ACADGILD

SESSION 7: Basic Statistics

Assignment 2

PROBLEM STATEMENT

1. Write a program to create **barplots** for all the categorical columns in **mtcars.**

2. Create a **scatterplot** matrix by gear types in **mtcars** dataset.

3. Write a program to create a **plot density** by class variable.

SOLUTION

1. Write a program to create barplots for all the categorical columns in mtcars.

The R-script for the given problem is as follows:

```
library(readr)
library(ggplot2)
mtcars <- read_csv("F:/ACADGILD - Online Course/1. DATA SETS/mtcars.csv")
View(mtcars)
mtcars
str(mtcars)
# change the categorical variables to factor
library(dplyr)
mtcars1 <- mutate(mtcars,
                  cyl = as.factor(cyl),
                  disp = as.factor(disp),
                  vs = as.factor(vs),
                  am = as.factor(am),
                  gear = as.factor(gear),
                  carb = as.factor(carb))
str(mtcars1)

is.fact <- sapply(mtcars1, is.factor)  # checking the categorical variables
mtcars2 <- mtcars1[,is.fact]   # creating dataframe of only factor class of variables
```

```
str(mtcars2)                # check structure
par(mfrow= c(2,3))          # Set plot area
lapply(lapply(mtcars2[,1:6], table), barplot)  # barplots for categorical variables

table1 <- table(mtcars$cyl, mtcars$gear, dnn=c("Cylinders", "Gears"))
# Creates a contingency table
addmargins(table1)          #Displays the table (Not necessary
barplot(table1, ylab="Frequency", xlab="Gears", main="Side-By-Side Bar Chart",
col=c("turquoise4", "turquoise2", "turquoise" ), beside=TRUE, width=.3)
legend("right", title="Cylinders", legend= sort(unique(mtcars$cyl)), fill
=c("turquoise4", "turquoise2", "turquoise" ), box.lty=0)
legend("right", title="Cylinders", legend= sort(unique(mtcars$cyl)), fill
=c("turquoise4", "turquoise2", "turquoise" ), box.lty=0)

 # Histogram on a Categorical variable
g <- ggplot(mpg, aes(manufacturer))
g + geom_bar(aes(fill=class), width = 0.5) + theme(axis.text.x =
element_text(angle=65, vjust=0.6)) + labs(title="Histogram on Categorical
Variable", subtitle="Manufacturer across Vehicle Classes")
```

The output of the R-Script (from Console window) is given as follows:

```
> library(readr)
> library(ggplot2)
> mtcars <- read_csv("F:/ACADGILD - Online Course/1. DATA
SETS/mtcars.csv")
Parsed with column specification:
cols(
  X1 = col_character(),
  mpg = col_double(),
  cyl = col_double(),
  disp = col_double(),
  hp = col_double(),
  drat = col_double(),
```

```
  wt = col_double(),
  qsec = col_double(),
  vs = col_double(),
  am = col_double(),
  gear = col_double(),
  carb = col_double()
)
> mtcars
# A tibble: 32 x 12
   X1               mpg   cyl  disp    hp  drat    wt  qsec    vs
am  gear  carb
   <chr>          <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
<dbl> <dbl> <dbl>
 1 Mazda RX4       21       6  160    110  3.9   2.62  16.5     0
1     4     4
 2 Mazda RX4 Wag   21       6  160    110  3.9   2.88  17.0     0
1     4     4
 3 Datsun 710      22.8     4  108     93  3.85  2.32  18.6     1
1     4     1
 4 Hornet 4 Drive  21.4     6  258    110  3.08  3.22  19.4     1
0     3     1
 5 Hornet Sporta~  18.7     8  360    175  3.15  3.44  17.0     0
0     3     2
 6 Valiant         18.1     6  225    105  2.76  3.46  20.2     1
0     3     1
 7 Duster 360      14.3     8  360    245  3.21  3.57  15.8     0
0     3     4
 8 Merc 240D       24.4     4  147.    62  3.69  3.19  20       1
0     4     2
 9 Merc 230        22.8     4  141.    95  3.92  3.15  22.9     1
0     4     2
10 Merc 280        19.2     6  168.   123  3.92  3.44  18.3     1
0     4     4
# ... with 22 more rows
> str(mtcars)
Classes ‘spec_tbl_df’, ‘tbl_df’, ‘tbl’ and 'data.frame': 32 obs.
of  12 variables:
 $ X1  : chr  "Mazda RX4" "Mazda RX4 Wag" "Datsun 710" "Hornet 4
Drive" ...
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
```

```
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
 - attr(*, "spec")=
  .. cols(
  ..   X1 = col_character(),
  ..   mpg = col_double(),
  ..   cyl = col_double(),
  ..   disp = col_double(),
  ..   hp = col_double(),
  ..   drat = col_double(),
  ..   wt = col_double(),
  ..   qsec = col_double(),
  ..   vs = col_double(),
  ..   am = col_double(),
  ..   gear = col_double(),
  ..   carb = col_double()
  .. )
> # change the categorical variables to factor
> library(dplyr)
> mtcars1 <- mutate(mtcars,
+                   cyl = as.factor(cyl),
+                   disp = as.factor(disp),
+                   vs = as.factor(vs),
+                   am = as.factor(am),
+                   gear = as.factor(gear),
+                   carb = as.factor(carb))
> str(mtcars1)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':   32 obs. of  12
variables:
 $ X1  : chr  "Mazda RX4" "Mazda RX4 Wag" "Datsun 710" "Hornet 4
Drive" ...
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
```

```
 $ disp: Factor w/ 27 levels "71.1","75.7",..: 13 13 6 16 23 15
23 12 10 14 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
 $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
 $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
 $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2
4 ...
> is.fact <- sapply(mtcars1, is.factor)  # checking the
categorical variables
> mtcars2 <- mtcars1[,is.fact]   # creating dataframe of only
factor class of variables
> str(mtcars2)                   # check structure
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':   32 obs. of  6
variables:
 $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
 $ disp: Factor w/ 27 levels "71.1","75.7",..: 13 13 6 16 23 15
23 12 10 14 ...
 $ vs  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
 $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
 $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
 $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2
4 ...
> par(mfrow= c(2,3))             # Set plot area
> lapply(lapply(mtcars2[,1:6], table), barplot)  # barplots for
categorical variables
$cyl
     [,1]
[1,]  0.7
[2,]  1.9
[3,]  3.1

$disp
      [,1]
 [1,]  0.7
 [2,]  1.9
 [3,]  3.1
 [4,]  4.3
```

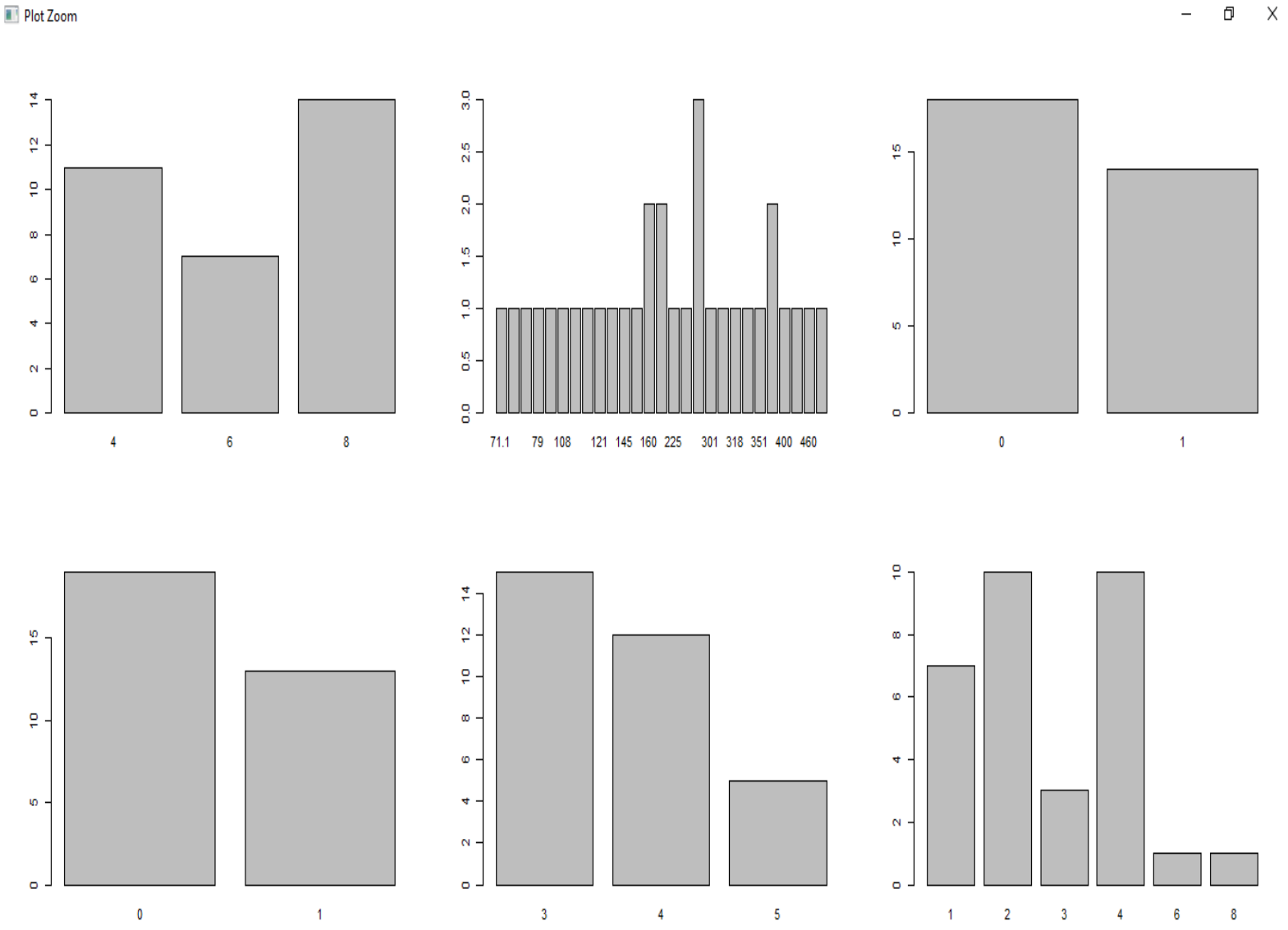
```
 [5,]  5.5
 [6,]  6.7
 [7,]  7.9
 [8,]  9.1
 [9,] 10.3
[10,] 11.5
[11,] 12.7
[12,] 13.9
[13,] 15.1
[14,] 16.3
[15,] 17.5
[16,] 18.7
[17,] 19.9
[18,] 21.1
[19,] 22.3
[20,] 23.5
[21,] 24.7
[22,] 25.9
[23,] 27.1
[24,] 28.3
[25,] 29.5
[26,] 30.7
[27,] 31.9

$vs
     [,1]
[1,]  0.7
[2,]  1.9

$am
     [,1]
[1,]  0.7
[2,]  1.9

$gear
     [,1]
[1,]  0.7
[2,]  1.9
[3,]  3.1
```

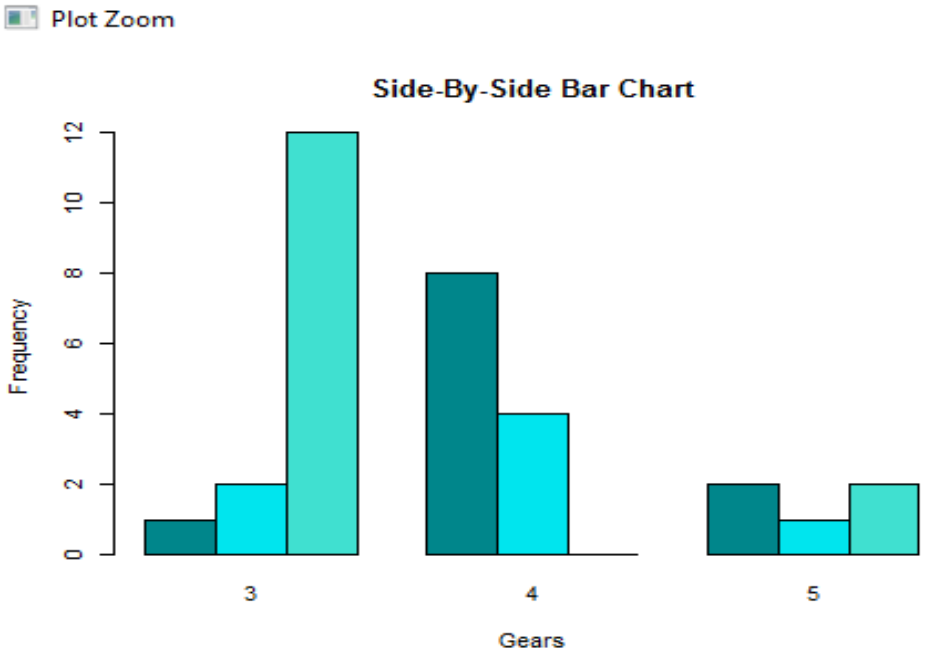
```
$carb
     [,1]
[1,]  0.7
[2,]  1.9
[3,]  3.1
[4,]  4.3
[5,]  5.5
[6,]  6.7
```



```
> table1 <- table(mtcars$cyl, mtcars$gear, dnn=c("Cylinders",
"Gears")) # Creates a contingency table
> addmargins(table1) #Displays the table (Not necessary
```

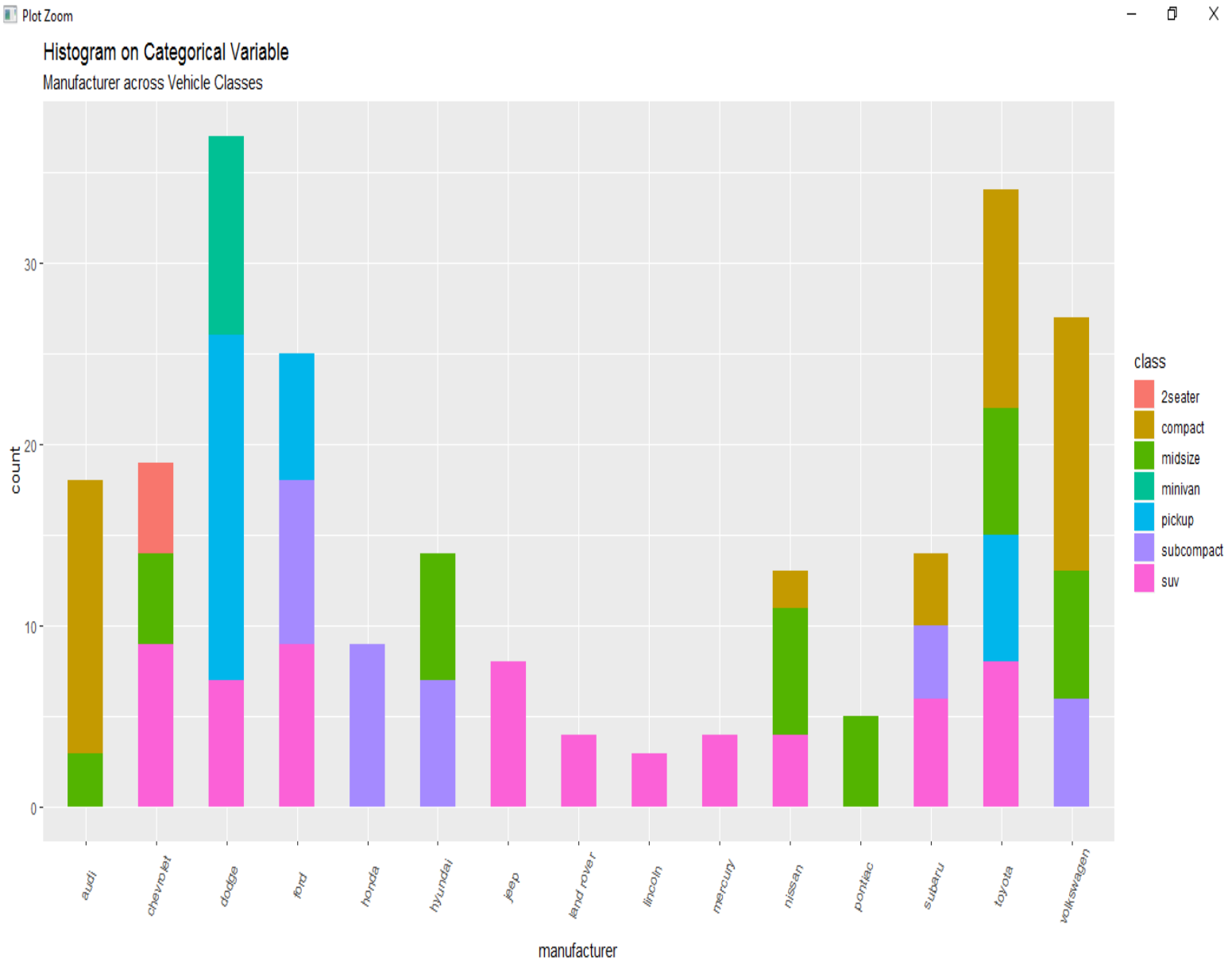
```
        Gears
Cylinders  3  4  5 Sum
      4    1  8  2  11
      6    2  4  1   7
      8   12  0  2  14
      Sum 15 12  5  32
```

```
> barplot(table1, ylab="Frequency", xlab="Gears", main="Side-By-Side
Bar Chart", col=c("turquoise4", "turquoise2", "turquoise" ),
beside=TRUE, width=.3)
```




```
> legend("right", title="Cylinders", legend=
sort(unique(mtcars$cyl)), fill =c("turquoise4", "turquoise2",
"turquoise" ), box.lty=0)
```

```
> # Histogram on a Categorical variable
> g <- ggplot(mpg, aes(manufacturer))
> g + geom_bar(aes(fill=class), width = 0.5) + theme(axis.text.x =
element_text(angle=65, vjust=0.6)) + labs(title="Histogram on
```




```
Categorical Variable", subtitle="Manufacturer across Vehicle
Classes")
```

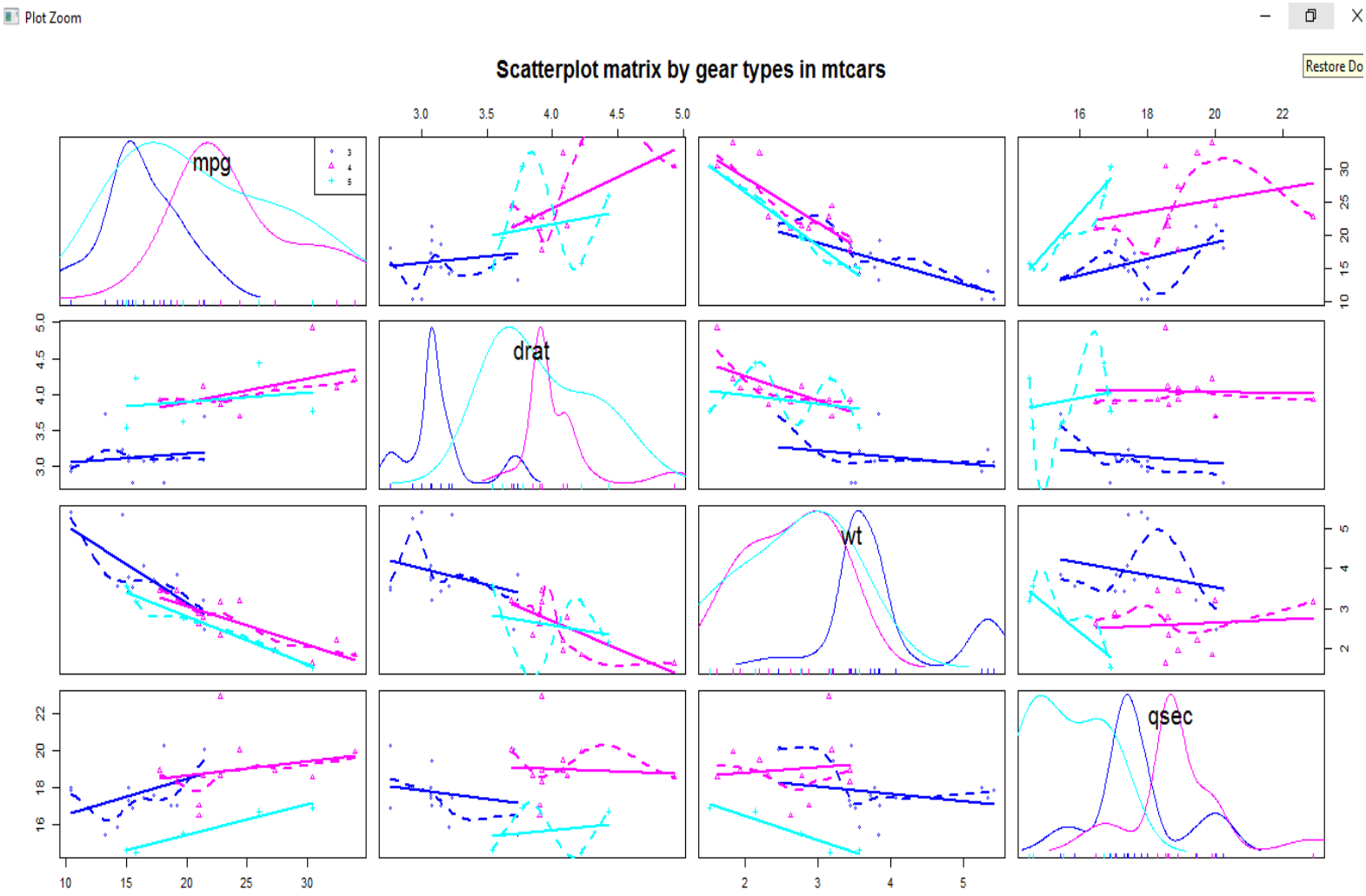
2. Create a scatterplot matrix by gear types in mtcars dataset.

The R-script for the given problem is as follows:

```
library(car)
str(mtcars)
scatterplotMatrix(~mpg+drat+wt+qsec|gear, data=mtcars,main="Scatterplot matrix
by gear types in mtcars")
```

The output of the R-Script (from Console window/Plot window) is given as follows:

```
> library(car)
> str(mtcars)
Classes ‘spec_tbl_df’, ‘tbl_df’, ‘tbl’ and 'data.frame': 32 obs.
of  12 variables:
 $ X1  : chr  "Mazda RX4" "Mazda RX4 Wag" "Datsun 710" "Hornet 4
Drive" ...
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
 - attr(*, "spec")=
  .. cols(
  ..   X1 = col_character(),
  ..   mpg = col_double(),
  ..   cyl = col_double(),
  ..   disp = col_double(),
  ..   hp = col_double(),
  ..   drat = col_double(),
  ..   wt = col_double(),
  ..   qsec = col_double(),
  ..   vs = col_double(),
  ..   am = col_double(),
  ..   gear = col_double(),
  ..   carb = col_double()
  .. )
> scatterplotMatrix(~mpg+drat+wt+qsec|gear,
data=mtcars,main="Scatterplot matrix by gear types in mtcars
```



3. Write a program to create a plot density by class variable

The R-script for the given problem is as follows:

```
par(mfrow = c(1,1))
x <- mtcars$mpg          # assign mpg to a object
h <- hist(x, breaks = 10, col = "pink",
          xlab = "MPG",
          main = "Density plot of mpg")        # plot histogram of the object
xfit <- seq(min(x), max(x), length = 40)       # create 40 points on x axis
yfit <- dnorm(xfit, mean = mean(x), sd= sd(x))    # normal plot of xfit
yfit <- yfit*diff(h$mids[1:2]*length(x))       # mids of the histogram with changing x
lines(xfit, yfit, col="Blue", lwd = 3)         # line plot for xfit and yfit
```

The output of the R-Script (from Console window/Plot window) is given as follows:

```
> par(mfrow = c(1,1))
> x <- mtcars$mpg          # assign mpg to a object
> h <- hist(x, breaks = 10, col = "pink",
+           xlab = "MPG",
+           main = "Density plot of mpg")        # plot histogram
of the object
> xfit <- seq(min(x), max(x), length = 40)       # create 40
points on x axis
> yfit <- dnorm(xfit, mean = mean(x), sd= sd(x)) # normal plot of
xfit
> yfit <- yfit*diff(h$mids[1:2]*length(x))       # mids of the
histogram with changing x
> lines(xfit, yfit, col="Blue", lwd = 3)         # line plot for
xfit and yfit
```