Prepared by

Dilan Mevis

Report Of Datasets

2023

gIRIS

Raporun Amaci:

Bu rapor, bitki yetiştiriciliği, sera etkisi, çim büyümesi ve karpuz üretimi gibi farklı konularla ilgili bes farklı veri setinin analizini sunmaktadır. Amacımız, her bir veri setinin ayrıntılı analizini yaparak sonuçlarını tartışmak ve bu veri setlerinden elde edilen bilgileri kullanarak urunlerimiz üzerinde ne tur işlemler yapildiginda daha iyi verim elde edeceğimizi gorecegiz. Raporun bu kısımları, bu konulara ilişkin bilgileri artırmak ve bitki yetiştiriciliği, sera etkisi, çim büyümesi ve karpuz üretimi gibi konularla ilgili araştırmaları desteklemek için önemlidir.

Analizler, R programlama dili kullanılarak gerçekleştirilmiştir.

Rapor, her bir veri setinin analizini içermektedir. İlk adım olarak, tanımlayıcı istatistikler kullanarak her veri setini daha yakından tanımaya çalışacağız. Bu, verilerin merkezi eğilim, dağılım ve diğer önemli özelliklerini belirlememize yardımcı olacaktır. Ayrıca, görselleştirme araçlarını kullanarak, verilerin özelliklerini daha net görebilir ve farklı değişkenler arasındaki ilişkileri görsel olarak analiz edebiliriz. Ardından, normallik testi uygulayarak, verilerin normal dağılım gösterip göstermediğini değerlendireceğiz.

COTTON dATASET

Bu veri seti "cotton" (pamuk) olarak adlandırılan bir bitki türü ile ilgilidir. Veri seti, Peru'nun Lima ve Pisco bölgelerinde gerçekleştirilen deneylerde toplanan pamuk verilerini içerir. 96 gözlem ve 5 değişkenden oluşmaktadır.

**"epoca"** değişkeni, pamukların hasat edildiği zamanı temsil eder. Sayısal bir değişkendir.

1: Birinci Hasat

2: Ikinci Hasat

**"yield"** değişkeni, pamuk verimini temsil eder. Sayısal bir değişkendir.pamuk uretiminde elde edilen hasat miktari

**"lineage"** değişkeni, pamuk bitkisinin soyunu belirten sayısal bir değişkendir. 4 farkli soyu temsil eder

**"block"** değişkeni, deneydeki blokları belirtir. "I", "II", "III", "IV", "V" ve "VI" seviyelerine sahip bir faktör değişkenidir.Bloklar, deneyin yapıldığı alanda homojen olmayan koşulların neden olduğu değişkenlikleri kontrol etmek için kullanılan bir yöntemdir. Deney alanı homojen olmadığından, farklı bölgelerdeki bitkilerin farklı büyüme koşullarına maruz kalması mümkündür. Bloklar, deney alanının farklı bölgelerindeki toprak yapısı, güneş ışığı, su kaynakları, bitki hastalıkları, böcek istilaları gibi koşullardaki değişkenliği en aza indirmek için kullanılır.

**"site"**değişkeni, deneyin yapıldığı alanı iki farklı bölgeye ayırır. "Lima" ve "Pisco" bölgesi arasında farklı iklim koşulları, toprak özellikleri, bitki hastalıkları, böcek istilaları vb. gibi değişkenler olabilir. Bu nedenle, deneyin sonuçlarını analiz ederken, farklı bölgeler arasındaki farklılıkları incelemek için "site" değişkeni kullanılabilir.

Cotton Veri Setinin Analizi

library(ggplot2)

p <- ggplot(cotton, aes(x = site, y = yield)) +

theme\_minimal() +

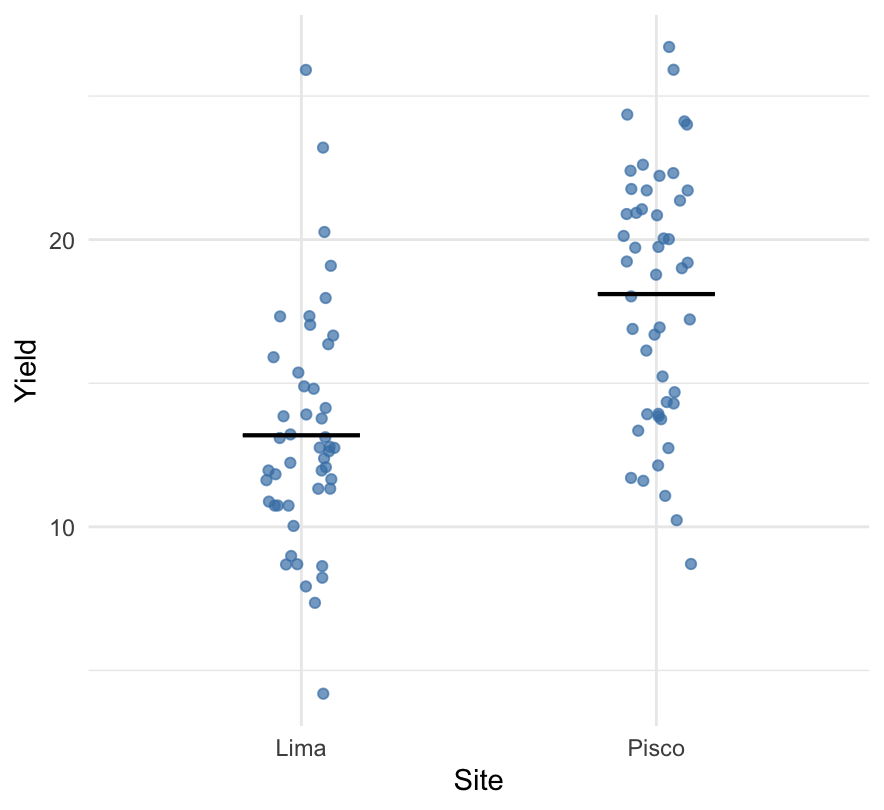
geom\_jitter(alpha = 0.7, width = 0.1, na.rm = T, color = "steelblue") +

stat\_summary(fun = "mean", fun.min = "mean", fun.max = "mean", size = 0.3, width = 0.33, geom = "crossbar") +

ylab("yield") +

xlab("site")

plot(p)



|  |
| --- |
| Bu grafik, pamuk üretimi verilerini görselleştirmektedir. X ekseni "Site" olarak belirlenmiş, Y ekseni ise "Yield" olarak belirlenmiştir. Her bir nokta, belirtilen sitelerde elde edilen ürün miktarını temsil etmektedir. Ayrıca, her bir site için ortalama ürün miktarı, çubuklar ve kesişen çizgilerle gösterilmiştir. Bu grafik, siteler arasında pamuk üretimindeki farklılıkları ve hangi sitelerin daha yüksek ürün elde ettiğini kolayca göstermektedir.  Bu hesaplamalar sonucunda, Pisco bölgesinde daha yüksek bir ortalama pamuk verimi olduğu görülmektedir. Dolayısıyla, Pisco bölgesinde daha çok pamuk elde edilmiş gibi görünmektedir. |

Tanimlayici istatistikler ile veriyi daha iyi taniyalim.

min(cotton$yield)

## [1] 4

“Veri Setinde en düşük pamuk verimi 4’türve Lima Bölgesine aittir.”

max(cotton$yield)

## [1] 27

“Veri Setinde en yüksek pamuk verimi 27’dir ve Pisco Bölgesine aittir.”

range(cotton$yield)

## [1] 4 27

“Veri setinde en yüksek ve en düşük değer”

mean(cotton$yield)

## [1] 15.64583

“Cotton veri setindeki pamuk veriminin ortalama değeri 15.64583'tür.”

median(cotton$yield)

“Cotton veri setinin pamuk verimlerine bakıldığında, ortanca verim değerinin 15.3 olduğu görülmüştür.

quantile(cotton$yield, 0.25)

## 12%

“Cotton veri setinde, verimlerin %25'inin 12'den küçük veya ona eşit olduğu görülmektedir.”

tablo <- table(cotton$yield)

mod <- as.numeric(names(which.max(tablo)))

print(mod)

## 12

“Cotton veri setinde en sık tekrar eden pamuk verimi 12'dir.”

IQR(cotton$yield)

## 8

“Cotton veri setindeki pamuk verimlerinin çeyreklikler arasındaki fark (IQR) 8'dir.”

sd(cotton$yield)

## [1] 4.860104

“Standart sapma, pamuk verimi değişkeninin ortalamadan ne kadar sapma gösterdiğini ölçen bir istatistik değeridir. Verilerin ortalamadan ne kadar uzaklaştığını gösterir. Cotton veri setinde pamuk verimi değişkeninin standart sapması 4.860104 olarak hesaplanmıştır. Bu, pamuk verimindeki farklılıkların ortalama değerden 4.860104 birim kadar sapma gösterdiği anlamına gelir. Yüksek standart sapma değerleri, verilerin dağılımının geniş olduğunu, düşük standart sapma değerleri ise verilerin daha homojen bir şekilde dağıldığını gösterir.”

var(cotton$yield)

## [1] 23.62061

“Veri setindeki pamuk verimlerinin dağılımının ortalamaya yakın olduğunu söyleyebiliriz. Çünkü standart sapma ve varyansın değeri, verilerin ortalama değerden ne kadar uzaklaştığını gösterir. Bu durumda, standart sapmanın 4.860104, varyansın ise 23.62061 olduğunu göz önünde bulundurarak, verilerin ortalama değerden biraz uzaklaştığını ancak dağılımın hala göreceli olarak düşük olduğunu söyleyebiliriz.”

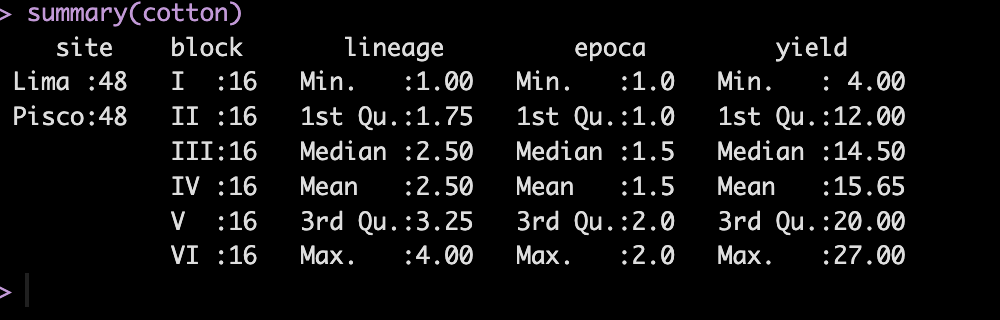
sd(cotton$yield) / mean(cotton$yield)

## [1] 0.3106325

Varyasyon Katsayisi Bu ifade, pamuk verimi değerlerinin standart sapmasının, ortalama değerin yaklaşık %31'ine eşit olduğunu gösterir. Bu oran, verilerin ne kadar değişken olduğunu ve birbirlerine ne kadar yayıldıklarını gösterir. Daha yüksek bir oran, daha fazla değişkenliği gösterirken, daha düşük bir oran daha az değişkenlik gösterir.

Herbir değeri bu sekilde tek tek yazip elde etmek yerine bazi komutlar ile bu değerleri doğrudan elde edebiliriz.

summary(cotton)



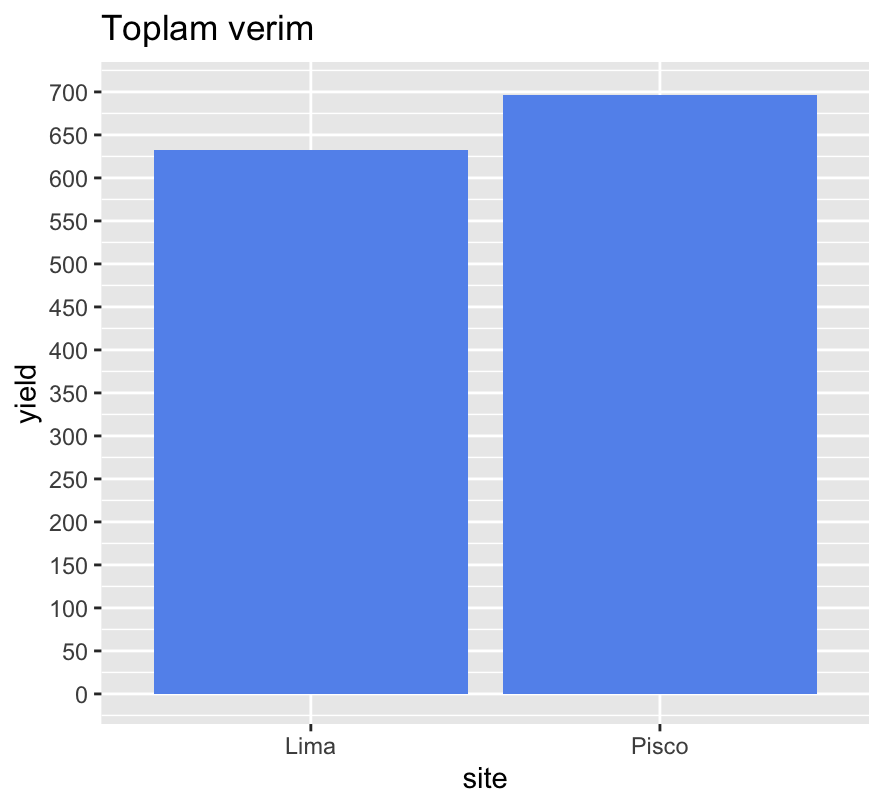
Verileri görselleştirelim

ggplot(cotton, aes(x = site, y = yield)) +

geom\_bar(stat = "identity", fill = 'cornflowerblue') +

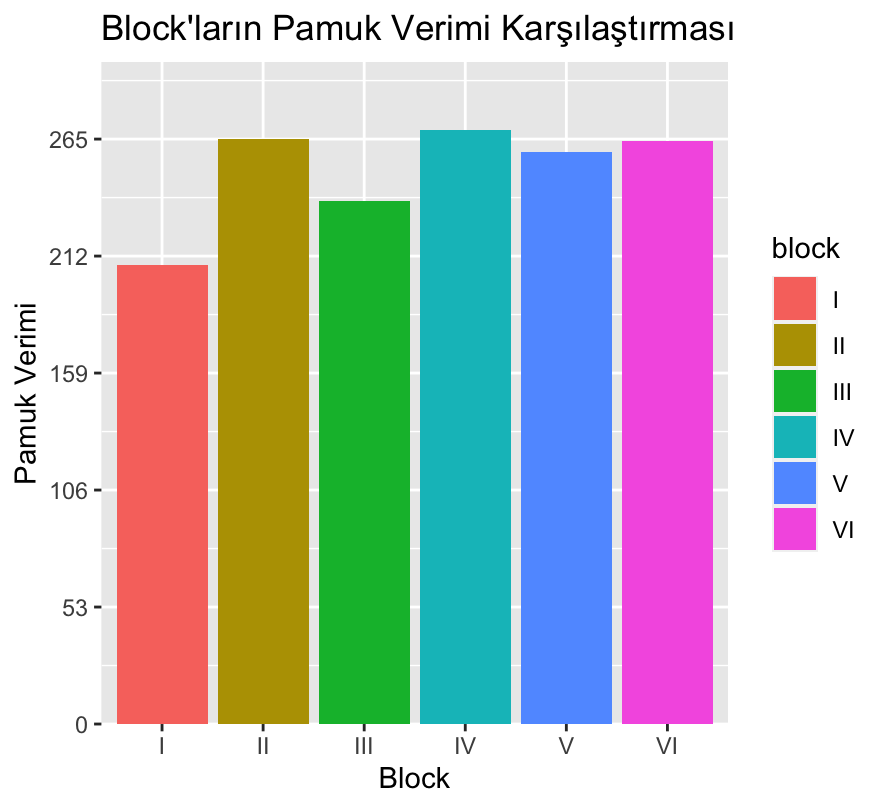
labs(x = "site", y = "yield", title = "Toplam verim") +

scale\_y\_continuous(limits = c(0, 700), breaks = seq(0, 700, by = 50))



Grafiğe baktığımızda, "pisco" bölgesindeki verim diğer bölgelere göre daha yüksek olduğu görülebilir. Bu, pisco bölgesinin diğer bölgelere göre daha elverişli koşullara sahip olabileceğini veya pisco bölgesindeki üreticilerin daha iyi yöntemler kullanarak daha yüksek verim elde ettiğini gösterebilir. Ancak, bu sonuçlar gözlem verilerine dayandığı için kesin bir neden-sonuç ilişkisi belirlemek için daha fazla veri analizi yapmak gerekebilir.

|  |
| --- |
| ggplot(cotton, aes(x = block, y = yield, fill = block)) +  geom\_bar(stat = "identity") +  scale\_y\_continuous(limits = c(0, 300), expand = c(0, 0), breaks = seq(0, 300, by = 53)) +  labs(title = "Block'ların Pamuk Verimi Karşılaştırması",  x = "Block",  y = "Pamuk Verimi") |



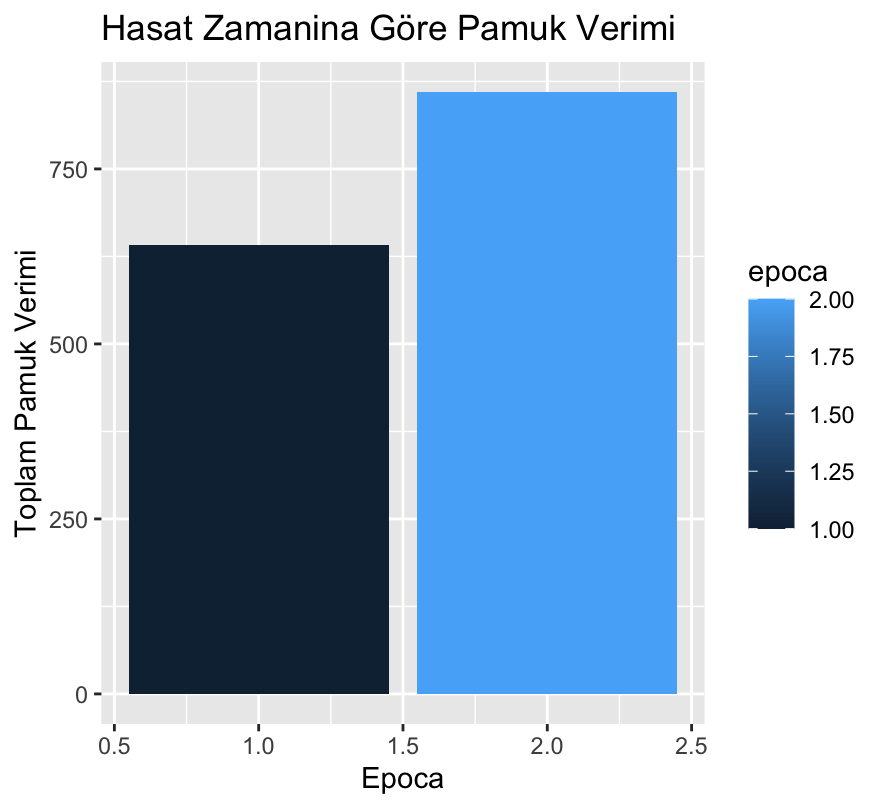
Hatirlayacak olursak Bloklar, deneyin yapıldığı alanda homojen olmayan koşulların neden olduğu değişkenlikleri kontrol etmek için kullanılan bir yöntemdir. Bu grafiğe göre, blok bazında en yüksek verim 265 ile II bloğunda elde edilmiştir. Devaminda 264 ile VI takip etmektedir

library(ggplot2)

ggplot(cotton, aes(x = epoca, y = yield, fill = epoca)) +

geom\_col() +

labs(x = "Epoca", y = "Toplam Pamuk Verimi", title = "Hasat Zamanina Göre Pamuk Verimi")



Grafikten de görüleceği üzere ikinci hasat zamaninda daha fazla pamuk elde edilmiştir.

lineage\_yields <- cotton %>%

group\_by(lineage) %>%

summarize(total\_yield = sum(yield))

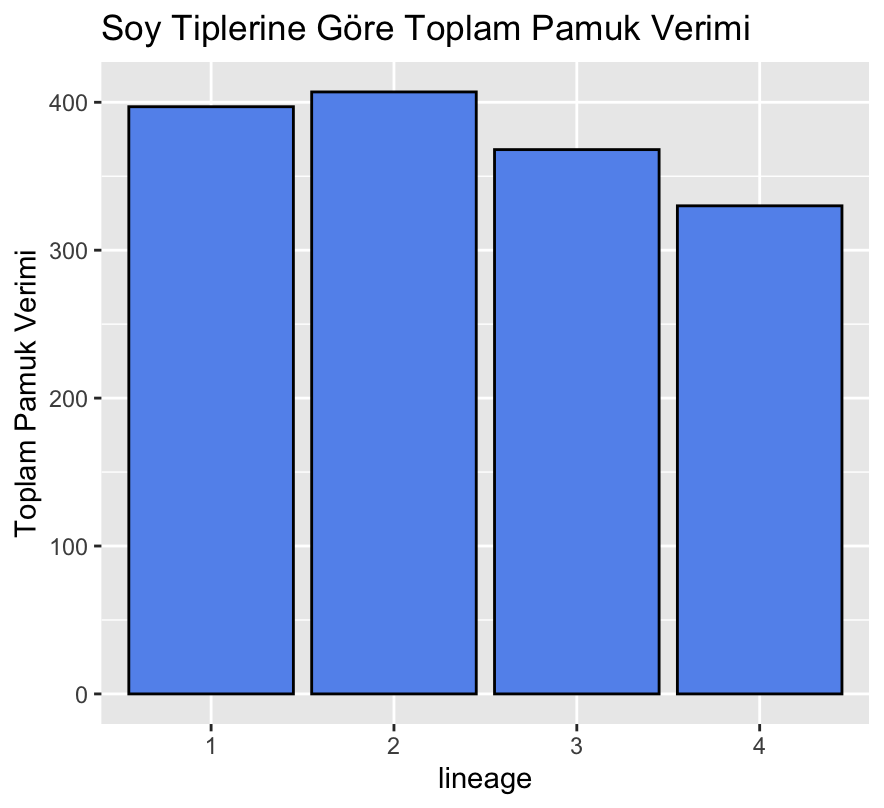
ggplot(lineage\_yields, aes(x = factor(lineage), y = total\_yield)) +

geom\_bar(stat = "identity", fill = "cornflowerblue", color = "black") +

labs(title = "Soy Tiplerine Göre Toplam Pamuk Verimi",

x = "lineage",

y = "Toplam Pamuk Verimi")



Buna göre iki numaralı soyun diğer üç soy türünden daha fazla pamuk ürettiği anlamına gelir. Bu veri, pamuk üreticilerinin bu iki soy türünü daha fazla yetiştirerek daha yüksek bir verim elde etme olasılığını gösterir.

Veri setine bagimli bagimsiz değişkenler arasinda normallik testi uygulamak istersek

|  |
| --- |
| x = cotton$lineage  y = cotton$yield  mod = lm(y ~ x) # y bagimli x bagimsiz degisken  summary(mod)  res = mod$residuals  hist(mod$residuals)  plot(res ~ x, pch = 21, bg = 'cyan', col = 'black', cex = 1.25)  abline(h = 0, col = 'maroon', lwd = 2)  shapiro.test(res)  qqnorm(res)  qqline(res) |

Hipotezler şöyledir:

H0: Veriler normal dağılıma uygun şekilde dağılmıştır.

H1: Veriler normal dağılıma uygun şekilde dağılmamıştır.

Test sonucu, W istatistiğinin değeri ve p-value değeridir. Testin sonucunu değerlendirmek için p-value değeri alfa (örneğin, 0.05) değerinden küçük mü diye bakılır. Eğer p-value, alfa değerinden küçükse H0 hipotezi reddedilir ve veriler normal dağılıma uygun değildir. Aksi takdirde, H0 hipotezi kabul edilir ve veriler normal dağılıma uygun olarak kabul edilir.

Bu örnekte, p-value değeri 0.01281'dir, bu da 0.05'ten küçüktür. Bu nedenle, H0 hipotezi reddedilir ve veriler normal dağılıma uygun değildir. Asagidaki Grafiktende anlasilacagi uzere. Burdaki kodlari calistirdigimizda asagidaki sonuclari elde ederiz.

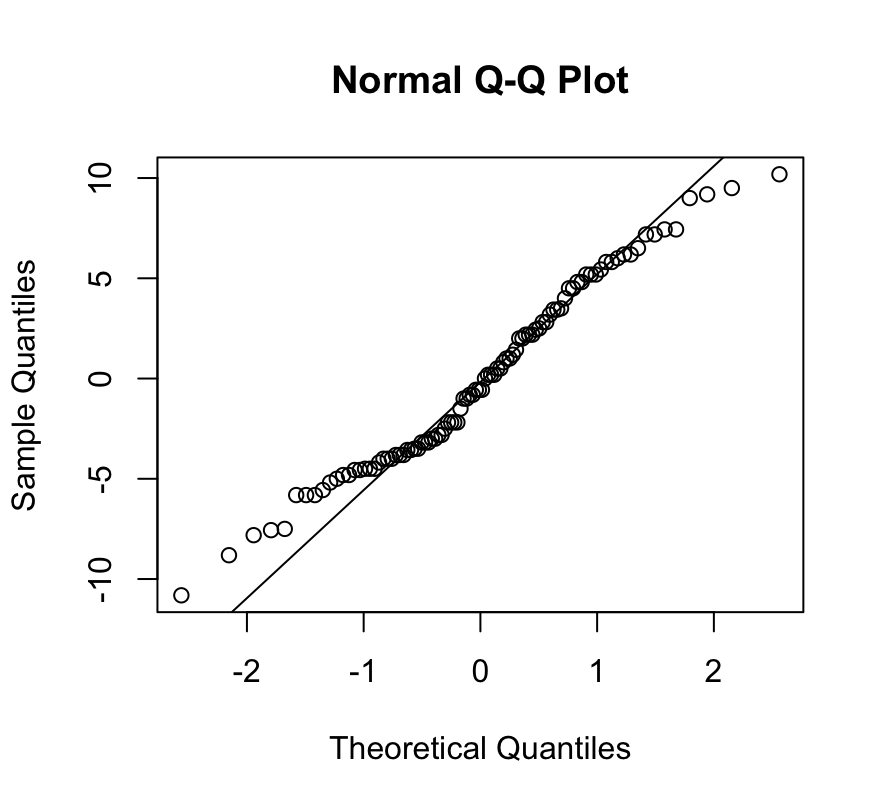
|  |
| --- |
| Shapiro-Wilk normality test  W = 0.96565, p-value = 0.01281 |

|  |
| --- |
| x = cotton$block  y = cotton$yield  mod = lm(y ~ x) # y bagimli x bagimsiz degisken  summary(mod)  res = mod$residuals  hist(mod$residuals)  plot(res ~ x, pch = 21, bg = 'cyan', col = 'black', cex = 1.25)  abline(h = 0, col = 'maroon', lwd = 2)  shapiro.test(res)  qqnorm(res)  qqline(res) |

Yukaridaki kodlari calistirdigimizda asagidaki sonuclari elde ederiz

|  |
| --- |
| Shapiro-Wilk normality test  data: res  W = 0.97571, p-value = 0.07151 |

Bu test sonucuna göre, hipotezimizi (veriler normal dağılır) reddedemiyoruz çünkü p-value değeri 0.05'ten büyük. Bu nedenle, verilerimizin normal olarak dağıldığı sonucuna varabiliriz. Ancak, 0.05'ten biraz daha büyük olan p-value değeri, normal dağılımın sağlamlığını tam olarak garanti etmez.



Yukaridaki analizlerin tamamina baktigimiz zaman cotton veri seti hakkında sunlari söyleyebiliriz.

“Pisco'da, ikinci hasat döneminde, II numaralı blokta ikinci soy tipini kullanmanın daha fazla pamuk elde etmeye olanak sağlayabileceği sonucuna varılabilir mi?”

Veri analizlerine dayanarak yapılan bir öngörü olduğu için "doğru" ya da "yanlış" şeklinde net bir cevap vermek doğru olmaz. Ancak yapılan analizler sonucunda pisco bölgesindeki 2 numaralı blokta ikinci hasat döneminde 2 numaralı soy tipi kullanıldığında daha yüksek bir pamuk verimi elde edildiği ortaya çıkmıştır. Bu nedenle, bu öngörü, veri analizlerinin sonuçlarına dayanarak mantıklı bir sonuç olarak yorumlanabilir. Ancak, bu öngörünün gerçekleştirilmesi için çeşitli faktörler (örneğin hava şartları, toprak koşulları, yetiştirme koşulları vb.) de göz önünde bulundurulmalıdır.

MELON dATASET

Melon veri seti, bir Latin kare deneyinde elde edilen verileri içermektedir. Veri seti, basit bir sıra ekimi yöntemi ile dört farklı kavun çeşidinin exudation sulama sistemi kullanılarak değerlendirilmesi sonucu elde edilen verileri içerir.

Bu veri seti, damlama sulama sisteminin dört çeşit kavun çeşidi üzerindeki etkisini incelemek için yapılan bir Latin kare deneyinin sonuçlarını içerir. Deney, BASİT SIRA ekim modelli kullanılarak gerçekleştirildi. Üç hibrid kavun çeşidi ve bir standart çeşit incelenmiştir. Veri setinde, sıra ve sütun bilgileri, kavun çeşitleri ve elde edilen verim değerleri yer almaktadır. Veri setinin kaynağı, Alberto Angeles L. tarafından yapılan "Evaluacion del sistema de riego por exudacion utilizando cuatro variedades de melon, bajo modalidad de siembra, SIMPLE HILERA" adlı tezdir ve Peru'nun Lima şehrindeki Universidad Agraria la Molina'da gerçekleştirilmiştir.

Veri seti, 16 gözlem ve 4 değişkenden oluşmaktadır.

"row" ve "col" sırasıyla satır ve sütun numaralarını temsil etmektedir ve deneyin yapıldığı Latin kare matrisinin satır ve sütunlarını ifade etmektedir. "row" (satır) ve "col" (sütun)

"variety", dört farklı kavun çeşidini temsil etmektedir ve "V1", "V2", "V3" ve "V4" olarak kodlanmıştır.

"yield", her bir gözlemde elde edilen kavun verimini ifade etmektedir ve sayısal bir değerdir.

library(ggplot2)

p <- ggplot(melon, aes(x = variety, y = yield)) +

theme\_minimal() +

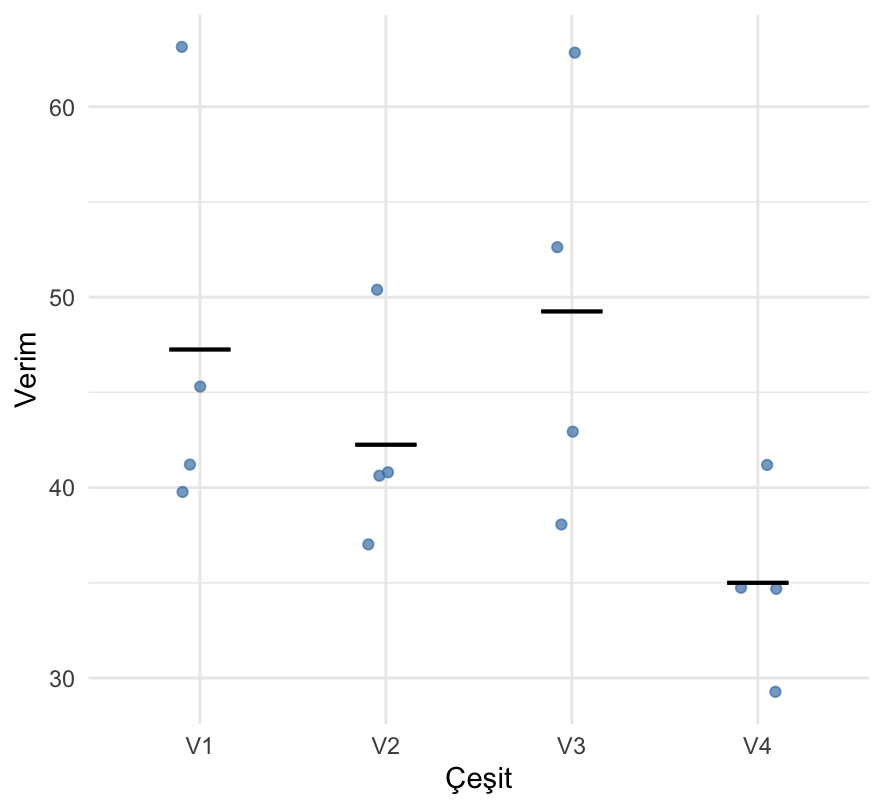
geom\_jitter(alpha = 0.7, width = 0.1, na.rm = T, color = "steelblue") +

stat\_summary(fun = "mean", fun.min = "mean", fun.max = "mean", size = 0.3, width = 0.33, geom = "crossbar") +

ylab("Verim +

xlab("Cesit")

plot(p)



Grafikteki siyah çizgiler, "stat\_summary" fonksiyonunun kullanımı sonucu oluşan hata çubuklarını (error bars) temsil eder. Hata çubukları, her grup için ortalama değeri gösterirken aynı zamanda standart sapmayı veya standart hata gibi ölçüleri de göstermek için kullanılabilir. Bu grafikte, her çubuk, ilgili varyeteye ait ortalama hasat verimini göstermektedir ve çubuk uzunlukları, bu ortalama hasat verimine karşılık gelen standart sapmayı gösterir.

Grafikte en fazla verim V3 çesidinde olduğunu goruyoruz.

Herbir çesitin toplam verimini elde etmek için asagidaki fonksiyonu da kullanabiliriz.

#

sum(melon$yield[melon$variety == "V1"]) #= 189

sum(melon$yield[melon$variety == "V2"]) #= 169

sum(melon$yield[melon$variety == "V3"]) #= 197

sum(melon$yield[melon$variety == "V4"]) #= 140

Tanimlayici istatistikler sayesinde veriyi daha iyi taniyalim.

min(melon$yield)

## [1] 29

“Veri Setinde en düşük karvun verimi 29’dur ve V4 çesidine aittir.”

max(melon$yield)

## [1] 63

“Veri Setinde en yüksek kavun verimi 63’dür. V1 ve V3 çesidine aittir.”

range(melon$yield)

## [1] 29 63

“Veri setinde en yüksek ve en düşük değer”

mean(melon$yield)

## [1] 43.4375

“melon veri setindeki kavun veriminin ortalama değeri 43.4375’dir.”

median(melon$yield)

## [1] 41

“melon veri setinin kavun verimlerine bakıldığında, ortanca verim değerinin 41 olduğu görülmüştür.

quantile(melon$yield, 0.25)

## 25%

37.75

“melon veri setinde, verimlerin %25'inin 37.75'den küçük veya ona eşit olduğu görülmektedir.”

tablo <- table(melon$yield)

mod <- as.numeric(names(which.max(tablo)))

print(mod)

## 41

“melon veri setinde en sık tekrar eden kavun verimi 41'dir.”

IQR(melon$yield)

## 8.5

“melon veri setindeki kavun verimlerinin çeyreklikler arasındaki fark (IQR) 8.5'dir.”

sd(melon$yield)

## [1] 9.521686

“Bu değer, melon veri setindeki verim değerlerinin ne kadar değişken olduğunu gösterir. Standart sapma, bir veri kümesinin ortalamadan ne kadar sapma gösterdiğini ölçer. Dolayısıyla, melon veri setindeki verim değerleri ortalamadan yaklaşık 9.52 birim kadar sapmaktadır. Bu, verim değerlerinin oldukça değişken olduğunu gösterir.”

var(melon$yield)

## [1] 90.6625

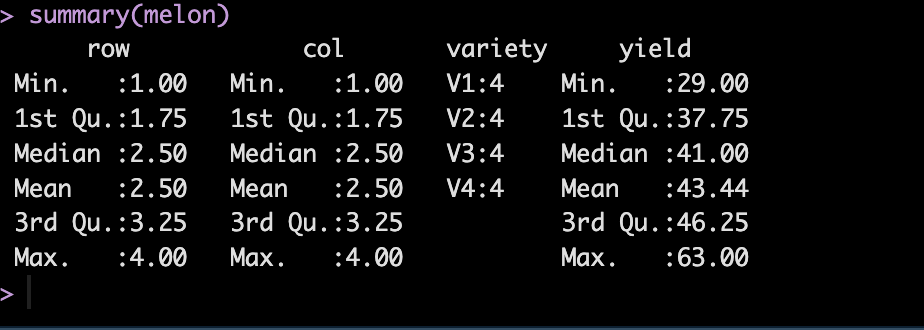
“Bu değer, melon veri setindeki verim değerlerinin varyansını temsil ediyor. Varyans, bir veri setindeki verilerin ne kadar farklı olduğunu ölçen bir istatistiksel terimdir. Melon veri setindeki verim değerlerinin varyansının 90.6625 olduğunu gösterir. Daha düşük bir varyans, verilerin birbirine daha yakın olduğunu ve daha homojen bir veri dağılımına sahip olduğunu gösterirken, daha yüksek bir varyans, verilerin daha dağınık olduğunu ve daha heterojen bir veri dağılımına sahip olduğunu gösterir.”

sd(melon$yield) / mean(melon$yield)

## [1] 0.2192043

Herbir değeri bu sekilde tek tek yazip elde etmek yerine bazi komutlar ile bu değerleri doğrudan elde edebiliriz.

summary(melon)



#melon veri setindeki row ve col matris haline getirme

matris\_melon <- reshape(agg\_melon, idvar = "row", timevar = "col", direction = "wide")

matris\_melon

#row x.1 x.2 x.3 x.4

1 45(V1) 29(V4) 37(V2) 38(V3) #150

2 50(V2) 53(V3) 41(V4) 40(V1) #186

3 43(V3) 41(V2) 41(V1) 35(V4) #163

4 35(V4) 63(V1) 63(V3) 41(V2) #206

#173 #186 #182 #154

Verileri matris haline getirdiğimiz zaman 4. Satir 2. Ve 3. Sütunda daha fazla kavun verimi elde edildiğini görüyoruz.

GROWTH dATASET

"growth" veri seti, Peru'daki Pucallpa'da yapılan bir deneyde, pijuayo ağaçlarının büyümesiyle ilgili verileri içerir. Veri seti 30 gözlem içermekte olup, her bir gözlem, farklı bir lokalitede yapılan ölçümlere karşılık gelmektedir. Veri setinde yer alan değişkenler şunlardır:

place: L1 ve L2 olmak üzere iki farklı lokalite

slime: pijuayo ağacının büyümesinde rol oynayan yoğunluk (slime) değeri. "Slime" kelimesi, ağaçların gövdesindeki reçine benzeri yapışkan bir maddenin birleştirici etkisini ifade eder ve ağacın büyümesini etkiler. Bu veri setindeki "slime" değerleri, ağaçların büyümesine katkıda bulunan faktörlerin etkisini ölçmek için ölçülmüş ve kaydedilmiştir.

height: pijuayo ağacının büyümesiyle ilgili yükseklik değeri

library(ggplot2)

ggplot(growth, aes(x = place, y = slime, fill = place)) +

geom\_bar(stat = "identity", alpha = 0.8) +

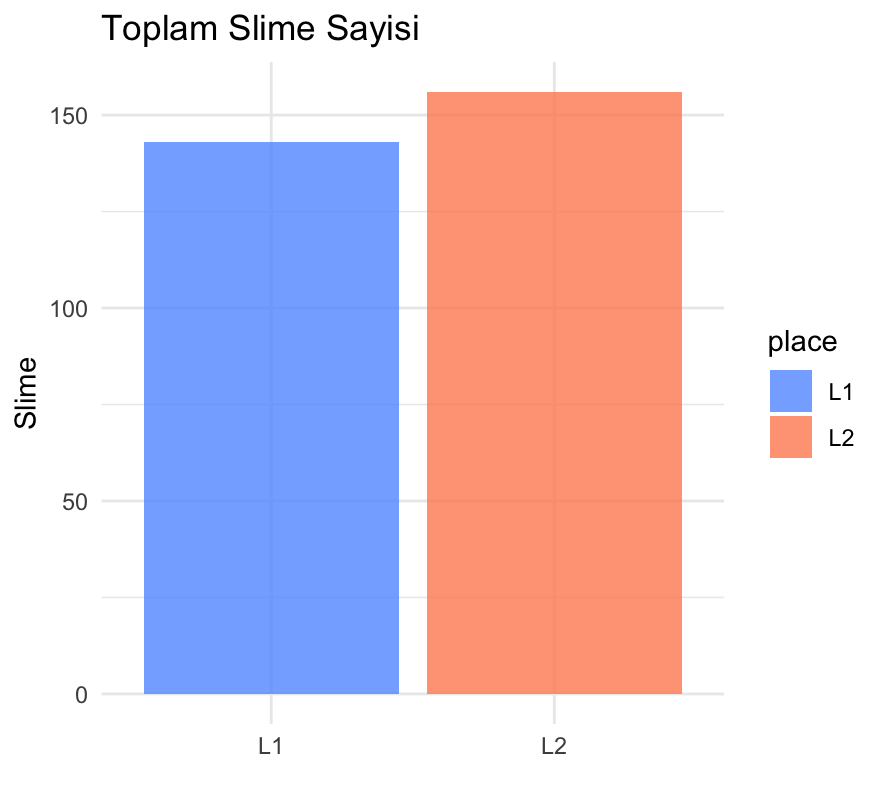
scale\_fill\_manual(values = c("#619CFF", "#FF8C61")) +

xlab("") +

ylab("Slime") +

ggtitle("Toplam Slime Sayisi") +

theme\_minimal()



Grafik L1 ve L2 deki toplam yoğunluğu(slime) gösteriyor.

library(ggplot2)

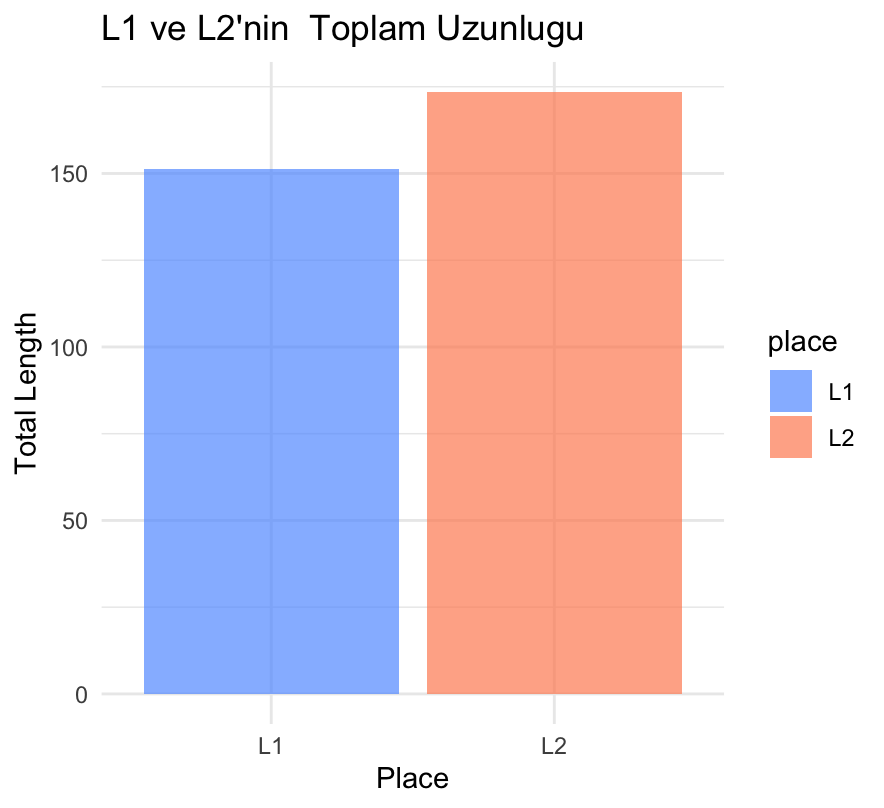
ggplot(growth, aes(x = place, y = height, fill = place)) +

geom\_histogram(stat = "identity", alpha = 0.7, bins = 10) +

scale\_fill\_manual(values = c("#619CFF", "#FF8C61")) +

labs(title = "L1 ve L2'nin Toplam Uzunlugu ", x = "Place", y = "Total Le ngth") +

theme\_minimal()



Grafik L1 ve L2 nin toplam yükseklik değerini gösteriyor.

Tanimlayici istatistikler sayesinde veriyi daha iyi taniyalim.

min(growth$height)

## [1] 6.5

“Veri Setinde en düşük yükseklik değeri 6.5’dir.”

max(growth$height)

## [1] 12.5

“Veri Setinde en yüksek yükseklik değeri 12.5’dur.”

range(growth$height)

## [1] 6.5 12.5

“Veri setinde en yüksek ve en düşük değer”

mean(growth$height)

## [1] 10.82333

“veri setindeki yüksekligin ortalama değeri 10.82333’dür.”

median(growth$height)

## [1] 11.15

“yükseklik değerinin ortancasi 11.15 olduğu görülmüştür.”

quantile(growth$height, 0.25)

## 25%

10.125

“veri setinde, verimlerin %25'inin 37.75'den küçük veya ona eşit olduğu görülmektedir.”

tablo <- table(growth$height)

mod <- as.numeric(names(which.max(tablo)))

print(mod)

## 11.7

“veri setinde en sık tekrar eden yükseklik 11.7’dir.”

IQR(growth$height)

## 1.65

“veri setindeki çeyreklikler arasındaki fark (IQR) 1.65'dir.”

sd(growth$height)

## [1] 1.401891

“Bu değer, "height" değişkenindeki verilerin ortalama değerden ne kadar uzaklaştığını ölçer. Bu durumda, standart sapmanın düşük olması, "height" değişkenindeki verilerin ortalama değere yakın olduğunu ve dağılımın düşük olduğunu gösterir. Yani, "height" değişkenindeki bitki boyu verileri arasındaki farkın genellikle küçük olduğu ve nispeten homojen bir şekilde dağıldığı söylenebilir..”

var(growth$height)

## [1] 1.965299

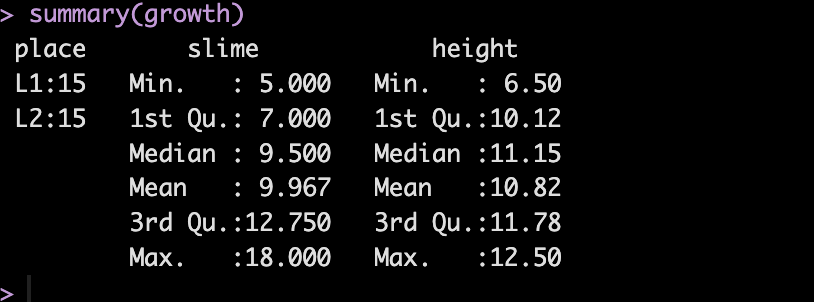
“Bu, height sütunundaki verilerin ortalamadan ne kadar uzaklaştığını ölçen birimlerin karesi cinsinden bir değerdir. Bu sonuca göre, height sütunundaki verilerin dağılımı ortalamadan biraz daha geniştir.”

sd(growth$height) / mean(growth$height)

## [1] 0.1295249

Bu veri kümesinde, yükseklik değerleri ortalamadan yüzde 12.95 daha az değişkenlik gösterir. Bu, veri kümesinin ne kadar homojen olduğunu gösteren bir ölçüttür. Düşük bir değişkenlik katsayısı daha homojen bir veri kümesine işaret ederken, yüksek bir değişkenlik katsayısı daha heterojen bir veri kümesine işaret eder.

summary(growth)



GRASS dATASET

Veri seti, yonca çimeni verimliliği üzerindeki etkisini incelemek için yapılan bir deneyde elde edilen verileri içermektedir.

Evaluation Her gözlemde, hakemlerin yonca çiminin verimliliğine yönelik değerlendirmelerinin ortalaması olan "evaluation" değişkeni de yer almaktadır.

Judge muameleler için on iki farklı hakem tarafından yapılan değerlendirmeler (judge1 – judge12) bulunmaktadır.

trt veri setinde kullanılan üç farklı yonca çimi muamelesini ifade eder.

library(ggplot2)

p <- ggplot(grass, aes(x = trt, y = evaluation)) +

theme\_minimal() +

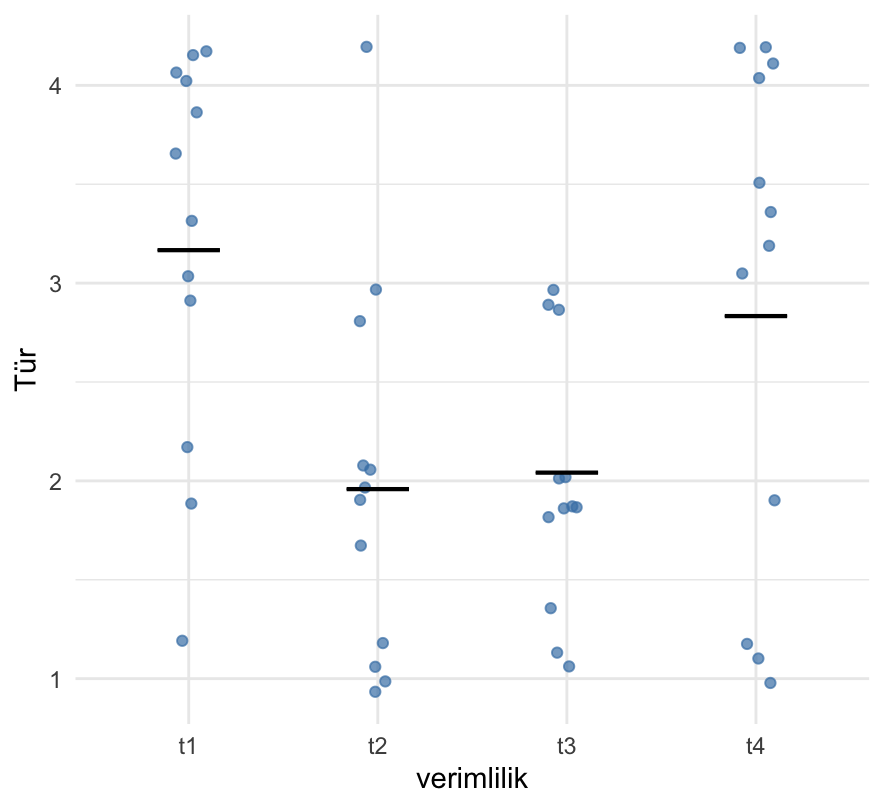
geom\_jitter(alpha = 0.7, width = 0.1, na.rm = T, color = "steelblue") +

stat\_summary(fun = "mean", fun.min = "mean", fun.max = "mean", size = 0.3, width = 0.33, geom = "crossbar") +

ylab("Tür") +

xlab("verimlilik")

plot(p)



library(ggplot2)

# Toplam değerleri hesapla

t1\_sum <- sum(grass$evaluation[grass$trt == "t1"])

t2\_sum <- sum(grass$evaluation[grass$trt == "t2"])

t3\_sum <- sum(grass$evaluation[grass$trt == "t3"])

t4\_sum <- sum(grass$evaluation[grass$trt == "t4"])

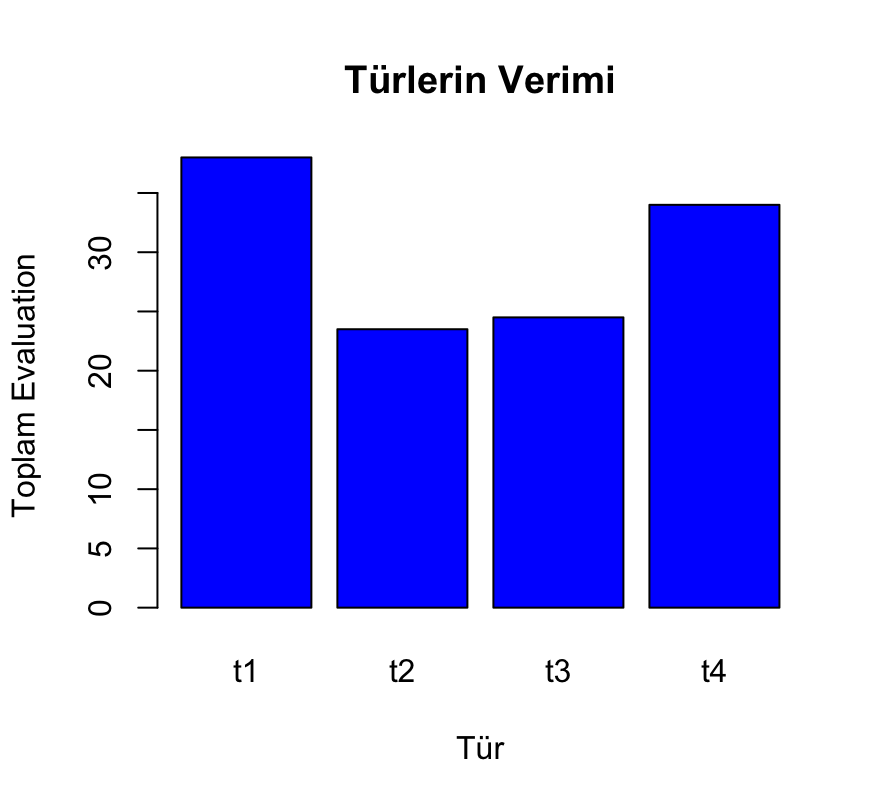
# Sütun grafiği oluştur

barplot(c(t1\_sum, t2\_sum, t3\_sum, t4\_sum),

names.arg = c("t1", "t2", "t3", "t4"),

xlab = "Tür", ylab = "Toplam Evaluation",

col = "blue", main = "Türlerin Verimi")



Tanimlayici istatistikler sayesinde veriyi daha iyi taniyalim.

min(grass$evaluation)

## [1] 1

“Veri Setinde en düşük verimlilik 1’dir.”

max(grass$evaluation)

## [1] 4

“Veri Setinde en yüksek verimlilik 4’dur.”

range(grass$evaluation)

## [1] 1 4

“Veri setinde en yüksek ve en düşük değer”

mean(grass$evaluation)

## [1] 2.5

“veri setindeki verimliliğin ortalama değer i2.5’dür.”

median(grass$evaluation)

## [1] 2

“ verimliliğin ortandca değeri 2 olduğu görülmüştür.”

quantile(grass$evaluation, 0.25)

## 25%

1.875

“veri setinde, verimlerin %25'inin 1.875'den küçük veya ona eşit olduğu görülmektedir.”

tablo <- table(grass$evaluation)

mod <- as.numeric(names(which.max(tablo)))

print(mod)

## 2

“veri setinde en sık tekrar eden sayi 2’dir.”

IQR(grass$evaluation)

## 1.625

“veri setindeki çeyreklikler arasındaki fark (IQR) 1.625'dir.”

sd(grass$evaluation)

## [1] 1.096416

“evaluation değerlerinin ortalama değerden ortalama 1.096 birim kadar farklı olduğunu gösterir.”

var(grass$evaluation)

## [1] 1.202128

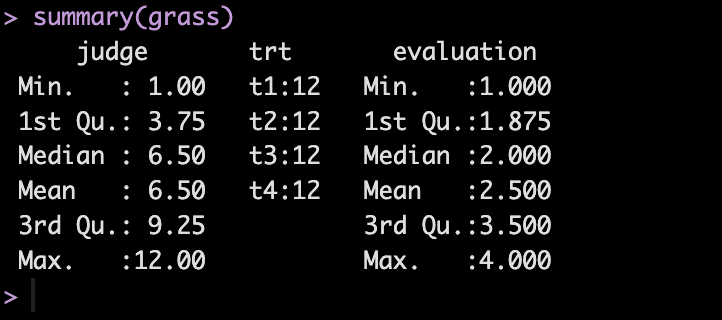
“Yüksek bir varyans veya standart sapma, verilerin geniş bir aralığa yayıldığını ve ortalamaya göre ne kadar farklılık gösterdiğini gösterir. Bu durumda, "evaluation" değişkeninin standart sapması ortalamaya kıyasla oldukça yüksek olduğundan, değerlerin birbirinden oldukça farklı olduğu ve dağılımın geniş olduğu söylenebilir.”

sd(grass$evaluation) / mean(grass$evaluation)

## [1] 0.4385663

Bu değer, "evaluation" değişkeninin ortalama değerine oranla standart sapmasının ne kadar olduğunu gösterir. Bu durumda, standart sapmanın ortalama değere oranı yaklaşık olarak %44'tür. Bu, "evaluation" değişkenindeki değerlerin ortalama değer etrafında ne kadar değiştiğini belirten bir ölçüttür. %44'ün altındaki bir oran, değişkenin oldukça homojen olduğunu, %44'ün üzerindeki bir oran ise değişkenin oldukça heterojen olduğunu gösterir. Ancak bu yorum, değişkenin ne olduğuna, ölçüm birimlerine ve hangi amaçla kullanıldığına bağlı olarak değişebilir.

summary(grass)



GREENHOUSE dATASET

Patates minituber üretimi için farklı yöntemlerin verimliliğini karşılaştırmak için kullanılir. Bu veri seti, bir sera ortamında patates minituber üretimine ilişkin verileri içerir. Veri seti, üç farklı patates çeşidinde, hidroponik, aeroponik, tencere ve bitki yatağı yöntemleriyle yapılan deneylerde elde edilen verileri içerir.

variety: Patates çeşitleri (Potato varieties)

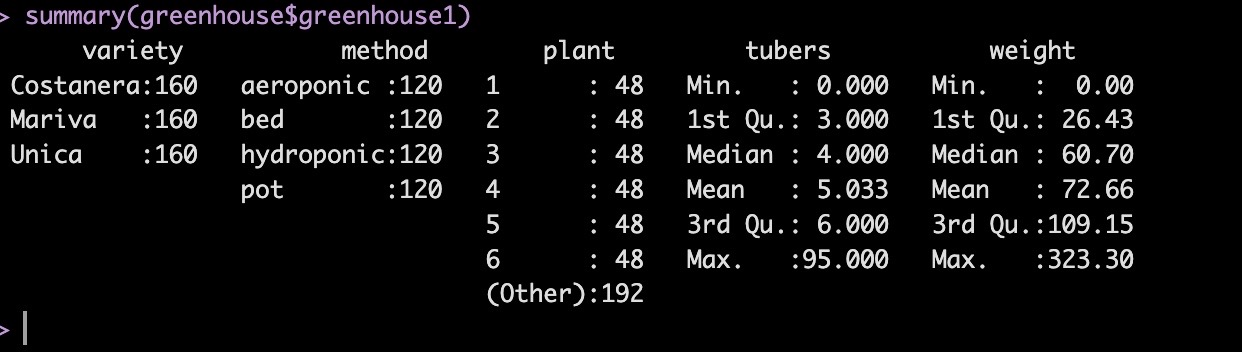
method: Yetiştirme yöntemleri (Hydroponics, Aeroponic, Pots and Plant beds)

plant: Bitki sayısı (Number of plants)

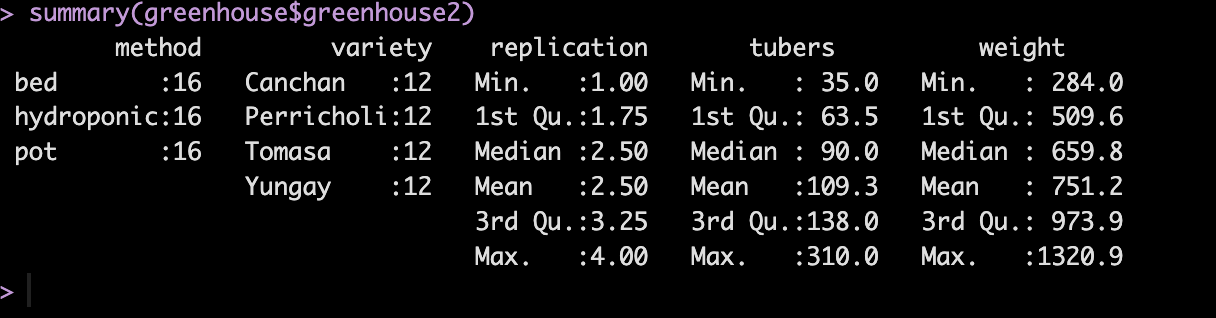
tubers: Bitki başına elde edilen minituber sayısı (Number of minitubers per plant)

weight: Toplam minituber ağırlığı (Total weight of minitubers)

summary(greenhouse$greenhouse1)



summary(greenhouse$greenhouse2)



summary(greenhouse$greenhouse3)

