Ethan Chang - ehc586

Question 1: (2 pts)

Question 2.1: (2 pts)

Question 2.2: (1 pts)

Question 2.3: (1 pts)

Question 2.4: (2 pts)

Question 2.5: (1 pts)

Question 3: (2pts)

HW 3

SDS322E

September 14, 2022

Ethan Chang - ehc586

Please submit as a PDF or HTML file on Canvas before the due date.

For all questions, include the R commands/functions that you used to find your answer. Answers without supporting code will not receive credit.

Review of how to submit this assignment

All homework assignments will be completed using R Markdown. These `.Rmd` files consist of >text/syntax (formatted using Markdown) alongside embedded R code. When you have completed the assignment (by adding R code inside codeblocks and supporting text outside of the codeblocks), create your document as follows (assuming you are using the edupod server and submitting HTML):

- Click the arrow next to the "Knit" button (above)
- Choose "Knit to HTML"
- Go to Files pane and put checkmark next to the correct HTML file
- Click on the blue gear icon ("More") and click Export
- Download the file and then upload to Canvas

> To submit a PDF, open your HTML file and print it to a pdf, then upload the pdf as your submission.

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).

```
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
```

Answer the following questions:

- How many distinct chicks are there (e.g., each row with the same value of `Chick` correspond to the same chick)?

- How many distinct time points?

- How many distinct diet conditions?

- How many chicks per diet condition at the start of the experiment (i.e., at Time 0)? For example, you could use some combination of `table()` and `length()` to answer these questions. *Do not forget to include the code you used to answer these questions!*

```
length(table(ChickWeight$Chick))
```

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

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

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

```
length(table(ChickWeight$Diet))
```

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

```
chickdiet_time0 <- ChickWeight$Diet[ChickWeight$Time ==
    0]
sum(chickdiet_time0 == 1)
```

```
## [1] 20
```

```
sum(chickdiet_time0 == 2)
```

```
## [1] 10
```

```
sum(chickdiet_time0 == 3)
```

```
## [1] 10
```

```
sum(chickdiet_time0 == 4)
```

```
## [1] 10
```

There are 50 distinct chicks, 12 distinct time points, and 4 distinct diet conditions. There is an average of 12.5 chicks per diet condition at the start of the experiment, with 20 chicks being on diet 1, and 10 chicks being on diets 2, 3, and 4 each (50 total chicks and 4 total diets at the start).

Question 2.1: (2 pts)

Using the ggplot2 package, create a simple scatterplot showing chick `weight` (y-axis) as a function of `Time` . This time, relabel the axes more informatively and give the plot a title.

```
library(tidyverse)
library(ggplot2)
ggplot(data = ChickWeight, aes(x = Time, y = weight)) +
    geom_point() + ggtitle("Chick Weight over Time") +
    xlab("Time (days)") + ylab("Weight (gm)")
```

Chick Weight over Time
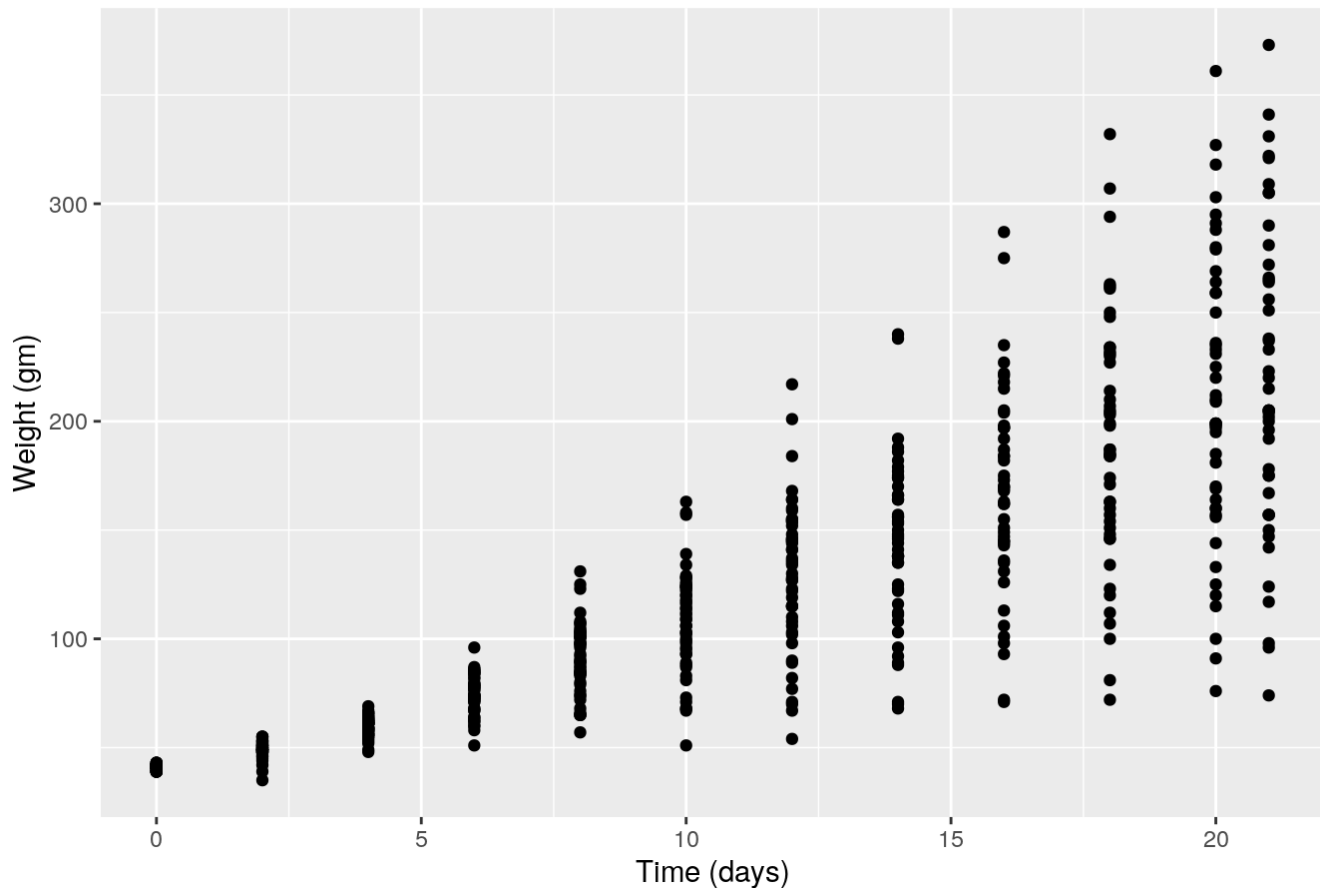
Question 2.2: (1 pts)

Building upon the previous plot, do the following:,

- 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. This can be done using `theme(legend.position = "none")`. Run `?theme` to see the options available for the themes.

```
ggplot(data = ChickWeight, aes(x = Time, y = weight,
    color = Chick)) + geom_point() + ggtitle("Chick Weight over Time") +
    xlab("Time (days)") + ylab("Weight (gm)") + geom_line() +
    theme(legend.position = "none")
```
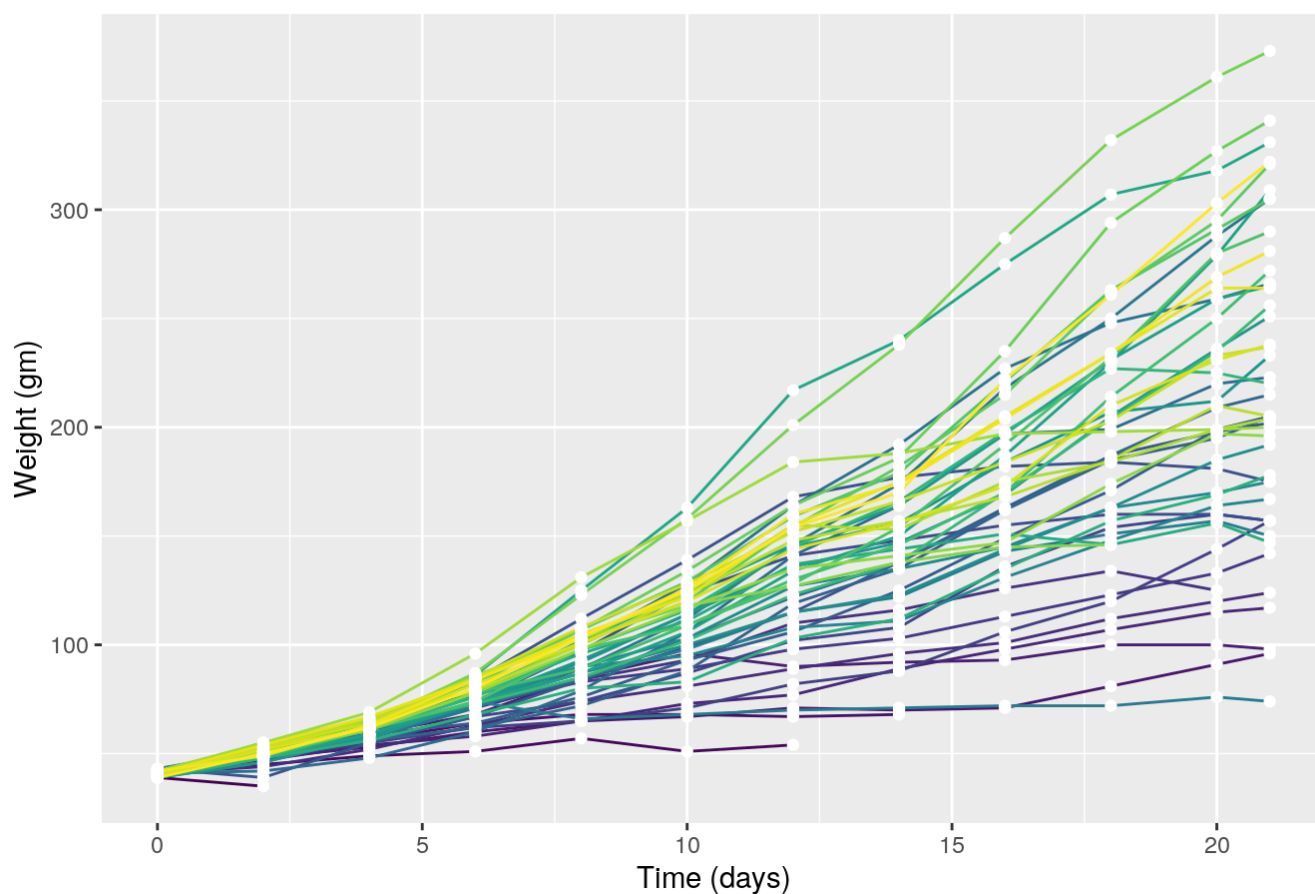
Chick Weight over Time
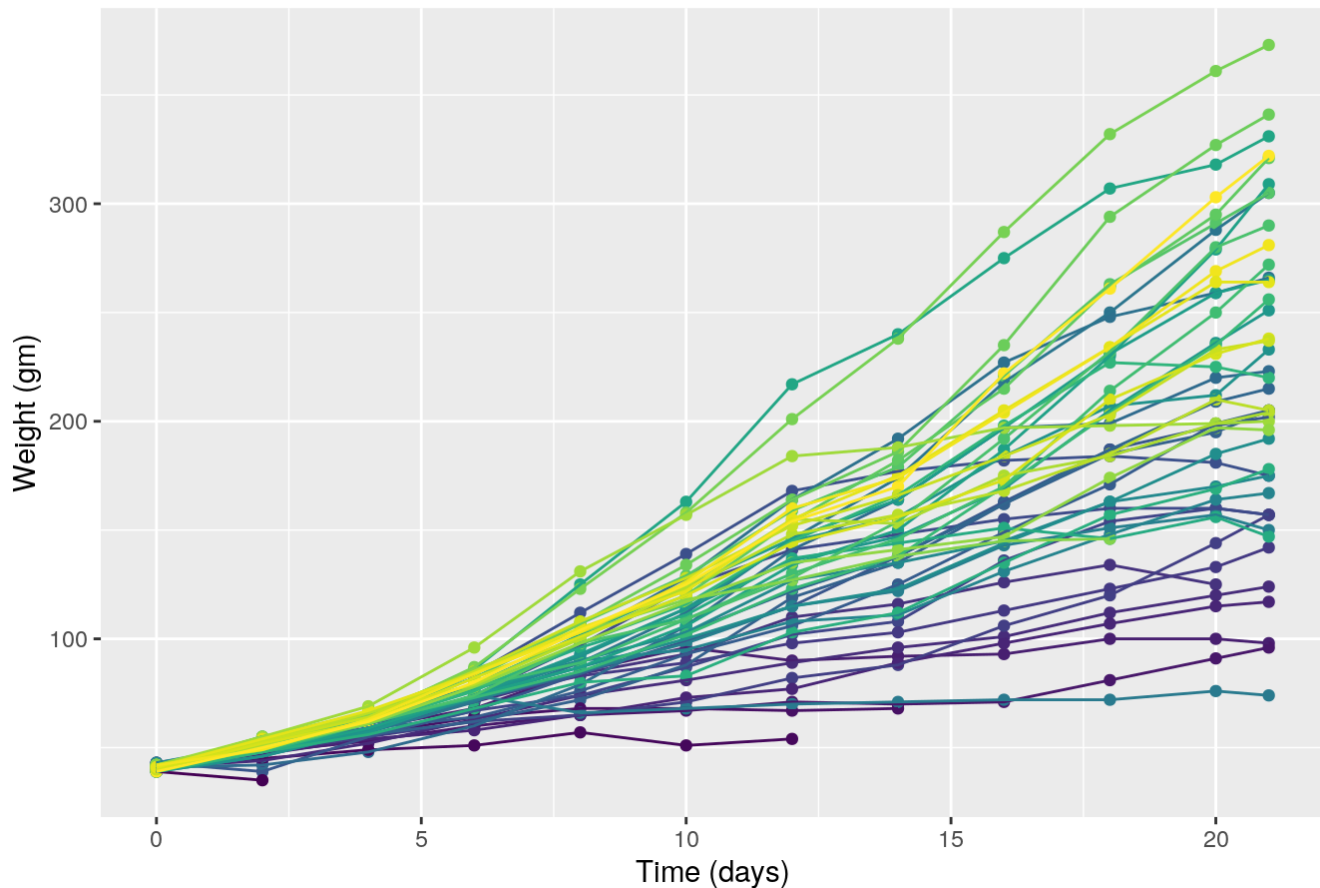

```
`?`(theme)
```

Question 2.3: (1 pts)

Take the plot you made in Question 2.2 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(data = ChickWeight, aes(x = Time, y = weight,
    color = Chick)) + ggtitle("Chick Weight over Time") +
    xlab("Time (days)") + ylab("Weight (gm)") + geom_line() +
    theme(legend.position = "none") + geom_point(color = "white")
```

Chick Weight over Time

Question 2.4: (2 pts)

Now, facet by diet. Is it clearer from this plot which diet is the best? (No real right answer: Just show me you are thinking about it.)

```
ggplot(data = ChickWeight, aes(x = Time, y = weight,
    color = Chick)) + ggtitle("Chick Weight over Time") +
    xlab("Time (days)") + ylab("Weight (gm)") + geom_line() +
    theme(legend.position = "none") + geom_point(color = "white") +
    facet_wrap("Diet")
```

Chick Weight over Time
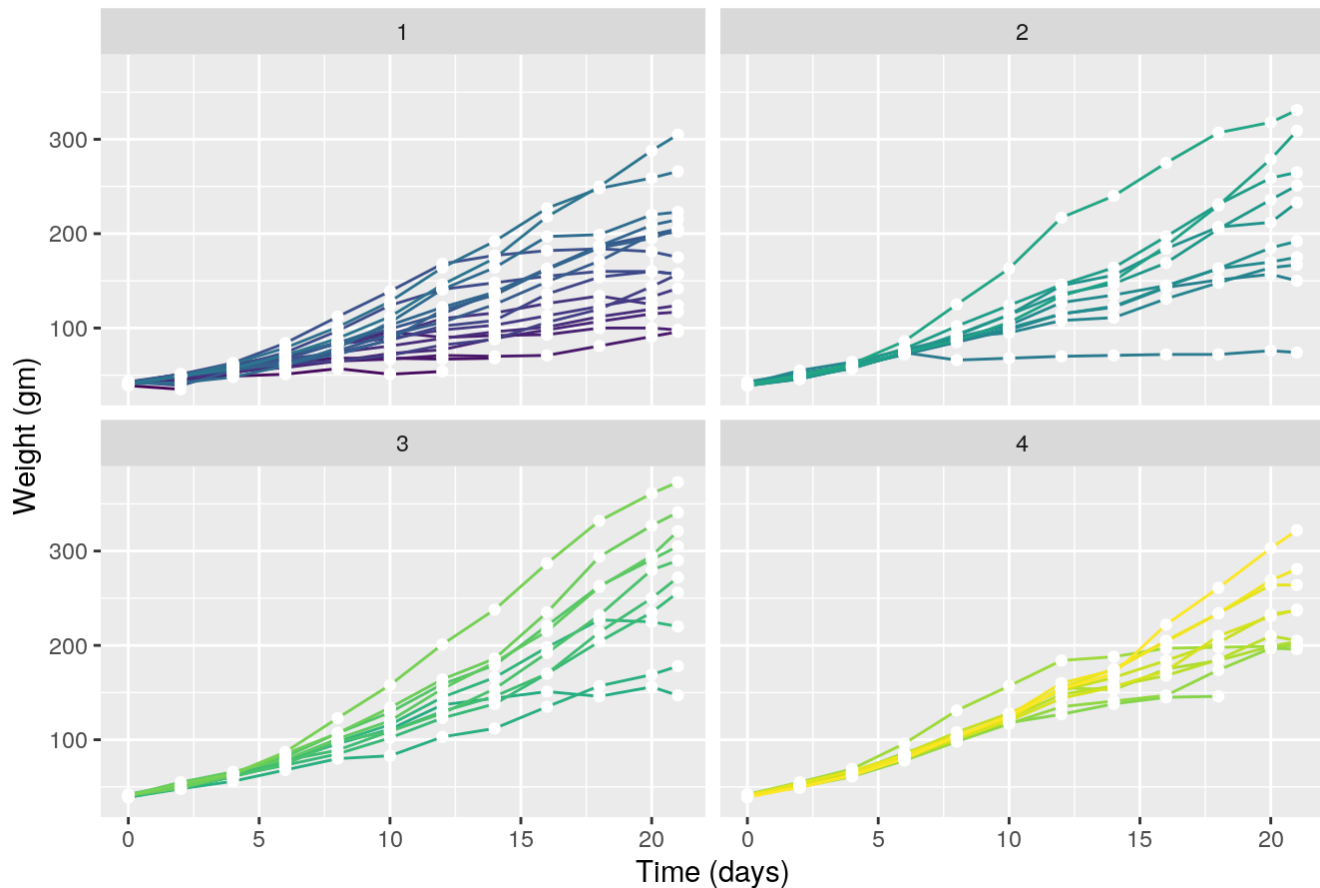
From this plot it can be inferred that diet 4 is the "best" as it appears to have the least variability per chick. This means that all chicks on this diet tend to have the same pattern for/most consistent weights per time, so it is better and more reliable than the others because of this.

Question 2.5: (1 pts)

Go back to your plot from question 2.1 and fit a *linear regression line* (not the default *loess*!) to the chicks in each diet with `geom_smooth(method = "lm")`. There should be 4 separate straight lines, one for each diet, each a separate color. **Now can you see more clearly which diet results in greater weight? Does the effect of diet on weight appear to depend on time?**

```
ggplot(data = ChickWeight, aes(x = Time, y = weight,
    color = Diet)) + geom_point() + ggtitle("Chick Weight over Time") +
    xlab("Time (days)") + ylab("Weight (gm)") + geom_smooth(method = "lm")
```

Chick Weight over Time
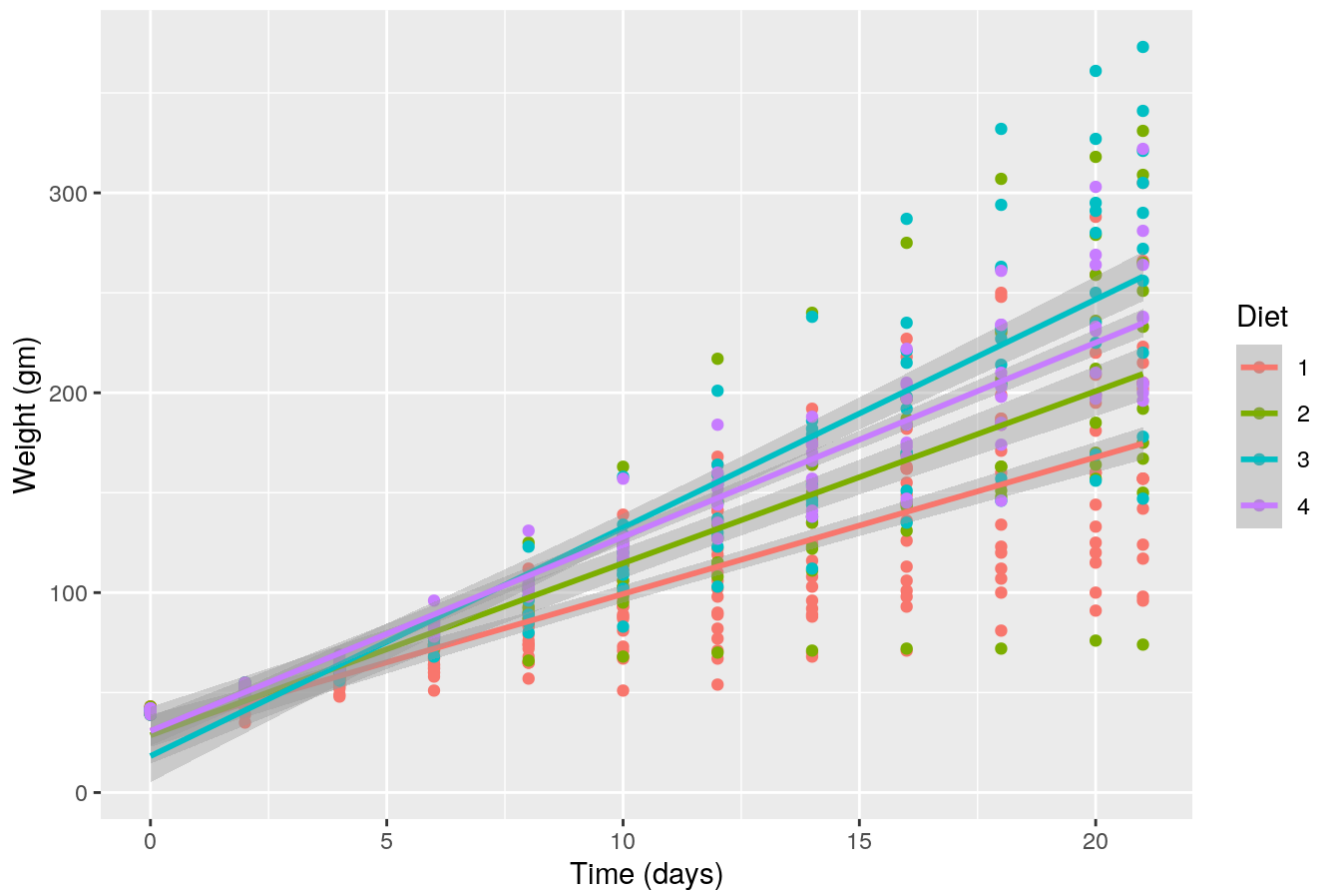
Answer: From the above graph, it can be seen that diet 3 appears to result in greater weight overall (according to the linear regression line). The graph also shows that the effect of diet on weight appears to depend on time as the average weight difference is shown to increase between diets as time goes on. In the beginning, the diets do not have much of an effect on the weight but after some time, the difference in weight is more apparent in the linear regression lines.

Question 3: (2pts)

Remake the figure from Question 2.5 with the following changes:

- Add the layer `facet_grid(~Diet)` to facet over `Diet`.
- Replace the scatterplot with a series of boxplots (one for each unique time).
- Change the theme (any `theme_` is fine as long as it is different from the default).
- The points above the whiskers on the boxplot are "outliers." Remove the outliers from your figure. (There are a few different ways to do this; one way to do so is to set `outlier.shape = NA` in `geom_boxplot()`; `?geom_boxplot` might give you some guidance.)

Hint: to have a different boxplot for each unique `Time` you can set `group = Time` in the aesthetic mapping for `geom_boxplot`. *But you don't want this grouping to be for all of the `geom`'s, just for `geom_boxplot`!!!* You can do this by feeding `mapping = aes(group = Time)` into `geom_boxplot()`.

```
ggplot(data = ChickWeight, aes(x = Time, y = weight,
    color = Diet)) + geom_boxplot(mapping = aes(group = Time),
    outlier.shape = NA) + ggtitle("Chick Weight over Time") +
    xlab("Time (days)") + ylab("Weight (gm)") + geom_smooth(method = "lm") +
    facet_grid(~Diet) + theme_bw()
```


Chick Weight over Time
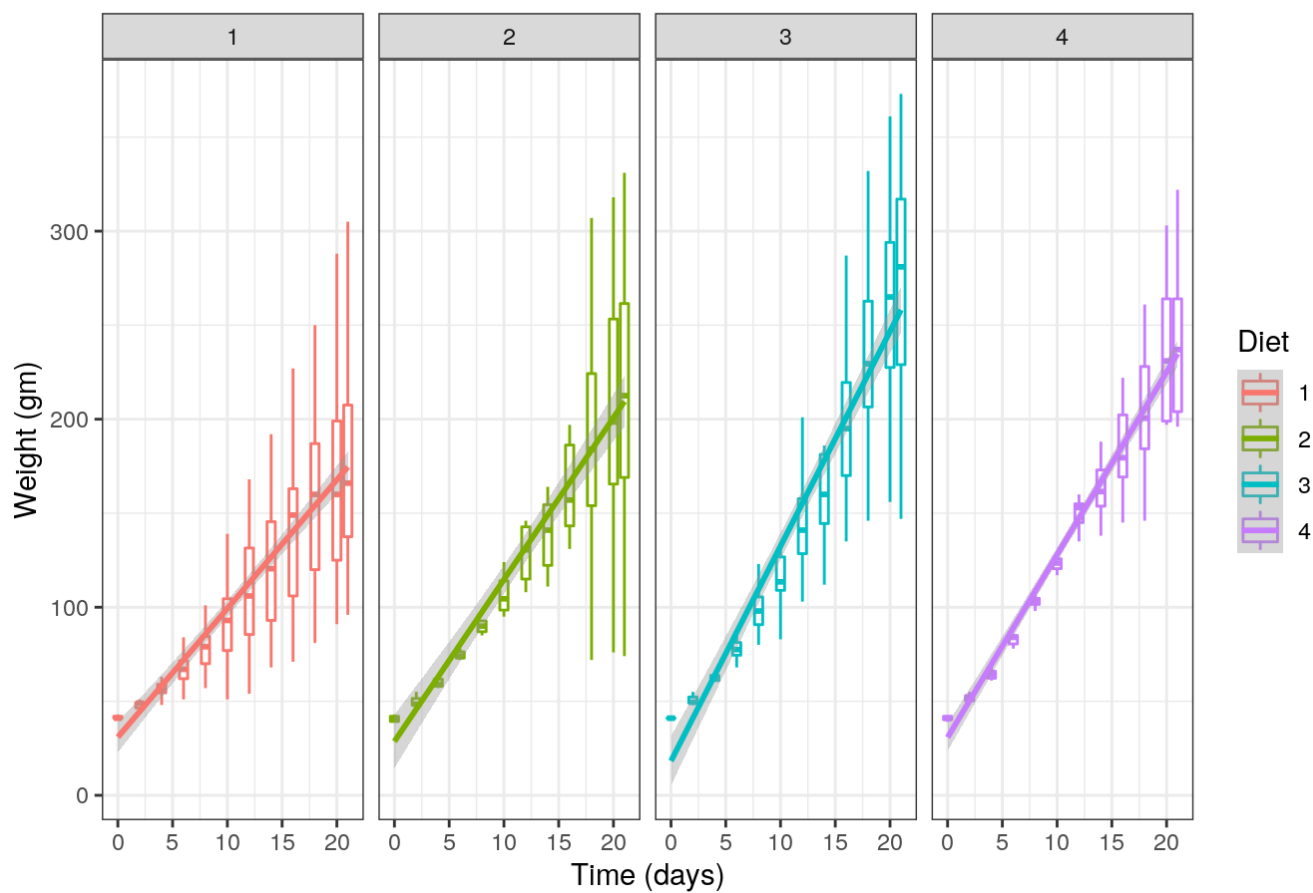
```
## R version 4.0.3 (2020-10-10)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.6 LTS
##
## Matrix products: default
## BLAS:    /stor/system/opt/R/R-4.0.3/lib/R/lib/libRblas.so
## LAPACK: /stor/system/opt/R/R-4.0.3/lib/R/lib/libRlapack.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8        LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8         LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8     LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8        LC_NAME=C
##  [9] LC_ADDRESS=C                LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats      graphics  grDevices utils     datasets  methods    base
##
## other attached packages:
## [1] forcats_0.5.1    stringr_1.4.0    dplyr_1.0.9      purrr_0.3.4
## [5] readr_2.1.2      tidyr_1.2.0      tibble_3.1.8     ggplot2_3.3.6
## [9] tidyverse_1.3.2
##
## loaded via a namespace (and not attached):
##  [1] lattice_0.20-45     lubridate_1.8.0     assertthat_0.2.1
##  [4] digest_0.6.29       utf8_1.2.2          R6_2.5.1
##  [7] cellranger_1.1.0    backports_1.4.1     reprex_2.0.1
## [10] evaluate_0.15       highr_0.9           httr_1.4.3
## [13] pillar_1.8.0        rlang_1.0.4         googlesheets4_1.0.0
## [16] readxl_1.4.0        rstudioapi_0.13     jquerylib_0.1.4
## [19] Matrix_1.4-1        rmarkdown_2.14      splines_4.0.3
## [22] labeling_0.4.2      googledrive_2.0.0   munsell_0.5.0
## [25] broom_1.0.0         compiler_4.0.3      modelr_0.1.8
## [28] xfun_0.31           pkgconfig_2.0.3     mgcv_1.8-40
## [31] htmltools_0.5.3     tidyselect_1.1.2    viridisLite_0.4.0
## [34] fansi_1.0.3         crayon_1.5.1        tzdb_0.3.0
## [37] dbplyr_2.2.1        withr_2.5.0         grid_4.0.3
## [40] nlme_3.1-158        jsonlite_1.8.0      gtable_0.3.0
## [43] lifecycle_1.0.1     DBI_1.1.3           magrittr_2.0.3
## [46] formatR_1.12        scales_1.2.0        cli_3.3.0
## [49] stringi_1.7.8       cachem_1.0.6        farver_2.1.1
## [52] fs_1.5.2            xml2_1.3.3          bslib_0.4.0
## [55] ellipsis_0.3.2      generics_0.1.3      vctrs_0.4.1
## [58] tools_4.0.3         glue_1.6.2          hms_1.1.1
## [61] fastmap_1.1.0       yaml_2.3.5          colorspace_2.0-3
## [64] gargle_1.2.0        rvest_1.0.2         knitr_1.39
## [67] haven_2.5.0         sass_0.4.2
```

```
## [1] "2022-09-14 14:58:23 CDT"
```

```
##                             sysname
##                              "Linux"
##                             release
##                   "4.15.0-191-generic"
##                             version
## "#202-Ubuntu SMP Thu Aug 4 01:49:29 UTC 2022"
##                             nodename
##            "educcomp04.ccbb.utexas.edu"
##                             machine
##                            "x86_64"
##                               login
##                           "unknown"
##                                user
##                           "ehc586"
##                       effective_user
##                           "ehc586"
```