Initial Analysis

Albert Hutiancong Wang, Chelsea Lee, Amanda Zhang

2023-05-25

Research Introduction

Loans offer many benefits if borrowed responsibly. They open pathways to new opportunities and growth that would have otherwise been difficult to obtain without borrowing money. They're also vital in stimulating economic growth. However, obtaining a loan is quite difficult since applicants must meet many requirements before a lender can offer a loan. Therefore, we wonder if various social factors can influence the dollar value of a loan. We believe these social factors: single male, graduate level education, not self employed, longer loan term, and high income will increase the chances of the applicant receiving a larger loan. If there is a trend between these social factors and the given loan amount, it can help predict an applicant's chances of receiving the loan amount they prefer. Knowing their chances can greatly help in the financial decision-making process and save time.

Research Design

We've decided to use this loan approval data set from Kaggle since it includes multiple characteristics of the applicant that are relevant to loan approval and the applicant's loan amount. The explanatory variables of the data are the various differences in applicants. Measured differences are gender, marital status, number of dependants, level of education, and annual income. There is also information about co-signers on applications, as well as their income levels. All of these factors should relate to changes in the amount of the loans received. All those explanatory variables of interest are measured in corresponding units; dollars for income & loans, yes/no for binary variables, and input answers for other categorical variables. Our approach to this research is to compare the value of these social factors and determine their impact on the loan outcome.

Initial Analysis

Load necessary packages

```
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(tidyr))
suppressPackageStartupMessages(library(ggplot2))
```

Load in data

```
data <- read.csv('loan_test_filtered.csv')
```

We wanna know which factor affects more on a candidate's loan amount, coapplicant income or loan term.

```
co_income_term <- lm(log(Loan_Amount) ~ log(Coapplicant_Income) + Term, data = data)
summary(co_income_term)
```

We can run a simple linear regression and take a look at the statistical significance

```
##
## Call:
## lm(formula = log(Loan_Amount) ~ log(Coapplicant_Income) + Term,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05274 -0.20854  0.00299  0.18678  0.93517
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              10.931779   1.163415   9.396 1.03e-12 ***
## log(Coapplicant_Income)   0.319290   0.083173   3.839 0.000343 ***
## Term                      0.004653   0.001126   4.132 0.000134 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3729 on 51 degrees of freedom
## Multiple R-squared:  0.3402, Adjusted R-squared:  0.3144
## F-statistic: 13.15 on 2 and 51 DF,  p-value: 2.48e-05
```

From the summary statistics, we can see that for every 1% increase in the coapplicant's income, there will be 0.319290% increase in average in the loan amount ceteris paribus. For every 1 unit increase in term, there will be 0.004653 increase in loan amount holding other factors constant. In general, Since Coapplicant_Income has a much higher value than that of Term, we conclude that the coapplicant income has a stronger effect on loan amount than term of loan does.

Do various social factors change the dollar value of a loan that can be achieved?

```
data <- cbind(
  data,
  model.matrix(~ Gender - 1, data = data),
  model.matrix(~ Married - 1, data = data),
  model.matrix(~ Education - 1, data = data),
  model.matrix(~ Self_Employed - 1, data = data),
  model.matrix(~ Area - 1, data = data)
)
```

Use model.matrix to make dummy variables

```
colnames(data)[16] <- "GraduateNo"
```

We realize that one-hot encoding leads to appearances of multiple correlated variables thus causes perfect multicollinearity.

```
model_loan <- lm(log(Loan_Amount) ~ GenderMale + MarriedYes + GraduateNo + Self_EmployedYes + log(Applic

summary(model_loan)
```

Now we can run a linear regression model based on the social factor variables we choose

```
##
## Call:
## lm(formula = log(Loan_Amount) ~ GenderMale + MarriedYes + GraduateNo +
##     Self_EmployedYes + log(Applicant_Income) + log(Coapplicant_Income) +
##     Term + AreaUrban, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60614 -0.12766  0.02905  0.18030  0.53240
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.760785   1.611195   1.714 0.093502 .
## GenderMale                0.036911   0.125385   0.294 0.769821
## MarriedYes                0.368409   0.175992   2.093 0.041987 *
## GraduateNo                0.107208   0.101105   1.060 0.294638
## Self_EmployedYes         -0.091527   0.112258  -0.815 0.419178
## log(Applicant_Income)     0.637746   0.105511   6.044 2.68e-07 ***
## log(Coapplicant_Income)   0.280684   0.068233   4.114 0.000163 ***
## Term                      0.004763   0.000846   5.629 1.10e-06 ***
## AreaUrban                -0.092994   0.078616  -1.183 0.243063
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2698 on 45 degrees of freedom
## Multiple R-squared:  0.6952, Adjusted R-squared:  0.641
## F-statistic: 12.83 on 8 and 45 DF,  p-value: 2.144e-09
```

We can conclude that ceteris paribus, being a male, being married, and having more income as applicant, having more income as coapplicant, would all have positive relationship on getting a higher amount of loan. Nevertheless, if the individual has not graduated, has not being self-employed, not living in urban area, those social factors would have negative effects on the loan amount the individuals can get.

What is the average loan amount difference between male and female?

```r
loan_diff <- data %>%
  group_by(Gender) %>%
  summarise(avg_loan_diff = mean(Loan_Amount))

loan_diff <- loan_diff %>%
  spread(Gender, avg_loan_diff) %>%
  mutate(diff_gender = Male - Female)

avg_diff <- loan_diff$diff_gender

avg_diff
```

```
## [1] -3072644
```

Not holding any other variables constant, we can conclude that the average loan amount differences between male and female is 3072644 dollars - female loans much more than men in general.

What is the average difference in applicant income between male and female?

```r
income_diff <- data %>%
  group_by(Gender) %>%
  summarize(avg_income_diff = mean(Applicant_Income))

income_diff <- income_diff %>%
  spread(Gender, avg_income_diff) %>%
  mutate(diff_male_female = Male - Female)

diff_in_income <- income_diff$diff_male_female

diff_in_income
```

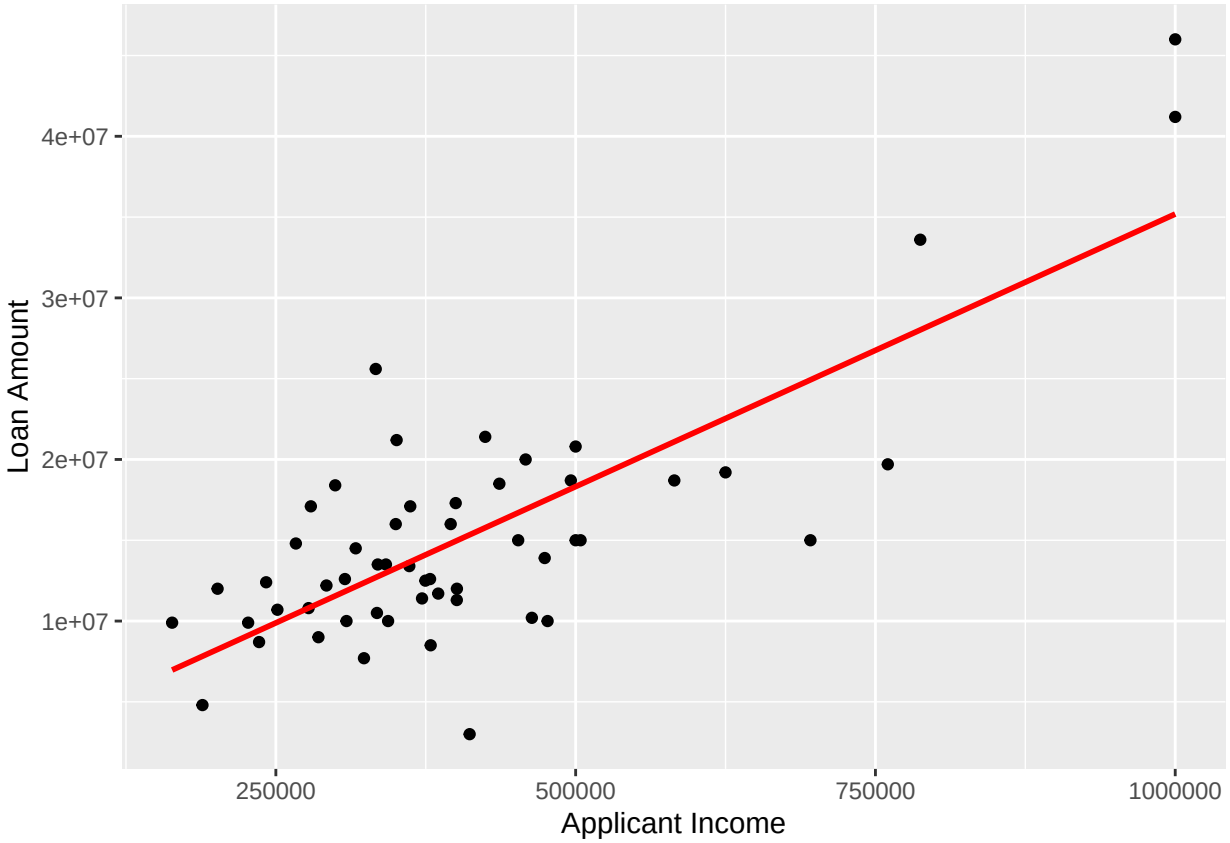
```
## [1] -17793.01
```

Not holding other social factors constant, we can conclude that in general, the average difference in applicant income between male and female is 17793.01 dollars - female earns much more than male.

plot 2

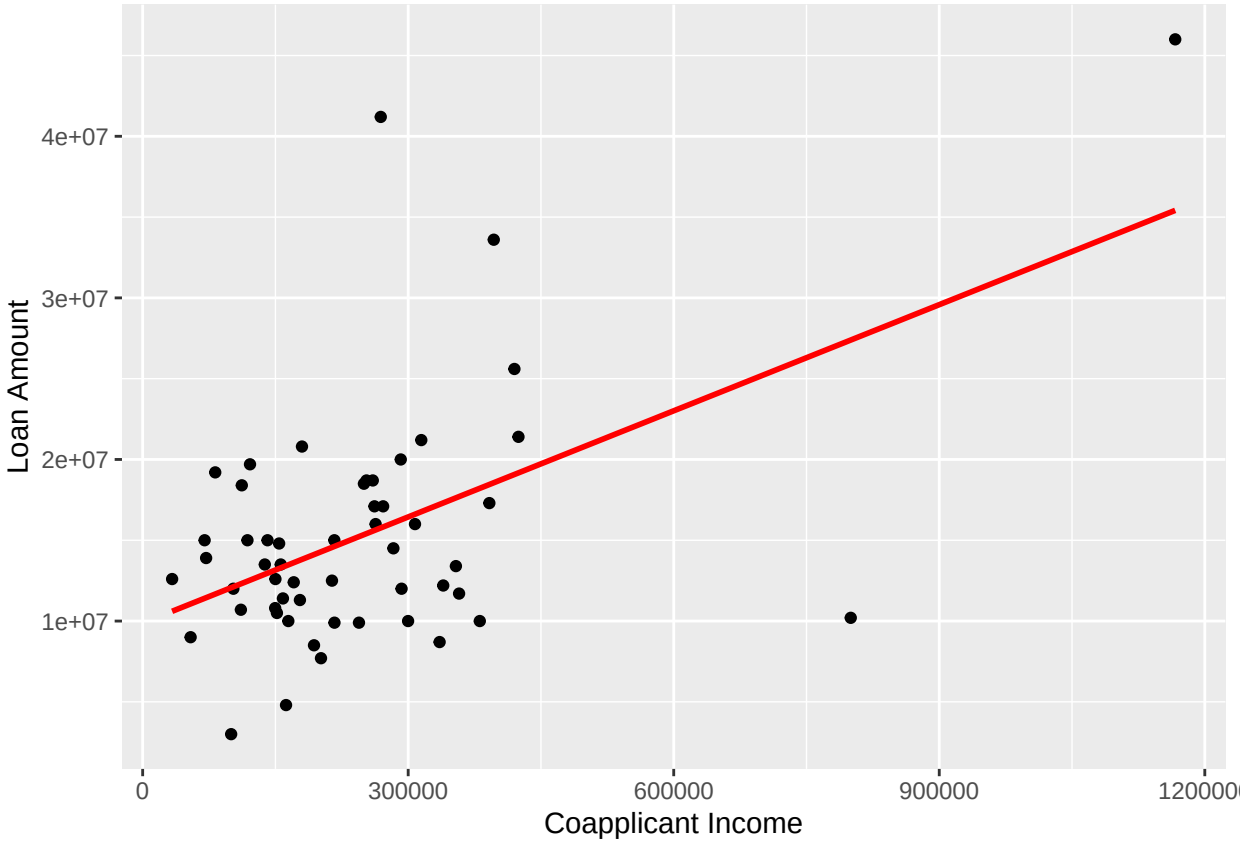
```r
# For Numerical Values
ggplot(data, aes(x = Applicant_Income, y = Loan_Amount)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(x = "Applicant Income", y = "Loan Amount")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



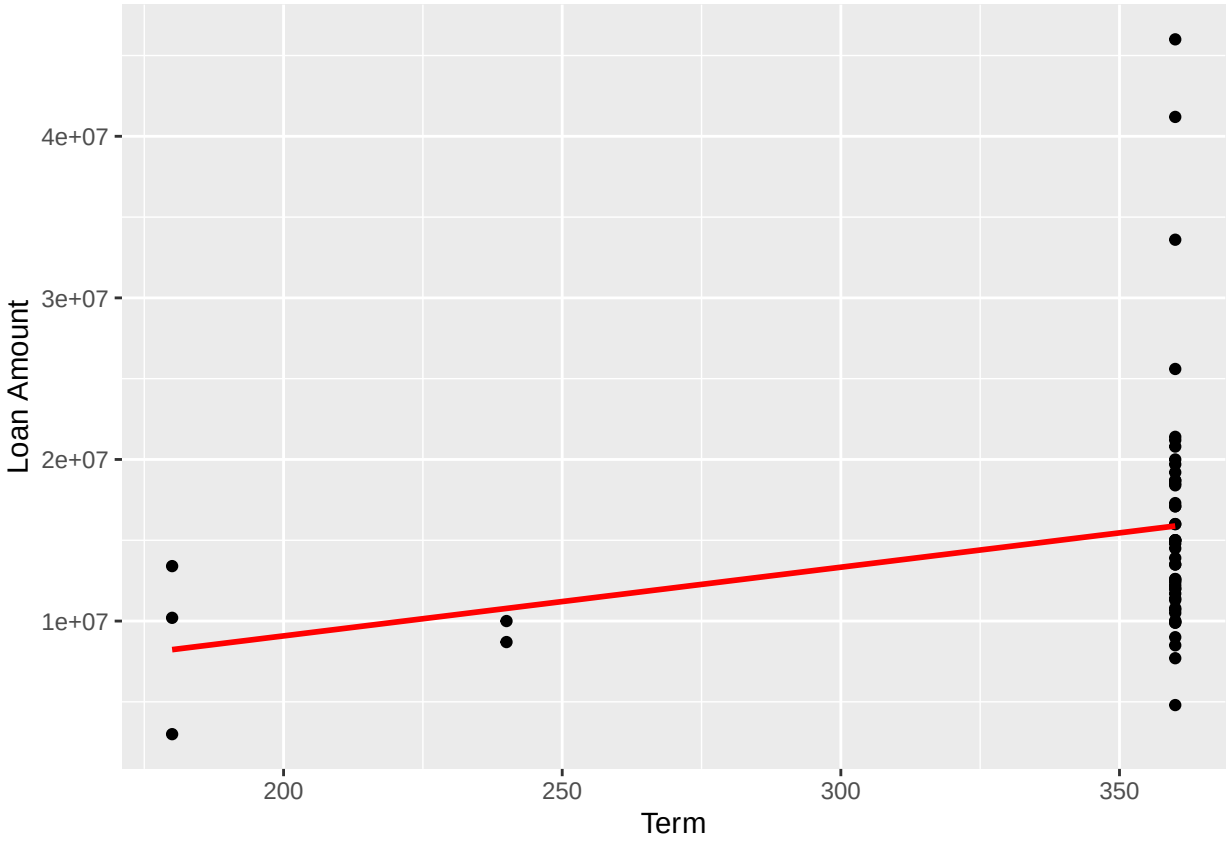

```
ggplot(data, aes(x = Coapplicant_Income, y = Loan_Amount)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(x = "Coapplicant Income", y = "Loan Amount")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

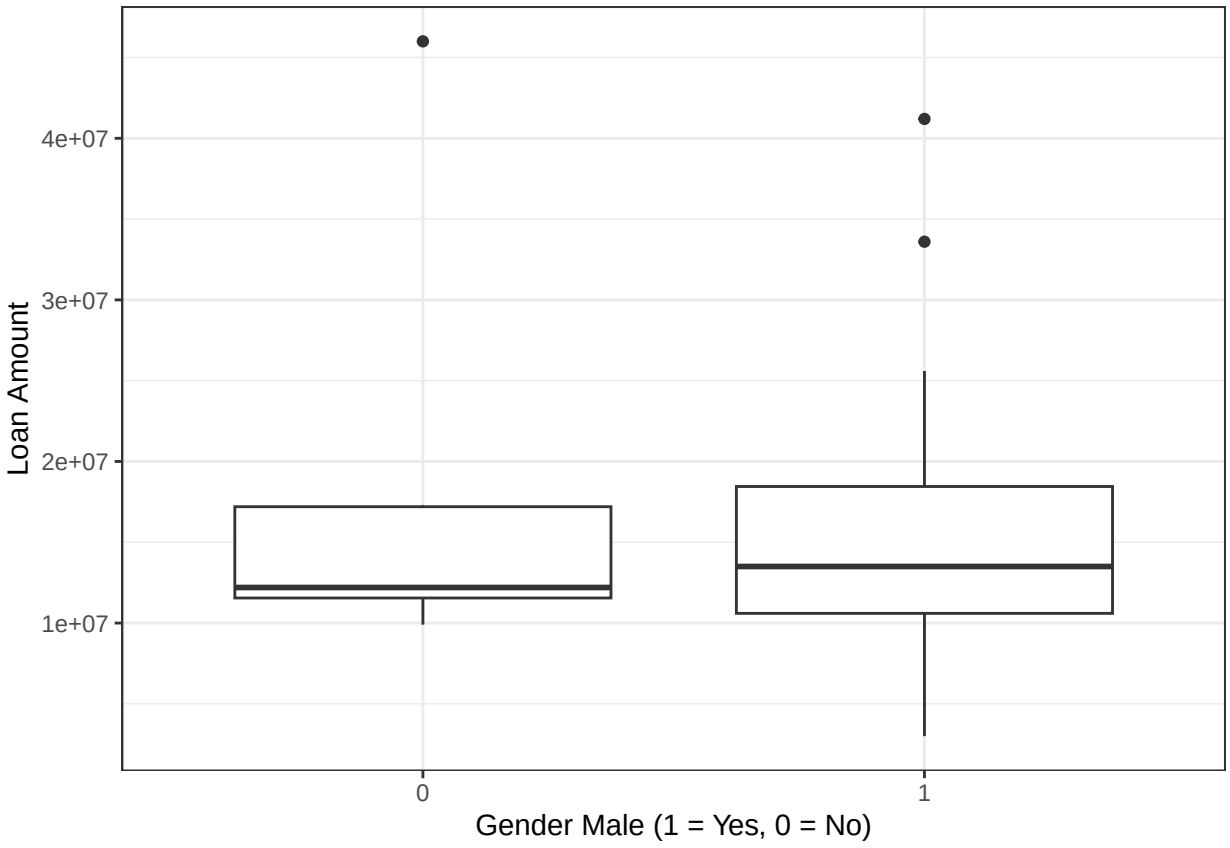



```
ggplot(data, aes(x = Term, y = Loan_Amount)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(x = "Term", y = "Loan Amount")
```

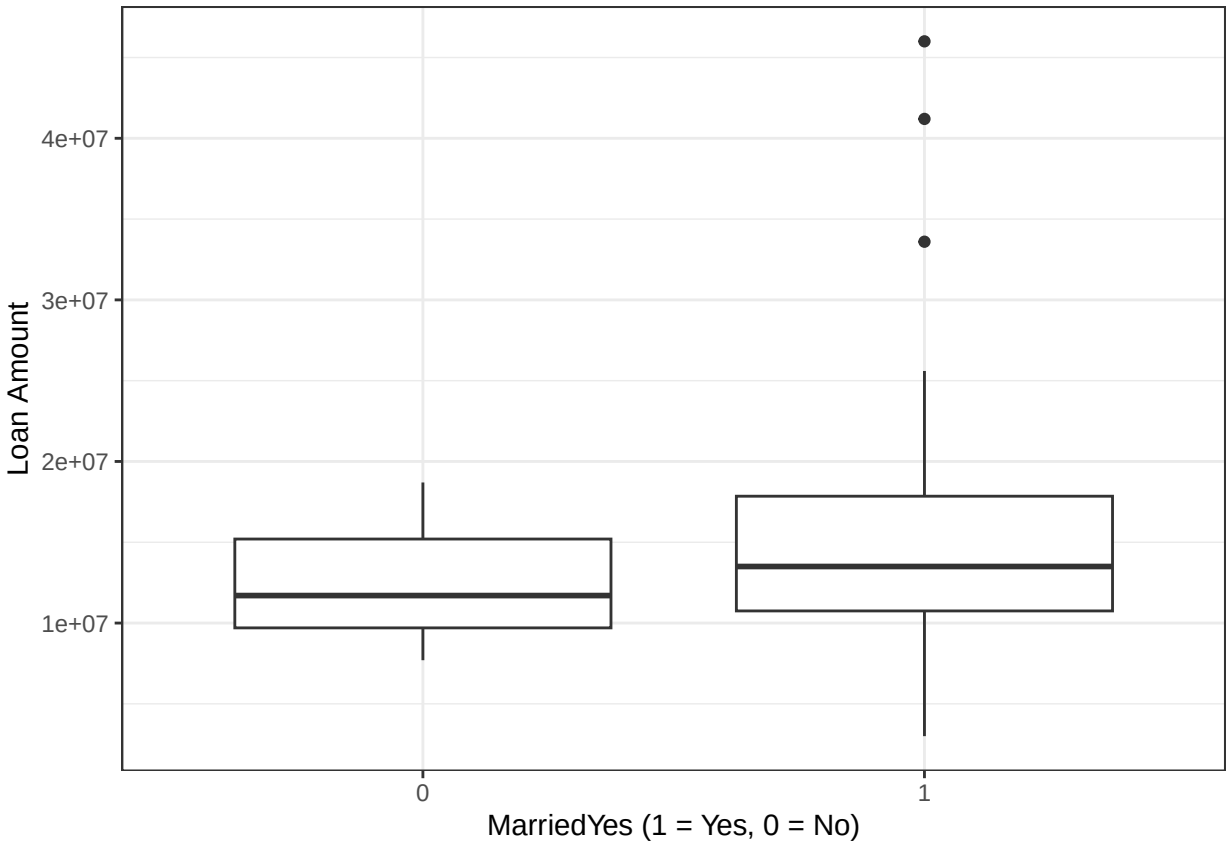
```
## `geom_smooth()` using formula = 'y ~ x'
```



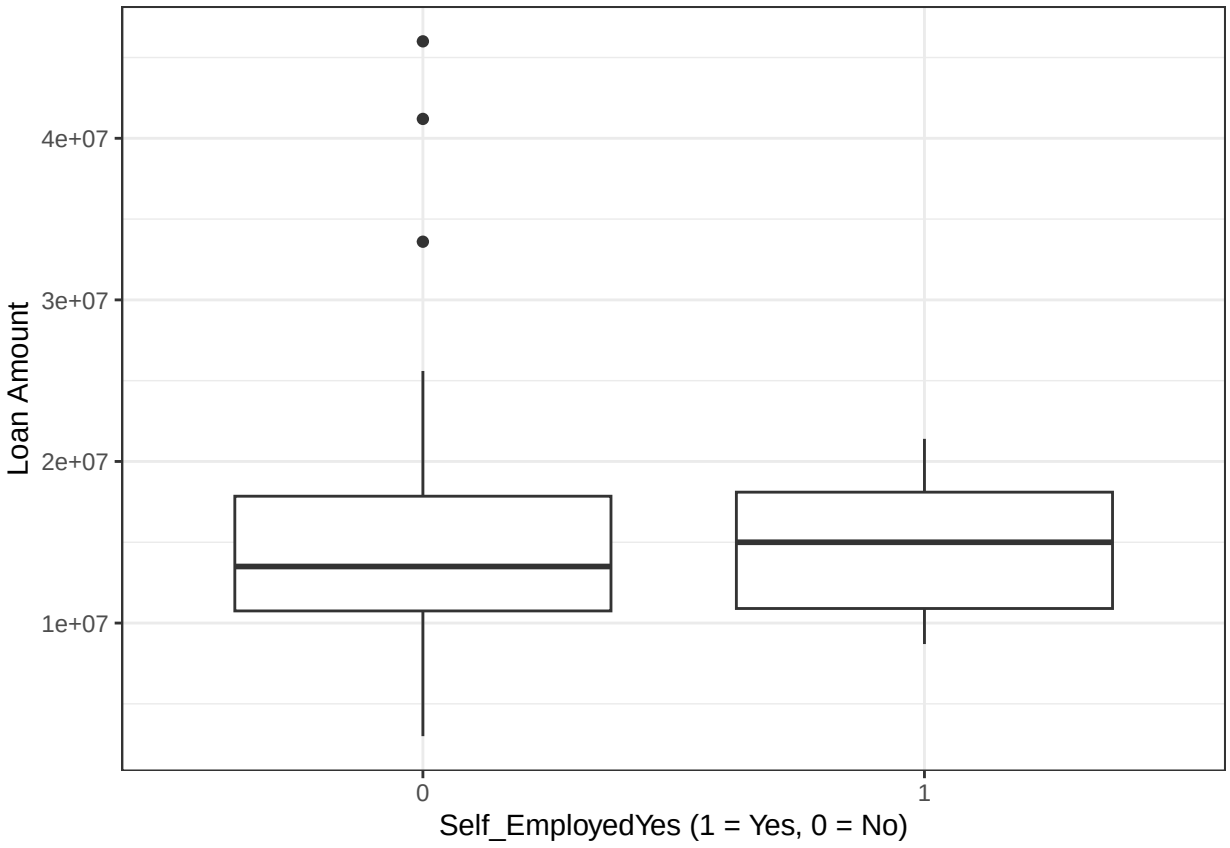

```
# For Categorical Values
ggplot(data, aes(x = as.factor(GenderMale), y = Loan_Amount)) +
  geom_boxplot() +
  theme_bw() +
  labs(x = "Gender Male (1 = Yes, 0 = No)", y = "Loan Amount")
```



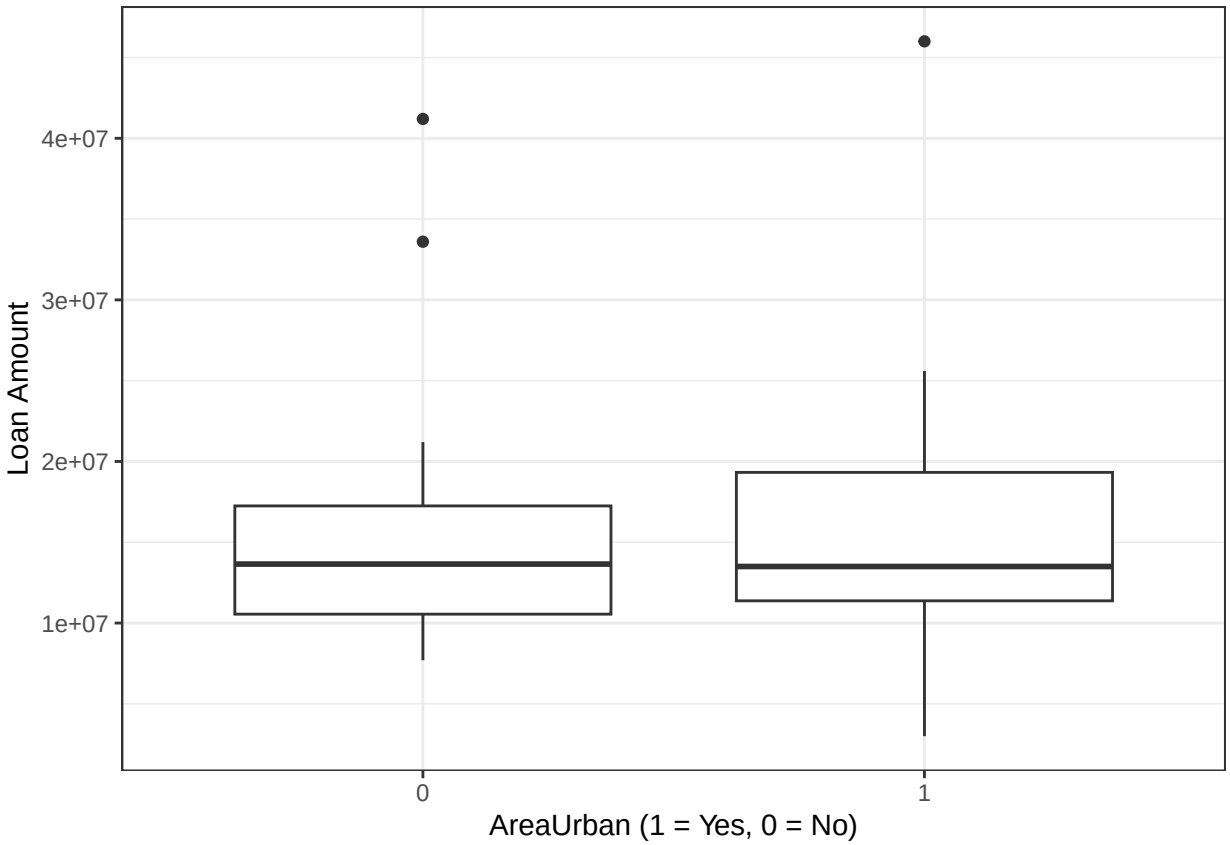

```
ggplot(data, aes(x = as.factor(MarriedYes), y = Loan_Amount)) +
  geom_boxplot() +
  theme_bw() +
  labs(x = "MarriedYes (1 = Yes, 0 = No)", y = "Loan Amount")
```




```
ggplot(data, aes(x = as.factor(Self_EmployedYes), y = Loan_Amount)) +
  geom_boxplot() +
  theme_bw() +
  labs(x = "Self_EmployedYes (1 = Yes, 0 = No)", y = "Loan Amount")
```




```
ggplot(data, aes(x = as.factor(AreaUrban), y = Loan_Amount)) +
  geom_boxplot() +
  theme_bw() +
  labs(x = "AreaUrban (1 = Yes, 0 = No)", y = "Loan Amount")
```



Loan Amount v.s. Applicant Income From the scatterplot, we can see that in general, there is a positive relationship between applicant income and loan amount. When applicant holds more income, the individual would generally be able to fetch a higher volume of loan. This relationship can fit approximately into a linear regression curve with couple outliers.

Loan Amount v.s. Coapplicant Income

The situation is similar to the graph from previous consideration, however, it is worth noting that the line does not necessarily well capture the observations, which can be due to the lack of observations for coapplicants who have income more than $350000. The data points are also concentrated close to the line, indicating that there is no evident linear relationship between those two. If we were to use high degree polynomial model then it might show a better relationship between loan amount and coapplicant income.

Loan Amount v.s. Term

With an increase in term, the applicants will have higher volume of loans, which is reasonable, and it also identifies that our plotting technique is accurate.

Loan Amount v.s. GenderMale or No

From the boxplots we can see that male usually has a higher chance of getting larger value loans compared to women with higher median and quantile values, ceteris paribus. However, this should be verified by getting more female data samples.

Loan Amount v.s. MarriedYes or No

From the boxplots we can see that people who are indeed married would have a higher chance of getting larger loans with higher median and quantile values, probably due to the fact that the newly developed familys would tend to take loans, and being in a marriage may indicate more financial stability.

Loan Amount v.s. Self_Employed_Yes or No

It can be seen that people who are self-employed would have a higher chance of getting larger loans with higher median and quantile values, probably due to the reason that self-employed ones tend to get higher potential income to pay back loans or with more collaterals.

Loan Amount v.s. AreaUrban or No

It can be seen that the upper and lower quantile values are higher from the ones who are in urban areas, but it is worth noting that the median value is the same for both. There are couple of reasons which can validate the phenomena like income inequality, living cost disparities, educational purposes, and etc.

Loan Amount v.s. GraduateNo or Yes

The plot shows that applicants who have not completed graduate education tend to receive larger loan amounts. This may be because in general there are less people who complete graduate level of education and more people who complete undergraduate education. Therefore, there is a higher chance of an applicant who has completed undergraduate education to apply for loans to pay off student tuition and debt. Also, people who have completed graduate level of education tend to have a stable job or other stable source of income that they can use to pay for their higher level education. On the other hand, those who have a undergraduate education tend to be in entry level jobs that require more job experience to be paid more which is why these types of people apply for loans to pay off their student debt.

Research Conclusion

All in all, our analysis shows that the following variables will result in higher loan amounts: larger applicant income, longer loan term, male applicant, applicants who are married, applicants who are self-employed, and applicants who have not completed graduate education. Our original hypothesis was not fully correct in the way that we believed that single applicants with a graduate level education would receive higher loans, but the belief that these social factors affect an applicant's loan amount in different degrees is valid. However, the effectiveness of the research result is insufficient since more data samples and other data resources need to be incorporated to reduce presentation biases from the dataset. In addition, it is difficult to control the lender's prejudice and biases of the applicant which can sway the results of the loan amount given. Knowing this, our research result should not be taken into full account but can provide some information to help applicants predict their loan chances.