Statistics Rapport

Study about oak trees

Table of Contents

Dataset

We make use of a custom dataset, founded in the eik.csv file.

Here we can find five variables:

- Boom = tree: the id of the tree
- Regio = region: the region the tree was measured at
- Grootte = width: the width of the area the tree grows in (100 km^2)
- Volume = volume: the volume of the tree (cm^3)
- Hoogte = height: the height of the tree (m)

Following columns are deleted from my dataset: {1, 3, 9, 11, 17}

Student number: 20213082

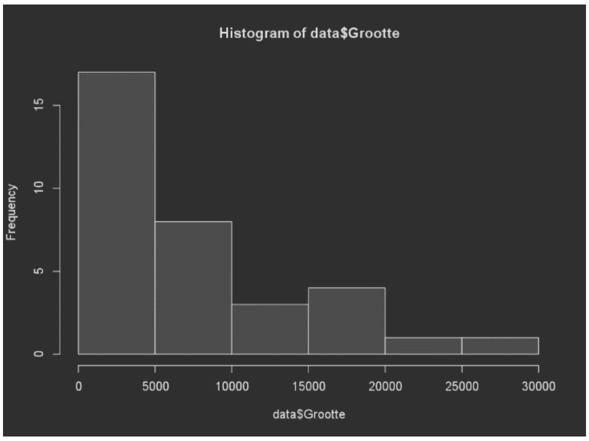
Questions

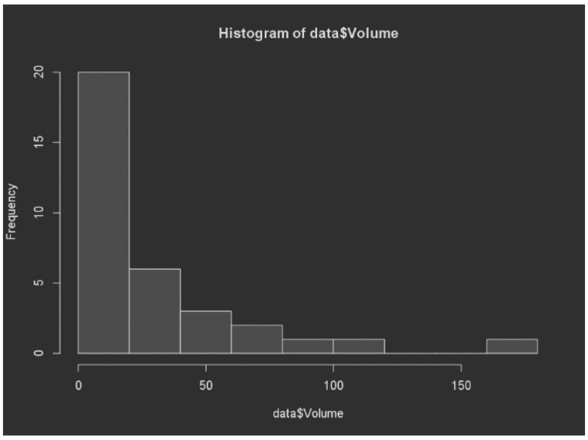
Question 1: *Study and discuss the distribution of the variables Volume and Width. To do this, discuss appropriate graphical representations. Also, formally determine whether the data is normally distributed. If this is not the case, in what way does the data deviate from normally distributed data? Discuss.*

Answer:

We make use of histograms to check whether our data is normally distributed.

If we look at both variables, we can see that both are positive skewed graphs.





The data deviates more to the right if we compare it to a normal distribution.

Now if we discuss about both histograms, we can conclude the following:

- We can see that there are more trees that grow in smaller areas.
- We can also see that more trees have a smaller volume.

Question 2: *Investigate whether there is a correlation between "thick acorns," which are oak trees whose acorn volume is at least 3 cm^3, and the area in which the tree occurs. To do this, create a new variable called "thick acorn." Then, perform an appropriate test to determine if there is a significant correlation between thick acorns and the tree's geographic location.*

Answer:

In this question we look at the corelation between the thickness of the acorn and the region they appear at.

We make an extra variable called “dikke_eikels”, it is one if it is bigger than 3 cubic centimetres and zero if not.

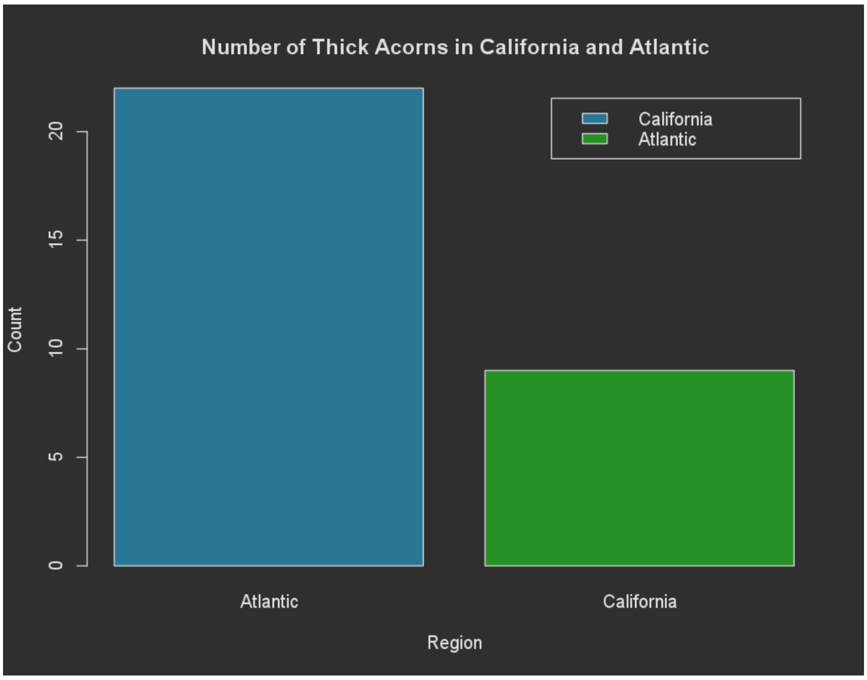
We make use of a chi-squared test to determine if there is a correlation between how thick the acorn is and what region it comes from. Our null hypothesis is that there is no correlation between how thick the acorns are and where they come from. We choose this as our null hypothesis because we want to prove that there is a correlation between these two variables.

After performing our chi-squared test we get the following results:

X-squared = 1.7702, df = 1, p-value = 0.1834

We can see that our p-value is higher than 0.05. This suggests that there is not enough evidence to conclude a significant association between "dikke eikels" (thick acorns) and the region where the tree occurs.

But if we look at bar plot, we can see that there is more thick acorns in Atlantic.

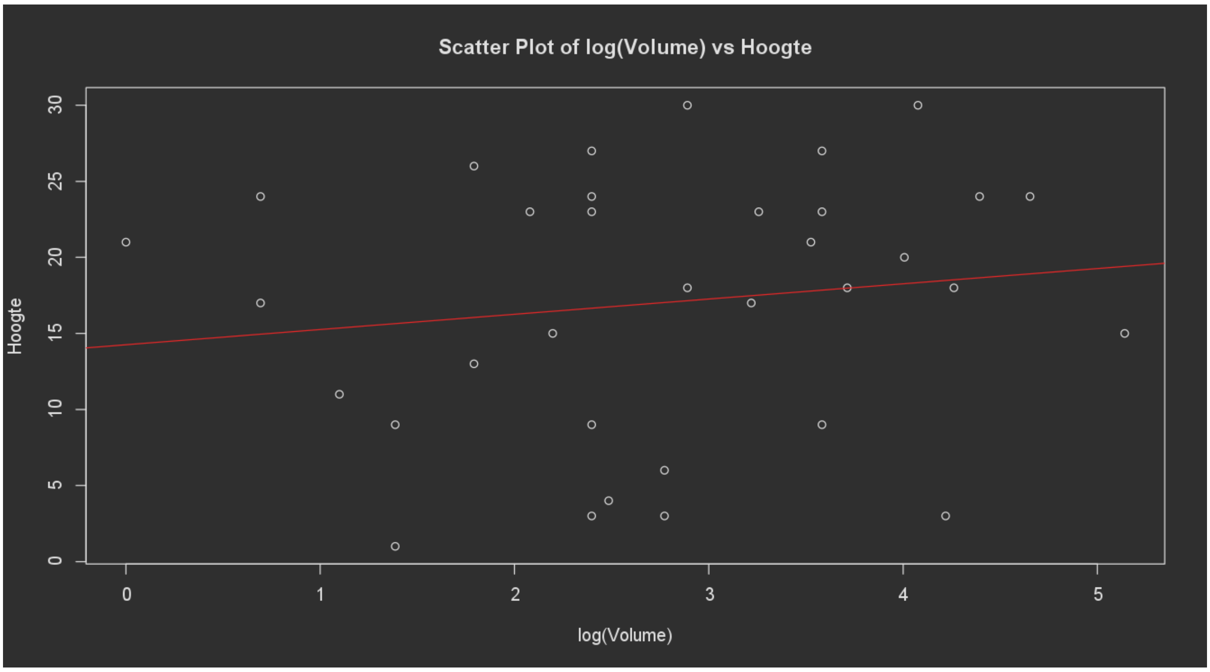


Question 3: *Can you predict Height from log(Volume)? Answer this question thoroughly and as completely as possible.*

Answer:

To determine if there is a relation between height and log(Volume), we make use of a scatter plot.

By doing so we can see if our data is a linear line, implying that there is a correlation between log(Volume) and height.



We can see that they don’t have any correlations with each other.

If we look at the plot more in detail:

- We can see that there is a linear line forming. Some points are on the line.
- More outliers than linear line

From these two points we mostly conclude that it’s not correlated with each other.