Module3

Joe Vargovich

8/26/2020

```
library(tidyverse)
library(MASS)
```

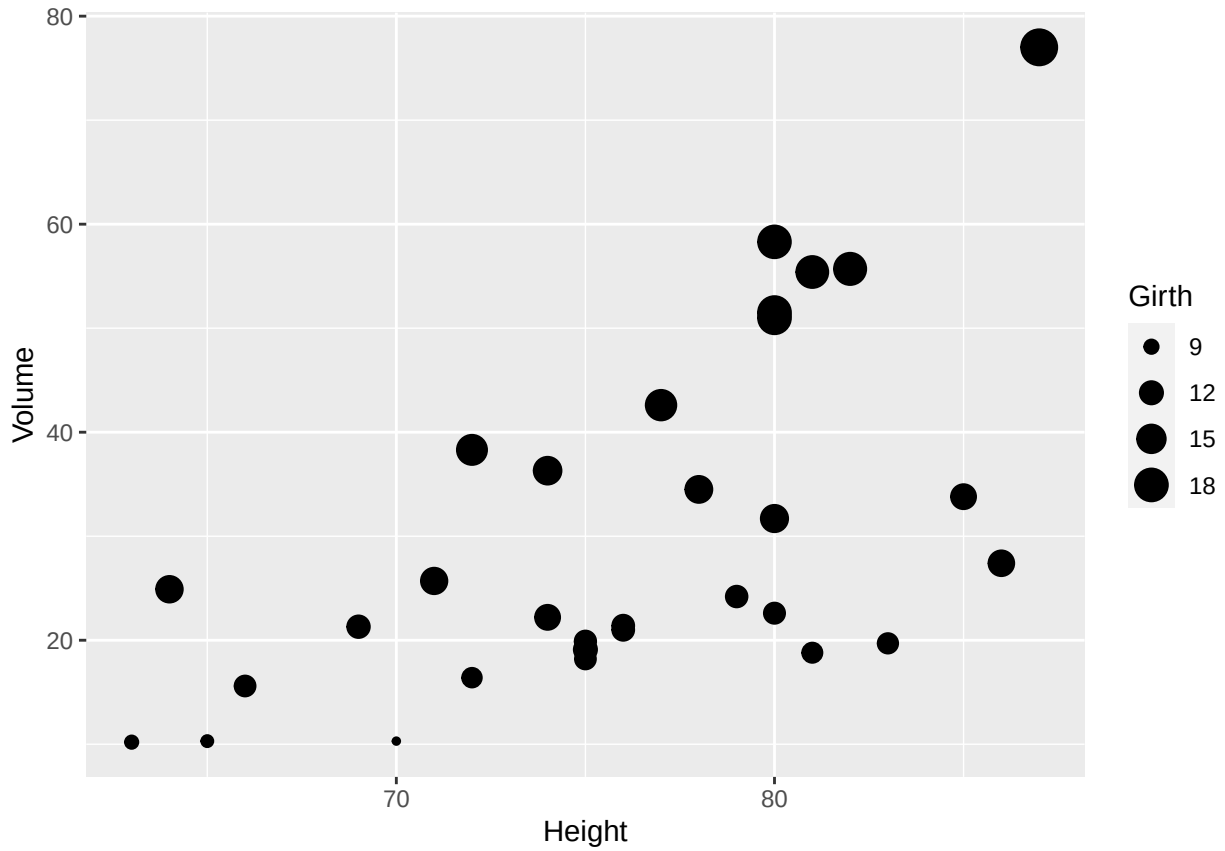
Exercise 1

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 scatterplot 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?

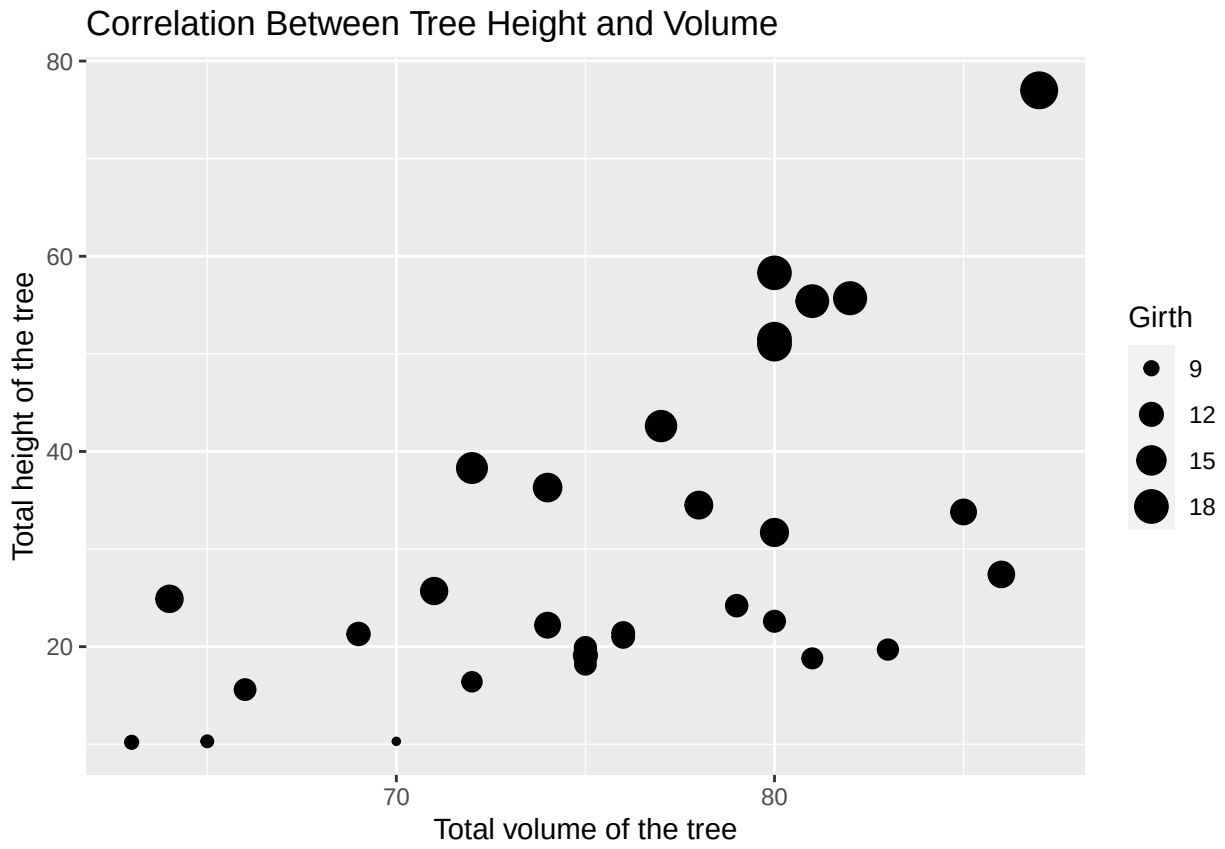
Size is far more intuitive than color, see the plot below.

```
treesPlot = ggplot(trees, aes(x = Height, y = Volume)) +
  geom_point(aes(size = Girth))
treesPlot
```



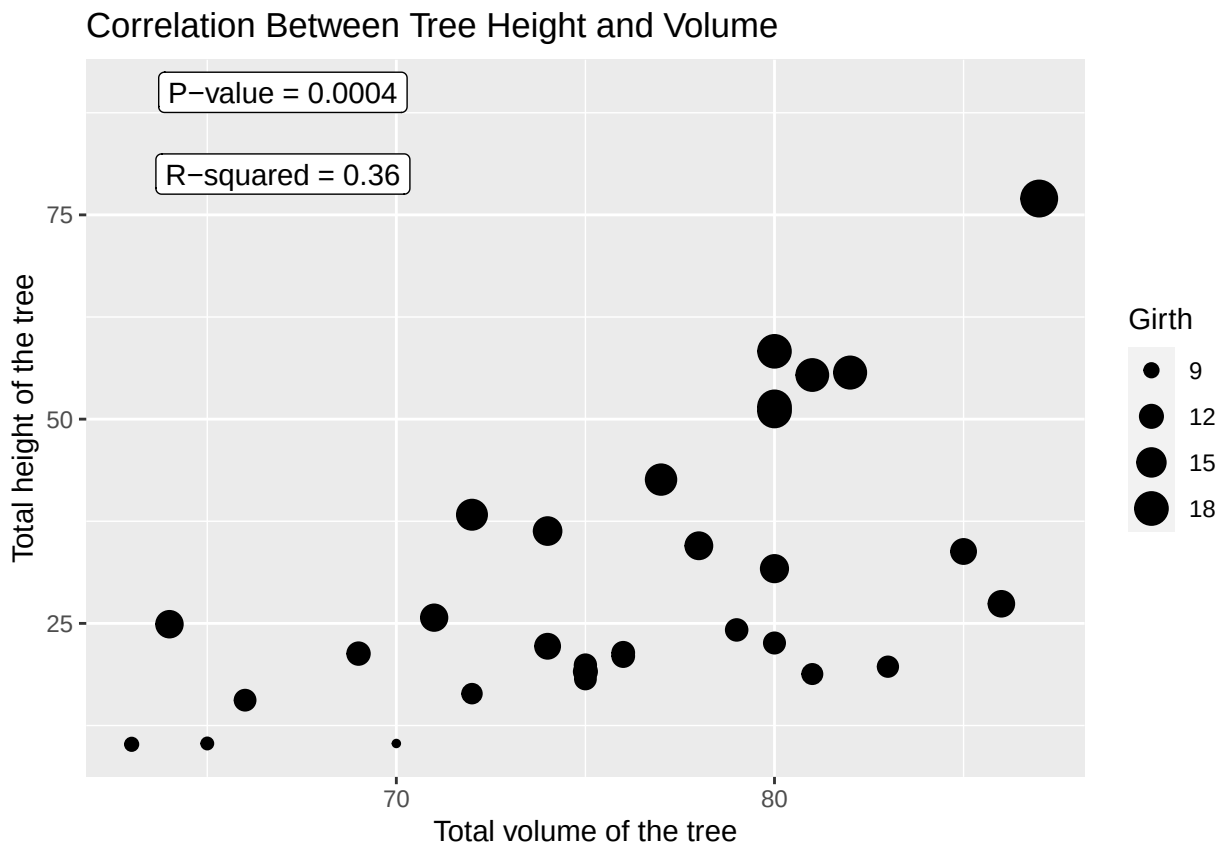
b. Add appropriate labels for the main title and the x and y axes.

```
treesPlotLabeled = treesPlot +
  labs(title= 'Correlation Between Tree Height and Volume') +
  labs(x = 'Total volume of the tree', y = 'Total height of the tree')
treesPlotLabeled
```



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.

```
treesPlotLabeled = treesPlotLabeled +
  annotate('label', x = 67, y=80, size=4, color='black', label="R-squared = 0.36") +
  annotate('label', x = 67, y=90, size=4, color='black', label="P-value = 0.0004")
treesPlotLabeled
```



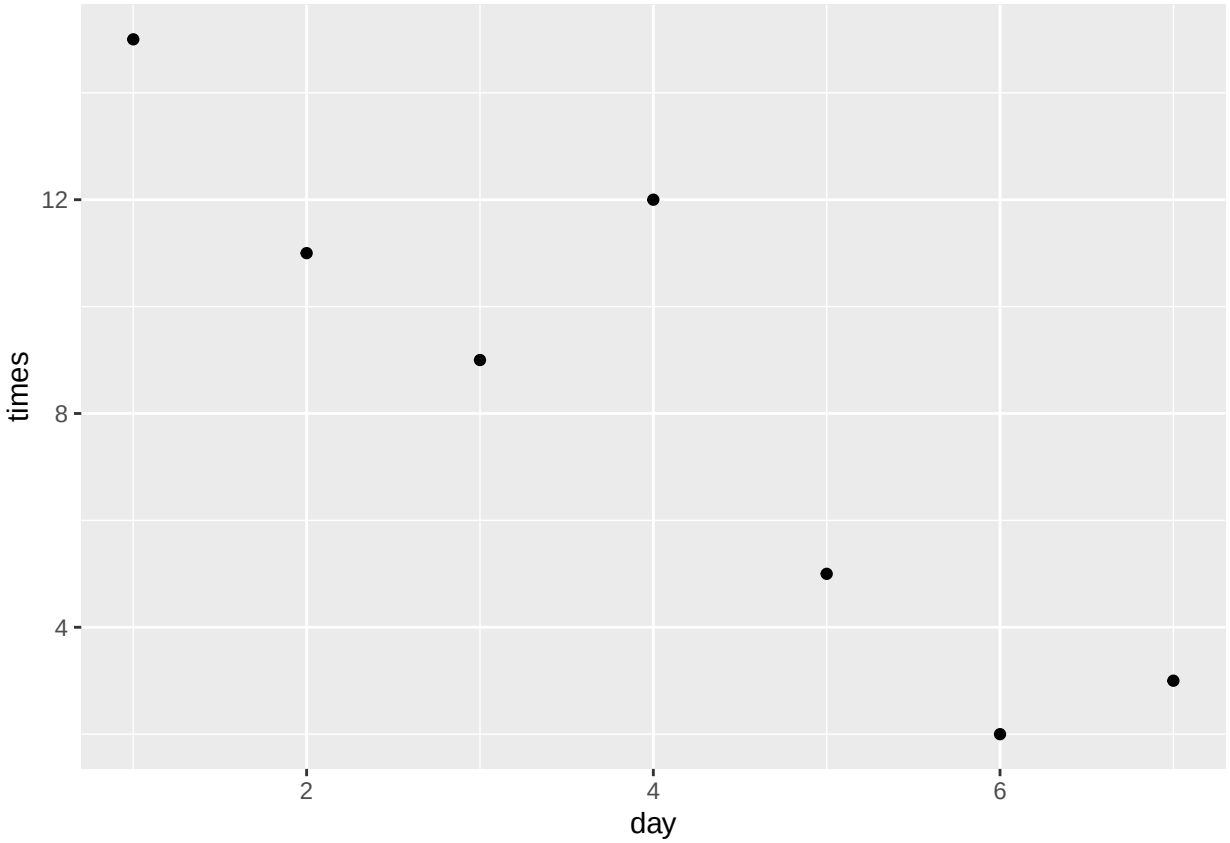
Exercise 2

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.

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

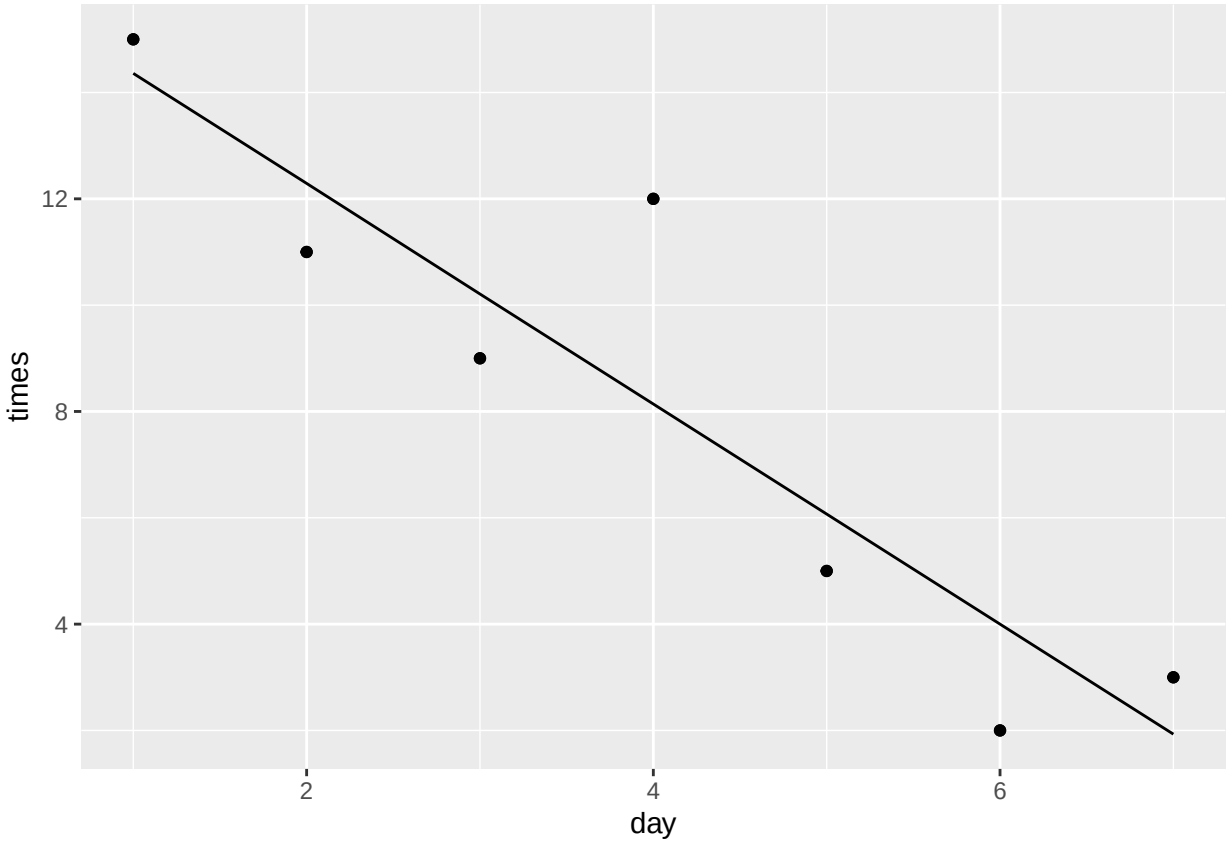
```
Rosy <- data.frame(
  times = c(15, 11, 9, 12, 5, 2, 3),
  day   = 1:7,
  yhat  = c(14.36, 12.29, 10.21, 8.14, 6.07, 4.00,  1.93),
  lwr   = c( 9.54,  8.5,   7.22, 5.47, 3.08, 0.22, -2.89),
  upr   = c(19.18, 16.07, 13.2, 10.82, 9.06, 7.78,  6.75))

rosyGraph = ggplot(Rosy, aes(x=day, y = times) ) +
  geom_point()
rosyGraph
```



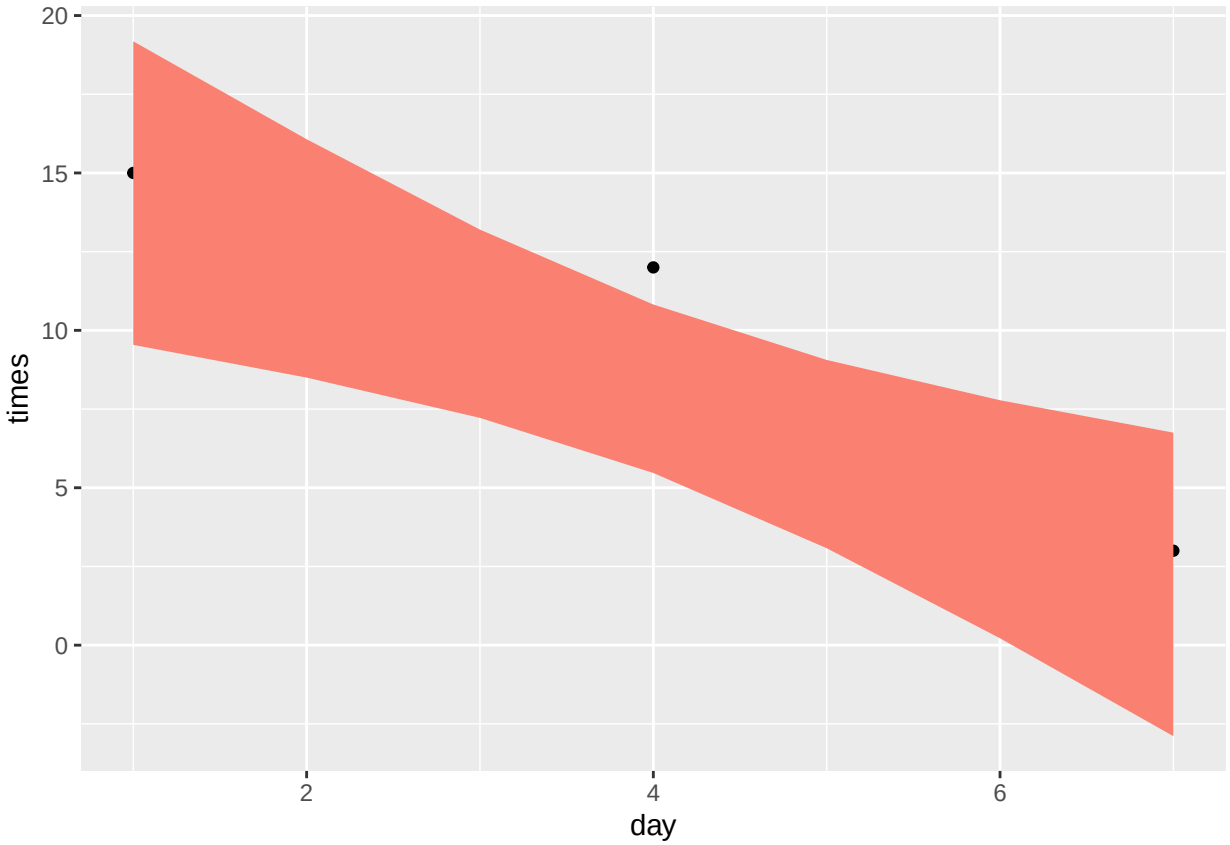
b. Add a line to the graph where the x-values are the day values but now the y-values are the predicted values which we've called yhat. Notice that you have to set the aesthetic y=times for the points and y=yhat for the line. Because each geom_ will accept an aes() command, you can specify the y attribute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

```
rosyGraph + geom_point(aes(y=times)) +
  geom_line( aes(y=yhat))
```



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.

```
rosyGraph + geom_ribbon( aes(ymin=lwr, ymax=upr), fill='salmon')
```

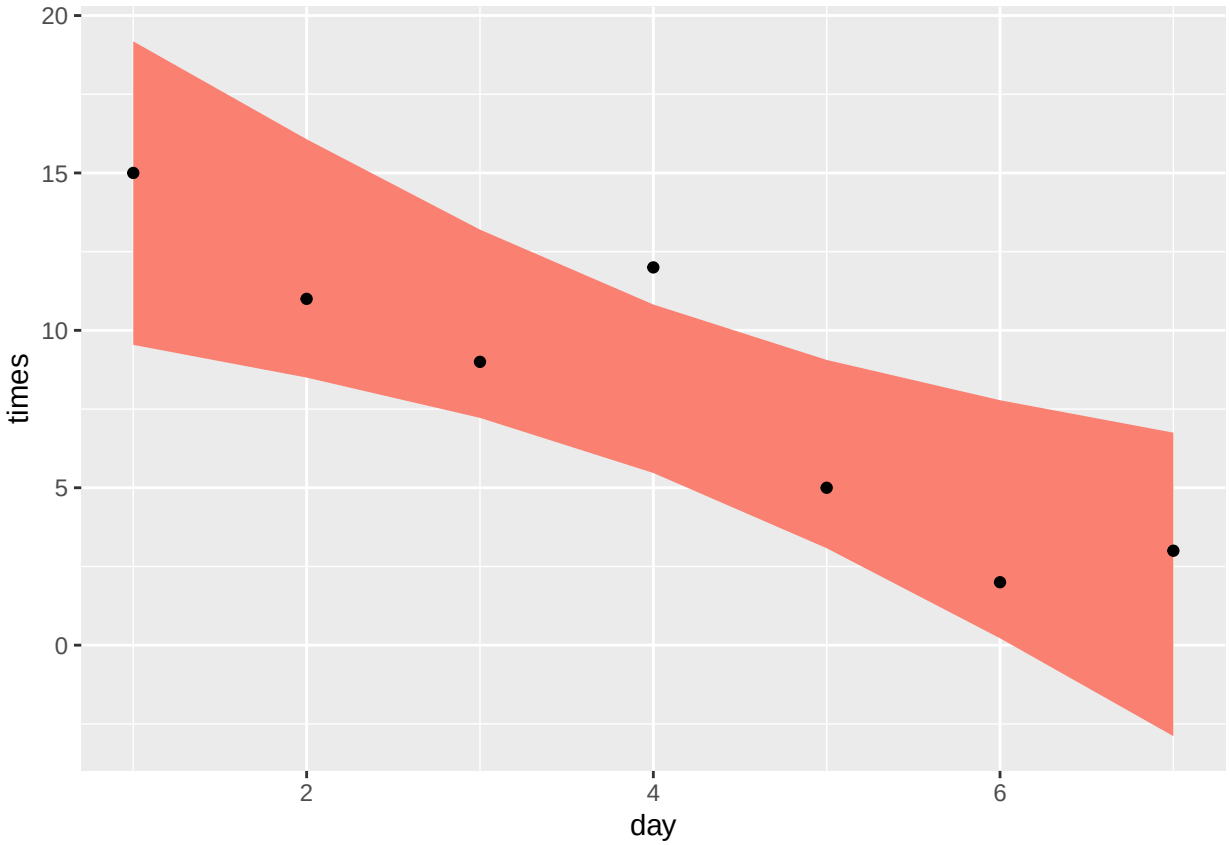


d. What happened when you added the ribbon? Did some points get hidden? If so, why?

The ribbon covers the data points as they lie under the range that is displayed by the solid color ribbon. The ribbon takes priority over the geompoints as it was added later than the geom_point statements.

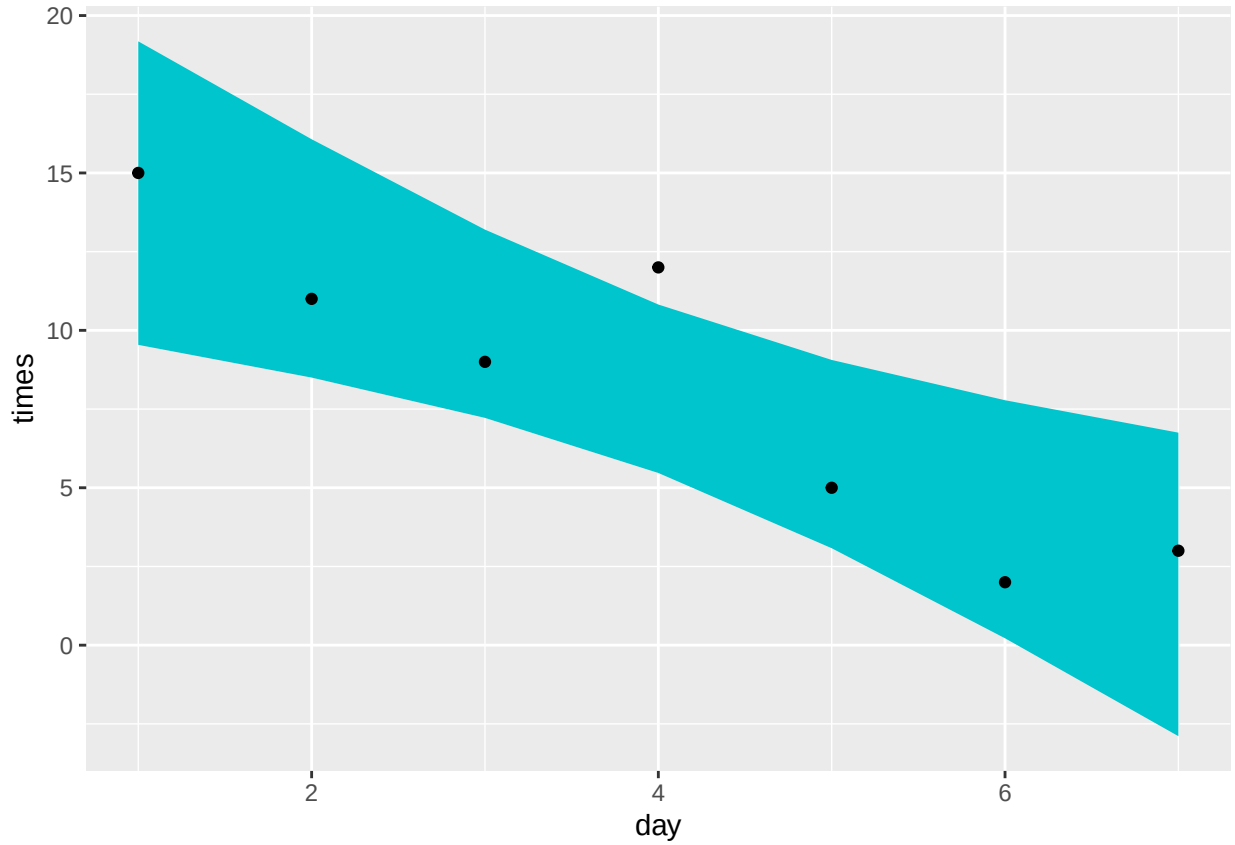
e. Reorder the statements that created the graph so that the ribbon is on the bottom and the data points are on top and the regression line is visible.

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



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.

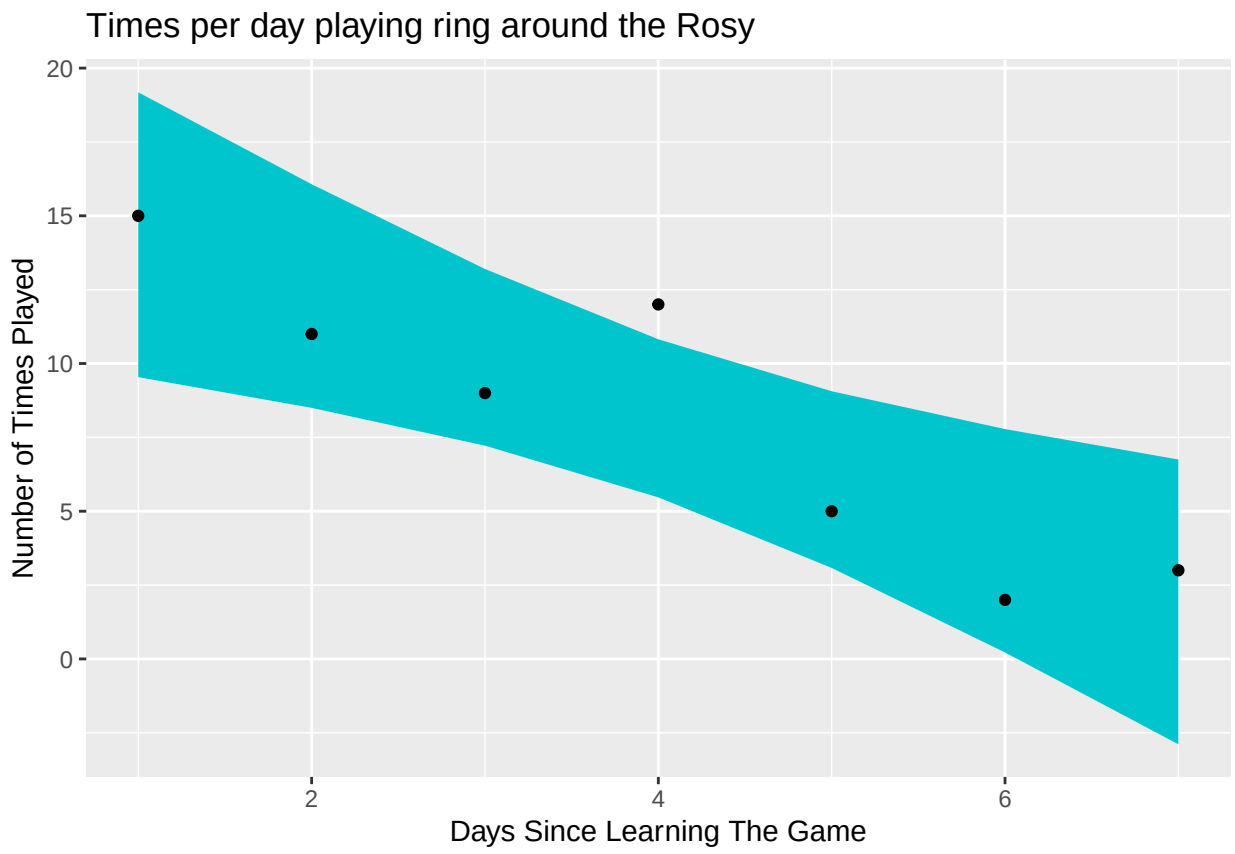
```
ggplot(Rosy, aes(x=day, y = times) ) +
  geom_ribbon( aes(ymin=lwr, ymax=upr), fill='turquoise3') +
  geom_point()
```



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

```
labeledPlot = ggplot(Rosy, aes(x=day, y = times) ) +
  geom_ribbon( aes(ymin=lwr, ymax=upr), fill='turquoise3') +
  geom_point() +
  labs(title= 'Times per day playing ring around the Rosy') +
  labs(x = 'Days Since Learning The Game', y = 'Number of Times Played')

labeledPlot
```



Exercise 3

We'll next make some density plots that relate several factors towards the birth weight of a child

a. Load the birthwt by either using the data() command or loading the MASS library

```
data(birthwt)
```

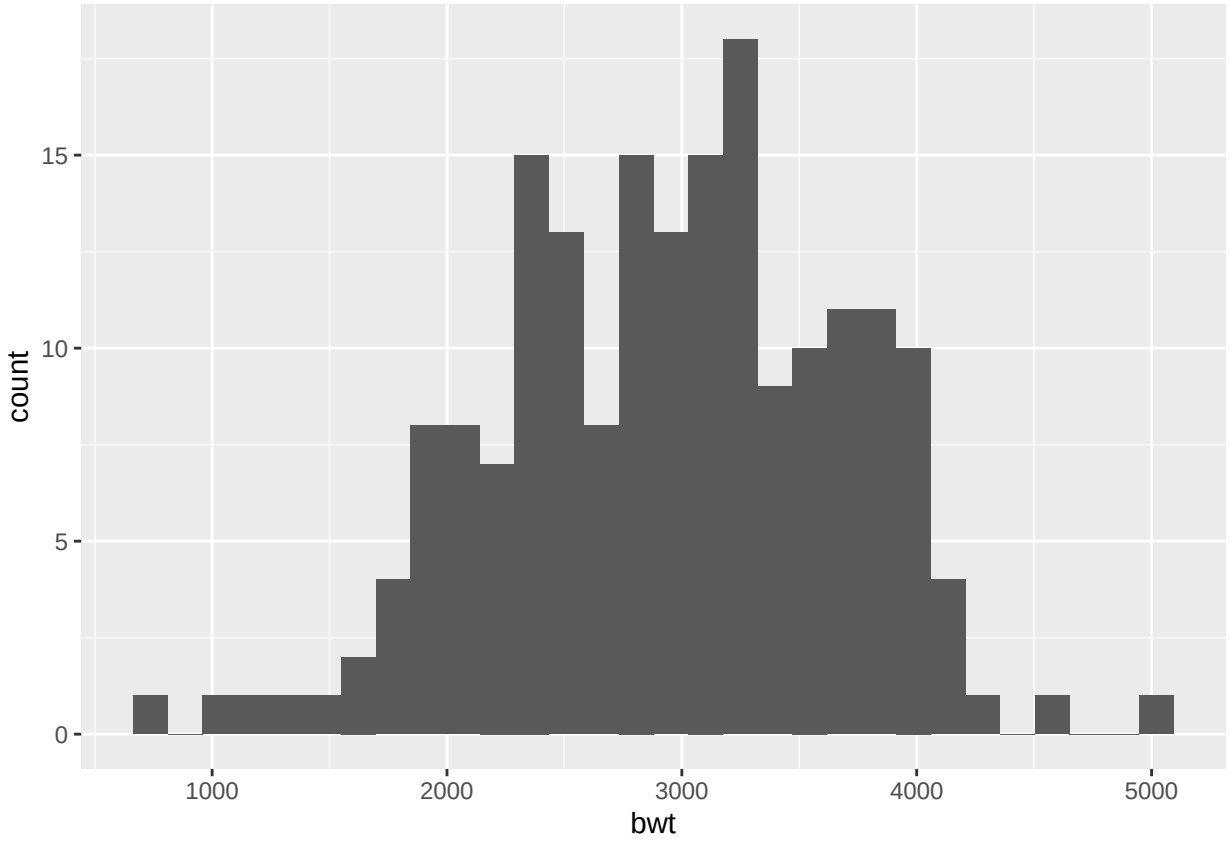
b. Read the help file for the dataset using MASS::birthwt. The covariates race and smoke are not stored in a user friendly manner. For example, smoking status is labeled using a 0 or a 1. Because it is not obvious which should represent that the mother smoked, we'll add better labels to the race and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

```
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

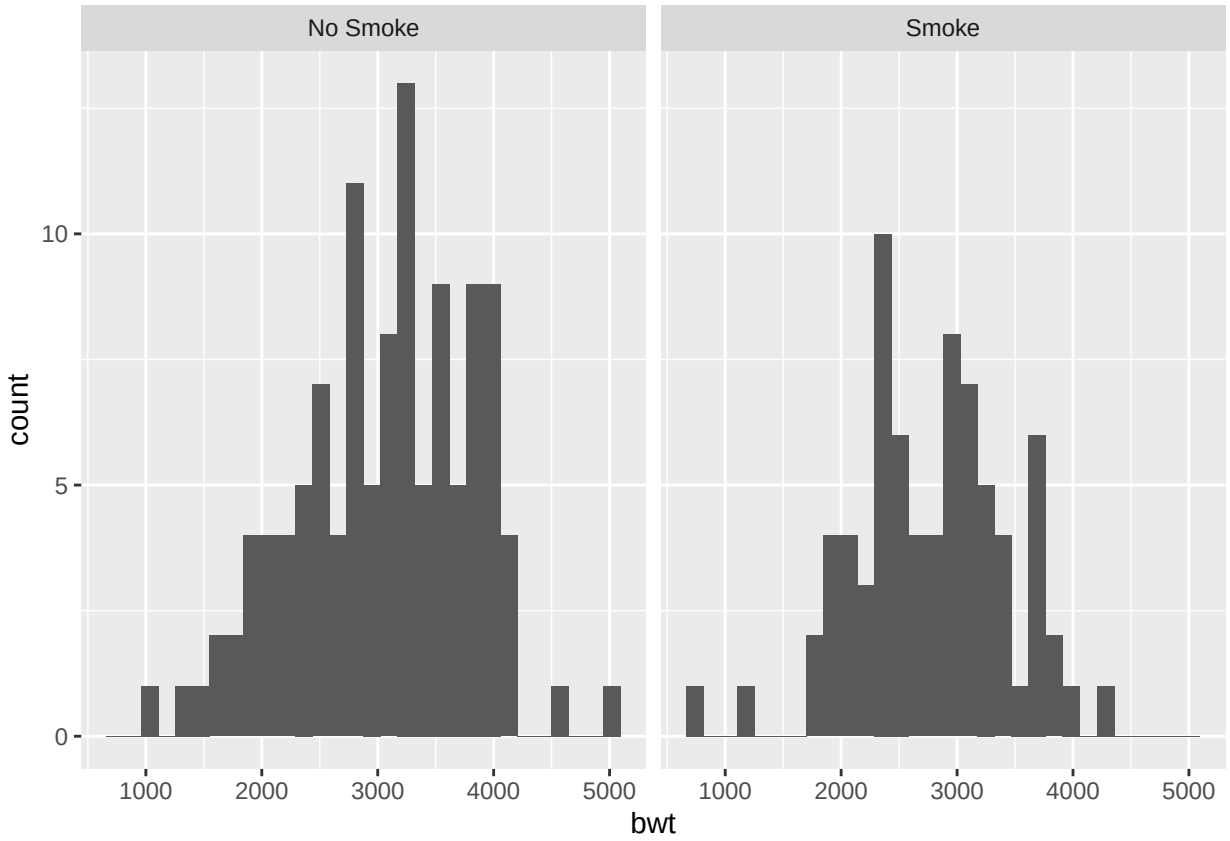
```
birthHistogram = ggplot(birthwt, aes(x=bwt)) +
  geom_histogram()

birthHistogram
```



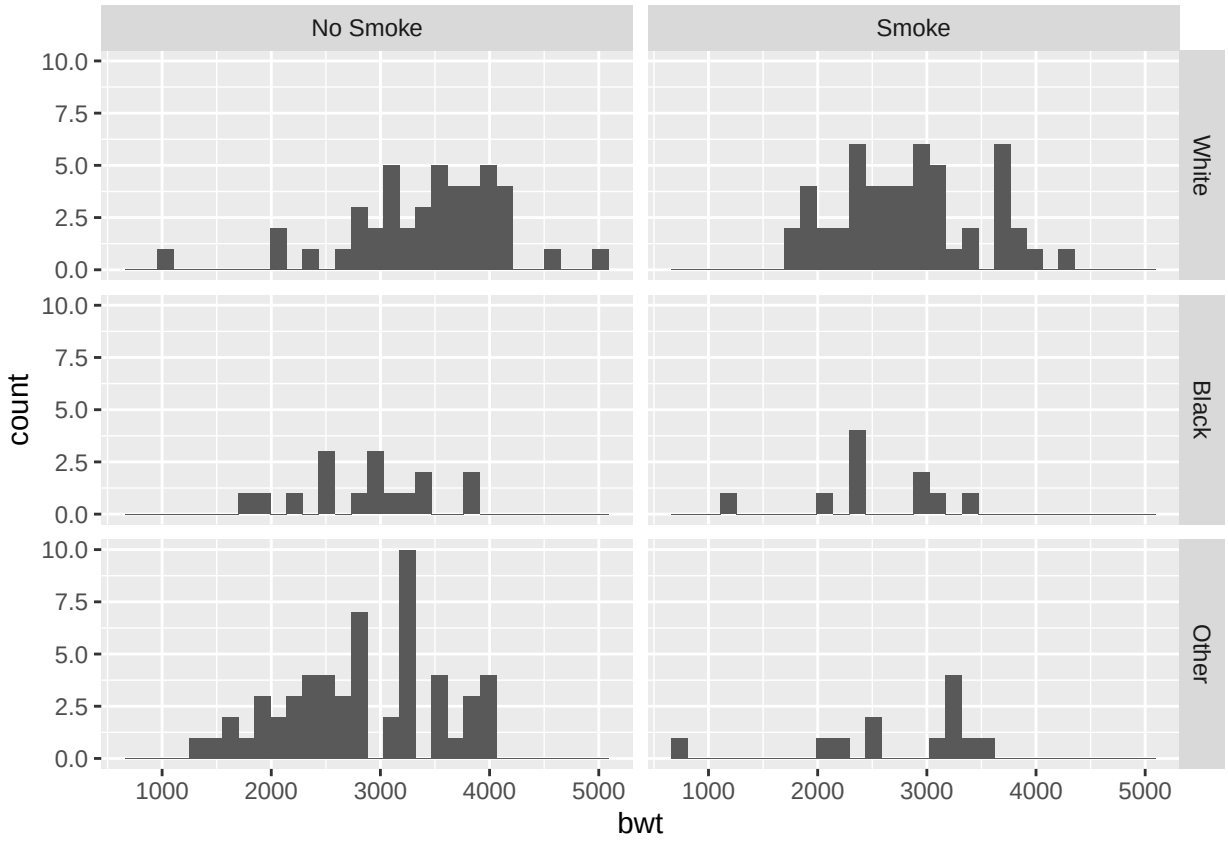
d. Make separate graphs that denote whether a mother smoked during pregnancy by appending + facet_grid() command to your original graphing command.

```
birthHistogram +
  facet_grid(~ smoke)
```



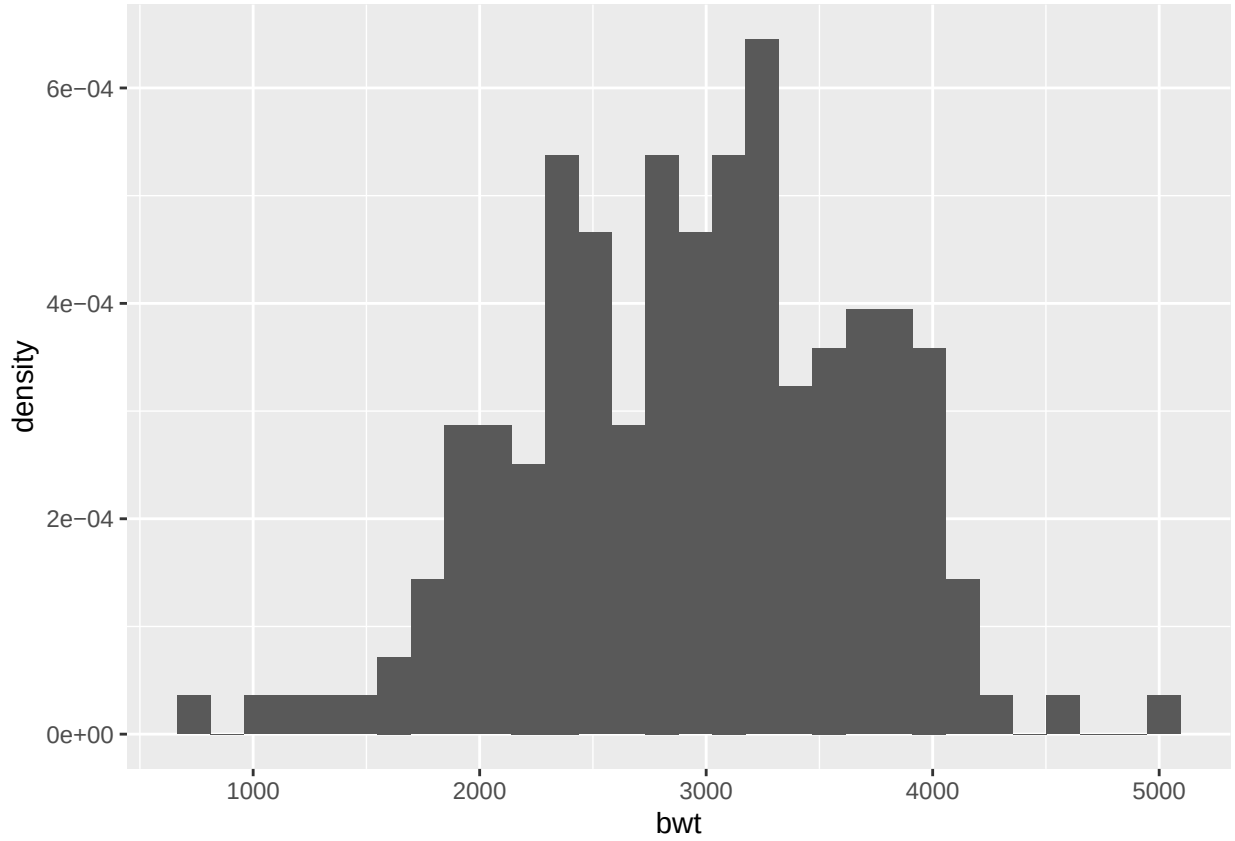
e. Perhaps race matters in relation to smoking. Make our grid of graphs vary with smoking status changing vertically, and race changing horizontally (that is the formula in facet_grid() should have smoking be the y variable and race as the x).

```
birthHistogram +
  facet_grid(race ~ smoke)
```



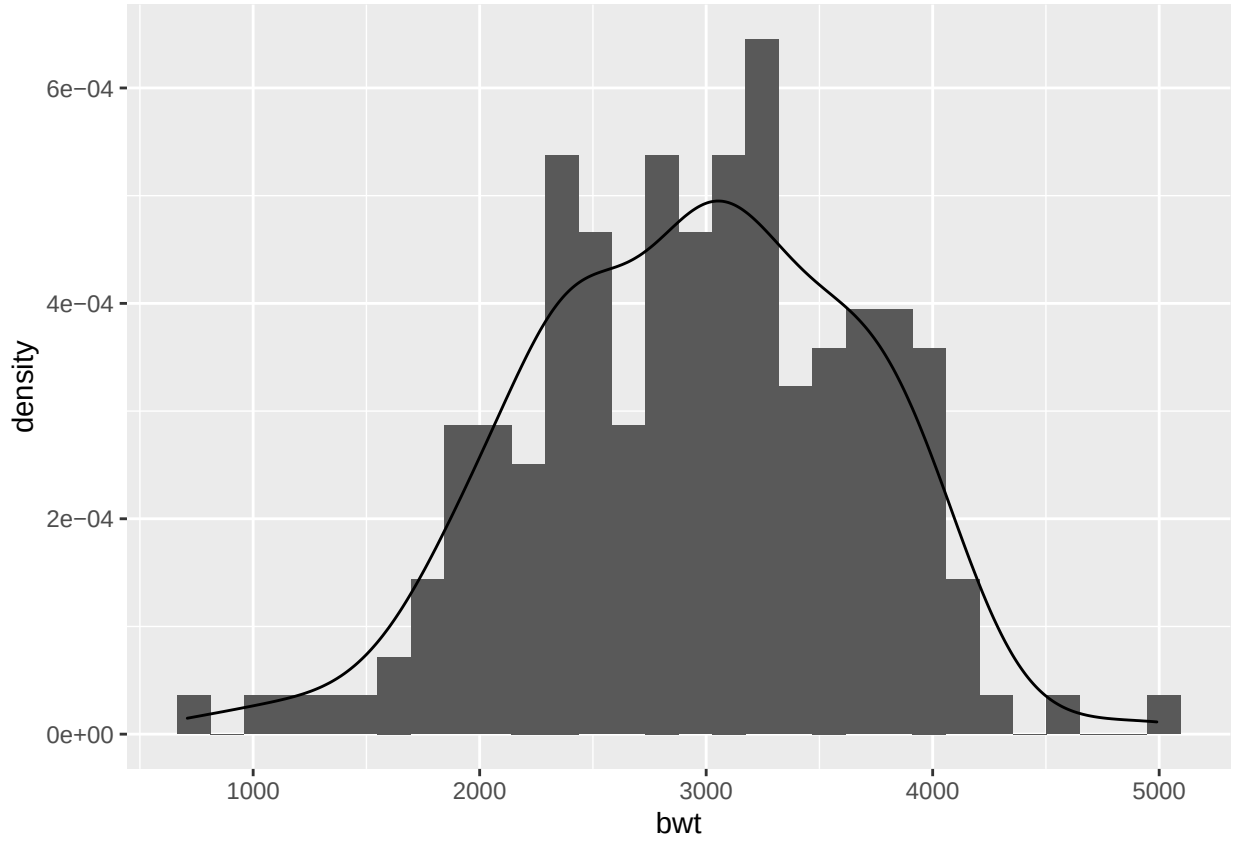
f. Remove race from the facet grid, (so go back to the graph you had in part d). I'd like to next add an estimated density line to the graphs, but to do that, I need to first change the y-axis to be density (instead of counts), which we do by using aes(y=..density..) in the ggplot() aesthetics command.

```
birthHistogramDens = ggplot(birthwt, aes(x=bwt, y=..density..)) +
  geom_histogram() +
  facet_grid()
birthHistogramDens
```



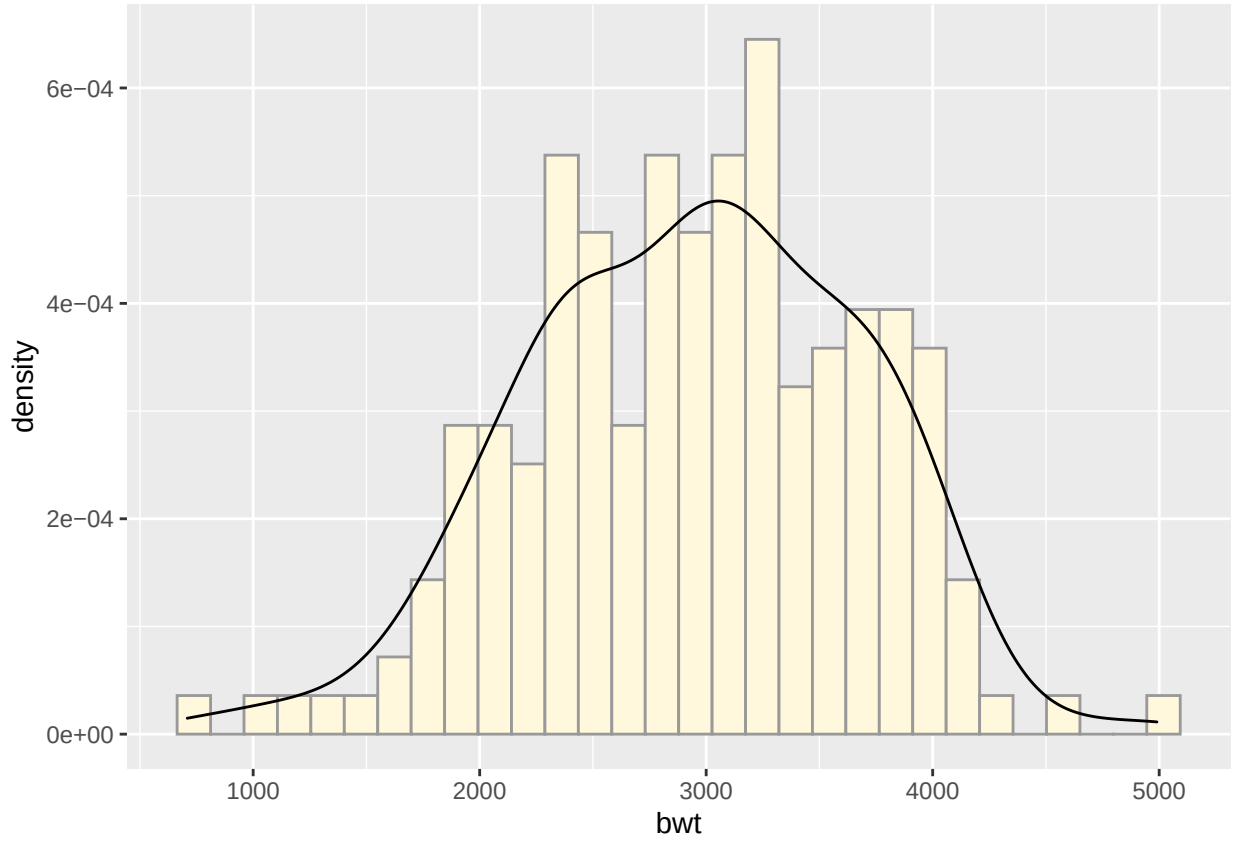
g. Next we can add the estimated smooth density using the geom_density() command.

```
birthHistogramDens = birthHistogramDens +
  geom_density()
birthHistogramDens
```



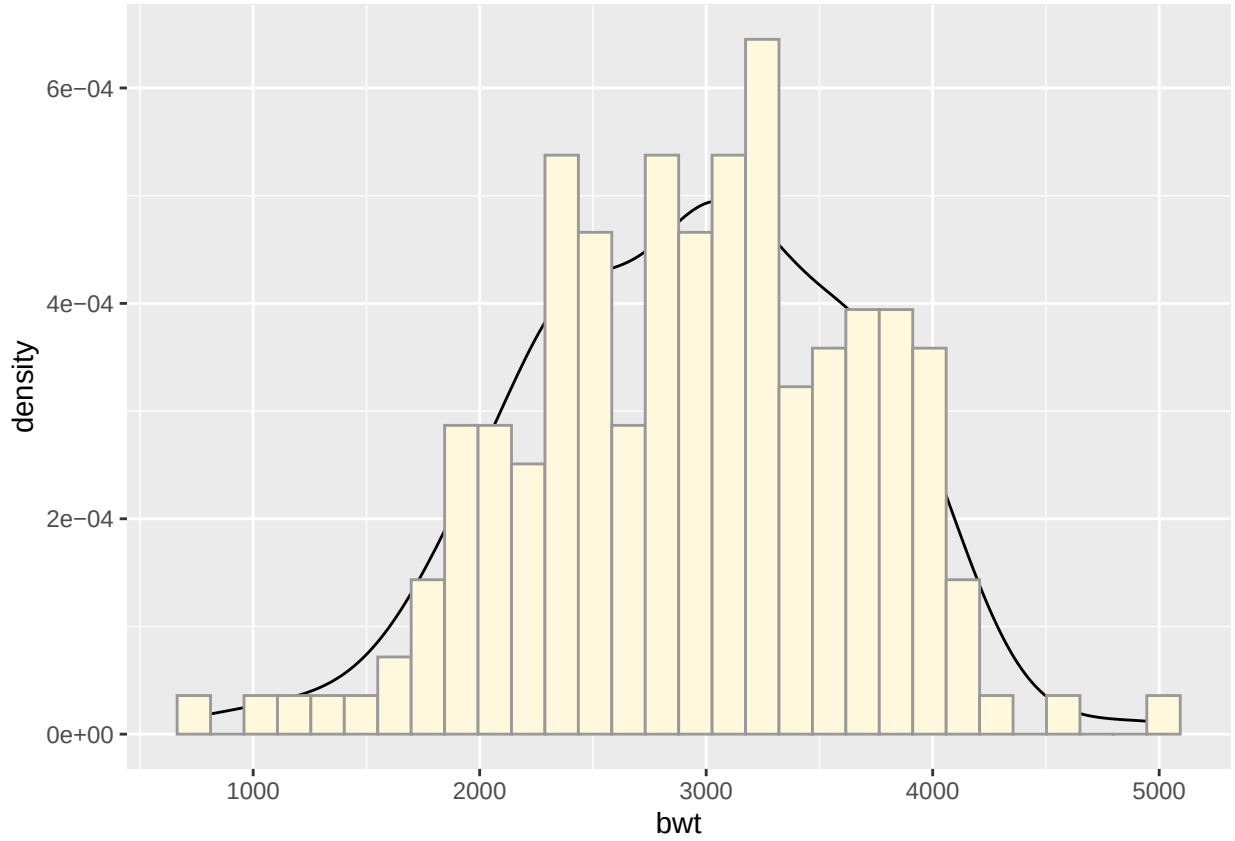
h. To really make this look nice, lets change the fill color of the histograms to be something less dark, lets use fill='cornsilk' and color='grey60'.

```
ggplot(birthwt, aes(x=bwt, y=..density..)) +
  geom_histogram(fill = "cornsilk", color = "grey60") +
  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?

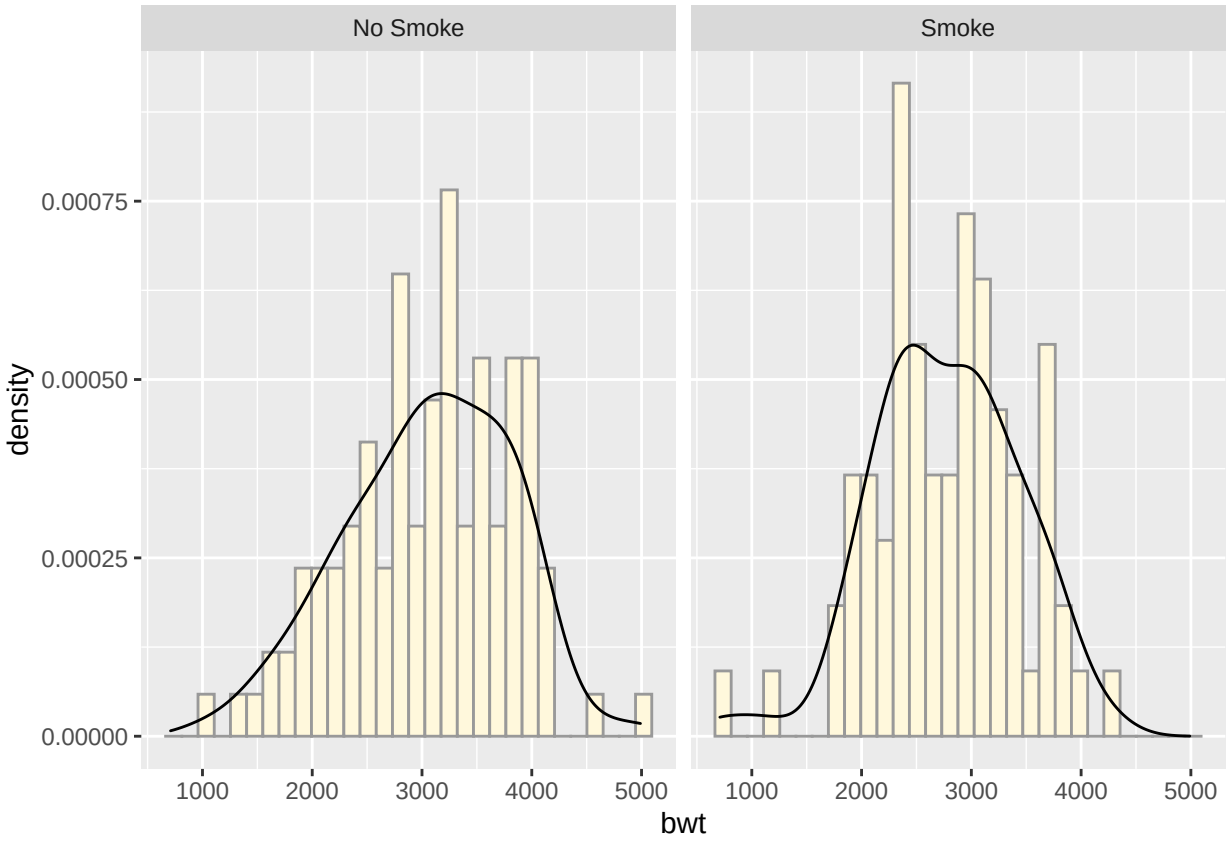
```
ggplot(birthwt, aes(x=bwt, y=..density..)) +
  geom_density() +
  geom_histogram(fill = "cornsilk", color = "grey60")
```



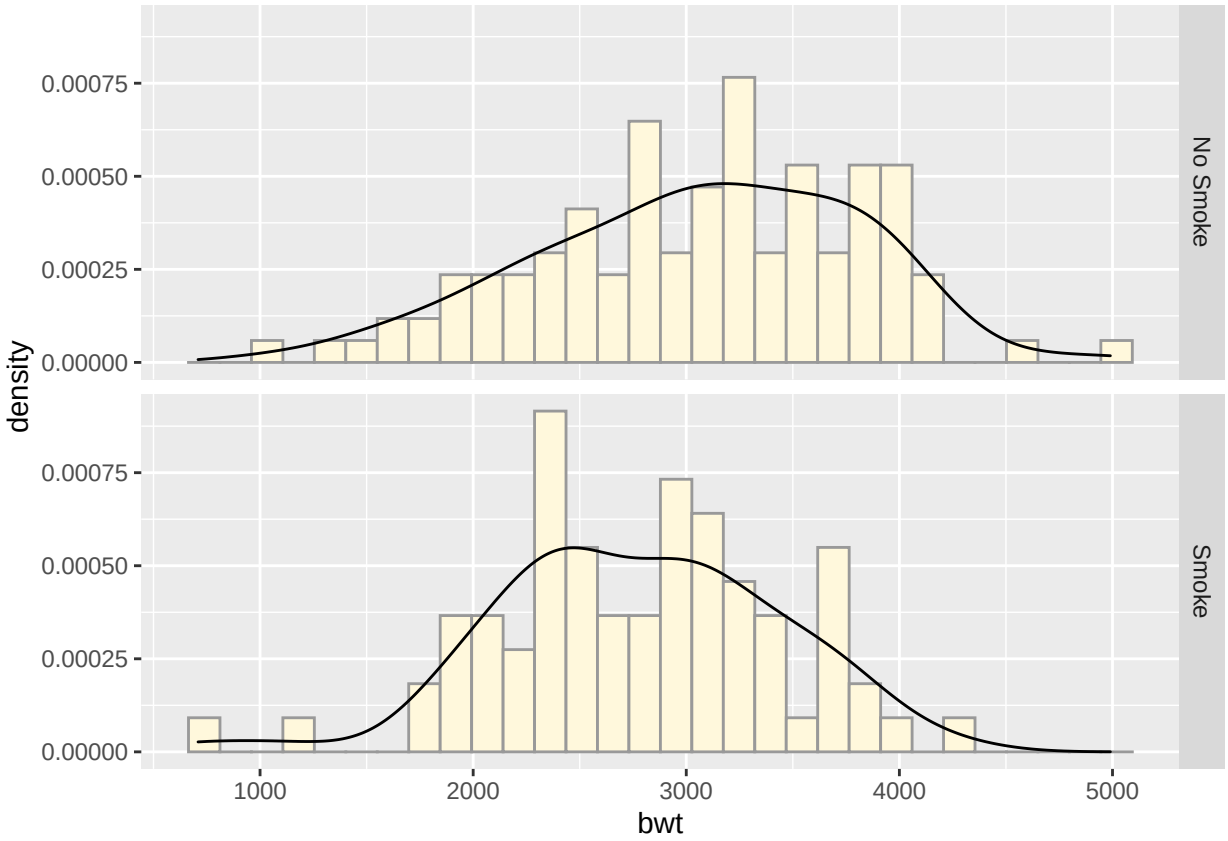
Personally I prefer the line to display over the graph, so I think it is more informative with the geom_density() call being placed after the histogram call.

j. Finally consider if you should have the histograms side-by-side or one ontop of the other (i.e. . ~ smoke or smoke ~ .). Which do you think better displayes the decrease in mean birthweight and why?

```
ggplot(birthwt, aes(x=bwt, y=..density..)) +
  geom_histogram(fill = "cornsilk", color = "grey60")+
  geom_density() +
  facet_grid(. ~ smoke)
```




```
ggplot(birthwt, aes(x=bwt, y=..density..)) +
  geom_histogram(fill = "cornsilk", color = "grey60")+
  geom_density() +
  facet_grid(smoke ~ .)
```



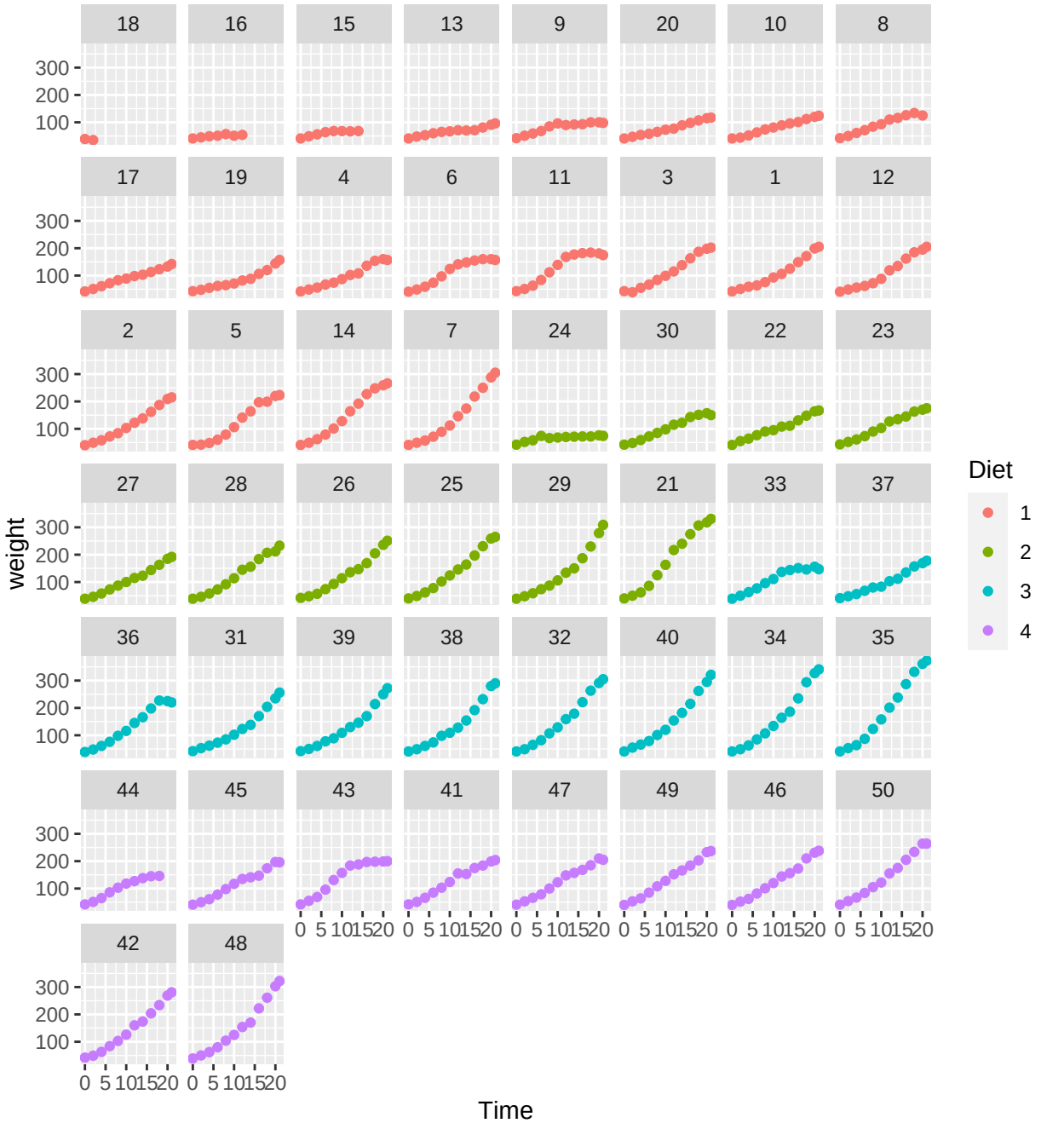
The version with each graph on top of each other, (. ~ smoke), better illustrates the density line. It is easily seen that No Smoke produces larger babies.

Exercise 4

Load the dataset ChickWeight which comes preloaded in R and get the background on the dataset 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.

Produce a separate scatter plot of weight vs age for each chick. Use color to distinguish the four different Diet treatments.

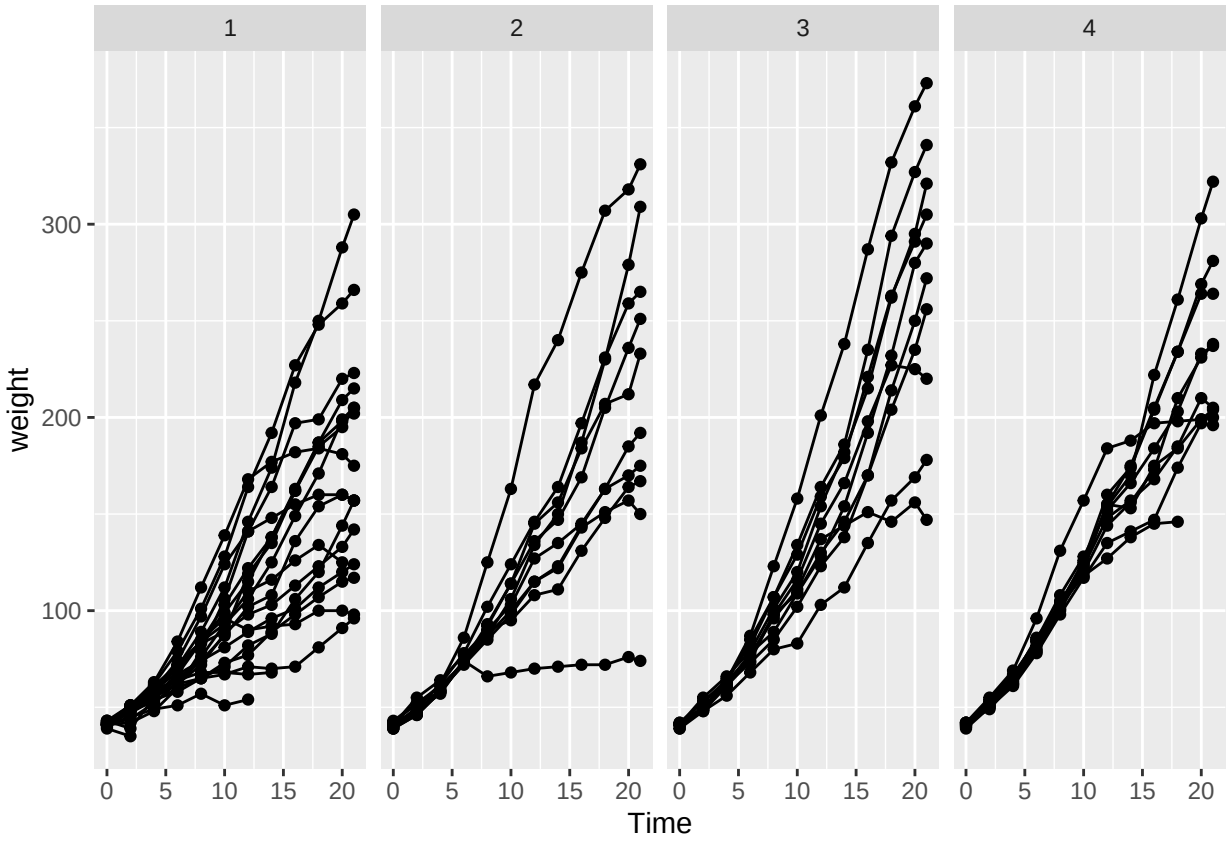
```
chickPlot = ggplot(ChickWeight, aes(x = Time, y = weight)) +
  geom_point(aes(color = Diet)) +
  facet_wrap( ~ Chick)
chickPlot
```



b. That graph above is large and annoying. Lets cut this down to split on diet.

Group chicks together using group = Chick within the aes call.

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



They even look like chicken feet! It looks cooler with color too. (add aes(color = Diet) to the geom_point call.)