Milestone 3

Berkeley D. Willis

2020-09-24

First the data needs to be loaded, and some basic sampling and summarization of the data would be helpful. It can help identify what could be cleaned, as well as give an idea of what data is available, record counts, and even possibly give an idea of what relationship may exist. Of course after the data cleaning operations this will become more obvious.

```r
# Now load the data that we want
hf_records_dt = fread("data/heart_failure_clinical_records_dataset.csv")

hf_records_dt %>% head
```

```
##    age anaemia creatinine_phosphokinase diabetes ejection_fraction
## 1:  75       0                      582        0                20
## 2:  55       0                     7861        0                38
## 3:  65       0                      146        0                20
## 4:  50       1                      111        0                20
## 5:  65       1                      160        1                20
## 6:  90       1                       47        0                40
##    high_blood_pressure platelets serum_creatinine serum_sodium sex smoking time
## 1:                   1    265000              1.9          130   1       0    4
## 2:                   0    263358              1.1          136   1       0    6
## 3:                   0    162000              1.3          129   1       1    7
## 4:                   0    210000              1.9          137   1       0    7
## 5:                   0    327000              2.7          116   0       0    8
## 6:                   1    204000              2.1          132   1       1    8
##    DEATH_EVENT
## 1:           1
## 2:           1
## 3:           1
## 4:           1
## 5:           1
## 6:           1
```

```r
# Let's start to see some of the possible data issues such as large
# numbers of NA's
hf_records_dt %>% summary
```

```
##       age           anaemia       creatinine_phosphokinase    diabetes
##  Min.   :40.00   Min.   :0.0000   Min.   :  23.0           Min.   :0.0000
##  1st Qu.:51.00   1st Qu.:0.0000   1st Qu.: 116.5           1st Qu.:0.0000
##  Median :60.00   Median :0.0000   Median : 250.0           Median :0.0000
##  Mean   :60.83   Mean   :0.4314   Mean   : 581.8           Mean   :0.4181
##  3rd Qu.:70.00   3rd Qu.:1.0000   3rd Qu.: 582.0           3rd Qu.:1.0000
##  Max.   :95.00   Max.   :1.0000   Max.   :7861.0           Max.   :1.0000
##  ejection_fraction high_blood_pressure   platelets      serum_creatinine
##  Min.   :14.00     Min.   :0.0000      Min.   : 25100   Min.   :0.500
##  1st Qu.:30.00     1st Qu.:0.0000      1st Qu.:212500   1st Qu.:0.900
##  Median :38.00     Median :0.0000      Median :262000   Median :1.100
##  Mean   :38.08     Mean   :0.3512      Mean   :263358   Mean   :1.394
##  3rd Qu.:45.00     3rd Qu.:1.0000      3rd Qu.:303500   3rd Qu.:1.400
##  Max.   :80.00     Max.   :1.0000      Max.   :850000   Max.   :9.400
##   serum_sodium        sex            smoking            time
##  Min.   :113.0   Min.   :0.0000   Min.   :0.0000   Min.   :  4.0
##  1st Qu.:134.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 73.0
##  Median :137.0   Median :1.0000   Median :0.0000   Median :115.0
##  Mean   :136.6   Mean   :0.6488   Mean   :0.3211   Mean   :130.3
##  3rd Qu.:140.0   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:203.0
##  Max.   :148.0   Max.   :1.0000   Max.   :1.0000   Max.   :285.0
##   DEATH_EVENT
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3211
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

What can be observed here is that the data is pretty clean with no major issues like missing data. However, some of the data types of the values though not ideal for giving a good view of the data. Some of these could be converted to integers and Boolean in order to make the data more clear on the summary and sampling, like a boolean as an indicator for diabetes isn't wrong to be 0/1 but will look simpler in the summary view if converted to a boolean. As well a few numerics like age, just because how R read the file, would look simpler and vizualize better as a non-continuous value.

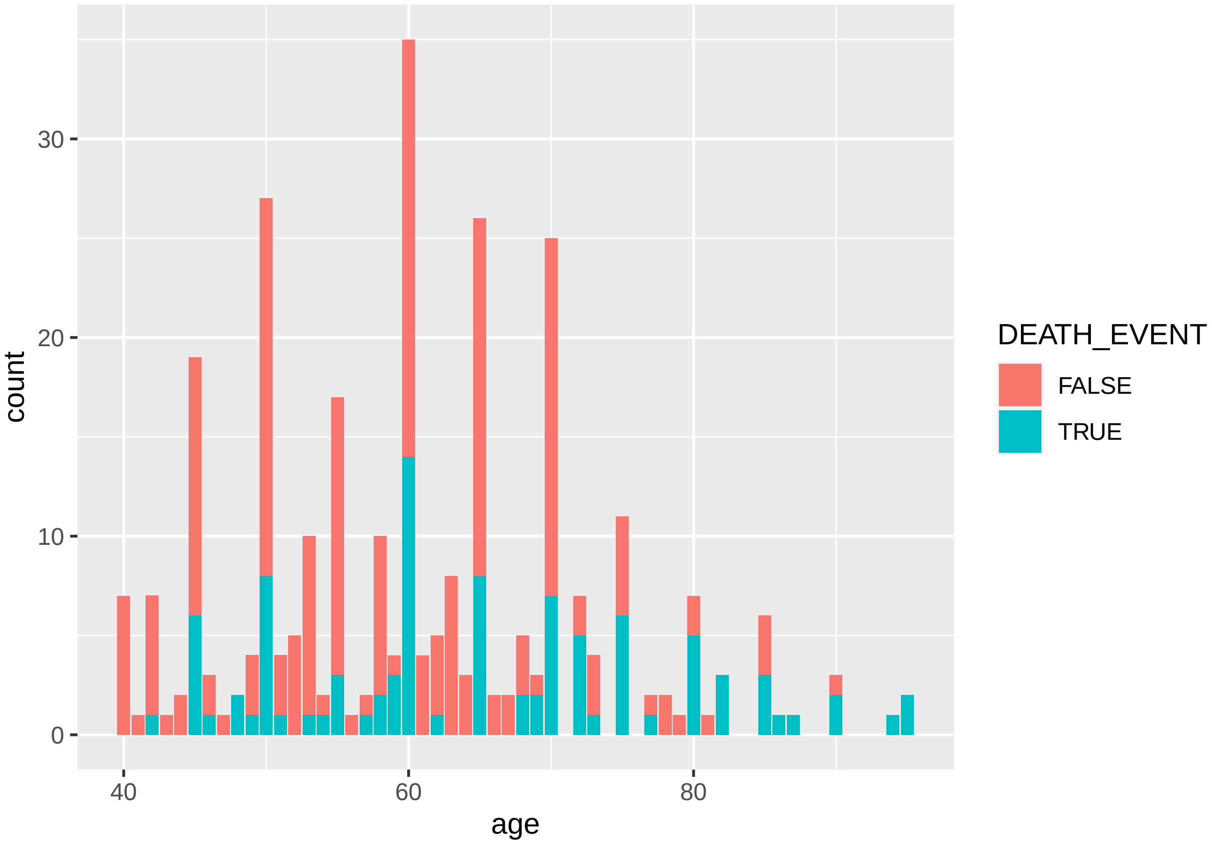
So the next steps will be to make what will be helpful conversions.

```r
# First convert those that are best to be used as booleans
hf_records_dt$anaemia = as.logical(hf_records_dt$anaemia)
hf_records_dt$diabetes = as.logical(hf_records_dt$diabetes)
hf_records_dt$high_blood_pressure = as.logical(hf_records_dt$high_blood_pressure)
hf_records_dt$smoking = as.logical(hf_records_dt$smoking)
hf_records_dt$DEATH_EVENT = as.logical(hf_records_dt$DEATH_EVENT)

# Sex currently represented as a number, won't be consider a classification so
# we'll change it to a factor
hf_records_dt$sex = as.factor(as.character(hf_records_dt$sex))

# Finally lets change numeric to integer for age
hf_records_dt$age = as.integer(hf_records_dt$age)

# Quick check of the data
hf_records_dt %>% summary
```

```
##       age          anaemia        creatinine_phosphokinase  diabetes
##  Min.   :40.00   Mode :logical   Min.   :  23.0           Mode :logical
##  1st Qu.:51.00   FALSE:170       1st Qu.: 116.5           FALSE:174
##  Median :60.00   TRUE :129       Median : 250.0           TRUE :125
##  Mean   :60.83                   Mean   : 581.8
##  3rd Qu.:70.00                   3rd Qu.: 582.0
##  Max.   :95.00                   Max.   :7861.0
##  ejection_fraction high_blood_pressure   platelets      serum_creatinine
##  Min.   :14.00     Mode :logical       Min.   : 25100   Min.   :0.500
##   1st Qu.:30.00     FALSE:194           1st Qu.:212500   1st Qu.:0.900
##   Median :38.00     TRUE :105           Median :262000   Median :1.100
##   Mean   :38.08                         Mean   :263358   Mean   :1.394
##   3rd Qu.:45.00                         3rd Qu.:303500   3rd Qu.:1.400
##   Max.   :80.00                         Max.   :850000   Max.   :9.400
##    serum_sodium   sex      smoking             time        DEATH_EVENT
##   Min.   :113.0   0:105   Mode :logical   Min.   :  4.0   Mode :logical
##   1st Qu.:134.0   1:194   FALSE:203       1st Qu.: 73.0   FALSE:203
##   Median :137.0           TRUE :96        Median :115.0   TRUE :96
##   Mean   :136.6                           Mean   :130.3
##   3rd Qu.:140.0                           3rd Qu.:203.0
##   Max.   :148.0                           Max.   :285.0
```

Though this isn't a lot of complex cleaning operations, the difference can be see in the summarization of the data. Now it is possible to get a preview and see how certain variables may have a relationship with the death indicator. As well those relationships can be explored further in a series of vizualizations.

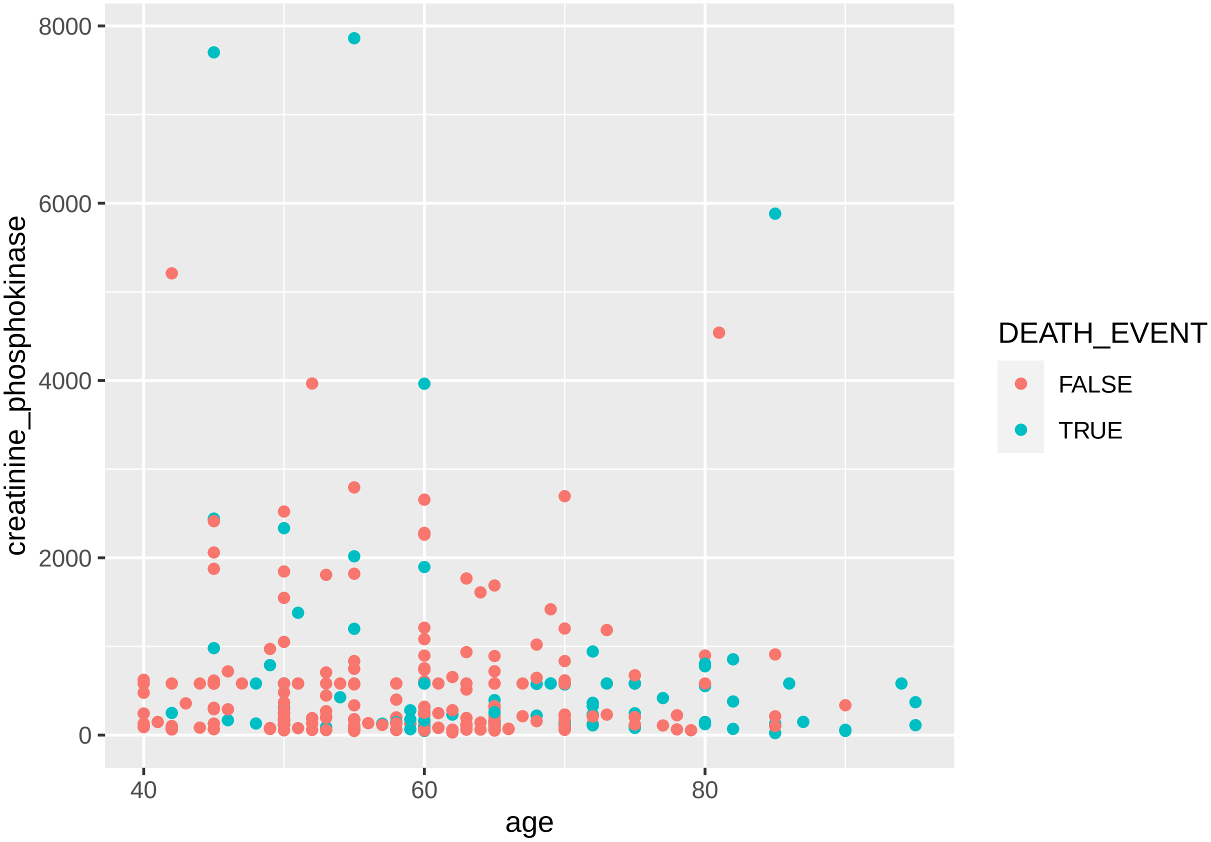
```r
# Take a look at the possibility of a relationship between death an age
hf_records_agg = setNames(
  aggregate(anaemia ~ age + DEATH_EVENT, hf_records_dt, length),
  c("age", "DEATH_EVENT", "count"))
ggplot(hf_records_agg, aes(x=age, y=count, fill=DEATH_EVENT)) +
  geom_bar(position="stack", stat="identity")
```



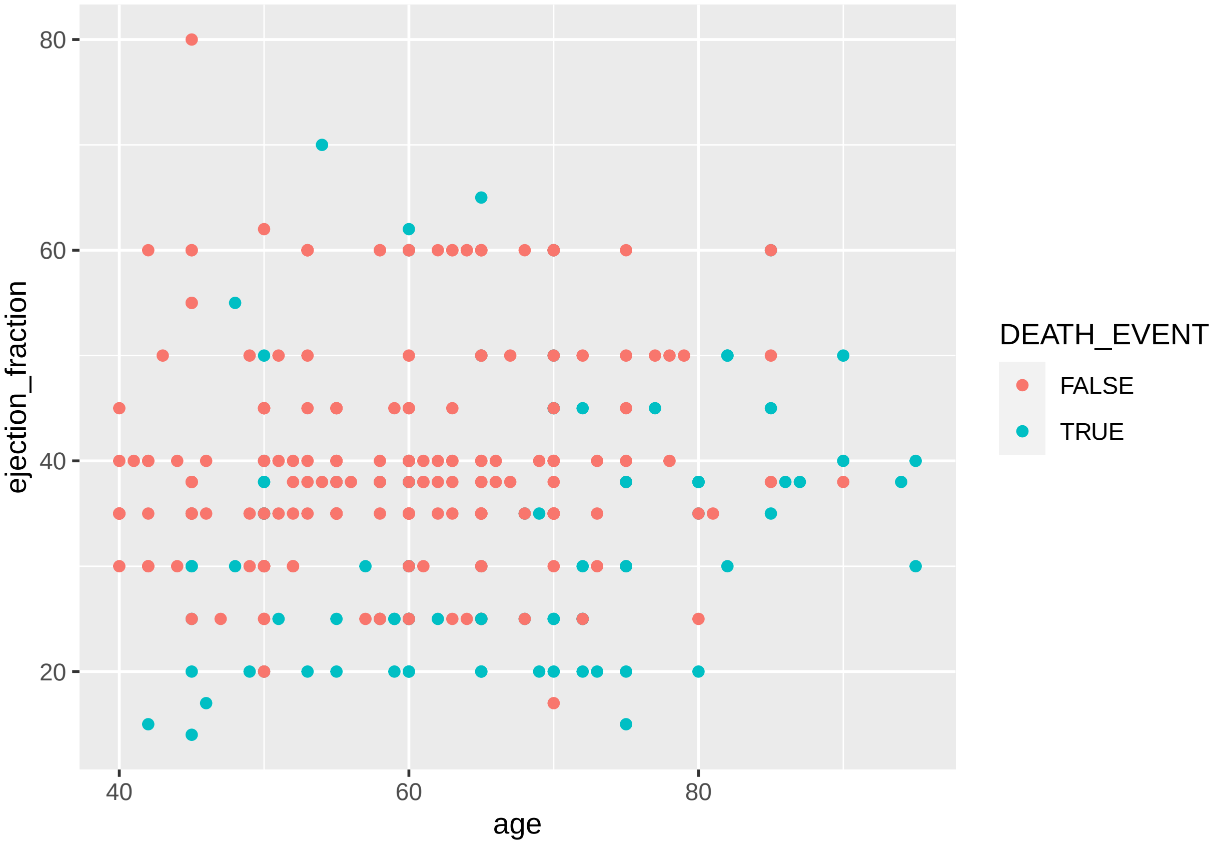
The obvious thing to point out is that of course as people age they are more likely to die, but the question here is whether or not it would be to heart failure despite these treatements. With the results of the visualization, it can be seen that there is likely a relationship between age and the count of deaths. Something of note here is that all of the ages in the records don't commonly have deaths until close to the age of 60.

For other vizualizations it might be useful to utilize scatter plots with age and whatever other variables might be relevant.

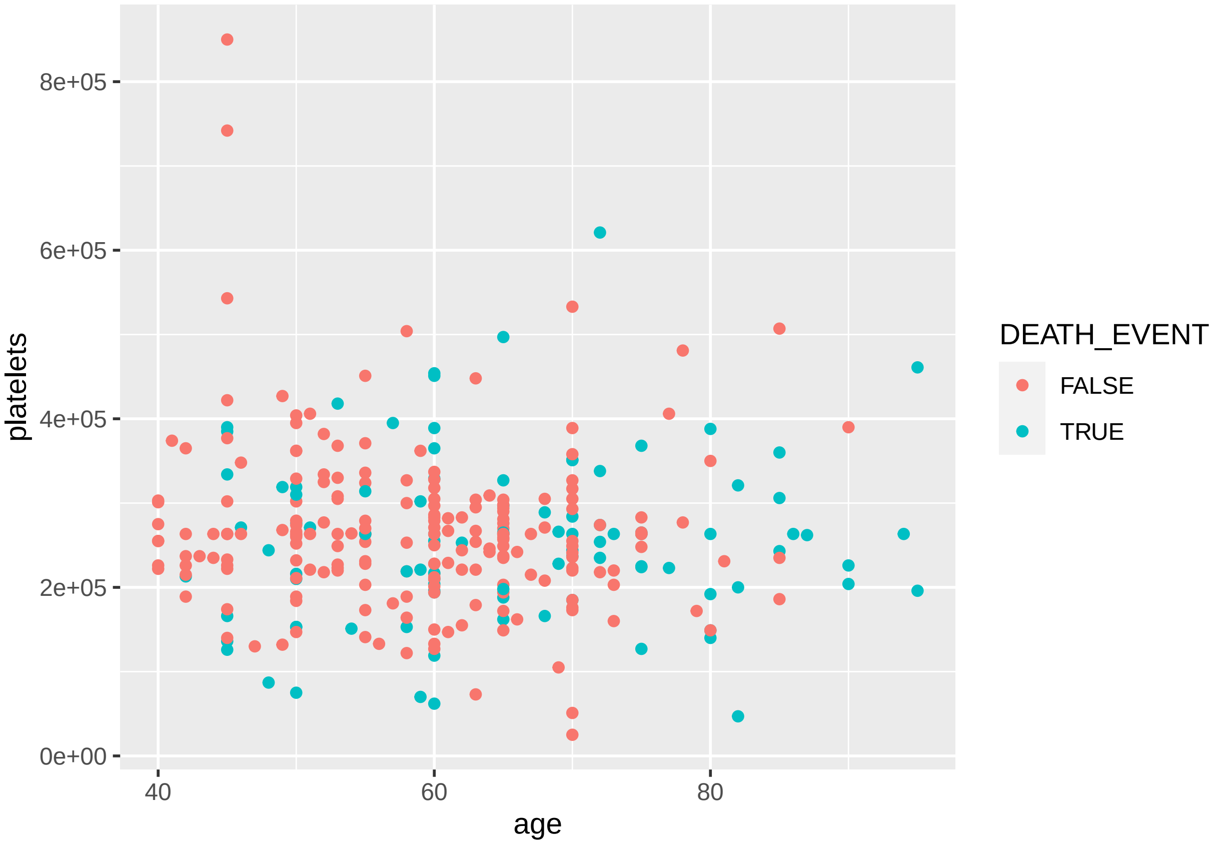
```
# Going to scroll through a few of these numeric values and see if there are
#  any good clusters
ggplot(hf_records_dt, aes(x=age,y=creatinine_phosphokinase, color=DEATH_EVENT))+
  geom_point()
```



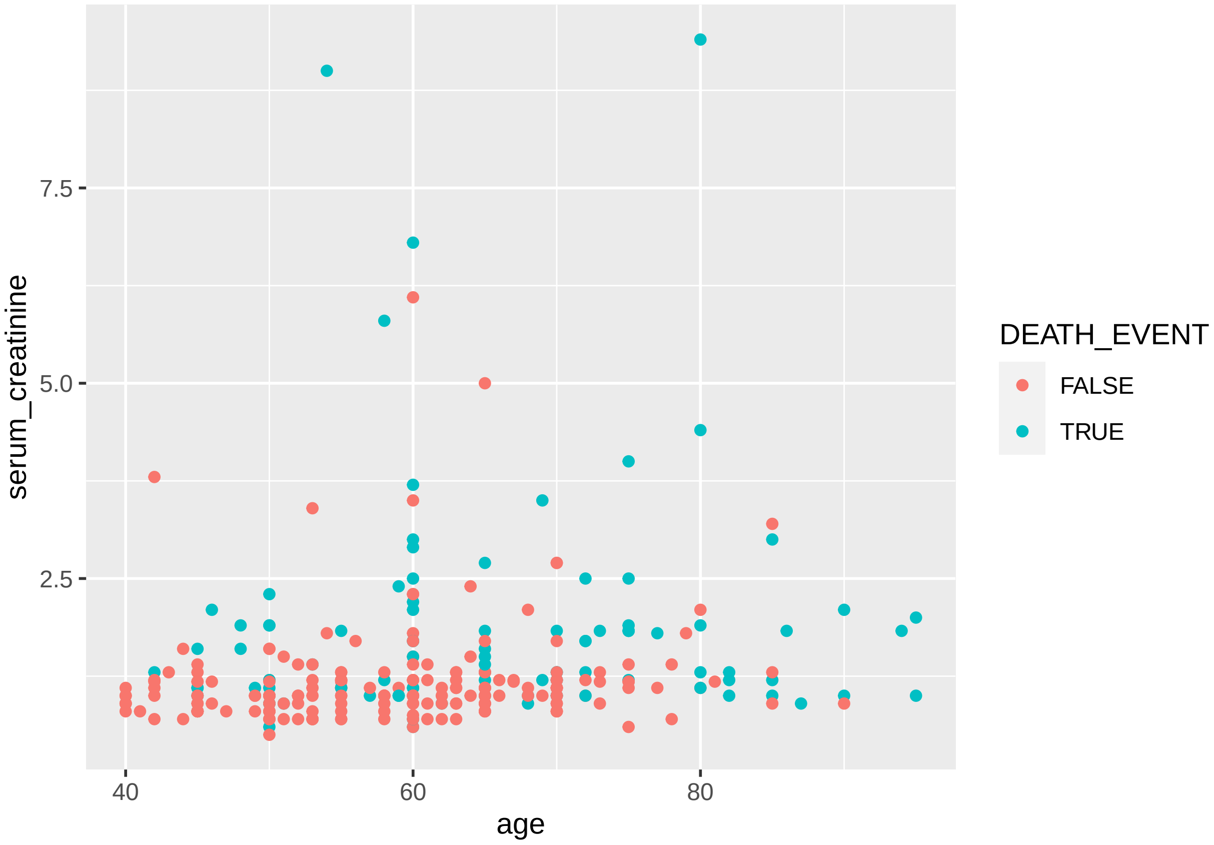
```
ggplot(hf_records_dt, aes(x=age, y=ejection_fraction, color=DEATH_EVENT)) +
  geom_point()
```



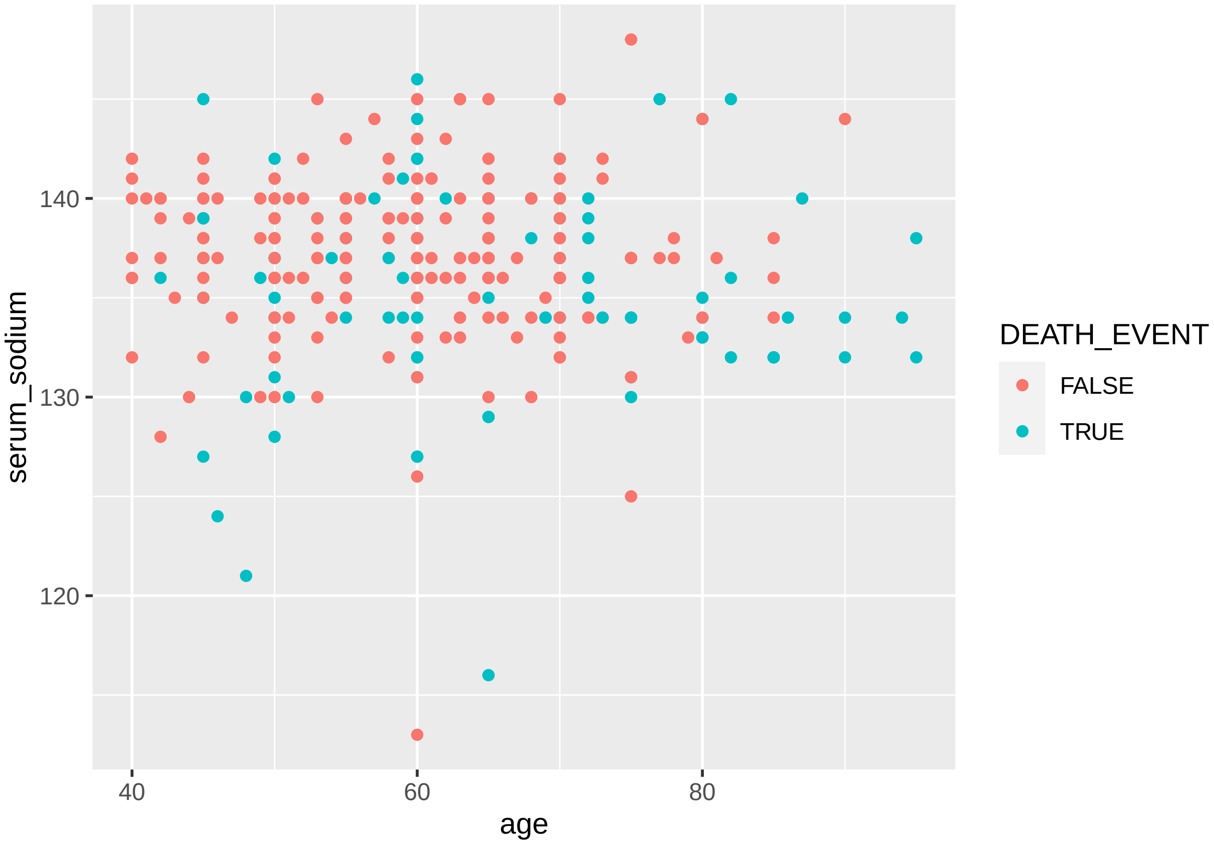
```r
ggplot(hf_records_dt, aes(x=age, y=platelets, color=DEATH_EVENT)) +
  geom_point()
```



```r
ggplot(hf_records_dt, aes(x=age, y=serum_creatinine, color=DEATH_EVENT)) +
  geom_point()
```



```r
ggplot(hf_records_dt, aes(x=age, y=serum_sodium, color=DEATH_EVENT)) +
  geom_point()
```



This same age relation does seem to hold and there are some small indications for clustering with all the variables currently explored. However, it would be nice to make this clearer if it were possible to get more records allowing for denser and clearer clusters.

One of the variables that may not be as helpful is ‘time’ which according to the data source is the number of days before follow-up. All the others are possibly helpful for building model. During the modeling process multiple combinations, with the exception of the ‘time’ variable, will be used and tested. Main concern is that there aren’t a lot of records, with only 299 available. The data will still be split out randomly for test data and validation but these will be fairly small, and there is the concern of overfitting. Having separate validation records should help diagnose if this does become an issue though.

Next milestone will include further exploration as models are created in oroder to find the most effective method of prediction.