STA 444 Exercise 3

Richard McCormick

9/6/2023

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.

```
?trees
```

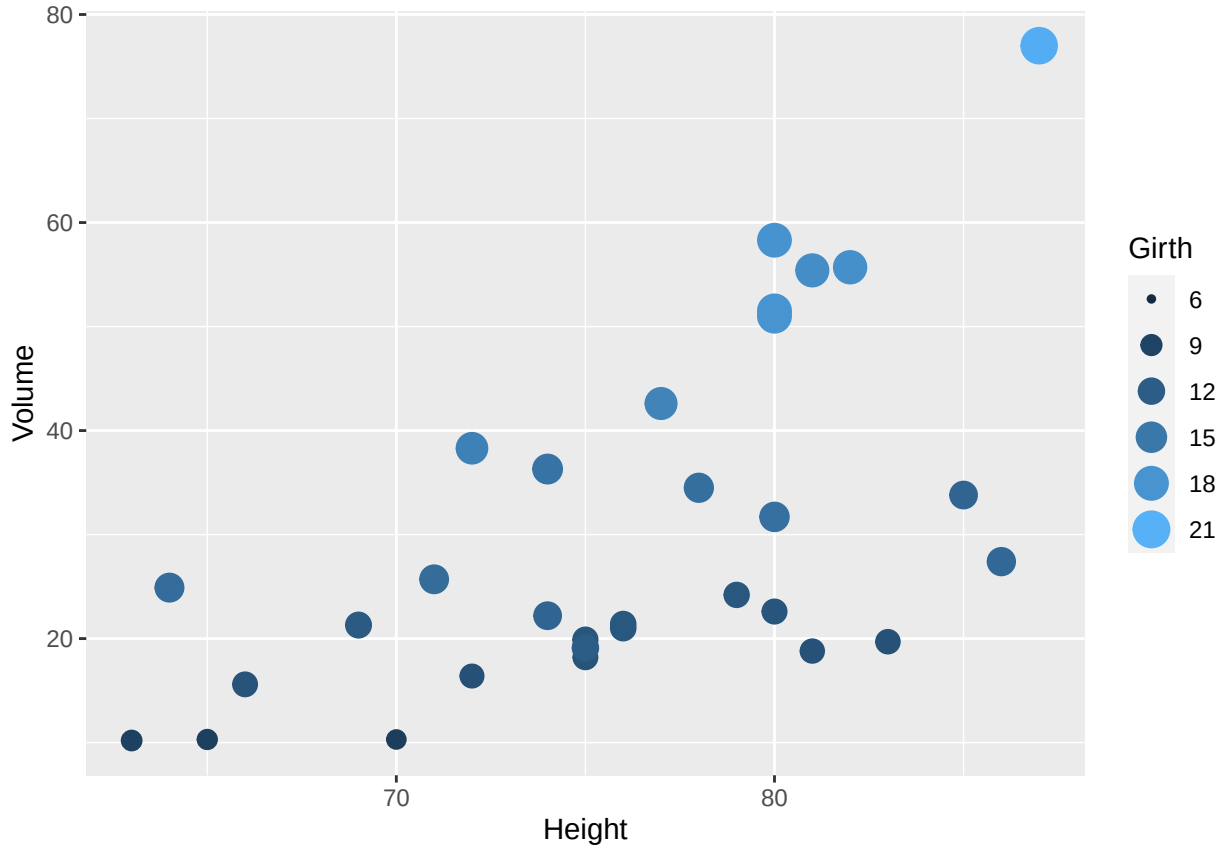
```
## starting httpd help server ... done
```

```
summary(trees)
```

```
##      Girth           Height       Volume
##  Min.   : 8.30   Min.   :63   Min.   :10.20
##  1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40
##  Median :12.90   Median :76   Median :24.20
##  Mean   :13.25   Mean   :76   Mean   :30.17
##  3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30
##  Max.   :20.60   Max.   :87   Max.   :77.00
```

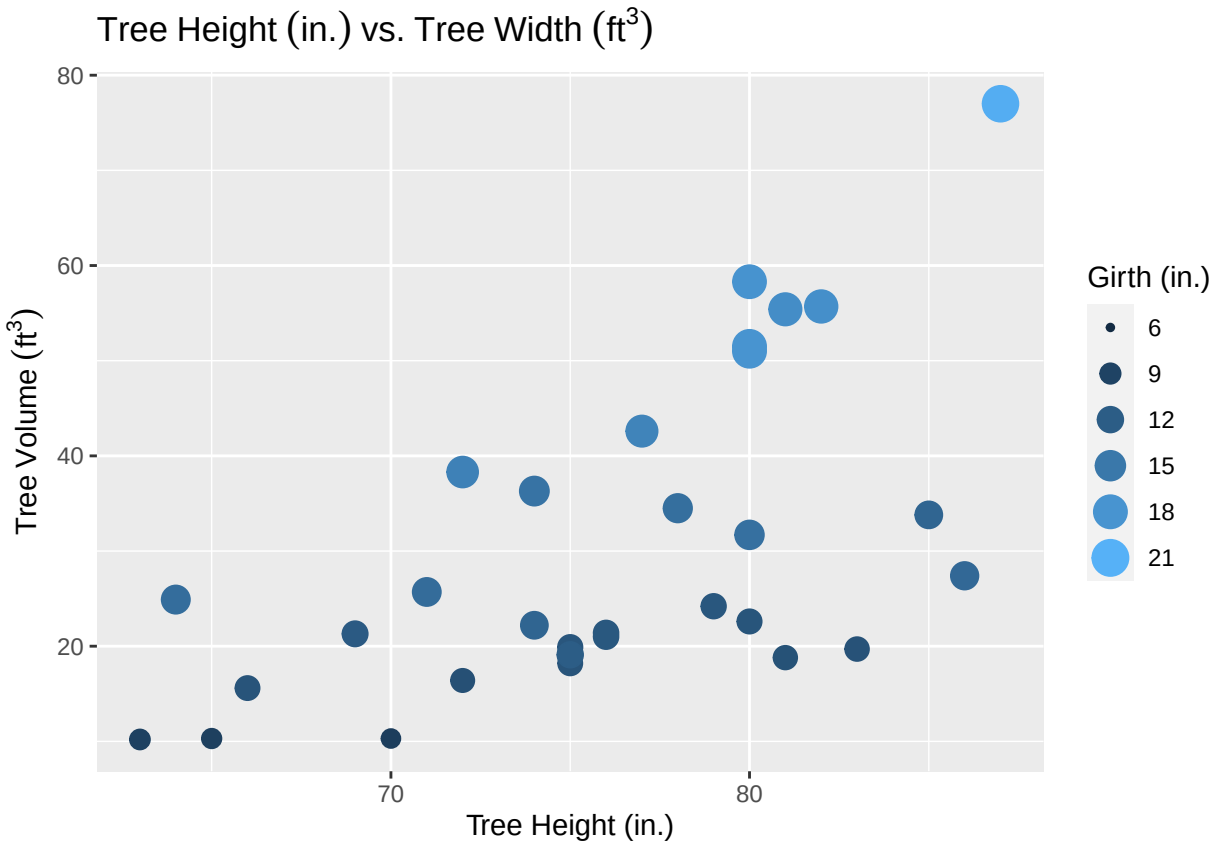
a. Create a graph using ggplot2 with Height on the x-axis, Volume on the y-axis, and Girth as the either the size of the data point or the color of the data point. Which do you think is a more intuitive representation?

```
graph <- ggplot(
  data = trees,
  aes(x=Height, y=Volume) ) +
  geom_point( aes(size=Girth, color=Girth) ) +
  scale_size_continuous(limits=c(6, 21), breaks=seq(6,21, by=3)) +
  scale_color_continuous(limits=c(6, 21), breaks=seq(6,21, by=3)) +
  guides(color= guide_legend(), size=guide_legend())

graph
```



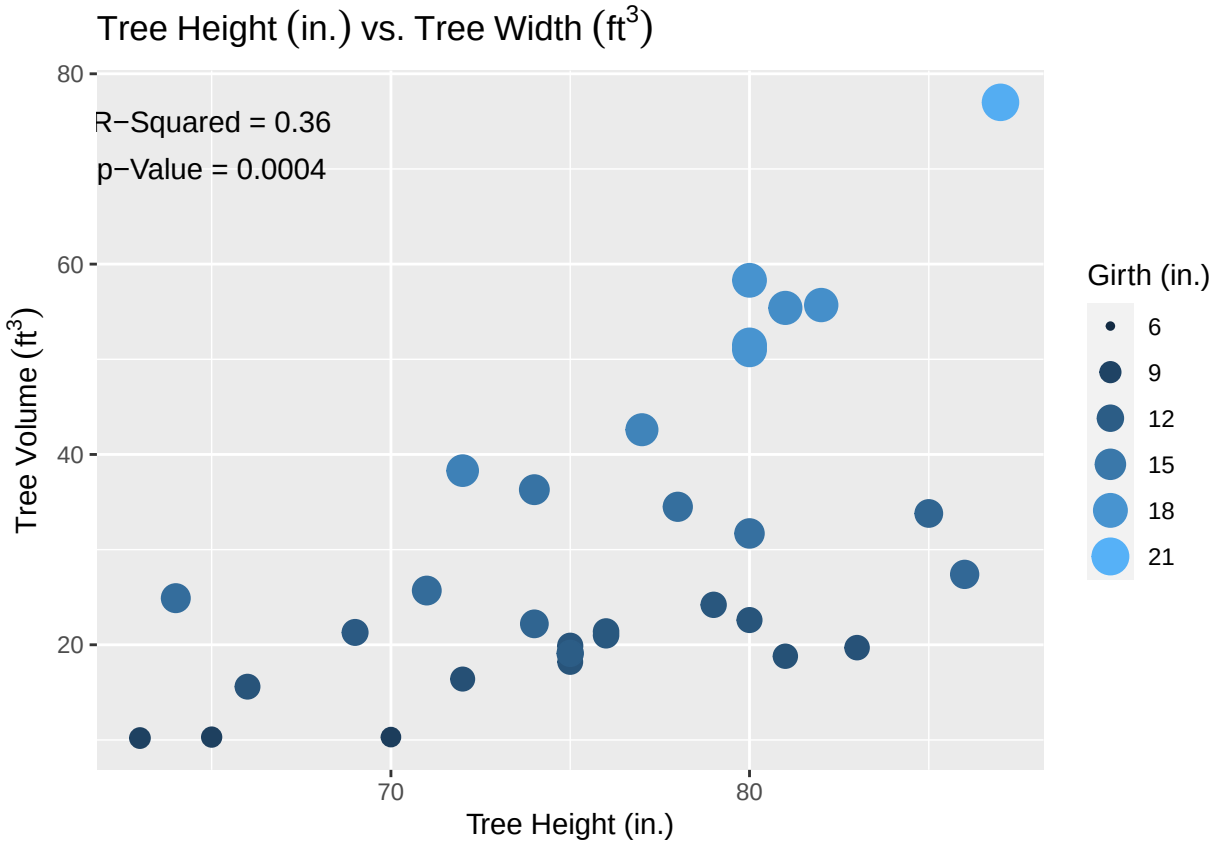
b. Add appropriate labels for the main title and the x and y axes.

```
title_lab <- expression(Tree ~ Height ~ (in.) ~ vs. ~ Tree ~ Width ~ (ft^3))
y_lab = expression(Tree ~ Volume ~ (ft^3) )

graph <- graph +
  labs( title= title_lab ) +
  labs( x='Tree Height (in.)' ) +
  labs( y= y_lab ) +
  labs( size= 'Girth (in.)' ) +
  labs( color= 'Girth (in.)' )

graph
```



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.

```
text_r = "R-Squared = 0.36"
text_p = "p-Value = 0.0004"

graph <- graph +
  annotate("text", x=65, y=70, label = text_p) +
  annotate("text", x=65, y=75, label = text_r)

graph
```

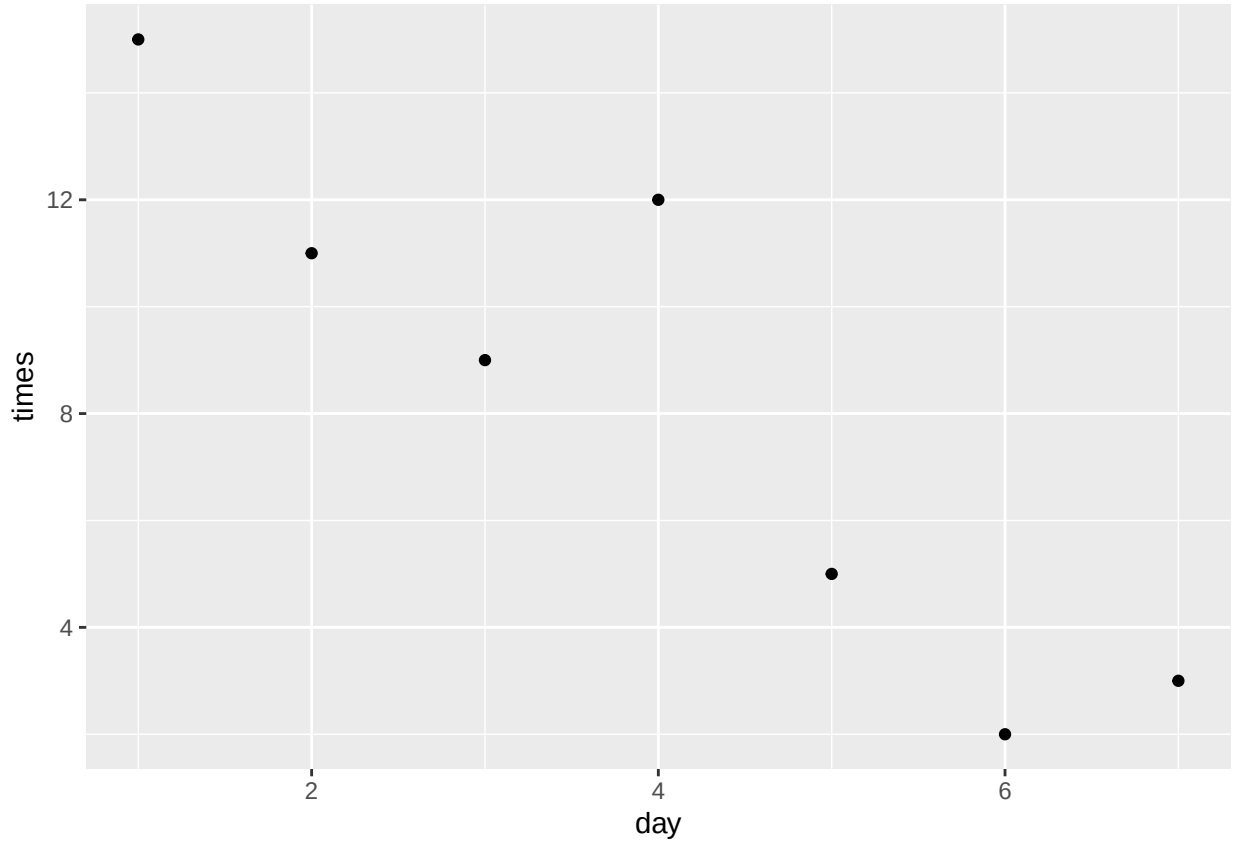


2. Consider the following small dataset that represents the number of times per day my wife played "Ring around the Rosy" with my daughter relative to the number of days since she has learned this game. The column yhat represents the best fitting line through the data, and lwr and upr represent a 95% 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))
```

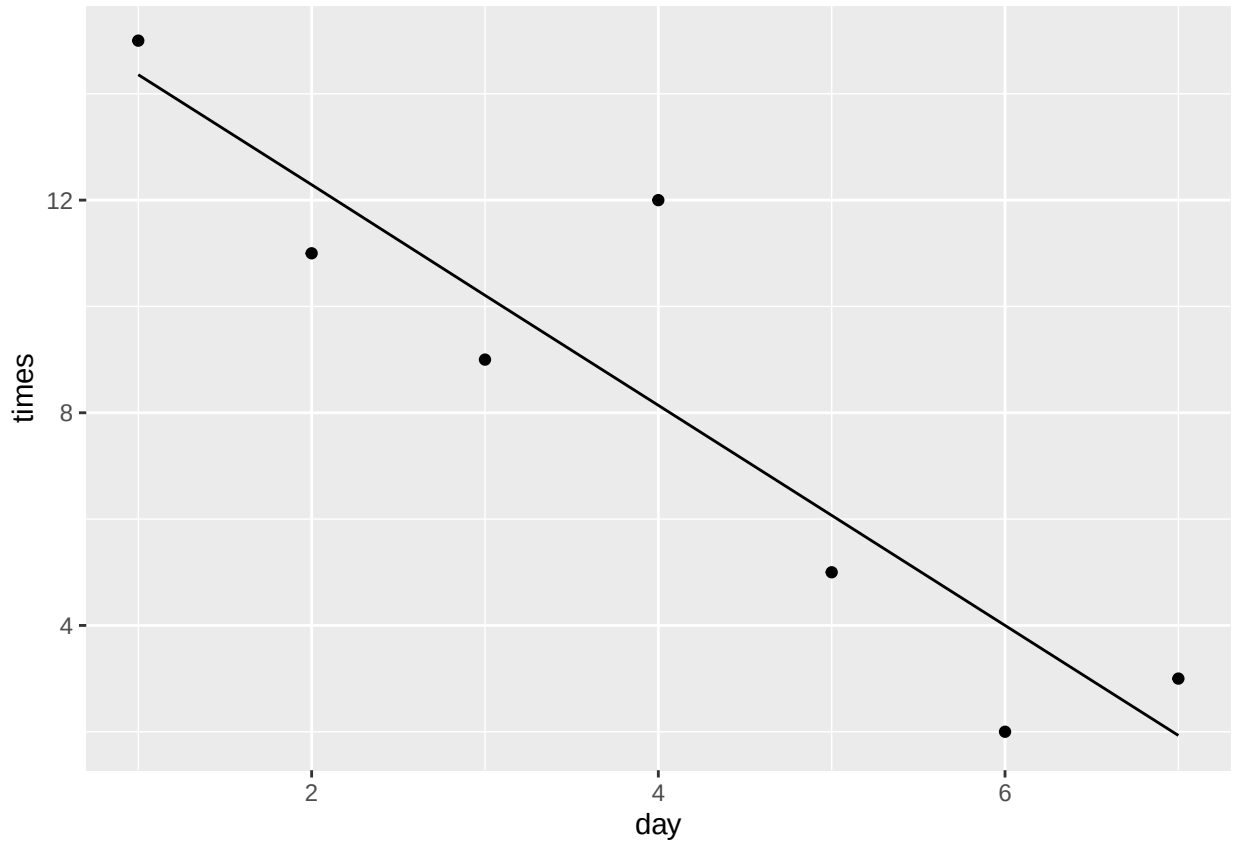
Using ggplot() and geom_point(), create a scatterplot with day along the x-axis and times along the y-axis.

```
scatter_plot <- ggplot( data=Rosy, aes( x=day, y=times ) ) +
  geom_point()

scatter_plot
```



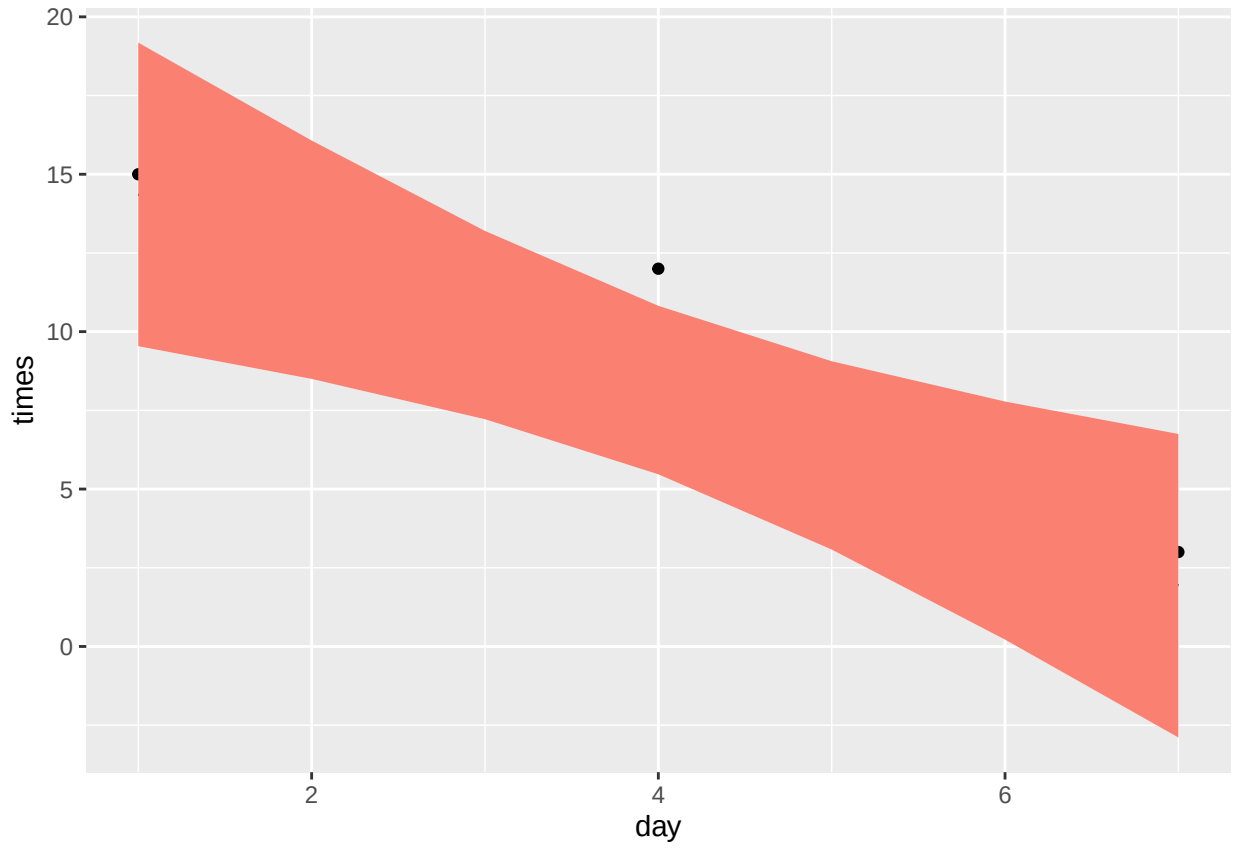
b. Add a line to the graph where the x-values are the day values but now the y-values are the predicted values which we've called yhat. Notice that you have to set the aesthetic y=times for the points and y=yhat for the line. Because each geom_ will accept an aes() command, you can specify the y attribute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

```
scatter_plot <- scatter_plot +
  geom_line( aes( x=day, y=yhat ) )

scatter_plot
```



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.

```
scatter_plot <- scatter_plot +
  geom_ribbon( aes( ymin=lwr, ymax=upr ), fill='salmon' )

scatter_plot
```

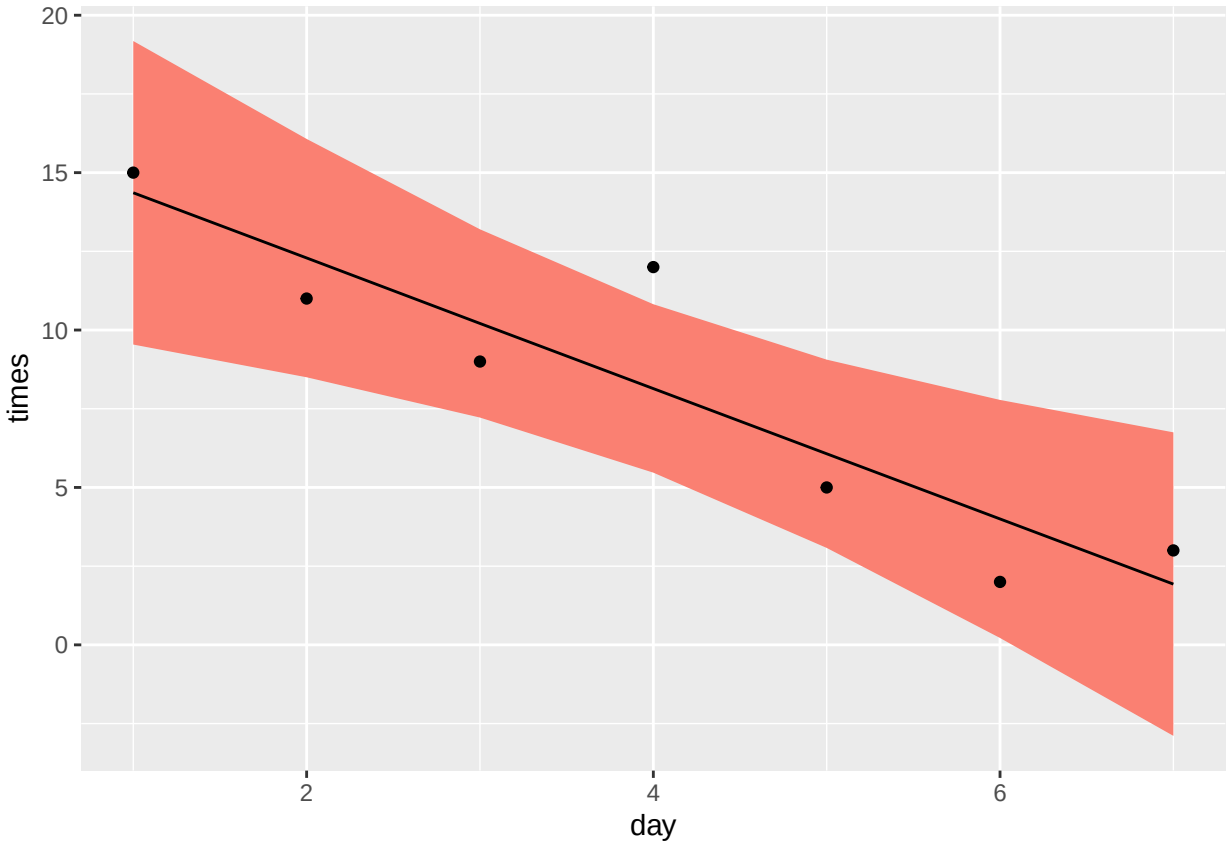


d. What happened when you added the ribbon? Did some points get hidden? If so, why?

All the points within the ribbon were covered up by the ribbon. The only point which was not covered up was an outlier which is outside of the ribbon. This most likely happened because the ribbon was placed on top of the points, instead of the points being added on top of the ribbon.

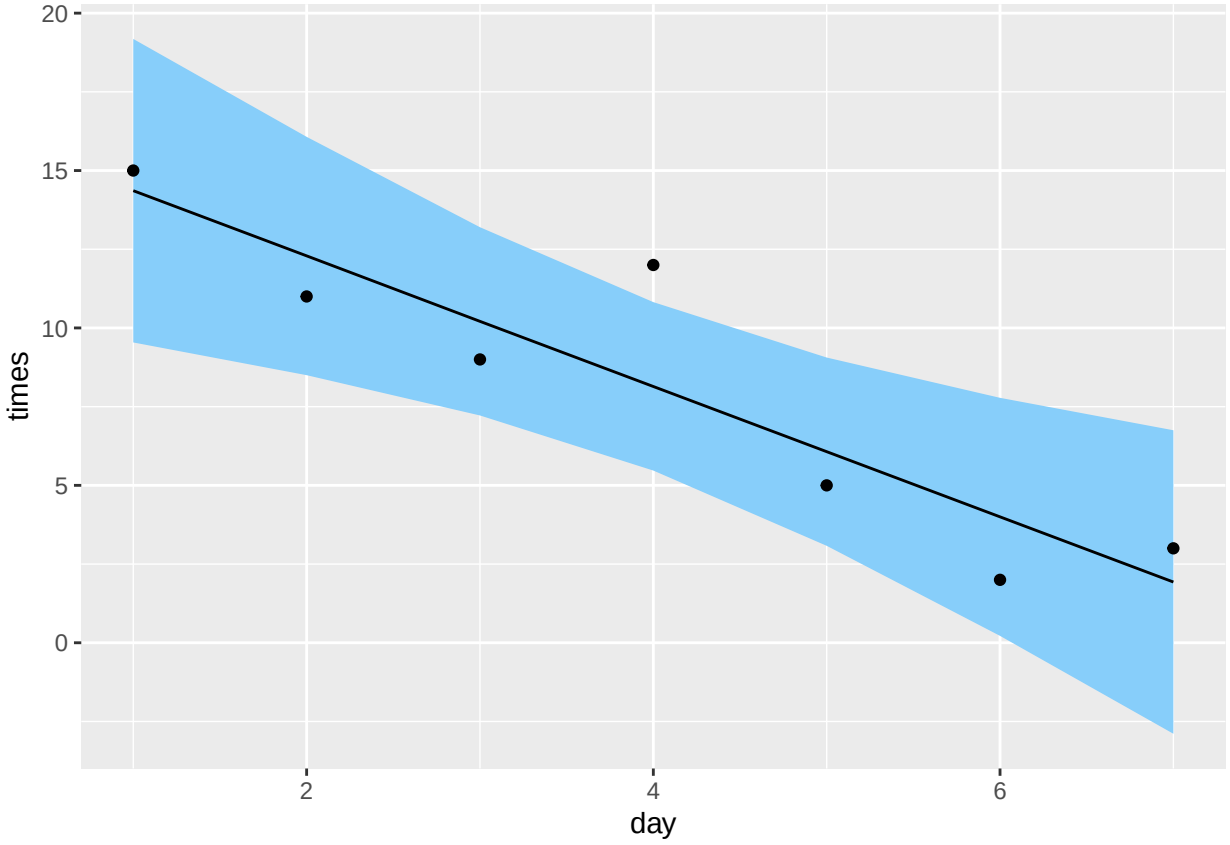
e. Reorder the statements that created the graph so that the ribbon is on the bottom and the data points are on top and the regression line is visible.

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

scatter_plot
```



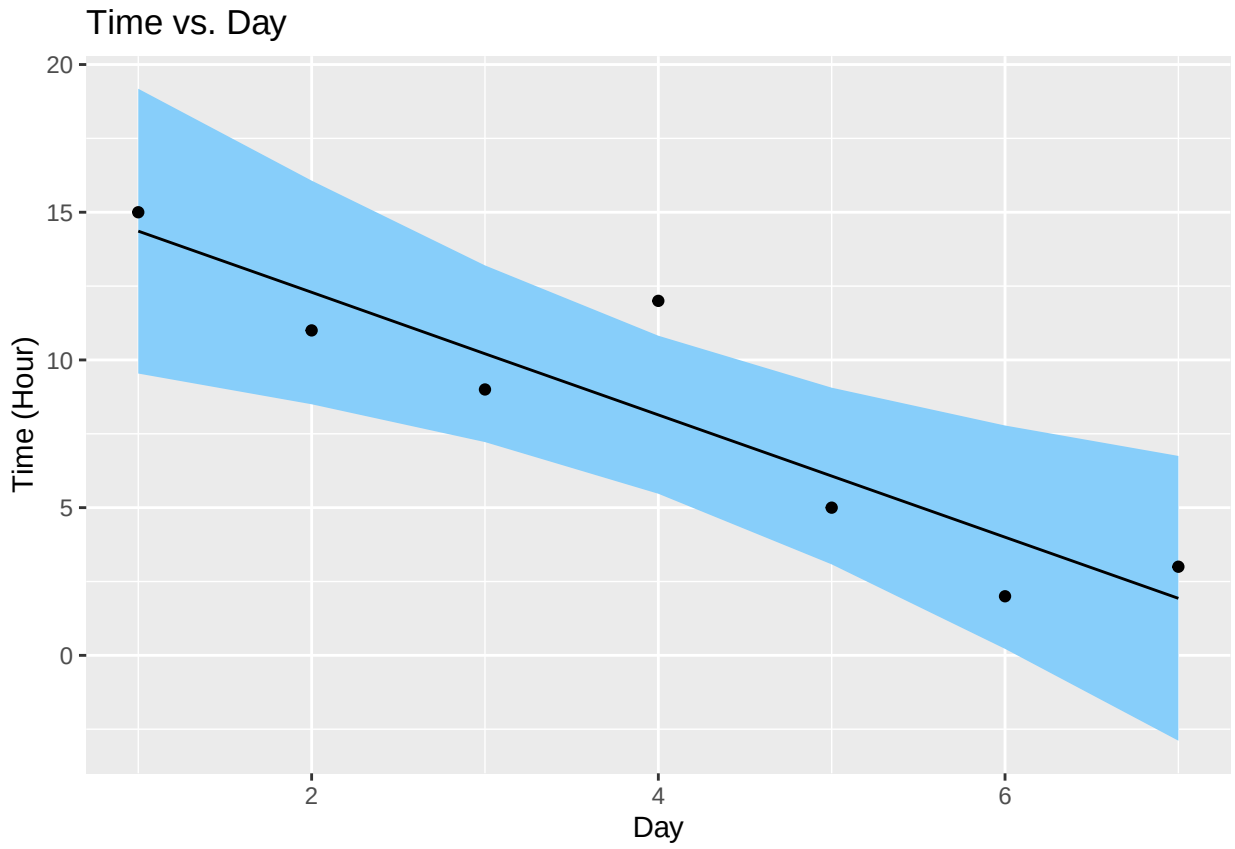
f. The color of the ribbon fill is ugly. Use Google to find a list of named colors available to ggplot2. For example, I googled "ggplot2 named colors" and found the following link: http://sape.inf.usi.ch/quick-reference/ggplot2/colour. Choose a color for the fill that is pleasing to you.

```
scatter_plot <- ggplot( data=Rosy, aes( x=day, y=times ) ) +
  geom_ribbon( aes( ymin=lwr, ymax=upr ), fill='lightskyblue' ) +
  geom_point() +
  geom_line( aes( x=day, y=yhat ) )

scatter_plot
```



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

```
scatter_plot <- scatter_plot +
  labs( title= 'Time vs. Day' ) +
  labs( x='Day' ) +
  labs( y= 'Time (Hour)' )

scatter_plot
```



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.

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.1.3
```

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)
```

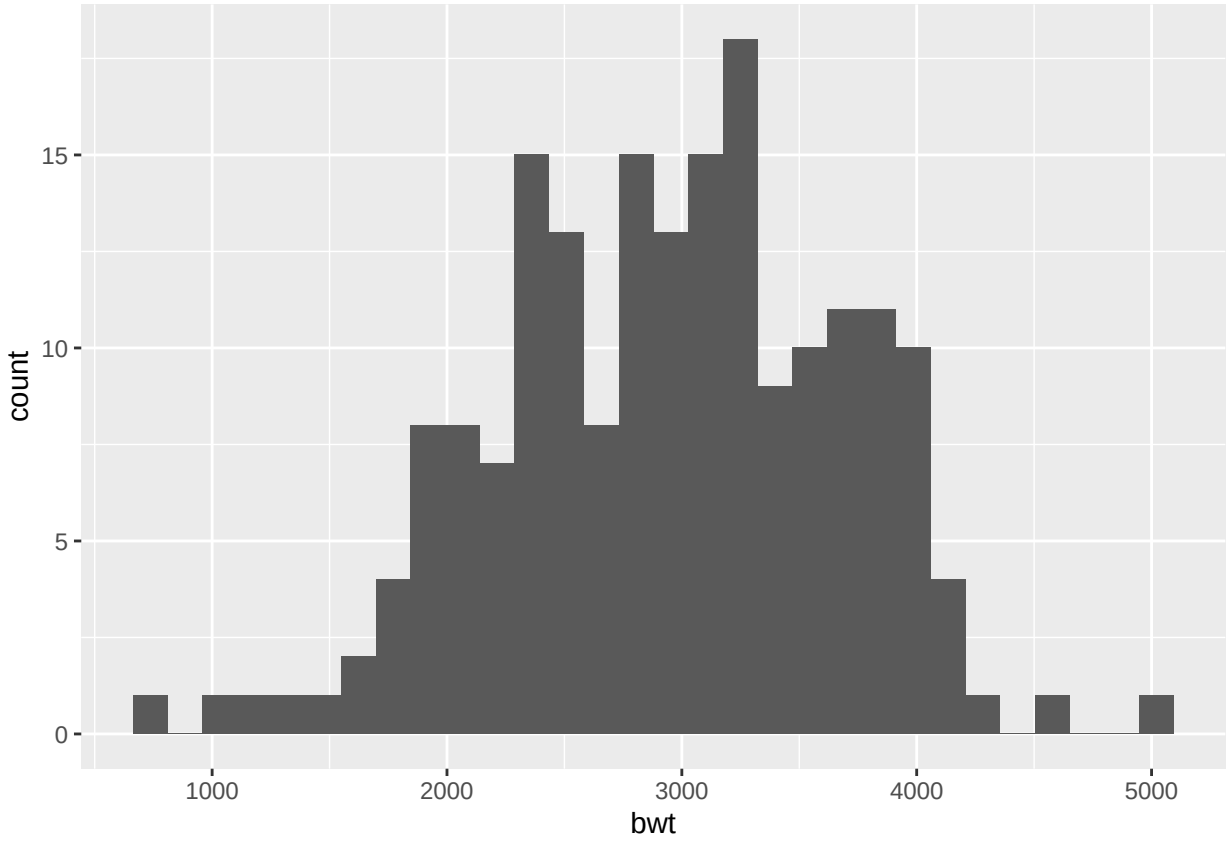
```
## Warning: package 'tidyverse' was built under R version 4.1.3
## Warning: package 'tibble' was built under R version 4.1.3
## Warning: package 'tidyr' was built under R version 4.1.3
## Warning: package 'readr' was built under R version 4.1.3
## Warning: package 'purrr' was built under R version 4.1.3
## Warning: package 'dplyr' was built under R version 4.1.3
## Warning: package 'stringr' was built under R version 4.1.3
## Warning: package 'forcats' was built under R version 4.1.3
## Warning: package 'lubridate' was built under R version 4.1.3
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.1     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.1     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

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

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

```
histogram <- ggplot( data=birthwt, aes( x=bwt ) ) +
  geom_histogram()

histogram
```

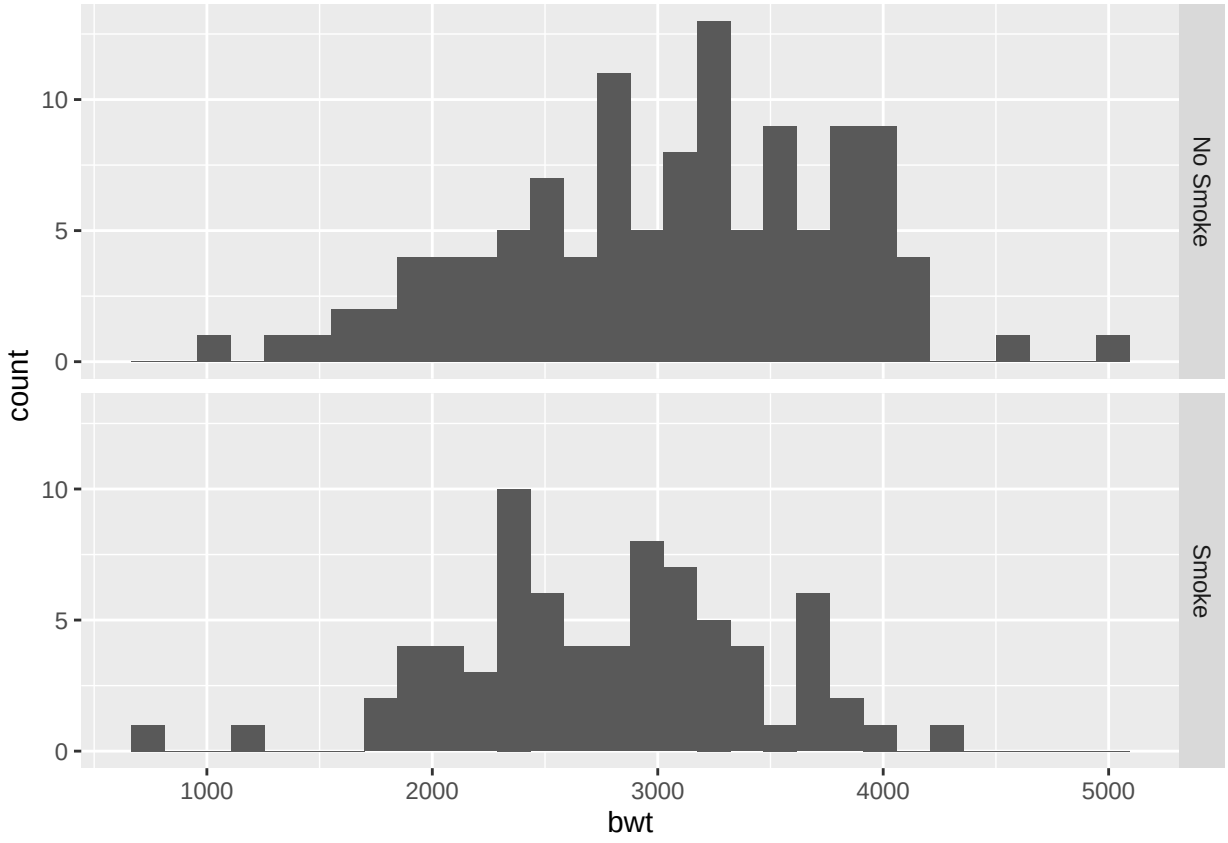
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
histogram <- histogram +
  facet_grid( vars(smoke) )

histogram
```

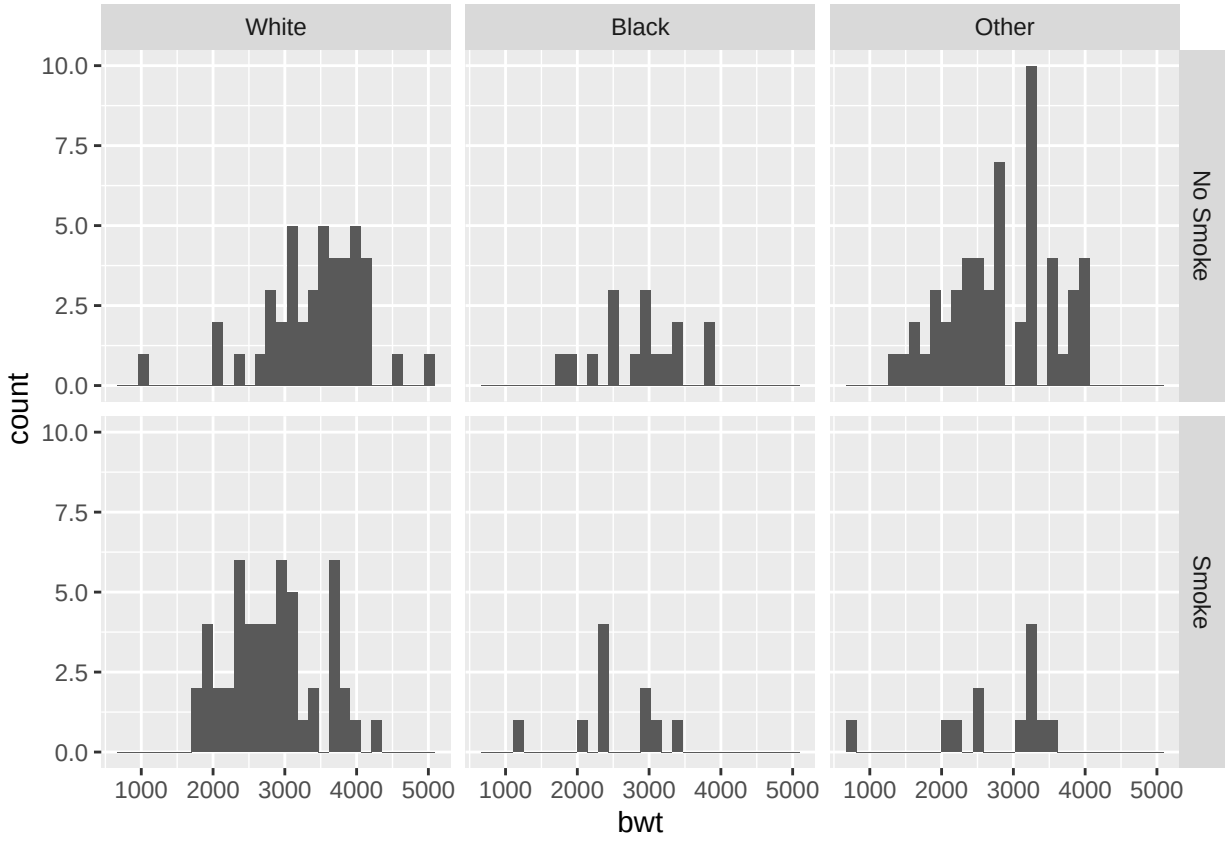
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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).

```
histogram <- histogram +
  facet_grid( rows=vars( smoke ), cols=vars( race ) )

histogram
```

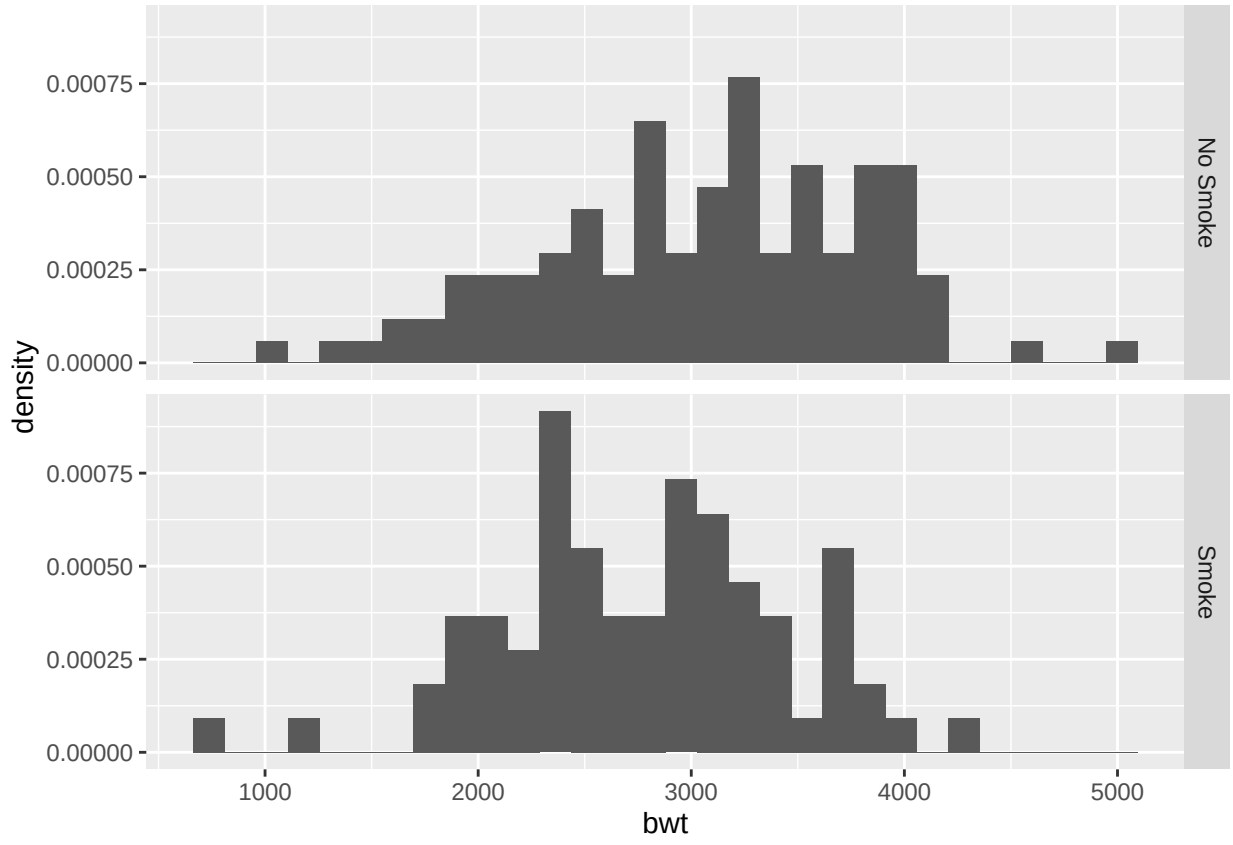
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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.

```
histogram <- ggplot( data=birthwt, aes( x=bwt, y=after_stat(density) ) ) +
  geom_histogram() +
  facet_grid( vars(smoke) )

histogram
```

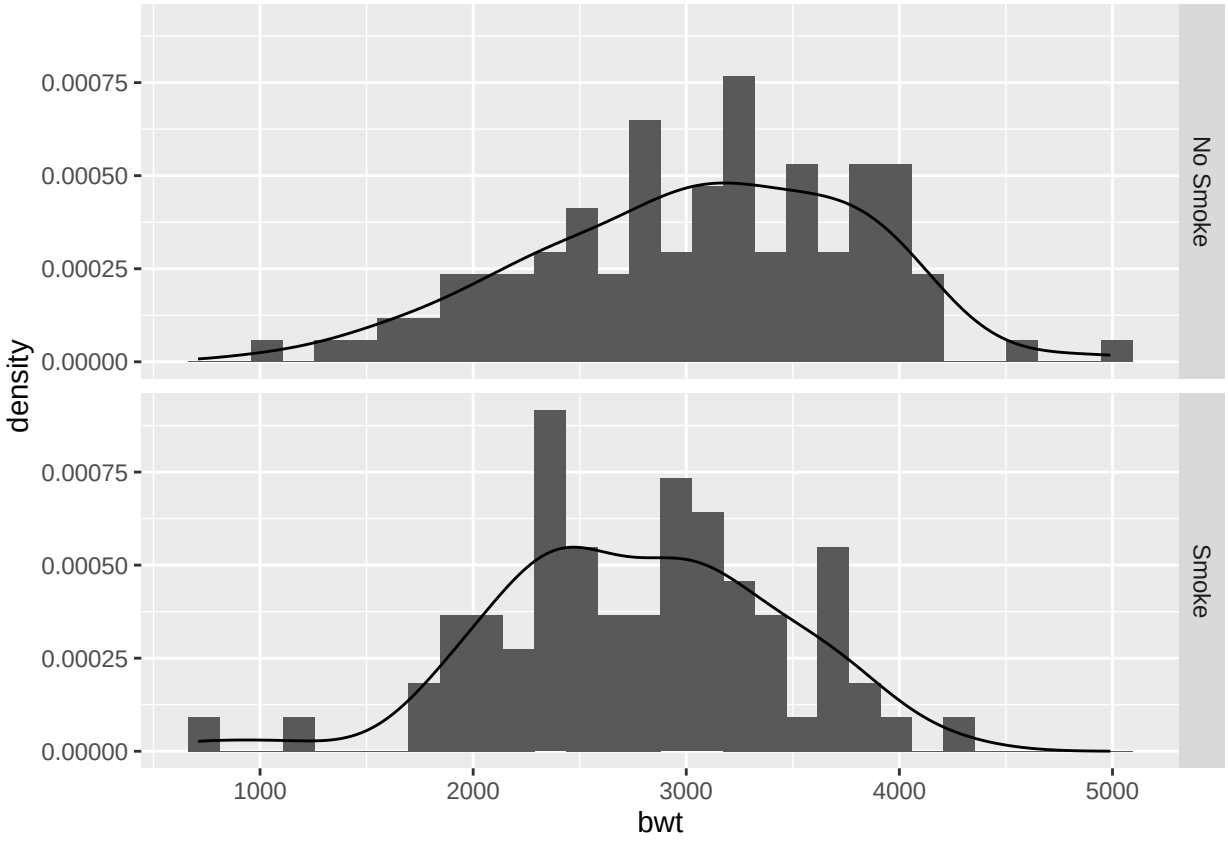
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
histogram <- histogram +
  geom_density()

histogram
```

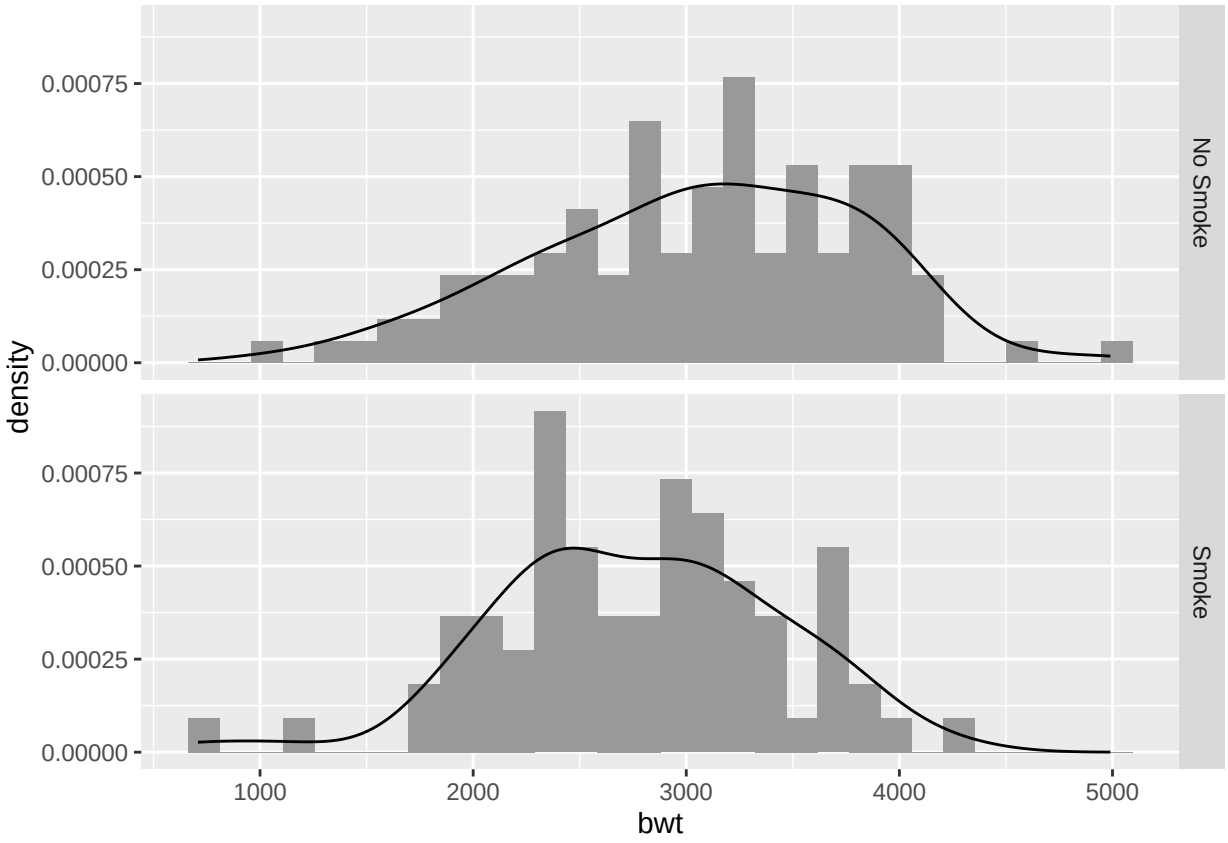
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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/.

```
histogram <- ggplot( data=birthwt, aes( x=bwt, y=after_stat( density ) ) ) +
  geom_histogram( fill='grey60' ) +
  facet_grid( vars(smoke) ) +
  geom_density()

histogram
```

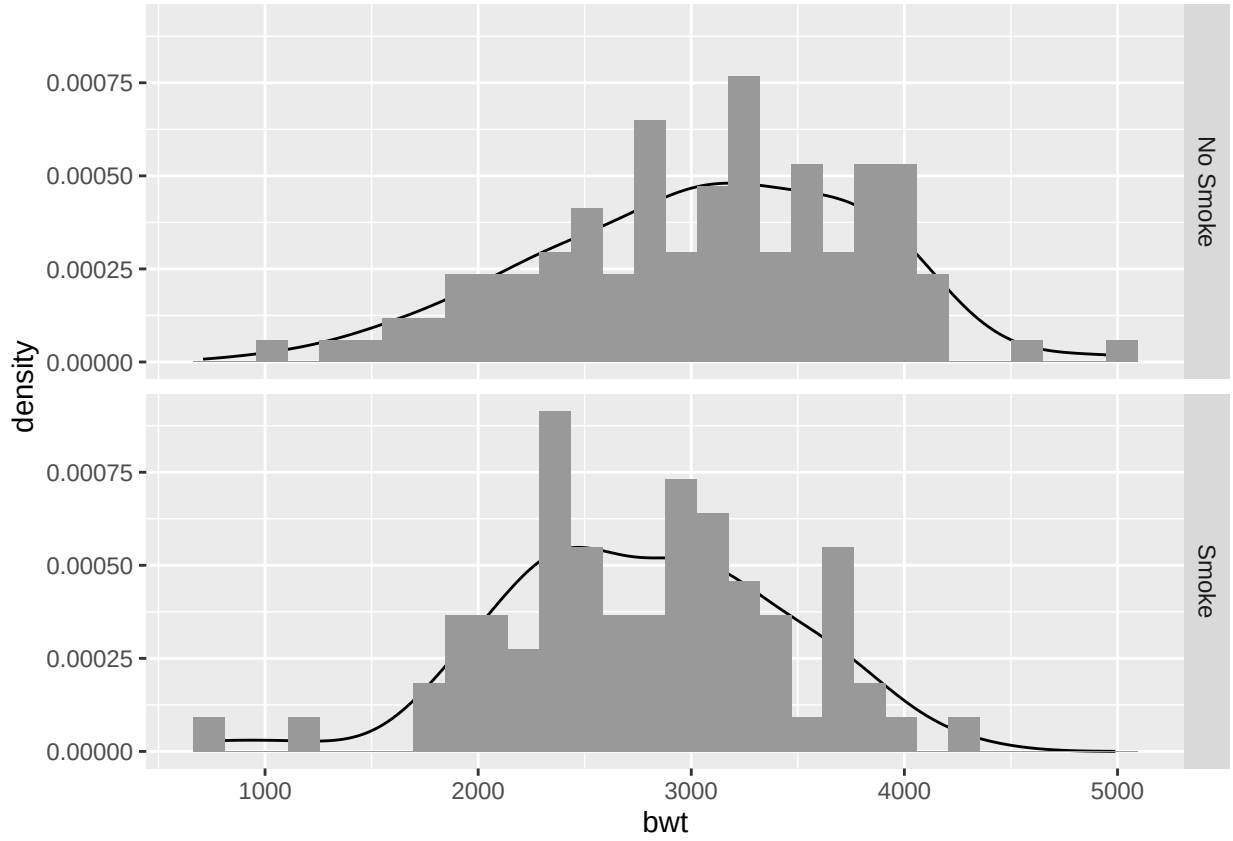
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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?

```
histogram <- ggplot( data=birthwt, aes( x=bwt, y=after_stat( density ) ) ) +
  geom_density() +
  geom_histogram( fill='grey60' ) +
  facet_grid( vars(smoke) )

histogram
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The order of the density line and the chart does matter, as having the chart on top of the density line obscures the line in many places. Personally, I prefer to have the line on top so it can always be seen.

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?

It is better to have the charts facetted vertically rather than horizontally. This allows their X-axes to line up, and it is visually apparent the different in birthweight between the two charts.

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.

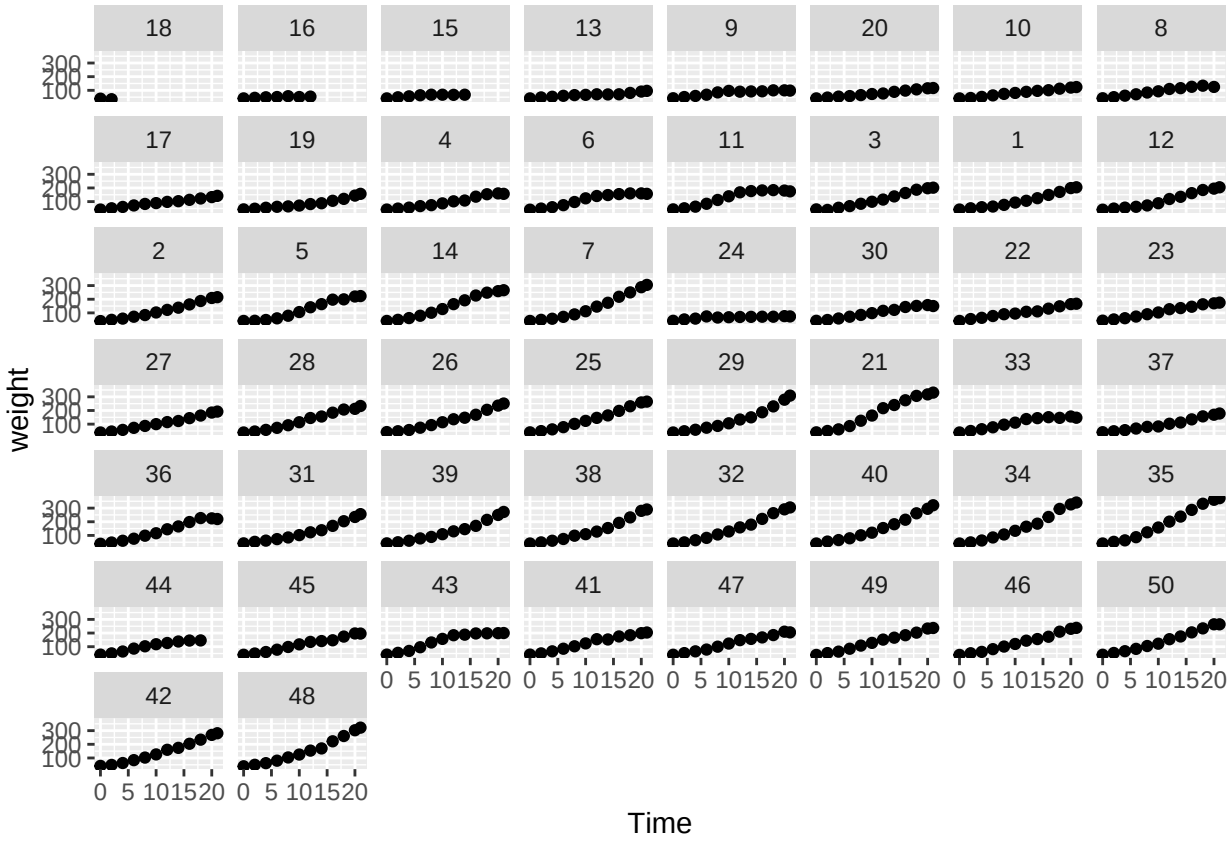
```
summary(ChickWeight)
```

```
##      weight           Time           Chick     Diet
##  Min.   : 35.0   Min.   : 0.00   13     : 12   1:220
##  1st Qu.: 63.0   1st Qu.: 4.00   9      : 12   2:120
##  Median :103.0   Median :10.00   20     : 12   3:120
##  Mean   :121.8   Mean   :10.72   10     : 12   4:118
##  3rd Qu.:163.8   3rd Qu.:16.00   17     : 12
##  Max.   :373.0   Max.   :21.00   19     : 12
##                                  (Other):506
```

```
?ChickWeight
```

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.

```
chick_scatter <- ggplot( data=ChickWeight, aes( x=Time, y=weight ) ) +
  geom_point() +
  facet_wrap( vars(Chick) )

chick_scatter
```



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)
```