KaplanMeier_HeartFailure

Antonio Pano

11/4/2022

https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+recorc (https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+recor

- All 299 patients had left ventricular systolic dysfunction

Initial Variables:

- age: age of the patient (years)
- anaemia: decrease of red blood cells or hemoglobin since last measure (boolean)
- high blood pressure: if the patient has hypertension (boolean)
- creatinine phosphokinase (CPK): level of the CPK enzyme in the blood (mcg/L)
- diabetes: if the patient has diabetes (boolean)
- ejection fraction: percentage of blood leaving the heart at each contraction (percentage)
- platelets: platelets in the blood (kiloplatelets/mL)
- sex: woman or man (binary)
- serum creatinine: level of serum creatinine in the blood (mg/dL)
- serum sodium: level of serum sodium in the blood (mEq/L)
- smoking: if the patient smokes or not (boolean)
- time: follow-up period (days)
- [target] death event: if the patient deceased during the follow-up period (boolean)

```
library(skimr)
```

```
## Warning: package 'skimr' was built under R version 4.2.2
```

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(survival)
library(survminer)
```

```
## Warning: package 'survminer' was built under R version 4.2.2
```

```
## Loading required package: ggpubr
```

```
## Warning: package 'ggpubr' was built under R version 4.2.2
```

```
##
## Attaching package: 'survminer'
```

```
## The following object is masked from 'package:survival':
##
##     myeloma
```

```
library(partykit)
```

```
## Warning: package 'partykit' was built under R version 4.2.2
```

```
## Loading required package: grid
```

```
## Loading required package: libcoin
```

```
## Warning: package 'libcoin' was built under R version 4.2.2
```

```
## Loading required package: mvtnorm
```

```
library(coin)
```

```
## Warning: package 'coin' was built under R version 4.2.2
```

```
library(survminer)
library(flexsurv)
```

```
## Warning: package 'flexsurv' was built under R version 4.2.2
```

Loading in the data

Creating Left Ventricular Ejection Fraction Groups set by Cardiology Experts (https://www.ncbi.nlm.nih.gov/books/NBK459131/). Rounding for averages instead of only using data for men and women.

```
HF <- read.csv("heart_failure_clinical_records_dataset.csv")

HF$anaemia = as.factor(HF$anaemia)
HF$diabetes = factor(HF$diabetes,levels=c(0,1),labels=c("Absent","Present"))
HF$hypertension = factor(HF$high_blood_pressure,levels=c(0,1),labels=c("Absent","Present"))

HF$sex = factor(HF$sex,levels=c(0,1),labels=c("Female","Male"))
HF$smoking = as.factor(HF$smoking)
HF$DEATH_EVENT = as.factor(HF$DEATH_EVENT)

HF <- HF %>%
  mutate(EF_Condition = cut(HF$ejection_fraction, breaks = c(0, 30, 40, 52, Inf),
          labels = c("Severe", "Moderate", "Mild", "Normal"), include.lowest = TRUE))


HF <- select(HF, -high_blood_pressure)

skim(HF)
```

Data summary

Name	HF
Number of rows	299
Number of columns	14
_____	
Column type frequency:	
factor	7
numeric	7
_____	
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
anaemia	0	1	FALSE	2	0: 170, 1: 129
diabetes	0	1	FALSE	2	Abs: 174, Pre: 125
sex	0	1	FALSE	2	Mal: 194, Fem: 105
smoking	0	1	FALSE	2	0: 203, 1: 96
DEATH_EVENT	0	1	FALSE	2	0: 203, 1: 96
hypertension	0	1	FALSE	2	Abs: 194, Pre: 105
EF_Condition	0	1	FALSE	4	Mod: 126, Sev: 93, Mil: 41, Nor: 39

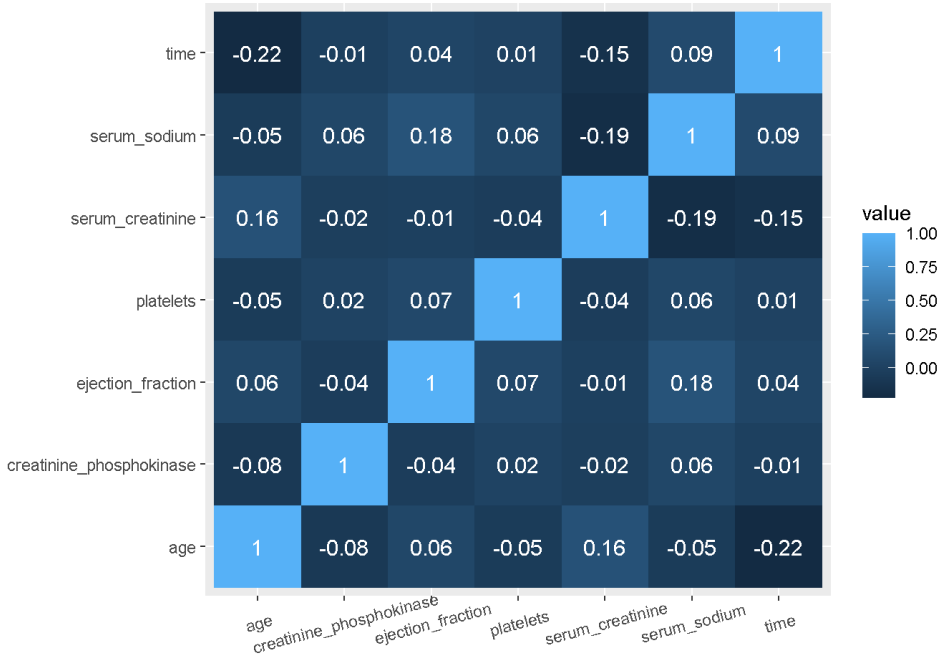
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	60.83	11.89	40.0	51.0	60.0	70.0	95.0	
creatinine_phosphokinase	0	1	581.84	970.29	23.0	116.5	250.0	582.0	7861.0	
ejection_fraction	0	1	38.08	11.83	14.0	30.0	38.0	45.0	80.0	
platelets	0	1	263358.03	97804.24	25100.0	212500.0	262000.0	303500.0	850000.0	
serum_creatinine	0	1	1.39	1.03	0.5	0.9	1.1	1.4	9.4	
serum_sodium	0	1	136.63	4.41	113.0	134.0	137.0	140.0	148.0	
time	0	1	130.26	77.61	4.0	73.0	115.0	203.0	285.0	

Correlation

Time and Serum_Creatinine have a correlation to Serum_Sodium of 0.15 & 0.19, respectively.

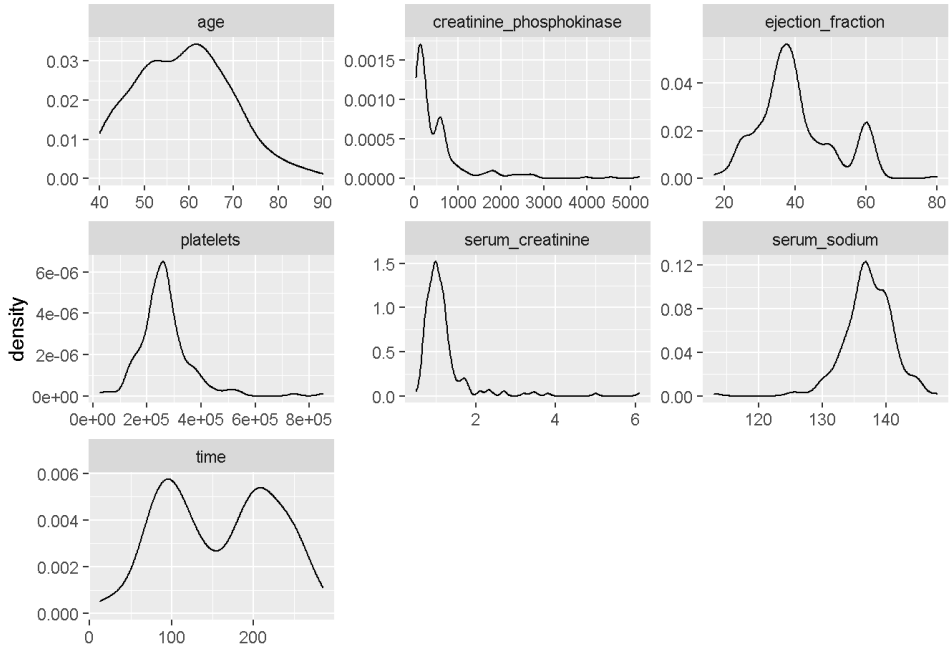
```
cormat <- HF %>% select(where(is.numeric)) %>% cor() %>% round(2)
melted_cormat <- reshape2::melt(cormat)

ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  geom_text(aes(Var2, Var1, label = value), color = "white", size = 4) +
   theme(axis.title.x=element_blank(),
        axis.title.y=element_blank(),
        axis.text.x = element_text(angle = 15, vjust = 0.8)
        )
```



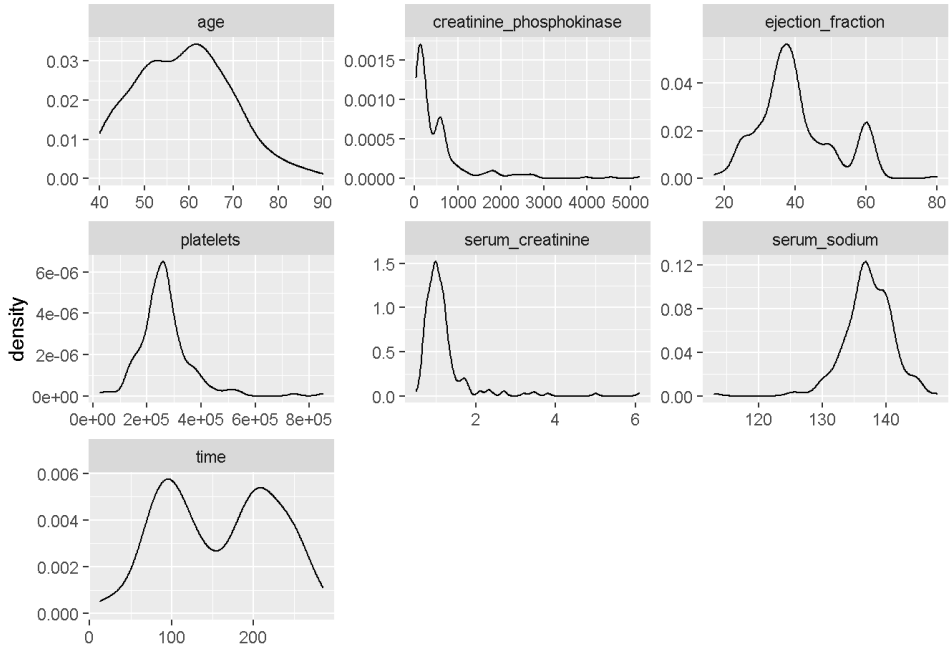
Choosing to grab distributions based on having hypertension– what's traditionally seen as a good indicator of heart failure.

Doing so to look at, specifically, Ejection Fraction right after to see if there is correlation.

```
HF %>% filter(DEATH_EVENT==0) %>%
  purrr::keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_density()
```



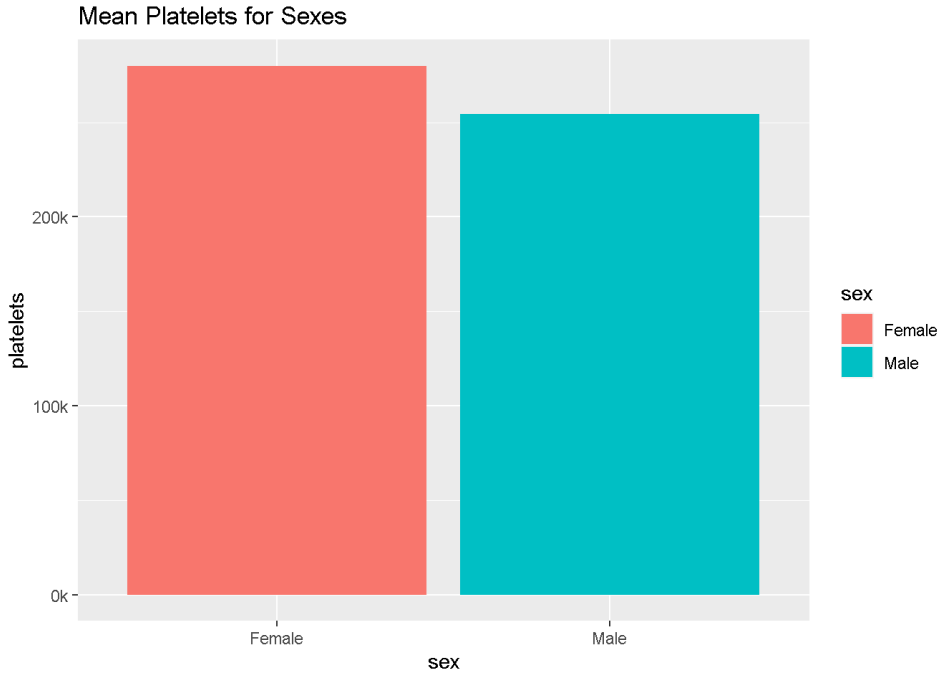

```
HF %>% filter(DEATH_EVENT==0) %>%
  purrr::keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_density()
```



Comparing `creatinine_phosphokinase` to Men & Women– those who smoke and those who do not.

- Noticing that the average `creatinine_phosphokinase` is higher for non-smokers.

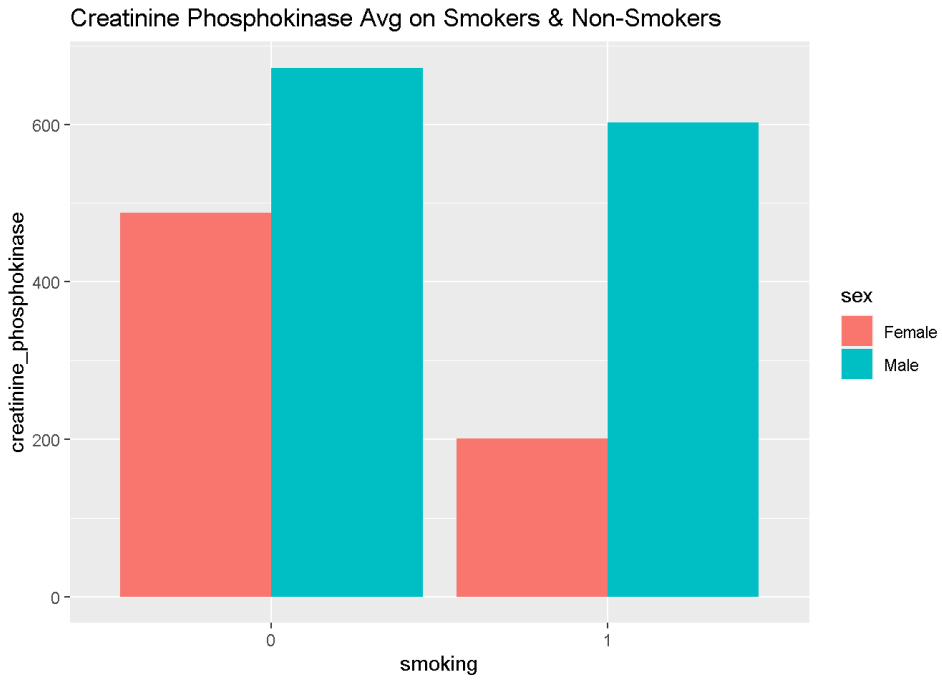
```
ggplot(HF, aes(x=sex, y=platelets, fill=sex)) +
  geom_bar(position = "dodge", stat="summary", fun="mean") +
  scale_y_continuous(labels = scales::label_number(suffix = "k", scale = 1e-3)) +
  ggtitle("Mean Platelets for Sexes")
```




```
HF %>% group_by(sex, DEATH_EVENT) %>%
  summarize(count = n(), .groups="drop")
```

```
## # A tibble: 4 × 3
##   sex    DEATH_EVENT count
##   <fct>  <fct>       <int>
## 1 Female 0              71
## 2 Female 1              34
## 3 Male   0             132
## 4 Male   1              62
```

```
ggplot(HF, aes(x=smoking, y=creatinine_phosphokinase, fill=sex)) +
  geom_bar(position = "dodge", stat="summary", fun="mean") +
  ggtitle("Creatinine Phosphokinase Avg on Smokers & Non-Smokers")
```

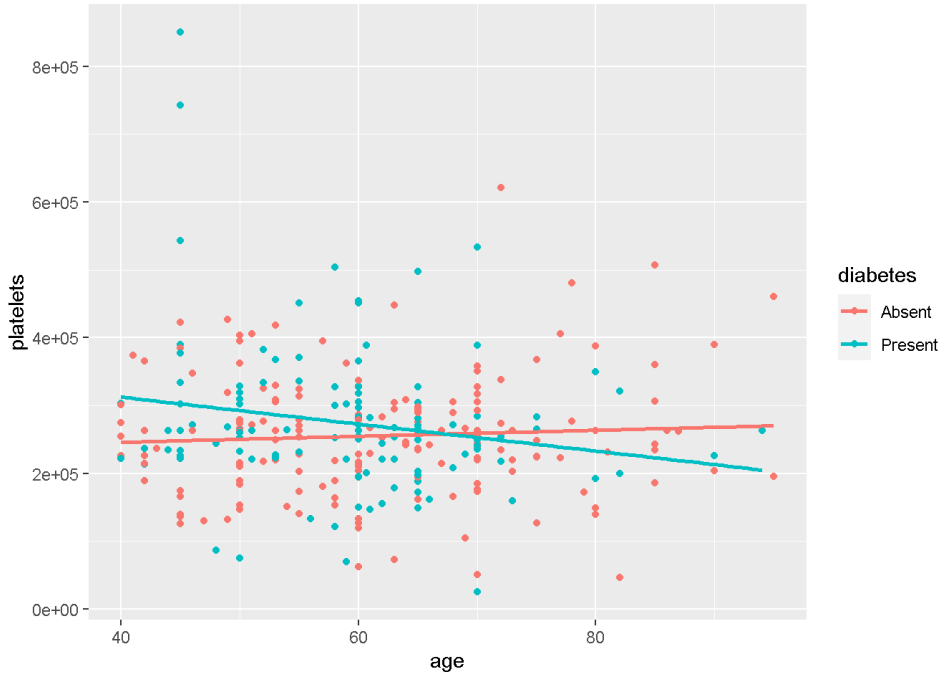


- Finding out that for those diabetic, plateletes are ensured to reduce as you age.
- For those who aren't diabetic, plateletes generally stay the same and potentially, increase by a marginal amount.

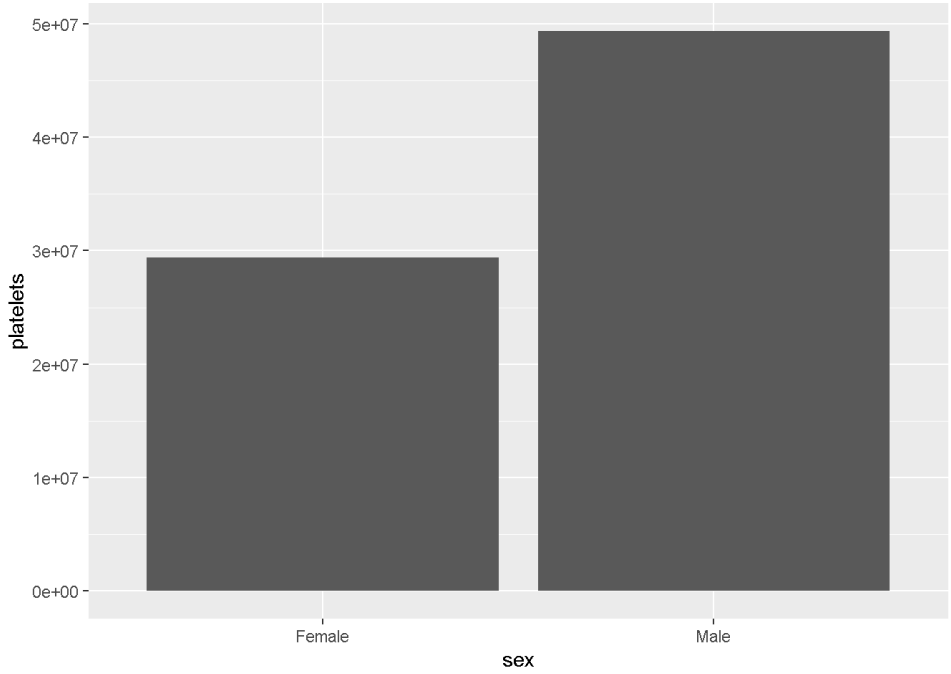
Plateletes are incredibly important. Having too few plateletes can lead to internal bleeding in intestines or stroke.

```
ggplot(HF, aes(x=age, y=platelets,color=diabetes)) + geom_point() +
  geom_smooth(method='lm', se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
ggplot(HF, aes(x=sex, y=platelets)) + geom_col()
```



Kaplan - Maeier Curve

- We can see we have remaining cases in which the person did was not declared deceased due to the ending of the curve not dropping down to 0%.

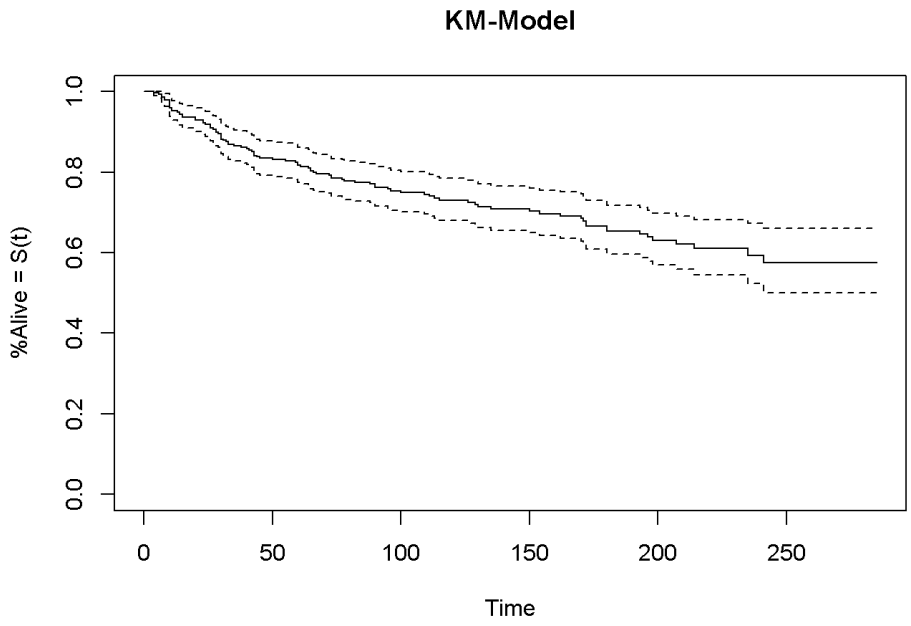
```
# Won't directly go from factor to numeric.
HF$DEATH_EVENT = as.numeric(as.character(HF$DEATH_EVENT))


# Predicting Survival without using any other variables besides Time.
fit <- survfit(Surv(time, DEATH_EVENT)~1, data=HF)

# Calling model itself returns:
# n:num of individuals
# nevent: number of decease events
# rmean*: no idea!
fit
```

```
## Call: survfit(formula = Surv(time, DEATH_EVENT) ~ 1, data = HF)
##
##        n events median 0.95LCL 0.95UCL
## [1,] 299     96     NA      NA      NA
```

```
# Calling the summary returns the hand calculations used to graph the curve.
# P((s0)): returns probabilty of surviving after `time`.
# P(1): No idea!
summary(fit)
```

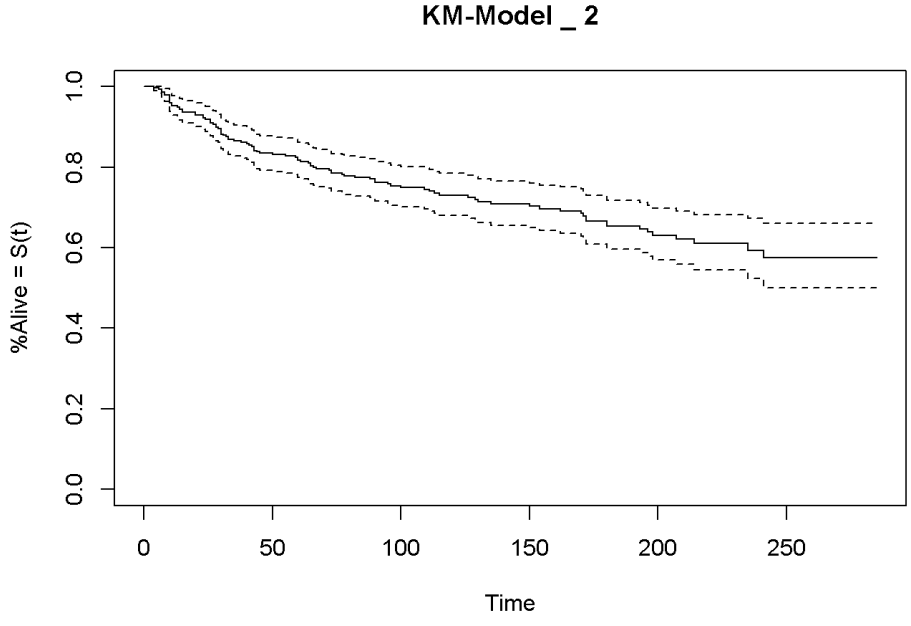
```
## Call: survfit(formula = Surv(time, DEATH_EVENT) ~ 1, data = HF)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     4    299       1    0.997 0.00334        0.990        1.000
##     6    298       1    0.993 0.00471        0.984        1.000
##     7    297       2    0.987 0.00664        0.974        1.000
##     8    295       2    0.980 0.00811        0.964        0.996
##    10    293       6    0.960 0.01135        0.938        0.982
##    11    287       2    0.953 0.01222        0.930        0.977
##    13    284       1    0.950 0.01263        0.925        0.975
##    14    283       2    0.943 0.01340        0.917        0.970
##    15    281       2    0.936 0.01412        0.909        0.964
##    20    278       2    0.930 0.01480        0.901        0.959
##    23    275       2    0.923 0.01545        0.893        0.954
##    24    273       1    0.920 0.01575        0.889        0.951
##    26    272       3    0.909 0.01663        0.877        0.943
##    27    269       1    0.906 0.01691        0.873        0.940
##    28    268       2    0.899 0.01745        0.866        0.934
##    29    266       1    0.896 0.01771        0.862        0.931
##    30    264       4    0.882 0.01869        0.846        0.920
##    31    259       1    0.879 0.01893        0.843        0.917
##    32    258       1    0.875 0.01916        0.839        0.914
##    33    257       2    0.869 0.01961        0.831        0.908
##    35    254       1    0.865 0.01983        0.827        0.905
##    38    253       1    0.862 0.02004        0.823        0.902
##    40    252       1    0.858 0.02025        0.820        0.899
##    41    251       1    0.855 0.02046        0.816        0.896
##    42    250       1    0.852 0.02066        0.812        0.893
##    43    249       3    0.841 0.02124        0.801        0.884
##    44    246       1    0.838 0.02143        0.797        0.881
##    45    245       1    0.834 0.02161        0.793        0.878
##    50    244       1    0.831 0.02179        0.789        0.875
##    55    241       1    0.828 0.02197        0.786        0.872
##    59    240       1    0.824 0.02215        0.782        0.869
##    60    239       2    0.817 0.02250        0.774        0.863
##    61    236       1    0.814 0.02267        0.771        0.859
##    64    234       1    0.810 0.02283        0.767        0.856
##    65    233       2    0.803 0.02316        0.759        0.850
##    66    231       1    0.800 0.02332        0.755        0.847
##    67    230       1    0.796 0.02348        0.752        0.844
##    72    227       1    0.793 0.02364        0.748        0.841
##    73    225       2    0.786 0.02394        0.740        0.834
##    77    217       1    0.782 0.02411        0.736        0.831
##    78    216       1    0.779 0.02427        0.732        0.828
##    82    207       1    0.775 0.02444        0.728        0.824
##    88    194       1    0.771 0.02464        0.724        0.821
##    90    189       2    0.763 0.02504        0.715        0.813
##    95    180       1    0.758 0.02526        0.711        0.810
##    96    175       1    0.754 0.02548        0.706        0.806
##   100    173       1    0.750 0.02571        0.701        0.802
##   109    159       1    0.745 0.02597        0.696        0.798
##   111    155       1    0.740 0.02625        0.691        0.794
##   113    152       1    0.735 0.02652        0.685        0.789
##   115    150       1    0.730 0.02679        0.680        0.785
##   126    136       1    0.725 0.02713        0.674        0.780
##   129    135       1    0.720 0.02746        0.668        0.776
##   130    134       1    0.714 0.02777        0.662        0.771
##   135    132       1    0.709 0.02808        0.656        0.766
##   150    118       1    0.703 0.02848        0.649        0.761
##   154    117       1    0.697 0.02886        0.643        0.756
##   162    116       1    0.691 0.02923        0.636        0.751
##   170    115       1    0.685 0.02959        0.629        0.745
##   171    114       1    0.679 0.02993        0.623        0.740
##   172    113       2    0.667 0.03059        0.610        0.730
##   180    106       2    0.654 0.03128        0.596        0.719
##   193     86       1    0.647 0.03183        0.587        0.712
##   196     83       1    0.639 0.03238        0.578        0.706
##   198     79       1    0.631 0.03297        0.569        0.699
##   207     71       1    0.622 0.03368        0.559        0.692
##   214     53       1    0.610 0.03503        0.545        0.683
##   235     37       1    0.594 0.03776        0.524        0.673
##   241     33       1    0.576 0.04068        0.501        0.661
```

```
plot(fit, xlab= "Time", ylab="%Alive = S(t)", main = "KM-Model")
```

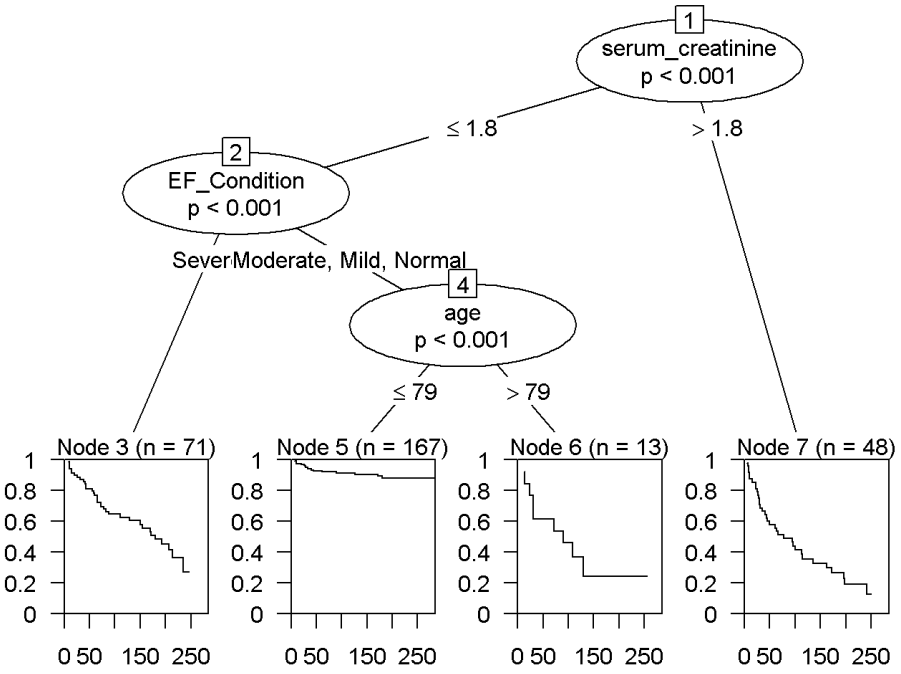


NEXT KM

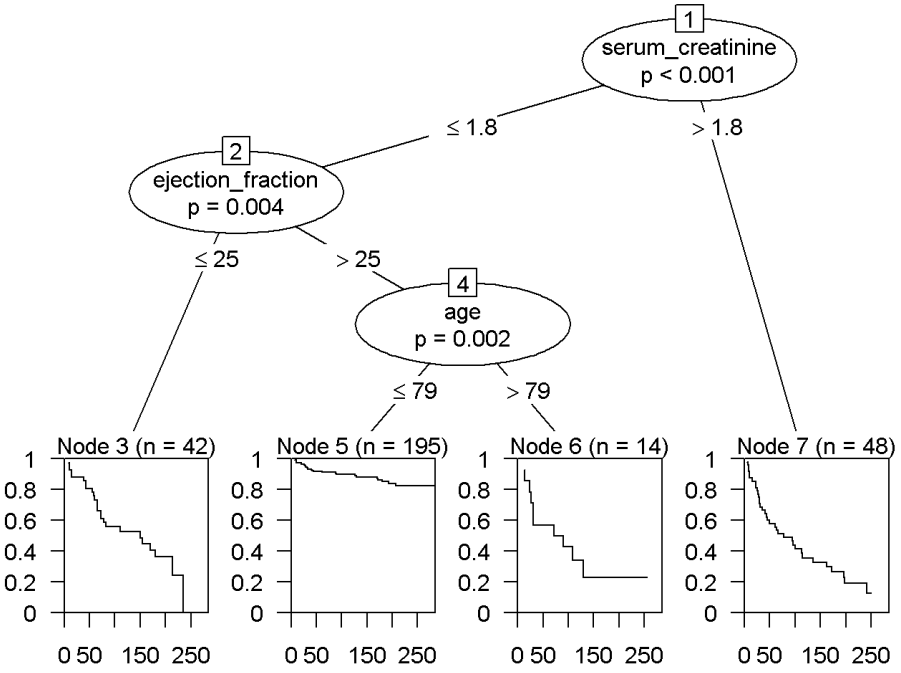
```
# The ~ 1 is our way ofletting R know that we aren't using any x variables. Just time and whether event occured which are both
 y variabes.
fit2 <- survfit(Surv(time, DEATH_EVENT) ~ 1, data=HF)


plot(fit2, xlab= "Time", ylab="%Alive = S(t)", main = "KM-Model _ 2")
```




```
stree <- ctree(Surv(time, DEATH_EVENT) ~ ., data = HF)


plot(stree)
```




```
# Dropping categorial Ejection Fraction.

HF <- HF %>% select(-EF_Condition)

yea <- ctree(Surv(time, DEATH_EVENT) ~ ., data = HF)


plot(yea)
```



```
K <- HF %>%
  filter(serum_creatinine <= 1.8, ejection_fraction > 25, age > 79)


# This one is best.
pred_k_surv <- survfit(Surv(time, DEATH_EVENT) ~ 1, data = K)


# The reason this still works is bc it's running KM at the end.
# the 'tree' part of it all isn't executing bc it doesn't find any variable that is # worth of being split on!
pred_k_tree <- ctree(Surv(time, DEATH_EVENT) ~ ., data = K)


pred_k_surv
```

```
## Call: survfit(formula = Surv(time, DEATH_EVENT) ~ 1, data = K)
##
##       n events median 0.95LCL 0.95UCL
## [1,] 14     10     81      30      NA
```

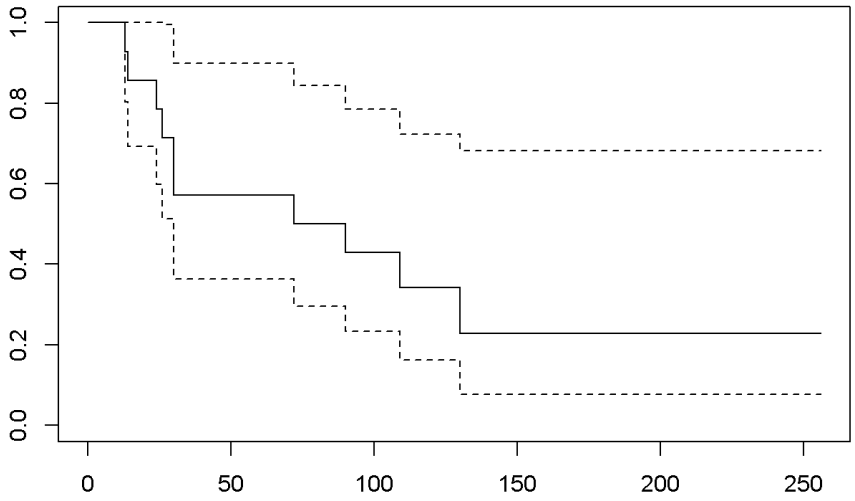
```
summary(pred_k_surv)
```

```
## Call: survfit(formula = Surv(time, DEATH_EVENT) ~ 1, data = K)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    13     14       1    0.929  0.0688        0.8030        1.000
##    14     13       1    0.857  0.0935        0.6921        1.000
##    24     12       1    0.786  0.1097        0.5977        1.000
##    26     11       1    0.714  0.1207        0.5129        0.995
##    30     10       2    0.571  0.1323        0.3630        0.899
##    72      8       1    0.500  0.1336        0.2961        0.844
##    90      7       1    0.429  0.1323        0.2341        0.785
##   109      5       1    0.343  0.1307        0.1624        0.724
##   130      3       1    0.229  0.1277        0.0765        0.683
```

```
summary(pred_k_surv, times=c(20, 45, 60, 80, 100, 10*(11:15)))
```

```
## Call: survfit(formula = Surv(time, DEATH_EVENT) ~ 1, data = K)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    20     12       2    0.857  0.0935        0.6921        1.000
##    45      8       4    0.571  0.1323        0.3630        0.899
##    60      8       0    0.571  0.1323        0.3630        0.899
##    80      7       1    0.500  0.1336        0.2961        0.844
##   100      6       1    0.429  0.1323        0.2341        0.785
##   110      4       1    0.343  0.1307        0.1624        0.724
##   120      4       0    0.343  0.1307        0.1624        0.724
##   130      3       1    0.229  0.1277        0.0765        0.683
##   140      2       0    0.229  0.1277        0.0765        0.683
##   150      2       0    0.229  0.1277        0.0765        0.683
```

```
plot(pred_k_surv)
```



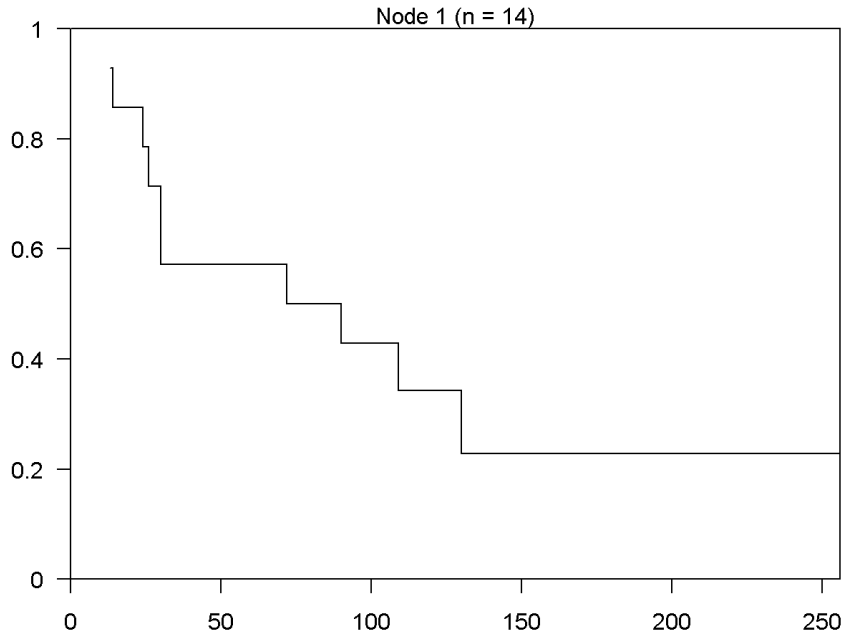

```
pred_k_tree
```

```
##
## Model formula:
## Surv(time, DEATH_EVENT) ~ age + anaemia + creatinine_phosphokinase +
##     diabetes + ejection_fraction + platelets + serum_creatinine +
##     serum_sodium + sex + smoking + hypertension
##
## Fitted party:
## [1] root: 90.000 (n = 14)
##
## Number of inner nodes:    0
## Number of terminal nodes: 1
```

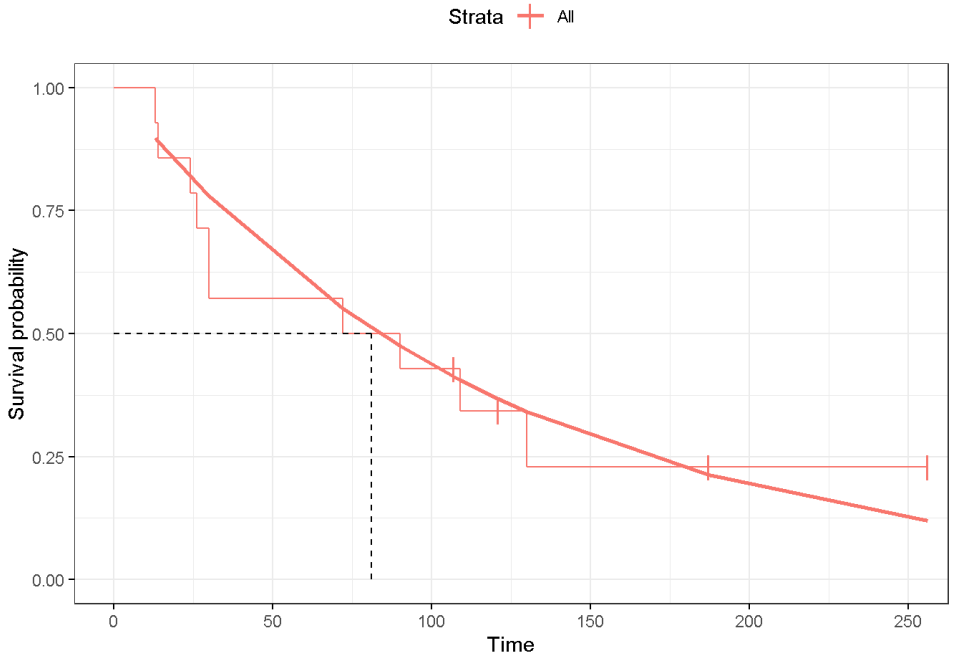
```
summary(pred_k_tree)
```

```
##   Length Class      Mode
## 1 1      constparty list
```

```
plot(pred_k_tree)
```




```
# Plotting the KM-Curve along with the exponential line that runs through it.

curve_exp <- flexsurvreg(Surv(time,DEATH_EVENT) ~ 1, data = K, dist = "exponential")

ggsurvplot(curve_exp, data = K,
           censor.shape="|",
           conf.int = FALSE, surv.median.line = "hv",
           ggtheme = theme_bw())
```



Cox Regression

- KM will make the curve based on event & time but that's all. We need to include the rest of the variables.

```
cox <- coxph(Surv(time, DEATH_EVENT) ~ ., data=K)
```

```
## Warning in coxph.fit(X, Y, istrat, offset, init, control, weights = weights, :
## Ran out of iterations and did not converge
```

```
cox
```

```
## Call:
## coxph(formula = Surv(time, DEATH_EVENT) ~ ., data = K)
##
##                               coef   exp(coef)    se(coef)         z      p
## age                       -9.732e+01   5.439e-43   1.768e-01  -550.516 <2e-16
## anaemia1                   9.857e+02         Inf   4.931e+02     1.999 0.0456
## creatinine_phosphokinase  -8.907e-02   9.148e-01   1.779e-03   -50.075 <2e-16
## diabetesPresent           -2.878e+02  1.005e-125   5.834e+02    -0.493 0.6218
## ejection_fraction         -5.116e-01   5.996e-01   2.256e+01    -0.023 0.9819
## platelets                 -1.795e-04   9.998e-01   1.489e-05   -12.059 <2e-16
## serum_creatinine          -2.513e+03   0.000e+00   1.474e+03    -1.705 0.0882
## serum_sodium              -1.302e+02   2.878e-57   1.286e-01 -1012.645 <2e-16
## sexMale                   -5.871e+02  1.012e-255   1.414e+00  -415.178 <2e-16
## smoking1                   4.498e+02  2.147e+195   4.784e+02     0.940 0.3472
## hypertensionPresent        1.543e+02   9.996e+66   1.414e+00   109.088 <2e-16
##
## Likelihood ratio test=41.25  on 11 df, p=2.179e-05
## n= 14, number of events= 10
```

```
summary(cox)
```

```
## Call:
## coxph(formula = Surv(time, DEATH_EVENT) ~ ., data = K)
##
##   n= 14, number of events= 10
##
##                               coef   exp(coef)    se(coef)         z Pr(>|z|)
## age                     -9.732e+01   5.439e-43   1.768e-01  -550.516   <2e-16
## anaemia1                 9.857e+02         Inf   4.931e+02     1.999   0.0456
## creatinine_phosphokinase -8.907e-02   9.148e-01   1.779e-03   -50.075   <2e-16
## diabetesPresent          -2.878e+02  1.005e-125   5.834e+02    -0.493   0.6218
## ejection_fraction        -5.116e-01   5.996e-01   2.256e+01    -0.023   0.9819
## platelets                -1.795e-04   9.998e-01   1.489e-05   -12.059   <2e-16
## serum_creatinine         -2.513e+03   0.000e+00   1.474e+03    -1.705   0.0882
## serum_sodium             -1.302e+02   2.878e-57   1.286e-01 -1012.645   <2e-16
## sexMale                  -5.871e+02  1.012e-255   1.414e+00  -415.178   <2e-16
## smoking1                  4.498e+02  2.147e+195   4.784e+02     0.940   0.3472
## hypertensionPresent       1.543e+02   9.996e+66   1.414e+00   109.088   <2e-16
##
## age                     ***
## anaemia1                *
## creatinine_phosphokinase ***
## diabetesPresent
## ejection_fraction
## platelets               ***
## serum_creatinine        .
## serum_sodium            ***
## sexMale                 ***
## smoking1
## hypertensionPresent     ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                          exp(coef) exp(-coef)  lower .95  upper .95
## age                      5.439e-43  1.839e+42  3.846e-43  7.691e-43
## anaemia1                       Inf  0.000e+00  2.537e+08        Inf
## creatinine_phosphokinase 9.148e-01  1.093e+00  9.116e-01  9.180e-01
## diabetesPresent          1.005e-125 9.946e+124  0.000e+00        Inf
## ejection_fraction        5.996e-01  1.668e+00  3.761e-20  9.557e+18
## platelets                9.998e-01  1.000e+00  9.998e-01  9.998e-01
## serum_creatinine         0.000e+00        Inf  0.000e+00 2.216e+163
## serum_sodium             2.878e-57  3.475e+56  2.237e-57  3.702e-57
## sexMale                 1.012e-255 9.883e+254 6.329e-257 1.618e-254
## smoking1                2.147e+195 4.658e-196 1.203e-212        Inf
## hypertensionPresent      9.996e+66  1.000e-67  6.253e+65  1.598e+68
##
## Concordance= 1  (se = 0 )
## Likelihood ratio test= 41.25  on 11 df,   p=2e-05
## Wald test            = 1512798  on 11 df,   p=<2e-16
## Score (logrank) test = 21.08  on 11 df,   p=0.03
```

NEXT STEPS: RELAXING RESTRICTIONS FOR THE TREE AND STARTING WITH OTHER PREDICTORS OVER SERUM_CREATININE