# Research Proposal

March 21 2022

Linh Hoang and Ayushi Singh

**DESCRIPTION OF DATA**

- The dataset 'trees' has 31 measurements of the volume of timber, girth(diameter) and heights of fallen Black Cherry trees.
- Citation (in Open dataset in R) : Ryan, T. A., Joiner, B. L. and Ryan, B. F. (1976) The Minitab Student Handbook. Duxbury Press.

**DESCRIPTION OF RELEVANT VARIABLES**

1. 'Girth' records the diameter (in inches) of the fallen tree. The diameter is measured across a tree at 4 feet 6 inches above the ground.
2. 'Volume' records the volume (in cubic feet) of a fallen tree's timber.
3. 'Height' records the height (in feet) of a fallen tree.

**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?

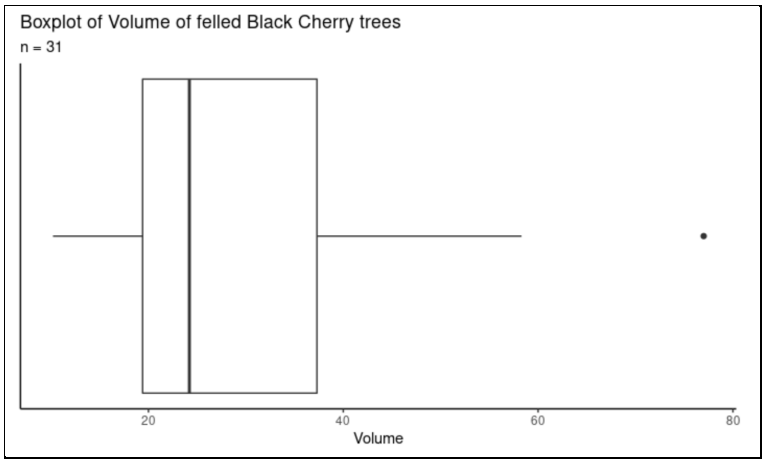
**ANTICIPATED DATA CLEANING AND REFINING**

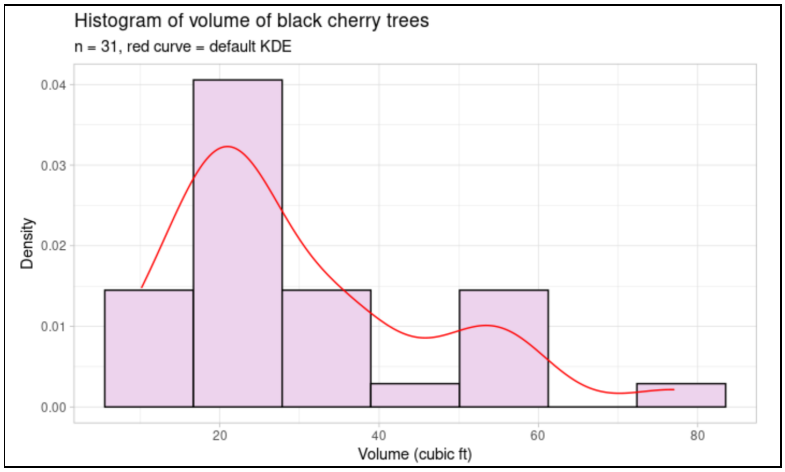
From our initial look into the dataset, There are no missing values from the samples. We plan to convert the values from the imperial to metric since most of the anticipated readers of our analysis would be more used to the metric system.

**5 SUMMARY STATISTICS**

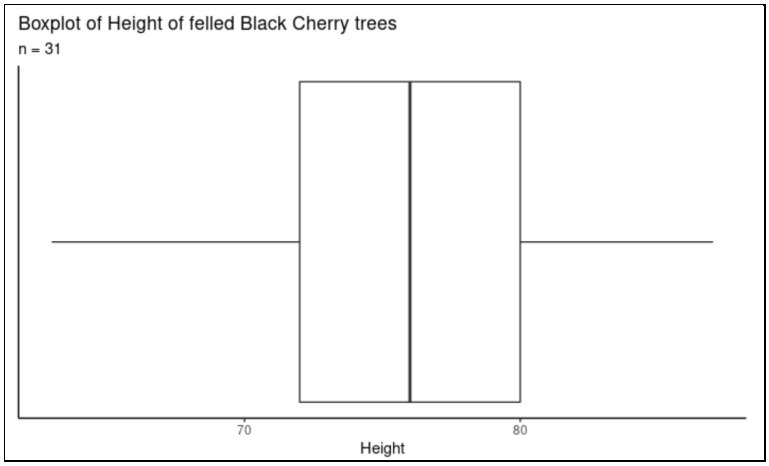
| | Girth | Height | Volume |
|---|---|---|---|
| Minimum | 8.30 | 63 | 10.20 |
| 1st quartile | 11.05 | 72 | 19.40 |
| Median | 12.90 | 76 | 24.20 |
| Mean | 13.25 | 76 | 30.17 |
| 3rd quartile | 15.25 | 80 | 37.30 |
| Max | 20.60 | 87 | 77.00 |

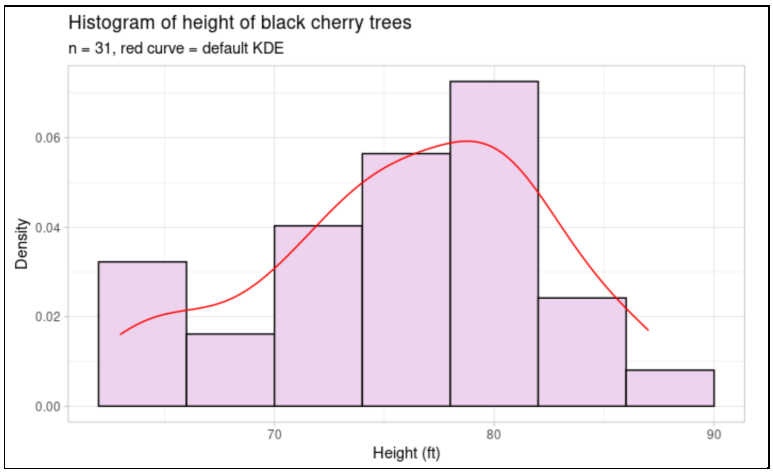
## EXPLORATORY DATA ANALYSIS



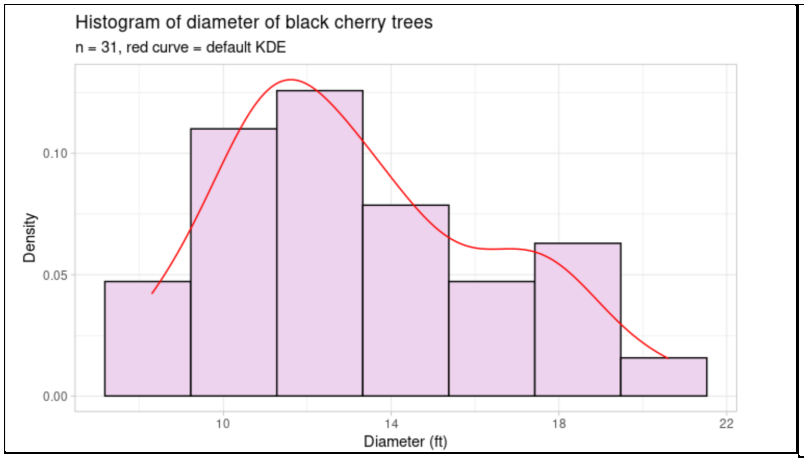



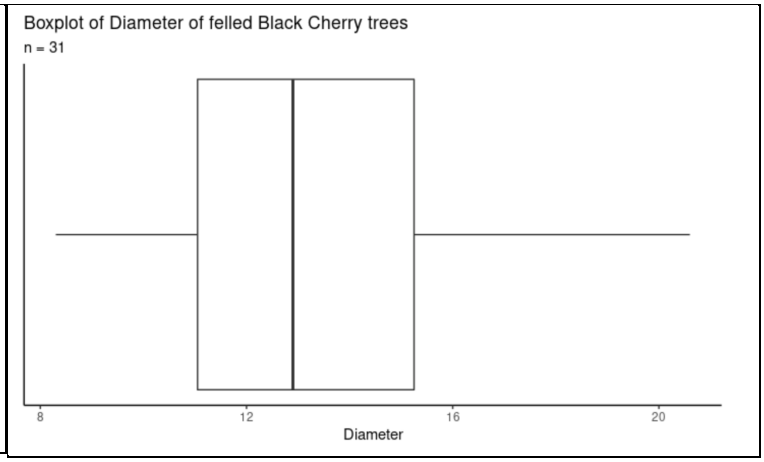

- Analysis : We can see that Q3 is at 40ft$^3$ in the box plot, which tells us that 75% of the data is below 40. We can also see that there is 1 outlier, which is at 80ft$^3$ (this outlier is also seen in the histogram as there is a “gap” between the last two bins). The data is also bimodal, with another peak at 50 cubic ft.




- Analysis : The box plot has its Q3 at around 80, indicating 75% of the sample heights have a value below 80 ft. The data has a singular peak at 80 ft. The KDE has a roughly normal shape, which may possibly indicate that it follows a normal distribution function.




- Analysis : The box plot indicates that Q3 is located around 15 and Q1 is located around 11, indicating that 50% of values are between 11 and 15ft. It is bimodal at 12ft and 17.5ft.

**LINEAR REGRESSION**

We will be answering RQ1 using linear regression. The question asks for the type of relationship between two quantitative variables. Using simple linear regression (which estimates and looks for any correlation between two variables) would fit well to help analyze evidence for some relation.

**METHODOLOGY 1**

We'll use empirical or parametric bootstrapping for research question 2. We'll decide which type of bootstrapping using QQ plot and goodness of fit (to find normality). Because we only have 31 data points and a fairly wide range of values (as seen in side-by-side histogram and boxplot for Height), there doesn't seem to be enough values to justify the sample mean is the population mean. Bootstrapping will show us how the sample mean can be distributed and if there's more evidence to suggest the sample mean is close to the true mean.

**METHODOLOGY 2**

We'll be using confidence intervals for our second statistical method to answer research question 2. As we are looking for the population mean of fallen black cherry tree heights, confidence intervals will help us determine how accurate our calculated mean may be to the population mean.