HW2

Jordan Hilton

April 8, 2019

Let's load our data:

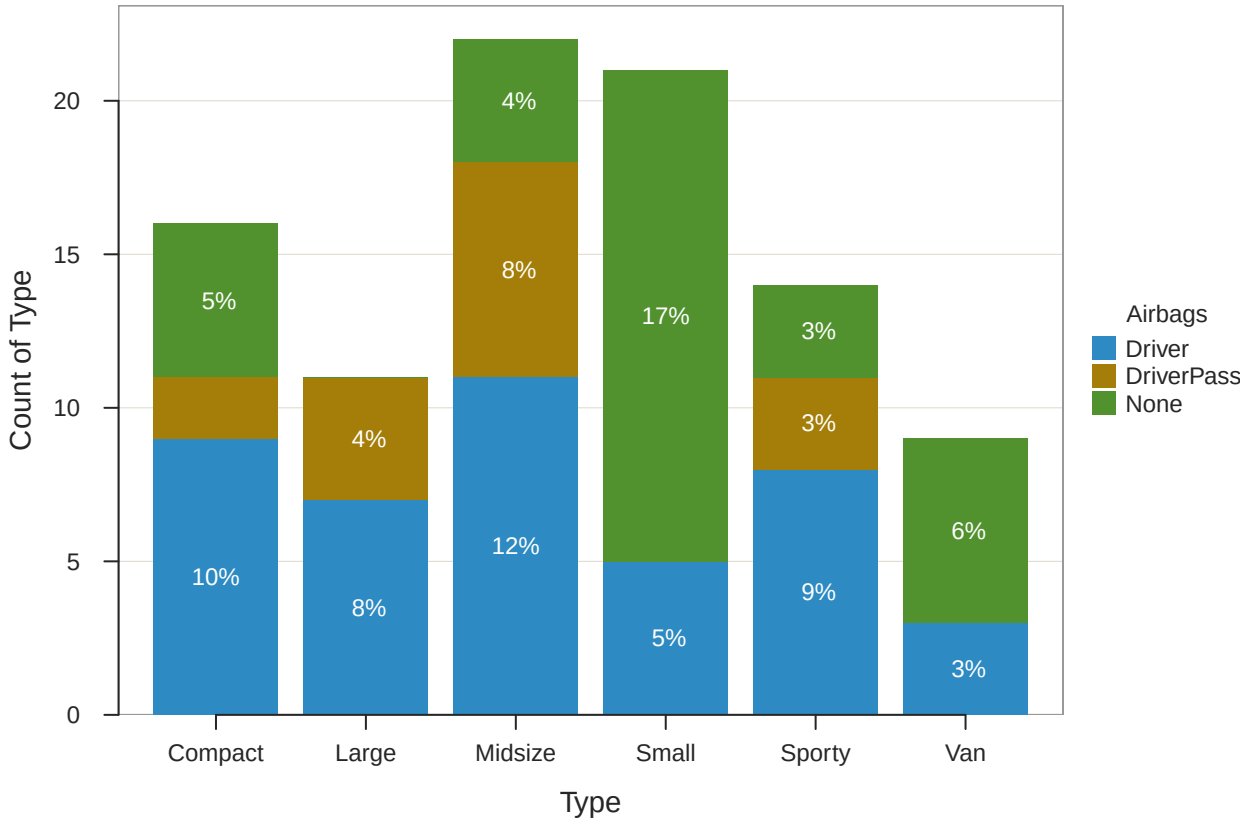
```
d <- rd("Cars93.csv", quiet=TRUE)
```

1 Bar Chart

a.

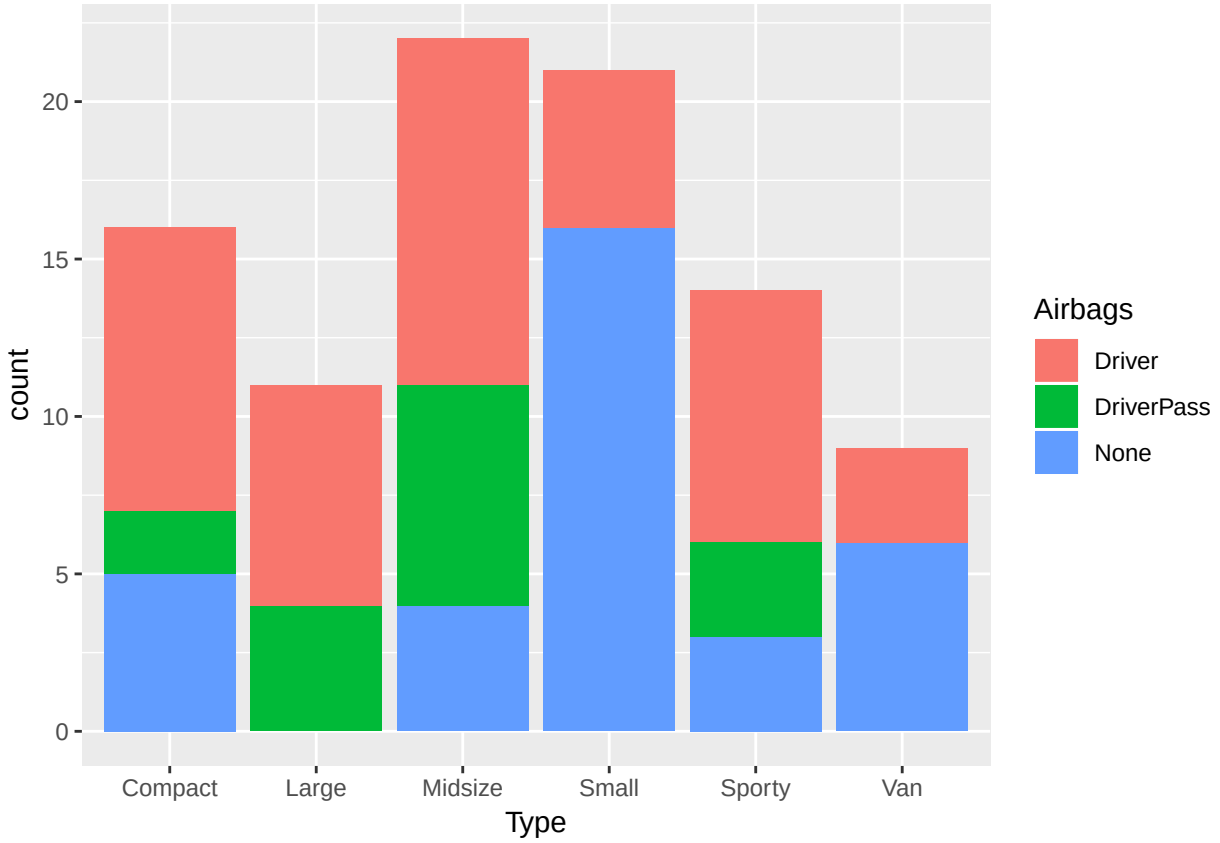
Here's the bar chart for type of car by airbag configuration in lessR:

```
bc(Type, by=Airbags, quiet=TRUE)
```



and ggplot2:

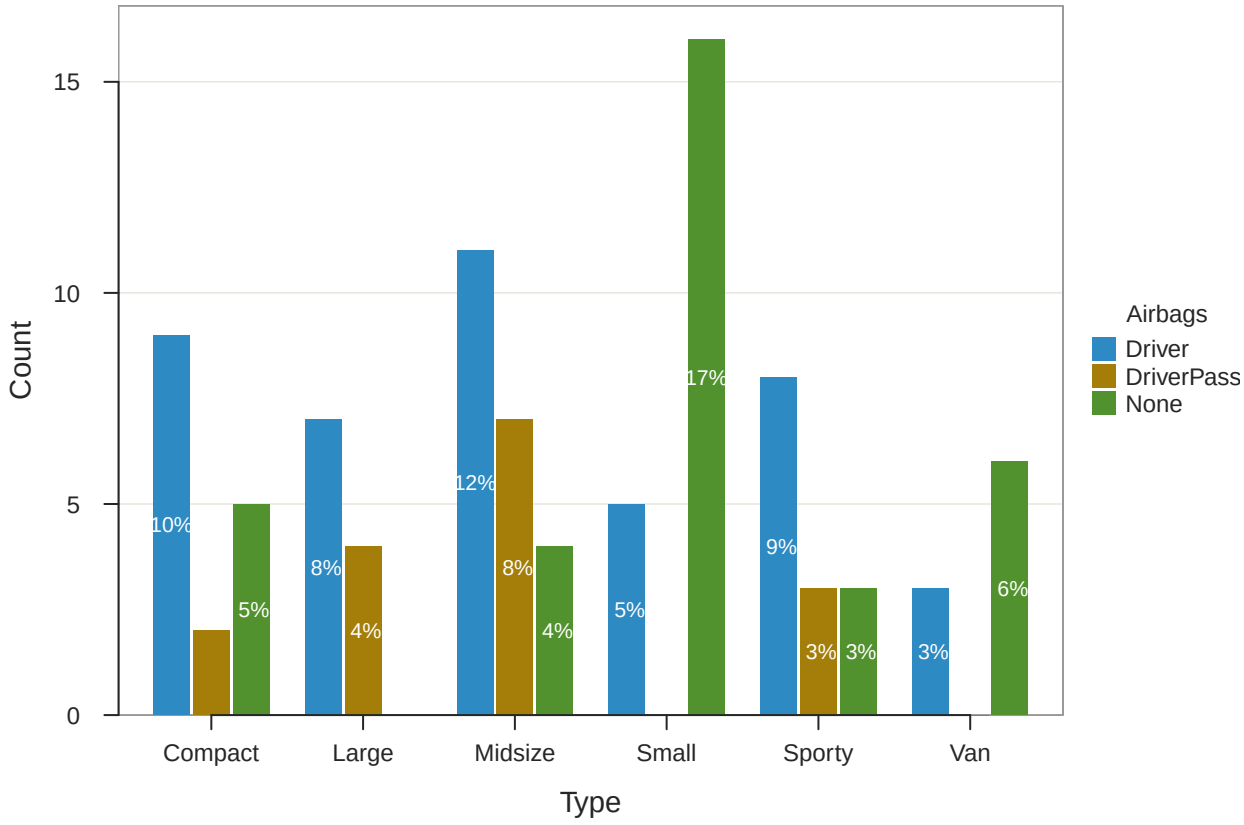
```
ggplot(d, aes(Type, fill=Airbags)) + geom_bar()
```



b.

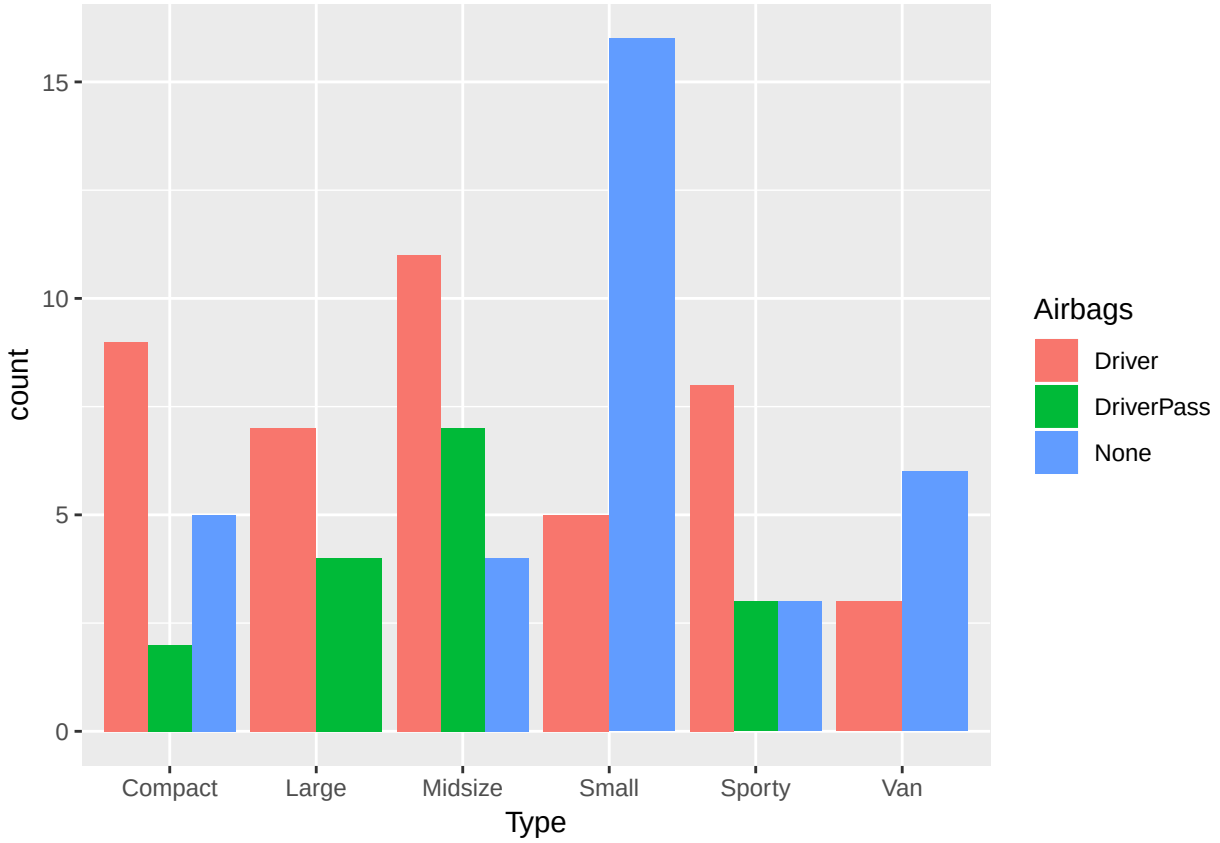
Here's the side-by-side bar chart in lessR for the same data:

```
bc(Type, by=Airbags, beside=TRUE, quiet=TRUE)
```



and ggplot2:

```
ggplot(d, aes(Type, fill=Airbags)) + geom_bar(position="dodge")
```



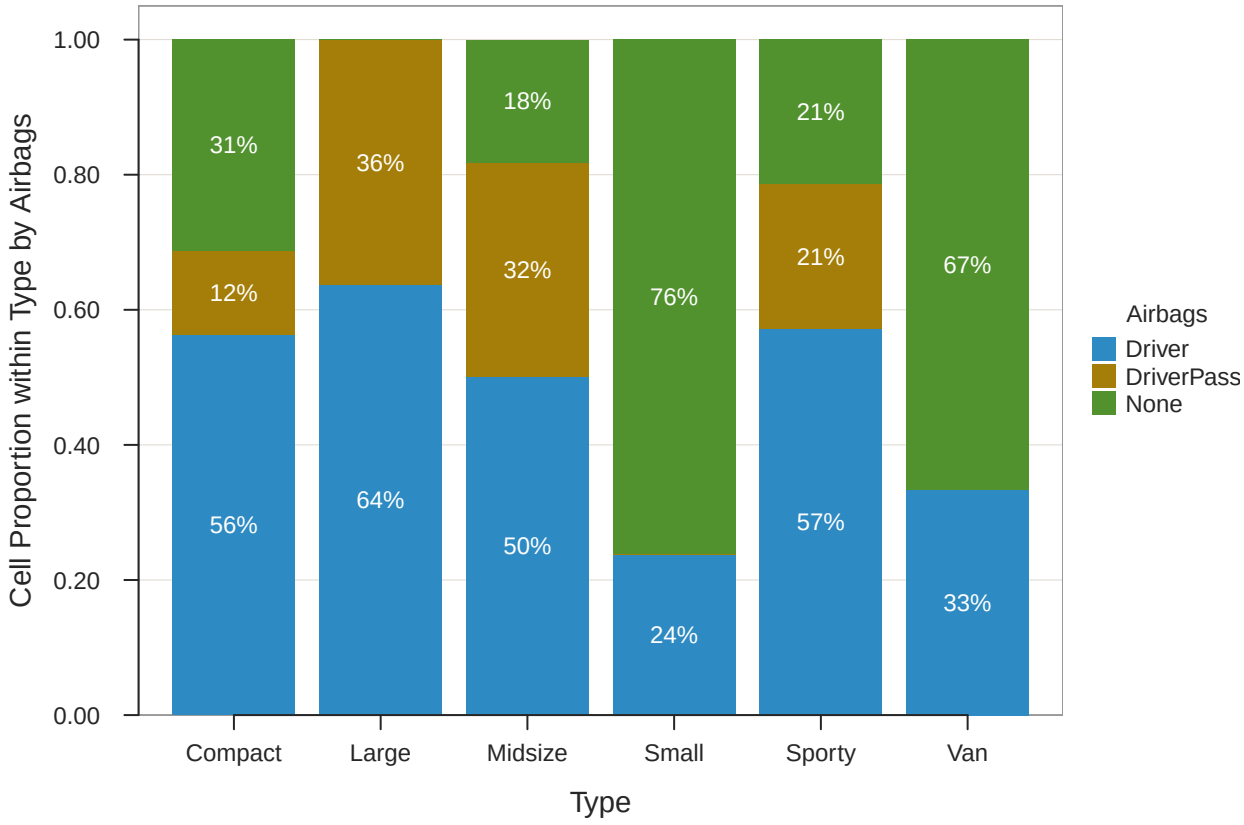
c.

Small cars frequently have no airbags- seems unsafe! It seems like midsize and large cars most frequently have both driver and passenger side airbags.

d.

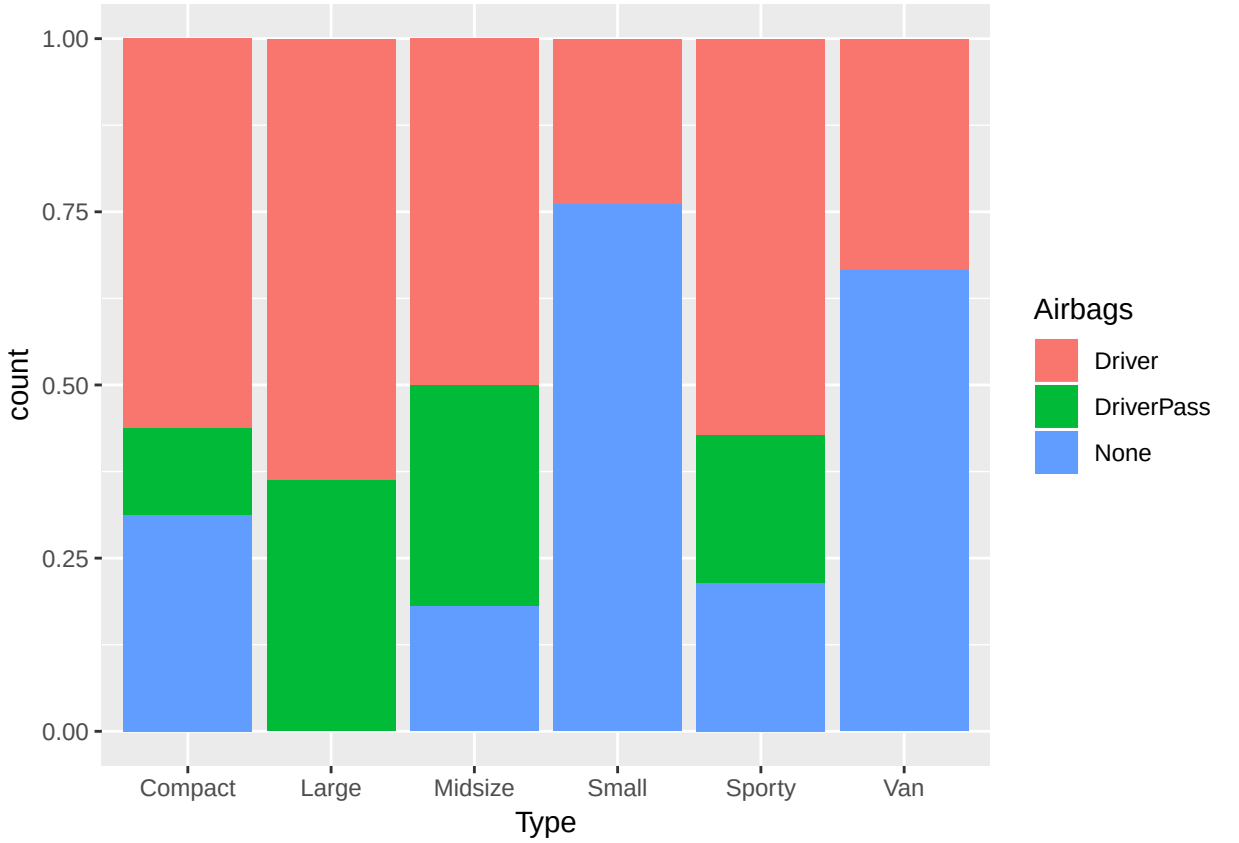
Here's the lessR bar chart by proportion:

```
bc(Type, by=Airbags, quiet=TRUE, stat.x="proportion")
```



and ggplot2:

```
ggplot(d, aes(Type, fill=Airbags)) + geom_bar(position="fill")
```



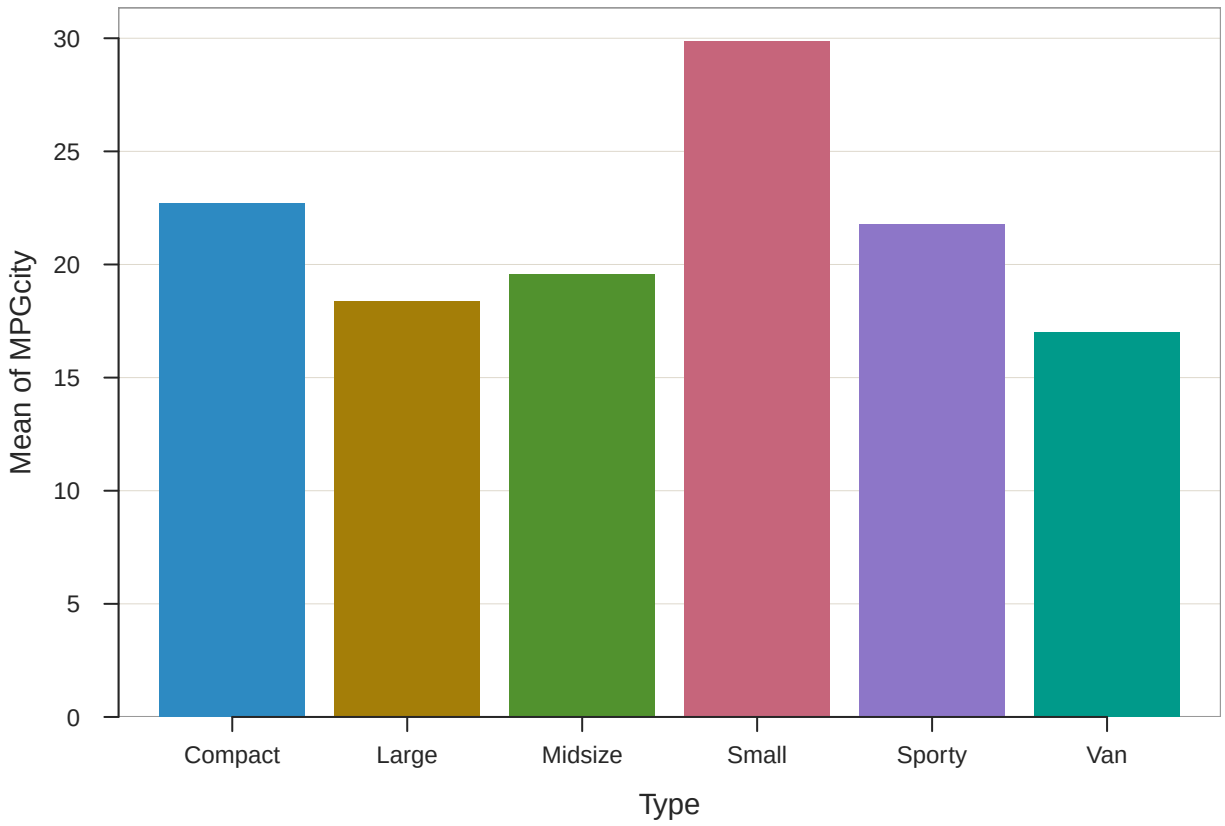
e.

Here the proportions are clearer- we can see that small cars and vans both have large proportions of vehicles with no airbags, and that large and midsize vehicles have the highest proportions of vehicles with both driver and passenger airbags.

f.

I'm interpreting this question to be asking for the mean of city MPG by type of car (since summing the city MPG of different models of cars doesn't make much sense); here's the relevant bar chart.

```
bc(Type, y=MPGcity, stat.yx="mean", quiet=TRUE)
```



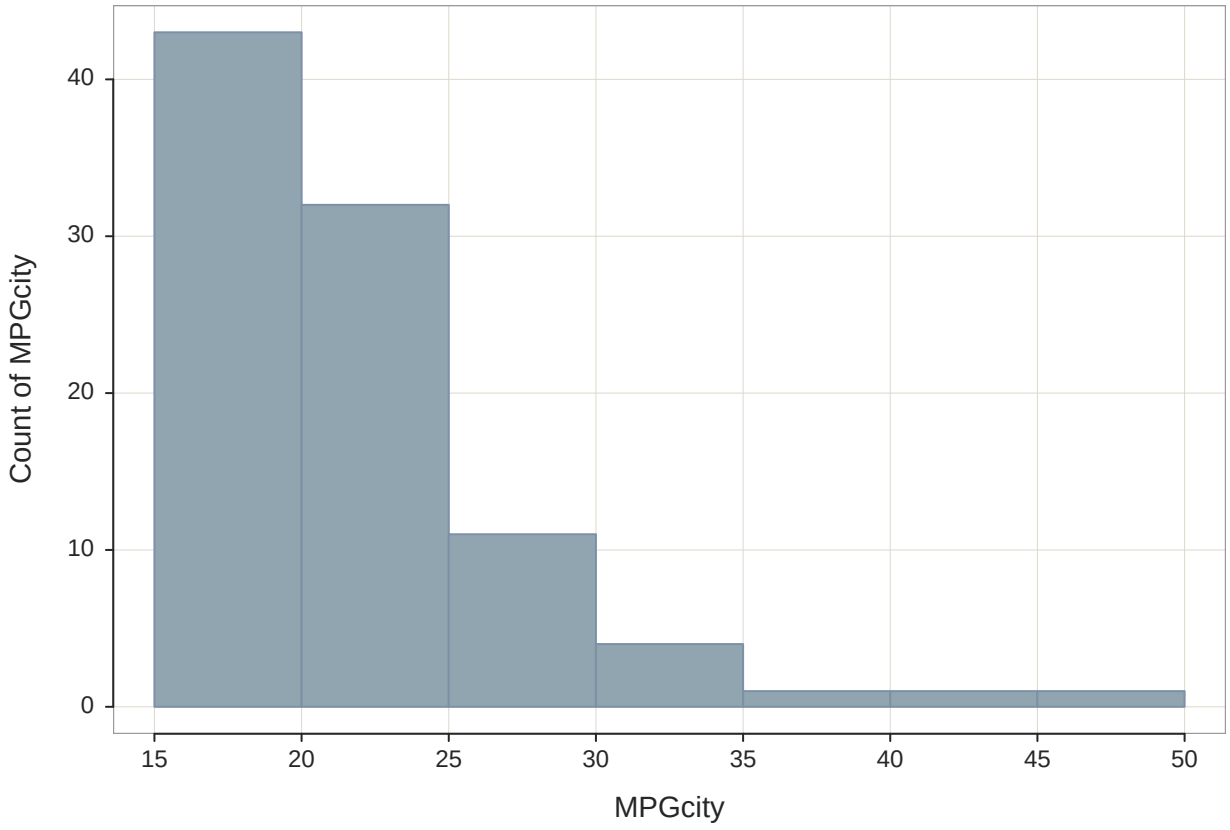
We can see that small cars have the best mileage, while vans have the worst.

2 Histogram

a.

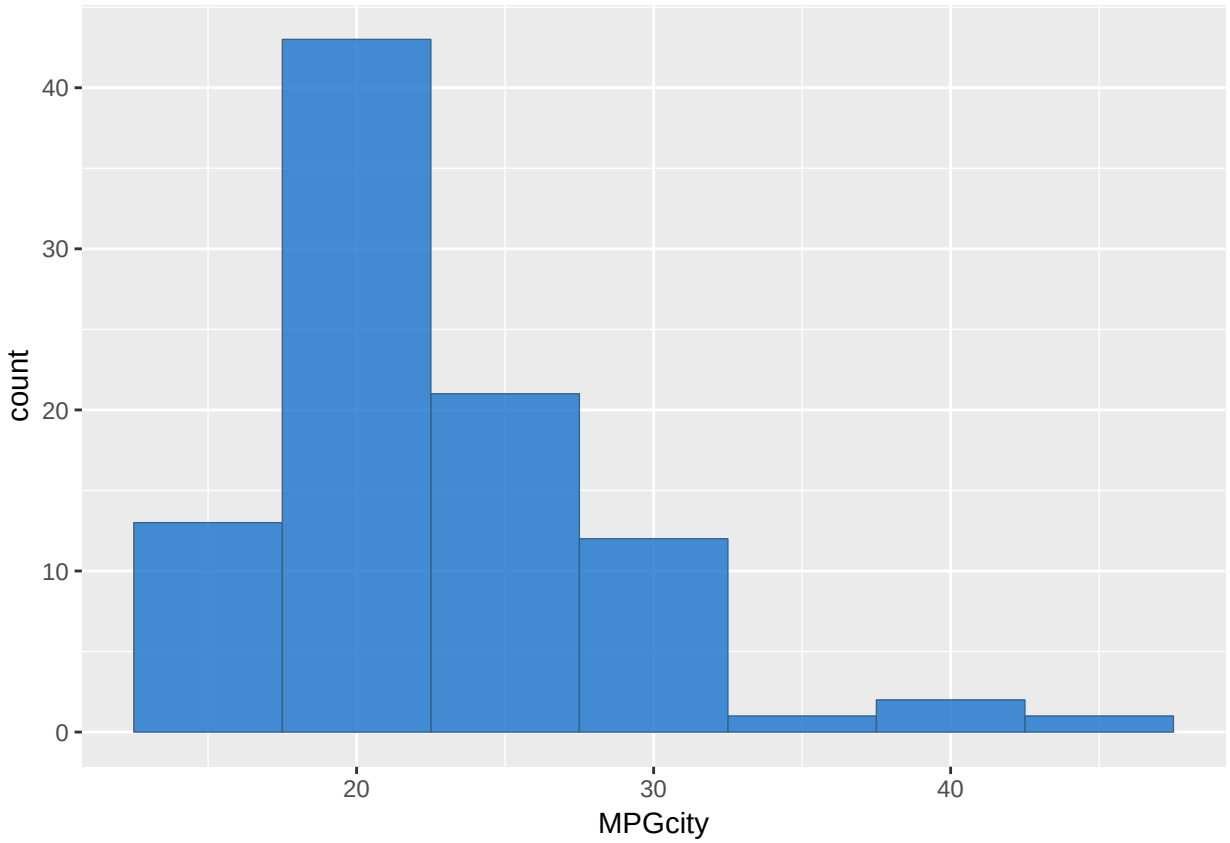
Here's the histogram for city MPG using lessR:

```
Histogram(MPGcity, quiet=TRUE)
```



and ggplot2, using the same bin width:

```
ggplot(d, aes(MPGcity)) +
  geom_histogram(binwidth=5, fill="dodgerblue3", color="steelblue4",
                 alpha=.8, size=.25)
```

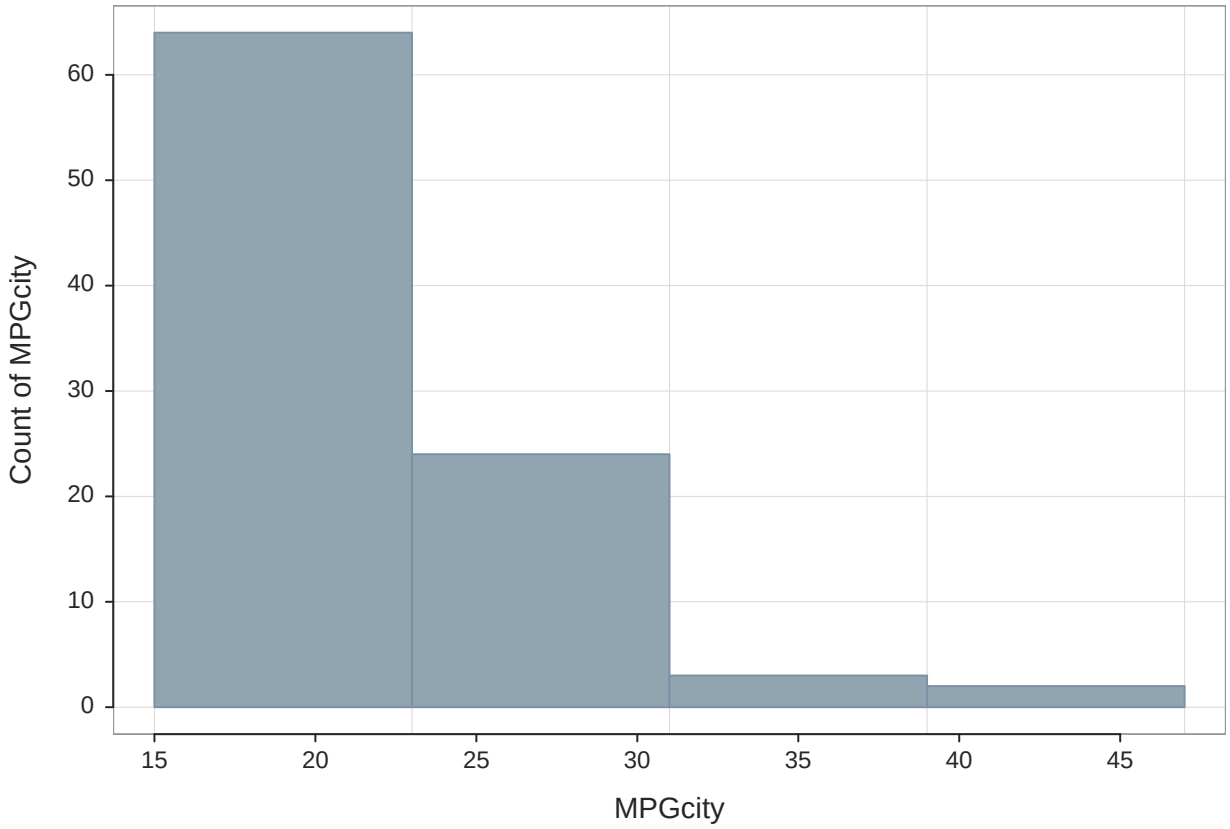


Note the steep dropoff after 25 MPG; this data may be from before hybrids were common.

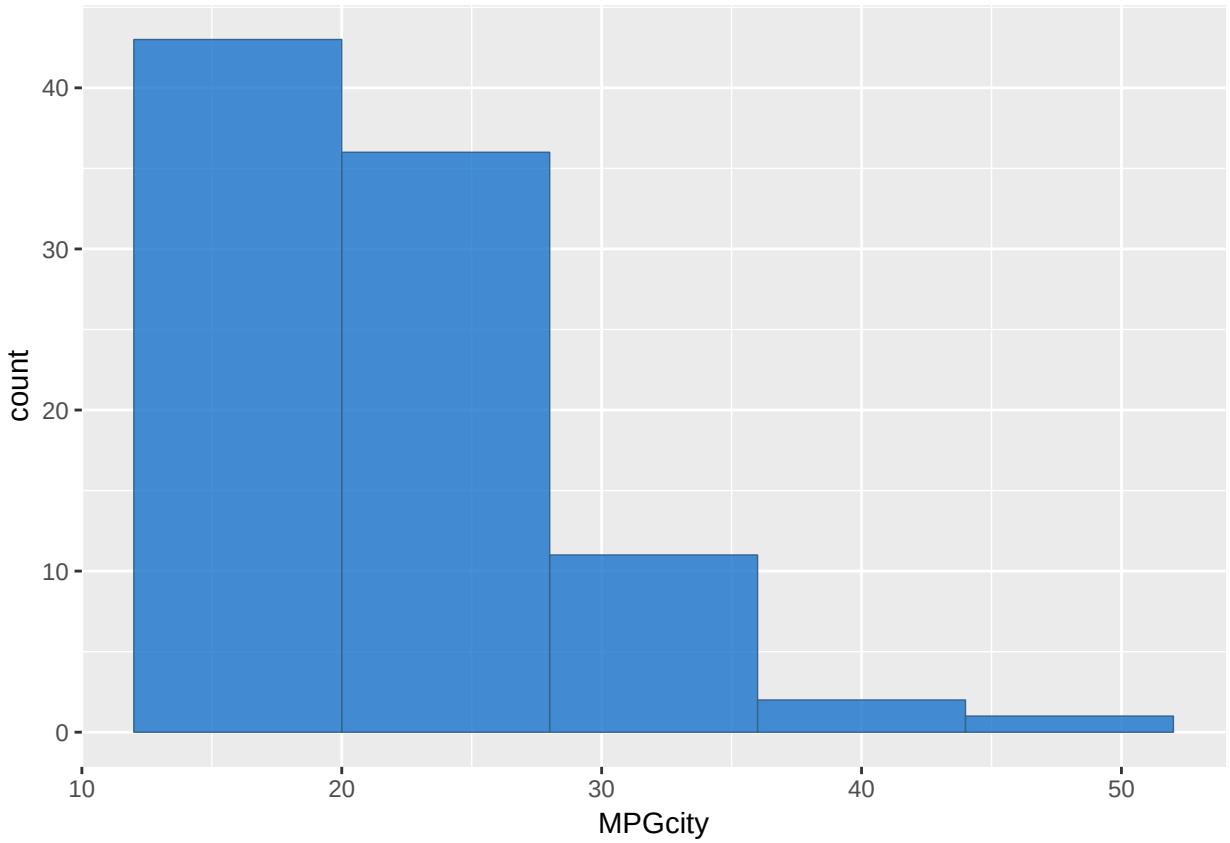
b.

Here are the same two plots with more appropriate bin widths. I've increased the width to 8 to more clearly show the divide between normal and high-mileage vehicles.

```
Histogram(MPGcity, bin.width=8, quiet=TRUE)
```



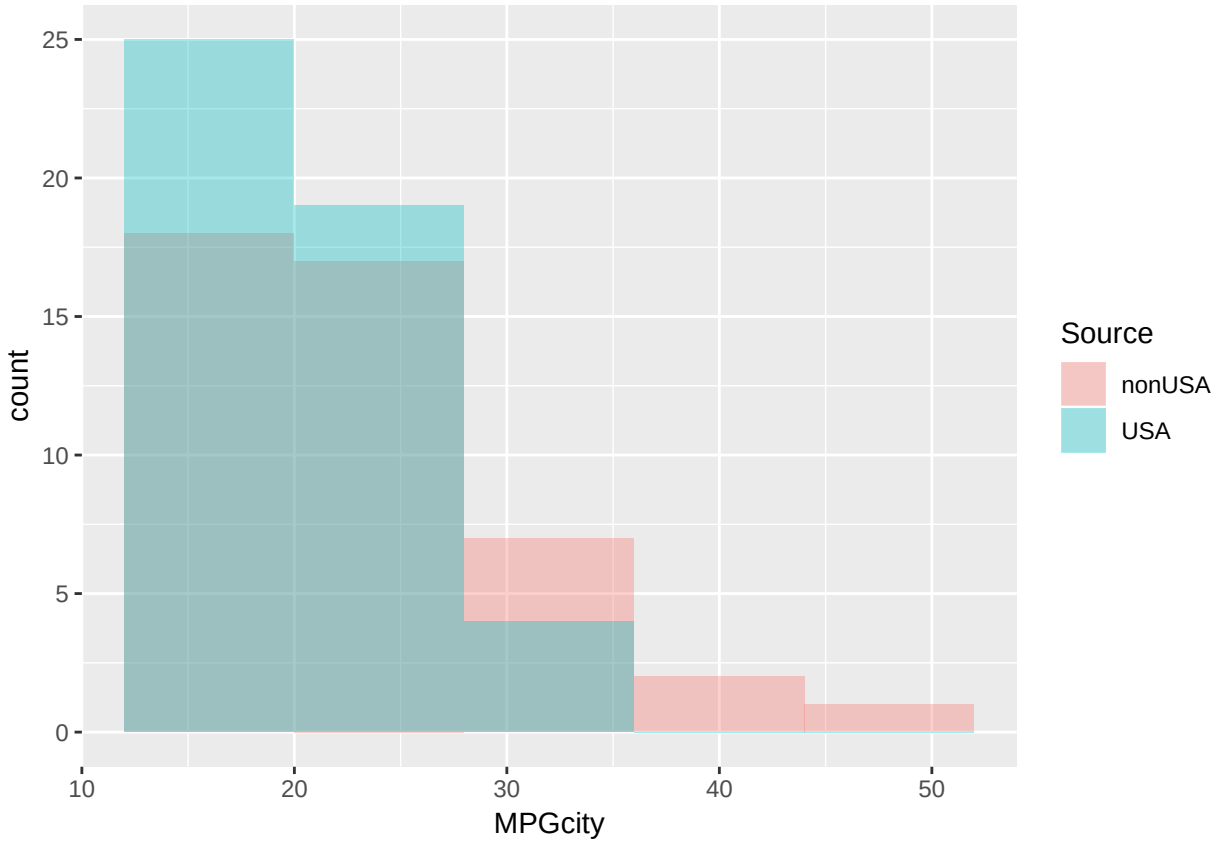
```
ggplot(d, aes(MPGcity)) +
  geom_histogram(binwidth=8, fill="dodgerblue3", color="steelblue4",
                 alpha=.8, size=.25)
```



c.

Here's the ggplot2 overlapping histogram for city MPG by source:

```
ggplot(d, aes(MPGcity, fill=Source)) +
  geom_histogram(position="identity",binwidth=8,
                 alpha=.35, size=.25)
```



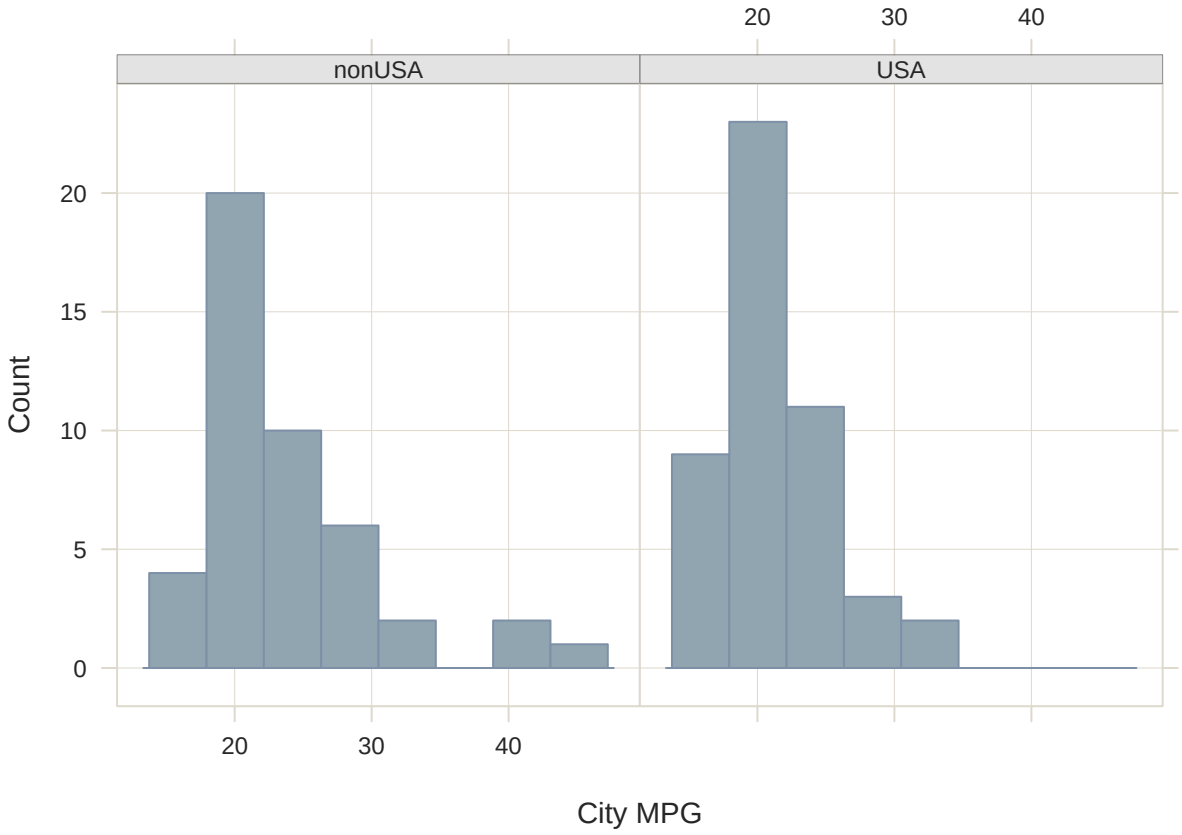
It looks like the non-USA cars in this sample skew toward being more fuel efficient.

d.

Here's the side-by-side histogram for city MPG by source from lessR:

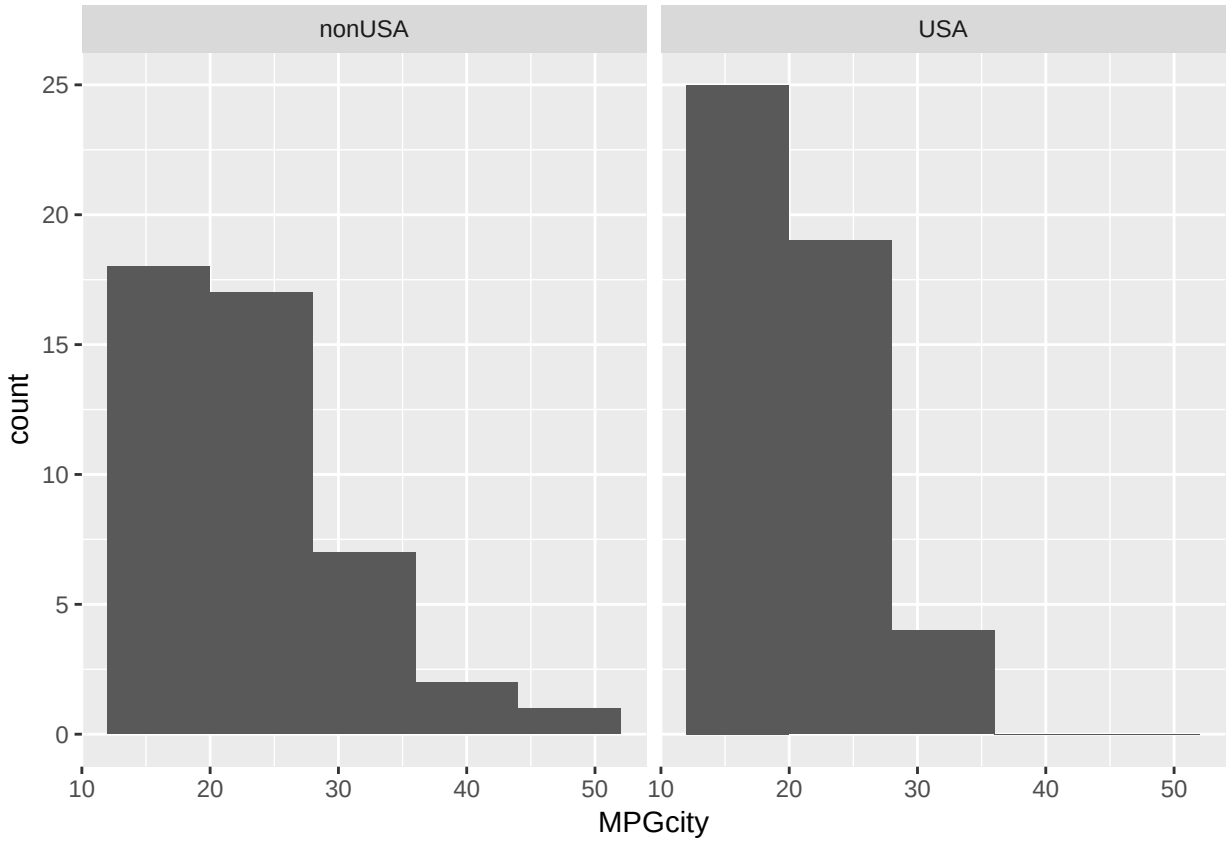
```
hs(MPGcity, by1=Source, quiet=TRUE, ylab="Count", xlab="City MPG")
```

```
## [Trellis graphics from Deepayan Sarkar's lattice package]
```



and ggplot2:

```
ggplot(d, aes(MPGcity)) +
  geom_histogram(binwidth=8) + facet_grid(cols=vars(Source))
```

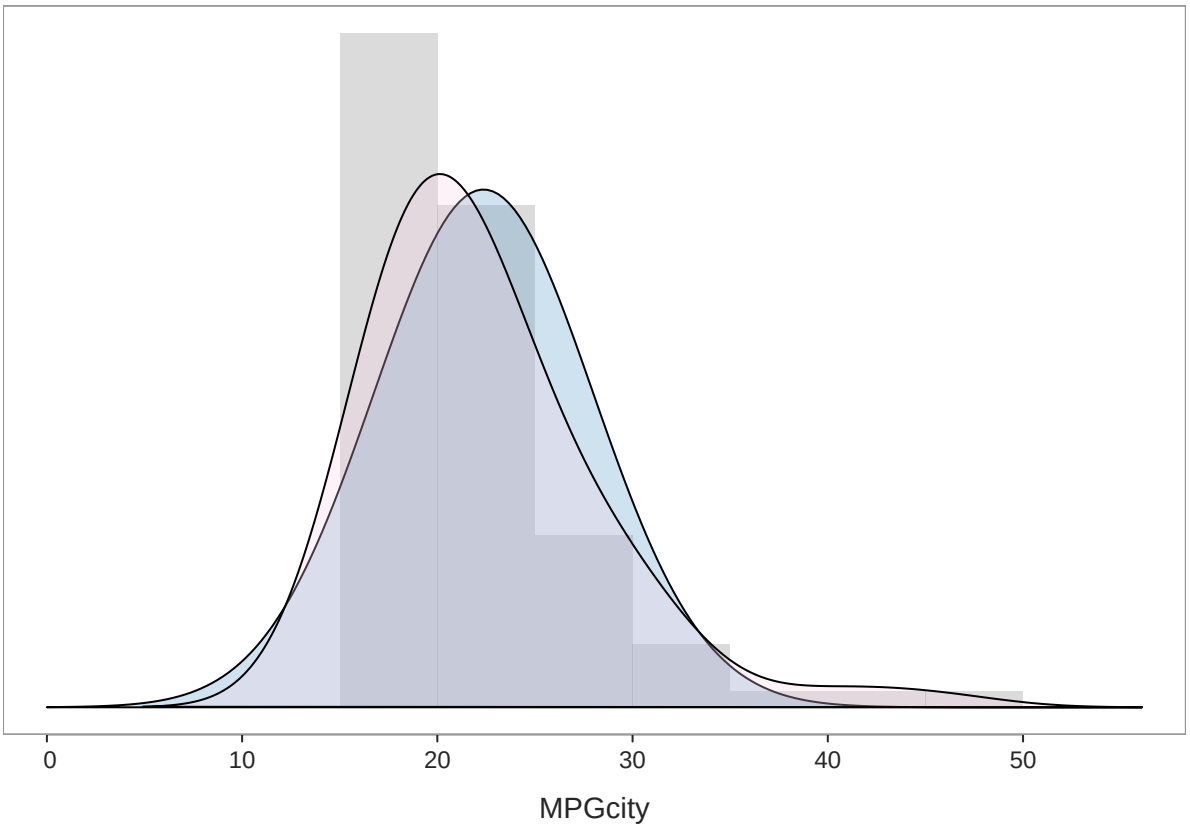


This shows the same comparison as in the overlapping histogram- the non-USA distribution skews more efficient.

e.

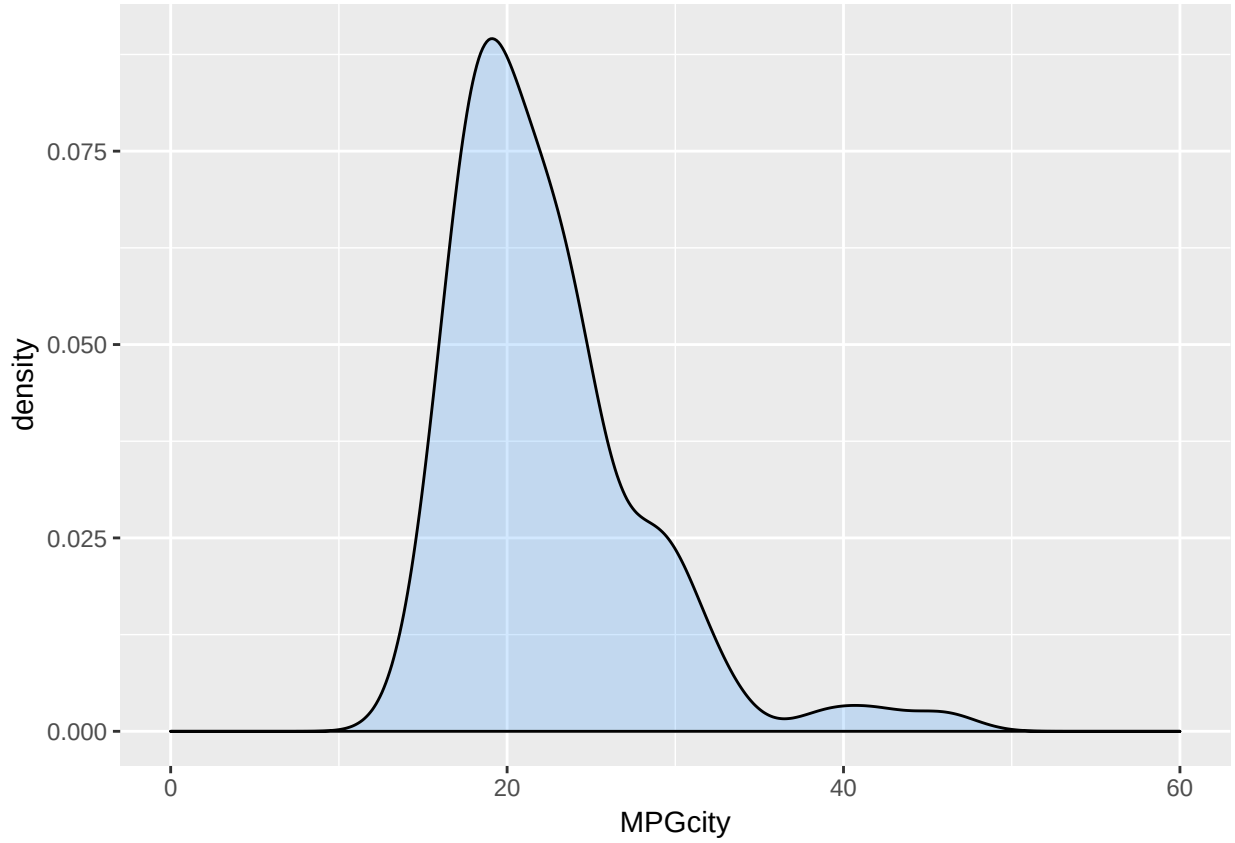
Here's the density curve for city MPG using lessR:

```
Density(MPGcity, x.min=0,quiet=TRUE)
```



and ggplot2:

```
ggplot(d, aes(MPGcity)) + geom_density(alpha=.2, fill="dodgerblue") + xlim(0,60)
```

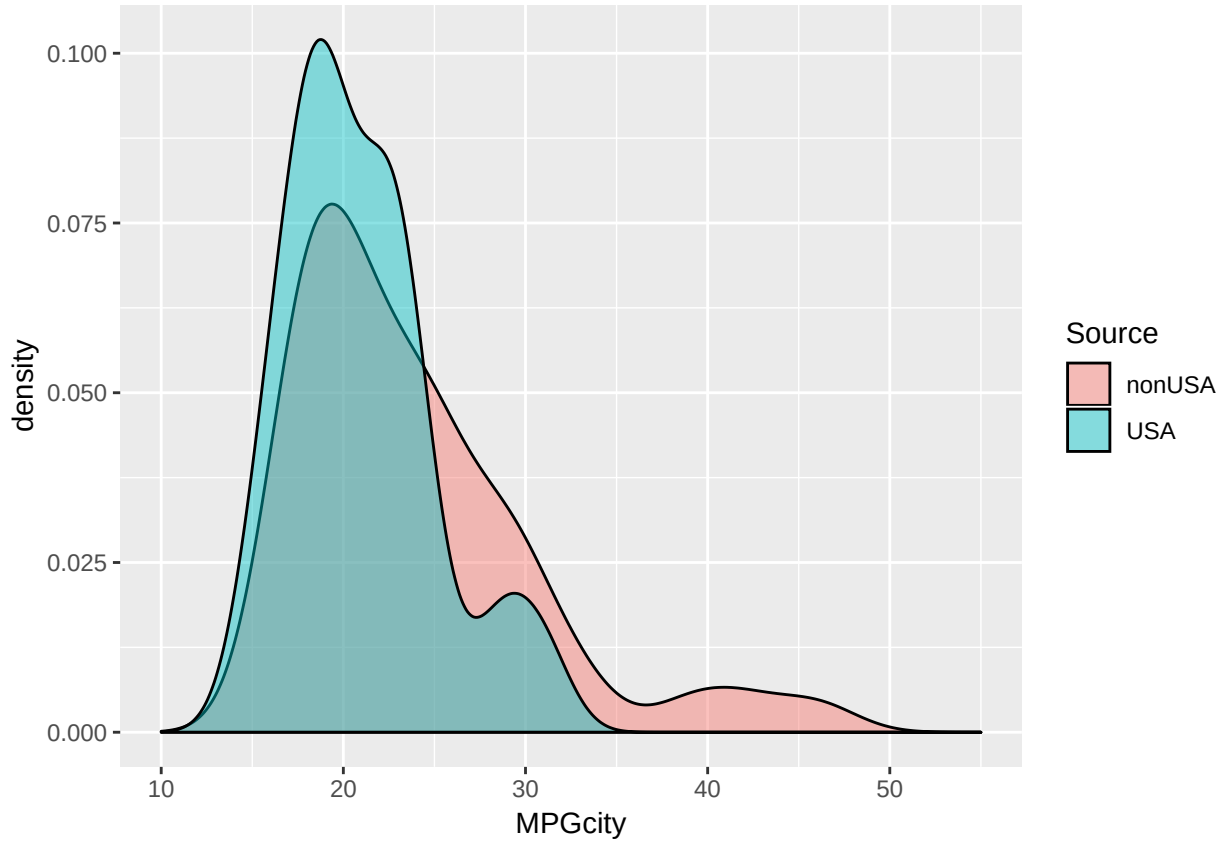


Again we can see the sharp peak in cars that get around 20 MPG in the city.

f.

Here's the overlapping density plot in ggplot2:

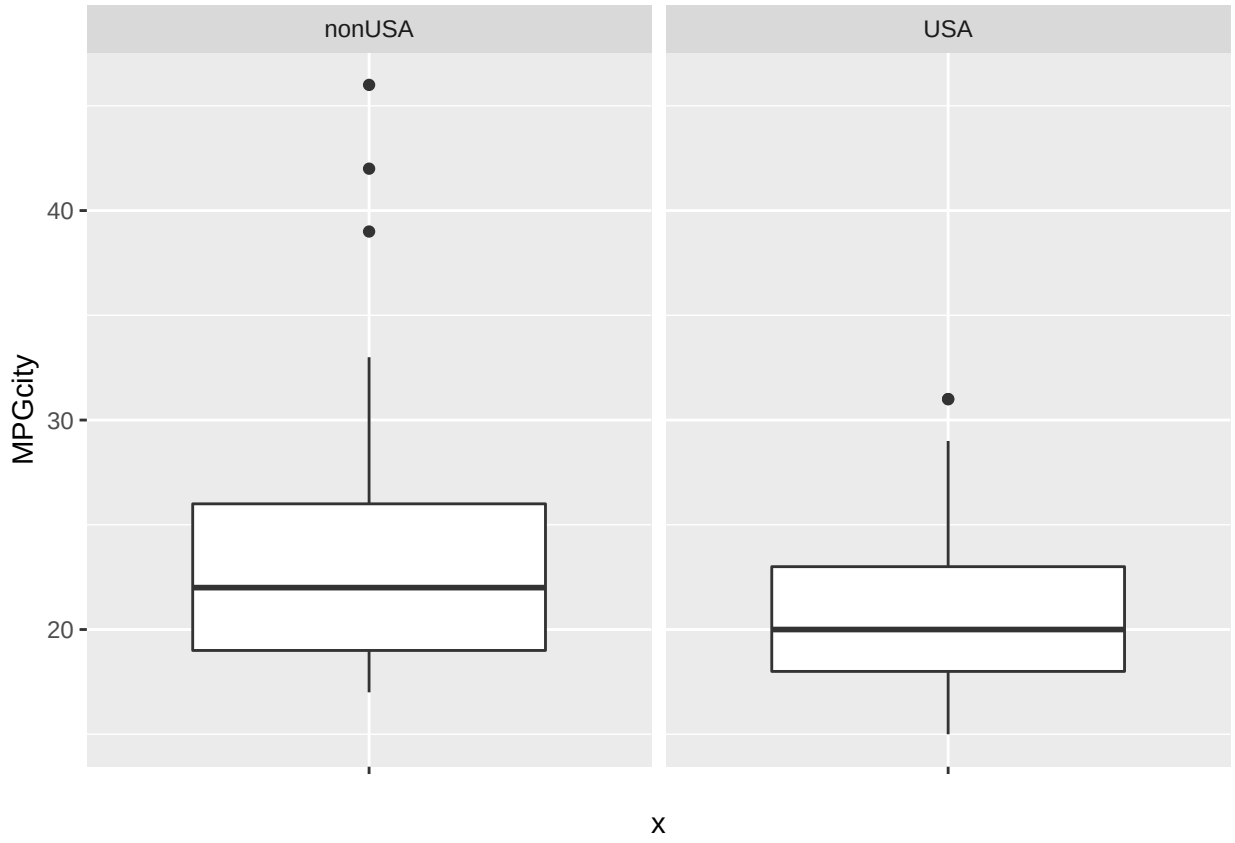
```
ggplot(d, aes(MPGcity, fill=Source)) +
  geom_density(position="identity", alpha=.45)+xlim(10,55)
```



g.

Here are the ggplot2 box plots for city MPG by source:

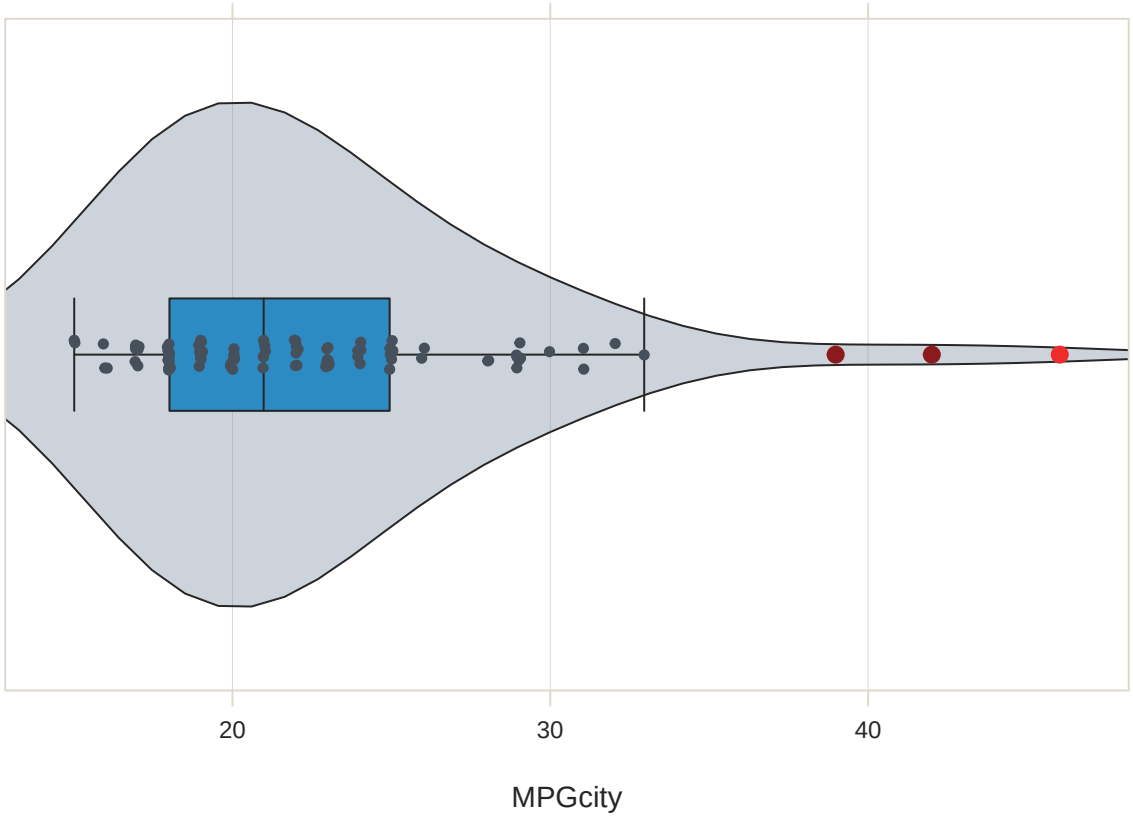
```
ggplot(d, aes(x="", y=MPGcity)) +
  geom_boxplot() + facet_grid(cols=vars(Source))
```



h.

Here's the integrated VBS for city MPG using lessR:

```
Plot(MPGcity, quiet=TRUE)
```



i.

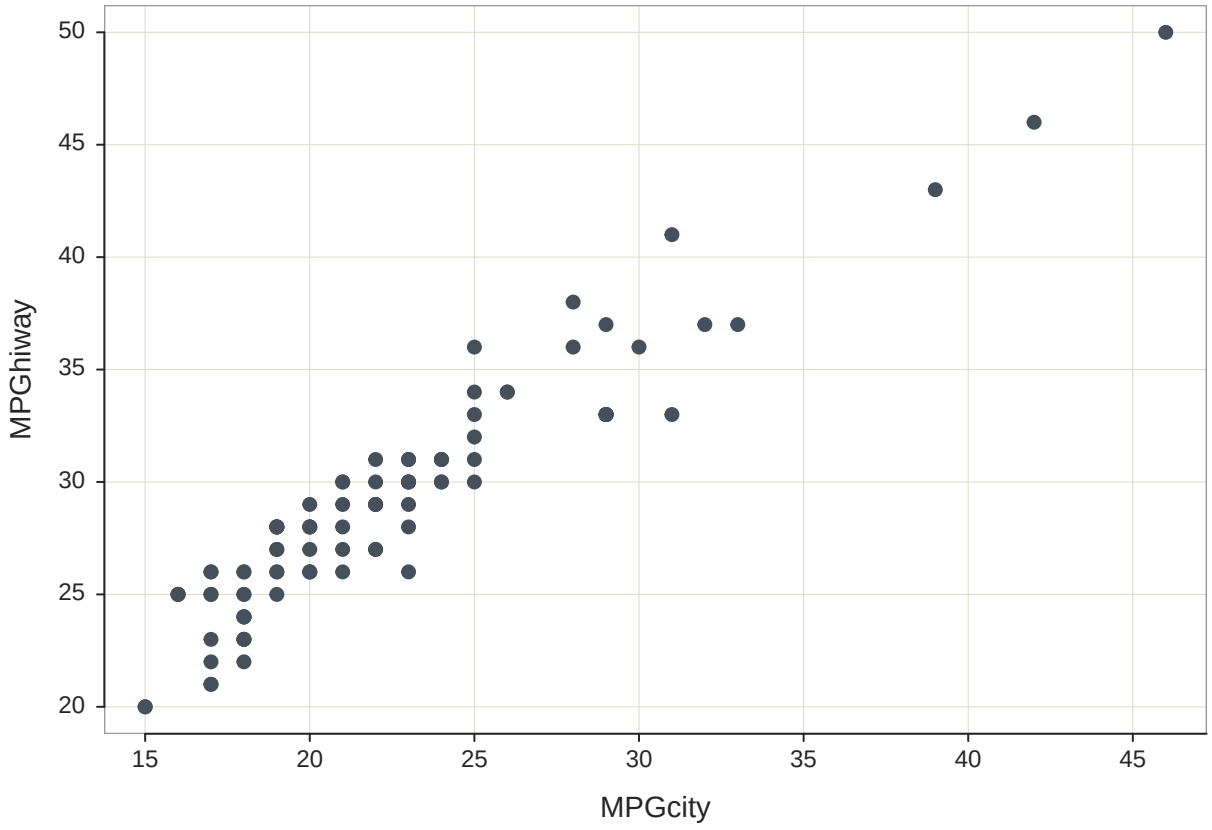
The full VBS plot presents the same distribution in so many different ways that you can get a lot more information from it- for instance, the identification of the extremity of the 3 outliers on the right, the density of the points around the mean, and how neatly most of the distribution fits within the range of the box plot. This level of detail also makes it much busier and more difficult to read- the important thing to get out of looking at this distribution is just that you have 3 outliers on the right and a bunch of points clumped around the mean, which the histogram communicates just as well and much more simply.

3. Scatterplot

a.

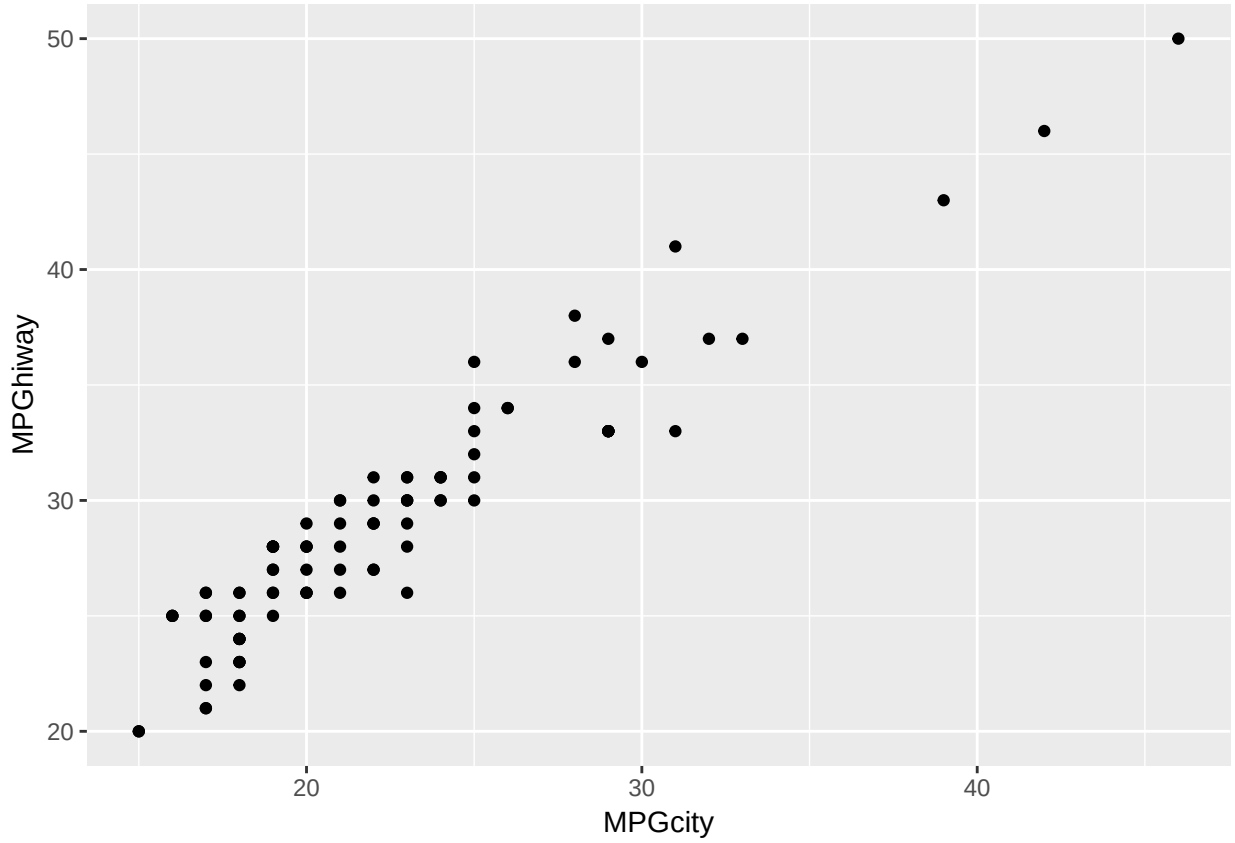
Here's the scatterplot for city and highway MPG in lessR:

```
Plot(MPGcity, MPGhiway, quiet=TRUE)
```



and the same thing in ggplot2:

```
ggplot(d, aes(MPGcity, MPGhiway)) + geom_point()
```



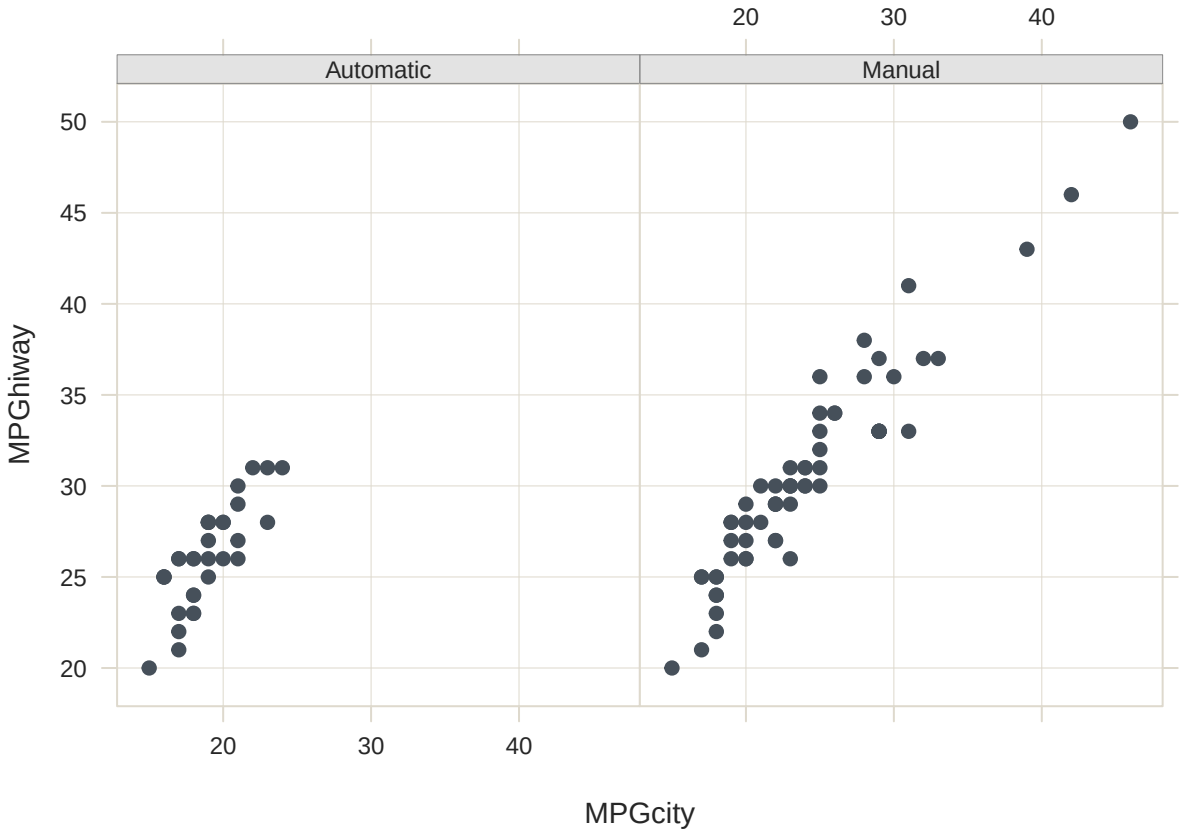
b.

Here's the trellis for the same plot split by transmission. I'm going to first convert the "Manual" variable into a factor as described in assignment 1.

```
d$Manual <- factor(d$Manual, levels=0:1, labels=c("Automatic", "Manual"))
```

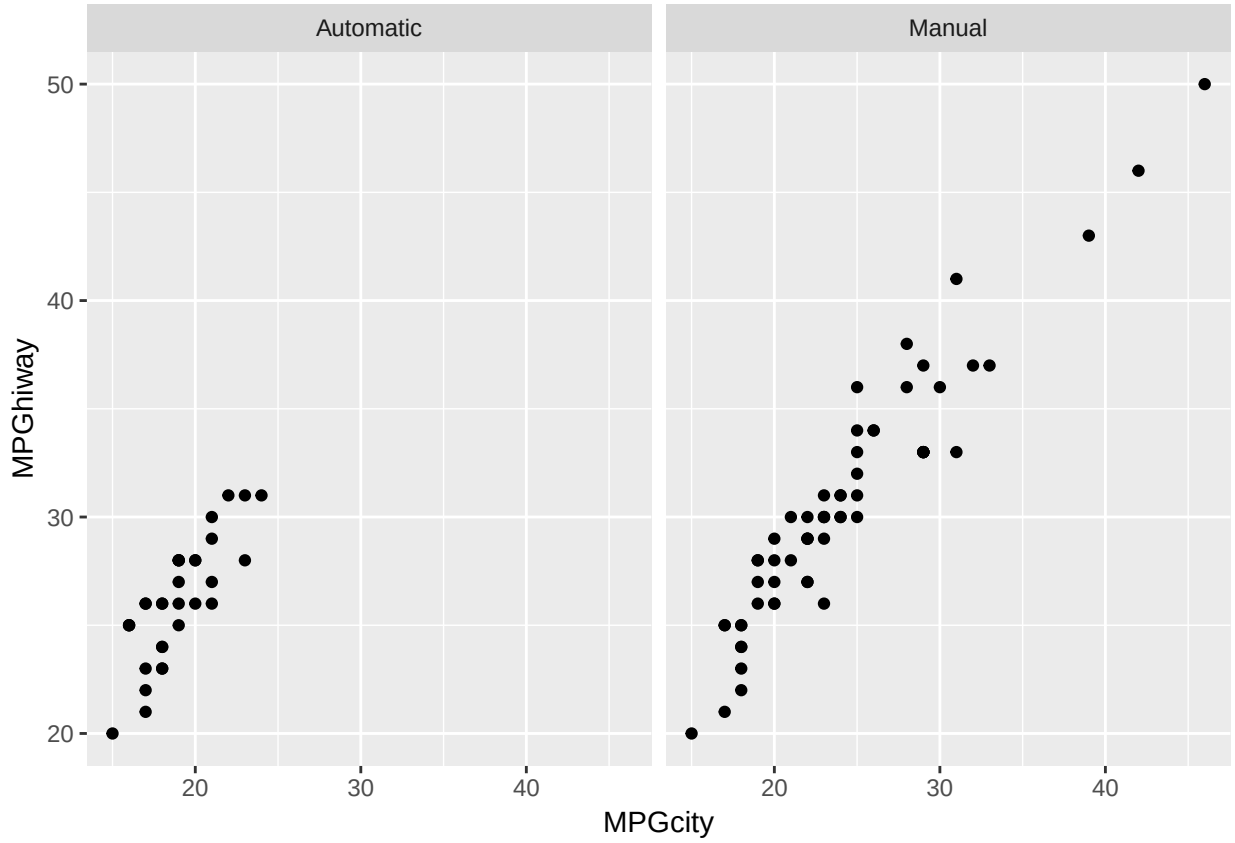
The plot in lessR:

```
Plot(MPGcity, MPGhiway, by1=Manual, n.col=2, quiet=TRUE)
```



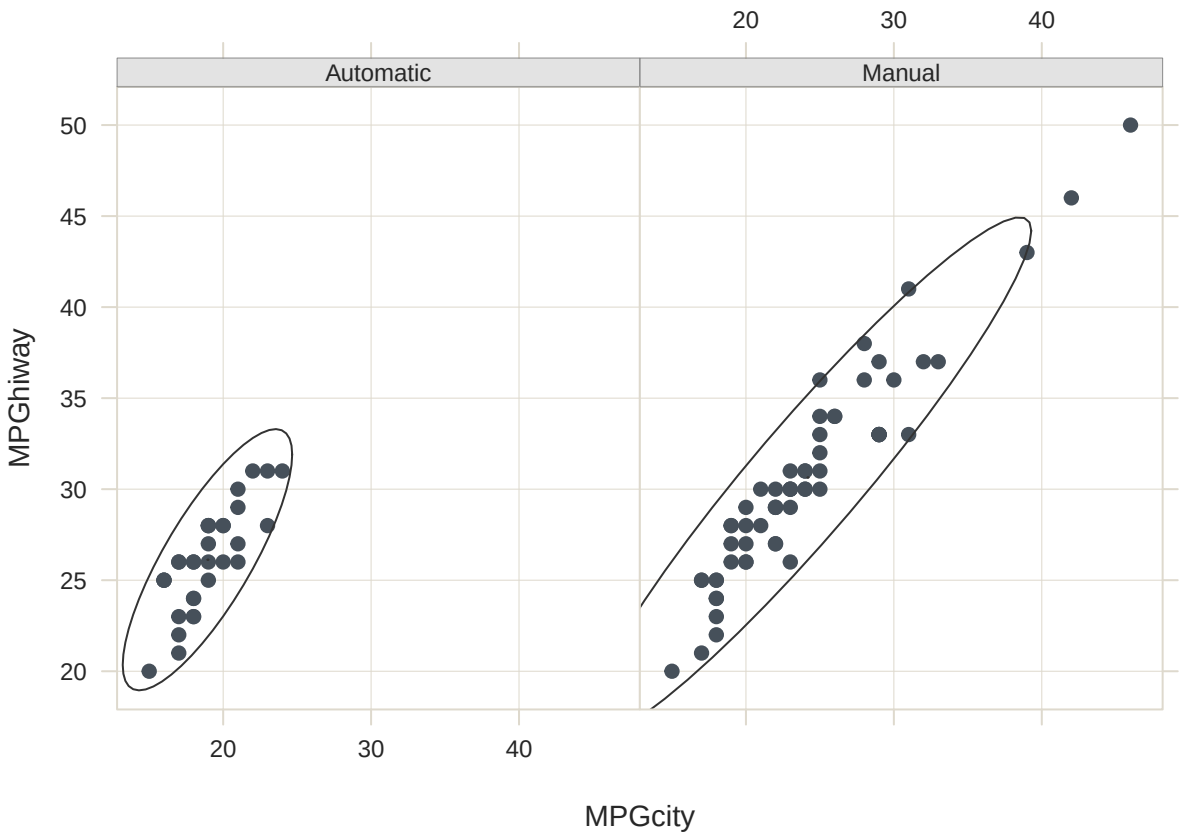
and in ggplot2:

```
ggplot(d, aes(MPGcity, MPGhiway)) + geom_point() + facet_grid(cols=vars(Manual))
```



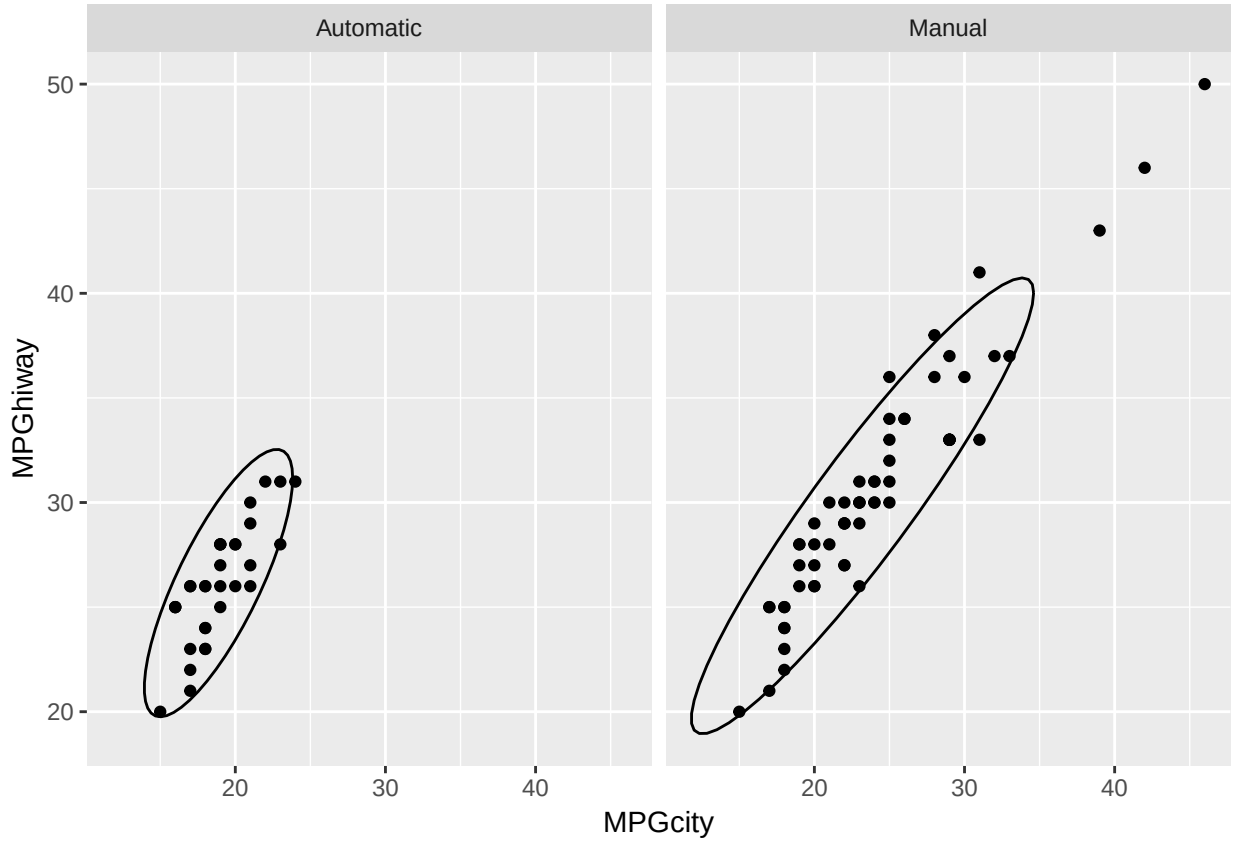
Now let's add ellipses to all of these plots:

```
Plot(MPGcity, MPGhiway, by1=Manual, n.col=2, ellipse=TRUE,  quiet=TRUE)
```



and ggplot2:

```
ggplot(d, aes(MPGcity, MPGhiway)) + geom_point() + facet_grid(cols=vars(Manual))+stat_ellipse()
```

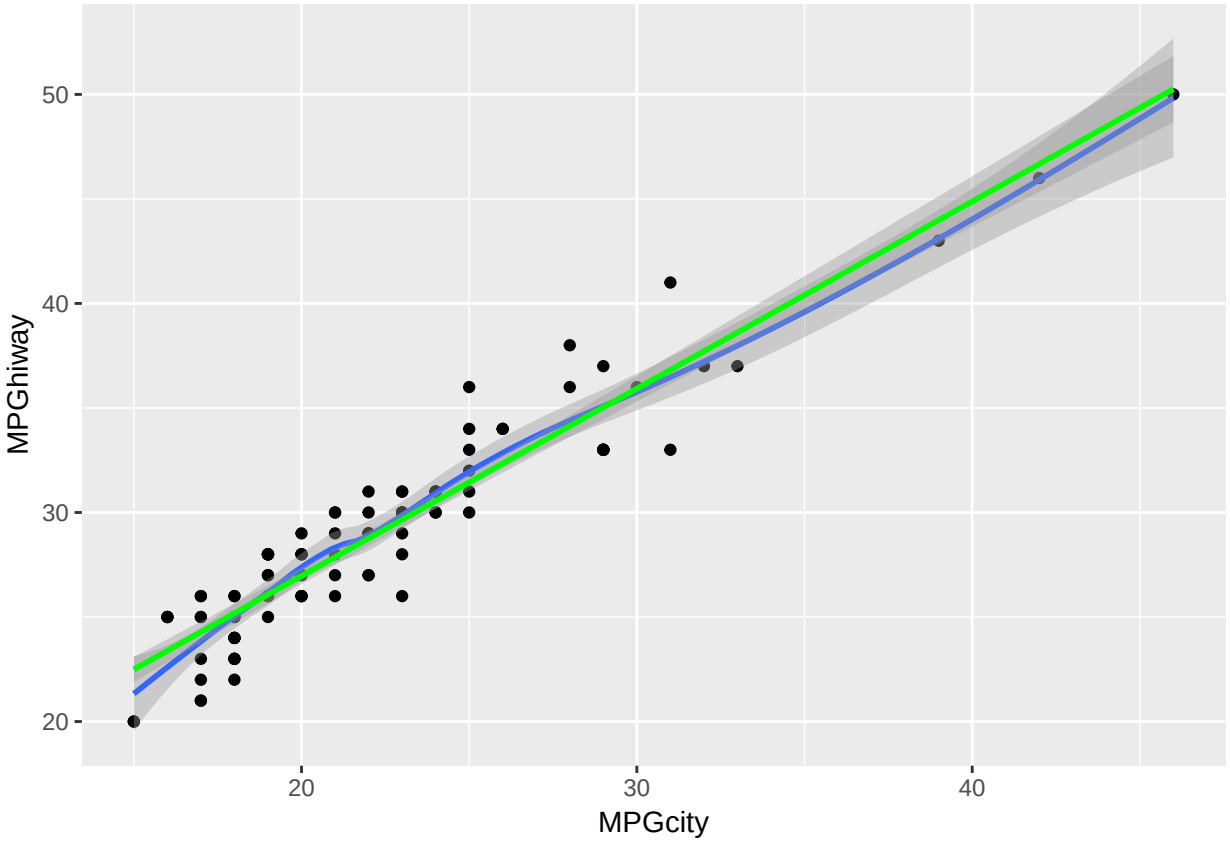


c.

Now we're adding nonlinear and linear fit lines to the ggplot2 scatterplot, along with a 95% CI:

```
ggplot(d, aes(MPGcity, MPGhiway)) + geom_point()+geom_smooth()+geom_smooth(method=lm, color="green", se=
```

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

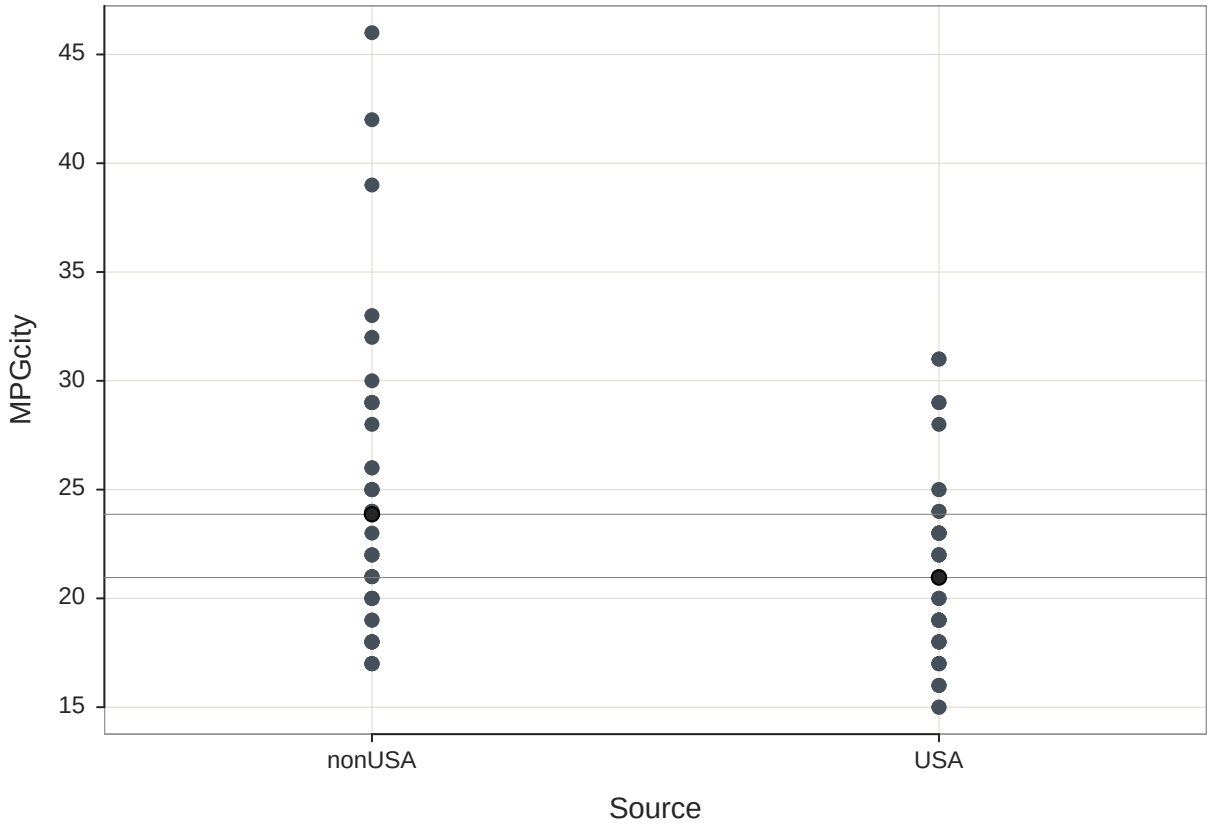


d.

Here are the scatter plots for city MPG for each source, including the mean, in lessR:

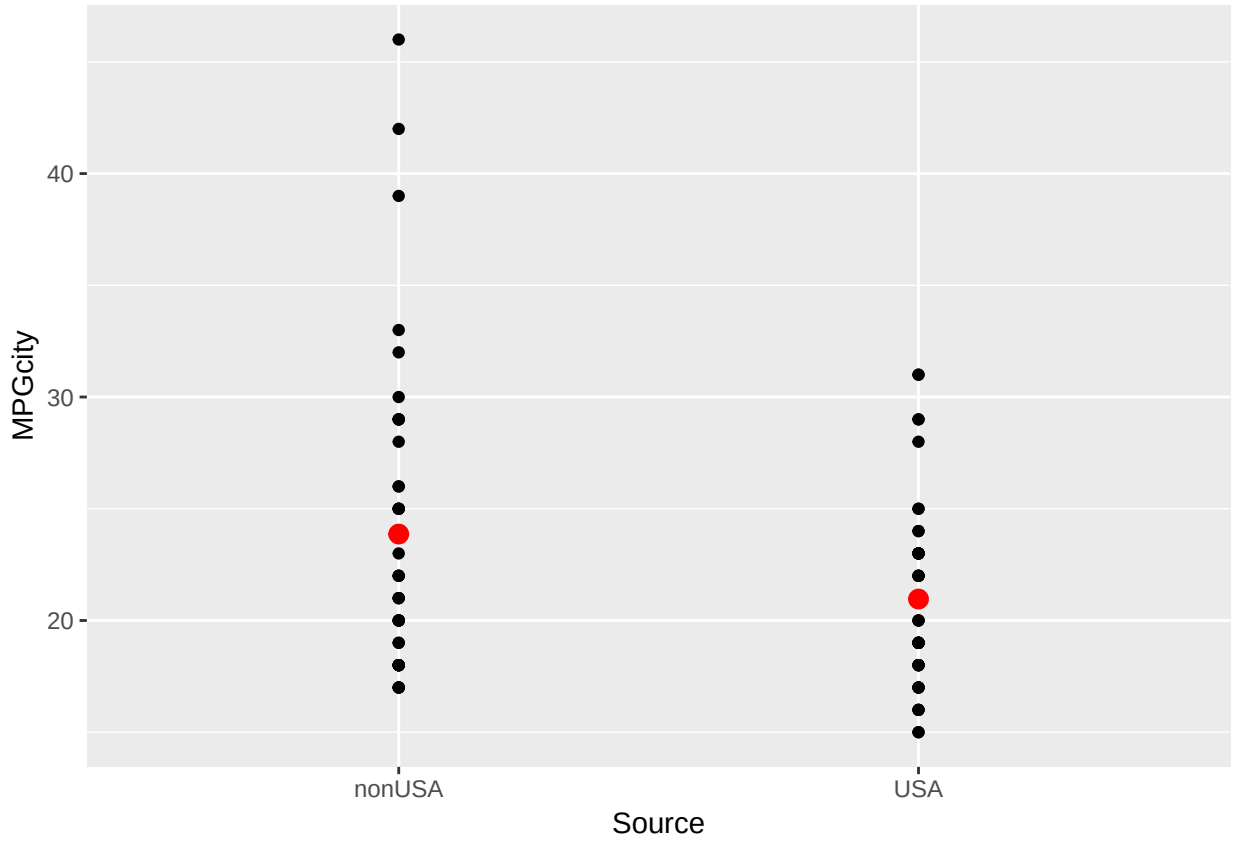
```
Plot(Source, MPGcity, quiet=TRUE)
```

```
##
## >>> Note
## The integrated Violin/Box/Scatterplot (VBS) for MPGcity
## at each level of Source is only obtained if the categorical
## variable is the variable listed second, that is, the y-variable.
##
## This ordering with Source listed first yields the
## scatterplot and the associated means, but no VBS plot.
```



and in ggplot2. I had trouble drawing the horizontal line so I've just marked the mean in red with a larger point:

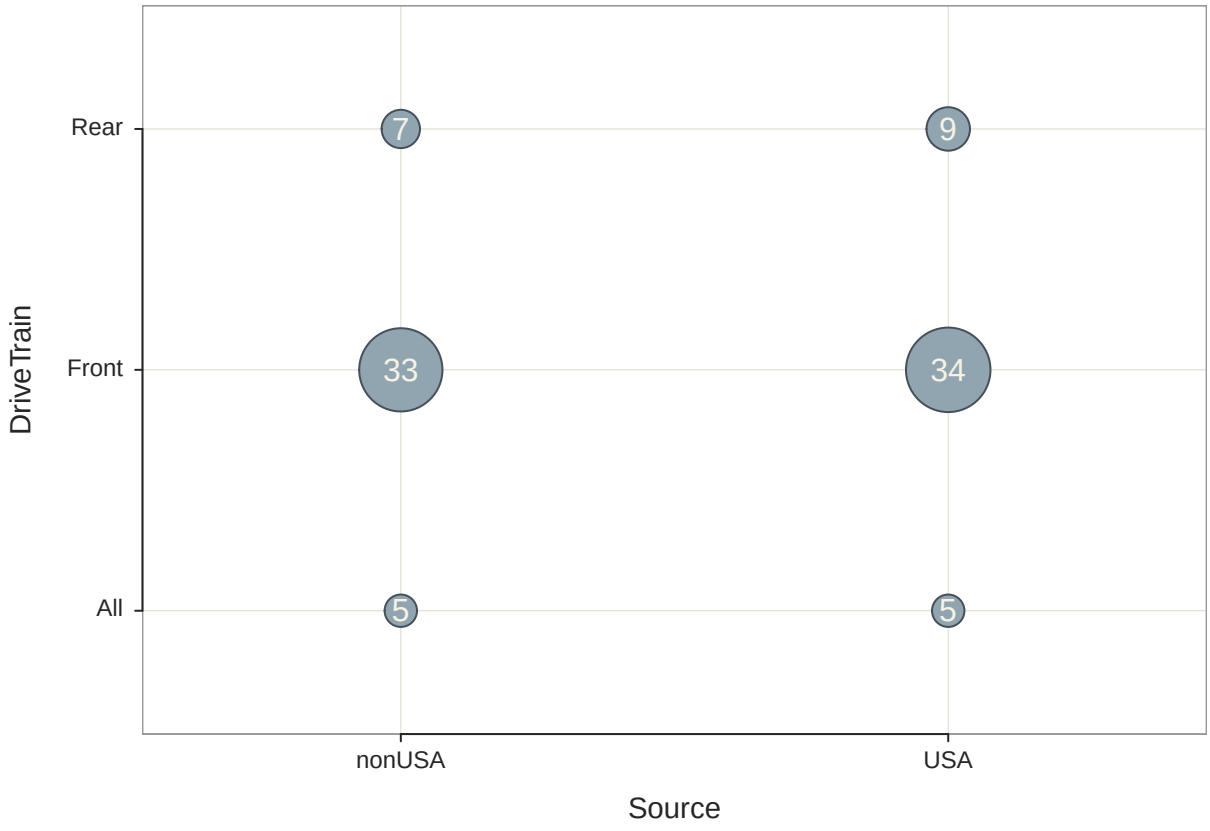
```
ggplot(d, aes(Source, MPGcity))+geom_point()+stat_summary(fun.y="mean",color="red", geom="point", size=3
```



e.

Here's the lessR bubble plot for source of car by frequency of drive train:

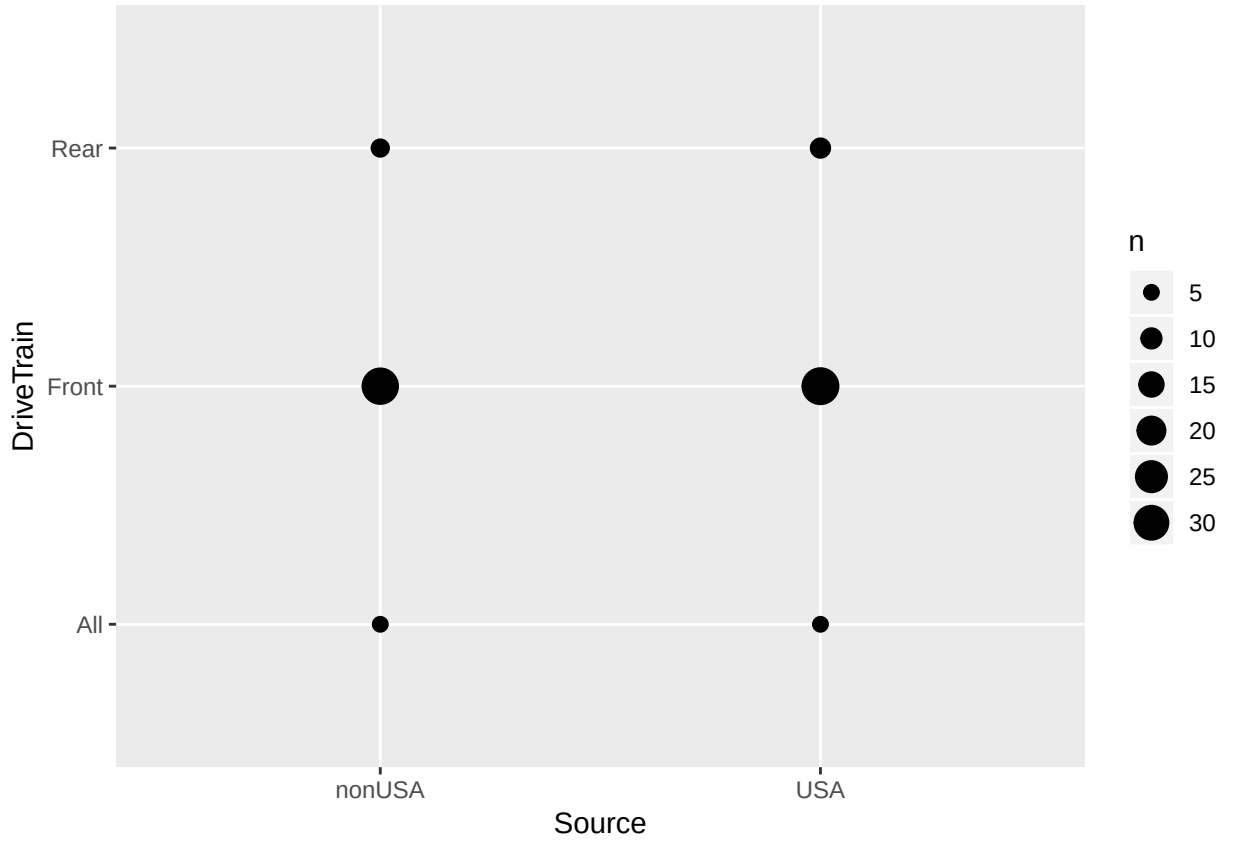
```
Plot(Source, DriveTrain)
```




```
## >>> Suggestions
## Plot(Source, DriveTrain, size.cut=FALSE)
## Plot(Source, DriveTrain, trans=.8, bg="off", grid="off")
## SummaryStats(Source, DriveTrain)  # or ss
##
##
## Joint and Marginal Frequencies
## ------------------------------
##
##          Source
## DriveTrain   nonUSA USA Sum
##   All            5   5  10
##   Front         33  34  67
##   Rear           7   9  16
##   Sum           45  48  93
##
##
## Cramer's V: 0.043
##
## Chi-square Test:  Chisq = 0.168, df = 2, p-value = 0.919
## >>> Low cell expected frequencies, chi-squared approximation may not be accurate
```

and in ggplot2:

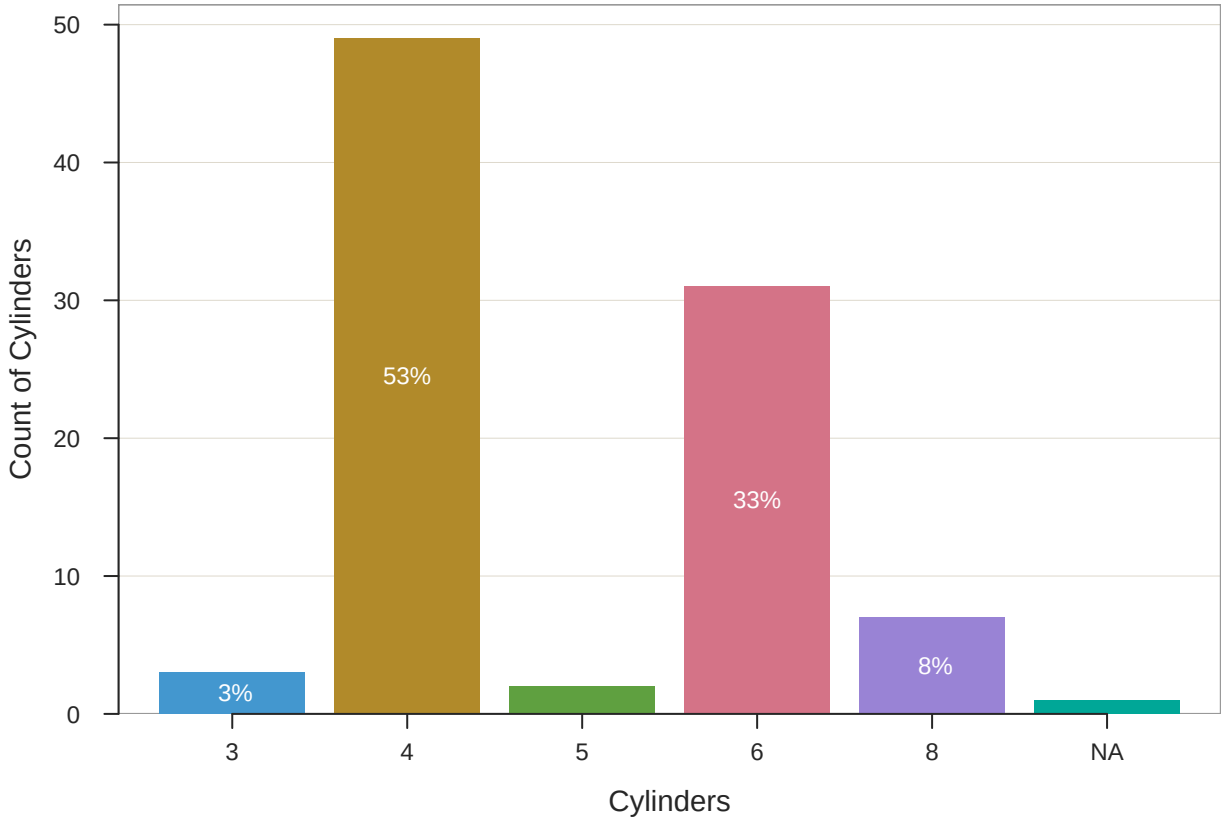
```
ggplot(d, aes(Source, DriveTrain)) + geom_count() +
  scale_size_area()
```



4. Colors

a. Here's a bar chart of number of cylinders where I'm using qualitative HCL colors with adjusted luminosity:

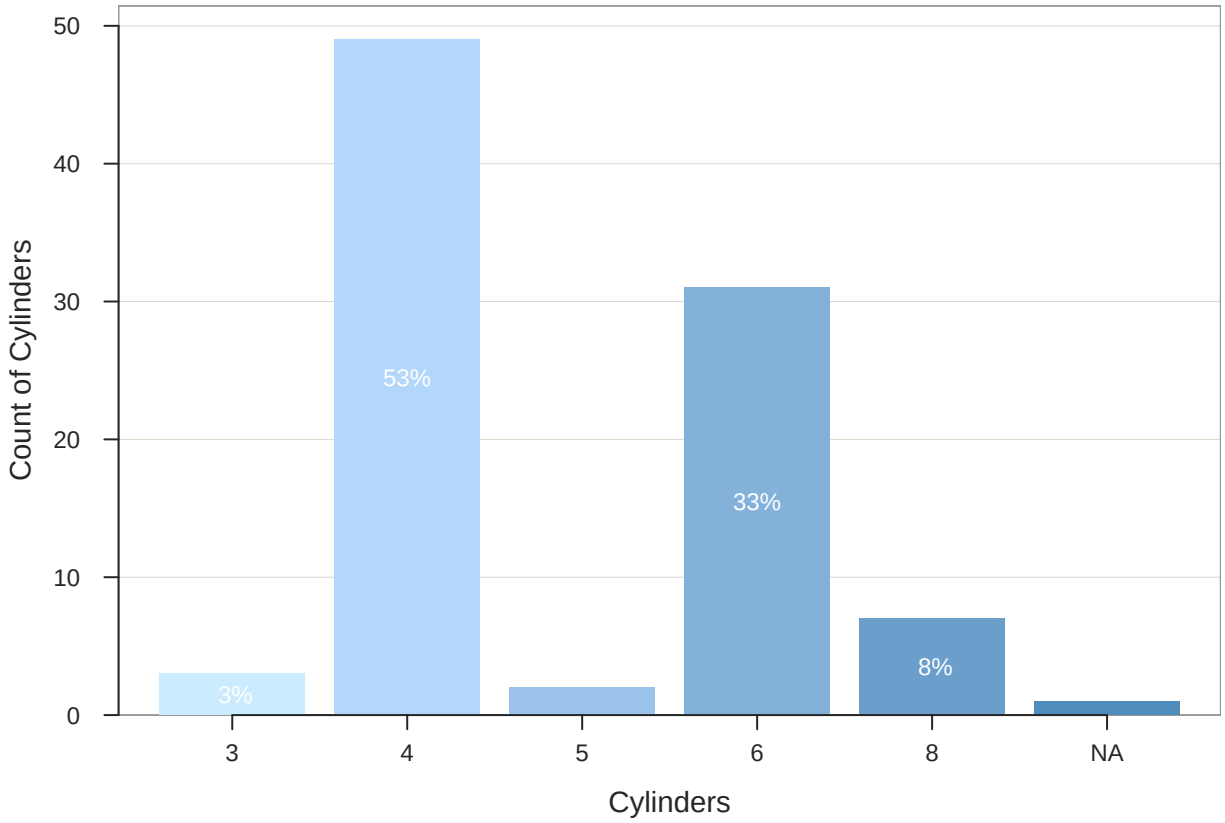
```
bc(Cylinders, fill=getColors(l=60), quiet=TRUE)
```



b.

Here are progressive colors with a blue hue:

```
bc(Cylinders, fill=getColors("blues"))
```




```
## >>> Suggestions
## BarChart(Cylinders, horiz=TRUE)  # horizontal bar chart
## BarChart(Cylinders, fill="greens")  # sequential green bars
## PieChart(Cylinders)  # doughnut (ring) chart
## Plot(Cylinders)  # bubble plot
## Plot(Cylinders, stat="count")  # lollipop plot
##
##
## --- Cylinders ---
##
##
## Missing Values of Cylinders: 0
##
##
##                    3      4      5      6      8     NA    Total
## Frequencies:       3     49      2     31      7      1       93
## Proportions:   0.032  0.527  0.022  0.333  0.075  0.011    1.000
##
##
## Chi-squared test of null hypothesis of equal probabilities
##   Chisq = 127.968, df = 5, p-value = 0.000
```

c.

Here's the same chart with the base R heat map color scheme:

```
bc(Cylinders, fill=getColors("heat"), quiet=TRUE)
```