STAT201 Assignment 4

ANOVA

Due by 3pm on Friday 1 April 2022

Your assignment is the answers to the questions below. You will need to copy the relevant output and graphs from R and put them into a Word document (or another editor). You can copy your graphs in R using export, and copy to clipboard.

Write your assignment so it can be read easily. You need to include your R code in your assignment, but do not put it in the main part of your assignment. Instead put your R code at the end as an appendix. You can save your R code from the script screen.

Submit your completed assignment by uploading it on the Learn webpage where you downloaded this question sheet. It must be uploaded before the due date.

You can upload more than one version of your assignment, and the most recent version is the one that will be marked.

Question 1

The dataset, animals.csv, has the weight of farm animals that were fed four different feed stock, A. B, C and D.

a. Download the data file, animals.csv, from Learn into a folder, and then, in R set the working directory to this. Import the dataset into R. Download the data file. Check the data has read in correctly by printing out the top 6 rows using `head()`, the bottom 6 rows using `tail()`.

b. Use a suitable graph to look at differences in animal weights among the feed stock, and write a sentence about the information displayed in the graph:

```
boxplot(Weight~Feed, data = animals, pch = 19)
```

c. Use analysis of variance to investigate if there are statistically significant differences amongst the feed stock:

```
animals.lm1<-lm(Weight~Feed, data = animals)
```

Print out your model with the summary `summary(animals.lm1)` and the anova `anova(animals.lm1)` option in R. Is there evidence of differences among feed stocks?

d. Look at the three residual plots, as in the previous question, for your model and comment on each.

e. Use Tukey's honest significant difference (HSD) test to help you explain in a few sentences about the differences in average animal weight with the feed stock. In R, the code for TukeyHSD needs a model that has been made with the code, aov rather than lm. Refit your final model using aov but check that it gives you the same model as with lm.

```
animals.lm2<-aov(Weight~Feed, data = animals)
anova(animals.lm1)
anova(animals.lm2)
TukeyHSD(animals.lm2)
```

Are all four feed stock different to each other or are two similar to each other? Which feed stock would you recommend?

Question 2

The dataset, health.csv, contains information from a small study of patients who were treated for an illness. Patients either received treatment A, or B, or C. They participated from two health facilities, named North and South, in equal numbers. Patients gave consent to participate but were not told which treatment they were being given. The time for recovery was measured, in days.

a. Download the data file, health.csv, from Learn into a folder and import it into R. Check the data has read in correctly by printing out the top 6 rows using `head()`, the bottom 6 rows using `tail()`. The variables are Treatment, Facility, and Days.

b. Explore the data with suitable graphs that shows how the treatment and health-facility are related to the time to recover. You can use this R code below. Write a few sentences about what you see in your graphs. Does how you interpret the box plots change with the different ordering?

```
interaction.plot(health$Treatment, health$Facility, health$Days)
boxplot(Days~Treatment*Facility, col=c("white", "steelblue"), data
=health)
boxplot(Days~Facility*Treatment, col=c("white", "steelblue"), data
=health)
```

c. Create a linear regression model to recovery time for patients with the different treatments from the two facilities, using the lm function in R:

```
health.lm1<-lm(Days~Treatment*Facility, data = health)
```

Use the summary and anova options in R, to print out your model. Explain why you can reduce your model to a simpler model.

d. Reduce your model to a simpler model by removing the interaction. Fit a reduced new model:

```
health.lm2<-lm(Days~Treatment+Facility, data = health)
```

e. Use the summary and anova options in R, to print out your new model. Explain, in a few sentences, what information is in the output. Look at the graphs you made earlier and comment on what you can see in these in terms of your new, simpler model.

f. Look at, and comment on, the residual plots of your simpler model.

g. Can you reduce your model ever further by removing either of the main effects (i.e. can you remove Facility or Treatment)?

h. Use Tukey's honest significant difference (HSD) test to help you explain in a few sentences more about the differences in average recovery time among treatments. Are all three treatments different to each other or are two similar to each other? Which treatment would you prefer if you had this illness?

In R, the code for TukeyHSD needs a model that has been made with the code, aov rather than lm. Refit your final model using aov:

```
health.lm3<-aov(Days~Treatment+Facility, data = health)
anova(health.lm3)
TukeyHSD(health.lm3)
```