HW 2

Enter your name and EID here: Austine Do (ahd589)

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.

Part 1

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

```
# Take a look at 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
```

Question 1: (2 pts)

Answer the following questions using code:

- How many distinct chicks are there?
- How many distinct time points?
- How many distinct diet conditions?
- How many chicks per diet condition at the beginning of the experiment?

Hint: the functions `length()` and `table()` might be handy!

```r
# Saving data frame
ahd589 <- ChickWeight

# the distinct chicks in the data frame
length(unique(ahd589$Chick))
```

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

```r
# the distinct time points in the data frame
length(unique(ahd589$Time))
```

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

```r
# the distinct type of diets
length(unique(ahd589$Diet))
```

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

```r
# count of chicks per diet condition
beginning_experiment <- ahd589[ahd589$Time == 0,]
table(beginning_experiment$Diet)
```

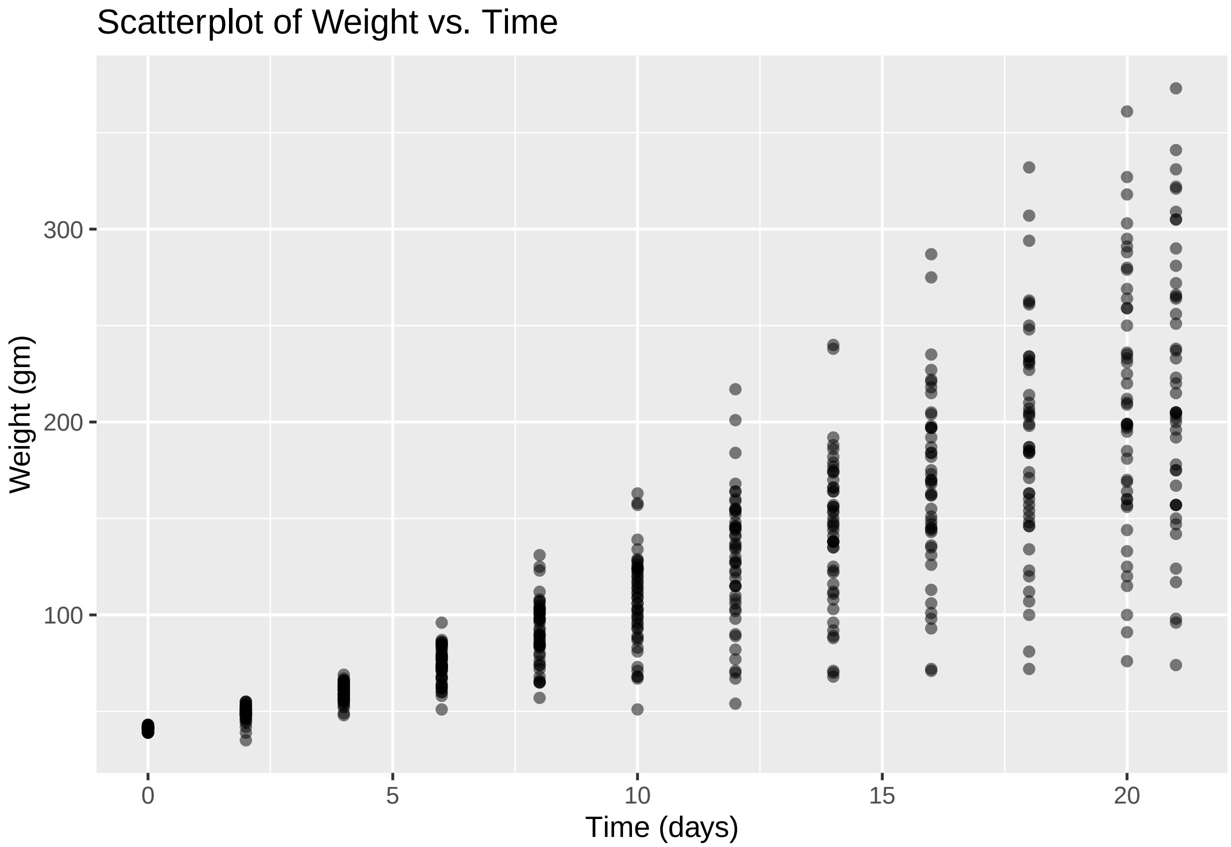
```
##
##  1  2  3  4
## 20 10 10 10
```

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

Question 2: (3 pts)

Using a `ggplot`, create a scatterplot showing chick `weight` (on the y-axis) depending on `Time`. Add a title to the plot and label the axes, including the units of the variables.

```r
# Scatter plot of weight as a function of time
ggplot(data = ahd589) +
    geom_point(aes(x = Time, y = weight), alpha = 0.50) +
    labs(title = 'Scatterplot of Weight vs. Time',
         x = 'Time (days)',
         y = 'Weight (gm)')
```

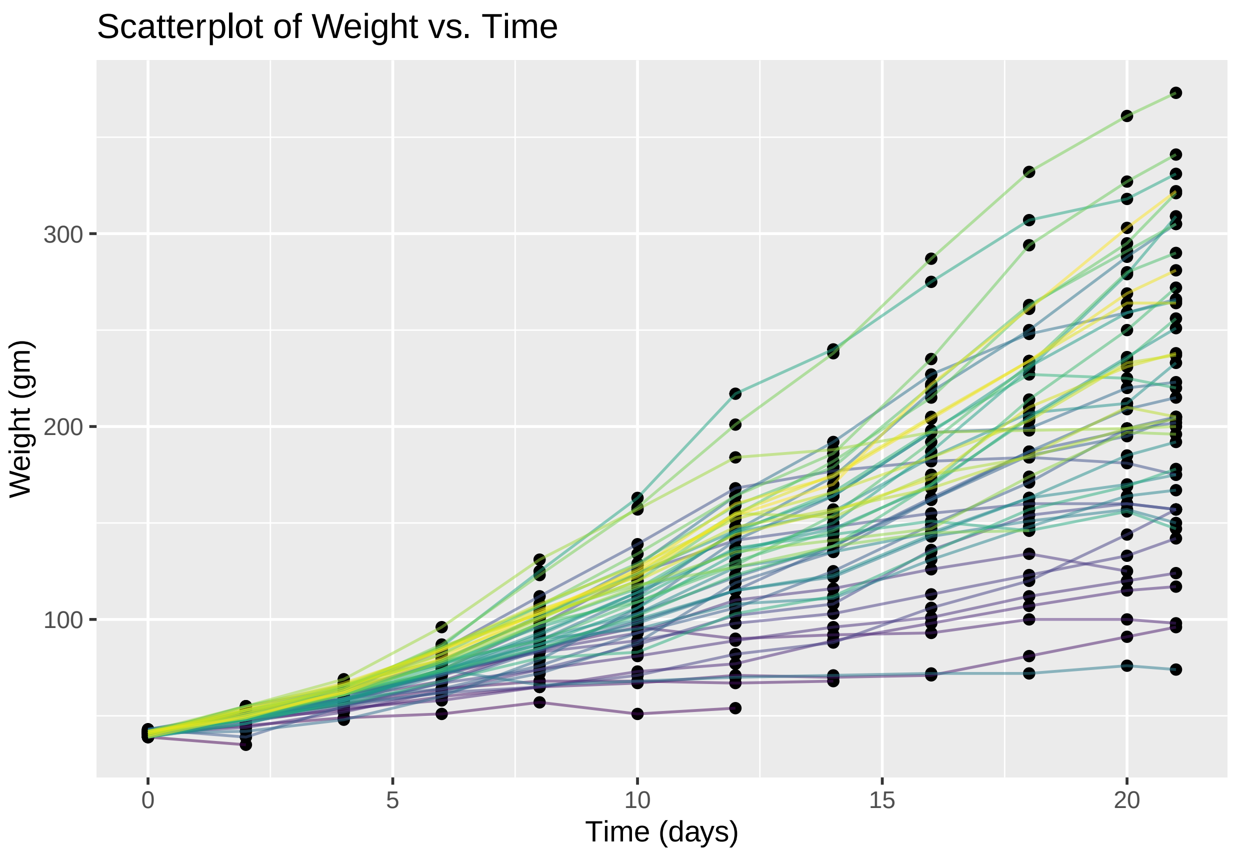


How does chick `weight` change over `Time`?

After visually inspecting the graph it seems that the trend is that as time increases weight typically increases as well. Weight and Time have have positive relationship.

Building upon the previous plot, add lines that connect each chick's points together with `geom_line()` and are represented with a different color per chick. Make sure the points representing the chicks are on top of the lines (each point should still be black by the way). Finally, remove the legend.

```r
#
ggplot(data = ahd589) +
    geom_point(aes(x = Time, y = weight)) +
    geom_line(aes(x = Time, y = weight, group = Chick, color = Chick), alpha = 0.5) +
    theme(legend.position='none') +
    labs(title = 'Scatterplot of Weight vs. Time',
         x = 'Time (days)',
         y = 'Weight (gm)')
```



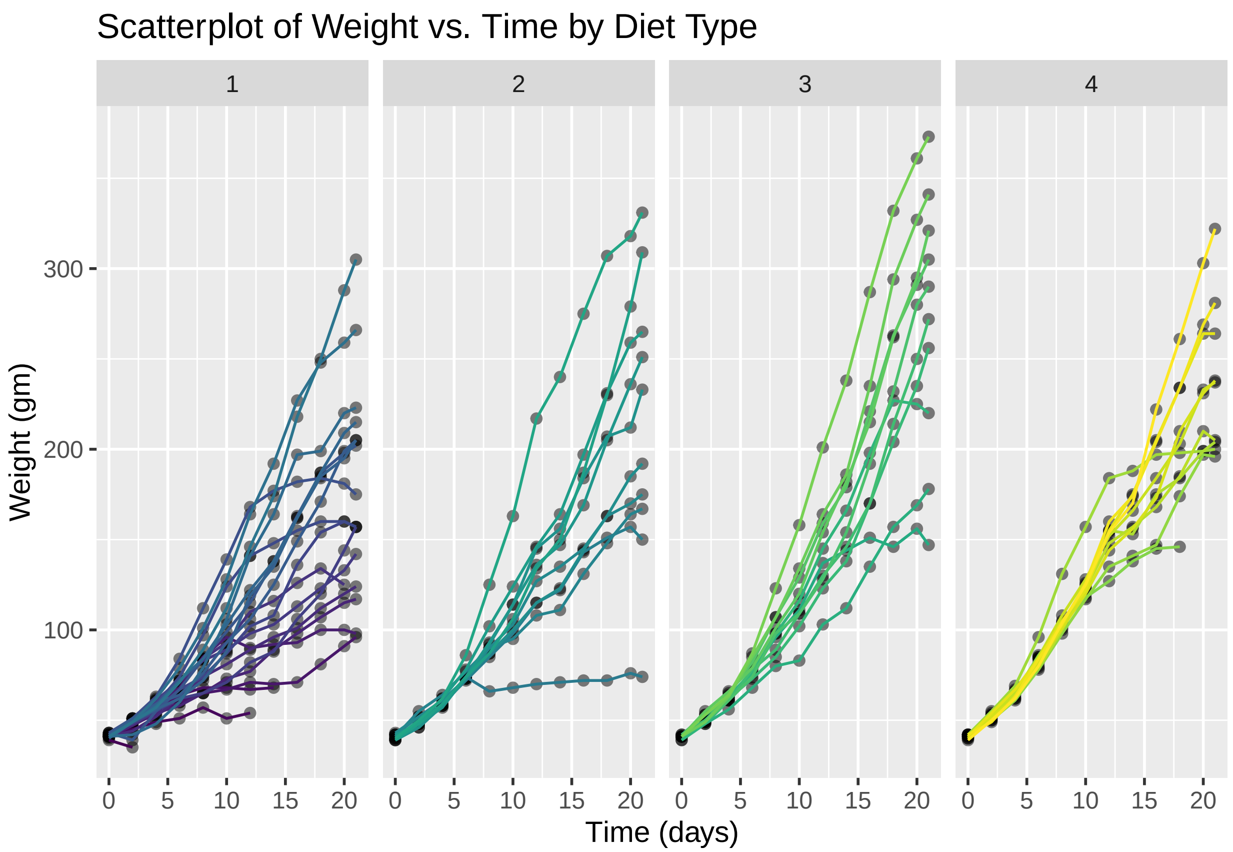
Do all chicks seem to gain weight in the same manner? Why/Why not?

No the chicks do not seem to gain weight in the same manner because from the graph we can see that as time increases, although weight does typically increase for all chicks, the spread of the weight also increases and it is evident that the chicks don't all gain weight in the same manner. If all chicks were to gain weight in the same manner, the weight on the later days (i.e. days 15-21) wouldn't vary as much like in the first few days (i.e. days 0-5).

Question 3: (2 pts)

Now, facet your last plot by diet.

```
# your code goes below (replace this comment with something meaningul)
ggplot(data = ahd589) +
    geom_point(aes(x = Time, y = weight), alpha = 0.50) +
    geom_line(aes(x = Time, y = weight, color = Chick)) +
    theme(legend.position='none') +
    facet_wrap(~Diet, nrow = 1) +
    labs(title = 'Scatterplot of Weight vs. Time by Diet Type',
         x = 'Time (days)',
         y = 'Weight (gm)')
```



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.

No you cannot explicitly tell which diet results in a greater weight. The relationship between weight and time remains a positive relationship across all diets.

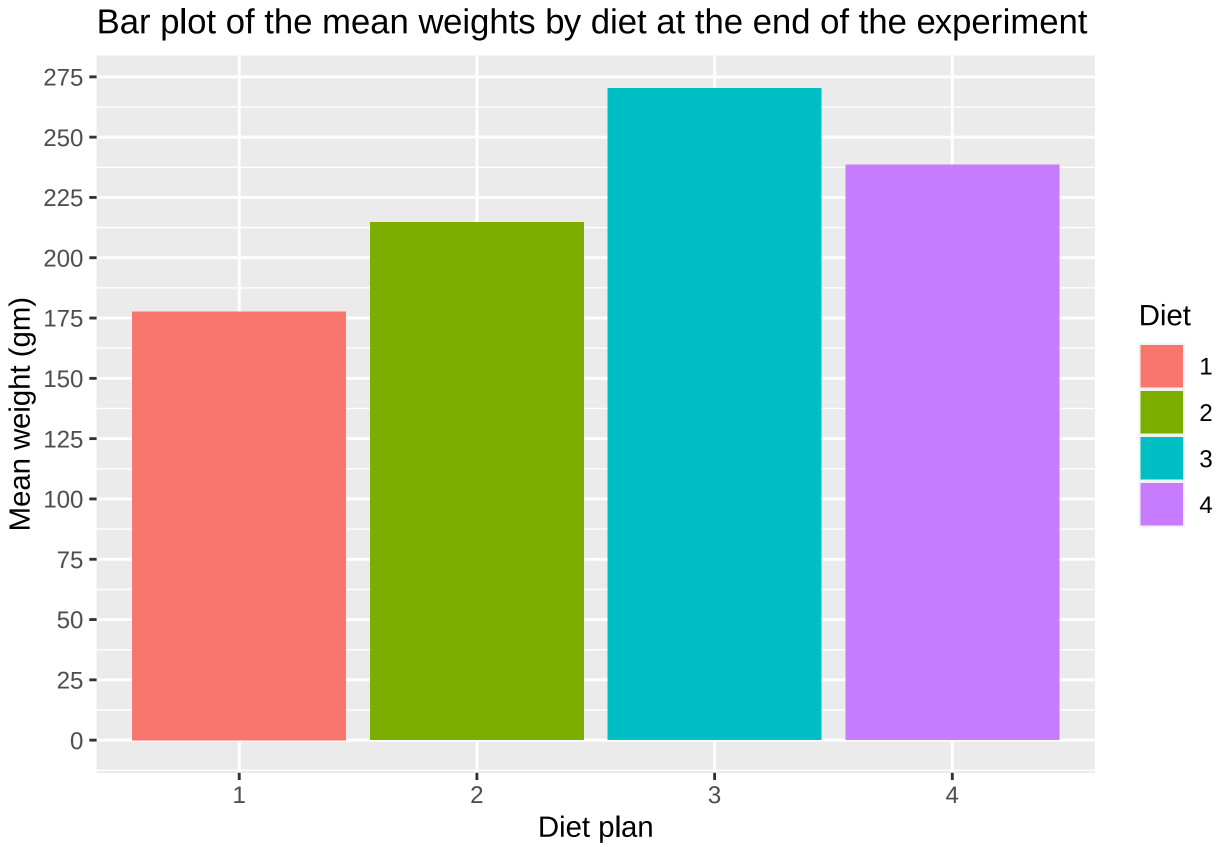
Question 4: (2 pts)

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.

Only focus on the last `Time` point. *Hint: find the maximum value of `Time`.* Then create a `ggplot` using `geom_bar` where each bar's height corresponds to a statistic: the mean weight for each of the four diet conditions. Label the y-axis to include units and make the major tick marks go from 0 to 300 by 25.

```
# This creates a bar plot that displays the mean weight at the end of the experiment by diet
end_of_experiment <- ahd589[ahd589$Time == max(ahd589$Time),]

ggplot(data = end_of_experiment, aes(x = Diet, y = weight, fill = Diet)) +
    geom_bar(stat = 'summary', fun = 'mean') +
    scale_y_continuous(breaks = seq(0,300,25)) +
    labs(title = 'Bar plot of the mean weights by diet at the end of the experiment',
         y = 'Mean weight (gm)',
         x = 'Diet plan')
```



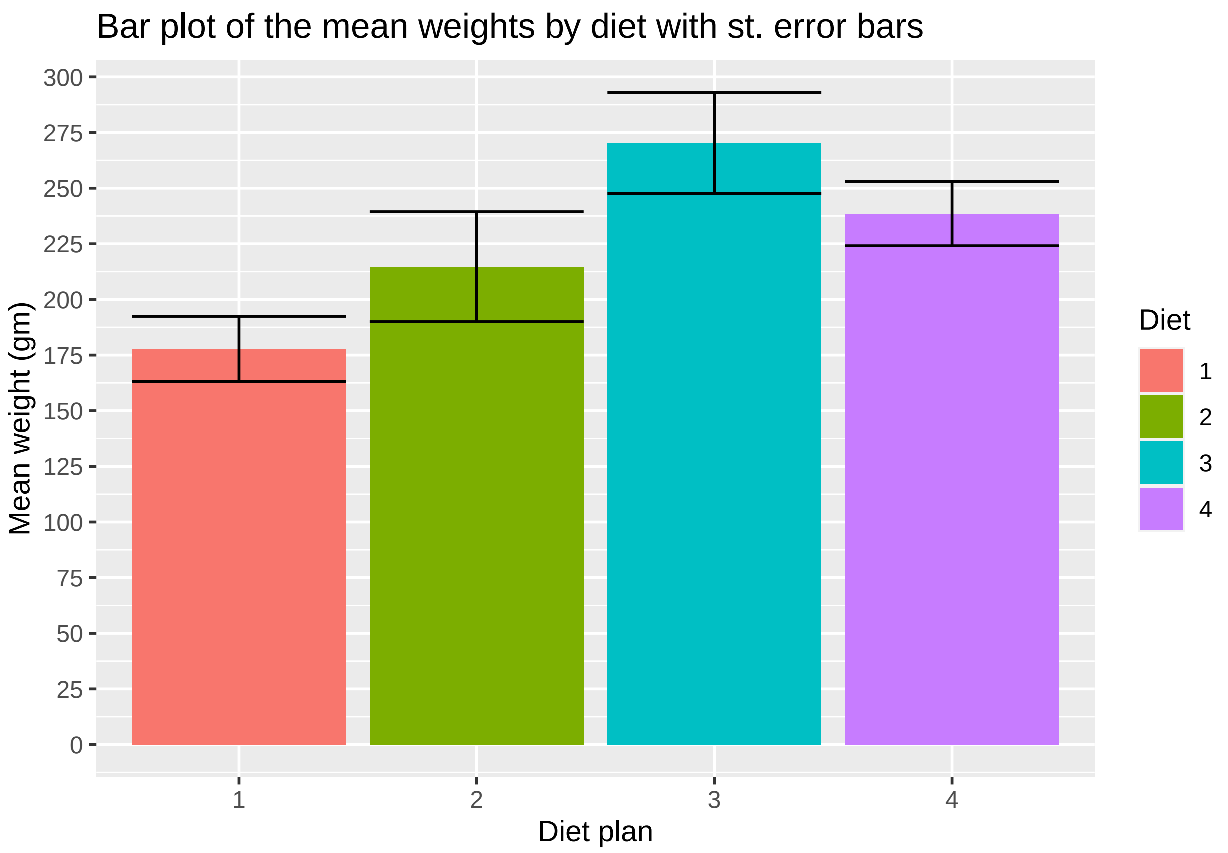
Which diet has the highest mean `weight`?

The diet that has highest mean weight is diet 3.

Question 5: (2 pts)

Building on the previous graph, add error bars showing + or - 1 standard error (). Fill the bars (not the error bars, but the bar graph bars) by diet.

```
# This code adds error bars to the plot above
ggplot(data = end_of_experiment,aes(x = Diet, y = weight, fill = Diet)) +
    geom_bar(stat = 'summary', fun = 'mean') +
    scale_y_continuous(breaks = seq(0,300,25)) +
    labs(title = 'Bar plot of the mean weights by diet with st. error bars',
         y = 'Mean weight (gm)',
         x = 'Diet plan') +
    geom_errorbar(stat = 'summary', fun.data = 'mean_se')
```



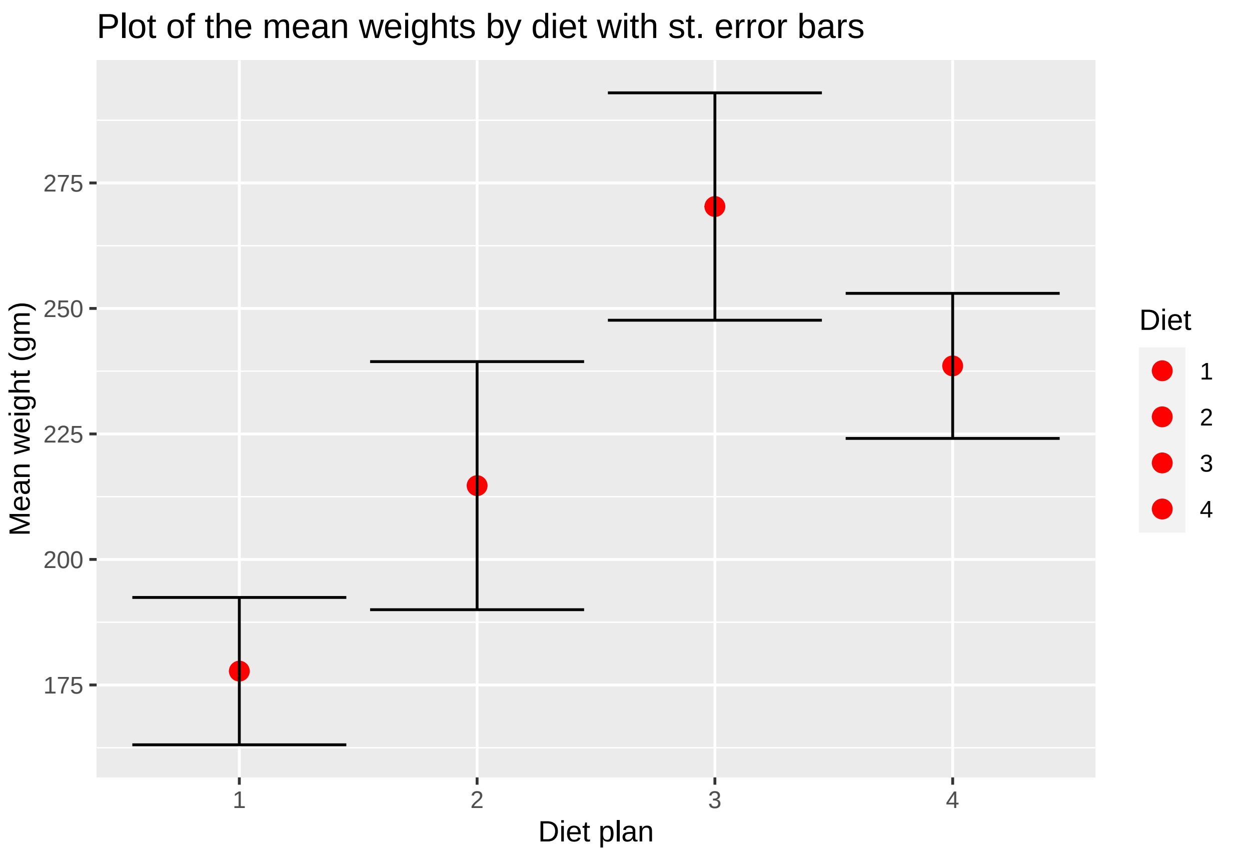
Compare the different diets in terms of variation in `weight`.

The diets with the most variation in weight are diet 2 and diet 3 while diet 1 and diet 4 seem to have similar and lower variation in weight compared to diet 2 an diet 3.

Question 6: (2 pts)

Copy your code from the previous question and replace `geom_bar()` with `geom_point()`. Make the points larger and color them all in red. Put them *on top of* the error bars.

```
# your code goes below (replace this comment with something meaningul)
ggplot(data = end_of_experiment,aes(x = Diet, y = weight, fill = Diet)) +
    geom_point(stat = 'summary', fun = 'mean', color = 'red', size = 3) +
    scale_y_continuous(breaks = seq(0,300,25)) +
    labs(title = 'Plot of the mean weights by diet with st. error bars',
         y = 'Mean weight (gm)',
         x = 'Diet plan') +
    geom_errorbar(stat = 'summary', fun.data = 'mean_se')
```



```
# note: [38;5;255mRemoved 8 rows containing non-finite values (`stat_summary()`).[39m
# when using c(0,300) as the limit
```

Does the mean chick weight seem to differ based on the diet? *Note: Informally state if they seem to differ and if so, how.*

It seems that the mean weight differs across diet 1, 2, and 3 since the range of the error bars do not overlap greatly, indicating that it is more likely that the mean weights across the 3 diets are different. This is also the case for diets 3 and 4. On the other hand the mean weights between diet 2 and diet 4 don't seem to differ and the range error bars do overlap more, indicating that the mean weights might not differ greatly.

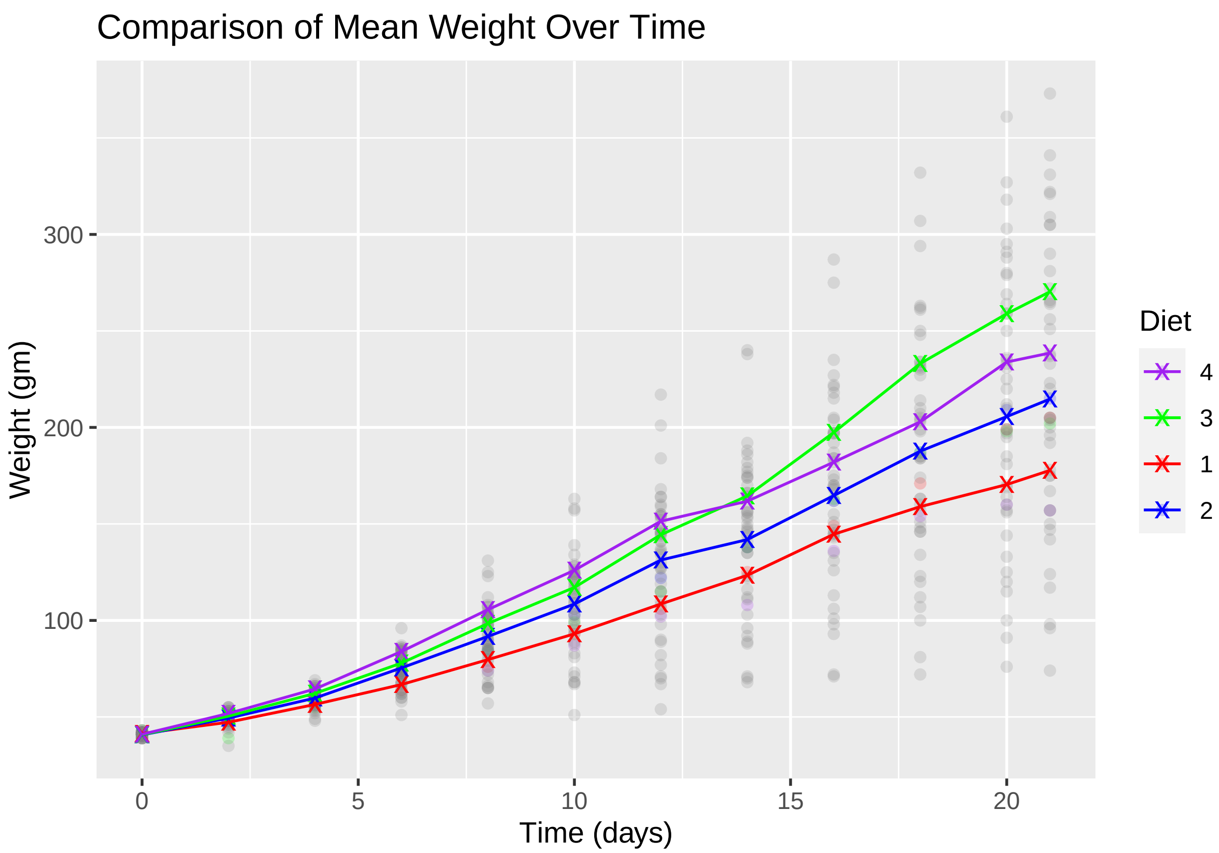
Question 7: (2 pts)

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 across time. Also represent the weight of each chick over time. Change the shape of the points representing the means to be `x` symbols and make these points bigger. To better distinguish between data values and mean values make the points representing the weight of each chick 20% transparent. *Giving you a hint anyway: use some `stat` options in the geoms and define aesthetics wisely (overall or within a geom)!*

```r
# This creates a plot that draws lines for the mean weights over time for each diet over a plot
# of the weight progression of each chick

custom_colors_lines <- c("1" = "red", "2" = "blue", "3" = "green", "4" = "purple")

ggplot(data = ahd589, aes(x = Time, y = weight)) +
    geom_point(aes(color = Chick), alpha = 0.2) +
    geom_point(stat = 'summary', fun = 'mean', shape = 'x', size = 4, aes(color = Diet, group = Diet)) +
    geom_line(stat = 'summary', fun = 'mean', aes(color = Diet, group = Diet)) +
    labs(title = 'Comparison of Mean Weight Over Time',
       x = 'Time (days)',
       y = 'Weight (gm)') +
    scale_color_manual(values = custom_colors_lines, name = "Diet")
```



Which diet has a constantly lower mean weight over time?

The diet that constantly had a lower mean weight over time was diet 1.

Part 2

Recall the context about the Internet clothing retailer Stitch Fix wanting to develop a new model for selling clothes to people online (see HW 1). Their basic approach is to send people a box of 5–6 items of clothing and allow them to try the clothes on. Customers keep (and pay for) what they like while mailing back the remaining clothes. Stitch Fix then sends customers a new box of clothes a month later.

You built an intake survey distributed to customers when they first sign up for the service. The intake survey had about 20 questions and you will use this data later to help you develop an algorithm for choosing clothes to send to each customer.

Suppose you are now in charge of producing a report for the department of Inventory Management at Stitch Fix so that clothes can be ordered to stock the inventory before starting to send clothes to each customer.

Question 8: (2 pts)

What are some steps you should undertake to produce the report for the department of Inventory Management? What would be the goal of these steps? Come up with at least 2 steps and comment on their goals. *To give you an idea of what I am asking here: making visualizations of the intake survey data would be an example of a step but what would be some possible goals for making such visualizations? I can think of at least 2!*

The first step of the report I would take is data cleaning and pre-processing. The goal of this step is not only to report clean, consistent, and accurate survey data, but ensure that the quality of data is good for business insights that may later be extracted in the future. The second step I would take is forecasting demand. The goal with this to accurately predict and therefore stock the correct items in inventory during different periods of time to meet customer demands as demands change. The last step I would take is exploratory data analysis on the survey data. The goal of this step is to gain insights on popular styles, fashion trends, and more that would keep inventory matching the general consumer preference.

Question 9: (2 pts)

Consider the data from the intake survey with the following variable collected for each customer: size (S, M, L, XL, XXL), favorite color to wear (blue, yellow, burgundy, . . .), preferred type of clothes (skirt, pants, dress, . . .), and hip size (in cm). What visualizations and summary statistics should you report for the department of Inventory Management? Come up with at least 2 visualizations and corresponding summary statistics using 3 of the variables.

The first visualization I would make is a histogram of the distribution of hip size and report median, mode, and IQR to summarize the distribution of hip sizes from customers. The second visualization I would make is a bar plot of color preference and report frequency/count and proportion to summarize the color preference of customers. The last visualization I would make, if possible, is a grouped bar plot of types of clothes and size, reporting the frequency and proportion of sizes per clothing type to better gauge what is the average or most occurring size per clothing type of the customers. These visualization will help inventory maintain adequate stock of certain items in specific sizes and help minimize costs associated with inventory.

Formatting: (1 pt)

Knit your file! You can knit into html and once it knits in html, click on `Open in Browser` at the top left of the window that pops out. **Print** your html file into pdf from your browser.

Is it working? If not, try to decipher the error message: look up the error message, consult websites such as stackoverflow or crossvalidated.

Finally, remember to select pages for each question when submitting your pdf to Gradescope.