2.4 Veriyi Anlamak

Veriyi anlamak, veri ile çalışan bütün disiplinler için en başta gelmektedir. Veri araştırması, verilerin istatistiksel ve görselleştirme teknikleriyle tanımlanması ile ilgilidir. Veri araştırması için herhangi bir kısayol yoktur. Makine Öğrenmesi ile bir süre uğraştıktan sonra, modelin doğruluğunu geliştirme konusunda mücadele ettiğinizin farkına varacaksınız. Böyle bir durumda veri araştırması teknikleri aklınıza gelecektir.

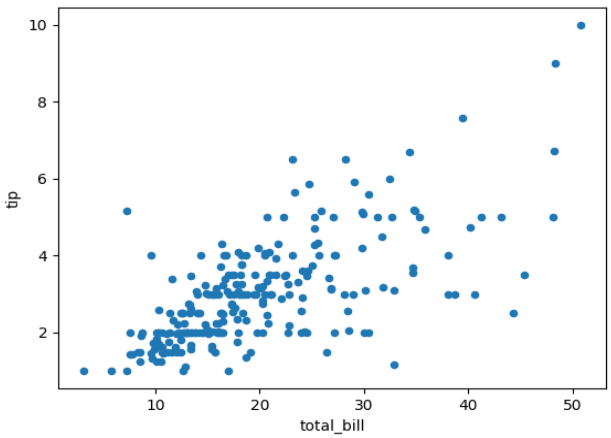
Girdi verilerinizin kalitesinin çıktılarınızın kalitesine karar verdiğini unutmayın. Veri araştırması, temizleme ve hazırlama toplam proje süresinin %70'ine kadar çıkabilir. Makine öğrenmesi modelini oluşturmak için verilerinizi anlama, temizleme ve hazırlama adımlarından bazılarını şöyle sıralayabiliriz.

1. Değişken Tanımlama
2. Tek Değişkenli Analiz
3. İki Değişkenli Analiz
4. Eksik Değer Düzenleme
5. Aykırı Veri Düzenleme
6. Değişken Dönüşümü
7. Değişken Oluşturma

2.4.1 Veriyi Görselleştirmek

Şimdi ödenen hesap ve verilen bahşiş arasındaki ilişkiyi daha rahat görebilmek için dağılım grafiğina bakalım.

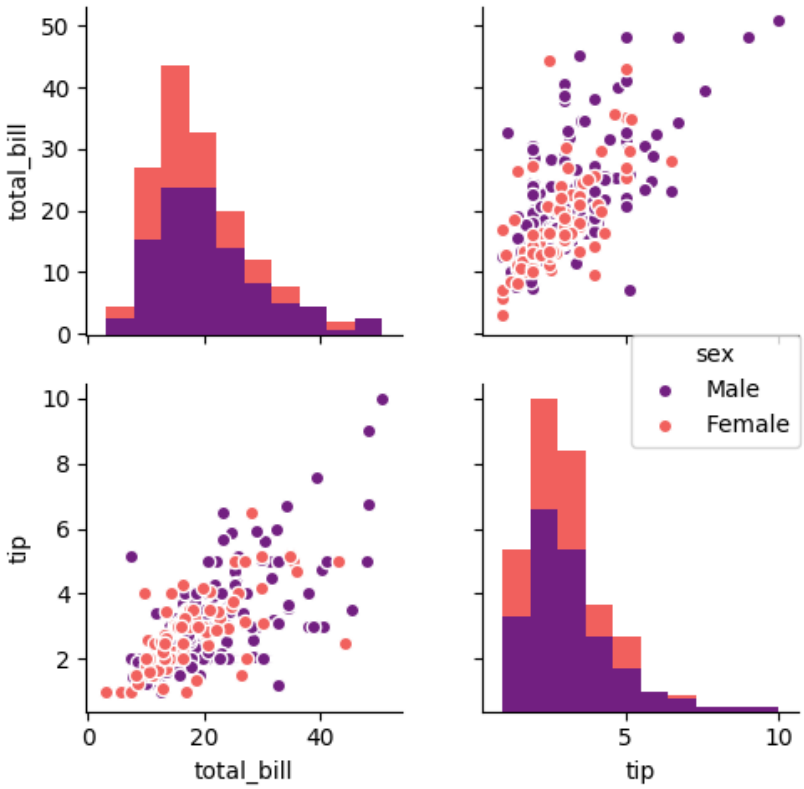
```
veri_seti.plot(x='total_bill', y='tip', kind='scatter')
plt.show()
```



Dağılım grafiğine baktığımızda ödenen hesaba göre verilen bahşişler nasıl değişir sorusuna bir varsayımdan daha çok gerçek bir cevap verebiliyoruz, ödenen hesaba göre verilen bahşişler genelde artmaktadır. Bu durumda aslında hesap ile bahşiş miktarı arasında bir ilişki olduğunu açıkça görebiliyoruz.

Şimdi biraz daha detaylı görsellere bakabiliriz.

```
sns.pairplot(veri_seti, hue='sex', palette='magma')
plt.show()
```



Yukarıdaki grafikleri incelediğimizde, hesap ve bahşiş arasındaki ilişkiyi biraz daha detaylandırarak, kadın ve erkekler ile ilgili ilişkilerini görebiliyoruz. Veriyi görselleştirmek veriyi anlamamızı kolaylaştırır.

2.4.2 Veriler Arasındaki İlişkiyi İncelemek

Verilerin aralarındaki ilişkiyi daha iyi tanımlayabilmek için her bir çift özellik arasındaki standart korelasyon katsayısını hesaplayalım.

```
print veri_seti.corr()
```

```
            total_bill       tip      size
total_bill    1.000000  0.675734  0.598315
tip           0.675734  1.000000  0.489299
size          0.598315  0.489299  1.000000
```

Korelasyon Katsayısı

Korelasyon katsayısı -1 ile 1 arasında değişir.

- 1'e yakın olduğunda güçlü bir pozitif korelasyon olduğu anlamına gelir.
- Katsayısı -1'e yakın olduğunda, güçlü bir negatif korelasyon olduğu anlamına gelir.
- Son olarak, sıfıra yakın katsayılar, doğrusal bir korelasyon bulunmadığı anlamına gelir.

Hesap ve bahşiş arasındaki katsayıya baktığımızda 0.68 değerinde olduğunu görüyoruz. Yani aralarında güçlü bir pozitif korelasyon olduğunu söyleyebiliriz.