**R-PRACTICAL 2 – MATH1712 - REPORT**

Task 1:  
When first glancing at our data, we notice there are 1,368 entries with 16 headings. However, it becomes apparent that rows are duplicated in the original spreadsheet; in order to overcome this, we deleted the column ‘Licence Category’ and used the “unique()” command. More specifically, we assigned the variable x as “x<-unique.data.frame(X)” where X is our original data. As always, we then attached x and used the View() command; from this we see that there are now 1,155 entries which suggests 213 rows were removed as duplicates.

In order to gain a general overview and a better understanding of the data, we used the summary statistics for each of the 16 headings. Using “t=x$’Header’ > summary(as.factor(t))” where ‘Header’ is one of the non-numerical 16 headings and “summary(x$’Header2’)” where ‘Header2’ is one of the numerical headings. Repeating this process for the 16 headings we were able to construct the following table:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | MIN | 1ST QU | MODE | MEAN | 3RD QU | MAX |
| Administrative Port | n/a | n/a | BELFAST | n/a | n/a | n/a |
| Home Port | n/a | n/a | UNKNOWN | n/a | n/a | n/a |
| Port Letters and Numbers | n/a | n/a | n/a | n/a | n/a | n/a |
| Vessel Name | n/a | n/a | KINGFISHER | n/a | n/a | n/a |
| Registry of Shipping and Seamen Number | n/a | n/a | n/a | n/a | n/a | n/a |
| Licence Number | n/a | n/a | n/a | n/a | n/a | n/a |
| Fish Producer Organisation | n/a | n/a | NON-SECTOR | n/a | n/a | n/a |
| Overall Length | 10.01 | 11.90 | n/a | 18.46 | 18.46 | 113.97 |
| Registered Tonnage | 3.94 | 18.76 | n/a | 145.18 | 150.00 | 5579.00 |
| Engine Power | 35.81 | 145.00 | n/a | 395.11 | 405.75 | 7320.00 |
| Vessel Capacity Units | 49.79 | 124.92 | n/a | 304.18 | 322.98 | 5140.14 |
| Year Built | 1896 | n/a | 1988 | 1988 | n/a | 2018 |
| Hull | n/a | n/a | S | n/a | n/a | n/a |
| County of Construction | n/a | n/a | GBR | n/a | n/a | n/a |
| Shellfish Licence Y/N | n/a | n/a | N | n/a | n/a | n/a |
| Scallop Licence Y/N | n/a | n/a | N | n/a | n/a | n/a |

Despite not being able to calculate certain values due to some columns being non-numeric, we can see that this dataset is *very* spread. For Registered Tonnage, Engine Power and Vessel Capacity Units, we see their minimum measurement in the tens to their maximum measurement in the thousands. We can also observe that on average, most vessels don’t have a Shellfish or Scallop License and their County of Construction is mainly based in Great Britain. Summarising the data in the table above allows us to make quick observations and observe any potential differences between categories.

Using the information in our table above, we see that the most common vessel name is ‘Kingfisher’. We can use the “t=x$‘Vessel.name’” and “summary(as.factor(t))” to ensure this is correct; this specific command outputs a table with vessel names in order of most occurrence. R Studio indeed tells us that ‘Kingfisher’ is the most common vessel name, with 5 vessels under the name Kingfisher.

Task 2:  
Since we are only interested in the vessels ‘Ardglass’ and ‘Kilkeel’, we assigned the variables ‘y’ and ‘z’ to the data of ‘Ardglass’ and ‘Kilkeel’ respectively. We were able to do this with the following R Commands: “y<-x[x$Home.port == "ARDGLASS",]” and “z<-x[x$Home.port == "KILKEEL",]”. Using the “attach()” and “View()” command, we were able to observe that 21 vessels from the original 1155 had Ardglass as their home port and 57 vessels had Kilkeel as their home port.

Using a similar process above, we can construct a numerical summary table for the two home ports and compare our findings to our overall summary table above. If both tables show similarities, we will be able to conclude whether or not the two vessels are ‘typical’.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ARDGLASS** | MEAN | VAR | MIN | MAX | 1ST QU | 3RD QU |
| Overall Length | 16.56 | 4.27 | 12.90 | 19.97 | 14.95 | 17.77 |
| Registered Tonnage | 55.85 | 472.11 | 25.53 | 106.00 | 39.63 | 64.58 |
| Engine Power | 225.40 | 5073.26 | 82.00 | 373.00 | 184.20 | 243.00 |
| Vessel Capacity Units | 193.90 | 2088.79 | 105.30 | 292.40 | 169.70 | 215.20 |

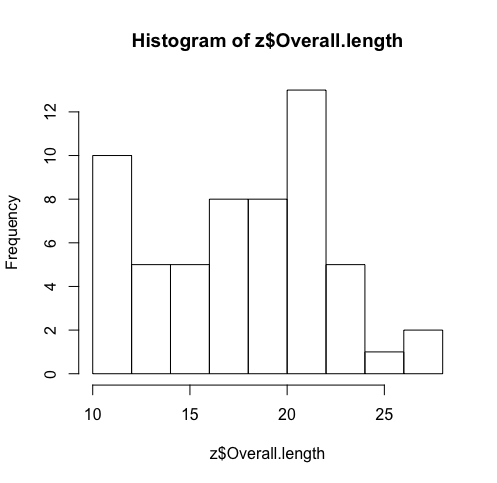
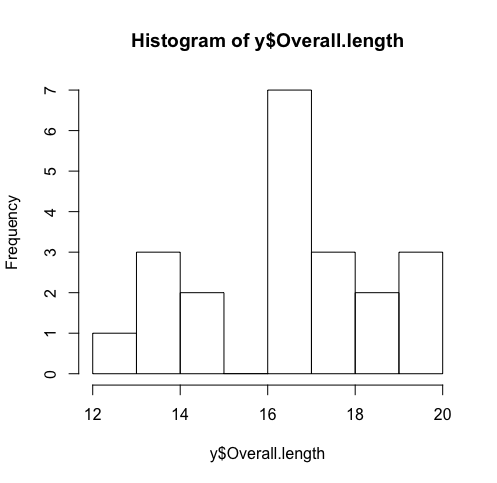
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **KILKEEL** | MEAN | VAR | MIN | MAX | 1ST QU | 3RD QU |
| Overall Length | 17.67 | 18.79 | 10.19 | 27.65 | 13.95 | 20.87 |
| Registered Tonnage | 87.73 | 4094.90 | 9.55 | 267.30 | 31.00 | 127.00 |
| Engine Power | 303.90 | 24456.5 | 73.0 | 675.0 | 171.0 | 438.0 |
| Vessel Capacity Units | 245.27 | 12300.23 | 77.43 | 473.00 | 138.72 | 331.94 |

When drawing comparisons between these two tables and the overall table above, we see that the overall lengths of the vessels on each port is slightly lower than the average for all fishing vessels. We can also observe that their registered tonnage, engine power and vessel capacity units is considerably lower than the average. From our observations, we can conclude that the two home ports of Ardglass and Kilkeel are slightly below average and are therefore not ‘typical’ of a vessel in the UK with Ardglass being on average, lower in all areas than Kilkeel. We may be hasten to add that perhaps the vessels at Ardglass and Kilkeel are smaller than other vessels recorded in this data.

Task 3:  
In this specific task we were asked to compare the overall lengths and engine powers of vessels based in the two ports. Since we have two constructed numerical tables above we are able to compare the two table’s columns ‘Overall Lengths’ and ‘Engine Powers’.

We can observe that on average, Kilkeel vessels have a larger overall length than those of Ardglass. However, Ardglass has a smaller variance, implying that Kilkeel’s data is spread. This can be further supported by the range of Kilkeel being 17.55 compared to the range of Ardglass at 7.07. In addition, the same pattern develops for Engine Power, with Kilkeel having the larger mean of the two vessels but also with the largest variance and range.

We are able to construct histograms using the following R command “hist(y$Overall.length)” and “hist(y$Engine.power)” and then repeat the same process for our variable ‘z’. Doing this we get the following Histograms:



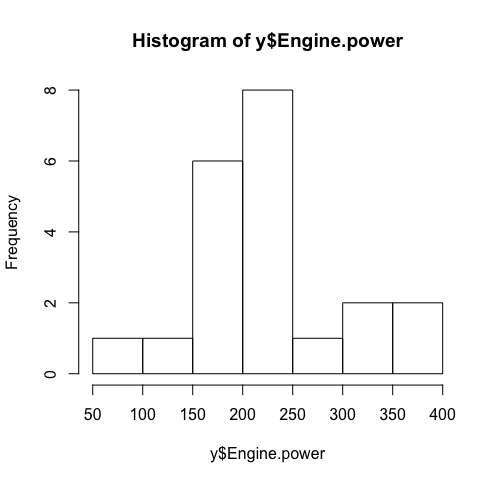
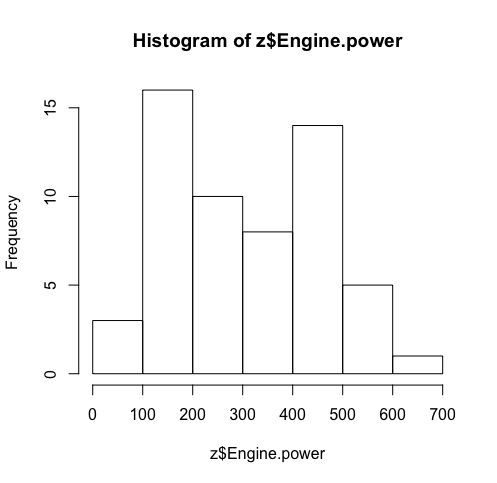
**Histogram of Overall Length of Kilkeel**

**Histogram of Overall Length of Ardglass**

Overall Length of Vessels at Kilkeel

Overall Length of Vessels at Ardglass

Comparing histograms, we notice that the vessels with homeport at Kilkeel has a larger spread of data than vessels with homeport at Ardglass. In addition to this, the mode for Ardglass is in the 16-17 category whereas the mode for Kilkeel is in the 20-22 category. From this particular observation, we could make the conclusion that vessels with homeport at Kilkeel have, on average, a larger overall length than vessels with homeport at Ardglass.



**Histogram of Overall Engine Power of Kilkeel**

**Histogram of Overall Engine Power of Ardglass**

Overall Engine Power of Vessels at Kilkeel

Overall Engine Power of Vessels at Ardglass

Immediately, looking at the histogram for the overall engine power of vessels at Ardglass, we can conclude that the engine power is consistently, on average, around the 150-250 mark. However, the distribution of engine power for vessels at Kilkeel appear to have an inconsistent distribution with two peaks. In addition, the spread of the data for Kilkeel is considerably higher than that of Ardglass, with the largest class boundary extending up to 700, compared to the 400 for Ardglass. From this, we can conclude that the overall engine power of vessels at Kilkeel are on average greater but less consistent than vessels at Ardglass.

Task 4

For the first part of task 4, we are asked to use a t-test to the following hypotheses at the 5% significance level:

H0: Overall vessel lengths at both ports have the same mean *OR* H0: µ1 = µ2  
H1: Overall vessel lengths at both ports have different means *OR* H1: µ1 ≠ µ2

Before using R Studio to calculate the whole test statistic, we must carry out the calculations individually and observe if we arrive at the same result. We assign the variables ‘A’ and ‘B’ to the vessel lengths at each port by inputting the following R commands: “A = z$Overall.length” and “B = y$Overall.length”. Then using simple mean() and var() commands we can construct the following table for our data A and data B.

|  |  |  |
| --- | --- | --- |
|  | Data A (Overall Vessel Length for Kilkeel) | Data B (Overall Vessel Length for Ardglass) |
| Mean | 17.67158 | 16.56048 |
| Variance | 18.78559 | 1.522009 |
| Length of Data | 57 | 21 |

From this we observe that the two datasets have different variances and so, we must carry out Welch’s t-test. Using the data above and carrying out the calculations individually we observe:

t = “(mean(a)-mean(b))/sqrt((var(a)/57)+var(b)/21)” = 1.522009

We then carry out the entire process on R Studio by inputting “t.test(a,b, alternative = "two.sided", mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)”.

From this command we obtain that p = 0.1325 and t = 1.522 which is consistent with our calculated t-value above. Since we are calculating a Welch’s Two Sample t-test, we can use the calculated p value and compare it against our 2.5% significance level. For this particular t-test for the overall lengths, we obtain a test statistic of 1.522 and a p-value of 13.25% and since 13.25% > 2.5%, there is insufficient evidence to reject H0 at the 5% significance level. We can therefore suggest that the mean lengths of vessels in Kilkeel is the same as Ardglass.

For the second part of this task, we are asked to use a t-test to the following hypotheses at the 5% significance level:

H0: Vessel Engine Powers at both ports have the same mean *OR* H0: µ1 = µ2  
H1: Vessel Engine Powers at both ports have different means *OR* H1: µ1 ≠ µ2

Similar to above, we assign the variables ‘C’ and ‘D’ to the engine powers of the vessels at each port by inputting the following R commands: “C = z$Engine.power” and “D = y$Engine.power”. Then using simple mean() and var() commands we can construct the following table for our data C and data D.

|  |  |  |
| --- | --- | --- |
|  | Data C (Engine Power for Kilkeel) | Data D (Engine Power for Ardglass) |
| Mean | 303.866 | 225.4314 |
| Variance | 24456.49 | 5073.264 |
| Length of Data | 57 | 21 |

From this we observe that the two datasets have different variances and so, we must carry out Welch’s t-test. Using the data above and carrying out the calculations individually we observe:

t = “(mean(c)-mean(d))/sqrt((var(c)/57)+var(d)/21)” = 3.028732

We then carry out the entire process on R Studio by inputting “t.test(c,d, alternative = "two.sided", mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)”.

From this command we obtain that p = 0.003401 and t = 3.0287 which is consistent with our calculated t-value above. Since we are calculating a Welch’s Two Sample t-test, we can use the calculated p value and compare it against our 2.5% significance level. For this particular t-test for the overall lengths, we obtain a test statistic of 3.0287 and a p-value of 0.3401% and since 0.3401% < 2.5%, there is sufficient evidence to reject H0 at the 5% significance level. We can therefore suggest that the mean engine power for vessels in Kilkeel differs from vessels in Ardglass.