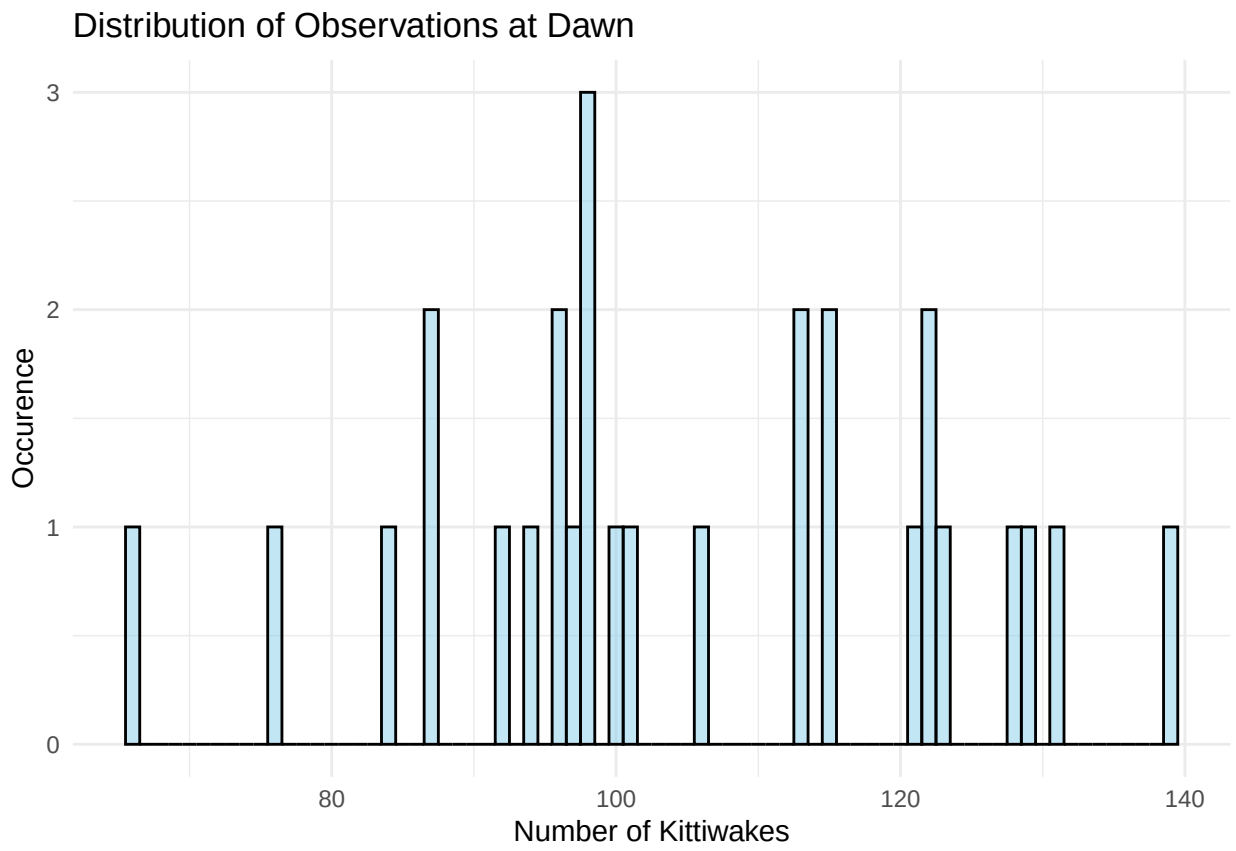
Assignment_2

Rohith Ganesan

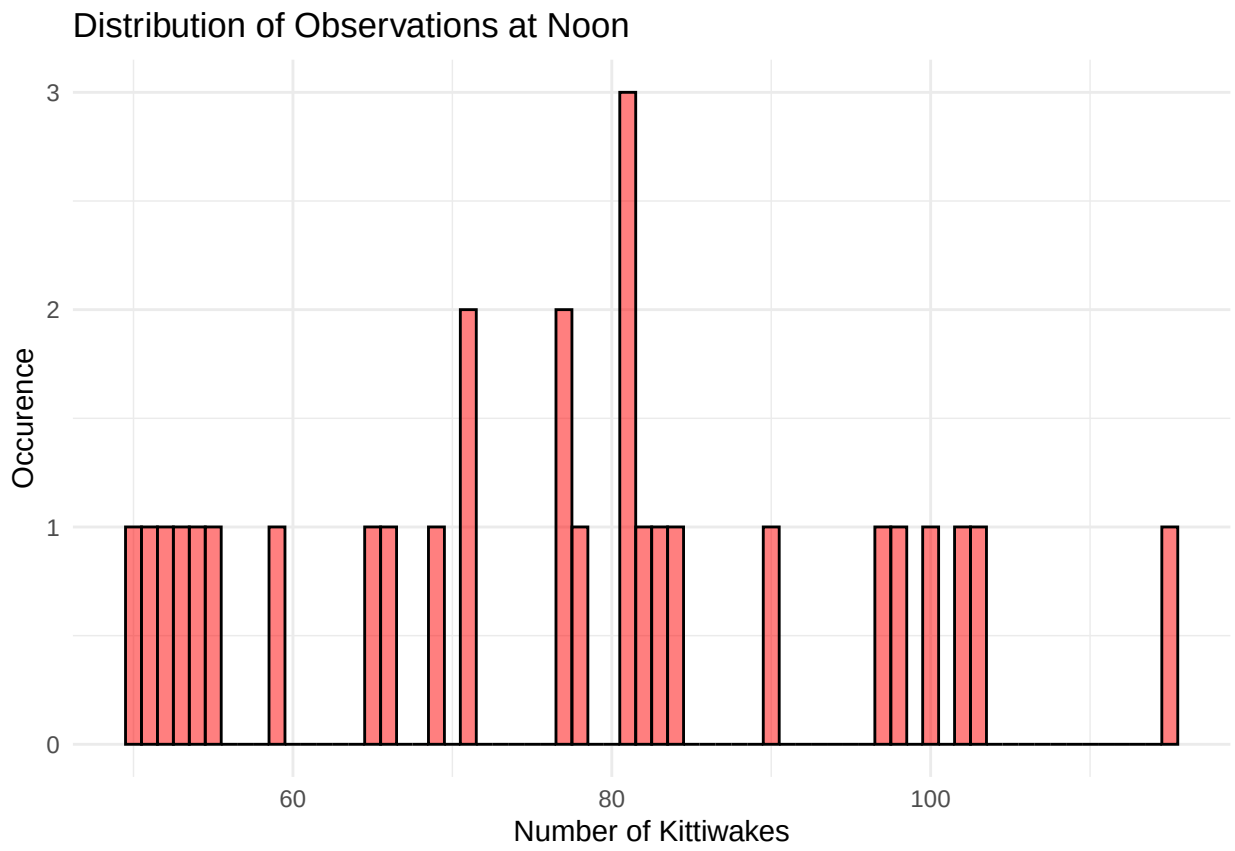
14/12/2023

```
library(ggplot2)
obs_data<-read.csv("Observation_20553375(Rohith).csv")
loc_data<-read.csv("Location_20553375(Rohith).csv")
meas_data<-read.csv("Measurement_20553375(Rohith).csv")
hist_data<-read.csv("Historical_20553375(Rohith).csv")
```

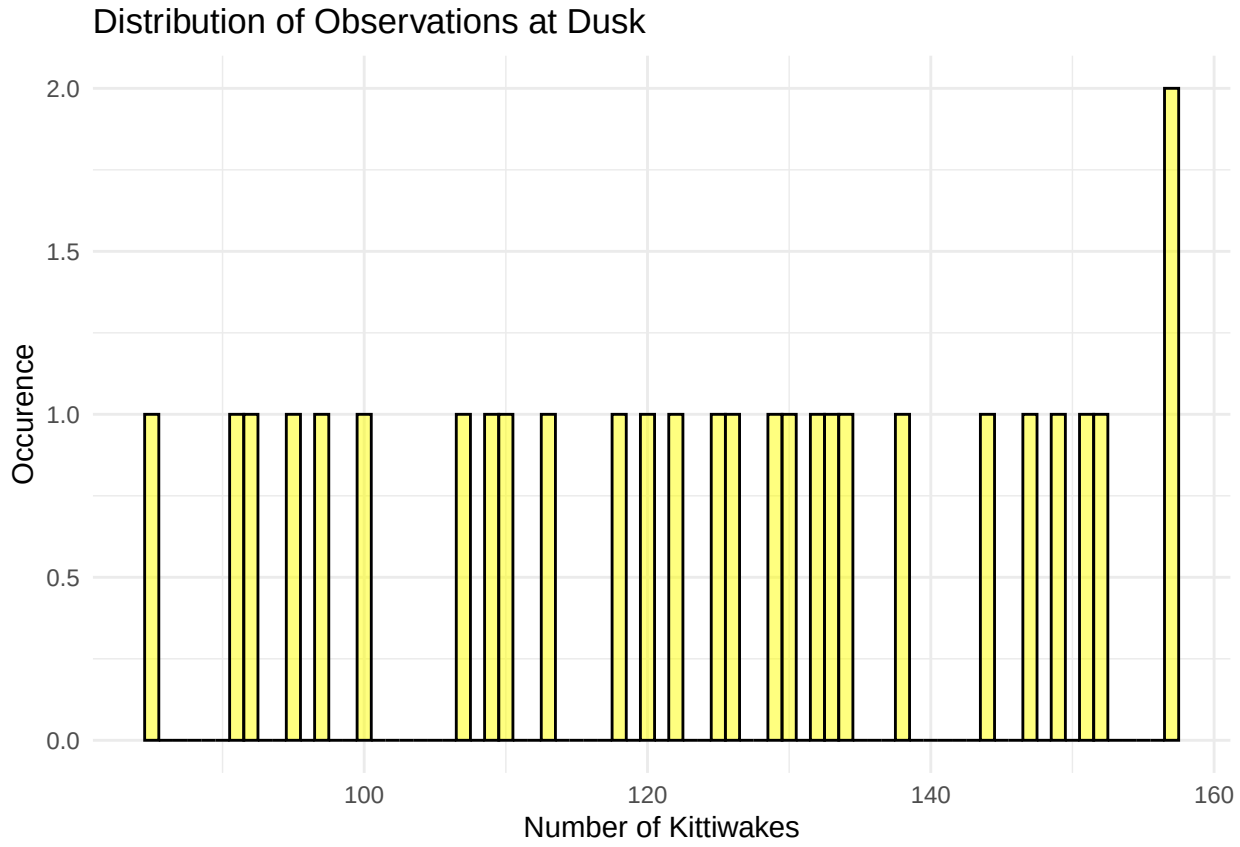
```
#Visualization & Exploratory Analysis for Q1:

if (!requireNamespace("ggplot2", quietly = TRUE)) {
  install.packages("ggplot2")
}

library(ggplot2)



ggplot(obs_data, aes(x = dawn)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black", alpha = 0.5) +
  labs(title = "Distribution of Observations at Dawn",
       x = "Number of Kittiwakes",
       y = "Occurence") +
  theme_minimal()
```




```
ggplot(obs_data, aes(x = noon)) +
  geom_histogram(binwidth = 1, fill = "red", color = "black", alpha = 0.5) +
  labs(title = "Distribution of Observations at Noon",
       x = "Number of Kittiwakes",
       y = "Occurence") +
  theme_minimal()
```




```
ggplot(obs_data, aes(x = dusk)) +
  geom_histogram(binwidth = 1, fill = "yellow", color = "black", alpha = 0.5) +
  labs(title = "Distribution of Observations at Dusk",
       x = "Number of Kittiwakes",
       y = "Occurence") +
  theme_minimal()
```




```
summary(obs_data)
```

```
##       dawn            noon        mid.afternoon         dusk
##  Min.   : 66.0   Min.   : 50.00   Min.   : 59.00   Min.   : 85.0
##  1st Qu.: 95.5   1st Qu.: 63.50   1st Qu.: 80.50   1st Qu.:108.5
##  Median :100.5   Median : 77.50   Median :100.00   Median :125.5
##  Mean   :105.2   Mean   : 76.61   Mean   : 96.75   Mean   :123.7
##  3rd Qu.:121.2   3rd Qu.: 85.50   3rd Qu.:115.00   3rd Qu.:139.5
##  Max.   :139.0   Max.   :115.00   Max.   :135.00   Max.   :157.0
```

```
summary(hist_data)
```

```
##        X            Site.A          Site.B          Site.C          Site.D
##  Min.   :2000   Min.   :27.0   Min.   :40.0   Min.   :40.0   Min.   :25
##  1st Qu.:2005   1st Qu.:32.0   1st Qu.:50.0   1st Qu.:40.0   1st Qu.:28
##  Median :2010   Median :34.0   Median :55.0   Median :40.0   Median :33
##  Mean   :2010   Mean   :38.4   Mean   :54.6   Mean   :42.2   Mean   :33
##  3rd Qu.:2015   3rd Qu.:42.0   3rd Qu.:58.0   3rd Qu.:45.0   3rd Qu.:39
##  Max.   :2020   Max.   :57.0   Max.   :70.0   Max.   :46.0   Max.   :40
##      Site.E
##  Min.   :52
##  1st Qu.:61
##  Median :61
##  Mean   :61
##  3rd Qu.:62
```

```
##  Max.   :69
```

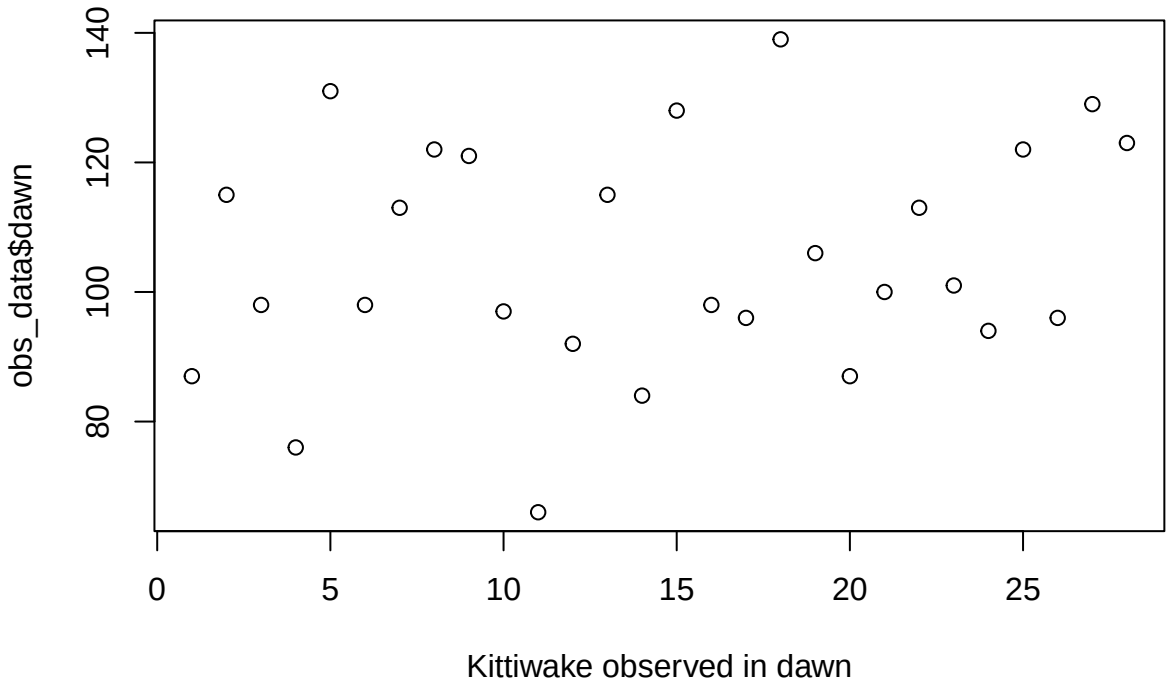
```r
summary(meas_data)
```

```
##  Sub.species            Weight         Wingspan          Culmen
##  Length:32          Min.   :339.0   Min.   : 88.00   Min.   :29.00
##  Class :character   1st Qu.:372.2   1st Qu.: 95.75   1st Qu.:33.00
##  Mode  :character   Median :382.5   Median : 98.50   Median :36.00
##                     Mean   :382.2   Mean   : 98.62   Mean   :36.78
##                     3rd Qu.:389.8   3rd Qu.:103.00   3rd Qu.:41.00
##                     Max.   :432.0   Max.   :109.00   Max.   :47.00
```

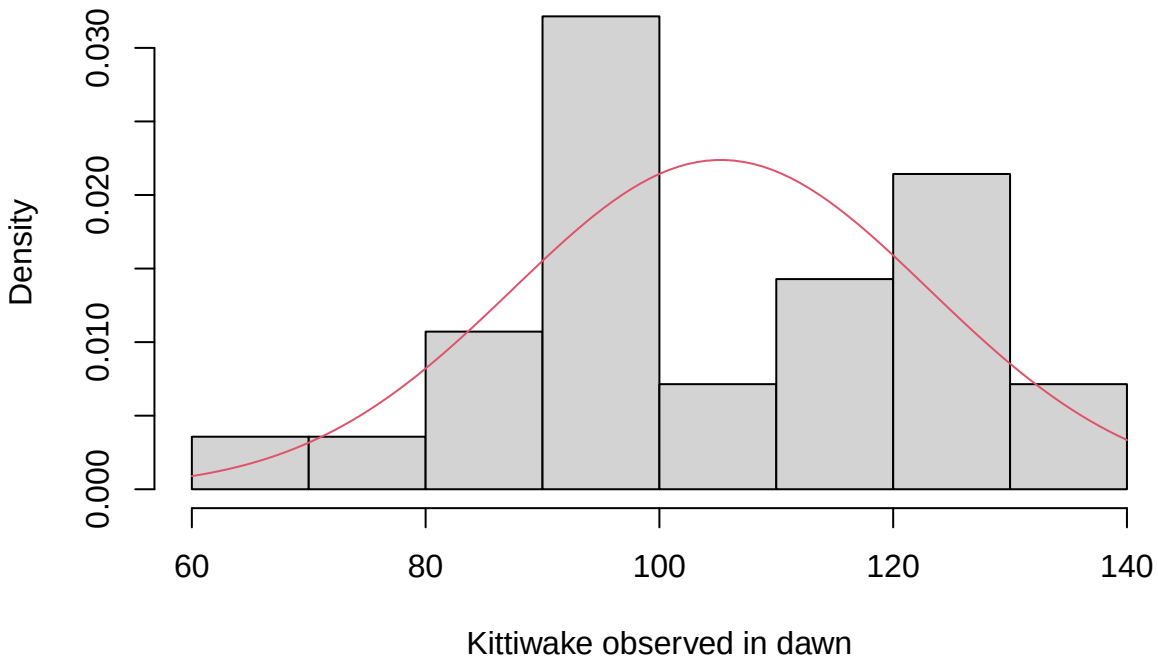
```r
summary(loc_data)
```

```
##  Coast.direction       sandeel       Summer.temp     cliff.height
##  Length:29          Min.   :0.560   Min.   :19.60   Min.   :2.980
##  Class :character   1st Qu.:1.360   1st Qu.:21.80   1st Qu.:3.480
##  Mode  :character   Median :2.170   Median :23.00   Median :3.830
##                     Mean   :1.953   Mean   :22.84   Mean   :3.831
##                     3rd Qu.:2.720   3rd Qu.:23.90   3rd Qu.:4.110
##                     Max.   :2.940   Max.   :28.30   Max.   :4.730
##  Breeding.pairs
##  Min.   : 84.0
##  1st Qu.:157.0
##  Median :219.0
##  Mean   :255.5
##  3rd Qu.:332.0
##  Max.   :608.0
```

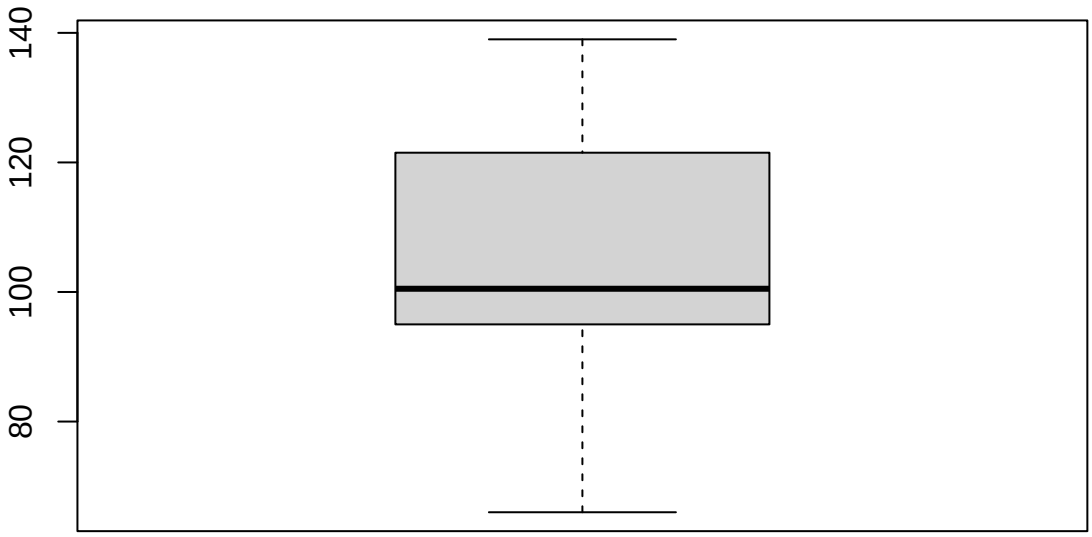
```r
#Generating 99% confidence interval for the the Dawn observations
plot(obs_data$dawn, main = "", xlab = "Kittiwake observed in dawn")
```




```
hist(obs_data$dawn, main = "", xlab = "Kittiwake observed in dawn" ,freq = FALSE)
curve(dnorm(x,mean(obs_data$dawn),sd(obs_data$dawn)),col=2,add=TRUE)
```




```
boxplot(obs_data$dawn)
```




```
#Since the data is normaly distributed we can use t-test to calculate the confidence interval for the m
t.test(obs_data$dawn, conf.level = 0.99)$conf.int
```

```
## [1]  95.91746 114.58254
## attr(,"conf.level")
## [1] 0.99
```

```
mean(obs_data$dawn)
```

```
## [1] 105.25
```

```
#Question 2
```

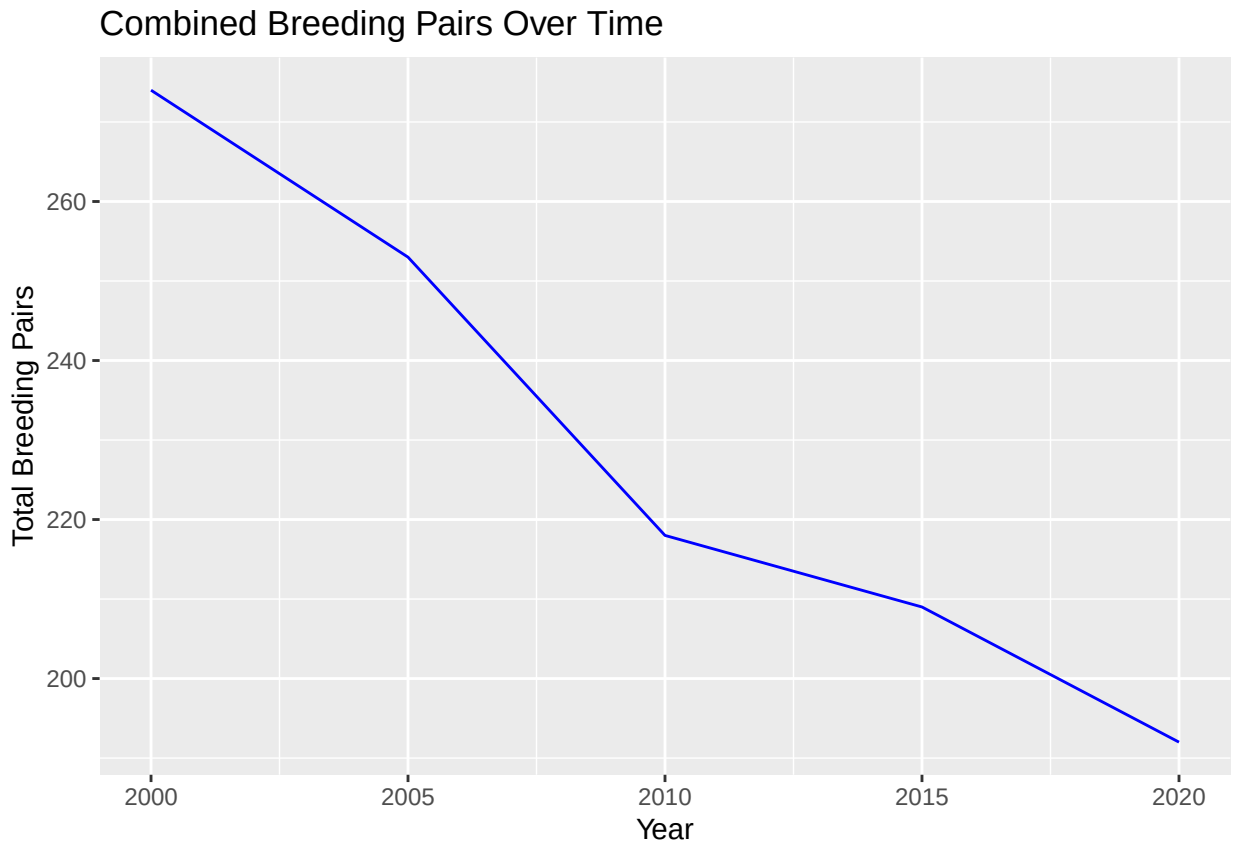
```
hist_data
```

```
##      X Site.A Site.B Site.C Site.D Site.E
## 1 2000     57     70     46     40     61
## 2 2005     42     58     45     39     69
## 3 2010     34     55     40     28     61
## 4 2015     32     50     40     25     62
## 5 2020     27     40     40     33     52
```

```r
column_names <- names(hist_data)

# Display the column names
print(column_names)
```

```
## [1] "X"      "Site.A" "Site.B" "Site.C" "Site.D" "Site.E"
```

```r
total_breeding_pairs <- rowSums(hist_data[, c("Site.A", "Site.B", "Site.C", "Site.D", "Site.E")])


ggplot(data = data.frame(X = hist_data$X, TotalBreedingPairs = total_breeding_pairs)) +
  geom_line(aes(x = X, y = TotalBreedingPairs), color = "blue") +
  labs(x = "Year", y = "Total Breeding Pairs", title = "Combined Breeding Pairs Over Time")
```




```r
chi_square_result <- chisq.test(hist_data)
print(chi_square_result)
```

```
##
##  Pearson's Chi-squared test
##
## data:  hist_data
## X-squared = 30.513, df = 20, p-value = 0.06196
```

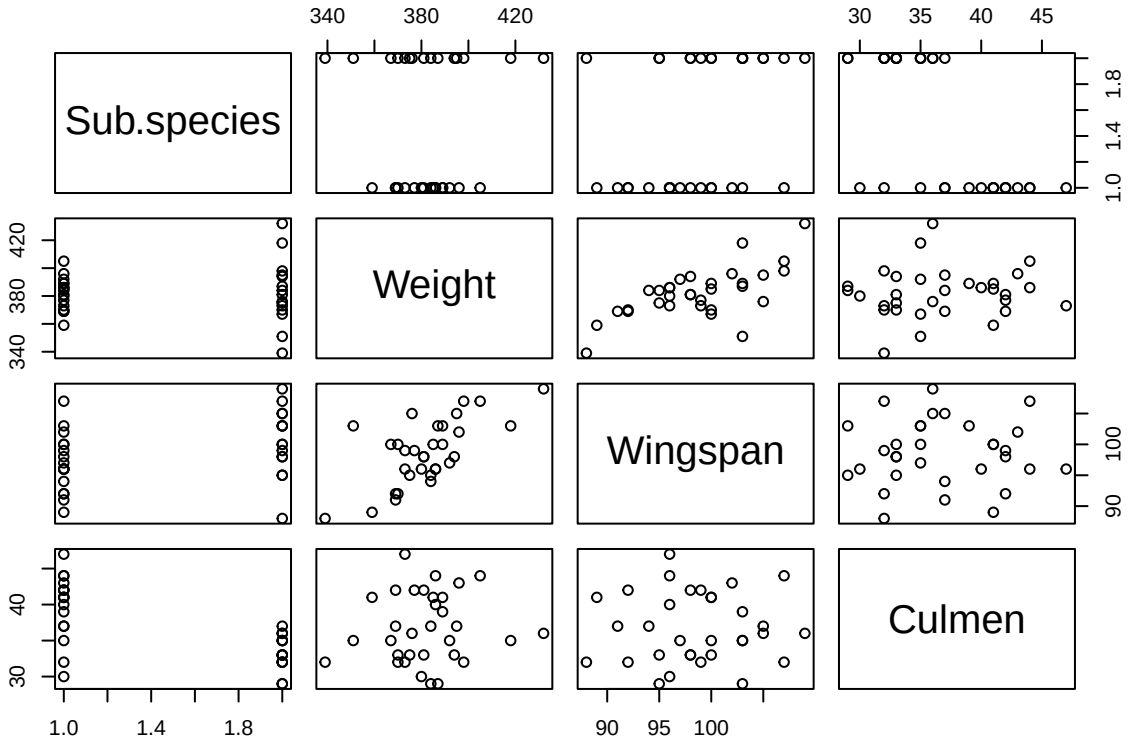
```r
# Question 2b Estimating 2006

site_e <- hist_data[, c("X", "Site.E")]


data_2005 <- site_e[site_e$X == 2005, "Site.E"]
data_2010 <- site_e[site_e$X == 2010, "Site.E"]


li_data_2006 <- data_2005 + (2006 - 2005) * (data_2010 - data_2005) / (2010 - 2005)

print(li_data_2006)
```

```
## [1] 67.4
```

```r
#Question 3a
head(meas_data)
```

```
##    Sub.species Weight Wingspan Culmen
## 1 Black-legged    405      107     44
## 2   Red-legged    351      103     35
## 3   Red-legged    394       98     33
## 4 Black-legged    389      100     41
## 5 Black-legged    384       94     37
## 6   Red-legged    395      105     37
```

```r
plot(meas_data)
```




```
summary(meas_data)
```

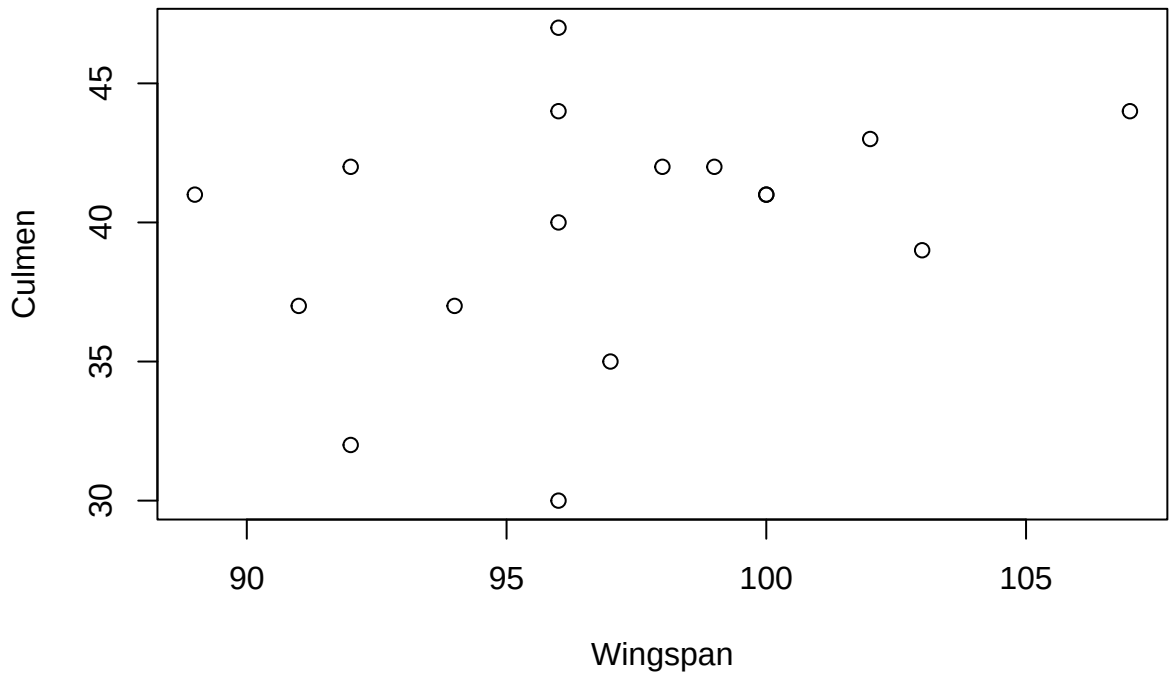
```
##  Sub.species            Weight         Wingspan         Culmen
##  Length:32          Min.   :339.0   Min.   : 88.00   Min.   :29.00
##  Class :character   1st Qu.:372.2   1st Qu.: 95.75   1st Qu.:33.00
##  Mode  :character   Median :382.5   Median : 98.50   Median :36.00
##                     Mean   :382.2   Mean   : 98.62   Mean   :36.78
##                     3rd Qu.:389.8   3rd Qu.:103.00   3rd Qu.:41.00
##                     Max.   :432.0   Max.   :109.00   Max.   :47.00
```

```
#Question 3b
#Correlation Test for "BLACK LEGGED" sub species
Sub_species1 <- subset(meas_data, Sub.species == 'Black-legged')
Sub_species1
```

```
##     Sub.species Weight Wingspan Culmen
## 1  Black-legged    405      107     44
## 4  Black-legged    389      100     41
## 5  Black-legged    384       94     37
## 7  Black-legged    369       91     37
## 8  Black-legged    392       97     35
## 9  Black-legged    386       96     40
## 11 Black-legged    377       99     42
## 13 Black-legged    396      102     43
## 17 Black-legged    381       98     42
```

```
## 18 Black-legged    385      100     41
## 19 Black-legged    359       89     41
## 20 Black-legged    389      103     39
## 22 Black-legged    380       96     30
## 23 Black-legged    373       96     47
## 24 Black-legged    370       92     32
## 26 Black-legged    369       92     42
## 29 Black-legged    386       96     44
```

```
plot(Sub_species1$Wingspan,Sub_species1$Culmen,xlab="Wingspan",ylab="Culmen")
```



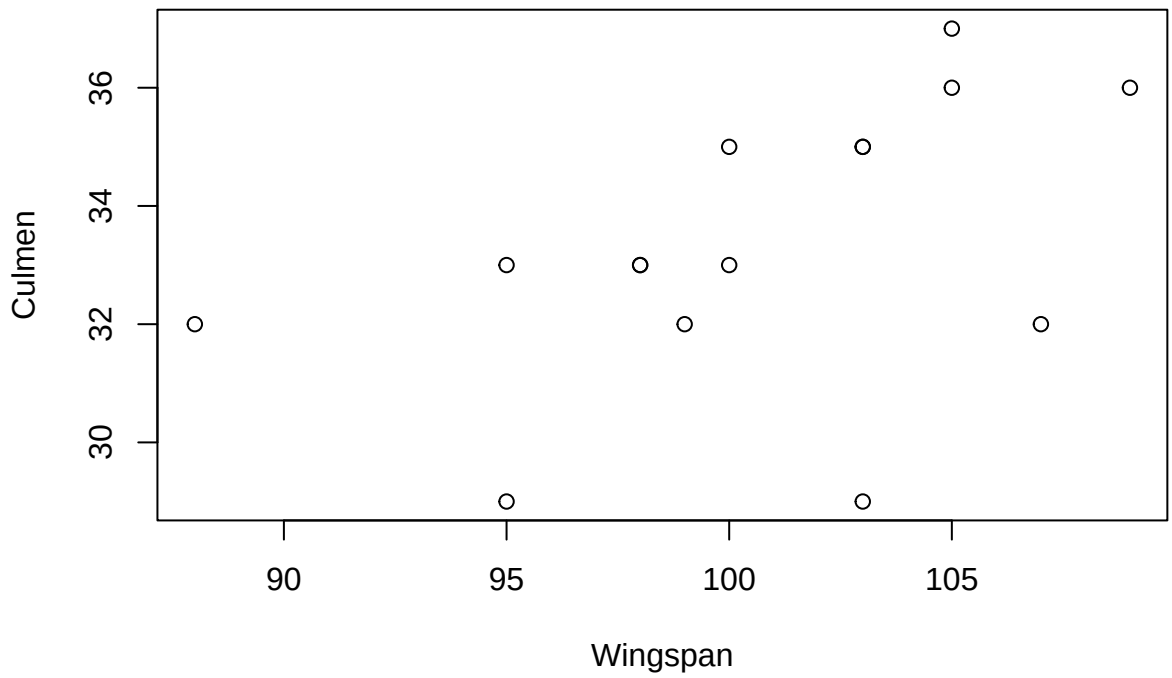

```
cor.test(Sub_species1$Wingspan,Sub_species1$Culmen)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Sub_species1$Wingspan and Sub_species1$Culmen
## t = 1.3563, df = 15, p-value = 0.1951
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1784845  0.6999630
## sample estimates:
##       cor
## 0.3305143
```

```r
#Question 3b
#Correlation Test for "RED LEGGED" sub species
Sub_species2 <- subset(meas_data, Sub.species == 'Red-legged')
Sub_species2
```

```
##    Sub.species Weight Wingspan Culmen
## 2   Red-legged    351      103     35
## 3   Red-legged    394       98     33
## 6   Red-legged    395      105     37
## 10  Red-legged    398      107     32
## 12  Red-legged    381       98     33
## 14  Red-legged    367      100     35
## 15  Red-legged    370      100     33
## 16  Red-legged    375       95     33
## 21  Red-legged    376      105     36
## 25  Red-legged    418      103     35
## 27  Red-legged    339       88     32
## 28  Red-legged    387      103     29
## 30  Red-legged    432      109     36
## 31  Red-legged    373       99     32
## 32  Red-legged    384       95     29
```

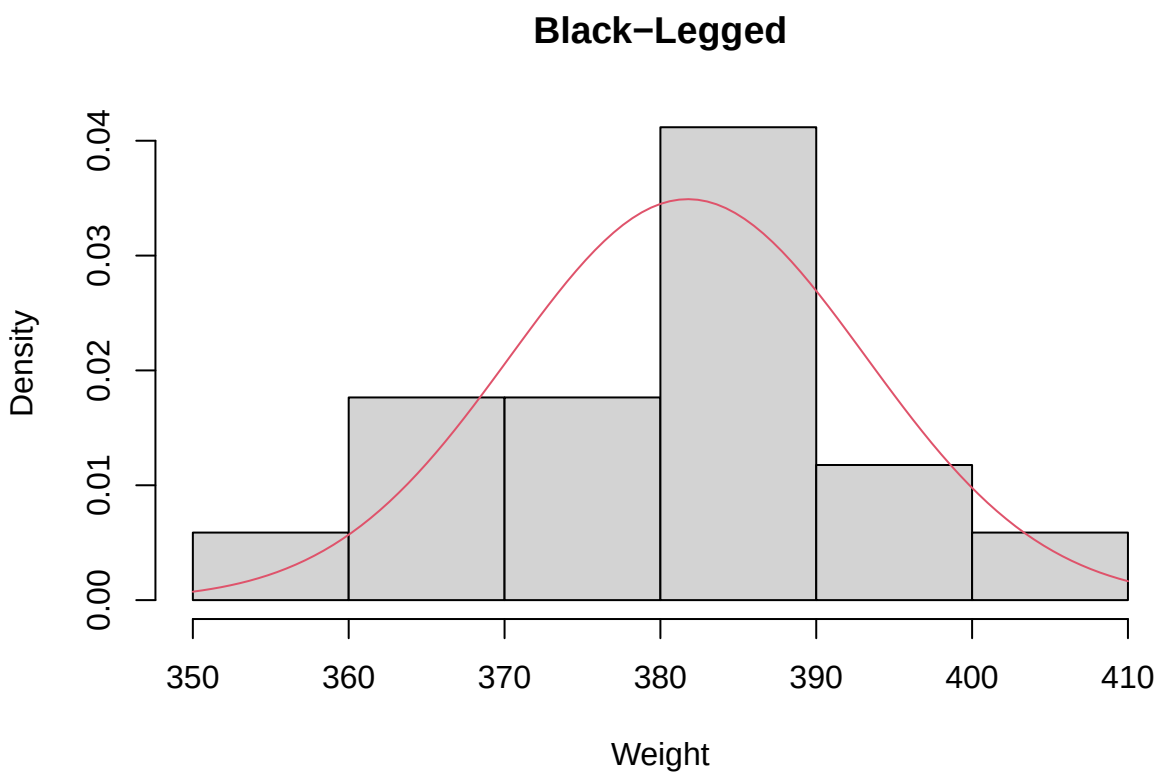
```r
plot(Sub_species2$Wingspan,Sub_species2$Culmen,xlab="Wingspan",ylab="Culmen")
```



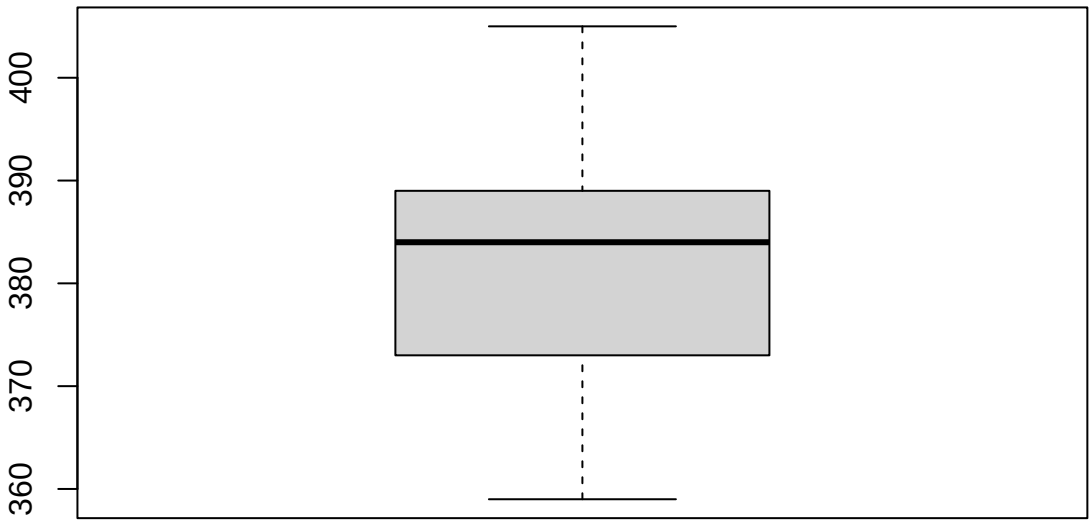
```r
cor.test(Sub_species2$Wingspan,Sub_species2$Culmen)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Sub_species2$Wingspan and Sub_species2$Culmen
## t = 1.9496, df = 13, p-value = 0.07314
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.04843176  0.79435179
## sample estimates:
##       cor
## 0.4756314
```

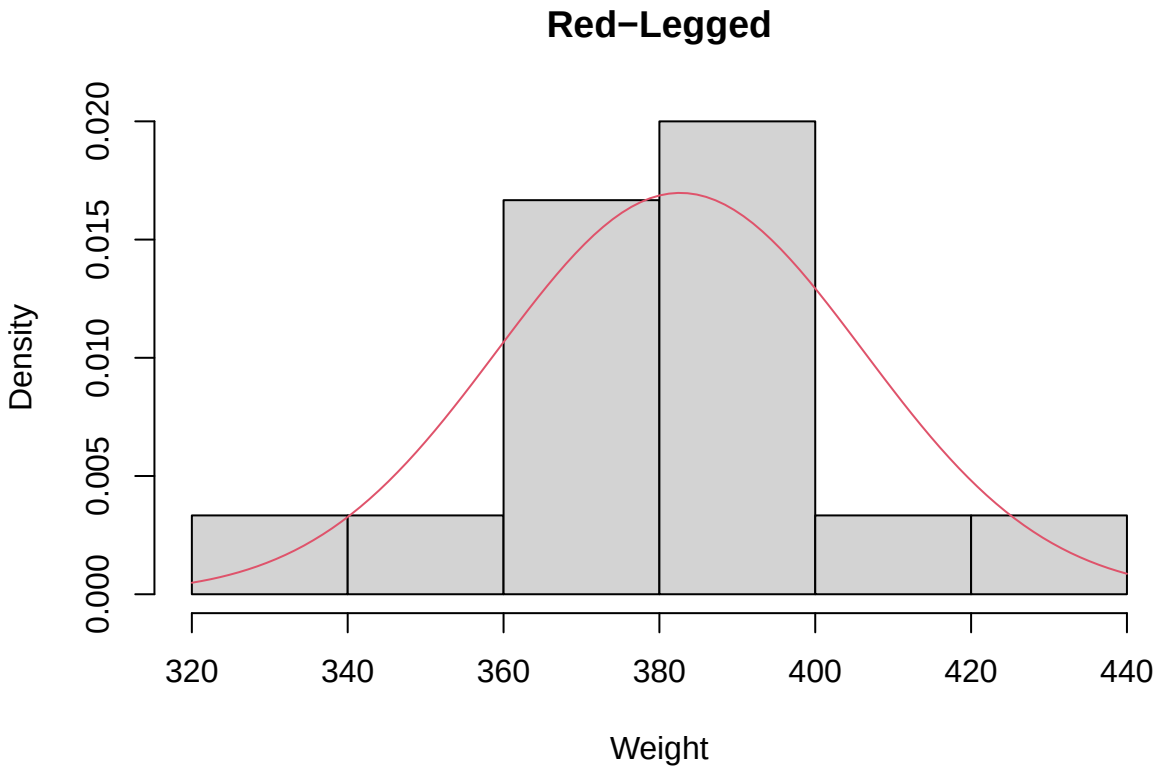
```r
#Question 3c
#Ftest to test if variances are equal
hist(Sub_species1$Weight, main = "Black-Legged", xlab = "Weight" ,freq = FALSE)
curve(dnorm(x,mean(Sub_species1$Weight),sd(Sub_species1$Weight)),col=2,add=TRUE)
```



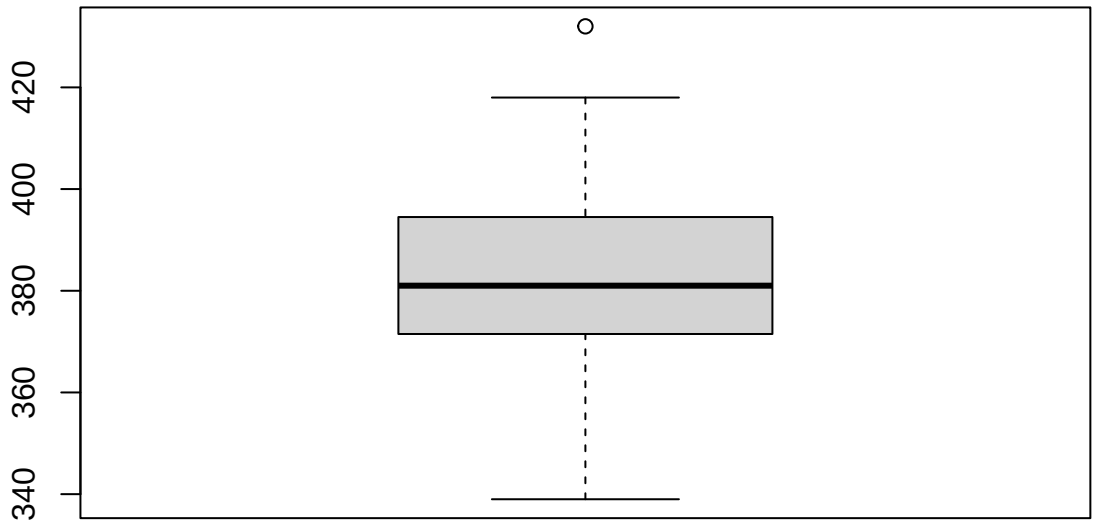

```r
boxplot(Sub_species1$Weight)
```




```
hist(Sub_species2$Weight, main = "Red-Legged", xlab = "Weight" ,freq = FALSE)
curve(dnorm(x,mean(Sub_species2$Weight),sd(Sub_species2$Weight)),col=2,add=TRUE)
```




```
boxplot(Sub_species2$Weight)
```




```
var.test(Sub_species1$Weight,Sub_species2$Weight)
```

```
##
##  F test to compare two variances
##
## data:  Sub_species1$Weight and Sub_species2$Weight
## F = 0.23637, num df = 16, denom df = 14, p-value = 0.007228
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.08085459 0.66585795
## sample estimates:
## ratio of variances
##          0.2363698
```

```
#welch t-Test to check for difference in weights in two species since theres a significant variance bet
welch_test_result <- t.test(Sub_species1$Weight,Sub_species2$Weight, var.equal = FALSE)
welch_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  Sub_species1$Weight and Sub_species2$Weight
## t = -0.1352, df = 19.699, p-value = 0.8938
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -14.83162  13.02770
## sample estimates:
## mean of x mean of y
##  381.7647  382.6667
```

```
# In summary, based on these results, there is no strong evidence to reject the null hypothesis that th
```

```
#Question 3d

manova_result <- manova(cbind(Weight, Wingspan, Culmen) ~ Sub.species, data = meas_data)

print(summary(manova_result))
```

```
##             Df  Pillai approx F num Df den Df    Pr(>F)
## Sub.species  1 0.62463   15.531      3     28 3.851e-06 ***
## Residuals   30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Question 4
head(loc_data)
```

```
##   Coast.direction sandeel Summer.temp cliff.height Breeding.pairs
## 1            West    1.45        23.3         3.54            165
## 2            East    2.47        23.9         4.24            367
## 3            East    0.65        21.8         4.00            271
## 4           North    1.93        19.8         4.56            504
## 5           North    1.51        20.3         4.27            350
## 6           South    1.62        23.0         4.06            280
```

```
model = lm(loc_data$Breeding.pairs~.,data = loc_data)
summary(model)
```

```
##
## Call:
## lm(formula = loc_data$Breeding.pairs ~ ., data = loc_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.673 -17.053  -7.547   9.319  97.321
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1017.633    138.201  -7.363 2.27e-07 ***
## Coast.directionNorth    24.278     22.748   1.067    0.297
## Coast.directionSouth    -1.450     24.730  -0.059    0.954
## Coast.directionWest      3.043     26.374   0.115    0.909
## sandeel                  7.446     10.063   0.740    0.467
## Summer.temp              4.605      4.566   1.009    0.324
## cliff.height           299.219     20.210  14.806 6.38e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 40.11 on 22 degrees of freedom
## Multiple R-squared:  0.9349, Adjusted R-squared:  0.9171
## F-statistic: 52.64 on 6 and 22 DF,  p-value: 6.158e-12
```

```
bestmodel = step(model)
```

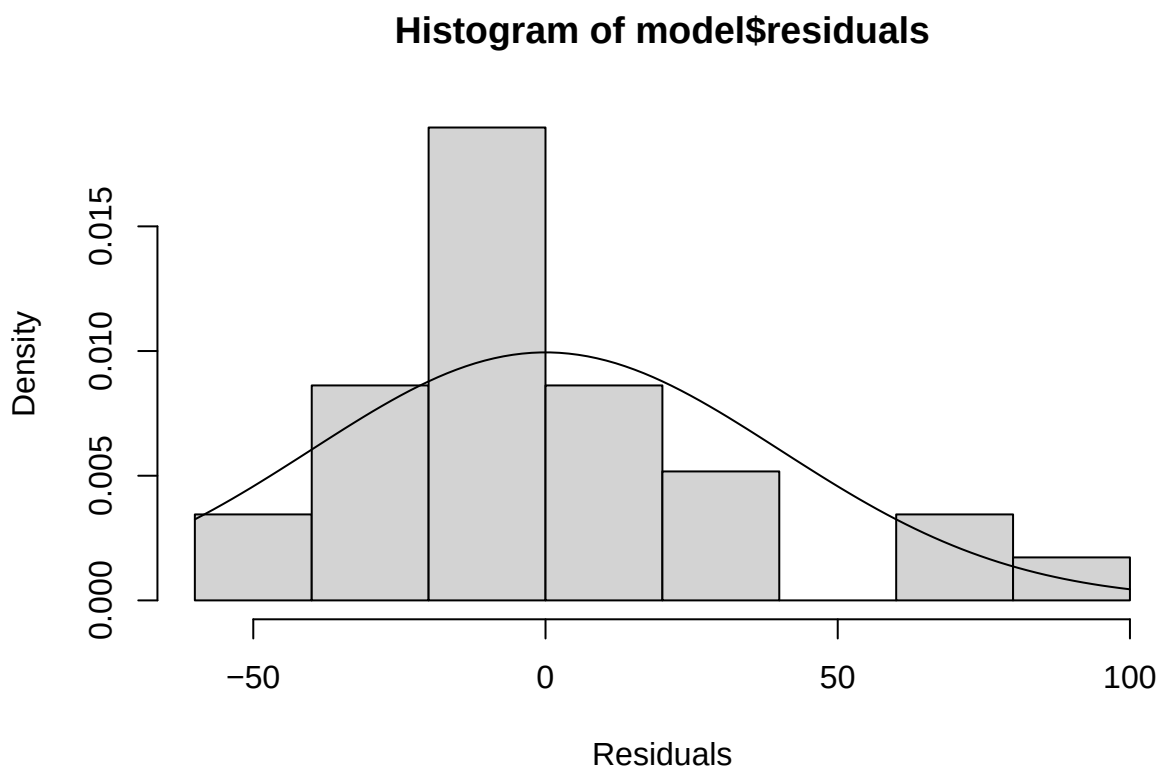
```
## Start:  AIC=220.1
## loc_data$Breeding.pairs ~ Coast.direction + sandeel + Summer.temp +
##     cliff.height
##
##                   Df Sum of Sq    RSS    AIC
## - Coast.direction  3      3136  38528 216.56
## - sandeel          1       881  36273 218.81
## - Summer.temp      1      1636  37028 219.41
## <none>                          35392 220.10
## - cliff.height     1    352650 388042 287.55
##
## Step:  AIC=216.56
## loc_data$Breeding.pairs ~ sandeel + Summer.temp + cliff.height
##
##                Df Sum of Sq    RSS    AIC
## - Summer.temp   1      1227  39755 215.47
## - sandeel       1      1457  39985 215.64
## <none>                        38528 216.56
## - cliff.height  1    489817 528345 290.50
##
## Step:  AIC=215.47
## loc_data$Breeding.pairs ~ sandeel + cliff.height
##
##                Df Sum of Sq    RSS    AIC
## - sandeel       1      1383  41138 214.46
## <none>                        39755 215.47
## - cliff.height  1    495740 535495 288.89
##
## Step:  AIC=214.46
## loc_data$Breeding.pairs ~ cliff.height
##
##                Df Sum of Sq    RSS    AIC
## <none>                        41138 214.46
## - cliff.height  1    502358 543495 287.32
```

```
summary(bestmodel)
```

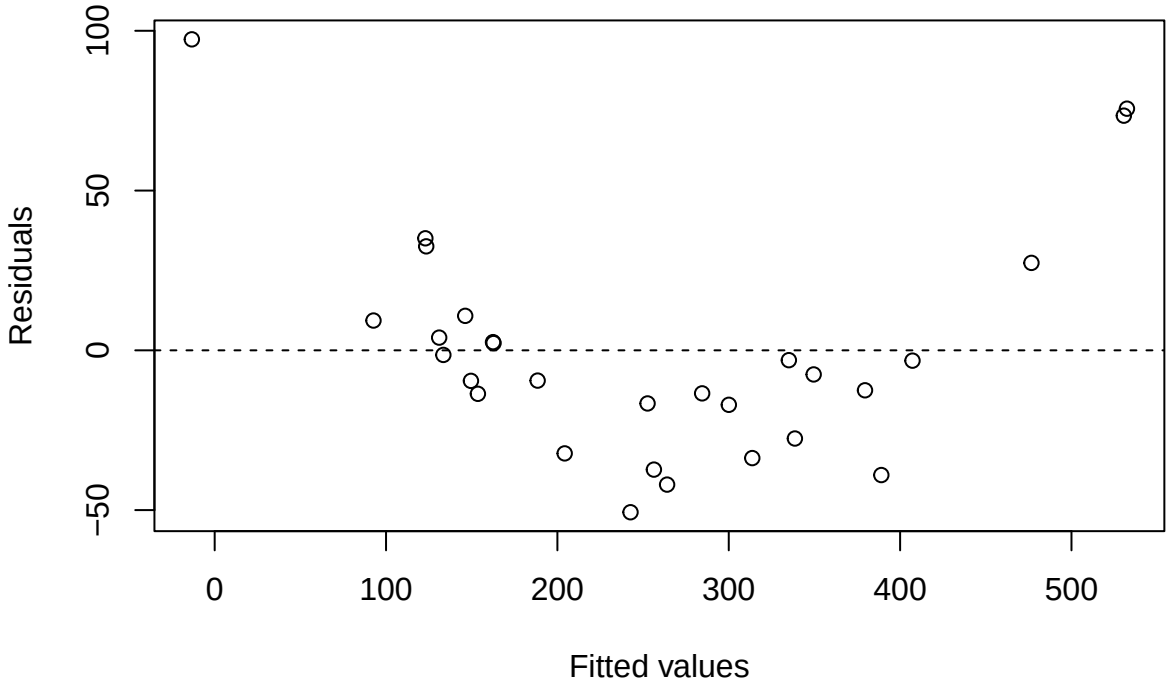
```
##
## Call:
## lm(formula = loc_data$Breeding.pairs ~ cliff.height, data = loc_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.246 -19.311  -7.702  20.168  98.612
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -889.53      63.48  -14.01  6.6e-14 ***
## cliff.height   298.91      16.46   18.16  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.03 on 27 degrees of freedom
## Multiple R-squared:  0.9243, Adjusted R-squared:  0.9215
## F-statistic: 329.7 on 1 and 27 DF,  p-value: < 2.2e-16
```

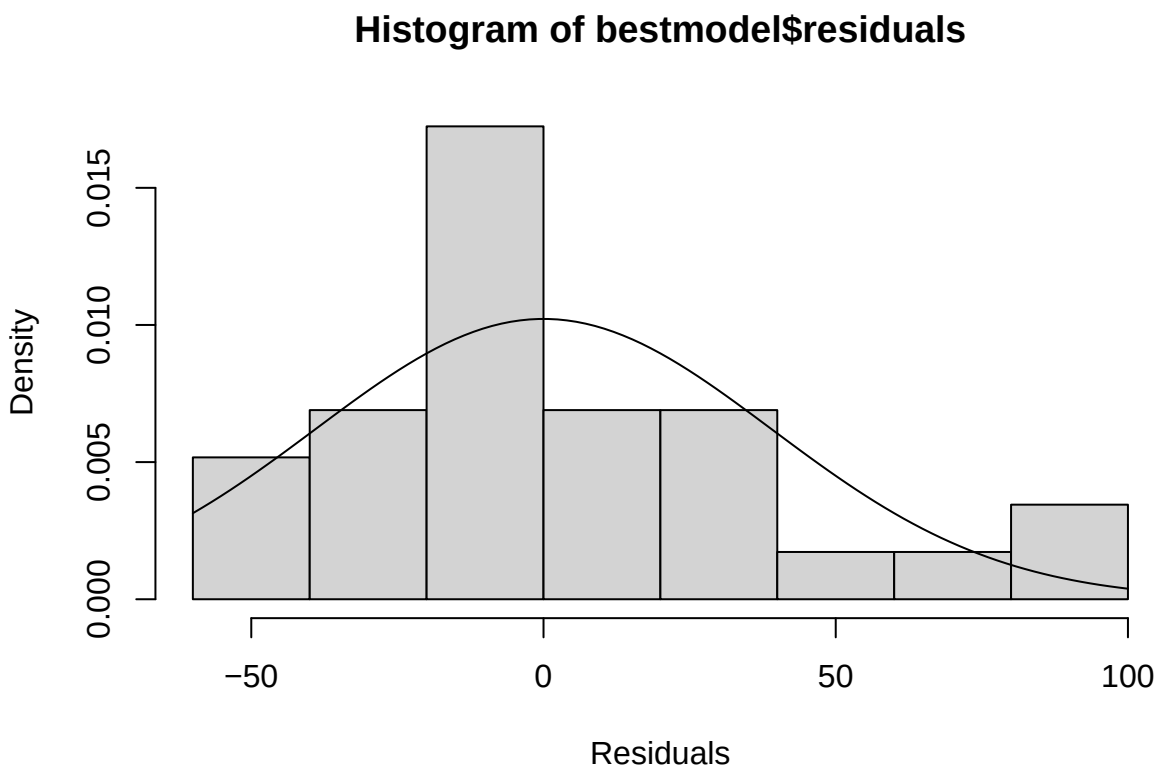
```
hist(model$residuals,xlab="Residuals",freq=FALSE)
curve(dnorm(x,0,summary(model)$sigma),lty=1,add=TRUE)
```



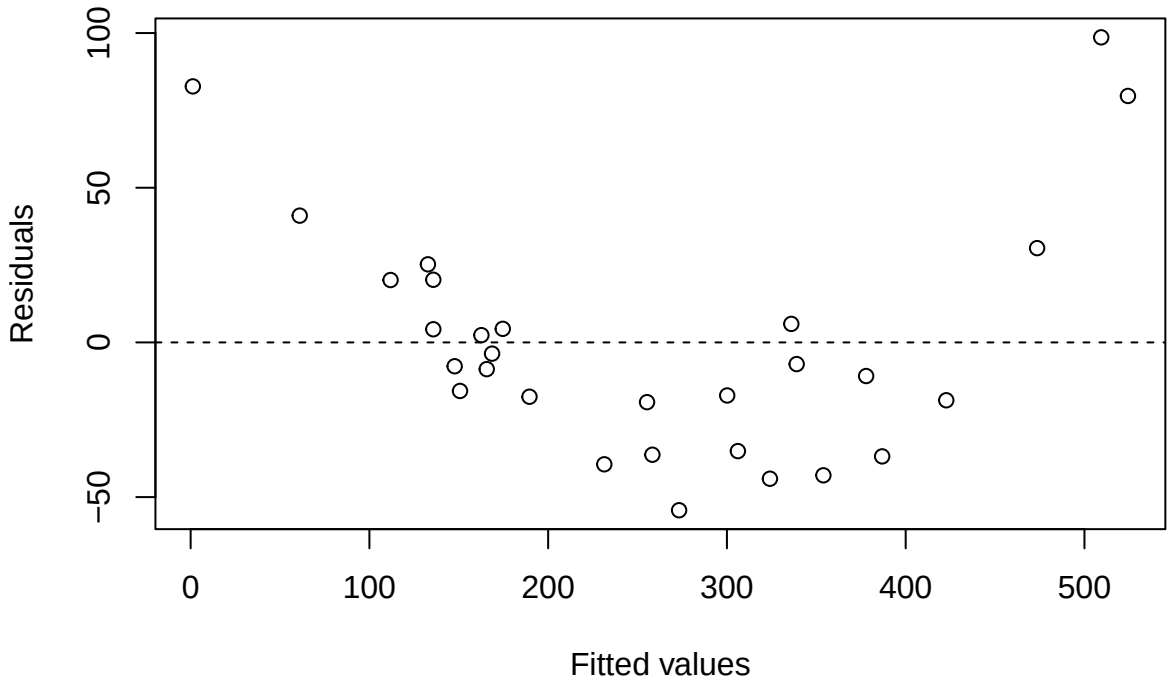

```
plot(model$fitted.values,model$residuals,xlab="Fitted values",ylab="Residuals")
abline(h=0,lty=2)
```




```
hist(bestmodel$residuals,xlab="Residuals",freq=FALSE)
curve(dnorm(x,0,summary(bestmodel)$sigma),lty=1,add=TRUE)
```



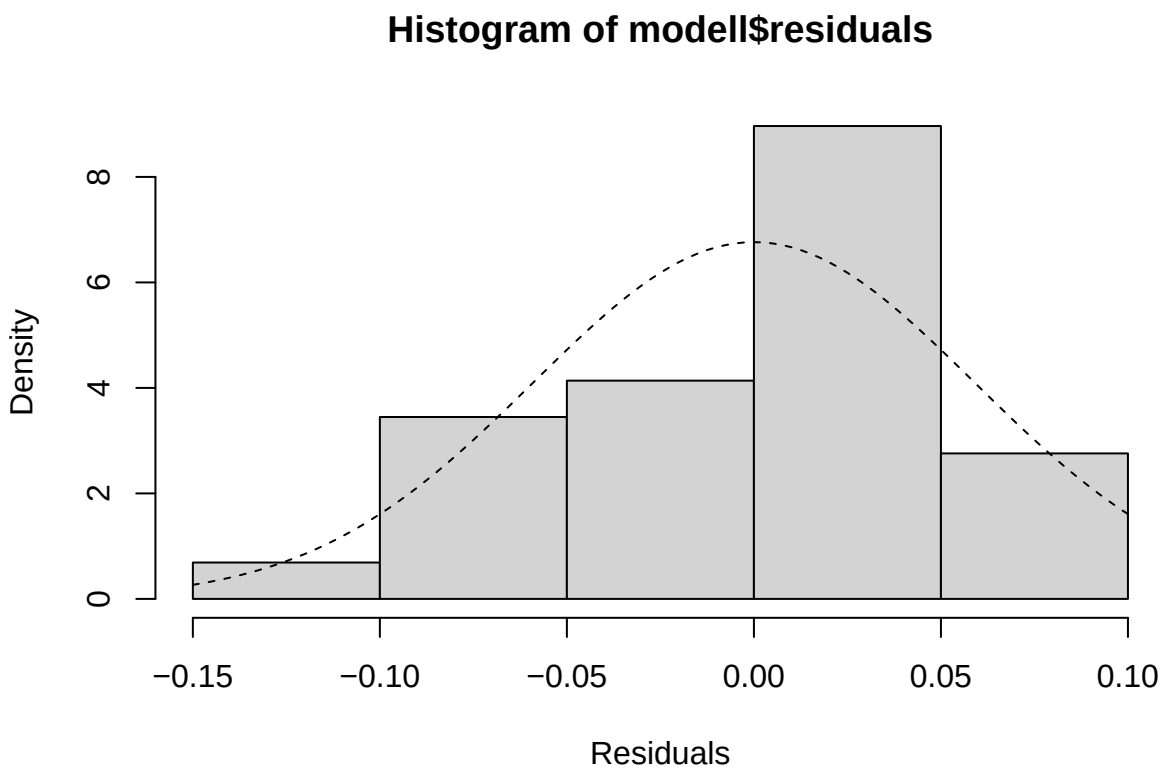

```
plot(bestmodel$fitted.values,bestmodel$residuals,xlab="Fitted values",ylab="Residuals")
abline(h=0,lty=2)
```



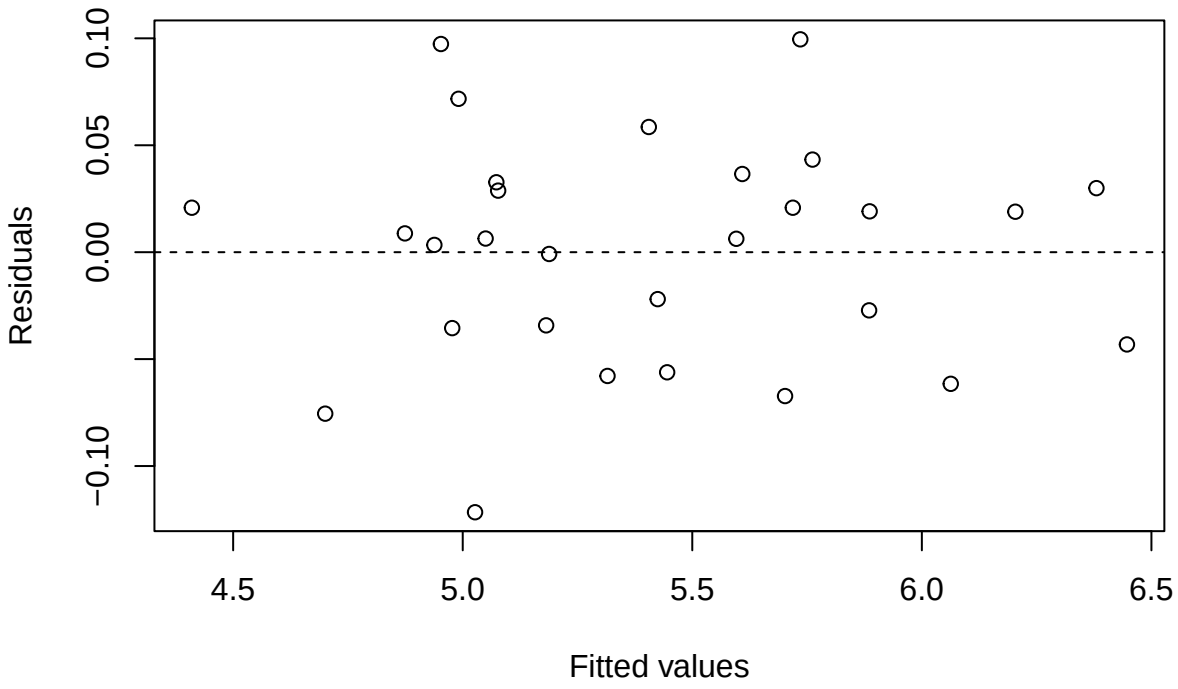

```
modell = lm(log(loc_data$Breeding.pairs)~.,data = loc_data)
summary(modell)
```

```
##
## Call:
## lm(formula = log(loc_data$Breeding.pairs) ~ ., data = loc_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.121602 -0.035505  0.006377  0.029950  0.099520
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.669014   0.203226   3.292  0.00332 **
## Coast.directionNorth  0.014891   0.033452   0.445  0.66056
## Coast.directionSouth  0.027646   0.036365   0.760  0.45519
## Coast.directionWest  -0.010085   0.038783  -0.260  0.79726
## sandeel              -0.010364   0.014798  -0.700  0.49104
## Summer.temp           0.016732   0.006714   2.492  0.02073 *
## cliff.height          1.142206   0.029719  38.434  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05898 on 22 degrees of freedom
## Multiple R-squared:  0.9895, Adjusted R-squared:  0.9866
## F-statistic: 345.3 on 6 and 22 DF,  p-value: < 2.2e-16
```

```r
hist(modell$residuals,xlab="Residuals",freq=FALSE)
curve(dnorm(x,0,summary(modell)$sigma),lty=2,add=TRUE)
```




```r
plot(modell$fitted.values,modell$residuals,xlab="Fitted values",ylab="Residuals")
abline(h=0,lty=2)
```




```
pred = predict(bestmodel,newdata=data.frame(Coast.direction = 'South', sandeel = 1.36, Summer.temp = 23
pred
```

```
##        fit      lwr     upr
## 1 303.1373 250.8686 355.406
```

```
predc = predict(bestmodel,newdata=data.frame(Coast.direction = 'South', sandeel = 1.36, Summer.temp = 23
predc
```

```
##        fit      lwr      upr
## 1 303.1373 293.0111 313.2636
```