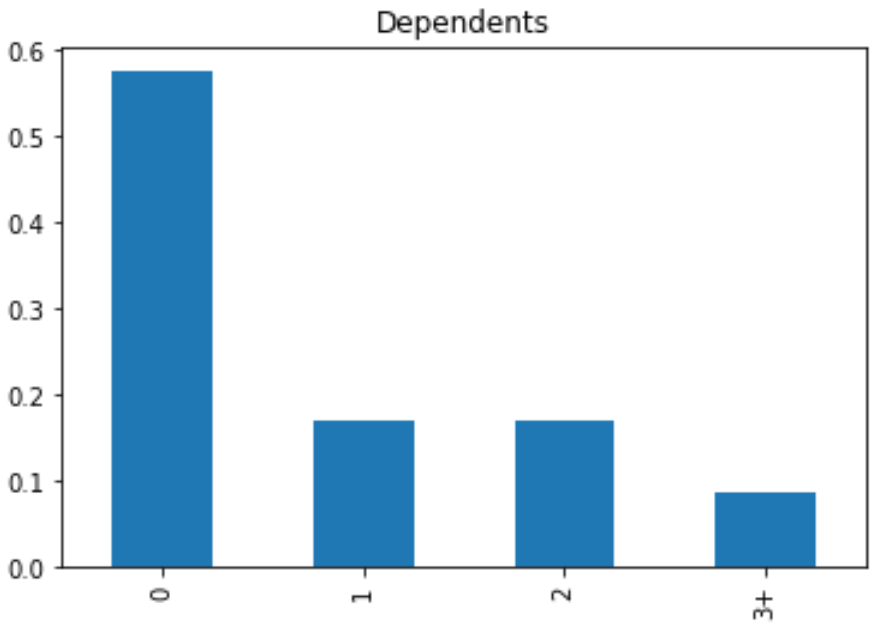
Graph 5- Ratio of credit history of loan applicants:



The above data has been normalized and interprets that 84.22% of loan applicants have a credit history and 15.78% of loan applicants do not have a credit history out of the total 564 loan applications.

I.e Out of the recorded data (564 in total), 475 loan applications have a credit history and 89 do not.

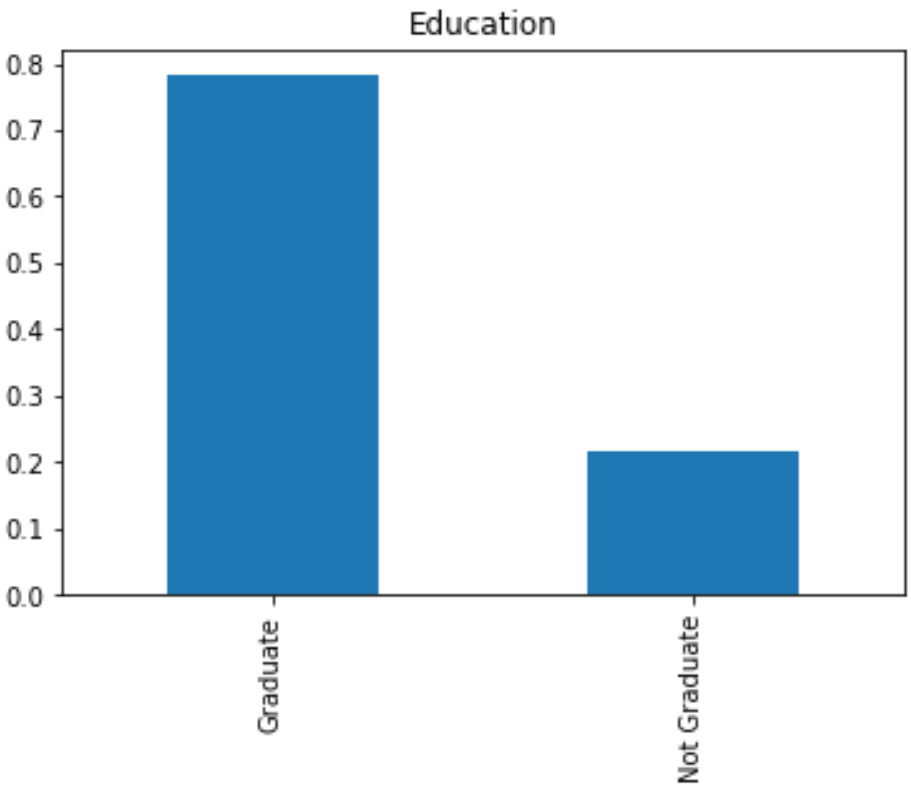
Graph 6- Ratio of dependents on the loan applicant:



The above data has been normalized and interprets that 57.60% of the loan applicants don't have any dependents, 17.03% of the loan applicants have one dependent, 16.86% of the loan applicants have two dependents and 8.51% of the loan applicants have three or more dependents out of the total 599 loan applications.

I.e Out of the recorded data (599 in total), 345 loan applications have no dependents, 102 have one, 101 have three dependents and 51 have three or more.

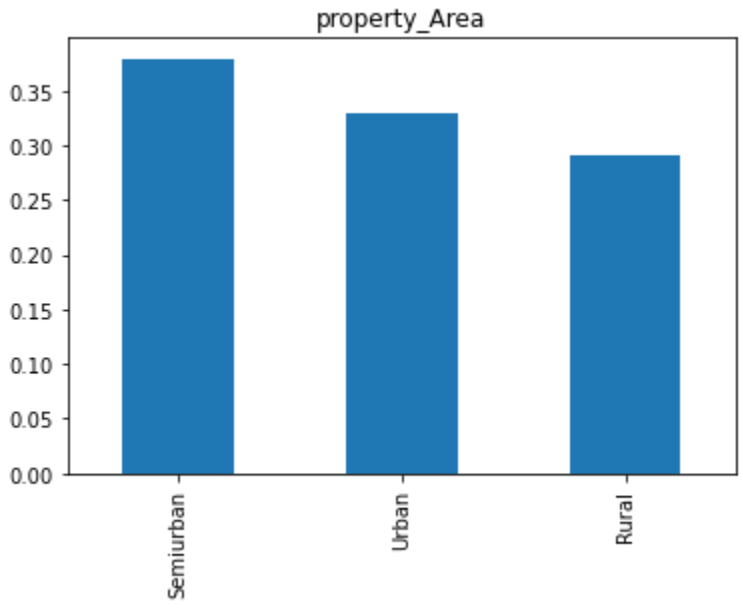
Graph 7- Ratio of graduate loan applicants:



The above data has been normalized and interprets that 78.18% of the loan applicants are graduates and 21.82% are non-graduates out of the total 614 loan applications.

I.e Out of the recorded data (614 in total), 480 loan applications are graduates and 134 aren't.

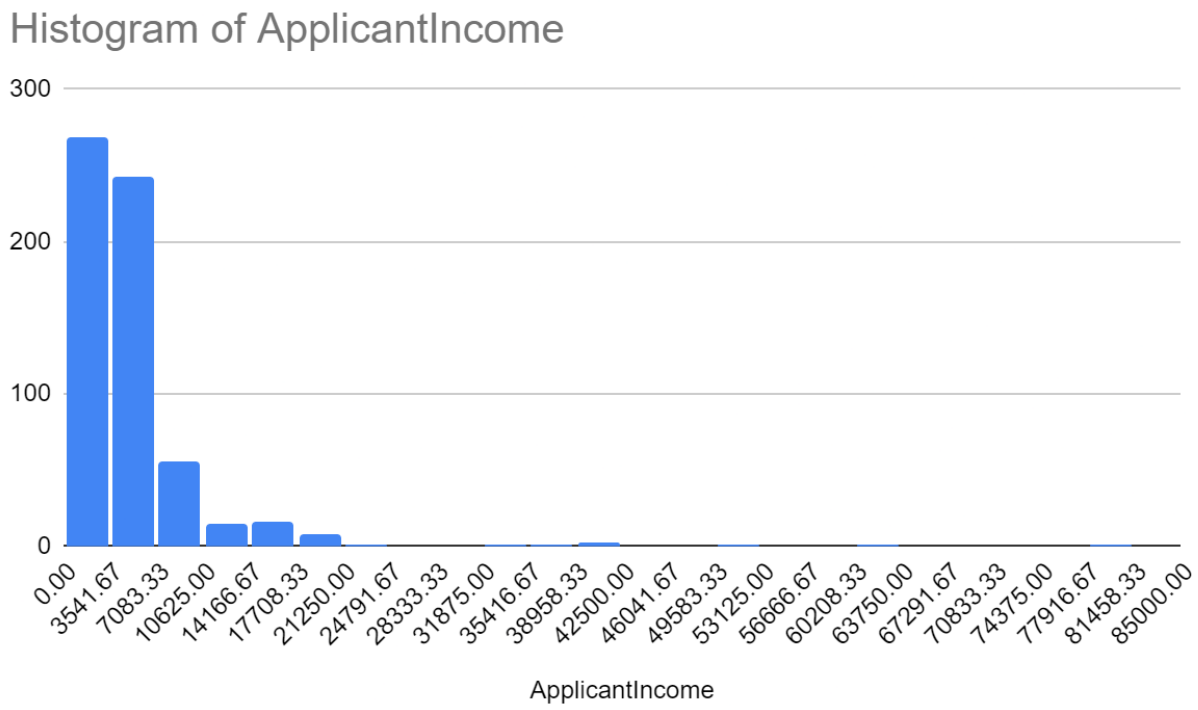
Graph 8- Ratio of property owned by loan applicants in Urban, Semi-Urban, and Rural areas:



The above data has been normalized and interprets that 37.95% of loan applicants own property in a Semi-Urban area, 32.90% of loan applicants own property in an Urban area, and 29.15% of loan applicants own property in a Rural area out of the total 614 loan applications.

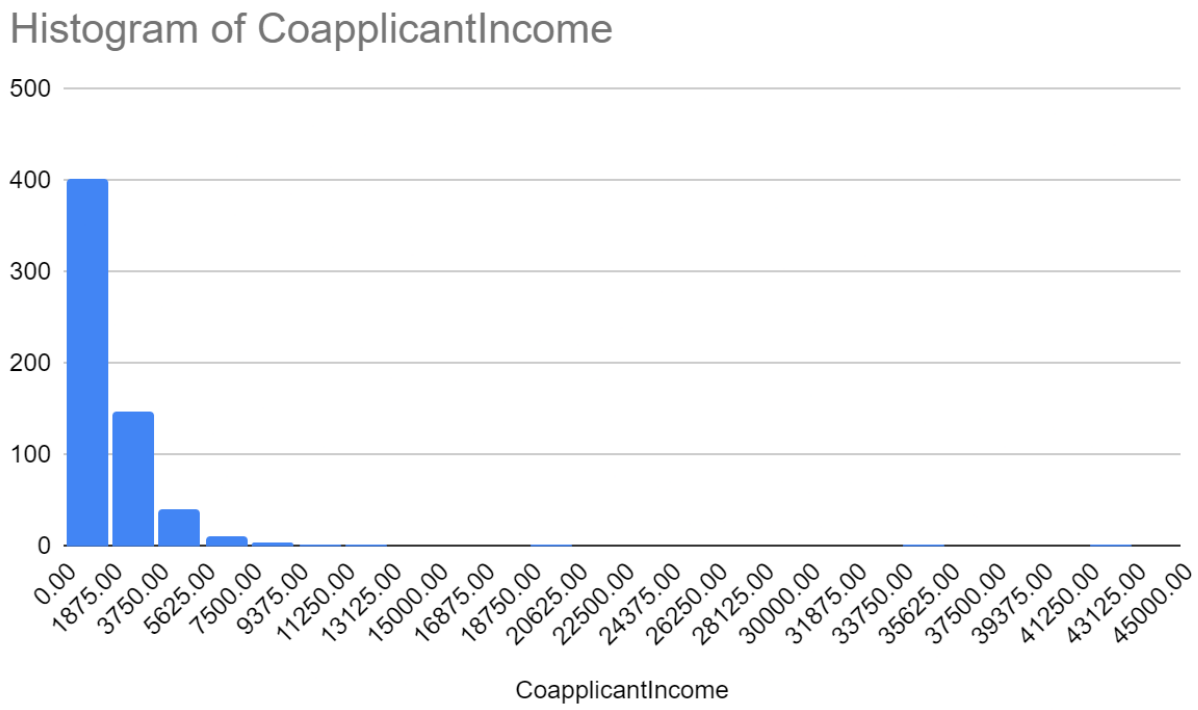
I.e Out of the recorded data (614 in total), 233 loan applicants own property in a Semi-Urban area, 202 loan applicants own property in an Urban area, and 179 loan applicants own property in a Rural area out of the total 614 loan applications.

Graph 9- Applicant Income Histogram:



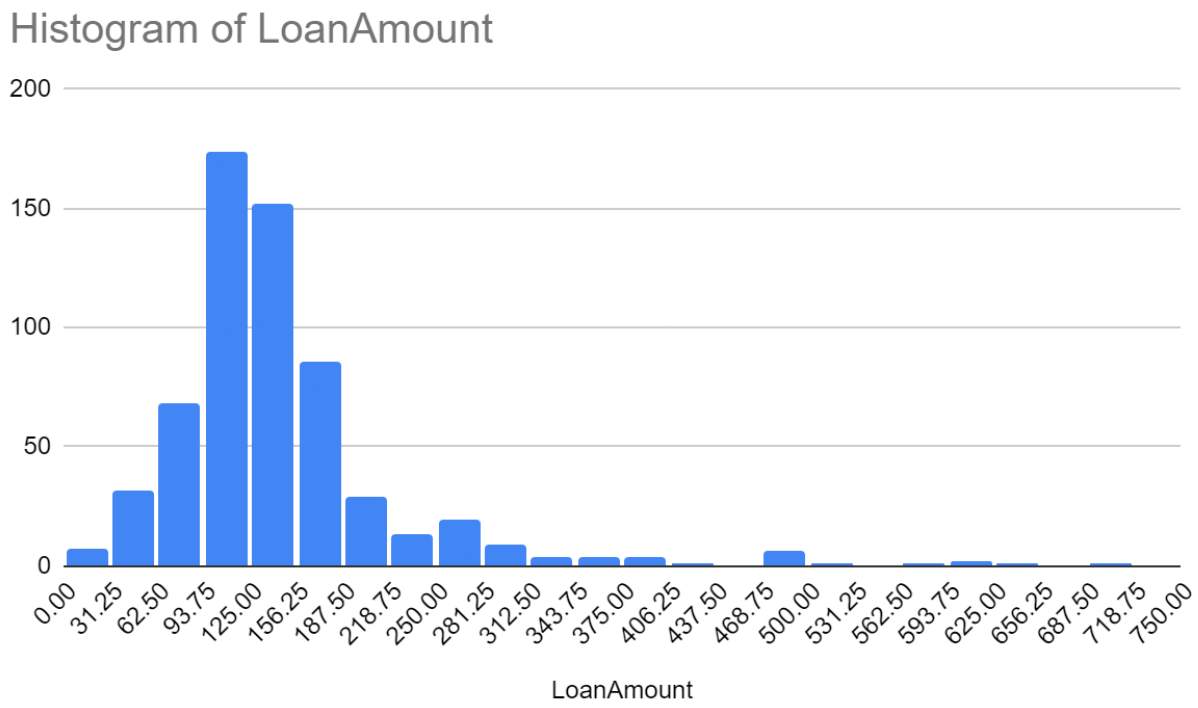
In the above graph, we can see that the Applicant Income histogram is positively skewed. The average applicant income is $5403.45 per month with skewness of +6.54 and kurtosis of 60.54. The mode applicant's income is $2,500.

Graph 10- Co-applicant Income Histogram:



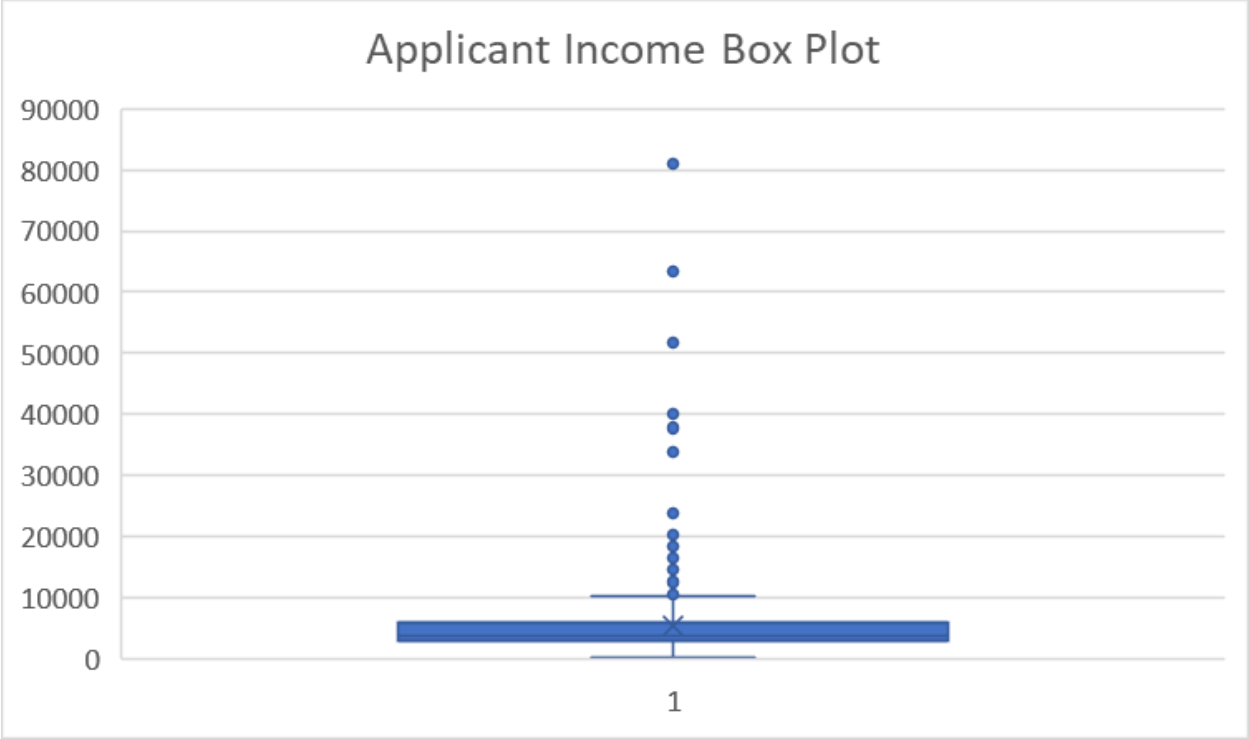
In the above graph, we can see that the Co-applicant Income histogram is positively skewed. The average co-applicant income is $1,621.25 per month with skewness of +7.49 and kurtosis of 84.96. The mode applicant income is $0 (i.e. In most cases the co-applicant is not employed).
The co-applicant income is higher than the applicant's income in 72 out of 614 cases.

Graph 11- Loan Amount Histogram:



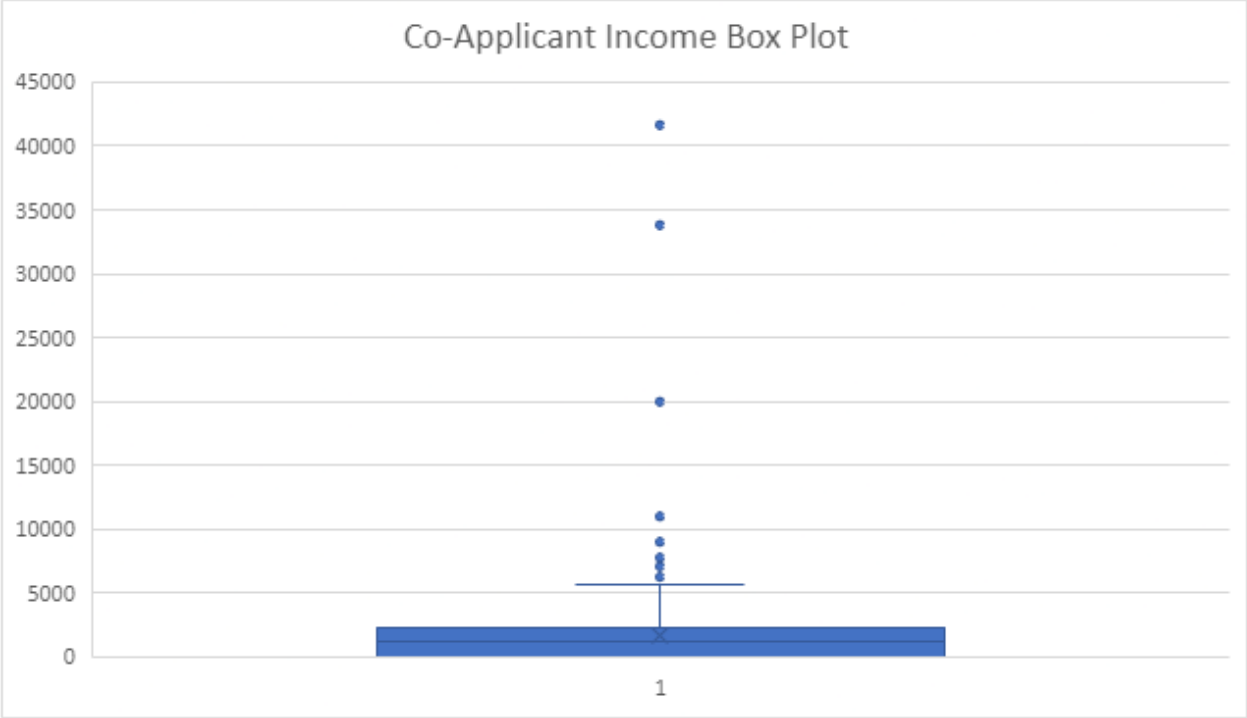
In the above graph, we can see that the loan amount histogram is positively skewed. The maximum loan amount is $700,000 and the minimum loan amount is $9,000. The average loan amount is $146,410 with skewness of +2.73 and kurtosis of 10.90. The mode loan amount is Rs.1,20,000.

Graph 12- Applicant Income Box Plot:



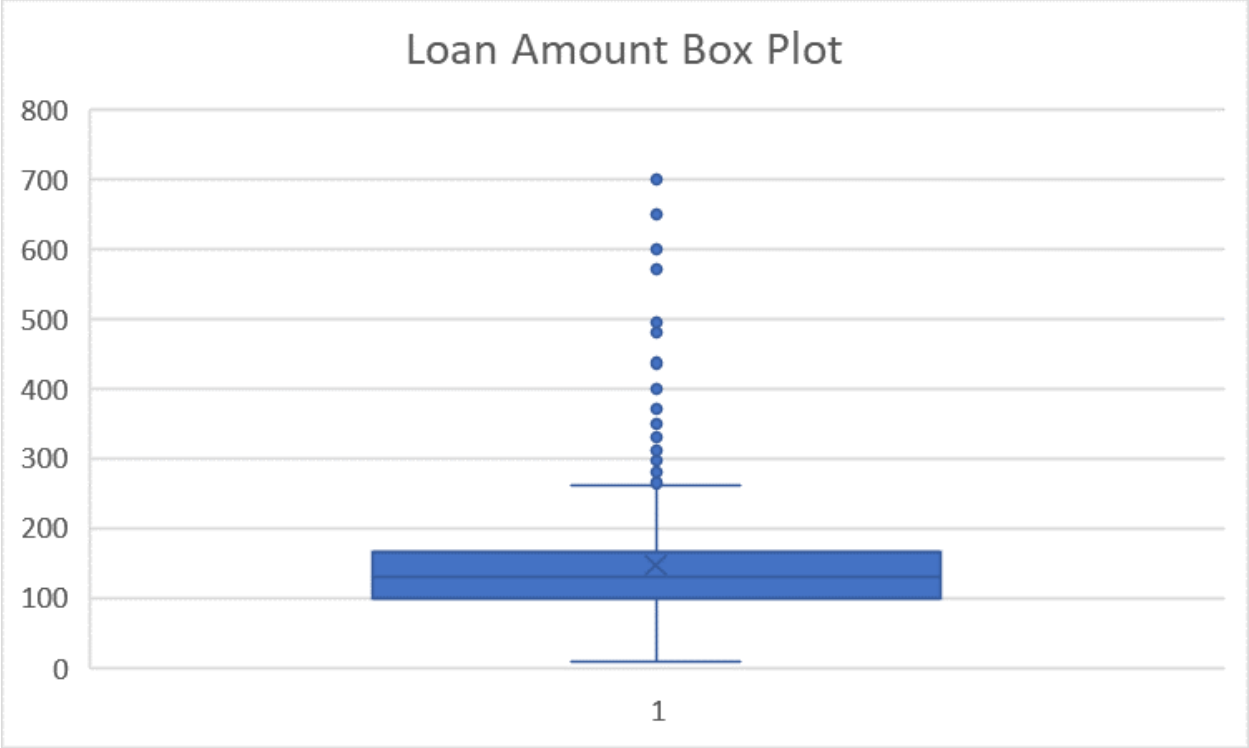
We can see from this box plot that the maximum applicant income is $81,000 per month and the minimum applicant income is $150 per month. The median applicant income is $3,812.5 per month. The first quartile is $2,877.5 per month and the third quartile is $5,795 per month. IQR is $2,917.5 per month. The low outliers lie below -$1,498.75 per month (i.e. no low outliers) and high outliers lie above $10,171.25 per month (50 high outliers).

Graph 13- Co-Applicant Income Box Plot:



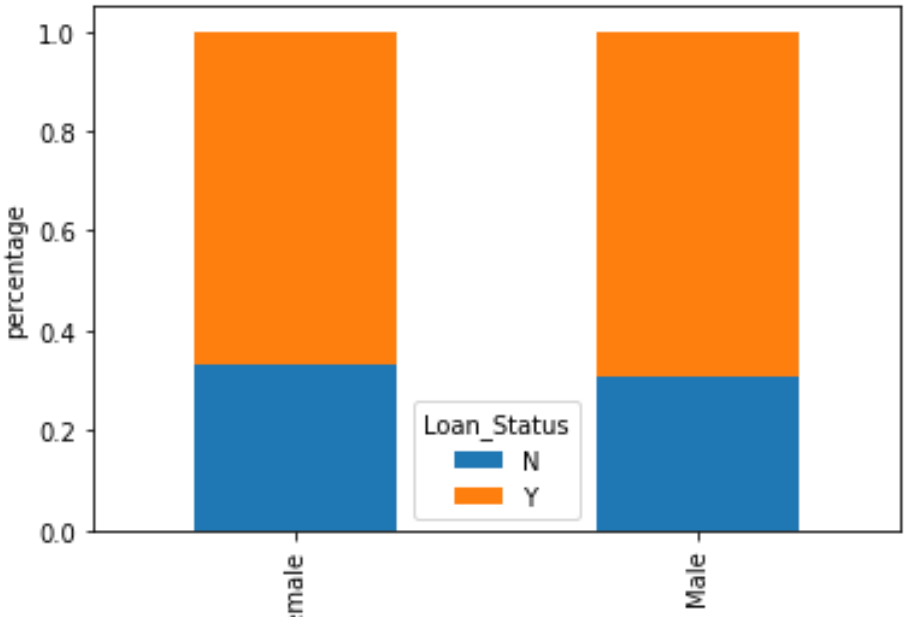
We can see from this box plot that the maximum co-applicant income is $41,667 per month and the minimum co-applicant income is $0 per month. The median co-applicant income is $1,188,5 per month. The first quartile is $0 per month and the third quartile is $2,297.25 per month. IQR is $2,297.25 per month. The low outliers lie below -$3445.875 per month (i.e. no low outliers) and high outliers lie above $5743.125 per month (18 high outliers).

Graph 14- Loan Amount Box Plot:



We can see from this box plot that the maximum loan amount is $700,000 and the minimum loan amount is $9,000. The median loan amount is $129,000. The first quartile is $100,250 and the third quartile is $164.75. IQR is $64.5. The low outliers lie below $3.5 (i.e. no low outliers) and high outliers lie above $261.5 (i.e. 41 high outliers).

Graph 15- Graph of Proportion of loan approval for males and females:

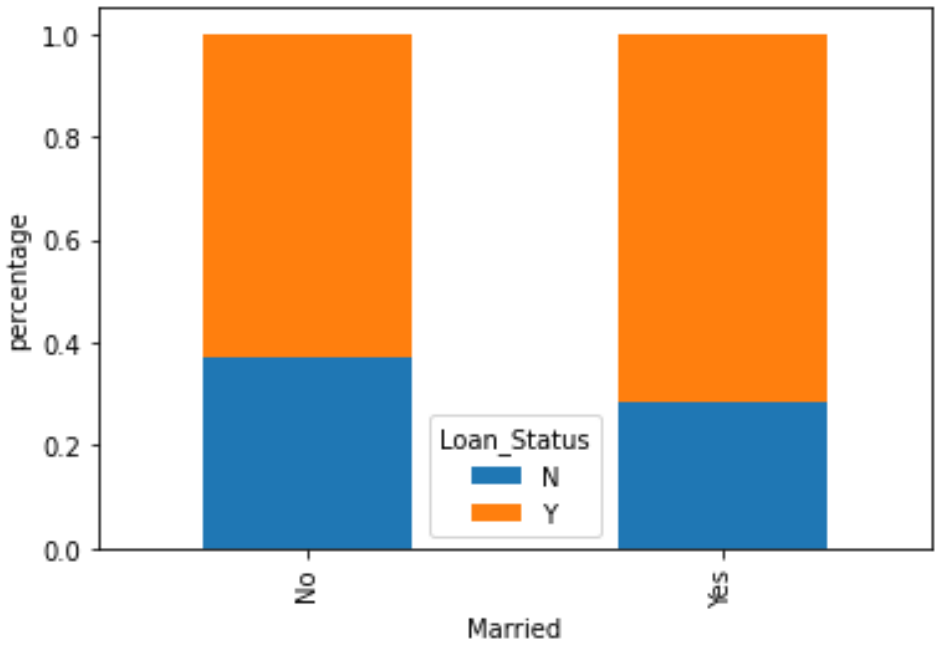


The graph above indicates that about 69.32% of applications of male applicants are approved and 66.96% of applications of female applicants are approved.

i.e 339 out of 489 total applications of male applicants are approved and 75 out of 112 applications of female applicants are approved.

The p-value of the two-tailed z-test (where H_0 is $p_1 = p_2$ and H_1 is $p_1 \neq p_2$) is 0.624 and as it is greater than 0.05, we can conclude that the percentage of approval for male and female applications is nearly equal.

Graph 16- Graph of Proportion of loan approval based on their married status:

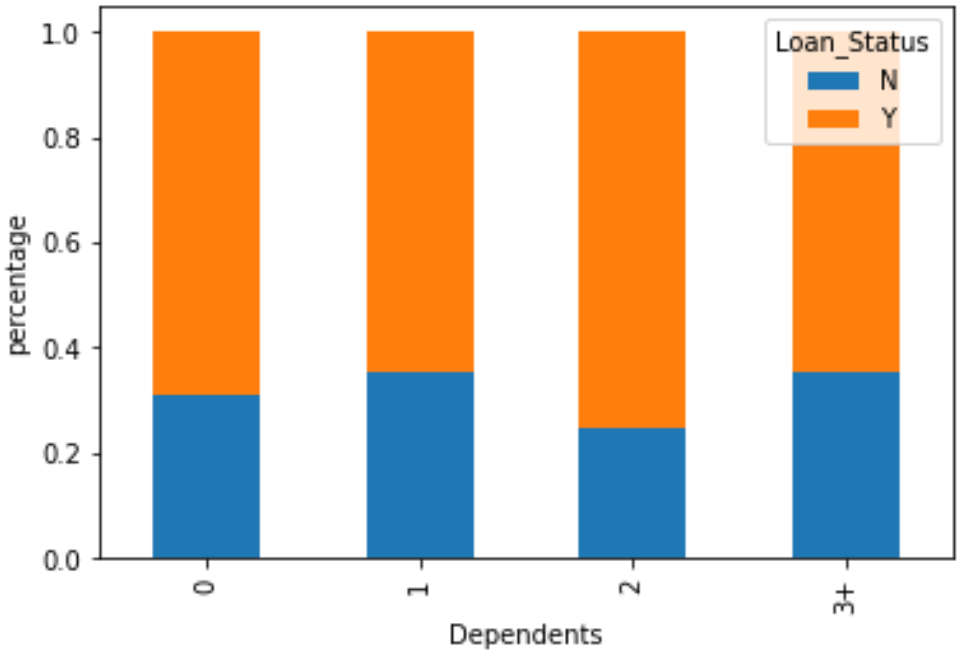


From the above graph we can see that about 71.60% of applications of married applicants are approved while 62.91% of applications of unmarried applicants are approved.

i.e 285 out of 398 applications of married applicants are approved and 134 out of 213 applications of unmarried applicants are approved.

The p-value of the one-tailed z-test is 0.0139 (where H_0 is $p_1 = p_2$ and H_1 is $p_1 \neq p_2$) and as it is smaller than 0.05, there is a significantly higher probability of loan approval for married applicants than unmarried applicants.

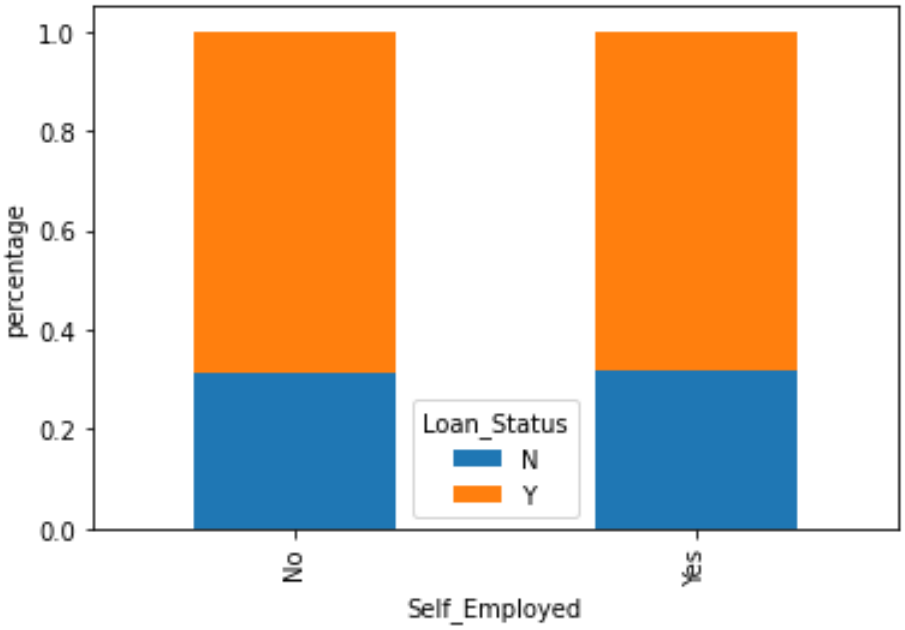
Graph 17: Graph of Proportion of loan approval based on the number of dependents:



From the above graph we can see that about 68.99% of applications of applicants not having any dependents are approved, 64.71% of applications of applicants having 1 dependent are approved, 75.25% applications of applicants having 2 dependents are approved, 64.71% applications of applicants having 3 dependents are approved.

i.e 238 out of 345 applications of applicants with zero dependents are approved, 66 out of 102 applications of applicants with 1 dependent are approved while 76 out of 101 applications of applicants with 2 dependents are approved and 33 out of 51 applications of applicants with zero dependents are approved.
As can be seen in the graph, loan approval is independent of the number of dependents.

Graph 18: Graph about the Proportion of loan approval based on whether they are self-employed or not:



From the above graph we can see that about 68.6% of applications of applicants that are not self-employed are approved while 68.29% of applications of self-employed applicants are approved.

i.e 343 out of 500 applications of applicants that are not self-employed are approved and 56 out of 82 applications of applicants that are self-employed are approved.

The p-value of the two-tailed z-test is 0.95 (where H_0 is $p_1 = p_2$ and H_1 is $p_1 \neq p_2$) and as it is greater than 0.05, we can conclude that the percentage of approval for self-employed and not self-employed applicants is the same.

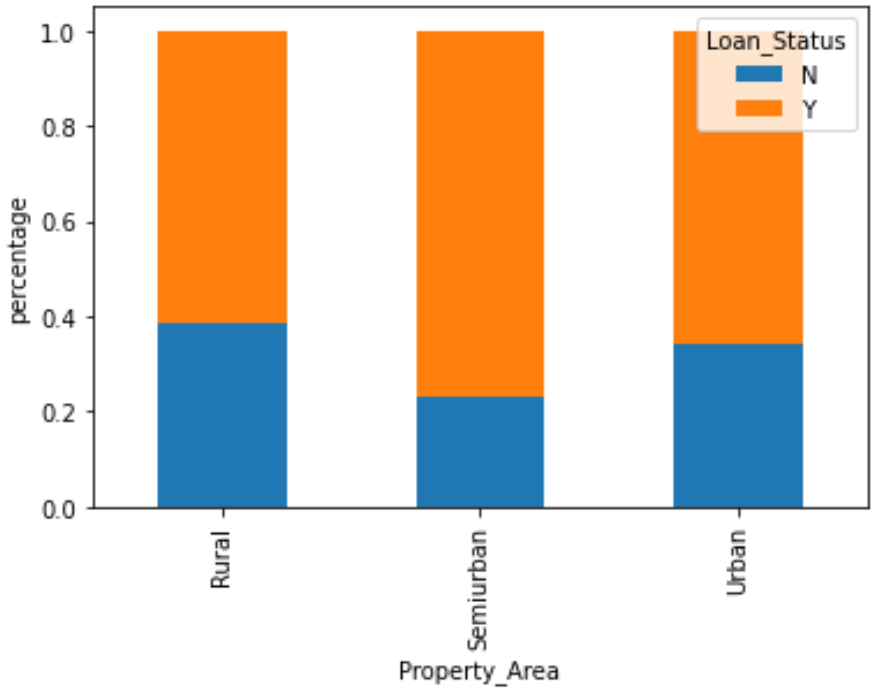
Graph 19: Graph about the Proportion of loan approval based on credit history:

From the above graph we can see that about 79.57% of applications from applicants with a credit history are approved while 7.86% of applications from applicants with no credit history are approved.

i.e 378 out of 475 applications of applicants with a credit history are approved and 7 out of 89 applications of applicants with no credit history are approved.

The p-value of the one-tailed z-test (where H_0 is $p_1 = p_2$ and H_1 is $p_1 > p_2$) is 0.00001 and as it is smaller than 0.05, there is a significantly higher probability of loan approval for applicants with credit history than those with no credit history.

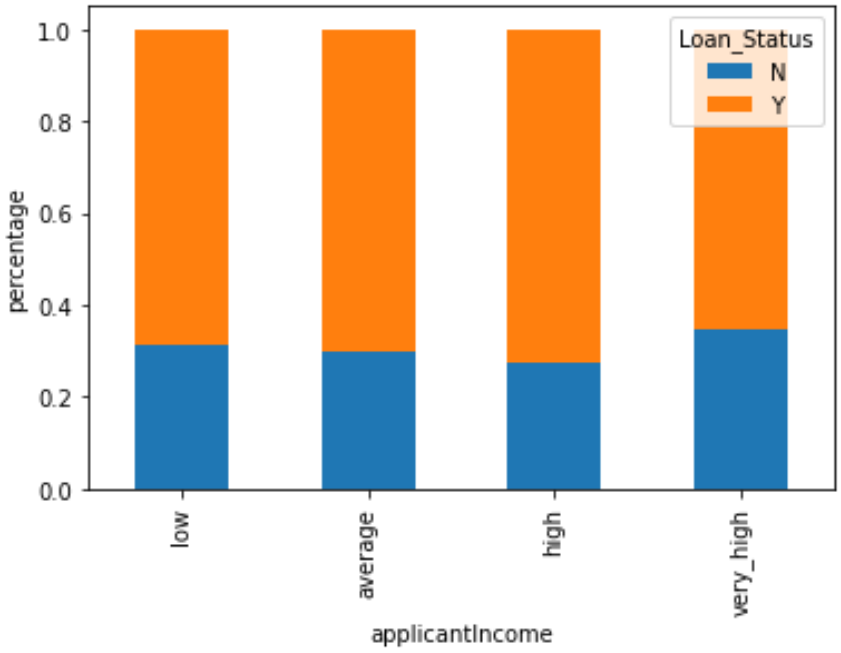
Graph 20: Graph about the proportion of loan approval based on the locality:



From the above graph we can see that about 61.45% of applications of applicants with property in rural areas are approved, 76.82% of applications of applicants with property in semi-urban areas are approved, 65.84% applications of applicants with property in urban areas are approved.

i.e 110 out of 179 applications of applicants with property in rural areas are approved, 179 out of 233 applications of applicants with property in semi-urban areas are approved, and 133 out of 202 applications of applicants with property in urban areas are approved.
The p-value of the one-tailed z-test (where H_0 is $p_1 = p_2$ and H_1 is $p_1 > p_2$) is 0.00036 and as it is smaller than 0.05, there is a significantly higher probability of loan approval for applicants with property in semi-urban areas than in the rest.

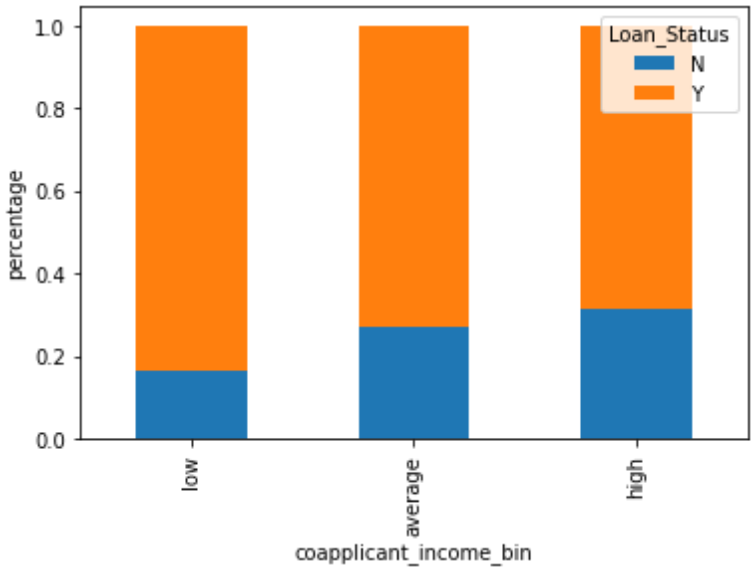
Graph 21: Graph about the proportion of loan approval based on the type of Income Category:



From the above graph we can see that about 68.67% of applications of applicants with low income are approved, 70,33% of applications of applicants with average income are approved, 72.36% of applications of applicants with high income are approved, 65.17% of applications of applicants with very high income are approved.

i.e 57 out of 83 applications of applicants with low income are approved, 147 out of 209 applications of applicants with average income are approved, 55 out of 76 applications of applicants with high income are approved, and 73 out of 112 applications of applicants with very high income are approved.
As can be seen in the graph, loan approval is independent of the applicant's income.

Graph 22: Graph about the proportion of loan approval based on the bins we created for Co-Applicant Income:

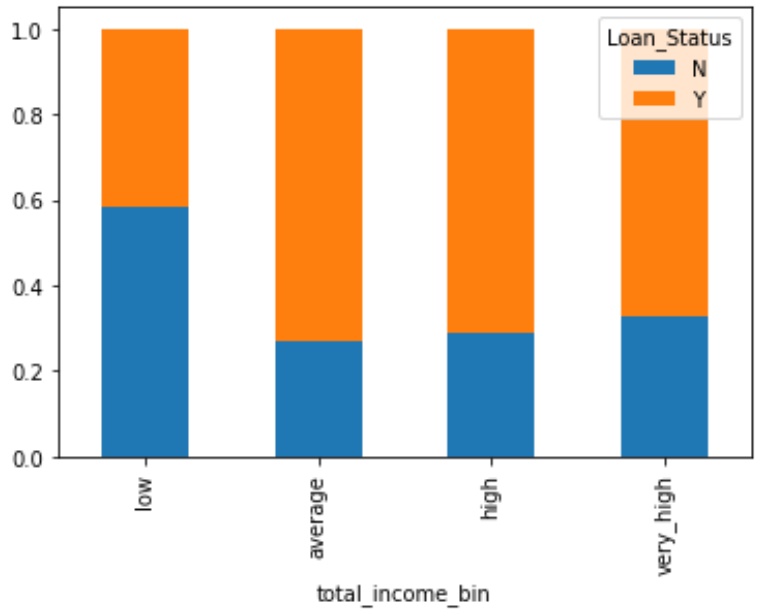


From the above graph we can see that about 83.33 % of applications of applicants with low co-applicant income are approved, 72.78 % of applications of applicants with an average co-applicant income are approved, 68.83 % applications of applicants with high co-applicant income are approved.

i.e 15 out of 18 applications of low co-applicant income are approved, 123 out of 169 applications of average co-applicant income are approved, and 53 out of 77 applications of high co-applicant income are approved.

The p-value of the two-tailed z-test (where H_0 is $p_1 = p_2$ and H_1 is $p_1 \neq p_2$) is 0.28014 and as it is greater than 0.05, there is no significant difference between the percentages of approval of applicants from low co-applicant income bin and that of applicants from average and high co-applicant income bins combined.

Graph 23: Graph about the proportion of loan approval based on the bins we created for Total Income:

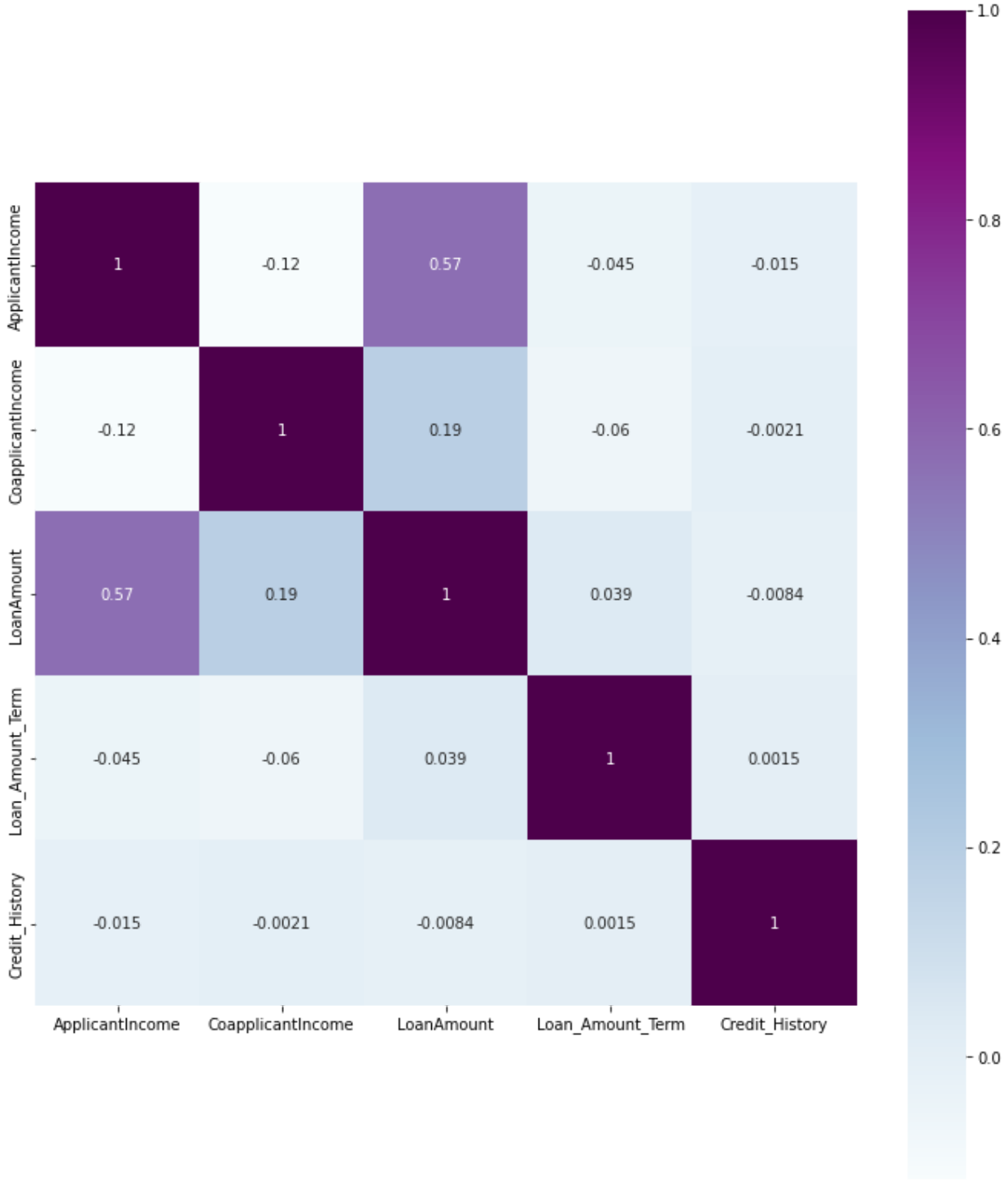


From the above graph we can see that about 41.66% of applications of applicants with low total income are approved, 73.10% of applications of applicants with average total income are approved, 70.98% applications of applicants with high total income are approved, 67.20% applications of applicants with very high total income are approved.

i.e 10 out of 24 applications of applicants with low total income are approved, 87 out of 119 applications of applicants with average total income are approved, 159 out of 224 applications of applicants with high total income are approved, 166 out of 247 applications of applicants with very high total income are approved.
The p-value of the one-tailed z-test(where H_0 is $p_1 = p_2$ and H_1 is $p_1 > p_2$) is 0.00175 and as it is smaller than 0.05, there is a significantly lower probability of loan approval for applicants from the low total income bin than the rest.

Heatmap of the variables:

This is the correlation matrix of the variables.



Applicant Income has a moderate positive correlation (0.57) with the loan amount. The rest have negligible correlations with others.

Using logistic regression to predict whether the loan would be approved or not:

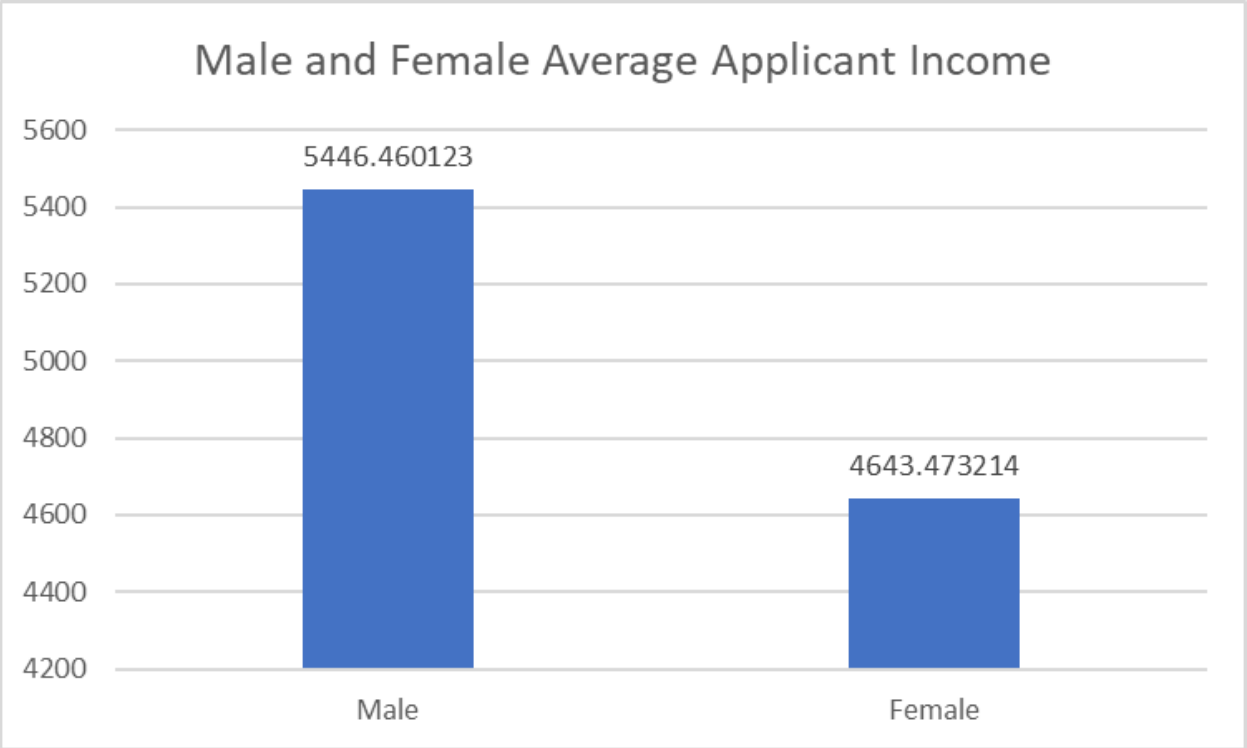
```
LR = LogisticRegression()
LR.fit(X_train, y_train)

y_predict = LR.predict(X_test)

#  prediction Summary by species
print(classification_report(y_test, y_predict))

# Accuracy score
LR_SC = accuracy_score(y_predict,y_test)
print('accuracy is',accuracy_score(y_predict,y_test))
```

```
accuracy is 0.9224324324324325
```

We have used logistic regression from the sklearn library to predict whether the loan will be approved or not based on the variables available in our dataset and we have got an accuracy of 92%.
(92% accuracy means that 92 out of 100 times we are able to correctly predict the outcome from the test dataset)
Test dataset: A test dataset is not used for training purposes. It is only used to know whether our model is performing well or not.

Accuracy = number of correct predictions / total number of predictions

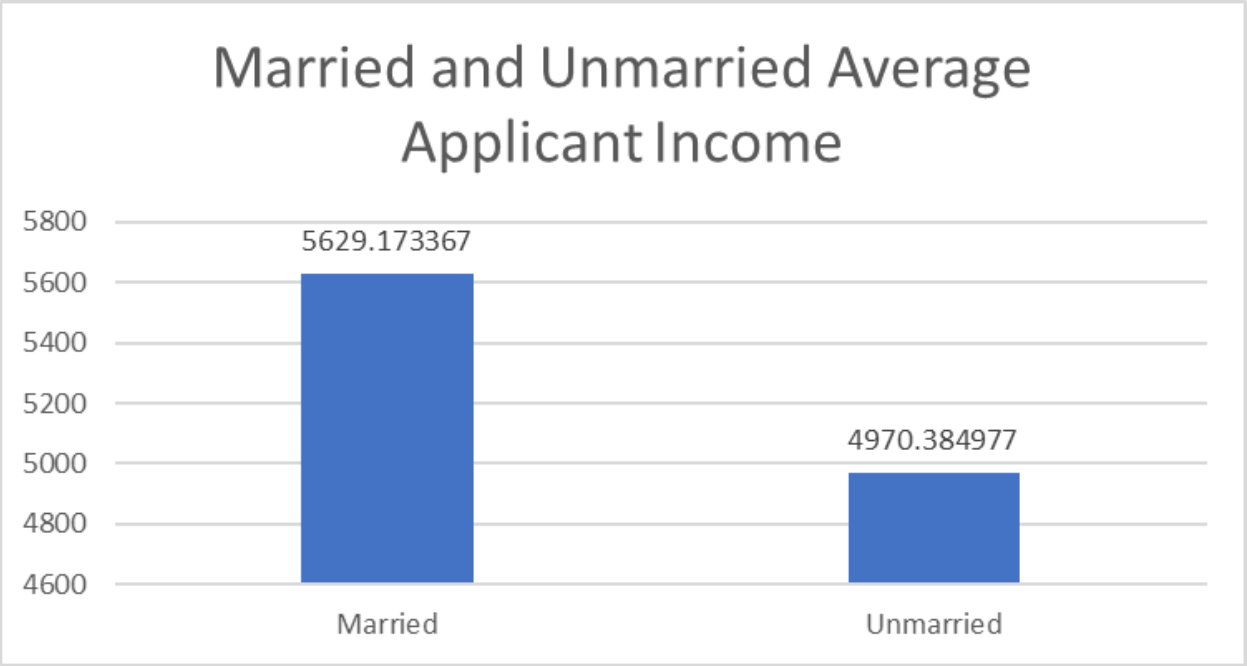
Graph 24- Average Income of Male and Female Applicants:



In the above graph, we can see that the average male and female applicant income is $5446.46 per month and $4643.47 per month respectively.

The p-value of the two-tailed paired t-test (where H_0 is $\mu_1 = \mu_2$ and H_1 is $\mu_1 \neq \mu_2$) is 0.1857 and as it is greater than 0.05, we can conclude that the male applicant's income and the female applicant's income are the same.

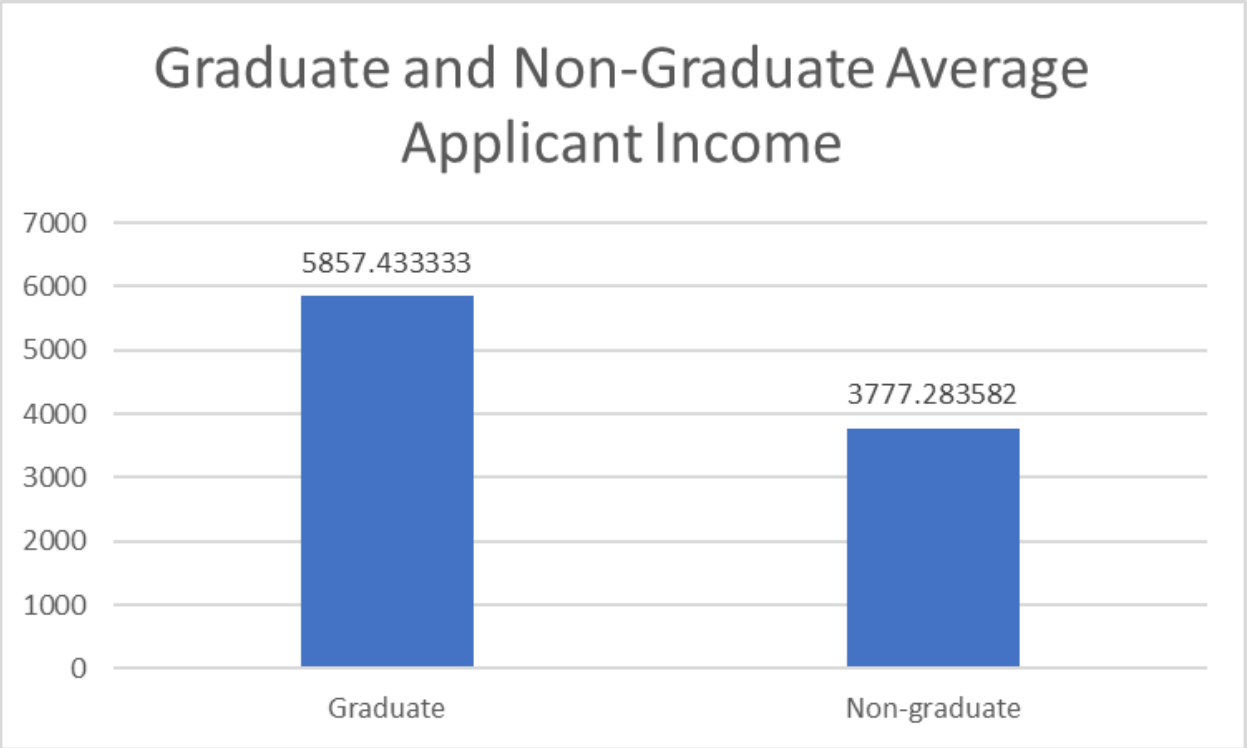
Graph 25- Average Income of Married and Unmarried Applicants:



In the above graph, we can see that the average married and unmarried applicant income is $5629.17 per month and $4970.38 per month respectively.

The p-value of the two-tailed paired t-test (where H_0 is $\mu_1 = \mu_2$ and H_1 is $\mu_1 \neq \mu_2$) is 0.2045 and as it is greater than 0.05, we can conclude that the married applicant's income and the unmarried applicant's income are the same.

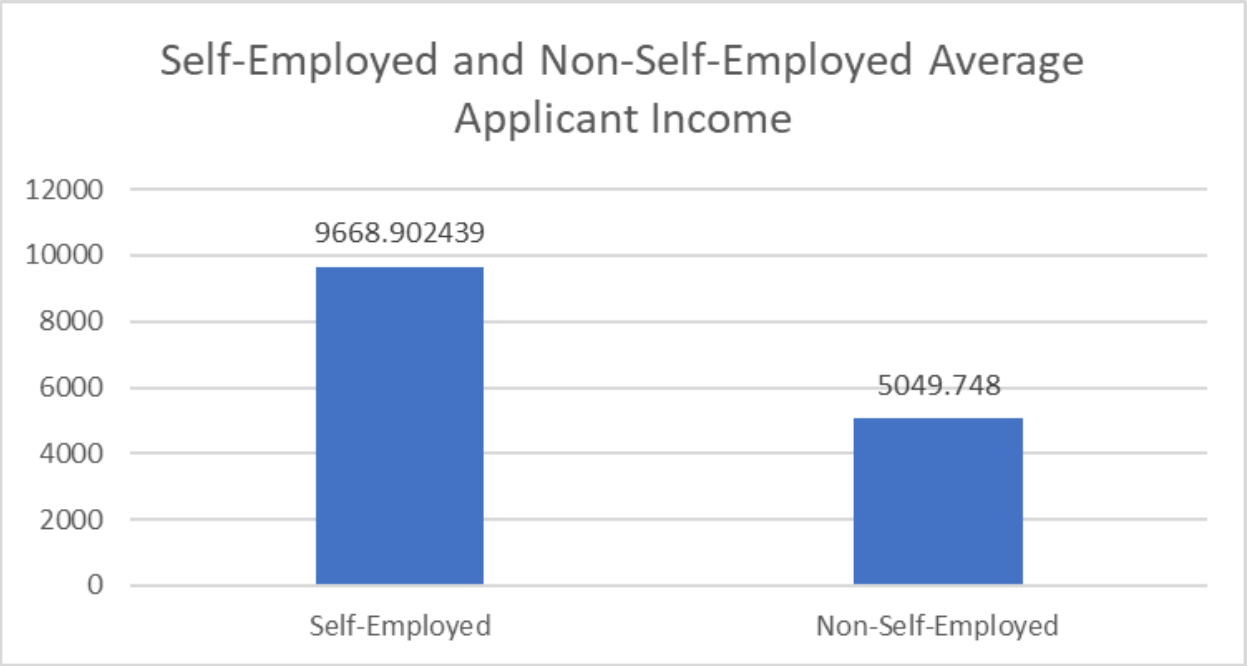
Graph 26- Average Income of Graduate and Non-graduate Applicants:



In the above graph, we can see that the average graduate and non-graduate applicant income is $5857.43 per month and $3777.28 per month respectively.

The p-value of the one-tailed paired t-test (where H_0 is $\mu_1 = \mu_2$ and H_1 is $\mu_1 > \mu_2$) is 0.0005 and as it is smaller than 0.05, we can conclude that the graduate applicant's income is higher than the non-graduate applicant's income.

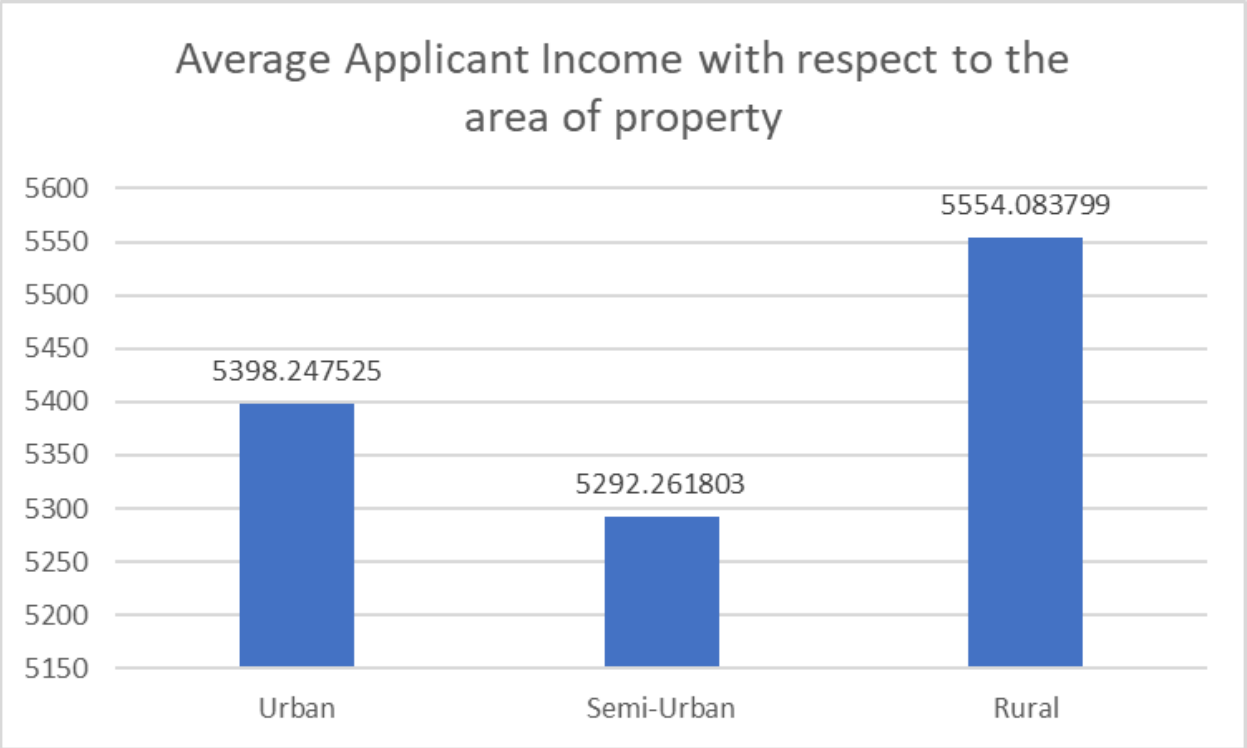
Graph 27- Average Income of Self-Employed and Non-Self-Employed Applicants:



In the above graph, we can see that the average self-employed and non-self-employed applicant income is $9668.90 per month and $5049.75 per month respectively.

The p-value of the one-tailed paired t-test (where H_0 is $\mu_1 = \mu_2$ and H_1 is $\mu_1 > \mu_2$) is 0.0001 and as it is smaller than 0.05, we can conclude that the self-employed applicant's income is higher than the non-self-employed applicant's income.

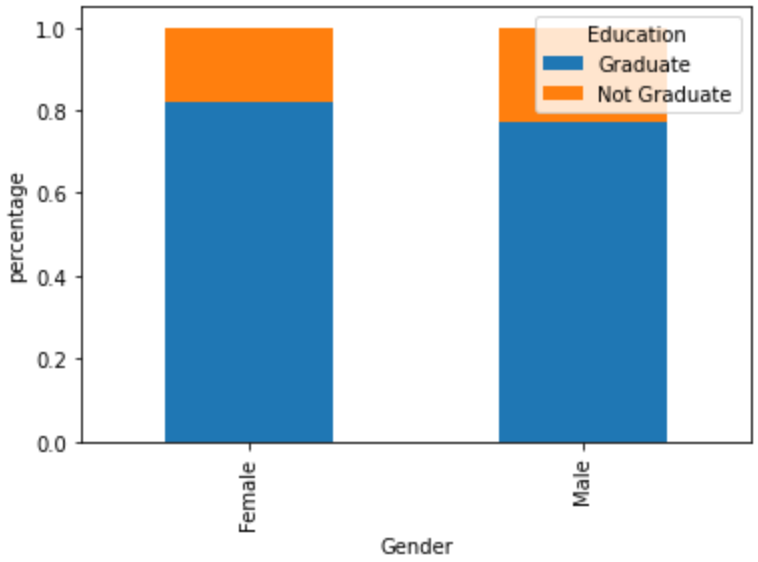
Graph 28- Average Income of Applicants with Property in Urban, Semi-Urban, and Rural areas:



In the above graph, we can see that the average self-employed and non-self-employed applicant income is $9668.90 per month and $5049.75 per month respectively.

As all the three p-values of the two-tailed paired t-tests (where H_0 is $\mu_1 = \mu_2$ and H_1 is $\mu_1 \neq \mu_2$) are greater than 0.05, we can conclude that the applicant's income is independent of the applicant's area of the property.

Graph 29: Graph about the proportion of graduates based on the gender of the applicant:

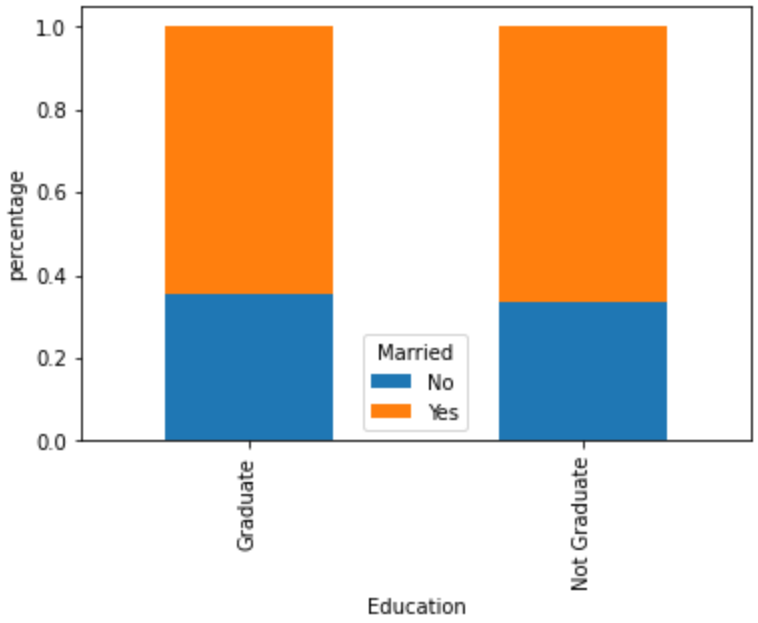


From the above graph, we can see that about 82.14% of female applicants are graduates and 76.89% of male applicants are graduates.

i.e 92 out of 112 female applicants are graduates and 376 out of 489 male applicants are graduates.

The p-value of the two-tailed z-test (where H_0 is $p_1 = p_2$ and H_1 is $p_1 \neq p_2$) is 0.2263 and as it is greater than 0.05, there is no significant difference between the male graduate ratio and female graduate ratio.

Graph 30: Graph about the proportion of married applicants based on the education of the applicant:

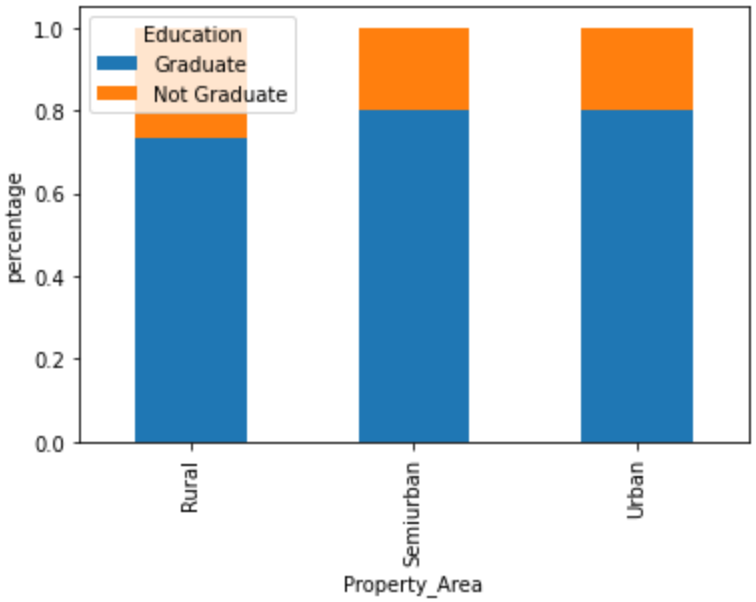


From the above graph, we can see that about 69.13% of graduate applicants are married and 66.42% of non-graduate applicants are married.

i.e 309 out of 447 graduate applicants are married and 89 out of 134 non-graduate applicants are married.

The p-value of the two-tailed z-test (where H_0 is $p_1 = p_2$ and H_1 is $p_1 \neq p_2$) is 0.5552 and as it is greater than 0.05, there is no significant difference between the graduate marriage ratio and non-graduate marriage ratio.

Graph 31: Graph about the proportion of graduates based on the area of property of the applicant:



From the above graph, we can see that about 73.18% of applicants with property in a rural area are graduates, 80.26% of applicants with property in a semi-urban area are graduates, and 80.20% of applicants with property in an urban area are graduates.

i.e 131 out of 179 applicants with property in a rural area are graduates, 187 out of 233 applicants with property in a semi-urban area are graduates, and 162 out of 202 applicants with property in an urban area are graduates.

The p-values of all three two-tailed z-tests (where H_0 is $p_1 = p_2$ and H_1 is $p_1 \neq p_2$) are greater than 0.05, so the graduate ratio is independent of the area of property of the applicant.

Applications:

The loan to be paid never goes down and always keeps on increasing. Our aim was to predict whether the company should approve the loan to a person or not. It is a quite difficult task to analyze all the data of customers manually and decide if their loan should be approved. So, by performing an intense data analysis on the previous record, we can apply the machine learning algorithm and that will give the result of their loan approval. This will help the company to minimize the time of loan approval.

Our main objective of making the process of loan approval simple and easy is complete and this will reduce the company's manpower and also make it easy for the customer.
Here we target all the banks and companies that give loans to customers.

Softwares Used for EDA:

- Google Colab
- Kaggle
- Excel
- Google Sheets