Assignment3

Ella Buxton

2023-09-07

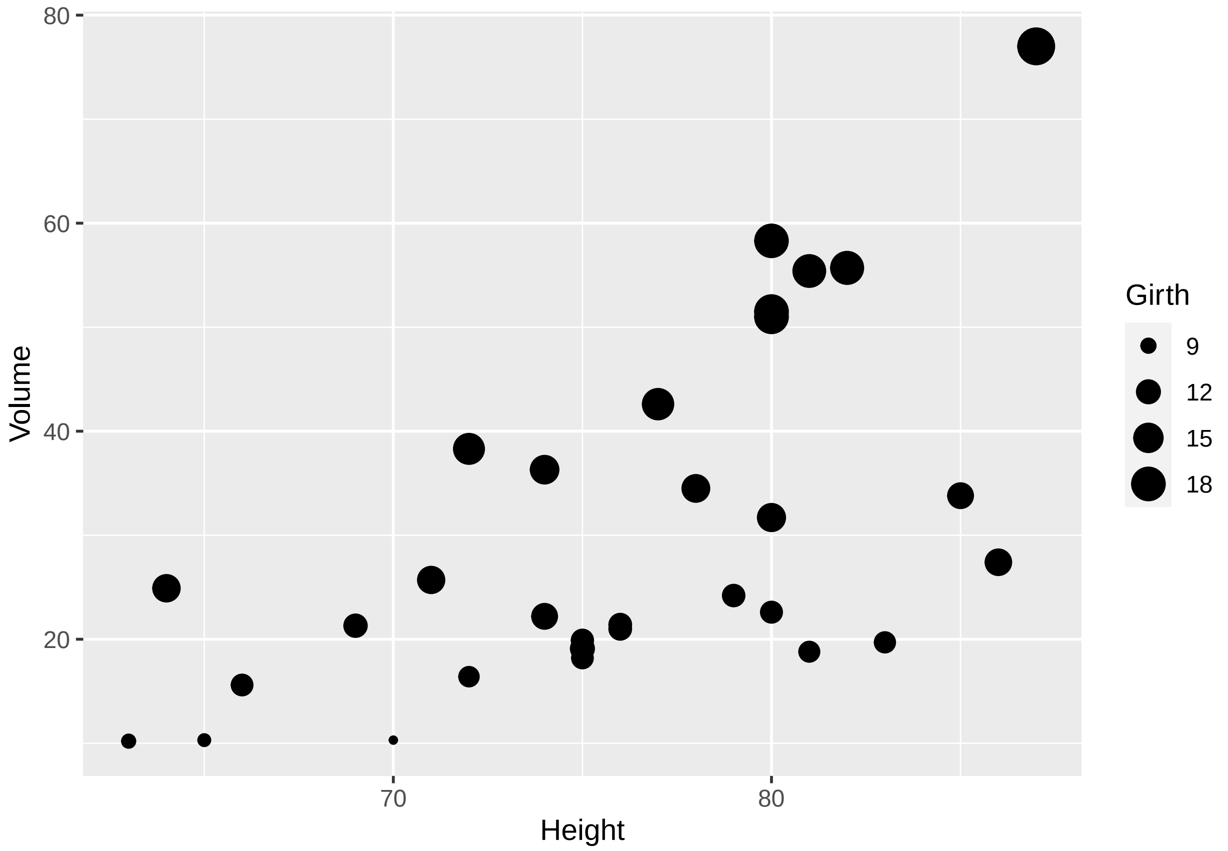
Exercise 1

1. Examine the dataset `trees`, which should already be pre-loaded. Look at the help file using `?trees` for more information about this data set. We wish to build a scatter plot that compares the height and girth of these cherry trees to the volume of lumber that was produced.

    a) Create a graph using `ggplot2` with Height on the x-axis, Volume on the y-axis, and Girth as the either the size of the data point or the color of the data point. Which do you think is a more intuitive representation?

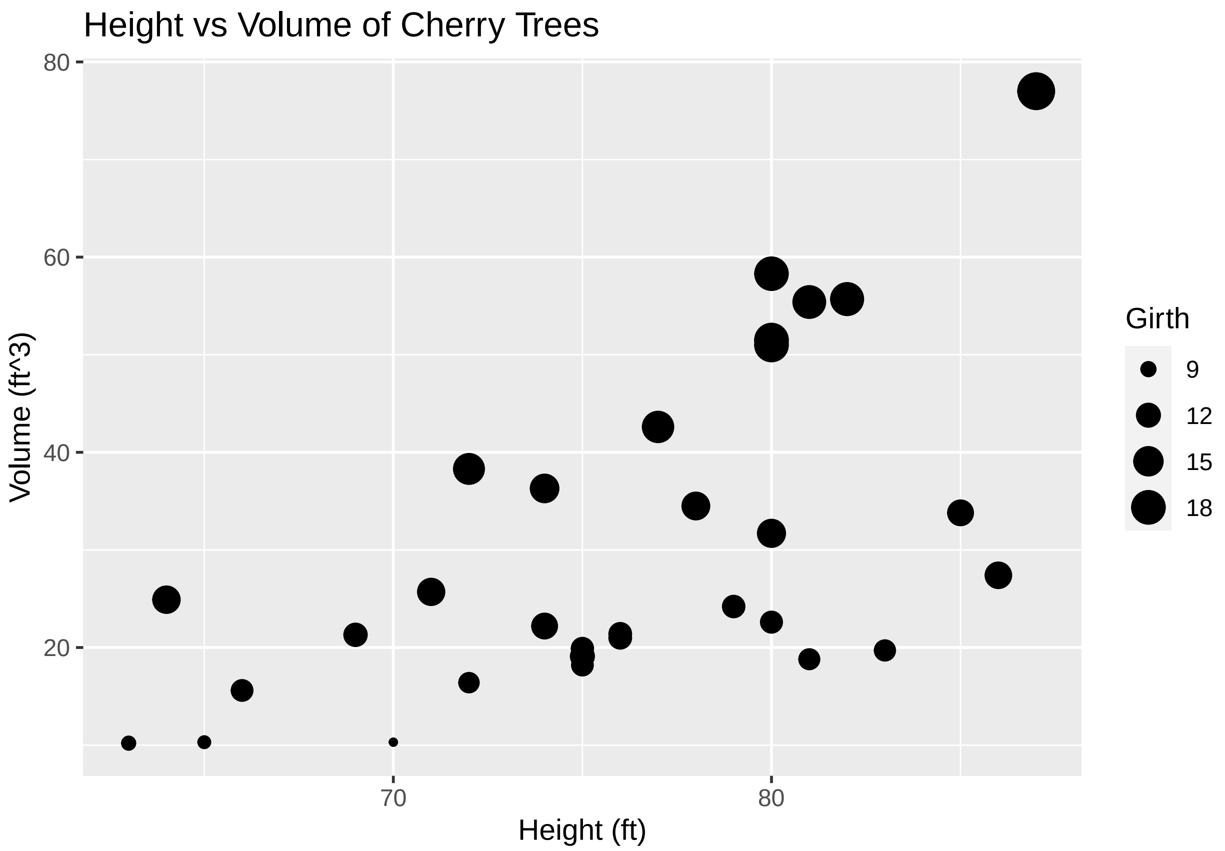
    I think that the point size is a more intuitive representation because the color was more difficult to differentiate between.

```
ggplot( trees, aes(x=Height, y=Volume)) + # scatterplot
  geom_point(aes(size=Girth)) # size of points
```



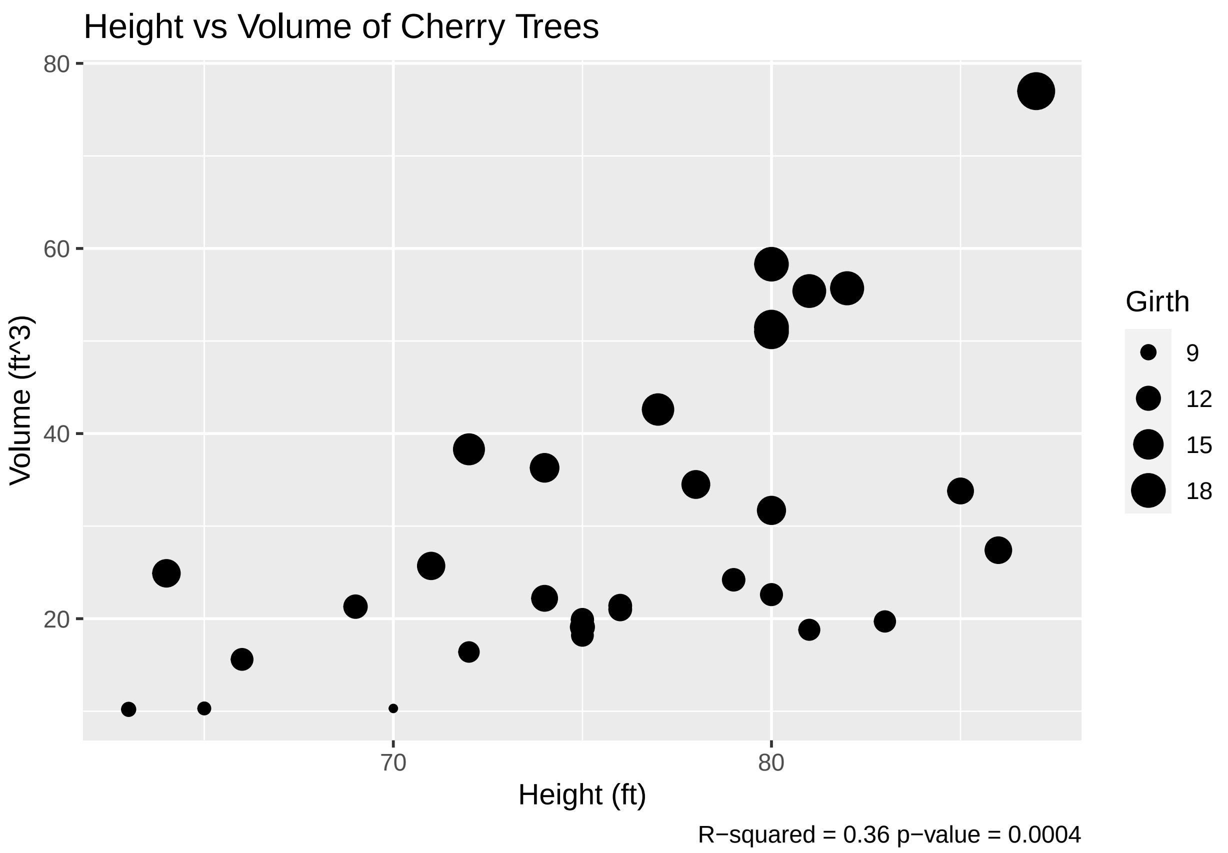
b) Add appropriate labels for the main title and the x and y axes.

```r
ggplot( trees, aes(x=Height, y=Volume)) + # scatterplot
  geom_point(aes(size=Girth)) + # point size
  labs( title= 'Height vs Volume of Cherry Trees') + # labels
  labs(x="Height (ft)", y="Volume (ft^3)" )
```



c) The R-squared value for a regression through these points is 0.36 and the p-value for the statistical significance of height is 0.00038. Add text labels "R-squared = 0.36" and "p-value = 0.0004" somewhere on the graph.

```r
ggplot( trees, aes(x=Height, y=Volume)) + # scatterplot
  geom_point(aes(size=Girth)) + # point size
  labs( title= 'Height vs Volume of Cherry Trees') + # labels
  labs(x="Height (ft)", y="Volume (ft^3)" ) +
  labs(caption= "R-squared = 0.36 p-value = 0.0004")
```



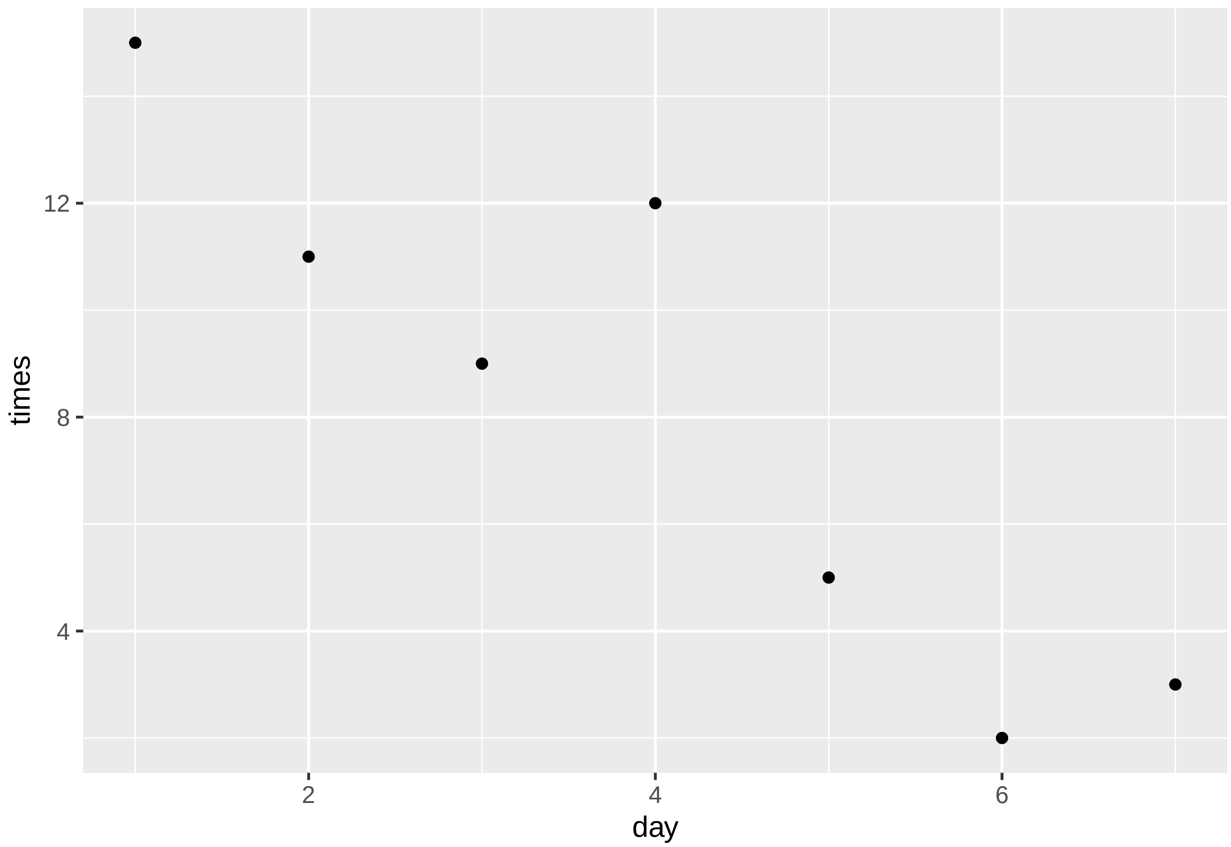
Exercise 2

2. Consider the following small dataset that represents the number of times per day my wife played "Ring around the Rosy" with my daughter relative to the number of days since she has learned this game. The column `yhat` represents the best fitting line through the data, and `lwr` and `upr` represent a 95% confidence interval for the predicted value on that day. *Because these questions ask you to produce several graphs and evaluate which is better and why, please include each graph and response with each sub-question.*

```
Rosy <- data.frame(
  times = c(15, 11, 9, 12, 5, 2, 3),
  day   = 1:7,
  yhat  = c(14.36, 12.29, 10.21, 8.14, 6.07, 4.00,  1.93),
  lwr   = c( 9.54,  8.5,   7.22, 5.47, 3.08, 0.22, -2.89),
  upr   = c(19.18, 16.07, 13.2, 10.82, 9.06, 7.78,  6.75))
```

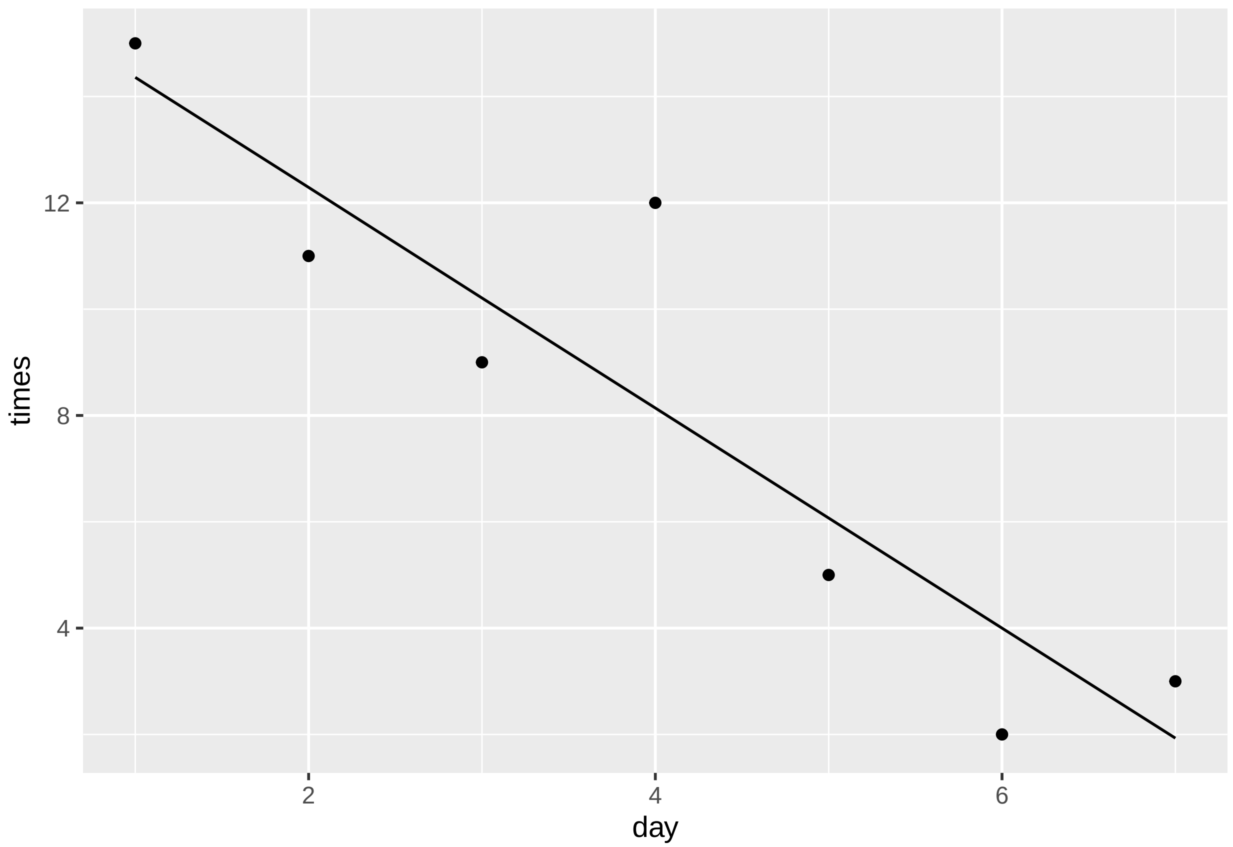
a) Using `ggplot()` and `geom_point()`, create a scatterplot with `day` along the x-axis and `times` along the y-axis.

```
Rosy <- data.frame(
  times = c(15, 11, 9, 12, 5, 2, 3),
  day   = 1:7,
  yhat  = c(14.36, 12.29, 10.21, 8.14, 6.07, 4.00,  1.93),
  lwr   = c( 9.54,  8.5,   7.22, 5.47, 3.08, 0.22, -2.89),
      upr   = c(19.18, 16.07, 13.2, 10.82, 9.06, 7.78,  6.75))
ggplot( data=Rosy, aes(x=day, y=times)) + # scatterplot
  geom_point()
```



b) Add a line to the graph where the x-values are the ‘day‘ values but now the y-values are the predicted values which we’ve called ‘yhat‘. Notice that you have to set the aesthetic ‘y=times‘ for the points and ‘y=yhat‘ for the line. Because each ‘geom_‘ will accept an ‘aes()‘ command, you can specify the ‘y‘ attribute to be different for different layers of the graph.

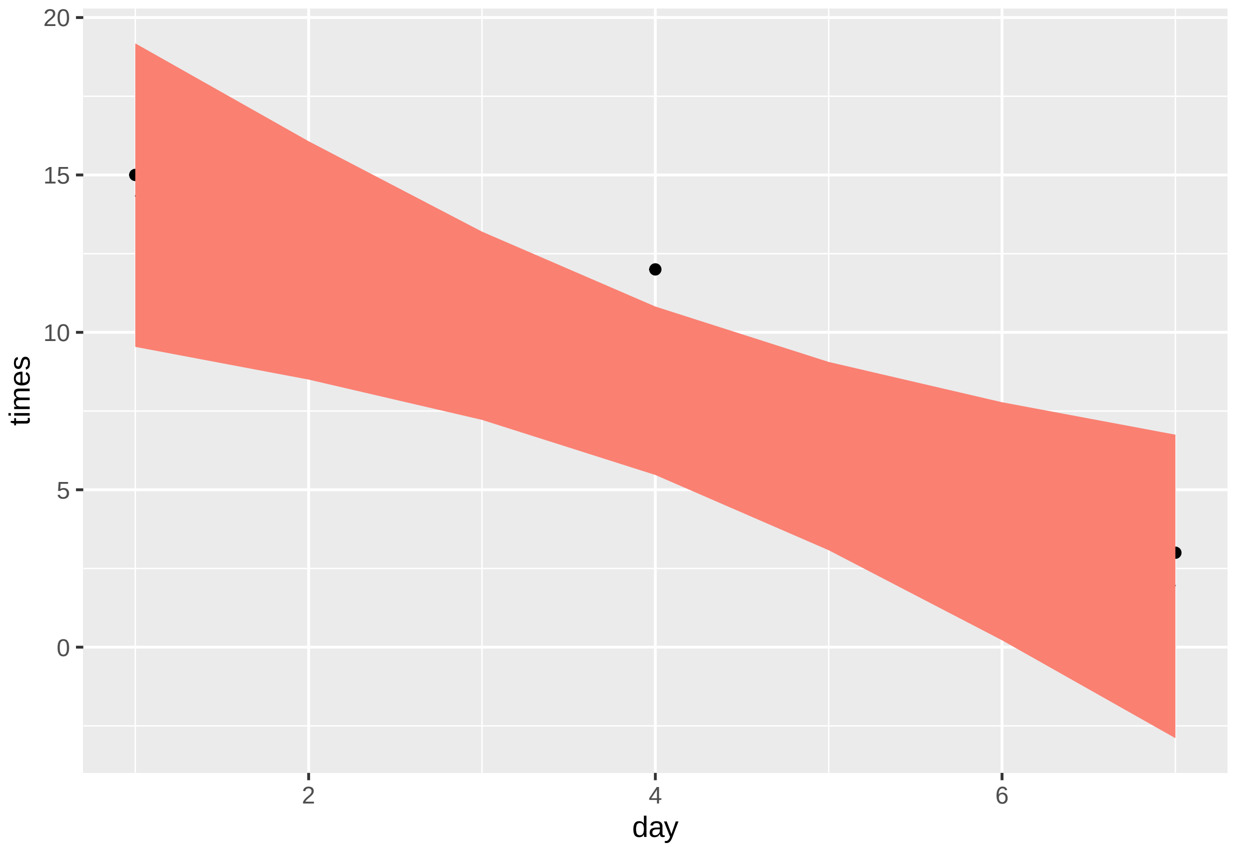
```
    Rosy <- data.frame(
      times = c(15, 11, 9, 12, 5, 2, 3),
      day   = 1:7,
      yhat  = c(14.36, 12.29, 10.21, 8.14, 6.07, 4.00,  1.93),
      lwr   = c( 9.54,  8.5,   7.22, 5.47, 3.08, 0.22, -2.89),
      upr   = c(19.18, 16.07, 13.2, 10.82, 9.06, 7.78,  6.75))
ggplot( data=Rosy, aes(x=day, y=times)) + # scatterplot
  geom_point() +
  geom_line(aes(x=day, y=yhat)) # best fit line
```



c) Add a ribbon that represents the confidence region of the regression line. The `geom_ribbon()` function requires an `x`, `ymin`, and `ymax` columns to be defined. For examples of using `geom_ribbon()` see the online documentation: http://docs.ggplot2.org/current/geom__ribbon.html.

```
ggplot(Rosy, aes(x=day)) +
  geom_point(aes(y=times)) +
  geom_line( aes(y=yhat)) +
  geom_ribbon( aes(ymin=lwr, ymax=upr), fill='salmon')
```

```
    Rosy <- data.frame(
      times = c(15, 11, 9, 12, 5, 2, 3),
      day   = 1:7,
      yhat  = c(14.36, 12.29, 10.21, 8.14, 6.07, 4.00,  1.93),
      lwr   = c( 9.54,  8.5,   7.22, 5.47, 3.08, 0.22, -2.89),
      upr   = c(19.18, 16.07, 13.2, 10.82, 9.06, 7.78,  6.75))
ggplot( data=Rosy, aes(x=day, y=times)) + # scatterplot
  geom_point() +
  geom_line(aes(x=day, y=yhat)) + # best fit line
  geom_ribbon(aes(ymin=lwr, ymax=upr), fill='salmon')
```

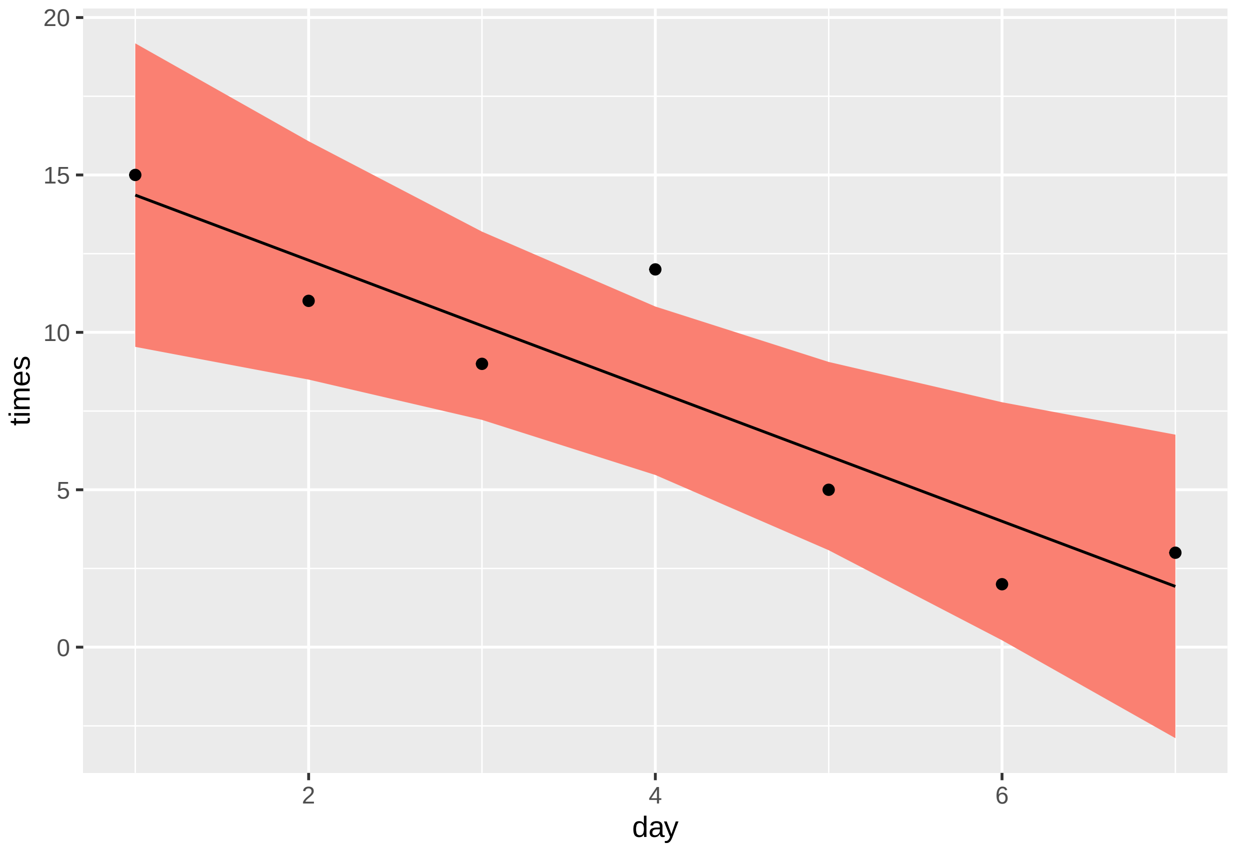


```
  # ribbon of confidence region
```

d) What happened when you added the ribbon? Did some points get hidden? If so, why?
When I added the ribbon most of the points became hidden because they are within the confidence region that was added.

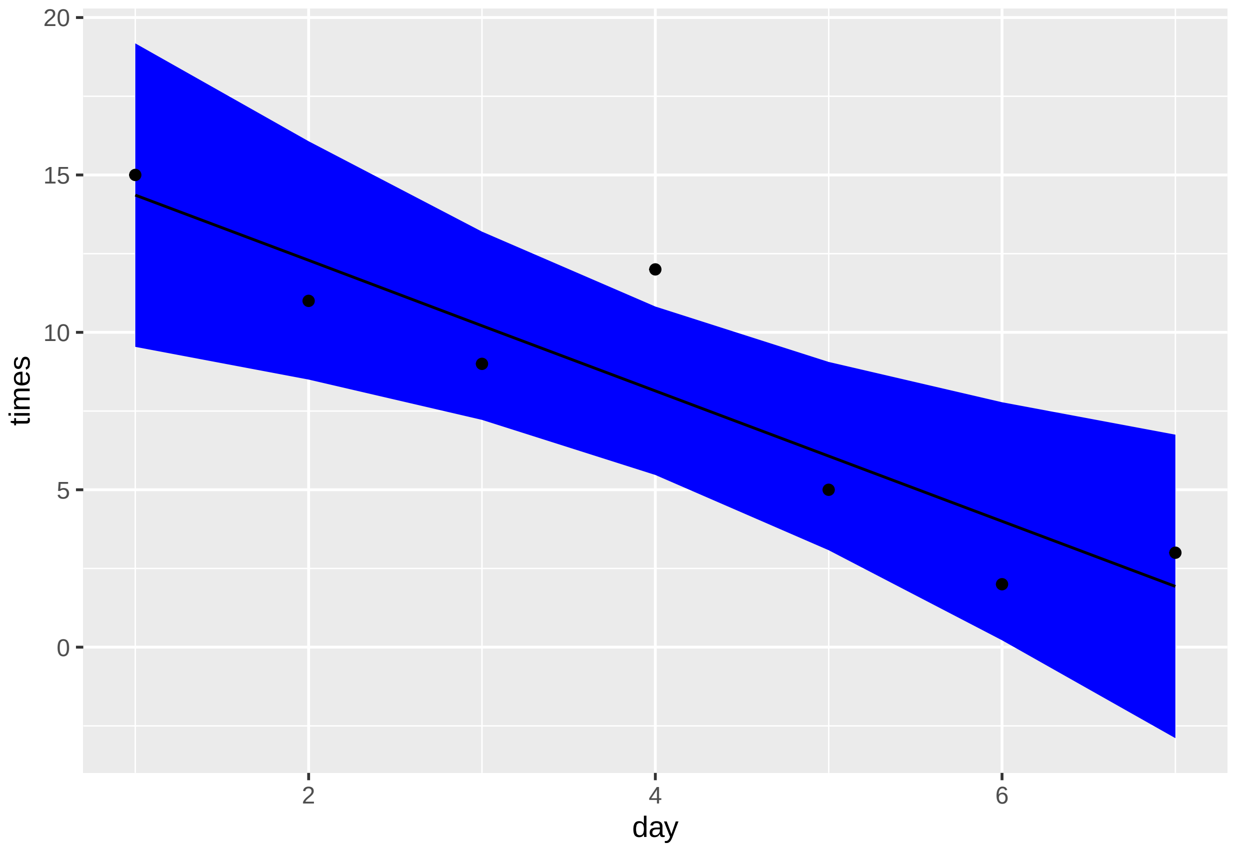
e) Reorder the statements that created the graph so that the ribbon is on the bottom and the data points are on top and the regression line is visible.

```
    Rosy <- data.frame(
      times = c(15, 11, 9, 12, 5, 2, 3),
      day   = 1:7,
      yhat  = c(14.36, 12.29, 10.21, 8.14, 6.07, 4.00,  1.93),
      lwr   = c( 9.54,  8.5,   7.22, 5.47, 3.08, 0.22, -2.89),
      upr   = c(19.18, 16.07, 13.2, 10.82, 9.06, 7.78,  6.75))
ggplot( data=Rosy, aes(x=day, y=times)) + # scatterplot
  geom_ribbon(aes(ymin=lwr, ymax=upr), fill='salmon') +  # layer 1
  geom_point() + # layer 2
  geom_line(aes(x=day, y=yhat)) # layer 3
```



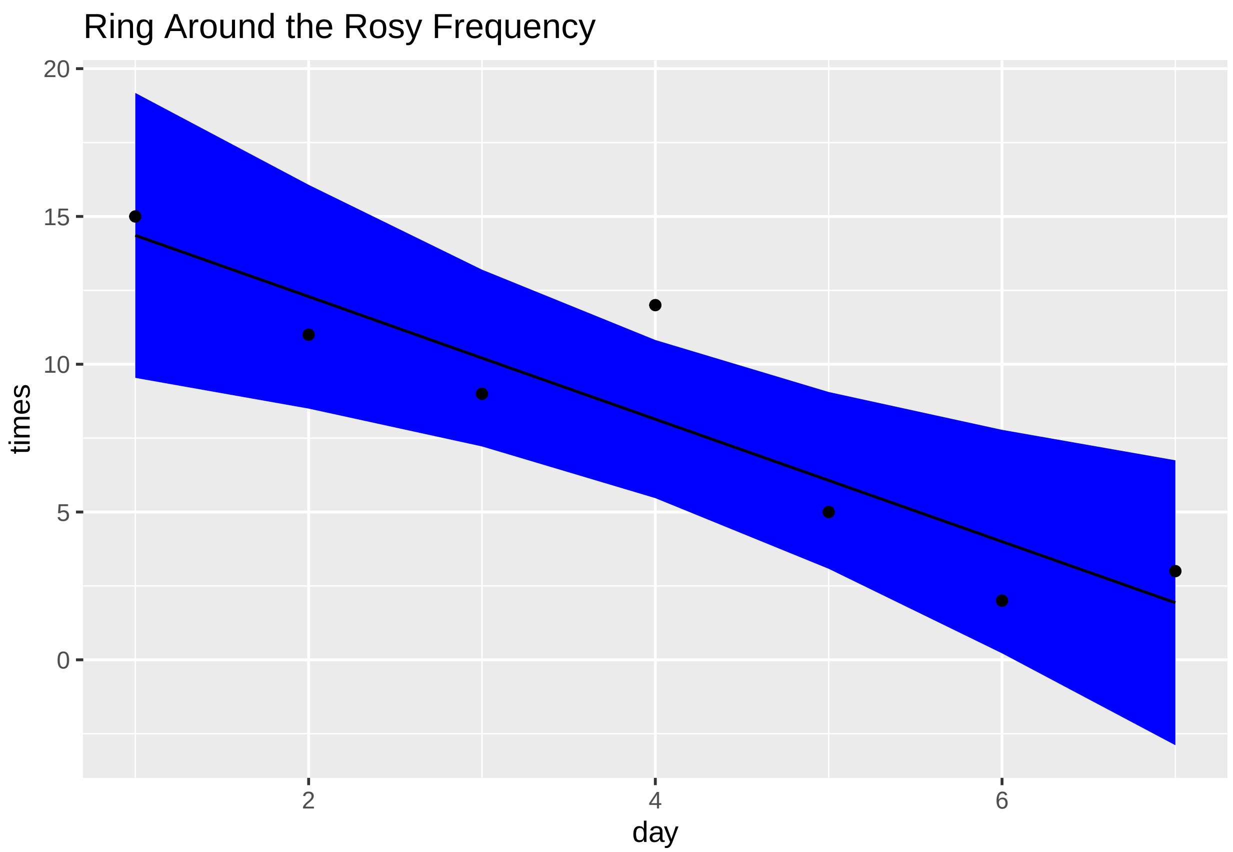
f) The color of the ribbon fill is ugly. Use Google to find a list of   named colors available t
colors" and found the following link:

```
    Rosy <- data.frame(
      times = c(15, 11, 9, 12, 5, 2, 3),
      day   = 1:7,
      yhat  = c(14.36, 12.29, 10.21, 8.14, 6.07, 4.00,  1.93),
      lwr   = c( 9.54,  8.5,   7.22, 5.47, 3.08, 0.22, -2.89),
      upr   = c(19.18, 16.07, 13.2, 10.82, 9.06, 7.78,  6.75))
ggplot( data=Rosy, aes(x=day, y=times)) + # scatterplot
  geom_ribbon(aes(ymin=lwr, ymax=upr), fill='blue') +  # color change
  geom_point() +
  geom_line(aes(x=day, y=yhat))
```



g) Add labels for the x-axis and y-axis that are appropriate along with a main title.

```
    Rosy <- data.frame(
      times = c(15, 11, 9, 12, 5, 2, 3),
      day   = 1:7,
      yhat  = c(14.36, 12.29, 10.21, 8.14, 6.07, 4.00,  1.93),
      lwr   = c( 9.54,  8.5,   7.22, 5.47, 3.08, 0.22, -2.89),
      upr   = c(19.18, 16.07, 13.2, 10.82, 9.06, 7.78,  6.75))
ggplot( data=Rosy, aes(x=day, y=times)) + # scatterplot
  geom_ribbon(aes(ymin=lwr, ymax=upr), fill='blue') +  # color change
  geom_point() +
  geom_line(aes(x=day, y=yhat)) +
  labs( title= 'Ring Around the Rosy Frequency') + # labels
  labs(x="day", y="times" )
```



3. We'll next make some density plots that relate several factors towards the birth weight of a child. *Because these questions ask you to produce several graphs and evaluate which is better and why, please include each graph and response with each sub-question.*

   a) The `MASS` package contains a dataset called `birthwt` which contains information about 189 babies and their mothers. In particular there are columns for the mother's race and smoking status during the pregnancy. Load the `birthwt` by either using the `data()` command or loading the `MASS` library.

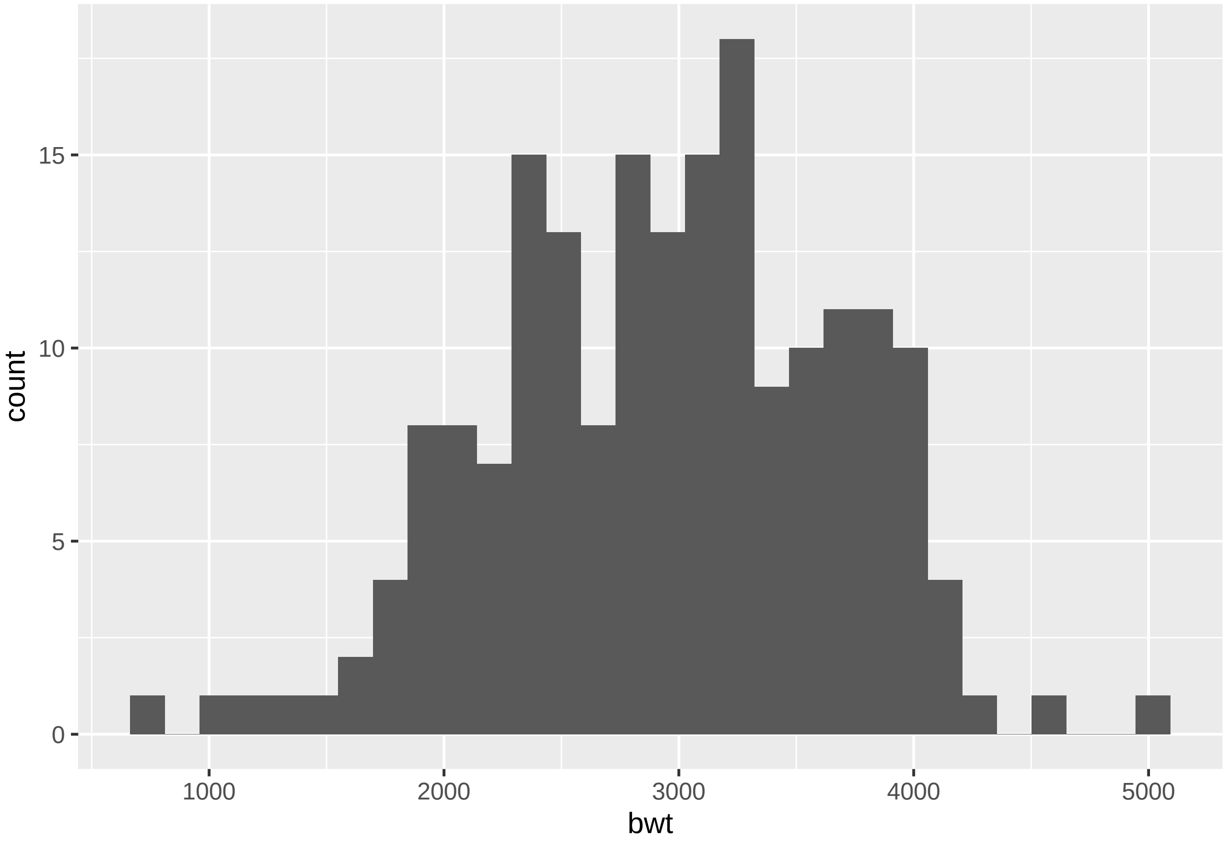
   b) Read the help file for the dataset using `MASS::birthwt`. The covariates `race` and `smoke` are not stored in a user friendly manner. For example, smoking status is labeled using a 0 or a 1. Because it is not obvious which should represent that the mother smoked, we'll add better labels to the `race` and `smoke` variables. For more information about dealing with factors and their levels, see the `Factors` chapter in these notes.

```
library(tidyverse)
data('birthwt', package='MASS')
birthwt <- birthwt %>% mutate(
  race  = factor(race,  labels=c('White','Black','Other')),
  smoke = factor(smoke, labels=c('No Smoke', 'Smoke')))
```

```
?MASS::birthwt # loaded help file
```

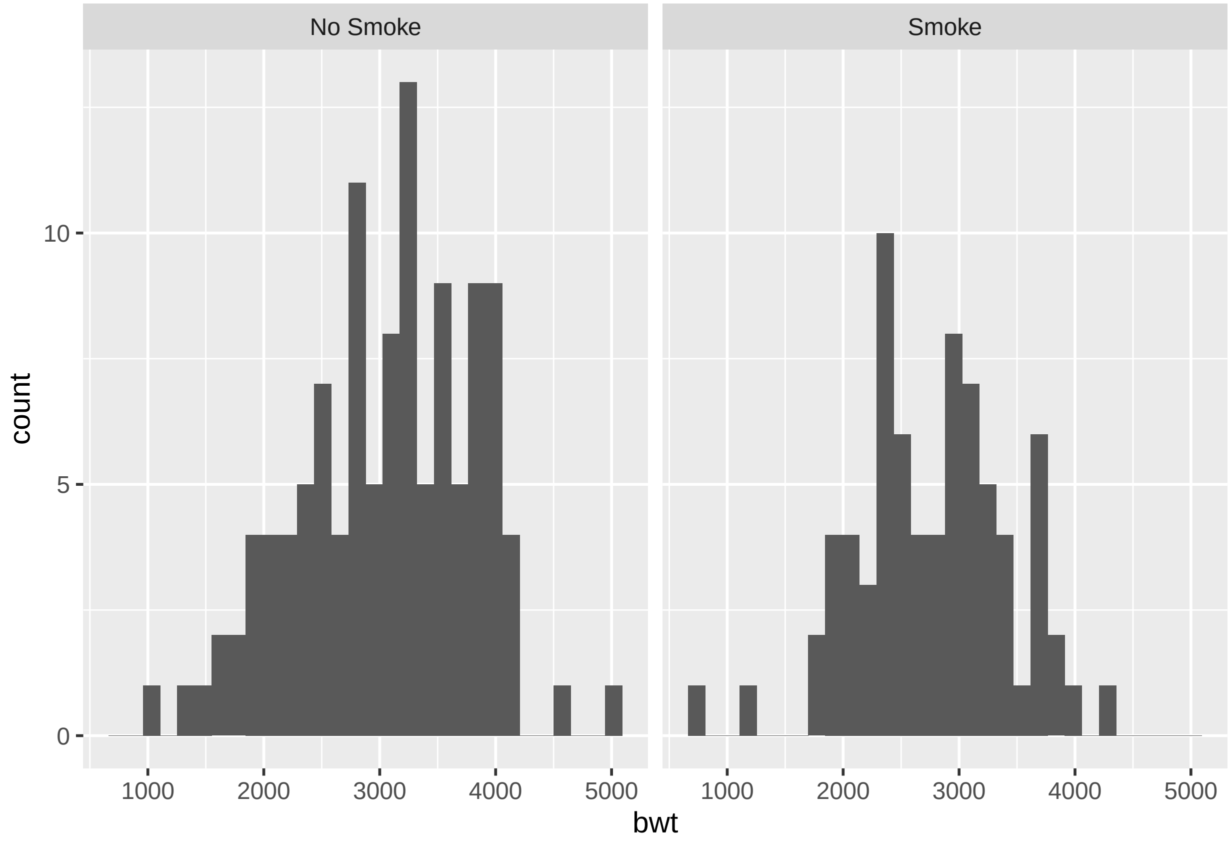
```
c)  Graph a histogram of the birth weights ‘bwt‘ using
‘ggplot(birthwt, aes(x=bwt)) + geom_histogram()‘.
```

```
library(tidyverse)
data('birthwt', package='MASS')
birthwt <- birthwt %>% mutate(
  race  = factor(race,  labels=c('White','Black','Other')),
  smoke = factor(smoke, labels=c('No Smoke', 'Smoke')))
ggplot(birthwt, aes(x=bwt)) + geom_histogram() # histogram
```



d) Make separate graphs that denote whether a mother smoked during pregnancy by appending '+ facet_grid()' command to your original graphing command.

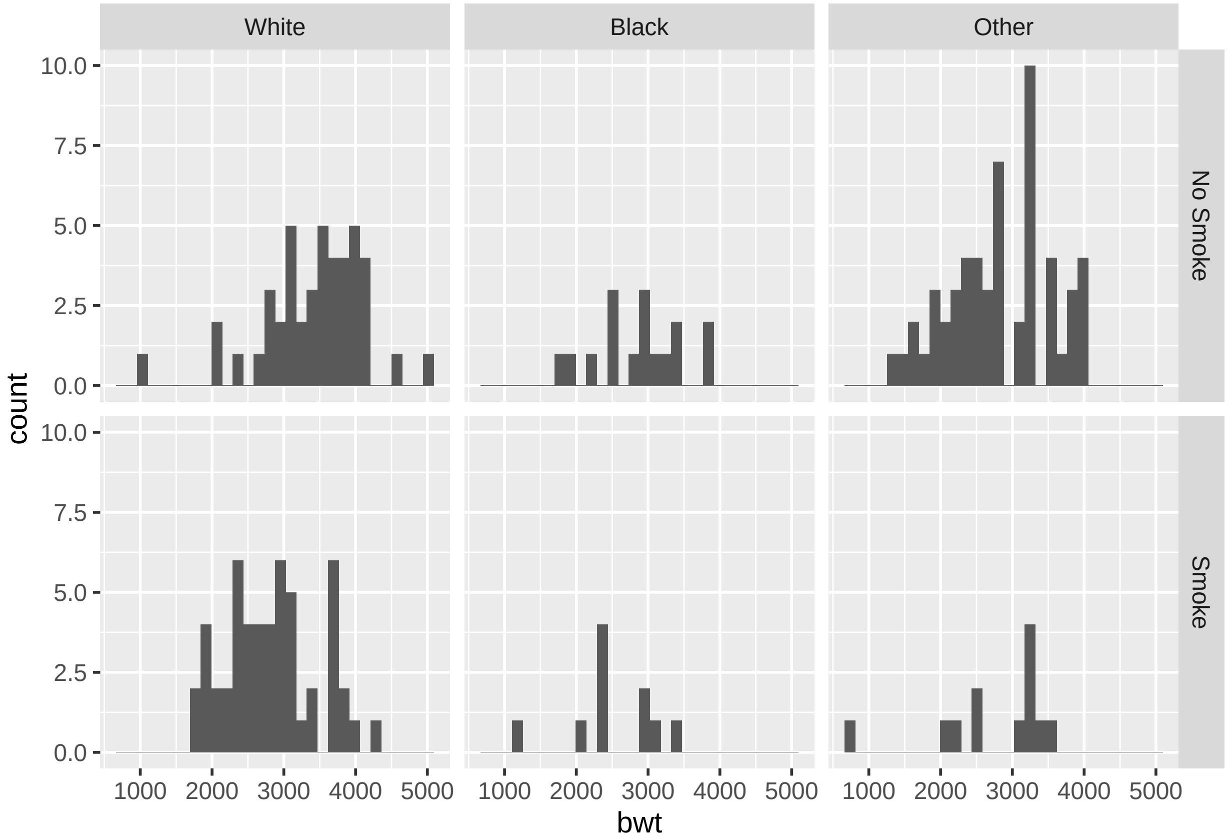
```
library(tidyverse)
data('birthwt', package='MASS')
birthwt <- birthwt %>% mutate(
  race  = factor(race,  labels=c('White','Black','Other')),
  smoke = factor(smoke, labels=c('No Smoke', 'Smoke')))
ggplot(birthwt, aes(x=bwt)) + geom_histogram() +
  facet_grid( cols = vars(smoke))
```



```
# 2 seperate graphs for smoke status
```

e) Perhaps race matters in relation to smoking. Make our grid of graphs vary with smoking status changing vertically, and race changing horizontally (that is the formula in ‘facet_grid()‘ should have smoking be the y variable and race as the x).

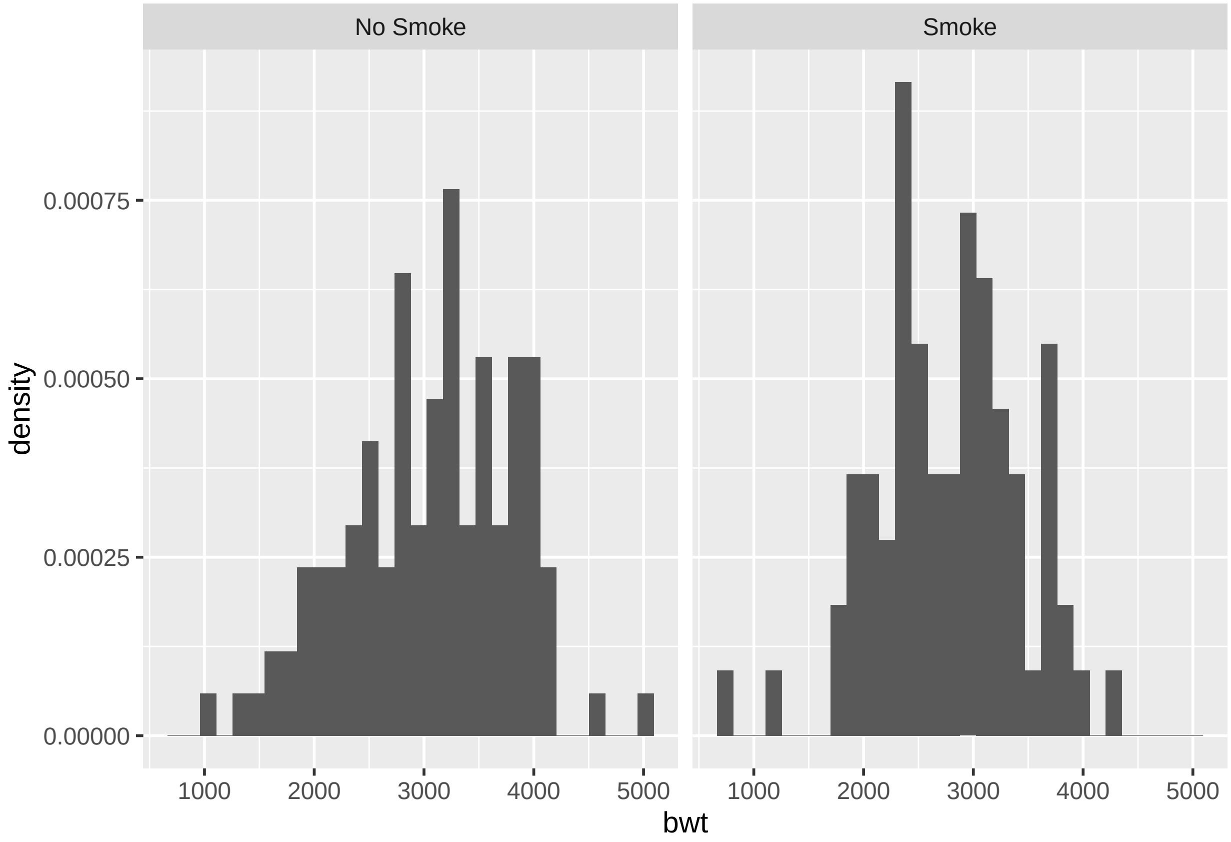
```
library(tidyverse)
data('birthwt', package='MASS')
birthwt <- birthwt %>% mutate(
  race  = factor(race,  labels=c('White','Black','Other')),
  smoke = factor(smoke, labels=c('No Smoke', 'Smoke')))
ggplot(birthwt, aes(x=bwt)) + geom_histogram() +
  facet_grid( smoke ~ race )
```



```
# seperate graphs for smoke status vs race
```

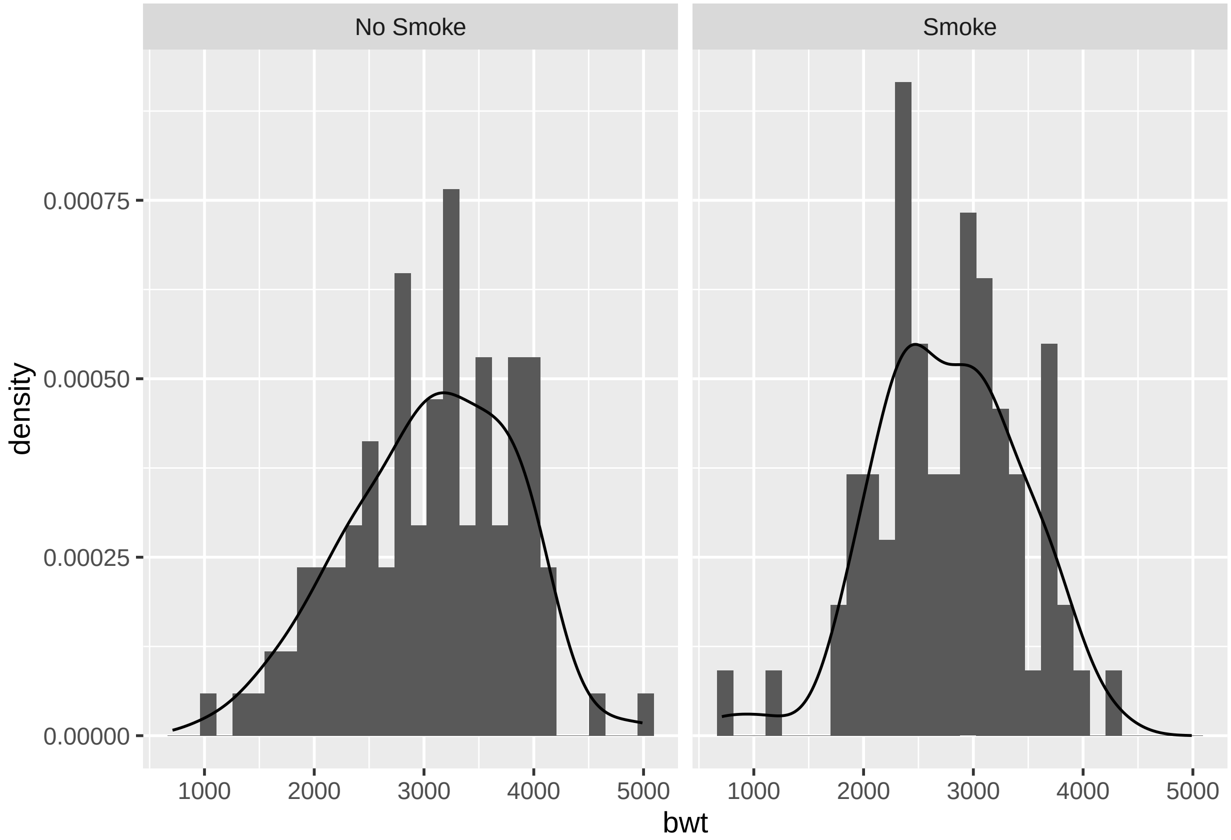
f) Remove ‘race‘ from the facet grid, (so go back to the graph you had in part d). I’d like to next add an estimated density line to the graphs, but to do that, I need to first change the y-axis to be density (instead of counts), which we do by using ‘aes(y=..density..)‘ in the ‘ggplot()‘ aesthetics command.

```
library(tidyverse)
data('birthwt', package='MASS')
birthwt <- birthwt %>% mutate(
  race  = factor(race,  labels=c('White','Black','Other')),
  smoke = factor(smoke, labels=c('No Smoke', 'Smoke')))
ggplot(birthwt, aes(x=bwt, y=..density..)) + geom_histogram() +
  facet_grid( cols = vars(smoke)) # density on y-axis
```



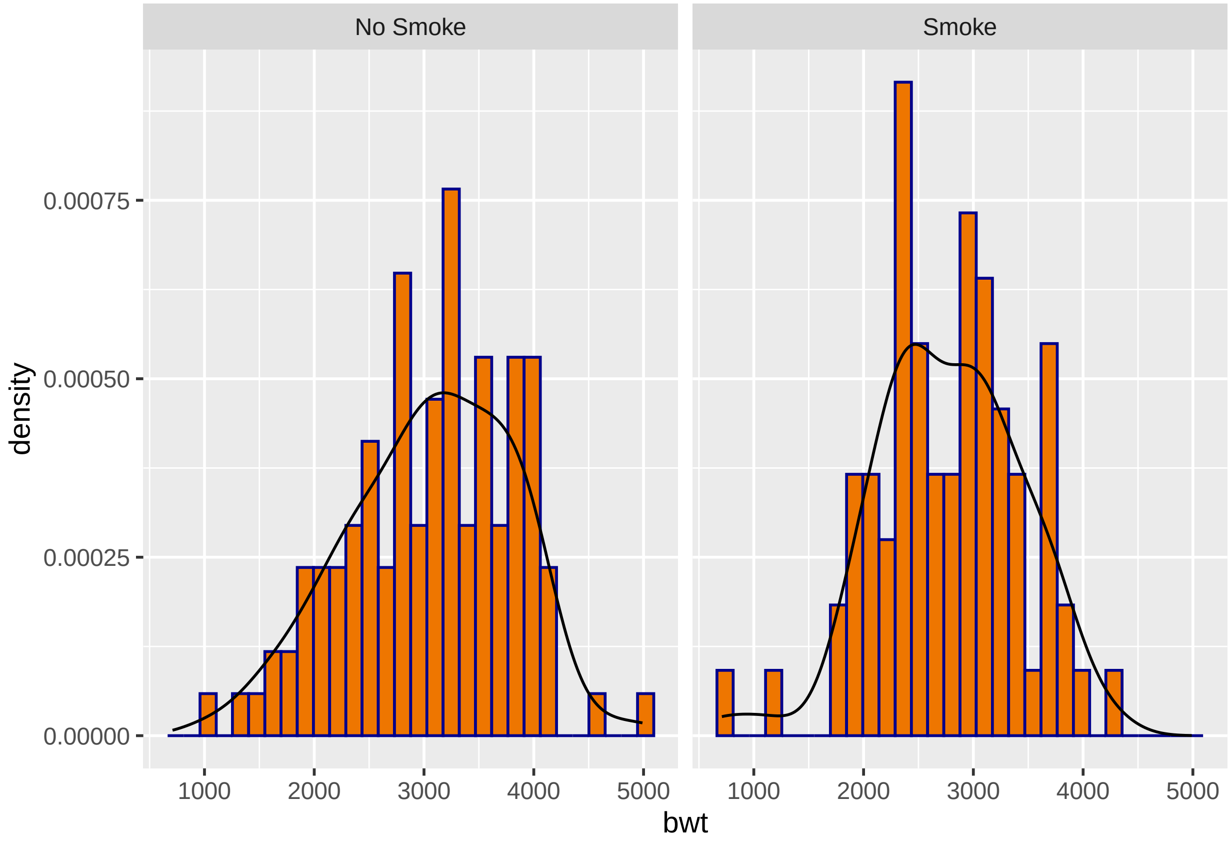
g) Next we can add the estimated smooth density using the ‘geom_density()‘ command.

```r
library(tidyverse)
data('birthwt', package='MASS')
birthwt <- birthwt %>% mutate(
  race  = factor(race,  labels=c('White','Black','Other')),
  smoke = factor(smoke, labels=c('No Smoke', 'Smoke')))
ggplot(birthwt, aes(x=bwt, y=..density..)) + geom_histogram() +
  facet_grid( cols = vars(smoke)) + geom_density() # add density line
```



```
h)  To really make this look nice, lets change the fill color of the
histograms to be something less dark, lets use ‘fill=’cornsilk’‘ and
‘color=’grey60’‘.
    To play with different colors that have names, check out the following:
    [https://www.datanovia.com/en/blog/awesome-list-of-657-r-color-names/](https://www.datanovia.com/en/
```

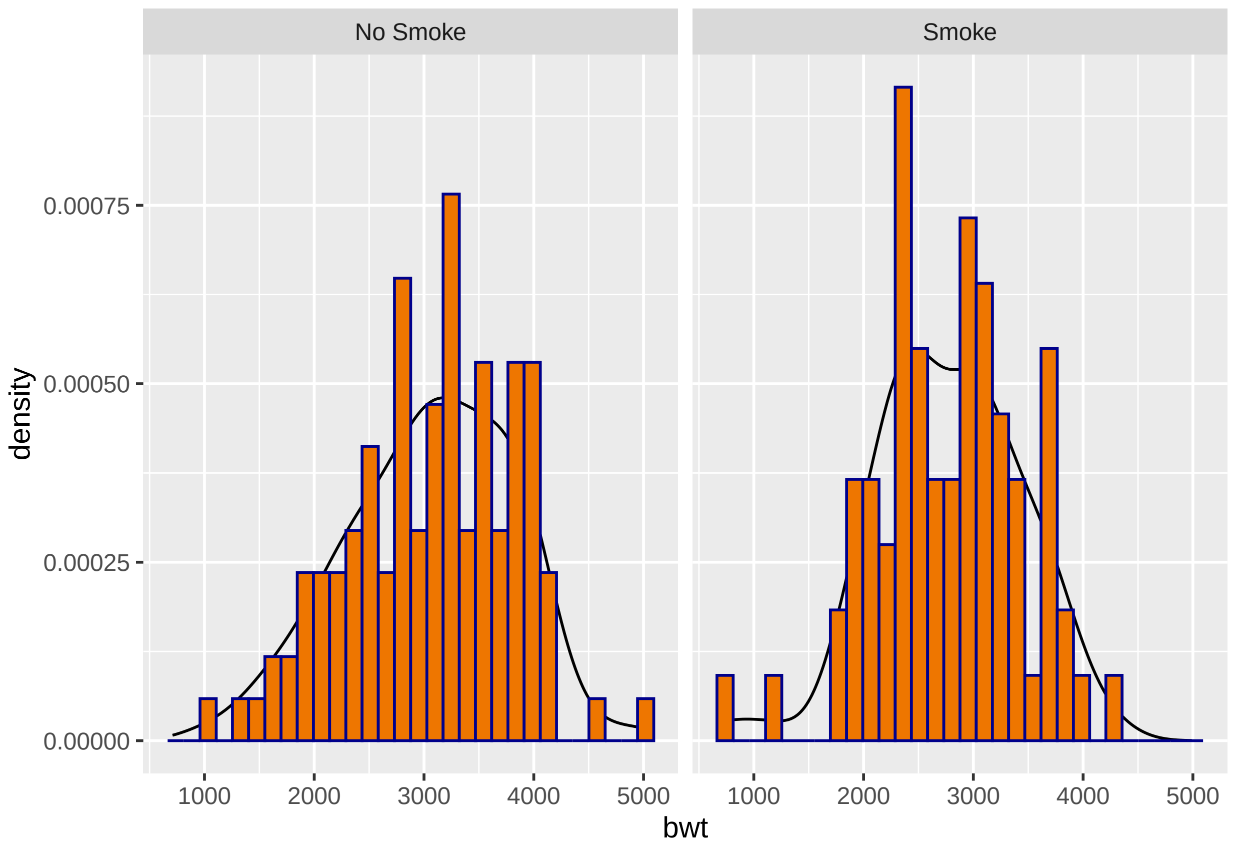
```r
library(tidyverse)
data('birthwt', package='MASS')
birthwt <- birthwt %>% mutate(
  race  = factor(race,  labels=c('White','Black','Other')),
  smoke = factor(smoke, labels=c('No Smoke', 'Smoke')))
ggplot(birthwt, aes(x=bwt, y=..density..)) +
  geom_histogram( fill='darkorange2', color='darkblue') + # change color
  facet_grid( cols = vars(smoke)) + geom_density()
```



```
i)  Change the order in which the histogram and the density line are added
to the plot. Does it matter and which do you prefer?
```

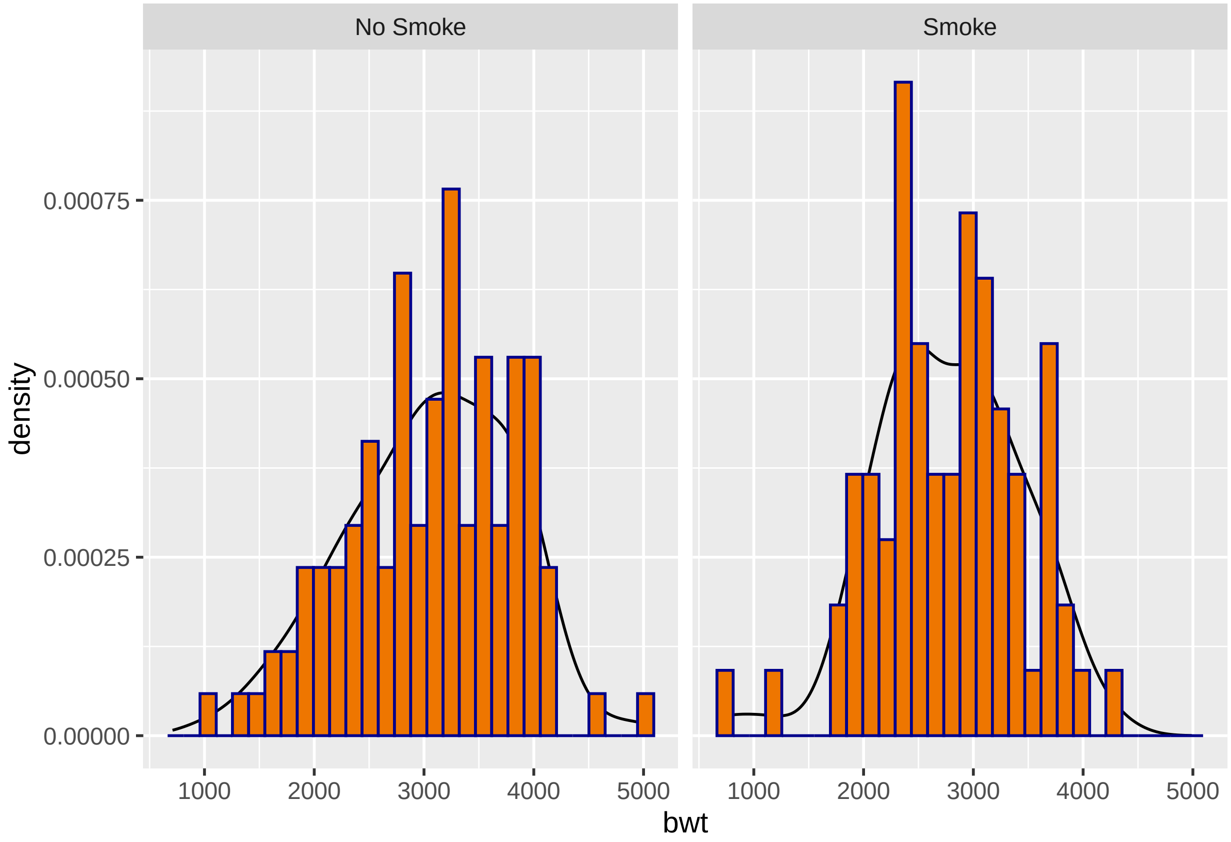
```
It doesn’t really matter if the density line is added under the histogram
in this scenario but I could see how it could make a difference if the
density line covered up certain critical values. I personally prefer the
density line layered over the histogram.
```

```r
library(tidyverse)
data('birthwt', package='MASS')
birthwt <- birthwt %>% mutate(
  race  = factor(race,  labels=c('White','Black','Other')),
  smoke = factor(smoke, labels=c('No Smoke', 'Smoke')))
ggplot(birthwt, aes(x=bwt, y=..density..)) +
  geom_density() + geom_histogram( fill='darkorange2', color='darkblue') +
  facet_grid( cols = vars(smoke)) # layered density line under histogram
```



j) Finally consider if you should have the histograms side-by-side or one on top of the other (i.e. ‘. ~ smoke‘ or ‘smoke ~ .‘). Which do you think better displays the decrease in mean birth weight and why? The side by side comparison better displays the decrease in mean birth weight because it makes it easier to compare the heights in the bars when placed next to eachother.

```
library(tidyverse)
data('birthwt', package='MASS')
birthwt <- birthwt %>% mutate(
  race  = factor(race,  labels=c('White','Black','Other')),
  smoke = factor(smoke, labels=c('No Smoke', 'Smoke')))
ggplot(birthwt, aes(x=bwt, y=..density..)) +
  geom_density() + geom_histogram( fill='darkorange2', color='darkblue') +
  facet_grid(. ~ smoke) # histograms side by side
```

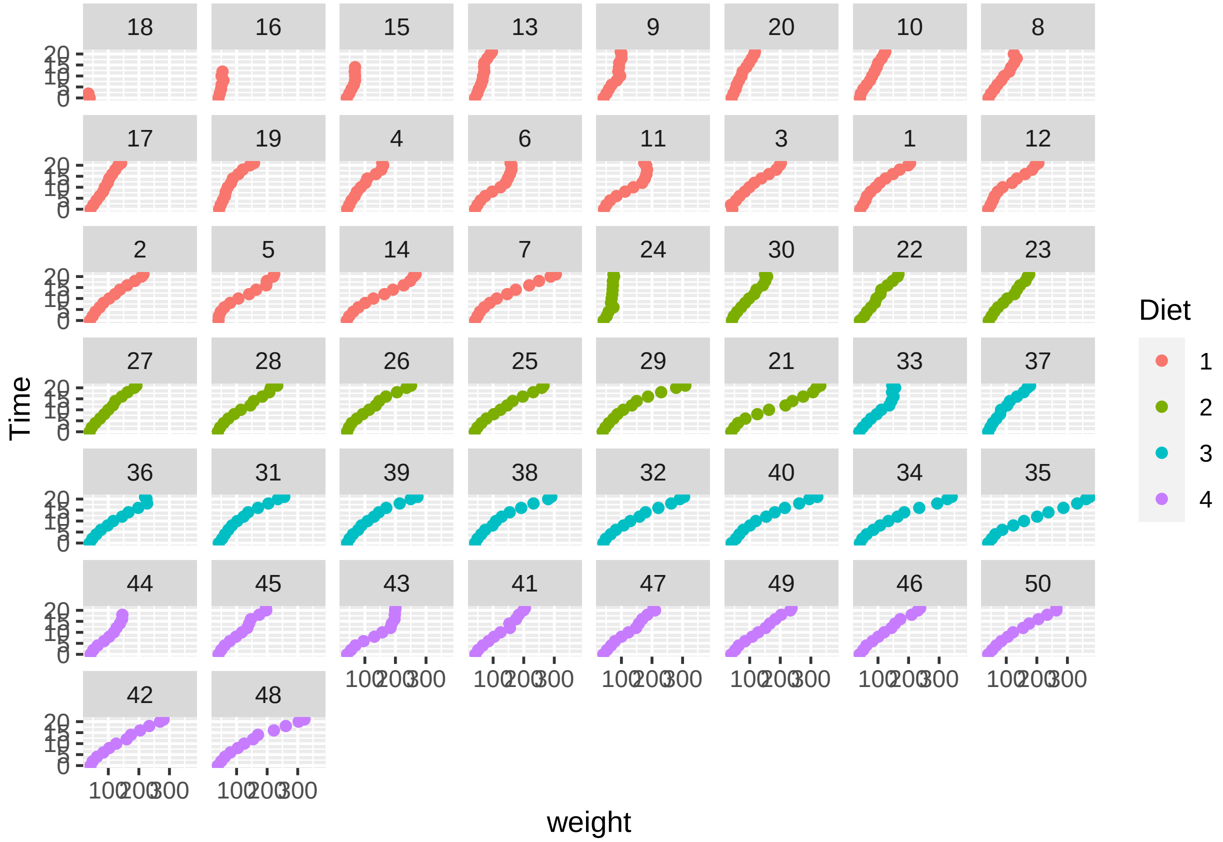


4. Load the data set `ChickWeight`, which comes pre-loaded in R, and get the background on the data set by reading the manual page `?ChickWeight`. *Because these questions ask you to produce several graphs and evaluate which is better and why, please include each graph and response with each sub-question.*

```
data('ChickWeight') # loaded data set
?ChickWeight # loaded manual page
```

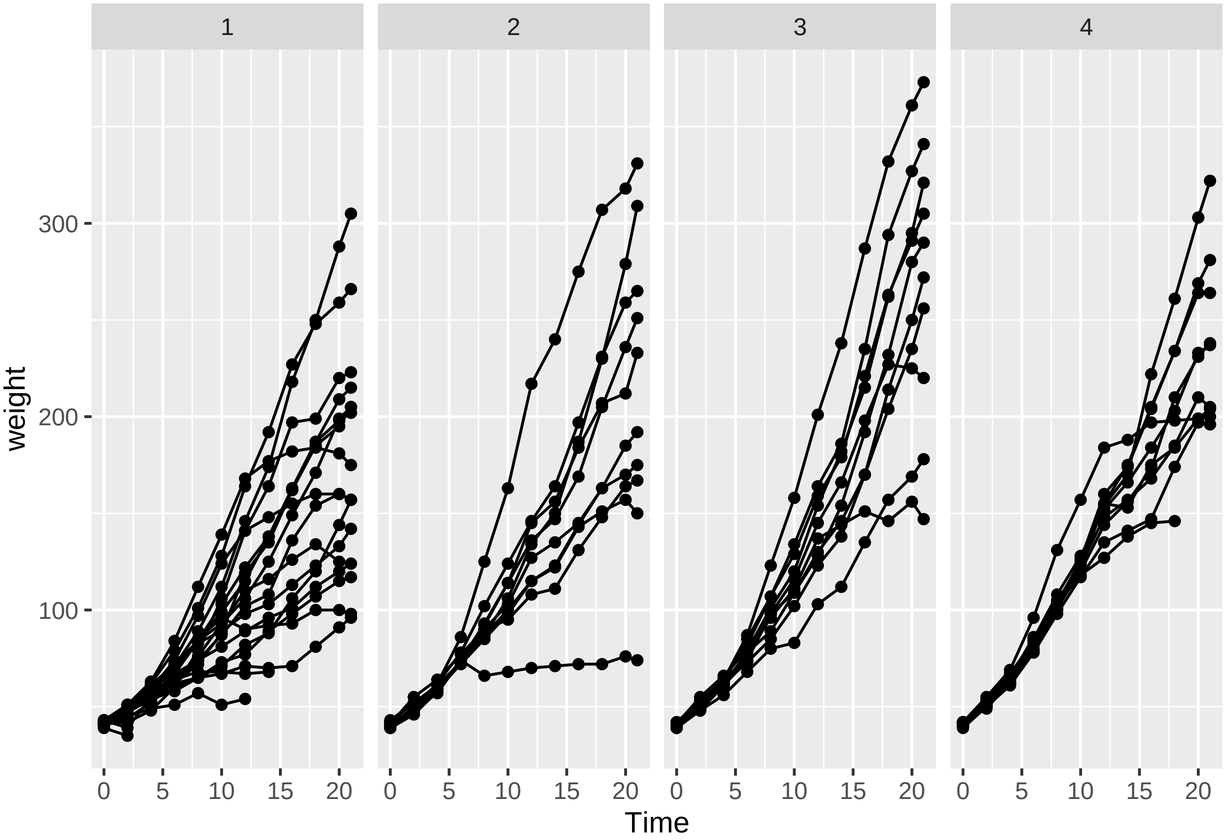
```
a)  Produce a separate scatter plot of weight vs age for each chick. Use
color to distinguish the four different ‘Diet‘ treatments. *Note, this*
    *question should produce 50 separate graphs! If the graphs are too squished*
    *you should consider how to arrange them so that the graphs wrap to a new*
    *row of graphs in the resulting output figure.*
```

```
data('ChickWeight')
ggplot(ChickWeight, aes(x= weight, y= Time)) + # plotted weight vs age
  geom_point(aes(color=Diet)) + # used different color for each diet
  facet_wrap(. ~ Chick) # used facet_wrap to better display graphs
```



b) We could examine these data by producing a scatter plot for each diet. Most of the code below is readable, but if we don't add the ‘group‘ aesthetic the lines would not connect the dots for each Chick but would instead connect the dots across different chicks.

```
data(ChickWeight)
ggplot(ChickWeight, aes(x=Time, y=weight, group=Chick )) +
  geom_point() + geom_line() +
  facet_grid( ~ Diet)
```



Notice in the code chunk above, if you don't remove the `eval=FALSE` *in the chunk header, that the code will be displayed, but it won't be run and no plot will be produced in your final output document.*