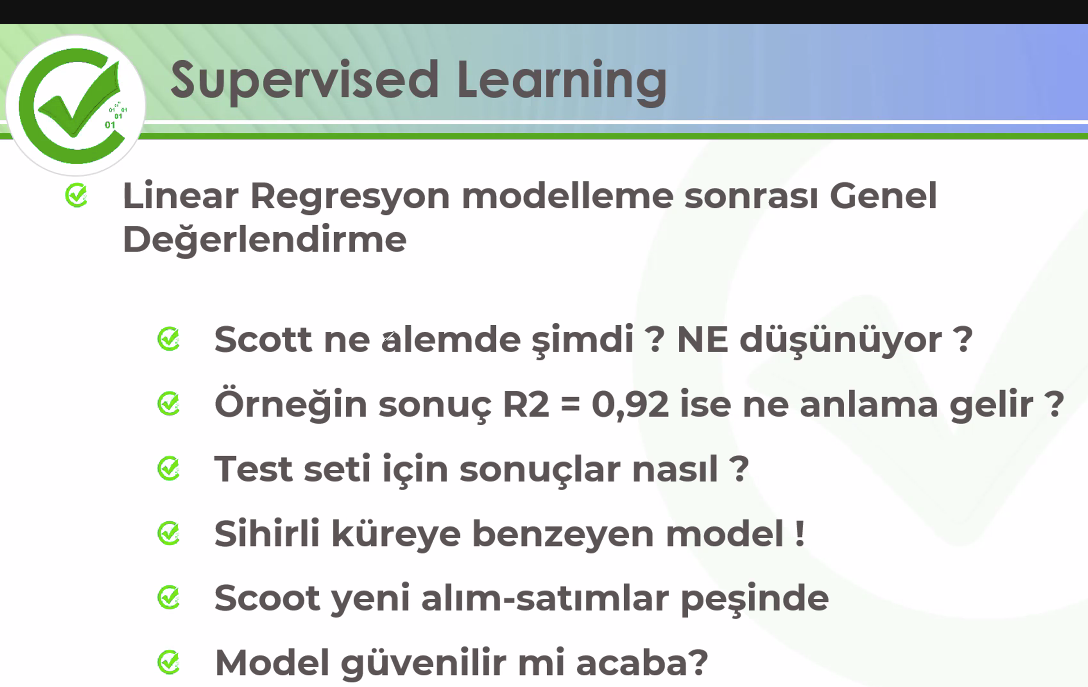
**ML 05.07.2023**

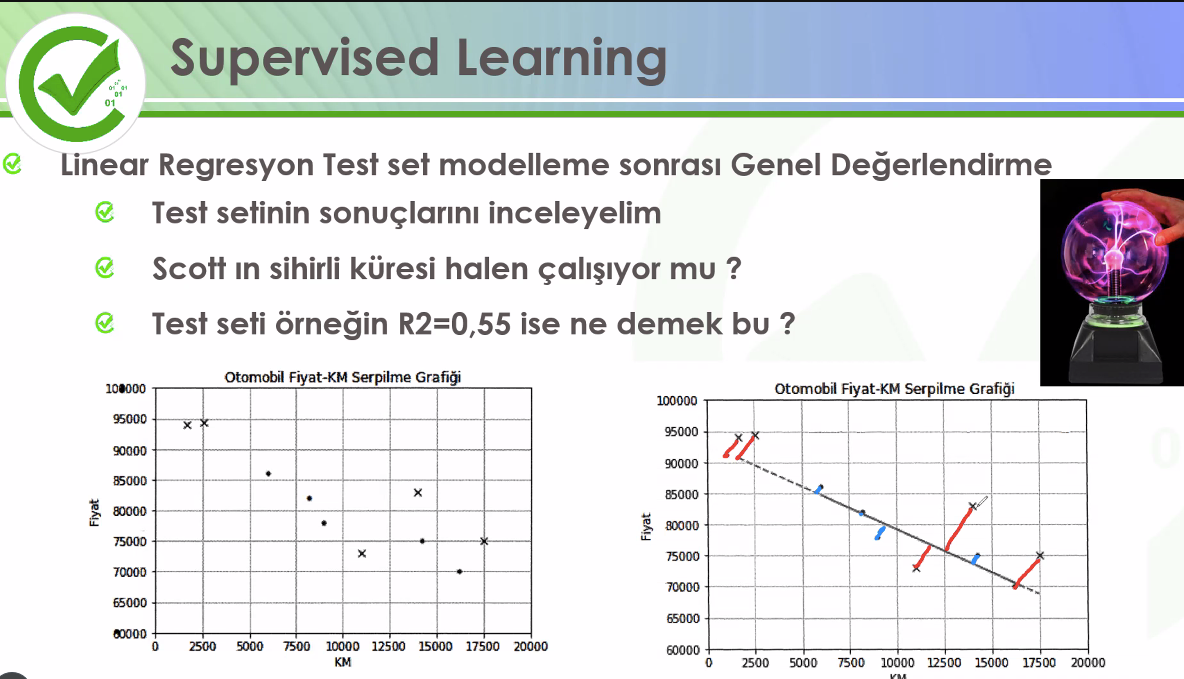


R2 ‘nin 0.92 olmasi demek--> Araclarin km degerlerinin 2. el arac fiyatlarindaki degisimin %92sini acikladigini ifade eder

ML’de asil mesele, train verisi uzerinden olusturulan modelin yeni datalari yani test datalarini tahmin edip edemedigine odaklanmaktir.

Train datasi uzerinde basarili bir tahminde bulunuyor ama test datasi icin

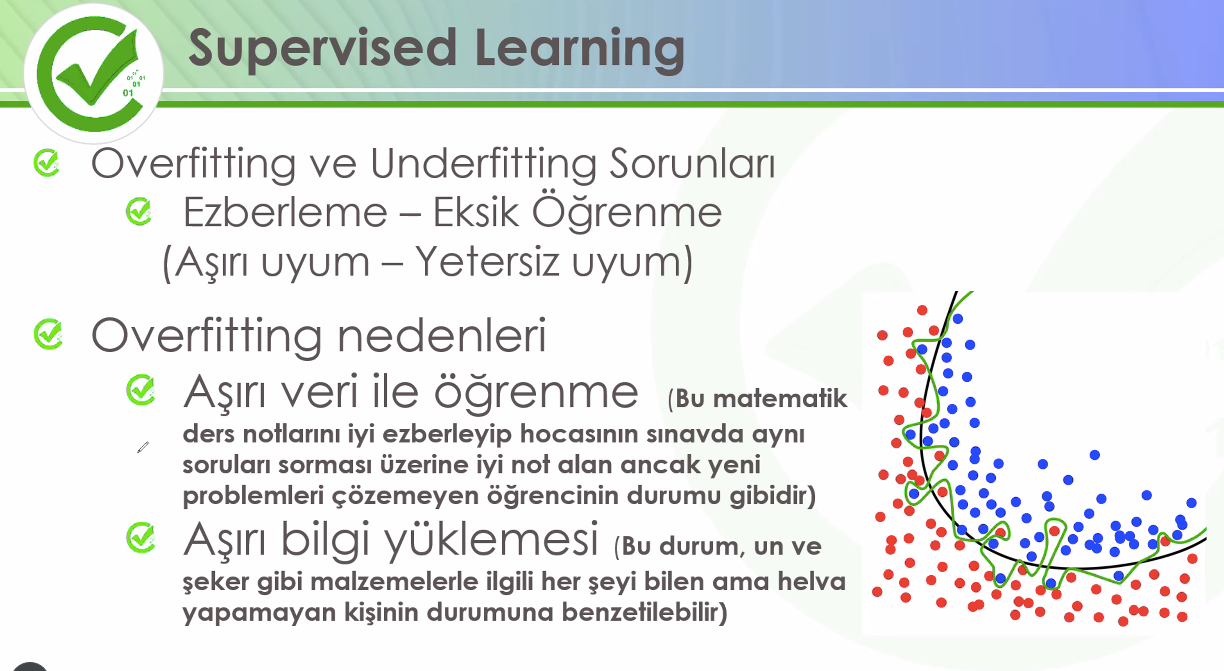
R2’nin train dataseti icin 0.92 cikmasinin onemi cok da muhim degildir. Onemli olan ML’de test datasinin dogru tahmin edilebilmesidir.



X isareti ile gorulenler test datasinin oldugu yerler.

Train set uzerinden uretilen simple linear regresyon mmodeli, yeni durumlari olcmek icin ne kadar guvenli?bu regresyon cizgisini test datasetinde baktigimizda kirmizi cizgilerle gosterilen residuallarin oldugunu goruyoruz.

Bilgi kaybinin yuksek oldugu anlamina gelir, yani maliyet yukselmis demektir. Train set uzerinden olusturgun basarili model, test datasetinde ayni basariyi gostermiyor.



Overfitting--> Model arametrelerinin train setinde cok iyi ama test datasetinde kotu sonuclar vermesi. Genelde iyasada karsilasilan durum budur. Asiri ezberleme. Asiri veri ile ogrenmekten kaynaklanabilir.

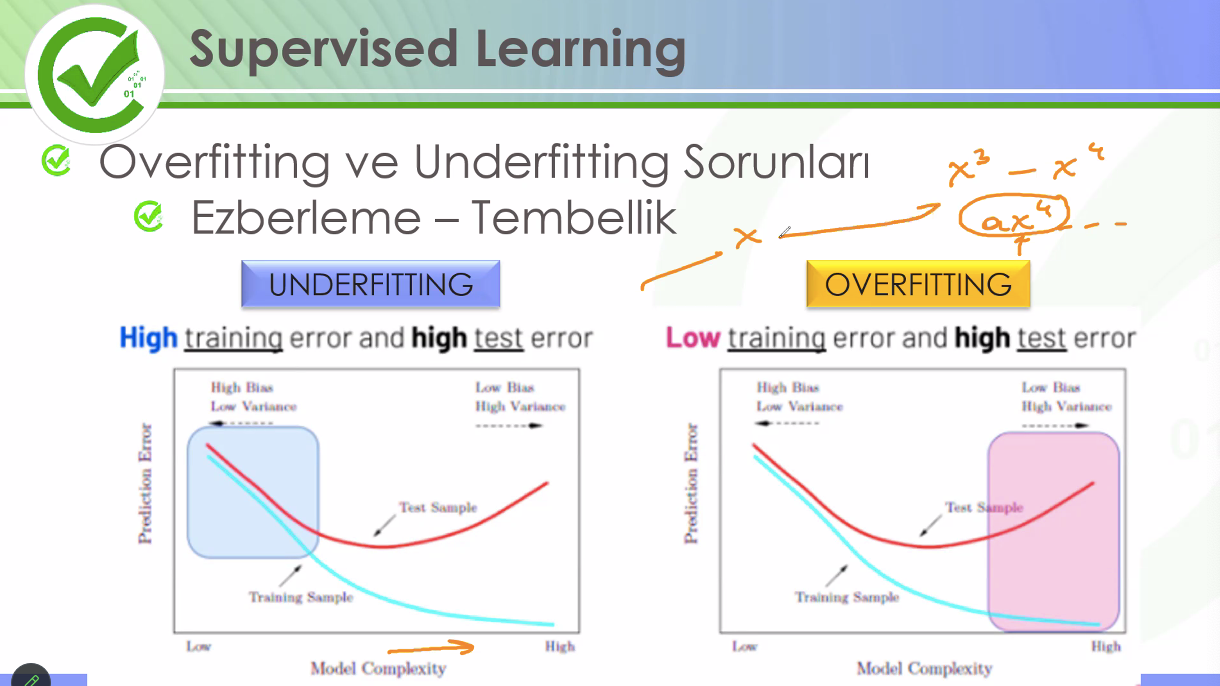
Underfitting--> eksik ogrenme ya da ogrenememe. Dogal olarak traini bile dogru tahmin edemeyen bir model testi de dogru tahmin edemeyecek demektir.

Datasetindeki veri seti dusuk olabilir mesela bir nedeni. Veri sayisini artirarak underfitting sorunu asilabilir.

Modelin kompleksligini artirdiginizda sorun cozulebilir.

**Overfitting neden olur**

1. Cok fazla veri ile ogrenmesinden kaynaklanabilir yani. Buna gore model olusturuyor ve yeni durumlara karsi esneyip uyarlanamiyor.
2. Feature fazla olursa, ezbere dogru gider.



**Bias-variance trade-off**

Modelin kompleksligi fazla ise--> mesela x degil de x3 ile x4 ile iliskili ise y.

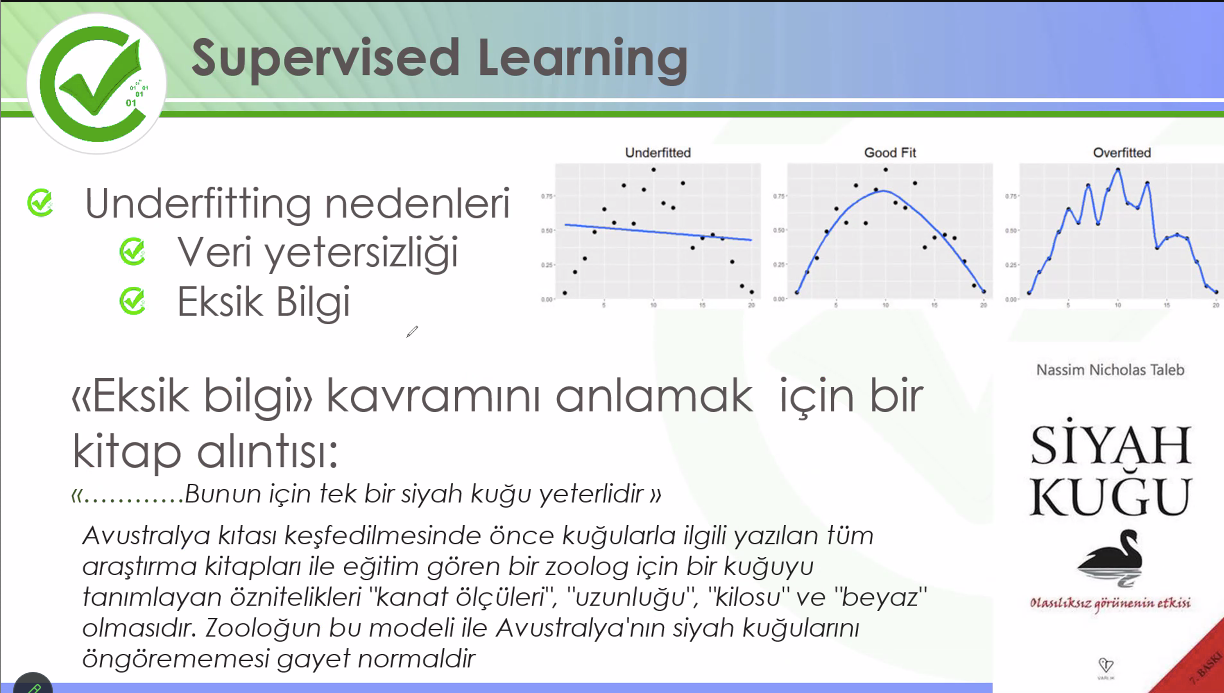
Bir yerden sonra test setin errorlari ile trainin errorlari aradaki makas aciliyor. Bunu gordugumuzde soyleyecegimiz sey-->overfitting

Mavi ile gosterilen alan--> underfitting alanidir. Burda eksik ogrenme vardir. Mesela aracin fiyatini sadece km belirlemez.

Regularizasyon

Cross-validation

Complexity’



**Underfitting neden olur**

1. Olayla ilgisi olmayan feature’lar olabilir.

2. Veri sayisi eksik olabilir

3. Modelin kompleksligi yeterli olmayabilir.

**ML\_Notebook\_1-Simple Linear Regression**

**Supervised**--> meselemiz prediction’dir. En buyuk ayrildigi nokta da budur.

**Unsupervised** --> target yoktur, amaci veri setindeki desenleri patternlari kesfetmektir. Egitim surecinde geri bildirim saglanmaz, cunku targetimiz yok. Analiz soz konusudur. Kumeleme algoritmasi yapiyorsun mesela. Subjektifir. Mesela 3 kume cikti, ama bunu 5’e bolebilirsin istersen de.

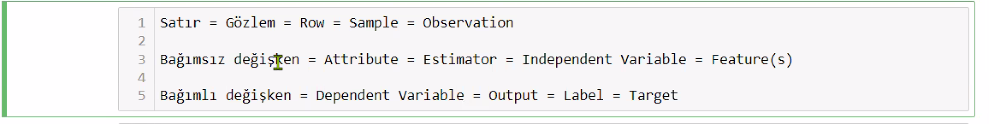
Datasetini okumazsa--> df = pd.read\_csv("Advertising.csv", encoding="utf-8")

Or

encoding="ISO-8859-1"

Or

encoding='latin1'



**Satır = Gözlem = Row = Sample = Observation**

**Bağımsız değişken = Attribute = Estimator = Independent Variable = Feature(s)**

**Bağımlı değişken = Dependent Variable = Output = Label = Target**

**df.describe** --> numeric datalarin istatiksel ozetini verir.

Standard sapma-->Ortalamaya gore sapma miktarini verir.

Count-->satir sayisini verir.

%50 bize ayni zamanda-->median degerini verir.

Max deger ile %75 arasinda cok fark olmasi--> outlier hakkinda bilgi verebilir.

Min deger ile %25 arasinda cok fark olmasi da ayni sekilde

**Simple linear regression**

y = ax + b şeklinde bir lineer regresyon modeli oluşturmak için ( 1 tane y değeri, 1 tane x değeri )  böyle yaptık

Bazi sartlarin saglanmasi gerekir.

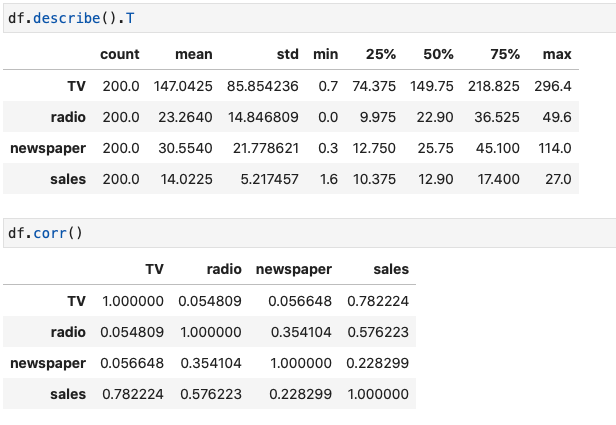
1. Bagimli degiskenin bagimsiz degiskenle **dogrusal iliski olmasi** gerekir. Yani korelasyon olmasi gerekir degiskenler arasinda. Korelasyon bir iliski analizidir. Regresyon ise etki analizidir. Eger korelasyon yok ise zaten regresyondan bahsedemezsin. Birinin bir birim artisindan otekinin etkilenmesi icin oncelikle aralarinda korelasyon olmali.
2. bagimsiz degiskenler arasinda cok yuksek korelasyon olmamali. Iki degisken arasinda ben mesela 0.8 korelasyon goruyorsam, bagimsiz degiskenler kendi aralarinda bagiml demektir. -->**multicollinearity** sorunu olmalali yani.
3. Residuallarin pattern gostermemesi gerekir. Pattern gosterirse homojenlikten bahsedemeyiz.

**Korelasyon**

1. Yon
2. Kuvvet ve guc

Korelasyon katsayisi--> -1 ile +1 arasinda yer alir. 0’dan her iki yone ilerledikce artar.

- yonde olmasli negatif korelasyon, ters yonlu iliski demektir.



#diagonal cizginin altinda ve ustunde kalan degerler dogal olarak aynidir.

#sales ile TV arasindaki yuksek korelasyon aralarinda yuksek iliski oldugunu gosterir.

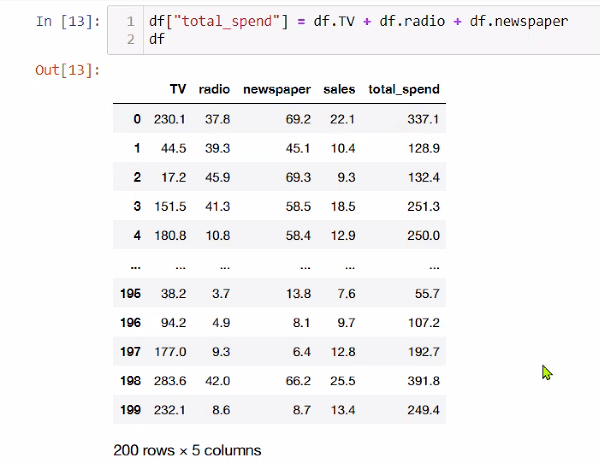
**features= TV, sales, newspaper**

**target= sales**

**Scatter Plot Yorumlamasi**

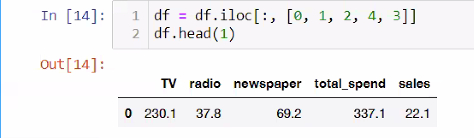






Sales’i reklam harcamalari uzerinden degerlendirmek istiyorum. Bunun icin tv, radio ve newspapaer’i toplar yeni bir feature olusturabilirim.

**df["total\_spend"] = df.TV + df.radio + df.newspaper**



Sales’in yani target’in en sonda olmasini istersem yerini degistirirm:

Sns.pairplot ile birbiri ile korelasyonlarini goruyorum.

Target ile feature’larim arasinda lienar bir korelasyon var mi ona bakacagim.

Total\_spend olarak olusturdugum yeni feature ile sales arasinda 0,86 korelasyon var.

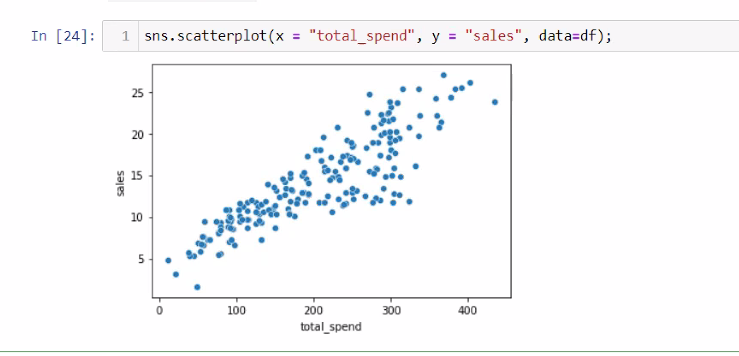
Feature’lar buyuk X target ise kucuk y ile gosterilecek.

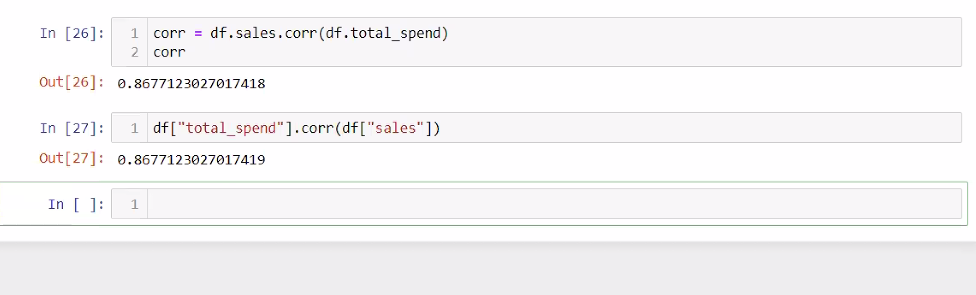
Normalizasyon, standardizasyon gibi islemleri target’ta yapmayacagiz.

O yuzden ilk islememiz target ile diger verileri ayirmak olacak.

O yuzden targeti droplayacagiz.

Total\_spend ile sales arasinda yuksek korelasyonu bir de scatterplot ile gorelim:





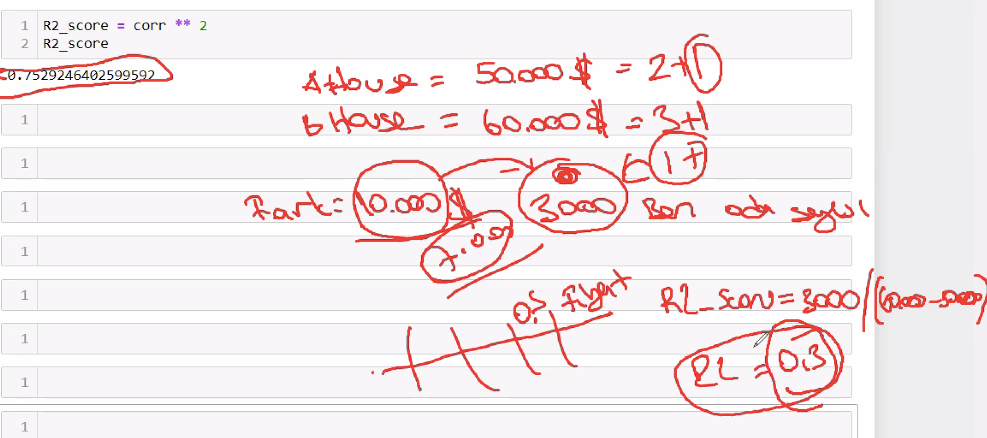
**Coefficant of Determination(R2)**

Korelasyonun karesi=R2

**R2\_score=corr\*\*2**

R2=0,75 ciktiysa total reklam harcamalari, salesideki degisimlerin %75 ini acikliyor demektir.

Bagimsiz degiskenin bagimli degiskeni ne kadarini aciklayabildigini gosterir.



Mesela evin fiyati oda sayisi 1 artinca 10 bin dolar artmis gorunuyor. Ama bu 10bin dolarlik artisin tek sebebi oda sayisi olamaz. Oda sayisi o fiyat artisinin 3000 dolarini acikliyor diyelim, o zaman R2=0.3 demektir. Ama kalan 7000 dolarlik kisim da cesitli diger degiskenlerle aciklanabiliyor aslinda. Bizim elimizde toplu tasimaya yakinlik, kacinci kat oldugu gibi ek bilgiler de olsaydi o zaman bu feature’lar mesela fiyat degisiminin %70’ini aciklayacakti mesela.

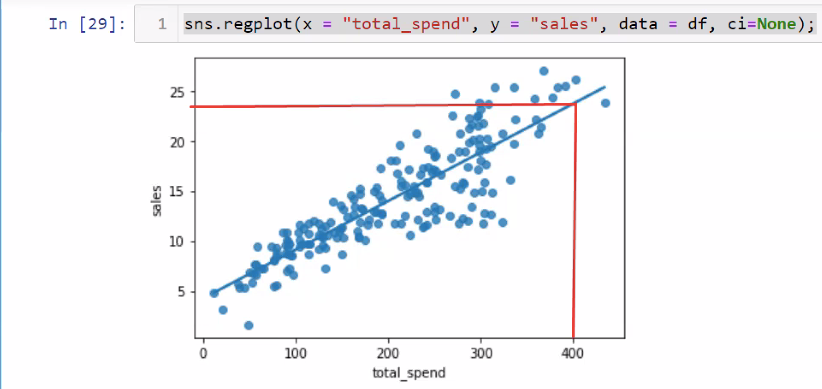
Daha fazla feature olursa o kadar iyi sonuc cikar. Ama amacimiz en az feature ile maksimum sonucu elde etmektir. Mesela 10 feature ile %96’sini tahmin ediyorsan, 100 feature ile %97’sini tahmin ediyorsan, makbulu 10 feature ile %96’sini tahmin etmektir.

**Linear Regresyon**

Best fit--> en optimum cizgi

Linear regresyonun mantiginda best fit line’in dogrusal olmasi.

sns.regplot(x = "total\_spend", y = "sales", data = df, ci=None);



Regression line’i, butun noktalara en optimum uzakliktaki noktadan gecirdi.

**Least square error-**-> Hatalari topluyoruz, bazilari best line’in altinda ve error negatif.

karelerini almasi, negatif degerleri de artiya cevirmesini sagliyor.

Bu cizginin ustunde ve altinda kalan degerlerin toplami 0’dir best line icin.

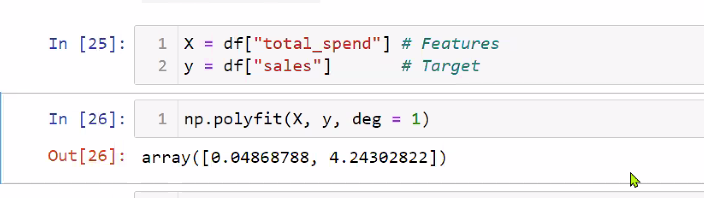
**y = b1 \* X +b0**

**y = Target = Bağımlı değişken = dependent variable**

**b1 = Katsayı/eğim, slope 0< b1 < 1 and -1 < b1 < 0**

**X = Feature(s), bağımsız değişken, independent variable**

**b0 = Intercept, otonom parametre, başlangıç parametresi -∞ < 0 < +∞**



Bastan x’i ve y’i ayiriyoruz.

Feature’lar buyuk X target ise kucuk y ile gosterilecek.

Ilk cikan deger egim,

ikincisi ise intercept--> yani x 0 iken y’nin aldigi deger.

b0=0.04--> slope

b1=4.24-->intercept

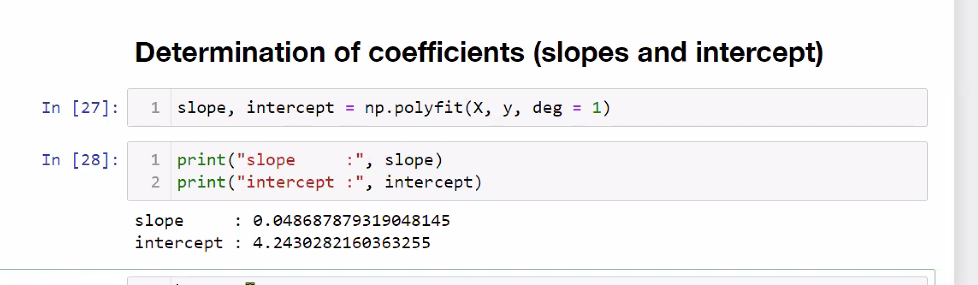
Determination of coefficients (slopes and intercept)

y = b0x + b1

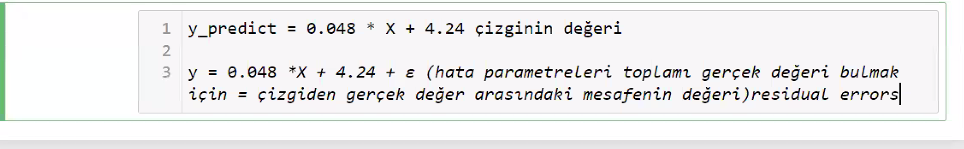
b1 = Katsayı/eğim, slope 0< b1 < 1 and -1 < b1 < 0

Degree=1 demek bizim icin linear regresyon modeli demek oluyor.

**Determination of coefficients (slopes and intercept)**



Bu sekilde de hesaplatabiliyoruz.



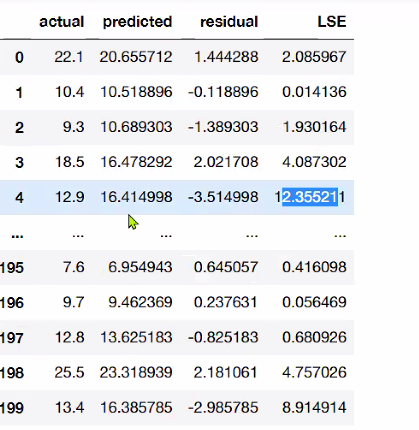
Tabloda

Gercek degeri gormek istiyorum

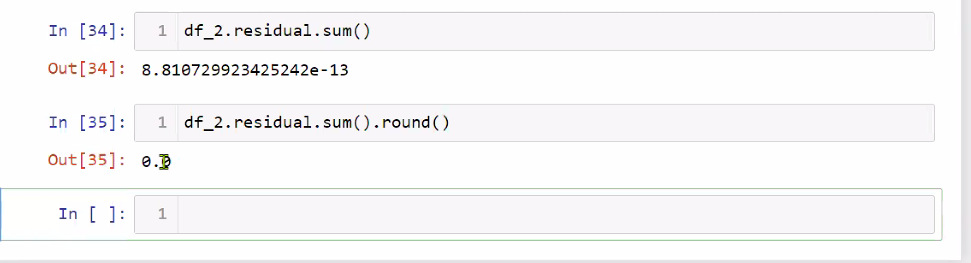
Tahmin ettigim degeri girmek istiyorum

Residual’i gormek istiyorum

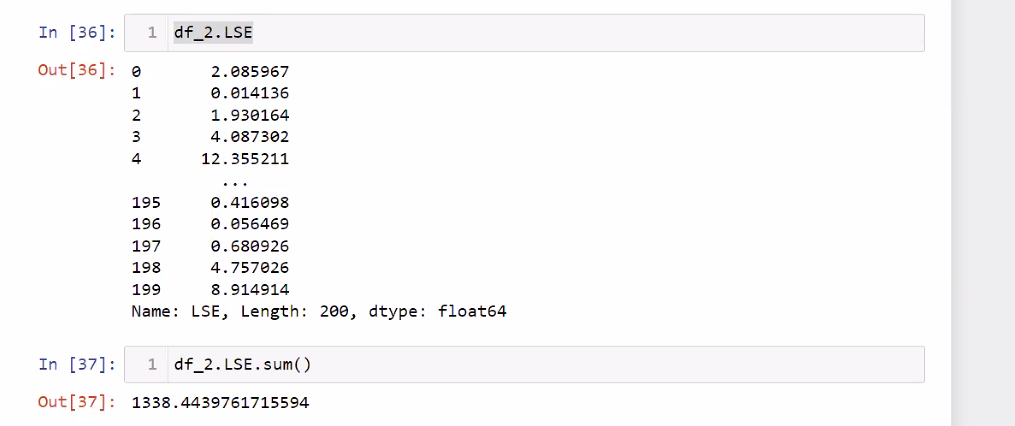
LSE’i gormek istiyorum



Modelimin nasil calistigini burdaki hatalarin toplamina gore degerlendirecegim.



Hatalari topladigimda 0 cikiyor. O yuzden biz LSE kullaniyoruz. Karelerini alarak - olanlarin da hata degerlerini hesaba katmis oluyoruz.



Best line--> target ve feature’dan olusan data ile best line cizgisi cekiyoruz.

Neden LSE kullaniyoruz-->Karelerini alarak - olanlarin da hata degerlerini hesaba katmis oluyoruz. Yoksa hatalari toplayinca 0 cikar.

**Prediction with simple linear regression**

Task-->0 ile 500 arasında 100 adet eşit aralıklarla bölünmüş bir potential\_spend datası oluşturalım ve bu dataya yönelik predicton alalım