Final Project: Fallen Black Cherry Trees

Ayushi Singh and Linh Hoang

18/04/2022

Introduction and Research Questions

The purpose of this report is to look into fallen black cherry tree data and answer the following two research questions:

1. What is the relationship between volume and diameter of felled black cherry trees?
2. What is the true population mean height of fallen black cherry trees?

To answer these questions, the sample data under the name "trees" from the Open dataset in R was used. This data set contains data from 31 black cherry trees, with each tree having measurements in the variables girth, height and volume. Girth records the diameter measurement (in inches) of a tree at 4 feet 6 inches above the ground. Height records the height measurement (in feet) of a tree. Finally, Volume records the measured volume of timber produced (in cubic feet) by a tree. As our first research question is looking into any possible correlation between volume and diameter of fallen black cherry trees, the Volume and Girth measurements will be used for this question. Meanwhile, the second research question focuses entirely on the true population mean height of fallen cherry trees. Thus, for the second question, we will be using the Height values from the dataset, assuming they are representative of the population of fallen black cherry trees.

We would like to note that for our report, we have converted the values from imperial to metric. This is mainly to make it easier to understand any significance in values found for our readers (who are likely to be be more used to the metric system). As such, inches will be converted to centimetres, feet to metres, and cubic feet to cubic metres. The process gone for this conversion can be found within the appendix.

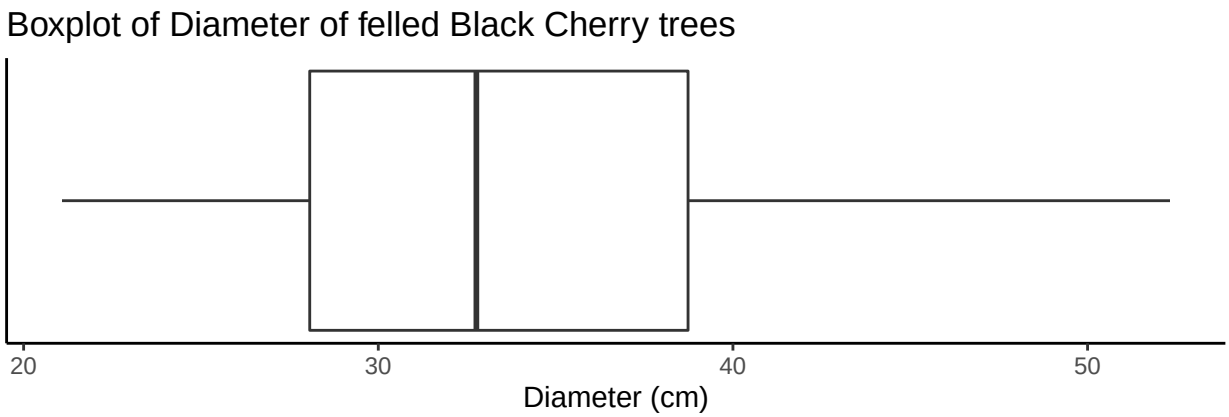
Exploratory Data Analysis

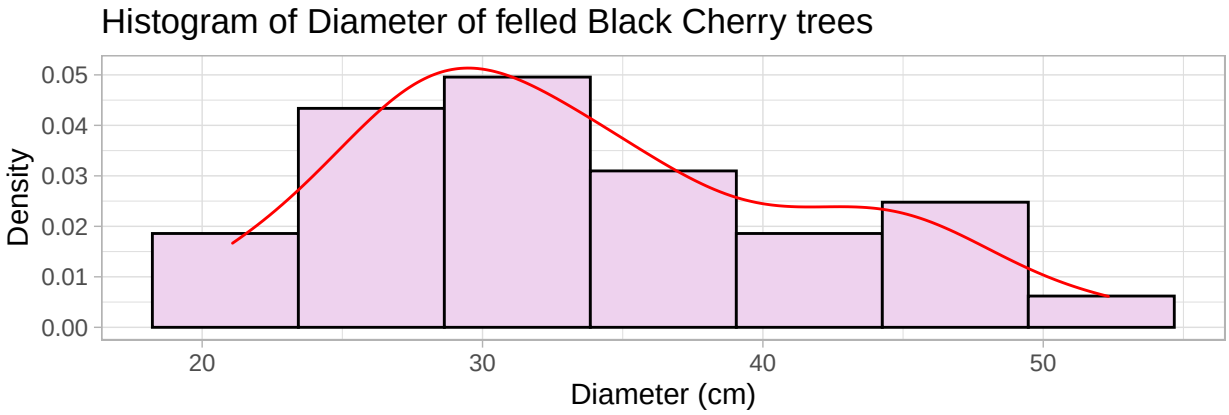
This sections will look into some initial data analysis for the data set. Calculations and graphical displays shown were done in R and can be found in the appendix.

Five Summary Statistics

	Diameter (cm)	Height(m)	Volume(m^3)
Minimum	21.08	19.20	0.2888
First Quartile	28.07	21.95	0.5493
Median	32.77	23.16	0.6853
Mean	33.65	23.16	0.8543
Third Quartile	38.73	24.38	1.0562
Maximum	52.32	26.52	2.1804

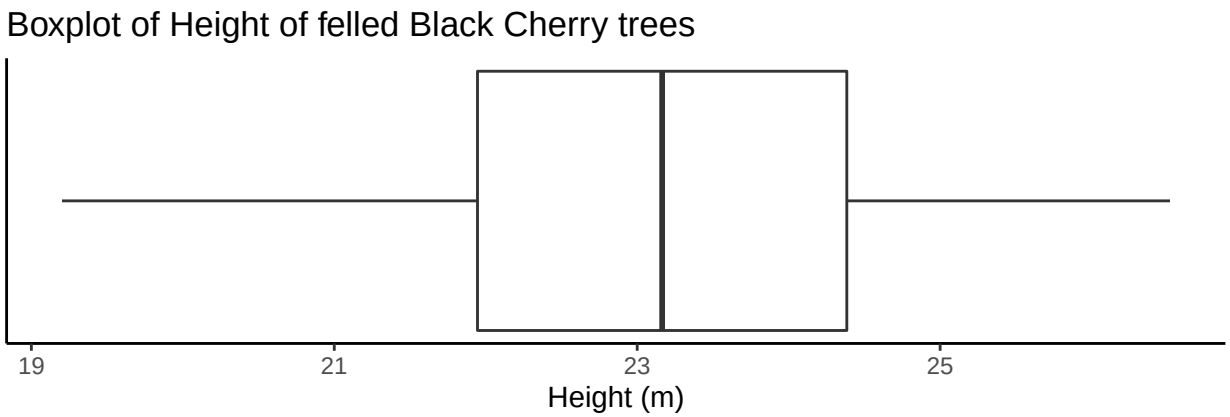
Distribution of the Data for Diameter

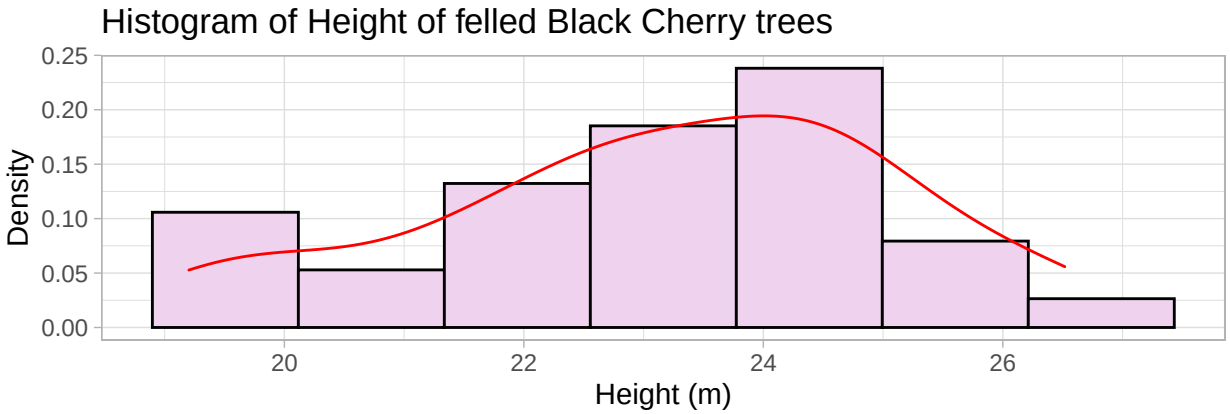




The box plot for diameter indicates that Q3 is located just under 40 cm and Q1 is located just under 30 cm, indicating that 50% of values are between 30 and 40 cm. The diameter histogram indicates a bimodal shape, with bumps at 30 and 45 cm.

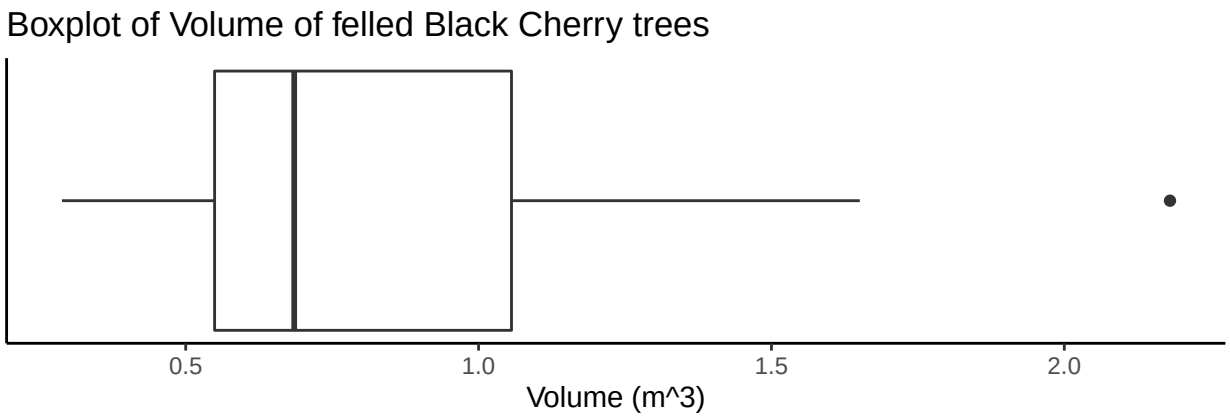
Distribution of the Data for Height

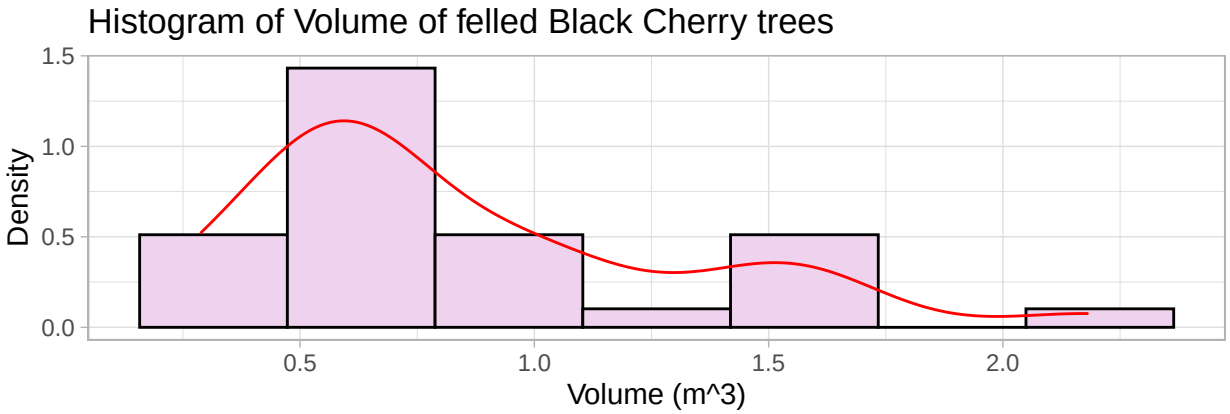




The box plot for height has its third quartile at around 24 meters, indicating 75% of the sample heights have a value below 24 meters. The data has a singular peak at 24 meters. The Kernel density estimation (red line) has a roughly normal shape, which possibly indicates that the height data follows a normal distribution function.

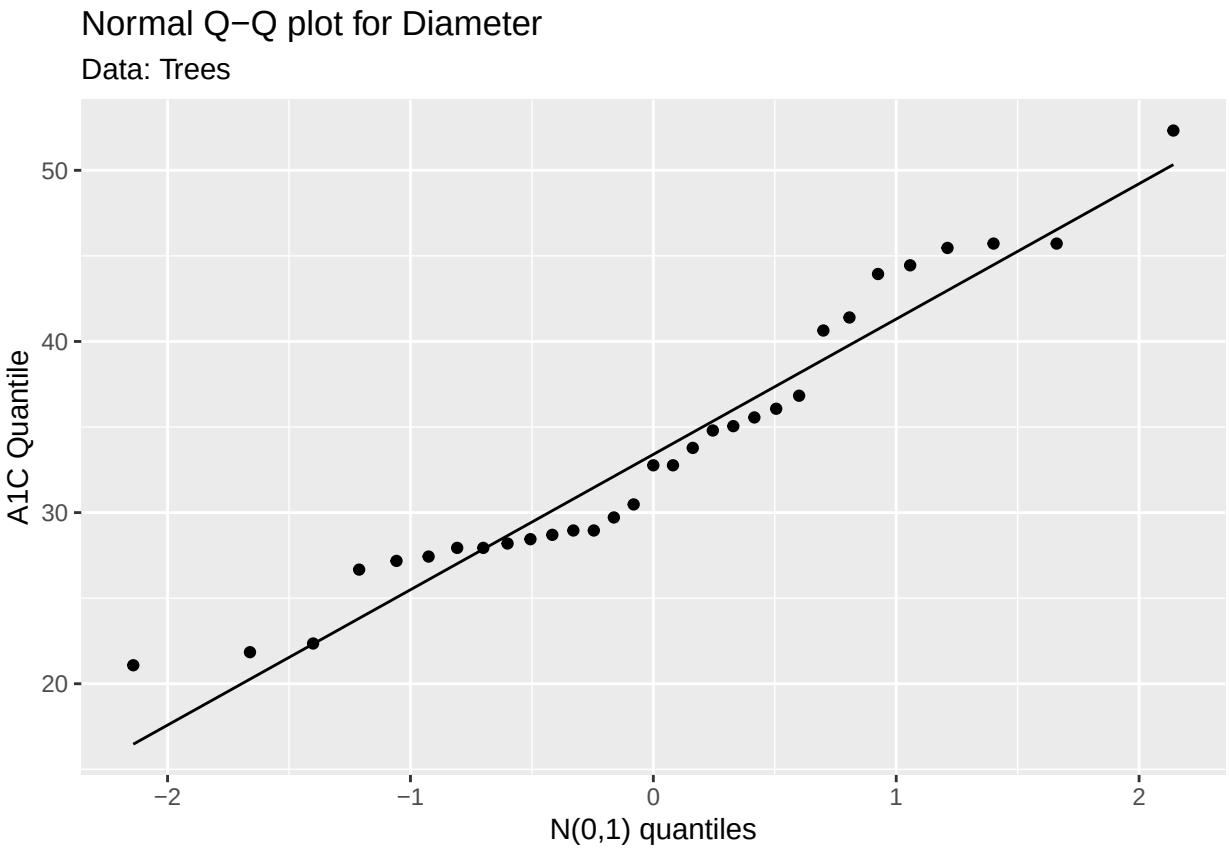
Distribution of the Data for Volume

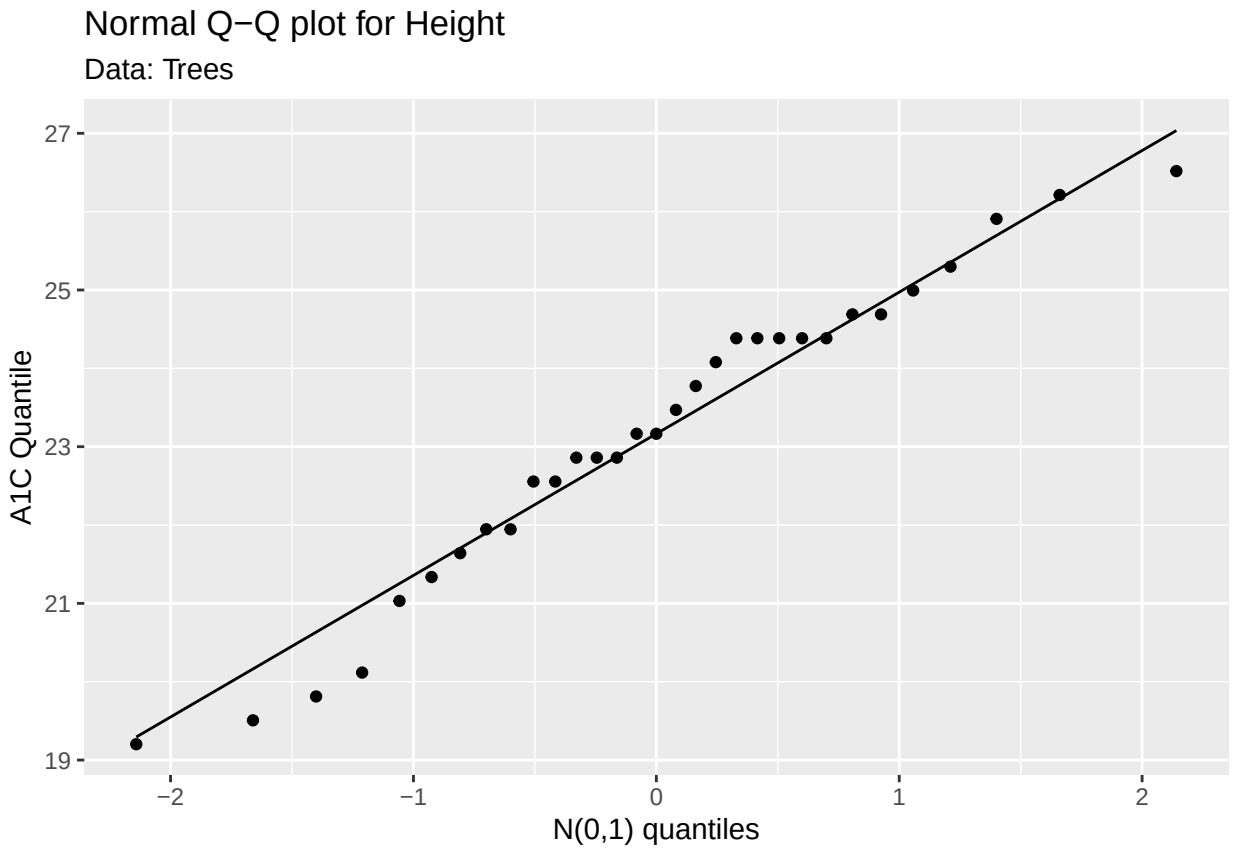


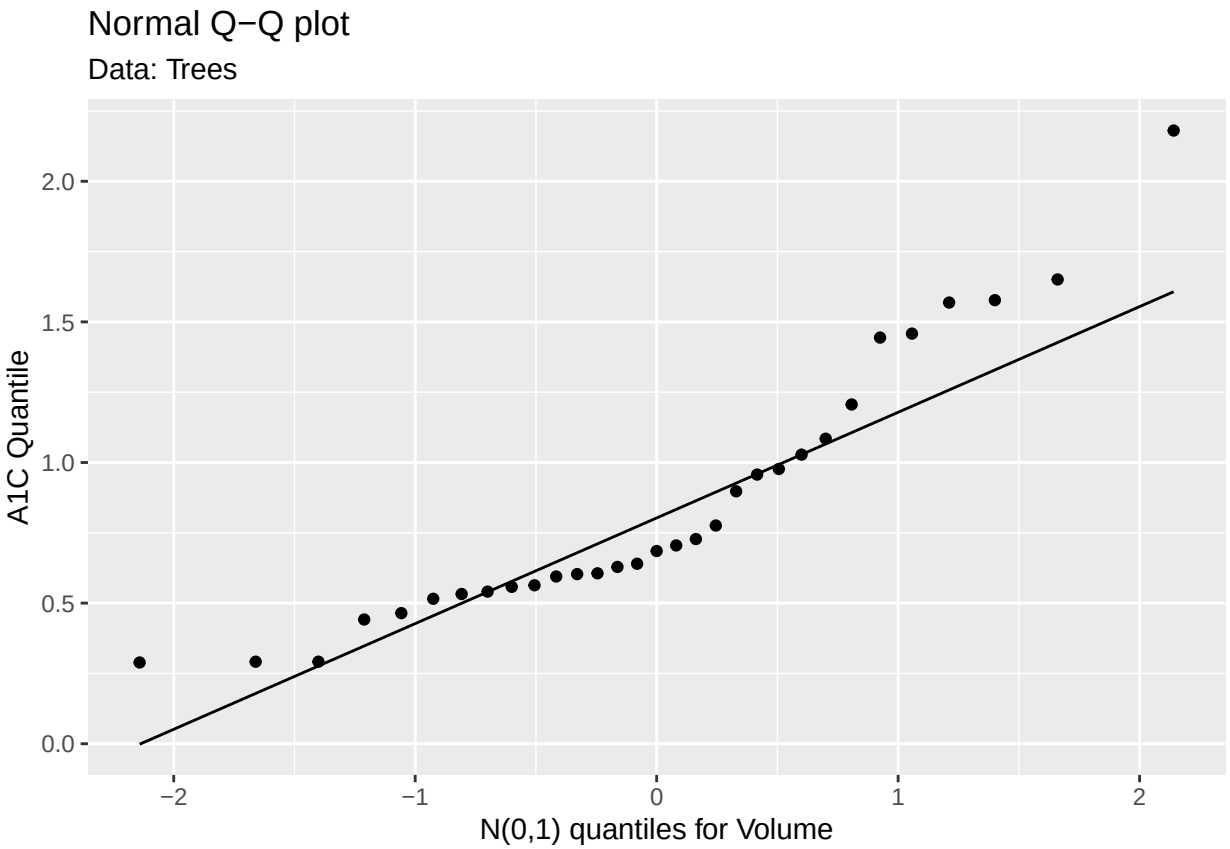


We can see that Q3 is just above 1 m^3 in the box plot for volume, which tells us that 75% of the data is below 1 m^3. We can also see that there is 1 outlier, which is at about 2.4 m^3. This outlier is also seen in the histogram as there is a "gap" between the last two bins. The data for volume is bimodal, with "bumps" at around 0.5 and 1.5 m^3.

Since the distributions were still ambiguous, we performed a QQ plot test for all 3 types of data to validate the distribution types. If the data were normally distributed, the graphical display of the QQ plot should have the plotted points fall fairly close to the line. We created three QQ plots, each depicting the Diameter, Height and Volume parameters.


Normal Q−Q plot for Diameter
Data: Trees
A1C Quantile
50
40
30
20
−2
−1
0
1
2
N(0,1) quantiles


Normal Q−Q plot for Height
Data: Trees
A1C Quantile
N(0,1) quantiles
27
25
23
21
19
−2
−1
0
1
2



The trend seemed to be fairly normally distributed for Diameter, Height and Volume. The data points for all 3 plots were fairly well situated with the linear line with higher variation on the tails (This is especially the case with Volume).

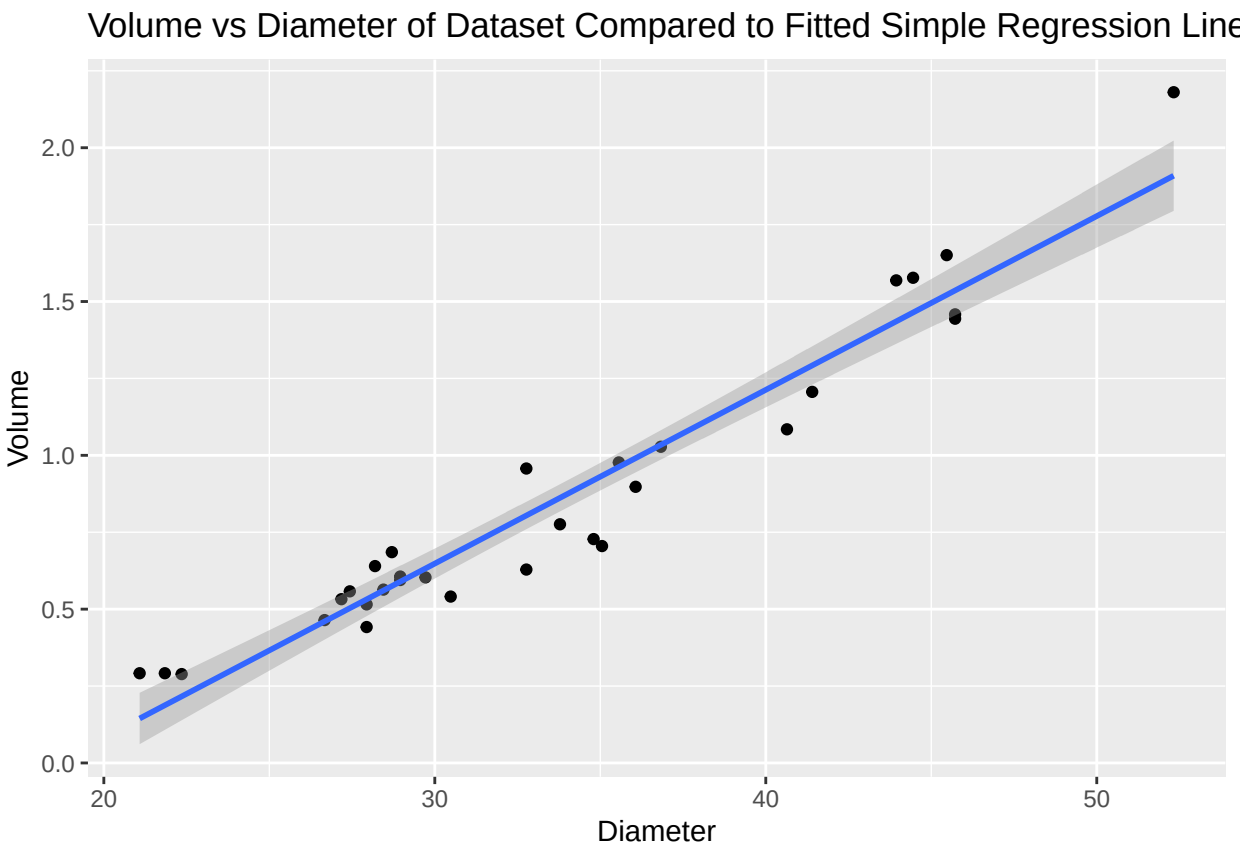
Research Question 1: Simple Linear Regression and Findings

(Note the code used to get these results can be found in the appendix).
Since the first research question asks for the relation between the volume and diameter of fallen black cherry trees, it would make sense to use simple linear regression. Simple linear regression looks for possible correlation between two values (if A increases, does B increase/decrease/remain unaffected?). Simple linear regression fits a linear equation to the data, with the linear equation being $y = \beta_0 + \beta_1 x$, where y is expected value for volume, β_0 is the y-intercept of the line (what we may expect the volume to be given a diameter of 0), β_1 is the slope of the line (rate the volume increases/decreases) and x is the value for the diameter we are finding a volume for.

Using R, we fitted a line to the data's volume and diameter values. The fitted line had the following estimated parameters:
$\beta_0 = -1.04612$ $\beta_1 = 0.05648$
Taking these estimations to reflect the true regression line for this data and putting these values into our linear equation, we get:
$V = -1.04612 + 0.05648d$ where V = volume of a fallen black cherry tree and d = diameter of a fallen black cherry tree.

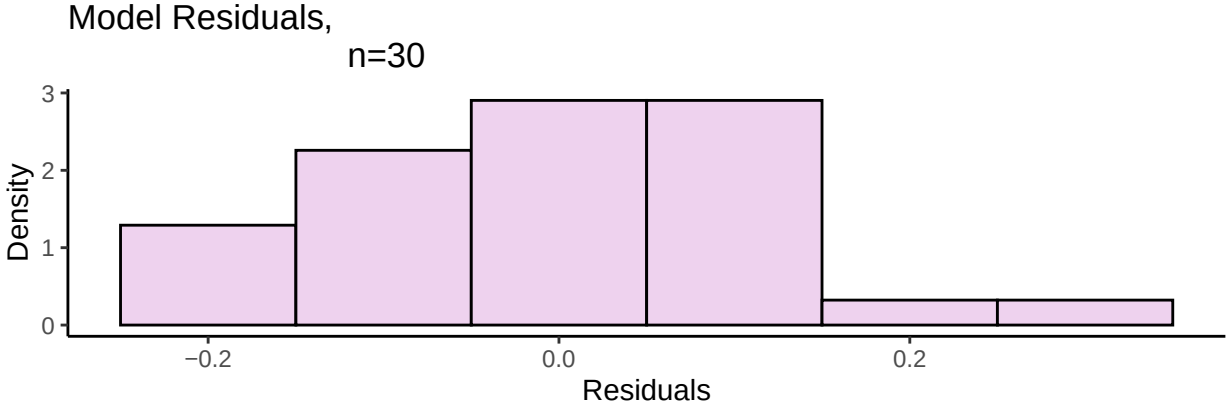
Graphing our linear regression line against the data set:

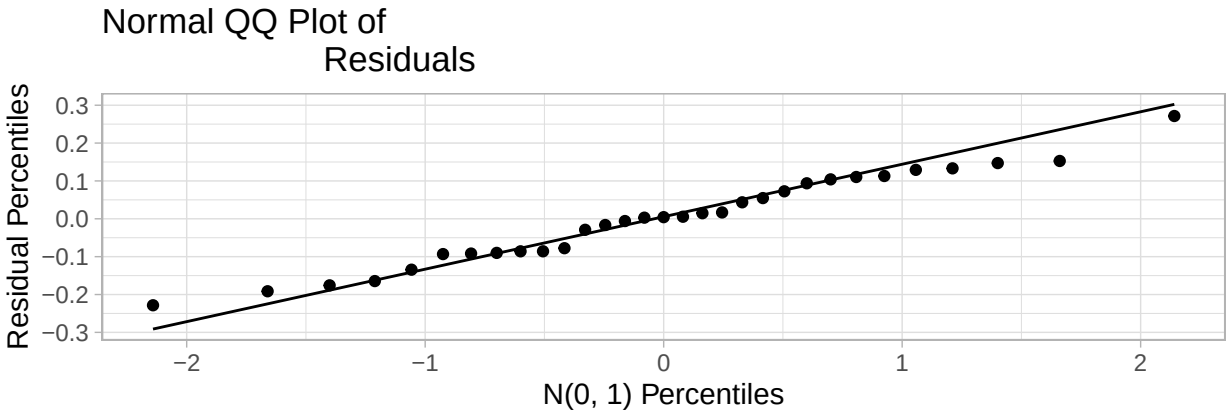
```
## `geom_smooth()` using formula 'y ~ x'
```

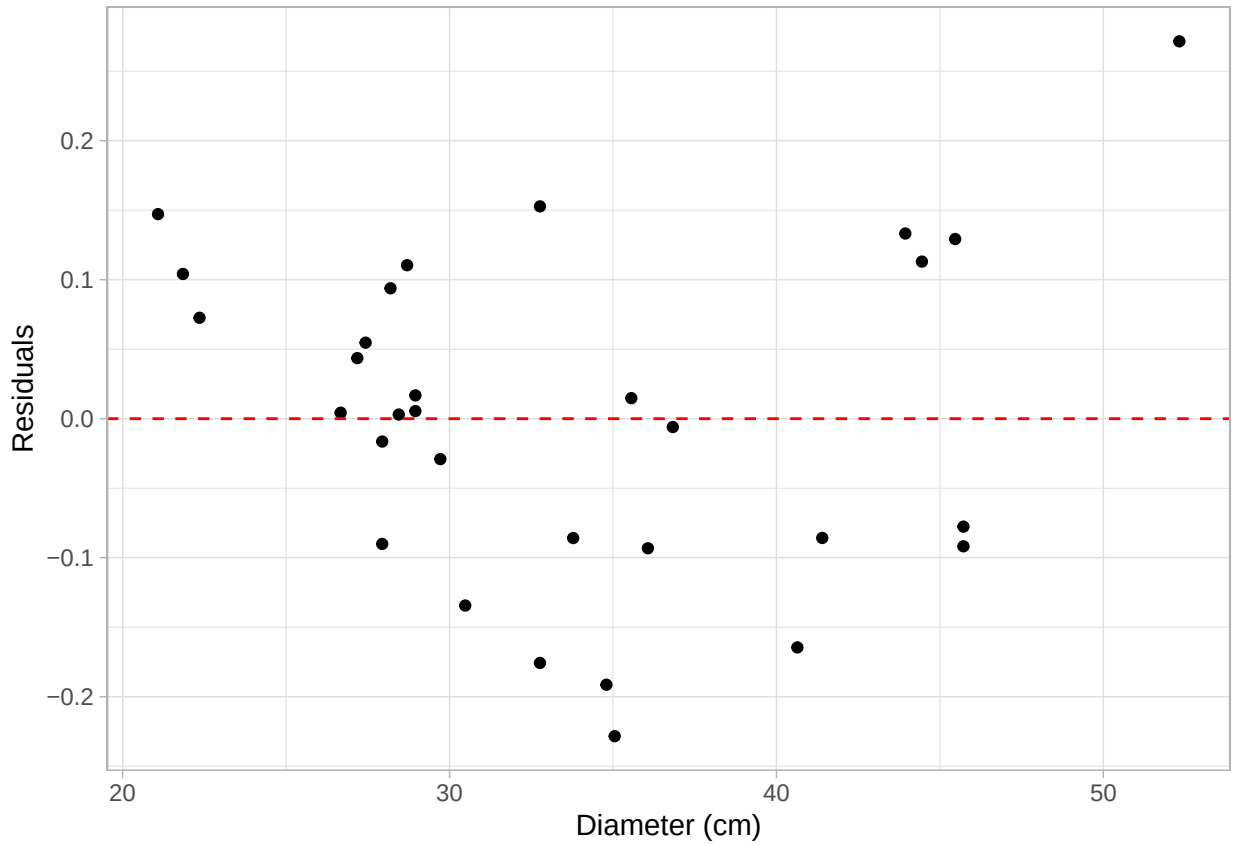


We also will need to analyze the residuals (the differences between the expected values from the regression line and the actual values from our data set) to see if this regression line with the estimated parameters are valid (and follow the needed assumptions).

Here are the graphs for analyzing the residuals and what they mean for our regression line:








```
## [1] 2.352019e-18
```

We also found the residuals to have a mean of $2.352019 * 10^{-18}$, which is small enough where we can round it down to 0.

From our regression line equation, we see the volume of timber from a fallen black cherry tree increases by at a rate of 0.05648 m^3 per cm, meaning there is a positive correlation between volume and diameter from this line (the volume of timber is related to the diameter of the tree with an increase of diameter equaling to an increase in volume). Our regression line and data points graph indicates that many of the data points from our data set are quite close to its expected value using the regression line. We note from our residual analysis, that the assumption where the mean of the residuals is 0 does hold. While the QQ plot does seem to indicate points at around the ends of the percentiles violate normality (they are quite a distance away from the line), the number of points in that area are few and the histogram indicates they could be potential outliers in an otherwise normal distribution. So, we believe the normality assumption has not been violated. The independence assumption of does seem to be true from our plot of the residuals vs diameter, as there doesn't appear to be an immediate pattern from our points. The constant variance assumption does seem to hold as well, as we can see the points do not seem to increase or decrease based on where we are in diameter. We do note that variance around the center of the plot (where diameter is 35 cm) does appear somewhat larger and there is a point in the top right corner with a large variance. The larger variance around the center could be due normal distribution having more data points in that area (therefore introducing a higher chance of finding differences in their values compared to the mean). The one really large point could be seen as a potential outlier that can be discarded, as seen with how we treated that point in the QQ plot as an outlier as well. Since all assumptions hold, we can conclude that our regression line with the estimated parameters do reflect the true relation of the volume and diameter of fallen black cherry trees.

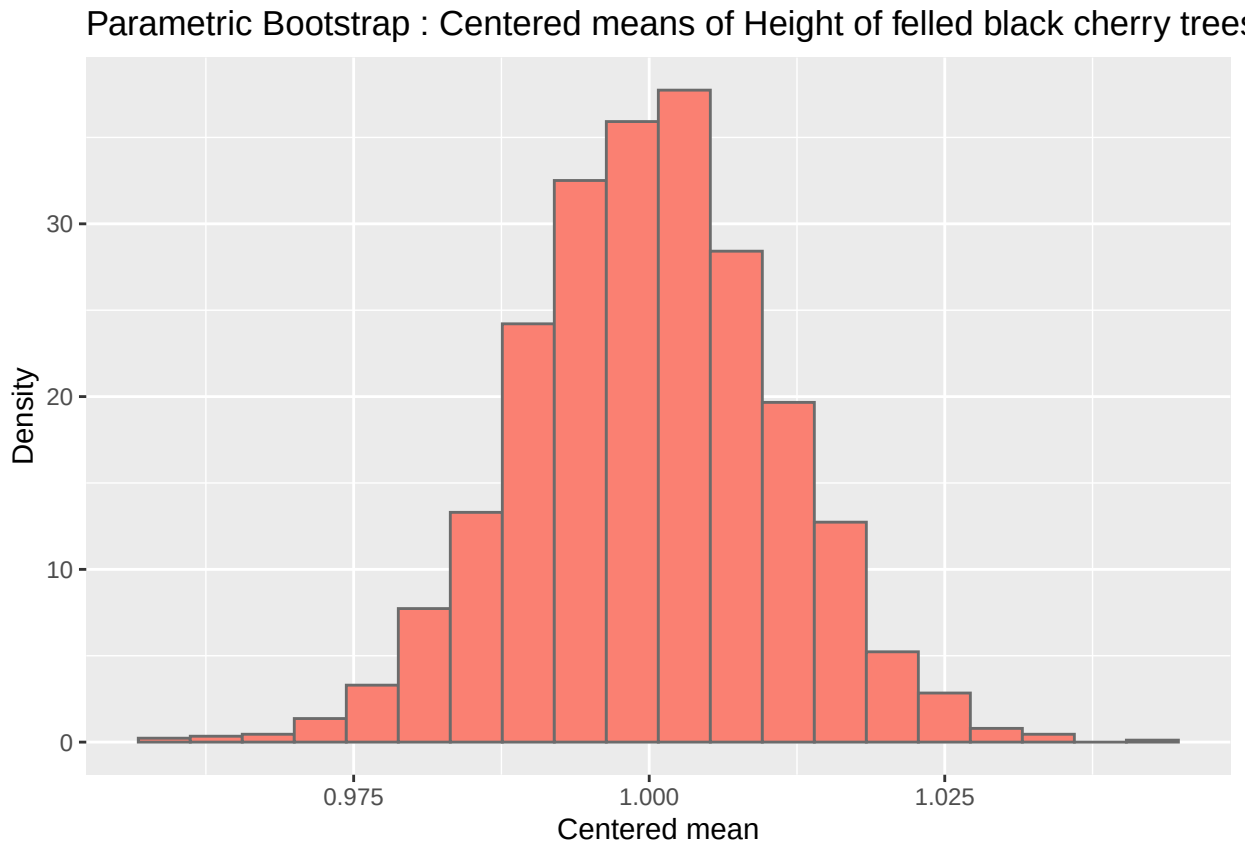
Research Question 1: Conclusion

Based on our findings, we were able to conclude that our linear regression model is valid such that any conclusions drawn are likely to accurately reflect on the true answer to the relationship. We can see from graph of the fitted line and the data points that, the regression line has formed a position relation between the volume of timber and the diameter of fallen black cherry trees. This means an increase in diameter would lead to an increase in volume. As such, we conclude that there is strong evidence that tells us there is likely a relationship between the volume and diameter of fallen black cherry trees, and that this relation is positive.

Our conclusion, however, is based on the decision that there were outliers near the minimum and maximum diameters and they could be ignored. Taking in more sample data values may indicate that these values are not outliers and thus would cause multiple assumptions simple linear regression relies on to not be valid. This would in turn cause any conclusions we make from the our line to be invalid and lead to a inconclusive answer to the research question.

Research Question 2: Bootstrapping

To answer our second research question, we will assume our sample data is representative of the population (the 31 data points we have reflect the values we would expect if we recorded the measurements of every fallen black cherry tree in the world). From our height QQ plot in the EDA analysis, we are able to find that height has a fairly normal distribution, as seen with how the points representing our data are fairly close to where the line in the plot is. We don't, however, know the exact values/parameters (i.e. true mean, true variance) to replicate the distribution. Since we know the distribution but not the parameters, we will be using parametric bootstrapping.

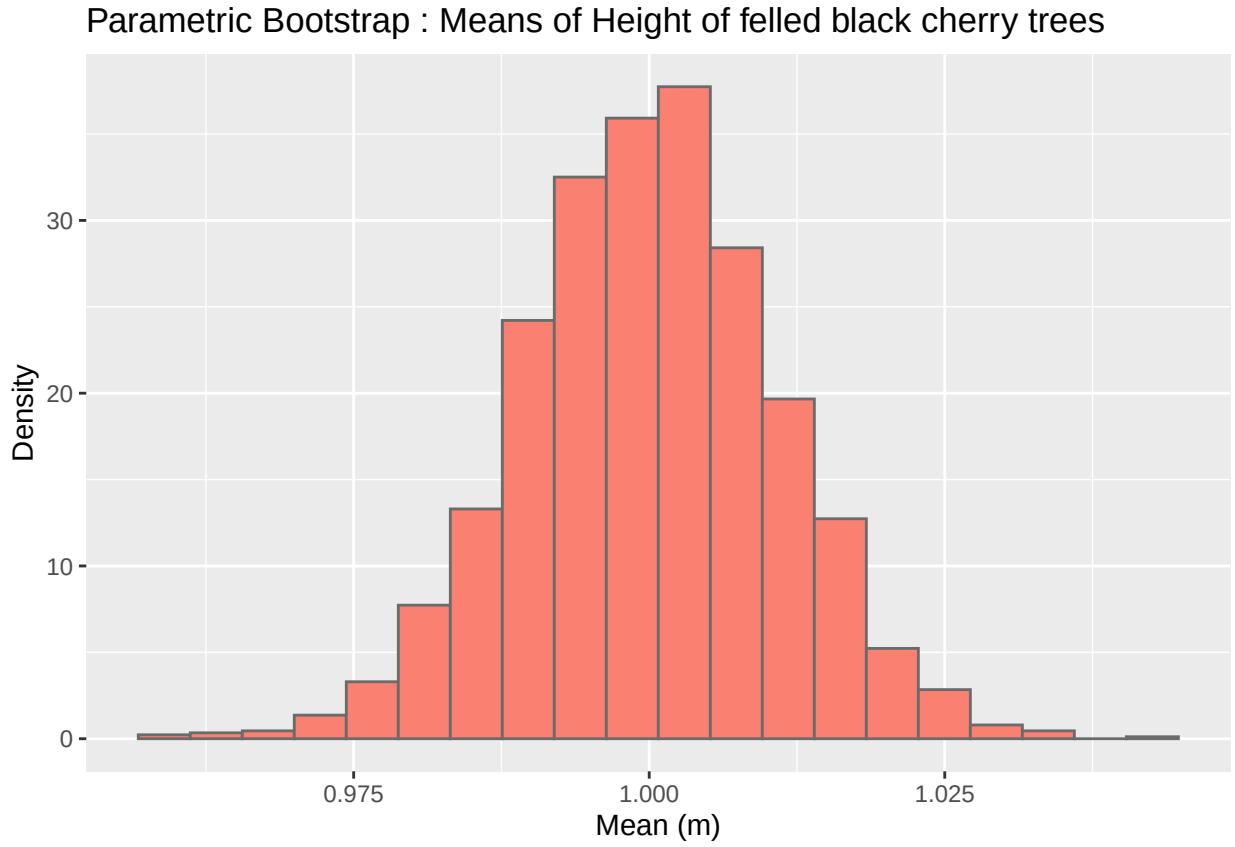



```
## [1] 0.975
```

The above results depicts the bootstrapped distribution of centered means of the height of felled Black Cherry trees. We have computed the probability that sample mean can estimate the true mean to 95% of it's scale(the sample mean will be between 0.975 and 1.025 times the true mean's value) to be 0.975. The centered mean seem to be normally distributed as seen in the histogram.

Research Question 2: Confidence Intervals

Next, we will use confidence intervals to find the range in which we are confident in our true mean will fall within




```
##       5%      95%
## 22.75698 23.56681
```

```
##     0.5%    99.5%
## 22.48346 23.80264
```

Taking our bootstrap from the previous method, we find that we are 95% confident that a true mean height from the fallen black cherry tree population will be between 22.75698 to 23.56681 meters and we are 99% confident that our true mean height from the fallen black cherry tree population will be between 22.48346 23.80264 meters. We note how close this range is despite have a very high confidence level and how our sample height from this data set lie within these intervals.

Research Question 2: Conclusion

Using the an estimator for mean height and parametric bootstrapping, we found that (assuming the sample we have is representative of the population) we have a high chance of having a sample mean height that is 0.975 to 1.025 times the true mean. This leads us to believe that our estimator is very accurate and that our sample mean height from the data could be representative (or close to representative) of the true population mean height.
We also employed the use of confidence intervals for our second statistical method to discover the true population mean height of fallen black cherry trees. This resulted in us being 99% confidant that the true mean height of the population is between 22.48346 to 23.80264 meters and 95% confidant that the true mean height of the population is between 22.75698 to 23.56681 meters.

Combining both the findings from our estimator analysis with parametric bootstrapping and confidence intervals, we are very confident to say that the true mean height of the population of fallen black cherry trees is between 22.48346 to 23.80264 meters and that the mean height we calculated from the sample data set is likely to be very close to what the true mean height is.

It must be noted for both methods used that the probability and confidence interval found does not show possible bias in collection of data or flawed experiment design. Both rely heavily on the fact that the data set we use represent the population (in other words, the 31 trees measured represent every single fallen black cherry tree in the world). Given the possibility that our data is bias and does not represent the population, our findings and conclusions would no longer be valid.

Citation of Dataset

Ryan, T. A., Joiner, B. L. and Ryan, B. F. (1976) *The Minitab Student Handbook*. Duxbury Press.

Appendix

Code to convert data from imperial to metric:

```
# conversion rates sourced from google
Girth <- trees$Girth * 2.54 #inches to cm
Height <- trees$Height * 0.3048 #feet to meters
Volume <- trees$Volume * 0.0283168 #ft^3 to m^3

data <- data.frame(Girth, Height, Volume)
```

Code for 5 summary statistics:

```
#The summary is used to find the 5 summary statistics information.
summary(data)
```

Code for Girth Distribution:

```
# Here, we have created the boxplot
# ggplot was used, with the x variable representing the studied parameter
box_diameter <- ggplot(data) +
  geom_boxplot(aes(x=Girth)) +
  labs(x='Diameter',
       title='Boxplot of Diameter of felled Black Cherry trees') +
  theme_classic() +
  scale_y_continuous(breaks=NULL)

# Here, we have created the historgam
# ggplot was used, with the x variable representing the studied parameter and the y representing Densit
his_diameter <- ggplot(data, aes(x=Girth))+
  geom_histogram(aes(y=..density..),
                 bins=7,
                 colour='black',
                 fill='thistle2')+
  geom_density(colour='red')+
  labs(x='Diameter',
       y='Density',
       title='Histogram of Diameter of felled Black Cherry trees')+
  theme_light()

grid.arrange(box_diameter, his_diameter)
```

Code for Height Distribution:

```
# Here, we have created the boxplots
# ggplot was used, with the x variable representing the studied parameter
box_height <- ggplot(data) +
  geom_boxplot(aes(x=Height)) +
  labs(x='Height',
```

```
       title='Boxplot of Height of felled Black Cherry trees') +
  theme_classic() +
  scale_y_continuous(breaks=NULL)

# Here, we have created the historgam
# ggplot was used, with the x variable representing the studied parameter and the y representing Densit
his_height <- ggplot(data, aes(x=Height))+
  geom_histogram(aes(y=..density..),
                 bins=7,
                 colour='black',
                 fill='thistle2')+
  geom_density(colour='red')+
  labs(x='Height',
       y='Density',
       title='Histogram of Height of felled Black Cherry trees')+
  theme_light()

grid.arrange(box_height, his_height)
```

Code for Volume Distribution:

```
# Here, we have created the boxplot
# ggplot was used, with the x variable representing the studied parameter
box_volume <- ggplot(data) +
  geom_boxplot(aes(x=Volume)) +
  labs(x='Volume',
       title='Boxplot of Volume of felled Black Cherry trees') +
  theme_classic() +
  scale_y_continuous(breaks=NULL)

# Here, we have created the histogram
# ggplot was used, with the x variable representing the studied parameter and the y representing Densit
his_volume <- ggplot(data, aes(x=Volume))+
  geom_histogram(aes(y=..density..),
                 bins=7,
                 colour='black',
                 fill='thistle2')+
  geom_density(colour='red')+
  labs(x='Diameter',
       y='Volume',
       title='Histogram of Volume of felled Black Cherry trees')+
  theme_light()

grid.arrange(box_volume, his_volume)
```

Code for the QQ plots:

```
# making QQ plots for all 3 variables in trees dataset
qq_diameter <- data %>%
  ggplot(aes(sample = Girth)) +
  geom_qq() +
  geom_qq_line() +
  labs(x = 'N(0,1) quantiles',
```

```
       y = 'A1C Quantile',
       title = 'Normal Q-Q plot for Diameter',
       subtitle = 'Data: Trees')

qq_height<- data %>%
  ggplot(aes(sample = Height)) +
  geom_qq() +
  geom_qq_line() +
  labs(x = 'N(0,1) quantiles',
       y = 'A1C Quantile',
       title = 'Normal Q-Q plot for Height',
       subtitle = 'Data: Trees')

qq_volume <- data %>%
  ggplot(aes(sample = Volume)) +
  geom_qq() +
  geom_qq_line() +
  labs(x = 'N(0,1) quantiles for Volume',
       y = 'A1C Quantile',
       title = 'Normal Q-Q plot',
       subtitle = 'Data: Trees')
# display plots
qq_diameter
qq_height
qq_volume
```

Code for Simple Linear Regression:

```
# use lm to get parameters for a linear fit/linear regression line
r_line <- lm(Volume ~ Girth, data = data)
r_line
```

```
# graph of data points and the linear regression line
data %>%
  ggplot(aes(Girth, Volume)) +
  geom_point() +
  stat_smooth(method = lm) +
  labs(x = 'Diameter', y = 'Volume',
       title = 'Volume vs Diameter of Dataset Compared to Fitted Simple Regression Line')
```

Code for Residual Analysis:

```
# put residuals from regression line into the dataset
r_line <- lm(Volume ~ Girth, data = data)
data$res <- r_line$residuals
data$fit <- r_line$fitted.values

# plot a density histogram of the residuals
res_hist <- ggplot(data)+
  geom_histogram(aes(x=res, y=..density..),
                 fill='thistle2', colour='black',
                 bins=6)+ labs(x='Residuals', y='Density',
                               title='Model Residuals,
```

```
                              n=30')+
  theme_classic()

# plot a QQplot of the residuals to check for normality
res_qq <- ggplot(data, aes(sample=res))+ geom_qq()+
  geom_qq_line()+ labs(x='N(0, 1) Percentiles',
                       y='Residual Percentiles',
                       title='Normal QQ Plot of
                       Residuals')+
  theme_light()

# find the mean of the residuals
res_mean <- mean(r_line$residuals)

# plot a scatter plot of residuals vs Girth and a line with where the residual mean is
res_var <- ggplot(data, aes(x=Girth, y=res))+
  geom_point()+ geom_hline(yintercept= res_mean,
                           colour='red', lty=2)+
  labs(x='Diameter (cm)',y='Residuals')+
  theme_light()

# display graphs and mean
grid.arrange(res_hist, res_qq)
res_var
res_mean
```

Code for Parametric Bootstrapping:

```
# we know from class that our MLE for the true mean for a normal distribution is just the sum of all sa

m <- mean(data$Height)
s2 <- sd(data$Height)
B = 2000
n <- length(trees$Height)

# set seed to replicate results
set.seed(123)

# taking a bootstrap sample of size 31, 2000 times
# the matrix method was used to perform
# parametric bootstrapping with normal data
boot.dat <- matrix(rnorm(B*n, m, sqrt(s2)),
                   nrow = B, ncol = n)

# computing point estimates for each sample
boot.m <- apply(boot.dat, 1, mean)

centered <- boot.m/m #the data is centered


# The histogram showing centered data bootstrap
ggplot(tibble(centered))+
  geom_histogram(aes(x=centered, y=..density..),
```

```
                fill = 'salmon', bins= 20,
                color = 'grey42')+
  labs(x = 'Centered mean', y = 'Density',
       title = 'Parametric Bootstrap : Centered means of Height of felled black cherry trees')

# calculate the probability that our sample can estimate our true population mean
prob <- sum(centered >= 0.99 & centered <= 1.01)/B
prob
```

Code for Confidence Intervals:

```
m <- mean(data$Height)
s2 <- sd(data$Height)
B = 2000
n <- length(data$Height)
# set seed to replicate results
set.seed(4321)
# taking a bootstrap sample of size 31, 2000 times
# the matrix method was used to perform
# parametric bootstrapping with normal data
boot.dat <- matrix(rnorm(B*n, m, sqrt(s2)),
                   nrow = B, ncol = n)

# computing point estimates for each sample
boot.m <- apply(boot.dat, 1, mean)

# The histogram showing data bootstrap
ggplot(tibble(boot.m))+
  geom_histogram(aes(x=centered, y=..density..),
                fill = 'salmon', bins= 20,
                color = 'grey42')+
  labs(x = 'Mean (m)', y = 'Density',
       title = 'Parametric Bootstrap : Means of Height of felled black cherry trees')

# 90% confidence interval
cf90 <- quantile(boot.m, prob=c(0.05, 0.95))
cf90
# 99% confidence interval
cf99 <- quantile(boot.m, prob=c(0.005, 0.995))
cf99
```