İST 155 İSTATİSTİĞE GİRİŞ 1 DÖNEM SONU ÖDEVİ

Ad Soyad: Hasan Efe KOCASU
Okul Numarası: 2240329066
Şube: 02

Veri Seti'ni Tanıyalım

Veri seti, 2021-2022 yılındaki Amerikan Ulusal Basketbol Ligi olan NBA'deki çaylak (rookie) oyuncuların belirli başlıklardaki verilerini kapsıyor. Bu ödevde ise oyuncuların maç başına atmış oldukları ortalama puan olan "points" değişkeni üzerinde çalışılacaktır.

Veri Setindeki değişkenin R'da tanımlanması:

Veri seti .csv (virgülle ayrılmış text dosyası) formatında olduğundan "File > Import Dataset > From text (base)" sekmesinden dosyayı RStudio'ya aktarıyoruz. Aşağıdaki kod satırı ile veri setini tanımlıyoruz:

Kod satırı:

```
veriseti <- c(nba_rookie_data_2021.2022)
```

Çıktı:

```
> veriseti <- c(nba_rookie_data_2021.2022)
```

Tanımladıktan sonra "points" değişkeni üzerinde çalışacağız. Daha kolay çalışmak için "points" sütununu başka bir vektöre atıyoruz:

Kod satırı:

```
veri <- c(veriseti$points)
```

Çıktı:

```
> veri <- c(veriseti$points)
```

"veri" vektörümüz :

```
> veri
  [1] 17.5 17.0 15.5 15.2 14.9 13.1 12.5 11.9 11.2 10.4  9.9  9.6  9.6
 [14]  9.6  9.5  9.1  8.7  8.5  8.3  8.1  8.1  7.8  7.7  7.6  7.5  7.1
 [27]  6.7  6.5  6.2  6.0  5.9  5.9  5.8  5.6  5.5  5.4  5.3  5.1  4.9
 [40]  4.8  4.8  4.4  3.8  3.5  3.2  3.2  2.2  2.2  2.2  2.1  2.0  1.1
```

Değişkenimizi de tanımladığımıza göre konum ölçülerinin hesaplanması ve yorumlanması kısmına geçebiliriz.

Ortalama (Mean)

Kod satırı:

```
mean(veri)
```

Çıktı:

```
> mean(veri)
[1] 7.503846
```

Yorum:

Veri'nin ortalaması 7.50'ymiş bu NBA 2021-2022 sezonu çaylak basketbol oyuncularının maç başına ortalama 7.50 puan sayı kaydettiklerini gösteriyor.

Tepe Değeri (Mod)

Kod satırı:

```
mod_veri <- table(veri)
names(mod_veri)[which(mod_veri==max(mod_veri))]
```

İlk satırdaki kod ile mod hesaplamak için "veri" tablosu oluşturduk ve ikinci satır ile de en çok tekrar eden değeri yani tepe değeri (mod) bulmuş olduk.

Çıktı:

```
> mod_veri <- table(veri)
> names(mod_veri)[which(mod_veri==max(mod_veri))]
[1] "2.2" "9.6"
```

Yorum:

Veri'nin tepe değeri (modu) 2 taneymiş bunlar "2.2" ve "9.6" bu da çaylak basketbolcular arasında maç başına 2.2 ve 9.6 puan sayı kaydeden oyuncular diğer değerlerde kaydedenlerden daha fazla olduğu ve birden fazla kez tekrarlandığını gösteriyor.

Medyan (Ortanca)

Kod satırı:

```
median(veri)
```

Çıktı:

```
> median(veri)
[1] 6.9
```

Yorum:

Medyan 6.9'muş bu da NBA'deki çaylak basketbol oyuncularının maç başına ortalama atılan puan sayıları arasındaki ortanca değerin 6.9 olduğunu verir. Veriler arasındaki orta değer 6.9'muş. 52 veri olduğu ve 52 çift bir sayı olduğu içinse 26. ve 27. değerlerin toplamının 2'ye bölünmesiyle hesaplanır.

Genel Yorum:(Aslında Dağılım,Simetri-Asimetri Ölçülerinde var olan bir yorum)

Ortalama, ortanca ve mod (tepe değeri) arasındaki ilişkiye bakarak verinin dağılımının nereye çarpık olduğunu yorumlayabiliriz.

Ortalama: 7.50 Ortanca: 6.90 Mod: 5.9 (2 tane olduğu için aritmetik ortalamaları)

Ortalama > Ortanca > Mod 7.5 > 6.9 > 5.9

Verilerimizin dağılımı sağa çarpıktır ('+' artı yöne çarpık) diyebiliriz.

Yüzdelikler (Percentiles)

Hesaplanacak yüzdelikler; P_{25}, P_{75}, P_{90}, P_{10} :

Kod satırı:

```
quantile(veri, probs = c(0.25, 0.75, 0.90, 0.10))
```

Çıktı:

```
> quantile(veri, probs = c(0.25, 0.75, 0.90, 0.10))
   25%    75%    90%    10%
 4.875  9.600 13.040  2.300
```

Yorum:

P_{25} için: Verilerin %25'i 4.875 değerinden daha küçük değer alıyormuş. Bu da çaylak basketbol oyuncuların %25'inin 4.875'ten daha az puan kaydettiği anlamına geliyor. Aynı zamanda %75'inin de 4.875'ten daha fazla puan kaydettiği anlamına da geliyor.

P_{75} için: Verilerin %75'i 9.600 değerinden daha küçük değer alıyormuş. Bu da çaylak basketbol oyuncuların %75'inin 9.6'dan daha az puan kaydettiği anlamına geliyor. Aynı zamanda %25'inin de 9.6'dan daha fazla puan kaydettiği anlamına da geliyor.

P_{90} için: Verilerin %90'ı 13.040 değerinden daha küçük değer alıyormuş. Bu da çaylak basketbol oyuncuların %90'ının 13.04'ten daha az puan kaydettiği anlamına geliyor. Aynı zamanda %10'unun da 13.04'ten daha fazla puan kaydettiği anlamına da geliyor.

P_{10} için: Verilerin %10'u 2.3 değerinden daha küçük değer alıyormuş. Bu da çaylak basketbol oyuncuların %10'unun 2.3'ten daha az puan kaydettiği anlamına geliyor. Aynı zamanda %90'ının da 2.3'ten daha fazla puan kaydettiği anlamına da geliyor.

Konum ölçülerini yukarıda inceledik, şimdi de Değişim Ölçülerini inceleyelim.

Varyans

Kod satırı:

```
var(veri)
```

Çıktı:

```
> var(veri)
[1] 15.96548
```

Yorum:

Varyans, değerlerin ortalamadan ne kadar saptığını, ortalamayla aralarındaki uzaklığı ölçer. Veri grubunun varyansı 15.96548'miş. Varyans değeri küçük olduğundan ortalamadan çok sapmadığı ve daha düzenli bir veri grubu olduğunu söyleyebiliriz.Çok sağlıklı olmasa da aynı zamanda aykırı değerinde bulunmadığı da buradan çıkartılabilir.

Standart Sapma

Kod satırı:

```
sd(veri)
```

Çıktı:

```
> sd(veri)
[1] 3.995682
```

Yorum:

Standart sapma, verilerin ortalamadan ne kadar sapma gösterdiğini de bize verir. Varyansın karekökünün alınmasıyla hesaplanır. Standart sapma ne kadar küçükse gözlemler o kadar ortalamaya yakındır. 3.99 değeri de küçük bir değerdir. Bu nedenle verilerin ortalama etrafında toplandığını söyleyebiliriz.

Standart Hata

Kod satırı:

```
sd(veri)/sqrt(length(veri))
```

Çıktı:

```
> sd(veri)/sqrt(length(veri))
[1] 0.5541014
```

Yorum:

Standart hata, örneklem ortalamalarının standart sapmasıdır. Çalışmamızı tek örneklem üzerinde yaptığımız için kitleye erişimimiz olmadığından dolayı da **standart hata adına yorum yapamayız**.

Çarpıklık Katsayısı(Skewness)

Kod satırı:

```
library(moments)
skewness(veri)
```

Çıktı:

```
> skewness(veri)
[1] 0.7070536
```

Yorum:

Çarpıklık katsayısı(skewness); normal simetrik dağılımdan ne kadar saptığını belli etmek için, ölçebilmek için kullanılır. Ve verimizin çarpıklık katsayısı 0.707'miş bu da küçük bir değer olduğu için simetrik bir dağılıma oldukça yakın olduğu yorumunu verebilir bize. Pozitif olmasından dolayısıyla da sağa ('+' artı yöne) çarpık diyebiliriz.

Kod satırını yorumlayacak olursak ilk satır olmadan "skewness()" fonksiyonunu kullanamayız, default R paketlerinde dahil değildir çünkü. İlk başta "moments" paketini manuel olarak "packages" sekmesinden ekleyip ilk satırdaki kod satırı sayesinde fonksiyonu içeren kütüphaneyi çağırıp fonksiyonu kullanılabilir hale getiriyoruz.

Basıklık Katsayısı(Kurtosis)

Kod satırı:

```
kurtosis(veri)
```

Çıktı:

```
> kurtosis(veri)
[1] 3.011578
```

Yorum:

Basıklık katsayısı(kurtosis); **tek tepeli** olan dağılımların diklik-basıklık ölçüsünü verir. Standart normal dağılımın basıklık katsayısı 3'e eşittir ."points" değerlerimiz ise çift tepeli, 2 mod değerine sahip (2.2 ve 9.6) bu yüzden basıklık katsayısı hakkında yorum yapamayız ancak R'ın bize verdiği değeri yorumlayacak olursak 3 değerinden büyük olduğu için "lepokurtic"(standart normal dağılıma göre daha sivri) diyebilir ve 3 değerine oldukça yakın olduğu için de standart normal dağılıma çok yakındır diyebiliriz.

Kod satırını yorumlayacak olursak bir önceki hesaplamada dahil ettiğimiz "moments" kütüphanesi sayesinde "kurtosis()" fonksiyonunu kullanabiliyoruz. Bir önceki işlemde zaten kütüphaneyi çağırdığımız için de şuan tekrar çağırmamız gerekmiyor.

Değişim ölçülerini yukarıda inceledik şimdi ise Simetri-Asimetri Ölçülerini inceleyelim.

Bowley'in Asimetri Ölçüsü

$$BÇ = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

BÇ ~ 0 ise dağılım simetriktir

BÇ < 0 ise dağılım sola çarpıktır

BÇ > 0 ise dağılım sağa çarpıktır

Bowley'in asimetri ölçüsünü hesaplayabilmek için çeyrekliklere (quantile) ihtiyacımız olucak önce onları hesaplayalım.

Çeyreklikler(Quantiles)

Kod satırı:

```
quantile(veri)
```

Çıktı:

```
> quantile(veri)
    0%    25%    50%    75%   100%
 1.100  4.875  6.900  9.600 17.500
```

$Q_1 = P_{25} = 4.875$ $Q_2 = P_{50} = 6.900$ $Q_3 = P_{75} = 9.600$

Bowley asimetri ölçüsünü R'da hesaplayalım.

Kod satırı:

```
Q1 <- 4.875
Q2 <- 6.900
Q3 <- 9.600
bowley <- (Q3+Q1-2*Q2)/(Q3-Q1)
bowley
```

Çıktı:

```
> bowley
[1] 0.1428571
```

Yorum:

Verilerimiz için Bowley'in Asimetri Ölçüsü 0.1428571'miş 0'dan büyük olduğu için dağılımın sağa ('+' artı, pozitif yöne) çarpık olduğunu ve 0'a çok yakın olduğu içinde simetrik dağılıma çok ama çok yakın olduğunuda bize gösteriyor.

Kod satırını inceleyecek olursak ilk önce çeyrekliklerimiz (quantile) hesaplayıp sırayla vektörlerine tanımladık böylece Bowley'in Asimetri Ölçüsü formülünü tanımlayıp hesaplayabildik.

Pearson'ın Asimetri Ölçüsü

$$PÇ = \frac{\bar{X} - mod}{S}$$

$$PÇ = \frac{3(\bar{X} - medyan)}{S}$$

PÇ ~ 0 ise dağılım simetriktir

PÇ < 0 ise dağılım sola çarpıktır

PÇ > 0 is dağılım sağa çarpıktır

Pearson'ın asimetri ölçüsünü hesaplarken 2 formül, 2 eşitlik vardır ve bunlardan hangisi veri setimize verilerimize daha uygunsa onu kullanabiliriz ancak bizim verilerimiz çift tepeli yani 2 mod deeğeri bulunduğundan 1.eşitliği hesaplayamayız bu yüzden sadece 2.eşitliği hesaplayacağız. Zaten daha önceden hesapladığımız yukarıda var olan ve eşitlikte kullanacağımız değerleri/ölçüleri hatırlayalım.

Medyan = 6.9 Ortalama(Mean) = 7.50 Standart Sapma(S) = 3.99

Kod satırı:

```
std_sapma <- sd(veri)
std_sapma
ortalama <- mean(veri)
ortalama
medyan <- median(veri)
medyan
pearson2 <- 3*(ortalama - medyan)/(std_sapma)
pearson2
```

Çıktı:

```
> std_sapma <- sd(veri)
> std_sapma
[1] 3.995682
> ortalama <- mean(veri)
> ortalama
[1] 7.503846
> medyan <- median(veri)
> medyan
[1] 6.9
> pearson2 <- 3*(ortalama - medyan)/(std_sapma)
> pearson2
[1] 0.453374
>
```

Yorum:

Pearson'ın asimetri ölçüsünün 2.eşitliğinden hesapladık ve değerimiz 0.453374 çıktı bu da 0'dan büyük olduğu için dağılım sağa ('+' artı, pozitif yöne) çarpıktır diyebilir ve 0'a çok yakın olduğu için de oldukça simetrik dağılıma yakın bir dağılımdır da diyebiliriz.

Kod satırını yorumlayacak olursak eşitliği daha kolay hesaplayabilmek için denklem değerlerimizi ayrı ayrı vektörlere tanımlayıp bu şekilde Pearson'ın asimetri ölçüsünün 2.eşitliğini tanımlayıp hesapladık.

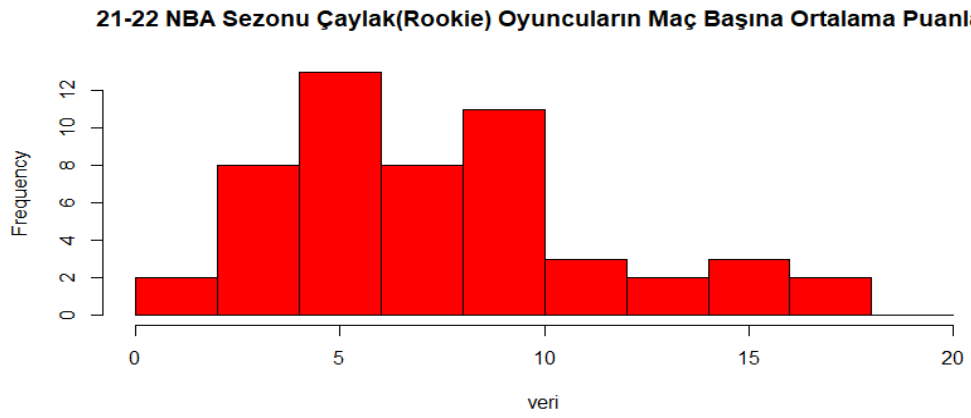
Grafikler ile verilerin incelenmesi

Histogram Grafiği:

Kod satırı:

```
hist(veri, main = "21-22 NBA Sezonu Çaylak(Rookie) Oyuncuların Maç
Başına Ortalama Puanlar",breaks=seq(0, 20, by=2),col = "red")
```

Çıktı:



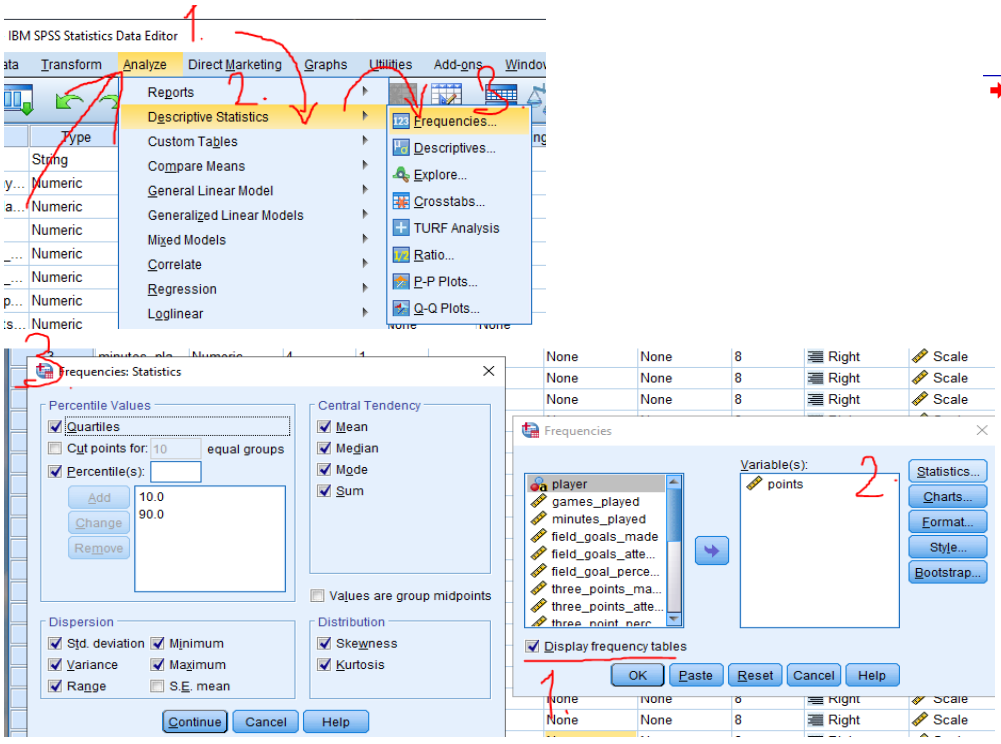
Yorum:

Histogram grafiğine bakıldığında verilerimizin dağılımını görüyoruz ve bu grafikten anlaşılabileceği üzere 2 mod(tepe değeri) sahibi olduğu görülüyor.

Kod satırını inceleyecek olursak “main” parametrei ile grafik başlığı,”breaks” parametresi ile x eksenine ait sınırları ve “col” parametresi ile de grafiğin rengini düzenleyebiliyoruz.

Birikimli Sıklık (Ogive) Grafiği:

Birikimli sıklık (Ogive) grafiği için sıklık tablosuna ihtiyacımız olucak bu yüzden sıklık tablosunu düzenlemeliyiz bunun için SPSS programını kullanacağız.



Frequencies

Statistics

points

N	Valid	52
	Missing	0
Mean		7.504
Median		6.900
Mode		2.2[a]
Std. Deviation		3.9957
Variance		15.965
Skewness		.728
Std. Error of Skewness		.330
Kurtosis		.138
Std. Error of Kurtosis		.650
Range		16.4
Minimum		1.1
Maximum		17.5
Sum		390.2
Percentiles	10	2.200
	25	4.825
	50	6.900
	75	9.600
	90	14.360

a. Multiple modes exist. The smallest value is shown

"Statistics" tablosundaki bilgilerle sıklık tablosunu ihtiyacımız olanlar doğrultusunda oluşturabiliriz.

$$Sınıf\ sayısı(k) = 1 + 3{,}3\log n \qquad 1 + 3{,}3log52 = 6{,}66 \cong 7$$

$$Sınıf\ aralığı(c) = \frac{Dağılım\ genişliği(Range)\ +\ son\ haneye\ 1\ ekle}{k}$$

$$Sınıf\ aralığı(c) = \frac{16{,}4 + 0{,}1}{7} = 2{,}35$$

Sınıf sayısı(k) 6.66'ymış en yakın tam değer 7 dir. Sınıf aralığı ise 2.35'miş. Bu bilgiler doğrultusunda sıklık grafiği oluşturup birikimli sıklığı bulabilicek ve R'da grafiğini çizebileceğiz.

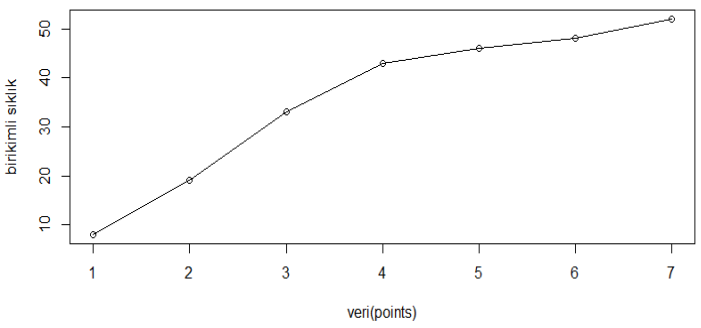
Sınıflar	f_i(Sıklık)	F_i(Birikimli Sıklık)
1,10 - 3,44	8	8
3,45 – 5,79	11	19
5,80 - 8,14	14	33
8,15 – 10,49	10	43
10,50 – 12,84	3	46
12,85 – 15,19	2	48
15,20 – 17,54	4	52

Sıklık tablomuzu oluşturup sıklık ve birikimli sıklık verilerimizi elde ettiğimize göre R programında Birikimli Sıklık Grafiği(Ogive) çizebiliriz.

Kod satırı:

```
f <- c(8,11,14,10,3,2,4)
plot(cumsum(f), type="o", xlab="veri(points)",
ylab="birikimli sıklık")
```

Çıktı:



Yorum:

Verilerin Birikimli Sıklık(Ogive) grafiği yanda görüldüğü üzere 7 sınıf olduğu ve 7 sınıfta 52 değer olduğu görülüyor. Kayıp değer yoktur da denebilir. Kod satırını inceleyecek olursak 1.satırda sıklık değerlerimizi girdik her biri

bir sınıfta kaç gözlem olduğunu veriyor. 2.satır ile de plot(çizgi) grafiğini çizdiriyoruz.

Dal-yaprak grafiği:

Kod satırı:

```
stem(veri, scale=2)
```

Çıktı:

```
> stem(veri, scale=2)

  The decimal point is at the |

   1 | 1
   2 | 01222
   3 | 2258
   4 | 4889
   5 | 13456899
   6 | 0257
   7 | 15678
   8 | 11357
   9 | 156669
  10 | 4
  11 | 29
  12 | 5
  13 | 1
  14 | 9
  15 | 25
  16 |
  17 | 05
```

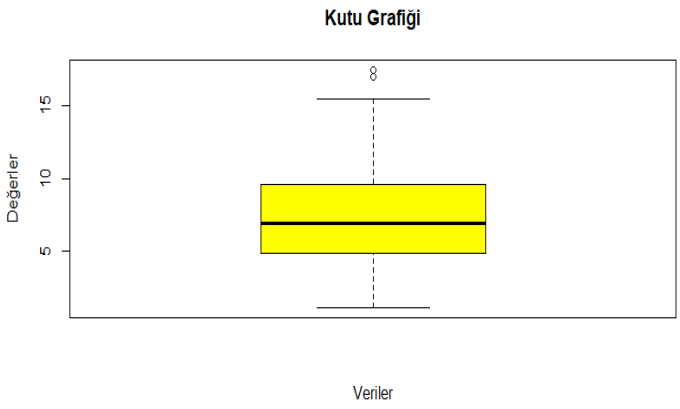
Dal yaprak grafiğinin sol tarafındaki sayılar gözlemlerin tam kısımlarının çizginin sağ tarafındaki her bir basamak ise ayrı birer gözlemi belirtiyor. Dal yaprak grafiği aynı zamanda bize en çok gözlem sayısının nerede yoğunlaştığını gösteriyor ve verilerin küçükten büyüğe düzenlenmiş halini de veriyor.

Kutu Grafiği(Boxplot) ve Aykırı Değer

Kod satırı:

```
boxplot(veri, main="Kutu Grafiği", col="yellow",
xlab="Veriler", ylab="Değerler")
```

Çıktı:



Yorum:

Kutu grafiğinin aşağı doğru yani sağa çarpık olduğu görülüyor, kutu grafiğinde alt ve üst siyah çizgiler "kuyruklar" aykırı değerlerden arındırılmış dağılımın alt ve üst sınırlarını temsil ediyor kutunun ortasındaki kalın çizgi ise medyan(ortanca) değeridir, Q_2 değeridir ve verimizde 6.9'dur grafikten de doğrulanabilir. Kutunun alt sınırı Q_1 değerini verir. Q_1 değeri verimizde 4.875'dir grafikten de sarı kutunun alt çizgisi bu değere denk geldiği görülüyor. Sarı kutunun üst çizgisi ise Q_3 değerini verir ve verimizde Q_3 değeri 9.6'dır grafikten de bu sonuç çıkarılabilir. Dağılımın dışında görülen 2 yuvarlak nokta ise gözlemlerimiz içindeki aykırı değerleri temsil ediyor. Aykırı değer ise bize verimizi yorumlarken hangi ölçüleri kullanmamız gerektiğini belirlememiz ve hepsine güvenemeyeceğimiz gibi bazı sonuçlara neden oluyor. Mesela aykırı değerler ortalamayı çok yukarıya çekebilir ya da normalden daha aşağı çekebilirler bu da verimizi analiz ederken bizi hatalara sürükleyebilir. Peki verimizde böyle aykırı değerler varken ne yapmamız gerek; ölçüler konusunda çeyrekliklere, yüzdeliklere, kesikli "tream"li ortalamayı hesaplayıp onun üzerinden yorumlayabiliriz gibi vb.

Değişim Katsayısı

Kod satırı:

```
degisim_katsayisi<-(sd(veri)/mean(veri))*100
```

Çıktı:

```
> degisim_katsayisi
[1] 53.24845
```

Yorum:

Değişim katsayısı farklı ölçeklerdeki verilerin homojenliğini kıyaslamak için doğru bir kullanımdır ancak tek başına da veri hakkında yorum yapabilmemizi de sağlar. Gözlemlerimizin değişim katsayisinin 53.24845 olması da verilerin standart sapmasının ortalamaya göre oldukça yüksek olduğu söylenebilir. Bu veri setinin geniş bir dağılıma sahip olduğu anlamına da gelir.

KAYNAKÇA

- Verilerin, veri setinin kaynağı : https://www.kaggle.com/datasets/davidgdong/20212022-nba-rookie-dataset-roty-analysis