Eksik Verilerle Baş Etme Stratejileri: Bir Veri Bilimcinin En Sessiz Düşmanı

# Giriş: Eksik Veri Neden Önemlidir?

Gerçek hayattaki veriler hiçbir zaman kusursuz değildir. Anketlerde boş geçilen sorular, arızalı sensörlerden gelen eksik kayıtlar ya da sistem hataları... Hepsi elimizdeki veriyi eksik hale getirir.  
  
Ama önemli bir soru var:  
Bu eksik verilerle ne yapmalıyız?  
Yok saymalı mıyız? Tahmin mi etmeliyiz? Sadece ortalama ile mi doldurmalıyız?  
  
Bu yazıda eksik verilerle başa çıkmanın farklı yollarını ve bunların ne zaman, nasıl kullanılacağını inceleyeceğiz.

# 1. Eksik Veri Türlerini Tanımak

Eksik verileri anlamak, onları düzeltmeden önceki en önemli adımdır:  
  
- MCAR (Missing Completely at Random): Veri tamamen rastgele bir şekilde kayıp.  
- MAR (Missing at Random): Kaybın bir nedeni var ama bu neden başka bir değişkenle ilişkili.  
- MNAR (Missing Not at Random): Kaybın nedeni kayıp olan verinin kendisiyle ilgili.

# 2. Eksik Veriyi Tespit Etmek

Python'da eksik verileri şöyle tespit edebiliriz:

import pandas as pd  
  
df = pd.read\_csv("your\_dataset.csv")  
print(df.isnull().sum())  
print(df.isnull().mean())

# 3. Ne Yapmalı? Eksik Veriyle Başa Çıkma Yöntemleri

## a) Silmek (Drop)

Avantaj: Basit ve hızlıdır.  
Dezavantaj: Değerli bilgileri kaybetme riski yüksek.

## b) Doldurmak (Imputation)

1. Sabit Değerle Doldurmak  
2. Ortalama / Medyan / Mod ile Doldurmak  
3. Gruplandırarak Doldurma  
4. Tahmin ile Doldurmak

df["age"] = df.groupby("city")["age"].transform(lambda x: x.fillna(x.mean()))

# 4. İleri Seviye: Çoklu İmputasyon (Multiple Imputation)

Tek bir tahmin yerine, birden fazla tahmin yapılır ve belirsizlik dikkate alınır. Özellikle istatistiksel analizlerde daha güvenilir sonuçlar verir.  
Python'da fancyimpute veya sklearn.experimental.IterativeImputer gibi araçlarla yapılabilir.

# 5. Hangi Yöntemi Ne Zaman Seçmeliyim?

Senaryo - Önerilen Yöntem  
Çok az eksik veri - Silme (drop)  
Sayısal veriler, MCAR varsayımı - Ortalama / medyan doldurma  
Kategorik veriler - Mod / sabit değerle doldurma  
Eksik veriler karmaşık yapıda - Tahmine dayalı doldurma  
Araştırma / analiz istatistikselse - Çoklu imputasyon

# Sonuç: Eksik Veriyle Dans Etmeyi Öğrenmek

Eksik veri her zaman kötü değildir; bazen onun neden eksik olduğuna bakarak da içgörü kazanabiliriz. Ama onu görmezden gelmek çoğu zaman bir modelin başarısını düşürür.  
  
Doğru stratejiyi seçmek, iyi bir veri bilimcisinin fark yarattığı noktalardan biridir.