20.03.2024

Veri Görselleştirme

Hafta 4: Dağılımların Görselleştirilmesi

Giriş

Dağılım nedir?

Dağılımların Görselleştirilmesi

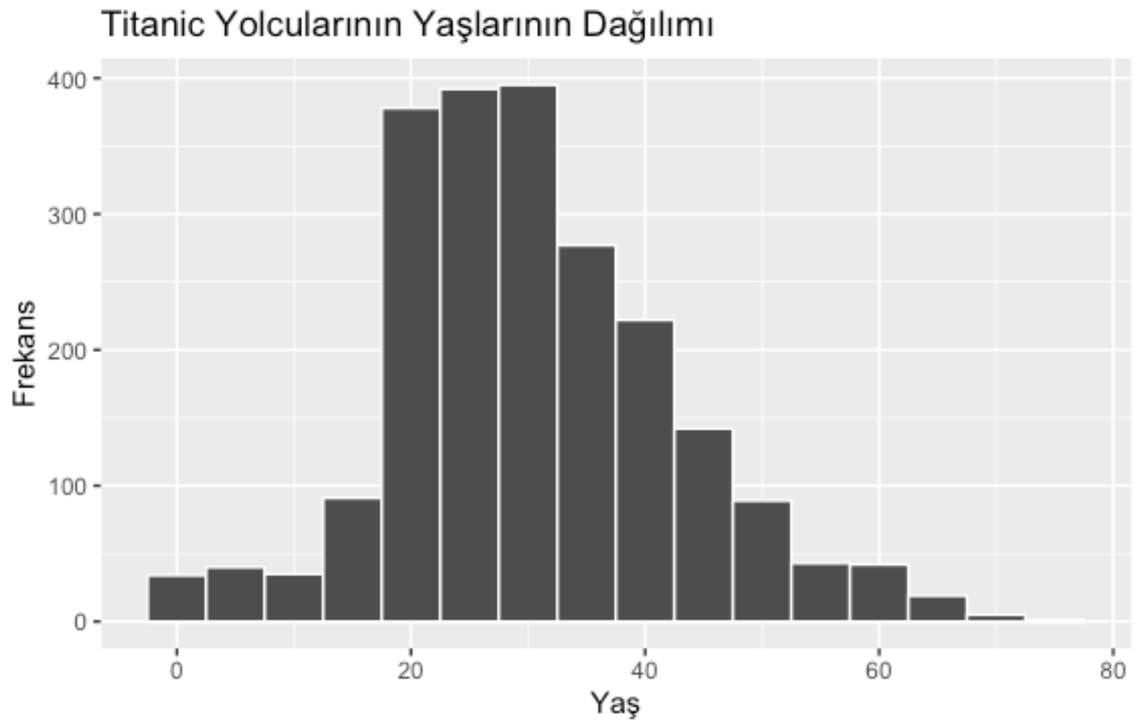
Bir değişkenin dağılımının görselleştirilmesi için:

- Histogram
- Kernel yoğunluk tahmini

kullanılır.

1. Histogram

Gözlem değerlerinin sabit kutu genişliklerine göre gruplandırılarak görselleştirilmesi ile oluşturulur.



1. Histogram

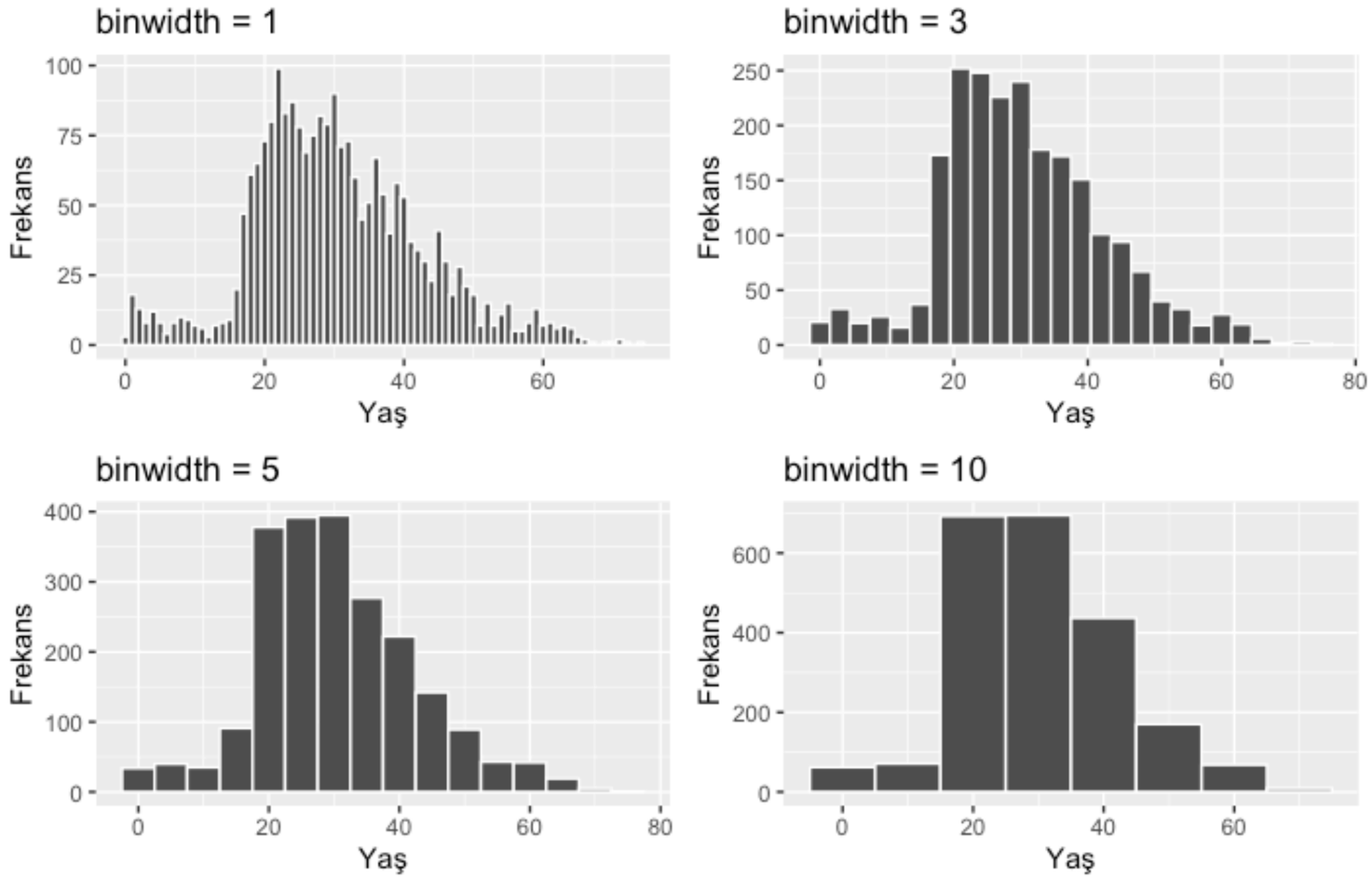
Histogram oluşturulmasında en önemli sorun, görünümünün seçilen kutu genişliğine bağlı olmasıdır.

- Eğer kutu genişliği olması gerektiğinden daha küçük seçilirse, histogramda aşırı pik değerler gözlemlenir ve yorumlanması zorlaşır.

- Olması gerektiğinden daha geniş seçilirlerse, küçük aralıklardaki önemli değişimler histogramda kaybolur ve tespiti mümkün olmayabilir.

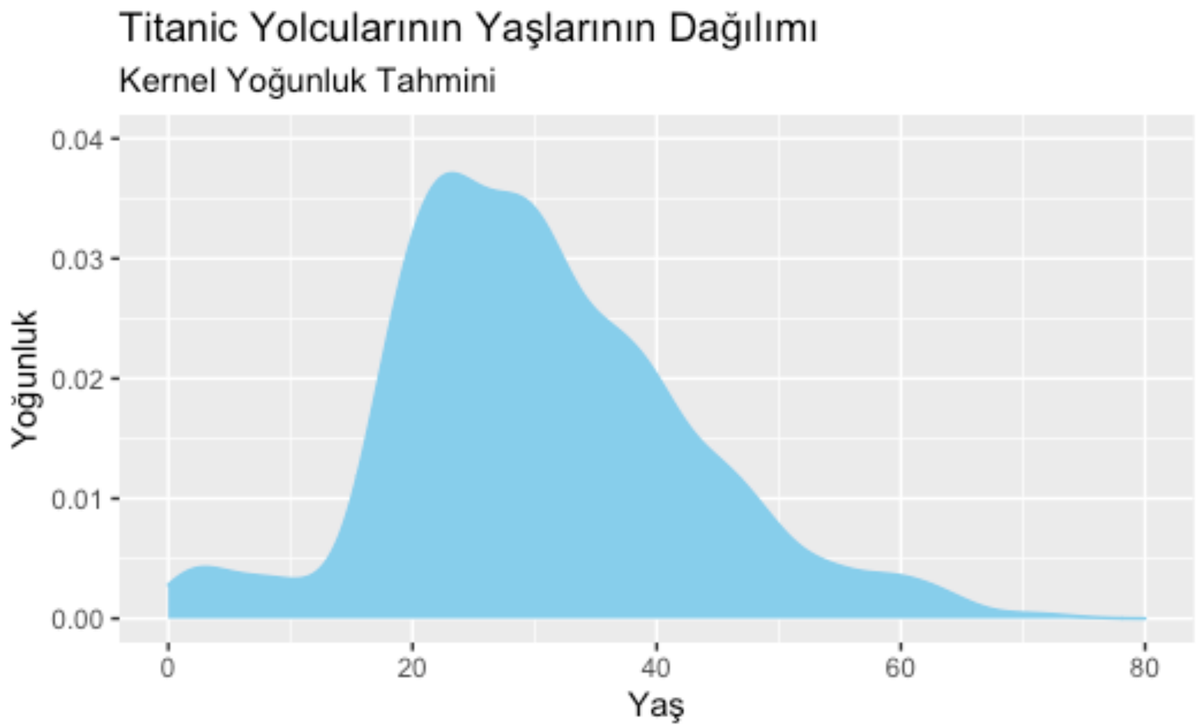
Uygun kutu genişliğinin bulunması, farklı kutu genişliklerinin denenerek en uygununa karar verilmesi ile mümkündür.

1. Histogram


binwidth = 1
Frekans
Yaş
binwidth = 3
Frekans
Yaş
binwidth = 5
Frekans
Yaş
binwidth = 10
Frekans
Yaş

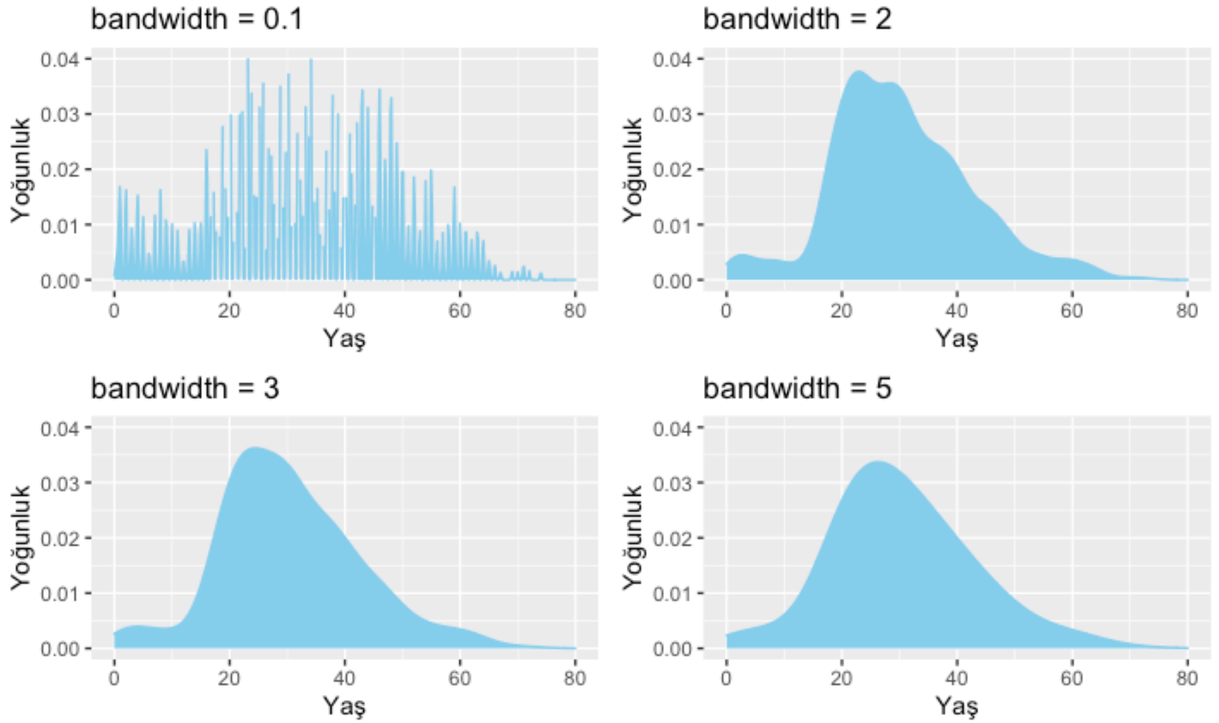
2. Kernel Yoğunluk Tahmini

Pratikte histogram daha sık tercih ediliyor olsa da, kernel yoğunluk tahmininin kullanımı son yıllarda artmıştır.



2. Kernel Yoğunluk Tahmini

Kernel yoğunluk tahminlerinin en önemli sorunu hiç bir gözlem bulunmayan noktalarda gözlem varmış gibi bir görsel ortaya çıkarabilmesidir. Örneğin yaş değişkeni gibi negatif değerler almayan bir değişkenin görselleştirilmesinde negatif bir yaş değeri ile karşılaşılabilir. Bu gibi durumlara karşı dikkatli olunması gerekmektedir.



Birden Fazla Değişkenin Dağılımının Görselleştirilmesi

Sıklıkla birden fazla değişkenin dağılımının görselleştirilmesinin gerektiği durumlarla karşılaşabiliriz.

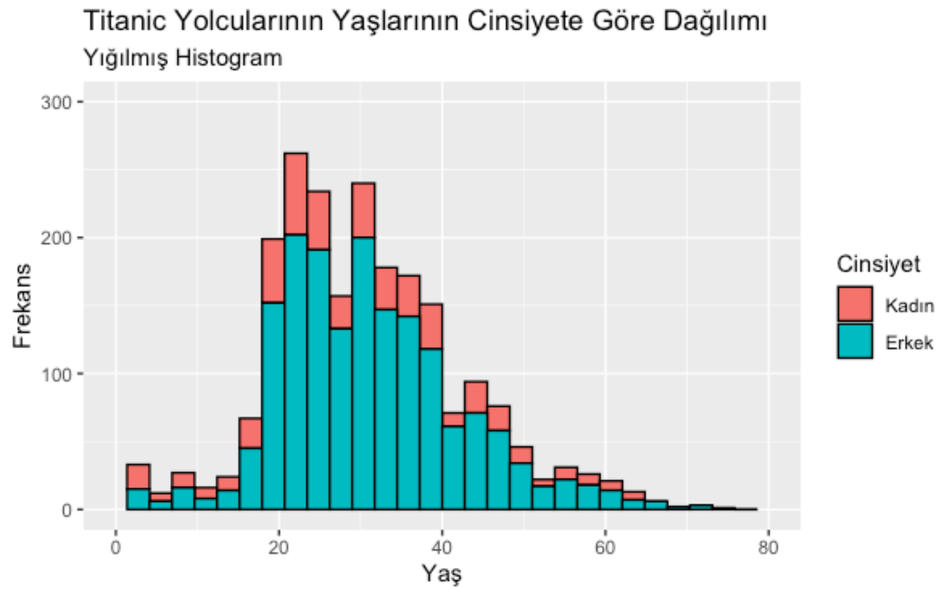
Örneğin, Titanic yolcularının yaşlarının cinsiyete göre dağılımlarını incelememiz, aşağıdaki sorulara yanıt verme ihtiyacı duyabiliriz:

- Erkek ve kadın yolcuların ortalama yaşları benzer miydi?
- Cinsiyetlere göre yolcu yaşları arasında bir fark var mıydı?

Bu gibi durumlarda iki cinsiyet grubu için ayrı ayrı histogramlar oluşturulabilir ya da yığılmış histogram kullanılabilir.

1. Yığılmış (stacked) Histogram

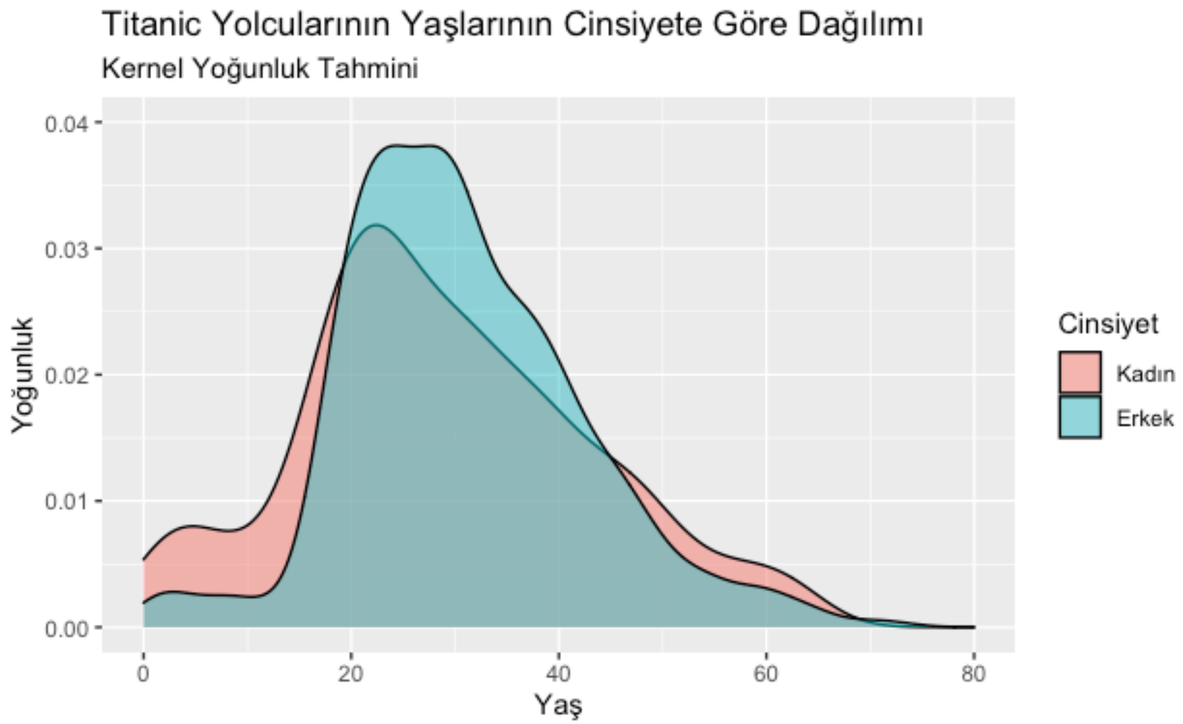
Yığılmış histogram, grupları temsil eden çubukların farklı renkler ile üst üste çizilmesidir.



Grafikte yer alan iki önemli sorun nedir?

2. Kernel Yoğunluk Tahmini

Yığılmış histogramın sınırlılıklarından dolayı birden fazla grup için kernel yoğunluk tahminini kullanmak daha iyi bir çözümdür.



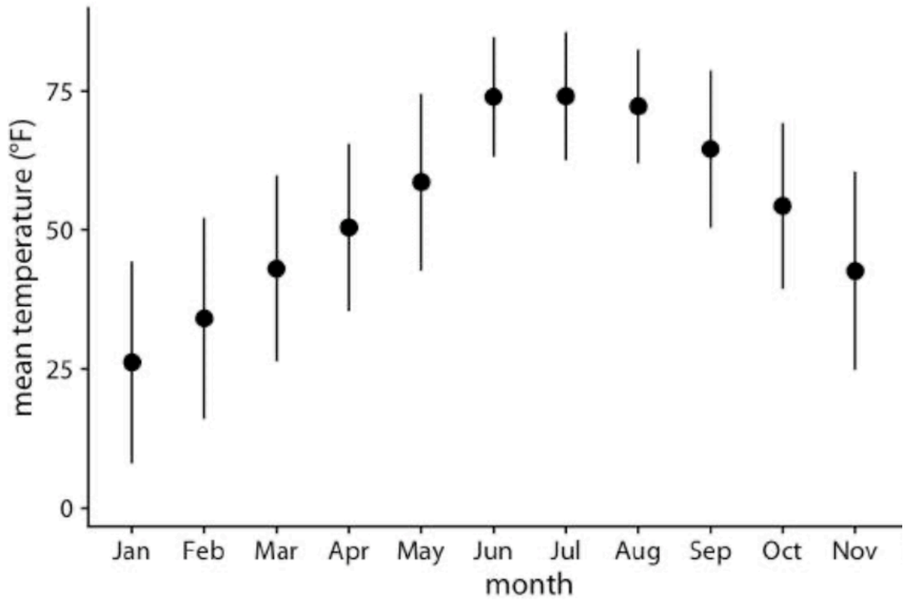
Birçok Değişkenin Dağılımının Görselleştirilmesi

Aynı anda birçok değişkenin dağılımının görselleştirilmesi gereken durumlarla karşılaşılabilir:

- Aylık hava sıcaklıklarının dağılımı
- Ülkelerin kişi başına gelirlerinin dağılımı
- …

1. Hata Çubukları

Bir çok dağılımı aynı anda görselleştirmenin en basit yolu hata çubuklarını kullanmaktır. Hata çubukları farklı şekillerde oluşturulabilir. Bu yollardan biri, medyanın nokta, medyanın bir standart sapma uzaklığını da çubuklar ile göstermektir.



1. Hata Çubukları

Ancak hata çubuklarının bazı sınırlılıkları vardır:

- Yalnızca medyan be standart sapmayı görebildiğimiz için **çok fazla bilgi içermemektedir**.
- Nokta ve çizgilerin neyi temsil ettiği herkes tarafından bilinemeyeceği için **açıklanmasına ihtiyaç vardır**.
- Verinin dağılımındaki **simetri veya asimetriyi göstermez**.

2. Kutu-Bıyık (Box-and-whisker) Grafiği

Veriyi 5 nokta (min, first quantile, median, third quantile, max) ile özetleyerek görselleştirir.

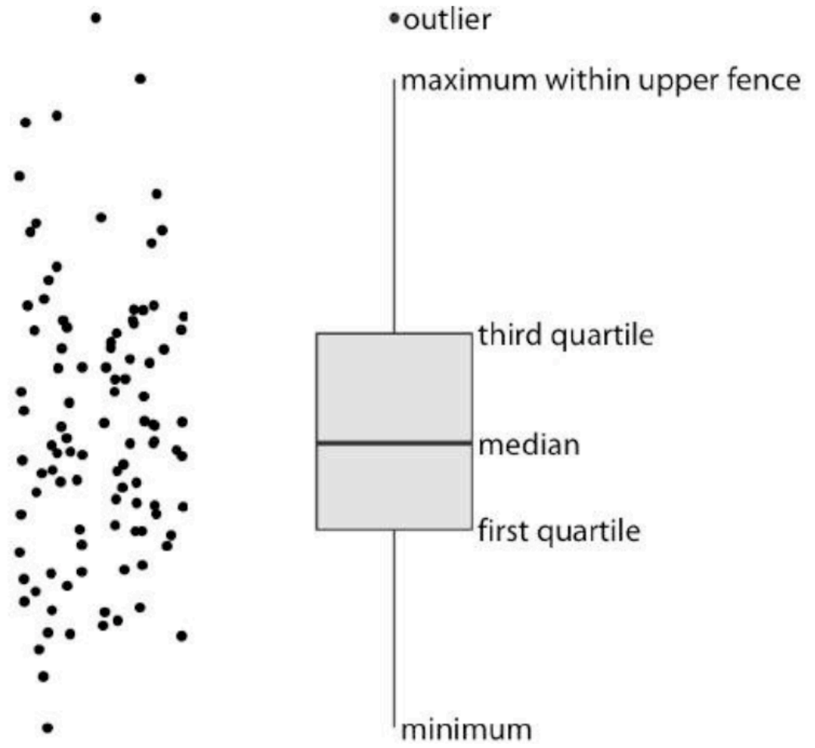
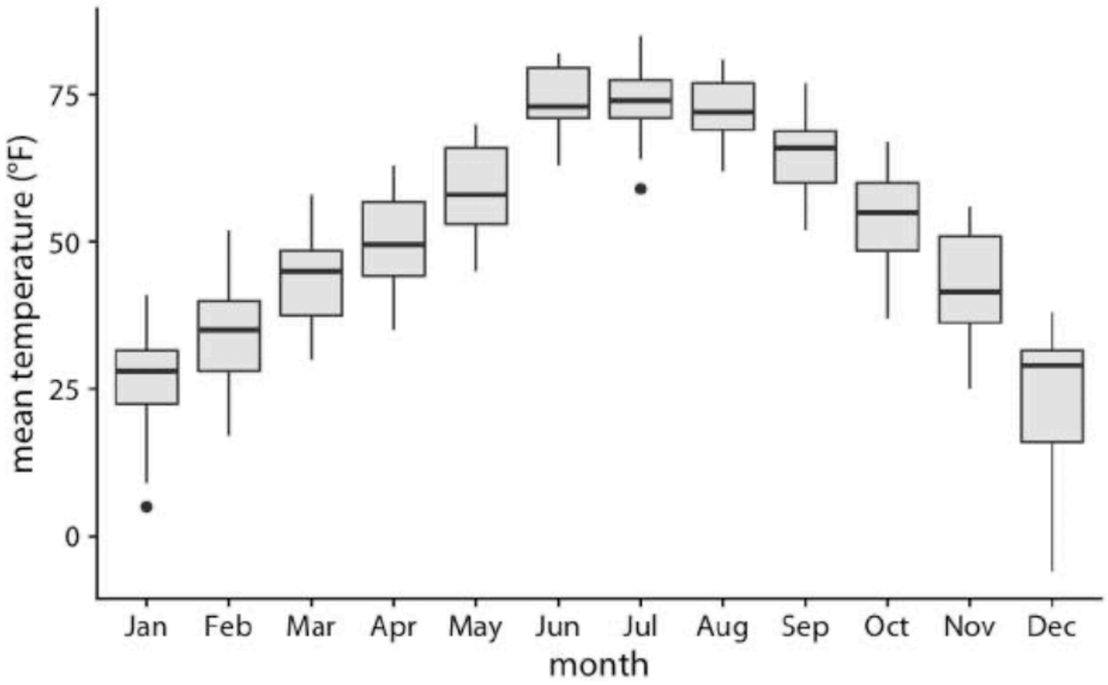


Figure 9-2. Anatomy of a boxplot. Shown are a cloud of points (left) and the corresponding boxplot (right).

2. Kutu-Bıyık (Box-and-whisker) Grafiği

Birçok dağılımı görselleştirmek için kutu grafikleri yan yana kullanılabilirler.



3. Keman (Violin) Grafiği

Kutu grafiklerinin en önemli sınırlılığı iki modlu dağılımları görselleştirememesidir. Bu gibi durumlarda keman grafiği iyi bir alternatiftir.

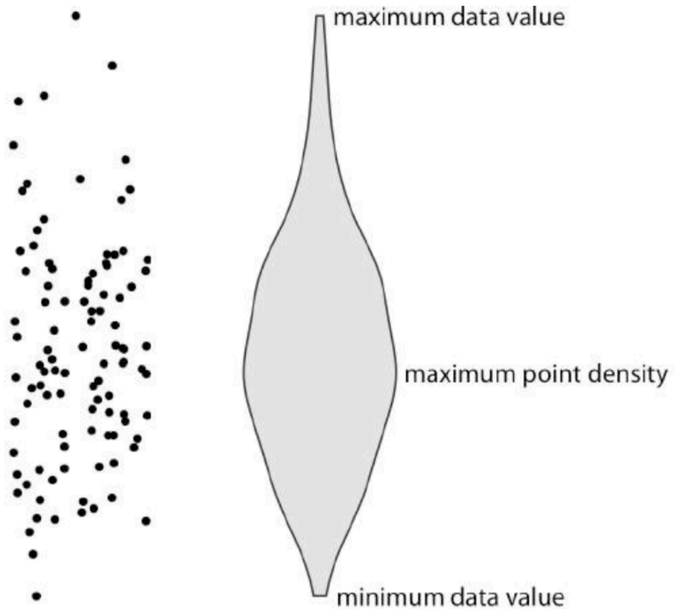


Figure 9-4. Anatomy of a violin plot. Shown are a cloud of points (left) and the corresponding violin plot (right).

- Kemanın genişliği, o noktadaki gözlem değeri yoğunluğunu temsil eder.
- Gözlem değerleri minimum noktasında başlar ve maksimum noktasında biter.
- Keman grafiği kullanmadan önce yeterli sayıda gözlem değerinin olduğundan emin olunması gerekir.

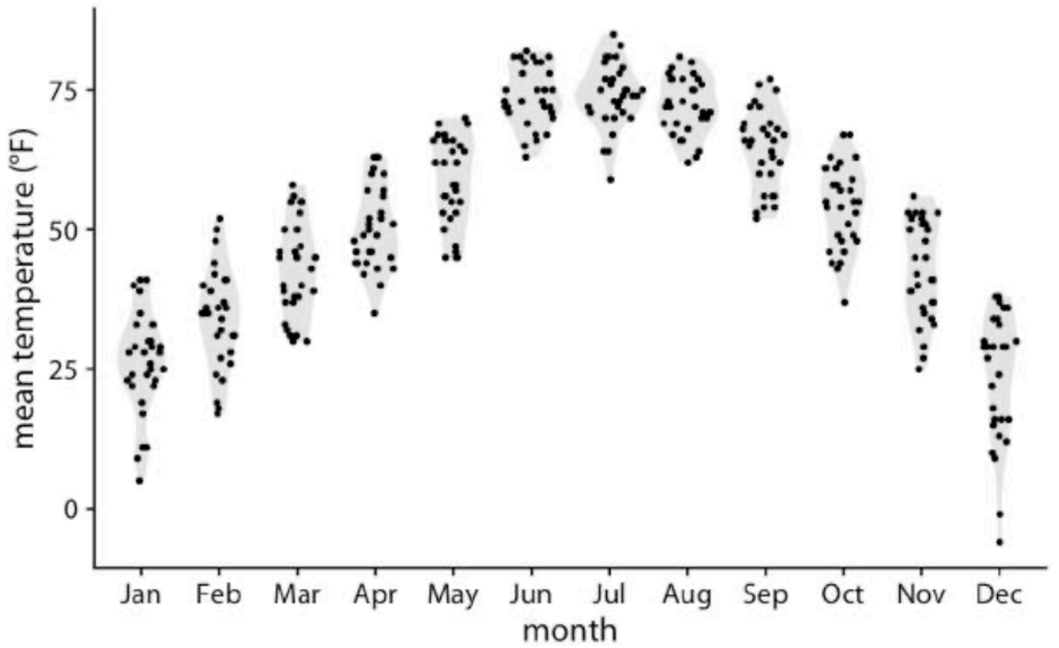
3. Keman (Violin) Grafiği

Keman grafikleri, yoğunluk tahminlerinden türetildiği için bazı sınırlılıkları vardır:

- Hiç bir gözlemin olmadığı yerde gözlem varmış gibi görünebilir.
- Çok az gözlemin olduğu yerde gözlemlerin çok yoğun olduğu görünümü verebilirler.

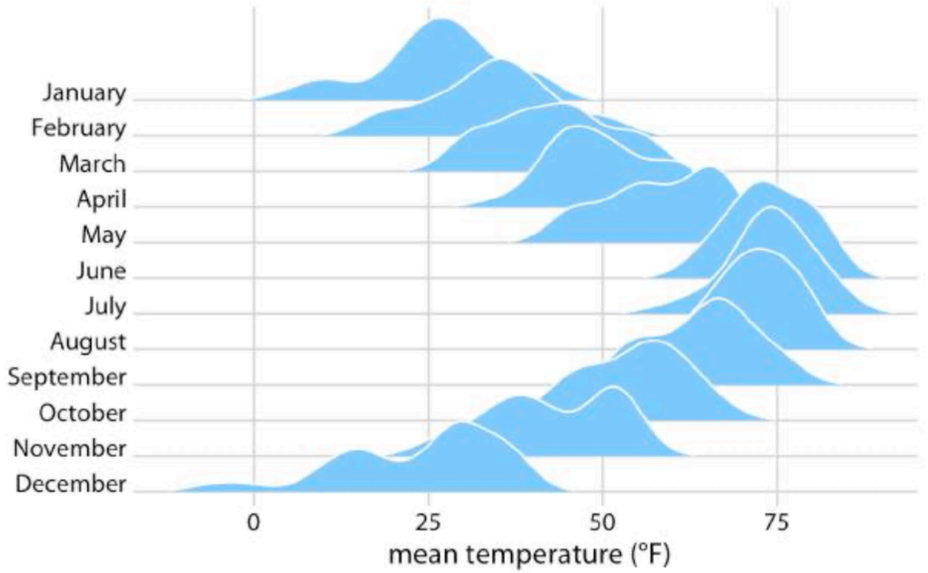
Bu gibi sınırlılıkların önüne geçmek için şerit (strip) grafikleri ile birlikte kullanılabilirler.

3. Keman (Violin) Grafiği



4. Ridgeline Grafiği

Yatay eksende sürekli ve dikey eksende çok düzeyli bir kategorik (genellikle zaman) değişkeninin yerleştirilmesiyle oluşturulur. Bu grafik türünün genel kullanım amacı zaman içerisinde ilgili değişkenin dağılımındaki değişimi gözlemlemektir.



4. Ridgeline Grafiği

Bir değişkenin uzun yıllar boyunca olan değişimini görselleştirmek için kullanılabilir.

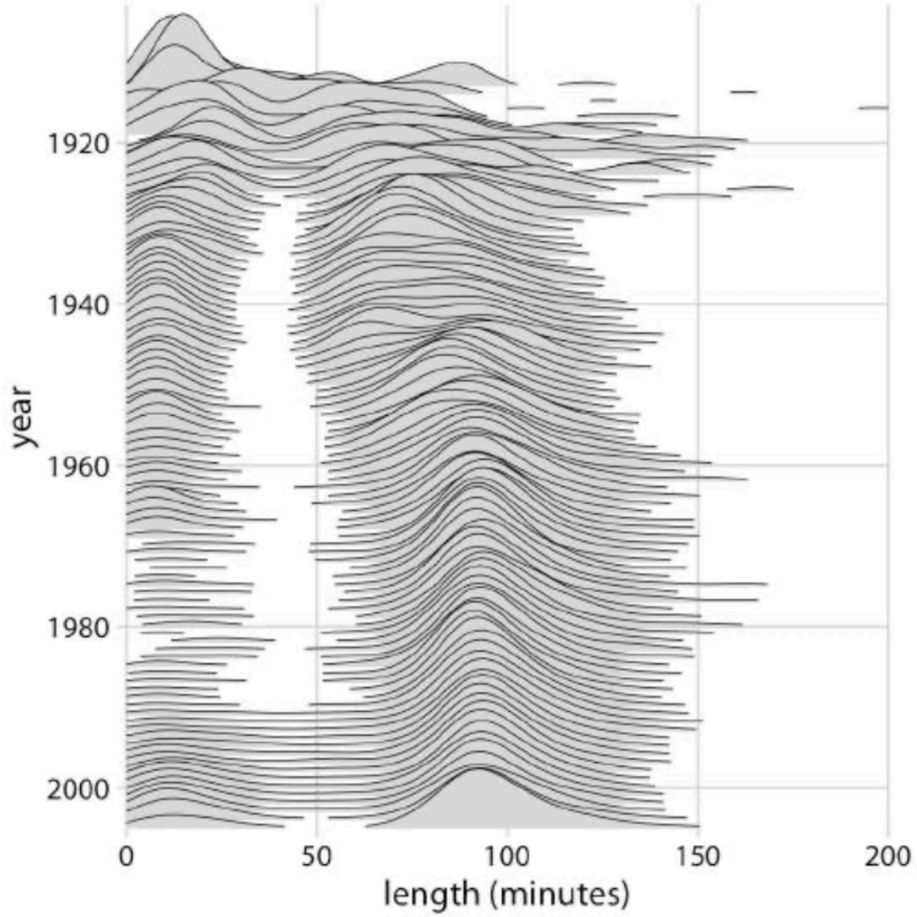
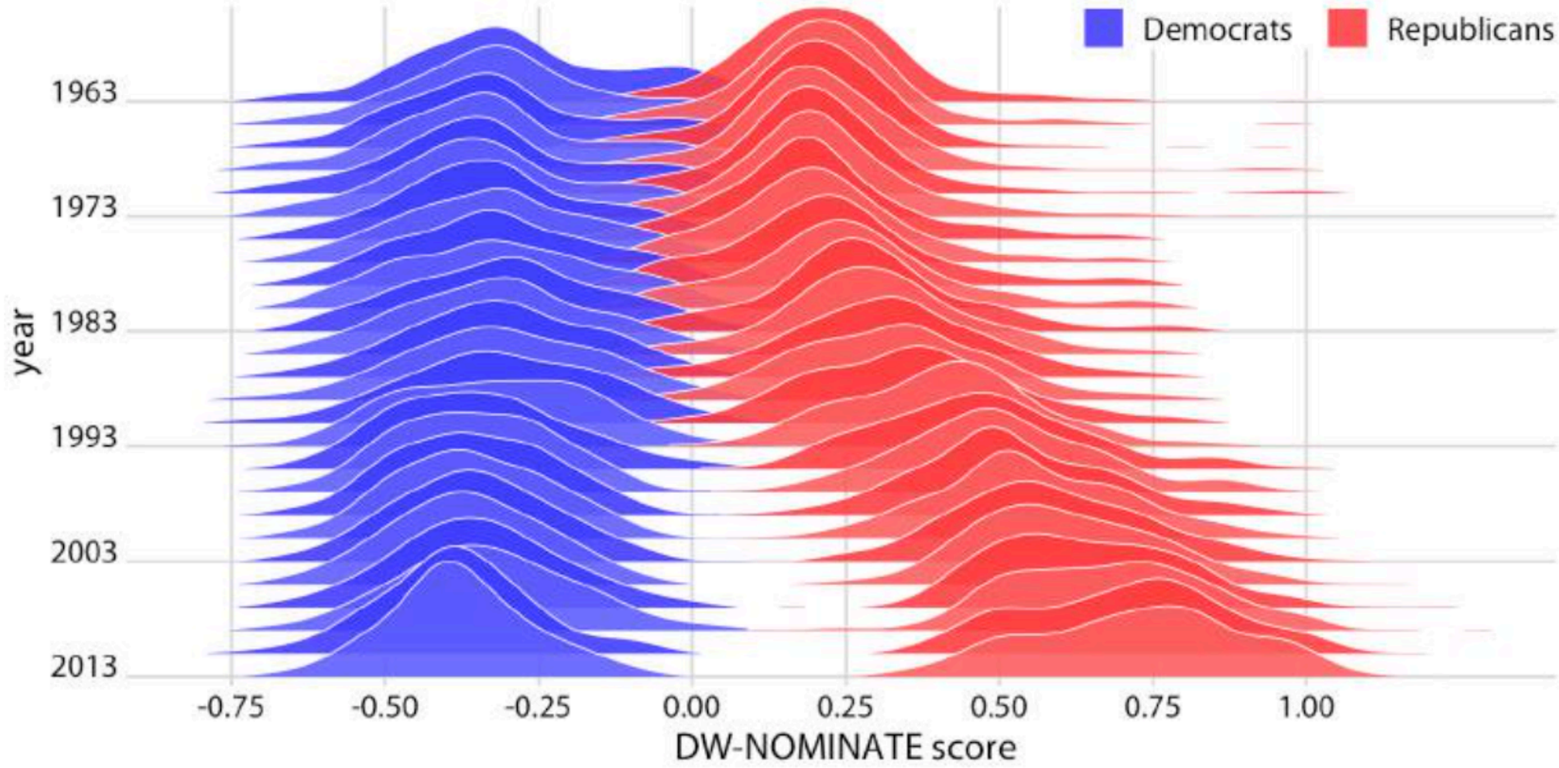


Figure 9-11. Evolution of movie lengths over time. Since the 1960s, the majority of all movies have been approximately 90 minutes long. Data source: Internet Movie Database (IMDB).

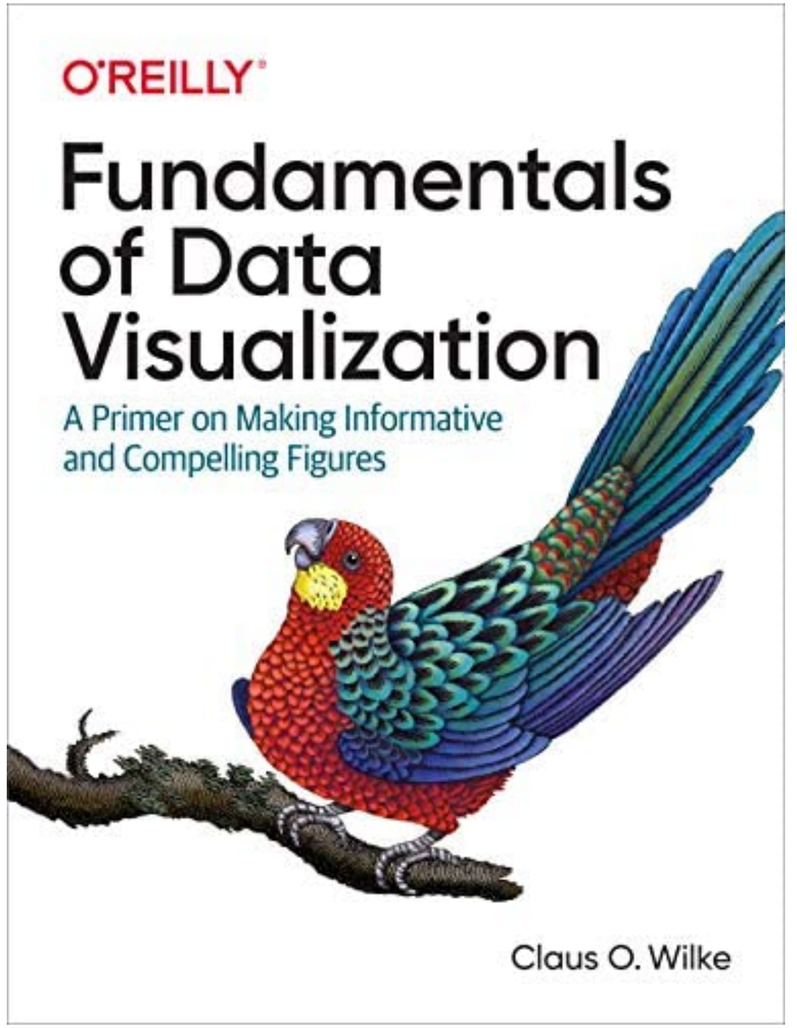
4. Ridgeline Grafiği

Zaman içerisinde iki eğilimi karşılaştırmak için kullanılabilir.



Kaynak

Bu derste yer alan not ve görseller, Claus O. Wilke'nin "Fundamentals of Data Visualization" isimli kitabından derlenmiştir.



Ders materyallerine Mergen üzerinden erişebilirsiniz.
Herhangi bir sorunuz olması durumunda mustafacavus@eskisehir.edu.tr adresini üzerinden e-posta ile bana ulaşabilirsiniz.