Revised: 2/15/2021

Data Visualization I

Michael C. Hackett

Assistant Professor, Computer Science

Community College of Philadelphia

ggplot2

- ggplot2 is a data visualization package in R
- Allows for declaratively creating graphics
  - Based on the text [The Grammar of Graphics]
- *"You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details."*
  - Project homepage: https://ggplot2.tidyverse.org/

ggplot2

- ggplot2 is installed along with the tidyverse:

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

- Can be installed as a stand-alone package:

```
install.packages("ggplot2")
```

- Extensions:

  https://exts.ggplot2.tidyverse.org/gallery/

ggplot2

- ggplot2 is loaded along with the rest of the tidyverse:

```
library(tidyverse)
```

- Can be loaded by itself:

```
library(ggplot2)
```

- ggplot2 has a sample data frame for demonstration purposes
  - The **mpg** dataset contains observations collected by the US Environmental Protection Agency on 38 models of cars

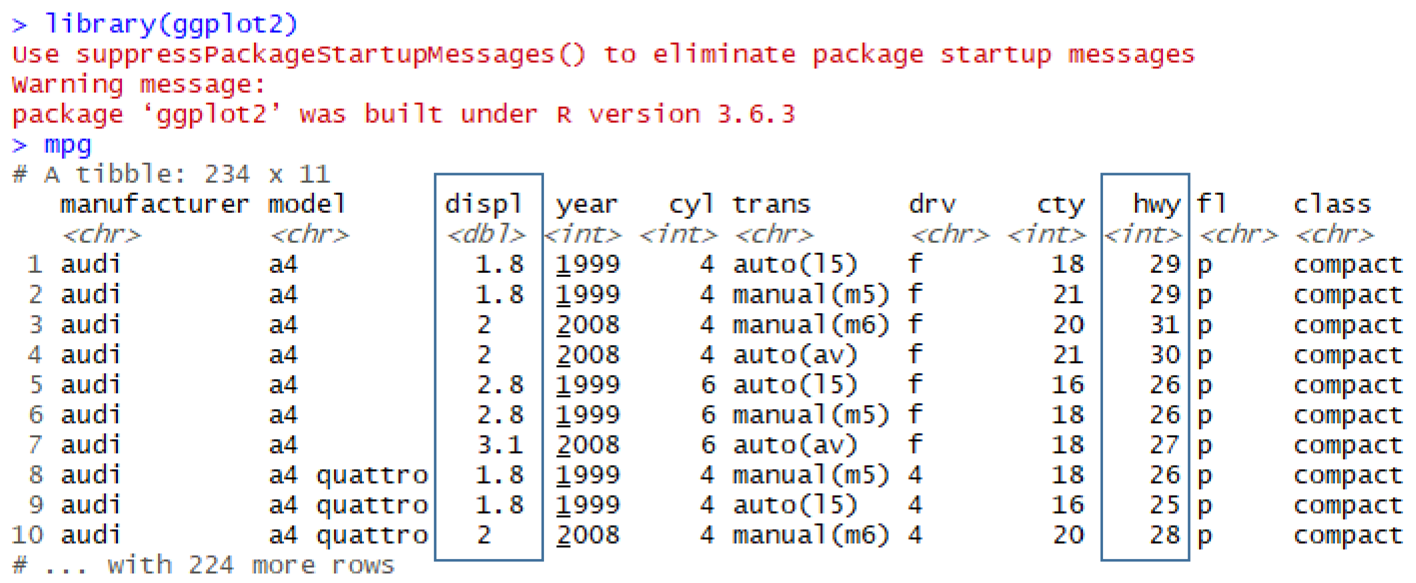
ggplot2

- If tidyverse was loaded:
  ```
  library(tidyverse)
  ggplot2::mpg
  ```

- If ggplot2 was loaded by itself:
  ```
  library(ggplot2)
  mpg
  ```

- We'll assume ggplot2 was loaded by itself for the remainder of the lecture

ggplot2



```
> library(ggplot2)
Use suppressPackageStartupMessages() to eliminate package startup messages
Warning message:
package ‘ggplot2’ was built under R version 3.6.3
> mpg
# A tibble: 234 x 11
   manufacturer model      displ  year   cyl trans      drv     cty   hwy fl    class  
   <chr>        <chr>      <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>  
 1 audi         a4           1.8  1999     4 auto(l5)   f        18    29 p     compact
 2 audi         a4           1.8  1999     4 manual(m5) f        21    29 p     compact
 3 audi         a4           2    2008     4 manual(m6) f        20    31 p     compact
 4 audi         a4           2    2008     4 auto(av)   f        21    30 p     compact
 5 audi         a4           2.8  1999     6 auto(l5)   f        16    26 p     compact
 6 audi         a4           2.8  1999     6 manual(m5) f        18    26 p     compact
 7 audi         a4           3.1  2008     6 auto(av)   f        18    27 p     compact
 8 audi         a4 quattro   1.8  1999     4 manual(m5) 4        18    26 p     compact
 9 audi         a4 quattro   1.8  1999     4 auto(l5)   4        16    25 p     compact
10 audi         a4 quattro   2    2008     4 manual(m6) 4        20    28 p     compact
# ... with 224 more rows
```

- The **displ** column is the engine size (in liters)
- The **hwy** column is the highway gas milage in miles-per-gallon

ggplot2

- We begin creating a plot with the **`ggplot()`** function
  - This creates a coordinate system that layers can be added on to

- The first argument to the ggplot function is the data we wish to plot

```
ggplot(data = mpg)
```

- Now that the plot has its data, layers are added that specify how to data is to be displayed.
  - Layers of data are referred to as *geometries* or ***geoms***

ggplot2

- ggplot2’s **`geom_point()`** function adds a layer of points to a plot
  - Effectively creating a scatterplot

- The first argument to the geom_point function is the mapping
  - Will define how variables are mapped to different aesthetics or visual properties on this layer

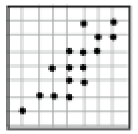
```
geom_point(mapping = aes())
```

  - The aes function shown here specifies the aesthetics of the layer

ggplot2 - Aesthetics

- The x and y arguments to the aes function specify which columns to use for the x and y axes of our scatterplot

```
geom_point(mapping = aes(x = displ , y = hwy))
```



f + geom_point()
x, y, alpha, color, fill, shape, size

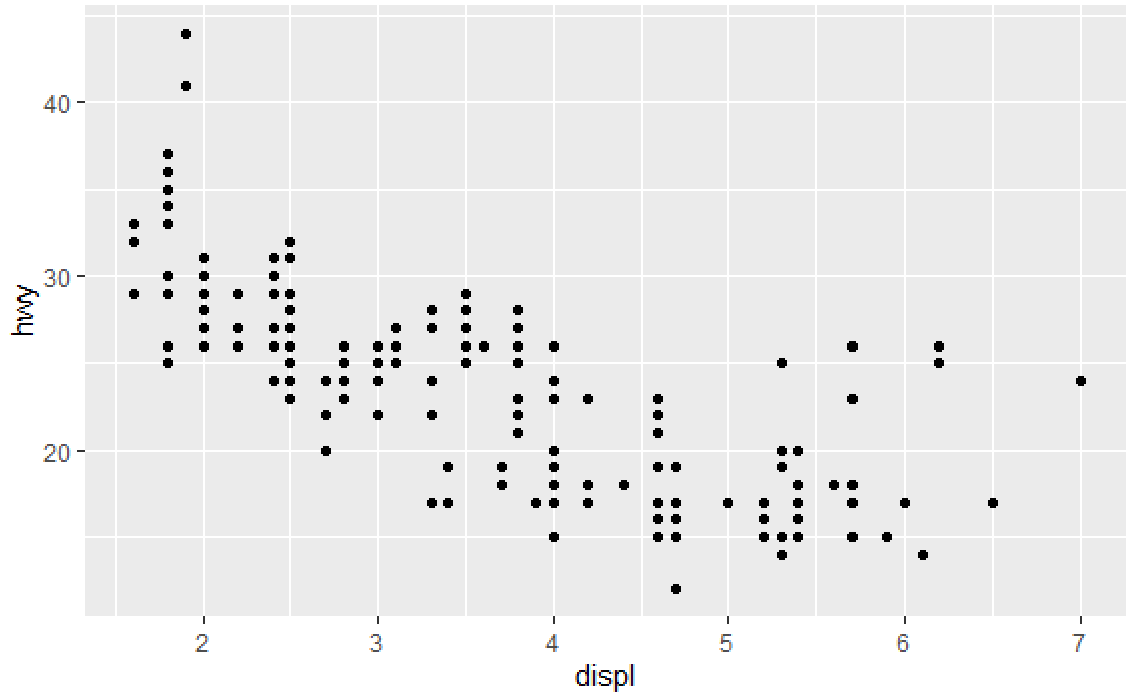
Aesthetics for geom_point from cheat sheet

- Now, we simply add this layer to the plot:

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

ggplot2 - Aesthetics

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

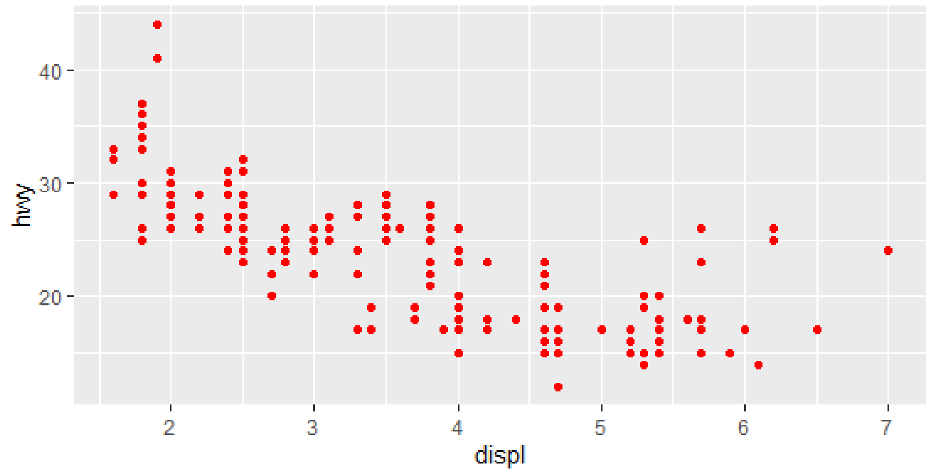


ggplot2 - Aesthetics

- The color argument to the geom_point function will specify the color of all the data points

```
geom_point(mapping = aes(x = displ , y = hwy), color = "red")
```

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



ggplot2 - Aesthetics

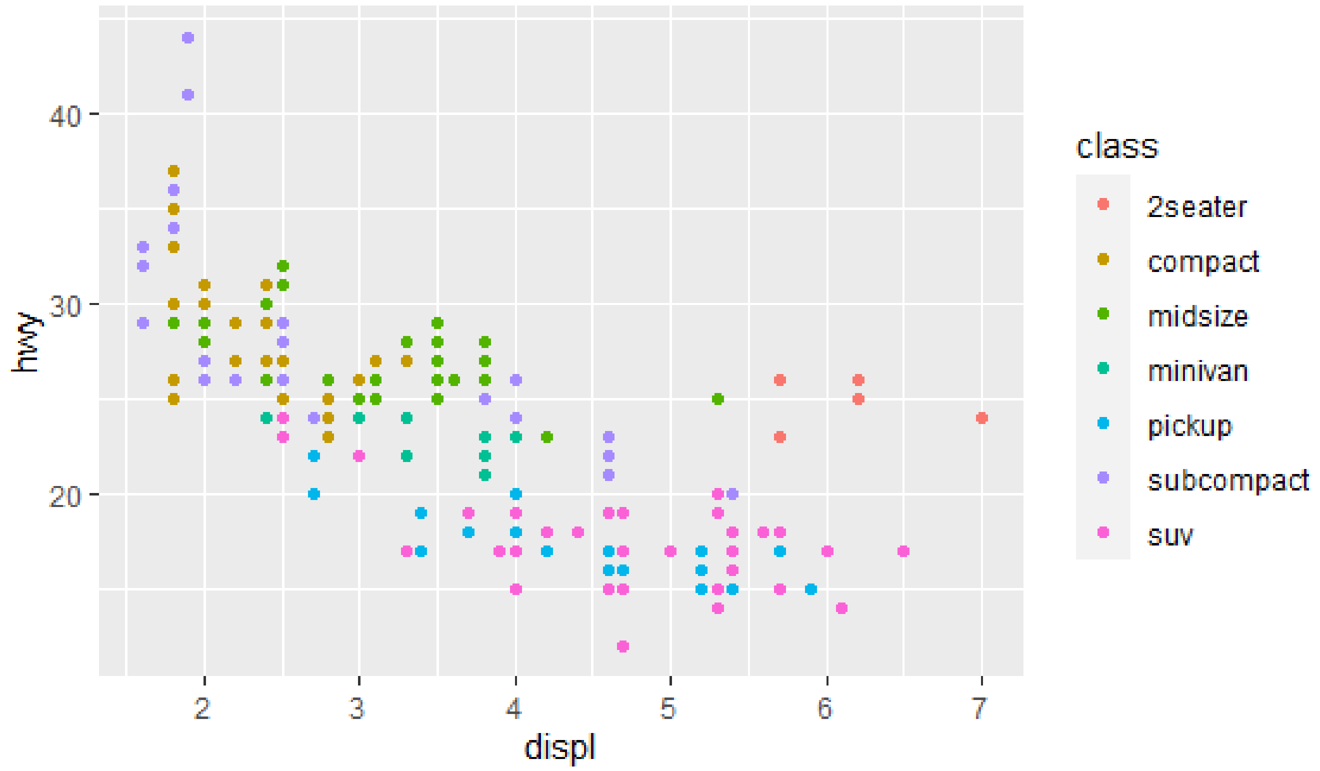
- The color argument to the *aes function* will specify how to color the individual data points

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

  - This will apply a unique color to each point, based on the class column of each observation

ggplot2 - Aesthetics

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



ggplot2 - Aesthetics

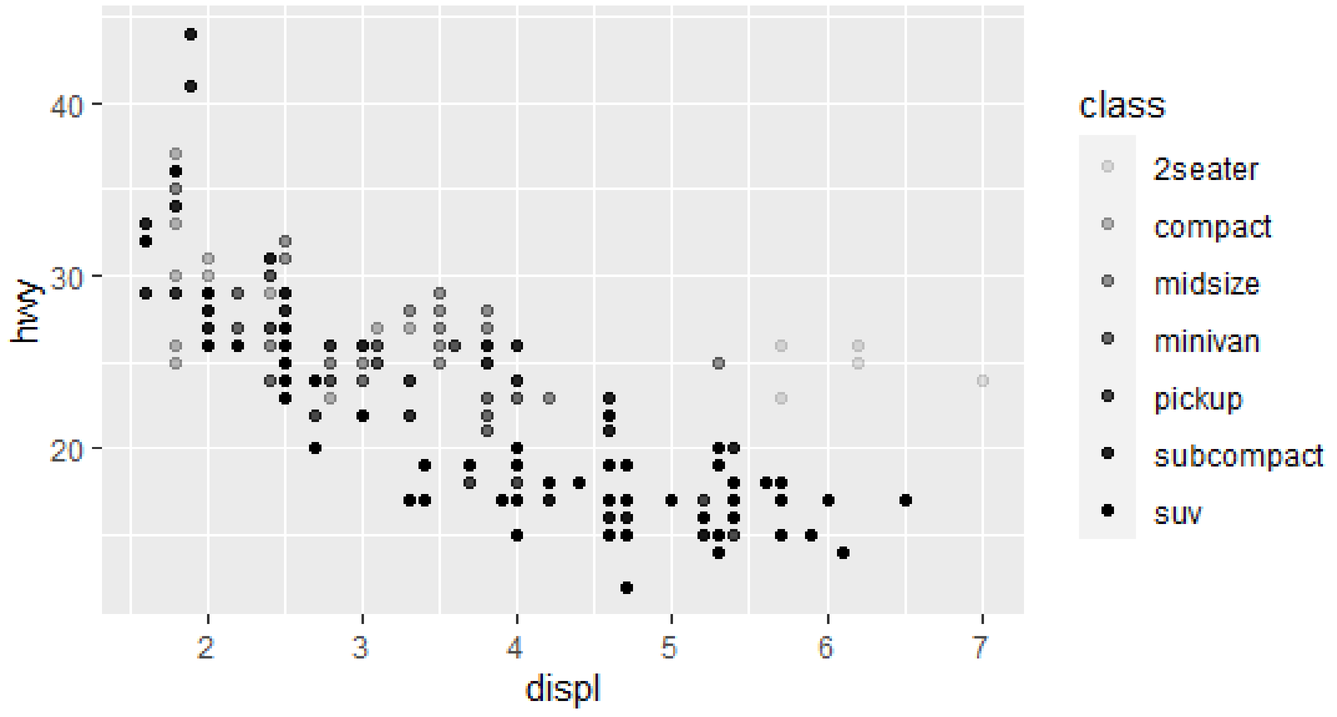
- The alpha argument to the aes function will specify the opacity of the data points

```
geom_point(mapping = aes(x = displ , y = hwy, alpha = class))
```

  - This will apply an opacity to each point, based on the class column of each observation

ggplot2 - Aesthetics

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



ggplot2 - Aesthetics

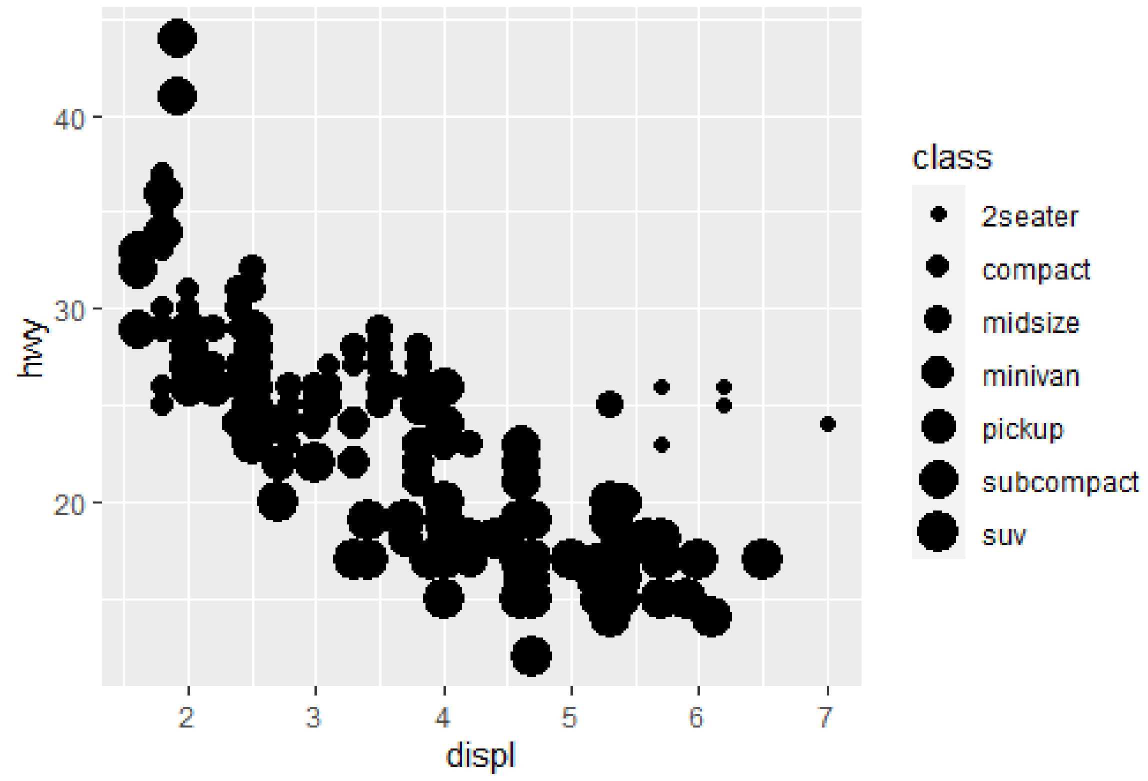
- The size argument to the aes function will specify the size of the data points

  ```
  geom_point(mapping = aes(x = displ , y = hwy, size = class))
  ```

  - This will apply a size to each point, based on the class column of each observation

ggplot2 - Aesthetics

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



ggplot2 - Aesthetics

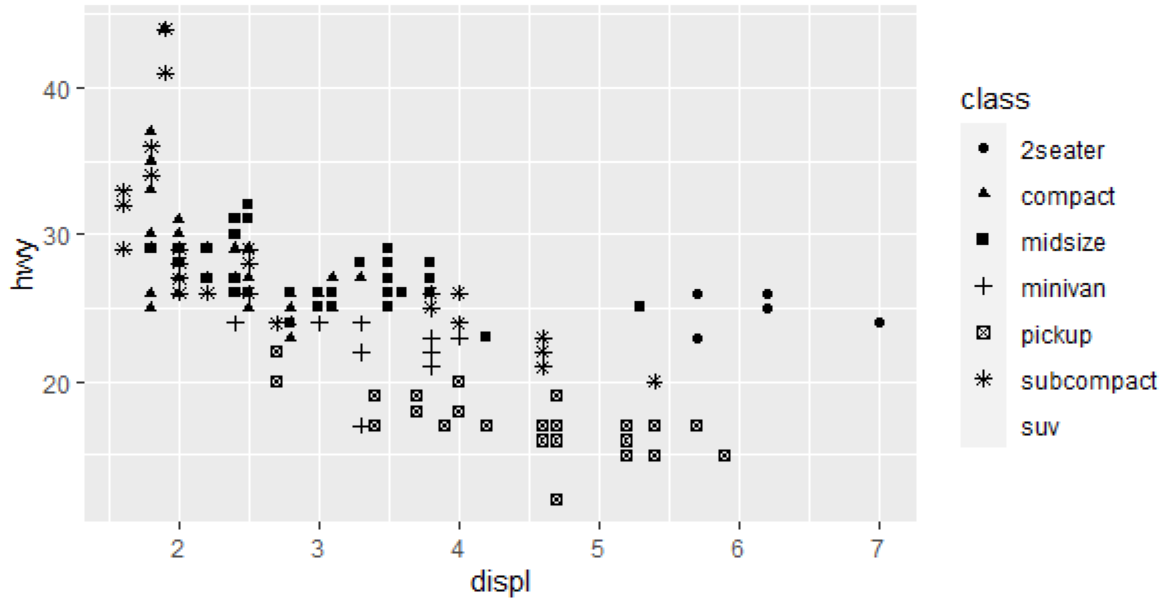
- The shape argument to the aes function will specify the shapes of the data points

```
geom_point(mapping = aes(x = displ , y = hwy, shape = class))
```

  - This will apply a shape to each point, based on the class column of each observation

ggplot2 - Aesthetics

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

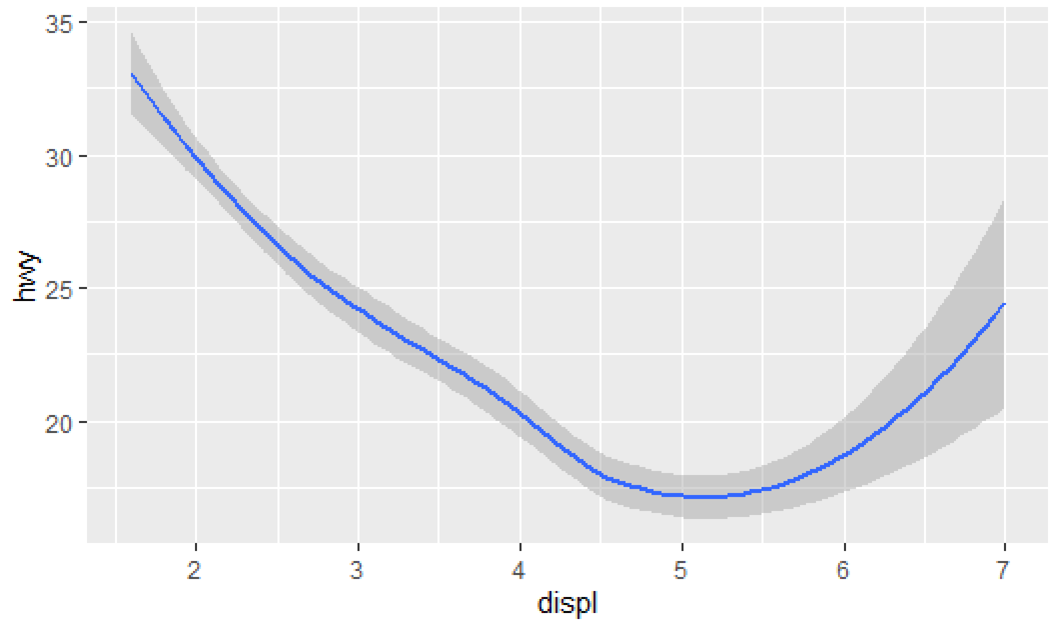


- Note: Only displays up to 6 shapes (hence why SUV is not shown)

ggplot2 - Geoms

- The geom_smooth function plots a layer of a smooth, fitted line

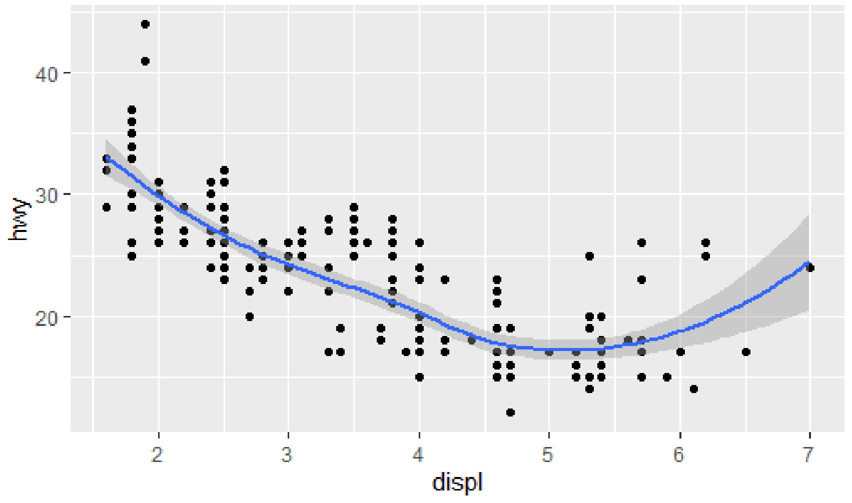
```
ggplot(data = mpg) + geom_smooth(mapping = aes(x = displ , y = hwy))
```



ggplot2 - Geoms

- The layer can be plotted on top of other layers

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

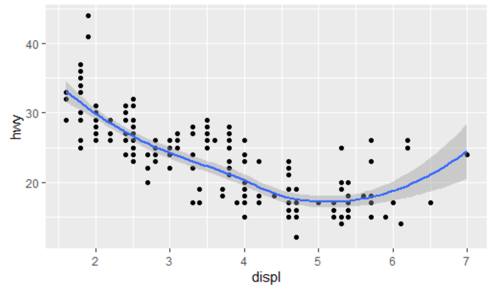


ggplot2 - Geoms

- There is some redundancy in the last slide
  - Both layers needed to be told the x and y values.

- Instead, the mapping can be moved into the original ggplot
  - Now, the data and a default mapping/aesthetic is passed up to any added layers

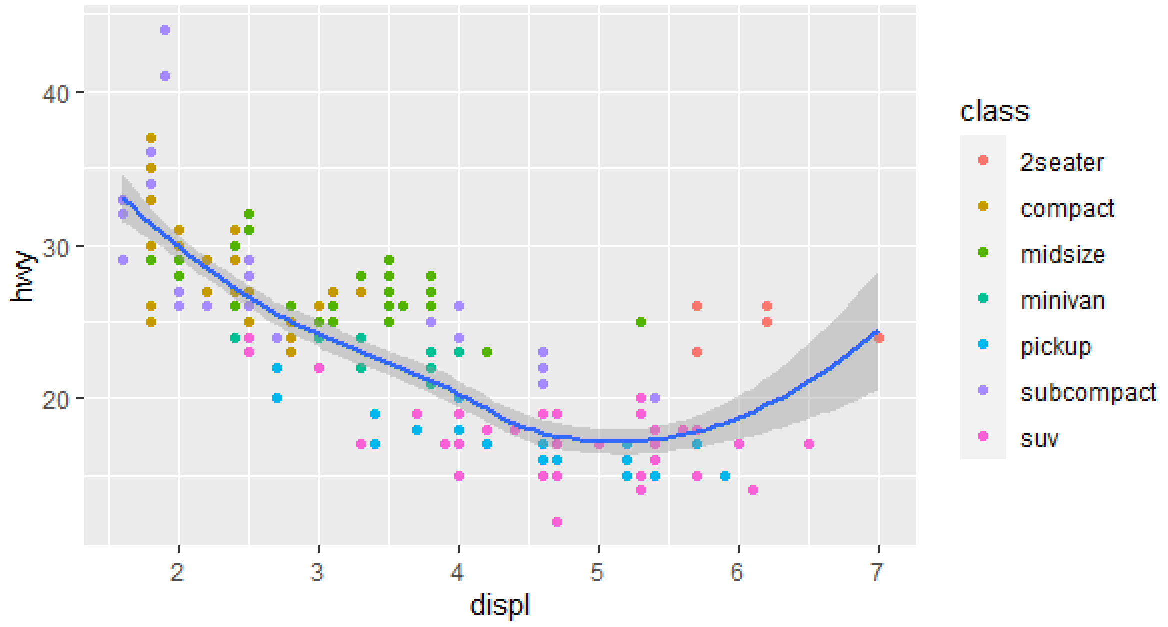
```
ggplot(data = mpg, mapping = aes(x = displ , y = hwy)) +
geom_point() +
geom_smooth()
```



(Same plot)

ggplot2 - Geoms

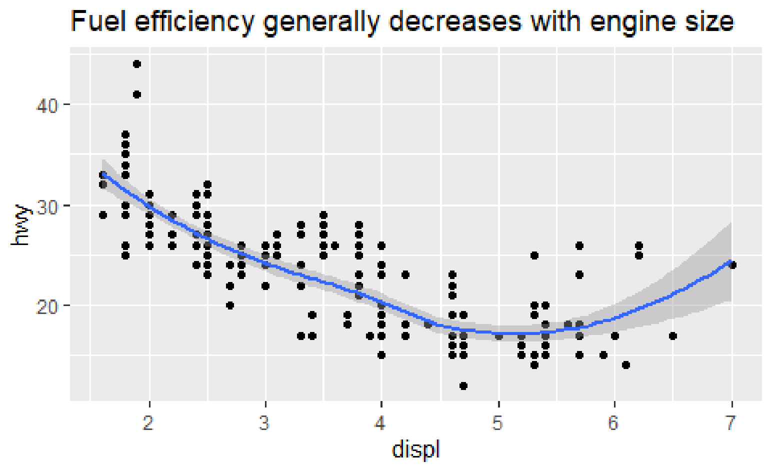
```
ggplot(data = mpg, mapping = aes(x = displ , y = hwy)) +
geom_point(mapping = aes(color = class)) +
geom_smooth()
```



ggplot2 - Labels

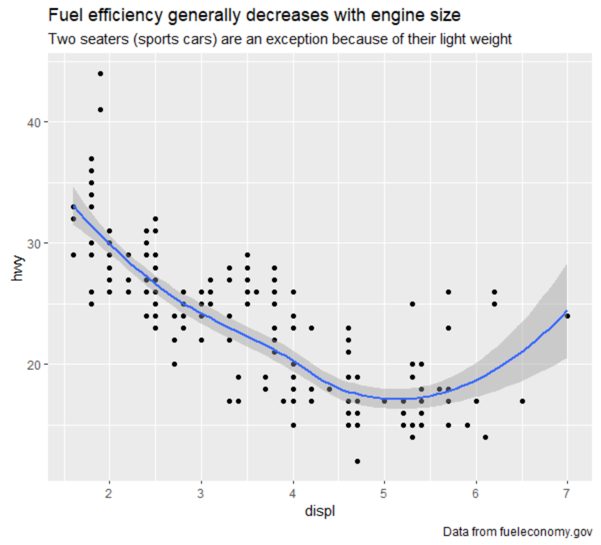
- Labels such as title, subtitle, captions, and axis labels can be added using ggplot2's lab function

```
ggplot(data = mpg, mapping = aes(x = displ , y = hwy)) +
geom_point() +
geom_smooth() +
labs(title = "Fuel efficiency generally decreases with engine size")
```



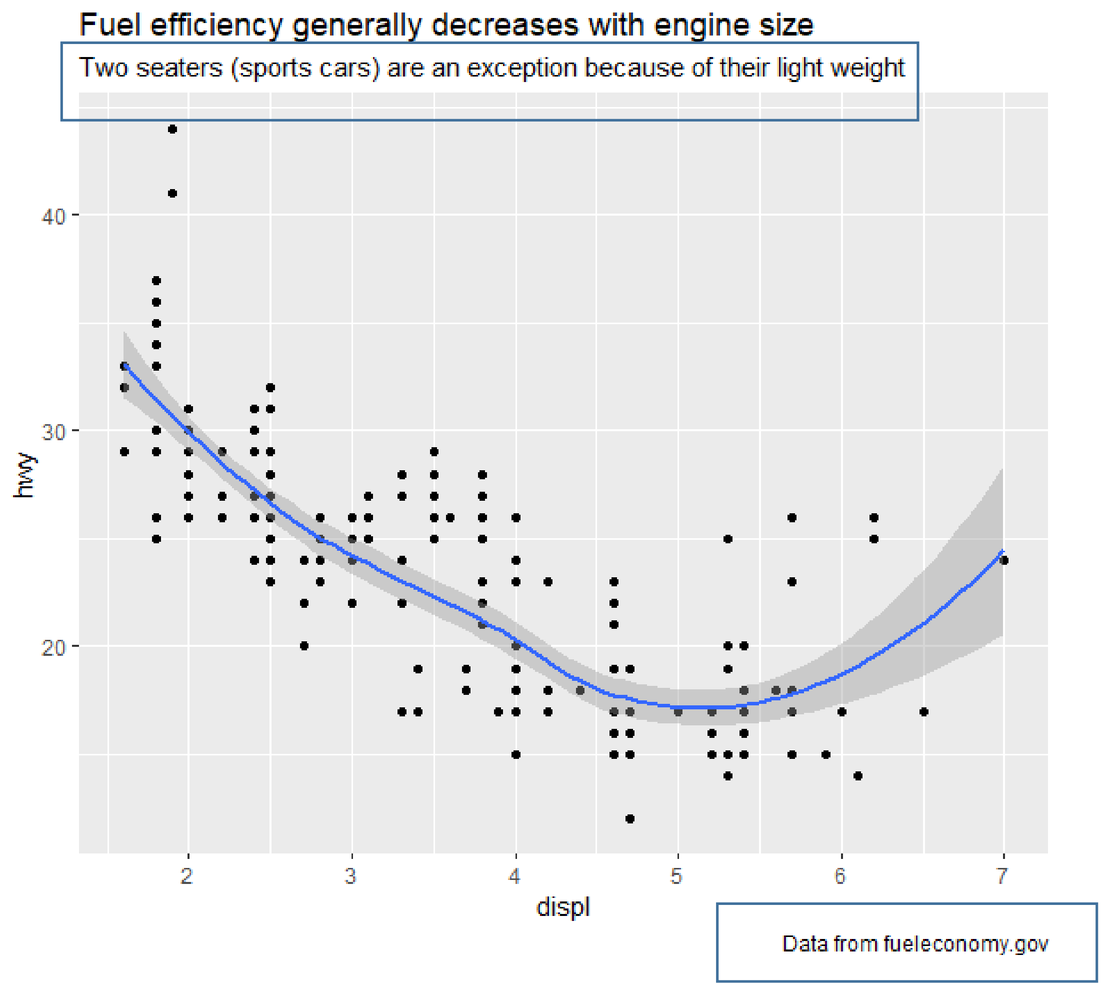
ggplot2 - Labels

```
ggplot(data = mpg, mapping = aes(x = displ , y = hwy)) + geom_point() + geom_smooth() +
labs(title = "Fuel efficiency generally decreases with engine size",
     subtitle = "Two seaters (sports cars) are an exception because of their light weight",
     caption = "Data from fueleconomy.gov")
```



(Shown larger on next slide)

ggplot2 - Labels

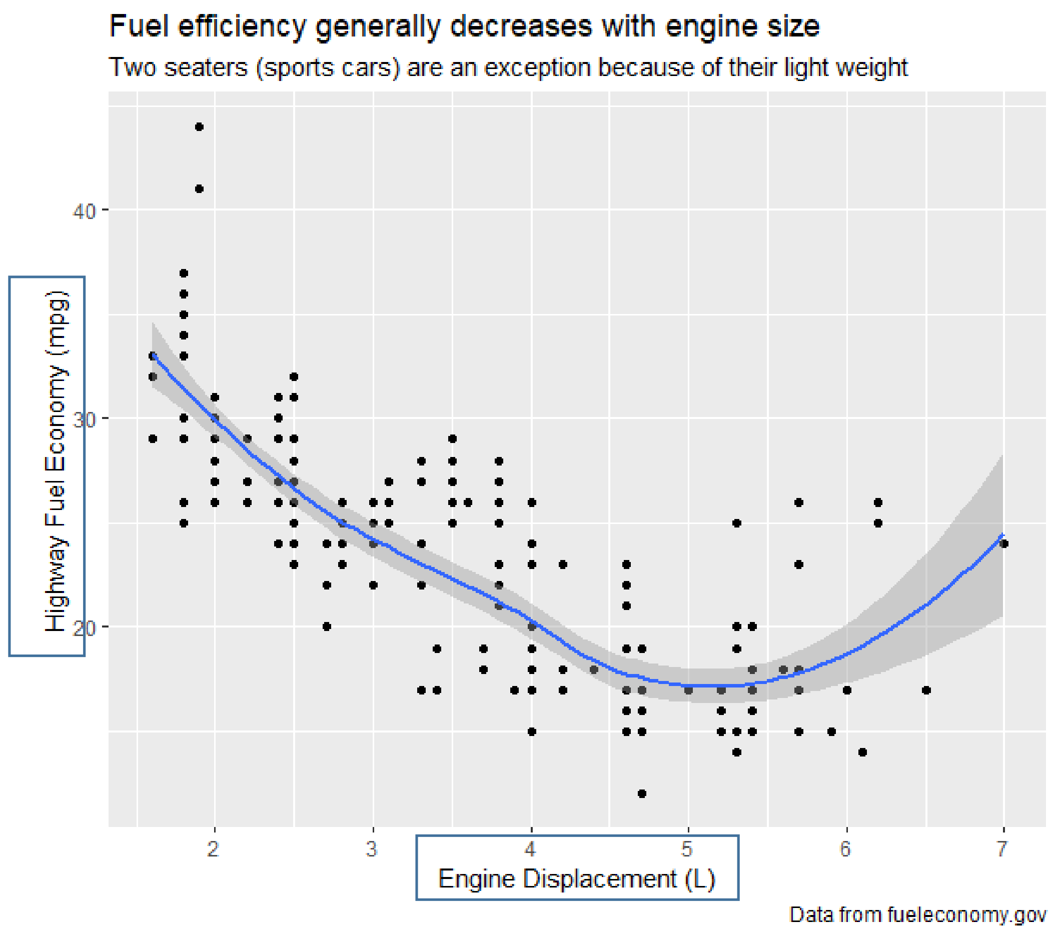


ggplot2 - Labels

```
ggplot(data = mpg, mapping = aes(x = displ , y = hwy)) + geom_point() + geom_smooth() +
labs(title = "Fuel efficiency generally decreases with engine size",
     subtitle = "Two seaters (sports cars) are an exception because of their light weight",
     caption = "Data from fueleconomy.gov",
     x = "Engine Displacement (L)", y = "Highway Fuel Economy (mpg)")
```

(Shown on next slide)

ggplot2 - Labels



ggplot2 - Scales

- Scales in ggplot2 control how a plot maps data values to the visual values of an aesthetic.

- The general format of a scale function is

```
scale_*_function()
```

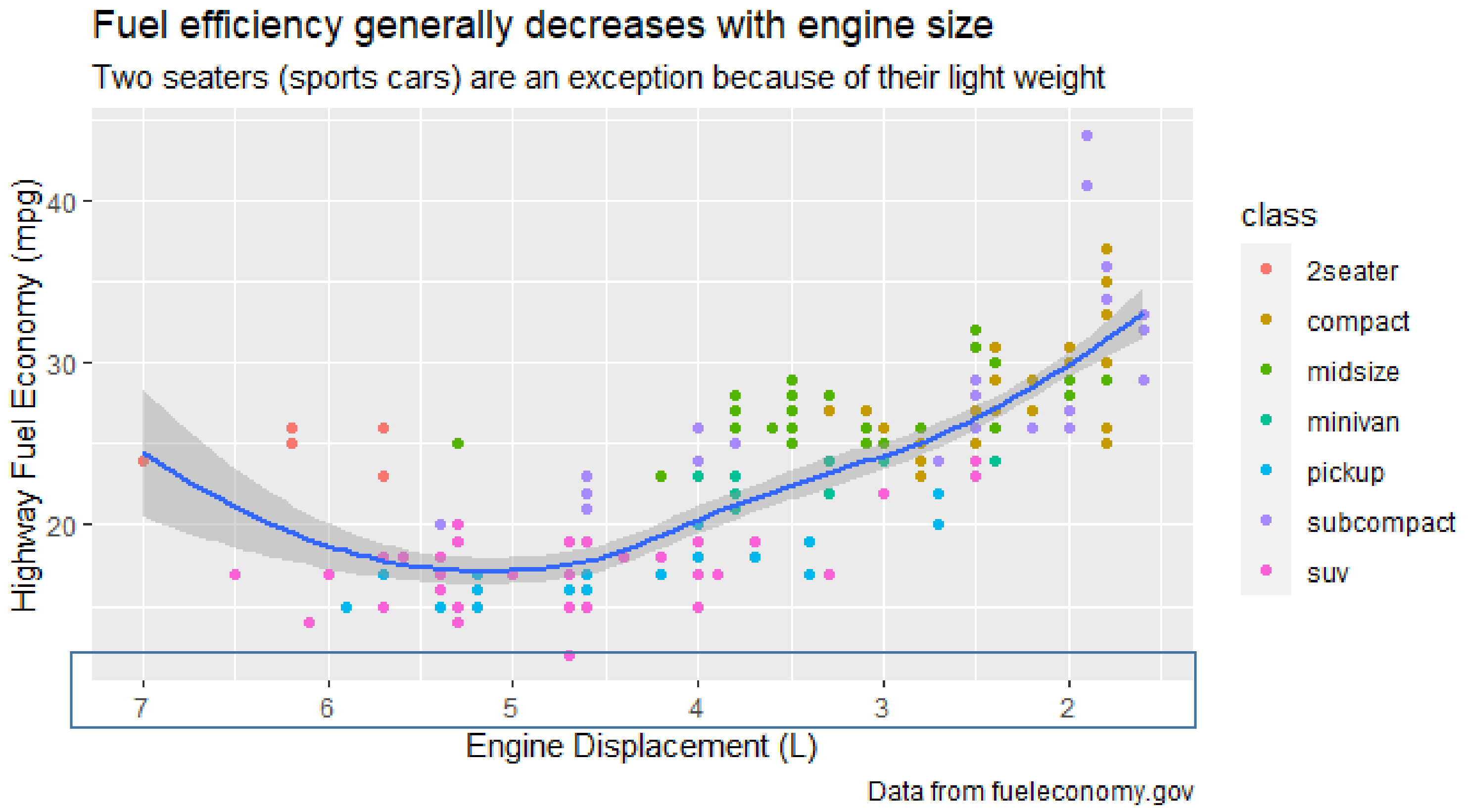
- For example, **`scale_x_reverse`** will reverse the x axis

ggplot2 - Scales

```
ggplot(data = mpg, mapping = aes(x = displ , y = hwy)) +
  geom_point() +
  geom_smooth() +
  labs(title = "Fuel efficiency generally decreases with engine size",
     subtitle = "Two seaters (sports cars) are an exception because of their light weight",
     caption = "Data from fueleconomy.gov",
     x = "Engine Displacement (L)", y = "Highway Fuel Economy (mpg)") +
  scale_x_reverse()
```

(Shown on next slide)

ggplot2 - Scales



(Not really helpful for this graph, but it demonstrates a scale in gglplot2)

ggplot2 - Scales

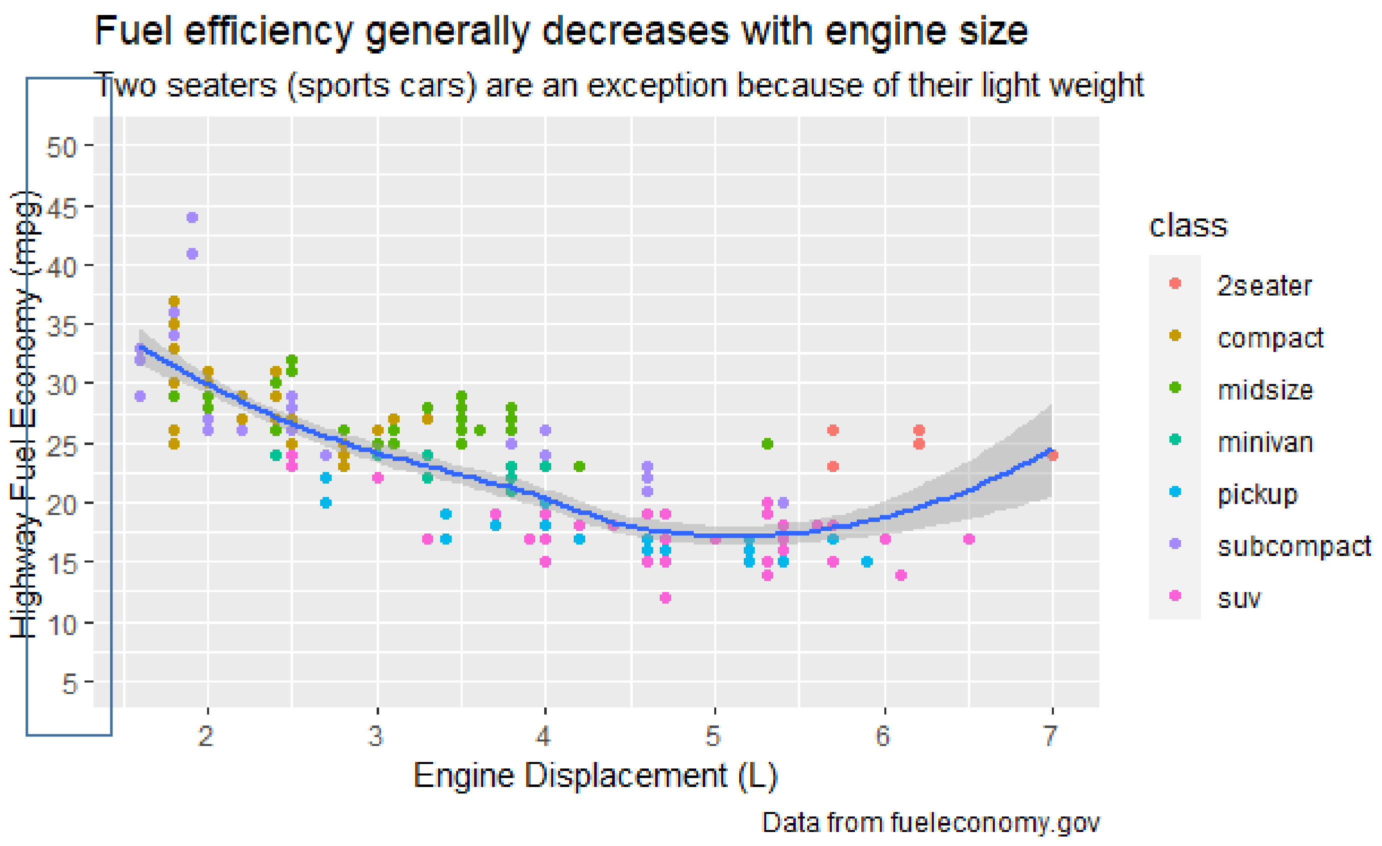
- A scale can be used to adjust the tick marks for the scatterplot.
- The **`scale_y_continuous`** function will allow us to change the breaks and limits for the y axis
  - **`scale_x_continuous`** would do the same for the x axis
  - Use **`scale_x_discrete`** and **`scale_y_discrete`** for axes that use discrete variables (like text)
- The **`breaks`** argument is a vector that specifies the labels
- The **`limits`** argument is a vector that specifies the axis' range.

ggplot2 - Scales

```
ggplot(data = mpg, mapping = aes(x = displ , y = hwy)) +
  geom_point() +
  geom_smooth() +
  labs(title = "Fuel efficiency generally decreases with engine size",
     subtitle = "Two seaters (sports cars) are an exception because of their light weight",
     caption = "Data from fueleconomy.gov",
     x = "Engine Displacement (L)", y = "Highway Fuel Economy (mpg)") +
  scale_y_continuous(breaks=seq(0, 50, 5), limits=c(5,50))
```

(Shown on next slide)

ggplot2 - Scales

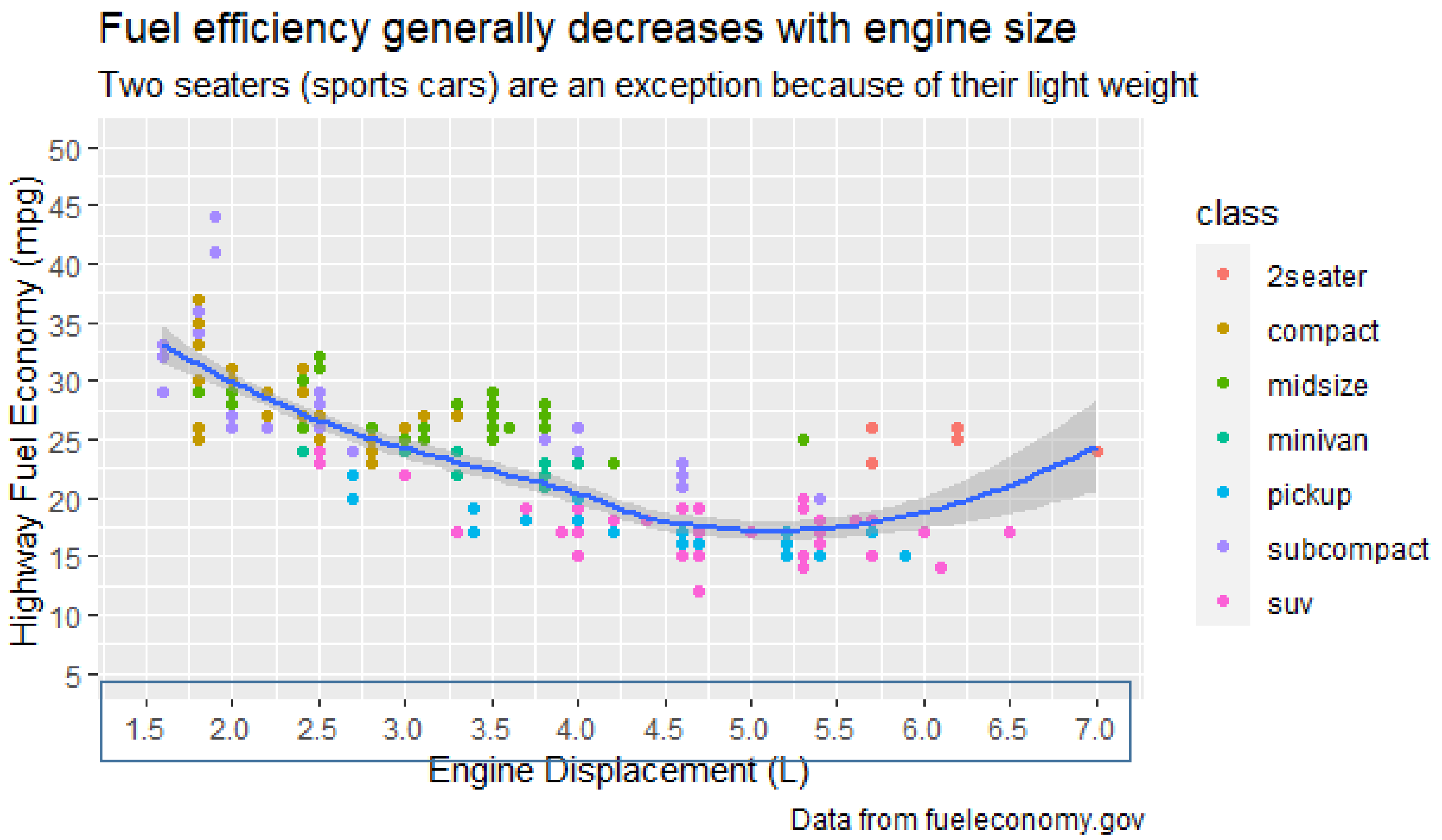


ggplot2 - Scales

```
ggplot(data = mpg, mapping = aes(x = displ , y = hwy)) +
  geom_point() +
  geom_smooth() +
  labs(title = "Fuel efficiency generally decreases with engine size",
     subtitle = "Two seaters (sports cars) are an exception because of their light weight",
     caption = "Data from fueleconomy.gov",
     x = "Engine Displacement (L)", y = "Highway Fuel Economy (mpg)") +
  scale_y_continuous(breaks=seq(0, 50, 5), limits=c(5,50)) +
  scale_x_continuous(breaks=seq(1.5, 7.1, 0.5), limits=c(1.5, 7))
```

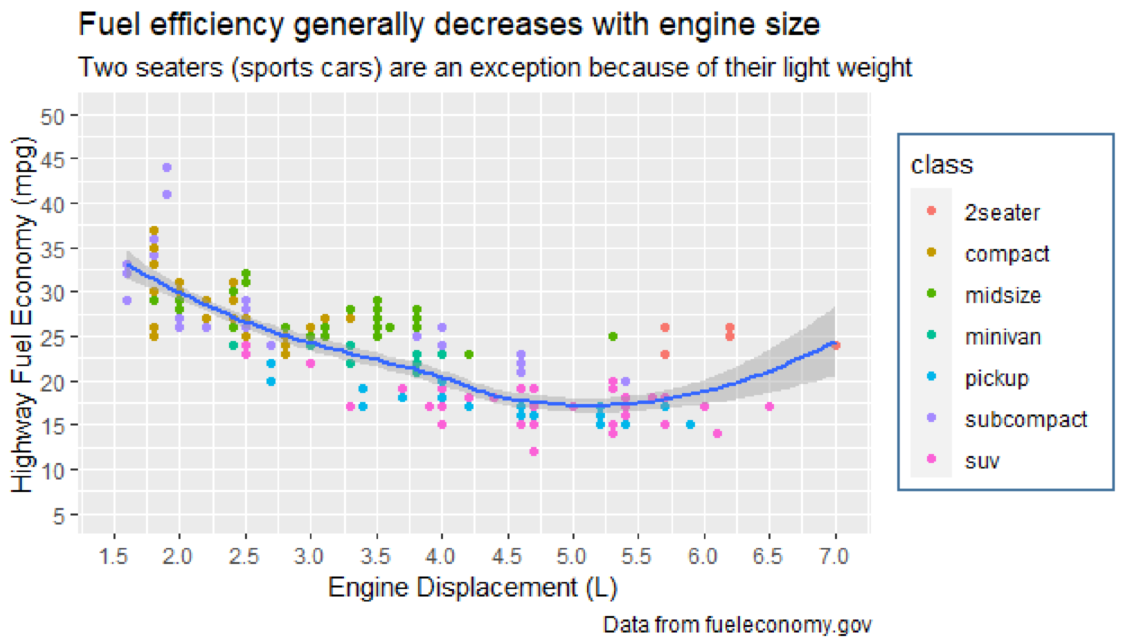
(Shown on next slide)

ggplot2 - Scales



ggplot2 - Legends

- There are a number of ways to control how the legend is displayed.



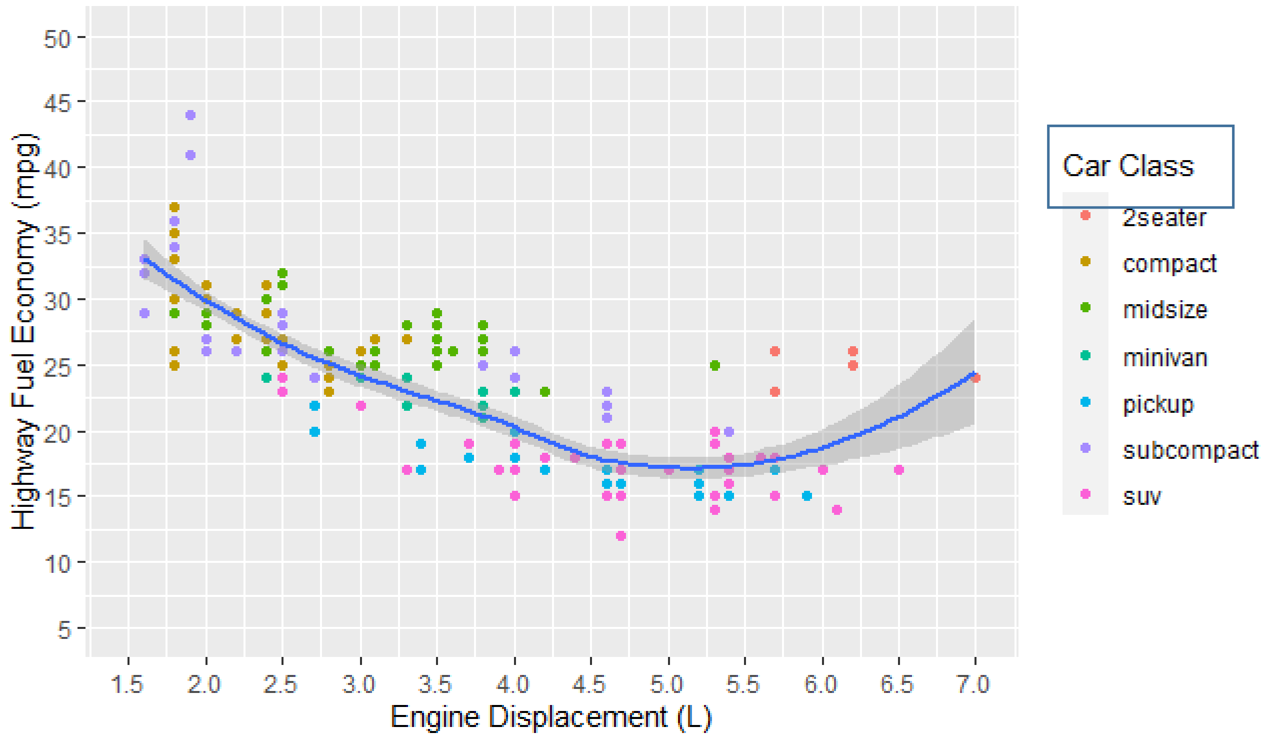
ggplot2 - Legends

- Changing the Legend title:

```
ggplot(data = mpg, mapping = aes(x = displ , y = hwy)) +
  geom_point() +
  geom_smooth() +
  labs(x = "Engine Displacement (L)", y = "Highway Fuel Economy (mpg)",
       color = "Car Class") +
  scale_y_continuous(breaks=seq(0, 50, 5), limits=c(5,50)) +
  scale_x_continuous(breaks=seq(1.5, 7.1, 0.5), limits=c(1.5, 7))
```

(Shown on next slide)

ggplot2 - Legends



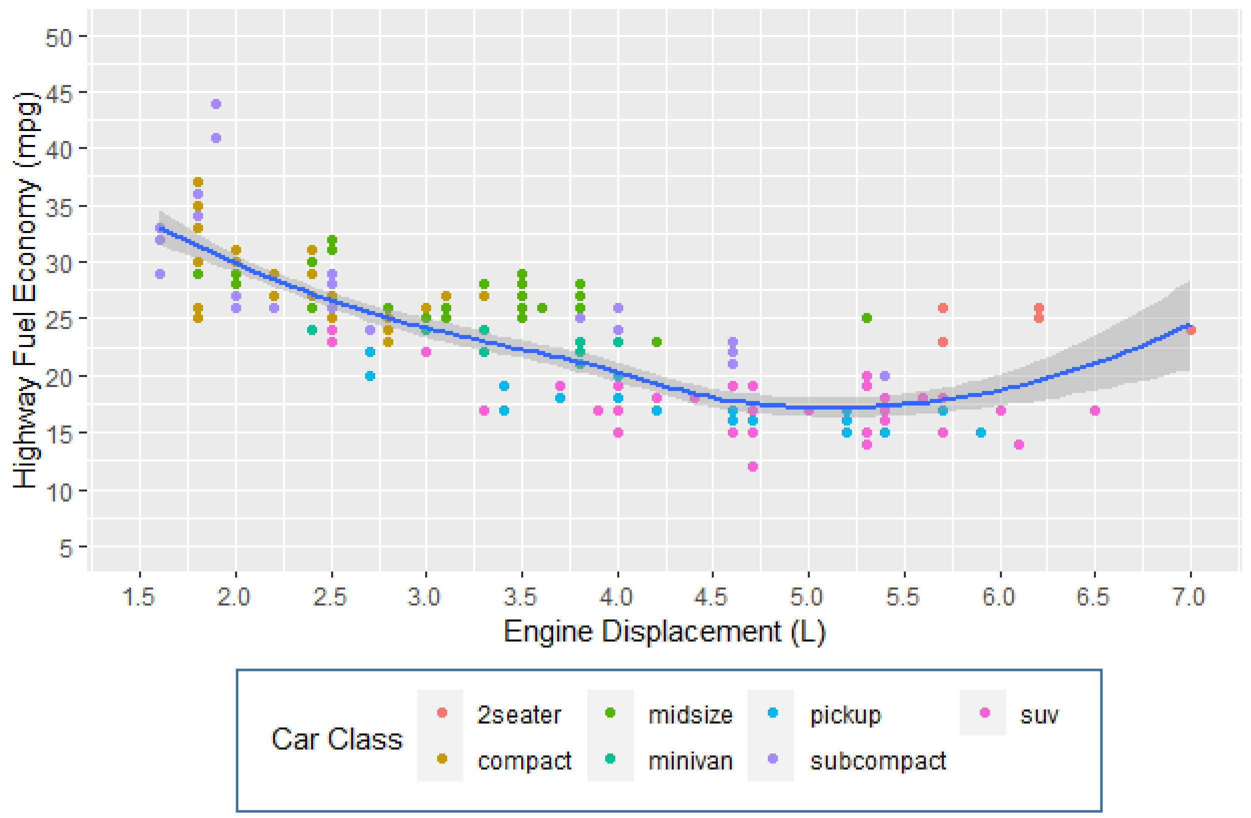
ggplot2 - Legends

- Changing the Legend position:
  - “bottom”, “top”, “left”, or “right”

```
ggplot(data = mpg, mapping = aes(x = displ , y = hwy)) +
  geom_point() +
  geom_smooth() +
  labs(x = "Engine Displacement (L)", y = "Highway Fuel Economy (mpg)",
       color = "Car Class") +
  scale_y_continuous(breaks=seq(0, 50, 5), limits=c(5,50)) +
  scale_x_continuous(breaks=seq(1.5, 7.1, 0.5), limits=c(1.5, 7)) +
  theme(legend.position = "bottom")
```

(Shown on next slide)

ggplot2 - Legends



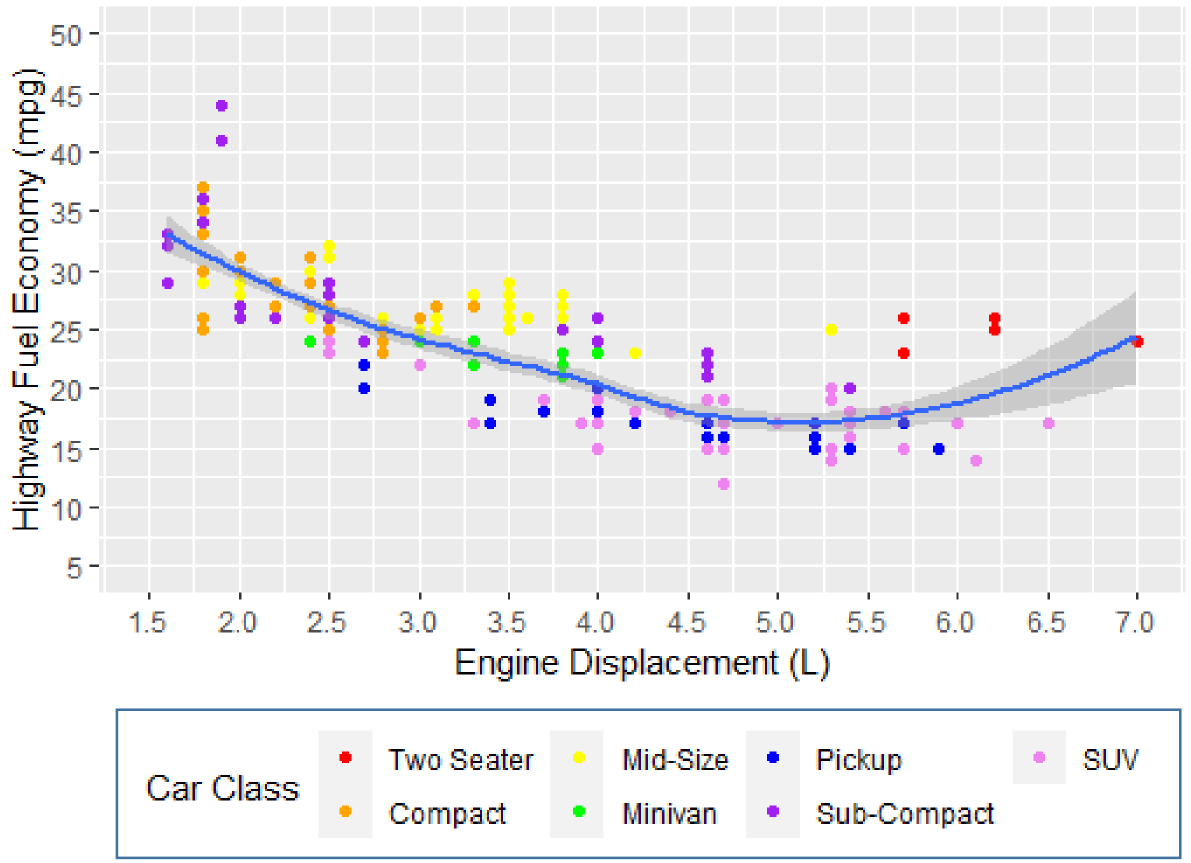
ggplot2 - Legends

- We can change the labels in the legend, but it will also require specifying the colors of each:

```
ggplot(data = mpg, mapping = aes(x = displ , y = hwy)) +
  geom_point() +
  geom_smooth() +
  labs(x = "Engine Displacement (L)", y = "Highway Fuel Economy (mpg)",
       color = "Car Class") +
  scale_y_continuous(breaks=seq(0, 50, 5), limits=c(5,50)) +
  scale_x_continuous(breaks=seq(1.5, 7.1, 0.5), limits=c(1.5, 7)) +
  theme(legend.position = "bottom") +
  scale_color_manual(labels=c("Two Seater", "Compact", "Mid-Size",
                              "Minivan", "Pickup", "Sub-Compact", "SUV"),
                     values=c("red", "orange", "yellow", "green",
                              "blue", "purple", "violet"))
```

(Shown on next slide)

ggplot2 - Legends


Highway Fuel Economy (mpg)
Engine Displacement (L)
Car Class
Two Seater
Compact
Mid-Size
Minivan
Pickup
Sub-Compact
SUV