DataScience Learning

Kalidoss

April 2, 2017

Data Visualization using R ggplot

Loading the packages required which has the some dataset we can use for visualization

```
#install.packages("tidyverse")
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.3.3
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Warning: package 'purrr' was built under R version 3.3.3
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

Now, Just view the data in table format

```
mpg
```

```
## # A tibble: 234 × 11
##    manufacturer      model displ  year   cyl      trans   drv   cty   hwy
##           <chr>      <chr> <dbl> <int> <int>      <chr> <chr> <int> <int>
## 1          audi         a4   1.8  1999     4   auto(l5)     f    18    29
## 2          audi         a4   1.8  1999     4 manual(m5)     f    21    29
## 3          audi         a4   2.0  2008     4 manual(m6)     f    20    31
## 4          audi         a4   2.0  2008     4   auto(av)     f    21    30
## 5          audi         a4   2.8  1999     6   auto(l5)     f    16    26
## 6          audi         a4   2.8  1999     6 manual(m5)     f    18    26
## 7          audi         a4   3.1  2008     6   auto(av)     f    18    27
## 8          audi a4 quattro   1.8  1999     4 manual(m5)     4    18    26
## 9          audi a4 quattro   1.8  1999     4   auto(l5)     4    16    25
## 10         audi a4 quattro   2.0  2008     4 manual(m6)     4    20    28
## # ... with 224 more rows, and 2 more variables: fl <chr>, class <chr>
```

```
summary(mpg)
```

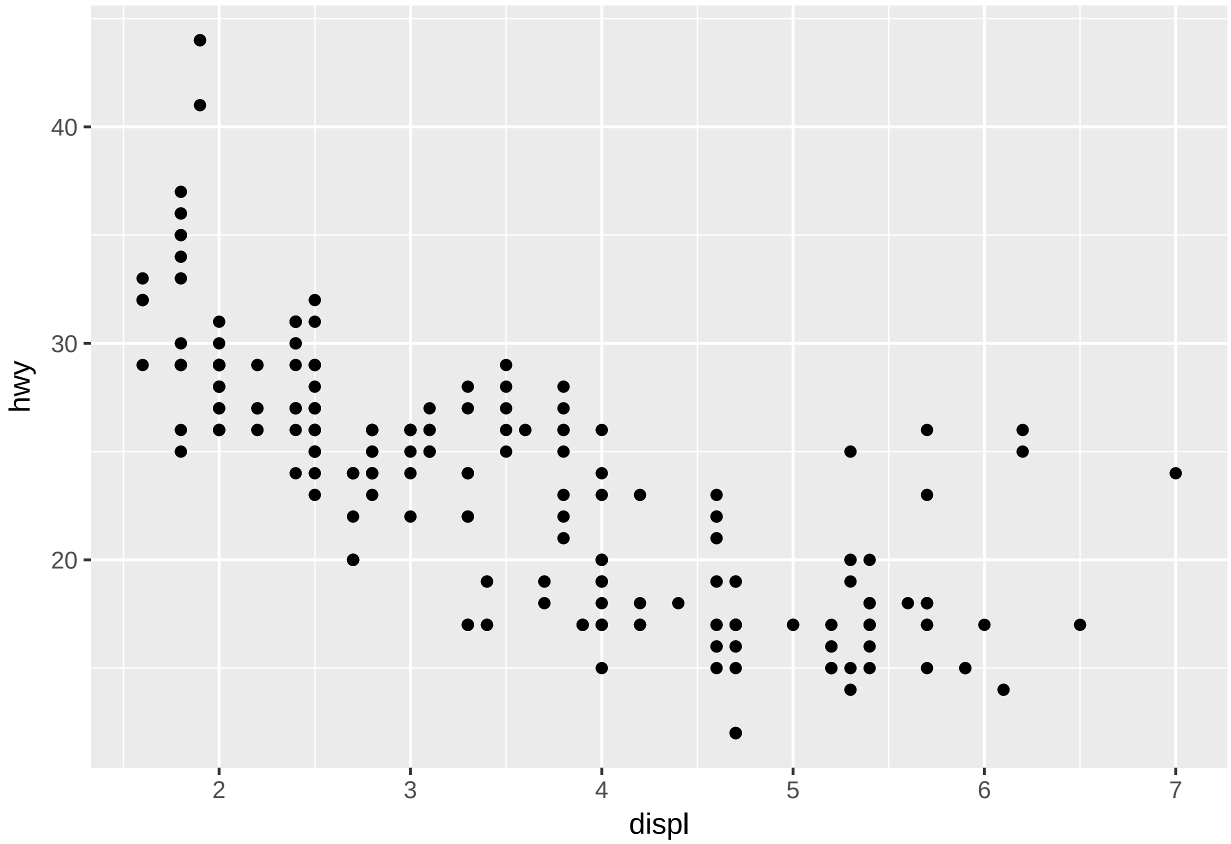
```
##  manufacturer          model               displ            year     
##  Length:234         Length:234         Min.   :1.600   Min.   :1999  
##  Class :character   Class :character   1st Qu.:2.400   1st Qu.:1999  
##  Mode  :character   Mode  :character   Median :3.300   Median :2004  
##                                        Mean   :3.472   Mean   :2004  
##                                        3rd Qu.:4.600   3rd Qu.:2008  
##                                        Max.   :7.000   Max.   :2008  
##       cyl           trans               drv                 cty       
##  Min.   :4.000   Length:234         Length:234         Min.   : 9.00  
##  1st Qu.:4.000   Class :character   Class :character   1st Qu.:14.00  
##  Median :6.000   Mode  :character   Mode  :character   Median :17.00  
##  Mean   :5.889                                         Mean   :16.86  
##  3rd Qu.:8.000                                         3rd Qu.:19.00  
##  Max.   :8.000                                         Max.   :35.00  
##       hwy             fl               class          
##  Min.   :12.00   Length:234         Length:234        
##  1st Qu.:18.00   Class :character   Class :character  
##  Median :24.00   Mode  :character   Mode  :character  
##  Mean   :23.44                                        
##  3rd Qu.:27.00                                        
##  Max.   :44.00                                        
```

```
str(mpg)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    234 obs. of  11 variables:
##  $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
##  $ model       : chr  "a4" "a4" "a4" "a4" ...
##  $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr  "f" "f" "f" "f" ...
##  $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr  "p" "p" "p" "p" ...
##  $ class       : chr  "compact" "compact" "compact" "compact" ...
```

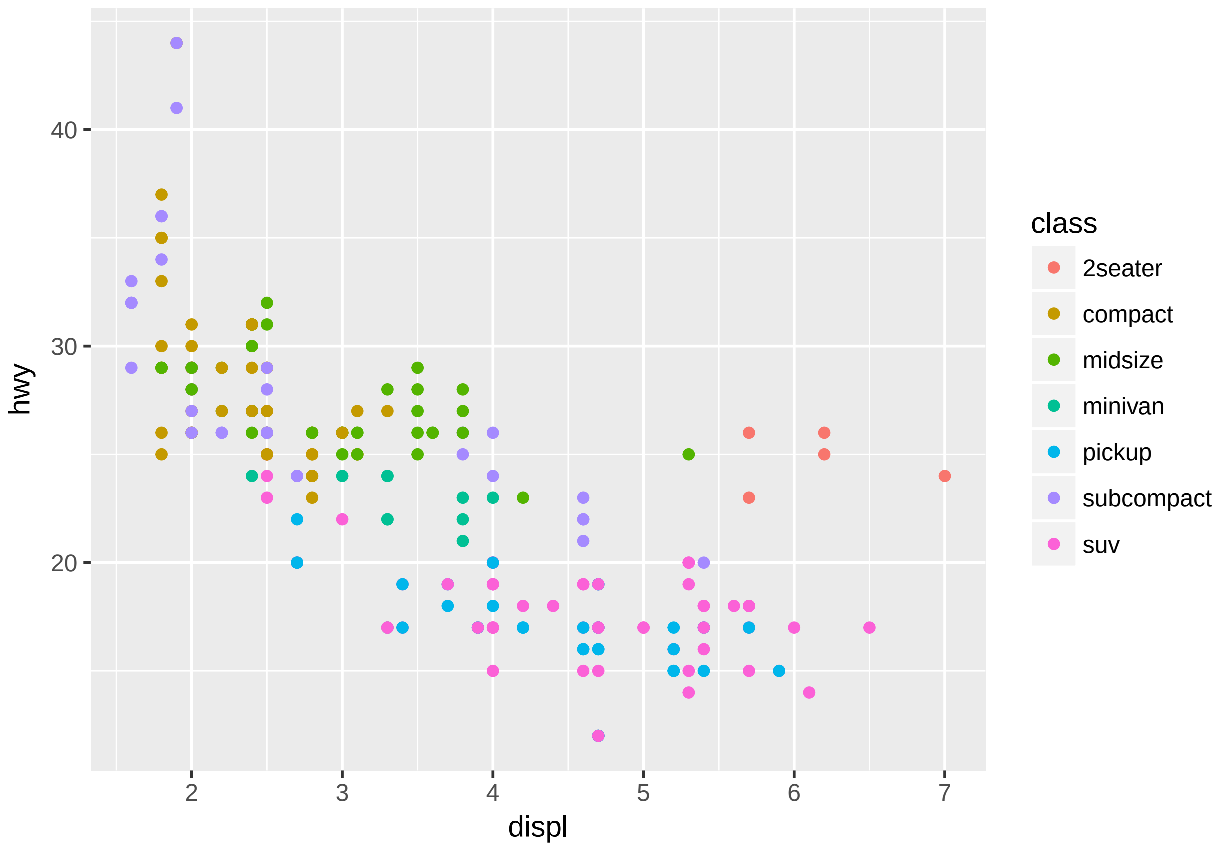
Let's start with simple scatter plot

```
ggplot(mpg, aes(x=displ, hwy) ) +
  geom_point()
```



This plot shows that it has some negative correlation of mileage with respect to Engine size

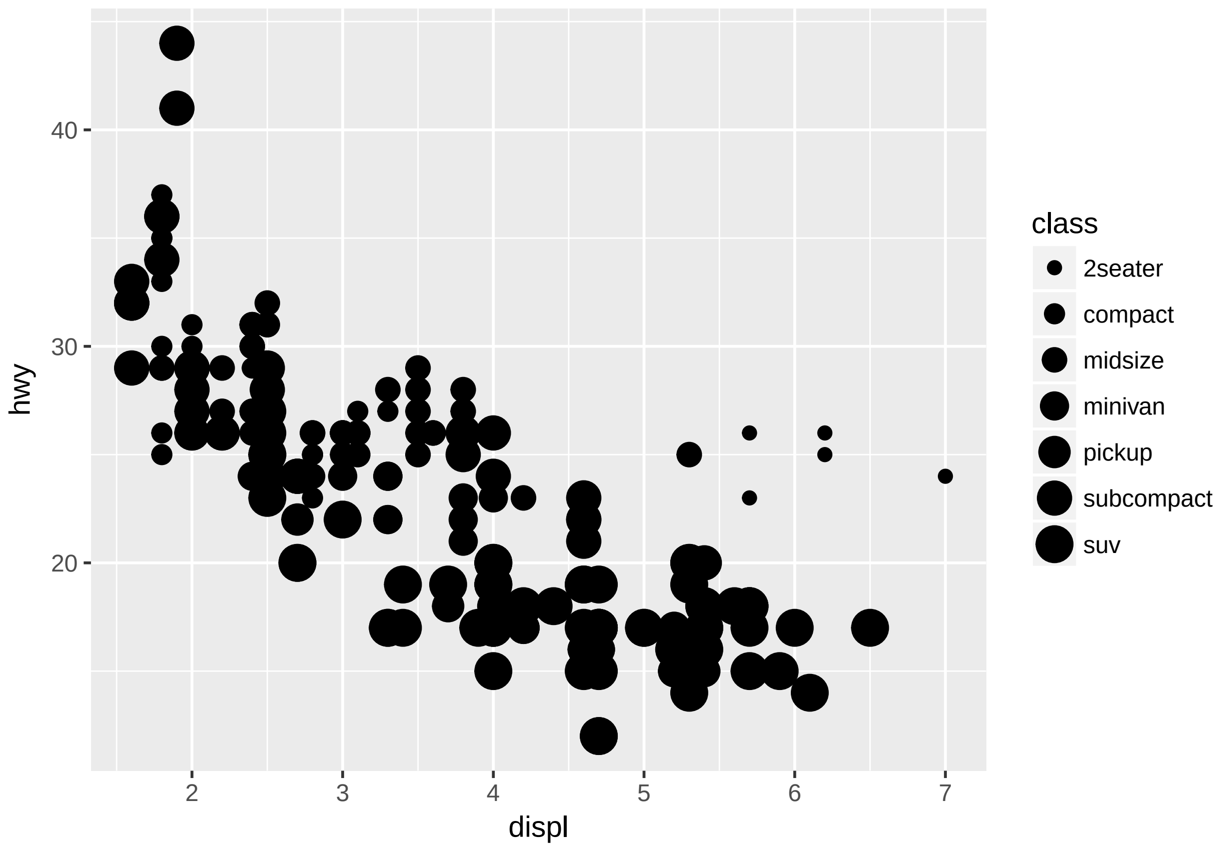
```
ggplot(mpg ) +
  geom_point( aes(x=displ, y=hwy, color=class ))
```



Let us visualze in other way using size aesthetic instead of color

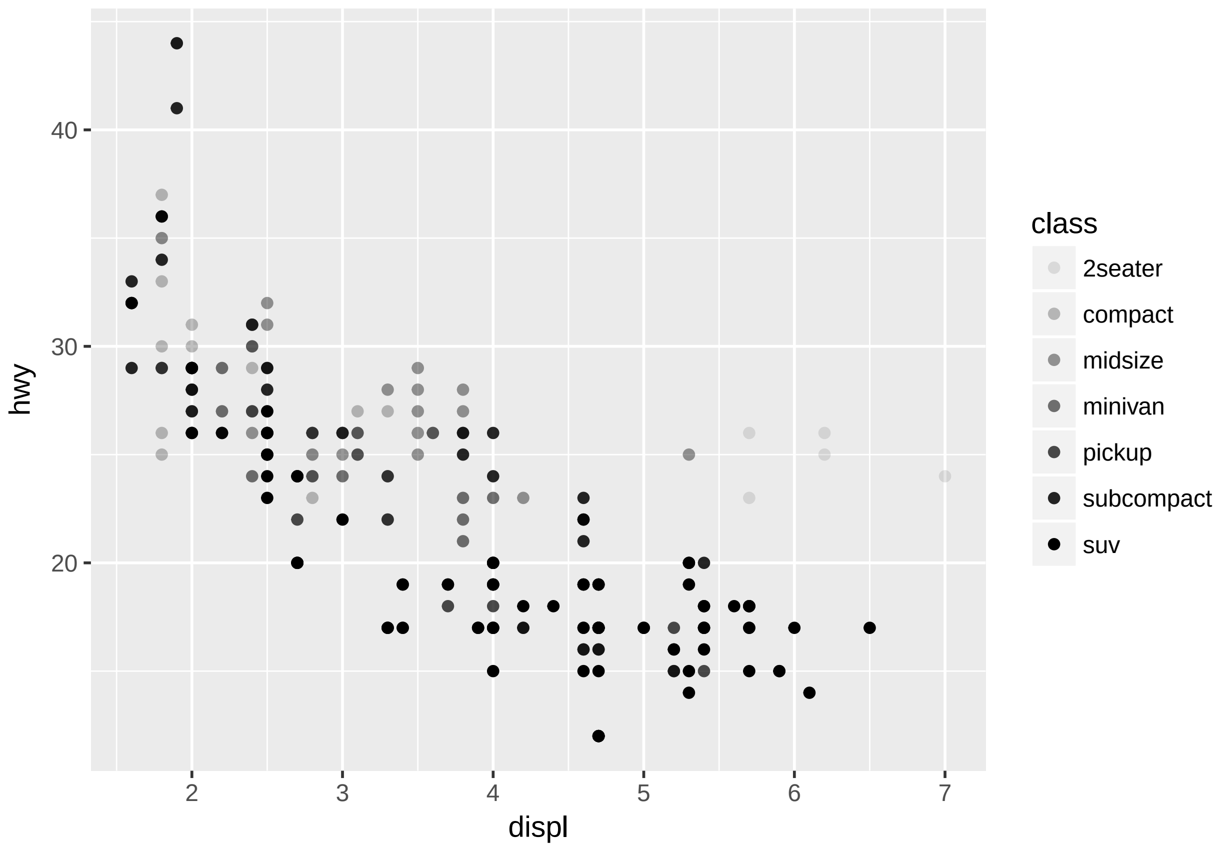
```
ggplot(mpg ) +
  geom_point( aes(x=displ, y=hwy, size=class ))
```

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



Other way of visualizing the plot such as alpha and shape

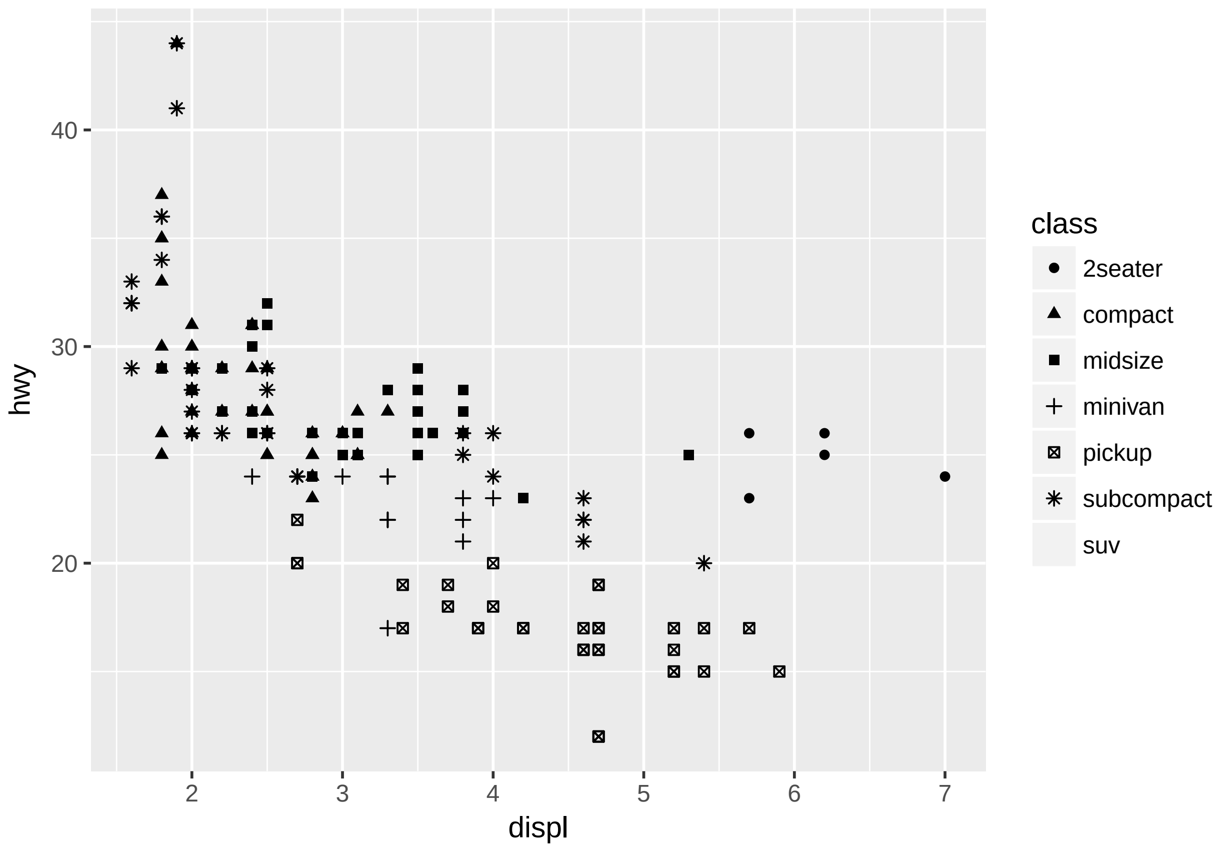
```
ggplot(mpg ) +
  geom_point( aes(x=displ, y=hwy, alpha=class ))
```



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

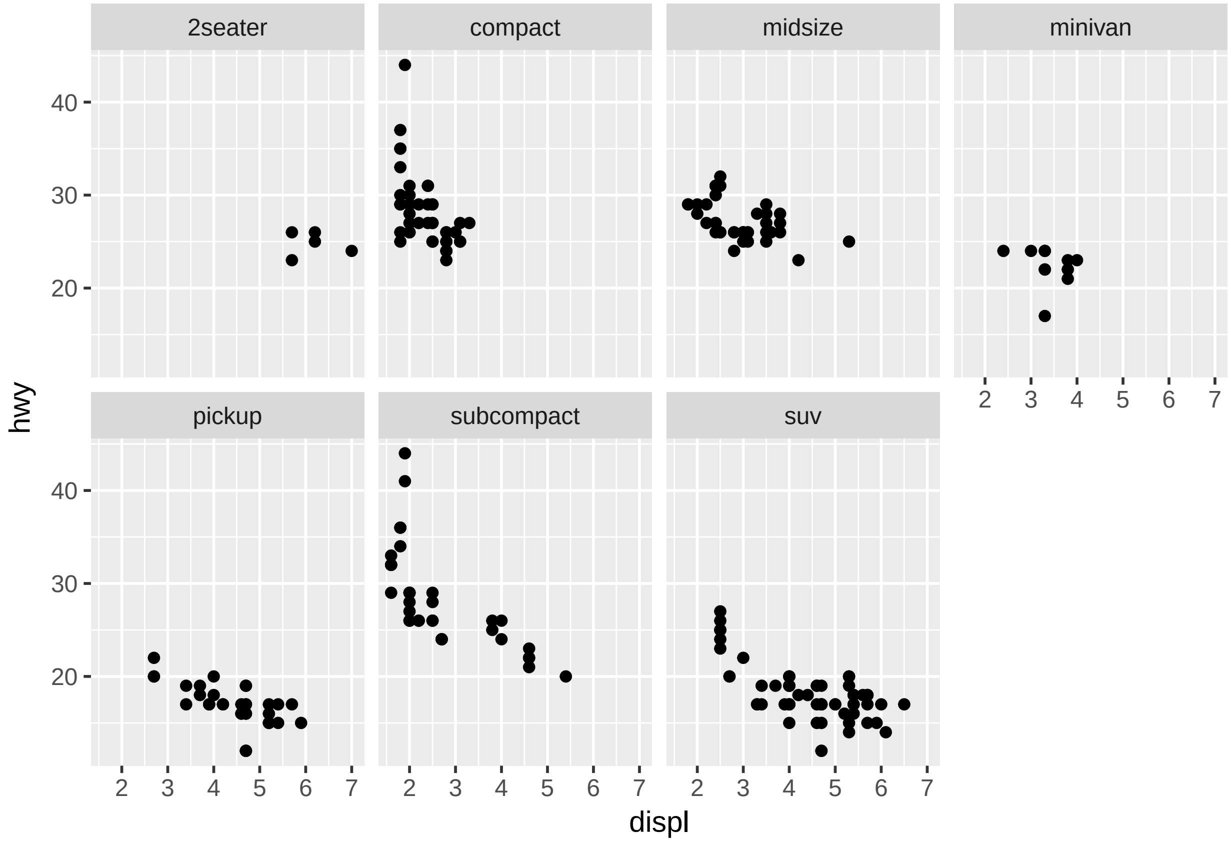
```
## Warning: The shape palette can deal with a maximum of 6 discrete values
## because more than 6 becomes difficult to discriminate; you have 7.
## Consider specifying shapes manually if you must have them.

## Warning: Removed 62 rows containing missing values (geom_point).
```



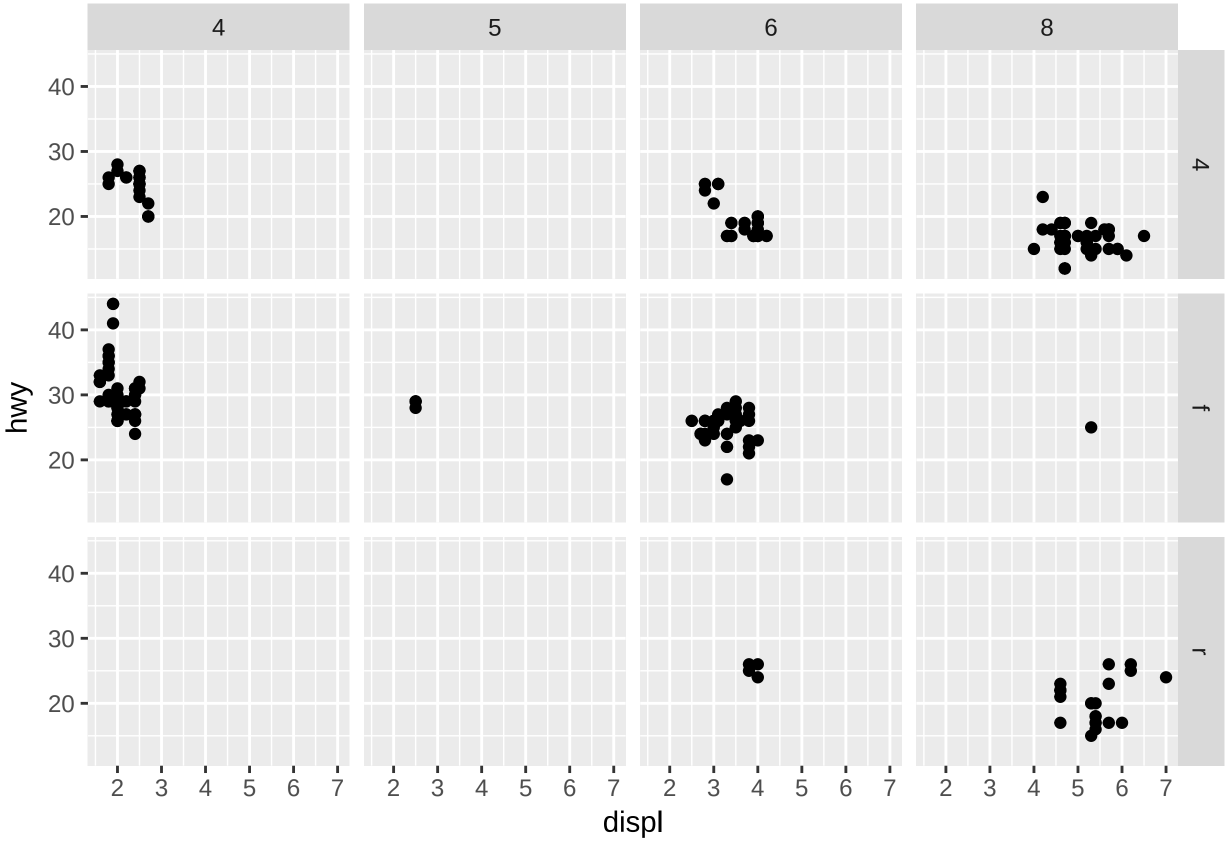
Visualize the plot using Facets, Using this we can include additional varaible into the plot particularly categorical variables

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



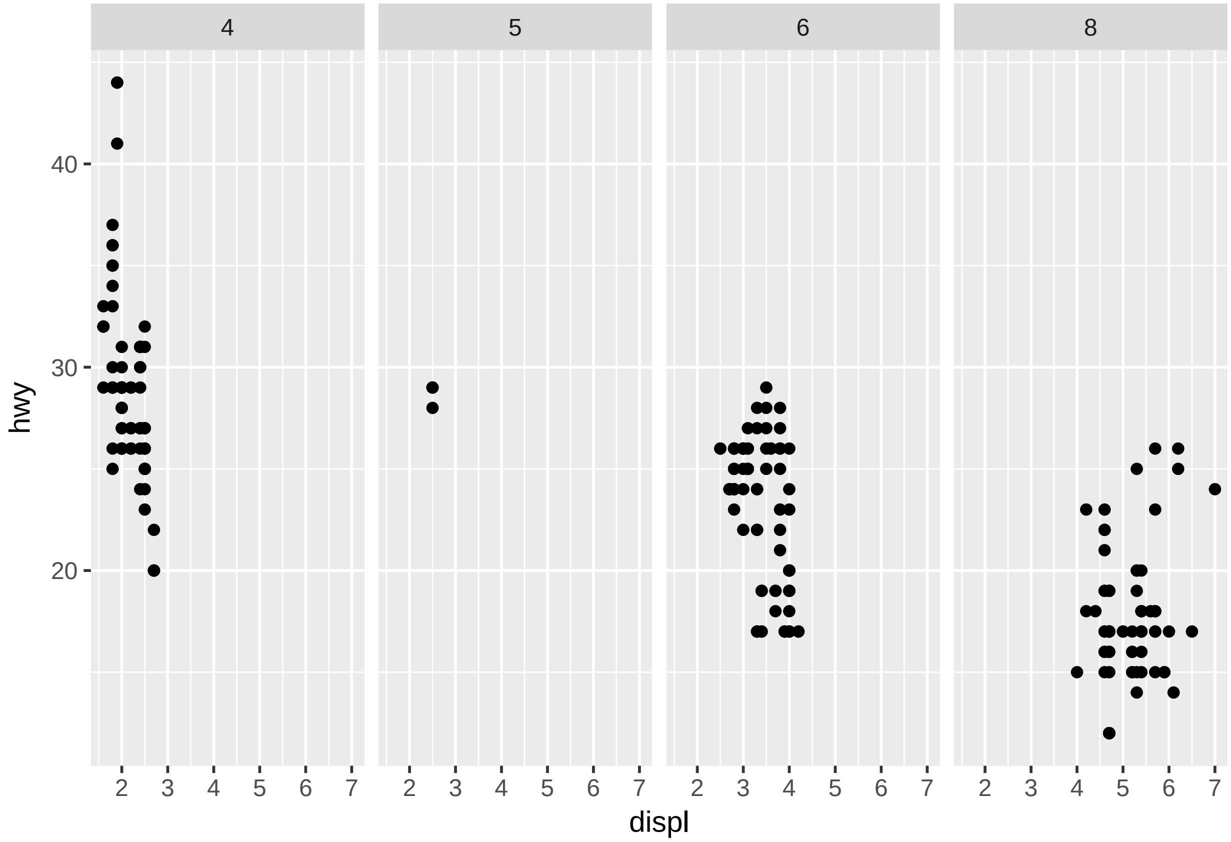
Facet plotting on the combination of two variables

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



Not to facet on rows or column, use . instead of variable name as below

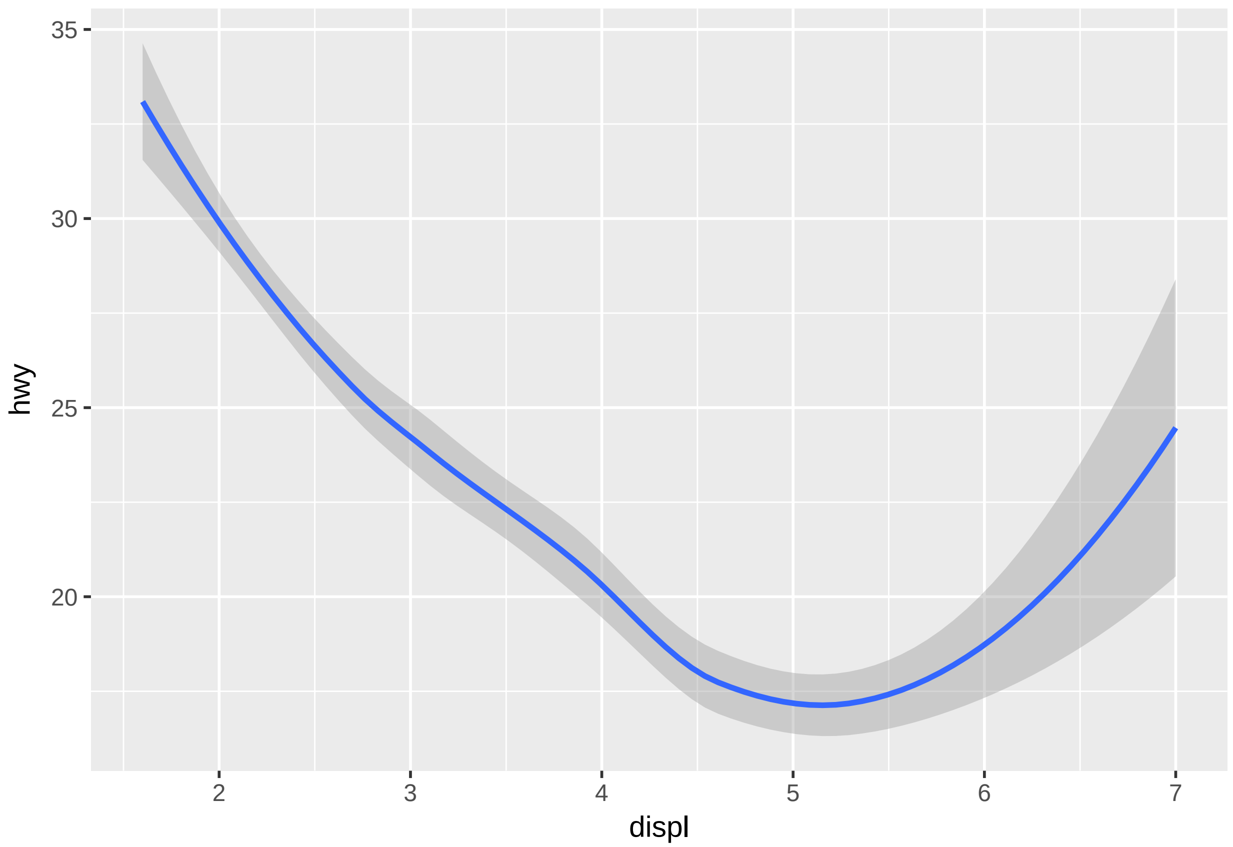
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_grid(. ~ cyl)
```



## Using Geometric Objects in ggplot

```
ggplot(data=mpg) +
  geom_smooth(aes(x=displ, y=hwy))
```

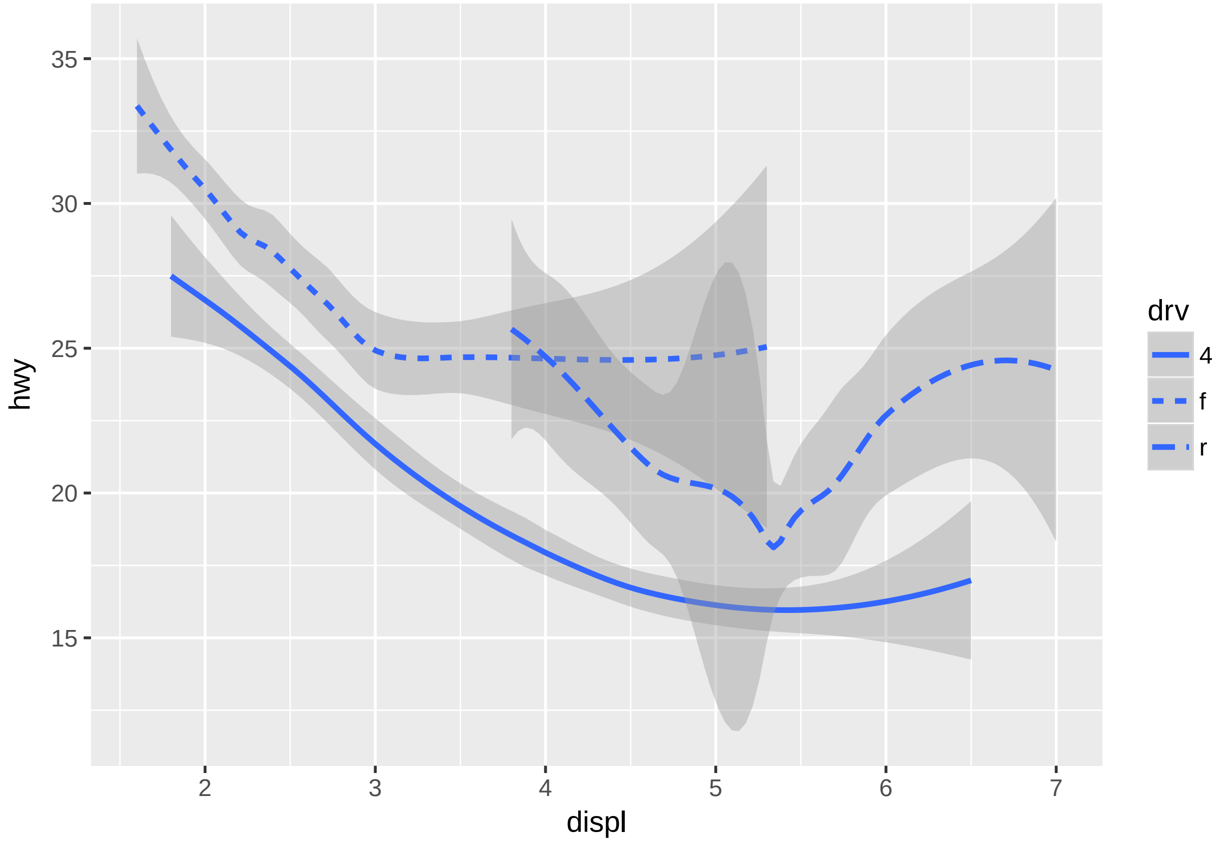
```
## `geom_smooth()` using method = 'loess'
```



Adding categorical varaible (drv) in this plot, This categorical variable says that * ‘f’ - front-wheel drive * ‘r’ - rear wheel drive * ‘4’ - four wheel drive

```
ggplot(data = mpg) +
  geom_smooth(mapping = aes(x = displ, y = hwy, linetype=drv ))
```

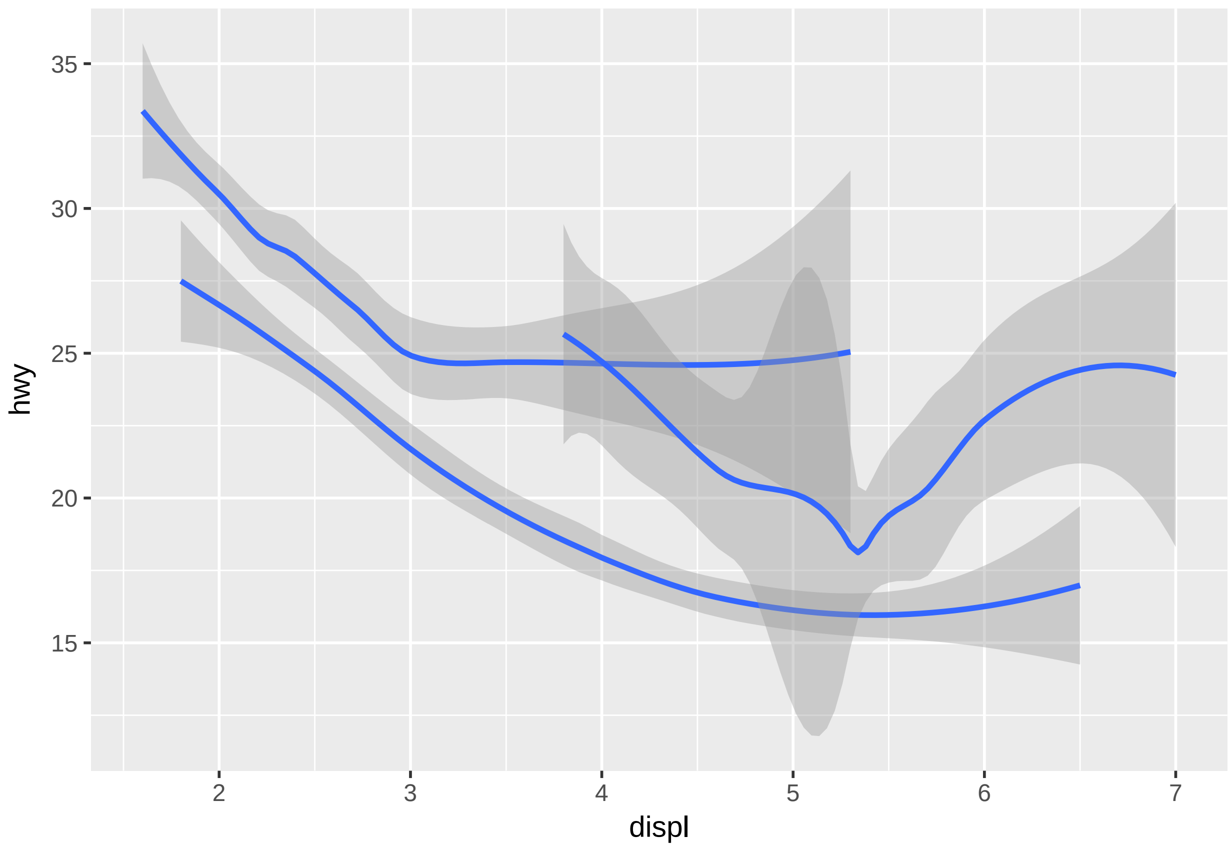
```
## `geom_smooth()` using method = 'loess'
```



Same plot can be achieved using aesthetic parma ‘group’ instead of ‘linetype’. But ‘group’ will not add Legend to the plot

```
ggplot(data = mpg) +
  geom_smooth(mapping = aes(x = displ, y = hwy, group=drv ))
```

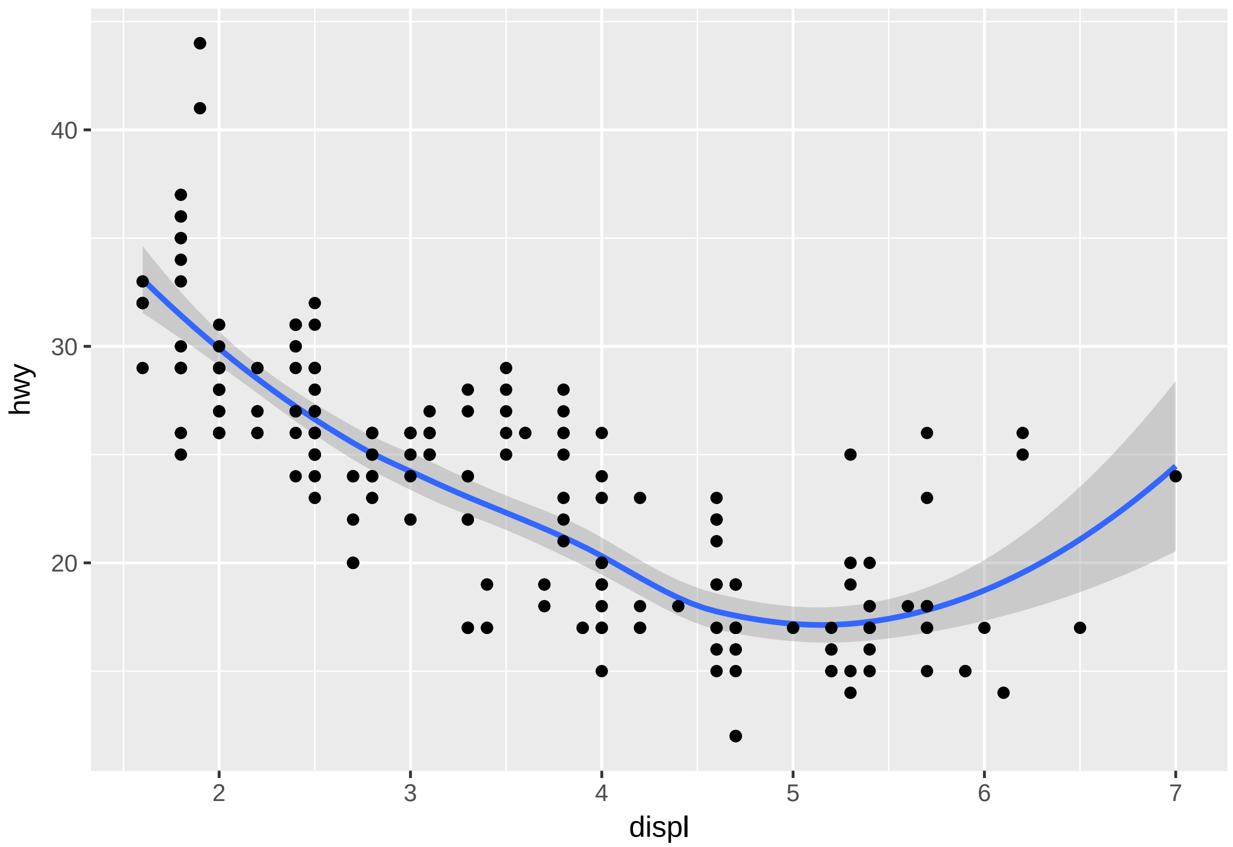
```
## `geom_smooth()` using method = 'loess'
```



Using Multiple geom object to the plot to create more meaning full visualization.

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

```
## `geom_smooth()` using method = 'loess'
```


40
30
20
hwy
2
3
4
5
6
7
displ