HİKAYE

Problem: Müşteri Harcama Tahmini

Bir pazarlama şirketi, müşteri harcama davranışlarını analiz ederek müşterilerinin harcama miktarlarını tahmin etmek istemektedir. Şirket, müşterilerin yaşadığı şehir nüfusu, geçmiş harcama miktarı, internet reklamlarına verdikleri tepki ve web sitesinde geçirdikleri süre gibi faktörleri kullanarak, müşterilerin harcama miktarını tahmin etmek istemektedir.

Veri setinde her bir müşteri için aşağıdaki değişkenler bulunmaktadır:

- *Bağımsız Değişkenler:*
    1. *Müşterinin yaşadığı şehir nüfusu (binlerce kişi)*
    2. *Müşterinin son bir yılda yaptığı toplam harcama (TL)*
    3. *Müşterinin internet reklamlarına tıklama sayısı*
    4. *Müşterinin web sitesinde geçirdiği ortalama süre (dakika)*
- *Bağımlı Değişken:*
    1. *Müşterinin harcama miktarı (TL)*

Pazarlama şirketi, bu verilere dayanarak çoklu lineer regresyon analizi yaparak müşterilerin harcama miktarını tahmin etmek istemektedir.

Müşterilerin yaşadığı şehir nüfusu, geçmiş harcama miktarı, reklamlara verdikleri tepki ve web sitesinde geçirdikleri süre gibi faktörleri kullanarak, müşterilerin harcama miktarını tahmin etmek, pazarlama şirketine daha etkili bir pazarlama stratejisi oluşturma ve kaynakları doğru şekilde dağıtma konusunda yardımcı olacaktır.

Şirket yaşanılan şehir , son bir yılda yapılan harcama (TL) , internet reklamlarına tıklama sayısı , ve web sitesinde geçirilen ortalama süreye bakarak müşterinin yapacağı harcama miktarını tahmin ederek , ona göre bir yatırım planı yapmayı hedeflemektedir.

Yani müşterilerin harcama miktarını , birden fazla değişkeni kullanarak tahmin etmeye çalışacağız. Fakat öncelikle bu problemdeki verilerimize çoklu lineer regresyon analizi uygulayıp uygulayamacağımızı test etmeliyiz. Gerekli varsayım kontrollerini yapıp eğer veri setimiz varsayım kontrollerinde gerekli şartları sağlıyorsa çoklu lineer regresyon modelimizi oluşturalım.

VERİ TANITIMI

Verilerimiz şekildeki tabloda görüldüğü gibidir.

	A	B	C	D	E	F
1	Şehir Nüfusu	Harcama	Reklam Tıklama	Ortalama Süre	Harcama Miktarı	
2	500	10000	50	5	12000	
3	1000	15000	100	7	18000	
4	2000	20000	80	10	22000	
5	800	12000	70	6	14000	
6	1500	18000	90	8	20000	
7	1200	16000	120	9	17000	
8	1800	19000	110	7	21000	
9	900	11000	60	6	13000	
10	1300	14000	95	8	16000	
11	1100	13000	85	7	15000	
12						
13						

Bu veri setimizde bağımlı değişkenimiz Harcama Miktarı, bağımsız değişkenlerimiz ise tüm diğer değişkenlerdir.

Yani Şehir Nüfusu, Toplam Harcama, Reklama Tıklama sayısı ve internette geçirilen ortalama süreye bağlı olarak değişen müşterinin harcama miktarını inceleyeceğiz. Bunu çoklu lineer regresyon ile incelemek için veri setimiz bazı varsayımları sağlamak zorundadır. Bu varsayımlar şu şekildedir.

VARSAYIMLAR

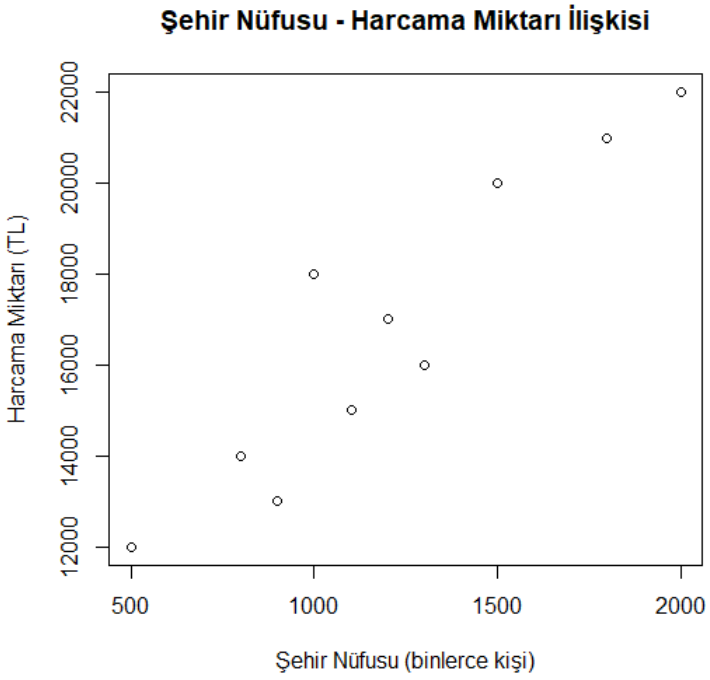
- Lineer İlişki Varsayımı
- Bağımsızlık Varsayımı
- Normallik Varsayımı

Lineer ilişki varsayımını incelemek için bağımlı değişkenin her bir bağımsız değişkenle olan scatter-plot grafiklerine bakıp lineer bir ilişki içinde olup olmadığına bakabiliriz.

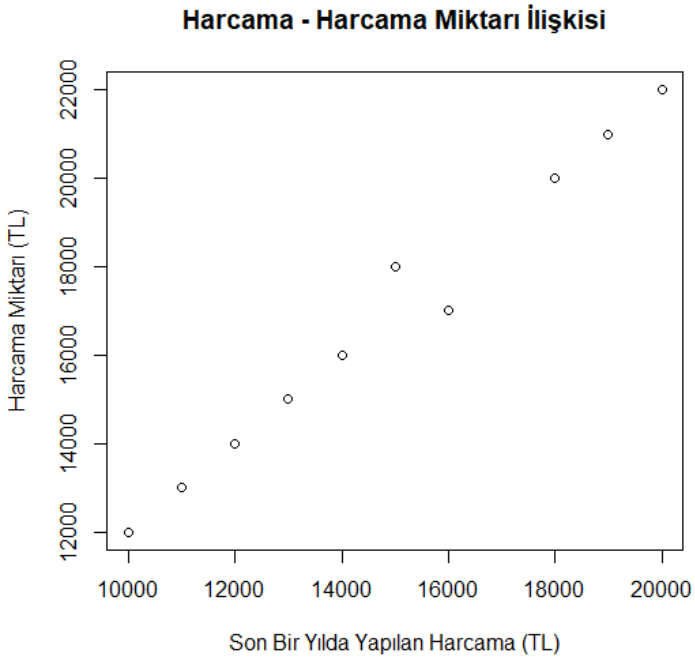
Bağımsızlık varsayımı için bağımsız değişkenler arasındaki korelasyonu inceleyeceğiz.Korelasyon iki değişken arasındaki ilişkiyi ve bu ilişkinin gücünü gösterir.Eğer iki bağımsız değişken arasında çok güçlü bir ilişki varsa bu durum varsayımımızı bozar.

```r
# Gerekli kütüphaneleri içe aktarın
library(stats)

# Müşteri veri setini oluşturun
sehir_nufusu <- c(500, 1000, 2000, 800, 1500, 1200, 1800, 900, 1300, 1100)
harcama <- c(10000, 15000, 20000, 12000, 18000, 16000, 19000, 11000, 14000, 13000)
reklam_tiklama <- c(50, 100, 80, 70, 90, 120, 110, 60, 95, 85)
ortalama_sure <- c(5, 7, 10, 6, 8, 9, 7, 6, 8, 7)
harcama_miktari <- c(12000, 18000, 22000, 14000, 20000, 17000, 21000, 13000, 16000, 15000)

# Veri setini birleştirin
veri <- data.frame(sehir_nufusu, harcama, reklam_tiklama, ortalama_sure, harcama_miktari)

# Scatter plot (nokta grafiği) ile bağımsız değişkenler ile bağımlı değişken arasındaki ilişkiyi görselleştirin
plot(veri$sehir_nufusu, veri$harcama_miktari, xlab = "Şehir Nüfusu (binlerce kişi)", ylab = "Harcama Miktarı (TL)", main = "Şehir Nüfusu - Harcama Miktarı İlişkisi")

# Diğer bağımsız değişkenler için de scatter plot'lar oluşturun
plot(veri$harcama, veri$harcama_miktari, xlab = "Son Bir Yılda Yapılan Harcama (TL)", ylab = "Harcama Miktarı (TL)", main = "Harcama - Harcama Miktarı İlişkisi")
plot(veri$reklam_tiklama, veri$harcama_miktari, xlab = "Reklam Tıklama Sayısı", ylab = "Harcama Miktarı (TL)", main = "Reklam Tıklama - Harcama Miktarı İlişkisi")
plot(veri$ortalama_sure, veri$harcama_miktari, xlab = "Web Sitesinde Geçirilen Ortalama Süre (dakika)", ylab = "Harcama Miktarı (TL)", main = "Ortalama Süre - Harcama Miktarı İlişkisi")
```

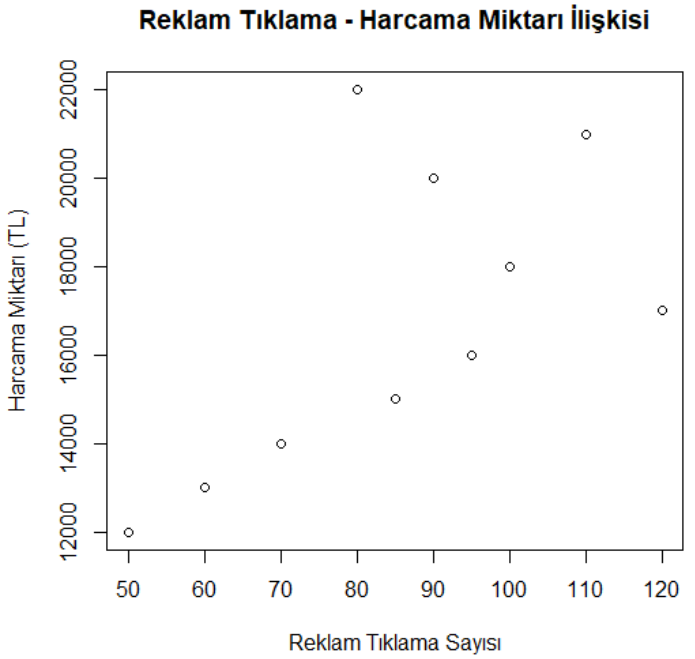
Şekildeki R kodlarını çalıştırdık.Bu kodlarda veri setimizi R ortamına aktardık ve Lineer İlişki varsayımımız için gerekli scatter-plot grafiklerimizi çağırdık.Bu grafikler şu şekildedir.



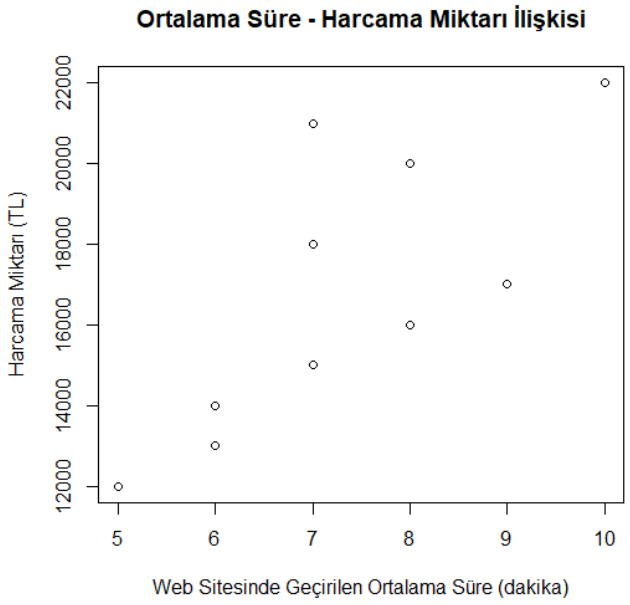
Görüdlüğü üzere lineer bir artış gözlenmekte



Görüldüğü üzere burda da çok net bir lineer artış gözlenmekte



Burada da bir lineerlik söz konusu, bazı ekstrem değerler olsada bunlar göz ardı edilebilir seviyede



Burada da bir lineer olma durumu söz konusu

Veri setimiz lineer ilişki varsayımını sağladı şimdi bağımsızlık varsayımını incelemek için korelasyon grafiğimizi çizdirelim.

```
# Korelasyon matrisini hesaplayın
cor_matrix <- cor(veri[, c("sehir_nufusu", "harcama", "reklam_tiklama", "ortalama_sure")])

# Korelasyon matrisini görüntüleyin
cor_matrix
```

```
> cor_matrix
              sehir_nufusu   harcama reklam_tiklama ortalama_sure
sehir_nufusu     1.0000000 0.9430968      0.5376513     0.7903418
harcama          0.9430968 1.0000000      0.6492633     0.7944217
reklam_tiklama   0.5376513 0.6492633      1.0000000     0.6027630
ortalama_sure    0.7903418 0.7944217      0.6027630     1.0000000
>
```

Bağımsız değişkenler arasındaki korelasyonu incelediğimizde şehir nüfusu ve harcama arasında çok güçlü bir ilişki olduğunu görüyoruz.Bu denli yüksek bir ilişki modelimizin anlamlılığına zarar verebilir bu yüzden modelimizi kurarken iki bağımsız değişkenden biri çıkarıp 3 bağımsız değişkenle devam edeceğiz.

Şimdi normallik testimizi yapalım. Hata terimlerimizin normal dağılıp dağılmadığını test etmeliyiz. Bunun anlamı gözlenen değer ile tahmin değeri arasındaki farkların dağılımının normal olmasıdır.Bu varsayımı test etmek için R'da shapiro wilk testimizi gerçekleştirelim.

```r
# Hata terimlerini hesaplayın
hata_terimleri <- residuals(model)

# Shapiro-Wilk testini uygulayın
shapiro.test(hata_terimleri)
```

```
        Shapiro-Wilk normality test

data:  hata_terimleri
W = 0.96643, p-value = 0.856

>
```

p-value değerimiz görüldüğü üzere 0.05'ten büyük olup H_0 Ret edilemez yani hata terimlerimiz normal dağılmaktadır. Üç varsayımımızı da kontrol ettik korelasyon matrisimizde güçlü ilişkisi olan değişkenlerimiz çıktı , aralarında 0.8'den büyük olup çok güçlü ilişkiye sahip olan değişkenlerimizden birini varsayımlarımızı ihmal ettiği ve modelimizin anlamlılığını riske attığı için modelimize eklemedik. Diğer varsayım kontrollerimizde herhangi bir problemle karşılaşmadık. Şimdi model özetimize bakıp yorumlarımızı yapalım.

```r
# Çoklu lineer regresyon modelini oluşturun
model <- lm(harcama_miktari ~ sehir_nufusu + reklam_tiklama + ortalama_sure, data = veri)

# Modelin özetini görüntüleyin
summary(model)
```

```
> summary(model)

Call:
lm(formula = harcama_miktari ~ sehir_nufusu + reklam_tiklama +
    ortalama_sure, data = veri)

Residuals:
     Min       1Q   Median       3Q      Max
-1570.19  -949.98   -67.09   568.79  2136.58

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    7234.629   2654.668   2.725  0.03440 *
sehir_nufusu      6.462      1.712   3.774  0.00924 **
reklam_tiklama   28.072     27.588   1.018  0.34815
ortalama_sure   -91.510    554.677  -0.165  0.87438
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1430 on 6 degrees of freedom
Multiple R-squared:  0.8837,    Adjusted R-squared:  0.8256
F-statistic:  15.2 on 3 and 6 DF,  p-value: 0.003284

>
```

Burada görüldüğü üzere p-value değerimiz 0.003 çıkmıştır. Bunun anlamı yaptığımız çoklu lineer regresyon modelimiz istatistiksel olarak %95 güvenle anlamlıdır yani bu modelimizi kullanarak müşterinin yapacağı harcama miktarını tahmin edebiliriz.

ANALİZ

Öncelikle karşımıza çıkan problemi anladık. Müşteri harcama davranışlarını inceleyen şirket, müşterinin harcamalarını bazı değişkenlerle tahmin edip ona göre bi politika izlemek istiyor.Bu harcamayı tahmin etmek içinde çoklu lineer regresyon analizine başvuruyoruz.Bu çoklu lineer regresyon analizini gerçekleştirebilmek için veriler toplanıyor ve bize sunuluyor.Bu verileri incelediğimizde öncelikle bu modeli uygulayabilmemiz için gerekli varsayım sınamalarını gerçekleştiriyoruz.İlk olarak bağımlı değişkenimiz ile bağımsız değişkenlerimiz arasındaki lineer ilişkiyi R'da scatter-plot grafiklerine bakarak analiz ediyoruz ve lineerlik dışına taşan ekstrem bir durum gözlemlenmiyor.İkinci olarak bağımlı değişkenimiz ve bağımsız değişkenlerimiz arasında çok güçlü bir ilişki olup olmadığını test ediyoruz.Bunun için korelasyon matrisimize bakıp 0.8'den büyük bir değerimiz olup olmadığını kontrol ediyoruz.Maalesef harcama ile şehir nüfusu 0.94 değerinde çok güçlü bir korelasyon değerimiz olup bu güçlü ilişki modelimizin anlamlılığını tehdit etmekte.Bundan dolayı iki değişkenden birini modelimizden çıkarıp varsayımlarımızı sağlayan değerler ile yolumuza devam ediyoruz.Üçüncü ve son sınamamız olan hata terimlerinin dağılımının normal olup olmadığına shapiro-wilks testi ile bakıyor ve burda da bir problem yaşamıyoruz.Tüm bu sınamaları gerçekleştirdikten sonra en nihayetinde çoklu lineer regresyon modelimizi çalıştırıp p-value değerini kontrol ediyoruz.Burda ki p-value değerimiz 0.05'ten küçük oldupu için modelimiz anlamlı çıktı yani modeldeki değerleri kullanabiliriz.O zaman modelimizi şu şekilde tablomuza bakarak fit edelim

Fit edilen model : $Y = 7234.629 + 6.462B_0 + 28.072B_1 - 91.510B_2$

Buradaki B_0 ,B_1 , B_2 Katsayılarımızın her biri bir bağımsız değişkenimizi ifade etmekte. Bu katsayılarımıza değer vererek müşterilerin bu verilere göre yapacağı harcamayı istatistiksel olarak tahmin edebiliriz.

Ayrıca Adjusted R-squared değerimiz 0.8256 olup bağımsız değişkenlerimiz , bağımlı değişkenimizi %82 oranında açıkladığını görüyoruz ki bu gayet iyi bir değer.

İsterseniz şimdi bir tahmin yapalım Şehir nüfusu 1500, reklam tıklama 95, ortalam süre 8 dakika olsun. Bu değerlere sahip bir müşterinin yapacağı harcama nasıl olur fit ettiğimiz modelden yararlanarak buna bir göz atalım.

```
43  # Yeni bir müşterinin verilerini kullanarak tahmin yapın
44  yeni_veri <- data.frame(sehir_nufusu = 1500, reklam_tiklama = 95, ortalama_sure = 8)
45  tahmin <- predict(model, newdata = yeni_veri)
46  tahmin
47
48
```

```
> tahmin
       1
18862.61
>
```

İşte görüldüğü gibi istatistiksel olarak anlamlı bir tahmin, söylemiş olduğumuz değerlere sahip bir müşterinin yapacağı harcama 18862.6 TL'dir

SONUÇ

Sonuç olarak elimize gelen veri setinden çoklu lineer regresyon analizi yaparak istatistiksel olarak anlamlı bir sonuç elde ettik.

Müşterinin harcama miktarını , müşterinin yaşadığı şehir nüfusu, müşterinin internet reklamlarına tıklama sayısı ve müşterinin web sitesinde geçirdiği ortalam süreye bakarak istatistisel olarak anlamlı bir şekilde tahmin ettik.

Artık şirket bu fit edilen model yardımıyla politikalarına yön verebilir ve satışlarını arttırabilir.