

**KEŞİFSEL VERİ ANALİZİ VE VERİ DÖNEM PROJESİ**

**TUĞBA KARADENİZ-200301025**

Bu proje çalışmamda veri seti olarak Türkiye İstatistik Kurumu’ndan aldığım 1927-2000 yılları arası genel nüfus sayımı sonuçları ve 2007-2022 arası adrese dayalı nüfus kayıt sistemi sonuçlarını kullandım. Veriye aşağıdaki linkten ulaşabilirsiniz.Veriyi indirirken bilgisayarıma .xlsx olarak kaydediyorum.

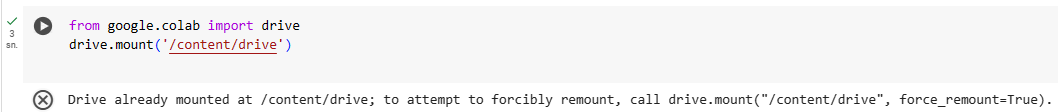
[Nüfus İstatistikleri Portalı (tuik.gov.tr)](https://nip.tuik.gov.tr/)

metin, ekran görüntüsü, sayı, numara, yazılım içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Google Collab Ortamına Veri Yüklenmesi**

Veriyi excell üzerinde 1927’den 2022’ye olacak biçimde düzenleyip Google Drive üzerinde 01\_MHT353\_Data ismi ile açtığım klasörün içine Data.xlsx olarak kaydediyorum.Ve aşağıdaki kodu çalıştırarak Google Colab ortamını drive'a bağlıyorum.



Verimi içeren Data.xlsx dosyasını aşağıdaki kodun ilk satırında import ettiğim pandas paketindeki "read\_excel" fonksiyonunu kullanarak içeri aldım.



**Verinin İncelenmesi**

Verimin ilk 5 satırına bakıyorum:

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Verim 1927 yılından başlıyor.

Verimin son 5 satırına bakıyorum:

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Verim 2022 yılında sona eriyor.

Verimin tamamına bakıyorum:

metin, ekran görüntüsü, doküman, belge, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, sayı, numara, doküman, belge içeren bir resim

Açıklama otomatik olarak oluşturuldu

Verimin içinde 3 değişken var.Yıl,düzey,toplam,erkek,kadın olmak üzere bu verilerin tiplerine bakıyorum:

metin, yazı tipi, sayı, numara, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Hipotez Kurma**

Bu aşamada verimin toplam nüfus kısmı için H0 yokluk hipotezi (sıfır hipotezi) ve H1 araştırma tek yönlü hipotezlerini kuruyorum.

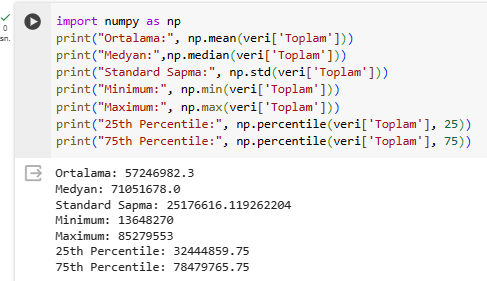
H0:X¯¯¯¯≤50000000

H1:X¯¯¯¯>50000000

Burada H0 hipotezi X dağılımının ortalamasının 50000000'a eşit yada 50000000'dan küçük olduğunu test eder.

**Verinin Dağılımı**

Hesaplamalara başlamadan önce Python’da matematik hesapları yapmak için kullanılan numpy paketini yüklüyorum.



metin, elektronik donanım, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

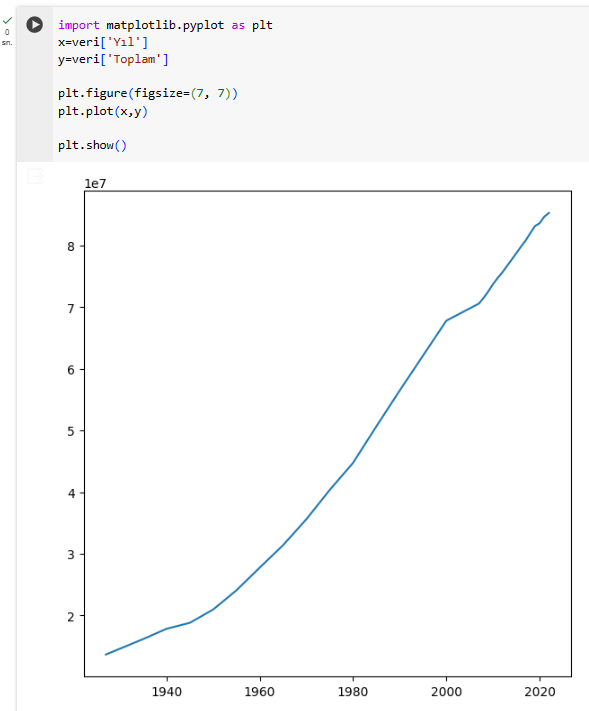
Numpy standart sapmayı hesaplarken verilen veriyi anakütle kabul eder, ancak pandas'ın .describe() fonksiyonu verilen veriyi örneklem kabul eder bu nedenle standart sapmaları farklıdır.

metin, çizgi, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Veri Görselleştirme (Çizgi Grafik)**

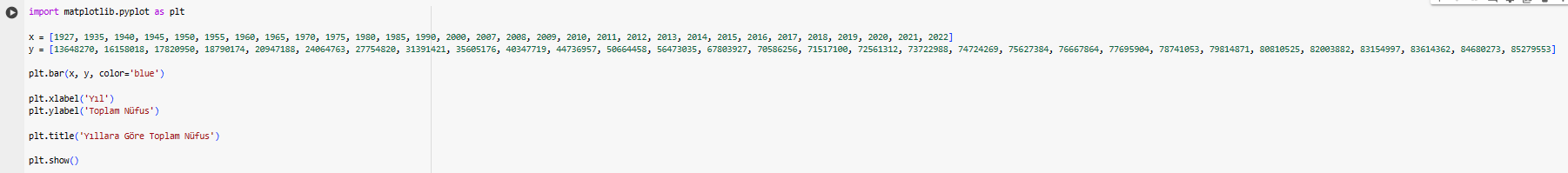
Verim içerisindeki toplam,erkek,kadın sayılarının zaman serisi grafiklerini çizdirmek için **matplotlib** adı verilen görselleştirme paketini çağırıyorum.



metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Veri Görselleştirme(Sütun Grafik)**

****

**metin, ekran görüntüsü, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**Veri Görselleştirme (Serpilme Diyagramı Ve Korelasyon Analizi)**

**metin, ekran görüntüsü, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**metin, ekran görüntüsü, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu**

İki değişken arasındaki ilişkinin bayağı yakın olduğunu gördüm.**numpy** içerisindeki **corrcoef()** fonksiyonunu kullanarak ikisinin arasındaki korelasyonu hesaplıyorum.

metin, yazı tipi, çizgi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Değişkenler arasındaki korelasyonu **pandas** altındaki **corr()** fonksiyonu ile de hesaplıyorum.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Burada kadın ve erkek nüfuslarını serpilme diyagramı ve korelasyon katsayıları aracılığıyla karşılaştıtracak olursam aralarında serpilme diyagramında pozitif bir korelasyon durumunun olduğunu ve korelasyon katsayısının da 0.999695 çıkarak bu pozitif korelasyonu desteklediğini görüyorum.Bu iki değişken arasında %99 oranında bir korelasyon vardır.

Korelasyon sıcaklık haritasını oluşturabilmek için **seaborn** adı verilen biraz daha gelişmiş özellikleri bulunan bir paketi yüklüyorum.

metin, ekran görüntüsü, sayı, numara, yazılım içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Veri Görselleştirme (Box-Plot)**

metin, ekran görüntüsü, ekran, görüntüleme, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, doküman, belge, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Veri setimi incelediğimde şimdilik bir outlier veri saptayamadım.Boxplot'uma farklı estetik özellikler de ekliyorum.

metin, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Veri Görselleştirme (Histogram)**

**metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Verimde her veri bir kez geçtiği ve sürekli artış olduğu için bir mod yok.Şimdi histogramıma bazı estetik öğeler ekliyorum.Örneğin ortalama, medyan, mod değerlerini birer çizgi ile gösterip başlıkları ekliyorum.

**metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Standart sapma değerine bakıyorum.

**metin, yazı tipi, ekran görüntüsü, beyaz içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**Tek Örneklemin Ortalamasının Bir Değerden Büyüklüğü Testi - Tek Kuyruk**

Bu aşamada hipotez kurma aşamasında kurguladığım H0 yokluk hipotezi (sıfır hipotezi) ve H1 araştırma tek yönlü hipotezlerini toplam nüfus veri setinden seçtiğim bir örneklem ile t testi aracılığıyla test ediyorum.

H0:X¯¯¯¯≤50000000

H1:X¯¯¯¯>50000000

Burada H0 hipotezi X dağılımının ortalamasının 50000000'a eşit yada 50000000'dan küçük olduğunu test eder.

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

Açıklama otomatik olarak oluşturuldu**

**ekran görüntüsü, çizgi, tasarım içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Formülde değerleri yerine koyarsak:

t=(65455642,06-50000000)/(14799112,97/√16)=4,18

Yukarıdaki hesaplamada görüldüğü gibi test istatistiğimi 4,18 olarak buldum. Bu aşamada ortalaması 0 olan çan eğrisi grafiğime 4,18'i işaretleyeceğim. Aslında standartlaştırılmış bir fark ölçüsü olan bu değer 65455642,06 değerinin 50000000 değerinden anlamlı derecede farklı olup olmadığını test etme amacı işler. Şimdi 4,18'in yeterince güçlü bir fark olup olmadığını değerlendirmek için bir kritik değere ihtiyacım var.Kritik değer standartlaştırılmış düzlemde öyle bir değerdir ki bu değerin üzerinde bir t istatistiğimiz var ise yukarıda test ettiğim 65455642,06 ile 50000000 arasındaki farkın anlamsız olması kritik değerin üzerinde kalan alan kadar bir olasılıktır.Örneğin %95 güven için tek kuyruk olasılıklı t tablosu kritik değerinin 1,64 olduğunu düşünürsek bu değer üzerinde herhangi bir noktada iki ortalama arasındaki farkın istatistiksel olarak anlamsız olmasının ancak %5 şansı vardır.

metin, çizgi, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

Testim tek yönlü ortalamanın 50000000 üzerinde olup olmadığı hipotezi için yapıldığından sağ etekten %5 hata payı bırakıyorum. Göründüğü üzere t testi istatistiğim (standartlaştırılmış fark) 1,64 olduğu için gayet anlamlıdır. Örneklem ortalamam 50000000'in haylice üzerindedir.Sonuç olarak H0 hipotezi reddedilir.