193.301 Statistics project 2018

Introduction

A recent study at Massey investigated the effect of feeding and body condition during late pregnancy and lactation on lamb performance until weaning.

For this assignment we'll be looking at a dataset on the weaning weight of twin lambs.

On day 141 of pregnancy, ewes body condition score was measured, and the ewes were randomly allocated to 'Low', 'Medium' or 'High' feeding treatments until weaning at day 79 of lactation. Of interest is whether there is a difference in the average weaning weight of their lambs between the treatments, and whether this is dependent on body condition score or other factors.

The data set consists of the following variables from 540 twin lambs from 270 ewes.

Variable	Description
EweTag	Identifying number for the ewe.
EweFeed	The feeding treatment for the ewe (Low, Medium, High).
EweBCS	The body condition score of the ewe (BCS2, BCS2.5, BCS3).
LambTag	Identifying number for the lamb.
Paddock	The paddock (A-L) where the ewes were kept.
Sex	The sex of the lamb (Ewe, Ram).
WeaningWeight	The weight of the lamb at weaning (in kg).
WeaningAge	Age of lamb at weaning (in days).

This markdown file contains a number of R code blocks to read in the data, do some exploratory analysis and then do some inference using a couple of potential linear models.

Expectations

- Your task is to **alter this document** to choose and clean plots, answer questions and add comments.
- Make sure you can 'Knit' this document to an HTML or Word file **before** you start altering it. This ensures you have everything needed setup on your computer.
- You should clearly answer all questions by typing your answer into this document below each question.
- You're welcome to remove the questions themselves once answered, but please leave all formatting and code blocks in place (e.g. the `### Question 2` lines).
- Once complete, you should check that you can 'Knit' your R notebook to an HTML or Word document to make sure there are no mistakes with code.
- You must submit **this file** (`project.Rmd`, **not** the HTML or Word document) to stream.
- You are welcome to discuss work with other students, but **all work submitted must be your own**.

Marking

- There are a total of 65 marks available.
- Each of the 12 questions are worth 5 marks.
- An additional 5 marks are available for demonstrating excellence.

Exploratory analysis

Question 1

This first code block loads some libraries we'll be using, reads the data in, and runs a summary

```
library(ggplot2); theme_set(theme_bw())
library(visreg)
weaning_all = read.csv("http://www.massey.ac.nz/~jcmarsha/193301/data/weaning.csv")
weaning_all$EweFeed = factor(weaning_all$EweFeed, levels=c("Low", "Medium", "High"))
summary(weaning_all)
```

```
##      EweTag        EweFeed       EweBCS       LambTag         Paddock
##  Min.   :  1.0   Low   :180   BCS2  :171   Min.   :  1.0   L      : 66
##  1st Qu.: 73.0   Medium:185   BCS2.5:186   1st Qu.:140.2   C      : 52
##  Median :150.0   High  :174   BCS3  :182   Median :283.5   G      : 52
##  Mean   :148.7                             Mean   :290.4   B      : 50
##  3rd Qu.:225.0                             3rd Qu.:437.8   H      : 44
##  Max.   :297.0                             Max.   :600.0   J      : 43
##                                            NA's   :1       (Other):232
##   Sex      WeaningWeight     WeaningAge
##  Ewe:271   Min.   :14.00   Min.   :57.0
##  Ram:268   1st Qu.:26.50   1st Qu.:64.0
##            Median :29.00   Median :69.0
##            Mean   :28.93   Mean   :68.8
##            3rd Qu.:31.50   3rd Qu.:73.0
##            Max.   :43.00   Max.   :79.0
##            NA's   :56
```

```
weaning = na.omit(weaning_all)
```

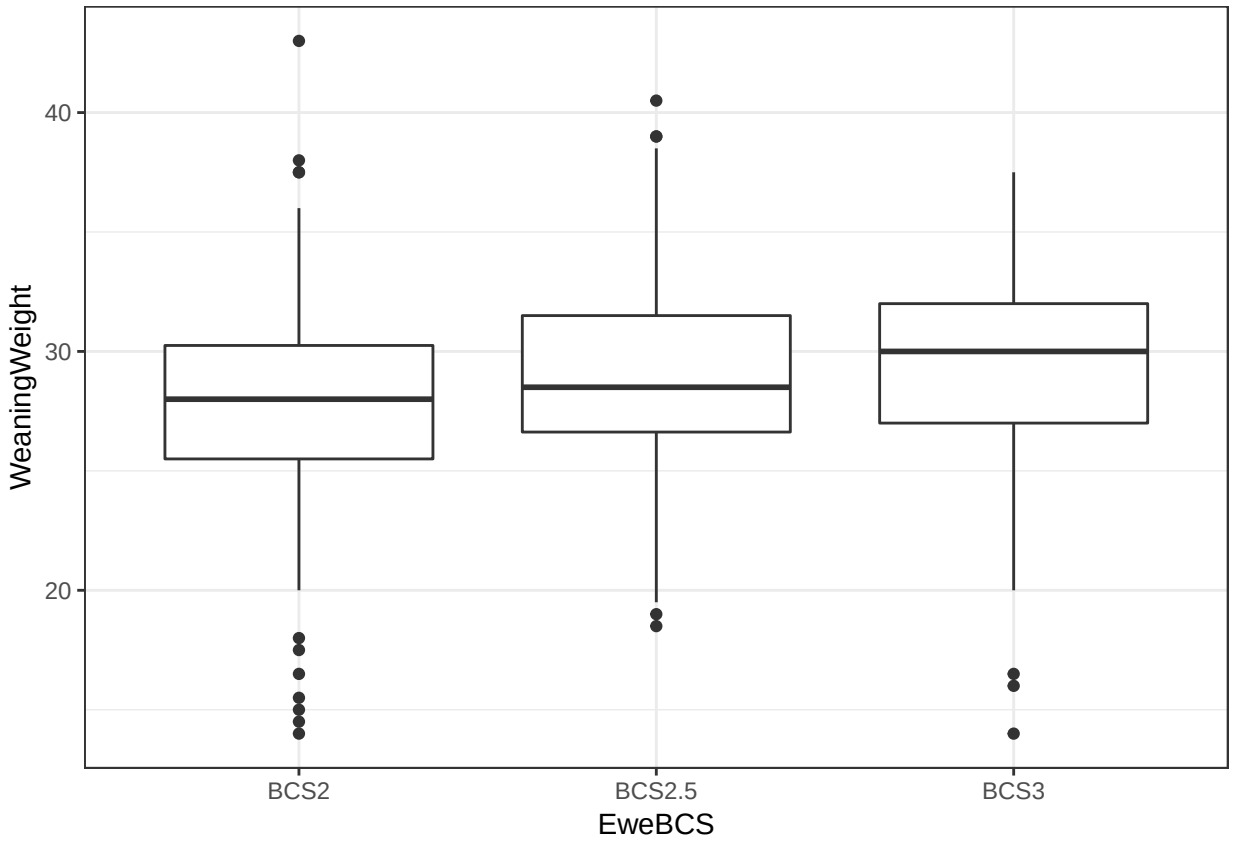
Answer the following questions:

- What is the `factor` command doing?
- Which columns contain missing values?
- Which one has the largest number of missing values? Why do you think that would be?
- What is the `na.omit` command doing?

Question 2

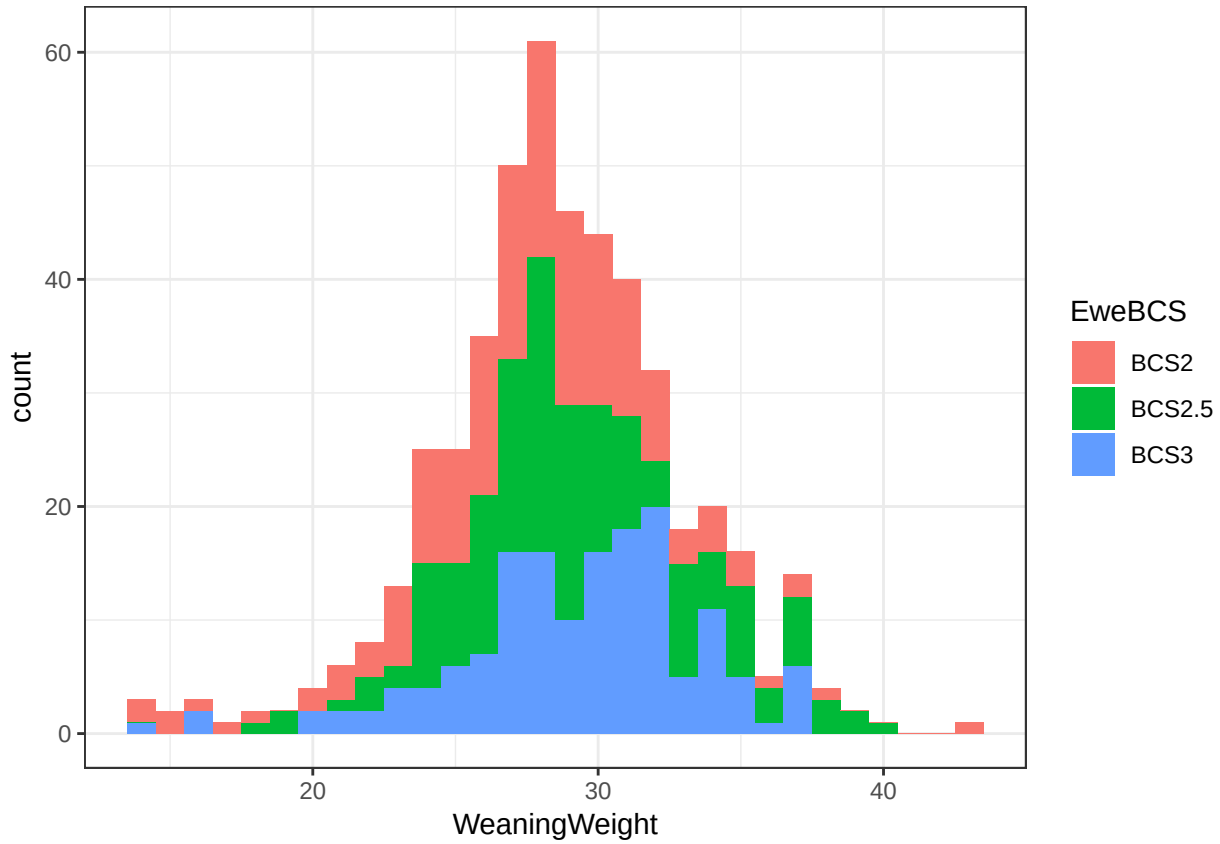
The code block below contains several plotting options for the distribution of weaning weight between the body condition scores.

```
ggplot(weaning, aes(x=EweBCS, y=WeaningWeight)) + geom_boxplot()
```

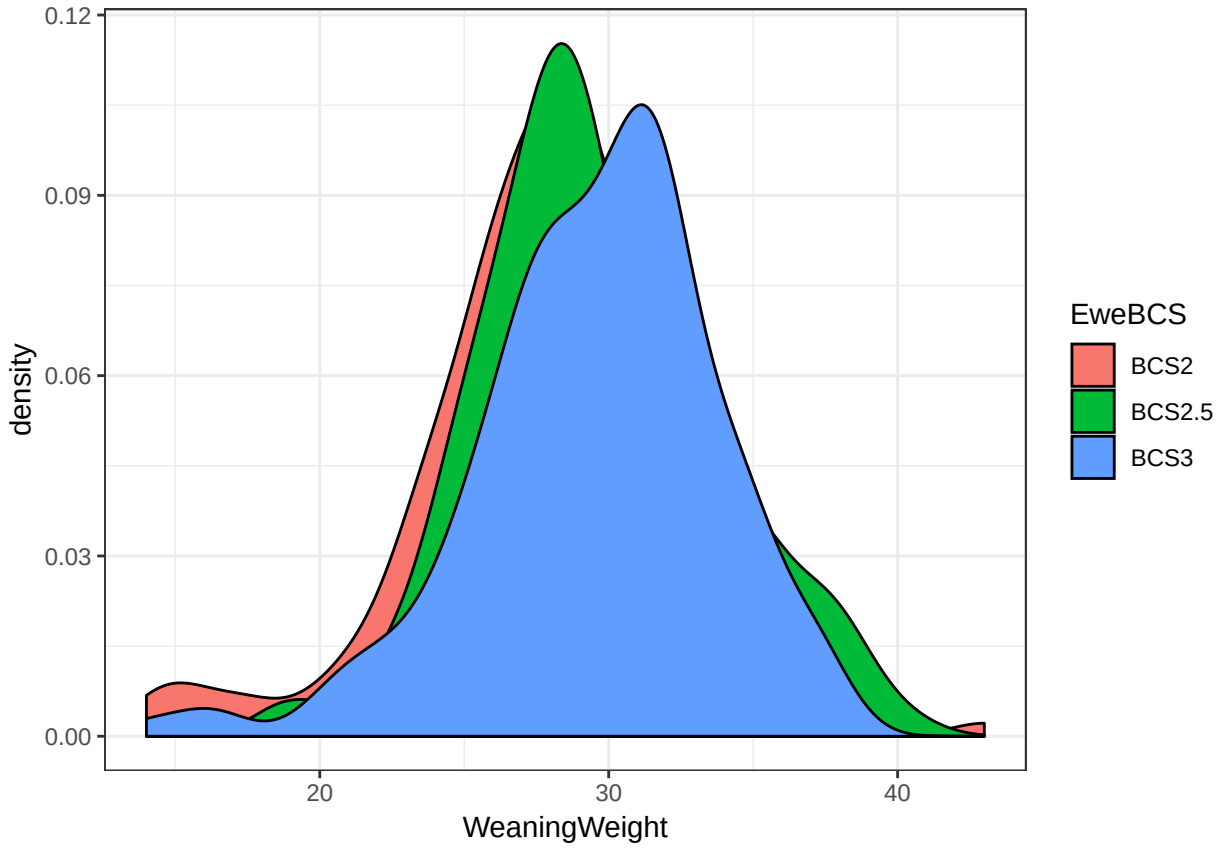



```
ggplot(weaning, aes(x=WeaningWeight, fill=EweBCS)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(weaning, aes(x=WeaningWeight, fill=EweBCS)) + geom_density()
```



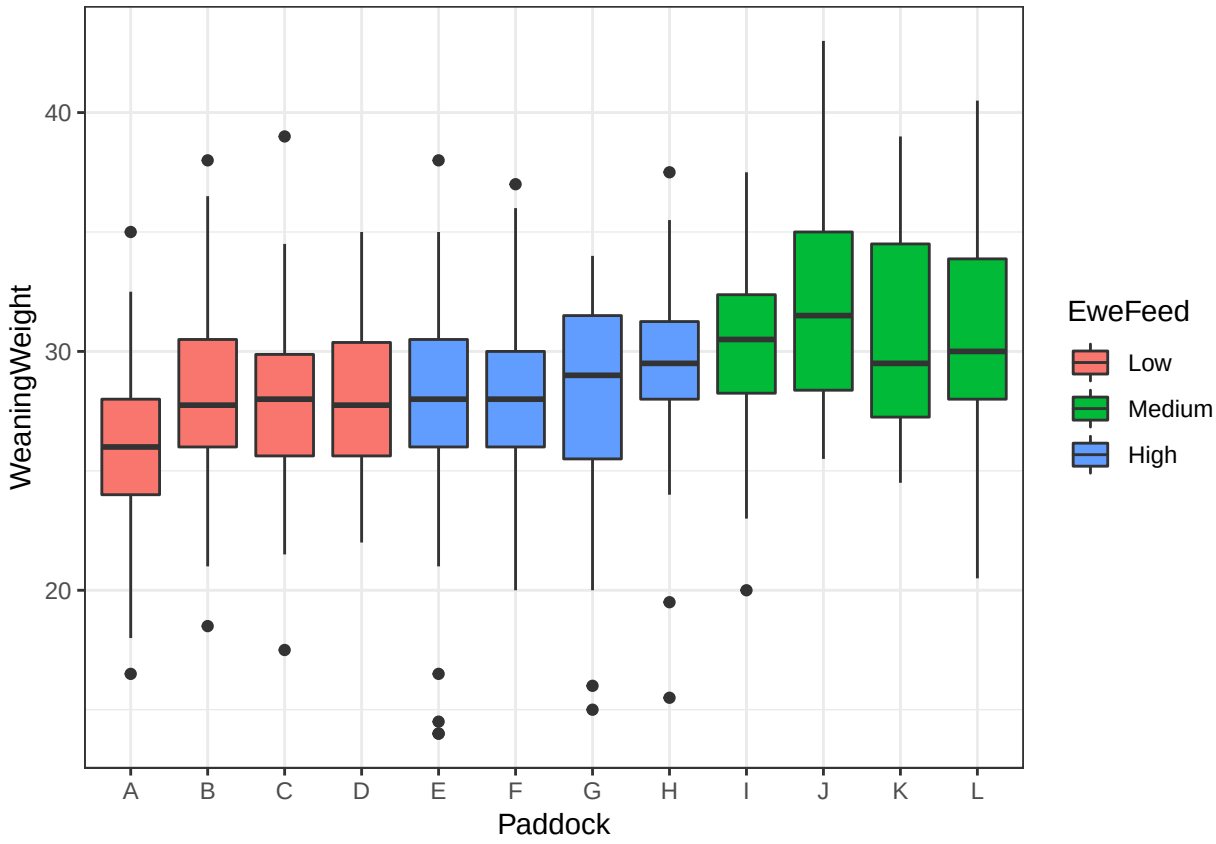
Alter the code block above to choose **one** plot that you feel best describes how the distribution of weaning weight differs between the ewe body condition.

- Make any changes to tidy up or alter the presentation before making your choice.
- You're welcome to choose an alternate plot to the above options that you think is more useful.
- Write a brief description of the distribution of weaning weights across ewe body condition scores.

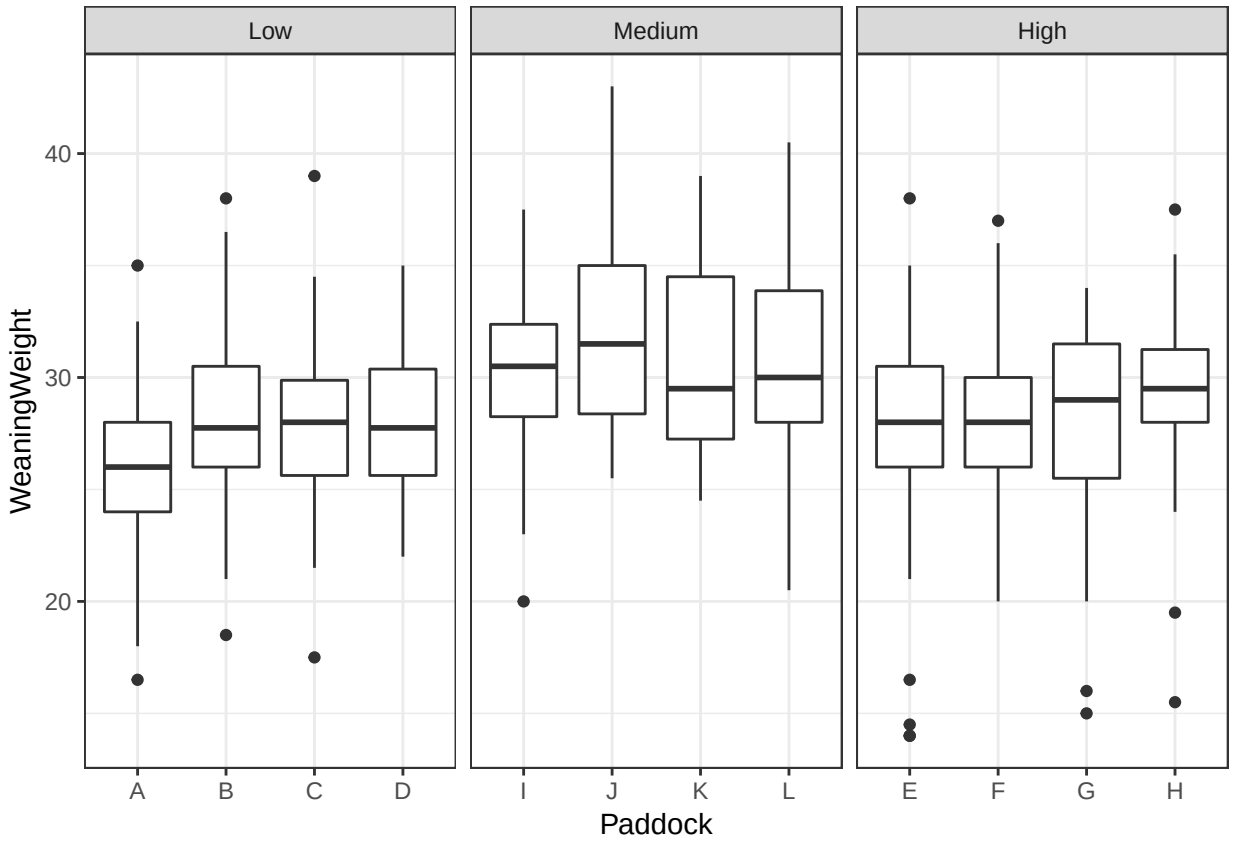
Question 3

The code block below contains several plotting options for the distribution of weaning weight between the feed treatment groups. The ewes from each feeding treatment were each kept in one of four paddocks.

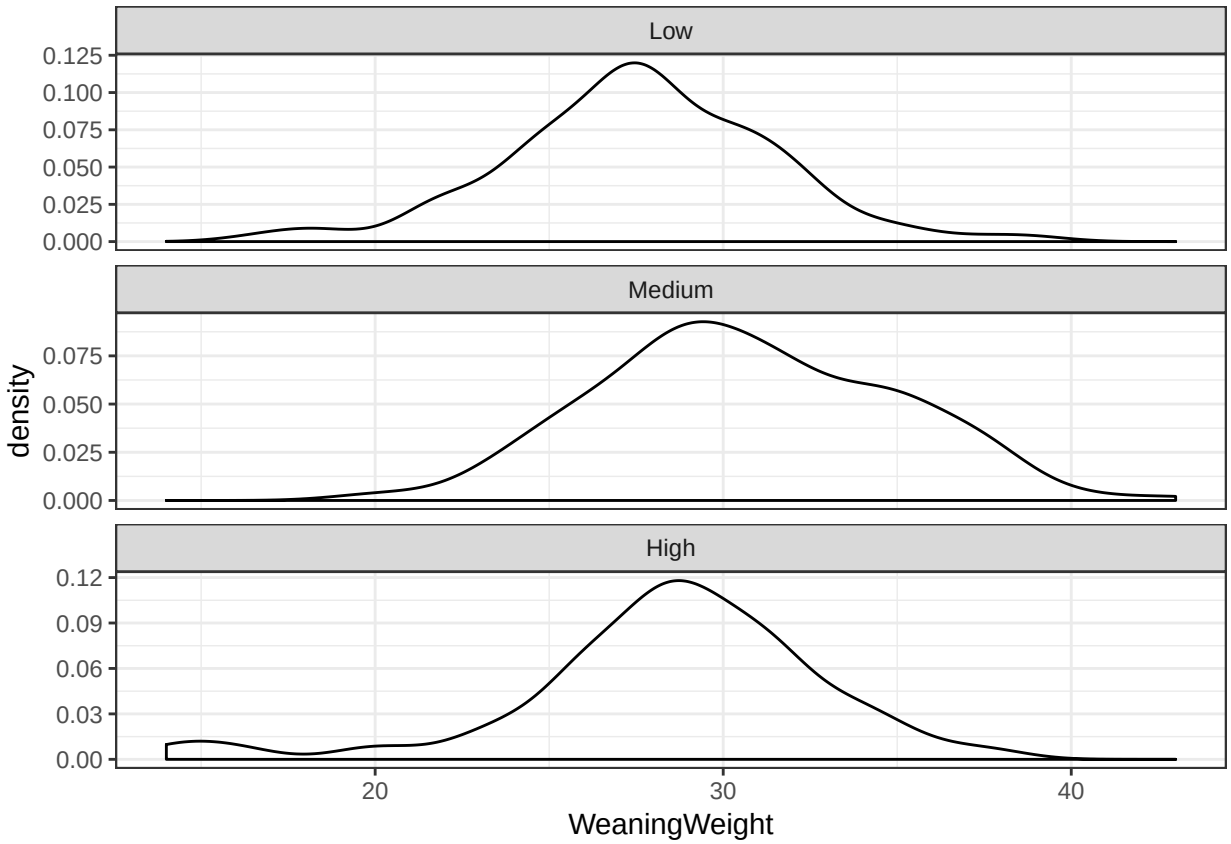
```
ggplot(weaning, aes(x=Paddock, y=WeaningWeight, fill=EweFeed)) + geom_boxplot()
```




```
ggplot(weaning, aes(x=Paddock, y=WeaningWeight)) + geom_boxplot() +
  facet_wrap(~EweFeed, scales='free_x')
```




```
ggplot(weaning, aes(x=WeaningWeight)) + geom_density() +
  facet_wrap(~EweFeed, scales='free_y', ncol=1)
```



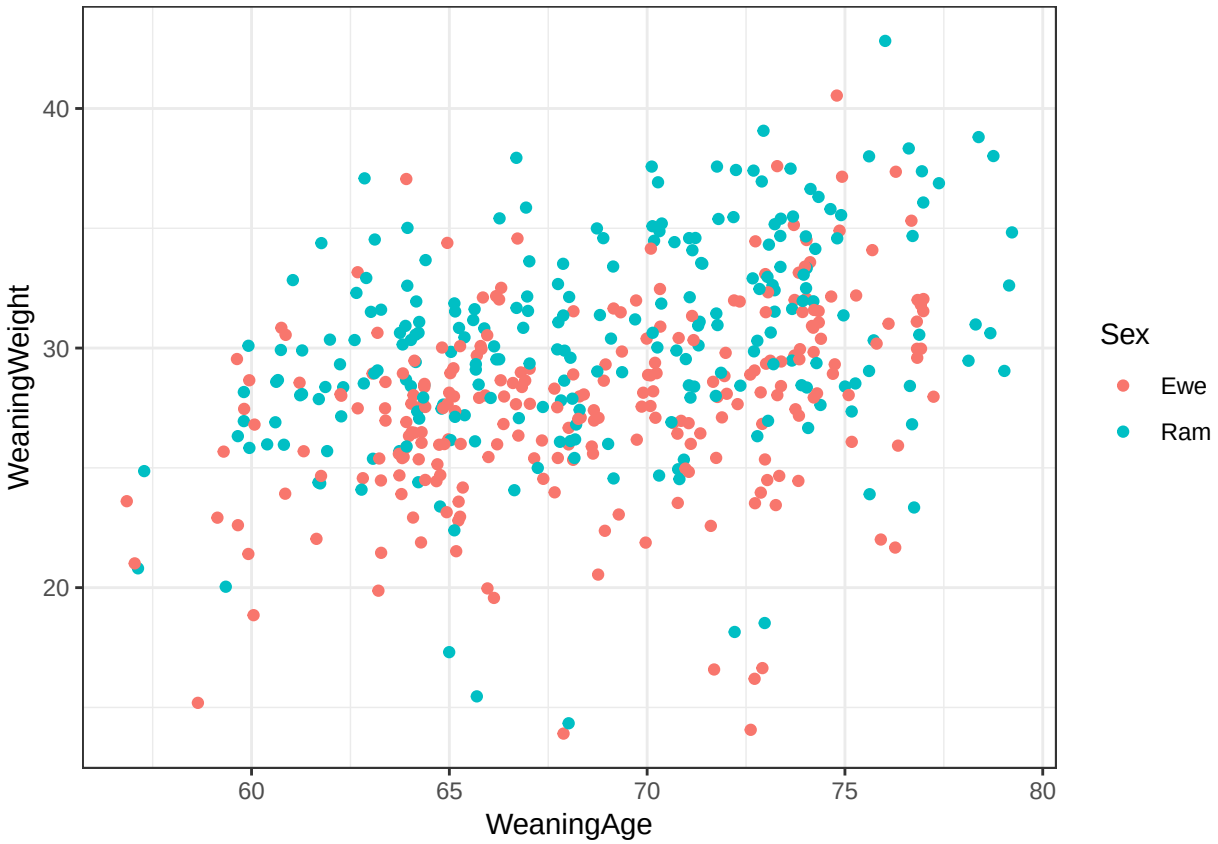
Alter the code block above to choose **one** plot that you feel best describes how the distribution of weaning weight differs between the feed treatments.

- Make any changes to tidy up or alter the presentation before making your choice.
- You're welcome to choose an alternate plot to the above options if you wish.
- Write a brief description of the distribution of weaning weights across the feeding treatments.

Question 4

The plot below shows the relationship between date of birth and weaning weight.

```
ggplot(weaning, aes(x=WeaningAge, y=WeaningWeight, col=Sex)) + geom_point(position='jitter')
```



For this question:

- Tidy the plot up (e.g. cleanup the labels) as needed.
- What is the purpose of the `position='jitter'` command and why is it used here?
- Briefly describe the relationship between weaning weight versus date of birth for ewe and ram lambs. Does the relationship make sense to you?

Linear modelling

Question 1

The code block below fits a multivariable linear model to the data, and produces an `anova` and `summary` table.

```
mod1 = lm(WeaningWeight ~ WeaningAge + Sex + EweBCS + EweFeed, data=weaning)
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: WeaningWeight
##             Df Sum Sq Mean Sq F value    Pr(>F)
## WeaningAge   1 1126.8 1126.79 85.0729 < 2.2e-16 ***
## Sex          1  765.6  765.64 57.8057 1.555e-13 ***
## EweBCS       2  229.2  114.62  8.6542 0.0002033 ***
## EweFeed      2  683.9  341.93 25.8159 2.268e-11 ***
## Residuals  476 6304.6   13.24
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mod1)
```

```
##
## Call:
## lm(formula = WeaningWeight ~ WeaningAge + Sex + EweBCS + EweFeed,
##     data = weaning)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.1850  -1.7510   0.2363   2.1440  10.6299
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.39790    2.33608   2.311 0.021278 *
## WeaningAge     0.29308    0.03372   8.691  < 2e-16 ***
## SexRam         2.33597    0.33261   7.023 7.54e-12 ***
## EweBCSBCS2.5   1.47520    0.40171   3.672 0.000268 ***
## EweBCSBCS3     1.47268    0.41281   3.567 0.000397 ***
## EweFeedMedium  2.84207    0.40432   7.029 7.24e-12 ***
## EweFeedHigh    0.91968    0.40998   2.243 0.025342 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.639 on 476 degrees of freedom
## Multiple R-squared:  0.308,  Adjusted R-squared:  0.2992
## F-statistic:  35.3 on 6 and 476 DF,  p-value: < 2.2e-16
```

Summarise your conclusions from this model:

- Which variables are important for weaning weight? What evidence do you have for each one?
- For the important variables, what are the effect sizes? What do they mean?
- Is the intercept here useful? Explain what it represents.
- Is this model useful for predicting the weight of individuals? What about predicting averages?

Question 2

The code block below adds an additional term to the above linear model.

```
mod2 = lm(WeaningWeight ~ WeaningAge + Sex + EweBCS + EweFeed + WeaningAge:EweBCS, data=weaning)
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: WeaningWeight
##                     Df Sum Sq Mean Sq F value    Pr(>F)
## WeaningAge           1 1126.8 1126.79 86.3816 < 2.2e-16 ***
## Sex                  1  765.6  765.64 58.6950 1.046e-13 ***
## EweBCS               2  229.2  114.62  8.7874  0.000179 ***
## EweFeed              2  683.9  341.93 26.2130 1.594e-11 ***
## WeaningAge:EweBCS    2  121.6   60.80  4.6612  0.009893 **
## Residuals          474 6183.0   13.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

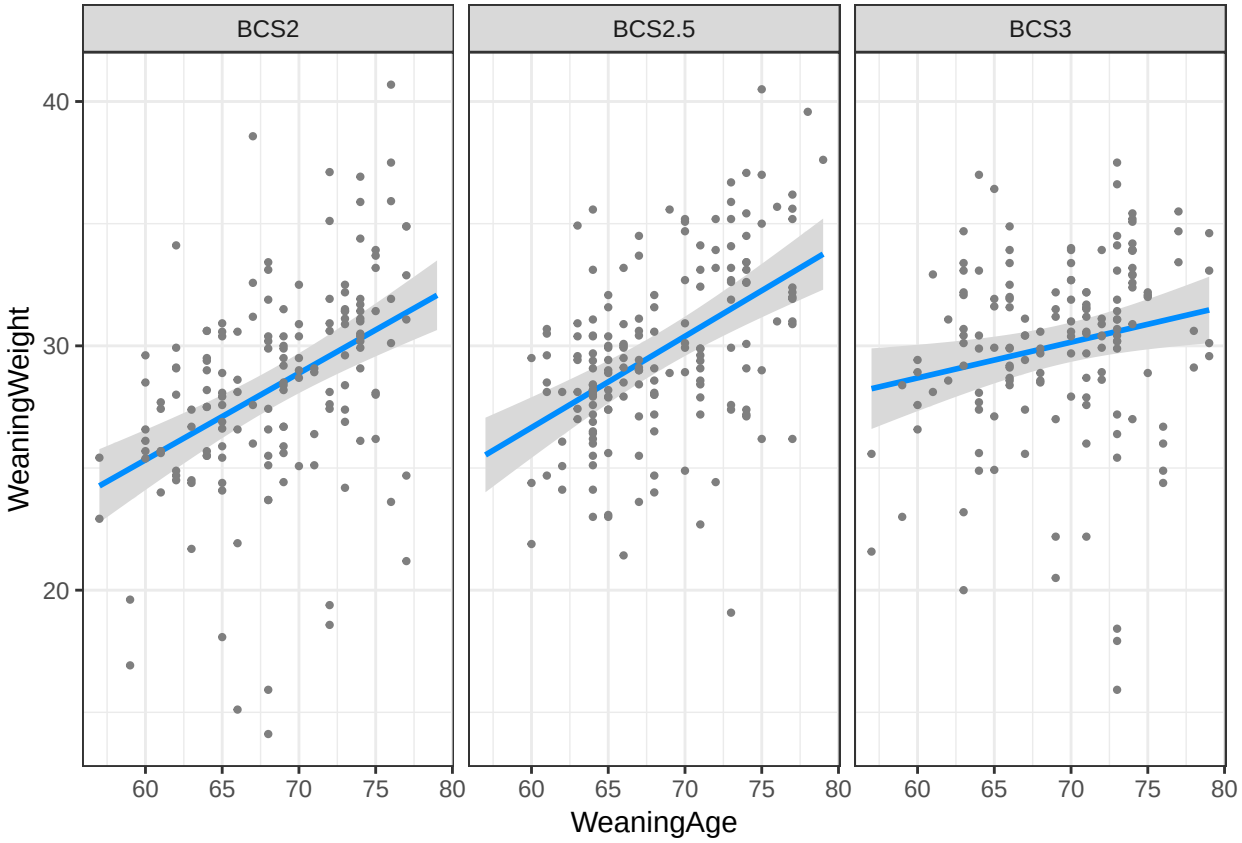
Answer the following:

- What is the purpose of the additional term in this model?
- Is the additional term important for weaning weight? What is your conclusion from this?
- The second model (`mod2`) would be better for predicting the average weaning weight of lambs. Why?

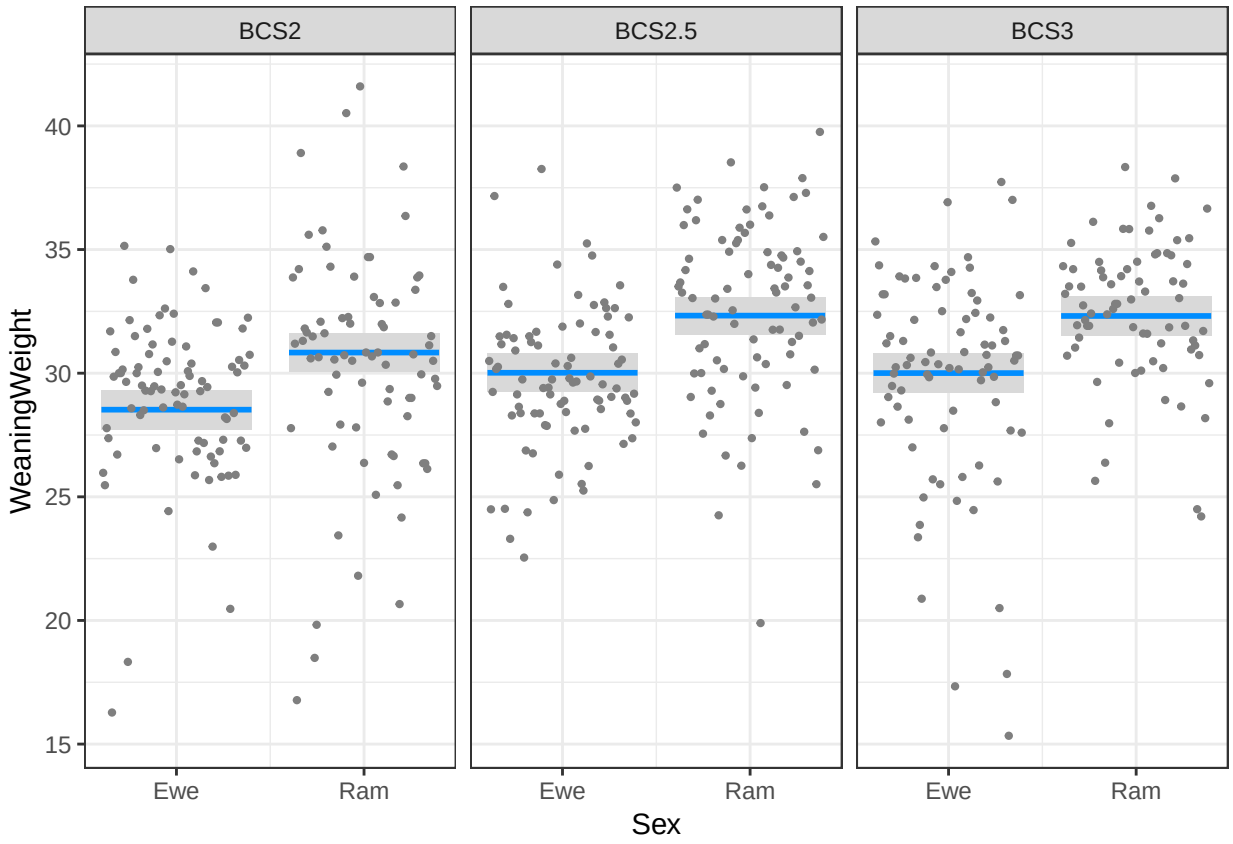
Question 3

To help evaluate the effects from this second model, the following plots were produced

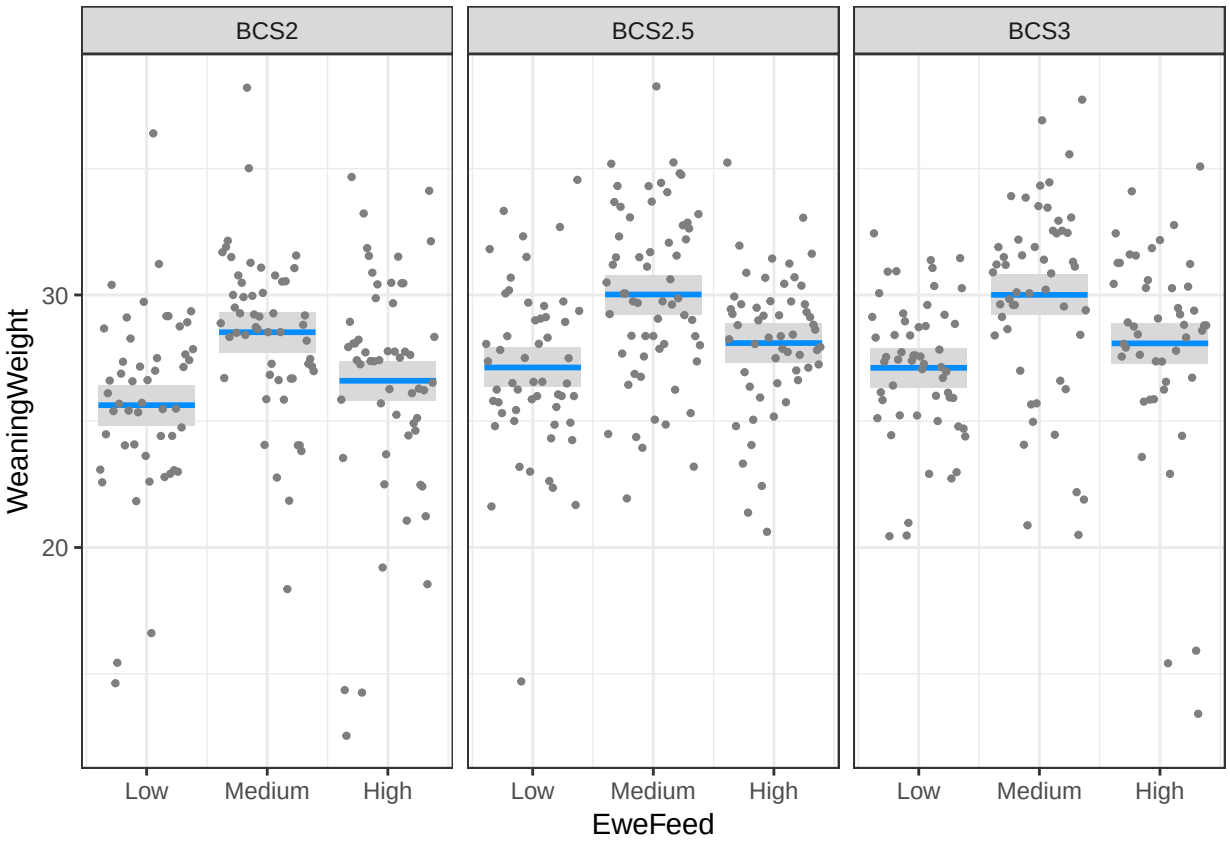
```
visreg(mod2, "WeaningAge", by="EweBCS", gg=TRUE)
```




```
visreg(mod2, "Sex", by="EweBCS", gg=TRUE)
```




```
visreg(mod2, "EweFeed", by="EweBCS", gg=TRUE)
```



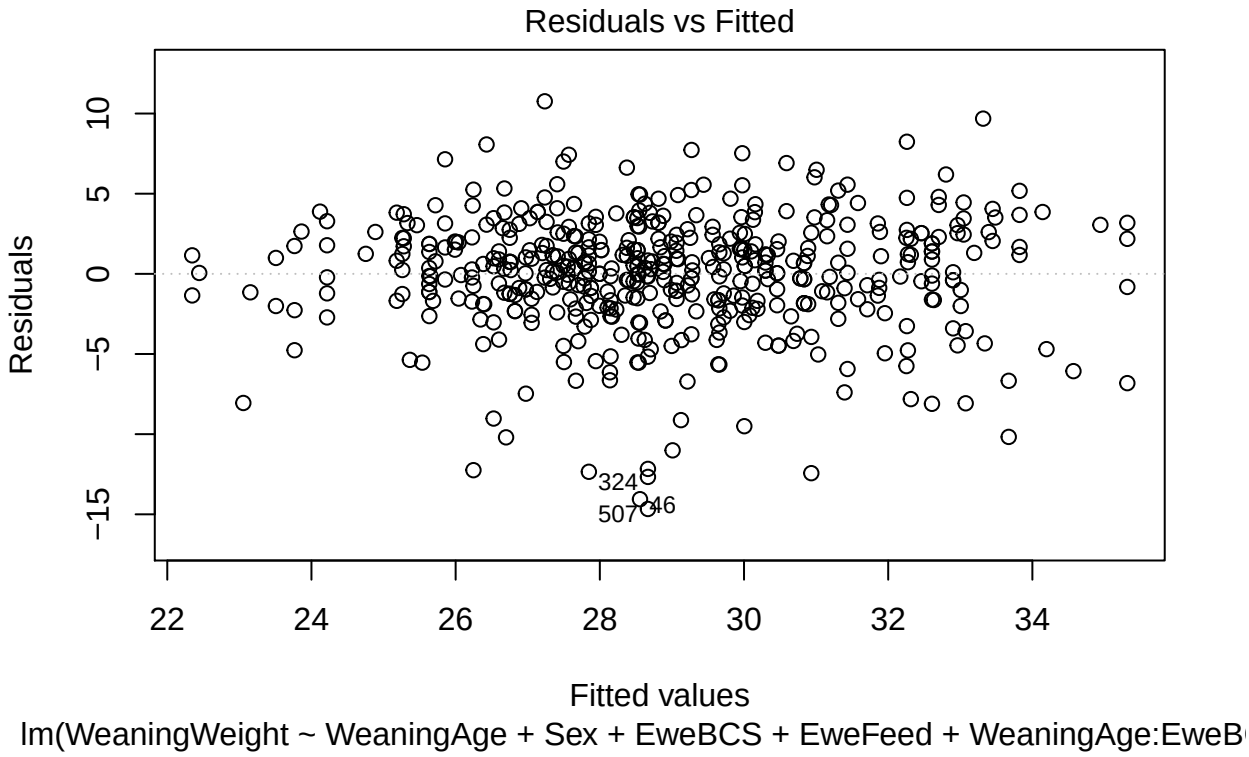
If needed, tidy up the above plots, then answer:

- What is your conclusion from each plot?
- You should notice the pattern in each of the panels of the feed and sex plots are the same. Why is this?

Question 4

The following plot was produced from the second model.

```
plot(mod2, which=1, add.smooth = FALSE)
```



Answer the following:

- What is the purpose of producing this plot?
- What is your conclusion from the plot?

Question 5

A farmer is interested in the average weight of ewe and ram lambs produced from ewes with body condition score 2.5 on a median feeding treatment that are weaned at day 75. The following code computes this.

```
new_data = data.frame(Sex = c("Ewe", "Ram"), EweBCS = "BCS2.5", EweFeed = "Medium", WeaningAge=75)
predict(mod2, new_data)
```

```
##        1        2
## 32.25932 34.56954
```

For this question:

- Alter the above so that the appropriate uncertainty interval is included.
- Briefly interpret these results in words a client may understand.

Question 6

The linear model assumption of independence is unlikely to be met. Clearly explain why this is, and what the consequences might be.