Logistik Regresyon

Hakan Mehmetcik

Logisitik Regresyon

Lojistik regresyon, bir bağımlı değişkenin **ikili (binary)** veya **çoklu (multinomial)** kategorilerden oluştuğu durumlarda, bu bağımlı değişken ile bir veya birden fazla bağımsız değişken arasındaki ilişkiyi modellemek için kullanılan bir istatistiksel yöntemdir.

Özellikleri:

1. **Bağımlı Değişken:**
   - İkili lojistik regresyon: Bağımlı değişken iki kategorilidir (ör., 0 ve 1).
   - Çoklu lojistik regresyon: Bağımlı değişken birden fazla kategorilidir (ör., A, B, C).
2. **Bağımsız Değişkenler:**
   - Sürekli veya kategorik olabilir.
   - Örneğin, yaş, cinsiyet, eğitim düzeyi.
3. **Amaç:**
   - Bağımlı değişkenin belirli bir kategoriye ait olma olasılığını tahmin etmek (ör., "Evet" oyu verme olasılığı).
   - Bağımsız değişkenlerin bağımlı değişken üzerindeki etkisini ölçmek.

1. **Logit Fonksiyonu:**
   - Lojistik regresyon, bağımlı değişkenin **log-odds**'larını bağımsız değişkenlerle ilişkilendirir:
   $$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$
     - ppp: Bağımlı değişkenin 1 olma olasılığı.

- 1−p: Bağımlı değişkenin 0 olma olasılığı.
- 0: Sabit terim (intercept).
- i: Bağımsız değişkenlerin katsayıları.

2. **Odds ve Odds Oranı:**

- **Odds:** Bir olayın gerçekleşme olasılığı ile gerçekleşmeme olasılığı arasındaki oran:

$$\text{Odds} = \frac{p}{1-p}$$

- **Odds Oranı (Odds Ratio):** Bir bağımsız değişkendeki bir birimlik artışın bağımlı değişken üzerindeki etkisini gösterir:

$$\text{Odds Ratio} = e^{\beta}$$

R'da Lojistik regresyon modeli

logit_model <- glm(y ~ x1 + x2 + x3, family = binomial, data = my_data)

summary(logit_model)

Örnek 1:

Bir kişinin kredi alıp alamaması (`kredi_onay` = 0 veya 1) ile geliri (`gelir`) ve kredi puanı (`kredi_puani`) arasındaki ilişkiyi incelemek istiyoruz.

```r
# Basit bir veri seti
data <- data.frame(
  kredi_onay = c(0, 1, 0, 1, 0, 1, 1, 0, 1, 0),  # 0 = reddedildi, 1 = onaylandı
  gelir = c(3000, 5000, 2500, 8000, 2000, 7000, 6000, 2200, 9000, 1500),  # Gelir düzeyi
  kredi_puani = c(600, 750, 580, 800, 550, 720, 700, 590, 820, 510)  # Kredi puanı
)

# Veri setini görüntüleme
print(data)
```

```
   kredi_onay gelir kredi_puani
1           0  3000         600
2           1  5000         750
3           0  2500         580
```

```
4           1  8000         800
5           0  2000         550
6           1  7000         720
7           1  6000         700
8           0  2200         590
9           1  9000         820
10          0  1500         510
```

Lojistik Regresyon Modeli

Bağımlı değişken: **kredi_onay**
Bağımsız değişkenler: **gelir** ve **kredi_puani**

```r
# Lojistik regresyon modeli
logit_model <- glm(kredi_onay ~ gelir + kredi_puani, family = binomial, data = data)
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
# Modelin özetini görüntüleme
summary(logit_model)
```

```
Call:
glm(formula = kredi_onay ~ gelir + kredi_puani, family = binomial,
    data = data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.696e+02  8.265e+05       0        1
gelir        8.688e-03  6.538e+01       0        1
kredi_puani  2.009e-01  1.637e+03       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.3863e+01  on 9  degrees of freedom
Residual deviance: 4.2672e-10  on 7  degrees of freedom
AIC: 6

Number of Fisher Scoring iterations: 24
```

Katsayılar, bağımsız değişkenlerin bağımlı değişken üzerindeki etkilerini temsil eder. Ancak burada:

- **Intercept (-169.6):** Çok yüksek bir negatif değerde, ancak bu modelin sorunlu olduğunun bir göstergesi.
- **`gelir` (0.0087):** Katsayı neredeyse sıfır.
- **`kredi_puani` (0.2009):** Katsayı yine anlamlı değil.

Pr(>|z|) sütunundaki değerler:

- Tüm değişkenler için **1**, yani bu değişkenlerin bağımlı değişken üzerindeki etkileri istatistiksel olarak anlamlı değil.
- Bu, modelin bağımsız değişkenleri kullanarak bağımlı değişkeni etkili bir şekilde açıklayamadığını gösteriyor.

Bu model, veri setinizin küçük olması ve bağımsız değişkenlerle bağımlı değişkenin mükemmel ayrışması nedeniyle güvenilir sonuçlar üretememiştir. Şimdi gerçek hayattan bir örnekle logistik regresyon'u inceleyelim.

Örnek 2:

Bu alıştırmada "The Scottish Social Attitudes (SSA) survey" ismiyle bilinen ve 1999 yılından bu yana İskoçya'daki insanların sosyal, politik ve ahlaki tutumlarındaki değişiklikleri takip eden bir anket serisinin 2019 yılı anket sonuçlarını kullanacağız.

Bu anket verisi içinde yer alan değişkenler aşağıdaki şekilde:

- **pserial:** Katılımcıya özgü bir tanımlayıcı.
- **rsex:** Cinsiyet (ör., "Male" veya "Female").
- **rage:** Katılımcının yaş kategorisi (ör., 31, 41 gibi).
- **incsour:** Gelir kaynağı (ör., "Wages/private income source", "State Benefits").
- **leftrigh:** Sol-sağ siyasi eğilim skalası (ör., 1.2, 2.4 gibi sürekli değerler).
- **libauth:** Liberal-otoriter eğilim skalası (ör., 2.833, 4.0 gibi sürekli değerler).
- **employment:** Katılımcının istihdam durumu (ör., "3. Employed", "5. Permanently Sick/Disabled").
- **employmentdum:** İstihdam durumunun ikili kategorik versiyonu (ör., "1. Employed", "NaN").
- **PtyAllgS_NoNa:** Parti bağlılığına ilişkin bir ölçüm (ör., 4.0, 5.0 gibi sürekli değerler).
- **GovTrust_NoNa:** Hükümete güven düzeyine ilişkin bir ölçüm (ör., 3.0, 4.0 gibi sürekli değerler).

- **TaxSpend_NoNa:** Vergi ve harcama politikalarına ilişkin görüşler (ör., 4.0, 5.0 gibi sürekli değerler).
- **ECPolicy_NoNa:** Ekonomi politikalarına ilişkin görüşler (ör., 3.0, 5.0 gibi sürekli değerler).
- **LetIn_NoNa:** Göçmen kabulüne ilişkin görüşler (ör., 4.0, 5.0 gibi sürekli değerler).
- **EvCameron_NoNa:** David Cameron hakkındaki değerlendirmeler (ör., 3.0, 4.0 gibi sürekli değerler).
- **EvSalmond_NoNa:** Alex Salmond hakkındaki değerlendirmeler (ör., 2.0, 4.0 gibi sürekli değerler).
- **Knowind_NoNa:** Bağımsızlık referandumu hakkında bilgi düzeyi (ör., 4.0, 5.0 gibi sürekli değerler).
- **liklyvt_NoNa:** Oy kullanma olasılığı (ör., 10.0).
- **Refvote_NoNa:** Referandumda hangi yönde oy kullanıldığı (ör., 1.0, 3.0 gibi kategoriler).
- **SEBenGB_NoNa:** İskoçya'nın Birleşik Krallık'a ekonomik faydalarına dair görüşler (ör., 1.0, 2.0 gibi sürekli değerler).
- **RefvoteDum:** Referandum oyunun ikili kategorisi (ör., "1. Vote Yes", "0. Vote No").
- **UKSpenGB_NoNa:** Birleşik Krallık harcamalarına ilişkin görüşler (ör., 4.0, 5.0 gibi sürekli değerler).
- **ScotID_NoNa:** İskoç kimliğine dair görüşler (ör., 6.0, 7.0 gibi sürekli değerler).
- **HEdQual2_NoNa:** Katılımcının eğitim düzeyi (ör., 4.0, 8.0 gibi kategoriler).
- **Party_Labels:** Katılımcının siyasi parti tercihi (ör., "2. Labour").
- **PartyFW_NoNa:** Siyasi partilere yönelik görüşlerin bir ölçütü (ör., 2.0, 3.0 gibi sürekli değerler).
- **Dole_NoNa:** İşsizlik yardımlarına dair görüşler (ör., 1.0, 2.0 gibi sürekli değerler)

1. Veri Setini Okuma

```
library(haven)
library(here)
```

```
here() starts at /Users/kobain/Desktop/IST2083
```

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2

-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```r
# Veri setini okuma
scottish <- read_dta((here("data", "scottish.dta")))

# İlk birkaç satıra göz atalım
head(scottish)
```

```
# A tibble: 6 x 26
  pserial rsex       rage      incsour leftrigh libauth employment employmentdum
    <dbl> <dbl+lbl>  <dbl+lbl> <dbl+l> <dbl+lb> <dbl+l> <dbl+lbl>  <dbl+lbl>
1  151343 1 [Male]   31        2 [Wag~ 1.2      2.83    3 [3. Emp~  1 [1. Emplo~
2  151856 2 [Female] 41        2 [Wag~ 1.8      3.5     3 [3. Emp~  1 [1. Emplo~
3  151537 1 [Male]   53        2 [Wag~ 2        4       3 [3. Emp~  1 [1. Emplo~
4  151369 1 [Male]   39        1 [Sta~ 1.8      4.67    5 [5. Per~ NA
5  152010 2 [Female] 43        1 [Sta~ 2.4      3.83    7 [7. Hom~ NA
6  151793 2 [Female] 60        2 [Wag~ 1.8      3.83    3 [3. Emp~  1 [1. Emplo~
# i 18 more variables: PtyAllgS_NoNa <dbl>, GovTrust_NoNa <dbl>,
#   TaxSpend_NoNa <dbl>, ECPolicy_NoNa <dbl>, LetIn_NoNa <dbl>,
#   EvCameron_NoNa <dbl>, EvSalmond_NoNa <dbl>, Knowind_NoNa <dbl>,
#   liklyvt_NoNa <dbl>, Refvote_NoNa <dbl>, SEBenGB_NoNa <dbl>,
#   RefvoteDum <dbl+lbl>, UKSpenGB_NoNa <dbl>, ScotID_NoNa <dbl>,
#   HEdQual2_NoNa <dbl>, Party_Labels <dbl+lbl>, PartyFW_NoNa <dbl>,
#   Dole_NoNa <dbl>
```

2. Veri Setinin Genel Yapısını Anlama

```r
# Veri setinin yapısını inceleme
glimpse(scottish)
```

```
Rows: 1,501
Columns: 26
$ pserial        <dbl> 151343, 151856, 151537, 151369, 152010, 151793, 151589,~
$ rsex           <dbl+lbl> 1, 2, 1, 1, 2, 2, 2, 1, 1, 2, 2, 2, 2, 1, 2, 2, 2, ~
$ rage           <dbl+lbl> 31, 41, 53, 39, 43, 60, 86, 49, 43, 53, 53, 49, 42,~
$ incsour        <dbl+lbl>  2,  2,  2,  1,  1,  2,  1,  2,  1,  2,  2,  2,  2,~
$ leftrigh       <dbl+lbl>  1.2,  1.8,  2.0,  1.8,  2.4,  1.8,  1.0,  2.4,  2.~
$ libauth        <dbl+lbl>  2.833333,  3.500000,  4.000000,  4.666667,  3.8333~
$ employment     <dbl+lbl> 3, 3, 3, 5, 7, 3, 6, 3, 4, 3, 3, 3, 3, 3, 4, 3, 3, ~
$ employmentdum  <dbl+lbl>  1,  1,  1, NA, NA,  1, NA,  1,  0,  1,  1,  1,  1,~
$ PtyAllgS_NoNa  <dbl> 5, 5, 4, 5, NA, NA, 4, 4, 12, 12, NA, 10, 11, 11, NA, 1~
$ GovTrust_NoNa  <dbl> 4, 3, 4, 4, 4, NA, 4, 3, 4, 4, 4, 4, 4, 3, 2, 1, 3, 3, ~
$ TaxSpend_NoNa  <dbl> 2, 3, 2, 3, 2, 3, 3, 3, NA, 3, NA, 3, 3, 3, 3, 2, 2, 1,~
$ ECPolicy_NoNa  <dbl> 5, NA, 4, 1, 2, 2, 2, 2, 2, 2, NA, 1, 1, 1, 5, 3, 1, NA~
$ LetIn_NoNa     <dbl> 5, 4, 5, 5, 1, 5, 5, 3, 5, 5, 5, 5, 5, 4, 4, 3, 5, 5, 5~
$ EvCameron_NoNa <dbl> 2, 2, 5, 0, 4, 5, 0, 5, 0, 0, 0, 0, 2, 3, 0, 9, 1, 5, 7~
$ EvSalmond_NoNa <dbl> 5, 4, 3, 5, 5, 6, 0, 1, 6, 0, 0, 7, 5, 7, NA, 10, 1, 7,~
$ Knowind_NoNa   <dbl> 4, 4, 2, 3, 4, 4, 5, 3, 3, 4, 5, 3, 3, 3, 4, 4, 3, 4, 5~
$ liklyvt_NoNa   <dbl> 10, 10, 10, 10, 10, 7, 10, 10, 10, 0, 1, 10, 10, 10, 5,~
$ Refvote_NoNa   <dbl> 1, 3, 2, 1, 3, 3, 2, 2, 3, 2, 3, 1, 1, NA, 2, 3, 2, 3, ~
$ SEBenGB_NoNa   <dbl> 1, 1, 1, 1, 1, 1, 1, 2, 1, NA, 1, 1, 1, 2, 3, NA, 1, 1,~
$ RefvoteDum     <dbl+lbl>  1, NA,  0,  1, NA, NA,  0,  0, NA,  0, NA,  1,  1,~
$ UKSpenGB_NoNa  <dbl> 5, 4, 4, 5, 5, NA, 5, 3, 5, 5, 5, 4, 4, 5, NA, 3, NA, 4~
$ ScotID_NoNa    <dbl> 7, 6, 4, 6, 7, 7, 7, 7, 7, 7, 7, 7, 4, 7, 1, 4, 7, 7, 7~
$ HEdQual2_NoNa  <dbl> 4, 8, 8, 5, NA, 8, 8, 4, 8, 8, 8, 3, 5, 8, 8, 8, 4, 5, ~
$ Party_Labels   <dbl+lbl>  2,  2,  2,  2, NA, NA,  2,  2,  4,  4, NA,  4,  4,~
$ PartyFW_NoNa   <dbl> 2, 2, 2, 2, NA, NA, 2, 2, 4, 4, 10, 4, 4, 4, NA, 4, NA,~
$ Dole_NoNa      <dbl> 1, NA, 2, 1, 1, NA, 2, 2, 1, NA, 3, 1, NA, 3, NA, 2, NA~
```

```r
# Sütun isimlerini ve sayısını kontrol etme
colnames(scottish)
```

```
 [1] "pserial"        "rsex"           "rage"           "incsour"
 [5] "leftrigh"       "libauth"        "employment"     "employmentdum"
 [9] "PtyAllgS_NoNa"  "GovTrust_NoNa"  "TaxSpend_NoNa"  "ECPolicy_NoNa"
[13] "LetIn_NoNa"     "EvCameron_NoNa" "EvSalmond_NoNa" "Knowind_NoNa"
[17] "liklyvt_NoNa"   "Refvote_NoNa"   "SEBenGB_NoNa"   "RefvoteDum"
[21] "UKSpenGB_NoNa"  "ScotID_NoNa"    "HEdQual2_NoNa"  "Party_Labels"
```

```
[25] "PartyFW_NoNa"   "Dole_NoNa"
```

Temel veri manüplasyonu:

```
scottish <- scottish %>%
  mutate(refvote = as_factor(RefvoteDum, levels="labels"),
                 pid = as_factor(Party_Labels, levels="labels"),
                 scot = ScotID_NoNa,
                 trust = as.numeric(fct_rev(as_factor(GovTrust_NoNa))),
                 age = rage,
                 edu1 = na_if(HEdQual2_NoNa,7),#this is a filler step
                 edu2 = recode(edu1, "8='7'"),
                 edu = as.numeric(fct_rev(as_factor(edu2)))) |>
  mutate(pid1 = recode_factor(pid,
          "1. Conservative" = "1. Conservative",
          "2. Labour" = "2. Labour",
          "4. SNP" = "3. SNP",
          .default = NA_character_)) |>
  mutate(trust_ordfac = ordered(as_factor(trust)))
```

```
Warning: There was 1 warning in `mutate()`.
i In argument: `edu2 = recode(edu1, "8='7'")`.
Caused by warning:
! Unreplaced values treated as NA as `.x` is not compatible.
Please specify replacements exhaustively or supply `.default`.
```

3. Eksik Verilerin Kontrolü

```
# Eksik veri yüzdesini hesaplama
scottish %>%
  summarise(across(everything(), ~ mean(is.na(.)) * 100)) %>%
  pivot_longer(cols = everything(), names_to = "Column", values_to = "Missing_Percentage") %>
  arrange(desc(Missing_Percentage))
```

```
# A tibble: 36 x 2
   Column        Missing_Percentage
   <chr>                      <dbl>
 1 edu2                        94.3
 2 edu                         94.3
```

```
 3 employmentdum               42.5
 4 pid1                        33.6
 5 RefvoteDum                  31.1
 6 refvote                     31.1
 7 PtyAllgS_NoNa               28.9
 8 Party_Labels                23.7
 9 pid                         23.7
10 PartyFW_NoNa                11.0
# i 26 more rows
```

Eksik Verilerin İşlenmesi

Eksik veriler için temel stratejiler:

- Eksik verileri silme:
- Eksik verileri doldurma (ör:, sütun medyanı/ortalaması ile)

Biz ikinci yöntemi kullanalım:

```
scottish_clean <- scottish %>%
  mutate(across(where(is.numeric), ~ ifelse(is.na(.), median(., na.rm = TRUE), .)))
```

Hala bazı verilerde "Not applicable" ve "Missing Values" ibareleri var. Bunların da temizlenmesi ya da ilgili değişkenin kullanımında farklı tekniklerin kullanılması gerekiyor.

```
glimpse(scottish_clean)
```

```
Rows: 1,501
Columns: 36
$ pserial        <dbl> 151343, 151856, 151537, 151369, 152010, 151793, 151589,~
$ rsex           <dbl> 1, 2, 1, 1, 2, 2, 2, 1, 1, 2, 2, 2, 2, 1, 2, 2, 2, 1, 1~
$ rage           <dbl> 31, 41, 53, 39, 43, 60, 86, 49, 43, 53, 53, 49, 42, 50,~
$ incsour        <dbl> 2, 2, 2, 1, 1, 2, 1, 2, 1, 2, 2, 2, 2, 2, 1, 2, 1, 1, 1~
$ leftrigh       <dbl> 1.2, 1.8, 2.0, 1.8, 2.4, 1.8, 1.0, 2.4, 2.0, 2.4, 1.0, ~
$ libauth        <dbl> 2.833333, 3.500000, 4.000000, 4.666667, 3.833333, 3.833~
$ employment     <dbl> 3, 3, 3, 5, 7, 3, 6, 3, 4, 3, 3, 3, 3, 3, 4, 3, 3, 6, 4~
$ employmentdum  <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0~
$ PtyAllgS_NoNa  <dbl> 5, 5, 4, 5, 6, 6, 4, 4, 12, 12, 6, 10, 11, 11, 6, 12, 6~
$ GovTrust_NoNa  <dbl> 4, 3, 4, 4, 4, 3, 4, 3, 4, 4, 4, 4, 4, 3, 2, 1, 3, 3, 4~
$ TaxSpend_NoNa  <dbl> 2, 3, 2, 3, 2, 3, 3, 3, 2, 3, 2, 3, 3, 3, 3, 2, 2, 1, 3~
$ ECPolicy_NoNa  <dbl> 5, 2, 4, 1, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 5, 3, 1, 2, 2~
$ LetIn_NoNa     <dbl> 5, 4, 5, 5, 1, 5, 5, 3, 5, 5, 5, 5, 5, 4, 4, 3, 5, 5, 5~
```

```
$ EvCameron_NoNa <dbl> 2, 2, 5, 0, 4, 5, 0, 5, 0, 0, 0, 0, 2, 3, 0, 9, 1, 5, 7~
$ EvSalmond_NoNa <dbl> 5, 4, 3, 5, 5, 6, 0, 1, 6, 0, 0, 7, 5, 7, 5, 10, 1, 7, ~
$ Knowind_NoNa   <dbl> 4, 4, 2, 3, 4, 4, 5, 3, 3, 4, 5, 3, 3, 3, 4, 4, 3, 4, 5~
$ liklyvt_NoNa   <dbl> 10, 10, 10, 10, 10, 7, 10, 10, 10, 0, 1, 10, 10, 10, 5,~
$ Refvote_NoNa   <dbl> 1, 3, 2, 1, 3, 3, 2, 2, 3, 2, 3, 1, 1, 2, 2, 3, 2, 3, 3~
$ SEBenGB_NoNa   <dbl> 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 1, 1, 1, 2, 3, 2, 1, 1, 3~
$ RefvoteDum     <dbl> 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0~
$ UKSpenGB_NoNa  <dbl> 5, 4, 4, 5, 5, 3, 5, 3, 5, 5, 5, 4, 4, 5, 3, 3, 3, 4, 4~
$ ScotID_NoNa    <dbl> 7, 6, 4, 6, 7, 7, 7, 7, 7, 7, 7, 7, 4, 7, 1, 4, 7, 7, 7~
$ HEdQual2_NoNa  <dbl> 4, 8, 8, 5, 4, 8, 8, 4, 8, 8, 8, 3, 5, 8, 8, 8, 4, 5, 8~
$ Party_Labels   <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 4, 4, 2, 4, 4, 4, 2, 4, 2, 4, 2~
$ PartyFW_NoNa   <dbl> 2, 2, 2, 2, 3, 3, 2, 2, 4, 4, 10, 4, 4, 4, 3, 4, 3, 4, ~
$ Dole_NoNa      <dbl> 1, 2, 2, 1, 1, 2, 2, 2, 1, 2, 3, 1, 2, 3, 2, 2, 2, 3, 1~
$ refvote        <fct> 1. Vote Yes, NA, 0. Vote No, 1. Vote Yes, NA, NA, 0. Vo~
$ pid             <fct> 2. Labour, 2. Labour, 2. Labour, 2. Labour, NA, NA, 2. ~
$ scot           <dbl> 7, 6, 4, 6, 7, 7, 7, 7, 7, 7, 7, 7, 4, 7, 1, 4, 7, 7, 7~
$ trust          <dbl> 1, 2, 1, 1, 1, 2, 1, 2, 1, 1, 1, 1, 1, 2, 3, 4, 2, 2, 1~
$ age            <dbl> 31, 41, 53, 39, 43, 60, 86, 49, 43, 53, 53, 49, 42, 50,~
$ edu1           <dbl> 4, 8, 8, 5, 4, 8, 8, 4, 8, 8, 8, 3, 5, 8, 8, 8, 4, 5, 8~
$ edu2           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ edu             <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ pid1           <fct> 2. Labour, 2. Labour, 2. Labour, 2. Labour, NA, NA, 2. ~
$ trust_ordfac   <ord> 1, 2, 1, 1, 1, NA, 1, 2, 1, 1, 1, 1, 1, 2, 3, 4, 2, 2, ~
```

4. Temel Özet İstatistikler

Analiz yapmadan önce, hangi sütunların kategorik veya sürekli olduğunu anlamak önemlidir. Bu bilgi dönüşüm veya filtreleme işlemlerinde yardımcı olur.

```r
# Sayısal değişkenleri seçme
numeric_vars <- scottish_clean %>%
  select(where(is.numeric))

# Sayısal değişkenlerin isimleri
colnames(numeric_vars)
```

```
 [1] "pserial"        "rsex"           "rage"           "incsour"
 [5] "leftrigh"       "libauth"        "employment"     "employmentdum"
 [9] "PtyAllgS_NoNa"  "GovTrust_NoNa"  "TaxSpend_NoNa"  "ECPolicy_NoNa"
[13] "LetIn_NoNa"     "EvCameron_NoNa" "EvSalmond_NoNa" "Knowind_NoNa"
[17] "liklyvt_NoNa"   "Refvote_NoNa"   "SEBenGB_NoNa"   "RefvoteDum"
```

```
[21] "UKSpenGB_NoNa"  "ScotID_NoNa"    "HEdQual2_NoNa"  "Party_Labels"
[25] "PartyFW_NoNa"   "Dole_NoNa"      "scot"           "trust"
[29] "age"            "edu1"           "edu"
```

```r
# Kategorik değişkenleri seçme
categorical_vars <- scottish_clean %>%
  select(where(is.factor))

# Kategorik değişkenlerin isimleri
colnames(categorical_vars)
```

```
[1] "refvote"      "pid"          "pid1"         "trust_ordfac"
```

5. Sayısal Değişkenler için Dağılım ve Özet İstatistikler:

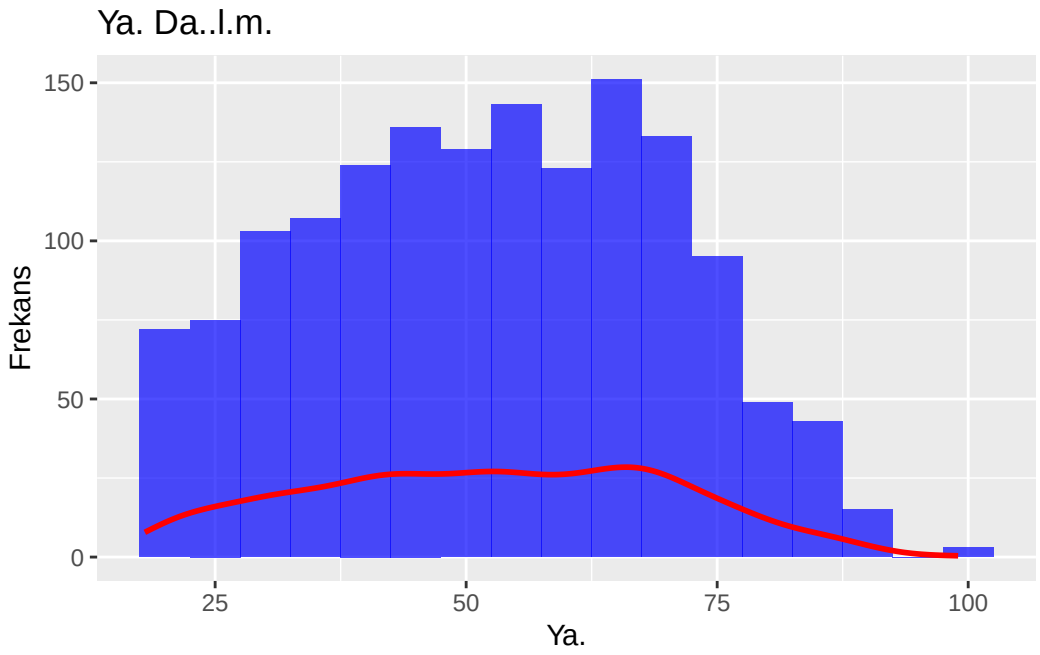
```r
# Temel istatistikler
scottish_clean %>%
  summarise(
    mean_age = mean(age, na.rm = TRUE),
    mean_trust = mean(trust, na.rm = TRUE),
    mean_edu = mean(edu, na.rm = TRUE),
    sd_age = sd(age, na.rm = TRUE),
    sd_trust = sd(trust, na.rm = TRUE),
    sd_edu = sd(edu, na.rm = TRUE)
  )
```

```
# A tibble: 1 x 6
  mean_age mean_trust mean_edu sd_age sd_trust sd_edu
     <dbl>      <dbl>    <dbl>  <dbl>    <dbl>  <dbl>
1     52.3       1.76        1   17.8    0.736      0
```

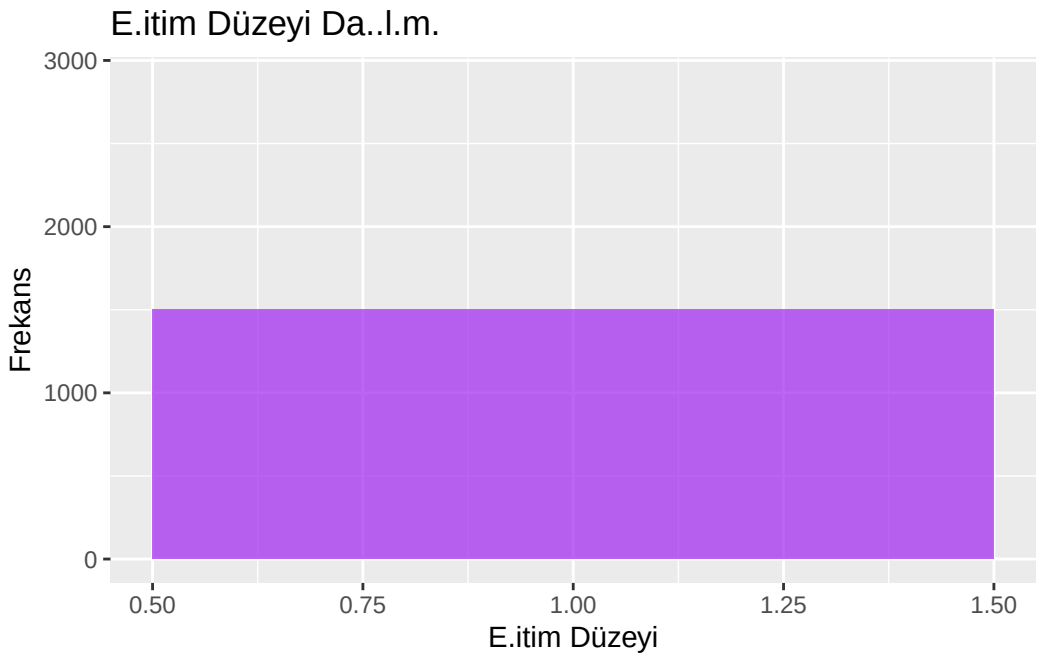
```r
library(ggplot2)

# Yaş histogramı ve yoğunluk grafiği
ggplot(scottish_clean, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "blue", alpha = 0.7) +
  geom_density(aes(y = ..count..), color = "red", size = 1) +
  labs(title = "Yaş Dağılımı", x = "Yaş", y = "Frekans")
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

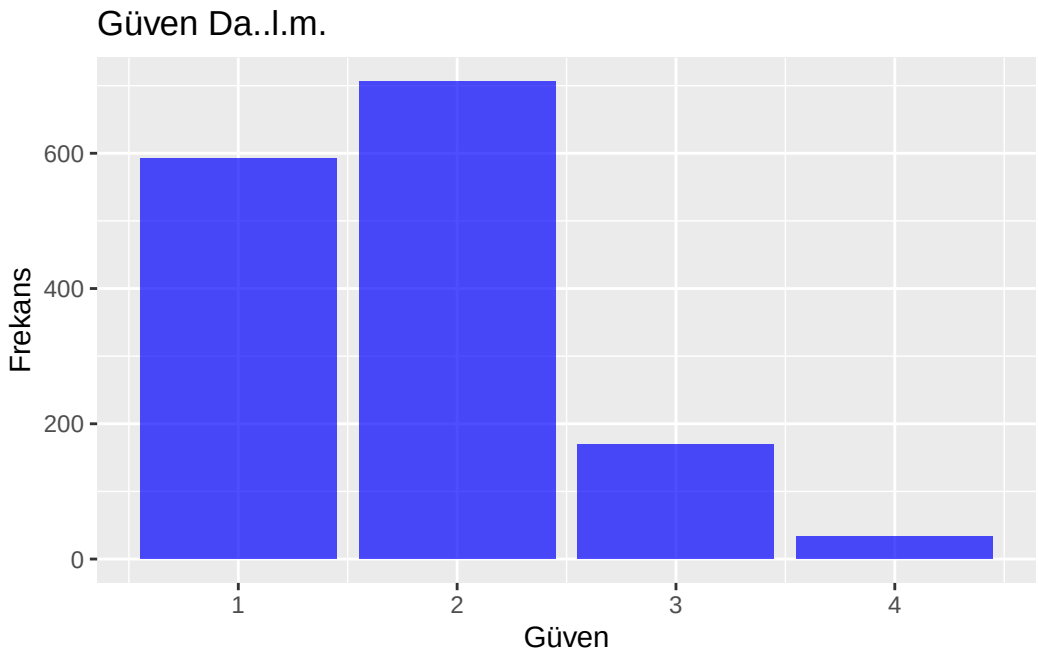
```
Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
i Please use `after_stat(count)` instead.
```



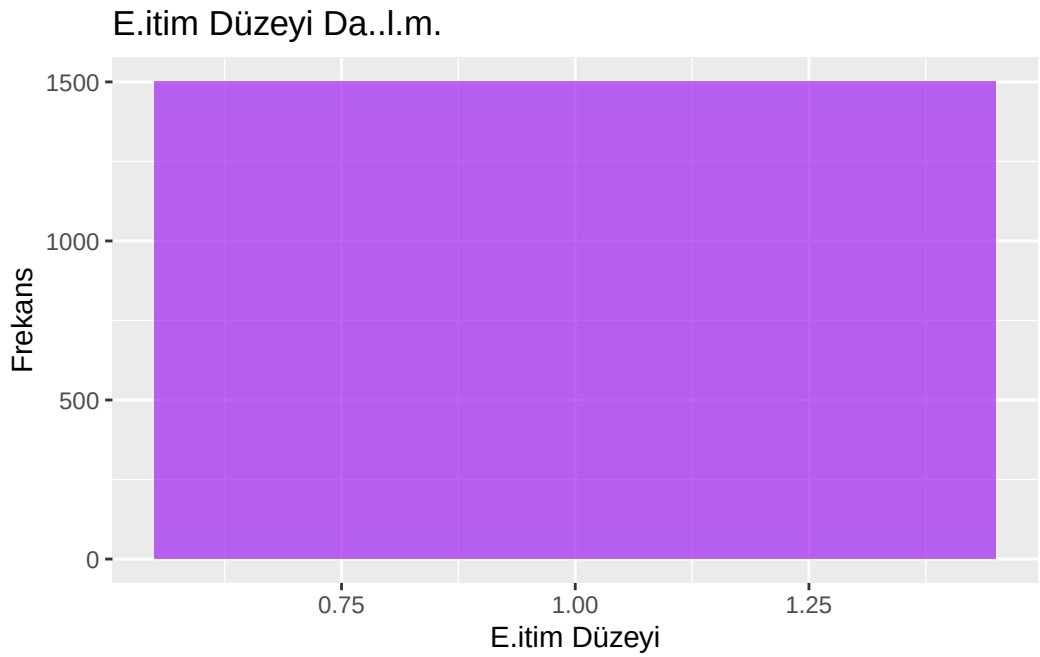

```r
# Eğitim histogramı ve yoğunluk grafiği
ggplot(scottish_clean, aes(x = edu)) +
  geom_histogram(binwidth = 1, fill = "purple", alpha = 0.7) +
  geom_density(aes(y = ..count..), color = "red", size = 1) +
  labs(title = "Eğitim Düzeyi Dağılımı", x = "Eğitim Düzeyi", y = "Frekans")
```




```
library(ggplot2)

# Güven değişkeni bar grafiği
ggplot(scottish_clean, aes(x = trust)) +
  geom_bar(fill = "blue", alpha = 0.7) +
  labs(title = "Güven Dağılımı", x = "Güven", y = "Frekans")
```




```
# Eğitim değişkeni bar grafiği
ggplot(scottish_clean, aes(x = edu)) +
  geom_bar(fill = "purple", alpha = 0.7) +
  labs(title = "Eğitim Düzeyi Dağılımı", x = "Eğitim Düzeyi", y = "Frekans")
```




```
# Spearman korelasyonu
scottish_clean %>%
  select(trust, edu, age) %>%
  cor(method = "spearman", use = "complete.obs")
```

```
Warning in cor(., method = "spearman", use = "complete.obs"): the standard
deviation is zero
```

```
             trust edu         age
trust  1.00000000  NA -0.01071422
edu            NA   1          NA
age   -0.01071422  NA  1.00000000
```

```
# Kendall tau korelasyonu
scottish_clean %>%
  select(trust, edu, age) %>%
  cor(method = "kendall", use = "complete.obs")
```

```
Warning in cor(., method = "kendall", use = "complete.obs"): the standard
deviation is zero
```

```
              trust edu          age
trust  1.000000000  NA -0.008841251
edu              NA   1           NA
age   -0.008841251  NA  1.000000000
```

```
scottish_clean %>%
  group_by(edu) %>%
  summarise(mean_trust = mean(trust, na.rm = TRUE))
```

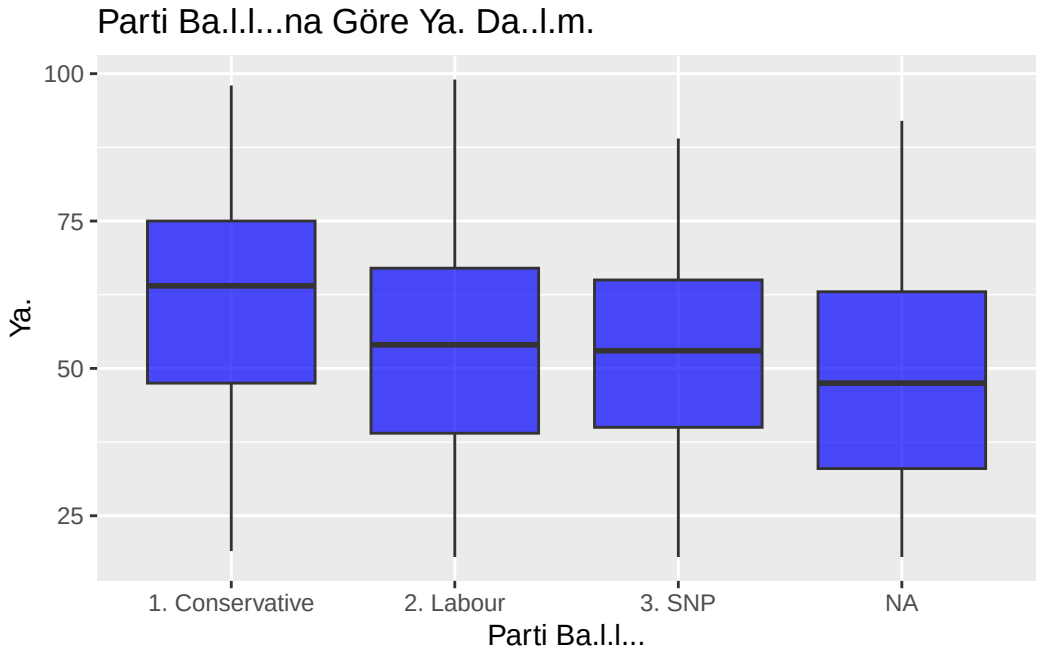
```
# A tibble: 1 x 2
    edu mean_trust
  <dbl>      <dbl>
1     1       1.76
```

- Güven (`trust`) ile eğitim (`edu`) arasında pozitif bir ilişki var, ancak bu ilişki oldukça zayıf.
- Eğitim seviyesi arttıkça güven düzeyinin biraz artma eğiliminde olduğu söylenebilir, ancak bu etki istatistiksel olarak anlamlı olmayabilir.
- Güven (`trust`) ile yaş (`age`) arasında çok zayıf ve neredeyse sıfır bir ilişki var.
- Yaşın, bireylerin güven düzeyine belirgin bir etkisi olmadığı söylenebilir.

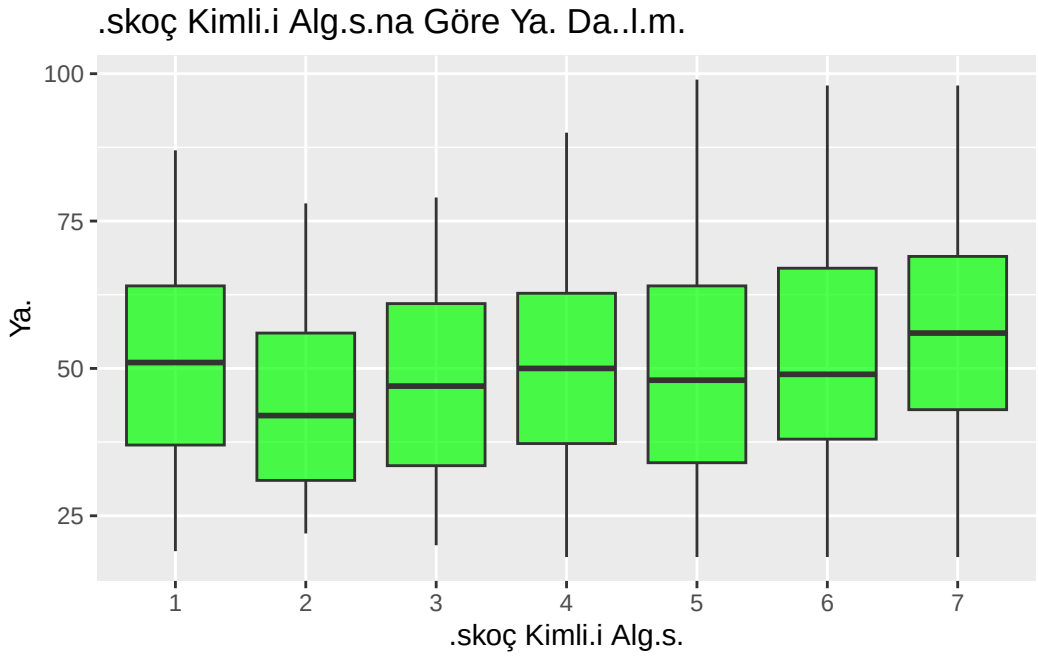
```
# Parti bağlılığına göre yaş ortalamaları
scottish_clean %>%
  group_by(pid1) %>%
  summarise(
    mean_age = mean(age, na.rm = TRUE),
    sd_age = sd(age, na.rm = TRUE),
    count = n()
  )
```

```
# A tibble: 4 x 4
  pid1            mean_age sd_age count
  <fct>              <dbl>  <dbl> <int>
1 1. Conservative     60.7   17.2   191
2 2. Labour           53.6   17.7   443
3 3. SNP              51.5   16.1   363
4 <NA>                48.5   18.2   504
```

```r
ggplot(scottish_clean, aes(x = pid1, y = age)) +
  geom_boxplot(fill = "blue", alpha = 0.7) +
  labs(title = "Parti Bağlılığına Göre Yaş Dağılımı", x = "Parti Bağlılığı", y = "Yaş")
```




```r
ggplot(scottish_clean, aes(x = factor(ScotID_NoNa), y = age)) +
  geom_boxplot(fill = "green", alpha = 0.7) +
  labs(title = "İskoç Kimliği Algısına Göre Yaş Dağılımı", x = "İskoç Kimliği Algısı", y = "
```

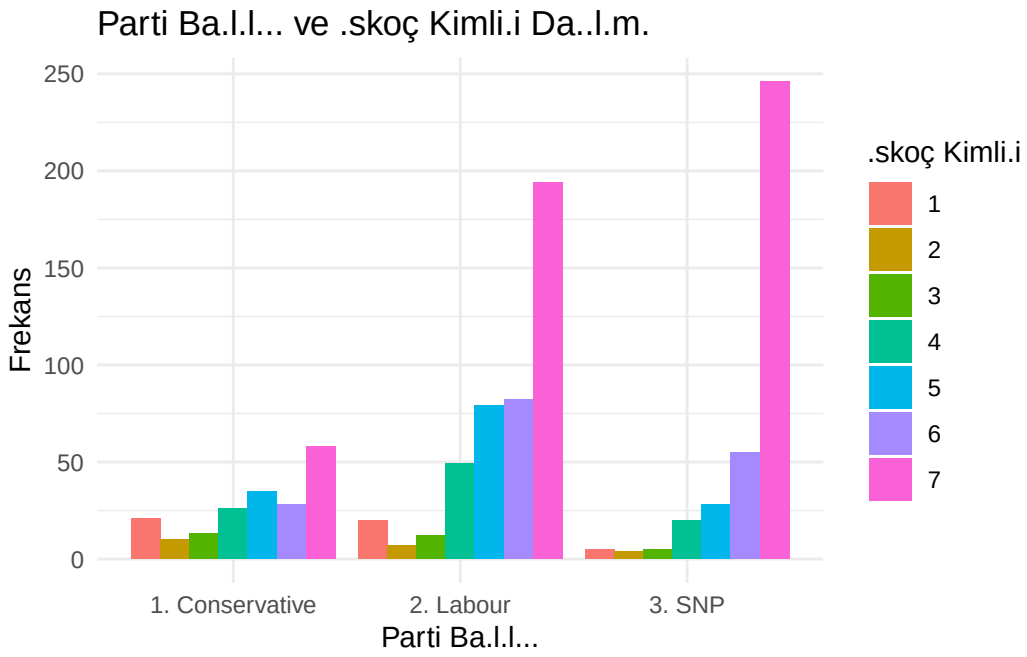



```r
# Parti bağlılığı ve İskoç kimliği algısı arasındaki ilişki
# Çapraz tablo
cross_table <- table(scottish_clean$pid1, scottish_clean$ScotID_NoNa)

# Satır ve sütun toplamlarını ekleme
cross_table_with_totals <- addmargins(cross_table)

# Yüzde dağılımları (satır yüzdesi)
cross_table_row_percent <- prop.table(cross_table, margin = 1) * 100

# Yüzde dağılımları (sütun yüzdesi)
cross_table_col_percent <- prop.table(cross_table, margin = 2) * 100
```

```r
library(dplyr)
library(tidyr)

# Çapraz tablo
cross_table <- table(scottish_clean$pid1, scottish_clean$ScotID_NoNa)

# Çapraz tabloyu bir veri çerçevesine dönüştürme
cross_table_df <- as.data.frame(cross_table)

# Satır toplamlarını ekleme
```

```r
cross_table_totals <- cross_table_df %>%
  group_by(Var1) %>%
  summarise(
    Total = sum(Freq)
  )

# Sütun toplamlarını ekleme
column_totals <- cross_table_df %>%
  group_by(Var2) %>%
  summarise(
    Total = sum(Freq)
  )

# Çapraz tabloyu görselleştirme için düzenleme
combined_table <- cross_table_df %>%
  spread(key = Var2, value = Freq) %>%
  left_join(cross_table_totals, by = c("Var1" = "Var1"))
```

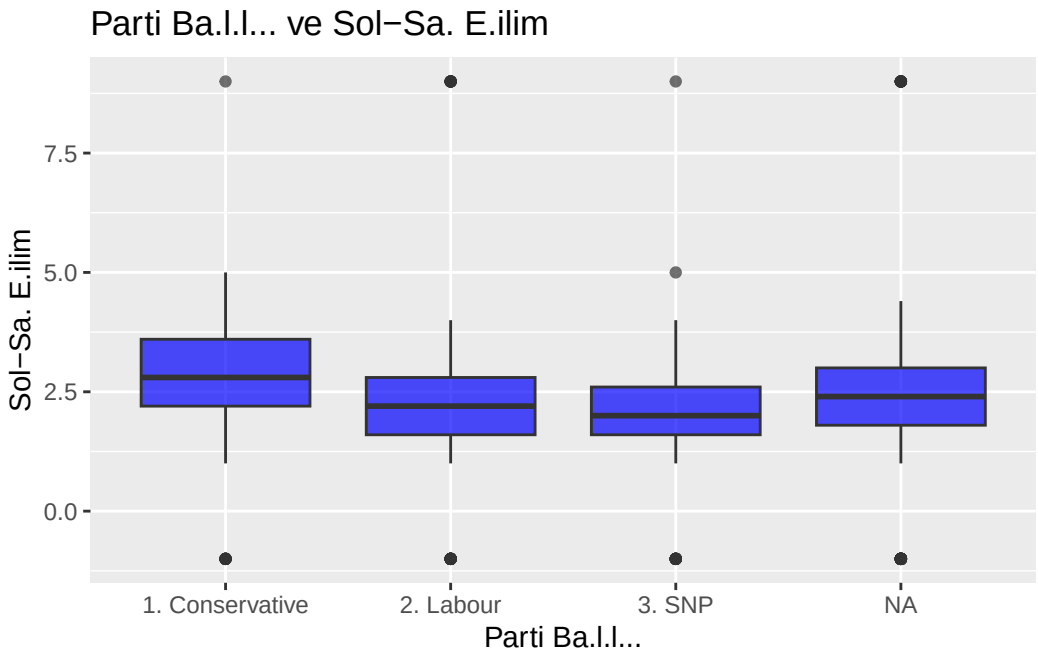
```r
ggplot(cross_table_df, aes(x = Var1, y = Freq, fill = Var2)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "Parti Bağlılığı ve İskoç Kimliği Dağılımı",
    x = "Parti Bağlılığı",
    y = "Frekans",
    fill = "İskoç Kimliği"
  ) +
  theme_minimal()
```



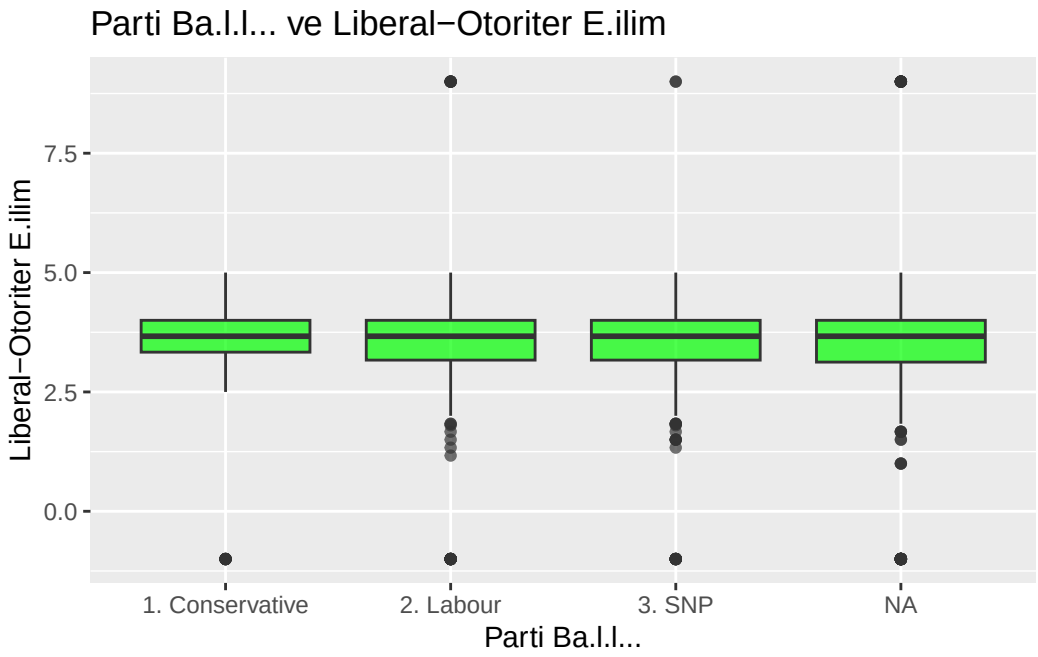

```
# Parti bağlılığına göre ortalamalar
scottish_clean %>%
  group_by(pid1) %>%
  summarise(
    mean_leftrigh = mean(leftrigh, na.rm = TRUE),
    mean_libauth = mean(libauth, na.rm = TRUE),
    count = n()
  )
```

```
# A tibble: 4 x 4
  pid1            mean_leftrigh mean_libauth count
  <fct>                   <dbl>        <dbl> <int>
1 1. Conservative          2.75         3.48   191
2 2. Labour                2.22         3.47   443
3 3. SNP                   2.01         3.43   363
4 <NA>                     2.36         3.42   504
```

```
library(ggplot2)

# Parti bağlılığına göre leftrigh kutu grafiği
ggplot(scottish_clean, aes(x = pid1, y = leftrigh)) +
  geom_boxplot(fill = "blue", alpha = 0.7) +
  labs(title = "Parti Bağlılığı ve Sol-Sağ Eğilim", x = "Parti Bağlılığı", y = "Sol-Sağ Eğili
```




```
# Parti bağlılığına göre libauth kutu grafiği
ggplot(scottish_clean, aes(x = pid1, y = libauth)) +
  geom_boxplot(fill = "green", alpha = 0.7) +
  labs(title = "Parti Bağlılığı ve Liberal-Otoriter Eğilim", x = "Parti Bağlılığı", y = "Libe
```



employmentdum (çalışma durumu: 0 = işsiz, 1 = çalışıyor) ve **edu** (eğitim düzeyi) arasında bir ilişki olup olmadığını test edebiliriz.

```
# Satır yüzdeleri
prop.table(table(scottish_clean$employmentdum, scottish_clean$edu), margin = 1) * 100
```

```
   1
0 100
1 100
```

Genel olarak işsizler, daha düşük eğitim seviyelerinde yoğunlaşıyor. İşsiz bireylerin en büyük oranı **3. Eğitim Düzeyinde (%29.2)** ve ardından **1. Eğitim Düzeyinde (%25)**.

Çalışanlar, her eğitim düzeyinde temsil edilirken, daha düşük seviyelerden yüksek seviyelere doğru azalan bir eğilim gösteriyor.

```
# Ki-kare testi
chi_test <- chisq.test(table(scottish_clean$employmentdum, scottish_clean$edu))
chi_test
```

```
    Chi-squared test for given probabilities

data:  table(scottish_clean$employmentdum, scottish_clean$edu)
X-squared = 1226.8, df = 1, p-value < 2.2e-16
```

- **Null Hipotez (H):** `employmentdum` (çalışma durumu) ile `edu` (eğitim düzeyi) arasında bir ilişki yoktur. Bu iki değişken birbirinden bağımsızdır.
- **Alternatif Hipotez (H):** `employmentdum` ile `edu` arasında bir ilişki vardır. Bu iki değişken bağımsız değildir.
- **p-değeri < 0.05** olduğundan, null hipotezi reddediyoruz.
- Bu, **çalışma durumu** (`employmentdum`) ile **eğitim düzeyi** (`edu`) arasında **istatistiksel olarak anlamlı bir ilişki olduğunu** gösterir.
- Eğitim düzeyindeki değişim, bireylerin çalışma durumunu etkileyebilir veya en azından bu iki değişken bağımsız değildir.

Bağımlı değişkenin (**`RefvoteDum`**) diğer bağımsız değişkenler tarafından nasıl etkilendiğini modellemek için bir **lojistik regresyon modeli** kullanabiliriz. Çünkü bağımlı değişkenimiz ikili (binary: 0 = Hayır, 1 = Evet) bir değişkendir.

- Bağımsız değişkenler: **`pid1` (parti bağlılığı), `edu` (eğitim düzeyi), `age` (yaş), `rsex` (cinsiyet), `libauth` (liberal-otoriter eğilim), `employmentdum` (istihdam durumu)**.

```r
# Lojistik regresyon modeli
logit_model <- glm(RefvoteDum ~ pid1 + edu + age + rsex + libauth + employmentdum,
                   data = scottish_clean, family = binomial)

# Modelin özetini görüntüleme
summary(logit_model)
```

```
Call:
glm(formula = RefvoteDum ~ pid1 + edu + age + rsex + libauth +
    employmentdum, family = binomial, data = scottish_clean)

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.664161   0.704507  -2.362 0.018169 *
pid12. Labour  1.921618   0.474518   4.050 5.13e-05 ***
pid13. SNP     4.040473   0.469735   8.602  < 2e-16 ***
edu                 NA         NA      NA       NA
```

```
age           -0.004490   0.004958  -0.906 0.365165
rsex          -0.636351   0.166844  -3.814 0.000137 ***
libauth       -0.148647   0.065965  -2.253 0.024232 *
employmentdum -0.259014   0.379392  -0.683 0.494791
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1207.55  on 996  degrees of freedom
Residual deviance:  894.82  on 990  degrees of freedom
  (504 observations deleted due to missingness)
AIC: 908.82

Number of Fisher Scoring iterations: 6
```

1. Modelin Genel Değerlendirmesi

Null deviance ile residual deviance arasındaki fark modelin açıklama gücünü gösterir. Modelin residual deviance'ındaki önemli düşüş, bağımsız değişkenlerin bağımlı değişkeni açıklamada etkili olduğunu gösterir.

2. Bağımsız Değişkenlerin Yorumu

Her bir bağımsız değişkenin **katsayı (Estimate)**, **p-değeri (Pr(>|z|))** ve anlamlılığına göre yorumları:

Intercept (Sabit Terim):

- **Katsayı = -1.760**, **p = 0.020**
  - Tüm bağımsız değişkenlerin etkisi sıfır olduğunda "Evet" oyu verme olasılığı negatif (log-odds olarak).

`pid1` (Parti Bağlılığı):

- **Labour (Katsayı = 1.932, p < 0.001):**
  - Labour Partisi'ne bağlılık, "Evet" oyu verme olasılığını artırır.
  - Odds oranı: $e^{1.932} \approx 6.91$, yani Labour'a bağlı bireylerin "Evet" oyu verme olasılığı, Conservative'e bağlı bireylere kıyasla yaklaşık 7 kat daha fazladır.

- **SNP (Katsayı = 4.051, p < 0.001):**
  - SNP'ye bağlılık, "Evet" oyu verme olasılığını çok güçlü bir şekilde artırır.
  - Odds oranı: $e^{4.051} \approx 57.4$, yani SNP'ye bağlı bireylerin "Evet" oyu verme olasılığı Conservative'e bağlı bireylerden yaklaşık 57 kat daha fazladır.

`edu` (Eğitim Düzeyi):

- **Katsayı = 0.0167, p = 0.729:**
  - Eğitim düzeyinin "Evet" oyu verme üzerinde anlamlı bir etkisi yoktur ($p > 0.05$).

`age` (Yaş):

- **Katsayı = -0.0039, p = 0.461:**
  - Yaşın "Evet" oyu verme üzerinde anlamlı bir etkisi yoktur ($p > 0.05$).

`rsex` (Cinsiyet):

- **Katsayı = -0.637, p < 0.001:**
  - Erkeklerin "Evet" oyu verme olasılığı kadınlara göre daha düşüktür.
  - Odds oranı: $e^{-0.637} \approx 0.53$, yani erkeklerin "Evet" oyu verme olasılığı kadınlara göre %47 daha azdır.

`libauth` (Liberal-Otoriter Eğilim):

- **Katsayı = -0.146, p = 0.029:**
  - Daha otoriter bir eğilime sahip bireylerin "Evet" oyu verme olasılığı azalır.
  - Odds oranı: $e^{-0.146} \approx 0.86$, yani otoriter eğilimdeki bireylerin "Evet" oyu verme olasılığı liberal eğilimde olanlara kıyasla %14 daha düşüktür.

`employmentdum` (İstihdam Durumu):

- **Katsayı = -0.277, p = 0.470:**
  - İstihdam durumunun "Evet" oyu verme üzerinde anlamlı bir etkisi yoktur ($p > 0.05$).

```
# Tahmin edilen olasılıkları hesaplama
predicted_prob <- predict(logit_model, type = "response")

# İlk birkaç tahmin
head(predicted_prob)
```

```
        1         2         3         4         7         8
0.2317852 0.1214676 0.1868734 0.1814289 0.0992647 0.1821207
```

- Modelimiz, her bir birey için "Evet" oyu verme olasılıklarını tahmin ediyor.
- İlk birkaç tahminde, bireylerin genellikle "Hayır" oyu verme olasılığının daha yüksek olduğunu görebiliyoruz.