SQL / 1 Giriş

Popülasyon: ilgilendiğimiz ana kitle

Örneklem: popülasyon içinden seçilen bir alt kümedir, popülasyonun özelliklerini çok iyi temsil etmelidir.

Gözlem Birimi: örneklem içerisindeki her bir elemana verilen isim, tablodaki her bir satır

Araç Fiyatı	KM	Vites Türü	Hasar Durumu	Marka	Model
10000	300000	Manuel	Evet	A	A1
54000	40000	Manuel	Hayır	A	A2
46999	90000	Manuel	Evet	B	B1
89000	1000000	Otomatik	Evet	C	C1
70000	78000	Otomatik	Evet	B	B2
50000	30000	Manuel	Hayır	C	C2
NA	600000	Manuel	Hayır	C	C3
68900	50000	Otomatik	Hayır	B	B2
12000	200000	Manuel	Hayır	A	A1

Değişken: Birimden birime farklı değerler alan niceliktir.

“Cinsiyet” kategorik değişkendir. Ve “Kadın” ve “Erkek” ise bu kategorik değişkenin sınıflarıdır.

Ölçek Türleri: Bir değişkenin değerlerini anlayabilmemiz için kullanılır.

- Sayısal değişkenler için: Aralık ve Oran
- Kategorik değişkenler için: Nominal ve Ordinal
- Nominal: “Kadın” ve “Erkek” sınıfları arasında fark olmadığı için bu değişken nominal ölçek türüne sahiptir. 0, 1, 2 gibi sayılara indirgenebilir.
- Ordinal: “Rütbe” isimli bir değişkenimiz olsun. Sınıfları onbaşı<yüzbaşı<binbaşı<albay gibi derece farkları varsa ordinaldir. Ya da eğitim durumu örnek olabilir.

Merkezi Eğilim Ölçüleri:

- Aritmetik Ortalama: Tüm değerlerin toplanıp gözlem sayısına bölünmesiyle elde edilir. Simetrik veri setlerini analiz etmek için uygundur.
- Medyan (Ortanca): Sıralı verilerin tam ortasındaki değer; büyükten küçüğe ya da küçükten büyüğe fark etmez.

  NOT: Aşırı değerlerin olduğu veri setlerinde merkezi eğilimi daha iyi yansıtır.
- Mod (Tepe Değer): Verilerde en çok tekrar eden değer, unique değer sayısı

Bu ölçüler, veri setinin merkezini ve yoğunlaştığı noktaları anlamamıza yardımcı olur.

NOT: Ortalama ile medyanın birbirine çok yakın olması dağılımın homojen olduğunu gösterir.

Kartiller:

Verileri dört eşit parçaya bölen değerlerdir.

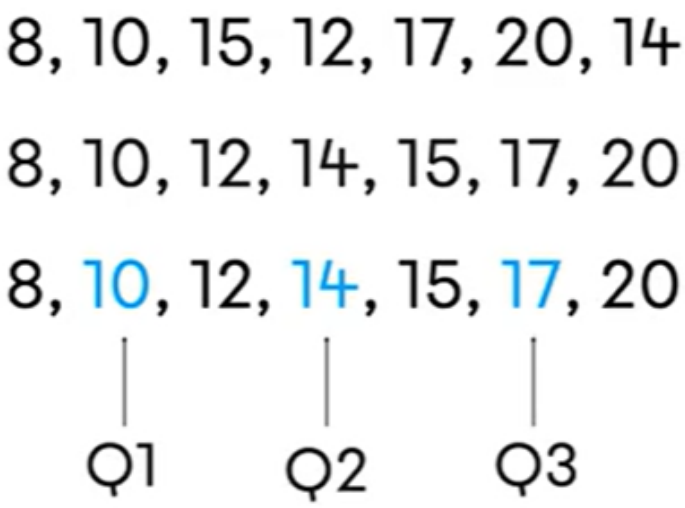
- Q1 (1. Kartil): Verilerin %25'ini altında bırakan değer
- Q2 (2. Kartil): Medyan değeridir, verilerin %50'sini altında bırakan değer
- Q3 (3. Kartil): Verilerin %75'ini altında bırakan değer

Kartiller arası aralık (IQR) = Q3 - Q1 formülü ile hesaplanır ve veri setinin yayılımını gösterir.

NOT:

- Q1 = 1/4 (n + 1) * terim
- Q3 = 3/4 (n + 1) * terim
- Q2 = Q3 - Q1 = MEDYAN

n: gözlem sayısı



Dağılım Ölçüleri:

- Varyans: Verilerin ortalamadan ne kadar uzaklaştığını gösteren ölçüdür. Sapmaların karelerinin ortalamasıdır.
- Varyans Kullanım Alanları:
  - Veri setindeki değişkenliği ölçer ve risk analizinde kullanılır
  - Finansal analizlerde yatırım riskini değerlendirmek için önemlidir
  - Kalite kontrol süreçlerinde ürün tutarlılığını ölçmede kullanılır
  - İstatistiksel testlerde ve model değerlendirmelerinde temel bir ölçüdür
- Standart Sapma: Varyansın karekköküdür. Verilerin ortalamadan uzaklığının ortalamasını gösterir.
- Değişim Katsayısı: Standart sapmanın ortalamaya bölümüdür. Farklı veri setlerini karşılaştırmak için kullanılır.
- Çarpıklık: Dağılımın simetrisini ölçer. Pozitif çarpıklık sağa, negatif çarpıklık sola çarpık dağılımı gösterir.

💡 Pearson Çarpıklık Katsayısı:
• Eğer katsayı 0'a yakınsa, dağılım simetriktir
• Katsayı > 0 ise, dağılım sağa çarpıktır
• Katsayı < 0 ise, dağılım sola çarpıktır

Not: -1 < p < +1 arasında olabilir.

Formül: 3 × (Ortalama - Medyan) / Standart Sapma

- Basıklık: Dağılımın sivriliğini ölçer. Normal dağılıma göre daha sivri veya daha basık olmasını gösterir.

💡 Basıklık (Kurtosis) Formülü:
$K = [\Sigma(x - \mu)^4/n] / \sigma^4$

Burada:
• x: Veri noktaları, her bir terim
• μ: Ortalama
• σ: Standart sapma
• n: Gözlem sayısı

Normal dağılım için K = 3'tür.
K > 3 ise dağılım sivridir (leptokurtik)
K < 3 ise dağılım basıktır (platikurtik)

İstatistiksel Düşünce Modelleri

Tanımlayıcı İstatistik (Descriptive Statistics)

Tanımlayıcı istatistik, verileri özetlemek ve anlamak için kullanılır. Bu model, verilerin merkezi eğilimini, dağılımını ve şeklini tanımlar.

- **Merkezi Eğilim:** Ortalama (mean), medyan ve mod gibi ölçümler.
- **Dağılım:** Standart sapma, varyans, aralık (range) ve çeyrekler (quartiles).
- **Görselleştirme:** Grafikler, histogramlar, kutu grafikler (box plot) gibi araçlar kullanılarak verilerin görsel sunumu.

Çıkarımsal İstatistik (Inferential Statistics)

Çıkarımsal istatistik, örneklem verilerinden genel popülasyon hakkında sonuçlar çıkarmak için kullanılır. Bu model, hipotez testleri ve güven aralıkları ile popülasyon hakkında tahminler yapar.

- **Hipotez Testleri:** T-test, ANOVA, ki-kare testleri ve diğer testler.
- **Güven Aralıkları:** Bir popülasyon parametresinin tahmini aralığını belirlemek için kullanılır.
- **Regresyon Analizi:** İki veya daha fazla değişken arasındaki ilişkiyi modellemek için kullanılır.

Korelasyon ve Regresyon

Korelasyon ve regresyon analizleri, değişkenler arasındaki ilişkileri inceler. Korelasyon, iki değişken arasındaki ilişkinin gücünü ve yönünü ölçer; regresyon ise bu ilişkinin doğrusal modelini oluşturur.

- **Korelasyon Katsayısı (r):** İki değişken arasındaki ilişkinin gücünü ölçer.
- **Doğrusal Regresyon:** Y = a + bX şeklinde bir doğrusal ilişkiyi modelleyen analiz.

Çok Değişkenli İstatistik (Multivariate Statistics)

Çok değişkenli istatistik, birden fazla değişkenin analizini içerir. Bu model, karmaşık veri setlerini analiz etmek ve çoklu ilişkileri anlamak için kullanılır.

- **Faktör Analizi:** Verideki gizli yapıları keşfetmek için kullanılır.
- **Kümeleme (Clustering):** Veriyi benzer gruplara ayırmak için kullanılır.
- **Başlıca Bileşenler Analizi (PCA):** Verinin boyutunu azaltmak ve temel bileşenleri keşfetmek için kullanılır.

Zaman Serisi Analizi (Time Series Analysis)

Zaman serisi analizi, düzenli aralıklarla toplanan verileri inceler. Bu model, verideki desenleri ve eğilimleri anlayıp gelecek değerleri tahmin etmek için kullanılır.

- **Hareketli Ortalama (Moving Average):** Verideki kısa vadeli dalgalanmaları düzeltir.
- **ARIMA (AutoRegressive Integrated Moving Average):** Zaman serisi tahmini için kullanılan gelişmiş bir model.

Bayesçi İstatistik (Bayesian Statistics)

Bayesçi istatistik, olasılık dağılımlarını kullanarak verileri analiz eder. Bu model, önceki bilgiler ile yeni verileri birleştirerek tahminler yapar.

- **Bayes Teoremi:** Olasılıkların güncellenmesini sağlar.
- **Bayesçi Çıkarım:** Önsel (prior) ve ardıl (posterior) olasılıkları kullanarak tahmin yapar.

MONEEY Modeli

MONEEY modeli, istatistiksel analizlerde kullanılan önemli bir yaklaşımdır:

- **M**easure (Ölç): Veriyi doğru şekilde topla ve ölç
- **O**bserve (Gözlemle): Verideki desenleri ve eğilimleri gözlemle
- **N**otice (Fark Et): Önemli değişimleri ve anomalileri belirle
- **E**valuate (Değerlendir): Bulguları detaylı şekilde analiz et
- **E**xplain (Açıkla): Sonuçları açık ve anlaşılır şekilde ifade et
- **Y**ield (Sonuç Al): Analiz sonuçlarından actionable insights elde et

💡

İstatistiksel Düşünce Düzeyleri

- **Seviye 1: Kişiye Özgü (Okuyamaz):** Bu seviyede olan kişiler, verileri genellikle kişisel deneyimlerine ve ön yargılarına göre yorumlarlar. Verilerin objektif bir analizi yerine, kendi inançlarına uygun sonuçlar çıkarırlar.
- **Seviye 2: Geçici (Biraz Okuyabilir):** Bu seviyedeki kişiler, verileri biraz daha sistematik bir şekilde incelemeye başlarlar. Ancak, verilerin arkasındaki istatistiksel kavramları tam olarak anlamazlar ve basit grafikleri bile yorumlamakta zorlanabilirler.
- **Seviye 3: Nicel (Okuyabilir):** Bu seviyede olan kişiler, verileri sayısal olarak değerlendirmeye başlarlar. Ortalama, medyan gibi temel istatistiksel kavramları kullanabilirler. Ancak, verilerin arasındaki ilişkileri ve daha karmaşık analizleri yapmakta zorlanabilirler.
- **Seviye 4: Analitik (Okuyabilir ve Analiz Edebilir):** Bu seviyedeki kişiler, verileri derinlemesine analiz edebilirler. İstatistiksel yöntemleri kullanarak veriler arasında anlamlı ilişkiler kurarlar, hipotezler oluşturur ve bu hipotezleri test ederler.

Veri Analitiği Türleri

Betimleyici Analitik (Descriptive Analytics):

- **Ne oldu?** sorusuna cevap verir.
- Geçmiş verileri kullanarak mevcut durumu özetler.
- Raporlama, grafikler ve tablolar gibi görselleştirme yöntemleri kullanılır.
- Örneğin, bir şirketin aylık satış rakamlarını analiz etmek açıklayıcı analitiğe örnektir.

Tanısal Analitik (Diagnostic Analytics):

- **Neden oldu?** sorusuna cevap verir.
- Açıklayıcı analitik sonuçlarının nedenlerini araştırır.
- Kök neden analizleri, korelasyon analizleri gibi yöntemler kullanılır.
- Örneğin, satışlarda düşüş yaşayan bir ürünün nedenlerini bulmak tanısal analitiğe örnektir.

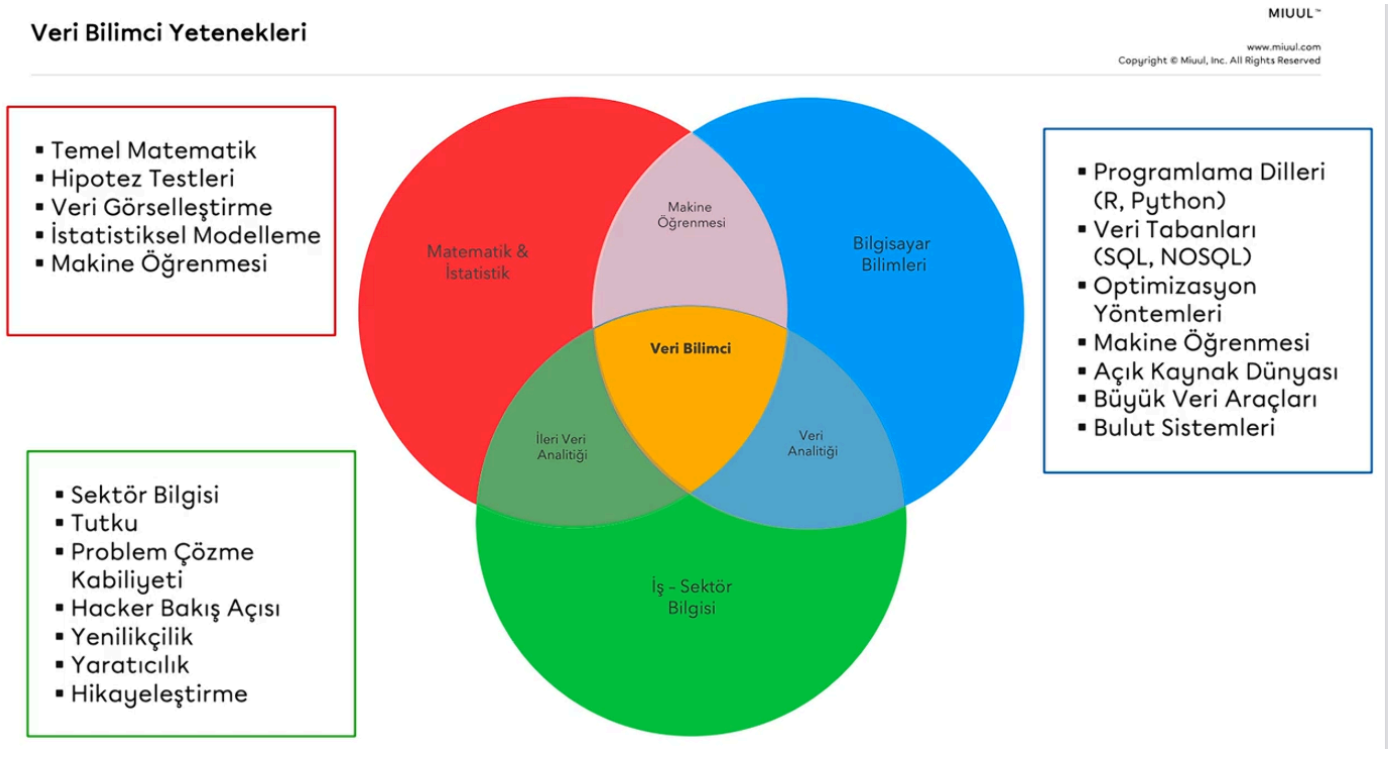
Tahmine Dayalı Analitik (Predictive Analytics):

- **Ne olacak?** sorusuna cevap verir.
- Geçmiş verileri kullanarak gelecekteki olayları tahmin eder.
- İstatistiksel modeller, makine öğrenimi algoritmaları gibi yöntemler kullanılır.
- Örneğin, müşteri kaybını tahmin etmek, ürün satışlarını öngörmek tahmine dayalı analitiğe örnektir.

Öngörüye Dayalı Analitik (Prescriptive Analytics):

- **Ne yapmalıyız?** sorusuna cevap verir.
- Tahminlere dayanarak en iyi eylem planını önerir.
- Simülasyonlar, optimizasyon algoritmaları gibi yöntemler kullanılır.
- Örneğin, bir üretim sürecindeki verimliliği artırmak için en uygun parametreleri belirlemek öngörüye dayalı analitiğe örnektir.

Gerekli Yetenekler



Olası Proje Alanları