

BÖLÜM 2

VERİ SETİNİN HAZIRLANMASI VE DÜZENLENMESİ

Veri seti; satırlarında gözlem birimleri, sütunlarında ise değişkenler bulunan iki boyutlu bir matristir. Satır ve sütunların kesişim bölgelerine 'hücre (cell)' adı verilir. Her bir hücreye, gözlemlerin değerleri yani yapılan ölçmenin sonuçları sayı veya sembol olarak girilir. Bir hücrede herhangi bir sayı veya sembol olmaması, amaçlanan gözlemin bulunmadığını ifade eder. Veri setinde yer alan bu tür boşluklar 'kayıp veri (missing value)' olarak adlandırılır.

		Değişkenler						
		A	B	C	D	E	F	G
Gözlem birimleri	1							
	2							
	3							
	4							
	5							
	6							
	7							
	8							
	9							
	10							
	11							
	12							
	13							
	14							
	15							
	16							
	17							

Veri setinin hazırlanmasından önce, yerine getirilmesi gereken işlemler yani gözlem aşamaları bulunmaktadır. Öncelikle gözlenmek istenen özelliğin ve bu özellik ile ilişkili değişkenlerin 'iyi' tanımlanması gerekmektedir. Bu aşama kuramsal ya da kavramsal çerçevenin oluşturulması aşamasıdır. Amaçlı gözlemlerin yapılması ve ilgilenilen özelliğe yönelik verilerin toplanması, sonraki aşamalardan biridir.

Ölçmelerin hassasiyetini ve duyarlılığını artırmak için gözlemlerde sıklıkla bir 'ölçme aracı' ya da 'veri toplama aracı' kullanılır. Farklı özellikler, farklı ölçme araçları ile gözlenir. Örneğin birey özelliklerine yönelik olarak, başarı, zeka gibi bilişsel özellikler başarı testi, zeka testi gibi ölçme araçları ile; tutum, ilgi, kişilik, görüş, inanç gibi duyuşsal özellikler ölçek, envanter, anket gibi ölçme araçları ile; psikomotor beceriler ise kontrol listesi, gözlem formu gibi ölçme araçları ile gözlenebilir. Ölçme aracı kullanmak, ilgilenilen özellik hakkında veri toplamanın yaygın yollarından biridir.

Veri setinin hazırlanması, ancak amaçlanan gözlemin ya da gözlemlerin yapılmasından ve ilgilenilen özelliğe yönelik verilerin 'iyi' tanımlanmış değişkenler düzeyinde toplanmasından sonra, toplanan bu verilerin belli bir düzen içerisinde bir matris ya da çizelge üzerinde gösterilmesi işidir.

Değişkenlerin 'iyi' tanımlanmış olması iki aşamaya vurgudur. Öncelikle ilgilenilen temel özellik hakkında bilgi sağlayacak 'gerekli' ve 'doğru' değişkenlerin belirlenmesi gerekmektedir. İkinci olarak bu değişkenlerin en 'iyi' bilgi verecek şekilde yani mümkün olan en üst ölçek düzeyinde ölçeklenmesi gerekmektedir.

ÖRNEK

Bir sınıftaki 20 öğrencinin derse karşı ilgileri gözlenmek istenmektedir.

Öğrencilerin ilgi düzeyleri, bu amaçla geliştirilmiş bir ölçek kullanılarak ölçülmüş ve 0 ile 60 puan arasındaki toplam ölçek puanları ile tanımlanmıştır. Bu durumda ilgi puanı olarak tanımlanan değişken, nicel, sürekli ve eşit aralıklı ölçek düzeyinde bir değişkendir. (Bu değişken veri seti üzerinde gösterilirken 'ILGI' kısaltması kullanılacaktır.)

Ayrıca derse karşı ilgileri ile ilişkili olabileceği dikkate alınarak, öğrencilerin cinsiyetleri, annelerinin eğitim düzeyleri ve ders başarıları hakkında da veri toplanmıştır.

- Cinsiyet (CINS): Cinsiyet kız öğrenciler için K, erkek öğrenciler için E olmak üzere iki kategorili olarak ölçeklenmiştir. Nitel, kesikli ve sınıflama ölçeği düzeyinde bir değişkendir.
- Anne Eğitim Düzeyi (AED): Anne eğitim düzeyleri, okuryazar değil için 1, okuryazar için 2, ilkökul mezunu için 3, ortaokul mezunu için 4, lise mezunu için 5, yüksekokul mezunu için 6, üniversite/lisans mezunu için 7 ve lisansüstü eğitim mezunu için 9 olmak üzere çok kategorili olarak ölçeklenmiştir. Nitel, kesikli ve sıralama ölçeği düzeyinde bir değişkendir.
- Ders Başarısı (BAS): Ders başarıları olarak yazılı sınav notları dikkate alınmıştır. Yazılı sınav notları 0 ile 100 arasında değişmektedir. Nicel, sürekli ve eşit aralıklı ölçek düzeyinde bir değişkendir.

Görüldüğü gibi öncelikle gözlenecek olan özellik ve bu özelliklerle ilişkili diğer özellikler ayrı birer değişken olarak ayrıntılı bir şekilde tanımlanmaktadır. Her bir değişkenin alabileceği olası değerler yani o değişkene yönelik gözlemlerin sayı ve sembollerle eşleştirilme şekli, gözlem yapılmadan ve veriler toplanmadan önce belirlenmektedir. Ayrıca veri seti üzerinde gösterimde değişkenler için kullanılacak kısaltma ve semboller de önceden belirtilmektedir.

Bu örnekte gözlem birimleri öğrencilerden oluşmaktadır ve toplam 20 gözlem birimi bulunmaktadır. O halde hazırlanacak veri setinde başlık satırı hariç 20 satır bulunmalıdır. İlgi puanları, cinsiyet, anne eğitim düzeyi ve ders başarı notları olmak üzere 4 tane de değişken bulunmaktadır. O halde hazırlanacak veri setinde başlık sütunu hariç 4 sütun bulunmalıdır. Sonuç olarak veri seti, başlık satır ve sütunları dahil 21x5'lik bir matris olacaktır.

Gözlem sonuçları her bir değişken düzeyinde, gözlem birimleri yani öğrencilere göre sıralı olarak aşağıda verilmiştir:

ILGI : 45; 50; 40; 45; 30; 35; 33; 34; 30; 40; 42; 40; 45; 38; 41; 42; 40; 45; 37; 35

CINS : K; K; K; E; E; K; K; K; E; K; E; K; K; E; E; E; K; K

AED : 1; 1; 3; 5; 7; 5; 6; 4; 2; 3; 3; 5; 5; 6; 8; 7; 6; 5; 5; 4

BAS : 80; 90; 85; 85; 60; 60; 70; 72; 75; 55; 83; 80; 75; 70; 65; 70; 80; 90; 75; 70

Şimdi bu gözlemden elde edilen ve yukarıda verilen verileri kullanarak bir veri seti oluşturalım.

	ILGI	CINS	AED	BAS
1	45	K	1	80
2	50	K	1	90
3	40	K	3	85
4	45	E	5	85
5	30	E	7	60
6	35	K	5	60
7	33	K	6	70
8	34	K	4	72
9	30	E	2	75
10	40	K	3	55
11	42	E	3	83
12	40	K	5	80
13	45	E	5	75
14	38	K	6	70
15	41	K	8	65
16	42	E	7	70
17	40	E	6	80
18	45	E	5	90
19	37	K	5	75
20	35	K	4	70

Görüldüğü gibi veri setinde tüm hücreler doludur. Yani amaçlanan her bir gözlem yapılabilmıştır ve bulunmaktadır. Bu tür veri setlerine 'eksiksiz (complete) veri seti' denilmektedir. Bazı gözlemler yapılamamış olsa idi örneğin bazı öğrencilerin annelerinin eğitim düzeyleri hakkında, bazı öğrencilerin cinsiyetleri hakkında ve benzeri veriler elde edilememiş olsa idi, aşağıdaki gibi bir veri seti ortaya çıkabilirdi.

	ILGI	CINS	AED	BAS
1	45	K	1	80
2	50	K	1	90
3	40	K	3	85
4	45	E	5	85
5	30	E	7	60
6	35	K	5	60
7	33	K	6	70
8	34	K		72
9	30	E	2	75
10	40		3	55
11	42	E	3	
12	40	K	5	80
13	45	F		75
14	38	K	6	70
15	41	K	8	65
16	42		7	
17	40		6	80
18	45	E	5	90
19	37	K		75
20	35	K	4	70

Kayıp veriler

Yukarıda görüldüğü gibi veri setinin bazı hücrelerinde boşluklar bulunmaktadır. Bu boşluklar ilgili gözlemin bulunmadığını ifade etmektedir. Örneğin 10. öğrencinin cinsiyeti hakkında bir bilgi bulunmamaktadır. Yine 13. öğrencinin annesinin eğitim düzeyi hakkında bilgi bulunmamaktadır. Bu boşluklar 'kayıp veri' olarak adlandırılmaktadır. Kayıplar içeren bu tür veri setlerine 'eksik (incomplete) veri seti' denilmektedir.

Bir veri setindeki kayıplar çoğaldıkça, amaçlanan özelliğe yönelik olarak elde edilen veriler azalır. Bu durumda verilerden özelliğe yönelik betimleme ve çıkarımların yapılması güçleşir. Bir takım istatistiksel düzeltmelerle belli bir orana kadar kayıplar ihmal edilebilmektedir. Kayıplar çoğaldıkça bu tür düzeltmeler de işe yaramayabilir. Bu durumda eksik gözlemlerin tamamlanması için çaba gösterilmesi belki de ölçme işleminin tekrar edilmesi gerekebilir.

Veri seti hazırlama, elle yapılabileceği gibi bu amaçla kullanılabilecek bilgisayar programları ve yazılımları da bulunmaktadır. Örneğin bir Excell programı, veri seti hazırlamada kullanılabilecek işlevsel bir programdır. Bunun yanı sıra istatistiksel analizlere özel olarak hazırlanmış SPSS, MatLab gibi paket programlarda da veri seti hazırlanabilmektedir. Örneğin yukarıda verilen gözlem sonuçlarının gösterildiği bir veri seti Excell programında düzenlendiği şekliyle aşağıda gösterilmektedir.

	A	B	C	D	E	F
1		ILGI	CINS	AED	BAS	
2	1	45	K	1	80	
3	2	50	K	1	90	
4	3	40	K	3	85	
5	4	45	E	5	85	
6	5	30	E	7	60	
7	6	35	K	5	60	
8	7	33	K	6	70	
9	8	34	K	4	72	
10	9	30	E	2	75	
11	10	40	K	3	55	
12	11	42	E	3	83	
13	12	40	K	5	80	
14	13	45	E	5	75	
15	14	38	K	6	70	
16	15	41	K	8	65	
17	16	42	E	7	70	
18	17	40	E	6	80	
19	18	45	E	5	90	
20	19	37	K	5	75	
21	20	35	K	4	70	
22						

Görüldüğü gibi Excel programında veri seti hazırlarken ilk satır başlık satırı olarak düzenlenmektedir. O nedenle satır sayısını gösteren sayılara bakarak kaç gözlem birimi olduğunu belirlerken dikkatli olmak gerekir. Ayrıca Excell programında veri seti hazırlarken ikinci bir sayfada değişkenleri ve değişkenlerin kategorilerini tanımlamak, ilerleyen aşamalarda yapılacak betimleme ve çıkarım işlemlerinde kolaylık sağlayacaktır.

	A	B	C	D
1	Değişken	Kategoriler		Açıklama
2				
3	ILGI	0-60 puan arası puanlar		Derse Karşı İlgil Ölçeği toplam puanları
4				
5	CINS	K	Kız	Öğrencilerin cinsiyetleri
6		E	Erkek	
7				
8	AED	1	Okuryazar değil	Öğrencilerin annelerinin eğitim düzeyleri
9		2	Okuryazar	
10		3	İlkokul mezunu	
11		4	Ortaokul mezunu	
12		5	Lise mezunu	
13		6	Yüksekokul mezunu	
14		7	Üniversite mezunu	
15		8	Lisansüstü eğitim mezunu	
16				
17	BAS	0-100 arası notlar		Öğrencilerin ders başarıları
18				
19				

Benzer şekilde SPSS, MatLab gibi özel programlarda da veri seti hazırlamak mümkündür. Özellikle gözlem birimlerinin ve değişkenlerin fazla olduğu yani veri setinin büyük olduğu ölçmelerde, bu tür bilgisayar tabanlı uygulamalardan yararlanmak, büyük kolaylık sağlamaktadır.