Intelligent Data Analytics

Abhishek Kumar Gupta

ggplot (R) Visualizations

Example 1

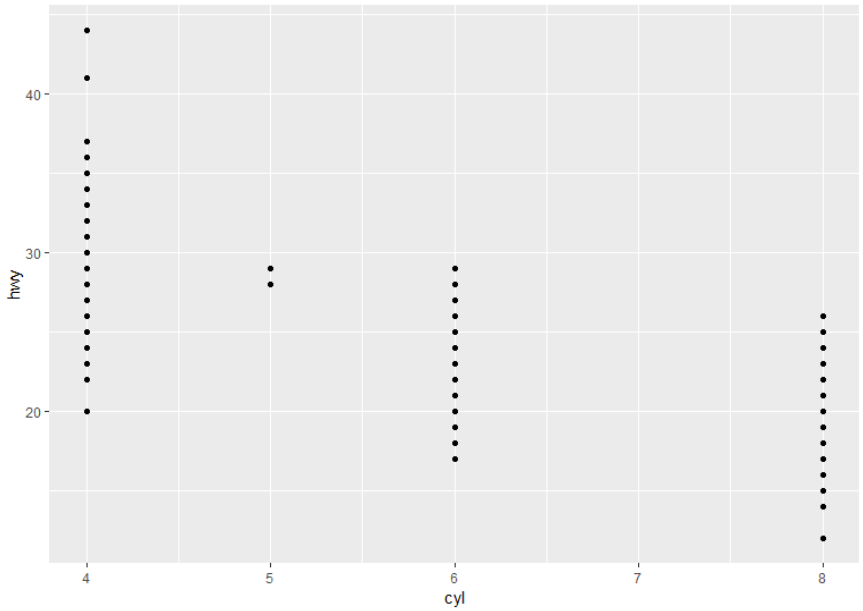
library(tidyverse)

data(mpg)

#1.(a, 3.2.4 Exercises#4)

Make a scatterplot of hwy vs cyl

ggplot(data = mpg)+ geom_point(mapping = aes(x=cyl, y=hwy))



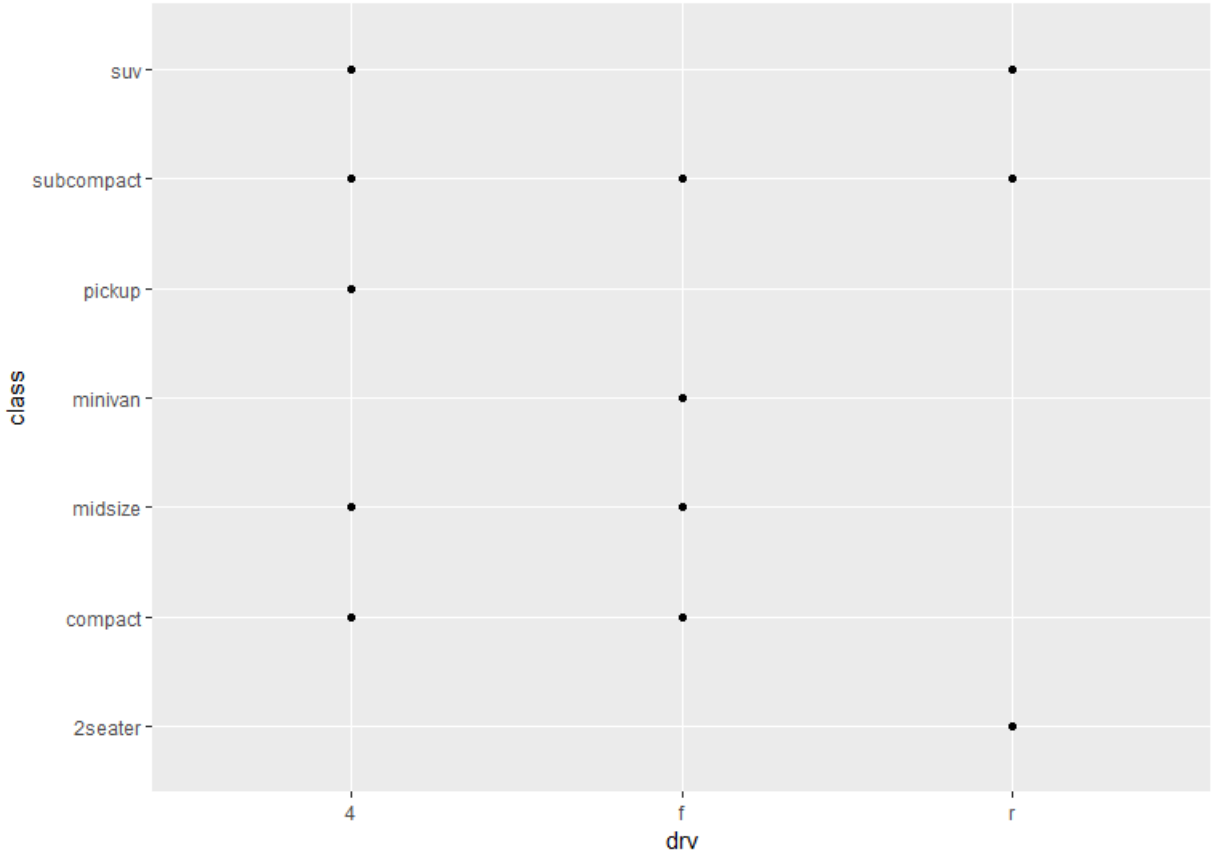
#1.(a, 3.2.4 #5)

What happens if you make a scatterplot of class vs drv ?

#Why is the plot not useful?

ggplot(data = mpg)+ geom_point(mapping = aes(x=drv, y=class))

The plot is not much useful because there is not any trend in the plot and nothing can be concluded out of the plot.



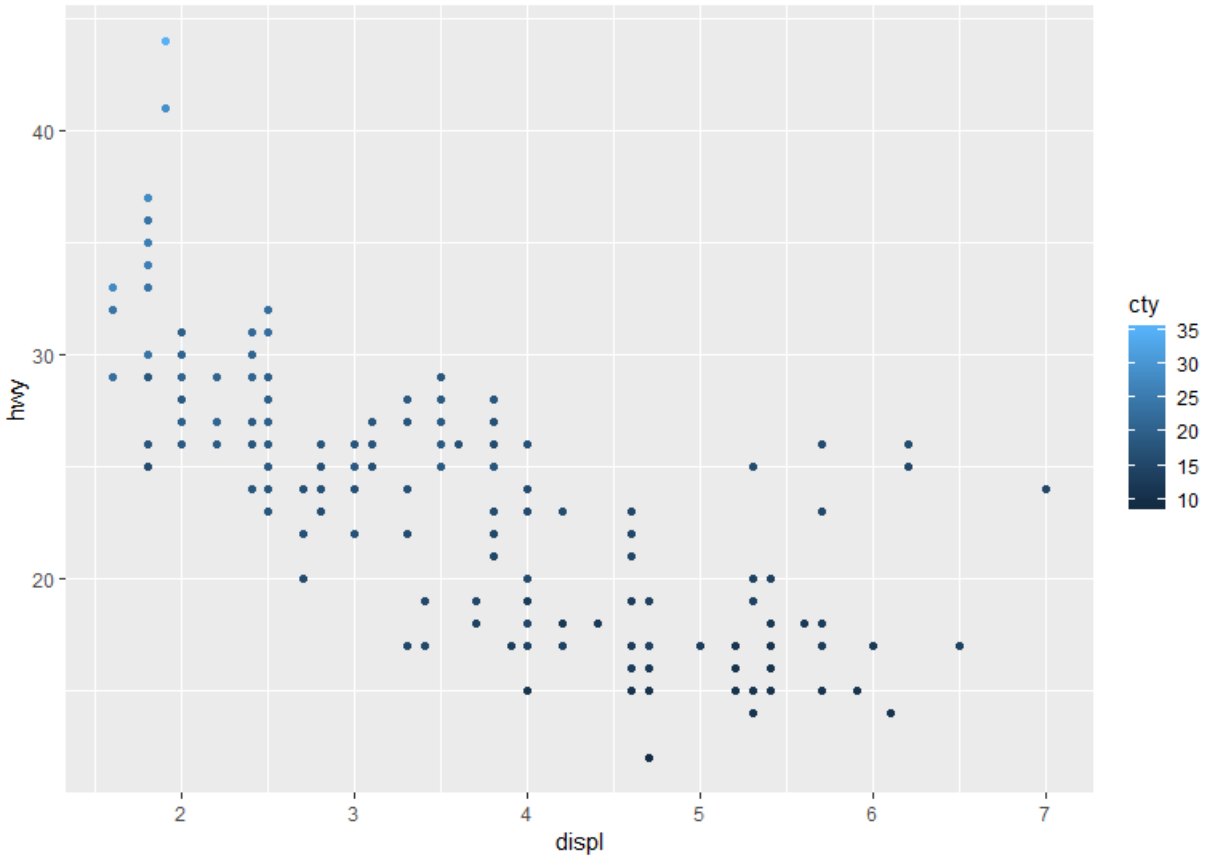
#1.(a, 3.3.1 Exercises#3)

Map a continuous & categorical variable to color, size and shape

#continuous variable map to color

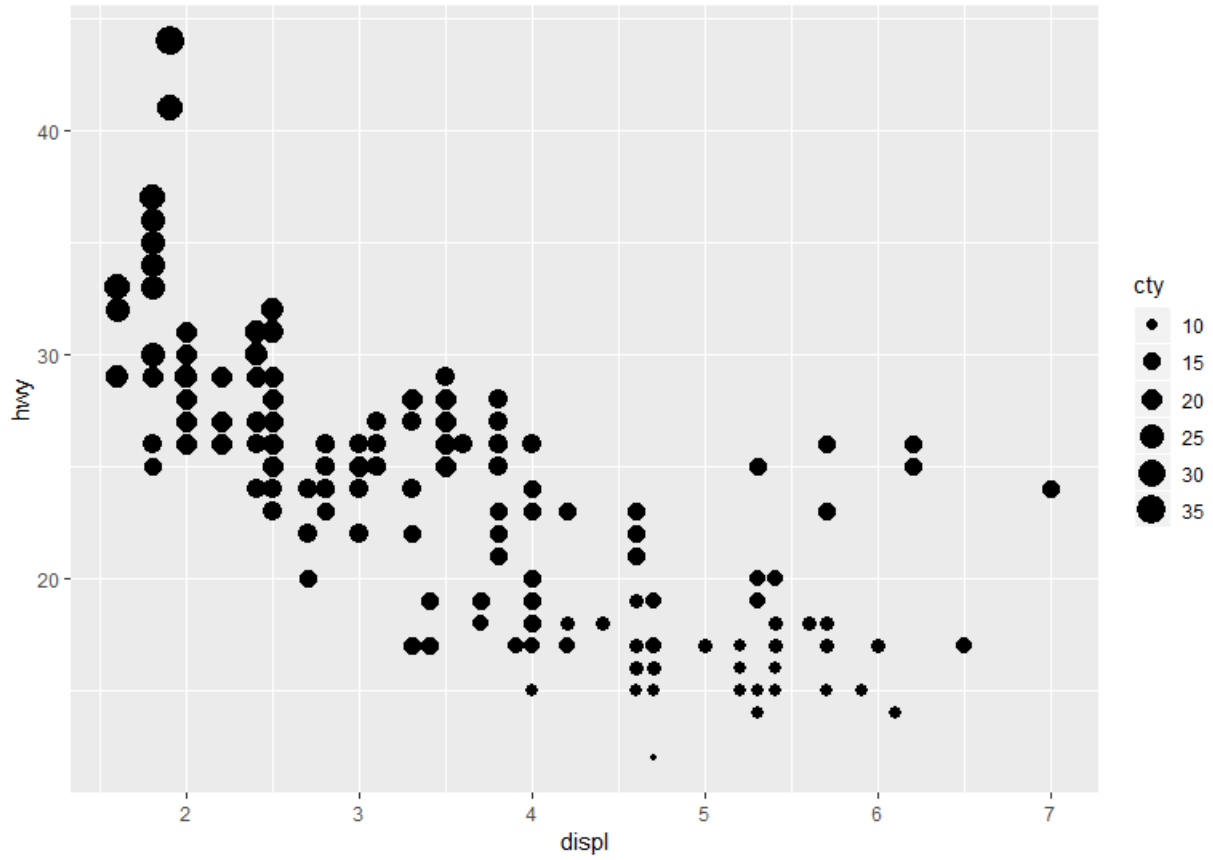
ggplot(data = mpg)+

geom_point(mapping = aes(x=displ, y=hwy, color=cty))



#continuous variable map to size

```
ggplot(data = mpg)+
  geom_point(mapping = aes(x=displ, y=hwy, size=cty))
```

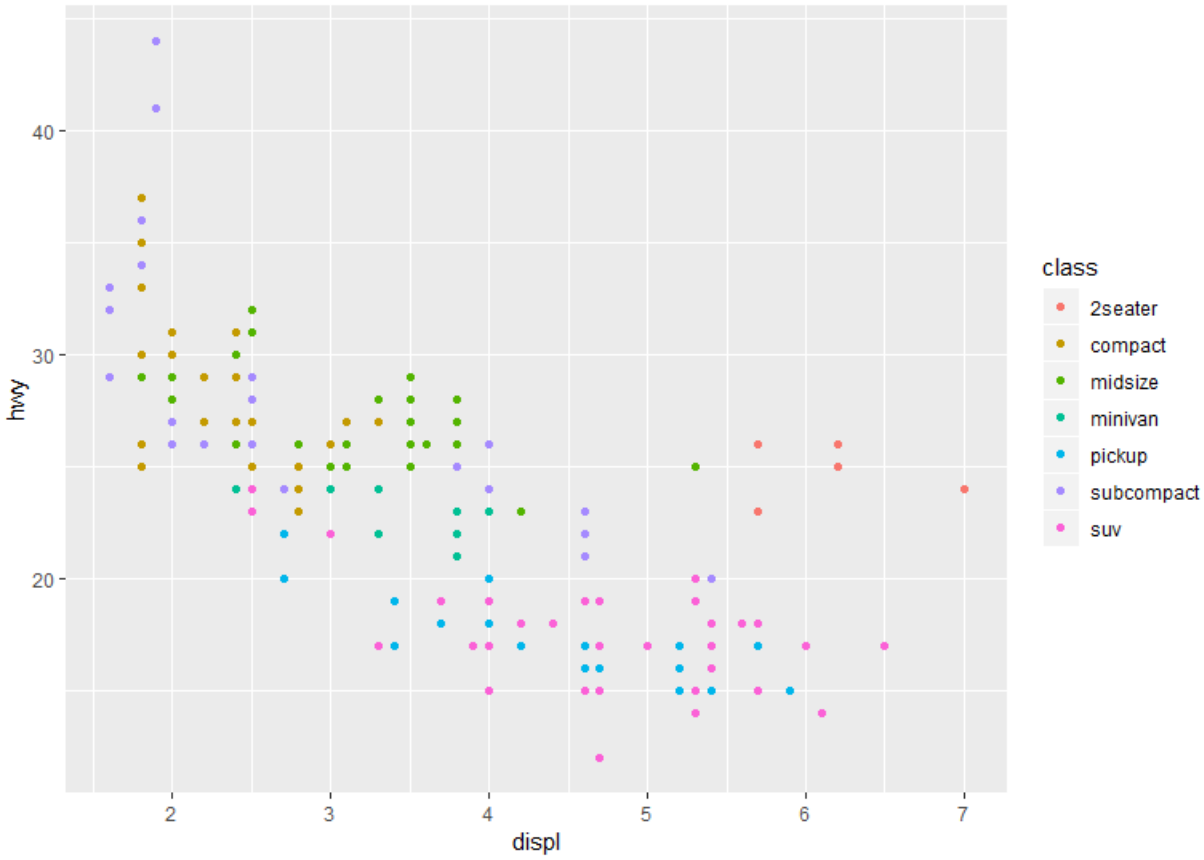


#continuous variable map to shape

```
ggplot(data = mpg)+
  geom_point(mapping = aes(x=displ, y=hwy, shape=cty))
```

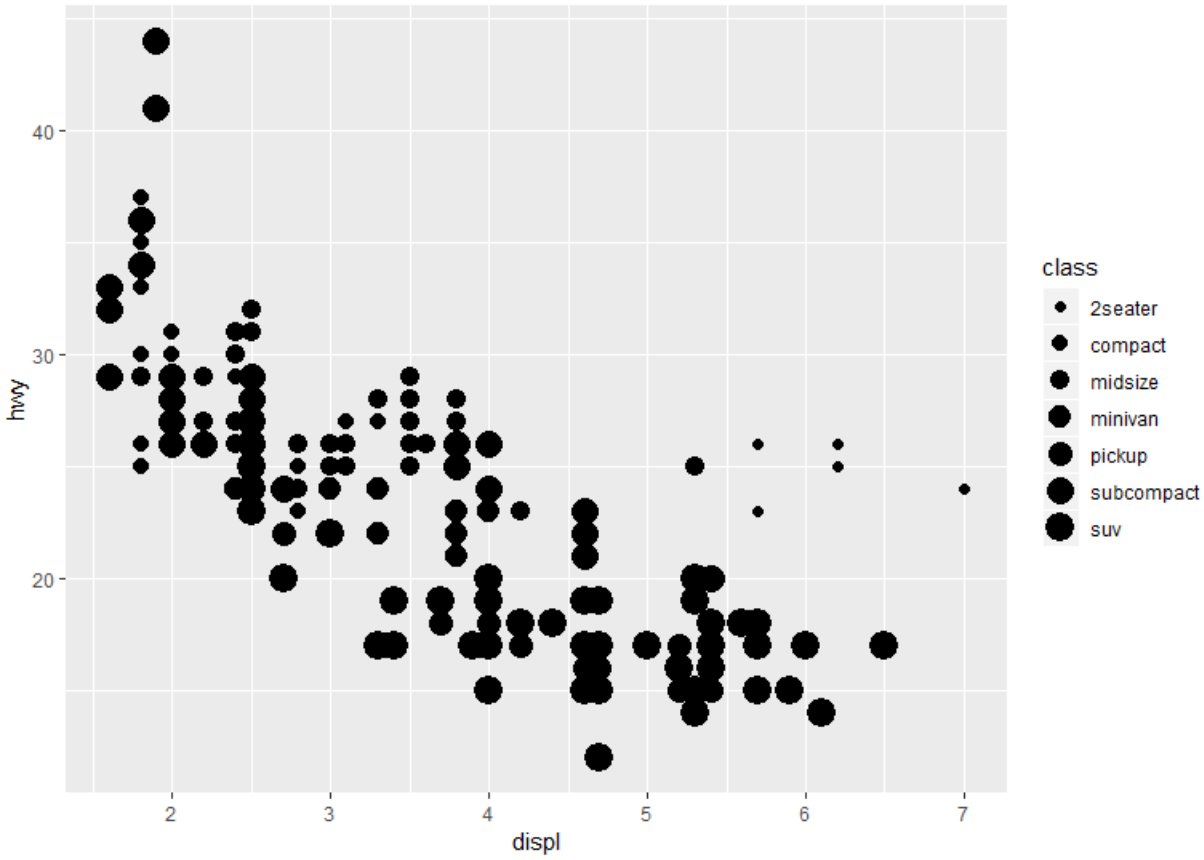
#categorical variable map to color

```
ggplot(data = mpg)+
  geom_point(mapping = aes(x=displ, y=hwy, color=class))
```



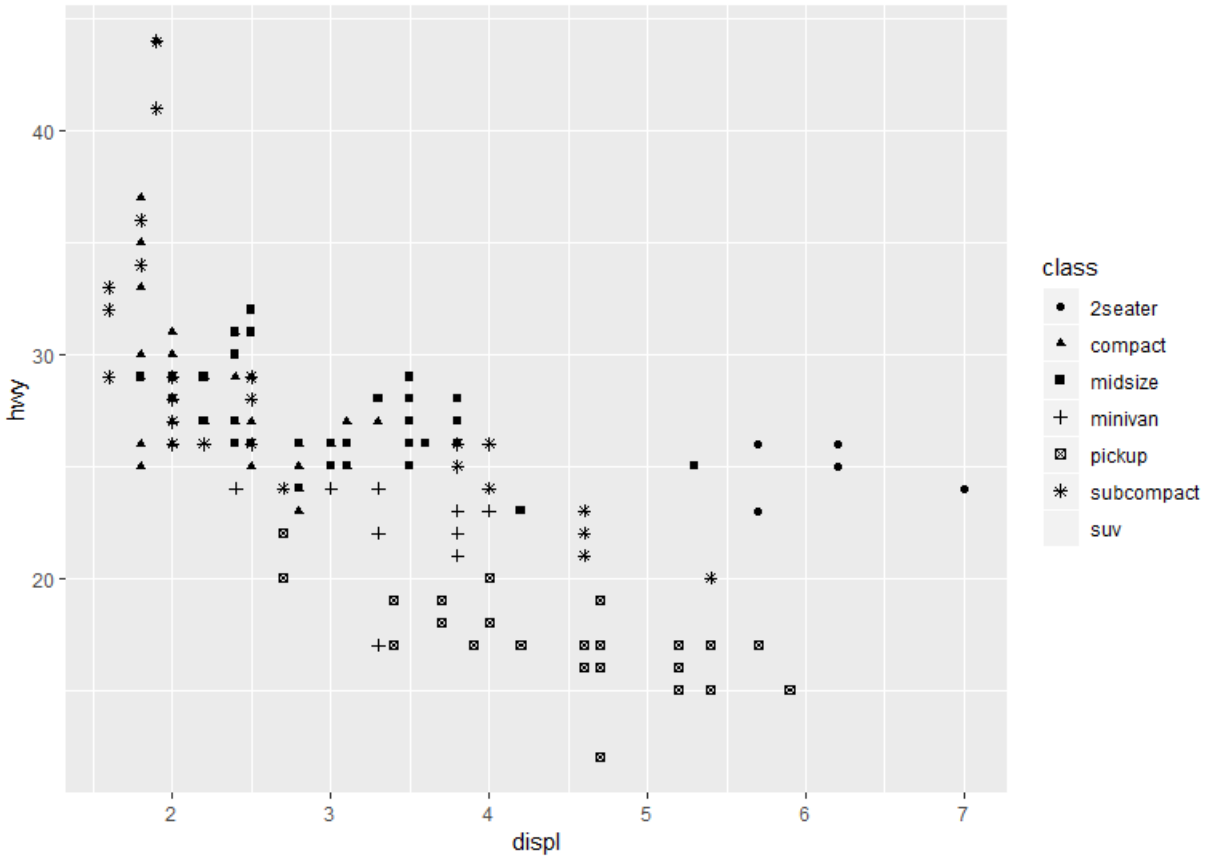
#categorical variable map to size

```
ggplot(data = mpg)+
  geom_point(mapping = aes(x=displ, y=hwy, size=class))
```



#categorical varible map to shape

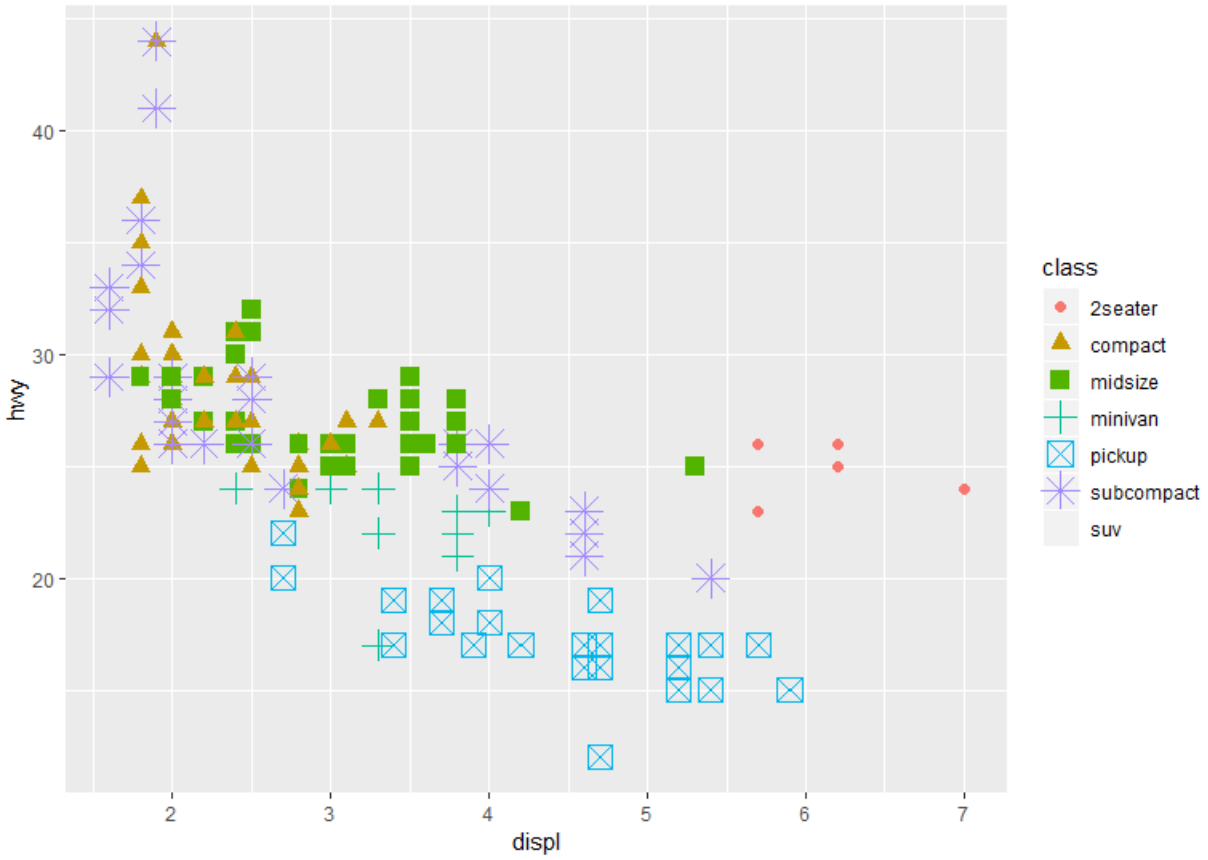
```
ggplot(data = mpg)+
  geom_point(mapping = aes(x=displ, y=hwy, shape=class))
```



#1.(a, 3.3.1 Exercises#4)

What happens if you map the same variable to multiple aesthetics?

```
ggplot(data = mpg)+
  geom_point(mapping = aes(x=displ, y=hwy, color=class, size=class, shape=class))
#We get a more refined plot where each feature is distinctly shown by different aesthetics
```



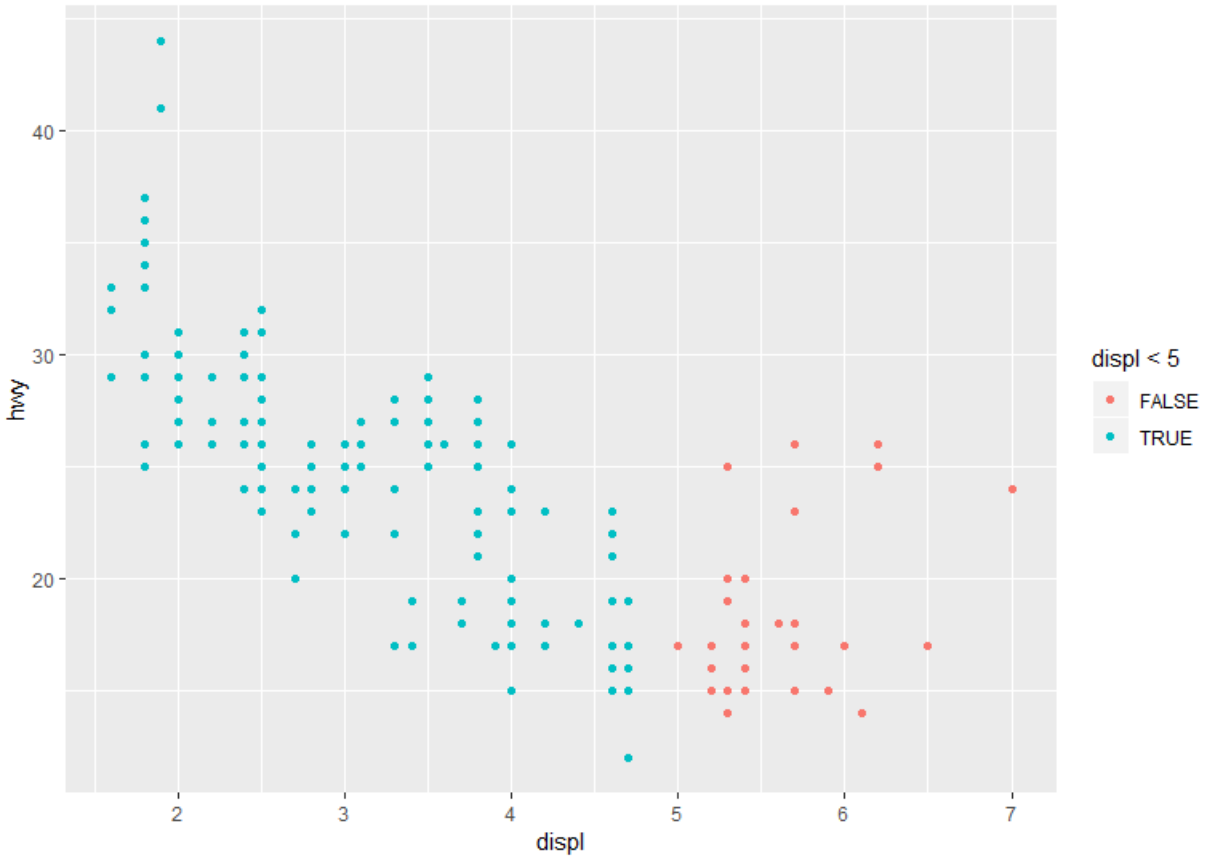
1.(a, 3.3.1 Exercises#6).

What happens if you map an aesthetic to something other than a variable name, like aes(colour = displ < 5) ?

ggplot(data = mpg)+

geom_point(mapping = aes(x=displ, y=hwy, color=displ<5))

here it shows different color for values which are true i.e <5 and different for >5



1.(a, 3.5.1 Exercises#4)

What are the advantages to using faceting instead of the colour aesthetic?

#What are the disadvantages?

#How might the balance change if you had a larger dataset?

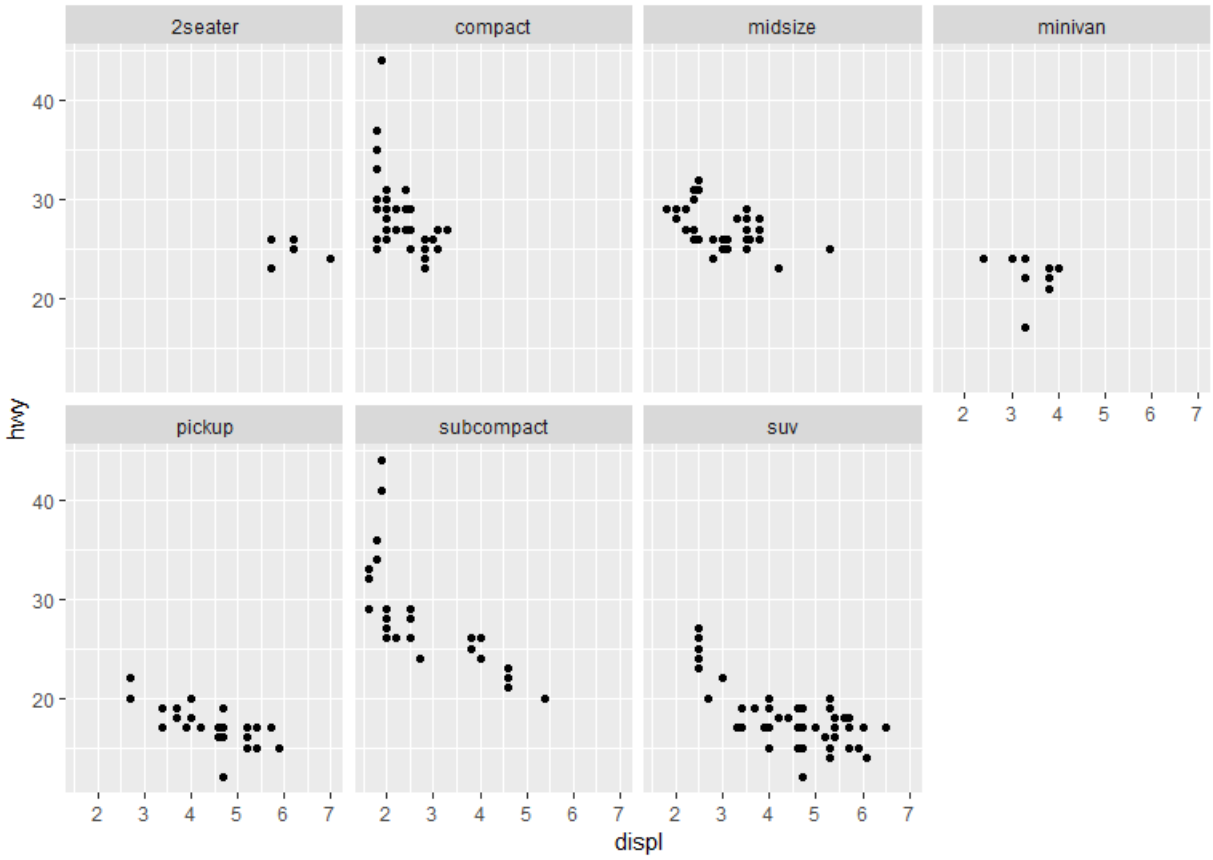
ggplot(data = mpg) +

geom_point(mapping = aes(x = displ, y = hwy)) + facet_wrap(~ class, nrow = 2)

Answer:

The faceted plot gives the plot of two variables for each of the third variable. Its main advantage is that it gives the glimpse of the correlation between two parameters across all the third category. However, color schemes od the same thing but faceted gives more clear picture compared to color for small data set.

Disadvantages: As the dataset of the third variable increase, the number of plots will increase and it will become difficult to visualize of get the clear picture out of those plots.

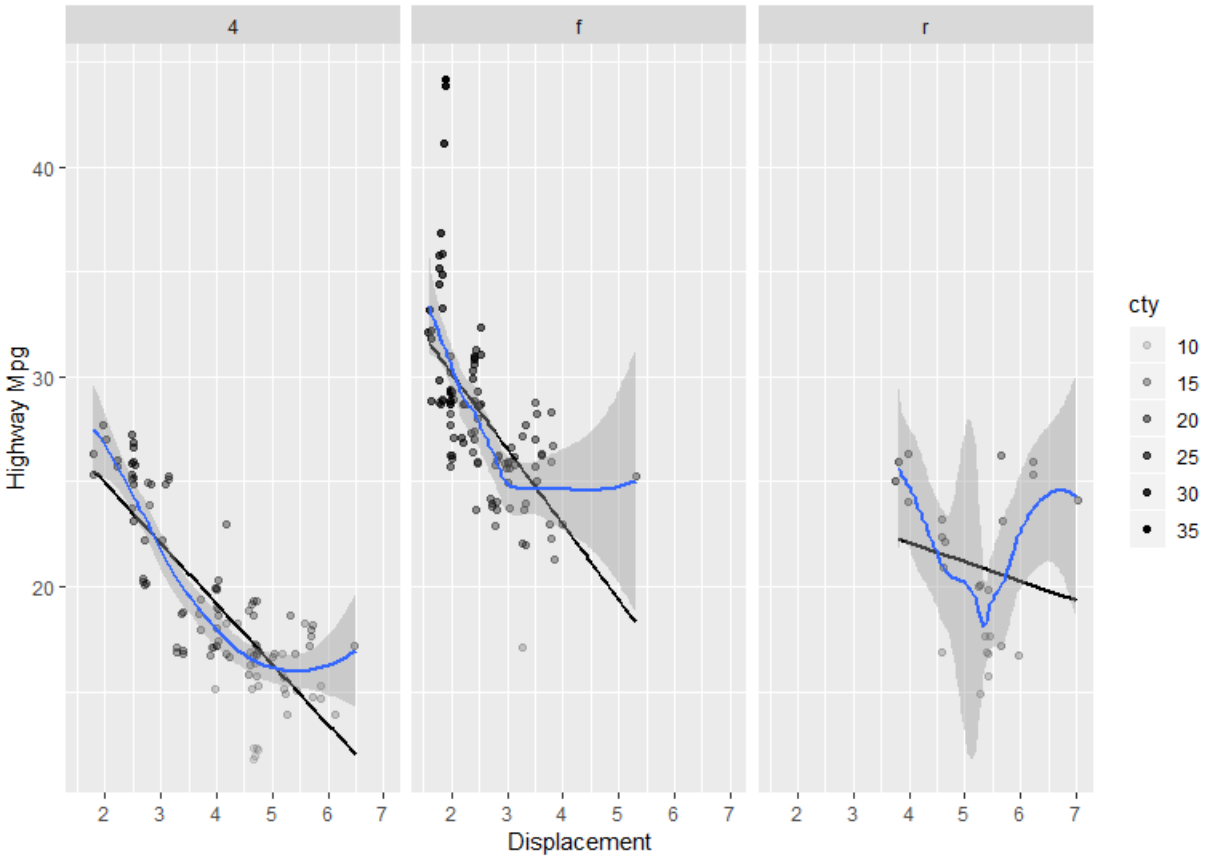


#1(b)

After reading this chapter, you should be ready to reproduce the plot in Figure 1

#using the same mpg data from above. Please do so.

```
p<-ggplot(data = mpg) +
  geom_point(mapping = aes(x=displ, y=hwy, alpha=cty), position = "jitter") +
  geom_smooth(mapping = aes(x = displ, y = hwy),method=lm,color="black", se=FALSE)+
  geom_smooth(mapping = aes(x = displ, y = hwy))+
  facet_wrap(~drv)
p+labs(x="Displacement", y="Highway Mpg")
```



Example 2(a.)

I have used random normal distribution, random poison distribution, random binomial distribution and random chi square distribution. gather function used to join all the four variables.

df<- data.frame(a=rnorm(500),b=rpois(500, lambda=3),c=rbinom(500,20,0.5),d=rchisq(500, df=3))

head(df)

```
           a  b  c      d
1  0.88591702  2 10 4.2808489
2  1.51094657  4  8 5.1385624
3  0.05073460  4  8 6.4322059
4 -0.02881551  2  7 1.2478455
5 -0.80646961  0 14 0.3320644
6  0.73189364  5 13 0.5103549
```

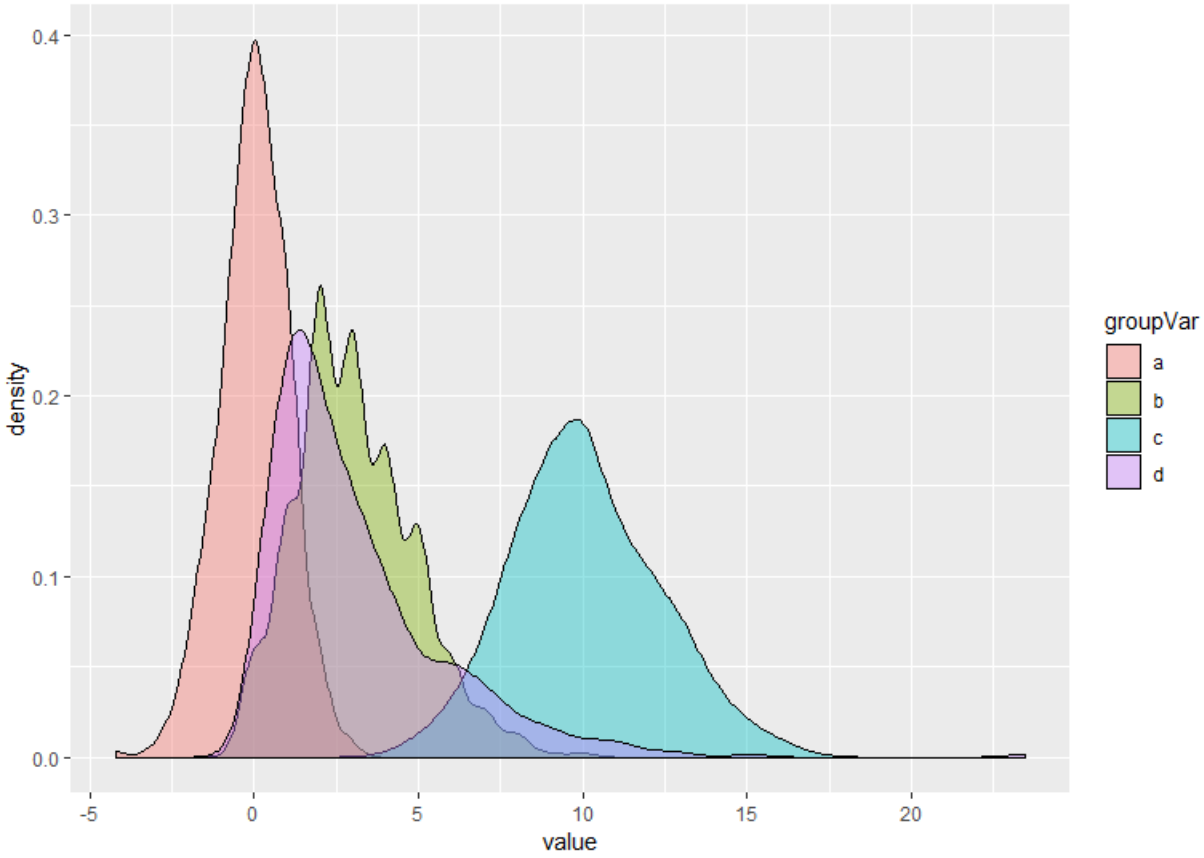
```
library(dplyr)

df2<-gather(df,key="groupVar", value="value")

head(df2)
```

```
  groupVar       value
1        a  0.88591702
2        a  1.51094657
3        a  0.05073460
4        a -0.02881551
5        a -0.80646961
6        a  0.73189364
```

Example 2(b.)

Plot the densities of each distribution overlaid on each other on one plot

```
library(ggplot2)

ggplot(df2,mapping= aes(fill = groupVar, x=value))+ geom_density(alpha=0.4)
```



Example -3

Housing proce data visualization

#read the CSV file form the working folder

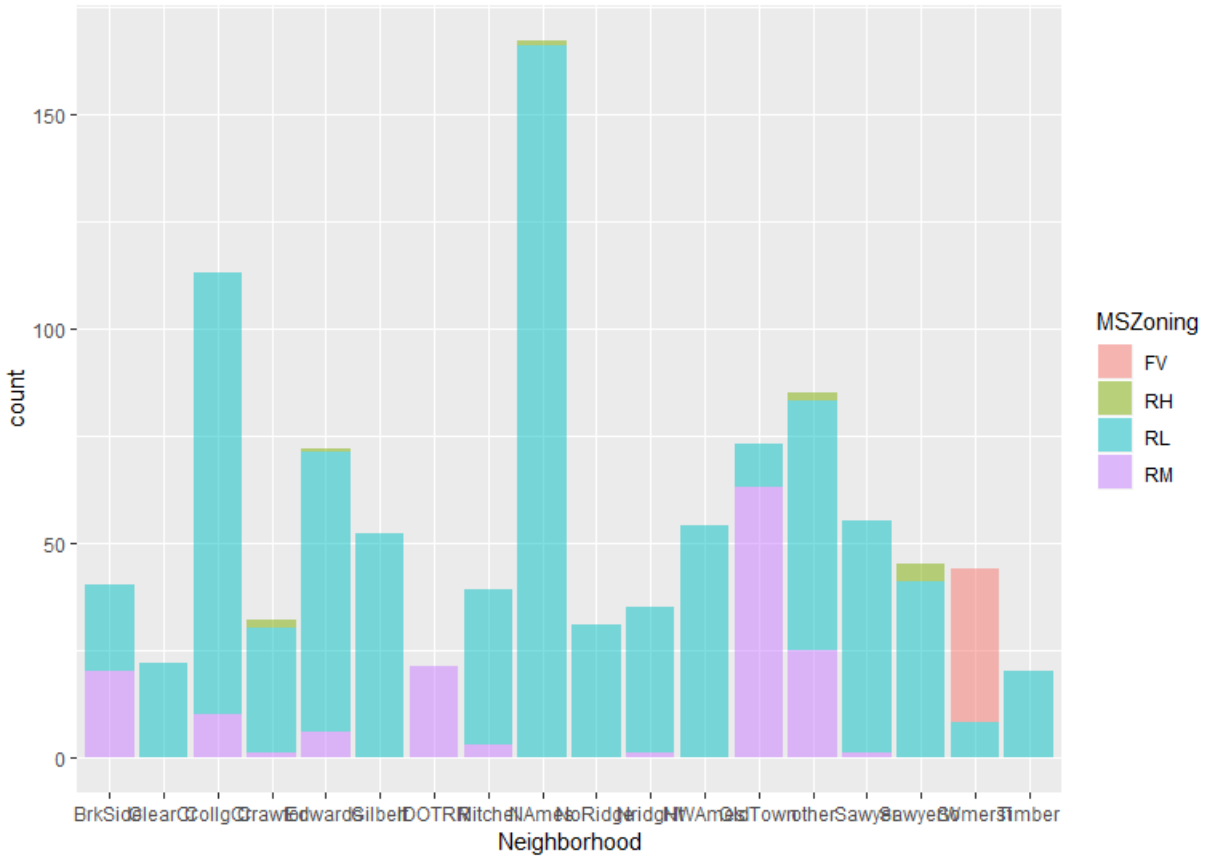
myData<-read.csv("housingData.csv")

scatter plot of sale price vs. lot area with housestyle as a color mapping

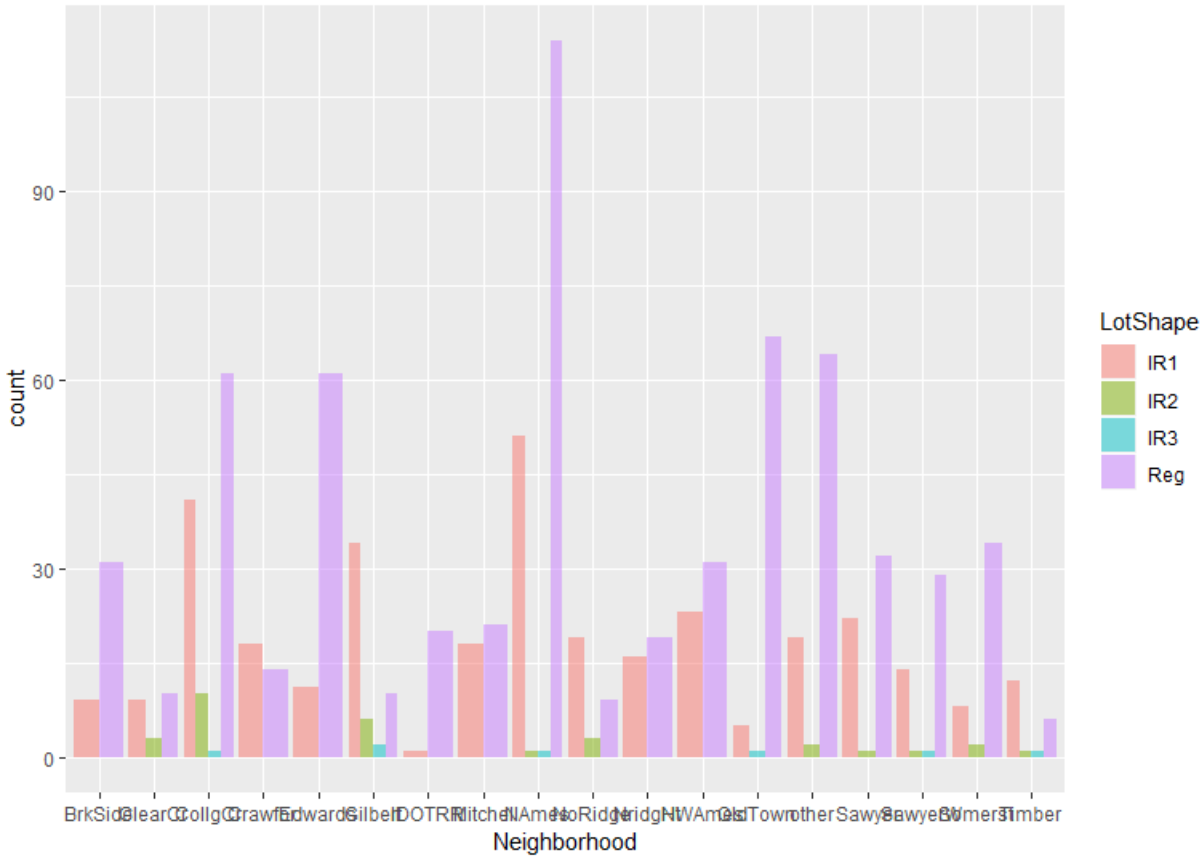
```
ggplot(data = myData)+
  geom_point(mapping = aes(x=LotArea, y=SalePrice, color=HouseStyle))
```

bar grap plot of neoghborhood with fill of MSZoning

```
ggplot(data = myData, mapping = aes(x=Neighborhood, fill=MSZoning))+
  geom_bar(alpha=0.5)
```



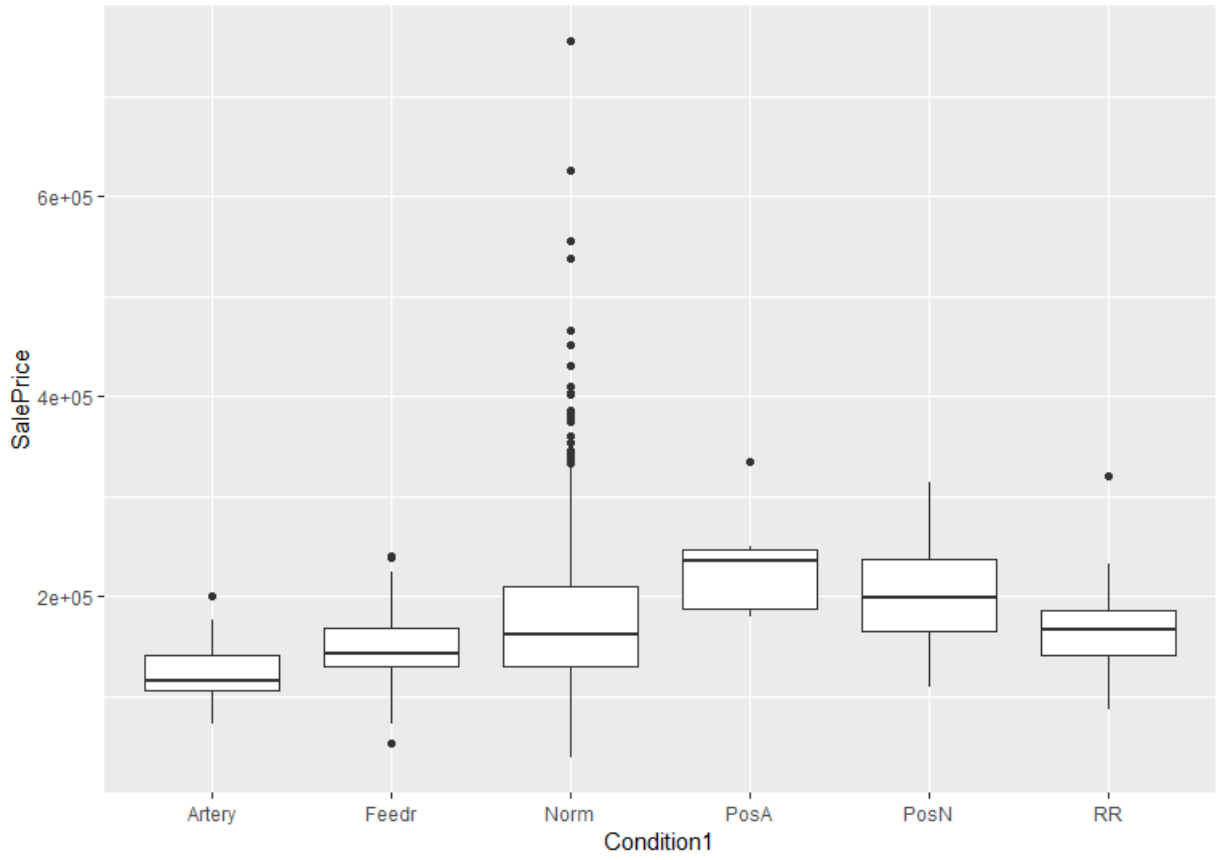

```
# bar grap plot of neoghborhood  with fill of lotshape and dodge aesthetic
ggplot(data = myData, mapping = aes(x=Neighborhood, fill=LotShape))+
  geom_bar(alpha=0.5, position = "dodge")
```



box plot of condition1 with saleprice

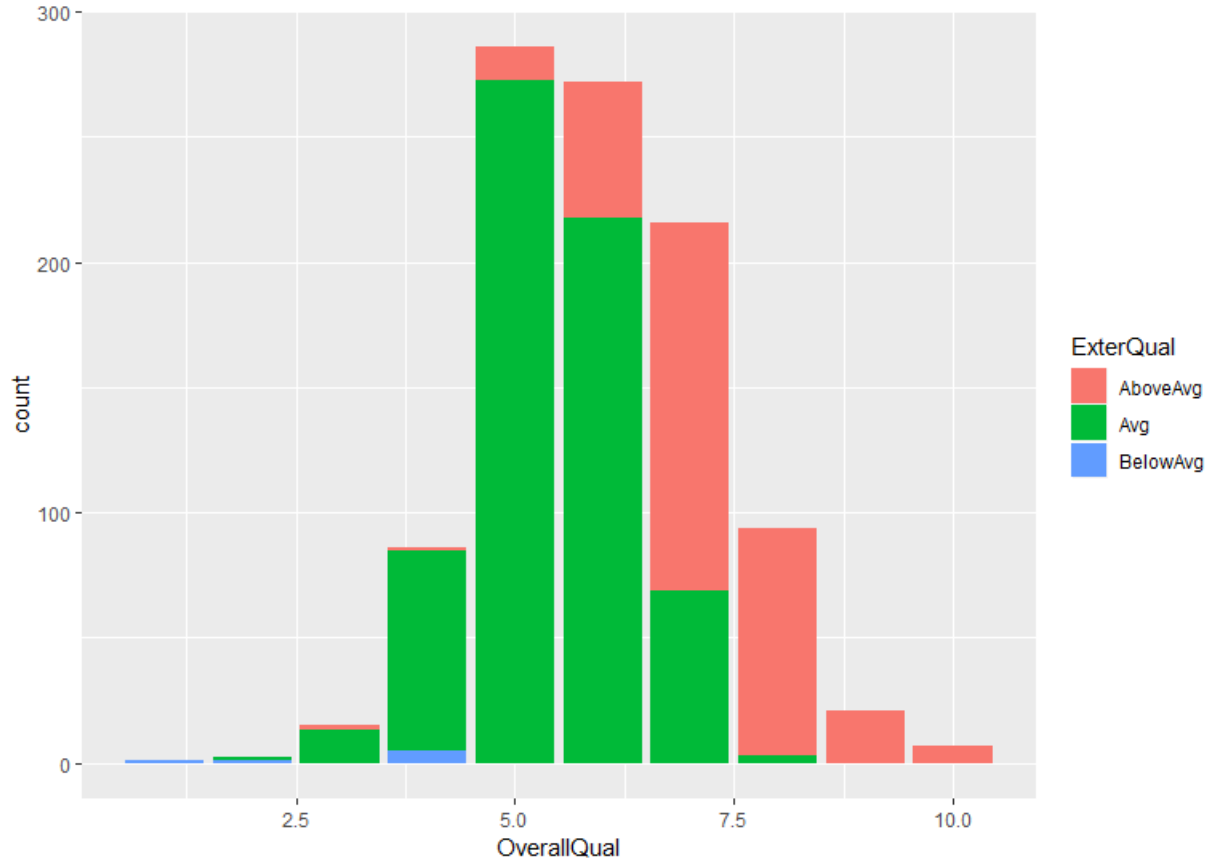
ggplot(data = myData, mapping = aes(x=Condition1, y=SalePrice))+

  geom_boxplot()

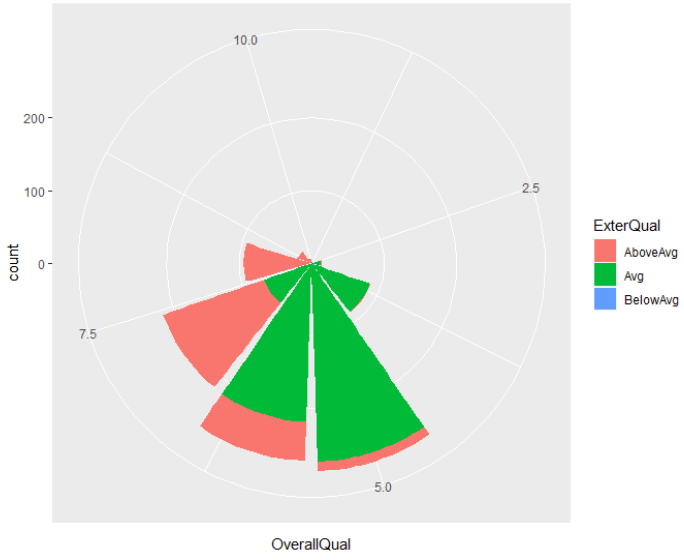


bar grap plot of overall quality with fill of external qualtity along with plot in polar coordinates

```
bar<- ggplot(data=myData)+
  geom_bar(mapping=aes(x=OverallQual, fill=ExterQual))
bar
```



bar+coord_polar()



Example 4: Missing Data Handling

#(4a.)

Explore the missingness of the data

#load the library Amelia and data freetrade

library(Amelia)

data(freetrade)

#load package mice for missingness analysis

install.packages("mice")

library(mice)

gives the number of observation per variable pair

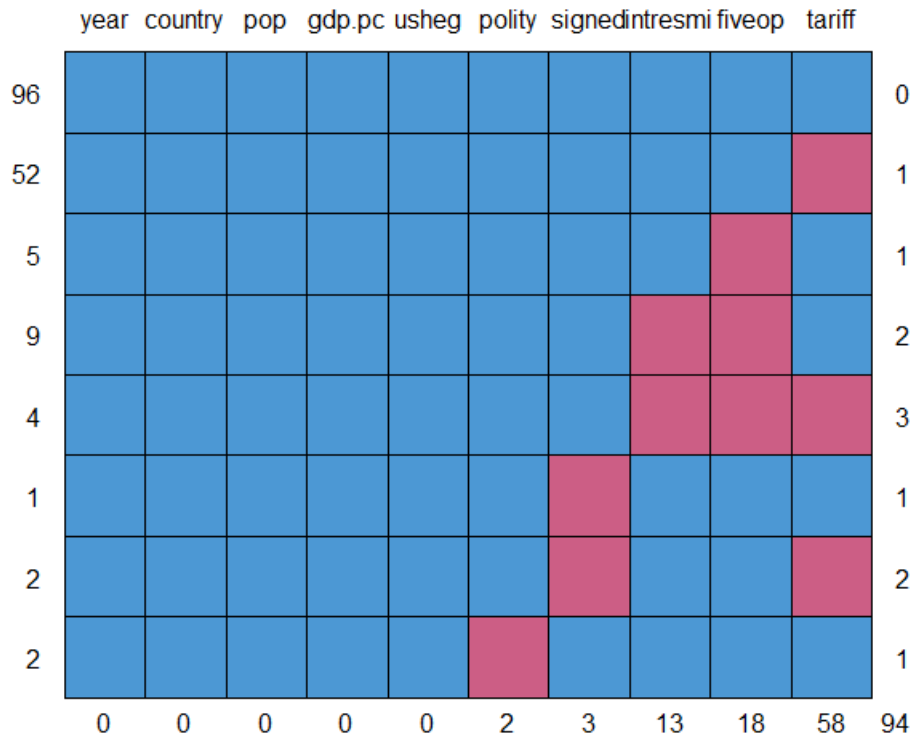
md.pairs(freetrade)

```
> md.pairs(freetrade)
$`rr`
         year country tariff polity pop gdp.pc intresmi signed fiveop usheg
year      171     171    113    169 171    171      158    168    153   171
country   171     171    113    169 171    171      158    168    153   171
tariff    113     113    113    111 113    113      104    112     99   113
polity    169     169    111    169 169    169      156    166    151   169
pop       171     171    113    169 171    171      158    168    153   171
gdp.pc    171     171    113    169 171    171      158    168    153   171
intresmi  158     158    104    156 158    158      158    155    153   158
signed    168     168    112    166 168    168      155    168    150   168
fiveop    153     153     99    151 153    153      153    150    153   153
usheg     171     171    113    169 171    171      158    168    153   171

$rm
         year country tariff polity pop gdp.pc intresmi signed fiveop usheg
year        0       0     58       2   0      0       13      3     18     0
country     0       0     58       2   0      0       13      3     18     0
tariff      0       0      0       2   0      0        9      1     14     0
polity      0       0     58       0   0      0       13      3     18     0
pop         0       0     58       2   0      0       13      3     18     0
```

```
gdp.pc      0       0     58       2   0       0       13       3     18     0
intresmi    0       0     54       2   0       0        0       3      5     0
signed      0       0     56       2   0       0       13       0     18     0
fiveop      0       0     54       2   0       0        0       3      0     0
usheg       0       0     58       2   0       0       13       3     18     0

$mr
         year country tariff polity pop gdp.pc intresmi signed fiveop usheg
year        0       0      0      0   0      0        0      0      0     0
country     0       0      0      0   0      0        0      0      0     0
tariff     58      58      0     58  58     58       54     56     54    58
polity      2       2      2      0   2      2        2      2      2     2
pop         0       0      0      0   0      0        0      0      0     0
gdp.pc      0       0      0      0   0      0        0      0      0     0
intresmi   13      13      9     13  13     13        0     13      0    13
signed      3       3      1      3   3      3        3      0      3     3
fiveop     18      18     14     18  18     18        5     18      0    18
usheg       0       0      0      0   0      0        0      0      0     0

$mm
         year country tariff polity pop gdp.pc intresmi signed fiveop usheg
year        0       0      0      0   0      0        0      0      0     0
country     0       0      0      0   0      0        0      0      0     0
tariff      0       0     58      0   0      0        4      2      4     0
polity      0       0      0      2   0      0        0      0      0     0
pop         0       0      0      0   0      0        0      0      0     0
gdp.pc      0       0      0      0   0      0        0      0      0     0
intresmi    0       0      4      0   0      0       13      0     13     0
signed      0       0      2      0   0      0        0      3      0     0
fiveop      0       0      4      0   0      0       13      0     18     0
usheg       0       0      0      0   0      0        0      0      0     0
```

display missing data pattern

md.pattern(freetrade)

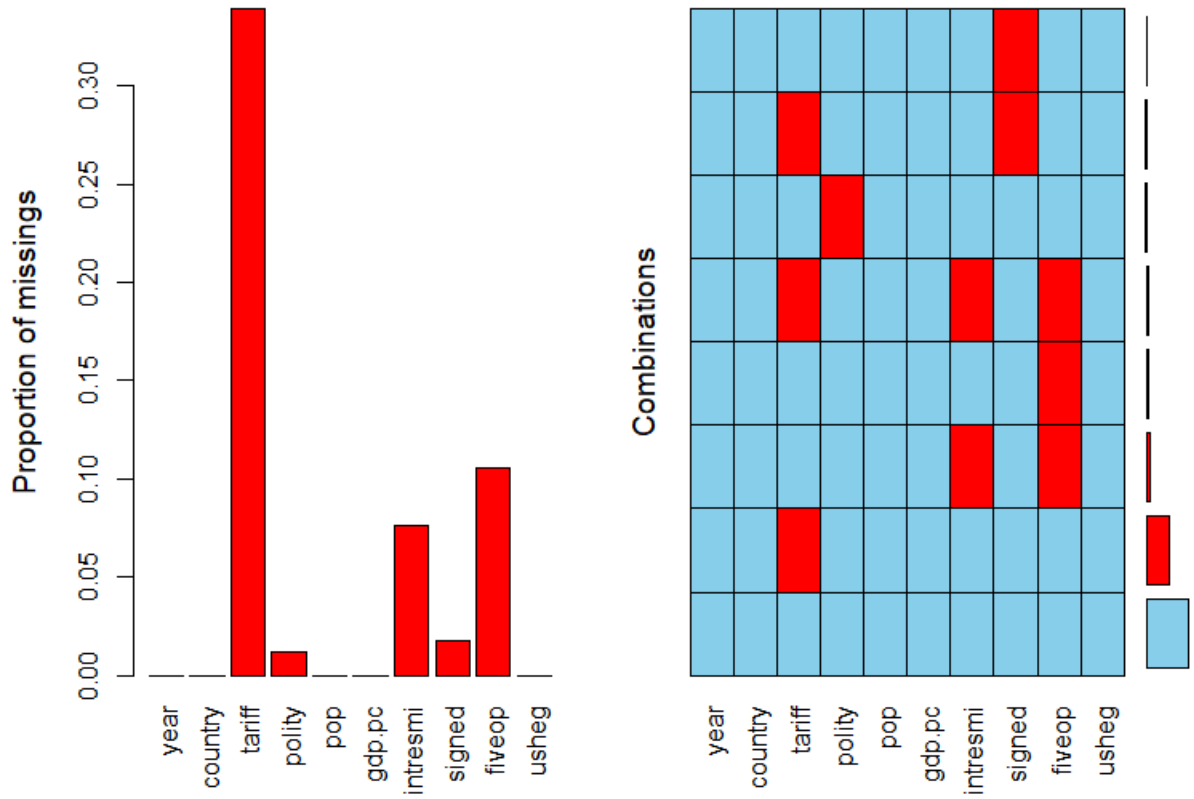


#load package VIM for missingness analysis

library(VIM)

#used VIM's "aggr" function to also get overall information on missing

a<-aggr(freetrade)



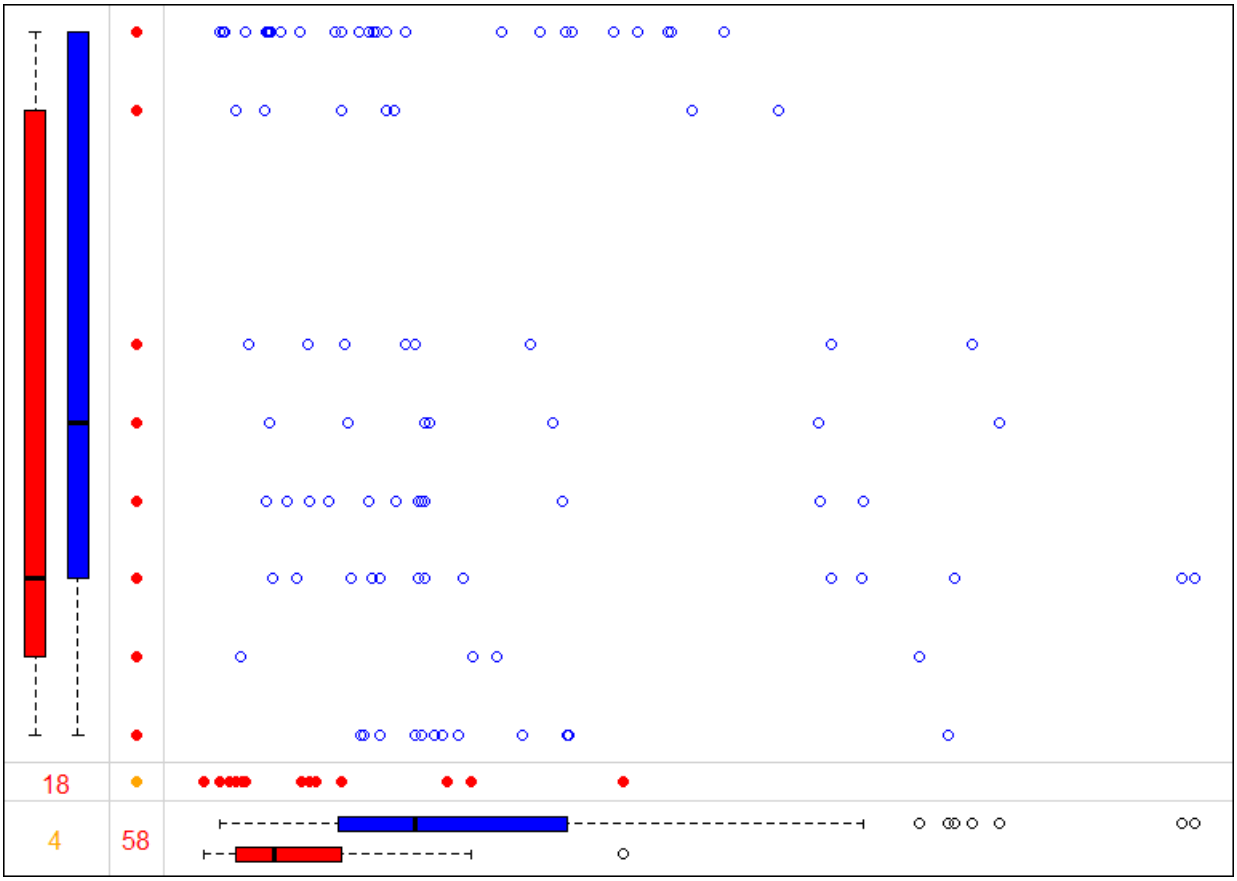
summary(a)

```
> #used VIM's "aggr" function to also get overall information on missing
> a<-aggr(freetrade)
> summary(a)

 Missings per variable:
 Variable Count
     year     0
  country     0
   tariff    58
   polity     2
      pop     0
   gdp.pc     0
 intresmi    13
   signed     3
   fiveop    18
    usheg     0

 Missings in combinations of variables:
        Combinations Count    Percent
 0:0:0:0:0:0:0:0:0:0    96 56.1403509
```

```
0:0:0:0:0:0:0:0:1:0      5  2.9239766
0:0:0:0:0:0:0:1:0:0      1  0.5847953
0:0:0:0:0:0:1:0:1:0      9  5.2631579
0:0:0:1:0:0:0:0:0:0      2  1.1695906
0:0:1:0:0:0:0:0:0:0     52 30.4093567
0:0:1:0:0:0:0:1:0:0      2  1.1695906
0:0:1:0:0:0:1:0:1:0      4  2.3391813
```

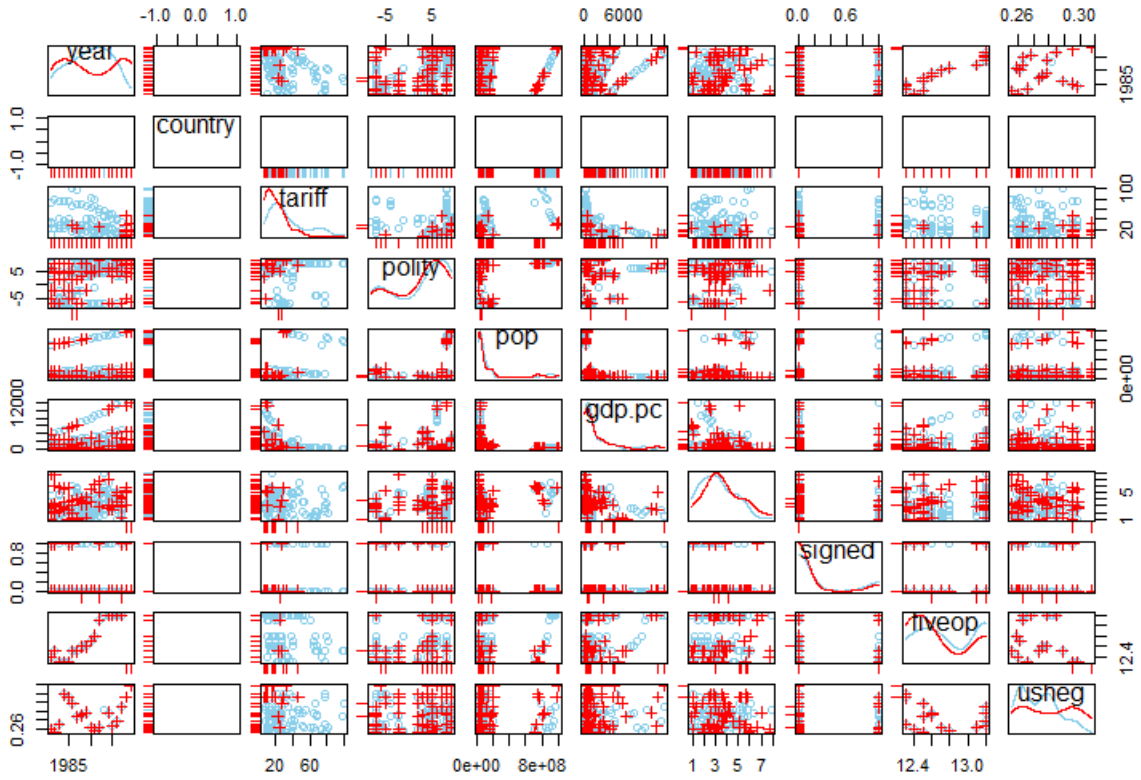
use VIM function "marginplot" to get a scatter plot that includes information on missing values

marginplot(freetrade[c("tariff","fiveop")], col=c("blue","red","orange"))



#looking at all of the plots with Missing Information

scattmatrixMiss(freetrade)



#(4b.)

statistical test chi-square used to determine if the missingness in the

#tariff variable is independent with the country variable

chisq.test(freetrade$tariff,freetrade$country)

```
Pearson's Chi-squared test

data:  freetrade$tariff and freetrade$country
X-squared = 831.96, df = 736, p-value = 0.007819
```

Since, p value is almost zero, we reject the null that the missingness of tariff is independent of county. So, they are dependent. It becomes clear when we remove Nepal and Phillipines, we see change in p-value.

#removed Nepal from the data and again prformed Chisq test

```
freetrade=freetrade[which(freetrade$country!="Nepal"), ]

chisq.test(freetrade$tariff,freetrade$country)

Pearson's Chi-squared test

data:  freetrade$tariff and freetrade$country
X-squared = 831.96, df = 736, p-value = 0.007819
```

#removed Philippines from the data and again prformed Chisq test

```
library(Amelia)

data(freetrade)

freetrade=freetrade[which(freetrade$country!="Philippines"), ]

chisq.test(freetrade$tariff,freetrade$country)


Pearson's Chi-squared test

data:  freetrade$tariff and freetrade$country
X-squared = 639.33, df = 574, p-value = 0.03012
```

Hence we see the missingness of tariff affects the p-value in the three cases. Initially, it was almost zero (0.007) indicating the dependence of both the parameter. When we remove Nepal p-value changes to (0.1) and when Philippines removed(p-value=0.03).