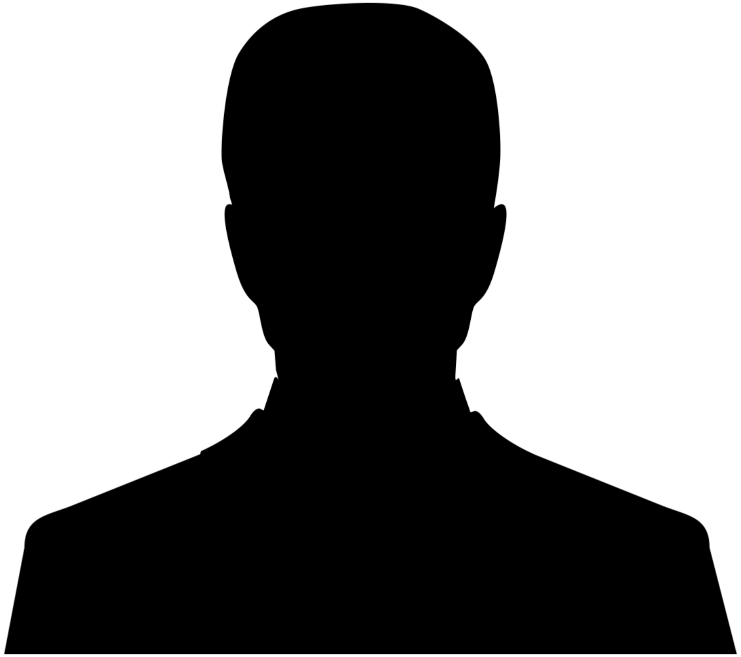
Data Visualization

Lesson Preview

- Learn how to use **base graphics**
- Learn how to use **base ggplot2**
- Understand **basic graph types** and when to use them

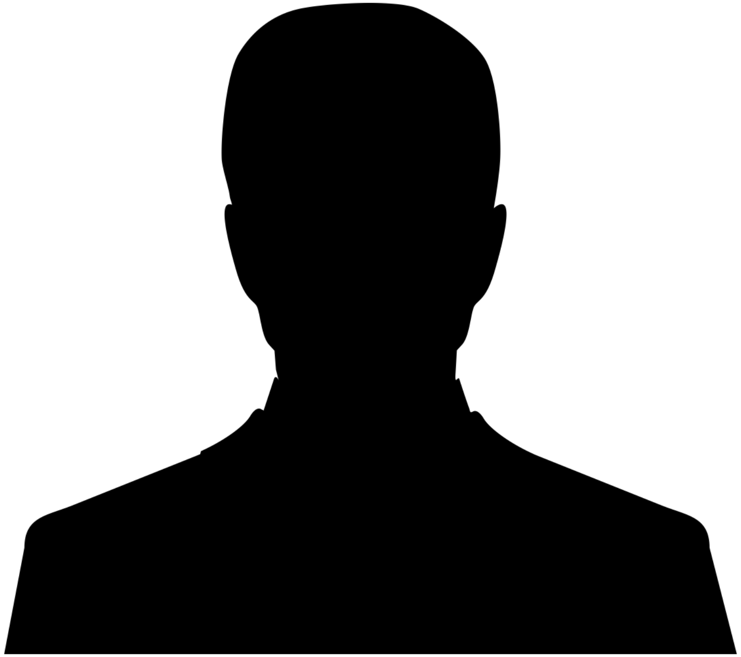


Data Visualization

Lesson Preview

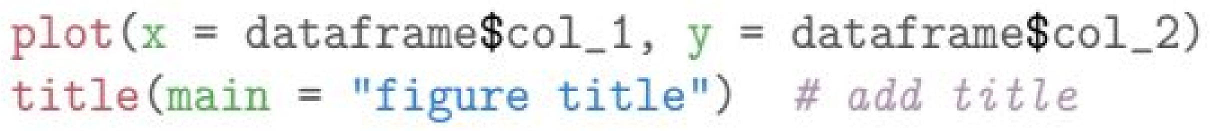
Four parts of this lesson:

- base graphics,
- ggplot2,
- datasets,
- basic graph types and case studies.



Base Graphics

Base graphics syntax: plot function followed by helper functions for annotating the graph.



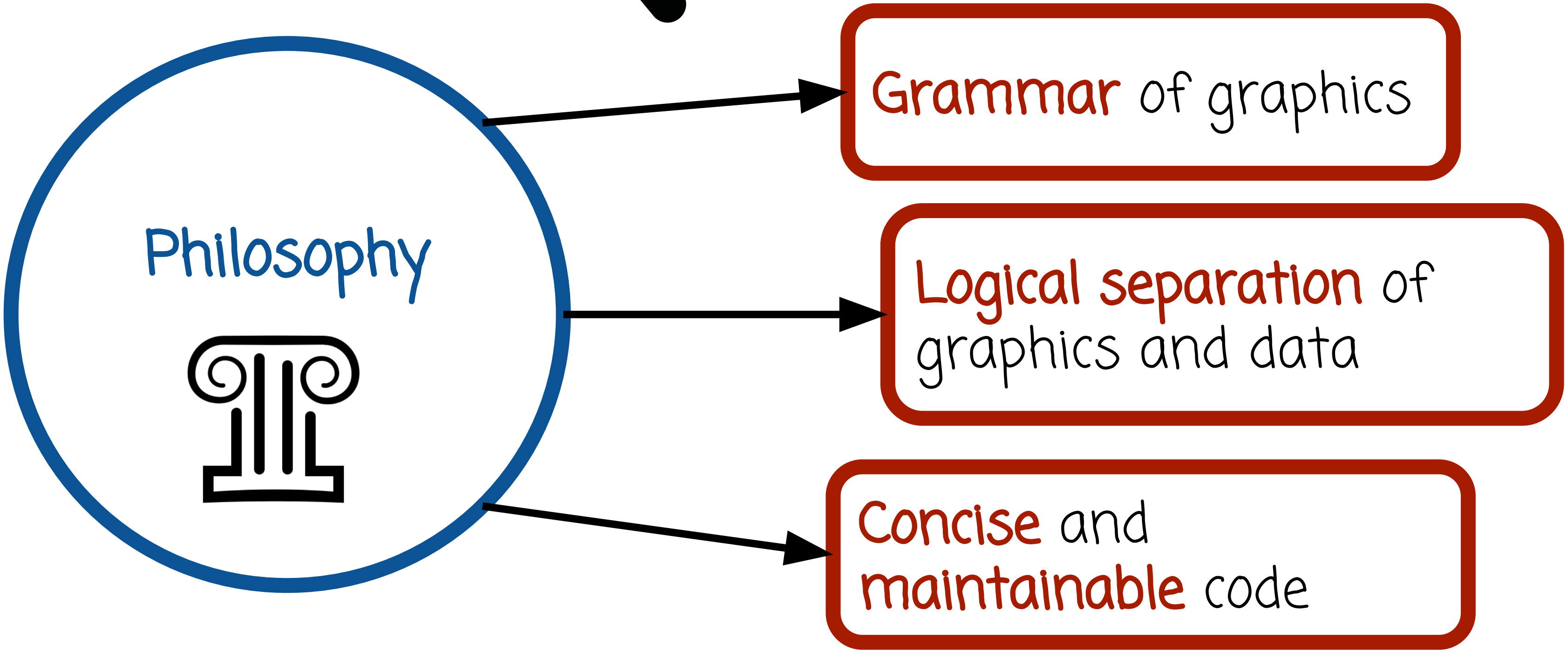
```
plot(x = dataframe$col_1, y = dataframe$col_2)
title(main = "figure title")  # add title
```

Base Graphics

Examples of low-level functions in the graphics package are:

- title adds or modifies labels of title and axes,
- grid adds a grid to the current figure,
- legend displays a legend connecting symbols, colors, and line-types to descriptive strings
- lines adds a line plot to an existing graph.

GGPLOT2



Philosophy

Grammar of graphics

Logical separation of graphics and data

Concise and maintainable code

GGPLOT2

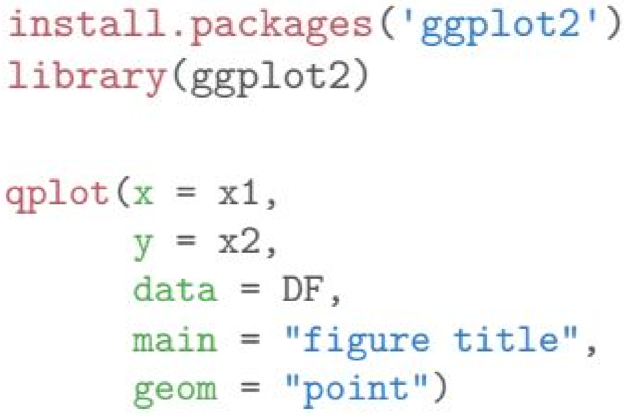
Option 1

- Use the **qplot function**.
- **Pass data frame** column names, data frame name, geometry, and graphing options

GGPLOT2

Option 1

Remember to install and load package using:



```
install.packages('ggplot2')
library(ggplot2)

qplot(x = x1,
      y = x2,
        data = DF,
        main = "figure title",
        geom = "point")
```

GGPLOT2

Option 2

- Use the ggplot function.
- Pass data frame, column names through aes function.

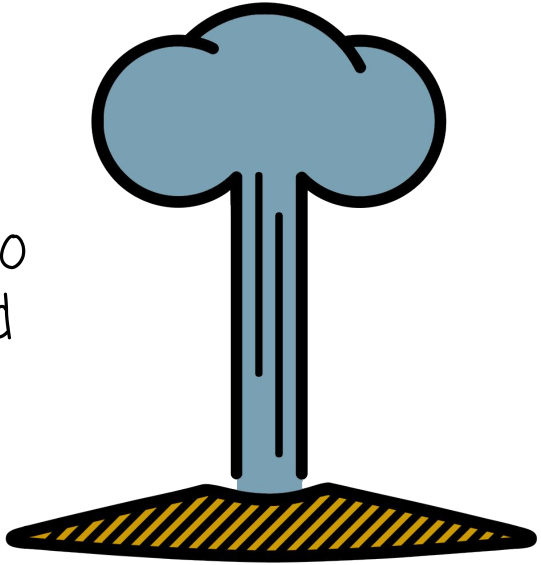
Compose function output with additional layers using + operator.

Function (and addition operator) returns an object that can be printed or saved for later.

Strip Plots

Dataset

faithful: eruption time and waiting time to next eruption (both in minutes) of the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.



```
names(faithful)
## [1] "eruptions" "waiting"
```

Strip Plots Quiz

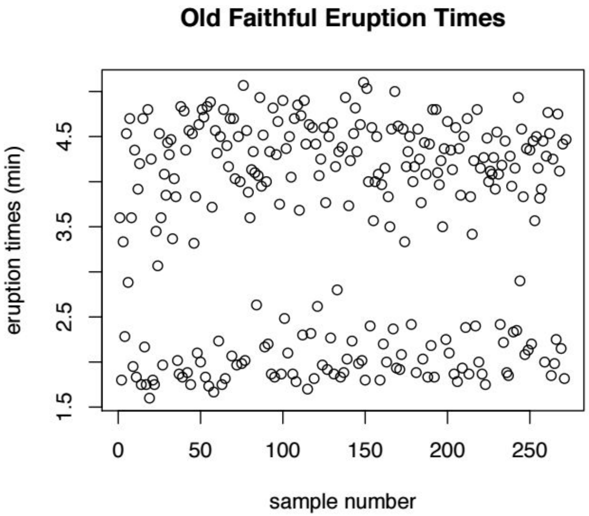
Strip plots graph one-dimensional numeric data as points in a two-dimensional space, with one coordinate corresponding to the index of the data point, and the other coordinate corresponding to its value. **Fill in the blanks to create a strip plot** given the following information:

```
names(faithful)
## [1] "eruptions" "waiting"
```

```
plot( faithful $ eruptions , xlab = “sample number”, ylab
= “eruption times (min)”, main = “Old Faithful Eruption
Times”)
```

Strip Plots Analysis Quiz

```
plot(faithful$eruptions, xlab = “sample number”, ylab =
“eruption times (min)”, main = “Old Faithful Eruption
Times”)
```

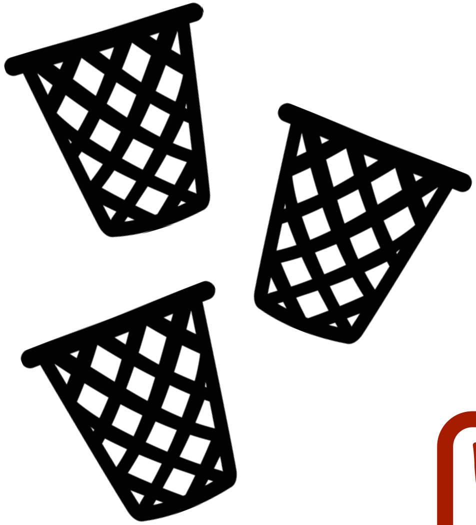


What can we conclude from the plot? **Select the true statements.**

☑ Old Faithful has two typical eruption times.

□ The order in which the data frame rows are stored is related to the eruption variable.

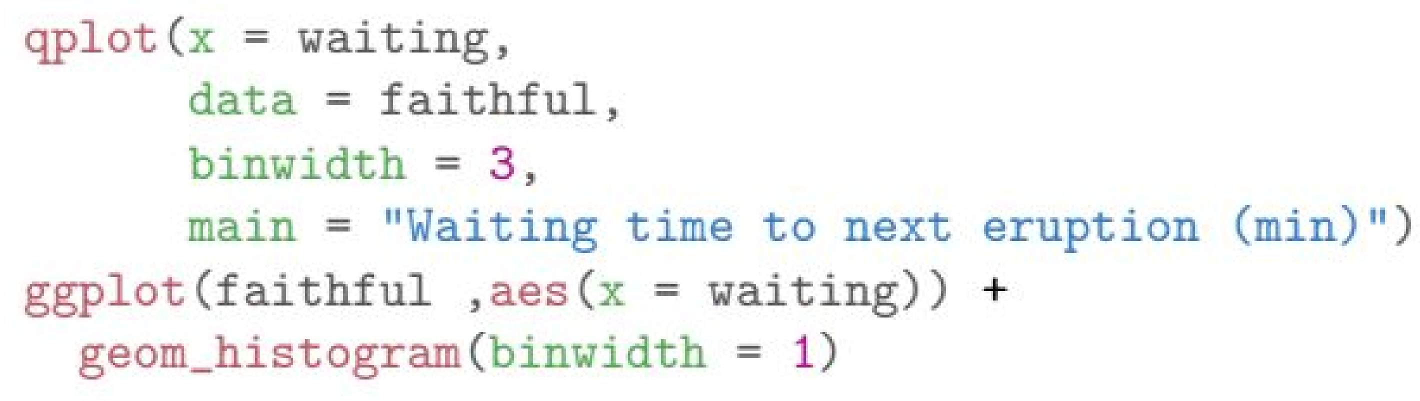
Histograms



Histograms **graph one-dimensional numeric data by dividing the range into bins** and counting number of occurrences in each bin.

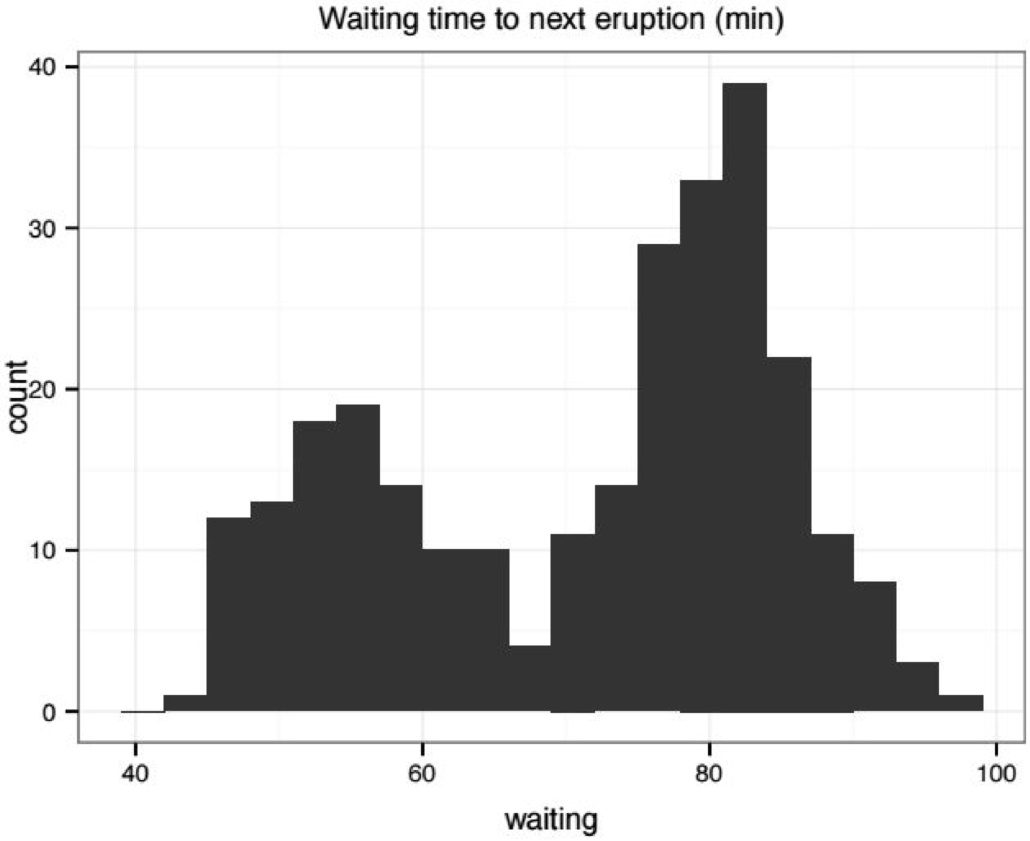
It is critical to set the bin width value correctly.

Histograms



```
qplot(x = waiting,
      data = faithful,
      binwidth = 3,
      main = "Waiting time to next eruption (min)")
ggplot(faithful ,aes(x = waiting)) +
  geom_histogram(binwidth = 1)
```

Histograms


Waiting time to next eruption (min)
count
40
30
20
10
0
40
60
80
100
waiting

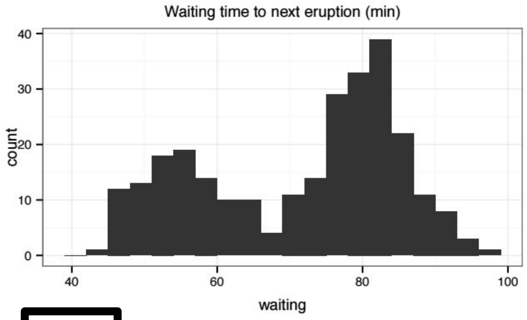
Histograms

y values can be **replaced with probability/frequency** using the following syntax

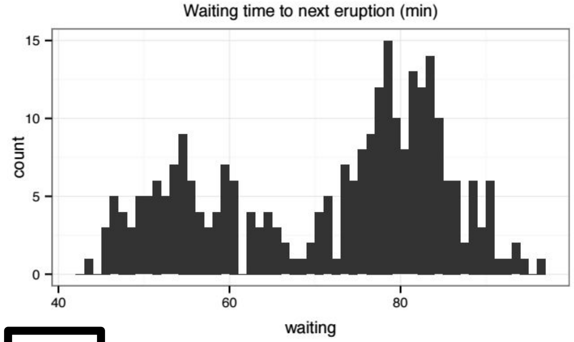
```
ggplot(faithful, aes(x =
waiting, y = ..density..)) +
geom_histogram(binwidth = #)
```

Histogram Quiz

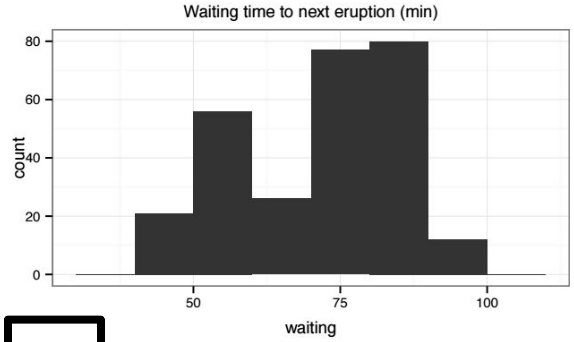
Given the following plots with different bin widths, **Match the description to the plot.**



C



B



C

A: good bin width - shows important signal in data (two modes) but not too much noise.

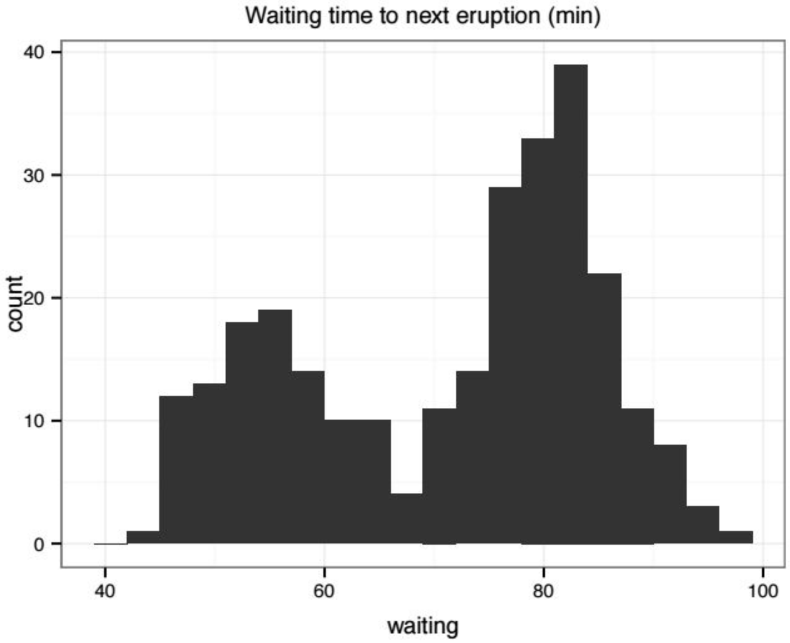
B: bin width is too small

C: bin width is too big

Histogram Quiz

Given the following plots with different bin widths, **Match the description to the plot.**

```
ggplot(faithful, aes(x
= waiting, y = ..
density..)) +
geom_histogram(binwidth
= 4)
```

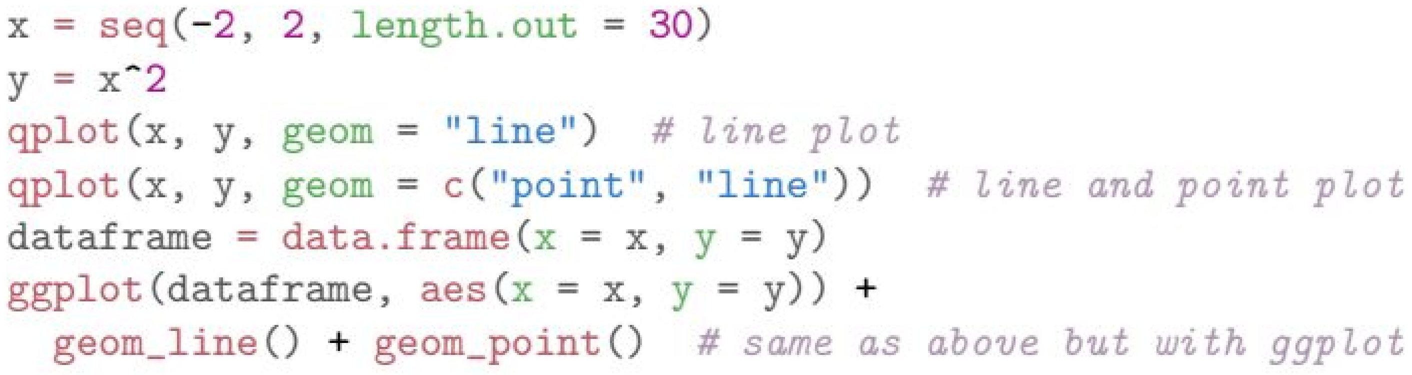


Line Plots

Line plot: a graph displaying a relation between x and y as a line in a Cartesian coordinate system.

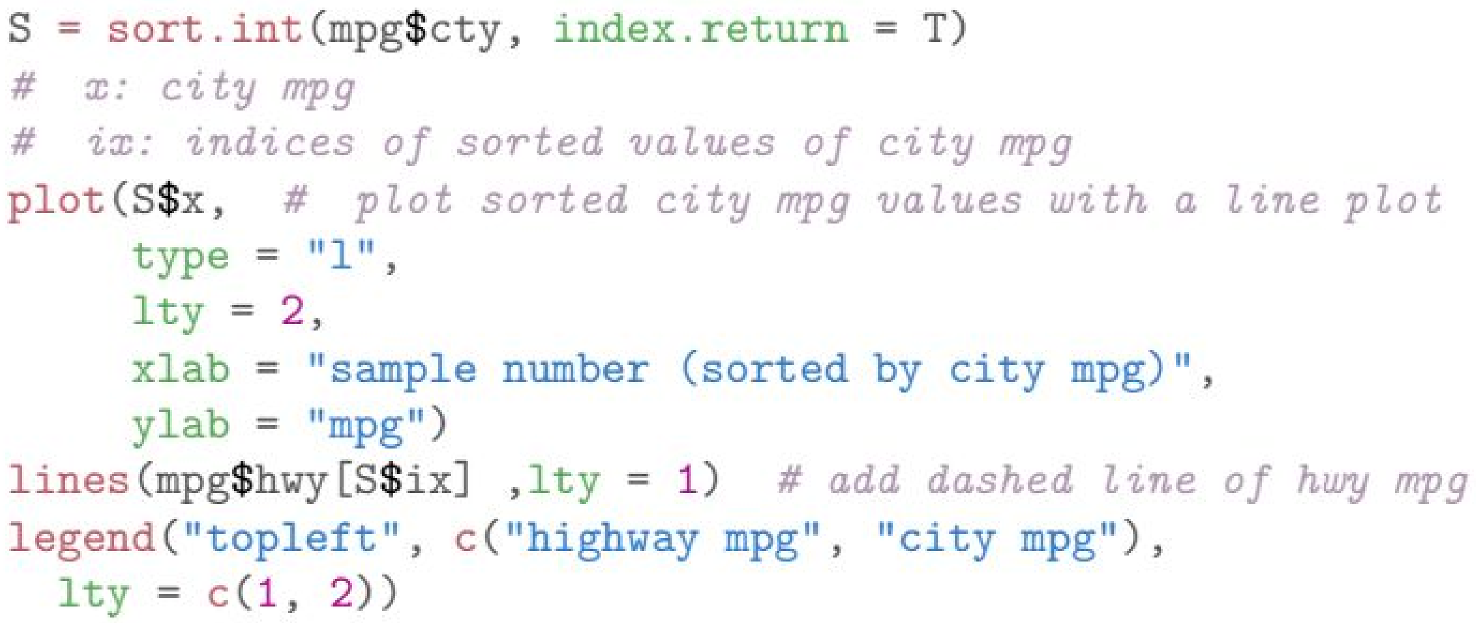
The relation may **correspond to an abstract mathematical function** or to a relation between two samples (for example, dataframe columns)

Line Plots



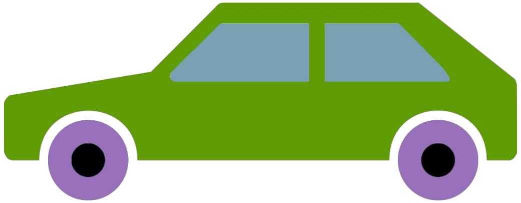
```
x = seq(-2, 2, length.out = 30)
y = x^2
qplot(x, y, geom = "line")  # line plot
qplot(x, y, geom = c("point", "line"))  # line and point plot
dataframe = data.frame(x = x, y = y)
ggplot(dataframe, aes(x = x, y = y)) +
  geom_line() + geom_point()  # same as above but with ggplot
```

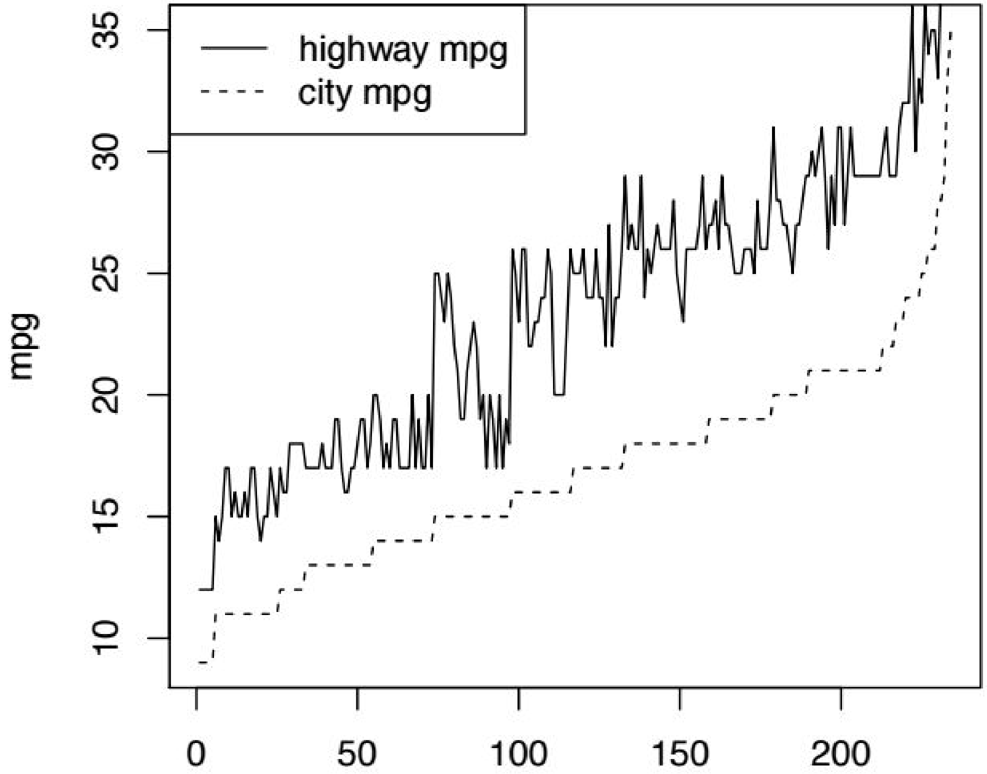
Line Plots



```
S = sort.int(mpg$cty, index.return = T)
#  x: city mpg
#  ix: indices of sorted values of city mpg
plot(S$x,  #  plot sorted city mpg values with a line plot
     type = "l",
     lty = 2,
     xlab = "sample number (sorted by city mpg)",
     ylab = "mpg")
lines(mpg$hwy[S$ix] ,lty = 1)  # add dashed line of hwy mpg
legend("topleft", c("highway mpg", "city mpg"),
  lty = c(1, 2))
```

Line Plots





Smoothed Histograms

Denoting n values by $x^{(1)}, \dots, x^{(n)}$, the smoothed histogram is the following function f_h: $\mathbb{R} \to \mathbb{R}_+$

$$f_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x^{(i)})$$

Where the kernel function K_h: $\mathbb{R} \to \mathbb{R}$ typically achieves it maximum at 0, and decreases as $|x-x^{(i)}|$ increases. We also assume that the kernel function integrates to one $\int K_h(x)dx=1$ and satisfies the relation

$$K_h(r) = h^{-1}K_1(r/h).$$

We refer to K_1 as the base form of the kernel and denote it as K.

Smoothed Histograms

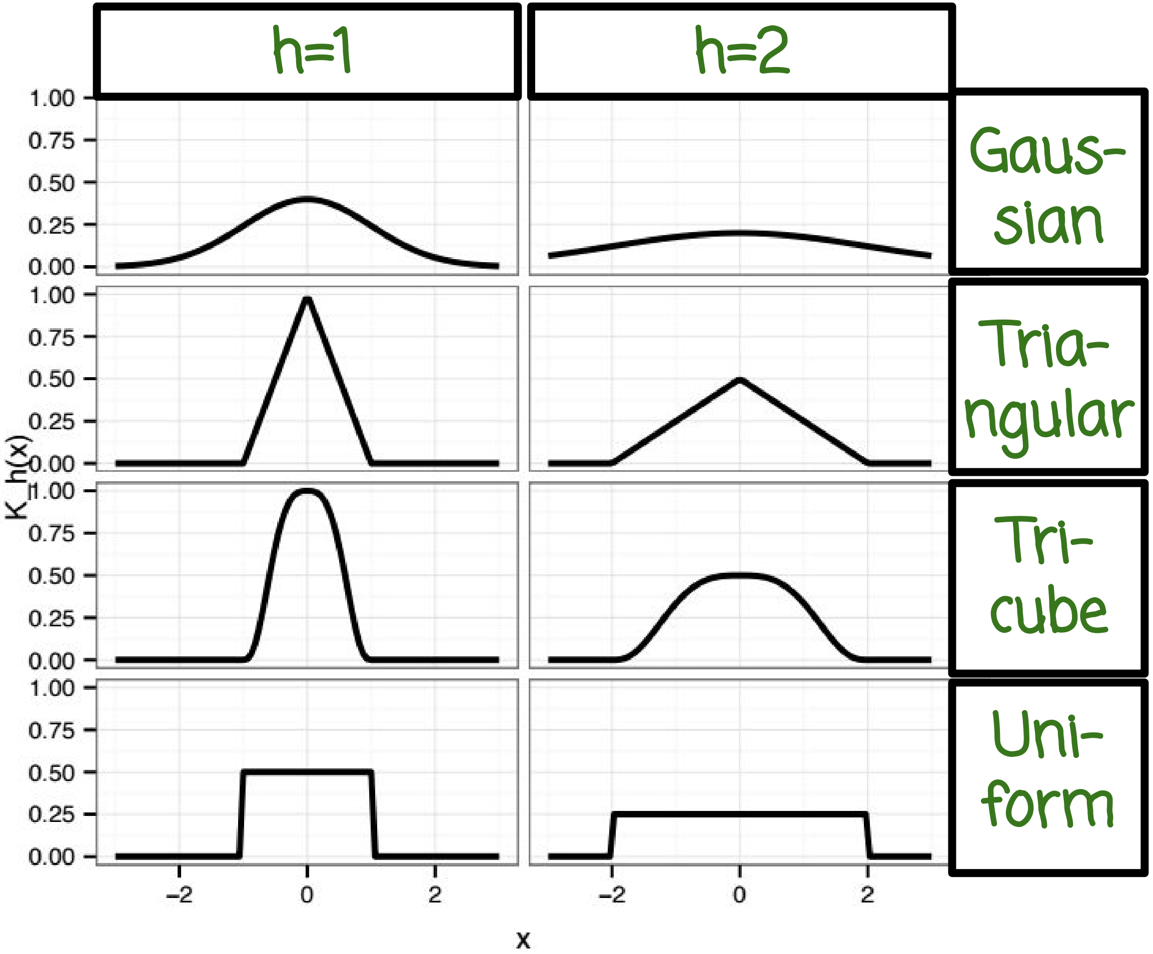
Four popular kernel choices are the **tricube, triangular, uniform, and Gaussian kernels**, defines as $K_h(r)=x^{-1}K(r/h)$ where the $K(\cdot)$ functions are respectively

$$
\begin{aligned}
K(r) &= (1-|r|^3)^3 \cdot 1_{\{|r|<1\}} && \text{(Tricube)}\\
K(r) &= (1-|r|) \cdot 1_{\{|r|<1\}} && \text{(Triangular)}\\
K(r) &= 2^{-1} \cdot 1_{\{|r|<1\}} && \text{(Uniform)}\\
K(r) &= \exp(-x^2/2)/\sqrt{2\pi} && \text{(Gaussian).}
\end{aligned}
$$

As h increases the kernel functions K_h become wider.

Smoothed Histogram Quiz

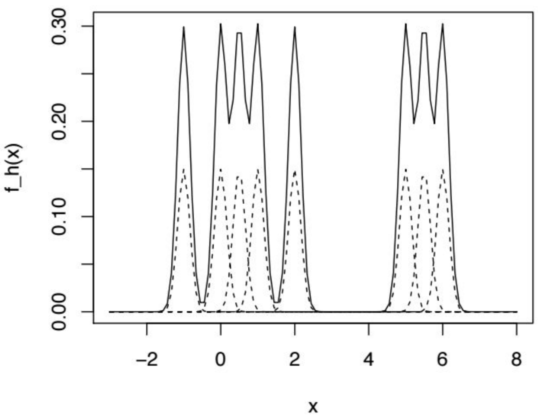
Fill in the blanks with the size of h (1 or 2) and the name of each kernel: Tricube, Triangular, Uniform, Gaussian



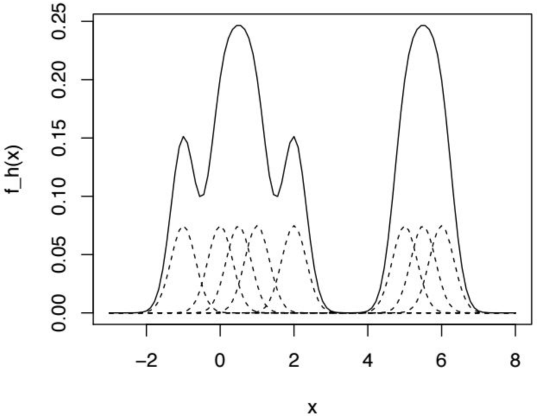
Varied h Quiz

Select the 'h' that was used to generate each plot. The choices are: **⅙, ⅓, 1.**

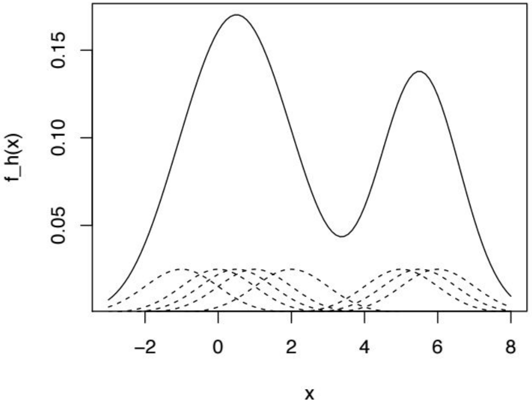
h= ⅙



h= ⅓



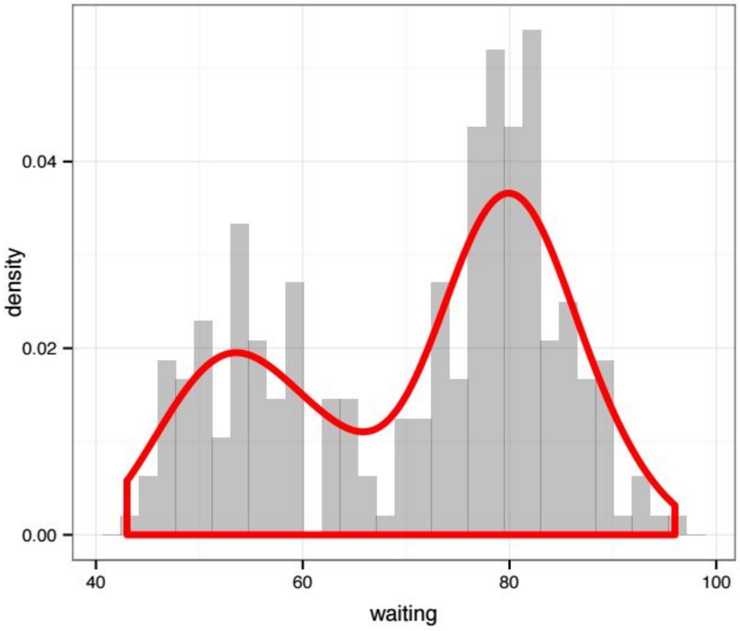
h= 1



Old Faithful Histogram Quiz

Using the Old Faithful dataset, write the command the will create the histogram.

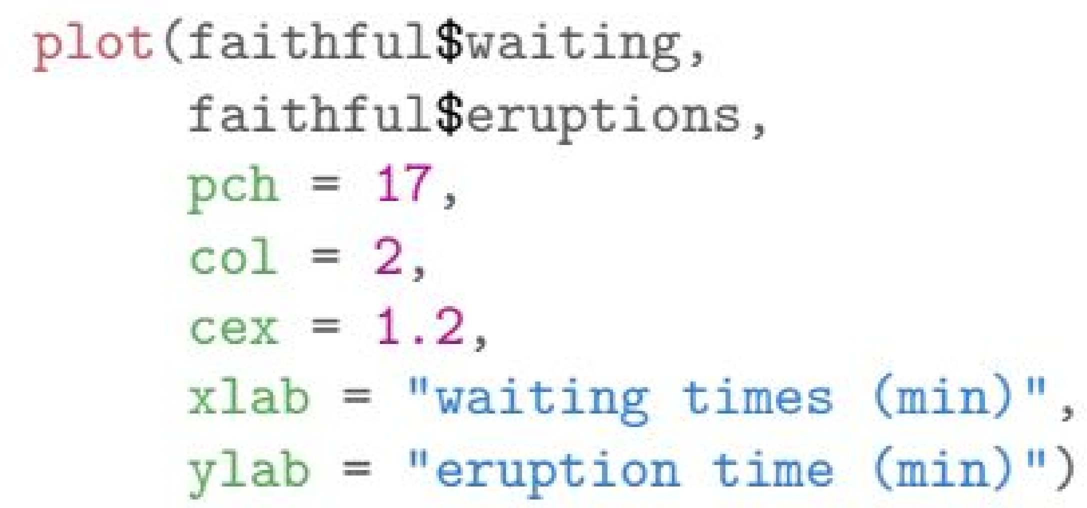
```
ggplot( faithful , aes
(x = waiting, y = ..
density..)) +
geom_histogram(alpha =
 0.3 ) + geom_density
(size = 1.5 , color
= " red ")
```



Scatter Plot

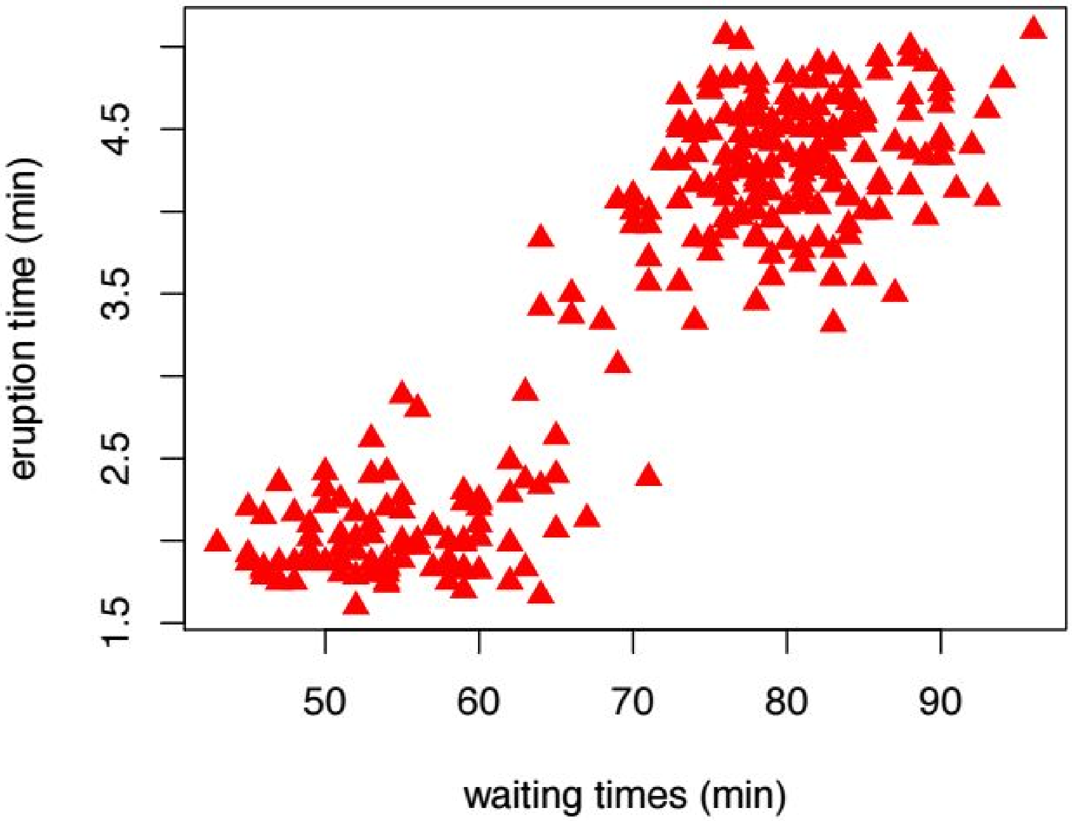
A scatter plot graphs the relationships between two numeric variables. It graphs each pair of variables as a point in a two dimensional space whose coordinates are the corresponding x, y values.

Scatter Plot



```
plot(faithful$waiting,
     faithful$eruptions,
     pch = 17,
     col = 2,
     cex = 1.2,
     xlab = "waiting times (min)",
     ylab = "eruption time (min)")
```

Scatter Plot



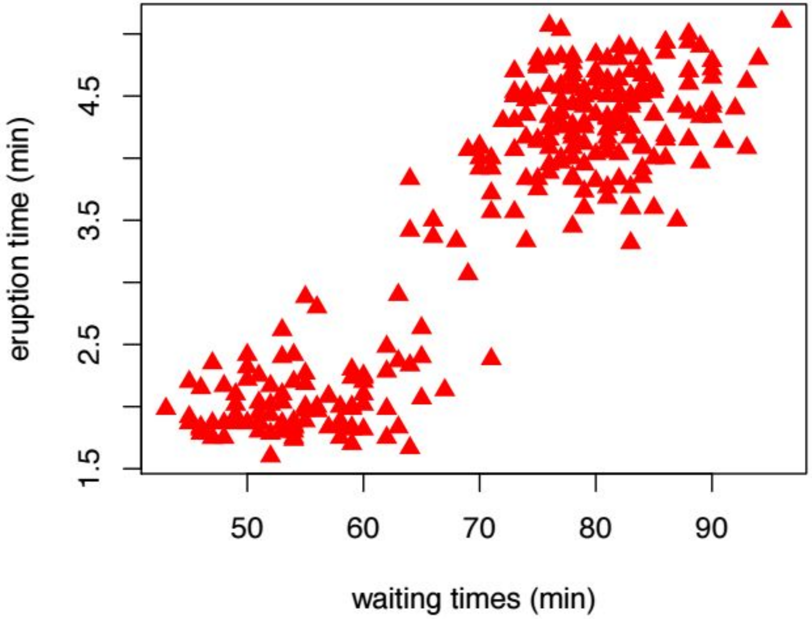
Scatter Plot Quiz

Based on the scatter plot for Old Faithful **answer the following questions.**

Number of distinct cases of eruption duration = $\boxed{2}$.

Short wait times generally result in $\boxed{\text{short}}$ eruptions.

Long wait times generally result in $\boxed{\text{long}}$ eruptions.



Variable Relationships and Scatter Plots

The relationship between two numeric variables and a categorical variable **can be graphed using a scatter plot where the categorical variable controls** the size, color, or shape of the markers.

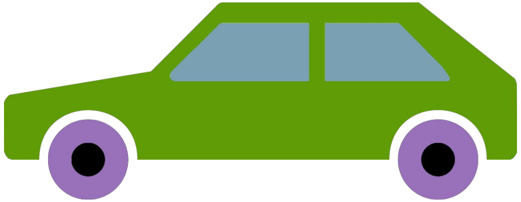
Line Plots

mtcars: model name, weight, horsepower, fuel efficiency, and transmission type of cars from 1974 Motor Trend magazine.

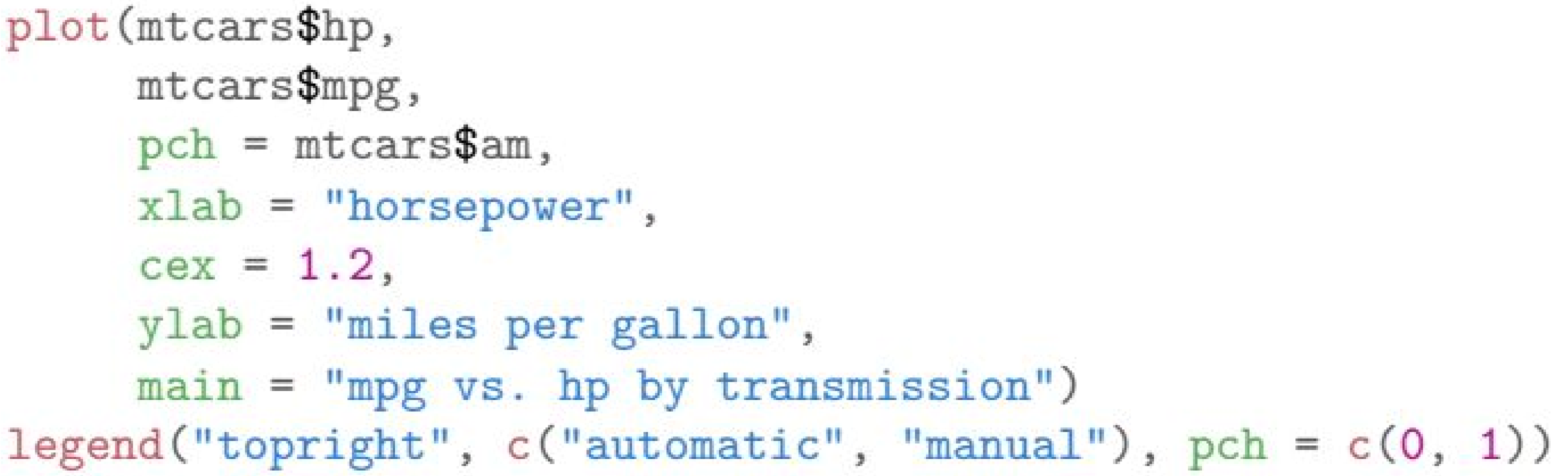
```
names(mtcars)
```

```
## [1] "mpg" "cyl" "disp" "hp" "drat" "wt" "qsec"
"vs" "am" "gear"

## [11] "carb"
```

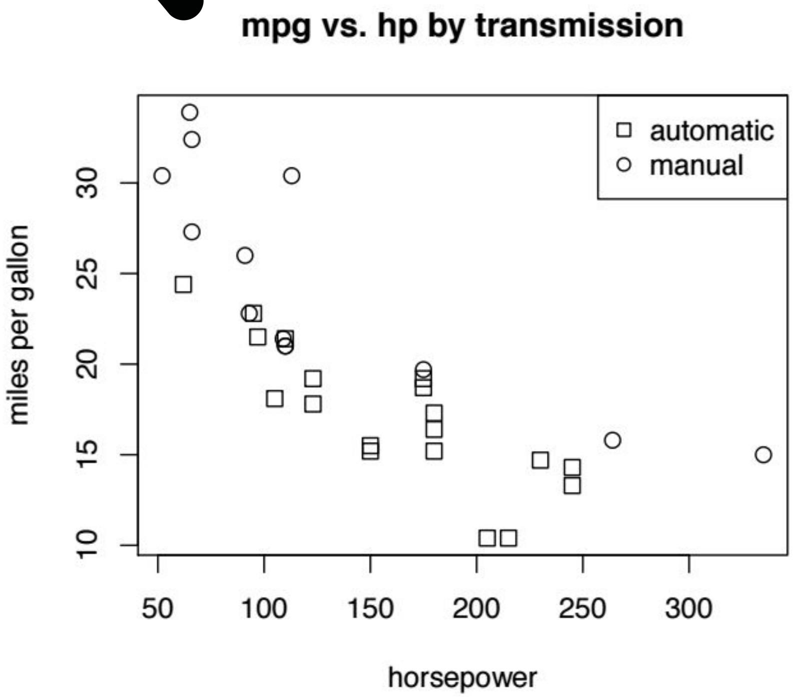


Variable Relationships and Scatter Plots



```
plot(mtcars$hp,
     mtcars$mpg,
     pch = mtcars$am,
     xlab = "horsepower",
     cex = 1.2,
     ylab = "miles per gallon",
     main = "mpg vs. hp by transmission")
legend("topright", c("automatic", "manual"), pch = c(0, 1))
```

Variable Relationships and Scatter Plots

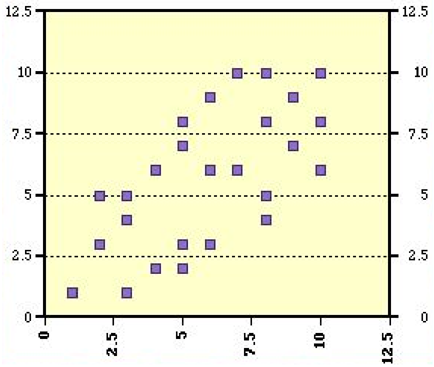


- There is an **inverse relationship** between horsepower and mpg.
- For a given horsepower amount, **manual transmission cars are generally more fuel efficient.**
- Cars with the **highest horsepower tend to be manual**.

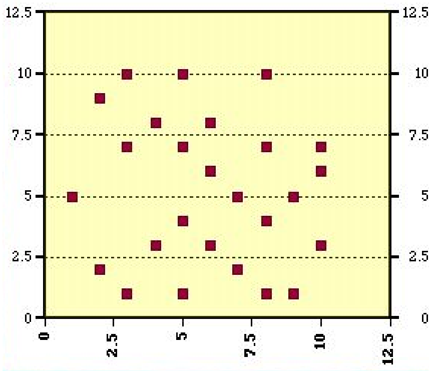
Scatter Plot Correlation Quiz

Determine the correlation for each plot: high negative, low negative, low positive, and no correlation.

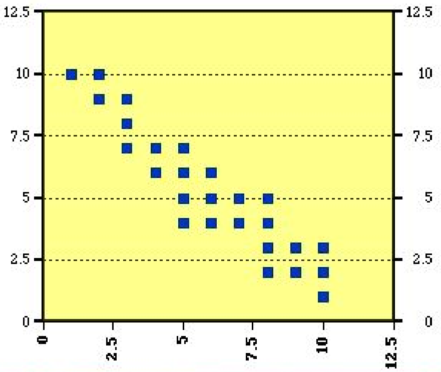
Low positive



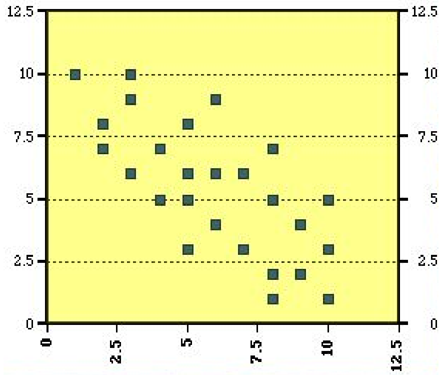
No correlation



High negative



Low negative



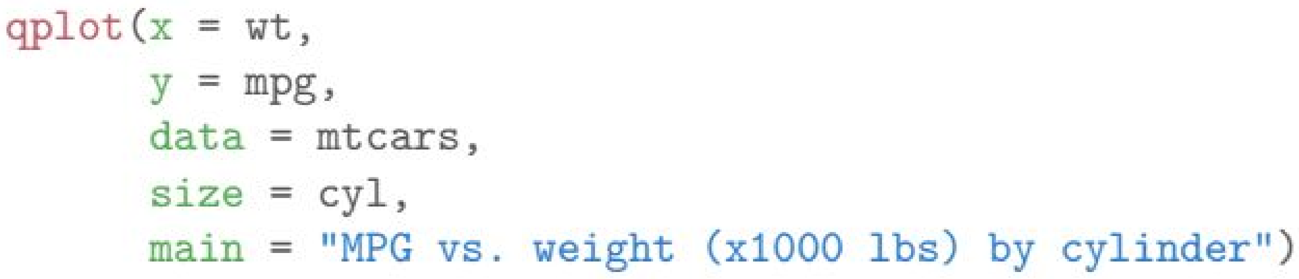
Multivariable Scatter Plots

mpg: fuel economy and other car attributes from http://fueleconomy.gov (similar to mtcars but larger and newer).

```
names(mpg)
## [1] "manufacturer" "model" "displ" "year"
## [5] "cyl" "trans" "drv" "cty"
## [9] "hwy" "fl" "class"
```

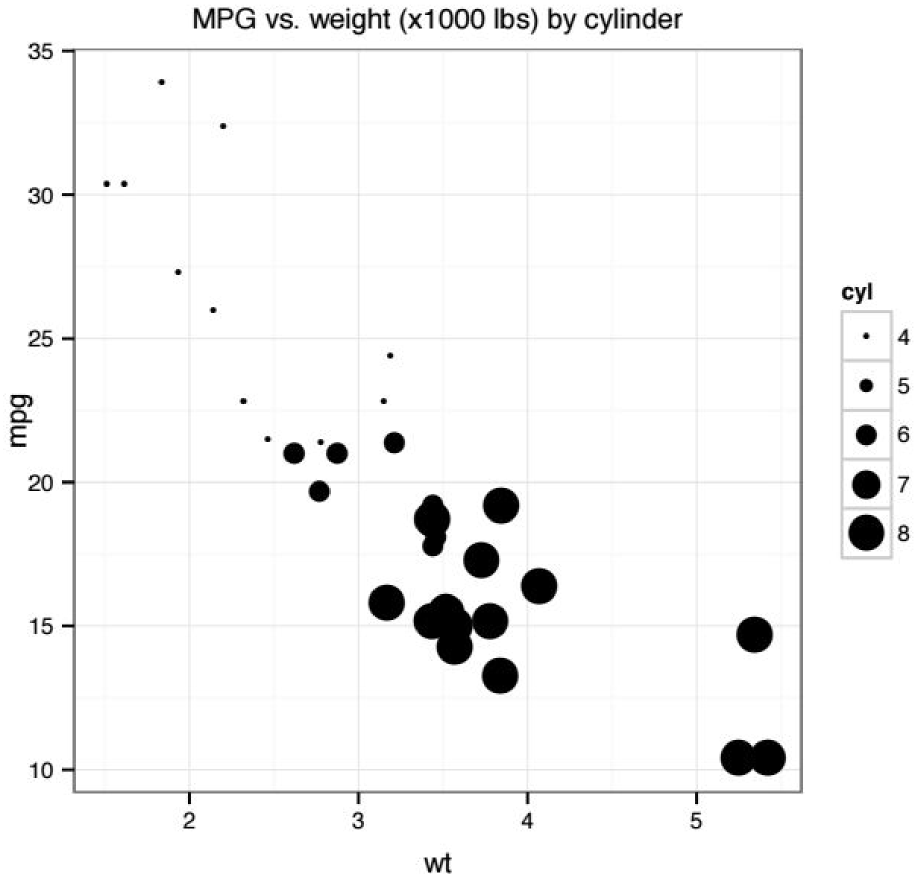
Multivariable Scatter Plots

Changing marker size in a scatter plot

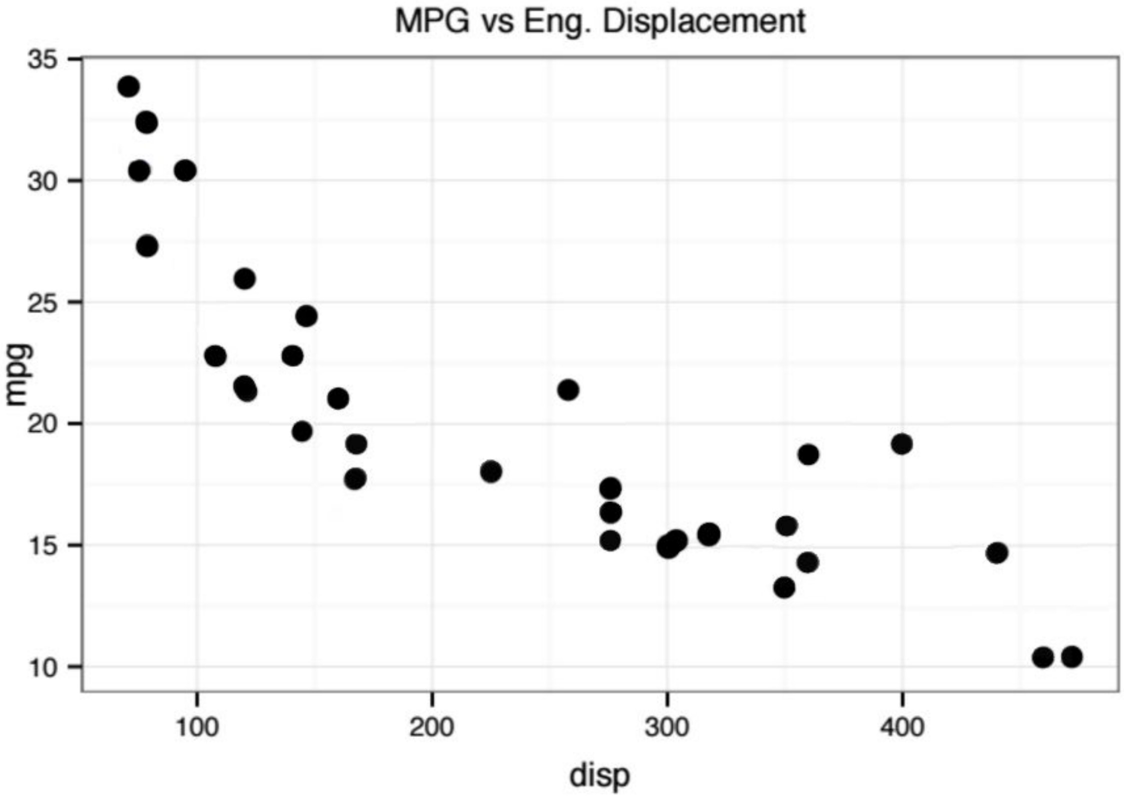


```
qplot(x = wt,
      y = mpg,
      data = mtcars,
      size = cyl,
      main = "MPG vs. weight (x1000 lbs) by cylinder")
```

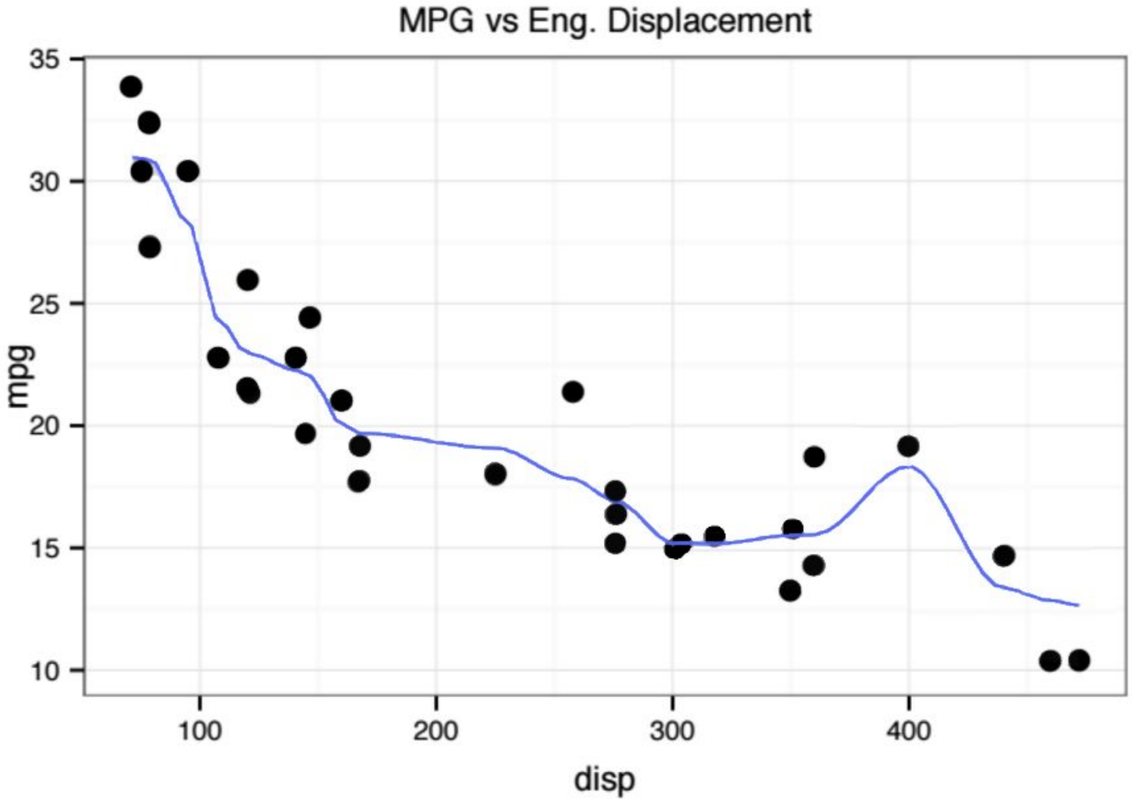
Multivariable Scatter Plots



Noisy Data


MPG vs Eng. Displacement
35
30
25
20
15
10
mpg
100
200
300
400
disp

Noisy Data



MPG vs Eng. Displacement
mpg
35
30
25
20
15
10
100
200
300
400
disp

Noisy Data

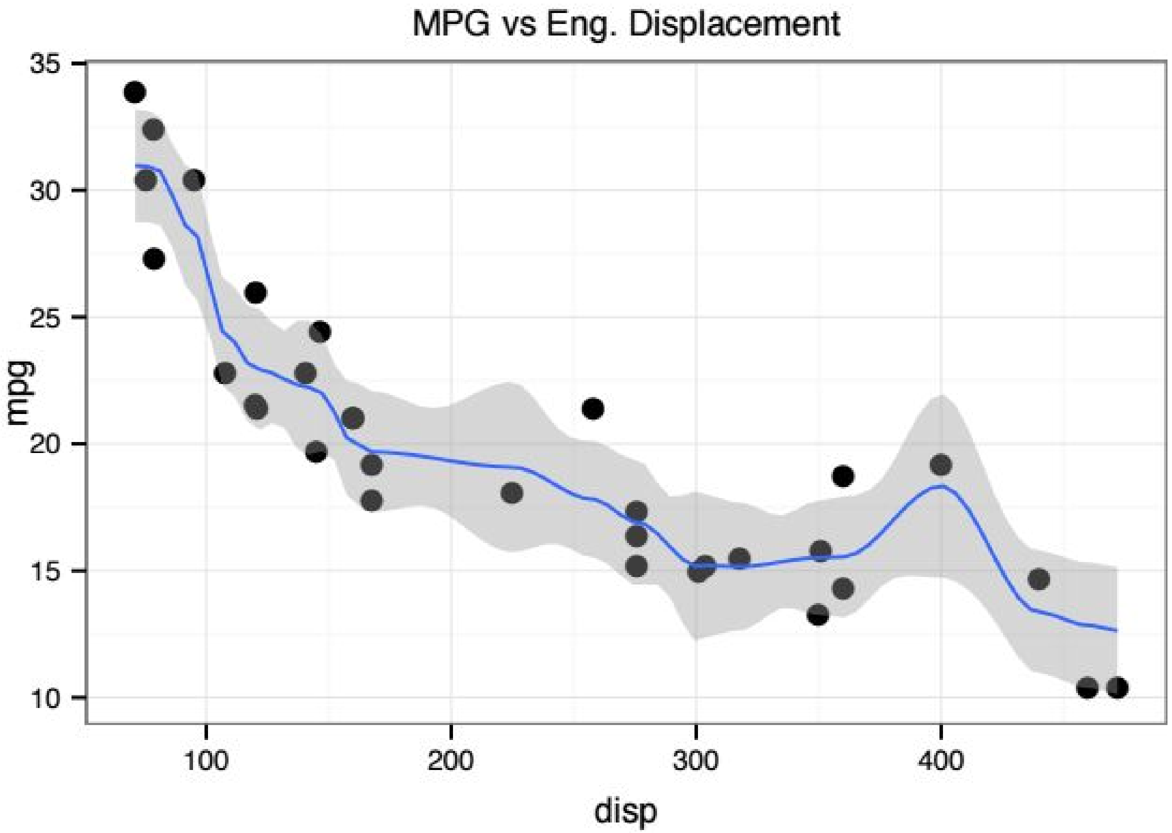
Adding a smooth line curve y_s, which a weighted average of the original data. $(y^{(i)}, x^{(i)})$ $i = 1, \ldots, n$:

$$ys(x) = \sum_{i=1}^{n} \frac{K_h(x - x^{(i)})}{\sum_{i=1}^{n} K_h(x - x^{(i)})} y^{(i)}.$$

Where the K_h functions above are the kernel functions described earlier.

- $y_s(x)$ = a weighted average of $y^{(i)}$ values
- The denominator ensures the weights sum to 1.

Noisy Data

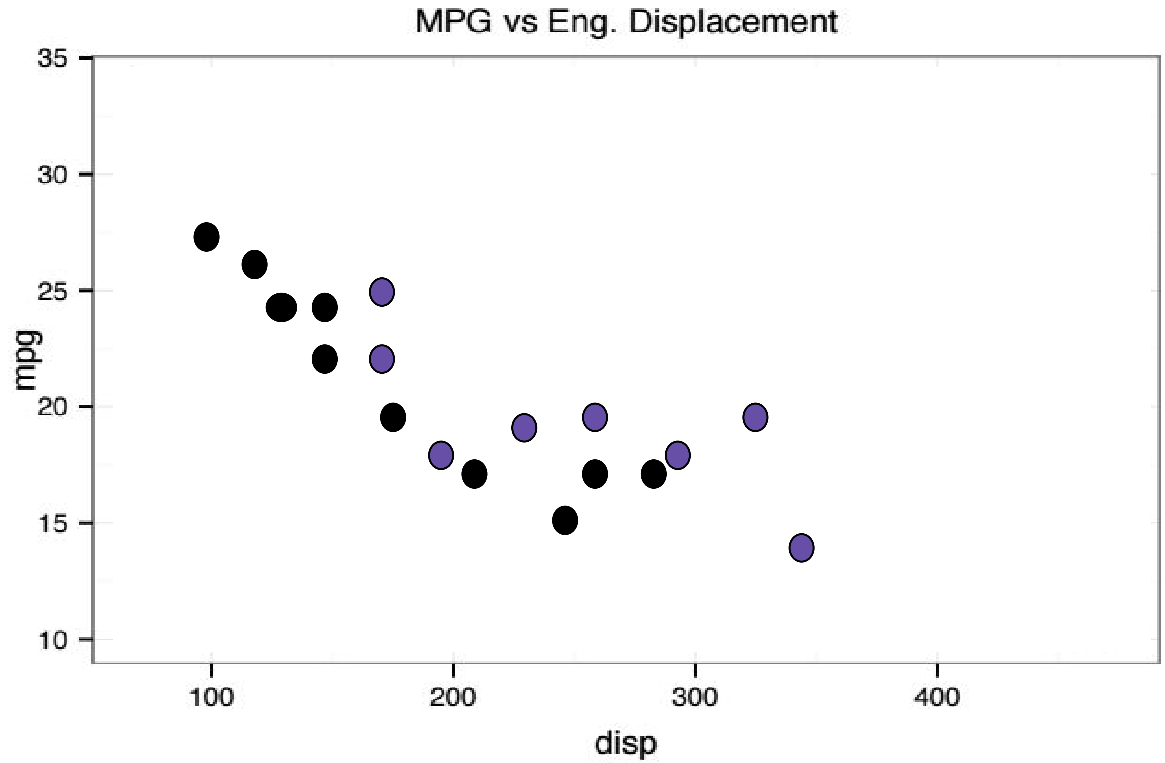

MPG vs Eng. Displacement
mpg
35
30
25
20
15
10
disp
100
200
300
400

Noisy Data Quiz

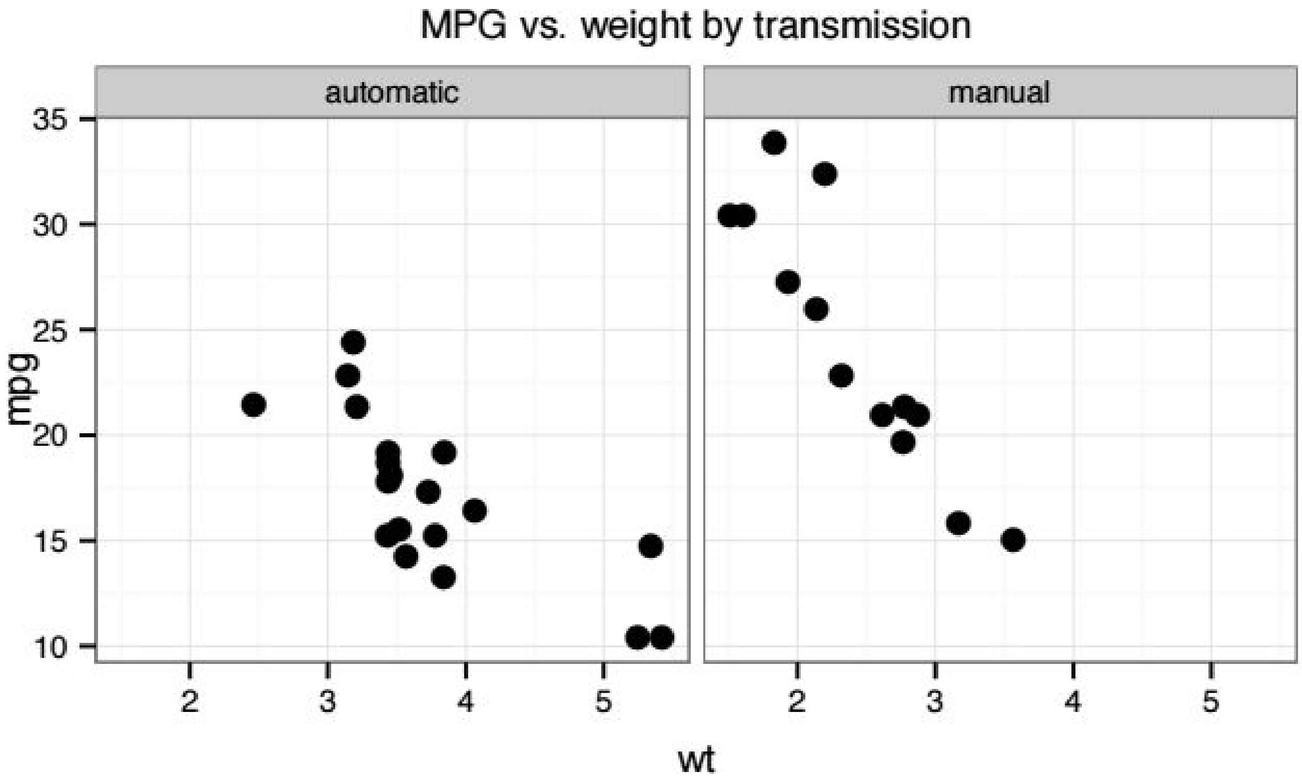
Fill in the blanks to add a line and standard errors to the MPG vs Displacement plot.

```
qplot( disp , mpg , data = mtcars , main =
" MPG vs. Eng. Displacement ") + stat_smooth(method =
" loess ", degree = 0 , span = 0.2 , se =
 TRUE )
```

Facets


MPG vs Eng. Displacement
mpg
35
30
25
20
15
10
100
200
300
400
disp

Facets


MPG vs. weight by transmission
automatic
manual
mpg
35
30
25
20
15
10
wt
2
3
4
5
2
3
4
5

Facets

The argument facets in qplot or ggplot **takes a formula a~b where a and b specify the variables** according to which way the column and rows are organized.

Note

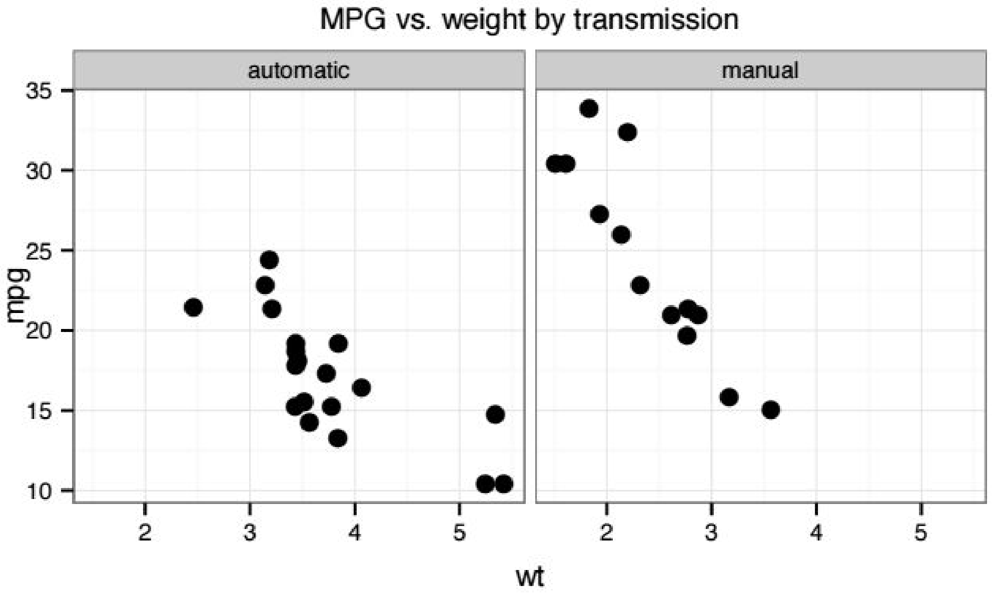
Before the following quizzes we need to modify the mtcars dataframe to have new columns with more appropriate names for better axes labeling:

```
mtcars$amf[mtcars$am==0] = 'automatic'
mtcars$amf[mtcars$am==1] = 'manual'
mtcars$vsf[mtcars$vs==0] = 'flat'
mtcars$vsf[mtcars$vs==1] = 'V-shape'
```

Facets Quiz

Setup the commands to create the graphs shown.

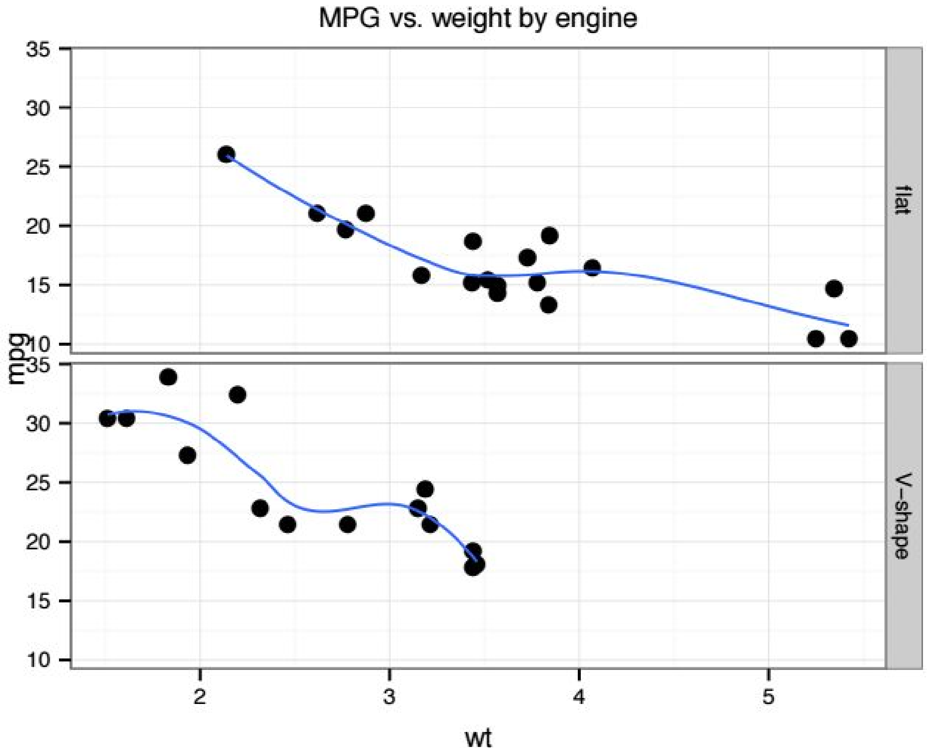
```
qplot(x= wt ,
y = mpg ,
facets = .~amf ,
data = mtcars ,
main = "MPG vs. weight by transmission")
```



Facets Quiz 2

Write a command to create the graphs shown.

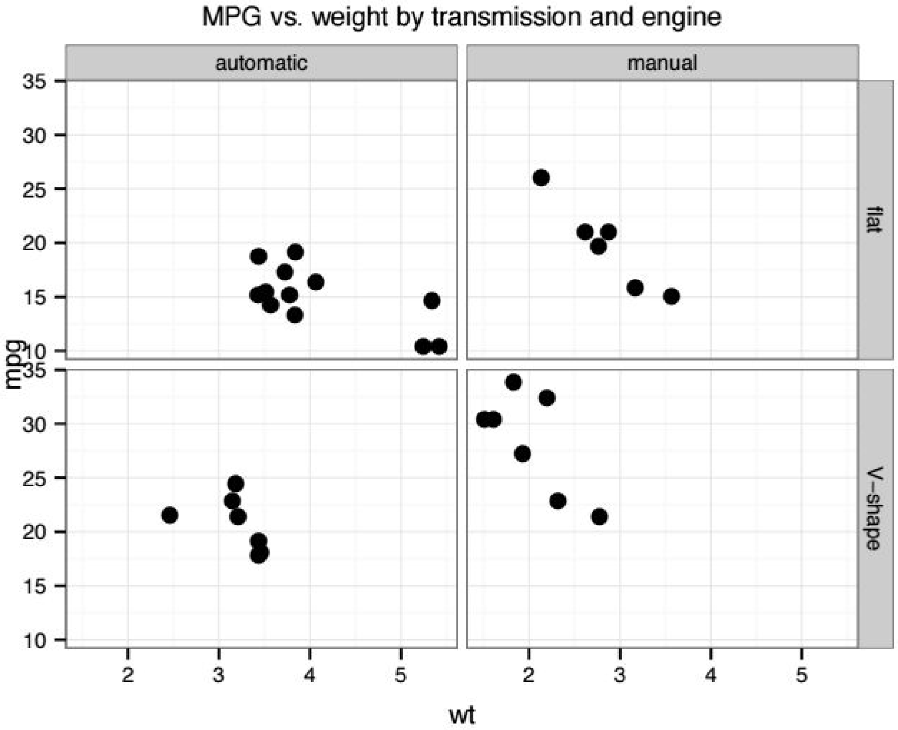
```
qplot(x= wt ,
y = mpg ,
facets = vsf~. ,
data = mtcars ,
main = " "MPG vs. ..." ")
```



Facets Quiz 3

Write a command to create the graphs shown.

```
qplot(x = wt ,
y= mpg ,
facets= vsf~amf ,
data= mtcars ,
main = “ "MPG vs. …" ”)
```



Facets Data Inference Quiz

From graphs created from the mtcars data, which type of car has the following characteristics?

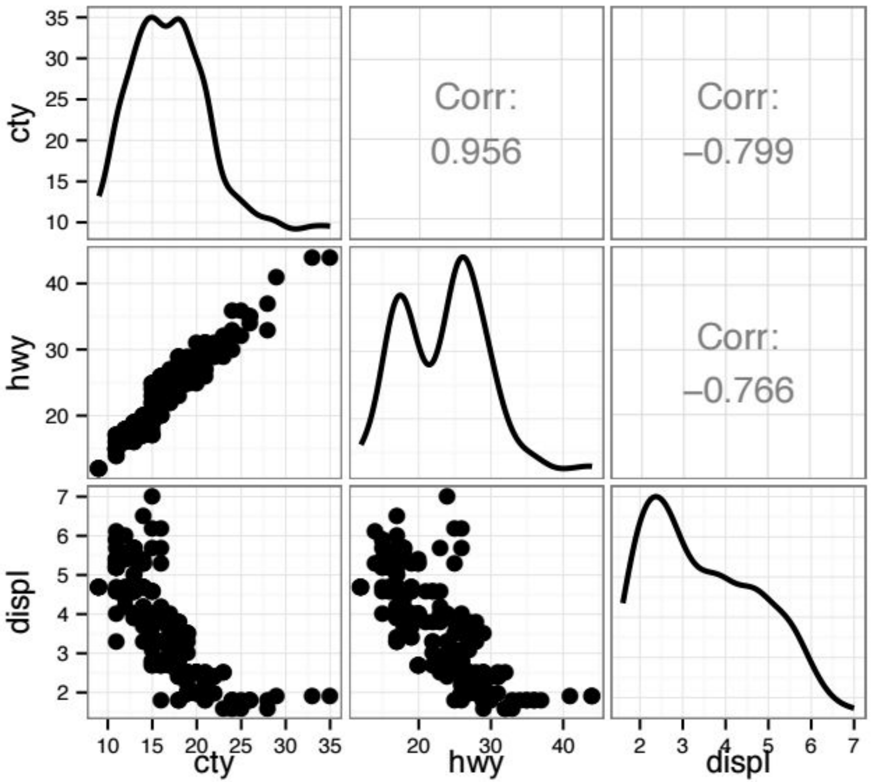
Type M for manual, V for v-shape, A for automatic, F for flat.

M,V	Tend to have lower weights and be more fuel efficient
A,F	Tend to be heavier and less fuel efficient

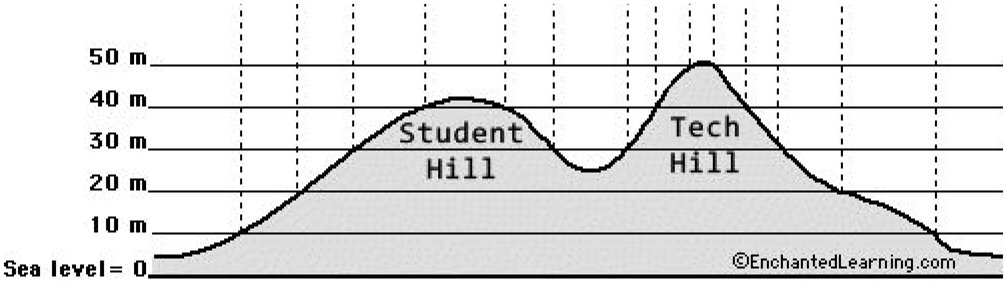
Facets Quiz 4

Complete the command to create this set of graphs.

```
DF = mpg[,
    c( cty , hwy , displ )]
library( GGally )
ggpairs( DF )
```



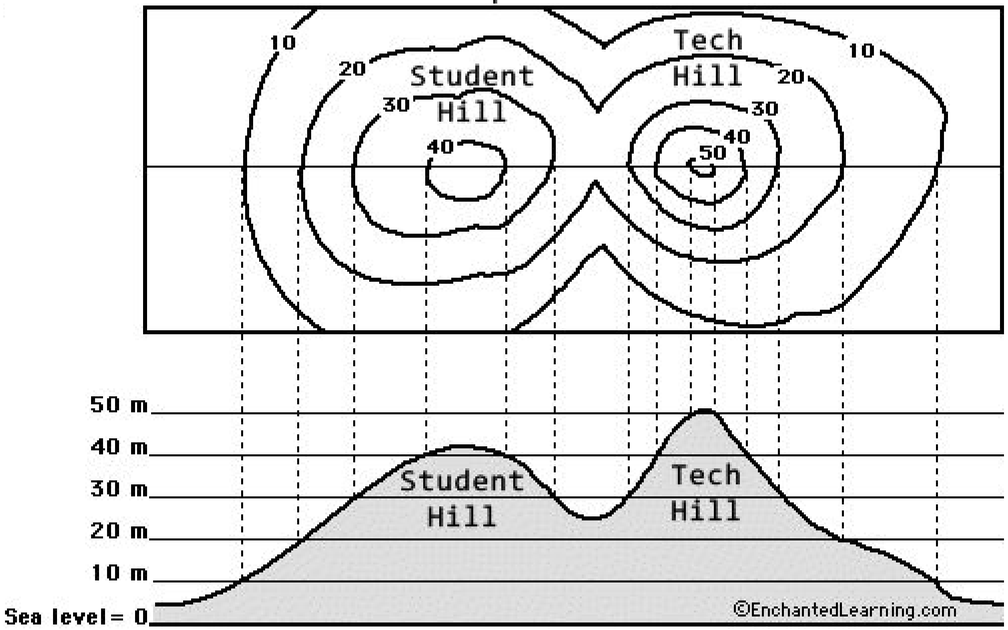
Contour Plots



The two hills seen from the side, with elevations marked and dotted lines pointing to the corresponding contour lines.

Contour Plots

Topographic Map
(with contour lines that show points that are on the same level)



The two hills seen from the side, with elevations marked and dotted lines pointing to the corresponding contour lines.

Contour Plots

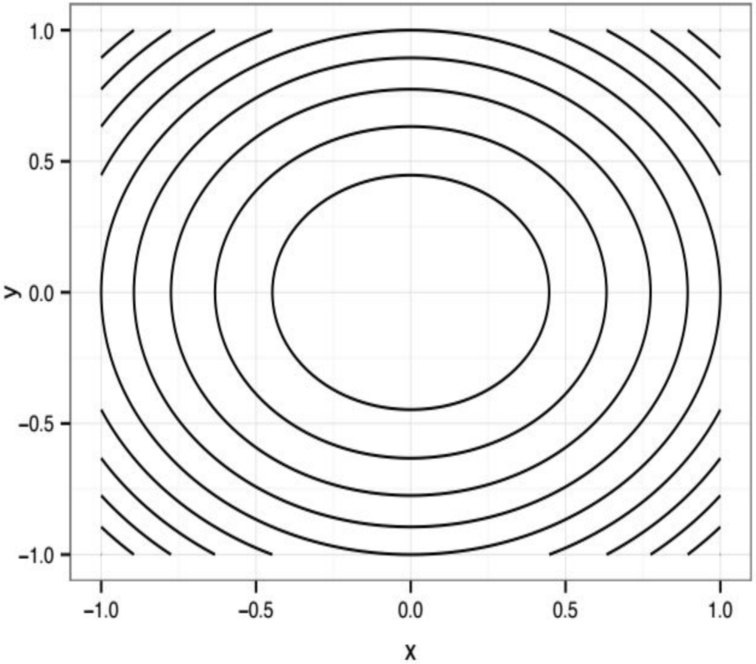
Creating a Contour Plot:

- **Step 1:** create a grid for x values
- **Step 2:** create a grid of y values
- **Step 3:** Create an expanded x by y grid
- **Step 4:** Compute values of z on the expanded grid
- **Step 5:** Graph the data

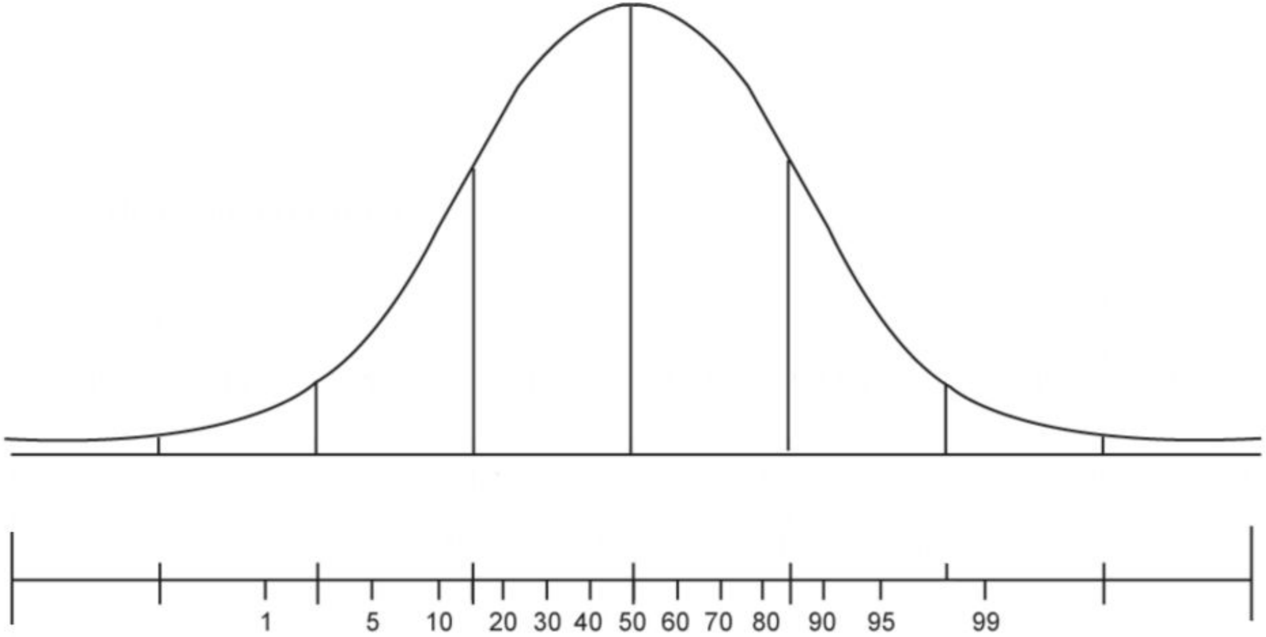
Contour Plots Quiz

Complete the command to generate the plot shown.

```
x_grid = seq (-1,1,length.out
= 100)
y_grid = x_grid
R = expand.grid (x_grid, y_grid)
names(R) = c(’ x ’,’ y ’)
R$z = R$x^2 + R$y^2
ggplot (R, aes (x=x, y=y,
z=z)) + stat_contour ()
```

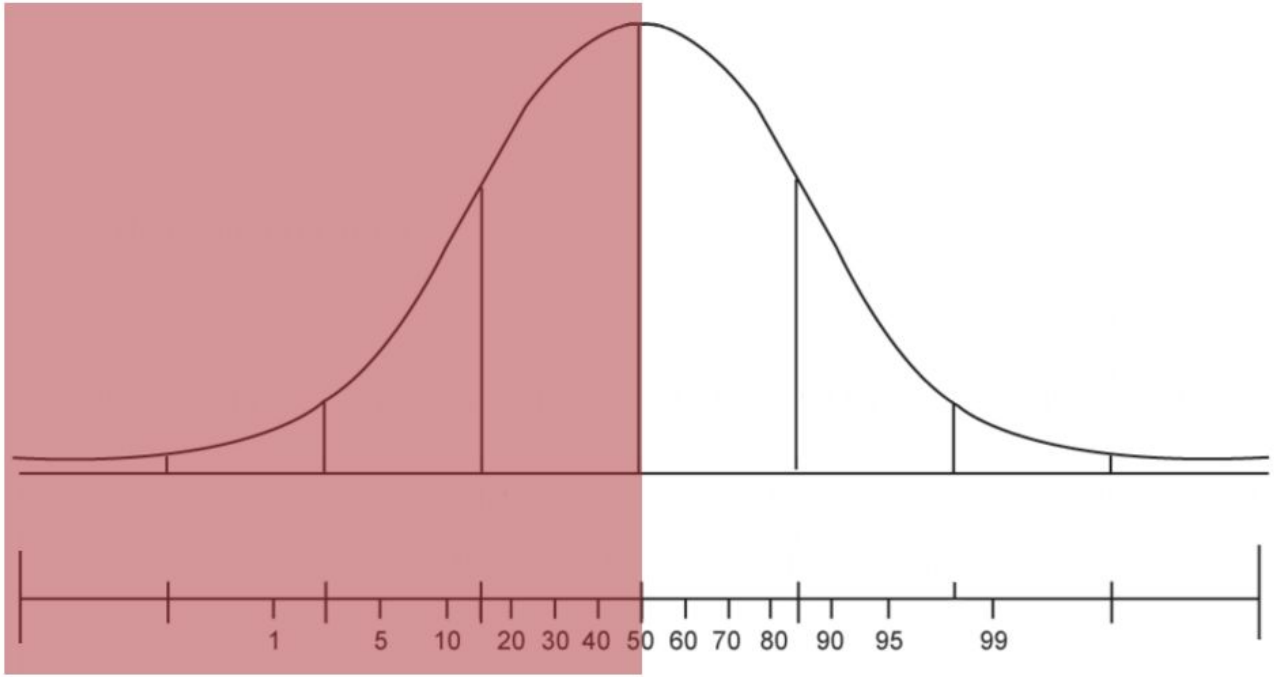


Box Plots



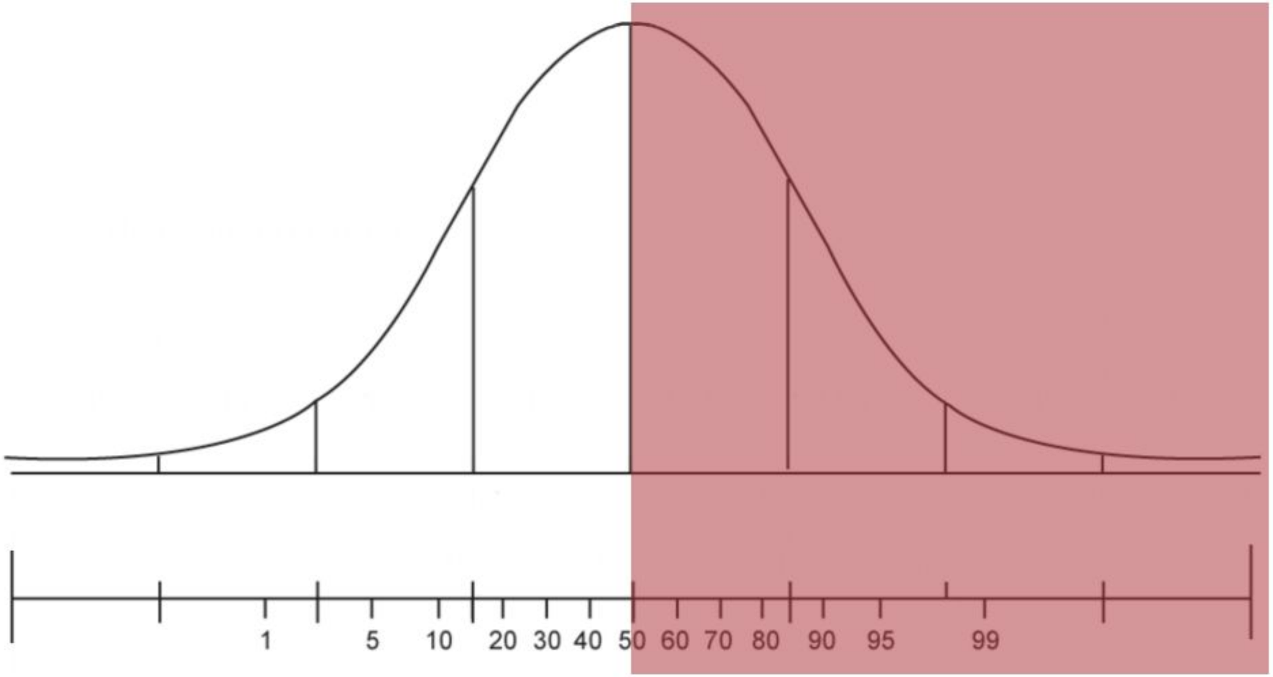
Box Plots

The Median



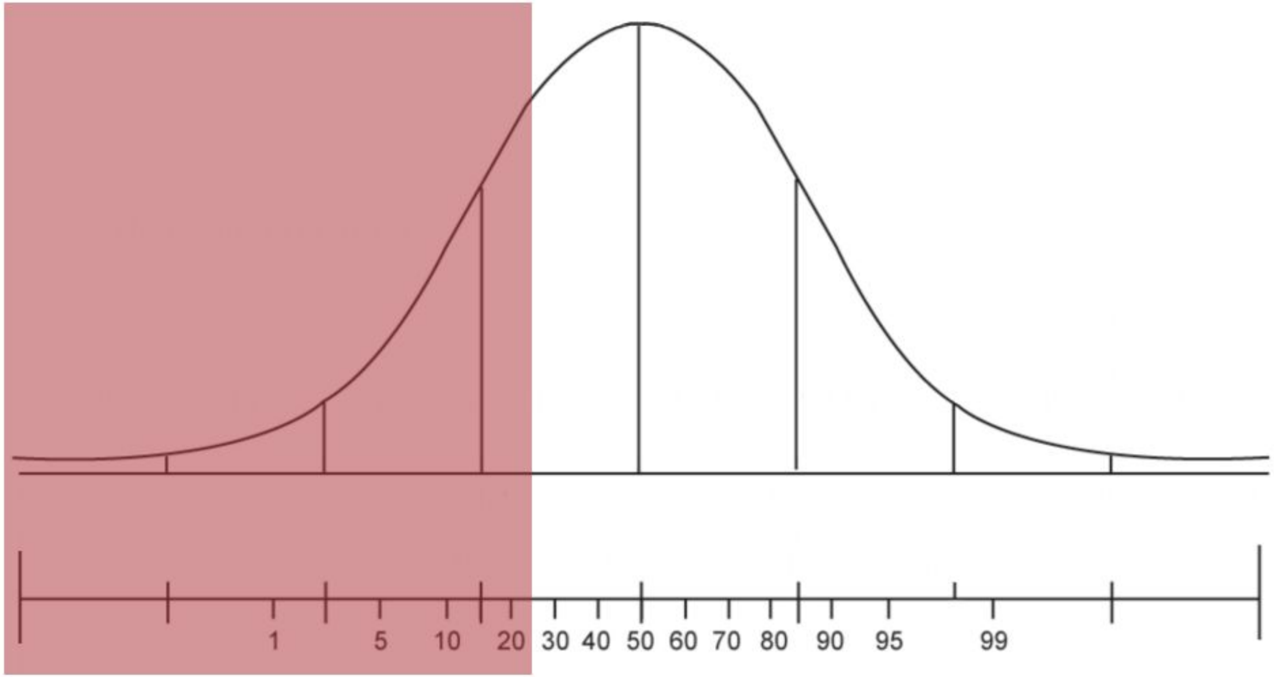
Box Plots

The Median



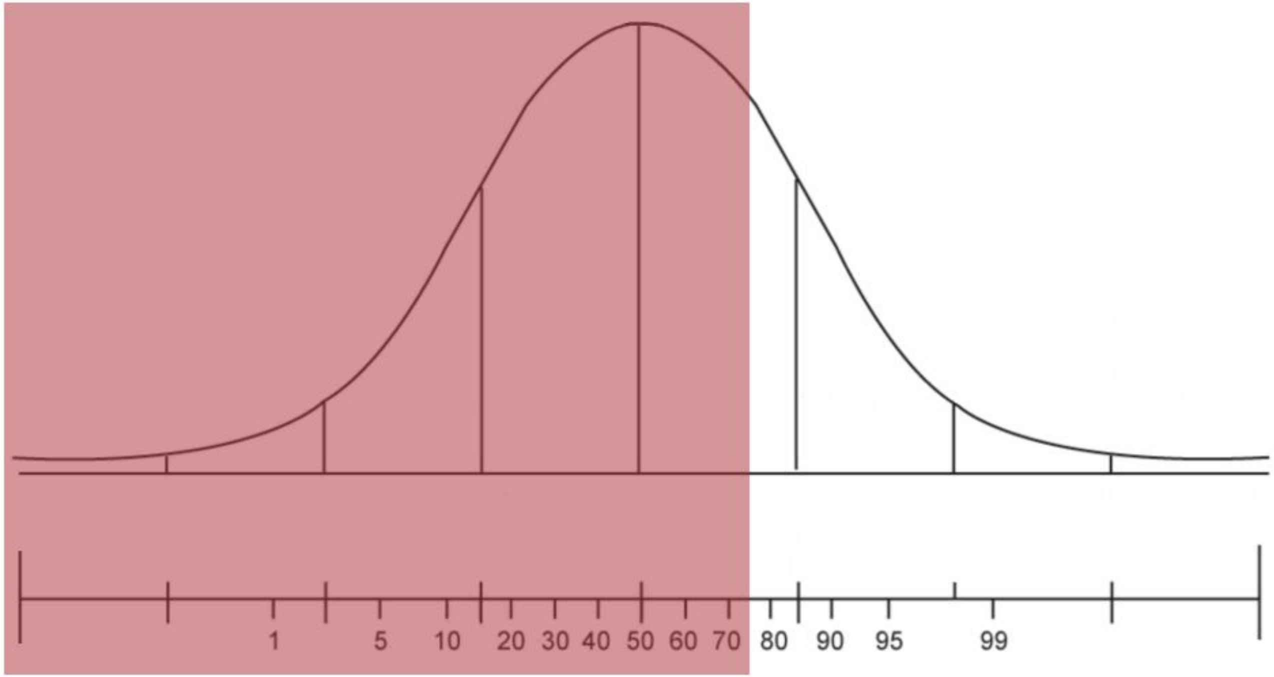
Box Plots

The 25 Percentile



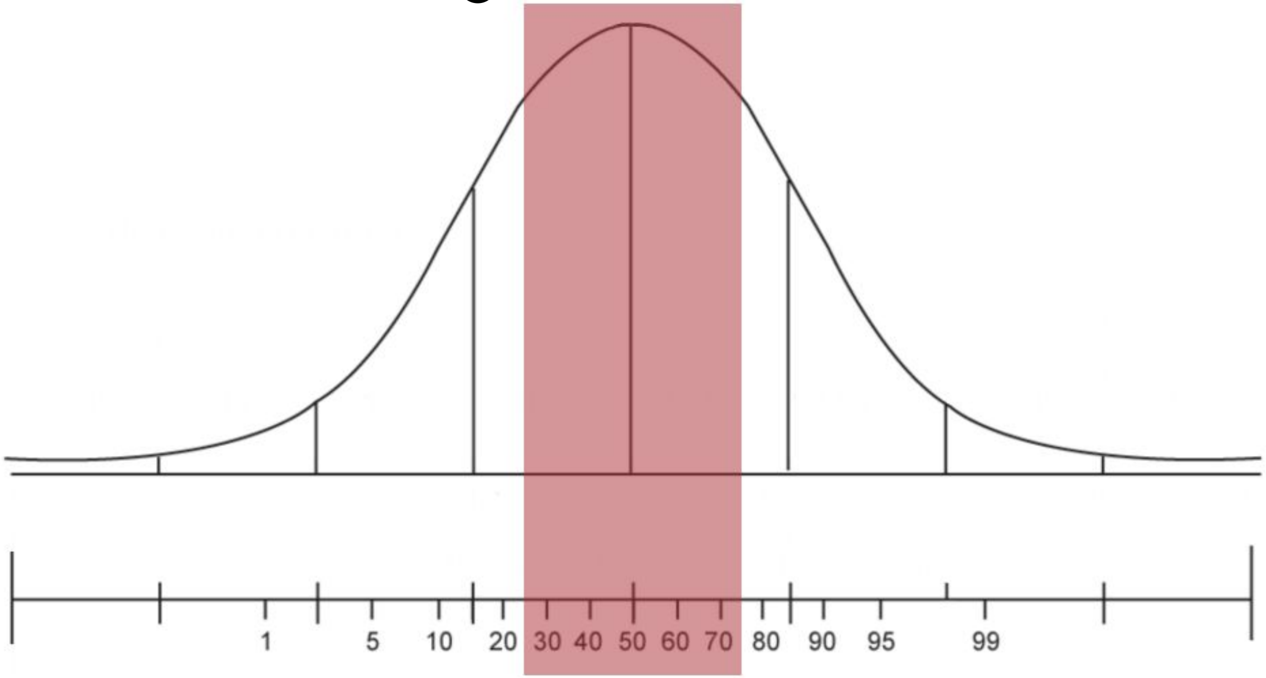
Box Plots

The 75 Percentile



Box Plots

IQR: Inter-Quartile Range



Box Plot Quiz

Complete the command that creates the following:

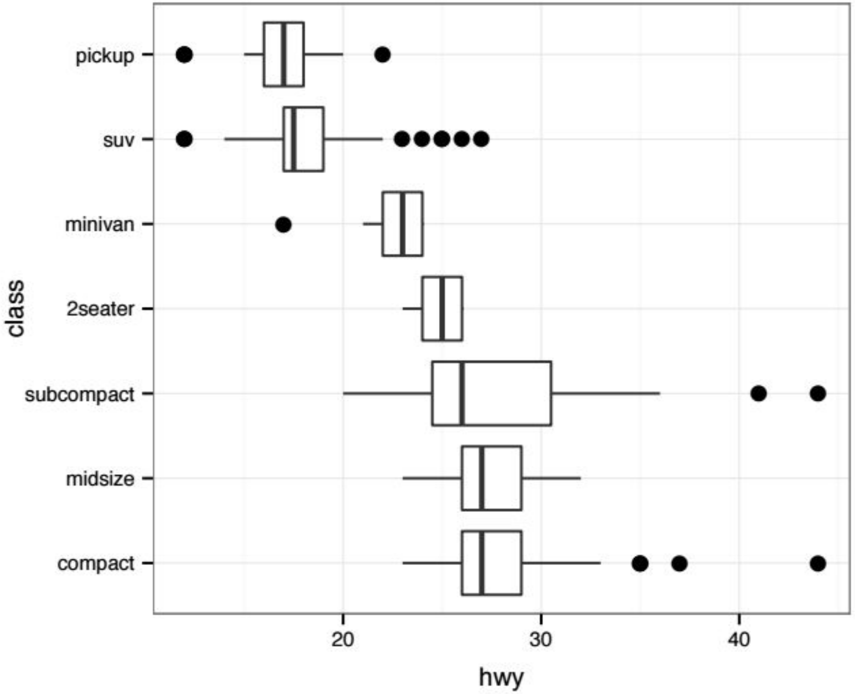
- box denoting the IQR
- An inner line bisecting the box denoting the median
- Whiskers extending to the most extreme point no further tan 1.5 times IQR length away from the edges of the box
- points outside the box and whiskers marked as outliers

```
ggplot(mpg, aes (“”,
hwy)) +
geom_boxplot() +
coord_flip() +
scale_x_discrete (“”)
```

Box Plot Quiz 2

Complete the command to create the shown box plot.

```
ggplot(mpg,
aes( reorder (class, -
hwy, median), hwy)) +
geom_boxplot () +
coord_flip () +
scale_x_discrete( "class" )
```



Box Plot Inference Quiz

Rank the vehicle classes by fuel efficiency. Use a '1' for the most fuel efficient, and a '7' for the least.

1 compacts

7 pickups

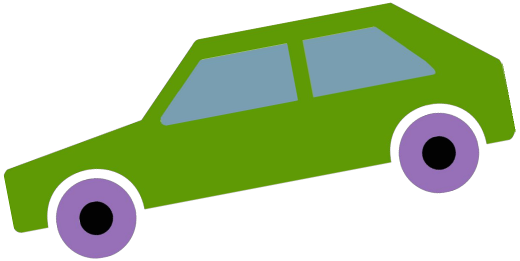
4 2-seaters

6 SUV

5 minivans

3 sub-compacts

2 midsizes



Box Plot Inference Quiz 2

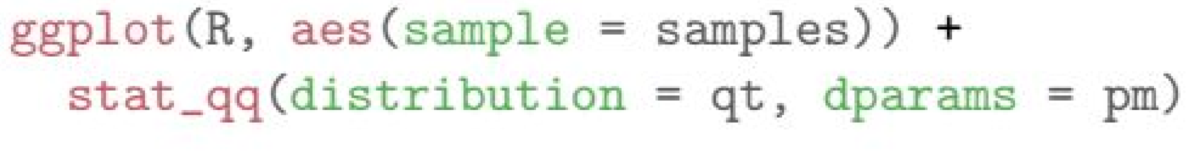
Fill in the blanks with the following answers:
SUV, 2-seater, subcompact, mid sizes, minivans, pickups, compacts

Subcompacts have the highest spread of the categories.

Almost all 2-seaters have a higher highway mpg than SUVs

SUVs and 2-seaters have almost disjoint values.

QQPlots



```
ggplot(R, aes(sample = samples)) +
  stat_qq(distribution = qt, dparams = pm)
```

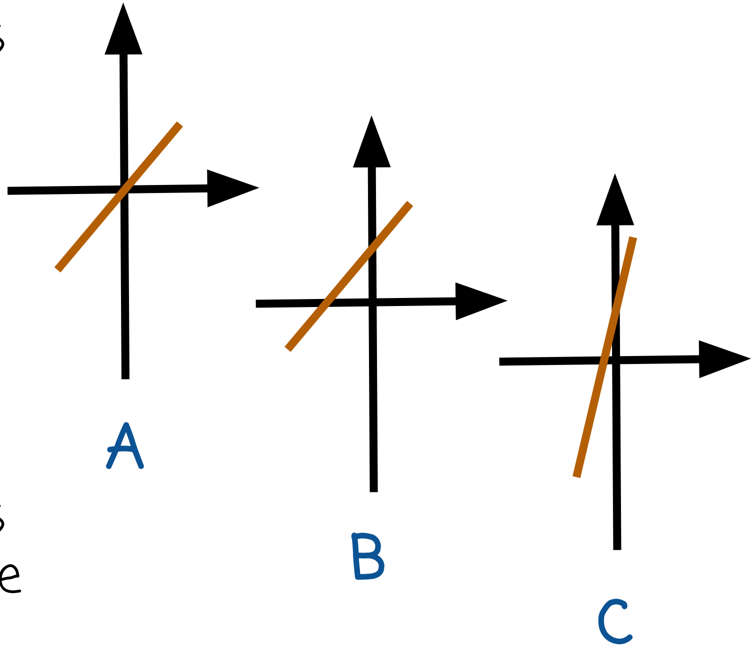
QQPlots Quiz

Match the description to its corresponding plot.

C The two datasets have distributions possessing similar shapes but one is translated and scaled with respect to the other.

A The two datasets have identical quantiles, and therefore they are sampled from the same distribution.

B The two datasets have distributions of similar shape and spread, but one is shifted with respect to the other

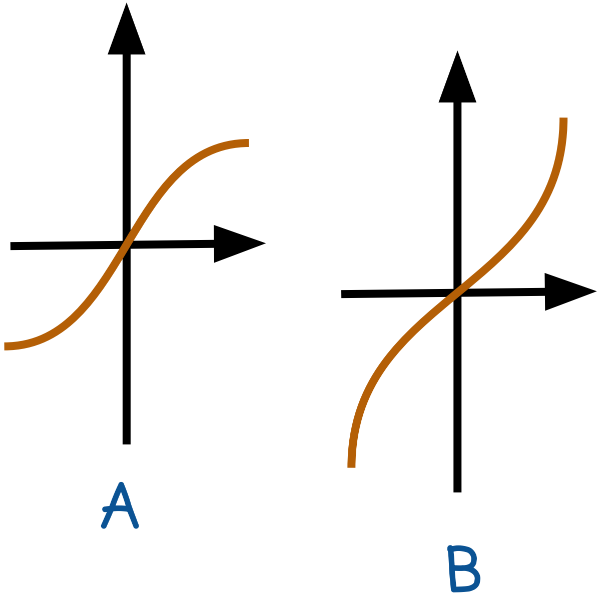


QQPlots Quiz 2

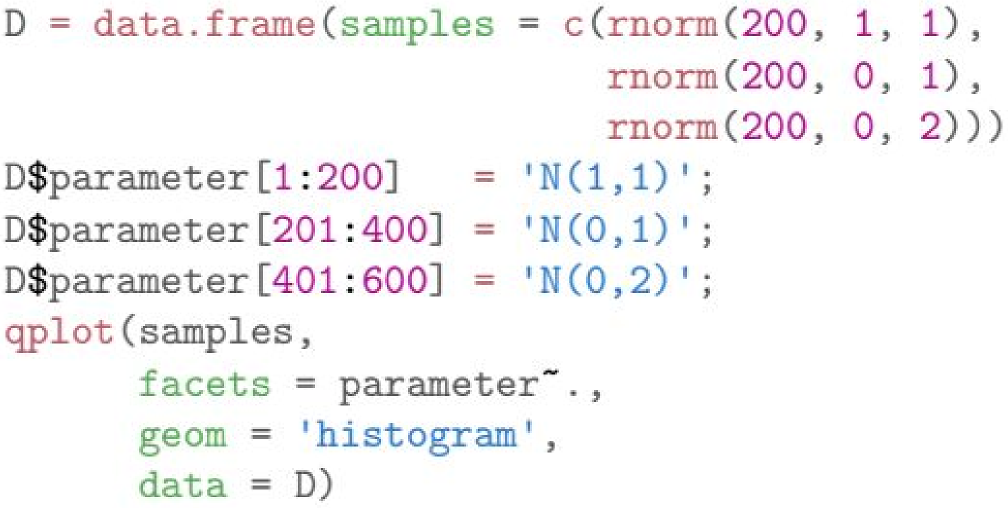
Match the plots the corresponding description.

□ B The dataset whose quantiles correspond to the y-axis is drawn from a distribution having heavier tails than the other dataset.

□ A The dataset corresponding to the x-axis is sampled from a distribution with heavier tails than the other dataset.

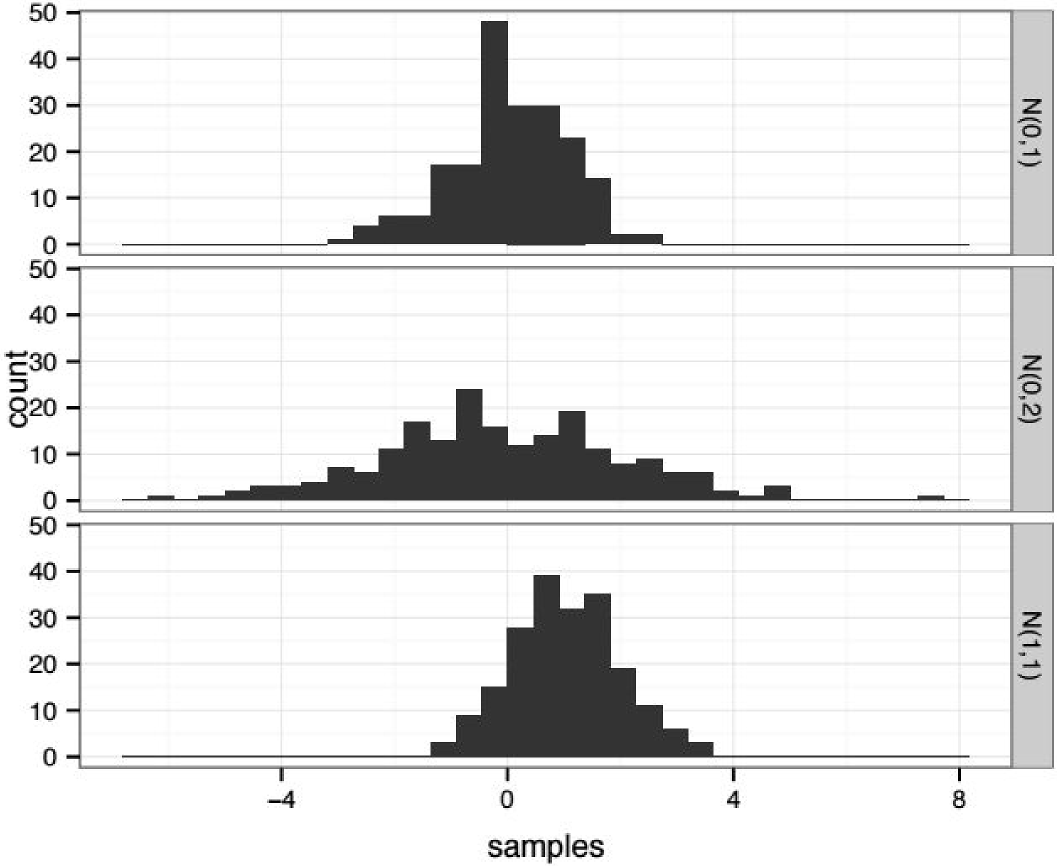


qqplot Example

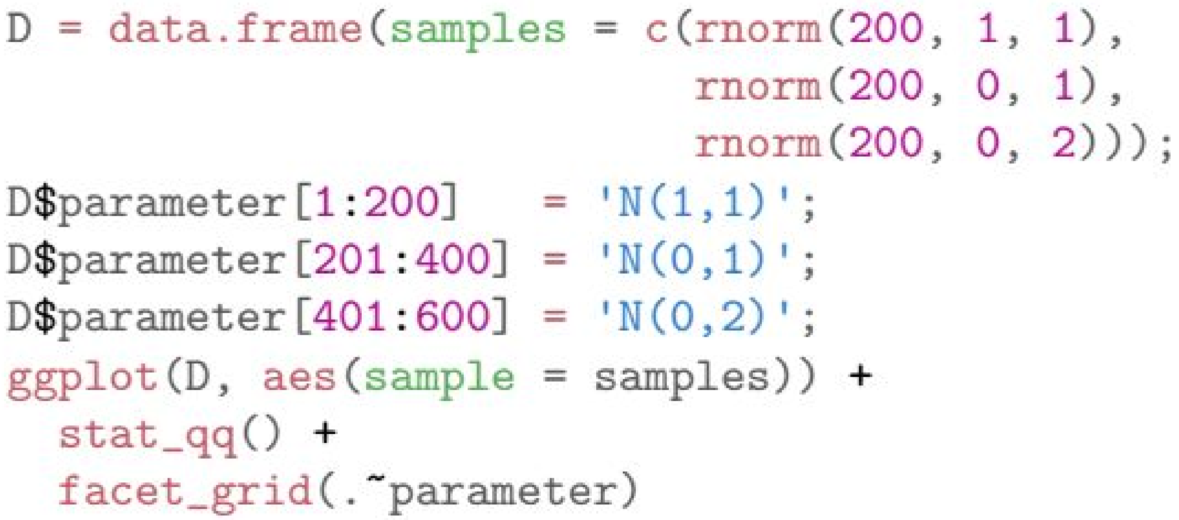


```
D = data.frame(samples = c(rnorm(200, 1, 1),
                           rnorm(200, 0, 1),
                           rnorm(200, 0, 2)))
D$parameter[1:200]   = 'N(1,1)';
D$parameter[201:400] = 'N(0,1)';
D$parameter[401:600] = 'N(0,2)';
qplot(samples,
      facets = parameter˜.,
      geom = 'histogram',
      data = D)
```

qqplot Example

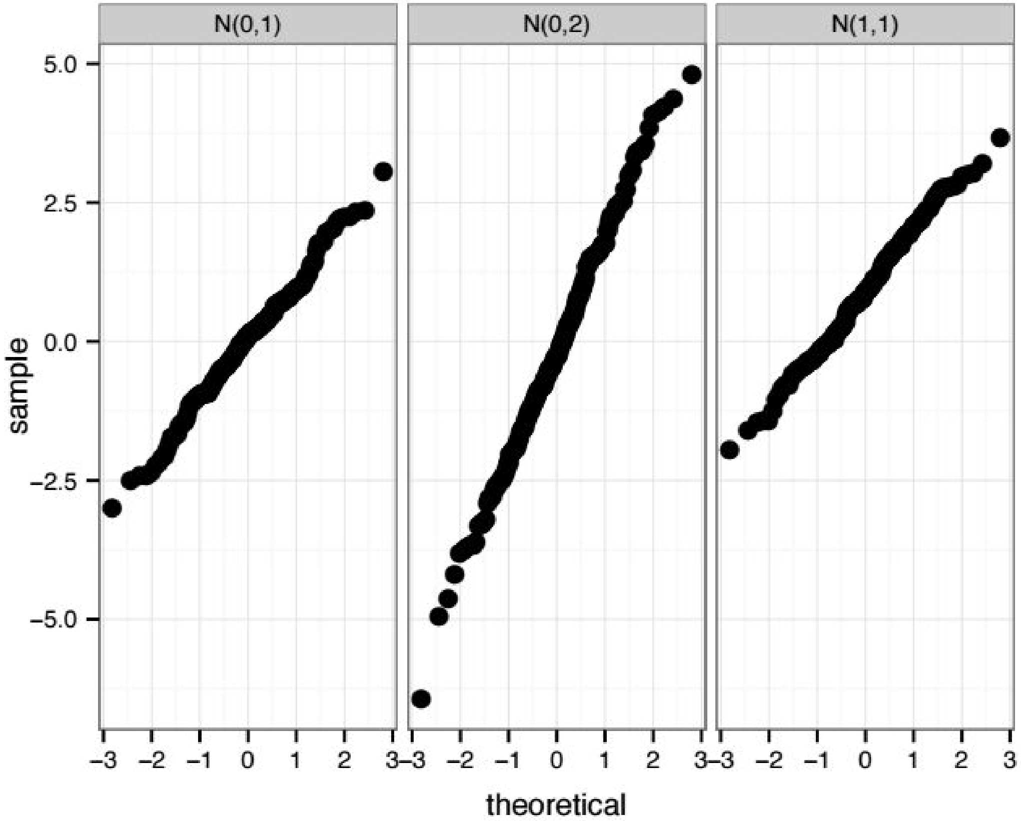


qqplot Example

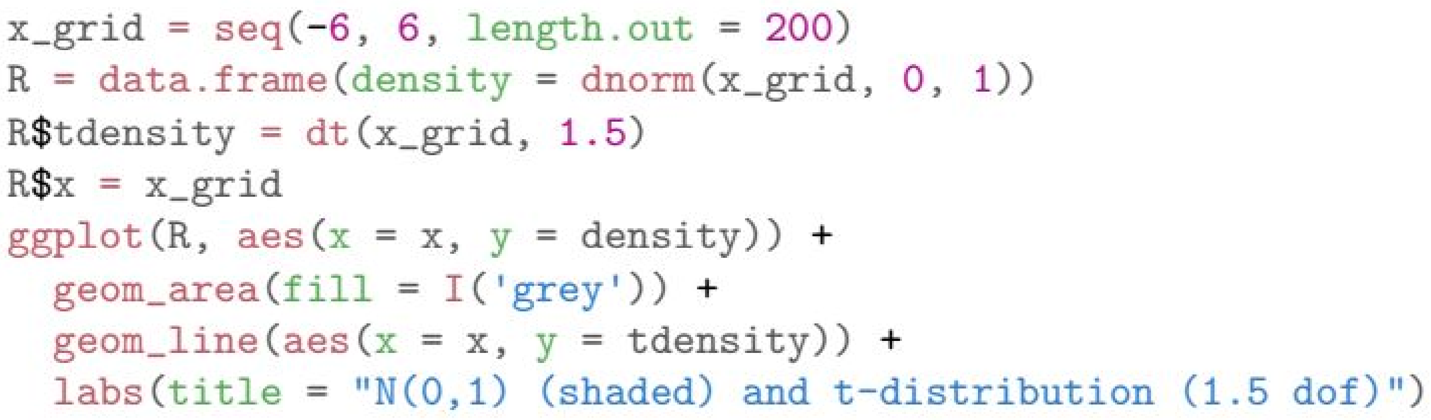


```
D = data.frame(samples = c(rnorm(200, 1, 1),
                           rnorm(200, 0, 1),
                           rnorm(200, 0, 2)));
D$parameter[1:200]   = 'N(1,1)';
D$parameter[201:400] = 'N(0,1)';
D$parameter[401:600] = 'N(0,2)';
ggplot(D, aes(sample = samples)) +
  stat_qq() +
  facet_grid(.~parameter)
```

qqplot Example

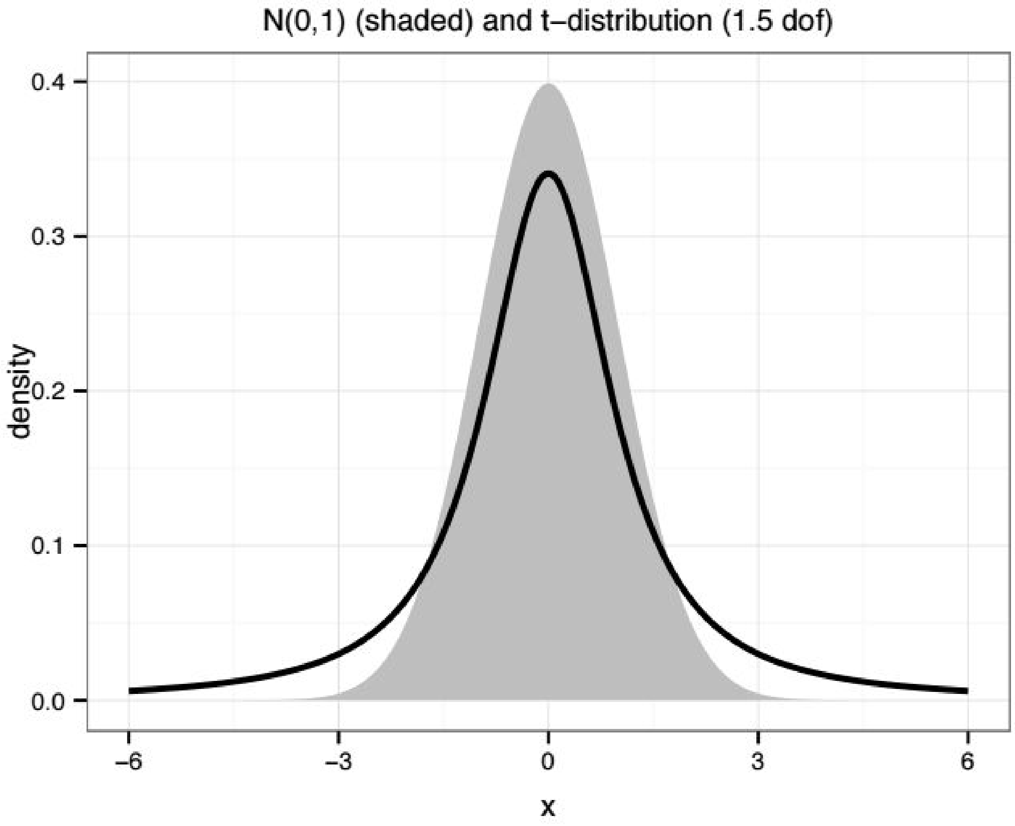


ggplot Example

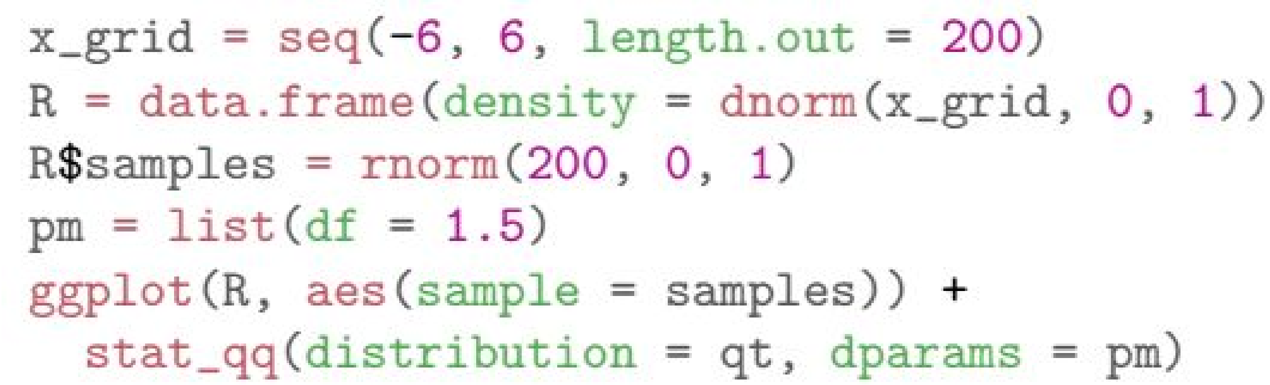


```
x_grid = seq(-6, 6, length.out = 200)
R = data.frame(density = dnorm(x_grid, 0, 1))
R$tdensity = dt(x_grid, 1.5)
R$x = x_grid
ggplot(R, aes(x = x, y = density)) +
  geom_area(fill = I('grey')) +
  geom_line(aes(x = x, y = tdensity)) +
  labs(title = "N(0,1) (shaded) and t-distribution (1.5 dof)")
```

qqplot Example

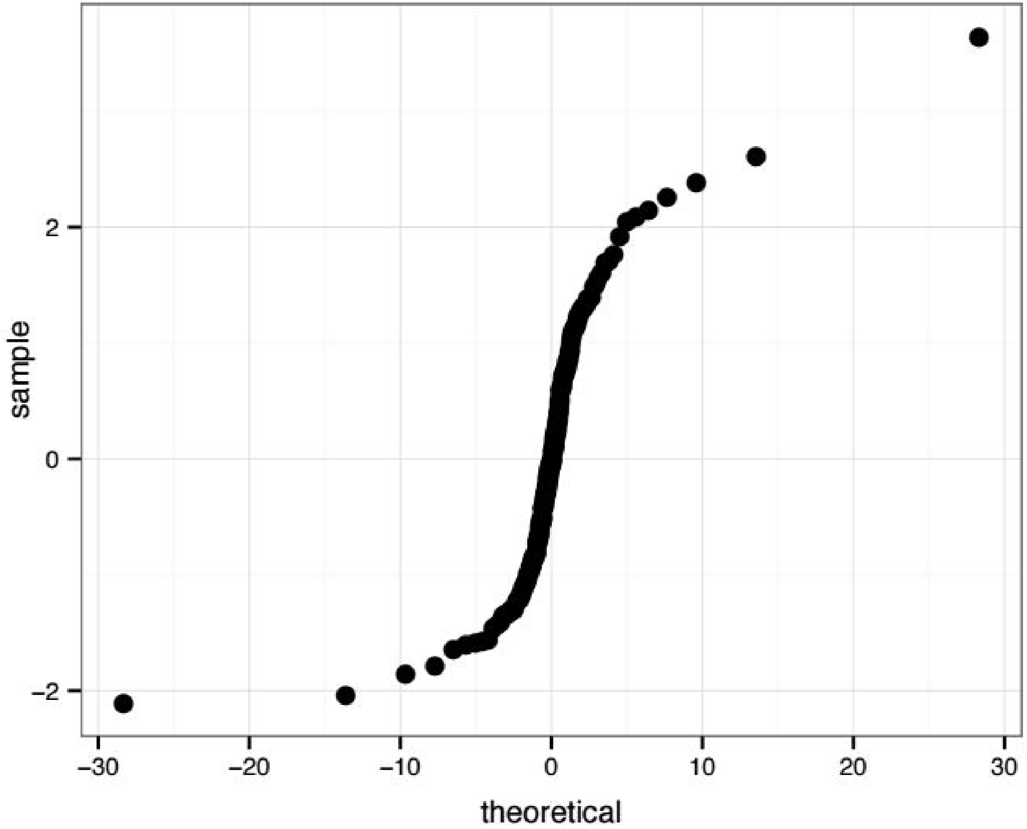


qqplot Example



```
x_grid = seq(-6, 6, length.out = 200)
R = data.frame(density = dnorm(x_grid, 0, 1))
R$samples = rnorm(200, 0, 1)
pm = list(df = 1.5)
ggplot(R, aes(sample = samples)) +
  stat_qq(distribution = qt, dparams = pm)
```

qqplot Example


sample
theoretical