GETTING STARTED

Check the Jotter/Wiki for a list of packages to install and update.

Log into sharelatex.com - make sure you know your username and password.

Review the assumptions and hypotheses for the various tests from last week.

CHRISTOPHER PRENER, PH.D.
FALL, 2017

WEEK 08
LECTURE 10

QUANTITATIVE ANALYSIS

DIFFERENCE OF MEANS (PART 2)

AGENDA

1 FRONT MATTER

ANNOUNCEMENTS

No video lectures next week!

Lab-08 & PS-06 due Monday, 10/30 by 4:15pm

Handout on papers in LaTeX will be posted on GitHub.

PS-07 will be a data cleaning puzzle, due Monday 10/30 as well

Midterm grade repots will be sent via GitHub

Lab-09 will be waived

2 GETTING STARTED WITH LaTeX

LAH-*tekh*

LAY-*tekh*

WHAT IS LaTeX?

- LaTeX is a *typesetting system* designed for professional documents
- It excels at documents that require a standard format.
  - Academic publishers often use it to create the format of journal articles
  - All of the documents (except the slides) for this course are produced using LaTeX.

```
\documentclass{article}
\usepackage[utf8]{inputenc}

\title{Learning \LaTeX{}}
\author{Christopher Prener, Ph.D.}
\date{October 12\textsuperscript{th}, 2017}

\begin{document}

\maketitle

\section{Introduction}
The quick brown fox jumped over the lazy sociologist.

\end{document}
```

Learning LaTeX

Christopher Prener, Ph.D.

October 12th, 2017

1 Introduction

The quick brown fox jumped over the lazy sociologist.

WHAT IS LaTeX?

- LaTeX, like R, has many components, including a core set of code, packages, and various graphic user interfaces for producing output.
    - Like R, it is really an *ecosystem* rather than a single thing you download and use
- Like Markdown, it uses "markup" syntax to indicate how text should be formatted.

```
\documentclass{article}
\usepackage[utf8]{inputenc}

\title{Learning \LaTeX{}}
\author{Christopher Prener, Ph.D.}
\date{October 12\textsuperscript{th}, 2017}

\begin{document}

\maketitle

\section{Introduction}
The quick brown fox jumped over the lazy sociologist.

\end{document}
```

Learning LaTeX

Christopher Prener, Ph.D.

October 12$^{\text{th}}$, 2017

1 Introduction

The quick brown fox jumped over the lazy sociologist.

WHY BOTHER?

- LATEX, like R, has a specific syntax and logical structure.
  - Like R, LATEX has a reputation for being difficult.
- This syntax separates *content* from *formatting*, and makes it easy to alter the *structure* of a document
- For the most part, we can dispense with concern with *formatting* and rather focus on our strong suit, which is developing *content*.

```
\documentclass{article}
\usepackage[utf8]{inputenc}

\title{Learning \LaTeX{}}
\author{Christopher Prener, Ph.D.}
\date{October 12\textsuperscript{th}, 2017}

\begin{document}

\maketitle

\section{Introduction}
The quick brown fox jumped over the lazy sociologist.

\end{document}
```

Learning LATEX

Christopher Prener, Ph.D.

October 12th, 2017

1 Introduction

The quick brown fox jumped over the lazy sociologist.

WHY BOTHER?

- LaTeX, like RMarkdown, ultimately should *simplify* rather than *complicate* your work:
  - It can keep track of your table of contents, page references, index items, figures, tables, and references
  - Table output can be produced by R packages
  - Equations can be written and reproduced easily

```
$ \bar { x } =
   \frac { \sum _{ i=1 }^{ n }{ { x }_{ i } }  }
   { n }  $
```

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

PREAMBLE COMMANDS

`\documentclass[`*classOptions*`]{`*className*`}`

Using the `article` class:

```
\documentclass[11pt]{article}
```

This is a *required* part of the document's **preamble**; the default body font size is 10 point font. Take care of my aging 👀!

PREAMBLE COMMANDS

`\usepackage[`*`packageOptions`*`]{`*`packageName`*`}`

Using the `inputenc` package with the `utf8` option:

```
\usepackage[utf8]{inputenc}
```

The `inputenc` package allows you to add accented characters (i.e. é) to your document.

PREAMBLE COMMANDS

```
\title{Title}
\author{Author Name}
\date{Month dd, yyyy}
```

Specifying title elements:

```
\title{My First Article}
\author{Christopher Prener, Ph.D.}
\date{October 16, 2017}
```

These require a separate command to be returned as output.

THE DOCUMENT BODY

```
\begin{document}
% insert body text here
\end{document}
```

Specifying title elements:

```
\begin{document}
The quick brown fox jumps over the lazy sociologist.
\end{document}
```

This is a *required* part of the document's **body**. The percent symbol (%) is for **comments**.

INSERT YOUR TITLE

```
\maketitle
```

Add the title specified in your preamble:

```
\maketitle
```

Will only return the title elements you have specified (i.e. if you do not add a date, one will not be included in your output).

PUTTING IT ALL TOGETHER

```
\documentclass{article}
\usepackage[utf8]{inputenc}

\title{My First Article}
\author{Christopher Prener, Ph.D.}
\date{October 16, 2017}

\begin{document}

\maketitle

The quick brown fox jumped over the lazy sociologist.
\end{document}
```

PUTTING IT ALL TOGETHER

My First Article

Christopher Prener, Ph.D.

October 16, 2017

The quick brown fox jumped over the lazy sociologist.

TEXT FORMATTING: ITALICS

```
\textit{text}
```

Italicizing parts of a sentence:

```
The \textit{quick} brown fox jumped over the
    \textit{lazy} sociologist.
```

The *quick* brown fox jumped over the *lazy* sociologist.

TEXT FORMATTING: BOLD

```
\textbf{text}
```

Bolding parts of a sentence:

```
The \textbf{quick} brown fox jumped over the
    \textbf{lazy} sociologist.
```

The **quick** brown fox jumped over the **lazy** sociologist.

TEXT FORMATTING: MIXING STYLES

```
\textbf{\textit{text}}
```

Italicizing parts of a bolded sentence:

```
\textbf{The \textit{quick} brown fox jumped over the
    \textit{lazy} sociologist.}
```

The *quick* brown fox jumped over the *lazy* sociologist.

TEXT FORMATTING: TYPEWRITER FONT

`\texttt{text}`

Adding typewriter font to parts of a sentence:

```
The \texttt{quick} brown fox jumped over the
    \texttt{lazy} sociologist.
```

The `quick` brown fox jumped over the `lazy` sociologist.

TEXT FORMATTING: SANS SERIF FONT

```
\textsf{text}
```

Adding sans serif font to parts of a sentence:

```
The \textsf{quick} brown fox jumped over the
    \texsf{lazy} sociologist.
```

The quick brown fox jumped over the lazy sociologist.

TEXT FORMATTING: SPECIAL CHARACTERS

The following characters have special meaning in LaTeX:

& % $ # _ { } ~ ^ \

If you do not “escape” these characters, they will generate errors.

The first seven of these can be “escaped” using a backslash (\):

\& \% \$ \# _ \{ \}

TEXT FORMATTING: SPECIAL CHARACTERS

`\%`

Adding a percentage symbol to a sentence:

```
The quick brown fox jumped over 25\% of the lazy
    sociologists.
```

The quick brown fox jumped over 25% of the lazy sociologists.

TEXT FORMATTING: SPECIAL CHARACTERS

The following characters have special meaning in LaTeX:

& % $ # _ { } ~ ^ \

If you do not "escape" these characters, they will generate errors.

The last three of these have dedicated "macros":

```
\textasciitilde
\textasciicircum
\textbackslash
```

DOCUMENT FORMATTING: HEADINGS

```
\section{Section Title}
```

Add a top-level heading to an `article` class document:

```
\section{Introduction}
The quick brown fox jumped over the lazy sociologist.
```

1 Introduction

The *quick* brown fox jumped over the *lazy* sociologist.

DOCUMENT FORMATTING: SUBHEADINGS

`\subsection{Subsection Title}`

Add a second-level heading to an `article` class document:

```
\subsection{Background}
The quick brown fox jumped over the lazy sociologist.
```

1.1 Background

The *quick* brown fox jumped over the *lazy* sociologist.

DOCUMENT FORMATTING: PARAGRAPHS

```
\par Paragraph text
```

Add paragraph breaks to body text:

```
\section{Introduction}
\par The quick brown fox jumped over the lazy
    sociologist. The sociologist was sleeping after
    reading Durkheim.
\par What the sociologist really needed that
    afternoon was a break from classical theory. Not
    all sociologists love social theory the way that
    some do.
```

DOCUMENT FORMATTING: PARAGRAPHS

1 Introduction

The quick brown fox jumped over the lazy sociologist. The sociologist was sleeping after reading Durkheim.

What the sociologist really needed that afternoon was a break from classical theory. Not all sociologists love social theory the way that some do.

WE NEED A BINARY VARIABLE!

```
library(tidyverse)

mpg %>%
  mutate(foreign = ifelse(manufacturer == "audi" |
                           manufacturer == "honda" |
                           manufacturer == "hyundai" |
                           manufacturer == "land rover" |
                           manufacturer == "nissan" |
                           manufacturer == "subaru" |
                           manufacturer == "toyota" |
                           manufacturer == "volkswagen",
                         TRUE, FALSE)) %>%
  select(cty, hwy, foreign) -> autoData
```

LaTeX TABLES FROM R

`stargazer::stargazer(dataFrame, title = “title”)`

Basic usage:

```
> stargazer(autoData, title = “Descriptive
    Statistics”)
# LaTeX output will be returned
```

If your data are stored as tibbles, you will need to coerce them back to data frames using `base::as.data.frame()`!

Only output for numeric and logical variables will be returned.

LaTeX TABLES FROM R

```
> stargazer(as.data.frame(autoData), title = "Descriptive Statistics")

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Oct 12, 2017 - 13:34:57
\begin{table}[!htbp] \centering
  \caption{Descriptive Statistics}
  \label{}
\begin{tabular}{@{\extracolsep{5pt}}lccccc}
\\[-1.8ex]\hline
\hline \\[-1.8ex]
Statistic & \multicolumn{1}{c}{N} & \multicolumn{1}{c}{Mean} & \multicolumn{1}{c}{St. Dev.} &
\multicolumn{1}{c}{Min} & \multicolumn{1}{c}{Max} \\
\hline \\[-1.8ex]
cty & 234 & 16.859 & 4.256 & 9 & 35 \\
hwy & 234 & 23.440 & 5.955 & 12 & 44 \\
foreign & 234 & 0.568 & 0.496 & 0 & 1 \\
\hline \\[-1.8ex]
\end{tabular}
\end{table}
```

LaTeX TABLES FROM R

```
> stargazer(as.data.frame(autoData), title = ‘‘Descriptive Statistics")

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Oct 12, 2017 - 13:34:57
\begin{table}[!htbp] \centering
  \caption{Descriptive Statistics}
  \label{}
\begin{tabular}{@{\extracolsep{5pt}}lccccc}
\\[-1.8ex]\hline
\hline \\[-1.8ex]
Statistic & \multicolumn{1}{c}{N} & \multicolumn{1}{c}{Mean} & \multicolumn{1}{c}{St. Dev.} &
\multicolumn{1}{c}{Min} & \multicolumn{1}{c}{Max} \\
\hline \\[-1.8ex]
cty & 234 & 16.859 & 4.256 & 9 & 35 \\
hwy & 234 & 23.440 & 5.955 & 12 & 44 \\
foreign & 234 & 0.568 & 0.496 & 0 & 1 \\
\hline \\[-1.8ex]
\end{tabular}
\end{table}
```

LaTeX TABLES FROM R

2 Descriptive Statistics

Table 1: Descriptive Statistics

Statistic	N	Mean	St. Dev.	Min	Max
cty	234	16.859	4.256	9	35
hwy	234	23.440	5.955	12	44
foreign	234	0.568	0.496	0	1

3 VARIANCE TESTING

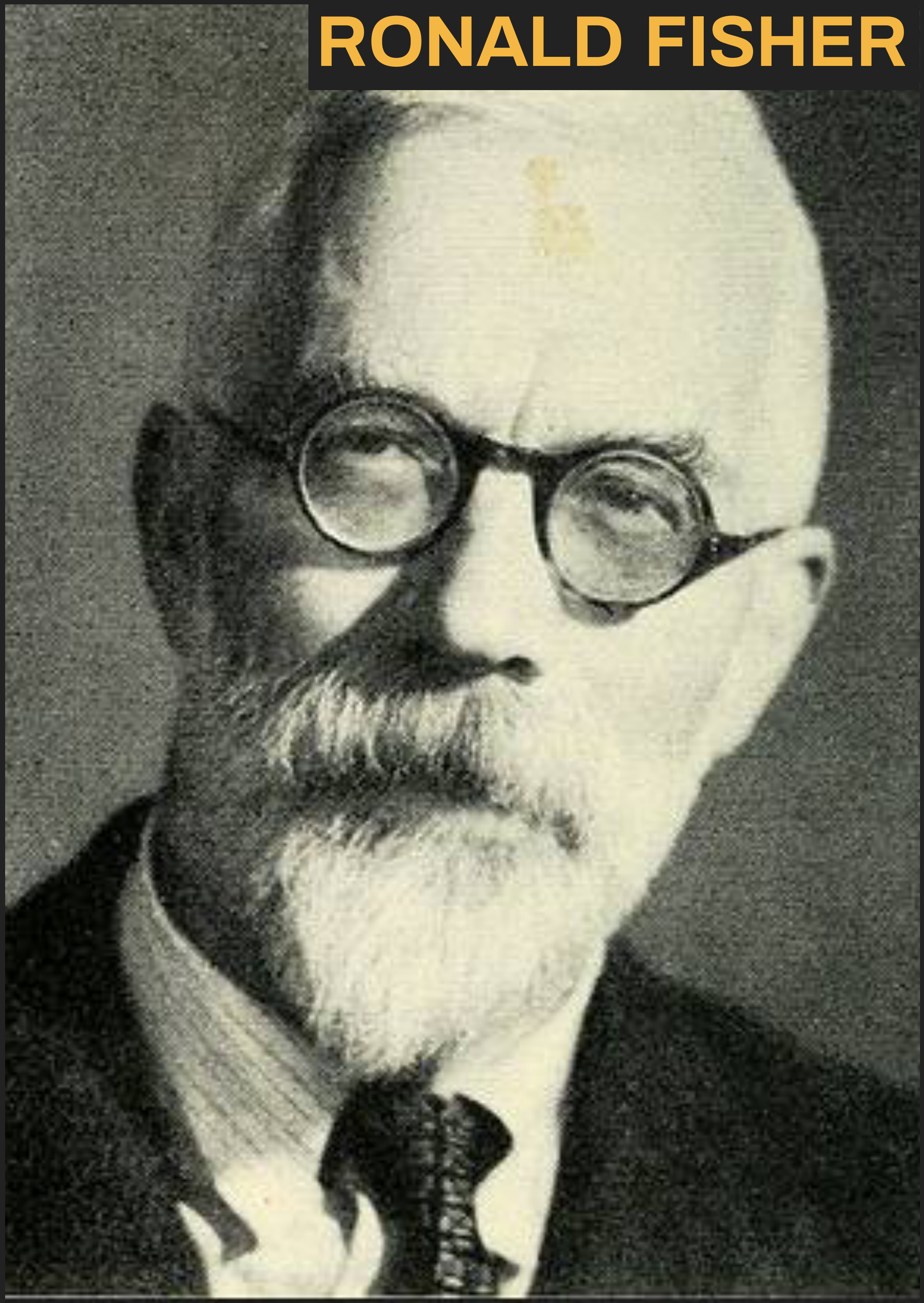
QUICK REVIEW

What does the Levne's test accomplish?

- The Levene's test is used for assessing the homogeneity of variance assumption.
  - $\mathrm{H}_0 = \text{The two variances are approximately equal.}$
  - $\mathrm{H}_1 = \text{The two variances are unequal.}$
- `R`'s implementation of the Levene's test uses the median, rather than the mean, for this comparison.

F-DISTRUBTION

- Named in honor of Ronald Fisher
- Models the distribution of the ratio between two groups based on their variance
- Used to test whether two estimates of variance can be assumed to come from the same population
- Not symmetrical like t, and its shape varies based on the given degrees of freedom



LEVENE'S TEST

```
car::leveneTest(yVar ~ xVar, data = dataFrame)
```

Using the new `foreign` variable and `hwy` from `ggplot2`'s `mpg` data:

```
> leveneTest(hwy ~ foreign, data = autoData)
# test output returned
```

The `leveneTest()` function will temporarily convert string or logical variables to factors to compute the test.

LEVENE'S TEST

```
> leveneTest(hwy ~ foreign, data = autoData)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  0.5867 0.4445
      232
Warning message:
In leveneTest.default(y = y, group = group, ...) : group coerced to
factor.
```

How would you interpret this result?

LEVENE'S TEST

```
> leveneTest(hwy ~ foreign, data = autoData)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  0.5867 0.4445
      232
Warning message:
In leveneTest.default(y = y, group = group, ...) : group coerced to
factor.
```

The results of the Levene's Test ($f = 0.587,\ p = 0.445$) suggest that the variance in highway fuel efficiency for domestic cars is approximately the same as the variance in highway fuel efficiency for foreign cars.

MODELING IN R

```
car::leveneTest(yVar ~ xVar, data = dataFrame)
```

The accent symbol (~) is used to separate the lefthand side of a model's equation from the righthand side.

The lefthand side is always for the dependent variable - the main outcome we are interested in understanding. We always call this variable y.

The righthand side is for our dependent variables, which we always refer to as x variables.

4 ONE OR TWO SAMPLES

QUICK REVIEW

What is the one-sample t test used for?

- The one-sample t test is used for assessing whether the sample is drawn from a population by comparing their means.

  - H_0 = The difference between the sample mean and the population's (i.e. the "true" mean) is approximately zero.

  - H_1 = difference between the sample mean and the population's (i.e. the "true" mean) is substantively different from zero.

ONE-SAMPLE T TEST

`stats::t.test(`*`dataFrame$yVar`*`, mu = `*`val`*`)`

Using the `hwy` variable from `ggplot2`'s `mpg` data:

```
> t.test(autoData$hwy, mu = 24.25)
# returns test output
```

μ (`mu`) is the population mean.

ONE-SAMPLE T TEST

```
> t.test(autoData$hwy, mu = 24.25)

    One Sample t-test

data:  autoData$hwy
t = -2.0804, df = 233, p-value = 0.03858
alternative hypothesis: true mean is not equal to 24.25
95 percent confidence interval:
 22.67324 24.20710
sample estimates:
mean of x
 23.44017
```

ONE-SAMPLE T TEST

```
> t.test(autoData$hwy, mu = 24.25)

    One Sample t-test

data:  autoData$hwy
t = -2.0804, df = 233, p-value = 0.03858
```

? How would you interpret this result?

ONE-SAMPLE T TEST

```
> t.test(autoData$hwy, mu = 24.25)

	One Sample t-test

data:  autoData$hwy
t = -2.0804, df = 233, p-value = 0.03858
```

The results of the one-sample t test (t = -2.080, $p = 0.039$) suggest that the sample mean (23.440) is not drawn from a population where $\mu = 24.250$.

QUICK REVIEW

What is the two-sample (independent) t test used for?

- The two-sample (independent) t test is used for assessing whether the mean of y for one group is approximately equal to the mean of y for another.
  - $\mathrm{H}_0 =$ The difference in means is approximately zero.
  - $\mathrm{H}_1 =$ The difference in means is substantively greater than zero.

TWO-SAMPLE T TEST

```
stats::t.test(dataFrame$yVar ~ dataFrame$xVar,
    var.equal = FALSE)
```

Using the new `foreign` variable and `hwy` from `ggplot2`'s `mpg` data:

```
> t.test(autoData$hwy ~ autoData$foreign,
    var.equal = TRUE)
# returns test output
```

Remember that x should be a logical variable. If var.equal is `FALSE`, Welch's corrected degrees of freedom are used.

TWO-SAMPLE T TEST

```
> t.test(autoData$hwy ~ autoData$foreign, var.equal = TRUE)

	Two Sample t-test

data:  autoData$hwy by autoData$foreign
t = -11.178, df = 232, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.348850 -5.846788
sample estimates:
mean in group FALSE  mean in group TRUE 
           19.40594            26.50376 
```

TWO-SAMPLE T TEST

```
> t.test(autoData$hwy ~ autoData$foreign, var.equal = TRUE)

   Two Sample t-test

data:  autoData$hwy by autoData$foreign
t = -11.178, df = 232, p-value < 2.2e-16
```

How would you interpret this result?

TWO-SAMPLE T TEST

```
> t.test(autoData$hwy ~ autoData$foreign, var.equal = TRUE)

	Two Sample t-test

data:  autoData$hwy by autoData$foreign
t = -11.178, df = 232, p-value < 2.2e-16
```

The results of the independent t test ($t =$ -11.178, $p < 0.001$) suggest that the mean fuel efficiency for domestic cars (19.406 miles per gallon) is not equal to the mean fuel efficiency for foreign cars (26.504 miles per gallon). Foreign cars are more fuel efficient than domestic vehicles.

5 DEPENDENT SAMPLES

EXAMPLE DATA

```
> library(stlData)
> library(tidyr)
> income <- as_tibble(stlIncome)
> income
# A tibble: 106 x 8
        geoID tractCE          nameLSAD  mi10 mi10_moe mi10_inflate  mi15 mi15_moe
        <dbl>   <int>            <fctr> <dbl>    <dbl>        <dbl> <dbl>    <dbl>
 1 29510101100  101100 Census Tract 1011 45530     9265     49106.38 56169     6278
 2 29510101200  101200 Census Tract 1012 58684     9715     63293.63 54464     7495
 3 29510101300  101300 Census Tract 1013 44403     6734     47890.86 49808     9626
 4 29510101400  101400 Census Tract 1014 40100     9341     43249.86 39183     5966
 5 29510101500  101500 Census Tract 1015 30266     5736     32643.39 30346     8053
 6 29510101800  101800 Census Tract 1018 27439     5485     29594.33 36424     6061
 7 29510102100  102100 Census Tract 1021 35475     2864     38261.56 45775     5000
 8 29510102200  102200 Census Tract 1022 57303     3319     61804.15 67534     9711
 9 29510102300  102300 Census Tract 1023 53277    10920     57461.91 49969     3984
10 29510102400  102400 Census Tract 1024 39191     7145     42269.45 39479     5512
# ... with 96 more rows
```

QUICK REVIEW

What is the difference between wide and long data? Are the `stlIncome` data wide or long?

- Wide data include a row for each observation and multiple columns for different time points or groupings.
- Long data include multiple rows for each observation, one for each time point or grouping.
- The `stlIncome` data are wide.

BEFORE RESHAPING...

```
> library(dplyr)
> income <- select(income, geoID, mi10_inflate, mi15)
> income
# A tibble: 106 x 3
         geoID mi10_inflate  mi15
         <dbl>        <dbl> <dbl>
 1 29510101100     49106.38 56169
 2 29510101200     63293.63 54464
 3 29510101300     47890.86 49808
 4 29510101400     43249.86 39183
 5 29510101500     32643.39 30346
 6 29510101800     29594.33 36424
 7 29510102100     38261.56 45775
 8 29510102200     61804.15 67534
 9 29510102300     57461.91 49969
10 29510102400     42269.45 39479
# ... with 96 more rows
```

RESHAPING DATA TO LONG

f(x)

`tidyr::gather(dataFrame, key, value, ...)`

Using the `stlIncome` data:

```
> incomeLong <- gather(income, period, estimate,
    mi10_inflate, mi15)
```

After you reshape, reordering observations (using `dplyr::arrange()`) and recoding the key (using `dplyr::mutate()`) are good practices.

RESHAPING DATA TO LONG

```
> incomeLong <- gather(income, period, estimate, mi10_inflate, mi15)
> incomeLong
# A tibble: 212 x 3
         geoID       period estimate
         <dbl>        <chr>    <dbl>
 1 29510101100 mi10_inflate 49106.38
 2 29510101200 mi10_inflate 63293.63
 3 29510101300 mi10_inflate 47890.86
 4 29510101400 mi10_inflate 43249.86
 5 29510101500 mi10_inflate 32643.39
 6 29510101800 mi10_inflate 29594.33
 7 29510102100 mi10_inflate 38261.56
 8 29510102200 mi10_inflate 61804.15
 9 29510102300 mi10_inflate 57461.91
10 29510102400 mi10_inflate 42269.45
# ... with 202 more rows
```

RESHAPING DATA TO WIDE

f(x) `tidyr::spread(dataFrame, key, value)`

Using the `stlIncome` data:

```
> incomeWide <- spread(incomeLong, period, estimate)
```

After you reshape, reordering observations (using `dplyr::arrange()`) and recoding the key (using `dplyr::mutate()`) are good practices.

RESHAPING DATA TO WIDE

```
> incomeWide <- spread(incomeLong, period, estimate)
> incomeWide
# A tibble: 106 x 3
         geoID mi10_inflate  mi15
 *       <dbl>        <dbl> <dbl>
 1 29510101100     49106.38 56169
 2 29510101200     63293.63 54464
 3 29510101300     47890.86 49808
 4 29510101400     43249.86 39183
 5 29510101500     32643.39 30346
 6 29510101800     29594.33 36424
 7 29510102100     38261.56 45775
 8 29510102200     61804.15 67534
 9 29510102300     57461.91 49969
10 29510102400     42269.45 39479
# ... with 96 more rows
```

WHAT TO USE WHEN

f(x) The `t.test()` function requires wide data.

Plots from `ggplot2` require long data.

QUICK REVIEW

What does the dependent t test accomplish?

- The dependent t test is used for assessing the difference means between two groups or time periods where probabilistic independence cannot be assumed.
  - $\text{H}_0 = \text{The difference in means is approximately zero.}$
  - $\text{H}_1 = \text{The difference in means is substantively greater than zero.}$

ASSUMPTION CHECKS

One of the assumptions is that the difference between y_1 and y_2 is normally distributed. You need to manually create a variable with those differences in your *wide* data to test this assumption.

```
> income <- mutate(income, yDiff = mi15-mi10_inflate)
```

DEPENDENT SAMPLES T TEST

```
stats::t.test(y1, y2, paired = TRUE)
```

Using the `stlIncome` data:

```
> t.test(income$mi10_inflate, income$mi15,
    paired = TRUE)
# returns test output
```

The order of y_1 and y_2 is not substantively important.

DEPENDENT SAMPLES T TEST

```
> t.test(income$mi10_inflate, income$mi15, paired = TRUE)

	Paired t-test

data:  income$mi10_inflate and income$mi15
t = 2.6556, df = 105, p-value = 0.009151
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  486.0955 3351.4629
sample estimates:
mean of the differences
               1918.779
```

DEPENDENT SAMPLES T TEST

```
> t.test(income$mi10_inflate, income$mi15, paired = TRUE)

	Paired t-test

data:  income$mi10_inflate and income$mi15
t = 2.6556, df = 105, p-value = 0.009151
```

? How would you interpret this result?

DEPENDENT SAMPLES T TEST

```
> t.test(income$mi10_inflate, income$mi15, paired = TRUE)

	Paired t-test

data:  income$mi10_inflate and income$mi15
t = 2.6556, df = 105, p-value = 0.009151
```

The results of the independent t test ($t = 2.656$, $p = 0.009$) suggest that the average median income in 2010 ($\$36{,}006.88$) is not equal to the average median income in 2015 ($\$34{,}088.10$). There has been a substantive drop in median income by census tract in St. Louis, Missouri over this period in time.

TEST OUTPUT IS...MESSY

```
> t.test(income$mi10_inflate, income$mi15, paired = TRUE)

	Paired t-test

data:  income$mi10_inflate and income$mi15
t = 2.6556, df = 105, p-value = 0.009151
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  486.0955 3351.4629
sample estimates:
mean of the differences
               1918.779
```

TIDY OUTPUT

f(x) `broom::tidy(`*`testFunction`*`)`

Using the `stlIncome` data:

```
> ttestResult1 <- tidy(t.test(income$mi10_inflate,
    income$mi15, paired = TRUE))
```

! Will not return any output if successful.

TIDY OUTPUT

```
> ttestResult1 <- tidy(t.test(income$mi10_inflate, income$mi15,
paired = TRUE))

> print(ttestResult1$statistic)
[1] 2.655565

> print(ttestResult1$p.value)
[1] 0.009151292
```

6 EFFECT SIZES

QUICK REVIEW

What is an effect size?

- An effect size shows use the “real world” significance as opposed to the statistical significance - is the final a “small”, “medium”, or “large” effect?

COHEN'S d

```
effsize::cohen.d(dataFrame$yVar ~ dataFrame$xVar,
    pooled = TRUE, paired = FALSE)
```

Using the new `foreign` variable and `hwy` from `ggplot2`'s `mpg` data:

```
> cohen.d(autoData$hwy ~ autoData$foreign, pooled =
    TRUE, paired = FALSE)
# returns test output
```

The `cohen.d()` function will temporarily convert string or logical variables to factors to compute the test.

COHEN'S d

```
> cohen.d(autoData$hwy ~ autoData$foreign, pooled = TRUE, paired =
FALSE)

Cohen's d

d estimate: 1.51912 (large)
95 percent confidence interval:
     inf      sup
1.224565 1.813675
Warning message:
In cohen.d.formula(autoData$hwy ~ autoData$foreign, pooled = TRUE,  :
  Cohercing rhs of formula to factor
```

COHEN'S d

```
> cohen.d(autoData$hwy ~ autoData$foreign, pooled = TRUE, paired =
FALSE)


Cohen's d


d estimate: 1.51912 (large)
```

How would you interpret this result?

COHEN'S d

```
> cohen.d(autoData$hwy ~ autoData$foreign, pooled = TRUE, paired =
FALSE)

Cohen's d

d estimate: 1.51912 (large)
```

The Cohen's D effect size ($D = 1.519$) is a large effect - the difference in mean fuel efficiency between foreign and domestic cars is notable in addition to being statistically significant.

COHEN'S d

```
effsize::cohen.d(dataFrame$y1, dataFrame$y2,
    paired = TRUE)
```

Using the new `foreign` variable and `hwy` from `ggplot2`'s `mpg` data:

```
> cohen.d(income$mi10_inflate, income$mi15,
    paired = TRUE)
# returns test output
```

The `pooled` parameter is not needed with paired data.

COHEN'S d

```
> cohen.d(income$mi10_inflate, income$mi15, paired = TRUE)

Cohen's d

d estimate: 0.2579313 (small)
95 percent confidence interval:
        inf         sup
-0.01397459  0.52983716
```

7 SAMPLE SIZE ESTIMATE

REVIEW: STATISTICAL POWER

$$p(\text{Type II}) = \beta$$
$$1\text{-}\beta = \text{power}$$

Sample	Population	
	$\mu = \mu_0$	$\mu \neq \mu_0$
Not Reject	yes	Type II
Reject	Type I	yes

*The null hypothesis is that $\mu = \mu_0$

$$p(\text{Type I}) = \boldsymbol{\alpha}$$

FINDING n

`pwr::pwr.t.test(`*d*, *power*, *sigLevel*, *type*, *alternative*`)`

A moderate effect size ($d = .5$) with statistical power of $.9$:

```
> pwr.t.test(d = .5, power = .9, sig.level = .05,
    type = "two.sample", alternative = "two.sided")
# returns test output
```

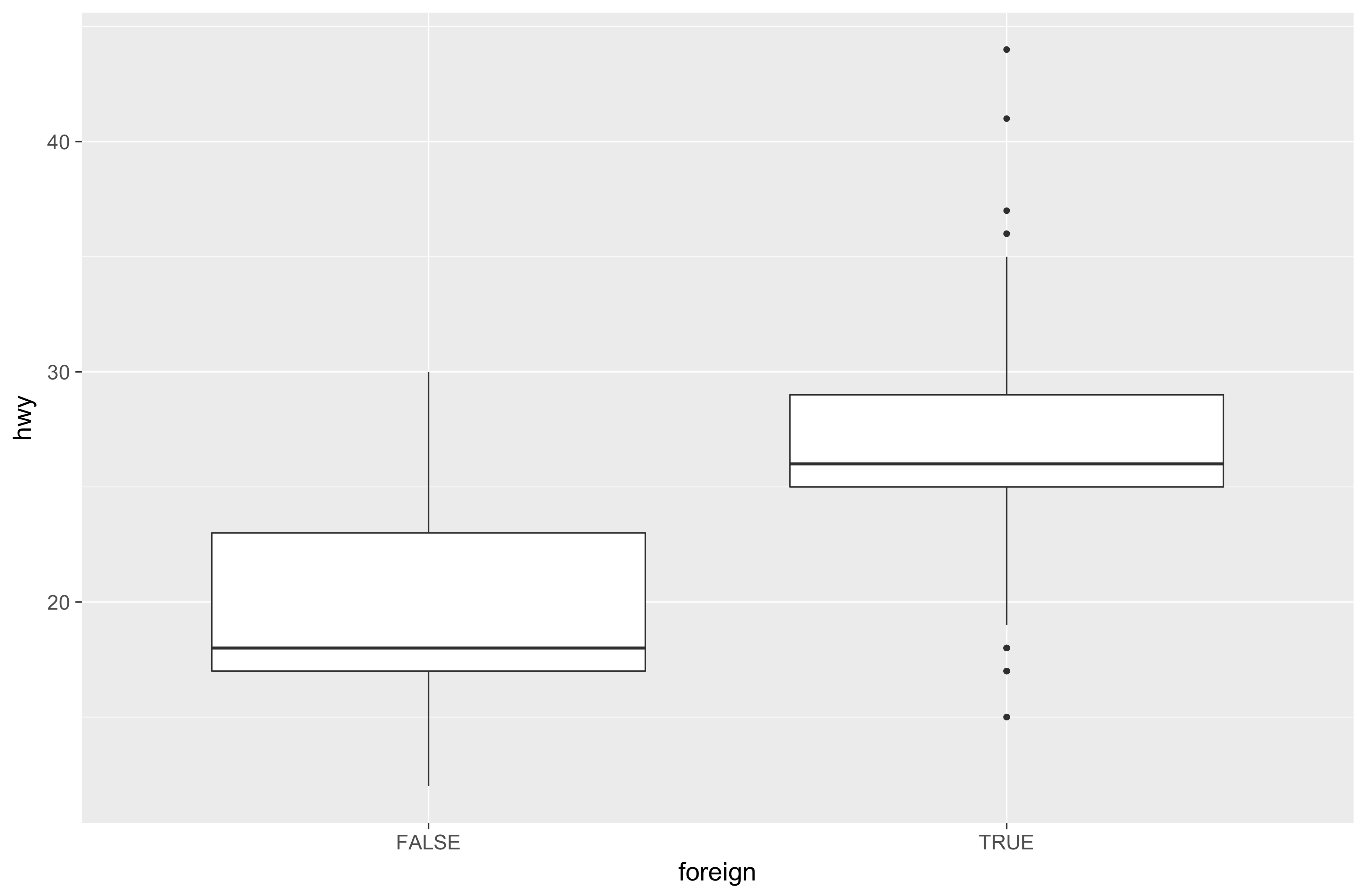
Other options for `type` are "`one.sample`" and "`paired`".

FINDING n

```
> pwr.t.test(d = .5, power = .9, sig.level = .05, type =
"two.sample", alternative = "two.sided")

     Two-sample t test power calculation

              n = 85.03128
              d = 0.5
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

8 PLOTS FOR MEAN DIFFERENCE

BOX PLOT



BOX PLOT

```
ggplot2::geom_boxplot(mapping = aes(aesthetic))
```

Using the `hwy` and `foreign` variables created earlier from `ggplot2`'s `mpg` data:

```
> ggplot(data = autoData) +
    geom_boxplot(mapping = aes(x = foreign, y = hwy))
```

The x variable should be discrete (binary, factor, or character), and the y variable should be continuous.

BOX PLOT

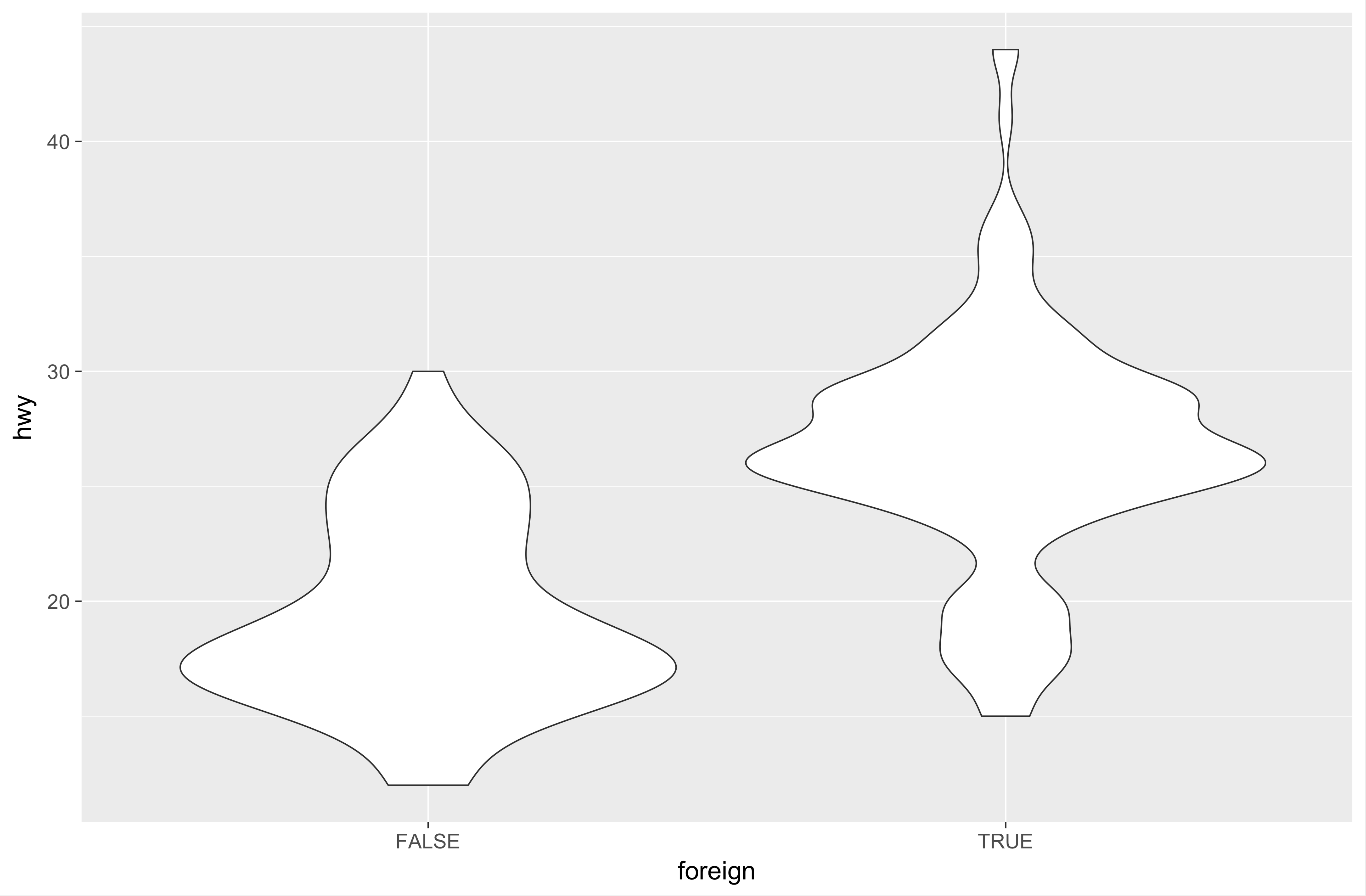
`ggplot2::geom_boxplot(mapping = aes(`*`aesthetic`*`))`

Using the `hwy` and `foreign` variables created earlier from `ggplot2`'s `mpg` data:

```
> ggplot(data = autoData) +
    geom_boxplot(mapping = aes(x = foreign, y = hwy))
```

Box plots are important parts of exploratory data analysis, but are less ideal for lay consumption.

VIOLIN PLOT



VIOLIN PLOT

```
ggplot2::geom_violin(mapping = aes(aesthetic))
```

Using the `hwy` and `foreign` variables created earlier from `ggplot2`'s `mpg` data:

```
> ggplot(data = autoData) +
    geom_violin(mapping = aes(x = foreign, y = hwy))
```

The x variable should be discrete (binary, factor, or character), and the y variable should be continuous.

VIOLIN PLOT

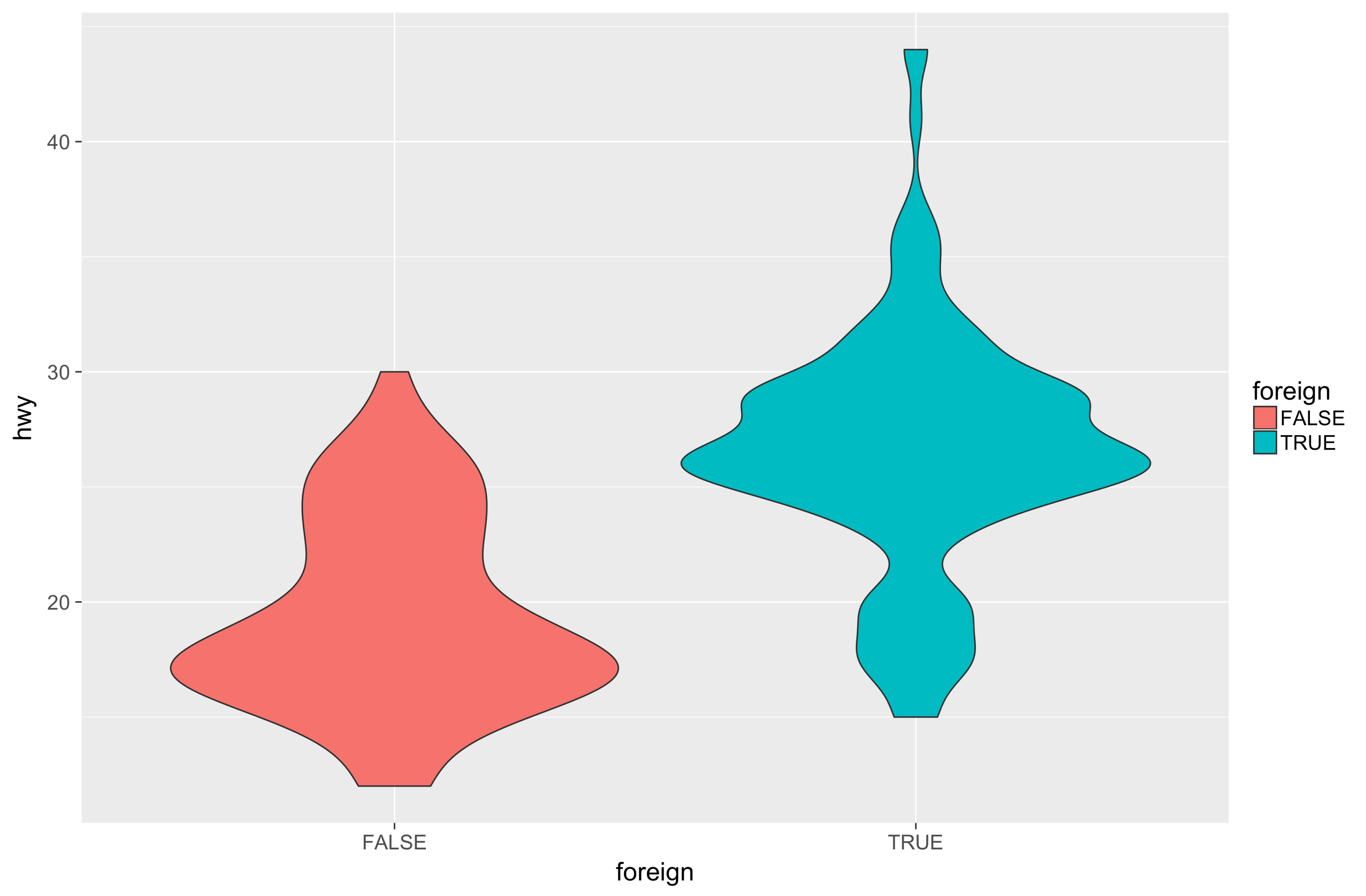
```
ggplot2::geom_violin(mapping = aes(aesthetic))
```

Using the `hwy` and `foreign` variables created earlier from `ggplot2`'s `mpg` data:

```
> ggplot(data = autoData) +
    geom_violin(mapping = aes(x = foreign, y = hwy,
      fill = foreign))
```

The x variable should be discrete (binary, factor, or character), and the y variable should be continuous.

VIOLIN PLOT



VIOLIN PLOT WITH MEAN POINTS

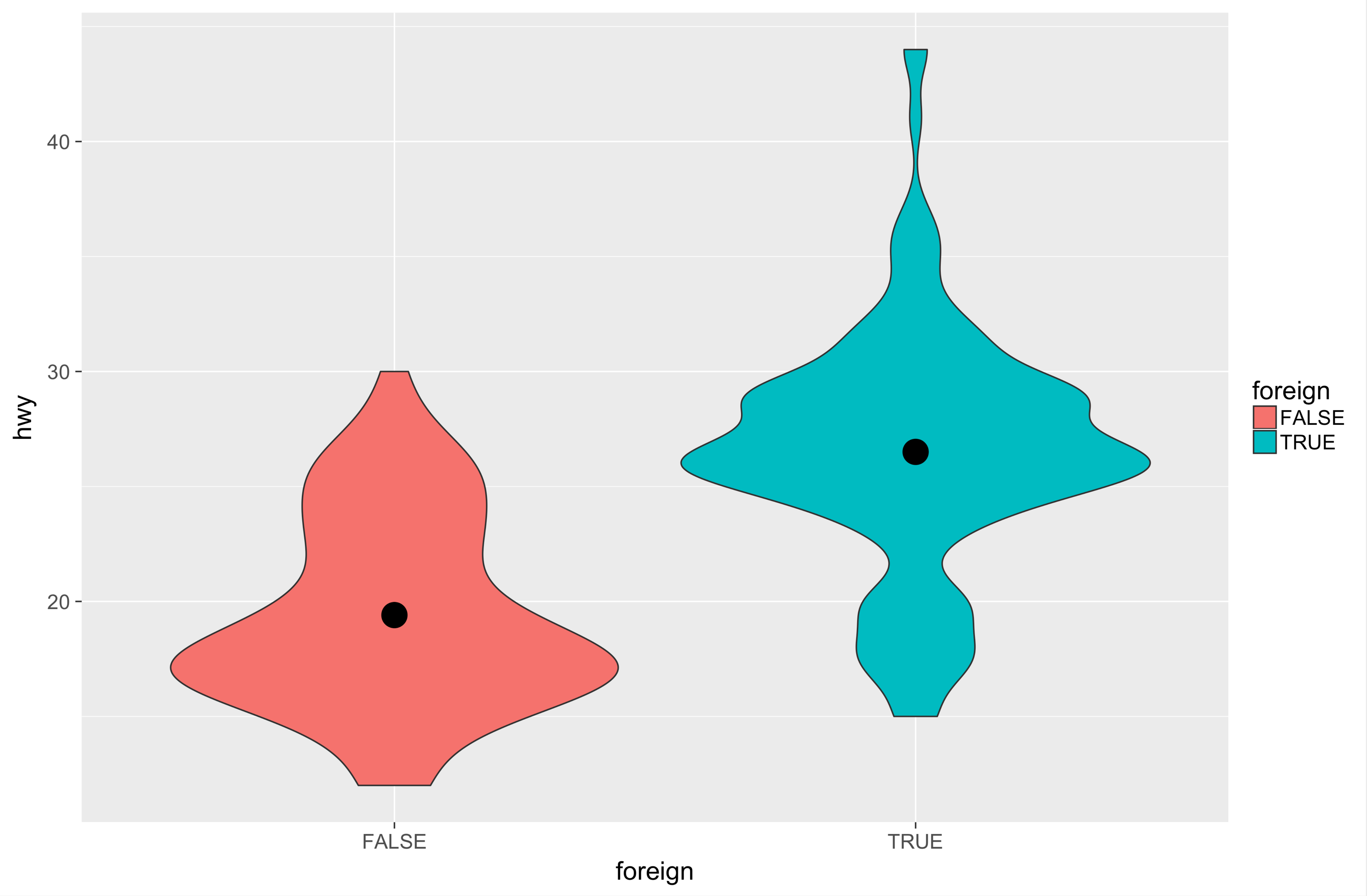
```
ggplot2::geom_violin(mapping = aes(aesthetic)) +
ggplot2::stat_summary(fun.y = mean, geom = "point")
```

Using the `hwy` and `foreign` variables created earlier from `ggplot2`'s `mpg` data:

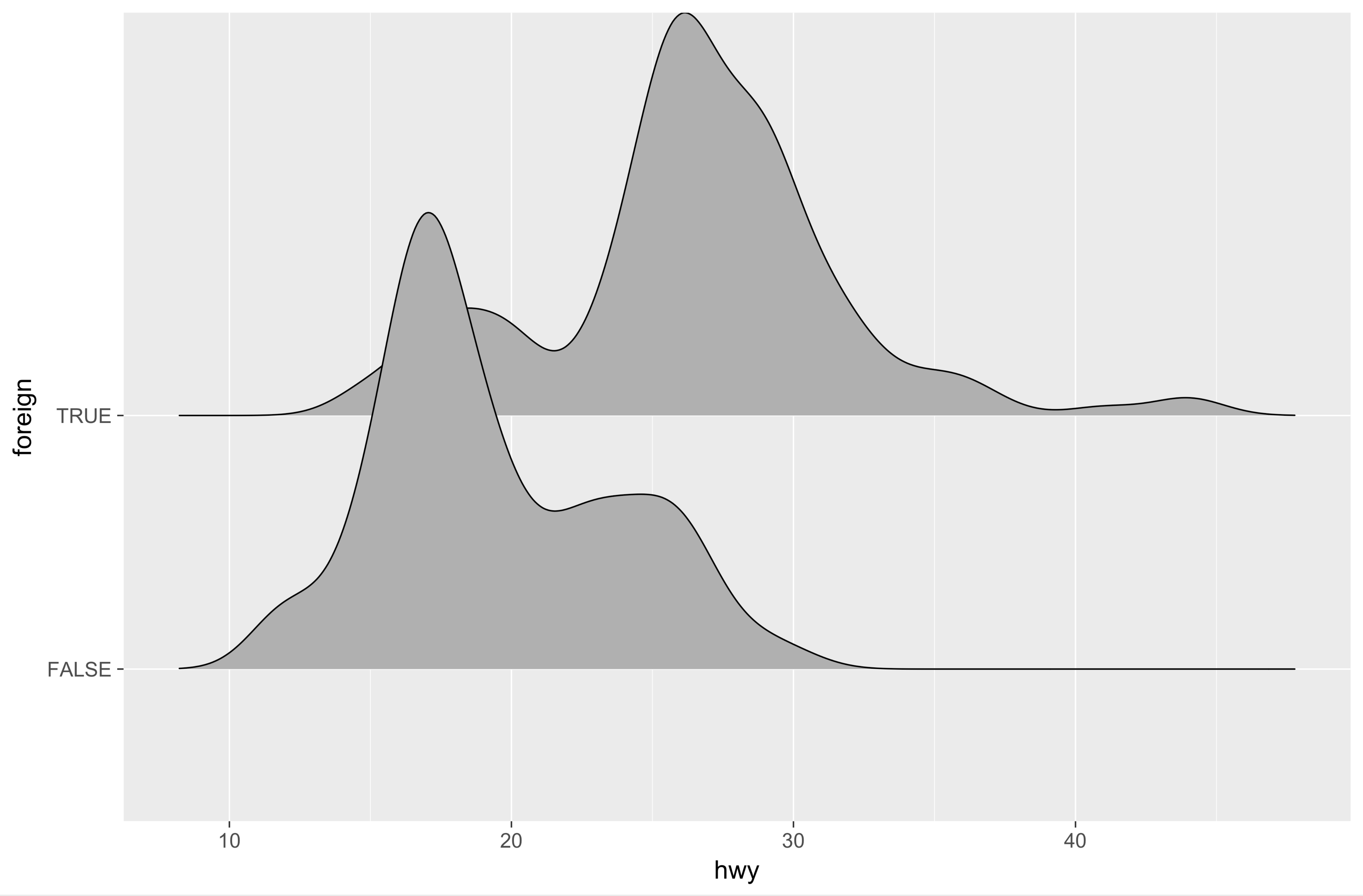
```
> ggplot(data = autoData,
    mapping = aes(x = foreign, y = hwy)) +
    geom_violin(mapping = aes(fill = foreign)) +
    stat_summary(fun.y = mean, geom = "point",
      size = 2)
```

The aesthetic mapping must appear in the initial `ggplot()` call.

VIOLIN PLOT WITH MEAN POINTS



RIDGE PLOT



RIDGE PLOT

`ggridges::geom_density_ridges(mapping = aes(aesthetic))`

Using the `hwy` and `foreign` variables created earlier from `ggplot2`'s `mpg` data:

```
> ggplot(data = autoData) +
    geom_density_ridges(mapping = aes(x = hwy,
      y = foreign))
```

The x and y variables are reversed here because of the way the ridge plot is oriented.

RIDGE PLOT

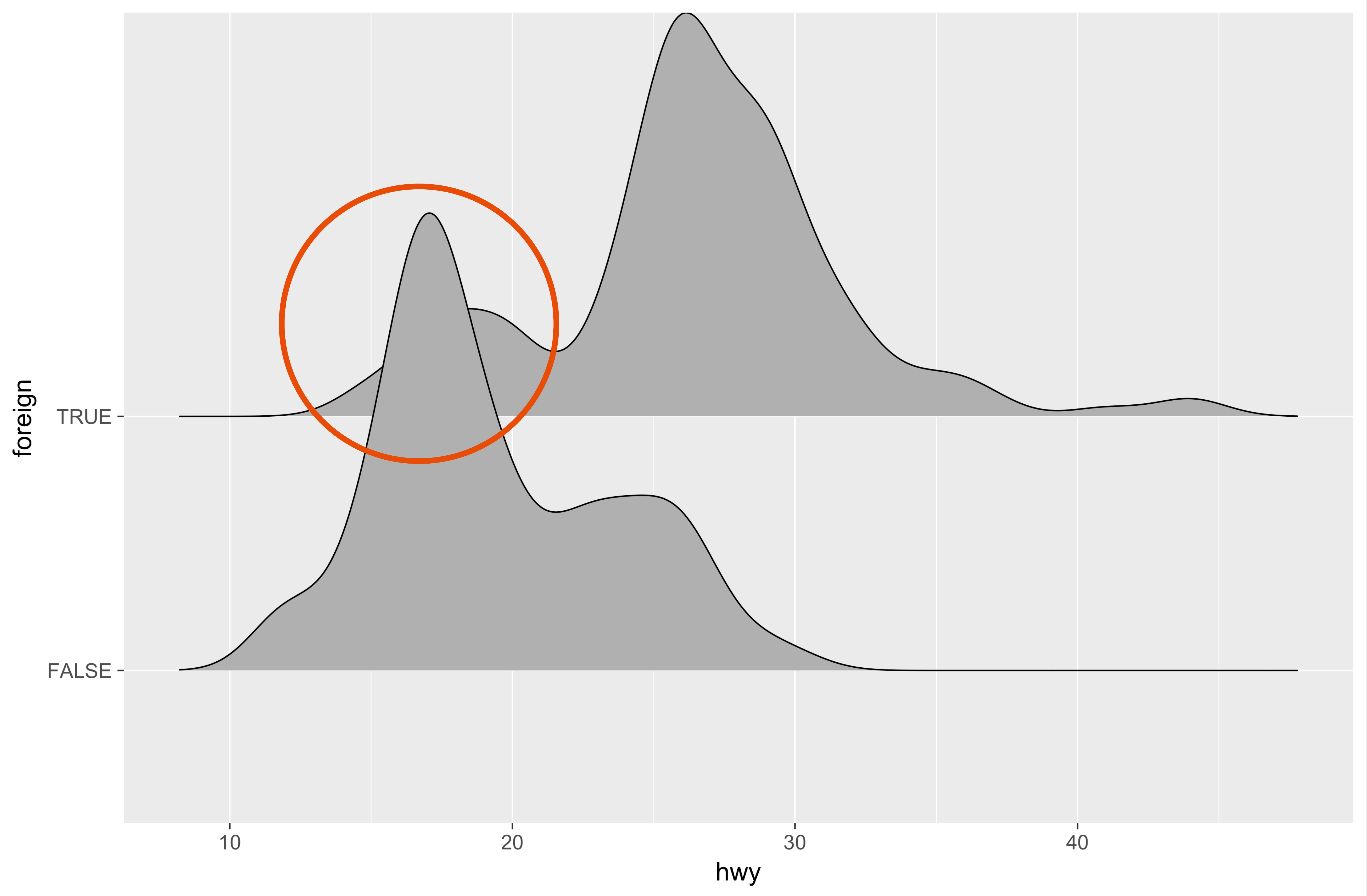
`ggridges::geom_density_ridges(mapping = aes(`*`aesthetic`*`))`

Using the `hwy` and `foreign` variables created earlier from `ggplot2`’s `mpg` data:

```
> ggplot(data = autoData) +
    geom_density_ridges(mapping = aes(x = hwy,
      y = foreign))
```

The design of these plots will obscure some aspects of your distributions.

RIDGE PLOT



RIDGE PLOT

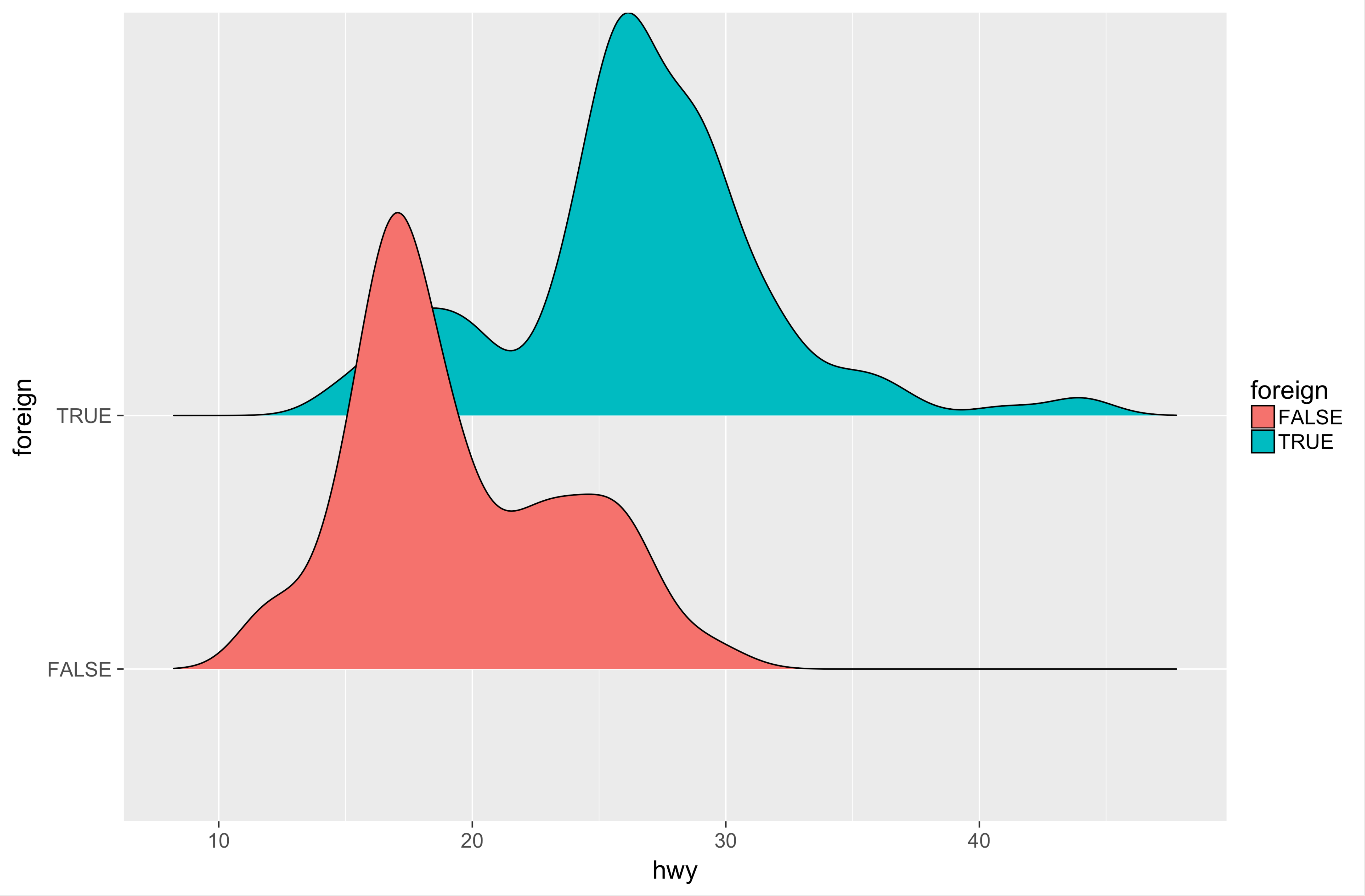
```
ggridges::geom_density_ridges(mapping = aes(aesthetic))
```

Using the `hwy` and `foreign` variables created earlier from `ggplot2`'s `mpg` data:

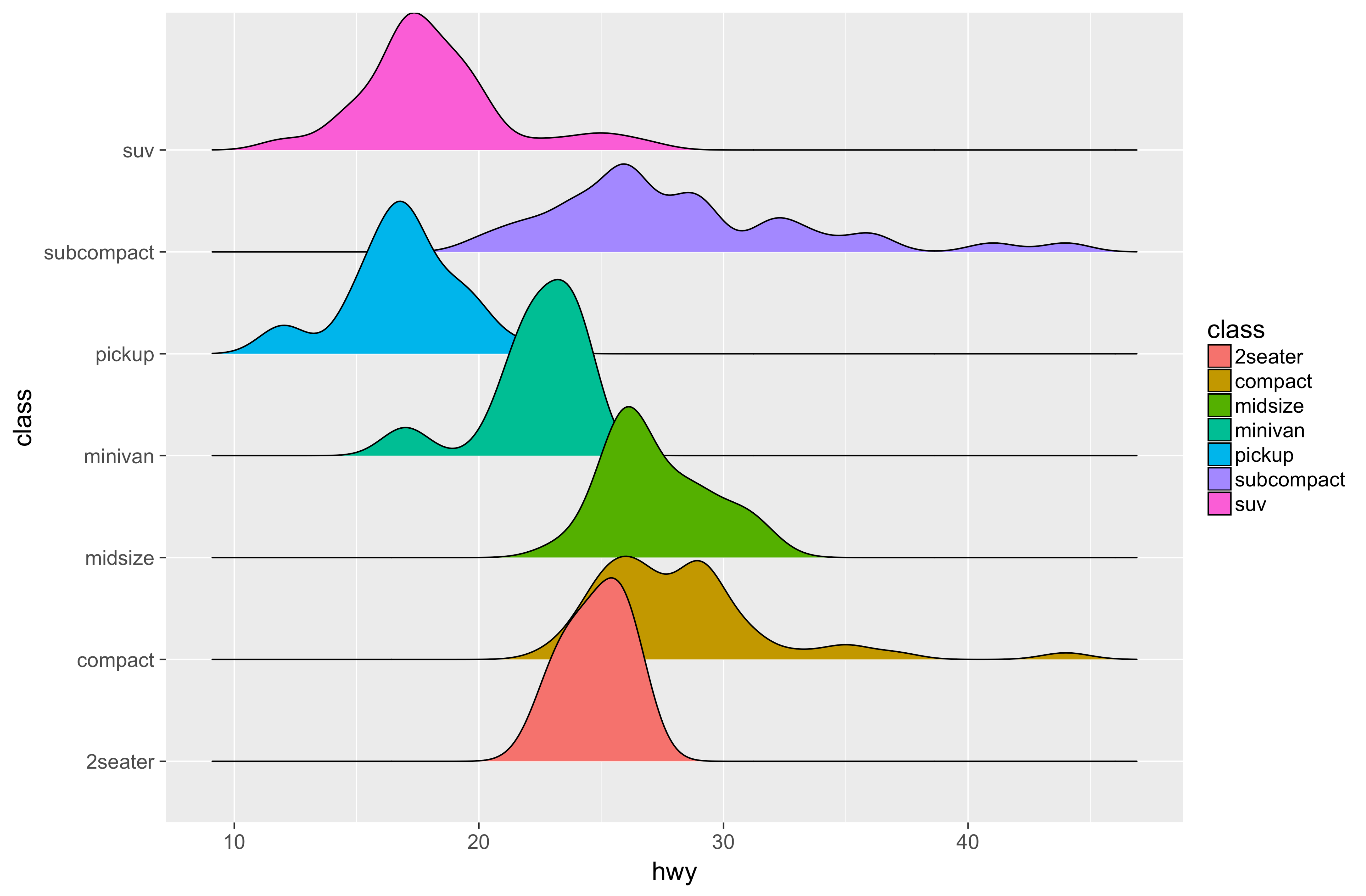
```
> ggplot(data = autoData) +
    geom_density_ridges(mapping = aes(x = hwy,
      y = foreign, fill = foreign))
```

The x and y variables are reversed here because of the way the ridge plot is oriented.

RIDGE PLOT



RIDGE PLOT



suv
subcompact
pickup
minivan
midsize
compact
2seater
class
10
20
30
40
hwy
class
2seater
compact
midsize
minivan
pickup
subcompact
suv

9 BACK MATTER

WHAT WE COVERED TODAY

2. Getting Started with LaTeX
3. Variance Testing
4. One and Two Samples
5. Dependent Samples
6. Effect Sizes
7. Sample Size Estimate
8. Plots for Mean Difference

REMINDERS

No video lectures next week!

Lab-08 & PS-06 due Monday, 10/30 by 4:15pm

Handout on papers in $\LaTeX$ will be posted on GitHub.

PS-07 will be a data cleaning puzzle, due Monday 10/30 as well

Midterm grade repots will be sent via GitHub

Lab-09 will be waived