

# Data Processing

Team 2

2025-11-03

## Data reading

```
data <- read.csv("00_ProstateCancer_Data.csv", header=T)
head(data)
```

```
## Hasta_ID Yas Tani_Tarihi PSA_Tani Klinik_Evre Biyopsi_Gleason Risk_Grubu
## 1      0  59  2022-12-09    41.3      cT3a             3+5           3
## 2      1  69  2023-08-26    12.9      cT2b             4+3           2
## 3      2  66  2021-11-17    32.0      cT3b             3+5           3
## 4      3  59  2022-01-15   142.7      cT3b             4+4           3
## 5      4  70  2021-07-07    16.2      cT2b             3+4           2
## 6      5  73  2023-06-17    55.7      cT3b             5+4           3
## Albumin Lenfosit  CRP NLR CALLY_Index Komorbidite_Skor Tedavi_Tipi
## 1      4.1      2227 0.58 1.7          1.6              2           1
## 2      4.6      1168 0.12 1.7          4.5              0           1
## 3      4.1      1125 0.16 3.5          2.9              1           4
## 4      3.5      1623 1.76 1.7          0.3              0           2
## 5      4.3      1399 0.53 2.3          1.1              2           2
## 6      4.2      2103 0.10 2.2          8.8              4           2
## Tedavi_Tarihi RT_Dozu ADT_Tipi ADT_Suresi Patolojik_Evre Cerrahi_Sinir
## 1  2023-02-14      NA      NA      NA      pT2c              0
## 2  2023-10-08      NA      NA      NA      pT2a              1
## 3  2022-01-19      70       1      12              NA
## 4  2022-03-30      74      NA      NA              NA
## 5  2021-07-28      70      NA      NA              NA
## 6  2023-07-28      70      NA      NA              NA
## Final_Gleason PSA_Nadir PSA_Takip_3ay PSA_Takip_6ay PSA_Takip_12ay BCR_Durum
## 1      3+5      0.14      0.14      0.17      0.21      True
## 2      4+5      0.04      0.04      0.04      0.06      False
## 3      0.30      1.03      0.33      0.38      False
## 4      0.20      0.61      0.63      0.71      False
## 5      0.28      0.90      0.38      0.46      False
## 6      0.81      2.38      1.57      4.47      True
## BCR_Tarihi Metastaz_Durum Metastaz_Tarihi Son_Durum Son_Takip_Tarihi
## 1 2028-01-11      0      NA      1      2029-05-22
## 2      0      NA      1      2029-06-12
## 3      0      NA      0      2022-03-05
## 4      0      NA      1      2023-11-06
## 5      0      NA      1      2027-11-16
## 6 2025-04-06      0      NA      1      2029-04-19
```

```
str(data)
```

```
## 'data.frame':    600 obs. of  31 variables:
## $ Hasta_ID      : int  0 1 2 3 4 5 6 7 8 9 ...
## $ Yas           : int  59 69 66 59 70 73 72 72 63 57 ...
## $ Tani_Tarihi   : chr  "2022-12-09" "2023-08-26" "2021-11-17" "2022-01-15" ...
## $ PSA_Tani      : num  41.3 12.9 32 142.7 16.2 ...
## $ Klinik_Evre   : chr  "cT3a" "cT2b" "cT3b" "cT3b" ...
## $ Biyopsi_Gleason : chr  "3+5" "4+3" "3+5" "4+4" ...
## $ Risk_Grubu    : int  3 2 3 3 2 3 3 3 3 3 ...
## $ Albumin       : num  4.1 4.6 4.1 3.5 4.3 4.2 4.8 4.2 4 4.5 ...
## $ Lenfosit      : int  2227 1168 1125 1623 1399 2103 2038 1418 1936 1348 ...
## $ CRP           : num  0.58 0.12 0.16 1.76 0.53 0.1 0.46 0.27 0.11 0.62 ...
## $ NLR           : num  1.7 1.7 3.5 1.7 2.3 2.2 4 2.8 2.3 1.8 ...
## $ CALLY_Index   : num  1.6 4.5 2.9 0.3 1.1 8.8 2.1 2.2 7 1 ...
## $ Komorbidite_Skor: int  2 0 1 0 2 4 0 2 3 5 ...
## $ Tedavi_Tipi   : int  1 1 4 2 2 2 2 1 1 3 ...
## $ Tedavi_Tarihi : chr  "2023-02-14" "2023-10-08" "2022-01-19" "2022-03-30" ...
## $ RT_Dozu       : num  NA NA 70 74 70 70 76 NA NA NA ...
## $ ADT_Tipi      : num  NA NA 1 NA NA NA NA NA NA 2 ...
## $ ADT_Suresi    : num  NA NA 12 NA NA NA NA NA NA 12 ...
## $ Patolojik_Evre : chr  "pT2c" "pT2a" "" "" ...
## $ Cerrahi_Sinir : num  0 1 NA NA NA NA NA 1 1 NA ...
## $ Final_Gleason : chr  "3+5" "4+5" "" "" ...
## $ PSA_Nadir     : num  0.14 0.04 0.3 0.2 0.28 0.81 0.36 0.04 0.04 0.39 ...
## $ PSA_Takip_3ay : num  0.14 0.04 1.03 0.61 0.9 2.38 1.28 0.04 0.05 1.5 ...
## $ PSA_Takip_6ay : num  0.17 0.04 0.33 0.63 0.38 1.57 0.94 0.04 0.04 0.82 ...
## $ PSA_Takip_12ay : num  0.21 0.06 0.38 0.71 0.46 4.47 0.56 0.08 0.04 0.39 ...
## $ BCR_Durum     : chr  "True" "False" "False" "False" ...
## $ BCR_Tarihi    : chr  "2028-01-11" "" "" "" ...
## $ Metastaz_Durum : int  0 0 0 0 0 0 0 0 0 0 ...
## $ Metastaz_Tarihi : logi  NA NA NA NA NA NA ...
## $ Son_Durum     : int  1 1 0 1 1 1 1 1 1 1 ...
## $ Son_Takip_Tarihi: chr  "2029-05-22" "2029-06-12" "2022-03-05" "2023-11-06" ...
```

```
summary(data)
```

```
##      Hasta_ID      Yas      Tani_Tarihi      PSA_Tani
## Min.   : 0.0   Min.   :50.00   Length:600   Min.    : 2.50
## 1st Qu.:149.8   1st Qu.:62.00   Class :character   1st Qu.: 18.10
## Median :299.5   Median :67.00   Mode  :character   Median : 59.35
## Mean   :299.5   Mean   :67.09                      Mean   : 64.70
## 3rd Qu.:449.2   3rd Qu.:72.00                      3rd Qu.:108.42
## Max.   :599.0   Max.   :85.00                      Max.   :150.00
##
## Klinik_Evre      Biyopsi_Gleason      Risk_Grubu      Albumin
## Length:600      Length:600      Min.    :1.00   Min.    :3.500
## Class :character   Class :character   1st Qu.:2.00   1st Qu.:4.000
## Mode  :character   Mode  :character   Median :3.00   Median :4.200
##                      Mean   :2.62   Mean   :4.219
##                      3rd Qu.:3.00   3rd Qu.:4.500
##                      Max.   :3.00   Max.   :5.000
##
```

```

##      Lenfosit          CRP          NLR          CALLY_Index
##  Min.   :1000    Min.   :0.100    Min.   :1.000    Min.   : 0.200
## 1st Qu.:1561    1st Qu.:0.130    1st Qu.:2.000    1st Qu.: 1.100
## Median :1834    Median :0.345    Median :2.500    Median : 2.100
## Mean   :1827    Mean   :0.518    Mean   :2.503    Mean   : 3.346
## 3rd Qu.:2093    3rd Qu.:0.690    3rd Qu.:3.000    3rd Qu.: 5.425
## Max.   :2936    Max.   :3.590    Max.   :5.000    Max.   :13.000
##
## Komorbidite_Skor  Tedavi_Tipi  Tedavi_Tarihi  RT_Dozu
##  Min.   :0.000    Min.   :1.000    Length:600    Min.   :70.00
## 1st Qu.:1.000    1st Qu.:1.000    Class :character 1st Qu.:74.00
## Median :3.000    Median :2.000    Mode  :character Median :76.00
## Mean   :2.515    Mean   :2.005                      Mean  :74.63
## 3rd Qu.:4.000    3rd Qu.:3.000                      3rd Qu.:76.00
## Max.   :5.000    Max.   :4.000                      Max.   :78.00
##                                     NA's   :325
##      ADT_Tipi      ADT_Suresi  Patolojik_Evre  Cerrahi_Sinir
##  Min.   :1.000    Min.   : 6    Length:600    Min.   :0.0000
## 1st Qu.:1.000    1st Qu.:12    Class :character 1st Qu.:0.0000
## Median :2.000    Median :24    Mode  :character Median :0.0000
## Mean   :2.012    Mean   :21                      Mean  :0.4764
## 3rd Qu.:3.000    3rd Qu.:36                      3rd Qu.:1.0000
## Max.   :3.000    Max.   :36                      Max.   :1.0000
## NA's   :436      NA's   :436                      NA's   :346
## Final_Gleason      PSA_Nadir      PSA_Takip_3ay  PSA_Takip_6ay
## Length:600          Min.   :0.0100    Min.   :0.0100    Min.   :0.0100
## Class :character    1st Qu.:0.0400    1st Qu.:0.0500    1st Qu.:0.0700
## Mode  :character    Median :0.1900    Median :0.7300    Median :0.4700
##                      Mean   :0.2847    Mean   :0.7633    Mean   :0.7016
##                      3rd Qu.:0.4400    3rd Qu.:1.2800    3rd Qu.:0.8925
##                      Max.   :1.0000    Max.   :2.4900    Max.   :2.9900
##
## PSA_Takip_12ay      BCR_Durum      BCR_Tarihi      Metastaz_Durum
##  Min.   :0.0100    Length:600    Length:600    Min.   :0
## 1st Qu.:0.0800    Class :character  Class :character 1st Qu.:0
## Median :0.4100    Mode  :character  Mode  :character Median :0
## Mean   :0.9256                      Mean   :0
## 3rd Qu.:0.6700                      3rd Qu.:0
## Max.   :4.9700                      Max.   :0
##
## Metastaz_Tarihi      Son_Durum      Son_Takip_Tarihi
## Mode:logical          Min.   :0.0000    Length:600
## NA's:600              1st Qu.:1.0000    Class :character
##                      Median :1.0000    Mode  :character
##                      Mean   :0.8817
##                      3rd Qu.:1.0000
##                      Max.   :1.0000
##

```

## Dataset Variables

### Initial Diagnosis

**Hasta\_ID (Categorical):** Patient ID

**Yas (Discrete):** Age

**Tani\_Tarihi (Date):** Diagnosis Date

**PSA\_Tani (Continuous):** Serum Prostate-Specific Antigen (PSA) level at diagnosis (ng/mL)

**Klinik\_Evre (Ordinal/Categorical):** Clinical cT-Stage determined by pre-treatment examinations (cT1c < cT2a < cT2b < cT2c < cT3a < cT3b for increasing extent of tumor invasion)

**Biyopsi\_Gleason (Ordinal/Categorical):** Biopsy Gleason Score (3+3 < 3+4 < 4+3 < 3+5 < 4+4 < 4+5 < 5+4 < 5+5, higher score indicates higher aggressiveness)

**Risk\_Grubu (Ordinal/Categorical):** Risk Group Classification (1 for Low, 2 for Intermediate, 3 for High)

### Risk Factors

**Albumin (Continuous):** A Serum albumin level (g/dL). Indicator of nutritional status and systemic health

**Lenfosit (Discrete):** Lymphocyte (Immune system component) Count

**CRP (Continuous):** C-Reactive Protein (mg/L). Indicator of inflammation

**NLR (Continuous):** Neutrophil-to-Lymphocyte Ratio. A prognostic indicator for systemic inflammation and cancer aggressiveness.

**CALLY\_Index (Continuous):** CALLY Index. A composite index, likely related to inflammation or blood components.

**Komorbidite\_Skor (Ordinal/Categorical):** Comorbidity Score indicating the severity of other co-existing chronic diseases ( 0 (No comorbidities) < ... < 5 (Severe comorbidities))

### Treatment Information

**Tedavi\_Tipi (Categorical):** Main Treatment Type received (1 for Radical Prostatectomy, 2 for Radiotherapy/RT, 3 for Hormone Therapy, 4 for Combination of Radiotherapy and Hormone Therapy)

**Tedavi\_Tarihi (Date):** Treatment Date

**RT\_Dozu (Continuous):** Total Radiation Dose (in Gy), if radiotherapy was performed

**ADT\_Tipi (Categorical):** Androgen Deprivation Therapy (ADT, hormone therapy) Type used

**ADT\_Suresi (Continuous):** ADT(hormone therapy) Duration

### Pathological Markers

**Patolojik\_Evre (Ordinal/Categorical):** Final Tumor Pathological Stage determined after surgery on the removed tissue (pT2a < pT2b < pT2c < pT3a < pT3b < pT4, NaN indicates patient did not undergo surgery)

**Cerrahi\_Sinir (Binary/Categorical):** Surgical Margin Status indicating if cancer cells were present at the edge of the removed tissue. Crucial for recurrence prediction (0: Negative, 1: Positive, NaN indicates patient did not undergo surgery).

**Final\_Gleason (Ordinal/Categorical):** Final Gleason Score confirmed from the final excised tissue (3+3 < 3+4 < 4+3 < 3+5 < 4+4 < 4+5 < 5+4 < 5+5, higher score indicates higher aggressiveness)

## Follow-up & Outcomes

**PSA\_Nadir (Continuous):** The lowest PSA level reached after treatment (ng/mL). A lower nadir generally indicates better treatment success

**PSA\_Takip\_3ay / 6ay / 12ay (Continuous):** Follow-up PSA levels (ng/mL) measured at at 3/6/12 Months

**BCR\_Durum (Binary/Categorical):** Biochemical Recurrence (BCR) Status whether the PSA level rise above a recurrence threshold? (True for Recurrence occurred, False for no Recurrence occurred)

**BCR\_Tarihi (Date):** Date when biochemical recurrence was confirmed

**Metastaz\_Durum (Binary/Categorical):** Metastasis Status whether distant metastasis occur during follow-up? (0 for No, 1 for Yes)

**Metastaz\_Tarihi (Date):** Date when metastasis was confirmed

**Son\_Durum (Binary/Categorical):** Patient's Survival Status at the last follow-up (0 for Alive, 1 for Deceased)

**Son\_Takip\_Tarihi (Date):** Date of the last recorded patient information.

## Data handling 1: Remove variables that we will not use for analysis

```
cols1 <- c("Tani_Tarihi", "Tedavi_Tarihi", "PSA_Takip_3ay", "PSA_Takip_6ay", "PSA_Takip_12ay", "BCR_Tar")
data <- data[, !(names(data) %in% cols1)]
```

## Data handling 2: Turning all categorical variables to factors

```
cols2 <- c("Klinik_Evre", "Biyopsi_Gleason", "Risk_Grubu", "Komorbidite_Skor", "Tedavi_Tipi", "ADT_Tipi")
data[cols2] <- lapply(data[cols2], as.factor)
```

```
str(data)
```

```
## 'data.frame':    600 obs. of  23 variables:
## $ Hasta_ID      : int  0 1 2 3 4 5 6 7 8 9 ...
## $ Yas           : int  59 69 66 59 70 73 72 72 63 57 ...
## $ PSA_Tani      : num  41.3 12.9 32 142.7 16.2 ...
## $ Klinik_Evre   : Factor w/ 6 levels "cT1c","cT2a",...: 5 3 6 6 3 6 6 6 5 6 ...
## $ Biyopsi_Gleason : Factor w/ 9 levels "3+3","3+4","3+5",...: 3 4 3 5 2 8 7 5 6 5 ...
## $ Risk_Grubu     : Factor w/ 3 levels "1","2","3": 3 2 3 3 2 3 3 3 3 3 ...
## $ Albumin       : num  4.1 4.6 4.1 3.5 4.3 4.2 4.8 4.2 4 4.5 ...
## $ Lenfosit      : int  2227 1168 1125 1623 1399 2103 2038 1418 1936 1348 ...
## $ CRP           : num  0.58 0.12 0.16 1.76 0.53 0.1 0.46 0.27 0.11 0.62 ...
## $ NLR           : num  1.7 1.7 3.5 1.7 2.3 2.2 4 2.8 2.3 1.8 ...
## $ CALLY_Index    : num  1.6 4.5 2.9 0.3 1.1 8.8 2.1 2.2 7 1 ...
## $ Komorbidite_Skor: Factor w/ 6 levels "0","1","2","3",...: 3 1 2 1 3 5 1 3 4 6 ...
## $ Tedavi_Tipi    : Factor w/ 4 levels "1","2","3","4": 1 1 4 2 2 2 2 1 1 3 ...
## $ RT_Dozu       : num  NA NA 70 74 70 70 76 NA NA NA ...
## $ ADT_Tipi       : Factor w/ 3 levels "1","2","3": NA NA 1 NA NA NA NA NA NA 2 ...
```

```
## $ ADT_Suresi      : num  NA NA 12 NA NA NA NA NA 12 ...
## $ Patolojik_Evre  : Factor w/ 7 levels "", "pT2a", "pT2b",...: 4 2 1 1 1 1 2 6 1 ...
## $ Cerrahi_Sinir   : Factor w/ 2 levels "0", "1": 1 2 NA NA NA NA NA 2 2 NA ...
## $ Final_Gleason   : Factor w/ 10 levels "", "3+3", "3+4",...: 4 7 1 1 1 1 1 7 2 1 ...
## $ PSA_Nadir       : num  0.14 0.04 0.3 0.2 0.28 0.81 0.36 0.04 0.04 0.39 ...
## $ BCR_Durum       : Factor w/ 2 levels "False", "True": 2 1 1 1 1 2 1 1 1 1 ...
## $ Metastaz_Durum  : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
## $ Son_Durum       : Factor w/ 2 levels "0", "1": 2 2 1 2 2 2 2 2 2 2 ...
```

## Data Visualization 1: Frequency of Treatment Type (Tedavi\_Tipi)

```
sum(is.na(data$Tedavi_Tipi))
```

```
## [1] 0
```

```
table(data$Tedavi_Tipi)
```

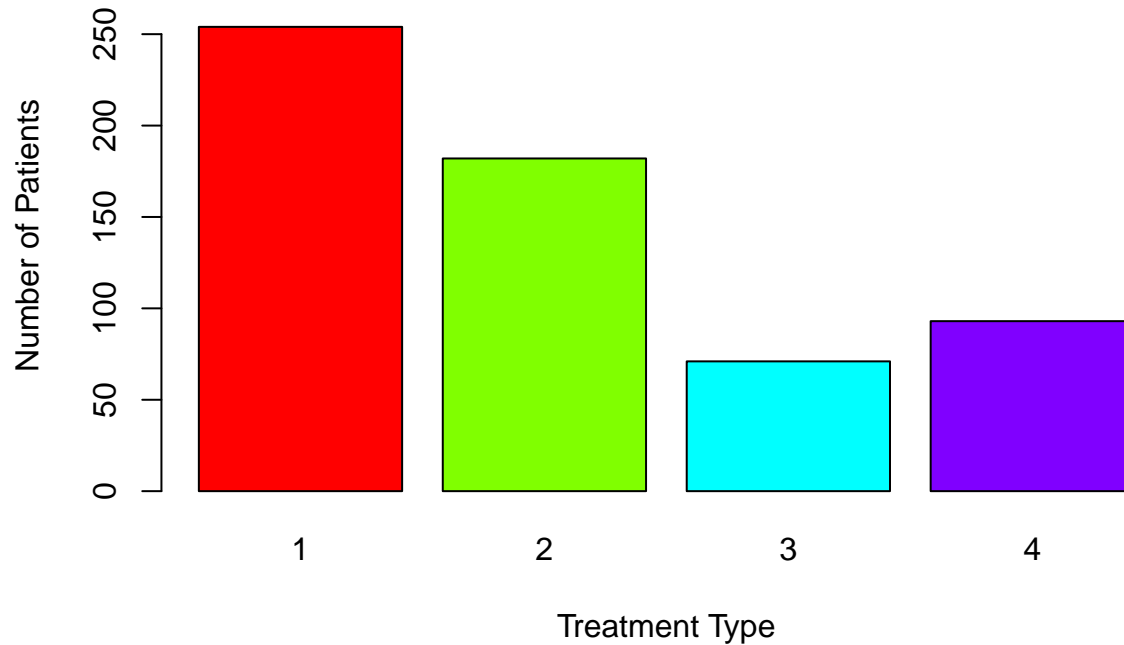
```
##
##  1  2  3  4
## 254 182 71 93
```

```
round(prop.table(table(data$Tedavi_Tipi)), 3)
```

```
##
##      1      2      3      4
## 0.423 0.303 0.118 0.155
```

```
frequencies <- table(data$Tedavi_Tipi)
barplot(
  frequencies,
  main = "Frequency of Treatment Type (Tedavi_Tipi)",
  xlab = "Treatment Type",
  ylab = "Number of Patients",
  col = rainbow(4),
  ylim = c(0, max(frequencies) * 1.1)
)
```

## Frequency of Treatment Type (Tedavi\_Tipi)



## Data Visualization 2: Frequency of Biochemical Recurrence Status (BCR\_Durum)

```
sum(is.na(data$BCR_Durum))
```

```
## [1] 0
```

```
table(data$BCR_Durum)
```

```
##  
## False  True  
##   416   184
```

```
round(prop.table(table(data$BCR_Durum)), 3)
```

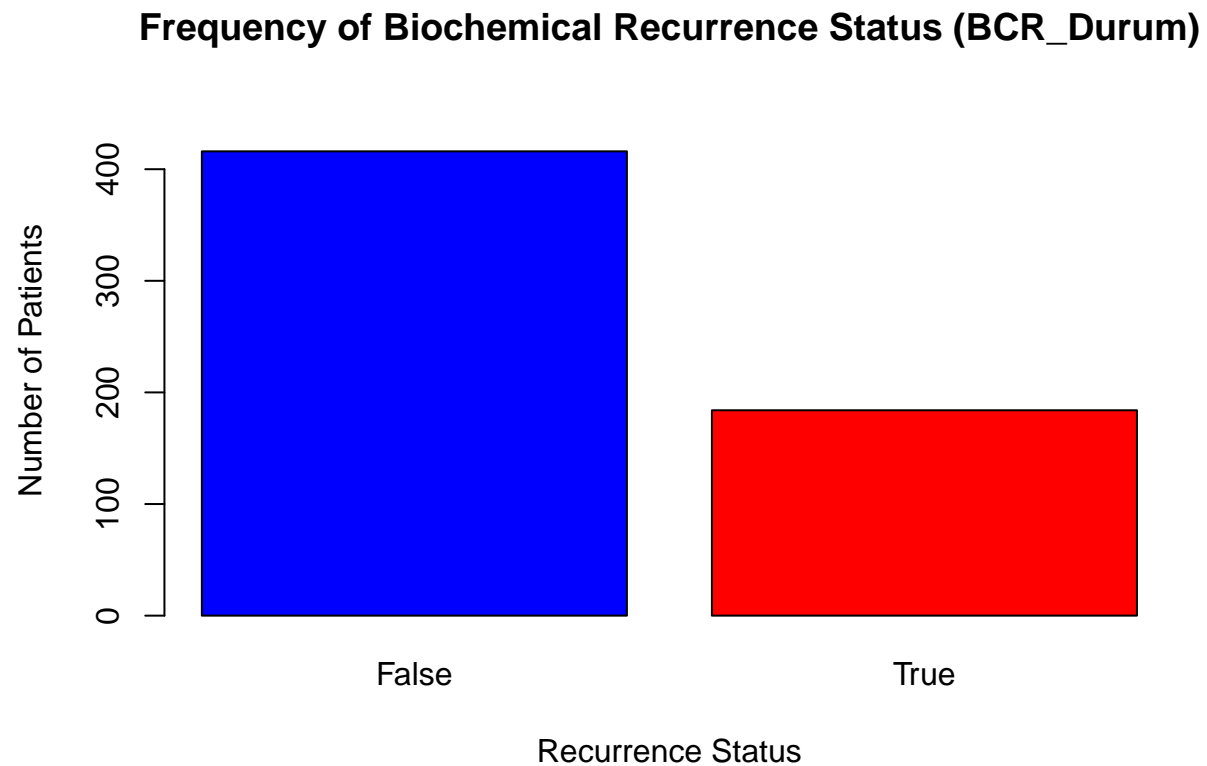
```
##  
## False  True  
## 0.693 0.307
```

```
frequencies <- table(data$BCR_Durum)  
barplot(  
  frequencies,
```

```

main = "Frequency of Biochemical Recurrence Status (BCR_Durum)",
xlab = "Recurrence Status",
ylab = "Number of Patients",
col = c("blue", "red"),
ylim = c(0, max(frequencies) * 1.1)
)

```



### Data Visualization 3: Frequency of Survival Status (Son\_Durum)

```
sum(is.na(data$Son_Durum))
```

```
## [1] 0
```

```
table(data$Son_Durum)
```

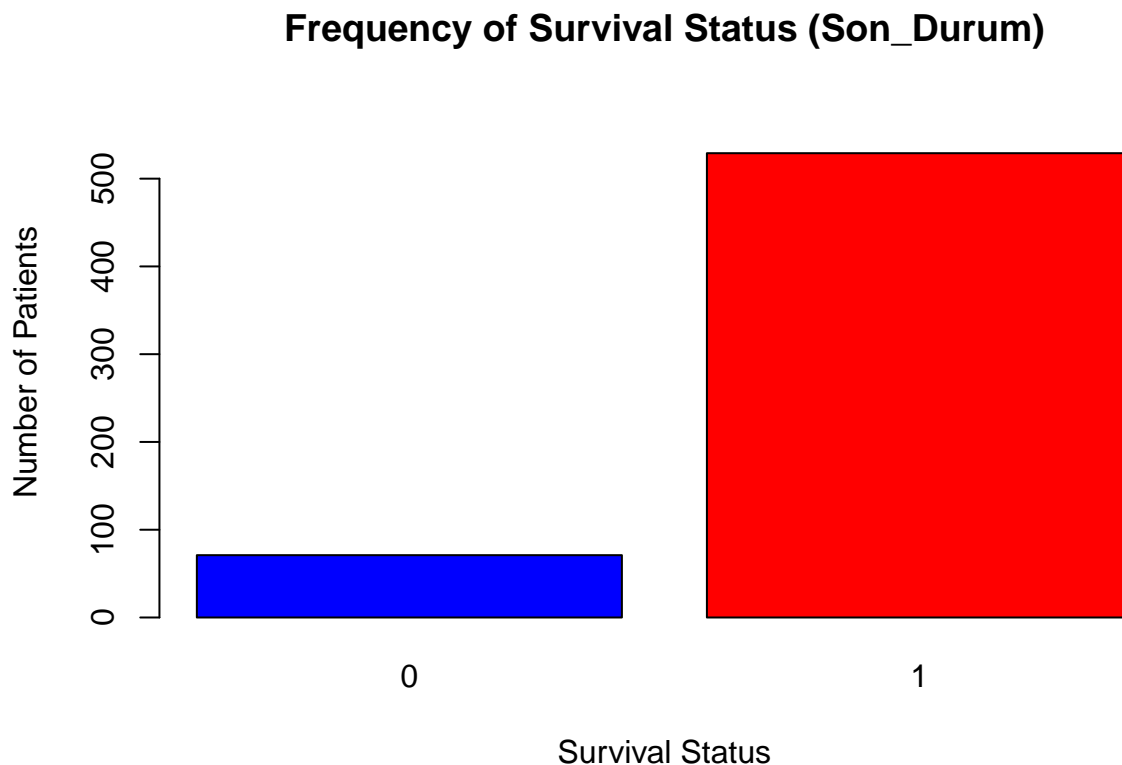
```
##
##  0  1
## 71 529
```

```
round(prop.table(table(data$Son_Durum)), 3)
```



```
##
##      0      1
## 0.118 0.882
```

```
frequencies <- table(data$Son_Durum)
barplot(
  frequencies,
  main = "Frequency of Survival Status (Son_Durum)",
  xlab = "Survival Status",
  ylab = "Number of Patients",
  col = c("blue", "red"),
  ylim = c(0, max(frequencies) * 1.1)
)
```



#### Data Visualization 4: Serum Prostate-Specific Antigen (PSA) level at diagnosis & after treatment

```
PSA <- c("PSA_Tani", "PSA_Nadir")
sapply(data[, PSA], function(x) sum(is.na(x)))
```

```
## PSA_Tani PSA_Nadir
##      0      0
```

```
summary(data[, PSA])
```

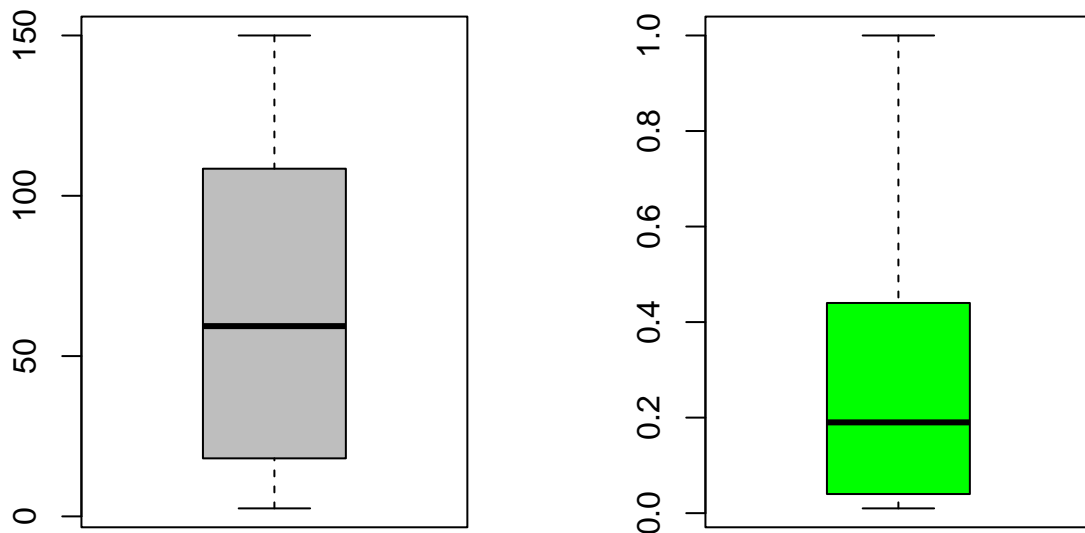
```
##      PSA_Tani      PSA_Nadir  
## Min.   : 2.50   Min.   :0.0100  
## 1st Qu.: 18.10   1st Qu.:0.0400  
## Median : 59.35   Median :0.1900  
## Mean   : 64.70   Mean    :0.2847  
## 3rd Qu.:108.42   3rd Qu.:0.4400  
## Max.   :150.00   Max.    :1.0000
```

```
par(mfrow = c(1, 2))
```

```
boxplot(data[["PSA_Tani"]],  
        main = "Box Plot of PSA level at diagnosis",  
        col = "grey")
```

```
boxplot(data[["PSA_Nadir"]],  
        main = "Box Plot of PSA level after treatment",  
        col = "green")
```

**Box Plot of PSA level at diagnosis    Box Plot of PSA level after treatment**



**Data Visualization 5: Summary & Boxplots of Risk Factors**

```
risk_factors <- c("Albumin", "Lenfosit", "CRP", "NLR", "CALLY_Index", "Komorbidite_Skor")
sapply(data[, risk_factors], function(x) sum(is.na(x)))
```

```
##           Albumin           Lenfosit           CRP           NLR
##           0           0           0           0
##      CALLY_Index Komorbidite_Skor
##           0           0
```

```
summary(data[, risk_factors])
```

```
##      Albumin           Lenfosit           CRP           NLR
##  Min.   :3.500   Min.   :1000   Min.   :0.100   Min.   :1.000
## 1st Qu.:4.000   1st Qu.:1561   1st Qu.:0.130   1st Qu.:2.000
## Median :4.200   Median :1834   Median :0.345   Median :2.500
## Mean   :4.219   Mean   :1827   Mean   :0.518   Mean   :2.503
## 3rd Qu.:4.500   3rd Qu.:2093   3rd Qu.:0.690   3rd Qu.:3.000
## Max.   :5.000   Max.   :2936   Max.   :3.590   Max.   :5.000
##      CALLY_Index      Komorbidite_Skor
##  Min.   : 0.200      0:100
## 1st Qu.: 1.100      1:101
## Median : 2.100      2: 87
## Mean   : 3.346      3:102
## 3rd Qu.: 5.425      4:122
## Max.   :13.000      5: 88
```

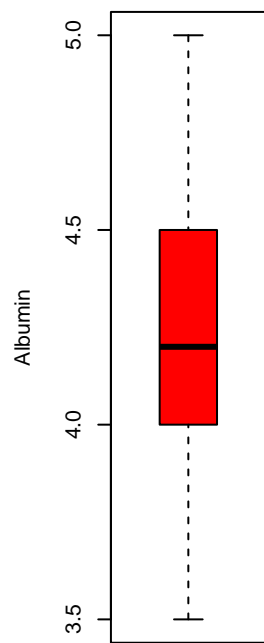
```
continuous_risks <- c("Albumin", "CRP", "NLR", "CALLY_Index")
plot_colors <- rainbow(4)
```

```
par(mfrow = c(1, 4))
```

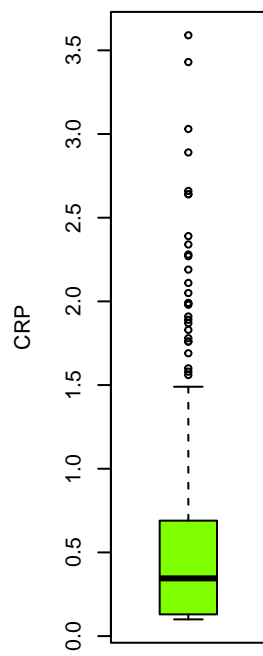
```
for (i in 1:length(continuous_risks)) {
  var_name <- continuous_risks[i]
  var_data <- data[[var_name]]

  boxplot(var_data,
    main = paste("Box Plot of", var_name),
    ylab = var_name,
    col = plot_colors[i])
}
```

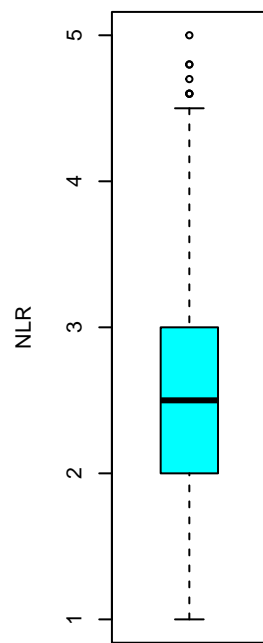
**Box Plot of Albumin**



**Box Plot of CRP**



**Box Plot of NLR**



**Box Plot of CALLY\_Ind**

