Jason Moens
COSC 5610 Data Mining

Data Exploration and Visualization

The Heart Failure Prediction data set has 299 separate entries of applicants to this clinical data set, with 13 categories ranging from whether the patient has diabetes to the determining if the patient had high levels of certain elements in their body, such as creatine-phosphokinase. These all culminated with a final variable, "DEATH_EVENT" which is the status of the patient by their next interaction with the doctor. All 299 sets were complete, so there were no entries that needed to be removed. However, two applicants put their age as 60.667 rather than the normal method of stating your age like 60 or 61. This was the only "blemish" in the code, so to combat it, their ages were truncated to 60.

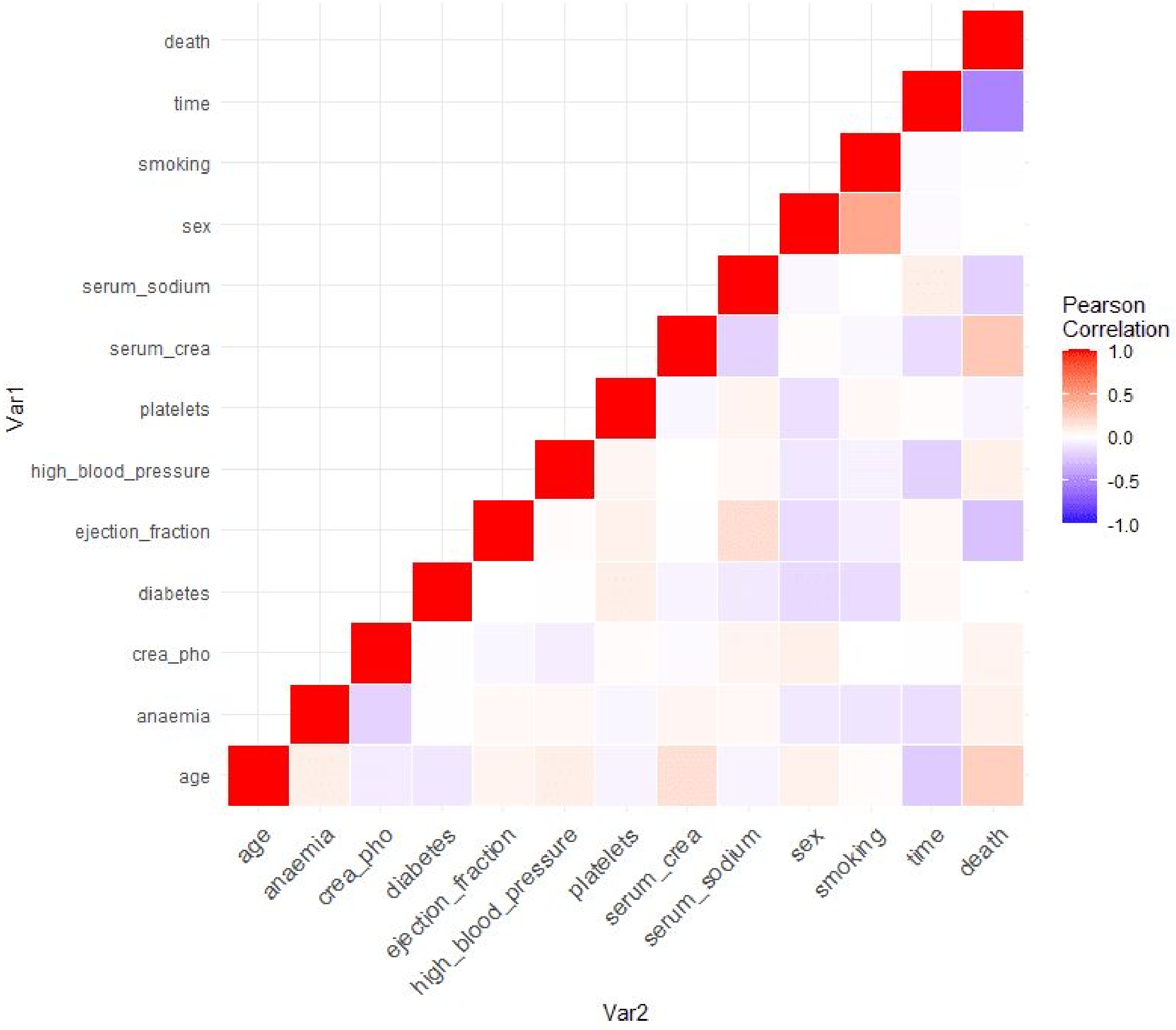

Figure 1. This is a heatmap of the correlation between all 13 values.

As I am not a medical professional, I do not know much about some of these values, such as the meaning of creatinine phosphokinase (crea_pho), so I began by putting the 13 different variables into a correlation matrix to see which stood out. Since we are researching whether certain conditions have an affect on the heart failure of a patient (death), we need to see which correlations are more extreme compared to the final column “death”. Since some of our values are not categorical, we need to make sure that we take both extremes into account. From this we can gather that “age”, “ejection_fraction”, “serum_crea”, “serum_sodium”, and “time” are significant factors.

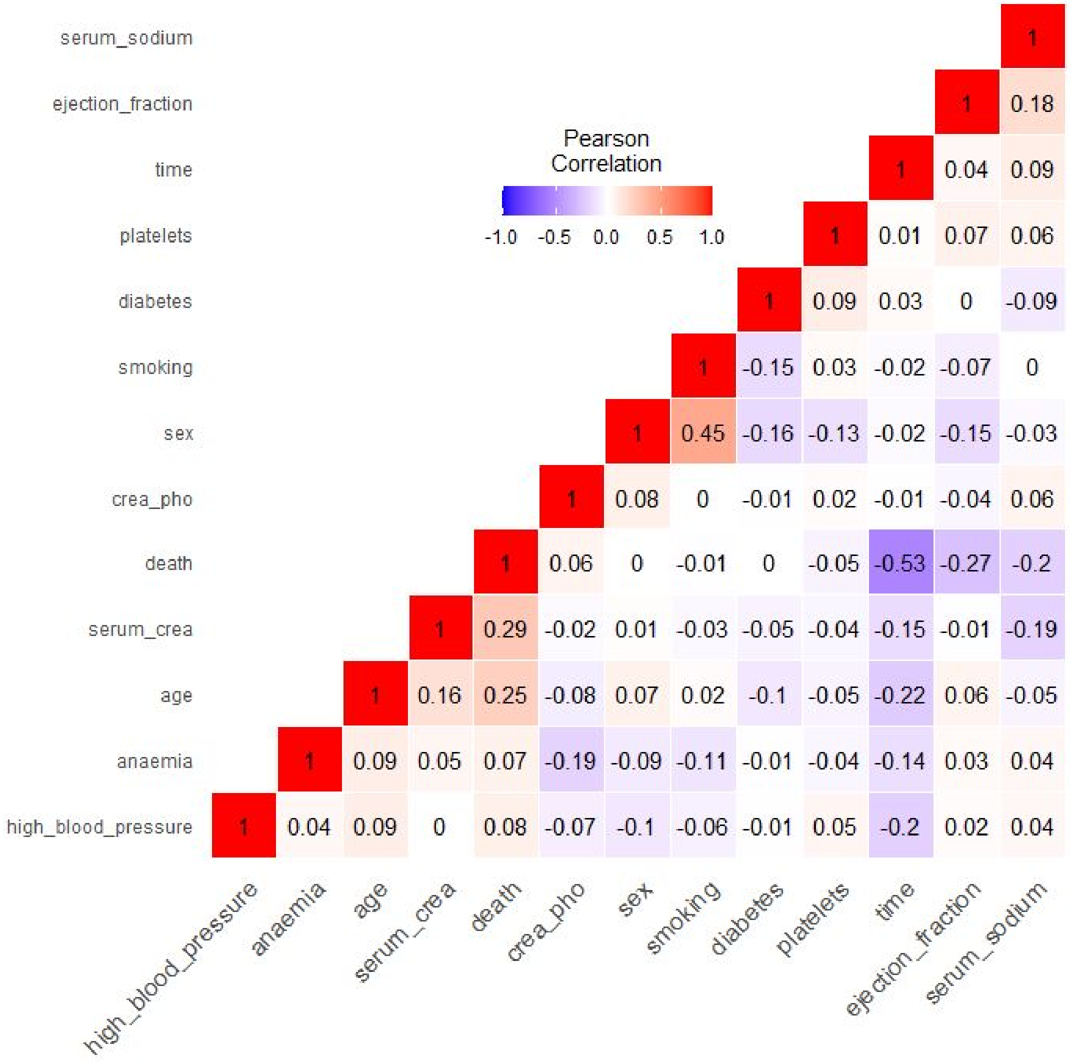

Figure 2. Correlation matrix is reordered and numerically labeled

Following Figure 1, Figure 2 gives a more numerical interpretation of the matrix showing us significant correlations between two values. We are still interested in the “death” column, and now row as the values have changed locations to better represent the correlations. Now that numerical values are given, we can see that the significance of the previous variables has an order: time, serum_crea, ejection_fraction, age, serum_sodium, in that order.

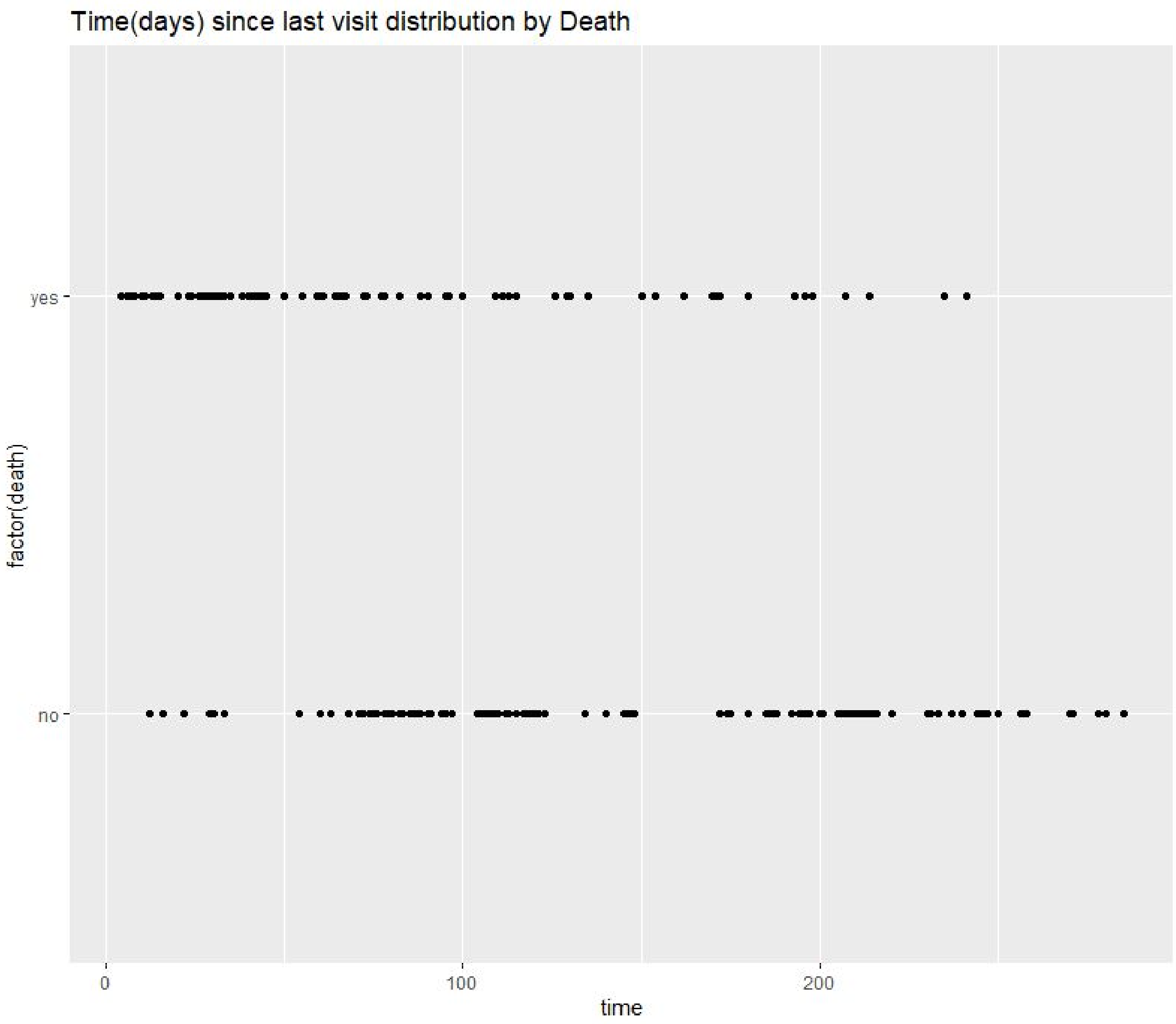

Figure 3. Time since last visit, separated by the whether the patient died or not

Our first variable, time, seems like it may have an incredibly high correlation with death. However, something needs to be taken into consideration first. How does the patient come for a checkup if they were to die after their initial checkup? I believe that this happens because either a relative or caretaker calls the doctor to say that the patient has passed, or they are found to be dead by the doctor. I believe that this time variable is actually not needed at all, and if used in a model, will skew the model heavily without actually providing any valuable input. A majority of these patients that come to the doctor are likely exhibiting symptoms that would prompt a visit to a doctor. So this variable should be eliminated from the model to preserve the model's accuracy.

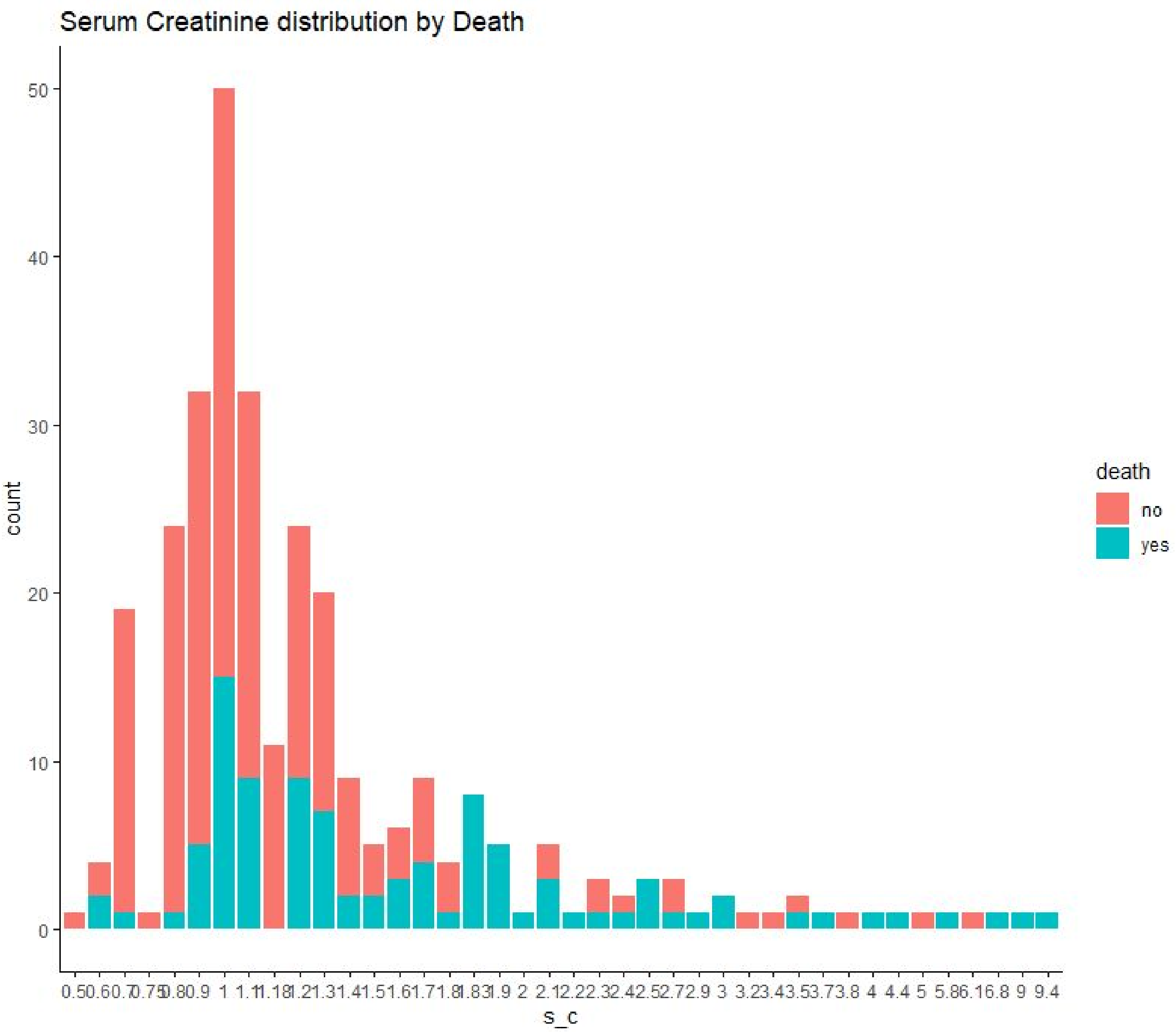

Figure 4. Distribution of Serum_Creatinine found in patients split by death outcome.

The next most death-related variable was the serum_creatinine within a patient. As we can see from Figure 4., a majority of the patients had a level of serum creatinine from 0.8 to 1.3. However, this graph stretches to 9.4 at the top end. We can see here that the higher the levels of serum creatinine are in a patient's body, the higher likelihood they will die, as the further right you go on the graph, the more blue and less red the graph gets. This is reflected by our correlation value from the heat map of 0.29. The higher it gets, the higher likelihood of the death event taking place.

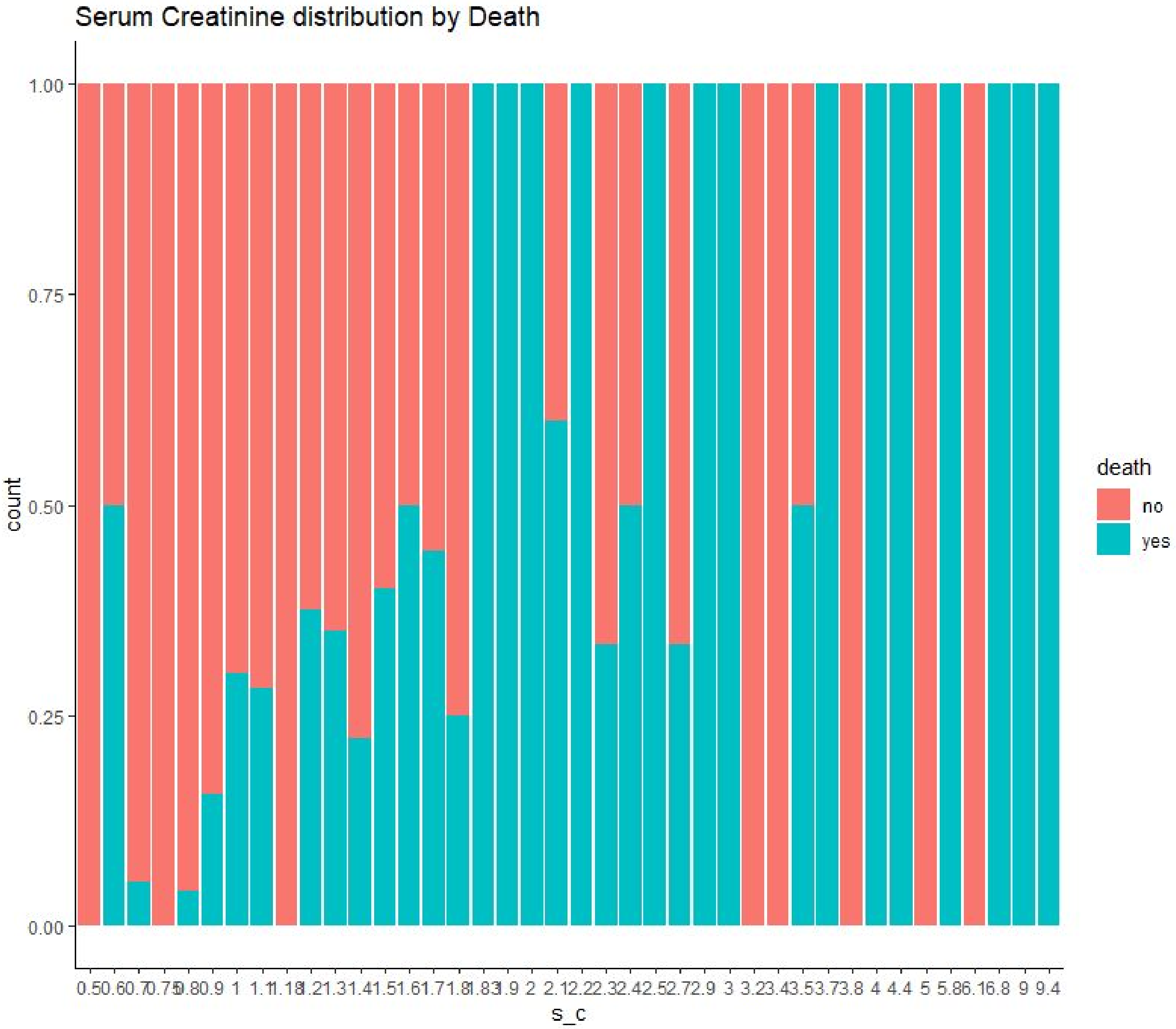

Figure 5. Percentage death distribution of the serum creatinine variable

Figure 5. Shows us further that a majority of the deaths come from higher levels of serum creatinine in the patient's body. The late solid lines with 100% 0 cases of death can be accounted for by single patients, as evident in Figure 4.

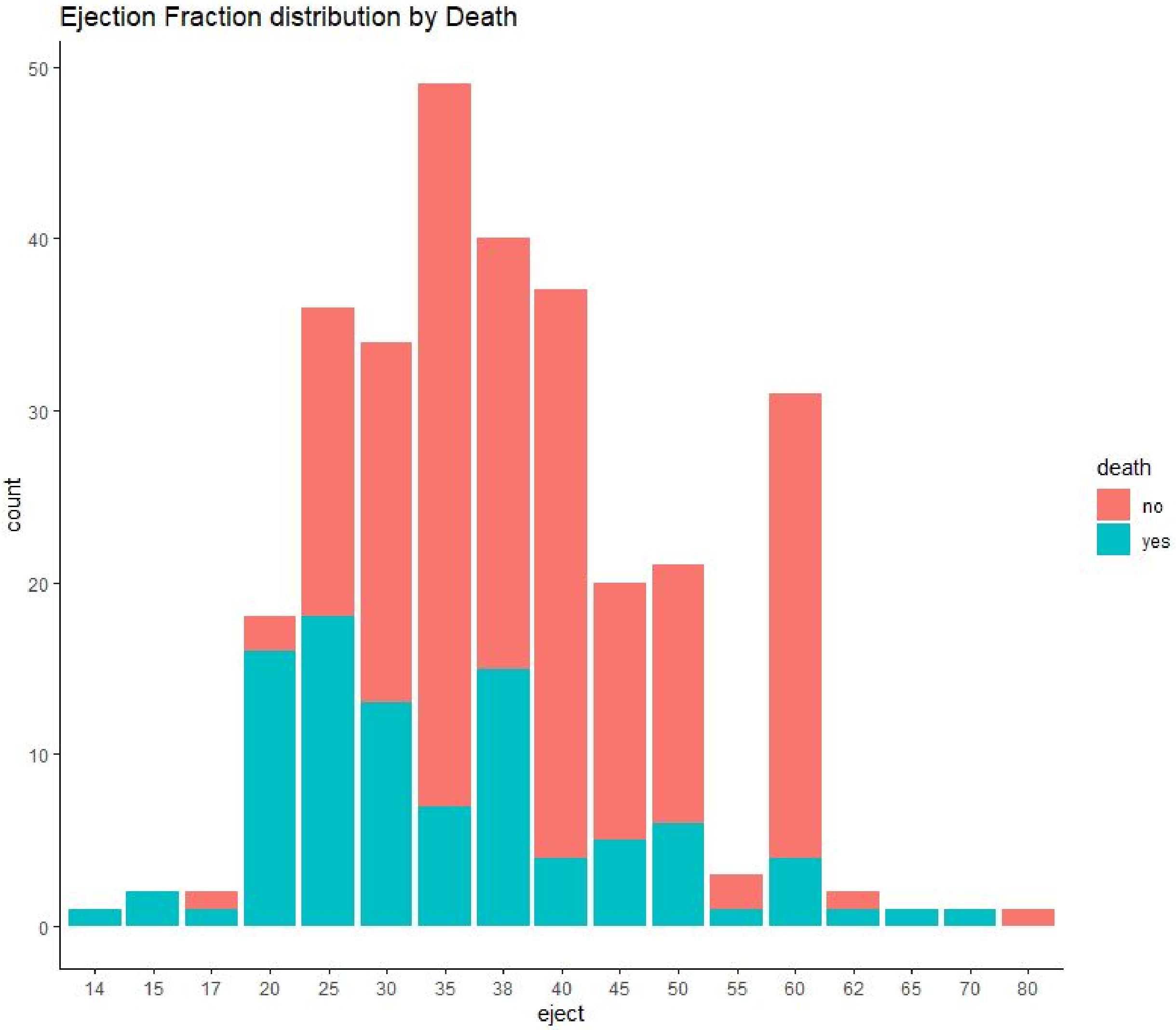

Figure 6. Ejection Fraction distribution, separated by death event

As can be seen by Fig. 6., the average ejection fraction lies somewhere in the middle, around 38. We can also see that both extremes tend to lead to death, while having a heart that does not pump out as much blood is a clearer sign of heart failure. This is evident by our -0.27 correlation value. The lower the ejection fraction is, the higher the chance of death.

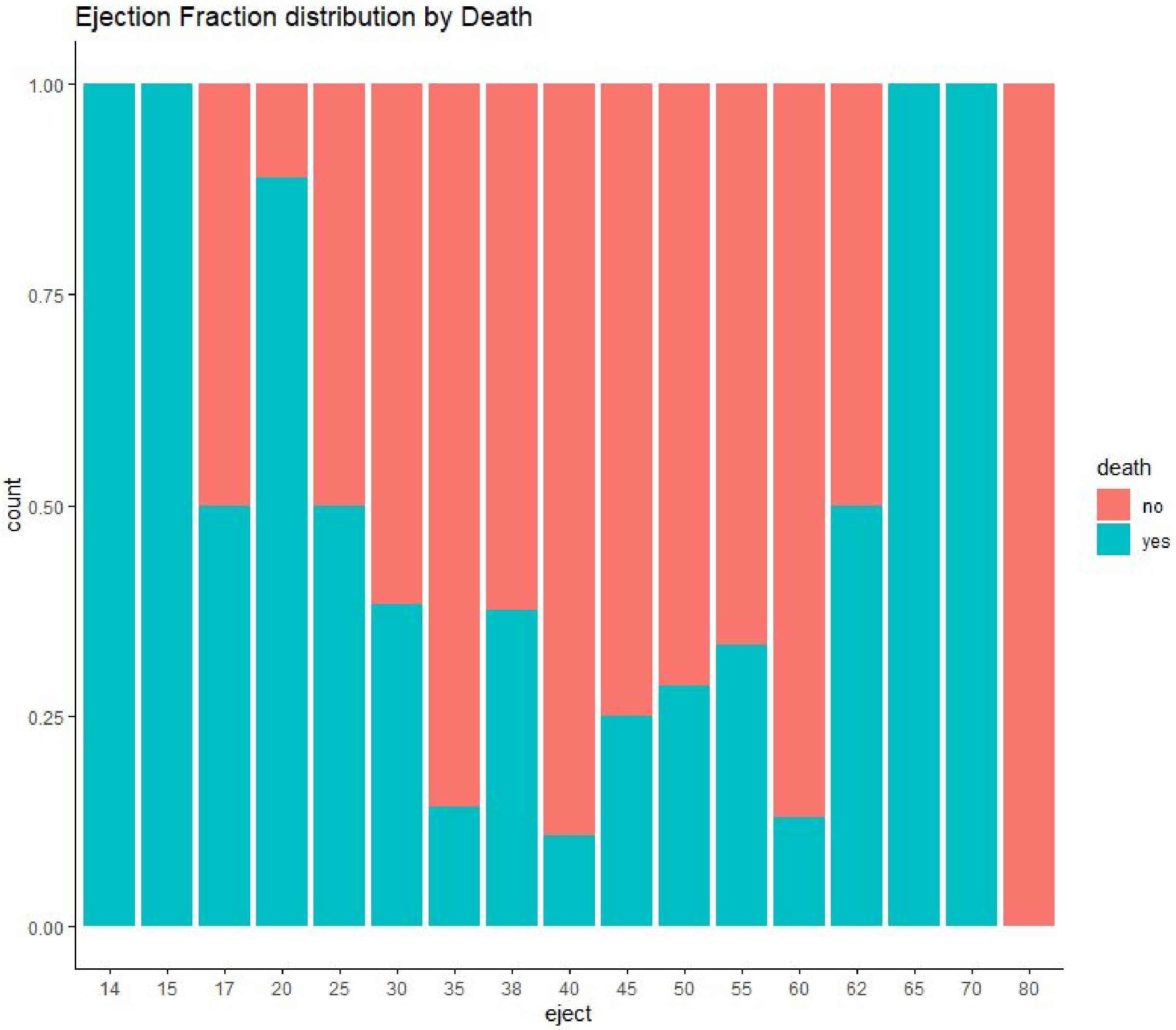

Fig 7. Ejection fraction percentages distribution.

Similar to the previous percentage distribution, Figure 5, the 80 value's 100% non death can be attributed to one case, and as such should likely be ignored while talking about the distribution. We can see that the average ejections, sitting near the middle of the graph, tend to produce the least deaths. This is likely because ejection fractions mean that the heart is not pumping as much blood out, that it is weaker or dying, while hearts that are pumping out too much blood are overworking themselves, quite literally to death.

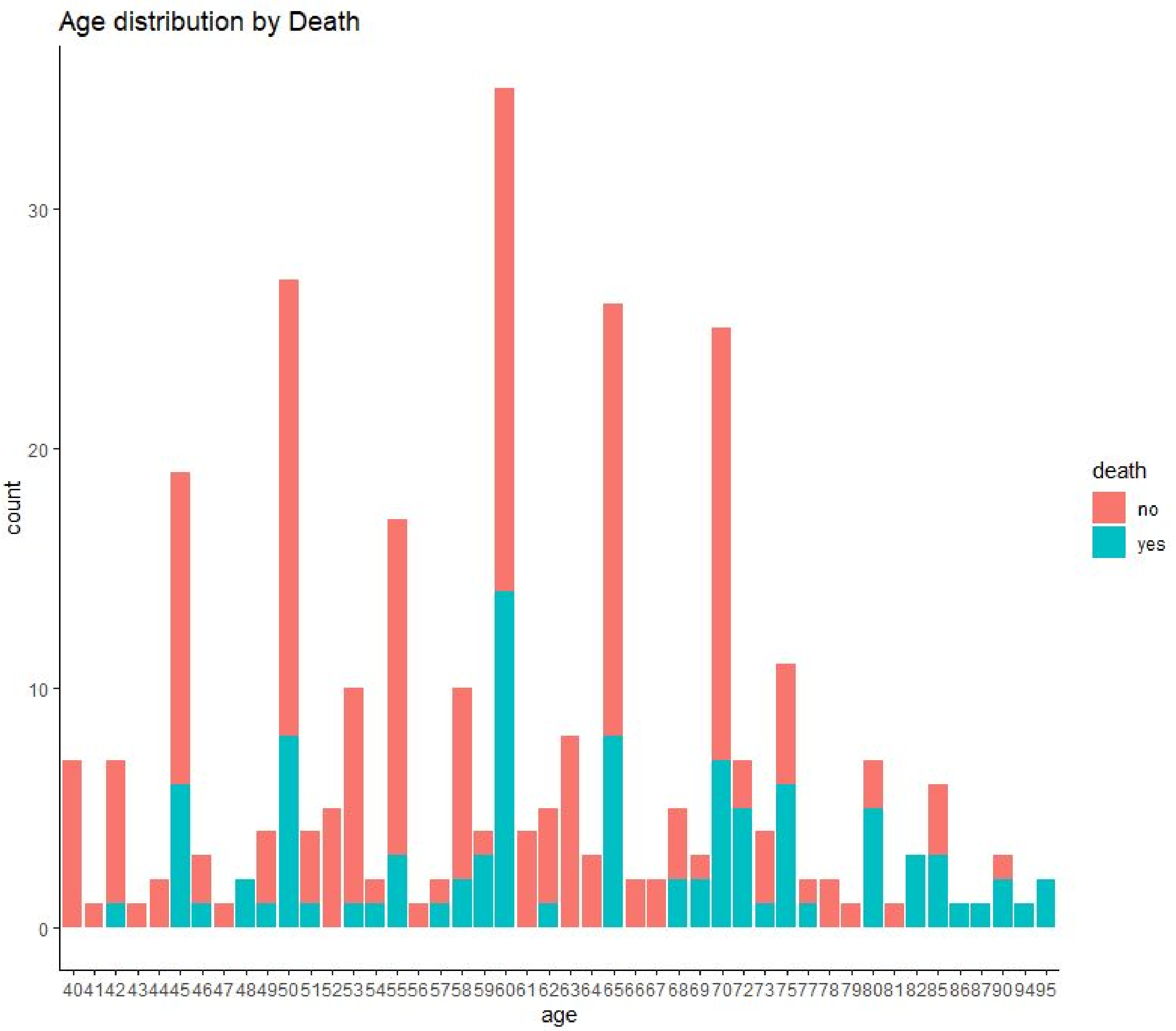

Figure 8. Age distribution, separated by death event

The age distribution is much more separated than any of the other variables that we have taken a look at so far. As the correlation matrix indicates with its 0.25 correlation, the older you are the more likely there is a death event. One thing to note is that this death does not take into other considerations other than the variables listed. As people get older, lots of other things begin to stop working, so it is plausible that some of these patients did not die related to a heart issue, but were counted as such since they went in for a checkup at some point before death.

.

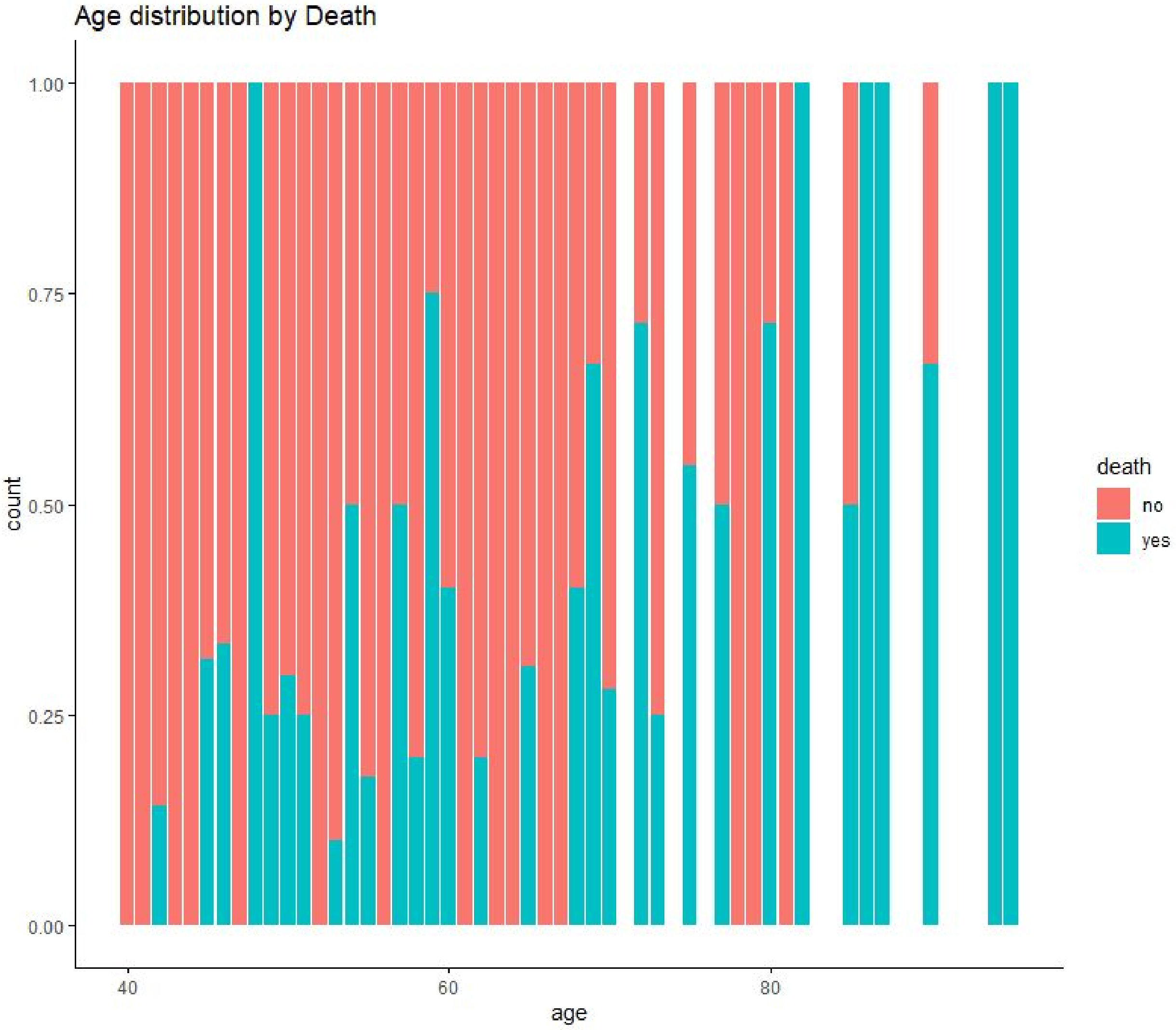

Figure 9. Age distribution by percentages and by death

To continue from Figure 8. We can very blatantly see that the farther right you go, the more likely it is you die. Compared to the other percentage graphs, these 100% death events can be attributed to the small sample size at that particular age. However, in this case there are a few samples at some of the later ages. Each blank white spot indicates an age that was not recorded by any of the patients.

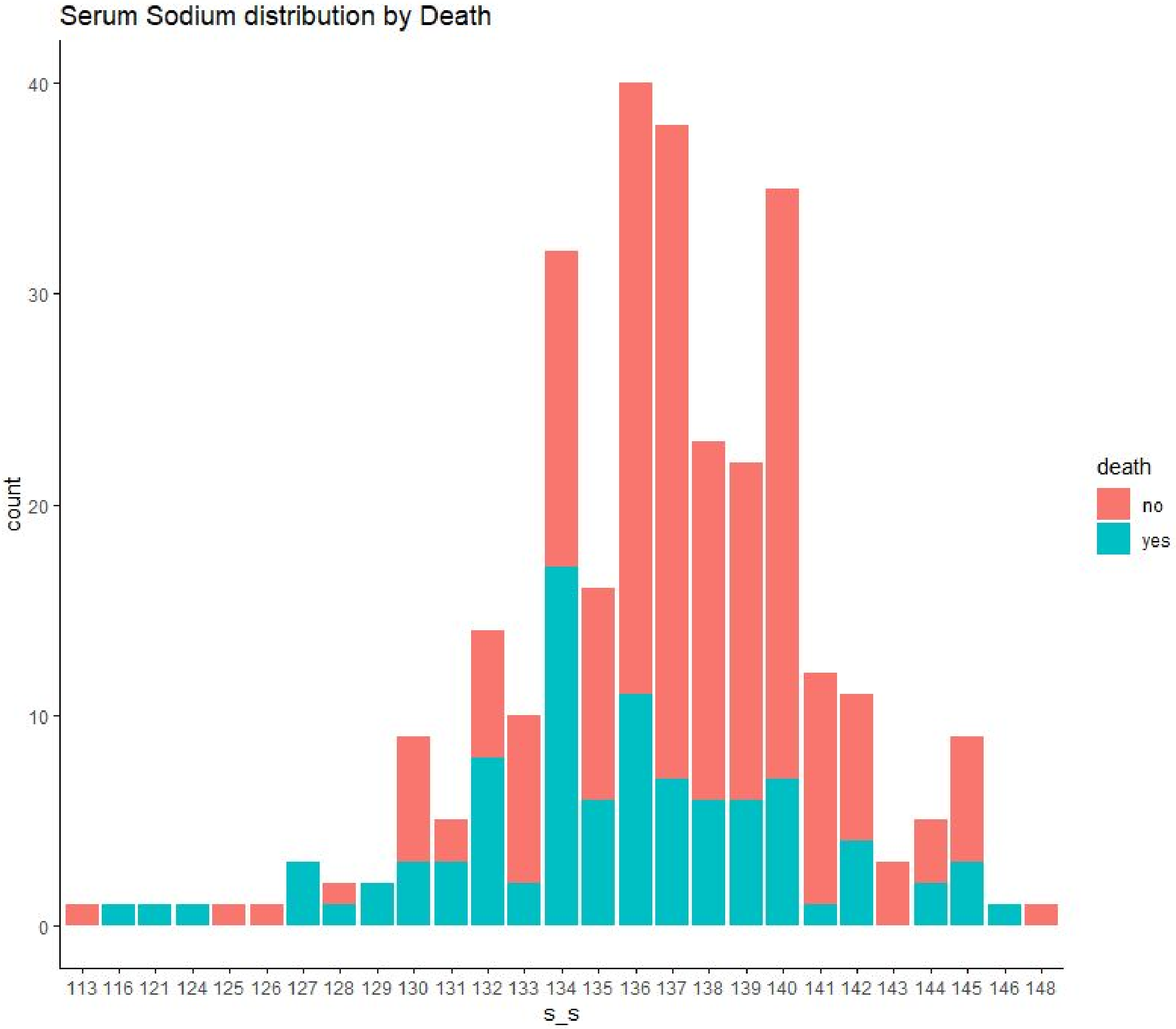

Figure 10. Serum sodium distribution between patients separated by death result

The final variable is the level of serum sodium found in the patient's body. In the above figure, we can see that there is a reasonable distribution between the patients. However, we can see that the lower the serum sodium is in a patient, the higher likelihood they will die. This is reflected by our heat map as serum_sodium and death had a -0.2 correlation, that less serum sodium in a person, the higher chance they would die.

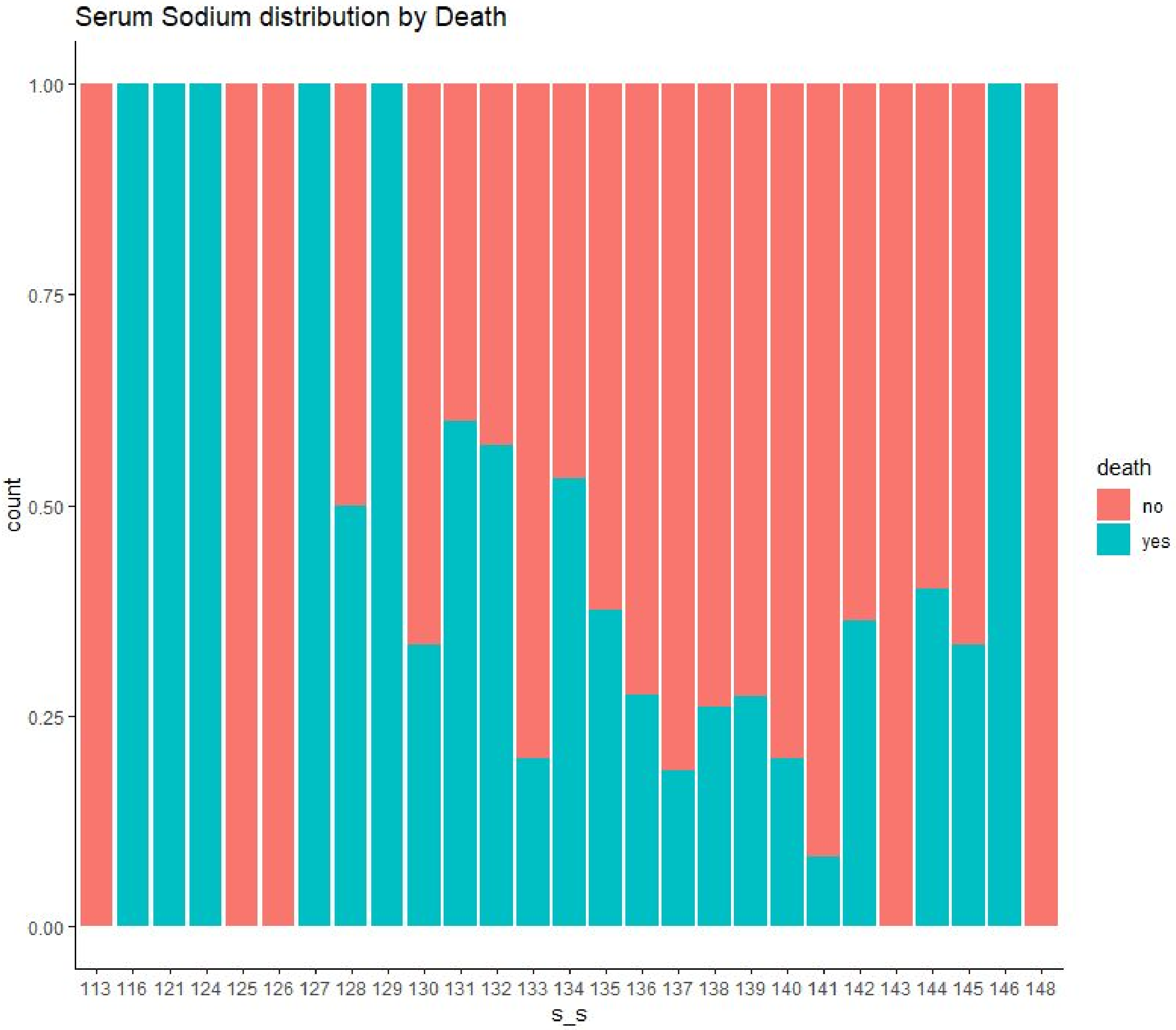

Figure 11. Percentage distribution of the serum sodium distribution

Once again, similarly to the previous figures depicting percentages, we can see that there is a general trend that both extremes lead to a higher likelihood of death. However, when we take a look at the -0.2 correlation we can see that the death portion of the graph is leaning to the left, as the percentage goes down, the likelihood of death increases.

References:

Anon. Bar Chart & Histogram in R (with Example). Retrieved November 4, 2020 from https://www.guru99.com/r-bar-chart-histogram.html

Anon. Correlation Test Between Two Variables in R. Retrieved November 4, 2020 from http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r

Anon. ggplot2 : Quick correlation matrix heatmap - R software and data visualization. Retrieved November 4, 2020 from http://www.sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-software-and-data-visualization

Larxel. 2020. Heart Failure Prediction. (June 2020). Retrieved November 4, 2020 from https://www.kaggle.com/andrewmvd/heart-failure-clinical-data

Anon. Reading and Writing CSV Files. Retrieved November 4, 2020 from https://swcarpentry.github.io/r-novice-inflammation/11-supp-read-write-csv/