**VERİ OKURYAZARLIĞI**

Günlük hayatta veriyle temas ettiğimiz ilk anlardaki basit veri yorumlama kabiliyetleridir.

Veri okuryazarlığı; her türden veri tipini, değişken ve ölçek türlerini tanımlayabilme, betimsel istatistikleri ve istatistiksel grafikleri kullanarak veri değerlendirebilme yeteneğidir.

**Popülasyon ve Örneklem**

Population: Popülasyon ilgilendiğimiz ana kitlemiz, hedef kitlemizdir. - verinin tamamı

Sample: Bu popülasyon, ana kitle içerisinden seçilen bir alt kümedir. – verileri temsil eden alt küme

Örneklem, popülasyonu çok iyi temsil etmelidir.

**Gözlem Birimi (Observation Unit)**

Araştırmada incelediğimiz birimlerdir.

İlgilendiğimiz ana kitle içerisinden bu ana kitleyi temsil eden alt kümenin içerisindeki her bir eleman, incelediğimiz odaklandığımız ana birim/ gözlem birimi olarak adlandırılır.

Bir veri setinde sütunlar değişkenlerdir. Her bir satır ise gözlem birimidir.

**Değişkenler ve Değişken Türleri (Variables and Variable Types)**

Veri setindeki her bir satır gözlem birimidir. - Odaklandığımız birimler

Değişken: Birimden birime farklı değerler alan niceliktir. Sütunlar.

Değişken Türleri:

* Sayısal Değişkenler (nicel, kantitatif)
* Kategorik Değişkenler (nitel, kalitatif)

“Cinsiyet” kategorik değişkendir. “Kadın”, “Erkek” ise bu kategorik değişkenin sınıflarıdır.

“Rütbe” kategorik değişkeninin sınıfları: Onbaşı<Yüzbaşı<Binbaşı<Albay

**Ölçek Türleri (Scales of Measurement)**

* Sayısal değişkenler için: Aralık (Interval) ve Oran (Ratio)
* Kategorik değişkenler için: Nominal ve Ordinal

Başlangıç noktası sıfır olmayan sayısal değişkenlerin ölçek türü **aralıktır.**

Başlangıç noktasını sıfır kabul eden sayısal değişkenlerin ölçek türü **orandır.**

Kategorik değişkenler string veri tipindedir. Bu kategorik değişkenin sınıfları arasında fark yoksa **nominal** ölçek türüdür. kategorik değişkenin sınıfları arasında fark varsa **ordinal** ölçek türüdür.

“Cinsiyet” kategorik değişkendir.

“Kadın”, “Erkek” ise bu kategorik değişkenin sınıflarıdır.

“Kadın” ve “Erkek” sınıfları arasında fark olmadığı için bu değişken **nominal** ölçek türüne sahiptir.

“Rütbe” kategorik bir değişkendir.

Bu kategorik değişkenin sınıfları: Onbaşı < Yüzbaşı < Binbaşı < Albay

Değişkenin sınıfları arasında fark olduğu için “Rütbe” değişkeni **ordinaldır**.

Veri

Negatif değerler alabiliyorsa aralık

Başlangıç noktası 0 ise oran

Sınıflar arası fark var ise ordinal

Sınıflar arası fark yok ise nominal

Kategorik

Sayısal

Oran

Aralık

Nominal

Ordinal

1. **Merkezi Eğilim Ölçüleri**
   1. **Aritmetik Ortalama (Arithmetic Mean)**

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

13, 10, 15, 12, 17, 13 => (13+10+15+12+17+13)/6 = 13.33

Aritmetik ortalama, elimizdeki bir değişkenin ya da serinin ortalamasını temsil edebilmek adına kullandığımız merkezi eğilim ölçüsüdür.

* 1. **Medyan (Median)**

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

* Gözlem birimi sayısı (n) tek ise:

Medyan=((n+1)/2). terim

13, 10, 15, 12, 17 => 10, 12, 13, 15, 17 => ((5+1)/2). terim=3. terim=13

* Gözlem birimi sayısı (n) çift ise:

Medyan=(((n/2). terim + ((n/2)+1). terim)/2)

13, 10, 15, 12, 17, 13 => 10, 12, 13, 13, 15, 17 => (terim3+terim4)/2=(13+13)/2=13

**Medyan ve Ortalama Karşılaştırması**

Aritmetik ortalama, seri dağılımının (değişkenin dağılımının) simetrik olduğu bilindiğinde kullanılabilir. Eğer elimizdeki seri dağılımı (değişkenin dağılımı) simetrik değilse, medyan hesaplanır. Çünkü elimizdeki serinin merkezini temsil etmek için bir değer arıyoruz. Eğer elimizdeki seri dağılımı simetrik değilse (simetrik olmamak, değişken içinde aykırı değerler olduğu anlamına gelir), ortalama yanıltıcı olacaktır.

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

Aritmetik Ortalama: (13+10+15+12+17+12+19+18+11+12+190)/11=29.9

Medyan: 10, 11, 12, 12, 12, 13, 15, 17, 18, 19, 190 => 13

Bu seriyi temsil etmek için medyan seçilir. Çünkü serinin içerisinde aritmetik ortalamaya (29 değerine) yakın herhangi bir değer bulunmamaktadır. Serinin içerisinde 1 tane 190 değeri bulunduğundan simetriyi/dağılımı bozar ve ortalamayı yukarıya doğru çekmiştir. Medyanı (13) incelediğimizde ise serinin içerisinde medyana yakın değerler (11, 12) bulunmaktadır. Dolayısıyla, bu seriyi temsil etmek için medyan seçilir.

Dağılım simetrikse, aritmetik ortalama kullanılır. Eğer serinin içerisinde aykırı değerler bulunuyorsa, simetrik değildir; bu yüzden bu değişkenin temsili değeri için medyan kullanılır.

Aritmetik ortalamayla medyanın birbirine çok yakın olması dağılımın düzgün olduğu, homojen olduğu anlamına gelir.

* 1. **Mod (Mode)**

Bir seride (değişkende) en çok tekrar eden değere **mod** adı verilir.

13, 10, 13, 12, 17, 13, 14 => bu seride en fazla tekrar eden 13 olduğundan mod, 13’tür.

* 1. **Kartiller (Quartiles)**

Küçükten büyüğe sıralanan bir seriyi 4 parçaya ayıran değerlere **kartiller** denir.

8, 10, 15, 12, 17, 20, 14 => 8, 10 (Q1), 12, 14 (Q2), 15, 17 (Q3), 20

Q1 = 1/4(n+1). terim => 1/4(7+1). terim=2. terim=10

Q3 = 3/4(n+1). terim => 6. terim=17

Q2 = (Q3-Q1). terim => 6-2=4. terim=14 (medyan)

**Merkezi Eğilimin Önemini Anlamak**

Portföy temsil istatistiği ?

A – Emlak

* Aritmetik ortalama: (900,000+670,000+800,000+700,000+500,000+700,000+890,000+550,000+640,000)/9=705555.555556
* Medyan: 700,000

Medyan ile aritmetik ortalama birbirine yakın olması nedeniyle herhangi biriyle temsil edilebilir.

B – Emlak

* Aritmetik ortalama: (400,000+480,000+360,000+700,000+240,000+310,000+400,000+3,500,000+200,000)/9=732222.222222
* Medyan: 400,000

3,500,000 değerindeki ev, dağılımın homojenliği bozmuştur. Aykırı bir değerdir. Aritmetik ortalamayı yukarı çekmiştir. Bu yüzden sadece medyan temsil edebilir.

|  |  |
| --- | --- |
| A - Emlak | B - Emlak |
| 900,000 | 400,000 |
| 670,000 | 480,000 |
| 800,000 | 360,000 |
| 700,000 | 700,000 |
| 500,000 | 240,000 |
| 700,000 | 310,000 |
| 890,000 | 400,000 |
| 550,000 | 3,500,000 |
| 640,000 | 200,000 |

Dolayısıyla ikisinin **medyan**

**değerlerine bakarak kıyaslama**

yapabiliriz.

1. **Dağılım Ölçüleri (Measures of Dispersion)**

Dağılım, değişkenin değerlerinin ne şekilde dağıldığını ifade eder. Değişkenin max ve min değerleri, ortalama etrafındaki durumunun ne olduğunu anlamaya çalışmak amacıyla kullanılan ölçülerdir.

* 1. **Değişim Aralığı (Range)**

Bir değişkendeki veya serideki max değerden min değeri çıkardığımızda elde ettiğimiz değerdir.

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

Değişim aralığı küçükse dağılım daha homojendir. Değerler birbirine daha yakındır.

* 1. **Standart Sapma (Standard Deviation)**

Ortalamadan olan sapmaların genel bir ölçüsüdür.

12, 15, 20, 30, 45, 22

(12+15+20+30+45+22)/6=24 Aritmetik ortalama

12-24=-12 -> 144

(144+81+16+36+441+4)/6 = 120

√120=11 Standart sapma

15-24=-9 -> 81

20-24=-4 -> 16

30-24=6 -> 36

45-24=21 -> 441

22-24=-2 -> 4

* 1. **Varyans (Variance)**

Varyans, standart sapmanın karesidir. Ortalamadan olan sapmaların karelerinin ortalamasıdır.

Birden fazla değişkenin dağılımını birbirleriyle kıyaslamak için kullanabiliriz.

12, 15, 20, 30, 45, 22

(12+15+20+30+45+22)/6=24 Aritmetik ortalama

12-24=-12 -> 144

15-24=-9 -> 81

20-24=-4 -> 16

(144+81+16+36+441+4)/6 = 120 Varyans

30-24=6 -> 36

45-24=21 -> 441

22-24=-2 -> 4

* 1. **Çarpıklık (Skewness)**

Çarpıklık bir değişkenin dağılımının simetrik olamayışıdır.

Çarpıklık durumunda medyan değeriyle merkezi eğilimi temsil etmemiz gerekiyor.

**Pearson Çarpıklık Katsayısı:**

3((aritmetik ortalama) - medyan)/(standart sapma)

Aritmetik ortalama ile medyan arasındaki fark; normal şartlar altında, simetrik bir dağılımda neredeyse yok denecek kadar azdır. Eğer dağılım simetrikse, verileri küçükten büyüğe sıraladığımızda medyan ve aritmetik ortalama aynı değeri verir. Dolayısıyla, ortalama ile medyan arasındaki fark, dağılımın simetrik olup olmadığını gösterir. Bu fark çarpıklığı ifade eder ve 3 ile çarparak standart sapmaya böleriz, böylece karşılaştırılabilir bir değer elde ederiz. Bu yöntem sayesinde dağılımın yapısını anlayabiliriz.

PÇK < 0 🡪 Negatif Çarpık (soldan)

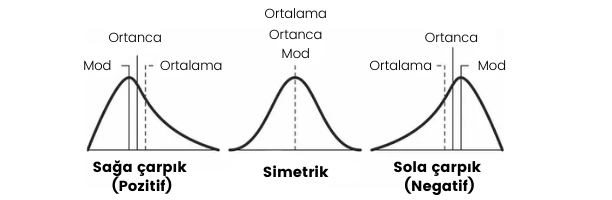
Pearson Çarpıklık katsayısı -1 ile +1 arasında değer alır. Bu katsayı +1'e yaklaştıkça sağa çarpıklık artar, -1'e yaklaştıkça sola çarpıklık artar. Katsayı sıfır olduğunda ise dizi simetrik kabul edilir.



PÇK > 0 🡪 Pozitif Çarpık (sağdan)

PÇK = 0 🡪Simetrik





12, 15, 20, 30, 45, 22

12, 15, 20, 22, 30, 45

Medyan: (20+22)/2=21

Pearson Çarpıklık Katsayısı:

3\*(24 - 21)/11 = 0.82

PÇK > 0 olduğundan sağa çarpıktır (pozitif). Ayrıca +1’ e çok yakın olduğundan yüksek sağa çarpıktır.

(12+15+20+30+45+22)/6=24 Aritmetik ortalama

12-24=-12 -> 144

(144+81+16+36+441+4)/6 = 120

√120=11 Standart sapma

15-24=-9 -> 81

20-24=-4 -> 16

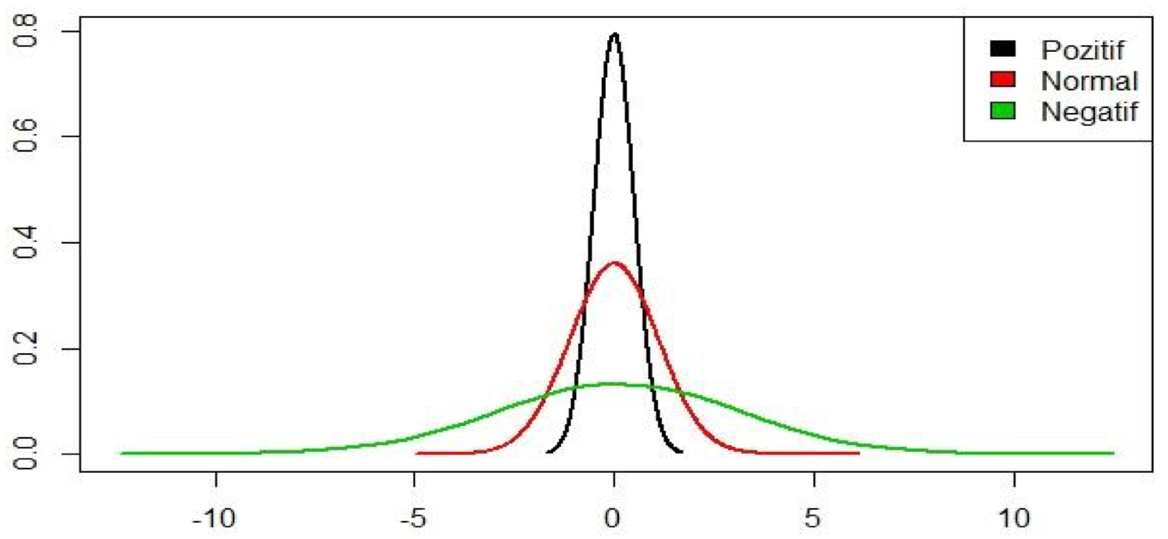
30-24=6 -> 36

45-24=21 -> 441

22-24=-2 -> 4

* 1. **Basıklık (Kurtosis)**

Dağılımın basıklığını/ sivriliğini gösterir.



Normal: Standart dağılıma uygun olan dağılımdır.

Pozitif: Normale göre dağılım sivridir.

Negatif: Normale göre daha aşağıda olduğundan dağılım basıktır.

Bu grafik, bir sürekli değişimin dağılımını göstermektedir. Üç durumda da 0 (sıfır) merkezlenmiştir. Sıfırın etrafındaki dağılıma baktığımızda, pozitif durumda sıfıra çok yakındır, negatif durumda ise ortalamanın pozitif durumla aynı değerde olduğunu görürüz. Ancak negatif dağılıma baktığımızda, dağılımın yanlara doğru yayıldığı ve basık olduğu görülür; standart sapmasının yüksek olduğu görülür.

**Basıklık Katsayısı**= (ortalamaya göre 4. moment)/(standart sapmanın 4. kuvveti)

Ortalamaya göre 4. moment= terimlerin ortalamadan olan uzaklıklarının 4. kuvveti alınır, toplanır ve toplam terim sayısına bölünür.

BK = 3 ise dağılım standart normal dağılıma uygundur. -Normal

BK > 3 ise dağılım sivridir. -Pozitif

BK < 3 ise dağılım basıktır. -Negatif

12, 15, 20, 30, 45, 22

(12+15+20+30+45+22)/6=24 Aritmetik ortalama

terimlerin ortalamadan olan uzaklıklarının

4. Kuvveti ve 2. Kuvveti

Ortalamaya göre 4. moment:

(20736+6561+256+1296+194481+16)/6 = 37224

Standart sapma:

(144+81+16+36+441+4)/6 = 120 => √120=11

4. kuvveti: 11^4=14641

12-24=-12 -> 20736, 144

15-24=-9 -> 6561, 81

20-24=-4 -> 256, 16

30-24=6 -> 1296, 36

45-24=21 -> 194481, 441

22-24=-2 -> 16, 4

BK = 37224/14641 = 2,54

BK < 3 olduğundan dağılım basıktır.

1. **İstatistiksel Düşünce Modelleri (Statistical Thinking Models)**

Veri okuryazarlığından veri analitiğine giden yolu modelleyen yol göstericilerdir.

Analitik düşünce becerilerini veri analitiği kapsamında belirli bir programatik şemayla ele almayı sağlayan modellerdir.

Bu modellerin genel amacı, bir bireyin veriye ilk dokunduğu andan son aşaması olan veriyle ilgili yorumlar yapabilme, analitik çıkarımlarda bulunabilme süreçlerini modelleyen teorik çalışmalardır.

* 1. **Mooney Modeli (2002)**

Mooney modeli 4 basamaktan oluşur:

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

Verinin Tanımlanması:

Veri okuryazarlığını kapsar.

Verinin Organize Edilmesi ve İndirgenmesi:

Analitik anlamda veriden faydalı bilgi çıkarma sürecidir. Veriyi önişlemlerden geçiririz.

Veri Gösterimi:

Veriyi görselleştirmedir. Veri okuryazarlığının bileşenlerinden olan istatiksel grafik yorumlamayı ifade eder.

Bu maddelerin her birisi kendi altında birçok sorudan oluşup, ölçüm modeline dönüştürülür.

Mooney modeline göre analitik düşünce becerilerinin ölçülmeye çalışıldığını düşünürsek, verinin tanımlanmasına ilişkin bir soru seti, Verinin Organize Edilmesi ve İndirgenmesi ile ilgili bir soru seti, Veri Gösteriminde istatistiksel suiistimal grafiklerin verilmesi ile ilgili bir soru seti ve Verinin Analiz Edilmesi ve Yorumlanmasında teorik kavramların gerçek hayatla ilişkilendirilip ilişkilendirilemeyeceğinin ölçülmesine yönelik bir soru seti gibi soru setleri kullanılır. Bu soru setlerine her biri için bir ağırlık verilir ve analitik düşünce becerileri (istatistiksel düşünce becerileri) bu şekilde modellenir.

**İstatistiksel Düşünce Düzeyleri (Statistical Thinking Levels)**

Mooney modelindeki her bir aşamayı ölçülebilir metriklere çevirebilirsek, bunun sonucunda ölçüm yaptıktan sonra elde ettiğimiz değerleri bireye özgü, geçici, nicel ve analitik açıdan değerlendiririz.

* Kişiye Özgülük (Seviye 1)
* Geçici (Seviye 2)
* Nicel (Seviye 3)
* Analitik (Seviye 4)

Kişiye Özgülük:

Genellikle bu basamakta kişiler veri setlerine bakıp değişkenlerin, aritmetik ortalamanın, standart sapmanın vb. ne anlama geldiğini bilemiyor. Veri okuryazarlığını yapamamasıdır.

Geçici:

Nicel düşünmenin öneminin fark edilmeye başlandığı basamaktır. Merkezi eğilim ve yayılım ölçüleri yorumlanabilir ve sayılarla ifade edilebilir, ancak veri üzerindeki bakış açısı tek yönlüdür ve veri temsilleri veya veri analizine yönelik olarak bağlantılar kurulamamaktadır.

Nicel:

Merkezi eğilim ve yayılım ölçülerinin doğru bir şekilde anlaşılmaya başlandığı basamaktır. İstatistiksel kararlar alabilmek için nicel verilere başvurabilme yeteneğinin pekiştiği bir aşamadır. Bağlam ve verilerin ikisine de bilinçli bir şekilde yaklaşılması ve bu kavramlar arasında ufak ufak ilişkiler kurulmaya başlanmasıdır.

Analitik:

Veriyi inceleme, analiz etme, yorumlama konusunda tam anlamıyla analitik bir yaklaşım kurulur. Ortalama nedir, medyan nedir ve bu kavramlar arasındaki farklar nelerdir, standart sapma nedir, varyans nedir gibi konular gerçek hayatla ilişkilendirildiğinde ne anlama geldikleri anlaşılabilir hale gelir. Dolayısıyla, veri ve bağlam arasında ilişki kurulabilir ve bu ilişkiye kanıtlar gösterilebilir.

İstatistiksel düşünce modelleri, ele aldığımız veri okuryazarlığı konularıyla birlikte ve burada öğrenilen kavramların gerçek hayatla ilişkilendirilebilmesi ve analitik anlamda yorumlar yapabilme yolunu modelleyen yaklaşımlardır.

* + 1. **Verinin Tanımlanması**

Bir şirket internet kullanımı ile ilgili yaptığı araştırmada şu açıklamaları yayınlamıştır: Anketi yanıtlayan 2000 kişinin %43.4’ünü erkekler, %66.4’sini kadınlar oluşturmaktadır. Anketi yanıtlayanların %80’i 15-27 yaş aralığındadır. Kadınların %72’si İngilizcesi bilmektedir. Erkekler günde ortalama 3 saat, kadınlar ise 4 saat internette zaman geçirmektedir. Erkeklerin %72’si, kadınların ise %75’i üniversite mezunudur.

**Soru 1)** Çalışmada ölçülmeye çalışılan değişkenler nelerdir ?

* Anketi yanıtlayan 2000 kişinin %43.4’ünü erkekler, %66.4’sini kadınlar oluşturmaktadır. Bu cümleden 1. değişkenin **cinsiyet** olduğu anlaşılır.
* Anketi yanıtlayanların %80’i 15-27 yaş aralığındadır. Bu cümleden 2. değişkenin **yaş** olduğu anlaşılır.
* Kadınların %72’si İngilizcesi bilmektedir. Bu cümleden 3. değişkenin **İngilizce bilme** olduğu anlaşılır.
* Erkekler günde ortalama 3 saat, kadınlar ise 4 saat internette zaman geçirmektedir. Bu cümleden 4. değişkenin **internette geçirilen zaman** olduğu anlaşılır.
* Erkeklerin %72’si, kadınların ise %75’i üniversite mezunudur. Bu cümleden 5. değişkenin **eğitim durumu** olduğu anlaşılır.

**Soru 2)** Değişkenlerin türleri nelerdir ?

* Cinsiyet: Kategorik
* Yaş: Sayısal
* İngilizce Bilme: Kategorik
* İnternette Geçirilen Zaman: Sayısal
* Eğitim Durumu: Kategorik

**Soru 3)** Belirlediğiniz değişkenlerin hangi ölçekle ölçüldüğünü belirtiniz.

* Cinsiyet: Kategorik – Nominal (Kadın ve erkek olmak üzere iki sınıftan oluşur ve bu sınıflar arasında bir fark olmadığından, nominaldir.)
* Yaş: Sayısal – Oran (Başlangıç noktası sıfırdır. Negatif değerler alamaz.)
* İngilizce Bilme: Kategorik – Nominal (bilir, bilmez)
* İnternette Geçirilen Zaman: Sayısal – Oran
* Eğitim Durumu: Kategorik – Ordinal (Sınıflar arasında bir fark olduğundan, bu değişken ordinaldir: ilkokul < ortaokul < lise < üniversite.)

**Soru 4)** Metni ilk okuduğunuzda dikkatinizi çeken bir anormallik var mı ?

* Anketi yanıtlayan 2000 kişinin %43.4’ünü erkekler, %66.4’sini kadınlar oluşturmaktadır. Burada belirtilen oranların toplamı 100’e eşitlenmediği görülmektedir.
* Kadınların %72’si İngilizcesi bilmektedir. Bu ifade ise yeterli şekilde açıklanmamıştır. İngilizceye kıyasla başka bir dil var mı, yoksa başka dillerin arasından İngilizce mi ya da İngilizce bilmek ile bilmemek arasında mı bir değerlendirme yapılmaktadır? Bu sorular bu cümlede cevaplanmamaktadır. Bu da bir anormalliktir.

**Soru 5)** Bu veri seti üzerinde birkaç dakika yorum yapabilir misiniz ?

* Anketi yanıtlayan 2000 kişi olduğundan, **gözlem sayısı (n) 2000’dır.** Eğer bu 2000 kişi yansız bir örneklem ise, bilimsel anlamda Türkiye’yi temsil etme konusunda yeterli bir sayıdır.
* Anketi yanıtlayan 2000 kişinin %43.4’ünü erkekler, %66.4’sini kadınlar oluşturmaktadır. Bu durum anormaldir, çünkü Türkiye'deki kadın-erkek oranıyla örtüşmemektedir. Oranlar %50 civarında olmalıydı, fakat birbirine yakın oranlar olduğundan çok anormal bir durum yoktur.
* Anketi yanıtlayanların %80'i 15-27 yaş aralığındadır. Bu durumda şirketin odak kitlesinin 15-27 yaş aralığındaki gençler olduğu görülür. Genç nüfusa yönelik bir araştırmanın sonucu değerlendirileceğinden, anketi yanıtlayan kişi sayısı 2000 olsa da bunun tüm Türkiye ile ilişkilendirilemeyiz. Burada sadece genç nüfusla ilgili genellemeler yapılabilir.
* Kadınların %72’si İngilizcesi bilmektedir Kadınların %28'i ne biliyor? Peki erkekler ne biliyor? Erkeklerin %28'i başka bir şey mi biliyor? Gibi sorular yanıtsız kaldığından, bu problemli bir maddedir.
* Erkekler günde ortalama 3 saat, kadınlar ise 4 saat internette zaman geçirmektedir. Genç bir kitleyi ele aldığımızdan ve bu kişiler çoğunlukla öğrenci olduğundan bu sonuçlar çıkarılabilir.
* Erkeklerin %72'si, kadınların ise %75'i üniversite mezunudur. Erkekleri ve kadınları üniversite mezunu olanlar ve olmayanlar şeklinde ikiye ayırabiliriz. Elimizdeki 15-27 yaş aralığındaki kitlede, ağırlığı 24 yaşından büyük kişilerin oluşturduğu bir durum söz konusu olabilir. Çünkü üniversite mezunlarının hem kadın hem erkek içindeki dağılımı çok daha fazla olduğu görülmektedir.
  + 1. **Verinin Organize Edilmesi ve İndirgenmesi**

Tabloya nasıl bir düzenleme yapılmalıdır, kısaca açıklayınız.

Belirli bir saat aralığı belirleyip bu aralıktaki kazançları toplulaştırarak sunmak ve bunu bütün tabloya yaparak daha az gözlemden oluşacak şekilde düzenlenebilir.

|  |  |  |  |
| --- | --- | --- | --- |
| Saat | Kazanç (TL) | Saat | Kazanç (TL) |
| 06:50 | 12 | 12:48 | 17 |
| 07:10 | 5 | 13:44 | 10 |
| 07:15 | 8 | 14:10 | 5 |
| 07:30 | 22 | 17:22 | 55 |
| 08:42 | 14 | 18:05 | 2 |
| 09:21 | 9 | 19:48 | 16 |
| 09:26 | 4 | 20:22 | 25 |
| 10:02 | 18 | 20:49 | 21 |
| 11:56 | 12 | 22:40 | 12 |

Saatleri 3’er saatlik dilimlerle olacak şekilde düzenledi.

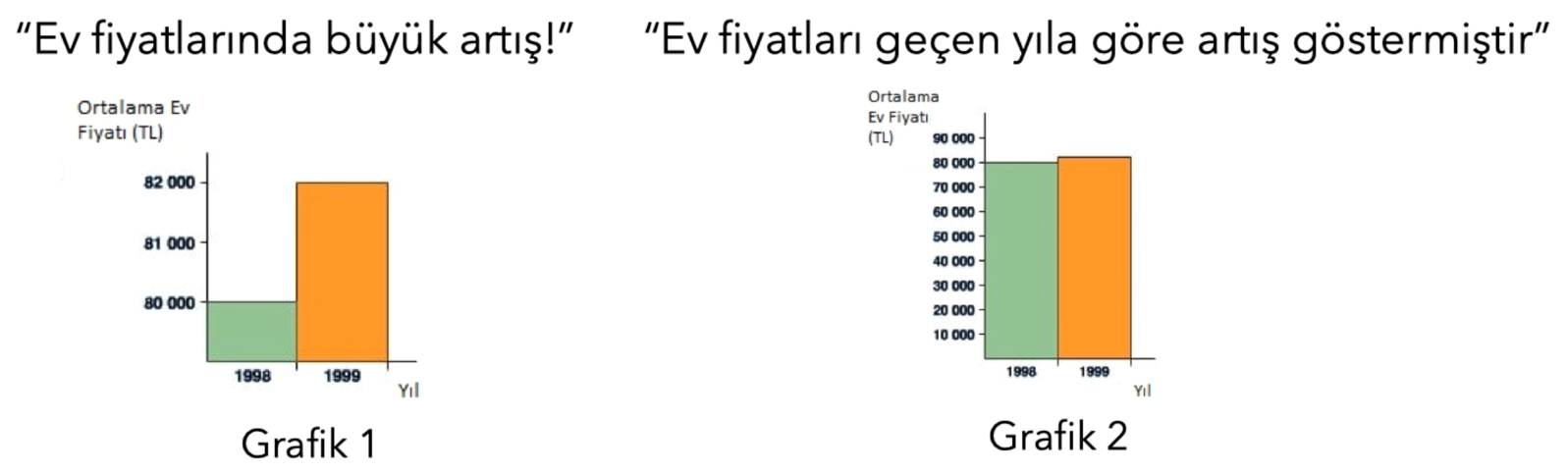
Bu tablonun bir bakkala ait olduğunu düşünelim.

|  |  |
| --- | --- |
| Saatler | Kazanç (TL) |
| 06:00 - 09:00 | 61 |
| 09:00 - 12:00 | 43 |
| 12:00 - 15:00 | 32 |
| 15:00 - 18:00 | 55 |
| 18:00 - 21:00 | 64 |
| 21:00 - 24:00 | 12 |

Eğer bu bakkal iş yerine yakın bir lokasyonda olsaydı, 12:00 ile 15:00 arasındaki satışların diğer tüm aralıklardan daha fazla olması beklenirdi. Çünkü ara verildiğinde insanlar bakkal veya markete giderdi. Bu bakkal iş yerinin işe gidiş-geliş yollarının üzerinde olabileceği veya bir mahallede bulunabileceği düşünülebilir. Çünkü sabah işe gidiş saatleriyle akşam işten dönüş saatleri arasındaki satışlar, diğer saatlere göre en fazla olan saatlerdir.

* + 1. **Verinin Gösterimi**

Bir medya şirketinde, ortalama ev fiyatları için oluşturulmuş aşağıdaki grafik ve yorumlardan hangisinin seçileceği tartışma konusu olmuştur.



**Soru 1)** Grafikler hangi konuda bilgi vermektedir?

Grafikler ev fiyatlarındaki artış konusunda bilgi vermektedir.

**Soru 2)** Hangi grafik ve yorum doğrudur? Seçim sebebinizi açıklayınız.

Grafik 1'in y ekseni 79000'den başladığı görülür. Grafik 2'de ise y ekseninin başlangıç noktası sıfırdır. Eksenin başlangıç noktasının 0 olması gerektiği için Grafik 2'nin başlangıç noktası doğru gösterilmiştir.

Grafik 2'nin ölçeğine baktığımızda değerlerin 10 biner 10 biner arttığı görülür. Grafik 1'in ise biner biner arttığı görülüyor.

Ev fiyatlarını göz önünde bulundurduğumuzda, ev fiyatlarıyla ilgili artıştan bahsedebilmek ve haberleştirebilmek düşünüldüğünde, 1000 TL, 2000 TL gibi fiyatlar ev fiyatının kendi içerisindeki varyansı veya yayılımı olarak adlandırılan durumu göz önünde bulundurarak mantıklı değildir. Ortalama ev fiyatlarının 1000'den 2000'e çıkması, 'Ev fiyatlarında büyük artış!' yorumunu desteklemez.

Grafik 2 ve yorumu doğrudur.

* + 1. **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.

|  |  |  |
| --- | --- | --- |
|  | Ortalama | Standart Sapma |
| Gönüllüler | 70 | 3 |
| Ünlüler | 74 | 8 |

Dayanıklılıkla ilgili bir skor var ve ortalamanın yüksek olması iyi, düşük olması kötü bir durumdur.

Standart sapma bilgimiz olmasaydı ortalaması en yüksek olan ünlüler takımını seçebilirdik.

Standart sapma burada bize grup içi davranış biçimlerinin birbirine yakınlığını söylemektedir. Eğer standart sapma küçükse grup içi daha homojendir, veriler birbirine benzerdir.

Standart sapma, ortalama etrafındaki dağılımın bir ölçüsüdür.

Ünlülerin ortalaması daha yüksek olsa da standart sapması da yüksektir. Bu durum, 80'den daha yüksek skorların mümkün olduğunu söyler. Aynı zamanda, 64, 65, 66 civarında skora sahip kişiler de bulunabilir. Ortalama etrafındaki dağılım geniştir. Ortalama 74 olmasına rağmen skoru 80 olanlar da vardır, ancak ortalamadan düşük skorlar da bulunur.

Gönüllülere baktığımızda ise ortalama 70 ama skorlar 73 ile 67 arasında değişim gösteriyor. Grup olarak birbirine yetenekler konusunda daha fazla benzer bir grup olduğu görülür.

Gönüllüler seçilir çünkü daha homojendir.

Üçüncü günümde, veri analizi ve istatistiksel kavramlar alanında yaptığım çalışmaları aktaracağım. Bu raporda, stajımın üçüncü gününde öğrendiğim ve üzerine çalıştığım istatistiksel kavramları detaylı bir şekilde açıklayacağım.

3. Gün Çalışmaları: Temel İstatistik Kavramları ve Analizi

Stajımın üçüncü gününde, veri okuryazarlığının önemini kavrayarak temel istatistik kavramlarına odaklandım. Bu kavramlar, veri analizi ve yorumlamasında sağlam bir temel oluşturmanın temelini oluşturuyor.

Öğrenilen ve Çalışılan Kavramlar:

1. Popülasyon ve Örneklem: Bir popülasyon, incelenen grubun tamamını ifade ederken, örneklem, bu popülasyonun temsil edici bir alt grubudur. Örneklem, genel popülasyon hakkında çıkarımlar yapmamızı sağlar.

2. Gözlem Birimi: Veri toplamanın temel taşlarından biridir. Bir gözlem birimi, veri setindeki herhangi bir öğeyi temsil eder. Örneğin, bir öğrenci, bir ürün veya bir şirket birer gözlem birimi olabilir.

3. Değişkenler ve Türleri: Değişkenler, veri setindeki özellikleri ifade eder. Sayısal değişkenler ve kategorik değişkenler olarak iki ana türü bulunur. Sayısal değişkenler sayısal değerlerle ifade edilirken, kategorik değişkenler kategorilerle sınıflandırılır.

4. Ölçek Türleri: Veri ölçekleri, değişkenlerin niteliklerini ifade eder. Nominal, ordinal, aralık ve oran ölçekleri temel ölçek türleridir.

5. Merkezi Eğilim Ölçüleri: Merkezi eğilim ölçüleri, veri setinin merkezini veya yoğunlaştığı noktayı ifade eden istatistiksel ölçülerdir. Bu ölçüler, veri dağılımının ortalamasını veya temsil edici bir değeri belirlemeye yardımcı olur. Temel merkezi eğilim ölçüleri şunlardır:

1. Aritmetik Ortalama: Veri setindeki tüm değerlerin toplanıp, toplamın veri sayısına bölünmesi ile elde edilir. Bu ölçü, veri setinin genel "ortalaması"nı temsil eder. Ancak, veri setindeki aykırı değerler bu ölçüyü etkileyebilir.

2. Medyan: Veri setini küçükten büyüğe sıralandığında, ortada yer alan değeri ifade eder. Medyan, aykırı değerlerin etkisine daha dirençlidir ve veri setinin dağılımını daha iyi temsil edebilir.

3. Mod: Veri setinde en sık görülen değeri ifade eder. Mod, nominal veya kategorik verilerde sıklıkla kullanılır. Bir veri seti birden fazla moda sahipse, "çoklu mod" veya "çok modlu" denir.

4. Kartiller (Çeyreklikler): Veri setini dört eşit parçaya bölen değerlerdir. Veri seti sıralandığında, birinci kartil alt çeyrek, ikinci kartil medyan ve üçüncü kartil üst çeyrek değeri temsil eder.

6. Dağılım Ölçüleri: Dağılım ölçüleri, veri setinin yayılma veya dağılım derecesini ifade eden istatistiksel ölçülerdir. Bu ölçüler, verilerin ne kadar homojen veya heterojen dağıldığını anlamak için kullanılır. Temel dağılım ölçüleri şunlardır:

1. Değişim Aralığı: En büyük ve en küçük değer arasındaki farkı ifade eder. Bu ölçü, veri setinin dağılımını yüzeyel bir şekilde belirtir, ancak veri setindeki diğer değerleri göz ardı eder.

2. Standart Sapma: Veri noktalarının ortalama değere olan uzaklıklarının ölçüsünü ifade eder. Daha yüksek standart sapma, veri noktalarının ortalama değerden daha uzak olduğunu gösterir.

3. Varyans: Standart sapmanın karesi olarak hesaplanır. Veri noktalarının ortalama değere olan varyasyonunun karesini ifade eder.

4. Çarpıklık (Skewness): Veri dağılımının simetrisini ifade eder. Pozitif çarpıklık, veri setinin sağa çarpık olduğunu, negatif çarpıklık ise sola çarpık olduğunu gösterir.

5. Basıklık (Kurtosis): Veri dağılımının zirve noktasının yüksekliğini ve kuyruklarının kalınlığını ifade eder. Normal dağılıma yakın veri setleri, düşük basıklığa sahip olacaktır.

Bu ölçüler, veri analizi ve modelleme sırasında verilerin dağılımını anlamada ve yorumlamada yardımcı olur. Her bir ölçü, farklı yönlerden verileri değerlendirmenizi sağlar ve genel veri analizini daha anlamli bir hale getirir.

Kazanımlar ve Sonuç:

Üçüncü günümde, temel istatistik kavramları üzerine yoğunlaşarak veri okuryazarlığını geliştirmeye başladım. Bu kavramlar, veri analizi sürecinin temel taşlarıdır ve veri setlerini anlamamızı, yorumlamamızı ve modellememizi sağlar. Bu çalışmalar, ileride gerçekleştireceğim daha karmaşık veri analizi projeleri için sağlam bir temel oluşturacak.