İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ 2010

EĞİTİMDE BÜYÜK VERİ UYGULAMALARI VE ÖĞRENME ANALİTİKLERİ

4005 TÜBİTAK

educatioANalitics

EĞİTİMDE BÜYÜK VERİ UYGULAMALARI VE ÖĞRENME ANALİTİKLERİ

BÜYÜK VERİLERDE VERİ ÖN İŞLEME

Prof. Dr. Ferhan ELMALI
Öğr. Gör. Dr. Berhan ÇOBAN

R Studio

4005 TÜBİTAK

Tübitak 4005 Yenilikçi Eğitim Uygulamaları Destekleme Programı

BÜYÜK VERİ (*Big Data*)

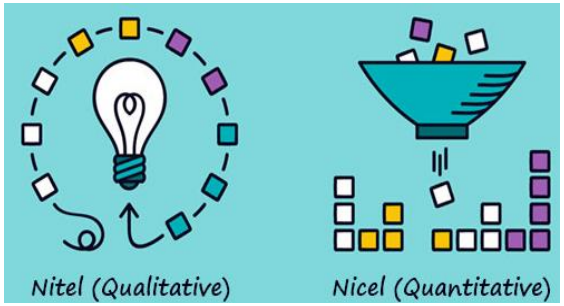
❖VERİ Nedir ?

işlenmemiş gerçek enformasyon parçacığına verilen addır. Veriler

- ölçüm, sayım,
- deney, gözlem
- araştırma yolu ile elde edilmektedir.

Ölçüm ya da sayım yolu ile toplanan ve sayısal bir değer bildiren veriler **nicel veriler**, sayısal bir değer bildirmeyen veriler de **nitel veriler** olarak adlandırılmaktadır.

İstatistiksel Veri Türleri



1. Sınıflama (Nominal) **ölçek**. → Örn: Cinsiyet, Meslek
2. Sıralama (Ordinal) **ölçek**. → Örn: öğrenim durumu, önem derecesi
3. Eşit aralıklı (Interval) **ölçek**. → Örn: Sıcaklık, zeka /başarı puanı
4. Oranlı (Ratio) **ölçek**. → Örn: «0» anlamlı. Boy, yaş, uzunluk

BÜYÜK VERİ (*Big Data*)

❖ BÜYÜK VERİ (*Big Data*) nedir ?

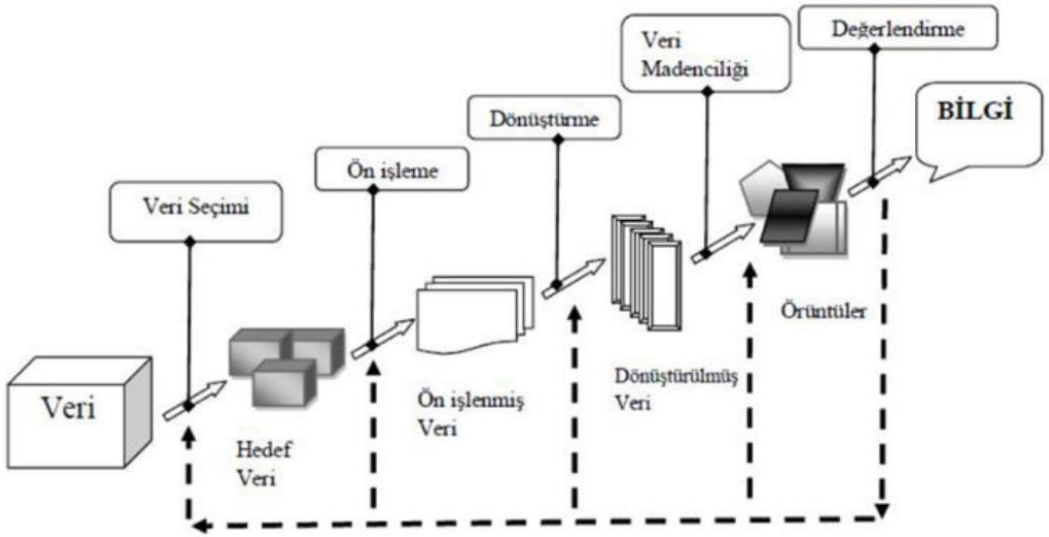
Bilgi teknolojilerinin gelişimi, sensörler, internet, sosyal medya, arama motorları vb. araçların ürettiği yüksek hacimli, yüksek hızlı ve / veya çok çeşitli veri türlerine denir.

❖ *Büyük Veri Bileşenleri nelerdir ?*

- çeşitlilik (variety),
- hız(velocity),
- verinin hacmi (volume),
- doğrulama (verification),
- doğruluk (veracity) ve
- değerden(value) oluşmaktadır

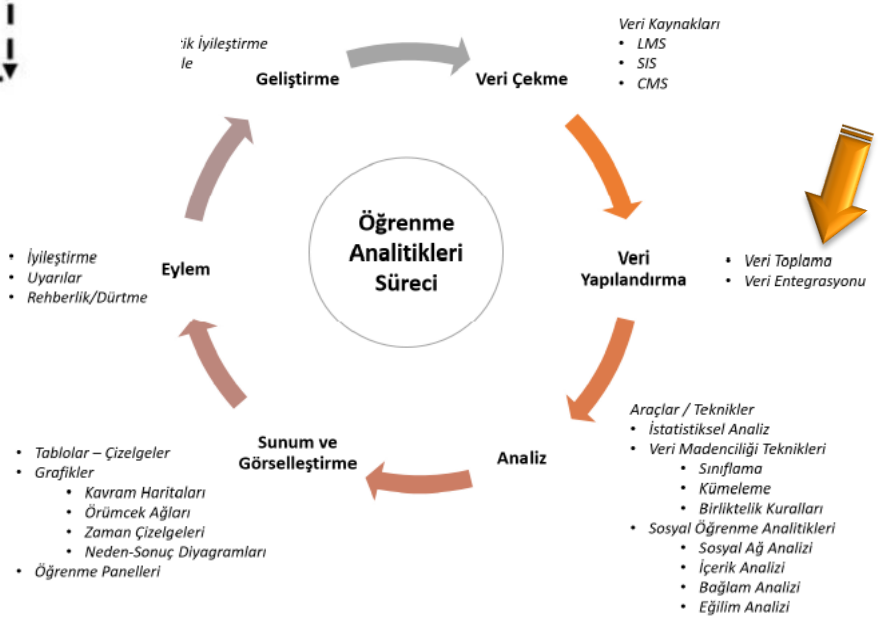
BÜYÜK VERİ (*Big Data*)

Büyük Veri İşleme Aşamaları



Suhirman, Zain, & Herawan, 2014

Öğrenme Analitiğinde Verinin Yeri



Şekil 1.3. Öğrenme Analitikleri Süreci (Lal, 2014'den uyarlanmıştır.)

Veri birçok alanda olduğu gibi öğrenme Analitiği Çalışmalarında da ilk basamakta yer almaktadır.

Doğru, zamanında ve yeterli verinin kullanılması

Çalışmaların kalitesini, modellerin geçerliliğini olumlu yönde etkileyecektir.

BÜYÜK VERİ (*Big Data*)

Büyük Veri kullanımına Yönelik Sektörel Örnekler

İletişim, Medya ve Eğlence Sektörlerinde

Devlet Hizmetlerinde

Sağlık Hizmetinde

Sigortacılıkta

Ulaşımda

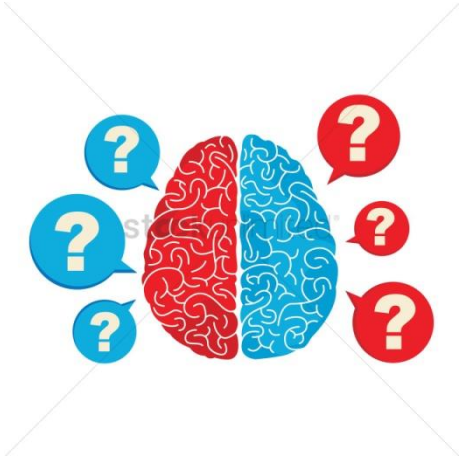
Ulaşımda

Büyük Veri kullanımına Yönelik Eğitim Alanı Örnekleri

Mobil uygulamalar

Öğrenci takip programları

Başarı İzleme uygulamaları



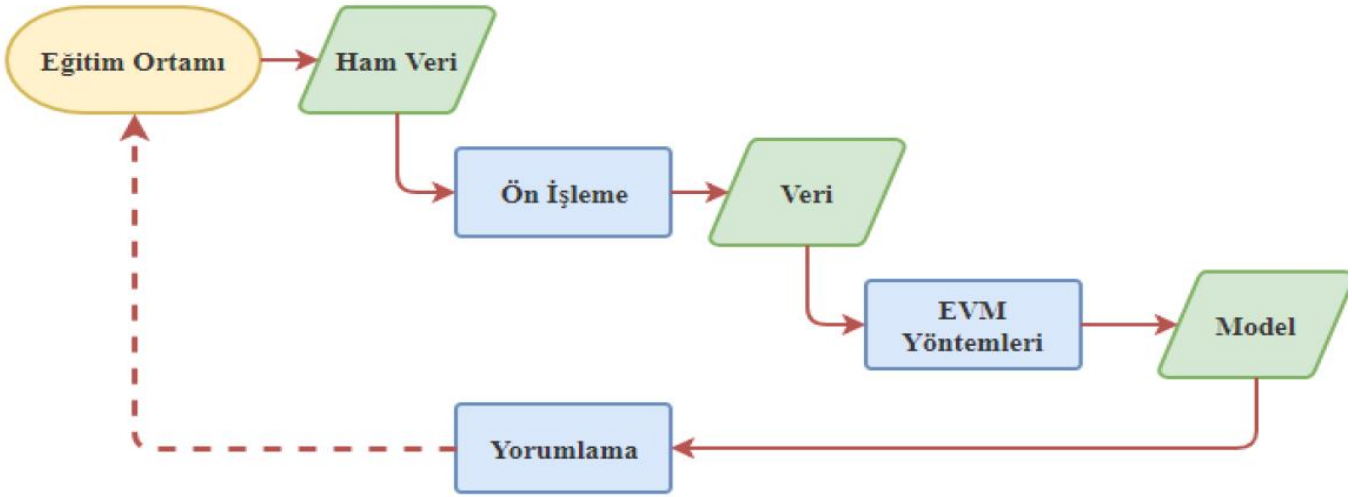
Veri Önişleme

Veri Analizinde Önişleme Neden Gerekli ?

Verilerin yapılandırılmış, ilişkisel formatlarda yer alabilmesi için uygun biçim ve özelliklere sahip olması gerekmektedir. Eksik değerler, Aykırı değerler, Mantıksız değerler Kullanılması düşünülen modele uygun olmayan girdi formatı

Uygun formata sahip olmayan verilerin analiz edilmesi mümkün olamayacağından yığın olarak kalmaya devam edecek ve bir değere dönüşemeyecektir.

Bu sistemin yavaşlaması, önemli bilginin kaybolması, modellerin yetersiz kurulması gibi bir çok yan etkilere sebep olacaktır.



Eğitsel Veri Madenciliğinde Veri Önişleme

Linan ve Perez (2015)'e göre, EVM uygulamalarının genelleştirilmiş süreci sırasıyla aşağıdaki gibi açıklanmaktadır:

1. İlk olarak yapılacak çalışmanın amacı ve bu amaca yönelik toplanılacak veriler belirlenerek, ilgili eğitim ortamından çıkartılır. Bahsedilen eğitim ortamı geleneksel, çevrimiçi veya karma-hibrit eğitim ortamı olabilir.

2. Genellikle eğitim ortamlarından gelen veriler farklı formatlarda ve hiyerarşi seviyelerinde tutuldukları için, analizde kullanılacak şekilde ön-işlemeye tabi tutulmaları gerekmektedir.

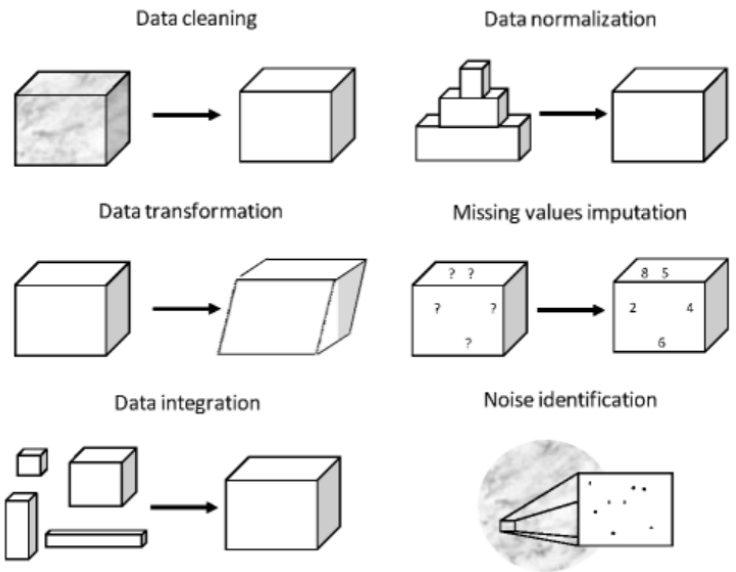
3. Ön işlemden geçirilmiş ve analize hazır verilere EVM yöntemlerinin uygulanmasıyla, yorumlanması gereken modeller ve ya desenler elde edilir.

4. Sonuçlar yorumlanır. Elde edilen bilgiye göre eğitim ortamını, öğrenim ve öğretim süreçlerini iyileştirecek önerilerde bulunulur.

KISACA....

Veri ön işleme; veri madenciliği modelleri kurulmadan önce veri seti üzerinde yapılan bir takım düzeltme, eksik veriyi tamamlama, tekrarlanan verileri kaldırma, dönüştürme, bütünleştirme, temizleme, normalleştirme, boyut indirgeme vb. işlemlerdir.

1. Veri Temizleme
2. Veri Birleştirme
3. Veri Dönüştürme
4. Veri İndirgeme



Veri Ön İşleme Teknikleri

1. Veri Temizleme

Veri temizleme, eksik verilerin tamamlanması, aykırı değerlerin teşhis edilmesi amacıyla gürültünün düzeltilmesi ve verilerdeki tutarsızlıkların giderilmesi gibi işlemleri içermektedir.

Örnekler;

Eksik veri: formdaki bazı özelliklerin girilmemiş olması (dersin notu, öğrencinin boyu)

Aykırı veri: mantıksız veya aykırı değerlerin olması (gelir:-20, yaş:112)

Tutarsız veri: farklı kaynaklardan gelen aynı verideki tutarsızlık (ders notu: 100 lük - . harf sistemi)

İhmal etme, elle doldurma, otomatik olarak doldurmak, ortalamanın yazılması, …

2. Veri Birleştirme

Veri tabanlarındaki veriler tek bir çatı altında birleştirilirler. Farklı veri tabanlarındaki verilerin tek bir veri tabanında birleştirilmesiyle şema birleştirme hataları (schema integration errors) oluşur.

Örneğin, bir veri tabanında girişler “tüketici-ID” şeklinde yapılmışken, bir diğerinde “tüketici-numarası” şeklinde olabilir.
Veriye ilişkin veri olan meta veri oluşturularak çözülebilir. Aynı bilgiyi veren sütünlrdan biri dikkate alınıp diğerleri silinebilir.

Bu tip sorunlarla karşılaşmamak için veri tabanları tanımları önceden itina ile yapılmalıdır.

Öğrenci No:
Ders Saati:
Ders Notu:
Sözlü Notu:
Bölümü: X

3. Veri Dönüştürme

Veri dönüştürme ile veriler, veri madenciliği için uygun formlara dönüştürülürler.

Veri dönüştürme; düzeltme, birleştirme, genelleştirme ve normalleştirme gibi değişik işlemlerden biri veya bir kaçını içerebilir.

1. Min-Max
2. Z Skor
3. Ondalık Ölçekleme

4. Veri İndirgeme

Veri indirgeme teknikleri, daha küçük hacimli olarak ve veri kümesinin indirgenmiş bir örneğinin elde edilmesi amacıyla uygulanır. Bu sayede elde edilen indirgenmiş veri kümesine veri madenciliği teknikleri uygulanarak daha etkin sonuçlar elde edilebilir.

1. Veri Birleştirme veya Veri Küpü (Data Aggregation or Data Cube)
2. Boyut indirgeme (Dimension Reduction)
3. Veri Sıkıştırma (Data Compression)
4. Kesikli hale getirme (Discretization)

Veri Ön İşleme Teknikleri R Uygulamaları

R, 1970lerin sonunda Bell laboratuarlarında geliştirilen S programının devamı niteliğinde olan ücretsiz, kod esaslı, açık kodlu ve çoğunlukla istatistiki analiz ve veri madenciliği alanında kullanılan bir programdır.

Yardım

?t.test
?histogram … vb.

https://www.rstudio.com/products/rstudio/download/

https://cran.r-project.org/bin/windows/base/

Veri Ön İşleme Teknikleri R Uygulamaları

Temel R programı Bilgileri

```
t<-c(1,5,7,9,8,5,11,23,15,8,7)
print(t)
plot(t)
```

```
çarp=function(x,y) x*y
çarp(5,6)
```

```
dairealanı=function(r){
  p=3.14
  alan=r*r*p
  print(alan)}
```

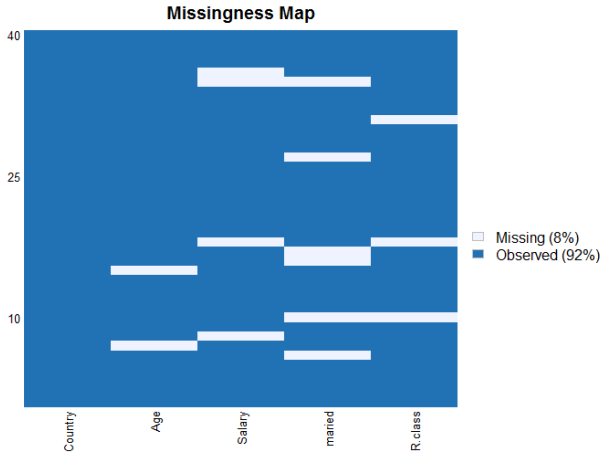
```
x[,4] # 4th column of matrix
x[3,] # 3rd row of matrix
x[2:4,1:3] # rows 2,3,4 of columns 1,2,3
```

```
ort=function(k)
  sum(k)/length(k)
k=2:20
ort(k)

mean(k)
```

```
median(data)
var(data) # Variance
sd(data) # Standard deviation
max(data) # Max value
min(data) # Min value
range(data) # Range
```

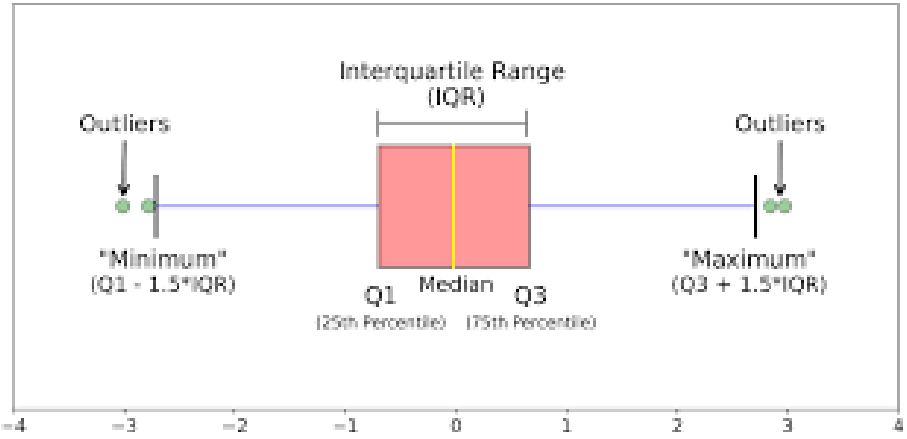
Eksik Veri Uygulamaları

```
library(xlsx)
Library(Amelia)

data=read.xlsx("C:/Users/User/Desktop/R/predata.xlsx",1)   # veriyi okuyor
data=data.frame(data)

View(data)                # veri tablosu

sum(is.na(data))          # kaç tane eksik gözlem var onu gösterir
require(Amelia)
missmap(data, rank.order = F)

mean(data$Age)            # NA eksik veri olduğundan hesaplamadı
mean(data$Age, na.rm = T)

mean(data$Salary)         # NA eksik veri olduğundan hesaplamadı
mean(data$Salary, na.rm = T)   #eksik gözlemi dikkate almadan ort. hesaplıyor

satırsil=na.omit(data)    # eksik veri olan satırı siliyor
satırsil


data$Age = ifelse(is.na(data$Age),
                  ave(data$Age, FUN = function (x)mean(x, na.rm = TRUE)),data$Age)
data$Salary = ifelse(is.na(data$Salary),
                     ave(data$Salary, FUN = function (x)mean(x, na.rm = TRUE)),data$Salary)
```



Aykırı Veri Uygulamaları

```
boxplot(data$Age)
hist(data$Age)
stem(data$Age)
```




```
set.seed(10)                    # herkesin aynı veriyi üretebilmesi için bir kod

veri <- c(rnorm(100, mean = 15, sd = 5), -5,30,55)

boxplot(veri, horizontal = TRUE)
hist(veri)
stem(veri)
```

Veri Normalizasyonu Uygulamaları

0-1 standartlaştırması

```
x=data$Salary
a=min(data$Salary)
b=max(data$Salary)
newsalary= function(x){(x-a)/(b-a)}
newsalary(x)
```

z score standartlaştırması

```
a=sd(x)
b=mean(x)
znewsalary= function(x){(x-b)/a}
znewsalary(x)

plot(znewsalary(x), type="o", col="green")
```

```
> data$Salary
 [1] 72000.00 48000.00 54000.00 61000.00 44444.44 44444.44 58000.00 52000.00 79000.00 83000.00 67000.00 45000.00 52000.00 51000.00
[15] 32000.00 75000.00 23000.00 36000.00 33000.00 22000.00 42000.00 33000.00 44444.44 19000.00 50000.00 40000.00 47000.00 48000.00
[29] 29000.00 38000.00 28000.00 39000.00 44444.44 25000.00 26000.00 34000.00 39000.00 38000.00 37000.00 45000.00
> x=data$Salary
> a=min(data$Salary)
> b=max(data$Salary)
> newsalary= function(x){(x-a)/(b-a)}
> newsalary(x)
 [1] 0.8281250 0.4531250 0.5468750 0.6562500 0.3975694 0.3975694 0.6093750 0.5156250 0.9375000 1.0000000 0.7500000 0.4062500
[13] 0.5156250 0.5000000 0.2031250 0.8750000 0.0625000 0.2656250 0.2187500 0.0468750 0.3593750 0.2187500 0.3975694 0.0000000
[25] 0.4843750 0.3281250 0.4375000 0.4531250 0.1562500 0.2968750 0.1406250 0.3125000 0.3975694 0.0937500 0.1093750 0.2343750
[37] 0.3125000 0.2968750 0.2812500 0.4062500
> a
[1] 19000
> b
[1] 83000
```

Veri Ön İşleme Teknikleri R Uygulamaları

Aynı sonucu veren iki sütundan veya satırdan birini silme

A=matrix(c(1,2,3,4,5,6), 3, 2, dimnames = list(c("X","Y","Z"), c("A","B")))

```
  A B
X 1 4
Y 2 5
Z 3 6
```

A = cbind(A, c(10,11,12))

colnames(A)[3]="C"

```
  A B  C
X 1 4 10
Y 2 5 11
Z 3 6 12
```

```
  A B  C
X 1 4 10
Y 2 5 11
Z 3 6 12
```

A = rbind(A, c(20,21,22))

rownames(A)[4]="W"

```
  A  B  C
X 1  4 10
Y 2  5 11
Z 3  6 12
W 20 21 22
```

A =A[-4,] **satırdan birini**

A =A[,-3] **sütundan birini**

```
> data
   Country       Age   Salary maried R.class
1   France  44.00000 72000.00     No       5
2    Spain  27.00000 48000.00    Yes       2
3  Germany  30.00000 54000.00     No       1
4    Spain  38.00000 61000.00     No       4
5  Germany  40.00000 44444.44    Yes       2
6   Turkey  36.00000 44444.44   <NA>       4
7   France  35.00000 58000.00    Yes       3
8    Spain  28.00000 52000.00     No       5
9   France  48.00000 79000.00    Yes       5
10 Germany  50.00000 83000.00     No      NA
11  France  37.00000 67000.00    Yes       2
12  France  41.00000 45000.00     No       2
13   Spain  52.00000 52000.00     No       3
14 Germany  36.00000 51000.00   <NA>       4
15  Turkey  70.00000 32000.00    Yes       1
16   Spain  85.00000 75000.00     No       2
```

data[,-1]

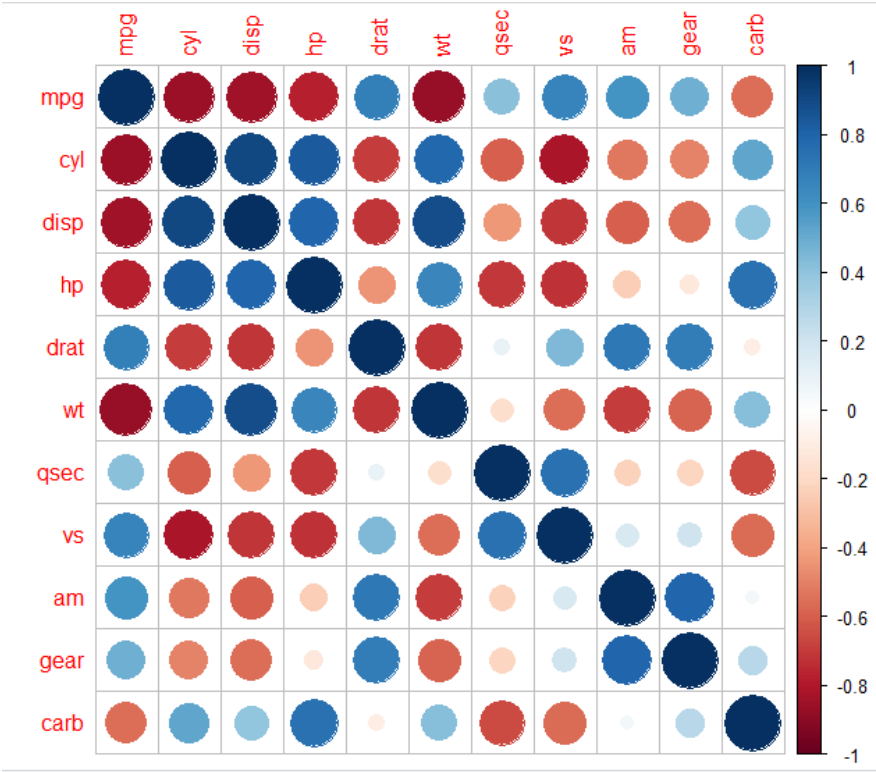
```
       Age   Salary maried R.class
1   44.00000 72000.00     No       5
2   27.00000 48000.00    Yes       2
3   30.00000 54000.00     No       1
4   38.00000 61000.00     No       4
5   40.00000 44444.44    Yes       2
6   36.00000 44444.44   <NA>       4
7   35.00000 58000.00    Yes       3
8   28.00000 52000.00     No       5
9   48.00000 79000.00    Yes       5
10  50.00000 83000.00     No      NA
11  37.00000 67000.00    Yes       2
12  41.00000 45000.00     No       2
13  52.00000 52000.00     No       3
14  36.00000 51000.00   <NA>       4
15  70.00000 32000.00    Yes       1
16  85.00000 75000.00     No       2
17 112.00000 23000.00     no       1
18  32.00000 36000.00    Yes       2
```

Principal Component Analizi (PCA) – Temel Bileşen Analizi

PCA nın boyut indirgemesi veri setindeki mevcut aralarında korelasyon olan değişkenleri bazı linear transformasyonlarla aynı sayıda ama aralarında korelasyon olmayan (ortogonal) değişkenlere dönüştürmeye dayanmaktadır.

Korelasyon

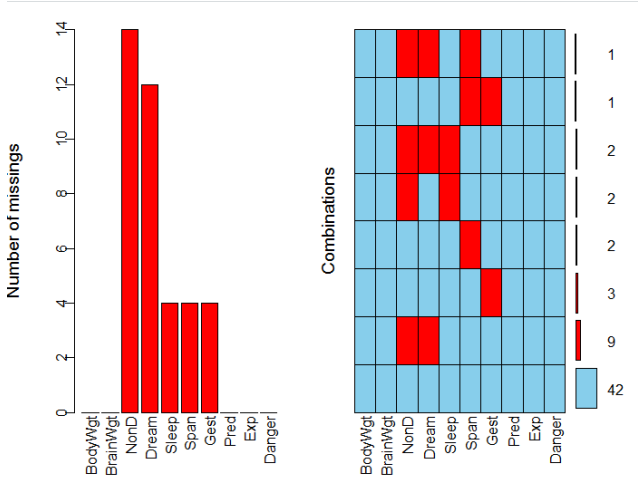
```
summary(mtcars)
head(mtcars)
                # mpg: Mil/Galon
                # cyl: Silindir sayısı
                # disp Silindir Hacmi
                # hp: Toplam Beygir Gücü
                # drat: Arka Aks Oranı
                # wt: Ağırlık (lb/1000)
                # qsec 1/4 mil zamanı
                # vs: V/S motor silindir çeşidi
                # am: Vites Türü (0 = otomatik, 1 = Manuel)
                # gear: İleri vites sayısı
                # carb: Karbürator sayısı
korelasyon_matris <- cor(mtcars)
korelasyon_matris
                # install.packages("corrplot")
library(corrplot)
corrplot(korelasyon_matris, method = "circle")
```

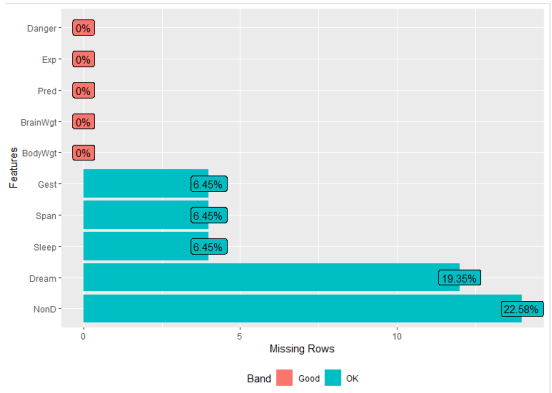


Veri Ön İşleme Teknikleri R Uygulamaları

Mice, Amelia paketleri

```
library(mice)
library(Amelia)
library(missForest)
library(Hmisc)
library(mi)
library(VIM)
library(DataExplorer)
library(ggplot2)
library(caret)
library(formattable)
introduce(sleep)

sleep[!complete.cases(sleep),]   #eksik verileri gösteriyor

missmap(sleep, col=c('grey', 'steelblue'), y.cex=0.5, x.cex=0.8)
aggr(sleep, prop = F, numbers = T)                              #grafikler
plot_missing(sleep)

tamamlanmis <- mice(sleep, method = "cart")  #karar ağacı metodu vb.
tamamlanmis
densityplot(tamamlanmis)                    #üretilen eksik verinin bilgisi
plot(tamamlanmis)

tamamlanmis$imp          #üretilen eksik veriler

formattable(head(mice::complete(tamamlanmis,2),10)) # üretilen veriyi birleştirdi.
                                    #   2. tercihi uygun gördük
```





Yeni Veri Setini Kaydetme

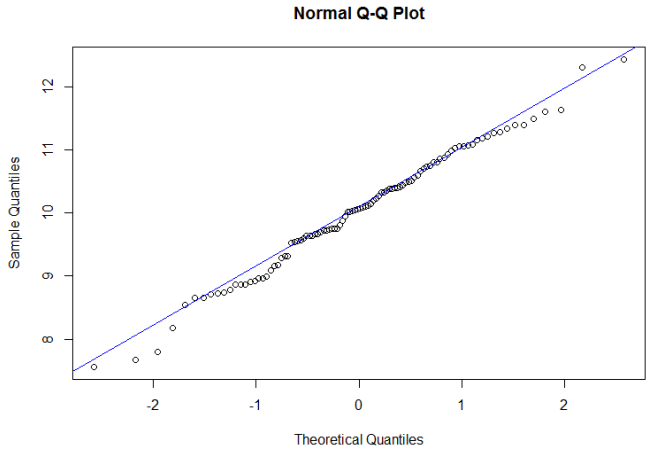
```
library(readr)

write.csv(data,"C:/Users/User/Desktop/R/yenidat.csv")
```

midastouch	any	Weighted predictive mean matching
sample	any	Random sample from observed values
cart	any	Classification and regression trees
rf	any	Random forest imputations
mean	numeric	Unconditional mean imputation
norm	numeric	Bayesian linear regression
norm.nob	numeric	Linear regression ignoring model error
norm.boot	numeric	Linear regression using bootstrap
norm.predict	numeric	Linear regression, predicted values
quadratic	numeric	Imputation of quadratic terms
ri	numeric	Random indicator for nonignorable data
logreg	binary	Logistic regression
logreg.boot	binary	Logistic regression with bootstrap
polr	ordered	Proportional odds model
polyreg	unordered	Polytomous logistic regression
lda	unordered	Linear discriminant analysis
2l.norm	numeric	Level-1 normal heteroscedastic
2l.lmer	numeric	Level-1 normal homoscedastic, lmer
2l.pan	numeric	Level-1 normal homoscedastic, pan
2l.bin	binary	Level-1 logistic, glmer
2lonly.mean	numeric	Level-2 class mean
2lonly.norm	numeric	Level-2 class normal
2lonly.pmm	any	Level-2 class predictive mean matching

Normal dağılım varsayımı

Sapiro-Wilk test

```
library("ggpubr")
veri5 <- c(rnorm(100, mean = 10, sd = 1))
hist(veri5)
shapiro.test(veri5)

 qqnorm(veri5)
 qqline(veri5, col = "blue")
```




```
library("ggpubr")
veri5 <- c(rnorm(100, mean = 10, sd =
5),100,85,63)
hist(veri5)
shapiro.test(veri5)

 qqnorm(veri5)
 qqline(veri5, col = "blue")
```

Normal olmayan veri

Ho: veri normal dağılmaktadır.

EĞİTİMDE BÜYÜK VERİ UYGULAMALARI VE ÖĞRENME ANALİTİKLERİ

KATILIMINIZ İÇİN TEŞEKKÜRLER

4005 TÜBİTAK

Tübitak 4005 Yenilikçi Eğitim Uygulamaları Destekleme Programı

ÇİĞLİ BELEDİYESİ

TRABZON ÜNİVERSİTESİ 2018

ANKARA ÜNİVERSİTESİ 1946

Kaynakça

Suhirman, Zain, J., & Herawan, T. (2014). Data Mining for Education Desicion Support: A Review. *International Journal of Emerging Technologies in Learning.*

Calvet Liñán, L., Juan Pérez, Á.A. Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. *Int J Educ Technol High Educ* **12,** 98–112 (2015). https://doi.org/10.7238/rusc.v12i3.2515

Kokoç, M. (2016). *E-Öğrenme Ortamlarında Bir Öğrenme Analitiği Aracı Olarak Öğrenme Panelleri İle Etkileşimin Öğrenme Çıktılarıyla İlişkisi. Doktora Tezi.*

RStudio Cloud: https://login.rstudio.cloud/
The R Project for Statistical Computing. https://www.r-project.org/

Oğuzlar,A.,(2003), Veri Ön İşleme, Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, Sayı: 21, Temmuz-Aralık 2003, ss. 67-76.

HAN, J. and M. KAMBER (2001), **Data Mining: Concepts and Techniques,**
Morgan Kaufmann Publishers, USA, 550p.

ROIGER, R. J. and M. W. GEATZ (2003), Data Mining A Tutorial-Based
Primer, Addison Wesley, USA, 350p.

Buuren, Stef van, Karin Groothuis-Oudshoorn, Alexander Robitzsch, Gerko Vink, Lisa Doove, ve Shahab Jolani. 2015. “mice: Multivariate Imputation by Chained Equations”. https://cran.r-project.org/package=mice.

https://www.veribilimiokulu.com/eksik-veri-eksik-veride-kullanilabilecek-algoritmalar/

https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/

https://tevfikbulut.com/2020/11/29/makine-ogrenme-yontemleri-kullanarak-eksik-verilere-atama-yapilmasi-uzerine-bir-vaka-calismasi-a-case-study-on-assigning-missing-data-using-machine-learning-ml-methods/

Tutgun Ünal, A. (2014). *Büyük veri ve eğitimsel veri madenciliğinin eğitim alanına katkılarının incelenmesi, 8. Uluslararası Bilgisayar ve Öğretim Teknolojileri Sempozyumu, Edirne.*