amsun Üniversitesi
Yazılım Mühendisliği Bölümü

# VERİ MADENCİLİĞİ (DATA MİNİNG)

Dr. Öğr. Üyesi Alper Talha Karadeniz

2025-2-26

# Hafta 4

## Veri Görselleştirme ve Keşifsel Veri Analizi (EDA)

Samsun Üniversitesi
Yazılım Mühendisliği Bölümü

**01**

Histogram, kutu grafiği, scatter plot, pairplot

**02**

EDA'nın modelleme öncesi rolü

**03**

Uygulama

# Histogram Nedir?

Histogram, verilerin frekans dağılımını görselleştirmek için kullanılan bir grafiktir. Veri setindeki sayısal değerleri belirli aralıklara (bin) böler ve her aralıkta kaç tane veri olduğunu çubuk yüksekliği ile gösterir.

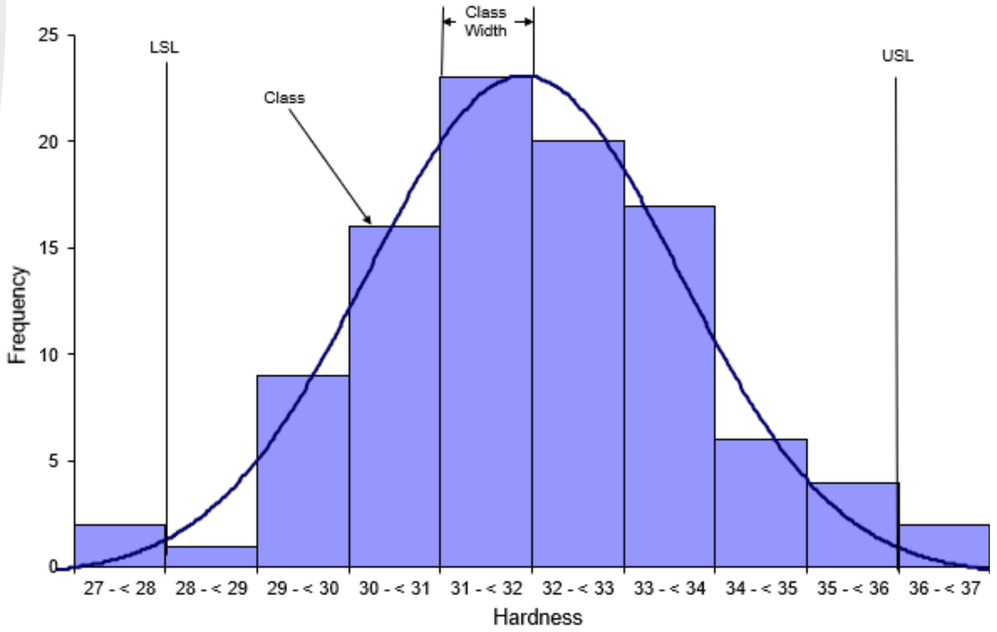
## Histogram Nerelerde Kullanılır?

- Verinin dağılımını görmek (ör. normal dağılıyor mu, sağa/sola çarpık mı).
- Veri ön işlemde, normalizasyon/standardizasyon öncesi dağılım analizi.
- Aykırı değer olup olmadığını anlamak.

## Histogram Nasıl Yorumlanır?

- Simetrik ise veri ortalama etrafında dengeli dağılmıştır.
- Sağa çarpık (right skew): Ortalama > Ortanca, büyük değerler kuyrukta.
- Sola çarpık (left skew): Ortalama < Ortanca, küçük değerler kuyrukta.

Samsun Üniversitesi
Yazılım Mühendisliği Bölümü



# Kutu Grafiği Nedir?

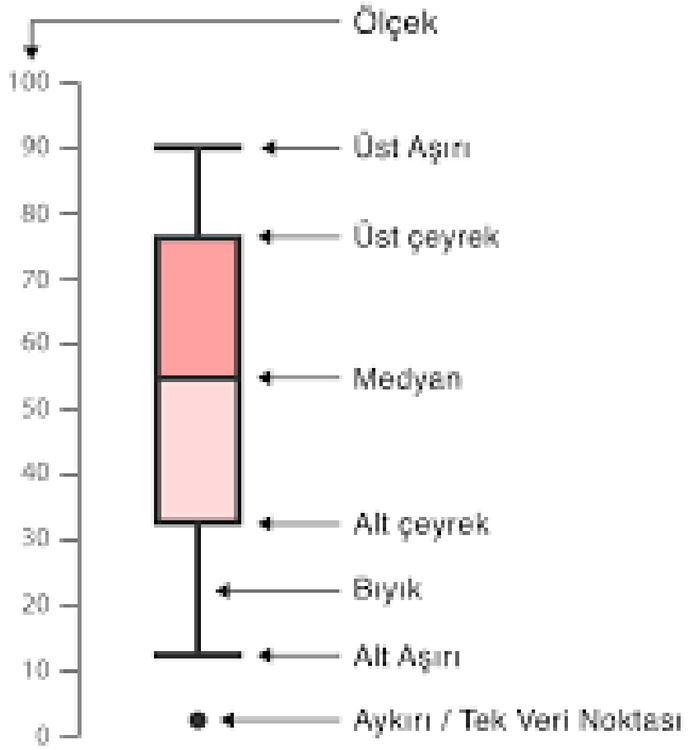
bir veri setinin dağılımını, merkezi eğilimini ve değişkenliğini görselleştiren bir grafik türüdür. Ortanca (median), alt ve üst çeyrekler (Q1 ve Q3) ile veri aralığı kutu ve kollar (whiskers) aracılığıyla gösterilir. Ayrıca, veri setindeki aykırı değerler kutu dışındaki noktalar olarak işaretlenir. Bu sayede, verinin simetrisi, yayılımı ve uç değerleri tek bakışta anlaşılabilir

## Kutu Grafiği Nerelerde Kullanılır?

- Aykırı değer tespiti.
- Veri temizleme öncesi kritik.

## Kutu Grafiği Nasıl Yorumlanır?

- Kutu uzunluğu (IQR) ne kadar büyükse veri o kadar dağınık.
- Simetrik kutu → veri simetrik.
- Medyan kutunun ortasındaysa simetri var, yoksa çarpıklık var.
- Noktalar → olası aykırı değerler.




Samsun Üniversitesi
Yazılım Mühendisliği Bölümü

# Scatter Plot Nedir?

Scatter plot (saçılım grafiği), iki değişken arasındaki ilişkiyi görselleştirmek için kullanılan bir grafik türüdür. Grafikteki her nokta, x ekseninde bir değişkenin, y ekseninde diğer değişkenin değerini temsil eder. Bu sayede değişkenler arasındaki korelasyon, genel eğilimler, kümelenmeler ve aykırı değerler kolayca fark edilebilir. Özellikle veri analizi ve istatistikte, ilişkilerin yönünü (pozitif, negatif) ve gücünü gözlemlemek için sıkça kullanılır.

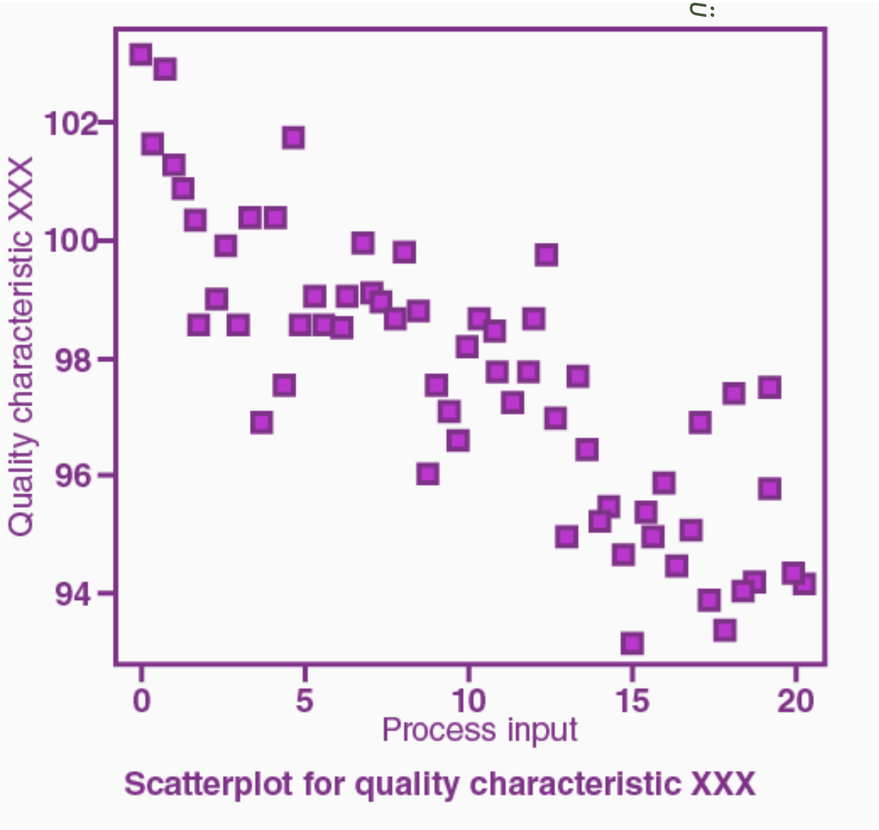
## Scatter Plot  Nerelerde Kullanılır?

- İki değişken arasındaki ilişkiyi görmek (pozitif/negatif korelasyon).
- Regresyon analizinde bağımlı-bağımsız değişken ilişkisi.
- Kümeleşme (clustering) eğilimi var mı anlamak.

## Scatter Plot Nasıl Yorumlanır?

- Eğilim yukarı → pozitif korelasyon.
- Eğilim aşağı → negatif korelasyon.
- Dağınık, düzen yok → korelasyon yok.
- Yoğun kümeler → alt gruplar olabilir.

*Samsun Üniversitesi*
*Yazılım Mühendisliği Bölümü*



# Pairplot Nedir?

Pairplot, bir veri setindeki tüm sayısal değişken çiftlerinin birbirleriyle olan ilişkilerini görselleştiren bir grafik türüdür. Her değişkenin kendi dağılımını gösteren histogram veya yoğunluk grafikleri, diyagonal üzerinde yer alırken, değişken çiftlerinin scatter plotları diyagonalın dışındaki hücrelerde bulunur. Böylece, veri içindeki değişkenler arasındaki korelasyonlar, dağılımlar ve olası kümelenmeler topluca ve kolayca analiz edilebilir. Özellikle keşifsel veri analizi (EDA) aşamasında çok kullanışlıdır.

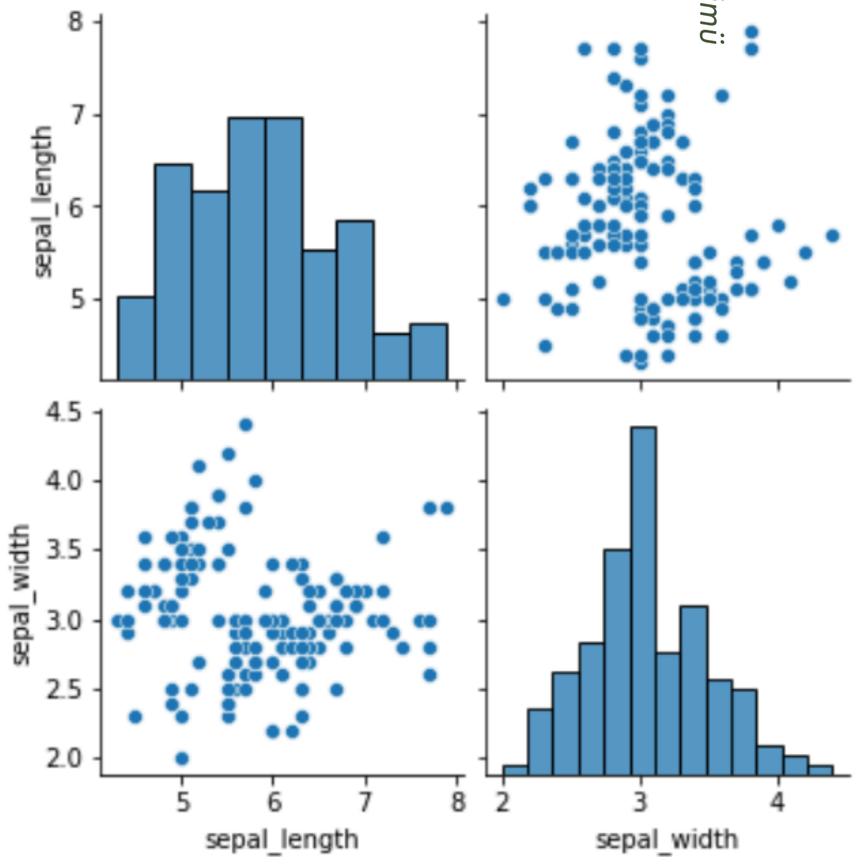
## Pairplot  Nerelerde Kullanılır?

- Keşifsel veri analizi (EDA) sırasında çok etkili.
- Hangi değişkenler ilişkili → modelde seçilecek özellikler için fikir verir.

## Pairplot Nasıl Yorumlanır?

- Diyagonal üzerindeki grafikler → her değişkenin dağılımı (histogram/kde).
- Diğer hücreler → iki değişkenin scatter plot'u.
- Belirgin doğrusal veya eğrisel ilişkiler → korelasyon var demek.

*Samsun Üniversitesi*
*Yazılım Mühendisliği Bölümü*



# EDA'nın Modelleme Öncesi Rolü

Keşifsel Veri Analizi (Exploratory Data Analysis – EDA), veri bilimi sürecinde modelleme aşamasına geçmeden önce gerçekleştirilen en kritik adımlardan biridir. Bu süreç, veriyi derinlemesine inceleyerek yapısını, özelliklerini, olası sorunlarını ve ilişkilerini anlamamızı sağlar. EDA'nın temel amacı, verinin doğasını anlamak, veri setinde eksik değerleri, aykırı gözlemleri ve tutarsızlıkları tespit etmek, ayrıca değişkenler arasındaki ilişkileri ortaya çıkarmaktır. Böylece, modelleme sürecinde karşılaşılabilecek problemlerin önüne geçilir ve daha doğru, güvenilir sonuçlar elde edilebilir.

EDA sayesinde, veri seti hakkında görsel ve istatistiksel özetler elde edilir. Grafikler, dağılımlar ve korelasyon analizleri aracılığıyla veri içindeki gizli paternler ve önemli değişkenler belirlenir. Bu süreç, hangi özelliklerin modelde daha etkili olabileceğini anlamamıza yardımcı olur. Dolayısıyla, modelin başarısını artırmak, gereksiz değişkenleri elemek ve doğru ön işleme adımlarını belirlemek için EDA, modelleme öncesi vazgeçilmez bir basamaktır.

Samsun Üniversitesi
Yazılım Mühendisliği Bölümü

# Keşifsel Veri Analizi (EDA) Aşamaları

## *1- Veri Toplama ve İlk İnceleme*

Modelleme sürecinin ilk adımı, elimizdeki veri setini anlamaktır. Bu aşamada, verinin yapısı, değişkenlerin türleri ve veri setinin büyüklüğü gibi temel bilgiler incelenir.

**Yöntemler:**

- Veri setini inceleme (head, info, describe komutları)
- Değişken tiplerini kontrol etme
- Örnekleme yaparak veri yapısını anlama

## *2-Eksik Veri Analizi ve İşleme*

Veri setlerinde genellikle bazı gözlemler eksik olabilir. Eksik veriler, modelin doğru öğrenmesini engeller ve hatalara yol açabilir. Bu yüzden, hangi değişkenlerde ne kadar eksik veri olduğu tespit edilir. Eksik veriler uygun yöntemlerle doldurulabilir

**Yöntemler:**

- Eksik veri analizi (null değer sayısı)
- Eksik değer doldurma (ortalama, medyan, ileri doldurma) veya silme
- Tutarsız veya hatalı verilerin düzeltilmesi

# Keşifsel Veri Analizi (EDA) Aşamaları

## *3- Aykırı Değerlerin Tespiti ve İşlenmesi*

Aykırı değerler, veri setindeki genel dağılımdan çok farklı olan uç değerlerdir. Bu değerler, modelin hatalı öğrenmesine neden olabilir. Aykırı değerler, görsel yöntemlerle (örneğin kutu grafiği) veya istatistiksel yöntemlerle (Z-skoru, IQR) tespit edilir. Tespit edilen aykırı değerler, modele etkisine göre çıkarılabilir, sınırlandırılabilir veya dönüştürülebilir.

**Yöntemler:**

- Kutu grafiği (boxplot) ile görsel inceleme
- İstatistiksel yöntemler (Z-skoru, IQR yöntemi)
- Aykırı değerleri kaldırma veya dönüştürme

## *4-Veri Dağılımının İncelenmesi*

Her değişkenin dağılımı incelenir; bu, verinin normal dağılıma uygun olup olmadığını anlamak için önemlidir. Histogram ve yoğunluk grafikleri gibi görseller, verinin simetrik mi, sağa ya da sola çarpık mı olduğunu gösterir. Ayrıca, değişkenlerin ortalama, medyan ve varyans gibi temel istatistikleri hesaplanır.

**Yöntemler:**

- Histogram ve yoğunluk grafikleri (density plots)
- Q-Q plot
- Betimsel istatistikler (ortalama, medyan, çeyrekler)

# Keşifsel Veri Analizi (EDA) Aşamaları

## *5- Değişkenler Arası İlişki Analizi*

Model performansı için değişkenler arasındaki ilişkiler iyi anlaşılmalıdır. Korelasyon matrisi ve ısı haritaları, sayısal değişkenler arasındaki doğrusal ilişkileri gösterir. Scatter plot ve pairplot gibi grafikler ise değişken çiftlerinin dağılımını görselleştirerek, ilişkilerin yapısını ve olası kümelenmeleri ortaya koyar.

**Yöntemler:**

- Scatter plot (saçılım grafiği)
- Korelasyon matrisi ve ısı haritası (heatmap)
- Pairplot

## *6- Özellik Mühendisliği ve Veri Dönüşümleri*

Veriler, modele uygun hale getirilmelidir. Kategorik veriler sayısal formata dönüştürülürken, sayısal veriler genellikle ölçeklendirilir. Ayrıca, sürekli veriler kesikli aralıklara bölünebilir veya bazı dönüşümler uygulanabilir. Bu işlemler, modelin daha iyi öğrenmesini ve performansını artırmayı sağlar.

**Yöntemler:**

- Kategorik değişkenlerin kodlanması (one-hot, label encoding)
- Ölçeklendirme (normalizasyon, standardizasyon)
- Kesikli hale getirme (discretization)

# Keşifsel Veri Analizi (EDA) Aşamaları

## *7- Veri Görselleştirme*

Veri görselleştirme, veriyi daha iyi anlamak ve içindeki örüntüleri, trendleri ya da anormallikleri fark etmek için kullanılır. Histogram, kutu grafiği, saçılım grafiği ve pairplot gibi grafiklerle veri farklı açılardan incelenir. Bu görseller, hem veri bilimciler hem de karar vericiler için değerli bilgiler sunar.

**Yöntemler:**

- Bar chart, pie chart
- Box plot, histogram
- Scatter plot, pairplot

## *8- Sonuçların Yorumlanması ve Raporlama*

EDA sürecinde elde edilen tüm bulgular değerlendirilir ve modelleme için strateji belirlenir. Veri temizliği, dönüşümler ve önemli değişkenlerin seçimi bu aşamada netleşir. Elde edilen bilgiler, yazılı ya da görsel raporlarla paylaşılır ve sonraki aşamalara sağlam bir temel oluşturur.

**Yöntemler:**

- Önemli değişkenlerin seçimi
- Veri temizliği ve dönüşümü planı
- Model seçimi önerileri

# UYGULAMA

## Temel İstatistikler

Kod Parçası

```python
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# 1️⃣ Veri Setini Yükleme
iris = sns.load_dataset("iris")

# Temel istatistiksel özet
print("\n🔹 Temel İstatistikler:")
print(iris.describe())
```

Kod Çıktısı

```
🔹 Temel İstatistikler:
       sepal_length  sepal_width  petal_length  petal_width
count    150.000000   150.000000    150.000000   150.000000
mean       5.843333     3.057333      3.758000     1.199333
std        0.828066     0.435866      1.765298     0.762238
min        4.300000     2.000000      1.000000     0.100000
25%        5.100000     2.800000      1.600000     0.300000
50%        5.800000     3.000000      4.350000     1.300000
75%        6.400000     3.300000      5.100000     1.800000
max        7.900000     4.400000      6.900000     2.500000
```

# UYGULAMA

## CountPlot

### Kod Parçası

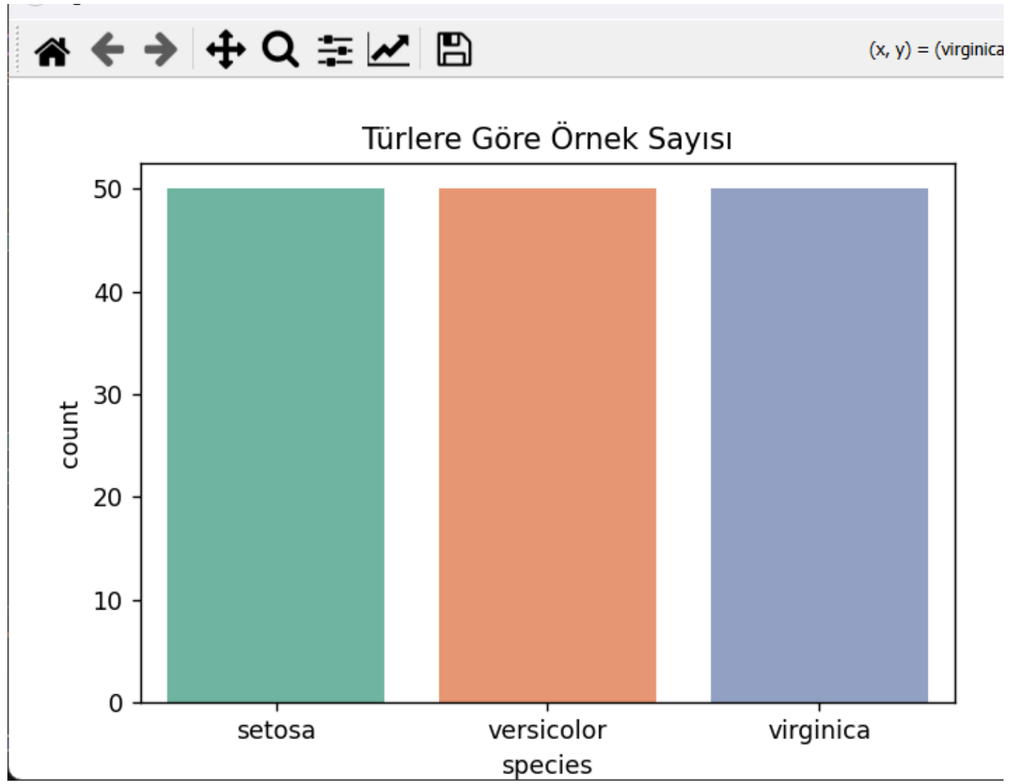
```python
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# 1️⃣ Veri Setini Yükleme
iris = sns.load_dataset("iris")

# 2️⃣ Türlere göre dağılım (countplot)
plt.figure(figsize=(6,4))
sns.countplot(x="species", data=iris, palette="Set2")
plt.title("Türlere Göre Örnek Sayısı")
plt.show()

```

### Kod Çıktısı



# UYGULAMA

## Histogram

Kod Parçası

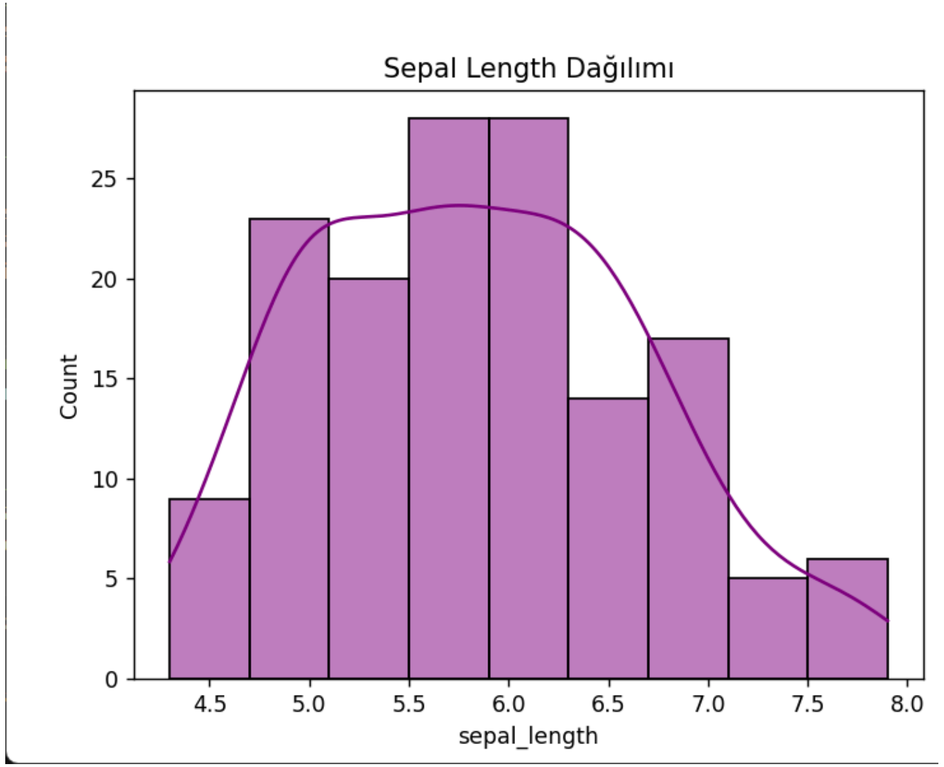
```python
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# 1️⃣ Veri Setini Yükleme
iris = sns.load_dataset("iris")

#  Sayısal değişkenlerin dağılımı (histogram)
sns.histplot(iris["sepal_length"], kde=True, color="purple")
plt.title("Sepal Length Dağılımı")
plt.show()

```

Kod Çıktısı



# UYGULAMA

## Sepal Ölçüleri

### Kod Parçası

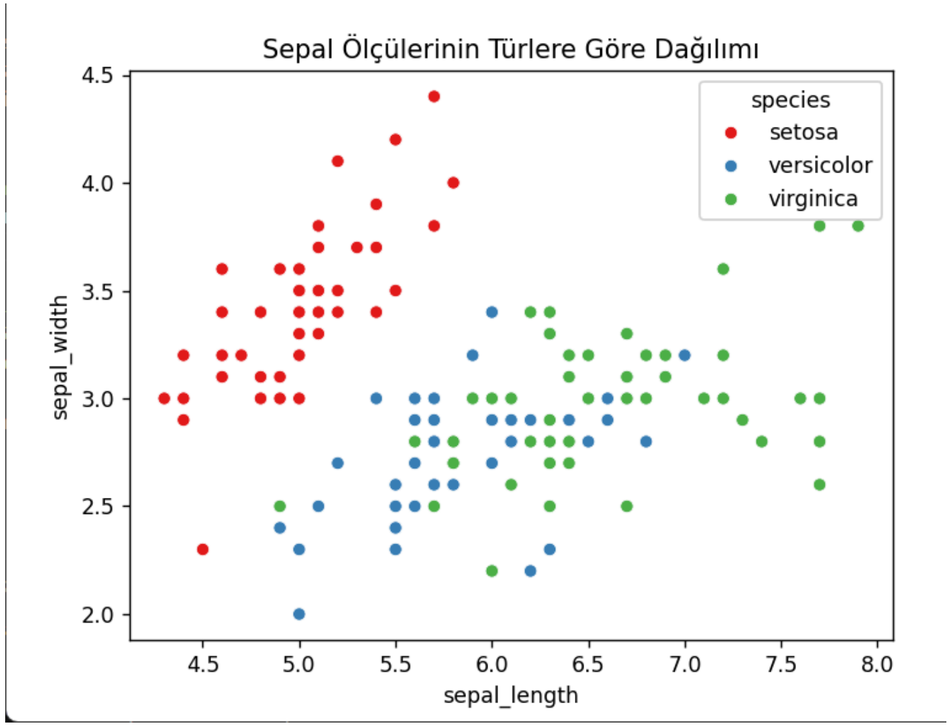
```python
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# 1️⃣ Veri Setini Yükleme
iris = sns.load_dataset("iris")

#  Türlere göre sepal_length ve sepal_width karşılaştırması (scatter plot)
sns.scatterplot(x="sepal_length", y="sepal_width", hue="species", data=iris, palette="Set1")
plt.title("Sepal Ölçülerinin Türlere Göre Dağılımı")
plt.show()

```

### Kod Çıktısı



# UYGULAMA

## Pair Plot

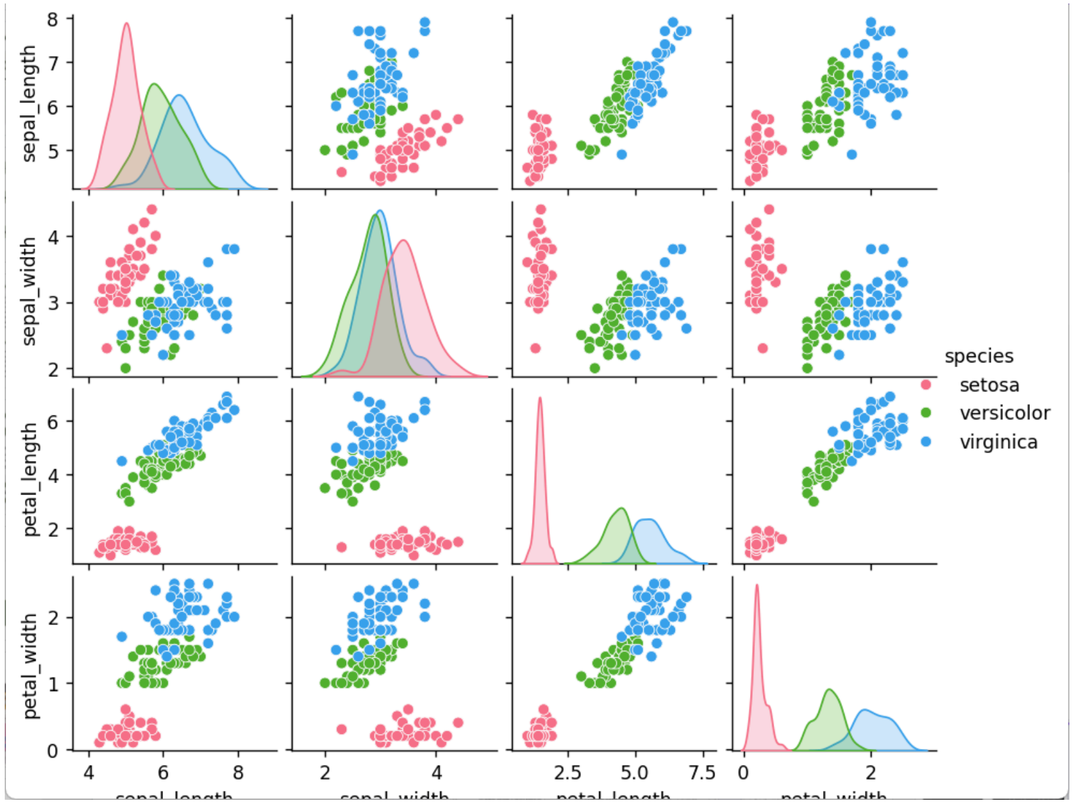
### Kod Parçası

```python
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# 1️⃣ Veri Setini Yükleme
iris = sns.load_dataset("iris")

# Tüm değişkenler arası ilişkiler (pairplot)
sns.pairplot(iris, hue="species", palette="husl")
plt.show()
```

### Kod Çıktısı



# UYGULAMA

## Pair Plot

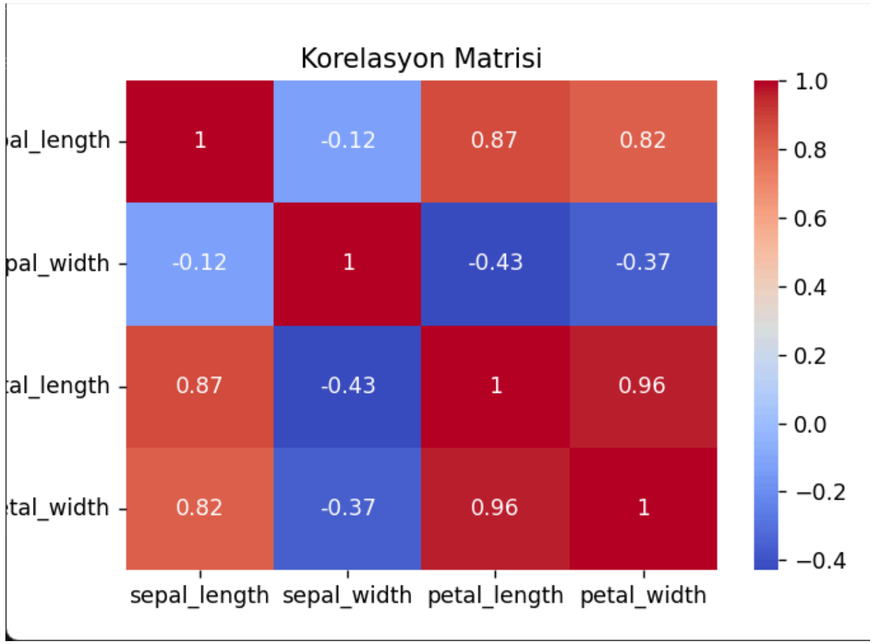
### Kod Parçası

```python
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# 1️⃣ Veri Setini Yükleme
iris = sns.load_dataset("iris")

#  Korelasyon matrisi
plt.figure(figsize=(6,4))
sns.heatmap(iris.drop("species", axis=1).corr(), annot=True, cmap="coolwarm")
plt.title("Korelasyon Matrisi")
plt.show()
```

### Kod Çıktısı



amsun Üniversitesi
Yazılım Mühendisliği Bölümü

# VERİ MADENCİLİĞİ (DATA MİNİNG)

Dr. Öğr. Üyesi Alper Talha Karadeniz

2025-2-26