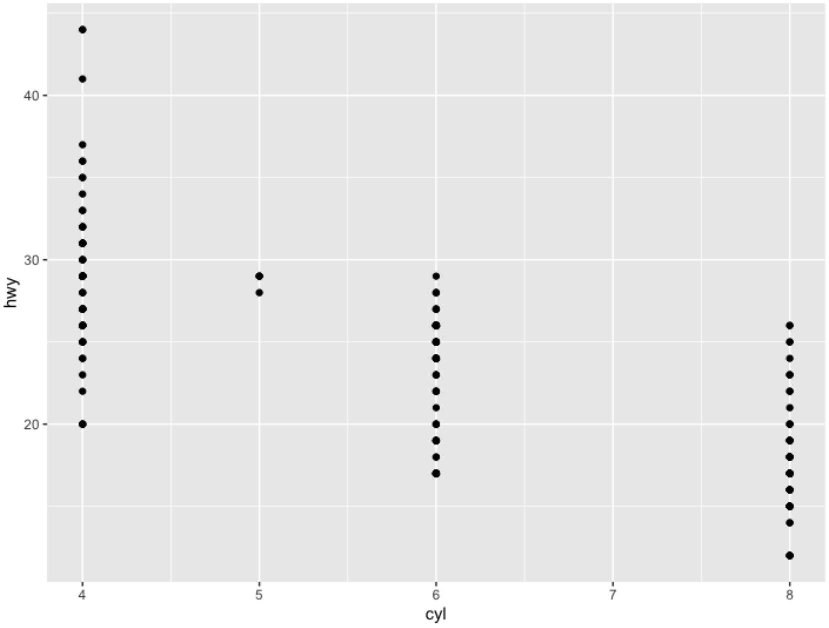
Homework2

Q1.(a)3.2.4 #4

Rcode:

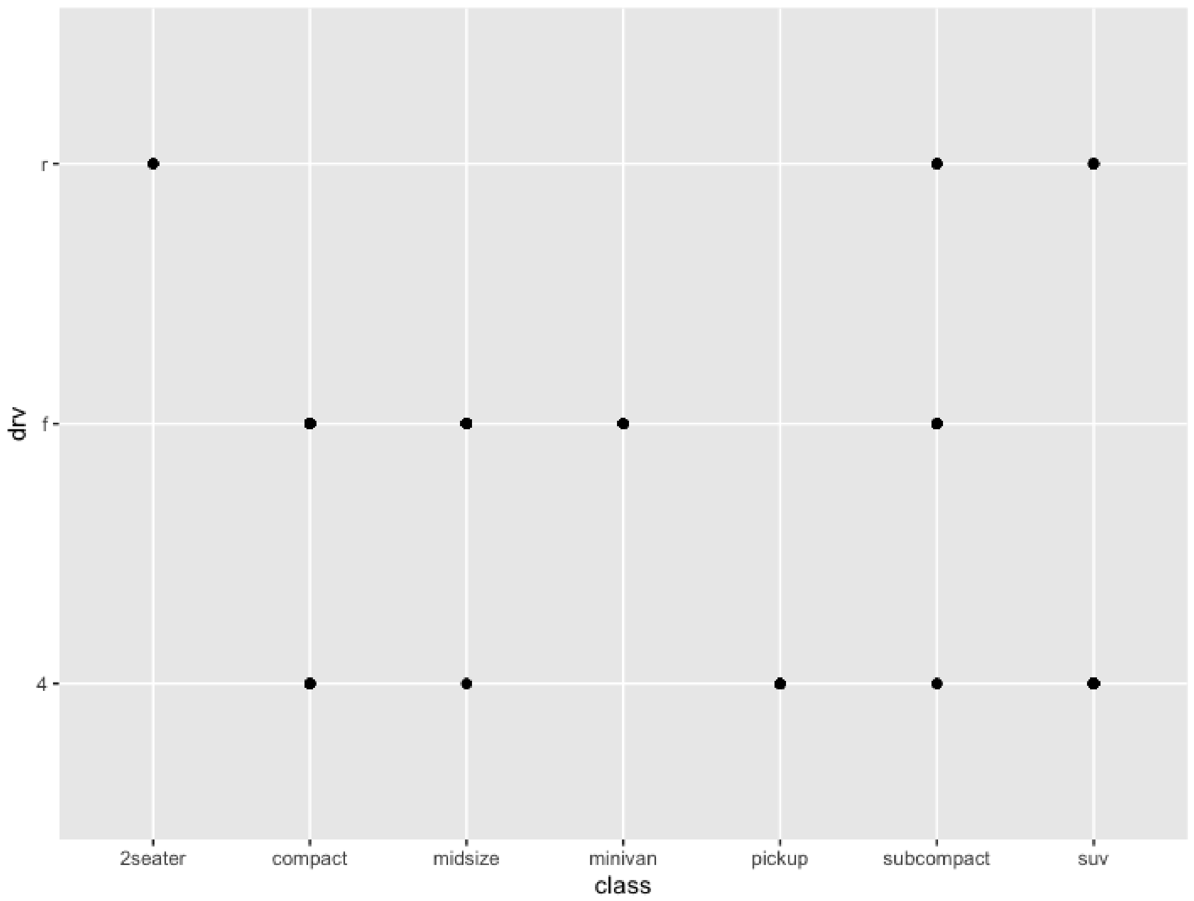
```
ggplot(data=mpg)+
geom_point(mapping=aes(x=cyl,y=hwy))
```



The graph shows the scatterplot of hwy vs cyl
X axis represent number of cylinders
Y axis represent highway miles per gallon

Q1(a)3.2.4 #5

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



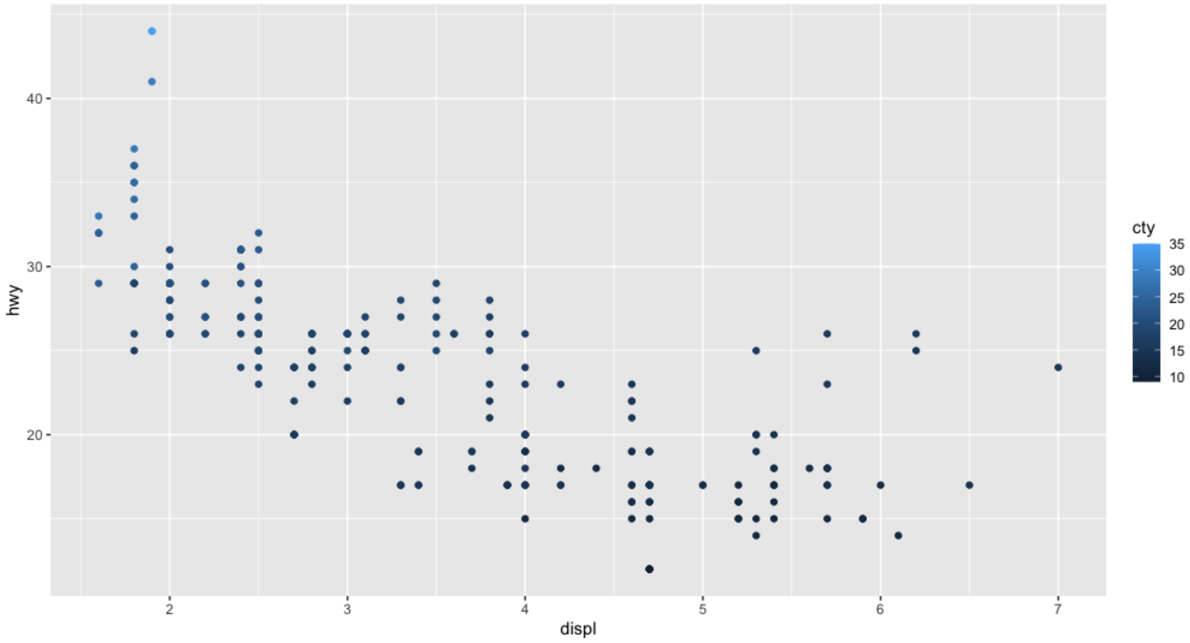
When we make a scatter plot of class vs drv for the data given we just get a few values defining both the variables.This plot is not useful because both the variables are categorial and the data is limited that it is difficult to reach any conclusion based on this data. Only there are 21 values that are possible to plot on scatter plot of drv vs class.we can see only 12 values in this plot.

Q1(a) 3.3.1 Exercise #3:

Continuous variable

Taking a continuous variable year into consideration got the following scatter plots

Color

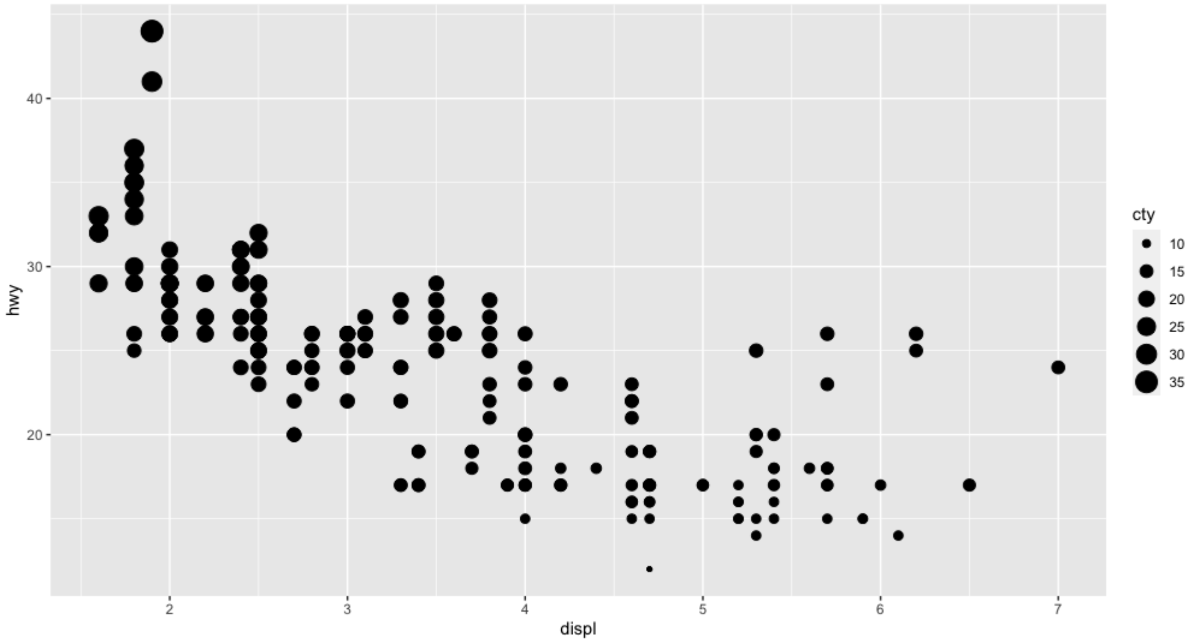


```
ggplot(data=mpg)+
        geom_point(mapping=aes(x=displ , y=hwy , colour=cty))
```

The variable cty is city highway miles per gallon

Size

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



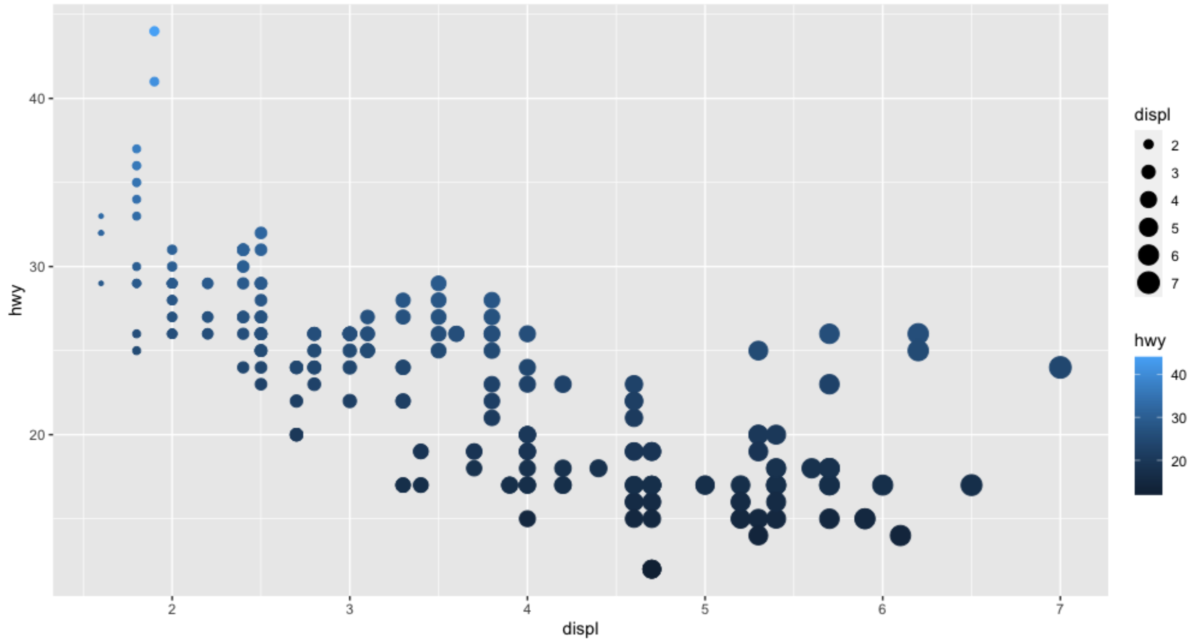
Shape

```
Error in `scale_f()`:
! A continuous variable can not be mapped to shape.
```

Thus, by looking at the results we can say that when a continuous variable is mapped to shape it throws error saying it cannot be mapped.We can split continuous variable into discrete categories and use a shape asthetic but again if we check conceptually it will not make sense.A numerical variable can be defined in a order but when it comes to colour shape and size we can say that greater size has greater value but we cant say that circle is greater than square or not. Whereas categorial value can be mapped to shape

Q1(a) 3.3.1 Exercise #4

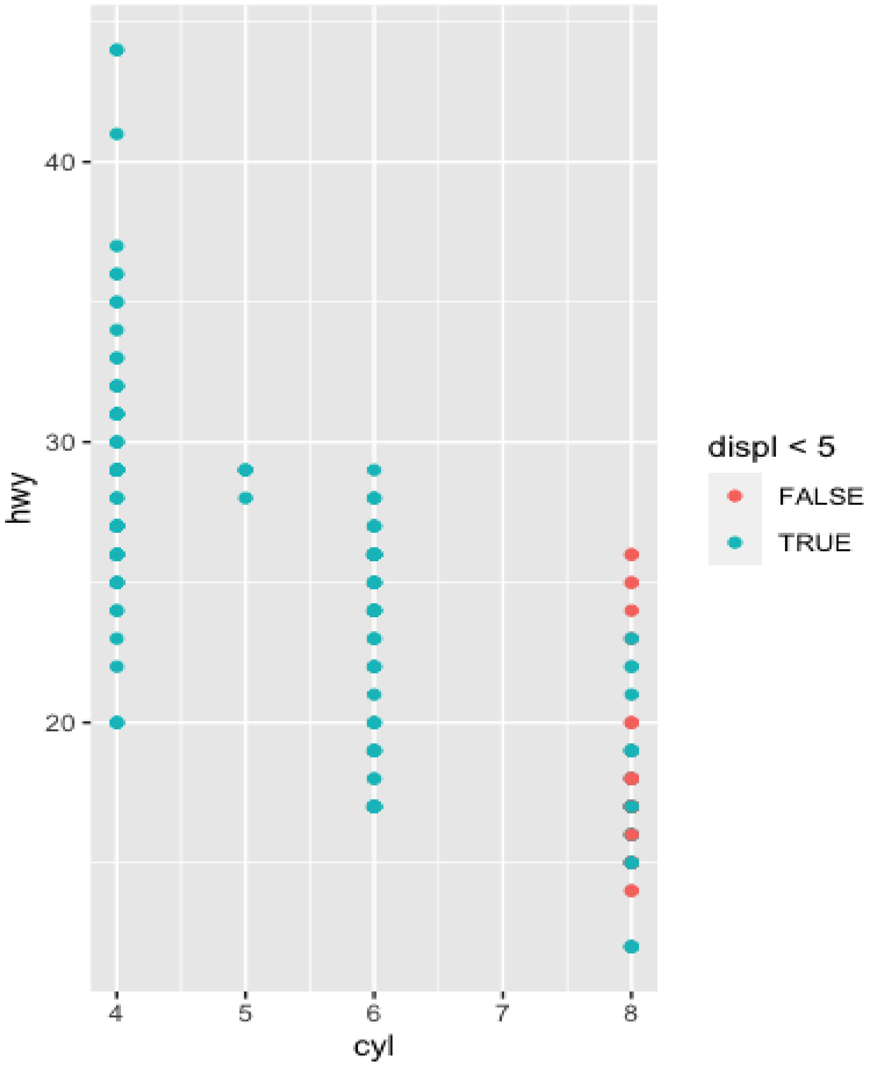
```
ggplot(data=mpg)+
      geom_point(mapping=aes(x=displ, y=hwy,size=displ,colour=hwy))
```



In the above graph engine displacement is mapped to both x axis and size, highway miles per gallon is mapped to y axis and colour and we can see we are able to plot the graph successfully.Anyways we are using the variable in the graph to analyse and again using it with multiple asthetics is redundant.

Q1(a) 3.3.1 Exercise #6

```
ggplot(data=mpg)+
      geom_point(mapping=aes(x=cyl,y=hwy, color=displ<5))
```



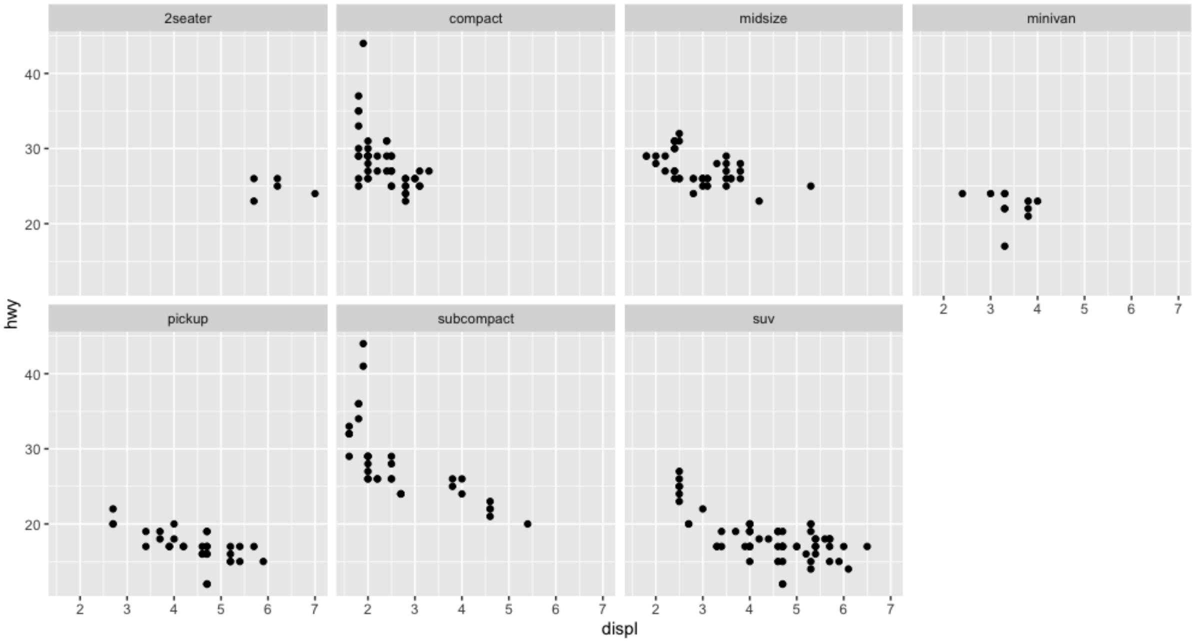
This graph shows the cyl and fuel efficiency of cars with engine capacity less than 5 litres. The blue dots represent the engine capacity is less than 5 for the cars is true whereas the red dots represent the engine capacity with more than or equal to 5 for those cars is false. displ<5 is a logical variable which takes true and false

Q1(a) 3.5.1 Exercise #4:

The advantages of using faceting instead of the colour asthetic is it divides the plot into each categories of variables into each separate plot on the face of the graph so we can compare the basic x and y value

The disadvantages of using it is since all the properties on the same plot have separate plots so the large number values may overlap. Also it creates confusion when ready and difficult to read each plot separately and more precisely.

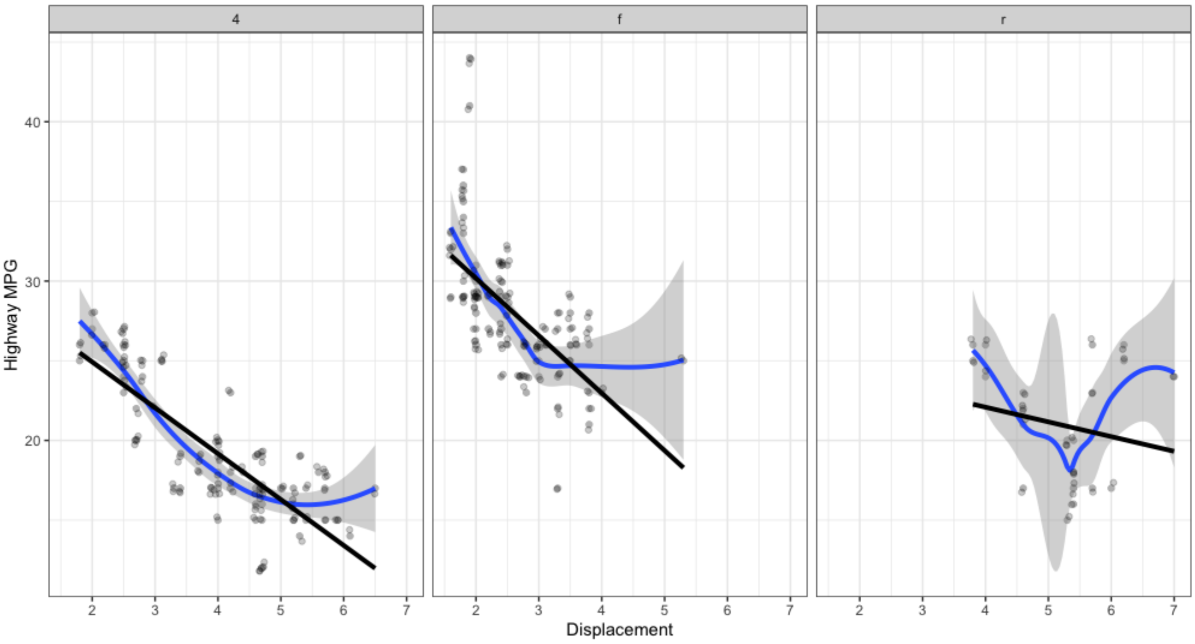
Also if we use different colours there are so many colours that it may create confusion in reading the large number of data



Q1(b)

Plot is clear and labeled.

```
ggplot(mpg, aes(displ, hwy)) + facet_wrap(~drv, ncol = 3) + labs(y="Highway MPG",
x="Displacement")+
  geom_point(color='grey') + geom_smooth(method = loess , formula = y ~ x,size = 1.40) +
  geom_jitter(position = "jitter", alpha = 0.21) +
  geom_smooth(method = lm, se = FALSE,color ='black',size = 1.40) + theme_bw()
```



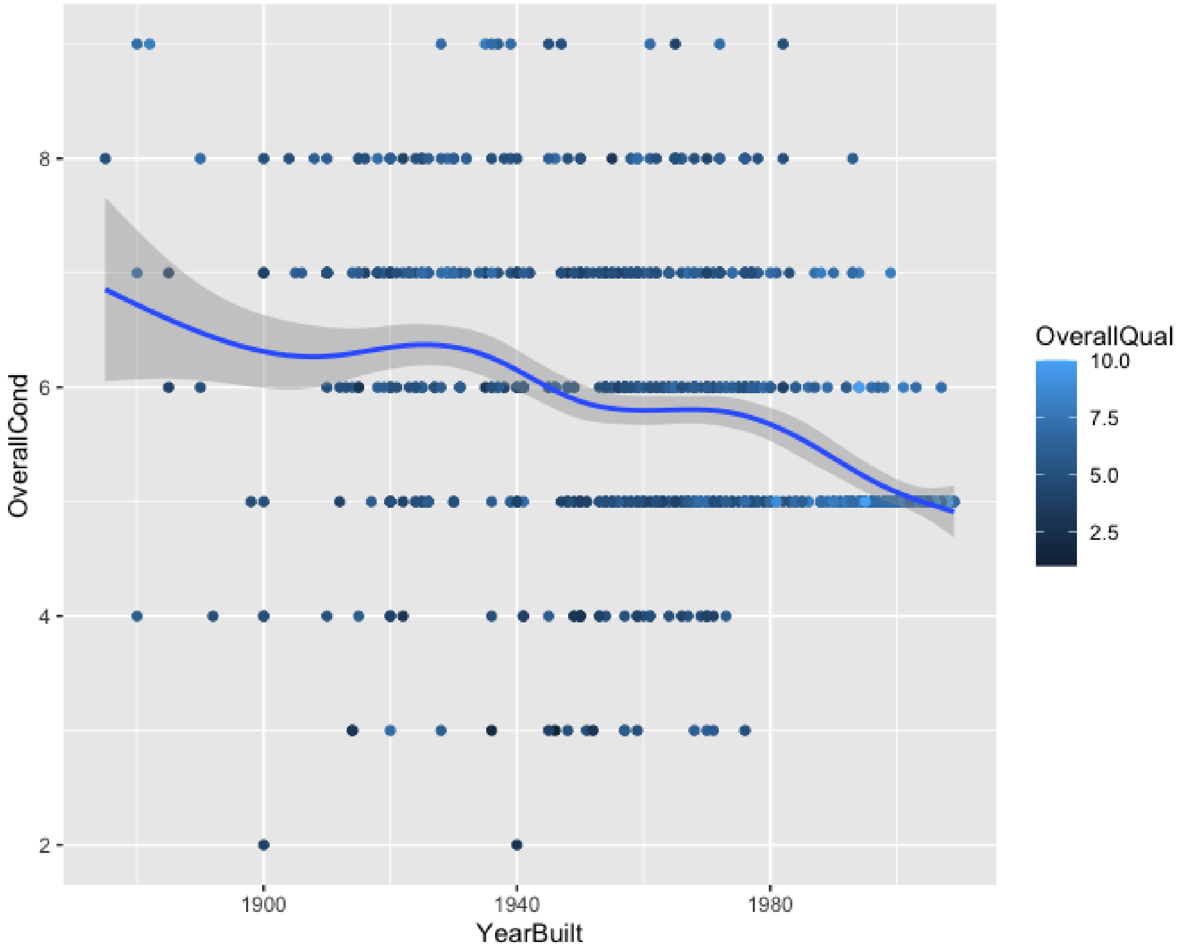
Q2 House Price visualisations

5 good visualisations with appropriate commentary

1.This Scatter and line plot shows the overall cond of building with respect to year built

Checking the plot shows the condition of older buildings are better than the condition of newer building

```
ggplot(data=housingData)+
  geom_point(mapping=aes(x=YearBuilt,y=OverallCond,colour=OverallQual))+
  geom_smooth(mapping=aes(x=YearBuilt,y=OverallCond,colour=OverallQual))
```

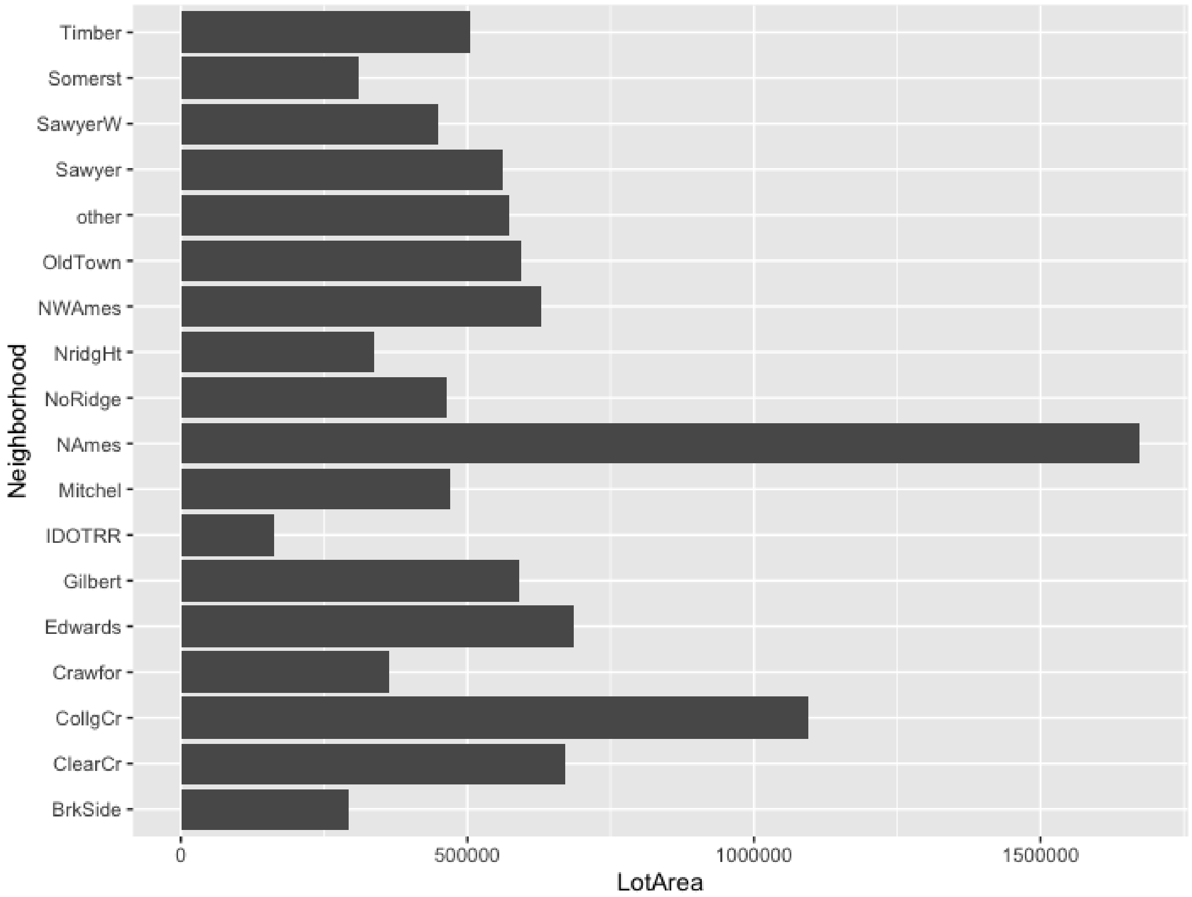


2.

This bargraph shows the graph between lot area and Neighborhood

We can see NAmes neighborhood has the highest area unlike IDOTRR

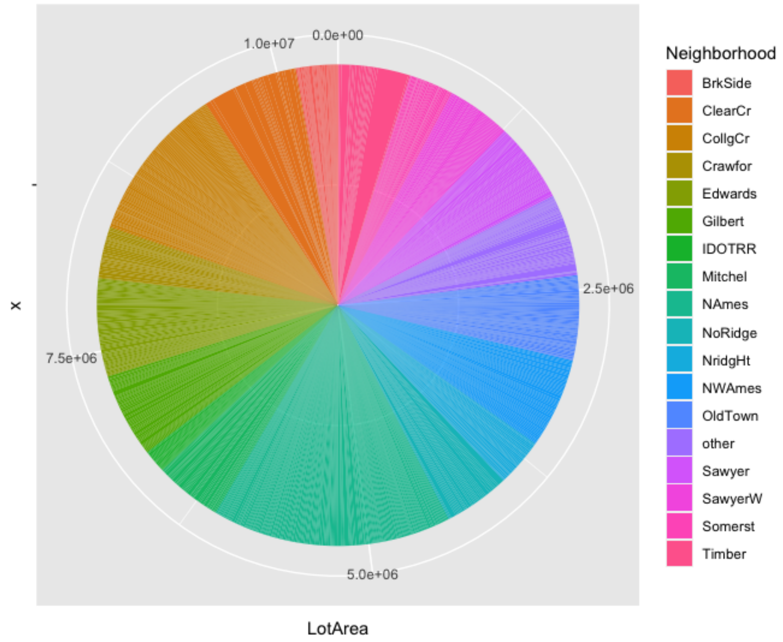
```
ggplot(data = housingData) +
  geom_bar(mapping = aes(x=LotArea,y=Neighborhood),stat="identity")
```



3.

This pie shows the Lot area verses Neighbourhood chart.The following observations are made:
NAmes has the biggest lot area when compared to others
IDOTRR has the least lot area when compared to others

```
ggplot(housingData, aes(x=” ”, y=LotArea, fill=Neighborhood))+
geom_col()+
coord_polar(theta=”y”)
```

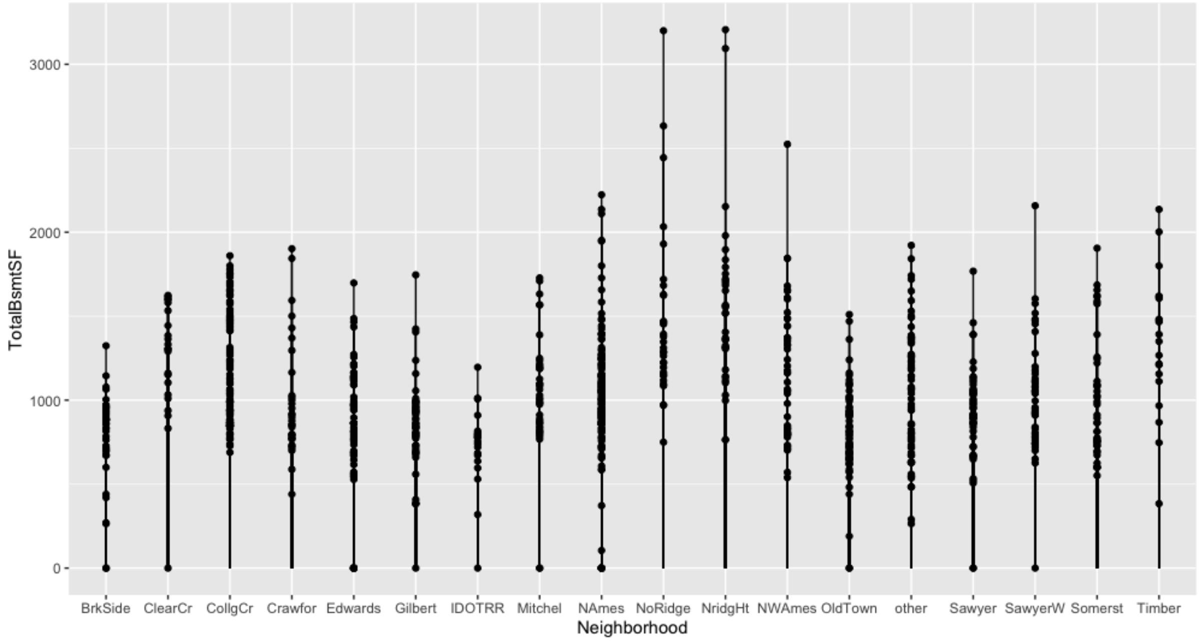


4.

This Lollipop chart represents the total square feet of basement area verses the neighborhood.The following observations made:

Neighborhood NoRidge and NridgHt has the highest basement area whereas the IDOTRR has the least

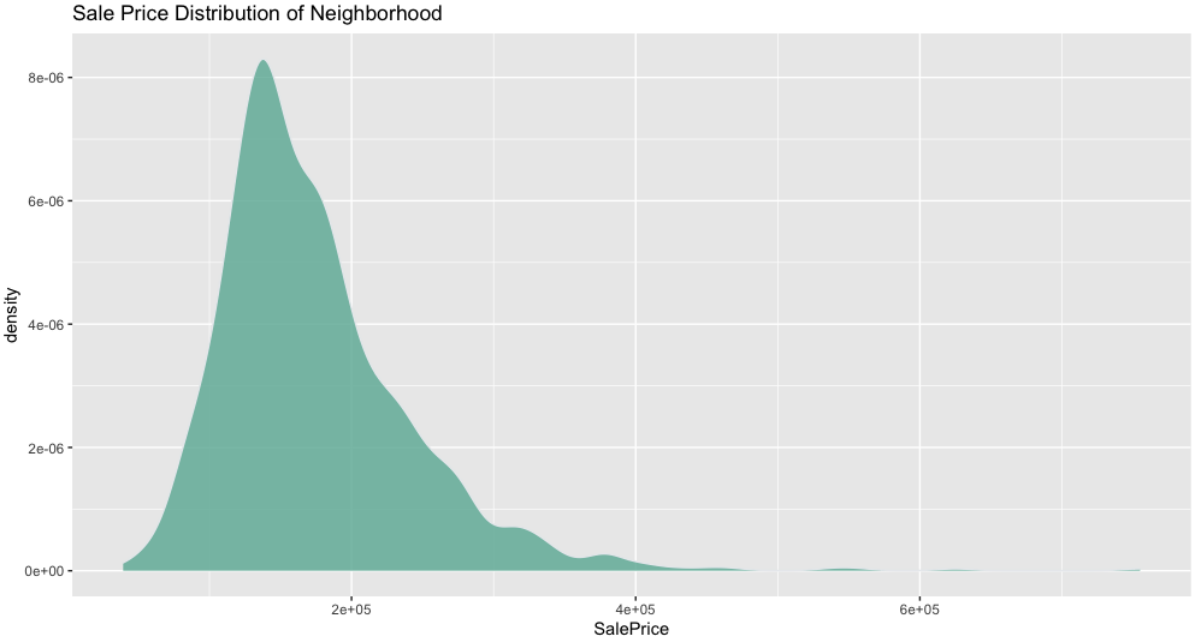
```
ggplot(housingData, aes(x = Neighborhood, y = TotalBsmtSF)) +
   geom_segment(aes(x =Neighborhood, xend =Neighborhood, y = 0, yend = TotalBsmtSF)) +
   geom_point()
```



5.

This graph shows the sales price distribution of different neighbourhood

```
 ggplot( housingData,aes(x=SalePrice)) +
   geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8) +
ggtitle("Sale Price Distribution of Neighborhood")
```


Sale Price Distribution of Neighborhood
density
8e-06
6e-06
4e-06
2e-06
0e+00
2e+05
4e+05
6e+05
SalePrice