BATCH : **B 150** Data Science
LESSON : **STATISTICS-2**
DATE : 16.06.2023
SUBJECT : **Scatter Plot**
Covariance
Correlation

techproeducation
techproeducation
techproeducation
techproeducation
techproedu

techproeducation.com info@techproeducation.com +1 (917) 768-7466

STATISTICS - 1

Data Science Program

Session -3

Session - 3 Content

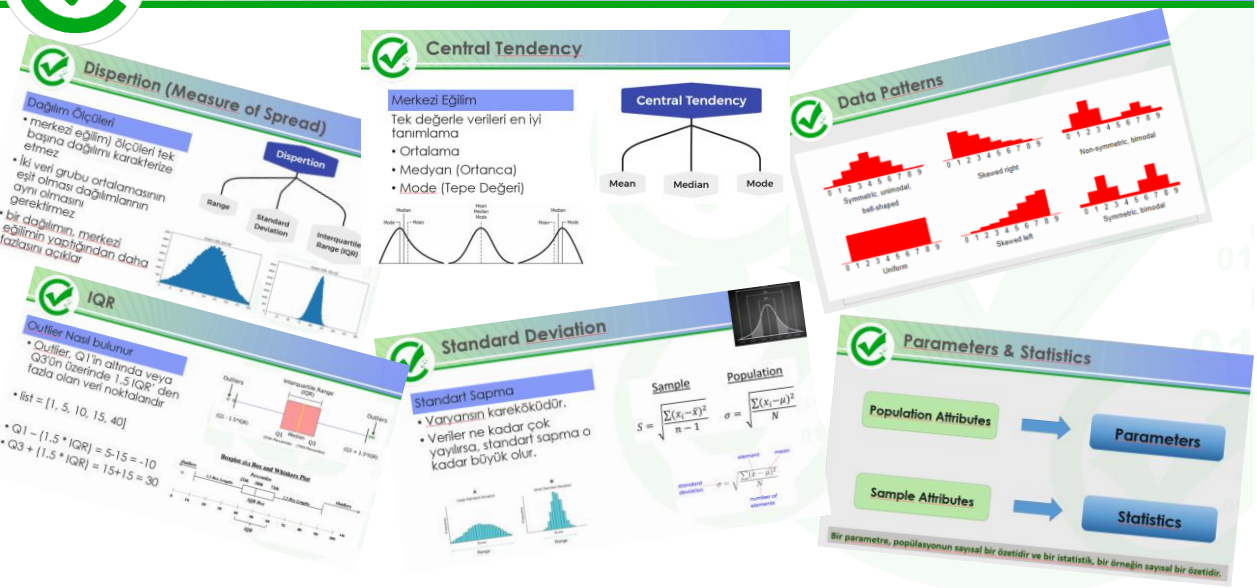
Content

- Scatter Plot
- Box Plot
- Covariance
- Correlation

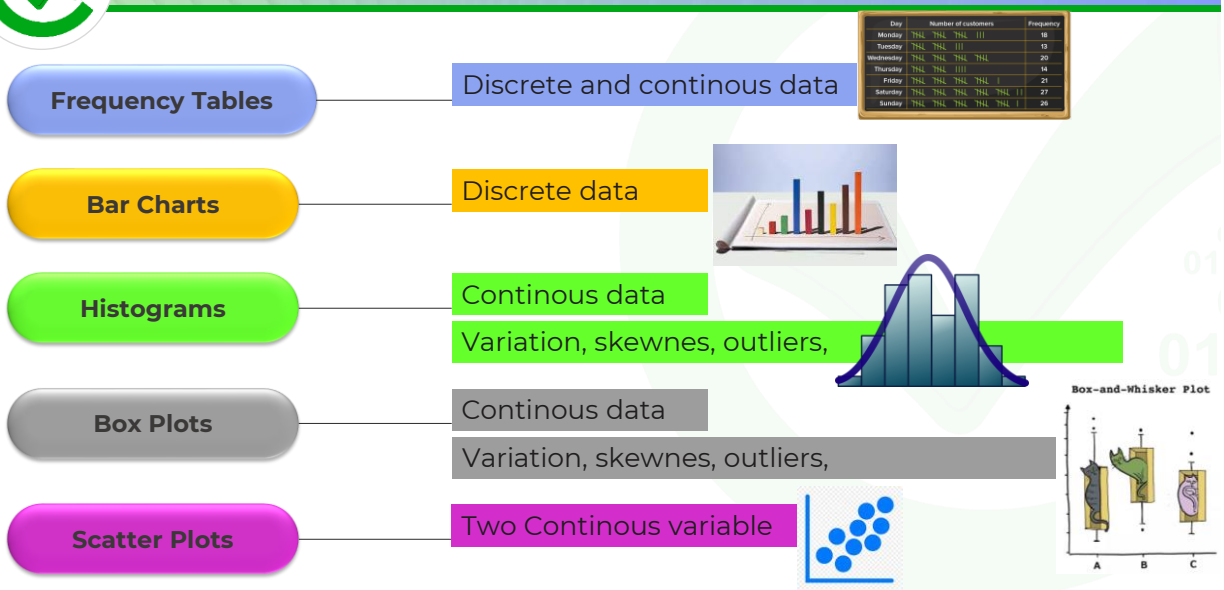
RECAP

Herkes önceki dersten hatırladığı 1 cümle yazabilir mi?

Recap – Previous Lesson



Graphical Summarization for Data

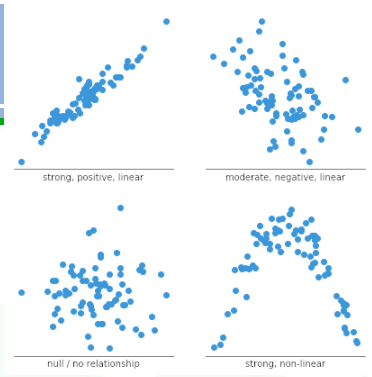


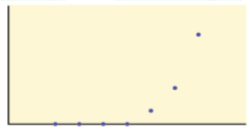
Scatter Plot

Saçılım Grafiği – Serpilme Diyagramı

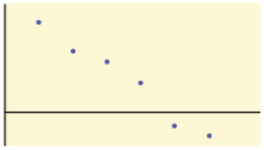
- İki değişkenli bir scatter plot, Y eksenindeki bir değişkenin değerlerini ve X eksenindeki diğer değişkenin değerlerini gösterir.

- değişkenler arasındaki ilişkinin yönünü ve büyüklüğünü gösterir




(a) Exponential growth pattern

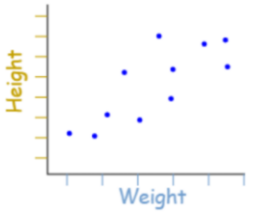

(a) Positive linear pattern (strong)


(a) Negative linear pattern (strong)

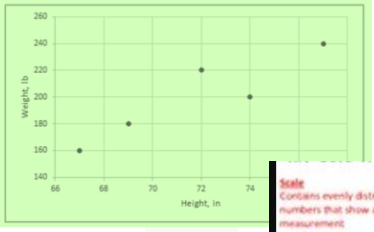

(b) Negative linear pattern (weak)

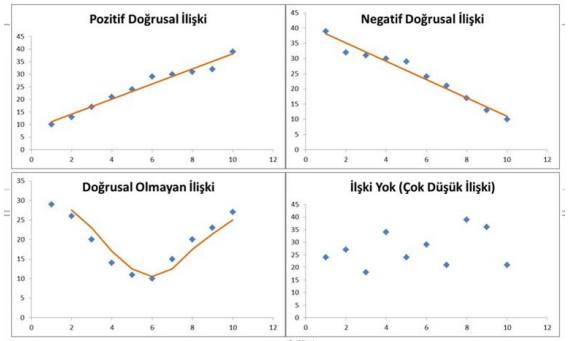

(b) No pattern

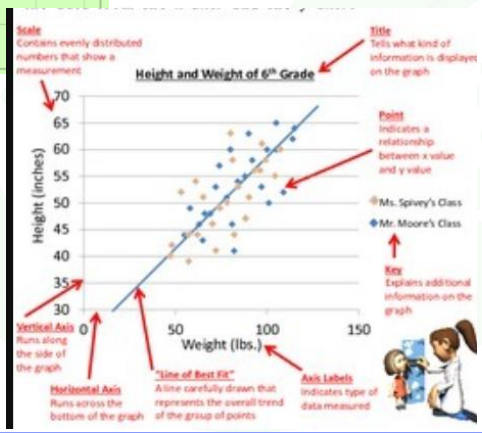
Scatter Plot



Height, in	Weight, lb
67	160
72	220
77	240
74	200
69	180



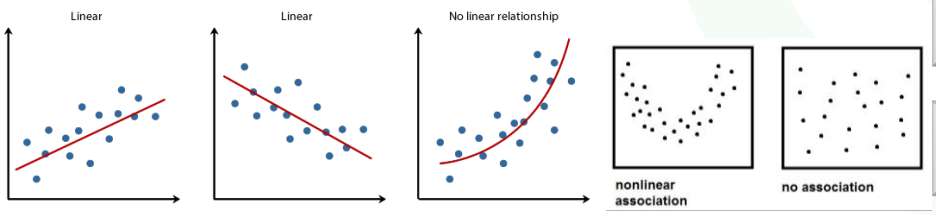




Scatter Plot Pattern

Linearity

- Lineerlik, bir veri modelinin lineer - doğrusal (düz) veya non-lineer -doğrusal olmayan (curve -eğri) olup olmadığını ifade eder.



Linearity

Slope

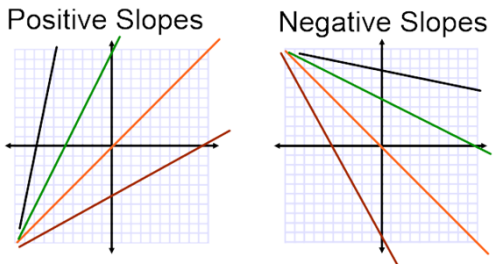
Strength

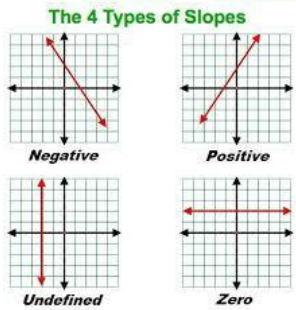
Unusual Features

Scatter Plot Pattern

Slope

- Slope yani eğim, X değişkeni büyüdüğünde Y değişkenindeki değişimin yönünü ifade eder.





Linearity

Slope

Strength

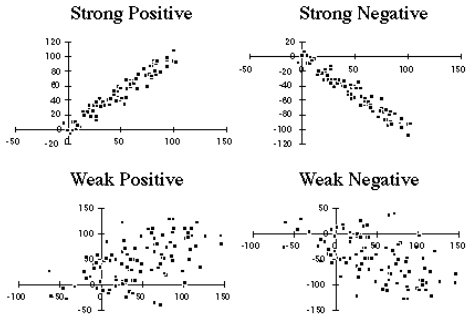
Unusual Features

Scatter Plot Pattern

Strength

- Strength, grafikteki dağılmanın veya saçılmanın derecesini - gücünü

ifade eder



Linearity

Slope

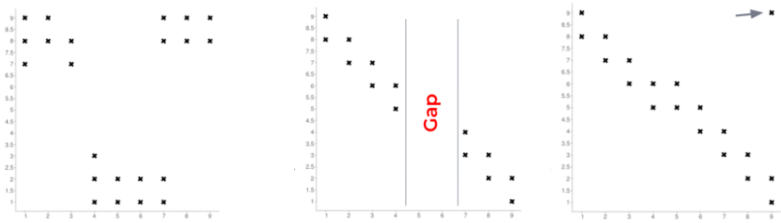
Strength

Unusual Features

Scatter Plot Pattern

Unusual Features

- Clusters (Kümelenme)
- Gaps (Boşluklar)
- Outliers (Aykırı Değerler)



Linearity

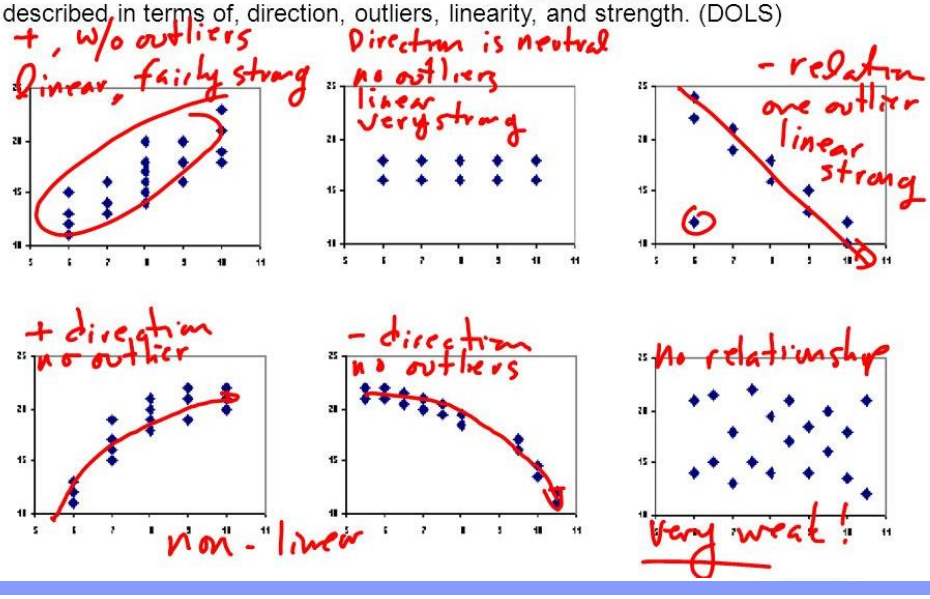
Slope

Strength

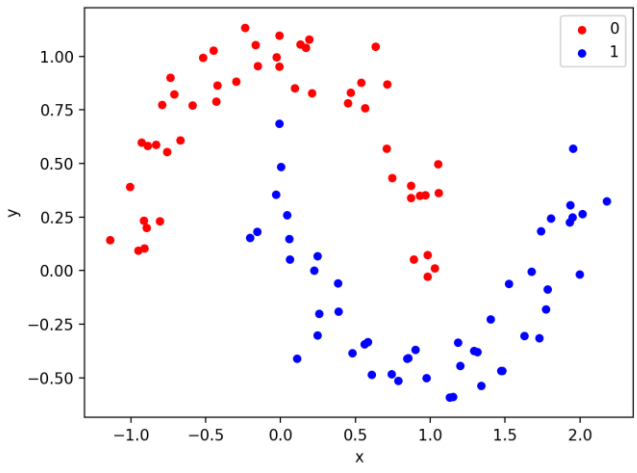
Unusual Features

Pattern of Data in Scatterplot

described in terms of, direction, outliers, linearity, and strength. (DOLS)



Peardeck Question



- ***Linearity***

  Linear Nonlinear

Slope

Positive Negative Zero

Strength

Weak Strong

Unusual Features

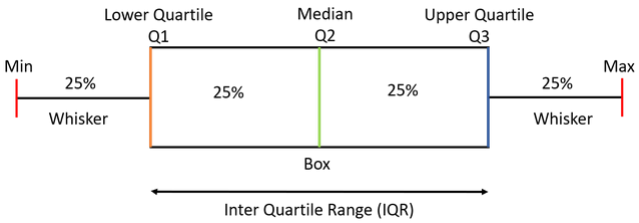
Cluster Gap Outlier None

Box Plot

Box Plot

Box Plot – Kutu Grafiği

- Bir veri kümesinin en etkili grafik özetlerinden biri olan box plot genellikle ortalama, medyan, 25. ve 75. yüzdelikler ve outlier'ları gösterir.



Quantiles same as percentiles except for scale

- Common quantiles have special names, such as **quartiles** (four groups), **deciles** (ten groups), and **percentiles** (100 groups).

Percentiles

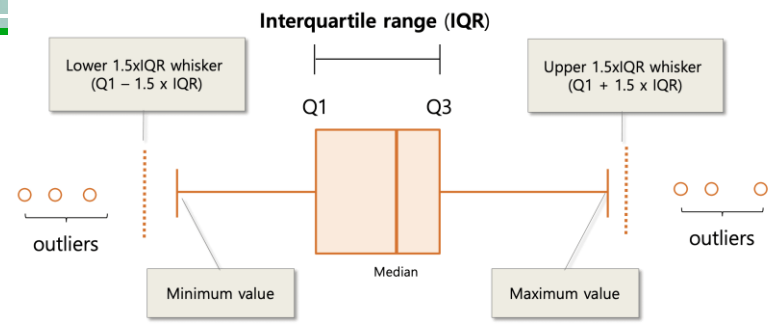
- For data, the pth percentile is the value of x such that p% of the data is less than or equal to x

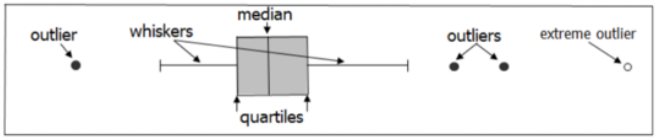
Percentiles & Quartiles & IQR

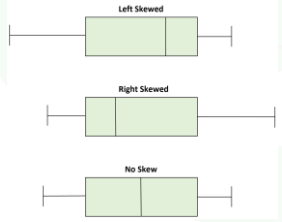
- Special percentiles:
  - Minimum: 0th percentile
  - Median: 50th percentile
  - Maximum: 100th percentile
- Quartiles: 25th and 75th percentiles
  - Sometimes called: "lower fourth" and "upper fourth"
- Interquartile Range (IQR):
  - IQR = 75th percentile - 25th percentile
  - Sometimes IQR is known as the "fourth spread"

Box Plot

- Ortadaki quartiles denen kısım benim datamın tamamının %50'si
- Q3+**1.5** IQR sonrası outlier, Q3+**3** IQR extreme outlier.

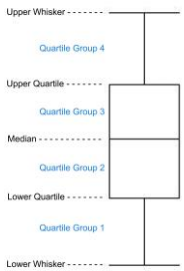






Box Plot

- Box-and-whisker plot bir veri setinin önemli özelliklerini vurgulayan bir EDA keşif veri analizi aracıdır
- Beş tane değer, grafiği çizmek için kullanılır:
  - minimum veri
  - Q1
  - Q2 (medyan)
  - Q3
  - maximum veri

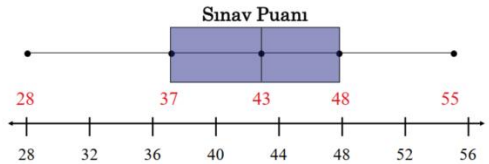


- Örnek
- Box-and-whisker plot çizmek için 15 sınav puanından verileri kullanın

28 30 33 37 37 38 42 43 43 44 45 48 48 51 55

5 tane değer

- minimum veri 28
- Q_1 37
- Q_2 (median) 43
- Q_3 48
- maximum veri 55



Box Plot – Min & Max Values

Weight, kg
38
25
37
28
35
29
35
29
34
30

Step 1: Order the data from smallest to largest.

25 28 29 29 30 34 35 35 37 38

Step 2: Find the median.

25 28 29 29 30 34 35 35 37 38

Median = 32

Step 3: Find the quartiles.

25 28 29 29 30 34 35 35 37 38

Q1 = 29 *Q3 = 35*

Step 4: Find the min and the max.

Min = 25 *Max = 38*

Min	25
Q1	29
Median	32
Q3	35
Max	38

Step 1: Scale / label an axis that fits the five-number.

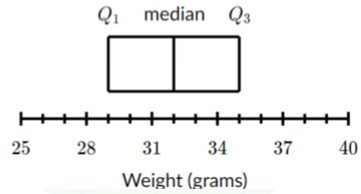
25 28 29 29 30 34 35 35 37 38

Min = 25 *Max = 38*

Min	25
Q1	29
Median	32
Q3	35
Max	38

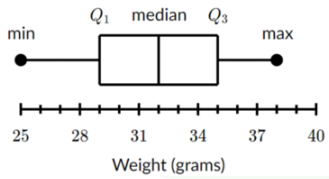
Step 2: Draw a box from *Q1* to *Q3* with a vertical line through the *median*.

25 28 29 29 30 34 35 35 37 38



Step 3: Draw a whisker from *Q1* to the *min* and from *Q3* to the *max*.

25 28 29 29 30 34 35 35 37 38

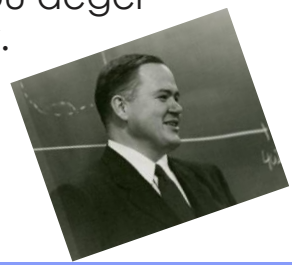


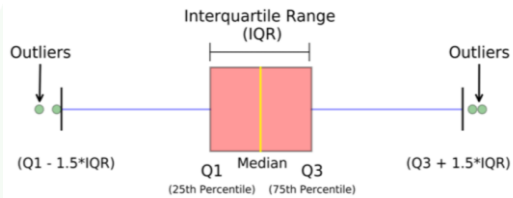
Outliers Detection

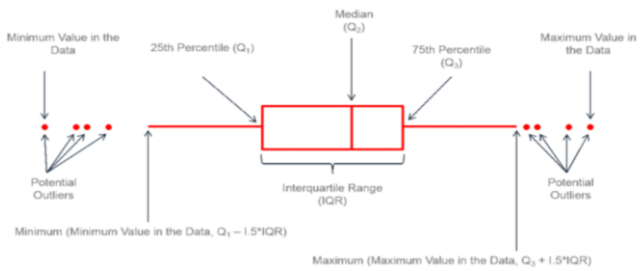
1.5*IQR Kuralı

İstatistikçi John Tukey'e göre,

- Eğer gözlem değeri Q1'in altında ve Q3'ün üzerinde 1.5*IQR dan daha fazla düşerse bu değer outlier'dır.







Box Plot – IQR

5, 7, 10, 15, 19, 21, 21, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25

Step 1: Find the median, quartiles, and interquartile range.

Median = 23
Q1 = 20
Q3 = 23.5
IQR = Q3 - Q1 = 23.5 - 20 = 3.5

```
[33] np.median(a)
     23.0

[34] np.percentile(a, 25)
     20.0

[35] np.percentile(a, 75)
     23.5
```

5, 7, 10, 15, 19, 21, 21, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25

Step 1: Find the median, quartiles, and interquartile range.

Step 2: Calculate 1.5 x IQR below the first quartile and check for low outliers.

Q1 - 1.5 x IQR = 20 - 1.5 x 3.5
= 14.75

Low Outliers: 5 7 10

5, 7, 10, 15, 19, 21, 21, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25

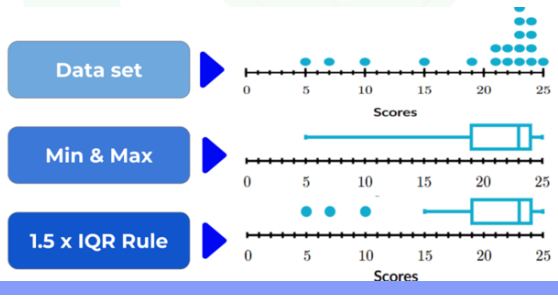
Step 1: Find the median, quartiles, and interquartile range.

Step 2: Calculate 1.5 x IQR below the first quartile and check for low outliers.

Step 3: Calculate 1.5 x IQR above the third quartile and check for high outliers.

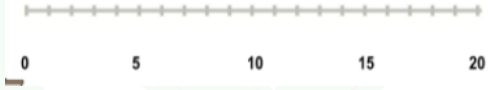
Q3 + 1.5 x IQR = 23.5 + 1.5 x 5
= 28.75

High Outliers: None



PRACTICE

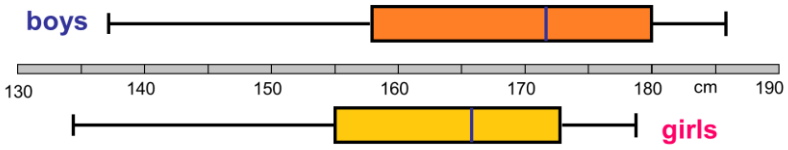
3	6	8	9	9	10	12	14	19



Box Plot'u çiziniz.

Box Plot Comments ??

boxplot of student's heights:



which are true and why?

1. the girls are taller on average
2. the boys are taller on average
3. the girls show less spread in height
4. the boys show less spread in height
5. the shortest person is a girl
6. the tallest person is a boy
7. both data sets are skewed to the left
8. half the boys are over 172 cm tall
9. half the girls are under 165cm tall

Covariance & Correlation

Covariance

Kovaryans

- İki veri arasındaki değişimin yönünü gösterir
- Değişkenlerin birlikte nasıl değiştiğini görmek önemlidir

Covariance Formula

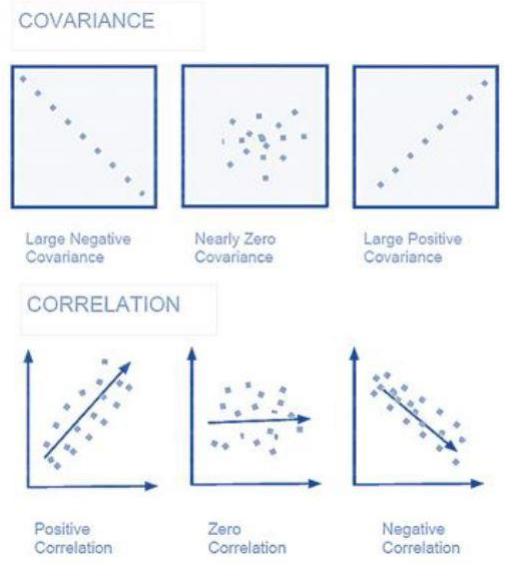
For Population

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

For Sample

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{(N-1)}$$

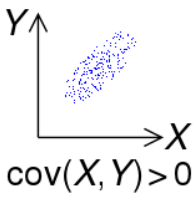
$$Cov(x,y) = \sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$



Covariance

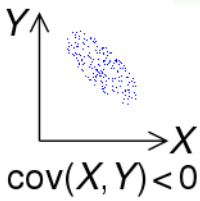
Cov (x,y) > 0

- İlişki pozitiftir.
- X artarken Y de artar



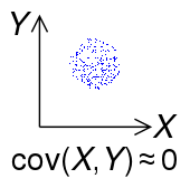
Cov (x,y) < 0

- İlişki negatiftir.
- X artarken Y azalır



Cov (x,y) = 0

- İki değişkenin arasında ilişki yoktur, birbirinden bağımısızlar.

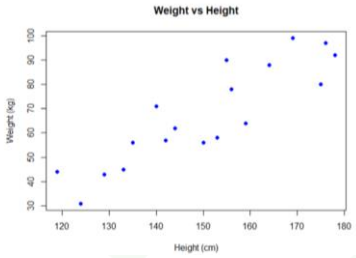


Correlation

Korelasyon

- iki değişken arasındaki ilişkinin derecesini verir.
- Bu değer -1 ile 1 arasındadır.
- -1 mutlak strong negative ilişkinin varlığını, +1 mutlak strong pozitif ilişkinin varlığını söyler

Height	Weight	Height	Weight
119	44	153	58
124	31	155	90
129	43	156	78
133	45	159	64
135	56	164	88
140	71	169	99
142	57	175	80
144	62	176	97
150	56	178	92



Correlation doesn't imply causation

Correlation

- Correlation (r): measures the direction and strength of the **linear** relationship between two quantitative variables

r = correlation

r < 0 Negative association

r > 0 Positive association

r = 0 No correlation

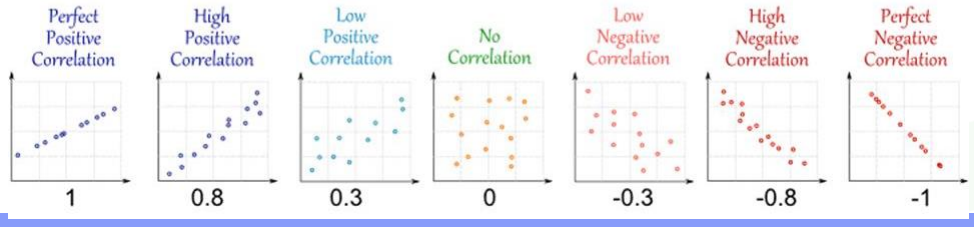
Correlation does NOT equal slope!

Sample Correlation

$$r = \frac{Cov(x,y)}{s_X s_Y}$$

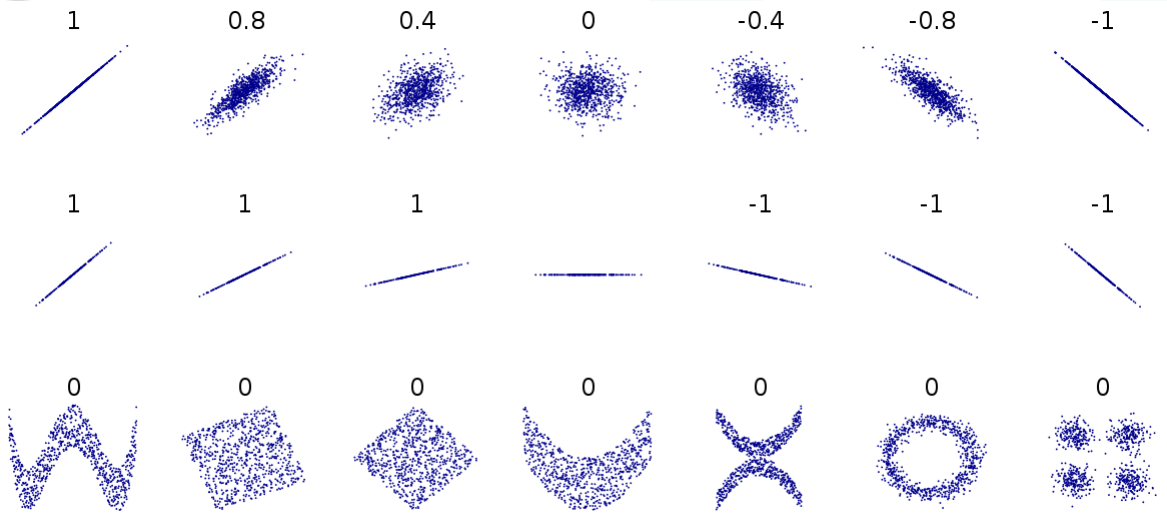
Population Correlation

$$\rho = \frac{Cov(x,y)}{\sigma_X \sigma_Y}$$



Hayattan correlation örnekleri veriniz.

Correlation

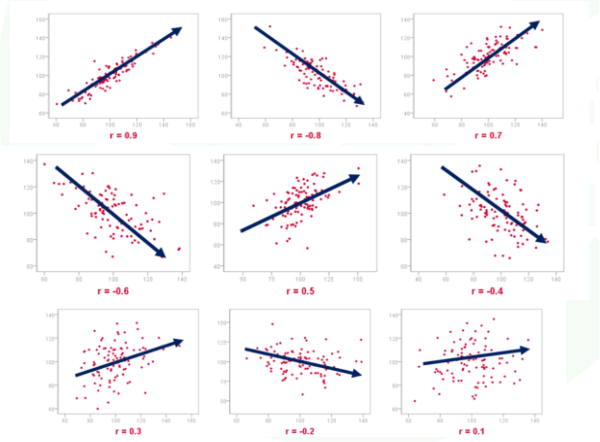


Pearson Correlation Coefficient

Pearson Katsayısı

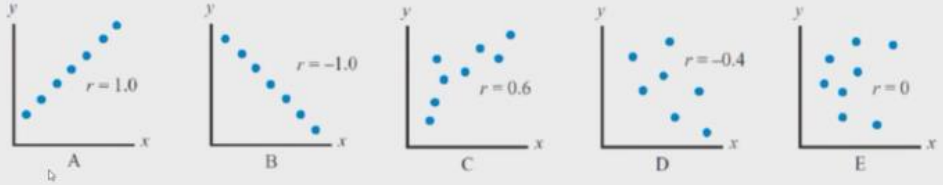
- İki değişken arasındaki korelasyon katsayısını hesaplamak için farklı yöntemler vardır. En ünlüsü Pearson Korelasyon Katsayısı. **Sample için r ile, Populasyon için R (veya ρ) ile gösterilir**
- İlişkinin gücünü gösteren -1 ile 1 arasında bir sayıdır.

$$r = \frac{\sum(x-\mu_x)(y-\mu_y)}{\sqrt{\sum(x-\mu_x)^2 \sum(y-\mu_y)^2}}$$

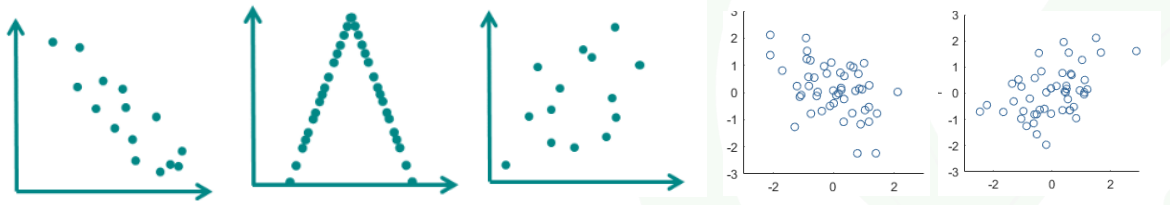


Correlation – Linear Relationship

Examples of Approximate r Value



Graph A (r = 1.0):	perfect positive correlation between x and y
Graph B (r = -1.0):	perfect negative correlation between x and y
Graph C (r = 0.6):	a moderately positive relationship: y tends to increase as x increases, but not necessarily at the steady rate we observed in Graph A
Graph D (r = -0.4):	a relatively weak negative relationship: the correlation coefficient is closer to zero, negative r value so y tends to decrease as x increases
Graph E (r = 0):	no relationship between x and y



r = - 0,5 | r = - 0,8 | r = 0 | r = +0,2 | r = +0,4

Üstteki r değerlerini alttaki plot'lar ile eşleştiriniz

Correlation - r Calculation

Cigarette (X)	Lung Capacity (Y)
0	45
5	42
10	33
15	31
20	29

$$r = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

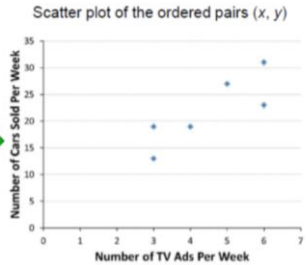
$$r_{xy} = \frac{(5)(1585) - (50)(180)}{\sqrt{\left[(5)(750) - 50^2\right]\left[(5)(6680) - 180^2\right]}}$$

$$= \frac{7925 - 9000}{\sqrt{(3750 - 2500)(33400 - 32400)}}$$

$$= \frac{-1075}{\sqrt{(1250)(1000)}} = -.9615$$

Week	Number of TV Ads x	Number of Cars Sold y	xy	x^2	y^2
1	3	13	39	9	169
2	6	31	186	36	961
3	4	19	76	16	361
4	5	27	135	25	729
5	6	23	138	36	529
6	3	19	57	9	361
	$\Sigma x = 27$	$\Sigma y = 132$	$\Sigma xy = 631$	$\Sigma x^2 = 131$	$\Sigma y^2 = 3110$

Week	Number of TV Ads x	Number of Cars Sold y
1	3	13
2	6	31
3	4	19
4	5	27
5	6	23
6	3	19



$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}}$$

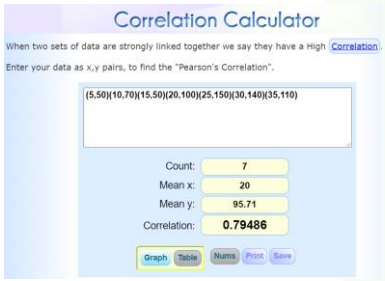
$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} = \frac{(6)(631) - (27)(132)}{\sqrt{\left[(6)(131) - (27)^2\right]\left[(6)(3110) - (132)^2\right]}}$$

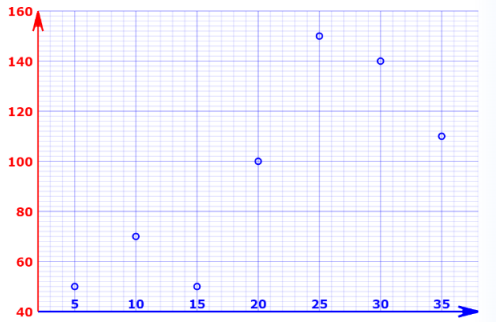
$$= \frac{222}{\sqrt{[57]\,[1236]}} = \frac{222}{265.43} = 0.836$$

Online Calculator

Online Calculator link

Variable 1	Variable 2
5	50
10	70
15	100
20	100
25	150
30	140
35	110

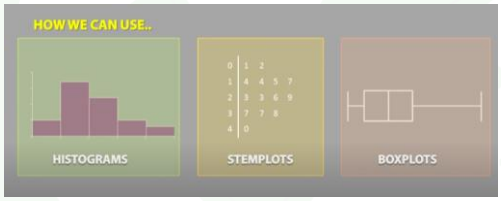




YOUTUBE VIDEO ONERI

https://www.youtube.com/watch?v=DAH8DyLXdjM

- Explanatory and Response Variables, Correlation (2.1)



Python Calculation

input :

```
import numpy as np

temp=[93,84,82,78,98,70]

number_of_people=[13,10, 11, 8, 15, 9]

print("covariance: ", np.cov(temp, number_of_people))

print("correlation: ", np.corrcoef(temp, number_of_people))
```

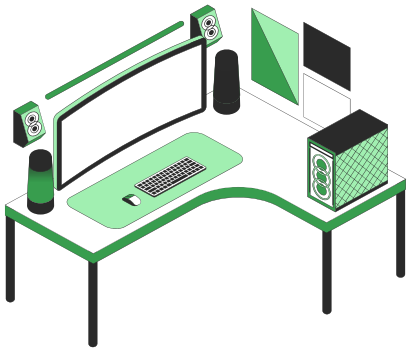
output :

```
covariance:  [[102.56666667  24.        ]
              [ 24.           6.8       ]]

correlation:  [[1.         0.90876934]
               [0.90876934 1.        ]]
```

Peardeck Interaction

Bugünkü ders verimli geçti



Do you have any questions?

Send it to us! We hope you learned something new.