Oscar Beresford 201722873

Task 1

The researcher initially started by using the built in functions into R Studio to import the *vessels.csv* dataset. They then altered the imported dataset by removing the Licence.Category column as they were not interested in the data it held. They then used the `unique()` function to remove duplicated excess rows. They then wrote the cleaned data-set to a new csv file labelled *cleaned_vessels.csv* .

```
df <- subset(ves, select = -Licence.Category)

data <- unique(df)

write.csv(data, "cleaned_vessels.csv")
```

They found that the largest and smallest lengths of boats in the dataset were *199.65 m* and *10.04 m* respectively

```
longest_length <- max(cleaned_vessels$Overall.length)

longest_length

min_length <- min(cleaned_vessels$Overall.length)

min_length


199.65

10.04
```

They found that *99* duplicates were removed from the original data set using the `unique()`and `nrow()` functions. The unique() function removed the duplicates, then they used the `nrow()` function on both the original and the cleaned data set finding the difference in rows was 99 and hence the number of rows removed as duplicates.

Furthermore, they found the most common vessel name to be *KINGFISHER.* They used the table() function to create a frequency table, from which they were able to calculate the name of the vessel which appeared the most in the *Vessel.name* collum using the `max()` function.

```
vessel_freq <-table(cleaned_vessels$Vessel.name)

mode_vessel<-names(vessel_freq)[which.max(vessel_freq)]

mode_vessel

KINGFISHER
```

Task 2

The Researcher was tasked with splitting the data up into two subsets based on the location of the home port. They used the `subset()` function in R to create a subset of *cleaned_vessles*, `Home.port== " … "` creates the subset based on the name of the home port. In this case, the desired home ports were *ARGLASS* and *NEWLYN.*

```
Ardglass<- subset(cleaned_vessels,Home.port== "ARDGLASS")

Ardglass
```

```
Newlyn <- subset(cleaned_vessels,Home.port =="NEWLYN")

Newlyn
```

Then they had to find the number of vessels that Ardglass or Newlyn as the home port. Simply using the `nrow()` function on the newly created subsets was the easiest way to find the required information. It was found that Ardglass had *18* vessels that identified it as the home port while Newlyn had a much greater *52* vessels.

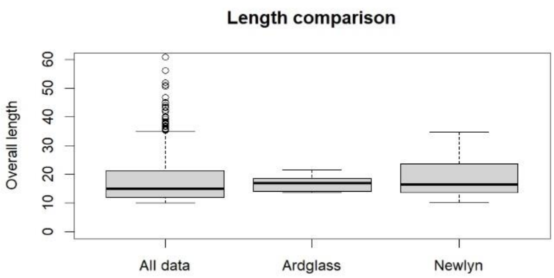
```
num_rows_arglass <- nrow(Ardglass)

num_rows_arglass

num_rows_newlyn <- nrow(Newlyn)

num_rows_newlyn
```

```
18

52
```

The researcher chose to use a boxplot to compare the different data-sets. They limited the y variable between 0 and 1500, as it led to the boxplot not being able to be properly analysed.

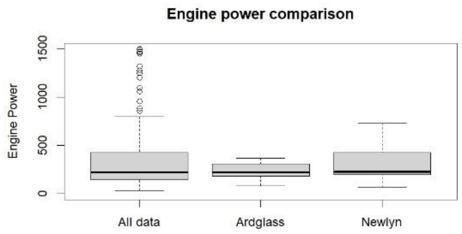
From the length boxplot, they determined that Newlyn was fairly standard compared to the complete data. However, they noted that Ardglass had a slightly greater mean length while also having a much tighter interquartile range, implying low dispersion in the central portion of the dataset.

```
boxplot(cleaned_vessels$Overall.length, Ardglass$Overall.length,
Newlyn$Overall.length, names=c("All data", "Ardglass","Newlyn"),
ylim=c(0,60), main="Length comparison", ylab="Overall length")
```



For engine power, again it’s important to note that Ardglass had a much tighter interquartile range and a smaller range. For Newlyn, the researcher comments on the clear negative skew pulling the mean towards the bottom of the interquartile range implying there may be a few very small values pulling the mean down. Hence the researcher has established Newlyn to be suitably typical.

```
boxplot(cleaned_vessels$Overall.length, Ardglass$Overall.length,
Newlyn$Overall.length, names=c("All data", "Ardglass","Newlyn"),
ylim=c(0,60), main="Length comparison", ylab="Overall length")
```

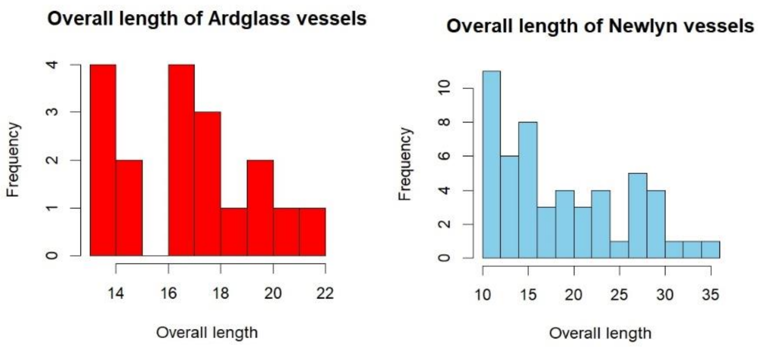


Task 3

Here the researcher was tasked to compare overall lengths and engine powers of vessels in the two ports. They used histograms and specifically didn't restrict the ranges of any of the values as they wanted to have an accurate portrayal of the data in hand. They used *10* breaks for all the histograms to keep it constant across all the data.

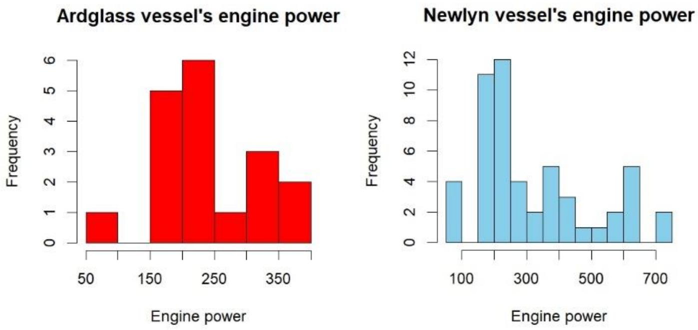
```
hist(Ardglass$Overall.length, main="Overall length of Ardglass vessels", xlab="Overall length", col="red", breaks = 10)

hist(Newlyn$Overall.length, main="Overall length of Newlyn vessels", xlab="Overall length", col="skyblue", breaks = 10)

hist(Ardglass$Engine.power, main="Ardglass vessel's engine power", xlab="Overall length", col="red", breaks = 10)

hist(Newlyn$Engine.power,main="Newlyn vessel's engine power", xlab="Overall length", col="skyblue", breaks = 10)
```

They came to the conclusion that despite having different number of datapoints, both the vessels' lengths are similarly distributed, both showing greater frequencies at the lower lengths then a tapering off as the lengths increase, implying for both ports that there are greater numbers of boats in the *10-16* range



For the engine power, they found that both the ports had a very similar distribution with the largest peaks around *150-250* range. No vessels in the *100-150* range and then a taper off in the number of vessels as the lengths increase after the 2 large peaks.

It is clear there is also a strong relation between the engine power and the overall length of the vessels, as the histograms all follow the same patterns and trends clearly showing how engine power and length of a vessel are linked together.



Task4

For the final task the researcher had to preform a welch t-test two different ways. First calculating the test statistic manually and then secondly using the built in `t.test()` function.

Firstly, for the engine power of the vessels they found that p-value was *0.0074832* and hence there was sufficient evidence at the *5%* level to reject the null hypothesis and instead accept the alternative hypothesis. Using the equation for the test statistic and assigning variables to the related values, they found the t value was *-2.763119* which relates to the p-value of *0.0074832*.

```
Ard_mean <- mean(Ardglass$Engine.power)

New_mean <- mean(Newlyn$Engine.power)

Ard_var <- var(Ardglass$Engine.power)

New_var <- var(Newlyn$Engine.power)

result <- (Ard_mean - New_mean) / ((Ard_var)/18 + (New_var)/52)^0.5

result


-2.76319
```

Next using the built in `t.test()` function they found that the t-value was *2.7632* and the relating p-value was *0.00748*. Here, `var.equal = FALSE` means the function doesn't assume the variances are equal which is critical when preforming a t-test.

```
result <- t.test(Ardglass$Engine.power, Newlyn$Engine.power,
var.equal = FALSE)

p_value<- result$p.value

p_value


0.00748
```

The same approach was again taken to perform a t-test on the overall length of the vessels. Instead of using `Engine.power` they used `Overall.length` and found a p-value of *0.0912844* and *0.09128* , respectively to the two methods, finding there was insufficient evidence at the *5%* level to reject the null hypothesis.

School of Mathematics
FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

UNIVERSITY OF LEEDS

Academic integrity statement

You must sign this Oscar Beresford and include it with each piece of work you submit.

I am aware that the University defines plagiarism as presenting someone else's work, in whole or in part, as your own. Work means any intellectual output, and typically includes text, data, images, sound or performance.

I promise that in the attached submission I have not presented anyone else's work, in whole or in part, as my own and I have not colluded with others in the preparation of this work. Where I have taken advantage of the work of others, I have given full acknowledgement. I have not resubmitted my own work or part thereof without specific written permission to do so from the University staff concerned when any of this work has been or is being submitted for marks or credits even if in a different module or for a different qualification or completed prior to entry to the University. I have read and understood the University's published rules on plagiarism and also any more detailed rules specified at School or module level. I know that if I commit plagiarism I can be expelled from the University and that it is my responsibility to be aware of the University's regulations on plagiarism and their importance.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to monitor breaches of regulations, to verify whether my work contains plagiarised material, and for quality assurance purposes.

I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and that I wish to have taken into account. I am aware of the University's policy on mitigation and the School's procedures for the submission of statements and evidence of mitigation. I am aware of the penalties imposed for the late submission of coursework.

Name	Oscar William Beresford
Student ID	201722873