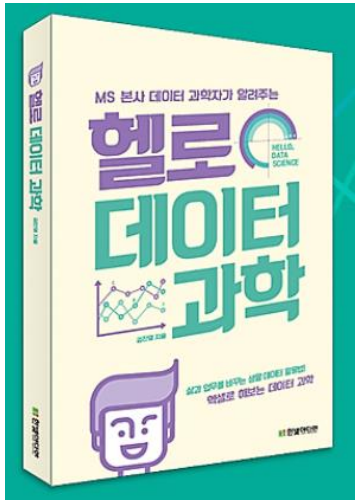
[1]Data Science 개요

- ✓ Data Science 정의
- ✓ Data Science 사례
- ✓ Data Science Process
- ✓ 분석도구 R

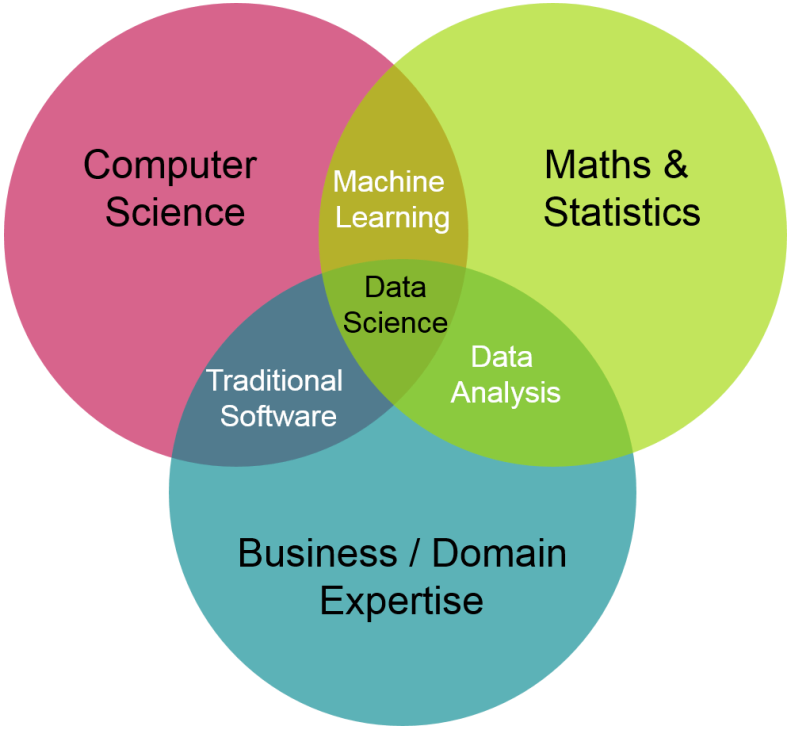
https://dataninja.me/ipds-kr/slides-ppt/



http://www.hellodatascience.com/

Data Science ?

- 정의
  - 컴퓨터 도구를 효율적으로 이용하고, 적절한 통계학 방법을 사용하여 실제적인 문제에 답을 내리는 활동



사례 1: (국가 경제)
국가별 경제 수준과 의료 수준의 상관관계 분석

- 데이터정의**(data definition)**
  - 국가별 경제 수준 : 일인당 **GDP**
  - 의료 수준 : 평균 기대 수명
- 데이터 취득(data acquisition) : 갭마인더(Gapminder)
  - 나라별, 연도별 평균기대수명, 일인당 GDP, 인구수 수집데이터
- 데이터 가공**(data processing)**
  - 각 기관으로부터 수집된 데이터를 통합하고 필요한 변수들을 테이블로 정리

- country: 142개의 다른 값(levels)을 가진 인자(factor) 변수
- continent: 다섯 가지 값을 가진 인자 변수
- year: 숫자형의 연도 변수. 1952년과 2007년 사이 5년 간격
- lifeExp: 이 해에 태어난 이들의 평균 기대 수명
- pop: 인구
- gdpPercap: 일인당 국민소득(GDP per capita)

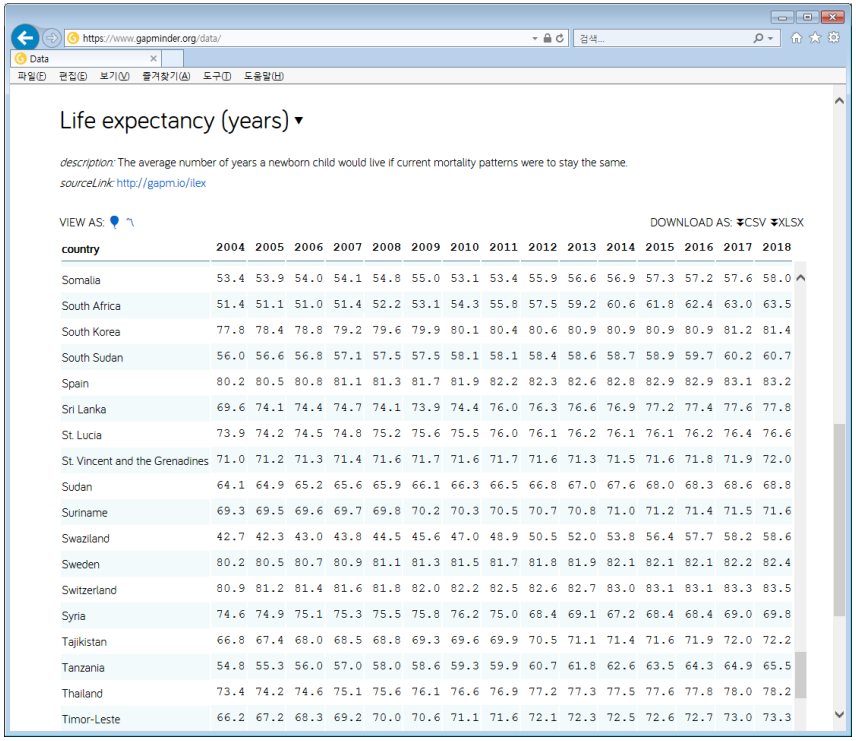
갭마인더 데이터는 위의 변수가 각 나라와 연도별로 수집된 것을 보여준다.

no.	country	continent	year	lifeExp	pop	gdpPercap
1	Afghanistan	Asia	1952	28.801	8425333	779.4453
2	Afghanistan	Asia	1957	30.332	9240934	820.8530
3	Afghanistan	Asia	1962	31.997	10267083	853.1007
...						
1699	Zimbabwe	Africa	1982	60.363	7636524	788.8550
1700	Zimbabwe	Africa	1987	62.351	9216418	706.1573
1701	Zimbabwe	Africa	1992	60.377	10704340	693.4208

사례 1: (국가 경제)
국가별 경제 수준과 의료 수준의 상관관계 분석

Gapminder World 지표 (http://www.gapminder.org/data)

- 기대 수명(Life Expectation)
- 1 인당 소득(income per person)



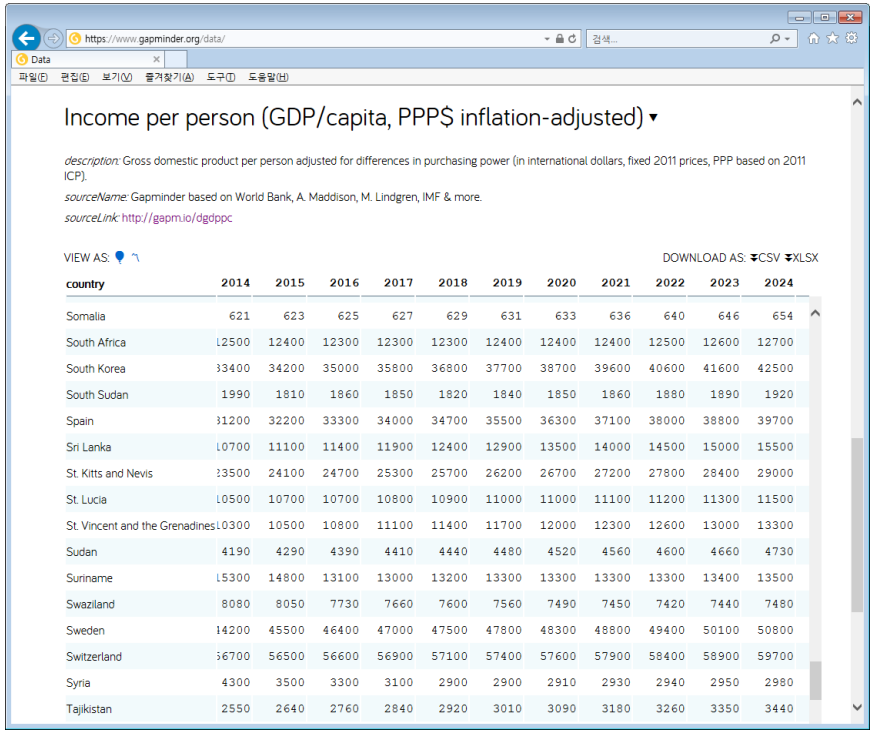
https://www.gapminder.org/data/

Life expectancy (years)

description: The average number of years a newborn child would live if current mortality patterns were to stay the same.
sourceLink: http://gapm.io/ilex

VIEW AS: DOWNLOAD AS: CSV XLSX

country	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Somalia	53.4	53.9	54.0	54.1	54.8	55.0	53.1	53.4	55.9	56.6	56.9	57.3	57.2	57.6	58.0
South Africa	51.4	51.1	51.0	51.4	52.2	53.1	54.3	55.8	57.5	59.2	60.6	61.8	62.4	63.0	63.5
South Korea	77.8	78.4	78.8	79.2	79.6	79.9	80.1	80.4	80.6	80.9	80.9	80.9	80.9	81.2	81.4
South Sudan	56.0	56.6	56.8	57.1	57.5	57.5	58.1	58.1	58.4	58.6	58.7	58.9	59.7	60.2	60.7
Spain	80.2	80.5	80.8	81.1	81.3	81.7	81.9	82.2	82.3	82.6	82.8	82.9	82.9	83.1	83.2
Sri Lanka	69.6	74.1	74.4	74.7	74.1	73.9	74.4	76.0	76.3	76.6	76.9	77.2	77.4	77.6	77.8
St. Lucia	73.9	74.2	74.5	74.8	75.2	75.6	75.5	76.0	76.1	76.2	76.1	76.1	76.2	76.4	76.6
St. Vincent and the Grenadines	71.0	71.2	71.3	71.4	71.6	71.7	71.6	71.7	71.6	71.3	71.5	71.6	71.8	71.9	72.0
Sudan	64.1	64.9	65.2	65.6	65.9	66.1	66.3	66.5	66.8	67.0	67.6	68.0	68.3	68.6	68.8
Suriname	69.3	69.5	69.6	69.7	69.8	70.2	70.3	70.5	70.7	70.8	71.0	71.2	71.4	71.5	71.6
Swaziland	42.7	42.3	43.0	43.8	44.5	45.6	47.0	48.9	50.5	52.0	53.8	56.4	57.7	58.2	58.6
Sweden	80.2	80.5	80.7	80.9	81.1	81.3	81.5	81.7	81.8	81.9	82.1	82.1	82.1	82.2	82.4
Switzerland	80.9	81.2	81.4	81.6	81.8	82.0	82.2	82.5	82.6	82.7	83.0	83.1	83.1	83.3	83.5
Syria	74.6	74.9	75.1	75.3	75.5	75.8	76.2	75.0	68.4	69.1	67.2	68.4	68.4	69.0	69.8
Tajikistan	66.8	67.4	68.0	68.5	68.8	69.3	69.6	69.9	70.5	71.1	71.4	71.6	71.9	72.0	72.2
Tanzania	54.8	55.3	56.0	57.0	58.0	58.6	59.3	59.9	60.7	61.8	62.6	63.5	64.3	64.9	65.5
Thailand	73.4	74.2	74.6	75.1	75.6	76.1	76.6	76.9	77.2	77.3	77.5	77.6	77.8	78.0	78.2
Timor-Leste	66.2	67.2	68.3	69.2	70.0	70.6	71.1	71.6	72.1	72.3	72.5	72.6	72.7	73.0	73.3



https://www.gapminder.org/data/

Income per person (GDP/capita, PPP$ inflation-adjusted)

description: Gross domestic product per person adjusted for differences in purchasing power (in international dollars, fixed 2011 prices, PPP based on 2011 ICP).
sourceName: Gapminder based on World Bank, A. Maddison, M. Lindgren, IMF & more.
sourceLink: http://gapm.io/dgdppc

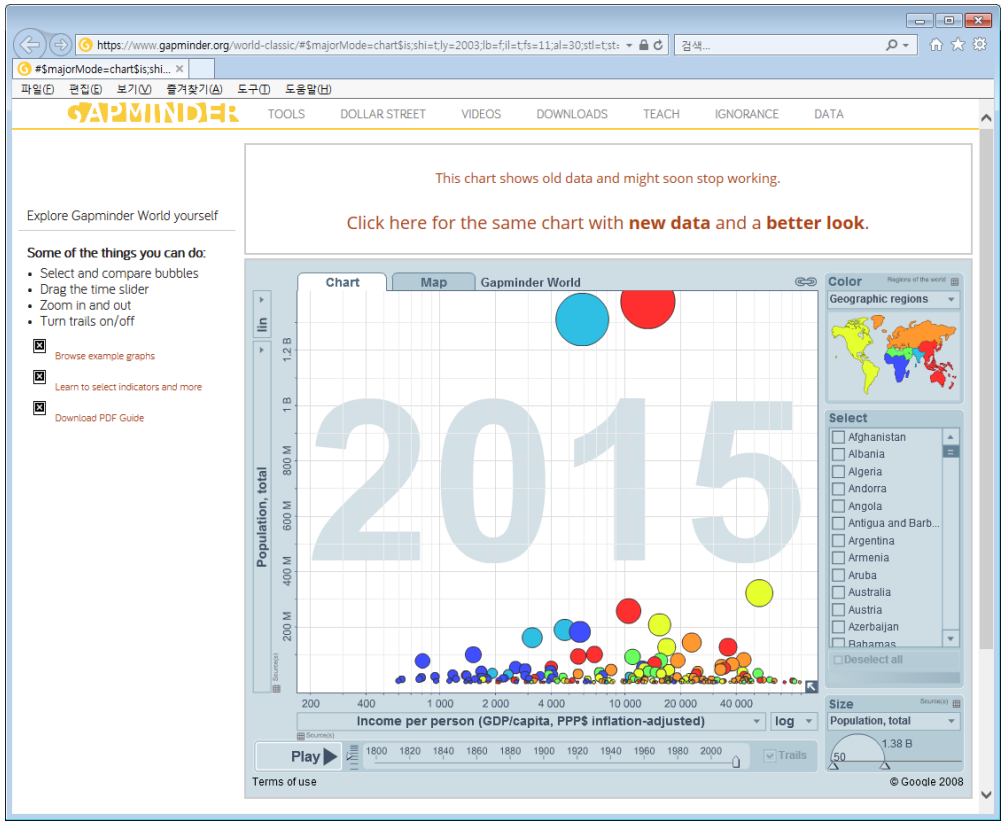
VIEW AS: DOWNLOAD AS: CSV XLSX

country	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Somalia	621	623	625	627	629	631	633	636	640	646	654
South Africa	12500	12400	12300	12300	12300	12400	12400	12400	12500	12600	12700
South Korea	33400	34200	35000	35800	36800	37700	38700	39600	40600	41600	42500
South Sudan	1990	1810	1860	1850	1820	1840	1850	1860	1880	1890	1920
Spain	31200	32200	33300	34000	34700	35500	36300	37100	38000	38800	39700
Sri Lanka	10700	11100	11400	11900	12400	12900	13500	14000	14500	15000	15500
St. Kitts and Nevis	23500	24100	24700	25300	25700	26200	26700	27200	27800	28400	29000
St. Lucia	10500	10700	10700	10800	10900	11000	11000	11100	11200	11300	11500
St. Vincent and the Grenadines	10300	10500	10800	11100	11400	11700	12000	12300	12600	13000	13300
Sudan	4190	4290	4390	4410	4440	4480	4520	4560	4600	4660	4730
Suriname	15300	14800	13100	13000	13200	13300	13300	13300	13300	13400	13500
Swaziland	8080	8050	7730	7660	7600	7560	7490	7450	7420	7440	7480
Sweden	44200	45500	46400	47000	47500	47800	48300	48800	49400	50100	50800
Switzerland	56700	56500	56600	56900	57100	57400	57600	57900	58400	58900	59700
Syria	4300	3500	3300	3100	2900	2900	2910	2930	2940	2950	2980
Tajikistan	2550	2640	2760	2840	2920	3010	3090	3180	3260	3350	3440

사례 1: (국가 경제)
국가별 경제 수준과 의료 수준의 상관관계 분석

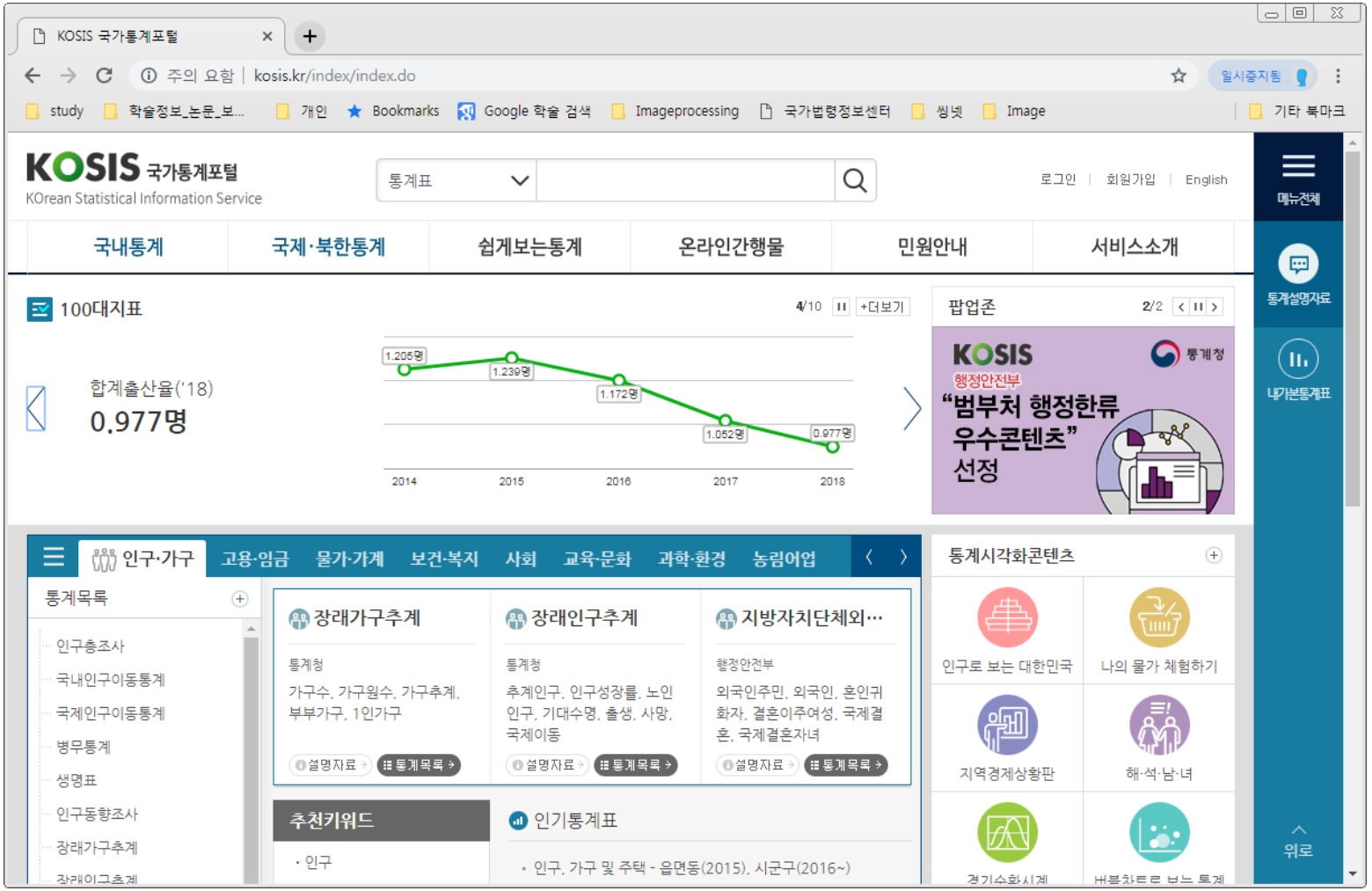
Gapminder World 지표를 이용한 시각화 예

- 1인당 소득수준과 인구수



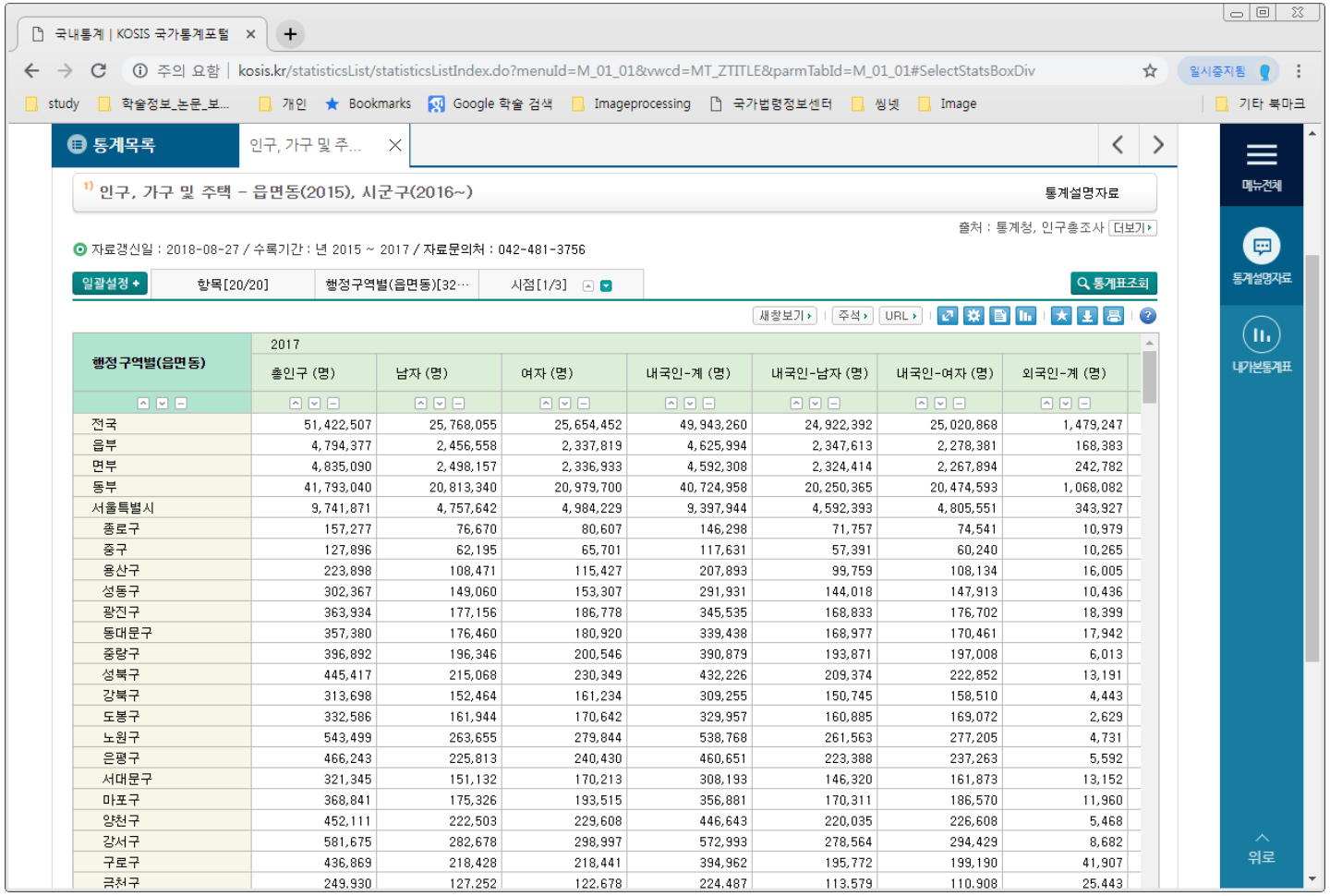
국내통계자료

국가통계포털 (http://kosis.kr/index/index.do)



국내통계자료

통계자료 : 국내 총 인구 수



행정구역별(읍면동)	2017						
	총인구 (명)	남자 (명)	여자 (명)	내국인-계 (명)	내국인-남자 (명)	내국인-여자 (명)	외국인-계 (명)
전국	51,422,507	25,768,055	25,654,452	49,943,260	24,922,392	25,020,868	1,479,247
읍부	4,794,377	2,456,558	2,337,819	4,625,994	2,347,613	2,278,381	168,383
면부	4,835,090	2,498,157	2,336,933	4,592,308	2,324,414	2,267,894	242,782
동부	41,793,040	20,813,340	20,979,700	40,724,958	20,250,365	20,474,593	1,068,082
서울특별시	9,741,871	4,757,642	4,984,229	9,397,944	4,592,393	4,805,551	343,927
종로구	157,277	76,670	80,607	146,298	71,757	74,541	10,979
중구	127,896	62,195	65,701	117,631	57,391	60,240	10,265
용산구	223,898	108,471	115,427	207,893	99,759	108,134	16,005
성동구	302,367	149,060	153,307	291,931	144,018	147,913	10,436
광진구	363,934	177,156	186,778	345,535	168,833	176,702	18,399
동대문구	357,380	176,460	180,920	339,438	168,977	170,461	17,942
중랑구	396,892	196,346	200,546	390,879	193,871	197,008	6,013
성북구	445,417	215,068	230,349	432,226	209,374	222,852	13,191
강북구	313,698	152,464	161,234	309,255	150,745	158,510	4,443
도봉구	332,586	161,944	170,642	329,957	160,885	169,072	2,629
노원구	543,499	263,655	279,844	538,768	261,563	277,205	4,731
은평구	466,243	225,813	240,430	460,651	223,388	237,263	5,592
서대문구	321,345	151,132	170,213	308,193	146,320	161,873	13,152
마포구	368,841	175,326	193,515	356,881	170,311	186,570	11,960
양천구	452,111	222,503	229,608	446,643	220,035	226,608	5,468
강서구	581,675	282,678	298,997	572,993	278,564	294,429	8,682
구로구	436,869	218,428	218,441	394,962	195,772	199,190	41,907
금천구	249,930	127,252	122,678	224,487	113,579	110,908	25,443

사례2: (부동산 경제) 주택가격 예측

- 보스턴 주택 데이터세트(Boston house-price dataset) [Belsley, et al. (1980)]
- 변수 14개 , 506개 데이터

  - crim: 범죄발생률
  - zn: 주거지 중 25000 ft^2 이상 크기의 대형주택이 차지하는 비율
  - indus: 소매상 이외의 상업지구의 면적 비율
  - chas: 찰스강과 접한 지역은 1, 아니면 0인 더미변수(dummy variable)
  - nox: 산화질소 오염도
  - rm: 주거지당 평균 방 개수
  - age: 소유자 주거지(전세 혹은 월세가 아닌) 중 1940년 이전에 지어진 집들의 비율
  - dis: 보스턴의 5대 고용 중심으로부터의 가중 평균 거리
  - rad: 도시 순환 고속도로에의 접근 용이 지수
  - tax: 만 달러당 주택 재산세율
  - ptratio: 학생-선생 비율
  - black: 흑인 인구 비율(Bk)이 지역 평균인 0.63과 다른 정도의 제곱, $1000(Bk - 0.63)^2$
  - lstat: 저소득 주민들의 비율 퍼센트
  - medv: 소유자 주거지(비 전세/월세) 주택 가격

사례2: (부동산 경제) 주택가격 예측

- 수치형 값을 예측하기 위해 회귀분석 사용
- 선형회귀분석(linear regression)
  - 주택가격에 영향을 주는 각 변수에 대한 가중치 곱의 합으로 예측
  - 알려지지 않은 가중치 값(β)을 추정(model fitting)

$$\text{주택 가격} \approx \beta_0 + \beta_{\text{crim}}\ x_{\text{crim}} + \ldots + \beta_{\text{lstat}}\ x_{\text{lstat}}$$

$$\text{주택 가격} = \beta_0 + \beta_{\text{crim}}\ x_{\text{crim}} + \ldots + \beta_{\text{lstat}}\ x_{\text{lstat}} + \text{잡음}$$

Data Science Process

1. 문제 정의(problem definition)
   - 현상의 이해: 탐험적 데이터 분석(Exploratory Data Analysis)
     - 각종 통계값(statistics)의 계산, 데이터 시각화, 상관도(correlation) 분석
   - 현상의 일반화: 통계적 추론
   - 현상의 예측: 기계 학습

<문제 정의>
•문제의 목표는 무엇인가?
•문제의 범위는 정확히 어디까지인가?
•문제 해결의 성공 / 실패 기준은 무엇인가?
•문제 해결에 있어서의 제약조건은 무었인가? (시간과 비용 등)

2. 데이터 정의(data definition)

<데이터 문제정의>
•문제와 관련된 데이터에 포함되어야 하는 요인은 무엇인가?
•문제 해결에 필요한 데이터를 어떻게 수집할 수 있는가?
•데이터 처리 및 분석을 위한 최적의 방법과 도구는 무엇인가?
•최종 결과물은 어떤 형태로 누구에게 전달되어야 하는가?

Data Science Process

3. 실험 계획(design of experiment)
   - 데이터 직접 수집하는 경우에 필요
   - 표본화, 표본의 크기 결정
     - 개별항목의 속성값을 정확하게 측정
     - 관찰하려는 현상을 대표할만한 환경에서 수집
       - 관찰형 연구(Observational Study) : 자연 그대로 수집
       - 통제형 실험(Controlled Experiments) : 조건에 따라 수집
4. 데이터 취득(data acquisition)
   - 문제에 정의된 관심을 가지는 현상을 데이터로 만드는 과정
     - 기존 데이터셋 사용
     - 실험 계획에 따라 새로운 데이터 직접 수집
   - 일관성, 유연성, 무작위성(randomness)

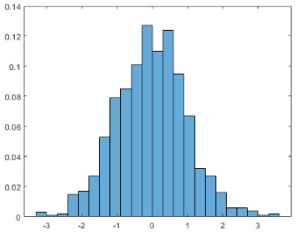
Data Science Process

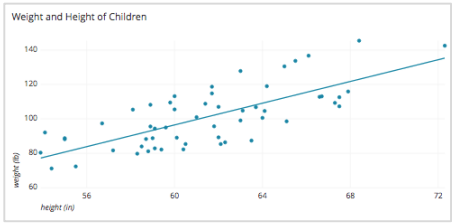
5. 데이터 가공(data processing, data wrangling)
   - 분석을 위한 표준 테이블 형태로 변환
     - 각 행은 개별 관찰 항목
     - 각 열은 개별 속성
     - 각 테이블에는 단일 유형의 데이터로 구성
     - 여러 테이블이 존재하는 경우 개별 테이블을 연결할 수 있는 공통된 속성 필요
   - 데이터의 품질 점검
     - 완전성, 정확성,일관성
   - R에서는 JSON, CSV, XML등 널리 사용되는 형식의 파일을 테이블 형태로 불러오는 라이브러리를 제공

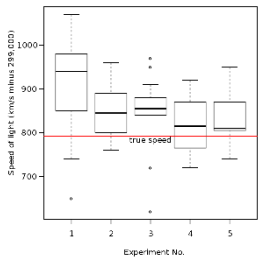
Data Science Process

6.탐색적 분석과 데이터 시각화
(exploratory data analysis(EDA), data visualization)

- 데이터의 분포 및 값을 검토
- 이상값(outlier), Missing value, 속성간의 관계(상관도) 등 파악
- 다양한 요약 통계값 (statistics), 시각화로 확인.
- 요약 통계 지표(summary statistics)
  - 데이터의 중심 : 평균(mean) 및 중앙값(median), 최 빈값(mode)
  - 데이터의 분산도 : 범위(range), 분산(variance)
  - 데이터 분포 : skewness
- 시각화(data visualization)
  - Histogram, scatter plot, box plot



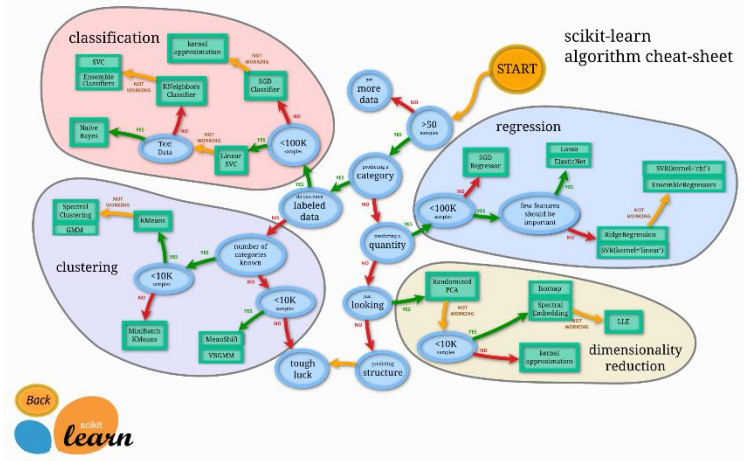




Data Science Process

7.모형화(modeling)

- 통계적 추론(statistical inference)
  - 표본을 바탕으로 모집단의 특성에 대한 결론을 유도
  - 통계이론을 바탕으로 한 현상을 일반화
- 기계학습(machine learning)
  - 데이터마이닝(data mining)
  - 지능적인 방법(기계학습 알고리즘)을 적용하여 데이터 모델링
  - 예측(Prediction), 분류(Classification), 군집(clustering), 연관규칙(Association Rule)



8. 분석 결과 정리(reporting)

Data Science Process

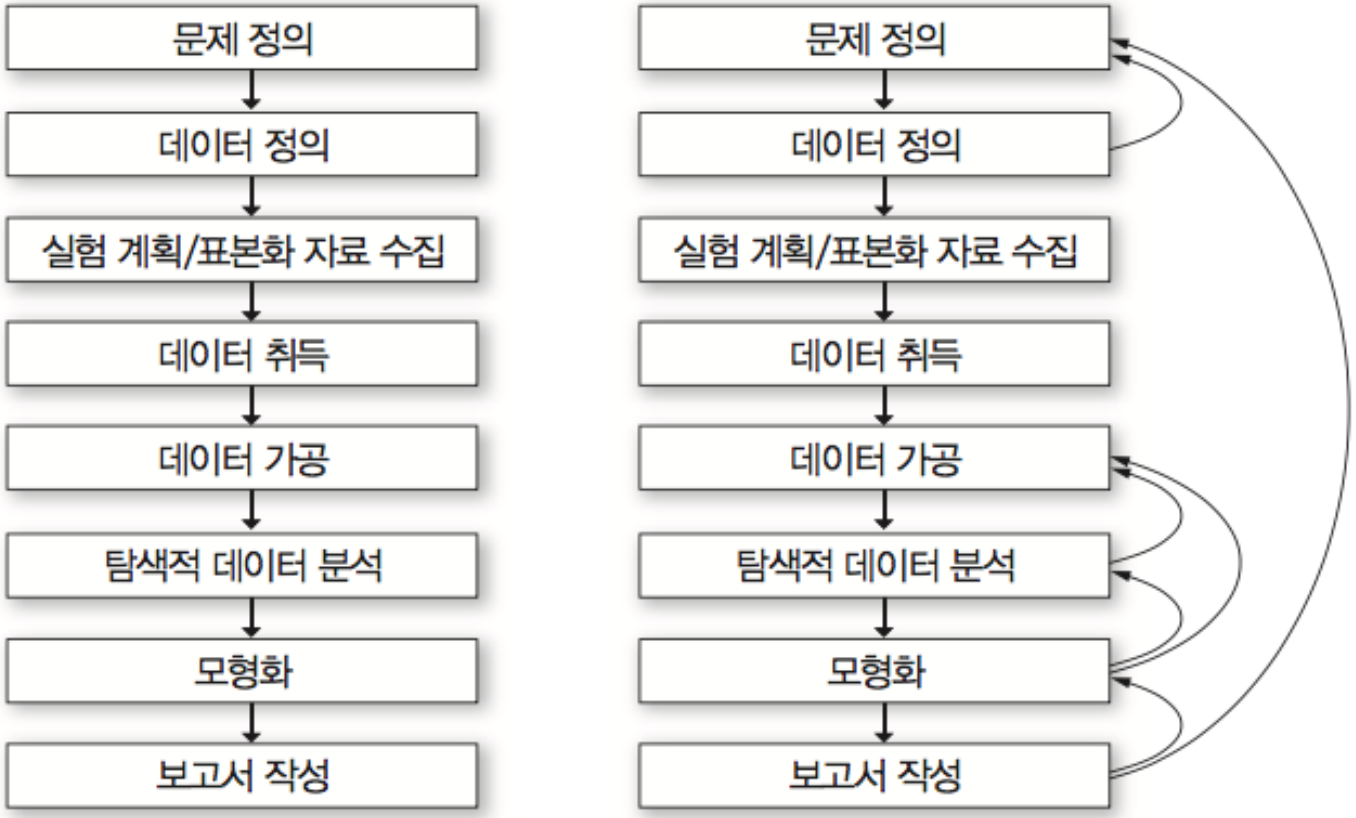


그림 1-3 데이터 분석 과정에 대한 이상적 관점(왼쪽)과 현실적 관점(오른쪽)

R ?

- 데이터 분석을 위한 통계 및 그래픽스를 지원하는 공개용 소프트웨어
- 1996년 뉴질랜드 Auckland 대학에서 Ross Ihaka와 Robert Gentleman이 개발
- CRAN(the Comprehensive R Archive Network)에서 제공 : http://cran.r-project.org
- 사용자 제작 패키지를 통하여 확장 가능
  - 핵심적인 패키지는 R과 함께 설치
  - 다양한 분야(통계, 머신러닝, 금융, 바이오인포메틱스, 그래픽스 등)의 패키지를 제공 : https://cran.r-project.org/web/packages/index.html

R 특징

- 오픈 소스 기반의 객체지향 언어
- 메모리 기반으로 동작하므로 데이터 처리 속도가 빠르며 하드웨어 메모리 크기가 처리 시간에 영향을 줌.
- 모든 플랫폼(Windows, MacOS, UNIX, Linux)에서 운영 가능.
- SAS나 SPSS 등 다른 통계분석 소프트웨어에서 플러그-인 형태 등으로 R의 스크립트 이용가능.
- 다른 언어로 작성된 프로그램을 통합하는 인터페이스를 제공(C, C++, C#, Fortran, Perl, Python, JAVA)