Data Visualiza

Kylie A. Bem

Northeastern Uni
Khoury College of Comput

Northeastern U

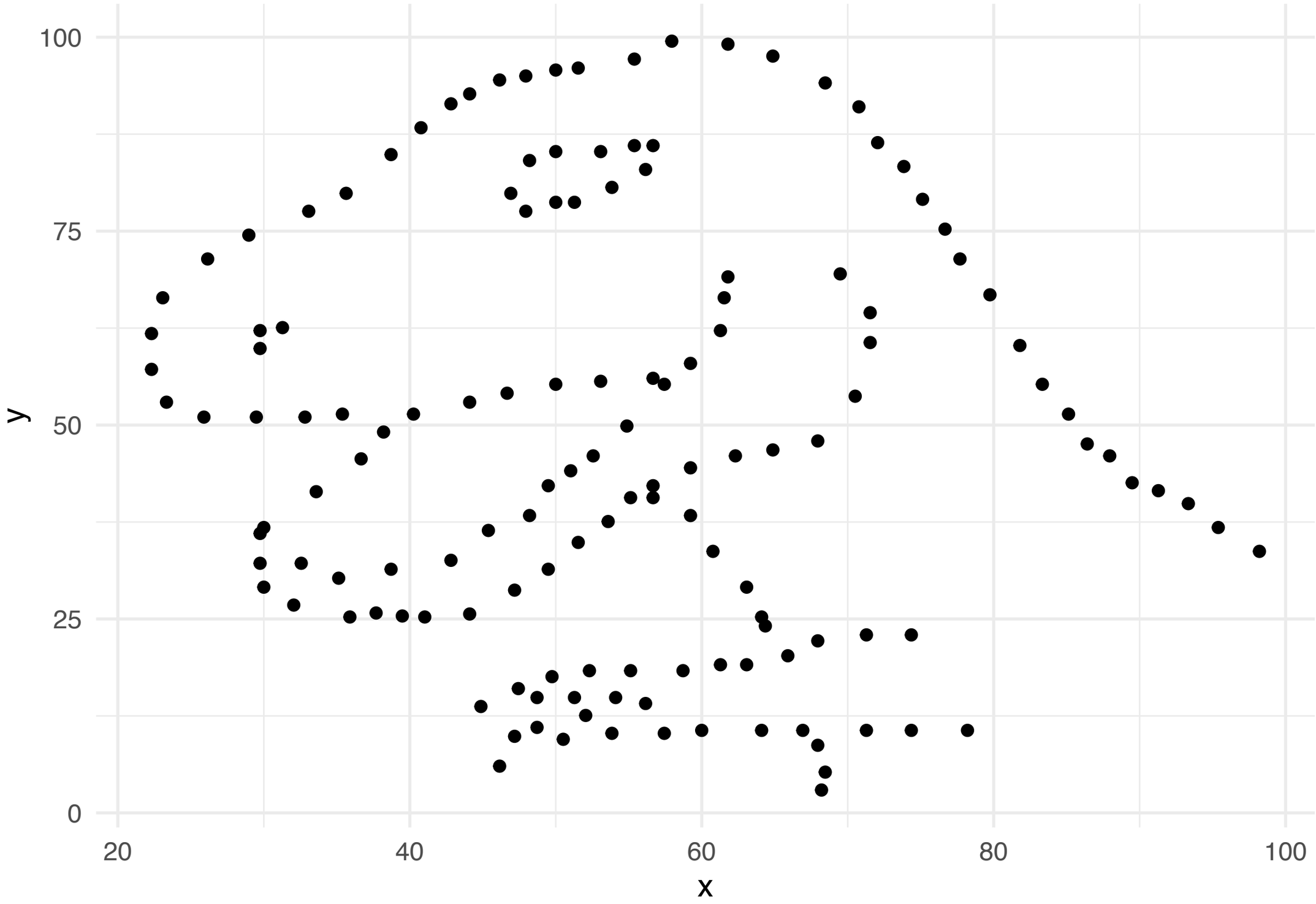
Learning go

- What are common stat
- How to look at data
- Key ingredients of usefu
- A grammar of graphics

STATISTICAL G
HOW TO LOOK

Why do we look

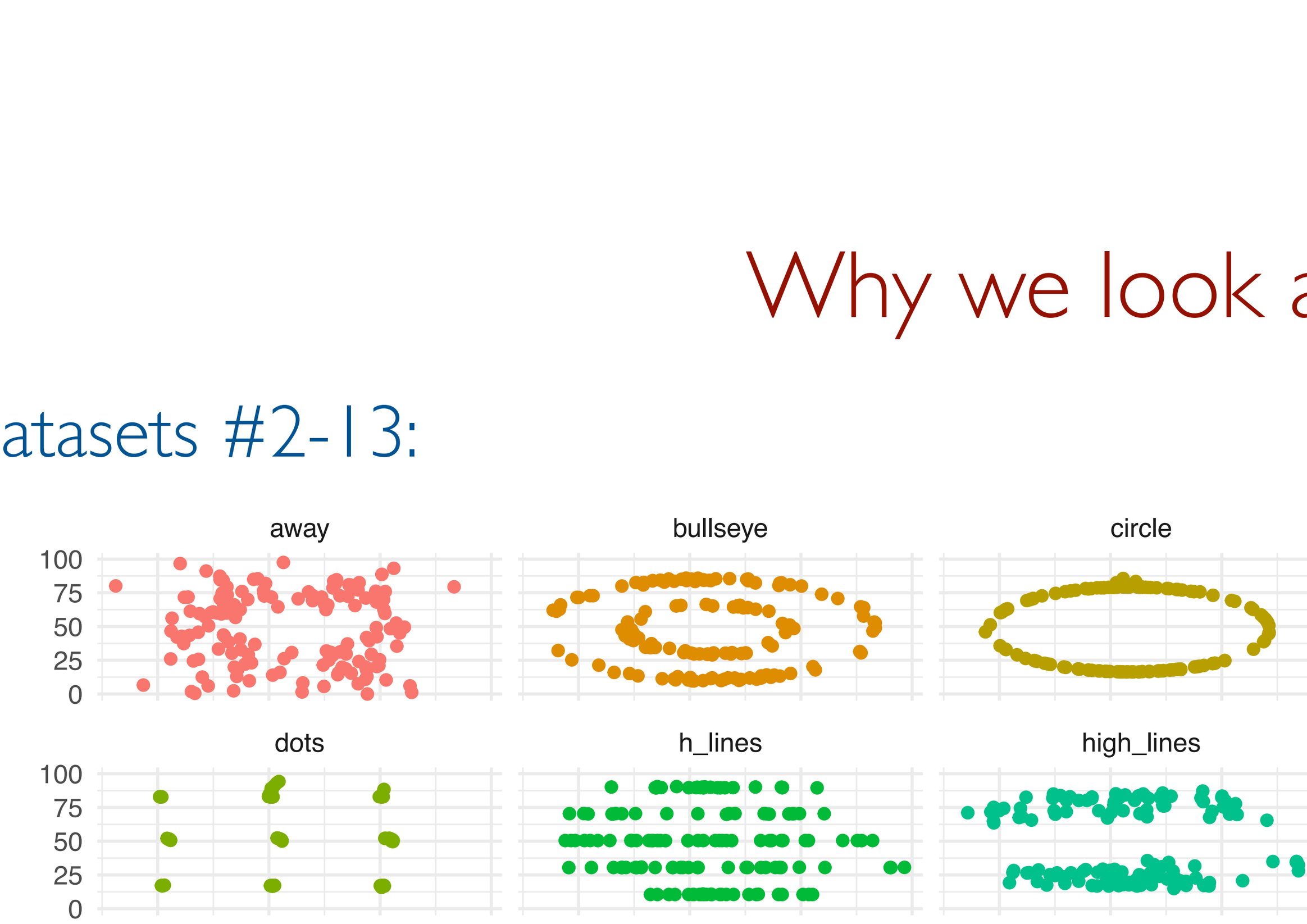
Dataset #1:



How similar are the other 12 datasets?

Why we look a

Datasets #2-13:



"The Datasaurus Dozen": https://www.autodeskrese

Looking at data is

- Summary statistics don’
- Easily find patterns
- Identify potential outlier
- Check model assumptio
- Intuitively display results

What are some common ways

Some common statis

- Scatter plot
- Line plot
- Box-and-whisker plot
- Histogram
- Bar plot

Roles of statistical

One variable

- Histogram
- Bar plot
- Box plot
- Pie chart

Roles of statistical

Distributions

- Histogram
- Bar plot
- Box plot
- Pie chart

SINGLE-VARIABLE

Example data: Ga

Life expectancy, GDP per capita, an

```
gapminder
```

```
## #   tibble: 1,704 x 6
##     country      continent  year lif
##     <fct>        <fct>     <int>   <
##  1  fghanistan   sia        1952
##  2  fghanistan   sia        1957
##  3  fghanistan   sia        1962
##  4  fghanistan   sia        1967
```

http://www.gapminder.or

Example data: Fuel

Fuel economy on 38 popular models o

```
mpg
```

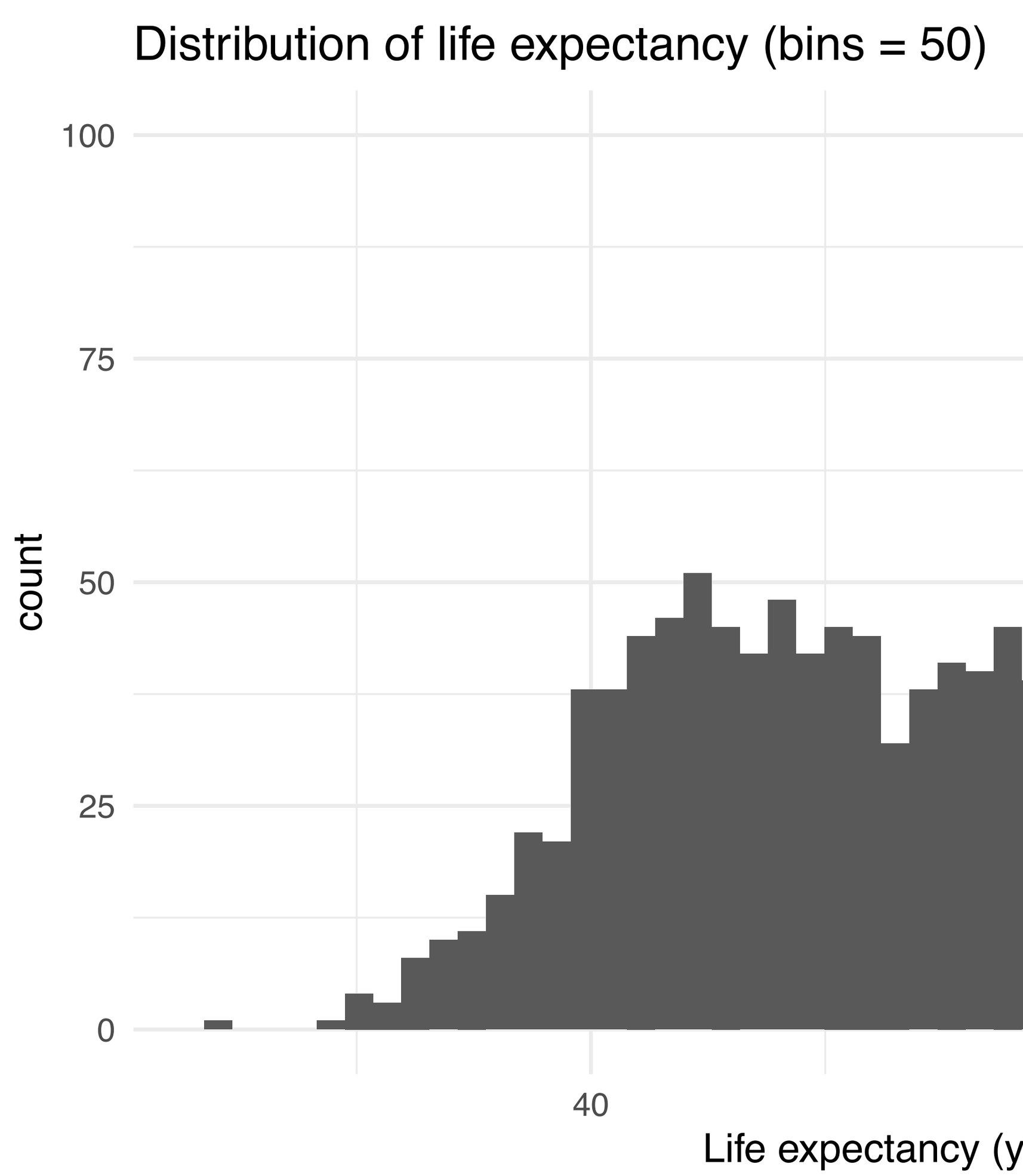
```
## #   tibble: 234 x 11
##    manufacturer model    displ  year   cyl tran
##    <chr>        <chr>    <dbl> <int> <int> <chr
##  1 audi         a4         1.8  1999     4 auto
##  2 audi         a4         1.8  1999     4 manu
##  3 audi         a4         2    2008     4 manu
##  4 audi         a4         2    2008     4 auto
##  5 audi         a4         2.8  1999     6 auto
##  6 audi         a4         2.8  1999     6 manu
```

http://fueleconomy.g

Looking at a single

- What is the distribution
  - Location - e.g., mean, median, m
  - Spread - e.g., variance
  - Shape - symmetric vs. skewed
- Are there outliers?
- What is notable about t

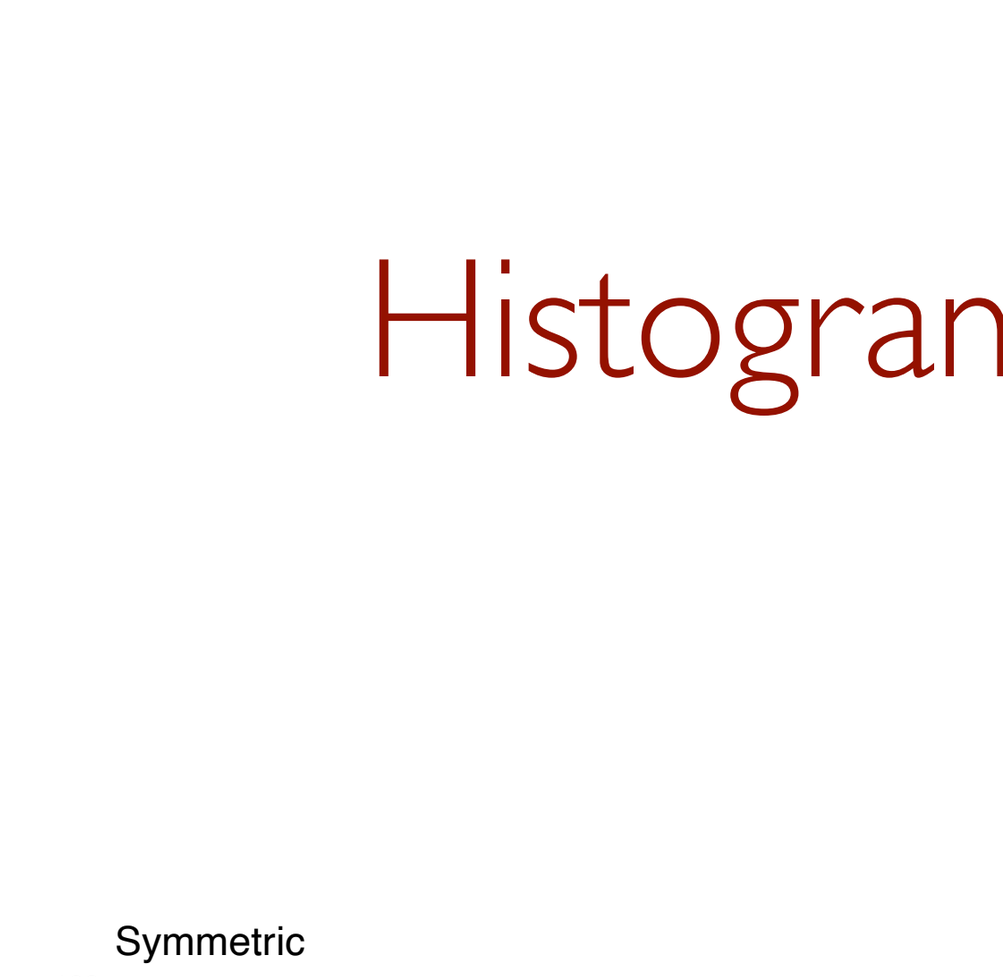
Histogram


Distribution of life expectancy (bins = 50)
100
75
50
25
0
count
40
Life expectancy (y

Histogram

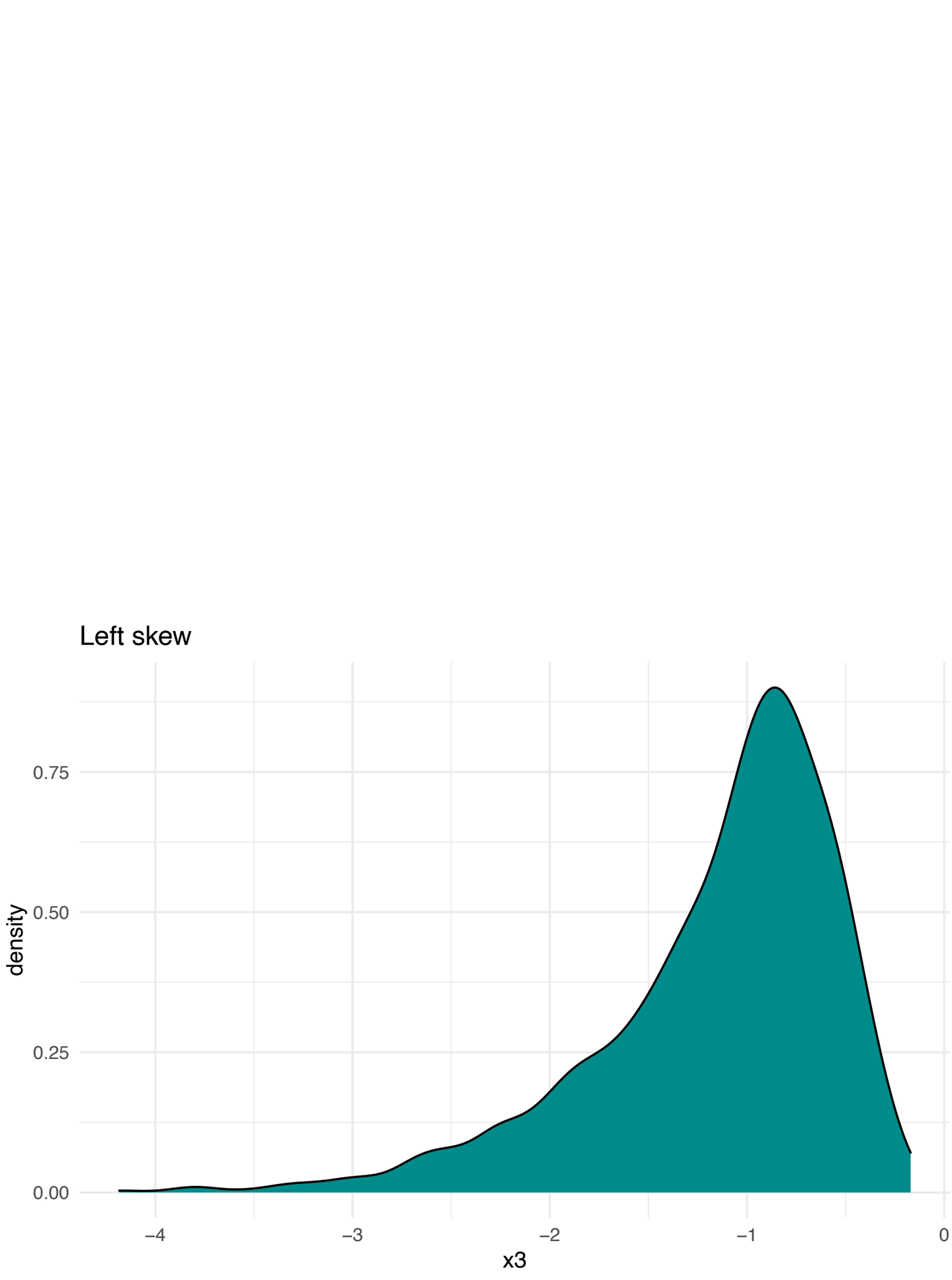
Distribution of life expectancy (binwidth = 1
count
75
50
25
0
40
Life expectancy (y

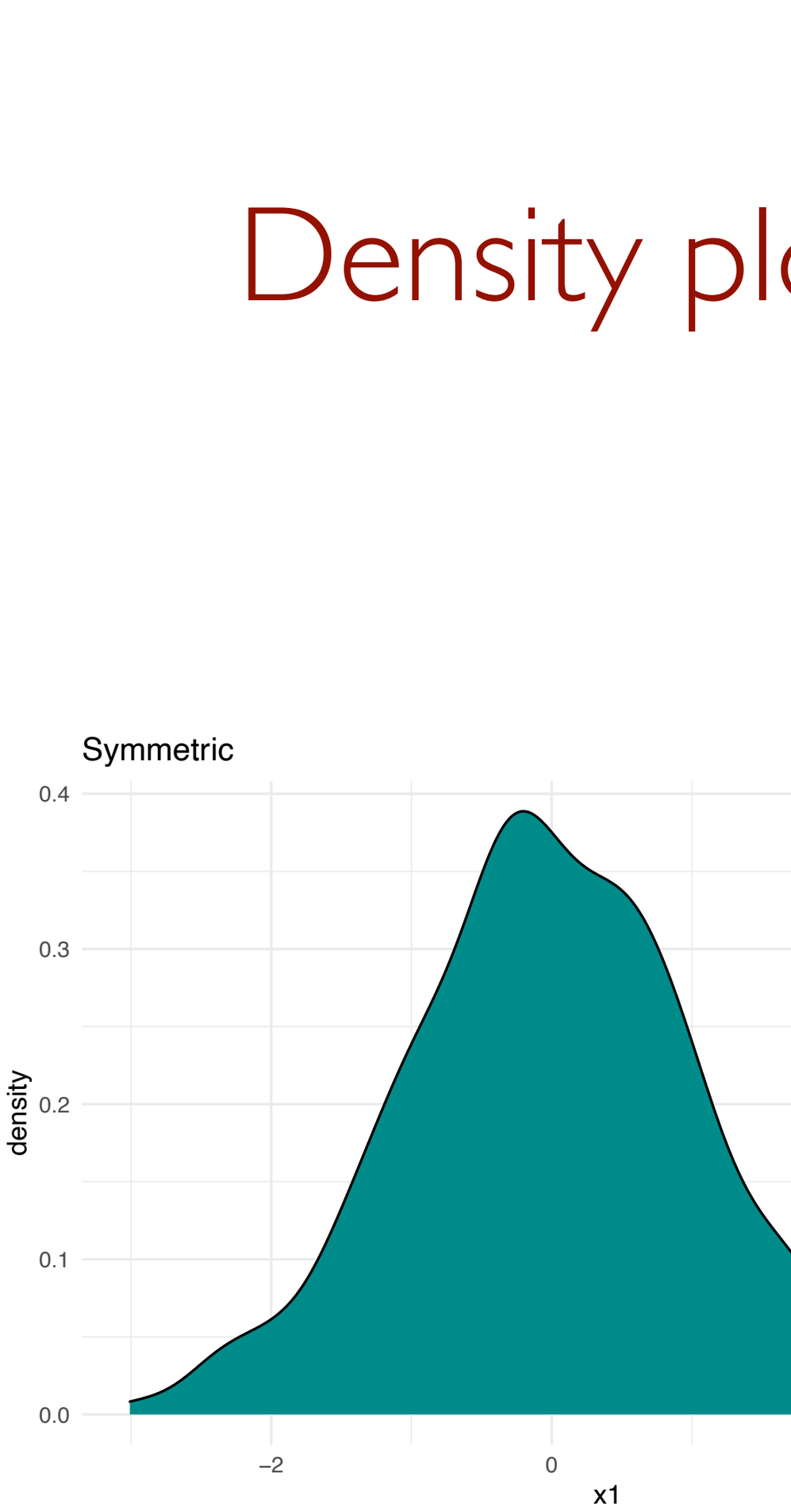
Histogram



simulated da
with different distr

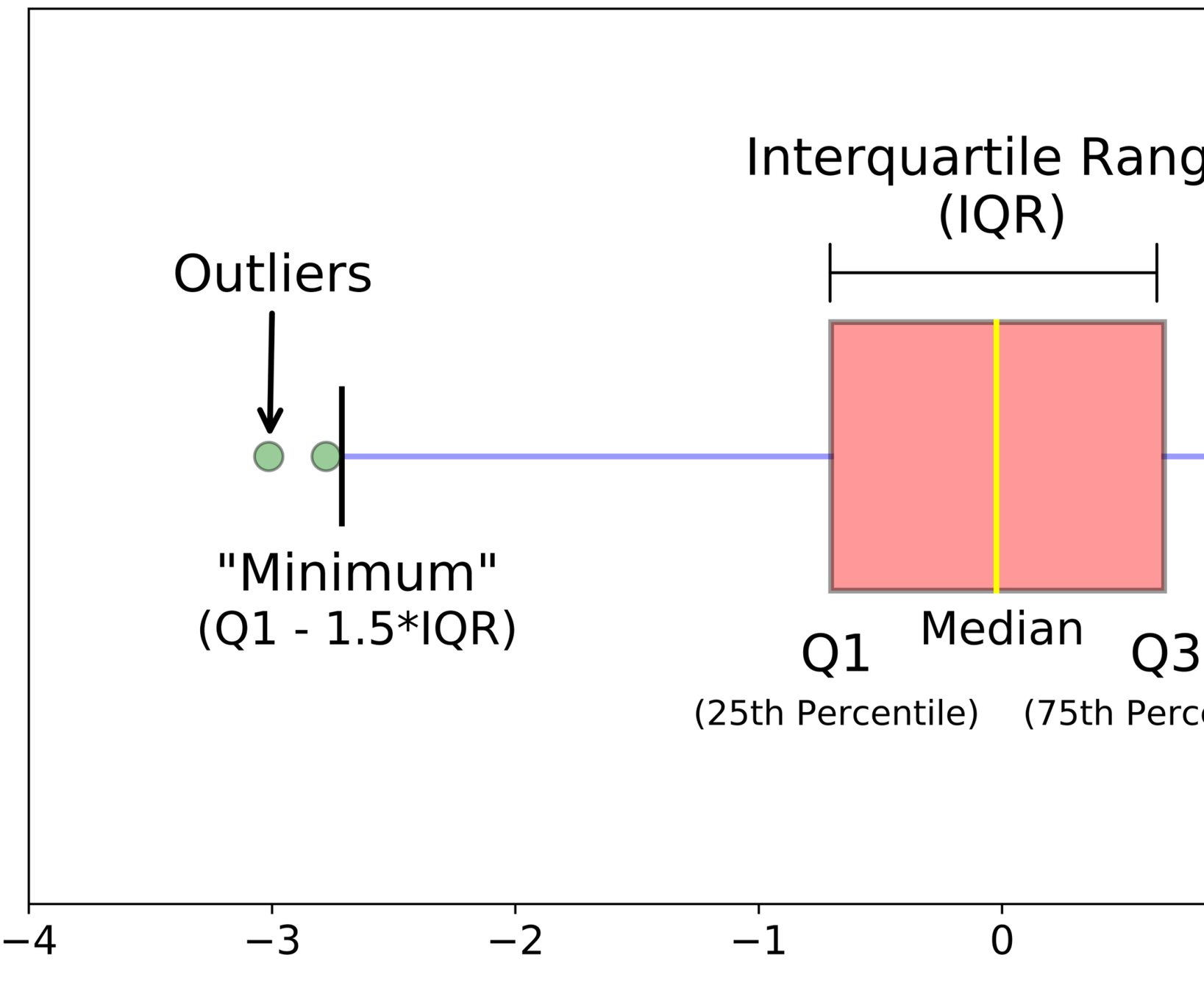
Density plo



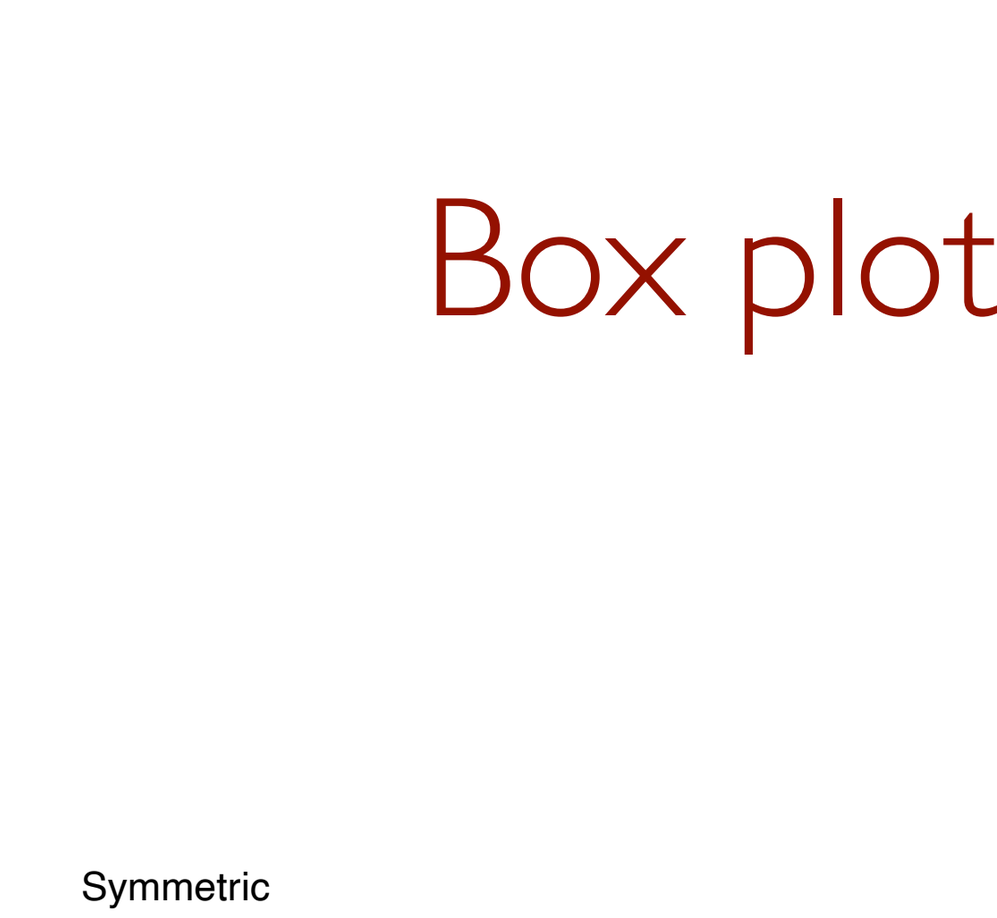


simulated da
with different distr

Box plot

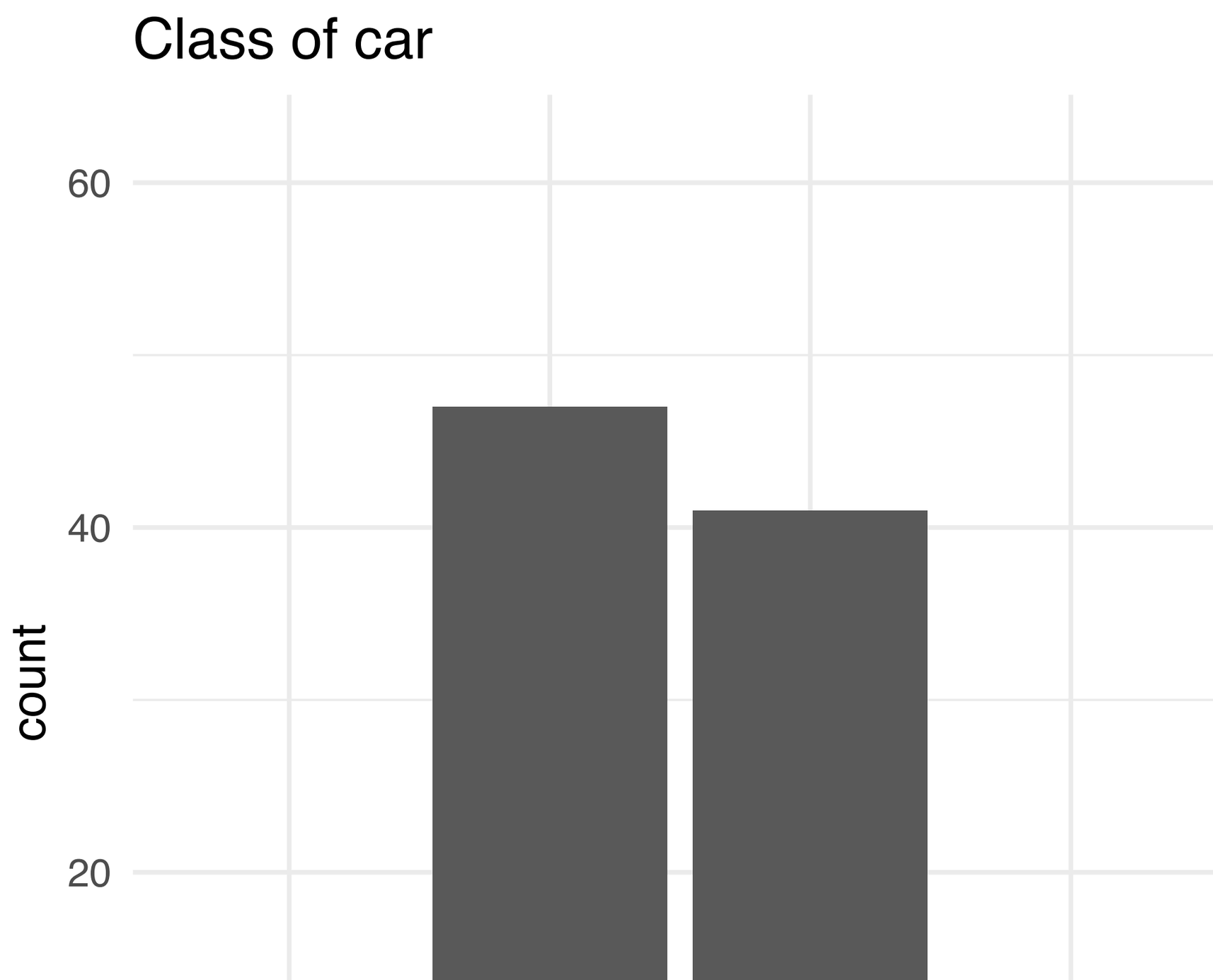

Interquartile Rang
(IQR)
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Median
Q1
(25th Percentile)
Q3
(75th Perc
−4
−3
−2
−1
0

Box plot

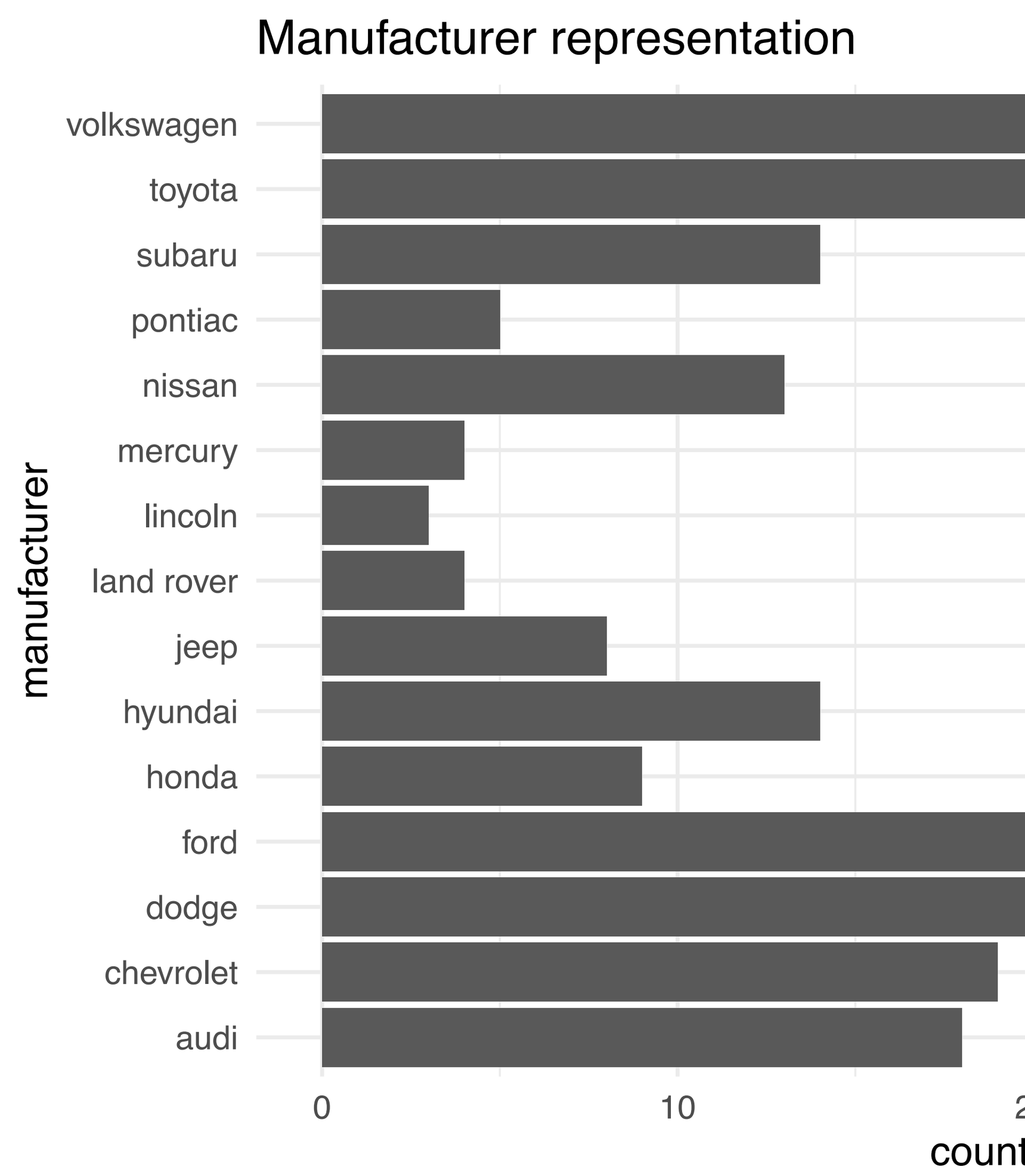


simulated da
with different distr

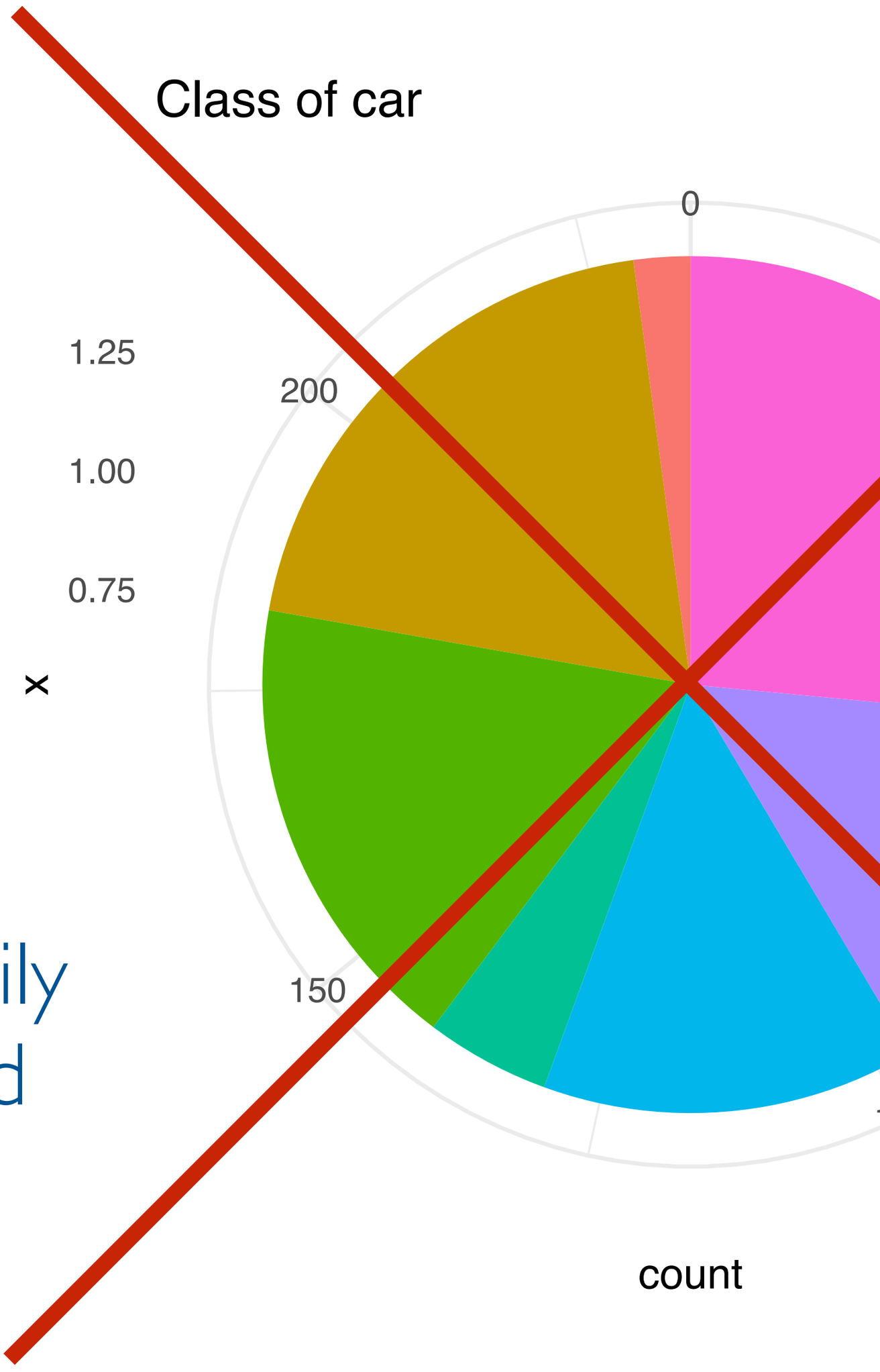
Bar plot


Class of car
60
40
20
0
count
2seater
compact
midsize
minivan
class

Bar plot


Manufacturer representation
volkswagen
toyota
subaru
pontiac
nissan
mercury
lincoln
land rover
jeep
hyundai
honda
ford
dodge
chevrolet
audi
manufacturer
0
10
count

Pie chart

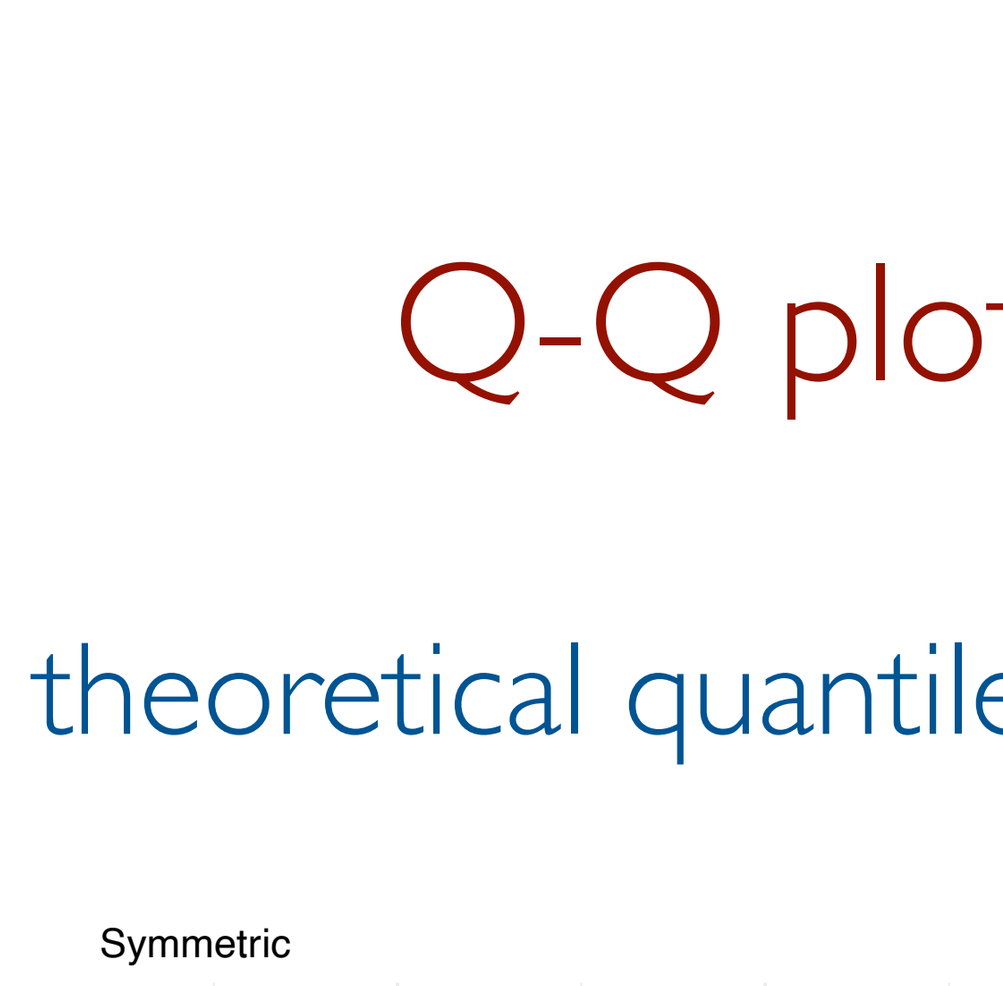


difficult to read easily

not recommended

Q-Q plot

sample quantiles versus theoretical quantile



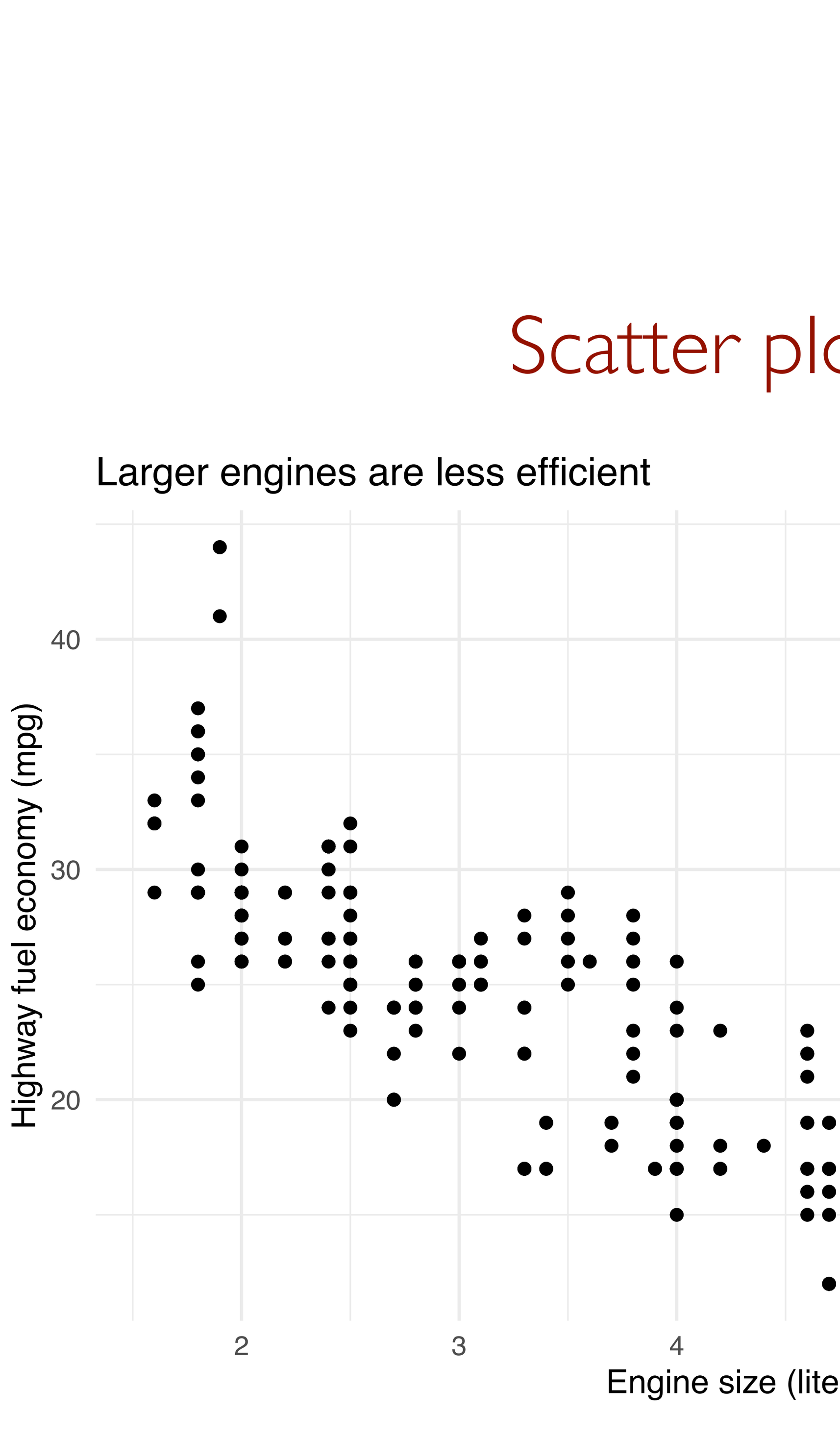
simulated da
with different distr

MULTI-VARIABLE

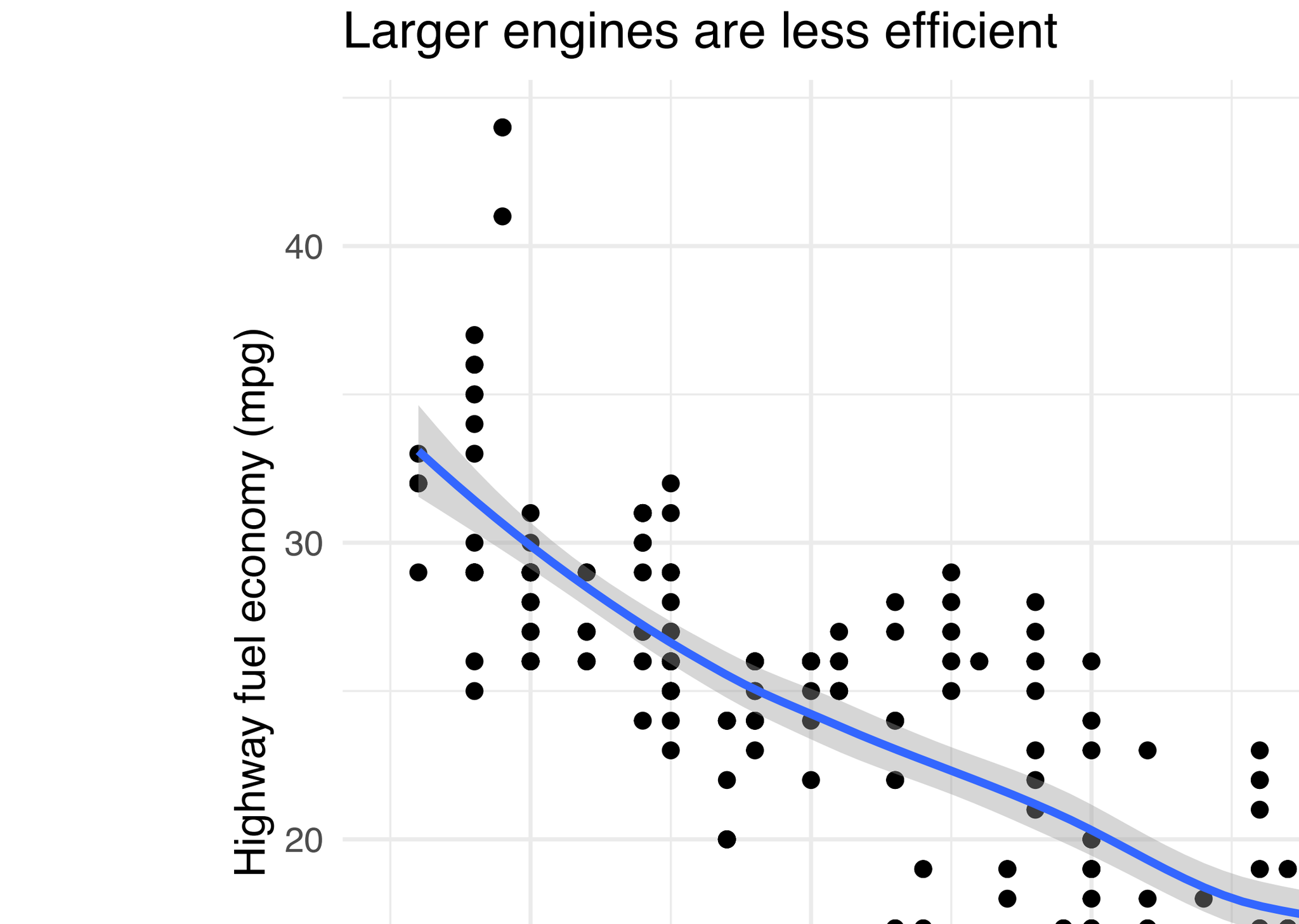
Looking at multiple

- How are the variables r
  - Is there a relationship?
  - Type of relationship (e.g., linear
  - Direction (positive vs. negative
- Are there outliers?
- Does the relationship **c**

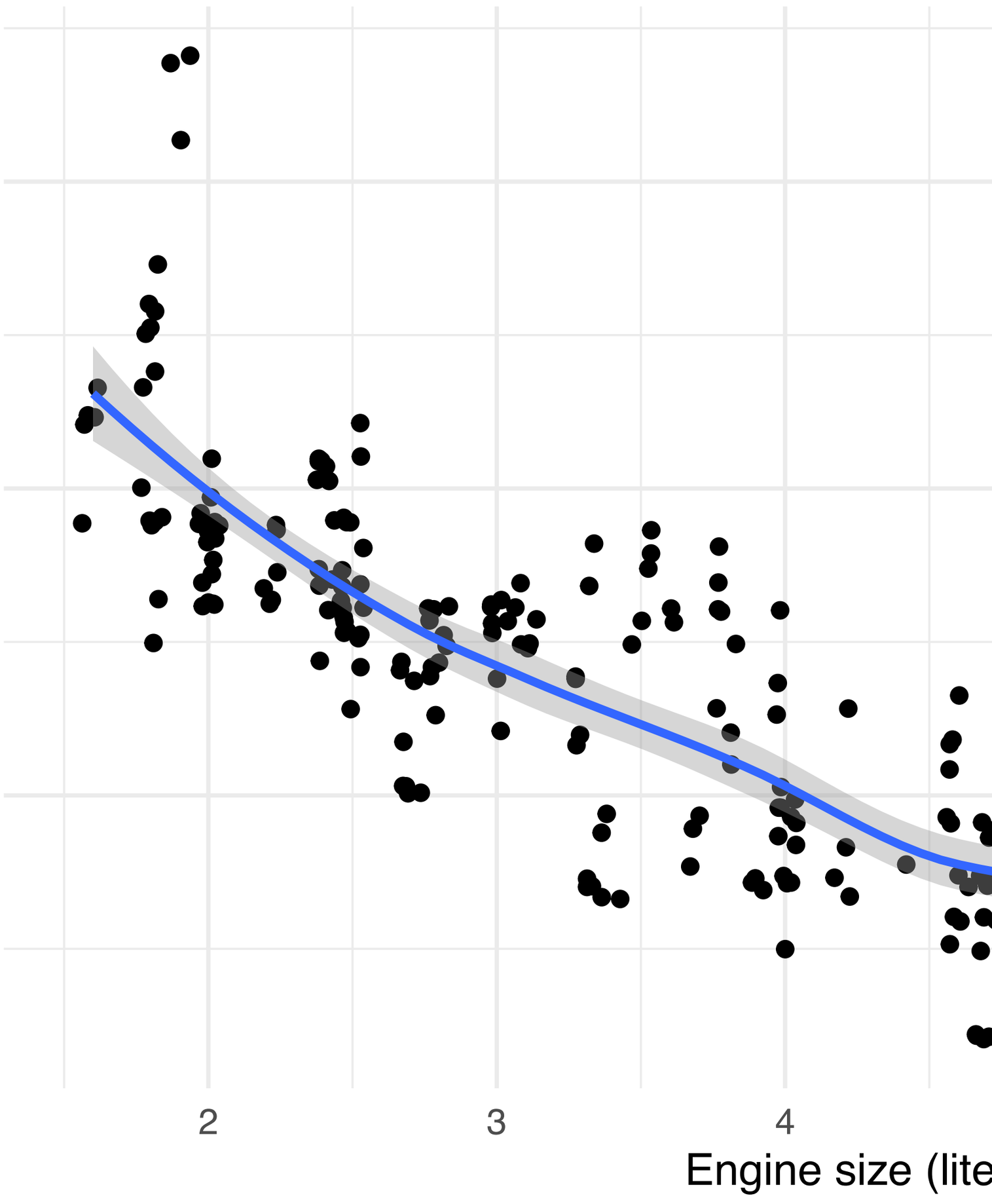
Scatter plo


Larger engines are less efficient
40
30
20
Highway fuel economy (mpg)
2
3
4
Engine size (lite

Scatter plo

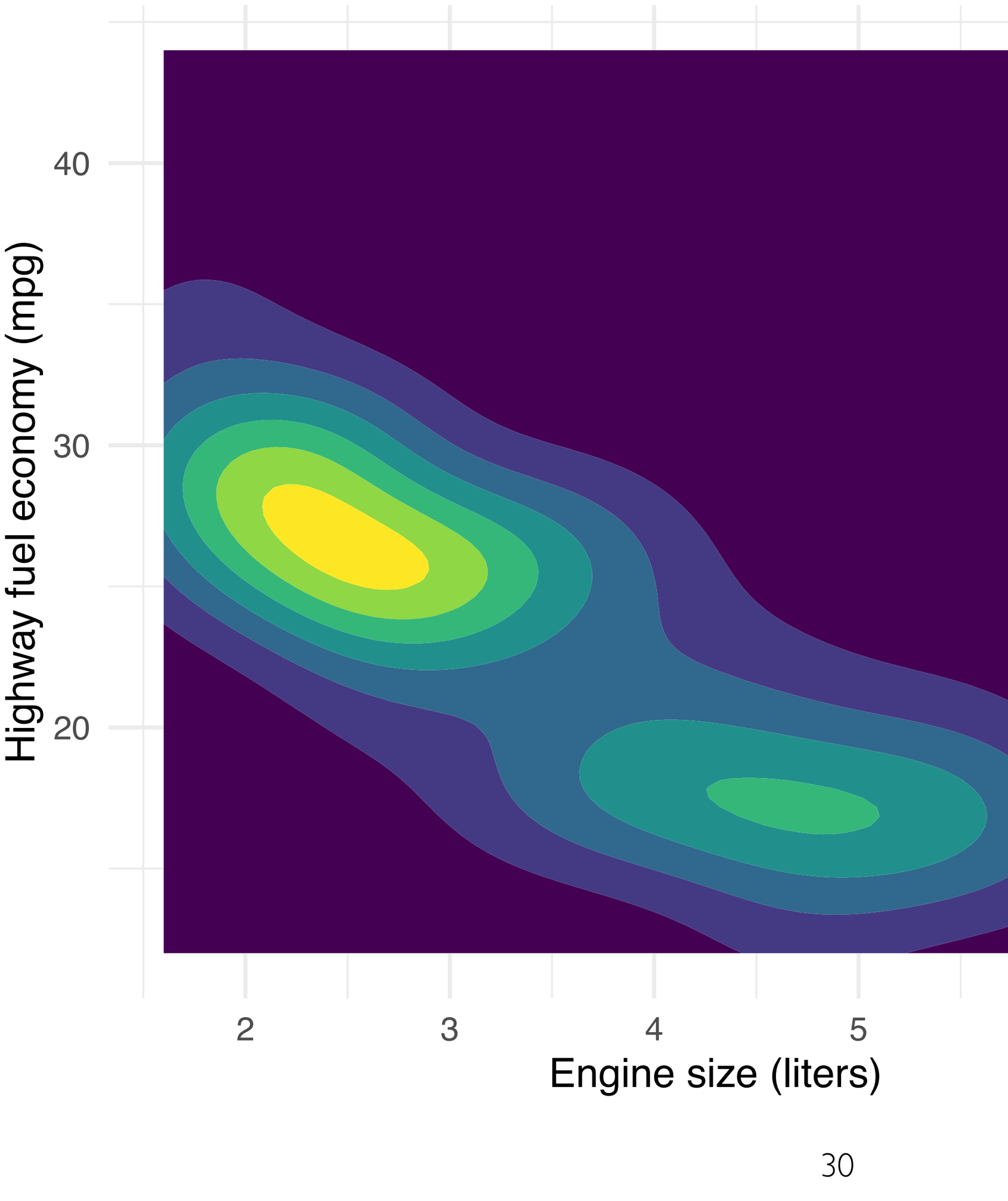

Larger engines are less efficient
Highway fuel economy (mpg)
40
30
20
2
3
4
Engine size (lite

Scatter plo

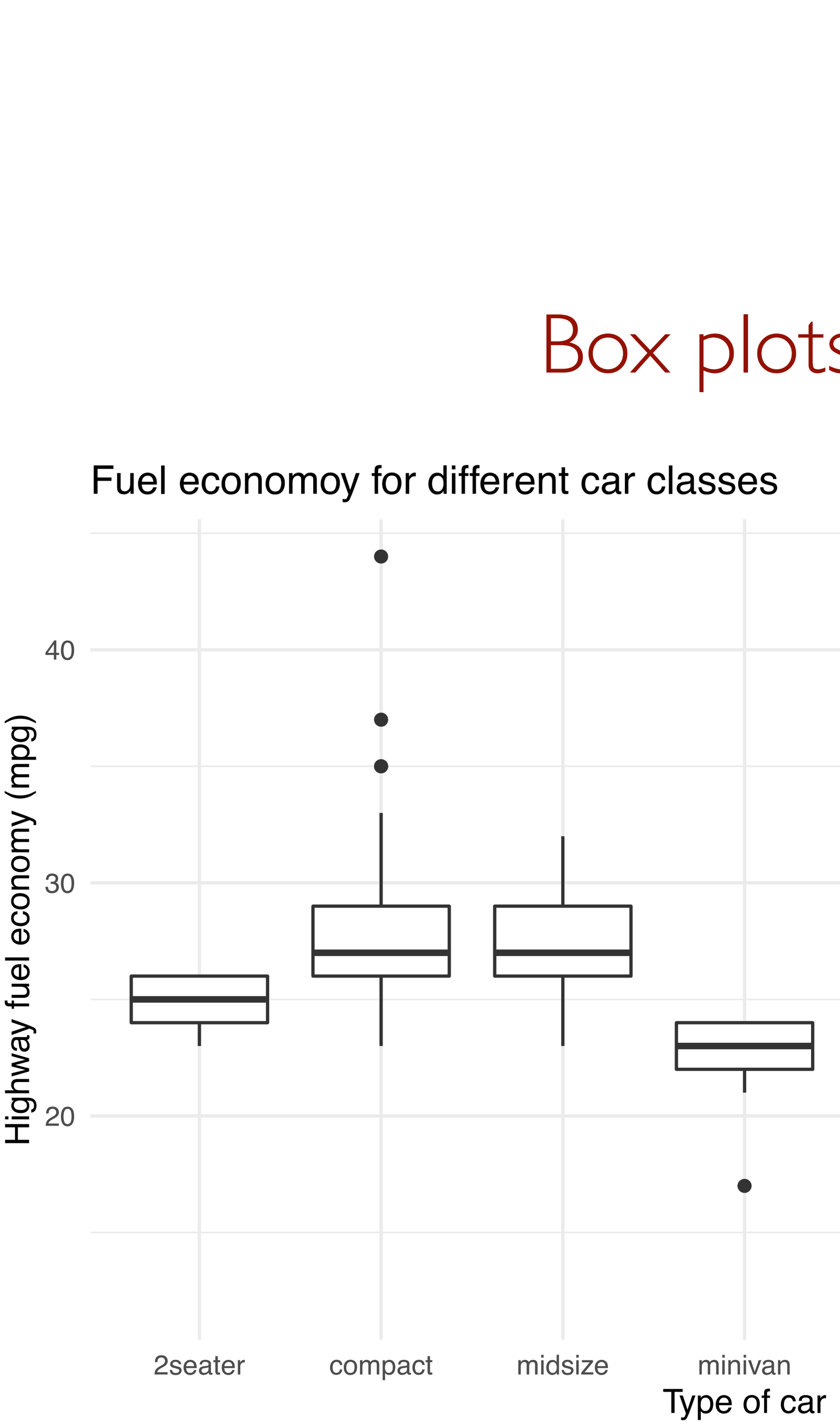

Larger engines are less efficient
40
30
20
Highway fuel economy (mpg)
2
3
4
Engine size (lite

2D density p

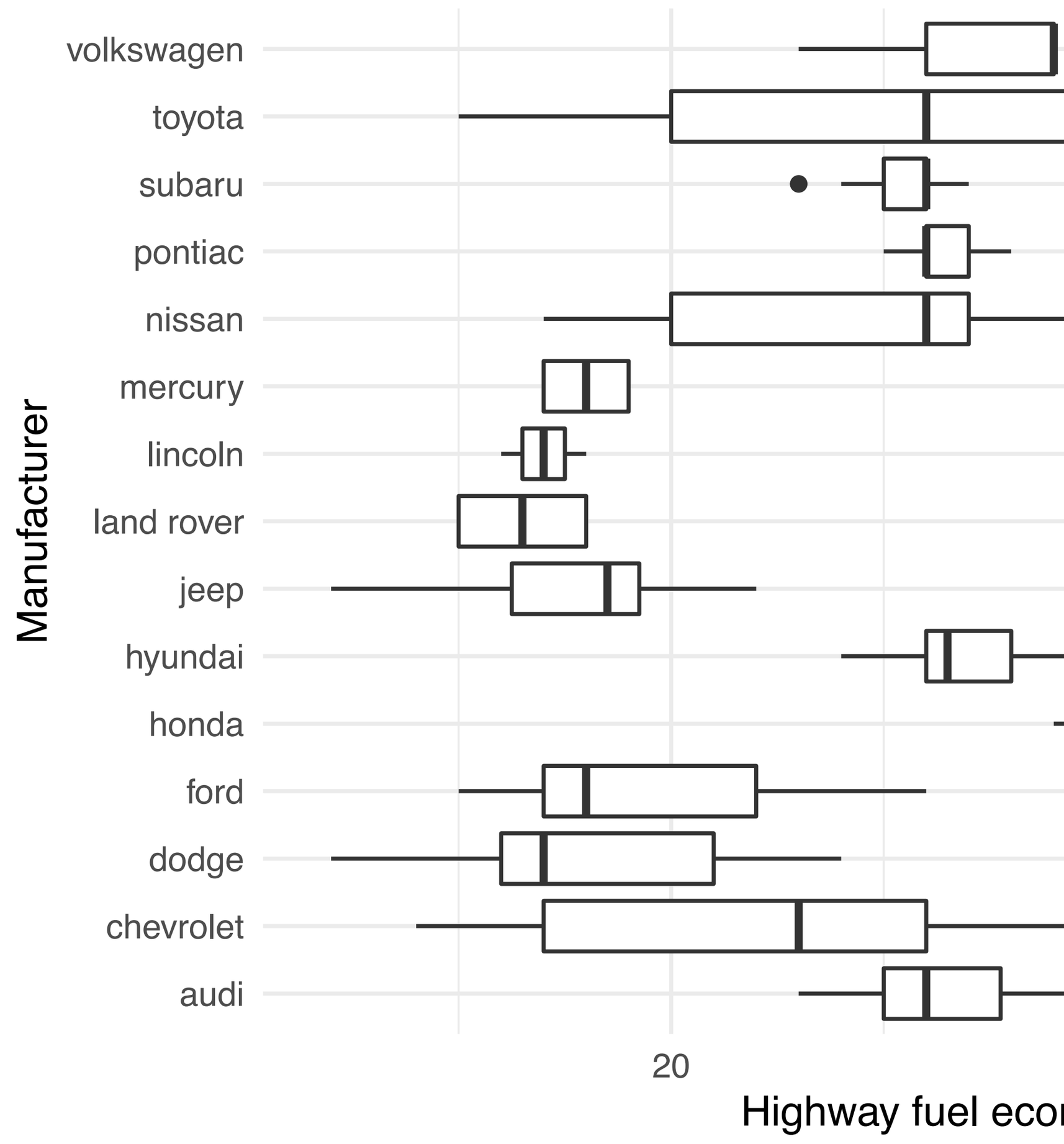
Larger engines are less efficient



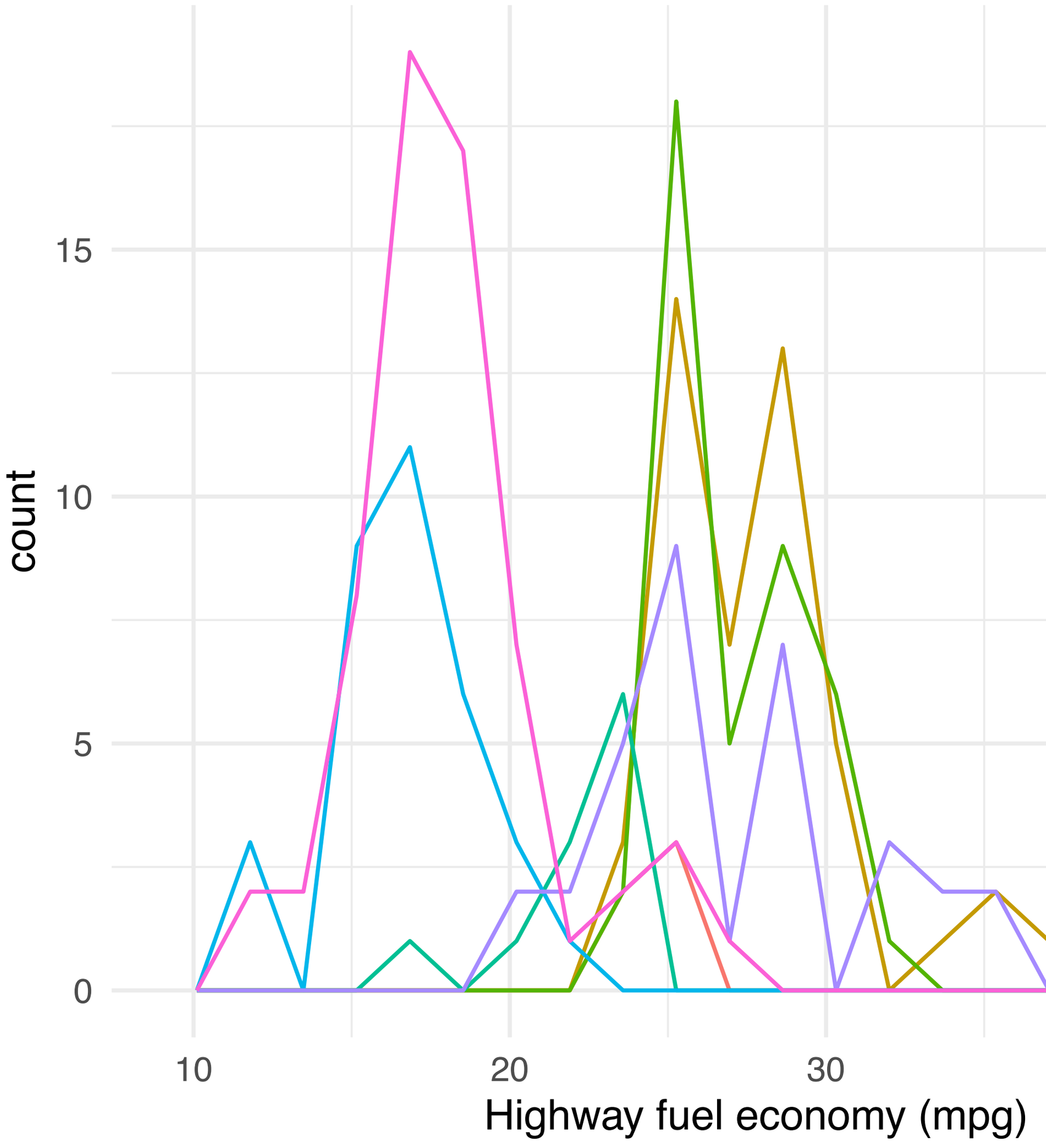
Box plots


Fuel economoy for different car classes
Highway fuel economy (mpg)
40
30
20
2seater
compact
midsize
minivan
Type of car

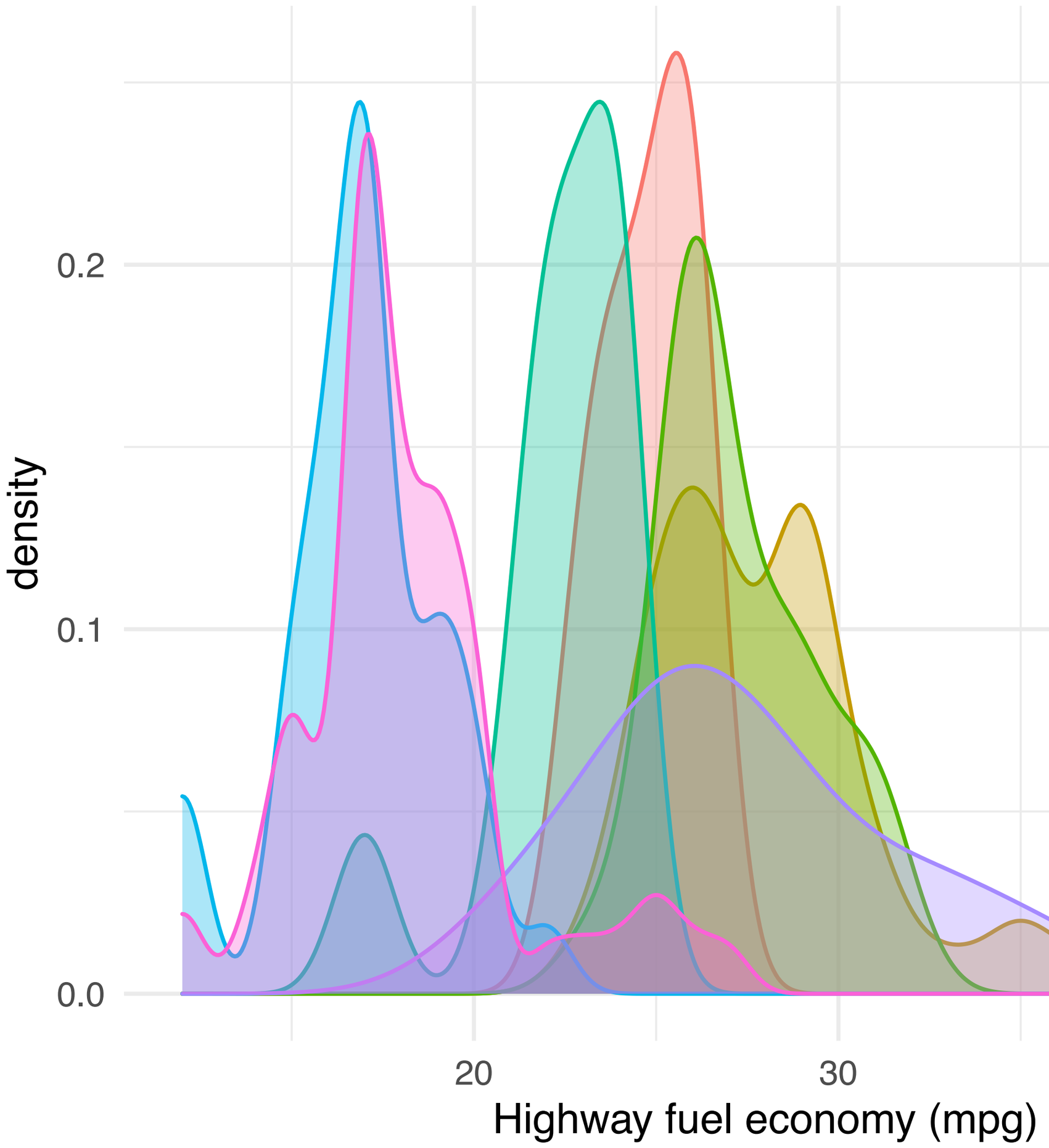
Box plots


Fuel economoy for different manufac
volkswagen
toyota
subaru
pontiac
nissan
mercury
lincoln
land rover
jeep
hyundai
honda
ford
dodge
chevrolet
audi
Manufacturer
20
Highway fuel econ

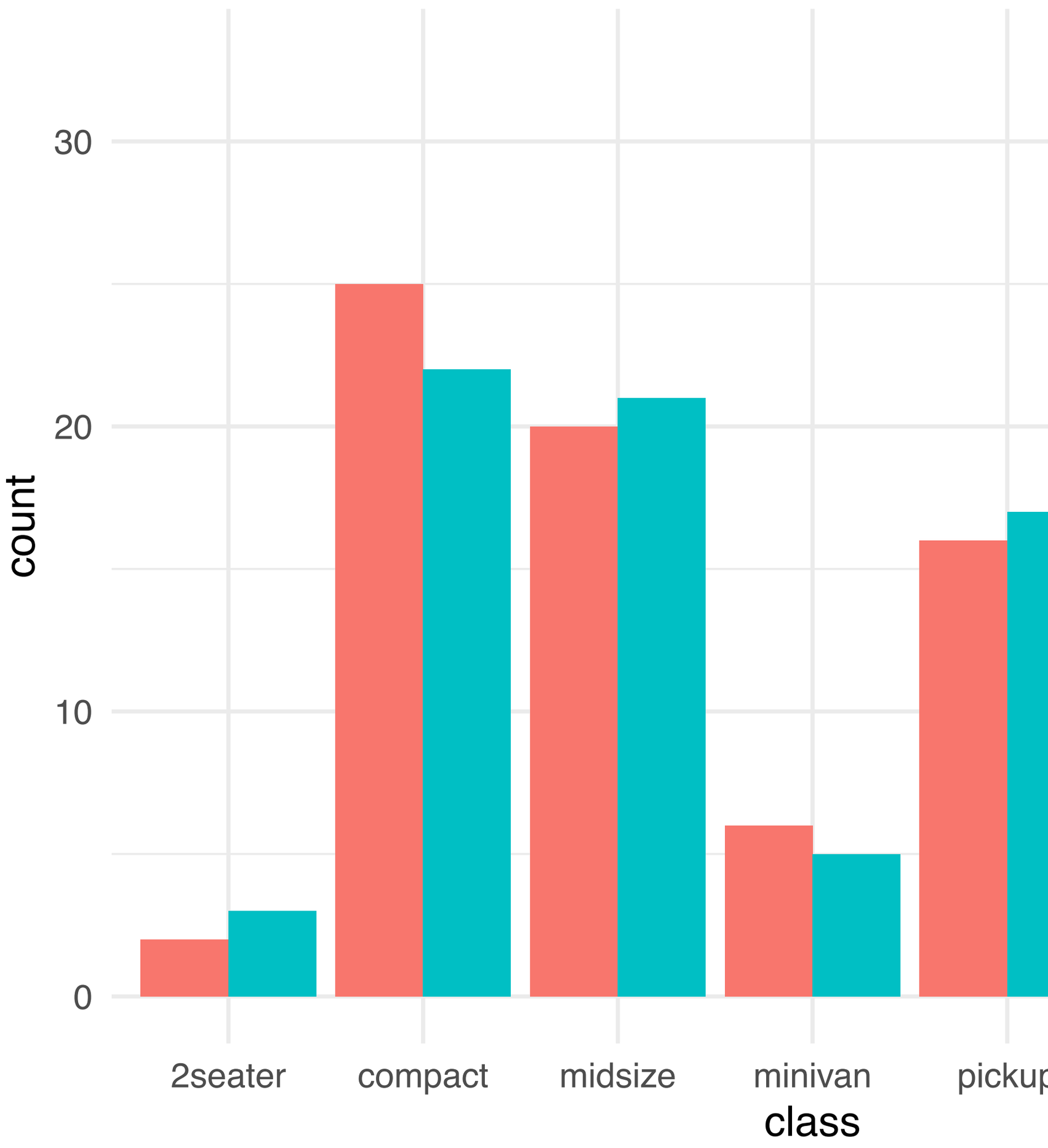
Histogram


Fuel economoy for different car classes
count
15
10
5
0
10
20
30
Highway fuel economy (mpg)

Density plo


Fuel economoy for different car classes
density
0.2
0.1
0.0
20
30
Highway fuel economy (mpg)

Bar plots

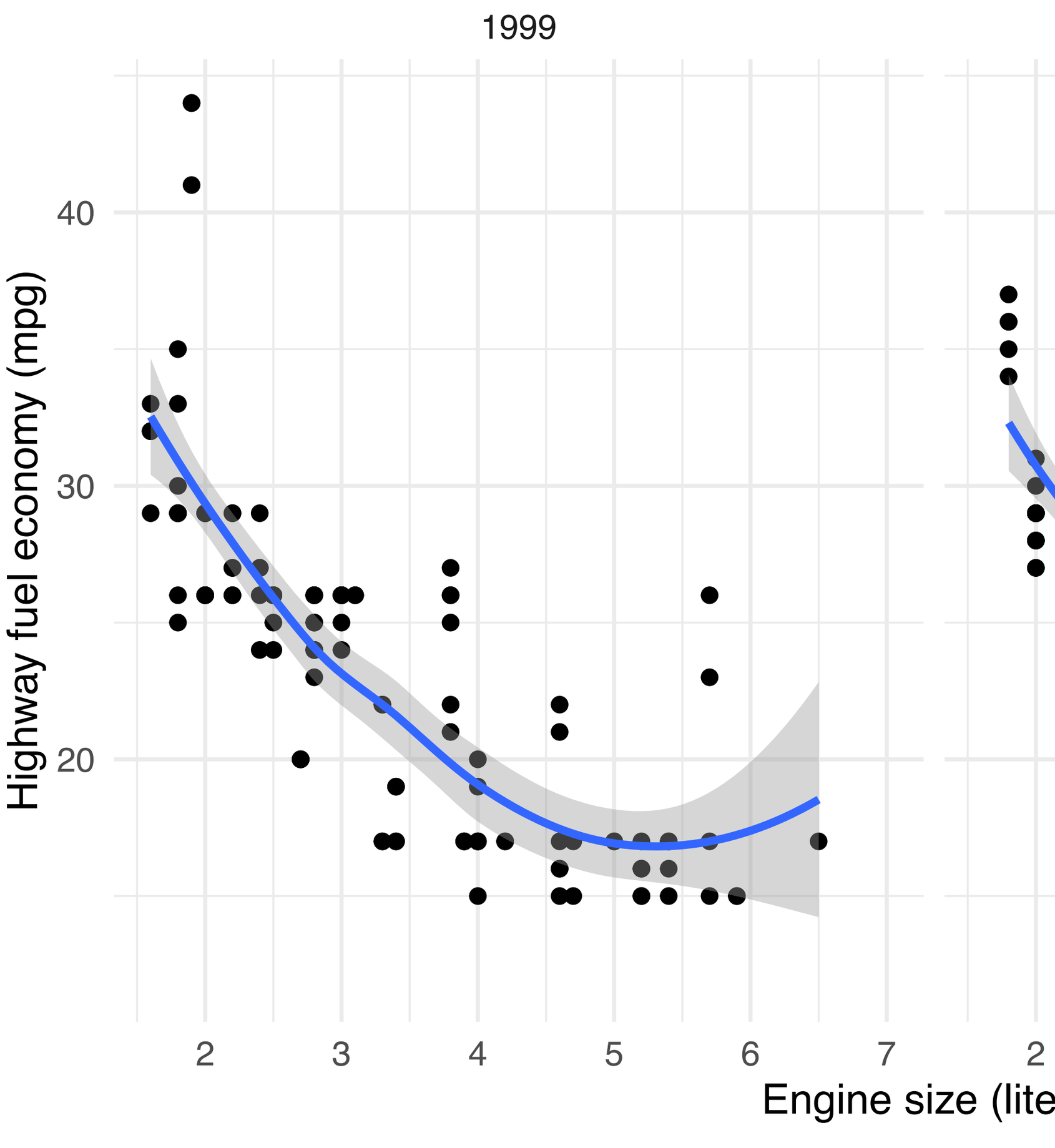

Breakdown of car types by year
count
30
20
10
0
2seater
compact
midsize
minivan
pickup
class

FACETIN

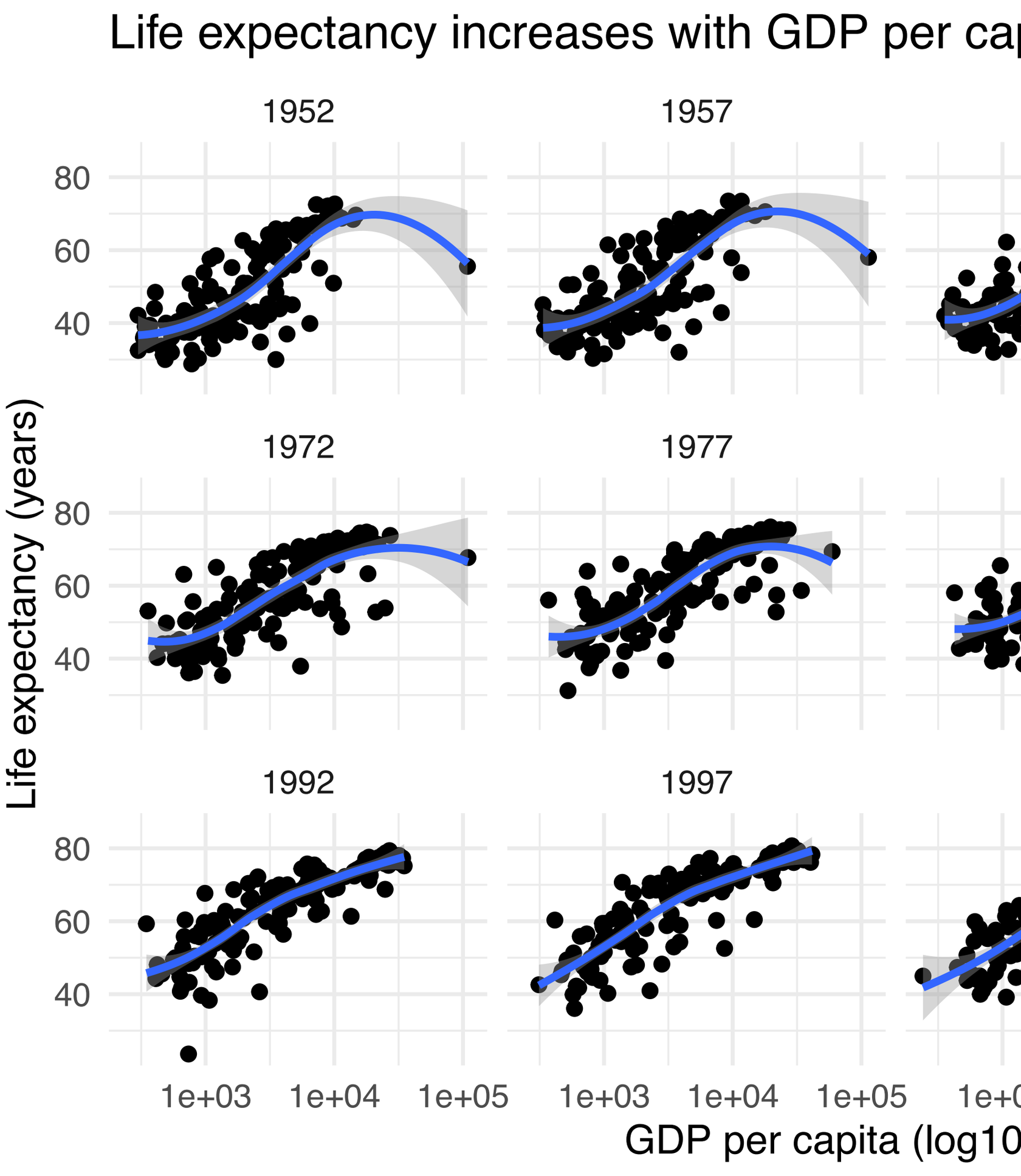
What is facet

- Based on idea of "small
- Condition on levels of s
- Split data into subsets b
- Create sub-plots for ea
  - Sub-plots share same scales an
  - Easily compare between sub-p

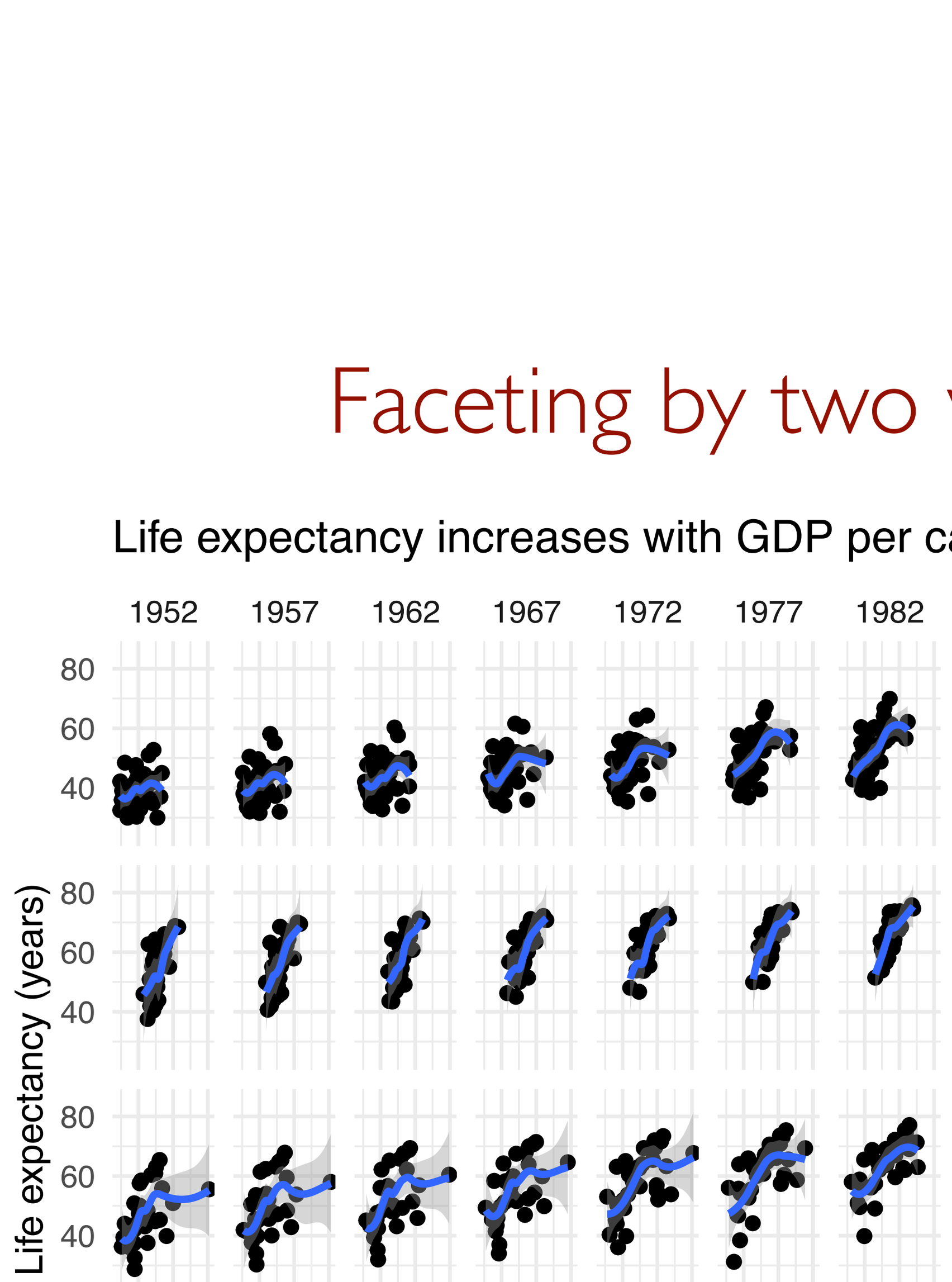
Faceting by one



Faceting by one



Faceting by two v



A GRAMMAR OF

How do we plo

- By using the "name" of
  - Scatter plot
  - Box plot
  - Histogram
- Using "base" R (and sim
  - `plot()` - scatter plot
  - `boxplot()` - box plot
  - `hist()` - histogram

What are some common ingredien

Recipes for common sta

- Scatter plot
  - Maps variables to x- and y- axes
  - Uses points to represent observations
- Line plot
  - Maps variables to x- and y- axes
  - Uses lines to connect observations
- Box plot
  - Maps 5-number summary to x- or y-axis
  - Uses boxes and whispers to show this
- Hist
  - Ma
  - Use
- Bar
  - Ma
  - Use
- Pie
  - Ma
  - Use

Key ingredients for stat

- Some kind of data
- Encodings from data to
  - Marks ("**geometric objects**", e.
  - Channels ("**aesthetics**", e.g., co
- Statistical transformatio
- Coordinate system
- Scales and annotations

Building a p

Consider a simple dataset:

A	B	C
2	3	4
1	2	1
4	5	15
9	10	80

http://vita.had.co.nz/papers/layere

How do we create a scatter plot of A versus C

Wh

Building a p

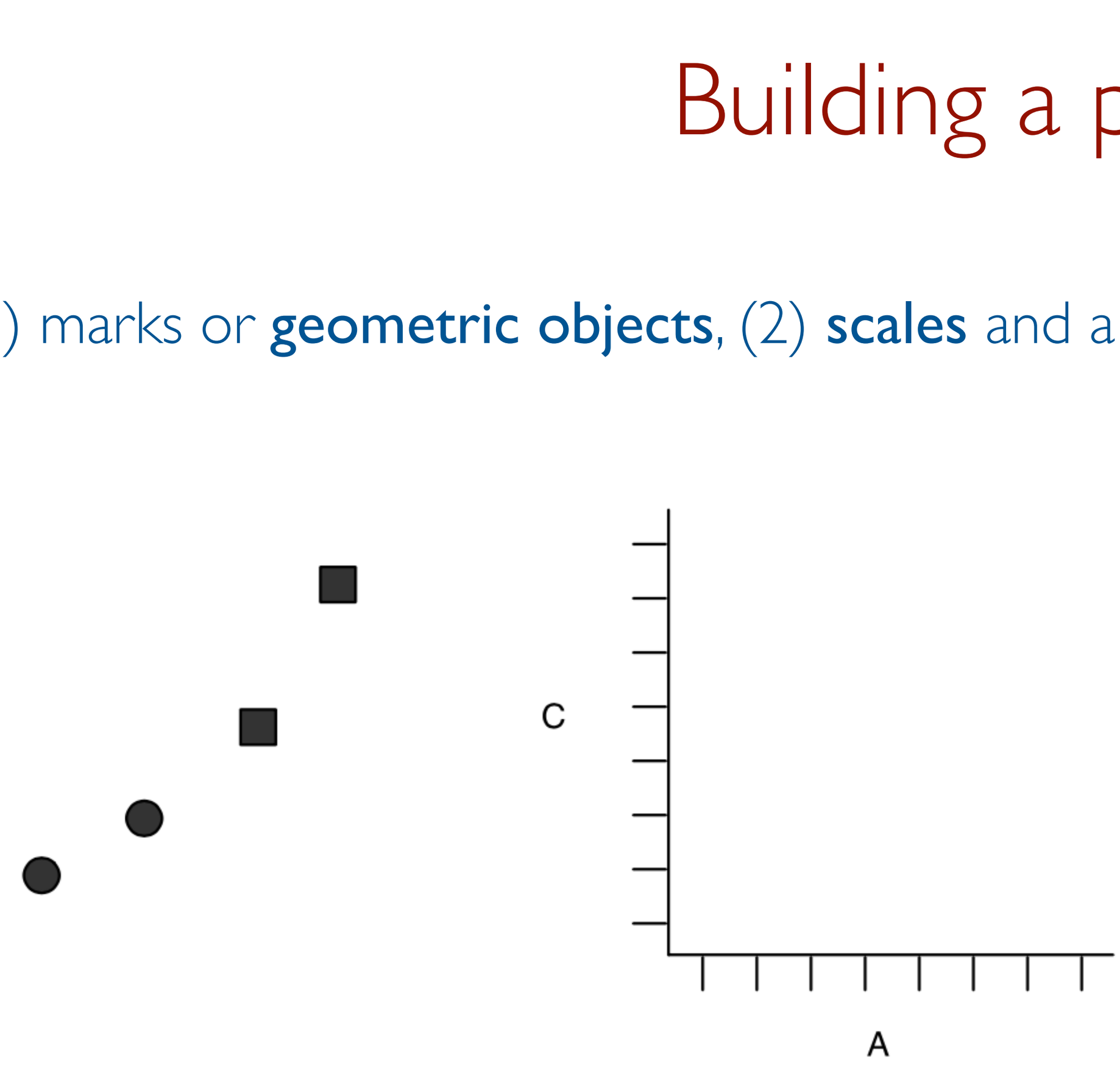
We map the x-axis to **A**, the y-axis to **C**, and shape to **D**

x	y	S
2	4	ci
1	1	ci
4	15	sq
9	80	sq

http://vita.had.co.nz/papers/layere

Building a p

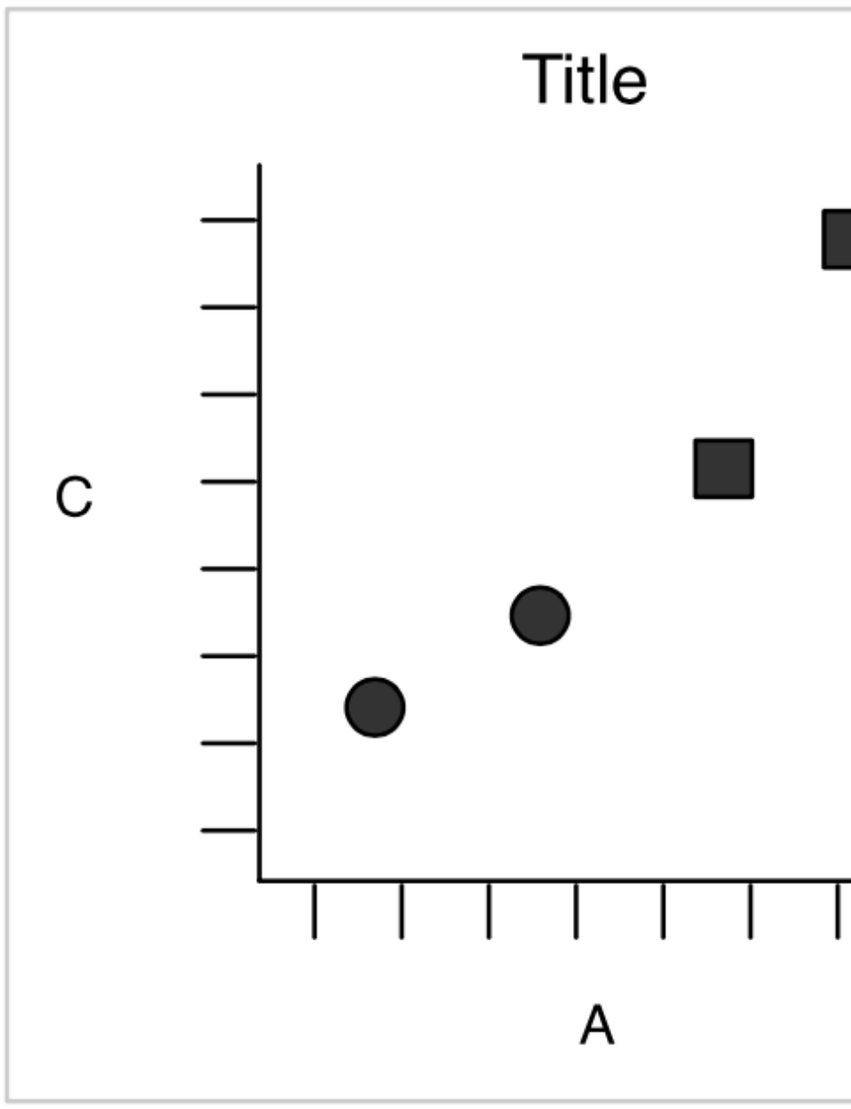
We have (1) marks or **geometric objects**, (2) **scales** and a



http://vita.had.co.nz/papers/layere

Building a p

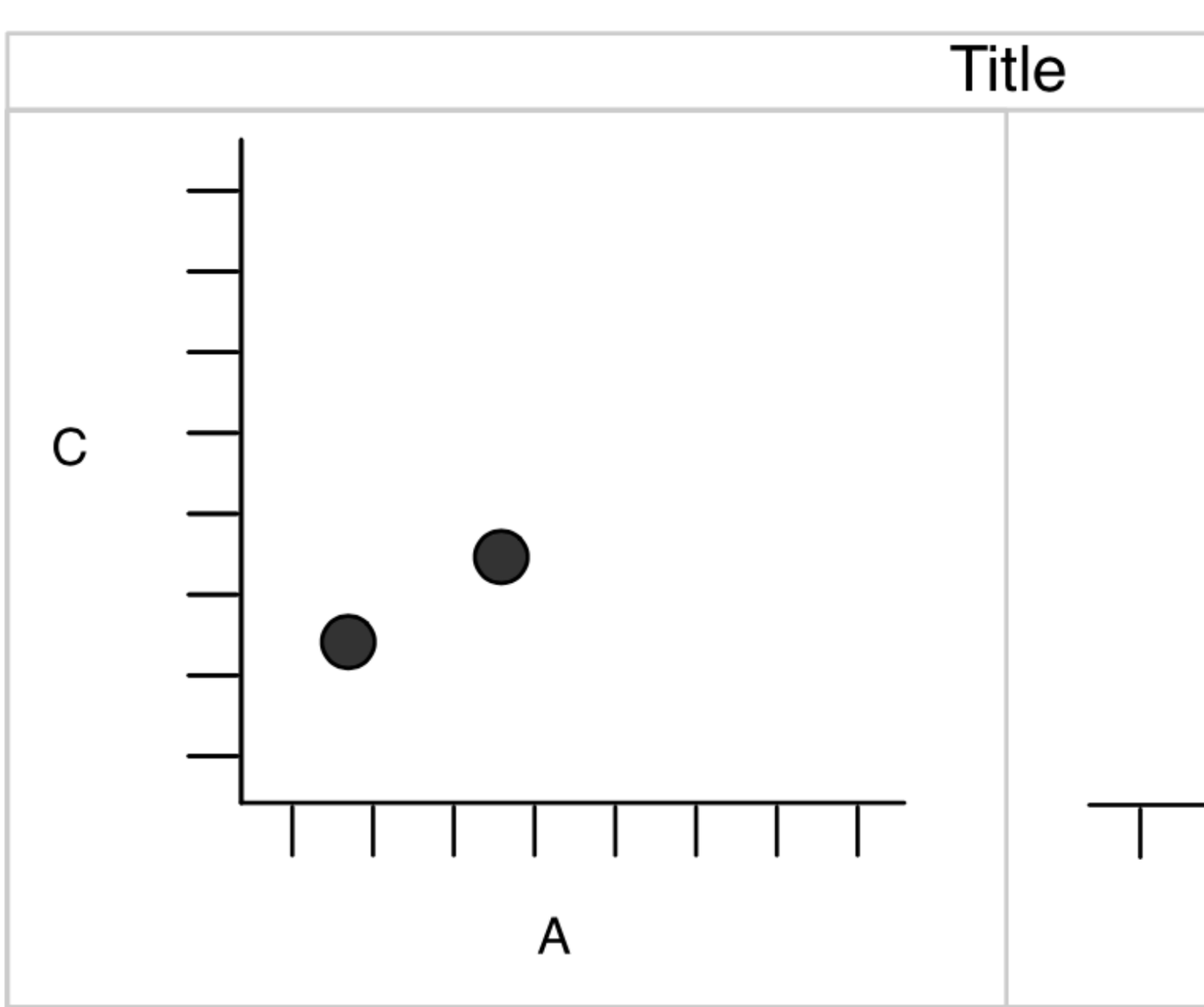
Putting the ingredients together, we have a plot:



http://vita.had.co.nz/papers/layere

Building a p

If we want to compare the relationship between **A** and



http://vita.had.co.nz/papers/layere

Faceting splits the data into subsets and crea

Building a voca

We can build more complicated plots by adding to

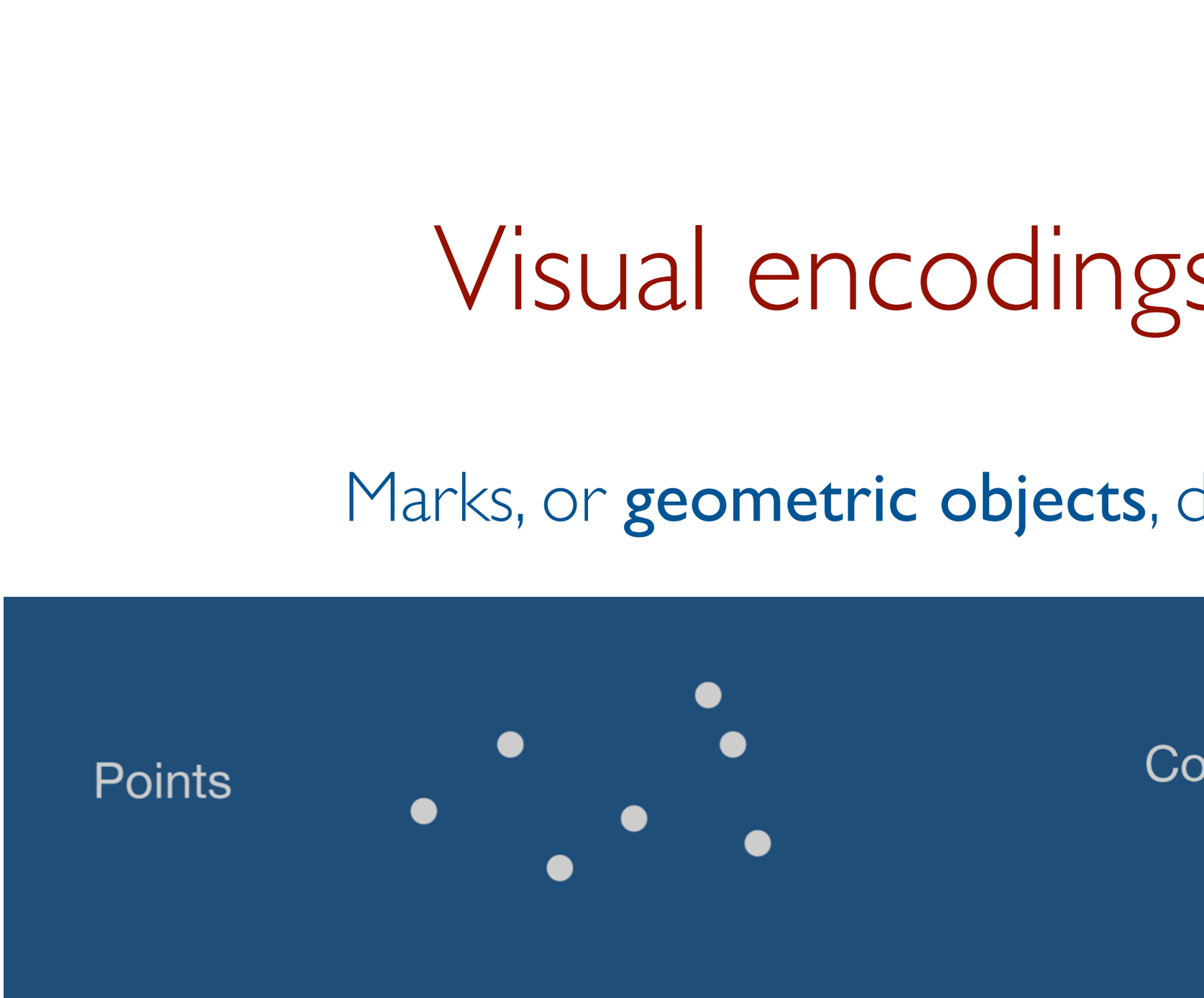
- Layers to overlay plots on to
- Multiple datasets on the sam
- Apply statistical transformatio
- Apply position adjustments (
- *A way to build such plots prog*

A layered grammar

- Default dataset
- Default set of mappings from varial
- One or more layers, each having:
  - Mark, or geometric object
  - Statistical transformation
  - Position adjustment
  - (Optionally) new dataset
  - (Optionally) new set of aesthetic m
- Scale for each mapped aesthetic
- Facet specification

Visual encodings

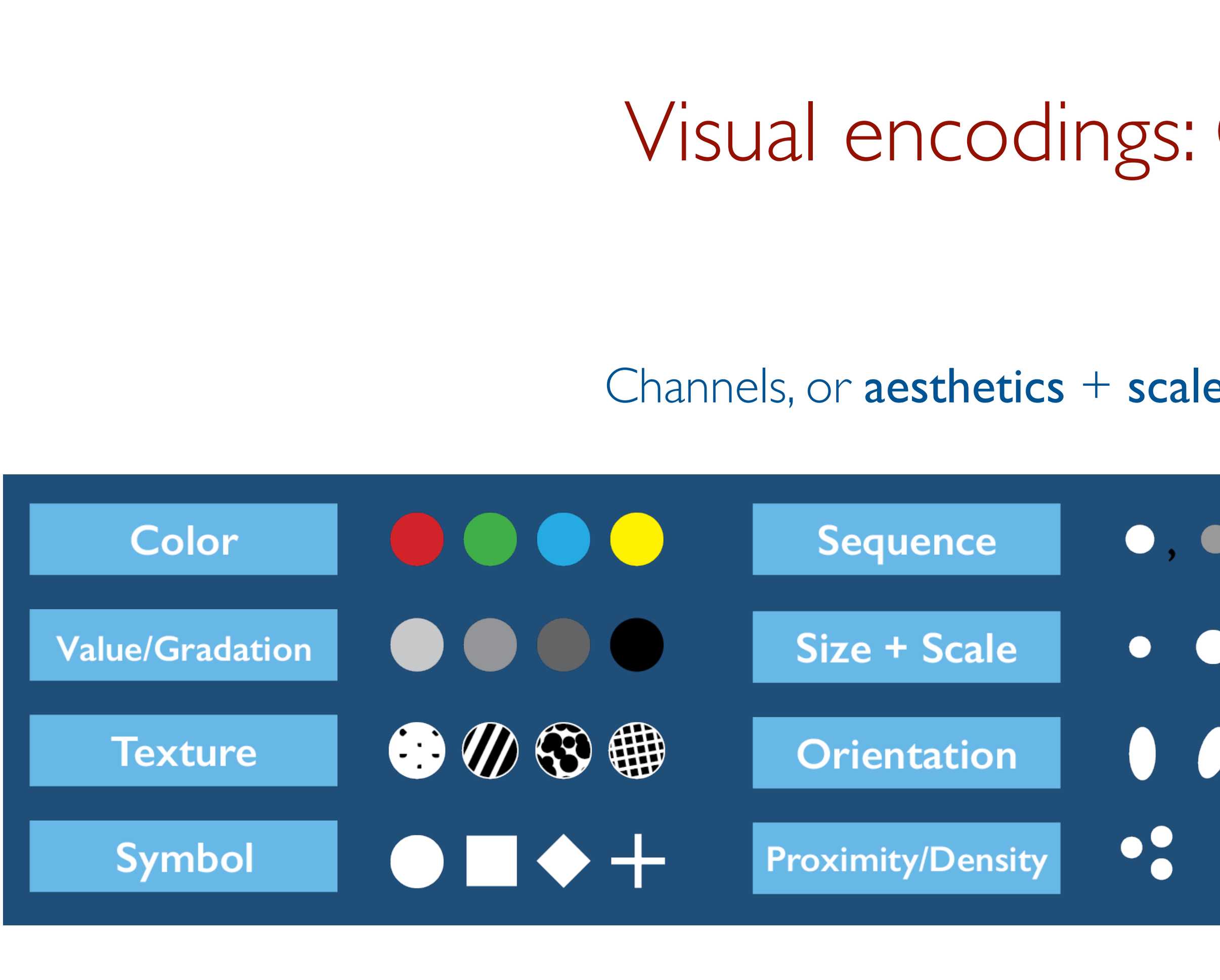
Marks, or **geometric objects**, d



Courtesy of Steven Braun, CAMD

Visual encodings: C

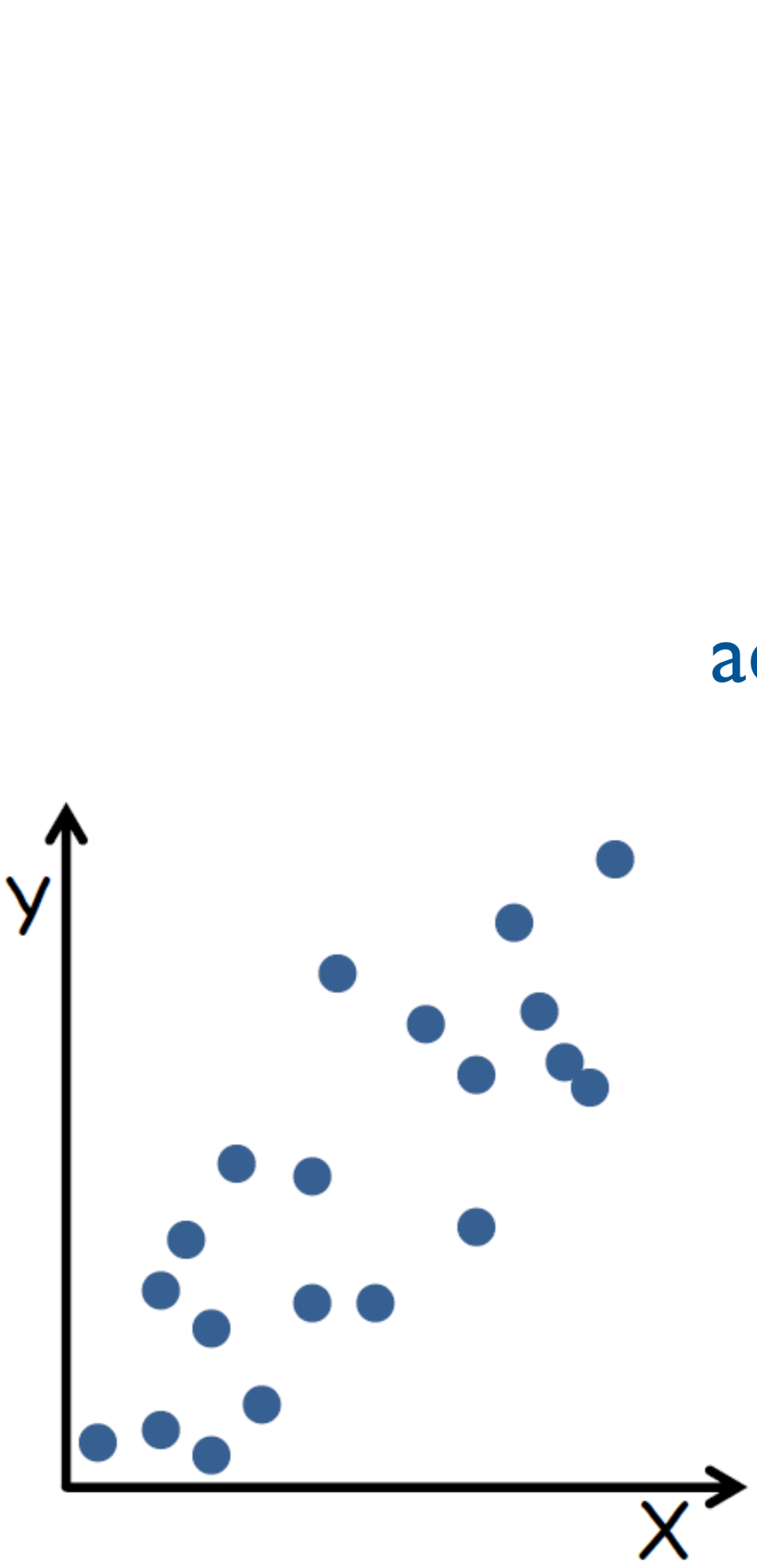
Channels, or **aesthetics** + **scale**



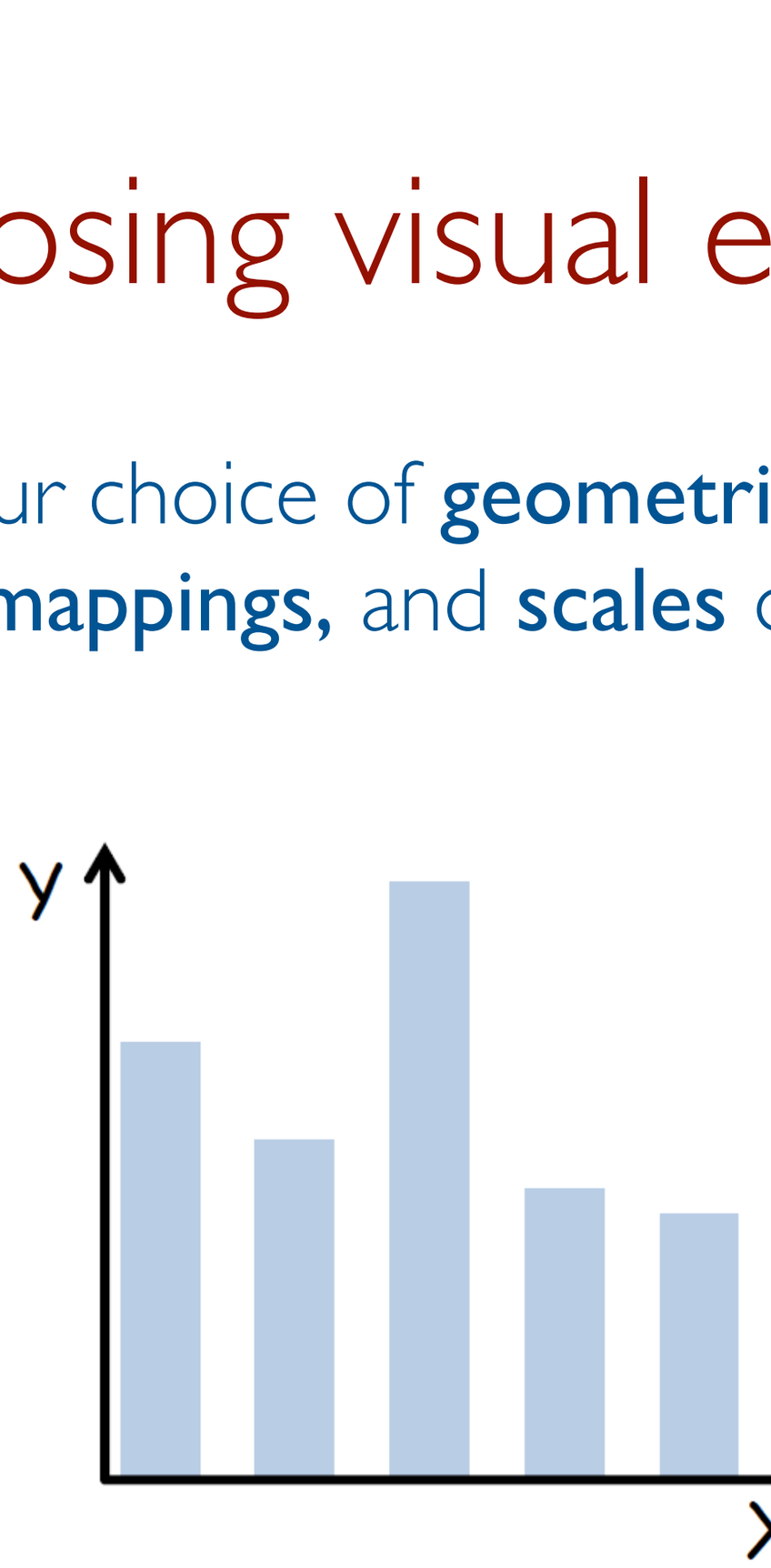
Courtesy of Steven Braun, CAMD

Choosing visual e

Your choice of **geometri**
aesthetic mappings, and **scales** c



Marks: points
Channels: position



Marks: lines
Channels: length, po

Courtesy of Steven Braun, CAMD

Statistical transfor

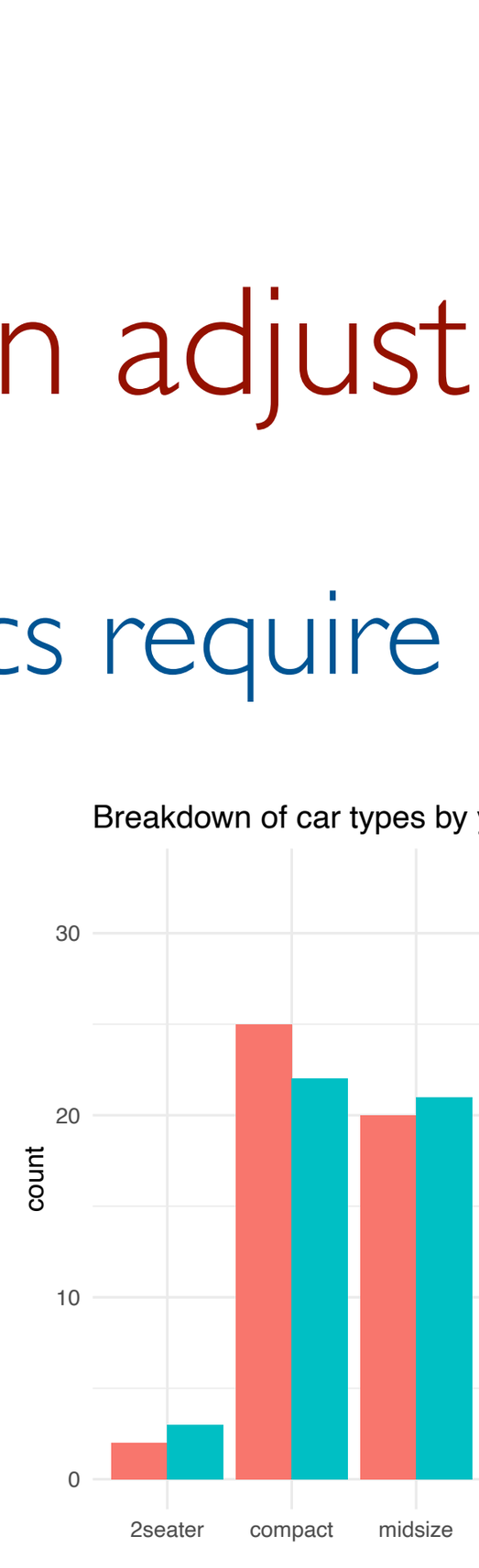
Many statistical graphics utilize **stat**

- Box plot
  - Five-number summary + outlie
- Histogram
  - Binning
- Bar plot
  - Counting

Position adjust

Many statistical graphics require

- Scatter plot
  - Jitter
- Bar plot
  - Dodge
  - Stack



Coordinate sy

Some graphics may require differe

- Cartesian
- Polar
- Map

Implementing a gramm

A version of the "layered grammar of gr


```
ggplot(data = <D T SET>,
       mapping =  es(<M PPINGS
  l yer(geom = <GEOM>,
        stat = <ST T>,
        position = <POSITION>)
  <SC LE_FUNCTION>() +
  <COORDIN TE_FUNCTION>() +
  <F CET_FUNCTION>()
```

or, more simply

```
ggplot(data = <D T SET>,
       mapping =  es(<M PPINGS
  <GEOM_FUNCTION>()
```

Recipes for common sta

- Scatter plot
  - Geom = “point”
  - Stat = “identity”
- Line plot
  - Geom = “line”
  - Stat = “identity”
- Box plot
  - Geom = “boxplot”
  - Stat = “boxplot”

GGPLOT