

VERİ BİLİMİ için R PROGRAMLAMA

Temel Kavramlar

R Programlama Temelleri – R ile Veri Ön İşleme

R ile Veri Görselleştirme Temelleri – R ile Tahmin Analitiğine Giriş

1) Başlamadan Önce

Neler Öğreneceksiniz?

R Programramlama Temelleri

Kurulum

Proje Oluşturma – Çalışma Dizini Kullanma

Konsol ve Script Kullanımı

Değişken Tipleri

Mantık Değişkenleri (if-else)

Döngüler (For – While Loops)

Vektörler ve Listeler

Fonksiyonlar (Kendi fonksiyonumuz da)

Kütüphaneler – Paketler

R ile Veri Hazırlamaya Giriş

Matrisler ve [] Kullanımı (Kendi matrisimiz)

Subsetting (Alt Küme alma)

Data Import (R'a Veriseti alma)

DataFrameler ve \$ Kullanımı (Kendi DFimiz)

Veri Keşfetmeye Giriş

DF işlemleri + Filtreleme + Birleştirme

R'da Faktörler (Kategorik Değişkenler)

Eksik değerler (Missing Values)

Apply fonksiyonları (mapply, apply, sapply, tapply)

R ile Veri Görselleştirmeye Giriş

plot() fonksiyonu ve arguments kullanımı

barplot() ile sütun grafikleri

hist() ile histogram grafiği

boxplot() ile kutu grafikleri

par() ile grafik ızgarası (çoklu grafik gösterme)

Kütüphaneler ile görselleştirmeye giriş(lattice – ggplot2)

R ile Tahmin Analitiğine Giriş

Veri Ön İşleme – Hazırlama

Regresyon modellemesi (Numerik değer tahmini)

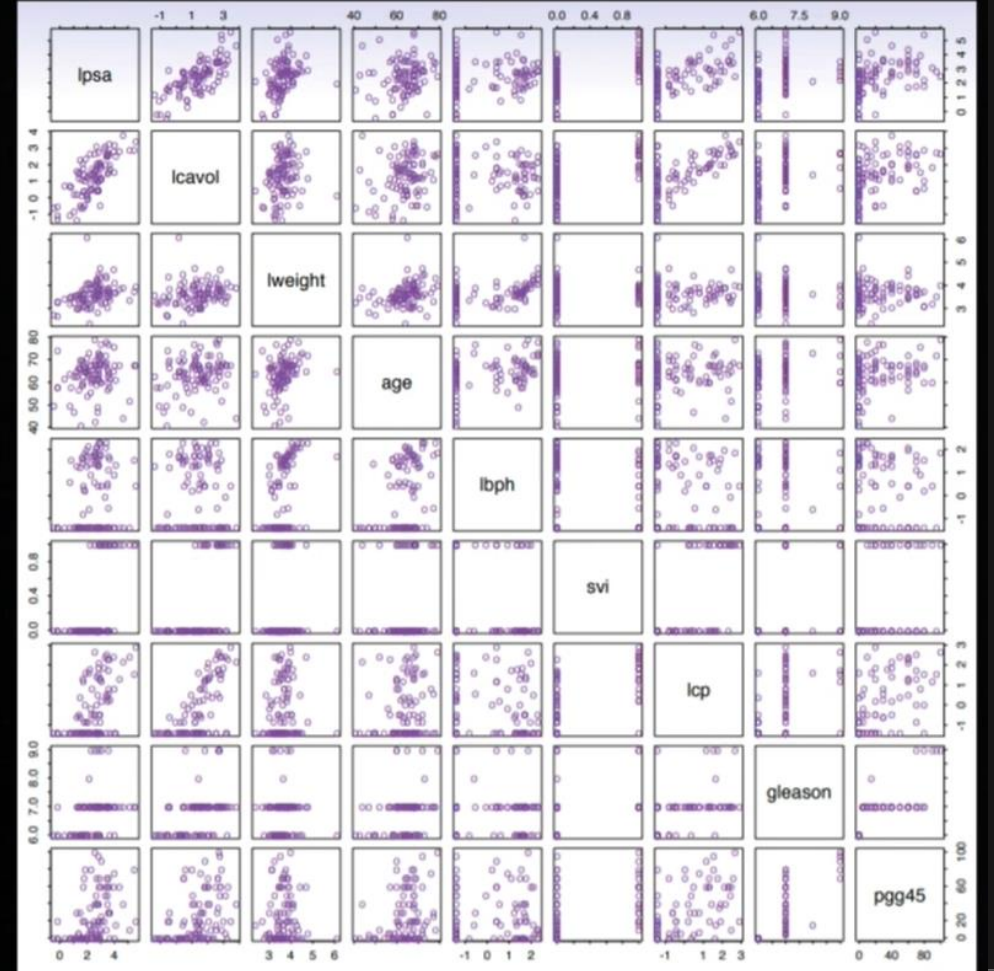
Sınıflandırma modellemesi (Classification)

Keywordler

1. Proje-Çalışma dizini
2. Konsol-Script
3. Paket-Kütüphaneler
4. Fonksiyon
5. Satır-Sütun-boyut
6. Değişken
7. Vektör
8. Index(İndis)
9. Matris
10. (Veri) Tablo(su)
11. [], \$
12. Faktörler(Levels)
13. Missing Value
14. Filtreleme
15. ?
16. Veri Keşfetme
17. Grafikler-Görselleştirme
18. Veri hazırlama
19. Modelleme
20. Sınıflandırma
21. Tahminleme
22. Supervised learning – Unsupervised learning
23. Machine learning (makine öğrenimi)
24. Train-Validation-Test Split

İstatistikî Öğrenim Örnekleri 1

Kanser Teshisi



İstatistiki Öğrenim Örnekleri 2

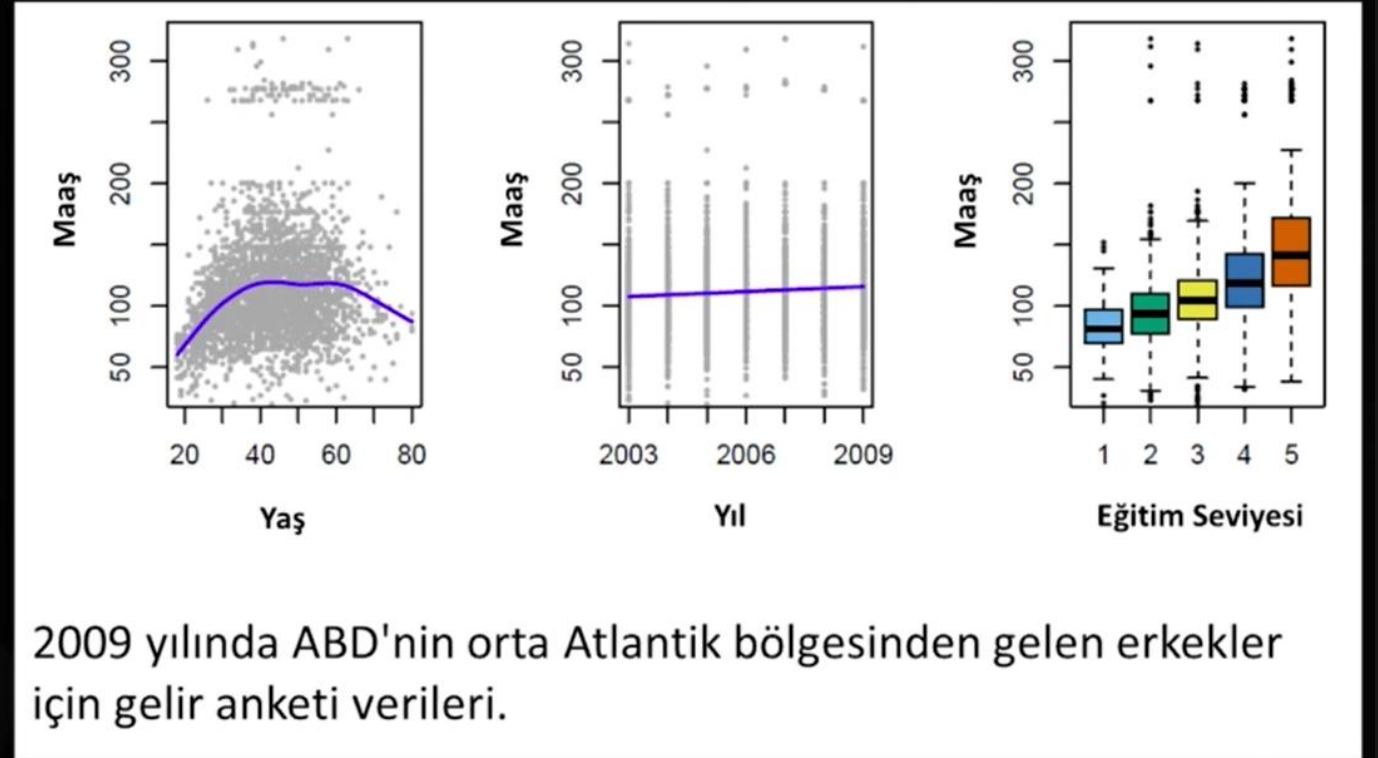
👤 Kanser Teshisi

👤 El yazısı ile yazılan
sayıların teshisi

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9

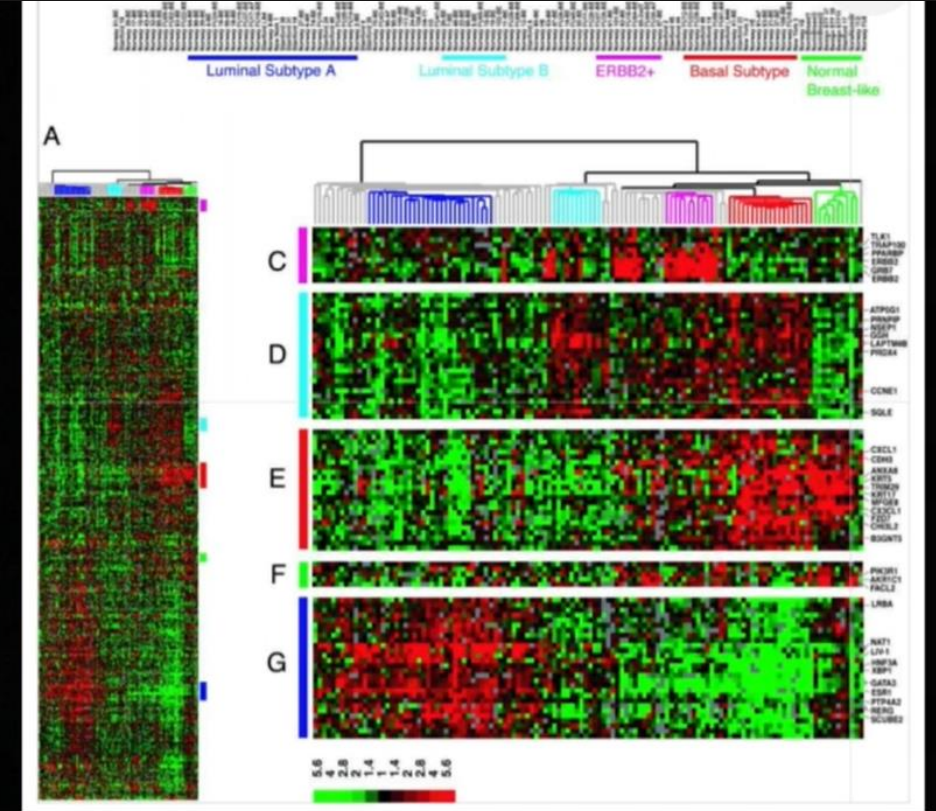
İstatistikî Öğrenim Örnekleri 3

- 👉 Kanser Teşhisi
- 👉 El yazısı ile yazılan sayıların teşhisi
- 👉 Maaş ~ Demografik Veriler



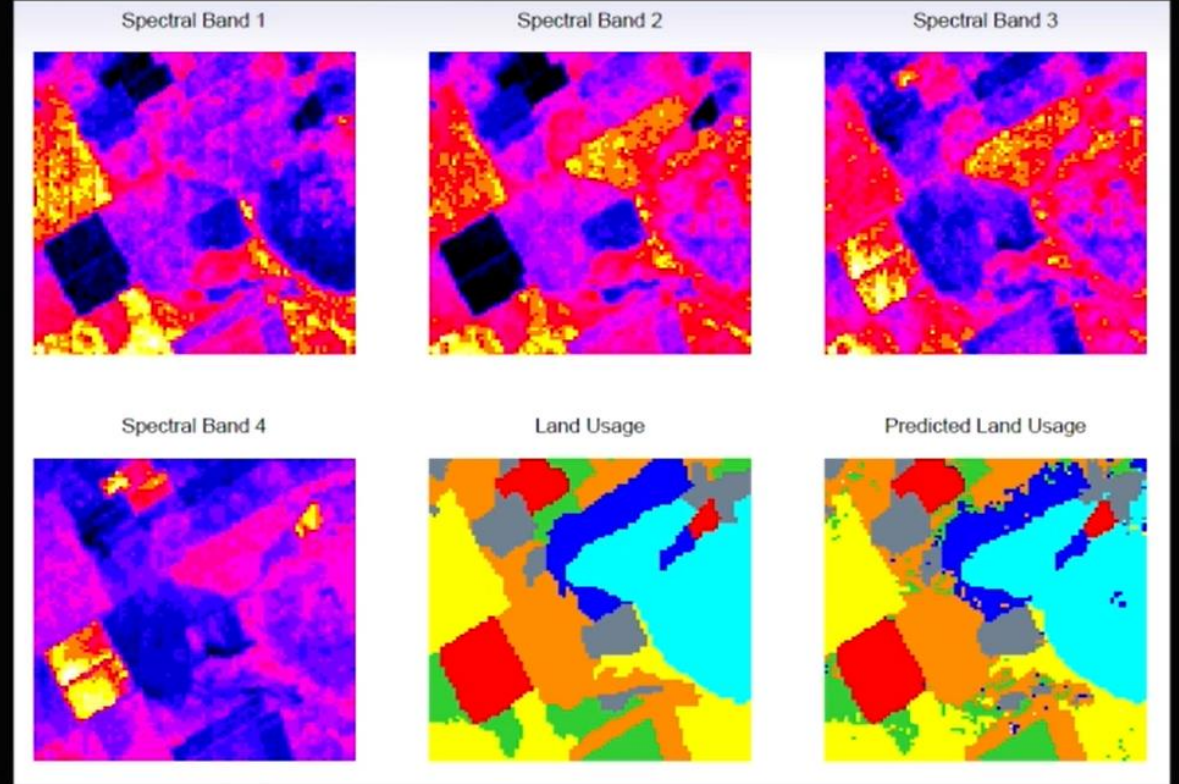
İstatistikî Öğrenim Örnekleri 4

- El yazısı ile yazılan sayıların tespiti
- Maaş ~ Demografik Veriler
- Genetik Bilgilerle → Kanserli Dokü Tespiti



İstatistikî Öğrenim Örnekleri 5

- ☺ Kansar Teshisi
- ☺ El yazısı ile yazılan sayıların teshisi
- ☺ Maaş ~ Demografik Veriler
- ☺ Genetik Bilgilerle → Kanserli Doku Teshisi
- ☺ Uydu görüntülerinden arazi kullanımı tespiti



Supervised Learning

- Sonuç Ölçütü: $Y \rightarrow$ dependent variable (DV)
response - target
- Tahmin edici ölçütler: $X \rightarrow$ independent variable (IV)
inputs - regressors - features - covariates
- Training Data $\rightarrow (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
(Eğitim Seti)

1) Regresyon (Numerik) Modelleme 2) Classification (Sınıflandırma)

Unsupervised Learning

Kümeleme (Clustering)

- 👤 Hedef değişkenimiz (Y) yok. Input (bağımsız) değişkenler var.
- 👤 Amaçlarımız daha belirsiz. → Benzer davranan gözlemlerin gruplanması. değişken kombinasyonlarını belirlemek
- 👤 Performans ölçütleri daha belirsiz.



Unsupervised Learning

Kümelene (Clustering)

- 1. Hedef değişkenimiz (Y) yok. Input (bağımsız) değişkenler var.
- 2. Amaçlarımız daha belirsiz. → Benzer davranan gözlemlerin gruplanması. değişken kombinasyonlarını belirlemek
- 3. Performans ölçütleri daha belirsiz.



Kümeleme Örnekleri ve Metrikler

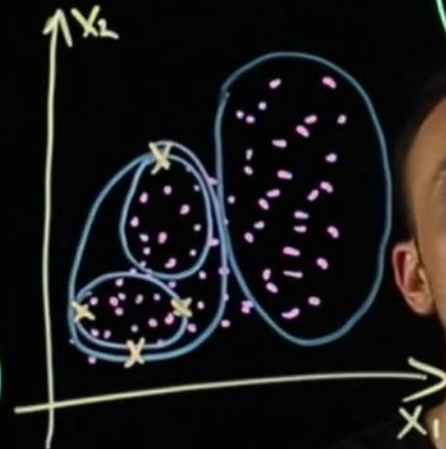
- Kümeleme

(1) Partitioning



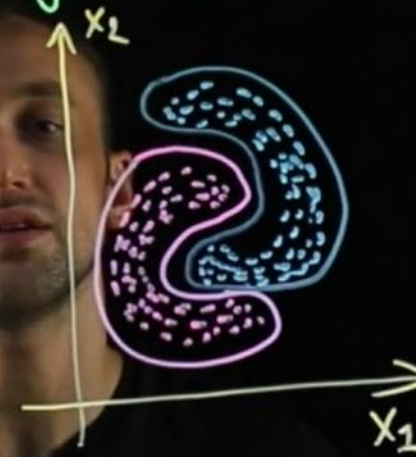
- * K-Means
- * K-Medoids
- * CLARANS

(2) Hierarchical



- * BIRCH
- * Chameleon

(3) Density based
(Yoğunluk Tabanlı)



- * DBSCAN
- * HDBSCAN

Performans Metrikleri:

- * Merkezlerin konumu
- * Kümeler arası uzaklık
- * Küme içi mesafe

Machine Learning vs Statistical Learning

(Makine Öğrenimi) vs. (İstatistiksel Öğrenme)

Machine Learning df Statistical Learning

- ① Yapay zekanın alt dalı olarak yükseldi / popülerleşti
- ② İstatistiğin alt dalı olarak popülerleşti / yükseldi.
- ③ Ortak alanları çok fazla → Supervised & Unsupervised
- ④ Modelleme, yorumlanabilirlik odaklı.
- ⑤ Büyük ölçeklilik, doğru tahmin odaklılık
- ⑥ Sınırlar belirsizleşmeye başladı → Çapraz çalışmalar (Cross-fertilization)
- ⑦ Machine Learning → Marketing ↑


Tavsiyeler

- 🧠 Keyword'lere dikkat! 🙌
- 🧠 Kodların ne işe yaradığını sorgulayın.
- 🧠 Gerçek hayatla bağ kurun ve üstüne ekleyin.
- 🧠 Kodları beraber çalıştırın
(mantığını anlamadıysanız tekrar izleyin).
- 🧠 Becerilerinizi farklı verisetlerinde deneyin.

- 🧠 Aynı kodlarla farklı pratikler üretin.
→ ?'den faydalanın.
- 🧠 Becerilerinizle iş hayatında veya günlük hayatınızda neleri değiştirebileceğinizi hayal edin.
- 🧠 Ödev ve çözümleri takip edin!
- 🧠 Anlaşılmayan kısımlar için iletişime geçin.

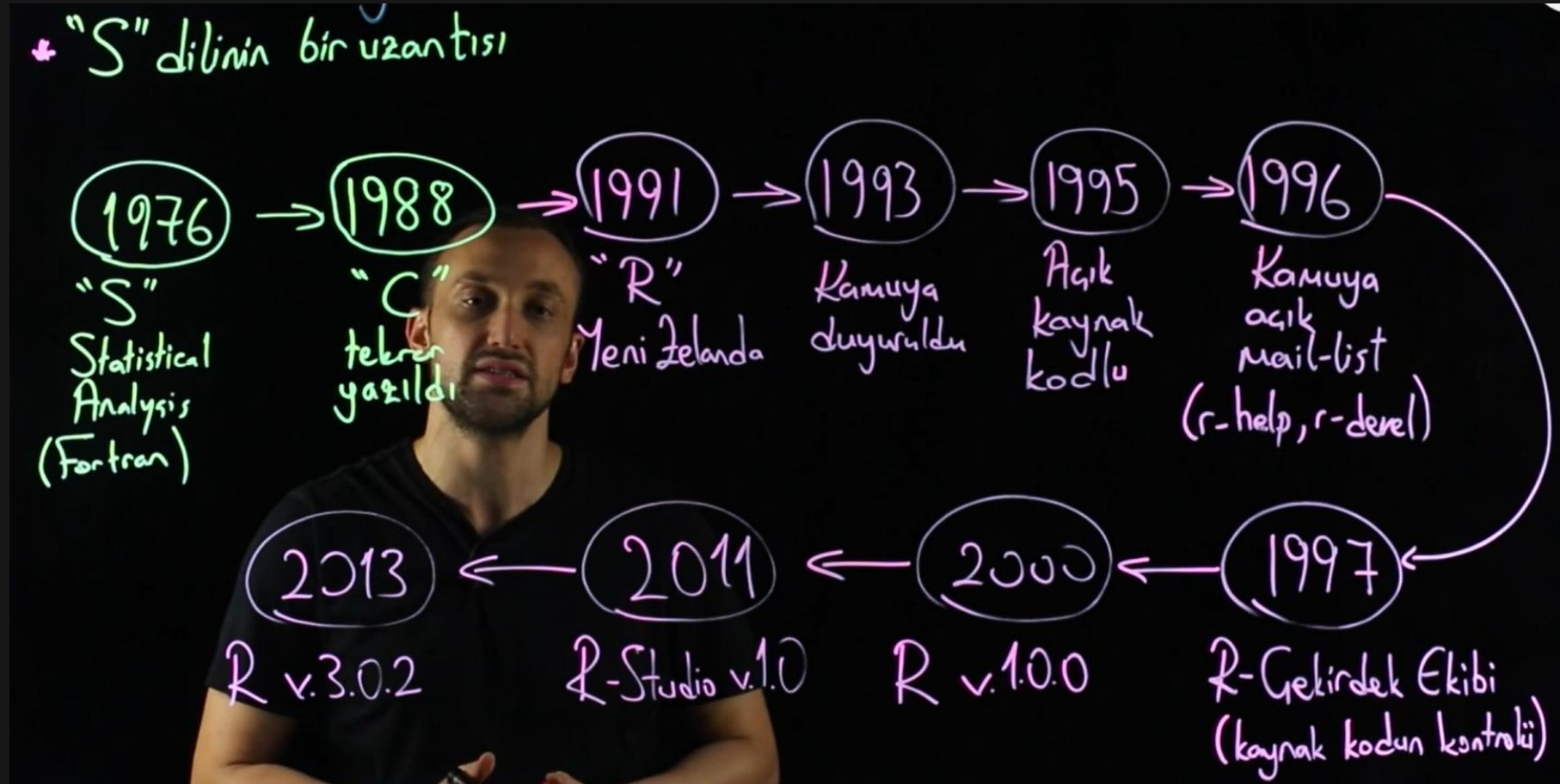
2) R Programlama Temelleri

R Programlama Diline Giriş



- * "Ücretsiz" → Açık kaynak kodlu (cran.r-project.org)
 - Geliştirmek
 - Adaptasyon
 - Paylaşmak
 - Geliştirmek
- * "S" programlama dili uzantısı → "C" tekrar
- * Herhangi bir hesaplama ortamında çalışabiliyor (PS 3 :))
- * Sık versiyon yükseltmeleri - hata (bug) düzeltmeleri
- * Yalındır (İstatistik-Matematik) → 4000 paket kütüphaneler
- * GÖRSELLEŞTİRME #1
- * Aktif kullanıcılar → Yardımlaşma

R Programlama Dilinin Tarihi



R ve R-Studio Kurulumu

- R -> Çekirdek (Core)
- R-Studio -> Kullanıcı Arayüzü (GUI)

Çalışma Dizini ve Proje

- `getwd()`: Anlık çalışma dizini öğrenme
- `setwd()`: Çalışma dizini belirleme
- Proje
 - Çalışma dizini
 - Dizindeki veriler
 - Kodlar
 - Kodlama sonucu oluşan bütün objeler

R Help?!

- ? Fonksiyonu ile Help penceresinin kullanabilirsiniz
- Stackoverflow
- R-Bloggers

R-Studio Konsol ve Script

- Konsol -> Kodları direk çalıştırabilirsiniz (ENTER)
- Script -> Kodlarınız revize ederek çalıştırabilirsiniz (CTRL + ENTER)

R'da Değişken Tipleri

- Integer
- Double
- Logic
- Character
- Logic

R'da Mantık İfadeleri

- TRUE (T) = DOGRU
- FALSE (F) = YANLIS
- > , <
- >= , <=
- == , != , !

Birleştirmek için

- | = OR (ya da)
- & = AND (ve)
- isTRUE(x) = DOGRUDUR

R'da Koşul İfadeleri

- IF
- ELSE IF
- ELSE
- Chained Conditionals (Zincir Koşulları)

R'da Döngüler

- FOR - WHILE
- NESTED LOOPS (Ic ice Donguler)

- FOR

```
for(iterasyon sayisi){  
    aksiyon  
}
```

- WHILE

```
while(logical expression){  
    action statement  
    counter  
}
```


R'da Döngüler

Vektörler (Vectors)

- Tek tip veri içerebilir

Listeler (Lists)

- Çeşitli tipte veri içerebilir

R'da Döngüler

Vektörler (Vectors)

- Tek tip veri içerebilir

Listeler (Lists)

- Çeşitli tipte veri içerebilir

R'da Basit Fonksiyonlar

- print()
- paste()
- is.numeric()
- is.character()
- typeof()
- c()
- seq()
- rep()
- sqrt()
- runif() # rnorm()

R'da Kendi Fonksiyonumuzu Yazmak

```
fonksiyon_adi <- function(x){  
  x parametresi ile ilgili islemlerimiz  
}
```

R'da Paketler ve Kütüphaneler

PACKAGE: Data(Veriler) - Function(Fonksiyonlar)

- `install.packages()`

LIBRARIES: Paketlerin bulunduğu dizinler(konumlar)

- `library()`

R'da Basit Fonksiyonlar

PACKAGE: Data(Veriler) - Function(Fonksiyonlar)

- `install.packages()`

LIBRARIES: Paketlerin bulunduğu dizinler(konumlar)

- `library()`

R Programlama Temelleri Özet

2.16. Bu Bölümde Neler Öğrendik?

• R Programlamaya Giriş & Tarihçesi

• R Kurulumu → R (Core)
↳ R-Studio (Arayüz)

• R-Help → Google
↳ r-bloggers, stackoverflow
vs..
↳ ? (?dcast())

• Çalışma Dizini → "PROJE"

• R Console & Script
↳ komutlar çalışır
↳ Defter

• Değişken Tipleri
↳ numeric
↳ character
↳ logical

• Logical (Mantık) & Condition (Koşul)
↳ >, <, ==, !=, !
↳ if, ifelse, else
& (and), | (or)

• Döngüler (Loops)
↳ for * sonsuz kadar
↳ while ↳ sınırsız ihtimalleri olabilir!

• Vectors & List →

6	7	8	9	10
---	---	---	---	----

• Fonksiyonlar → R (base) Fonksiyonlar
↳ Kullanıcıların Fonksiyonları

• R Packages / Libraries



R ile Veri Ön İşleme

R'da [] Kullanımı

- R Programlama Dilinde indisler(pozisyonlar) 1'den başlar
- Örnekler
 - `x[2]`
 - `x[-4]`
 - `x[1:4]`
 - `y[c(2,4)]`
 - `y[c(-1,-3)]`
 - `v1[-3:-5]`

R'da [] Kullanımı

- R Programlama Dilinde indisler(pozisyonlar) 1'den başlar
- Örnekler
 - `x[2]`
 - `x[-4]`
 - `x[1:4]`
 - `y[c(2,4)]`
 - `y[c(-1,-3)]`
 - `v1[-3:-5]`

Matrisler

X (Vektör Adı)				
Y (Matris Adı)	[,1]	[,2]	[,3]	[,4]
[1,]				
[2,]				
[3,]				

- $X[3] +$
- $Y[3] ?$
- Indeks: $Y[2,] - Y[,3] - Y[2,3]$
- Aynı tipten değerler ve değişkenler içerebilir

R'da Kendi Matrisimizi Oluşturma

- `B <- matrix(veri,3,4)`
- `C <- matrix(veri,3,4, byrow = T)`
- `rbind()` / Satir(row) Baglama(binding)
- `cbind()` / Sütun(Column) Baglama(binding)

R'da Boyut İsimlendirme

- Vektör İsimlendirme

- `names(vektor1) <- c("x", "y", "z", "t")`
- `names(vektor1) <- NULL`

Matris Boyutu İsimlendirme (Naming Matrix Dimensions)

- `row.names()` / satir isimleri
- `col.names()` / sutun isimlerini

R'da Matris İşlemleri

- `matrix_B / matrix_A`
- `C = round(matrix_B / matrix_A,1)`
- `D = round(matrix_A+matrix_B,1)`

R'da Subsetting

- Subsetting Vectors
- Matrix Subset
 - `a_matrix[satir_numarasi, sutun numarasi]`
 - `a_matrix[1,]` # satir ismi kayboldu (Matrix degil, Vektore donustu)
 - `is.matrix(a_matrix[1,,drop=F])` (Matrix olarak kalir)
- SUBSETTING EXTRA
 - List subset
 - Subset Nested Items / Ic ice Elemanlardan Alt Kume Olusturma
 - Subset Partial Matching / Kism Eslesme ile Alt Kume Olusturma

R Data Import

- 1. Yol: Manual Selection / Dosyayi Elle Secmek
- 2. Yol: Set Working Directory / Calisma Dizini
- 3. Yol: File Directory / Dosyanin Kendi Dizini

Dataframes

Y (Matris Adı)	\$Yas [,1]	\$Cinsiyet	\$Egitim
[1,]	30	K	Yukse
[2,]	26	K	Unv
[3,]	742	E	Lise

- Özelleşmiş Listelerdir -> Her değişken listenin bir elemanı
- Aynı uzunluğa sahiptir (matris yapısının satır sayısı)
- Y\$Yas, Y\$Cinsiyet, Y\$...
- Farklı tipten değer ve değişkenler içerebilir
- row.names(): Satır adı atanıp kullanılabilir

R'da Veri Keşfine Giriş

- number of rows / satir sayisi. 891
- number of columns / sutun sayisi.
- head()
- tail()
- str()
- summary()

R'da \$ Kullanımı

```
data_table[,"Degisken"] = data_table$ Degisken
```

R'da Kendi Data Frame'imizi Oluşturma

- dataframe() fonksiyonu
- Cbind – Rbind Kullanımı

R'da Data Frame İşlemleri

- Tek boyutlu dataframe kontrolü
- Sutunlarda İşlemler
- Sutun Ekleme
- Hatırlatma: Recycling Vectors
- Sutun silme

R'da Data Frame Filtreleme

Örnekler

- `filtre <- data_table$Age > 20`
- `filtre`
- `data_table[filtre,]`
- `data_table[data_table$Cabin == "" & data_table$Survived == 1, c("Age", "Sex")]`

R'da Data Frame Birleştirme

```
merge(data_table, degiskenler, by.x = "x_ID", by.y = "y_ID")
```

R'da Faktörler

- `factor(c("erkek", "kadin", "kadin", "erkek"))`
- `as.factor(data_table$degisken)`

R'da Eksik Değerler

- NA: Hersey için
- NAN: Tanimsiz matematik islemleri için (NaN)
- Missing Value Cleaning (Data Cleaning) / Eksik Degerleri Temizleme
 - `is.na(a)`
 - `complete.cases(a)`

R'da APPLY Fonksiyon Ailesi

- Lapply: Listeler üzerinde apply fonksiyonu çalıştırır
- Sapply: Simplified lapply / Basitleştirilmiş lapply()
- Apply: Loop ile aynı sureyi alabilir fakat tek satırda okuma kolaylığı sağlar
- Mapply: Multivariate apply / Çok değişkenli apply fonksiyonu uygulamamızı sağlar
- Tapply: Fonksiyonu Vektörün Alt Kumelerine uygulamamızı sağlar

R'da Veri Ön İşleme Özeti

3.19. Bölüm Özeti

- Matrisler ve `[]` kullanımı
 - Kendi matrisimizi oluşturma
 - Satır ve sütun isimlendirme
 - Matris işlemleri
 - Subsetting (Alt küme)
- Data Import (R'a veri yükleme)
- Dataframe ve `$$` kullanımı
 - Kendi DF'mizi oluşturma
 - DF işlemleri
 - DF filtreleme
 - DF birleştirme
- R'da faktörler (kategorik metin değişkenleri)
 - `levels()`
- Missing Values (eksik değerler)
- Apply fonksiyonları (Döngü fonksiyonları)
 - `lapply ~ sapply`
 - `mapply`
 - `tapply`
 - `apply`

R'da Veri Görselleştirme Temelleri

R'da Veri Görselleştirme Temelleri

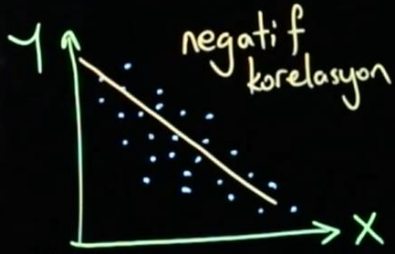
- `plot()` fonksiyonu ve *arguments* kullanımı
 - Tek değişken
 - İki değişken (Korelasyon)
 - Bütün değişkenler (Korelasyon matrisi)
- Sütun grafikleri
- Histogram
 - Örtüşen Histogramlar
- Boxplot (Kutu Grafikleri)
- Grid of Plots (Grafik Izgarası)
- Görselleştirme Kütüphaneleri
 - Lattice
 - ggplot2

R'da Veri Görselleştirme Temelleri Özeti

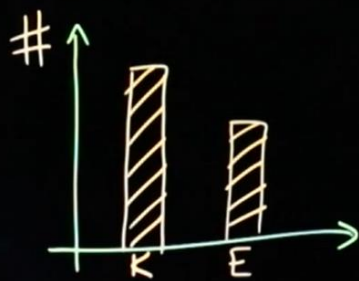
4.8. Bölüm Özeti

- `plot()` - arguments kullanımı

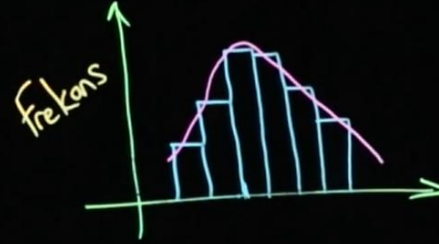
- Scatter plot (Serpme Grafikleri)



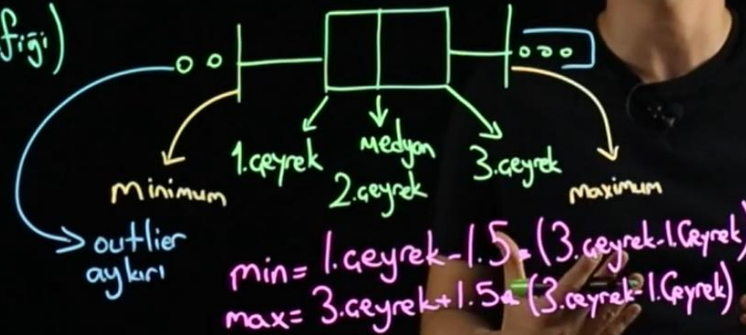
- `barplot()` - Sütun / Çubuk Grafiği



- `hist()` - Histogram



- `boxplot()` - Kutu grafikleri



- `par()` - Grafik ırgarası ile çoklu gösterim.

- Görselleştirme kütüphanelerini kullanmak (lattice - ggplot2)

R ile Tahmin Analitiğine Giriş

R ile Veri Ön İşleme Örneği

- Imputation
- Normalizasyon

R ile Linear Regresyon

Simple Linear Regression

- `model1<- lm(Y~X1)`
- `plot(Y~X1)`
 - `abline(model1, col="blue", lwd=3)`
- `tahmin1<- predict(model1,data.frame("X1"=10))`

Multiple Linear Regression

- `model3 <- lm(Y ~ X1 + X2 + X3, data = data_table)`

R ile k-NN Sınıflandırma

k-NN algoritması için komsu sayısı genelde satır sayısının karekökü kadar belirlenir

```
model5<- knn(train = data_table_train,  
             test = data_table_test,  
             cl=hedef_degisken, k=komsu)
```

R ile Tahmin Analitiğine Giriş Özeti

- Veri Ön İşleme-Hazırlama (Data Preprocessing – Data Preparation)
 - Imputation (Eksik Değerleri doldurma)
 - Normalizasyon (Farklı değer aralıklarında olan değişkenler arasındaki farkları normelleştirme)
- Regresyon Modellemesi
 - Numerik değer tahmin etme
 - Bir değişken ile tahminleme (Simple Linear Regression)
 - Birden çok değişken ile tahminleme (Multiple Linear Regression)
- Sınıflandırma Modellemesi
 - Sınıf tahmin etme
 - Train-test set ayrımı -> Genellenebilir modeller

R ile Tahmin Analitiğine Giriş Özeti

5.4. Bölüm Özeti

- Veri Ön İşleme (Hazırlama)
Data Preprocessing (Preparation)
 - Imputation (Eksik değer doldurma)
 - Normalizasyon

- Regresyon Modellemesi
 - Numerik değer tahmini
 - Bir değişken ile regresyon
(Simple Linear Reg.)
 - Birden çok değişken ile Regresyon
(Multiple Linear Reg.)

- Sınıflandırma Modellemesi
 - Hangi ay → Sınıf
 - Train-Test set ayrımı
 - ↓
 - Genellenebilir Modeller