



5주차: 데이터 가공

ChulSoo Park

School of Computer Engineering & Information Technology
Korea National University of Transportation

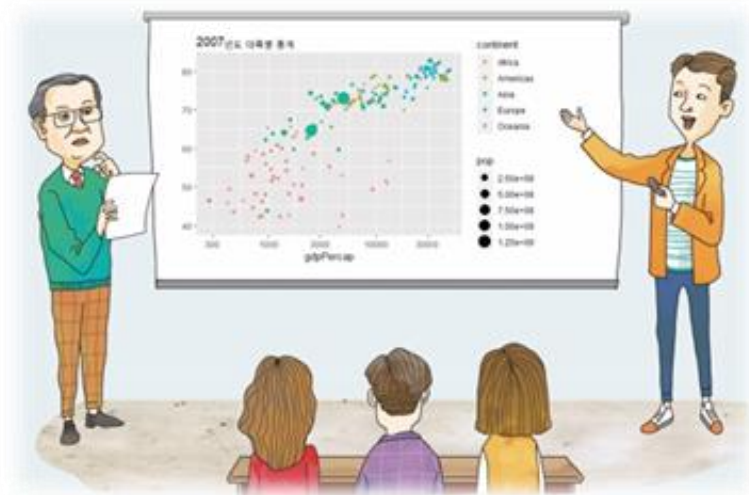
학습목표 (5주차)

- ❖ 데이터 가공의 개념 이해
- ❖ 베이스 R을 이용한 데이터 가공
- ❖ dplyr 라이브러리를 이용한 데이터 가공
- ❖ 대량의 데이터 가공 실습
- ❖ 데이터 가공 실 사례 학습

05

CHAPTER

데이터 가공



CONTENTS

- 5.1 데이터 가공이란?
- 5.2 베이스 R을 이용한 데이터 가공
- 5.3 dplyr 라이브러리를 이용한 데이터 가공
- 5.4 대량의 데이터 가공
- 5.5 데이터 가공 사례 학습
 - 요약

5.4 데이터 가공의 실제 : 모델링을 위한 가공

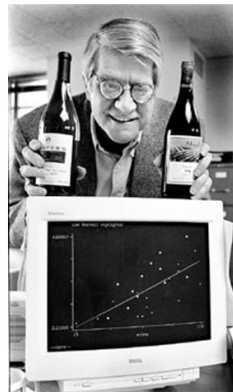
❖ Bordeaux Wine quality as a function of weather condition?

X



- temperature
- sunshine
- precipitation
- humidity
- ...

Y



- price

$$Y = f(X)$$

Vine quality as a function of weather condition?



[Orley Ashenfelter]

$$\begin{aligned} \text{Quality} = & 12.145 + 0.06140 * \text{avg growing season temperature} \\ & + 0.00117 * \text{winter rainfall} - 0.00386 * \text{harvest rainfall} \end{aligned}$$

5.4 데이터 가공의 실제 : 모델링을 위한 가공



입지 정보수집

- ◆ 각종 정보 활동(신문, 잡지 등)
- ◆ 직접 수집(친지, 부동산 중개인, 건축업자)

후보 입지 조사

- ◆ 공부서류 조사(토지, 건물등기부 등본, 건축물관리대장, 도시계획 확인원 등)
- ◆ 점포현장 조사(형태, 면적, 높이, 기동위치 등)

위치 조사

- ◆ 지리적 위치(지형, 지세, 시계성, 홍보성, 교통시설 등)
- ◆ 기능적 위치 조사(지역의 주요기능, 활성화도, 인구유발기능 등)

상권 조사 분석

- ◆ 상권조사(인구, 세대수, 통행량, 상권수준, 경쟁업체, 전망 등)
- ◆ 상권분석(상권 범위설정, 실질 상권 규모 추정 등)

사업타당성 분석

- ◆ 투자비용, 손익분석(투자규모, 매출액, 비용 등)
- ◆ 매출 예측(상권의 흡인력, 1일 고객수, 경쟁점 조사 및 분석)

조건 협의

- ◆ 시설 조건(전기용량, 간판위치, 건물주 협조 공사부문)
- ◆ 계약 조건(임차보증금, 임차료, 권리금, 지불조건, 계약기간)
- ◆ 기타 특약사항, 인상조건 등)

입지 확정 (점포 계약)

5.4 데이터 가공의 실제 : 모델링을 위한 가공

- UCI 리퍼지토리는 기계 학습 알고리즘의 실험적 분석용 데이터의 보관소
- <https://archive.ics.uci.edu/ml/datasets/Wine>에서 wine.data.txt를 다운로드
- 한 개의 관측값은 14개의 수치형 데이터로 구성
- 와인의 종류를 표시하는 첫 번째 열을 제외한 나머지 13개의 수치형 데이터는 와인의 화학적 성분을 분석한 결과다.
- 와인의 종류와 성분 분석치의 상호 연관성을 모델링하는 대표적인 좋은 사례
- 데이터 프레임 내에 측정 속성, 즉 열 이름이 header로 기록되어 있지 않다.
- 해당 정보는 별도의 파일에 설명되어 있어 사용자가 데이터 프레임 내에 통합하거나 수치 데이터와는 별도로 활용해야 한다.

5.4 데이터 가공의 실제 : 모델링을 위한 가공

캘리포니아 남부 태평양 연안에 위치한 도시이다. 북서쪽으로 14km 떨어진 곳에 카운티의 행정중심지인 산타아나(Santa Ana)가 있다. U.S. 고속도로 5번과 주립도로 73번·133번·241번·261번·405번 등이 지난다.

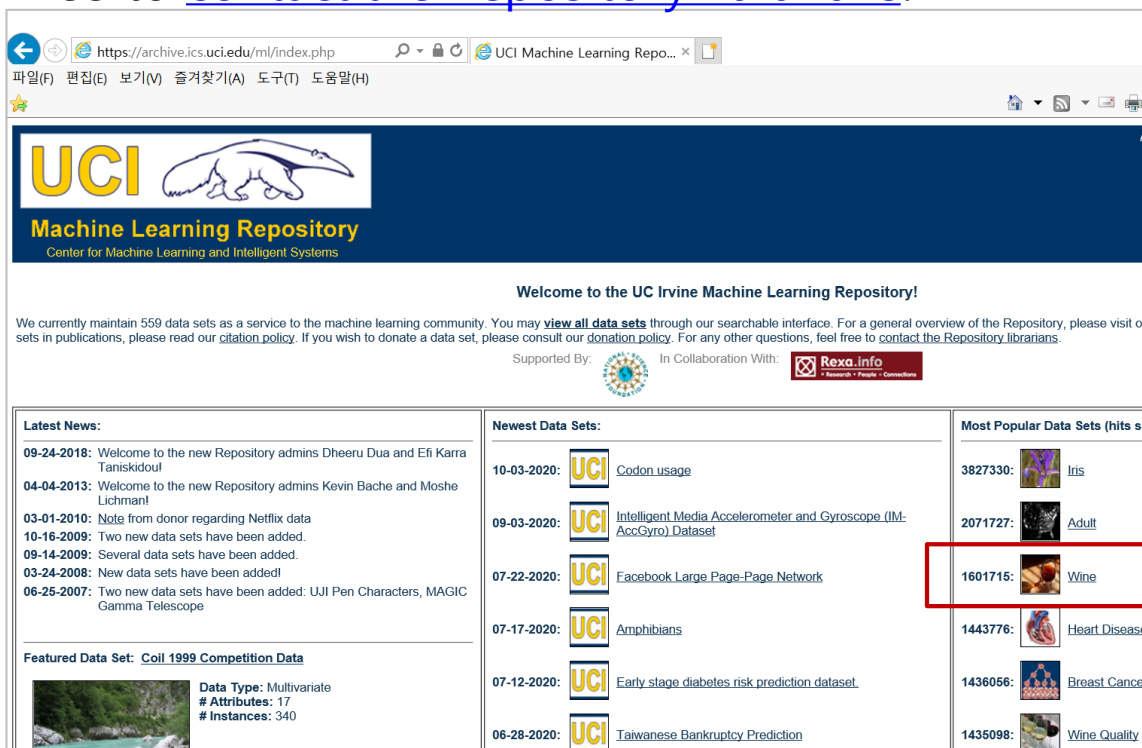


A screenshot of the UCI website. The header includes the UCI logo and navigation links: About, Admissions, Academics, Research, Community, and a user icon. A search bar and 'Web'/'People' filters are present. Below the header is a horizontal menu with links: Top, The Buzz, News, Who We Are, Visit, Events, Arts & Athletics, Initiatives, Health, Alumni & Giving, and Resources. A prominent orange banner contains a COVID-19 notice: 'COVID-19 Notice: For timely updates on campus safety precautions and other public health advisories, visit UCI Forward for COVID-19 and the CDC's COVID-19 site.' Below this are two yellow buttons: 'Support UCI's COVID-19 Efforts' and 'Learn How UCI is Taking Action'. A teal banner below features the text 'BRILLIANT FUTURE The Campaign for UCI' and a 'Learn More' button. The main content area has a background image of students and a large blue box with the text 'Congrats, Anteater admits!' followed by two links: 'Learn why the Class of 2025 is choosing UCI' and 'Take a virtual tour of campus'.

5.4 데이터 가공의 실제 : 모델링을 위한 가공

Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 559 data sets as a service to the machine learning community. You may [view all data sets](#) through our searchable interface. For a general overview of the Repository, please visit our [About page](#). For information about citing data sets in publications, please read our [citation policy](#). If you wish to donate a data set, please consult our [donation policy](#). For any other questions, feel free to [contact the Repository librarians](#).



Latest News:

- 09-24-2018: Welcome to the new Repository admins Dheeru Dua and Efi Karra Taniskidou!
- 04-04-2013: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!
- 03-01-2010: Note from donor regarding Netflix data
- 10-16-2009: Two new data sets have been added.
- 09-14-2009: Several data sets have been added.
- 03-24-2008: New data sets have been added!
- 06-25-2007: Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope

Featured Data Set: [Coil 1999 Competition Data](#)

Data Type: Multivariate
Attributes: 17
Instances: 340

Newest Data Sets:

- 10-03-2020: [UCI Codon usage](#)
- 09-03-2020: [UCI Intelligent Media Accelerometer and Gyroscope \(IM-AccGyro\) Dataset](#)
- 07-22-2020: [UCI Facebook Large Page-Page Network](#)
- 07-17-2020: [UCI Amphibians](#)
- 07-12-2020: [UCI Early stage diabetes risk prediction dataset](#)
- 06-28-2020: [UCI Taiwanese Bankruptcy Prediction](#)

Most Popular Data Sets (hits since 2010):

- 3827330: [Iris](#)
- 2071727: [Adult](#)
- 1601715: [Wine](#)
- 1443776: [Heart Disease](#)
- 1436056: [Breast Cancer](#)
- 1435098: [Wine Quality](#)



UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Wine Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: Using chemical analysis determine the origin of wines



Data Set Characteristics:	Multivariate	Number of Instances:	178	Area:	Physical
Attribute Characteristics:	Integer, Real	Number of Attributes:	13	Date Donated	1991-07-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	1601720

Source:

Original Owners:

Forina, M. et al, PARVUS - An Extendible Package for Data Exploration, Classification and Correlation. Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy.

Donor:

Stefan Aeberhard, email: stefan.aeberhard@coral.cs.ciu.edu.au

■ 데이터 프레임의 열 이름 읽고 쓰기(1)

- 다음 내용을 wine.name.txt라는 별도의 파일에 저장해놓자.

Wine Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: Using chemical analysis determine the origin of wines



Data Set Characteristics:	Multivariate	Number of Instances:	178	Area:	Physical
Attribute Characteristics:	Integer, Real	Number of Attributes:	13	Date Donated	1991-07-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	1601730

Source:

Original Owners:

Forina, M. et al, PARVUS -
An Extendible Package for Data Exploration, Classification and Correlation.
Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno,
16147 Genoa, Italy.

Donor:

Stefan Aeberhard, email: stefan '@' coral.cs.jcu.edu.au

Data Set Information:

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different varieties.

I think that the initial data set had around 30 variables, but for some reason I only have the 13 dimensional version. I have no idea what happened.

The attributes are (donated by Riccardo Leardi, riclea '@' anchem.unige.it)


- 1) Alcohol
- 2) Malic acid
- 3) Ash
- 4) Alcalinity of ash
- 5) Magnesium
- 6) Total phenols
- 7) Flavanoids
- 8) Nonflavanoid phenols
- 9) Proanthocyanins
- 10) Color intensity
- 11) Hue
- 12) OD280/OD315 of diluted wines
- 13) Proline

wine_name - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

- 1) Alcohol
- 2) Malic acid
- 3) Ash
- 4) Alcalinity of ash
- 5) Magnesium
- 6) Total phenols
- 7) Flavanoids
- 8) Nonflavanoid phenols
- 9) Proanthocyanins
- 10) Color intensity
- 11) Hue
- 12) OD280/OD315 of diluted wines
- 13) Proline

■ 데이터 살펴 보기

 wine_data - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

1,14.23,1.71,2.43,15.6,127,2.8,3.06,28,2.29,5.64,1.04,3.92,1065
1,13.2,1.78,2.14,11.2,100,2.65,2.76,26,1.28,4.38,1.05,3.4,1050
1,13.16,2.36,2.67,18.6,101,2.8,3.24,3,2.81,5.68,1.03,3.17,1185
1,14.37,1.95,2.5,16.8,113,3.85,3.49,24,2.18,7.8,86,3.45,1480
1,13.24,2.59,2.87,21,118,2.8,2.69,39,1.82,4.32,1.04,2.93,735
1,14.2,1.76,2.45,15.2,112,3.27,3.39,34,1.97,6.75,1.05,2.85,1450
1,14.39,1.87,2.45,14.6,96,2.5,2.52,3,1.98,5.25,1.02,3.58,1290
1,14.06,2.15,2.61,17.6,121,2.6,2.51,31,1.25,5.05,1.06,3.58,1295
1,14.83,1.64,2.17,14,97,2.8,2.98,29,1.98,5.2,1.08,2.85,1045
1,13.86,1.35,2.27,16,98,2.98,3.15,22,1.85,7.22,1.01,3.55,1045
1,14.1,2.16,2.3,18,105,2.95,3.32,22,2.38,5.75,1.25,3.17,1510
1,14.12,1.48,2.32,16.8,95,2.2,2.43,26,1.57,5,1.17,2.82,1280
1,13.75,1.73,2.41,16,89,2.6,2.76,29,1.81,5.6,1.15,2.9,1320
1,14.75,1.73,2.39,11.4,91,3.1,3.69,43,2.81,5.4,1.25,2.73,1150
1,14.38,1.87,2.38,12,102,3.3,3.64,29,2.96,7.5,1.2,3,1547
1,13.63,1.81,2.7,17.2,112,2.85,2.91,3,1.46,7.3,1.28,2.88,1310
1,14.3,1.92,2.72,20,120,2.8,3.14,33,1.97,6.2,1.07,2.65,1280
1,13.83,1.57,2.62,20,115,2.95,3.4,4,1.72,6.6,1.13,2.57,1130
1,14.12,1.52,2.42,16.5,102,2.3,2.62,22,1.26,6.7,1.22,2.62,1620

Wine Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: Using chemical analysis determine the origin of wines



Data Set Characteristics:	Multivariate	Number of Instances:	178	Area:	Physical
Attribute Characteristics:	Integer, Real	Number of Attributes:	13	Date Donated	1991-07-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	1601730

Source:

Original Owners:

Forina, M. et al, PARVUS -
An Extendible Package for Data Exploration, Classification and Correlation.
Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno,
16147 Genoa, Italy.

Donor:

Stefan Aeberhard, email: stefan '@' coral.cs.jcu.edu.au

Data Set Information:

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different c

I think that the initial data set had around 30 variables, but for some reason I only have the 13 dimensional version. I had a li

The attributes are (donated by Riccardo Leardi, riclea '@' anchem.unige.it)

- 1) Alcohol
- 2) Malic acid
- 3) Ash
- 4) Alcalinity of ash
- 5) Magnesium
- 6) Total phenols
- 7) Flavanoids
- 8) Nonflavanoid phenols
- 9) Proanthocyanins
- 10) Color intensity
- 11) Hue
- 12) OD280/OD315 of diluted wines
- 13) Proline

5.4 데이터 가공의 실제 : 모델링을 위한 가공

■ 데이터 프레임의 열 이름 읽고 쓰기(2)

- wine.name.txt 파일을 읽어들이어 wine 데이터의 열 이름으로 지정한다.
- 문자열 일부를 추출하기 위해 **substr 함수** 사용
- 속성 지정이 끝나면 wine 데이터를 보다 효과적으로 탐색하고 모델링 가능

```
Console C:/RSources/
> wine = read.table("c:/rdata/wine_data.txt",header=FALSE,sep=",")
> head(wine,3)
  V1    V2    V3    V4    V5    V6    V7    V8    V9   V10   V11   V12   V13   V14
1  1 14.23  1.71  2.43 15.6 127  2.80  3.06  0.28  2.29  5.64  1.04  3.92 1065
2  1 13.20  1.78  2.14 11.2 100  2.65  2.76  0.26  1.28  4.38  1.05  3.40 1050
3  1 13.16  2.36  2.67 18.6 101  2.80  3.24  0.30  2.81  5.68  1.03  3.17 1185
> name = readLines("c:/rdata/wine_name.txt")
> head(name,3)
[1] "1) Alcohol "      "2) Malic acid " "3) Ash "
> names(wine)[2:14]<-substr(name, 4, nchar(name))
> names(wine)
[1] "V1"                "Alcohol "
[3] "Malic acid "       "Ash "
[5] "Alcalinity of ash " "Magnesium "
[7] "Total phenols "    "Flavanoids "
[9] "Nonflavanoid phenols " "Proanthocyanins "
[11] "Color intensity "  "Hue "
[13] "OD280/OD315 of diluted wines " "Proline"
> head(wine,3)
  V1 Alcohol Malic acid Ash Alcalinity of ash Magnesium Total phenols Flavanoids
1  1    14.23      1.71 2.43          15.6          127          2.80          3.06
2  1    13.20      1.78 2.14          11.2          100          2.65          2.76
3  1    13.16      2.36 2.67          18.6          101          2.80          3.24
  Nonflavanoid phenols Proanthocyanins Color intensity Hue OD280/OD315 of diluted wines
1           0.28           2.29           5.64           1.04           3.92
2           0.26           1.28           4.38           1.05           3.40
3           0.30           2.81           5.68           1.03           3.17
```

substr(x, start, stop)

5.4 데이터 가공의 실제 : 모델링을 위한 가공

■ 데이터 프레임의 열 이름 읽고 쓰기(2)

Source

Console C:/Rsources/ ↗

```
> name = readLines("c:/rdata/wine_name.txt")
```

경고메시지(들):

```
In readLines("c:/rdata/wine_name.txt") :
```

```
'c:/rdata/wine_name.txt'에서 불완전한 마지막 행이 발견되었습니다
```

Console C:/Rsources/ ↗

```
> wine = read.table("c:/rdata/wine_data.txt",header=FALSE,sep=",")
```

```
> name = readLines("c:/rdata/wine_name.txt")
```

```
> names(wine)[1:14]<-substr(name, 4, nchar(name))
```

```
> head(wine,5)
```

	No Alcohol	Malic acid	Ash	Alcalinity of ash	Magnesium	Total phenols	Flavanoids	
1	1	14.23	1.71	2.43	15.6	127	2.80	3.06
2	1	13.20	1.78	2.14	11.2	100	2.65	2.76
3	1	13.16	2.36	2.67	18.6	101	2.80	3.24
4	1	14.37	1.95	2.50	16.8	113	3.85	3.49
5	1	13.24	2.59	2.87	21.0	118	2.80	2.69

	Nonflavanoid phenols	Proanthocyanins	Color intensity	Hue
1	0.28	2.29	5.64	1.04
2	0.26	1.28	4.38	1.05
3	0.30	2.81	5.68	1.03
4	0.24	2.18	7.80	0.86
5	0.39	1.82	4.32	1.04

	OD280/OD315 of diluted wines	Proline
1	3.92	1065
2	3.40	1050
3	3.17	1185
4	3.45	1480
5	2.93	735

```
> |
```

wine_name - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
1) Alcohol
2) Malic acid
3) Ash
4) Alcalinity of ash
5) Magnesium
6) Total phenols
7) Flavanoids
8) Nonflavanoid phenols
9) Proanthocyanins
10)Color intensity
11)Hue
12)OD280/OD315 of diluted wines
13)Proline
```

wine_name - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
1) No
2) Alcohol
3) Malic acid
4) Ash
5) Alcalinity of ash
6) Magnesium
7) Total phenols
8) Flavanoids
9) Nonflavanoid phenols
10) Proanthocyanins
11)Color intensity
12)Hue
13)OD280/OD315 of diluted wines
14)Proline
```

5.4 데이터 가공의 실제 : 모델링을 위한 가공

■ 데이터 셋 분할하기

- 모델링을 학습하는 데 필요한 학습 데이터, 구해진 모델이 적절한지 검증하기 위한 테스트 데이터는 주어진 데이터 셋을 일정 비율로 분할하여 얻게 된다.

Wine type

Console C:/RSources/ ↗

```
> str(wine)
'data.frame': 178 obs. of 14 variables:
 $ No : int 1 1 1 1 1 1 1 1 1 1 ...
 $ Alcohol : num 14.2 13.2 13.2 14.4 13.2 ...
 $ Malic acid : num 1.71 1.78 2.36 1.95 2.59 1.76 1.87 2.15 1.64 1.35 ...
 $ Ash : num 2.43 2.14 2.67 2.5 2.87 2.45 2.45 2.61 2.17 2.27 ...
 $ Alcalinity of ash : num 15.6 11.2 18.6 16.8 21 15.2 14.6 17.6 14 16 ...
 $ Magnesium : int 127 100 101 113 118 112 96 121 97 98 ...
 $ Total phenols : num 2.8 2.65 2.8 3.85 2.8 3.27 2.5 2.6 2.8 2.98 ...
 $ Flavanoids : num 3.06 2.76 3.24 3.49 2.69 3.39 2.52 2.51 2.98 3.15 ...
 $ Nonflavanoid phenols : num 0.28 0.26 0.3 0.24 0.39 0.34 0.3 0.31 0.29 0.22 ...
 $ Proanthocyanins : num 2.29 1.28 2.81 2.18 1.82 1.97 1.98 1.25 1.98 1.85 ...
 $ Color intensity : num 5.64 4.38 5.68 7.8 4.32 6.75 5.25 5.05 5.2 7.22 ...
 $ Hue : num 1.04 1.05 1.03 0.86 1.04 1.05 1.02 1.06 1.08 1.01 ...
 $ OD280/OD315 of diluted wines : num 3.92 3.4 3.17 3.45 2.93 2.85 3.58 3.58 2.85 3.55 ...
 $ Proline : int 1065 1050 1185 1480 735 1450 1290 1295 1045 1045 ...
> |
```

■ 데이터 셋 분할하기

- 임의의 샘플을 취하여 분할하는 것이 중요함 !
- dplyr에서 제공하는 sample_frac나 sample_n 함수를 사용하면 간편

Console C:/RSources/ ↗

```
> library(dplyr)
> train_set = sample_frac(wine, 0.7)
> test_set = setdiff(wine, train_set)
> str(train_set)
'data.frame': 125 obs. of 14 variables:
 $ No                : int  3 2 3 2 1 3 3 2 1 2 ...
 $ Alcohol            : num  12.9 12.5 13.7 13.1 13.4 ...
 $ Malic acid         : num  2.99 1.52 3.26 1.01 3.84 5.65 1.24 3.74 1.81 5.8 ...
 $ Ash                : num  2.4 2.2 2.54 1.7 2.12 2.45 2.25 1.82 2.41 2.13 ...
 $ Alcalinity of ash  : num  20 19 20 15 18.8 20.5 17.5 19.5 20.5 21.5 ...
 $ Magnesium          : int  104 162 107 78 90 95 85 107 100 86 ...
 $ Total phenols       : num  1.3 2.5 1.83 2.98 2.45 1.68 2 3.18 2.7 2.62 ...
 $ Flavanoids          : num  1.22 2.27 0.56 3.18 2.68 0.61 0.58 2.58 2.98 2.65 ...
 $ Nonflavanoid phenols : num  0.24 0.32 0.5 0.26 0.27 0.52 0.6 0.24 0.26 0.3 ...
 $ Proanthocyanins     : num  0.83 3.28 0.8 2.28 1.48 1.06 1.25 3.58 1.86 2.01 ...
 $ Color intensity     : num  5.4 2.6 5.88 5.3 4.28 7.7 5.45 2.9 5.1 2.6 ...
 $ Hue                 : num  0.74 1.16 0.96 1.12 0.91 0.64 0.75 0.75 1.04 0.73 ...
 $ OD280/OD315 of diluted wines : num  1.42 2.63 1.82 3.18 3 1.74 1.51 2.81 3.47 3.1 ...
 $ Proline             : int  530 937 680 502 1035 740 650 562 920 380 ...
> str(test_set)
'data.frame': 53 obs. of 14 variables:
 $ No                : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Alcohol            : num  14.2 14.8 13.9 14.1 14.8 ...
 $ Malic acid         : num  1.76 1.64 1.35 2.16 1.73 1.59 2.05 1.72 1.8 1.64 ...
 $ Ash                : num  2.45 2.17 2.27 2.2 2.2 2.2 2.48 2.22 2.14 2.65 ...
```

■ 데이터 구조 변경

- gapminder 패키지의 데이터는 gapminder 웹 사이트에서 제공하는 데이터 등 중 극히 일부분
- 여러 지표를 총체적으로 분석하려면 다양한 항목의 관측값을 정리하여 하나의 데이터 프레임으로 통합하는 가공 작업이 필요
- gapminder 사이트 : <https://www.gapminder.org/data/>

5.4 데이터 가공의 실제 : 데이터 구조 변경

- 데이터 취득 : <https://www.gapminder.org/data>

The image shows two overlapping screenshots of the Gapminder website. The top screenshot shows the 'Download the data' page with a search bar and a 'Donate' button. The bottom screenshot shows the 'Choose individual indicators' page, which lists all indicators available in Gapminder Tools. A search bar is present, and a dropdown menu is open, showing the search results for 'electric'. The dropdown menu lists 'Electricity' as the selected indicator, and a list of related indicators is shown on the right, including 'Electricity generation /person', 'Electricity generation, total', 'Electricity use /person', 'Residential electricity use', and 'Residential electricity use /...'. The 'Electricity' indicator is highlighted in the dropdown menu, and the 'Electricity generation /person' indicator is highlighted in the list on the right.

Download the data | Gapminder x +

gapminder.org/data/

Search

Home > Download the data

Download the data

Choose individual indicators

This menu lists all indicators available in Gapminder Tools. Select one to preview

Select an indicator *

electric

Electricity

Electricity generation /person

Electricity generation, total

Electricity use /person

Residential electricity use

Residential electricity use /...

Electricity generation, per person

Electricity generation per person (kilowatt-hours)

5.4 데이터 가공의 실제 : 데이터 구조 변경

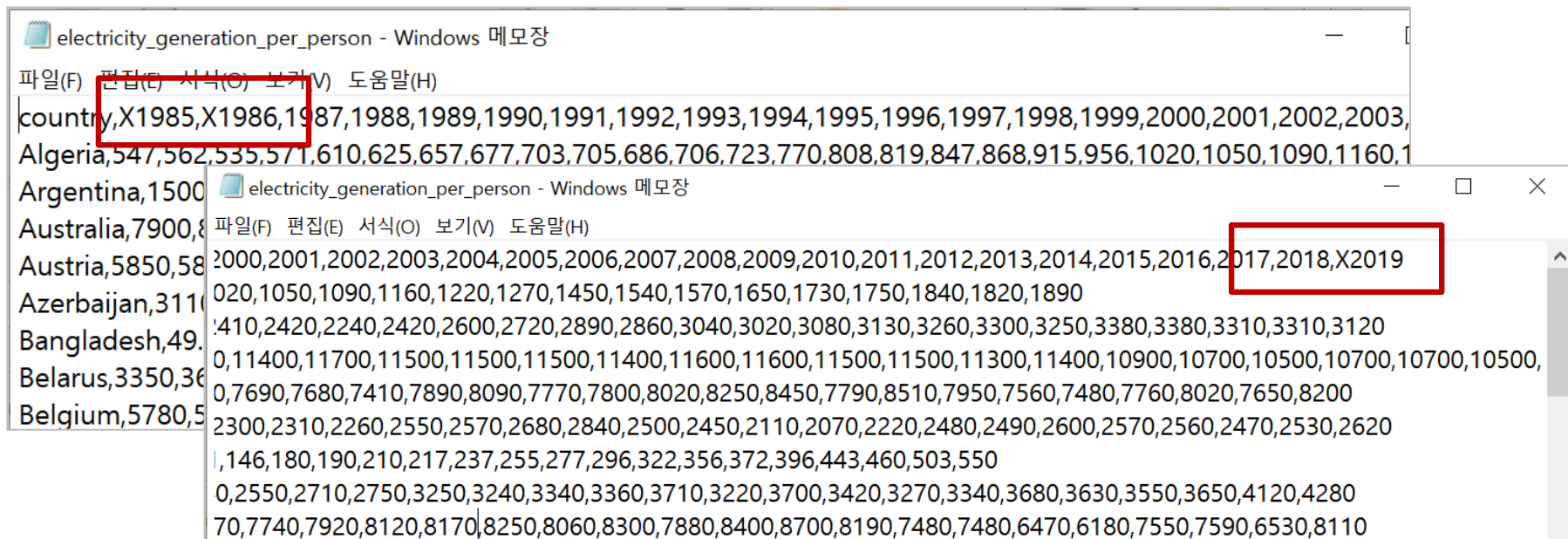
■ 1인당 전기 생산량 데이터 파일

(electricity_generation_per_person.csv)을 다운로드해 열어보자.

- 총 65개 국가에 대해 33년 동안(1985 ~ 2019)의 1인당 전기 생산량이 기록됨.

■ 열 이름이 되는 연도 앞에 문자 X가 의도치 않게 추가되어 있음.

- 텍스트 인코딩 문제로 인해 이런 일이 간혹 발생



5.4 데이터 가공의 실제 : 데이터 구조 변경

- 앞에서 학습한 `names`와 `substr` 함수를 이용해서 깔끔하게 정리해보자.

```
Console C:/RSources/
> elec_gen=read.csv("c:/rdata/5c/electricity_generation_per_person.csv",header = TRUE,sep=",")
> names(elec_gen)
 [1] "country" "x1985"   "x1986"   "x1987"   "x1988"   "x1989"
 [7] "x1990"   "x1991"   "x1992"   "x1993"   "x1994"   "x1995"
[13] "x1996"   "x1997"   "x1998"   "x1999"   "x2000"   "x2001"
[19] "x2002"   "x2003"   "x2004"   "x2005"   "x2006"   "x2007"
[25] "x2008"   "x2009"   "x2010"   "x2011"   "x2012"   "x2013"
[31] "x2014"   "x2015"   "x2016"   "x2017"   "x2018"   "x2019"
> names(elec_gen)=substr(names(elec_gen), 2, nchar(names(elec_gen)))
> names(elec_gen)
 [1] "ountry" "1985"    "1986"    "1987"    "1988"    "1989"    "1990"
 [8] "1991"    "1992"    "1993"    "1994"    "1995"    "1996"    "1997"
[15] "1998"    "1999"    "2000"    "2001"    "2002"    "2003"    "2004"
[22] "2005"    "2006"    "2007"    "2008"    "2009"    "2010"    "2011"
[29] "2012"    "2013"    "2014"    "2015"    "2016"    "2017"    "2018"
[36] "2019"
```

5.4 데이터 가공의 실제 : 데이터 구조 변경

- 같은 방법으로 전기 사용량 데이터를 파일 (electricity_use_per_person.csv)을 다운로드
- 연도 이름에서 불필요한 문자를 제거
- 총 138개국에 대해 56년 동안(1960~2014)의 1인당 전기 사용량이 기록되어 있다.



```
electricity_use_per_person - Windows 메모장
파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)
country,1960,1961,1962,1963,1964,1965,1966,1967,1968,1969,1970,1971,1972,1973,1974,1975,1976,1977,1978,
Albania,,,,,,,,,532,568,593,591,739,909,1070,1100,1070,1140,1120,1100,1070,1020,796,1420,1160,1010,1020,552
Algeria,,,,,,,,,134,143,159,171,196,220,233,279,315,330,363,406,417,441,466,485,455,482,522,532,532,560,548,55
Angola,,,,,,,,,92,101,114,136,139,60.3,58.3,56.3,58.6,60.7,58.6,60.3,63.3,52.2,60.6,58.5,56.8,55.2,53.6,53.2,54.5,53.5
Arg
electricity_use_per_person - Windows 메모장
파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)
Arn
Aus16,1997,1998,1999,2000,2001,2002,2003,2004,2005,2006,2007,2008,2009,2010,2011,2012,2013,2014
Aus940,2210,2120,2530,2310
Aze
Bah
1460,2770,2730,2880,2930,3000,2970,3070
10,9290,9780,10000,10200,10600,10800,10400,10600,10500,10500,11000,10700,10800,10700,10600,10400,10200,10100
0,6620,6710,6970,7080,7330,7430,7700,7810,7980,8240,8210,8230,7940,8380,8430,8550,8510,8360
```

5.4 데이터 가공의 실제 : 데이터 구조 변경

- 앞에서 학습한 `names`와 `substr` 함수를 이용해서 깔끔하게 정리해보자.

```
Console C:/RSources/
> elec_use=read.csv("c:/rdata/5c/electricity_use_per_person.csv",header = TRUE,sep=",")
> head(elec_use,3)
  country x1960 x1961 x1962 x1963 x1964 x1965 x1966 x1967 x1968 x1969
1 Albania   NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
2 Algeria   NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
3 Angola    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
x1970 x1971 x1972 x1973
1 532 568 593
2 134 143 159
3  92 101 114

Console C:/RSources/
> elec_use=read.csv("c:/rdata/5c/electricity_use_per_person.csv",header = TRUE,sep=",")
> names(elec_use)[2:56]=substr(names(elec_use)[2:56], 2, nchar(names(elec_use)[2:56]))
> head(elec_use,3)
  country 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973
1 Albania  NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA  532  568  593
2 Algeria  NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA  134  143  159
3 Angola   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   92  101  114
  1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985
1  591  739 909.0 1070.0 1100.0 1070.0 1140.0 1120.0 1100.0 1070.0 1020.0 796.0
2  171  196 220.0  233.0  279.0  315.0  330.0  363.0  406.0  417.0  441.0 466.0
3  136  139  60.3   58.3   56.3   58.6   60.7   58.6   60.3   63.3   52.2  60.6
  1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998
```

5.4 데이터 가공의 실제 : 데이터 구조 변경

- View(elec_gen), View(elec_use) 명령어로 data 확인

The screenshot displays two RStudio windows. The top window shows the 'elec_use' data frame with columns for country and years from 1960 to 1966. The bottom window shows the 'elec_gen' data frame with columns for country and years from 1985 to 1992. The console at the bottom shows the command `> View(elec_gen)` being executed.

elec_use Data Frame:

	country	X1960	X1961	X1962	X1963	X1964	X1965	X1966	X
1	Albania								
2	Algeria								
3	Angola								
4	Argentina								
5	Armenia								
6	Australia								
7	Austria								
8	Azerbaijan								

elec_gen Data Frame:

	country	1985	1986	1987	1988	1989	1990	1991	1992	1993
1	Algeria	547.0	562.0	535.0	571.0	610.0	625.0	657.0	677.0	
2	Argentina	1500.0	1600.0	1670.0	1660.0	1580.0	1560.0	1630.0	1670.0	
3	Australia	7900.0	8140.0	8390.0	8710.0	9060.0	9180.0	9190.0	9270.0	
4	Austria	5850.0	5860.0	6610.0	6400.0	6530.0	6530.0	6620.0	6540.0	
5	Azerbaijan	3110.0	3180.0	3320.0	3360.0	3270.0	3200.0	3190.0	2640.0	
6	Bangladesh	49.9	51.5	58.4	66.6	70.7	74.9	78.3	82.4	
7	Belarus	3350.0	3640.0	3740.0	3780.0	3800.0	3890.0	3810.0	3700.0	
8	Belgium	5780.0	5910.0	6370.0	6560.0	6770.0	7090.0	7170.0	7170.0	
9	Brazil	1120.0	1160.0	1110.0	1500.0	1500.0	1500.0	1550.0	1570.0	

Showing 1 to 9 of 77 entries, 36 total columns

Console C:/RSources/

```
> View(elec_gen)
> |
```

5.4 데이터 가공의 실제 : 데이터 구조 변경

■ 두 개의 데이터 프레임을 병합(1)

- 이제까지 보왔던 일반 데이터 프레임의 구성
 - 속성 하나를 열 하나에 배치하고
 - 행 하나가 샘플 한 개를 기록한 구성과 다름.
 - 각 측정 연도 값이 열의 이름이 되어 한 행에 여러 해의 데이터가 기록됨
- 국가 명, 연도, 전기 생산량과 소비량을 각각 하나의 열에 대응시키는 변형이 필요
- gather 함수 이용

5.4 데이터 가공의 실제 : 데이터 구조 변경

■ 두 개의 데이터 프레임을 병합(2)

- 새로 만들어질 데이터 프레임에서 구분자가 될 year를 key에, 측정값의 속성 이름을 value에 지정하면 gather 함수는 데이터 프레임을 다음과 같이 재구성한다.(원본 data 다름)

Console C:/RSources/

```
> elec_gen = read.csv("C:/rdata/5c/electricity_generation_per_person.csv", header = TRUE, sep = ",")
> names(elec_gen)[2:36] = substr(names(elec_gen)[2:36], 2, nchar(names(elec_gen)[2:36]))
> elec_use = read.csv("C:/rdata/5c/electricity_use_per_person.csv", header = TRUE, sep = ",")
> names(elec_use)[2:56] = substr(names(elec_use)[2:56], 2, nchar(names(elec_use)[2:56]))
> elec_gen_df = gather(elec_gen, -country, key = "year", value = "ElectricityGeneration")
> elec_use_df = gather(elec_use, -country, key = "year", value = "ElectricityUse")
>
```

	country	year	ElectricityGeneration
1	Algeria	1985	547.0
2	Argentina	1985	1500.0
3	Australia	1985	7900.0
4	Austria	1985	5850.0
5	Azerbaijan	1985	3110.0
6	Banqladesh	1985	49.9

Showing 1 to 6 of 2,695 entries, 3 total columns

	country	year	ElectricityUse
5658	Zimbabwe	2000	898.0
5659	Albania	2001	1350.0
5660	Algeria	2001	709.0
5661	Angola	2001	82.6
5662	Argentina	2001	2120.0
5663	Armenia	2001	1270.0

Showing 5,657 to 5,663 of 7,590 entries, 3 total columns

5.4 데이터 가공의 실제 : 데이터 구조 변경

■ 두 개의 데이터 프레임을 병합(2)

■ NA 제거

Console C:/RSources/ ↗

```
> elec_use_narm=na.omit(elec_use)
> elec_use_narm_df = gather(elec_use_narm, -country, key = "year", value = "ElectricityUse")
>
> view(elec_use_narm_df)
```

	country	year	ElectricityUse
1	Albania	1960	NA
2	Algeria	1960	NA
3	Angola	1960	NA
4	Argentina	1960	NA
5	Armenia	1960	NA
6	Australia	1960	1830
7	Austria	1960	1810

Showing 1 to 7 of 7,590 entries, 3 total columns



	country	year	ElectricityUse
1	Australia	1960	1830.0
2	Austria	1960	1810.0
3	Belgium	1960	1580.0
4	Canada	1960	5630.0
5	Denmark	1960	1090.0
6	Finland	1960	1870.0
7	France	1960	1460.0

Showing 1 to 7 of 1,375 entries, 3 total columns

5.4 데이터 가공의 실제 : 데이터 구조 변경

■ 두 개의 데이터 프레임을 병합(3)

- 재구성된 데이터 프레임을 merge 함수를 이용해 하나의 데이터 프레임으로 병합

```
Console C:/RSources/
> elec_use_narm=na.omit(elec_use)
> elec_use_narm_df = gather(elec_use_narm, -country, key = "year", value = "ElectricityUse")
>
> View(elec_use_narm_df)
> elec_gen_narm=na.omit(elec_gen)
> elec_gen_narm_df = gather(elec_gen_narm, -country, key = "year", value = "ElectricityGeneration")
> elec_gen_use_narm = merge(elec_gen_narm_df, elec_use_narm_df)
> View(elec_gen_use_narm)
```

	country	year	ElectricityGeneration	ElectricityUse
1	Australia	1985	7900	7010
2	Australia	1986	8140	7310
3	Australia	1987	8390	7530
4	Australia	1988	8710	7800
5	Australia	1989	9060	8130
6	Australia	1990	9180	8530

Showing 1 to 7 of 750 entries, 4 total columns

데이터 가공의 의의

- 데이터 영역을 추출하는 작업은 물론이고, 데이터의 직관적이고 쉬운 검색을 위한 여러 작업 포함
- 그 어떤 분석 기법보다도 오래된 데이터 과학의 기본 기술. 크고 복잡한 데이터 덩어리를 지속적으로 가공하며 의미가 파악되고 원하는 형태로 정리되어 가는 것을 보면 일종의 성취감도 맛볼 수 있음.



그림 5-10 데이터 가공이 갖는 의미

데이터 가공의 의의

- 데이터 변형은 데이터를 바라보는 관찰자의 생각과 관점을 바꾸는 것
- **데이터 가공은 기계적이거나 단순한 작업이 결코 아니며**, 데이터에 내포된 의미가 잘 드러나도록 불필요한 부분을 잘라내고, 신뢰성과 일관성을 부여하기 위한 끊임없는 사고력을 요구하는 작업
- 데이터에 대한 이해와 데이터의 가공은 같이 진행되는 것



그림 5-11 데이터 가공 = 데이터 관찰의 관점 변화

1. 데이터 가공이란?
2. 베이스 R을 이용한 데이터 가공
3. Dplyr 라이브러리를 이용한 데이터 가공
4. 데이터 가공 실습

Thank you

