Mini Project_1

2 September 2017

Table 1: Group 1

StudentName	StudentID	Contribution
Bhupendra Singh	201452020	Studying Relationships between variables of dataset
Vikas Singh	201452047	"
Vipin Sahu	201452051	"
Yeshwanth Chinna	201452055	"
Anand Rahul	201452017	Outlier Detection(False and True)=> Graphs and Plotting
Kenneth Tenny	201452066	"
Parul Bindal	201452045	"
Lalit Singh	201452049	Simulations
Aditya Raj	201451038	"
Aniket Raj	201451027	Validation

R Markdown

This report contains:
a) After data cleansing, we have done a detailed analysis of how people of different regions, income group, age group and gender responded to whether they support demonetisation or not. It also contains the relationship between parameters that we have used for the analysing the above data.
b) Analysing the distribution of monthly income by assuming it to be a Gamma Distribution, validating our assumptions and finally, simulating from the Gamma Distribution we analysed.

Data Cleansing

First we imported the csv file into our environment as data frame. This data set contains 8 variables.
a) Some of the values in Demonetisation column contains "not Yes" which were converted to "No".

b) Some of the entries in Urban column were NA(Not Available) which were removed as per requirement during analysis.

c) Two of the entries in age column contains value 0,1130 which are false outliers and cannot be true in any case. We used these row data entry only when we wanted to find relationships and plot graphs on the other parameters of this data entry.

Problem 1:

Exploratory Data analysis: Figure are more than thousands of words

We have done visualistion to understand the relationship between various variables of data set. Here is the summary of our data set. The summary of pid column does not make any sense because its a unique identity number of each person.

```
library(ggplot2)
require(gridExtra)
```

```
## Loading required package: gridExtra
```

```
Demon <- read.csv("Demon.csv", header = TRUE)
attach(Demon)

# converting all the "not yes" to "No"

Demonitisation[Demonitisation == 'not Yes'] <- 'No'
Data_ <- Demon[,-8]
Data <- cbind(Data_, Demonitisation)

DataSum = summary(Data)


# converting continuous variable age into categorical variable for visualisation.
row <- which(age == 1130)
Data <- Data[-47,]
attach(Data)
```

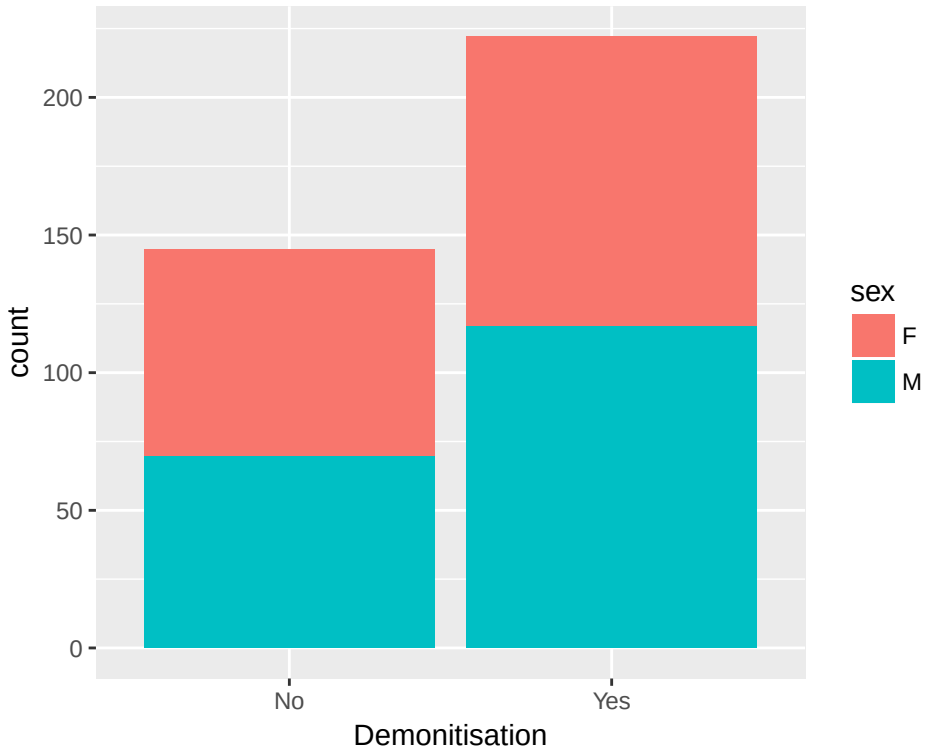
```
## The following object is masked _by_ .GlobalEnv:
##
##     Demonitisation
```

```
## The following objects are masked from Demon:
##
##     Demonitisation, Pid, Residence, Sl.No., Urban, age,
##     monthly.income, sex
```

```
agecat <- cut(age,c(18,40,60,80,100,136), labels = c("18-40", "40-60","60-80", "80-100",
                                                     "100-above"))
incomecat <- cut(monthly.income, c(-1,13000,23000,33000,43000,53000,63000,85000,500000),
                 labels = c("below 13k","13k-23k","23k-33k","33k-43k","43k-53k","53k-63k","63k-85k","85-

Data_T <- cbind(Data, incomecat,agecat)
# Transformed data which contains more categorical variable
Data_N <- Data[!is.na(Urban),]
#Urban_N <- Urban[!is.na(Urban)]

library(ggplot2)
NewDataSum = summary(Data_T)
```

Relationship between sex and Demonetisation

```
qplot(x = Demonitisation, data = Data, fill = sex, geom="bar")
```



Inference:

```
Data$Demonitisation <- droplevels(Data$Demonitisation)
sex_Demon<-with(Data, table(sex,Demonitisation))                ##### Sex Based Table
sex_Demon
```

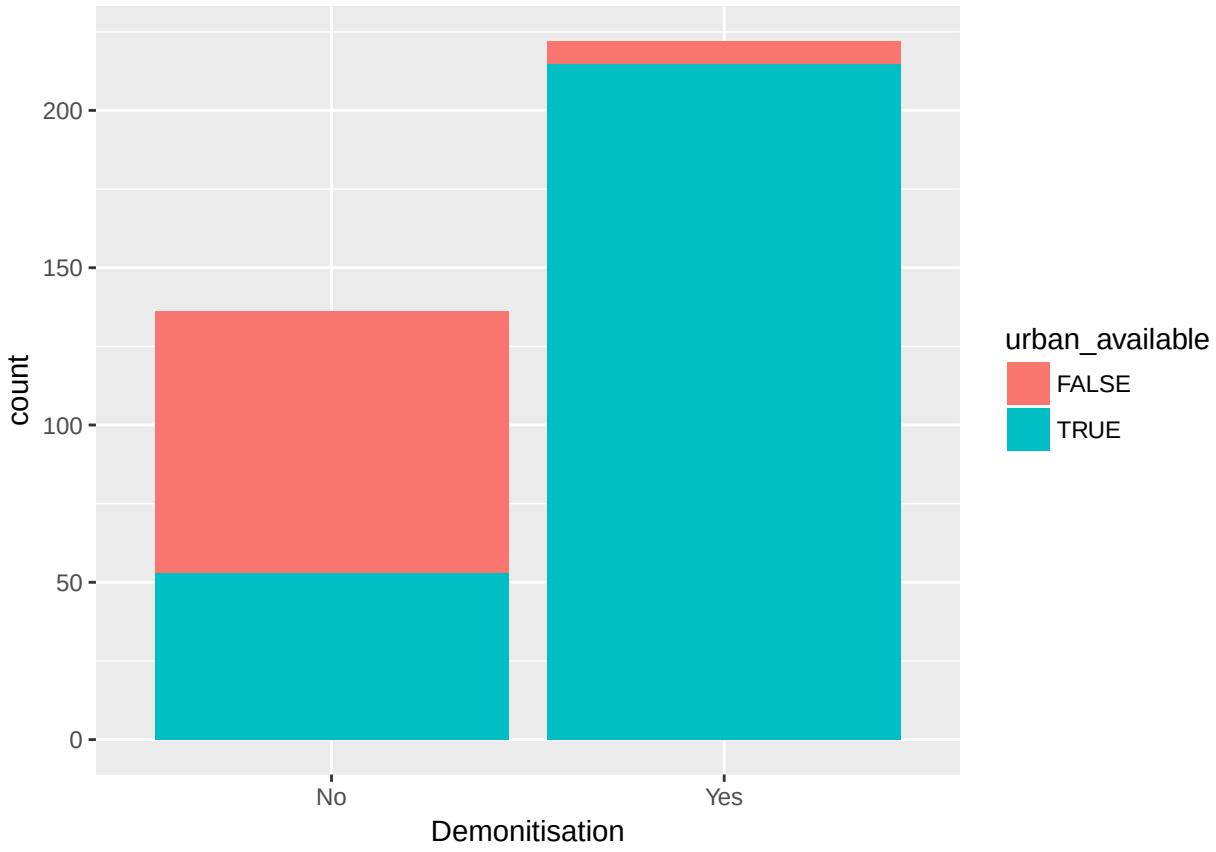
```
##    Demonitisation
## sex  No Yes
##   F  75 105
##   M  70 117
```

Inference:
1. 60% people are in favour of demonetisation. 2. 55% people who are in favour of Demonetisation are Men.
3. 52% people who are not in favour of Demonetisation are Women.

So, We can conclude that the majority of people who are in favour of Demonetisation are Men.

Relationship between Urban and Demonetisation

```
urban_available <- Data_N$Urban
qplot(x = Demonitisation, data = Data_N, fill = urban_available, geom = "bar")
```




```
urban_Demon<-with(Data, table(Urban,Demonitisation))
urban_Demon
```

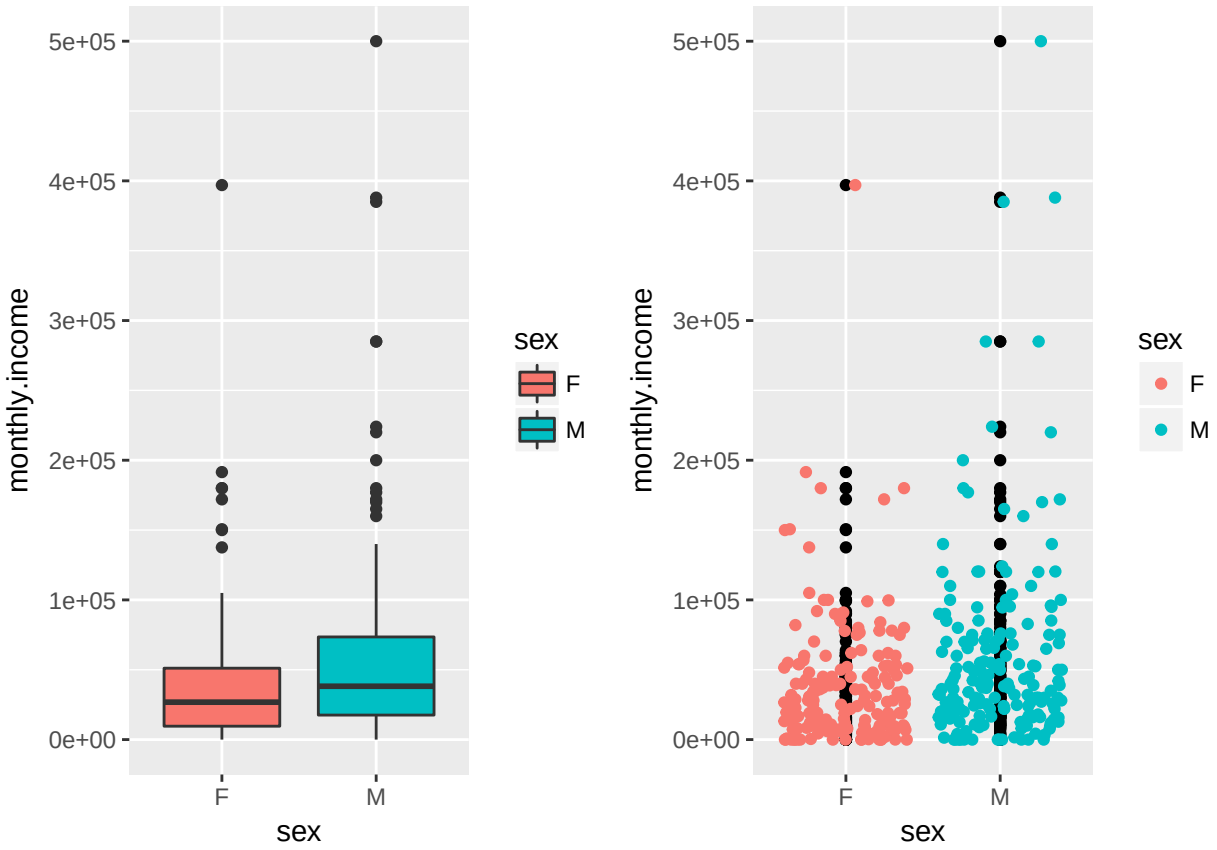
```
##        Demonitisation
## Urban     No Yes
##   FALSE  83   7
##   TRUE   53 215
```

Inference:
1. 96% people who are in favour of Demonetisation are living in Urban areas. 2. 61% people who are not in favour of Demonetisation are living in Rural Areas.

So, We can conclude that the majority of people who are in favour of Demonetisation are from Urban Areas.

Visualising Difference between male and female Average Income

```
require(gridExtra)
p1 <- qplot(x=sex, y = monthly.income, data=Data, fill=sex, geom = "boxplot")
p2 <- qplot(x=sex, y=monthly.income, data = Data) + geom_jitter(aes(color=sex))
grid.arrange(p1,p2, ncol=2)
```



Inference:

1. We can visualise the values of the (Minimum, Q1, Q2{Median}, Q3, Maximum) in the Box Plot after classifying the Gender in Data.
2. Although, There is no significstatistical difference between Incomes of Men and Women, There are more number of men in Q2-Q3(50-75 percentile) range, when compared to women. The number of Outliers in the above 75 percentile range are also more in case of Men

Relationship between Urban, sex and Demonetisation

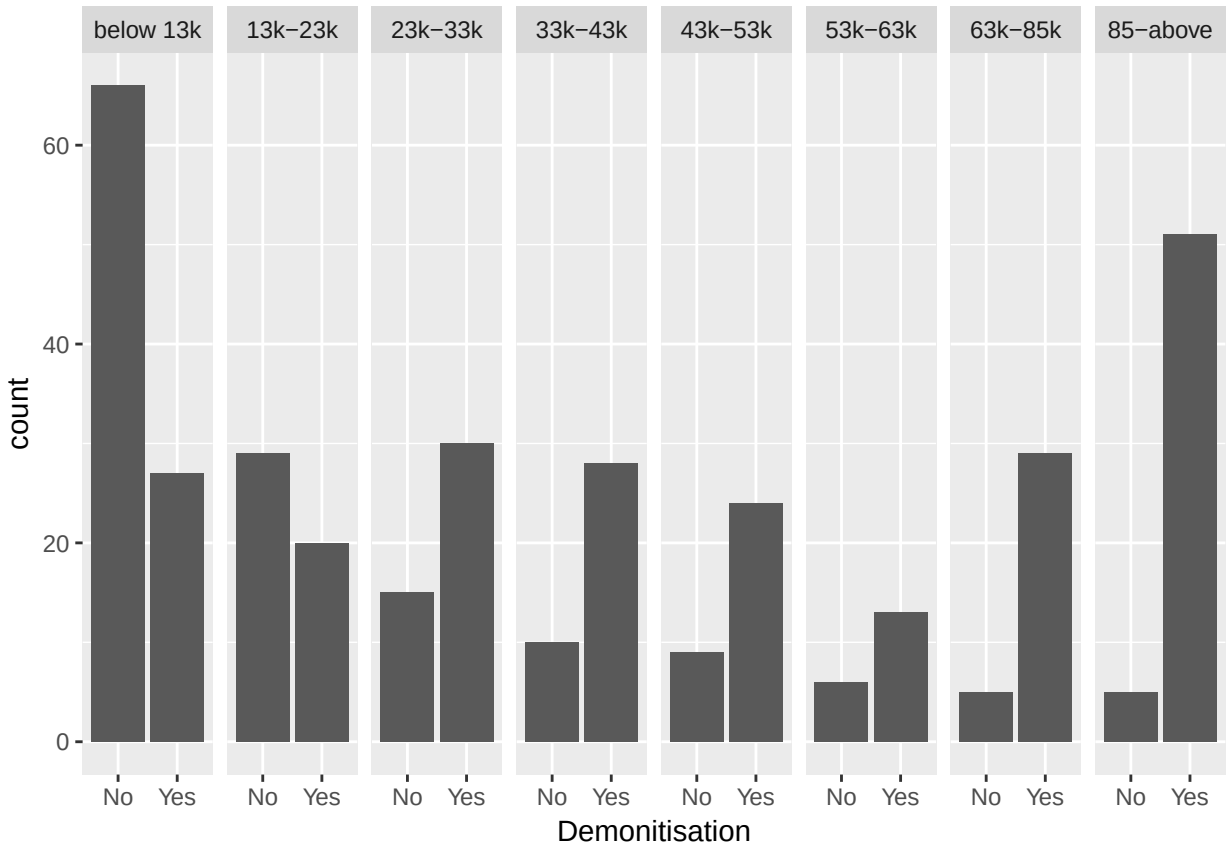
```
qplot(x = Demonitisation, data = Data[!is.na(Urban),], fill = sex,
      facets = .~Urban[!is.na(Urban)] )+ geom_bar()
```

Inference:

1. 59% of the people who are living in Rural areas and who are not in favour of Demonetisation are Women. 2.80% of the people who are living in Rural areas and who are in favour of Demonetisation are Women.
2. 58% of the people who are living in Urban areas and who are not in favour of Demonetisation are Men.
3. 56% of the people who are living in Urban areas and who are in favour of Demonetisation are Men.

Relationship between Income category and Demonetisation

```
qplot(x = Demonitisation, data = Data_T, facets = .~incomecat) + geom_bar()
```




```
income_Demon<-with(Data_T, table(incomecat,Demonitisation))     ##### Income Based Table
income_Demon
```

```
##            Demonitisation
## incomecat    No Yes not Yes
##   below 13k 66  27       0
##   13k-23k   29  20       0
##   23k-33k   15  30       0
##   33k-43k   10  28       0
##   43k-53k    9  24       0
##   53k-63k    6  13       0
##   63k-85k    5  29       0
##   85-above   5  51       0
```

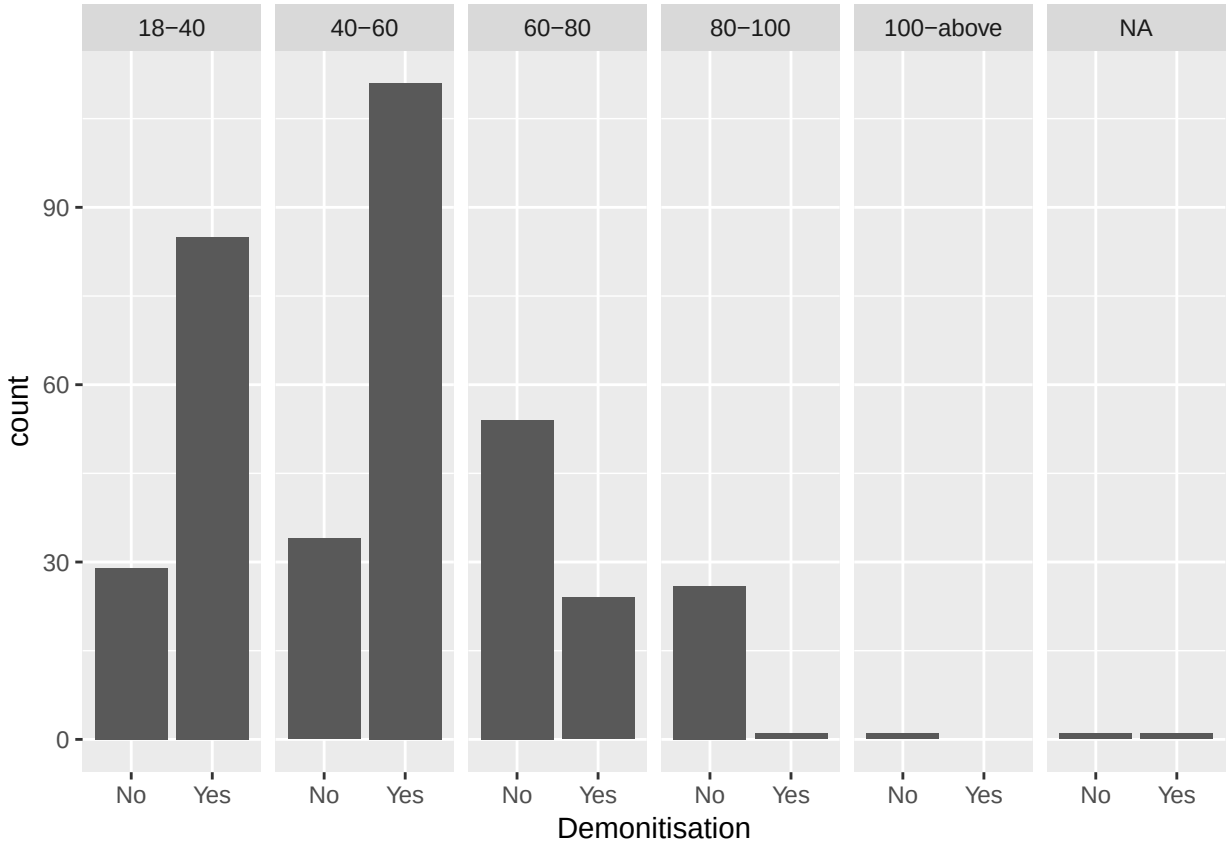
1. 26% of the people earning below Rs.13k support Demonetisation.
2. 41% of the people earning between Rs.13k and Rs.23k support Demonetisation.
3. 67% of the people earning between Rs.23k and Rs.33k support Demonetisation.
4. 74% of the people earning between Rs.33k and Rs.43k support Demonetisation.
5. 72% of the people earning between Rs.43k and Rs.53k support Demonetisation.
6. 70% of the people earning between Rs.53k and Rs.63k support Demonetisation.
7. 83% of the people earning between Rs.63k and Rs.85k support Demonetisation.
8. 91% of the people earning above Rs.85k support Demonetisation.

So, Of all the age groups, the people earning above Rs.85k have the maximum majority of the people supporting Demonetisation.

We can also say that as the incomes of the people are increasing, the percentage of the people in that income group supporting Demonetisation is also increasing.

Relationship between Age category and Demonetisation

```
qplot(x = Demonitisation, data = Data_T, facets = .~agecat) + geom_bar()
```




```
age_Demon<-with(Data_T, table(agecat,Demonitisation))
age_Demon
```

```
##            Demonitisation
## agecat       No Yes not Yes
##   18-40      29  85       0
##   40-60      34 111       0
##   60-80      54  24       0
##   80-100     26   1       0
##   100-above   1   0       0
```

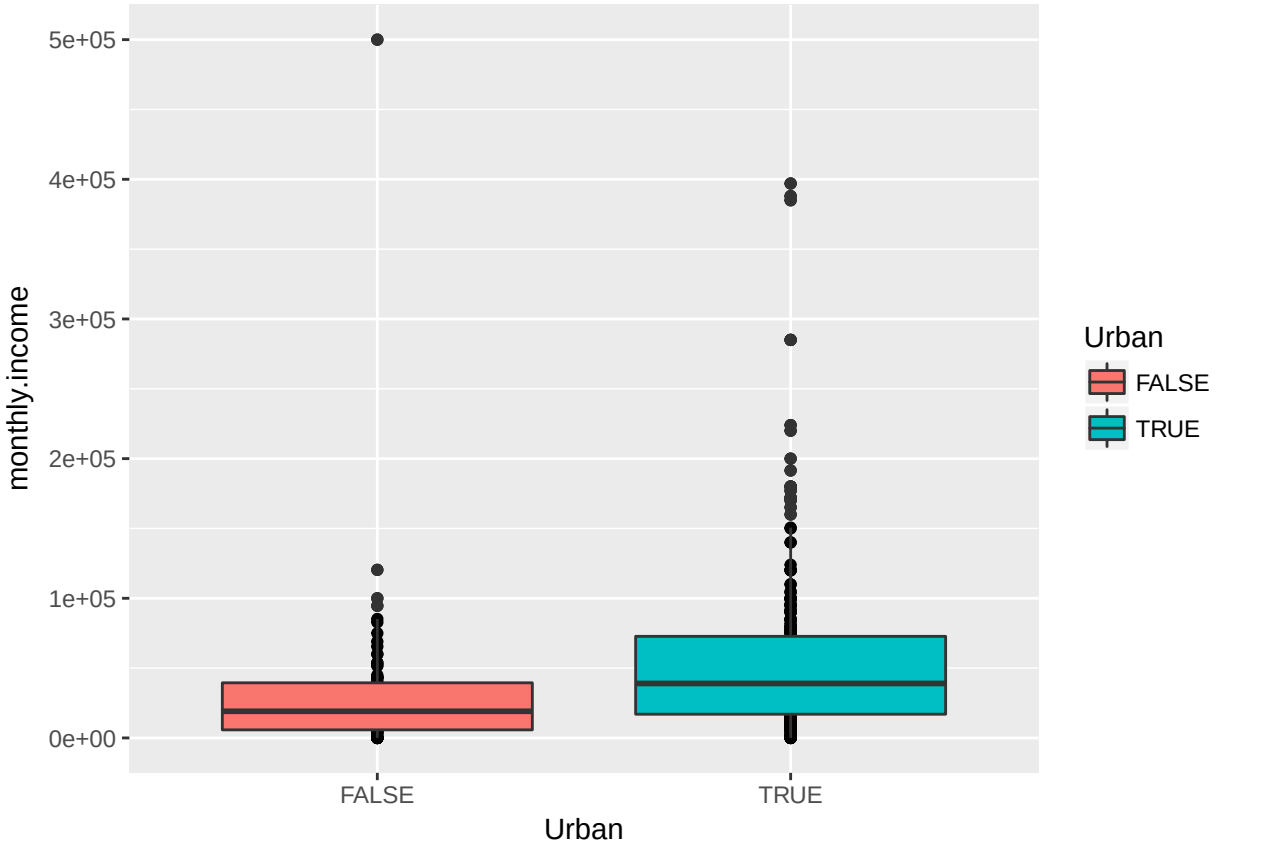
Inference:

1. 75% of the people having age between 18 and 40 support Demonetisation.
2. 76% of the people having age between 40 and 60 support Demonetisation.
3. 32% of the people having age between 60 and 80 support Demonetisation.
4. 7% of the people having age between 80 and 100 support Demonetisation.
5. 0% of the people having age above 100 support Demonetisation.

We can also say that as the Ages of the people are increasing, the percentage of the people in that Age range supporting Demonetisation is decreasing.

Visualising the difference between income of urban and rural people.

```
qplot(x = Urban, y = monthly.income, data = Data_N, fill=Urban) + geom_boxplot()
```



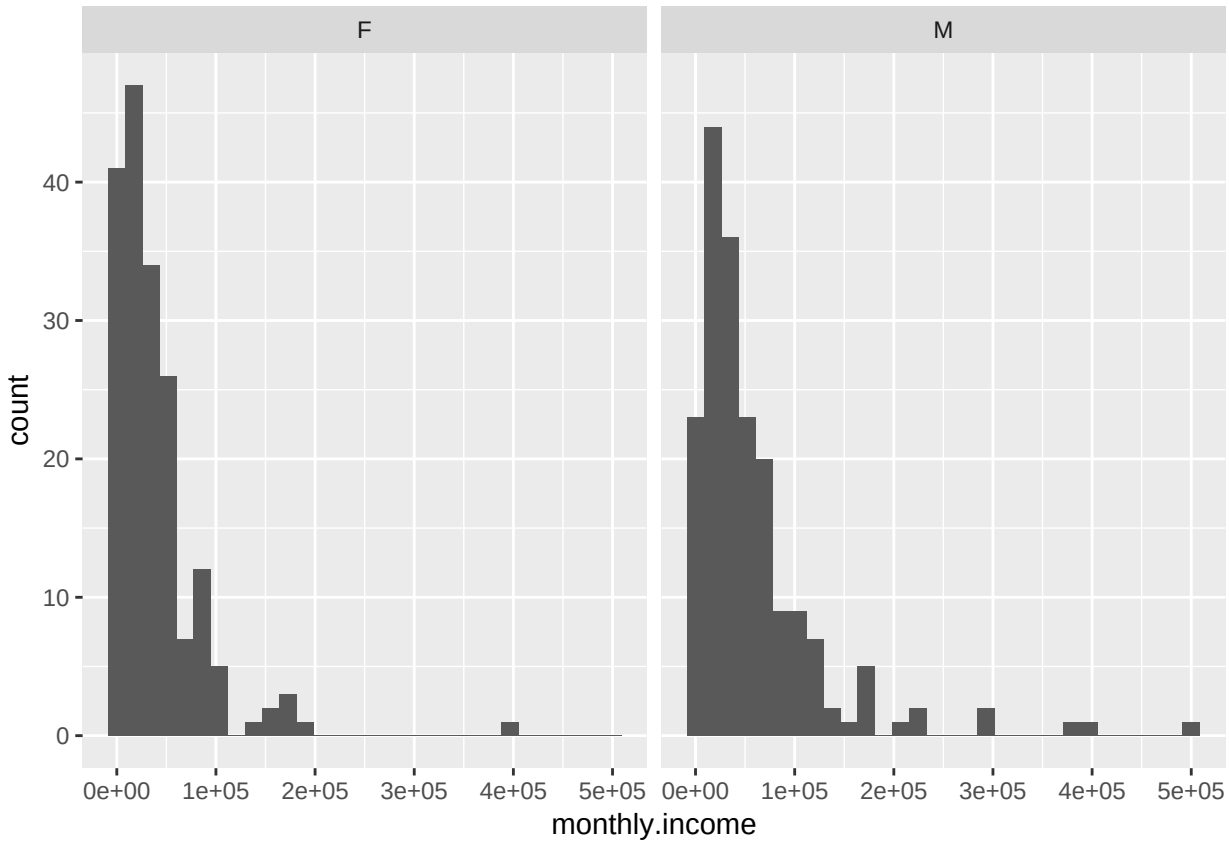
Inference:

1. There is a significant difference between the range of Incomes between Urban and Rural people.
2. Although the person earning the highest income belongs to the rural region, but The number of people earning more are significant in the case of Urban folks.

Visualising the distribution of income among male and female with histogram

```
qplot(x = monthly.income, data = Data_T, facets = .~sex) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



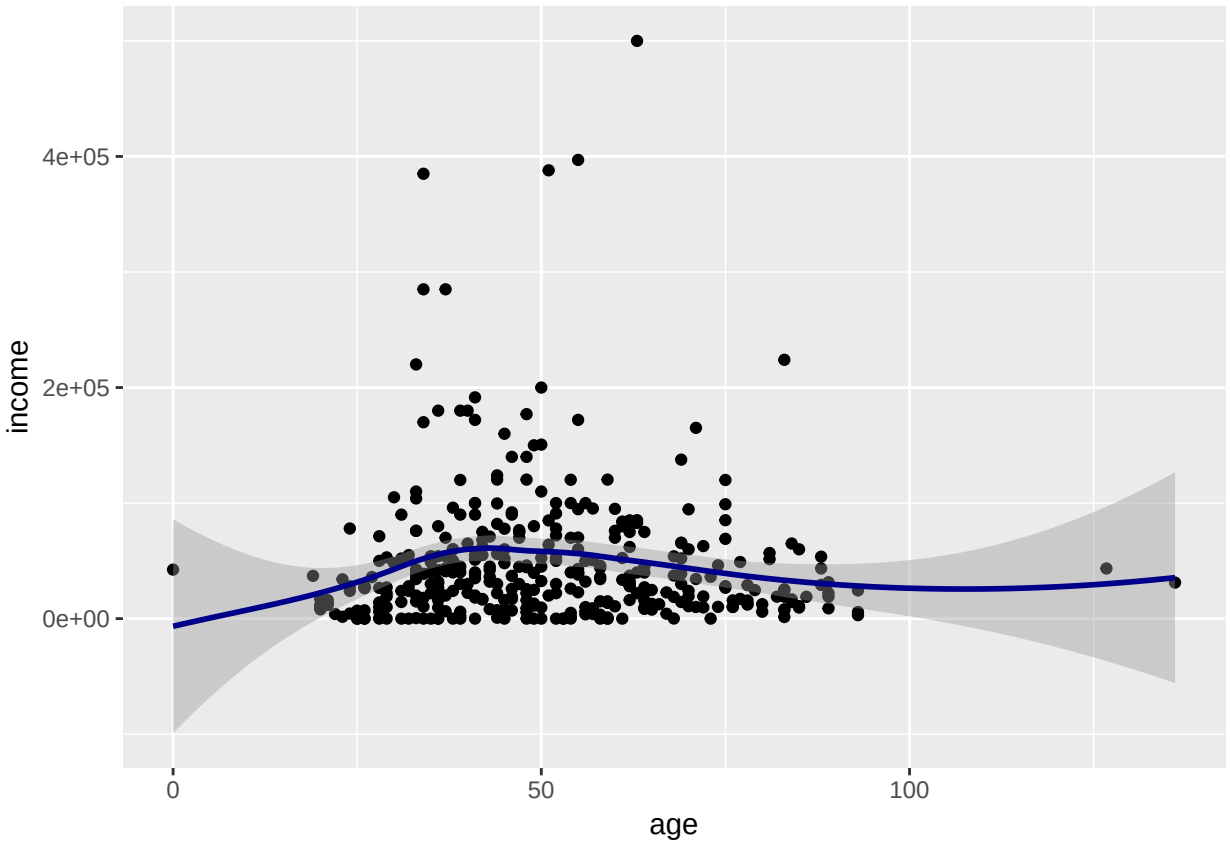
Inference:

1. The distribution of income approximately follows Gamma Distribution in case of Men and Women.
2. The distribution of the Income in case of Men is slightly skewed compared to Women.

Visualising the relationship between age and income

```
dat <- data.frame(age = Data_T$age,income = Data_T$monthly.income)
ggplot(data = dat, aes(x=age, y = income), colour = factor(sex))  +geom_point() +
  geom_smooth(col="darkblue")
```

```
## `geom_smooth()` using method = 'loess'
```



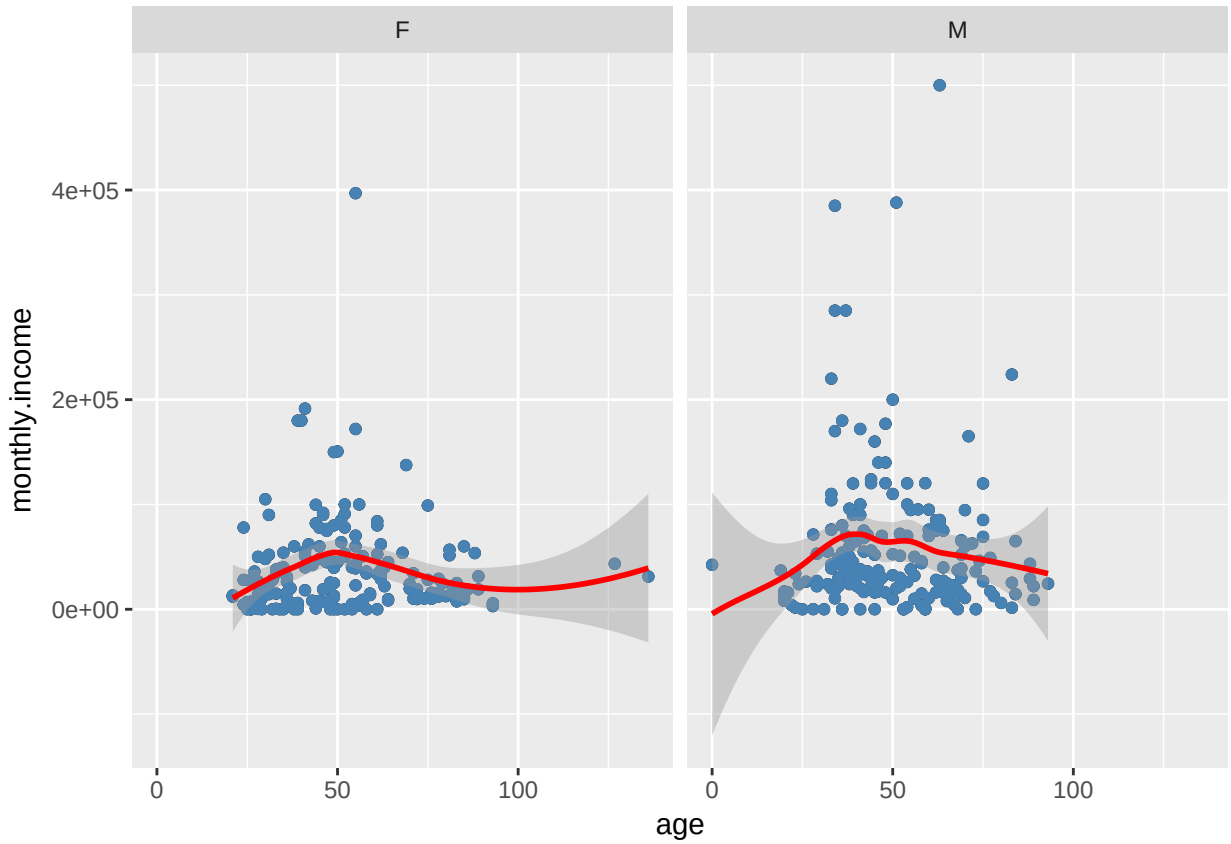
Inference:

1. There is no relation whatsoever when we compare the Age with the Income.

Visualising the realtionship between age and income among male and female

```
qplot(x=age,y=monthly.income, data=Data, facets = .~sex) + geom_point(col="steelblue") +  geom_smooth(c
```

```
## `geom_smooth()` using method = 'loess'
```

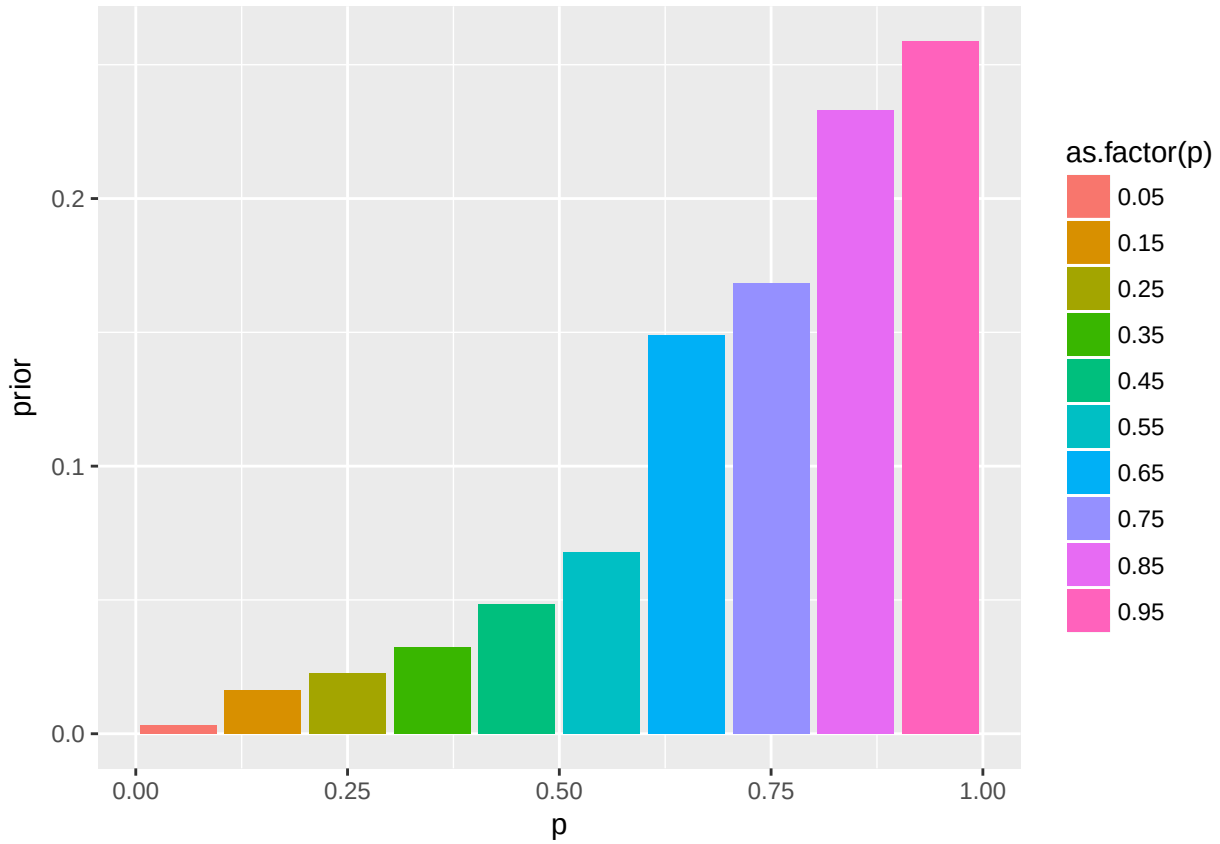


Inference:

1. Similar to the Histogram, Even this distribution of the incomes is somewhat similar in Males and Females.
2. When we compare the number of people earning more than the most of the people, we see that there are more males than compared to females.

Descrete distribution

```
library(LearnBayes)
p <-seq(0.05, 0.95, by = 0.1)
prior <- c( 0.1, 0.5, 0.7,1, 1.5, 2.1, 4.6, 5.2, 7.2, 8)
prior <- prior/sum(prior)
df<-data.frame(p,prior)
plt<-ggplot(df,aes(x=p,y = prior,fill=as.factor(p))) + geom_col()
print(plt)
```

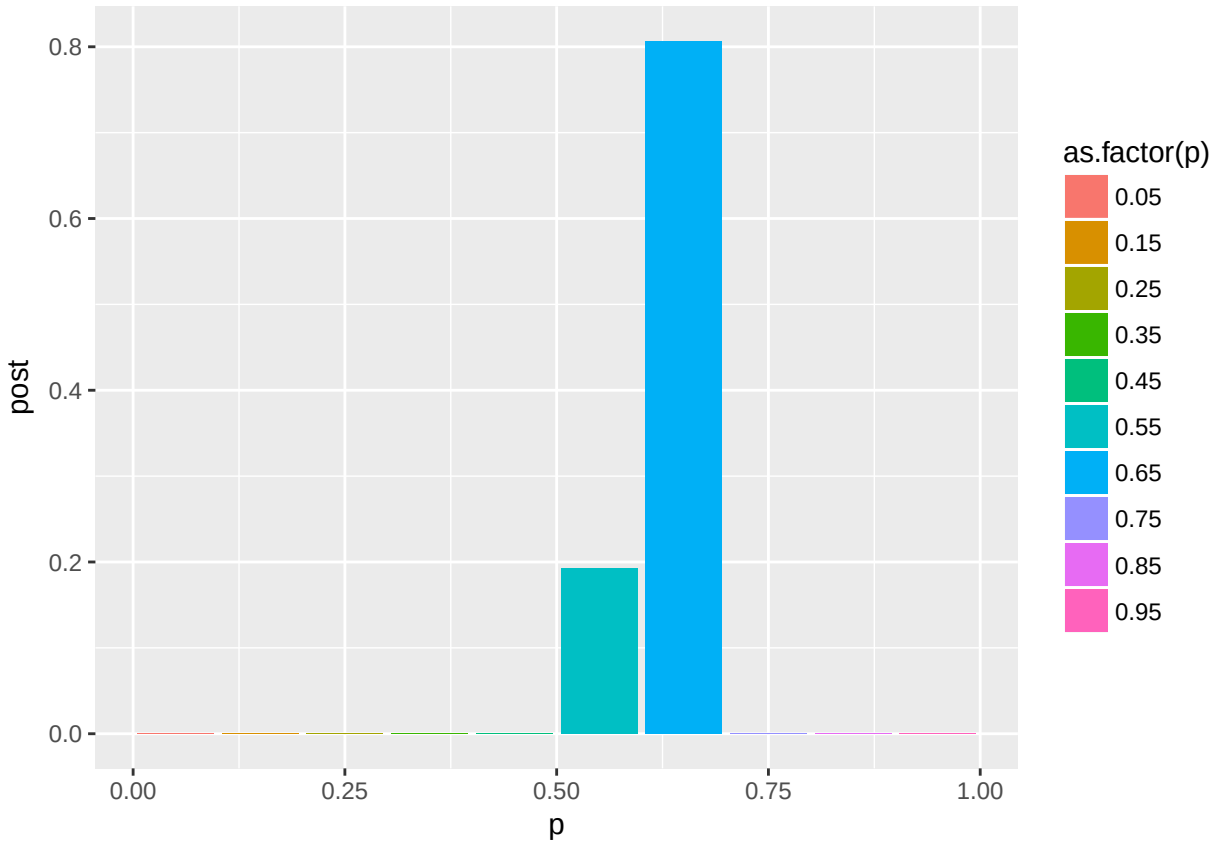



```
data<-c(nrow(Data[Data$Demonitisation=='Yes',]),nrow(Data[Data$Demonitisation=='No',]))

post = pdisc(p, prior, data)

df$post = post


plt1<-ggplot(df,aes(x=p,y = post,fill=as.factor(p))) + geom_col()
```

```
print(round(cbind(p, prior, post),2))
```

```
##          p prior post
##  [1,] 0.05  0.00 0.00
##  [2,] 0.15  0.02 0.00
##  [3,] 0.25  0.02 0.00
##  [4,] 0.35  0.03 0.00
##  [5,] 0.45  0.05 0.00
##  [6,] 0.55  0.07 0.19
##  [7,] 0.65  0.15 0.81
##  [8,] 0.75  0.17 0.00
##  [9,] 0.85  0.23 0.00
## [10,] 0.95  0.26 0.00
```

```
print(plt1)
```



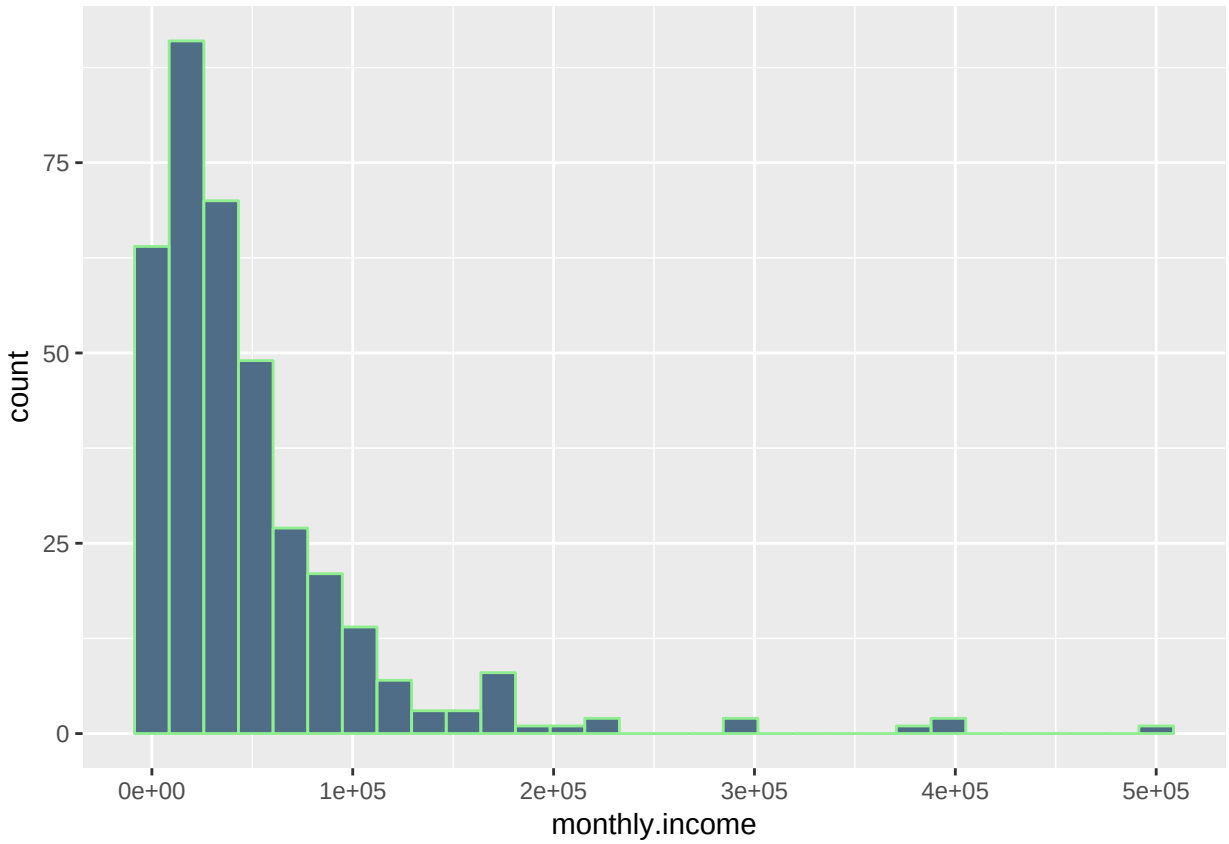
Problem 2:

Analysing the distribution on monthly income

We extracted monthly income parameter from the data set and analyze it to calculate the population parameter and validate its distribution as mentioned in the problem. First we analyze its distribution with histogram.

```
qplot(x = monthly.income) + geom_histogram(fill=
                              "steelblue", alpha = 1/2, col ='lightgreen')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The data seems to be highly skewed so there are possible chances of outliers which may disturb our parameter estimations. We analyze the outliers with three methods mentioned below:

Mehtod:1 Removing the data after 95th percentile

```
out <- quantile(monthly.income, 0.95)
monthly.income_ <- monthly.income[monthly.income < out]
data_h1 = data.frame(x = monthly.income_)
ggplot(data = data_h1, aes(x=monthly.income_, y = ..density..)) +
  geom_histogram(fill = "steelblue",
          alpha = 1/2) + geom_density(col="red", lwd=1.5)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


```
# Histogram showing data distribution after removing
#the lower and upper percentile vlaue data.
```

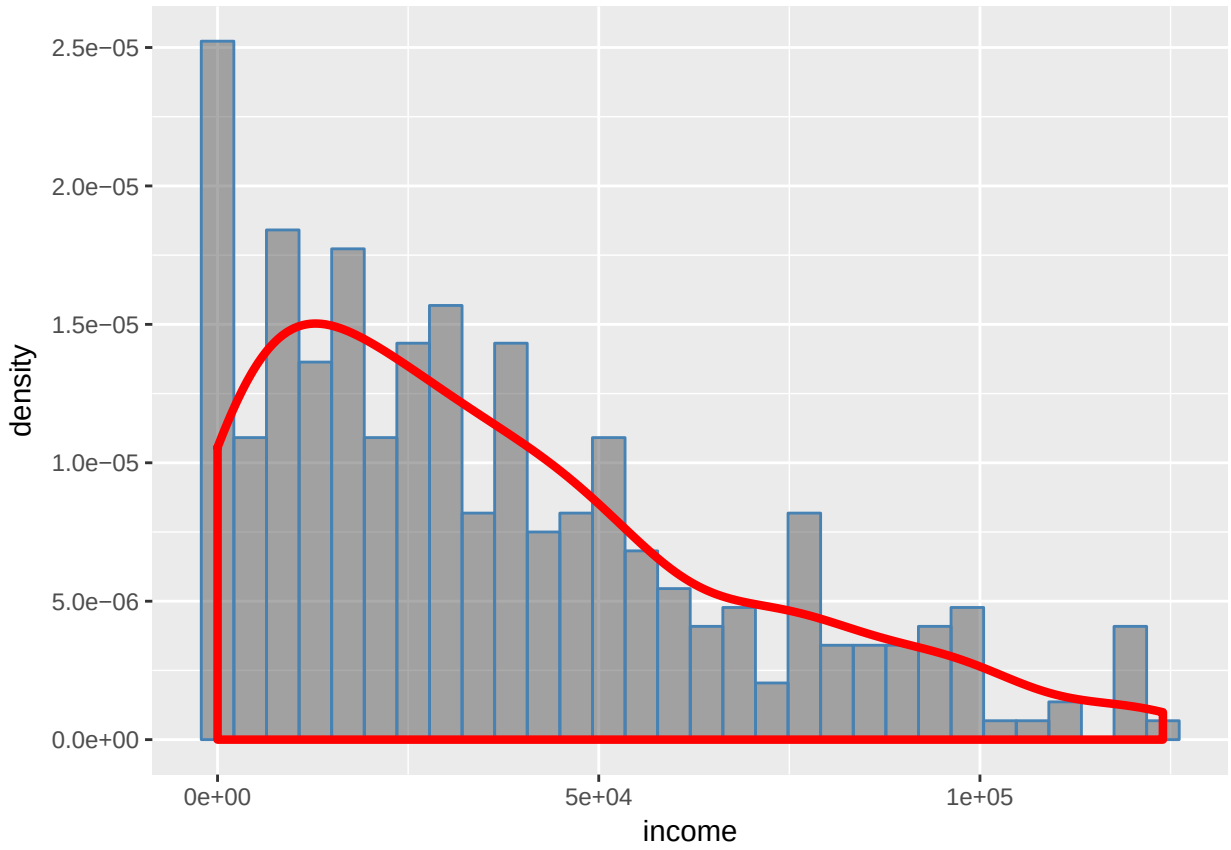
After removing all the data after 95th percentile, we checked the distribuiton. We found that the shape of distribution was distorted and a large number of sampel data was to be removed. Clipping the outliers to 95th percentile seems to distort the shape of tail. So we discarded this method.

Method:2 Retaining the 3/2 times Inter Quantile Range data This is the most general method for checking outliers. However its not appropiate for highly skewed data. We set the upper limit of data to be 3/4th quantile plus 3/2 times IQR and lower limit to be 1/4th minus 3/2 times IQR. This is general setting in most of the cases. However, removing outliers with this method badly distorted the shape of distribution as seen in below figure.

```
Q1 = quantile(monthly.income, 1/4)
Q2  = quantile(monthly.income, 3/4)

IR = quantile(monthly.income, 3/4) - quantile(monthly.income,1/4)
upper_limit = Q2 + 3/2*IR
lower_limit = Q1 - 3/2*IR

Income_ <- monthly.income[monthly.income < upper_limit  ]
data_h2 = data.frame(income = Income_)
ggplot(data = data_h2 ,aes(x = income, y= ..density.. )) +   geom_histogram(col="steelblue", alpha = 1/
    geom_density(col="red", lwd = 1.5)
```

```
## Warning: The plyr::rename operation has created duplicates for the
## following name(s): (`colour`)
```

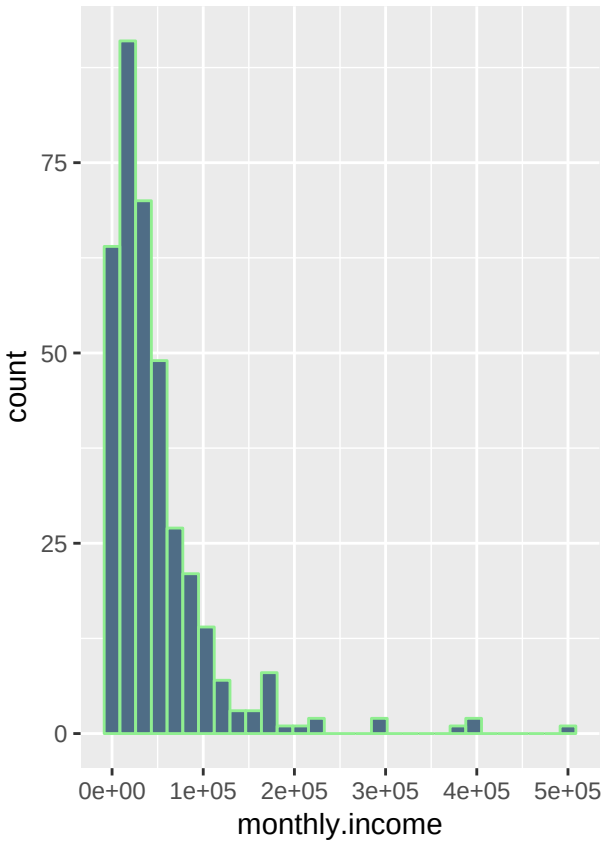
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

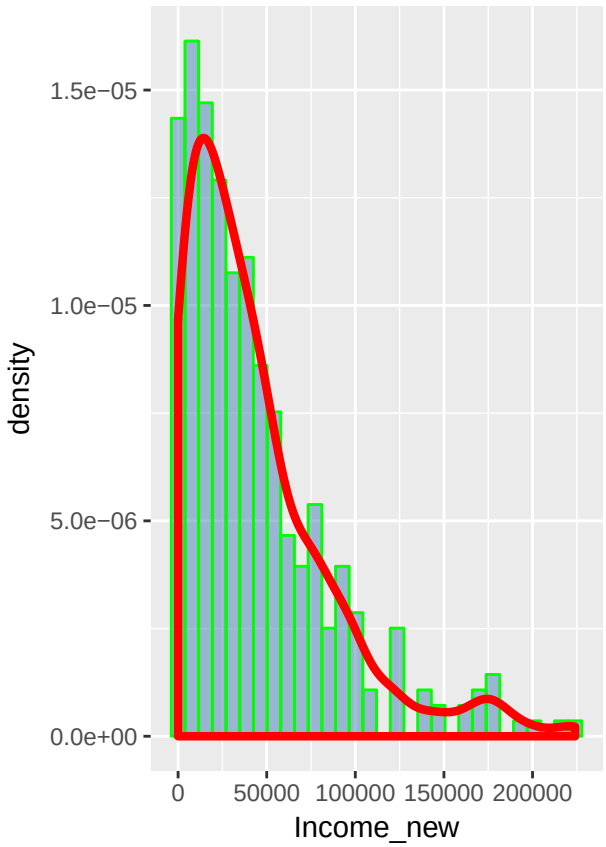


Method:3 Manually setting the limit for outliers The above two methods did not work wo we set the upper limit for outliers manually by closely observing the histogram and the distribution we get after removing the outliers was quite simillar to the original one.

```
s3 <- qplot(x = monthly.income) + geom_histogram(fill= "steelblue", alpha = 1/2,
                                                 col ='lightgreen')
# Histogram to check the outliers.
Income_new <- monthly.income[monthly.income < 250000]
# Outliers selected manually after observing the histogram.
data_s4 = data.frame(x = Income_new)
s4 <- ggplot(data=data_s4, aes(x=Income_new, y = ..density..)) + geom_histogram(
  alpha = 1/2, col = "green", fill="steelblue") +
                        geom_density(col="red", lwd=1.5)
grid.arrange(s3,s4, ncol=2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

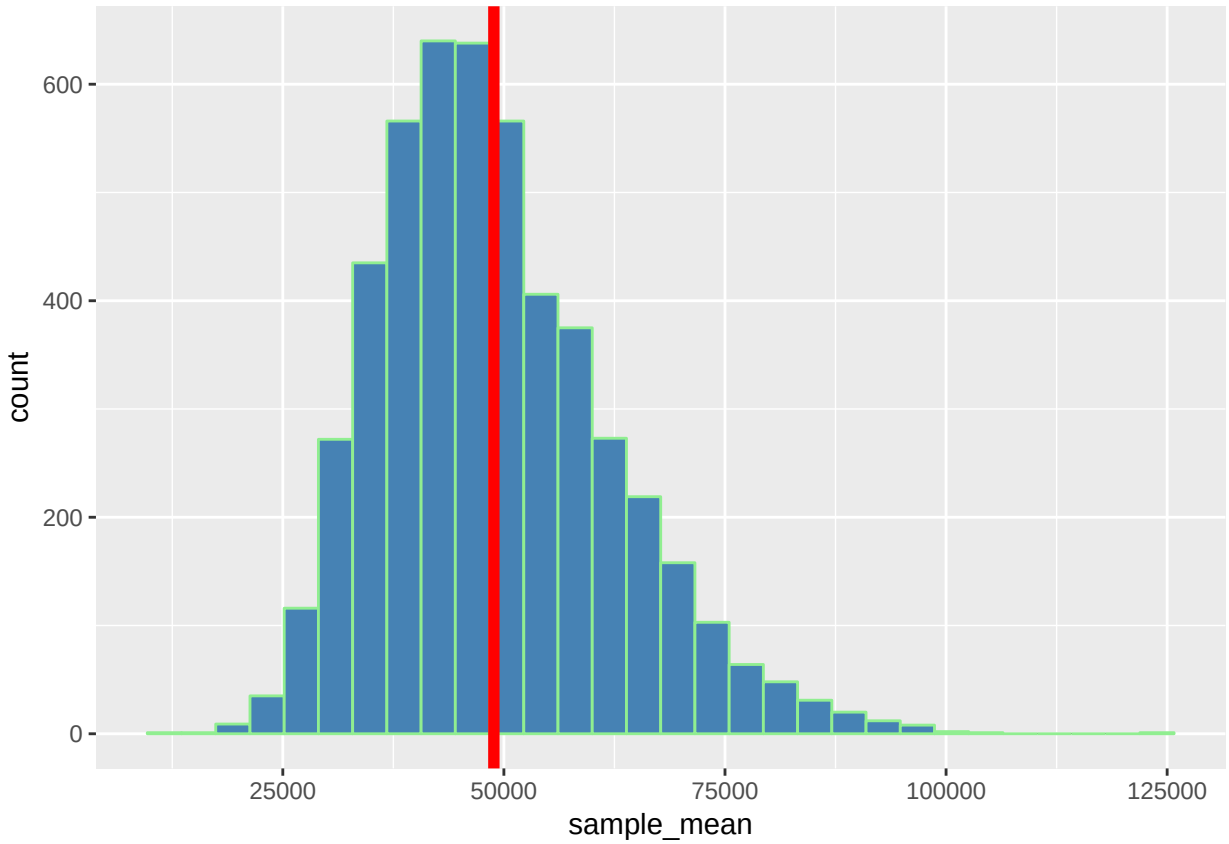




Prameter Estimation with Central Limit theorem After removing these outliers from the data,we estimated our population parameter through central limit theorem. The confidence interval of mean is (43001.50 54726.98). The estimated mean is 48864.24 . The below figure shows the distribution of sample mean and the dark vertical line shows the point estimate.

```
set.seed(1)
x <- Income_new
n = 5000 # number of samples to be collected
sample_mean = rep(NA, n) # vector to store all the sample means
for (i in 1:n){
  sample_mean[i] = mean(sample(monthly.income, 20))
}

population_mean_1 = mean(sample_mean) # From the central limit theorem
std_error = sd(sample_mean)/sqrt(20)

CI_ = population_mean_1 + c(-1, 1)*2*std_error  # confidence interval
qplot(x=sample_mean) + geom_histogram(col="lightgreen",
                                      fill="steelblue") +
  geom_vline(xintercept = population_mean_1, lwd = 2, col="red")
```

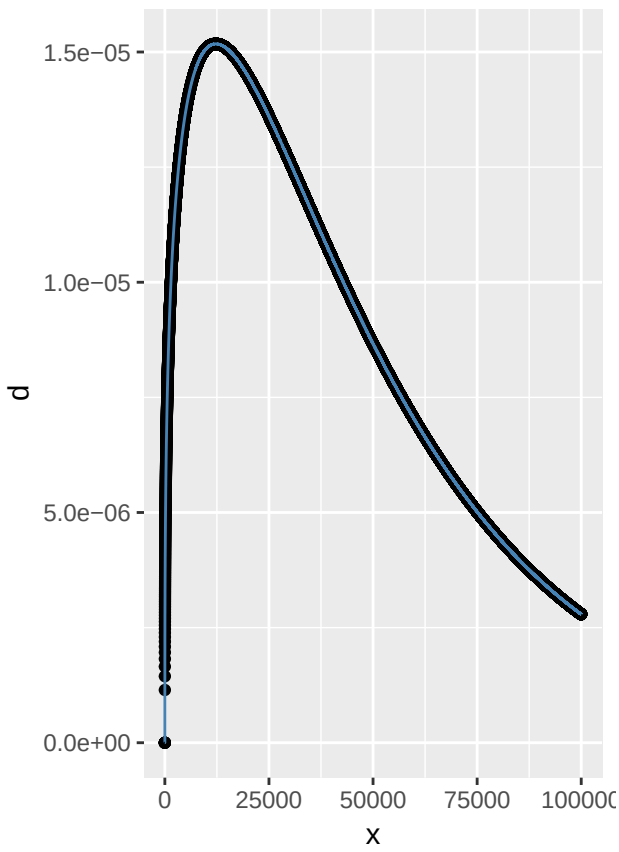
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

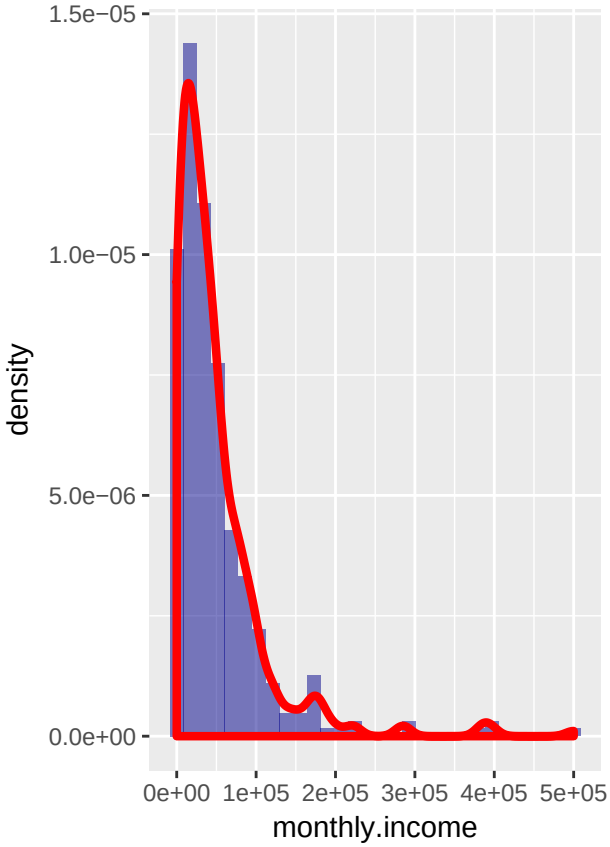


Since we calculated the mean and variance of the distribution, we were able to calculate the parameter of gamma distribution with the formul rate = mean/variance, shape = beta* mean. Below is the plot for theoretical distribution and sample population distribution.

```
var <- var(Income_new)
mean <- population_mean_1

beta = mean/var # shape of gamma distribution
alpha = mean*beta # rate of the gamma distribution

# plotting gamma distributions using the calcualted parameter
x = seq(-10,100000, by = 2)
d <- dgamma(x, shape = alpha, rate = beta)
p1 <- qplot(x = x, y = d) + geom_line(col = "steelblue")

p2 <- ggplot(data=Data, aes(x=monthly.income, y=..density..)) + geom_histogram(
  fill="darkblue",
            alpha=1/2) + geom_density(col="red",lwd=3/2, alpha=1/5)
require(gridExtra)
grid.arrange(p1,p2, ncol=2)
```

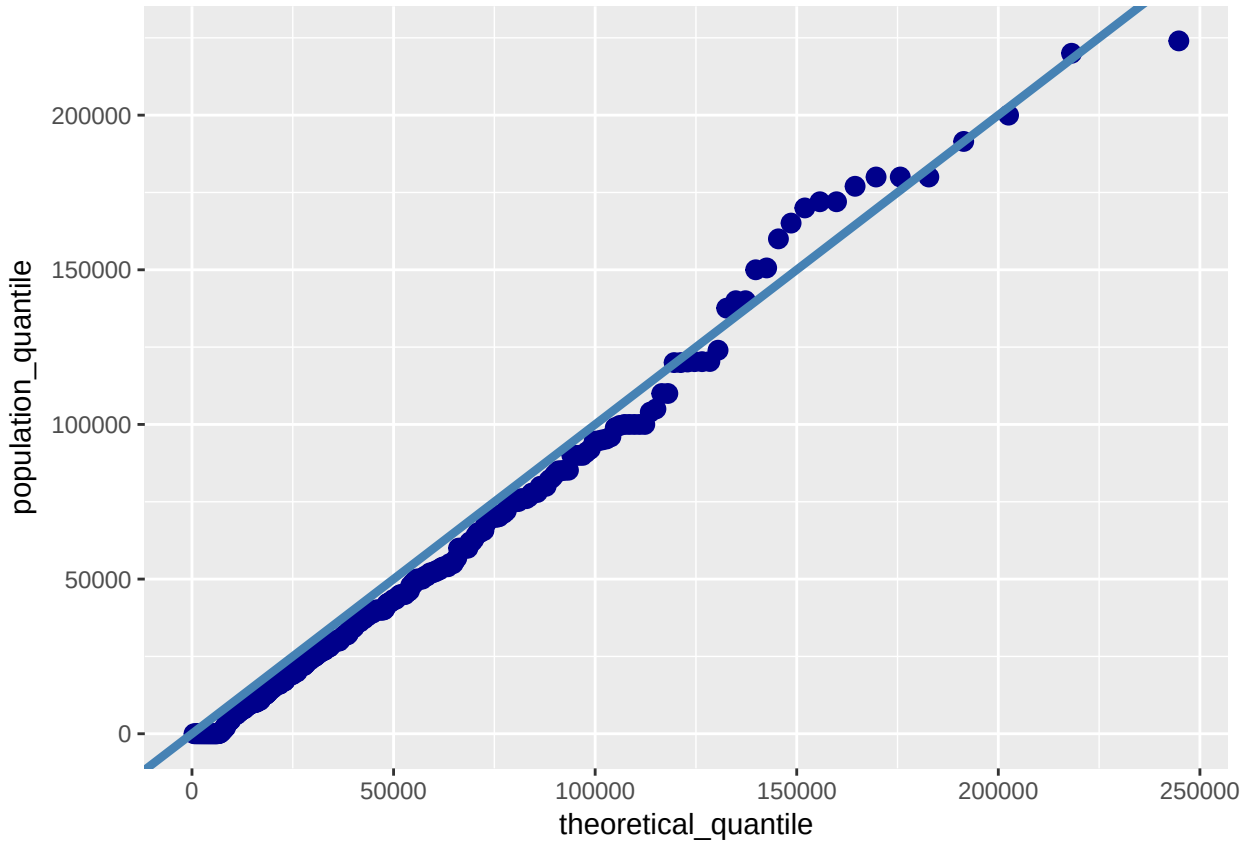
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





Validating our assumptions By observing the above plot we can conclude that our assumptions of gamma distribution was correct. However for getting sure, we plot its QQ plot. QQ plot is the of theoretical quantile of distribution on y axis and sample quantile on x axis.

```
n = length(Income_new)
probablities = 1:n/(n+1)
t_quantiles = qgamma(probablities, shape = alpha, rate = beta)
theoretical_quantile = sort(t_quantiles)
population_quantile = sort(Income_new)

qplot(y = population_quantile, x = theoretical_quantile) + geom_point(
  col="darkblue",size= 3) + geom_abline(slope = 1, intercept = 0, col="steelblue", lwd=1.5)
```




```
# The above plot validate our assumption of distribution being gamma as
# the line is almost straight.
```

The above QQ plot is almost straight line passing through origin which confirm our assumption. Our theoretical quantile matches smaple quantile with slight deviation.

Parameter Estimation with Maximum Likelihood Estimates

We estimated the parameter for gamma distribution with maximum likelihood technique. The parameters calculated were slightly different from above calculation.

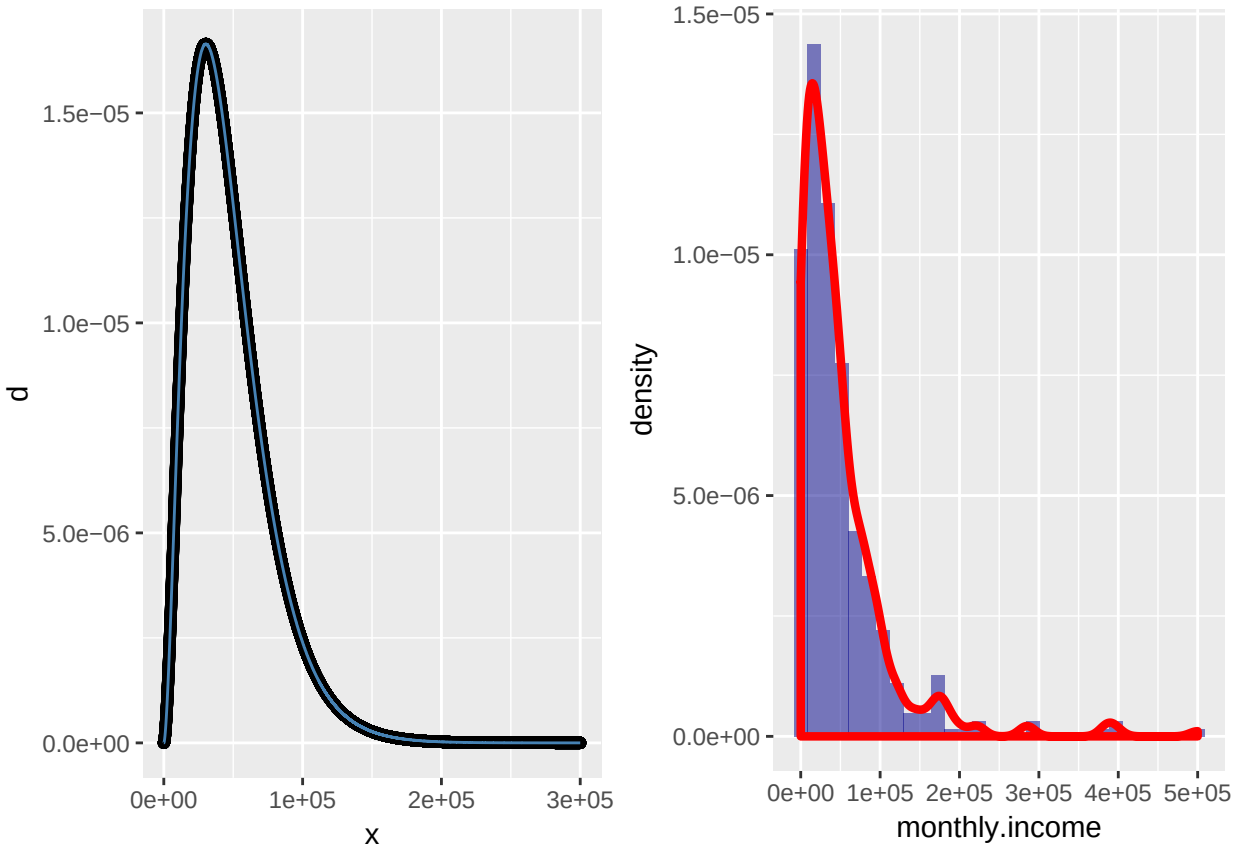
```
library(stats4)
y <- Income_new[Income_new > 0]
set.seed(1)
 LL <- function(par, data){
   R = dgamma(data, shape=par[1], rate = par[2], log = TRUE)
  return(-sum(R))
 }

mle <- optim(par=c(2, 1),fn = LL, data = y)

mle$par
```

```
## [1] 2.7567243224 0.0000579406
```

The estimated parameter for alpha and beta was 2.7567243224 0.0000579406 respectively. We plotted the gamma distribution corresponding to this parameter and compared with our original distribution.

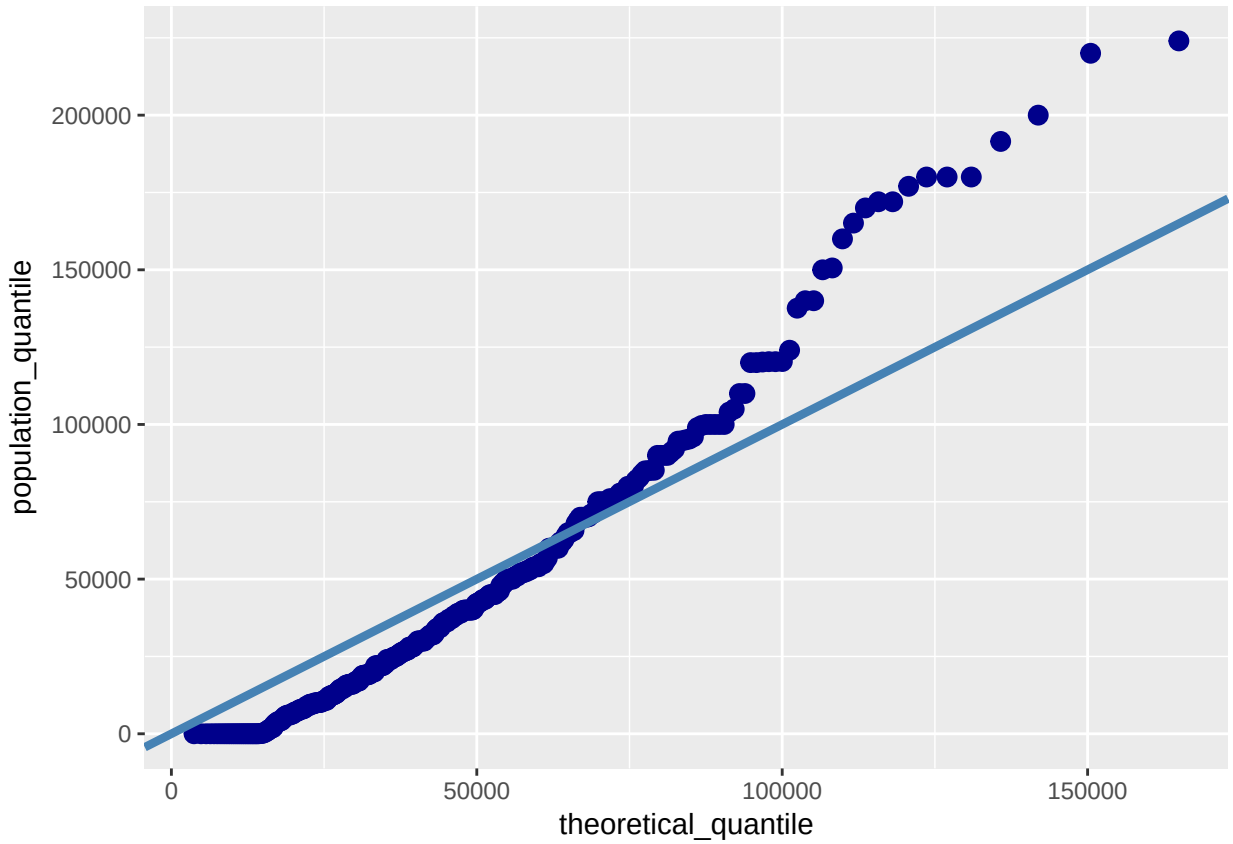
```
alpha2 = mle$par[1]
beta2 = mle$par[2]
x = seq(0,3e+05, by = 10)
d <- dgamma(x, shape = alpha2, rate = beta2)
p1 <- qplot(x = x, y = d) + geom_line(col = "steelblue")

p2 <- ggplot(data=Data, aes(x=monthly.income, y=..density..)) + geom_histogram(fill="darkblue",alpha=1/
    geom_density(col="red",lwd=3/2, alpha=1/5)
require(gridExtra)
grid.arrange(p1,p2, ncol=2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We also plotted QQ plot for this parameter estimate to validate our assumption

```
alpha1 = mle$par[1]
beta1 = mle$par[2]

n = length(Income_new)
probablities = 1:n/(n+1)
t_quantiles = qgamma(probablities, shape = alpha1, rate = beta1)
theoretical_quantile = sort(t_quantiles)
population_quantile = sort(Income_new)

qplot(y = population_quantile, x = theoretical_quantile) + geom_point(col="darkblue",    size= 3) + geo
```

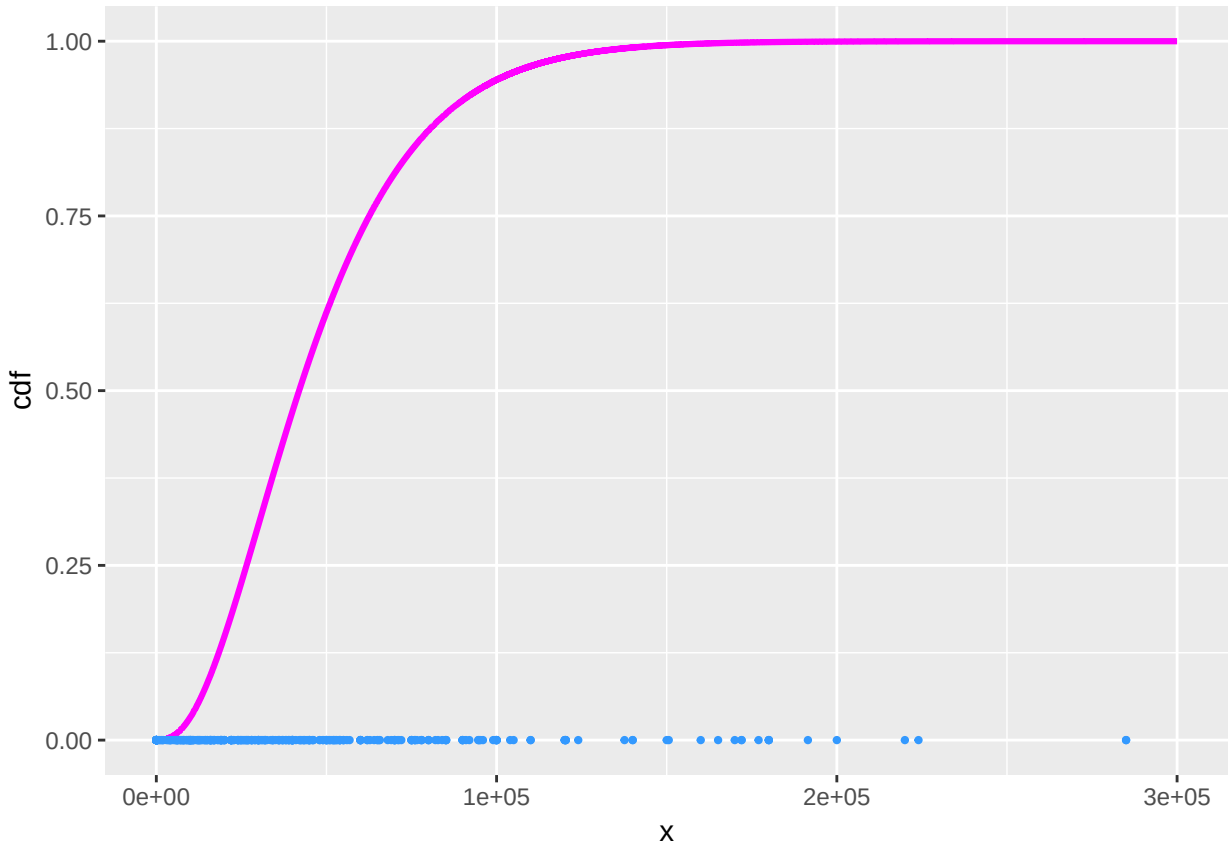



```
# The above plot validate our assumption of distribution being gamma as the line is almost straight.
```

The QQ plot was deviated from straight line with these parameters estimate as compared to parameters estimated with central limit theorem. Hence our method of parameter calculation was more accurate. This might be due to the reason that the maximum likelihood estimated are affected with the starter value of optimization. We might have been not able to pick good starter value. However we try many combinations.

Validating with CDF plot

```
demon_wo <- Data[monthly.income<=3e+05,]
cdf<-pgamma(x,alpha1, beta1)
df<-data.frame(x,cdf)
plt <- ggplot()
plt <- plt + geom_line(lwd = 1.1,data =df ,aes(x=x,y=cdf), color="#ff00ff")
plt <- plt + geom_point(size = 0.8,data = demon_wo,aes(x=monthly.income,y=0),color="#3399ff")
print(plt)
```

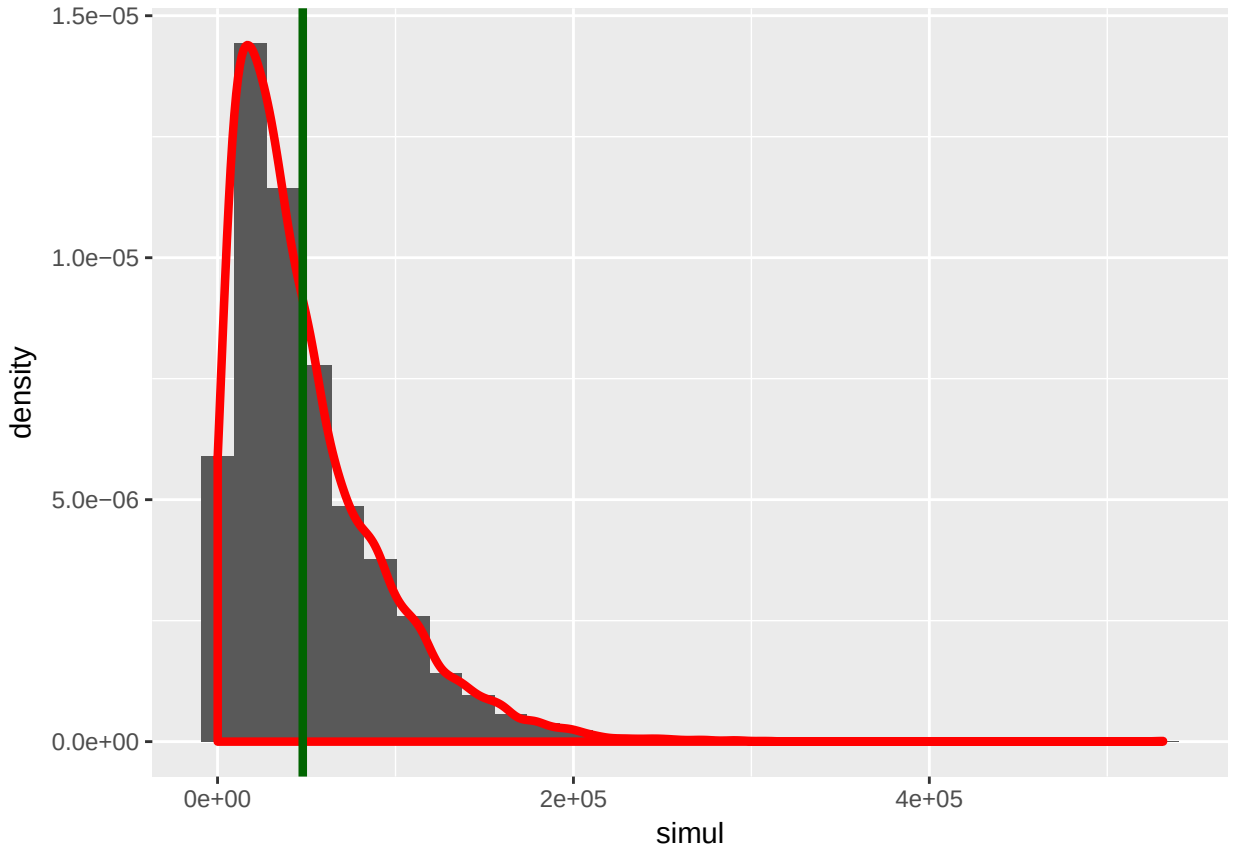


Simulations:

Below is the simulation of of 5000 samples with parameter for gamma distribution calculated using above methods.

```
simulations = 5000
prob = rgamma(simulations,shape = alpha, rate=beta)
perc_60 = quantile(prob, 0.6)
#h <- hist(prob,breaks=500,plot=FALSE)
#cuts <- cut(h$breaks, c(-100,perc_60,'Inf'))
cuts1 = cut(prob, c('-Inf',perc_60,'Inf'),c("Before 60the percentile",
                                            "after 60th percentile"))
data_prob = data.frame(simul=prob)
ggplot(data=data_prob, aes(x=simul, y = ..density..))+geom_histogram()  +geom_density(col="red", lwd = 
  geom_vline(xintercept = perc_60, col="darkgreen", lwd = 1.5)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The 60th percentile of the simulations of simulation comes out to be 47995.93. The vertical green line in the above maks tha data point of 60th percentile in the above density plot of distribution.

** The End **