ICU

Marc Palomer

2022-11-14

R Markdown

This exercise contains one dataset:

ICU dataset. This dataset contains clinical records for 32536 subjects. Database records contain results of laboratory tests, medications, ICD9 diagnoses, admitting notes, discharge summaries, and more. Each record contains data for a single subject, and many records span multiple ICU admissions for the same subject, including available medical history between ICU stays.

Perform the following actions:

Task 1:

Convert your main variables to its correct format (factors, numeric, etc).

```
library(tidyverse)
library(dplyr)
ICU <- read_delim("ICU.csv", delim=",", show_col_types = FALSE)

a = c()

#Here we detect the variables with less than 4 unique values, meaning that probably these are factors

for (element in names(ICU)){
  if (count(unique(head(ICU[element])))<=3){
    a = c(a, element)
  }
}

#Preselection of variables of interest
variables_interest <- c("SUBJECT_ID","ICUSTAY_ID","ICUSTAY_ADMIT_AGE","BMI_ADMIT","ICUSTAY_ADMIT_SAPS",'

#Factorisation
ICU_factorised<-ICU %>%
  mutate_at(vars(a), list(factor))%>%
  select(all_of(variables_interest))

#Overview of variables inside ICU_factorised
str(ICU_factorised)
```

```
## tibble [15,690 x 34] (S3: tbl_df/tbl/data.frame)
##  $ SUBJECT_ID              : num [1:15690] 3 9 12 13 17 20 21 23 25 26 ...
##  $ ICUSTAY_ID              : num [1:15690] 4 10 13 14 18 22 23 26 28 29 ...
##  $ ICUSTAY_ADMIT_AGE       : num [1:15690] 76.5 41.8 72.4 39.9 47.5 75.9 87.4 71.1 59 72 ...
##  $ BMI_ADMIT               : num [1:15690] 30.2 31.1 29.9 35.1 24.5 ...
```

```
##  $ ICUSTAY_ADMIT_SAPS      : num [1:15690] 28 16 15 14 15 20 17 14 16 6 ...
##  $ ICUSTAY_ADMIT_SOFA      : num [1:15690] 14 9 9 6 7 8 8 10 5 2 ...
##  $ NUM_SIRS_COND           : Factor w/ 5 levels "0","1","2","3",..: 4 3 2 4 4 3 3 3 4 2 ...
##  $ HR_ADMIT                : num [1:15690] 95 85 86 80 93 80 84 90 72 61 ...
##  $ HR_MAX                  : num [1:15690] 168 111 105 102 98 80 84 100 104 65 ...
##  $ TEMP_MAX                : num [1:15690] 99.3 100.2 99.8 99.1 98.6 ...
##  $ WBC_MAX                 : num [1:15690] 24.4 13.7 8.4 19.3 24 17.5 23.5 9.4 13 8.2 ...
##  $ HR_MIN                  : num [1:15690] 75 82 71 60 74 67 60 85 49 60 ...
##  $ TEMP_MIN                : num [1:15690] 96.9 95.9 95.9 96.7 96.8 96.6 95.7 95 96.3 99 ...
##  $ WBC_MIN                 : num [1:15690] 11.3 13.7 7.8 16.6 10.5 17.5 21.1 7.6 11.6 8.2 ...
##  $ CHRONIC_PULMONARY       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ DIABETES_UNCOMPLICATED  : Factor w/ 2 levels "0","1": 1 1 1 2 1 2 2 1 2 1 ...
##  $ LIVER_DISEASE           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ AIDS                    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ALCOHOL_ABUSE           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ DRUG_ABUSE              : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ oneyear_num             : Factor w/ 2 levels "1","2": 2 2 2 1 1 1 2 1 1 1 ...
##  $ hospital_mort_num       : Factor w/ 2 levels "1","2": 1 2 2 1 1 1 1 1 1 1 ...
##  $ gender_num              : Factor w/ 2 levels "1","2": 2 2 2 1 1 1 2 2 2 2 ...
##  $ ICUstay_num             : num [1:15690] 4 5 5 1 2 2 1 2 1 1 ...
##  $ OBESITY                 : Factor w/ 2 levels "0","1": 2 2 1 2 1 2 1 1 1 2 ...
##  $ avghr                   : num [1:15690] 92 86 88 98 83 79 77 90 76 71 ...
##  $ stddevhr                : num [1:15690] 20 6 9 15 11 2 19 7 10 8 ...
##  $ numoverload             : Factor w/ 16 levels "0","1","2","3",..: 1 3 1 1 2 2 1 2 3 1 ...
##  $ maxcumvol               : num [1:15690] 17118 8413 19998 2474 6185 ...
##  $ mincumvol               : num [1:15690] 9246 3144 8490 -610 2220 ...
##  $ Creatmax                : num [1:15690] 2.5 2 1.7 0.8 0.8 0.8 4.6 0.7 1.6 1.4 ...
##  $ Bicarbmin               : num [1:15690] 11 21 10 20 22 19 15 22 21 24 ...
##  $ Bcmaxratio              : num [1:15690] 17.2 16.5 24.1 22.5 13.8 ...
##  $ CONGESTIVE_HEART_FAILURE: Factor w/ 2 levels "0","1": 2 2 1 1 1 1 1 1 1 2 ...
```

There are 73 variables, and some of theme were converted to factors. IDs have been considered to be optimal to be in number type. (Later on xlsx archive with variable types was found and coherence was checked).

The table below shows the number of patients not duplicated (False) and the number of patients duplicated (True)

```
table(duplicated(ICU_factorised$SUBJECT_ID))
```

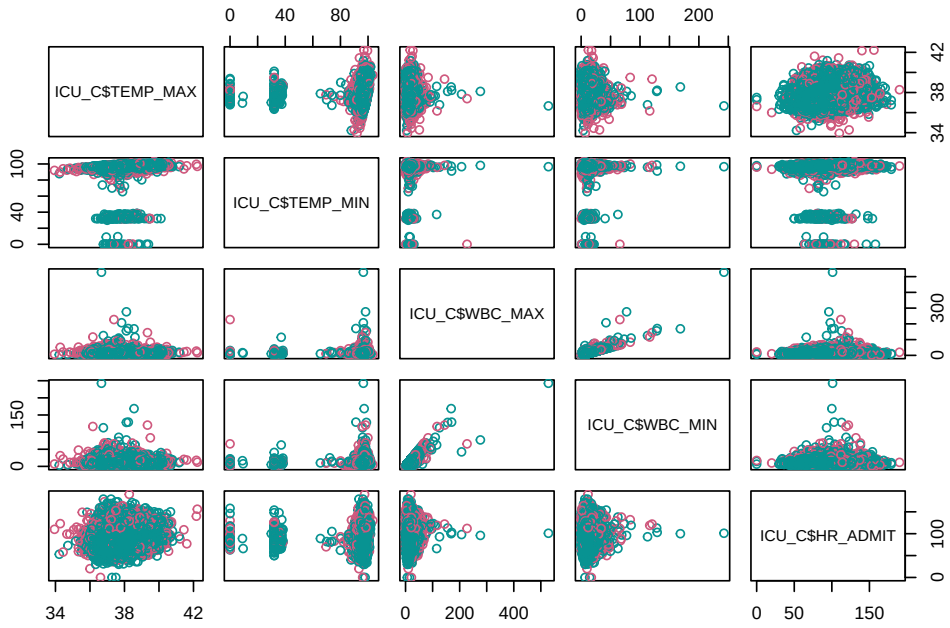
```
##
## FALSE
## 15690
```

There are no duplicated patients.

Task 2

Think and generate a number of plots so that you learn something from the data using standard plotting and ggplot2 package

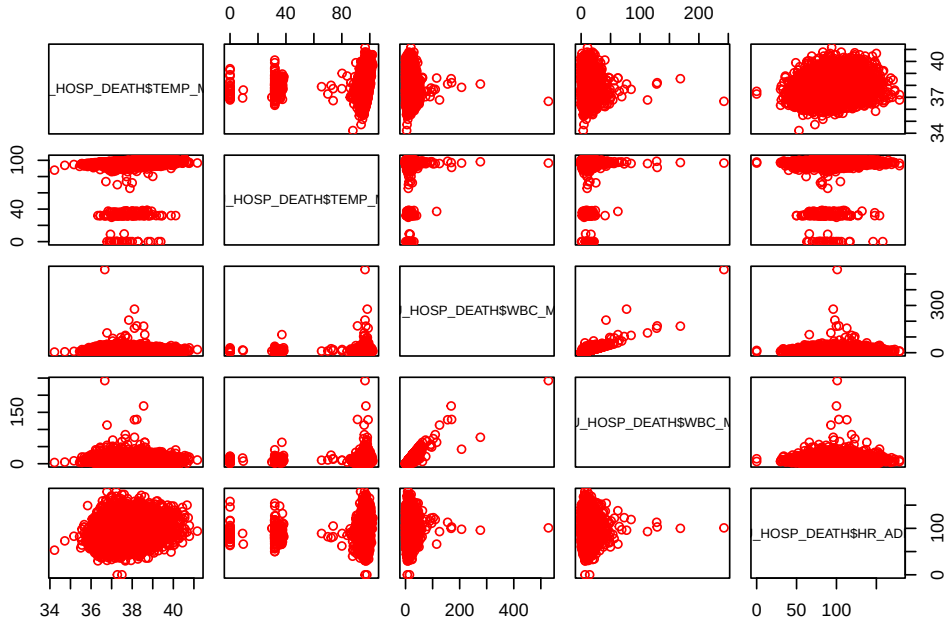
Glimpse of the data:



In this table of scatter plots we can have a glimpse of how the main five numerical variables of the dataset relate to each other. We can extract some simple information from this graphics in order to decide what to further explore in the plots to come

```
ICU_HOSP_DEATH <- ICU_C%>%
  filter(ICU_C$hospital_mort_num==1)
pairs(~ICU_HOSP_DEATH$TEMP_MAX + ICU_HOSP_DEATH$TEMP_MIN + ICU_HOSP_DEATH$WBC_MAX + ICU_HOSP_DEATH$WBC_
```

Glimpse but for in hospital deaths:

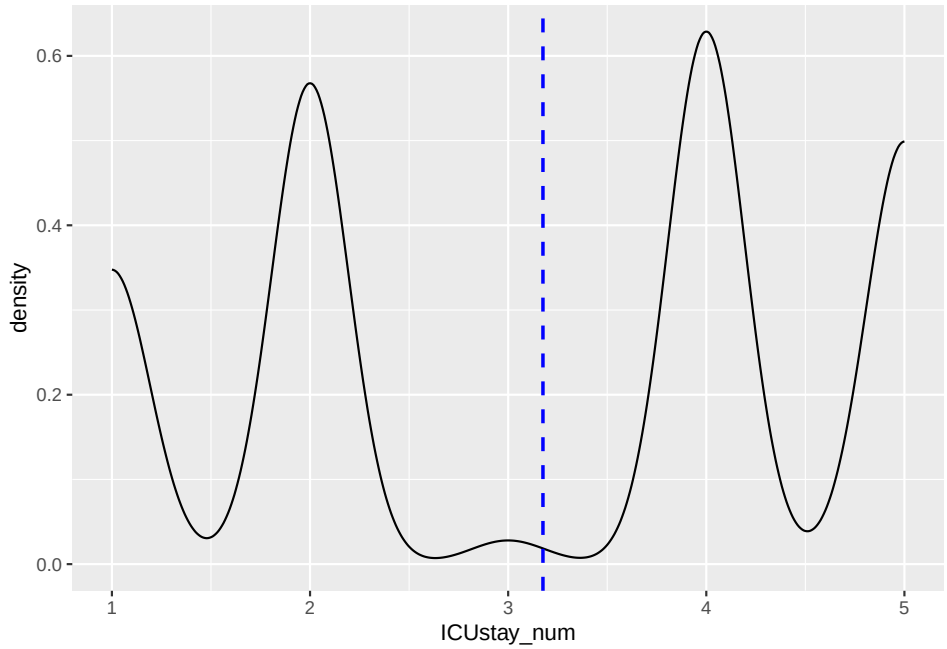


After seeing the two pairs plots, it is not visible at first sight that patients that suffered inhospital death (in red) and those that didn't (in black) have different distributions respect to the plotted variables. On the other hand it is visible that the minimal temperature variable is distributed in three agrupations which make no sense (-20, 0 and ~35 degrees Celcius). Apart from this I can only see that, maybe, patients with

extreme maximum temperature die more.

```
library(ggplot2)
p <- ggplot(ICU_C, aes(x=ICUstay_num)) +
  geom_density()
p+ geom_vline(aes(xintercept=mean(ICUstay_num)),
            color="blue", linetype="dashed", linewidth=0.8)
```

Distribution of days at the UCI:

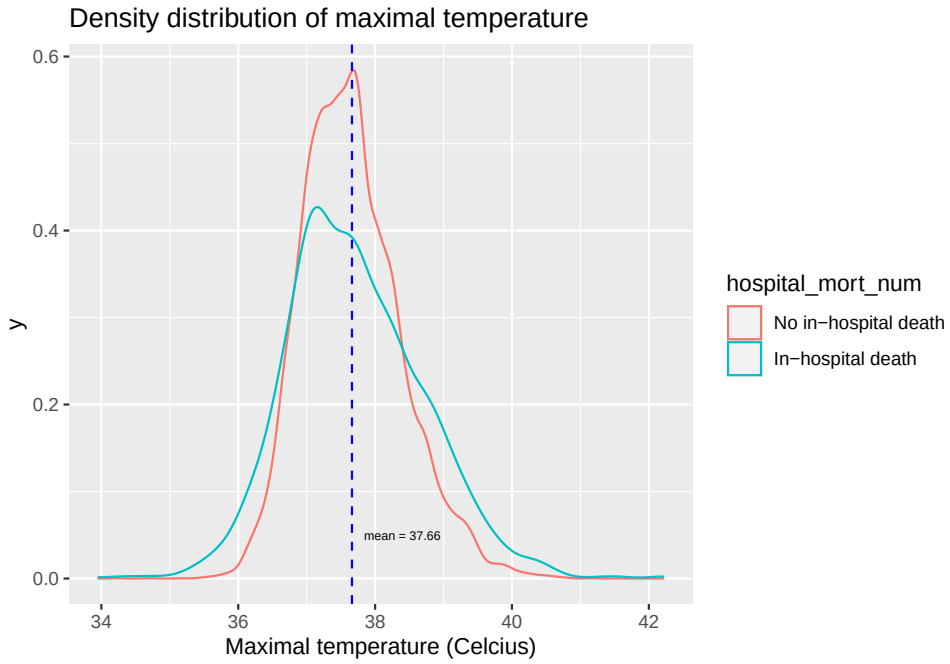


The prob. distribution function of patients as a function of the days they stay in the ICU and the mean stay by patient.

```
#We visualize the distribution of minimal temperatures to acknowledge what was previously commented abo
p_tmin <-ggplot(ICU_C, aes(x=TEMP_MIN)) +
  geom_density()+
  geom_vline(aes(xintercept=mean(TEMP_MIN)),
            color="blue", linetype="dashed", size=0.5)

#Relebelling of inhospital mortality factor
levels(ICU_C$hospital_mort_num)<- c("No in-hospital death","In-hospital death" )
levels(ICU_C$CONGESTIVE_HEART_FAILURE)<- c("No","Yes" )
levels(ICU_C$AIDS)<- c("No","Yes" )
levels(ICU_C$OBESITY)<- c("No","Yes" )

#Density function of maximal temperature
p_tmax <- ggplot(ICU_C, aes(x=TEMP_MAX, col = hospital_mort_num)) +
  geom_density()+
  geom_vline(aes(xintercept=mean(TEMP_MAX)),
            color="blue", linetype="dashed", size=0.5)+
  labs(x = "Maximal temperature (Celcius)", title ="Density distribution of maximal temperature")+
  annotate(geom = "text", x = 38.4 , y = 0.05, label = "mean = 37.66", size = 2)
```

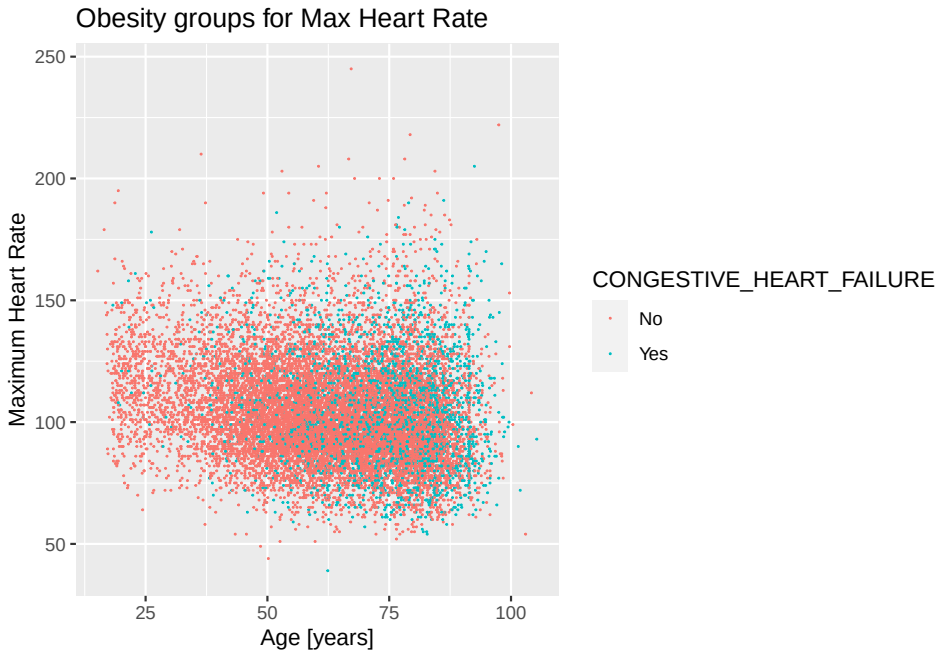
```
p_tmax
```

Distribution of maximal temperatures as there seems to be a relationship with in-hospital death.



Distribution of congestive heart failure in a plot of maximum heart rate versus the age of the patients.

```
ggplot(ICU_C, aes(x = ICUSTAY_ADMIT_AGE,y = HR_MAX, col = CONGESTIVE_HEART_FAILURE))+
 geom_point(size = 0.01)+
  labs(x = "Age [years]", y = "Maximum Heart Rate", legend = "Congestive heart failure")+
  labs(title = "Obesity groups for Max Heart Rate")
```



```
  facet_wrap(ICU_C$OBESITY, )
```

```
## <ggproto object: Class FacetWrap, Facet, gg>
##     compute_layout: function
##     draw_back: function
##     draw_front: function
##     draw_labels: function
##     draw_panels: function
##     finish_data: function
##     init_scales: function
##     map_data: function
##     params: list
##     setup_data: function
##     setup_params: function
##     shrink: TRUE
##     train_scales: function
##     vars: function
##     super:  <ggproto object: Class FacetWrap, Facet, gg>
```

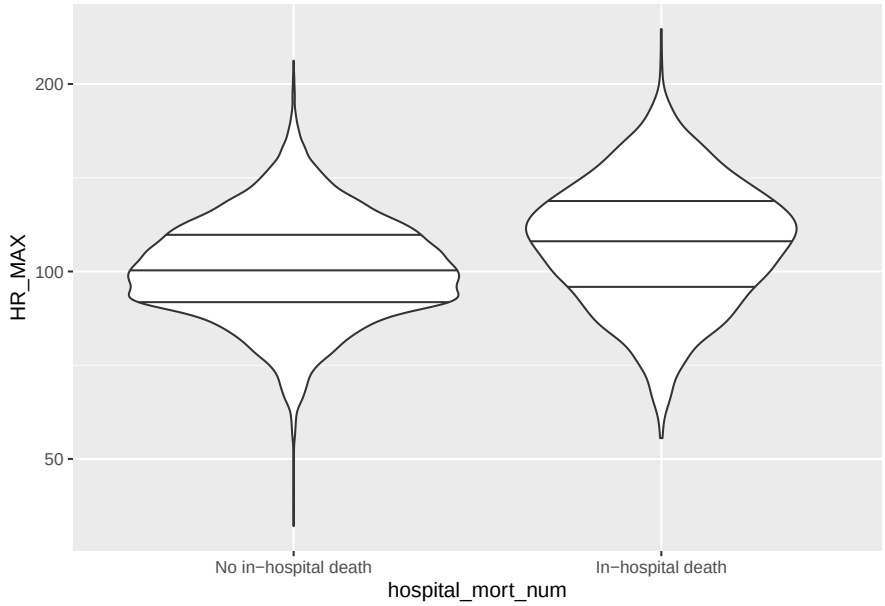
Congestive heart failure does not seem to be related to the maximum HR over the ICU stay, but of course it increases in frequency as patients grow old. Also, young obesity might be related to heart failure

Task 3

Focus on plotting survival data values against the variables that you suspect might be relevant. Check the relationship between obesity and survival.

```
ggplot(ICU_C, aes(x = hospital_mort_num, y= HR_MAX)) +
  geom_violin(draw_quantiles = c(0.25, 0.5, 0.75))+
  scale_y_log10()
```

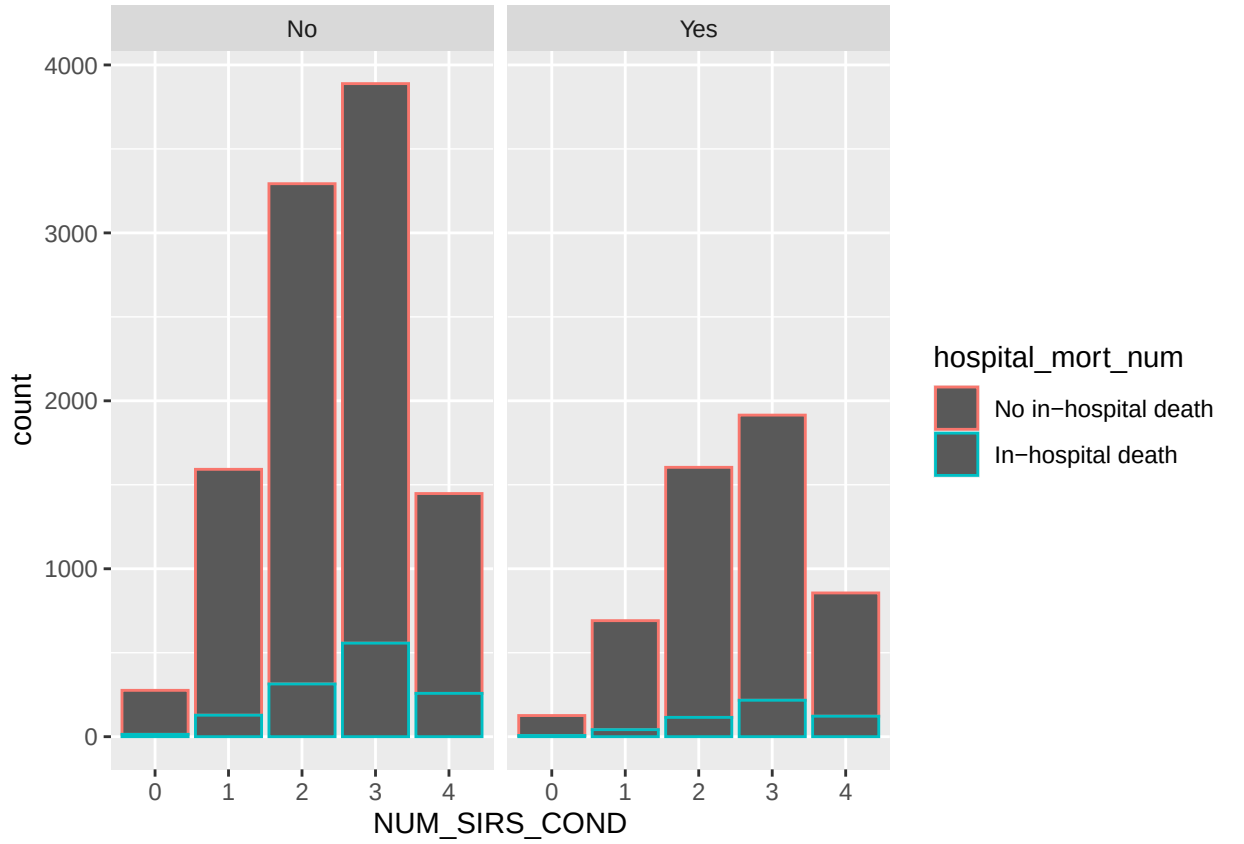
Maximum heart rate distribution compared for in-hospital deaths and those that did not die.



SIRS criteria is a discontinuous score that goes from 0 to 5 defined depending on how many of this conditions

are met: tachycardia (heart rate >90 beats/min), tachypnea (respiratory rate >20 breaths/min), fever or hypothermia (temperature >38 or <36 °C), and leukocytosis, leukopenia, or bandemia (white blood cells >1,200/mm^3, <4,000/mm^3 or bandemia 10%). We are going to study the relationship of this score with in-hospital death.

```
ggplot(ICU_C, aes(x=NUM_SIRS_COND, col = hospital_mort_num, xlabel="SIRS NUMBER")) +
  geom_bar()+
  facet_wrap(ICU_C$OBESITY)
```




```
  labs(x ="SIRS NUMBER")
```

```
## $x
## [1] "SIRS NUMBER"
##
## attr(,"class")
## [1] "labels"
```

```
probability <- prop.table(table(ICU_C$NUM_SIRS_COND, ICU_C$hospital_mort_num), margin = 1)

probability
```

```
##
##     No in-hospital death In-hospital death
##   0           0.94776119        0.05223881
##   1           0.92553657        0.07446343
##   2           0.91239534        0.08760466
##   3           0.86664369        0.13335631
##   4           0.83506944        0.16493056
```

When a SIRS number is given to a patient when entering the UCI it has this given probability of in hospital death as shown below:

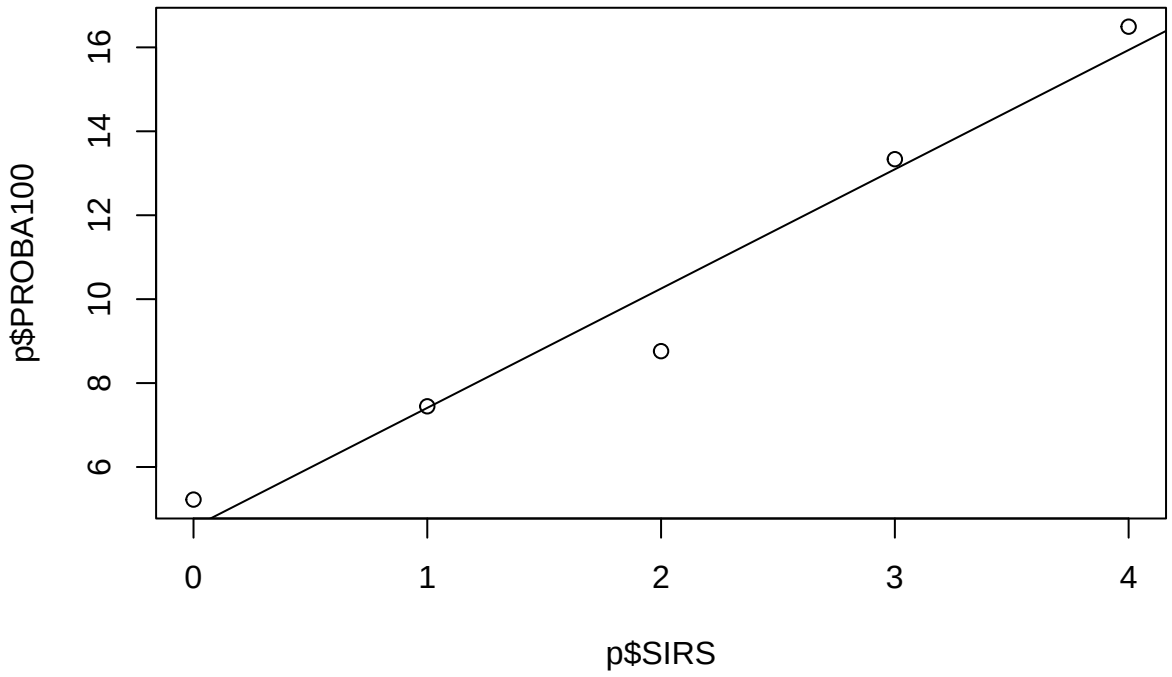
```
p <- data.frame(100*probability[,2],c(0,1,2,3,4))
names(p)<- c("PROBA100", "SIRS")

m<-lm(p$PROBA100~ p$SIRS, data = p)
coef(m)
```

```
## (Intercept)      p$SIRS
##    4.566347    2.842764
```

```
summary(m)
```

```
##
## Call:
## lm(formula = p$PROBA100 ~ p$SIRS, data = p)
##
## Residuals:
##        0        1        2        3        4
##  0.65753  0.03723 -1.49141  0.24099  0.55565
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.5663     0.7778   5.871  0.00986 **
## p$SIRS        2.8428     0.3175   8.953  0.00294 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.004 on 3 degrees of freedom
## Multiple R-squared:  0.9639, Adjusted R-squared:  0.9519
## F-statistic: 80.15 on 1 and 3 DF,  p-value: 0.002941
```

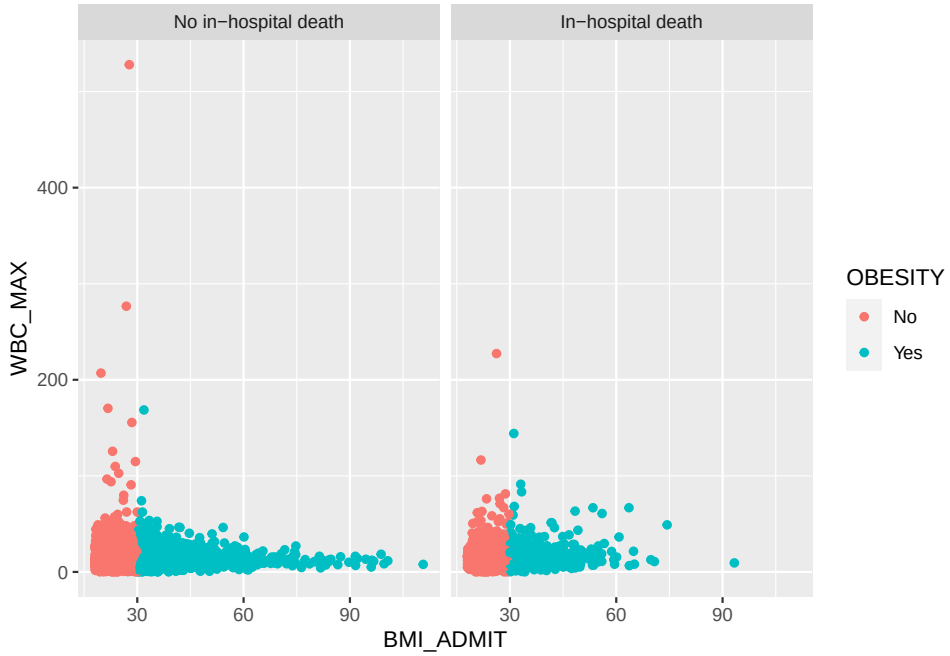
```
plot(p$PROBA100~p$SIRS, )
abline(m)
```



Relationship between obesity and survival

```
ggplot(ICU_C, aes(x = BMI_ADMIT,y=WBC_MAX, col = OBESITY))+
  geom_point()+
  facet_wrap(ICU_C$hospital_mort_num)
```

Dot plot of infection biomarkers as a function of BMI of admission.



As we can see in both the dot plot and the violin plot, there is no direct relationship between obesity (BMI>30) and in-hospital death. Actually the mean BMI of survivals is actually higher than that of non survivors, meaning that an increased BMI could be a benefit when surviving in the UCI.