CSCM27 – Assignment 1

Dataset and Task

The dataset I have chosen is to do with heart disease. It is composed of 14 data columns which include:

- Age (ordinal)
- Sex (nominal)
- Chest pain type (nominal)
- Resting blood pressure (quantitative)
- Cholesterol (quantitative)
- Fasting blood sugar > 120 mg/dl (nominal)
- Resting ECG result (nominal)
- Maximum heart rate achieved during exercise (ordinal)
- If reported angina was exercise induced (nominal)
- Diagnosis (nominal)

I thought it would be a good idea to ignore some of the data that I didn't understand the relevance of, as I do not have a medical background. The data remaining should still have some visible relationships within it if portrayed correctly. The version of the dataset I have used is the Cleveland one.

I thought a suitable task would be for a user to estimate the risk of a potential heart disease diagnosis, given a value for several of the factors mentioned above.

Link to the dataset: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Designs

Design 1

Given how many dimensions the dataset has, I thought that a variety of graphs comparing different elements could be helpful with visualising it. Given the task that the user is expected to perform, the diagnosis feature is almost certainly the most important one to include in graphs, so the other elements of the data that most heavily relate to that one are important for comparing against it.

I looked for resources on heart disease, particularly the most prominent factors contributing to it. I found an article (NHS, 2022) that stated a list of 5 of these factors that I had data on, so I could work around that.

The idea I had for this design was to produce a selection of heatmaps, each featuring one of these contributing factors compared against another. This way a user would be able to tell at a glance how high the risk might be for a diagnosis, as each plot would be broken down into separate blocks of differing colour intensity, in turn highlighting the most "dangerous" values. I thought this would be a good idea as heatmaps make extensive use of colour, which the user is likely to be very receptive to, as colour intensity (or saturation) can be more evident than positional attributes when creating graphs.

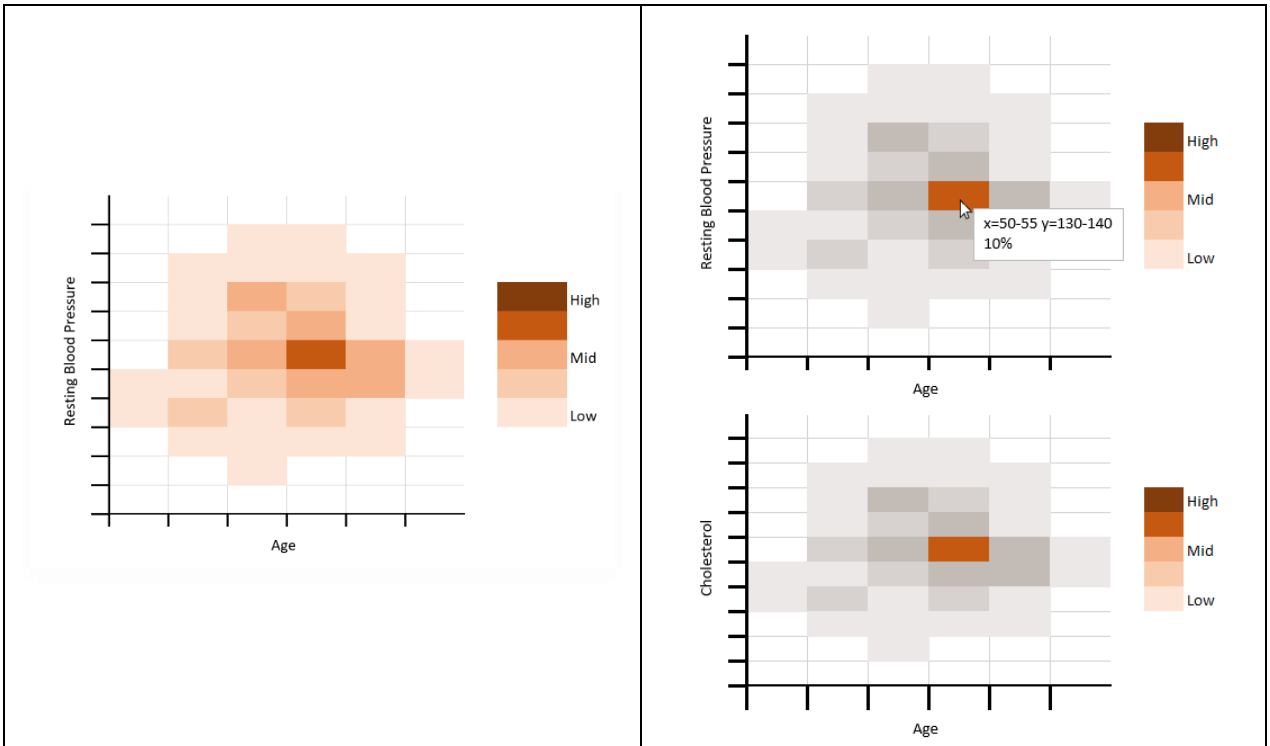

Figure 1: An example heatmap.

Figure 2: An example interaction showing multiple views.

The interaction I wanted to add for this would involve the user clicking on a given block, which would reveal the percentage of patients from the data (who received a positive diagnosis) that fell within that group as well as the values from the axes according to their position. The action would also lower the colour saturation of all the other blocks on the chart, focusing the user's view and conveying which element had been selected.

Again, as users are likely to be receptive to this change in saturation, I thought it would be a good idea to add that. As for the values that appear, I thought this would be useful if the user had mistakenly selected the wrong part of the plot, and by informing them exactly where they are they would be able to correct that. The percentage data however may not be as good of an idea. I thought it would be helpful for assessing the likelihood that their given values would receive a positive diagnosis, but realistically I think it would be misleading, especially when considering the small sample size of the data.

Another interaction would be co-ordinating the multiple views that the user is presented with, so when they hover over a given value on one heatmap, an associated value (with a matching value) would be highlighted on another plot.

I thought this would be useful given the fact that multiple graphs would be an element of this design, so showing the connections between them would be helpful. However, when examining the data and building a few initial versions of it using Altair, I saw that the individual factors behind heart disease (or even 2 or 3 factors) weren't enough to produce a convincing or useful enough correlation that someone could perform a task with them. The mix of similar yet slightly different graphs, and the weak relationships shown by them was mostly just confusing to look at. In the end, the dataset dictated that a different design be used.

Design 2

After the issues with the previous design, I decided that a compound value (one created by combining columns from the original data) might be more helpful as far as seeing actual trends within the data, and helping the user perform their task. As much as this is bad practice for high-dimensional data, I found through testing that it makes the relationship between the risk factors and rate of diagnosis much easier to see.

I did this very simply using Altair's `transform_calculate` function, using this formula:

$$age \times \frac{restbp}{100} \times \frac{cholesterol}{100} \times \frac{1}{maxhr} = CRV$$

With CRV here meaning compound risk value. With some research I probably could've found more accurate ways to weight each of these variables, but since that's an optimisation problem and this formula was producing good enough results to correlate, I didn't want to waste any time.

I still wanted to include multiple plots, as there were so many more columns in the data to look at, and the coordinated views seemed interesting to implement. I had the idea of having 3 separate plots, one for correlation within each sex and then also a combined plot to show the trend of risk factors against diagnosis.

The idea for this design was a series of line charts that showed risk value against number of diagnoses, each chart showing a given sex and then a final chart which would have them both combined into one. Line charts are basic when compared to the heatmap design, but I felt it would show the data much better in this case, where the sample size for the data is relatively small, and the differences in values can be subtle.

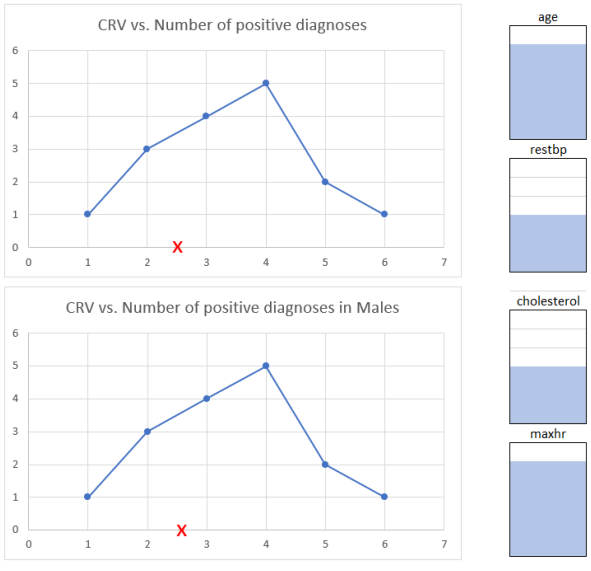

Figure 3: An example layout.

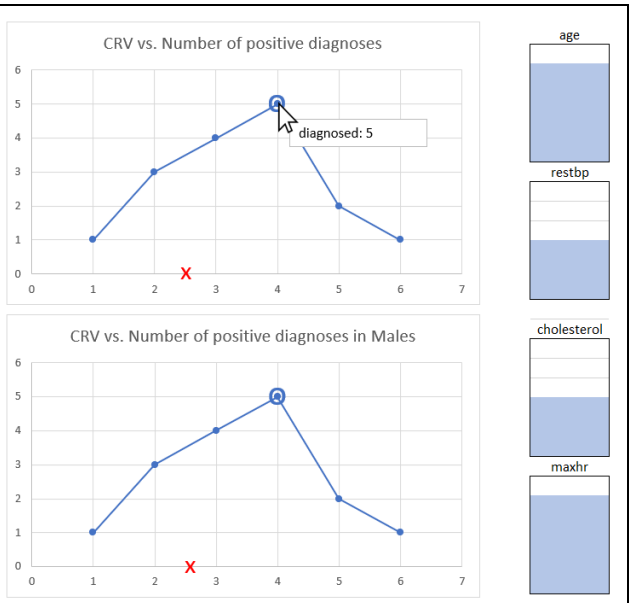

Figure 4: An example interaction showing multiple views.

As far as the interaction, I wanted to keep all the ideas from the previous design, as I didn't see any issues with those. Showing the user important values relative to where they had clicked on the chart as well as co-ordinating the various views seemed like valuable tools.

I decided against showing any kind of percentage, as it could be misleading to the user, in case they interpreted it as anything other than the percentage of diagnoses that fall into a certain bin. This could be dangerous if the user were to interpret it as something like "the percentage chance that a given value will be positive". Instead, I opted to only display it as the number of records at that given point. I liked the idea of highlighting the important data though, so I wanted to keep an element of that.

Another idea I had for an interaction involved sliders for each element in the CRV formula, allowing the user to input values using them to show a graph element along the line at a given calculated CRV. This would allow the user to complete their task more easily. This is an ambitious interaction, but it would greatly simplify the confusing issue of the combined value.

Implementation

For my implementation, I was mostly able to build my second design, except for the slider interaction that calculates the risk value. This proved particularly difficult to re-create in Altair. I made some changes on the initial design that would hopefully show some interesting features within the data, such as including the negative diagnoses overlayed with the positive ones.

The first interesting feature I noticed was that the peak for negative diagnoses had much lower risk values than that of positive ones. This should be a given, as they are risk factors for a reason, but it at least shows that this dataset lines up with the advice given by the NHS website. I think my visualisation shows this well, with each plot effectively dominating a separate side of the screen, making it clear that low risk factors are more likely to result in a negative diagnosis.

The next interesting feature was that the female graph had far less data. Initially I thought this might be because rates of heart disease were lower in women, but after some research on this I found that the answer was more complicated than that, so in this case it's most likely because of bias within the original dataset. Whether it's an issue with the dataset or not, I think that the fact that someone could think it from looking at my visualisation shows that it's not very good. This could be slightly fixed by adjusting the axes on the graph to be more in line with the other two.

This does lead on to the next interesting feature though, which is that the peak in positive diagnoses on the female chart is much further along on the risk value axis than it is on the male one, which is likely because heart disease develops later in women (Appelman, 2010). I'm not sure as to whether my visualisation shows this well or not, I would say that it does not, as the lines can be confusing to look at, especially when they intersect as closely as they do on the female only chart.

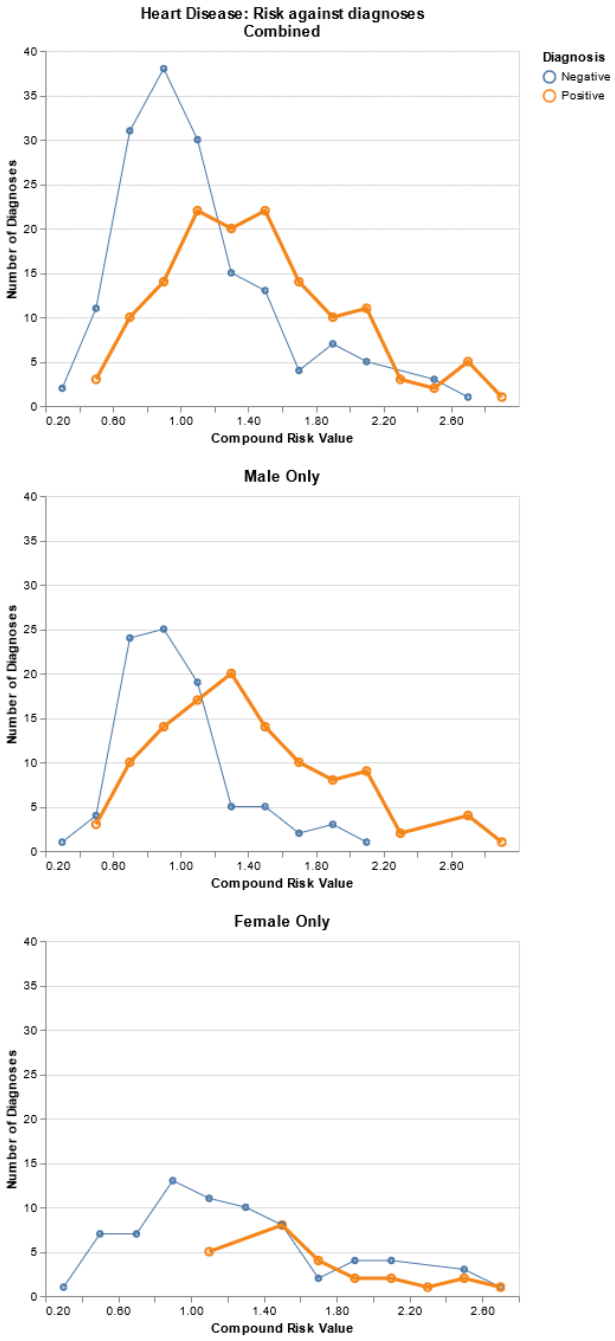

Figure 5: Completed implementation

Conclusion

Overall, I don't think my visualisation is great. I think that this is in part due to the dataset, as it was too small to establish a great deal of meaningful trends from, but mostly due to my poor design choices. While it can highlight some interesting features of the data, I don't think it would help a user very much with their initial task of estimating a diagnosis risk given relevant data. Looking back now, I think that a better choice for representing the higher-dimensional data could've been to use a tool like SPLOM (Scatterplot Matrix) to make sense of it.

References

Maas, A. and Appelman, Y. (2010). Gender differences in coronary heart disease. *Netherlands Heart Journal*.

NHS. (2022, 11 8). *Cardiovascular disease*. Retrieved from nhs.uk: https://www.nhs.uk/conditions/cardiovascular-disease/