MA4413: Statistics for Computing:

R Project Report:

Question 1(ii): Graphical and Numerical Summaries:

Information on Midearly:

Mean:	63.01
Standard deviation(sd):	13.8148
Quartile 1(Q1):	56.700
Quartile 2(Q2):	61.700
Quartile 3(Q3):	69.175
Quartile 4(Q4):	83.300
IQR:	12.475
Minimum:	40
Maximum:	66.7

Information on Midlate:

Mean:	75.67
Standard deviation(sd):	10.0472
Quartile 1 (Q1):	67.525
Quartile 2 (Q2):	73.300
Quartile 3 (Q3):	76.700
Quartile 4 (Q4):	93.300
IQR:	9.175
Minimum:	66.7
Maximum:	93.3

From the above output, we can determine that:

Both groups are of symmetrical data as the median is approximately equivalent to Q2.

The data centres around 9.175 (IQR) for midearly and the data centres around 12.475 (IQR) for midlate.

The overall spread of data for midearly is 26.6 (range) and the overall spread of data for midlate is 43.3 (range), showing the variability of the data shown.

Question 1(iii): Check for Normality of Data:

Midlate seems to be normally distributed as it only has 3 outliers.

In comparison, midearly does not seem to be normally distributed as it has many outliers.

From looking at the histograms and boxplots, we can also see that midlate seems to be normally distributed and that midearly does not seem to be normally distributed.

From the Shapiro-Wilk tests, we can see that the p-value of midearly is low (0.0149), meaning that there is a low chance that midearly is normally distributed. On the other hand, we can see the p-value of midlate is high (0.6032), meaning there is a high possibility that it is from normally distributed data.

Question 1(iv): Confidence Intervals and Hypothesis Testing:

a)

Ho: The overall mean midterm score is equal to 50.

Ha: The overall mean midterm score is not equal to 50.

From the t-test, we can see that the mean is not equal to fifty as the estimated mean is equal to 69.34 and the value of 50 is outside the 95% confidence interval of [63.05402, 75.63598], so I must reject the null hypothesis (Ho) and accept the alternative hypothesis (Ha).

```
        One Sample t-test


data:  midsample$score
t = 23.088, df = 19, p-value = 2.307e-15
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 63.05402 75.62598
sample estimates:
mean of x
    69.34
```

b)

Ho: There is no difference between the variances in each group.

Ha: That there is a difference between the variances in each group.

From the variance test, the p-value (0.3567) is greater than 0.05, therefore we can accept the null hypothesis (Ho) that the is no differences between the variances in each group.

```
F test to compare two variances

data:  midearly and midlate
F = 0.52894, num df = 9, denom df = 9, p-value = 0.3567
alternative hypothesis: true ratio of variances is not equal
to 1
95 percent confidence interval:
 0.131381 2.129505
sample estimates:
ratio of variances
          0.528939
```

c)

Ho: That there is no difference between the means in each group.

Ha: That there is a difference between the means in each group.

We can see from the t-test, we can see from the estimated means that there is a difference between the two mean, the estimate mean of midearly is 75.67 and the estimate mean of midlate is 63.01.

```
   Welch Two Sample t-test


data:  midearly and midlate
t = 2.3437, df = 16.44, p-value = 0.03196
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  1.233526 24.086474
sample estimates:
mean of x mean of y
    75.67     63.01
```

d)

Ho: When the samples are small, the t-test requires the assumption that data is normally distributed.

Ha: When the samples are small, the t-test requires the assumption that data is not normally distributed.

As we can see from the Wilcoxon rank sum test, we can see that the data is not normally distributed as the p-value is less that .05.

```
Wilcoxon rank sum test with continuity correction

data:  midearly and midlate
W = 77.5, p-value = 0.03905
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(midearly, midlate) :
  cannot compute exact p-value with ties
```

Question 1(v): (independent research) Simple Linear Regression:

a & b)

From the correlation coefficient and graph, we can see that there is a weak downhill (negative) linear relationship. It is of value (-0.4362).

c)

The estimated coefficient of day is (-2.517), meaning is a steady, downhill slope. From the summary, we can see that the p-value is ($p<0$) less than 0, meaning that it is a 100% confidence interval, that it is the right coefficient value.

d)

From the graph, we can see that the regression line does not fit well to the graph as there are many outliers and no points touching the line. We can see the values are scattered away from the line. From the value R squared, we can see that the graph makes up for (0.1903) or 19% of the variance.

Question 1(vi): Brief Summary of Analysis:

From the above questions (ii) to (v), we can see that there are two groups of symmetrical data, which have an wide overall spread and centres around the two medians.

We can see that the means are not equal to 50, that there are no differences in variances, that there is a difference in the means and that midearly is not normally distributed and midlate is normally distributed.

From the graph, we can see the data is very scattered and that there is a weak downhill linear relationship, with the value r squared making up 19% of the variance.

The conclusion, beginning that the data is symmetrical, they have a weak relationship and that there is an overall spread of data. Also, midearly is not normally distributed and midlate is normally distributed.

Question 3: Probability Questions:

i) 0.9888
ii) 0.3302
iii) 0.6698
iv) 0.1033
v) 0.1969
vi) 0.2606
vii) 0.0228
viii) 0.5793
ix) 0.0081
x) 0.4071

Group Members:

ID: 15167771 Name: Stephen King

ID: 15150763 Name: Daniel Lavin

ID:15143929 Name: Trevor Sherin

ID:15141225 Name: Michael Ryan