Visualizing Data

Author: Melina Padron

Question 1:

```
# Save dataset as a dataframe
ChickWeight <- as.data.frame(ChickWeight)

# Visuaize the first ten rows of the dataset
head(ChickWeight, 10)
```

The dataset `ChickWeight` contains information about the weights (in grams) of chicks on four different diets over time (measured at 2-day intervals) as the result of an experiment. The first few observations are listed below.

```
##    weight Time Chick Diet
## 1      42    0     1    1
## 2      51    2     1    1
## 3      59    4     1    1
## 4      64    6     1    1
## 5      76    8     1    1
## 6      93   10     1    1
## 7     106   12     1    1
## 8     125   14     1    1
## 9     149   16     1    1
## 10    171   18     1    1
```

Use some combination of `table()` and `length()` to answer the following questions:

- How many distinct chicks are there?
- How many distinct time points?
- How many distinct diet conditions?
- How many chicks per diet condition?

```
# used the length() and table() functions to view the amount of distinct chicks
length(table(ChickWeight$Chick))
```

```
## [1] 50
```

```r
# used the length() and table() functions to view the amount of time points
length(table(ChickWeight$Time))
```

```
## [1] 12
```

```r
# used the length() and table() functions to view the amount of diet conditions
length(table(ChickWeight$Diet))
```

```
## [1] 4
```

```r
# used length() and table() to view how many chicks are in each diet condition
table(ChickWeight[ChickWeight$Time == 0,]$Diet)
```

```
##
##  1  2  3  4
## 20 10 10 10
```

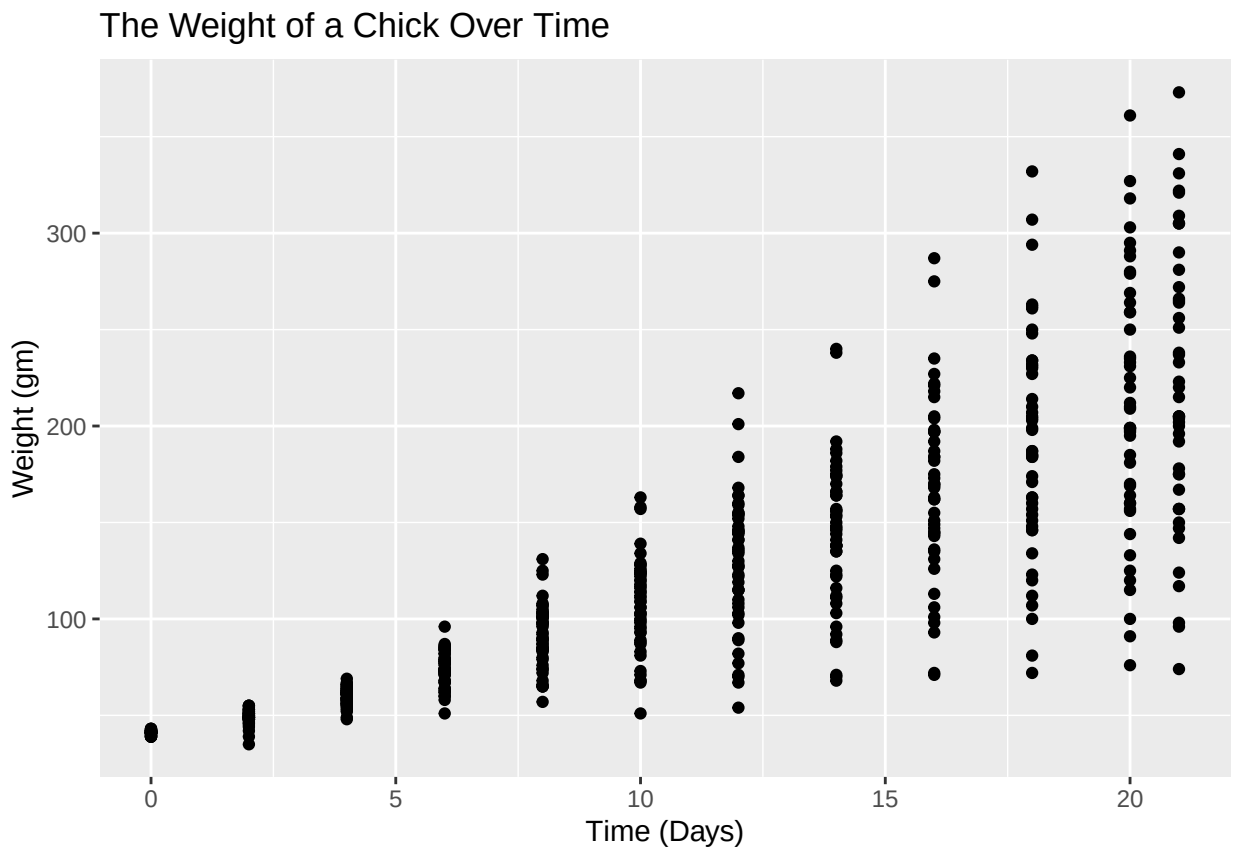
Answer: I used the length() and table() functions to find that there are 50 distinct chickens, 12 distinct time points, and 4 distinct diet conditions. Moreover, I used the table() and length() to discover that there are 20 chicks following diet condition 1, 10 chicks following diet condition 2, 10 chicks following diet condition 3, and 10 chicks following diet condition 4.

Question 2:

```r
# Load the tidyverse (including ggplot2)
library(tidyverse)

# created a scatterplot showing chick `weight` as a function of `Time`
# added necessary labels for the graph
ChickWeight %>%
  ggplot(aes(x = Time, y = weight)) +
  geom_point() +
  labs(title = 'The Weight of a Chick Over Time',
       y = 'Weight (gm)',
       x = 'Time (Days)')
```

Using the `ggplot2` package, create a simple scatterplot showing chick `weight` (on the y-axis) as a function of `Time`. Label the axes including the units of the variables and give the plot a title. How does chick `weight` change over `Time`?

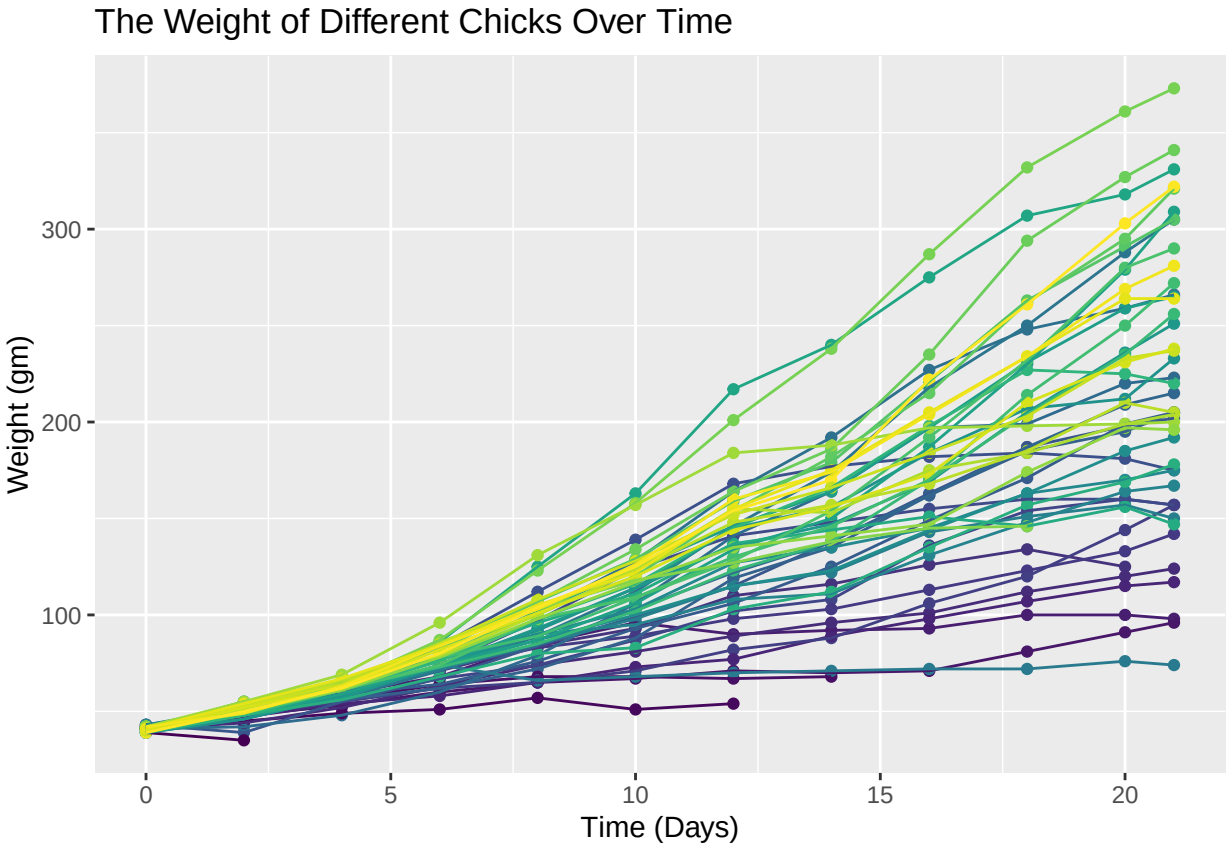


Answer: I made the above scatterplot by using the ggplot() and geom_point() functions to show chick 'weight' on the y-axis as a function of 'Time.'From the graph, we can conclude that chick 'weight' increases over time.

Question 3:

```
# used the above plot code that created a scatterplot of chick weight over time
# used color() to assign a color to each chick's data points
# used geom_line() to add lines that connect each chick's points together
# removed legend
ChickWeight %>%
  ggplot(aes(x = Time, y = weight, color = Chick)) +
  geom_point() +
  geom_line()+
  theme(legend.position = "none")+
  labs(title = 'The Weight of Different Chicks Over Time',
       y = 'Weight (gm)',
       x = 'Time (Days)')
```

Building upon the previous plot, map `Chick` to an aesthetic that assigns a color to each chick's data points. Add lines that connect each chick's points together with `geom_line()`. Finally, remove the legend. Do all chicks seem to gain weight in the same manner? Why/Why not?



Answer: I created the same scatter plot as the above question using the ggplot() and geom_point() functions, however, I made a couple of changes. First, I mapped 'Chick' to an aesthetic that assigns a color to each chick's data points using the color() function inside the ggplot() function. I also added lines that connect each chick's points together with the geom_line() function. Finally, I removed the legend using theme() and legend_position(). From the graph, we can conclude that not all chicks gain weight in the same manner. We can see this by looking at the bottom blue line and the top light green line. The chick represented by the blue line grew at a much slower pace compared to the chick represented by the green line, as seen by the difference in the steepness the of slopes. Thus, the chicks gain weight in different manners.

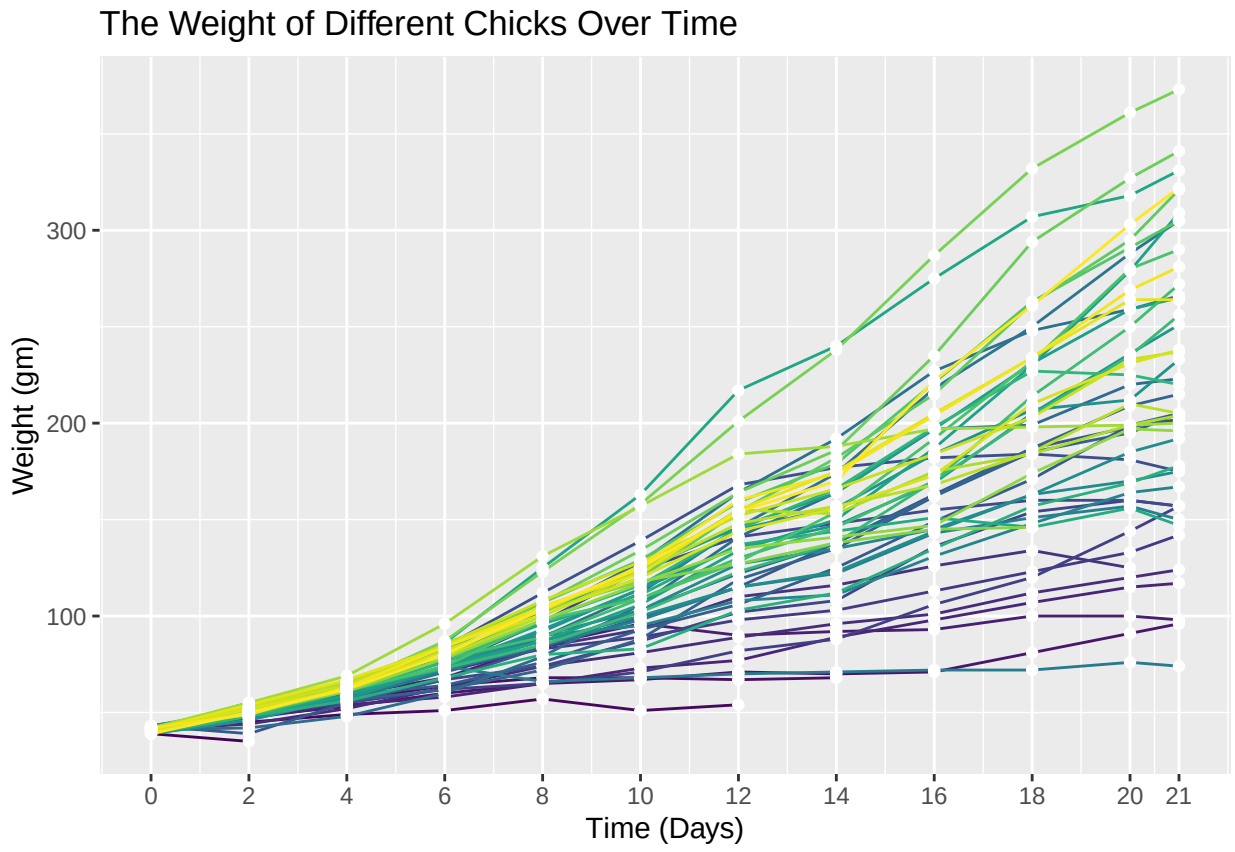
Question 4:

Continue modifying the same plot by

- removing the color from the points only
- make all of the points white
- leave the lines colored by chick
- Put the points *on top of* the lines

On which day was the last value of the chicks' weight recorded?

```r
# used the above plot code that created a scatterplot of chick weight over time
# changed the points to white with color() inside geom_point()
# put the geom_line() function first to put points on top
# used scale_x_continous to the last day that the chicks' weight was recorded
ChickWeight %>%
  ggplot(aes(x = Time, y = weight, color = Chick)) +
  geom_line()+
  geom_point(color = 'white') +
  theme(legend.position = "none")+
  labs(title = 'The Weight of Different Chicks Over Time',
       y = 'Weight (gm)',
       x = 'Time (Days)')+
  scale_x_continuous(breaks = unique(ChickWeight$Time))
```



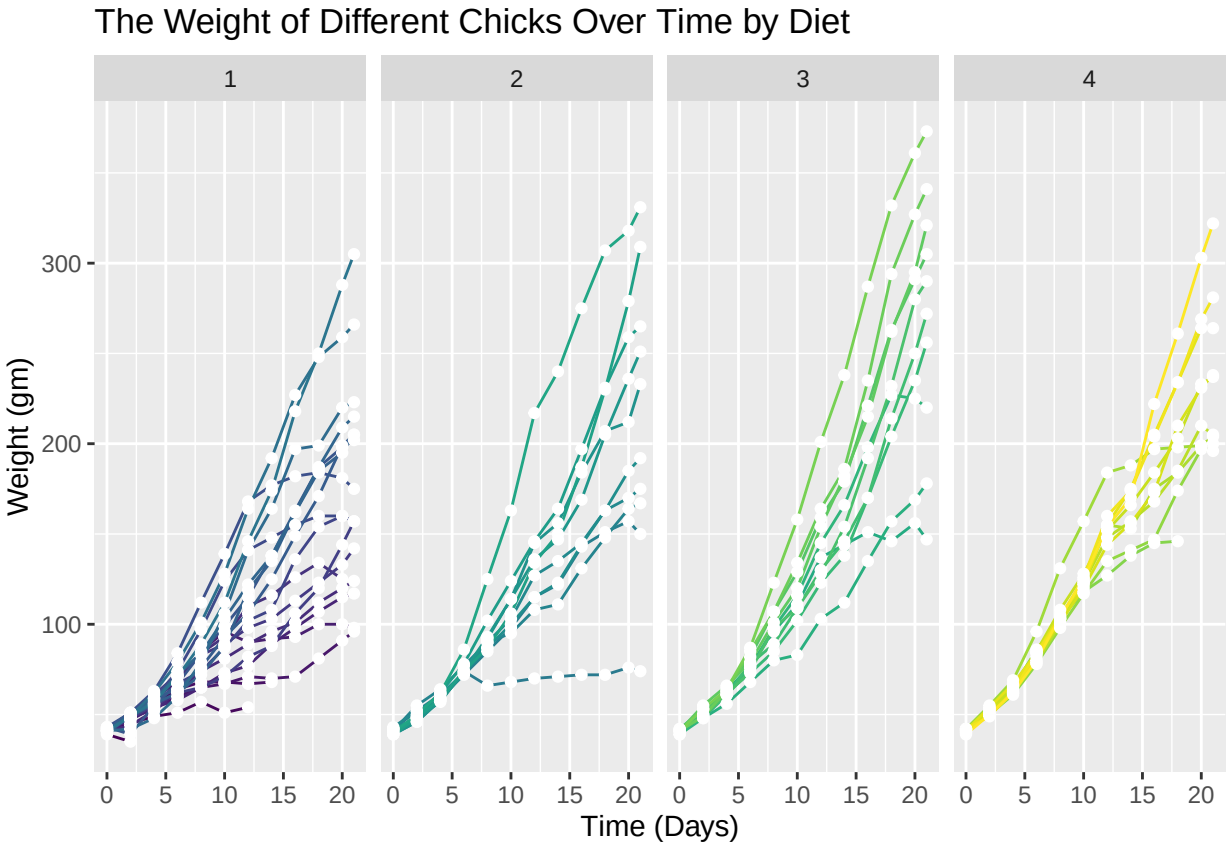
Answer: I used the scatterplot from the previous problem and then made all of the points white using color() inside geom_point(). Then, I put the points on top of the lines by swapping the position of geom_point() and geom_line(), putting geom_line() first. Moreover, I left the lines colored by the 'Chick' variable by keeping the color() function inside ggplot. I also used scale_x_continous() and breaks() to see when the last day the chicks' weight were recorded. In this case, the 21st day was the last day that the weight of the chicks was recorded.

Question 5:

Now, facet this plot by diet. Can you tell from this new plot which diet results in greater weight? Explain.

```
# used the above scatterplot that was made from ggplot and geom_point
# used facet_grid() to create this same scatterplot for each diet condition

ChickWeight %>%
  ggplot(aes(x = Time, y = weight, color = Chick)) +
  geom_line()+
  geom_point(color = 'white')+
  facet_grid(col= vars(Diet))+
  theme(legend.position = "none")+
  labs(title = 'The Weight of Different Chicks Over Time by Diet',
       y = 'Weight (gm)',
       x = 'Time (Days)')
```

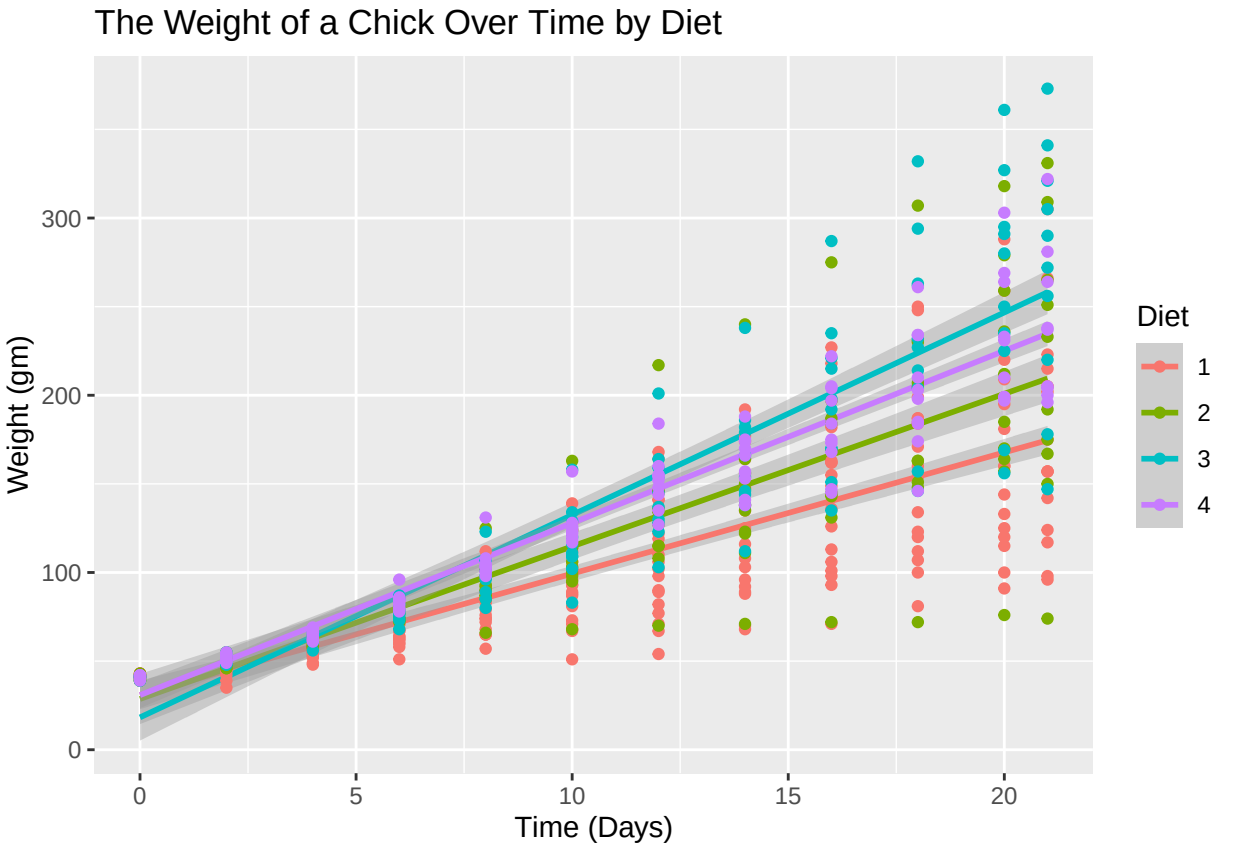


Answer: I used the scatterplot from above that was made using ggplot() and geom_point() and then used facet_grid() to make this same scatter plot for each diet condition. From this new graph created by facet_grid(), we can see which diet results in greater weight, but it is not extremely clear. Since each graph is placed in separate columns, we can see that diet condition 3 had a few chicks that reach a greater weight gain compared to the diet plans. However, the overall pattern can be displayed in a more noticeable way.

Question 6:

Go back to your plot from question 2 and fit a *linear regression line* (using `lm`) to the chicks in each diet with `geom_smooth()`. There should be 4 separate regression lines, one for each diet, each a separate color. Can you see more clearly which diet results in greater weight? Explain.

```
# created a scatterplot showing chick `weight` (on the y-axis) as a function of `Time`
# used color() for the Diet variable inside ggplot to assign color based on Diet
# used geom_smooth to create regression lines for each diet
ChickWeight %>%
  ggplot(aes(x = Time, y = weight, color = Diet)) +
  geom_smooth(method = lm)+
  geom_point() +
  labs(title = 'The Weight of a Chick Over Time by Diet',
       y = 'Weight (gm)',
       x = 'Time (Days)')
```



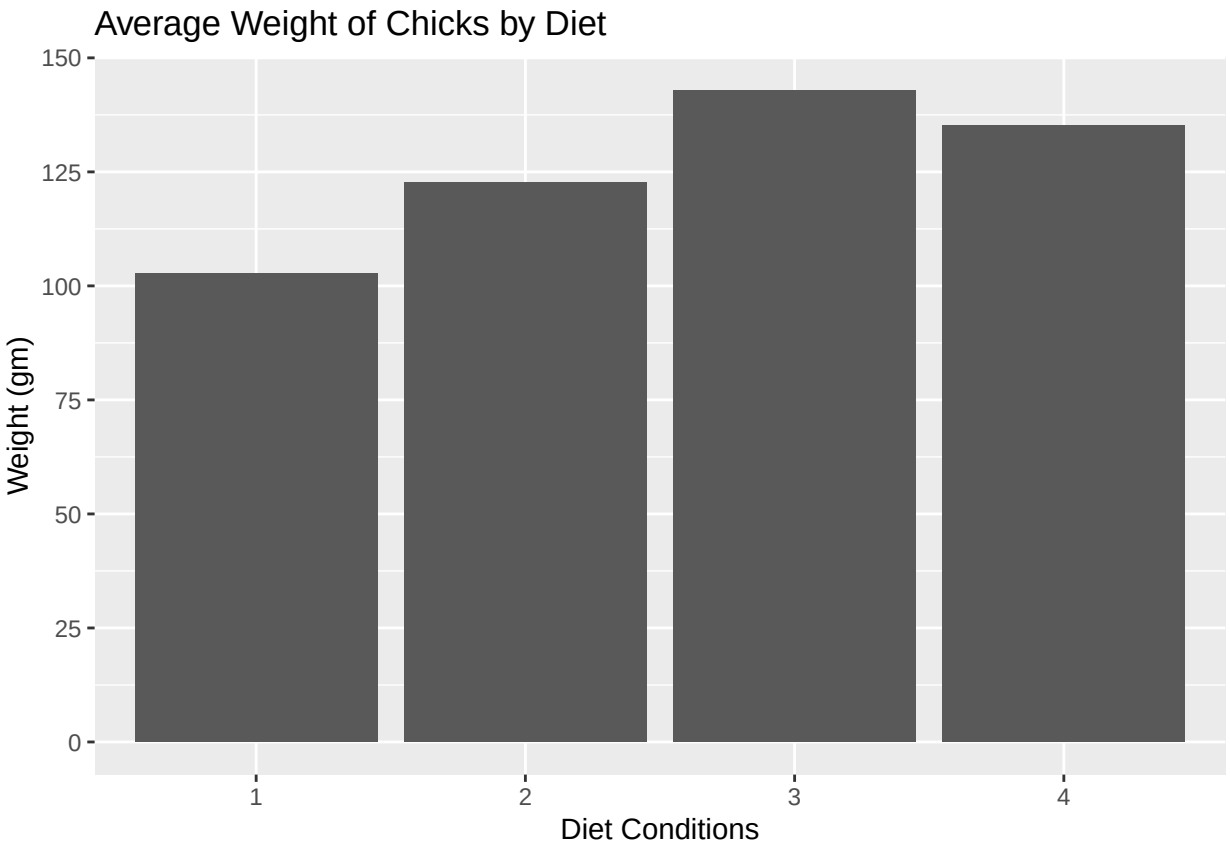
I used the scatterplot that I created from question 2 and assigned color based on diet using color() inside ggplot(). I then used geom_smooth() and 'lm' to create four different linear regression lines for each diet condition. With this new graph, I can more clearly see that diet condition 3 results in greater weight because the regression lines show the average weight of the chicks in each diet plan over time. Diet condition 3 is followed by 4, 2, and finally 1 where diet condition 1 ranks the lowest in weight gain. This makes the overall trend easier to see compared to the graph in question 2 that doesn't show which diet plan is being followed.

Question 7:

A scatterplot might not be the best way to visualize this data: it calls attention to the relationship between weight and time, but it can be hard to see the differences between diets. A more traditional approach for exploring the effect of diet would be to construct a barplot representing group means with standard error bars showing +/- 1 standard error.

Create a plot using `geom_bar()` where each bar's height corresponds to the average chick weight for each of the four diet conditions. Rename the y-axis to include units (e.g., with scale_y_continuous(`name=...`)) and make the major tick marks go from 0 to 150 by 25 (e.g., with scale_y_continuous(`breaks=...`)). Which diet has the highest mean `weight`?

```
# grouped the data by Diet and calculated the mean weight for each diet condition
# created bar plot with weight on the y axis
# renamed y axis using scale_y_continuous and adjusted its major breaks
ChickWeight %>%
  group_by(Diet) %>%
  summarize(mean_weight = mean(weight)) %>%
  ggplot(aes(x = Diet, y = mean_weight)) +
  geom_bar(stat = "identity") +
  labs(title = 'Average Weight of Chicks by Diet',
       x = 'Diet Conditions') +
  scale_y_continuous(name = "Weight (gm)", breaks = seq(0, 150, 25))
```



Answer: In order to create a bar plot of the average weight of chicks by diet I first had to separate the data by 'Diet.' I then calculated the mean weight of each diet using summarize()

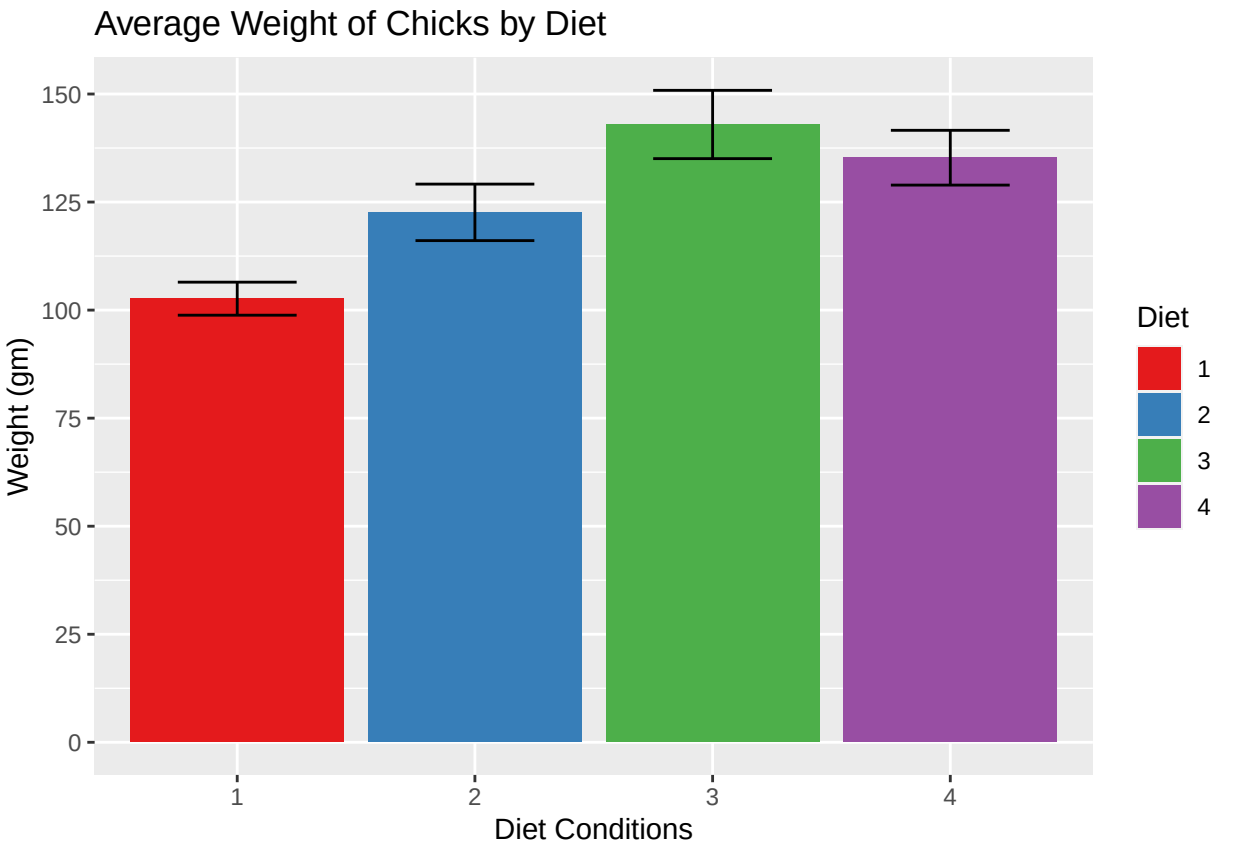
and mean(). Afterwards, I created the bar plot using ggplot() and geom_bar(). I then adjusted the y axis to have the necessary axis title and major breaks using scale_y_continuous(), name(), and breaks(). From this graph, it is clear to see that diet condition '3' had the highest mean weight.

Question 8:

Add error bars showing + or - 1 SE using `geom_errorbar(stat = "summary")`. Make the error-bars skinnier by adding a `width =0.5` argument. Color the bars (not the error bars, but the barplot bars) by diet and change from the default color scheme using a `scale_fill_` or a `scale_color_`. diet seems to have the most variation in `weight`? The least variation?

```
# used the above barplot that was created with ggplot() and geom_bar()
# calculated SE using summarize()
# added error bars showing + or - 1 SE using `geom_errorbar(stat = "summary")`
# colored the bars by Diet using fill() inside ggplot()

ChickWeight %>%
  group_by(Diet) %>%
  summarize(mean_weight = mean(weight), se = sd(weight)/sqrt(n())) %>%
  ggplot(aes(x = Diet, y = mean_weight, fill = Diet)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin = mean_weight - se,
                    ymax = mean_weight + se), width = 0.5) +
  labs(title = 'Average Weight of Chicks by Diet',
       x = 'Diet Conditions') +
  scale_y_continuous(name = "Weight (gm)",
                     breaks = seq(0, 150, 25)) +
  scale_fill_brewer(palette = "Set1")
```



Answer: I used the same bar plot as the previous question but made some modifications. First, I calculated the standard error using group_by() and summarize(). I then created the bar plot using ggplot(),geom_bar(), and used fill() to make sure the bars were colored by diet plan. I then added the error bars to each bar on the plot using geom_errorbar() and changed the width to 0.5. The rest was the same as the previous plot. From this new graph, we can now conclude that diet condition '3' has the most variation in 'weight,'while diet condition '1' has the least variation in weight.

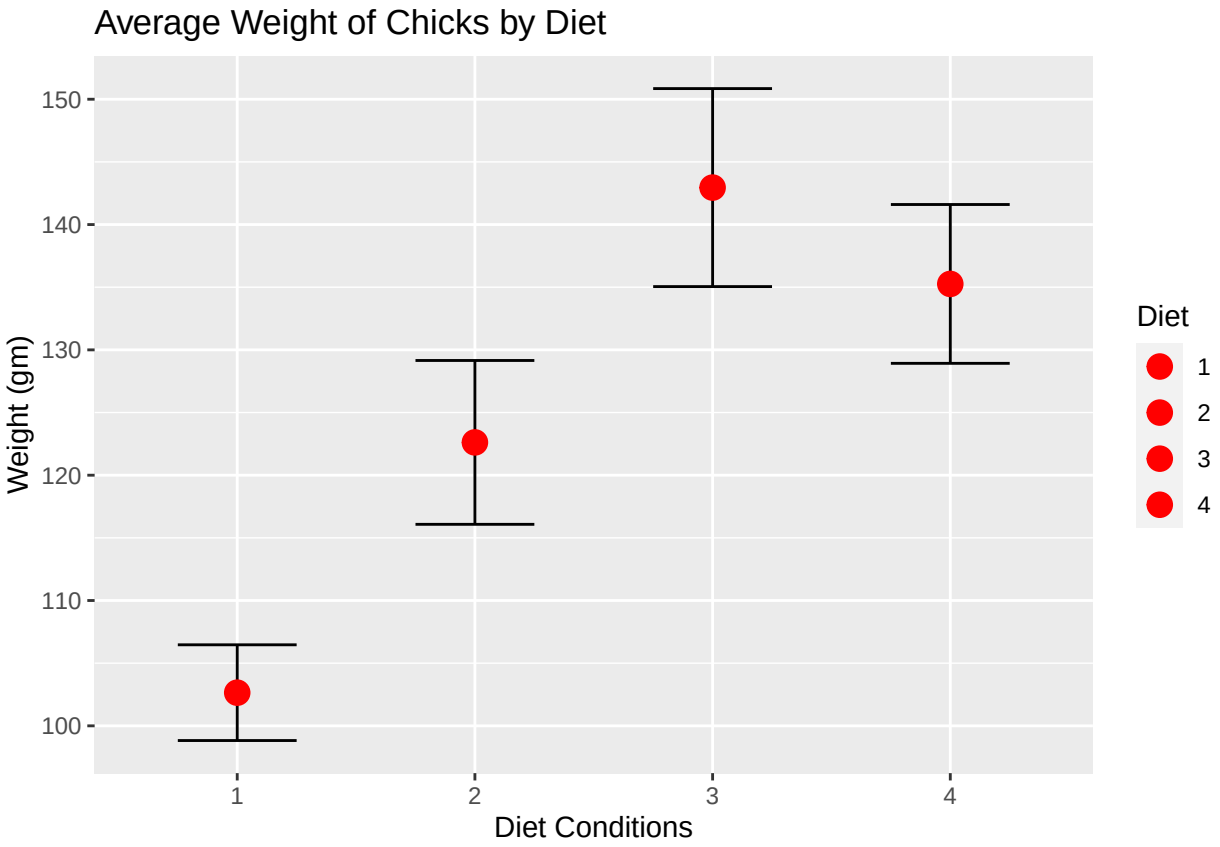
Question 9:

Take your code from question 8 and replace `geom_bar()` with `geom_point()`. Remove the `breaks=` argument from `scale_y_continuous`. Make the points larger and color them all red. Put them *on top of* the error bars. Does the mean chick weight seem to differ based on the diet? *I am not asking to conduct hypothesis testing but informally state if they seem to differ and if so, how.*

```
# used the above barplot that was created with ggplot() and geom_bar()
# replaced geom_bar with geom_point
# removed the breaks argument
# made points red, larger, and put them above the error bars

ChickWeight %>%
  group_by(Diet) %>%
```

```
  summarize(mean_weight = mean(weight), se = sd(weight)/sqrt(n())) %>%
  ggplot(aes(x = Diet, y = mean_weight, fill = Diet)) +
  geom_errorbar(aes(ymin = mean_weight - se, ymax = mean_weight + se),
                width = 0.5)+
  geom_point(stat = "identity", color = "red", size = 4) +
  labs(title = 'Average Weight of Chicks by Diet', x = 'Diet Conditions') +
  scale_y_continuous(name = "Weight (gm)") +
  scale_fill_brewer(palette = "Set1")
```



Answer: After starting with the same code from the previous question, I then replaced geom_bar() with geom_point(). Next, I removed the breaks argument from scale_y_continuous() and made the points on the graph red. I then increased the size of the points using the size argument and placed them on top of the error bars by putting the geom_errorbar() code above the geom_point() function. Based off the graphs created in this lab, I would conclude that the mean chick weight differs based on the diet. I would assume this because from the graphs we can see that there is a significant difference in mean weight depending on which diet plan is used. For instance, diet condition 3 resulted in the highest mean weight gain followed by 4, 2, and finally 1.
