

# 4\_hafta\_temel\_istatistik

Hakan Mehmetcik

2024-10-28

## Temel İstatistik Kavramlarına Giriş

### Tanımlayıcı/Betimsel İstatistik

**İstatistiğin Amacı:** İstatistik, verilerden sorularımızı yanıtlamamıza yardımcı olmak için vardır. Yani, verileri analiz ederek, onları anlamaya çalışırız. Fakat ilk olarak verileri anlamamız ve özetlememiz gereklidir.

Verileri özetlemenin bir yolu, **tanımlayıcı/betimsel istatistikler** kullanmaktır. Bu istatistikler, veri setindeki genel durumu anlamamıza yardımcı olur. Ancak, tanımlayıcı istatistiklerin sonuçlarına dayanarak kesin kararlar vermememiz gerekir. Bunun yerine, tanımlayıcı istatistikler bize veri setindeki değişkenler arasındaki ilginç ilişkileri keşfetme fırsatı sunar.

#### Note

Veri analizi bağlamında, **tanımlayıcı/betimsel istatistik** ve **yorumlayıcı/çıkarımsal istatistik** iki önemli kavramdır. Bu iki istatistik türü, verileri farklı şekillerde kullanmamıza olanak tanır.

**Tanımlayıcı/betimsel İstatistik** Tanımlayıcı istatistikler, bir veri setinin genel özelliklerini özetlemeye yönelik yöntemlerdir. Bu istatistikler, ortalama, medyan, mod, varyans gibi ölçümleri içerir ve verilerin dağılımı hakkında bilgi verir. Ancak, tanımlayıcı istatistiklerden kesin kararlar vermek mümkün değildir; bunlar yalnızca veri setindeki genel durumu anlamamıza yardımcı olur.

**Yorumlayıcı/çıkarımsal İstatistik** Yorumlayıcı istatistik ise verileri analiz etmekle ilgilidir; yani verileri özetlemek yerine, örnek verilerden tüm popülasyon hakkında çıkarımlar yapmayı amaçlar. Yani, yorumlayıcı istatistik, örnek verileri kullanarak popülasyon hakkında sonuçlar çıkarmak veya çıkarımlarda bulunmak için kullanılır. Bu bağlamda, yorumlayıcı istatistik, popülasyon hakkında ne kadar güvenilir sonuçlar çıkarabileceğimizi belirlemek için korelasyonlar, olasılık, regresyon gibi çeşitli istatistiksel yöntemler kullanır.

### Özetle:

- **Tanımlayıcı İstatistik:** Veri setinin genel özelliklerini özetler, kesin kararlar vermez.
- **Yorumlayıcı İstatistik:** Verileri analiz eder ve örnek verilerden popülasyon hakkında çıkarımlar yapar.

## Kategorik Değişkenlerle Kullanılabilen Tanımlayıcı İstatistikler

### 1. Frekans (Frequency):

- Kategorik değişkenlerin her bir seviyesinin kaç kez tekrarlandığını gösterir. Örneğin, bir anket sonucunda “evet” ve “hayır” cevaplarının sayısı.

### 2. Yüzde (Percentage):

- Her bir kategorinin toplam içindeki oranını gösterir. Frekansların toplam gözlem sayısına bölünmesiyle elde edilir. Örneğin, “evet” cevabının yüzdesi, “evet” sayısının toplam cevap sayısına bölünmesi ile hesaplanır.

$$p = \frac{\text{kategori içindeki birey sayısı}}{\text{örneklem büyüklüğü}}$$

### 3. Mod (Mode):

- En sık rastlanan kategori veya değer. Kategorik veriler için en yaygın olan seviyeyi belirtir. Örneğin, bir sınıftaki en çok tercih edilen renk.

### 4. Çapraz Tablo (Contingency Table):

- İki veya daha fazla kategorik değişken arasındaki ilişkiyi gösterir. Her bir kategorinin kesişimindeki frekansları içerir. Örneğin, cinsiyet ve hayatta kalma durumu arasındaki ilişkiyi gösteren bir tablo.

## Örnek 1: Kitap okuma Oranları ve Bilim Haberlerine İlgisi

```
# Kitap okuma kategorileri ve sayıları
books_readers <- c("no_books"=395, "print_only"=577, "digital_only"=91, "print_and_digital"=)
books_readers # Kitap okuyanların sayısını görüntüle
```

no_books	print_only	digital_only	print_and_digital
395	577	91	425

```
# Sadece basılı kitap okuyanların örnek oranını hesaplama
```

### 1. Frekans:

**Kitap Okuma Kategorileri ve Sayıları:**

- Basılı Kitap Okumayan (no\_books): 395
- Sadece Basılı Kitap Okuyan (print\_only): 577
- Sadece Dijital Kitap Okuyan (digital\_only): 91
- Hem Basılı Hem Dijital Kitap Okuyan (print\_and\_digital): 425

Bu frekanslar, her kitap okuma kategorisinde kaç okuyucu bulunduğunu gösterir.

### 2. Yüzde

**Toplam Okuyucu Sayısı:**

```
toplam_okuyucu <- sum(books_readers) # Toplam okuyucu sayısını hesapla
print_only_proportion <- books_readers["print_only"] / toplam_okuyucu # Sadece basılı kitap
print_only_proportion # Oranı görüntüle
```

```
print_only
0.3877688
```

```
# Kesirli tablolar, iki veya daha fazla kategorik değişken arasındaki tüm olası varyasyonları
# Burada, kitap okuma kategorilerinin oranlarını içeren bir tablo oluşturuluyor
oran_tablosu <- books_readers / toplam_okuyucu
oran_tablosu # Oluşturulan oran tablosunu görüntüle
```

no_books	print_only	digital_only	print_and_digital
0.26545699	0.38776882	0.06115591	0.28561828

### 3. Mod

- **En Sık Görülen Kategori:** “Sadece Basılı Kitap Okuyan” (577). Bu, en çok okunan kitap türüdür.

#### 4. Çapraz Tablo

##### Etnik Gruplar ve Bilim Haberlerine İlgi:

```
# Bilim haberlerini aktif, sıradan veya ilgisiz bir şekilde tüketiyor musunuz?
# Farklı etnik grupların bilim haberlerine olan ilgisini göstermek için kategoriler ve sayılar
white <- c("active"=487, "casual"=916, "uninterested"=1431, "no_answer"=28)
black <- c("active"=59, "casual"=98, "uninterested"=227, "no_answer"=8)
hispanic <- c("active"=89, "casual"=152, "uninterested"=183, "no_answer"=23)

# Elde edilen verileri bir veri çerçevesi (data frame) olarak birleştirme
my_table <- as.data.frame(rbind(white, black, hispanic))

# Her bir grubun toplamını hesaplama
my_table$rowsum <- rowSums(my_table) # Her bir satırın toplamını hesapla
my_table["colsum",] <- colSums(my_table) # Her bir sütunun toplamını hesapla

# Sonuçları görüntüleme
my_table # Hesaplanan tabloyu görüntüle
```

	active	casual	uninterested	no_answer	rowsum
white	487	916	1431	28	2862
black	59	98	227	8	392
hispanic	89	152	183	23	447
colsum	635	1166	1841	59	3701

##### Yorumlar:

###### 1. Sadece Basılı Kitap Okuyanların Oranı:

- **Oran:** 0.3877688 (yaklaşık %38.78)
- Bu, toplam okuyucu sayısının yaklaşık %38.78'inin yalnızca basılı kitap okuduğunu göstermektedir. Bu oran, okuyucuların çoğunluğunun basılı kitapları tercih ettiğini ortaya koymaktadır.

###### 2. Diğer Kategoriler:

- **Hiç Kitap Okumayanlar (no\_books):** Yaklaşık %26.55 kişi, hiç kitap okumadığını belirtmiştir.
- **Yalnızca Basılı Kitap Okuyanlar (print\_only):** %38.78, yalnızca basılı kitap okuduğunu ifade etmiştir.

- **Yalnızca Dijital Kitap Okuyanlar (digital\_only):** %6.12, yalnızca dijital kitap okuduğunu belirtmiştir.
- **Hem Basılı Hem de Dijital Kitap Okuyanlar (print\_and\_digital):** %28.56, hem basılı hem de dijital kitap okuduğunu söylemiştir.

Bu oranlar, kitap okuma alışkanlıklarının dağılımını daha net bir şekilde gösterir. Yalnızca basılı kitap okuma oranı en yüksekken, yalnızca dijital kitap okuma oranı oldukça düşüktür.

### 3. Bilim Haberlerine Tüketim:

- Her üç grup arasında “ilgili” ve “sıradan” olarak sınıflandırılan birey sayısının yüksek olması dikkat çekicidir. Ancak “ilgisiz” olanların sayısı en yüksektir; bu durum, bilim haberlerine olan ilgisizliği göstermektedir.

### 4. Etnik Gruba Göre Dağılım:

- Beyaz grubun, diğer gruplara göre bilim haberlerine daha fazla ilgi gösterdiği gözlemlenmektedir. Siyah ve Hispanik grupların “aktif” oranları daha düşüktür.

### 5. Cevap Vermeyenler:

- Cevap vermeyenlerin sayısı oldukça düşük kalmış, bu da verilerin güvenilirliğini artırmaktadır.

## Sonuç

Bu analiz, hem kitap okuma kategorileri hem de etnik grupların bilim haberlerine olan ilgisini anlamak için çeşitli tanımlayıcı istatistikler kullanmaktadır.

- **Frekanslar:** Her kategorideki okuyucu sayısını göstermektedir.
- **Yüzdelere:** Belirli grupların toplam içindeki oranlarını belirtmektedir.
- **Mod:** En sık rastlanan kategoriyi ifade etmektedir.
- **Çapraz Tablo:** İki veya daha fazla değişken arasındaki ilişkileri ortaya koymaktadır.

## Örnek 2: Titanik'te Hayatta Kalma Oranları

```
# Titanic verisini okuma
titanic <- read.csv("https://raw.githubusercontent.com/bio304-class/bio304-course-notes/master/titanic.csv")

# Gerekli kütüphaneleri yükleyin
library(tidyverse)
```

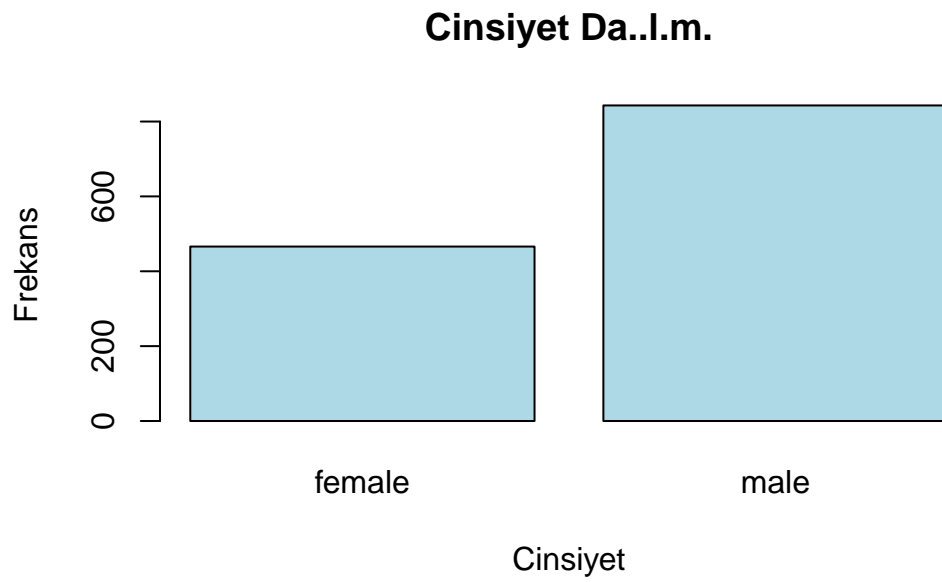
```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr      1.1.4      v readr      2.1.5
v forcats    1.0.0      v stringr    1.5.1
v ggplot2     3.5.1      v tibble     3.2.1
v lubridate  1.9.3      v tidyr      1.3.1
v purrr       1.0.2

-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()     masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# Cinsiyet dağılımı tablosu
table(titanic$sex)
```

```
female    male
   466     843
```

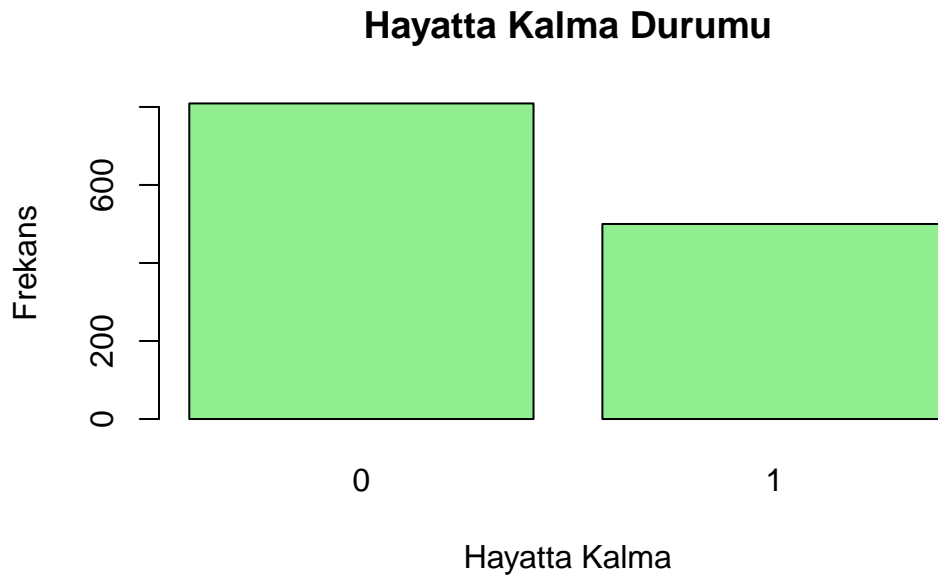
```
# Cinsiyet dağılımını gösteren çubuk grafiği
barplot(table(titanic$sex), main = "Cinsiyet Dağılımı", xlab = "Cinsiyet", ylab = "Frekans",
```



```
# Hayatta kalma durumu tablosu  
table(titanic$survived)
```

```
0    1  
809 500
```

```
# Hayatta kalma durumunu gösteren çubuk grafiği  
barplot(table(titanic$survived), main = "Hayatta Kalma Durumu", xlab = "Hayatta Kalma", ylab = "Frekans")
```



```
# Cinsiyet ve hayatta kalma durumu arasındaki tablo
table_1 <- table(titanic$sex, titanic$survived)

# Toplamları ekleyerek tabloyu gösterme
addmargins(table_1)
```

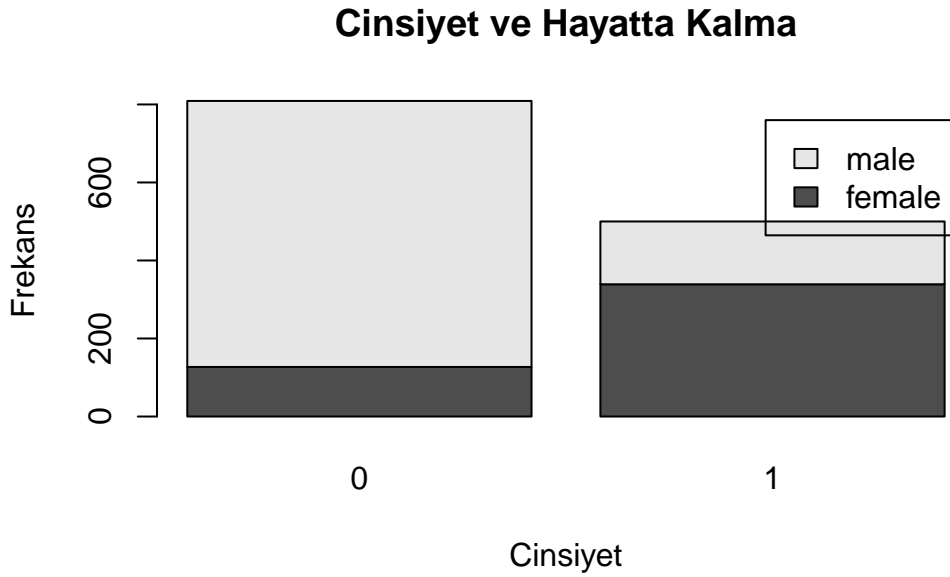
	0	1	Sum
female	127	339	466
male	682	161	843
Sum	809	500	1309

```
# Oran tablosunu gösterme
prop.table(table_1, margin = 2)
```

	0	1
female	0.1569839	0.6780000
male	0.8430161	0.3220000



```
# Cinsiyet ve hayatta kalma durumu için çubuk grafiği
barplot(table_1, legend.text = TRUE, main = "Cinsiyet ve Hayatta Kalma", xlab = "Cinsiyet", ylab = "Frekans")
```



```
# Sınıf ve hayatta kalma durumu arasındaki tablo
table_2 <- table(titanic$pclass, titanic$survived)

# Toplamları ekleyerek tabloyu gösterme
addmargins(table_2)
```

	0	1	Sum
1	123	200	323
2	158	119	277
3	528	181	709
Sum	809	500	1309

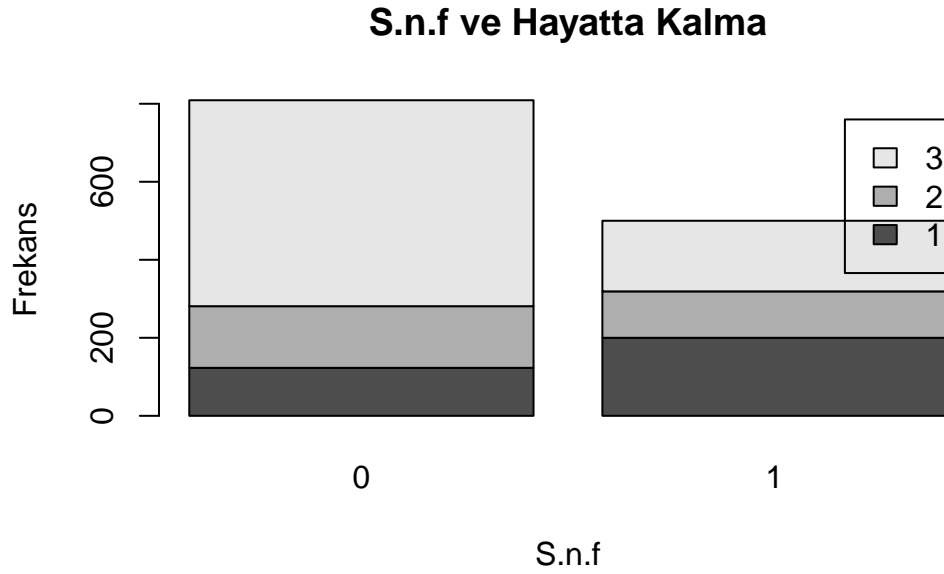
```
# Oran tablosunu gösterme
prop.table(table_2)
```

	0	1
1	0.152	0.250
2	0.195	0.148
3	0.653	0.222

```
1 0.09396486 0.15278839
2 0.12070283 0.09090909
3 0.40336134 0.13827349
```

```
# Sınıf ve hayatta kalma durumu için çubuk grafiği
```

```
barplot(table_2, legend.text = TRUE, main = "Sınıf ve Hayatta Kalma", xlab = "Sınıf", ylab =
```



```
# Yaş ve hayatta kalma durumu arasındaki tablo
table_3 <- table(titanic$age, titanic$survived)
```

```
# Toplamları ekleyerek tabloyu gösterme
addmargins(table_3)
```

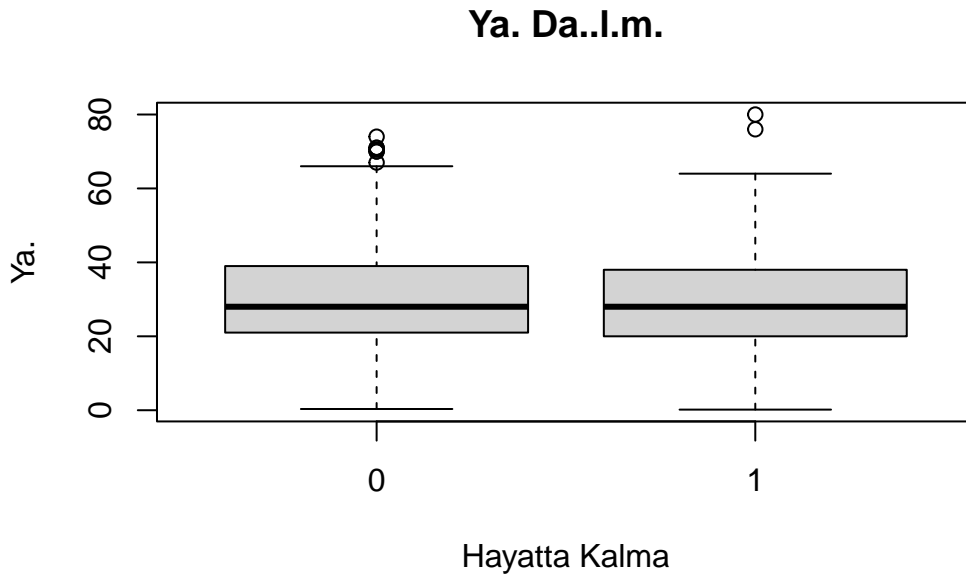
	0	1	Sum
0.1667	0	1	1
0.3333	1	0	1
0.4167	0	1	1
0.6667	0	1	1
0.75	1	2	3
0.8333	0	3	3
0.9167	0	2	2

1	3	7	10
2	8	4	12
3	2	5	7
4	3	7	10
5	1	4	5
6	3	3	6
7	2	2	4
8	2	4	6
9	6	4	10
10	4	0	4
11	3	1	4
11.5	1	0	1
12	0	3	3
13	2	3	5
14	4	4	8
14.5	2	0	2
15	1	5	6
16	11	8	19
17	13	7	20
18	25	14	39
18.5	3	0	3
19	18	11	29
20	15	8	23
20.5	1	0	1
21	30	11	41
22	23	20	43
22.5	1	0	1
23	16	10	26
23.5	1	0	1
24	25	22	47
24.5	1	0	1
25	23	11	34
26	19	11	30
26.5	1	0	1
27	17	13	30
28	24	8	32
28.5	3	0	3
29	17	13	30
30	25	15	40
30.5	2	0	2
31	11	12	23
32	13	11	24
32.5	3	1	4

33	12	9	21
34	10	6	16
34.5	2	0	2
35	10	13	23
36	17	14	31
36.5	1	1	2
37	7	2	9
38	8	6	14
38.5	1	0	1
39	12	8	20
40	12	6	18
40.5	3	0	3
41	9	2	11
42	12	6	18
43	6	3	9
44	7	3	10
45	7	14	21
45.5	2	0	2
46	6	0	6
47	11	3	14
48	4	10	14
49	4	5	9
50	9	6	15
51	5	3	8
52	3	3	6
53	0	4	4
54	5	5	10
55	4	4	8
55.5	1	0	1
56	2	2	4
57	5	0	5
58	2	4	6
59	2	1	3
60	3	4	7
60.5	1	0	1
61	5	0	5
62	3	2	5
63	2	2	4
64	3	2	5
65	3	0	3
66	1	0	1
67	1	0	1
70	2	0	2

70.5	1	0	1
71	2	0	2
74	1	0	1
76	0	1	1
80	0	1	1
Sum	619	427	1046

```
# Yaş dağılımını gösteren kutu grafiği
boxplot(titanic$age ~ titanic$survived, main = "Yaş Dağılımı", xlab = "Hayatta Kalma", ylab = "Yaş")
```



#### Yorumlar:

- **Grup 1:** 123 kişi ölmüş, 200 kişi hayatta kalmış. Toplamda 323 kişi.
- **Grup 2:** 158 kişi ölmüş, 119 kişi hayatta kalmış. Toplamda 277 kişi.
- **Grup 3:** 528 kişi ölmüş, 181 kişi hayatta kalmış. Toplamda 709 kişi.
- **Toplamlar:**
  - Tüm gruplarda toplam 1309 kişi gözlemlenmiştir.
  - Ölenlerin toplamı 809, hayatta kalanların toplamı ise 500'dür.
- **Hayatta Kalma Oranı:**

- En yüksek hayatta kalma sayısına sahip grup 1’dir (200 kişi hayatta), en yüksek ölüm sayısına sahip grup ise 3’tür (528 kişi ölmüş).

- **Hayatta Kalma Oranları:** Kadınların hayatta kalma oranı (67.8%) erkeklerin hayatta kalma oranından (32.2%) oldukça yüksektir. Bu, kadınların Titanic faciasında erkeklere göre daha yüksek bir hayatta kalma oranına sahip olduğunu göstermektedir.
- **Ölüm Oranları:** Erkekler için ölüm oranı (84.3%) oldukça yüksekken, kadınlar için bu oran çok daha düşüktür (15.7%). Bu durum, kadınların daha iyi korunmuş olabileceğini veya bazı sosyal faktörlerin etkisiyle hayatta kalma şanslarının artmış olabileceğini düşündürmektedir.

## Sayısal Değişkenlerle Kullanılabilen Tanımlayıcı İstatistikler

Tanımlayıcı istatistiklerde veriyi tanımlamak için sıkça kullanılan istatistiksel ölçümler şunlardır:

1. **Merkezi Eğilim Ölçüleri:** Ortalama, Medyan, Mod
2. **Merkezi Dağılım Ölçüleri:** Aralık, standart sapma, varyans
3. **Eğrilik ve Basıklık:** Dağılım grafiklerinin normal dağılımdan farklılaşması
4. **Korelasyon:** İki değişken arasındaki ilişkinin yönü ve gücü.

### 1. Merkezi Eğilim Ölçüleri:

#### 1.1 Ortalama (Mean):

- **Tanım:** Verilerin aritmetik ortalaması, tüm değerlerin toplamının gözlem sayısına bölünmesi ile hesaplanır.
- **Matematiksel Gösterim:**

$$\text{Ortalama} = \frac{\sum_{i=1}^n x_i}{n}$$

- **R Formülü:**

```
veri <- c(34, 67, 23, 45, 89, 12, 56, 78, 99, 5, 62, 48, 39, 75, 80, 22, 90, 11, 36, 50)
mean(veri) # veri, ortalamasını almak istediğiniz sayısal vektördür.
```

```
[1] 51.05
```

#### 1.2 Medyan (Median):

- **Tanım:** Verilerin sıralandıktan sonra ortadaki değeri. Özellikle aşırı değerlerin etkisini azaltır.
- **Matematiksel Gösterim:**
  - Eğer  $n$  tek ise:

$$\text{Medyan} = x_{(\frac{n+1}{2})}$$

- Eğer  $n$  çift ise:

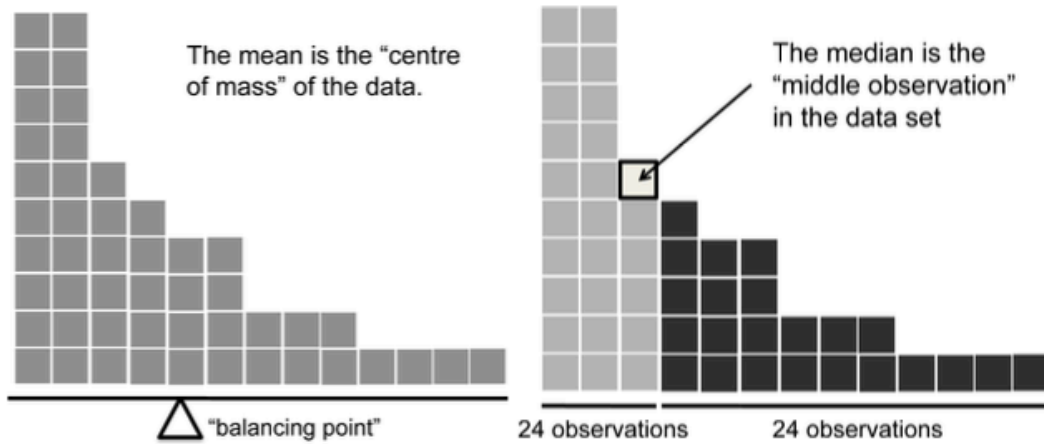
$$\text{Medyan} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

```
median(veri) # veri, medyanını almak istediğiniz sayısal vektördür.
```

[1] 49

#### **i** Note

Ortalama, veri setinin “ağırlık merkezi”dir: verilerin histogramını katı bir cisim olarak hayal ederseniz, onu dengeleyebileceğiniz nokta (bir tahterevalli gibi) ortalamadır. Buna karşılık, medyan ortadaki gözlemdir. Gözlemlerin yarısı daha küçüktür ve yarısı daha büyüktür.



### 1.3 Mod (Mode):

- **Tanım:** En sık rastlanan sayısal değer. Sayısal veriler için de kullanılabilir, ancak genellikle kategorik verilerle daha yaygındır.

- **R Formülü:**

```
mode_function <- function(x) {
  uniq_x <- unique(x)
  uniq_x[which.max(tabulate(match(x, uniq_x)))]
}
mode_function(veri) # veri, modunu almak istediğiniz sayısal vektördür.
```

[1] 34

## 2. Merkezi Dağılım Ölçüleri

### 2.1 Aralık (Range):

- **Matematiksel Gösterim:**

$$\text{Aralık} = \max(x) - \min(x)$$

- **R Formülü:**

```
range(veri) # veri, aralığını almak istediğiniz sayısal vektördür.
```

[1] 5 99

```
max(veri) - min(veri) # Aralığın hesaplanması
```

[1] 94

### 2.2 Standart Sapma (Standard Deviation):

- **Tanım:** Verilerin ortalamadan ne kadar yayıldığını gösterir. Verilerin ne kadar değişken olduğunu ölçer.
- **Matematiksel Gösterim:**

$$\text{Standart Sapma} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

- **R Formülü**



```
sd(veri) # veri, standart sapmasını almak istediğiniz sayısal vektördür.
```

```
[1] 28.45398
```

### 2.3 Varyans (Variance):

- **Tanım:** Verilerin ortalamadan ne kadar saptığını gösteren bir ölçüdür. Standart sapmanın karesidir. Veriler arasındaki dağılımın ne kadar farklı olduğunu anlamamıza yardımcı olur. **Matematiksel Gösterim:**

$$\text{Varyans} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

- **R Formülü**

```
var(veri) # veri, varyansını almak istediğiniz sayısal vektördür.
```

```
[1] 809.6289
```

Population	Sample
$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$ <p>X - The Value in the data distribution <math>\mu</math> - The population Mean N - Total Number of Observations</p>	$s = \sqrt{\frac{\sum (X - \bar{x})^2}{n - 1}}$ <p>X - The Value in the data distribution <math>\bar{x}</math> - The Sample Mean n - Total Number of Observations</p>

$x_1$	$x_1 - \bar{x}$	$(x_1 - \bar{x})^2$
2	-1.5	2.25
6	2.5	6.25
5	1.5	2.25
3	-0.5	0.25
2	-1.5	2.25
3	-0.5	0.25
<b>= 13.5</b>		

#### 2.4 Çeyrekler Aralığı (Interquartile Range, IQR):

- **Tanım:** Verilerin ortasındaki yarısını kapsayan yayılımı gösterir. 1. çeyrek (Q1) ile 3. çeyrek (Q3) arasındaki farktır.

- **Matematiksel Gösterim:**

$$IQR = Q3 - Q1$$

- **R Formülü**

```
IQR(veri) # veri, çeyrekler aralığını almak istediğiniz sayısal vektördür.
```

```
[1] 44.5
```

#### 2.5 Küçük ve Büyük Çeyrekler (Quantiles):

- **Tanım:** Verilerin belirli bir yüzdesine karşılık gelen değerlerdir. Örneğin, %25'lik çeyrek (Q1) ve %75'lik çeyrek (Q3).

- **R Formülü:**

```
quantile(veri, probs = c(0.25, 0.75)) # %25 ve %75'lik çeyrekler
```

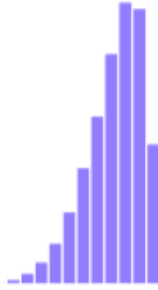
```
25%    75%
31.25  75.75
```

### 3. Dağılımda Sapma: Eğrilik ve Basıklık

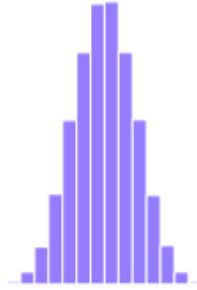
#### 3.1 Skewness (Eğrilik):

- **Tanım:** Verinin dağılımının simetrik olup olmadığını gösterir. Pozitif veya negatif eğrilik değerleri, veri dağılımının sağa veya sola kaydığını belirtir.

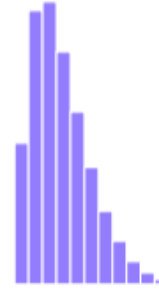
Negative Skew



No Skew

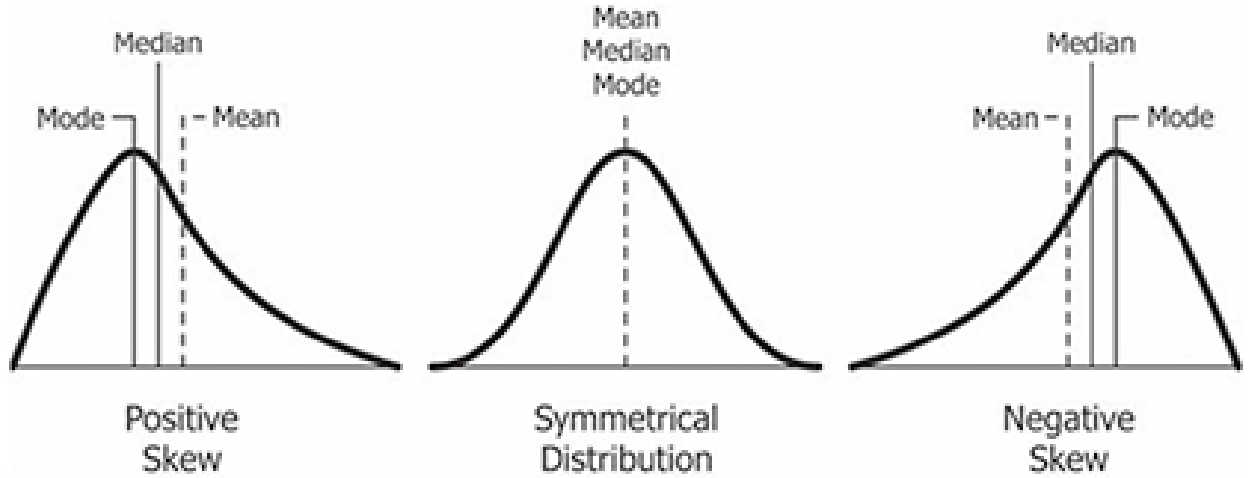


Positive Skew



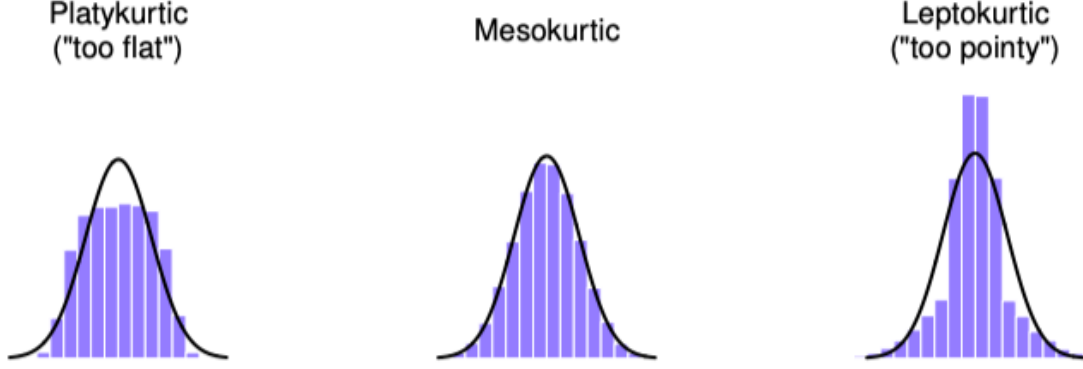
#### **i** Note

Hertür eğrilikte Median daha merkezi bir tanımlayıcı istatistiğe dönüşür.



#### 3.2 Kurtosis (Basıklık):

- **Tanım:** Verinin dağılımının ne kadar “keskin” veya “düz” olduğunu ölçer. Yüksek kurtosis, veri dağılımının keskin zirvelere sahip olduğunu gösterirken, düşük kurtosis daha düz bir dağılımı işaret eder.



#### **i** Note

Ortalama, Medyan, Mod, IQR ve Aralık (Range) Kullanım Rehberi  
Veri analizi sırasında, ortalama, medyan, mod, çeyrekler aralığı (IQR) ve aralık (range) gibi tanımlayıcı istatistikleri kullanarak verilerimizi özetleyebiliriz. Ancak, her bir istatistiğin kullanımı, veri setinin özelliklerine bağlıdır. Aşağıda, bu istatistiklerin hangi durumlarda kullanılacağına dair detaylı bir rehber sunulmaktadır.

### 1. Ortalama (Mean)

- **Kullanım Durumu:** Sayısal verilerin merkezi eğilimini ölçmek için kullanılır.
- **Uygun Durumlar:**
  - Veriler normal dağılım gösteriyorsa.
  - Aşırı değerler (outlier) yoksa.
- **Örnek:** Bir sınıfın notları 70, 75, 80, 85, 90 olduğunda ortalama 80'dir. Ancak, eğer sınıfta 0 alan bir öğrenci varsa (notlar: 0, 70, 75, 80, 85, 90), ortalama 50 olur ki bu yanıltıcıdır.

### 2. Medyan (Median)

- **Kullanım Durumu:** Veri setinin ortasındaki değeri bulmak için kullanılır.

- **Uygun Durumlar:**

- Veriler asimetrik dağılım gösteriyorsa.
- Aşırı değerler mevcutsa, çünkü medyan bu değerlerden etkilenmez.

- **Örnek:** Bir ev fiyatları seti 100.000, 120.000, 130.000, 1.000.000, 1.200.000 olduğunda, ortalama fiyat 510.000, ancak medyan 130.000'dir. Medyan, veri setinin çoğunluğunu daha iyi temsil eder.

### 3. Mod (Mode)

- **Kullanım Durumu:** En sık rastlanan değeri bulmak için kullanılır.

- **Uygun Durumlar:**

- Kategorik verilerde kullanılır.
- Verilerin birden fazla mod değeri (bimodal veya multimodal) varsa.

- **Örnek:** Bir anket sonucu olarak, katılımcıların tercih ettiği meyveler: Elma, Muz, Elma, Portakal, Elma, Muz. Burada mod "Elma"dır çünkü en sık rastlanan değerdir.

### 4. Çeyrekler Aralığı (IQR)

- **Kullanım Durumu:** Verilerin yayılımını ölçmek için kullanılır.

- **Uygun Durumlar:**

- Verilerin ortasındaki yayılımı anlamak için.
- Aşırı değerlere karşı dayanıklıdır, bu yüzden özellikle asimetrik dağılımlarda tercih edilir.

- **Örnek:** Bir test notları setinde, 60, 70, 75, 80, 85, 90, 95, 100 değerleri mevcutsa, IQR ( $Q3 - Q1$ ) hesaplanarak veri setinin ortasında yer alan yayılım gösterilir.  $Q1 = 72.5$  ve  $Q3 = 90$  ise,  $IQR = 90 - 72.5 = 17.5$  olacaktır.

### 5. Aralık (Range)

- **Kullanım Durumu:** Verilerin en küçük ve en büyük değeri arasındaki farkı bulmak için kullanılır.

- **Uygun Durumlar:**

- Veri setinin yayılımını genel olarak anlamak için.
- Aşırı değerlere dikkat edilmelidir, çünkü aralık bu değerlerden etkilenir.

- **Örnek:** Bir grup öğrencinin notları 50, 60, 70, 80, 90 olduğunda, aralık  $90 - 50 = 40$ 'tır. Ancak, 0 alan bir öğrenci eklenirse (notlar: 0, 50, 60, 70, 80, 90), aralık  $90 - 0 = 90$  olacaktır.

### Hangi İstatistiğin Kullanılacağına Dair Karar Verirken Dikkat Edilmesi Gerekenler

- **Veri Dağılımı:** Verilerin normal mi yoksa asimetrik mi dağıldığına bakın. Normal dağılımda ortalama kullanılabilirken, asimetrik dağılımda medyan daha iyi bir temsil sunar.
- **Aşırı Değerler:** Veri setinde aşırı değerlerin (outlier) olup olmadığına dikkat edin. Aşırı değerlerin varlığında medyan ve IQR daha güvenilir ölçümler sağlar.
- **Veri Türü:** Verilerin sayısal mı yoksa kategorik mi olduğunu belirleyin. Kategorik veriler için mod, sayısal veriler için ortalama ve medyan daha uygundur.
- **Analiz Amacı:** Hangi bilgiyi elde etmek istediğinizi düşünün. Eğer yayılımı ölçmek istiyorsanız, IQR veya aralık kullanın; merkezi eğilimi ölçmek istiyorsanız, ortalama veya medyan tercih edin.

### Örnek 3: Avustralya Futbol Ligi

```
library(here)
```

here() starts at /Users/kobain/Desktop/IST2083

```
load(here("data", "aflsmall.Rdata"))
```

# Bu iki veri, Avustralya futbol ligi ile ilgilidir. afl.margins, 176 oyunun kazanç farkını

table(afl.finalists) # Bu, bir vektördeki bir girişin kaç kez görüldüğünü hesaplamak için ta

```
afl.finalists
```

Adelaide	Brisbane	Carlton	Collingwood
26	25	26	28
Essendon	Fitzroy	Fremantle	Geelong
32	0	6	39

Hawthorn	Melbourne	North Melbourne	Port Adelaide
27	28	28	17
Richmond	St Kilda	Sydney	West Coast
6	24	26	38
Western Bulldogs			
24			

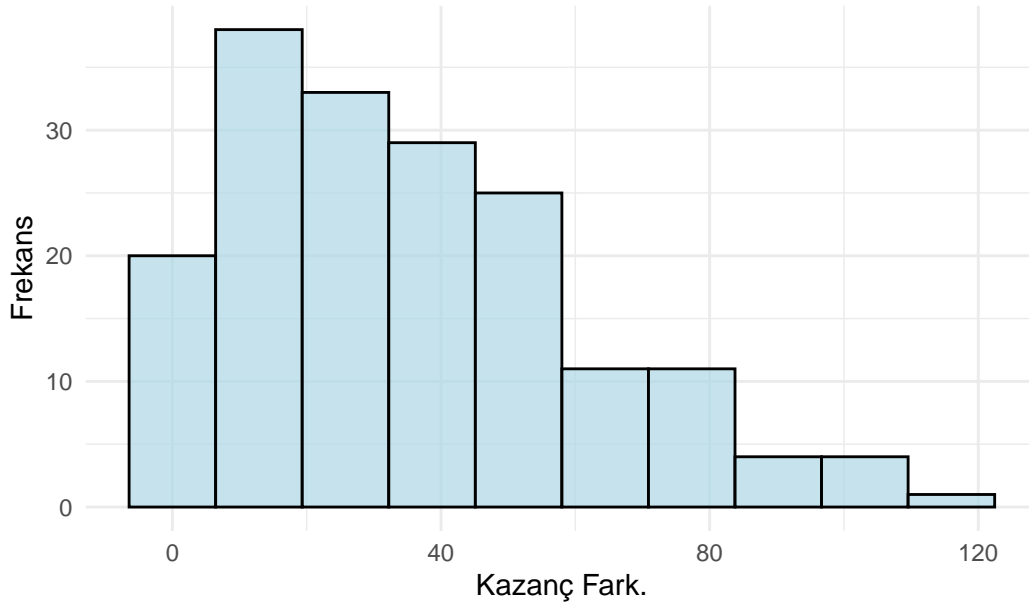
```
print(afl.margins)
```

```
[1] 56 31 56 8 32 14 36 56 19 1 3 104 43 44 72 9 28 25
[19] 27 55 20 16 16 7 23 40 48 64 22 55 95 15 49 52 50 10
[37] 65 12 39 36 3 26 23 20 43 108 53 38 4 8 3 13 66 67
[55] 50 61 36 38 29 9 81 3 26 12 36 37 70 1 35 12 50 35
[73] 9 54 47 8 47 2 29 61 38 41 23 24 1 9 11 10 29 47
[91] 71 38 49 65 18 0 16 9 19 36 60 24 25 44 55 3 57 83
[109] 84 35 4 35 26 22 2 14 19 30 19 68 11 75 48 32 36 39
[127] 50 11 0 63 82 26 3 82 73 19 33 48 8 10 53 20 71 75
[145] 76 54 44 5 22 94 29 8 98 9 89 1 101 7 21 52 42 21
[163] 116 3 44 29 27 16 6 44 3 28 38 29 10 10
```

Verileri tanımlamak için bir histogram ile başlayalım. Bu, tanımlamaya çalıştığımız verilerin nasıl görüldüğüne dair bir fikir edinmenizi sağlayacaktır.

```
# Histogram oluşturma
ggplot(data = data.frame(margins = afl.margins), aes(x = margins)) +
  geom_histogram(bins = 10, fill = "lightblue", color = "black", alpha = 0.7) +
  labs(title = "Kazanç Farklarının Histogramı", x = "Kazanç Farkı", y = "Frekans") +
  theme_minimal()
```

Kazanç Farklar.n.n Histogram.



```
# Tanımlayıcı istatistikleri hesaplama
ortalama <- mean(afl.margins)
medyan <- median(afl.margins)

# Mod fonksiyonu tanımlama
mode_function <- function(x) {
  uniq_x <- unique(x)
  uniq_x[which.max(tabulate(match(x, uniq_x)))]
}
mod <- mode_function(afl.margins)

varyans <- var(afl.margins)
standart_sapma <- sd(afl.margins)
IQR_val <- IQR(afl.margins)

# Sonuçları görüntüleme
result <- data.frame(
  İstatistik = c("Ortalama", "Medyan", "Mod", "Varyans", "Standart Sapma", "Çeyrekler Aralığı"),
  Değer = c(ortalama, medyan, mod, varyans, standart_sapma, IQR_val)
)

print(result)
```



	İstatistik	Değer
1	Ortalama	35.30114
2	Medyan	30.50000
3	Mod	3.00000
4	Varyans	679.83451
5	Standart Sapma	26.07364
6	Çeyrekler Aralığı	37.75000

#### 4. Değişkenler Arası İlişki: Korelasyon

Korelasyon, iki sayısal değişkenin doğrusal olarak ne derece ilişkili olduğunu ifade eden istatistiksel bir ölçüttür (yani, iki değişkenin birlikte sabit bir oranda değişip değişmediğini gösterir). Korelasyon, basit ilişkileri tanımlamak için sıkça kullanılır ancak bu ilişki hakkında bir nedensellik iddiasında bulunmaz; yalnızca iki değişkenin nasıl birlikte hareket ettiğini gösterir.

Örneğin, sıcaklık arttıkça dondurma satışlarının da artması bir korelasyon gösterebilir, ancak bu dondurma satışlarının artmasının sebebinin sıcaklık olduğunu kesin olarak kanıtlamaz. Korelasyon, genellikle -1 ile +1 arasında değişen bir korelasyon katsayısı ( $r$ ) ile ifade edilir.

- **Pozitif Korelasyon ( $r > 0$ ):** Bir değişken arttıkça diğer değişken de artar. Örneğin, eğitim seviyesi ile gelir arasındaki ilişki genellikle pozitif korelasyon gösterir.
- **Negatif Korelasyon ( $r < 0$ ):** Bir değişken arttıkça diğer değişken azalır. Örneğin, bir araç ne kadar hızlı yaparsa gidebileceği mesafe o kadar azalır.
- **Korelasyon Yok ( $r = 0$ ):** İki değişken arasında anlamlı bir doğrusal ilişki yoktur.

#### Örnek 4: Ebeveynlik

Korelasyon hakkında daha ayrıntılı inceleme yapmak için yeni bir veri seti kullanarak örnekleme yapalım:

```
# veriyi yükle
load(here("data", "parenthood.Rdata"))

# veriye göz at
str(parenthood)
```

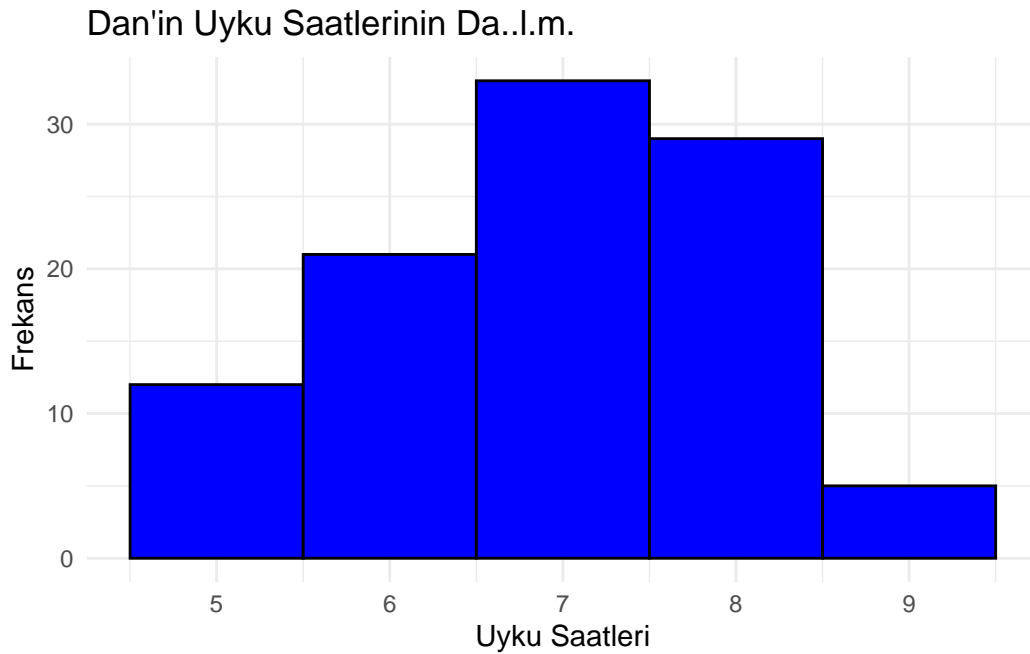
```
'data.frame': 100 obs. of 4 variables:
 $ dan.sleep : num 7.59 7.91 5.14 7.71 6.68 5.99 8.19 7.19 7.4 6.58 ...
 $ baby.sleep: num 10.18 11.66 7.92 9.61 9.75 ...
 $ dan.grump : num 56 60 82 55 67 72 53 60 60 71 ...
 $ day : int 1 2 3 4 5 6 7 8 9 10 ...
```

Öncelikle, `summary()` fonksiyonu ile veri setinin özetini alabiliriz. Bu, her bir değişken için temel tanımlayıcı istatistikleri (ortalama, medyan, minimum, maksimum, çeyrekler gibi) sağlar. Ardından, bir değişkenin dağılımını görselleştirmek için `ggplot2` kütüphanesini kullanarak histogram çizebiliriz.

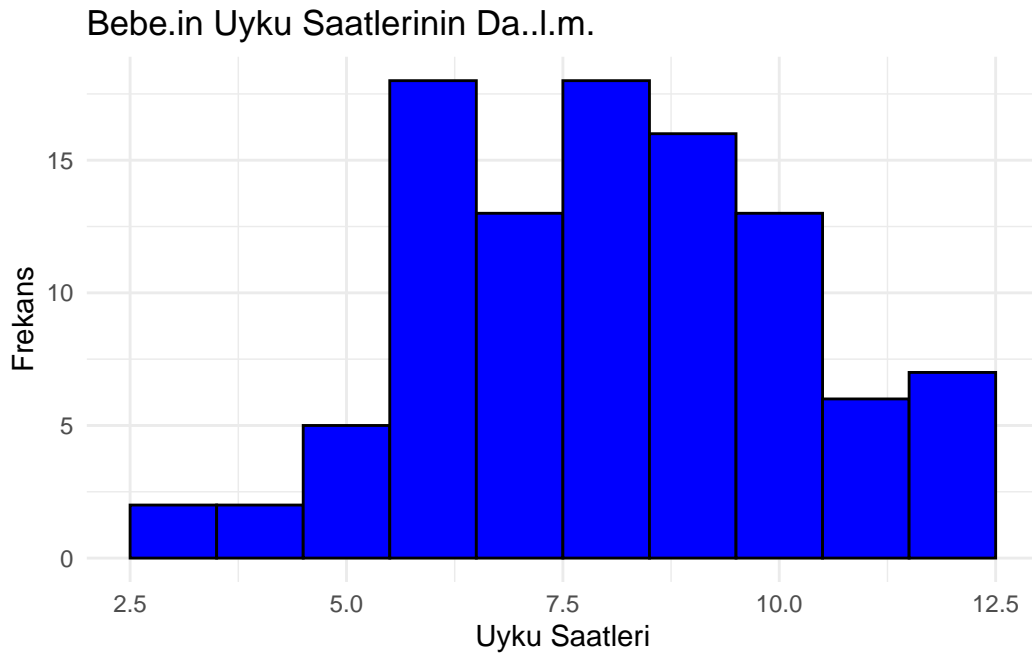
```
# Veri setinin özetini al
summary(parenthood)
```

dan.sleep	baby.sleep	dan.grump	day
Min. :4.840	Min. : 3.250	Min. :41.00	Min. : 1.00
1st Qu.:6.293	1st Qu.: 6.425	1st Qu.:57.00	1st Qu.: 25.75
Median :7.030	Median : 7.950	Median :62.00	Median : 50.50
Mean :6.965	Mean : 8.049	Mean :63.71	Mean : 50.50
3rd Qu.:7.740	3rd Qu.: 9.635	3rd Qu.:71.00	3rd Qu.: 75.25
Max. :9.000	Max. :12.070	Max. :91.00	Max. :100.00

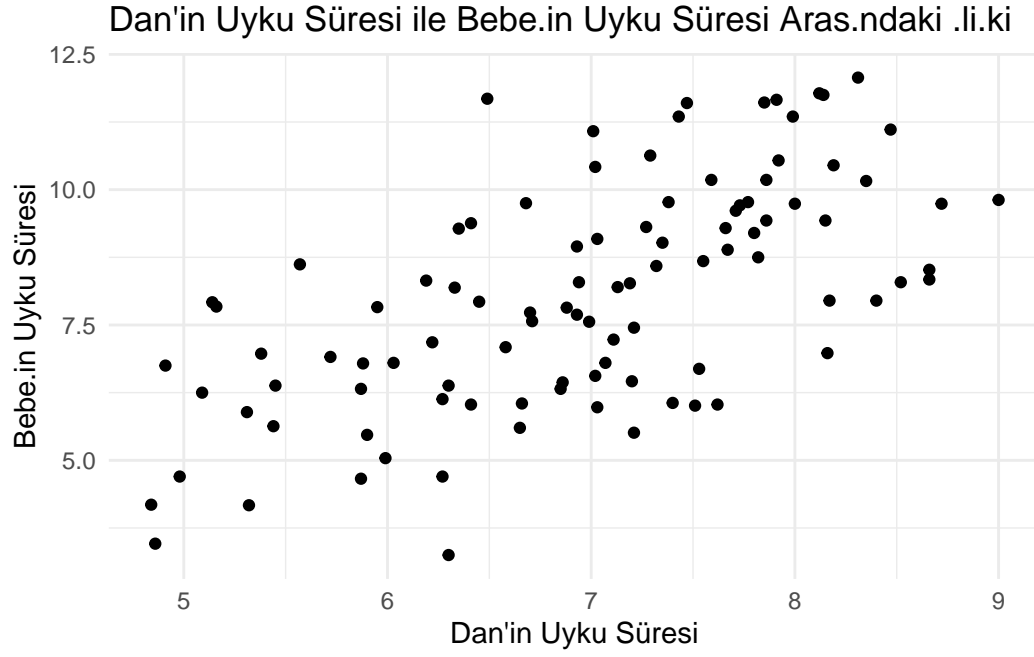
```
# Dan'in uyku süresi dağılımı histogramı
ggplot(data=parenthood, aes(x=dan.sleep)) +
  geom_histogram(binwidth=1, fill="blue", color="black") +
  labs(title="Dan'in Uyku Saatlerinin Dağılımı", x="Uyku Saatleri", y="Frekans") +
  theme_minimal()
```



```
# Bebeğin uyku süresi dağılımı histogramı
ggplot(data=parenthood, aes(x=baby.sleep)) +
  geom_histogram(binwidth=1, fill="blue", color="black") +
  labs(title="Bebeğin Uyku Saatlerinin Dağılımı", x="Uyku Saatleri", y="Frekans") +
  theme_minimal()
```

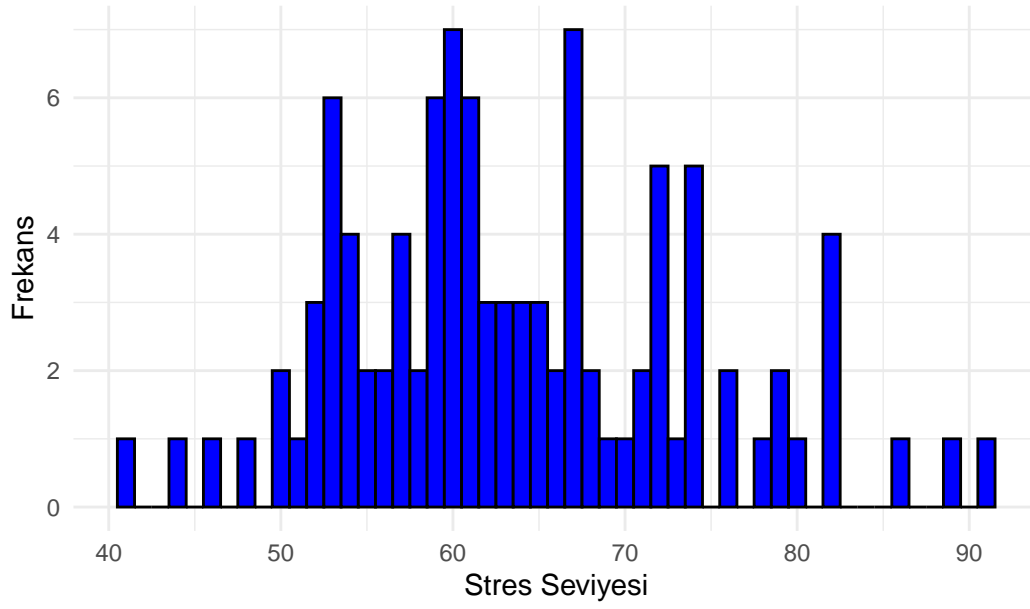


```
# Dan'in uyku süresi ile bebeğin uyku süresi arasındaki ilişki (korelasyon)
ggplot(data=parenthood, aes(x=dan.sleep, y=baby.sleep)) +
  geom_point() +
  labs(title="Dan'in Uyku Süresi ile Bebeğin Uyku Süresi Arasındaki İlişki", x="Dan'in Uyku Süresi", y="Bebeğin Uyku Süresi") +
  theme_minimal()
```



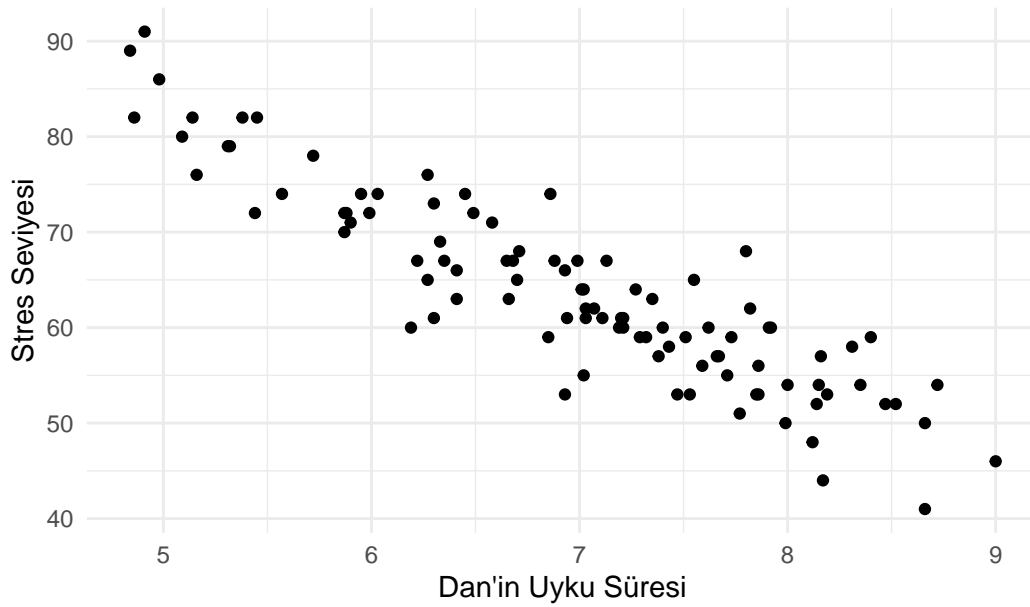
```
# Dan'in stres (grumpiness) seviyesinin dağılımı histogramı
ggplot(data=parenthood, aes(x=dan.grump)) +
  geom_histogram(binwidth=1, fill="blue", color="black") +
  labs(title="Dan'in Stres Seviyesinin Dağılımı", x="Stres Seviyesi", y="Frekans") +
  theme_minimal()
```

### Dan'in Stres Seviyesinin Dağılımı



```
# Dan'in uyku süresi ile stres seviyesi arasındaki ilişki (korelasyon)
ggplot(data=parenthood, aes(x=dan.sleep, y=dan.grump)) +
  geom_point() +
  labs(title="Dan'in Uyku Süresi ile Stres Seviyesi Arasındaki İlişki", x="Dan'in Uyku Süresi", y="Dan'in Stres Seviyesi")
theme_minimal()
```

### Dan'in Uyku Süresi ile Stres Seviyesi Arasındaki İlişki



```
# parenthood veri setindeki tüm değişkenler arasındaki korelasyonları hesapla
korelasyon_matrisi <- cor(parenthood)

# Korelasyon matrisini yazdır
print(korelasyon_matrisi)
```

	dan.sleep	baby.sleep	dan.grump	day
dan.sleep	1.00000000	0.62794934	-0.90338404	-0.09840768
baby.sleep	0.62794934	1.00000000	-0.56596373	-0.01043394
dan.grump	-0.90338404	-0.56596373	1.00000000	0.07647926
day	-0.09840768	-0.01043394	0.07647926	1.00000000

#### Note



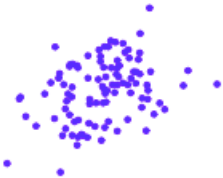



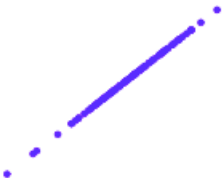

##### **Korelasyon Katsayısı:**

Korelasyon katsayısı, iki değişken arasındaki doğrusal ilişkinin yönünü ve gücünü ölçen bir istatistiksel ölçüdür. -1 ile +1 arasında değerler alır.

- **+1:** Mükemmel pozitif korelasyon. Bir değişken arttıkça, diğer değişken de aynı oranda artar. Görselde en alttaki sol grafikte görüldüğü gibi, tüm noktalar yükselen bir doğru üzerinde yer alır.
- **+0.66:** Güçlü pozitif korelasyon. Bir değişken arttıkça, diğer değişken de genellikle

artar, ancak mükemmel bir ilişki yoktur. Görselde ortadaki sol grafikte görüldüğü gibi, noktalar yükselen bir doğru etrafında kümelenmiştir, ancak doğrudan üzerinde değildir.

- **+0.33:** Zayıf pozitif korelasyon. Bir değişken arttıkça, diğer değişkenin artma eğilimi vardır, ancak ilişki daha az belirgindir. Görselde üstteki sol grafikte görüldüğü gibi, noktalar dağınık bir şekilde yükselen bir trend gösterir.
- **0:** Korelasyon yok. Değişkenler arasında doğrusal bir ilişki yoktur. Değişkenlerden birinin değeri değiştiğinde, diğerinin değeri üzerinde tahmin edilebilir bir etkisi olmaz.
- **-0.33:** Zayıf negatif korelasyon. Bir değişken arttıkça, diğer değişkenin azalma eğilimi vardır. Görselde üstteki sağ grafikte görüldüğü gibi, noktalar dağınık bir şekilde azalan bir trend gösterir.
- **-0.66:** Güçlü negatif korelasyon. Bir değişken arttıkça, diğer değişken genellikle azalır. Görselde ortadaki sağ grafikte görüldüğü gibi, noktalar azalan bir doğru etrafında kümelenmiştir.
- **-1:** Mükemmel negatif korelasyon. Bir değişken arttıkça, diğer değişken aynı oranda azalır. Görselde en alttaki sağ grafikte görüldüğü gibi, tüm noktalar azalan bir doğru üzerinde yer alır.

positive correlations		negative correlations	
correlation	example	correlation	example
0.0		0.0	
0.33		-0.33	
0.66		-0.66	
1.0		-1.0	



## Korelasyon bir nedensellik değildir

Korelasyon, iki değişken arasındaki ilişkinin yönünü ve gücünü ifade eder. Pozitif korelasyon, bir değişken arttığında diğerinin de artma eğiliminde olduğunu, negatif korelasyon ise bir değişken arttığında diğerinin azalma eğiliminde olduğunu gösterir.

**Önemli nokta:** Korelasyon, nedensellik anlamına gelmez. Yani, iki değişken arasında korelasyon olması, birinin diğerine neden olduğu anlamına gelmez.

**Örnek:** Dondurma satışları ile denizde boğulma vakaları arasında pozitif bir korelasyon vardır. Yani, dondurma satışları arttığında, boğulma vakaları da artar. Ancak bu, dondurma yemek insanların boğulmasına neden oluyor anlamına gelmez. Aslında, her iki değişken de sıcak hava gibi üçüncü bir faktörden etkilenir. Sıcak havalarda insanlar daha çok dondurma yerler ve denize girerler, bu da boğulma vakalarının artmasına neden olur.

**Başka bir örnek:** Ayakkabı numarası ile okuma becerileri arasında pozitif bir korelasyon olabilir. Yani, ayakkabı numarası büyük olan çocuklar genellikle daha iyi okuma becerilerine sahiptir. Ancak bu, büyük ayakların çocukların daha iyi okumasına neden olduğu anlamına gelmez. Aslında, her iki değişken de yaş gibi üçüncü bir faktörden etkilenir. Çocuklar büyüdükçe ayakları büyür ve okuma becerileri gelişir.

**Sonuç olarak:** Korelasyon, iki değişken arasında bir ilişki olduğunu gösterir, ancak bu ilişkinin nedensel olup olmadığını belirlemek için daha fazla araştırma yapmak gerekir.

```
# "here" fonksiyonunu kullanabilmek için "here" paketini yükle
if (!require("here")) install.packages("here")

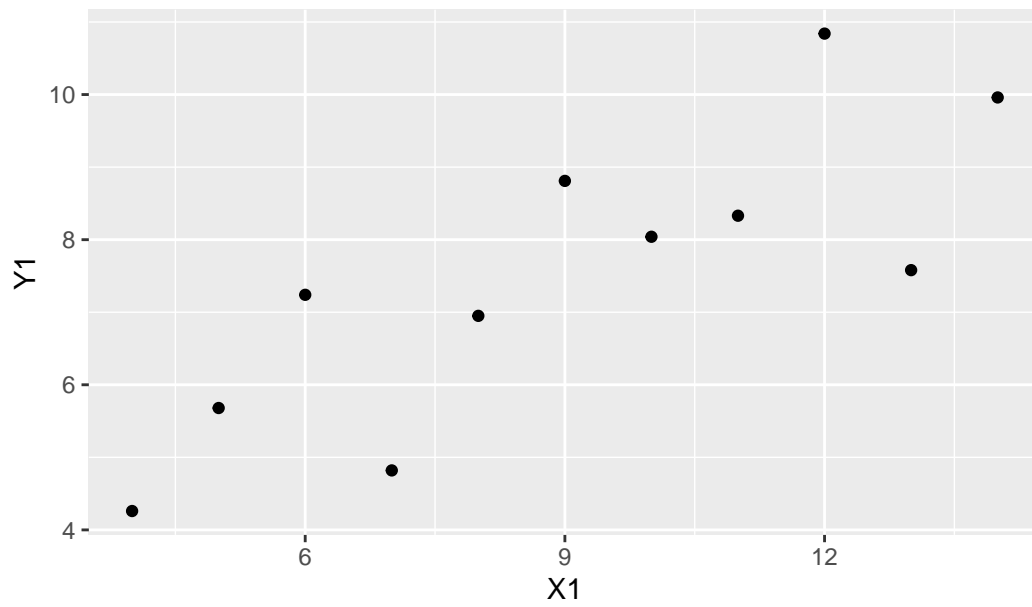
# Gerekli kütüphaneleri yükle
library(here)
library(ggplot2)

# Veri setini yükle
load(here("data", "anscombesquartet.Rdata"))

# X1, Y1, X2, Y2 vb. vektörlerini kullanarak bir data.frame oluştur
anscombesquartet <- data.frame(X1, Y1, X2, Y2, X3, Y3, X4, Y4)

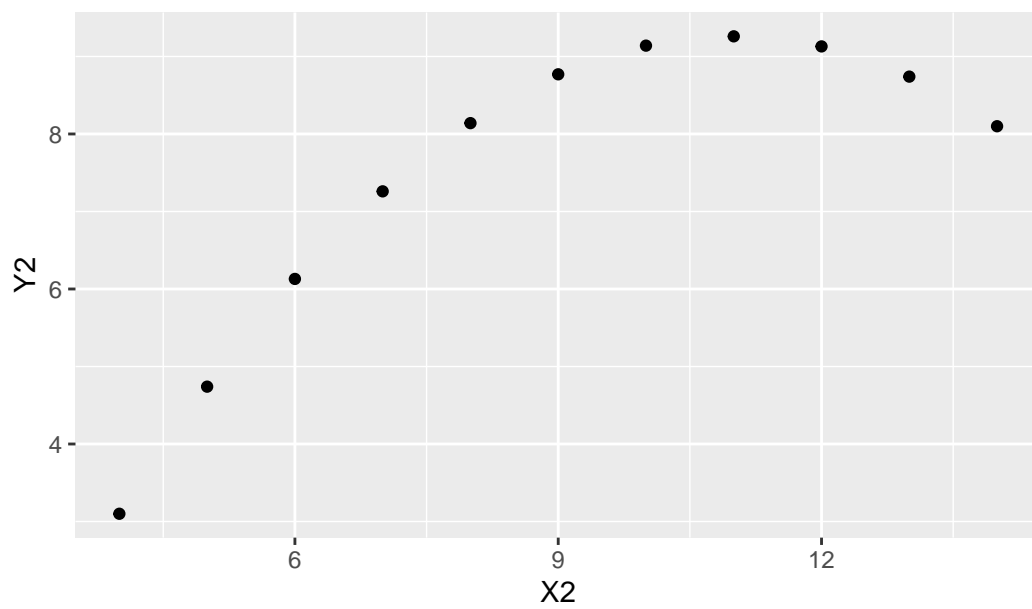
# Farklı ikililer arası scatter plotlar oluştur
ggplot(anscombesquartet, aes(x = X1, y = Y1)) +
  geom_point() +
  ggtitle("X1 ve Y1")
```

X1 ve Y1

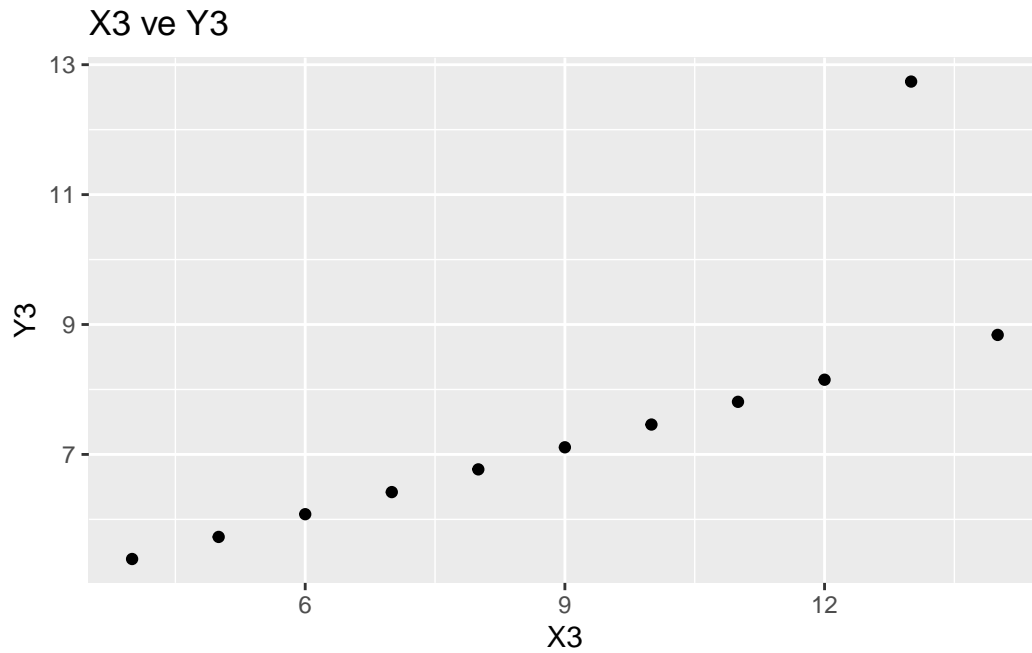


```
ggplot(anscombesquartet, aes(x = X2, y = Y2)) +
  geom_point() +
  ggtitle("X2 ve Y2")
```

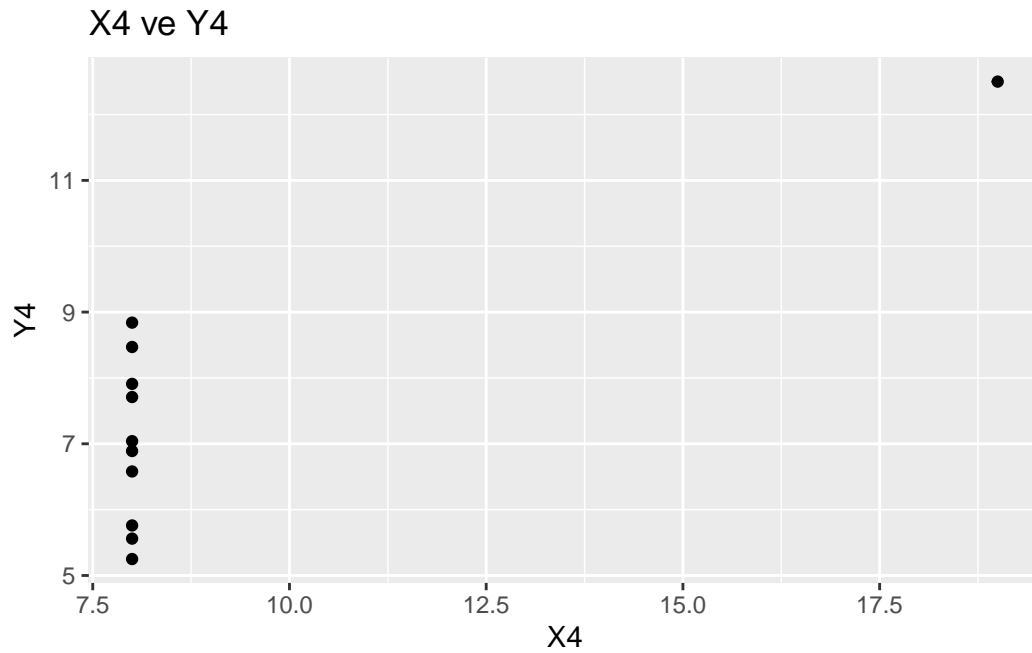
X2 ve Y2



```
ggplot(anscombesquartet, aes(x = X3, y = Y3)) +  
  geom_point() +  
  ggtitle("X3 ve Y3")
```



```
ggplot(anscombesquartet, aes(x = X4, y = Y4)) +  
  geom_point() +  
  ggtitle("X4 ve Y4")
```



```
# Korelasyon katsayılarını hesapla  
cor(anscombesquartetet$X1, anscombesquartetet$Y1)
```

```
[1] 0.8164205
```

```
cor(anscombesquartetet$X2, anscombesquartetet$Y2)
```

```
[1] 0.8162365
```

```
cor(anscombesquartetet$X3, anscombesquartetet$Y3)
```

```
[1] 0.8162867
```

```
cor(anscombesquartetet$X4, anscombesquartetet$Y4)
```

```
[1] 0.8165214
```

#### **i** Note

##### **Anscombe's Quartet veri seti neden önemli?**

Anscombe's Quartet, korelasyon katsayısının tek başına bir ilişkinin tam resmini veremeyeceğini gösteren önemli bir örnektir. Dört veri kümesinin de korelasyon katsayıları neredeyse aynıdır, ancak grafiklerine baktığımızda çok farklı ilişkiler olduğunu görürüz:

- X1 ve Y1: Doğrusal bir ilişki var gibi görünüyor.
- X2 ve Y2: Doğrusal bir ilişki yoktur, parabolik bir ilişki vardır.
- X3 ve Y3: Doğrusal bir ilişki vardır, ancak bir aykırı değer bu ilişkiyi etkilemektedir.
- X4 ve Y4: X4 neredeyse sabittir, tek bir aykırı değer Y4 ile yüksek bir korelasyon göstermektedir.

Bu nedenle, sadece korelasyon katsayısına bakmak yerine, verileri görselleştirmek ve ilişkinin doğasını anlamak çok önemlidir.

#### **i** Note

##### **cor() Fonksiyonunun method Parametresi**

R'daki `cor()` fonksiyonu, varsayılan olarak Pearson korelasyon katsayısını hesaplar. Ancak, farklı korelasyon türleri hesaplamak isteyebilirsiniz. İşte `method` parametresi ile kullanabileceğiniz seçenekler:

- **"pearson"**: Pearson korelasyon katsayısı (varsayılan). İki değişken arasındaki doğrusal ilişkinin gücünü ölçer.
- **"spearman"**: Spearman sıra korelasyon katsayısı. Değişkenlerin sıraları arasındaki korelasyonu ölçer. Doğrusal olmayan, monotonik ilişkilere duyarlıdır.
- **"kendall"**: Kendall korelasyon katsayısı. İki değişken arasındaki uyum (concordance) sayısını temel alır. Aykırı değerlere karşı daha dayanıklıdır.

Örnek olarak, Spearman sıra korelasyonunu hesaplamak için `cor()` fonksiyonunu şu şekilde kullanabilirsiniz:

```
cor(x, y, method = "spearman")
```

##### **Eksik Verilerle Başa Çıkma**

Veri setlerinde eksik veriler olması yaygın bir durumdur. `cor()` fonksiyonu, eksik verilerle nasıl başa çıkacağını belirlemek için `use` parametresini kullanır. İşte `use` parametresi ile kullanabileceğiniz seçenekler:

- **"everything"**: Tüm gözlemleri kullanır. Eksik veri içeren herhangi bir çift, NA (Not Available) değeri üretir.

- “**all.obs**”: Herhangi bir eksik veri varsa hata verir.
- “**complete.obs**”: Eksik veri içeren gözlemleri tamamen çıkarır ve sadece tam gözlemleri kullanarak korelasyonu hesaplar.
- “**na.or.complete**”: “complete.obs” ile aynıdır.
- “**pairwise.complete.obs**”: Her bir değişken çifti için, eksik veri içeren gözlemleri çıkarır ve kalan gözlemlerle korelasyonu hesaplar.

Örnek olarak, eksik verileri olan gözlemleri çıkararak Pearson korelasyonunu hesaplamak için `cor()` fonksiyonunu şu şekilde kullanabilirsiniz:

```
cor(x, y, use = "complete.obs")
```

Eksik verilerle nasıl başa çıkılacağı, veri setinin özelliklerine ve analiz amacına bağlıdır. Eksik verileri yok saymak veya yanlış yöntemlerle doldurmak, yanıltıcı sonuçlara yol açabilir. Bu nedenle, eksik verileri dikkatli bir şekilde ele almak önemlidir.

## Özet

Gerçek verileri analiz ederken yapacağınız ilk şeylerden biri temel betimsel istatistikleri hesaplamaktır ve betimsel istatistikler, çıkarımsal istatistiklerden çok daha kolay anlaşılır. Bu derste, aşağıdaki konuları ele aldık:

**Merkezi eğilim ölçüleri:** Bu ölçüler, veri setinin tipik veya merkezi bir değerini temsil etmeye çalışır. Ortalama, tüm değerlerin toplamının değer sayısına bölünmesiyle elde edilir. Medyan, sıralanmış veri setinin orta değeridir. Mod ise en sık görünen değerdir. Hangi merkezi eğilim ölçüsünün kullanılacağı, verilerin dağılımına ve analiz amacına bağlıdır.

**Merkezi dağılım ölçüleri:** Bu ölçüler, verilerin yayılımını veya dağılımını gösterir. Aralık, en büyük değer ile en küçük değer arasındaki farktır. Standart sapma, verilerin ortalamadan ne kadar uzakta olduğunu gösterir. Çeyrekler arası aralık ise, veri setinin %25'i ile %75'i arasındaki farktır.

**R’da değişkenlerin özetlerini alma:** Bu ders R’da veri analizi yapmaya odaklandığından, R’da betimsel istatistiklerin nasıl hesaplanacağı hakkında biraz zaman harcadık.

**Korelasyonlar:** İki sayısal değişken arasındaki ilişkinin yönünü ve boyutunu incelemek üzere korelasyon hesaplamaları yaptık

**Eksik veriler:** Eksik veriler, veri analizinde yaygın bir sorundur ve sonuçları etkileyebilir. Eksik verilerle başa çıkmak için çeşitli yöntemler vardır, örneğin eksik verileri silmek, ortalama veya medyan ile doldurmak veya daha gelişmiş istatistiksel yöntemler kullanmak.