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)

(Students must submit the soft copy as per the following segments within two hours of the practicals. The soft copy must be uploaded on Blackboard LMS or emailed to the concerned Lab in charge Faculties at the end of practical; in case Blackboard is not accessible)

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

B.1 Work done by student

(Paste your gather information and the comparison table)

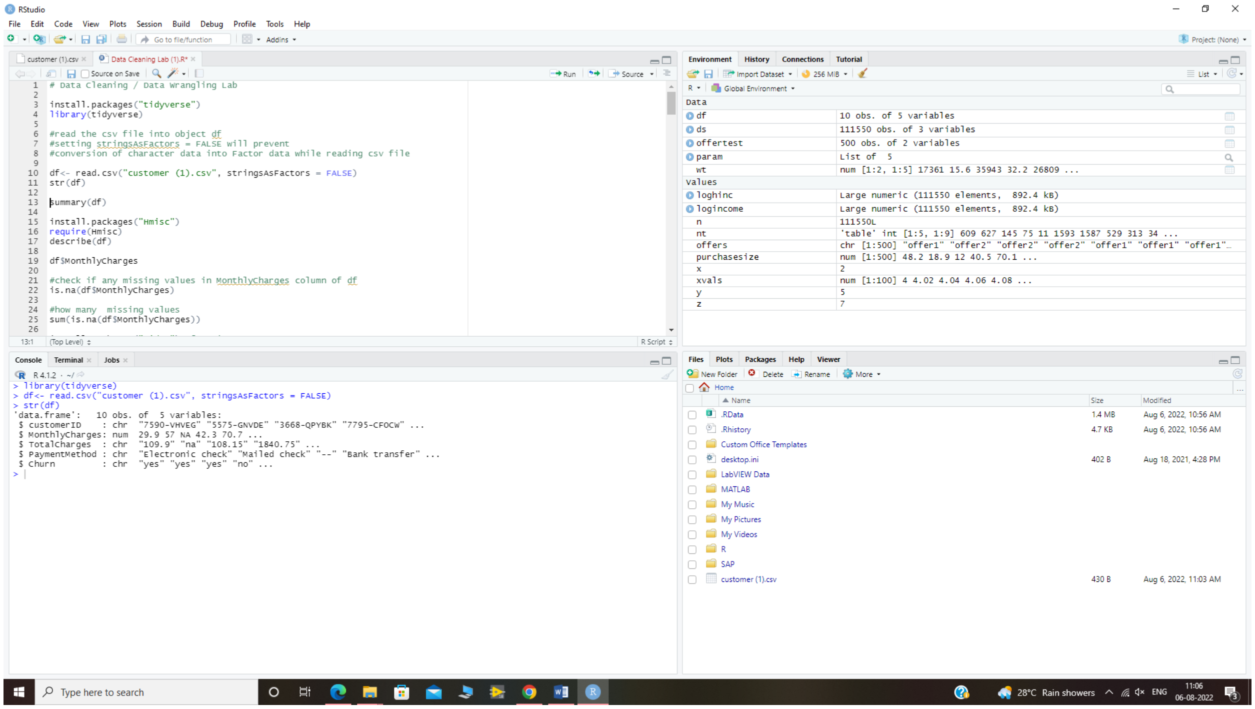
1. **Study the working of following commands in R from R documentation by typing them in the 'help' tab**
   a. **read.csv**

2. **Prepare working environment for the Lab and load data files**
   1. Set the working directory to where we have stored the data.
   2. Read customer.csv dataset using read.csv command

```
df<- read.csv("customer.csv", stringsAsFactors = FALSE)
```

   3. Display structure of dataframe df

```
str(df)
```



```
# Data Cleaning / Data Wrangling Lab

install.packages("tidyverse")
library(tidyverse)

#read the csv file into object df
#setting stringsAsFactors = FALSE will prevent
#conversion of character data into Factor data while reading csv file

df<- read.csv("customer (1).csv", stringsAsFactors = FALSE)
str(df)

summary(df)

install.packages("Hmisc")
require(Hmisc)
describe(df)

df$MonthlyCharges

#check if any missing values in MonthlyCharges column of df
is.na(df$MonthlyCharges)

#how many  missing values
sum(is.na(df$MonthlyCharges))
```

Data	
df	10 obs. of 5 variables
ds	111550 obs. of 3 variables
offertest	500 obs. of 2 variables
param	List of 5
wt	num [1:2, 1:5] 17361 15.6 35943 32.2 26809 ...
Values	
loghinc	Large numeric (111550 elements, 892.4 kB)
logincome	Large numeric (111550 elements, 892.4 kB)
n	111550L
nt	'table' int [1:5, 1:9] 609 627 145 75 11 1593 1587 529 313 34 ...
offers	chr [1:500] "offer1" "offer2" "offer2" "offer2" "offer1" "offer1" "offer1"…
purchasesize	num [1:500] 48.2 18.9 12 40.5 70.1 ...
x	2
xvals	num [1:100] 4 4.02 4.04 4.06 4.08 ...
y	5
z	7

```
> library(tidyverse)
> df<- read.csv("customer (1).csv", stringsAsFactors = FALSE)
> str(df)
'data.frame':   10 obs. of  5 variables:
 $ customerID    : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
 $ MonthlyCharges: num  29.9 57 NA 42.3 70.7 ...
 $ TotalCharges  : chr  "109.9" "na" "108.15" "1840.75" ...
 $ PaymentMethod : chr  "Electronic check" "Mailed check" "--" "Bank transfer" ...
 $ Churn         : chr  "yes" "yes" "yes" "no" ...
> 
```

Name	Size	Modified
.RData	1.4 MB	Aug 6, 2022, 10:56 AM
.Rhistory	4.7 KB	Aug 6, 2022, 10:56 AM
Custom Office Templates		
desktop.ini	402 B	Aug 18, 2021, 4:28 PM
LabVIEW Data		
MATLAB		
My Music		
My Pictures		
My Videos		
R		
SAP		
customer (1).csv	430 B	Aug 6, 2022, 11:03 AM

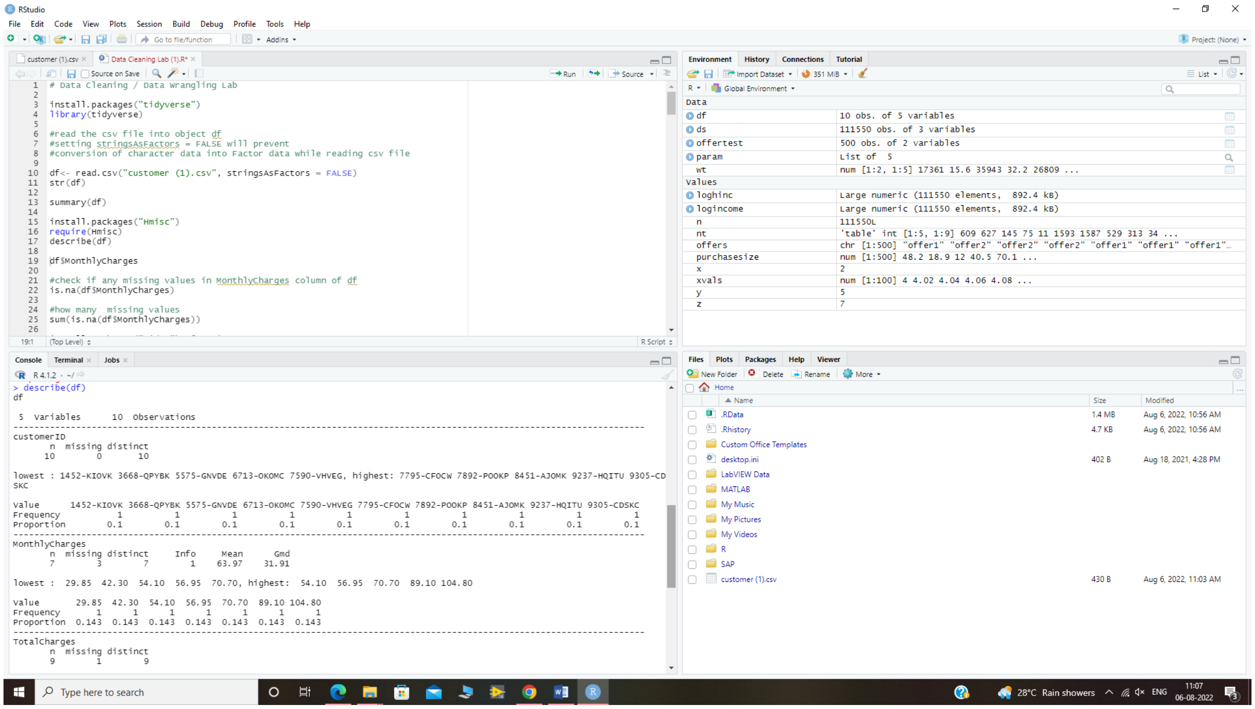
4. Display summary of dataframe to know five point summary of each attribute
5. Install package Harrell Miscellaneous (hmisc)

```
install.packages("Hmisc")
require(Hmisc)
```

This package contains many functions useful for data analysis, high-level graphics, utility operations, functions for computing sample size and power, importing and annotating datasets, imputing missing values, advanced table making, variable clustering, character string manipulation, conversion of R objects to LaTeX and html code, and recoding variables.
require(package) load the namespace of the package with name package and attach it on the search list. require is designed for use inside other functions; it
returns FALSE and gives a warning (rather than an error as library() does by default) if the package does not exist.

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



```
# Data Cleaning / Data Wrangling Lab

install.packages("tidyverse")
library(tidyverse)

#read the csv file into object df
#setting stringsAsFactors = FALSE will prevent
#conversion of character data into Factor data while reading csv file

df<- read.csv("customer (1).csv", stringsAsFactors = FALSE)
str(df)

summary(df)

install.packages("Hmisc")
require(Hmisc)
describe(df)

df$MonthlyCharges

#check if any missing values in MonthlyCharges column of df
is.na(df$MonthlyCharges)

#how many  missing values
sum(is.na(df$MonthlyCharges))

```

```
> describe(df)
df

 5  Variables      10  Observations
--------------------------------------------------------------------------------------------------------------------------
customerID
       n  missing distinct
      10        0       10

lowest : 1452-KIOVK 3668-QPYBK 5575-GNVDE 6713-OKOMC 7590-VHVEG, highest: 7795-CFOCW 7892-POOKP 8451-AJOMK 9237-HQITU 9305-CD
SKC

Value      1452-KIOVK 3668-QPYBK 5575-GNVDE 6713-OKOMC 7590-VHVEG 7795-CFOCW 7892-POOKP 8451-AJOMK 9237-HQITU 9305-CDSKC
Frequency           1          1          1          1          1          1          1          1          1          1
Proportion        0.1        0.1        0.1        0.1        0.1        0.1        0.1        0.1        0.1        0.1
--------------------------------------------------------------------------------------------------------------------------
MonthlyCharges
       n  missing distinct     Info     Mean      Gmd
       7        3        7        1    63.97    31.91

lowest :  29.85  42.30  54.10  56.95  70.70, highest:  54.10  56.95  70.70  89.10 104.80

Value       29.85  42.30  54.10  56.95  70.70  89.10 104.80
Frequency       1      1      1      1      1      1      1
Proportion  0.143  0.143  0.143  0.143  0.143  0.143  0.143
--------------------------------------------------------------------------------------------------------------------------
TotalCharges
       n  missing distinct
       9        1        9
```

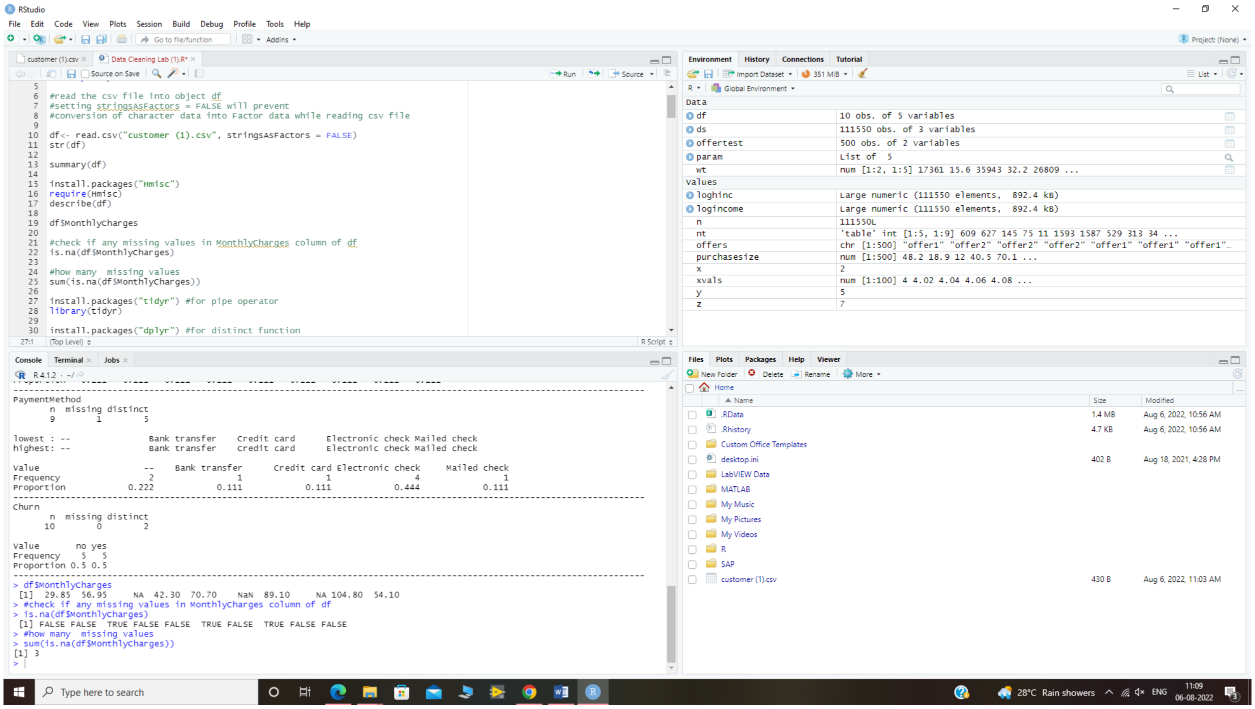
Data	
df	10 obs. of 5 variables
ds	111550 obs. of 3 variables
offertest	500 obs. of 2 variables
param	List of 5
wt	num [1:2, 1:5] 17361 15.6 35943 32.2 26809 ...
Values	
loghinc	Large numeric (111550 elements, 892.4 kB)
logincome	Large numeric (111550 elements, 892.4 kB)
n	111550L
nt	'table' int [1:5, 1:9] 609 627 145 75 11 1593 1587 529 313 34 ...
offers	chr [1:500] "offer1" "offer2" "offer2" "offer2" "offer1" "offer1" "offer1"…
purchasesize	num [1:500] 48.2 18.9 12 40.5 70.1 ...
x	2
xvals	num [1:100] 4 4.02 4.04 4.06 4.08 ...
y	5
z	7

Name	Size	Modified
.RData	1.4 MB	Aug 6, 2022, 10:56 AM
.Rhistory	4.7 KB	Aug 6, 2022, 10:56 AM
Custom Office Templates		
desktop.ini	402 B	Aug 18, 2021, 4:28 PM
LabVIEW Data		
MATLAB		
My Music		
My Pictures		
My Videos		
R		
SAP		
customer (1).csv	430 B	Aug 6, 2022, 11:03 AM

6. Generate a concise statistical description of dataframe using describe command

7. Display monthlycharges column
8. Check if any missing values are there in column
   `is.na(df$MonthlyCharges)`
9. Find how many missing values are there
   `sum(is.na(df$MonthlyCharges))`

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



```

#read the csv file into object df
#setting stringsAsFactors = FALSE will prevent
#conversion of character data into Factor data while reading csv file

df<- read.csv("customer (1).csv", stringsAsFactors = FALSE)
str(df)

summary(df)

install.packages("Hmisc")
require(Hmisc)
describe(df)

df$MonthlyCharges

#check if any missing values in MonthlyCharges column of df
is.na(df$MonthlyCharges)

#how many  missing values
sum(is.na(df$MonthlyCharges))

install.packages("tidyr") #for pipe operator
library(tidyr)

install.packages("dplyr") #for distinct function
```

```
--------------------------------------------------------------------------------------------------------------
PaymentMethod
       n  missing distinct
       9        1        5

lowest : --                 Bank transfer    Credit card      Electronic check Mailed check
highest: --                 Bank transfer    Credit card      Electronic check Mailed check

Value                    --    Bank transfer      Credit card Electronic check     Mailed check
Frequency                 2                1                1                4                1
Proportion            0.222            0.111            0.111            0.444            0.111
--------------------------------------------------------------------------------------------------------------
Churn
       n  missing distinct
      10        0        2

Value        no yes
Frequency     5   5
Proportion 0.5 0.5
--------------------------------------------------------------------------------------------------------------
> df$MonthlyCharges
 [1]  29.85  56.95     NA  42.30  70.70    NaN  89.10     NA 104.80  54.10
> #check if any missing values in MonthlyCharges column of df
> is.na(df$MonthlyCharges)
 [1] FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE
> #how many  missing values
> sum(is.na(df$MonthlyCharges))
[1] 3
> 
```

Data	
df	10 obs. of 5 variables
ds	111550 obs. of 3 variables
offertest	500 obs. of 2 variables
param	List of 5
wt	num [1:2, 1:5] 17361 15.6 35943 32.2 26809 ...
Values	
loghinc	Large numeric (111550 elements, 892.4 kB)
logincome	Large numeric (111550 elements, 892.4 kB)
n	111550L
nt	'table' int [1:5, 1:9] 609 627 145 75 11 1593 1587 529 313 34 ...
offers	chr [1:500] "offer1" "offer2" "offer2" "offer2" "offer1" "offer1" "offer1"…
purchasesize	num [1:500] 48.2 18.9 12 40.5 70.1 ...
x	2
xvals	num [1:100] 4 4.02 4.04 4.06 4.08 ...
y	5
z	7

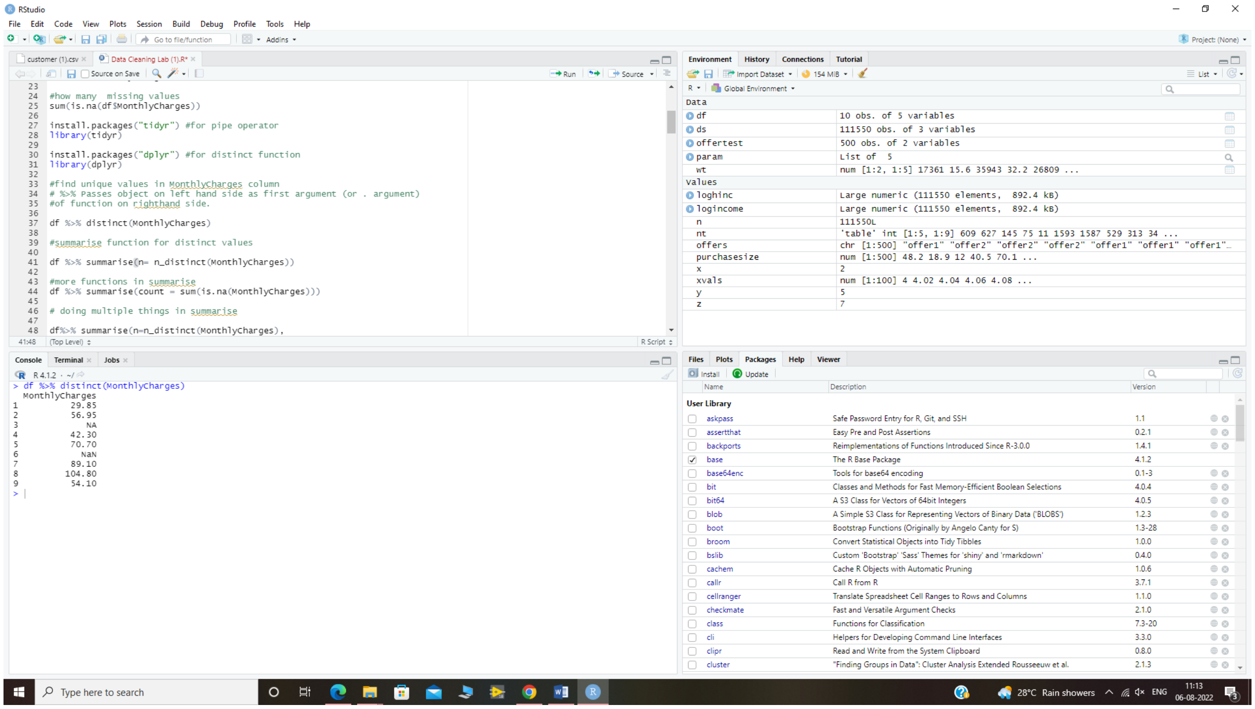
Name	Size	Modified
.RData	1.4 MB	Aug 6, 2022, 10:56 AM
.Rhistory	4.7 KB	Aug 6, 2022, 10:56 AM
Custom Office Templates		
desktop.ini	402 B	Aug 18, 2021, 4:28 PM
LabVIEW Data		
MATLAB		
My Music		
My Pictures		
My Videos		
R		
SAP		
customer (1).csv	430 B	Aug 6, 2022, 11:03 AM

10. Install package tidyr and dplyr

```
install.packages("tidyr") #for pipe operator
library(tidyr)
install.packages("dplyr") #for distinct function
library(dplyr)
```

11. find unique values in MonthlyCharges column

```
df %>% distinct(MonthlyCharges)
```



```
#how many  missing values
sum(is.na(df$MonthlyCharges))

install.packages("tidyr") #for pipe operator
library(tidyr)

install.packages("dplyr") #for distinct function
library(dplyr)

#find unique values in MonthlyCharges column
# %>% Passes object on left hand side as first argument (or . argument)
#of function on righthand side.

df %>% distinct(MonthlyCharges)

#summarise function for distinct values

df %>% summarise(n= n_distinct(MonthlyCharges))

#more functions in summarise
df %>% summarise(count = sum(is.na(MonthlyCharges)))

# doing multiple things in summarise

df%>% summarise(n=n_distinct(MonthlyCharges),
```

```
> df %>% distinct(MonthlyCharges)
  MonthlyCharges
1          29.85
2          56.95
3             NA
4          42.30
5          70.70
6            NaN
7          89.10
8         104.80
9          54.10
> 
```

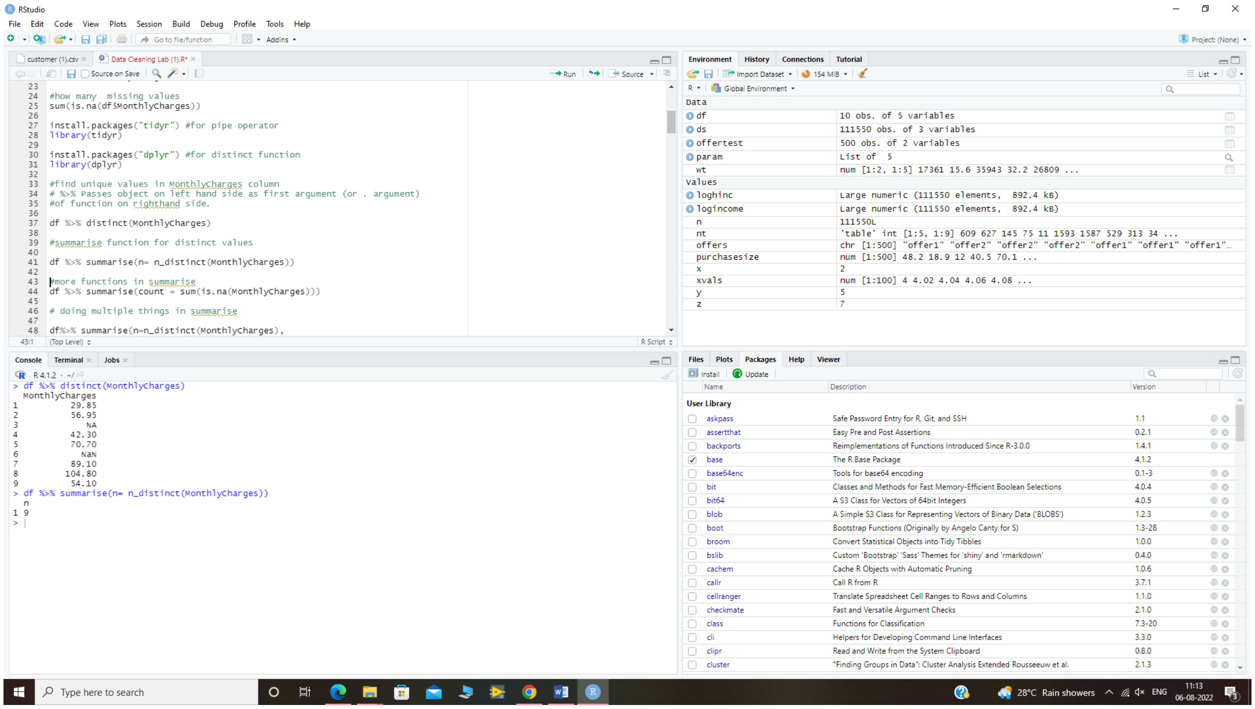
Data	
df	10 obs. of 5 variables
ds	111550 obs. of 3 variables
offertest	500 obs. of 2 variables
param	List of 5
wt	num [1:2, 1:5] 17361 15.6 35943 32.2 26809 ...
Values	
loghinc	Large numeric (111550 elements, 892.4 kB)
logincome	Large numeric (111550 elements, 892.4 kB)
n	111550L
nt	'table' int [1:5, 1:9] 609 627 145 75 11 1593 1587 529 313 34 ...
offers	chr [1:500] "offer1" "offer2" "offer2" "offer2" "offer1" "offer1" "offer1"…
purchasesize	num [1:500] 48.2 18.9 12 40.5 70.1 ...
x	2
xvals	num [1:100] 4 4.02 4.04 4.06 4.08 ...
y	5
z	7

	Name	Description	Version
	User Library		
□	askpass	Safe Password Entry for R, Git, and SSH	1.1
□	assertthat	Easy Pre and Post Assertions	0.2.1
□	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.4.1
☑	base	The R Base Package	4.1.2
□	base64enc	Tools for base64 encoding	0.1-3
□	bit	Classes and Methods for Fast Memory-Efficient Boolean Selections	4.0.4
□	bit64	A S3 Class for Vectors of 64bit Integers	4.0.5
□	blob	A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS')	1.2.3
□	boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-28
□	broom	Convert Statistical Objects into Tidy Tibbles	1.0.0
□	bslib	Custom 'Bootstrap' 'Sass' Themes for 'shiny' and 'rmarkdown'	0.4.0
□	cachem	Cache R Objects with Automatic Pruning	1.0.6
□	callr	Call R from R	3.7.1
□	cellranger	Translate Spreadsheet Cell Ranges to Rows and Columns	1.1.0
□	checkmate	Fast and Versatile Argument Checks	2.1.0
□	class	Functions for Classification	7.3-20
□	cli	Helpers for Developing Command Line Interfaces	3.3.0
□	clipr	Read and Write from the System Clipboard	0.8.0
□	cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.	2.1.3

12. summarise distinct values

```
df %>% summarise(n= n_distinct(MonthlyCharges))
```

'n_distinct' Efficiently count the number of unique values in a set of vector. **This is a faster and more concise equivalent of length(unique(x))**



```
#how many  missing values
sum(is.na(df$MonthlyCharges))

install.packages("tidyr") #for pipe operator
library(tidyr)

install.packages("dplyr") #for distinct function
library(dplyr)

#find unique values in MonthlyCharges column
# %>% Passes object on left hand side as first argument (or . argument)
#of function on righthand side.

df %>% distinct(MonthlyCharges)

#summarise function for distinct values

df %>% summarise(n= n_distinct(MonthlyCharges))

#more functions in summarise
df %>% summarise(count = sum(is.na(MonthlyCharges)))

# doing multiple things in summarise

df%>% summarise(n=n_distinct(MonthlyCharges),
```

Data	
df	10 obs. of 5 variables
ds	111550 obs. of 3 variables
offertest	500 obs. of 2 variables
param	List of 5
wt	num [1:2, 1:5] 17361 15.6 35943 32.2 26809 ...
Values	
loghinc	Large numeric (111550 elements, 892.4 kB)
logincome	Large numeric (111550 elements, 892.4 kB)
n	111550L
nt	'table' int [1:5, 1:9] 609 627 145 75 11 1593 1587 529 313 34 ...
offers	chr [1:500] "offer1" "offer2" "offer2" "offer2" "offer1" "offer1" "offer1"...
purchasesize	num [1:500] 48.2 18.9 12 40.5 70.1 ...
x	2
xvals	num [1:100] 4 4.02 4.04 4.06 4.08 ...
y	5
z	7

```
> df %>% distinct(MonthlyCharges)
  MonthlyCharges
1          29.85
2          56.95
3             NA
4          42.30
5          70.70
6            NaN
7          89.10
8         104.80
9          54.10
> df %>% summarise(n= n_distinct(MonthlyCharges))
  n
1 9
> |
```

	Name	Description	Version
	User Library		
☐	askpass	Safe Password Entry for R, Git, and SSH	1.1
☐	assertthat	Easy Pre and Post Assertions	0.2.1
☐	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.4.1
☑	base	The R Base Package	4.1.2
☐	base64enc	Tools for base64 encoding	0.1-3
☐	bit	Classes and Methods for Fast Memory-Efficient Boolean Selections	4.0.4
☐	bit64	A S3 Class for Vectors of 64bit Integers	4.0.5
☐	blob	A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS')	1.2.3
☐	boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-28
☐	broom	Convert Statistical Objects into Tidy Tibbles	1.0.0
☐	bslib	Custom 'Bootstrap' 'Sass' Themes for 'shiny' and 'rmarkdown'	0.4.0
☐	cachem	Cache R Objects with Automatic Pruning	1.0.6
☐	callr	Call R from R	3.7.1
☐	cellranger	Translate Spreadsheet Cell Ranges to Rows and Columns	1.1.0
☐	checkmate	Fast and Versatile Argument Checks	2.1.0
☐	class	Functions for Classification	7.3-20
☐	cli	Helpers for Developing Command Line Interfaces	3.3.0
☐	clipr	Read and Write from the System Clipboard	0.8.0
☐	cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.	2.1.3

13. doing multiple things in summarise

```
df%>% summarise(n=n_distinct(MonthlyCharges),
                count = sum(is.na(MonthlyCharges)),
                M = mean(MonthlyCharges, na.rm=TRUE))
```

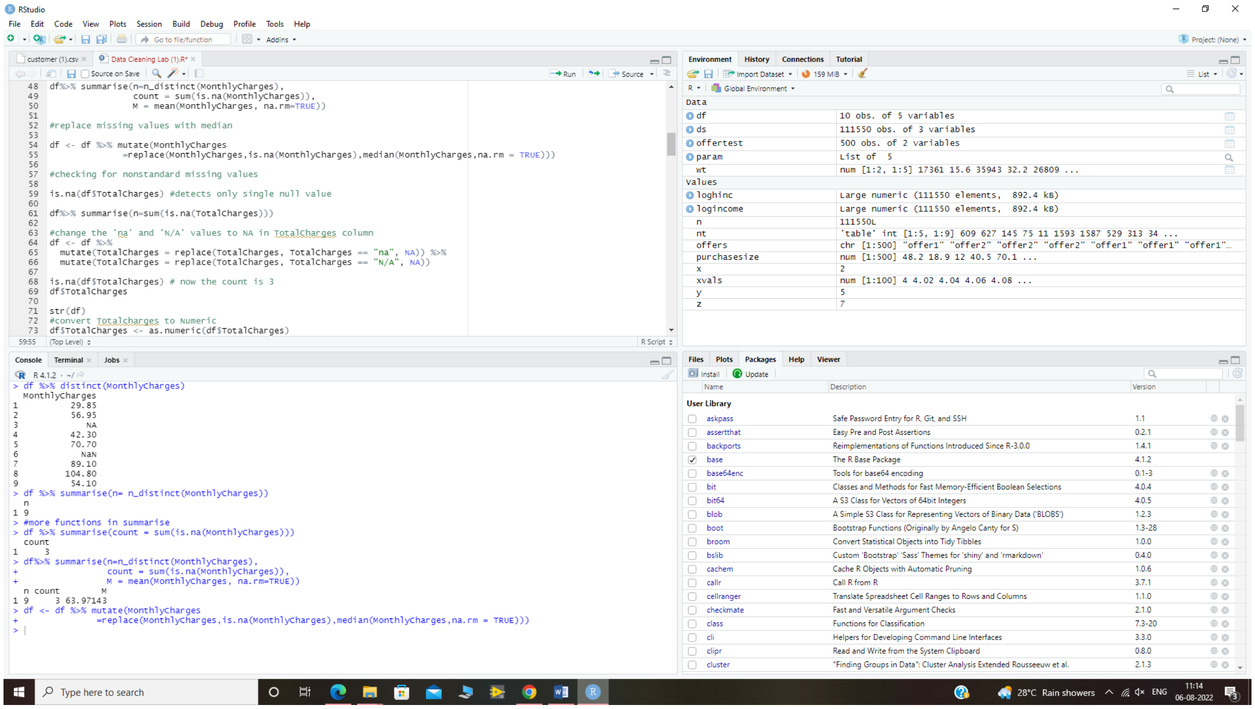
14. replace missing values with median

```
df <- df %>% mutate(MonthlyCharges
=replace(MonthlyCharges,is.na(MonthlyCharges),median(Mont
hlyCharges,na.rm = TRUE)))
```

15. checking for nonstandard missing values:

```
is.na(df$TotalCharges) #detects only single null value
df%>% summarise(n=sum(is.na(TotalCharges)))
```

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



```
df%>% summarise(n=n_distinct(MonthlyCharges),
                count = sum(is.na(MonthlyCharges)),
                M = mean(MonthlyCharges, na.rm=TRUE))

#replace missing values with median

df <- df %>% mutate(MonthlyCharges
              =replace(MonthlyCharges,is.na(MonthlyCharges),median(MonthlyCharges,na.rm = TRUE)))

#checking for nonstandard missing values

is.na(df$TotalCharges) #detects only single null value

df%>% summarise(n=sum(is.na(TotalCharges)))

#change the 'na' and 'N/A' values to NA in TotalCharges column
df <- df %>%
  mutate(TotalCharges = replace(TotalCharges, TotalCharges == "na", NA)) %>%
  mutate(TotalCharges = replace(TotalCharges, TotalCharges == "N/A", NA))

is.na(df$TotalCharges) # now the count is 3
df$TotalCharges

str(df)
#convert Totalcharges to Numeric
df$TotalCharges <- as.numeric(df$TotalCharges)
```

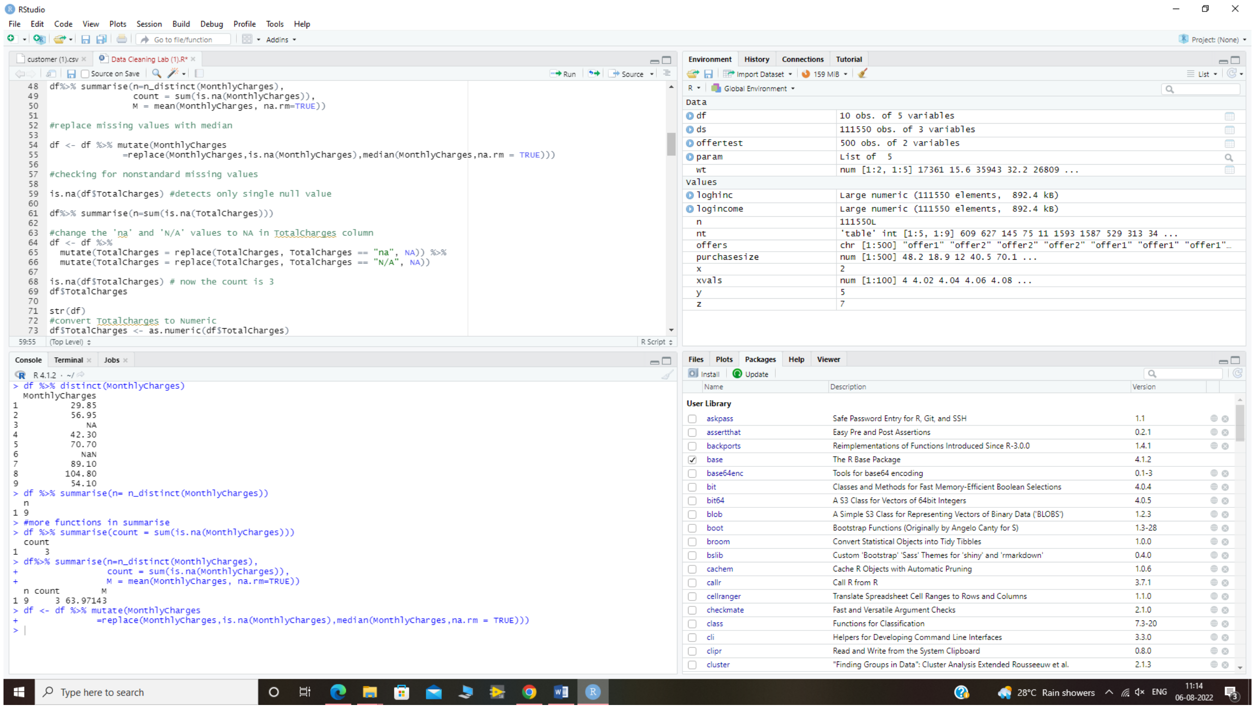
```
> df %>% distinct(MonthlyCharges)
  MonthlyCharges
1          29.85
2          56.95
3             NA
4          42.30
5          70.70
6            NaN
7          89.10
8         104.80
9          54.10
> df %>% summarise(n= n_distinct(MonthlyCharges))
  n
1 9
> #more functions in summarise
> df %>% summarise(count = sum(is.na(MonthlyCharges)))
  count
1     3
> df%>% summarise(n=n_distinct(MonthlyCharges),
+                 count = sum(is.na(MonthlyCharges)),
+                 M = mean(MonthlyCharges, na.rm=TRUE))
  n count        M
1 9     3 63.97143
> df <- df %>% mutate(MonthlyCharges
+               =replace(MonthlyCharges,is.na(MonthlyCharges),median(MonthlyCharges,na.rm = TRUE)))
> |
```

Data	
df	10 obs. of 5 variables
ds	111550 obs. of 3 variables
offertest	500 obs. of 2 variables
param	List of 5
wt	num [1:2, 1:5] 17361 15.6 35943 32.2 26809 ...
Values	
loghinc	Large numeric (111550 elements, 892.4 kB)
logincome	Large numeric (111550 elements, 892.4 kB)
n	111550L
nt	'table' int [1:5, 1:9] 609 627 145 75 11 1593 1587 529 313 34 ...
offers	chr [1:500] "offer1" "offer2" "offer2" "offer2" "offer1" "offer1" "offer1"…
purchasesize	num [1:500] 48.2 18.9 12 40.5 70.1 ...
x	2
xvals	num [1:100] 4 4.02 4.04 4.06 4.08 ...
y	5
z	7

	Name	Description	Version
User Library			
□	askpass	Safe Password Entry for R, Git, and SSH	1.1
□	assertthat	Easy Pre and Post Assertions	0.2.1
□	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.4.1
☑	base	The R Base Package	4.1.2
□	base64enc	Tools for base64 encoding	0.1-3
□	bit	Classes and Methods for Fast Memory-Efficient Boolean Selections	4.0.4
□	bit64	A S3 Class for Vectors of 64bit Integers	4.0.5
□	blob	A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS')	1.2.3
□	boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-28
□	broom	Convert Statistical Objects into Tidy Tibbles	1.0.0
□	bslib	Custom 'Bootstrap' 'Sass' Themes for 'shiny' and 'rmarkdown'	0.4.0
□	cachem	Cache R Objects with Automatic Pruning	1.0.6
□	callr	Call R from R	3.7.1
□	cellranger	Translate Spreadsheet Cell Ranges to Rows and Columns	1.1.0
□	checkmate	Fast and Versatile Argument Checks	2.1.0
□	class	Functions for Classification	7.3-20
□	cli	Helpers for Developing Command Line Interfaces	3.3.0
□	clipr	Read and Write from the System Clipboard	0.8.0
□	cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.	2.1.3

16. change the 'na' and 'N/A' values to NA in TotalCharges column. Then count and display null values in totalcharges column
17. Display all values in totalcharges column

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



```
df%>% summarise(n=n_distinct(MonthlyCharges),
                count = sum(is.na(MonthlyCharges)),
                M = mean(MonthlyCharges, na.rm=TRUE))

#replace missing values with median

df <- df %>% mutate(MonthlyCharges
              =replace(MonthlyCharges,is.na(MonthlyCharges),median(MonthlyCharges,na.rm = TRUE)))

#checking for nonstandard missing values

is.na(df$TotalCharges) #detects only single null value

df%>% summarise(n=sum(is.na(TotalCharges)))

#change the 'na' and 'N/A' values to NA in TotalCharges column
df <- df %>%
  mutate(TotalCharges = replace(TotalCharges, TotalCharges == "na", NA)) %>%
  mutate(TotalCharges = replace(TotalCharges, TotalCharges == "N/A", NA))

is.na(df$TotalCharges) # now the count is 3
df$TotalCharges

str(df)
#convert Totalcharges to Numeric
df$TotalCharges <- as.numeric(df$TotalCharges)
```

```
> df %>% distinct(MonthlyCharges)
  MonthlyCharges
1          29.85
2          56.95
3             NA
4          42.30
5          70.70
6            NaN
7          89.10
8         104.80
9          54.10
> df %>% summarise(n= n_distinct(MonthlyCharges))
  n
1 9
> #more functions in summarise
> df %>% summarise(count = sum(is.na(MonthlyCharges)))
  count
1     3
> df%>% summarise(n=n_distinct(MonthlyCharges),
+                 count = sum(is.na(MonthlyCharges)),
+                 M = mean(MonthlyCharges, na.rm=TRUE))
  n count        M
1 9     3 63.97143
> df <- df %>% mutate(MonthlyCharges
+               =replace(MonthlyCharges,is.na(MonthlyCharges),median(MonthlyCharges,na.rm = TRUE)))
> |
```

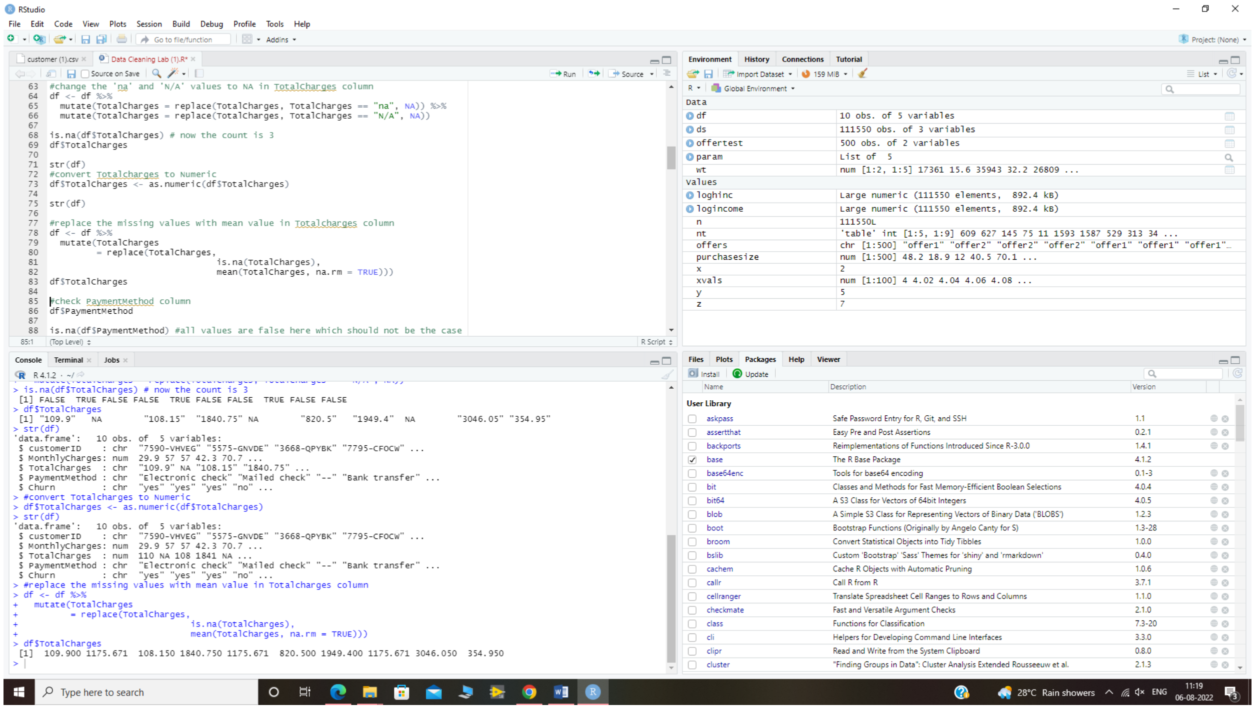
```
Data
df            10 obs. of 5 variables
ds            111550 obs. of 3 variables
offertest     500 obs. of 2 variables
param         List of  5
  wt          num [1:2, 1:5] 17361 15.6 35943 32.2 26809 ...
Values
loghinc       Large numeric (111550 elements,  892.4 kB)
logincome     Large numeric (111550 elements,  892.4 kB)
n             111550L
nt            'table' int [1:5, 1:9] 609 627 145 75 11 1593 1587 529 313 34 ...
offers        chr [1:500] "offer1" "offer2" "offer2" "offer2" "offer1" "offer1" "offer1"…
purchasesize  num [1:500] 48.2 18.9 12 40.5 70.1 ...
x             2
xvals         num [1:100] 4 4.02 4.04 4.06 4.08 ...
y             5
z             7
```

	Name	Description	Version
	User Library		
□	askpass	Safe Password Entry for R, Git, and SSH	1.1
□	assertthat	Easy Pre and Post Assertions	0.2.1
□	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.4.1
☑	base	The R Base Package	4.1.2
□	base64enc	Tools for base64 encoding	0.1-3
□	bit	Classes and Methods for Fast Memory-Efficient Boolean Selections	4.0.4
□	bit64	A S3 Class for Vectors of 64bit Integers	4.0.5
□	blob	A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS')	1.2.3
□	boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-28
□	broom	Convert Statistical Objects into Tidy Tibbles	1.0.0
□	bslib	Custom 'Bootstrap' 'Sass' Themes for 'shiny' and 'rmarkdown'	0.4.0
□	cachem	Cache R Objects with Automatic Pruning	1.0.6
□	callr	Call R from R	3.7.1
□	cellranger	Translate Spreadsheet Cell Ranges to Rows and Columns	1.1.0
□	checkmate	Fast and Versatile Argument Checks	2.1.0
□	class	Functions for Classification	7.3-20
□	cli	Helpers for Developing Command Line Interfaces	3.3.0
□	clipr	Read and Write from the System Clipboard	0.8.0
□	cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.	2.1.3

18. convert Totalcharges to Numeric using as.numeric command
19. Describe structure of dataframe
20. replace the missing values with mean value in Totalcharges column and display totalcharges column. Ignore null values while calculating mean value.

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



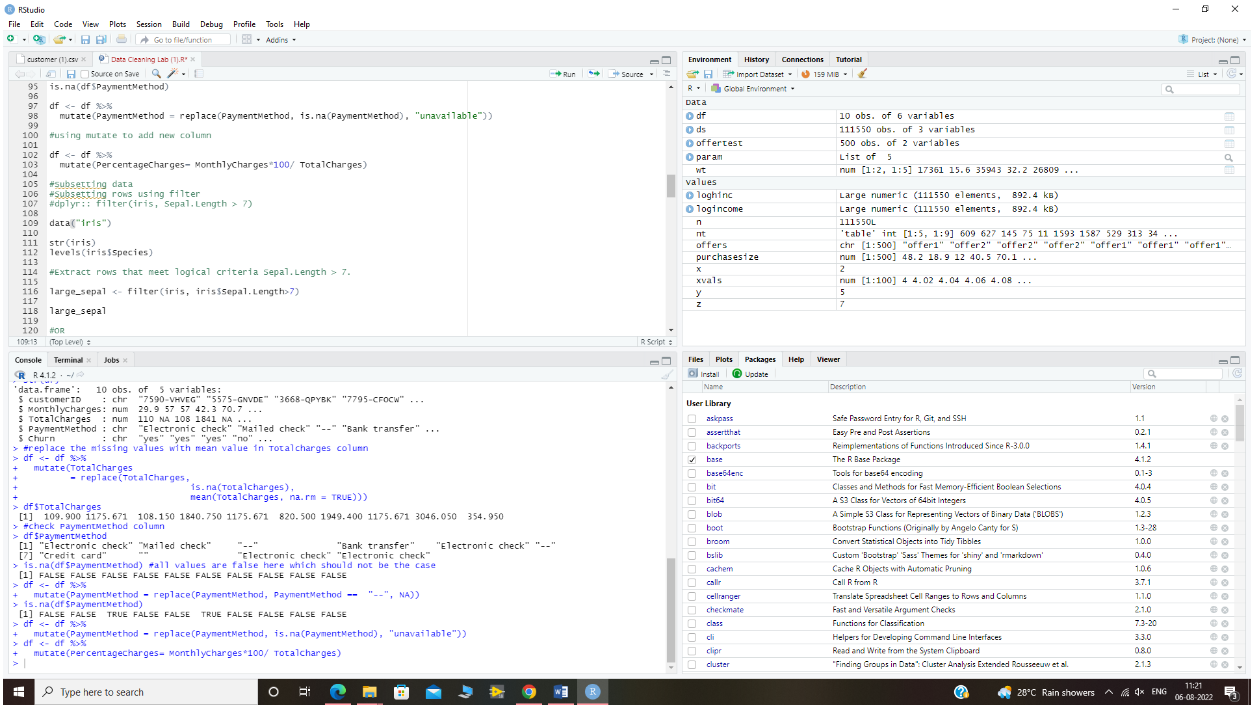
```
#change the 'na' and 'N/A' values to NA in TotalCharges column
df <- df %>%
  mutate(TotalCharges = replace(TotalCharges, TotalCharges == "na", NA)) %>%
  mutate(TotalCharges = replace(TotalCharges, TotalCharges == "N/A", NA))

is.na(df$TotalCharges) # now the count is 3
df$TotalCharges

str(df)
#convert Totalcharges to Numeric
df$TotalCharges <- as.numeric(df$TotalCharges)

str(df)

#replace the missing values with mean value in Totalcharges column
df <- df %>%
  mutate(TotalCharges
         = replace(TotalCharges,
                              is.na(TotalCharges),
                              mean(TotalCharges, na.rm = TRUE)))
df$TotalCharges

#check PaymentMethod column
df$PaymentMethod

is.na(df$PaymentMethod) #all values are false here which should not be the case
```

```
> is.na(df$TotalCharges) # now the count is 3
 [1] FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE
> df$TotalCharges
 [1] "109.9"   NA        "108.15"  "1840.75" NA        "820.5"   "1949.4"  NA        "3046.05" "354.95"
> str(df)
'data.frame':	10 obs. of  5 variables:
 $ customerID    : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
 $ MonthlyCharges: num  29.9 57 57 42.3 70.7 ...
 $ TotalCharges  : chr  "109.9" NA "108.15" "1840.75" ...
 $ PaymentMethod : chr  "Electronic check" "Mailed check" "--" "Bank transfer" ...
 $ Churn         : chr  "yes" "yes" "yes" "no" ...
> #convert Totalcharges to Numeric
> df$TotalCharges <- as.numeric(df$TotalCharges)
> str(df)
'data.frame':	10 obs. of  5 variables:
 $ customerID    : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
 $ MonthlyCharges: num  29.9 57 57 42.3 70.7 ...
 $ TotalCharges  : num  110 NA 108 1841 NA ...
 $ PaymentMethod : chr  "Electronic check" "Mailed check" "--" "Bank transfer" ...
 $ Churn         : chr  "yes" "yes" "yes" "no" ...
> #replace the missing values with mean value in Totalcharges column
> df <- df %>%
+   mutate(TotalCharges
+          = replace(TotalCharges,
+                               is.na(TotalCharges),
+                               mean(TotalCharges, na.rm = TRUE)))
> df$TotalCharges
 [1]  109.900 1175.671  108.150 1840.750 1175.671  820.500 1949.400 1175.671 3046.050  354.950
> 
```

Data	
df	10 obs. of 5 variables
ds	111550 obs. of 3 variables
offertest	500 obs. of 2 variables
param	List of 5
wt	num [1:2, 1:5] 17361 15.6 35943 32.2 26809 ...
Values	
loghinc	Large numeric (111550 elements, 892.4 kB)
logincome	Large numeric (111550 elements, 892.4 kB)
n	111550L
nt	'table' int [1:5, 1:9] 609 627 145 75 11 1593 1587 529 313 34 ...
offers	chr [1:500] "offer1" "offer2" "offer2" "offer2" "offer1" "offer1" "offer1"…
purchasesize	num [1:500] 48.2 18.9 12 40.5 70.1 ...
x	2
xvals	num [1:100] 4 4.02 4.04 4.06 4.08 ...
y	5
z	7

	Name	Description	Version
	User Library		
□	askpass	Safe Password Entry for R, Git, and SSH	1.1
□	assertthat	Easy Pre and Post Assertions	0.2.1
□	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.4.1
☑	base	The R Base Package	4.1.2
□	base64enc	Tools for base64 encoding	0.1-3
□	bit	Classes and Methods for Fast Memory-Efficient Boolean Selections	4.0.4
□	bit64	A S3 Class for Vectors of 64bit Integers	4.0.5
□	blob	A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS')	1.2.3
□	boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-28
□	broom	Convert Statistical Objects into Tidy Tibbles	1.0.0
□	bslib	Custom 'Bootstrap' 'Sass' Themes for 'shiny' and 'rmarkdown'	0.4.0
□	cachem	Cache R Objects with Automatic Pruning	1.0.6
□	callr	Call R from R	3.7.1
□	cellranger	Translate Spreadsheet Cell Ranges to Rows and Columns	1.1.0
□	checkmate	Fast and Versatile Argument Checks	2.1.0
□	class	Functions for Classification	7.3-20
□	cli	Helpers for Developing Command Line Interfaces	3.3.0
□	clipr	Read and Write from the System Clipboard	0.8.0
□	cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.	2.1.3

21. check ‘paymentmethod’ column for null values and comment on the result.
22. Replace ‘- -‘ by ‘NA’ and null value by ‘unavailable’
23. Add new column ‘percentagecharges’ using ‘mutate’ command



```r
is.na(df$PaymentMethod)

df <- df %>%
  mutate(PaymentMethod = replace(PaymentMethod, is.na(PaymentMethod), "unavailable"))

#using mutate to add new column

df <- df %>%
  mutate(PercentageCharges= MonthlyCharges*100/ TotalCharges)

#Subsetting data
#Subsetting rows using filter
#dplyr:: filter(iris, Sepal.Length > 7)

data("iris")

str(iris)
levels(iris$Species)

#Extract rows that meet logical criteria Sepal.Length > 7.

large_sepal <- filter(iris, iris$Sepal.Length>7)

large_sepal

#OR
```

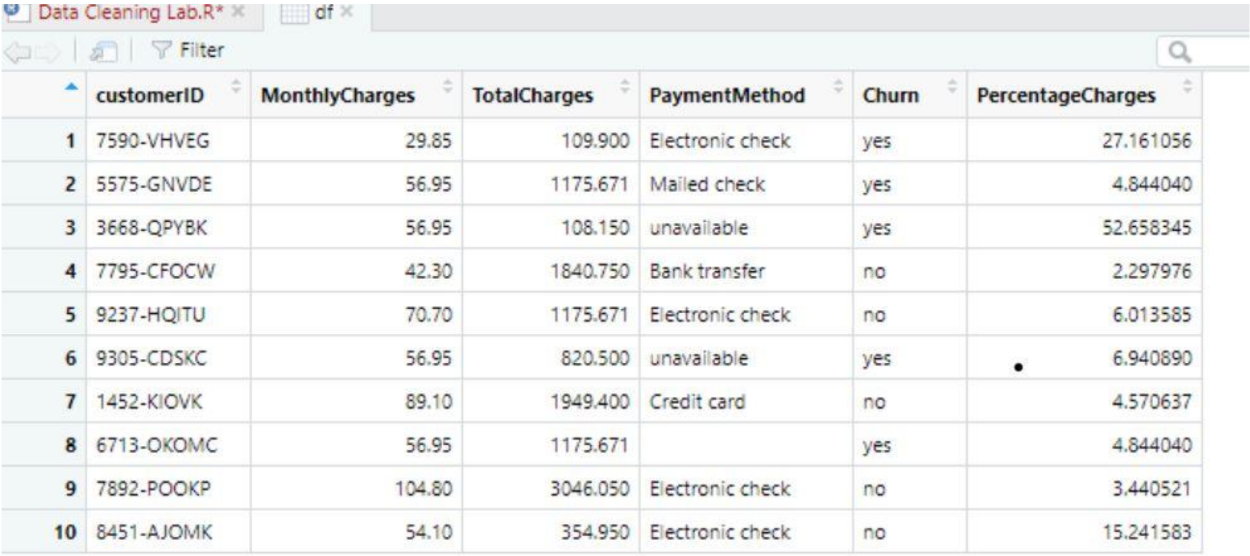
```
'data.frame':   10 obs. of  5 variables:
 $ customerID    : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
 $ MonthlyCharges: num  29.9 57 57 42.3 70.7 ...
 $ TotalCharges  : num  110 NA 108 1841 NA ...
 $ PaymentMethod : chr  "Electronic check" "Mailed check" "--" "Bank transfer" ...
 $ Churn         : chr  "yes" "yes" "yes" "no" ...
> #replace the missing values with mean value in Totalcharges column
> df <- df %>%
+   mutate(TotalCharges
+          = replace(TotalCharges,
+                    is.na(TotalCharges),
+                    mean(TotalCharges, na.rm = TRUE)))
> df$TotalCharges
 [1]  109.900 1175.671  108.150 1840.750 1175.671  820.500 1949.400 1175.671 3046.050  354.950
> #check PaymentMethod column
> df$PaymentMethod
 [1] "Electronic check" "Mailed check"     "--"                 "Bank transfer"    "Electronic check" "--"
 [7] "Credit card"      ""                 "Electronic check" "Electronic check"
> is.na(df$PaymentMethod) #all values are false here which should not be the case
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
> df <- df %>%
+   mutate(PaymentMethod = replace(PaymentMethod, PaymentMethod ==  "--", NA))
> is.na(df$PaymentMethod)
 [1] FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
> df <- df %>%
+   mutate(PaymentMethod = replace(PaymentMethod, is.na(PaymentMethod), "unavailable"))
> df <- df %>%
+   mutate(PercentageCharges= MonthlyCharges*100/ TotalCharges)
> |
```

```
Data
df            10 obs. of 6 variables
ds            111550 obs. of 3 variables
offertest     500 obs. of 2 variables
param         List of  5
  wt          num [1:2, 1:5] 17361 15.6 35943 32.2 26809 ...
Values
loghinc       Large numeric (111550 elements,  892.4 kB)
logincome     Large numeric (111550 elements,  892.4 kB)
  n           111550L
  nt          'table' int [1:5, 1:9] 609 627 145 75 11 1593 1587 529 313 34 ...
  offers      chr [1:500] "offer1" "offer2" "offer2" "offer2" "offer1" "offer1" "offer1"…
  purchasesize num [1:500] 48.2 18.9 12 40.5 70.1 ...
  x           2
  xvals       num [1:100] 4 4.02 4.04 4.06 4.08 ...
  y           5
  z           7
```

User Library

	Name	Description	Version
☐	askpass	Safe Password Entry for R, Git, and SSH	1.1
☐	assertthat	Easy Pre and Post Assertions	0.2.1
☐	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.4.1
☑	base	The R Base Package	4.1.2
☐	base64enc	Tools for base64 encoding	0.1-3
☐	bit	Classes and Methods for Fast Memory-Efficient Boolean Selections	4.0.4
☐	bit64	A S3 Class for Vectors of 64bit Integers	4.0.5
☐	blob	A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS')	1.2.3
☐	boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-28
☐	broom	Convert Statistical Objects into Tidy Tibbles	1.0.0
☐	bslib	Custom 'Bootstrap' 'Sass' Themes for 'shiny' and 'rmarkdown'	0.4.0
☐	cachem	Cache R Objects with Automatic Pruning	1.0.6
☐	callr	Call R from R	3.7.1
☐	cellranger	Translate Spreadsheet Cell Ranges to Rows and Columns	1.1.0
☐	checkmate	Fast and Versatile Argument Checks	2.1.0
☐	class	Functions for Classification	7.3-20
☐	cli	Helpers for Developing Command Line Interfaces	3.3.0
☐	clipr	Read and Write from the System Clipboard	0.8.0
☐	cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.	2.1.3

24. Display dataframe


Data Cleaning Lab.R*
df
Filter

	customerID	MonthlyCharges	TotalCharges	PaymentMethod	Churn	PercentageCharges
1	7590-VHVEG	29.85	109.900	Electronic check	yes	27.161056
2	5575-GNVDE	56.95	1175.671	Mailed check	yes	4.844040
3	3668-QPYBK	56.95	108.150	unavailable	yes	52.658345
4	7795-CFOCW	42.30	1840.750	Bank transfer	no	2.297976
5	9237-HQITU	70.70	1175.671	Electronic check	no	6.013585
6	9305-CDSKC	56.95	820.500	unavailable	yes	6.940890
7	1452-KIOVK	89.10	1949.400	Credit card	no	4.570637
8	6713-OKOMC	56.95	1175.671		yes	4.844040
9	7892-POOKP	104.80	3046.050	Electronic check	no	3.440521
10	8451-AJOMK	54.10	354.950	Electronic check	no	15.241583

B.2 Conclusion

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

After completing this experiment, I am able to Apply appropriate data cleaning techniques and improve data quality and to make it complete and consistent.