Data Visualization with ggplot2 (Part 2)

Lessons from DataCamp

Contents

Introduction

The following document outlines the written portion of the lessons from DataCamp's Data Visualization with ggplot2 (Part 2). This requires Intermediate R-Knowledge and understanding of Part 1 of the course.

As a note: All text is completely copied and pasted from the course. There are instances where the document refers to the "editor on the right". please note, that in this notebook document all of the instances are noted in the "r-chunks" (areas containing working r-code), which occurs below the text, rather than to the right. Furthermore, This lesson contained instructional videos at the beginning of new concepts that are not detailed in this document. However, even without these videos, the instructions are quite clear in indicating what the code is accomplishing.

If you have this document open on "R-Notebook", simply click "run" -> "Run all" (Or just press 'ctrl + alt + r'), let the "r-chunks" run (This might take a bit of time) then click "Preview". There are 5 necessary datasets to run this program, please create an r-project with this data or set a working directory (required files names are available in the "Required data for this session" section)

This document was created by Neil Yetz on 06/02/2018. Please send any questions or concerns in this document to Neil at ndyetz@gmail.com

Required packages for this session

Below are the install.packages and libraries you will need to have in order to run this session successfully.

```
library(ggplot2)
library(RColorBrewer)
library(car) #<- Vocab dataset
library(Hmisc)
```

Required data for this session

Course Description

This ggplot2 tutorial builds on your knowledge from the first course to produce meaningful explanatory plots. We'll explore the last four optional layers. Statistics will be calculated on the fly and we'll see how Coordinates and Facets aid in communication. Publication quality plots will be produced directly in R using the Themes layer. We'll also discuss details on data visualization best practices with ggplot2 to help make sure you have a sound understanding of what works and why. By the end of the course, you'll have all the

tools needed to make a custom plotting function to explore a large data set, combining statistics and excellent visuals.

Chapter 1: Statistics

In this chapter, we'll delve into how to use R ggplot2 as a tool for graphical data analysis, progressing from just plotting data to applying a variety of statistical methods. This includes a variety of linear models, descriptive and inferential statistics (mean, standard deviation and confidence intervals) and custom functions.

Smoothing

Welcome to the exercises for the second ggplot2 course!

To practice on the remaining four layers (statistics, coordinates, facets and themes), we'll continue working on several datasets that we already encountered in the first course.

The mtcars dataset contains information for 32 cars from Motor Trends magazine from 1973. This dataset is small, intuitive, and contains a variety of continuous and categorical (both nominal and ordinal) variables.

In the previous course we learned how to effectively use some basic geometries, such as point, bar and line. In the first chapter of this course we'll explore statistics associated with specific geoms, for example, smoothing and lines.

INSTRUCTIONS

Familiarize yourself again with the mtcars dataset using str(). Extend the first ggplot call: add a LOESS smooth to the scatter plot (which is the default) with geom_smooth(). We want to have the actual values and the smooth on the same plot.

Change the previous plot to use an ordinary linear model, by default it will be y ~ x, so you don't have to specify a formula. You should set the method argument to "lm".

Modify the previous plot to remove the 95% CI ribbon. You should set the se argument to FALSE.

Modify the previous plot to show only the model, and not the underlying dots.

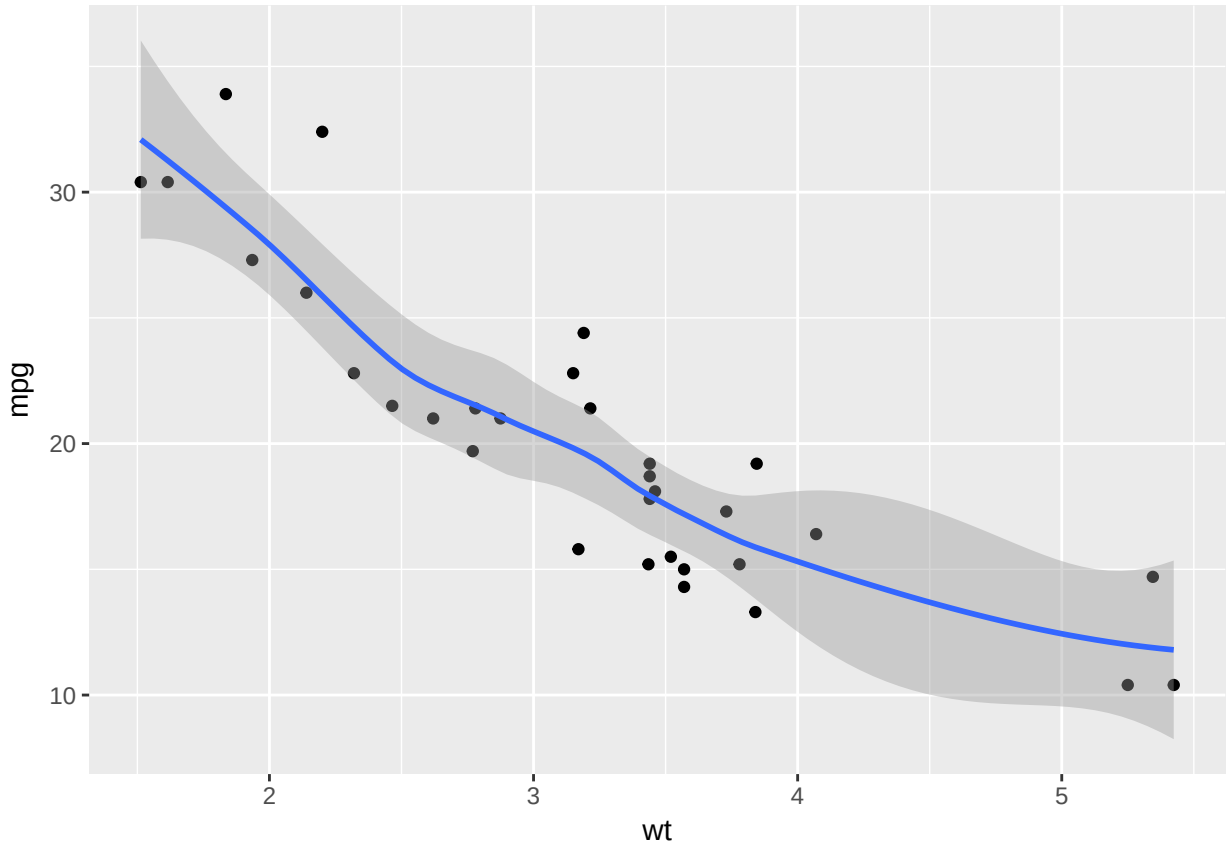
```
# ggplot2 is already loaded

# Explore the mtcars data frame with str()
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

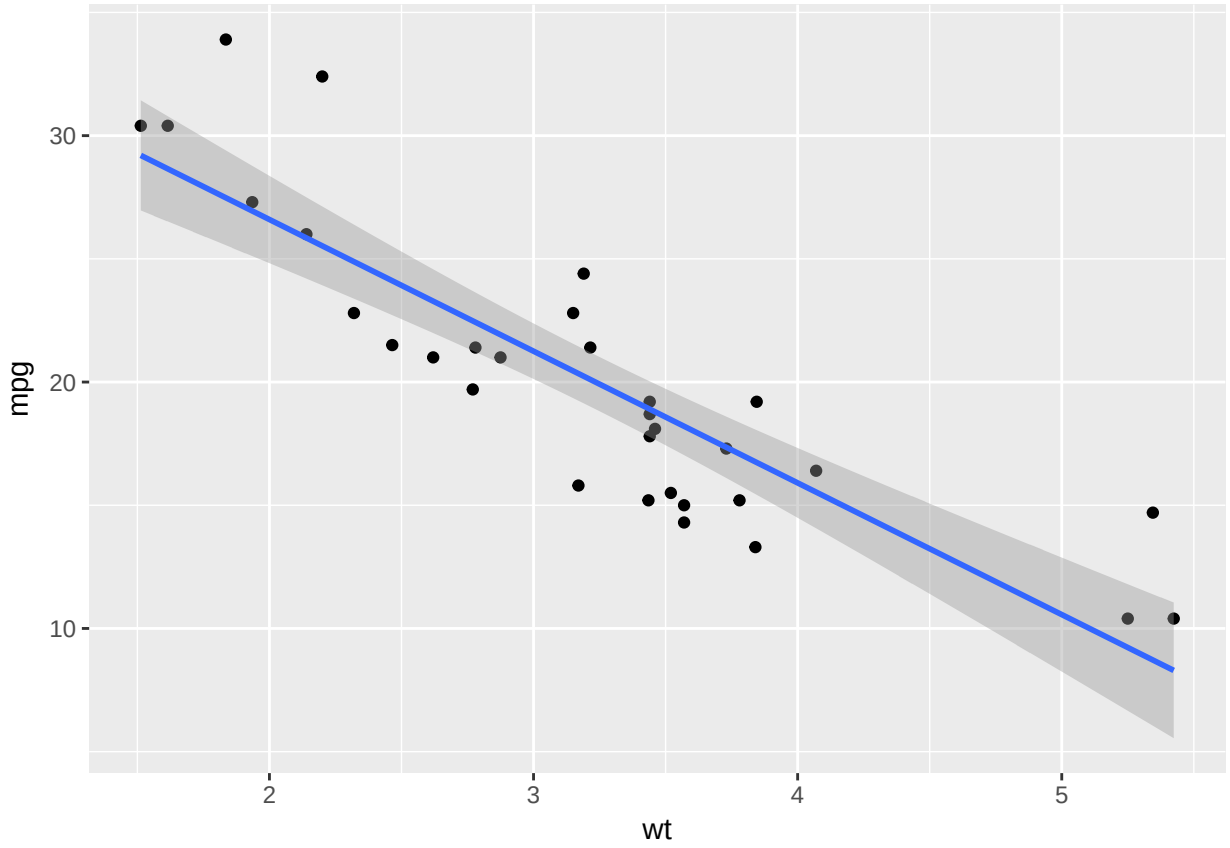
```
# A scatter plot with LOESS smooth
ggplot(mtcars, aes(x = wt, y = mpg)) +
```

```
  geom_point() +
  geom_smooth()
```

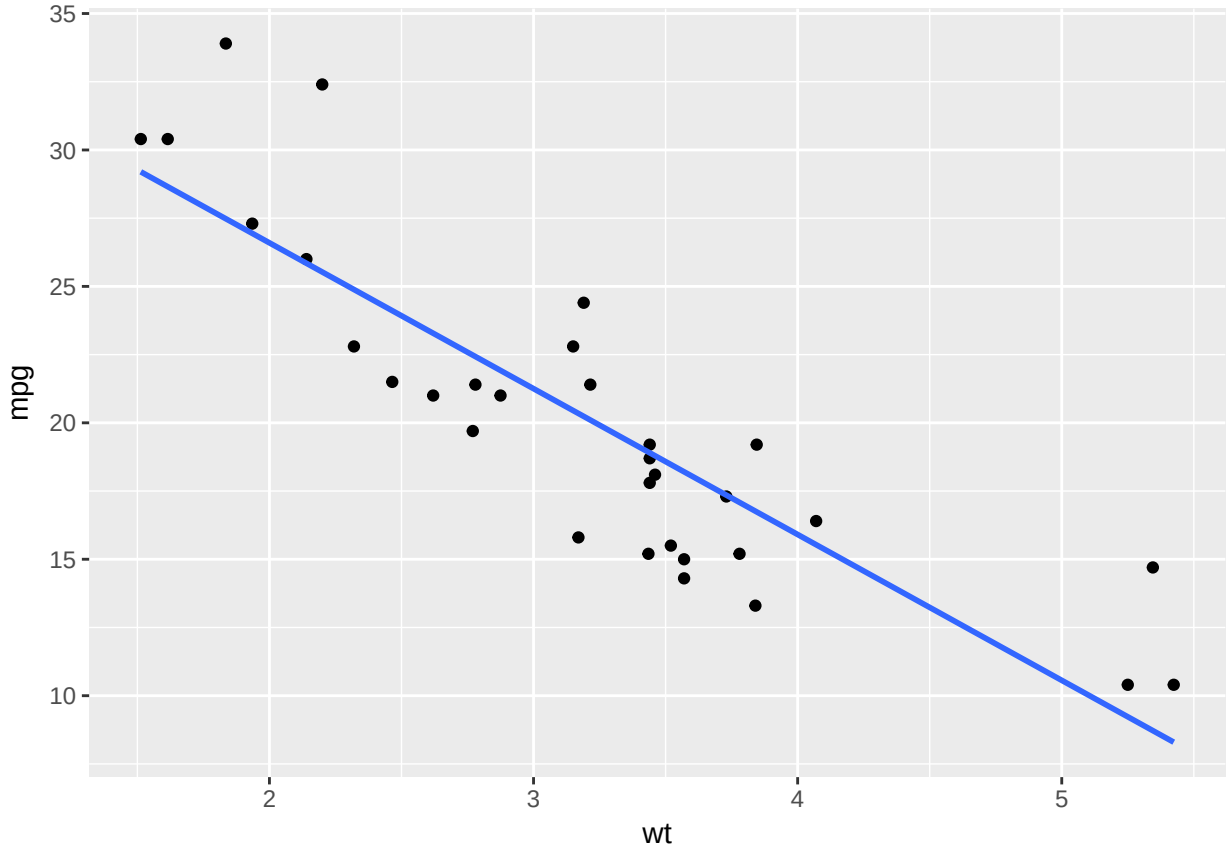
```
## `geom_smooth()` using method = 'loess'
```



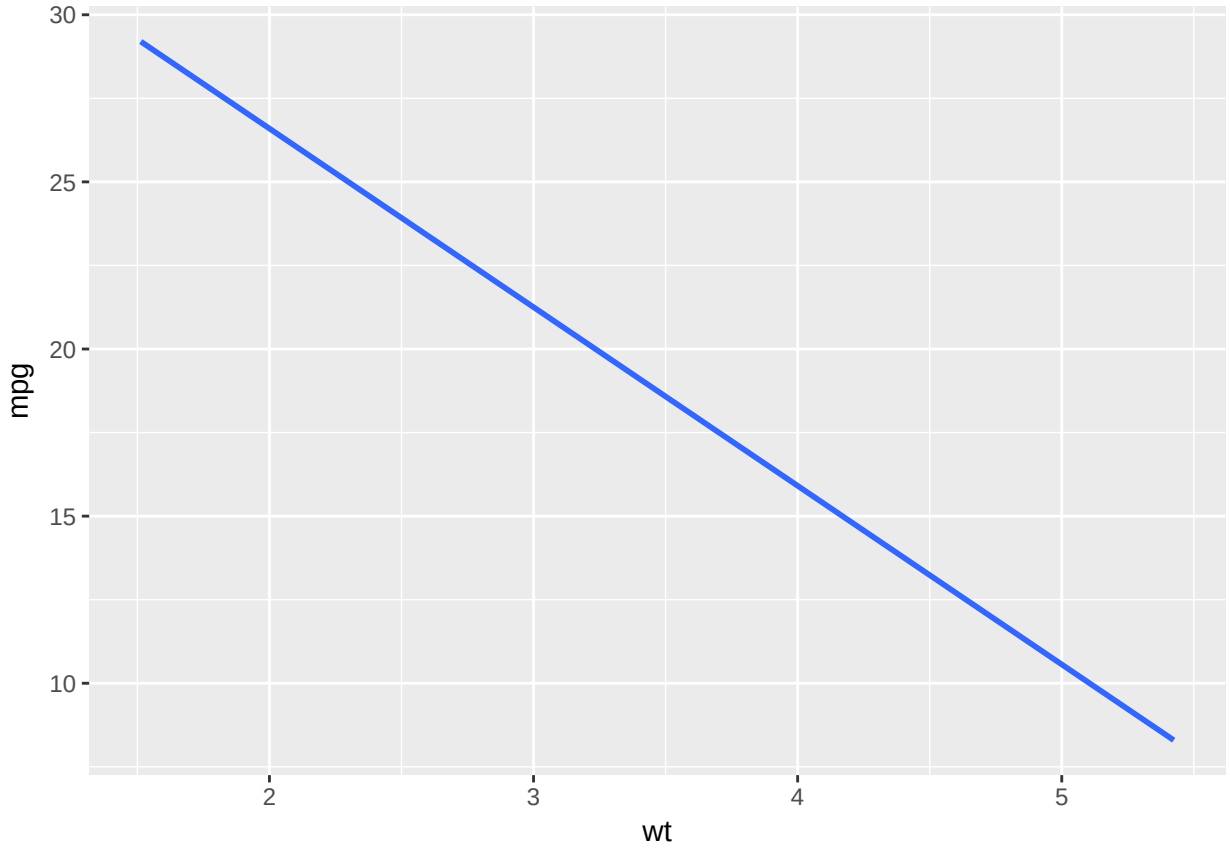

```
# A scatter plot with an ordinary Least Squares linear model
ggplot(mtcars, aes(x = wt, y = mpg)) +
  geom_point() +
  geom_smooth(method = "lm")
```




```
# The previous plot, without CI ribbon
ggplot(mtcars, aes(x = wt, y = mpg)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```




```
# The previous plot, without points
ggplot(mtcars, aes(x = wt, y = mpg)) +
  geom_smooth(method = "lm", se = FALSE)
```



Grouping variables

We'll continue with the previous exercise by considering the situation of looking at sub-groups in our dataset. For this we'll encounter the invisible group aesthetic.

INSTRUCTIONS

A plot that maps cyl onto the col aesthetic is already coded.

Change col so that factor(cyl) is mapped onto it instead of just cyl.

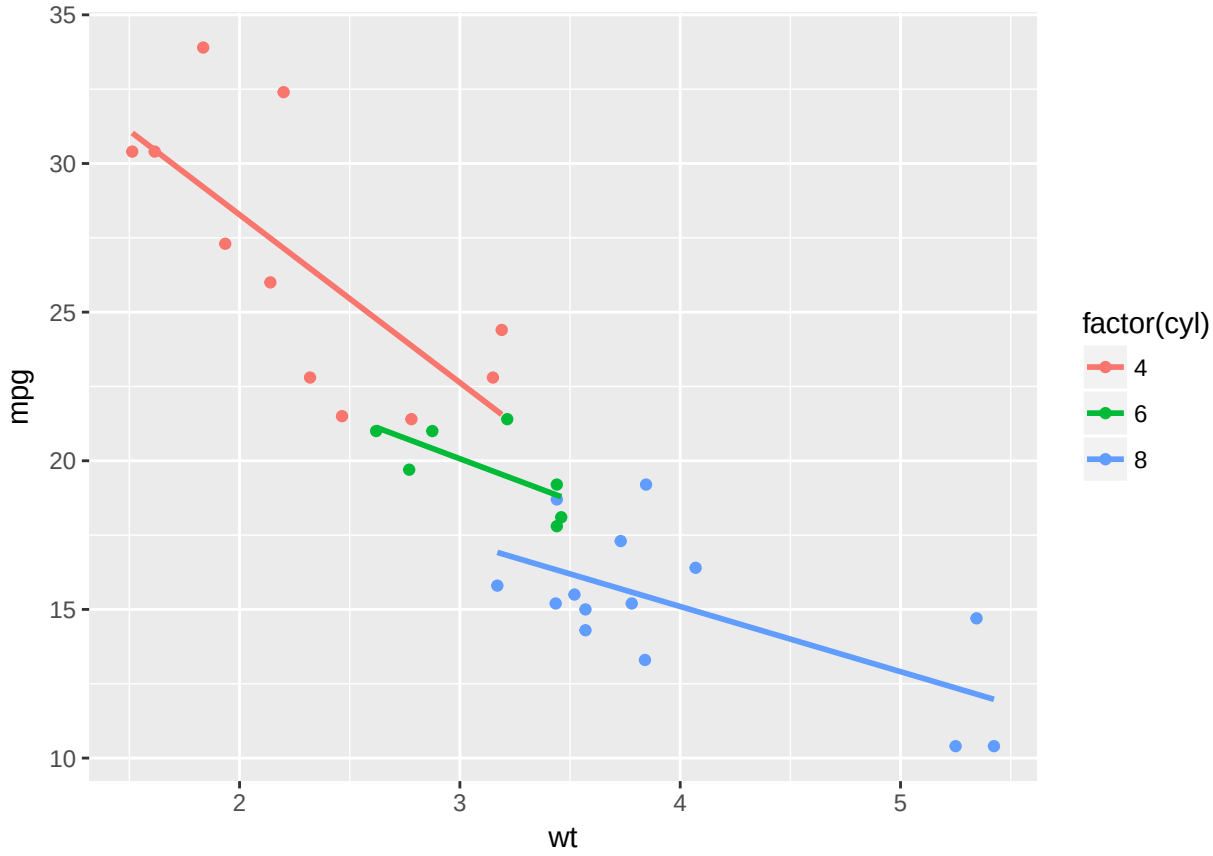
Note: In this ggplot command our smooth is calculated for each subgroup because there is an invisible aesthetic, group which inherits from col.

Complete the second ggplot command.

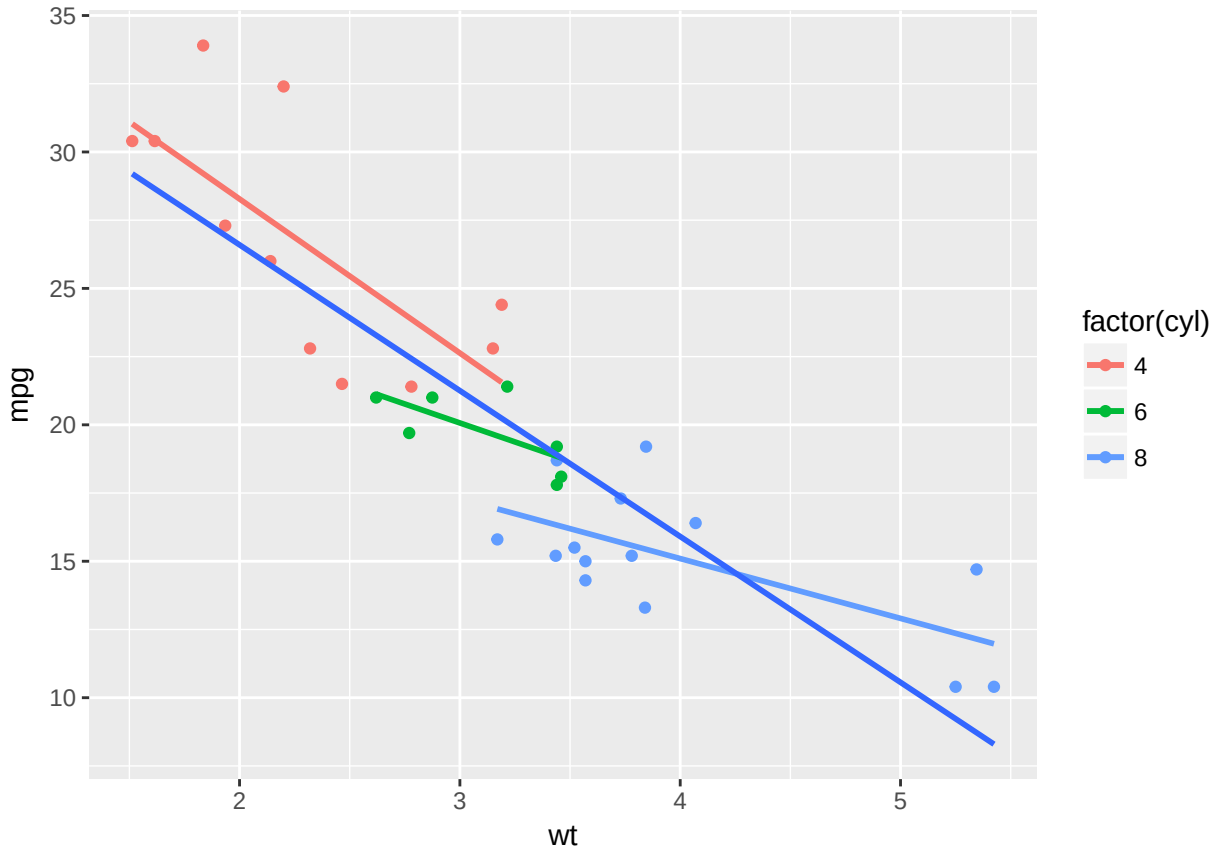
Add another stat_smooth() layer with exactly the same attributes (method set to "lm", se to FALSE).

Add a group aesthetic inside the aes() of this new stat_smooth(), set it to a dummy variable, 1.

```
# ggplot2 is already loaded

# 1 - Define cyl as a factor variable
ggplot(mtcars, aes(x = wt, y = mpg, col = factor(cyl))) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE)
```




```
# 2 - Plot 1, plus another stat_smooth() containing a nested aes()
ggplot(mtcars, aes(x = wt, y = mpg, col = factor(cyl))) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE) +
  stat_smooth(method = "lm", se = FALSE, aes(group = 1))
```



Modifying stat_smooth

In the previous exercise we used se = FALSE in stat_smooth() to remove the 95% Confidence Interval. Here we'll consider another argument, span, used in LOESS smoothing, and we'll take a look at a nice scenario of properly mapping different models.

ggplot2 is already loaded and several of the linear models we looked at in the two previous exercises are already given.

INSTRUCTIONS

Plot 1: Recall that LOESS smoothing is a non-parametric form of regression that uses a weighted, sliding-window, average to calculate a line of best fit. We can control the size of this window with the span argument.

Add span, set it to 0.7. Plot 2: In this plot, we set a linear model for the entire dataset as well as each subgroup, defined by cyl. In the second stat_smooth(),

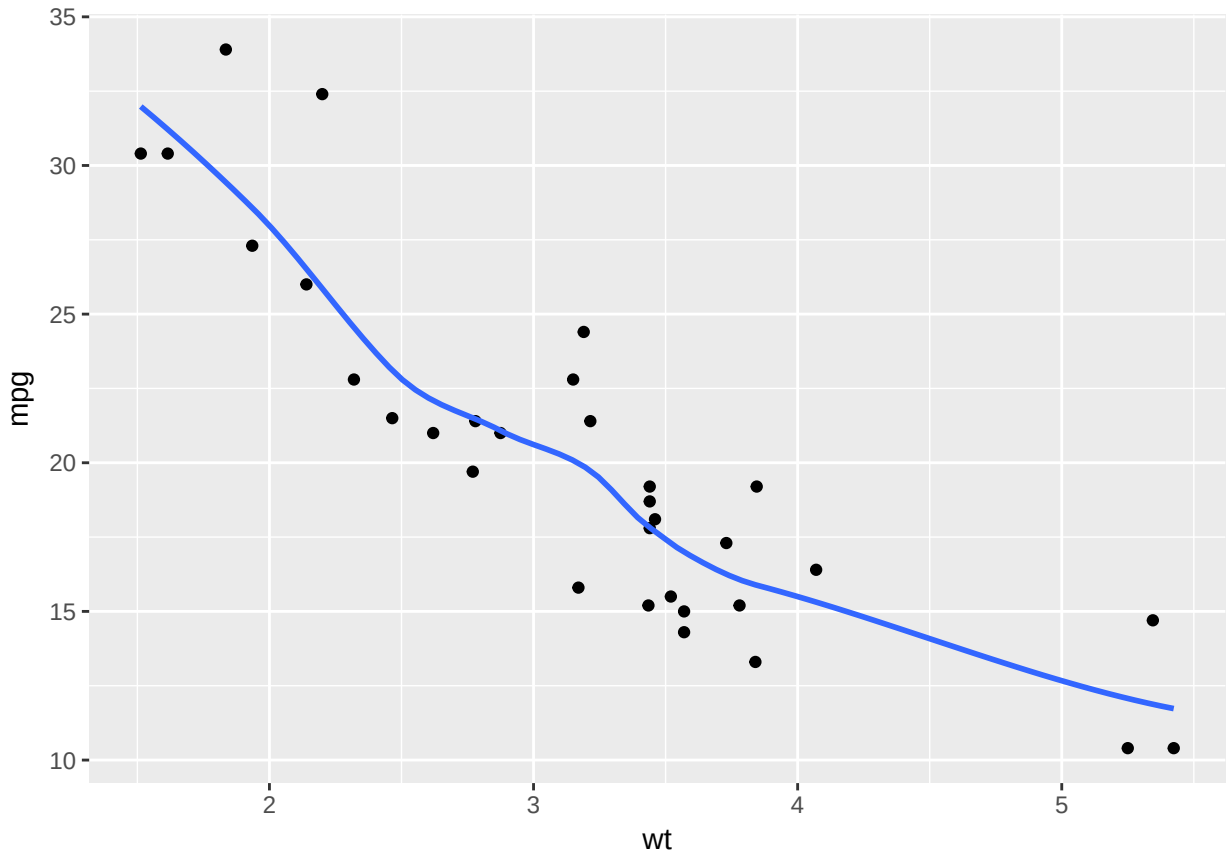
Set method to "loess" (this is the default with a small (n < 1000) data set, but we will specify it explicitly). Add span, set it to 0.7. Plot 3: Plot 2 presents a problem because there is a black line on our plot that is not included in the legend. To get this, we need to map something to col as an aesthetic, not just set col as an attribute.

Add col to the aes() function in the second stat_smooth(), set it to "All". This will name the line properly. Remove the col attribute in the second stat_smooth(). Otherwise, it will overwrite the col aesthetic. Plot 4: Now we should see our "All" model in the legend, but it's not black anymore.

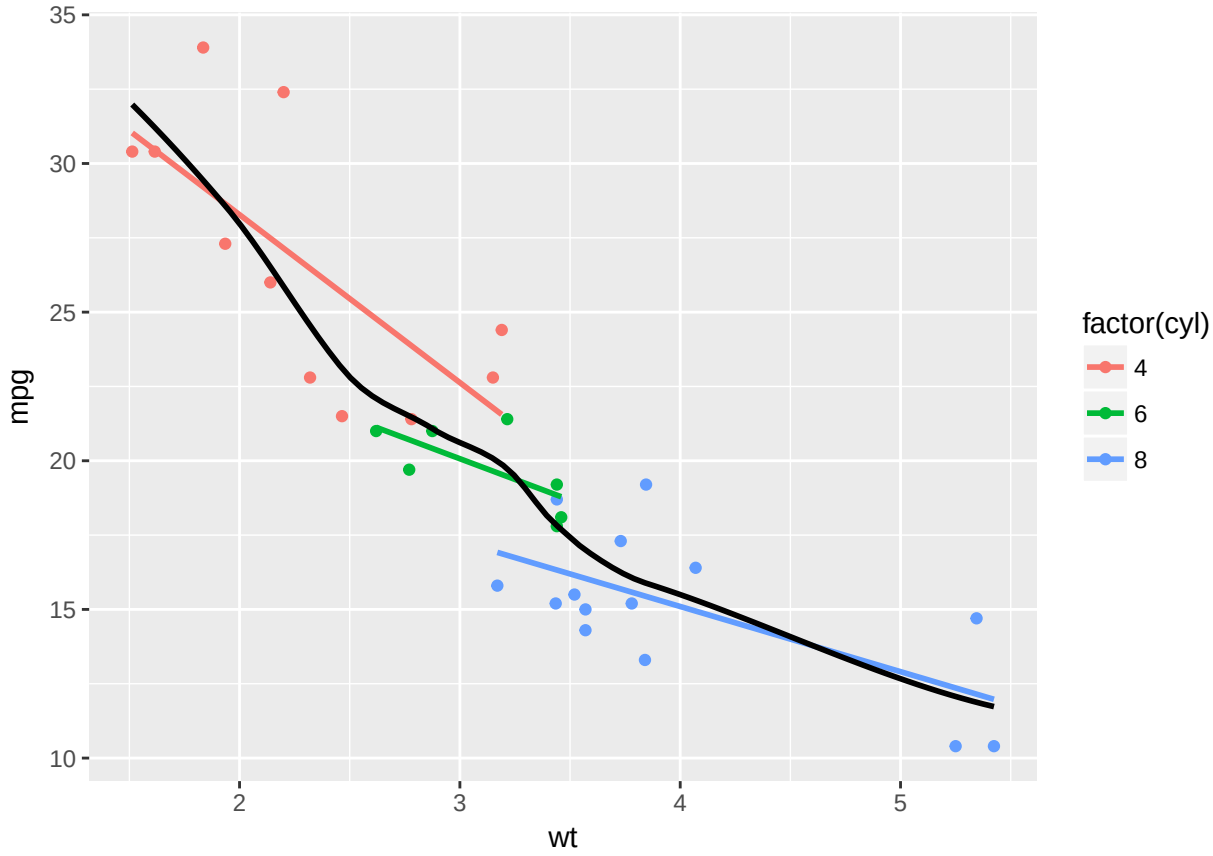
Add a scale layer: scale_color_manual() with the first argument set to “Cylinders” and values set to the predefined myColors variable.

```
# Plot 1: change the LOESS span
ggplot(mtcars, aes(x = wt, y = mpg)) +
  geom_point() +
  # Add span below
  geom_smooth(se = FALSE, span = 0.7)
```

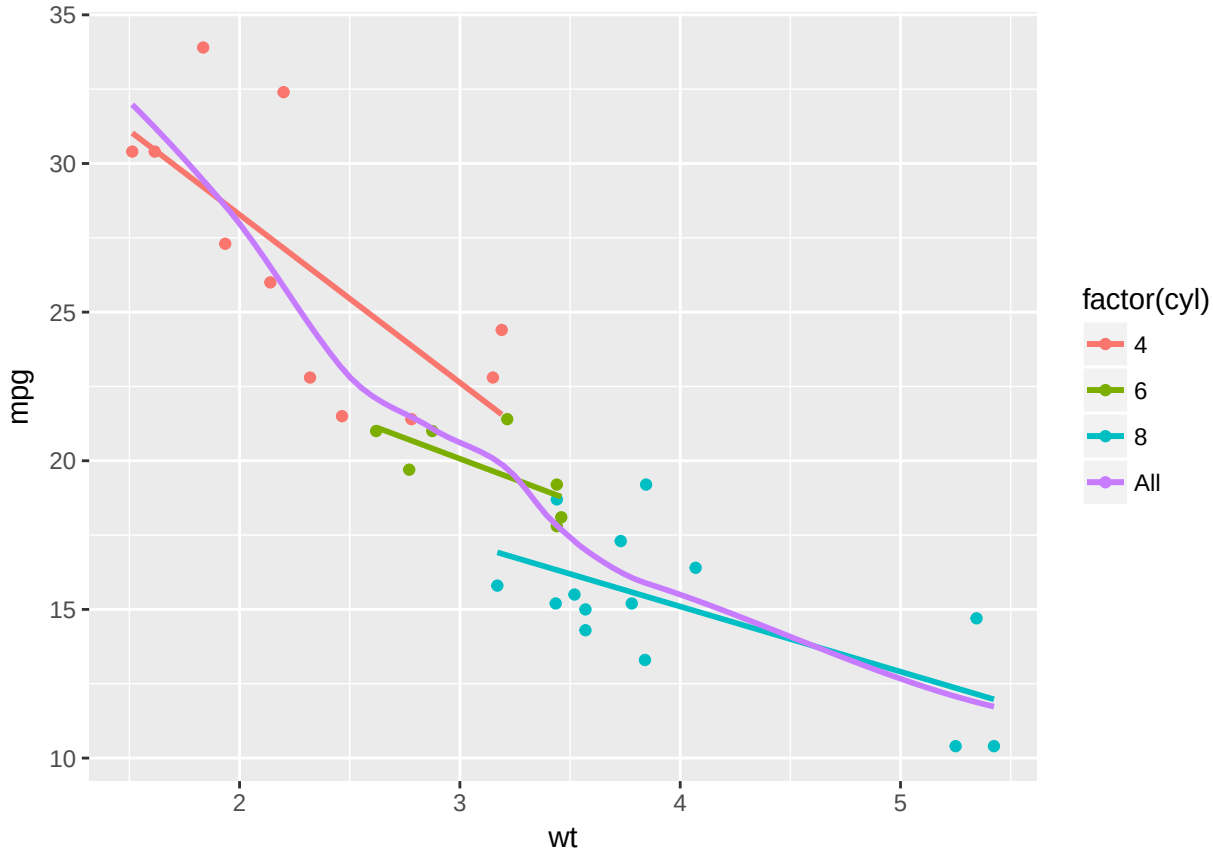
```
## `geom_smooth()` using method = 'loess'
```



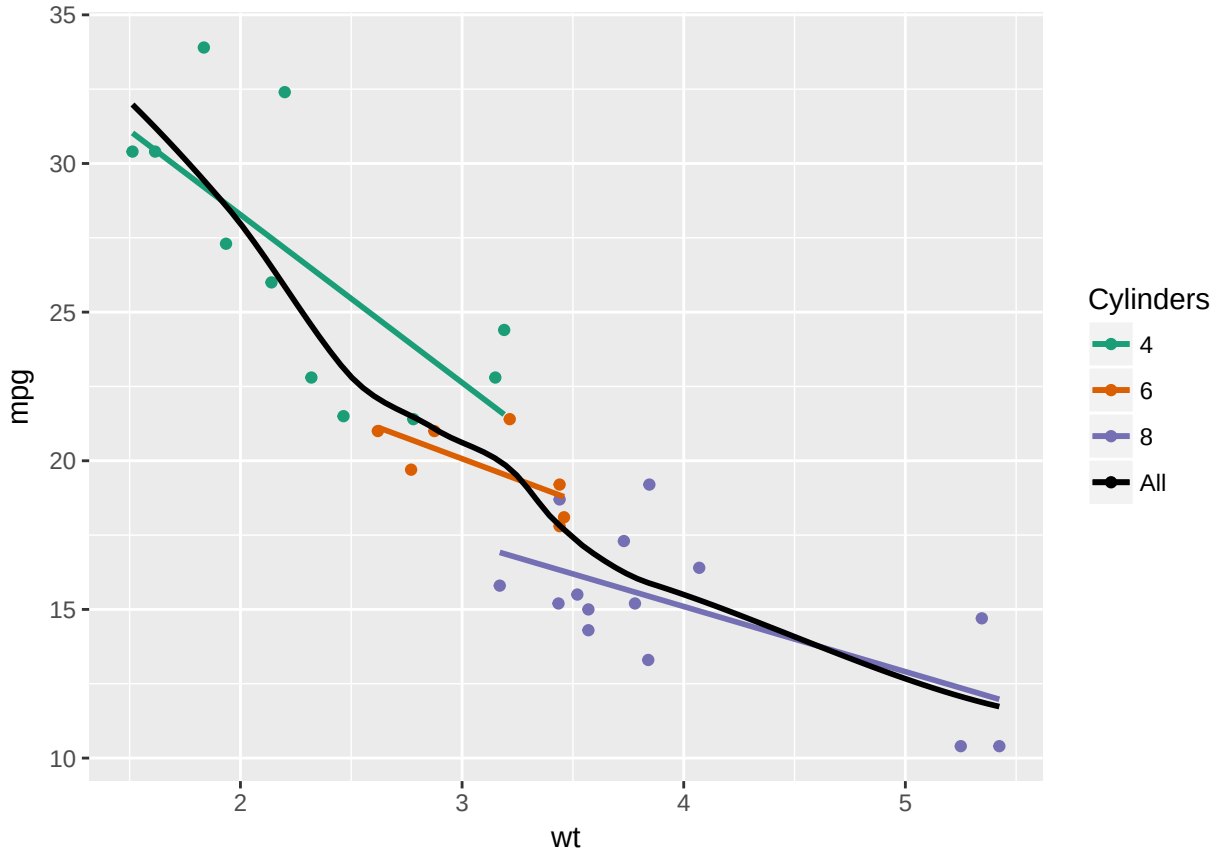

```
# Plot 2: Set the second stat_smooth() to use LOESS with a span of 0.7
ggplot(mtcars, aes(x = wt, y = mpg, col = factor(cyl))) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE) +
  # Change method and add span below
  stat_smooth(method = "loess", aes(group = 1),
              se = FALSE, col = "black", span = 0.7)
```




```
# Plot 3: Set col to "All", inside the aes layer of stat_smooth()
ggplot(mtcars, aes(x = wt, y = mpg, col = factor(cyl))) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE) +
  stat_smooth(method = "loess",
              # Add col inside aes()
              aes(group = 1, col = "All"),
              # Remove the col argument below
              se = FALSE, span = 0.7)
```




```
# Plot 4: Add scale_color_manual to change the colors
myColors <- c(brewer.pal(3, "Dark2"), "black")
ggplot(mtcars, aes(x = wt, y = mpg, col = factor(cyl))) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE, span = 0.7) +
  stat_smooth(method = "loess",
              aes(group = 1, col="All"),
              se = FALSE, span = 0.7) +
  # Add correct arguments to scale_color_manual
  scale_color_manual("Cylinders", values = myColors)
```



Modifying stat_smooth (2)

In this exercise we'll take a look at a more subtle example of defining and using linear models. ggplot2 and the Vocab data frame are already loaded for you.

INSTRUCTIONS 100 XP Plot 1: This code produces a jittered plot of vocabulary against education, variables from the Vocab data frame.

Add a stat_smooth() layer with method set to "lm". Hide the CI ribbons by using se = FALSE. Plot 2: Color by year.

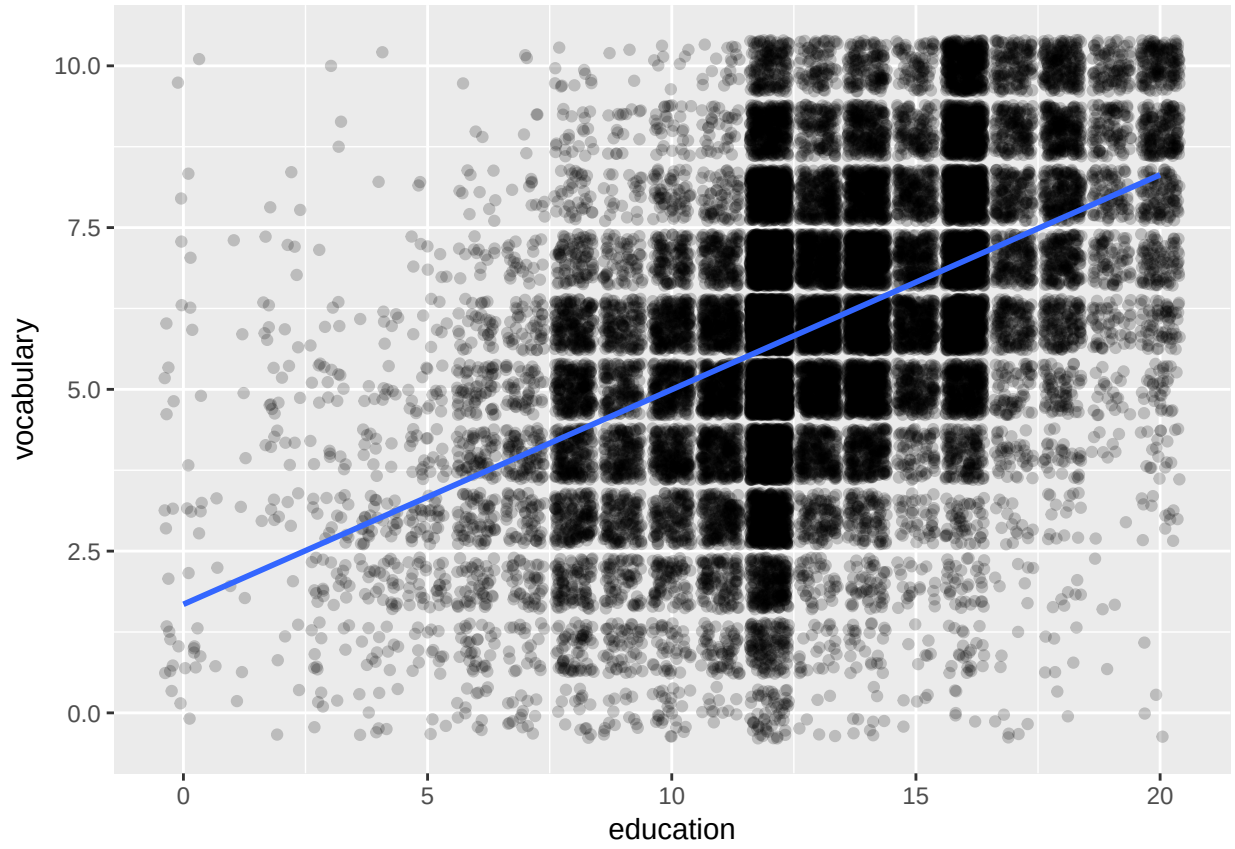
Specify the col = year aesthetic to the nested ggplot(aes()) function. To see why this works, try using only col = year, and adding points. Plot 3: Linear model for each year.

We need to specify year as a factor variable if we want to use it as a grouping variable for our linear models. Add the col = factor(year) aesthetic to the nested ggplot(aes()) function. Plot 4: Years are ordered, so use a sequential color palette.

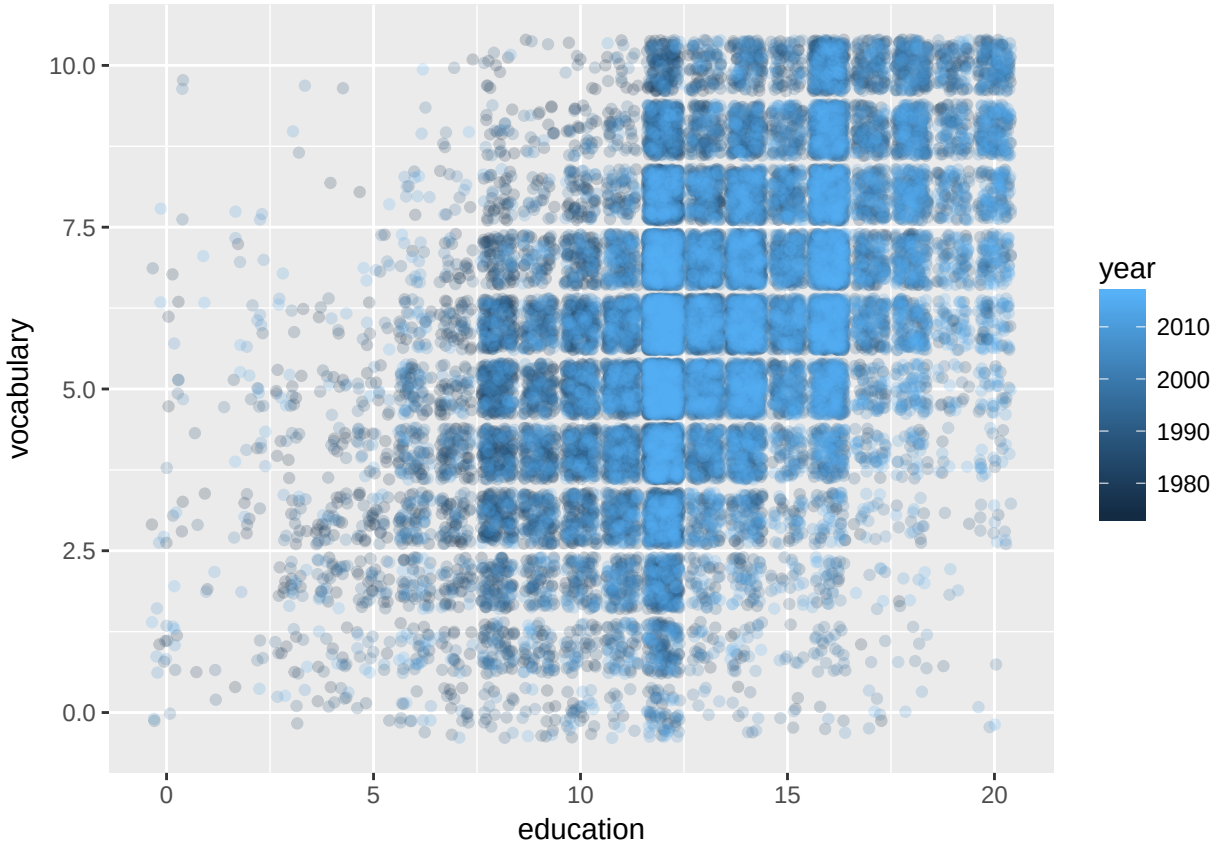
Add scale_color_brewer(). Don't add any arguments here. This results in a warning message, since the default palette, "Blues", only has 9 colors. Since we have 16 years, this is not a complete solution! Plot 5: To get the proper colors, we can use col = year, because the variable year is type integer and we want a continuous scale. However, we'll need to specify the invisible group aesthetic so that our linear models are still calculated appropriately. The scale layer, scale_color_gradientn(), has been provided for you - this allows us to map a continuous variable onto a colour scale.

Add group = factor(year) inside aes(). Inside stat_smooth(), set alpha = 0.6 and size = 2.

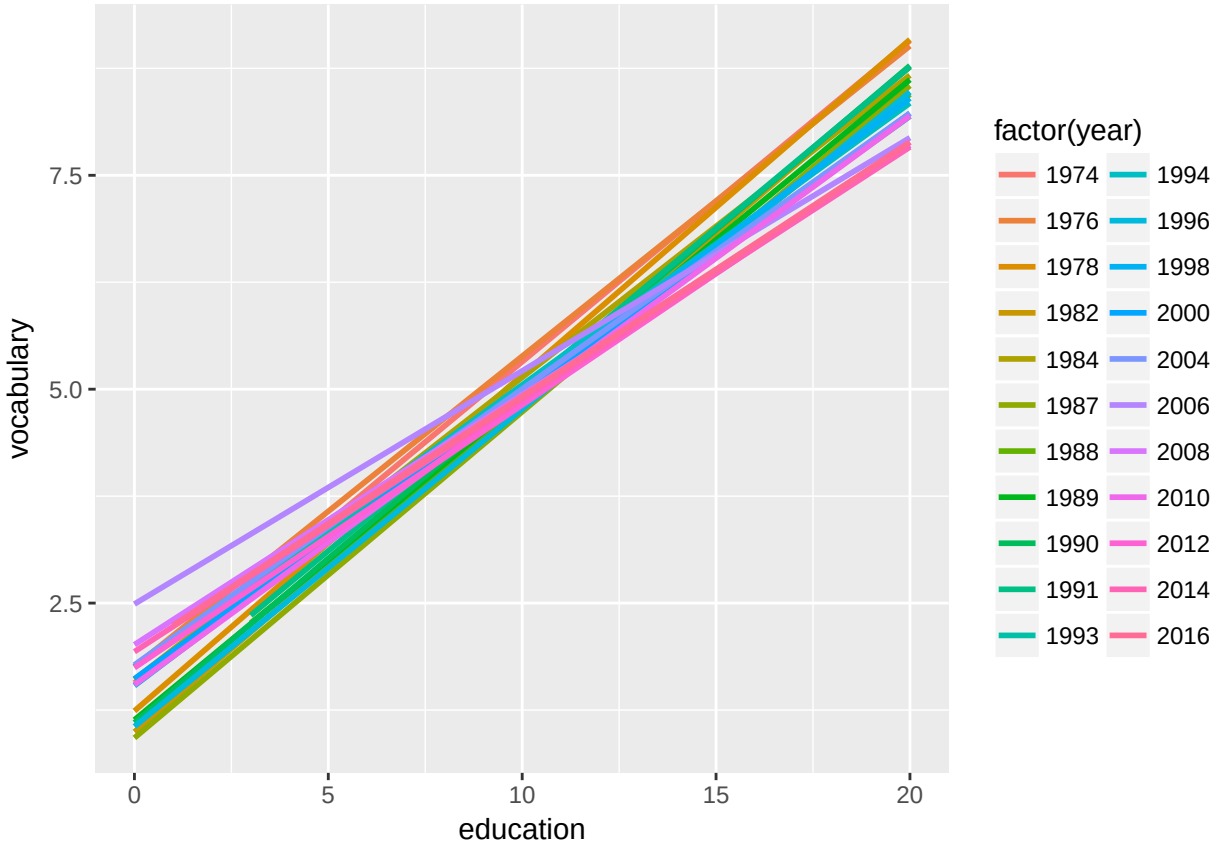
```r
# Plot 1: Jittered scatter plot, add a linear model (lm) smooth
ggplot(Vocab, aes(x = education, y = vocabulary)) +
  geom_jitter(alpha = 0.2) +
  stat_smooth(method = "lm", se = FALSE) # smooth
```




```r
# Plot 2: points, colored by year
ggplot(Vocab, aes(x = education, y = vocabulary, col = year)) +
  geom_jitter(alpha = 0.2)
```

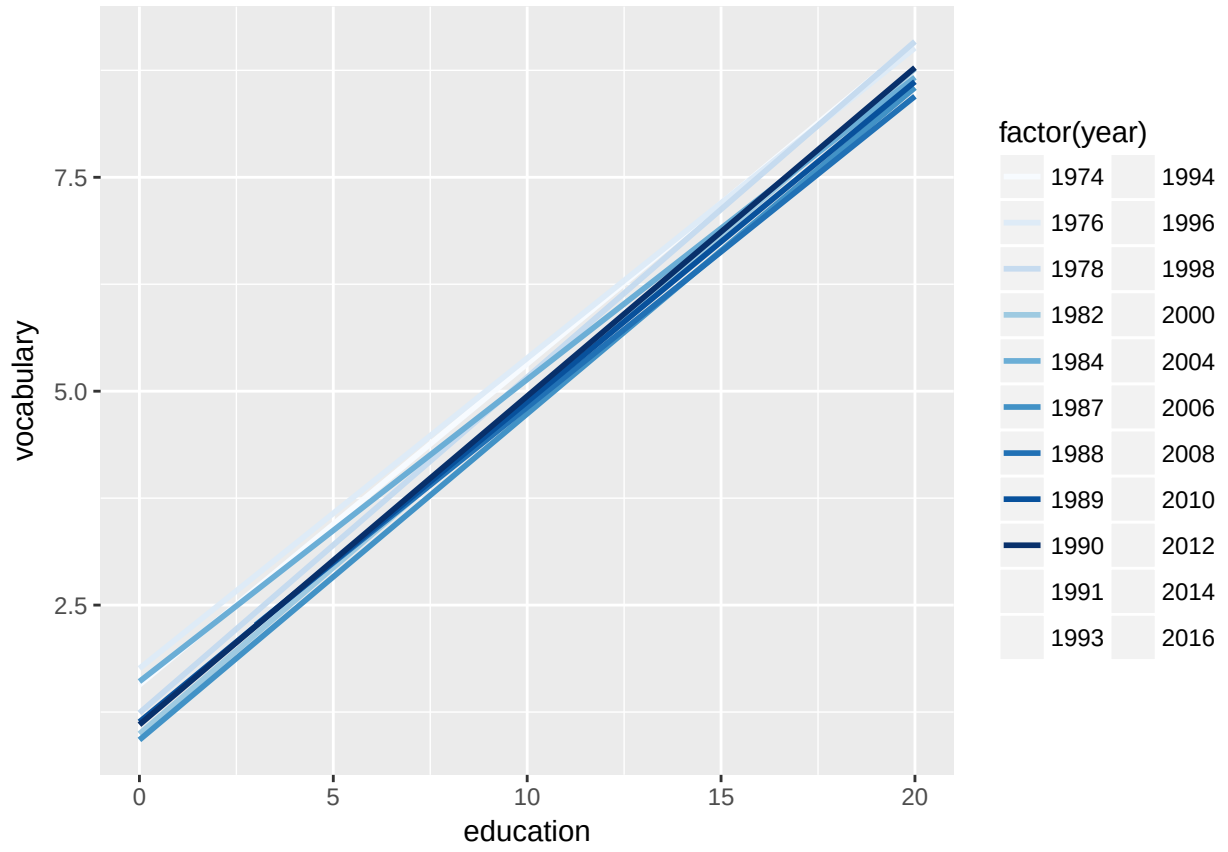



```
# Plot 3: lm, colored by year
ggplot(Vocab, aes(x = education, y = vocabulary, col = factor(year))) +
  stat_smooth(method = "lm", se = FALSE) # smooth
```

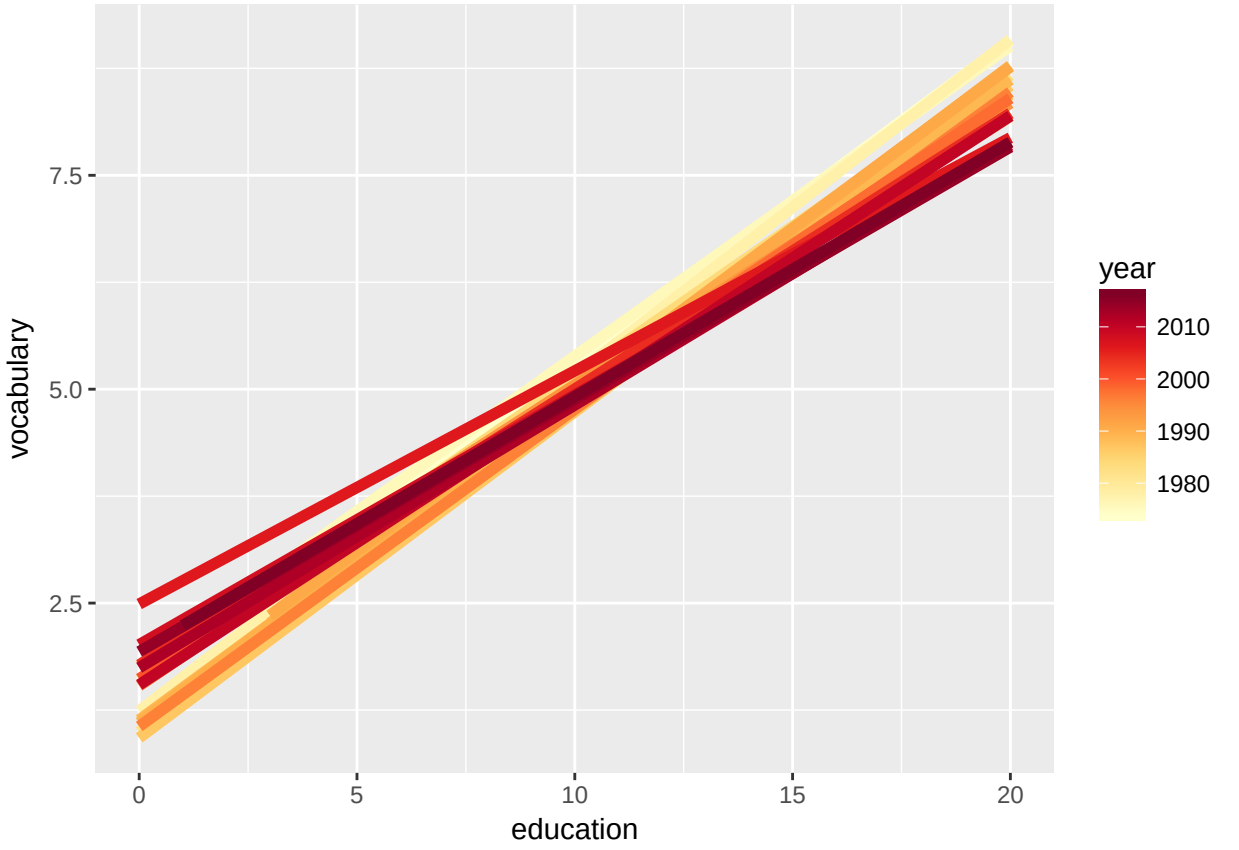



```
# Plot 4: Set a color brewer palette
ggplot(Vocab, aes(x = education, y = vocabulary, col = factor(year))) +
  stat_smooth(method = "lm", se = FALSE) +  # smooth
  scale_color_brewer()  # colors
```

```
## Warning in RColorBrewer::brewer.pal(n, pal): n too large, allowed maximum for palette Blues is 9
## Returning the palette you asked for with that many colors
```




```
# Plot 5: Add the group aes, specify alpha and size
ggplot(Vocab, aes(x = education, y = vocabulary, col = year, group = factor(year))) +
  stat_smooth(method = "lm", se = FALSE, alpha = 0.6, size = 2) +
  scale_color_gradientn(colors = brewer.pal(9, "YlOrRd"))
```



Quantiles

The previous example used the Vocab dataset and applied linear models describing vocabulary by education for different years. Here we'll continue with that example by using stat_quantile() to apply a quantile regression (method rq).

By default, the 1st, 2nd (i.e. median), and 3rd quartiles are modeled as a response to the predictor variable, in this case education. Specific quantiles can be specified with the quantiles argument.

If you want to specify many quantile and color according to year, then things get too busy. We'll explore ways of dealing with this in the next chapter.

INSTRUCTIONS

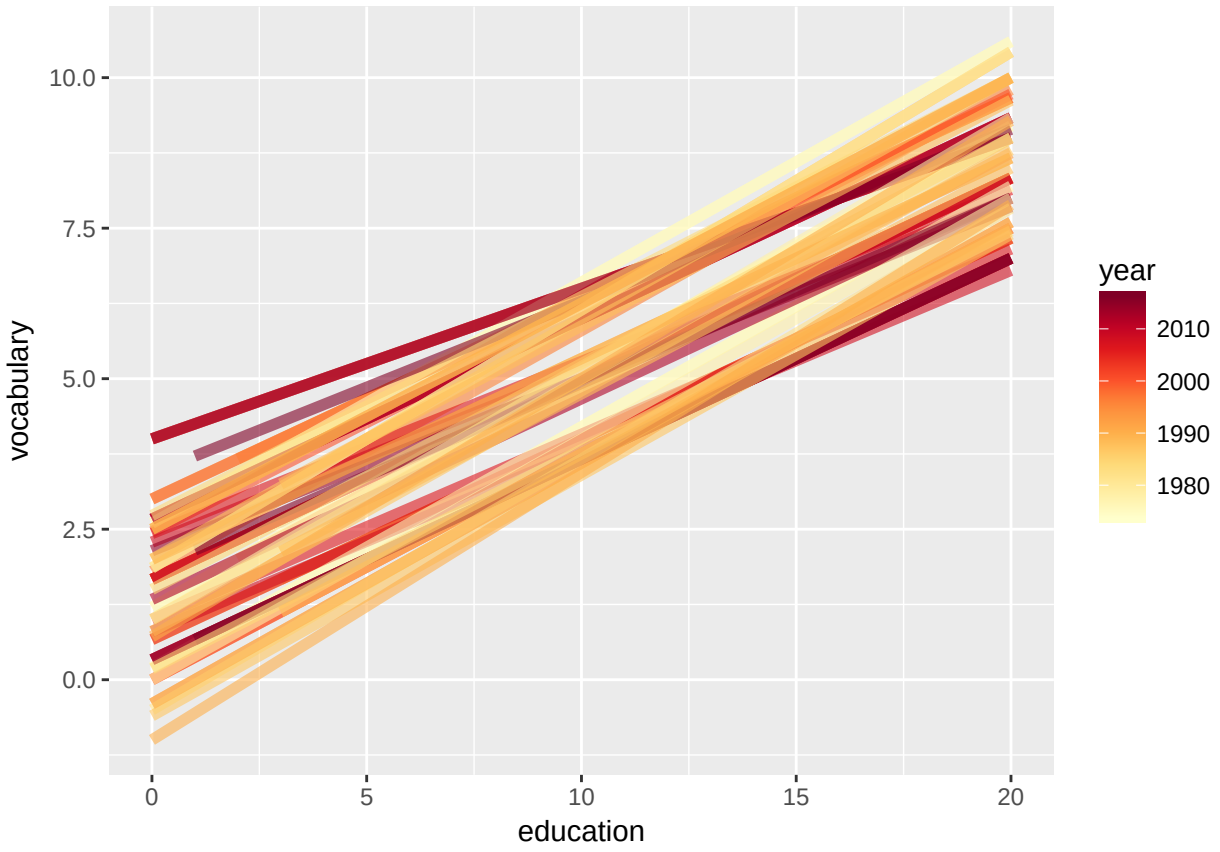
The code from the previous exercise, with the linear model and a suitable color palette, is already shown.

Update the plotting code. Change the stat function from stat_smooth() to stat_quantile(). Get rid of all the arguments except alpha and size. The resulting plot will be a mess, because there are three quartiles drawn by default. Copy the code for the previous instruction. Set the quantiles argument to 0.5 so that only the median is shown.

```
# Use stat_quantile instead of stat_smooth
ggplot(Vocab, aes(x = education, y = vocabulary, col = year, group = factor(year))) +
  stat_quantile(alpha = 0.6, size = 2) +
  scale_color_gradientn(colors = brewer.pal(9,"YlOrRd"))
```

```
## Loading required package: SparseM
```

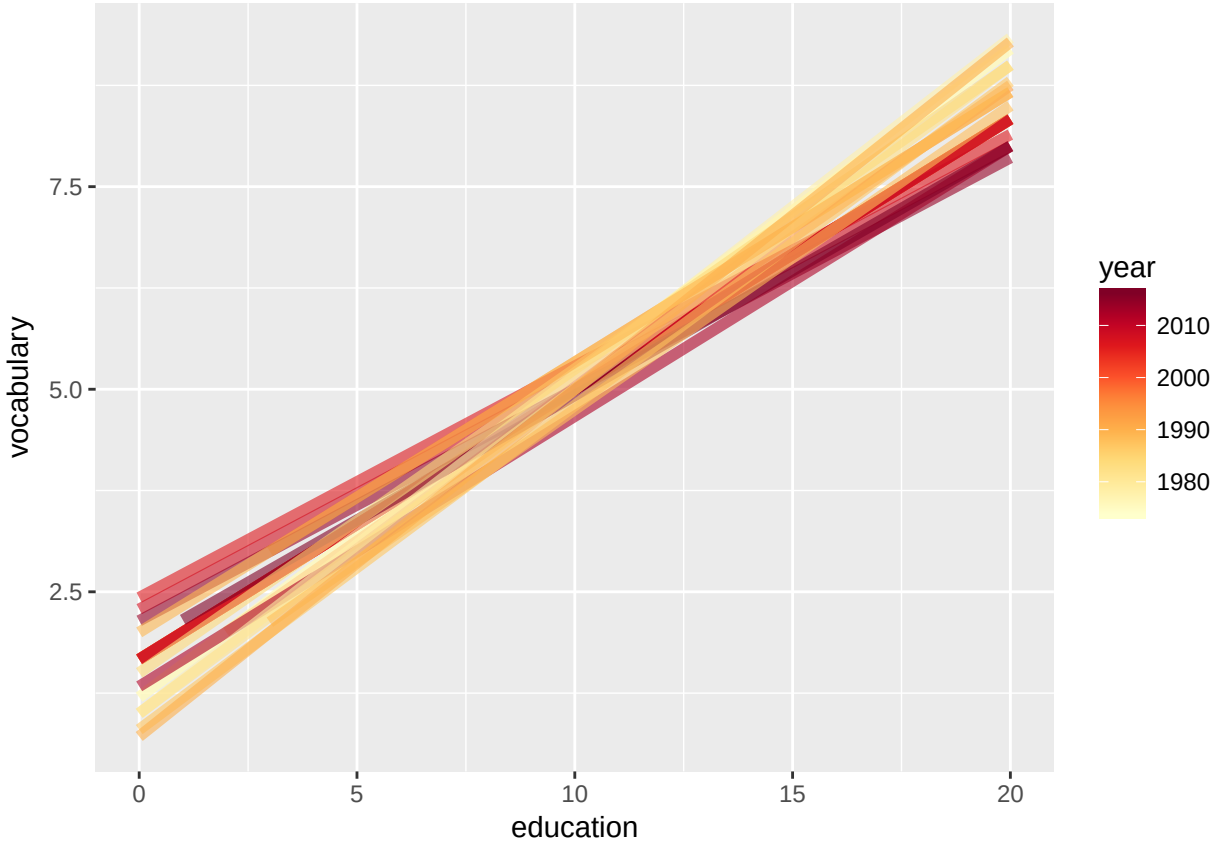
```
##
## Attaching package: 'SparseM'

## The following object is masked from 'package:base':
##
##     backsolve

##
## Attaching package: 'quantreg'

## The following object is masked from 'package:Hmisc':
##
##     latex

## The following object is masked from 'package:survival':
##
##     untangle.specials

## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x

## Warning in rq.fit.br(wx, wy, tau = tau, ...): Solution may be nonunique

## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x

## Warning in rq.fit.br(wx, wy, tau = tau, ...): Solution may be nonunique

## Smoothing formula not specified. Using: y ~ x

## Warning in rq.fit.br(wx, wy, tau = tau, ...): Solution may be nonunique


## Warning in rq.fit.br(wx, wy, tau = tau, ...): Solution may be nonunique

## Smoothing formula not specified. Using: y ~ x

## Warning in rq.fit.br(wx, wy, tau = tau, ...): Solution may be nonunique


## Warning in rq.fit.br(wx, wy, tau = tau, ...): Solution may be nonunique

## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x

## Warning in rq.fit.br(wx, wy, tau = tau, ...): Solution may be nonunique


## Warning in rq.fit.br(wx, wy, tau = tau, ...): Solution may be nonunique

## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
```




```
# Set quantile to 0.5
ggplot(Vocab, aes(x = education, y = vocabulary, col = year, group = factor(year))) +
  stat_quantile(alpha = 0.6, size = 2, quantiles = 0.5) +
  scale_color_gradientn(colors = brewer.pal(9,"YlOrRd"))
```

```
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Warning in rq.fit.br(wx, wy, tau = tau, ...): Solution may be nonunique
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Warning in rq.fit.br(wx, wy, tau = tau, ...): Solution may be nonunique
## Smoothing formula not specified. Using: y ~ x
```

```
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
```



Sum

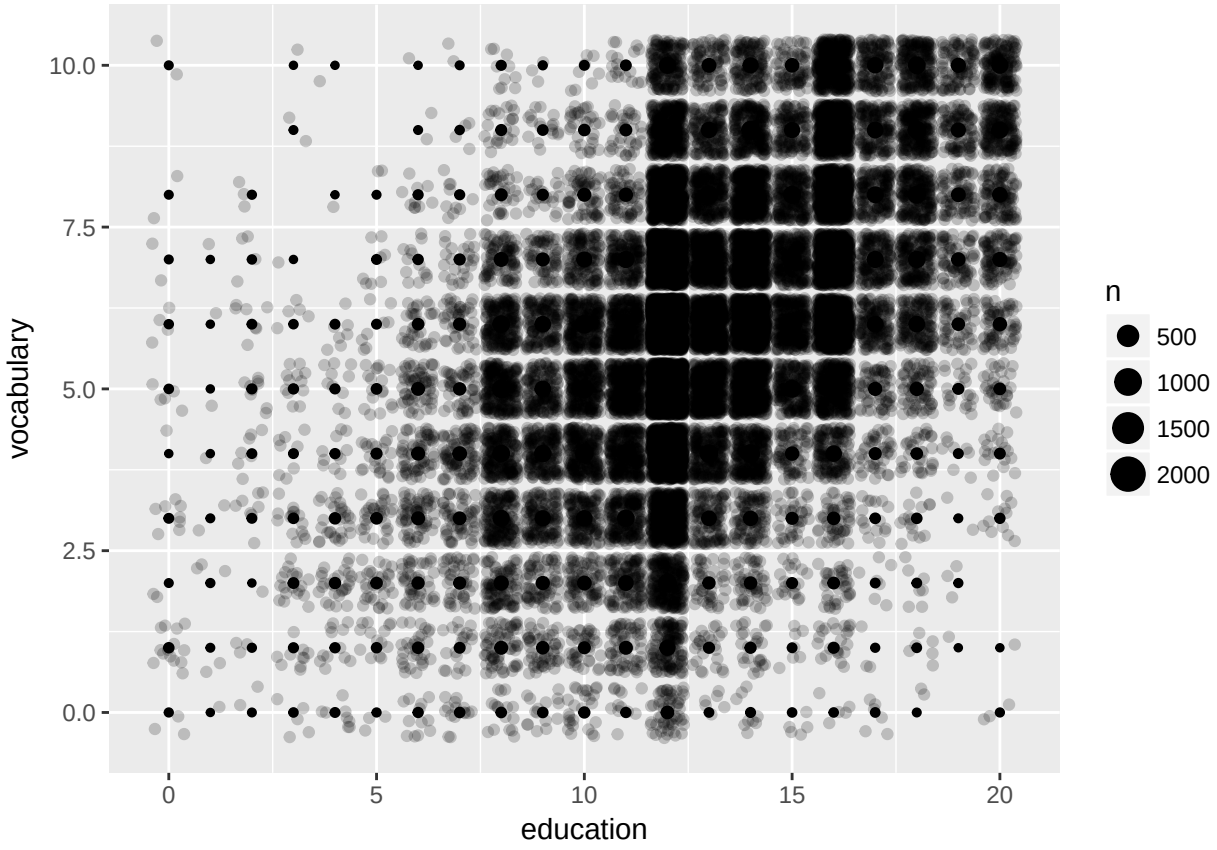
Another useful stat function is stat_sum(). This function calculates the total number of overlapping observations and is another good alternative to overplotting.

INSTRUCTIONS

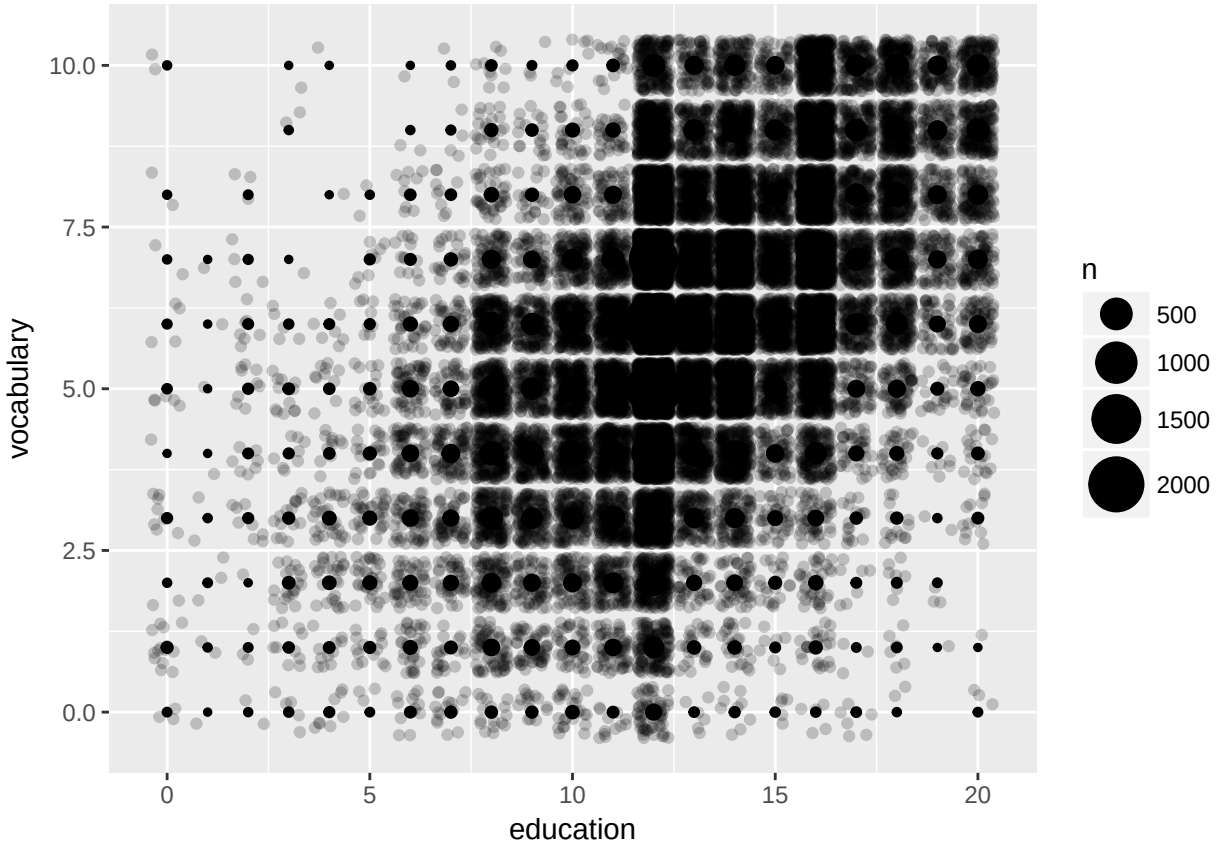
ggplot2 is already loaded. A plot showing jittered points is already provided and stored as p.

Add stat_sum() to this plotting object p. This maps the overall count of each dot onto size. You don't have to set any arguments; the aesthetics will be inherited from the base plot! Add the size scale with the generic scale_size() function. Use range to set the minimum and maximum dot sizes as c(1,10).

```
# Plot 1: Jittering only
p <- ggplot(Vocab, aes(x = education, y = vocabulary)) +
  geom_jitter(alpha = 0.2)

# Plot 2: Add stat_sum
p +
  stat_sum() # sum statistic
```




```
# Plot 3: Set size range
p +
  stat_sum() + # sum statistic
  scale_size(range = c(1, 10)) # set size scale
```



Preparations

Here we'll look at stat_summary() in action. We'll build up various plots one-by-one.

In this exercise we'll consider the preparations. That means we'll make sure the data is in the right format and that all the positions that we might use in our plots are defined. Lastly, we'll set the base layer for our plot. ggplot2 is already loaded, so you can get started straight away!

INSTRUCTIONS

Explore the structure of the mtcars dataset by executing str(mtcars). In mtcars, cyl and am are classified as continuous, but they are actually categorical. Previously we just used factor(), but here we'll modify the actual dataset. Change cyl and am to be categorical in the mtcars data frame using as.factor. Next we'll set three position objects with convenient names. This allows us to use the exact positions on multiple layers. Create: posn.d, using position_dodge() with a width of 0.1, posn.jd, using position_jitterdodge() with a jitter.width of 0.1 and a dodge.width of 0.2 posn.j, using position_jitter() with a width of 0.2. Finally, we'll make our base layers and store it in the object wt.cyl.am. Make the base call for ggplot mapping cyl to the x, wt to y, am to both col and fill. Also set group = am inside aes(). The reason for these redundancies will become clear later on.

```
# Display structure of mtcars
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
```

```
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
# Convert cyl and am to factors
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$am <- as.factor(mtcars$am)

# Define positions
posn.d <- position_dodge(width = 0.1)
posn.jd <- position_jitterdodge(jitter.width = 0.1, dodge.width = 0.2)
posn.j <- position_jitter(width = 0.2)

# Base layers
wt.cyl.am <- ggplot(mtcars, aes(x = cyl, y = wt, col = am, fill = am, group =am))
```

Plotting variations

Now that the preparation work is done, let's have a look at at stat_summary().

ggplot2 is already loaded, as is wt.cyl.am, which is defined as

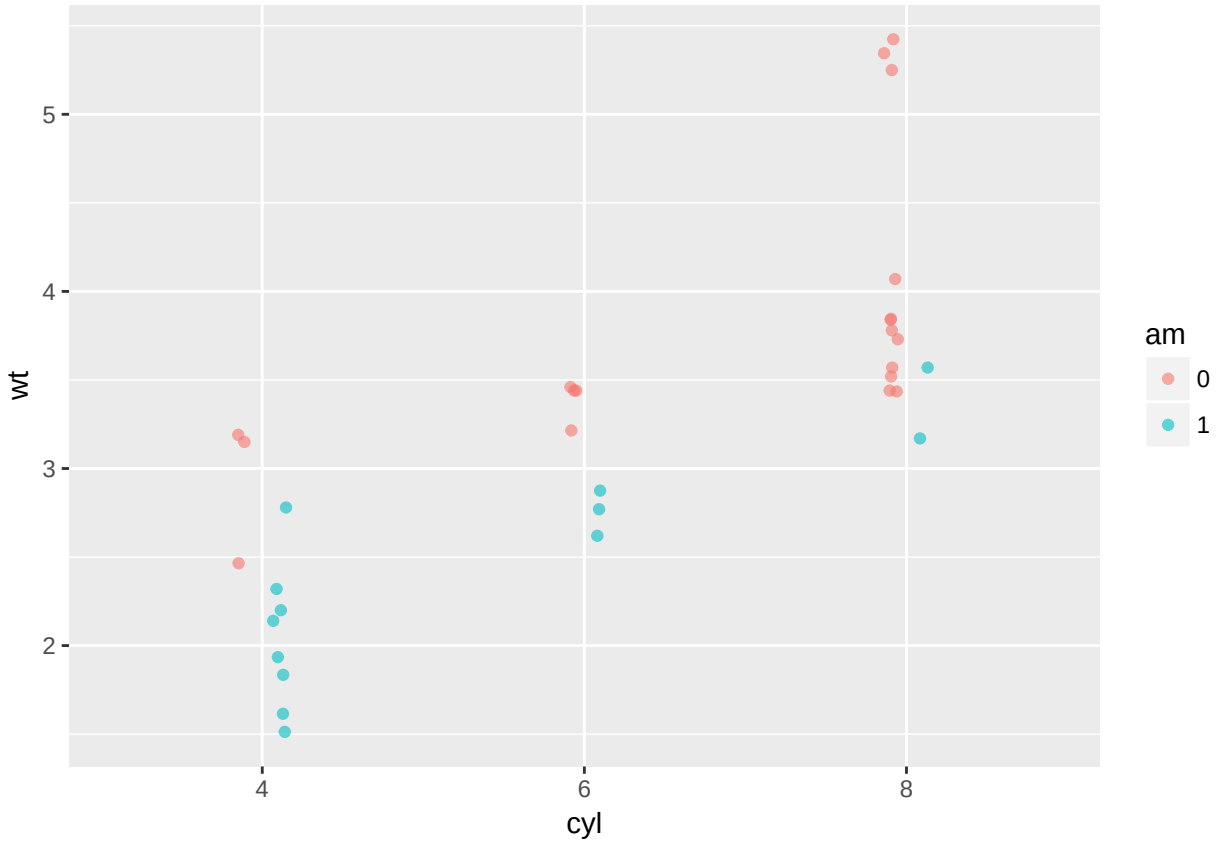
```
wt.cyl.am <- ggplot(mtcars, aes(x = cyl,  y = wt, col = am, fill = am, group = am))
```

Also all the position objects of the previous exercise, posn.d, posn.jd and posn.j, are available. For starters, Plot 1 is already coded for you.

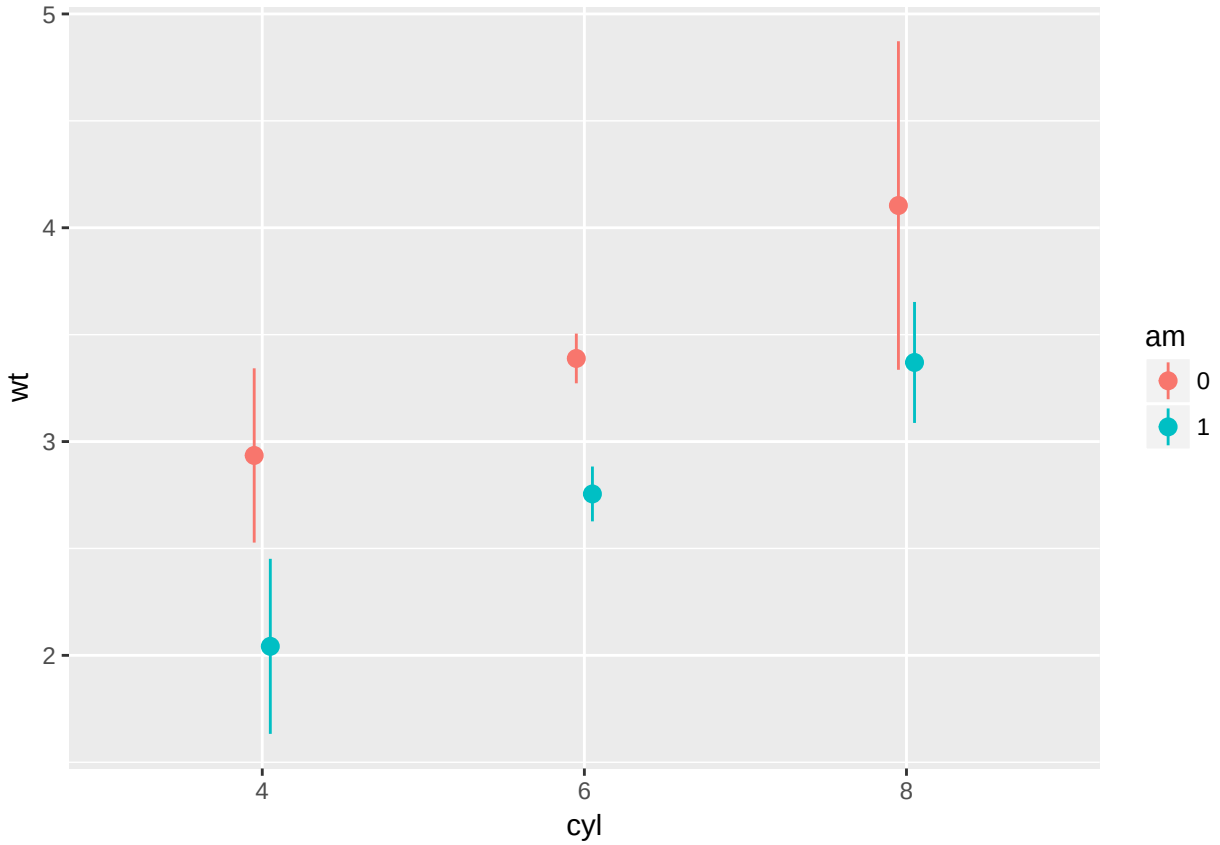
INSTRUCTIONS

Plot 2: Add a stat_summary() layer to wt.cyl.am and calculate the mean and standard deviation as we did in the video: set fun.data to mean_sdl and specify fun.args to be list(mult = 1). Set the position argument to posn.d. Plot 3: Repeat the previous plot, but use the 95% confidence interval instead of the standard deviation. You can use mean_cl_normal instead of mean_sdl this time. There's no need to specify fun.args in this case. Again, set position to posn.d. The above plots were simple because they implicitly used a default geom, which is geom_pointrange(). For Plot 4, fill in the blanks to calculate the mean and standard deviation separately with two stat_summary() functions: For the mean, use geom = "point" and set fun.y = mean. This time you should use fun.y because the point geom uses the y aesthetic behind the scenes. Add error bars with another stat_summary() function. Set geom = "errorbar" to get the real "T" tips. Set fun.data = mean_sdl.

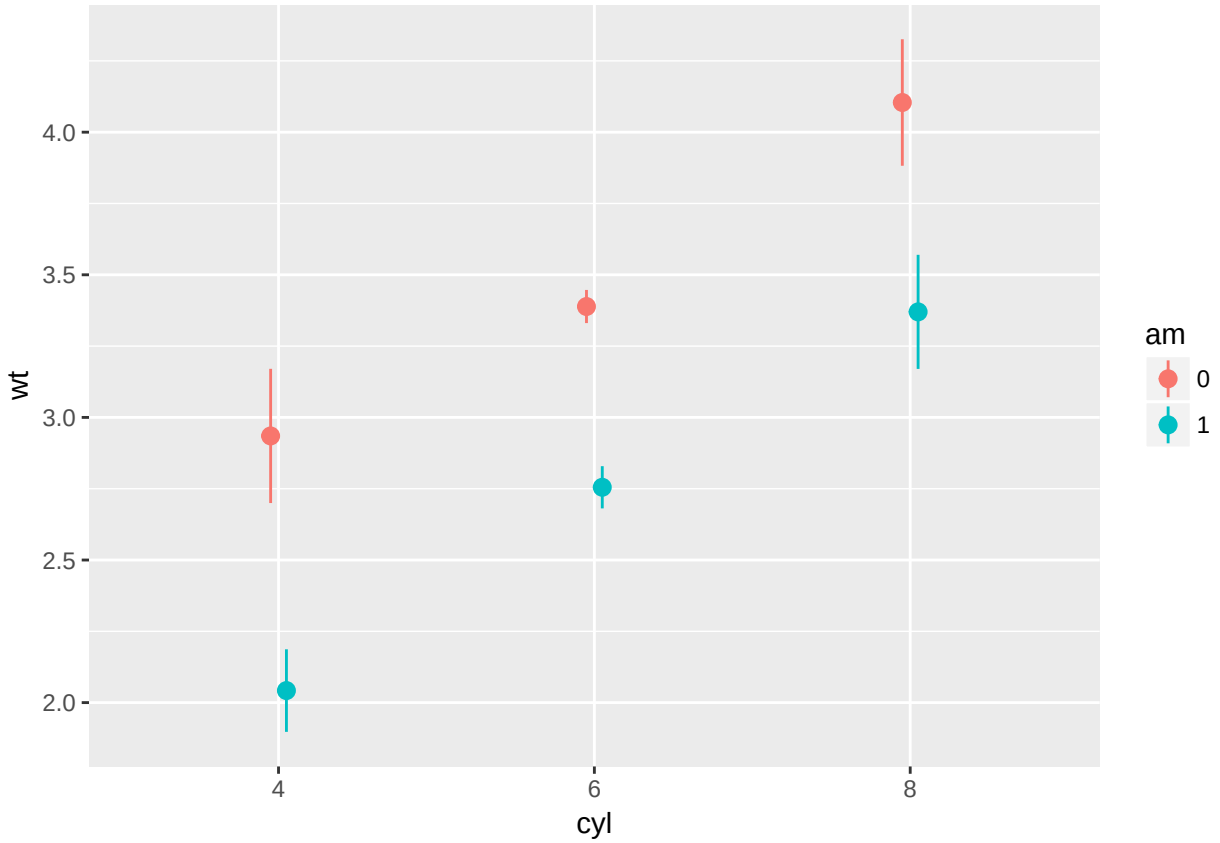
```
# wt.cyl.am, posn.d, posn.jd and posn.j are available

# Plot 1: Jittered, dodged scatter plot with transparent points
wt.cyl.am +
  geom_point(position = posn.jd, alpha = 0.6)
```



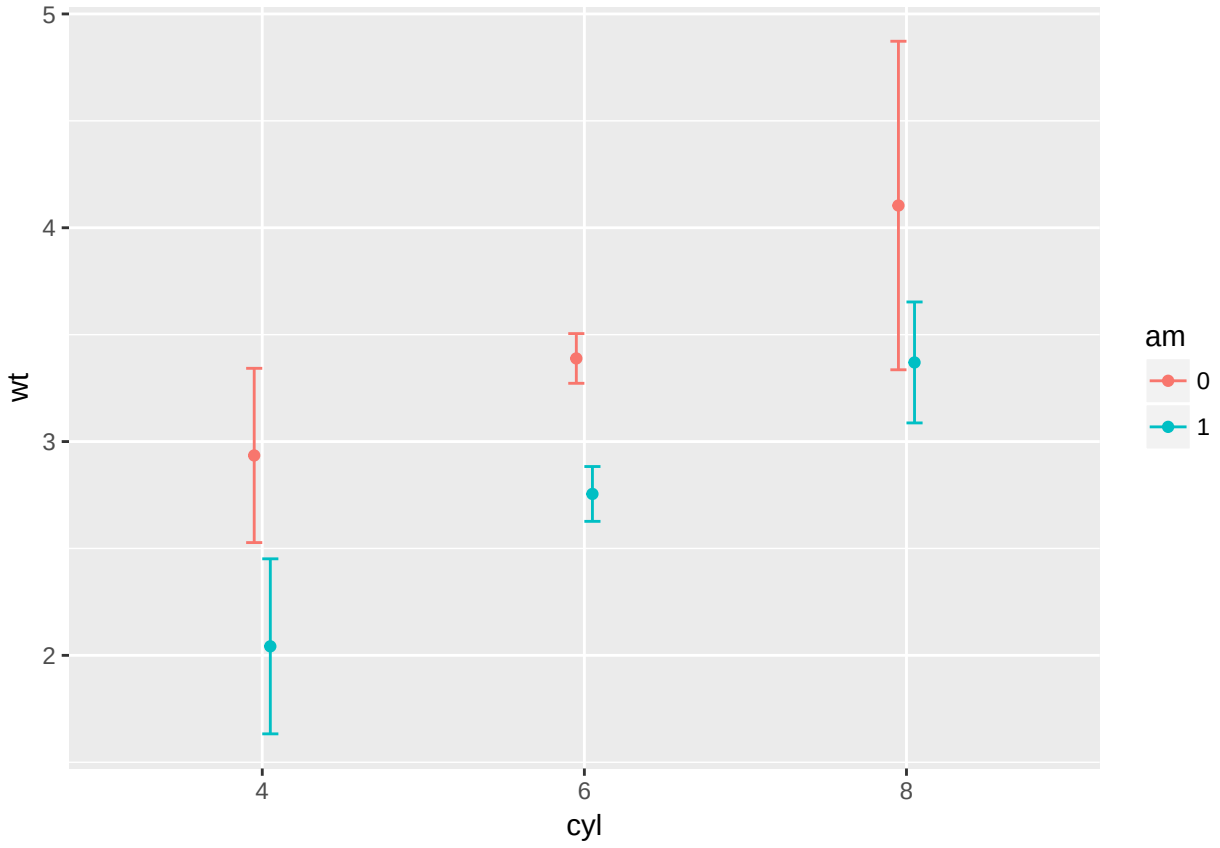

```
# Plot 2: Mean and SD - the easy way
wt.cyl.am +
  stat_summary(fun.data = mean_sdl, fun.args = list(mult = 1), position = posn.d)
```




```
# Plot 3: Mean and 95% CI - the easy way
wt.cyl.am +
  stat_summary(fun.data = mean_cl_normal, fun.args = list(mult = 1), position = posn.d)
```




```
# Plot 4: Mean and SD - with T-tipped error bars - fill in ___
wt.cyl.am +
  stat_summary(geom = "point", fun.y = mean,
               position = posn.d) +
  stat_summary(geom = "errorbar", fun.data = mean_sdl,
               position = posn.d, fun.args = list(mult = 1), width = 0.1)
```



Custom Functions

In the video we saw that the only difference between ggplot2::mean_sdl() and Hmisc::smean.sdl() is the naming convention. In order to use the results of a function directly in ggplot2 we need to ensure that the names of the variables match the aesthetics needed for our respective geoms.

Here we'll create two new functions in order to create the plot shown in the viewer. One function will measure the full range of the dataset and the other will measure the interquartile range.

A play vector, xx, has been created for you. Execute

```
mean_sdl(xx, mult = 1)
```

in the R Console and consider the format of the output. You'll have to produce functions which return similar outputs.

INSTRUCTIONS

First, change the arguments ymin and ymax inside the data.frame() call of gg_range().

ymin should be the minimum of x

ymax should be the maximum of x

Use min() and max(). Watch out, naming is important here. gg_range(xx) should now generate the required output.

Next, change the arguments y, ymin and ymax inside the data.frame() call of med_IQR().

y should be the median of x

ymin should be the first quartile

ymax should be the 3rd quartile.

You should use median() and quantile(). For example, quantile() can be used as follows to give the first quartile: quantile(x)[2]. med_IQR(xx) should now generate the required output.

```
#Define xx
xx <- seq(1:100)

# Play vector xx is available

# Function to save range for use in ggplot
gg_range <- function(x) {
  # Change x below to return the instructed values
  data.frame(ymin = min(x), # Min
             ymax = max(x)) # Max
}

gg_range(xx)
```

```
##   ymin ymax
## 1    1  100
```

```
# Required output
#   ymin ymax
# 1    1  100

# Function to Custom function
med_IQR <- function(x) {
  # Change x below to return the instructed values
  data.frame(y = median(x), # Median
             ymin = quantile(x)[2], # 1st quartile
             ymax = quantile(x)[4])  # 3rd quartile
}

med_IQR(xx)
```

```
##        y  ymin  ymax
## 25% 50.5 25.75 75.25
```

```
# Required output
#        y  ymin  ymax
# 25% 50.5 25.75 75.25
```

Custom Functions (2)

In the last exercise we created functions that will allow us to plot the so-called five-number summary (the minimum, 1st quartile, median, 3rd quartile, and the maximum). Here, we'll implement that into a unique plot type.

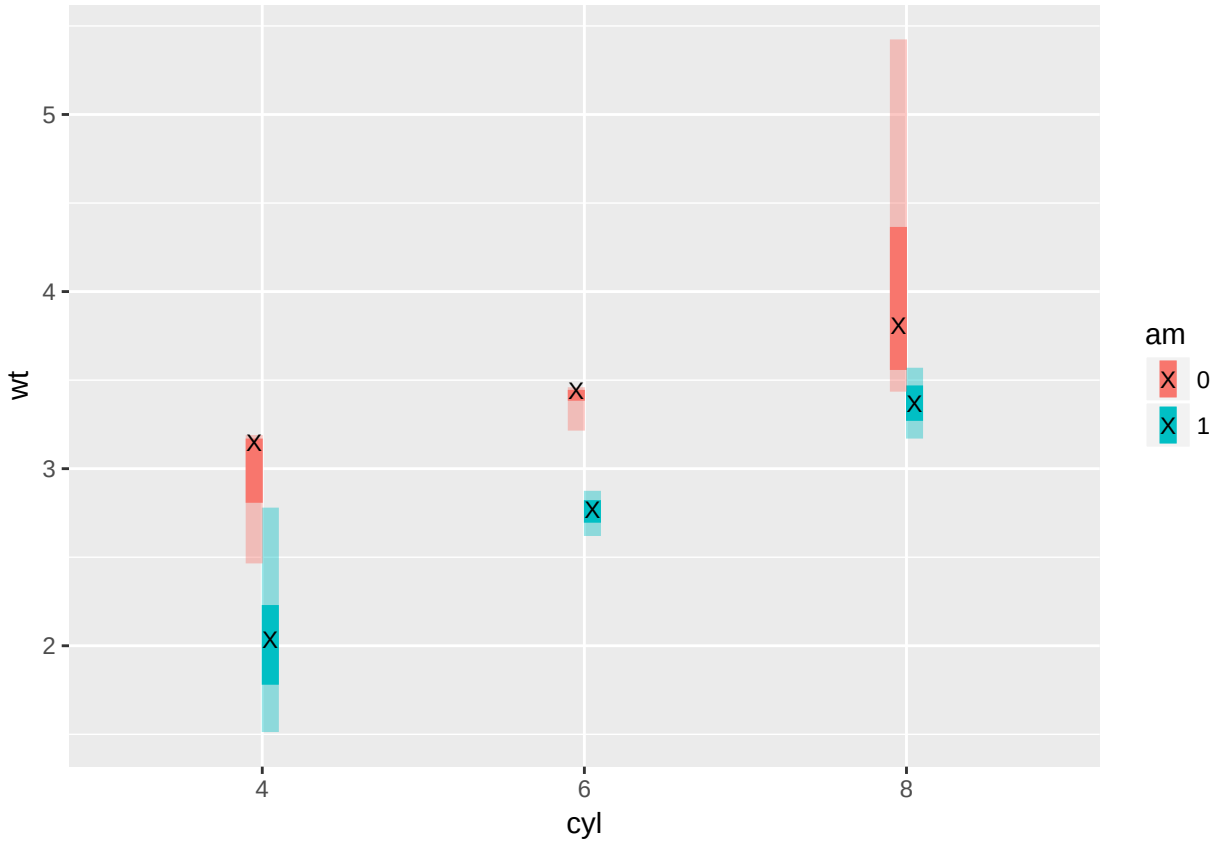
All the functions and objects from the previous exercise are available including the updated mtcars data frame, the position object posn.d, the base layers wt.cyl.am and the functions med_IQR() and gg_range().

The plot you'll end up with at the end of this exercise is shown on the right. When using stat_summary() recall that the fun.data argument requires a properly labelled 3-element long vector, which we saw in the previous exercises. The fun.y argument requires only a 1-element long vector.

INSTRUCTIONS 100 XP Complete the given stat_summary() functions, don't change the predefined arguments:

The first stat_summary() layer should have geom set to "linerange". fun.data argument should be set to med_IQR, the function you used in the previous exercise. The second stat_summary() layer also uses the "linerange" geom. This time fun.data should be gg_range, the other function you created. Also set alpha = 0.4. For the last stat_summary() layer, use geom = "point". The points should have col "black" and shape "X".

```
# The base ggplot command; you don't have to change this
wt.cyl.am <- ggplot(mtcars, aes(x = cyl,y = wt, col = am, fill = am, group = am))

# Add three stat_summary calls to wt.cyl.am
wt.cyl.am +
  stat_summary(geom = "linerange", fun.data = med_IQR,
               position = posn.d, size = 3) +
  stat_summary(geom = "linerange", fun.data = gg_range,
               position = posn.d, size = 3,
               alpha = 0.4) +
  stat_summary(geom = "point", fun.y = median,
               position = posn.d, size = 3,
               col = "black", shape = "X")
```



Chapter 2: Coordinates & Facets

The Coordinates and Facets layers offer specific and very useful tools for efficiently and accurately communicating data. In this chapter we'll look at the various ways of effectively using these two layers.

Zooming In

Aspect Ratio

Pie Charts

Facets: the basics

Many variables

Dropping levels

Chapter 3: Themes

Now that you've built high-quality plots, it's time to make them pretty. This is the last step in the data viz process. The Themes layer will enable you to make publication quality plots directly in R.

Rectangles

Lines

Text

Legends

Positions

Updating Themes

Exploring ggthemes

Chapter 4: Best Practices

Once you have the technical skill to make great visualizations, it's important that you make them as meaningful as possible. In this chapter we'll go over three plot types that are mostly discouraged in the data viz community - heat maps, pie charts and dynamite plots. We'll understand what the problems are with these plots and what the alternatives are.

Bar Plots (1)

Bar Plots (2)

Bar Plots (3)

Pie Charts (1)

Pie Charts (2)

Plot Matrix (1)

Plot Matrix (2)

Heat Maps

Heat Maps Alternatives (1)

Heat Maps Alternatives (2)

Chapter 5: Case Study

In this case study, we'll explore the large, publicly available California Health Interview Survey dataset from 2009. We'll go step-by-step through the development of a new plotting method - a mosaic plot - combining statistics and flexible visuals. At the end, we'll generalize our new plotting method to use on a variety of datasets we've seen throughout the first two courses.

Exploring Data

Unusual Values

Default Binwidths

Data Cleaning

Multiple Histograms

Alternatives

Do Things Manually

Marimekko/Mosaic Plot

Adding statistics

Adding text

Generalizations