HW2

NgocTran

9/6/2019

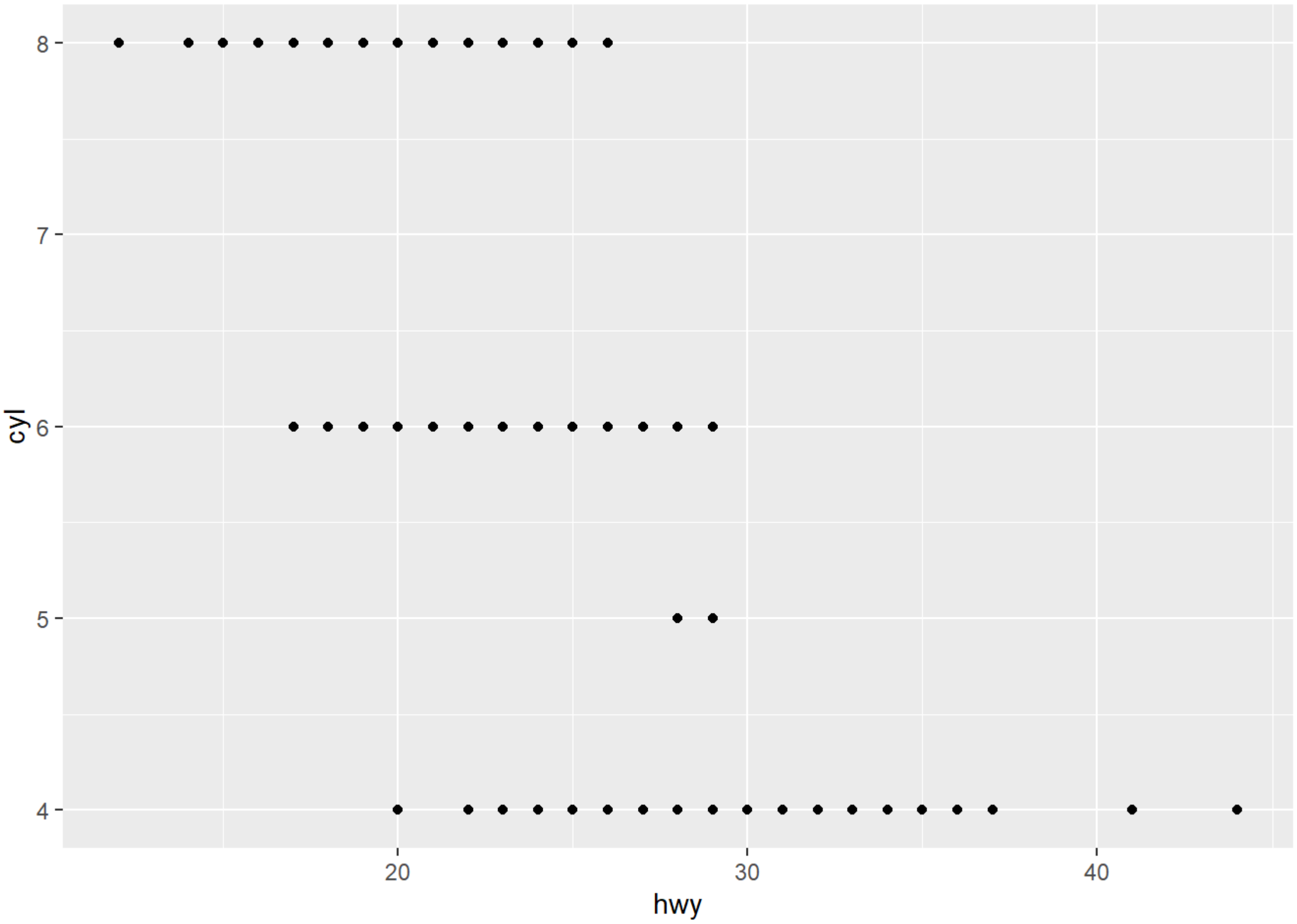
PROBLEM 1

a. *3.2.4 Exercise 4 and 5*

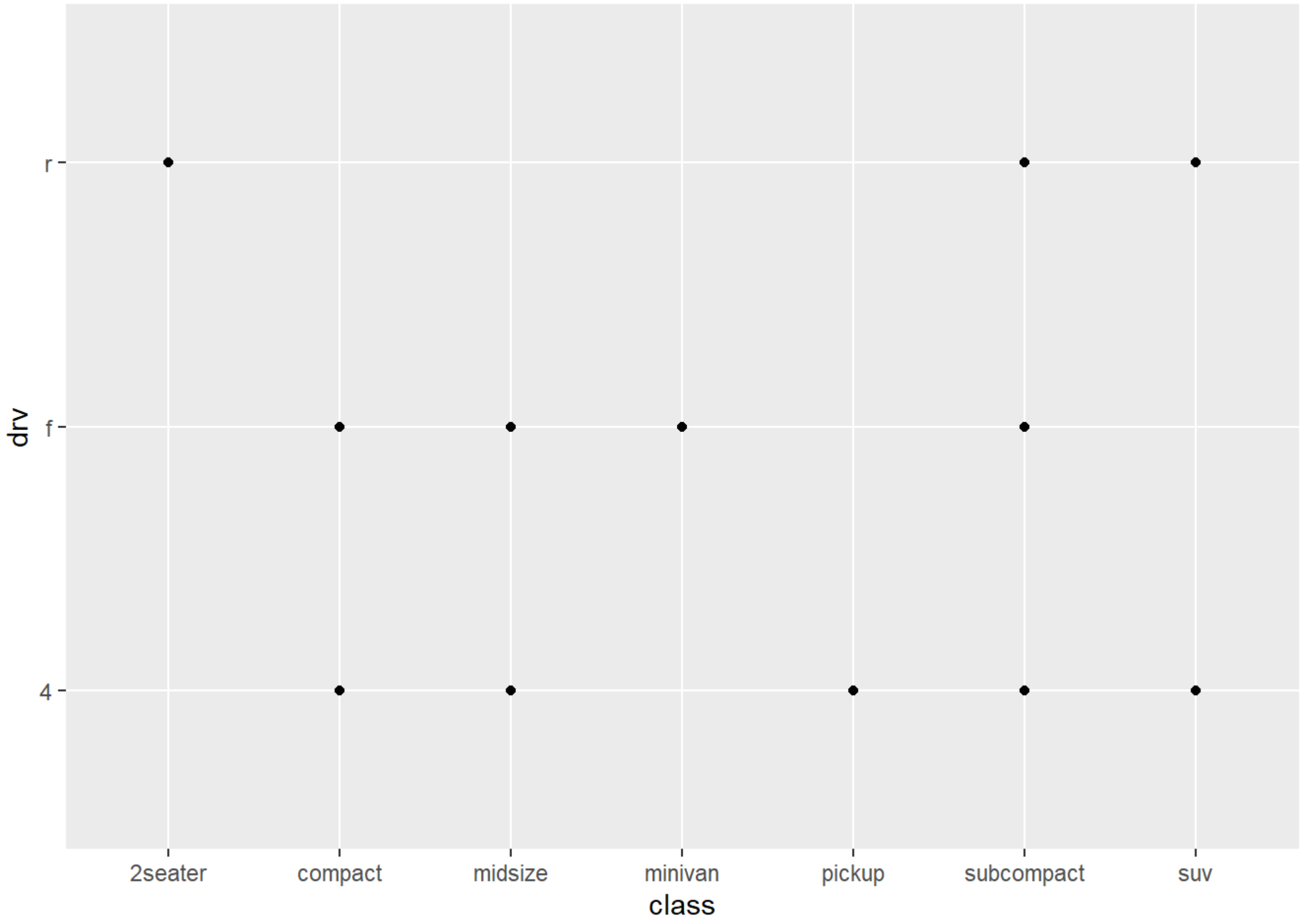
   Make scatter plotof hwy vs. cyl

   What happens if making the scatter plot of class vs drv ? why is the plot not sucessful

```
#Problem1a
#3.2.4 exercise 4 and 5
ggplot(mpg,aes(x=hwy,y=cyl))+geom_point() # scatterplot between hwy and cyl
```



```
ggplot(mpg,aes(x=class,y=drv))+geom_point() #scatterplot betweem class and drw
```



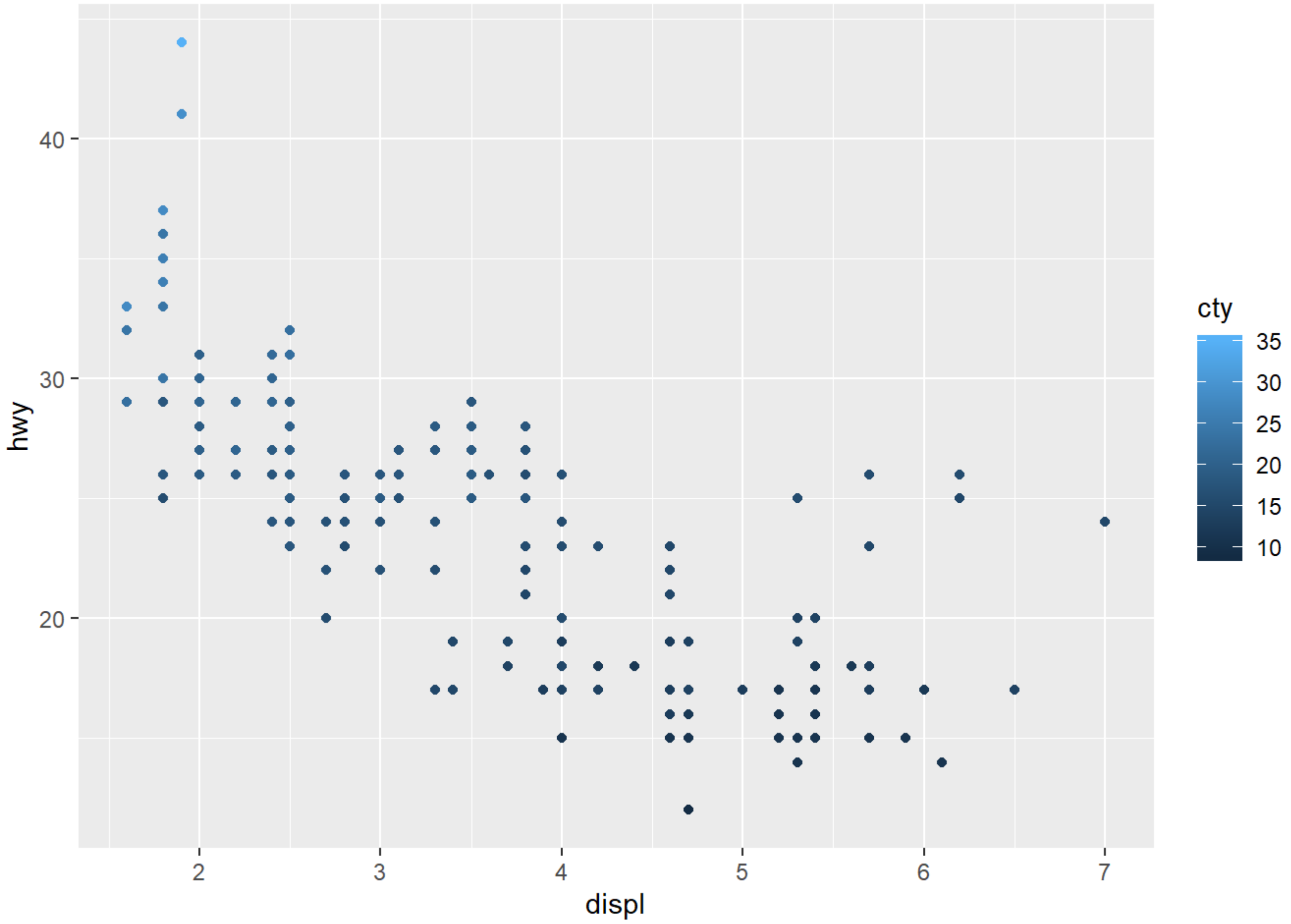
```
#the plot is not usefull because the scatterpoint doesnt show any trend
#not many datapoints
```

The second plot doesn’t have sufficient numbers of datapoints to make the scatterplot useful and the class values have many categories, which should be put on bin or box plots

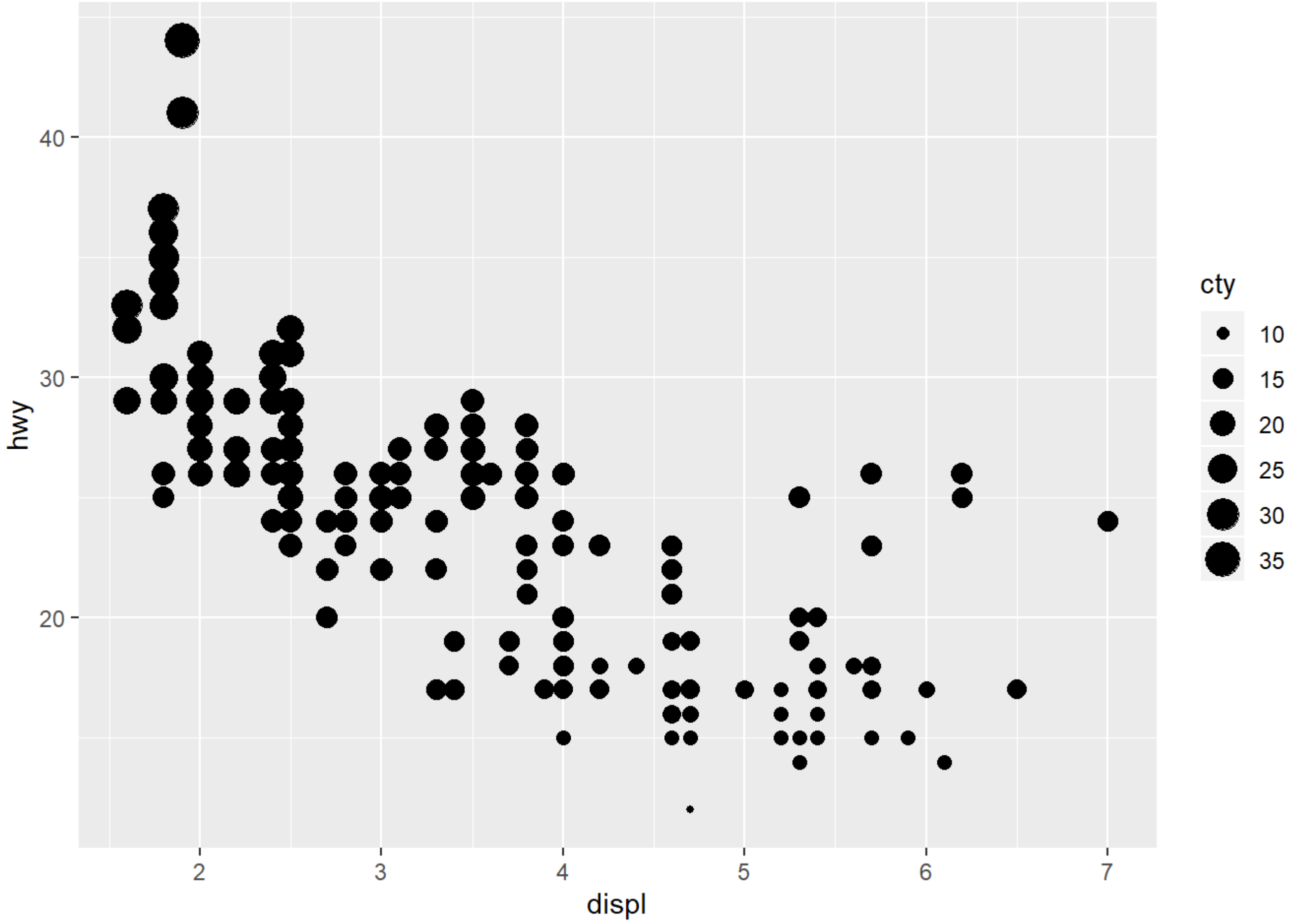
```
*3.3.1 Exercise 3*
```

Map a continuous variable to color, size, and shape. How do these aesthetics behave differently for categorical vs. continuous variables?

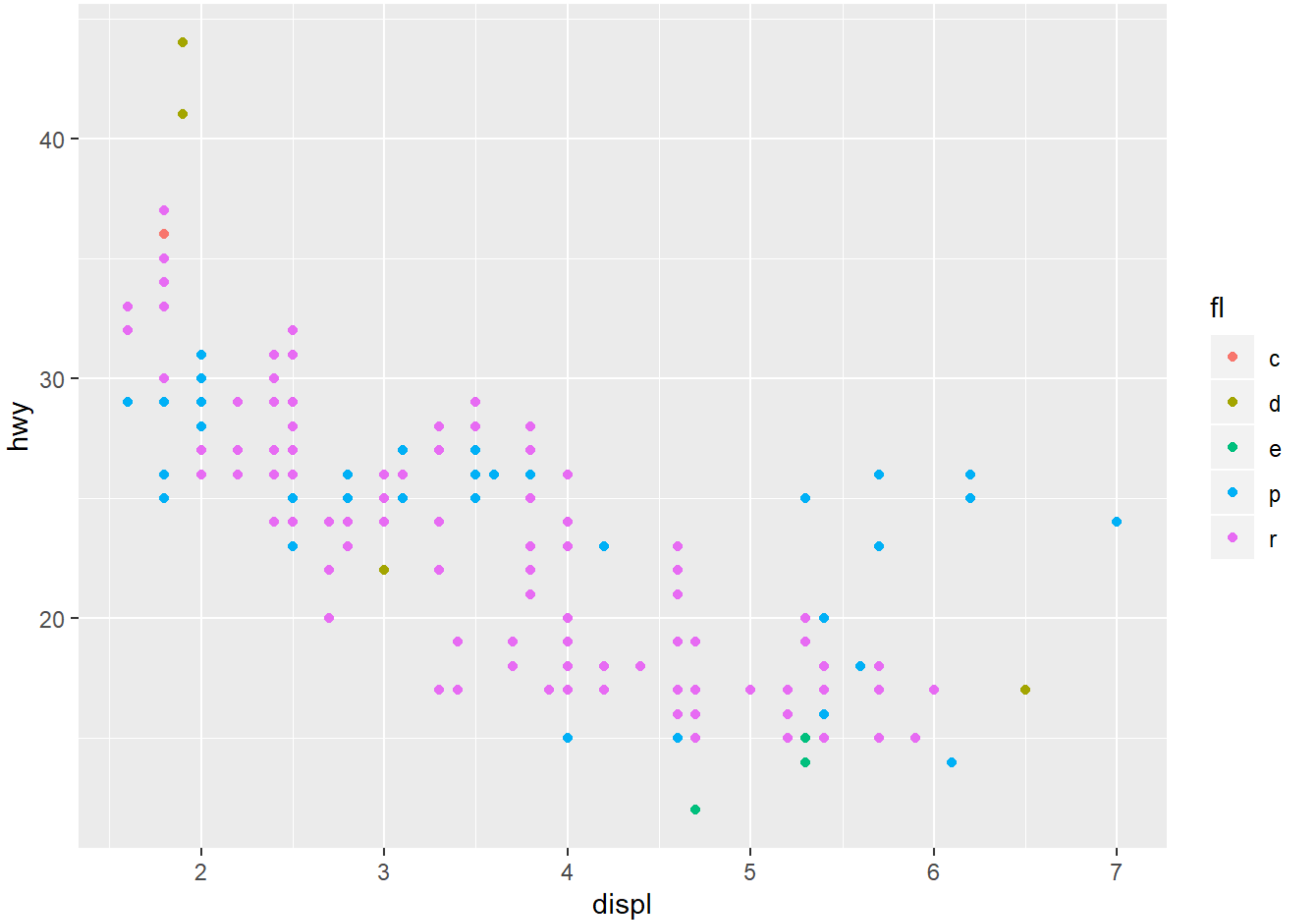
```
View(mpg)
#map continuous variables to color, size and shape
ggplot(data=mpg)+geom_point(mapping = aes(x=displ,y=hwy,col=cty)) #map cty to color
```



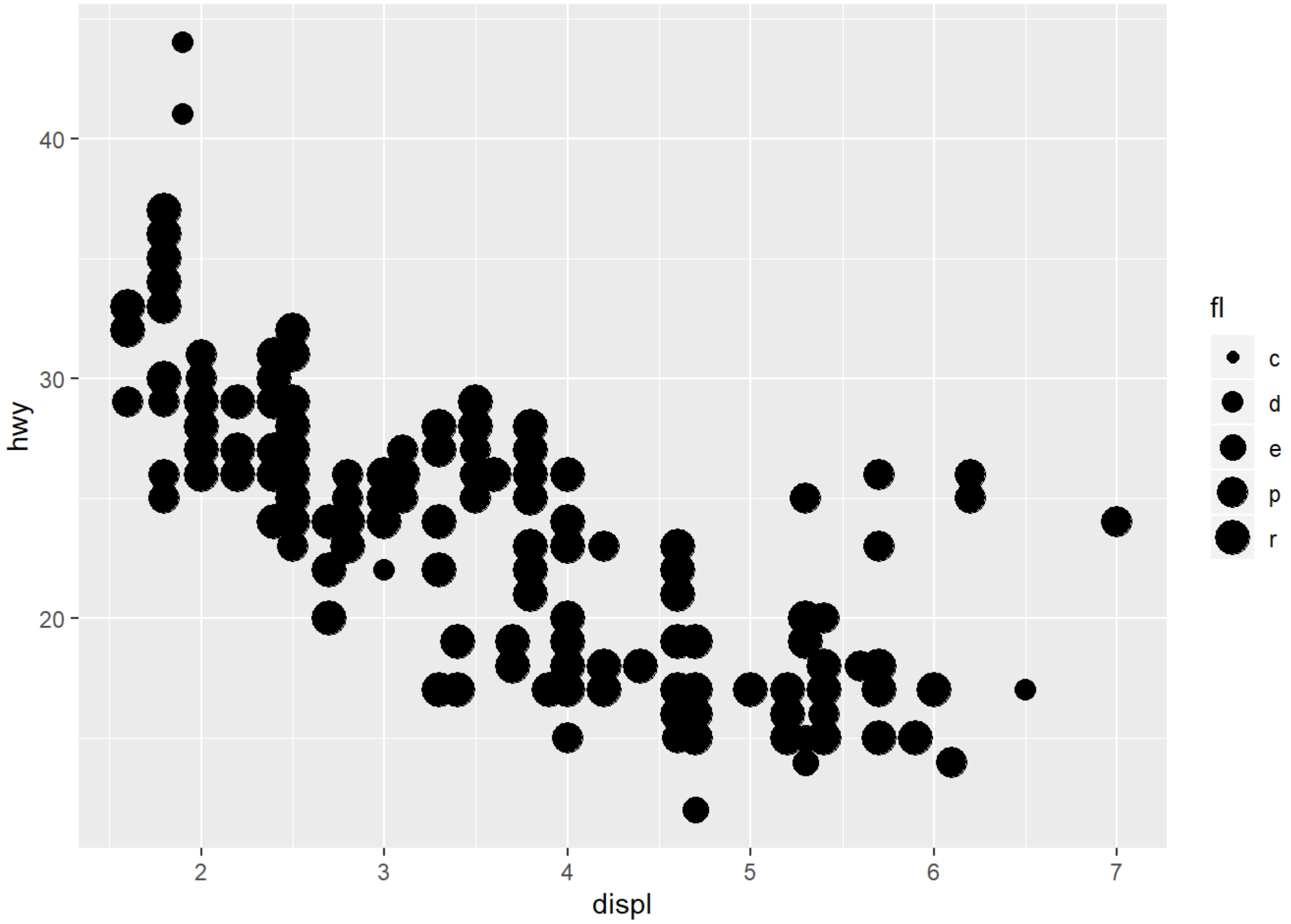
```
ggplot(data=mpg)+geom_point(mapping = aes(x=displ,y=hwy,size=cty)) #map cty to size
```



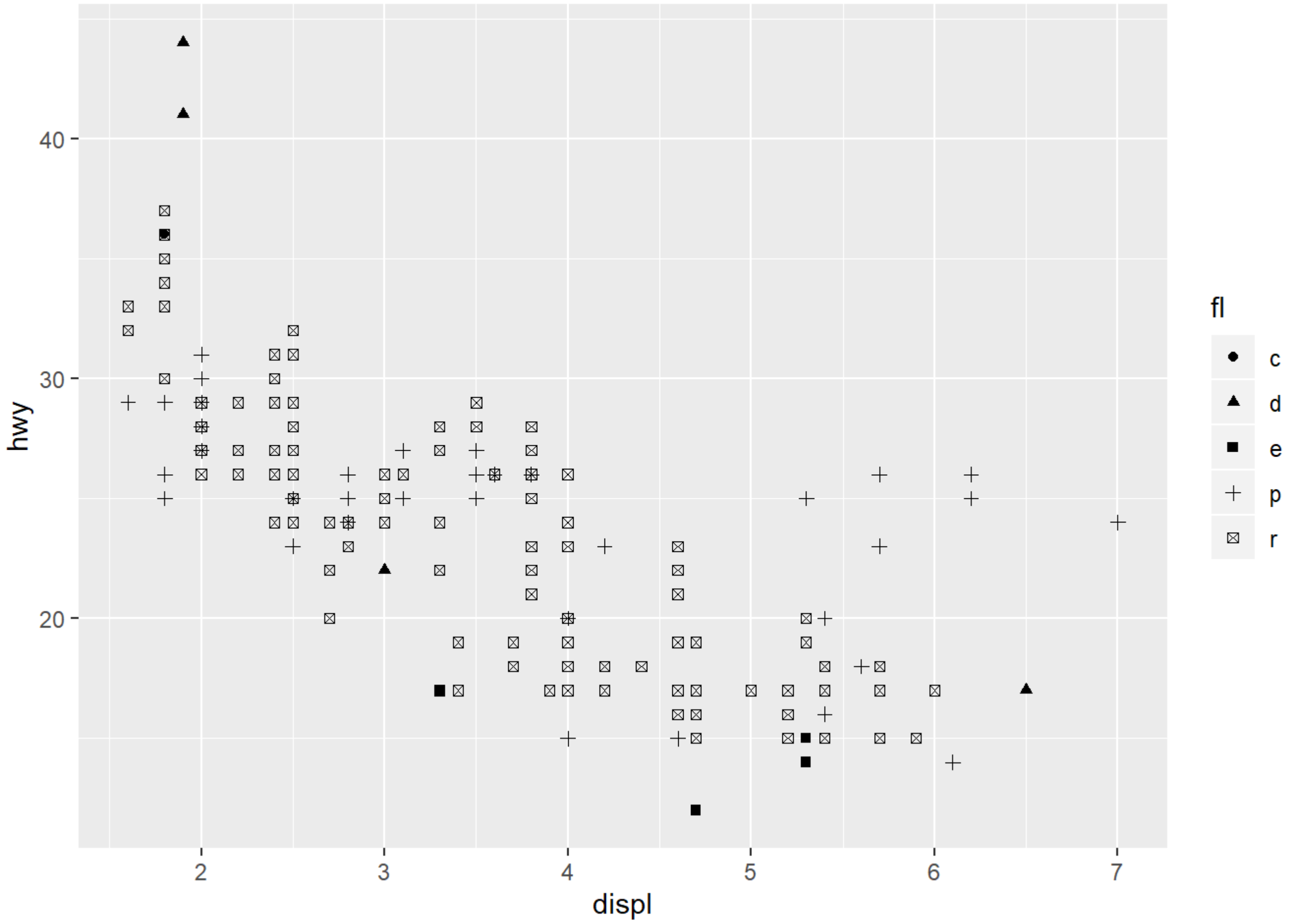
```
#ggplot(data=mpg)+geom_point(mapping = aes(x=displ,y=hwy,shape=cty)) #map cty to shape

#map categorical variables to color, size and shape
ggplot(data=mpg)+geom_point(mapping = aes(x=displ,y=hwy,col=fl)) #map fl to color
```



```
ggplot(data=mpg)+geom_point(mapping = aes(x=displ,y=hwy,size=fl)) #map fl to size
```



```
ggplot(data=mpg)+geom_point(mapping = aes(x=displ,y=hwy,shape=fl)) #map fl to shape
```



```
#aesthetic color and size works fine for both continuous and categorical variables
#aesthetic shape only works for categorical variables
```

Aesthetics color can be applicable for both categorical and continuous variables. For continuous variable, different shades of color will be mapped in continuous manner. As for categorical variable, the discreet color will be mapped in different values.

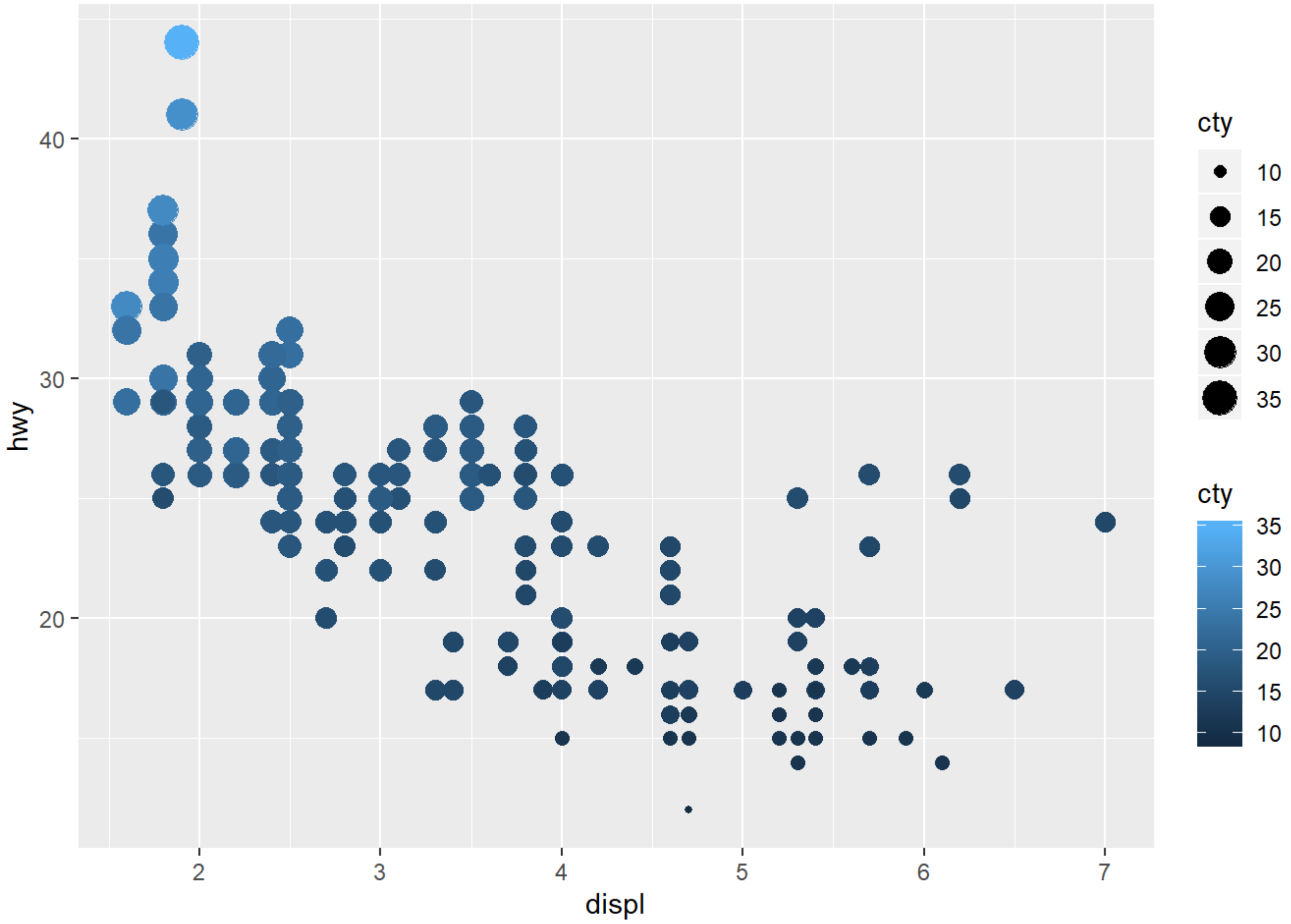
Aesthetic size looks similar for both categorical and continuous variable. Continuous variable could be more advantageous in this case because the size reflects high/low values. As for categorical variable with minimal differences, the size would be too necessary to used.

Aesthetics shape can only be used in categorical variables because it reflects different shape for discreet values.

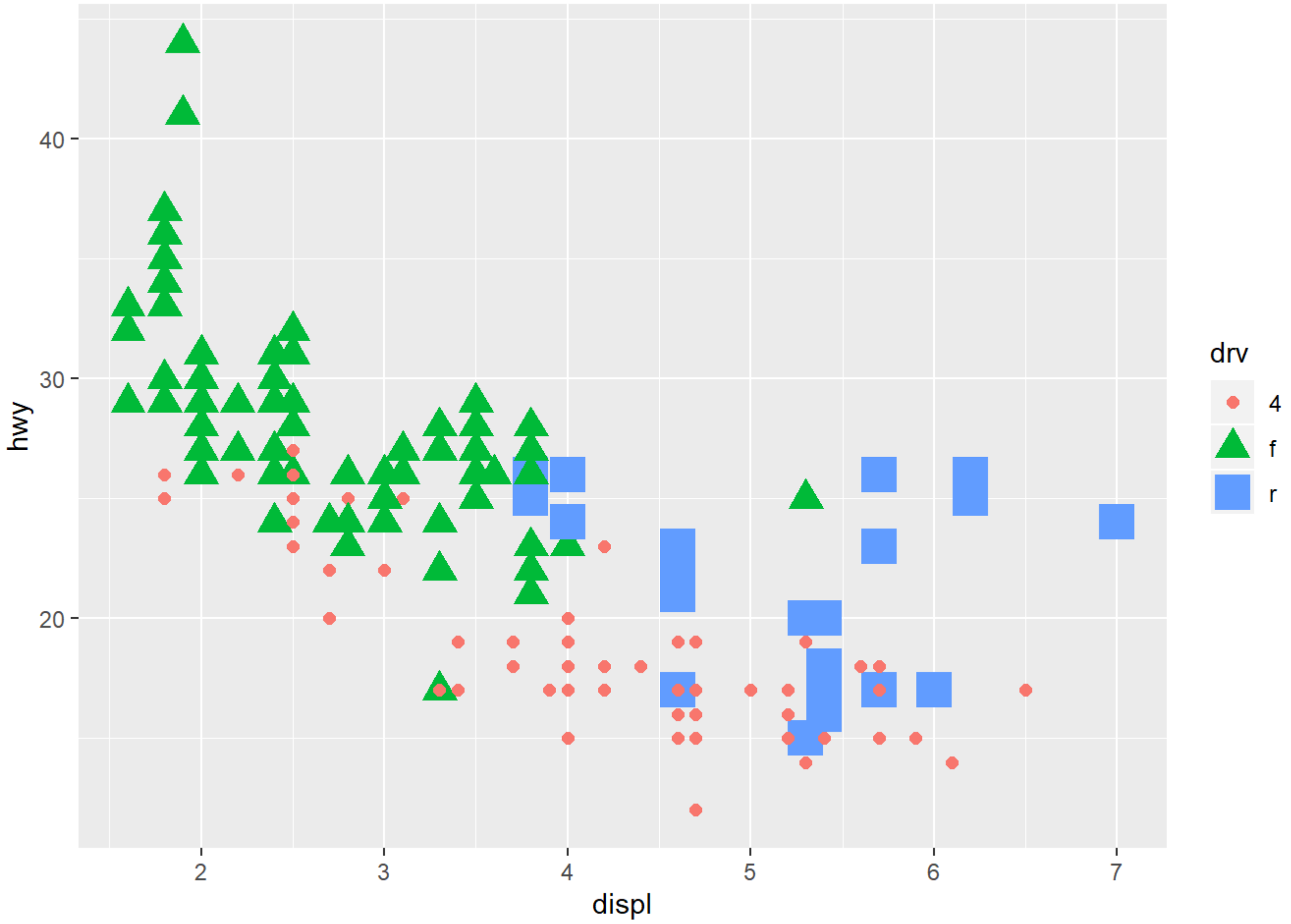
```
*3.3.1 Exercise 4*
```

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

```
#3.3.1 exercise 4
ggplot(data=mpg)+geom_point(mapping = aes(x=displ,y=hwy,size=cty,col=cty))
```



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



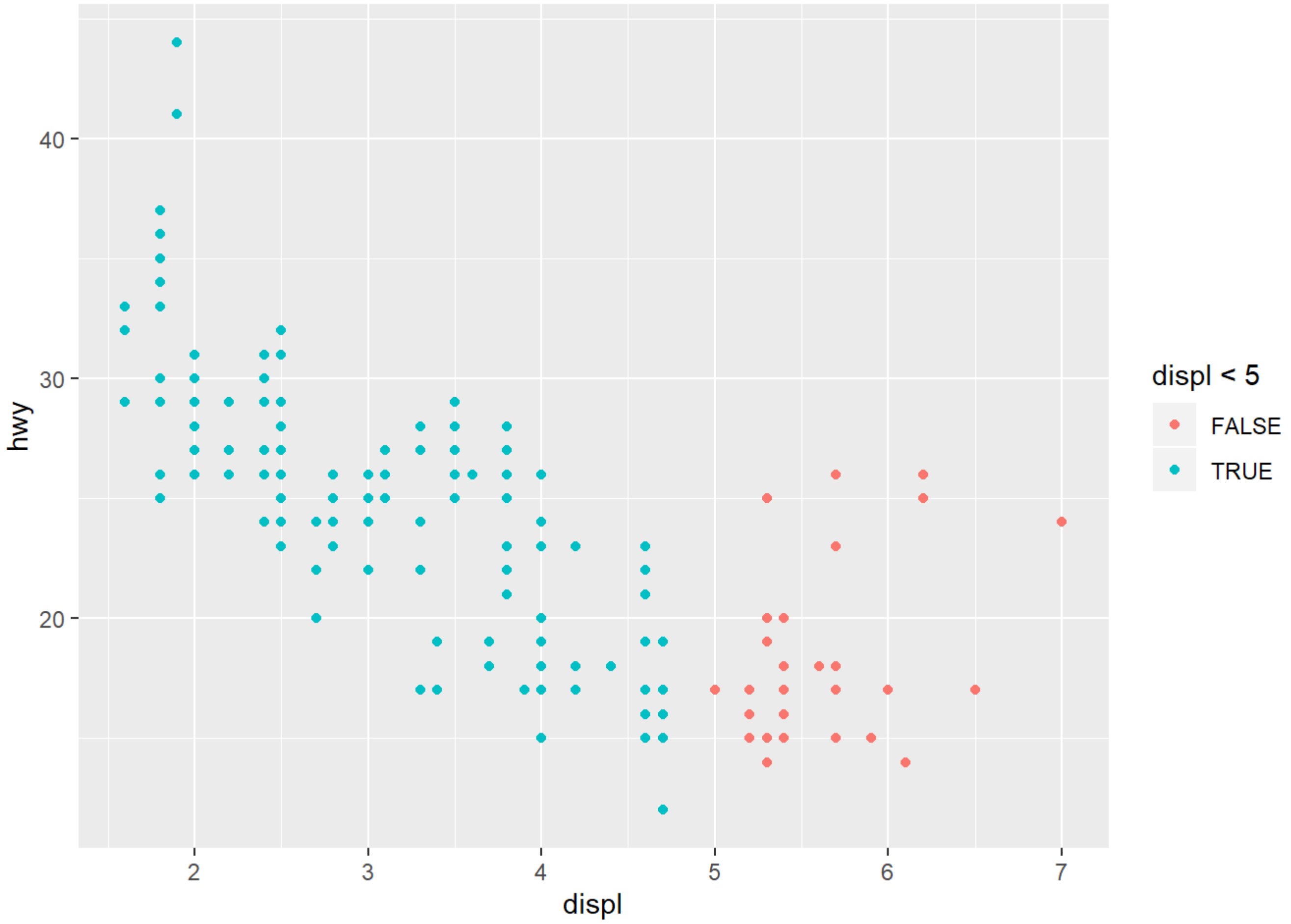
Mapping same variables to multiple-aesthetics, then the data points would be more recognized in 3rd aesthetics with continuous variables. But categorical variables, the data points would be distinguished by shape and color.

In this case, mapping same variables to multiple-aesthetics doesnt not make any different but it would help keeping information for large numbers of variables instead of black and white

```
*3.3.1 Exercise 6*
```

What happens if you map an aesthetic to something other than a variable name, like aes(colour = displ < 5)? Note, you’ll also need to specify x and y.

```
#3.3.1 exercise 6
ggplot(data=mpg)+geom_point(mapping = aes(x=displ,y=hwy,col=displ<5))
```

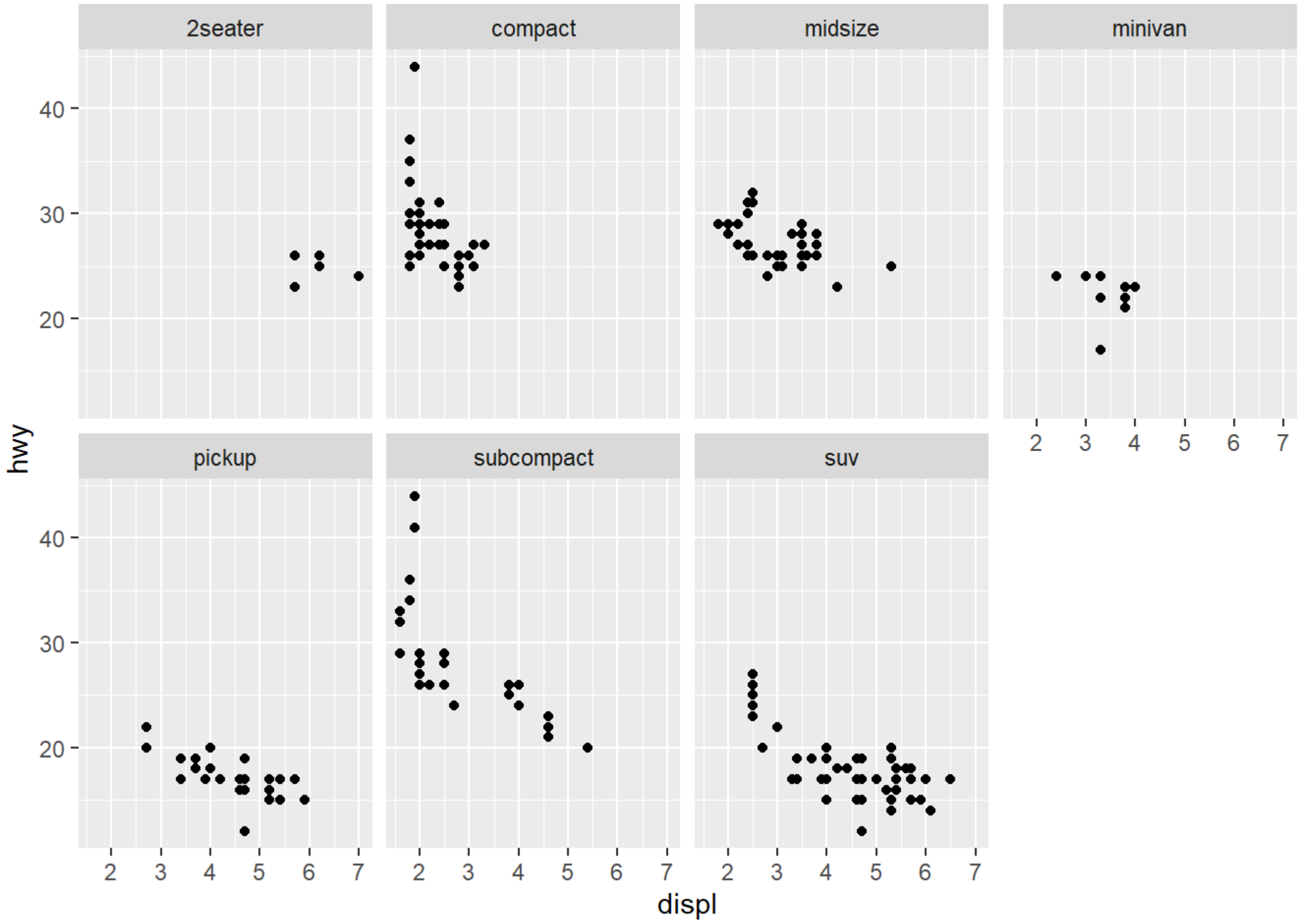


With displ<5, the legend shows the variable displ will characterized as Boolean values, TRUE and FALSE

```
*3.5.1 Exercise 6*
```

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?

```
#3.5.1 exercise 4

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



```
#the bins split to different numerical integer
```

The advantages of faceting:

- Better to investigate individual classes
- Split data into many subsets to create multiple variations in the same plot
- The aesthetic function in other layers override the defaults aesthetics for that layer only while faceting can wrap into multirow panel of plot and specify direction which split plots along rows and columns

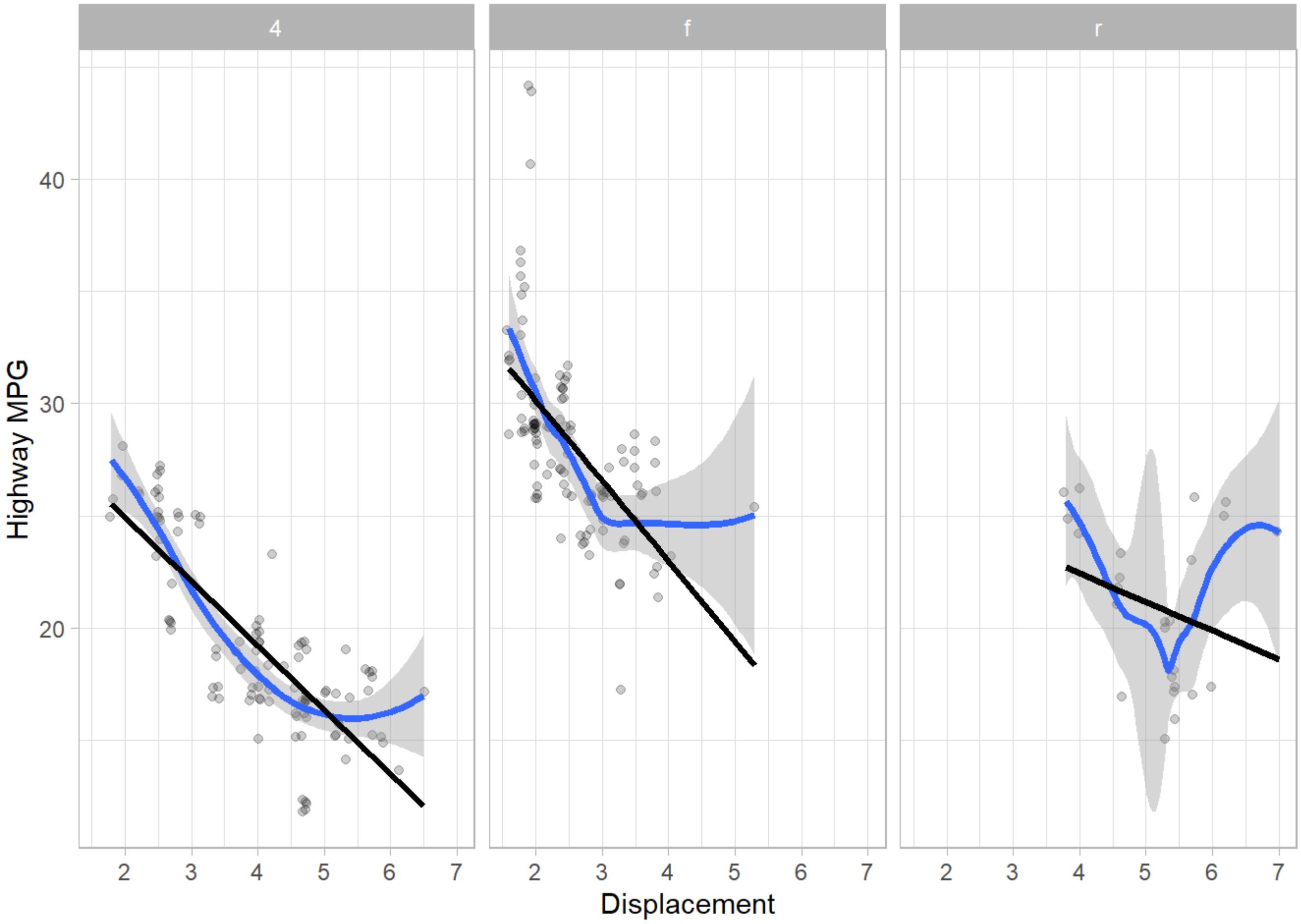
The disadvantages - aesthetic allows to using colors. With coloring, it is easier to see how the classes are clustereD

- the variables that are compared wouldn’t follow the dimension from facet

In big datasets, the coloring from aesthetics is usually better for overall clustering, but if more classes appear in dataset, faceting would be applied

b. Use ggplot to make the plot

```
#problem1b
library(lme4)
mixed<-lmer(hwy~displ+(1+displ|drv),data=mpg)
mpg$fit_mix<-predict(mixed)
ggplot(data=mpg,aes(x=displ,y=hwy))+geom_jitter(alpha=0.2)+facet_wrap(~drv)+
  geom_smooth(method="loess",size=1.3)+
  geom_line(aes(y=fit_mix),size=1.1)+
  labs(y="Highway MPG",x="Displacement")+
  theme_light()
```



Problem 2

a. Create dataframe df with 500 rows and 4 varaibles in different type of distribution

```
Then create dataframe df2 by reshaping data into GroupVar and value
```

```
#problem2a
x<-seq(-1,2,length=500) #generate random vector x
binom<-rbinom(500,5,0.2) # use random binominal distribution generate vector binom

d1<-rcauchy(500,scale=2) #generate random samples with cauchy distribution

#create dataframe df with 500 rows and each variables generate randomly from different types of
 distribution
df<-data.frame("a"=rnorm(500,mean=0,sd=10),"b"=rpois(500,lambda=4),"c"=binom,"d"=rcauchy(500,sca
le=2))
#norman distrubition, poisson distribution, binary distribution, cauchy distribution
nrow(df) #check datafram have 500 rows
```

```
## [1] 500
```

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

	groupVar <chr>	value <dbl>
1	a	-5.0460143
2	a	-0.7983234
3	a	4.6631078
4	a	6.4360393
5	a	9.7715860
6	a	10.1742503

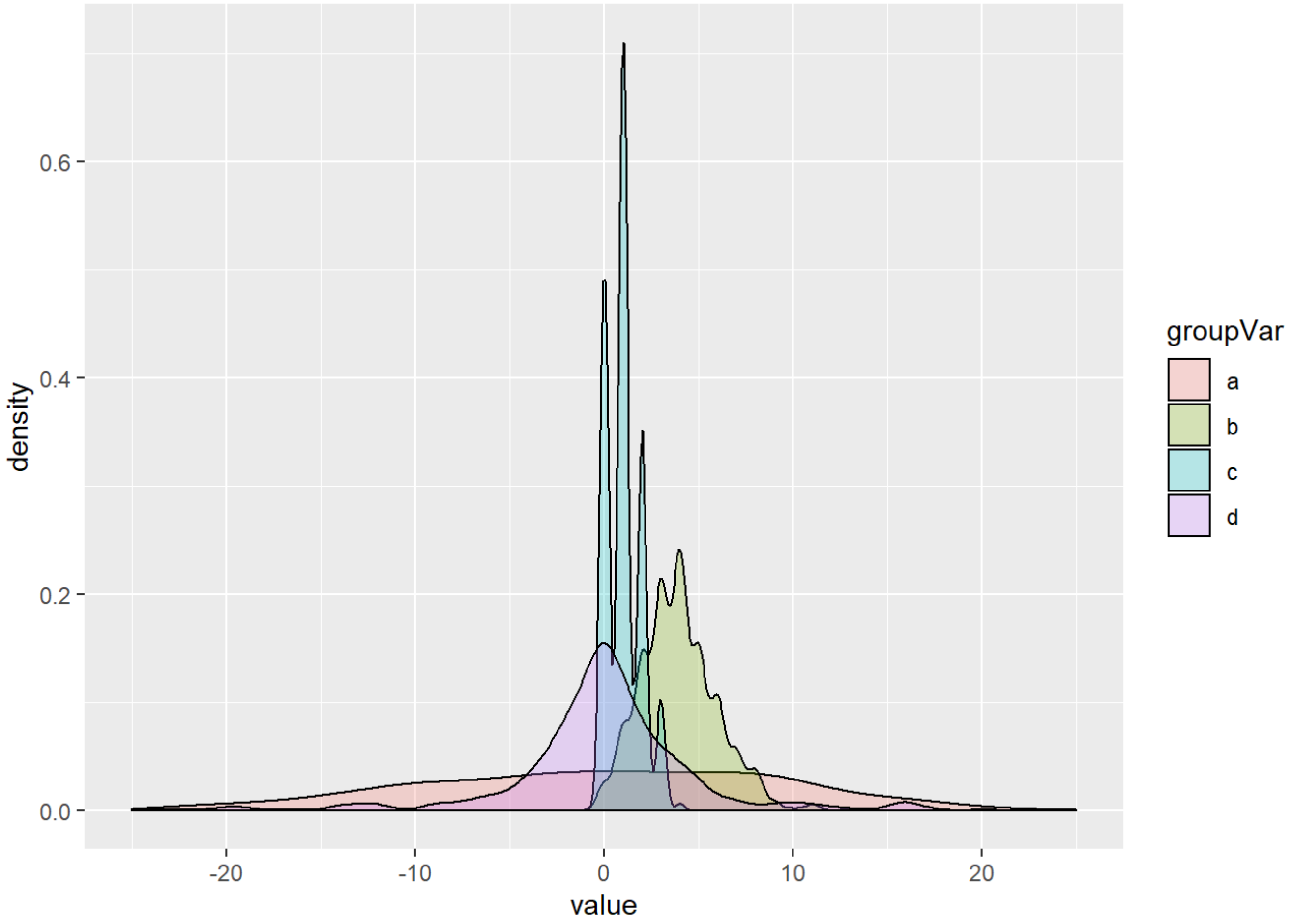
6 rows

```
nrow(df2) #check if 2000 rows
```

```
## [1] 2000
```

b.

```
#Problem2b
#draw density overlaid plot
ggplot(data=df2,aes(x=value,fill=groupVar))+geom_density(alpha=0.25)+scale_alpha(range=c(0,4,0.8
))+xlim(-25,25)
```



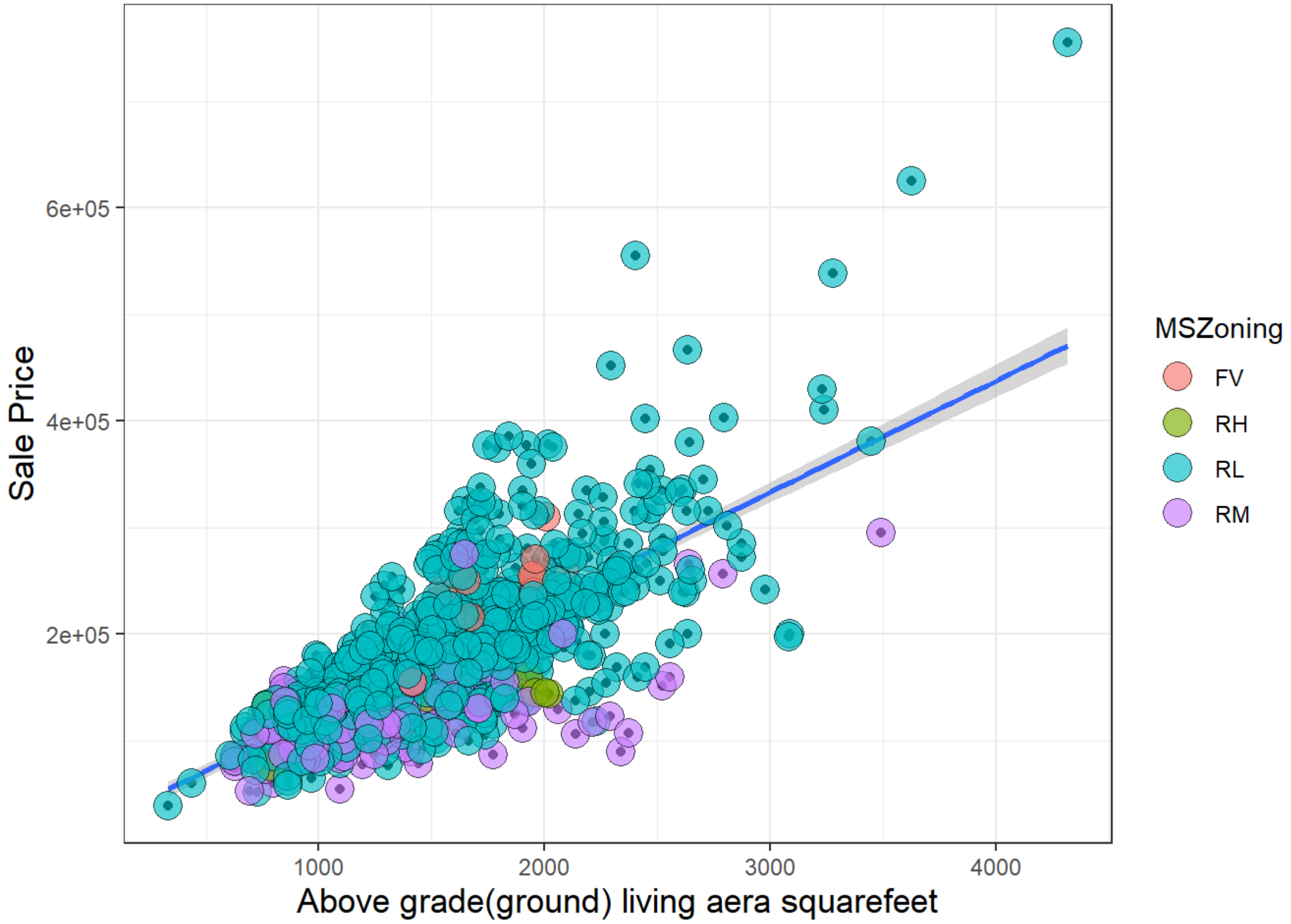
Problem 3

Create 5 visualizations from housing data

```
Load and read data
```

Visualization 1

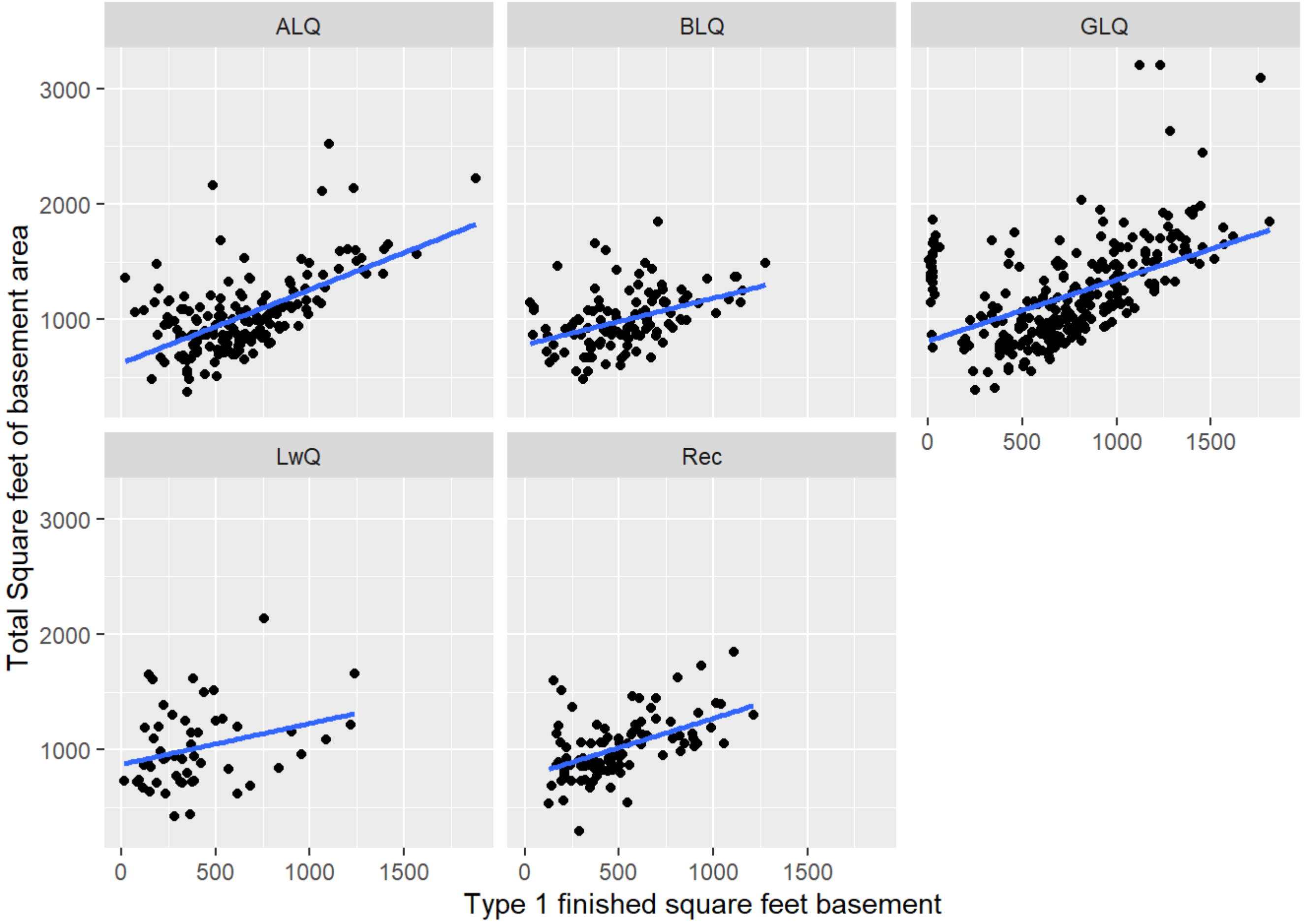
```
#Load and read data
housing<-read.csv(file="housingData.csv", header=TRUE,sep=",") #upload and read data
view(housing)
#str(housing)
#visualiation 1

#housing$SalePrice
ggplot(data=housing,aes(x=GrLivArea,y=SalePrice))+geom_point()+geom_smooth(method = "glm")+
  geom_point(aes(fill=MSZoning),
             alpha=I(.65),
             position = "jitter",
             colour="black",pch=21, size=5) +
  theme_bw() +
  labs(y = "Sale Price",
       x = "Above grade(ground) living aera squarefeet") +
  theme(legend.key=element_blank(),
        axis.title = element_text(size = 13))
```



The visual shows the generalized linear smooth between sale price and above grade(ground) ling area square feet. The studied variables are set into different groups of MSZoning. The linear model is utilized to understand the correlation between sale price and above grade(ground) living area square feet. Note that most of the zoning classification of the sale lies on residential low density zone.

Visualization 2

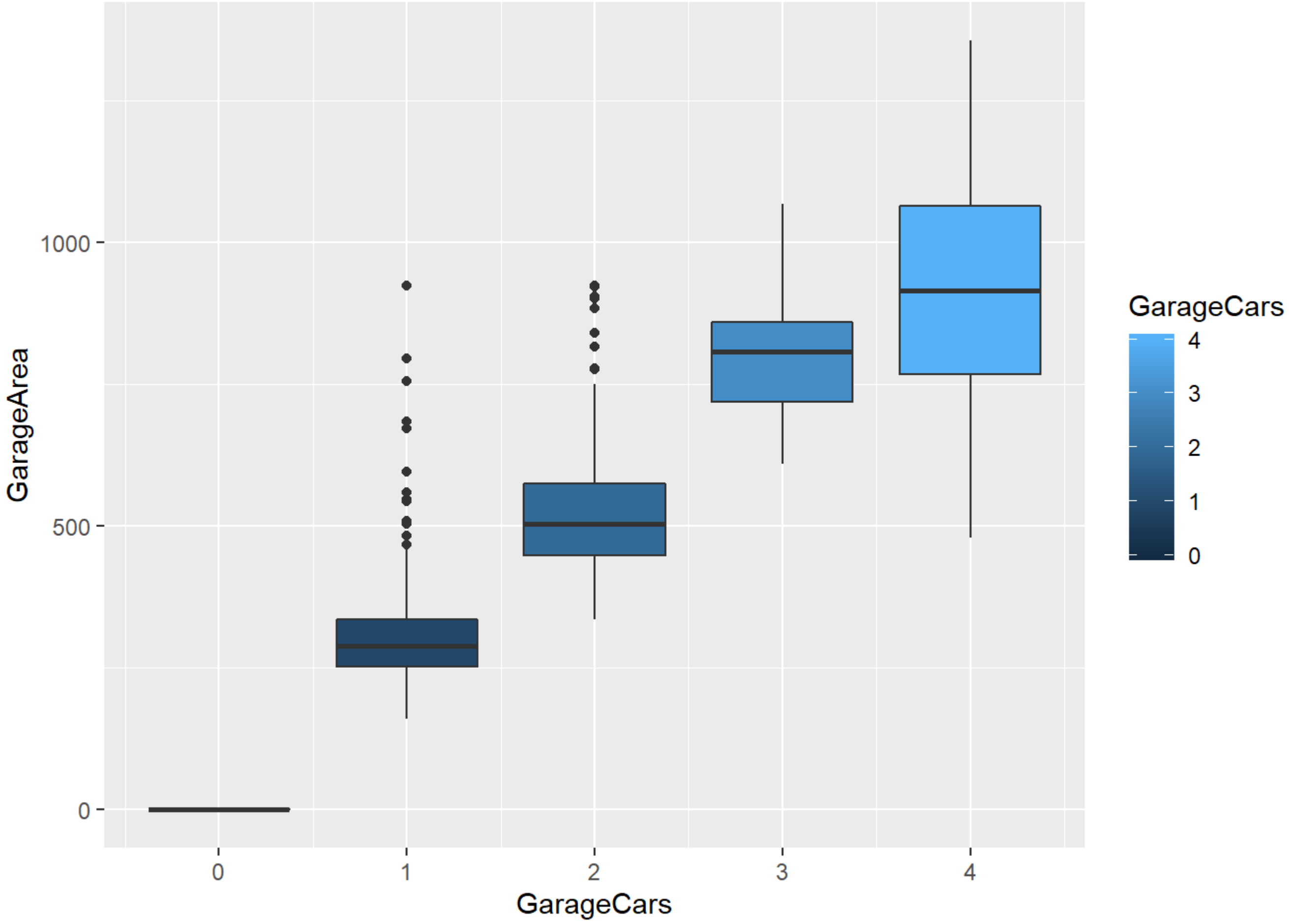
```
#visualization 2
housing2<-housing[!(housing$BsmtFinSF1==0),] #remove zero square foot basement type 1
ggplot(data=housing2,aes(x=BsmtFinSF1,y=TotalBsmtSF))+geom_point()+facet_wrap(~BsmtFinType1)+
  geom_smooth(method="lm",se=FALSE,span=1.5)+
  labs(x="Type 1 finished square feet basement",y="Total Square feet of basement area")+scale_fi
ll_discrete( labels = c( "Average Living Quarters ", "Below Average Living Quarters","Good Livin
g Quarters ","Low Quality","Average Rec Room","Unf"))
```



This graph shows the correlation between type 1 finished square feet basement and total square feet of basement area, divided in different groups of rating of basement finished area type 1. The linear correlation between total square feet of basement area with type 1 square feet of basement. It's also noticeable that most ratings for basement type 1 lies on good living quarters.

Visualization 3

```
ggplot(data=housing,aes(x=GarageCars,y=GarageArea))+geom_boxplot(aes(group=GarageCars,fill=Garag
eCars))
```

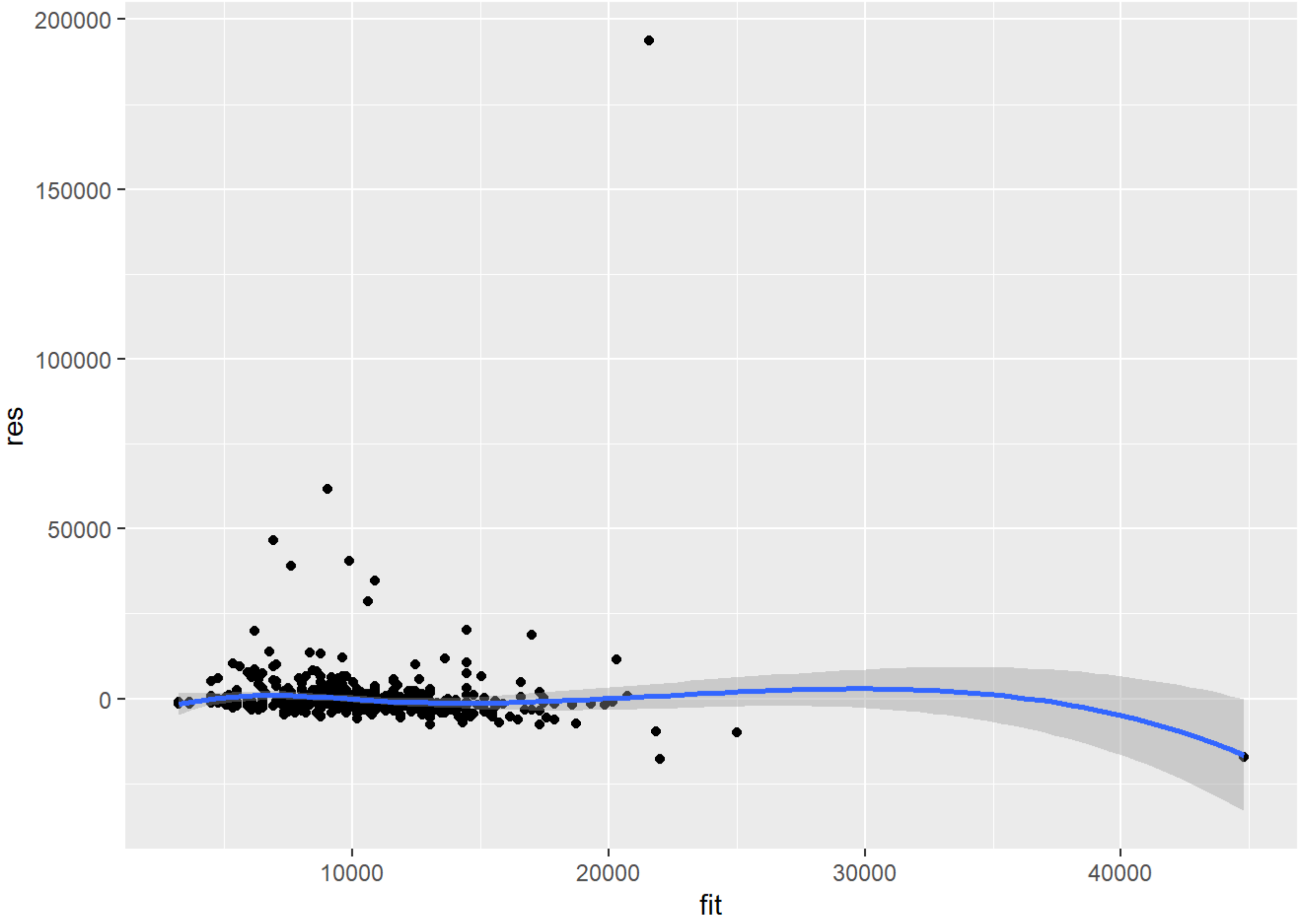


The box plot shows the classification of garage area based on different numbers of cars. The average garage area increases as the numbers of car increases. Most of the garage having more than 2 cars usually have the area capacity larger than 500 square feet.

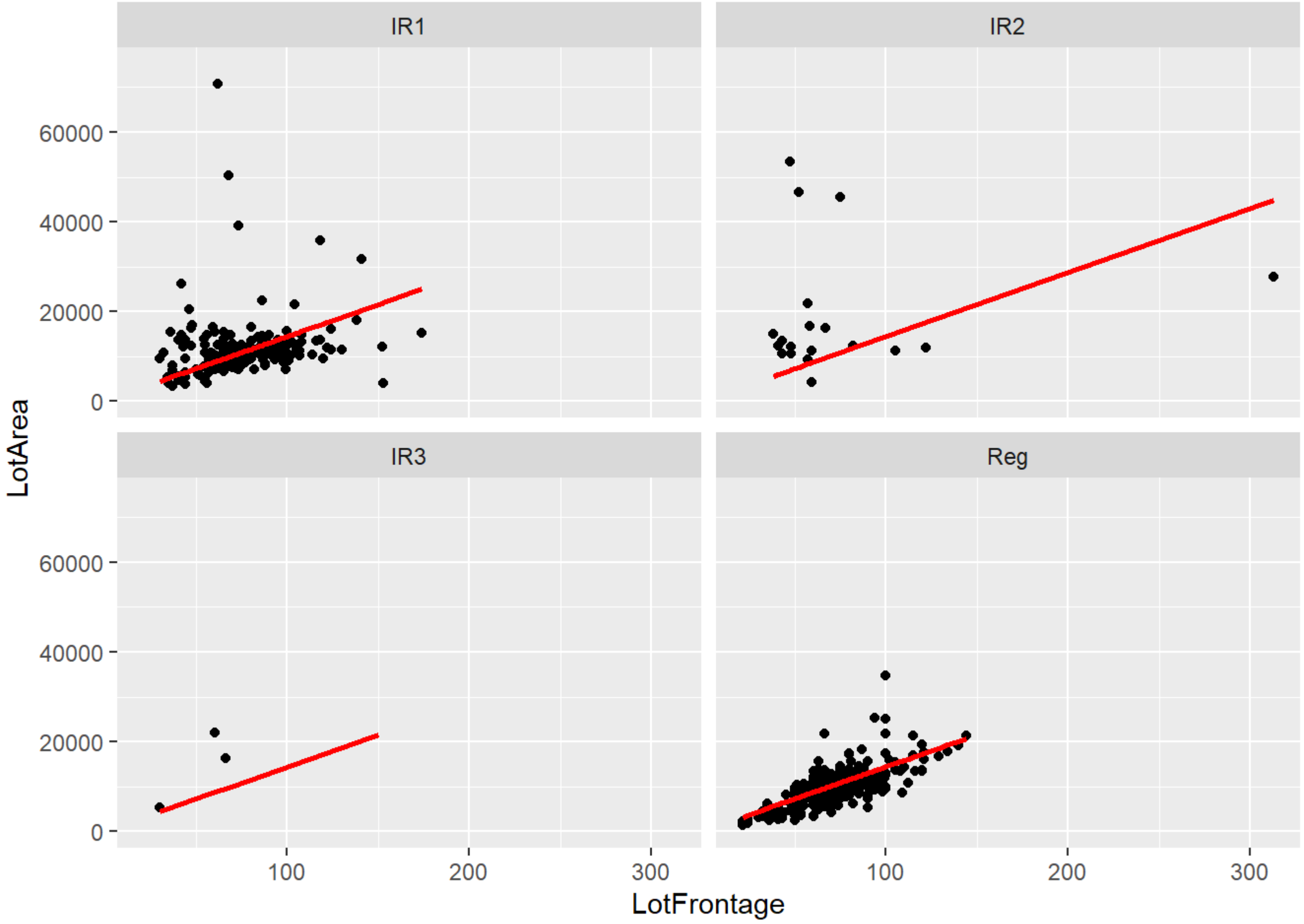
Visualization 4

```
housing4<-housing[!(is.na(housing$LotFrontage)),]
linModel<-lm(data=housing4,LotArea~LotFrontage)
housing4$res <- residuals(linModel)
housing4$fit <- predict(linModel)
ggplot(housing4, aes(x= fit, y=res)) + geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



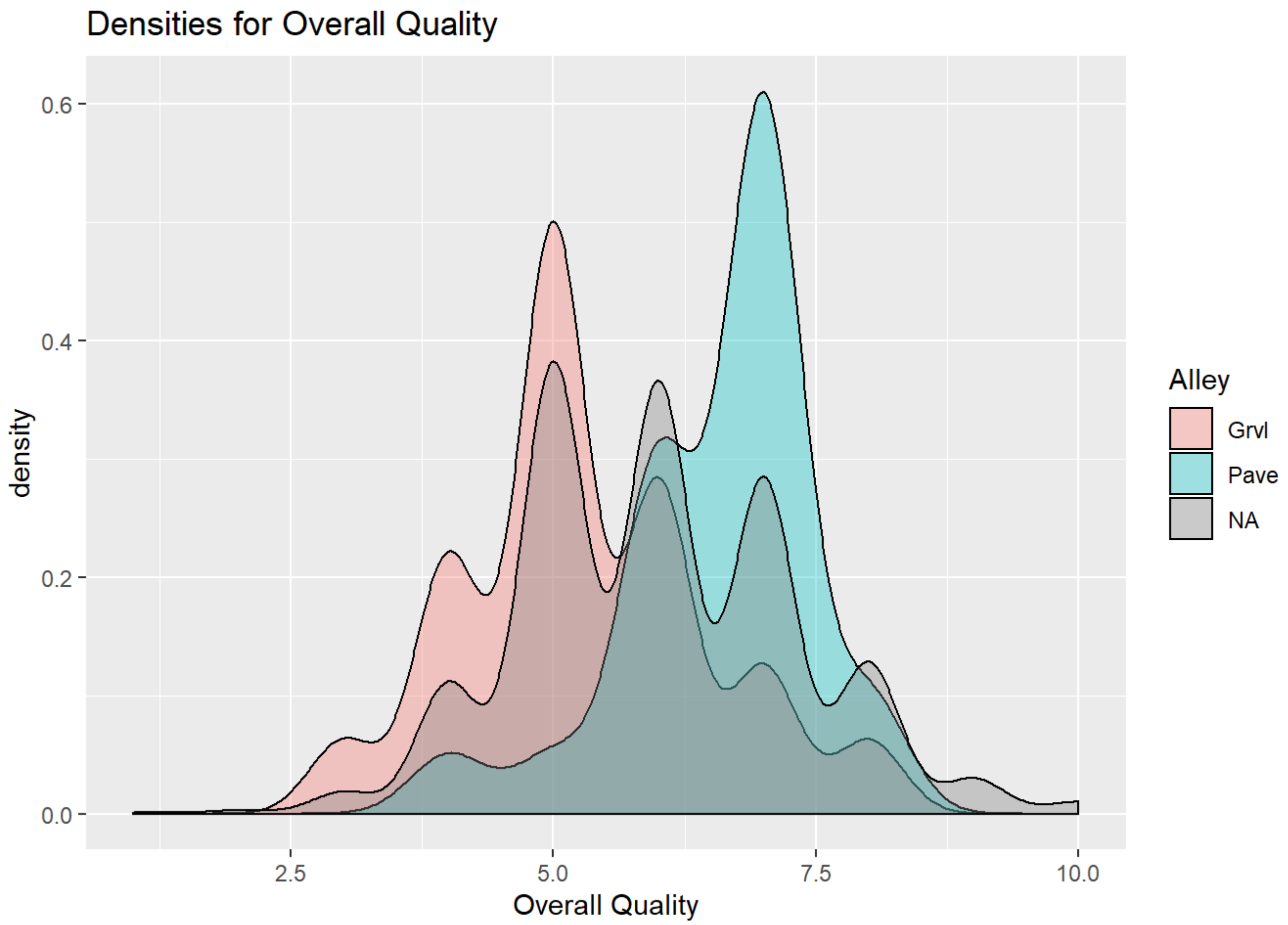
```
ggplot(data=housing4,aes(x=LotFrontage,y=LotArea))+geom_point()+geom_line(aes(y=fit),size=1,color="red")+facet_wrap(~LotShape)+ylim(0,75000)
```



The graph reveals the linear fit regression between area lot size and linear feet of street connected to the property. The studied variables are divided into different property shape groups. The linear fit model is utilized to understand the correlation between lot size and street length. Note that most lot were built on regular and slightly regular group

Visualization 5

```
ggplot(housing, aes(x=OverallQual, fill=Alley)) +
  geom_density(alpha=0.35) + labs(x = "Overall Quality",title = "Densities for Overall Quality")
```



The density plot reveals the distribution of overall quality in different categories of paved driveway. Generally, the paved alley has better overall quality score than gravel alley due to the paved alley have the score distribution more skewed to the right side

Problem 4

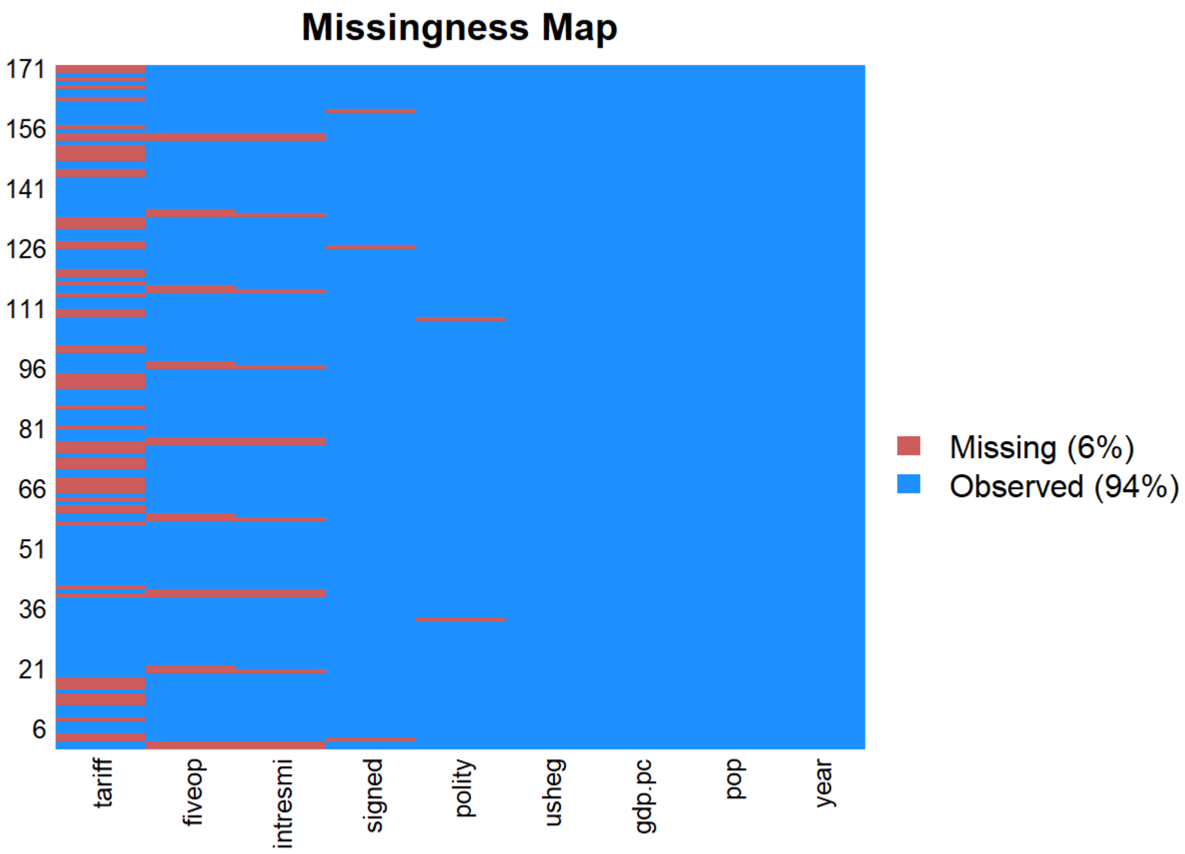
4.a Explore misisingness

```
Amelia packages
```

```
#problem4a: explore missingness
data(freetrade) #loaddata
summary(freetrade) #explore data
```

```
##       year        country              tariff          polity     
##  Min.   :1981   Length:171         Min.   :  7.10   Min.   :-8.000  
##  1st Qu.:1985   Class :character   1st Qu.: 16.30   1st Qu.:-2.000  
##  Median :1990   Mode  :character   Median : 25.20   Median : 5.000  
##  Mean   :1990                      Mean   : 31.65   Mean   : 2.905  
##  3rd Qu.:1995                      3rd Qu.: 40.80   3rd Qu.: 8.000  
##  Max.   :1999                      Max.   :100.00   Max.   : 9.000  
##                                    NA's   :58       NA's   :2       
##       pop                gdp.pc          intresmi          signed      
##  Min.   : 14105080   Min.   :  149.5   Min.   :0.9036   Min.   :0.0000  
##  1st Qu.: 19676715   1st Qu.:  420.1   1st Qu.:2.2231   1st Qu.:0.0000  
##  Median : 52799040   Median :  814.3   Median :3.1815   Median :0.0000  
##  Mean   :149904501   Mean   : 1867.3   Mean   :3.3752   Mean   :0.1548  
##  3rd Qu.:120888400   3rd Qu.: 2462.9   3rd Qu.:4.4063   3rd Qu.:0.0000  
##  Max.   :997515200   Max.   :12086.2   Max.   :7.9346   Max.   :1.0000  
##                                        NA's   :13       NA's   :3       
##      fiveop          usheg       
##  Min.   :12.30   Min.   :0.2558  
##  1st Qu.:12.50   1st Qu.:0.2623  
##  Median :12.60   Median :0.2756  
##  Mean   :12.74   Mean   :0.2764  
##  3rd Qu.:13.20   3rd Qu.:0.2887  
##  Max.   :13.20   Max.   :0.3083  
##  NA's   :18
```

```
#summary(lm(tariff ~ polity + pop + gdp.pc + year + country,data = freetrade)) #explore data
#remove the country because the non-numerical value could affect the validity of Amelia resulst
f_boottrap<-freetrade[,-2]
missmap(f_boottrap) #from Amelia packages
```



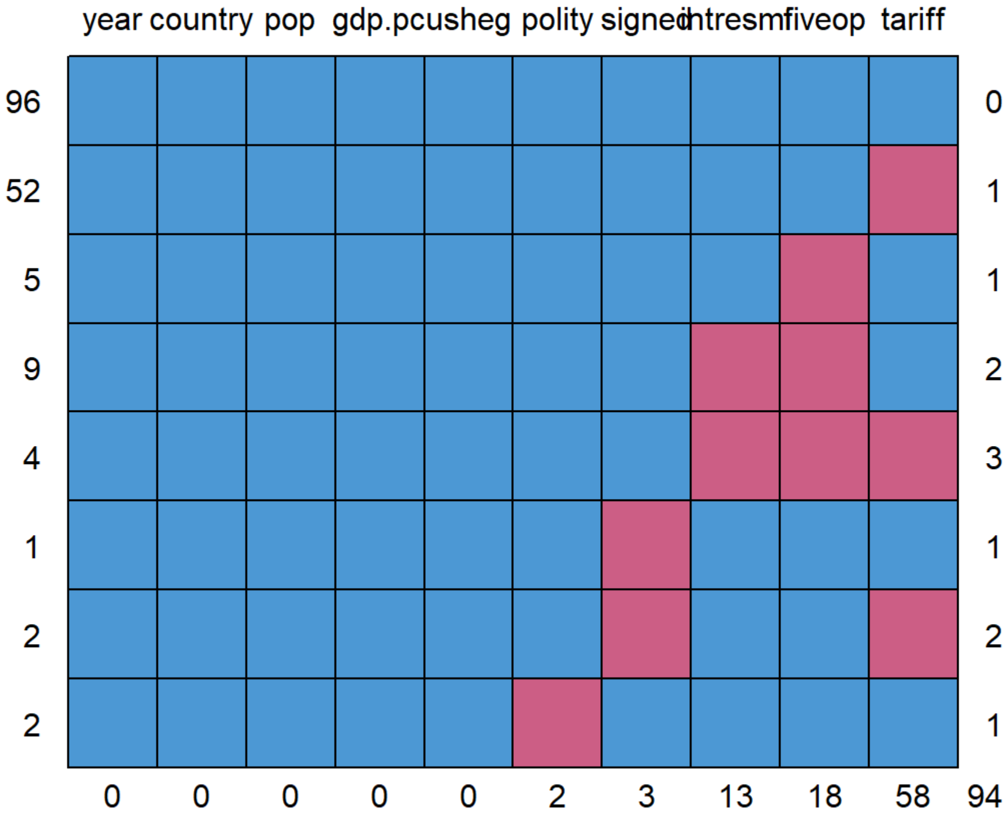
The data is bootstrapped to remove the variable country. This happens because Amelia package works well with numeric characters. The presence of non-numeric variable affect the results The missingness map from Amelia package reveal missing value in tan color and observed value in blue. The tariff rate is the most missing variables. Variables, (intresmi) and (fivop), are missing mostly of each cross section

```
mice packages
```

```
#problem4a: explore missingness mice packages
md.pairs(freetrade)
```

```
## $rr
##          year country tariff polity pop gdp.pc intresmi signed fiveop
## year      171     171    113    169 171    171      158    168    153
## country   171     171    113    169 171    171      158    168    153
## tariff    113     113    113    111 113    113      104    112     99
## polity    169     169    111    169 169    169      156    166    151
## pop       171     171    113    169 171    171      158    168    153
## gdp.pc    171     171    113    169 171    171      158    168    153
## intresmi  158     158    104    156 158    158      158    155    153
## signed    168     168    112    166 168    168      155    168    150
## fiveop    153     153     99    151 153    153      153    150    153
## usheg     171     171    113    169 171    171      158    168    153
##          usheg
## year       171
## country    171
## tariff     113
## polity     169
## pop        171
## gdp.pc     171
## intresmi   158
## signed     168
## fiveop     153
## usheg      171
## 
## $rm
##          year country tariff polity pop gdp.pc intresmi signed fiveop
## year        0       0     58      2   0      0       13      3     18
## country     0       0     58      2   0      0       13      3     18
## tariff      0       0      0      2   0      0        9      1     14
## polity      0       0     58      0   0      0       13      3     18
## pop         0       0     58      2   0      0       13      3     18
## gdp.pc      0       0     58      2   0      0       13      3     18
## intresmi    0       0     54      2   0      0        0      3      5
## signed      0       0     56      2   0      0       13      0     18
## fiveop      0       0     54      2   0      0        0      3      0
## usheg       0       0     58      2   0      0       13      3     18
##          usheg
## year         0
## country      0
## tariff       0
## polity       0
## pop          0
## gdp.pc       0
## intresmi     0
## signed       0
## fiveop       0
## usheg        0
## 
## $mr
##          year country tariff polity pop gdp.pc intresmi signed fiveop
## year        0       0      0      0   0      0        0      0      0
## country     0       0      0      0   0      0        0      0      0
## tariff     58      58      0     58  58     58       54     56     54
## polity      2       2      2      0   2      2        2      2      2
## pop         0       0      0      0   0      0        0      0      0
## gdp.pc      0       0      0      0   0      0        0      0      0
## intresmi   13      13      9     13  13     13        0     13      0
## signed      3       3      1      3   3      3        3      0      3
## fiveop     18      18     14     18  18     18        5     18      0
## usheg       0       0      0      0   0      0        0      0      0
##          usheg
## year         0
## country      0
## tariff      58
## polity       2
## pop          0
## gdp.pc       0
## intresmi    13
## signed       3
## fiveop      18
## usheg        0
## 
## $mm
##          year country tariff polity pop gdp.pc intresmi signed fiveop
## year        0       0      0      0   0      0        0      0      0
## country     0       0      0      0   0      0        0      0      0
## tariff      0       0     58      0   0      0        4      2      4
## polity      0       0      0      2   0      0        0      0      0
## pop         0       0      0      0   0      0        0      0      0
## gdp.pc      0       0      0      0   0      0        0      0      0
## intresmi    0       0      4      0   0      0       13      0     13
## signed      0       0      2      0   0      0        0      3      0
## fiveop      0       0      4      0   0      0       13      0     18
## usheg       0       0      0      0   0      0        0      0      0
##          usheg
## year         0
## country      0
## tariff       0
## polity       0
## pop          0
## gdp.pc       0
## intresmi     0
## signed       0
## fiveop       0
## usheg        0
```

```
md.pattern(freetrade)
```



```
##    year country pop gdp.pc usheg polity signed intresmi fiveop tariff
## 96    1       1   1      1     1      1      1        1      1      1  0
## 52    1       1   1      1     1      1      1        1      1      0  1
## 5     1       1   1      1     1      1      1        1      0      1  1
## 9     1       1   1      1     1      1      1        0      0      1  2
## 4     1       1   1      1     1      1      1        0      0      0  3
## 1     1       1   1      1     1      1      0        1      1      1  1
## 2     1       1   1      1     1      1      0        1      1      0  2
## 2     1       1   1      1     1      0      1        1      1      1  1
##       0       0   0      0     0      2      3       13     18     58 94
```

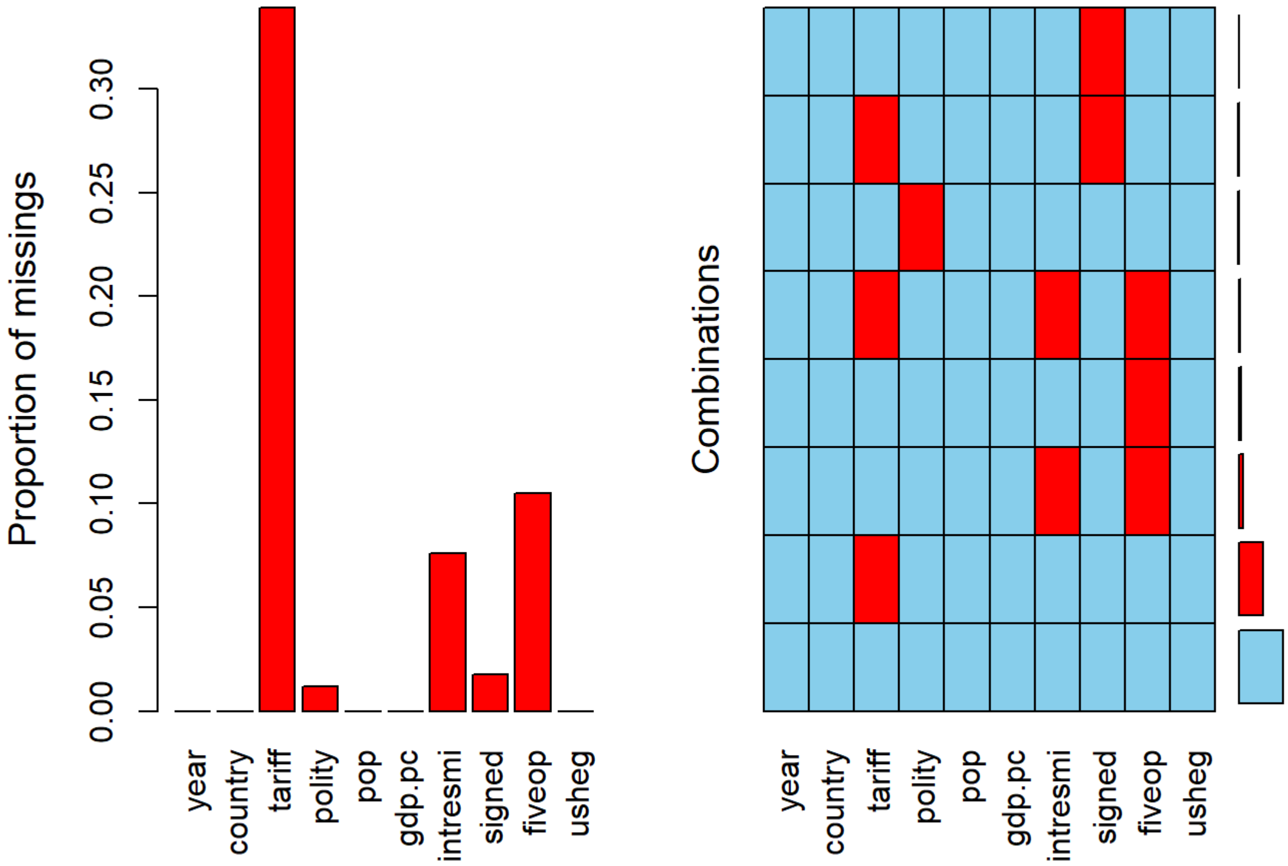
mice packages gives the pairs summary rr,rm,mr,mm (remain,missing) among the pairs of varaibles

mice packages gives the information for missingness of variables in respect to the other variables in the form of matrix

58 missing values for tariff variables, 18 missing values for fiveop variable, 13 missing values for intersmi variables

```
aggr frunction
```

```
# Use "aggr" function to also get overall information on missing

f<-aggr(freetrade)
```


Proportion of missings
0.00
0.05
0.10
0.15
0.20
0.25
0.30
year
country
tariff
polity
pop
gdp.pc
intresmi
signed
fiveop
usheg
Combinations
year
country
tariff
polity
pop
gdp.pc
intresmi
signed
fiveop
usheg

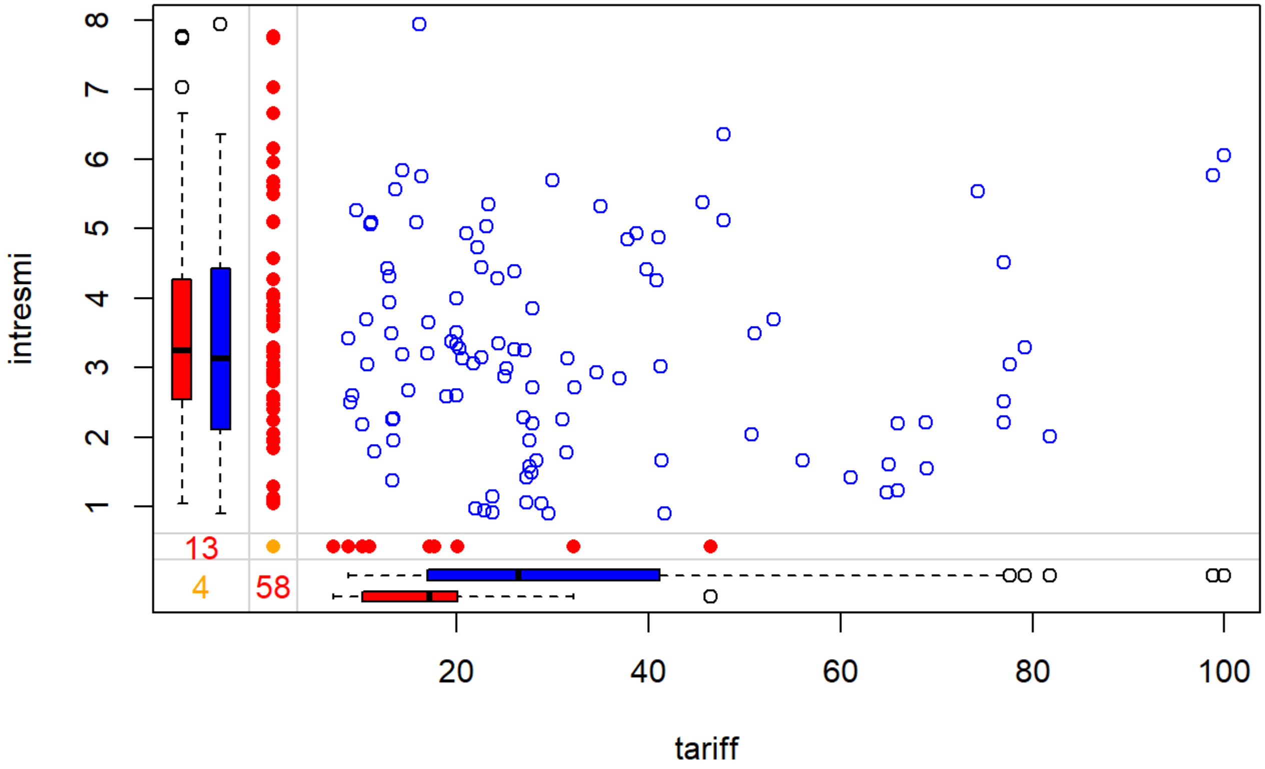
```
summary(f)
```

```
##
##  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
```

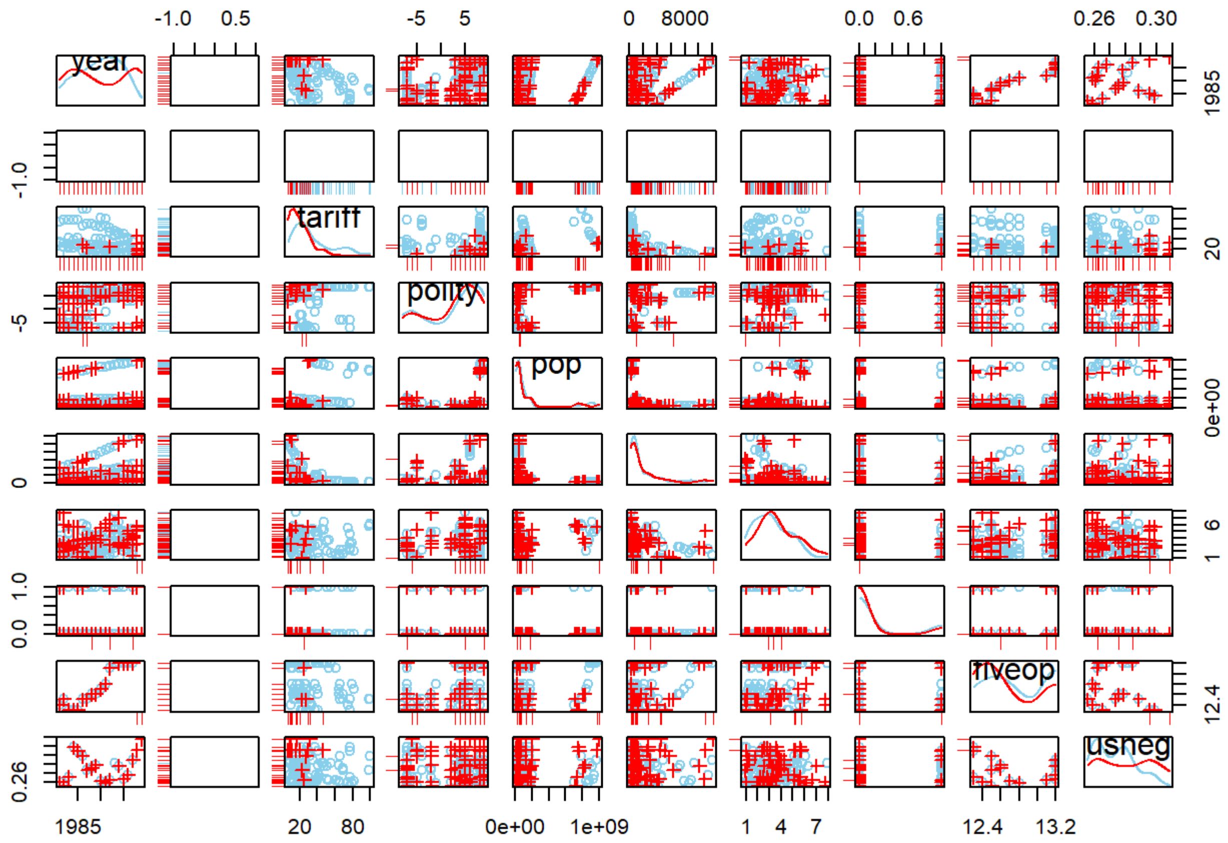
Use aggr funtion to get the overall information of missingness as in proportions of missing per variables and the combinations among variables as in the matrix. Missings is indicates in the red colors

Summary function gives the counting of missingness per varaibles and alos the percentages

```
# use VIM function "marginplot" to get a scatter plot that includes information on missing value
s
marginplot(freetrade[c("tariff","intresmi")], col = c("blue", "red", "orange"))
```



```
# all plots have Missing Information
scattmatrixMiss(freetrade)
```



Scatteplot include all missing information on missing values

The margin plot show the scatter plot containing all information and distribution of missing values. Red color box plot shows distribution of missing values in respect to the adjacent axis. Plots have all missing information but there are to many variables to observed

4.b Using Chi-square test

The Chi-square test is applied for analyzing discreet variables. The objective is to reveal whether the variables are dependent or independent

The column tariff and country is extracted to different table. Any misisng value in tarriff set to 0 (NA ->0) while the others set as 1. The Chi-square test on the new table

```
View(freetrade)
f_4b<-freetrade[,c(2,3)] #extract data to tariff and country
f_4b[is.na(f_4b$tariff),]$tariff<-0 # set NA values to zero
f_4b[!(f_4b$tariff==0),]$tariff<-1 # set non-missing values to 1
#create table and Chisq-test
table_f4b<-table(f_4b$tariff,f_4b$country)
chisq.test(table_f4b)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_f4b
## X-squared = 23.064, df = 8, p-value = 0.003283
```

the degree of freedom is 8 the p-value is 0.003283, which is less than 0.05. Then we reject hypothesis null, 2 varaibles are dependent

Remove Philippines

```
#remove Phillipines
f_4b1<-freetrade[,c(2,3)] #extract data to tariff and country
f_rm_phil<-f_4b1[(f_4b1$country!="Philippines"),]
country_numeric<-unclass(as.factor(f_rm_phil$country))
f_rm_phil[is.na(f_rm_phil$tariff),]$tariff<-0 # set NA values to zero
f_rm_phil[!(f_rm_phil$tariff==0),]$tariff<-1 # set non-missing values to 1
#create table and Chisq-test
table_f_rmphil<-table(f_rm_phil$tariff,country_numeric)
chisq.test(table_f_rmphil)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_f_rmphil
## X-squared = 11.486, df = 7, p-value = 0.1188
```

the degree of freedom is 7, x square decreases

the p-value is 0.1188, which is more than 0.1. 90% confidence. Then we accepts hypothesis null, 2 varaibles are independent

On the original table, variable Philippines does not contain any missing values in tariff. Therefore, when removing Philippines, all other variables contain the missing tarrif values

Removing Nepal

```
#remove Nepal
f_4b2<-freetrade[,c(2,3)] #extract data to tariff and country
f_rm_nep<-f_4b2[(f_4b2$country!="Nepal"),]
country_numeric<-unclass(as.factor(f_rm_nep$country))
f_rm_nep[is.na(f_rm_nep$tariff),]$tariff<-0 # set NA values to zero
f_rm_nep[!(f_rm_nep$tariff==0),]$tariff<-1 # set non-missing values to 1
#create table and Chisq-test
table_f_rmnep<-table(f_rm_nep$tariff,country_numeric)
chisq.test(table_f_rmnep)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_f_rmnep
## X-squared = 15.836, df = 7, p-value = 0.02666
```

the degree of freedom is 7, x square increases the p-value is 0.02666, which is lower than 95% confidence. Then we rejects hypothesis null, 2 varaibles are independent