

BURSA TEKNİK ÜNİVERSİTESİ
MÜHENDİSLİK VE DOĞA BİLİMLERİ FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ
SEMİNER DERSİ PROJESİ

R Dili ile Temel İstatistikler ve Problemler:
Veri Analizi ve Çözümleme

Adile AKKILIÇ

2023-2024 BAHAR DÖNEMİ

ÖNSÖZ

Bu rapor, "R Dili ile Temel İstatistikler ve Problemler: Veri Analizi ve Çözümleme" başlığı altında sunulmaktadır. R programlama dilini kullanarak temel istatistiksel işlemler ve problemleri ele alan bu rapor, veri analizi ve çözümleme süreçlerine odaklanmaktadır.

Raporun ilk bölümünde, R programlama dilinin tarihçesi ve temel özellikleri kısaca incelenmiştir. Ardından, R'de kullanılan veri yapıları ve temel işlemler hakkında bilgi verilmiştir. Temel istatistiksel işlemler bölümünde, veri aktarma, temel veri inceleme işlemleri, merkezi eğilim ve dağılım ölçüleri, olasılık dağılımları ve temel istatistiksel testler ele alınmıştır.

Raporun son bölümü, örnek problemler ve çözümleri ile veri görselleştirme ve analiz araçlarına odaklanmaktadır. Bu bölüm, okuyucuların R programlama dilini kullanarak temel istatistiksel problemleri çözme yeteneklerini geliştirmelerine yardımcı olmayı amaçlamaktadır.

Adile AKKILIÇ

Bursa 2024

İÇİNDEKİLER

ÖNSÖZ.....	2
İÇİNDEKİLER.....	3
ÖZET.....	4
1. R PROGRAMLAMA DİLİNE GENEL BAKIŞ.....	5
1.1. R Programlama Dilinin Tarihi ve Özellikleri.....	5
1.2. Veri Yapıları ve Temel İşlemler.....	6
2. TEMEL İSTATİSTİKSEL İŞLEMLER.....	9
2.1. Veri İçerme Aktarma ve Temel Veri İnceleme İşlemleri	9
2.2. Merkezi Eğilim ve Dağılım Ölçüleri	11
2.3. Olasılık Dağılımları ve Temel İstatistiksel Testler	13
3. RStudio Geliştirme Ortamı	16
3.1. Rstudio Kurulumu.....	16
3.2. İlk Bakış.....	18
3.3. RMarkdown Kullanımı.....	19
4. UYGULAMA.....	22
4.1. Proje Oluşturma.....	22
4.2. Tasarımını Oluşturma.....	24
4.3. Kodlama.....	25
5. SONUÇ.....	34
6. KAYNAKLAR.....	35

ÖZET

Bu rapor, "R Dili ile Temel İstatistikler ve Problemler: Veri Analizi ve Çözümleme" başlığı altında sunulan bir çalışmayı içermektedir. R programlama dilini kullanarak temel istatistiksel işlemler ve problemleri ele almaktadır.

Raporun içeriği, R dilinin temel özelliklerinden başlayarak veri analizi süreçlerini kapsamaktadır. Rapor, okuyucuların R dilini kullanarak temel istatistiksel problemleri anlamalarına ve çözmelerine yardımcı olmayı amaçlamaktadır.

1. R PROGRAMLAMA DİLİNE GENEL BAKIŞ

1.1. R Programlama Dilinin Tarihi ve Özellikleri

R programlama dili, istatistiksel hesaplama ve grafiksel görselleştirme için kullanılan açık kaynaklı bir yazılım ortamıdır. 1993 yılında Ross Ihaka ve Robert Gentleman tarafından geliştirilmeye başlanmıştır. R, S programlama dilinin bir türevidir ve GNU Genel Kamu Lisansı altında dağıtılmaktadır.

R dilinin bazı temel özellikleri şunlardır:

Açık Kaynaklı ve Ücretsiz: R, GNU Genel Kamu Lisansı altında serbestçe dağıtılmaktadır.

Geniş Paket Desteği: CRAN (Comprehensive R Archive Network) üzerinden erişilebilen binlerce paket ile zenginleştirilebilir.

İstatistiksel Hesaplama ve Modelleme: Çok çeşitli istatistiksel ve grafiksel teknikleri destekler.

Çapraz Platform Desteği: Windows, MacOS ve Linux gibi birçok işletim sisteminde çalışabilir.

Topluluk ve Destek: Aktif bir kullanıcı topluluğu ve geniş bir belge arşivi bulunmaktadır.

1.2. Veri Yapıları ve Temel İşlemler

R programlama dilinde veriler çeşitli veri yapıları ile temsil edilir. Bu veri yapılarından bazıları şunlardır:

- **Vektörler:** Aynı türdeki elemanlardan oluşan tek boyutlu veri yapılarıdır. Vektörler, en temel veri yapılarıdır ve çeşitli veri türlerini (numerik, karakter, mantıksal) içerebilir.Örneğin:

```
> numeric_vector <- c(1, 2, 3, 4)
> character_vector <- c("A", "B", "C")
> logical_vector <- c(TRUE, FALSE, TRUE)
> print(logical_vector)
[1] TRUE FALSE TRUE
> |
```

Şekil 1: Vektörler

- **Matrisler:** İki boyutlu, aynı türdeki elemanlardan oluşan veri yapılarıdır. Matrisler, sütun ve satırlardan oluşur ve genellikle matematiksel ve istatistiksel işlemler için kullanılır. Örneğin:

```
> matrix_data <- matrix(1:9, nrow = 3, ncol = 3)
> print(matrix_data)
      [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
> |
```

Şekil 2:Matrisler

- **Data Frame'ler:** Farklı türdeki elemanlardan oluşan, iki boyutlu veri yapılarıdır. Data frame'ler, özellikle veri analizi ve veri manipülasyonu işlemlerinde yaygın olarak kullanılır. Her sütun aynı uzunlukta olmalıdır. Örneğin:

```
> df <- data.frame(
+   ID = c(1, 2, 3),
+   Name = c("John", "Jane", "Doe"),
+   Age = c(23, 25, 28)
+ )
> print(df)
  ID Name Age
1  1 John  23
2  2 Jane  25
3  3 Doe   28
> |
```

Şekil 3: Data Frame'ler

- **Liste:** Farklı türdeki elemanlardan oluşan ve boyutları farklı olabilen veri yapılarıdır. Listeler, karmaşık veri yapılarının saklanması için kullanılır ve içinde vektörler, matrisler, data frame'ler ve hatta başka listeler barındırabilir. Örneğin:

```
> my_list <- list(
+   Name = "Alice",
+   Age = 26,
+   Scores = c(85, 90, 95),
+   Details = data.frame(Subject = c("Math", "Science"), Grade = c("A", "B"))
+ )
> print(my_list)
$Name
[1] "Alice"

$Age
[1] 26

$Scores
[1] 85 90 95

$Details
  Subject Grade
1   Math     A
2 Science     B
```

Şekil 4: Liste

R programlama dilinde veriler çeşitli veri yapıları ile temsil edilir. Bu veri yapılarından bazıları şunlardır:

- **Vektör İşlemleri:**

```
> vec <- c(10, 20, 30)
> sum_vec <- sum(vec) # Toplam
> mean_vec <- mean(vec) # Ortalama
> print(sum_vec)
[1] 60
> print(mean_vec)
[1] 20
```

Şekil 5: Vektör İşlemleri

- **Matris İşlemleri:**

```
> matrix_data <- matrix(1:9, nrow = 3, ncol = 3)
> transpose_matrix <- t(matrix_data) # Transpoz alma
> print(matrix_data)
      [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
> print(transpose_matrix)
      [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
```

Şekil 6: Matris İşlemleri

- **Data Frame İşlemleri:**

```
> df <- data.frame(
+   ID = c(1, 2, 3),
+   Name = c("John", "Jane", "Doe"),
+   Age = c(23, 25, 28)
+ )
> df$Name # "Name" sütununu seçme
[1] "John" "Jane" "Doe"
> summary(df) # Özet istatistikler
```

	ID	Name	Age
Min.	:1.0	Length:3	Min. :23.00
1st Qu.:	:1.5	Class :character	1st Qu.:24.00
Median :	:2.0	Mode :character	Median :25.00
Mean :	:2.0		Mean :25.33
3rd Qu.:	:2.5		3rd Qu.:26.50
Max. :	:3.0		Max. :28.00

Şekil 7: Data Frame İşlemleri

- **Liste İşlemleri:**

```
> my_list <- list(  
+   Name = "Alice",  
+   Age = 26,  
+   Scores = c(85, 90, 95)  
+ )  
> my_list$Name # Listenin "Name" elemanını seçme  
[1] "Alice"  
_ |
```

Şekil 8: Liste İşlemleri

Bu temel veri yapıları ve işlemler, R programlama dilinin güçlü yönlerinden sadece birkaçıdır ve veri analizinde geniş bir uygulama yelpazesine sahiptir.

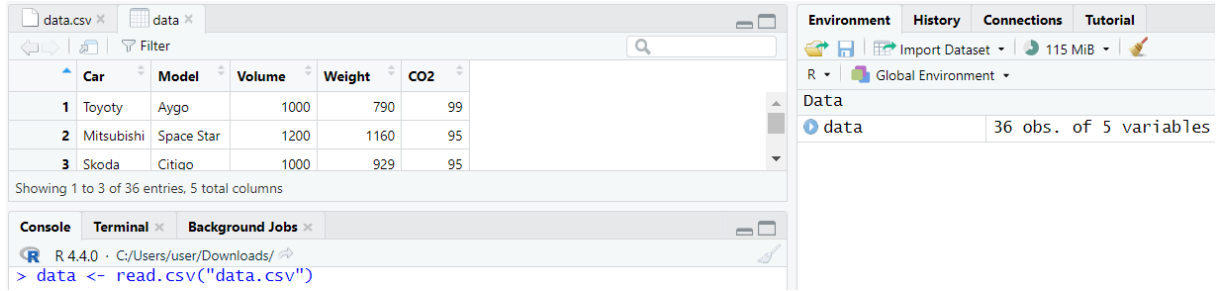
2. TEMEL İSTATİSTİKSEL İŞLEMLER

2.1. Veri İçe Aktarma ve Temel Veri İnceleme İşlemleri

R programlama dili, istatistiksel hesaplama ve grafiksel görselleştirme için kullanılan açık kaynaklı bir yazılım ortamıdır. 1993 yılında Ross Ihaka ve Robert Gentleman tarafından geliştirilmeye başlanmıştır. R, S programlama dilinin bir türevidir ve GNU Genel Kamu Lisansı altında dağıtılmaktadır.

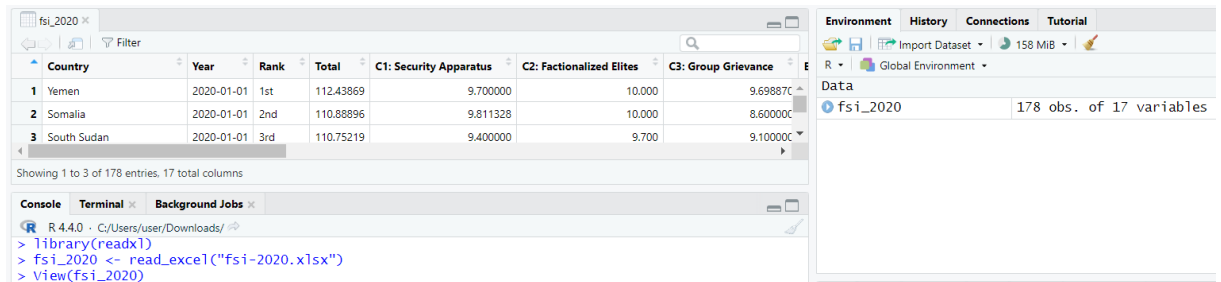
Veri İçe Aktarma:

- **CSV Dosyası İçe Aktarma:** CSV (Comma-Separated Values) dosyaları, veri setlerinin saklanması ve paylaşılması için yaygın bir formattır. `read.csv()` işlevi kullanılarak CSV dosyaları R'ye aktarılabilir.



Şekil 9: CSV Dosyası İçe Aktarma

- **Excel Dosyası İçe Aktarma:** Excel dosyaları .xls ve .xlsx formatlarında olabilir. `readxl` paketi kullanılarak Excel dosyaları içe aktarılabilir.



Şekil 10: Excel Dosyası İçe Aktarma

Temel Veri İnceleme İşlemleri:

- **Veri Yapısının İncelenmesi:** Veri setinin genel yapısını görmek için `str()` işlevi kullanılır.

```
R 4.4.0 · C:/Users/user/Downloads/
> str(data)
spec_tbl_ [36 × 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Car   : chr [1:36] "Toyoty" "Mitsubishi" "Skoda" "Fiat" ...
 $ Model : chr [1:36] "Aygo" "Space Star" "Citigo" "500" ...
 $ Volume: num [1:36] 1000 1200 1000 900 1500 1000 1400 1500 1500 1600 ...
 $ Weight: num [1:36] 790 1160 929 865 1140 ...
 $ CO2   : num [1:36] 99 95 95 90 105 105 90 92 98 99 ...
- attr(*, "spec")=
.. cols(
..   Car = col_character(),
..   Model = col_character(),
..   Volume = col_double(),
..   Weight = col_double(),
..   CO2 = col_double()
.. )
- attr(*, "problems")=<externalptr>
>
```

Şekil 11: Veri Yapısının İncelenmesi

- **Özet İstatistikler:** Verinin özet istatistiklerini almak için `summary()` işlevi kullanılır.

```
> summary(data)
      Car           Model           Volume           Weight           CO2
Length:36      Length:36      Min.   : 900      Min.   : 790      Min.   : 90.00
Class :character Class :character 1st Qu.:1475      1st Qu.:1117      1st Qu.: 97.75
Mode  :character Mode  :character  Median :1600      Median :1329      Median : 99.00
                                Mean  :1611      Mean  :1292      Mean  :102.03
                                3rd Qu.:2000      3rd Qu.:1418      3rd Qu.:105.00
                                Max.   :2500      Max.   :1746      Max.   :120.00
>
```

Şekil 12: Özel İstatistikler

- **İlk ve Son Kayıtların Görüntülenmesi:** Veri setinin ilk birkaç kaydını görmek için `head()` ve son birkaç kaydını görmek için `tail()` işlevleri kullanılır.

```
> head(data)
# A tibble: 6 × 5
  Car      Model      Volume Weight   CO2
  <chr>    <chr>    <dbl>  <dbl> <dbl>
1 Toyoty   Aygo        1000    790    99
2 Mitsubishi Space Star  1200    1160   95
3 Skoda    Citigo      1000    929   95
4 Fiat     500         900    865   90
5 Mini     Cooper      1500    1140  105
6 VW       Up!        1000    929   105

> tail(data)
# A tibble: 6 × 5
  Car      Model      Volume Weight   CO2
  <chr>    <chr>    <dbl>  <dbl> <dbl>
1 Mercedes E-Class      2100    1605  115
2 Volvo    XC70       2000    1746  117
3 Ford     B-Max       1600    1235  104
4 BMW      216        1600    1390  108
5 Opel     Zafira       1600    1405  109
6 Mercedes SLK       2500    1395  120
>
```

Şekil 13: İlk ve son kayıtların görüntülenmesi

- **Veri Boyutlarının Kontrolü:** Veri setinin boyutlarını (satır ve sütun sayısı) öğrenmek için `dim()` işlevi kullanılır.

```
> dim(data)
[1] 36  5
```

Şekil 14: Veri Boyutların Kontrolü

- **Sütun İsimlerinin İncelenmesi:** Veri setindeki sütun isimlerini görmek için `colnames()` işlevi kullanılır.

```
> colnames(data)
[1] "Car"      "Model"    "Volume"   "Weight"   "CO2"
```

Şekil 15: Sütun İsimlerinin İncelenmesi

Bu işlemler, veri setinin genel yapısını ve içeriğini anlamak için ilk adımlar olup, veri analizi sürecinin temelini oluşturur.

2.2. Merkezi Eğilim ve Dağılım Ölçüleri

Merkezi eğilim ve dağılım ölçüleri, veri setinin temel özelliklerini anlamak için kullanılan istatistiksel ölçülerdir. Merkezi eğilim ölçüleri, verilerin ortalama bir değer etrafında nasıl toplandığını gösterirken, dağılım ölçüleri, verilerin bu ortalama değerden ne kadar uzaklaştığını gösterir.

Merkezi Eğilim Ölçüleri:

- **Ortalama(Mean):** Verilerin aritmetik ortalamasıdır. Tüm verilerin toplamının veri sayısına bölünmesiyle elde edilir.

```
> mean_value <- mean(data$CO2, na.rm = TRUE)
> print(mean_value)
[1] 102.0278
```

Şekil 16: Ortalama(Mean)

- **Medyan (Median):** Verilerin sıralanmış haliyle ortanca değeridir. Veri sayısı tek ise ortanca değer, çift ise ortadaki iki değer ortalaması alınır.

```
> median_value <- median(data$CO2, na.rm = TRUE)
> print(median_value)
[1] 99
```

Şekil 17: Medyan (Median)

- **Mod (Mode):** Veriler içinde en sık tekrarlanan değerdir. R'de mod hesaplamak için özel bir işlev yoktur, bu nedenle aşağıdaki gibi özel bir fonksiyon yazabilirsiniz:

```
> get_mode <- function(v) {  
+   uniqv <- unique(v)  
+   uniqv[which.max(tabulate(match(v, uniqv)))]  
+ }  
> mode_value <- get_mode(data$C02)  
>  
> print(mode_value)  
[1] 99
```

Şekil 18: Mod (Mode)

Dağılım Ölçüleri:

- **Varyans (Variance):** Verilerin ortalamadan ne kadar saptığını ölçer. Ortalama sapmanın karesi olarak hesaplanır.

```
> variance_value <- var(data$C02)  
>  
> print(variance_value)  
[1] 55.57063
```

Şekil 19: Varyans (Variance)

- **Standart Sapma (Standard Deviation):** Varyansın karekökü olarak hesaplanır. Verilerin ortalamadan ne kadar saptığını gösterir.

```
> sd_value <- sd(data$C02)  
>  
> print(sd_value)  
[1] 7.454571
```

Şekil 20: Standart Sapma (Standard Deviation)

- **Aralık (Range):** Verilerin en büyük değeri ile en küçük değeri arasındaki farktır.

```
> range_value <- range(data$C02)  
> range_diff <- diff(range_value)  
> print(range_diff)  
[1] 30
```

Şekil 21: Aralık (Range)

- **Çeyrekler Açıklığı (Interquartile Range, IQR):** Verilerin orta yüzde ellisini kapsayan çeyrekler açıklığıdır. Üçüncü çeyrek (Q3) ile birinci çeyrek (Q1) arasındaki fark olarak hesaplanır.

```
> iqr_value <- IQR(data$CO2)
> print(iqr_value)
[1] 7.25
```

Şekil 22: Çeyrekler Açıklığı (Interquartile Range , IQR)

2.3. Olasılık Dağılımları ve Temel İstatistiksel Testler

Bu bölümde, R programlama dilini kullanarak çeşitli olasılık dağılımlarını inceleme ve bazı temel istatistiksel testleri nasıl uygulayacağınızı ele alacağız.

Olasılık Dağılımları

Olasılık dağılımları, rastgele bir olayın olası sonuçlarının dağılımını tanımlar. R, çeşitli dağılımlar için fonksiyonlar sunar:

1. Normal Dağılım:

- **rnorm():** Normal dağılımından rastgele sayılar üretir.
- **dnorm():** Belirli bir değerin normal dağılımın yoğunluk fonksiyonundaki değerini verir.
- **pnorm():** Bir değerin normal dağılım fonksiyonunun kümülatif değerini hesaplar.
- **qnorm():** Kümülatif dağılım fonksiyonunun belirli bir yüzdesine karşılık gelen değeri verir.

```
> # Normal dağılımdan 10 rastgele sayı üret
> random_numbers <- rnorm(10, mean = 0, sd = 1)
> print(random_numbers)
[1] -0.81171764 -0.77276743  0.01065366 -0.82375021  1.28616310 -0.42468699 -0.57396844
[8]  0.10713931  0.10623576 -0.09567683
```

Şekil 23: Random dağılım ile sayı üretme

2. Binom Dağılımı:

- **rbinom():** Binom dağılımından rastgele sayılar üretir.
- **dbinom():** Binom dağılımdaki bir değerin olasılığını hesaplar.
- **pbinom():** Binom dağılımında bir değerin kümülatif olasılığını hesaplar.

```
> # 5 denemede 3 başarı olasılığı (başarı olasılığı 0.5 olan binom dağılımı)
> prob_success <- dbinom(3, size = 5, prob = 0.5)
> print(prob_success)
[1] 0.3125
```

Şekil 24: dbinom

Temel İstatistiksel Testler

İstatistiksel testler, veriler üzerinden çıkarımlar yapmamızı sağlar. R'da sık kullanılan bazı temel testler:

1. t-Testi:

- Bir örneğin ortalamasının belirli bir değerden farklı olup olmadığını test eder.

```
> # Tek örnekli t-test
> t.test(data$CO2, mu = 20)

One Sample t-test

data: data$CO2
t = 66.022, df = 35, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 20
95 percent confidence interval:
 99.50551 104.55004
sample estimates:
mean of x
102.0278
```

Şekil 25: t-test

2. ANOVA (Varyans Analizi):

- Birden fazla grubun ortalamalarını karşılaştırır.

```
> # ANOVA testi
> anova_result <- aov(data$CO2 ~ data$Volume, data = data)
> summary(anova_result)
              Df Sum Sq Mean Sq F value    Pr(>F)    
data$Volume   1   681.8    681.8   18.35 0.000142 ***
Residuals    34  1263.1     37.2             
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Şekil 26: ANOVA

3. Korelasyon Testi:

- İki değişken arasındaki ilişkinin gücünü ve yönünü ölçer.

```
> # Pearson korelasyon testi
> cor_test <- cor.test(data$CO2, data$weight, method = "pearson")
> print(cor_test)

Pearson's product-moment correlation

data: data$CO2 and data$weight
t = 3.8616, df = 34, p-value = 0.0004806
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2731671 0.7454588
sample estimates:
      cor
0.55215
```

Şekil 27: Korelasyon Testi

Bu araçlar ve testler, veri analizi süreçlerinde temel rol oynar ve R dilinin güçlü istatistiksel yeteneklerini sergiler. Her bir fonksiyonun ve testin parametrelerini ve kullanımını, kendi veri setlerinize ve analiz ihtiyaçlarınıza göre ayarlamak önemlidir.

3. RStudio Geliştirme Ortamı

RStudio, R programlama dili için kullanılan popüler bir entegre geliştirme ortamıdır (IDE). Veri analizi, veri görselleştirme ve istatistiksel hesaplamalar gibi işlemler için güçlü araçlar sunar. RStudio, Windows, macOS ve Linux gibi çeşitli işletim sistemlerinde kullanılabilir.

RStudio, bir dizi özellik ve araçla birlikte gelir:

Script Editor: R kodlarını yazmak ve düzenlemek için kullanılır.

Console: R komutlarını doğrudan çalıştırmak için kullanılır.

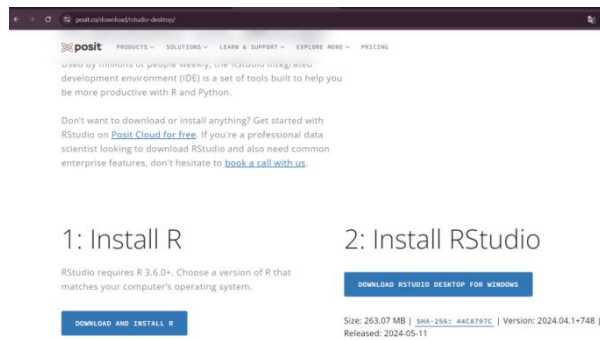
Environment/History: Çalışma alanını ve geçmiş komutları görüntüler.

Files/Plots/Packages/Help/Viewer: Dosyaları yönetmek, grafikler oluşturmak, paketleri yönetmek, yardım belgelerine erişmek ve web içeriğini görüntülemek için kullanılır.

Aşağıda, RStudio'nun bazı temel özellikleri ve kullanımı hakkında bilgiler bulunmaktadır.

3.1. RStudio Kurulumu

R, grafiksel bir arayüze sahip olmayan ve eski tip bir yazılım konsoluna benzeyen bir editör sunar. Bununla birlikte, ücretsiz olarak sunulan RStudio programı, kullanıcılarına daha modern, işlevsel ve kullanıcı dostu bir arayüz sağlar. RStudio'nun yanı sıra, Revolution Analytics, StatET ve ESS gibi diğer editörler de mevcuttur ve R kullanıcıları arasında popüler alternatiflerdir. RStudio programını indirmek ve kurmak için aşağıdaki bağlantıyı ziyaret edebilirsiniz. <https://posit.co/download/rstudio-desktop/>



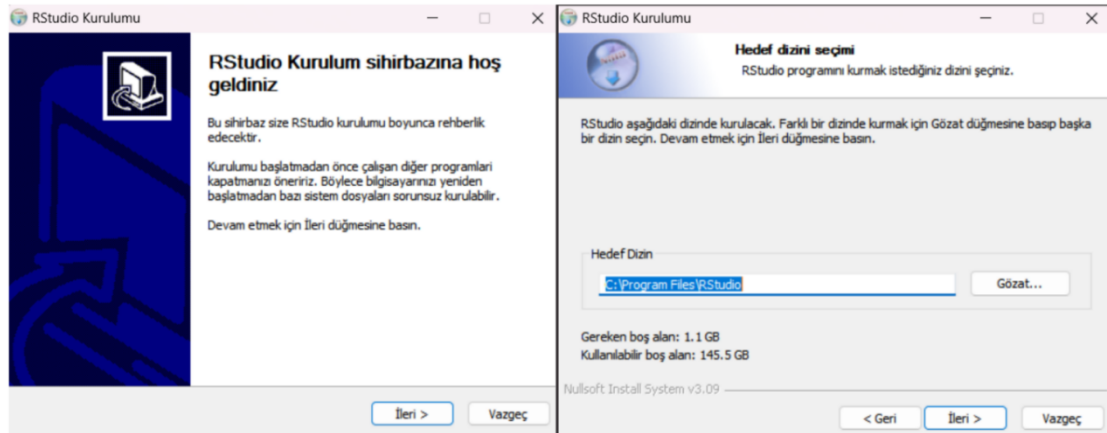
Şekil 28: R ve RStudio indirme ekranı

İşletim sisteminize uygun olan ve en son yayınlanan sürümü seçerek, RStudio platformunu indirip kurabiliriz. Bu, en güncel özellikleri ve en iyi performansı elde etmenizi sağlar, böylece RStudio'nun sunduğu modern ve kullanışlı arayüzden tam anlamıyla faydalanabilirsiniz.

OS	Download	Size	SHA-256
Windows 10/11	RSTUDIO-2024.04.1-748.EXE ↓	263.07 MB	44C8797C
macOS 12+	RSTUDIO-2024.04.1-748.DMG ↓	566.51 MB	A5EDA699
Ubuntu 20/Debian 11	RSTUDIO-2024.04.1-748-AMD64.DEB ↓	194.71 MB	505311AE
Ubuntu 22/Debian 12	RSTUDIO-2024.04.1-748-AMD64.DEB ↓	197.00 MB	88D485CD

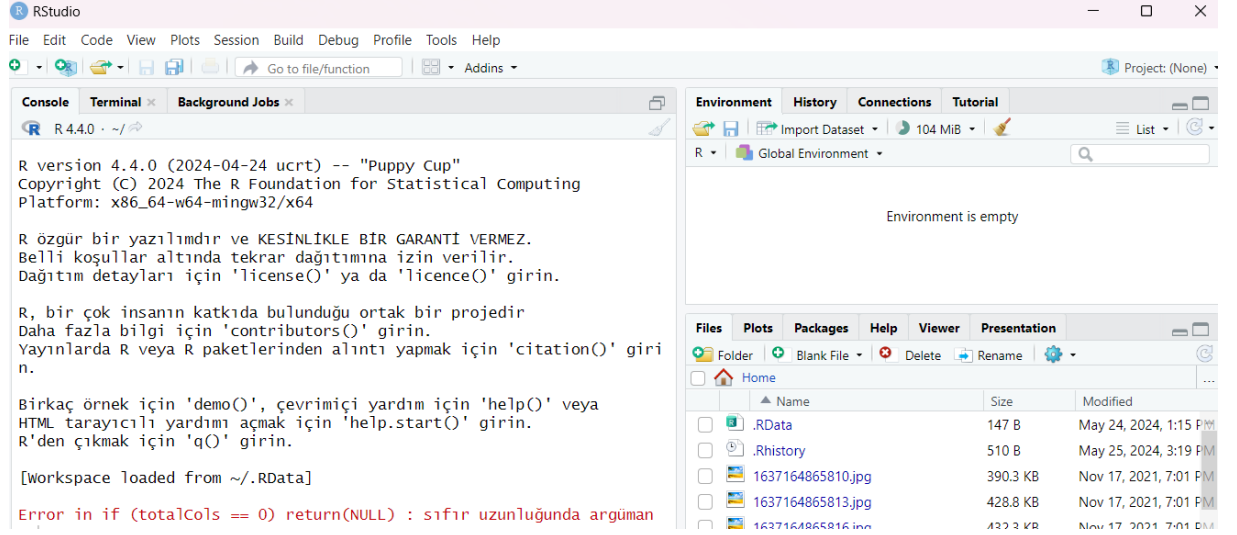
Şekil 29 : İndirebileceğimiz İşletim Sistemleri

Kurulum işlemini dikkatlice devam ettiririz ve ekrandaki talimatları adım adım izleriz. Kurulum dosyaları bilgisayarınıza kopyalanır ve gerekli ayarlamalar otomatik olarak yapılır. Bu aşamada, RStudio'nun sisteminize düzgün bir şekilde entegre olması sağlanır. Dosyaların kopyalanması ve gerekli yapılandırma işlemleri tamamlandığında, kurulum işlemi başarıyla sonlandırılır. Artık RStudio programını kullanmaya hazır hale gelmiş olursunuz, böylece veri analizi ve programlama projelerinizi modern ve kullanıcı dostu bir ortamda gerçekleştirebilirsiniz.



Şekil 30 : RStudio Kurulumu ekranları

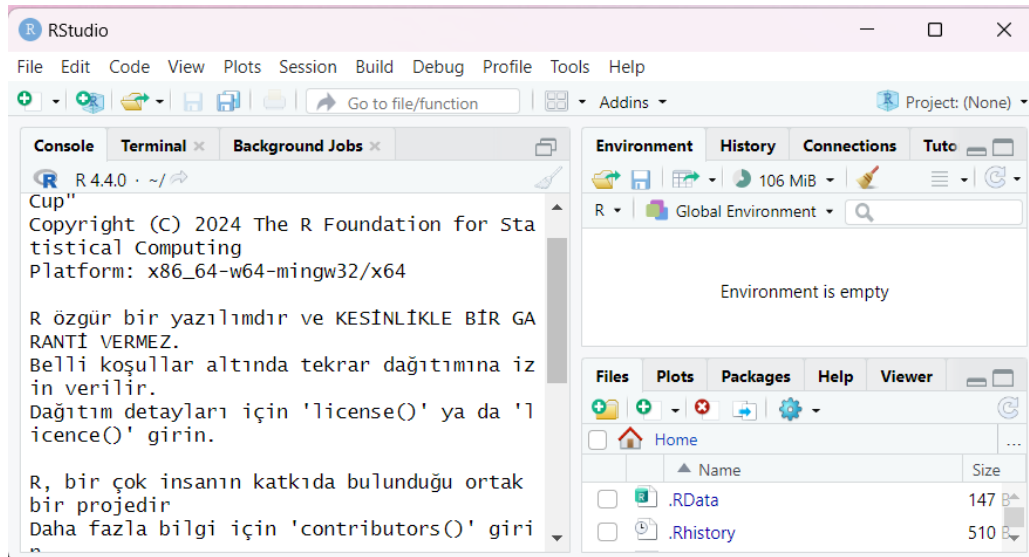
İndirme işlemi tamamlandıktan sonra, bir sonraki adım olan kurulum aşamasına geçeriz. RStudio programının kurulumunu başlatırız ve ekrandaki talimatları takip ederek adım adım ilerleriz. Kurulum sürecinin başlangıcında, kurulum dosyalarının kopyalanacağı yeri dikkatlice seçeriz, bu da programın sisteminizde nereye yükleneceğini belirler. Böylece, RStudio'nun bilgisayarınıza doğru bir şekilde kurulmasını sağlayarak, veri analizlerinizi ve programlama projelerinizi en verimli şekilde gerçekleştirebileceğiniz modern bir çalışma ortamı oluştururuz.



Şekil 31: RStudio Kurulduktan sonraki ilk ekran

3.2. İlk Bakış

RStudio kurulumunu tamamladığınızda ve RStudio'yu çalıştırdığınızda, sizi bir karşılama ekranı karşılayacaktır.



Şekil 32: RStudio Karşılama ekranı

Bu ekranda aşağıdaki seçenekleri göreceksiniz:

- 1. Konsol:** R komutlarını doğrudan çalıştırabileceğiniz yerdir. Konsolda yazdığınız komutlar anında R tarafından işlenir ve sonuçları hemen görebilirsiniz..
- 2. R Ortamı ve Geçmiş:** Bu alan, çalışma alanınızı ve komut geçmişinizi gösterir. Environment sekmesinde yüklediğiniz veri setlerini, tanımladığınız fonksiyonları ve oluşturduğunuz nesneleri görüntüleyebilirsiniz. History sekmesinde ise daha önce çalıştırdığınız komutları görebilirsiniz.
- 3. Dosyalar-Grafikler-Paketler-Yardım:** Bu alan çeşitli sekmelere ayrılmıştır:

Dosyalar: Dosyaları yönetmek için kullanılır. Bu sekmede çalışma dizinindeki dosyaları görebilir ve yönetebilirsiniz.

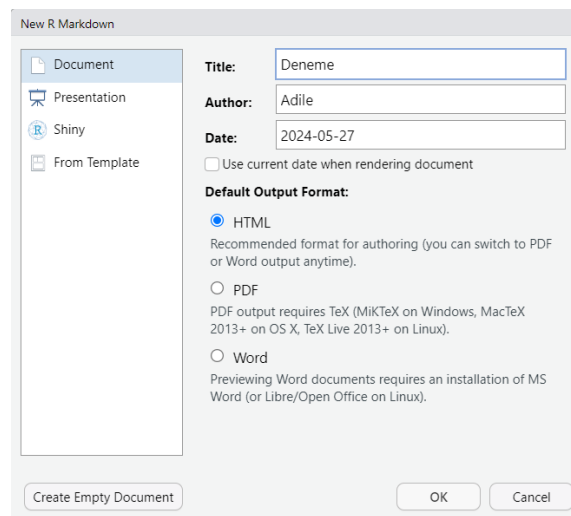
Grafikler: Grafiklerinizi görüntüler. Bu sekmede oluşturduğumuz grafikler ve görselleştirmeler yer alır.

Paketler: Yüklü paketleri listeler ve yeni paketler yüklemenizi sağlar. Bu sekmeden paketlerinizi yönetebilirsiniz.

Yardım: Yardım belgelerine erişmenizi sağlar. R fonksiyonları ve paketleri hakkında ayrıntılı bilgileri buradan bulabilirsiniz.

3.3. Rmarkdown Kullanımı

RStudio'yu açarak RStudio IDE'sini başlatın. Yeni bir RMarkdown belgesi oluşturmak için üst menüden "File" sekmesine tıklayın, ardından açılan menüden "New File" seçeneğini seçin ve "RMarkdown" seçeneğine tıklayın. Bu adımlar sizi yeni bir RMarkdown belgesi oluşturma arayüzüne yönlendirecektir.



Şekil 33: RMarkdown yeni dosya ekranı

Bu ekran görüntüsü, RStudio'da yeni bir RMarkdown belgesi oluşturmak için kullanılan bir arayüzü göstermektedir. Şimdi adım adım bu arayüzdeki seçenekleri açıklayalım:

1.Title (Başlık): Bu alana, oluşturmak istediğiniz RMarkdown belgesinin başlığını yazabilirsiniz. Bu örnekte, başlık “Deneme” olarak girilmiştir.

2.Author (Yazar): Bu alana, belgenin yazarının adını yazabilirsiniz. Bu örnekte, yazar adı “Adile” olarak girilmiştir.

3.Date (Tarih): Bu alana, belgenin oluşturulma tarihini yazabilirsiniz. Tarih otomatik olarak bugünün tarihi olarak ayarlanır, ancak isterseniz bu alanı manuel olarak değiştirebilirsiniz. Bu örnekte tarih “2024-05-27” olarak girilmiştir. “Use current date when rendering document” seçeneği işaretlendiğinde, belge her oluşturulduğunda güncel tarihi kullanır.

4.Default Output Format (Varsayılan Çıktı Formatı): Bu bölümde, belgenizin varsayılan çıktı formatını seçebilirsiniz. Üç seçenek vardır:

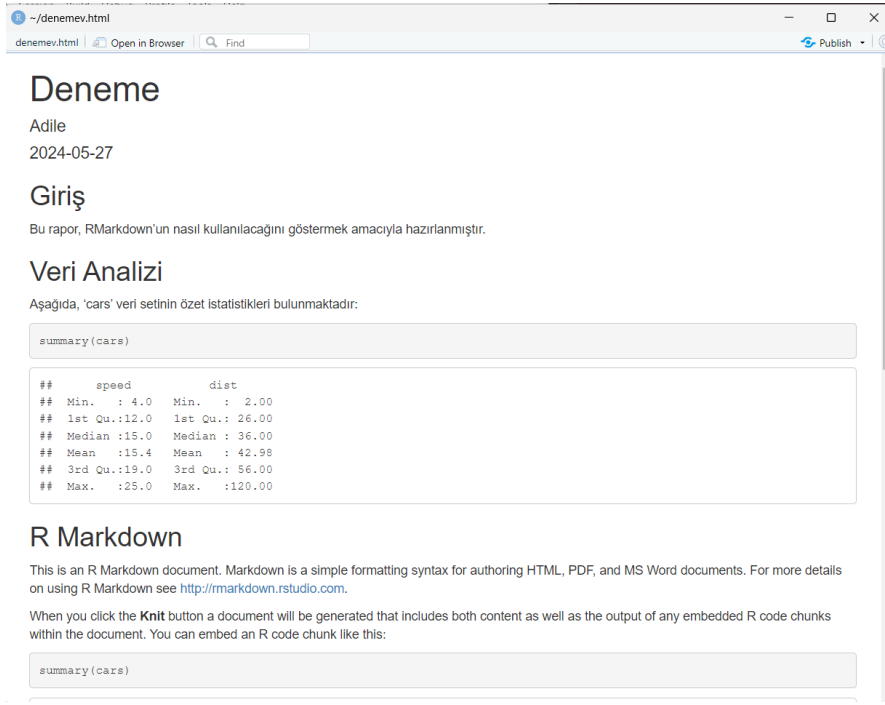
HTML: HTML formatında çıktı oluşturur. Bu format, belgeyi web tarayıcısında görüntülemek için uygundur.

PDF: PDF formatında çıktı oluşturur. Bu format belgeyi yazdırmak veya paylaşmak için uygundur. Ancak, bu formatı kullanmak için TeX (MiKTeX, MacTeX veya TeX Live) kurulmuş olmalıdır

Word: Microsoft Word formatında çıktı oluşturur. Bu format, belgeyi Word veya benzeri yazılımlarda düzenlemek için uygundur. Linux kullanıcıları için LibreOffice veya OpenOffice kurulu olmalıdır.

5.Create Empty Document (Boş Belge Oluşturur): Bu seçenek işaretlendiğinde, şablon kullanmadan boş bir RMarkdown belgesi oluşturulur.

6.OK Butonu: Seçeneklerinizi belirledikten sonra bu butona tıklayarak yeni bir RMarkdown belgesi oluşturabilirsiniz.



Şekil 34: RMarkdown Arayüzü

Bu ekran görüntüsü, RMarkdown kullanarak nasıl dinamik ve yeniden üretilebilir raporlar oluşturulacağını göstermektedir. RMarkdown, veri analizi ve görselleştirmeyi kolayca bir araya getiren güçlü bir araçtır. Bu örnek sayfa, 'cars' veri setinin özet istatistiklerini ve grafiklerini içerir. RStudio arayüzü ile kolayca R kodları yazabilir, çıktıları görüntüleyebilir ve sonuçları zenginleştirilmiş raporlar olarak paylaşabilirsiniz. Bu sayede, veri bilimi ve analitiği çalışmalarınızı daha etkili ve anlaşılır hale getirebilirsiniz.

4. UYGULAMA

Bu bölümde, R programlama dilini kullanarak “student-mat.csv” veri seti üzerinde temel istatistiksel analizler ve veri görselleştirme uygulamaları gerçekleştireceğiz . Analizlerimizde, veri setinin yüklenmesi ve incelenmesi, temel istatistiksel özetlerin çıkarılması, histogram , yoğunluk grafikleri ve kutu grafikleri gibi çeşitli görselleştirme teknikleri kullanılacaktır. Ayrıca, değişkenler arasındaki ilişkileri incelemek için korelasyon analizi yapılacak ve basit regrasyon modeli kurulacaktır. Bu uygulamalar, verilerin daha iyi anlaşılmasını sağlayacak ve analiz sonuçlarını görsel olarak sunarak, veriye dayalı kararların alınmasına yardımcı olacaktır.

4.1. Proje Oluşturma

Bu bölümde, analiz ve görselleştirme işlemlerine başlamadan önce veri setinin yüklenmesi ve gerekli paketlerin kurulumu gerçekleştirilmiştir. İlk olarak, analizde kullanacağımız R paketlerini yüklememiz gerekmektedir. Bu paketler arasında veri manipülasyonu için dplyr, görselleştirme için ggplot2, istatistiksel özetler için summarytools ve korelasyon analizi için corrrplot bulunmaktadır. Paketlerin yüklenmesi ve çağırılması işlemi aşağıdaki kod parçacıkları ile gerçekleştirilmiştir.

```
> # Gerekli Paketleri Yükle
> install.packages("ggplot2")
Error in install.packages : Updating loaded packages
> install.packages("dplyr")
Error in install.packages : Updating loaded packages
> install.packages("summarytools")
Error in install.packages : Updating loaded packages
> install.packages("corrrplot")
Error in install.packages : Updating loaded packages
>
> # Paketleri Çağırın
> library(ggplot2)
> library(dplyr)
> library(summarytools)
> library(corrrplot)
>

Restarting R session...

> install.packages("corrrplot")
Installing package into 'C:/Users/user/AppData/Local/R/win-library/4.4'
(as 'lib' is unspecified)
URL 'https://cran.rstudio.com/bin/windows/contrib/4.4/corrrplot_0.92.zip' deneniyor
Content type 'application/zip' length 3846610 bytes (3.7 MB)
downloaded 3.7 MB

package 'corrrplot' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:/Users/user/AppData/Local/Temp/Rtmpi6FgJA/downloaded_packages

Restarting R session...

> install.packages("summarytools")
```

Şekil 35: Gerekli Paketlerin Yüklenmesi

Bu kod parçasığında “install.packages” fonksiyonu kullanılarak gerekli paketler yüklenmiştir. Daha sonra “library” fonksiyonu ile bu paketler çağrılmıştır. Bu adım, analize başlamadan önce gerekli fonksiyonların kullanılabilir olmasını sağlar.

Bir sonraki adımda analiz edeceğimiz veri setini yüklememiz gerekmektedir. “student-mat.csv” adlı veri setini “read.csv” fonksiyonu ile yükleyerek “data” adlı bir veri çerçevesine atadık. Bu veri seti, öğrencilerin çeşitli akademik ve kişisel bilgilerini içermektedir ve analizin temelini oluşturacaktır.

```
> # Veri Setini Yükle
> data <- read.csv("C:/Users/user/Desktop/student-mat.csv", sep = ";")
>
> # Veri Setini İncele
> head(data)
```

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime
1	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2
2	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1
3	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1
4	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1
5	GP	F	16	U	GT3	T	3	3	other	other	home	father	1
6	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1

	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime
1	2	0	yes	no	no	no	yes	yes	no	no	4	3
2	2	0	no	yes	no	no	no	yes	yes	no	5	3
3	2	3	yes	no	yes	no	yes	yes	yes	no	4	3
4	3	0	no	yes	yes	yes	yes	yes	yes	yes	3	2
5	2	0	no	yes	yes	no	yes	yes	no	no	4	3
6	2	0	no	yes	yes	yes	yes	yes	yes	no	5	4

	goout	Dalc	walc	health	absences	G1	G2	G3
1	4	1	1	3	6	5	6	6
2	3	1	1	3	4	5	5	6
3	2	2	3	3	10	7	8	10
4	2	1	1	5	2	15	14	15
5	2	1	2	5	4	6	10	10
6	2	1	2	5	10	15	15	15

Şekil 36 : Veri Setini yükleme ve inceleme

Yukarıdaki kod parçasığında, “read.csv” fonksiyonu kullanılarak veri seti yüklenmiş ve “data” değişkenine atanmıştır. CSV dosyasının doğru bir şekilde yüklenebilmesi için “sep” argümanı kullanılarak verinin noktalı virgül ile ayrıldığı belirtilmiştir. “head” fonksiyonu ise veri setinin ilk birkaç satırını görüntüleyerek veri seti hakkında genel bir bilgi edinmemizi sağlar. Bu, veri setinin doğru bir şekilde yüklendiğini ve beklenen yapıda olduğunu doğrulamamıza yardımcı olur.

Yükleme ve ilk inceleme adımları tamamlandıktan sonra, veri seti üzerinde daha derinlemesine analizler yapmaya hazırız. Bu süreç, veri setinin genel özelliklerini anlamamıza ve daha ileri analizler için gerekli ön hazırlıkları yapmamıza olanak tanır.

4.2. Tasarımını Oluşturma

Bu aşamada, “student-mat.csv” veri seti üzerinde temel istatistiksel analizler gerçekleştirilmiş ve değişkenlerin ortalama ve standart sapmaları hesaplanmıştır. Temel istatistiksel analizler, veri setinin genel özelliklerini anlamamıza yardımcı olurken, ortalama ve standart sapmalar değişkenlerin merkezi eğilim ve yayılımını anlamamıza yardımcı olur.

İlk olarak, veri setinin temel istatistiksel özetlerini elde etmek için “summary” fonksiyonunu kullandık. Bu fonksiyon, her bir değişken için minimum, birinci çeyrek, medyan, ortalama, üçüncü çeyrek ve maksimum değerleri sağlar.

```
> # Temel İstatistiksel Analizler
> summary(data)
  school      sex      age      address      famsize
Length:395   Length:395   Min.   :15.0   Length:395   Length:395
Class :character   Class :character   1st Qu.:16.0   Class :character   Class :character
Mode  :character   Mode  :character   Median :17.0   Mode  :character   Mode  :character
                        Mean  :16.7
                        3rd Qu.:18.0
                        Max.  :22.0

  Pstatus      Medu      Fedu      Mjob      Fjob
Length:395     Min.   :0.000   Min.   :0.000   Length:395   Length:395
Class :character   1st Qu.:2.000   1st Qu.:2.000   Class :character   Class :character
Mode  :character   Median :3.000   Median :2.000   Mode  :character   Mode  :character
                        Mean  :2.749   Mean  :2.522
                        3rd Qu.:4.000   3rd Qu.:3.000
                        Max.  :4.000   Max.  :4.000

  reason      guardian      traveltime      studytime      failures      schoolsup
Length:395     Length:395     Min.   :1.000   Min.   :1.000   Min.   :0.0000   Length:395
Class :character   Class :character   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   Class :character
Mode  :character   Mode  :character   Median :1.000   Median :2.000   Median :0.0000   Mode  :character
                        Mean  :1.448   Mean  :2.035   Mean  :0.3342
                        3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
                        Max.  :4.000   Max.  :4.000   Max.  :3.0000

  famsup      paid      activities      nursery      higher
Length:395     Length:395     Length:395     Length:395     Length:395
Class :character   Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character

  internet      romantic      famrel      freetime      goout      Dalc
Length:395     Length:395     Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
Class :character   Class :character   1st Qu.:4.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000
Mode  :character   Mode  :character   Median :4.000   Median :3.000   Median :3.000   Median :1.000
                        Mean  :3.944   Mean  :3.235   Mean  :3.109   Mean  :1.481
                        3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000
                        Max.  :5.000   Max.  :5.000   Max.  :5.000   Max.  :5.000

  walc      health      absences      G1      G2      G3
Min.   :1.000   Min.   :1.000   Min.   :0.000   Min.   :3.00   Min.   :0.00   Min.   :0.00
1st Qu.:1.000   1st Qu.:3.000   1st Qu.:0.000   1st Qu.:8.00   1st Qu.:9.00   1st Qu.:8.00
Median :2.000   Median :4.000   Median :4.000   Median :11.00   Median :11.00   Median :11.00
```

Şekil 37 : Temel İstatistiksel Analizler

Yukarıdaki kod ile, “summary” fonksiyonu kullanılarak veri setindeki her bir değişkenin özet istatistikleri elde edilmiştir. Bu özet, veri setindeki değerlerin dağılımı hakkında genel bir fikir verir. Örneğin, her bir değişkenin merkezi eğilimi ve yayılımı hakkında bilgi sahibi olabiliriz.

Daha sonra, her bir değişkenin ortalama ve standart sapmalarını hesaplayarak, verinin merkezi eğilim ve yayılımını daha detaylı bir şekilde inceledik. Bu işlemi gerçekleştirmek için “dplyr” paketindeki “summarise” ve “across” fonksiyonlarını kullandık. “summarise” fonksiyonu, belirtilen özet fonksiyonlarını uygulayarak yeni bir veri çerçevesi döndürürken, “across” fonksiyonu ise belirli sütunlar veya tüm sütunlar üzerinde aynı işlemi uygulamak için kullanılır.

	age_mean	age_sd	Medu_mean	Medu_sd	Fedu_mean	Fedu_sd	traveltime_mean	traveltime_sd	studytime_mean	studytime_sd
1	16.6962	1.276043	2.749367	1.094735	2.521519	1.088201	1.448101	0.6975048	2.035443	0.8392403

```

> # Değişkenlerin Ortalama ve Standart Sapmalarını Hesaplayın
> mean_sd <- data %>%
+   select_if(is.numeric) %>%
+   summarise(across(everything(), list(mean = mean, sd = sd)))
> # Ortalamalar ve Standart Sapmaların Tablo Şeklinde Görüntülenmesi
> print(mean_sd)

```

Şekil 38 : Değişkenlerin Ortalaması ve Standart Sapma Hesaplama

Yukarıdaki kod ile her bir sayısal değişkenin ortalama ve standart sapmaları hesaplanmış ve tablo halinde gösterilmiştir. Bu özetler, her bir değişkenin merkezi eğilim ve yayılımı hakkında detaylı bilgi sağlar.

4.3. Kodlama

Bu bölümde, “student-mat.csv” veri seti üzerinde çeşitli istatistiksel analizler ve görselleştirmeler gerçekleştirilmiştir. İlk olarak, veri setinin yüklenmesi ve incelenmesi işlemleri yapılmıştır. Daha sonra, veri seti üzerinde temel istatistiksel analizler gerçekleştirilmiştir. Son olarak, veri seti üzerinde çeşitli görselleştirmeler yapılmıştır.

Statistic	Value
1 Ortalama	10.41519
2 Medyan	11.00000
3 Mod	10.00000

Showing 1 to 3 of 3 entries. 2 total columns

```

R 4.4.0 - C:/Users/user/Desktop/
> mean_grade <- mean(student$G3, na.rm = TRUE)
> print(paste("Ortalama:", mean_grade))
[1] "Ortalama: 10.4151898734177"
>
> # Medyan hesaplama
> median_grade <- median(student$G3, na.rm = TRUE)
> print(paste("Medyan:", median_grade))
[1] "Medyan: 11"
>
> # Mod hesaplama fonksiyonu
> get_mode <- function(v) {
+   uniqv <- unique(v)
+   uniqv[which.max(tabulate(match(v, uniqv)))]
+ }
>
> # Mod hesaplama
> mode_grade <- get_mode(student$G3)
> print(paste("Mod:", mode_grade))
[1] "Mod: 10"
>
> # Değerleri bir veri çerçevesine yazma
> central_tendency <- data.frame(
+   Statistic = c("Ortalama", "Medyan", "Mod"),
+   Value = c(mean_grade, median_grade, mode_grade)
+ )
>
> # Tabloyu görüntüleme
> print(central_tendency)
  Statistic  Value
1 Ortalama 10.41519
2 Medyan 11.00000
3 Mod 10.00000

```

Şekil 39 : Ortalama , Medyan ve Mod hesaplama

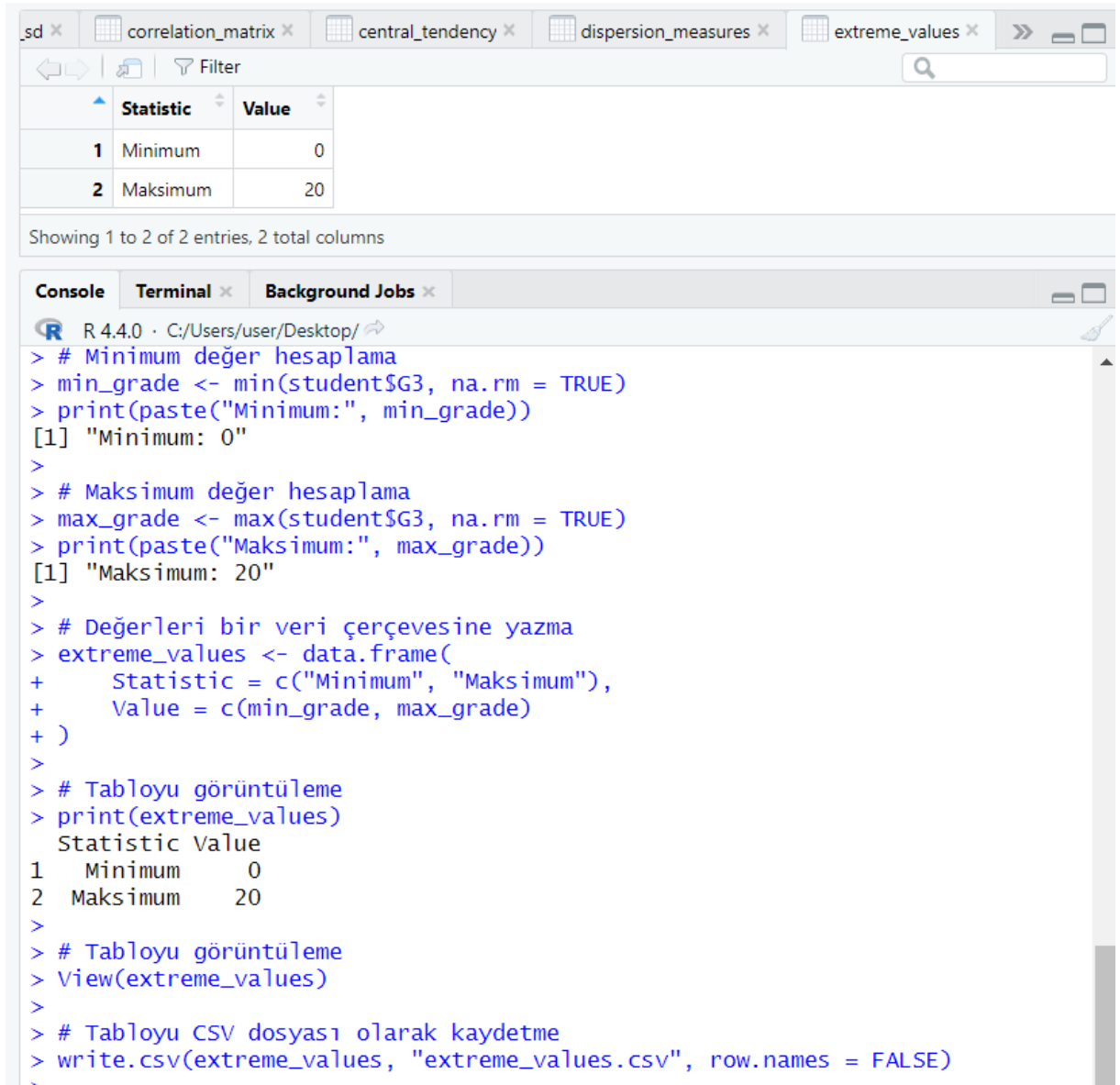
Bu çalışmada, öğrenci notlarının merkezi eğilim ölçüleri hesaplanmıştır. Ortalama, “mean” fonksiyonu kullanılarak 10.41519 olarak bulunmuştur. Medyan değer, “median” fonksiyonu ile 11 olarak hesaplanmıştır. Mod değeri ise en sık tekrar eden değeri bulan bir fonksiyon (“get_mode”) yardımıyla 10 olarak belirlenmiştir. Bu değerler bir veri çerçevesine yazılarak tablo halinde sunulmuştur. Hesaplamalarda eksik değerler dikkate alınmamıştır (“na.rm = TRUE”). Bu ölçüler, veri setindeki notların genel eğilimini anlamak için önemli bir özet sağlamaktadır.

The screenshot shows an RStudio window with several tabs at the top: 'student', 'data', 'mean_sd', 'correlation_matrix', 'central_tendency', 'dispersion_measures', and 'extreme_values'. The 'dispersion_measures' tab is active, displaying a table with two columns: 'Statistic' and 'Value'. The table contains two rows: 'Standart Sapma' with a value of 4.581443, and 'Varyans' with a value of 20.989616. Below the table, it says 'Showing 1 to 2 of 2 entries, 2 total columns'. The console at the bottom shows the following R code and output:

```
R 4.4.0 · C:/Users/user/Desktop/
> # Standart sapma hesaplama
> std_dev_grade <- sd(student$G3, na.rm = TRUE)
> print(paste("Standart Sapma:", std_dev_grade))
[1] "Standart Sapma: 4.58144261099784"
>
> # Varyans hesaplama
> variance_grade <- var(student$G3, na.rm = TRUE)
> print(paste("Varyans:", variance_grade))
[1] "Varyans: 20.9896163978667"
>
> # Tabloyu görüntüleme
> View(dispersion_measures)
Error in View : 'dispersion_measures' nesnesi bulunamadı
>
> # Tabloyu CSV dosyası olarak kaydetme
> write.csv(dispersion_measures, "dispersion_measures.csv", row.names = FALSE)
Error in eval(expr, p) : 'dispersion_measures' nesnesi bulunamadı
> # Değerleri bir veri çerçevesine yazma
> dispersion_measures <- data.frame(
+   Statistic = c("Standart Sapma", "Varyans"),
+   Value = c(std_dev_grade, variance_grade)
+ )
>
> # Tabloyu görüntüleme
> print(dispersion_measures)
  Statistic      Value
1 Standart Sapma 4.581443
2      Varyans 20.989616
>
> # Tabloyu görüntüleme
> View(dispersion_measures)
>
> # Tabloyu CSV dosyası olarak kaydetme
> write.csv(dispersion_measures, "dispersion_measures.csv", row.names = FALSE)
```

Şekil 40 : Standart Sapma ve Medyan hesaplama

Bu çalışmada, öğrenci notlarının merkezi eğilim ölçüleri hesaplanmıştır. Standart sapma değeri “sd” fonksiyonu kullanılarak 4.58144 olarak belirlenmiştir. Varyans değeri ise “var” fonksiyonu kullanılarak 20.98962 olarak hesaplanmıştır. Bu değerler, notların ne kadar yayılmış olduğunu ve veri setinin ortalamadan sapma miktarını anlamak için önemlidir. Hesaplanan standart sapma ve varyans değerleri bir veri çerçevesinde birleştirilmiş ve tablo halinde sunulmuştur. Hesaplamalarda eksik değerler dikkate alınmamıştır (“na.rm = TRUE”). Elde edilen tablo, veri setinin genel dağılımı hakkında önemli bilgiler sağlamaktadır.



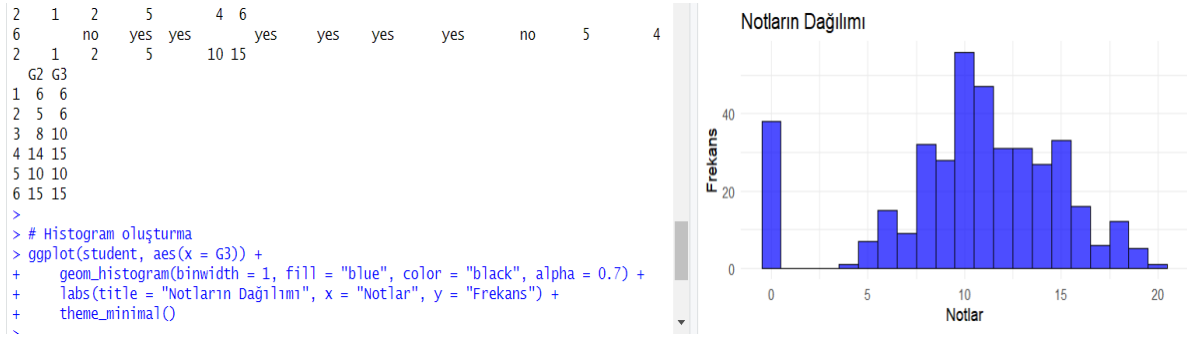
The screenshot displays the R Studio interface. At the top, there are tabs for various data analysis functions: 'sd', 'correlation_matrix', 'central_tendency', 'dispersion_measures', and 'extreme_values'. Below these tabs is a table with two columns: 'Statistic' and 'Value'. The table contains two rows: '1 Minimum 0' and '2 Maksimum 20'. Below the table, it says 'Showing 1 to 2 of 2 entries, 2 total columns'. At the bottom, there is a 'Console' tab showing R code and its output. The code calculates the minimum and maximum values of a variable 'G3' in a dataset 'student', stores them in 'min_grade' and 'max_grade', and then creates a data frame 'extreme_values' with these values. The output shows the minimum value as 0 and the maximum value as 20. The code also prints and views the 'extreme_values' data frame, and saves it as a CSV file.

	Statistic	Value
1	Minimum	0
2	Maksimum	20

```
R 4.4.0 · C:/Users/user/Desktop/
> # Minimum değeri hesaplama
> min_grade <- min(student$G3, na.rm = TRUE)
> print(paste("Minimum:", min_grade))
[1] "Minimum: 0"
>
> # Maksimum değeri hesaplama
> max_grade <- max(student$G3, na.rm = TRUE)
> print(paste("Maksimum:", max_grade))
[1] "Maksimum: 20"
>
> # Değerleri bir veri çerçevesine yazma
> extreme_values <- data.frame(
+   Statistic = c("Minimum", "Maksimum"),
+   Value = c(min_grade, max_grade)
+ )
>
> # Tabloyu görüntüleme
> print(extreme_values)
  Statistic Value
1  Minimum     0
2  Maksimum    20
>
> # Tabloyu görüntüleme
> View(extreme_values)
>
> # Tabloyu CSV dosyası olarak kaydetme
> write.csv(extreme_values, "extreme_values.csv", row.names = FALSE)
>
```

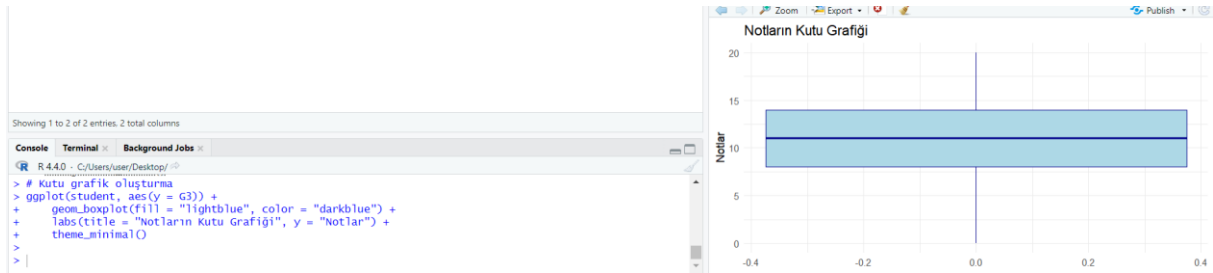
Şekil 41 : Minimum ve Maksimum değeri hesaplama

Bu çalışmada, öğrenci notlarının en düşük (minimum) ve en yüksek (maksimum) değerleri hesaplanmıştır. Minimum değer “min” fonksiyonu kullanılarak 0 olarak belirlenmiştir. Maksimum değer ise “max” fonksiyonu kullanılarak 20 olarak bulunmuştur. Bu değerler, veri setindeki notların uç noktalarını göstermektedir ve notların hangi aralıkta değiştiğini anlamak için kritiktir. Hesaplanan minimum ve maksimum değerler bir veri çerçevesinde birleştirilerek tablo halinde sunulmuştur. Hesaplamalarda eksik değerler dikkate alınmamıştır (“na.rm = TRUE”). Bu tablo, notların dağılımının sınırlarını belirlemek için önemli bilgiler sağlamaktadır.



Şekil 42 : Histogram Grafiği

Bu çalışmada, öğrenci notlarının dağılımını görselleştirmek için bir histogram oluşturulmuştur. Histogram, “ggplot2” kütüphanesi kullanılarak hazırlanmış ve notların frekansını göstermektedir. Histogramda, notların x ekseninde, frekanslarının ise y ekseninde gösterildiği görülmektedir. Mavi renk ile doldurulmuş sütunlar, notların hangi aralıklarda yoğunlaştığını ve dağıldığını açıkça ortaya koymaktadır. Histogram, öğrenci notlarının yaklaşık olarak normal bir dağılım sergilediğini ve ortalama notun etrafında yoğunlaştığını göstermektedir. Bu tür görselleştirmeler, veri setinin genel yapısını ve dağılımını anlamak için oldukça faydalıdır.



Şekil 43 : Kutu Grafiği

Bu çalışmada, öğrenci notlarının dağılımını ve olası aykırı değerleri görselleştirmek için bir kutu grafiği (boxplot) oluşturulmuştur. Kutu grafiği, “ggplot2” kütüphanesi kullanılarak hazırlanmış ve notların medyanı, çeyrekler arası aralığı (IQR) ve olası aykırı değerleri göstermektedir. Grafikte, kutunun içindeki kalın çizgi notların medyanını, kutunun üst ve alt sınırları ise çeyrekler arası aralığı temsil etmektedir. Çeyrekler arası aralığın dışında kalan çizgiler (whiskers) veri setindeki minimum ve maksimum değerleri göstermektedir. Bu tür görselleştirmeler, veri setinin merkezi eğilim ve dağılım özelliklerini daha iyi anlamak için oldukça faydalıdır.

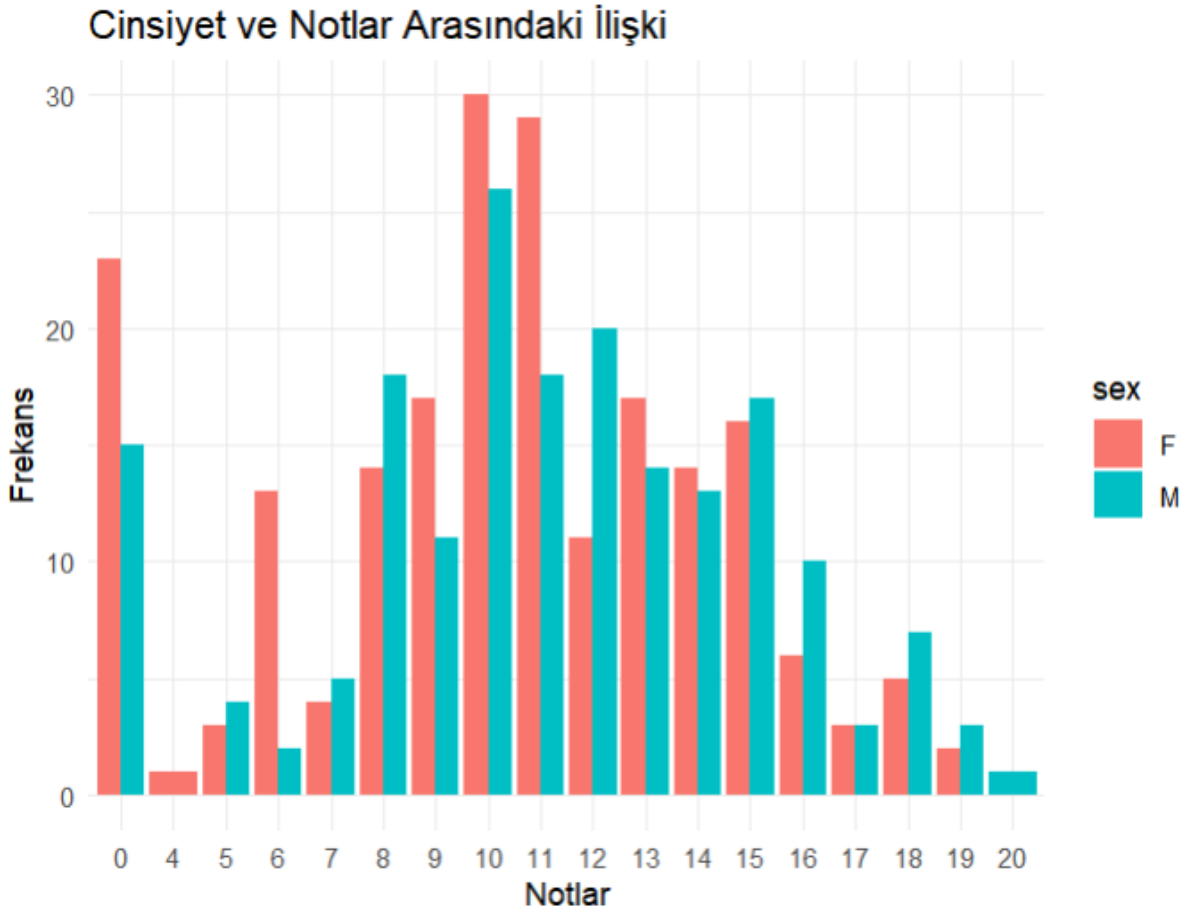
Bu çalışmada, öğrenci verilerindeki çeşitli değişkenler arasındaki ilişkileri görselleştirmek için bir korelasyon matrisi oluşturulmuştur. Korelasyon matrisi, her iki değişken çifti arasındaki Pearson korelasyon katsayısını göstermektedir. Genel olarak, notlar arasında yüksek pozitif korelasyonlar gözlemlenmiştir. Birinci dönem notları (G1) ile ikinci dönem notları (G2) arasında 0.85, ikinci dönem notları (G2) ile son notlar (G3) arasında 0.9 ve birinci dönem notları (G1) ile son notlar (G3) arasında 0.8'lik pozitif korelasyon bulunmaktadır. Bu, bir öğrencinin önceki dönem notlarının sonraki dönem notlarını güçlü bir şekilde öngördüğünü göstermektedir.

Ebeveyn eğitim seviyesinin (Medu ve Fedu) öğrenci notları ile pozitif bir ilişkisi olduğu görülmüştür. Anne eğitim seviyesi (Medu) ile son notlar (G3) arasındaki korelasyon 0.22, baba eğitim seviyesi (Fedu) ile son notlar (G3) arasındaki korelasyon ise 0.16'dır. Bu, ebeveynlerin eğitim seviyesinin öğrencinin akademik başarısı üzerinde olumlu bir etkisi olduğunu göstermektedir.

Başarısızlıklar (failures) ile notlar arasında negatif bir ilişki bulunmaktadır. Failures ile G3 arasındaki korelasyon -0.36 olup, başarısızlık sayısının artmasıyla birlikte öğrencinin notlarının düşme eğiliminde olduğunu göstermektedir. Haftalık alkol tüketimi (Walc) ile notlar arasında da negatif bir ilişki gözlemlenmiştir. Walc ile G3 arasındaki korelasyon -0.16 olup, yüksek alkol tüketiminin öğrencinin notlarını olumsuz etkileyebileceğini işaret etmektedir.

Diğer değişkenlerle ilişkiler incelendiğinde, anne ve baba eğitim seviyeleri (Medu ve Fedu) arasında 0.62'lik pozitif bir korelasyon bulunmuştur. Bu, ebeveynlerin eğitim seviyelerinin birbirine benzer olma eğiliminde olduğunu göstermektedir. Haftalık alkol tüketimi (Walc) ile günlük alkol tüketimi (Dalc) arasında güçlü pozitif bir korelasyon (0.65) bulunmaktadır. Bu, yüksek günlük alkol tüketiminin haftalık alkol tüketimiyle ilişkili olduğunu göstermektedir. Failures ile studytime (çalışma süresi) arasında -0.17'lik negatif bir korelasyon olup, çalışma süresinin artmasıyla başarısızlıkların azaldığını işaret etmektedir. Walc ile goout (dışarı çıkma sıklığı) arasında 0.42'lik pozitif bir korelasyon bulunmuştur, bu da daha sık dışarı çıkan öğrencilerin daha fazla alkol tüketme eğiliminde olduğunu göstermektedir.

Sağlık durumu (health) ile devamsızlıklar (absences) arasında anlamlı bir korelasyon bulunmamaktadır (0.11). Ayrıca, devamsızlıklar (absences) ile son notlar (G3) arasında da anlamlı bir korelasyon gözlenmemiştir (0.03). Bu korelasyon matrisi, veri setindeki değişkenler arasındaki ilişkileri görselleştirerek öğrencilerin akademik performanslarını etkileyen faktörleri daha iyi anlamamıza yardımcı olmaktadır. Pozitif ve negatif korelasyonların gücü, hangi faktörlerin öğrencilerin başarılarını olumlu ya da olumsuz etkilediğini belirlemede önemli ipuçları sağlamaktadır.



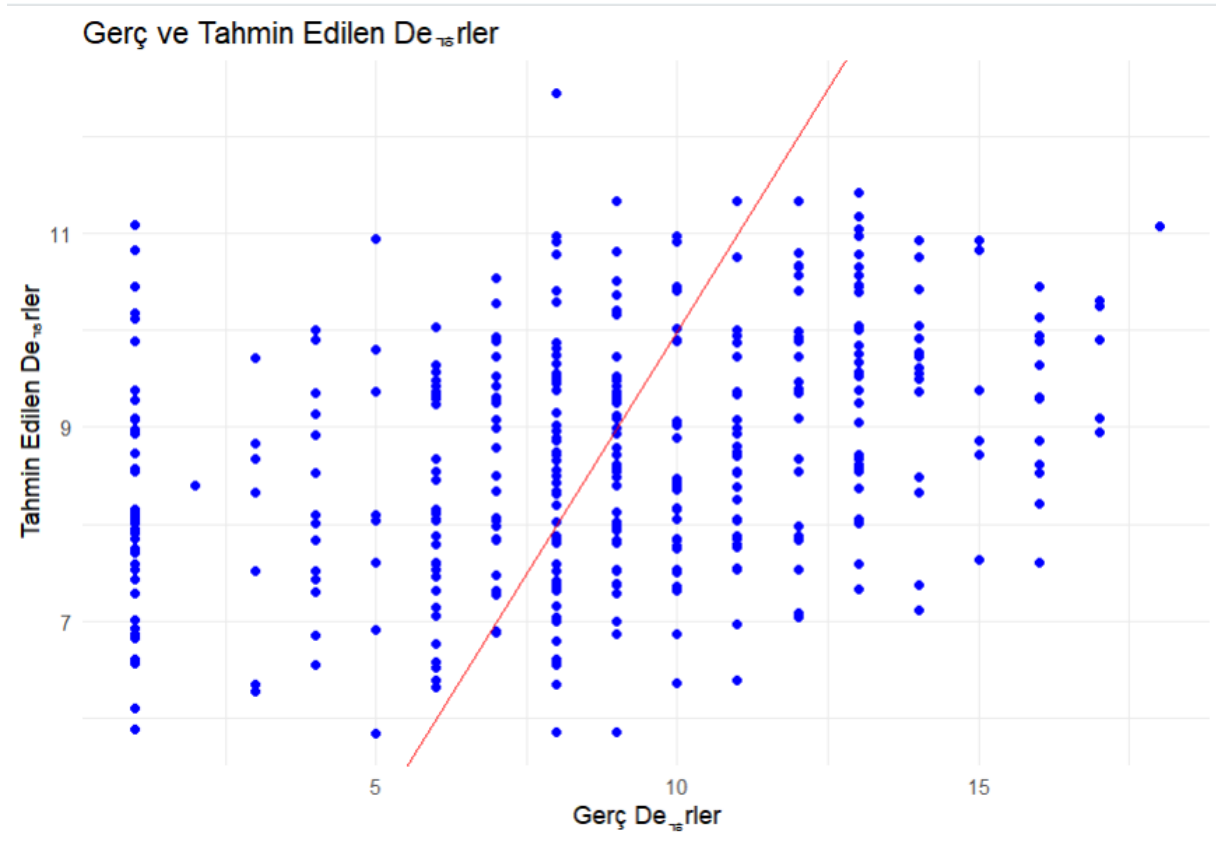
Şekil 44 : Çapraz Tablo Oluşturma

Bu çalışmada, cinsiyet ile öğrenci notları arasındaki ilişkiyi incelemek için bir histogram oluşturulmuştur. Histogram, kız (F) ve erkek (M) öğrencilerin notlarının frekans dağılımını göstermektedir. Grafikte, kız öğrenciler kırmızı, erkek öğrenciler ise mavi renk ile temsil edilmiştir.

Grafikteki gözlemlerimize göre, her iki cinsiyet için de notların genellikle 0 ile 20 arasında dağıldığı görülmektedir. En yüksek frekans, 10 ve 11 notlarında yoğunlaşmıştır ve her iki cinsiyet için de bu notlar oldukça yaygındır. Özellikle 0 notu, kız öğrenciler arasında daha yaygınken, erkek öğrencilerde 10 ve 11 notları daha fazla gözlemlenmiştir.

Grafik, cinsiyetler arasında belirgin bir performans farkı olmadığını, ancak belirli not aralıklarında küçük farklılıklar olabileceğini göstermektedir. Örneğin, erkek öğrenciler 11 ve 12 notlarında daha fazla yoğunlaşırken, kız öğrencilerde bu notlar biraz daha az frekansla temsil edilmiştir. Genel olarak, her iki cinsiyetin not dağılımı benzerdir ve belirli notlarda küçük değişiklikler gözlemlenmiştir.

Bu tür görselleştirmeler, cinsiyetin akademik performans üzerindeki etkilerini anlamak ve potansiyel cinsiyet farklarını belirlemek için önemli ipuçları sağlamaktadır.



Şekil 45 : Çoklu Regresyon

Bu çalışmada, öğrenci notlarının gerçek değerleri ile çoklu regresyon modeli kullanılarak tahmin edilen değerleri arasındaki ilişkiyi incelemek için bir dağılım grafiği (scatter plot) oluşturulmuştur. Grafikte, yatay eksen (x eksen) gerçek değerleri, dikey eksen (y eksen) ise tahmin edilen değerleri göstermektedir. Mavi noktalar her bir veri çiftini temsil ederken, kırmızı çizgi $y = x$ doğrusunu göstermektedir. Bu doğru, tahmin edilen değerlerin mükemmel bir şekilde gerçek değerlere eşit olduğu durumu ifade eder.

Grafikğin genel görünümüne baktığımızda, noktaların kırmızı çizgi etrafında dağılmakla birlikte belirli bir yayılım gösterdiği görülmektedir. Bu durum, tahmin edilen değerlerin gerçek değerlerle birebir örtüşmediğini, ancak genel bir eğilimi takip ettiğini göstermektedir. Noktaların kırmızı çizgiye ne kadar yakın olduğu, tahminlerin doğruluğunu ve modelin performansını yansıtır. Kırmızı çizgiden sapmalar ise tahminlerdeki hataları ve modelin öngörü gücündeki farklılıkları gösterir.

Grafikte, bazı noktaların kırmızı çizgiden oldukça uzaklaştığı, yani tahmin edilen değerlerin gerçek değerlerden belirgin bir şekilde farklılaştığı görülmektedir. Bu sapmalar, modelin belirli not aralıklarında daha az doğru tahminler yaptığını işaret eder. Özellikle 5 ile 10 arasındaki gerçek değerlerde, tahmin edilen değerlerin daha geniş bir yayılım gösterdiği fark edilmektedir.

Bu çoklu regresyon modeli, birden fazla bağımsız değişkenin öğrenci notları üzerindeki etkisini tahmin etmek için kullanılmıştır. Modelin doğruluğunu değerlendirmek ve iyileştirme alanlarını belirlemek için bu tür bir görselleştirme önemlidir. Modelin performansını artırmak için, hatalı tahminlerin nedenlerini analiz etmek ve modelin parametrelerini yeniden gözden

geçirmek gerekebilir. Ayrıca, veri setindeki olası eksiklikler veya önyargılar da tahmin doğruluğunu etkileyebilir, bu yüzden veri kalitesi ve model seçimi kritik öneme sahiptir. Bu analiz, modelin genel performansını anlamak ve iyileştirme fırsatlarını belirlemek için değerli bilgiler sağlamaktadır.

5. SONUÇ

Bu çalışmada, öğrenci notlarının istatistiksel analizi ve görselleştirilmesi yapılmıştır. Merkezi eğilim ve dağılım ölçüleri hesaplanarak veri setinin genel eğilimleri ve dağılımı analiz edilmiştir. Histogram ve kutu grafiği kullanılarak notların dağılımı ve aykırı değerler görselleştirilmiştir.

Korelasyon matrisi ile değişkenler arasındaki ilişkiler incelenmiş ve ebeveyn eğitim seviyesi, başarısızlık sayısı ve alkol tüketimi gibi faktörlerin öğrenci notları ile anlamlı korelasyonları olduğu bulunmuştur. Cinsiyet ile notlar arasındaki ilişki incelenmiş ve belirgin bir performans farkı olmadığı, ancak bazı not aralıklarında küçük farklılıklar olduğu görülmüştür.

Çoklu regresyon modeli kullanılarak notların tahmin edilen ve gerçek değerleri karşılaştırılmış, modelin bazı not aralıklarında daha az doğru tahminler yaptığı tespit edilmiştir. Bu analizler, öğrenci performansını etkileyen faktörleri anlamak ve modelin iyileştirilmesi için gerekli adımları belirlemek açısından önemli bilgiler sağlamıştır. Sonuçlar, veri analizi ve modelleme süreçlerinde iyileştirme fırsatları sunmuştur. bölümde, R programlama dilini kullanarak “student-mat.csv” veri seti üzerinde temel istatistiksel analizler ve veri görselleştirme uygulamaları gerçekleştireceğiz . Analizlerimizde, veri setinin yüklenmesi ve incelenmesi, temel istatistiksel özetlerin çık

6. KAYNAKLAR

1. URL [https://tr.wikipedia.org/wiki/R_\(programlama_dili\)](https://tr.wikipedia.org/wiki/R_(programlama_dili))
2. URL <file:///C:/Users/user/Downloads/kurulum%20bilgileri.pdf>
3. URL <https://medium.com/@emrekuru34/r-%C3%BCzeri%CC%87ndeki%CC%87-hazir-veri%CC%87-setleri%CC%87ne-nasil-eri%CC%87%C5%9Fi%CC%87li%CC%87r-877b4f7b9d2b>
4. URL <https://demir.pw/courses/r-ekonometri-1/rmarkdowndev/>
5. URL <https://cran.rstudio.com/>
6. URL <https://coderspace.io/blog/r-programlama-dili-nedir-kimler-neden-r-ogrenmelid/#:~:text=1993%20y%C4%B1%C4%B1nda%20Yeni%20Zelanda'n%C4%B1n,kaynakl%C4%B1%20%C3%B6zg%C3%BCr%20bir%20vaz%C4%B1%C4%B1m%20dildir.>
7. URL <https://archive.ics.uci.edu/dataset/320/student+performance>