DATA 606 Data Project Proposal

Md Forhad Akbar

Load Package

```
library(tidyverse)
```

Data Preparation

```
# load data
my_loan_data<- read.csv("https://raw.githubusercontent.com/forhadakbar/data606fall2019stat/master/Final%
```

```
head(my_loan_data)
```

```
##    Loan_ID Gender Married Dependents    Education Self_Employed
## 1 LP001002   Male      No          0     Graduate            No
## 2 LP001003   Male     Yes          1     Graduate            No
## 3 LP001005   Male     Yes          0     Graduate           Yes
## 4 LP001006   Male     Yes          0 Not Graduate            No
## 5 LP001008   Male      No          0     Graduate            No
## 6 LP001011   Male     Yes          2     Graduate           Yes
##   ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term
## 1            5849                 0         NA              360
## 2            4583              1508        128              360
## 3            3000                 0         66              360
## 4            2583              2358        120              360
## 5            6000                 0        141              360
## 6            5417              4196        267              360
##   Credit_History Property_Area Loan_Status
## 1              1         Urban           Y
## 2              1         Rural           N
## 3              1         Urban           Y
## 4              1         Urban           Y
## 5              1         Urban           Y
## 6              1         Urban           Y
```

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.
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.

Cases

What are the cases, and how many are there?

```r
dim(my_loan_data)
```

```
## [1] 614  13
```

There are 614 cases

Data collection

Describe the method of data collection.

This data source was given as part of a data science challenge or practice problem. I downloaded the data and loaded to my git-hub account. I will read the data into R from my git-hub account using raw link of the csv file using read.csv command.

Type of study

What type of study is this (observational/experiment)?

This is an observational study

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

Source: https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/

Dependent Variable

What is the response variable? Is it quantitative or qualitative?

Loan_Status is the response variable. It is a categorical variable which gives us yes and no for loan approval status.

Independent Variable

You should have two independent variables, one quantitative and one qualitative.

I have few independent variables that i will consider for now. I will choose the most appropiate variables after doing exploratory analysis.

Applicants took a loan before. Credit history is the variable which answers that.
Applicants with higher incomes. So, we might look at the applicant income variable.
Applicants with higher education.
Gender of the applicant.
Number of Dependens an applicant has.
Property area contains location information of the loan property applied for.

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```r
str(my_loan_data)
```

```
## 'data.frame':    614 obs. of  13 variables:
##  $ Loan_ID          : Factor w/ 614 levels "LP001002","LP001003",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Gender           : Factor w/ 3 levels "","Female","Male": 3 3 3 3 3 3 3 3 3 3 ...
##  $ Married          : Factor w/ 3 levels "","No","Yes": 2 3 3 3 2 3 3 3 3 3 ...
##  $ Dependents       : Factor w/ 5 levels "","0","1","2",..: 2 3 2 2 2 4 2 5 4 3 ...
##  $ Education        : Factor w/ 2 levels "Graduate","Not Graduate": 1 1 1 2 1 1 2 1 1 1 ...
##  $ Self_Employed    : Factor w/ 3 levels "","No","Yes": 2 2 3 2 2 3 2 2 2 2 ...
##  $ ApplicantIncome  : int  5849 4583 3000 2583 6000 5417 2333 3036 4006 12841 ...
##  $ CoapplicantIncome: num  0 1508 0 2358 0 ...
##  $ LoanAmount       : int  NA 128 66 120 141 267 95 158 168 349 ...
##  $ Loan_Amount_Term : int  360 360 360 360 360 360 360 360 360 360 ...
##  $ Credit_History   : int  1 1 1 1 1 1 1 0 1 1 ...
##  $ Property_Area    : Factor w/ 3 levels "Rural","Semiurban",..: 3 1 3 3 3 3 3 2 3 2 ...
##  $ Loan_Status      : Factor w/ 2 levels "N","Y": 2 1 2 2 2 2 2 1 2 1 ...
```

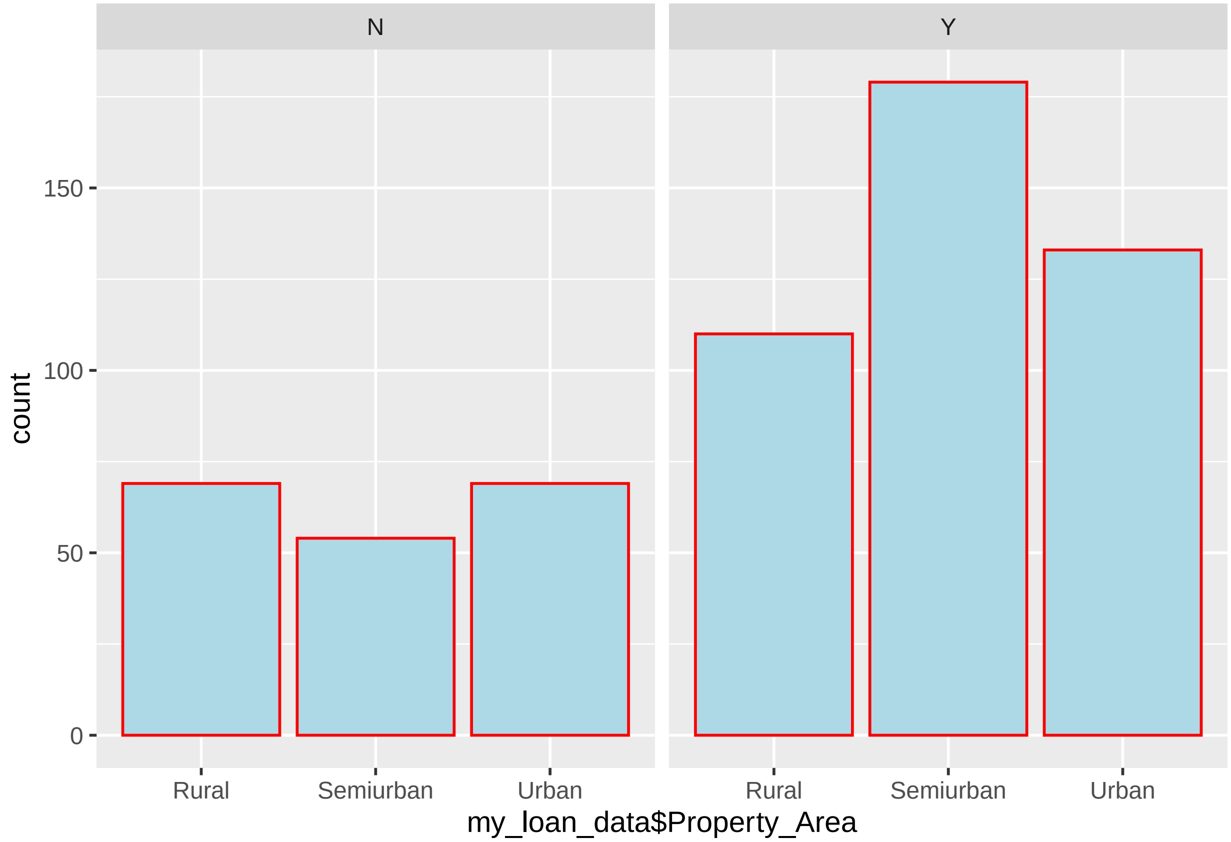
Property Area:

```r
summary(my_loan_data$Property_Area)
```

```
##     Rural Semiurban     Urban
##       179       233       202
```

```r
ggplot(data=my_loan_data, aes(my_loan_data$Property_Area)) +
  geom_histogram(col="red",fill="lightblue",stat="count" ) +
  facet_grid(~my_loan_data$Loan_Status)+
  scale_x_discrete()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

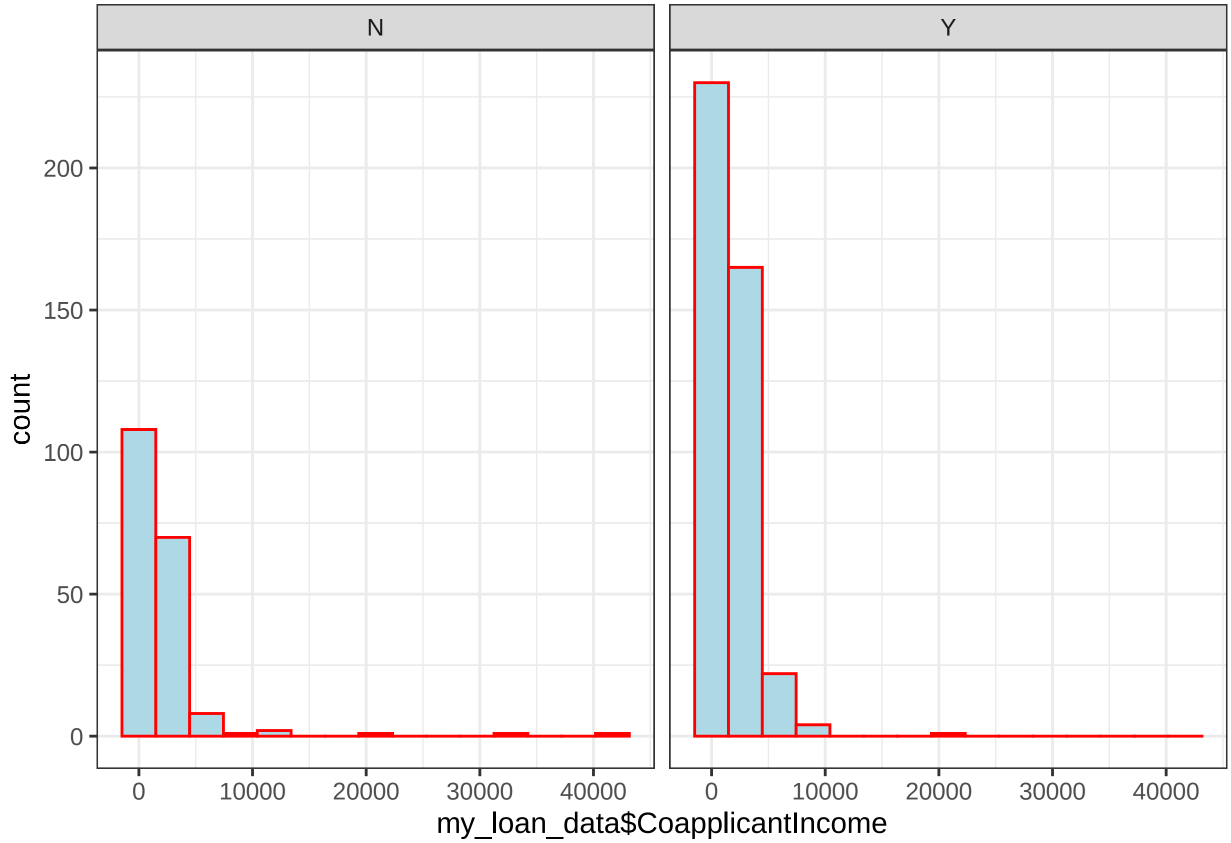


Coapplicant Income:

```
summary(my_loan_data$CoapplicantIncome)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       0       0    1188    1621    2297   41667
```

```
ggplot(data=my_loan_data, aes(x= my_loan_data$CoapplicantIncome)) +
  geom_histogram(col="red",fill="lightblue", bins = 15) +
  facet_grid(~my_loan_data$Loan_Status)+
  theme_bw()
```



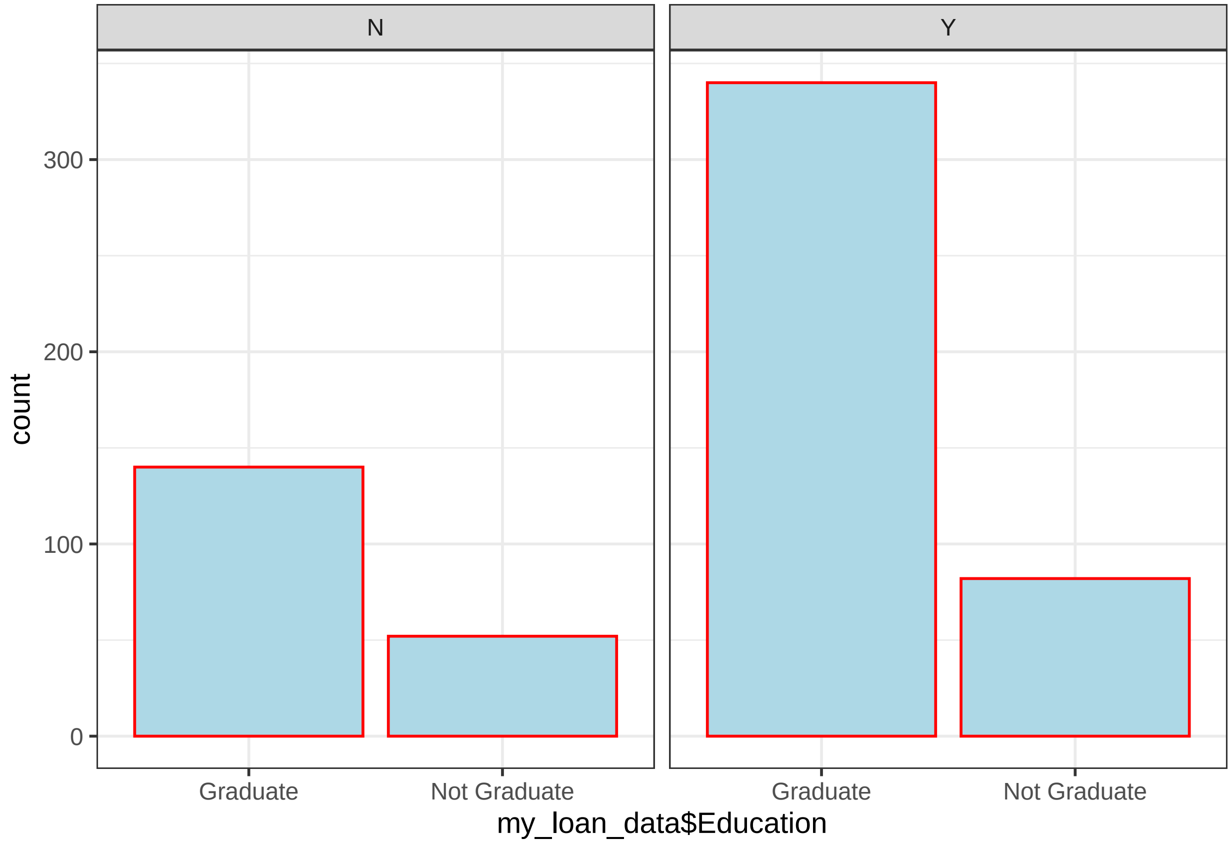
Education:

```
summary(my_loan_data$Education)
```

```
##     Graduate Not Graduate
##          480          134
```

```
ggplot(data=my_loan_data, aes(my_loan_data$Education)) +
  geom_histogram(col="red",fill="lightblue",stat="count" ) +
  facet_grid(~my_loan_data$Loan_Status)+
  scale_x_discrete()+
  theme_bw()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



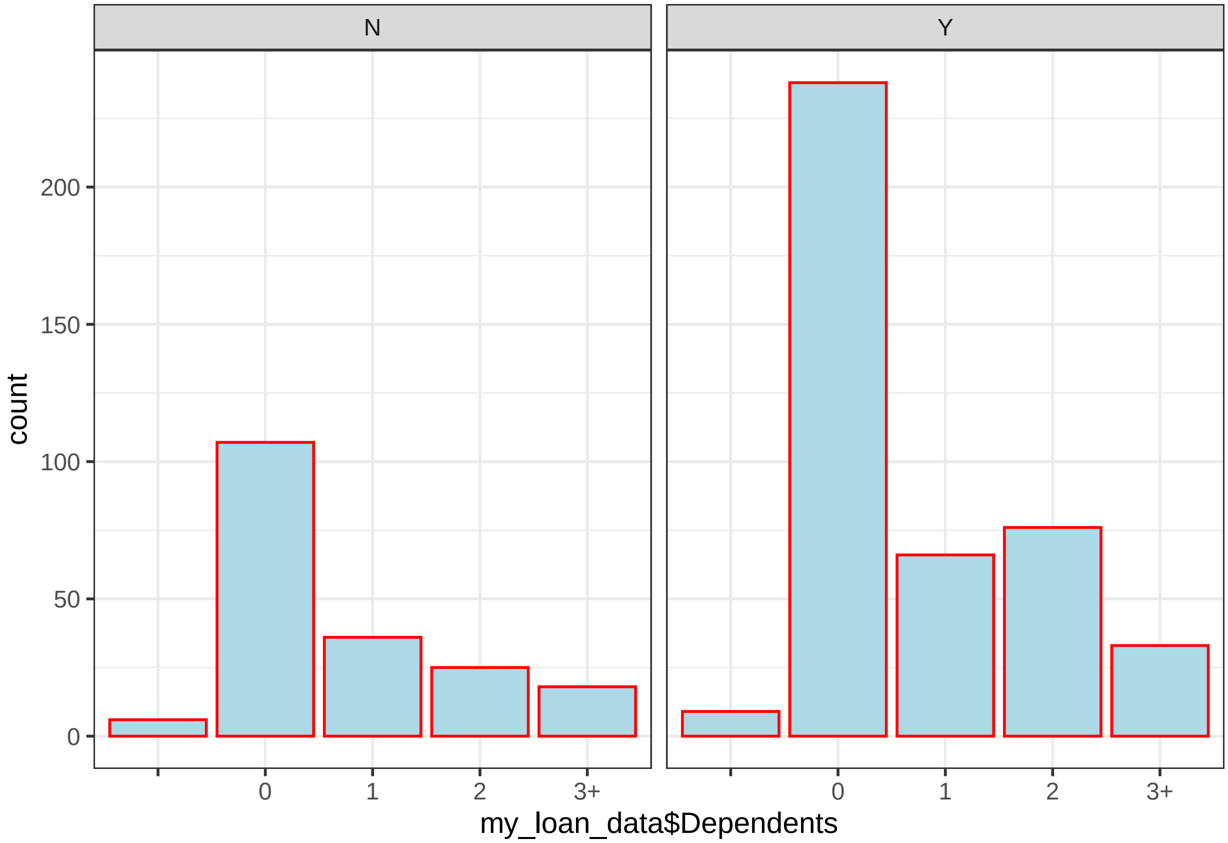
Number of Dependents:

```
summary(my_loan_data$Dependents)
```

```
##       0   1   2  3+
##  15 345 102 101  51
```

```
ggplot(data=my_loan_data, aes(my_loan_data$Dependents)) +
  geom_histogram(col="red",fill="lightblue",stat="count" ) +
  facet_grid(~my_loan_data$Loan_Status)+
  scale_x_discrete()+
  theme_bw()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



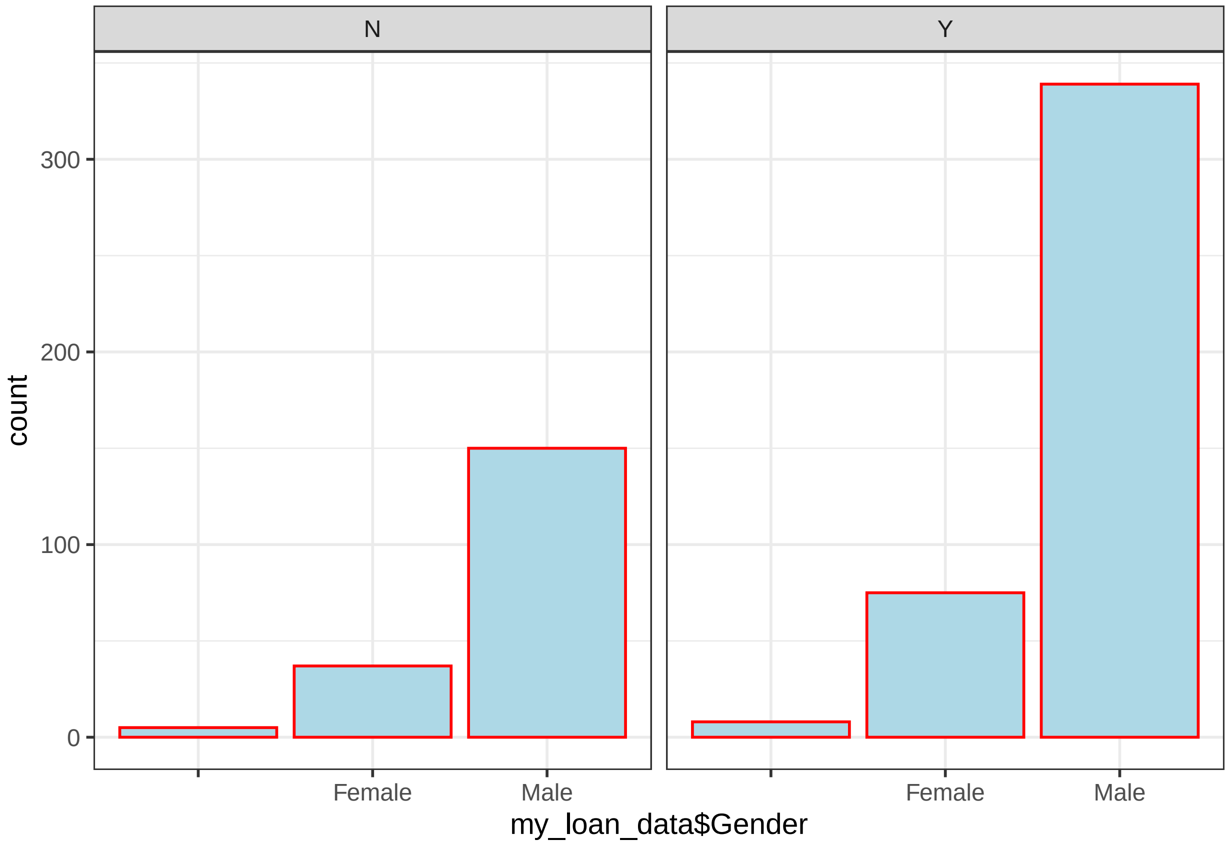
Gender:

```
summary(my_loan_data$Gender)
```

```
##        Female   Male
##     13    112    489
```

```
ggplot(data=my_loan_data, aes(my_loan_data$Gender)) +
  geom_histogram(col="red",fill="lightblue",stat="count") +
  facet_grid(~my_loan_data$Loan_Status)+
  scale_x_discrete()+
  theme_bw()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```


N
Y
count
300
200
100
0
Female
Male
Female
Male
my_loan_data$Gender