Data Analytics

Mini Project
Date- 06/09/17

Q-1.

The data given in “Demon.csv” presents people’s reaction on demonetisation. The data was collected from the states given in “Residence”. The people’s id was given so as to protect the identity of the individual. The variable “Urban” indicates whether the person is from the urban area of the state (TRUE) or rural area (FALSE). The age and monthly income of the persons included in the study are given in “age” and “monthly.income” respectively. The variable “Demonetisation” indicates the support. If the concerned person supports demonetisation policy then “Yes” is noted otherwise a “No”. Suppose you are a data scientist. Given the data set, extract as much information as you can think of from the dataset.

Ans. To extract the information, we divided the question in different parts. There are some exceptions in the dataset which is impossible(like having an age of 1130 years) so we made some assumptions, based on which we performed our extraction from the dataset. We also made some different categories to show the distribution of different attributes of the given dataset. These are as follows:

Assumption:

1. The person having opinion “Not Yes” have been assumed not supporting the demonetisation and all these cases are treated as “No”.

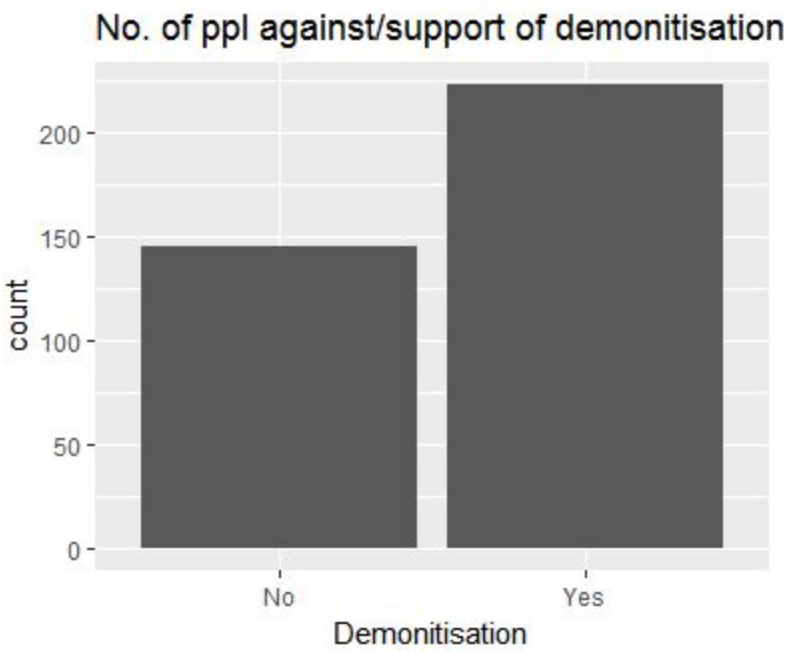
2. The person with the age of 1130 years has not been considered as part of data-set and has been removed while analyzing the dataset.

3. In age, we have made three categories:
    a. 0-21: Teen
    b. 21-60: Adult
    c. 60-150: Old

4. In monthly income, we have made three categories:

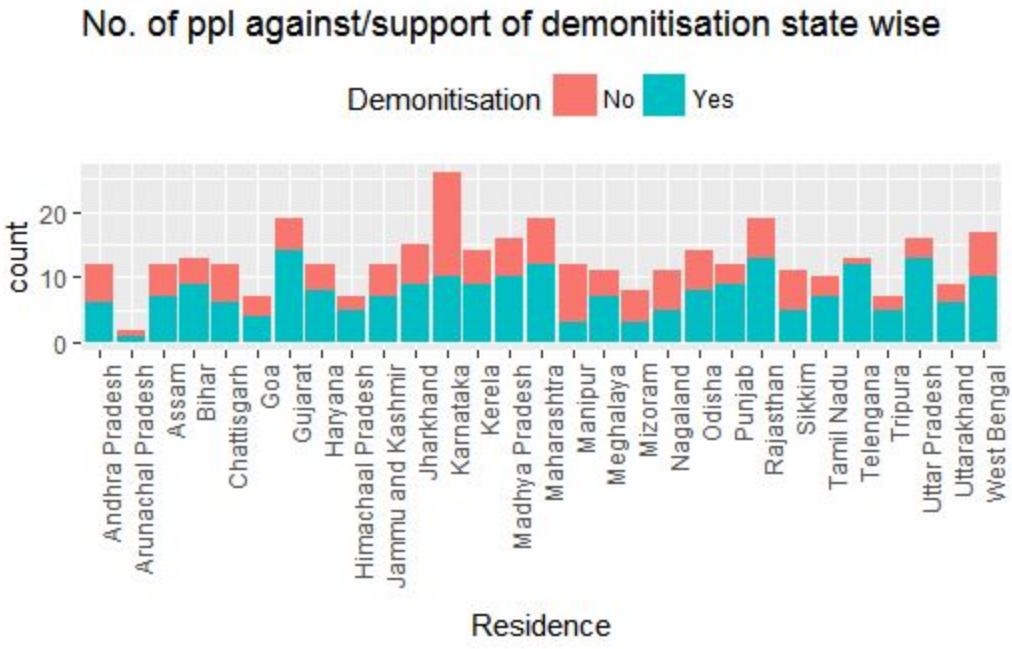
a. 0-30000rs: Poor
b. 30000-100000rs: Middle
c. Above 100000rs: Rich

1(a) Total number of people in favour or against demonetisation in the given dataset.



Observation: In the given data-set “Demon.csv”, which represents the statistics of the "Demonetisation" across India who is supporting or contradicting the idea of Demonetisation, with the given data, we concluded that among all the 368 people, '223'(approximately 60.6%) people are supporting and '145'(approximately 39.4%) people are against the Demonetisation.

1(b) Statewise opinion of people about demonetisation.



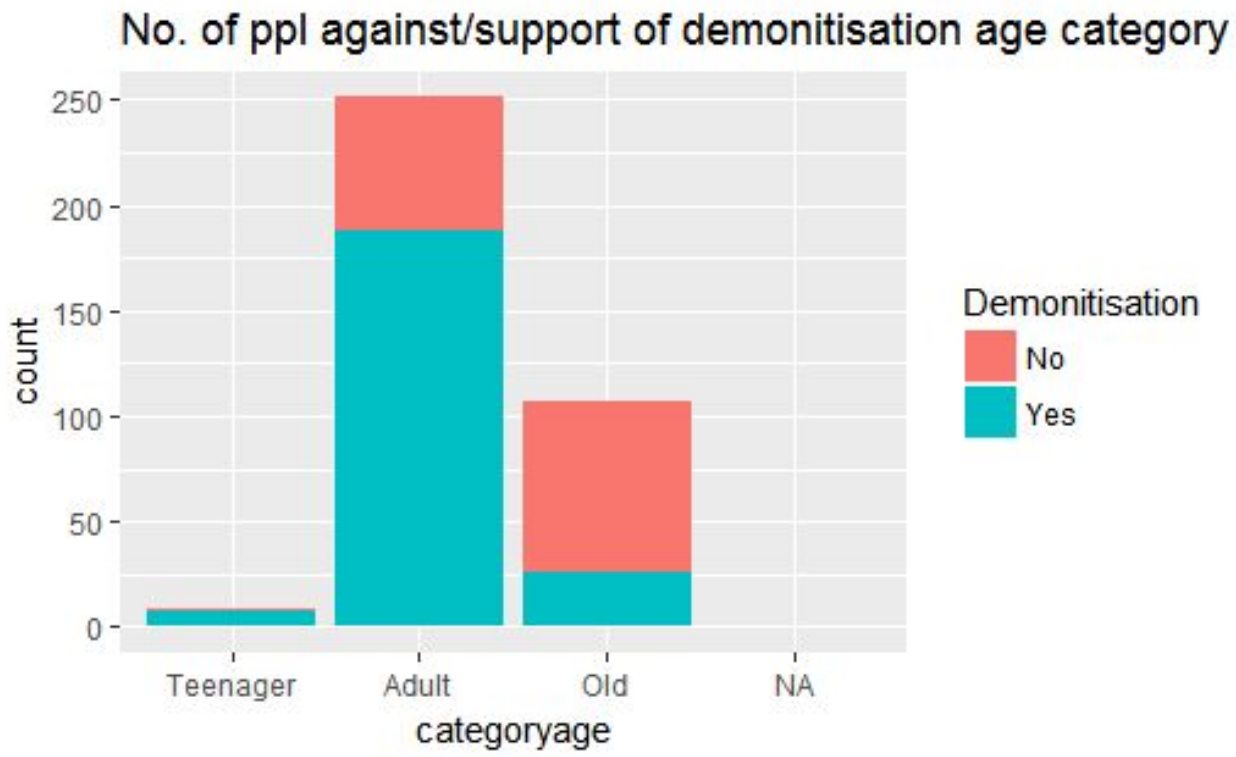
Data-Table:

Residence	No	Yes
Andhra Pradesh	6	6
Arunachal Pradesh	1	1
Assam	5	7
Bihar	4	9
Chhattisgarh	6	6
Goa	3	4

Gujarat	5	14
Haryana	4	8
Himachal Pradesh	2	5
Jammu and Kashmir	5	7
Jharkhand	6	9
Karnataka	16	10
Kerala	5	9
Madhya Pradesh	6	10
Maharashtra	7	12
Manipur	9	3
Meghalaya	4	7
Mizoram	5	3
Nagaland	6	5
Odisha	6	8
Punjab	3	9
Rajasthan	6	13
Sikkim	6	5
Tamil Nadu	3	7
Telengana	1	12
Tripura	2	5
Uttarakhand	3	13
Uttar Pradesh	3	6
West Bengal	7	10

Observation: From the above graph, we can safely conclude that except from Karnataka, Manipur, Mizoram and Nagaland, rest of the states have more people in support of demonetisation than against it. People of Telangana, show highest support for demonetisation as 92.30% of its people are in favour of demonetisation. People of Manipur show lowest support(25%) for demonetisation. We can say that people from southern states are more in favour of demonetisation than in eastern areas.

1(c) Opinion of people on demonetisation in different age categories.

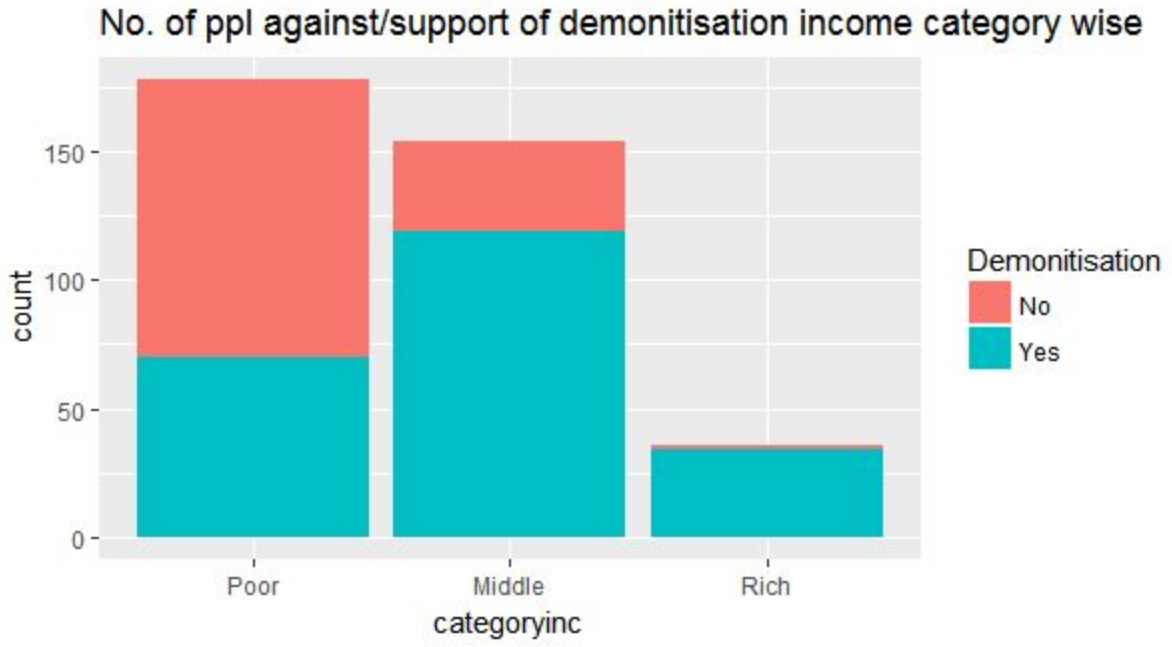


Data-Table:

	Teenager	Adult	Old
Yes	8(89%)	188(75%)	26(24%)
No	1(11%)	63(25%)	81(76%)

Observation : From the given bar graph, we can say that the percentage ratio who is supporting the Demonetisation is more in Teenager than in Adult category and the least in Old category. In Teenager, approx 89% is supporting the Demonetisation, whereas Adult who are supporting is approx 75% and 24% is supporting Demonetisation in the Old category.

1(d) Total number of people in support of demonetisation income category wise.



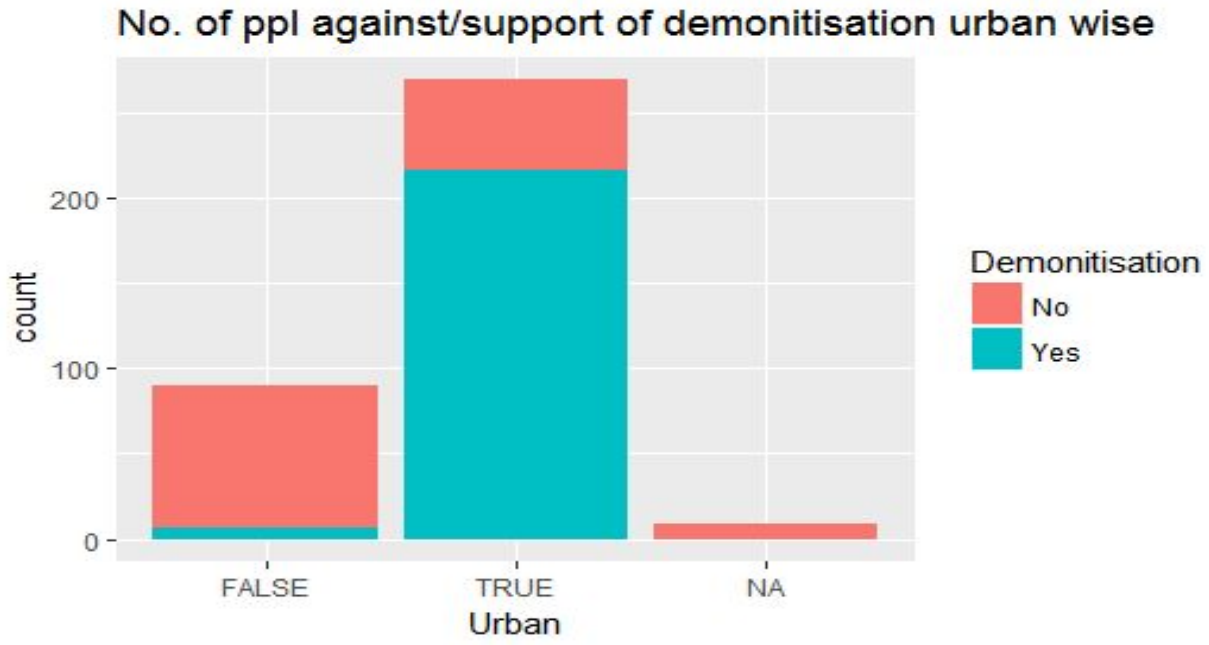
Data-Table:

	Poor	Middle	Rich
Yes	70(40%)	119(77%)	34(95%)
No	108(60%)	35(23%)	2(5%)

Observation : From the above graph, it can be seen that while most of the rich and middle income category are supporting demonetisation, there is no clear mandate

among the poor section of society. More than 60% of poor are against demonetisation whereas only 25% are against demonetisation in the middle income group and only 2 out of 36 in rich people.

1(e). Opinion of people about demonetisation region wise.

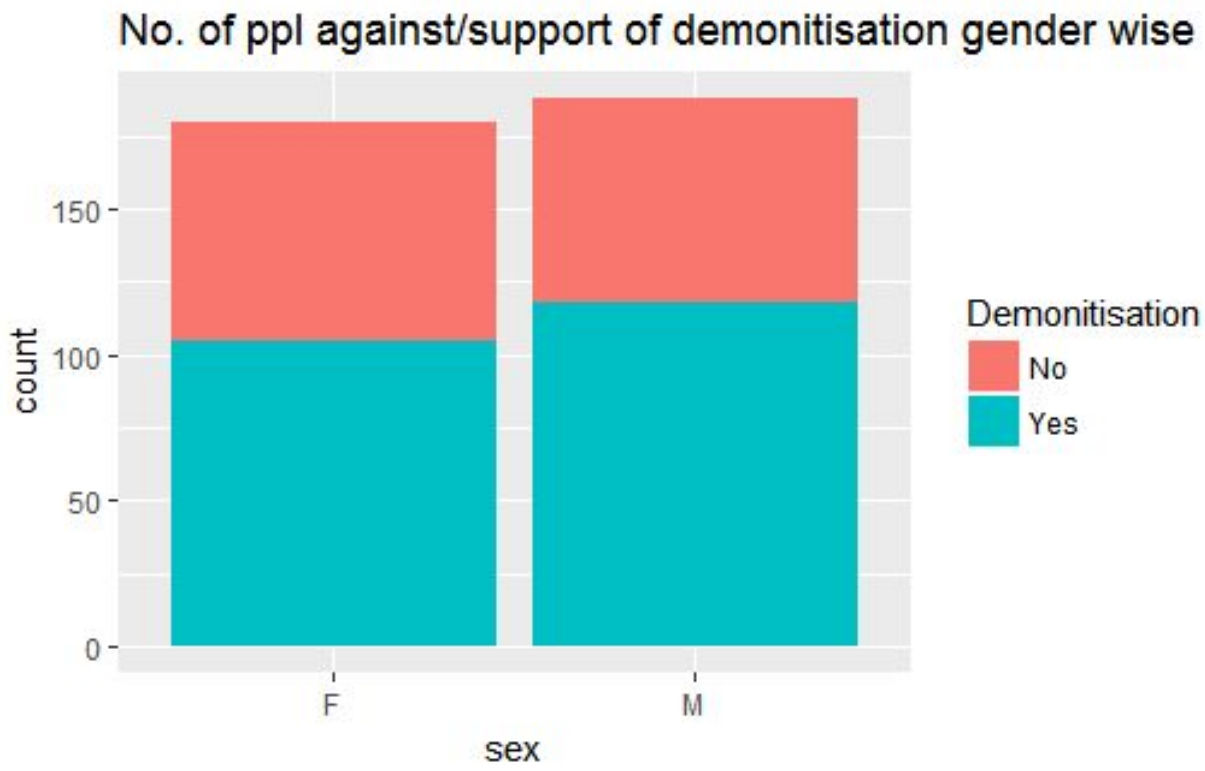


Data-Table:

	Urban	Rural
Yes	216(80%)	7(8%)
No	53(20%)	83(92%)

Observation: From the above graph, it can be clearly seen that in urban areas majority (80%) is supporting demonetisation while in rural areas majority(92%) is against demonetisation.

1(f). Gender-wise opinion of people on demonetisation.



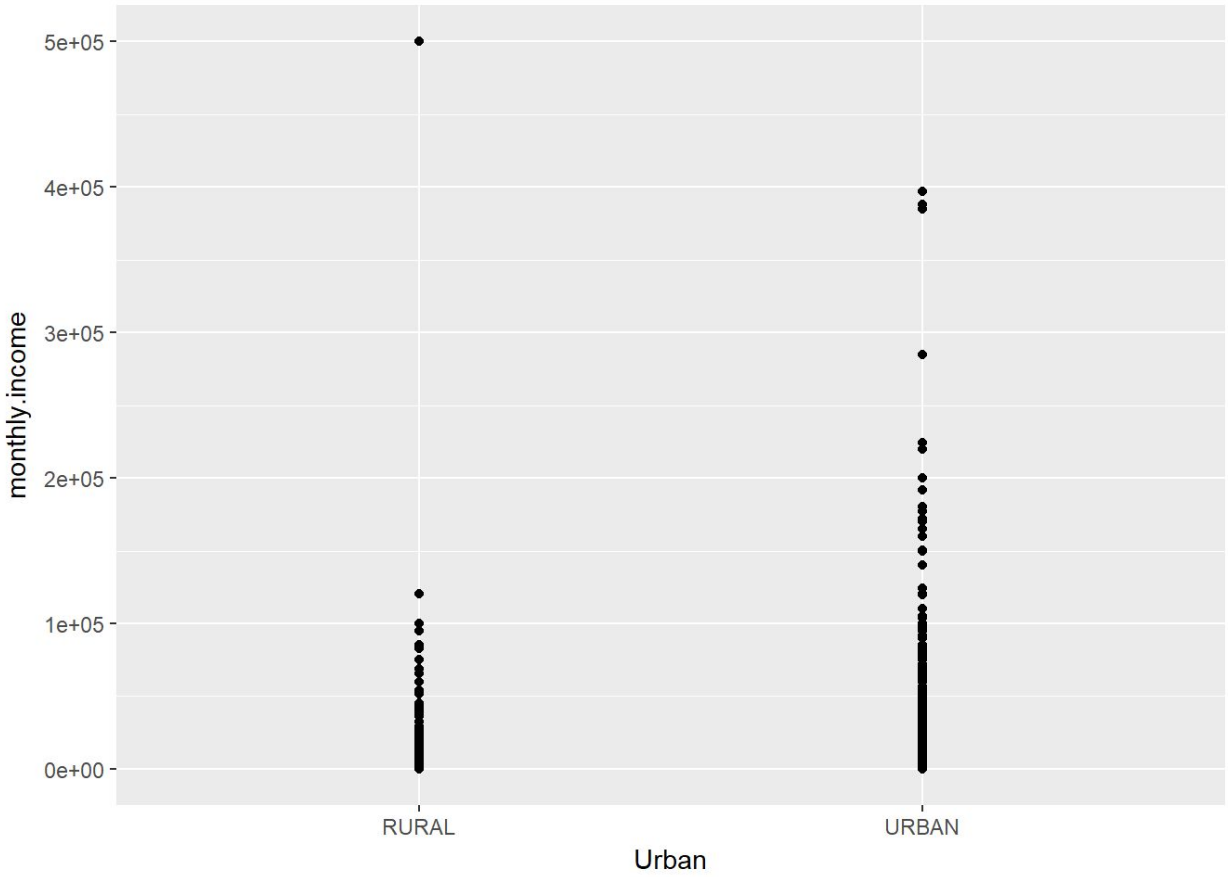
Data-Table:

	Female	Male

Yes	105(58%)	118(62%)
No	75(42%)	70(38%)

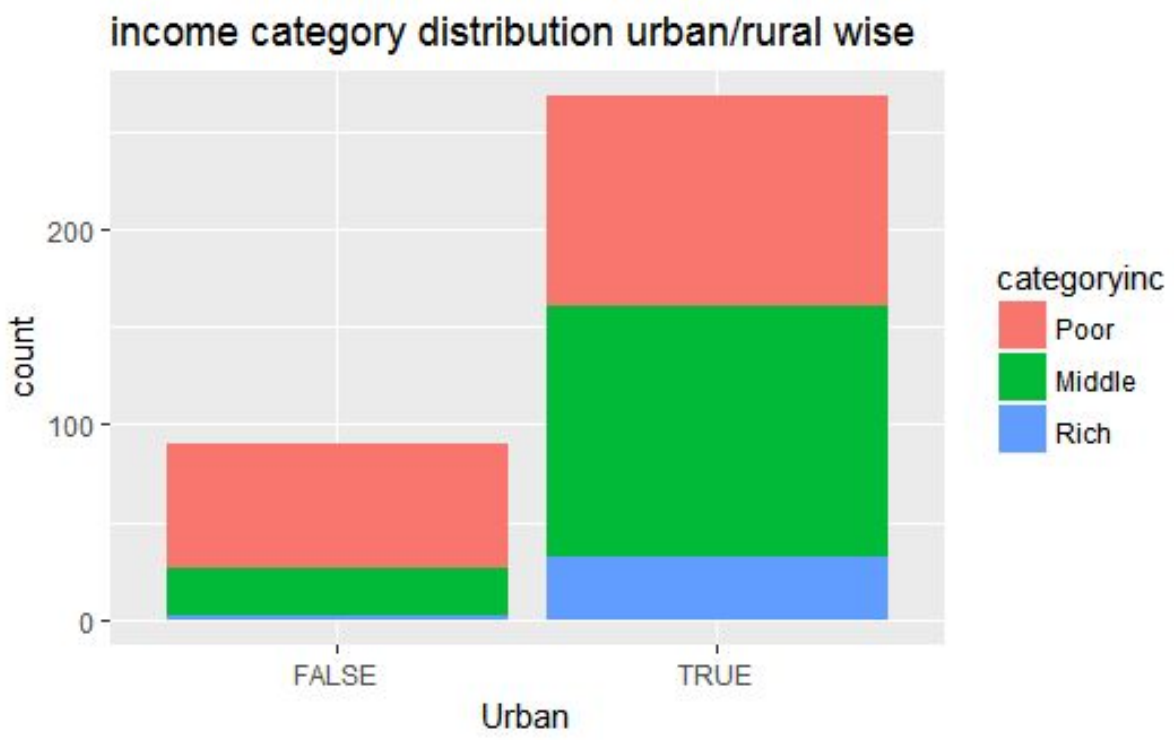
Observation: From the given bar graph, we can say that on an average both the genders are supporting Demonetisation are same or Males(approx 62%) are slightly more than Females(approx 58%).

1(g). Distribution of monthly income region wise.



Observation: With the given distribution, we can say that in rural areas except for two people, all are having monthly income between 0 - 100000Rs and mostly between 0 - 50000Rs so for this set 500000Rs value is the outlier. In urban areas, three people are having monthly income between 300000 - 400000Rs and between 200000 - 300000Rs most of people have monthly income between 0 - 100000Rs and there are many people having income between 100000 - 200000Rs.

1(h). Income distribution region wise.

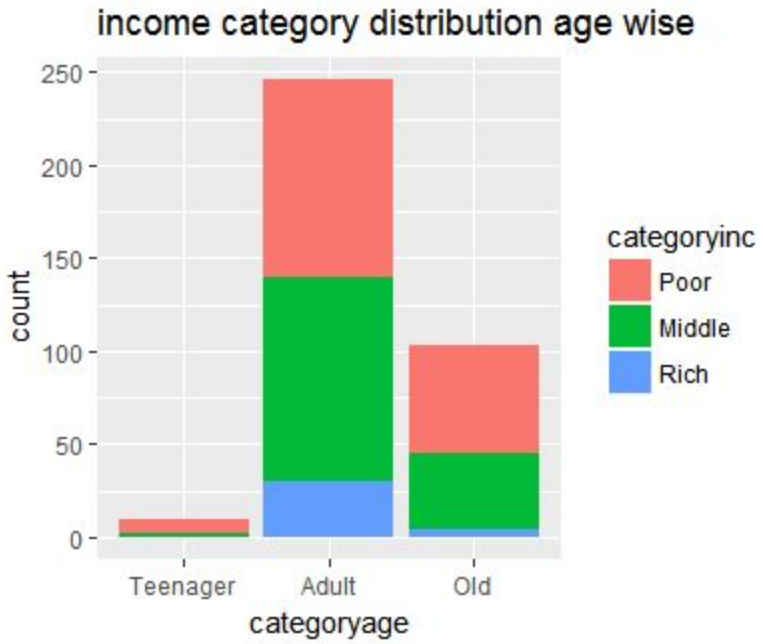


Data-table:

	Urban	Rural
Poor	107(40%)	64(71.1%)
Middle	129(48%)	24(26.7%)
Rich	32(12%)	2(2.2%)

Observation: From the given bar graph, we can state that in Rural, the majority people comes under the poor category which is around 71% whereas 26.7% comes under the middle category and only 2.2% comes under the rich category whereas in Urban, the majority people comes under the middle category approx 48%, 40% under the poor category and 12% falls in the rich category which is far better than rural area.

1(i). Distribution of income age-category wise.

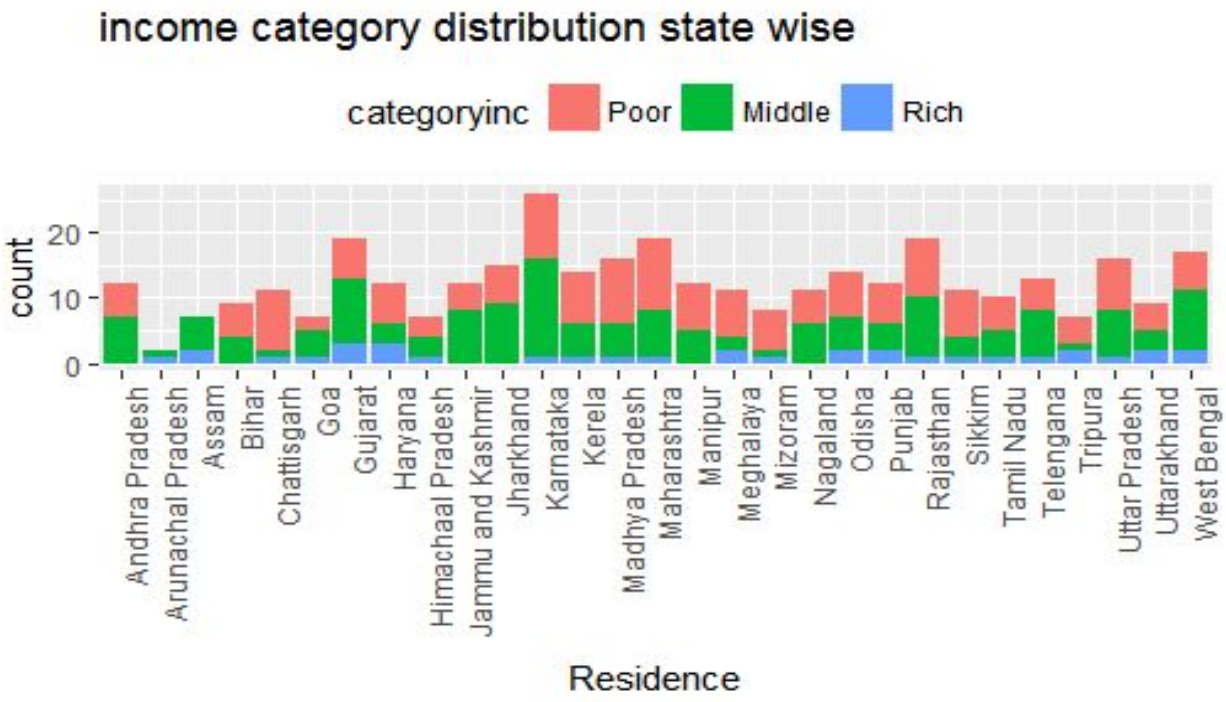


Data-Table:

Category	Teenager	Adult	Old
Poor	7(78%)	106(42%)	58(56%)
Middle	2(22%)	110(46%)	41(40%)
Rich	0(0%)	30(12%)	4(4%)

Observation: From the above graph it can be seen that the ratio of poor is largest in teenagers and lowest in adults. The ratio of rich is also largest in adults. It can be concluded that adults form the richest group followed by old people.

1(j). State-wise number of people falling under the different income categories.



Data-Table:

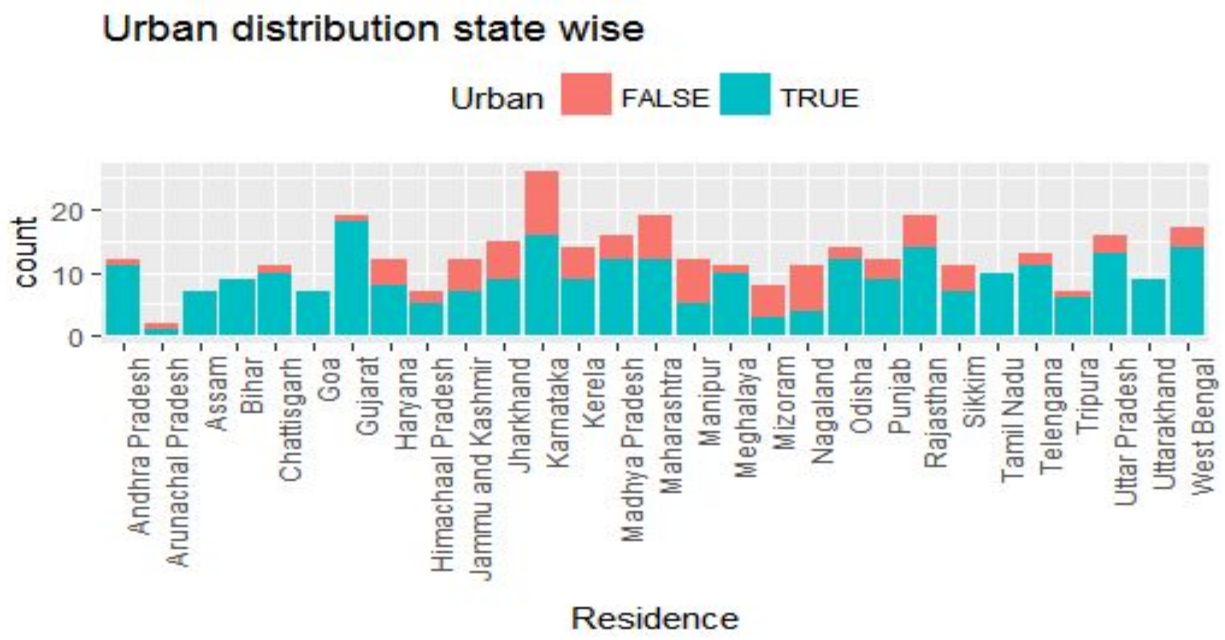
Location	Poor	Middle	Rich
Andhra Pradesh	5	7	0
Arunachal Pradesh	0	1	1
Assam	0	5	2
Bihar	5	4	0
Chhattisgarh	9	1	1
Goa	2	4	1
Gujarat	6	10	3
Haryana	6	3	3
Himachal Pradesh	3	3	1
Jammu Kashmir	4	8	0

Jharkhand	6	9	0
Karnataka	10	15	1
Kerala	8	5	1
Madhya Pradesh	10	5	1
Maharashtra	11	7	1
Manipur	7	5	0
Meghalaya	7	2	2
Mizoram	6	1	1
Nagaland	5	6	0
Odisha	7	5	2
Punjab	6	4	2
Rajasthan	9	9	1
Sikkim	7	3	1
Tamil Nadu	5	4	1
Telengana	5	7	1
Tripura	4	1	2
Uttarakhand	4	3	2
Uttar Pradesh	8	7	1
West Bengal	6	9	2

Observation: From the given bar graph, we can conclude that the state which is having lowest poor rate(no one belongs in poor category) are Arunachal Pradesh and Assam, the lowest middle category rate is in Chattisgarh, having 9.1% middle category, the lowest rich category rate(no one belongs in rich category) are Andhra Pradesh, Bihar, Jammu Kashmir, Jharkhand, Manipur and Nagaland.

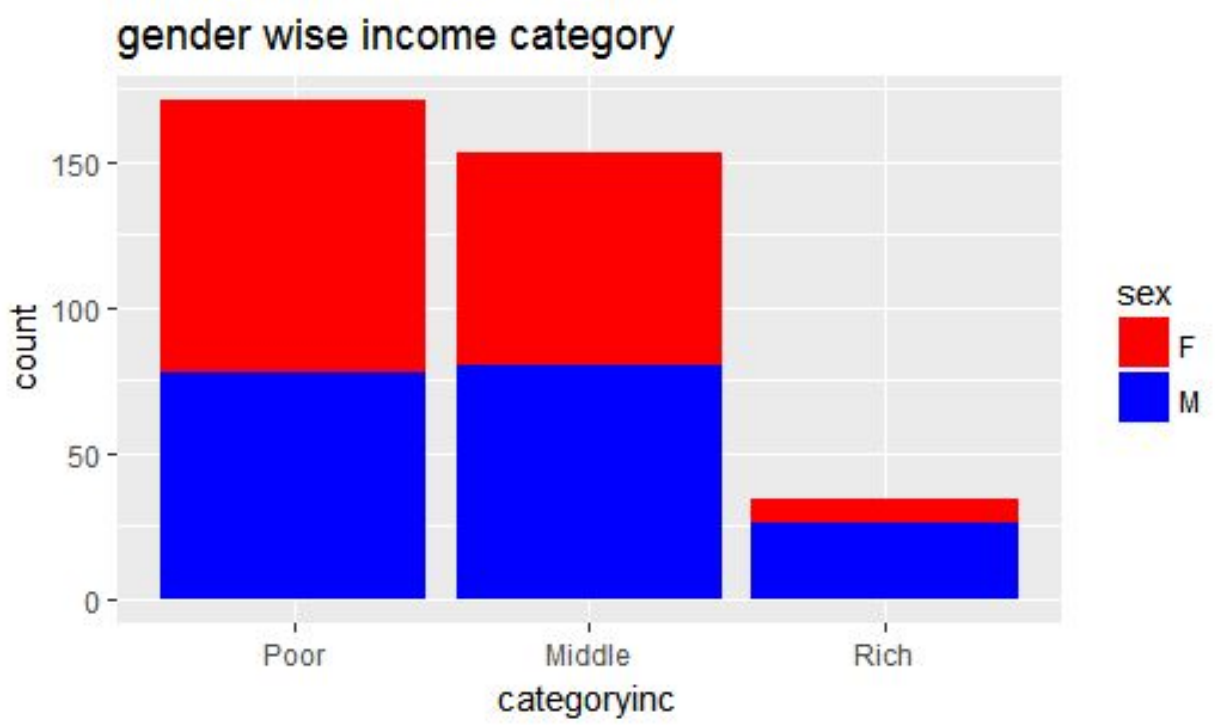
The state which is having highest poor rate is in Mizoram having 75% poor rate, the highest middle rate is in Andhra Pradesh having 8.3% middle rate, the highest rich state is in Arunachal Pradesh having 50% rich category.

1(k). State-wise distribution of urban/rural areas.



Observation: With the above graph, we can say that except for Manipur, Mizoram and Nagaland, all the states have more urban areas than rural. Among 29 states, five of them(Assam, Bihar, Goa, Tamil Nadu and Uttarakhand) do not have any rural areas so we can say for a large dataset too, number rural regions would be negligible as compared to urban in these five states.

1(I). Gender-wise distribution of people falling under different income categories.



Data-table:

	Poor	Middle	Rich
Female	93(55%)	73(47%)	8(24%)
Male	78(45%)	80(53%)	26(76%)

Observation: From the above graph it can be seen that income of male is greater than female. There are more female in poor category(55%), more male in middle income category, and very large number of males in rich income category(76%).

Q-2

Ans (a)

Gamma Distribution has two parameters Lambda and Alpha, Alpha tells us about the shape of the graph and Lambda tells us about rate.

First we find the mean and variance of income data, now from mean and variance we can find α and λ for this distribution

$$f(x) = \lambda^{\alpha}/\Gamma(\alpha) * x^{\alpha-1} * e^{-\lambda x}$$

Using the formulas:-

E(x)= α/λ

Var(x)= α/λ^2

Results that we got:-

Mean - 49149.54

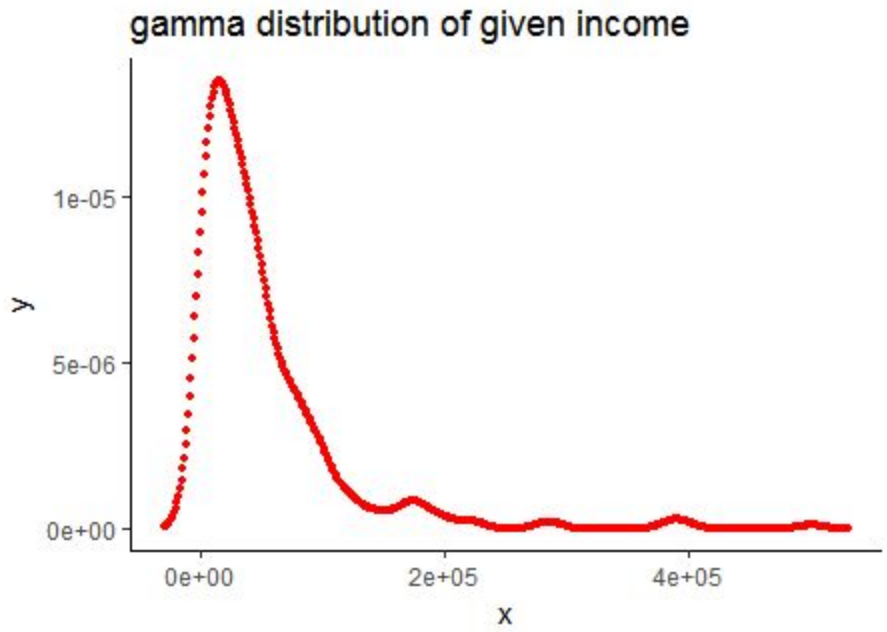
Variance-3656892321

Lambda-1.344025e-05

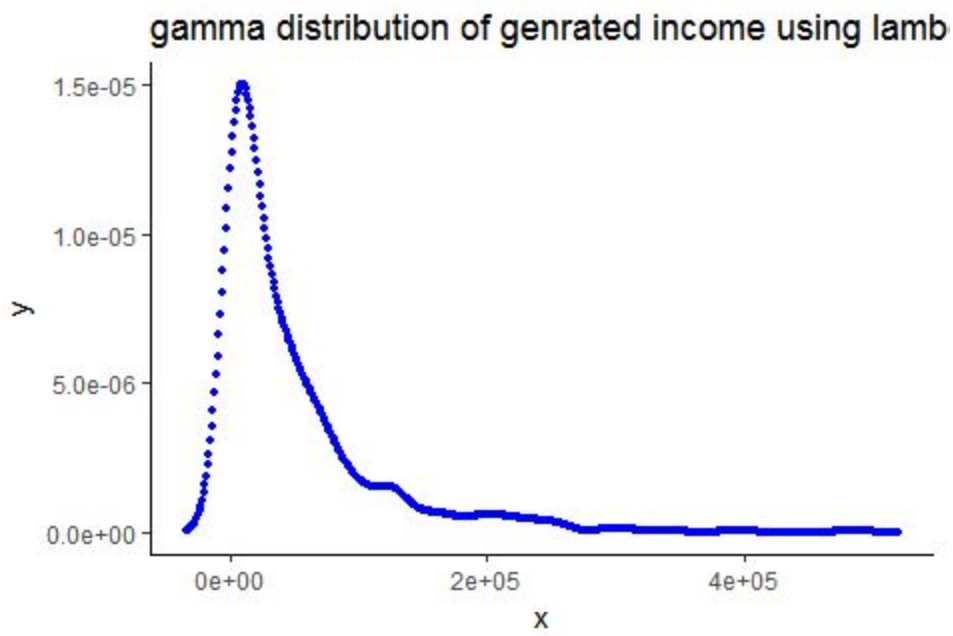
Alpha- 0.660582

2(b).

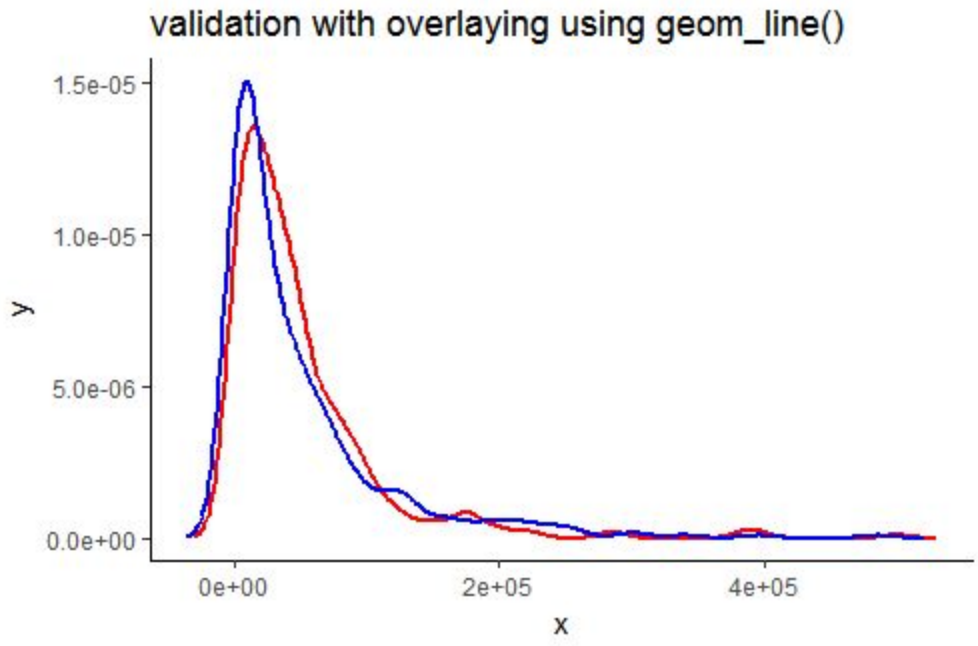
First we plotted gamma distribution for the given income data, as we got α<1, the resulted graph is desireable that is right skewed graph



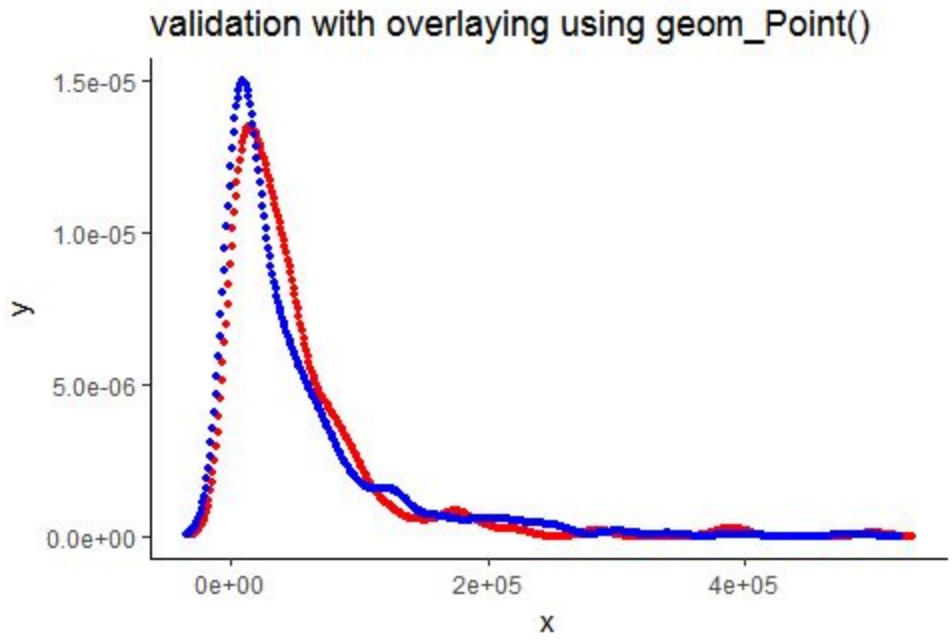
Then we generated random data(368 entries) using α and λ and plotted gamma distribution graph and its also right skewed graph.



For validation we generated overlaying graph using above two figure and we got the resulted figure like this



For more clarity we use geom_points here



It can be seen that if we generate random distribution of data using same α and λ it overlap with given data graph
At many points.

Ans c

In this part we generated 5000 random values using same α and λ and find its 60th percentile. The result is 37660.19 .

R code for part-1

```
library(ggplot2)            #including graph package ggplot2
data<- read.csv(file.choose(), header = T)  # Read and store data into a variable data
attach(data)
Demonitisation[Demonitisation == 'not Yes'] <- 'No'       # Convert Not Yes Into No
data1 <- data[,-8]
data2 <- cbind(data1, Demonitisation)
categoryinc <- cut(monthly.income, c(-1,30000,100000,500000), labels =
c("Poor","Middle","Rich")) #Categorize income into three groups
data3 <- cbind(data2, categoryinc) # join the column categoryinc to data
categoryage <- cut(age, c(-1,21,60,150), labels = c("Teenager", "Adult", "Old"))  #Categorize
age into three groups
data4 <- cbind(data3, categoryage)
data5=subset(data4,age<150)                 #Remove people with age more than 150
data6=subset(data5, Urban!="NA")               # Remove people with urban data not available

q1<- qplot(x=Demonitisation, data=data2, geom="bar")
q1+ggtitle("No. of ppl against/support of demonitisation")

q11<- qplot(x=categoryinc, data=data3,fill=Demonitisation, geom="bar")
q11+ggtitle("No. of ppl against/support of demonitisation income category wise")

q12<-qplot(x=Residence, data=data2,fill=Demonitisation,
geom="bar")+theme(legend.position="top",axis.text.x = element_text(angle=90, hjust=1))
q12+ggtitle("No. of ppl against/support of demonitisation state wise")

q13<- qplot(x=categoryage, data=data4,fill=Demonitisation, geom="bar")
q13+ggtitle("No. of ppl against/support of demonitisation age category wise")

q14<- qplot(x=sex, data=data2,fill=Demonitisation, geom="bar")
q14+ggtitle("No. of ppl against/support of demonitisation gender wise")

q15<- qplot(x=Urban, data=data2,fill=Demonitisation, geom="bar")
q15+ggtitle("No. of ppl against/support of demonitisation urban wise")

q16<- qplot(x=Urban, data=data6,fill=categoryinc, geom="bar")
q16+ggtitle("income category distribution urban/rural wise")

q17<- qplot(x=categoryage, data=data6,fill=categoryinc, geom="bar")
```

```
q17+ggtitle("income category distribution age wise")

q18<- qplot(x=Residence, data=data6,fill=categoryinc, geom="bar")+
theme(legend.position="top",axis.text.x = element_text(angle=90, hjust=1))
q18+ggtitle("income category distribution state wise")

q19<- qplot(x=Residence, data=data6,fill=Urban, geom="bar")+
theme(legend.position="top",axis.text.x = element_text(angle=90, hjust=1))
q19+ggtitle("Urban distribution state wise")

q20<- qplot(x=Residence, data=data6,fill=sex, geom="bar")+
theme(legend.position="top",axis.text.x = element_text(angle=90, hjust=1))
q20+ggtitle("sex ratio state wise")

q21 <- qplot(data=data6, x=categoryinc) +geom_bar(aes(fill=sex))+
scale_fill_manual(values=c("red","blue"))
q21+ggtitle("gender wise income category")
```

R code for part 2

```
library(ggplot2)
library(MASS)

data<- read.csv("Demon.csv", header = T)
attach(data)

mm<-mean(monthly.income)   # calculate mean
varm<-var(monthly.income)   # calculate variance
lambda<- mm/varm        #calculate lambda of gamma distribution with formula mean/variance
alpha<-lambda*mm            #calculate alpha of gamma distribution with formula mean*lambda

#six1<-quantile(monthly.income,c(.60))

den <- density(monthly.income)
dat <- data.frame(x = den$x, y = den$y)
g1<-ggplot(data = dat, aes(x = x, y = y)) + geom_point(size = 1, color="red") +theme_classic()
g1+ggtitle("gamma distribution of given income")
newg<- rgamma(368, alpha, lambda)      # generate gamma distribution with alpha and lambda
#six2<- quantile(newg,c(.60))

den2<- density(newg)
dat2<- data.frame(x = den2$x, y = den2$y)
g2<- ggplot(data = dat2, aes(x = x, y = y)) + geom_point(size = 1, color="blue")
+theme_classic()
g2+ggtitle("gamma distribution of genrated income using lambda and alpha")
overlay <- ggplot(data = dat2, aes(x = x, y = y)) + geom_line(data=dat, size = 1, color="red") +
geom_line(size = 1, color="blue") +theme_classic()
overlay+ggtitle("validation with overlaying using geom_line()")
overlay2 <- ggplot(data = dat2, aes(x = x, y = y)) + geom_point(data=dat, size = 1, color="red")
+ geom_point(size = 1, color="blue") +theme_classic()
overlay2+ggtitle("validation with overlaying using geom_Point()")

newg1<- rgamma(5000, alpha, lambda)          # generate 5000 values using gamma
distribution with alpha and lambda
six2<- quantile(newg,c(.60))
```
