



İST470 - Kategorik Veri Çözümlemesi

Yusuf Ziya Ateş 21935621 (yusufziyaates7@gmail.com)
Fatih Can Karaca 21936062 (karaca.fatihcan@gmail.com)
Yiğit Saçık 21936277 (yigitsacik@gmail.com)
Ahmet Serdar Candesteci 21935759
(aserdarcandesteci@hotmail.com)





Veri Hakkında Bilgiler

Bu veri setinde bir bankanın müşterilerine sattığı birikim hesabı, kredi kartı, yatırım vb. birden çok bankacılık ürünü bulunmaktadır. Kredi kartını hangi müşterisinin temin edeceğini belirlemek istemektedir. Aynı zamanda, müşterinin demografik detayları, bankacılık davranışları vb. ile ilgili çeşitli bilgilere de sahibiz.



- 0 - Id :(kimlik no)
- 1 – age :(yas)
- 2 -job : (işi ne?)
- 3 - marital : (medeni hali)
- 4 - education : (eğitim durumu)
- 5 - default: (kredisi var mı yok mu?)
- 6 - balance: (banka hesabındaki para (\$))
- 7 - housing: (ev kredi borcu var mı?)
- 8 – loan : (kişisel kredi borcu var mı?)
- 9 - contact: (iletişime geçme türü)
- 10 - day: (ay içerisinde en son iletişim kurulan gün tarihi)
- 11 - month: (yıl içerisinde en son iletişim kurulan ay)

- 12 - duration: (son görüşmede geçen süre (dk))
- 13 - campaign: (bu kampanya sırasında müşteriyle gerçekleştirilen görüşme sayısı)
- 14 - pdays: (son kampanyadan bu yana müşteriyle yapılan en son görüşmeden bugüne kadarki gün sayısı)
- 15 - previous: (bu kampanyadan önce ve bu müşteri için gerçekleştirilen görüşme sayısı)
- 16 - poutcome: (önceki pazarlama kampanyasının sonucu)
- 17 - y – (Müşteri vadeli mevduata abone oldu mu?)

Veri Seti Yükleme ve Örneklem Seçimi

Veri setimizde 45,211 gözlem; 6 nicel (age, balance, duration, campaign, pdays, previous) değişken ve 12 kategorik (id, job, marital, education, default, loan, housing, contact, day, month, poutcome, y) değişken mevcuttur.

```
> str(ist)
tibble [45,211 x 18] (S3: tbl_df/tbl/data.frame)
 $ Id      : num [1:45211] 1001 1002 1003 1004 1005 ...
 $ age     : num [1:45211] 999 44 33 47 33 35 28 NA 58 43 ...
 $ job     : chr [1:45211] "management" "technician" "entrepreneur" "blue-collar"
 $ marital : chr [1:45211] "married" "single" "married" "married" ...
 $ education: chr [1:45211] "tertiary" "secondary" "secondary" "unknown" ...
 $ default : chr [1:45211] "no" "no" "no" "no" ...
 $ balance : num [1:45211] 2143 29 2 1506 1 ...
 $ housing : chr [1:45211] "yes" "yes" "yes" "yes" ...
 $ loan    : chr [1:45211] "no" "no" "yes" "no" ...
 $ contact : chr [1:45211] "unknown" "unknown" "unknown" "unknown" ...
 $ day     : num [1:45211] 5 5 5 5 5 5 5 5 5 5 ...
 $ month   : chr [1:45211] "may" "may" "may" "may" ...
 $ duration: num [1:45211] 261 151 76 92 198 139 217 380 50 55 ...
 $ campaign: num [1:45211] 1 1 1 1 1 1 1 1 1 1 ...
 $ pdays  : num [1:45211] -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
 $ previous: num [1:45211] 0 0 0 0 0 0 0 0 0 0 ...
 $ poutcome: chr [1:45211] "unknown" "unknown" "unknown" "unknown" ...
 $ y       : chr [1:45211] "no" "no" "no" "no" ...
```



```
# Rastgele örneklem çekme  
orneklem <- ist[sample(nrow(ist), 300, replace = FALSE), ]
```

Veri setimizde çok fazla gözlem olduğu için basit rastgele örnekleme yöntemi ile 45,211 veriden 300 gözlem rasgele çekilmiştir. Analizlere bu 300 gözlem üzerinden devam edilecektir.

```
> str(orneklem)  
tibble [300 x 18] (S3: tbl_df/tbl/data.frame)  
$ Id      : num [1:300] 4508 26957 31710 37946 24760 ...  
$ age     : num [1:300] 48 52 47 53 30 41 36 27 59 48 ...  
$ job     : chr [1:300] "management" "admin." "management" "admin." ...  
$ marital : chr [1:300] "married" "married" "married" "divorced" ...  
$ education: chr [1:300] "unknown" "primary" "tertiary" "secondary" ...  
$ default : chr [1:300] "no" "no" "no" "no" ...  
$ balance : num [1:300] 1103 2390 0 28 1239 ...  
$ housing : chr [1:300] "yes" "no" "yes" "yes" ...  
$ loan    : chr [1:300] "no" "yes" "no" "yes" ...  
$ contact : chr [1:300] "unknown" "cellular" "telephone" "cellular" ...  
$ day     : num [1:300] 15 19 6 12 28 19 21 3 5 3 ...  
$ month   : chr [1:300] "may" "nov" "feb" "may" ...  
$ duration: num [1:300] 52 216 138 673 125 ...  
$ campaign: num [1:300] 1 2 4 4 5 1 1 1 2 1 ...  
$ pdays  : num [1:300] -1 -1 9 -1 -1 -1 -1 -1 -1 351 ...  
$ previous: num [1:300] 0 0 2 0 0 0 0 0 0 2 ...  
$ poutcome: chr [1:300] "unknown" "unknown" "other" "unknown" ...  
$ y       : chr [1:300] "no" "no" "no" "no" ...
```

```
> sapply(orneklem, function(x) sum(is.na(x)))
      Id      age      job marital education default balance housing  loan contact   day  month duration
      0         0         0         0         0         0         0         0         0         0         0         0
campaign pdays previous poutcome      y
      0         0         0         0         0
```

Seçtiğimiz örnekleme kayıp verimizin olup olmadığına baktık. Bunun sonucunda kayıp verimiz olmadığı için analizimize devam edebiliriz.

Tanımlayıcı İstatistikler

```
> summary(orneklem)
      Id      age      job      marital      education      default
Min.   : 1034  Min.   :20.00 Length:300 Length:300 Length:300 Length:300
1st Qu.:12582 1st Qu.:33.00 Class :character Class :character Class :character Class :character
Median :24070 Median :41.00 Mode  :character Mode  :character Mode  :character Mode  :character
Mean   :24455 Mean   :42.21
3rd Qu.:36783 3rd Qu.:50.00
Max.   :45604 Max.   :74.00

      balance      housing      loan      contact      day      month
Min.   : -1168.0 Length:300 Length:300 Length:300 Min.   : 1.00 Length:300
1st Qu.:   49.0 Class :character Class :character Class :character 1st Qu.: 8.00 Class :character
Median :  518.5 Mode  :character Mode  :character Mode  :character Median :15.00 Mode  :character
Mean   : 1474.5
3rd Qu.: 1404.5
Max.   :27696.0

      duration      campaign      pdays      previous      poutcome      y
Min.   :    7.0 Min.   : 1.00 Min.   : -1.00 Min.   : 0.0000 Length:300 Length:300
1st Qu.:  104.0 1st Qu.: 1.00 1st Qu.: -1.00 1st Qu.: 0.0000 Class :character Class :character
Median :  185.5 Median : 2.00 Median : -1.00 Median : 0.0000 Mode  :character Mode  :character
Mean   :  274.6 Mean   : 3.04 Mean   : 43.88 Mean   : 0.8133
3rd Qu.:  341.5 3rd Qu.: 3.00 3rd Qu.: -1.00 3rd Qu.: 0.0000
Max.   :1409.0 Max.   :36.00 Max.   :670.00 Max.   :58.0000
```




Age değişkeni için:	Kişilerin yaşları 20 ile 74 arasında dağılmaktadır. Ortalama yaş 42 yıl 2 aydır. Medyan (41) ortalamadan (42.2) daha küçük olduğundan sola çarpık bir dağılım söz konusu olabilir.
Balance değişkeni için:	Kişilerin hesaplarındaki para miktarı -1168\$ ile 27696\$ arasında dağılmaktadır. Katılımcıların ortalama para miktarları ise 1474.5'tir. Medyan (518.5) ortalamadan (1474.5) daha küçük olduğundan sola çarpık bir dağılım söz konusu olabilir.
Duration değişkeni için:	Kişilerin bankayla görüşme süreleri 7 dk ile 1409 dk arasında dağılmaktadır. Katılımcıların ortalama görüşme süresi ise 274.6 dk'dır . Medyan (185.5) ortalamadan (274.6) daha küçük olduğundan sola çarpık bir dağılım söz konusu olabilir.

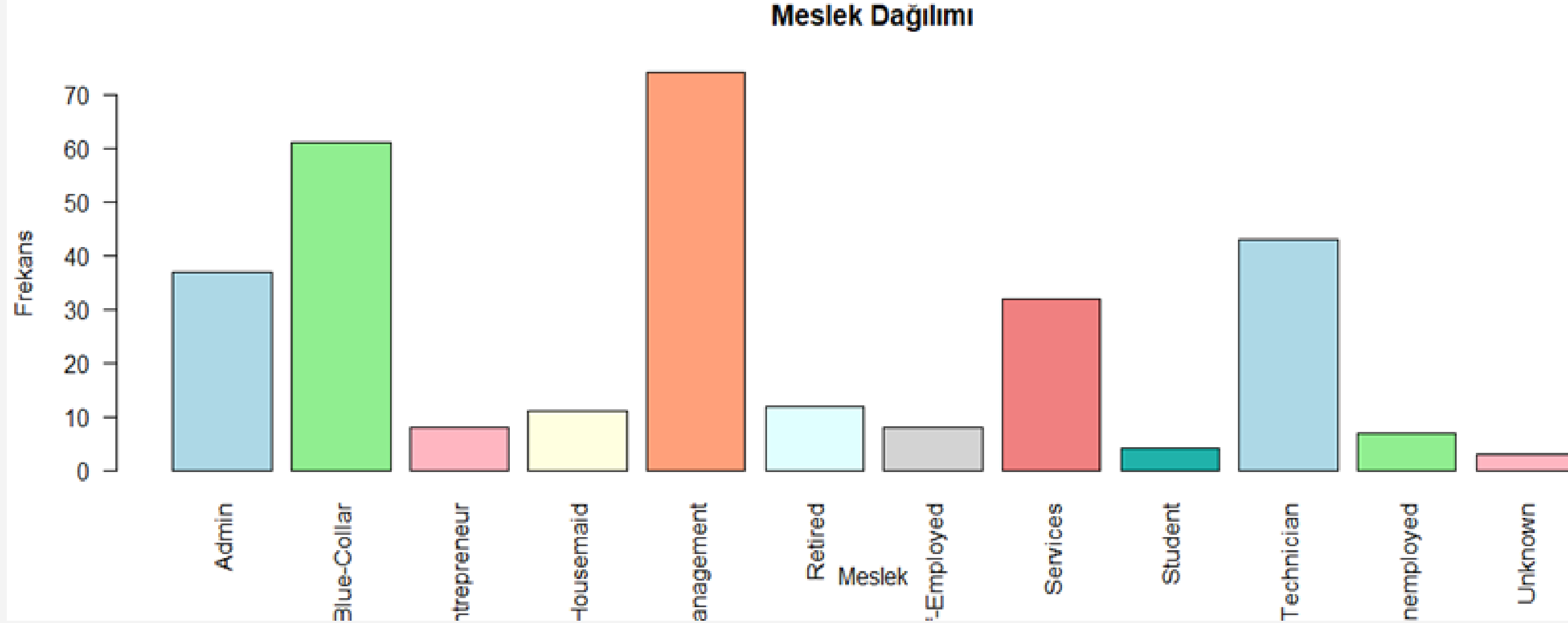
Campaign değişkeni için:	müşteriyle gerçekleştirilen görüşme sayısı 1 ile 36 arasında dağılmaktadır. Gerçekleştirilen ortalama görüşme sayısı 3.04`tür. Medyan (2) ortalamadan (3.04) daha küçük olduğundan sola çarpık bir dağılım söz konusu olabilir.
Pdays değişkeni için:	son kampanyadan bu yana müşteriyle yapılan en son görüşmeden bugüne kadarki gün sayısı -1 ile 670 gün arasında dağılmaktadır. Ortalama bu süre 43.88`dir. Medyan (-1) ortalamadan (43.88) daha küçük olduğundan sola çarpık bir dağılım söz konusu olabilir.
Previous değişkeni için:	bu kampanyadan önce ve bu müşteri için gerçekleştirilen görüşme sayısı 0 ile 58 sefer arasında dağılmaktadır. Ortalama bu sayı 0.81`dir. Medyan (0) ortalamadan (0.81) daha küçük olduğundan sola çarpık bir dağılım söz konusu olabilir.

Veri Görselleştirme

```
###HISTOGRAM GRAFIĞİ
frekanslar <- table(ornekleme$job)
meslekler <- c("Admin", "Blue-Collar", "Entrepreneur", "Housemaid", "Management", "Retired", "Self-Employed", "Services", "Student", "Technician", "Unemployed", "Unknown")

library(ggplot2)
# Renkler
renkler <- c("lightblue", "lightgreen", "lightpink", "lightyellow", "lightsalmon", "lightcyan", "lightgray", "lightcoral", "lightseagreen")

# Çubuk grafik oluşturma
barplot(frekanslar,
        main = "Meslek Dağılımı",
        xlab = "Meslek",
        ylab = "Frekans",
        col = renkler,
        names.arg = meslekler,
        las = 2)
```



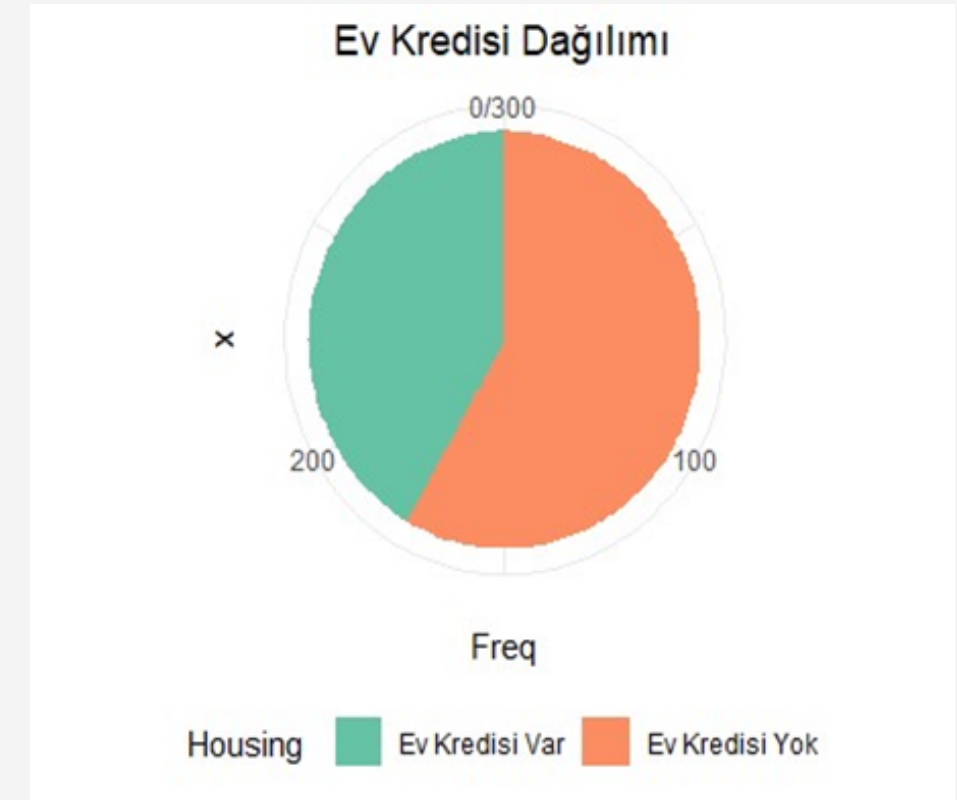
Örnekleminizdeki meslek dağılımı grafikte görülmektedir. Görüldüğü üzere örnekleminizde en fazla çalışan kişi sayısı management iş grubundadır.

```
###PASTA GRAFİĞİ
# Pasta grafiği için veri tablosunun oluşturulması
veri_tablosu <- table(orneklem$housing)

# Pasta grafiği çizimi
install.packages("ggplot2")
library(ggplot2)

# Pasta grafiği çizimi|
ggplot(data.frame(veri_tablosu), aes(x = "", y = Freq, fill = factor(veri_tablosu))) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar(theta = "y") +
  labs(fill = "Housing") +
  theme_minimal() +
  theme(legend.position = "bottom") +
  scale_fill_manual(values = c("#66C2A5", "#FC8D62"),
                    labels = c("Ev Kredisi Var", "Ev Kredisi Yok")) + # Renkleri özelleştirme ve açıklamaları ekleme
  ggtitle("Ev Kredisi Dağılımı") +
  theme(plot.title = element_text(hjust = 0.5))
```

Pasta grafiğini incelediğimizde ev kredisi olan ve ev kredisi olmayan kişilerin dağılımını görmekteyiz. Bu sonuca göre ev kredisine sahip olmayan kişilerin daha fazla olduğunu söyleyebiliriz.



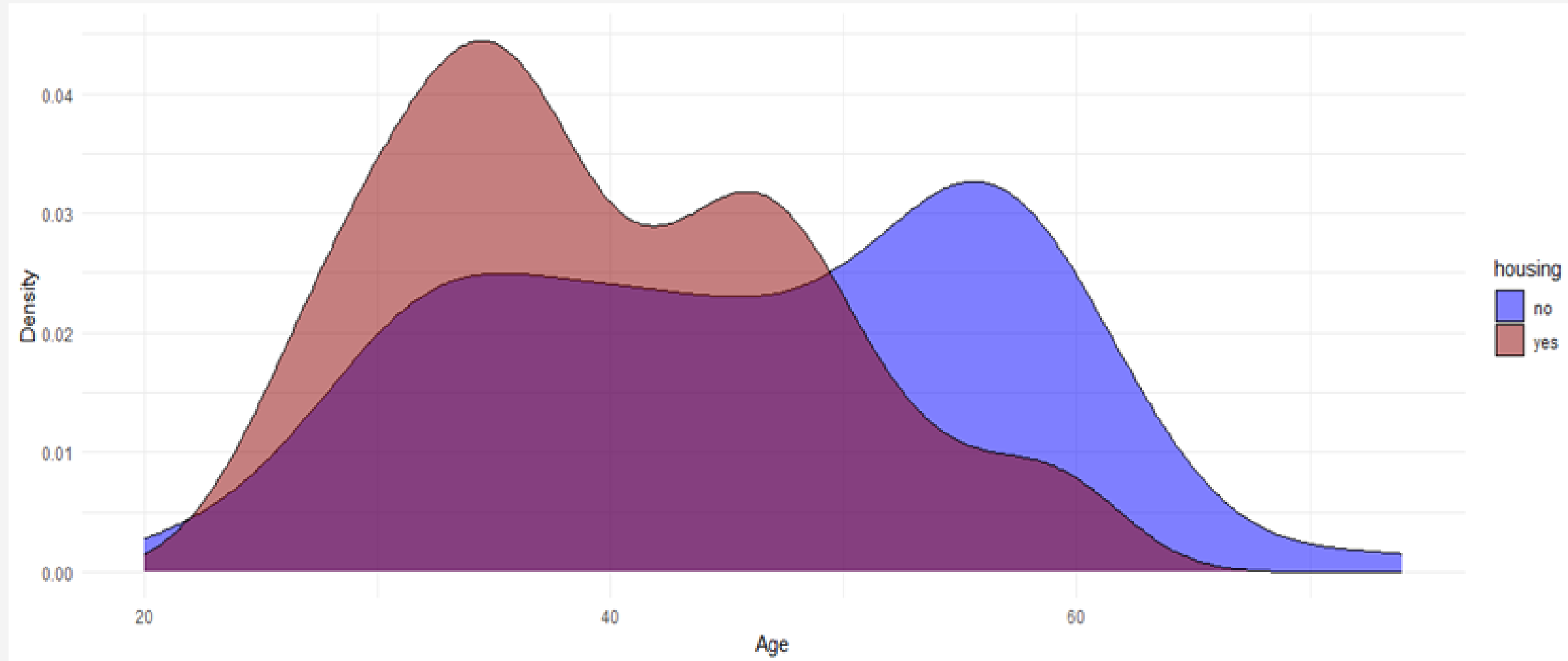
```
###DENSITY GRAFİĞİ ÇİZDİRME
# Grafik çizimi
install.packages("ggplot2")
library(ggplot2)

ggplot(data.frame(orneklem), aes(x = age, fill = housing)) +
  geom_density(alpha = 0.5) +
  labs(x = "Age", y = "Density") +
  theme_minimal() +
  scale_fill_manual(values = c("blue", "darkred"))
```


Aşağıda yaş değişkenine göre ev kredisine sahip olup olmama durumunun yoğunluk grafiği verilmiştir.

Ev kredisine sahip olanlar 20 ile 40 yaş arasında yoğunlaşmaktadır ayrıca yaş ilerledikçe ev kredisine sahip olma oranı düşmektedir.

Ev kredisine sahip olmayanlar ise 50 ile 60 yaş arasında yoğunlaşmaktadır.



İlişki Matrisi

```
# Korelasyon matrisini hesaplama
korelasyon_matrisi <- cor(orneklem[c("age", "balance", "duration")])

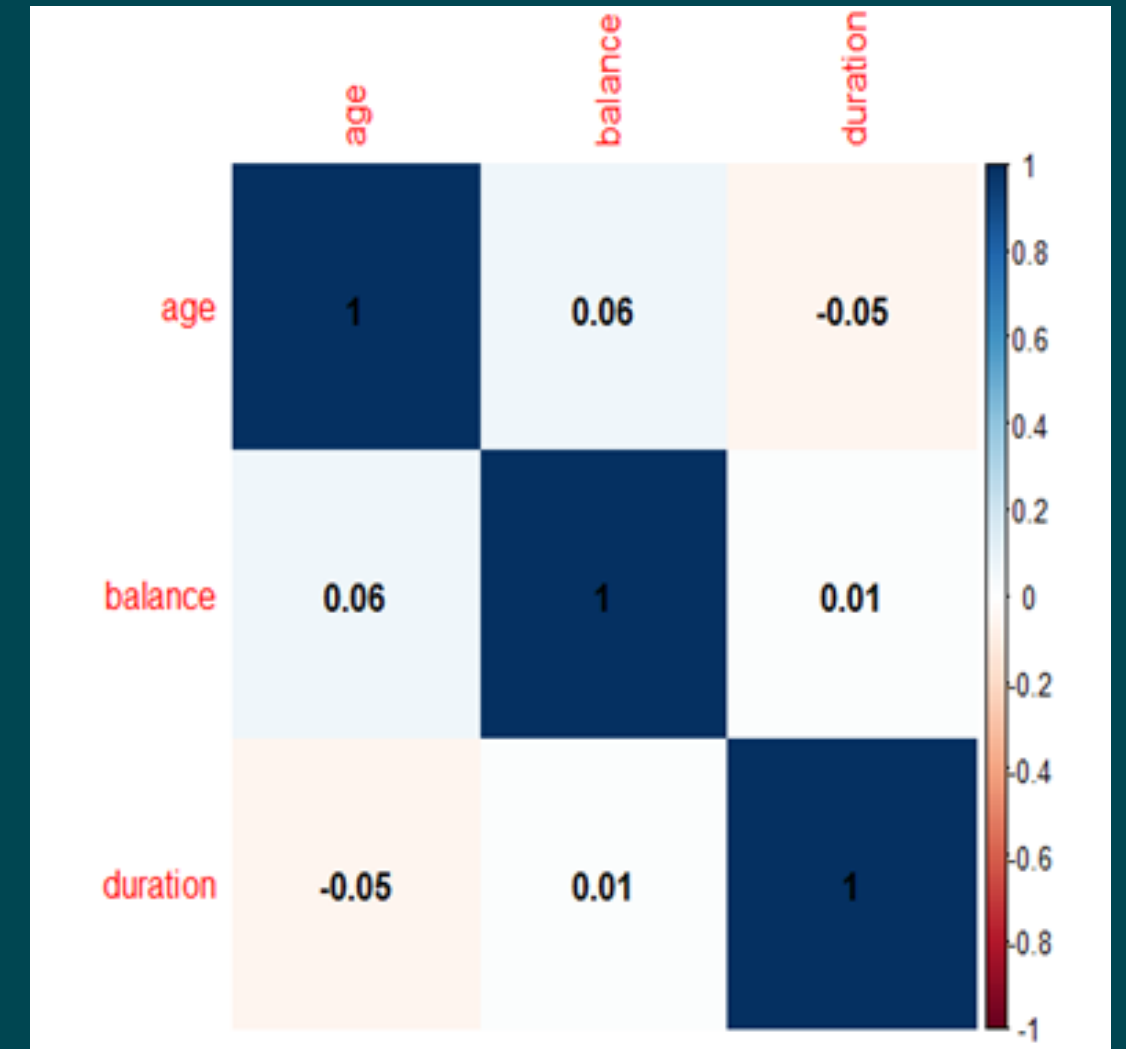
# Korelasyon matrisini yazdırma
print(korelasyon_matrisi)
```

```
> print(korelasyon_matrisi)
```

	age	balance	duration
age	1.000000000	0.06164574	-0.05082801
balance	0.06164574	1.000000000	0.01468036
duration	-0.05082801	0.01468036	1.000000000

Korelasyon Matrisi Görselleştirme

```
###KORELASYON MATRİSİ  
# Veri çerçevesi oluşturma  
sayisal_degiskenler <- c("age", "balance", "duration")  
alt_veri <- orneklem[, sayisal_degiskenler]  
# Korelasyon matrisi oluşturma  
korelasyon_matrisi <- cor(alt_veri)  
# Korelasyon matrisinin görselleştirilmesi  
install.packages("corrplot")  
library(corrplot)  
corrplot(korelasyon_matrisi, method = "color")
```



Satırda ve sütunda değişkenler olmak üzere korelasyon matrisi yukarıda verilmiştir. Buradan görebileceğimiz üzere duration ve age değişkenleri arasında orta negatif yönlü bir ilişki olduğunu balance ile age değişkenleri arasında ise orta derecede pozitif ilişki olduğunu söyleyebiliriz. Ayrıca balance ve duration arasında zayıf pozitif yönlü ilişki görülmektedir.

RXC Çözümlemesi ve Uyum Analizi

Eğitim düzeyi ve ev kredisi olup olmama arasında ilişki olduğunu düşündüğümüz için bu iki değişken arasında RxC çözümlemesi yapılması uygun görülmüştür.

Housing: “yes”, “no”

Education: “primary”, “secondary”, “tertiary”, “unknown”

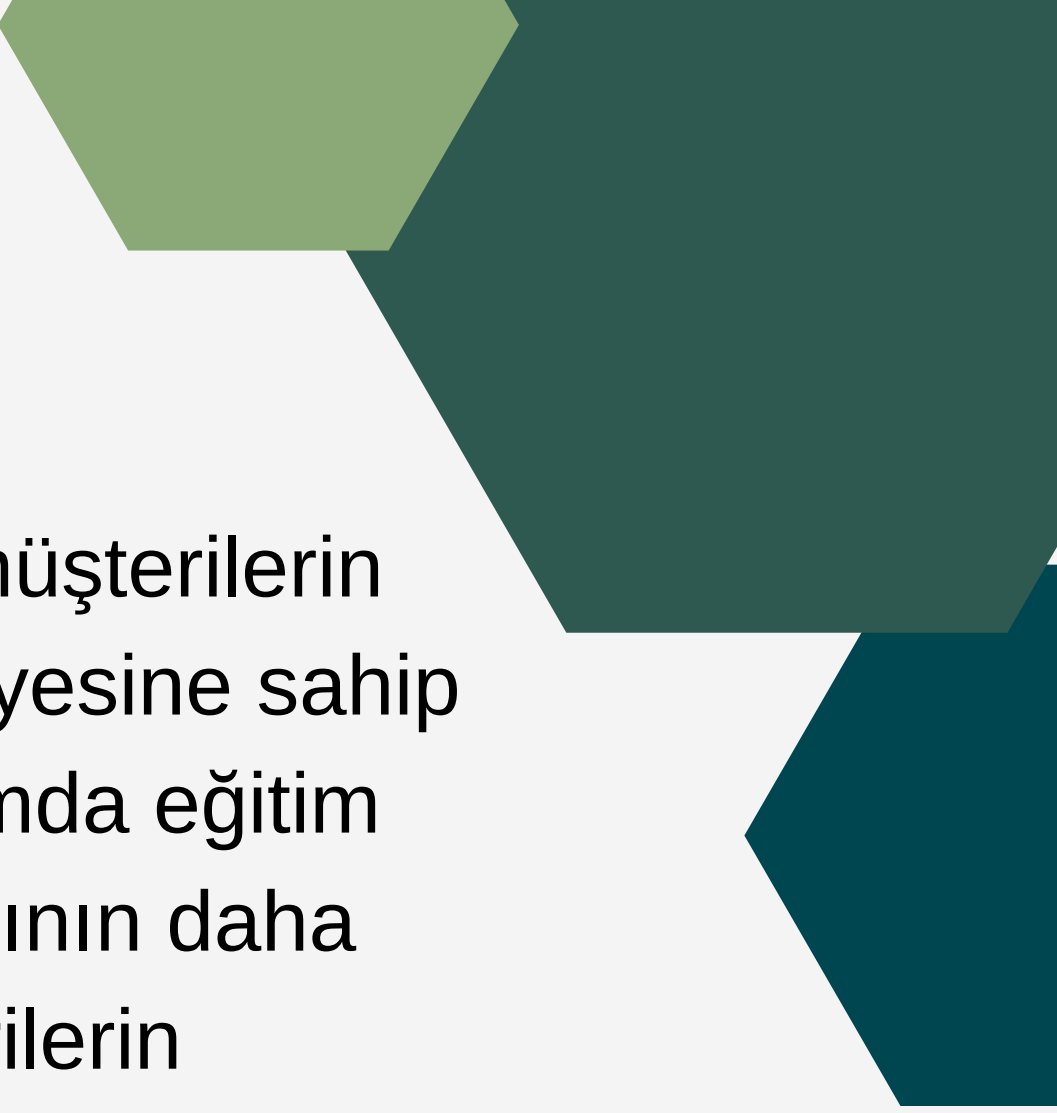
housing * education Crosstabulation

			education				Total
			primary	secondary	tertiary	unknown	
housing	no	Count	25	46	49	5	125
		Expected Count	22,5	61,3	36,7	4,6	125,0
		% within housing	20,0%	36,8%	39,2%	4,0%	100,0%
		% within education	46,3%	31,3%	55,7%	45,5%	41,7%
		% of Total	8,3%	15,3%	16,3%	1,7%	41,7%
	yes	Count	29	101	39	6	175
		Expected Count	31,5	85,8	51,3	6,4	175,0
		% within housing	16,6%	57,7%	22,3%	3,4%	100,0%
		% within education	53,7%	68,7%	44,3%	54,5%	58,3%
		% of Total	9,7%	33,7%	13,0%	2,0%	58,3%
Total	Count	54	147	88	11	300	
	Expected Count	54,0	147,0	88,0	11,0	300,0	
	% within housing	18,0%	49,0%	29,3%	3,7%	100,0%	
	% within education	100,0%	100,0%	100,0%	100,0%	100,0%	
	% of Total	18,0%	49,0%	29,3%	3,7%	100,0%	

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	14,162 ^a	3	,003
Likelihood Ratio	14,242	3	,003
N of Valid Cases	300		

a. 1 cells (12,5%) have expected count less than 5. The minimum expected count is 4,58.



"Education" deęiřkeninde "primary" eęitim seviyesine sahip müşterilerin %16,6'sı ev kredisi almaya yatkınken "secondary" eęitim seviyesine sahip müşterilerinin %57,7'si ev kredisi almaya yatkındır. Bu durumda eęitim seviyesi "secondary" olan müşterilerin ev kredisi alma olasılıęının daha yüksek olduğunu gösterir. Ev kredisine sahip olmayan müşterilerin çoęunluęunu "tertiary" eęitim seviyesine sahip kişiler oluşturmaktadır.



Ki-Kare Tablosu Yorum:

RxC tablosunda beklenen sıklık değerlerinden 3'ten büyük olduğu için veya beklenen sıklık yüzdesi sıfır olmadığı için 2×4'lük çapraz tablodaki verilerin χ^2 dağılımı gösterdiği söylenemez. Bu yüzden hipotez test edilirken Fisher' χ^2 istatistiğine karşılık gelen Likelihood Ratio ile ilgili değerler yorumlanır.

H_0 : Eğitim düzeyi ile ev kredisi alma arasında ilişki yoktur.

H_s : Eğitim düzeyi ile ev kredisi alma arasında ilişki vardır.

Likelihood p-value = 0,003 ve $\alpha = 0,05$ değerlerine baktığımızda H_0 hipotezimiz reddedilemez. Eğitim ile ev kredisi arasında ilişki olmadığını %95 güven düzeyinde söyleyebiliriz.



Yerel Bütünleşik Odds Oranları

```
> tablo <- table(orneklem$housing, orneklem$education)
> # Yerel odds oranını hesapla
> or<-local_odds_ratio <- loddsratio(tablo, correct = any(tablo == 0L), log = FALSE)
> # Sonucu yazdır
> print(local_odds_ratio)
odds ratios for  and

      primary:secondary secondary:tertiary  tertiary:unknown
              1.8928036              0.3624975              1.5076923
> confint(or)
              2.5 %    97.5 %
no:yes/primary:secondary 0.9995714 3.5842416
no:yes/secondary:tertiary 0.2099188 0.6259775
no:yes/tertiary:unknown  0.4280314 5.3106763
```

θ_{11} : 1,89 olarak bulunmuştur. Ev kredisine sahip olup olmamalarına göre eğitim durumunun “primary” olma olasılığı “secondary” olma olasılığından 1,89 kat daha fazladır.

θ_{12} : 0,36 olarak bulunmuştur. Ev kredisine sahip olup olmamalarına göre eğitim durumunun “secondary” olması olasılığı “tertiary” olması olasılığından 0,36 kat fazladır. Ama $0,36 < 1$ olduğu için odds oranının tersi alınarak yorum yapılması daha doğrudur. Yani ev kredisine sahip olanların ev kredisine sahip olmayanlara göre eğitim durumunun “secondary” olması olasılığı “tertiary” olması olasılığından 2,77 kat daha fazladır. ($1/0,36$)

θ_{13} : 1,50 olarak bulunmuştur. Ev kredisine sahip olup olmamalarına göre eğitim durumunun “tertiary” olma olasılığı “unknown” olma olasılığından 1,50 kat daha fazladır diyebiliriz.

Önem Kontrolü

$$H_0: \theta_{11} = 0$$

%95 güven aralığı [0,99 ; 3,58] içinde “1” değeri yer aldığı için reddedilemez ve istatistiksel olarak anlamlı olmadığı söylenebilir.

$$H_0: \theta_{12} = 0$$

%95 güven aralığı [0,20 ; 0,62] içinde “1” değeri yer almadığı için reddedilir ve istatistiksel olarak anlamlı olduğu söylenebilir.

$$H_0: \theta_{13} = 0$$

için ise %95 güven aralığı [0,42 ; 5,31] içinde “1” değeri yer aldığı için reddedilemez ve istatistiksel olarak anlamlı olmadığı söylenebilir.

```
> confint(or)
              2.5 %    97.5 %
no:yes/primary:secondary 0.9995714 3.5842416
no:yes/secondary:tertiary 0.2099188 0.6259775
no:yes/tertiary:unknown  0.4280314 5.3106763
```

SATIR-SÜTUN ETKİ MODELLERİNİN ÇÖZÜMLENMESİ

Satır Etki Modeli (Row Effects Model) ($N \times O$)

RxC tablolarında;

X = Satır değişkeni (Housing)

Y = Sütun değişkeni (Education)

Satır değişkeni sınıflandırılabilir ve sütun değişkeni sıralanabilir nitel değişken olduğu için çözümlemede satır etkisi modeli kullanılacaktır.

```
> # Satır etkisi modeli
> library(vcd)
> model_satir_etki <- assocstats(table(orneklem$housing, orneklem$education))
> # Sonuçları yazdırın
> print(model_satir_etki)
```

	X ²	df	P(> X ²)
Likelihood Ratio	14.242	3	0.0025942
Pearson	14.162	3	0.0026929

H_0 : Satır etki modeline uyum vardır.

H_s : Satır etki modeline uyum yoktur.

Uyum iyiliği sonuçlarına göre yokluk hipotezi kabul edilemez ($p < 0,05$). Bu sebepten, satır etki modeline uyum olmadığını %95 güven düzeyinde söyleyebiliriz. Bu sebepten analizime devam edemeyiz.

Satır Etki Modeli (Row Effects Model) (NxO)

RxC tablolarında;

X = Sütun değişkeni (Education)

Y = Satır değişkeni (Housing)

Goodness-of-Fit Tests^{a,b}

	Value	df	Sig.
Likelihood Ratio	11,569	2	,003
Pearson Chi-Square	11,549	2	,003

a. Model: Poisson

b. Design: Constant + education + housing + T1

H_0 : Sütun etki modeline uyum vardır.

H_s : Sütun etki modeline uyum yoktur.

Uyum iyiliği sonuçlarına göre yokluk hipotezi reddedilir ($p < 0,05$) ve sütun etki modeline uyum olmadığı söylenebilir.

RxCxK TABLOLARININ ÇÖZÜMLENMESİ

Bu çözümlemede amacımız banka hesabındaki paranın, ev kredi borcuna ve eğitim durumuna göre nasıl bir değişim gösterdiğini incelemektir. Bu bağlamda çözümleme için seçilen değişkenler ise şu şekildedir:

Housing: “yes”, “no”

Education: “primary”, “secondary”, “tertiary”, “unknown”

Orneklem\$katgoriler: “Düşük”, “Orta”, “Yüksek”


```

> ###SINIFLANDIRMA
> # Kategorileri belirleyin ve verileri kategorilere ayırın
> kategori <- cut(orneklem$balance, breaks = c(-Inf, 0, 2000, Inf), labels = c("Dusuk", "Orta", "Yuksek"))
> # Kategorilere ayrılmış verileri örneklem veri setine ekleyin
> orneklem$ kategoriler <- kategori
> # Tabloyu oluşturma
> tablo <- xtabs(~housing + education + orneklem$ kategoriler, data = orneklem)
> # Tabloyu görüntüleme
> print(tablo)
, , orneklem$ kategoriler = Dusuk

      education
housing 1  2  3  4
no       3 11  5  0
yes      8 17  6  1

, , orneklem$ kategoriler = Orta

      education
housing 1  2  3  4
no      20 29 24  2
yes     17 68 26  5

, , orneklem$ kategoriler = Yuksek

      education
housing 1  2  3  4
no       2  6 20  3
yes      4 16  7  0

```

X = Satır Değişkeni (Housing) = Sınıflanabilir

Y = Sütun Değişkeni (Education) = Sınıflanabilir

Z = Tabaka Değişkeni (Orneklem\$ kategoriler) = Sınıflanabilir

```

> ftable(tablo)
      orneklem$ kategoriler Dusuk Orta Yuksek
housing education
no      1      3      20      2
        2     11     29      6
        3      5     24     20
        4      0      2      3
yes     1      8     17      4
        2     17     68     16
        3      6     26      7
        4      1      5      0

```

Kurulan modeller aşağıdaki gibidir:

```
#Bağımsız Modeli (M0)
model0 <- glm(Freq ~ housing + education + orneklem.kategoriler, family = poisson, data = tablo_df)

#Kısmi Bağımsız Modeller
model1 <- glm(Freq ~ housing + education + orneklem.kategoriler +
              housing*orneklem.kategoriler, family = poisson, data = tablo_df)
model2 <- glm(Freq ~ housing + education + orneklem.kategoriler +
              housing*education, family = poisson, data = tablo_df)
model3 <- glm(Freq ~ housing + education + orneklem.kategoriler +
              education*orneklem.kategoriler, family = poisson, data = tablo_df)

#Koşullu Bağımsız Modeller
model4 <- glm(Freq ~ housing + education + orneklem.kategoriler + housing*education +
              housing*orneklem.kategoriler, family = poisson, data = tablo_df)
model5 <- glm(Freq ~ housing + education + orneklem.kategoriler + housing*orneklem.kategoriler +
              education*orneklem.kategoriler, family = poisson, data = tablo_df)
model6 <- glm(Freq ~ housing + education + orneklem.kategoriler + housing*education +
              education*orneklem.kategoriler, family = poisson, data = tablo_df)

#Karşılıklı Bağımsız Modeller
model7 <- glm(Freq ~ housing + education + orneklem.kategoriler + housing*education + housing*orneklem.kategoriler +
              education*orneklem.kategoriler, family = poisson, data = tablo_df)


#Doygun Model
model8 <- glm(Freq ~ housing + education + orneklem.kategoriler + housing*education + housing*orneklem.kategoriler +
              education*orneklem.kategoriler + housing*education*orneklem.kategoriler, family = poisson, data = tablo_df)
```

Modellerin uyumunun test edilmesi için kullanılan R çıktısı ise aşağıdaki gibidir:

```
> LRstats(model0, model1, model2, model3, model4, model5, model6, model7, model8)
Likelihood summary table:
```


	AIC	BIC	LR	Chisq	Df	Pr(>Chisq)	
model0	144.85	153.10	43.038	17	0.0004740	***	
model1	144.73	155.33	38.914	15	0.0006601	***	
model2	136.61	148.39	28.797	14	0.0111328	*	
model3	144.44	159.75	30.625	11	0.0012628	**	
model4	136.48	150.62	24.672	12	0.0164543	*	
model5	144.31	161.98	26.501	9	0.0016905	**	
model6	136.20	155.04	16.384	8	0.0372028	*	
model7	137.88	159.08	14.063	6	0.0289441	*	
model8	135.81	164.09	0.000	0	< 2.2e-16	***	

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Çıktıya göre, en iyi model "model8" olarak belirlenmiştir. Bu model, en düşük AIC ve BIC değerlerine sahip olduğu için tercih edilmiştir. Ayrıca, LR Chisq değeri 0.000 olduğu için diğer modellerle karşılaştırıldığında anlamlı bir şekilde daha iyi uyum sağladığı görülmektedir. Bu nedenlerle, model8 en iyi model olarak değerlendirilmiştir.

8.model=Sabit+ housing + education + orneklem.kategoriler + housing*education + housing*orneklem.kategoriler + education*orneklem.kategoriler + housing*education*orneklem.kategoriler



SIRALANABİLİR KATEGORİK DEĞİŞKENLERDE LOGARİTMİK DOĞRUSAL MODELLER

NxNxO modeli için seçtiğimiz değişkenler şunlardır:

N: Age

N: Education

O: Marital

Modeli kurabilmek için öncelikle verilerin hazırlanması gerekmektedir. Kategorik değişkenlerin faktör olarak tanımlanması bu hazırlık sürecinin bir parçasıdır.

[illegible]

Frekans değerlerini elde edebilmek için ftable() fonksiyonu kullanılır.

```
model_veri <- NxNxO %>%  
  dplyr::select(AGE, EDUCATION, MARITAL) %>%  
  ftable() %>%  
  as.data.frame()  
  
model_veri$MARITAL_skor <- as.numeric(factor(model_veri$MARITAL, levels = c("married", "divorced", "single"),  
                                           labels = c(1, 2, 3)))
```

```
> head(model_veri)
```

	AGE	EDUCATION	MARITAL	Freq	MARITAL_skor
1	Genç	primary	divorced	2	2
2	Orta	primary	divorced	3	2
3	Yaşlı	primary	divorced	1	2
4	Genç	secondary	divorced	6	2
5	Orta	secondary	divorced	16	2
6	Yaşlı	secondary	divorced	0	2

Modelimizde kullanılacak veri setimiz yandaki şekildedir.

```

> model <- glm(formula = Freq ~ MARITAL * AGE + MARITAL * EDUCATION, data = model_veri, family = poisson)
> summary(model)

Call:
glm(formula = Freq ~ MARITAL * AGE + MARITAL * EDUCATION, family = poisson,
    data = model_veri)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.73205  -0.49166   0.00038   0.38912   1.49467

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.4055     0.4846   0.837  0.40277
MARITALmarried    2.2734     0.5184   4.385 1.16e-05 ***
MARITALsingle     0.9566     0.6623   1.444  0.14868
AGEOrta           1.0033     0.3525   2.846  0.00442 **
AGEYaşlı         -1.2993     0.6513  -1.995  0.04607 *
EDUCATIONsecondary 1.2993     0.4606   2.821  0.00479 **
EDUCATIONtertiary  0.8473     0.4880   1.736  0.08249 .
EDUCATIONunknown  -1.0986     0.8165  -1.346  0.17846
MARITALmarried:AGEOrta -0.5253     0.3878  -1.355  0.17556
MARITALsingle:AGEOrta -2.2738     0.4520  -5.031 4.89e-07 ***
MARITALmarried:AGEYaşlı  0.2167     0.6986   0.310  0.75643
MARITALsingle:AGEYaşlı -20.1745    3697.3685  -0.005  0.99565
MARITALmarried:EDUCATIONsecondary -0.6061     0.4970  -1.220  0.22262
MARITALsingle:EDUCATIONsecondary  0.7548     0.6616   1.141  0.25392
MARITALmarried:EDUCATIONtertiary -0.7584     0.5316  -1.426  0.15373
MARITALsingle:EDUCATIONtertiary  0.8391     0.6893   1.217  0.22348
MARITALmarried:EDUCATIONunknown -0.7167     0.9126  -0.785  0.43225
MARITALsingle:EDUCATIONunknown  0.1823     1.1690   0.156  0.87607
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 446.673  on 35  degrees of freedom
Residual deviance:  15.661  on 18  degrees of freedom
AIC: 156.36

Number of Fisher Scoring iterations: 17

```

Modelin anlamlılığının test edilmesi için uyum iyiliği istatistiklerinin elde edilmesi gerekmektedir.

```
> LRstats(model)
Likelihood summary table:
      AIC      BIC LR Chisq Df Pr(>Chisq)
model 156.36 184.86  15.661 18    0.6162
```

H_0 : Kısmi ilişki modeline uyum vardır.

H_s : Kısmi ilişki modeline uyum yoktur.

%95 güven düzeyinde $p\text{-value} > 0.05$ olduğundan yokluk hipotezi reddedilmez.

Kısmi ilişki modeline uyum vardır diyebiliriz.

Modele uyum olduğu için satır etki parametreleri, ilişki parametresi üstel değerleri yerel odds oranlarına eşittir.

Parametrelere ait odds oranları:

```
> coef(model)
```

(Intercept)	MARITALmarried	MARITALsingle	AGEOrta
0.4054651	2.2733832	0.9565646	1.0033021
AGEYaşlı	EDUCATIONsecondary	EDUCATIONtertiary	EDUCATIONunknown
-1.2992830	1.2992830	0.8472979	-1.0986123
MARITALmarried:AGEOrta	MARITALsingle:AGEOrta	MARITALmarried:AGEYaşlı	MARITALsingle:AGEYaşlı
-0.5252663	-2.2737647	0.2166710	-20.1745215
MARITALmarried:EDUCATIONsecondary	MARITALsingle:EDUCATIONsecondary	MARITALmarried:EDUCATIONtertiary	MARITALsingle:EDUCATIONtertiary
-0.6061358	0.7548407	-0.7583504	0.8391011
MARITALmarried:EDUCATIONunknown	MARITALsingle:EDUCATIONunknown		
-0.7166777	0.1823216		

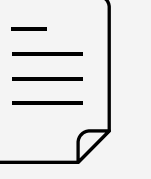
MARITALmarried:	Evli olmanın, bağımlı değişken üzerinde pozitif bir etkisi olduğunu gösterir.
MARITALsingle:	Bekar olmanın, bağımlı değişken üzerinde pozitif bir etkisi olduğunu gösterir.
AGEOrta:	Orta yaş aralığındaki bireylerin, bağımlı değişken üzerinde pozitif bir etkisi olduğunu gösterir.
AGEYaşlı:	Yaşlı bireylerin, bağımlı değişken üzerinde negatif bir etkisi olduğunu gösterir.
EDUCATIONsecondary:	İkincil eğitime sahip olmanın, bağımlı değişken üzerinde pozitif bir etkisi olduğunu gösterir.
EDUCATIONtertiary:	Üçüncül eğitime sahip olmanın, bağımlı değişken üzerinde pozitif bir etkisi olduğunu gösterir.
EDUCATIONunknown:	Eğitim durumu bilinmeyen bireylerin, bağımlı değişken üzerinde negatif bir etkisi olduğunu gösterir.



MARITALmarried:AGEOrta, MARITALsingle:AGEOrta, MARITALmarried:AGEYaşlı, MARITALsingle:AGEYaşlı ve diğer etkileşim terimleri: Bu terimler, iki veya daha fazla bağımsız değişkenin etkileşimini gösterir.

Örneğin, MARITALmarried:AGEOrta terimi, evli olmanın ve orta yaş aralığındaki olmanın birlikte bağımlı değişken üzerindeki etkisini temsil eder.

Sonuç ve Tartışma



Sonuç olarak biz bu çalışmada bir bankanın verilerini kullanarak kişinin eğitim durumunun ev kredisine sahip olup olmaması durumunu incelemeye çalıştık. Bu analiz zarfında çeşitli teknikler ve modeller denenmiş anlamlı ve uygun olanlar belirtilmiştir. Bu çalışmanın temel amaçları, kişilerin banka hesabındaki paranın, ev kredisine sahip olup olmamasına ve eğitim durumuna göre nasıl bir değişim gösterdiğini incelemektir. Bu üç faktör, kişilerin eğitim durumuna göre değişen maddi durumlarının ev kredisine sahip olup olamama durumunu göstermektedir.