HW 3

Enter your name and EID here:

Jongho Yoo (jy23294)

You will submit this homework assignment as 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 four different diets over time (measured 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)

# Visualize the first ten rows of the dataset
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:

```
# your code goes below (make sure to edit comment)
length(unique(ChickWeight$Chick))
```

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

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

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

```
length(unique(ChickWeight$Diet))
```

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

```
summary(ChickWeight[ChickWeight$Time == 0, ])
```

```
##      weight           Time       Chick    Diet
##  Min.   :39.00   Min.   :0   18     : 1   1:20
##  1st Qu.:41.00   1st Qu.:0   16     : 1   2:10
##  Median :41.00   Median :0   15     : 1   3:10
##  Mean   :41.06   Mean   :0   13     : 1   4:10
##  3rd Qu.:42.00   3rd Qu.:0   9      : 1
##  Max.   :43.00   Max.   :0   20     : 1
##                              (Other):44
```

- How many distinct chicks are there? **50**
- How many distinct time points? **12**
- How many distinct diet conditions? **4**
- How many chicks per diet condition at the beginning of the experiment? **Diet 1: 20, Diet 2: 10, Diet 3: 10, Diet 4: 10**

There are 50 distinct chicks. There are 12 distinct time points. There are 4 distinct diet conditions. At the beginning of the experiment: diet 1 has 20 chicks, diet 2 has 10 chicks, diet 3 has 10 chicks, and diet 4 has 10 chicks

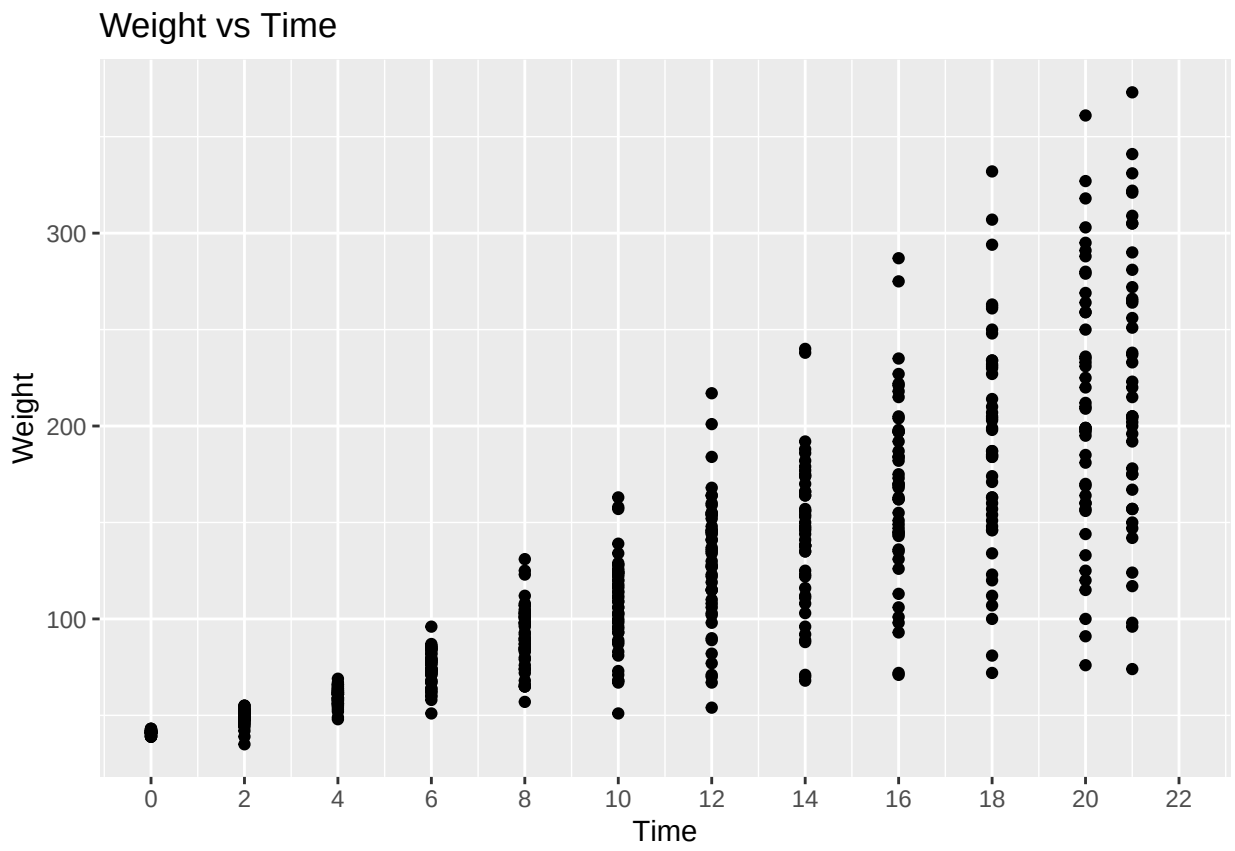
---

Question 2: (1 pt)

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

```
# Load package
library(ggplot2)

# scatterplot of Weight vs Time
ggplot(data = ChickWeight, mapping = aes(x = Time, y = weight)) +
  geom_point() +
  scale_x_continuous(breaks = seq(0,22,2), # adjust the tick marks of the x-axis
                     limits = c(0,22)) + # adjust the min/max tick marks of the x-axis
  labs(title = "Weight vs Time", x = "Time", y = "Weight")
```

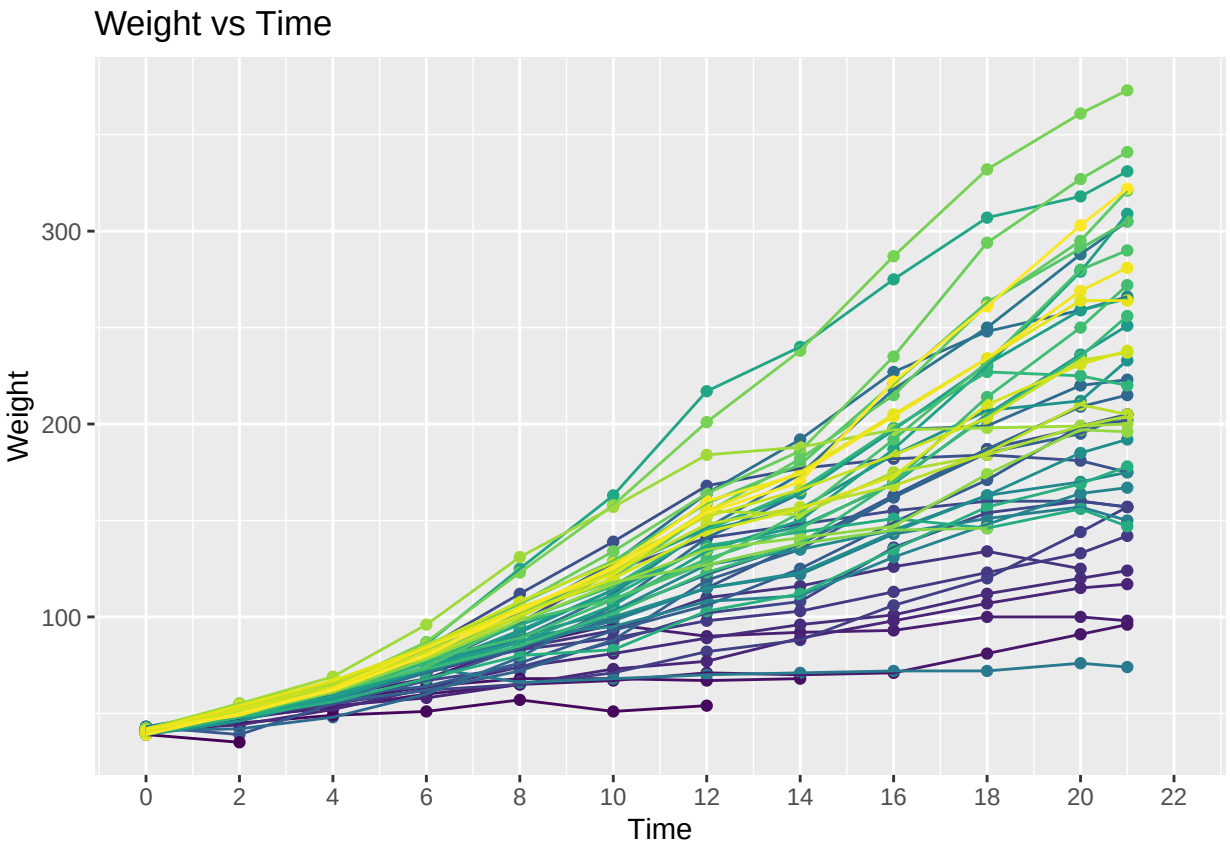


Weight seems to increase with time

Question 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 with `geom_line()` (also colored by each chick). Finally, remove the legend. Do all chicks seem to gain weight in the same manner? Why/Why not?

```
# scatterplot of Weight vs Time, with each chick point colored and connected with lines
ggplot(data = ChickWeight, mapping = aes(x = Time, y = weight, color = Chick)) +
  geom_point() +
  geom_line() +
  # remove legend
  theme(legend.position = "none") +
  # set x-scale
  scale_x_continuous(breaks = seq(0,22,2), limits = c(0,22)) +
  # plot labels
  labs(title = "Weight vs Time", x = "Time", y = "Weight")
```



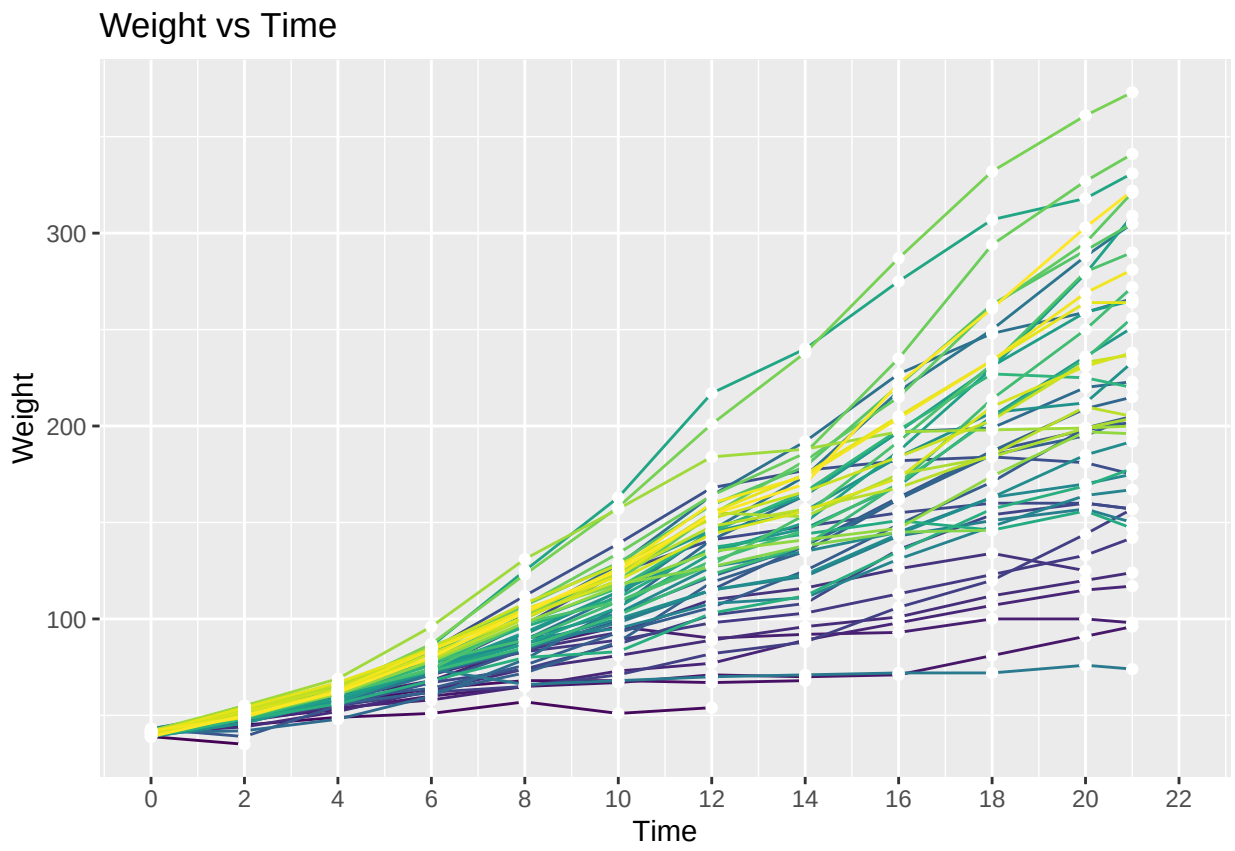
No, all chicks do not gain weight in the same manner. There are some chicks that barely gain weight over time, shown by the nearly horizontal slopes, whereas some have relatively strong positive slopes, meaning they gained lots of weight

---

Question 4: (1 pt)

Continue modifying the same graph by removing the color from the points only: leave the lines colored by chick, but make all of the points white. Make sure to put the points *on top of* the lines. On which day was the last value of the chicks' weight recorded?

```
# From graph in Q3, make color of points white and put in front of lines
ggplot(data = ChickWeight, mapping = aes(x = Time, y = weight)) +
  geom_line(aes(color = Chick)) +
  geom_point(color = "white") +
  theme(legend.position = "none") +
  scale_x_continuous(breaks = seq(0,22,2), limits = c(0,22)) +
  labs(title = "Weight vs Time", x = "Time", y = "Weight")
```

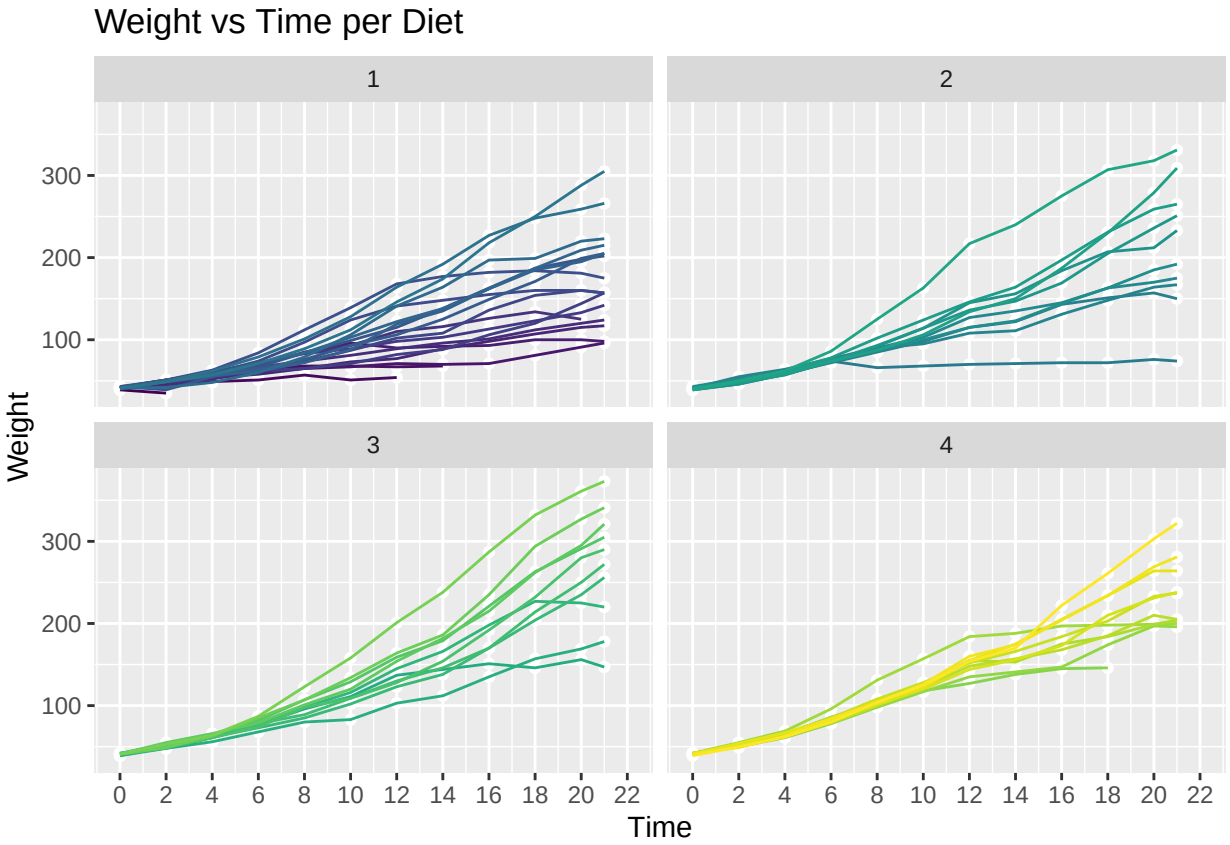


The last value of the chicks' weight was recorded on day 21

Question 5: (2 pts)

Now, facet this plot by diet. Can you tell from this new plot which diet results in greater weight? Describe how the relationship between `weight` and `Time` changes, or not, across the different diets.

```
# Facet wrap Q4 plot by the different diets
ggplot(data = ChickWeight, mapping = aes(x = Time, y = weight)) +
  geom_point(color = "white") +
  geom_line(aes(color = Chick)) +
  theme(legend.position = "none") +
  facet_wrap(~Diet) +
  scale_x_continuous(breaks = seq(0,22,2), limits = c(0,22)) +
  labs(title = "Weight vs Time per Diet", x = "Time", y = "Weight")
```

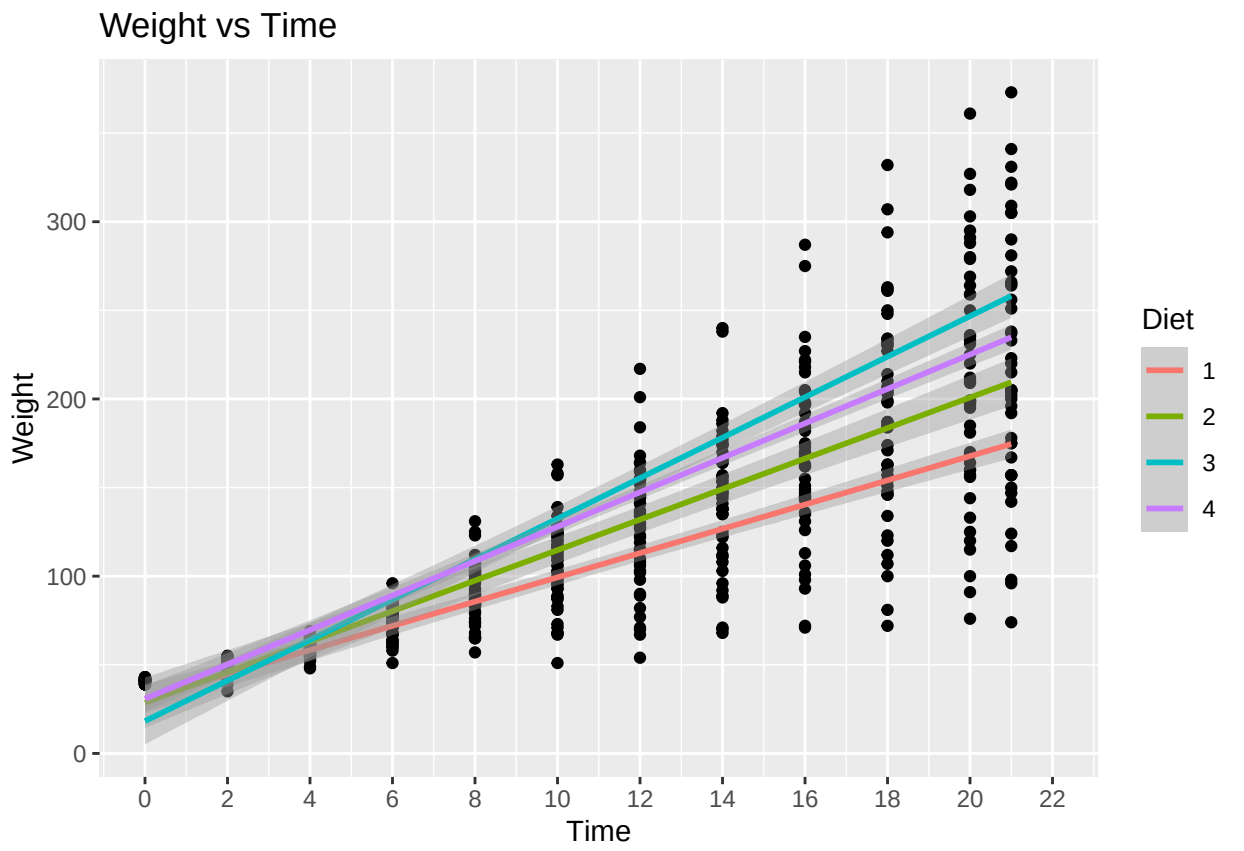


We cannot tell from this plot which diet results in greater weight. Each diet has multiple chicks with large positive slopes and chicks that weigh more than 300gm.

Question 6: (2 pts)

Go back to your plot from question 2 and fit a *linear regression line* (using `method = "lm"` in `geom_smooth()`) 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.

```
# Fit a linear regression line for each diet on Q2 plot
ggplot(data = ChickWeight, mapping = aes(x = Time, y = weight)) +
  geom_point() +
  geom_smooth(method = "lm", aes(color = Diet)) +
  scale_x_continuous(breaks = seq(0,22,2), limits = c(0,22)) +
  labs(title = "Weight vs Time", x = "Time", y = "Weight")
```



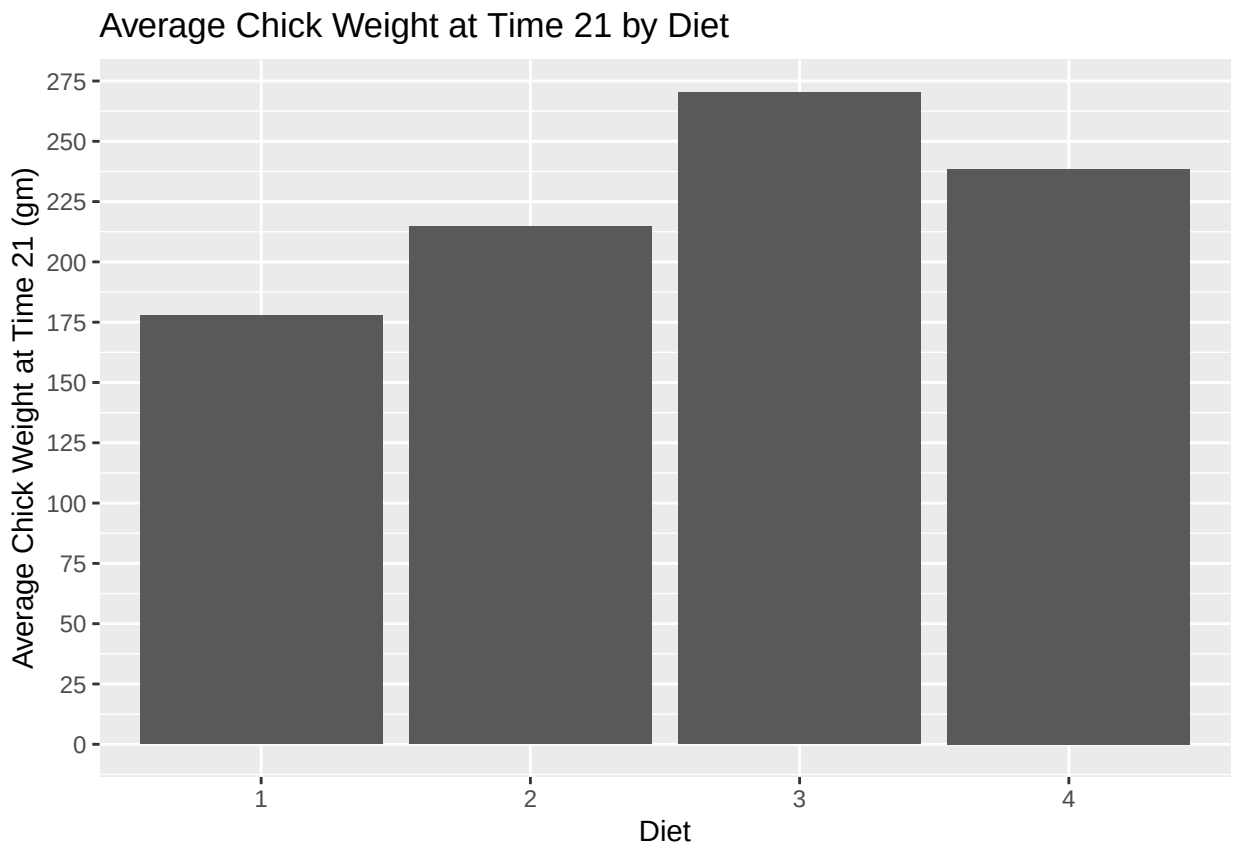
This goes make it easier to see which diet results in greater weight. It seems like diet 3 results in greatest weight compared to the other 3 diets, and diet 1 results in the lowest weight.

Question 7: (2 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 exploring the effect of diet would be to construct a bar graph representing group means at the end of the experiment.

Index the data in the `ggplot` function to focus on the last `Time` point. *Hint: Refer to Question 4 or find the max `Time`).* Then 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 300 by 25 (e.g., with scale_y_continuous(`breaks=`...)). Which diet has the highest mean `weight`?

```
# bar plot of average weight of each diet at the end of the experiment (time = 21)
ggplot(data = ChickWeight[ChickWeight$Time == 21, ], aes(x = Diet, y = weight)) +
  geom_bar(stat = "summary", fun = "mean") +
  scale_y_continuous(name = "Average Chick Weight at Time 21 (gm)", breaks = seq(0, 300, 25)) +
  labs(title = "Average Chick Weight at Time 21 by Diet", x = "Diet")
```

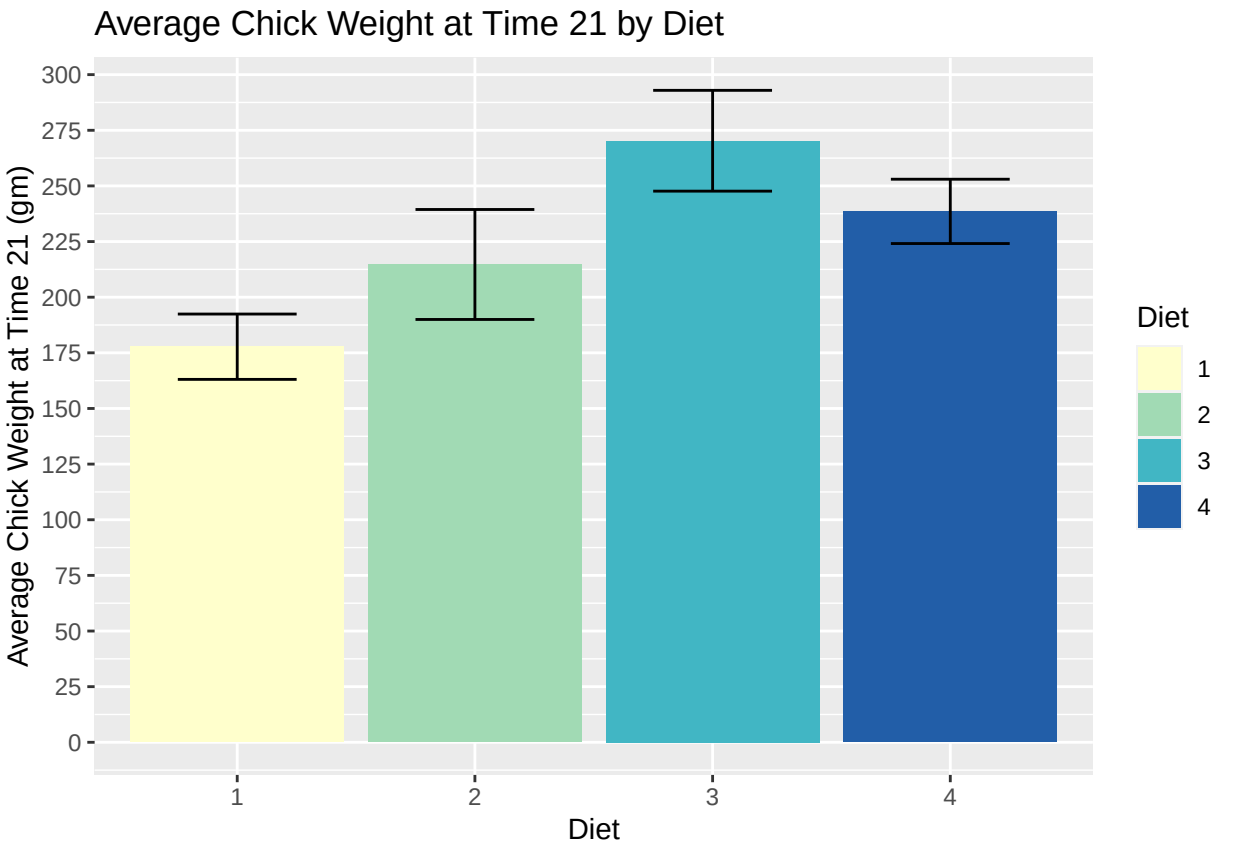


Diet 3 has the highest mean weight.

Question 8: (2 pts)

Building on the previous graph, 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 bar graph bars) by diet and change from the default color scheme using a `scale_fill_` or a `scale_color_`. Compare the different diets in terms of variation in `weight`.

```
# add error bars and color to Q7 plot
ggplot(data = ChickWeight[ChickWeight$Time == 21, ], aes(x = Diet, y = weight, fill = Diet)) +
  geom_bar(stat = "summary", fun = "mean") +
  geom_errorbar(stat = "summary", fun.data = "mean_se", width = 0.5) +
  scale_y_continuous(name = "Average Chick Weight at Time 21 (gm)", breaks = seq(0, 300, 25)) +
  labs(title = "Average Chick Weight at Time 21 by Diet", x = "Diet") +
  scale_fill_brewer(palette = "YlGnBu")
```

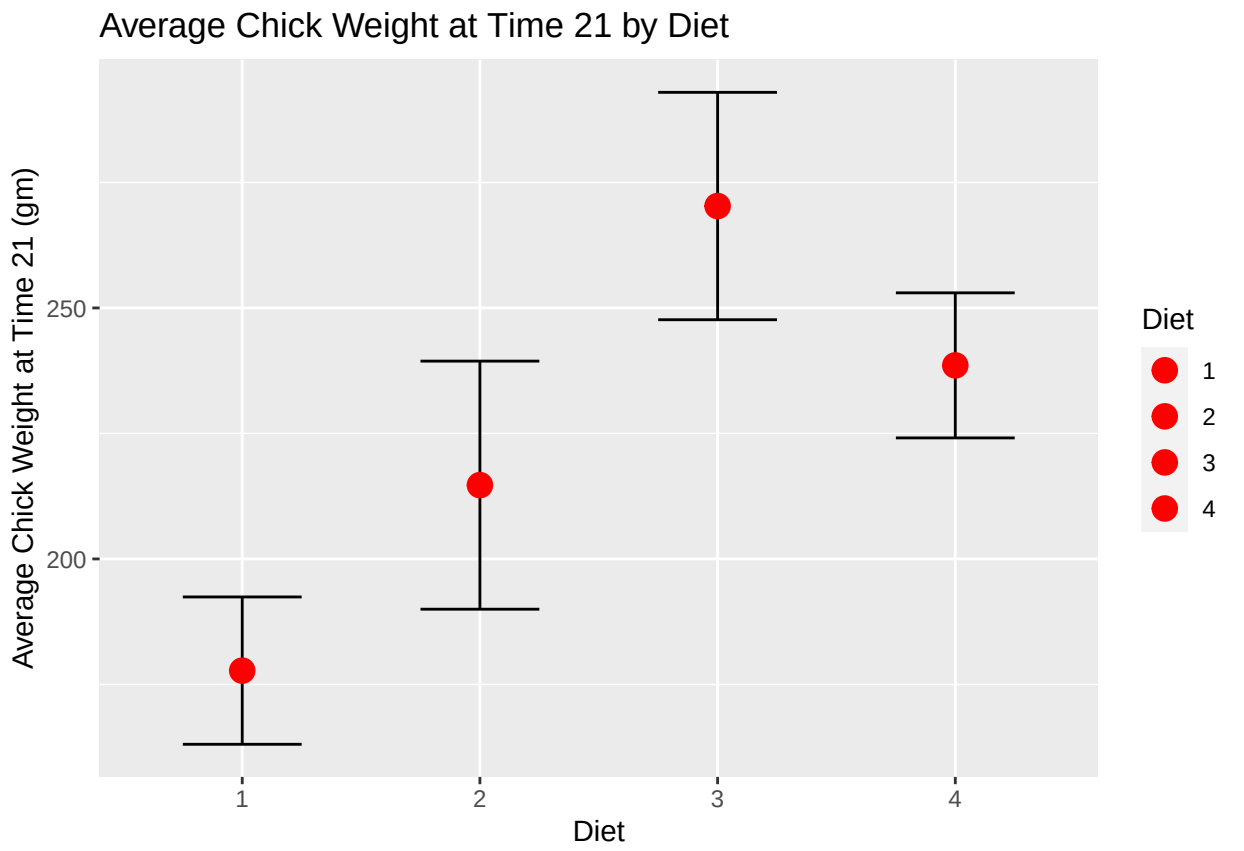


Looking at the error bars, diets 1 and 4 seem to have similar variation (about 25gm), and diets 2 and 3 seem to have similar variation (about 50gm). In addition, diets 2 and 3 have higher variation in weight than diets 1 and 4.

Question 9: (2 pts)

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? *Note: Informally state if they seem to differ and if so, how.*

```
# add points to indicate mean weight for each diet, and keep errors bars
ggplot(data = ChickWeight[ChickWeight$Time == 21, ], aes(x = Diet, y = weight, fill = Diet)) +
  geom_errorbar(stat = "summary", fun.data = "mean_se", width = 0.5) +
  geom_point(stat = "summary", fun = "mean", size = 4, color = "red") +
  scale_y_continuous(name = "Average Chick Weight at Time 21 (gm)") +
  labs(title = "Average Chick Weight at Time 21 by Diet", x = "Diet")
```

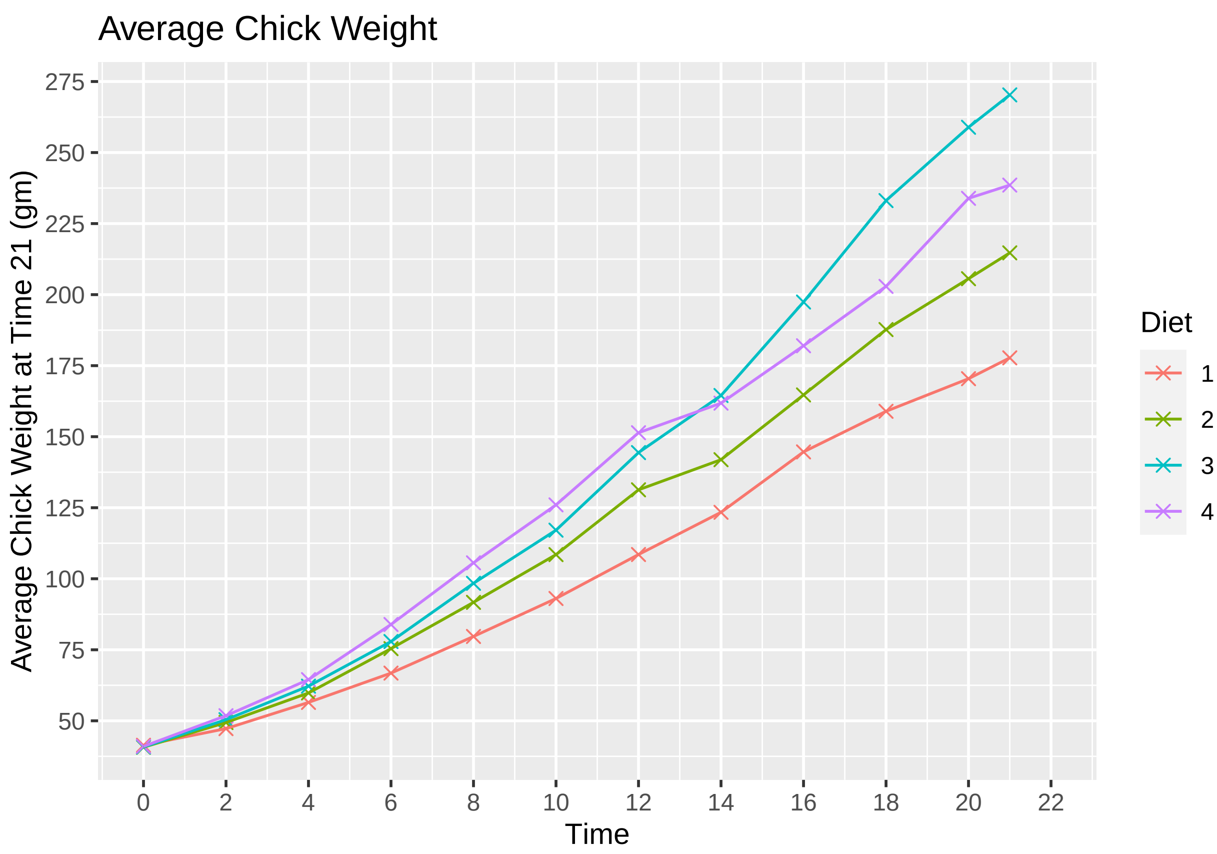


Yes, mean chick weight does seem to differ based on diet. We can clearly see that the mean weights rank from: Diet 3, diet 4, diet 2, and diet 1, ranking from highest to lowest.

Question 10: (2 pts)

One last graph! And a little less guided... It would be even more meaningful to compare the mean `weight` of each `Diet` over `Time`! Use `geom_point` to represent the mean weight over time and `geom_line` to connect the mean weights per diet. Change the shape of the points to be `x` symbols. *Giving you a hint anyway: color by diet and use some* ***`stat`*** *options in the geoms!* Which diet has a constantly lower mean weight over time?

```
# Plot of mean weight per diet condition over time
ggplot(data = ChickWeight, aes(x = Time, y = weight, color = Diet)) +
  geom_point(stat = "summary", fun = "mean", shape = 4, size = 2) +
  geom_line(stat = "summary", fun = "mean") +
  scale_y_continuous(name = "Average Chick Weight at Time 21 (gm)", breaks = seq(0, 300, 25)) +
  scale_x_continuous(breaks = seq(0,22,2), limits = c(0,22)) +
  labs(title = "Average Chick Weight", x = "Time")
```



Diet 1 has a constantly lower mean weight over time

Formatting: (2 pts)

Comment your code, write full sentences, and knit your file!

```
##                                                                                                sysr
##                                                                                               "Darw
##                                                                                                rele
##                                                                                               "21.3
##                                                                                                vers
## "Darwin Kernel Version 21.3.0: Wed Jan  5 21:37:58 PST 2022; root:xnu-8019.80.24~20/RELEASE_ARM64_T8
##                                                                                               noder
##                                                                                    "Stevens-MBP-2.l
##                                                                                                mach
##                                                                                                "arm
##                                                                                                  lo
##                                                                                                 "ro
##                                                                                                   u
##                                                                                          "steveny
```

```
##                                                                                                                effective_u
##                                                                                                                   "steveny
```