Assignment 1 (A1):

Information Visualisation

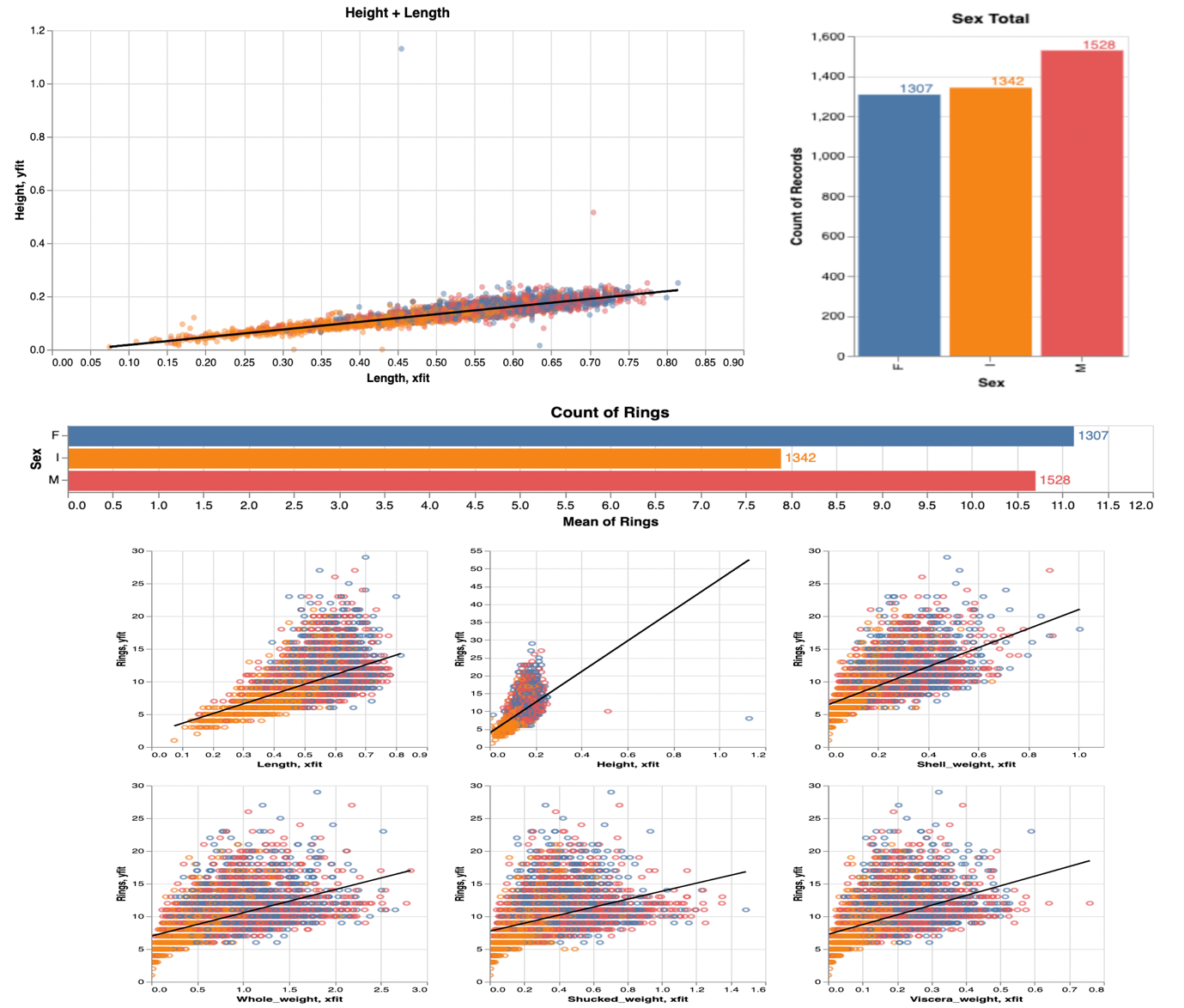
Data Set and Data Structure

Chosen data set: Abalone Data Set. This data set contains 8 attributes with 4177 instances. There a number of data types, these include: Quantitative; Nominal [4].

Task to undertake with data

The user wants to be able to guess the age of the abalone, which is determined by the number of rings, that are within it, when it has been cut open.

**Prototype 1**



Description and justification:

The starting point for this design is getting a functional, effective system that provides all of the required tasks of the user. [1] With this in mind, the main aim is to make sure that all required metrics are available for the user. This intention is to allow the user to complete their task as best as possible. Due to humans only being able to deal with things in chunks, the visualisations divided into horizontal sections. Each section has relevant information, which is chunked to make it easier for the user to remember and compare. Also, to try and not overwhelm them with much information in one go.

With eyes being better than a person’s memory, having visualisations side-by-side makes it easier to compare. [2] There will be a level of animation, but this will be kept short. Transitioning is used between changes of different states [2] when the user has selected a section within the main graph. This subsection is then what will be displayed in the other visualisations, to help get a clearer picture. Scatter charts are “Good for showing the relationship between two different variables where one correlates to another (or does not).” [3] An alternative possibility is to use a bubble chart instead, as they are suitable for expressing three numerical variables. [3]

For the nominal data, sex, and the quantitative value of the mean of the rings, bar charts will display the values. As categories are used to separate the values, by using bar charts, this allows the user to compare different values when specific values are essential, for example, the average number of rings the selected Abalone have. [3]

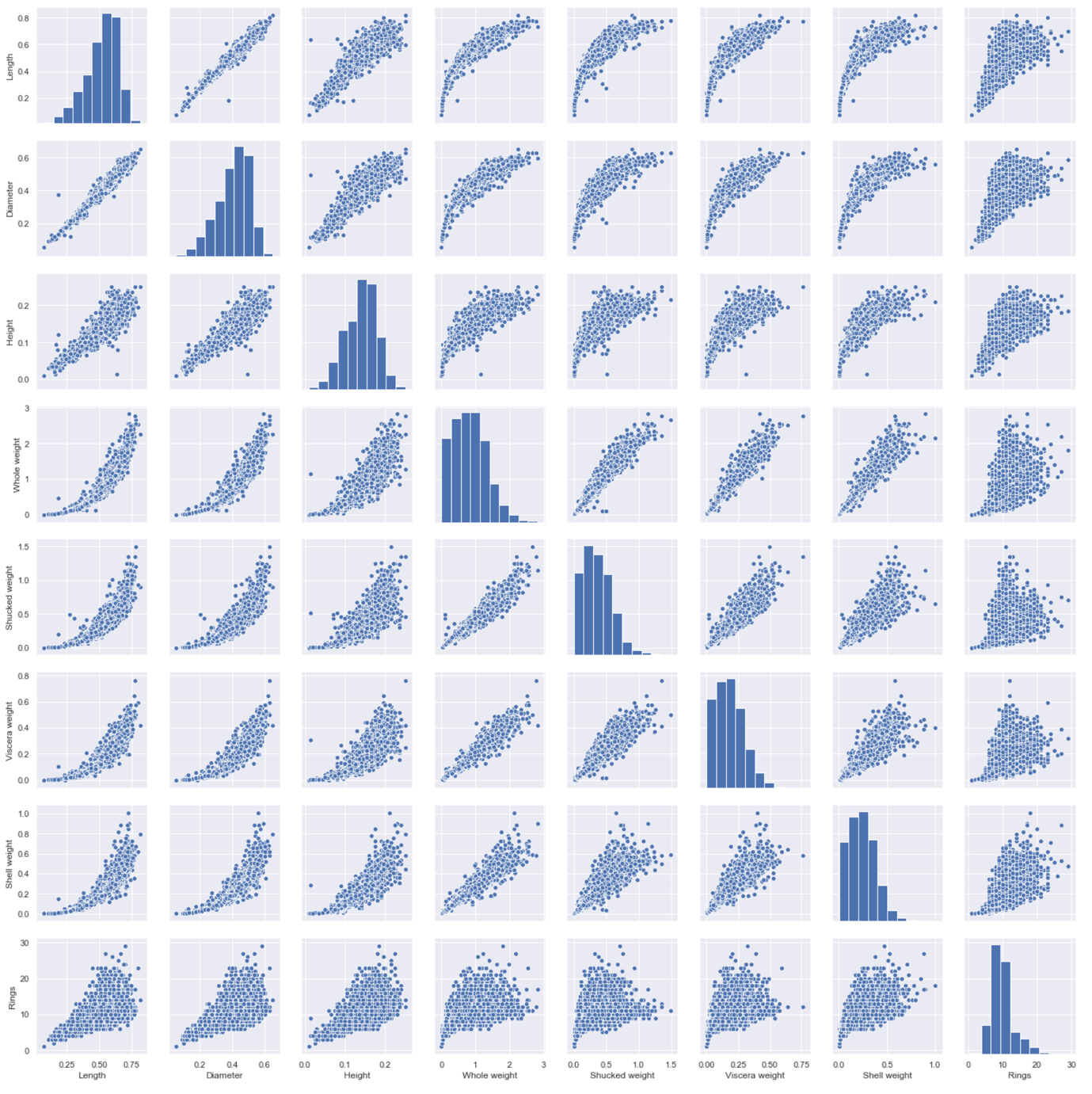
The scatter graphs will be placed together, in the bottom half of the view, as they will be using similar attributes. So by them being side by side, this will make it easier to compare the values. Through using coordinated views, the user can interact with the data in various ways, therefore allowing the exploration of the data visualisations easier. [8] Tamara states, “the strength of the small-multiple views is in making different partitions of the dataset simultaneously visible side by side, allowing the user to glance between them with minimal interaction cost and memory load.” [9]

Radio buttons are available to the user to allow them to start filtering out data and see what is more appropriate for completing their task. As if the user has, for example, decided the Abalone is female, then they can remove the noise generated by the male and infant data to allow them to have a clearer picture and make a better decision.

Each graph will use colours to differentiate the difference between the data points for Male, Female and Infants. This is to allow the user to be able to see the relevant data across all visualisations and know which data point is for what sex. The Hue of the colours will be bold, different colours to make this even more comfortable for the user, as if saturated colours were used, then it means it will make the user have to think and distract away from the task. Therefore, it might make more errors prone to happening. Colin Ware states, “A nominal code does not have to be orderable; it must be remembered and recognized. Color can be extremely effective when we wish to make it easy for some-one to classify visual symbols into separate categories; giving the objects distinctive colors is often the best solution.” [10]

The graphs will also display the data point values when the mouse hovers over the point or column. It will allow the user to have the required information they need at their fingertips, with minimal effort on their behalf. Again, this is to allow them to be able to focus on what their actual task is, guessing the age, not figuring out the data values.

Prototype 2



This design is first and foremost aiming to get all the potential measurement required by the user displayed functionally and effectively. The dashboard is presented again in vertical sections. This design will utilise the technique called Scatterplot Matrix (SPLOMS). SPLOMS are useful as they allow the user to be able to see multiple paired views. However, they take a lot of screen space. [6]

By using interaction, this will make sure that SPLOM is effective as possible for the user. It is enabling the user to be able to see the links between the different graphs while exploring the data. [Find references]

As SPLOMS create so many views, this potentially could be overwhelming for the user. In order to reduce the number of visualisations on the screen, the option of checkboxes will allow the user to select their desired sub selections. It is allowing them only to see data based on the features that they want. [Find references]

An analytical technique that will be applied is a dimension reduction technique called PCA. Due to the data being vast, with a lot of overlapping features. The aim will be, by using this technique, to reduce the amount of similarity but expand the spread of the variance. It is allowing there to be a better chance to decide on the outcome, which will be the number of rings. Dimensionality reduction aims to place similar features together, potentially allowing clusters to form which can then possibly discovered in lower dimensions like 2D. [7]

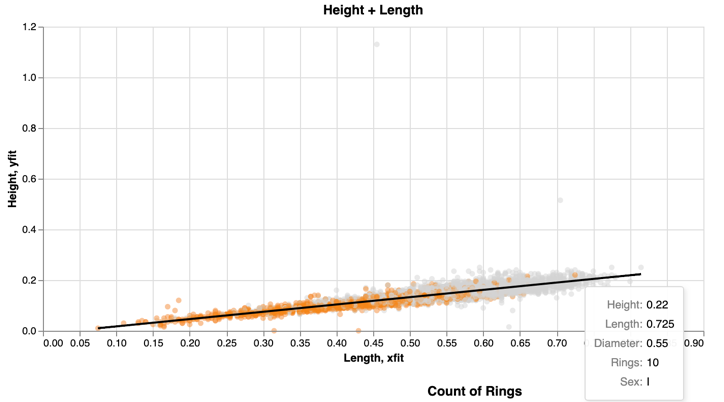
The colours will be using a Hue variance. In order to make sure the data is made clear as possible for the user, as there are three main groups of nominal data. Making it an excellent feature to set the charts colouring schemes on. (Why using Hue colouring?)

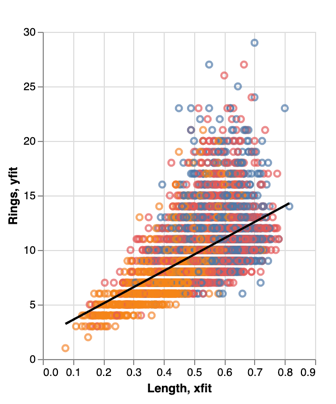
Altair Code Implementation

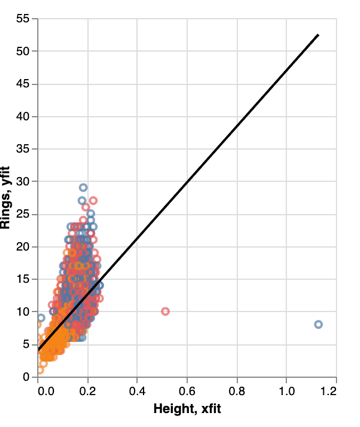
Analytic technique: SPLOM?? Maybe

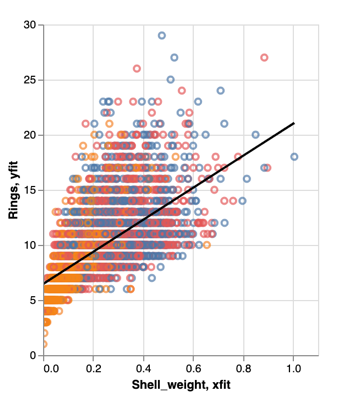
Data Discoveries

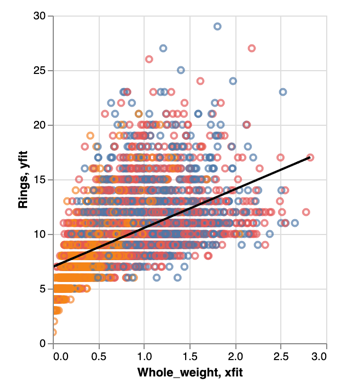
Charts can be cluttered. However, restricting data you don’t get a true result, however, too much data makes it hard to determine.

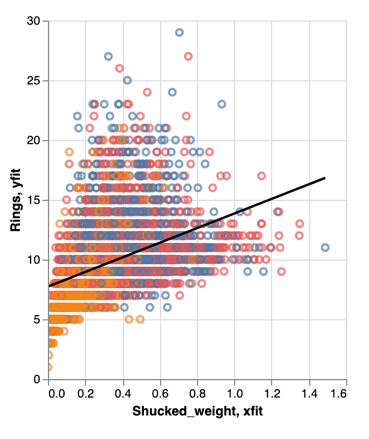


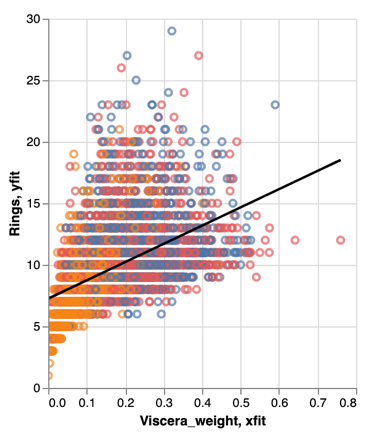












Talk about the category splits between M, F, I.

Also talk about the mean amount of rings for each sex.

The linear line matches up with same data in different charts. Think of example showed mum.

Within the data, there are several outliers. There are a few along the 0 y-axes, also there are a few extreme data points that potentially skew the data. This would create a few issues if machine learning techniques were being used as this would have a great impact on training the model and skew the training. If machine learning was going to be used, I would suggest these data points being removed. (Need image + references).

It is an assumption that the older an Abalone is, the bigger it is likely to be in both length and height. This could be said for all the other metrics. However, through exploring the data, being able to determine the sex of the creature is not as straight forward as determining the age and then using this as a factor to influence the decision on its chances of it being male or female. There are a few instances of the creatures being 11.5 years old and still being an infant. (need references and more images)

When comparing the weight metrics to the age of the Abalone, they all follow a very similar pattern. However, the whole weight of the snail and the shucked weight are very similar is data distribution, just different weights, but the viscera weight follows a similar pattern as the other two weights but differs slightly. The viscera weight seems to be a lot more condensed and the data points are closer together, with only a few that look like outliers. For looking at the weight, you can see that just being the heaviest doesn’t mean that the snail is the oldest or just older than the rest. (Need references + tidy up images)

Findings:

Kaggle: <https://www.kaggle.com/miksaas/abalone-eda-regression-pca-classification#Categorical-Feature>

- Many features are highly correlated

- length and diameter show linear correlation

- the length and weight features are quadratic correlated

- whole weight is linearly correlated with other weight features

- Number of Rings is positively corelated with almost all quadratic features

- Possible outliers in Height features

-Whole weight is almost linearly correlated with all the features except Rings

-Length is linearly correlated with Diameter

-From all the features excluding Rings, Height is least correlated with other features

- Rings feature has the highest correlation with Shell Weight followed by Height, Length and Diameter

-Distribution between Male and Female is similar

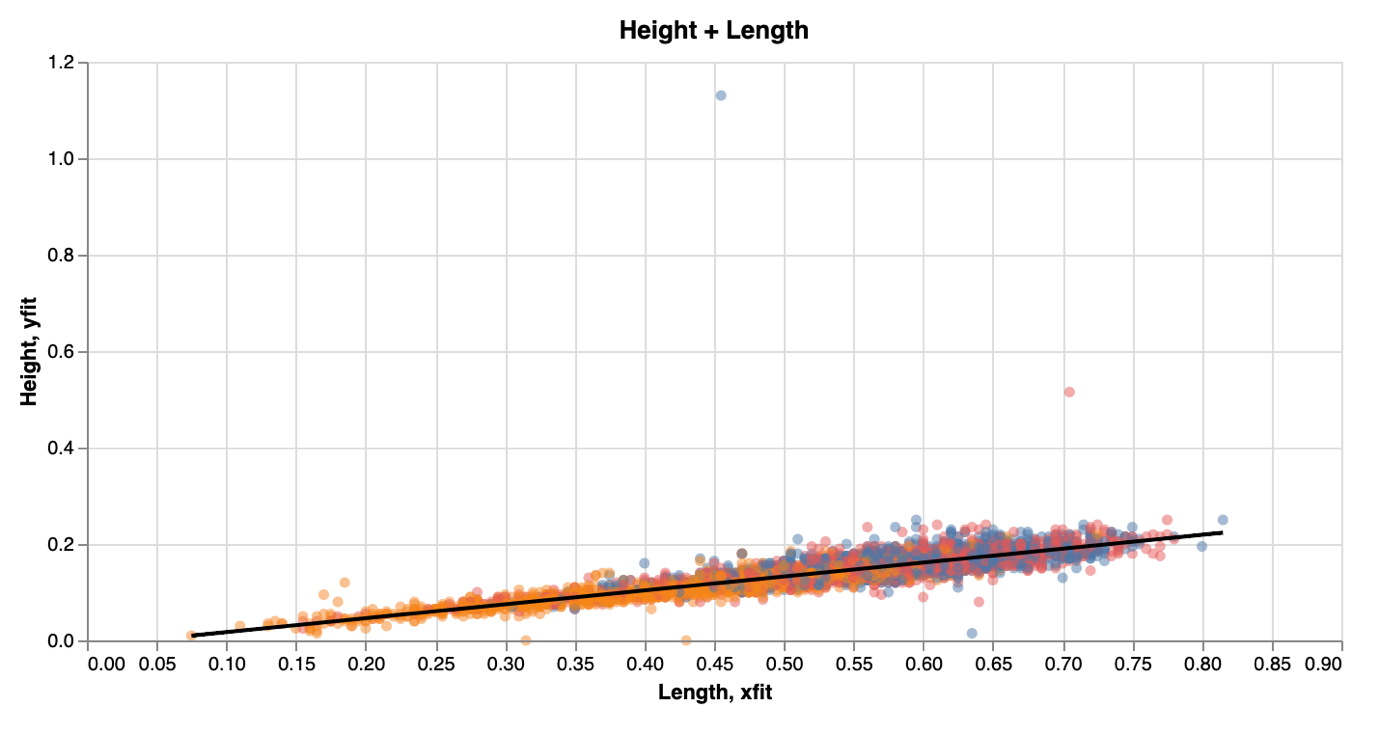
-Most of the Rings both for Male and Female are between 8 and 19

-Infants have mostly from 5 to 10 Rings

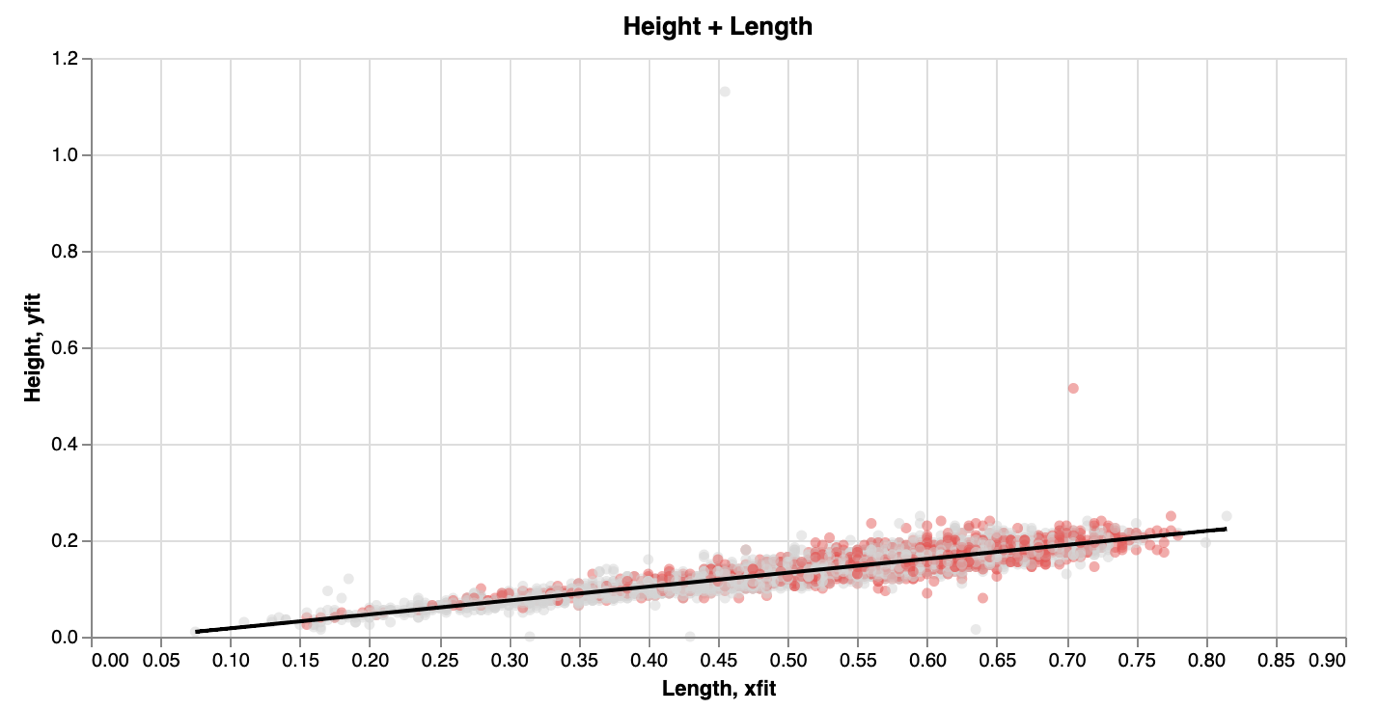
A picture containing screenshot

Description automatically generated

The image above shows that the data between the sexes is quite even. However, there is slightly more males compared to the rest.



The length and the height are very linearly correlated.



A close up of a white background

Description automatically generated

Average age of rings: Female live longer, then male.

A screenshot of a cell phone

Description automatically generated

# Bibliography

**There are no sources in the current document.**

[1] Lecture slides: Visualisation Rule of thumb: Daniel Archambault: 2019, Slide 6

[2] Lecture slides: Visualisation Rule of thumb: Daniel Archambault: 2019, Slide 17

[3] EasyBI - <https://eazybi.com/blog/data_visualization_and_chart_types/> - Accessed [6/11/19]

[4] Dataset (<https://archive.ics.uci.edu/ml/datasets/Abalone>)

[6] Lecture 7 – Dan Anc - slide 9

[7] Lecture 10 – Slide 15

[8] state of art jonathan Roberts – 2007 - <https://www.cs.kent.ac.uk/pubs/2007/2559/content.pdf> - [6/11/19]

[9] Tamara munzner - Munzner, Tamara – Visualisation Analysis & Design – 2015 – p274

[10] Ware, Colin – Information Visualisation: Perception for design 3rd edition (year?) – p122

Ware, Colin – Information Visualisation: Perception for design 3rd edition (year?)

Munzner, Tamara – Visualisation Analysis & Design – 2015

Chen Min, Feixas Miquel, Viola Ivan, Bardera Anton, Shen Han-Wei, Sbert Mateu – Information Theory Tools for Visualisation – 2017

Good link about chunking up data : <https://junkcharts.typepad.com/junk_charts/2013/01/ruining-the-cake-with-too-much-icing.html>

Find references to back up

Review, is it needed?

Worth a re-read – not 100% sure.