Assignment One (A1):

Information Visualisation (Individual)

David Saunders - 910995

Data Set [10 marks total]

The data set used in this assignment is the Contraceptive Method Choice Data Set from the UCI Machine Learning Repository (Dua and Graff, 2019). The data set has a total of 10 features; of which 2 are quantitative, 4 are ordinal, and 4 are nominal. Some of the features are binary, but can be thought of as nominal but with only 2 categories.

The description of the dataset on the repository claims the data should be used to predict what form of birth control a woman uses, however the information can also be used to see how women of different ages may have different approaches to contraception.

The dataset can be downloaded online at:
https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice (Dua and Graff, 2019)

Designs [40 marks total]

Design 1

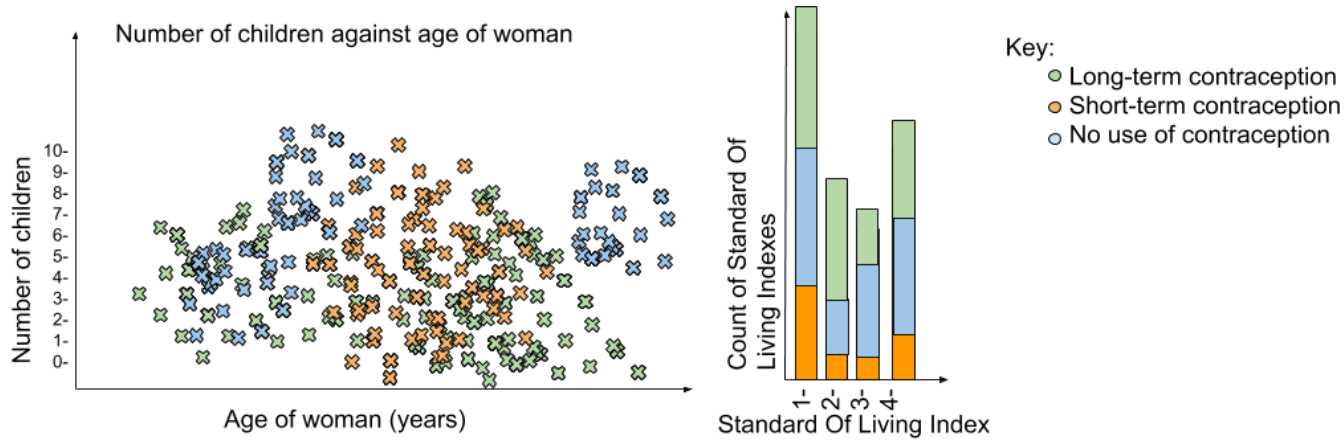

Figure 1 - The first visualisation design.

Effectiveness Argument

Since the dataset contains 10 features it could have been represented in multidimensional ways such as 3D visualisations. However they can be known to make it difficult to compare values. Instead it was decided to encode data only in the position and colour hue.

Position is a very interpretable dimension, so it was chosen to represent the following features, age of woman, number of children, and the count of SOL index. All of these values are quantitative and position is known to be a good method of representing quantitative data (Clevelands and McGill, 1984).

Colour hue represents the form of contraception as this is a nominal data type and colour is a very good way of representing that (Forrest, 2015). Colours were chosen that would be easily interpretable by having a range of hues. An online tool was used to ensure that the colours would be suitable for people with colour blindness also (Brewer and Harrower, 2013). For the 3 contraceptive categories the colour values of #66c2a5, #fc8d62, and #8da0cb were chosen.

To avoid the problem of overfitting the marks in the scatter plot are made slightly transparent and the points may be jittered slightly. To jitter the data a small amount of noise was added to the values of the age of the women, the small decrease in accuracy caused by jittering is offset by its effectiveness of solving the problem of overfitting(Few, 2008).

When plotting any data it must be considered how the information may be perceived by a user. It has been shown that an increased scale of the axis can have an effect on the perceived correlation of two variables being higher compared to an axis with a smaller scale (Cleveland and Diaconis, 1982). Therefore care was taken to ensure that the scale was large enough to show the distribution of marks, but small enough as to not introduce any artificial correlation.

Design 2

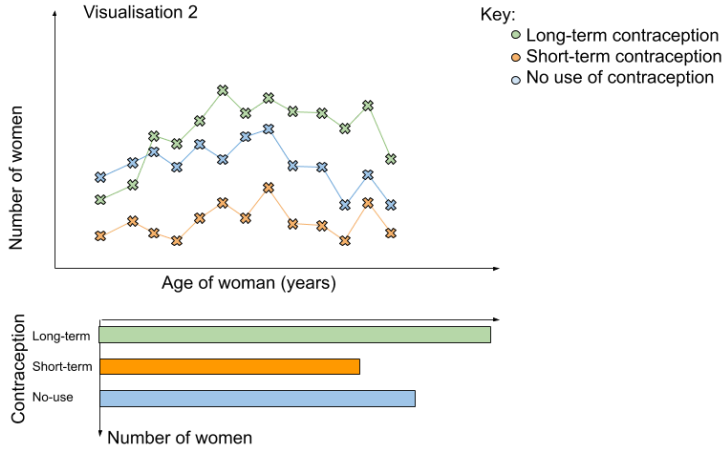

Figure 2 - The second visualisation design without user interaction.

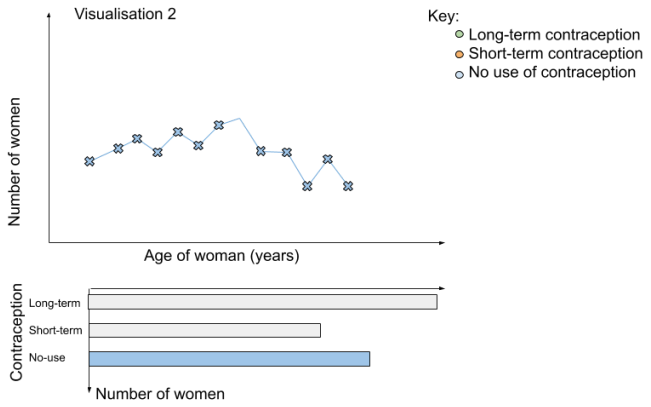

Figure 3 - The second visualisation design showing only the data from a specified form of contraception.

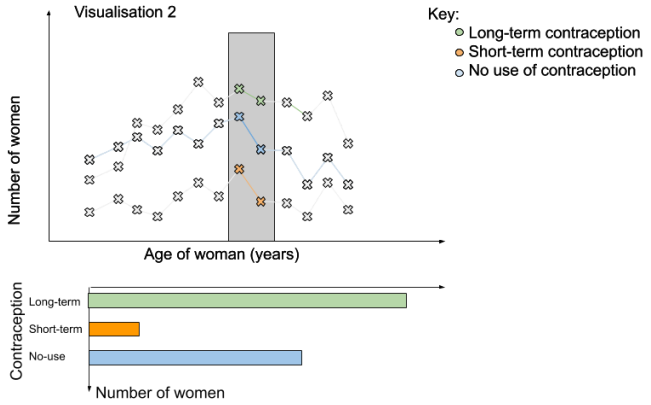

Figure 4 - The second visualisation design showing how data from a range on the x axis can be shown.

Effectiveness Argument

While the first design has its benefits, it is possible that all 1472 records may be too many to points to effectively visualise and jittering is not enough to make the data interpretable due to massive over-plotting. A method not yet mentioned to reduce the impact of over-plotting is to reduce the number of values being shown at once by aggregating the data (Few, 2008). Aggregating datasets is meaningful ways is a known and proven way of simplifying large quantities of data (Kelleher and Wagener, 2011). Figure 2 shows how many less points there will be compared to figure 1, due to aggregation. Additionally by using Altairs interactive features it is possible for a user to easily filter the data which can highlight certain trends.

In figure 3 it is shown how users can have the option to focus in on just one of the contraceptive methods used, removing emphasis from the others. The points and lines belonging to other methods are completely removed as to reduce the number of marks on the screen which would be unnecessary information when looking at just one form of contraception. Information that is not being used is redundant and can overcomplicated the visualisation, confusing the user.

Information that is not being used is redundant and can overcomplicate the visualisation, confusing the user (Kelleher and Wagener, 2011).Users may see trends showing the comparative use of different forms of contraception over time or compare those features for different age groups. Figure 4 shows how a horizontal range can be specified and only data inside that range will be coloured on the graph, and only that data will be shown on the bottom linked view histogram.

As with the first design colours were chosen that would be easily interpretable by having a range of hues. There was no need to change the colours in the second visualisation as the same categories of contraceptive method is being represented and the colour chosen was already well justified.

Implementation [50 marks total]

Implementation prototype [30 marks]

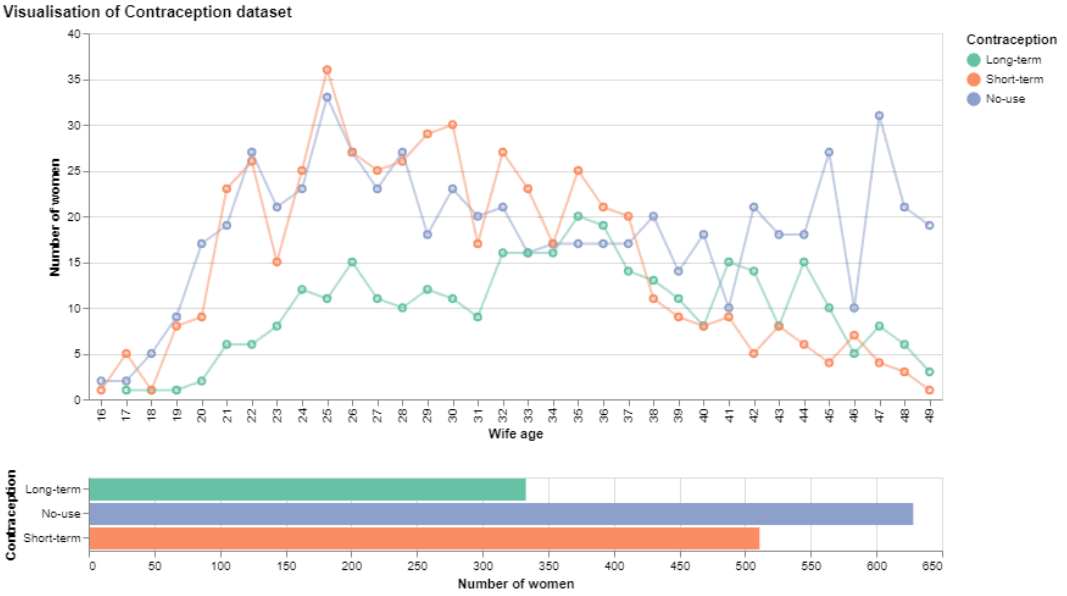

Figure 5 - The implementation prototype showing how data from a range on the x axis can be shown.

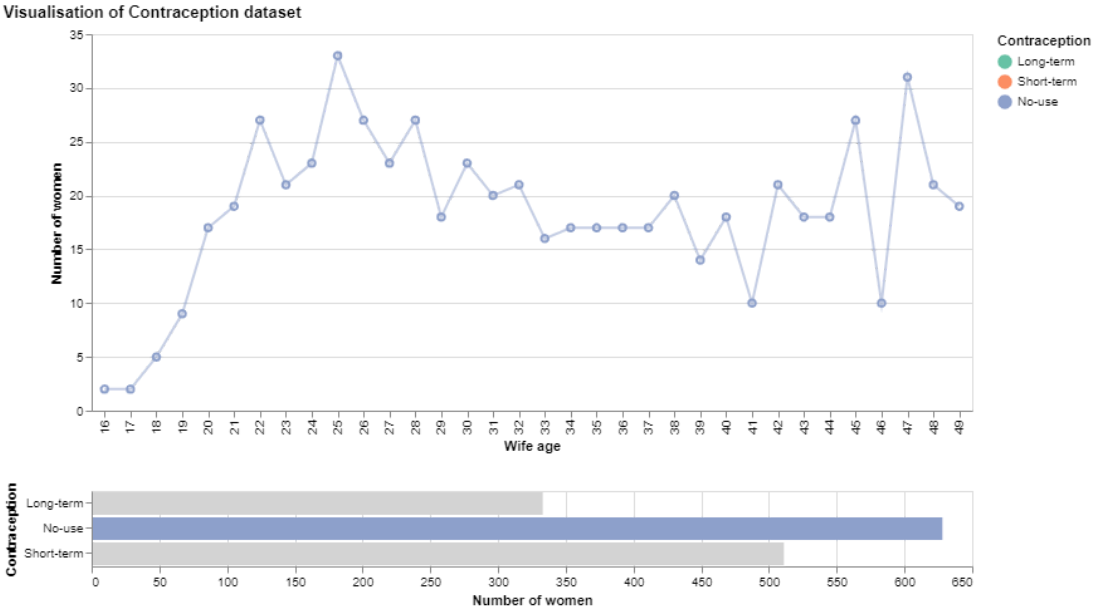

Figure 6 - The implementation prototype design showing only the data from a specified form of contraception.

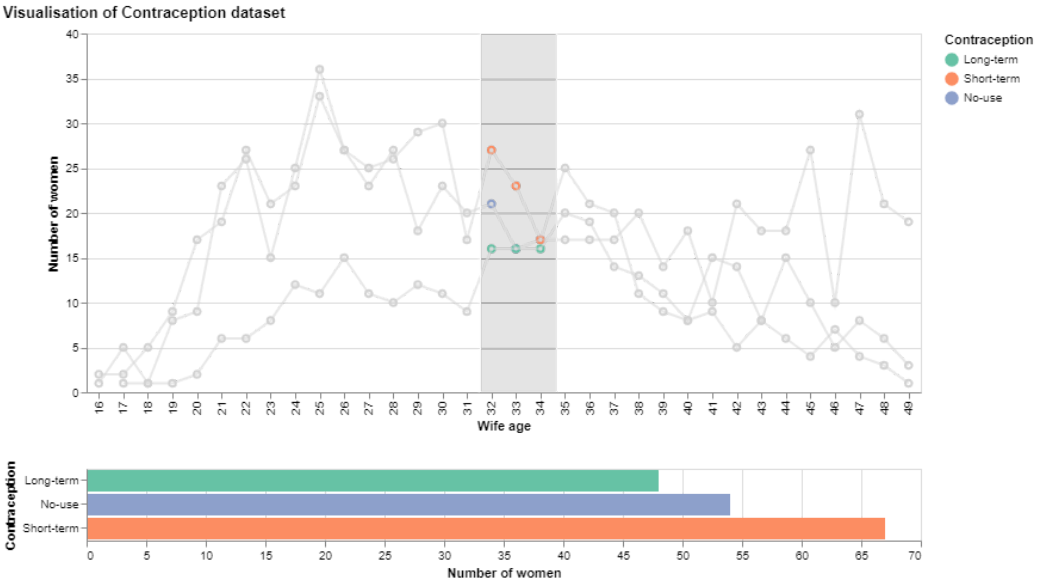

Figure 7 - The implementation prototype showing how data from a range on the x axis can be selected.

Description of discovered features

From looking at the visualization prototype with no interactive features as shown in figure 5 interesting trends about population can be seen.
When compared to the total population there is only a small number of women under the age of 21 on the survey, a large spike at the age of 25, and another dip around the age of 40.

Just by looking at the histogram beneath the main chart in figure 5 it is clear that for the Indonesian woman population surveyed no-use of contraception is the most popular choice of contraception, followed by short-term methods. Both of these are much more popular than the long-term form of contraception, which is used around half as much compared to no use of contraception.
When looking at the main plot of number of women against wife age in figure 5 trends in popularity of different methods for women of different age groups become apparent. Women aged under 35 appear to have no clear preference of method between short-term contraception and no-use of contraception as they are always close with the total of short-term being slightly greater than no-use as shown in figure 8. It is clear however that women in that age group seem to avoid long-term methods of contraception.

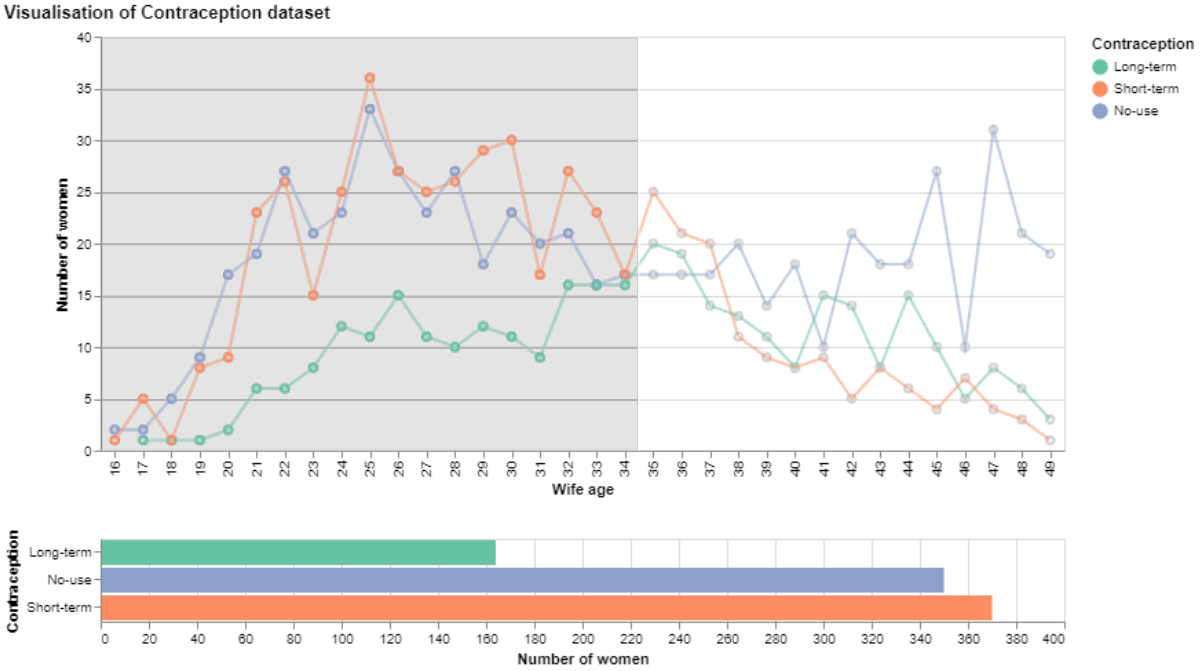

Figure 8 - The implementation prototype showing data for women under the age of 35.

When using the visualisation to examine women above the age of 34 it is shown in figure 9 that the group has a strong preference for using no contraceptive methods. The next most popular method is long-term with short-term contraceptive methods being the least used. This reveals an interesting divide in the dataset where women younger than 35 and women 35 and older have distinctly different preferences of contraceptive methods.

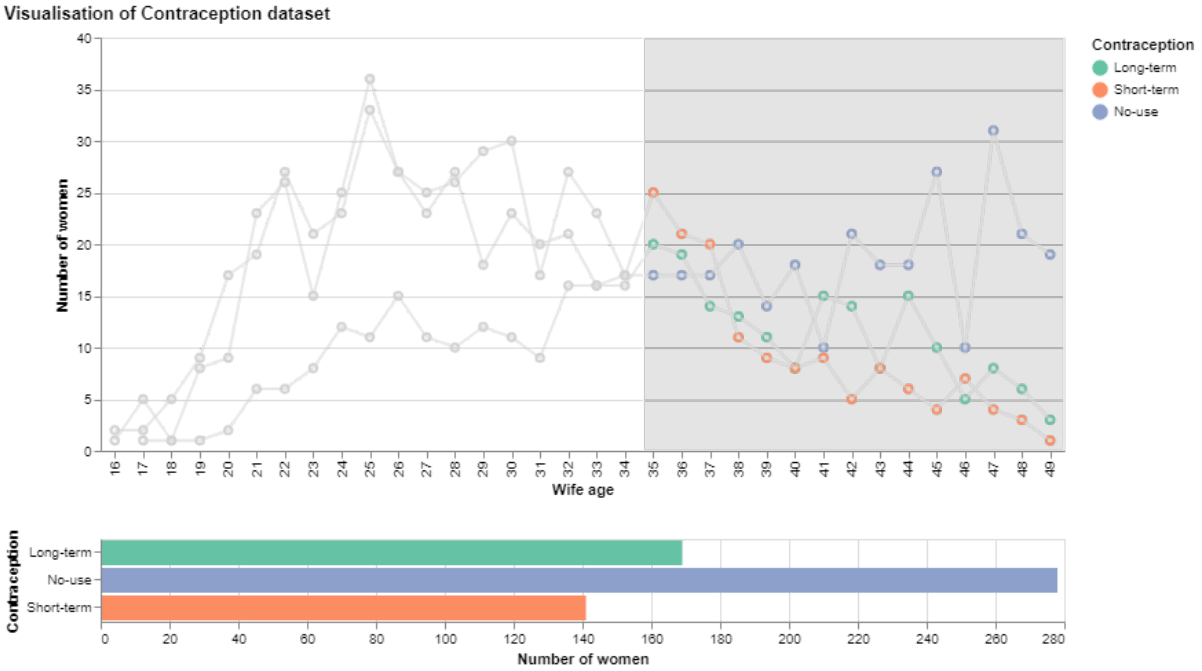

Figure 9 - *The implementation prototype showing data for women aged 35 and over.*

A report on the Contraceptive survey states that nearly half of the women are currently using a contraceptive method (Demographic and Health Surveys, 1987). In contradiction to this, the visualisation shows that the majority of women use some form of birth control, this may be explainable since the dataset that was used to create the visualisation was only a "subset" of the actual results so it is understandable for them to differ slightly (Dua, D. and Graff, 2019).

References:

Brewer, C. and Harrower, M. (2013). *ColorBrewer: Color Advice for Maps.* [online] Colorbrewer2.org. Available at: http://colorbrewer2.org/ [Accessed 8 Nov. 2019].

Cleveland, W.S., Diaconis, P. and McGill, R., 1982. Variables on scatterplots look more highly correlated when the scales are increased. *Science*, *216*(4550), pp.1138-1141.

Cleveland, W.S. and McGill, R., 1984. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387), pp.531-554.

Demographic and Health Surveys (1987). *National Indonesia Contraceptive Prevalence Survey 1987*. [online] p.18. Available at: https://dhsprogram.com/pubs/pdf/SR9/SR9.pdf [Accessed 8 Nov. 2019].

Dua, D. and Graff, C. (2019). *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science. Available at: https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice [Accessed 8 Nov. 2019].

Few, S. (2008). ‘Solutions to the Problem of Over-Plotting in Graphs’, *Visual Business Intelligence Newsletter*, October
Forrest,D. (2015). ‘International Encyclopedia of the Social & Behavioral Sciences’ (Second Edition), 2015, Pages 260-267

Kelleher, C. and Wagener, T. (2011). Ten guidelines for effective data visualization in scientific publications. *Environmental Modelling & Software*, 26(6), pp.822-8