ggplot2 tutorial, STAT 432

The objectives of this tutorial are to use `ggplot2` to

- create a scatter plot
- color points according to a categorical variable (factor)
- use `facet_wrap()` to form a matrix of scatter plots corresponding to the levels of a categorical variable
- color points according to a continuous variable using a variety of palettes

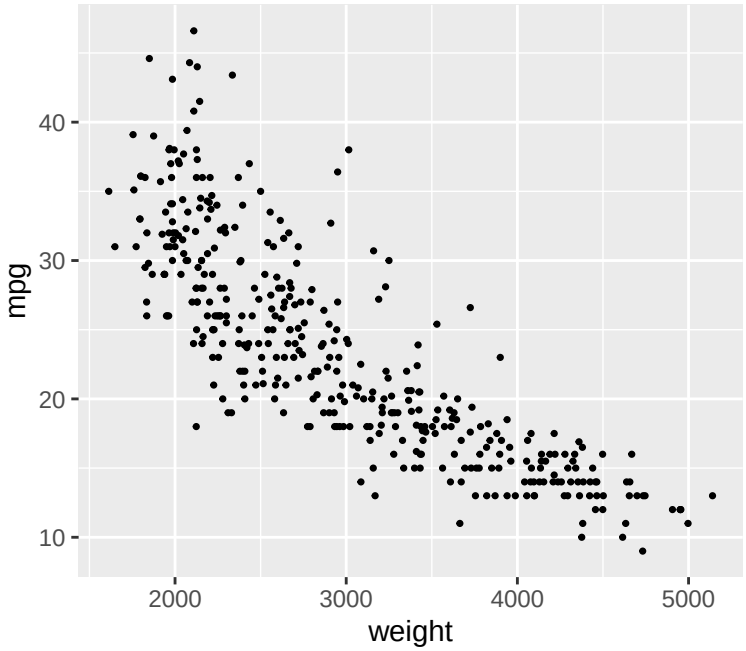
For this demonstration, we will use the `Auto` data set from the ISLR package.

```
library(ISLR)
library(ggplot2)
```

Basic Scatter Plot

Use `geom_point()` to make a scatter plot of `mpg` versus `weight`. The argument `size` is used to adjust the point size.

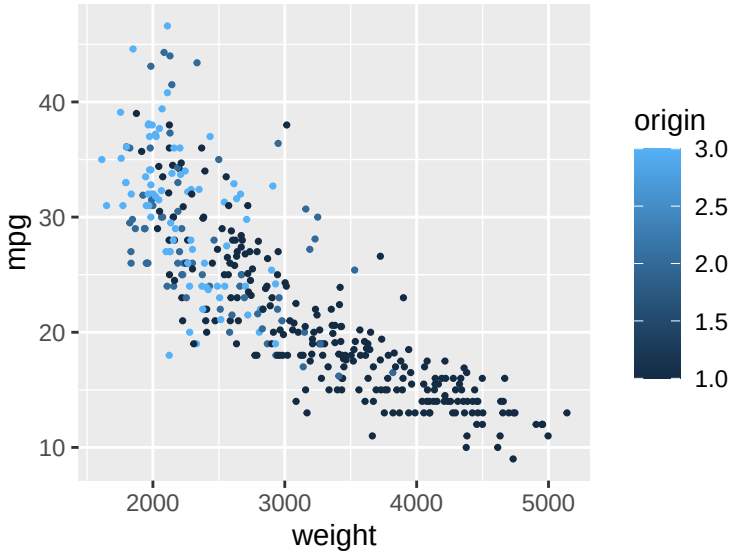
```
ggplot(Auto, aes(weight, mpg)) + geom_point(size=0.6)
```



Coloring Points with a Categorical Variable

Next, we color the points according to the variable `origin` which is coded as 1 for American, 2 for European, and 3 for Japanese. For the plot, we will convert `origin` to a factor type in R. This will tell `ggplot()` that `origin` is categorical, and so it will use a discrete color scale. The plots below demonstrate why the conversion to a factor type is necessary.

```
ggplot(Auto, aes(weight, mpg, color = origin)) + geom_point(size=0.6)
```



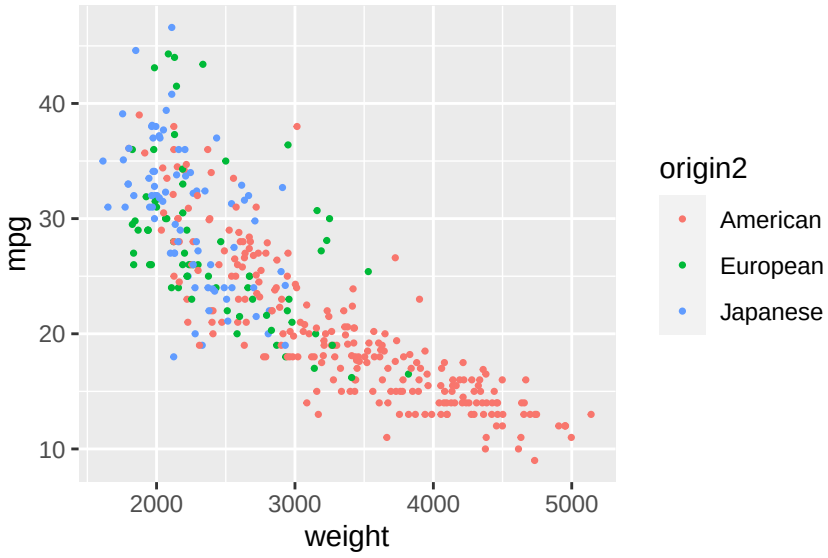

```
# create new variable origin2 that is a factor
Auto$origin2 <- factor(Auto$origin, levels=c(1, 2, 3), labels=c("American", "European", "Japanese"))
class(Auto$origin)
```

```
## [1] "numeric"
```

```
class(Auto$origin2)
```

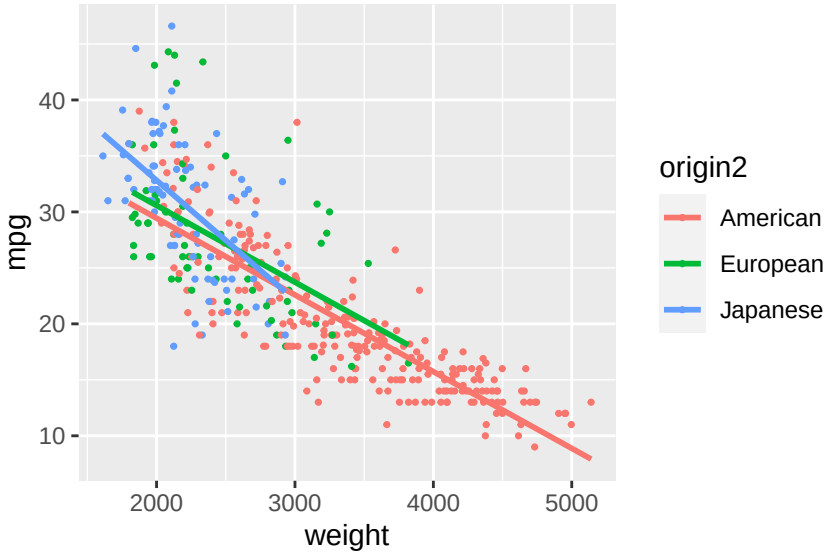
```
## [1] "factor"
```

```
ggplot(Auto, aes(weight, mpg, color = origin2)) + geom_point(size=0.6)
```



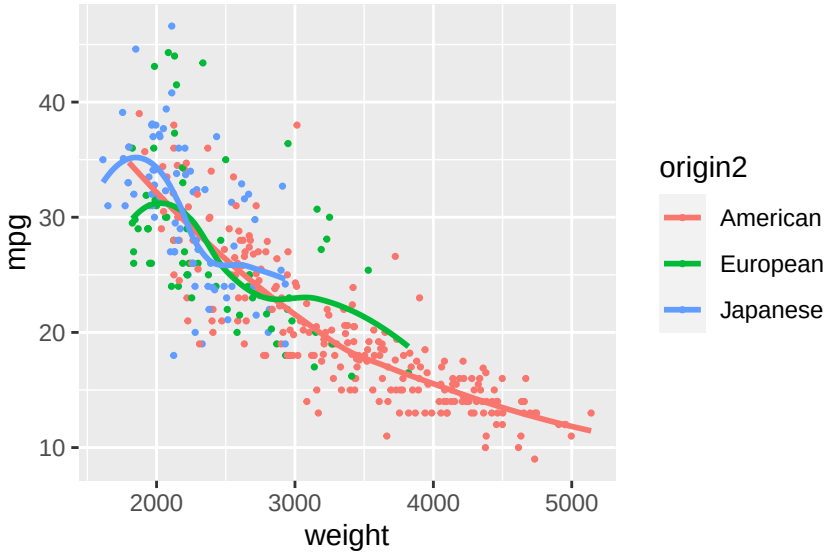
We can use `geom_smooth()` to add the least squares regression line for each category (country of origin). The argument `se` controls whether or not to add a confidence interval band around the line.

```
g1 <- ggplot(Auto, aes(weight, mpg, color = origin2)) + geom_point(size=0.6)
g1 + geom_smooth(method='lm', se=F)
```



We also can use `geom_smooth()` to add loess smoothers to evaluate nonlinearity in the data.

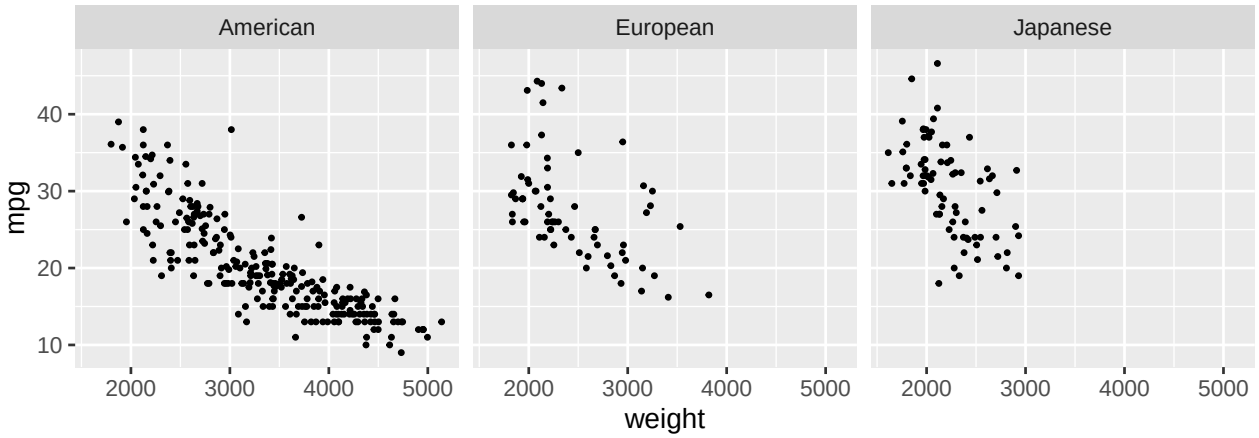
```
g1 + geom_smooth(method='loess', se=F)
```



Faceting Scatter Plots

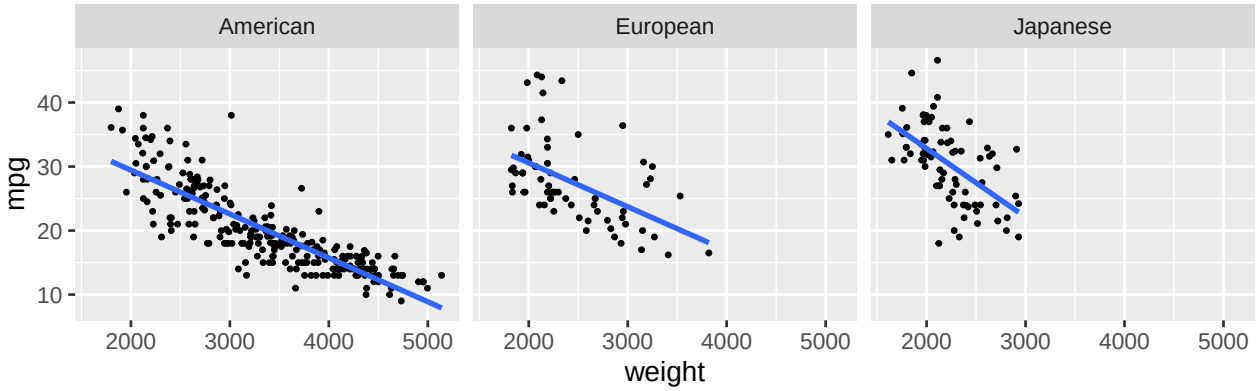
We can use `facet_wrap()` to form a matrix of scatter plots of `mpg` versus `weight` for each country of origin.

```
g2 <- ggplot(Auto, aes(weight, mpg)) + geom_point(size=0.6) +
  facet_wrap(vars(origin2))
g2
```

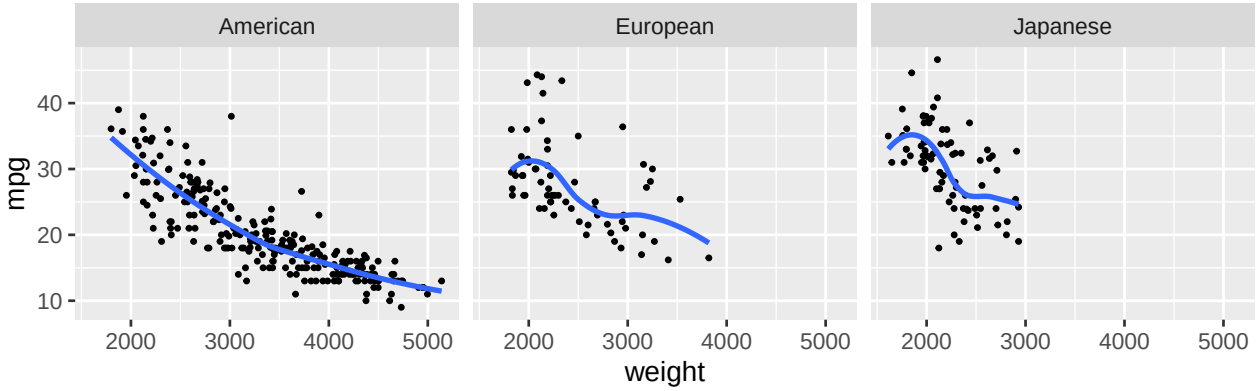


We can use `geom_smooth()` to add a regression line or loess curve to the scatter plot in each panel.

```
g2 + geom_smooth(method='lm', se=F)
```

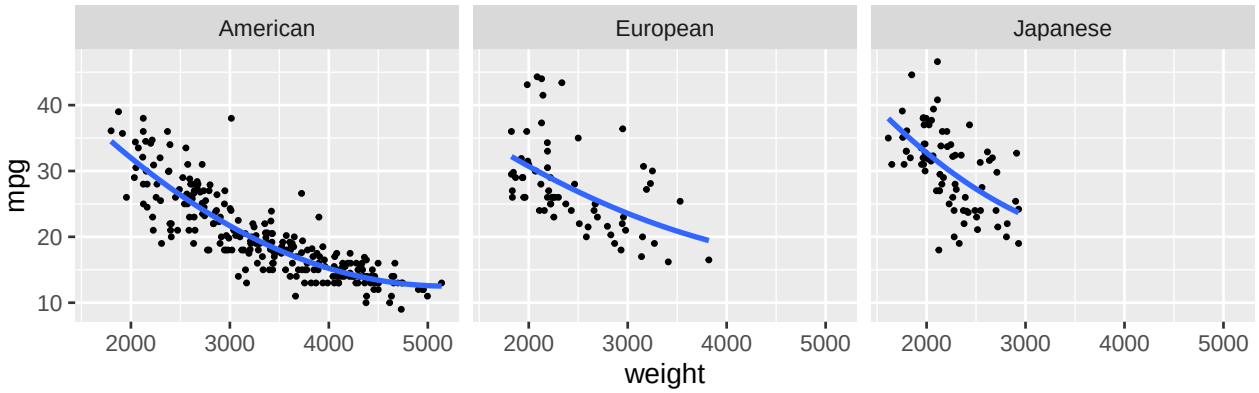



```
g2 + geom_smooth(method='loess', se=F)
```



`geom_smooth()` can also be used to add a quadratic regression curve to the scatter plot in each panel.

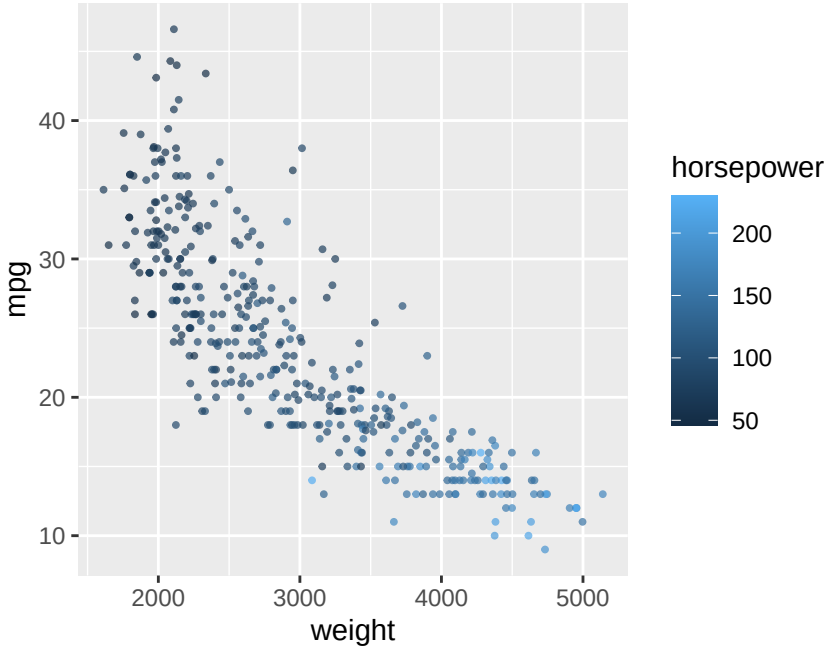
```
g2 + geom_smooth(method='lm', formula = y ~ poly(x,2), se=F)
```



Coloring Points with a Continuous Variable

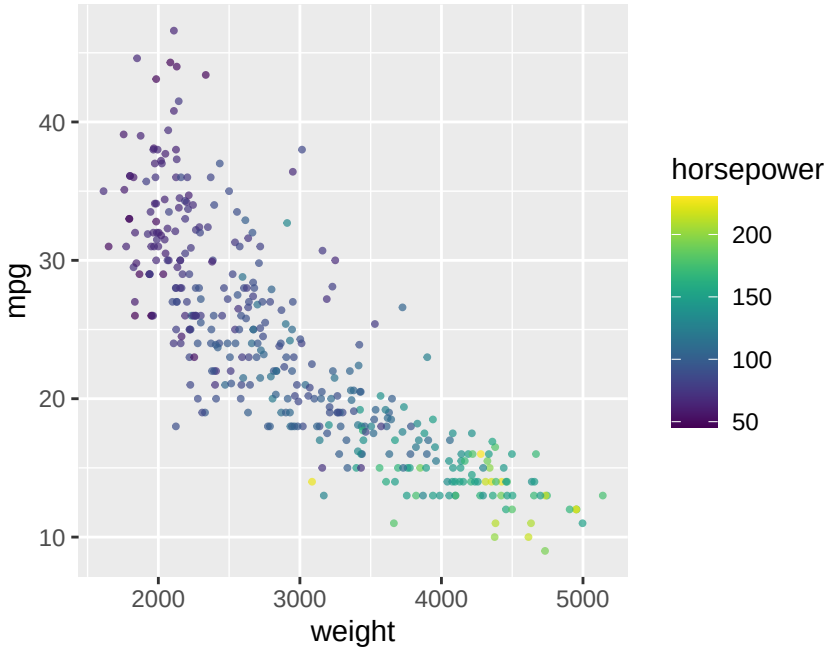
We can color the points in the scatter plot according to a continuous variable such as `horsepower`. The argument `alpha` can be used to adjust the transparency of the points.

```
ggplot(Auto, aes(weight, mpg, color = horsepower)) + geom_point(size=0.7, alpha=0.7)
```



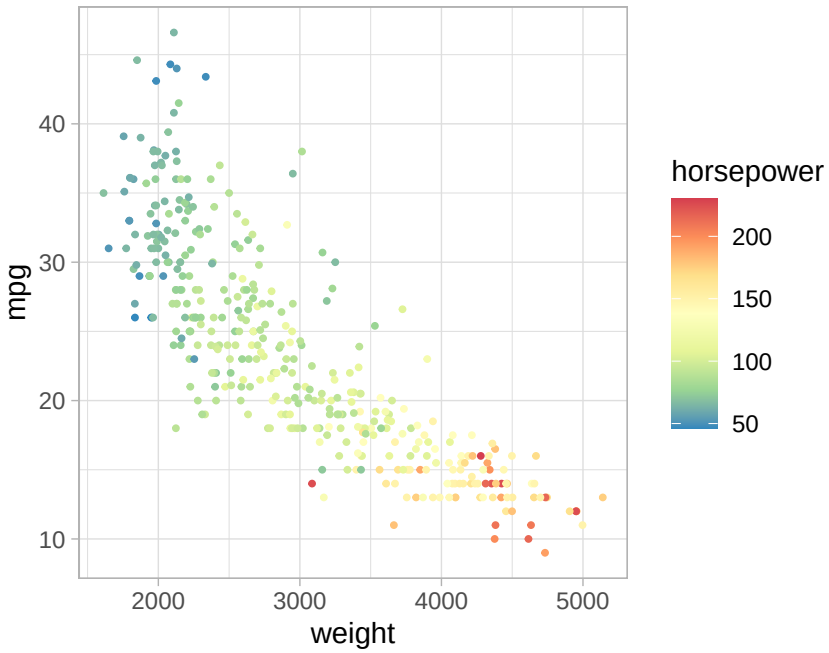
The default color palette is not that great. The `viridis` color palette is a nice alternative since it is designed to be perceptually uniform, robust to color blindness, and prints well in grey scale.

```
library(viridis)
ggplot(Auto, aes(weight, mpg, color = horsepower)) + geom_point(size=0.7, alpha=0.7) +
  scale_color_viridis()
```

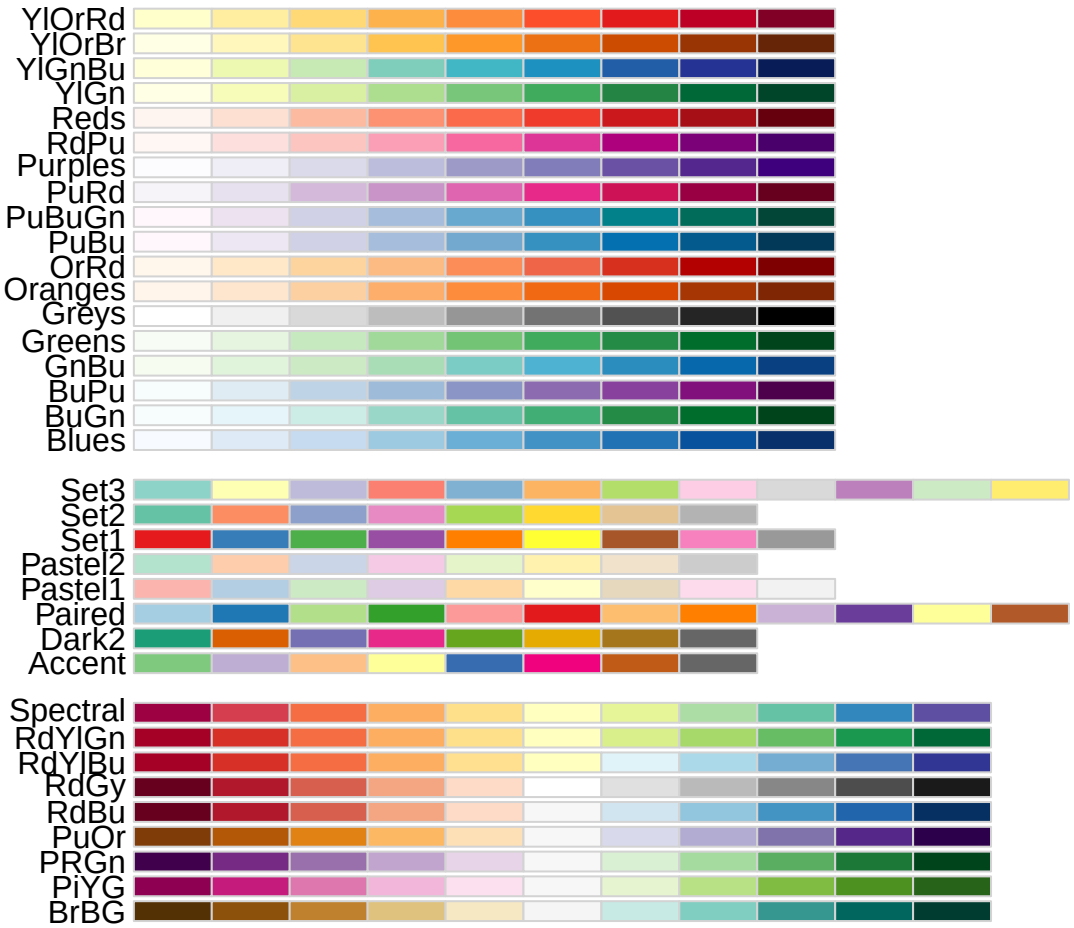


There are other color palettes as well.

```
ggplot(Auto, aes(weight, mpg, color = horsepower)) + geom_point(size=0.7) +
  scale_color_distiller(palette='Spectral') + theme_light()
```



```
library(RColorBrewer)
display.brewer.all()
```



Links

ggplot2 reference

Viridis color palettes