T.C.

ÜSKÜDAR ÜNİVERSİTESİ

MÜHENDİSLİK VE DOĞA BİLİMLERİ FAKÜLTESİ

ADLİ BİLİMLER BÖLÜMÜ

ABL440

Adli Bilimlerde Büyük Veri

FİNAL ÖDEVİ

Hazırlayan

Aysu İĞDİ

200207016

Doç.Dr. Kaan YILANCIOĞLU

İÇİNDEKİLER

Takibi Yapılan Kitabın Künyesi

Kitap Adı : R and Data Mining -- Examples and Case Studies

Yazar : Yanchang Zhao

Yayınevi : Academic Press, Elsevier

Yayım Tarihi : Aralık 2012

ISBN : 978-0-123-96963-7

Uygulama

GitHub : [igdiaysu/big-data-in-fs]

Konsoldan çalışma alanı ayarlandı.

```
> setwd("D:/GitRepository/big-data-in-fs/final")
```

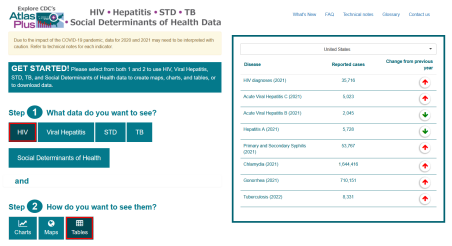
—

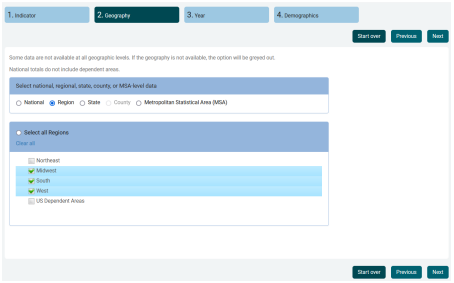
İlk bulunan dataset: [AIDS Deaths by State]

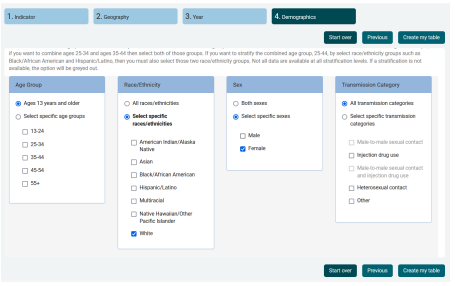
Source: [NCHHSTP AtlasPlus Portal]

Ayarlanan datasetin özellikleri:

- Gösterge (indicator) : HIV ölümleri
- Coğrafya : Bölge (region)
    - Midwest
    - West
    - South
- Yıl : 2008-2019
    - Veriler 2008-2021 arası olsa da sitede pandemi sebepli verilerle ilgili uyarı bulunmasından ötürü 2019 olarak daralttım.
- Yaş : 13+
- Irk/Etnik Köken : Beyaz
- Cinsiyet : Kadın
- Bulaşma Kategorisi : Tüm kategoriler
    - Enjeksiyonla uyuşturucu alımı
    - Heteroseksüel ilişki
    - Diğer







```
> library(readr)
> HIV_deaths2 <-
read_csv("D:/GitRepository/big-data-in-fs/final/HIV_deaths2.csv")
```

Veri Keşfi ve Görselleştirme (Data Exploration and Visualization)

Veri setinin boyutunu öğrenme: `dim()`

```
> dim(HIV_deaths2)
[1] 36  6
```

Veri setindeki değişken adlarını öğrenme: `names()`

```
> names(HIV_deaths2)
[1] "Indicator"       "Year"            "Geography"       "FIPS"            "Cases"           "Rate per 100000"
```

Veri setinin yapısını ve niteliklerini öğrenme: `str()` ve `attributes()`

```
> str(HIV_deaths2)
spc_tbl_ [36 × 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Indicator      : chr [1:36] "HIV deaths" "HIV deaths" "HIV deaths" "HIV deaths" ...
 $ Year           : num [1:36] 2008 2009 2010 2011 2012 ...
 $ Geography      : chr [1:36] "Midwest" "Midwest" "Midwest" "Midwest" ...
 $ FIPS           : num [1:36] 83002 83002 83002 83002 83002 ...
 $ Cases          : num [1:36] 114 98 114 106 123 93 83 83 115 124 ...
 $ Rate per 100000: num [1:36] 0.5 0.4 0.5 0.5 0.5 0.4 0.4 0.4 0.5 0.5 ...
 - attr(*, "spec")=
  .. cols(
  ..   Indicator = col_character(),
  ..   Year = col_double(),
  ..   Geography = col_character(),
  ..   FIPS = col_double(),
  ..   Cases = col_double(),
  ..   `Rate per 100000` = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

```
> attributes(HIV_deaths2)
$row.names
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36

$names
[1] "Indicator"       "Year"            "Geography"       "FIPS"            "Cases"           "Rate per 100000"

$spec
cols(
  Indicator = col_character(),
  Year = col_double(),
  Geography = col_character(),
  FIPS = col_double(),
  Cases = col_double(),
  `Rate per 100000` = col_double()
)

$problems
<pointer: 0x000002bc23a29cf0>

$class
[1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"

> |
```

Verinin minimum, maksimum, ortalama, medyan, birinci ve üçüncü çeyrek değerlerini içeren özeti inceleme: *summary()*

```
> summary(HIV_deaths2)
  Indicator              Year       Geography             FIPS          Cases        Rate per 100000
 Length:36          Min.   :2008   Length:36          Min.   :83002   Min.   : 83.0   Min.   :0.4
 Class :character   1st Qu.:2011   Class :character   1st Qu.:83002   1st Qu.:113.8   1st Qu.:0.5
 Mode  :character   Median :2014   Mode  :character   Median :83003   Median :126.5   Median :0.7
                    Mean   :2014                      Mean   :83003   Mean   :199.6   Mean   :0.8
                    3rd Qu.:2016                      3rd Qu.:83004   3rd Qu.:345.2   3rd Qu.:1.1
                    Max.   :2019                      Max.   :83004   Max.   :414.0   Max.   :1.3
```

Verinin çeyreklerini inceleme: *quantile()*

```
> quantile(HIV_deaths2$Cases)
    0%    25%    50%    75%   100%
 83.00 113.75 126.50 345.25 414.00
> quantile(HIV_deaths2$`Rate per 100000`)
  0%  25%  50%  75% 100%
 0.4  0.5  0.7  1.1  1.3
```

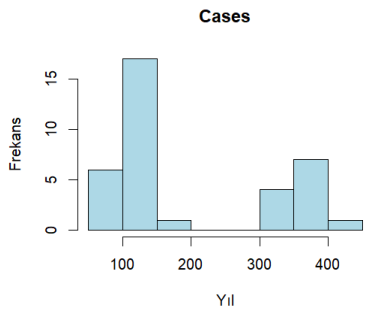
Verinin varyansını inceme: *var()*

100 bin kişi başına ölüm oranındaki sapmanın sonucu:

```
> var(HIV_deaths2$`Rate per 100000`)
[1] 0.09885714
```

Verinin histogram grafiği: *hist()*

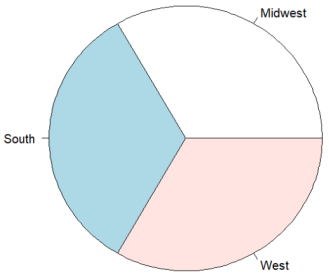
200-300 arasında vaka bulunan bölge olmadığını görüyoruz.




```
> hist(
+   HIV_deaths2$Cases,
+   main = "Cases",
+   xlab = "Yıl",
+   ylab = "Frekans",
+   col = "lightblue",
+ )
```

Verinin pie chartı: `pie(table())`

```
> pie(table(HIV_deaths2$Geography))
```

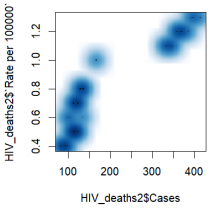


Verinin kovaryans ve korelasyon: `cov()` ve `cor()`

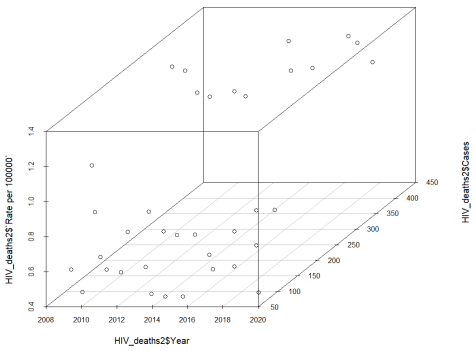
```
> cov(HIV_deaths2$Cases, HIV_deaths2$`Rate per 100000`)
[1] 36.04
> cor(HIV_deaths2$Cases, HIV_deaths2$`Rate per 100000`)
[1] 0.9409925
```

Verinin düzgün dağılım grafiği (renk yoğunluğu): `smoothScatter()`

```
> smoothScatter(HIV_deaths2$Cases, HIV_deaths2$`Rate per 100000`)
```



Verinin üç boyutlu dağılım grafiği: `scatterplot3d()`



Verinin ısı haritası: `heatmap()`

Isı haritası kullanımını gerektiren sayıda veri satırı bulunmamaktadır.

Decision Tree: train data

Başlangıçta `Geography`nin değişkenin türü değiştirildi: chracter → factor

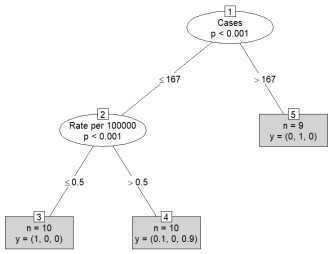
Tekrarlanabilirlik için seed girdisi yapıldı. Aynı sürümde farklı bilgisayarda bile aynı sonuçların tekrar edilebilmesi için (Aynı rastgelelik).

```
# HIV_deaths2$Geography<-as.factor(HIV_deaths2$Geography)

set.seed(444)
ind <- sample(2, nrow(HIV_deaths2), replace=TRUE, prob=c(0.7, 0.3))
trainData <- HIV_deaths2[ind==1,]
testData <- HIV_deaths2[ind==2,]

library(party)
myFormula <- Geography ~ Year + Cases + `Rate per 100000`
hiv_ctree <- ctree(myFormula, data=trainData)


# check the prediction >
table(predict(hiv_ctree), trainData$Geography)
```

```
table(predict(hiv_ctree), trainData$Geography)

         Midwest South West
Midwest       10     0    0
South          0     9    0
West           1     0    9
```

```
> print(hiv_ctree)

        Conditional inference tree with 3 terminal nodes

Response:  Geography
Inputs:  Year, Cases, Rate per 100000
Number of observations:  29

1) Cases <= 167; criterion = 1, statistic = 27.393
  2) Rate per 100000 <= 0.5; criterion = 1, statistic = 14.394
    3)*  weights = 10
  2) Rate per 100000 > 0.5
    4)*  weights = 10
1) Cases > 167
  5)*  weights = 9
```

```
plot(hiv_ctree, type="simple")
```



Decision Tree: test data

```
> # predict on test data >
> testPred <- predict(hiv_ctree, newdata = testData)
> table(testPred, testData$Geography)

testPred  Midwest South West
  Midwest       1     0    0
  South         0     3    0
  West          0     0    3
```

Tamamını doğru tahmin etti.

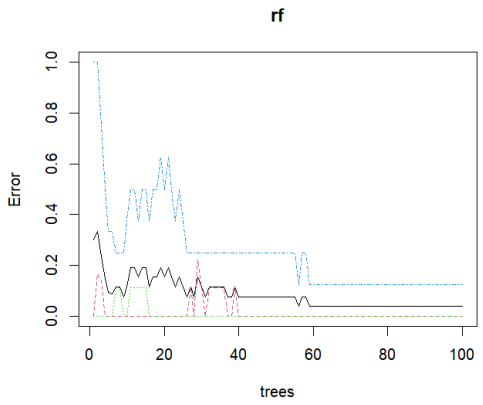
Random Forest

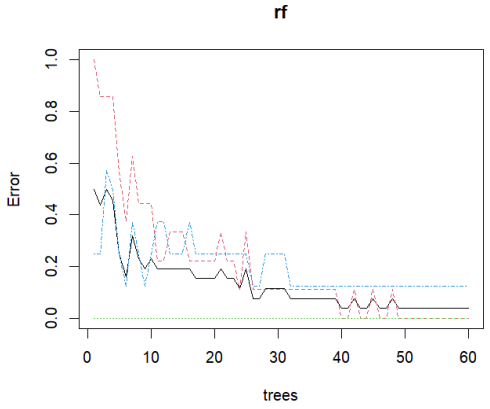
trainData işlemleri:

```
> HIV_deaths3$Geography<-as.factor(HIV_deaths3$Geography)
> ind <- sample(2, nrow(HIV_deaths3), replace=TRUE, prob=c(0.7, 0.3))
> trainData <- HIV_deaths3[ind==1,]
> testData <- HIV_deaths3[ind==2,]
> library(randomForest)
> rf <- randomForest(Geography ~ Year + Cases + `Rate`, data=trainData, ntree=100, proximity=TRUE)
> table(predict(rf), trainData$Geography)

          Midwest South West
  Midwest       9     0    1
  South         0     9    0
  West          0     0    7
```

```
> print(rf)

Call:
 randomForest(formula = Geography ~ Year + Cases + Rate, data = trainData,      ntree = 100, proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 1

        OOB estimate of  error rate: 3.85%
Confusion matrix:
        Midwest South West class.error
Midwest       9     0    0       0.000
South         0     9    0       0.000
West          1     0    7       0.125
```

```
> attributes(rf)
$names
 [1] "call"            "type"            "predicted"       "err.rate"        "confusion"       "votes"
 [7] "oob.times"       "classes"         "importance"      "importanceSD"    "localImportance" "proximity"
[13] "ntree"           "mtry"            "forest"          "y"               "test"            "inbag"
[19] "terms"

$class
[1] "randomForest.formula" "randomForest"
```





100 ağaç üretilerek çalıştırılmış modelde 60'tan sonra değişiklik olmadığı gözlemlenerek 60 ağaçla tekralanıp grafik gözlemlendi.

```
> importance(rf)
      MeanDecreaseGini
Year          1.485198
Cases         6.243888
Rate          8.856384
```

Tahmin etmede değişkenlerin önemine dair tablo.

Random Forest: test data

```
> hivPred <- predict(rf, newdata=testData)
> table(hivPred, testData$Geography)

hivPred   Midwest South West
  Midwest       2     0    0
  South         0     3    0
  West          1     0    4
```

3 Mindwest'ten 1 tanesini West olarak tahmin etti.

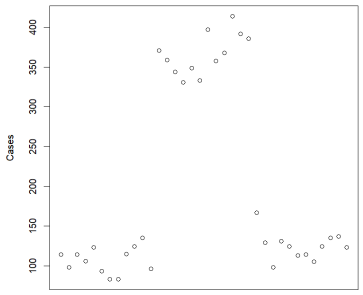
Regresyon

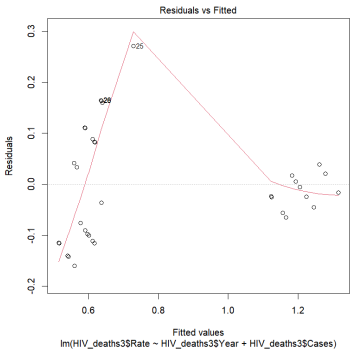
Regresyon, bağımlı bir değişkeni (yanıt olarak da adlandırılır) tahmin etmek için bağımsız değişkenlerin (tahmin ediciler olarak da bilinir) bir fonksiyonunu oluşturmaktır.

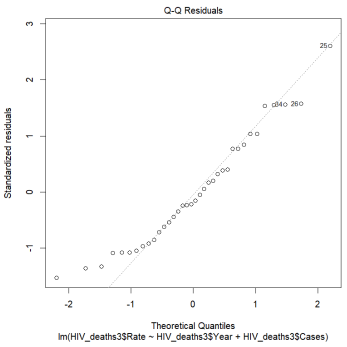
$$y = c_0 + c_1 x_1 + c_2 x_2 + \cdots + c_k x_k,$$

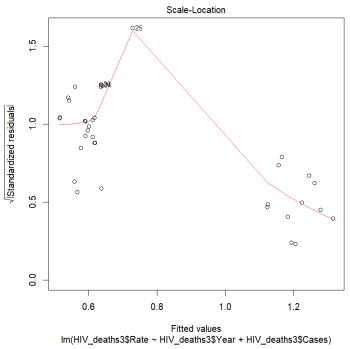
```
> reg <- lm(HIV_deaths3$Rate ~ HIV_deaths3$Year + HIV_deaths3$Cases)
> reg

Call:
lm(formula = HIV_deaths3$Rate ~ HIV_deaths3$Year + HIV_deaths3$Cases)

Coefficients:
      (Intercept)   HIV_deaths3$Year  HIV_deaths3$Cases
          3.31471           -0.00149            0.00243
```

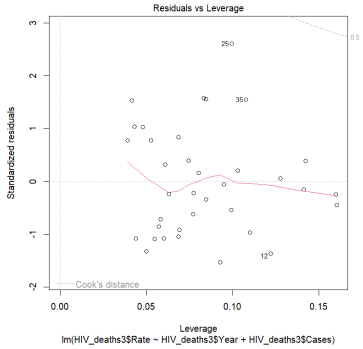
```
> summary(reg)

Call:
lm(formula = HIV_deaths3$Rate ~ HIV_deaths3$Year + HIV_deaths3$Cases)

Residuals:
     Min       1Q   Median       3Q      Max
-0.15975 -0.09226 -0.01945  0.08261  0.27106

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.314705  10.646352   0.311    0.757
HIV_deaths3$Year  -0.001490   0.005288  -0.282    0.780
HIV_deaths3$Cases  0.002430   0.000152  15.991   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1095 on 33 degrees of freedom
Multiple R-squared:  0.8857,    Adjusted R-squared:  0.8788
F-statistic: 127.9 on 2 and 33 DF,  p-value: 2.852e-16
```



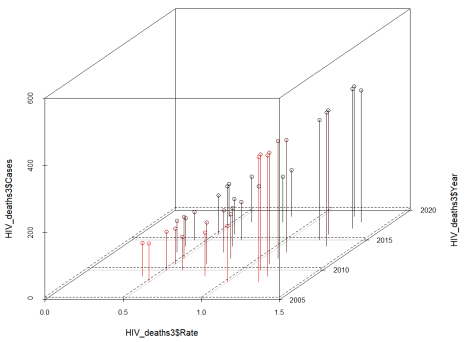









```
> library(scatterplot3d)
> s3d <- scatterplot3d(HIV_deaths3$Rate, HIV_deaths3
$Year, HIV_deaths3$Cases, highlight.3d=T, type="h", l
ab=c(2,3))
> s3d$plane3d(reg)
```



Kümeleme (Clustring)

```
> hiv2 <- HIV_deaths3
> hiv2$Geography <- NULL
> hiv2$Indicator <- NULL
> hiv2$FIPS <- NULL
> (kmeans.result <- kmeans(hiv2, 3))
K-means clustering with 3 clusters of sizes 8, 16, 12

Cluster means:
      Year    Cases     Rate
1 2013.375  95.2500 0.462500
2 2013.562 126.3750 0.675000
3 2013.500 366.8333 1.191667

Clustering vector:
 [1] 2 1 2 1 2 1 1 1 2 2 2 1 3 3 3 3 3 3 3 3 3 3 3 2 2 1 2 2 2 2 1 2 2 2 2

Within cluster sum of squares by cluster:
[1]  609.4337 2956.0175 7704.7358
 (between_SS / total_SS =  97.8 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
```

12-12-12 bulamadı.

```
> table(HIV_deaths3$Geography, kmeans.result$cluster)

           1  2  3
  Midwest  6  6  0
  South    0  0 12
  West     2 10  0
```

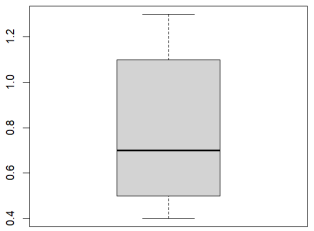
West ve Midwest'in iki ayrı kümeye dağılması hepsini doğru kümeyemediğini gösteriyor.

3 = South. Ama 1 ve 2'yi bilemeyiz.

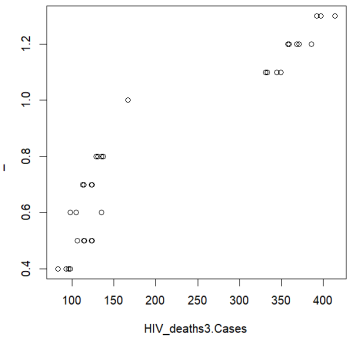
Verinin azlığı sebebiyle kmeans'tan verim alınamadığından diğer metotlar eklenmemiştir.

Aykırı Değer Tespiti (Outlier Detection)

```
> boxplot.stats(HIV_deaths3$Rate)$out
numeric(0)
> boxplot(HIV_deaths3$Rate)
```




```
> attach(df)
The following object is masked from df (pos = 3):

    HIV_deaths3.Rate

The following object is masked from df (pos = 4):

    HIV_deaths3.Rate

> (a <- which(HIV_deaths3$Cases %in% boxplot.stats(HIV_deaths3$Cases)$out))
integer(0)
> detach(df)
> (b <- which(HIV_deaths3$Rate %in% boxplot.stats(HIV_deaths3$Rate)$out))
integer(0)
> (outlier.list1 <- intersect(a,b))
integer(0)
```



Verimde outlier çıkmadığı için bölümün kalanı veri setine uygulanmadı.

Time Series Data in R

```
> # Time Series
> a <- ts(1:12, frequency=1, start=c(2008,1))
> a
Time Series:
Start = 2008
End = 2019
Frequency = 1
 [1]  1  2  3  4  5  6  7  8  9 10 11 12
>
```

Time Series Forecasting

```
install.packages("forecast")
library(forecast)

ts_trainData <- ts(HIV_deaths3$Cases, start = min(HIV_deaths3$Year), end = 2017, frequency = 1)
plot(ts_trainData, main = "Cases Over Time", xlab = "Year", ylab = "Number of Cases")

model <- arima(ts_trainData)
forecast_result <- forecast(model, h = 2)  # 2
plot(forecast_result, main = "Forecasted Cases Over Next 2 Year", xlab = "Year", ylab = "Number of Cases")
```

```
> forecast_result
     Point Forecast    Lo 80    Hi 80   Lo 95    Hi 95
2018          105.3 86.69277 123.9072 76.8427 133.7573
2019          105.3 86.69277 123.9072 76.8427 133.7573
```