Data Science for Operational Researchers Using R Online

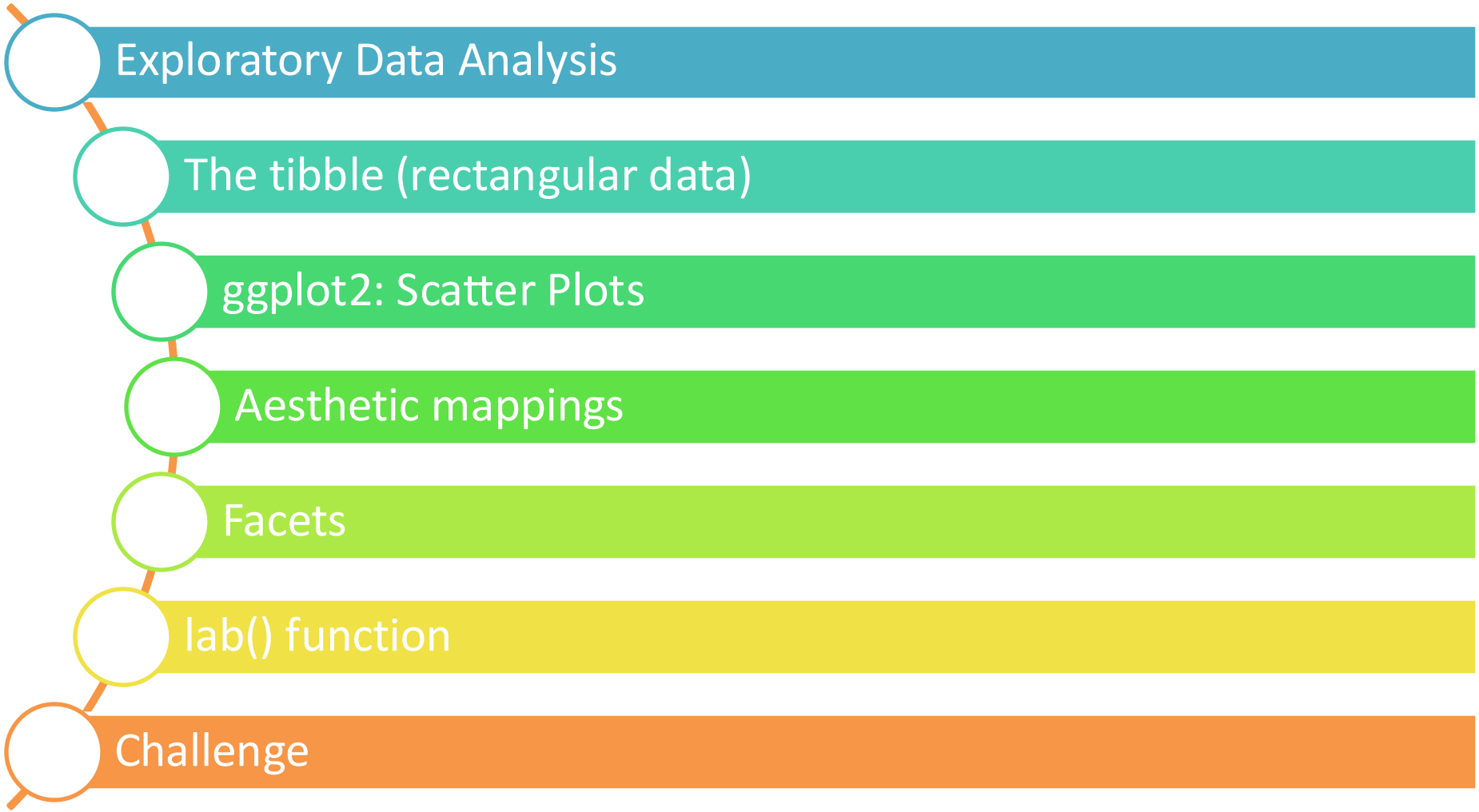
2. Exploratory Data Analysis with ggplot2

Prof. Jim Duggan,
School of Computer Science
University of Galway.

https://github.com/JimDuggan/explore_or

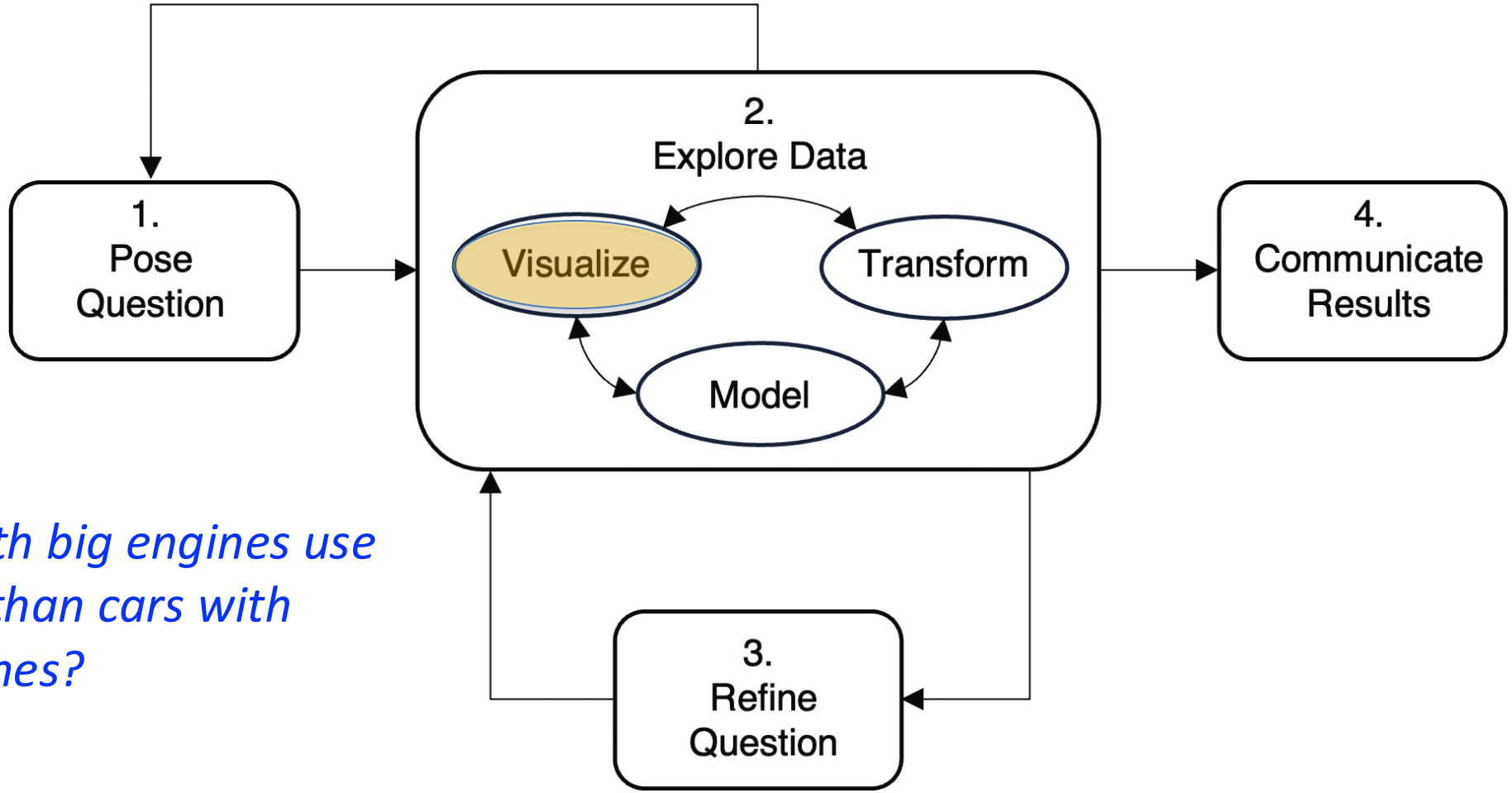
> Exploratory analysis is what you do to understand the data and figure out what might be noteworthy or interesting to highlight to others. When we do exploratory analysis, it's like hunting for pearls in oysters.
>
> — Cole Nussbaumer Knaflic (Knaflic, 2015)

Overview

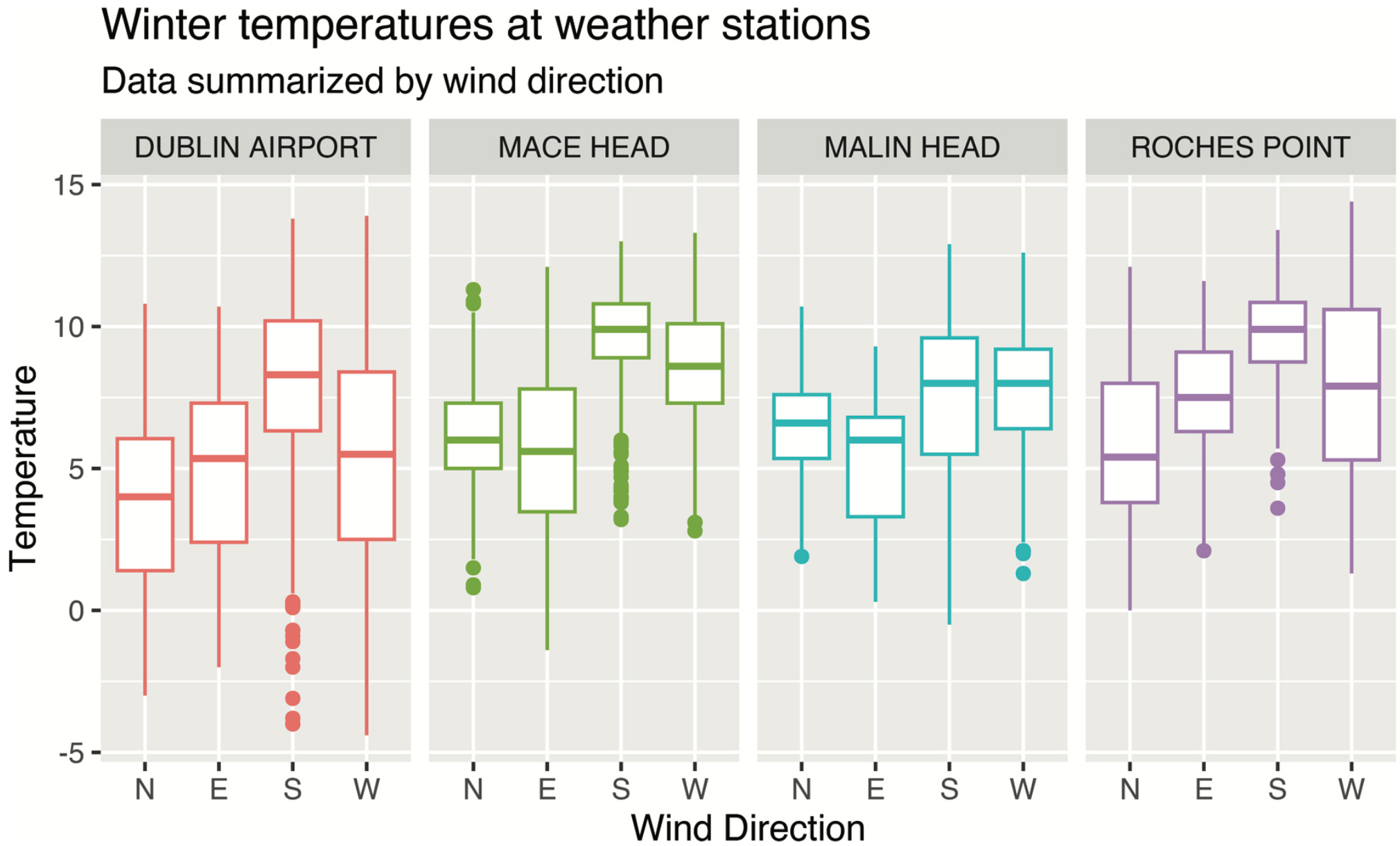

Exploratory Data Analysis
The tibble (rectangular data)
ggplot2: Scatter Plots
Aesthetic mappings
Facets
lab() function
Challenge

1. Exploratory Data Analysis

(Wickham and Grolemund 2016)



Do cars with big engines use more fuel than cars with small engines?


Winter temperatures at weather stations
Data summarized by wind direction
DUBLIN AIRPORT
MACE HEAD
MALIN HEAD
ROCHES POINT
Temperature
15
10
5
0
-5
N
E
S
W
Wind Direction

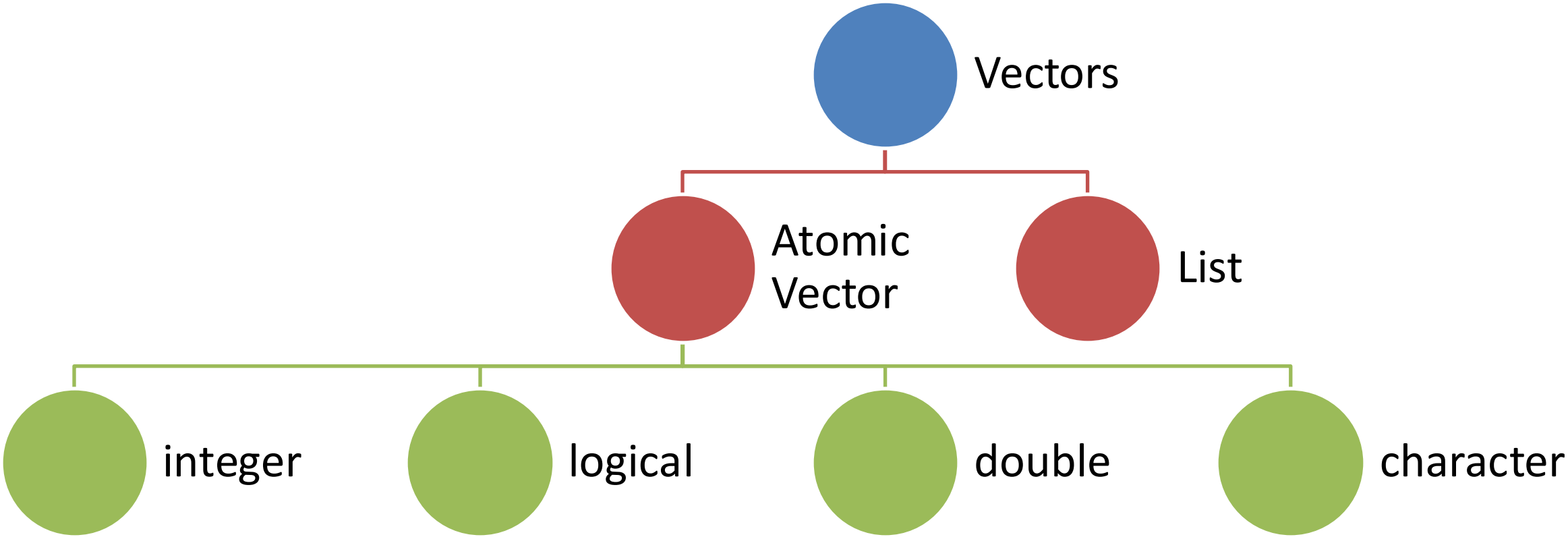
2. The tibble/data frame

> On an intuitive level, a *data frame* is like a matrix, with a two-dimensional rows-and-columns structure. However, it differs from a matrix in that each column may have a different type.
>
> — Norman Matloff (Matloff, 2011)

Definition

- A data frame is two-dimensional row and column structure, while on a technical level, a data frame is a list, with the elements of that list containing equal length vectors (Matloff, 2011).
- It's defined using the data.frame() or tibble() function
- The elements (columns) of a data frame can be of different types
- The data frame, with its row and column structure, will be familiar to anyone who has used a spreadsheet, where each column is a variable (feature), and every row is an observation.

Vectors in R (covered later)



library(ggplot2) has the tibbles mpg and diamonds

ggplot2::mpg, N = 234

Variable	Description
manufacturer	Manufacturer name
model	Model name
displ	Engine displacement (liters)
year	Year of manufacture
cyl	Number of cylinders
trans	Type of transmission
drv	Type of drive train (e.g. front wheel)
cty	City miles per gallon
hwy	Highway miles per gallon
fl	Fuel type
class	"type" of car (e.g. "compact")

ggplot2::diamonds, N = 53,940

Variable	Description
carat	Weight of the diamond
cut	Quality of the cut (categorical)
color	Diamond color (categorical)
clarity	Diamond clarity (categorical)
depth	Total depth percentage
table	Measure related to width of diamond top
price	Price in dollars
x	Length in mm
y	Width in mm
z	Depth in mm

View(mpg)

	manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class
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	2.0	2008	4	manual(m6)	f	20	31	p	compact
4	audi	a4	2.0	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	2.0	2008	4	manual(m6)	4	20	28	p	compact

Exploring at the console

```
> mpg
# A tibble: 234 × 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
# ℹ 224 more rows
# ℹ Use `print(n = ...)` to see more rows
```

summary(mpg)

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

Datasets and tidy data

- With tidy data, where every column is a variable, and every row is an observation.
- These are defined as (Wickham, 2016):
  - A variable is a quantity, quality, or property that you can measure, and will have a value at the time it is measured.
  - An observation is a set of measurements made under similar conditions (often at the same time)

head(mpg)

```
> head(mpg)
# A tibble: 6 × 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
```

tail(mpg)

```
> tail(mpg)
# A tibble: 6 × 11
  manufacturer model  displ  year   cyl trans      drv     cty   hwy fl    class
  <chr>        <chr>  <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
1 volkswagen   passat   1.8  1999     4 auto(l5)   f        18    29 p     midsize
2 volkswagen   passat   2    2008     4 auto(s6)   f        19    28 p     midsize
3 volkswagen   passat   2    2008     4 manual(m6) f        21    29 p     midsize
4 volkswagen   passat   2.8  1999     6 auto(l5)   f        16    26 p     midsize
5 volkswagen   passat   2.8  1999     6 manual(m5) f        18    26 p     midsize
6 volkswagen   passat   3.6  2008     6 auto(s6)   f        17    26 p     midsize
```

dplyr::sample_n()

```
> dplyr::sample_n(mpg,10)
# A tibble: 10 × 11
   manufacturer model                displ  year   cyl trans      drv     cty   hwy fl    class
   <chr>        <chr>                <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
 1 hyundai      tiburon                2.7  2008     6 manual(m6) f        16    24 r     subcompact
 2 nissan       pathfinder 4wd         5.6  2008     8 auto(s5)   4        12    18 p     suv
 3 dodge        durango 4wd            4.7  2008     8 auto(l5)   4         9    12 e     suv
 4 toyota       land cruiser wagon 4wd 4.7  1999     8 auto(l4)   4        11    15 r     suv
 5 hyundai      tiburon                2    2008     4 auto(l4)   f        20    27 r     subcompact
 6 volkswagen   jetta                  2    2008     4 auto(s6)   f        22    29 p     compact
 7 ford         mustang                4.6  2008     8 auto(l5)   r        15    22 r     subcompact
 8 pontiac      grand prix             3.1  1999     6 auto(l4)   f        18    26 r     midsize
 9 dodge        caravan 2wd            3.8  1999     6 auto(l4)   f        15    21 r     minivan
10 toyota       4runner 4wd            3.4  1999     6 auto(l4)   4        15    19 r     suv
```

dplyr::sample_frac()

```
> dplyr::sample_frac(mpg,.05)
# A tibble: 12 × 11
   manufacturer model              displ  year   cyl trans      drv     cty   hwy fl    class
   <chr>        <chr>              <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
 1 nissan       altima               2.4  1999     4 manual(m5) f        21    29 r     compact
 2 ford         explorer 4wd         5    1999     8 auto(l4)   4        13    17 r     suv
 3 hyundai      sonata               2.5  1999     6 auto(l4)   f        18    26 r     midsize
 4 dodge        caravan 2wd          3.3  1999     6 auto(l4)   f        16    22 r     minivan
 5 volkswagen   new beetle           2.5  2008     5 auto(s6)   f        20    29 r     subcompact
 6 toyota       toyota tacoma 4wd    2.7  1999     4 manual(m5) 4        15    20 r     pickup
 7 ford         mustang              4.6  2008     8 manual(m5) r        15    23 r     subcompact
 8 toyota       toyota tacoma 4wd    3.4  1999     6 auto(l4)   4        15    19 r     pickup
 9 jeep         grand cherokee 4wd   5.7  2008     8 auto(l5)   4        13    18 r     suv
10 volkswagen   gti                  2.8  1999     6 manual(m5) f        17    24 r     compact
11 subaru       impreza awd          2.5  1999     4 auto(l4)   4        19    26 r     subcompact
12 volkswagen   jetta                2.8  1999     6 auto(l4)   f        16    23 r     compact
```

set.seed(n) – replicate sample

```
> set.seed(100)
> dplyr::sample_n(mpg,5)
# A tibble: 5 × 11
  manufacturer model              displ  year   cyl trans      drv     cty   hwy fl    class
  <chr>        <chr>              <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
1 toyota       toyota tacoma 4wd    2.7  1999     4 auto(l4)   4        16    20 r     pickup
2 honda        civic                1.6  1999     4 manual(m5) f        25    32 r     subcompact
3 hyundai      sonata               2.4  2008     4 manual(m5) f        21    31 r     midsize
4 volkswagen   jetta                2    2008     4 manual(m6) f        21    29 p     compact
5 toyota       toyota tacoma 4wd    4    2008     6 manual(m6) 4        15    18 r     pickup
```

3. ggplot2 – Scatter Plots

> Data graphics provide one of the most accessible, compelling, and expressive modes to investigate and depict patterns in data.
>
> — Benjamin S. Baumer, Daniel T. Kaplan, and Nicholas J. Horton (Baumer et al., 2021)

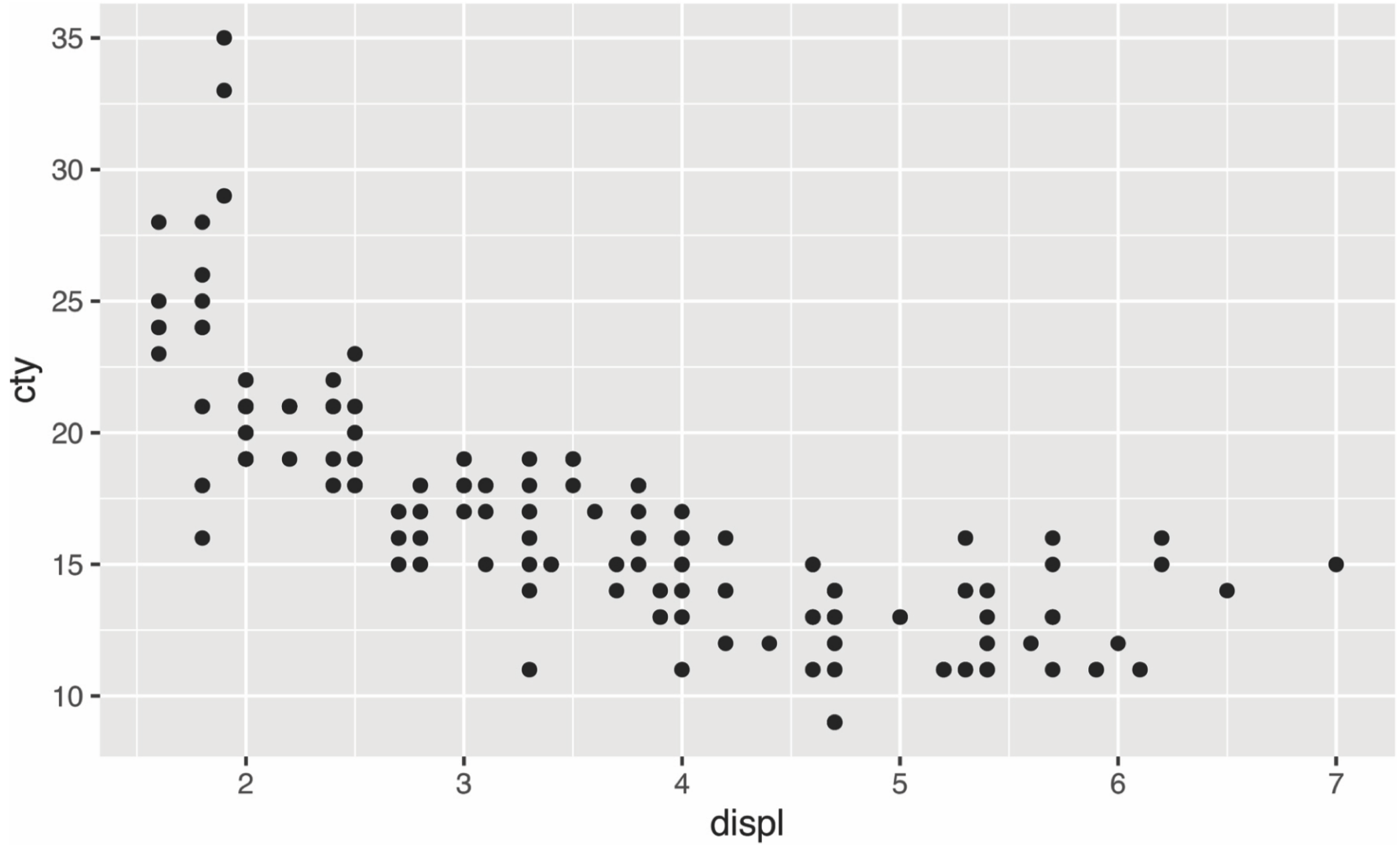
ggplot2

- A core part of any data analysis and modelling process is to visualize data and explore relationships between variables
- There are three important benefits of ggplot2:
  1. plots can be designed in a layered manner, where additional plotting details can be added using the + operator;
  2. a wide range of plots can be generated to support decision analysis, including scatterplots, histograms, and time series charts,
  3. once the analyst is familiar with the structure and syntax of ggplot2, charts can be developed rapidly, and this supports an iterative process of decision support.

Creating our first graph (scatterplot)

1. We call ggplot(data=mpg) which initializes a ggplot object, and this call also allows us to specify the tibble that contains that data.
2. We extend this call to include the x-axis and y-axis variables by including an addition argument (mapping) and the function aes() which 7.4 Aesthetic mappings describe how variables in data are mapped to the plot's visual properties.
3. To visualize the set of points on the graph, and we do this by calling the relevant geometric object, which is one that is designed to draw points, namely the function geom_point()

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

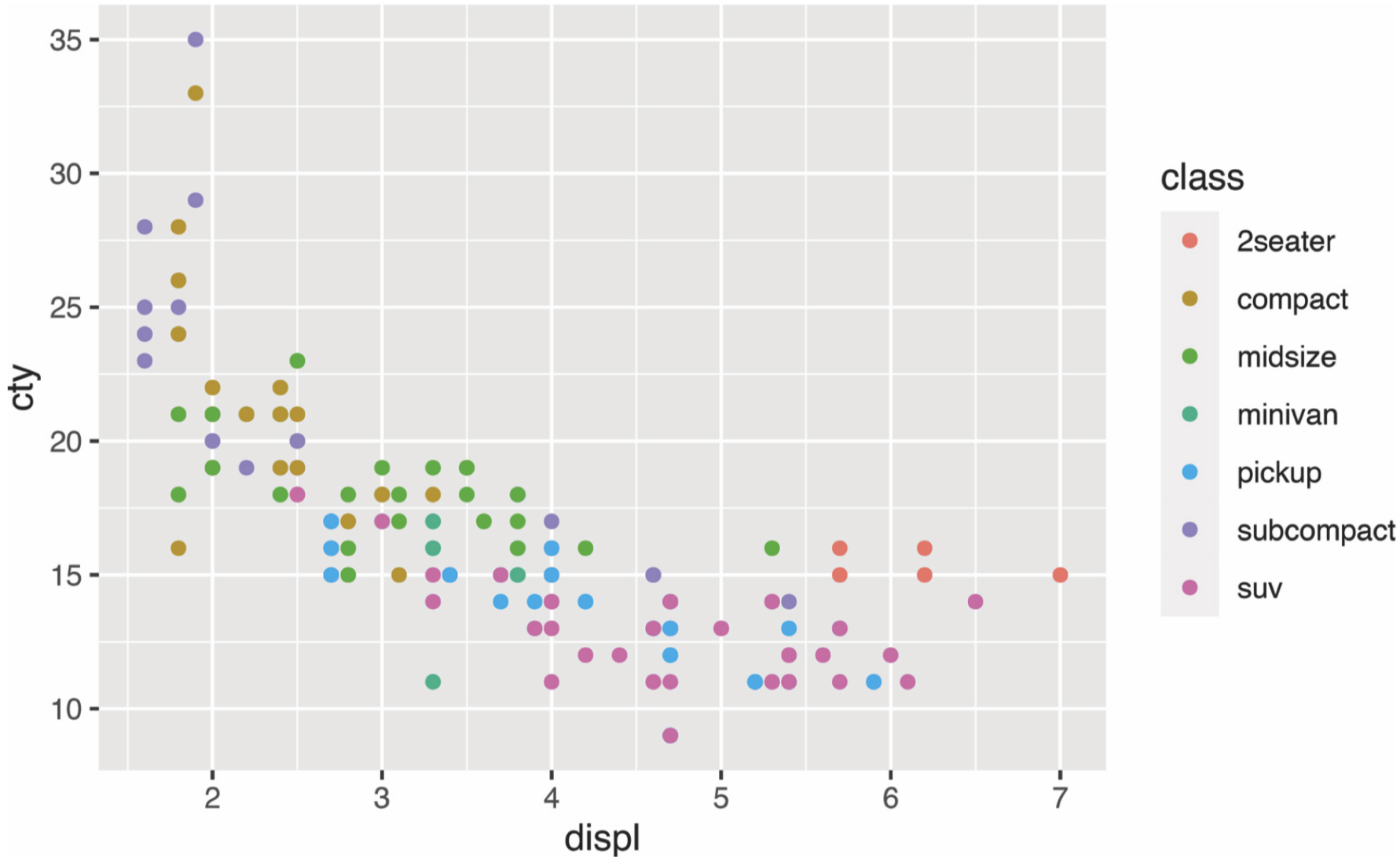


4. Aesthetic Mappings

- The aes() function has some nice additional features that can be used to add extra information to a plot by using data from other variables.
- For example, what if we also wanted to see which class of car each point belonged to?
- We can set the argument color to this class in the aes() function

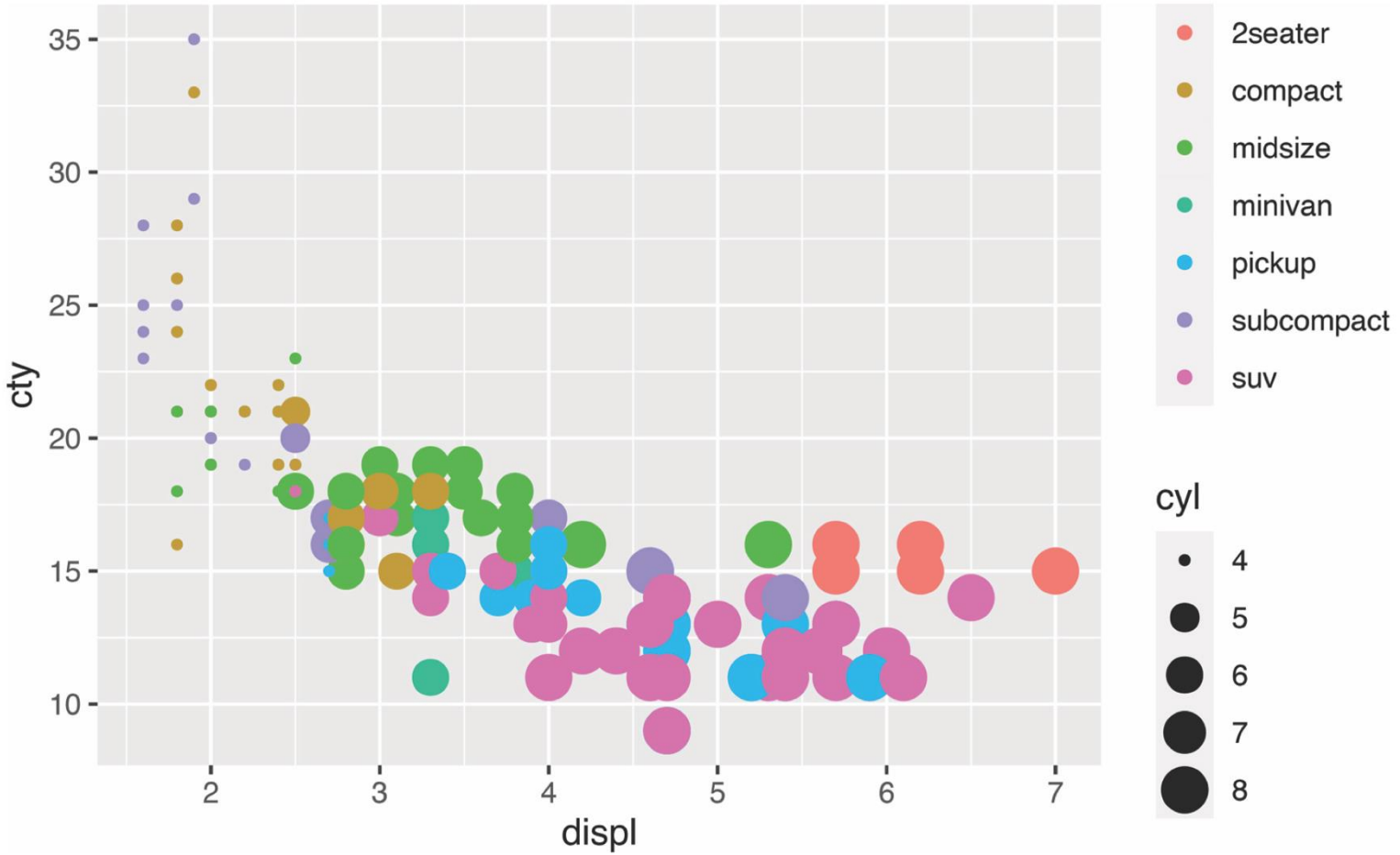
```
unique(mpg$class)
#> [1] "compact"    "midsize"    "suv"        "2seater"
#> [5] "minivan"    "pickup"     "subcompact"
```

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



The size argument

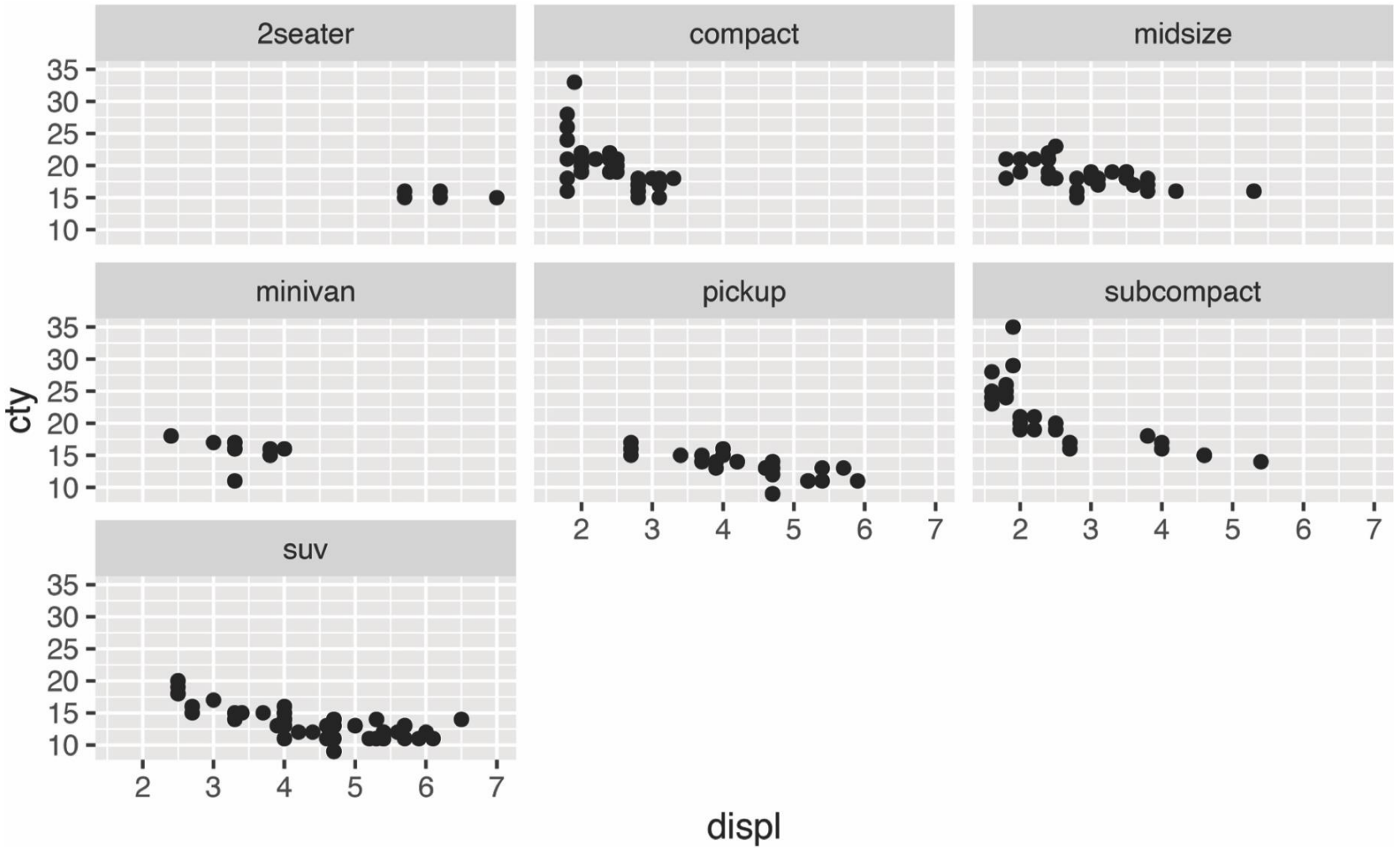
```
ggplot(data=mpg,mapping=aes(x=displ,y=cty,color=class,size=cyl))+
  geom_point()
```



5. Subplots with facets

- What if we needed to drill down on the plots and show, for example, the relationships for each class of car on separate plots?
- Or, in the more general case, sub-divide a plot into multiple plots based on another variable.
- The function facet_wrap() will do this in ggplot2, and all it needs as an argument is the variable for dividing the plots, which must be preceded by the tilde (~) operator.

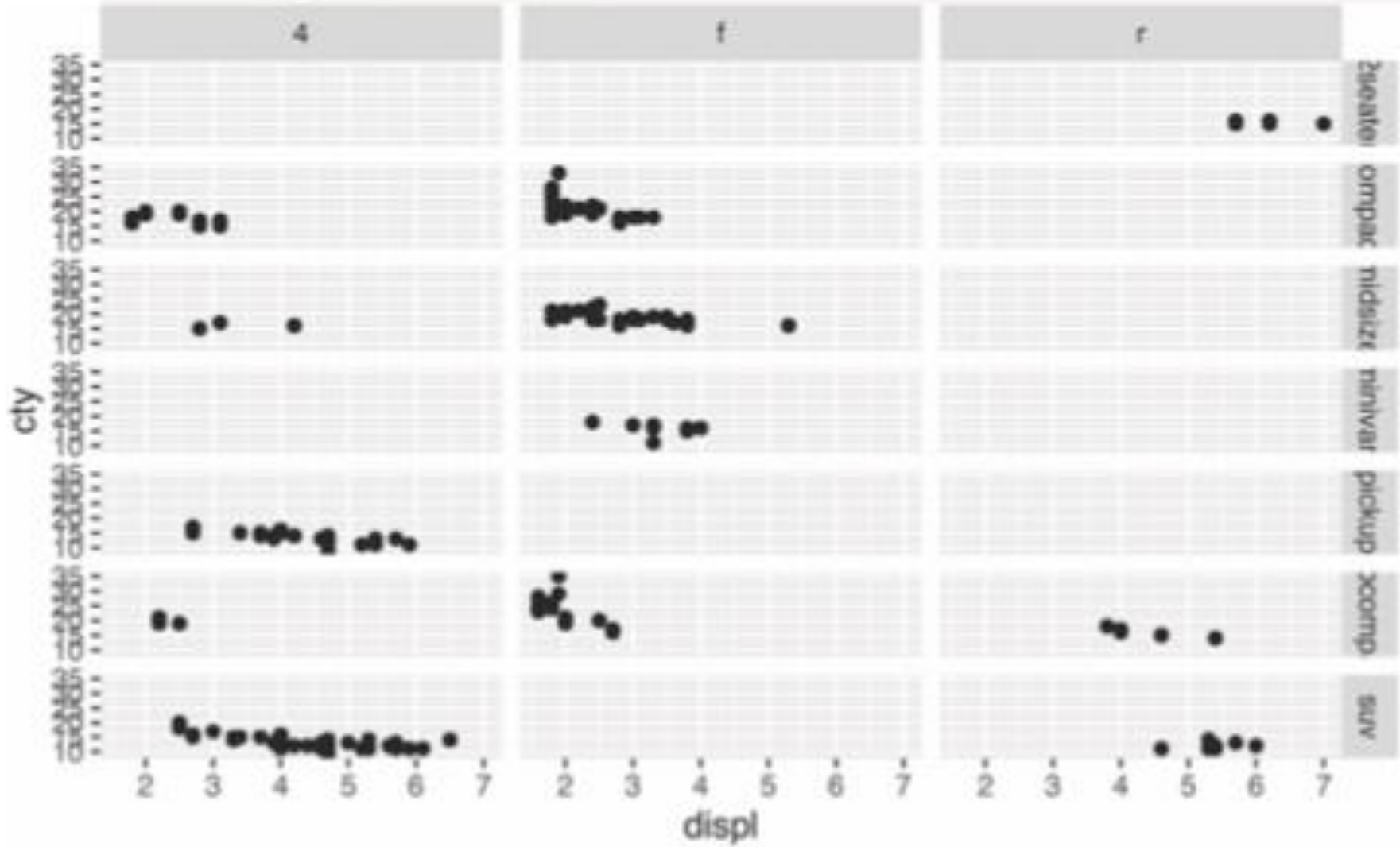
```
ggplot(data=mpg,aes(x=displ,y=cty))+
  geom_point()+
  facet_wrap(~class)
```



Using facet_grid()

- An extra variable can be added to the faceting process by using the related function facet_grid(), which takes two arguments, separated by the ~ operator.
- The first argument specifies which variable is to be mapped to each row, and the second argument identifies the variable to be represented on the columns.
- For example, we may want to generate 21 plots that show the type of drive (drv) on the columns, and the class of car (class) shown on each row.

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



The lab() function – name-value pairs

- title provides an overall title text for the plot.
- subtitle adds a subtitle text.
- color allows you to specify the legend name for the color attribute.
- caption inserts text on the lower right-hand side of the plot.
- size, where you can name the size attribute.
- x to name the x-axis.
- y to name the y-axis.
- tag, the text for the tag label to be displayed on the top left of the plot.

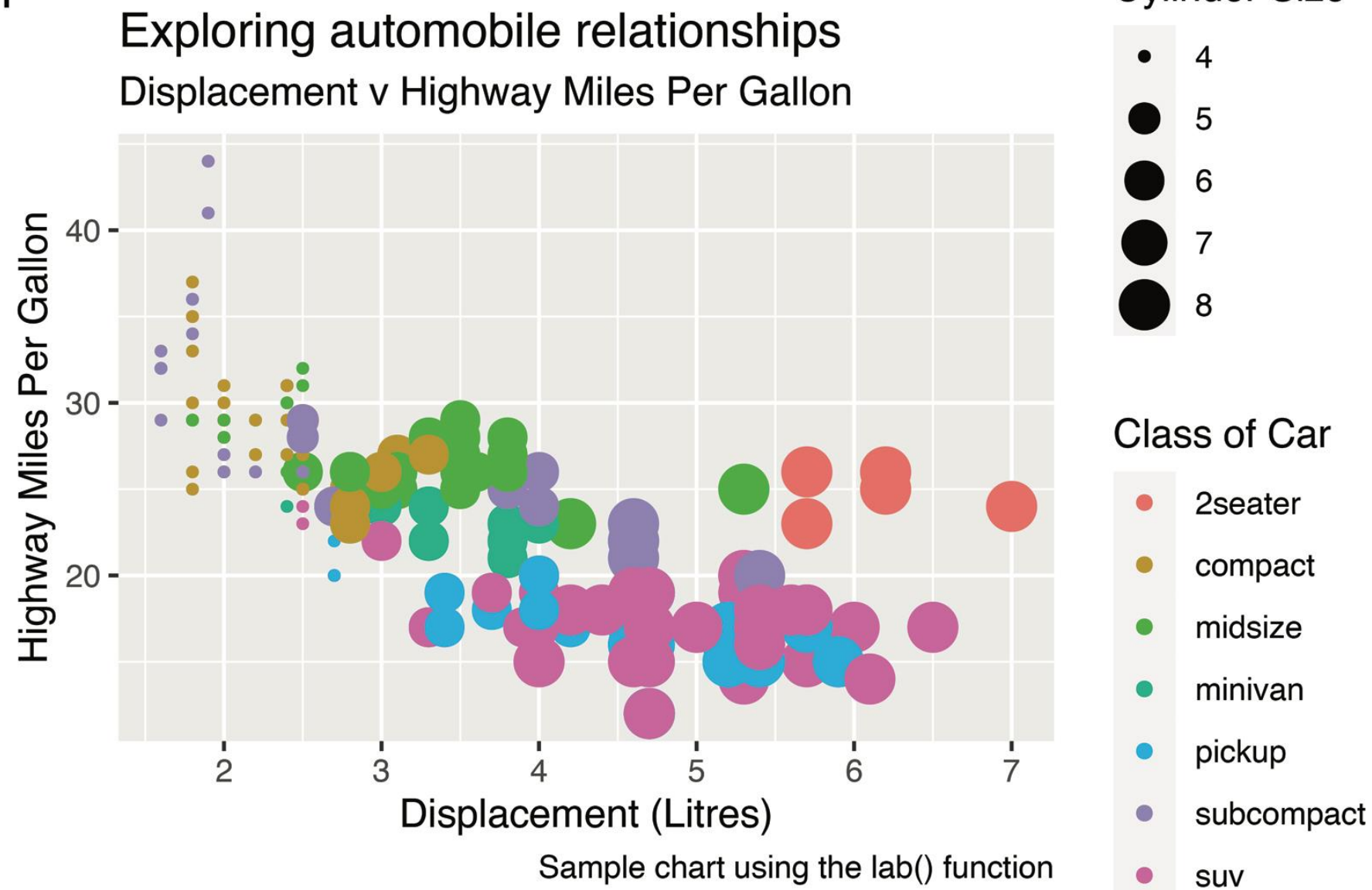
Note that plots can be variables (p1)

```
p1 <- ggplot(data=mpg,aes(x=displ,y=hwy,size=cyl,color=class))+
        geom_point()

p1 <- p1 +
       labs(
         title = "Exploring automobile relationships",
         subtitle = "Displacement v Highway Miles Per Gallon",
         color = "Class of Car",
         size = "Cylinder Size",
         caption = "Sample chart using the lab() function",
         tag = "Plot #1",
         x = "Displacement (Litres)",
         y = "Highway Miles Per Gallon"
       )

p1
```

Plot #1


Exploring automobile relationships
Displacement v Highway Miles Per Gallon
Highway Miles Per Gallon
40
30
20
2
3
4
5
6
7
Displacement (Litres)
Sample chart using the lab() function
Cylinder Size
4
5
6
7
8
Class of Car
2seater
compact
midsize
minivan
pickup
subcompact
suv

6. Challenge

1. Generate the following plot from the `mpg` tibble in `ggplot2`. The x-variable is `displ` and the y-variable `cty`. Make use of the `lab()` and `theme()` functions.