SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23

PART B

(PART B: TO BE COMPLETED BY STUDENTS)

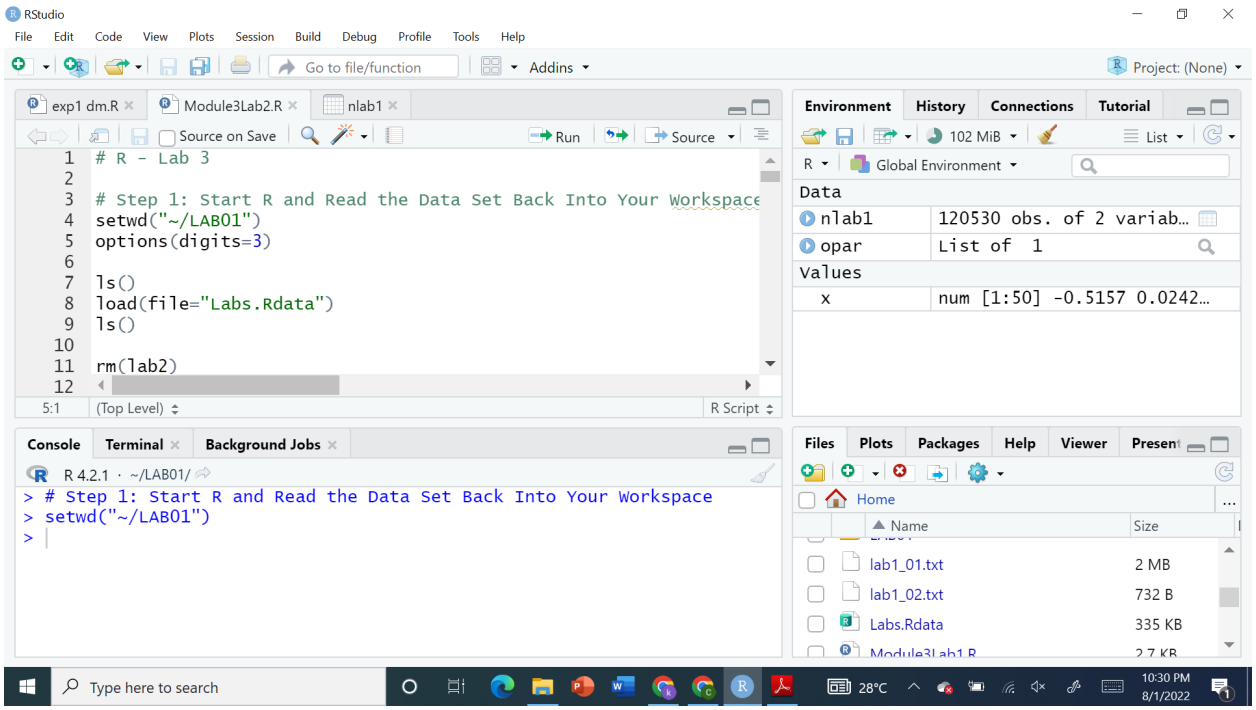
Roll No: C035	Name: Krisha Goti
Class: B	Batch: B1
Date of Experiment: 1/6/2022	Date of Submission: 13/08/2022
Grade	

B.1 Work done by student

(Paste your gather information and the comparison table)

1. **Prepare working environment for the Lab and load data files**
    1. Set the working directory to LAB01 where we have stored the data. On the console window type:

setwd("~/LAB01")

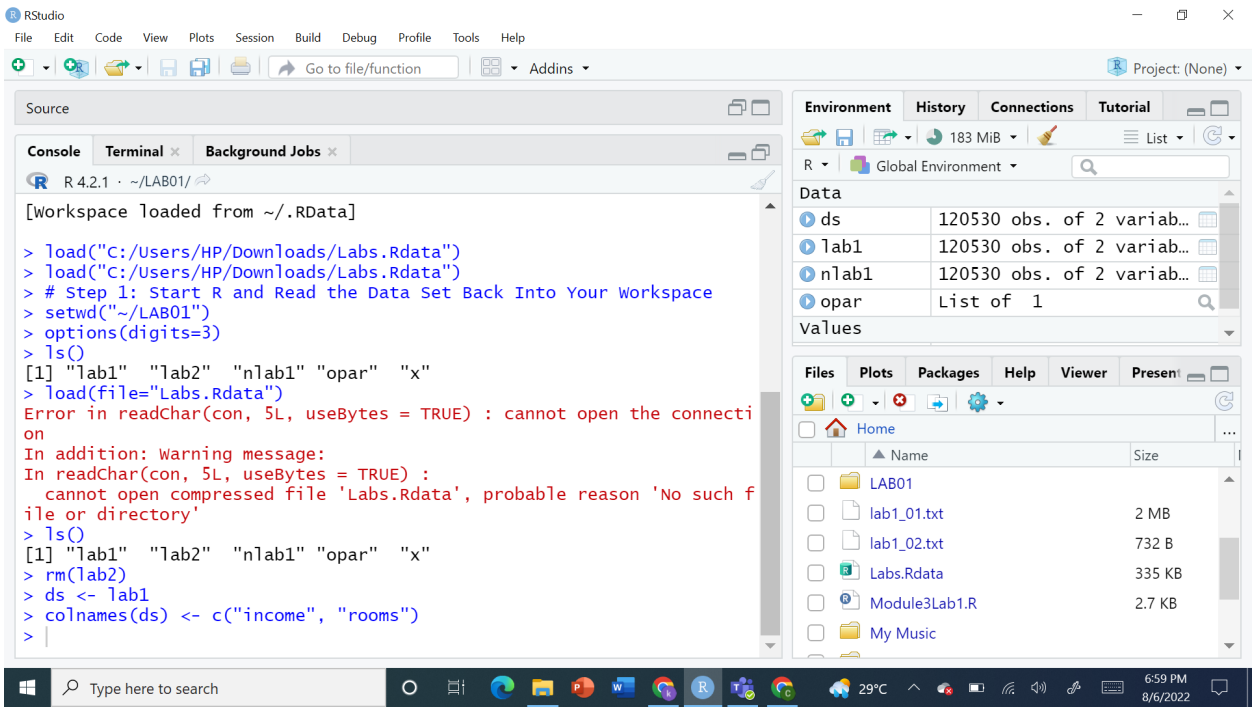


SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23

2. In the script window, open the script called "Module3Lab2.R". (Click on "File", "Open File" and Navigate to directory LAB03 and click on file "Module3Lab2.R").
   Start R and Read the Data Set Back Into Your Workspace:

3. Execute the following commands from the script window:
   ls()
   load(file="Labs.Rdata")
   ls()
   rm(lab2)
   ds <- lab1
   colnames(ds) <- c("income", "rooms")

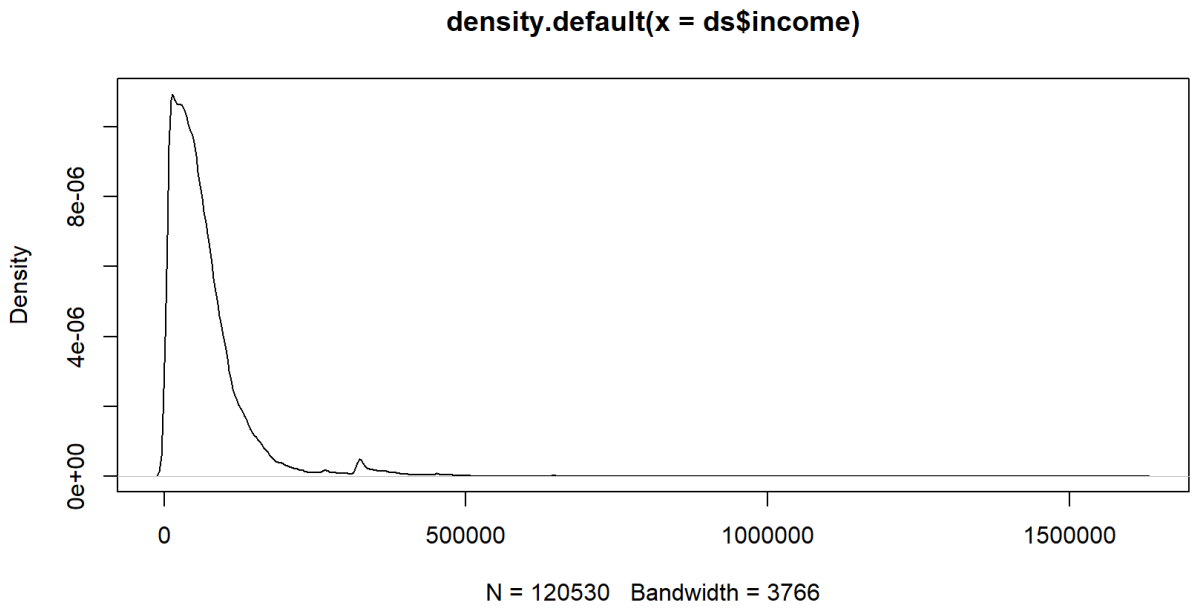


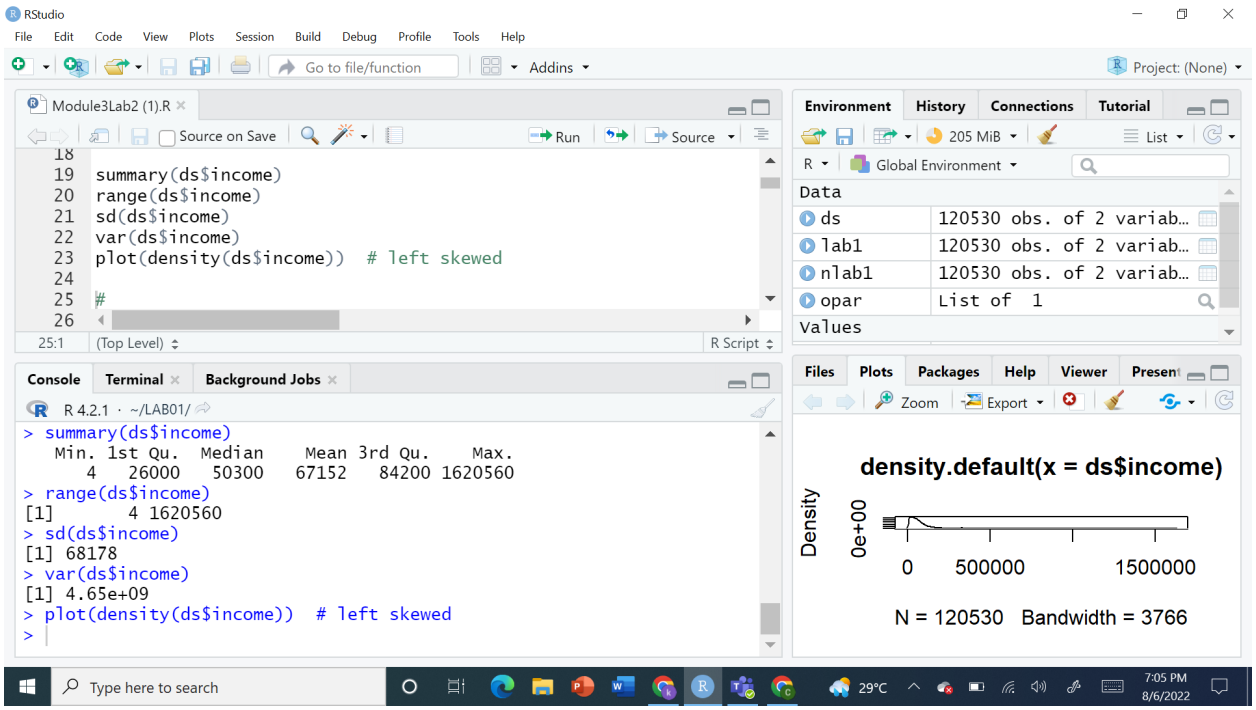
2. **Obtain summary statistics for Household Income and visualize data:**
   a. Execute the following commands from the script window:
   summary(ds$income)
   range(ds$income)
   sd(ds$income)

SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23

var(ds$income)
plot(density(ds$income)) # left skewed





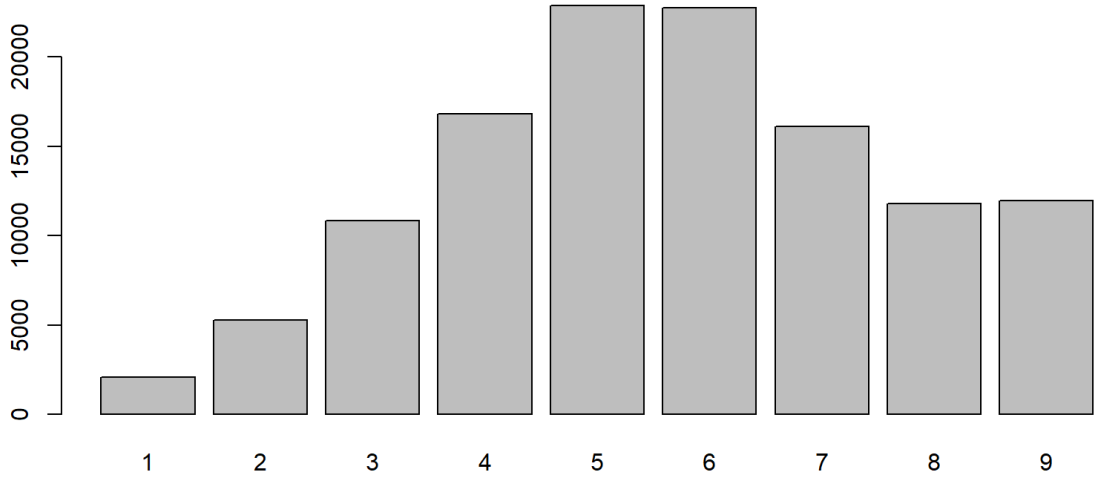
Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23

b. What is the mean? **67152**

c. What is the median? **50300**

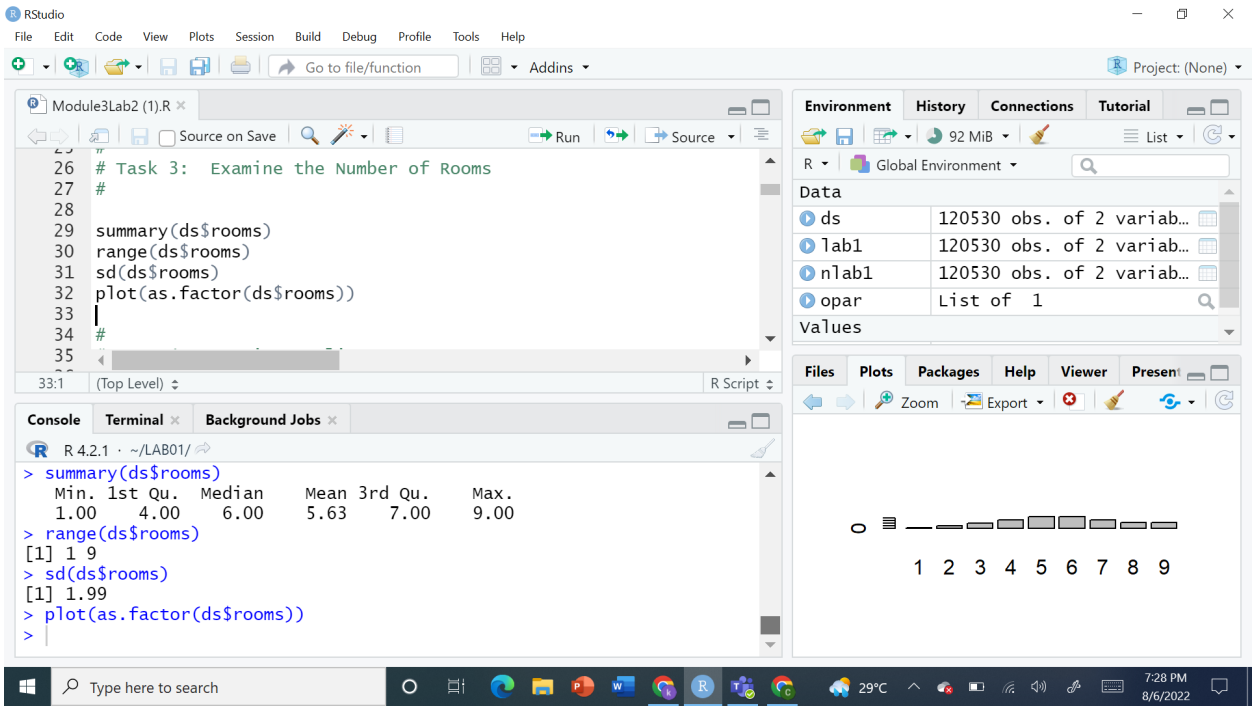
d. What is the standard deviation? **68178**

3. **Obtain summary statistics for Number of rooms and visualize data:**
   a. Execute the following commands from the script window:
   summary(ds$rooms)
   range(ds$rooms)
   sd(ds$rooms)
   plot(as.factor(ds$rooms))



SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23



b. What is the mean? **5.63**
c. What is the median? **6.00**
d. What is the standard deviation? **1.99**

4. **Remove Outliers**

In a previous lab, you recorded the range of income. You observed that the minimum household income is 4, and the maximum is 1,620,560.

1. Does this make sense to you? Why? *

The data used in the experiment are trim, so the values are different.

2. What happens if you throw out the top and bottom 10%? Execute the following line from the script window

(m <- mean(ds$income, trim=0.10))

3. How does this compare to the previous mean of this variable?

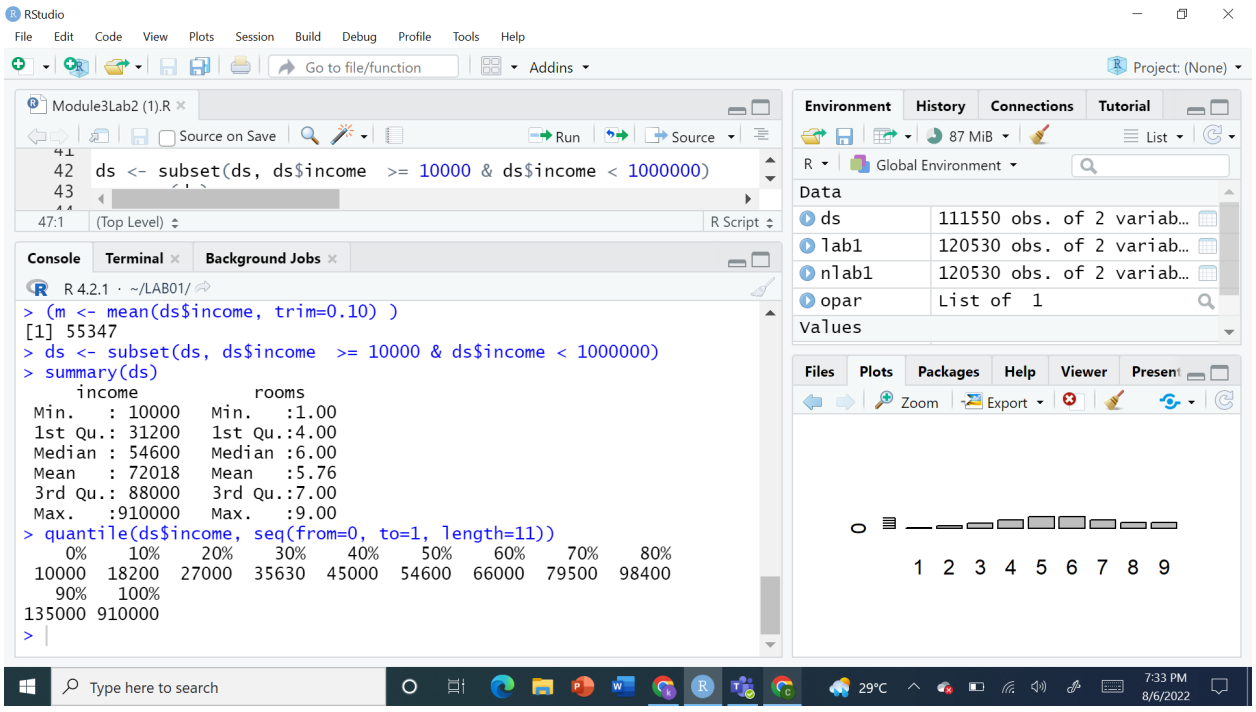
The mean value has increased

4. Execute the following commands from the script window:

SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23

ds <- subset(ds, ds$income >= 10000 & ds$income < 1000000)
summary(ds)
quantile(ds$income, seq(from=0, to=1, length=11))



5. How do these values vary from the values in the original data set?

Due to the range,the values of mean, median are more realistic compare to values of whole dataset

6. Do they make more sense? Yes
7. Which data set would you prefer to use? Trim Dataset

*We might consider the high and low value as outliers, and get rid of them. On the other hand, as we will discover, income is best described via a lognormal distribution, and hence these values are in the extreme ends +- 3 sds from the mean.

5. **Stratify Variable – Household Income and plot the results:**

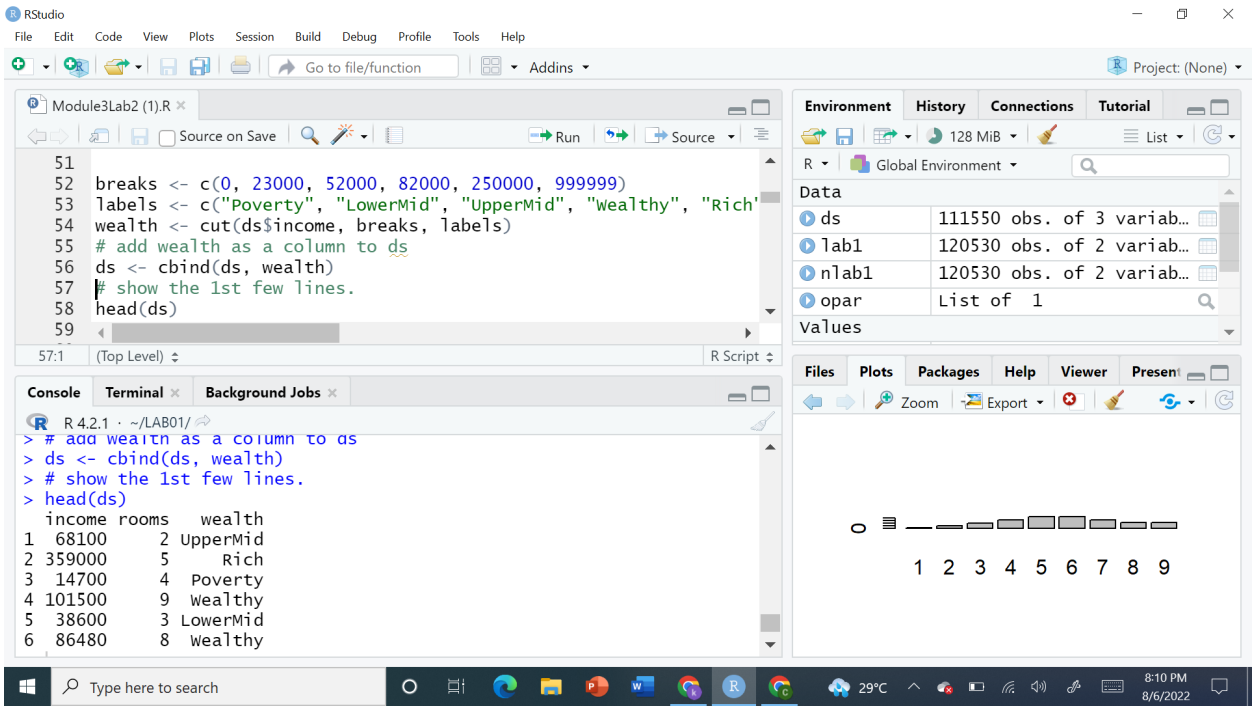
SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23

Stratify breaks that occur close to U.S. Guidelines for Poverty, Median Income, Wealth, and Rich (> $250k @ year)

1. Execute the following code (listed under comment heading "step 5" in the script file):

```
breaks <- c(0, 23000, 52000, 82000, 250000, 999999)
labels <- c("Poverty", "LowerMid", "UpperMid", "Wealthy", "Rich")
wealth <- cut(ds$income, breaks, labels)
# add wealth as a column to ds
ds <- cbind(ds, wealth)
# show the 1st few lines.
head(ds)
```

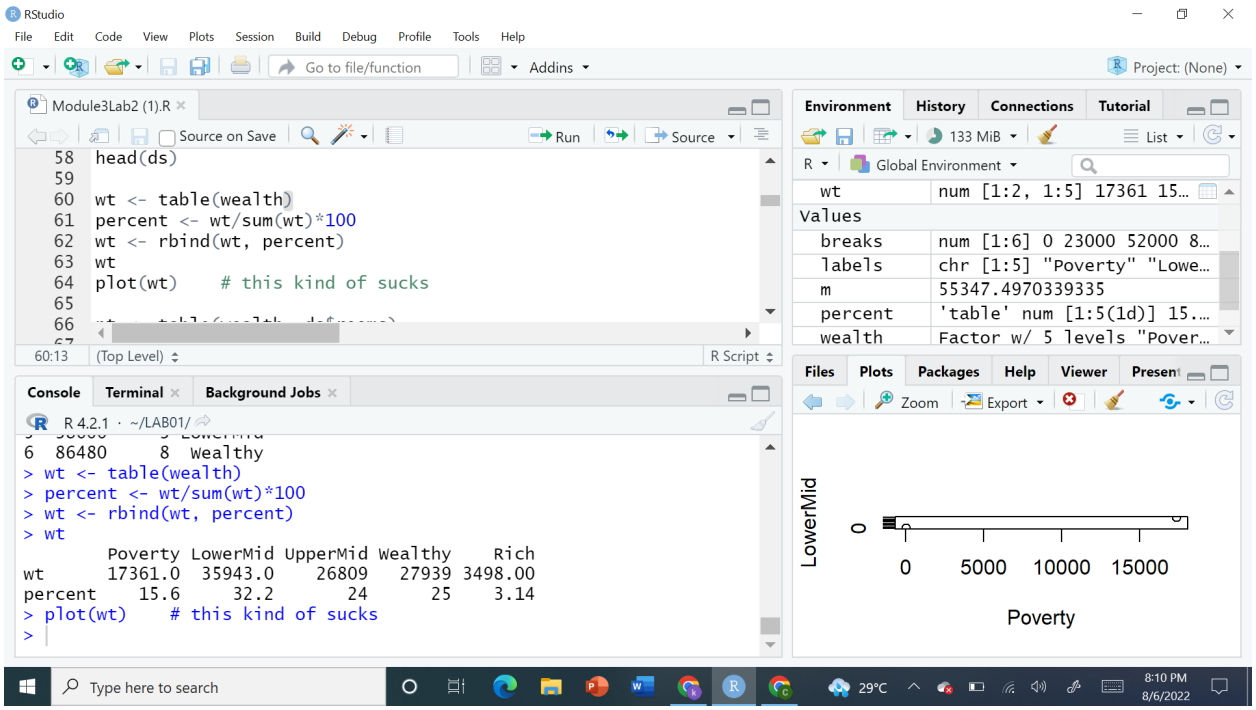


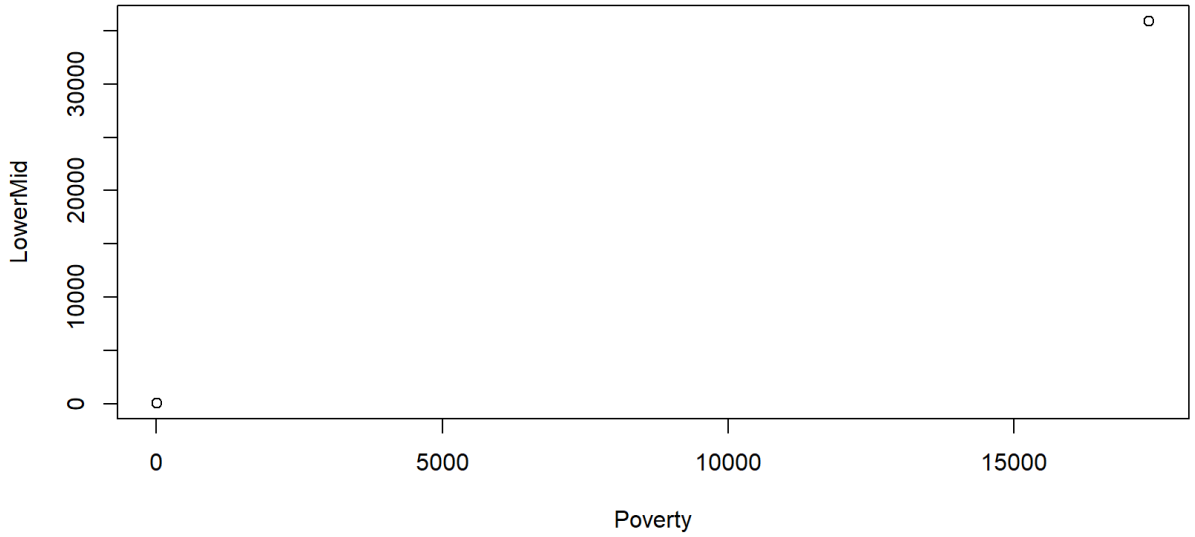
2. Continue to execute the remaining part of the code in Step 5

```
wt <- table(wealth)
percent <- wt/sum(wt)*100
wt <- rbind(wt, percent)
wt
plot(wt)
```

SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23



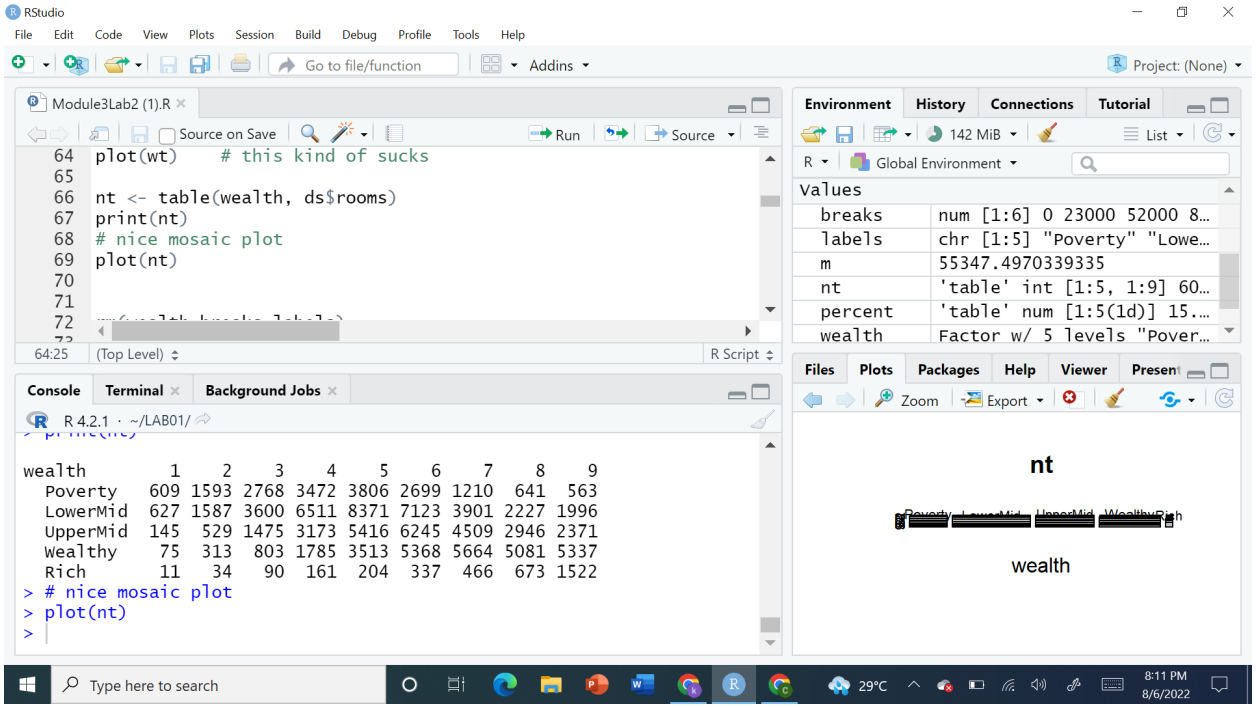


3. Take another look at the relationship between wealth and income. Execute the following lines:

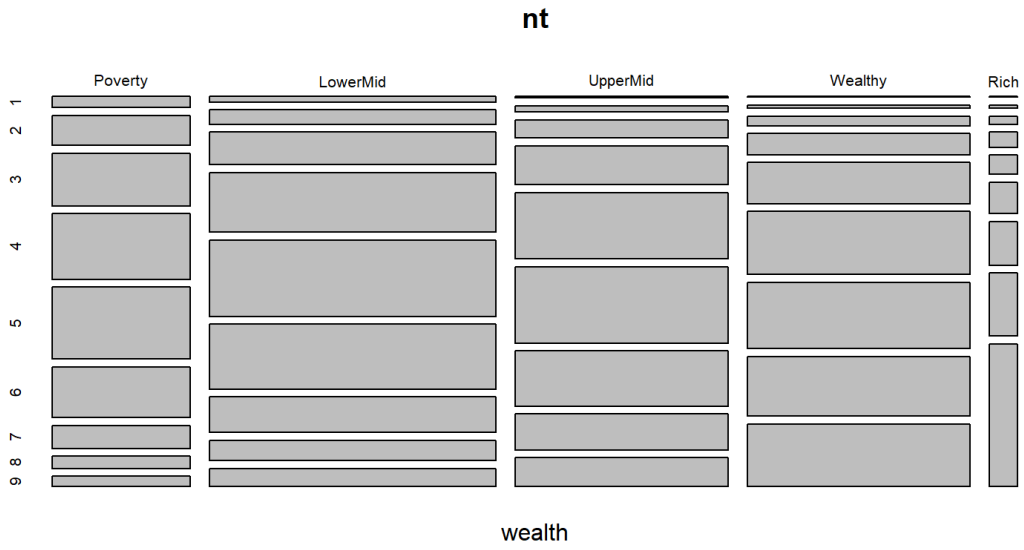
SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23

```
# take another look -- wealth by rooms
nt <- table(wealth, ds$rooms)
print(nt)
plot(nt)     # nice mosaic plot
```



Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23



4. Execute this code from the script file. These lines will remove the variables wealth, breaks and labels, and then save the variables data set and write into a file named "Census.Rdata".
   rm(wealth,breaks,labels)
   save(ds, wt, nt, file="Census.Rdata")

6. **Plot Histogram and Distributions:**

Problem: How do you represent income given the range of values?

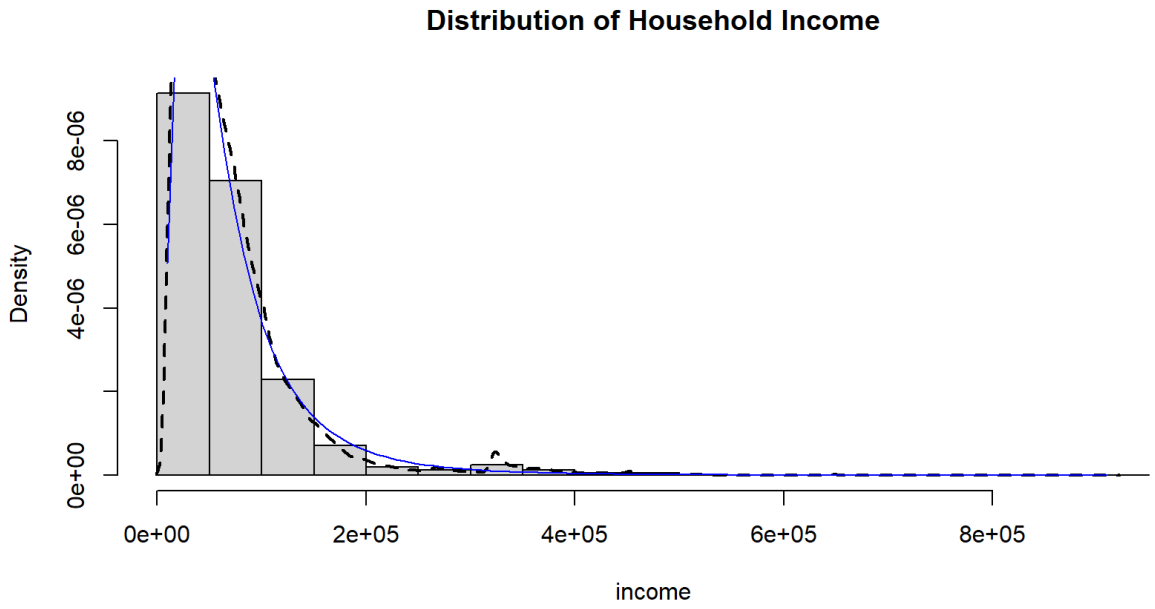
1. Select and execute the code under Step 6 Histograms and distributions in the script file.

```
library(MASS)

with(ds, {
hist(income, main="Distribution of Household Income", freq=FALSE)
lines(density(income), lty=2, lwd=2)
# line type (lty) 2 is dashed
xvals = seq(from=min(income), to=max(income), length=100)
param = fitdistr(income, "lognormal")
lines(xvals, dlnorm(xvals, meanlog = param$estimate[1], sdlog =
param$estimate[2]), col =”blue”)
        })
```

SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23

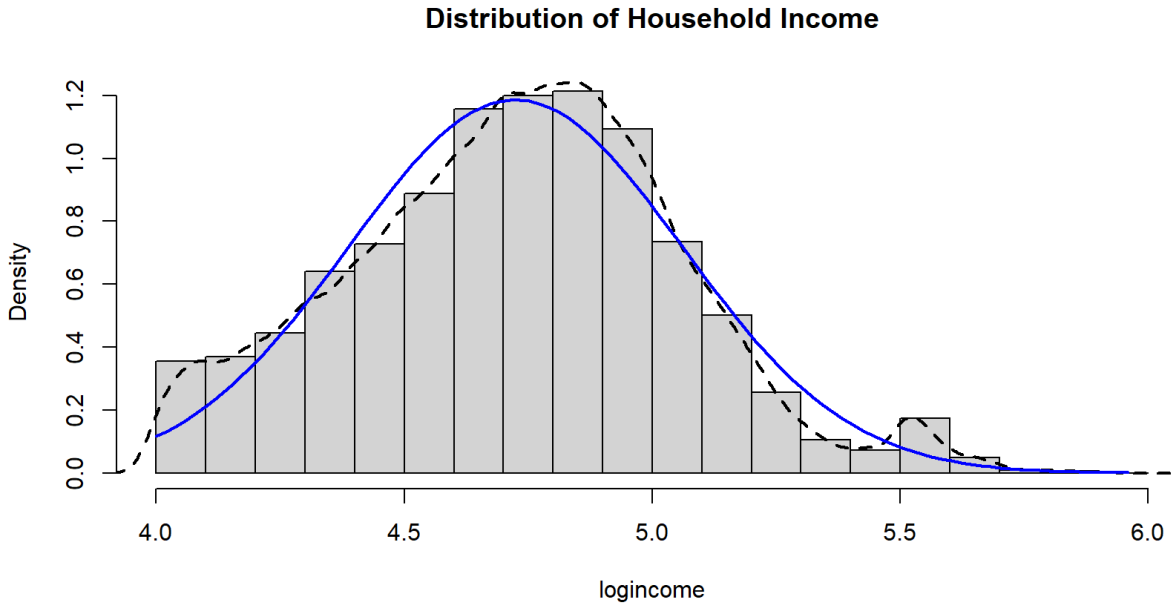


2. Now try the same thing with log10(income)

```
logincome = log10(ds$income)
hist(logincome, main="Distribution of Household Income", freq=FALSE)
# line type lty(2) is a dashed line
lines(density(logincome), lty=2, lwd=2)
xvals = seq(from=min(logincome), to=max(logincome), length=100)
param = fitdistr(logincome, "normal")
lines(xvals, dnorm(xvals, param$estimate[1], param$estimate[2]), lwd=2, col="blue")
```

SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23



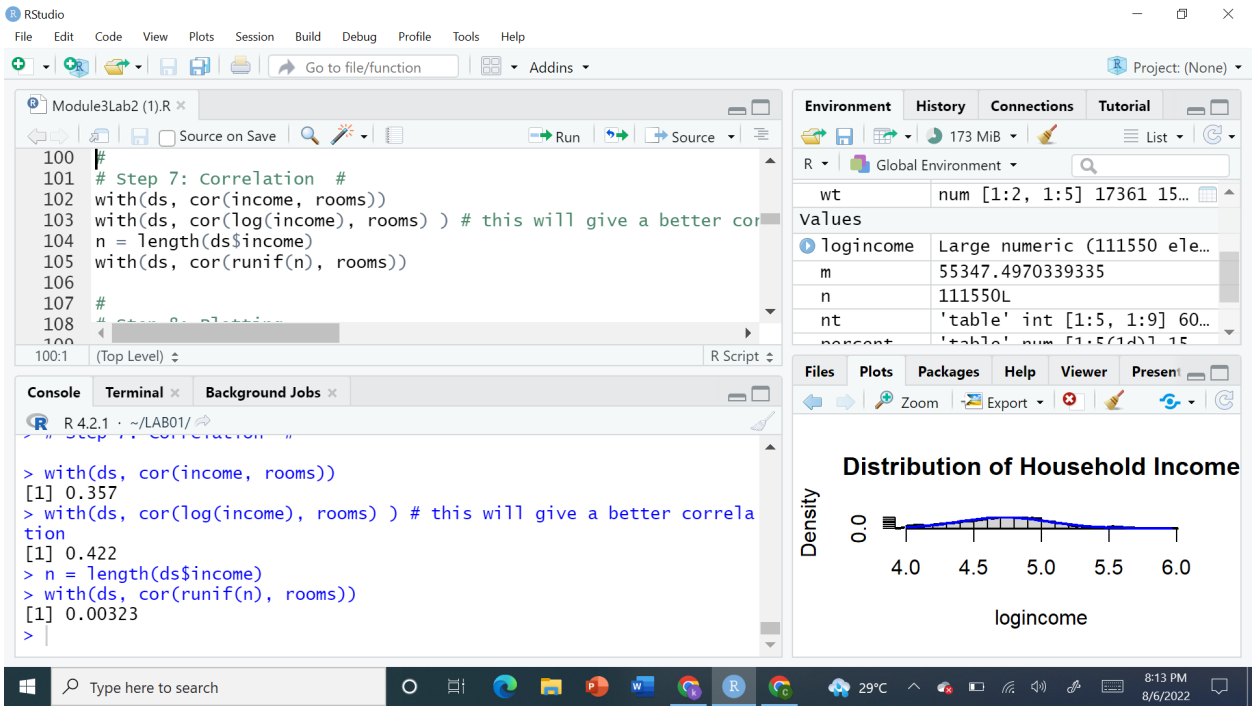
7. **Compute Correlation between income and number of rooms:**

1. You need to consider your hypothesis.
    1. Your hypothesis is that the number of rooms in a house is predicted by household income (the rich can buy bigger houses), e.g. *lm(rooms ~ income)*
    2. Therefore, our null hypothesis: no correlation between income and number of rooms.
    3. Alternate hypothesis: there is a correlation between income and the number of rooms.
2. Execute the following code (listed after the comment line "Step7 in the script file).
   with(ds, cor(income, rooms))
   with(ds, cor(log(income), rooms))) # this will give a better correlation

3. For comparison, correlate rooms with a completely unrelated variable.
   n = length(ds$income)
   with(ds, cor(runif(n), rooms))

SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23



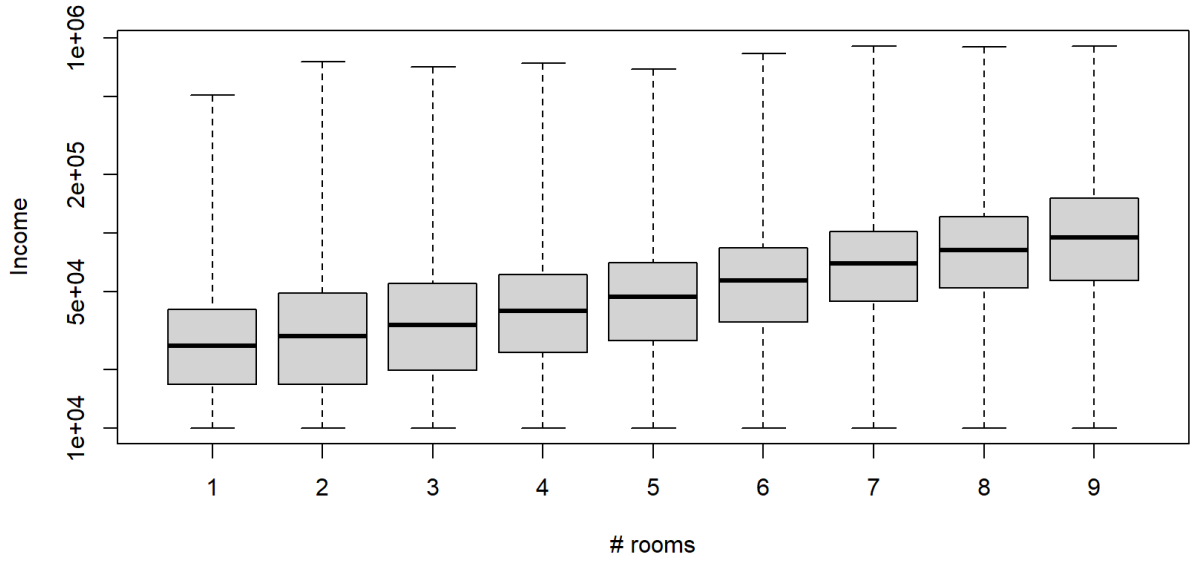
8. Create a Boxplot - Distribution of income as a factor of number of rooms:

1. Select and execute the code (Listed after the comment line "Step 8") in the script window.
2. Plot the distribution of income as a factor of # of rooms. 'log="y"' plots income on log scale. We will suppress the outlier points and let the whiskers cover the full range of the data.

   boxplot(income ~ as.factor(rooms), data=ds, range=0, outline=F, log="y", xlab="# rooms", ylab="Income")

SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

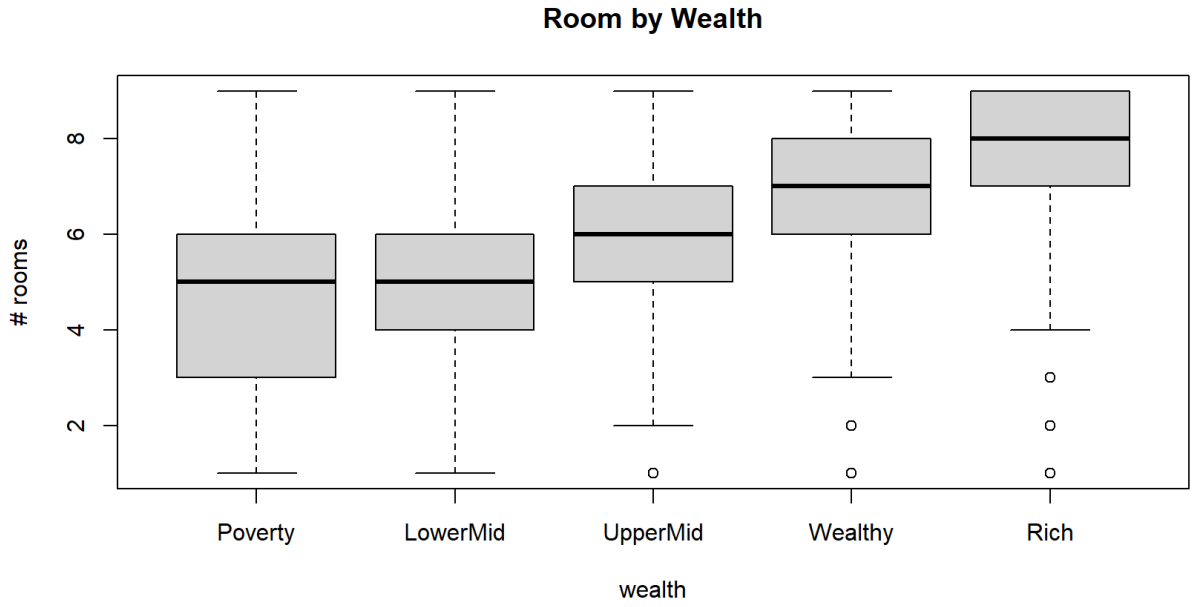
Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23



3. Plot the # of rooms as a function of wealth level.
   boxplot(rooms ~ wealth, data = ds, main="Room by Wealth", Xlab="Category", ylab="# rooms")

   # we‟ll keep the outlier points in this one

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23



B.2 Conclusion

After completing this experiment, I am able to- Apply appropriate analytic techniques and tools to analyze big data, create statistical models and identify insights leading to actionable results. Mine given data set using data mining tool.