Veri Okuryazarlığı

Veri Okuryazarlığı Nedir?

Günlük hayatta veriyle temas ettiğimiz ilk anlardaki basit veri yorumlama kabiliyetleridir. Her türden veri tipini, değişken ve ölçek türlerini tanımlayabilme, betimsel istatistikleri ve istatiksel grafikleri kullanarak veri değerlendirebilme yeteneğidir.

Popülasyon ve Örneklem

Popülasyon ilgilendiğimiz ana kitlemizdir. Örneklem popülasyon içerisinden seçilen alt kümedir. Verinin tamamı popülasyon onu temsil eden örneklemdir. Örneklem ana kitleyi çok iyi temsil etmelidir. Anketler en güzel örnektik. Bütün ülkeye değil ülkenin iyi bir temsilcisi olduğu düşünülen örneklem seçilir ve onun üzerinden tahminlerde bulunulur.

Gözlem Birimi

Araştırmada incelenen birimlerdir. Anketlerde soru sorulan her bir birey gözlem birimidir. Verilerdeki her bir satır gözlem birimidir.

Değişken ve Değişken Türleri

Değişken birimden birime farklı değerler alan niceliktir. Araç veri setinde KM, Araç Fiyatı, Vites Türü vb değişkendir. Değişkenler Sayısal(nicel, kantitatif) ve Kategorik(nitel, kalitatif) olarak ikiye ayrılır. Cinsiyet kategorik değişkendir. Erkek ve Kadın ise bunun sınıflarıdır.

Ölçek Türleri

Sayısal değişkenler için: Aralık ve Oran

Kategorik değişkenler için: Nomimal ve Ordinal

Aralık Ölçek

Başlangıç noktası sıfır olmayan sayısal değişken türüdür. – değerler alabilir.

Sıcaklık bir örnektir.

Oran Ölçek

Başlangıç noktasını sıfır kabul eden sayısal değişken türüdür.

Aracın kilometre bilgisi bir örnektir.

Nominal Ölçek

Sınıflar arasında fark olmadığı kategorik değişkenlerdir.

Kadın erkek bir örnektir.

Ordinal Ölçek

Sınıflar arasında fark olan kategorik değişkenlerdir.

Rütbeler bir örnektir. Albay > Binbaşı gibi.

Eğitim durumu da bir örnektir. Üniversite > Lise gibi.

Merkezi Eğilim Ölçüleri

Aritmetik Ortalama

Bir seride(değişkende) yer alan tüm değerlerin toplanması ve birim sayısına bölünmesi ile elde edilen istatistiktir.

Medyan

Bir seriyi küçükten büyüğe veya büyükten küçüğe sıraladığımızda tam orta noktadan seriyi iki eşit parçaya ayıran değere medyan adı verilir.

N tek ise: (n+1)/2

N çift ise: ((n/2). Terim + ((n/2)+1). Terim) /2

N tek: 10,12,13,15,17 = (5+1)/2 = 3. Terim Medyan = 13

N çift: 10,12,13,13,15,17 = ((6/2).terim + ((6/2)+1). Terim) /2 = (3.Terim + 4.Terim)/2 (13+13)/2 = Medyan = 13

Medyan ve Ortalama Karşılaştırması

Aritmetik ortalama, seri dağılımının(değişken dağılımının) simetrik olduğu bilindiğinde kullanılabilir.

Simetrik olmamak : Değişkenlerin arasında aykırı değer olmasıdır.

13,10,15,12,17,12,19,18,11,12,190

Aritmetik Ortalama = 28.5 Medyan = 13

Burada Aritmetik ortalama bu değerleri temsil edemez. Yakın herhangi bir değer yok. Bu yüzden medyan temsil eder.

Mod

Bir seride (değişkende) en çok tekrar eden değerdir. Normalde algoritmalar ile bulunur.

13,10,13,12,17,13,14 burada en çok tekrar eden 13 Mod = 13

Kartiller

Küçükten büyüğe sıralanan bir seriyi 4 parçaya ayıran değerlere katiller denir.

8,10,15,12,17,20,14 = 8,10,12,14,15,17,20

Q1 = 10 Q2 = 14 Q3= 17

Q1 = ¼ (n+1).terim

Q3 = ¾ (n+1).terim

Q2 = Q3 – Q1

Serinin merkezi eğilimiyle alakalı bilgi almak için kullanılır. Dağılım ölçülerinde de bilgi almak için kullanılabilir. Merkezinin sağında ve solundaki eğilimlerle alakalı da bilgi verir.

Merkezi Eğilimin Önemini Kavrama

1. Temsil Yönünü Kavramak
2. Doğru Kullanılması

2 emlakçı var portföyleri hakkında tartışıyorlar.

A ortalama = 706000

A medyan = 700000

B ortalama = 730000

B medyan = 400000

Medyan ve ortalama birbirine yakın olmalıdır. B ortalama daha yüksek olsa da ayrık değişkeni olduğu için ortalaması yüksektir. A emlakçısının portföyü daha iyidir. B için ortalama kullanılamaz.

Dağılım Ölçüleri

Değişkenin nasıl dağıldığını gösterir.

Değişim Aralığı

Bir serideki maksimum değerden minimum değer çıkartılınca elde edilen değerdir. İlk ilkel dağılım ölçüsüdür.

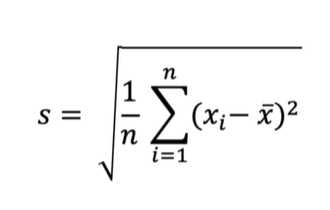
8,10,15,12,17,20,14

D.A = Max – Min = 20-8 = 12

Değişim aralığının küçük olması daha iyidir. Maaş örneği olarak düşünülebilir. Düşük değişim aralığında daha adil bir dağılım vardır.

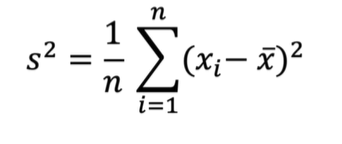
Standart Sapma

Ortalamadan olan sapmaların genel bir ölçüsüdür.

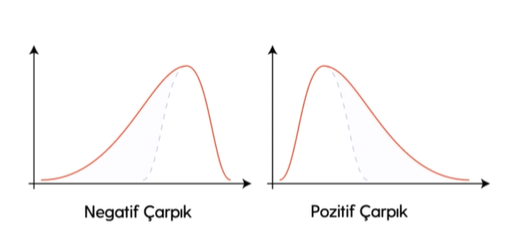


Varyans

Standart sapmanın karesidir.(Ortalamadan olan saplamaların karelerinin ortalamasıdır)



Çarpıklık

Bir değişkenin dağılımının simetrik olamayışıdır. 

Pearson Çarpıklık Katsayısı

(3 (Ortalama – medyan))/standart sapma

PÇK < 0 🡪Negatif çarpık(soldan)

PÇK > 0 🡪Pozitif çarpık(sağdan)

PÇK = 0 🡪Simetrik

Basıklık

Basıklık Katsayısı = m4 / s4

m4 = Ortalamaya göre 4. Moment. Terimlerin ortalamadan olan uzaklıklarının 4. Kuvvetini almak.

s4 = Standart sapmanın 4. Kuvveti

BK = 3 ise dağılım standart normal dağılıma uygundur.

BK > 3 ise dağılım sivridir.

BK < 3 ise dağılım basıktır.

Dağılımın basık olması veri setinin yapısı hakkında bilgi verir. Dağılım şekliyle ilgili ek bilgi verir.

İstatiksel Düşünce Modelleri

Veri okuryazarlığından veri analitiğine giden yolu modelleyen yol göstericilerdir.

Mooney

Verinin Tanımlanması 🡪 Verinin Organize Edilmesi ve İndirgenmesi 🡪 Veri Gösterimi 🡪 Verinin Analiz Edilmesi ve Yorumlanması

İstatiksel Düşünme Düzeyleri

Kişiye Özgülük (Seviye 1)

Veri setlerine bakıp verideki değişkenleri vb ilk temas noktalarında yorumlamaları yapamıyor. Ortalama falan var ama ne anlama geliyor bilmiyor.

Geçici (Seviye 2)

Veri üzerindeki bakış açısı tek yönlü.

Nicel(Seviye 3)

Merkezi eğilim ve yayılım doğru şekilde anlaşılır. İstatistiki kararlar alabilmek için genellikle bu verilere başvurabilme pekişir. Veri ve bağlam arasında ilişkiler kurabilir.

Analitik(Seviye 4)

Veri inceleme analiz etmede tamamıyla analitik yaklaşım kullanılıyor. Veri ve bağlam arasında ilişki kurabiliyor.

Verinin Tanımlanması

Bir şirket internet kullanımı ile ilgili yaptığı araştırmada şu açıklamaları yayınlamıştır: Anketi yanıtlayan 2000 kişinin %43,4’ünü erkekler , %66,4’ünü kadınlar oluşturuyor. Anketi yanıtlayanların %80’i 15-27 yaş aralığındadır. Kadınların %72’si İngilizce bilmektedir. Erkekler günde ortalama 4 saat, kadınlar 4 saat internette zaman geçirmektedir. Erkeklerin %72’si, kadınların %75’i üniversite mezunudur.

1. Çalışmada ölçülmeye çalışılan değişkenler nelerdir?
2. Cinsiyet
3. Yaş
4. İngilizce Bilmek
5. İnternette Geçirilen Zaman
6. Eğitim Durumu
7. Değişken türleri nelerdir?
   1. Cinsiyet - Kategorik
   2. Yaş - Sayısal
   3. İngilizce Bilmek - Kategorik
   4. İnternette Geçirilen Zaman - Sayısal
   5. Eğitim Durumu - Kategorik
8. Belirlediğiniz değişkenlerin hangi ölçekle ölçüldüğünü belirtiniz.
   1. Cinsiyet – Kategorik - Nominal
   2. Yaş – Sayısal - Oran
   3. İngilizce Bilmek – Kategorik - Nominal
   4. İnternette Geçirilen Zaman – Sayısal - Oran
   5. Eğitim Durumu – Kategorik - Ordinal
9. Metni ilk okuduğunuzda dikkatinizi çeken bir anormallik var mı?

Erkek kadın oranı birbirlerini tamamlamıyor %99.7 yapıyor.

Kadınların %72’si İngilizce biliyor peki ya erkekler? Başka hangi dilleri biliyorlar? Geri kalanı hiçbir dil bilmiyor mu?

1. Bu veri seti üzerinde birkaç dakika yorum yapabilir misiniz?

Gözlemimiz 2000. Yansız sağlam bir örneklem seçildiyse bu sayı yeterlidir. Türkiye’deki kadın erkek oranıyla çok uyuşmamakta ama iyi bir oran. Gençlere yönelik bir araştırma yapılmış. Genç nüfus ile alakalı genellemeler yapılabilir. İngilizce bilme durumu her açıdan problemlidir. Ters önermesi anlaşılamıyor. İnternette zaman geçirme oranı normaldir. Üniversite mezunluğu durumunda erkeklerin kadınların kalanı ne mezunu? Bilgi taşır biraz problemlidir.

Verinin Organize Edilmesi ve İndirgenmesi

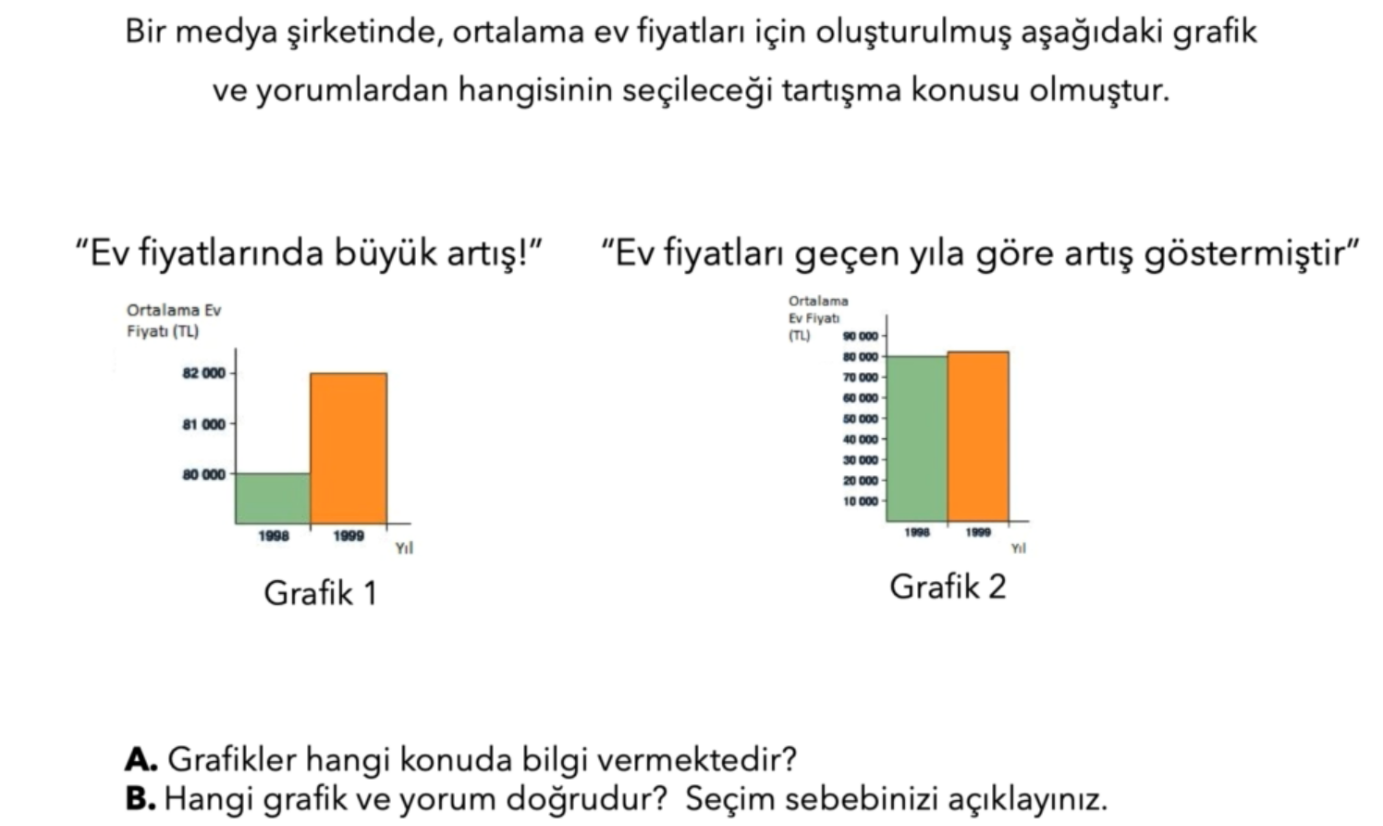


Veri ham durduğu için okunamıyor. Belirli saat aralığı belirleyip bunları toplulaştırıp sunmak faydalı olabilir.



Bu şekilde çok daha rahat okunulabilir ve yorumlanılabilir. İş yerlerine yakın olsaydı 12-15 arasında satış çok daha fazla olmalıydı. İşe gidiş geliş yerleri üstünde olabilir.

Verinin Gösterimi



1. Grafikler Hangi Konuda Bilgi Vermektedir?

Ev fiyatlarındaki artış konusunda bilgi vermektedir. Senelere göre değerlendirme yapılmıştır.

1. Hangi grafik ve yorum doğrudur? Seçim sebebinizi açıklayınız.

Soldaki grafiğin başlangıç noktası 0 değil sağdaki 0. Eksenlerin başlangıç noktaları 0 olmalıdır. Solda değerler 1000 er artarken sağda 10000 er artıyor. Ev fiyatları bazında 1000er artış doğru olmaz. Sağdaki grafik daha doğrudur.

Verinin Analiz Edilmesi ve Yorumlanması

Survivor yarışması için yeni dönem yarışmacıları seçilecektir. Ünlüler ve gönüllüler olarak yarışacak olan iki grubun fiziki ve ruhsal dayanıklılığının ölçülmesi için bir ölçek geliştirilmiş ve grupların dayanıklılıkları ölçülüp aşağıdaki değerler elde edilmiştir.



Uzun sürecek bir yarışma, açlık, doğayla mücadele ve her türlü çekişmenin olacağı bu ortamda hangi grupla beraber çalışıp liderlik etmek isterdiniz? Sebebiyle beraber açıklayınız

Standart sapmanın yüksek olması farklı durumlar olduğunu gösterir. Takım arasındaki fark daha düşük olduğu için, grup davranışları daha benzer daha yakın olan gönüllüler tercih edilmelidir.