INDIVIDUAL TASK COVER SHEET

JAMES COOK UNIVERSITY AUSTRALIA

MA5820 Statistical Methods for Data Scientists

Assessment Task	Assessment 2: Statistical Analysis
College	

Student: Please sign, date, and attach this cover sheet to the front of your assessment task for all hard copy submissions.

Student Family Name	**Student Given Name**	**JCU Student Number**							
MILLS	ADAM	1	3	8	1	3	6	1	7
Assessment Title	Assessment 2: Statistical Analysis								
Due Date	9 Aug. 2020								
Lecturer Name	Yvette Everingham								
Tutor Name									

Student Declaration

1. This assignment is my original work and no part has been copied/ reproduced from any other person's work or from any other source, except where acknowledgement has been made (see *Learning, Teaching and Assessment Policy 5.1*).
2. This work has not been submitted previously for assessment and received a grade OR concurrently for assessment, either in whole or part, for this subject (unless part of integrated assessment design/approved by the Subject Coordinator), any other subject or any other course (see *Learning, Teaching and Assessment Policy 5.9*).
3. This assignment has not been written for me.
4. We hold a copy of this assignment and can produce a copy if requested.
5. This work may be used for the purposes of moderation and identifying plagiarism.
6. We give permission for a copy of this marked assignment to be retained by the College for benchmarking and course review and accreditation purposes.

Learning, Teaching and Assessment Policy 5.1. A student who submits work containing plagiarised material for assessment will be subject to the provisions of the Student Academic Misconduct Requirements Policy.

Note the definition of plagiarism and self plagiarism in the Learning, Teaching and Assessment Policy:

Plagiarism: reproduction without acknowledgement of another person's words, work or expressed thoughts from any source. The definition of words, works and thoughts includes such representations as diagrams, drawings, sketches, pictures, objects, text, lecture hand-outs, artistic works and other such expressions of ideas, but hereafter the term 'work' is used to embrace all of these. Plagiarism comprises not only direct copying of aspects of another person's work but also the reproduction, even if slightly rewritten or adapted, of someone else's ideas. In both cases, someone else's work is presented as the student's own. Under the Australian *Copyright Act 1968* a copyright owner can take legal action in the courts against a party who has infringed their copyright.

Self Plagiarism: the use of one's own previously assessed material being resubmitted without acknowledgement or citing of the original.

Student Signature	Amills	Submission Date: 09-Aug-2020

Statistical Methods for Data Scientists
Assessment-2 [STATISTICAL ANALYSIS]

Author: A MILLS

Question 1:

(a) Find the probability that a single randomly chosen TV will last less than 4500 hours. Use R to assist with your computations.

```
pnorm(4500, mean = 4800, sd = 400)

0.2266274
```

(b) Find the probability that the mean lifetime of a random sample of 16 TVs is less than 4500 hours. Use R to assist with your computations

```
pnorm(4500, mean = 4800, sd = 400)

0.2266274
```

(c) Compare answers from (a) and (b)

Both answers are the identical because no matter the sample size the distribution remains normal and the because the mean and the standard deviation are the same for both questions the probability remains the same regardless if its 100 tv's or 16 tv's

Question 2:

Beta endorphins are morphine like substances produced by the body. They create a sense of wellbeing. It has been proposed that Beta endorphins increase with exercise. Test this hypothesis using the data in beta.csv which has Beta endorphin levels for 10 people measured for each person pre and post exercise. Using this sample, test if Beta endorphins increase with exercise. Adopt a 5% risk of committing a type I error.

1. Enter the data into R

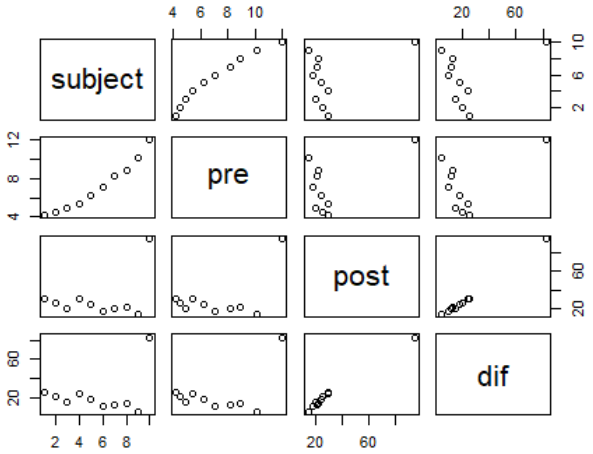
```r
library(readr)
beta <- read_csv("beta.csv")
```

2. Perform some exploratory data analysis procedures

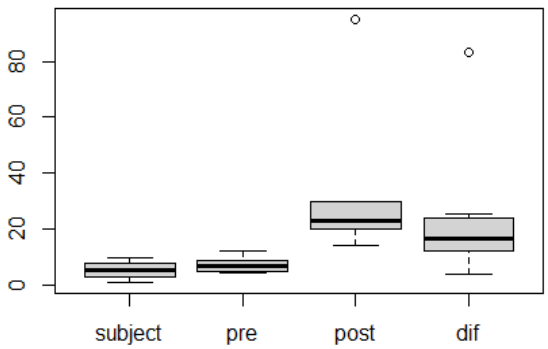
```r
# Output summary of descriptive statistics
describe(beta)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis
subject	1	10	5.5	3.03	5.5	5.5	3.71	1	10	9	0	-1.56
pre	2	10	7.15	2.61	6.65	6.91	2.89	4.2	12	7.8	0.48	-1.27
post	3	10	29.68	23.48	23	23.45	6.6	14.2	95	80.8	2.09	3.02
dif	4	10	22.53	22.23	16.55	17.27	8.3	4.1	83	78.9	1.91	2.54

```r
# Display multiplot for quicklook at relationships
pairs(beta)
```



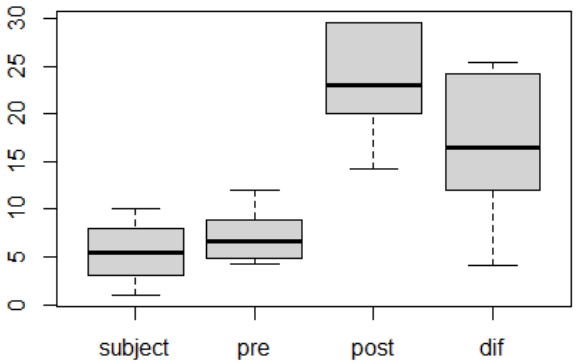

```r
# Diplay boxplot
boxplot(beta)
```



Outliers identified,

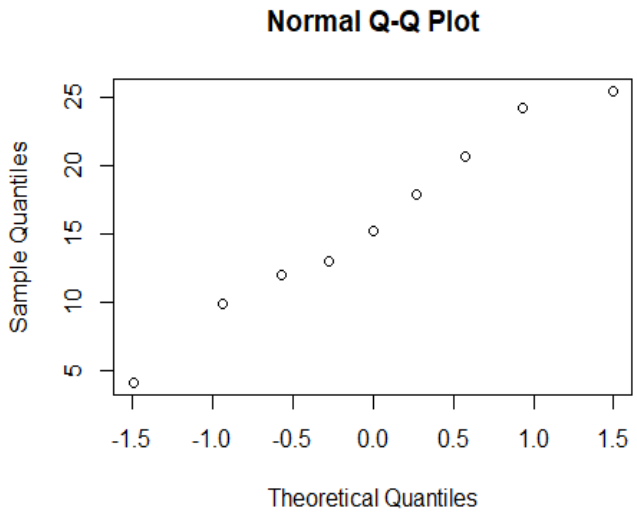
display boxplot **excluding outliers**

```
boxplot(beta, outline = FALSE)
```



Determine that row ten is probably useless. Remove row ten and save output to new name so that orginal data is preserved for comparison later.

```
beta_nooutliers <- beta[-c(10),]
beta_nooutliers

## # A tibble: 9 x 4
##   subject   pre  post   dif
##     <dbl> <dbl> <dbl> <dbl>
## 1       1   4.2  29.6  25.4
## 2       2   4.5  25.1  20.6
## 3       3   4.9  20.1  15.2
## 4       4   5.4  29.6  24.2
## 5       5   6.2  24.1  17.9
## 6       6   7.1  17     9.9
## 7       7   8.2  20.2  12
## 8       8   8.9  21.9  13
## 9       9  10.1  14.2   4.1

# Assess normality of dif variable after removal of outlier
qqnorm(beta_nooutliers$dif)
```



3. Perform an appropriate significance test

Testing if Beta endorphins increase with exercise. Adopting a 5% risk of committing a type I error.

Test Type: Single Sample t-test on *beta_nooutliers$diff*

Ho: The mean increase of endorphins is equal to 0 (H0:μ = 0)

HA: The mean increase of endorphins is greater than 0 (HA: μ > 0)

One tailed hypothesis. α = 0.05

```
t.test(beta_nooutliers$dif, conf.level = 0.95, alternative = "greater")

##
##  One Sample t-test
##
## data:  beta_nooutliers$dif
## t = 6.8419, df = 8, p-value = 6.604e-05
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  11.51382      Inf
## sample estimates:
## mean of x
##  15.81111
```

The resultant p-value of 6.604e-05 is smaller than the significance level of 0.05 therefore we reject the null hypothesis. There is evidence to conclude that Beta endorphins increase with exercise.

4. State any assumptions needed to support the validity of the procedure and where possible comment on the adequacy of these assumptions

- The data are a simple random sample. (This was not stated in the description)
- Observations come from a population that is normally distributed. The Q-Q plot of the sample data, shown above, indicates that this is a reasonable assumption.

5. If you were conducting this experiment, what would you try to do to minimise confounding?

This experiment does not have a control group provided. A control group would presumably provide evidence that not doing exercise does not increase Beta endorphins. To reduce the effect of a

confounding variable subjects that are homogenous in some way (sex,BMI,resting heart rate) could be grouped together.

Question 3:

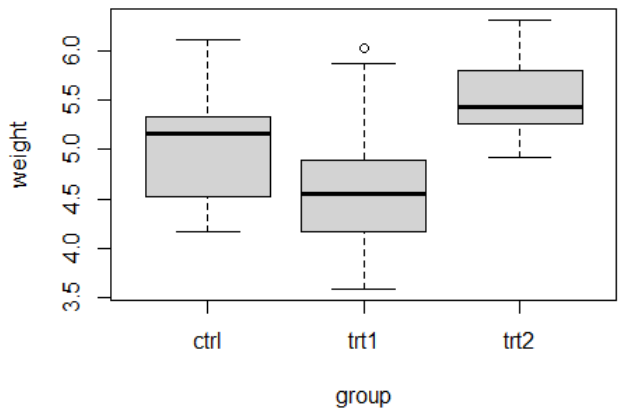
These data are obtained from an experiment to compare plant yields under a control and two other treatments. Test if there is a difference between the control group and treatment 2 on mean plant yields.

Loading & Exploratory Data Analysis

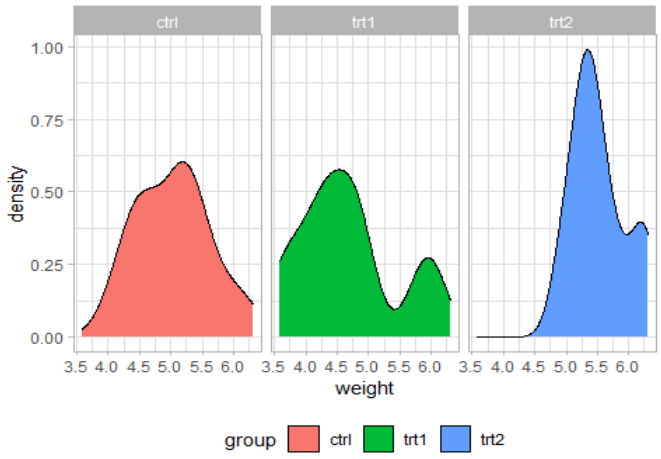
```
# Load plantgrowth dataset into plants variable
plants <- PlantGrowth

# Output summary of descriptive statistics
describeBy(plants, group = "group")
```

```
group: ctrl
       vars  n mean   sd median trimmed  mad  min  max range skew kurtosis   se
weight    1 10 5.03 0.58   5.15    5.00 0.72 4.17 6.11  1.94 0.23    -1.12 0.18
group*    2 10 1.00 0.00   1.00    1.00 0.00 1.00 1.00  0.00  NaN      NaN 0.00
------------------------------------------------------------ 
group: trt1
       vars  n mean   sd median trimmed  mad  min  max range skew kurtosis   se
weight    1 10 4.66 0.79   4.55    4.62 0.53 3.59 6.03  2.44 0.47    -1.1 0.25
group*    2 10 2.00 0.00   2.00    2.00 0.00 2.00 2.00  0.00  NaN      NaN 0.00
------------------------------------------------------------ 
group: trt2
       vars  n mean   sd median trimmed  mad  min  max range skew kurtosis   se
weight    1 10 5.53 0.44   5.44    5.50 0.36 4.92 6.31  1.39 0.48   -1.16 0.14
group*    2 10 3.00 0.00   3.00    3.00 0.00 3.00 3.00  0.00  NaN      NaN 0.00
```

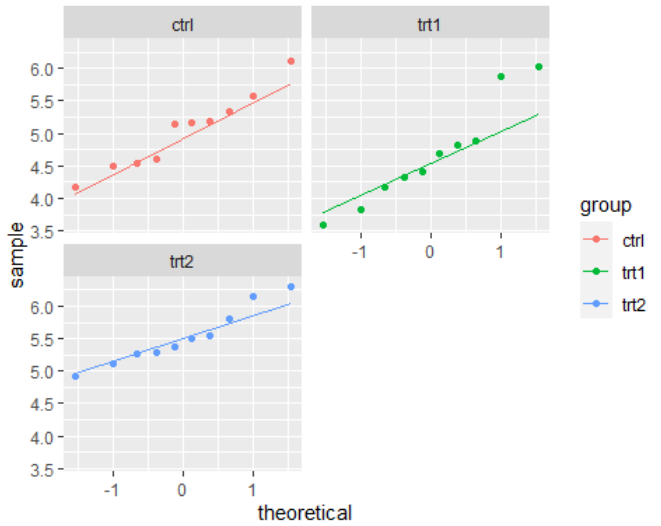
```
# Diaply boxplot
boxplot(weight~group, data = plants)
```




```
# Display density plots for each group
ggplot(plants) +
  aes(x = weight, fill = group) +
  geom_density(adjust = 1L) +
  scale_fill_hue() +
  theme_light() +
  theme(legend.position = "bottom") +
  facet_wrap(vars(group))
```




```
# Display QQ plot for each group
ggplot(plants, aes(sample = weight, colour = group)) +
    stat_qq() +
    stat_qq_line()+
    facet_wrap(nrow = 2, ncol = 2, vars(group), as.table=
TRUE)
```



Make new data frame with only the control group and trt2 group as variables

```
ctrl <- plants %>% filter(group == "ctrl") %>% select(wei
ght)
trt2 <- plants %>% filter(group == "trt2") %>% select(wei
ght)
ctrl <- as_tibble(ctrl)

ctrl <- rename(ctrl, ctrl = weight)
trt2 <- rename(trt2, trt2 = weight)
ctrl$trt2 <- trt2$trt2
main <- ctrl

# Display new dataframe
main

## # A tibble: 10 x 2
##     ctrl  trt2
##    <dbl> <dbl>
##  1  4.17  6.31
##  2  5.58  5.12
##  3  5.18  5.54
##  4  6.11  5.5
##  5  4.5   5.37
##  6  4.61  5.29
##  7  5.17  4.92
##  8  4.53  6.15
##  9  5.33  5.8
## 10  5.14  5.26
```

```r
cor.test(main$ctrl,main$trt2)

##
##  Pearson's product-moment correlation
##
## data:  main$ctrl and main$trt2
## t = -1.4955, df = 8, p-value = 0.1731
## alternative hypothesis: true correlati
on is not equal to 0
## 95 percent confidence interval:
##  -0.8476006  0.2298435
## sample estimates:
##        cor
## -0.4674267

# Potential negative correlation
```

Perform an appropriate significance test

Testing if there is a statistically significant difference between the control group and treatment group:2

Adopting a 5% risk of committing a type I error. Test Type: Paired t-test on *main*

Ho: The mean is equal to 0 (H0:μ = 0)

HA: The mean difference is greater or less than 0 (HA: μ > 0) Two tailed hypothesis. α = 0.05

```r
t.test(main$ctrl,main$trt2, paired = TRUE, conf.level = 0
.95, alternative = c("two.sided"))

##
##  Paired t-test
##
## data:  main$ctrl and main$trt2
## t = -1.7721, df = 9, p-value = 0.1101
## alternative hypothesis: true difference in means is no
t equal to 0
## 95 percent confidence interval:
##  -1.1246169  0.1366169
## sample estimates:
## mean of the differences
##                  -0.494
```

With a p-value of 0.11 being greater than the defined significance level of 0.05 it is apparent that there is not a statistically significant difference between the control group and treatment group two. It was observed in EDA that there was a possible negative correlation, but this has been debunked.

Assumptions

- The data are a simple random sample. (This was not stated in the description)
- Observations come from a population that is normally distributed. The Q-Q plot of the sample data, shown in the EDA step, indicates that this is a reasonable assumption.