20BDS0146

VENNELA G

Programming For Data Science

Lab Assessment No: 2

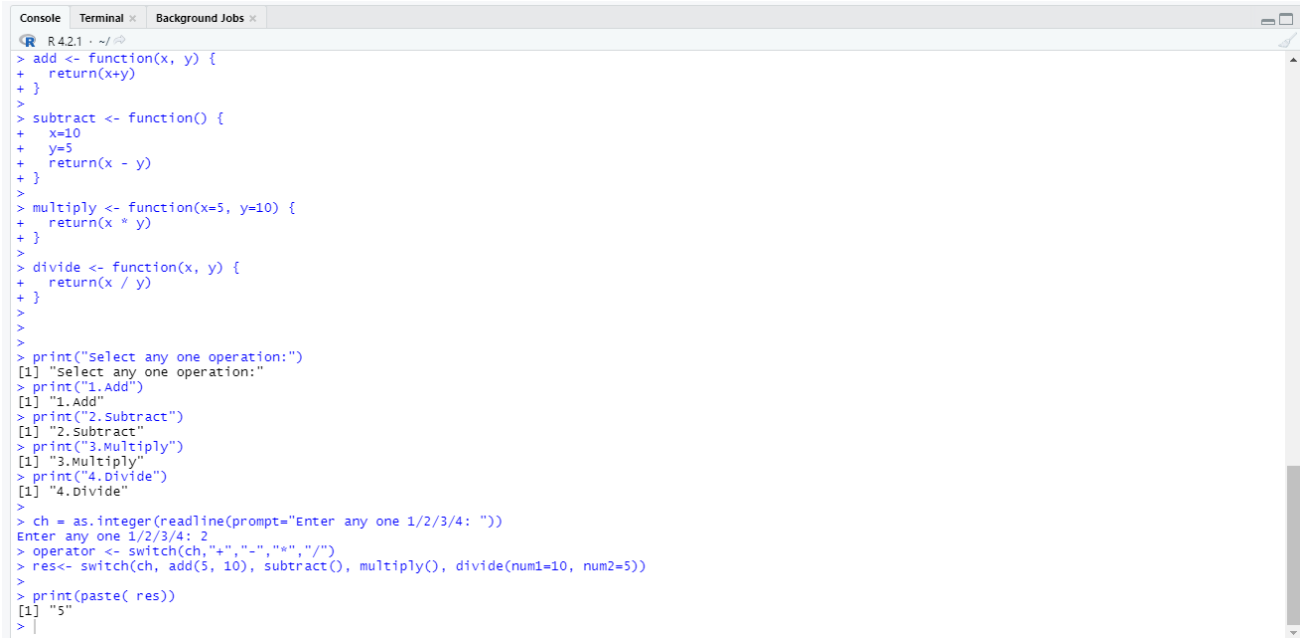
Data Manipulation Using R

1.Create user defined functions to perform various arithmetic operations and call each functions using menu driven format (try functions with and without parameters, functions with default argument)

CODE:

```
add <- function(x, y) {
  return(x+y)
}

subtract <- function() {
  x=10
  y=5
  return(x - y)
}

multiply <- function(x=5, y=10) {
  return(x * y)
}
```

```
divide <- function(x, y) {
  return(x / y)
}



print("Select any one operation:")
print("1.Add")
print("2.Subtract")
print("3.Multiply")
print("4.Divide")

ch = as.integer(readline(prompt="Enter any one 1/2/3/4: "))

operator <- switch(ch,"+","-","*","/")
res<- switch(ch, add(5, 10), subtract(), multiply(), divide(num1=10, num2=5))

print(paste( res))
```



2. Familiarize basic statistical operation on a random vector of 100 elements

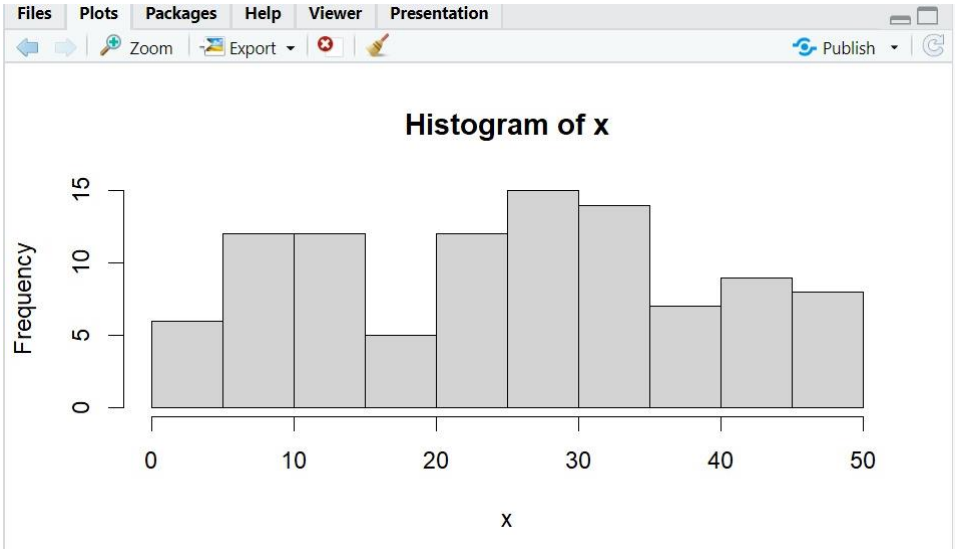
a. Mean
b. IQR
c. Table
b. Median
f. Standard deviation
c. Mode
g. Summary
d. Range
h. Histogram

CODE:

```
getmode <- function(x) {
uniqv <- unique(x)
  uniqv[which.max(tabulate(match(x, uniqv)))]
 }
 set.seed(123)
x <- sample(1:50, size = 100, replace=TRUE)
 print(paste("x=",x))
 result.mean <- mean(x)
print(paste("Mean=",result.mean))
 median.result <- median(x)
 print(paste("Median=",median.result))
result <- getmode(x)
 print(paste("Mode=",result))
 print(paste("Range=",diff(range(x))))
 print(paste("IQR=",IQR(x)))
 s<-sd(x)
print(paste("Standard deviation=",s))
print(paste("Summary=",summary(x)))
print(paste("Table=",table(x)))
hist(x)
```

OUTPUT:

```
> getmode <- function(x) {
+   uniqv <- unique(x)
+   uniqv[which.max(tabulate(match(x, uniqv)))]
+ }
> set.seed(123)
> x <- sample(1:50, size = 100, replace=TRUE)
> print(paste("x=",x))
  [1] "x= 31" "x= 15" "x= 14" "x= 3"  "x= 42" "x= 50" "x= 43" "x= 37" "x= 14" "x= 25" "x= 26" "x= 27" "x= 5"  "x= 27" "x= 28" "x= 9"  "x= 29" "x= 35"
 [19] "x= 8"  "x= 26" "x= 7"  "x= 42" "x= 9"  "x= 19" "x= 36" "x= 14" "x= 17" "x= 43" "x= 39" "x= 12" "x= 15" "x= 32" "x= 42" "x= 45" "x= 7"  "x= 9"
 [37] "x= 41" "x= 10" "x= 23" "x= 27" "x= 7"  "x= 27" "x= 32" "x= 38" "x= 25" "x= 34" "x= 29" "x= 5"  "x= 8"  "x= 12" "x= 13" "x= 18" "x= 33" "x= 27"
 [55] "x= 25" "x= 38" "x= 21" "x= 15" "x= 41" "x= 47" "x= 26" "x= 31" "x= 16" "x= 30" "x= 6"  "x= 43" "x= 8"  "x= 22" "x= 22" "x= 39" "x= 31" "x= 48"
 [73] "x= 17" "x= 50" "x= 49" "x= 34" "x= 4"  "x= 13" "x= 5"  "x= 25" "x= 22" "x= 25" "x= 32" "x= 46" "x= 25" "x= 23" "x= 35" "x= 40" "x= 48" "x= 30"
 [91] "x= 12" "x= 31" "x= 46" "x= 30" "x= 35" "x= 14" "x= 29" "x= 32" "x= 7"  "x= 3"
> result.mean <- mean(x)
> print(paste("Mean=",result.mean))
[1] "Mean= 25.62"
> median.result <- median(x)
> print(paste("Median=",median.result))
[1] "Median= 26.5"
>
> result <- getmode(x)
> print(paste("Mode=",result))
[1] "Mode= 25"
> print(paste("Range=",diff(range(x))))
[1] "Range= 47"
> print(paste("IQR=",IQR(x)))
[1] "IQR= 21"
> s<-sd(x)
> print(paste("Standard deviation=",s))
[1] "Standard deviation= 13.2014538532081"
> print(paste("Summary=",summary(x)))
[1] "Summary= 3"      "Summary= 14"     "Summary= 26.5"   "Summary= 25.62" "Summary= 35"     "Summary= 50"
>
> print(paste("Table=",table(x)))
 [1] "Table= 2" "Table= 1" "Table= 3" "Table= 1" "Table= 4" "Table= 3" "Table= 3" "Table= 1" "Table= 3" "Table= 2" "Table= 4" "Table= 3" "Table= 1"
[14] "Table= 2" "Table= 1" "Table= 1" "Table= 1" "Table= 3" "Table= 2" "Table= 6" "Table= 3" "Table= 5" "Table= 1" "Table= 3" "Table= 3" "Table= 4"
[27] "Table= 4" "Table= 1" "Table= 2" "Table= 3" "Table= 1" "Table= 1" "Table= 2" "Table= 2" "Table= 1" "Table= 2" "Table= 3" "Table= 3" "Table= 1"
[40] "Table= 2" "Table= 1" "Table= 2" "Table= 1" "Table= 2"
```



3. Perform given operations on a data frame
    a. Create a data frame
    b. Access a component ([, [[, $)
    c. Structure of data frame
    d. Add new column
    e. Add new row
    f. Delete column
    g. Delete specific row
    h. Order data frame (order, arrange)

CODE:

```
df=data.frame(col1=c(40,NA,NA,45),col2=c(47,NA,35,42))
print(df)
print(select.list(df$col1))
print(select.list(df[[2,1]]))
print(select.list(df[2]))
print(str(df))
df$col3=c(20,50,90,NA)
print(df)
df[nrow(df)+1]<-c(10,20)
```

```
print(df)

df<- subset(df,select=-col2)


print(df)

df4=df[-c(2),]

print(df4)

arrange(df,desc(col2))

order(df$col1)
```

OUTPUT:

```
4   42
> print(str(df))
'data.frame':   4 obs. of  2 variables:
 $ col1: num  40 NA NA 45
 $ col2: num  47 NA 35 42
NULL
```

```
> df$col3=c(20,50,90,NA)
> print(df)
  col1 col2 col3
1   40   47   20
2   NA   NA   50
3   NA   35   90
4   45   42   NA
> v
```

```
> df=data.frame(col1=c(40,NA,NA,45),col2=c(47,NA,35,42))
> df[nrow(df)+1,]<-c(10,20)
>
>
> print(df)
  col1 col2
1   40   47
2   NA   NA
3   NA   35
4   45   42
5   10   20
>
```

```
> df=data.frame(col1=c(40,NA,NA,45),col2=c(47,NA,35,42))
>
>
> df<- subset(df,select=-col2)
> print(df)
  col1
1   40
2   NA
3   NA
4   45
> df=data.frame(col1=c(40,NA,NA,45),col2=c(47,NA,35,42))
>
>
>
> df4=df[-c(2),]
> print(df4)
  col1 col2
1   40   47
3   NA   35
4   45   42
> arrange(df,desc(col2))
  col1 col2
1   40   47
2   45   42
3   NA   35
4   NA   NA
> order(df$col1)
[1] 1 4 2 3
>
```

4. Read Air quality dataset and handle the missing data using following technique
    a. Drop Row
    b. Drop Column
    c. Imputation (Replace with unknown, mean or Group mean)

CODE:

```
data=airquality

print(class(data))

na.exclude(head(data))

new_data=data[,colSums(is.na(data))==0]

print(head(new_data)

V=is.na(data)

data[V]= 0

print(head(data))

data$Ozone[is.na(data$ Ozone)]=mean(data$ Ozone,na.rm=TRUE)

print(head(data))

g=group_by(data, Ozone)

g$ Ozone [is.na(g$ Ozone)]=mean(g$ Ozone,na.rm=TRUE)

print(head(data))
```

OUTPUT:

```
> data=airquality
> print(class(data))
[1] "data.frame"
> na.exclude(head(data))
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
> new_data=data[,colSums(is.na(data))==0]
> print(head(new_data))
  Wind Temp Month Day
1  7.4   67     5   1
2  8.0   72     5   2
3 12.6   74     5   3
4 11.5   62     5   4
5 14.3   56     5   5
6 14.9   66     5   6
> V=is.na(data)
> data[V]= 0
> print(head(data))
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5     0       0 14.3   56     5   5
6    28       0 14.9   66     5   6
>
```

```
>
> data$Ozone[is.na(data$Ozone)]=mean(data$Ozone,na.rm=TRUE)
> print(head(data))
     Ozone Solar.R Wind Temp Month Day
1 41.00000     190  7.4   67     5   1
2 36.00000     118  8.0   72     5   2
3 12.00000     149 12.6   74     5   3
4 18.00000     313 11.5   62     5   4
5 42.12931      NA 14.3   56     5   5
6 28.00000      NA 14.9   66     5   6
>
```

```
>
>
> g=group_by(data,Ozone)
> g$Ozone[is.na(g$Ozone)]=mean(g$Ozone,na.rm=TRUE)
> print(head(data))
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
>
```