Data Analytics

Mini Project 1 - Report

[Group - 8]

Date: 05/09/2017

Team

Name	Student ID	Role
Jitendra Singh	201452005	R Programming,Solved Q2
Anshul Dawar	201452059	Solved Q1 - Report
Prem Chand Saini	201452019	Formalization of Q1
Aarush Anand	201452028	Formalization of Q1
Ravi Kumar Patel	201452056	Solved Q2 - Report
N Santosh	201451063	R Programming
Rajat Dangi	201451038	Formalization of Q1 - Report
Ankit Kumar	201451065	Formalization of Q1 -Report
Prasad Prakash Wajekar	201451002	Report

Table of contents:

1. Abstract

In this report, we are presenting the work done by our group as part of the 1st mini project. We followed an holistic approach in order to answer the question number one and the other part completely relied on the **R** programming. The data given in demon.csv represents sample of people's opinion on the demonetization along with the other details like gender, state, income, and area of residence. Looking at the data, we constructed a number of questions and analyzed them by plotting. In the later section, we calculated the parameters of gamma distribution of the Income.

2. Introduction

In Data Analytics, our job is to find trends in the given data. In order to solve the section-1, we formalized a number of questions with a sense of making queries over the dataset and then analysed it by plotting them in Rstudio.

All the questions along with the solution program and plot are given in the next section. In section-2, we took only one variable 'monthly.income' from the dataset 'Demon.csv' and drawn the observations from the Gamma distribution. Accordingly, we calculated the gamma parameter and validated the assumption, which can be found in the section 4 of this report.

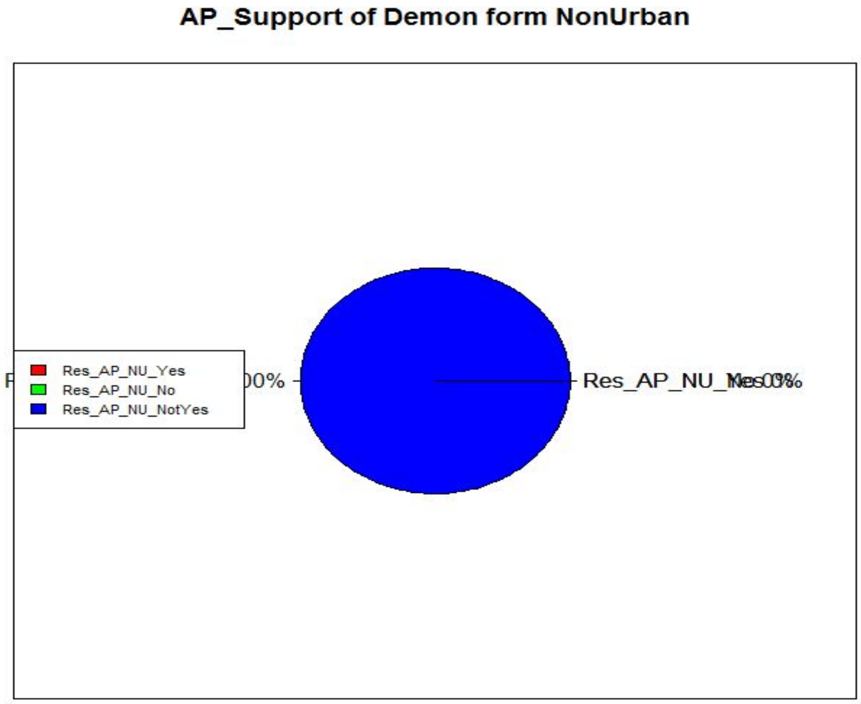
3. Extracted Information from Dataset [Q1]

Q1. How are many people from Andhra Pradesh from rural areas supports Demonetisation?

Code

```
library(plotly)
```

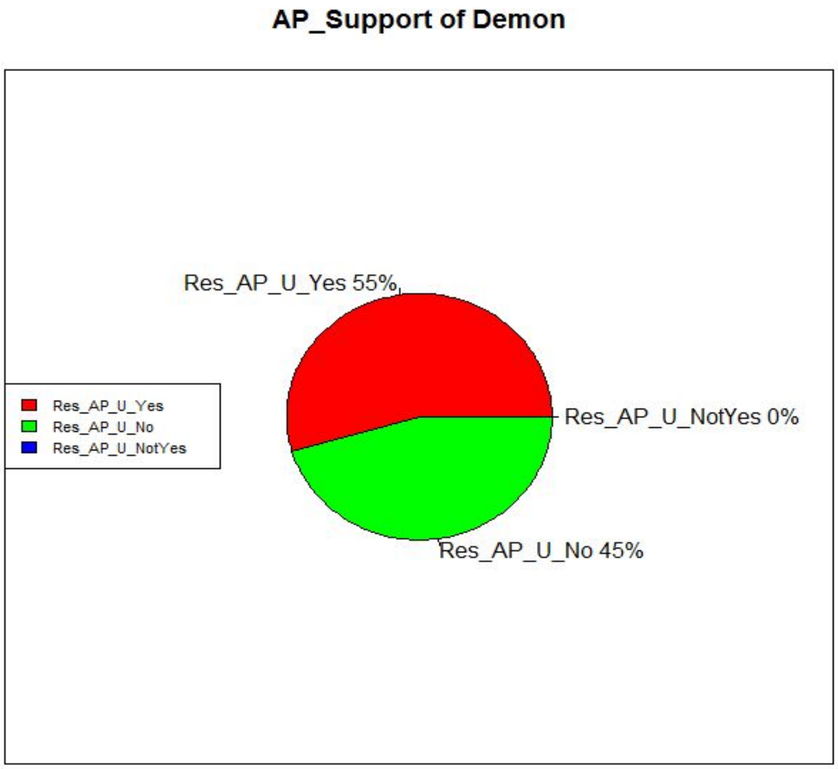
```
orig.data <- read.csv("Demon.csv")
attach(orig.data)
p4 <- sum(orig.data$Residence=="Andhra Pradesh" & Urban=="FALSE" & Demonitisation=="Yes")
p5 <- sum(orig.data$Residence=="Andhra Pradesh" & Urban=="FALSE" & Demonitisation=="No")
p6 <- sum(orig.data$Residence=="Andhra Pradesh" & Urban=="FALSE" & Demonitisation=="not Yes")

x1 <- c(p4,p5,p6)
lbles <- c("Res_AP_NU_Yes","Res_AP_NU_No","Res_AP_NU_NotYes")
pct <- round(x1/sum(x1)*100)
lbls <- paste(lbles, pct)
lbls <- paste(lbls,"%",sep="")
pie(x1,labels = lbls, col=rainbow(length(lbls)),
    main="AP_Support of Demon form NonUrban",radius = 4)
legend("topleft", c("Res_AP_NU_Yes","Res_AP_NU_No","Res_AP_NU_NotYes"), cex = 0.7,fill = rainbow(length(x1)))

box()
```



Q2. How are many people from Andhra Pradesh from urban areas supports Demonetisation?

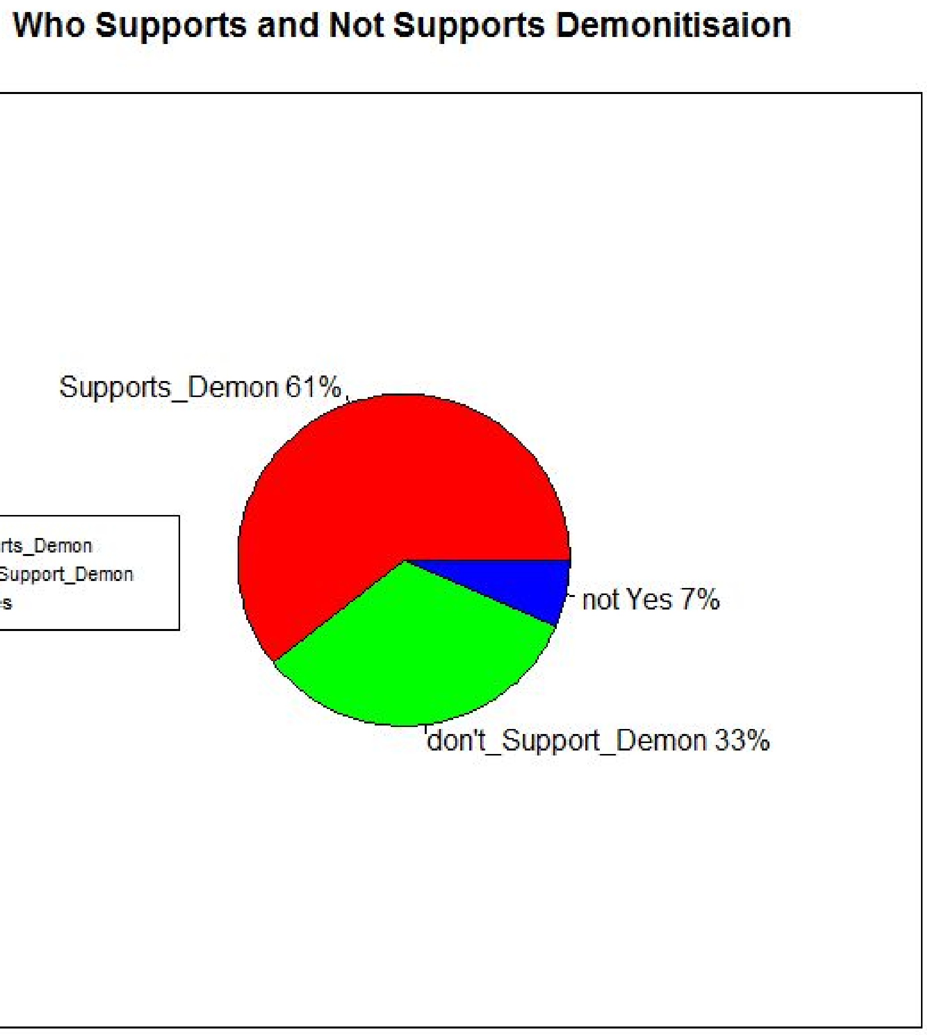
<u>Code</u>

```
p1 <- sum(orig.data$Residence=="Andhra Pradesh" & Urban=="TRUE" & Demonitisation=="Yes")

> p2 <- sum(orig.data$Residence=="Andhra Pradesh" & Urban=="TRUE" & Demonitisation=="No")

> p3 <- sum(orig.data$Residence=="Andhra Pradesh" & Urban=="TRUE" & Demonitisation=="not Yes")

> x1 <- c(p1,p2,p3)

> lbles <- c("Res_AP_U_Yes","Res_AP_U_No","Res_AP_U_NotYes")

> pct <- round(x1/sum(x1)*100)

> lbls <- paste(lbles, pct)

> lbls <- paste(lbls,"%",sep="")

> pie(x1,labels = lbls, col=rainbow(length(lbls)),
+     main="AP_Support of Demon",radius = 4)

> legend("topleft", c("Res_AP_U_Yes","Res_AP_U_No","Res_AP_U_NotYes"), cex = 0.7,fill = rainbow(length(x1)))

> box()
```



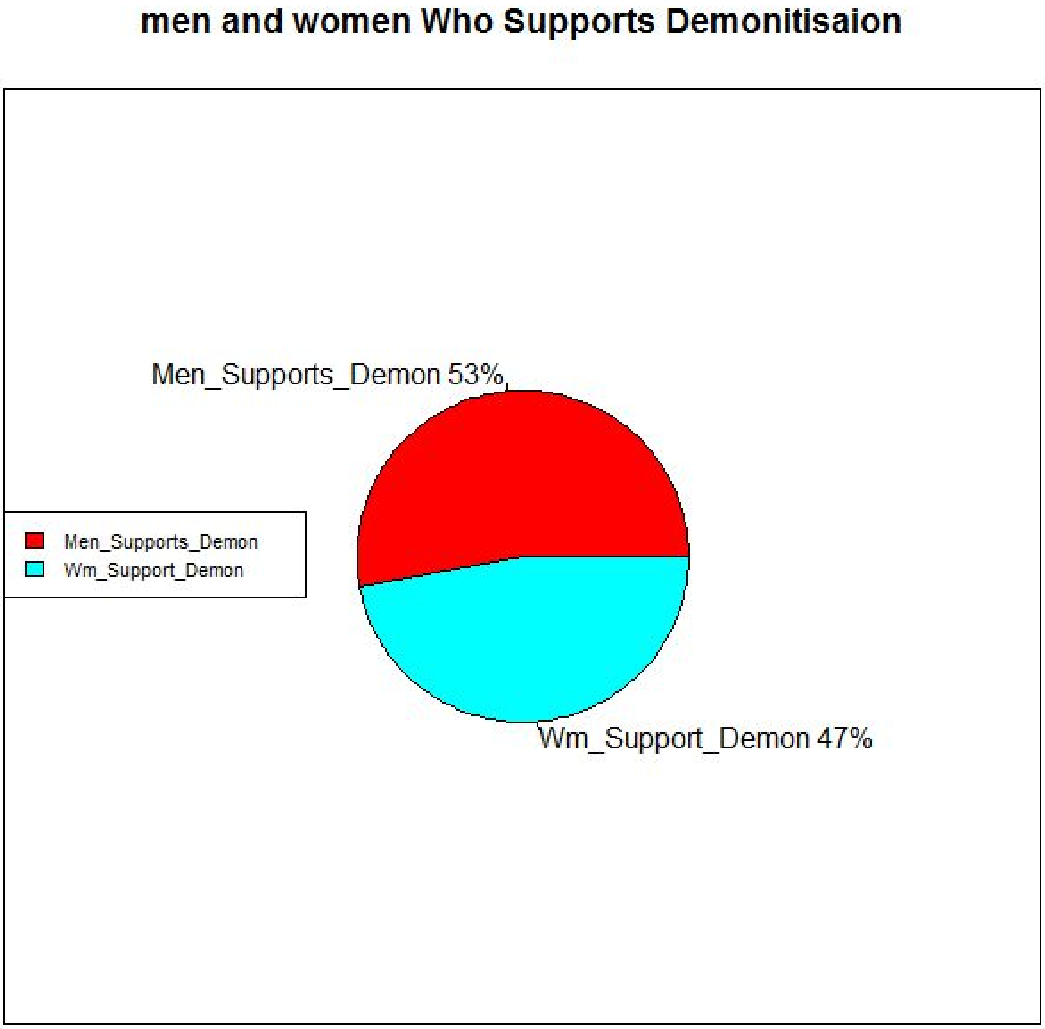
Q3. How many people supports Demonetisation and how many people do not support Demonetisation?

```
library(plotly)
orig.data <- read.csv("Demon.csv")
attach(orig.data)
p1 <- sum(orig.data$Demonitisation=="Yes")
p2 <- sum(orig.data$Demonitisation=="No")
p3 <- sum(orig.data$Demonitisation=="not Yes")
x <- c(p1,p2,p3)
lbles <- c("Supports_Demon","don't_Support_Demon","not Yes")
pct <- round(x/sum(x)*100)
lbls <- paste(lbles, pct)
lbls <- paste(lbls,"%",sep="")
pie(x,labels = lbls,col=rainbow(length(lbls)),main = "Who Supports and Not Supports Demonitisaion",radius = 4
)
legend("topleft", c("Supports_Demon","don't_Support_Demon","not Yes"), cex = 0.7,fill = rainbow(length(x)))

box()
```



Q4. How many men and women support the Demonetisation?

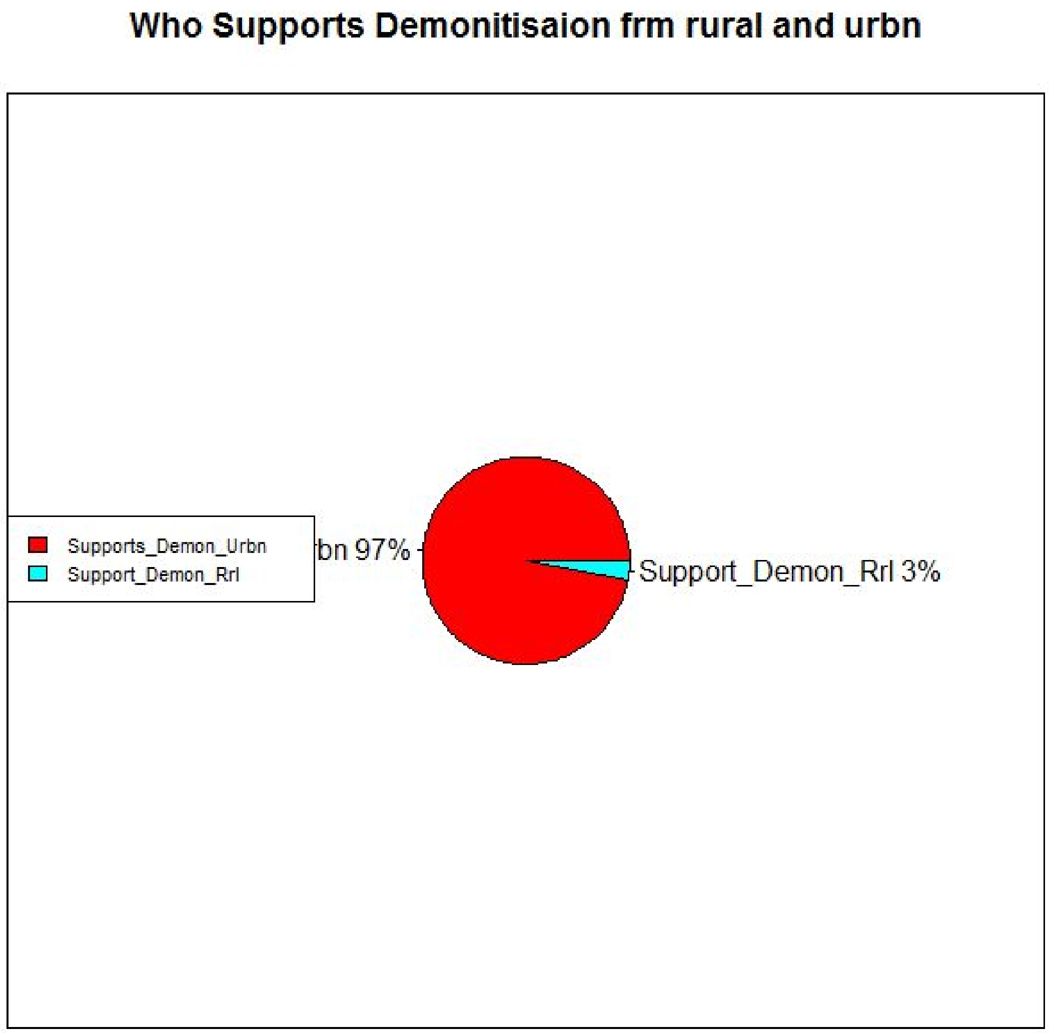
```
library(plotly)
orig.data <- read.csv("Demon.csv")
attach(orig.data)
p1 <- sum(orig.data$Demonitisation=="Yes" & sex=="M")
p2 <- sum(orig.data$Demonitisation=="Yes" & sex=="F")

x <- c(p1,p2)
lbles <- c("Men_Supports_Demon","Wm_Support_Demon")
pct <- round(x/sum(x)*100)
lbls <- paste(lbles, pct)
lbls <- paste(lbls,"%",sep="")
pie(x,labels = lbls,col=rainbow(length(lbls)),main = "men and women Who Supports Demonitisaion",radius = 4
)
legend("topleft", c("Men_Supports_Demon","Wm_Support_Demon"), cex = 0.7,fill = rainbow(length(x)))

box()
```



Q5. How many people supporting Demonetisation are belongs to urban area and How many people supporting Demonetisation are belongs to rural area ?

```
library(plotly)
orig.data <- read.csv("Demon.csv")
attach(orig.data)
```

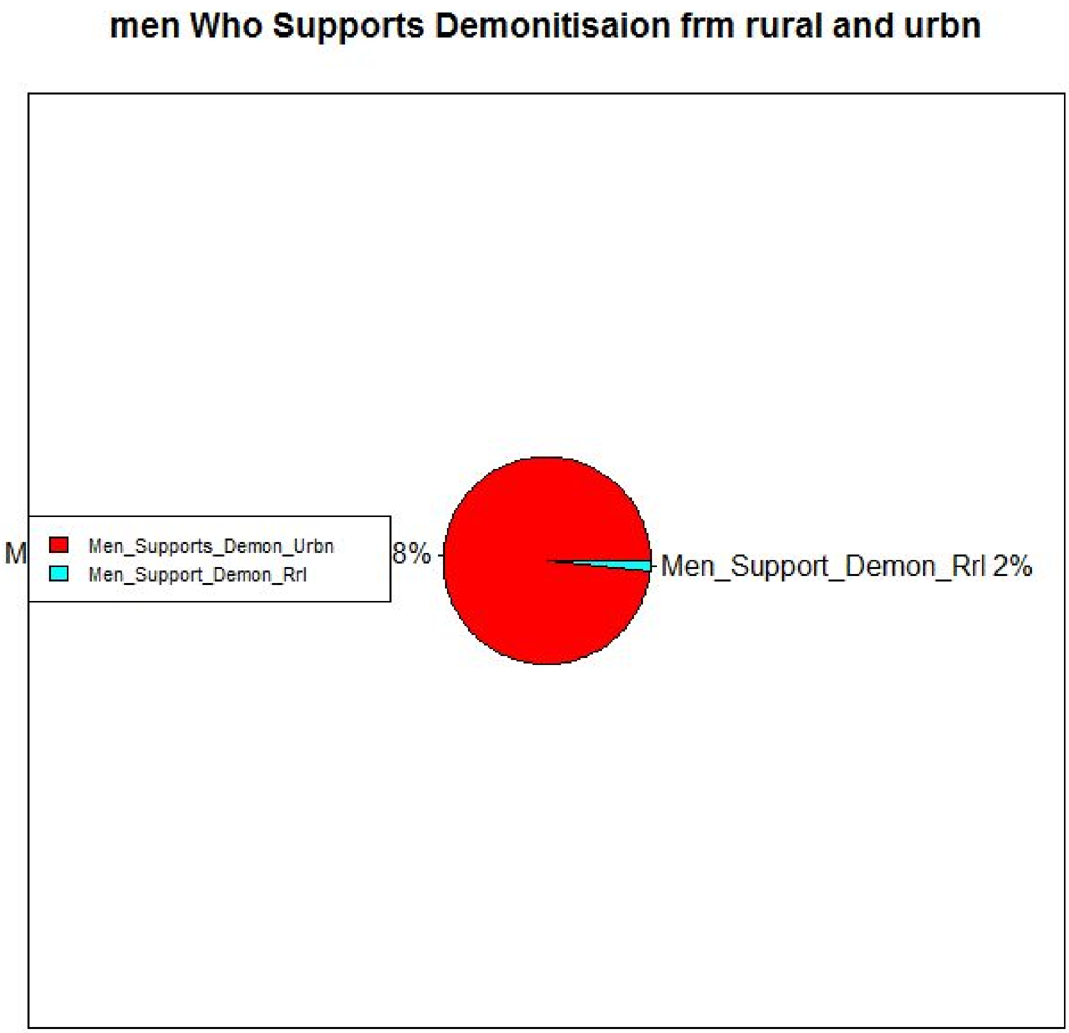
```
p1 <- sum(orig.data$Demonitisation=="Yes" & Urban=="TRUE")
p2 <- sum(orig.data$Demonitisation=="Yes" & Urban=="FALSE")

x <- c(p1,p2)
lbles <- c("Supports_Demon_Urbn","Support_Demon_Rrl")
pct <- round(x/sum(x)*100)
lbls <- paste(lbles, pct)
lbls <- paste(lbls,"%",sep="")
pie(x,labels = lbls,col=rainbow(length(lbls)),main = "Who Supports Demonitisaion frm rural and urbn",radius = 2.5
)
legend("topleft", c("Supports_Demon_Urbn","Support_Demon_Rrl"), cex = 0.7,fill = rainbow(length(x)))

box()
```



Q6. How many men supporting Demonetisation are belongs to urban area? How many men supporting Demonetisation are belongs to rural area?

```
library(plotly)
orig.data <- read.csv("Demon.csv")
attach(orig.data)
p1 <- sum(orig.data$Demonitisation=="Yes" & Urban=="TRUE" &sex=="M")
```

```
p2 <- sum(orig.data$Demonitisation=="Yes" & Urban=="FALSE"&sex=="M")

x <- c(p1,p2)
lbles <- c("Men_Supports_Demon_Urbn","Men_Support_Demon_Rrl")
pct <- round(x/sum(x)*100)
lbls <- paste(lbles, pct)
lbls <- paste(lbls,"%",sep="")
pie(x,labels = lbls,col=rainbow(length(lbls)),main = "men Who Supports Demonitisaion frm rural and urbn",radius = 2.5
)
legend("topleft", c("Men_Supports_Demon_Urbn","Men_Support_Demon_Rrl"), cex = 0.7,fill = rainbow(length(x)))

box()

```

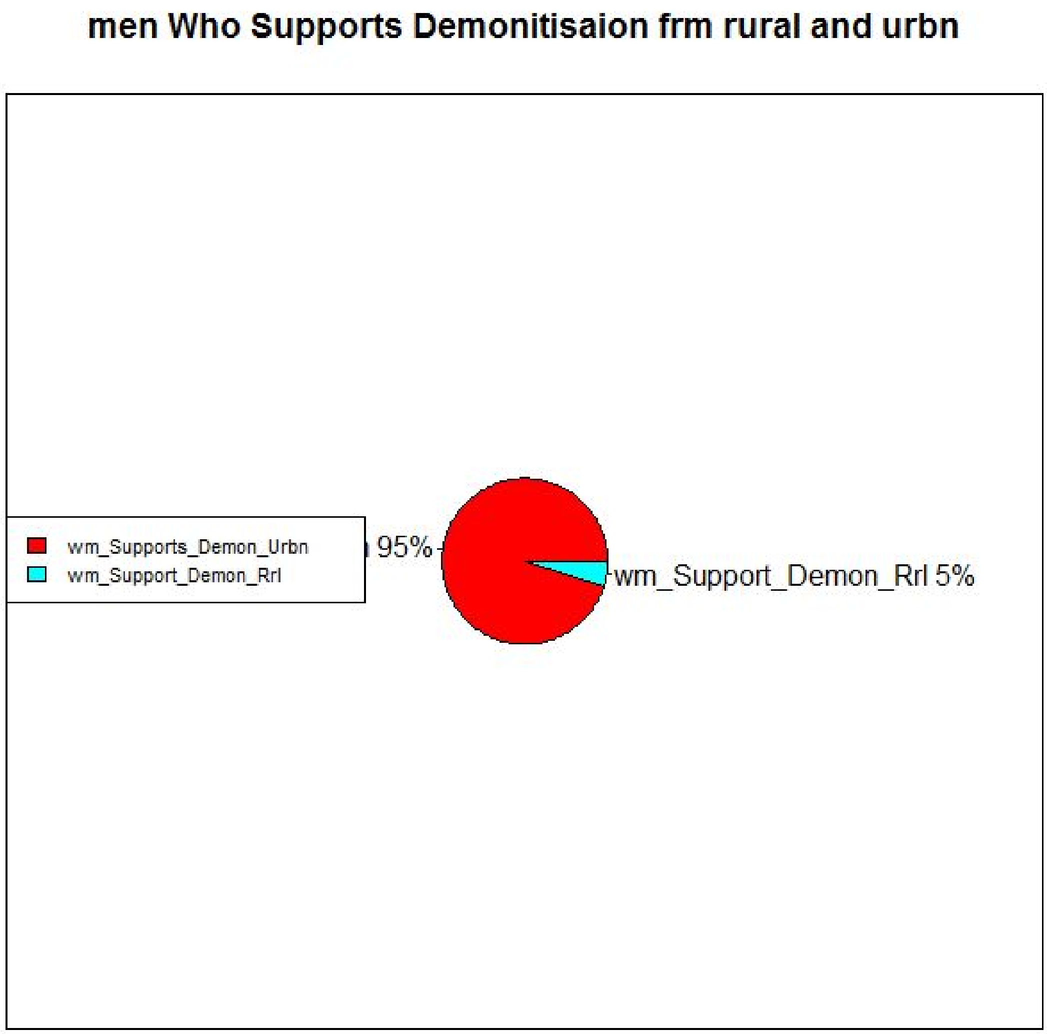


Q7. How many women supporting Demonetisation are belongs to urban area? How many women supporting Demonetisation are belongs to rural area?

```
library(plotly)
orig.data <- read.csv("Demon.csv")
attach(orig.data)
```

```
p1 <- sum(orig.data$Demonitisation=="Yes" & Urban=="TRUE" &sex=="F")
p2 <- sum(orig.data$Demonitisation=="Yes" & Urban=="FALSE"&sex=="F")

x <- c(p1,p2)
lbles <- c("Women_Supports_Demon_Urbn","Women_Support_Demon_Rrl")
pct <- round(x/sum(x)*100)
lbls <- paste(lbles, pct)
lbls <- paste(lbls,"%",sep="")
pie(x,labels = lbls,col=rainbow(length(lbls)),main = "women Who Supports Demonitisaion frm rural and urbn",radius = 2.5
)
legend("topleft", c("Women_Supports_Demon_Urbn","Women_Support_Demon_Rrl"), cex = 0.7,fill = rainbow(length(x)))

```

```
box()
```

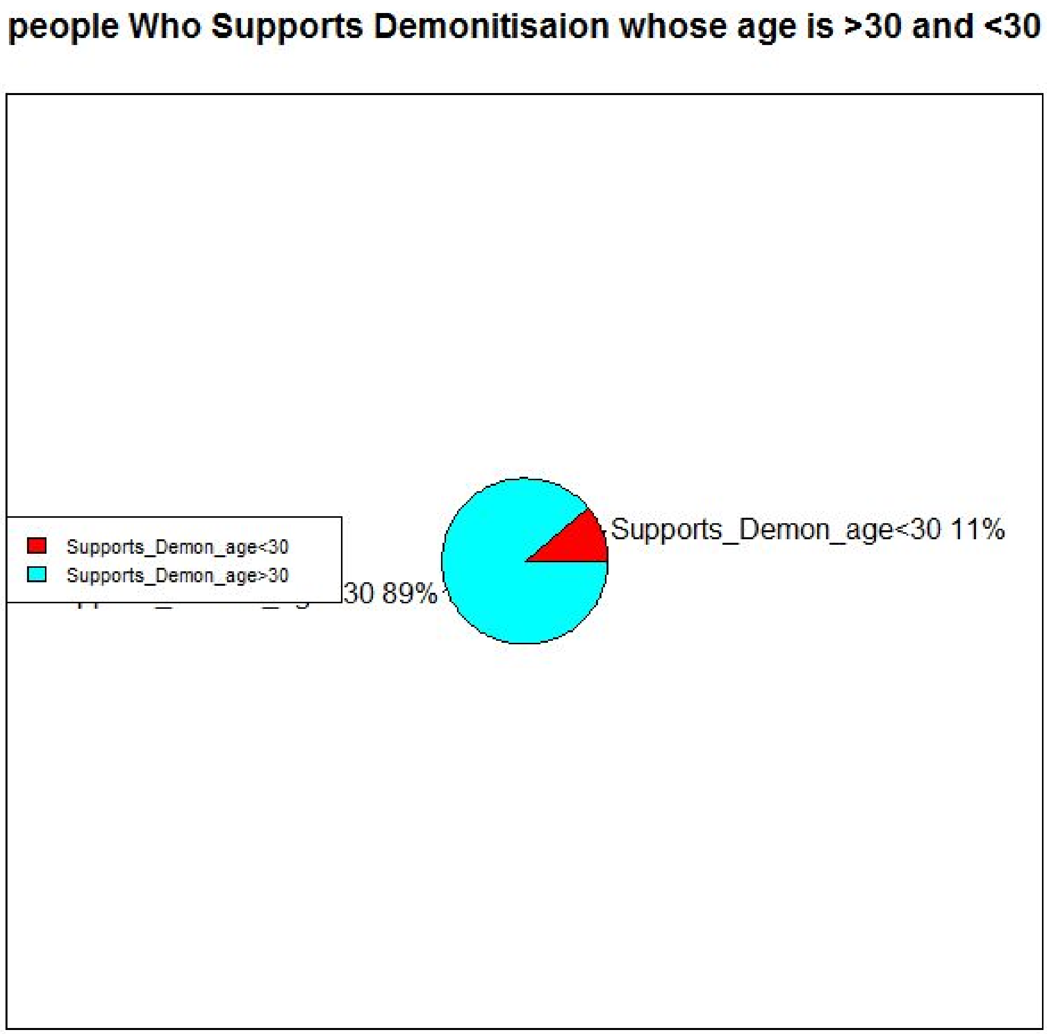


Q8. How many people supporting Demonetisation are older than 30 years? How many supporting Demonetisation are younger than 30 years?

```
library(plotly)
orig.data <- read.csv("Demon.csv")
attach(orig.data)
p1 <- sum(orig.data$Demonitisation=="Yes"&age<30)
```

```
p2 <- sum(orig.data$Demonitisation=="Yes"&age>30)

x <- c(p1,p2)
lbles <- c("Supports_Demon_age<30","Supports_Demon_age>30")
pct <- round(x/sum(x)*100)
lbls <- paste(lbles, pct)
lbls <- paste(lbls,"%",sep="")
pie(x,labels = lbls,col=rainbow(length(lbls)),main = "people Who Supports Demonitisaion whose age is >30 and <30",radius = 2
)
legend("topleft", c("Supports_Demon_age<30","Supports_Demon_age>30"), cex = 0.7,fill = rainbow(length(x)))

box()
```

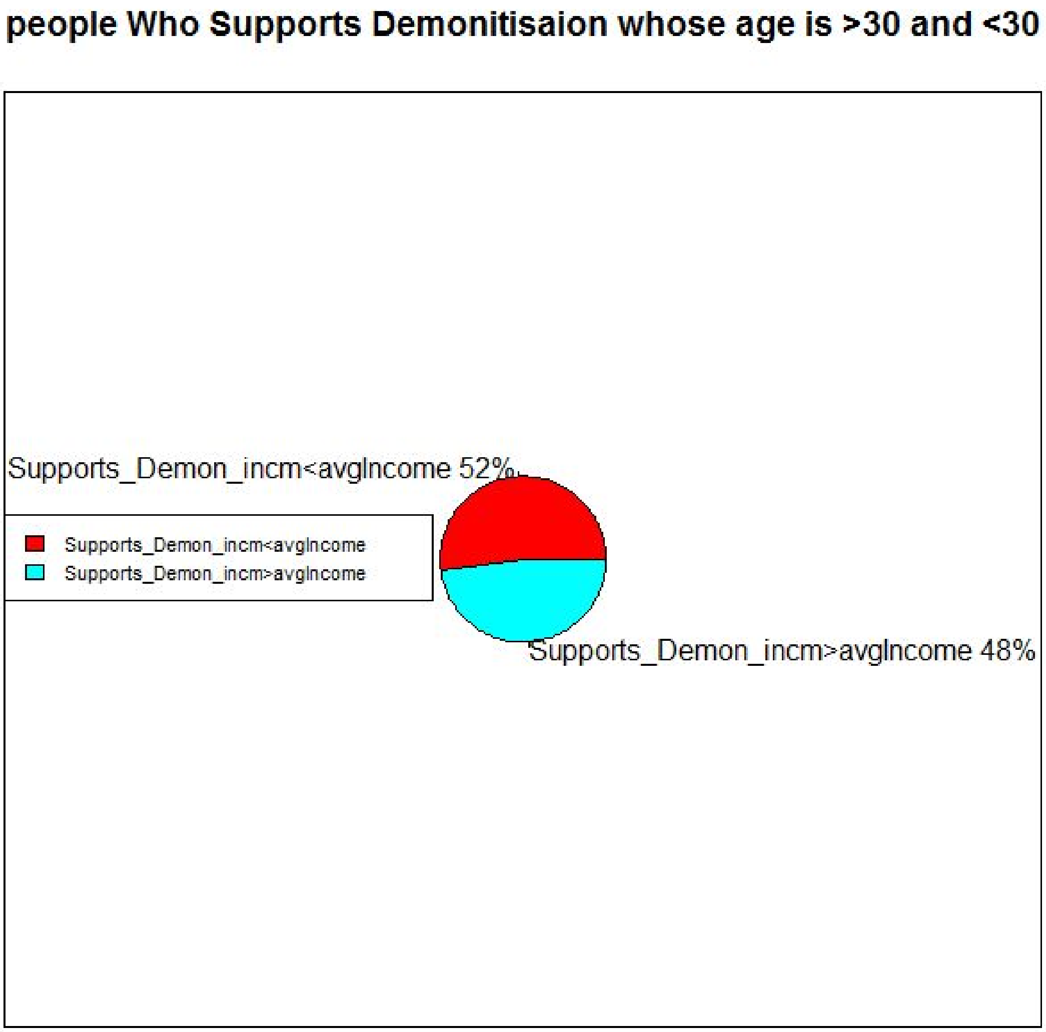
17.



Q9. How many people are support Demonetisation whose income is greater than average income and less than average income?

```
library(plotly)
```

```
orig.data <- read.csv("Demon.csv")
attach(orig.data)

p1 <- sum(orig.data$Demonitisation=="Yes"& monthly.income<avg_income)
p2 <- sum(orig.data$Demonitisation=="Yes"& monthly.income>avg_income)

x <- c(p1,p2)
lbles <- c("Supports_Demon_incm<avgIncome","Supports_Demon_incm>avgIncome")
pct <- round(x/sum(x)*100)
lbls <- paste(lbles, pct)
lbls <- paste(lbls,"%",sep="")
pie(x,labels = lbls,col=rainbow(length(lbls)),main = "people Who Supports Demonitisaion whose income is greater than average income and whose income is less than average income, radius = 2
)
legend("topleft", c("Supports_Demon_incm<avgIncome","Supports_Demon_incm>avgIncome"), cex = 0.7,fill = rainbow(length(x)))

box()
```



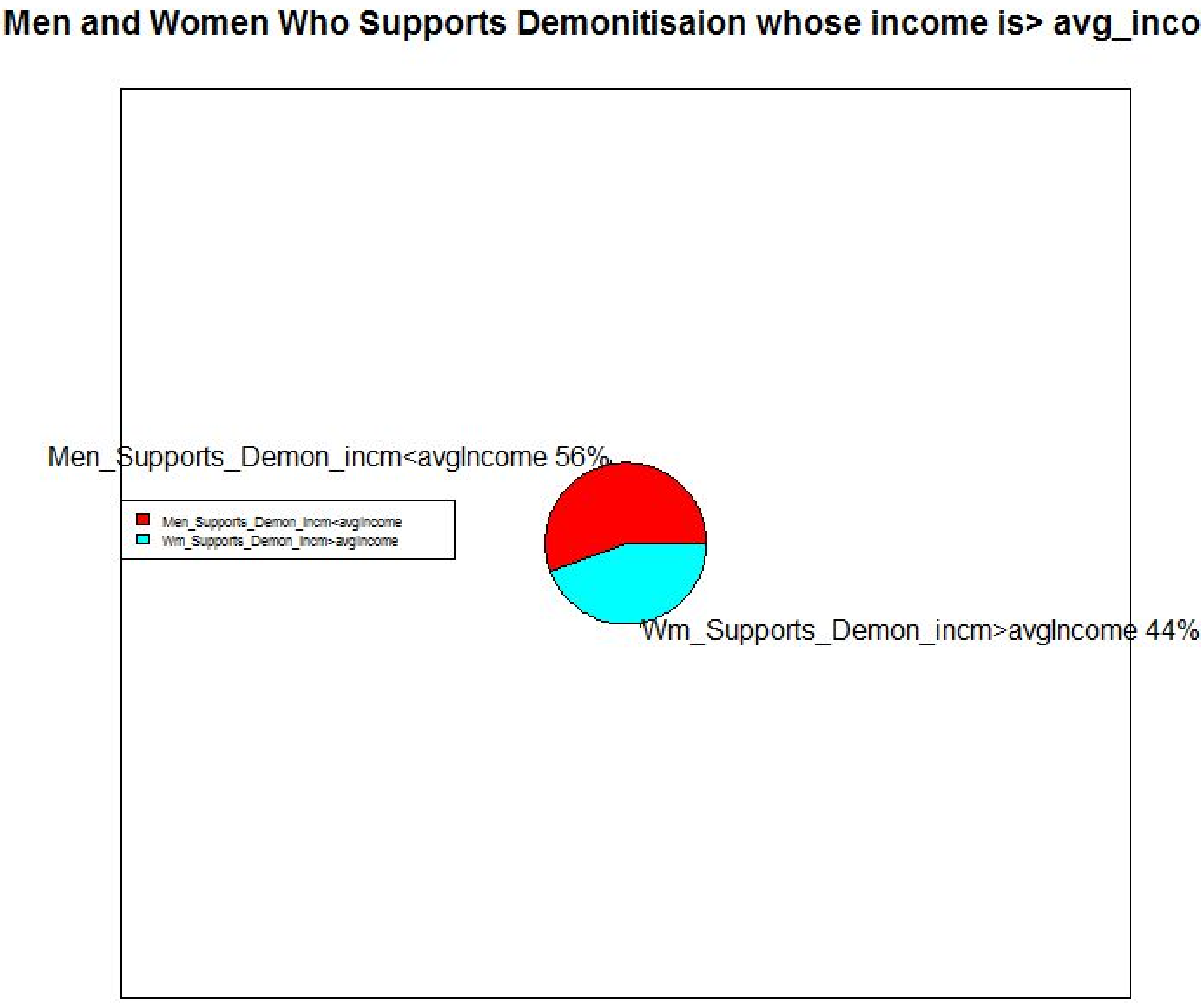
Q10. How many men support Demonetisation? How many women support Demonetisation?

<u>Code</u>

```
library(plotly)
orig.data <- read.csv("Demon.csv")
```

```
attach(orig.data)

p1 <- sum(orig.data$Demonitisation=="Yes"& monthly.income<avg_income & sex=="M")
p2 <- sum(orig.data$Demonitisation=="Yes"& monthly.income>avg_income& sex=="F")

x <- c(p1,p2)
lbles <- c("Men_Supports_Demon_incm<avgIncome","Wm_Supports_Demon_incm>avgIncome")
pct <- round(x/sum(x)*100)
lbls <- paste(lbles, pct)
lbls <- paste(lbls,"%",sep="")
pie(x,labels = lbls,col=rainbow(length(lbls)),main = "Men and Women Who Supports Demonitisaion whose income is> avg_income",radius = 2
)
legend("topleft", c("Men_Supports_Demon_incm<avgIncome","Wm_Supports_Demon_incm>avgIncome"), cex = 0.5,fill = rainbow(length(x)))

box()

```

19.


Men and Women Who Supports Demonitisaion whose income is> avg_inco
Men_Supports_Demon_incm<avgIncome 56%
🟥 Men_Supports_Demon_Incm<avgIncome
🟦 Wm_Supports_Demon_Incm>avgIncome
Wm_Supports_Demon_incm>avgIncome 44%

4. Solution of Q2

Copy only the variable “monthly.income” from the dataset “Demon.csv”.
Suppose you denote the monthly income as X and assume that the observations are drawn from a Gamma distribution.
(a) Find the parameters of the gamma distribution.

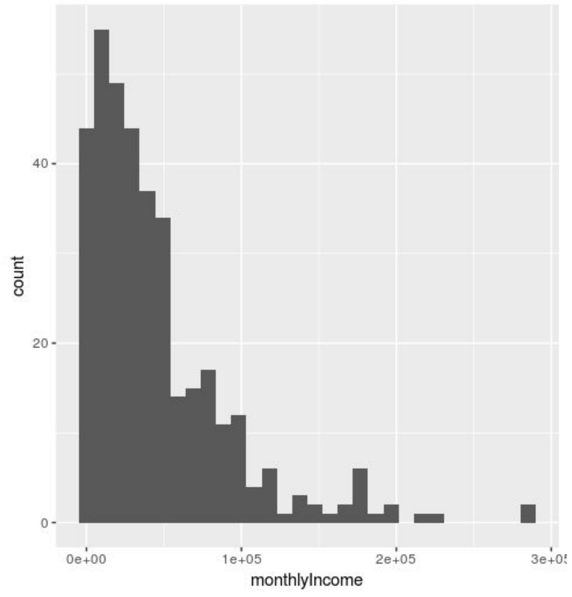
Before find the parameter we find that there are some outlier so we removed the outlier because it will effect on distribution
# Remove Outlier
monthly Income <- monthlyIncome[monthlyIncome < 3e+05]

After that we find the mean and variance from the monthly Income data set

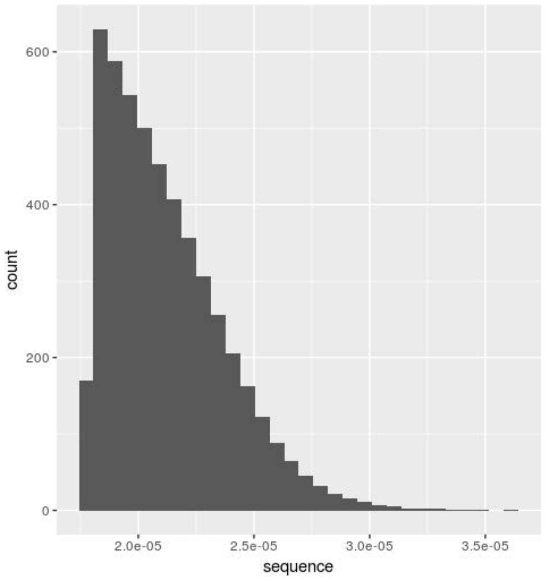
Mean =: 45101.73
Var =: 2160424771
Alpha =: 0.9415585
Beta =: 2.087633e-05

(b) For validating we plot the the histogram original value of monthly income and seq from 0 to 10000 with by 2 step

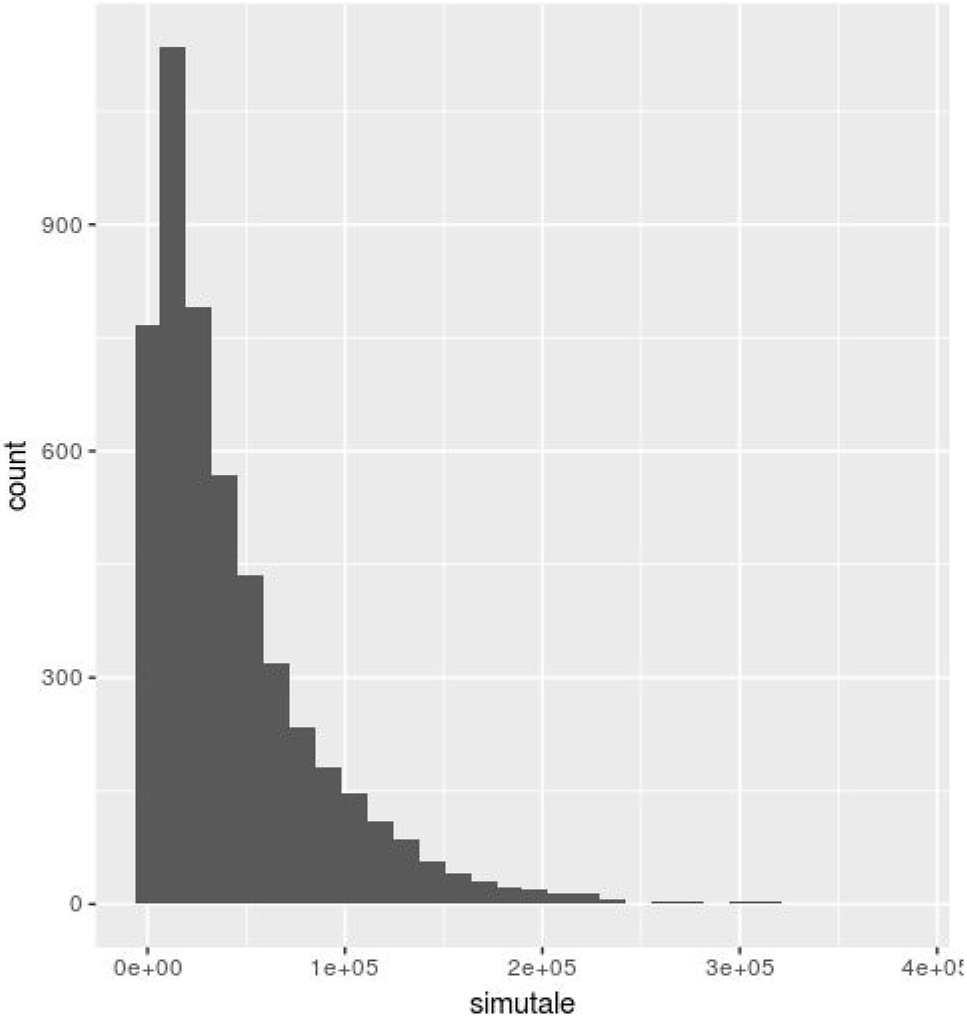
Assumption of origin plot



Gamma distribution plot

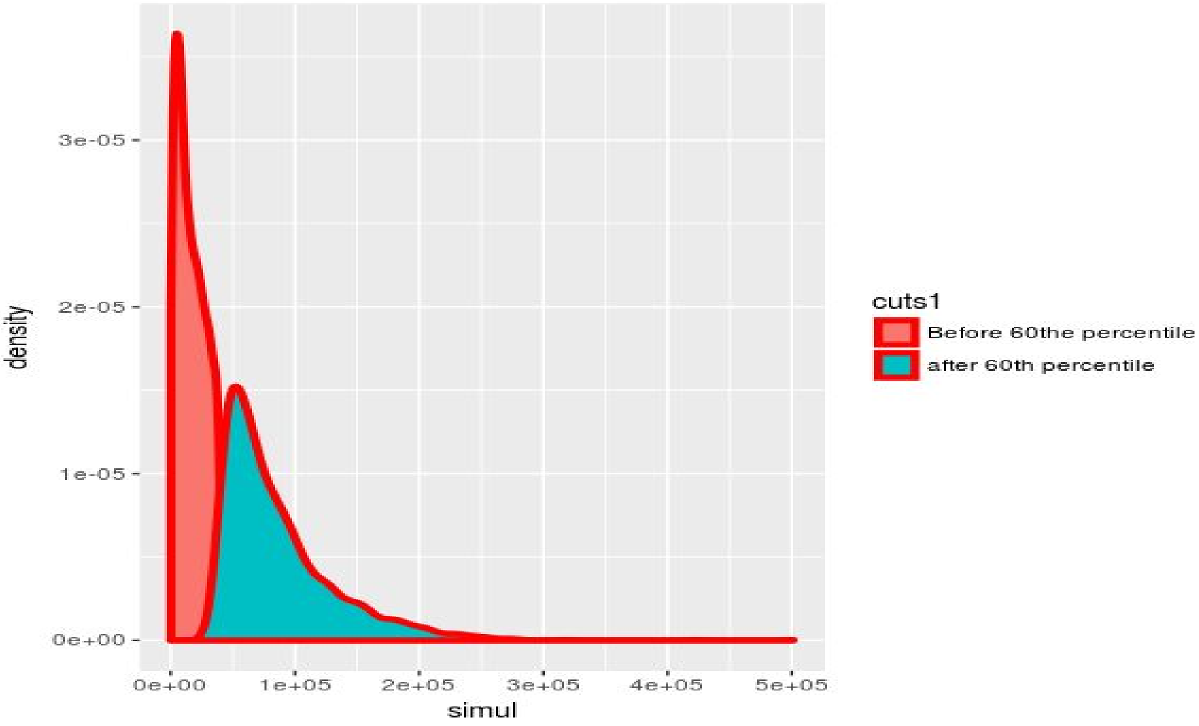


(c)
Simulate (simulation size 5000) from the gamma distribution with parameter values same as in (a) and find characteristics of the estimator of 60 th percentile.
Using the origin parameter alpha and beta and rgamma function and plot the graph



According the given question find the 60th percentile of simulated data
So using the quantile function we find out the values

Percentile 60% =: 95012.78



##### Code for 2nd problem

```
#get access data from file
data <- read.csv('Demon.csv')
# getting column form demon data set
monthlyIncome <- data$monthly.income

# Remove Outlier
monthlyIncome <- monthlyIncome[monthlyIncome < 3e+05]

# find the mean
mean <- mean(monthlyIncome)

#find the verience
var <- var(monthlyIncome)
# find the standard deviosion
```

```
sd <- sqrt(var
# find the alpha
alpha <- mean^2/var
# find the beta
beta  <- mean/var

##########  validation

qplot(x = monthlyIncome)+geom_histogram()

x1 = seq(0, 10000, by = 2)
d <- dgamma(x1,shape = alpha, rate = beta)
qplot(x=d)+geom_histogram()

###################### simulation ###
#Simulate (simulation size 5000) from the gamma distribution
with
#parameter values same as in (a) and find characteristics of the
estimator of

#60 th percentile.

simulation = 5000
n <- rgamma(simulations, shape =alpha, rate = beta)
qplot(x = n) + geom_histogram()
perc_60 = quantile(n, 0.6)
simulation <- ggplot(data=n, aes(x=simul, y = ..density..))
+geom_density(aes(fill= cuts1),col="red", lwd = 1.5)
print(perc_60)
## out put
# print(perc_60)
# 60%
# 95012.78
```