Exploratory Data Analysis and Hypothesis Testing

**Problem Statement:**

**About Company**

Dream Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan.

**Problem**

Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers.

**Introduction:**

The objective of this project is to do “Exploratory Analysis” and “Hypothesis Testing” on the features of this dataset to find various insights as to how each feature affects the chances of getting a loan.

**Dataset:**

The Loan Prediction dataset consists of 613 customer details and 13 features.

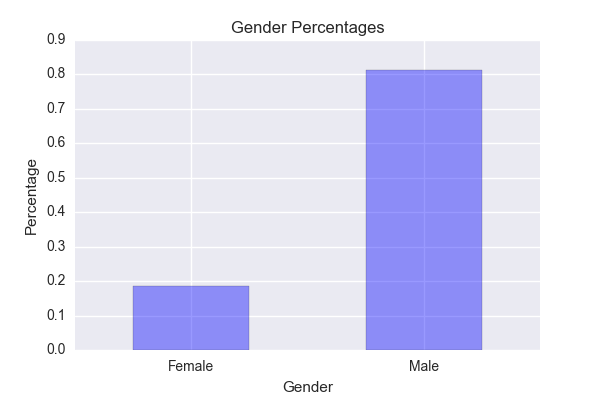
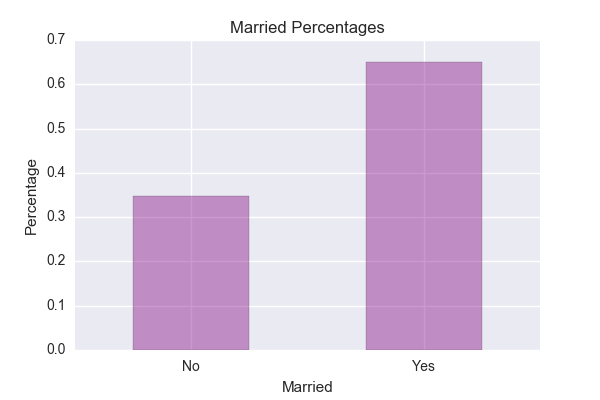
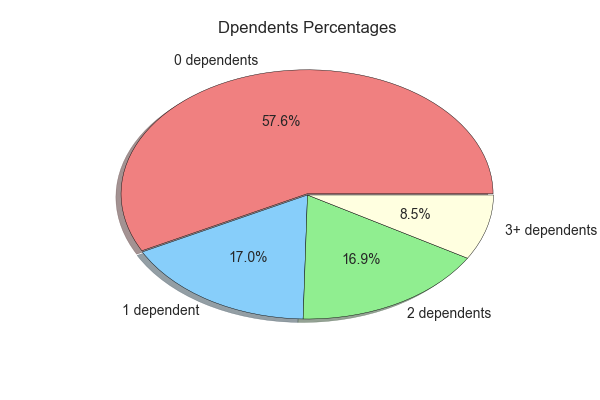
|  |  |
| --- | --- |
| **Variable** | **Description** |
|  |  |
| **Loan\_ID** | Unique Loan ID |
| **Gender** | Male/ Female |
| **Married** | Applicant married (Y/N) |
| **Dependents** | Number of dependents |
| **Education** | Applicant Education (Graduate/ Under Graduate) |
| **Self\_Employed** | Self employed (Y/N) |
| **ApplicantIncome** | Applicant income |
| **CoapplicantIncome** | Coapplicant income |
| **LoanAmount** | Loan amount in thousands |
| **Loan\_Amount\_Term** | Term of loan in months |
| **Credit\_History** | credit history meets guidelines |
| **Property\_Area** | Urban/ Semi Urban/ Rural |
|  |  |
| **Loan\_Status** | Loan approved (Y/N) 🡪🡪🡪🡪 Target Variable |

**SECTION 1:**

**Exploratory Data Analysis**

* 1. **Univariate Analysis**

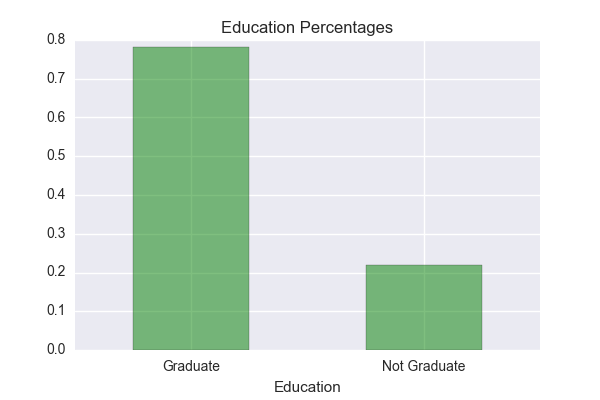
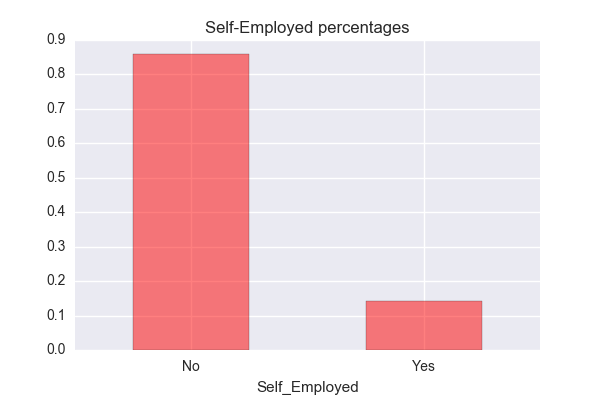
**Categorical Features:**

**** **** 

Gender: Males 🡪81.6% , Females 🡪18.3%

Married: YES 🡪65.1% 🡪, NO 🡪34.8%

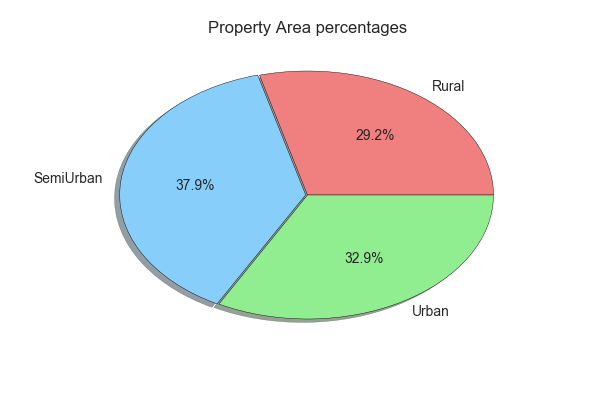
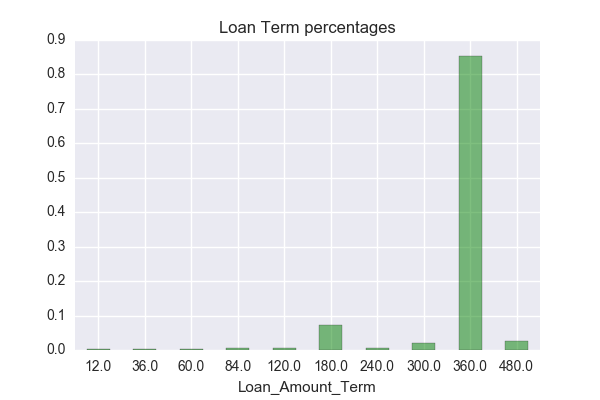
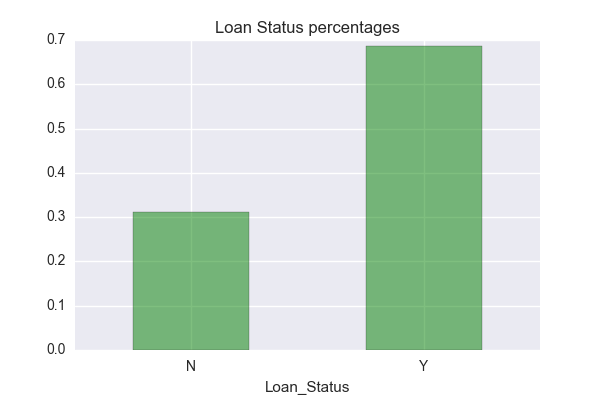
Dependents: 0 dependents 🡪 57.6%, 1 dependent 🡪17%, 2 dependents 🡪 16.9%, 3+ dependents 🡪 8.5%

**** **** 

Education: Graduate 🡪 78.1%, Not-graduate 🡪 21.8%

Self-Employed: NO 🡪 85.9%, YES 🡪 14.0%

Credit History: Yes 🡪 84.2%, NO 🡪15.7%

**** **** 

Property Area: Semi-urban 🡪 37.9%, Urban 🡪 32.9%, Rural 🡪 29.2%

Loan Term: 12.0 🡪 0.16%

36.0 🡪 0.33%

60.0 🡪 0.33%

84.0 🡪 0.66%

120.0 🡪 0.50%

180.0 🡪 7.33%

240.0 🡪 0.66%

300.0 🡪 2.16%

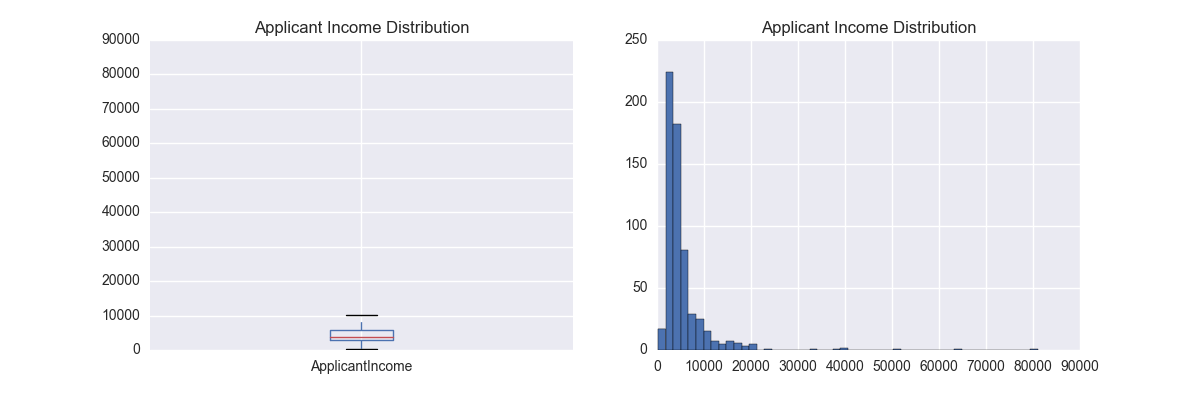
360.0 🡪 85.33%

480.0 🡪 2.50%

Loan Status: Yes 🡪 68.7%, NO 🡪 31.2%

**Numerical Features:**

1. **Applicant Income:**

****

mean = 5403

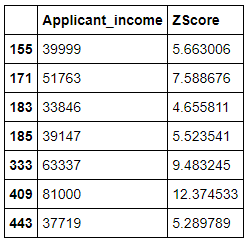
std = 6109

The standard deviation is too much, due to presence

of extreme values, as can be seen in the histogram.

75% of the income lies within 5,800.

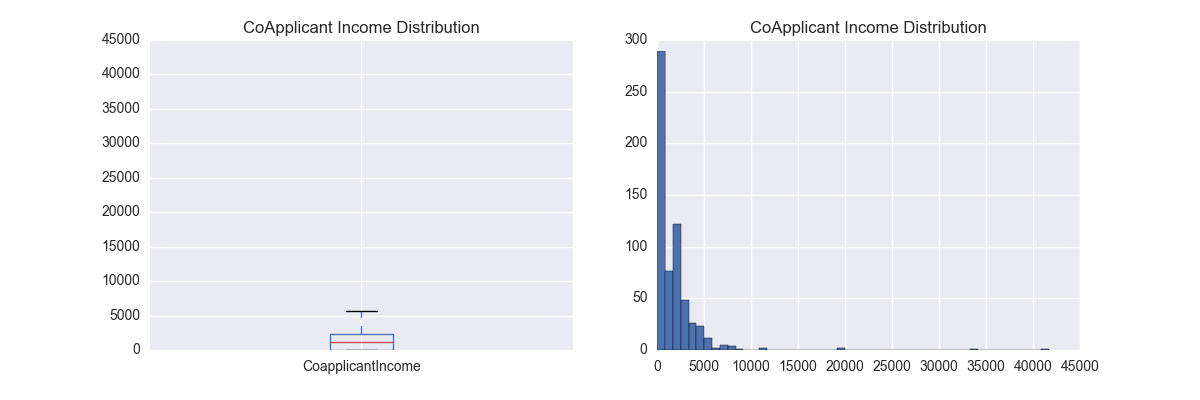
Z-scores of all incomes that are beyond 25,000:



These incomes are beyond 4 standard deviations from the mean. The highest point (i.e. 81000 income) is 12 standard deviations away.

We can call these points as extreme outliers and can be imputed/deleted from our model.

1. **Coapplicant Income:**



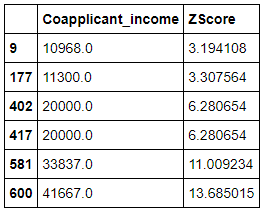
mean = 1621

std = 2926

The standard deviation of this column is too much, due to the presence of extreme outliers.

75% of the data lies within 2300.

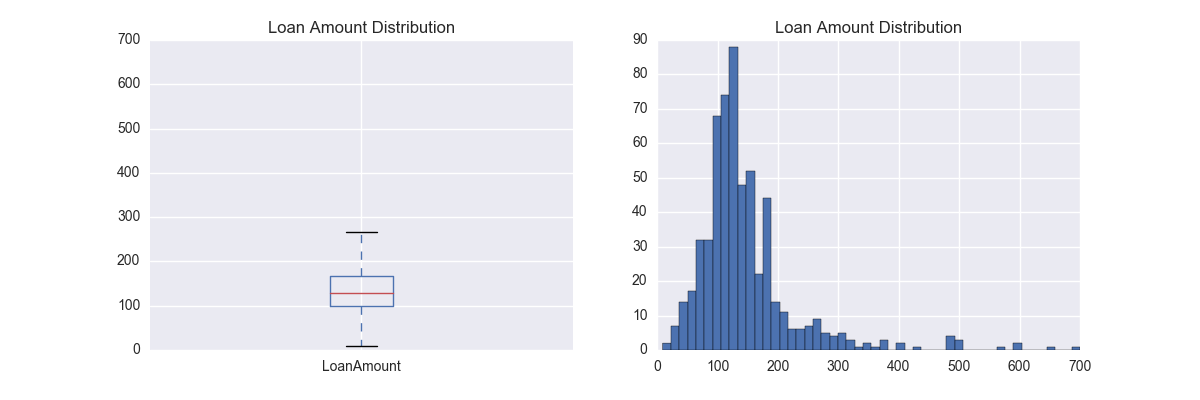
Z-scores of all incomes that are beyond 10,000:



The first 2 points are slightly above 3 standard deviations from the mean. They can be kept in our model.

But the rest of them are much farther away, above 6 standard deviations, they can be imputed /deleted from our model.

1. **Loan Amount:**



mean = 146

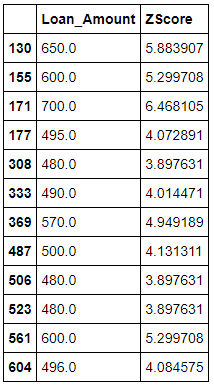
std = 86.

This distribution is positively skewed.

From the histogram, i can say that 90% of the data falls under 450.

Lets calculate the z-score for those points beyond 450.

Z-scores of all loan amounts that are beyond 450:

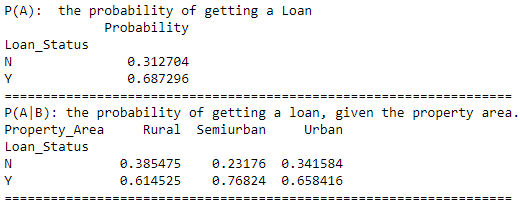
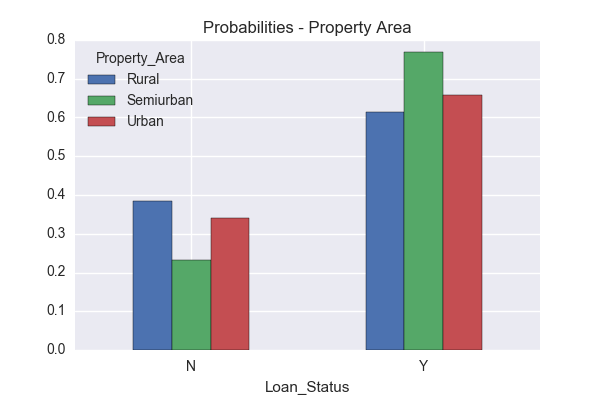


Yes, these are extreme outliers with standard deviations > 3.

* 1. **Bivariate and Multivariate Analysis**

**Categorical Features:**

**1. Property Area vs Loan Status: “**Question: is there an association between getting a Loan and the Property Area?**”**

****

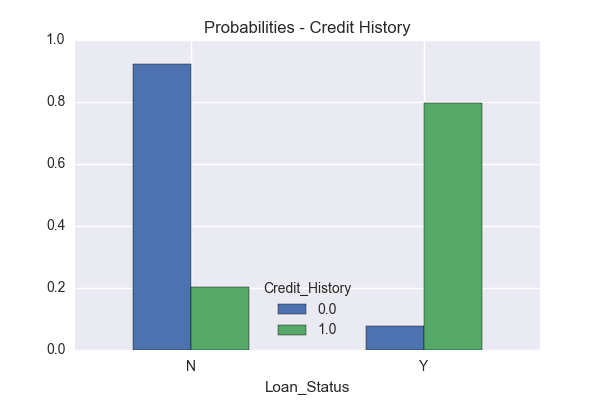
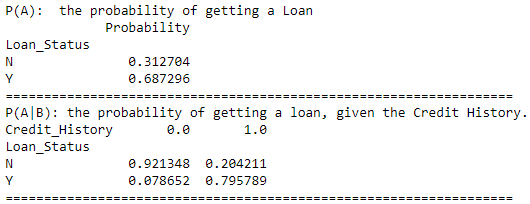
Since, P(A) != P(A/B), we can conclude that there is an association between getting a loan and the property area. The probability of getting a loan, regardless of

property area, is 68.7%.

However, examination of a contingency table containing both Loan Status and Property area showed that the conditional probability of getting a loan changes by property area.

Visual examination of the barplot shows the conditional probabilities of getting a loan are not equal across Property areas.

**2. Credit History vs Loan Status: “**Question: is there an association between getting a Loan and Credit History?**”**

**** ****

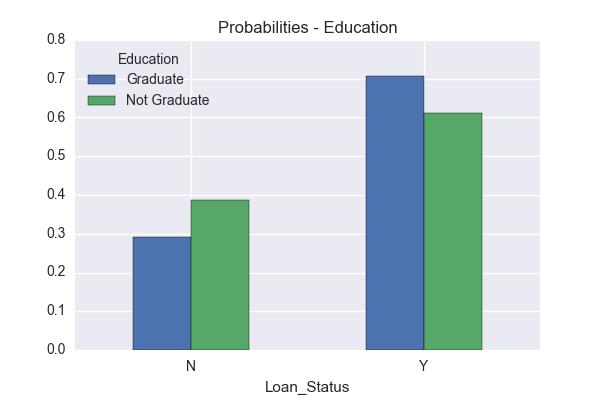
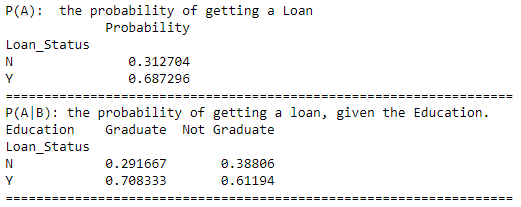
Since, P(A) != P(A/B), there is an association between getting a loan and the credit history. The probability of getting a loan, regardless of credit history, is 68.7%.

However, conditional probability showed that getting a loan changes by credit history.

From the barplot, it is very clear that those customers who have a credit card stands a better chance for getting a loan. We will do some statistical tests to confirm this

assumption later on this project.

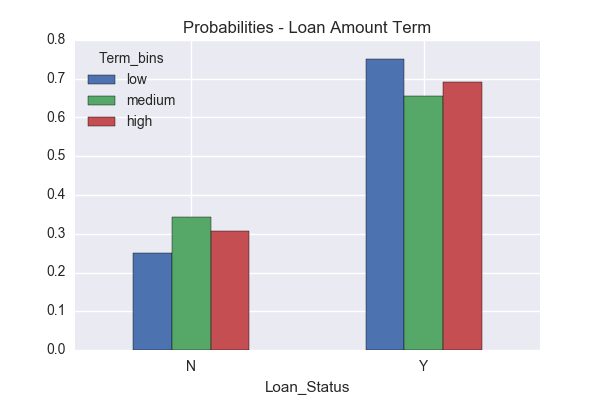
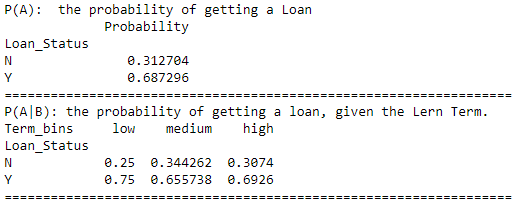
1. **Education vs Loan Status: “**Question: is there an association between getting a Loan and Education?**”**

**** ****

Since, P(A) != P(A/B), we can assume that there is an association between getting a loan and the Education. Although, the difference is just 9%, a statistical test would be helpful

to confirm this assumption.

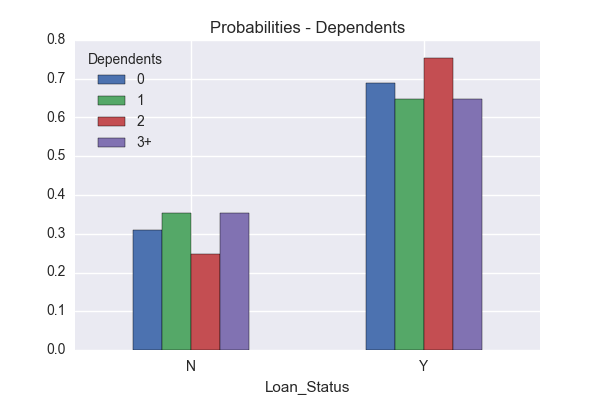
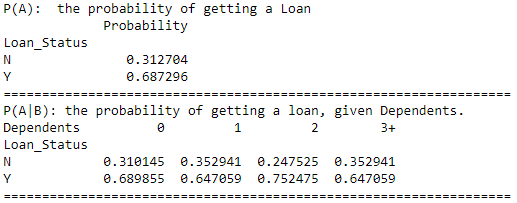
1. **Loan Amount Term vs Loan Status: “**Question: is there an association between getting a Loan and Loan Amount Term?**”**

**** ****

Since, P(A) != P(A/B), there seems to be an association. This needs to be confirmed with some statistical tests.

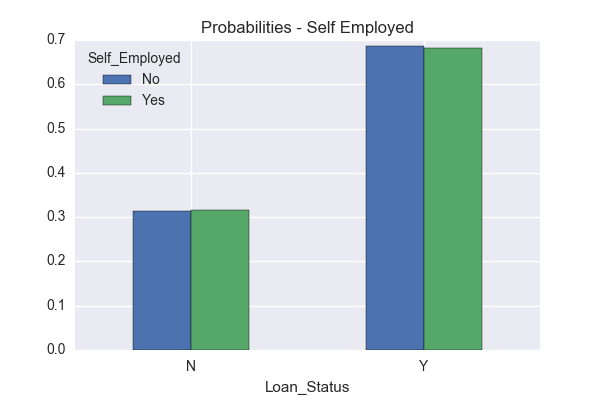
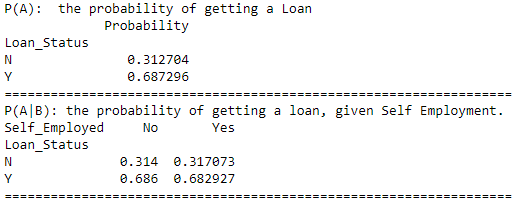
**NOTE:** Loan Term was divided into 3 bins: **low[0-160 months], medium[160-320 months], high[320-500 months]**

1. **Dependents VS Loan Status: “**Question: is there an association between getting a Loan and Dependents?**”**

**** ****

For Dependents == 0, P(A) = P(A/B), whereas, others seems to have an association.

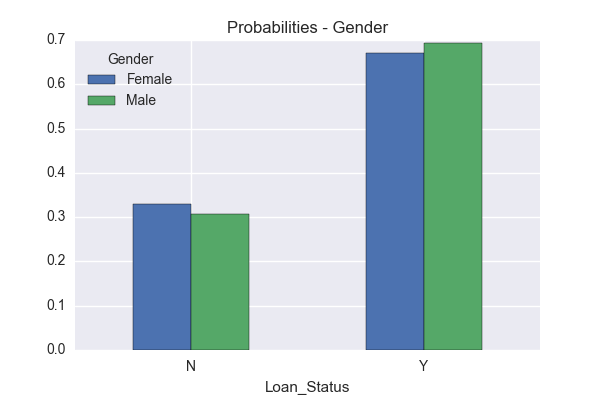
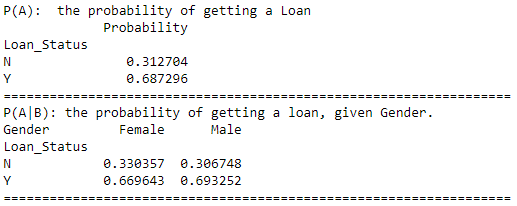
#### **Self Employed vs Loan Status: “**Question: is there an association between getting a Loan and Self Employment?**”**

Since P(A) is almost equal to P(A/B). Our initial assumption would be that there is no association between these two.

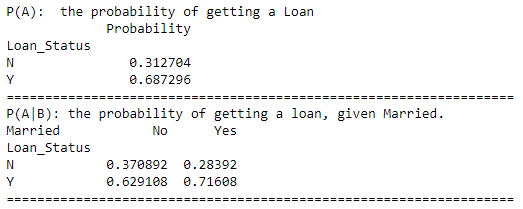
Need to conduct a statistical test to confirm.

1. **Gender vs Loan Status: “**Question: is there an association between getting a Loan and Gender?**”**

**** ****

Since P(A) and P(A/B) are almost equal, we would assume that there is no association between these two. A statistical test would confirm our assumption.

1. **Married vs Loan Status: “**Question: is there an association between getting a Loan and Married?**”**

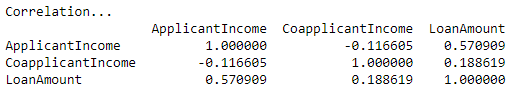
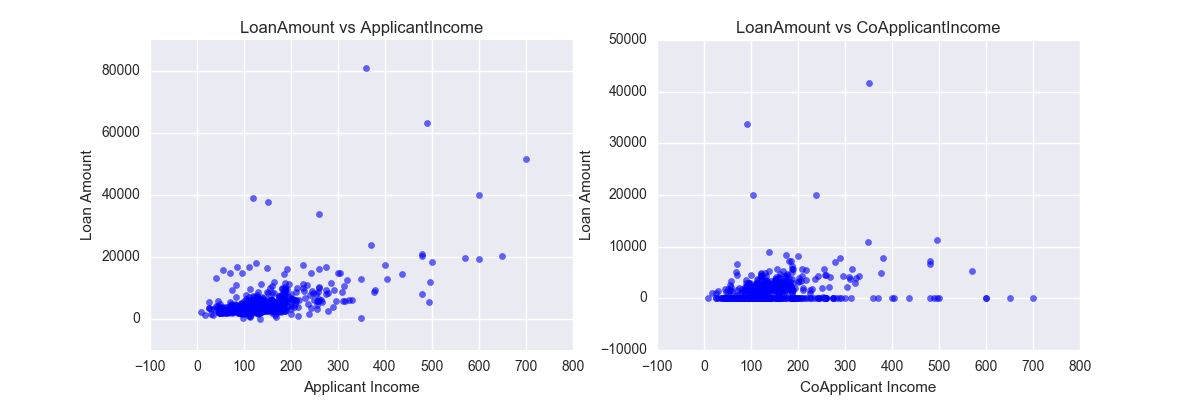
**** 

There is a difference of around 9%, so we would assume that there is an association. A statistical test would confirm their significance.

**Numerical Features:**

1. **ApplicantIncome, CoApplicantIncome vs LoanAmount:**

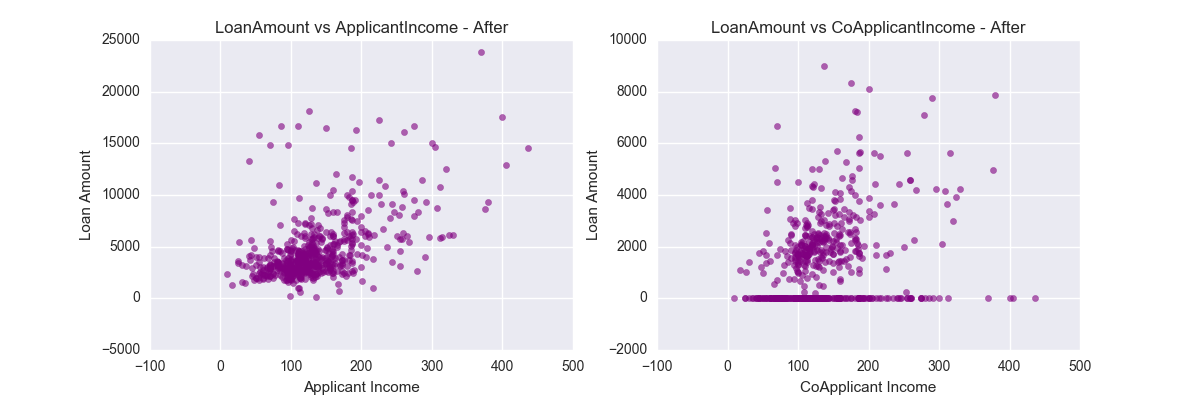
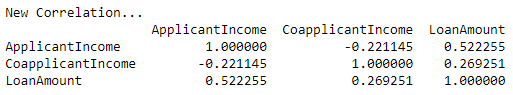
**“**Question:Does Loan Amount applied increases with increase in Applicant and Co-Applicant Incomes?**”**

****

Correlation factor for ApplicantIncome and LoanAmount accounts for 0.57, whereas for Co-applicant Income, it is very poor [just 0.18].

As we saw in Univariate analysis, there were few outliers, i.e. observations having standard deviations higher than 3. Lets delete those and recheck the

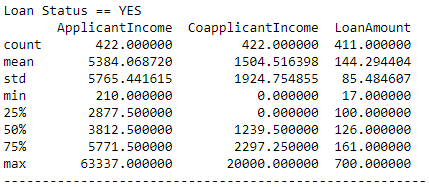
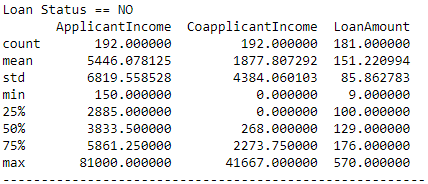
Correlation factors.

We can only see a slight improvement for Co-applicant Income, increasing to 0.26, whereas for ApplicantIncomes, it decreased by 0.05

**Numerical – Categorical Features:**

1. **LoanAmount, ApplicantIncome, Co-ApplicantIncome vs Loan Status:**

** **

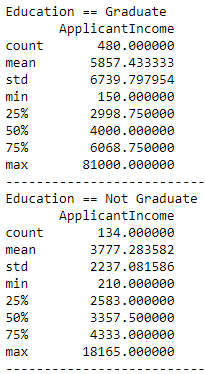
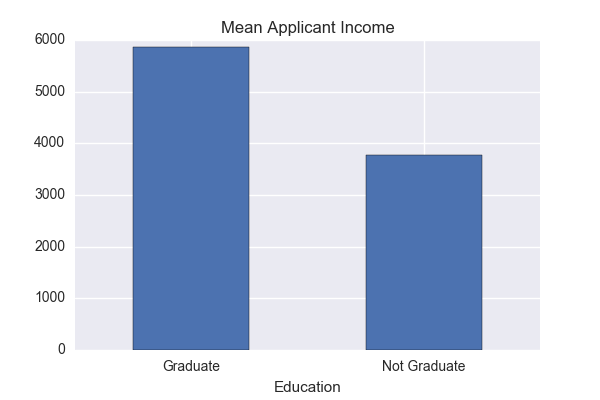
By comparing both summaries, it is seen that there is no enough evidence that getting a loan or not getting a loan is purely based on Income Range or loan amount,

as the means are almost same.

Although the standard deviations varies a lot, this is due to variation in count.

We need to conduct some statistical tests to determine whether there is actually any significance between the two groups.

1. **Education VS Applicant Income: “**Question: Does Graduates have higher chances of getting a high Salary?**”**

****

By comparing both the groups, initial assumption would be that a graduate has more chances of getting a higher salary than a non-graduate.

**SECTION 2:**

**Hypothesis Testing**

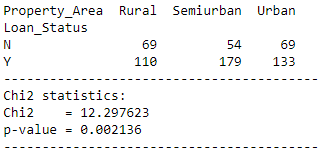
1. **Property Area vs Loan Status:**

As we saw in Bivariate Analysis that the conditional probability of getting a loan is not equal across different property area.

So now we need to conduct a statistical test to confirm this assumption.

**Null hypothesis** : every property area has equal chances of getting a loan; [m1 == m2]

**Alternate** **hypothesis** : not equal chances; [m1 != m2]



With chi2 = 12.29 and p-value < 0.05, we can reject the null hypothesis and confirm that there is actually no equal

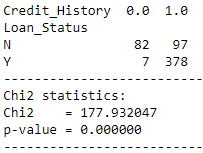
chances of getting a loan for different property area.

1. **Credit History vs Loan Status:**

As we saw in Bivariate Analysis that the conditional probability of getting a loan is not equal across different credit history.

**Null hypothesis** : equal chances of getting a loan for different credit history

**Alternate hypothesis** : not equal chances



Again with p-value < 0.05, we can reject null hypothesis and conclude that the chances are not equal.

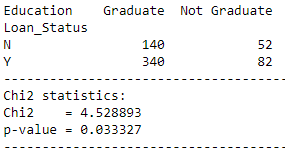
The chi2 statistics is 177.9, which means the difference between the 2 groups of credit\_history is a lot.

1. **Education vs Loan Status:**

We assumed that there is an association between getting a loan and Education, with a difference of just 9% in conditional probabilities.

**Null hypothesis** : equal chances of getting a loan for different education background

**Alternate hypothesis** : not equal chances



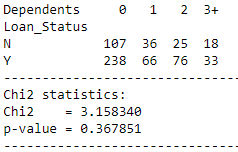
Although the chi2 stats = 4.52 and p-value = 0.03, which is slightly less than 0.05, it is enough to reject null hypothesis and conclude that their chances are not equal.

1. **Dependents vs Loan Status:**

Earlier we found that except dependents = 0, others had an association with Loan Status.

**Null hypothesis** : equal chances of getting a loan for different no. of Dependents

**Alternate hypothesis** : not equal chances



With p-value > 0.05, we fail to reject the null hypothesis.

There is no difference in chances of getting a loan based on number of dependents.

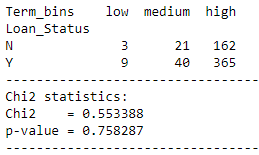
1. **Loan Amount Term vs Loan Status:**

Earlier we divided the Loan Amount Term into 3 categories and through bivariate tests we assumed that there is actually an association with getting a loan.

Lets confirm that assumption.

**Null hypothesis** : equal chances of getting a loan for different Loan Term

**Alternate hypothesis** : not equal chances



With p-value > 0.05, we fail to reject the null hypothesis.

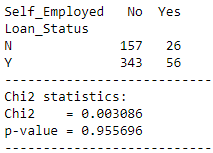
Our initial assumption was wrong and now we can conclude that there is no association between Loan Term and getting a loan.

1. **Self Employed vs Loan Status:**

Earlier we assumed that there is no association between Self\_Employed and Loan Status.

**Null hypothesis** : equal chances of getting a loan for Self\_employment

**Alternate hypothesis** : not equal chances



Yes, our initial assumption was correct. With p-value > 0.05, we fail to reject the null hypothesis.

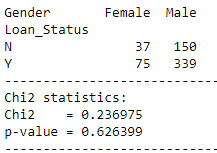
1. **Gender vs Loan Status:**

Null hypothesis : equal chances of getting a loan for Gender

Alternate hypothesis : not equal chances

With p-value > 0.05, we fail to reject the null hypothesis.

We conclude that there is no difference between the genders.

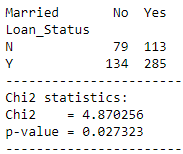


1. **Married vs Loan Status:**

Earlier we saw that probability of getting a loan for married people was high by around 9%.

Null hypothesis : equal chances of getting a loan for Married

Alternate hypothesis : not equal chances



With p-value < 0.05, we reject the null hypothesis.

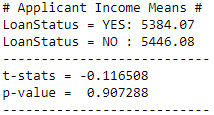
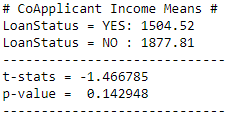
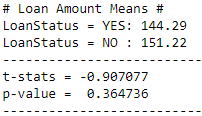
We conclude that there is a difference in getting a loan between the married and unmarried customers.

1. **ApplicantIncome, CoapplicantIncome, LoanAmount vs LoanStatus:**

As we saw in bivariate analysis, there is no enough evidence that getting a loan or not getting a loan is purely based on Income Range or loan amount, as the means are almost same. Lets conduct t-test taking into account the incomes and loan amounts against loan status.

**Null Hypothesis** : customers with loan\_status == yes and those whose loan\_status == no, have equal mean Applicant Income /Coapplicant Income/LoanAmount

**Alternate hypothesis** : not equal

In all the above cases, we fail to reject the null hypothesis as p-values > 0.05

**Conclusion:**

We did Exploratory data Analysis on the features of this dataset and saw how each feature is distributed. We also calculated probabilities to prove an association among the Independent variables and the Target variable. And based on the results, we assumed whether or not there is an association.

Lastly, Statistical Tests were conducted so as to confirm or deny the assumptions we made.