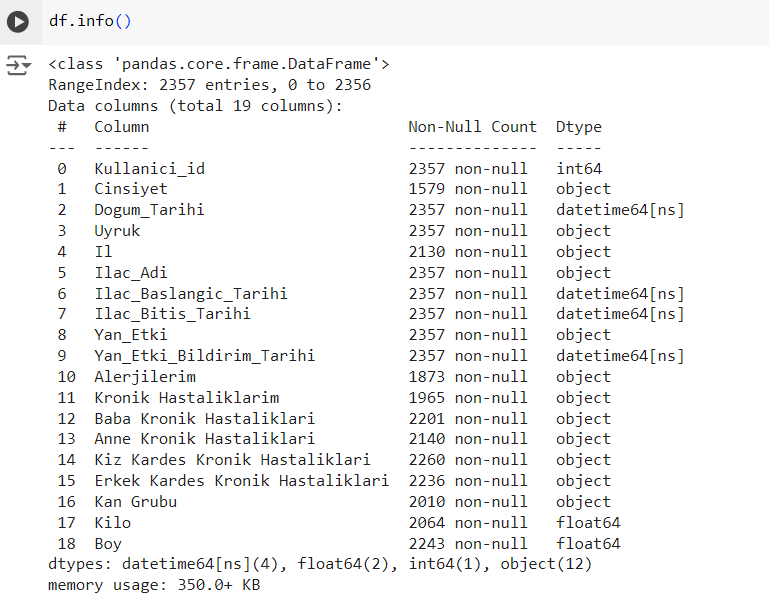
Semanur Özyılmaz

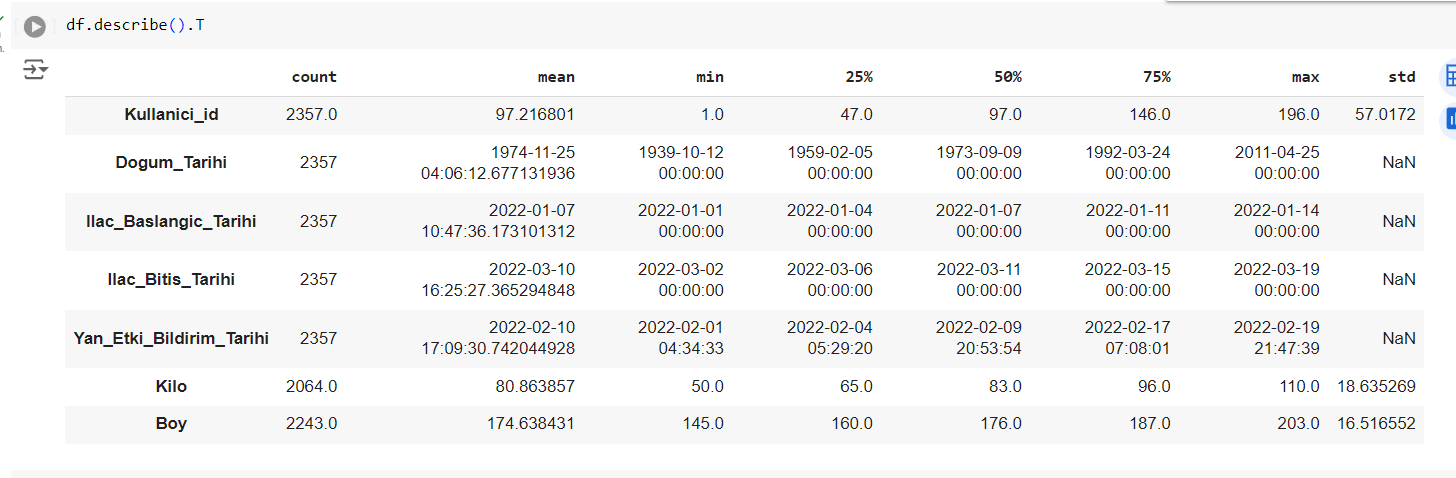
semanurozyilmaz9@gmail.com

İlaç Yan Etkileri Veri Seti Raporu

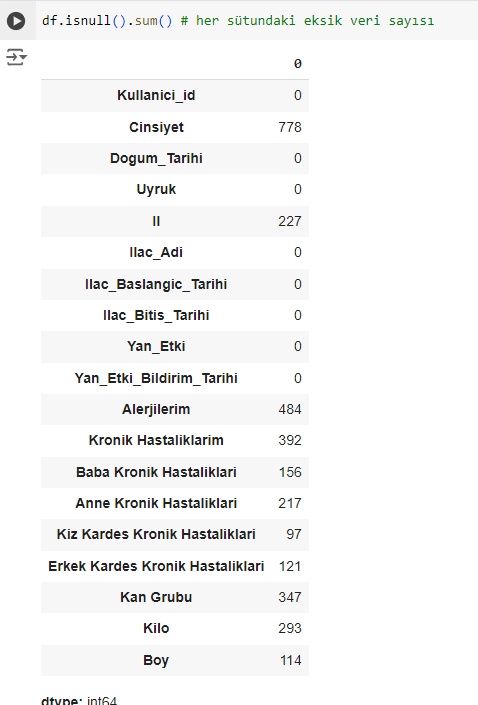
Öncelikle veri setine dair bilgiler incelendi.



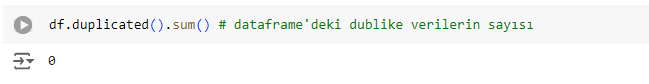
2357 adet veri var. Sütunların bazılarında eksik veriler bulunuyor. Tarih sütunları datetime türünde, kategorik sütunlar object türünde, sayısal sütunlar ise int ve float türlerinde.



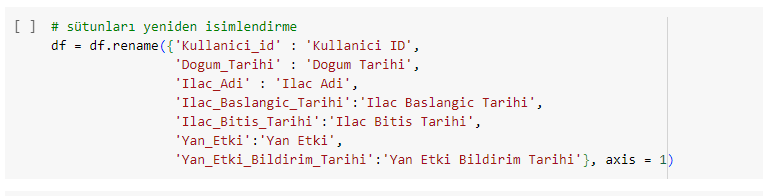
Burada kilo ve boy sütunları için anlamlı ortalama, minimum, maksimum, çeyrek ve yarı değerler, standart sapma bilgisi elde edilebilir. Ortalama kilo 80, boy 174.



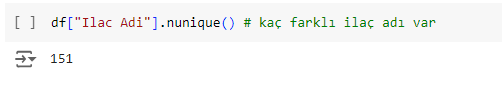
Hangi sütunda kaç tane eksik veri olduğu bilgisine bakıldı. Bu sütunlarda eksik veriler ile ilgili çalışma yapılmalı.



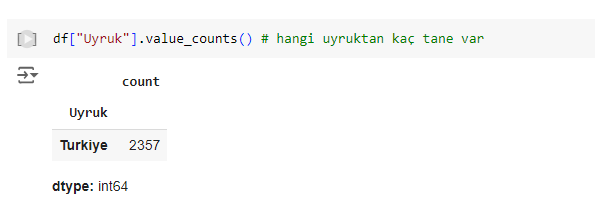
Veri setinde tekrar eden satırlar var mı diye kontrol edildi. Tekrar eden herhangi bir satır bulunamadı.



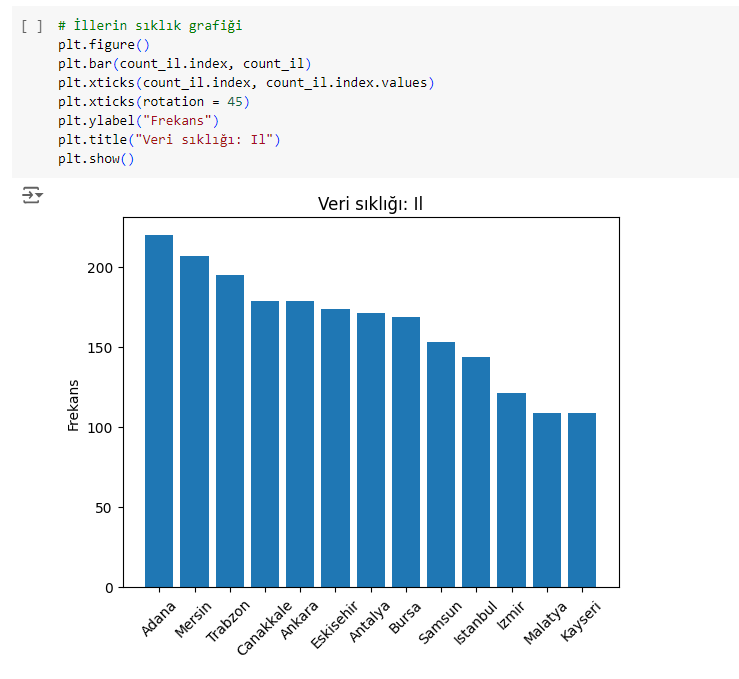
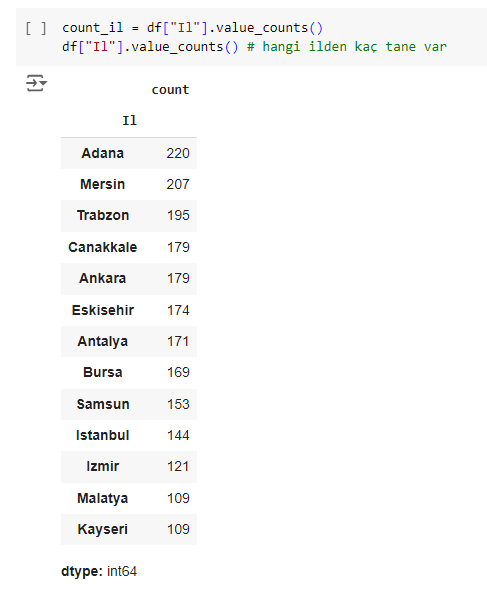
Bazı sütunlar yeniden isimlendirilerek diğer sütunlarla aynı isim formatında yapıldı.



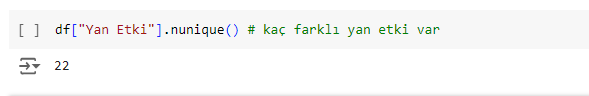
Veri setinde 151 farklı ilaç adı bulundu.



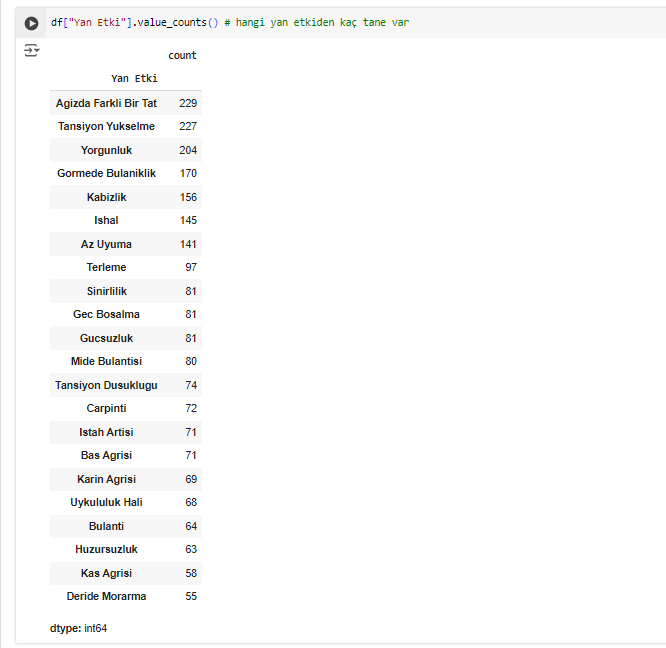
Veri setinde yan etki bildiren herkesin uyruğunun Türkiye olduğu belirlendi.



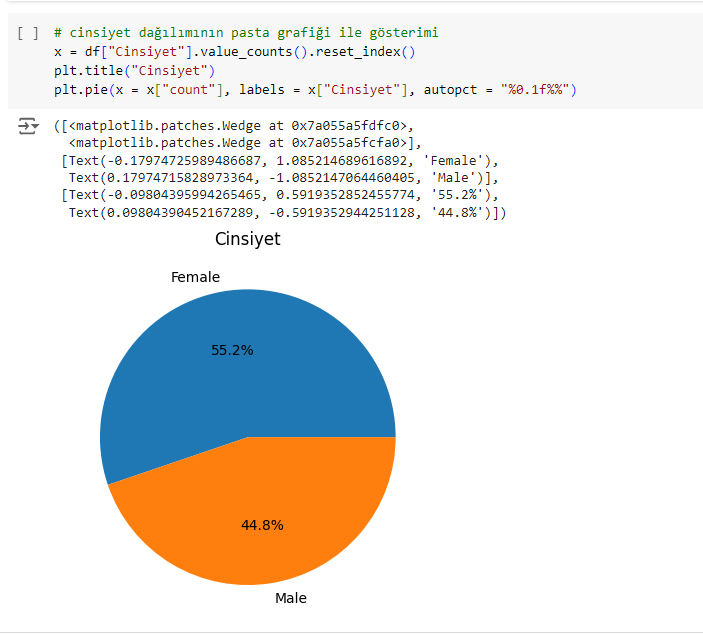
Veri setinde yan etki bildirenlerin yaşadığı 13 farklı il olduğu tespit edildi ve her ilde kaç kişinin yaşadığı belirlendi ve bu veri sütun grafiğine aktarıldı. Veri setinde en çok Adana’da yaşayan bireyler, en az ise Malatya ve Kayseri’de yaşayan bireyler bulunuyor.



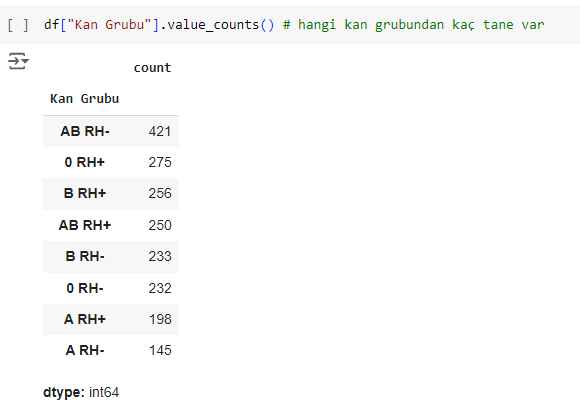
Veri setinde 22 farklı yan etki bulundu.



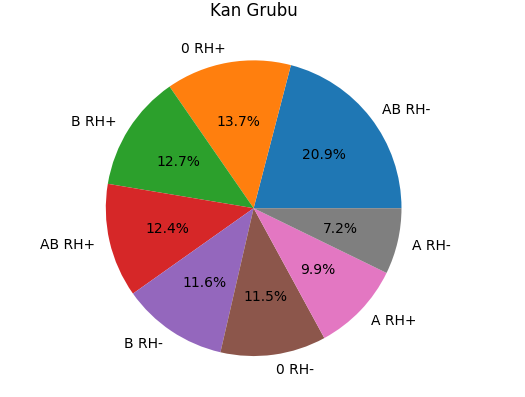
Her bir yan etkiden kaç tane olduğu belirlendi. Bireylerde en çok görülen yan etki ağızda farklı bir tat olması, en az görülen yan etki ise deride morarma.



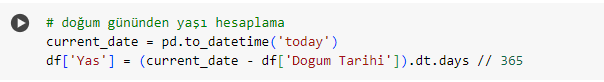
Veri setindeki cinsiyet dağılımı pasta grafiği ile gösterildi. Dağılım anormal değil. Yüzdeler birbirine yakın.



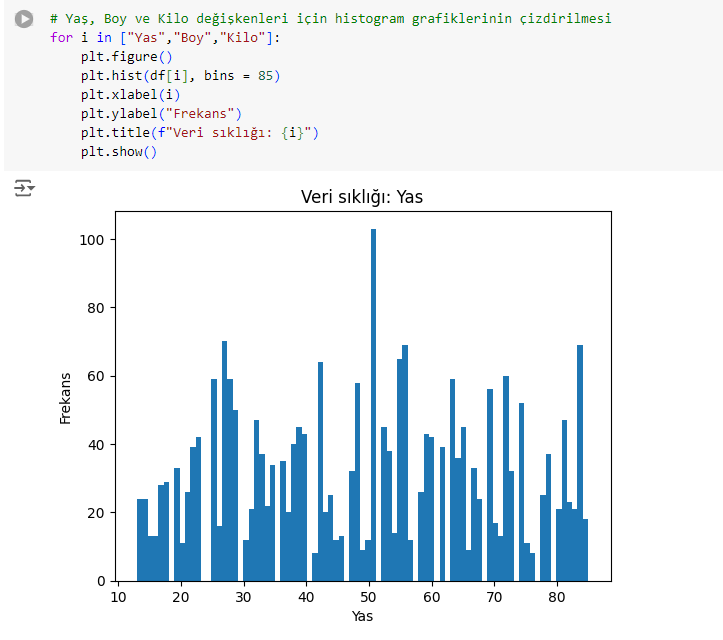
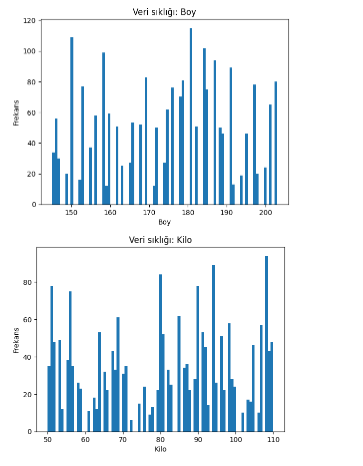




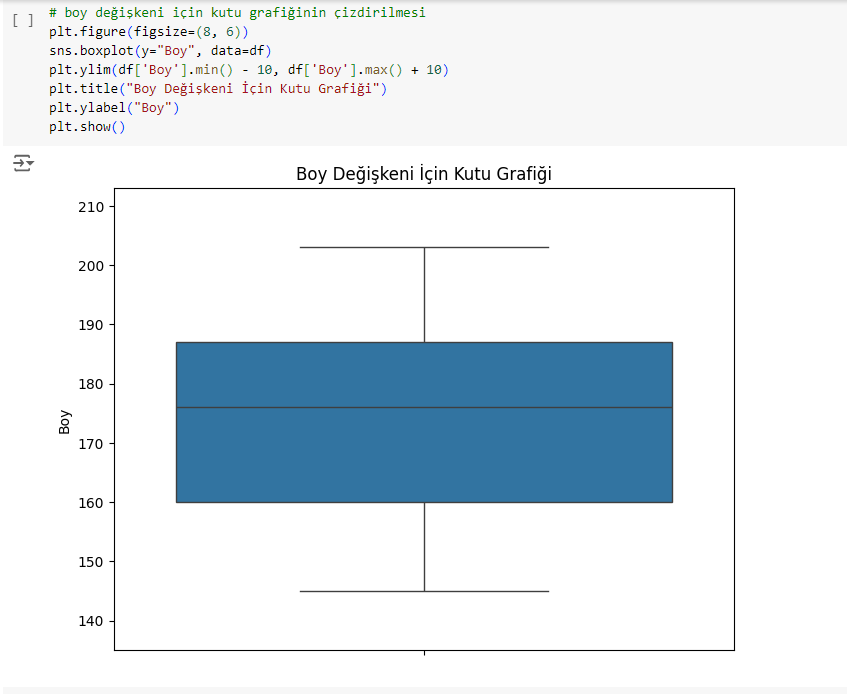
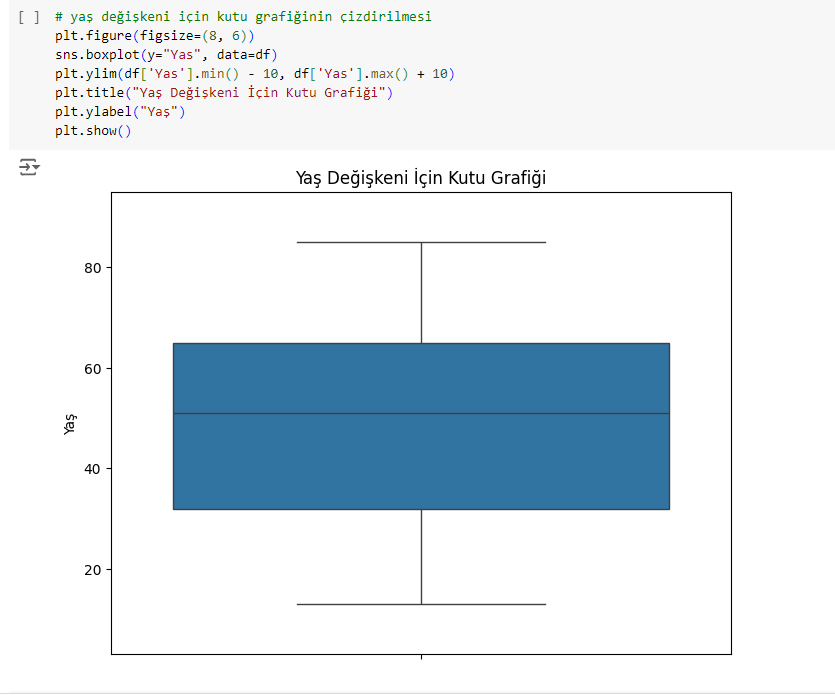
Veri setinde her kan grubuna sahip en az bir bireyin olduğu görüldü. Hangi kan grubundan kaç tane birey olduğu bulundu ve bu dağılım pasta grafiği ile gösterildi. En çok AB RH- kan grubuna sahip birey, en az ise A RH- kan grubuna sahip birey olduğu belirlendi.

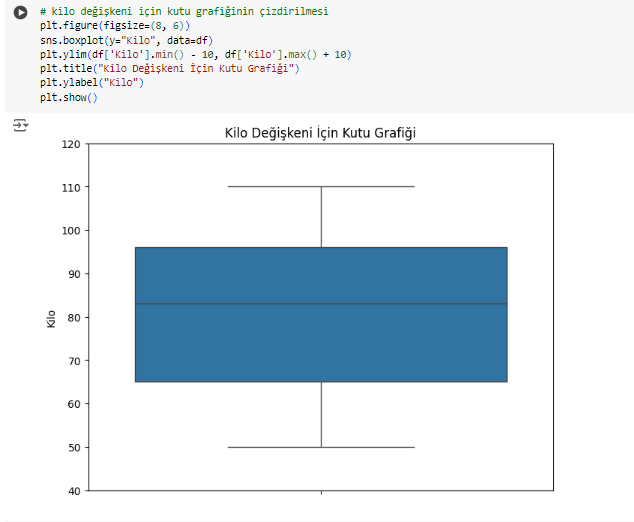


Veri setindeki bireylerin doğum tarihi bilgilerinden daha anlamlı bir bilgi olan yaş bilgileri bulundu. Dogum Tarihi sütununda eksik bir veri yoktu.

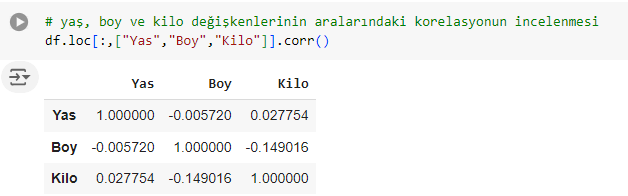
 

Veri setindeki yas, boy ve kilo sütunları için histogram grafikleri çizdirildi. 51 yaş, 181 boy ve 108 kilo değerlerinin en sık olduğu görüldü.





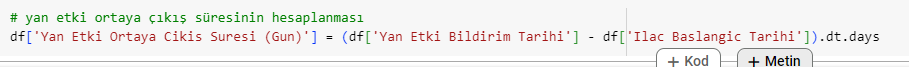
Yas, boy ve kilo değişkenleri için kutu grafikleri oluşturuldu. Burada aykırı değerlerin olmadığı görüldü. Burada describe metodu ile de görmüş olduğumuz alt çeyrek, üst çeyrek, medyan, minimum ve maksimum değerleri görülebilir.



Yas, boy ve kilo değişkenleri arasındaki korelasyonlara bakıldı. Yüksek bir korelasyon oranı bulunamadı. Korelasyon zayıf.



Ilac Baslangıc Tarihi ve Ilac Bitis Tarihi sütunlarından daha anlamlı bir veri olan Ilac Kullanim Suresi (Gun) sütunu oluşturuldu.



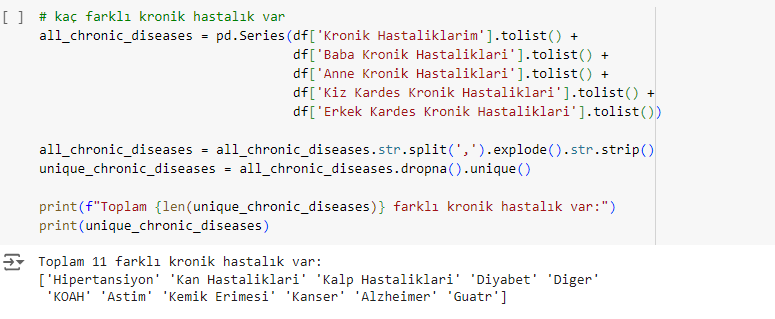
Yan Etki Bildirim Tarihi ve Ilac Baslangic Tarihi sütunlarından daha anlamlı bir sütun olan Yan Etki Ortaya Cikis Suresi (Gun) sütunu oluşturuldu.



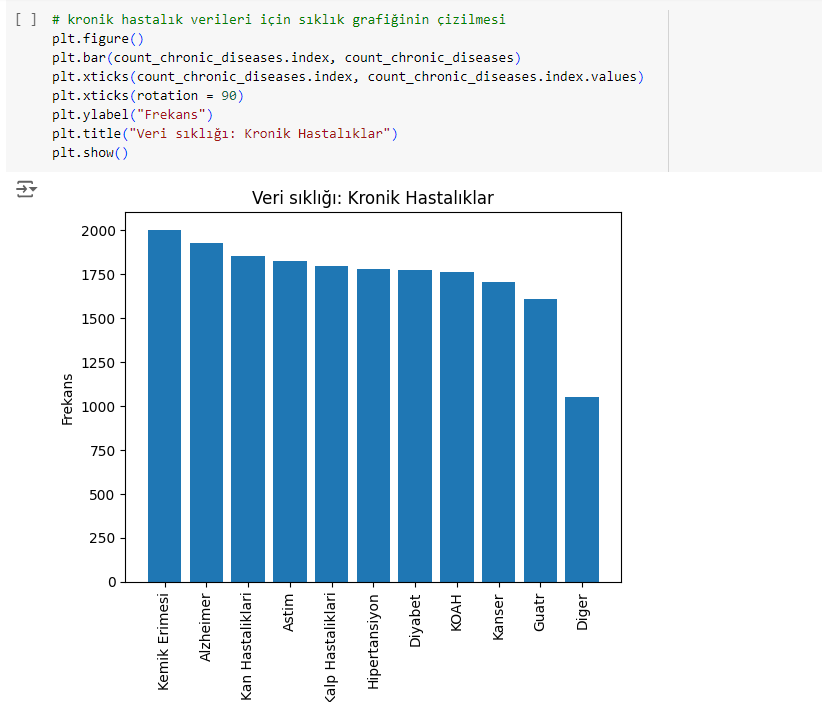
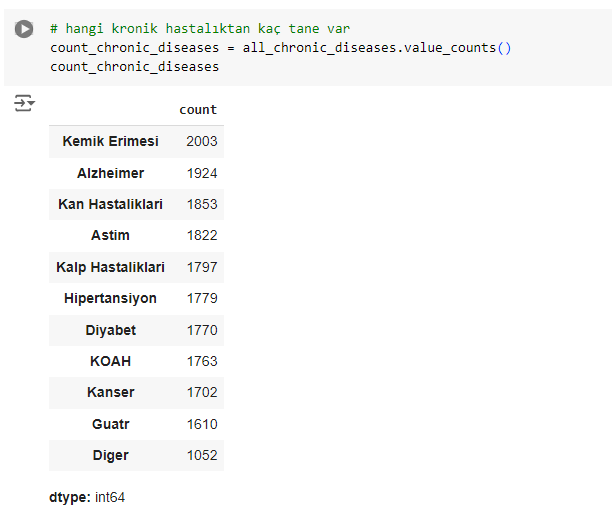
Analizde kullanılmayacak ve tahmin açısından da işe yaramayacak olan Kullanici ID ve Dogum Tarihi sütunları veri setinden kaldırıldı.



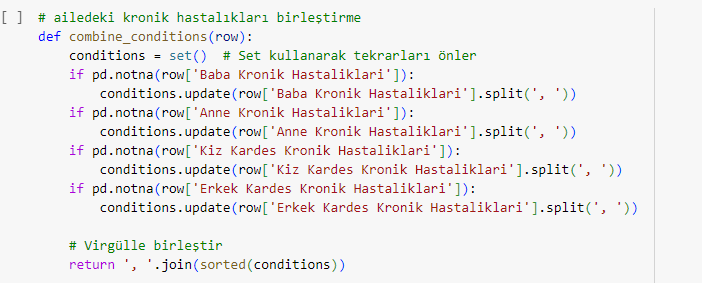
Veri setindeki boş ve NaN değerlerin birlikte incelenebilmesi için boş değerler de NaN olarak değiştirildi.



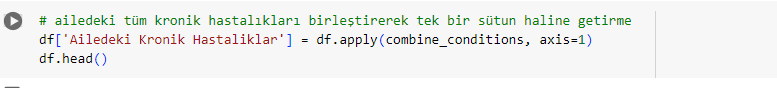
Veri setinde kronik hastalıklar ile ilgili sütunlardaki değerler birleştirilerek Nan değerler çıkarıldı ve kalan değerler unique hale getirildi. Bu işlem sonucunda veri setinde 11 farklı kronik hastalık olduğu bulundu ancak bu kronik hastalıklardan biri de ‘Diger’. Yani 10 kronik hastalık içinde olmayan diğer kronik hastalıkları temsil ediyor.



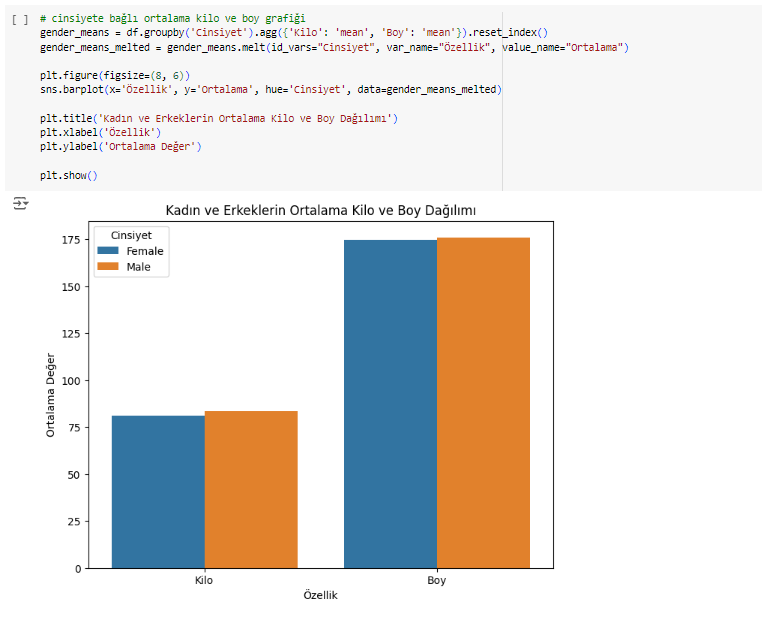
Hangi kronik hastalıktan kaç tane olduğu da tespit edildi. En çok bulunan kronik hastalık kemik erimesi iken, burada olmayan diğer kronik hastalıkların ise en az olduğu belirlendi. Bu verilere ait sütun grafiği de çizdirildi.



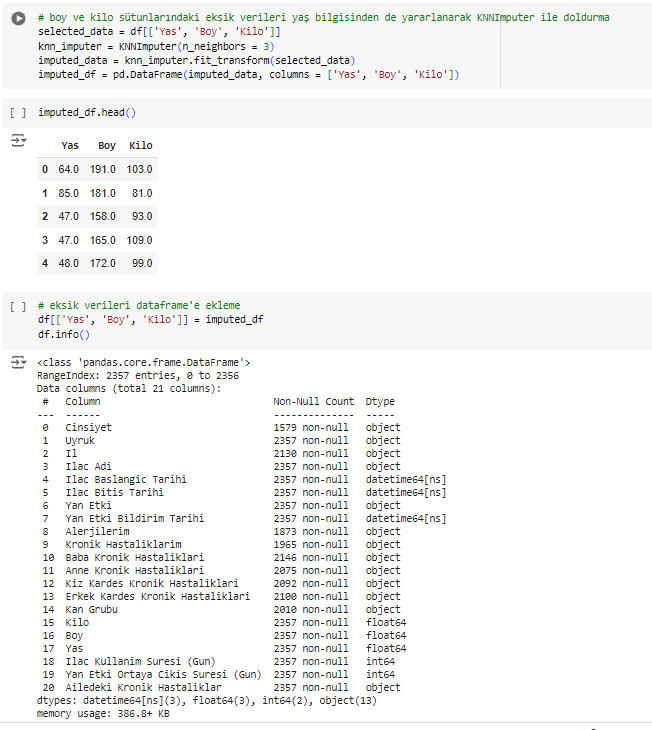
Veri setinde kişinin kendisine ait kronik hastalıklar dışında ailesine ait kronik hastalıklar da yer alıyor. Anne, baba, kız kardeş ve erkek kardeş kronik hastalıkları farklı sütunlarda bulunuyor. Kişinin kendisine ait kronik hastalıkların diğer kronik hastalıklara göre öncelikli ele alınması daha sağlıklı olacağı için kişinin kronik hastalıkları tek başına bir sütunda yer alabilir ancak diğer aile üyelerine ait kronik hastalıkları tek bir sütunda toplamak amacıyla bu sütunları birleştiren bir fonksiyon yazıldı.



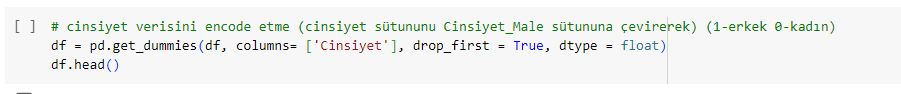
Daha sonra da bu fonksiyon kullanılarak Ailedeki Kronik Hastaliklar isimli yeni bir sütun oluşturuldu.



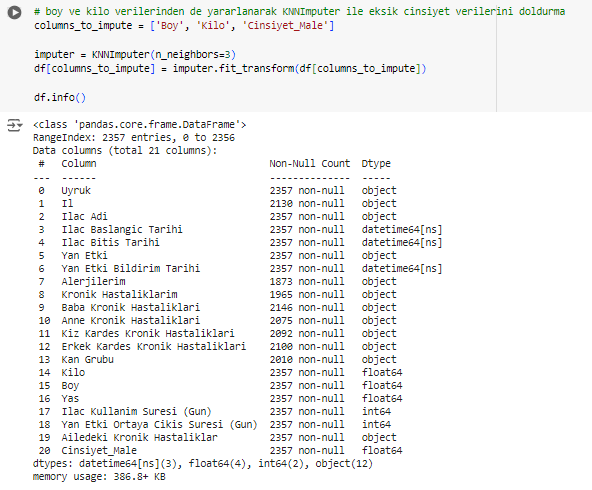
Cinsiyete bağlı ortalama kilo ve boy grafiği çizdirildi. Burada iki cinsiyet için de ortalama kilo ve boy değerlerinin çok yakın olduğu görüldü.



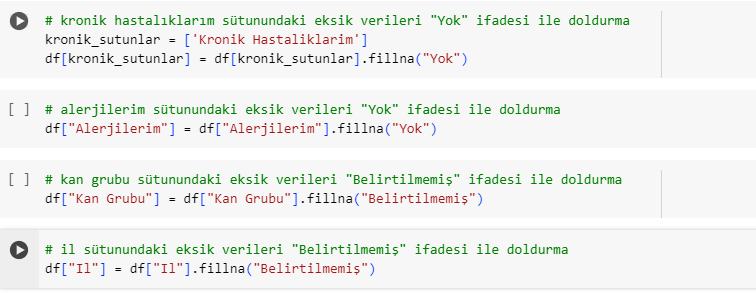
Boy ve kilo sütunlarındaki eksik verileri tamamlamak için KNNImputer kullanıldı. Burada en yakın 3 komşuya bakılarak tahminde bulunuldu. Yas değişkeninde eksik bir veri yoktu ancak boy ve kilo tahminine yardımcı olabileceği için yaş sütunu da kullanıldı. Doldurulan veriler veri setine de aktarıldı. Sonuçta boy ve kilo sütunlarında da eksik veri kalmamış oldu.



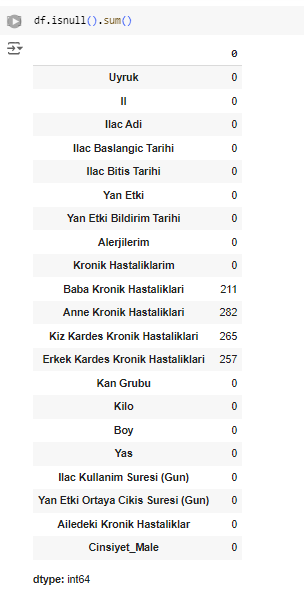
Cinsiyet sütunu kategorik bir veri olduğu için one-hot encode ile Cinsiyet\_Male sütununa dönüştürülerek bu sütuna cinsiyeti erkek olan satırlar için 1, kadın olan satırlar için 0 değeri verildi.



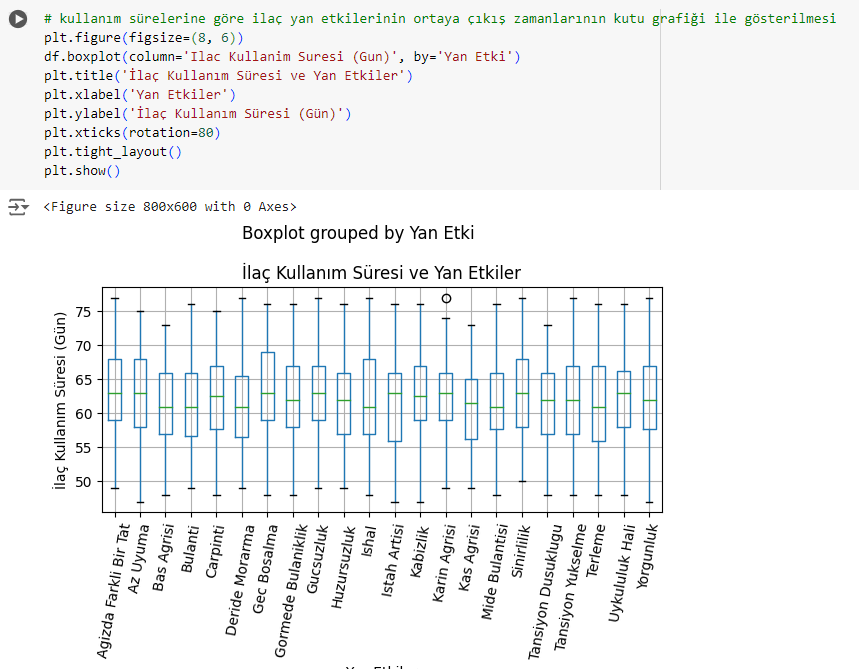
Daha sonra da cinsiyet verilerindeki eksiği tamamlamak için boy ve kilo verilerinden de yararlanılarak KNNImputer ile doldurma işlemi yapıldı. Burada da en yakın 3 komşuya bakılarak tahminde bulunuldu. Böylece cinsiyet verileri de tamamlanmış oldu.



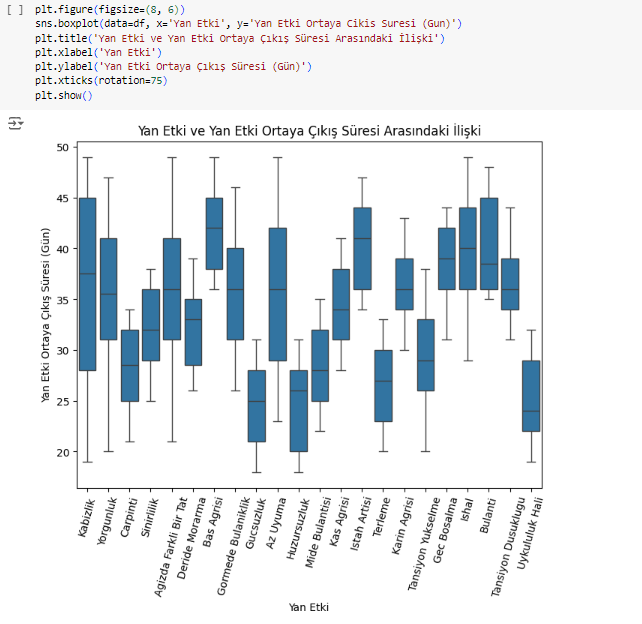
Kronik hastaliklarim ve Alerjilerim sütunlarındaki NaN olan ifadeler “Yok” ifadesi ile dolduruldu. Yani kişiye ait kronik hastalık veya alerji bilgisi girilmediyse yok olarak kabul edildi. Kan Grubu ve Il sütunlarında NaN olan ifadeler ise “Belirtilmemiş” ifadesi ile dolduruldu. Yani kişiye ait kan grubu veya il bilgisi girilmediyse belirtilmemiş şeklinde işaretlenmiş oldu.



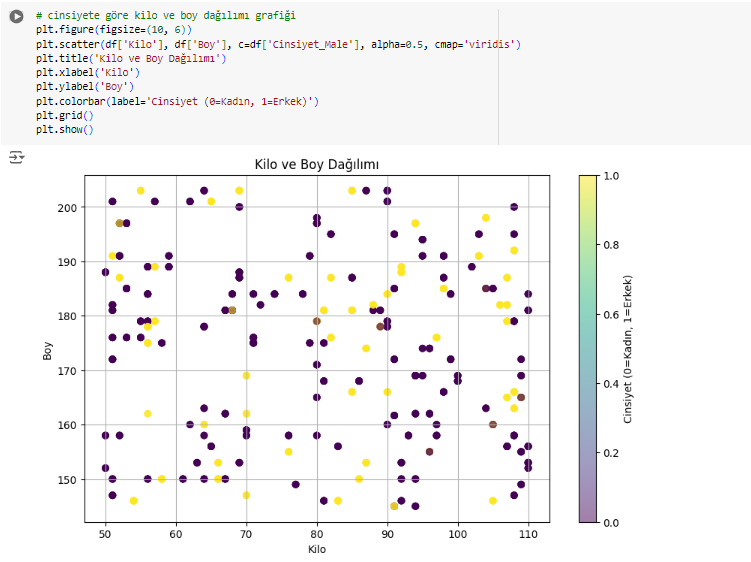
Bu işlemlerden sonra veri setine bakıldığında sadece ailedeki kronik hastalıklara ait sütunlarda eksik veriler kaldığı görüldü. Zaten bu sütunlardaki kronik hastalıklar yeni bir sütunda toplandığı için bu sütunlar veri setinden kaldırılacak.



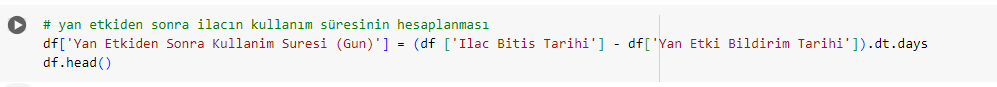
İlaçların kullanım sürelerine göre hangi yan etkilerin ne zaman ortaya çıktığını gösteren kutu grafiği çizdirildi. Ortalama sürelerin yakın olduğu görüldü.



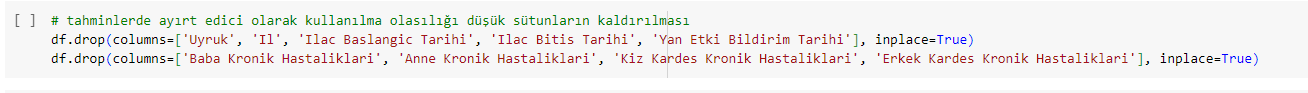
Yan etkilerin ortaya çıkış süresi kutu grafiği de çizdirildi. Burada ortaya çıkış aralığının en geniş olduğu yan etkinin kabızlık olduğu görüldü. Yan etkinin en hızlı ortaya çıktığı örneklerin güçsüzlük ve huzursuzlukta olduğu görüldü. En geç ortaya çıkan yan etki örnekleri ise kabızlık, ağızda farklı bir tat, baş ağrısı, az uyuma ve ishal.



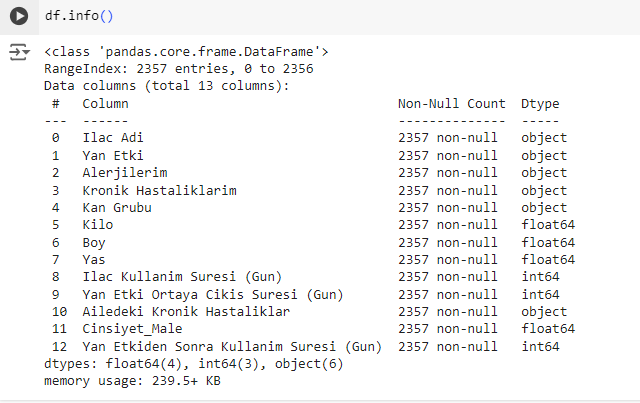
Cinsiyete göre kilo ve boy saçılım grafiği çizdirildi. Zaten cinsiyete göre kilo ve boy ortalamalarının çok yakın olduğu görülmüştü. Burada da kişilere ait kilo, boy ve cinsiyet verileri aynı anda gözlemlenmiş oldu. Boy ve kilo arasında belirgin bir doğrusal ilişki gözlemlenemedi. Hem kısa hem de uzun bireylerin farklı kilo aralıklarına dağıldığı gözlemlendi.



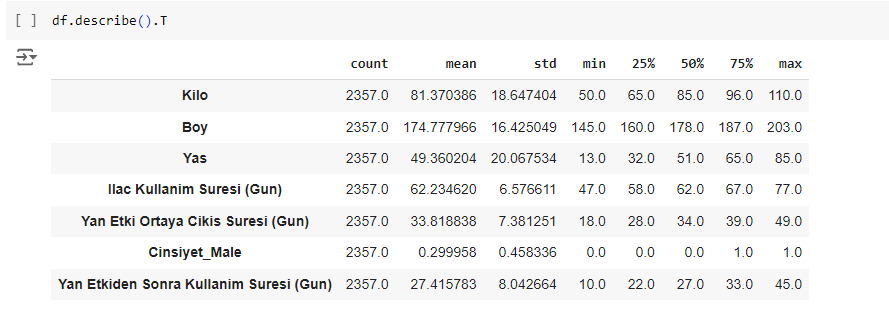
Ilac Bitis Tarihi ve Yan Etki Bildirim Tarihi sütunlarında da daha anlamlı bir sütun olan Yan Etkiden Sonra Kullanim Suresi (Gun) sütunu oluşturuldu.



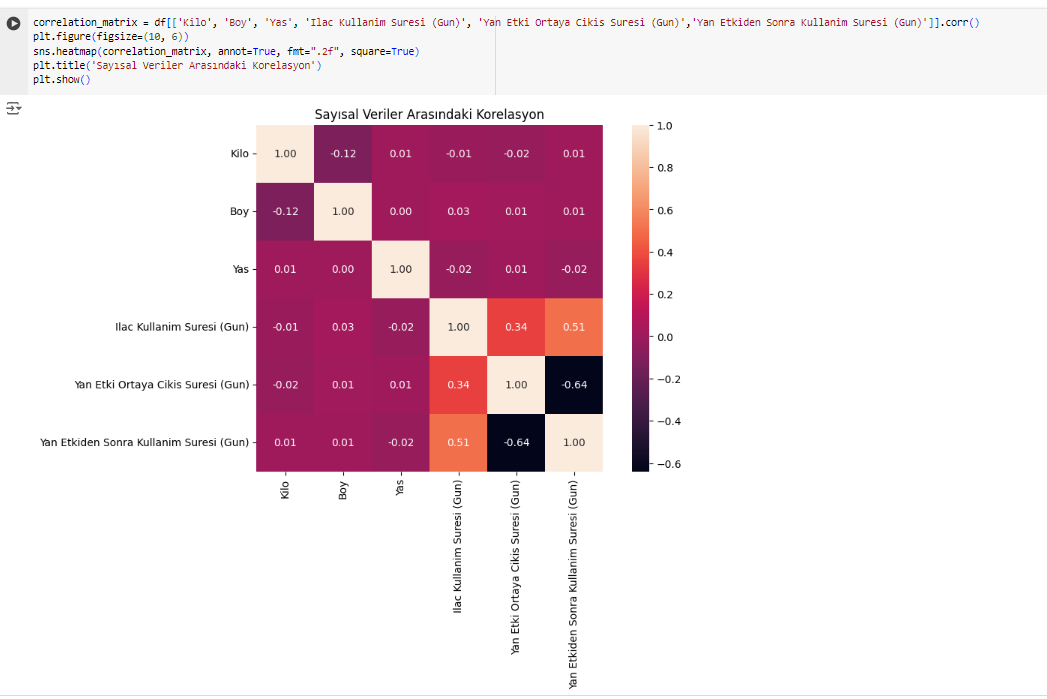
Daha sonra da analizleri yapılan Uyruk ve İl sütunları, ayrıca da analiz yapmaya uygun olmayan tarih sütunlarıyla tek bir sütunda birleştirdiğimiz ailedeki kronik hastalıklara ait sütunlar veri setinden kaldırıldı. Bu sütunların tahminlerde ayırt edici olma olasılığı da düşüktü.



Böylece veri setini temizlenmiş ve eksik veriler tamamlanmış oldu.

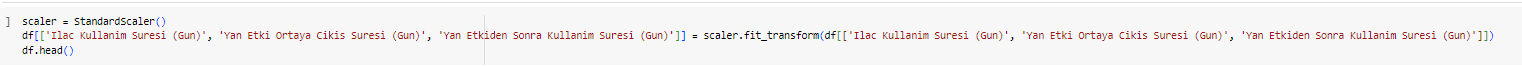


Kilo ve boy sütunlarında eksik veriler doldurulmuştu. Bu doldurmadan sonra ortalamalarda büyük bir değişim olup olmadığı incelendi. Kilo ortalamasının 80’den 81’e çıktığı, boy ortalamasının ise 174’te sabit kaldığı görüldü.

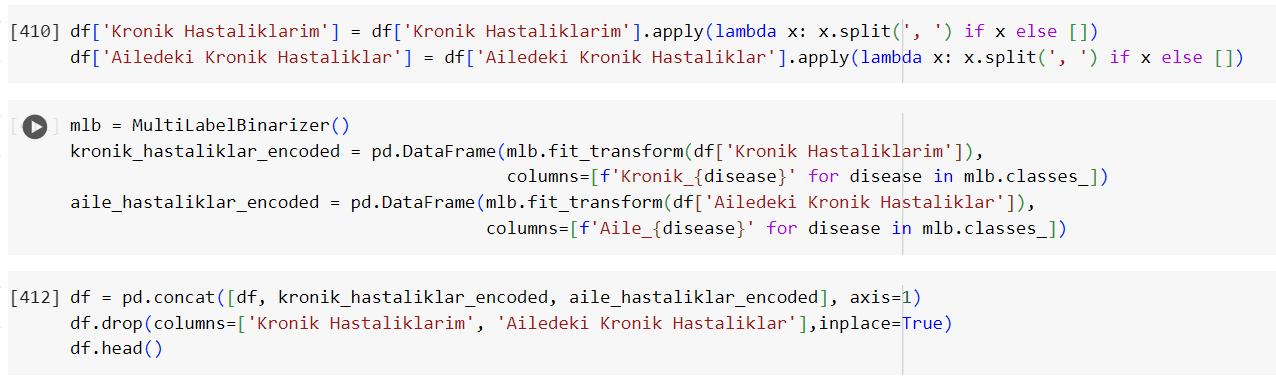


Sayısal veriler arasındaki korelasyonlar ısı grafiği ile gösterildi. En yüksek korelasyonun -0.64 ile yan etki ortaya çıkış süresi ve yan etkiden sonra kullanım süresi arasında olduğu görüldü. Bu iki değer arasında negatif bir korelasyon var, yani yan etkinin ortaya çıkış süresi arttıkça yan etkiden sonra kullanım süresi azalıyor. En yüksek ikinci korelasyonun ilaç kullanım süresi ve yan etkiden sonra kullanım süresi arasında ve 0.51’lik bir korelasyon olduğu görüldü, yani pozitif bir korelasyon var. Zaten ikisi de ilaç kullanımının toplam süresine bağlı olduğu için bu korelasyon normal. Üçüncü en yüksek korelasyon ise ilaç kullanım süresi ve yan etki ortaya çıkış süresi arasında 0.34’lük pozitif bir korelasyon. Bunun dışında diğer değişkenler arasındaki korelasyonlar çok zayıf.



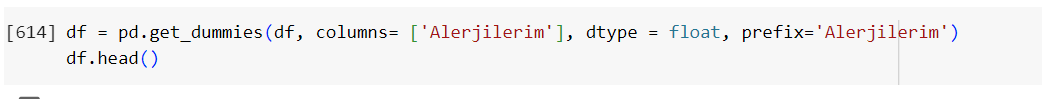


Veri setindeki sayısal veriler StandartScaler kullanılarak ortalamaları 0, standart sapmaları 1 olacak şekilde ölçeklendirildi.

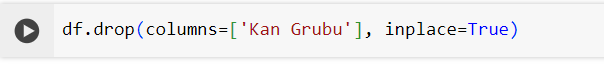


Kronik Hastaliklarim ve Ailedeki Kronik Hastaliklar sütunlarında virgülle ayrılmış birden çok değer olduğu için bu değerlere MultiLabelBinarizer metodu uygulanarak encode edilmiştir. Bu metod virgülle ayrılmış tüm değerleri ayrı ayrı encode eder. Daha sonra encode edilmiş sütunlar veri setine eklenmiş ve eski halleri de veri setinden kaldırılmıştır.





Ilac Adi sütunu kalacak şekilde diğer sütunlar da encode edildi. Yan Etki ve Alerjilerim sütunlarında nominal veriler olduğu için one-hot encoding kullanılmıştır.



Kan Grubu sütunu bu tahminlerde diğer sütunlara göre daha az ayırt edici olabilir. Duruma göre veri setinden kaldırılabilir.