Assignment 1 (A1):
Information Visualisation

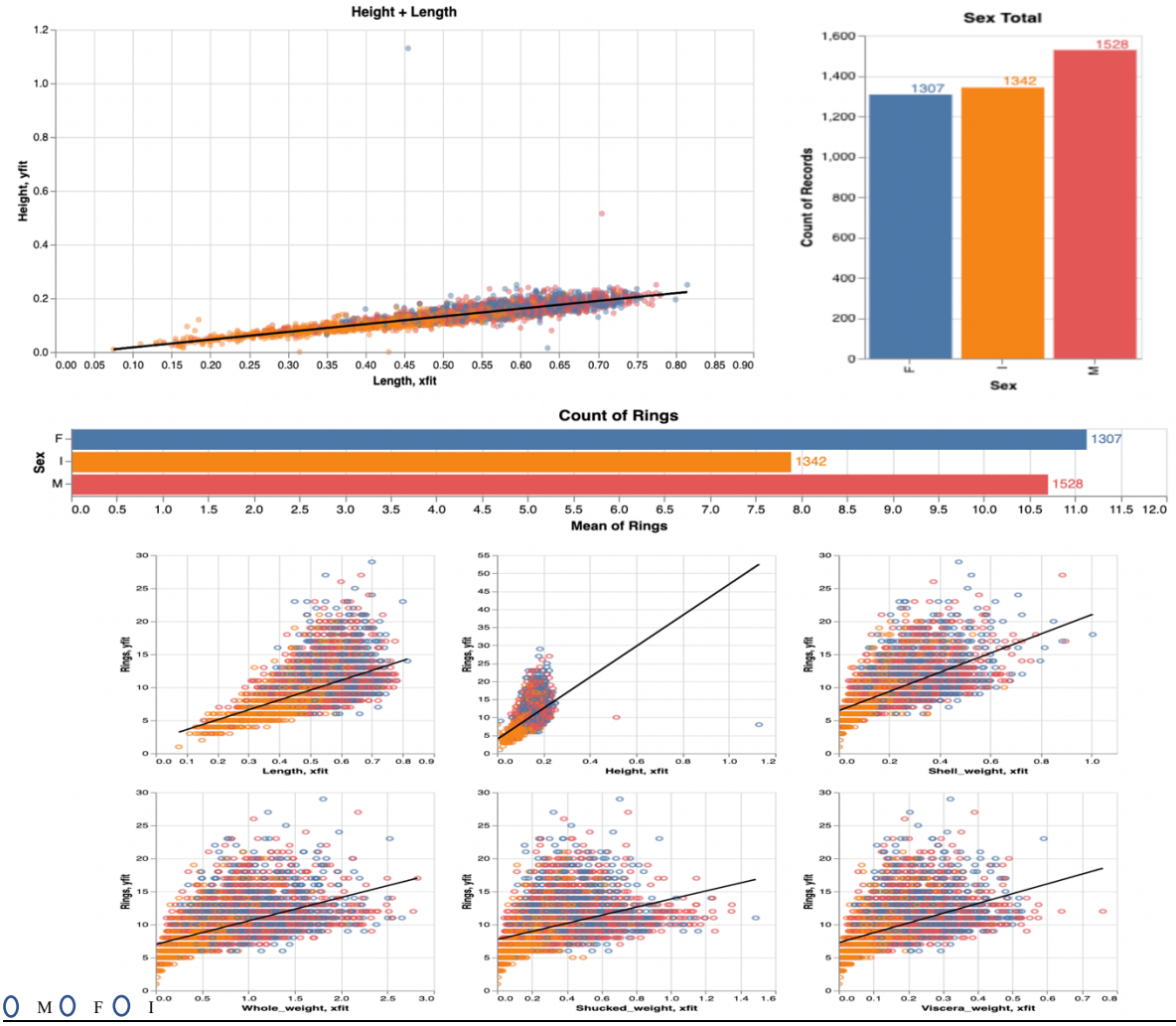
Data Set and Data Structure

Chosen data set: Abalone Data Set. This data set contains eight attributes with 4177 instances. There several data types, these include Quantitative; Nominal [1].

User Task

The user wants to be able to guess the age of the abalone. It is determined by the number of rings, that are within the creature when it has been cut open then add 1.5.

Prototype 1



M F I

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 [2]. The main aim is to make sure that all the necessary metrics are available for the user. 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 [3]. There will be a level of animation, but this will be kept short. Transitioning is used

between changes of different states [3] 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) [4].”

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 [4].

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 [5]. Munzner [6] 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.”

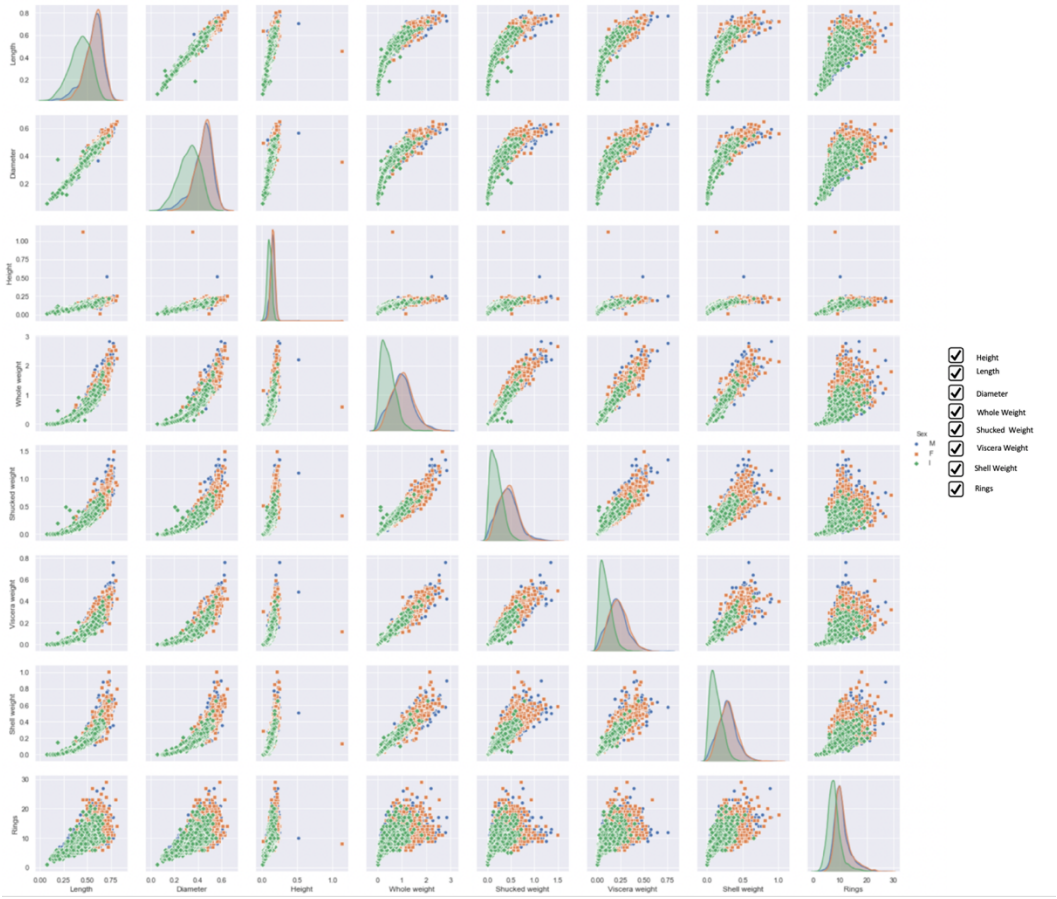
Radio buttons are available to the user to allow them to start filtering out data and see what is more appropriate for completing their task.

Each graph will use colours to differentiate the difference between the data points for Male, Female and Infants. This allows 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 distinctively different colours to make this even more comfortable for the user. Ware [7] states, “color can be extremely effective when we wish to make it easy for someone to classify visual symbols into separate categories; giving the objects distinctive colors is often the best solution.”

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.

The analytical technique used is linear regression. Linear regression aims to plot out the relationship between two variables to be able to observe data [8]. Regression lines appear on scatter point visualisations [9].

Prototype 2



Description and justification:

This design is first and foremost aiming to get all the potential measurement required by the user displayed functionally and effectively. This design will utilise the technique called Scatterplot Matrix (SPLOM). SPLOM are useful as they allow the user to be able to see multiple paired views. However, they take a lot of screen space [10].

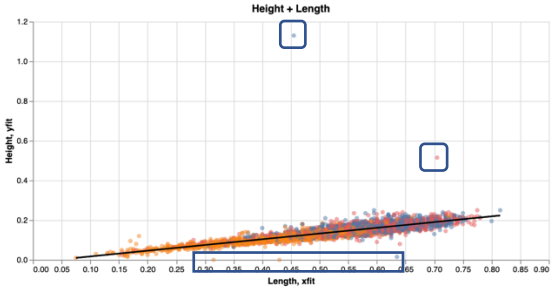
By using interaction, this will make sure that SPLOM is effective as possible for the user. It is enabling the user to be able to see the links between the different graphs while exploring the data. SPLOM show all the possible pairwise combinations of attributes, with the attributes as the rows and columns [11].

As SPLOM 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, and this is known as linked highlighting. The benefit of this is in being able to see data in one graph is viewed in another. It also allows data to be shown simultaneously but by reducing clutter on a single graph [12].

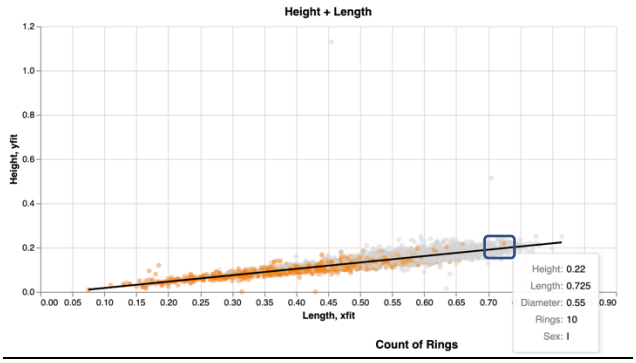
An analytical technique that will be applied is a dimension reduction technique called PCA. Due to the data being vast, with a lot of overlapping features. The aim will be, by using this technique, to reduce the amount of similarity but expand the spread of the variance. It is allowing there to be a 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. [13]

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.

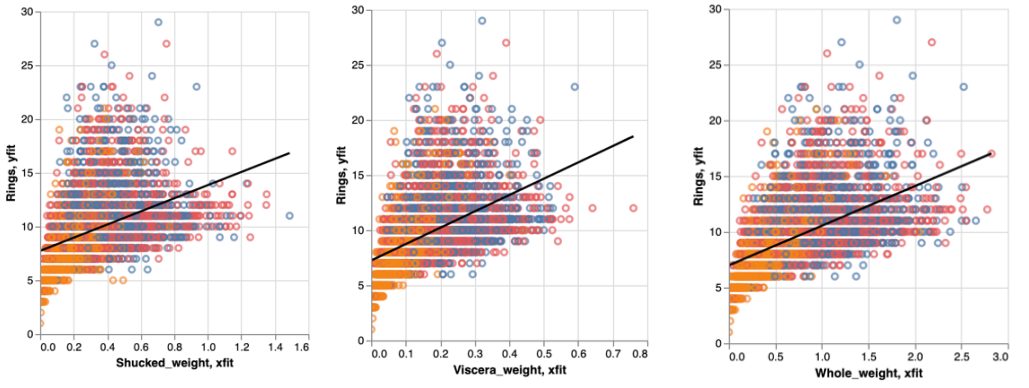
Data Discoveries



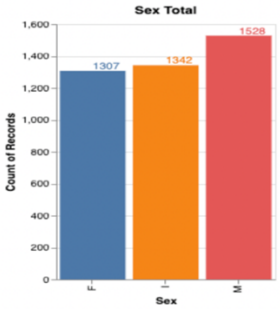
The length and height are very linearly correlated. There are also several outliers within the data. Also, there are a few extreme data points that potentially skew the data. The outliers would create a few issues if machine learning techniques were being used as this would have a significant impact on training the model and skew the training. If machine learning were going to be used, it would suggest that these data points be removed [14].

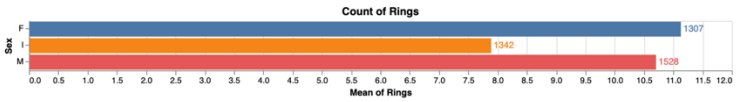


It is an assumption that the older an Abalone is, the bigger it is likely to be in both length and height. This assumption could be made for all the other metrics. However, through exploring the data, being able to determine the sex of the creature is not as straight forward as determining the age and 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.

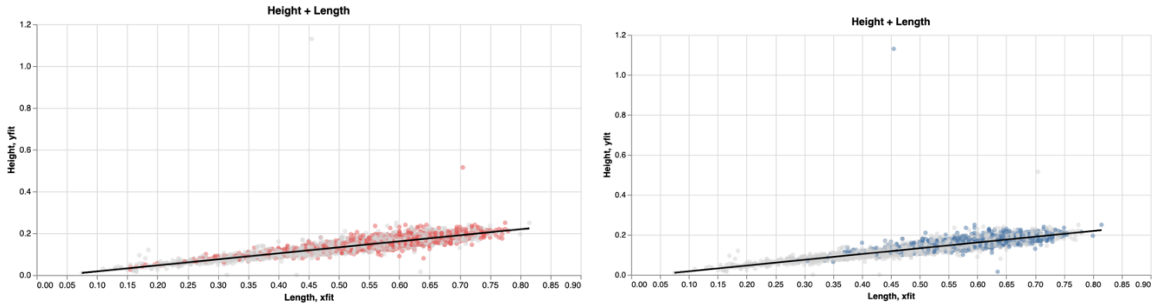


When comparing the weight metrics to the age of the Abalone, they all follow a very similar pattern. However, the whole weight of the snail and the shucked weight are very similar is data distribution, just different weights, but the viscera weight follows a similar pattern as the other two weights but 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.

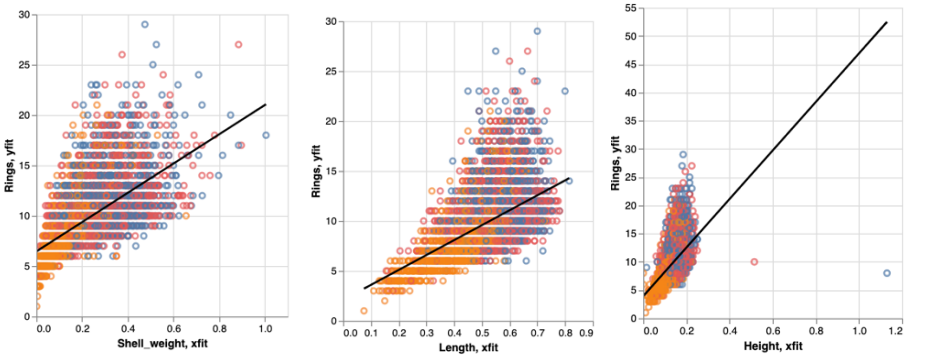




The image above shows that the data between the sexes is quite even. However, there is slightly more males compared to the rest. The average age of all the rings give a conclusion that females live longer, then male.



Using the size of the creature, both length and height, is not a clear metric to determine if it is male or female. Both sexes grow to roughly the same size, both length and height.



The rings feature of the data set has the highest correlation with shell weight, which is then followed by height and length. Research also suggests that this is then also followed finally by the diameter [14].

Bibliography

[1] UCI. Abalone Data Set https://archive.ics.uci.edu/ml/datasets/Abalone (accessed 03/11/19).
[2] Archambault, D. Visualisation Rule of thumb. Lecture slides: 2019. Slide 6.
[3] Archambault, D. Visualisation Rule of thumb. Lecture slides: 2019. Slide 17
[4] EasyBI. Data Visualization – How to Pick the Right Chart Type?
https://eazybi.com/blog/data_visualization_and_chart_types/ (accessed 06/11/19)
[5] Roberts, J. State of Art https://www.cs.kent.ac.uk/pubs/2007/2559/content.pdf (accessed 06/11/19)
[6] Munzner, T. Visualisation Analysis & Design; CRC Press, 2015. p274.

[7] Ware, C. Information Visualisation: Perception for design, 3rd ed.; Morgan Kaufmann: 2013. p122.
[8] Linear Regression http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm (accessed 11/11/19)
[9] Munzner, T. Visualisation Analysis & Design; CRC Press, 2015. p148.
[10] Archambault, D. High Dimensional Data. Lecture slides: 2019. Slide 9.
[11] Munzner, T. Visualisation Analysis & Design; CRC Press, 2015. p160.
[12] Munzner, T. Visualisation Analysis & Design; CRC Press, 2015. p267.
[13] Archambault, D. Introduction to Data Analytics Methods. Lecture slides: 2019. Slide 15
[14] Kaggle. Abalone - EDA, Regression, PCA, Classification
https://www.kaggle.com/miksaas/abalone-eda-regression-pca-classification#Categorical-Feature
(accessed 10/11/19)