Homework #2

Azadeh Gilanpour

September 12, 2018

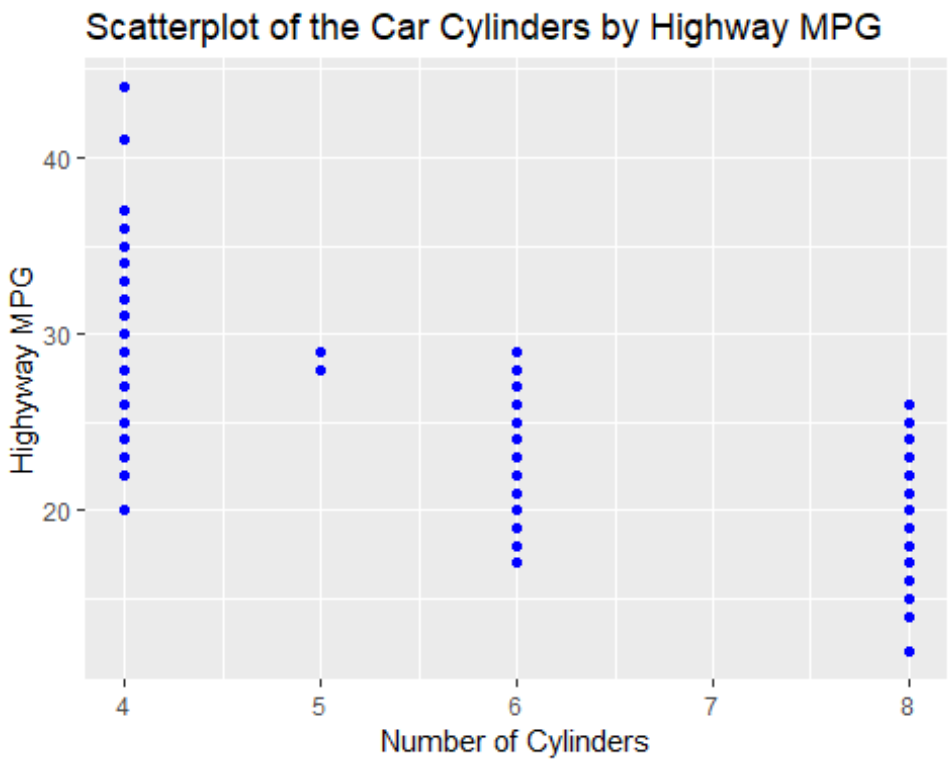
Problem 1: Learning ggplot2

(a)

- 3.2.4 Exercies 4:

Make a scatterplot of hwy vs cyl

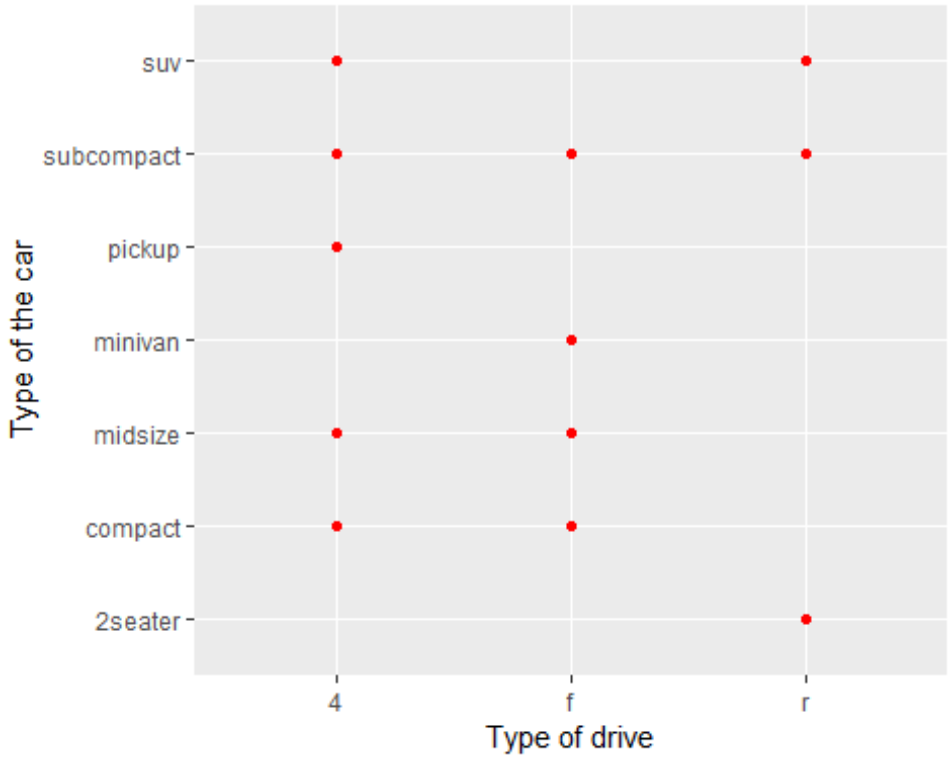
```
ggplot(data=mpg,mapping=aes(x=cyl,y=hwy))+
  geom_point(color="blue")+
  labs(x="Number of Cylinders",y="Highway MPG",title="Scatterplot of the Car Cylinders by Highway MPG")
```



- The plot shows a negative relationship between the number of cylinder (cyl) and fuel efficiency (hwy). That menas the car has a biger cylinder use more fuel.
- 3.2.4 Exercies 5:

Make a scatterplot of class vs drv

```
ggplot(data = mpg, mapping = aes(x=drv,y=class))+
  geom_point(color="red")+
  labs(x="Type of drive",y="Type of the car")
```



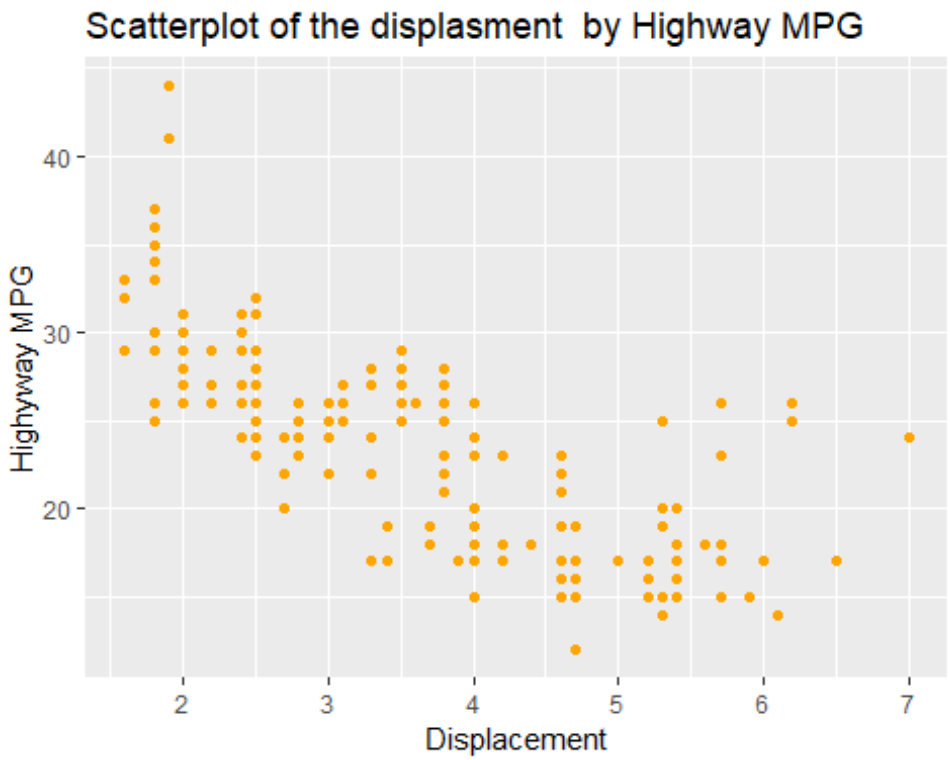
- This sactterplot is uesles beacuse comapre two categorical attribute of “drv” and “class” with eachother. Each points in the plot has overlap with the other.

(b)

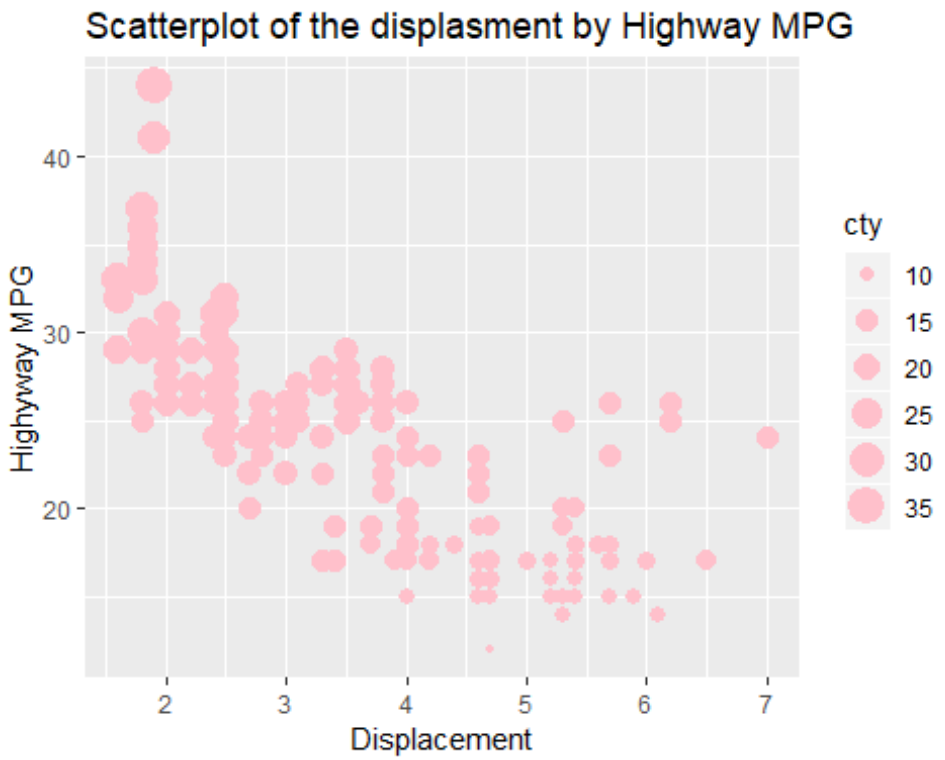
- 3.3.1 Exercies 3:

Map continoues variable to Color/Size/Shap

```
#continous variable
ggplot(data=mpg)+
  geom_point(mapping = aes(x=displ,y=hwy,color= cyl),color='orange')+
  labs(x="Displacement",y="Highway MPG",title="Scatterplot of the displasment  by Highway MPG")
```

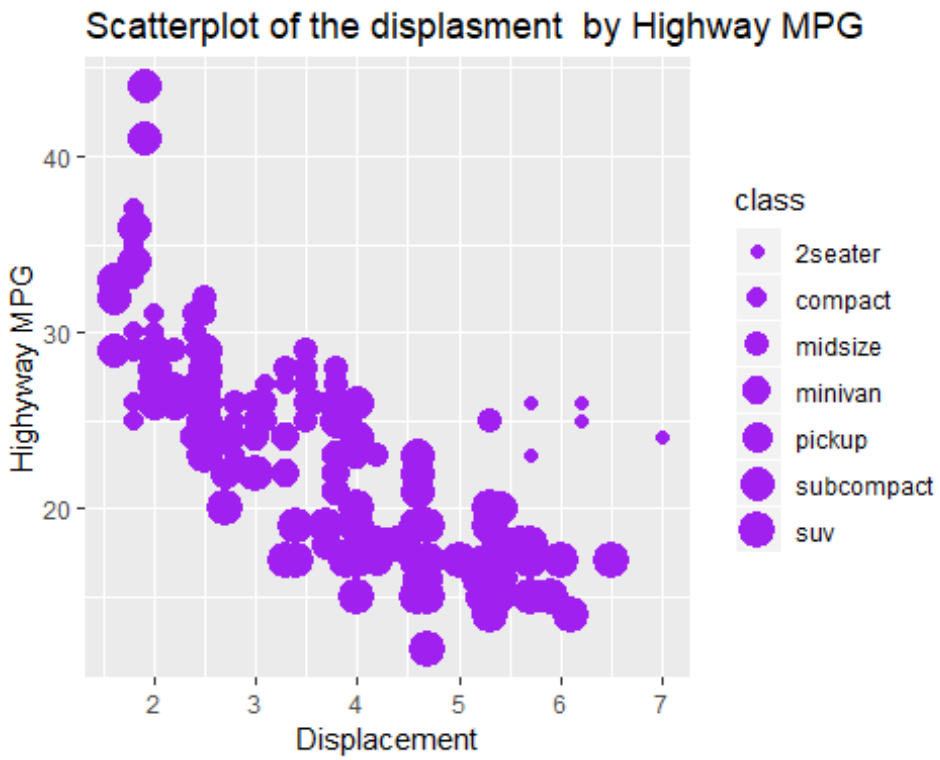



```
ggplot(data=mpg)+
  geom_point(mapping = aes(x=displ,y=hwy,size= cty),color='pink')+
  labs(x="Displacement",y="Highyway MPG",title="Scatterplot of the displasment by Highway
MPG")
```




```
#categorical variable
ggplot(data=mpg)+
  geom_point(mapping = aes(x=displ,y=hwy,size= class),color="purple")+
  labs(x="Displacement",y="Highyway MPG",title="Scatterplot of the displasment  by Highway
MPG")
```

```
## Warning: Using size for a discrete variable is not advised.
```



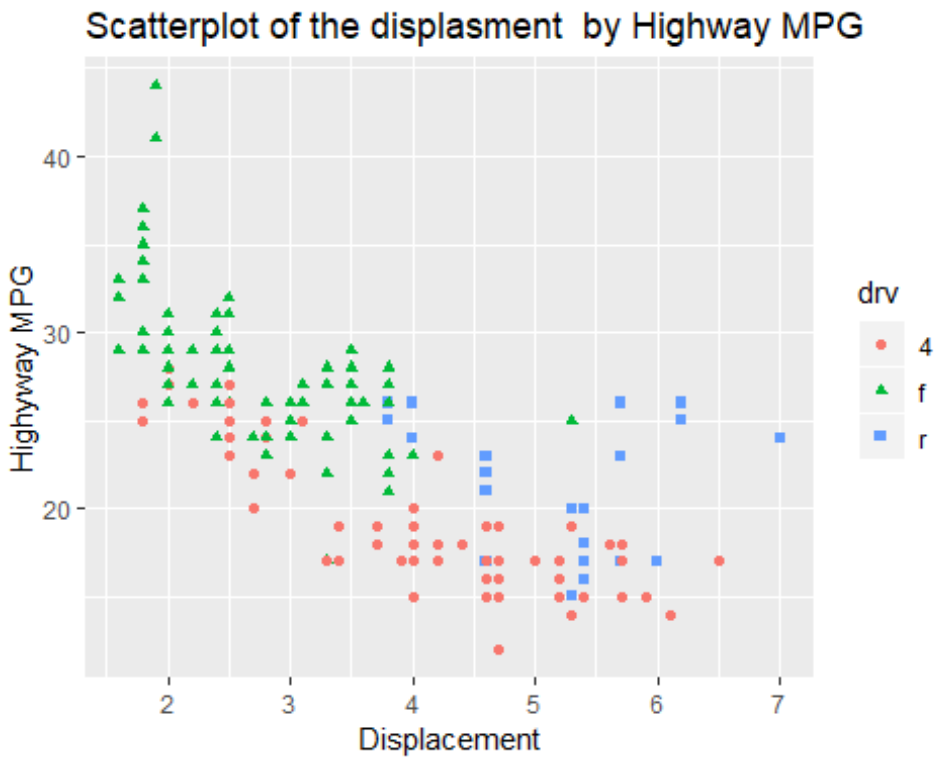
Behave of aesthetics for continous and categorical variables

- The result shows size and color aesthetic works for continous varibels while assiging shape aesthetis to continous varibale get error beacuse shape are not orderd while continous varibles are orderd. Also, using size aesthetic for categorical variable doesnot make sene baeacuse of non oridnal of categorical variables.

- 3.3.1 Exercise 4:

The result of mapping same varibles to multiple aesthetics

```
ggplot(data=mpg)+
  geom_point(mapping = aes(x=displ,y=hwy,color= drv,shape=drv))+
  labs(x="Displacement",y="Highway MPG",title="Scatterplot of the displasment  by Highway MPG")
```



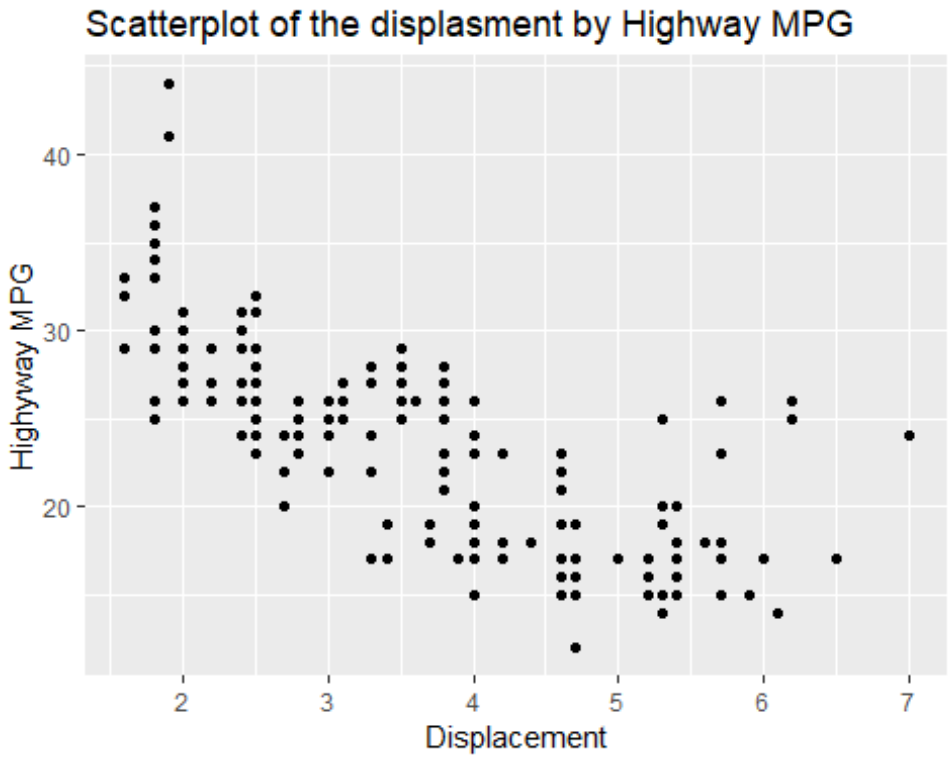
- We can map the same variable to multiple aesthetics.In some data even get us more insight to analyse data. But we just need consider that the aesthetics are compatibale with the type of the vaiable.

- 3.3.1 Exercise 6:

The result of mapping aesthetic to sth other than a varibale name.

```
ggplot(data=mpg)+
  geom_point(mapping = aes(x=displ,y=hwy,class= displ<5))+
  labs(x="Displacement",y="Highway MPG",title="Scatterplot of the displasment by Highway MPG")
```

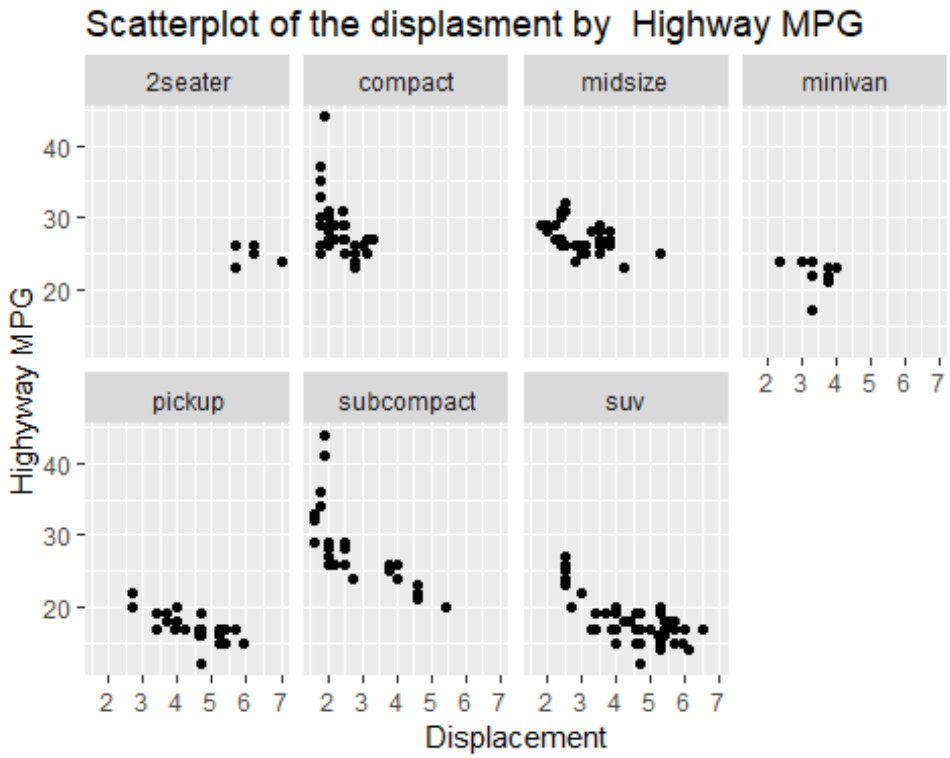
```
## Warning: Ignoring unknown aesthetics: class
```



- The above plot shows that if aesthetics map to sth other than varible name divded the observation into boolean variables.
- 3.5.1 Exercise 4:

The visualtion of using "Facet" insted of "color"

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_wrap(~ class, nrow = 2)+
  labs(x="Displacement",y="Highyway MPG",title="Scatterplot of the displasment by  Highway MPG")
```



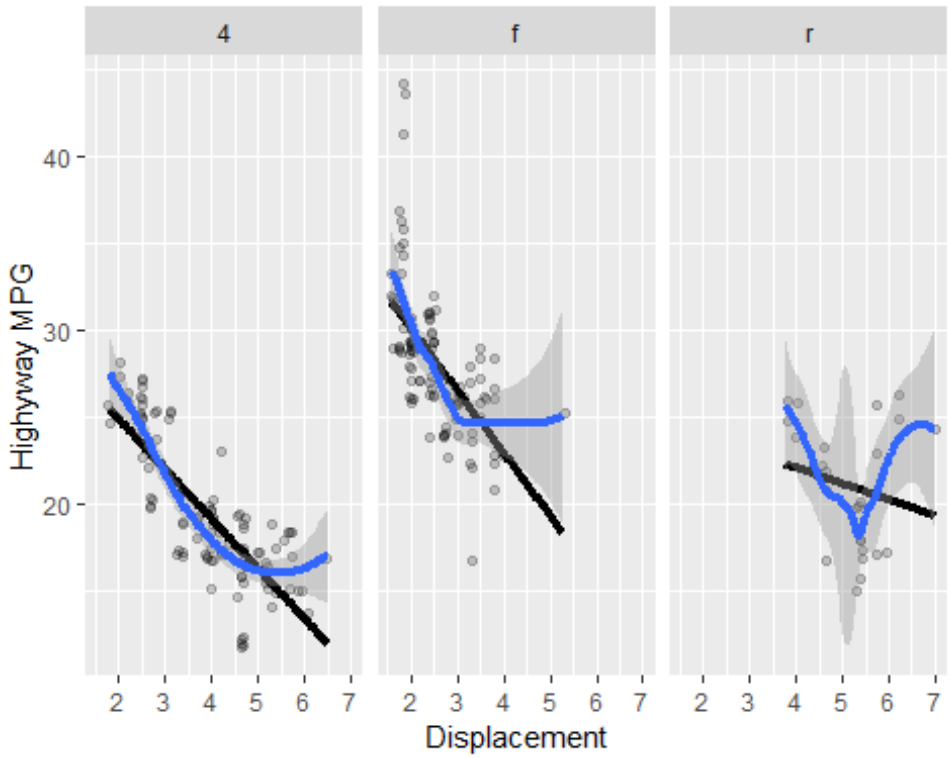
- Advantage: Advantages of using “facet” is allow us to encode more distinct categories insted of dealing with so several colors.
- Disadvantage: With “faceting” it is easier to examine the indivual classes while “coloring” make easier to see how the classes are clustered overall.
- With large dataset: When we have large dataset using “Color”it may become difficult to clearly distinguish points based on color and using “facet” provide simple result.

(b)

Provide reproduce visualization for the “mpg” data

```
ggplot(mpg, aes(x=displ, y=hwy)) +
  geom_point(alpha=0.2,position = "jitter")+
  geom_smooth(method="lm", color="black",span =0.89,se=F,size=1.5)+
  facet_wrap(~drv)+
  geom_smooth(size=1.5)+
  labs(x="Displacement",y="Highway MPG")
```

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



The plot shows a negative relationship between engine size (displ) and fuel efficiency (hwy). Also we can see that front-wheel drive with the little engine can be more effecincy.

Problem 2: Generating data and advanced density plots

(a)

Generate random variable.Four sets of random numbers a-d have been generated in different distributions.The variales combined into a data frame “df” and then reshape data with “gather” function and create “df2”data frame.

melted into a new data frame (df2

df data frame

```
df=data.frame(a,b,c,d)
head(df,2)
```

```
##           a        b       c         d
## 1 -0.5071735 0.3328360 5.127979 0.4846009
## 2 -1.7335064 0.3543254 2.322969 0.5682989
```

df2 data frame

```
df2<- gather(df, key="groupVar",value="value")
head(df2,4)
```

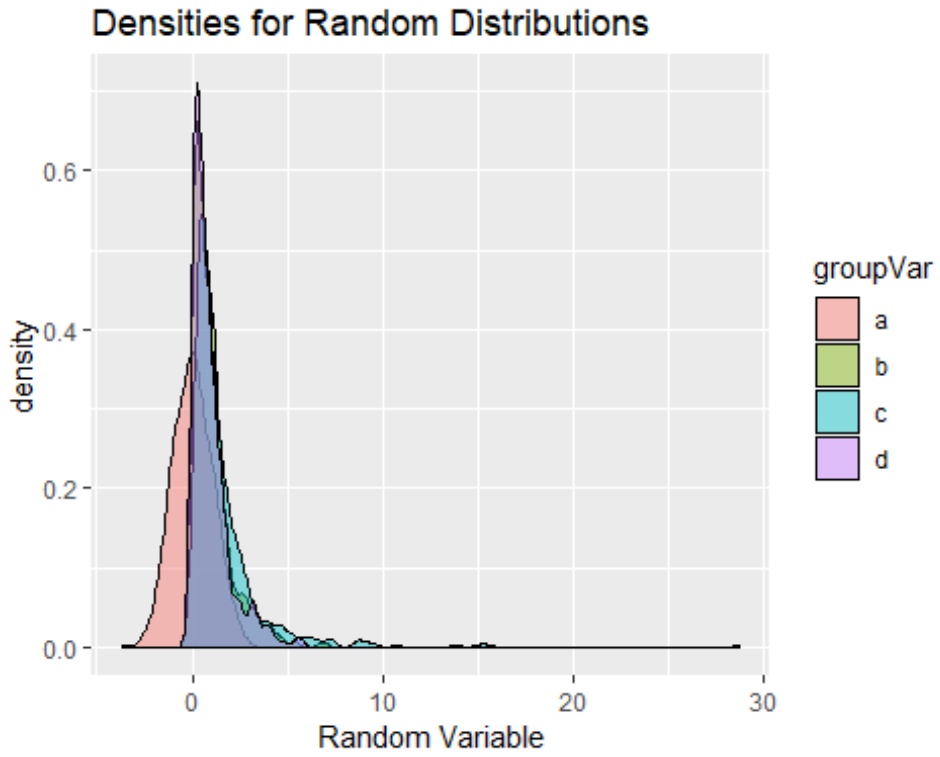
```
##   groupVar      value
## 1        a -0.5071735
## 2        a -1.7335064
## 3        a  0.3623584
## 4        a  0.3280874
```

- The figure shows the density function for the random data sets. Different distributions have different behavior as shown in the

(b)

Using ggplot to visulize the densities

```
ggplot(df2, aes(x = value, fill = groupVar)) +
  geom_density(alpha = 0.45) +
  labs(x = "Random Variable",
       title = "Densities for Random Distributions")
```



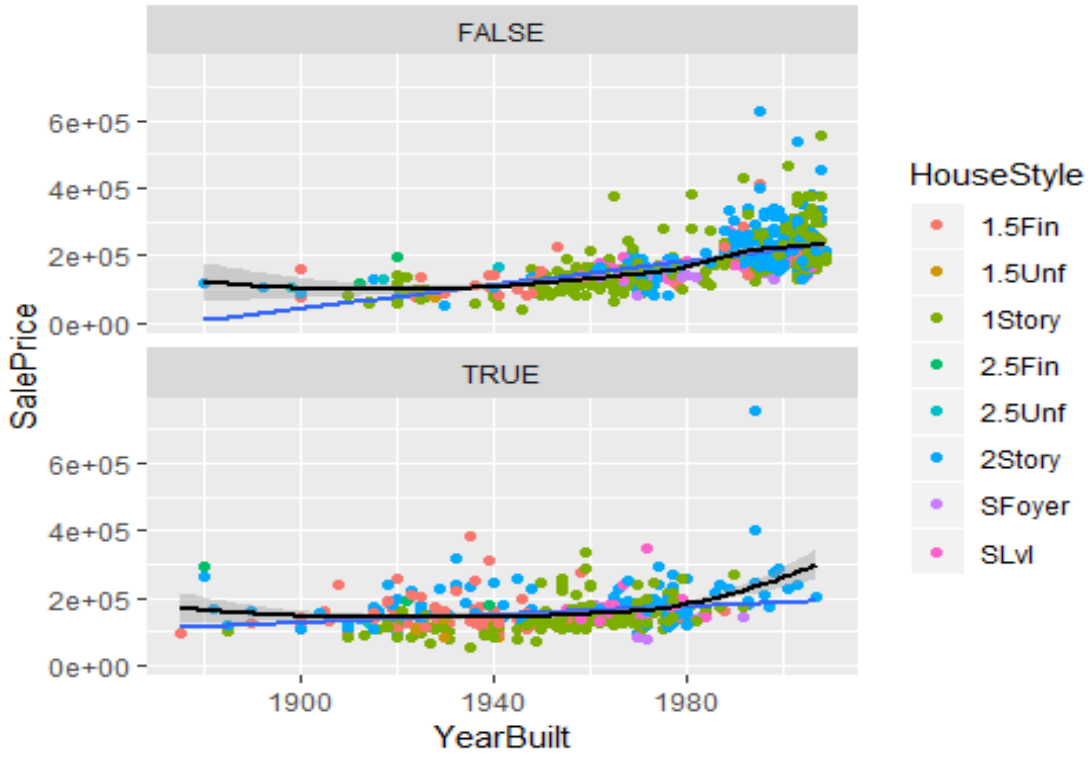
problem 3: house prices data

Read data from file“housingData.csv and visualize the housing dataset

```
housing<-read.csv("housingData.csv")

ggplot(housing, aes(x=YearBuilt, y=SalePrice)) +
  geom_point(aes(color=HouseStyle)) +
  geom_smooth(method="lm",se=F) +
  facet_wrap(~OverallCond>5,nrow = 2)+ geom_smooth(color="black")
```

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



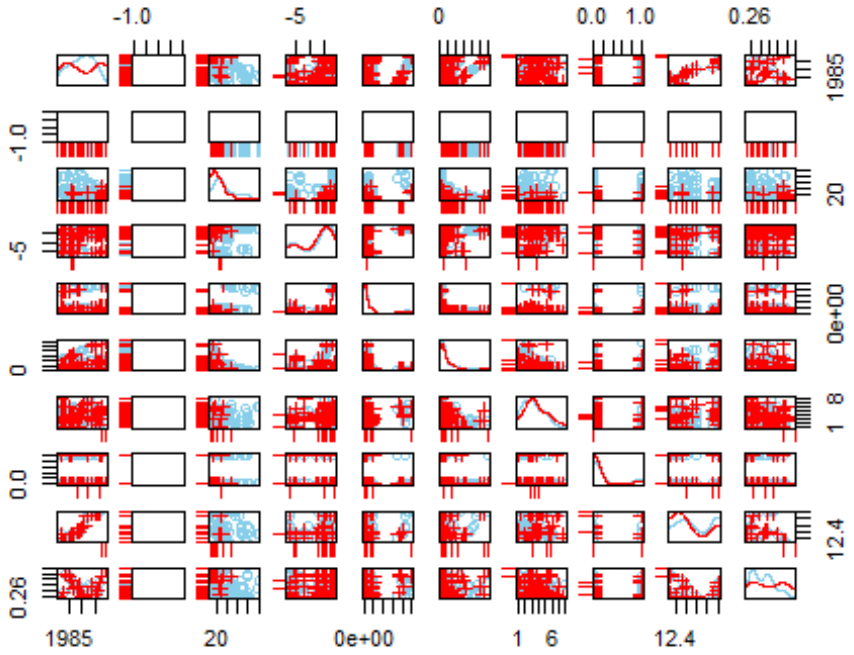
- The plot reveals that regardless of the houses condition during the 1950-1970 the 1story Style house with the min price of 200k were popular. This trend changes by the time over the first decad of 20 centurty. People pay more for same house style with the worse condition.

problem 4:Missing Data

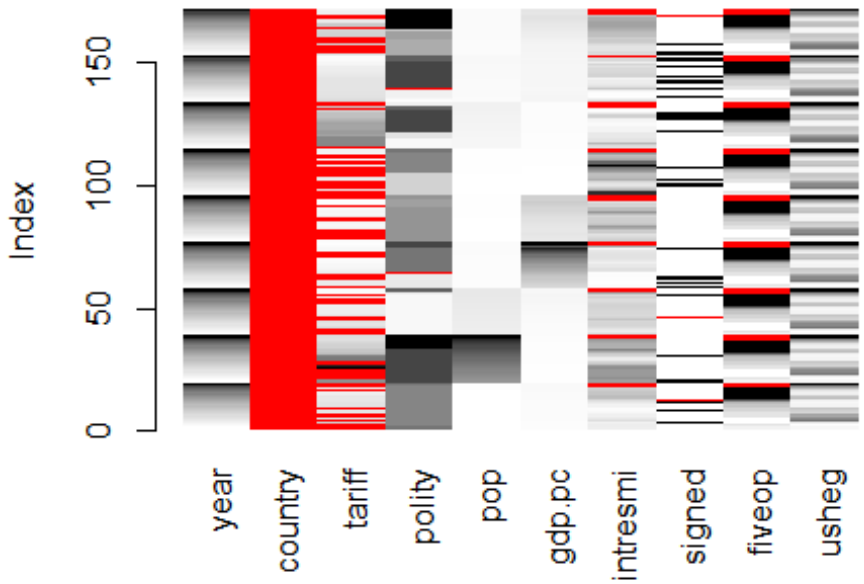
(a)

Explore the missingness on “freetrade” dataset

```
# sactterplot for missing values
scattmatrixMiss(freetrade)
```




```r
# Matrix plot for missing values
matrixplot(freetrade, interactive = F, sortby = "country")
```



- The Result of invistigation on “freetrade” data set shows that the “tariff” variable contain more mising value among the other values and “fiveop”and “intresmi” are in the second and third place of mising values.

(b)

- I used Chi -squred tset for this question and the result of it on orginal dataset and the dataset without “Nepal” and “Philipine” reveal that p-value is the same for all three conditions. Therefore we can asume that there is no significant change if we remove some data from dataset. So it can be concluded that the missing value in “tariff” is completely independed and there is no realtionship between the “country” and “tariff” valiable.