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
```

```r
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" ...
```

```r
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 forHigh)

Risk Factors

Albumin (Continuous): ASerum 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)
```

Data handling 3: Changing variable names

```
names(data) <- c("Patient ID", "Age", "PSA_before", "CTstage", "GleasonScore_before", "RiskClass", "Albu
```

```
str(data)
```

```
## 'data.frame':    600 obs. of  23 variables:
##  $ Patient ID         : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Age                : int  59 69 66 59 70 73 72 72 63 57 ...
##  $ PSA_before         : num  41.3 12.9 32 142.7 16.2 ...
##  $ CTstage            : Factor w/ 6 levels "cT1c","cT2a",..: 5 3 6 6 3 6 6 6 5 6 ...
##  $ GleasonScore_before: Factor w/ 9 levels "3+3","3+4","3+5",..: 3 4 3 5 2 8 7 5 6 5 ...
##  $ RiskClass          : 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 ...
##  $ Lymphocyte         : 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 ...
##  $ CallyIndex         : num  1.6 4.5 2.9 0.3 1.1 8.8 2.1 2.2 7 1 ...
##  $ ComorbidityScore   : Factor w/ 6 levels "0","1","2","3",..: 3 1 2 1 3 5 1 3 4 6 ...
##  $ Treatment          : Factor w/ 4 levels "1","2","3","4": 1 1 4 2 2 2 2 1 1 3 ...
##  $ RadiationDose      : num  NA NA 70 74 70 70 76 NA NA NA ...
##  $ HormoneType        : Factor w/ 3 levels "1","2","3": NA NA 1 NA NA NA NA NA NA 2 ...
##  $ HormonDuration     : num  NA NA 12 NA NA NA NA NA NA 12 ...
##  $ TumorSize          : Factor w/ 7 levels "","pT2a","pT2b",..: 4 2 1 1 1 1 1 2 6 1 ...
##  $ MarginStatus       : Factor w/ 2 levels "0","1": 1 2 NA NA NA NA NA 2 2 NA ...
##  $ GleasonScore_after : Factor w/ 10 levels "","3+3","3+4",..: 4 7 1 1 1 1 1 7 2 1 ...
##  $ PSA_after          : num  0.14 0.04 0.3 0.2 0.28 0.81 0.36 0.04 0.04 0.39 ...
##  $ BCR                : Factor w/ 2 levels "False","True": 2 1 1 1 1 2 1 1 1 1 ...
##  $ Metastasis         : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Survival           : Factor w/ 2 levels "0","1": 2 2 1 2 2 2 2 2 2 2 ...
```

Data Analysis 1: Chi-squared Test, Treatment vs BCR

Hypotheses

H0: Treatment Type and Biochemical Recurrence (BCR) Status are independent. (The recurrence rate is the same across all treatment groups)

Ha: Treatment Type and Biochemical Recurrence (BCR) Status are not independent. (The recurrence rate is significantly different for at least one treatment group)

```
bcr_table <- table(data$Treatment, data$BCR)
cat("Frequency Table (Counts):\n")
```

```
## Frequency Table (Counts):
```

```
print(bcr_table)
```

```
##
##     False True
##   1   189   65
##   2   123   59
##   3    40   31
##   4    64   29
```

```
bcr_test <- chisq.test(bcr_table)
cat("\nChi-squared Test Results:\n")
```

```
##
## Chi-squared Test Results:
```

```
print(bcr_test)
```

```
##
##  Pearson's Chi-squared test
##
## data:  bcr_table
## X-squared = 8.9915, df = 3, p-value = 0.0294
```

```
cat("\nBCR Rate (%) within each Treatment Group:\n")
```

```
##
## BCR Rate (%) within each Treatment Group:
```

```
round(prop.table(bcr_table, margin = 1) * 100, 1)
```

```
##
##     False True
##   1  74.4 25.6
##   2  67.6 32.4
##   3  56.3 43.7
##   4  68.8 31.2
```

Conclusion: Since the p-value (0.0294) is less than the significance level (a=0.05), we reject the null hypothesis (H0).

Interpretation: There is a statistically significant association between the type of treatment a patient receives and the likelihood of experiencing Biochemical Recurrence (BCR).

Highest Recurrence Rate: Treatment Type 3 (Hormone Therapy (ADT) Monotherapy) showed the highest biochemical recurrence rate at 43.7%.

Lowest Recurrence Rate: Treatment Type 1 (Radical Prostatectomy) showed the lowest recurrence rate at 25.6%.

Data Analysis 2: Chi-squared Test, Treatment vs Survival

Hypotheses

H0: Treatment Type and Survival Status are independent. (The survival rate is the same across all treatment groups)

Ha: Treatment Type and Survival Status are not independent. (The survival rate is significantly different for at least one treatment group)

```
survival_table <- table(data$Treatment, data$Survival)
cat("Frequency Table (Counts):\n")
```

```
## Frequency Table (Counts):
```

```
print(survival_table)
```

```
##
##       0   1
##   1  35 219
##   2  24 158
##   3   6  65
##   4   6  87
```

```r
survival_test <- chisq.test(survival_table)
cat("\nChi-squared Test Results:\n")
```

```
##
## Chi-squared Test Results:
```

```r
print(survival_test)
```

```
##
##  Pearson's Chi-squared test
##
## data:  survival_table
## X-squared = 4.6021, df = 3, p-value = 0.2034
```

```r
cat("\nSurvival Rate (%) within each Treatment Group (Survival = Alive):\n")
```

```
##
## Survival Rate (%) within each Treatment Group (Survival = Alive):
```

```r
round(prop.table(survival_table, margin = 1) * 100, 1)
```

```
##
##          0    1
##   1 13.8 86.2
##   2 13.2 86.8
##   3  8.5 91.5
##   4  6.5 93.5
```

Conclusion: Since the p-value (0.2034) is greater than the significance level (a=0.05), we fail to reject the null hypothesis (H0).

Interpretation: There is no statistically significant association between the type of treatment a patient receives and the likelihood of their Survival Status (Alive vs Deceased).

Data Analysis 3: ANOVA Test, Treatment vs PSA Difference

Hypotheses

H0: The mean change in PSA (Delta PSA) is the same across all four Treatment Type groups

Ha: The mean change in PSA (Delta PSA) is significantly different for at least one treatment group

```r
data$Delta_PSA <- data$PSA_before - data$PSA_after

anova_result <- aov(Delta_PSA ~ Treatment, data = data)
summary(anova_result)
```

```
##              Df  Sum Sq Mean Sq F value   Pr(>F)
## Treatment     3   60755   20252   9.774 2.66e-06 ***
## Residuals   596 1234905    2072
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(anova_result)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Delta_PSA ~ Treatment, data = data)
##
## $Treatment
##          diff        lwr       upr     p adj
## 2-1  14.59788   3.209034 25.986728 0.0055972
## 3-1  29.75452  14.011616 45.497414 0.0000085
## 4-1  16.52282   2.309503 30.736139 0.0151264
## 3-2  15.15663  -1.252454 31.565723 0.0821099
## 4-2   1.92494 -13.022893 16.872773 0.9873986
## 4-3 -13.23169 -31.713328  5.249938 0.2535812
```

Conclusion: Since the p-value (2.66e-06) is less than the significance level (a=0.05), we reject the null hypothesis (H0).

Interpretation: Interpretation: There is a statistically significant difference in the mean change in PSA (Delta PSA) among the different treatment groups.

Tukey's HSD Interpretation: The mean PSA change (Delta PSA, where a larger value indicates a greater reduction) in the Radical Prostatectomy (Treatment 1) group is significantly smaller than the mean change in all other treatment groups (2, 3, and 4).

Largest Mean Difference: Treatment 3 (Hormone Therapy Monotherapy) showed the largest positive mean difference (Delta PSA = 29.755) compared to Treatment 1, indicating the greatest average PSA reduction after treatment.

No Significant Difference: There is no significant difference in the mean PSA Defference among the three non-surgical groups (Treatment 2, 3, and 4).