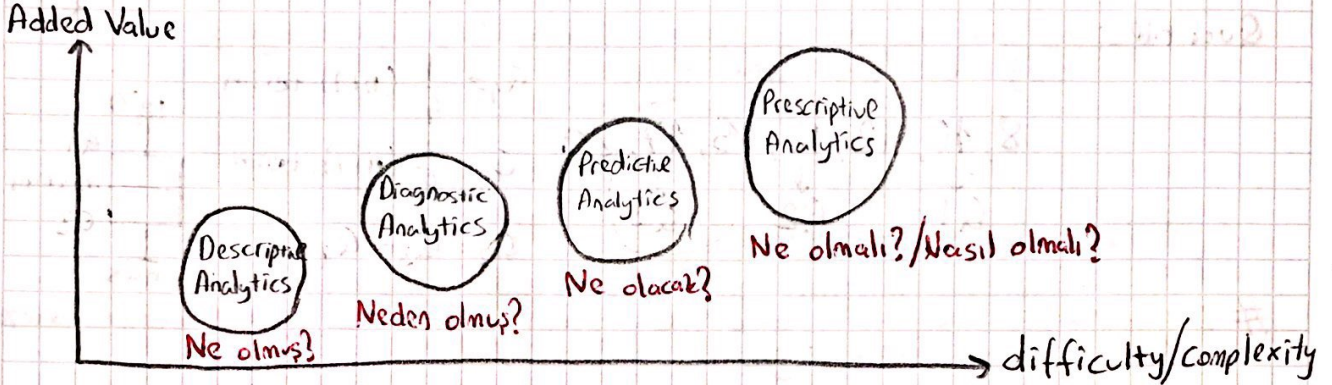
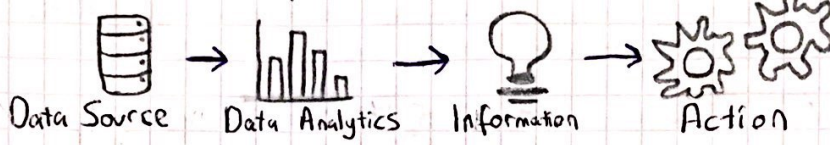


DATA LITERACY

Veri bilimci = The person who manages the process of extracting useful info. from data.



→ Veriler genellikle şirketlerin ilişkisel veri tabanları veya sigmadıysa big data ortamlarından veri kaynaklarından alınır. R veya Python gibi ortamlara taşınır. Analizi ve modellenmesi yapılır.

→ Üretilen model fonksiyonunun katsayılarını bulmaya parametre tahmin yöntemleri veya optimizasyon yöntemleri adı verilir.

→ Modelimizi canlıya çıkarmaya "deployment" denir.

Population = İlgilendirmiş ana kitlemiz, hedef kitlemiz. (Population)
Sample = Ana kitle içerisinde seçilen alt kümedir (örneklem)
Ana kitleyi temsil edebilecek ve yansız bir alt kümedir.

Observation Unit = Belirlediğimiz alt küme içerisinde belirlediğimiz her bir birim "gözlem birimi" olarak adlandırılır.

Variable = Birimden birime farklı değer alan nicelik (Araçların kolları)
İkiye ayrılır sayısal değişken ve kategorik değişken (nicel, nitel)
nicel = kantitatif nitel = kalitatif

Scales of Measurement Ölçek türleri

Sayısal değişkenler için: Aralık ve oran **başlangıç noktası sıfır**

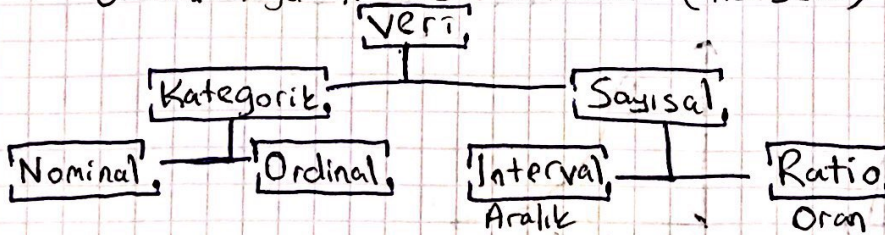
Kategorik değişkenler için: Nominal ve ordinal

Aralık: "+" ve "-" değerler alabilen sayısal değişkenler.

Oran: Başlangıç noktası sıfır olan sayısal değişkenler.

Nominal: Kategorik değişkenin sınıfları arasında hiyerarşi yoksun nominaldır.

Ordinal: Hiyerarşi varsa ordinaldir (Rütbe...)



Merkezi Eğilim Ölçüleri - Measures of Central Tendency

Arithmetic Mean: Tüm değerlerin toplanıp birim sayısına bölünmesidir.

Median: Seriyi küçükten büyüğe sıraladıktan sonra ortadaki değerdir. İki değer varsa onların aritmetik ortalaması medyan olur.

Mode: Bir seride en çok tekrar eden değere mod denir.

Quartiles: Küçükten büyüğe sıralanan seriyi 4 parçaya ayıran değerlere kartiller denir.

8, 10, 12, 14, 15, 17, 20

Q1

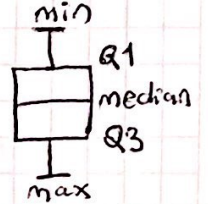
Q2

Q3

$$Q_1 = \frac{1}{4} (n+1). \text{terim}$$

$$Q_3 = \frac{3}{4} (n+1). \text{terim}$$

$$Q_2 = Q_3 - Q_1$$



Ortalama ile medyanın birbirine yakın olması dağılımın düzenli (homojen) olduğu anlamına gelir.

Kıyas yapıldığında ortalama ile medyan birbirinden uzak ise medyan esas alınır.

Measures of Dispersion - Dağılım Ölçüleri

Range: Bir serideki max değerden min değeri çıkardığımızda değişim aralığı (range)'nı elde ederiz.

Standard Deviation: Ortalamadan sapmanın genel bir ölçüsüdür.

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

	ana kitle	örneklem
Ortalama	μ (nü)	\bar{x}
S. sapma	σ (sigma)	s

Variance: Standart sapmanın karesidir. Ortalamadan olan sapmaların karelerinin ortalamasıdır.

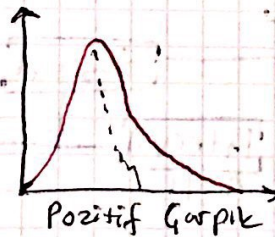
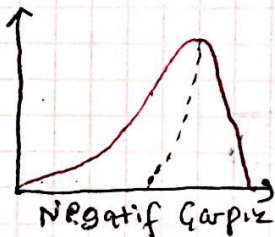
$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Skewness: Bir değişkenin dağılımının simetrik olmayaşına çarpıklık denir. Çarpıklık olduğunda medyan değeriyle merkezi eğilim temsil edilir.

Pearson
Çarpıklık
Katsayısı

$$\frac{3(\bar{x} - \text{medyan})}{\text{Standart sapma}}$$

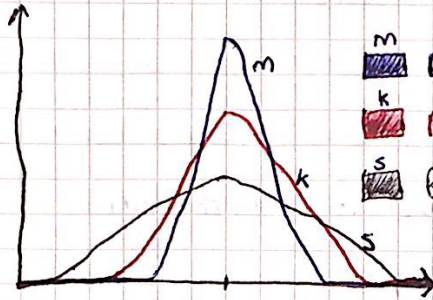
$P\&K < 0 \rightarrow$ Negatif çarpık (soldan)
 $P\&K > 0 \rightarrow$ Pozitif çarpık (sağdan)
 $P\&K = 0 \rightarrow$ Simetrik



Kurtosis: Elimizdeki değişkenin dağılımının sivri mi yoksa basık mı olduğunu ifade eden ölçme basıklık denir.

$$\text{basıklık katsayısı} = \frac{m_4}{s^4} = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{n}$$

Standart sapmanın 4. kuvveti



- $\text{BK} > 3$ ise dağılım sivridir.
- $\text{BK} = 3$ ise dağılım standart normal dağılıma uygundur.
- $\text{BK} < 3$ ise dağılım basıktır.

Statistical Thinking Models: Veri okuryazarlığından veri analitiğine giden yolu modelleyen yol göstericilerdir.

- Ben-Zvi ve Friedlander (1997)
- Wild ve Pfannkuch (1999)
- Jones ve Digerleri (2000)
- Hoerl ve Snee (2001)
- Mooney (2002)

Mooney Modeli: İş hayatına uygunluğuyla daha çok tercih edilen modeldir.

- ① Verinin Tanımlanması (değişkenleri belirleme, anormallikleri belirleme)
- ② Verinin Organize Edilmesi ve İndirgenmesi (veriyi düzenleme)
- ③ Veri Gösterimi (Görselleştirme ve sunum hazırlığı)
- ④ Verinin Analiz Edilmesi ve Yorumlanması

İstatistiksel Düşünce Seviyeleri

- Kişiyi Özgünlük (Seviye 1)
- Geçici (Seviye 2)
- Nicel (Seviye 3)
- Analitik (Seviye 4)

Merkezi eğilim ölçütleri kitleye ilişkin bir değişkenin bütün farklı değerlerinin çevresinde toplandığı merkezi bir değeri gösterirler. Dağılım ölçütleri ise aldığı değerlerin birbirinden ne kadar farklı olduğunu ölçsüdü.

Değişken

En az iki değer alabilen ve durumdan duruma değişebilen nitelik ve özelliklerdir. Uzunluk, boy, ağırlık, yetenek... Bunların hepsi bir değişken olabildiği gibi aynı nitelik farklı değişken özelliği de gösterebilirler.

Değişken Türleri

1) **Nicel Değişken:** Sayılarla ifade edilebilen değerlerdir. Burada anlam sayılarda gizlidir.

→ Ali sınavdan 85 aldı. → "85" niceliktir.

2) **Nitel Değişken:** ifadenin betimlenme durumudur. Sayılar kullanılabilir ancak buradaki sayıların ifadesi ölçme anlamında sayısal bir değeri yoktur. Sadece betimlemedir.

→ Ali başarılı bir öğrencidir. → "başarılı" niteliktir.

3) **Süreklili Değişken:** İki birim arasında sonsuz bölünebilirlik şansı varsa bu ifade süreklidir. Metre, kilogram, yaş bunlara örnektir. Örneğin 3 ile 4 yaş arasında ay, hafta, gün ya da saat özelliğine varana kadar araya birçok değişik ifade girebilmektedir. Ayrıca sayılarla ifade edildiği için aynı zamanda niceldir.

4) **Süreksiz Değişken:** İki birim arasında bölünme şansı yoksa süreksizdir. Cinsiyet, tam sayılar bunlara örnek teşkil eder. Bu ifadeler genelde nitel ifade olarak da karşımıza çıkar.

5) **Bağımlı Değişken:** Başka bir değişkene bağlı olarak değişen değişkenlere bağımlı değişken denir.

6) **Bağımsız Değişken:** Başka bir değişkene bağlı olmadan artan ya da azalan değişkenlere denir.

7) **Ara Değişken:** Bağımsız ifadenin bağımlı ifade üzerindeki etkisini artırması ya da azaltması durumudur.