HW 2

SDS348 Spring 2021

2021-02-08

Rose Hedderman rrh2298

This homework is due on Feb 8, 2021 at 8am. Submit a pdf file on Gradescope.

For all questions, include the R commands/functions that you used to find your answer (show R chunk). Answers without supporting code will not receive credit. Write full sentences to describe your findings.

Question 1: (2 pts)

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

```
# Save dataset as a dataframe
ChickWeight <- as.data.frame(ChickWeight)
head(ChickWeight,10)
```

```
##    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:

```
#?ChickWeight
length(ChickWeight$Chick)
```

```
## [1] 578
```

```
length(ChickWeight$Time)
```

```
## [1] 578
```

```
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          
```

```
chicksInDiet <- ChickWeight[ChickWeight$Time == "0",]
table(chicksInDiet$Diet)
```

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

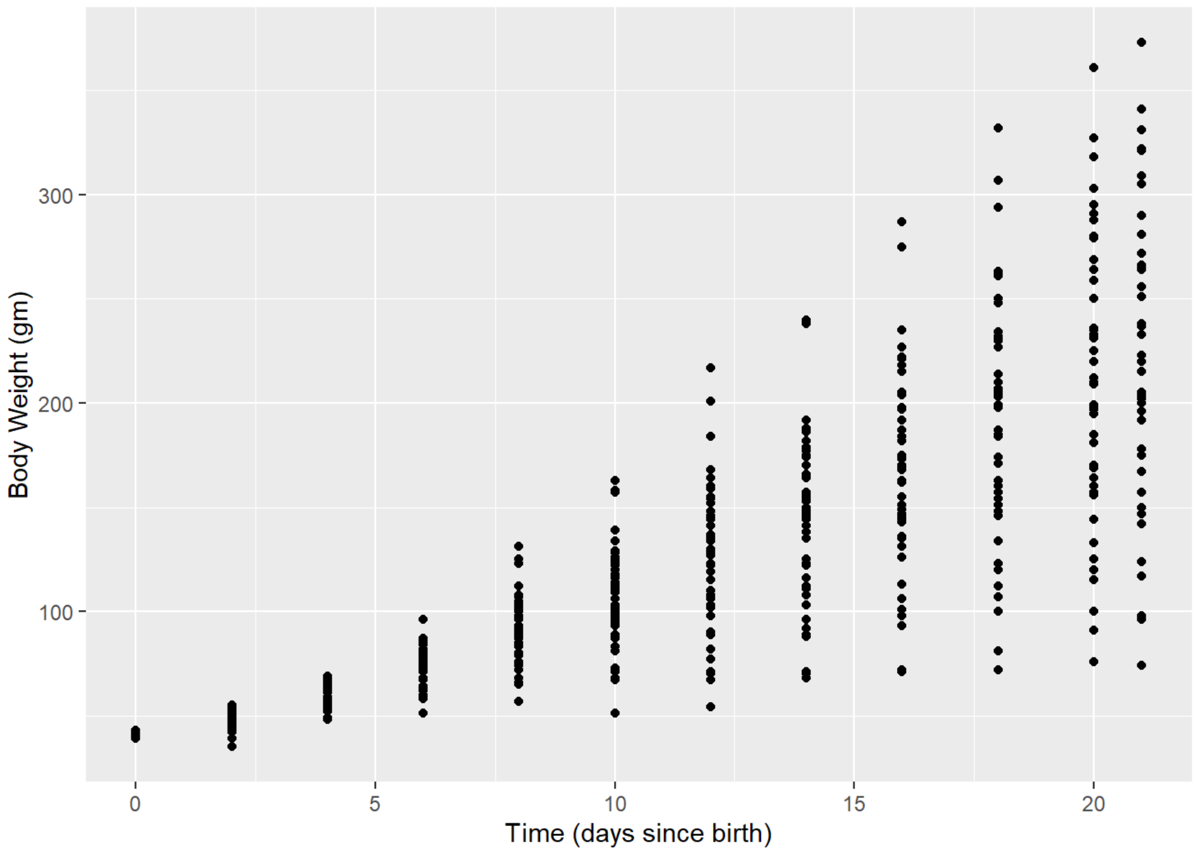
- How many distinct chicks are there?
- How many distinct time points?
- How many distinct diet conditions?
- How many chicks per diet condition?

There 578 chicks in the ChickWeight There are 578 distinct timepoints. There are 4 distinct diet conditions. There are 220 chicks on Diet 1, 120 chicks on Diet 2, 120 chicks on Diet 3, and 118 chicks on Diet 4.

Question 2: (12 pts)

2.1 (2 pts) 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.

```
library(ggplot2)
#?ggplot()
ggplot(ChickWeight, aes(Time, weight)) + geom_point() +
  xlab("Time (days since birth)") +
  ylab("Body Weight (gm)")
```

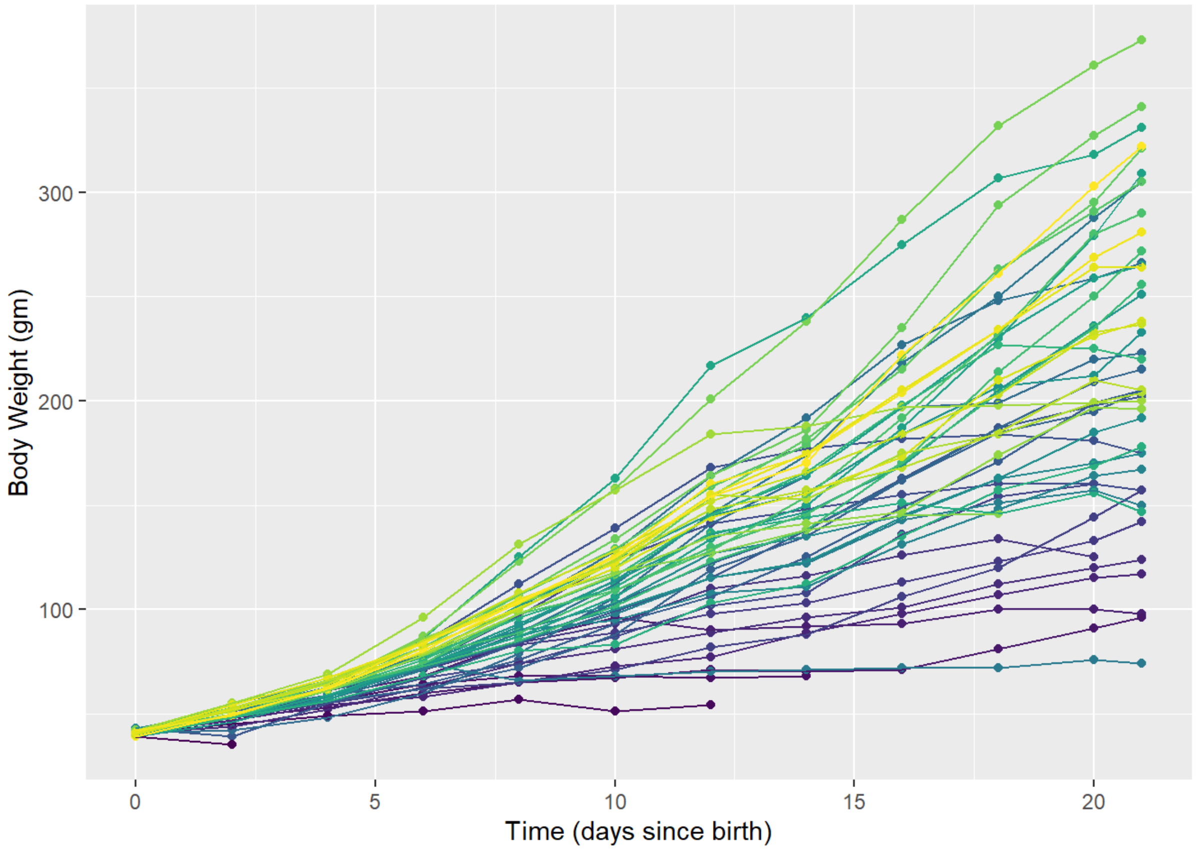


2.2 (1 pt) How does chick `weight` change over `Time`?

The ChickWeight rises linearly over the more days since birth.

2.3 (2 pts) 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. Finally, remove the legend.

```
ggplot(ChickWeight, aes(Time, weight, color = Chick)) + geom_point() +
  xlab("Time (days since birth)") +
  ylab("Body Weight (gm)") +
  geom_path()+
  theme(legend.position="none")
```

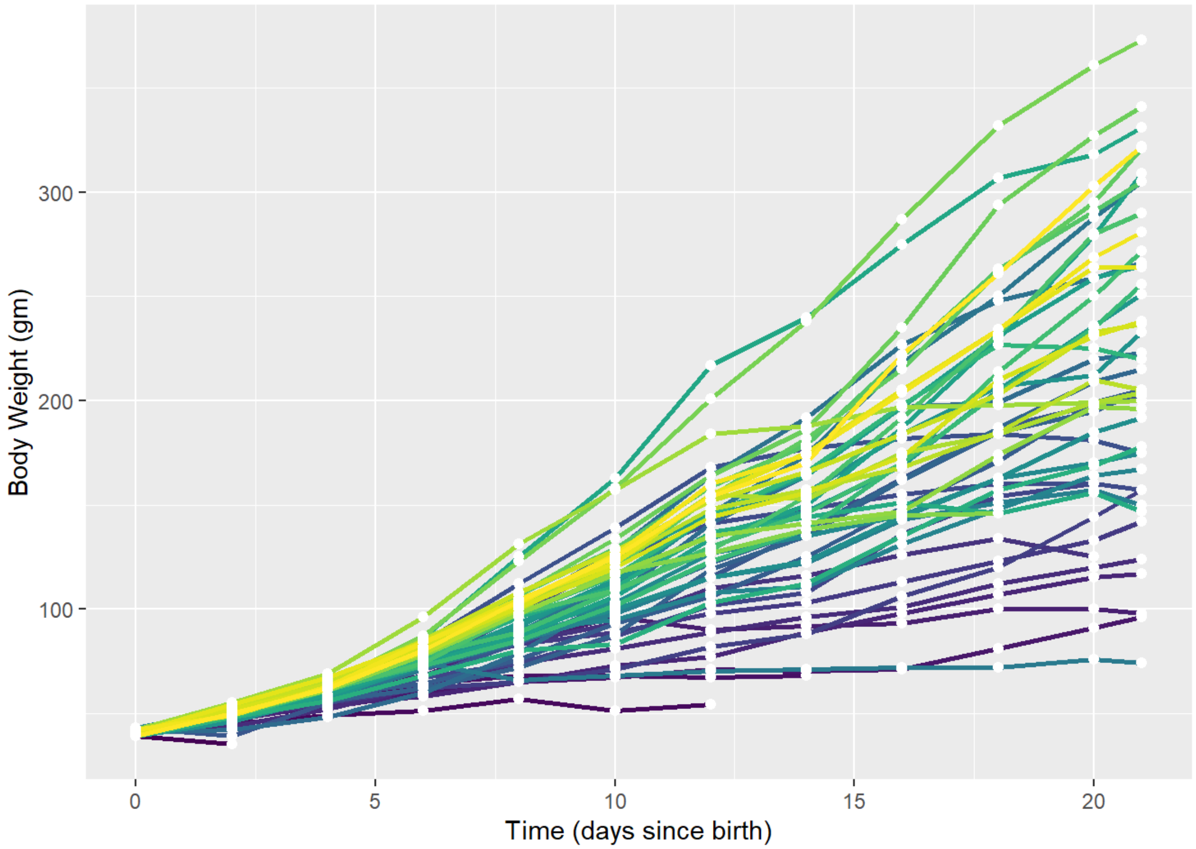


2.4 (1 pt) Do all chicks seem to gain weight in the same manner? Why/Why not?

No, some chicks barely gain weight while others gain weight consistently over the trial period. This is due to the four different diets.

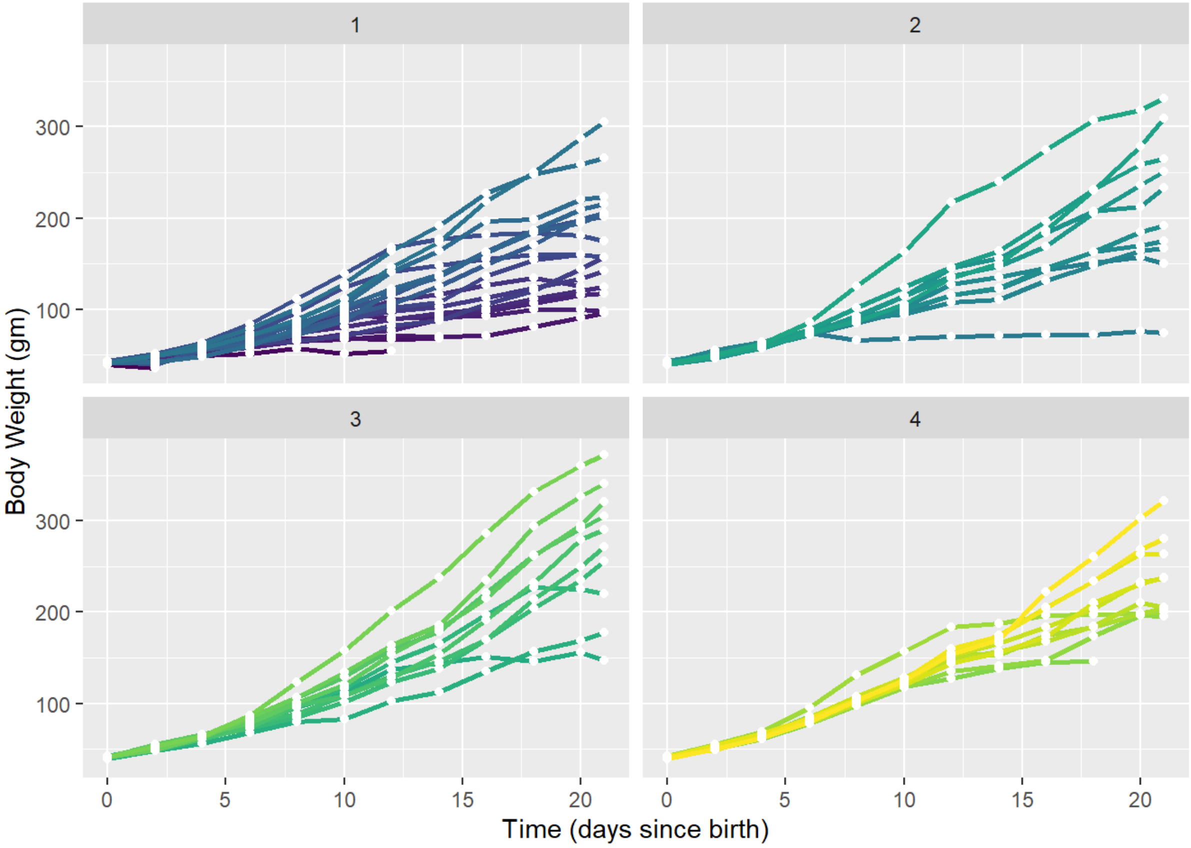
2.5 (1 pt) Take the plot you made in Question 2.3 and remove the color from the points only (leave the lines colored by chick, but make all of the points white). Put the points *on top of* the lines.

```
ggplot(ChickWeight, aes(Time, weight, color = Chick)) + geom_path(size = 1)+
  geom_point(size = 1.75,color = "white") +
  xlab("Time (days since birth)") +
  ylab("Body Weight (gm)") +
  theme(legend.position="none")
```



2.6 (2 pts) Facet by diet. Can you tell from this plot which diet is the best? Explain.

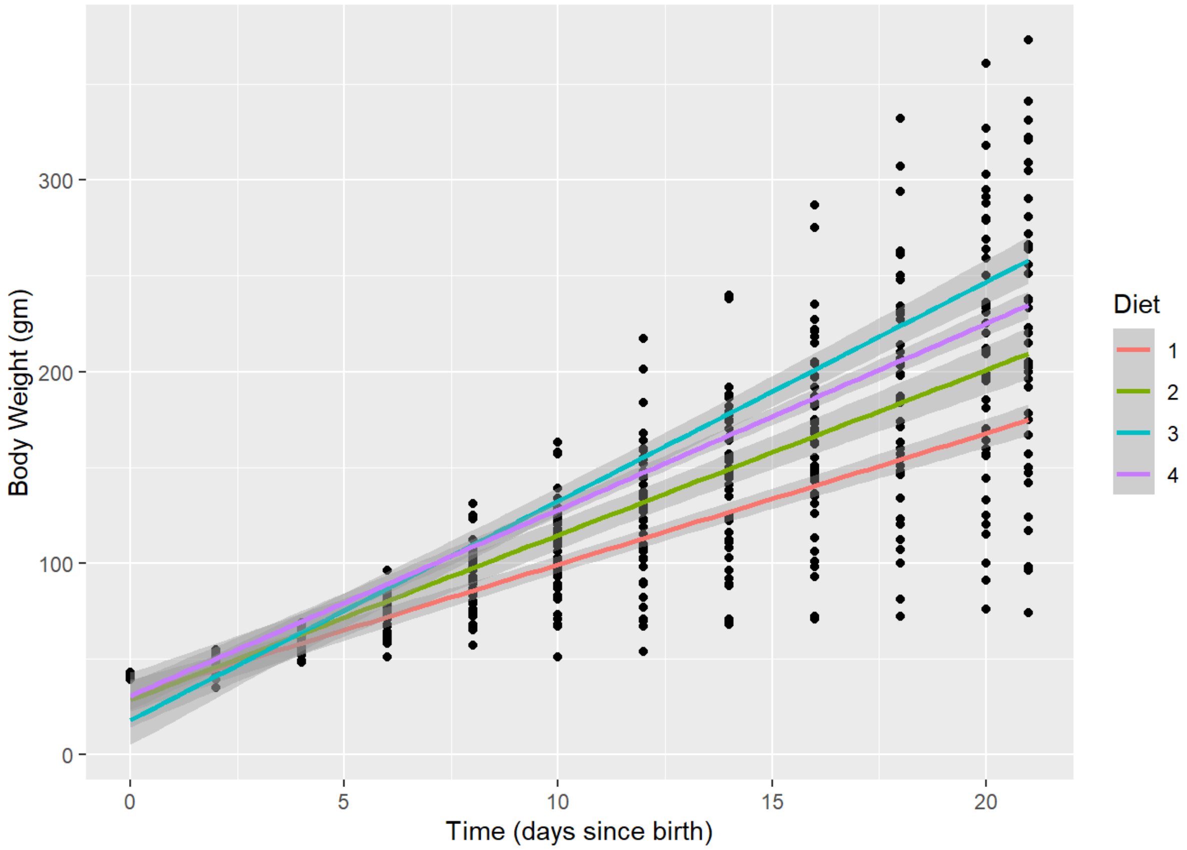
```
plot <- ggplot(ChickWeight, aes(Time, weight, color = Chick)) + geom_path(size = 1)+
  geom_point(size = 1.5,color = "white") +
  xlab("Time (days since birth)") +
  ylab("Body Weight (gm)") +
  theme(legend.position="none")
plot + facet_wrap(~Diet)
```



From these plots, Diet 3 looks the best because the chicks gain the most weight over the fewest amount of days.

2.7 (2 pts) Go back to your plot from question 2.1 and fit a *linear regression line* (not *loess*) to the chicks in each diet with `geom_smooth()` . There should be 4 separate lines, one for each diet, each a separate color.

```
ggplot(ChickWeight, aes(Time, weight)) + geom_point() +
  xlab("Time (days since birth)") +
  ylab("Body Weight (gm)") +
  geom_smooth(aes(color = Diet), method = "lm", formula = y ~ x)
```



2.8 (1 pt) Can you see more clearly which diet results in greater weight? Does the effect of diet on weight depend on time?

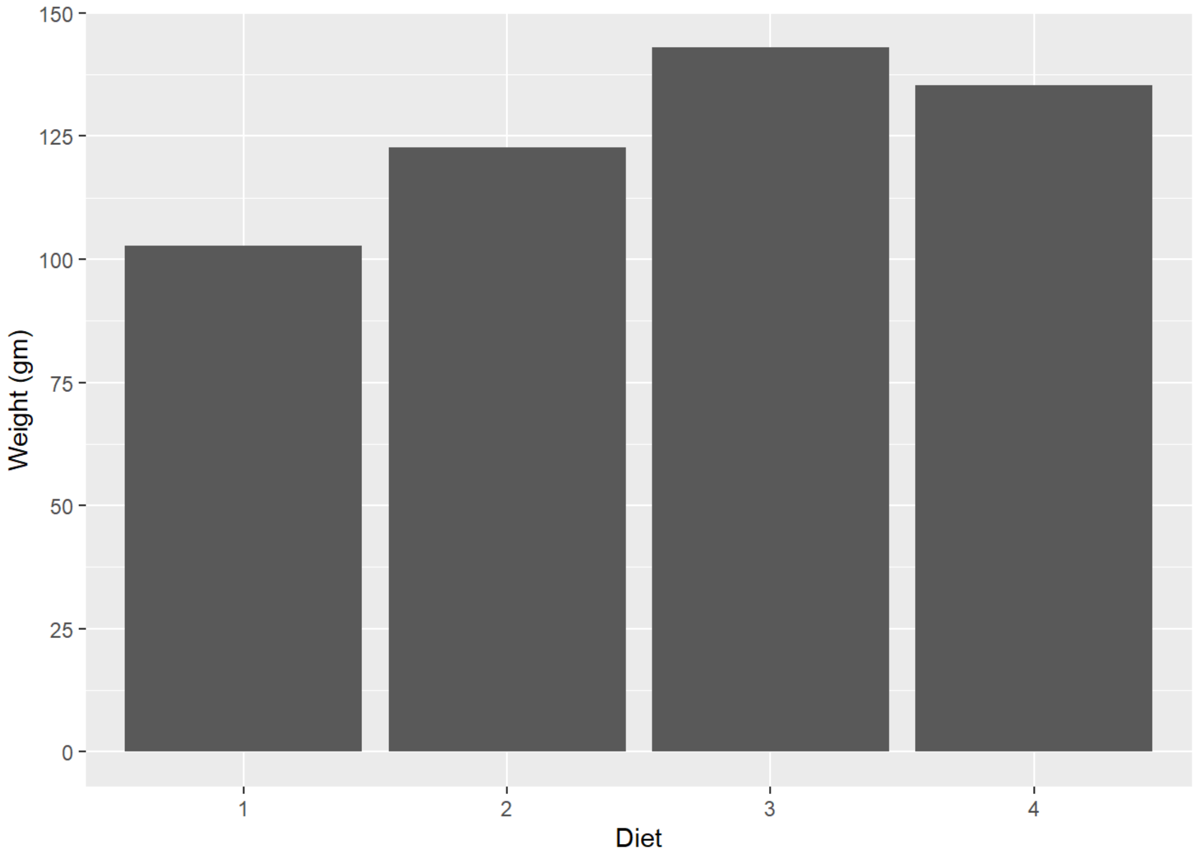
By separating the regression lines by weight, it is more clear that Diet 3 is the diet that results in the greatest weight. The effect of diet on weight does depend on time. This is sene in Diet 3 where it starts as one of the lowest weights and ends at the highest.

Question 3: (11 pts)

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 achieving this would be to construct a barplot of group means with standard error bars showing +/- 1 standard error.

3.1 (2 pts) 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=` …)).

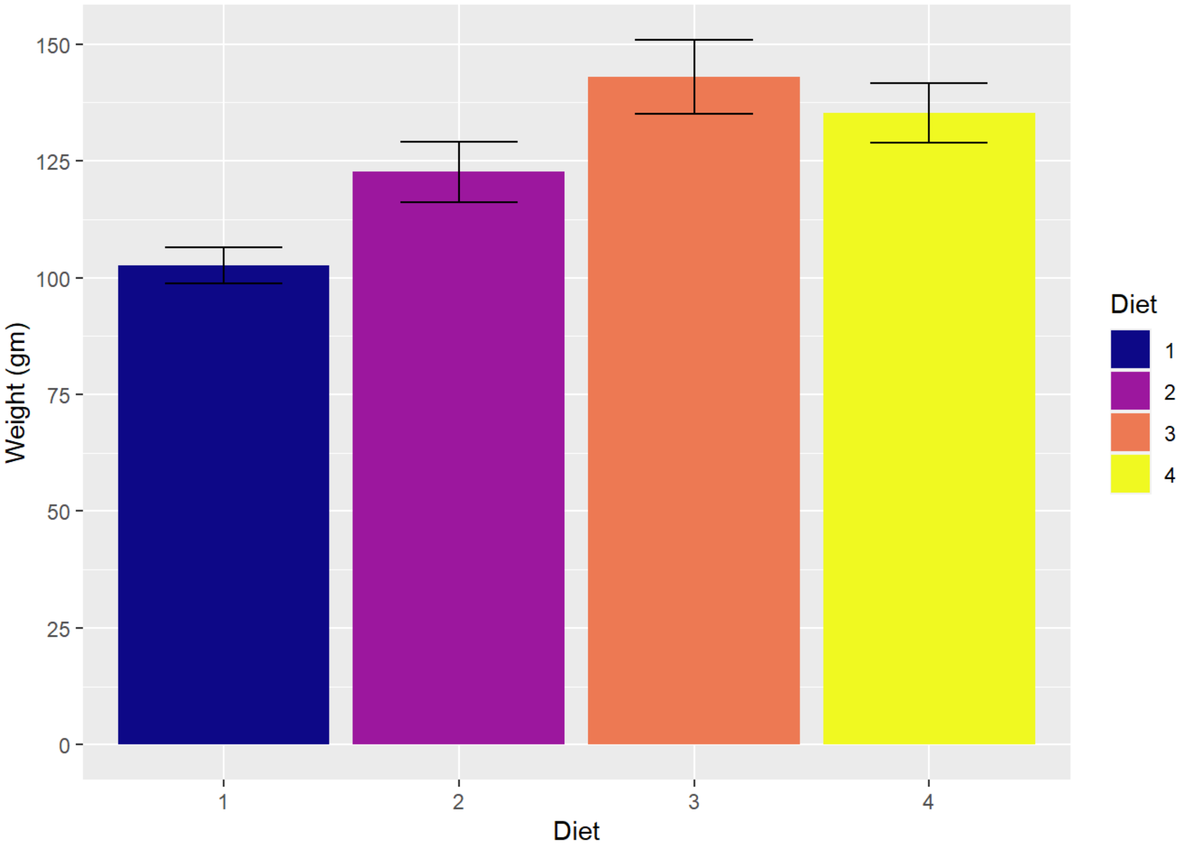
```
ggplot(ChickWeight, aes(x = Diet))+
  geom_bar(aes(y = weight), stat = "summary", fun = "mean") +
  scale_y_continuous(name = "Weight (gm)", breaks = seq(0,150,by = 25))
```



3.2 (3 pts) Add error bars showing $\pm 1\ SE$ 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_` .

```
c <- ggplot(ChickWeight, aes(x = Diet, y = weight))+
  geom_bar(aes(fill = Diet),stat = "summary", fun = "mean") +
  scale_y_continuous(name = "Weight (gm)", breaks = seq(0,150,by = 25))+
  geom_errorbar(stat="summary", width = 0.5, fun.data="mean_se")

c + scale_fill_viridis_d(option = "plasma")
```



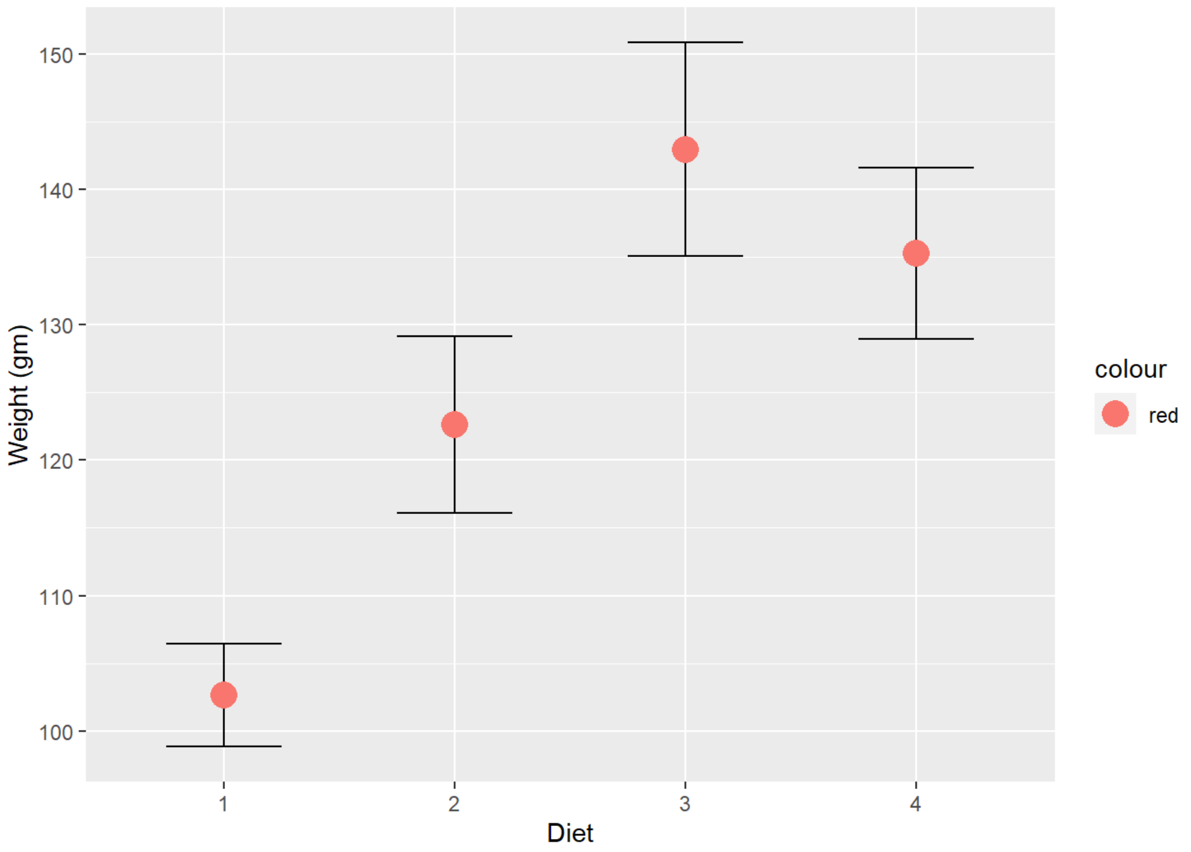
3.3 (2 pts) What diet seems to have the most variation in `weight` ? The least variation?

Diet 3 seems to have the most variation in mean weight while Diet 1 seems to have the least variation in mean weight.

3.4 (2 pts) Take your code from 3.2 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.

```
c <- ggplot(ChickWeight, aes(x = Diet, y = weight)) +
  geom_errorbar(stat="summary", width = 0.5, fun.data="mean_se") +
  geom_point(aes(color = "red"), stat = "summary", fun = "mean", size = 5) +
  scale_y_continuous(name = "Weight (gm)")

c + scale_fill_viridis_d(option = "plasma")
```



3.5 (2 pts) Does the mean chick weight seem to differ based on the diet? *No need to conduct hypothesis testing but informally state if they seem to differ and if so, how.*

The mean chick weight seems ot differ based in diet. However, the range of the first standard deviation of Diet 2, 3, and 4 all overlap so they do not differ very much. The standard error of the mean weight of chicks under Diet 1 is much lower than the others.

```
##        sysname        release        version       nodename        machine
##      "Windows"      "10 x64"  "build 18363"     "ROSE-XPS"       "x86-64"
##          login           user effective_user
##        "roseh"        "roseh"        "roseh"
```