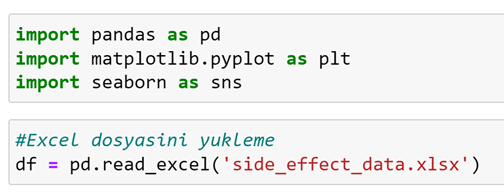
**Data Science Intern Case Study**

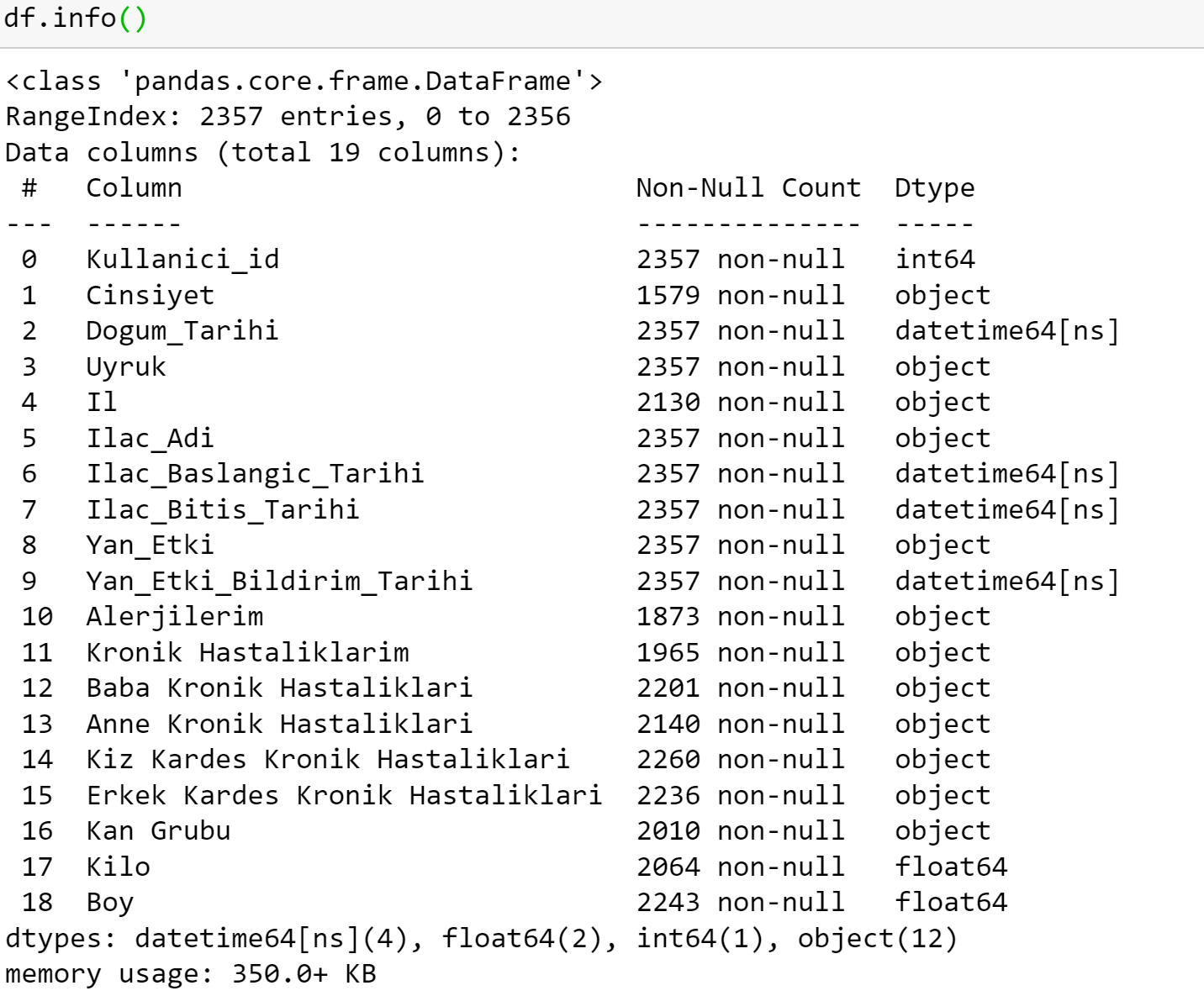
**Şule Aktaş**

**0suleaktas@gmail.com**

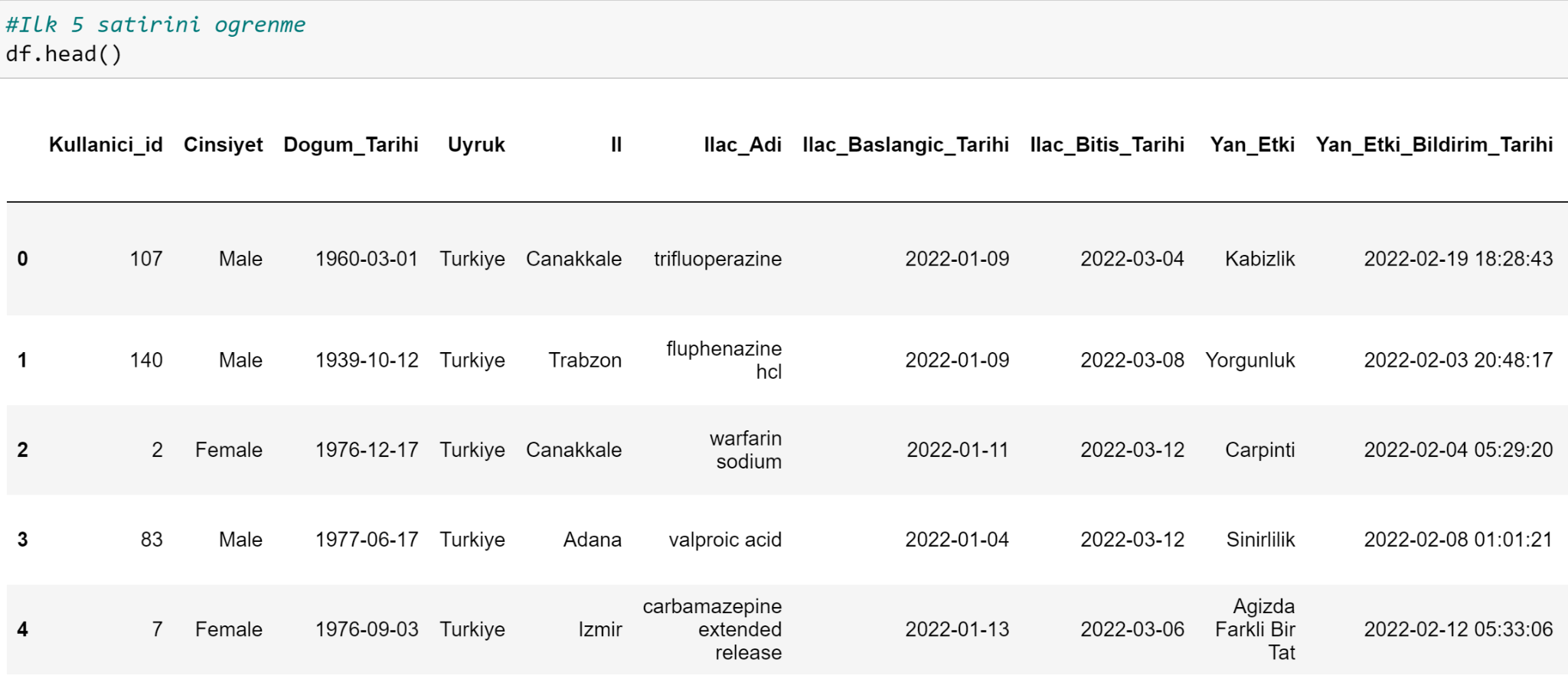
1. **Keşifsel Veri Analizi (Exploratory Data Analysis - EDA)**
   1. Python kullanılarak;
      1. Gerekli kütüphaneler ve Excel dosyası yüklendi.

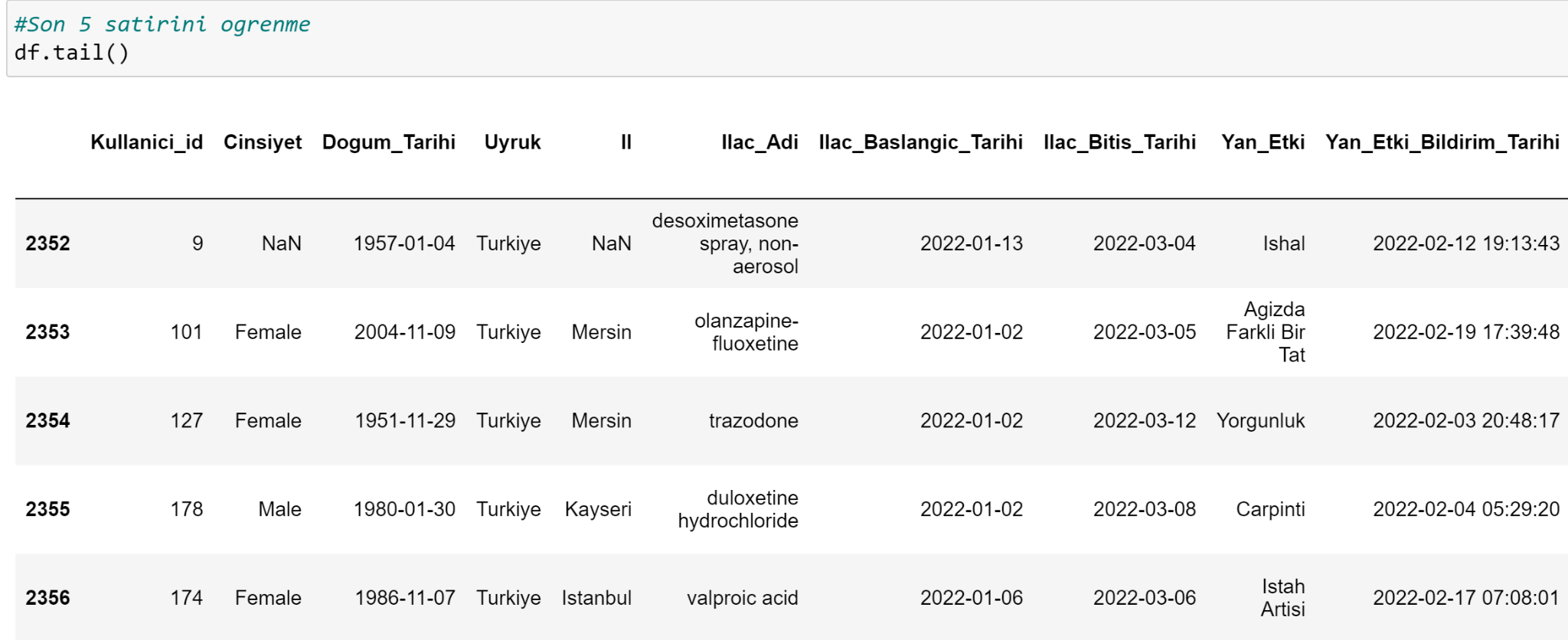


* + 1. Veri kümesi kapsamlı bir şekilde incelendi.
       1. Veri Seti hakkında genel bilgi almak için **info()** fonksiyonu kullanılarak aşağıdaki kod satırı yazıldı.

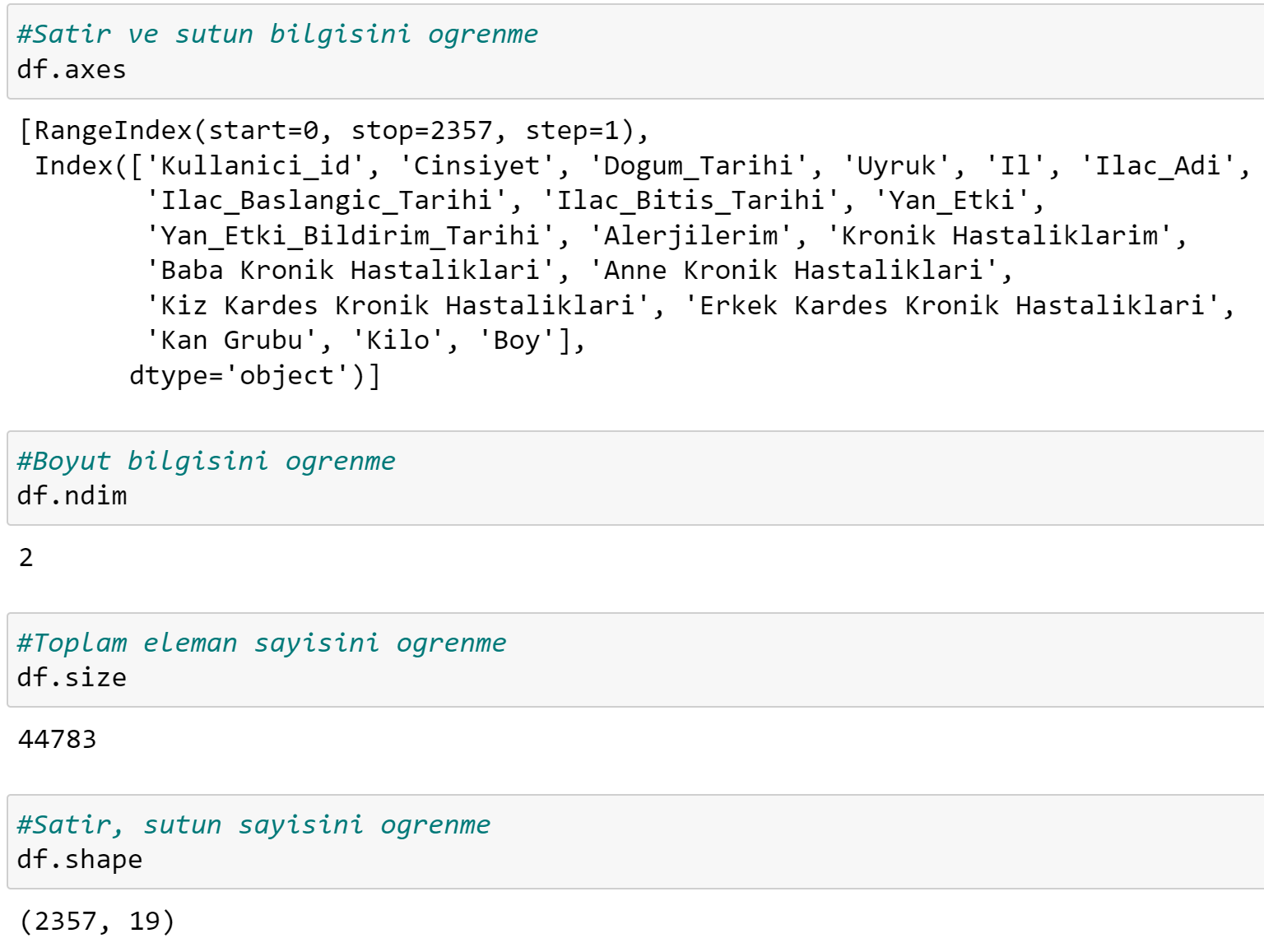


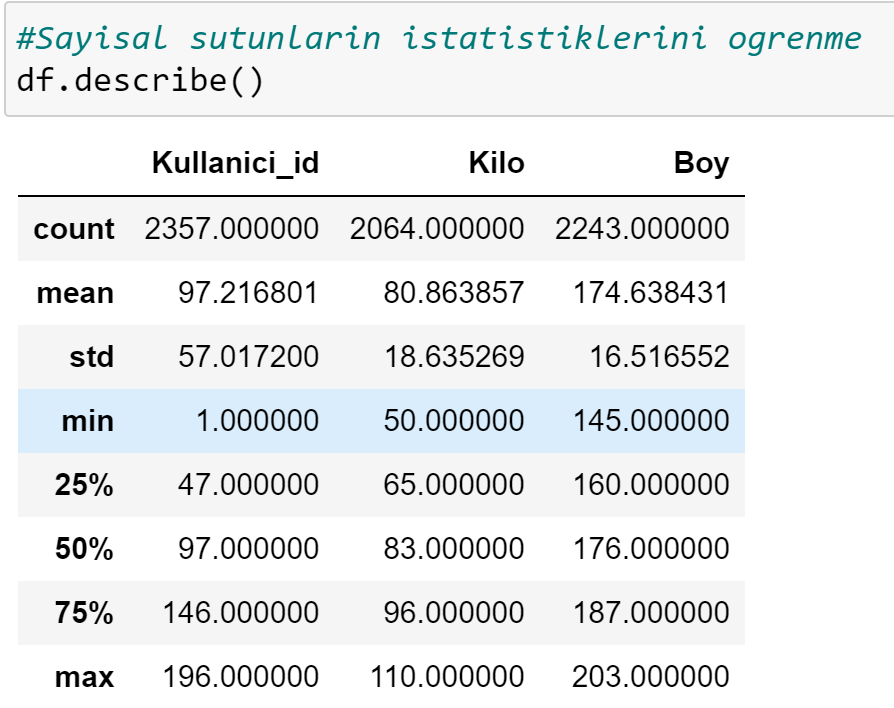
* + - 1. Veri Setindeki ilk 5 satırı görmek için **head()** fonksiyonu ve son 5 satırı görmek için **tail()** fonksiyonu kullanılarak aşağıdaki kod satırları yazıldı.



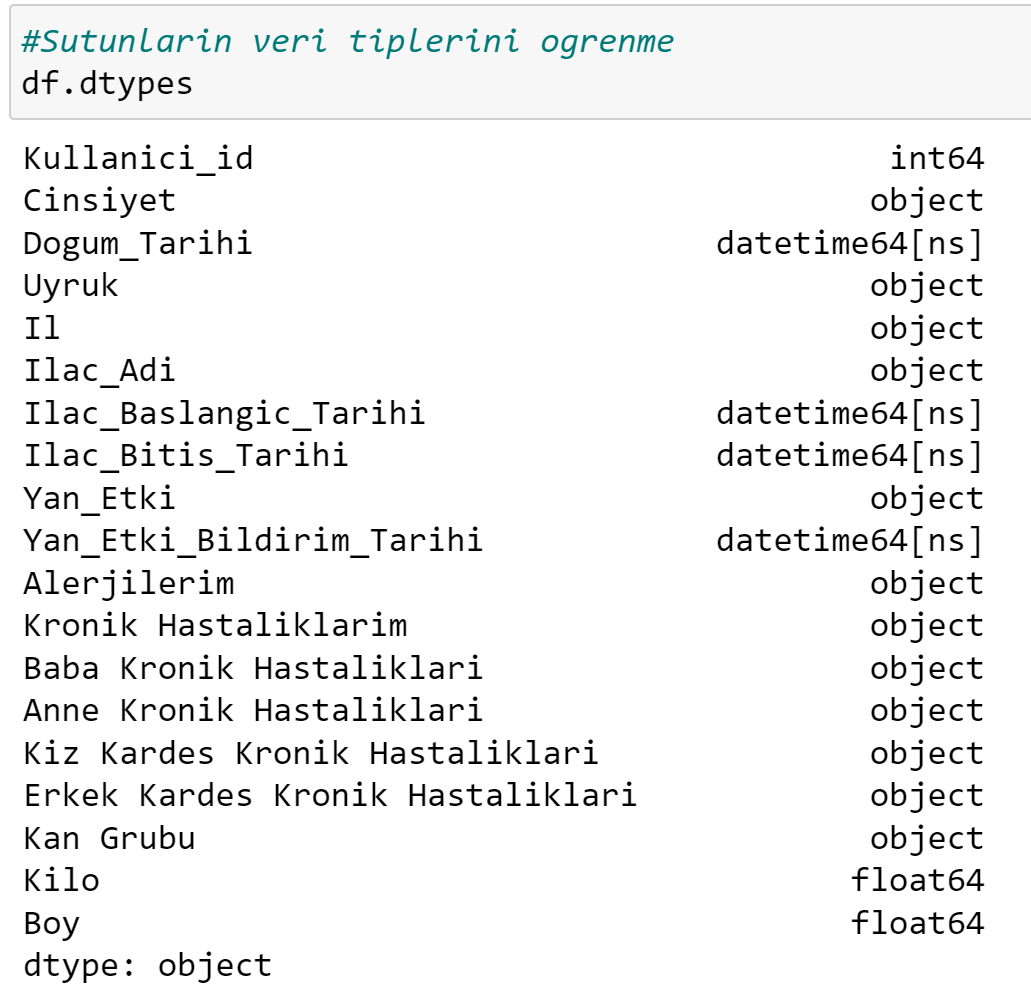


* + - 1. **axes(), ndim(), size() , shape() ve describe()** fonksiyonları kullanılarak veri setindeki satır ve sütun bilgisine, boyut bilgisine, toplam eleman sayısı bilgisine, satır ve sütun sayısı bilgisine, sayısal sütunların istatiksel bilgilerine aşağıdaki kod satırları yazılarak ulaşıldı.

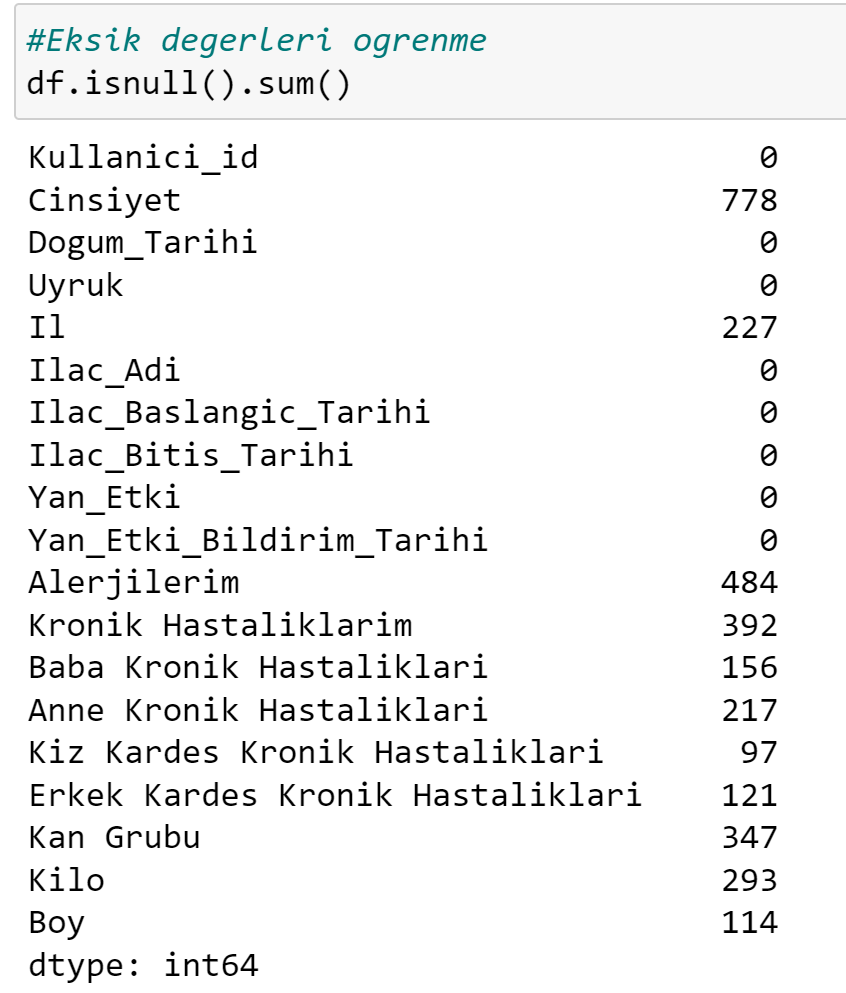


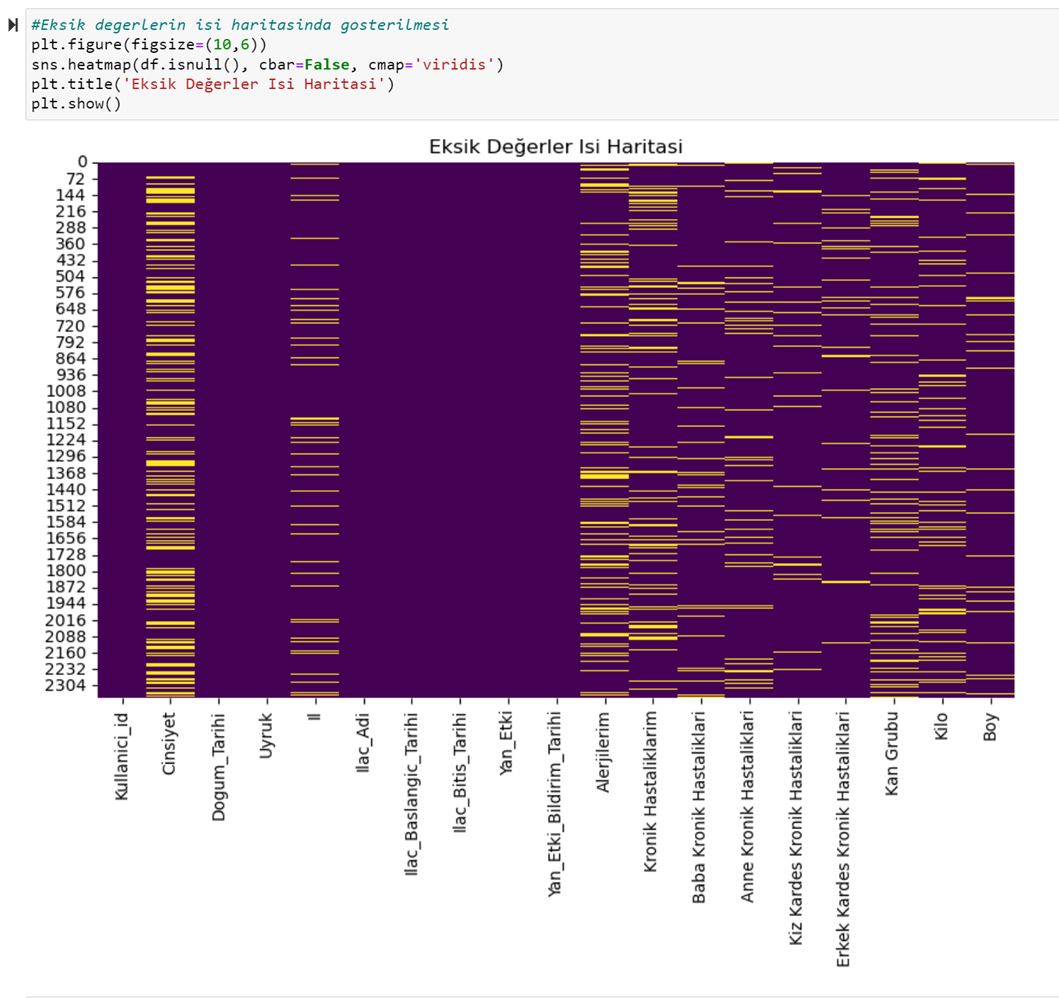


* + 1. Değişken türleri incelendi.
       1. Kategorilerin değişken tipleri öğrenmek için **dtypes()** fonksiyonu kullanılarak aşağıdaki kod satırı yazıldı.



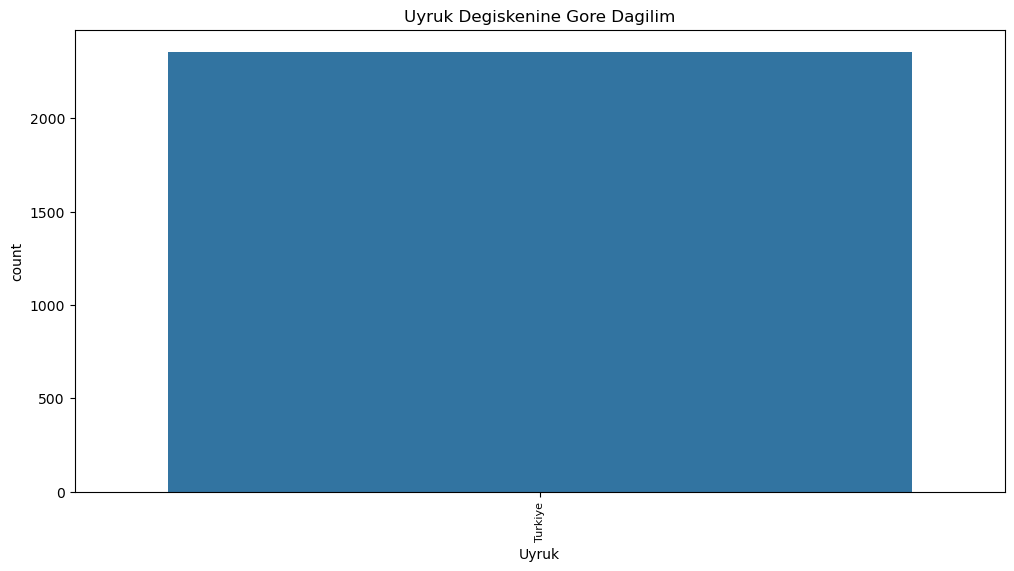
* + 1. Eksik veriler tespit edildi.
       1. Kategorilerdeki eksik verileri tespit etmek için **isnull().sum()** fonksiyonu kullanılarak aşağıdaki kod satırı yazıldı.

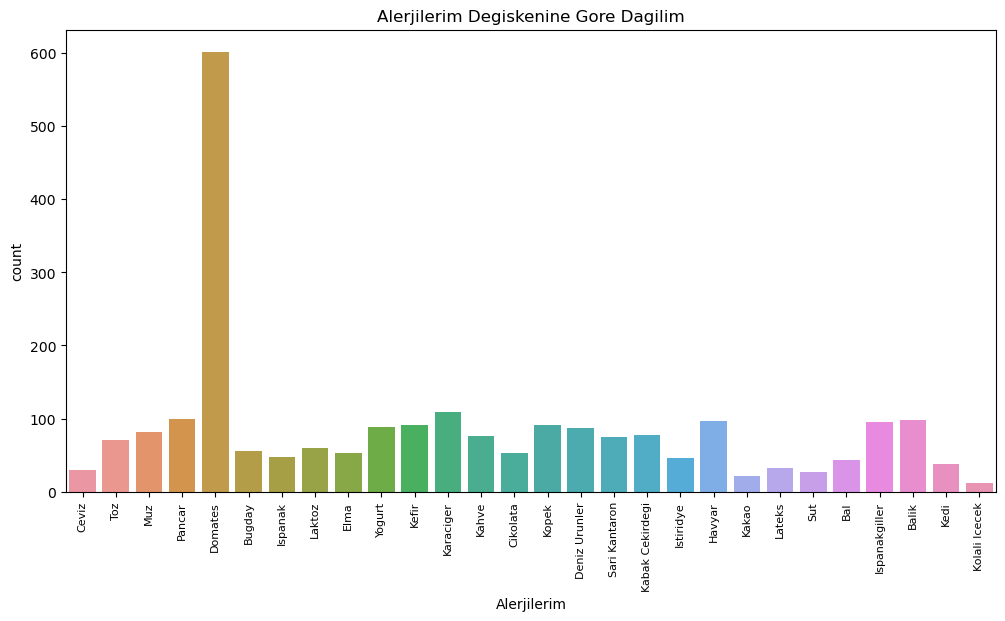


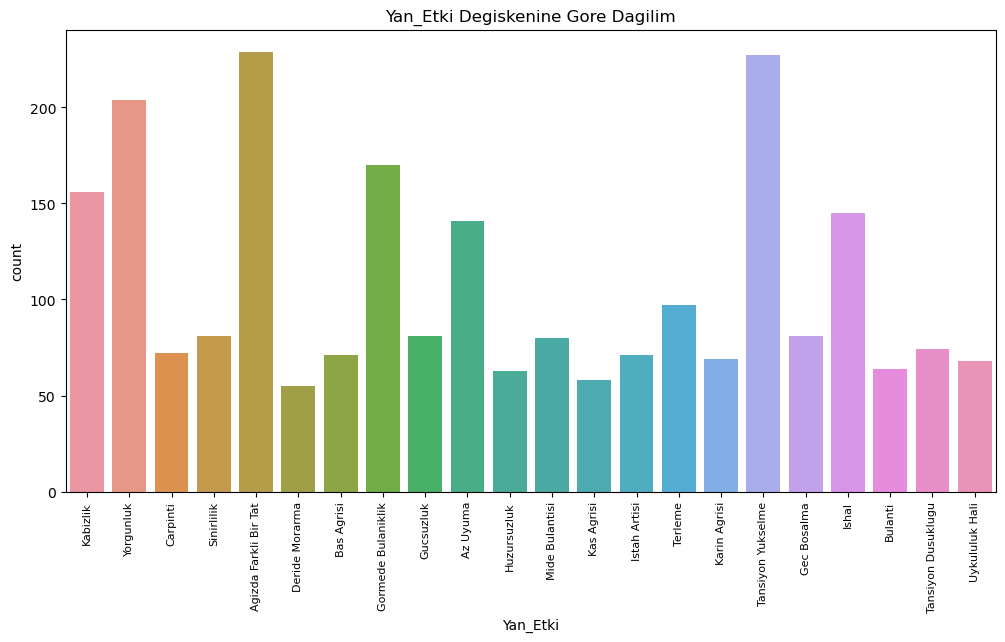
* + 1. Veri görselleştirme teknikleri kullanıldı. (Matplotlib, Seaborn)
       1. **heatmap()** fonksiyonu kullanılarak Isı Haritası oluşturuldu. Grafik için bir figür (çerçeve) yapıldı ve **figsize=(10,6)** argümanı ile bu figürün genişlik ve yükseklik boyutları ayarlandı.
       2. Bu, görselin 10 birim genişliğinde ve 6 birim yüksekliğinde olmasını sağladı. **df.isnull()** kısmı, DataFrame'deki (df) eksik değerleri (NaN ya da boş hücreler) tespit edildi. **isnull()** fonksiyonu ile veri setindeki her hücrede eksik değer olup olmadığı kontrol edildi ve True/False (1/0) döndürdü. **viridis**, Seaborn ve Matplotlib'deki popüler bir renk haritasıdır ve sarıdan mora doğru bir renk geçişi sunar. **title()** fonksiyonu ile Haritaya başlık belirlendi ve **show()** fonksiyonu ile grafik ekrana bastırıldı.
       3. **df.select\_dtypes(include=['object'])** Bu komut ile veri setindeki yalnızca kategorik değişkenleri seçildi. Kategorik değişkenler, genellikle object veri tipinde olan ve sayısal olmayan sütunlar (örneğin metin ya da kategorilere ayrılmış veri) içerir. Seaborn kütüphanesinden gelen **countplot()** fonksiyonu ile çubuk grafik oluşturuldu. **plt.xticks()** fonksiyonu ilex eksenindeki kategorik değerlerin (etiketlerin) yönünü ayarlandı. **rotation=90 ile** etiketler 90 derece döndürülerek dikey hale getirildi. **fontsize=8 ile** etiketlerin yazı boyutu 8 birim olarak ayarlandı. Örnek grafikler aşağıdaki gibidir.

A screenshot of a computer code

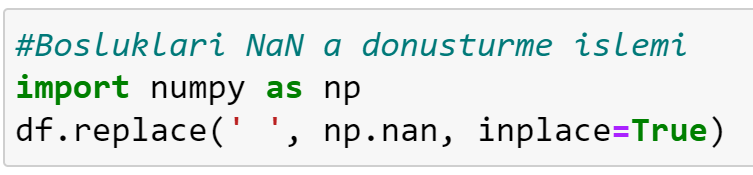
Description automatically generated



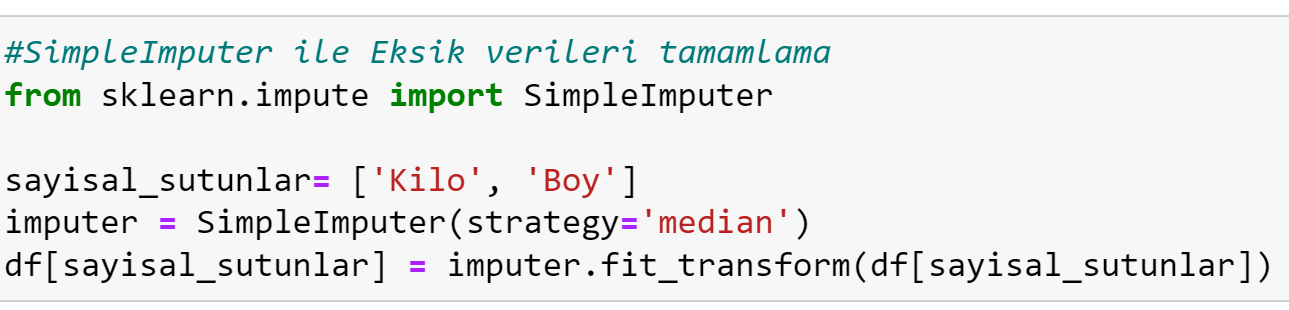




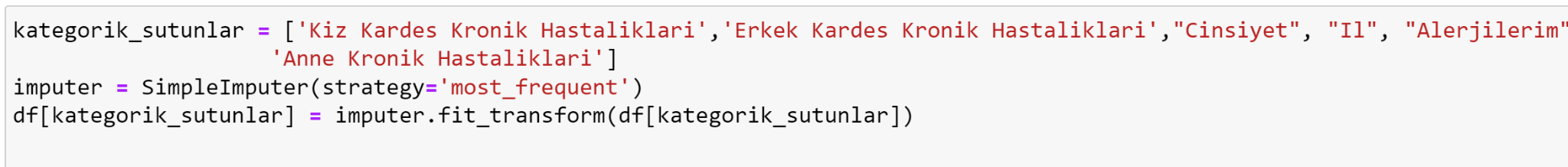
1. **Veri Ön İşleme (Data Pre-Processing)**
   1. Boşluk karakterlerini NaN'a dönüştürme işlemi yapıldı.
      1. **replace()** fonksiyonu kullanılarak **' '** (boşluk karakteri) ile eşleşen değerler **np.nan** ile değiştirilir



* 1. Eksik verileri tamamlama işlemi yapıldı.
     1. Sayısal sütunlardaki eksik veriler için **SimpleImputer** sınıfından bir nesne oluşturuldu ve stratejisini **'median'** olarak belirlendi. Bu strateji, eksik olan (NaN) hücreleri ortanca değerler ile belirledi. **sayisal\_sutunlar** değişkeni ile sayısal sütunlar seçildi. **fit()** fonksiyonu her sütun için en sık görülen değeri bulur ve model olarak saklar. **transform()** fonksiyonu bu bulduğu en sık görülen değerleri kullanarak, her sütundaki eksik hücreleri doldurur.



* + 1. Kategorik sütunlardaki eksik veriler için **SimpleImputer** sınıfından bir nesne oluşturuldu ve stratejisini **'most\_frequent'** olarak belirlendi. Bu strateji, eksik olan (NaN) hücreleri o sütunda en sık görülen değerle doldurdu. **kategorik\_sutunlar** değişkeni ile kategorik sütunlar seçildi.



* 1. Eksik veriler tespit edildi.
     1. Kategorilerdeki eksik verileri doldurduktan sonra test etmek için **isnull().sum()** fonksiyonu kullanılarak aşağıdaki kod satırı yazıldı.

