Statistics project

Andrew

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 represents categorical variables in R.

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

The paddock column. It is likely because ewes may have been moved around and/or accidentally mixed with one another.

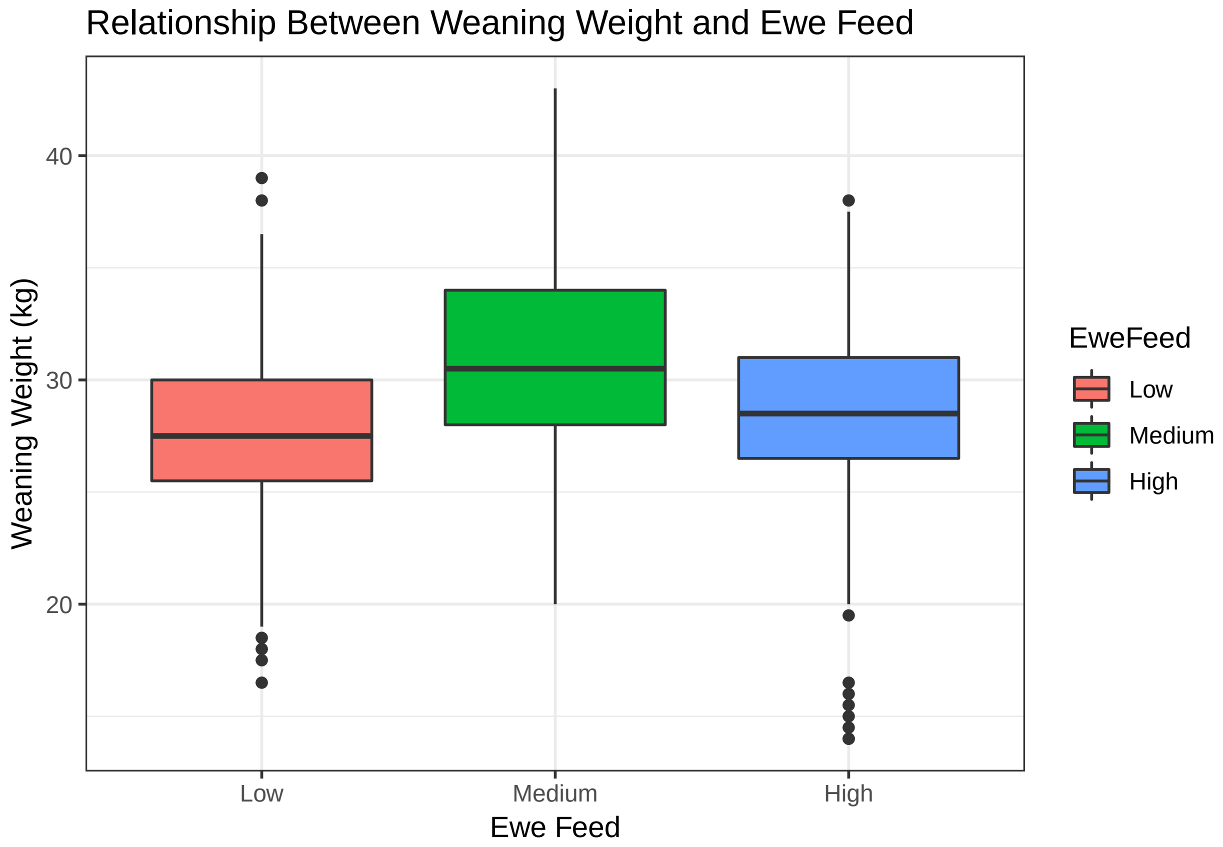
What is the `na.omit` command doing?

This command removes rows which have missing values. In this data set there are missing values for weaning weight and lamb tags so these rows of data were removed.

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=EweFeed, y=WeaningWeight, fill=EweFeed)) + geom_boxplot() + ggtitle("Relationship
```

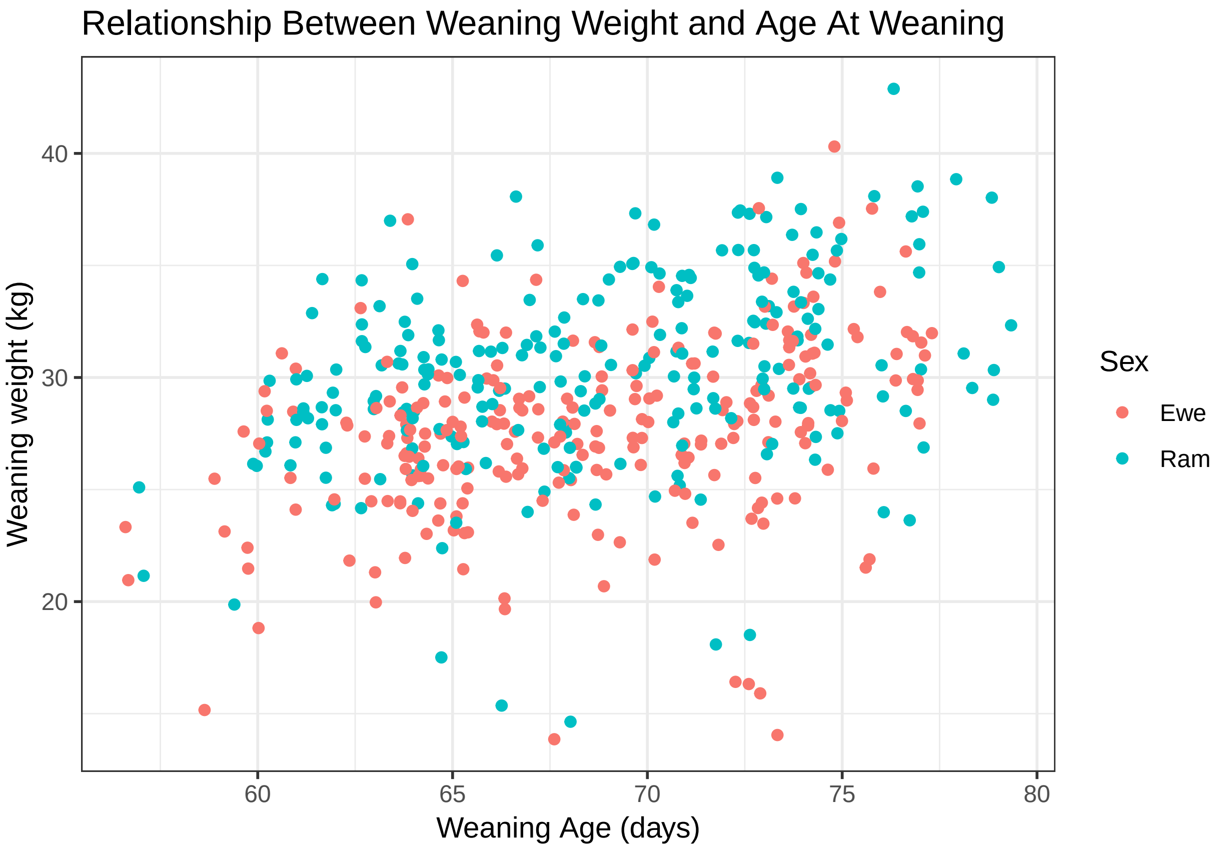


Overall, all feeding treatment groups are pretty symetrical with no noticeable skews. Ewes with high feeding treatment had a larger proprtion of outliers and the medium no outliers. The centre of the low and high feeding ewes were similar to eachother indicating similar weaning weights. The medium group apperaed to have a higher median than the other two.

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?

Jitter adds a small amount of noise to a numeric vector. It is used here to show a more apparent trend in the data. If jitter is not used then the data points look like one and there appears to be no trend when there may be one.

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

There is a weak positive relationship for ewe and ram lambs. This makes sense as the older they become in age before being weaned, the heavier they should become as they have more time to grow.