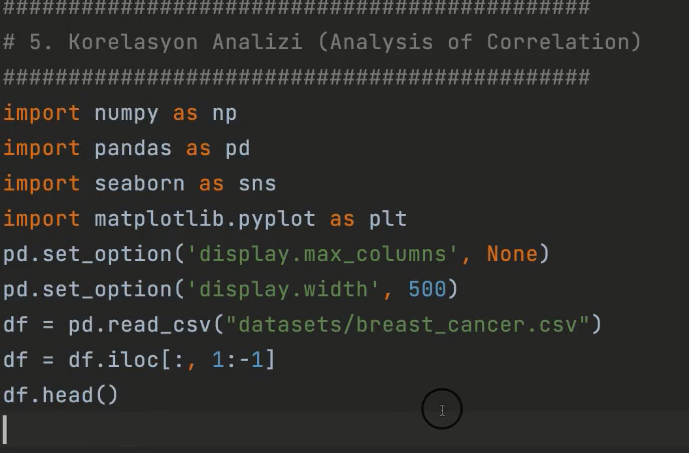
**Korelasyon Analizi**

Başlamadan önce

Korelasyon analizine başlamadan önce hatalı id’leri çıkartmak amacıyla df.iloc işlemi yapıyoruz.



Bu sayede hatalı değerler dışarıda kalacaktır.

Korelasyon amacında harekete geçecek olursak amaç şudur

Isı Haritası yardımıyla yüksek korelasyonlu değişkenleri tespit etmek ve bu değişkenleri dışarıda bırakmaktır.

**NOT**: Bu amaç her veri seti için geçerli değildir, sadece analiz aracı olarak kullanılmalıdır.

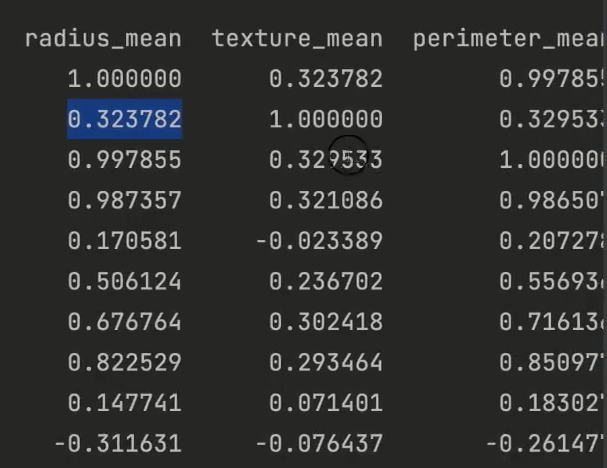
Şimdi sayısal değişkenleri seçmek için (korelasyon, sayısal değişkenlerde mevcuttur.)



Korelasyonları hesaplamak için “corr” fonksiyonundan yararlanalım.



Burada mevcut korelasyon değerlerini görmekteyiz (resultta çoklama olduğu için aynı değerler birden fazla kez tekrar ediyor)



Korelasyon, değişkenlerin birbiriyle ilişkisini ifade eden bir istatistiki ölçümdür.

Bu ölçüm değerleri +1 ve -1 aralığında değer almaktadır. Bunlara negatif ve pozitif korelasyon denilmektedir.

Değerler +1 veya -1’e yaklaştıkça ilişkinin şiddeti kuvvetlenir. (pozitif veya negatif bakılmaksızın)

**NOT**: Korelasyon değeri ne kadar sıfır (0) değerine yakınsa bu o kadar korelasyon olmadığına işarettir.

Korelasyon değeri ne kadar tam (bir) değerine yakınsa bu o kadar korelasyon olduğuna işarettir. (Birbirinin aynısını ifade eder)

Diğer bir örnekle 0.50 aralığında düşük korelasyon demek mümkündür.

Eğer ilişki pozitif yönlüyse buna Pozitif Korelasyon denir ve bir değişkenin değeri arttıkça ilişkili diğer değişkenin de değeri artar

Eğer ilişki negatif yönlüyse buna Negatif Korelasyon denir ve bir değişkenin değeri azalırken ilişkili diğer değişkenin de değeri azalır.

Bizler veri seti içerisinde bulunan yüksek korelasyonlu değerleri (0.90 > ) bir arada bulunsun istemeyiz.

Çünkü bu değişkenler yüksek korelasyon nedeniyle zaten birbirlerinin aynılarını ifade etmektedir.

Isı Haritası üzerinden yüksek korelasyon değerlerini incelemek istersek

metin, ekran görüntüsü, yazı tipi, yazılım içeren bir resim

Açıklama otomatik olarak oluşturuldu

ekran görüntüsü, kalıp, desen, düzen, metin, renklilik içeren bir resim

Açıklama otomatik olarak oluşturuldu

Koyu maviler pozitif korelasyonu, koyu kırmızılar ise negatiaf korelasyonu ifade etmektedir.

**Yüksek Korelasyonlu Değişkenlerin Silinmesi**

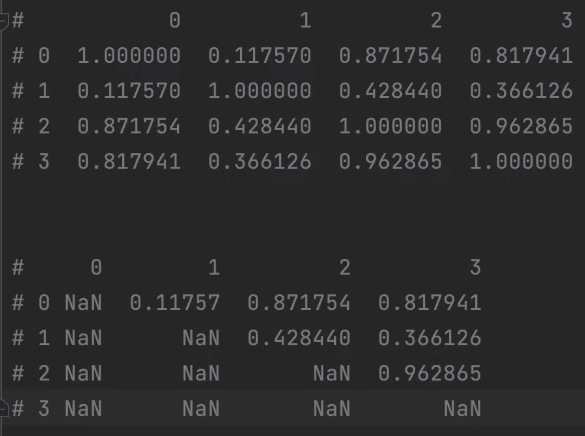
metin, yazı tipi, grafik, siyah içeren bir resim

Açıklama otomatik olarak oluşturuldu

**NOT**: Yüksek korelasyonlu ifadelerin pozitif veya negatif olduklarına bakılmaksızın işlem gerçekleştirdiğimiz için

ABS yani mutlak değer fonksiyonu kullanmaktayız.

Tüm işlemlerimiz bittiğinde result olarak elimizde bu türden bir çıktı olmasını bekliyoruz. (Çoklamadan kaynaklı değişk. Değerleri silinmiş)

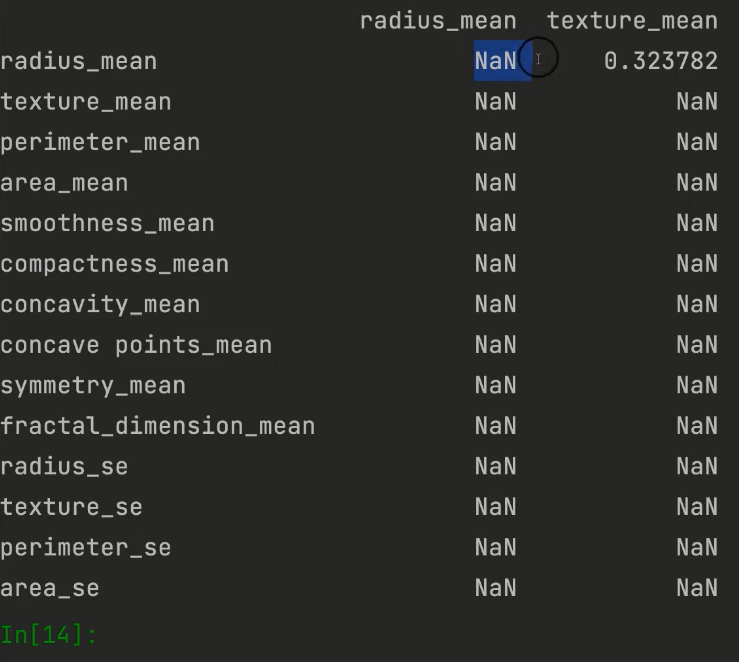


Silme işlemini gerçekleştirmek için NumPy’ın içerisinde yer alan bir fonksiyondan yararlanacağız.





Bu fonksiyon sayesinde çoklama nedeniyle oluşan değerler temizlendi



Daha sonrasında ise yüksek korelasyona sahip değerleri (0.90 >) temizleyeceğiz

**…**



Result olarak silinen değişkenlere baktığımızda bu değişkenlerin silinmiş olduğunu görüyoruz

metin, yazı tipi, ekran görüntüsü, tasarım içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, yazı tipi, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Bu silme işlemleri ardından 10 değişken silinmiş oldu

**Fonksiyonel Korelasyon Analizi İşlemleri**

metin, ekran görüntüsü, yazılım, işletim sistemi içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, yazılım, multimedya yazılımı içeren bir resim

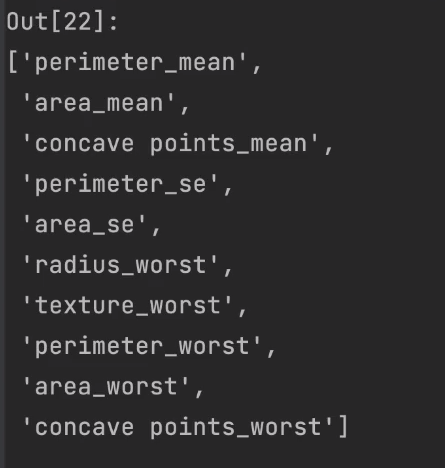
Açıklama otomatik olarak oluşturuldu

metin, yazı tipi, ekran görüntüsü, grafik içeren bir resim

Açıklama otomatik olarak oluşturuldu

Fonksiyona, tanımladığımız veri setini giriyoruz ve işleme tabii tutuyoruz.

Resultx”

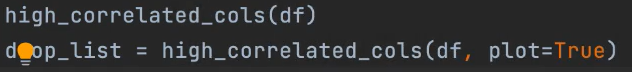


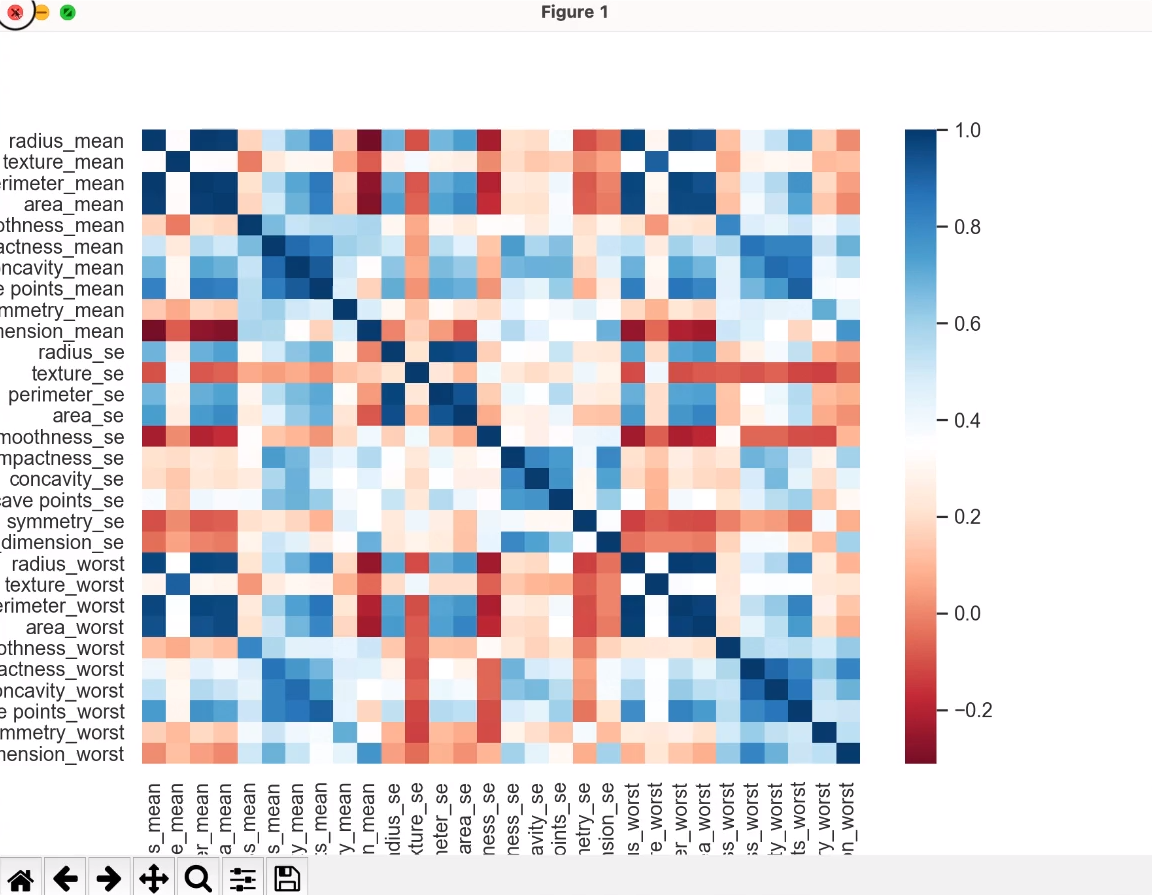
**NOT:** Veri setindeki yüksek korelasyonlu değişkenler olması makine öğrenmesi (model optimizasyonu, model deployment ve development), sürdürülebilirlik ve projenin el değiştirmesi süreçlerini

Özetle, negatif etkileyebilir. Bu yüzden yüksek korelasyona sahip değişkenlerden kurtulmamız gerekmektedir.

**NOT:** Yüksek korelasyona sahip değerleri kontrol etmeden asla silmemek gerekir

Yaptığımız işlemlerin sağlamasını yapmak için





Şimdi de silinmiş halini grafiğe ekleyelim



Koyu renkteki değişkenleri şimdi göremiyoruz, demek ki fonksiyon doğru çalışıyor.

metin, ekran görüntüsü, kare, tasarım içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Kaggle Yarışma Veri Seti Örneği ile Uygulamalar**

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Veri setinde görüldüğü üzere yüksek korelasyona sahip çok fazla değişken var

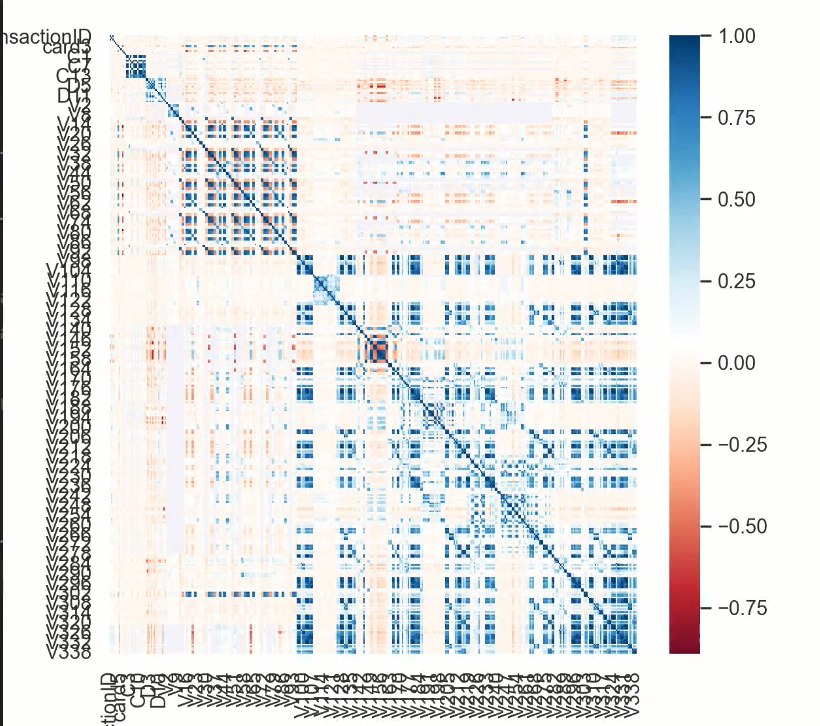
metin, ekran görüntüsü, menü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Bunları ısı haritasıyla bakarak da görelim



Isı haritası ile birlikte korelasyonların incelenmesi



Kontroller sonrası silme işlemlerini ve kaç adet olduğunu görelim (198 silindikten sonraki adet)



**NOT**: Bu şekilde işlemler yapmak bir maharet değildir.

Eğer veri setimizde el, göz yordamıyla analiz yapamayacak kadar yüksek korelasyona sahip değerler var ise bu fonksiyona, işlemlere başvurmalıyız.

Aksi takdirde veri setini darmadağınık bir hale getirmiş oluruz.