Analysis Of Survival Data - Lung Cancer Patients

Github link - https://github.com/jsv-datascientist/survival_analysis_in_R

What is the aim of the analysis?

We are interested in knowing how long a subject suffering from a particular disease is meant to survive. **The event of interest in this case is the death of a patient suffering from lung cancer**.

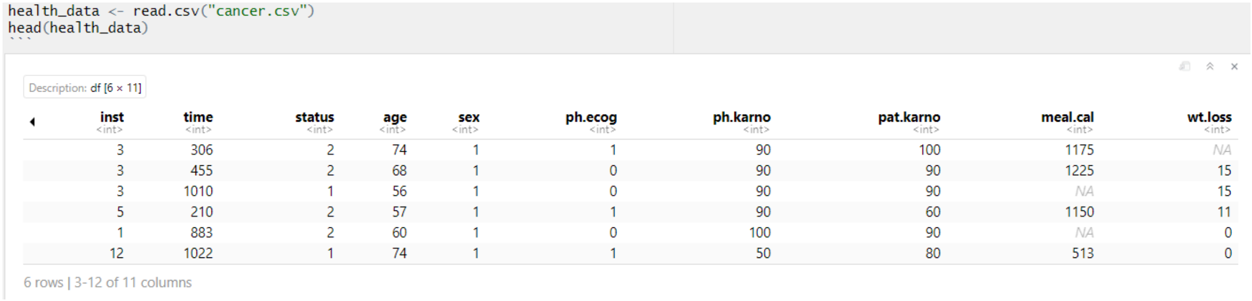
Here we are trying to accomplish below all for the given dataset.

Nonparametric Estimation - Kaplan-Meier: Suitable for survival function estimation.

Nonparametric Comparison - Log-Rank : Tests for differences between survival distributions.

Semi-Parametric Model - Cox Regression : Evaluates covariates' effects while assuming proportional hazards.

1. Data Exploration



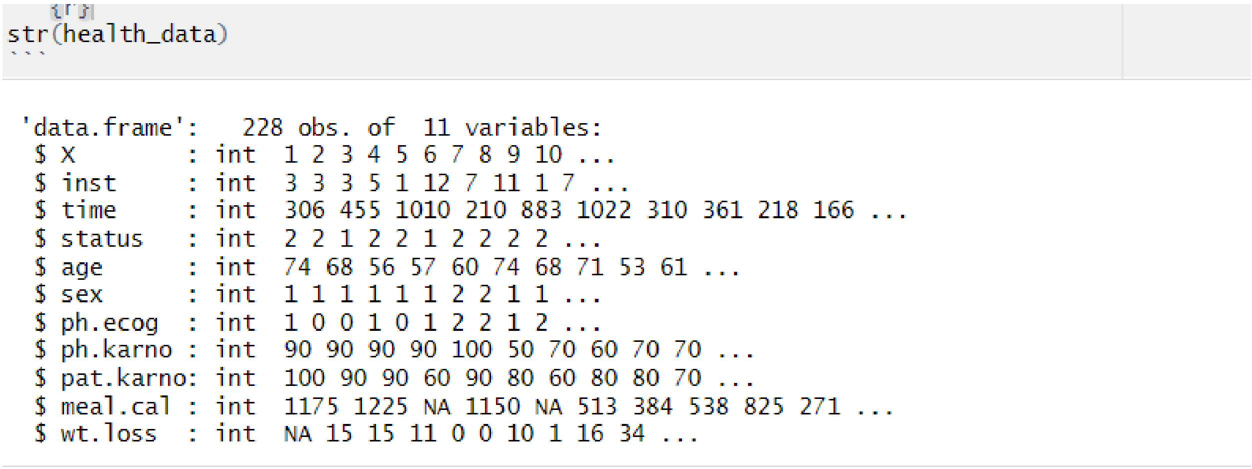
```
health_data <- read.csv("cancer.csv")
head(health_data)
```

Description: df [6 × 11]

inst <int>	time <int>	status <int>	age <int>	sex <int>	ph.ecog <int>	ph.karno <int>	pat.karno <int>	meal.cal <int>	wt.loss <int>
3	306	2	74	1	1	90	100	1175	NA
3	455	2	68	1	0	90	90	1225	15
3	1010	1	56	1	0	90	90	NA	15
5	210	2	57	1	1	90	60	1150	11
1	883	2	60	1	0	100	90	NA	0
12	1022	1	74	1	1	50	80	513	0

6 rows | 3-12 of 11 columns

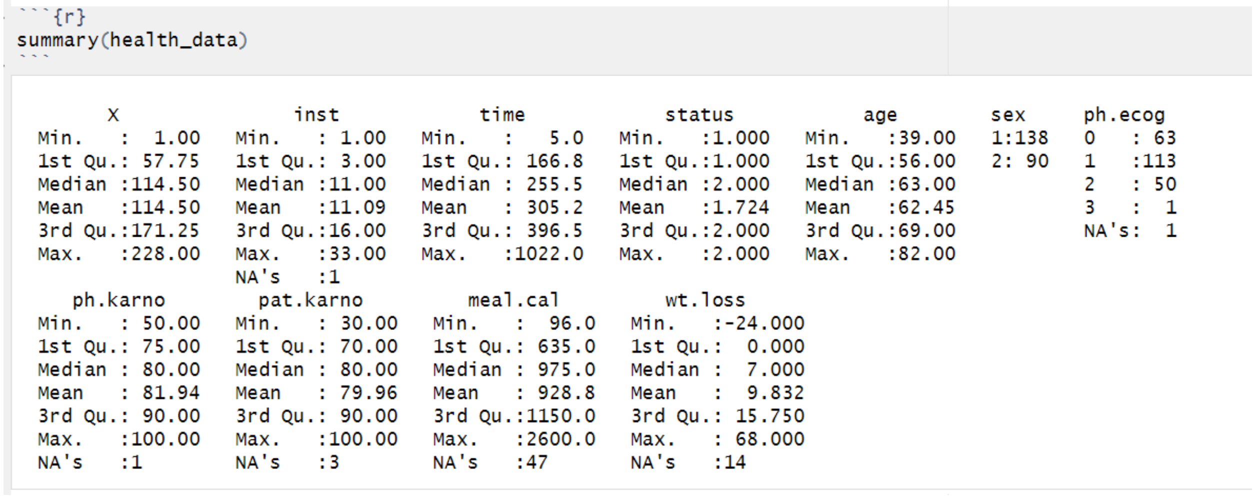
Explore the datatype of the model



```
str(health_data)
```

```
'data.frame':   228 obs. of  11 variables:
 $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ inst     : int  3 3 3 5 1 12 7 11 1 7 ...
 $ time     : int  306 455 1010 210 883 1022 310 361 218 166 ...
 $ status   : int  2 2 1 2 2 1 2 2 2 2 ...
 $ age      : int  74 68 56 57 60 74 68 71 53 61 ...
 $ sex      : int  1 1 1 1 1 1 2 2 1 1 ...
 $ ph.ecog  : int  1 0 0 1 0 1 2 2 1 2 ...
 $ ph.karno : int  90 90 90 90 100 50 70 60 70 70 ...
 $ pat.karno: int  100 90 90 60 90 80 60 80 80 70 ...
 $ meal.cal : int  1175 1225 NA 1150 NA 513 384 538 825 271 ...
 $ wt.loss  : int  NA 15 15 11 0 0 10 1 16 34 ...
```

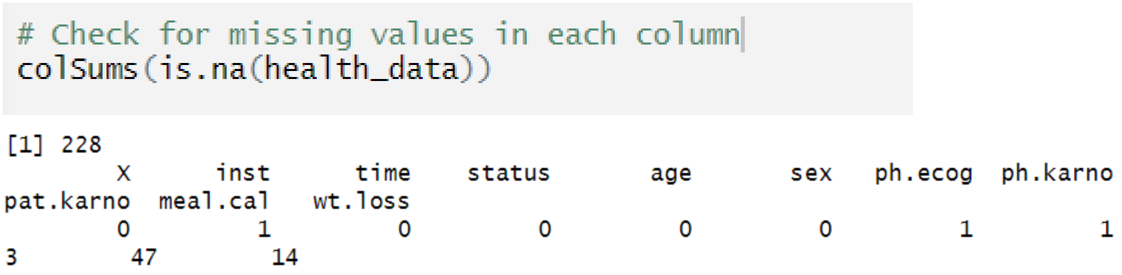
Summary of data



~~~
```{r}
summary(health_data)
```
~~~

```
       X              inst            time            status           age         sex     ph.ecog  
 Min.   :  1.00   Min.   : 1.00   Min.   :   5.0   Min.   :1.000   Min.   :39.00   1:138   0   : 63  
 1st Qu.: 57.75   1st Qu.: 3.00   1st Qu.: 166.8   1st Qu.:1.000   1st Qu.:56.00   2: 90   1   :113  
 Median :114.50   Median :11.00   Median : 255.5   Median :2.000   Median :63.00           2   : 50  
 Mean   :114.50   Mean   :11.09   Mean   : 305.2   Mean   :1.724   Mean   :62.45           3   :  1  
 3rd Qu.:171.25   3rd Qu.:16.00   3rd Qu.: 396.5   3rd Qu.:2.000   3rd Qu.:69.00           NA's:  1  
 Max.   :228.00   Max.   :33.00   Max.   :1022.0   Max.   :2.000   Max.   :82.00                     
                  NA's   :1                                                                           
    ph.karno        pat.karno         meal.cal         wt.loss       
 Min.   : 50.00   Min.   : 30.00   Min.   :  96.0   Min.   :-24.000  
 1st Qu.: 75.00   1st Qu.: 70.00   1st Qu.: 635.0   1st Qu.:  0.000  
 Median : 80.00   Median : 80.00   Median : 975.0   Median :  7.000  
 Mean   : 81.94   Mean   : 79.96   Mean   : 928.8   Mean   :  9.832  
 3rd Qu.: 90.00   3rd Qu.: 90.00   3rd Qu.:1150.0   3rd Qu.: 15.750  
 Max.   :100.00   Max.   :100.00   Max.   :2600.0   Max.   : 68.000  
 NA's   :1        NA's   :3        NA's   :47       NA's   :14       
```

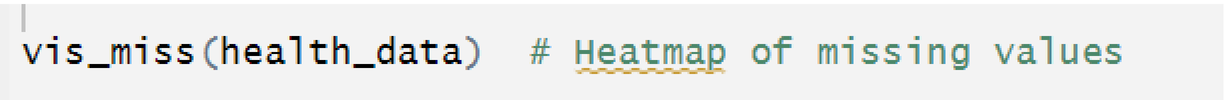
Check for missing values



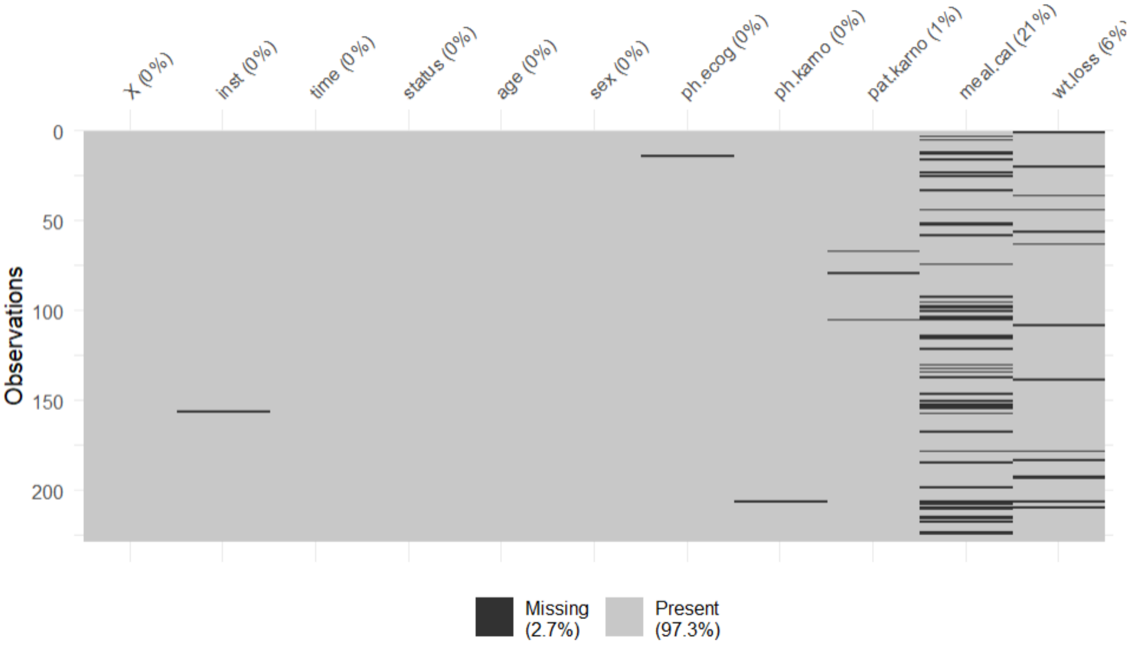
```
# Check for missing values in each column
colSums(is.na(health_data))
```

```
[1] 228
        X      inst      time    status       age       sex   ph.ecog  ph.karno 
pat.karno  meal.cal   wt.loss 
        0         1         0         0         0         0         1         1 
3        47        14 
```

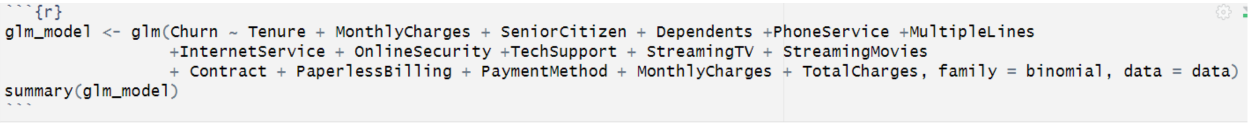
We can perform the MAR (Missing at Random), MNAR (Missing not at Random) test.



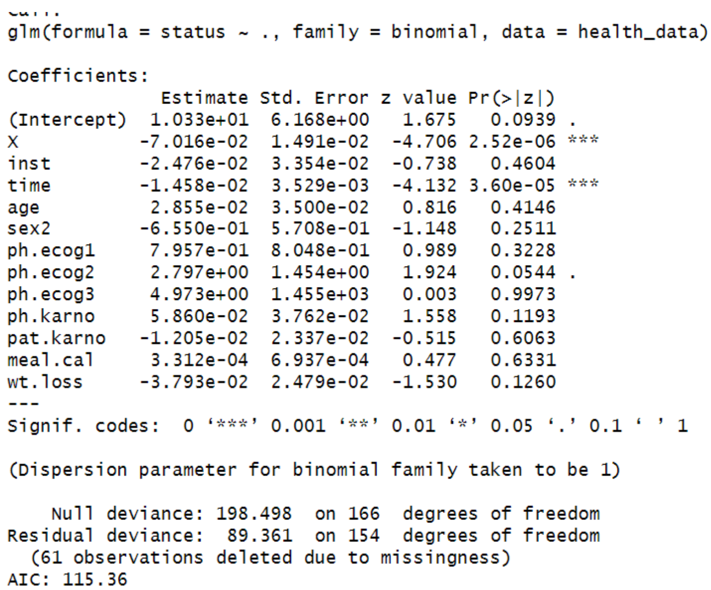
```
vis_miss(health_data)  # Heatmap of missing values
```



Check GLM to know the impacted variables



~~~
```{r}
glm_model <- glm(Churn ~ Tenure + MonthlyCharges + SeniorCitizen + Dependents +PhoneService +MultipleLines
                +InternetService + OnlineSecurity +TechSupport + StreamingTV + StreamingMovies
                + Contract + PaperlessBilling + PaymentMethod + MonthlyCharges + TotalCharges, family = binomial, data = data)
summary(glm_model)
```
~~~



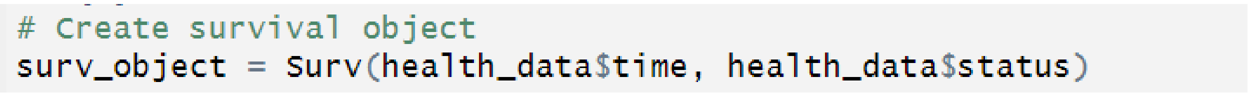
```
glm(formula = status ~ ., family = binomial, data = health_data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)  1.033e+01  6.168e+00   1.675   0.0939 .  
X           -7.016e-02  1.491e-02  -4.706 2.52e-06 ***
inst        -2.476e-02  3.354e-02  -0.738   0.4604    
time        -1.458e-02  3.529e-03  -4.132 3.60e-05 ***
age          2.855e-02  3.500e-02   0.816   0.4146    
sex2        -6.550e-01  5.708e-01  -1.148   0.2511    
ph.ecog1     7.957e-01  8.048e-01   0.989   0.3228    
ph.ecog2     2.797e+00  1.454e+00   1.924   0.0544 .  
ph.ecog3     4.973e+00  1.455e+03   0.003   0.9973    
ph.karno     5.860e-02  3.762e-02   1.558   0.1193    
pat.karno   -1.205e-02  2.337e-02  -0.515   0.6063    
meal.cal     3.312e-04  6.937e-04   0.477   0.6331    
wt.loss     -3.793e-02  2.479e-02  -1.530   0.1260    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 198.498  on 166  degrees of freedom
Residual deviance:  89.361  on 154  degrees of freedom
  (61 observations deleted due to missingness)
AIC: 115.36
```

This says time plays a vital role

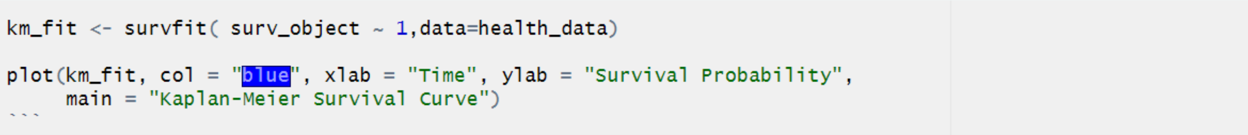
2. Nonparametric Estimation - Kaplan-Meier

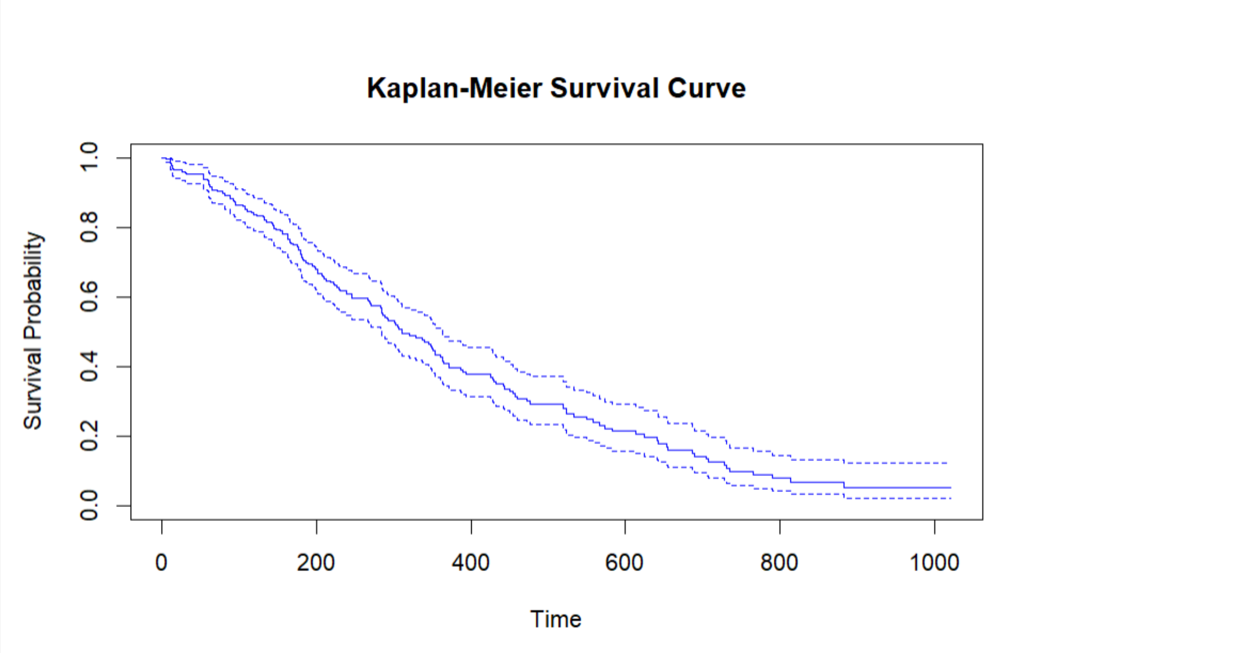
KM plot gives us visual curves.

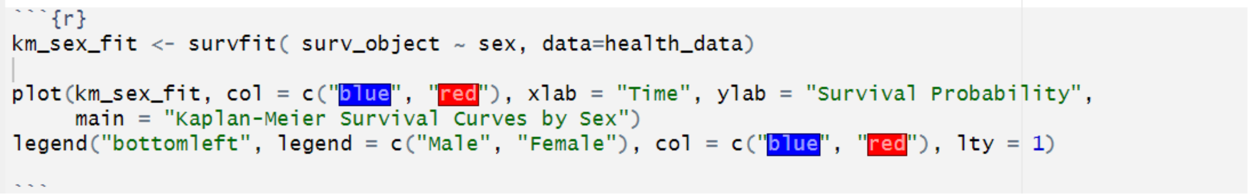
Create the Survival object

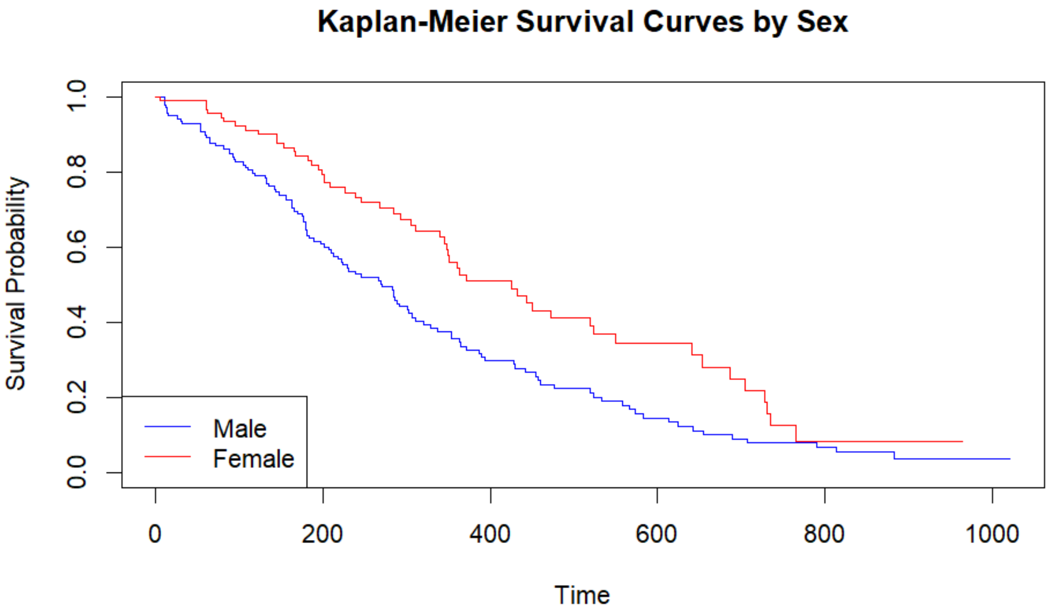
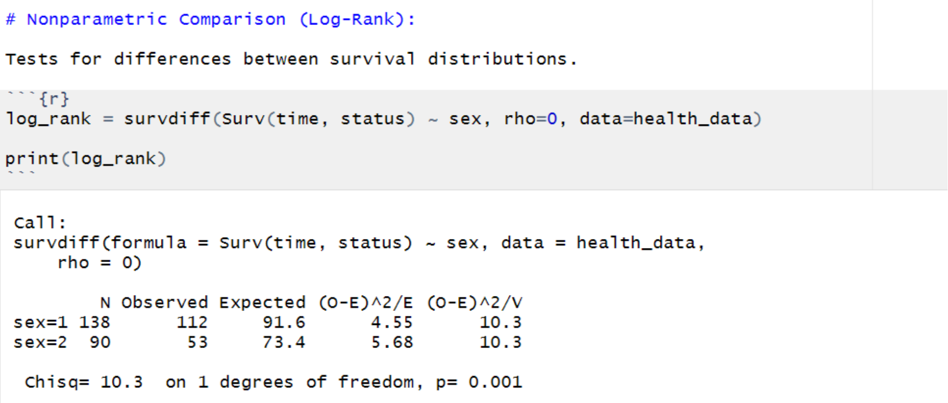


```
# Create survival object
surv_object = Surv(health_data$time, health_data$status)
```



```
km_fit <- survfit( surv_object ~ 1,data=health_data)

plot(km_fit, col = "blue", xlab = "Time", ylab = "Survival Probability",
     main = "Kaplan-Meier Survival Curve")
```
```



The survival probability drops below 20%, meaning most individuals have experienced the event ( death)

**KM plot with Sex/Gender factor**
```



~~~
```{r}
km_sex_fit <- survfit( surv_object ~ sex, data=health_data)

plot(km_sex_fit, col = c("blue", "red"), xlab = "Time", ylab = "Survival Probability",
     main = "Kaplan-Meier Survival Curves by Sex")
legend("bottomleft", legend = c("Male", "Female"), col = c("blue", "red"), lty = 1)

~~~
```


Kaplan-Meier Survival Curves by Sex
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
Time
Male
Female

Over time, the survival probability for females (red) is higher than for males at most points in time. This suggests that females tend to survive longer than males in the dataset.

3. Nonparametric Comparison - Log-Rank :

Log-rank tells us if the curves are statistically different.



```
# Nonparametric Comparison (Log-Rank):

Tests for differences between survival distributions.

```{r}
log_rank = survdiff(Surv(time, status) ~ sex, rho=0, data=health_data)

print(log_rank)
```

 Call:
 survdiff(formula = Surv(time, status) ~ sex, data = health_data,
     rho = 0)

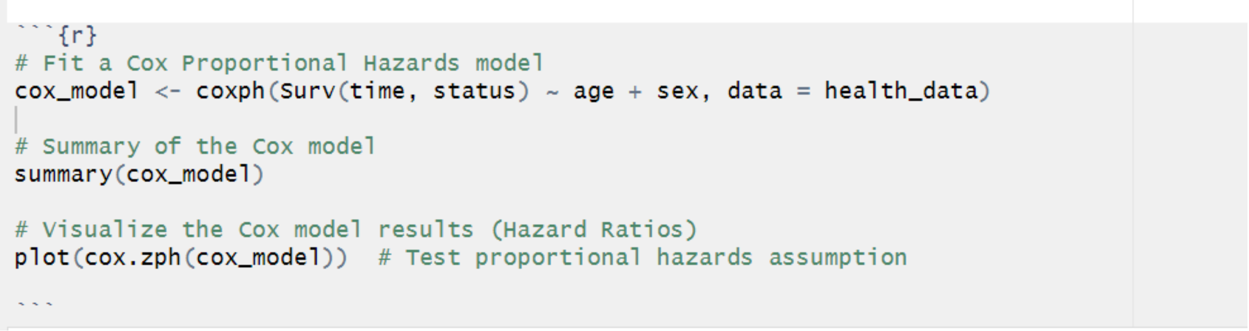
         N Observed Expected (O-E)^2/E (O-E)^2/V
 sex=1 138   112   91.6   4.55   10.3
 sex=2 90   53   73.4   5.68   10.3

  Chisq= 10.3 on 1 degrees of freedom, p= 0.001
```

A p-value below 0.05 indicates that the difference in survival between the groups is statistically significant
```

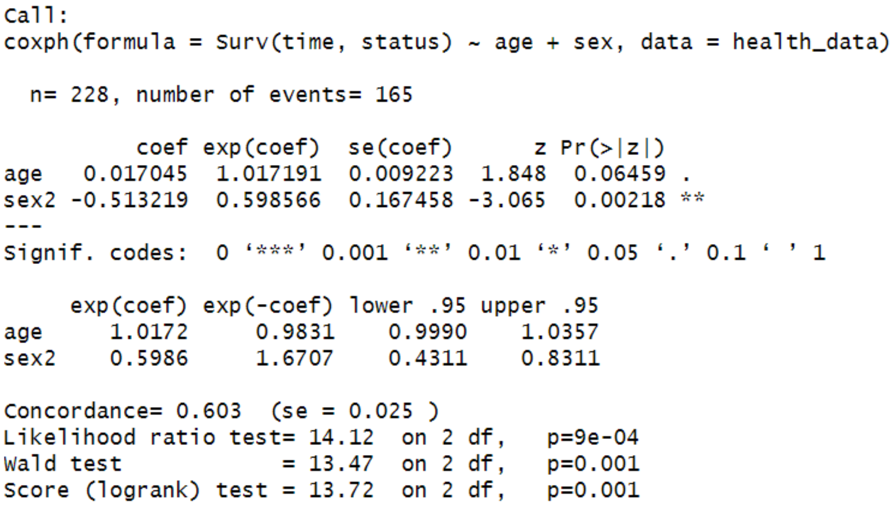
# 4. Semi-Parametric Model - Cox Regression

Cox regression explains why the curves are different and which factors matter most.

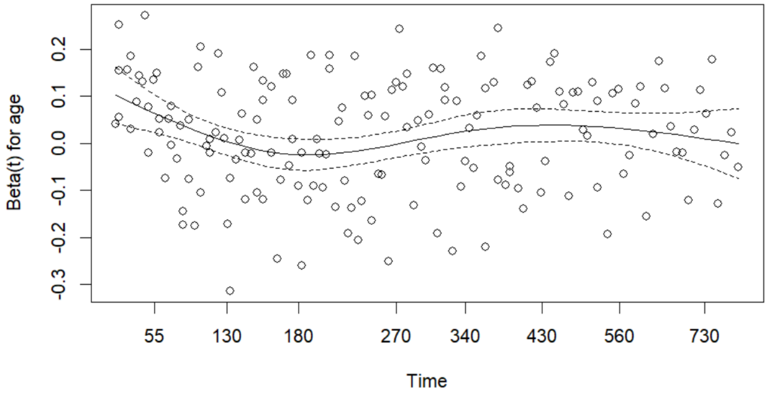
For Age and Sex, we apply coxph

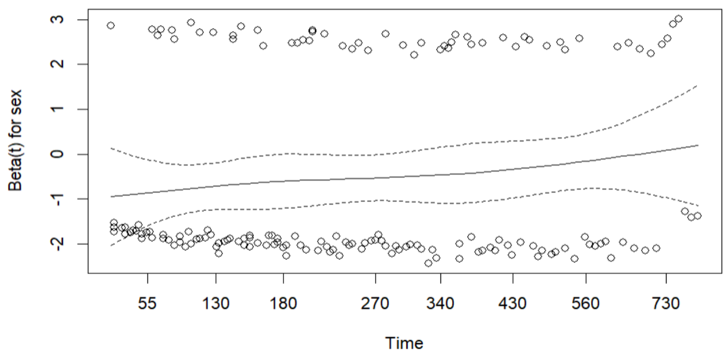


~~~
```{r}
# Fit a Cox Proportional Hazards model
cox_model <- coxph(Surv(time, status) ~ age + sex, data = health_data)

# Summary of the Cox model
summary(cox_model)

# Visualize the Cox model results (Hazard Ratios)
plot(cox.zph(cox_model))  # Test proportional hazards assumption

```
~~~



```
Call:
coxph(formula = Surv(time, status) ~ age + sex, data = health_data)

  n= 228, number of events= 165 

          coef exp(coef)  se(coef)      z Pr(>|z|)   
age   0.017045  1.017191  0.009223  1.848  0.06459 . 
sex2 -0.513219  0.598566  0.167458 -3.065  0.00218 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     exp(coef) exp(-coef) lower .95 upper .95
age     1.0172     0.9831    0.9990    1.0357
sex2    0.5986     1.6707    0.4311    0.8311

Concordance= 0.603  (se = 0.025 )
Likelihood ratio test= 14.12  on 2 df,   p=9e-04
Wald test            = 13.47  on 2 df,   p=0.001
Score (logrank) test = 13.72  on 2 df,   p=0.001
```





Below are the observations for age and sex in an event to occur.

age :

coef = 0.017045   The positive coefficient suggests that as age increases, the hazard slightly increases.

HR = 1.017, For every 1-year increase in age, the hazard increases by 1.7% (HR > 1 indicates higher hazard).

p = 0.06459): The p-value is slightly above 0.05, meaning the effect of **age on survival is not statistically significant at the 5% leve**l but may be borderline significant.

95% CI: 0.9990 – 1.0357, The confidence interval for age includes 1, suggesting that its effect **may not be strongly significant**

sex2:

coef = -0.513219, The negative coefficient for sex2 indicates that **females have a lower hazard (risk of the event) compared to males.**

HR = 0.599- Females have 59.9% of the hazard of males, meaning they are less likely to experience the event (death)

p = 0.00218, The effect of **sex on survival is highly significan**t, indicating that the survival difference between males and females is not due to chance.

95% CI: 0.4311 – 0.831, The confidence interval for **sex2 does not include 1**, supporting the conclusion that f**emales have a significantly lower hazard than males**

*This shows the **Sex (Females)** have **statistically significantly** lower risk of death compared to males and **age does not have enough statistically evident** to impact the event of death.*