

VERİ MADENCİLİĞİ

VİZE ÖDEVİ

Mustafa KURÇEREN



NOT: …………………..….

**ANKARA MEDİPOL ÜNİVERSİTESİ**

**BİLGİSAYAR PROGRAMCILIĞI BÖLÜMÜ**

VERİ MADENCİLİĞİ DERSİ ARA SINAV ÖDEVİ

**ADI SOYADI:……………………………………………… NO:……………………………**

Verilen “datamining.csv” veri seti üzerinde, gerekli temizleme ve veri türleri üzerinde değişiklikler yapılacak (yapılan her işlem anlatılacak);

1. Veri setinin özet analizi çıkarılacak, ve bu özet ekran görüntüsü ödevde gösterilecek,
2. Correlasyonu (Kendall, Spearman ve Pearson yöntemleri olarak tek tek ) hesaplanacak,

* Correlasyon ekran matrix görüntüsü (correlasyon sonucu) aynen yazılacak,
* Matrixte anlamlı ilişkisi olan veriler üzerinde sonuçlar değerlendirilecek ve yorum bölümüne yazılacak.

Correlasyonda anlamlı çıkan her ikili veri türü için;

1. Linear Regrasyon slope-intercept değerleri tespit edilecek,
   * Plotting olarak Grafiği Çizilecek
   * f(x) fonksiyonu bulunacak

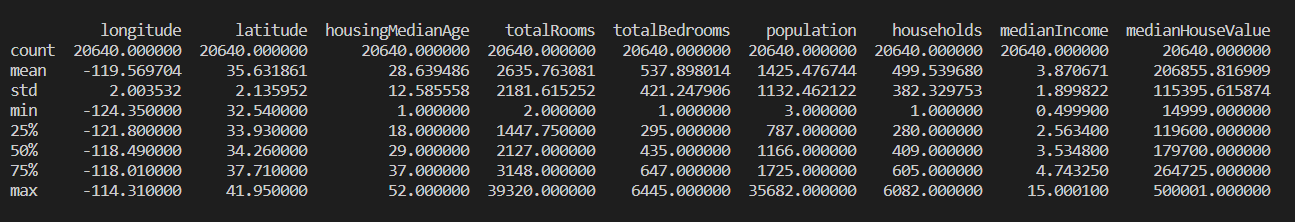
1. Linear Regrasyon yöntemi uygulanacak,
   * train-test datası olarak ayrı ayrı hesaplanacak
   * Plotting Grafiği oluşturulacak
   * Linear R2 değeri hesaplanacak

* Tüm sonuçlar değerlendirilerek yorum bölümüne yazılacak.

*Başarılar dilerim.*

Verilen “datamining.csv” veri seti üzerinde, gerekli temizleme ve veri türleri üzerinde değişiklikler yapılacak (yapılan her işlem anlatılacak);

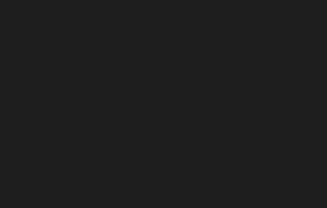
1. Veri setinin özet analizi çıkarılacak, ve bu özet ekran görüntüsü ödevde gösterilecek,

Desceribe Kullanılarak Analiz Yapma

**metin, levha, plaka içeren bir resim

Açıklama otomatik olarak oluşturuldu**Statistics Kullanılarak Analiz Yapma

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu****metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

1. Correlasyonu (Kendall, Spearman ve Pearson yöntemleri olarak tek tek ) hesaplanacak,

* Correlasyon ekran matrix görüntüsü (correlasyon sonucu) aynen yazılacak,
* Matrixte anlamlı ilişkisi olan veriler üzerinde sonuçlar değerlendirilecek ve yorum bölümüne

yazılacak.

Sadece Corr() ile Yapılan Korelasyon

**masa içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Kendall ile Yapılan Korelasyon

**masa içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Pearson ile Yapılan Korelasyon

**masa içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Spearman ile Yapılan Korelasyon

metin, bilgisayar içeren bir resim

Açıklama otomatik olarak oluşturuldu

Correlasyonda anlamlı çıkan her ikili veri türü için;

1. Linear Regrasyon slope-intercept değerleri tespit edilecek,
   * Plotting olarak Grafiği Çizilecek
   * f(x) fonksiyonu bulunacak

Linear Regrasyon Slope-Itercept Değerler

**metin içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Metin hali:

longitude latitude slope intercept değeri: [ -0.98577886 -82.23742528]

totalRooms totalBedrooms slope intercept değeri: [ 0.17955303 64.6387697 ]

totalRooms population slope intercept değeri: [ 0.44492845 252.75075521]

totalRooms households slope intercept değeri: [ 0.16096511 75.27377369]

totalBedrooms population slope intercept değeri: [ 2.36044072 155.80036918]

totalBedrooms households slope intercept değeri: [ 0.88930475 21.18442068]

population households slope intercept değeri: [ 0.30628668 62.93513925]

medianIncome medianHouseValue slope intercept değeri: [41793.8492019 45085.57670327]

f(x) Fonsiyonu Bulunacak

longitude-latitude Slope Intercept: f(x)=(-0.98577886 x) + ( -82.23742528)

totalRooms-totalBedrooms Slope Intercept: f(x)=( 0.17955303 x) + ( 64.6387697)

totalRooms-population Slope Intercept: f(x)=( 0.44492845 x) + ( 252.75075521)

totalRooms-households Slope Intercept: f(x)=( 0.16096511 x) + ( 75.27377369)

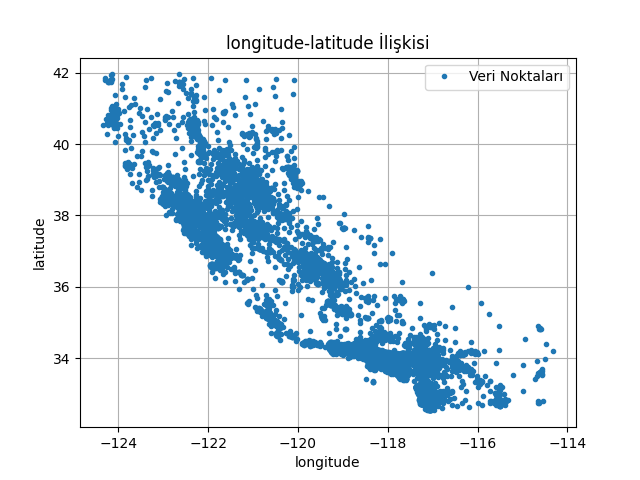
totalBedrooms population slope Intercept: f(x)=( 2.36044072 x) + (155.80036918)

totalBedrooms households slope Intercept f(x)=(2.36044072 x ) +(155.80036918)

population households Slope Intercept : f(X)= (0.30628668 x ) + (62.93513925)

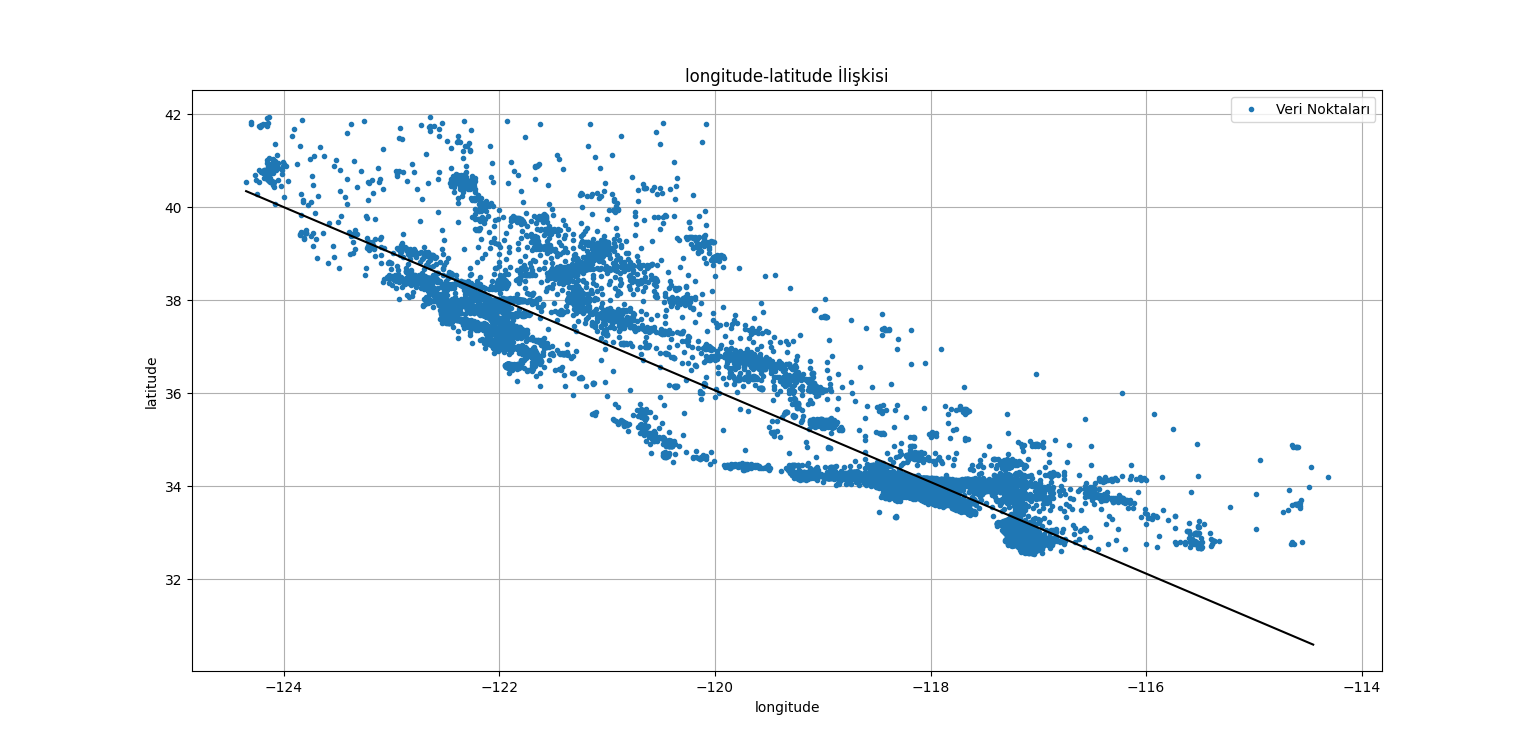
medianIncome-medianHouseValue Slope Interce: f(x)=( 41793.8492019 x) + ( 45085.57670327)

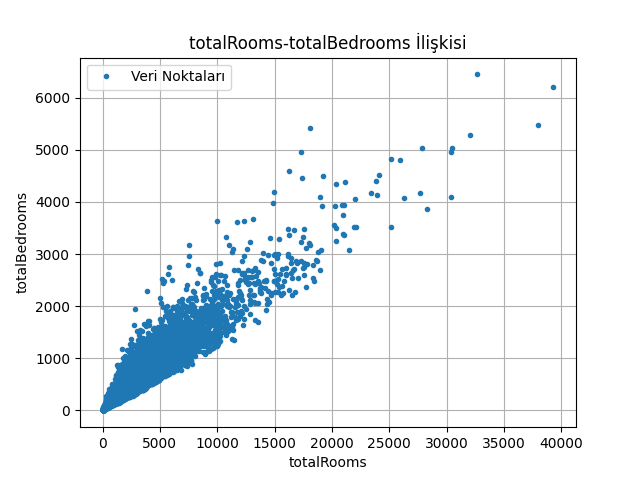
**Plotting Olarak Grafik**

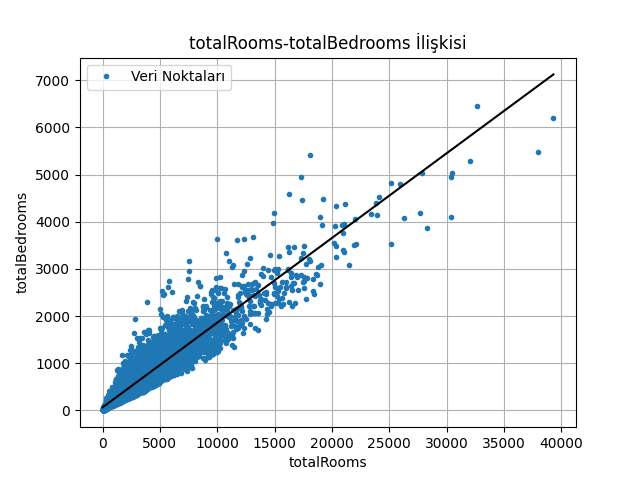
****a)“longitude” ve “latitude” Grafiği

a.1)“longitude” ve “latitude” Grafiği ve

Slope-Itercept ile Oluşturulmuş Doğru

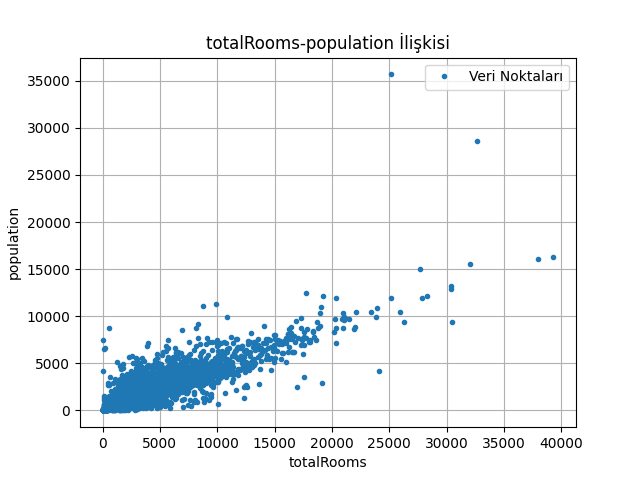
****

b)“totalRooms” ve “totalBedrooms” Grafiği

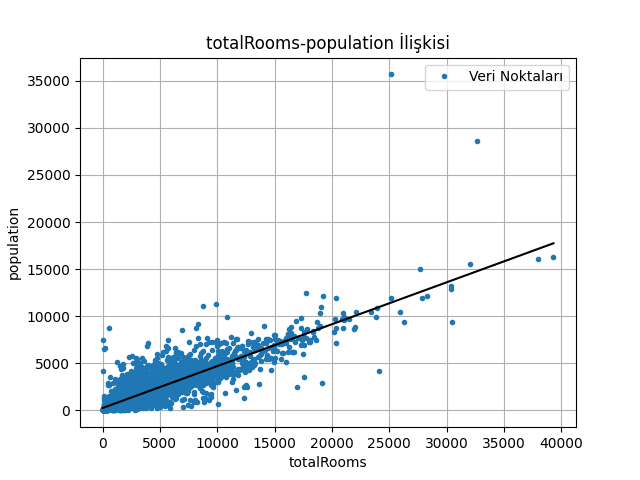


b.1)“totalRooms” ve “totalBedrooms” Grafiği ve

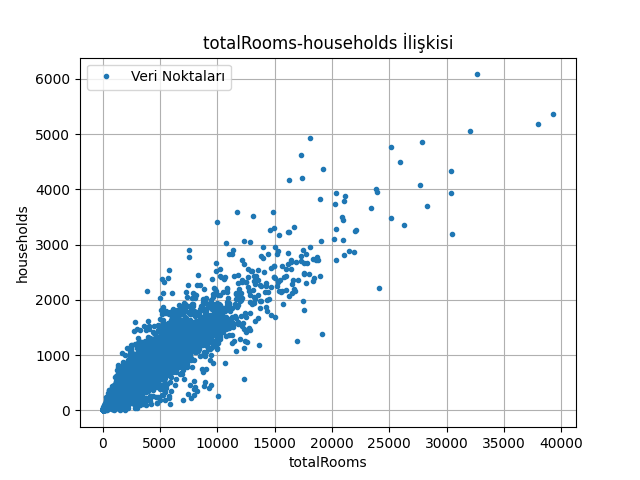
Slope-Itercept ile Oluşturulmuş Doğru

c)“totalRooms” ve “population” Grafiği

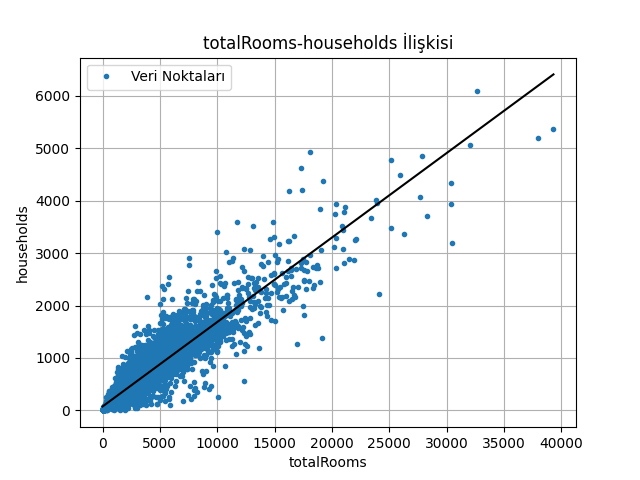
c.1)“totalRooms” ve “population” Grafiği ve

Slope-Itercept ile Oluşturulmuş Doğru

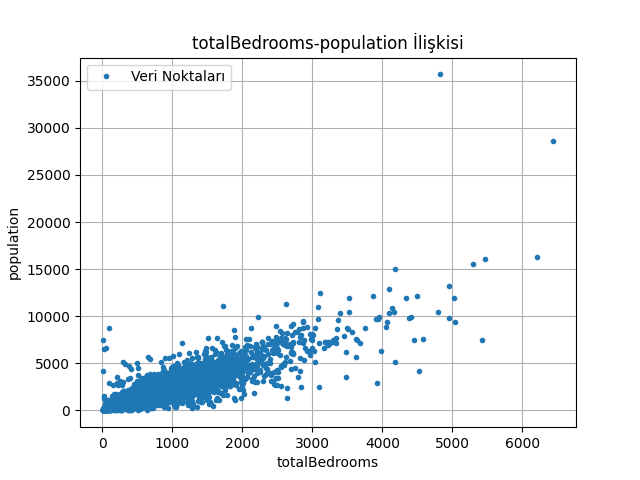
d)“totalRooms” ve “households” Grafiği



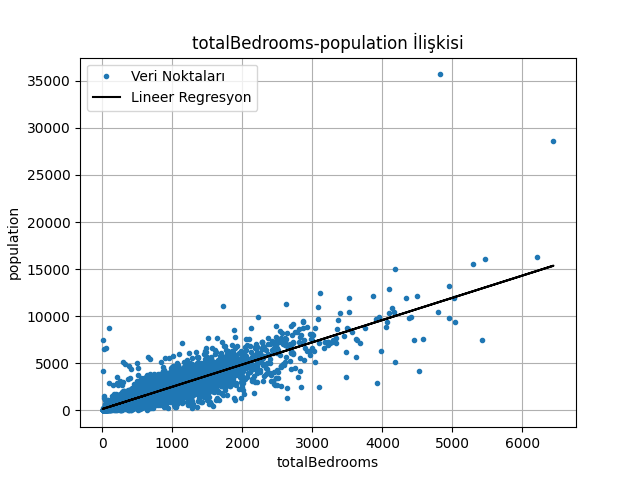
d.1)“totalRooms” ve “households” Grafiği ve

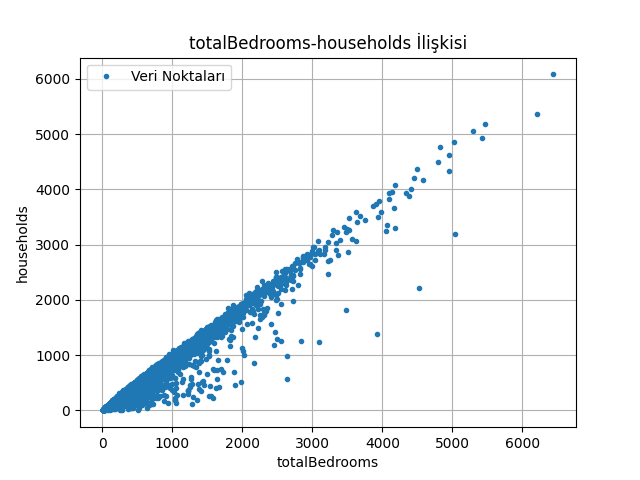
Slope-Itercept ile Oluşturulmuş Doğru

e)“totalBedrooms” ve “population” Grafiği



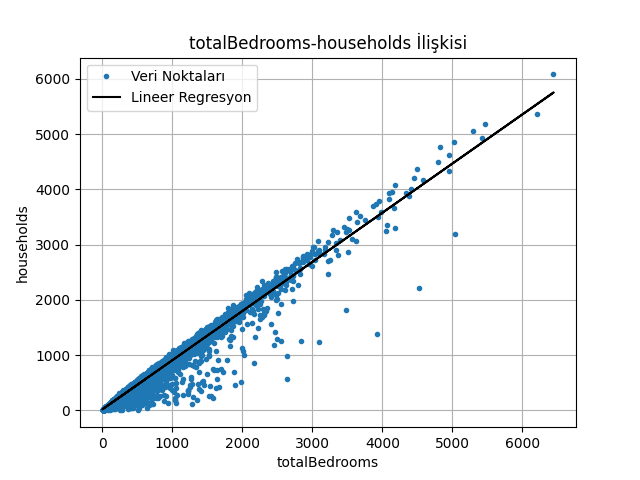
e.1)“totalBedrooms” ve “population” Grafiği ve

Slope-Itercept ile Oluşturulmuş Doğru

f)“totalBedrooms” ve “households” Grafiği

f.1)“totalBedrooms” ve “households” Grafiği ve

Slope-Itercept ile Oluşturulmuş Doğru



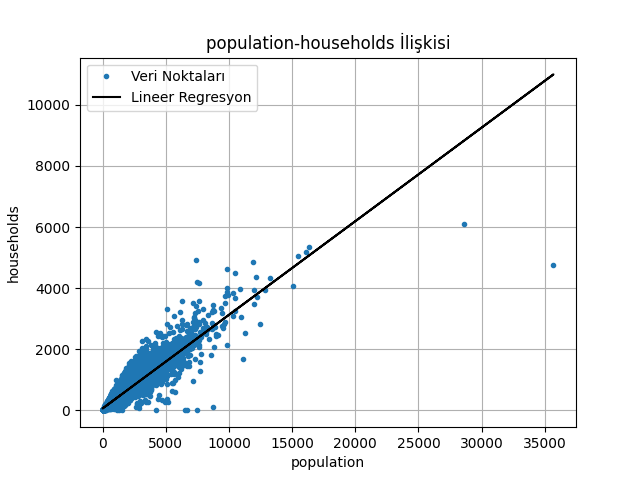
g)“ population” ve “households” Grafiği

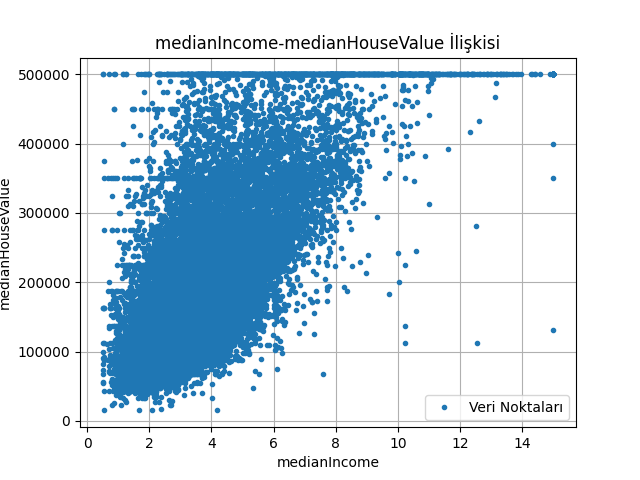
çizelge içeren bir resim

Açıklama otomatik olarak oluşturuldu

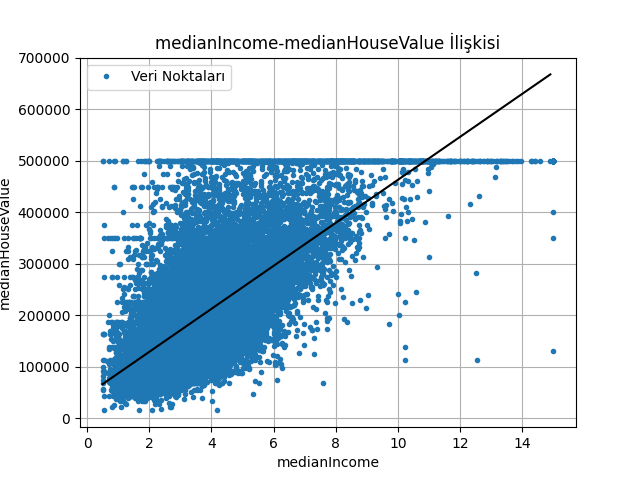
g.1)“ population” ve “households” Grafiği ve

Slope-Itercept ile Oluşturulmuş Doğru



h)“medianIncome” ve “medianHouseValue” Grafiği 

h.1)“medianIncome” ve “medianHouseValue”Grafiği ve

Slope-Itercept ile Oluşturulmuş Doğru

1. Linear Regrasyon yöntemi uygulanacak,
   * train-test datası olarak ayrı ayrı hesaplanacak
   * Plotting Grafiği oluşturulacak
   * Linear R2 değeri hesaplanacak

* Tüm sonuçlar değerlendirilerek yorum bölümüne yazılacak.

a)“longitude” ve “latitude”

Train-Test Datası Terminal Görüntüsü

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Linear R2 değeri

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Predict değeri

metin içeren bir resim

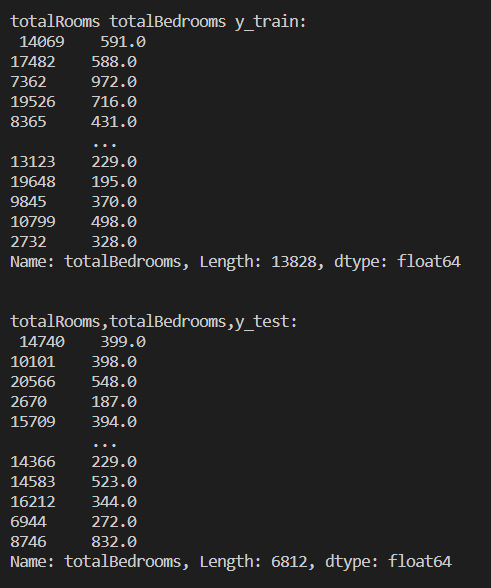
Açıklama otomatik olarak oluşturuldu

çizelge içeren bir resim

Açıklama otomatik olarak oluşturulduPlotting Grafiği

b)“totalRooms” ve “totalBedrooms”

Train-Test Datası Terminal Görüntüsü

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Linear R2 değeri

metin içeren bir resim

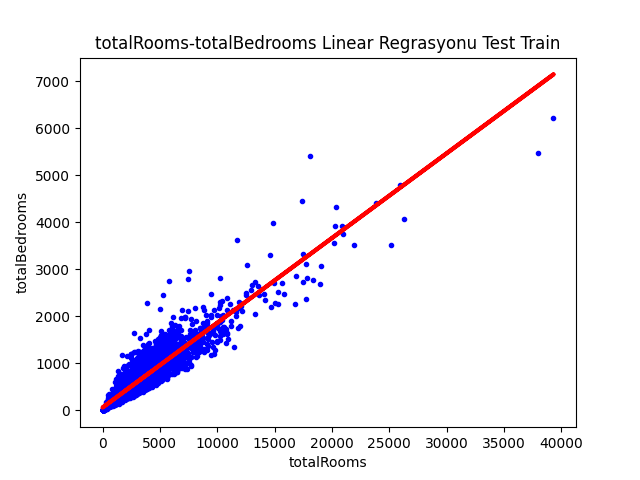
Açıklama otomatik olarak oluşturuldu

Predict değeri

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Plotting Grafiği



c)“totalRooms” ve “population”

Train-Test Datası Terminal Görüntüsü

metin içeren bir resim

Açıklama otomatik olarak oluşturuldumetin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Linear R2 değeri

metin içeren bir resim

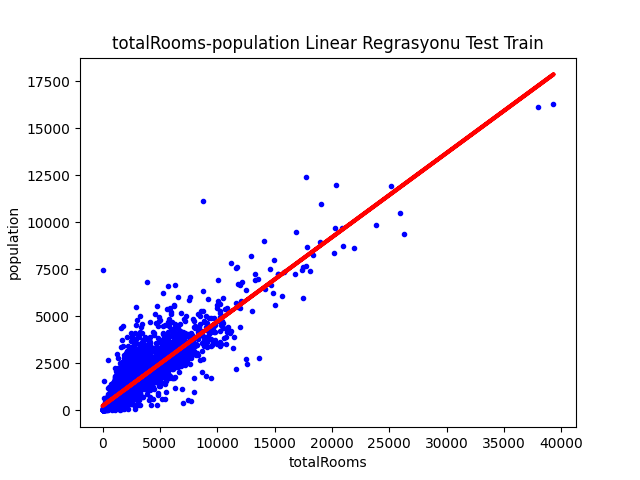
Açıklama otomatik olarak oluşturuldu

Predict değeri

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Plotting Grafiği



d)“totalRooms” ve “households”

Train-Test Datası Terminal Görüntüsü

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Linear R2 değeri

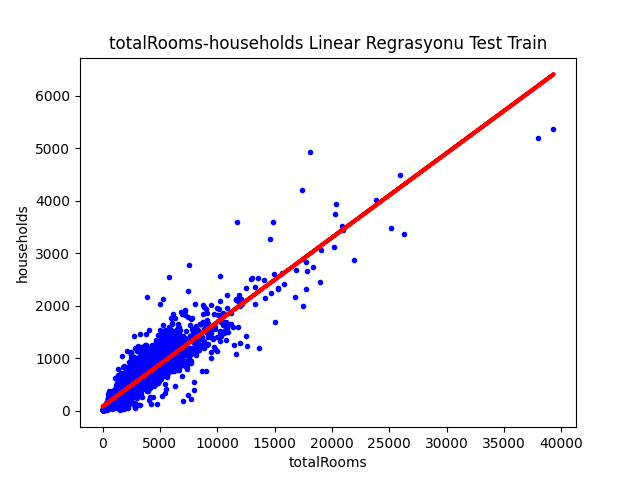
metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Predict değeri

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Plotting Grafiği

e)“totalBedrooms” ve “population”

Train-Test Datası Terminal Görüntüsü

metin içeren bir resim

Açıklama otomatik olarak oluşturuldumetin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Linear R2 değeri

metin içeren bir resim

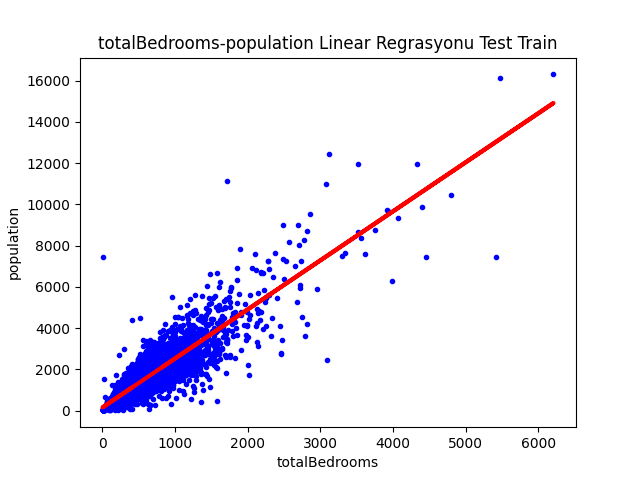
Açıklama otomatik olarak oluşturuldu

Predict değeri

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Plotting Grafiği



f)“totalBedrooms” ve “households”

Train-Test Datası Terminal Görüntüsü

metin içeren bir resim

Açıklama otomatik olarak oluşturuldumetin içeren bir resim

Açıklama otomatik olarak oluşturuldu

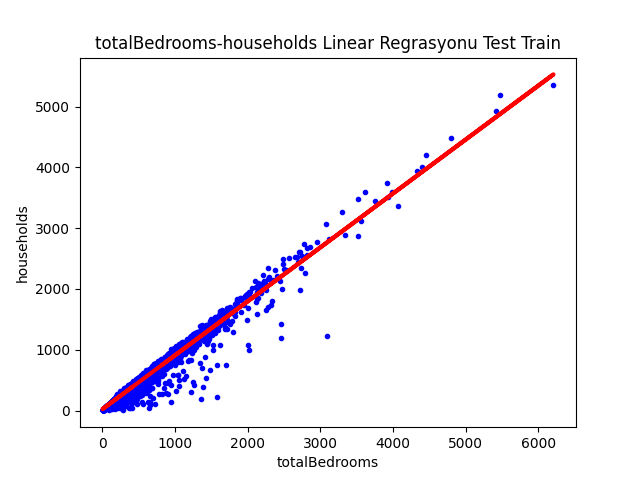
Linear R2 değeri



Predict değeri



Plotting Grafiği



g)“ population” ve “households”

Train-Test Datası Terminal Görüntüsü

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Linear R2 değeri



Predict değeri

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Plotting Grafiği



e)“medianIncome” ve “medianHouseValue”

Train-Test Datası Terminal Görüntüsü

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Linear R2 değeri

metin içeren bir resim

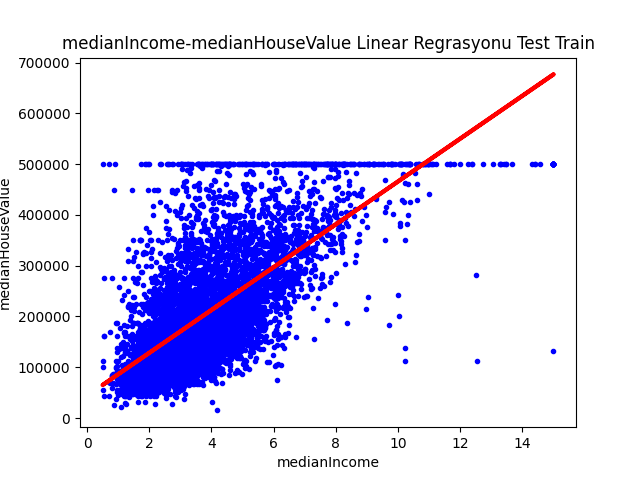
Açıklama otomatik olarak oluşturuldu

Predict değeri

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Plotting Grafiği



Yorumlar

**Veri Seti Üzerindeki Temizleme İşleme Yapılanlar ve Yorumlar**

1.soru: Verilen “datamining.csv” veri seti üzerinde, gerekli temizleme ve veri türleri üzerinde değişiklikler yapılacak

(yapılan her işlem anlatılacak);

Gerekli temizleme işlemleri için pandas kütüphanesinden yaralandım sonra csv dosyamı read.csv ile okuttum öncelikle hiçbir işlem yapmadan rowsayisi = df.shape[0] ve columsayisi = df.shape[1] ile kaç satıra ve sütuna sahip olduğumu öğrendiğim bu sayede yaptığım temizleme işlemleri ile datamın ne kadarı ile işlem yapacağımı öğrenebileceğim. Bu işlemlerden sonra öncelikle dropna(inplace=True) komutu ile boş veri bulunan satırları silerek temizleme işlemime başladım sonradan tekrardan satır kontrolü yaptığımda hiçbir değişim olmadığını gördüm bu datamın tüm verilerimde boş bir veri bulunmadığı anlamına geliyor. Sonra drop\_dublicates(inplace=True) komutu ile aynı olan satırları temizlemek istedim sonra tekrardan satır sayımı kontrol ettim ve satır sayımında hiçbir azalma olmadığını gördüm bu birbirini tekrar eden hiçbir verimin olmadığının göstergesidir. Böylelikle son işlemime geçtim. Df.info() komutu ile data tiplerini control ettim hepsinin float64 olduğunu gördüm. Burda herhangi bir değişiklik yapmaya gerek kalmadı çünkü ben sayısal verilerim olması gereken veri tipinde bu şekilde artık veri analizlerime başlayabilirim. Verimin bu işlemlerden sonra çıkaracağım sonuç %100 bir şekilde hiçbir veri kaybı yaşamadan analizlerime başlayabileceğimdir ve bu analizlerimin doğruluğunu olumlu yönde etkileyecektir.

**Veri Seti Özet Analizi Yorum**

2.soru: Veri setinin özet analizi çıkarılacak, ve bu özet ekran görüntüsü ödevde gösterilecek,

Burda mean komutu ile sütunların ortalamalarını gördüm bu şekilde birkaç tane çıkarıma ulaşabilirim öncelikle enlem ve boylam üzerine konuşabilirm ortalamasında gördüğüm değerler ile ve max min değerlerini google haritalarda arattığımda verilerimin Amerika Birleşik Devletleri’nin Kalifornia Şehrinde bulunan evler için olduğu bilgisine sahip oldum. Max değerleri ile min değerlerine baktığımda ise evlerin konumları daha çok dağu batı doğrultusu üzerinde olduğunu gözlemledim. Son olarak da standart sapmanın (std) 2 çıkması üzerine de bu veri adreslerinin birbirine çok yakın mahallerde olduğu sonucuna ulaştım.

HousingMedianAge sütunu ile ilgili yaptığım analizde ortalama olarak binaların 28 yaşında olduğunu bilgisine sahip oldum. Standart sapması 12 olması binaların yaşları arasında farkın olduğunu ve binaların orta yaşlı olduğunu olduğunu ama yaş dağılımının çok olduğunu söyleyebilirim.

TotalRooms burda ortalamanın 2635 olduğunu görüyorum bu da çok fazla odalı bölgelerin olduğu gibi daha az odalı mahallelerin olduğunu ve bu konudaki dağılımın fazla olmadığını söyleyebilirim. Çeşitliliğin az olmasını düşündüren etkenler std 2181 olması ve ortalamaya çok yakın olmasıdır. Min ve max değerler arasında büyük farklar vardır ama ortalamanın çok üstünde ve altındaki değerlerin az olduğunu söyleyebiliriz.

TotalBedrooms sütunu için yatak odası sayısı ortalaması 537 ve std 421 bu bize aşırı bir sayı çeşitliliği olmadığını söyler. Burda şu yorumu yapabiliriz yatak odası min değeri 1 dir ve yatak odası olmayan bir yerin olmadığını söyleyebiliriz.

Population kısmı için çok kalabalık mahallelerde bulunmaktadır ama ortalama 1425 insan yaşamaktadır ve max değeri 35682 olduğu için aşırı büyük mahallelerin nadir olduğunu ortalama kalabalıkta bir nüfusa sahip olduğunu söyleyebiliriz.

Households bina halkı sayısı ortalama 499 kişidir ve std 382 genel olarak 500 kişilik binalar yapılmıştır ve min 1 kişinin yaşadığı binalar da bulunmaktadır. Max 6082 kişilik olan bina ve bu kadar fazla sayıda insanın

yaşadığı binalar azdır.

Median Income bina halkı geliri ortalama 3.870 bin dolardır bu da orta derece bir gelir düzeyini temsil

eder. Zengin ve fazla gelir düzeyine sahip insanların da bulunduğu bu mahallede zenginler azınlıktadır çünkü ortalama değer min değerine yakındır

medianHauseValue binaların değerleri ile ilgilidir ortalama 206 855bin dolara Kalifornia’dan ev

alabilirsiniz. Ev fiyatları birbirine çok yakındır çünkü std değeri ortalamaya baya yakındır. 14bin dolara bile ev bulunmaktadır ama ucuz ev sayısı fazla değildir. Burdaki en pahalı ev 500 000bin dolardır.

**Korelasyon Analizi**

Correlasyonu (Kendall, Spearman ve Pearson yöntemleri olarak tek tek ) hesaplanacak,

* Correlasyon ekran matrix görüntüsü (correlasyon sonucu) aynen yazılacak,
* Matrixte anlamlı ilişkisi olan veriler üzerinde sonuçlar değerlendirilecek ve yorum bölümüne yazılacak.

**İyi Korelasyon:**

İyi korelasyon için bulduğumuz değerin “-1” ya da “1” e çok yakın olmasına bağlıdır ve “0.6” ya da “-0.6” korelasyon değerlerinin üstündeki ikililere iyi bir korelasyon diyebiliriz

**“**longitude” X “latitude” = -0.924664 ,bu çok iyi bir korelasyon ve bu enlem ile boylam arasında ters orantı olduğunu gösterir. Bunun yorumu boylam artarsa enlem azalacağı demektir.

“totalRooms” X “totalBedrooms”= 0.929893,bu çok iyi bir korelasyondur ve bu toplam oda sayısı ile

toplam yatak odası sayısı arasında doğru orantı olduğunu gösterir. Bunun yorumu toplam oda sayısı artarsa

toplam yatak odası sayısı da artacağını demektir.

“totalRooms” X “population” =0.857126, bu çok iyi bir korelasyondur ve bu toplam oda sayısı ile bina çevresi mahalle nüfusu sayısı arasında doğru orantı olduğunu gösterir. Bunun yorumu toplam oda sayısı artarsa toplam bina çevresi mahalle nüfusu da artacağını demektir.

“totalRooms” X “households”=0.918484 , bu çok iyi bir korelasyondur ve bu toplam oda sayısı ile bina nüfusu sayısı arasında doğru orantı olduğunu gösterir. Bunun yorumu toplam oda sayısı artarsa toplam bina

çevresi mahalle nüfusu da artacağını demektir.

“totalBedrooms” X “population”= 0.878026, bu çok iyi bir korelasyondur ve bu toplam yatak sayısı ile mahalle nüfusu arasında doğru orantı vardır. Bunun yorumu toplam yatak sayısı artarsa mahalle nüfusu da artacağını demektir.

“totalBedrooms” X “households”= 0.979829 neredeyse 1’e çok yakın yani aşırı iyi bir korelasyondur ve bu toplam yatak yatak sayısı ile bina nüfusu sayısı doğru orantılıdır. Bunun yorumu toplam yatak sayısı artarsa binadaki nüfus oranı da artacağını demektir.

“Population” X “households”= 0.907222 bu çoj iyi bir korelasyon ve bu mahalle nüfusu ile bina nüfusu

sayısı arasında doğru orantı vardır. Bunun yorumu bina nüfusu artarsa mahalle nüfusu da artacağını demektir.

“medianIncome” X “medianHouseValue”= 0.688075, bu iyi denebilecek bir korelasyondur ve binada yaşayanların gelir durumu ile binanın satış değeri arasında doğru orantılı olduğunu gösterir. Bunun yorumu

binada yaşayanların gelir durumunun artmasıyla binanın satış değerinin de artacağını demektir.

**Kötü Korelasyon:**

Kötü korelasyon için bulduğumuz değerin “-1” ya da “1” e çok uzak olmasına bağlıdır ve “0.6” ya da “-0.6” korelasyon değerlerinin altındaki ikililere kötü bir korelasyon diyebiliriz. Bulduğumuz iyi korelasyonlar haricindeki tüm ikililer kötü korelasyona sahiptir.

( Uyarı: **“**longitude”X**“**longitude”=1.0000 , bu sonuçlar hariçtir kendi kendine yapılan korelasyonlar zaten sonucunda “1” çıkar.)

Örnek vermek gerekirse **“**medianIncome” X “population”= 0.004834, bu “1”den çok uzak bir değerdir ve binada yaşayanların gelir durumu ile bina çevresi mahalle nüfusu arasında hiçbir ilişki olmadığı anlamına gelir.

**Linear Regrasyon Analizi Yorumu**

Correlasyonda anlamlı çıkan her ikili veri türü için;

4) Linear Regrasyon slope-intercept değerleri tespit edilecek,

* + Plotting olarak Grafiği Çizilecek
  + f(x) fonksiyonu bulunacak

Yaptığım korelasyon yönteminden birbiri ile ilişki 0.6’dan büyük olan ikilileri slope\_intercept değerlerini aldım bu aldığım değerler ile bir f(x) fonksiyonu oluşturdum oluşturmuş olduğum f(x) fonksiyonu bana Linear regrasyon elde etmemi sağladı. Bulduğum ikililerin plotting olarak grafiklerini oluşturdum bununla beraber bir de ikililerin grafiklerinin içine f(x) fonksiyonun grafiğini ekleyerek bir grafik oluşturdum ve bulduğum grafik çok güzel bir şekilde birbirinin üstünü örtüyordu sadece bir ikili hariç (“medianIncome” ve “medianHouseValue”) ikilisi grafiğe bakınca biraz uyumsuz olduğunu düşündüm zaten corr değeri de 1 fazla yakın değildi. Test Train grafiklerinin R2 sonuçlarında daha iyi göreceğiz her şeyi.

1. Linear Regrasyon yöntemi uygulanacak,
   * train-test datası olarak ayrı ayrı hesaplanacak
   * Plotting Grafiği oluşturulacak
   * Linear R2 değeri hesaplanacak

* Tüm sonuçlar değerlendirilerek yorum bölümüne yazılacak.

Burda ise aldığım ikililerin verilerinin bir kısmını alıp modeli eğitmek için kullandım ve bir kısmı ile bunu test ettim ve

grafiklerin görüntülerini aldım. Bunların sonucunu alıp plotting grafiğine döktüm ve elde ettiğim grafikler neredeyse birebir uyuyordu şüphelendiğim (“medianIncome” ve “medianHouseValue”) ikilisinin grafiğinden sonra Linear R2 değerini hesapladım

sonuçlar

longitude-latitude R2 değeri: 0.8530734992629649 bu değer 1’e yakın ve doğruluğu gayet yeterli

totalRooms-totalBedrooms R2 değeri: 0.8565628206396894 bu değer 1’e yakın ve doğruluğu yeterli

totalRooms-population R2 değeri: 0.7512234851734285 bu değer için biraz şüpheli bakabiliriz ama hala yeterince iyi

totalRooms-households R2 değeri: 0.8396946924819624 bu değer 1’e yakın ve doğruluğu yeterli

totalBedrooms-population R2 değeri: 0.7845703640993615 bu değer 1’e yakın diyebiliriz doğruluğu gayet iyi

totalBedrooms-households R2 değeri: 0.9643243067169097 bu değer 1’e çok yakın ve neredeyse birebir doğruluğa sahip

population-households R2 değeri: 0.8323149304375397 bu değer 1’e yakın ve doğruluğu yeterli

medianIncome-medianHouseValue R2 değeri: 0.45998761856422554 bu değer 1’e çok uzak ve bununla bir iş yapmak bizi çok yanıltıcı sonuçlarla karşı karşıya bırakabilir grafiğinden de şüphelendiğim gibi bizim için yeterli değil.

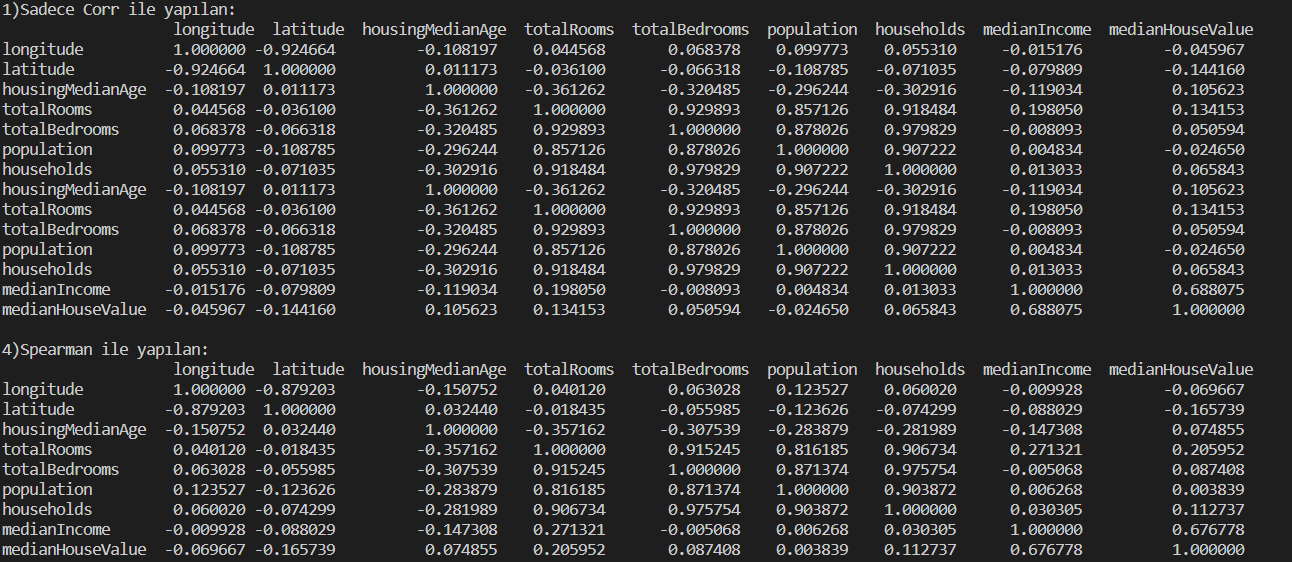
**Yaptığım Tespitler**

Sadece corr() komutunu kullandığımda elde ettiğim korelasyon pearson korelasyonu ile aynı çıktıyı elde ettim hiçbir korelasyon belirtmediğimde pearson ile çalıştığını tespit ettim. Ödevin içinde ikisinin de görselleri var.

Plt.plot() ile grafik için datamdan verileri yerleştirirken hiçbir bilgi girmediğimde color=”blue”,linestyle=”solid” olmaktadır. Noktasal bir grafik istediğimde linestyle kendim “none” yapmam gerekmektedir.

**Aldığım Bazı Hatalar**

**Korelasyonların print ederken uzun bir süre rastgele bir biçimde bazı korelasyonların çıktısını vermedi ve programı açıp katapmam gerekti. Birkaç keren fazla print alınca da aynı hata ile karşılaştım.**



**Burda tüm korelasyonları print ettirdim ama sadece bu ikisini verdi 2 ve 3 numarayı kendisi atladı.**

**Korelasyonda aldığım hatanın aynısı y\_test ve y\_train değerlerini print ederken de aldım sebebini bilmemekle beraber bilgisayarımın yetersiz olaması olabilir.**