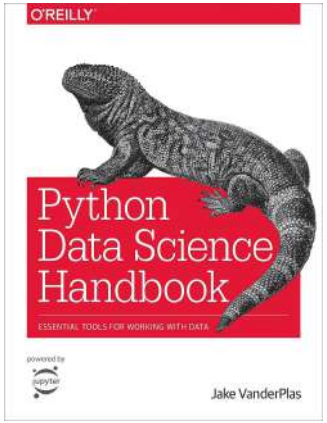
# Veri Analizi

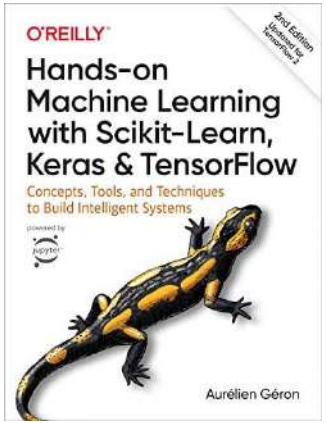
# Ömer Cengiz

**Github:** omercengiz
**Linkedin:** omercengiz96
**Twitter:** omerr.cengizz

# Kaynaklar




**Python Data Science Handbook**
Jake VanderPlas




**Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow**
Aurelien Geron

Stanford
University

**Stanford University stanford.edu/~shervine/**
Shervine Amidi

# Bugün Ne Konuşacağız?

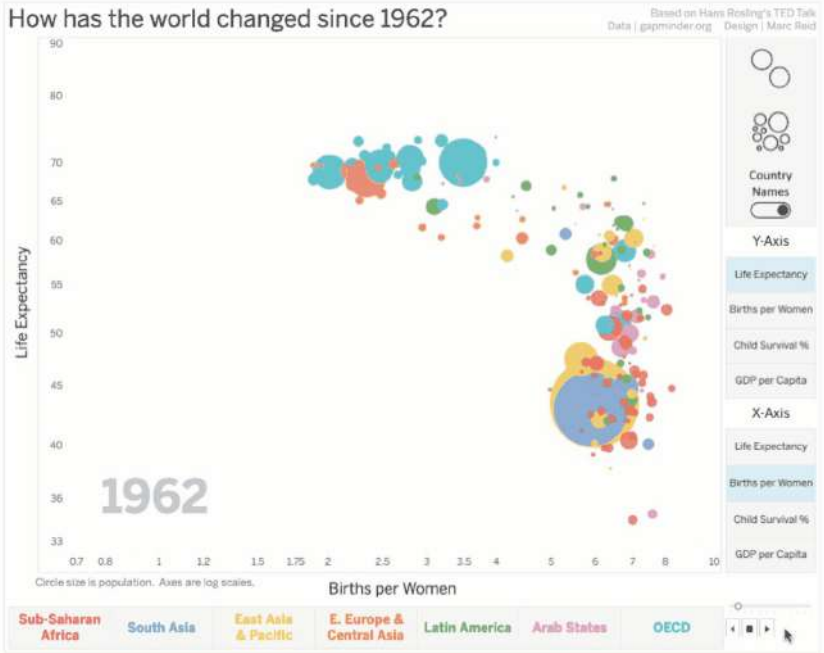
1. Günümüzün en önemli ve popüler kaynağı veri ile tanışacak, veri tiplerini görerek verileri anlamaya derin bir giriş yapacağız.
2. Veriyi ifade etmenin en iyi yöntemi olan grafik türlerini görerek veriyi grafiklerle nasıl ifade edeceğimizi öğreneceğiz.
3. Verileri anlamak ve kullanmak için önemli konulardan biri olan temel istatistik konusuna girecek ve verileri nasıl yorumlayabileceğimizi göreceğiz.
4. Veri Dağılımını anlamlandırabilmek ve anlayabilmek için hipotez testlerini öğreneceğiz.
5. Verilerde aykırı ve eksik değerleri tespit etmeyi öğrenecek ve bir ön işleme yöntemi olan veri temizleme operasyonunu göreceğiz.
6. Veri ön işleme ve analizi için Pandas kütüphanesinin sunmuş olduğu teknikleri ve fonksiyonları görerek, verimizi hem analize hem de makine öğrenmesi modelimize hazır hale getirmeyi öğreneceğiz.

# Veri Nedir?

Veriler, gözlem yoluyla toplanan, genellikle sayısal olan bilgi birimleridir.

- Bir veya birden fazla bilgiden oluşan kümedir.
- Genellikle araştırma, gözlem, deney, sayım, ölçüm yoluyla elde edilir

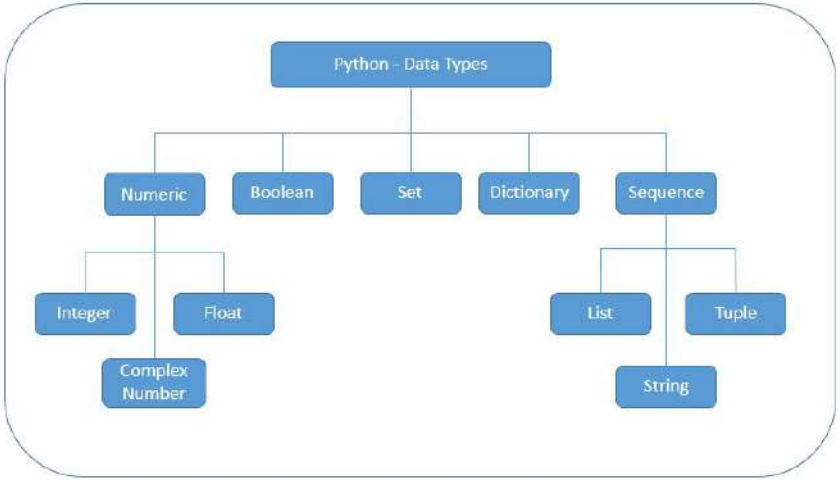
*Örneğin; bizi çevreleyen her şey bir tür veridir. Ev fiyatlarından araba kilometresine, hava tahmininden yürünen mesafeye kadar hayatımızdaki her şey devasa bir veri kümesidir.*



# Standart ve Özel Veri Tipleri

**Standart Veri Tipleri:**

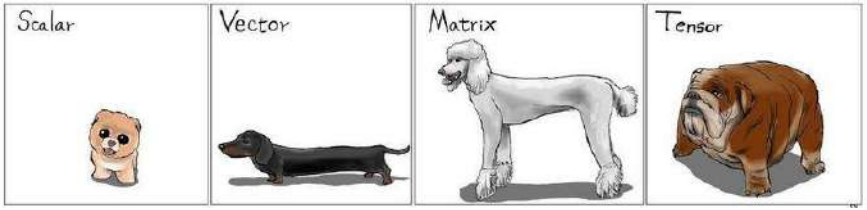
- **String**: Unicode text verilerini temsil eden veri türüdür
- **Integer**: Tam sayıları temsil eden veri türüdür
- **Float**: Ondalıklı sayıları temsil eden veri türüdür
- **Boolean**: Doğru/Yanlış (1/0) ikilisini temsil eden veri türüdür
- **List**: Birden çok veriyi saklamak için kullanılan veri türüdür
- **Dict**: Birden fazla veriyi key:value çiftleri halinde saklamak için kullanılan veri türüdür

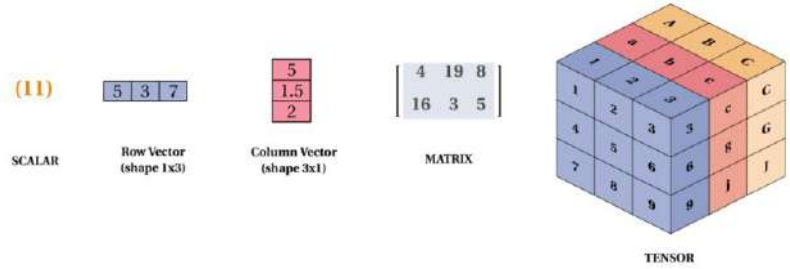


# Standart ve Özel Veri Tipleri

**Özel Veri Tipleri:**

- **NumPy Array (Dizi)**: Python listeleri gibi tek boyutlu koleksiyon objeleridir. Hem tek tipte veri barındırmaları hem de NumPy'ın lineer cebir için olan özel fonksiyonları sayesinde Python listelerine göre çok daha efektif çalışırlar
- **Pandas Series**: İndekslerin istenen şekilde tanımlanabildiği ve değer olarak NumPy array tutan tek boyutlu koleksiyon objeleridir
- **Pandas DataFrame**: İndekslerin ve satır isimlerinin istenen şekilde tanımlanabildiği 2 boyutlu (satır ve sütun) bir koleksiyon objeleridir



# Veri Türleri

**Categorical (Kategorik) Veri:** Kategorik veriler, bir kişinin cinsiyeti, medeni durumu, memleketi veya beğendiği film türleri gibi özellikleri temsil eder. Kategorik veriler sayısal değerler alabilir ("1" erkek ve "2" kadın anlamına gelebilir), ancak bu sayıların matematiksel anlamı yoktur.

**Ordinal (Sıralı) Veri:** Sayısal ve kategorik verilerin bir karışımıdır. Veriler kategorilere ayrılır, ancak kategorilere yerleştirilen sayıların bir anlamı vardır. Örneğin, bir restoranı 0 (en düşük) ila 4 (en yüksek) yıldız arasında derecelendirmek, sıralı veriler verir.

# Veri Türleri

**Interval (Aralıklı) Veri:** Her noktanın birbirinden eşit uzaklıkta (aralık) yerleştirildiği, bir ölçek boyunca ölçülebilen bir veri türüdür. Bu verilerde mutlak sıfır yoktur.

*Örnek: Celcius derece değerleri*

**Ratio (Oransal) Veri:** Bu tür veriler ölçülebilir ve sıralanabilir. Aralıklı veriden farkı, mutlak sıfırın olmasıdır. Mutlak sıfırın varlığı, oran verilerinde negatif değer alamayacağı anlamına gelir.

*Örnek: Kelvin cinsinden sıcaklık (0K ısı yok demek değildir), boy, hız, vb.*

# Yapılandırılmış vs. Yapılandırılmamış Veri

**Yapılandırılmış Veri:**

Açıkça tanımlanmış bir şekilde excel tablolarında, MySQL ve MSSQL gibi ilişkisel veri tabanlarında sabit satırlara ve sütunlara düzgün bir şekilde uyan ve tutulabilen verilerdir.

Daha az depolama alanı gerektirir ve yüksek düzeyde ölçeklenebilir.

*Tarihler, telefon numaraları, adresler (string, number ve date tipindeki veriler) yapılandırılmış veriye örnektir.*

**Yapılandırılmamış Veri:**

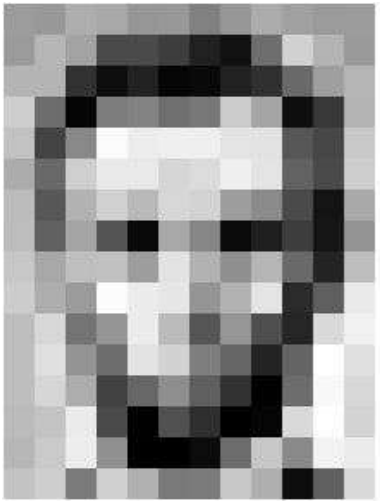
Ses, görüntü, metin gibi bir çok farklı formda bulunan ve excel tablolarında, ilişkisel veri tabanlarında tutulamayan verilerdir. NoSQL gibi ilişkisel olmayan veri tabanlarında tutulabilirler.

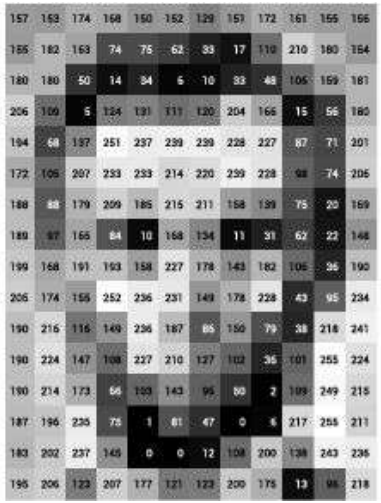
Daha fazla depolama alanı gerektirir ve ölçeklenmesi zordur.

*Metinler, mobil aktivite, sosyal medya gönderileri, sensör verileri yapılandırılmamış veriye örnektir.*

# Çağımızın Veri Tipi: Multimedya

Sayı, metin, resim, ses ve video ile ilgili herhangi bir şey içeren her türlü bilgi multimedya olarak tanımlanır. Türü ne olursa olsun, sonunda her zaman sayı cinsine dönüştürülerek saklanırlar.





| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 157 | 153 | 174 | 168 | 150 | 152 | 129 | 151 | 172 | 161 | 155 | 156 |
| 155 | 182 | 163 | 74 | 75 | 62 | 33 | 17 | 110 | 210 | 180 | 154 |
| 180 | 180 | 50 | 14 | 34 | 6 | 10 | 33 | 48 | 106 | 159 | 181 |
| 206 | 109 | 5 | 124 | 131 | 111 | 120 | 204 | 166 | 15 | 56 | 180 |
| 194 | 68 | 137 | 251 | 237 | 239 | 239 | 228 | 227 | 87 | 71 | 201 |
| 172 | 105 | 207 | 233 | 233 | 214 | 220 | 239 | 228 | 98 | 74 | 206 |
| 188 | 88 | 179 | 209 | 185 | 215 | 211 | 158 | 139 | 75 | 20 | 169 |
| 189 | 97 | 165 | 84 | 10 | 168 | 134 | 11 | 31 | 62 | 22 | 148 |
| 199 | 168 | 191 | 193 | 158 | 227 | 178 | 143 | 182 | 106 | 36 | 190 |
| 205 | 174 | 155 | 252 | 236 | 231 | 149 | 178 | 228 | 43 | 95 | 234 |
| 190 | 216 | 116 | 149 | 236 | 187 | 86 | 150 | 79 | 38 | 218 | 241 |
| 190 | 224 | 147 | 108 | 227 | 210 | 127 | 102 | 36 | 101 | 255 | 224 |
| 190 | 214 | 173 | 66 | 103 | 143 | 96 | 50 | 2 | 109 | 249 | 215 |
| 187 | 196 | 235 | 75 | 1 | 81 | 47 | 0 | 6 | 217 | 255 | 211 |
| 183 | 202 | 237 | 145 | 0 | 0 | 12 | 108 | 200 | 138 | 243 | 236 |
| 195 | 206 | 123 | 207 | 177 | 121 | 123 | 200 | 175 | 13 | 96 | 218 |

# Veri Görselleştirme

Görselleştirme Nedir?

Tufte'nin 6 Prensibi

Grafik Türleri

# Veri Görselleştirme Nedir?

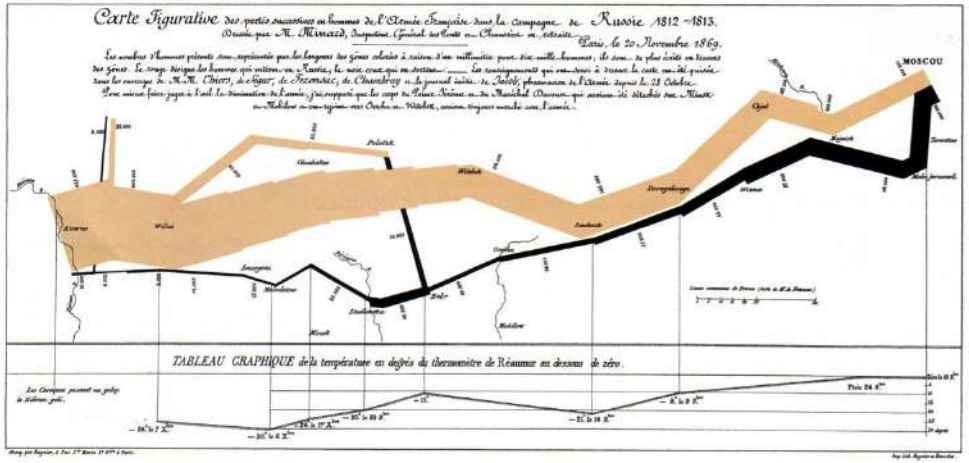
**Veri Görselleştirme,** her türlü boyuttaki verileri anlamamıza ve yorumlamamıza yardımcı olduğu için EDA (Keşifsel Veri Analizi)'nın en önemli ve temel aşamalarından biridir.

Veri Görselleştirme tekniklerini kullanarak Big Data adı verilen büyük boyutlu verilerden bile hem istatistiksel hem de betimsel yorumlamalar ve çıkarımlar gerçekleştirilebilir.

# Tufte'nin 6 Prensibi

Tufte'nin 6 prensibi, görsel öğeleri kullanarak verilerin nasıl doğru bir şekilde tasvir edileceği konusunda bir kılavuz görevi görür.

1. Karşılaştırma
2. Nedensellik
3. Çok Değişkenlilik
4. Bütünleştirme
5. Dokümantasyon
6. İçerik



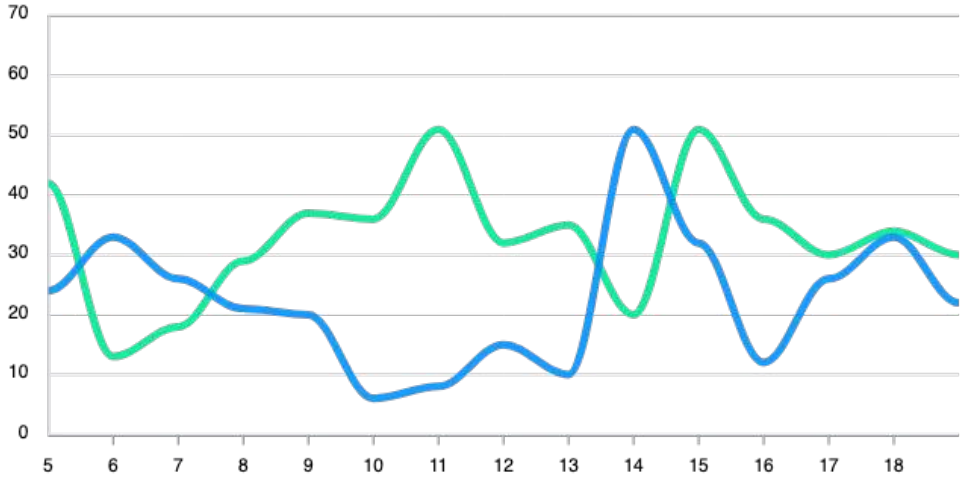
# Grafik Türleri

# Line Plot

Çizgi grafiği, temel olarak iki sayısal değer kümesi arasındaki ilişkiyi göstermek için kullanılır. Genellikle, iki bağımlı değişken arasında artan veya azalan bir eğilim göstermek için uygundur.

*Örneğin, 24 saatlik bir süre içinde havanın nasıl değiştiğini görmek istiyorsanız, x ekseninin saatlik bilgileri ve y ekseninin derece cinsinden hava durumunu içeren bir çizgi grafiği kullanabilirsiniz.*

Verideki trendleri anlamak için bilgileri hızla taramayı kolaylaştırır



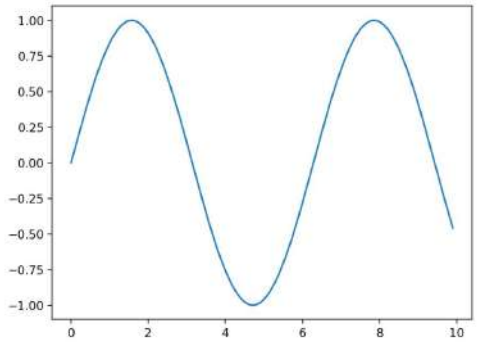
# Line Plot Uygulama

Çizgi grafikleri, gözlemler arasında bir sıralamanın olduğu herhangi bir dizi verisinin yanı sıra zaman serisi verilerinin sunulması için kullanışlıdır.

Aşağıdaki örnek, y eksenindeki gözlemler olarak x ekseni ve x ekseninin bir fonksiyonu olarak bir sinüs dalgası olarak 100 kayan nokta değerinden oluşan bir dizi oluşturur. Sonuçlar bir çizgi grafiği olarak çizilir.

Line Plot

```
1 from numpy import sin
2 from matplotlib import pyplot
3
4 x = [x*0.1 for x in range(100)]
5 y = sin(x)
6
7 pyplot.plot(x, y)
8 pyplot.show()
```



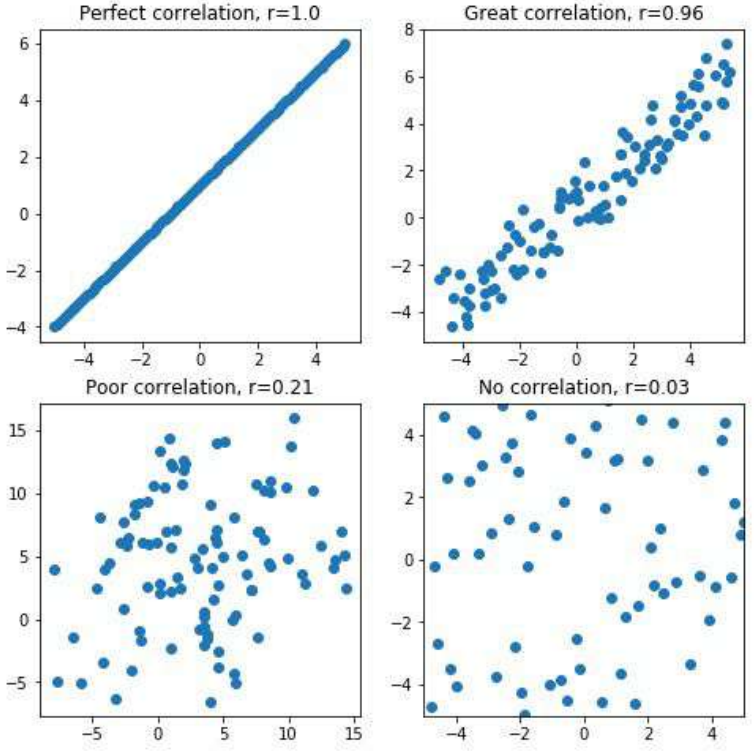
# Scatter Plot

Bir Dağılım grafiği, esas olarak iki sayısal grup arasındaki ilişkiyi dağınık noktalar şeklinde çizmek için kullanılır.

Grafikteki her nokta tek bir gözlemi temsil edecek şekilde gösterilir:

- X ekseni, örneğin bir özelliğini temsil eder
- Y ekseni, aynı örneğin farklı bir özelliğini temsil eder

*Dağılım grafikleri, iki değişken arasındaki ilişkiyi veya korelasyonu göstermek için kullanışlıdır*

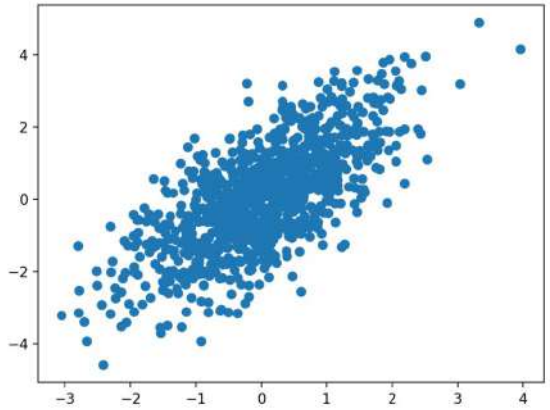


# Scatter Plot Uygulama

Aşağıdaki örnek, birbiriyle ilişkili iki veri örneği oluşturur. Birincisi, standart bir dağılımdan alınan rastgele sayıların bir örneğidir. İkincisi, birinci ölçünün değerine ikinci bir rastgele değer ekleyerek birinciye bağlıdır. Sol taraftaki görselde ise bu iki değişkenin korelasyonu ve dağılım grafiği gözükmektedir.

Scatter Plot

```
1 from numpy.random import seed
2 from numpy.random import randn
3 from matplotlib import pyplot
4
5 x = 20 * randn(1000) + 100
6 y = x + (10 * randn(1000) + 50)
7
8 pyplot.scatter(x, y)
9 pyplot.show()
```



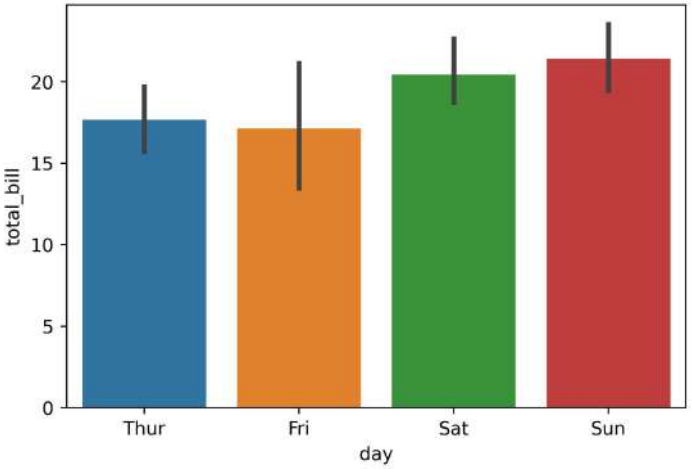
# Bar Plot

Çubuk grafiği; toplam, ortalama, medyan vb. gibi bir toplama işleviyle gruplanmış kategorik bir sütundaki benzersiz değerler arasındaki ilişkiyi çizmek için kullanılır.

Kategorik değerler x ekseni olarak iletilir ve karşılık gelen toplu sayısal değerler y ekseninde iletilir.

- Farklı kategorideki verileri karşılaştırmak için kullanılır
- Grafiğin bir ekseni karşılaştırılmakta olan belirli kategorileri gösterir ve diğer eksen ölçülen bir değeri temsil eder

Çubuk grafiğinin histogram ile karıştırılmaması gerekir! Boxplotlar kategorik, histogramlar ise sürekli değerler için kullanılır.

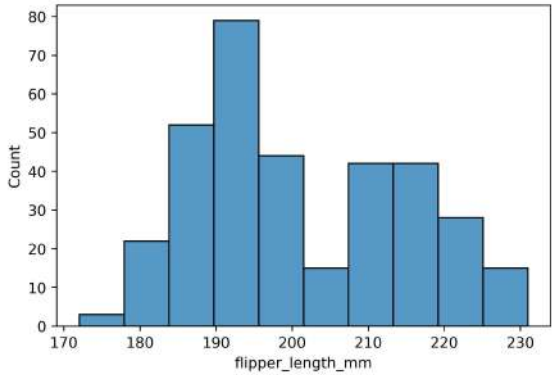


# Histogram

Histogramlar temel olarak sayısal bir öğe listesi için veri dağılımını görüntülemek için kullanılır.

Verilerin sürekli bir aralık veya belirli bir süre boyunca dağılımını gösteren bir veri görselleştirme grafiğidir.

- X ekseninde gösterilen sürekli değişken, ayrık aralıklara bölünür ve o ayrık aralıkta sahip olduğunuz veri sayısı, çubuğun yüksekliğini belirler
- Histogramlar, değerlerin nerede yoğunlaştığını, uç noktaların neler olduğunu ve veri kümesinde herhangi bir boşluk veya olağandışı değerler olup olmadığı konusunda bir tahmin verir




$$Bin\ width\ (h) = \frac{3.5 \times \sigma}{\sqrt[3]{n}}$$

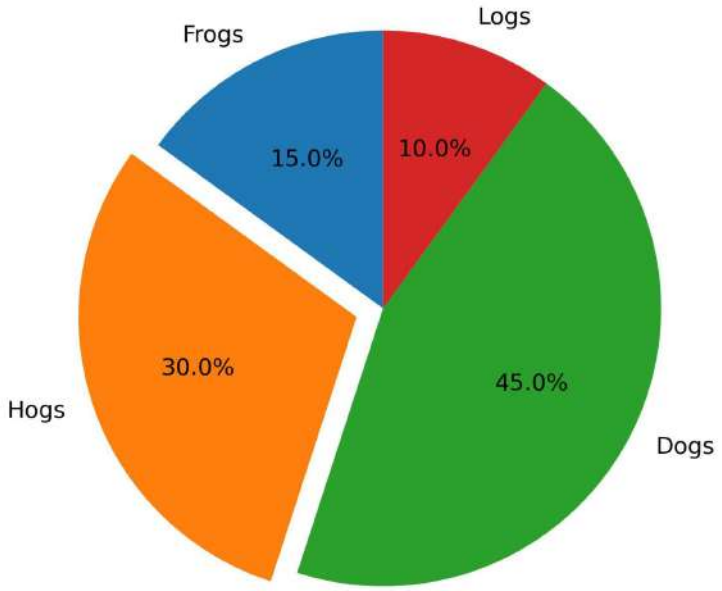
σ = Standard deviation of the data source

n = number of values in the data source

# Pie Charts

Pasta grafikler, adından da anlaşılacağı gibi, kategorik bir sütundaki değerlerin yüzde dağılımını gösterir.

- Bir bütünün parçaları arasındaki ilişki açık halde görülebilir
- Grafiğin parçaları, her kategorideki bütünün kesri ile orantılıdır

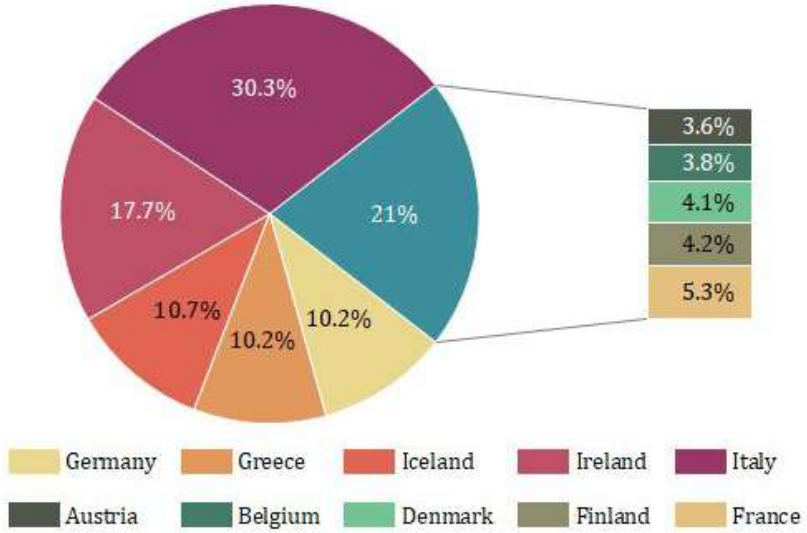


# Pie Charts

**Sık yapılan hatalar**

- Birkaç pasta grafiği karşılaştırılmamalı
- Grafik üzerinde oran gösteriliyorsa, oranlar toplamı 100 olmalıdır

**Her dilime açıklama yazılması:** Açıklamalar kısa, öz, bir veya birkaç kelime ile açıklanmalıdır

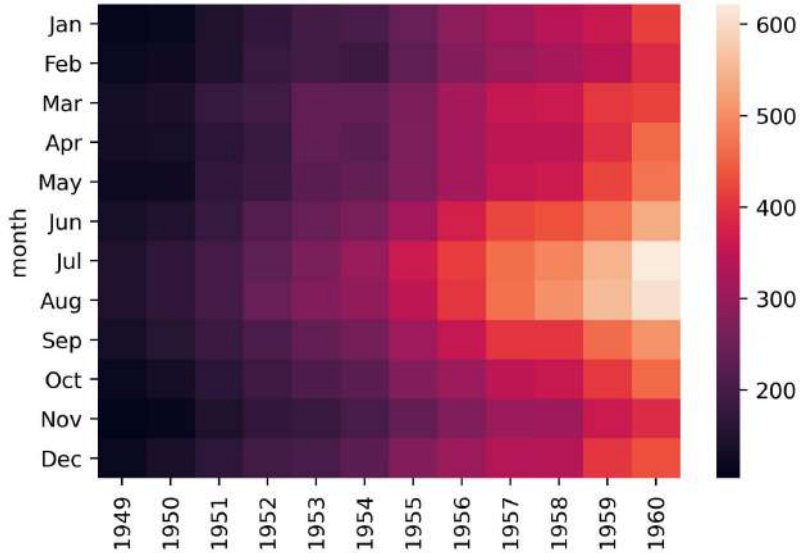


# Heatmap

Matrisin değerini görselleştirmek için renkleri kullanarak verilerin grafiksel bir temsilidir

Daha yaygın değerleri veya daha yüksek etkinlikleri temsil etmek için daha parlak renkler kullanılır daha az yaygın veya etkinlik değerlerini temsil etmek için daha koyu renkler tercih edilir.

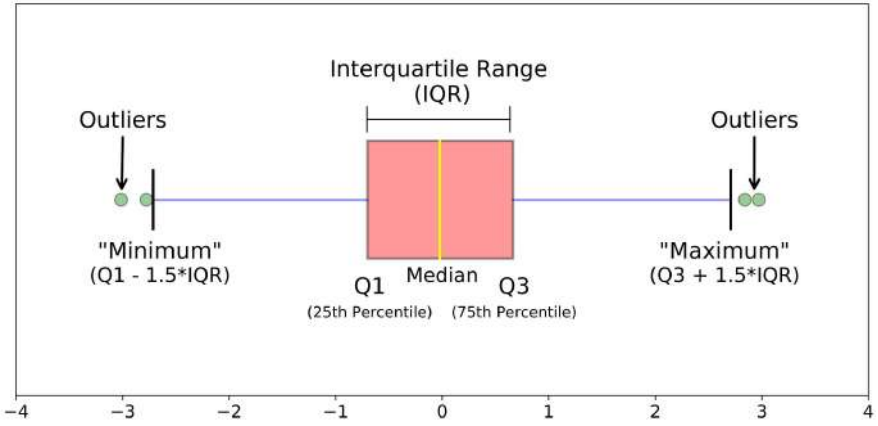
Korelasyon haritalarında veya karmaşıklık matrislerinde sıkça kullanılır.



# Box Plot

Genellikle gruplar arasında bir veri dağılımını göstermenin görsel bir temsilidir.

- En basit Box plot çizimleri, minimumdan maksimuma tüm varyasyon aralığını, olası varyasyon aralığını ve aykırı değerleri gösterir
- Box plot beş parçadan oluşur
  - Minimum
  - İlk çeyrek
  - Medyan (ikinci çeyrek)
  - Üçüncü çeyrek
  - Maksimum
  - Kartiller




$$\mathrm{IQR} = Q_3 - Q_1$$

# Kartil Nedir ve Nasıl Hesaplanır?

Kartiller, verileri çeyreklere bölen değerlerdir.
Veriler, sayı doğrusunda denk geldikleri yerlere göre dört parçaya ayrılır.

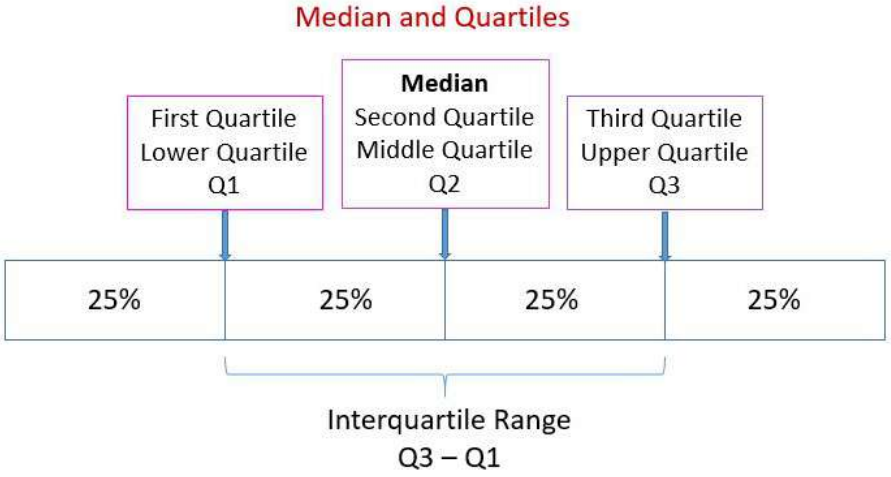
N: Toplam veri noktası sayısı

**Lower Quartile (Q1)** = (N+1) * 1 / 4

**Middle Quartile (Q2)** = (N+1) * 2 / 4

**Upper Quartile (Q3)** = (N+1) * 3 / 4

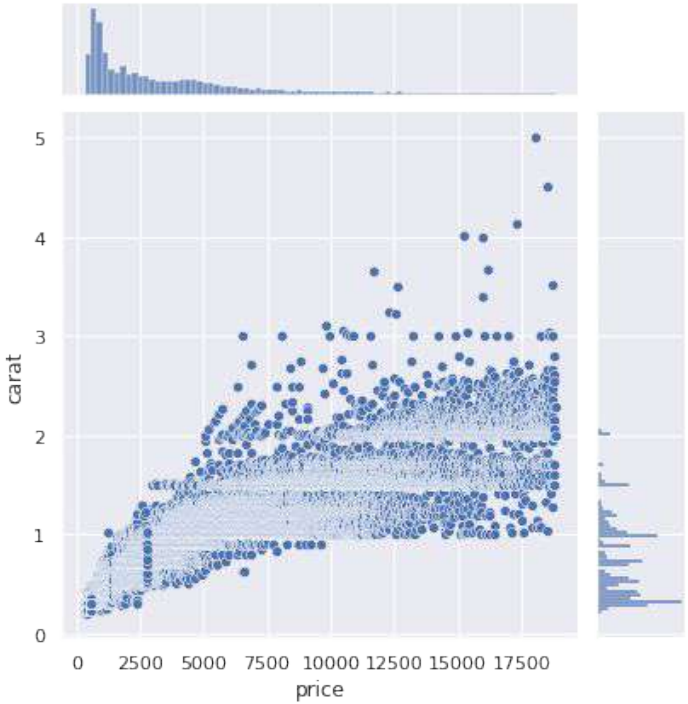
**Interquartile Range** = Q3 – Q1



# Ortak (Joint) Grafik

Ortak Grafikler, iki değişkenli veriler arasındaki ilişkileri ve bu verilerin bireysel dağılımlarını aynı anda araştırmak için kullanılır.

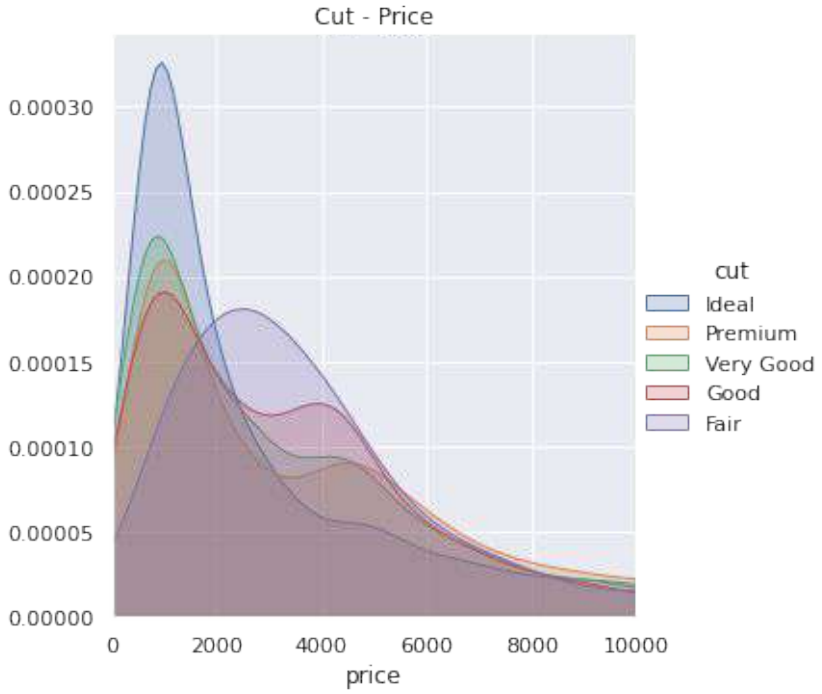
*Örneğin; yandaki örnekte **diamonds** veri setindeki **carat** ve **price** sürekli değişkenleri arasındaki ilişki ve her birinin ayrı ayrı yoğunluk grafiği verilmiştir.*



# KDE Plot (Kernel Density Estimate)

KDE, sürekli bir değişkenin farklı değerlerdeki olasılık yoğunluğunu gösterir. Başka bir deyişle, sayısal bir değerin dağılımını ifade etmek için kullanılır.

Verilerin dağılımının, ayrık histogram yerine sürekli (continuous) olarak görselleştirmesi istendiği durumlarda yararlı olabilir.



# Temel İstatistik

Betimsel İstatistik Kavramları

Çarpıklık Kavramı

Korelasyon ve Korelasyon Matrisi

Simpsons Paradox

Anscombe Quartet

Veri Dağılımı ve Hipotez Testleri

# Betimsel İstatistik Kavramları

- **Ortalama (Mean):** Bir veri setindeki tüm gözlemlerin toplamının toplam gözlem sayısına oranıdır
- **Medyan (Median):** Sıralanmış, artan veya azalan bir sayı listesinde ortadaki sayıdır
- **Varyans (Variance):** Varyans, veri noktalarının ortalama değerden ne kadar farklı olduğunun bir ölçüsüdür. Veri noktalarının ortalamaya olan uzaklıklarının karelerinin ortalamasıdır
- **Standart Sapma (Standard Deviation):** Standart sapma, bir veri kümesinin ortalamasına göre dağılımını ölçen bir istatistiktir. Varyansın karekökü alınarak bulunur ve orjinal verilerle aynı birimde ifade edilir
- **Mod (Mode):** Veri noktaları içerisinde en çok tekrar eden değerdir

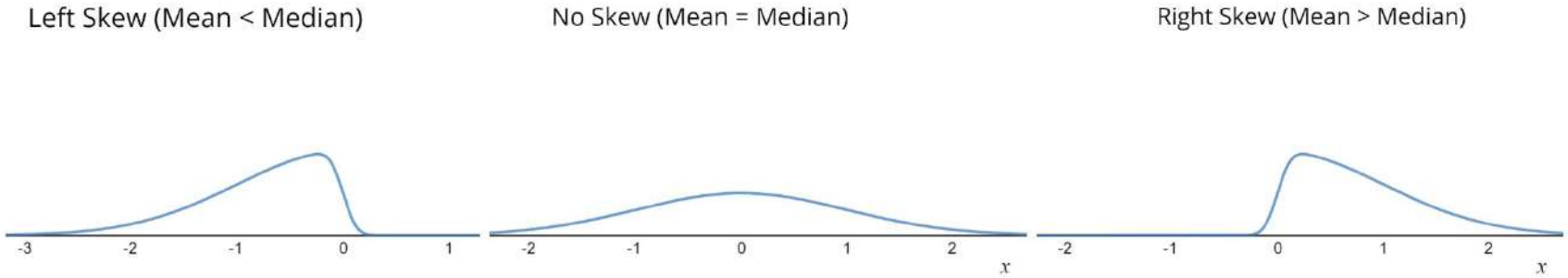
$$m = \frac{\text{sum of the terms}}{\text{number of terms}}$$

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

# Çarpıklık (Skewness)

Çarpıklık, bir veri setinde normal dağılımdan sapan bir bozulma veya asimetriyi ifade eder. Dağılım eğrisi sola veya sağa kaydırılırsa çarpıklıktan söz edilebilir. Çarpıklık, belirli bir dağılımın normal bir dağılımdan ne ölçüde farklılık gösterdiğinin bir temsili olarak ölçülebilir.

Asimetrik bir dağılımda; negatif bir çarpıklık sol taraftaki kuyruğun sağdakinden daha uzun olduğunu (left-skewness) gösterirken, pozitif bir çarpıklık sağ taraftaki kuyruğun soldan daha uzun olduğunu (right-skewness) gösterir.



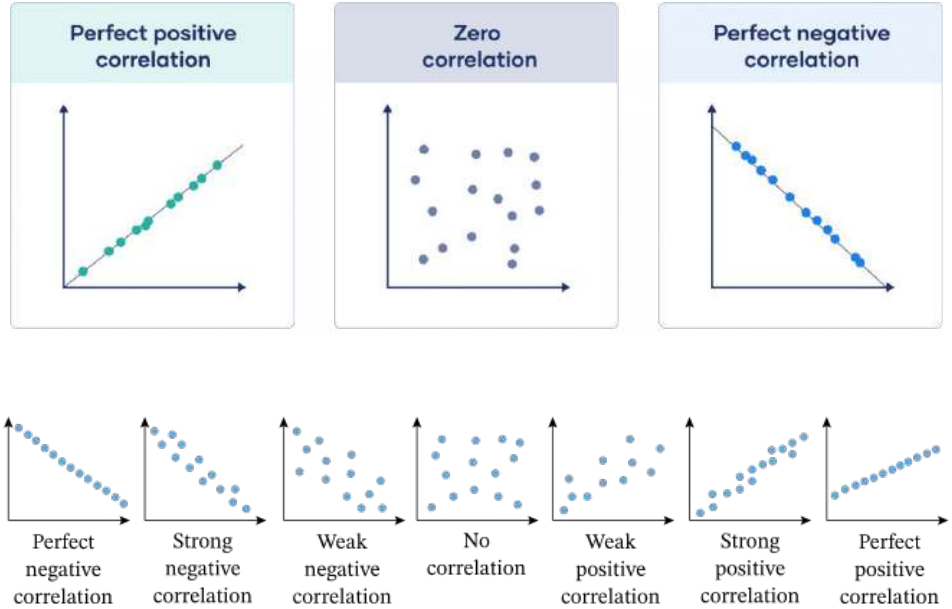
# Korelasyon Nedir?

Korelasyon, iki değişkenin doğrusal olarak ne ölçüde ilişkili olduğunu ifade eden istatistiksel bir ölçüdür. -1 ile 1 arasında değer alır.

-1, güçlü ters orantıyı temsil ederken 1, güçlü doğru orantıyı temsil etmektedir.

Korelasyon değeri 0'a yaklaştıkça iki sütun arasındaki ilişki o kadar zayıflar.

Verideki tüm özelliklerin birbirleri arasındaki korelasyonlarına bakmak istenirse, korelasyon matrisi kullanılabilir.

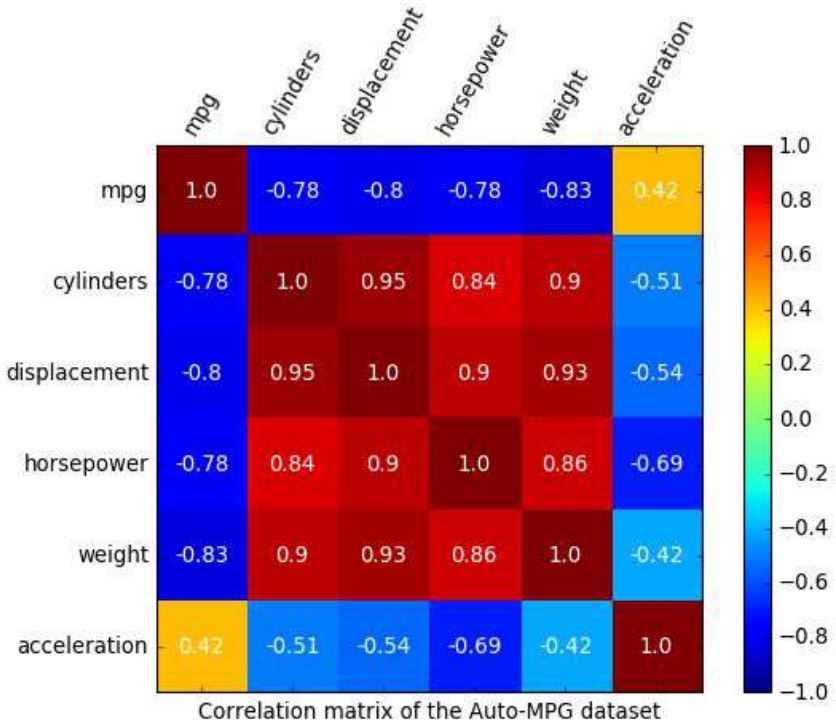


# Korelasyon Matrisi

Bir korelasyon matrisi, tüm sütunların birbirileri ile arasındaki korelasyonunu gösteren bir matristir.

Bu ilişkinin gücünü temsil etmek için hem sayısal değerleri hem de renkleri kullanarak güçlü bir görselleştirme ile bu bilgilerin raporlanmasına olanak sağlar.

*Korelasyon matrisi heatmap görselleştirme yöntemi ile görselleştirilebilir. Bu görselleştirme sonucu yandaki gibi bir matrix elde edilir.*



# Simpson Paradoksu

Simpson Paradoksu, değerler içerisindeki ilişki göz ardı edildiğinde sonucun yanıltıcı olabileceğine dayanır.

*Örneğin, bir okul arşivindeki tüm öğrencilerin kitap okuyor/okumuyor olarak sınıflandırıldığını, ve kitap okuyanların mı yoksa okumayanların mı daha çok arkadaşı olduğunu araştıracağımızı hayal edelim.*

*Bu tabloya baktığımızda kitap okumayanların daha çok arkadaşa sahip olduğunu düşünebiliriz.*

| Kitap | Toplam Sayı | Ortalama Arkadaş Sayısı |
|---|---|---|
| Okumuyor | 101 | 8.2 |
| Okuyor | 103 | 6.5 |

# Simpson Paradoksu

*Yeni bir boyut ekleyip öğrencilerin okul derecesine baktığımızda hem ilkokulda hem de ortaokulda kitap okuyanların daha çok arkadaşı olduğunu görebiliriz.*

*Yani, kitap okuma - ortalama arkadaş sayısı arasındaki korelasyon bir öncekinin tam tersi çıkmıştır.*

Bunun sebebi, korelasyonun, iki sütun arasındaki ilişkiyi diğer her şeyin eşit olduğunu kabul ederek ölçüyor olmasıdır.

| Kitap | Okul | Toplam Sayı | Ort. Arkadaş Sayısı |
|---|---|---|---|
| Okumuyor | Ortaokul | 35 | 3.1 |
| Okuyor | Ortaokul | 70 | 3.2 |
| Okumuyor | İlkokul | 66 | 10.9 |
| Okuyor | İlkokul | 33 | 13.4 |

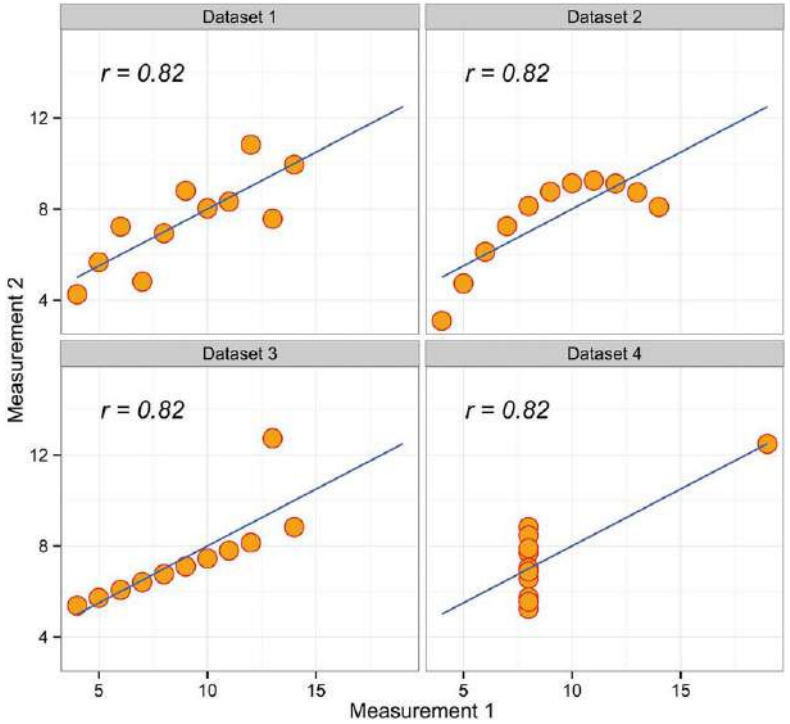
# Anscombe Dörtlüsü (Anscombe Quartet)

Anscombe'un dörtlüsü, hemen hemen aynı basit tanımlayıcı istatistiklere sahip (ortalama, varyans, korelasyon, regresyon), ancak çok farklı dağılımlarda olan ve grafiklendiğinde çok farklı görünen dört veri kümesinden oluşur.

Her veri kümesi 11 nokta (x,y) içerir.

*“Numerical calculations are exact, but graphs are rough”*

- Francis Anscombe



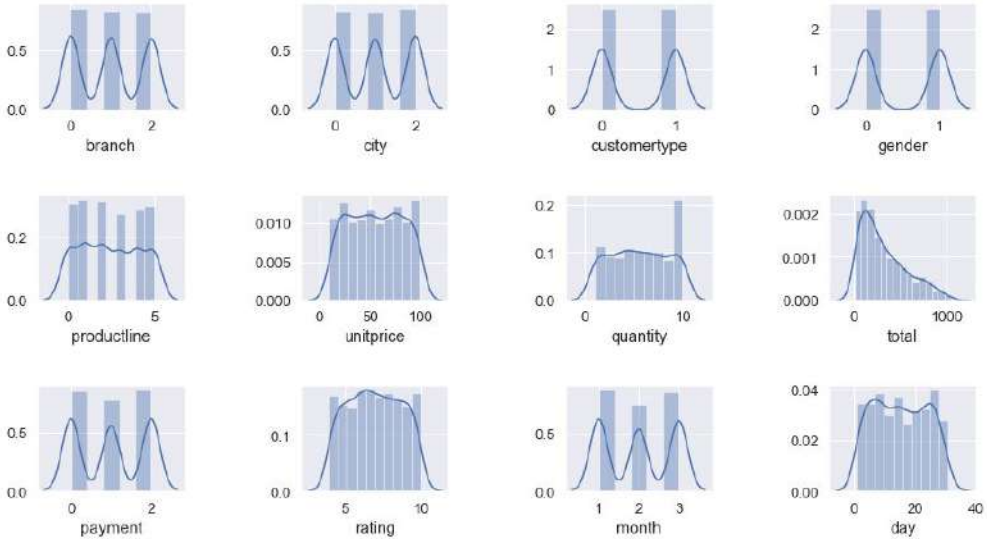
# Veri Dağılımı

# Veri ve Dağılım

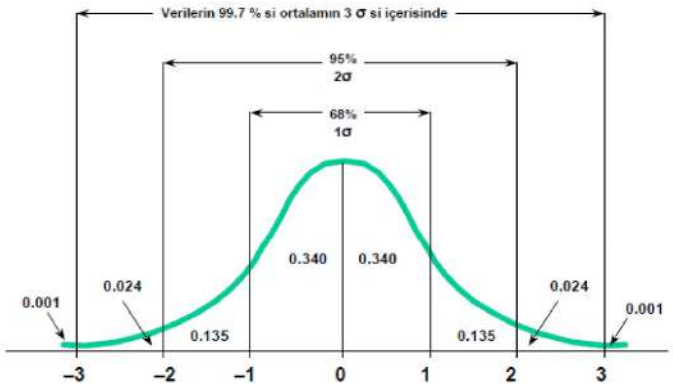
Veri dağılımı, verilerin tüm olası **değerlerini veya aralıklarını** gösterir. Ayrıca her bir değerin frekansını da nicelleştirir.

Verinin karakteristik özellikleri (dengeli, dengesiz vb) veri noktalarının frekansına veya dağılımına bakarak anlaşılabilir.

Dağılım analizi sonucunda, ihtiyaç olursa veri ön işlenir.

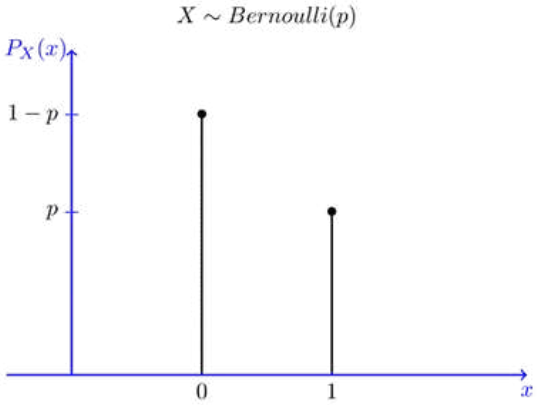


# Kesikli ve Sürekli Olasılık Dağılımları




Sürekli Olasılık Dağılımları

- Normal Dağılım
- Uniform Dağılım
- Üstel Dağılım




Kesikli Olasılık Dağılımları

- Bernoulli
- Binom
- Poison

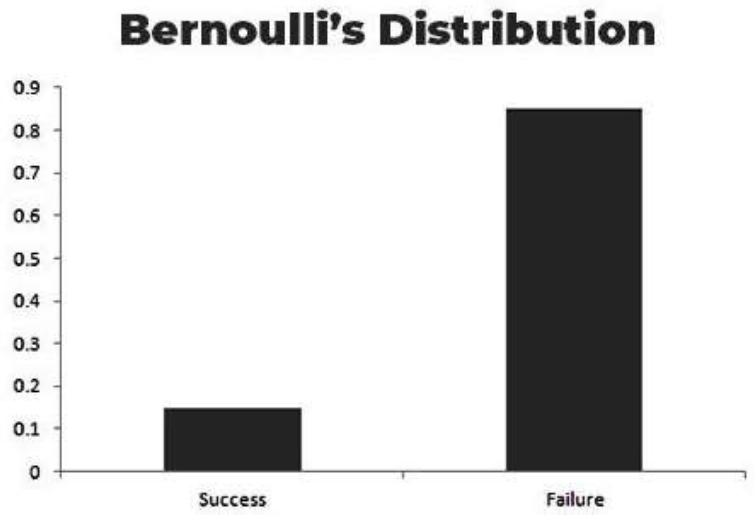
# Bernoulli's Dağılımı

Bernoulli's Dağılımı, n=1 (başarı) ve n= 0 (başarısızlık) olarak muhtemel iki sonucu olan bir dağılımdır.

n=1'in gerçekleşme olasılığı p, n=0'ın gerçekleşme olasılığı 1-p'dir.

*Örneğin, yazı tura atıldığında, bir sonucun tura gelme olasılığı p, sonuç olarak yazı gelme olasılığı (1-p) olur.*

$$f(x) = p^x (1-p)^{(1-x)} \quad \text{where } x \in (0,1)$$

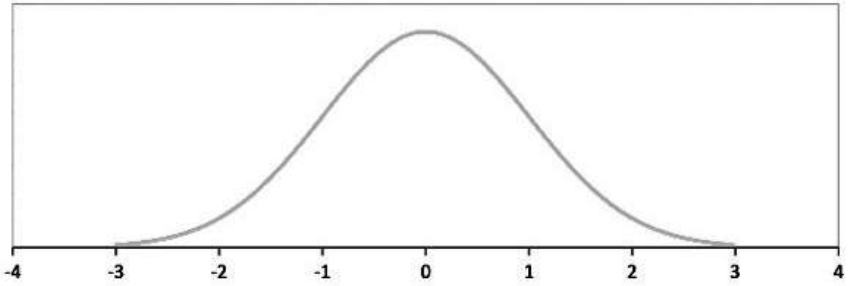


# Gaussian(Normal) Dağılım

Ortalaması etrafında simetrik olan sürekli bir olasılık dağılımı olan normal dağılım, bir değişkenin değerlerinin nasıl dağıldığını tanımlar.

Normal dağılım aşağıdaki özelliklere sahiptir;

- Ortalama, mod ve medyan birbiriyle örtüşür
- Dağılım, çan şeklinde bir dağılım eğrisine sahiptir
- Dağılım eğrisi merkeze simetriktir
- Eğrinin altında kalan alan 1'e eşittir




Formül notasyonunda;

- μ = Mean value
- σ = Olasılığın standart olasılık dağılımı
- x = Rastgele değişken

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

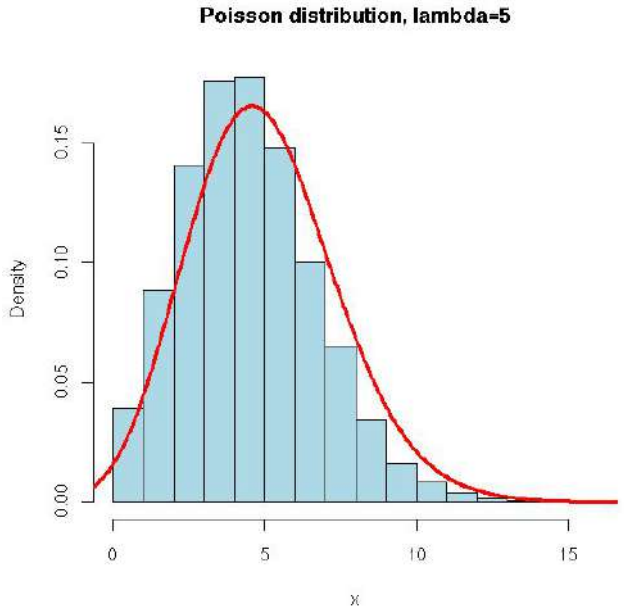
Ortalama (μ) = 0 ve standart sapma (σ) = 1 ise dağılım normal kabul edilir.

# Poisson Dağılımı

- Bir ayrık olasılık dağılımı olup belli bir sabit zaman birim aralığında meydana gelme sayısının olasılığını ifade eder.
- Bu zaman aralığında ortalama olay meydana gelme sayısının bilindiği ve herhangi bir olayla onu hemen takip eden olay arasındaki zaman farkının, önceki zaman farklarından bağımsız oluştuğu kabul edilir.

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$E(X) = \lambda \quad Var(X) = \lambda$$



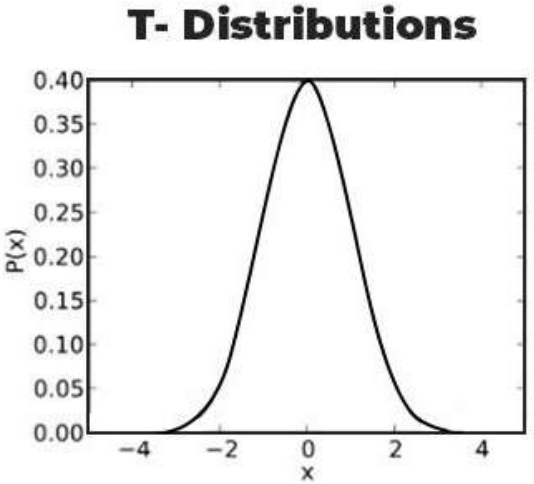
# t-Dağılımı

Popülasyonun standart sapması bilinmediğinde ve gözlemler normal olarak dağılmış bir popülasyondan geldiğinde betimsel istatistik için kullanılan bir sürekli olasılık dağılımıdır.

Örnek sayısı az olduğunda normal dağılım yerine kullanılır. Örneklem boyutu ne kadar büyük olursa, t dağılımı o kadar normal dağılıma benzer.

t dağılımı aşağıdaki özelliklere sahiptir;

- Normal dağılıma benzer şekilde, t-dağılımı çan şeklinde bir eğriye sahiptir ve ortalama sıfır olduğunda simetriktir
- Varyans değeri her zaman birden fazladır
- Dikkate alınan örneklem büyüklüğünün 30'dan büyük olduğu durumlarda normal dağılım gibi davranır

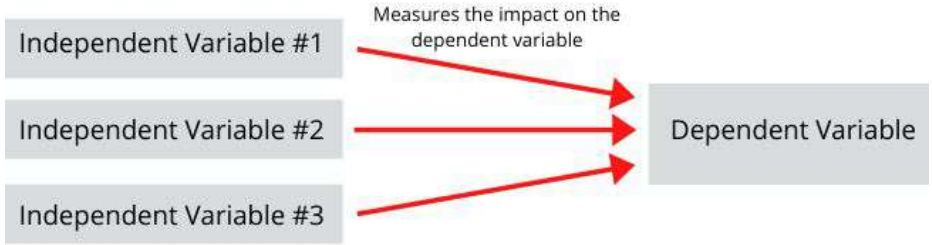


# Serbestlik Derecesi

Serbestlik Dereceleri, veri örneğinde değişkenlik gösterme özgürlüğüne sahip değerlerdir ve bağımsız değerlerin maksimum sayısını ifade eder.

*Örneğin; beş pozitif tam sayıdan oluşan bir veri örneğinde serbestlik derecesi 5'tir çünkü değerler değişimlerde birbirini etkilemez ve aralarında bir ilişki yoktur.*

*Ancak beş pozitif tam sayıdan sonuncusu, ilk 4 sayının ortalaması ise burada serbestlik derecemiz 4 olur. Veri örneklerimizin: 2,3,6,5 olduğu varsayılırsa son değer ortalama değeri yani 4 olmalıdır. Son değerin değişme özgürlüğü yoktur ve ilk 4 sayıya bağlıdır.*



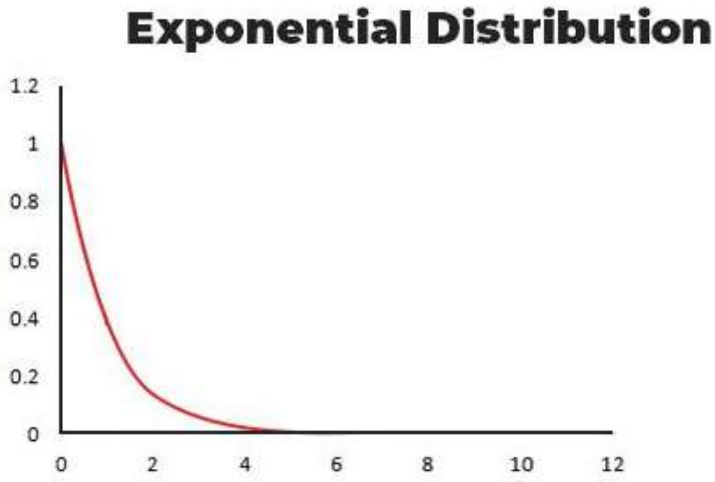
# Exponential Dağılım

Exponential Dağılım, bir olayın belli bir süre içerisinde gerçekleşip gerçekleşmeyeceğinin olasılığını veren bir dağılımdır. Bu süre saniye, saat, yıl veya yüzyıl gibi ölçütlerde olabilir.

*Örneğin bir mağazaya belli bir süre içinde müşterinin gelip gelmeyeceği olasılığı Exponential Dağılıma örnektir.*

λ: Poisson Dağılımı izleyen olaylar (başarılar) arasındaki ortalama zaman olarak tanımlanır. X hangi zaman birimindeyse, λ da o zaman biriminde olmalıdır

X: olayın gerçekleşeceği zamanı temsil eder




$$f(x) = \begin{cases} \lambda e^{-\lambda * x}, x \geq 0 \\ 0, x < 0 \end{cases}$$

# Exponential Dağılım Örneği

Müşterilerin saatte ortalama 15 kez kahve sipariş ettiği bir kahve dükkanında, sonraki kahve siparişinin 5 dakika sonra gelme olasılığının hesaplanmak istendiğini düşünelim.

x = 5 olacaktır. x dakika cinsinden olduğundan λ değerini dakika cinsinden bulmamız gerekir. Bir saatte 60 dakika olduğundan, 15/60'tan itibaren kahve siparişleri dört dakikada bir (1/4) gelir. Bu hesapla λ = 1/4 olacaktır.

λ=1/4 ve x=5 olan olasılık yoğunluk fonksiyonu, tam beşinci dakikada sipariş gelme olasılığını verir.

$$\lambda \cdot e^{-(\lambda x)} = \frac{1}{4} \cdot e^{-(\frac{1}{4}) \cdot 5} = 0.0716262$$

$$p\left(x = 5, \lambda = \frac{1}{4}\right) = 0.0716262$$

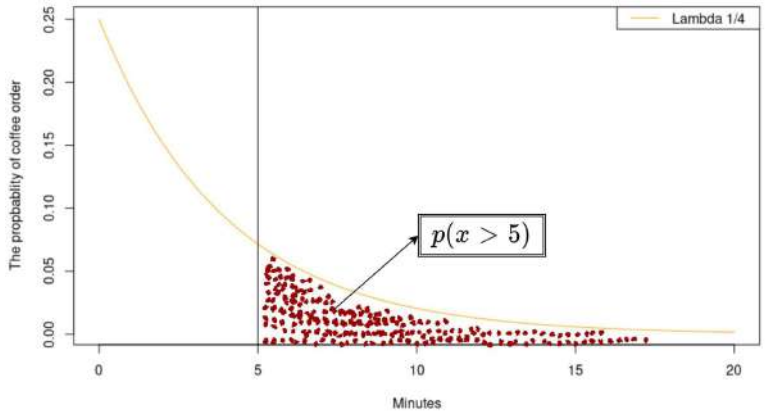
# Exponential Dağılım Örneği

Tam 5.dakikada sipariş gelme olasılığı 0.07 olarak bulunmuştur.

λ değerinin 1/4 olduğu tüm olası dakikalarda emir alma olasılık değerleri yandaki şekilde gösterilmektedir.

Beş dakika içinde sipariş alma olasılığını bulmak için 5. dakikadan sonra eğrinin altındaki alanın hesaplanması gerekir.

Bunun için integral alınabileceği gibi üstel dağılımın kümülatif olasılığını veren başka bir formül de kullanılabilir. Bu formül belirli bir olayın x biriminden önce meydana gelme olasılığını verir.




$$F(x, \lambda) = \begin{cases} 1 - e^{-(\lambda x)} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

# Exponential Dağılım Örneği

Örneğe göre bu formül siparişin 5 dakikadan önce gelme olasılığını vermektedir.

Bir siparişten 5 dakika sonra sipariş gelme olasılığı da bu formül sonucunun 1 sayısından çıkarılmasıyla bulunabilir.

Yandaki sonuca göre bir sonraki siparişin 5 dakika sonra gelme olasılığı 0.28 olarak elde edilmiştir. Bu, bir sonraki siparişi 5 dakika önce alma olasılık değerinin 5 dakika sonra alma olasılığından daha yüksek olduğu anlamına gelir.

$$f\left(x \leq 5, \lambda = \frac{1}{4}\right) = 1 - e^{-\frac{1}{4} \cdot 5}$$

$$f\left(x > 5, \lambda = \frac{1}{4}\right) = 1 - f\left(x \leq 5, \lambda = \frac{1}{4}\right)$$

$$= 1 - \left(1 - e^{-\frac{1}{4} \cdot 5}\right)$$

$$= e^{-\frac{1}{4} \cdot 5}$$

$$= 0.2865048$$