Statistics project

Brenda

Exploratory analysis

Question 1 (a)

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)
```

What is the `factor` command doing?

The factor command is sorting the 'EweFeed data' (the 3 different feeding treatments), into labelled groups/levels "Low", "Medium", & "High". By default it would be alphabetical.

Which one has the largest number of missing values? Why do you think that would be?

The 'WeaningWeight' column has the largest number of missing values (56), while the 'LambTag' column is only missing 1 value. The 56 missing values from the 'WeaningWeight' column may represent the number of lambs that died in this sample after birth, before their weaning weight was measured.

What is the `na.omit` command doing?

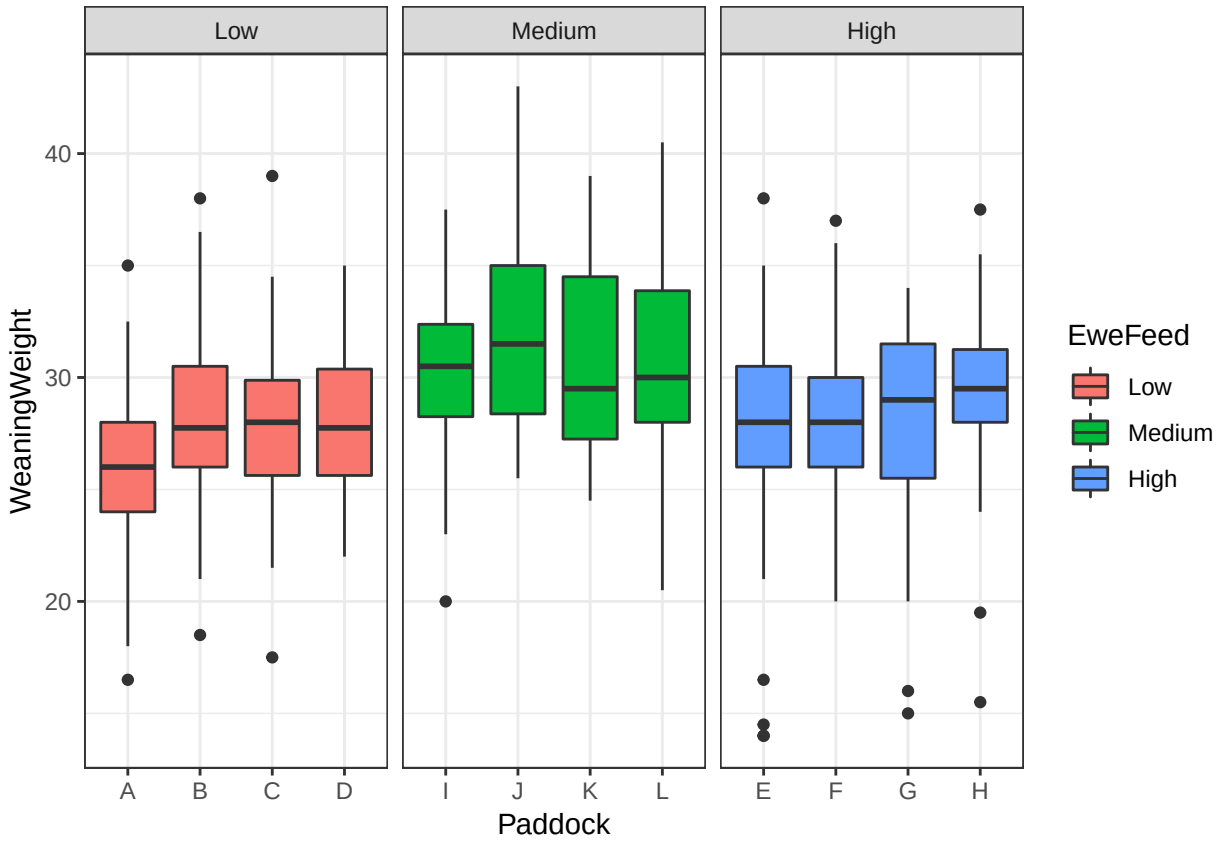
The na.omit command is removing all of the missing values from the 'weaning_all' dataset. This generates a new dataset ('weaning') that doesn't contain any missing values.

Question 1 (b)

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. 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, and write a brief description of the distribution of weaning weights across the feeding treatments.

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



Overall, the ewes with 'Low' feed treatments appear to wean lambs with lower weaning weights. The ewes with 'Medium' feed treatments appear to wean lambs with the highest weaning weights. The ewes with 'High' feed treatments appear to wean lambs with weaning weights that weren't much greater than the ewes with 'Low' feed treatments.

The plots within the 'Low' feed treatment group all have similar spread about the data. All plots appear to be reasonably symmetrical, although data from paddock D is slightey skewed to the right.

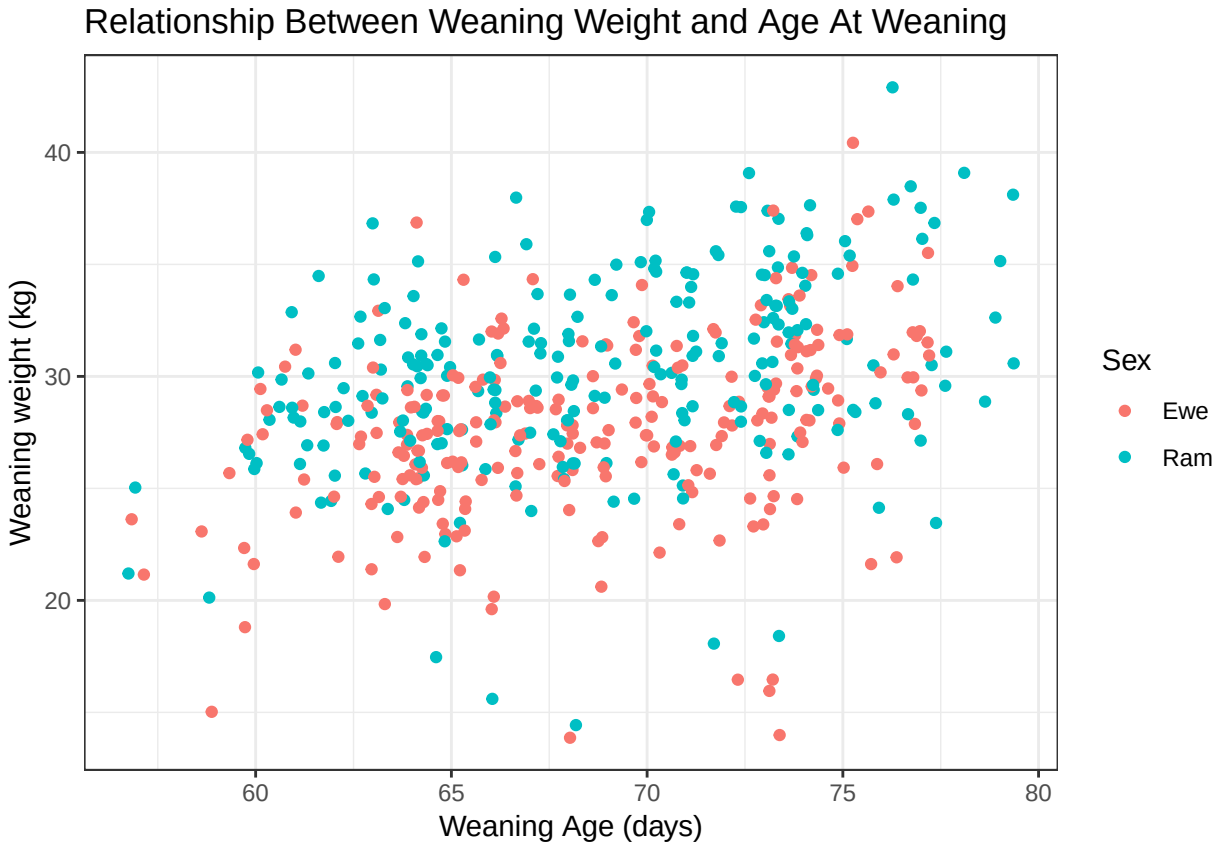
The plots within the 'Medium' feed treatment group have more variation in spread about the data. Data from paddocks I & L is reasonably symmetrical, while data from paddocks J & K is skewed to the right.

The plots within the 'High' feed treatment group have also have some variation in spread about the data. Data from paddocks E, F, & H is quite symmetrical, while data from paddock G is skewed to the left.

Question 1 (c)

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') +ggtitle("
```



What is the purpose of the `position='jitter'` command and why is it used here?

Some lambs have the same date of birth ('WeaningAge'), but we can't tell which data points these are on the scatterplot, as the points are just plotted on top of each other. By adding the 'position=jitter' command, points that are overlayed on top of each other are separated out on the scatterplot.

Briefly describe the relationship between weaning weight versus weaning age for ewe and ram lambs.

There does appear to be a positive relationship between weaning weight and weaning age, although the strength of this relationship is weak. Ram lambs tend to be heavier than ewes. This makes sense as the older the lamb is (i.e. the greater the 'WeaningAge'), the greater its 'WeaningWeight' would be expected to be. It also makes sense that ram lambs would have a greater 'WeaningWeight' than ewe lambs that are of the same 'WeaningAge'.