

01

빅데이터 분석도구 개요

BIG

DATA

1. 빅데이터 분석도구 개요



가. 빅데이터 분석도구의 필요성



매년 홍수처럼 불어나는
상당한 양의 정보들을
다루는 기술의 필요성 대두

1. 빅데이터 분석도구 개요



가. 빅데이터 분석도구의 필요성

빅데이터
5V

Volume

초대용량의 데이터 양

Variety

다양한 형태

Velocity

빠른 생성 속도

Veracity

정확성

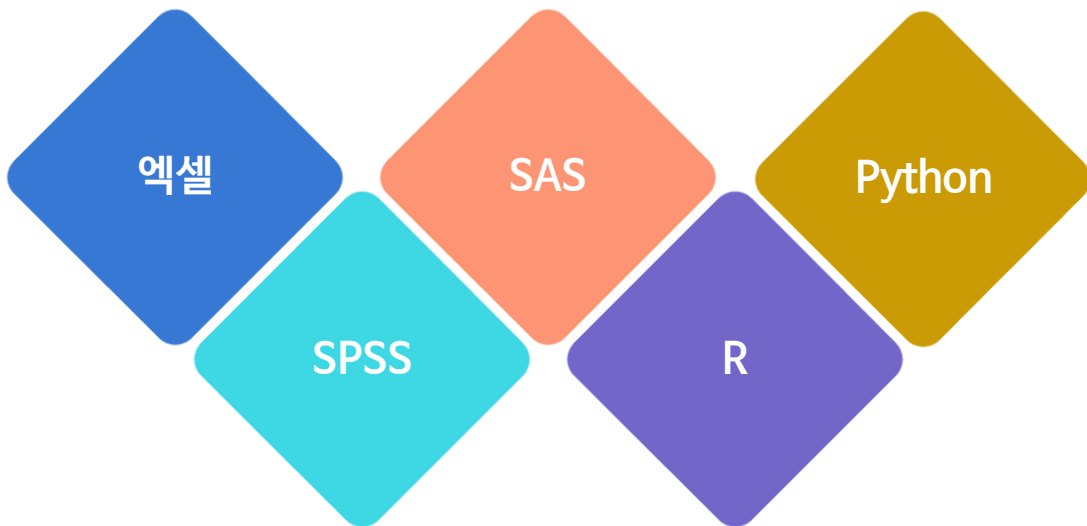
Value

가치

1. 빅데이터 분석도구 개요



가. 빅데이터 분석도구의 필요성



Value(가치) 창출을 위해 빅데이터 핸들링과 통계적 분석,
이를 뒷받침하는 통계적 분석 도구들이 필수적으로 요구

02

빅데이터 분석도구 및 기술 : 엑셀

BIG

DATA

2. 빅데이터 분석도구 및 기술 : 엑셀



가. 엑셀



마이크로소프트사에서 개발한
윈도 환경의 스프레드시트 프로그램

2. 빅데이터 분석도구 및 기술 : 엑셀



가. 엑셀



- 스프레드시트, 매크로, 그래픽, 데이터베이스 기능과 차트 및 문서 작성 등이 가능
- 현재 엑셀 2017 최신 버전 제공

2. 빅데이터 분석도구 및 기술 : 엑셀



나. 엑셀의 데이터 입력과 분석

1

데이터 입력 시 복잡한 명령어 없이 사용자가
직접 해당 셀에 원하는 데이터 입력 가능

2. 빅데이터 분석도구 및 기술 : 엑셀



나. 엑셀의 데이터 입력과 분석

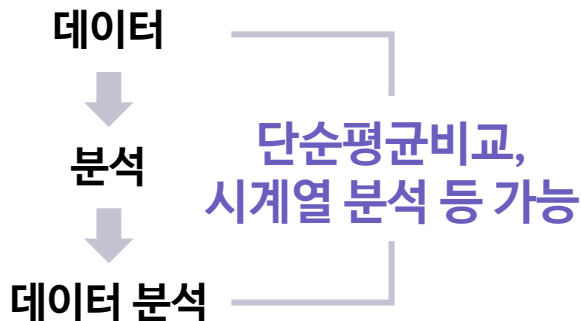
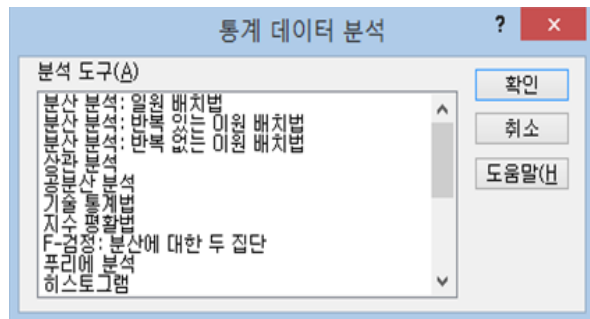
- 1 데이터 입력 시 복잡한 명령어 없이 사용자가 직접 해당 셀에 원하는 데이터 입력 가능
- 2 기존 데이터를 불러와 수정, 사용하는 방법이 있어 사용자의 상황에 따라 선택적 사용 가능

2. 빅데이터 분석도구 및 기술 : 엑셀



나. 엑셀의 데이터 입력과 분석

- 1 데이터 입력 시 복잡한 명령어 없이 사용자가 직접 해당 셀에 원하는 데이터 입력 가능
- 2 기존 데이터를 불러와 수정, 사용하는 방법이 있어 사용자의 상황에 따라 선택적 사용 가능
- 3 '데이터' 리본메뉴에서 제공되는 다양한 방법을 마우스 클릭으로 사용 가능



2. 빅데이터 분석도구 및 기술 : 엑셀



다. 엑셀의 데이터 분석 예제

Iris flower data set

위키피디아에서 무료로 제공되는 데이터세트를 불러온 뒤,
Iris setosa 종의 꽃받침의 가로와 세로 길이의 상관관계 분석

2. 빅데이터 분석도구 및 기술 : 엑셀



다. 엑셀의 데이터 분석 예제

Iris flower data set

Fisher's *Iris* Data

Sepal length ⇅	Sepal width ⇅	Petal length ⇅	Petal width ⇅	Species ⇅
5.1	3.5	1.4	0.2	<i>I. setosa</i>
4.9	3.0	1.4	0.2	<i>I. setosa</i>
4.7	3.2	1.3	0.2	<i>I. setosa</i>
4.6	3.4	1.4	0.3	<i>I. setosa</i>
5.0	3.6	1.4	0.1	<i>I. setosa</i>
5.4	3.9	1.7	0.4	<i>I. setosa</i>
4.6	3.4	1.4	0.3	<i>I. setosa</i>
5.0	3.4	1.5	0.2	<i>I. setosa</i>
4.4	2.9	1.4	0.2	<i>I. setosa</i>
4.9	3.1	1.5	0.1	<i>I. setosa</i>

데이터

〈출처 : http://en.wikipedia.org/wiki/Iris_flower_data_set〉

2. 빅데이터 분석도구 및 기술 : 엑셀



다. 엑셀의 데이터 분석 예제

Iris flower data set

Fisher's Iris Data

Sepal length	Species	Petal length	Petal width
5.1	<i>I. setosa</i>	1.4	0.2
4.9	<i>I. setosa</i>	1.4	0.2
4.7	<i>I. setosa</i>	1.3	0.2
4.6	<i>I. setosa</i>	1.3	0.2
5.0	<i>I. setosa</i>	1.4	0.2
5.4	<i>I. setosa</i>	1.5	0.2
4.6	<i>I. setosa</i>	1.4	0.3
5.0	<i>I. setosa</i>	1.5	0.3
4.4	<i>I. setosa</i>	1.4	0.1
4.9	<i>I. setosa</i>	1.5	0.1

Diagram illustrating the data analysis process:

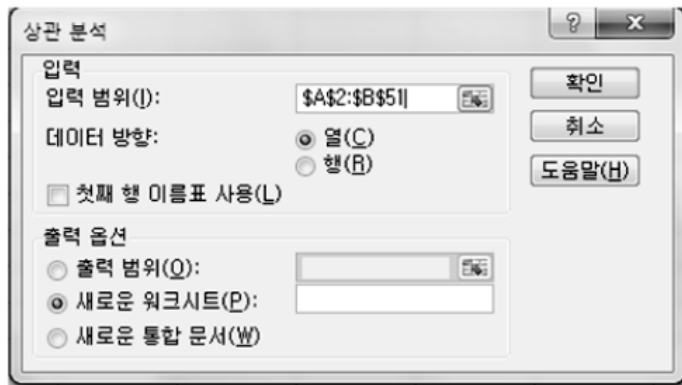
- 데이터 (Data)
- 분석 (Analysis)
- 데이터 분석 (Data Analysis)
- 상관분석 클릭 (Click Correlation Analysis)

〈출처 : http://en.wikipedia.org/wiki/Iris_flower_data_set〉

2. 빅데이터 분석도구 및 기술 : 엑셀

다. 엑셀의 데이터 분석 예제

Iris flower data set



	A	B	C
1		Column 1	Column 2
2	Column 1	1	
3	Column 2	0.742547	1

연속된 데이터 범위 입력 → 결과 출력될 공간에 대한 정보 지정 → 꽃받침의 가로와 세로 길이의 상관계수는 0.742547, 즉 양의 상관관계

〈출처 : http://en.wikipedia.org/wiki/Iris_flower_data_set〉

03

빅데이터 분석도구 및 기술 : SPSS

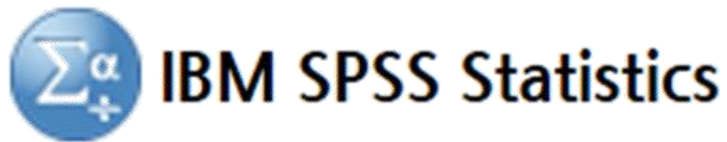
BIG

DATA

3. 빅데이터 분석도구 및 기술 : SPSS



가. SPSS

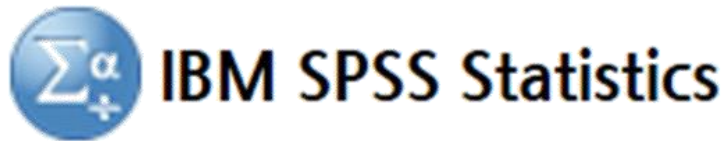


사회과학 자료분석을 위해
고안된 통계분석 전용 프로그램

3. 빅데이터 분석도구 및 기술 : SPSS



가. SPSS



- 1969년 사회과학 분야의 데이터 분석을 위해 시카고대학에서 컴퓨터 프로그램의 모음집으로 출발
- 2009년 IBM사에서 인수하면서 명칭이 IBM SPSS Statistics으로 개칭
- 현재 IBM SPSS Statistics 23 최신 버전 제공

3. 빅데이터 분석도구 및 기술 : SPSS



나. SPSS의 특징

1

비즈니스 사용자나 분석가 또는 통계프로그래머에게
적합한 프로그램

3. 빅데이터 분석도구 및 기술 : SPSS



나. SPSS의 특징

- 1 비즈니스 사용자나 분석가 또는 통계프로그래머에게 적합한 프로그램
- 2 사용이 간편하여 비전문가일지라도 단기간에 사용법 습득 가능

3. 빅데이터 분석도구 및 기술 : SPSS



나. SPSS의 특징

- 1 비즈니스 사용자나 분석가 또는 통계프로그래머에게 적합한 프로그램
- 2 사용이 간편하여 비전문가일지라도 단기간에 사용법 습득 가능
- 3 사용자가 속한 기관에 따라 교육기관, 의학연구기관, 공공기관, 병원, 일반기관 등으로 분류된 프로그램 제공

3. 빅데이터 분석도구 및 기술 : SPSS

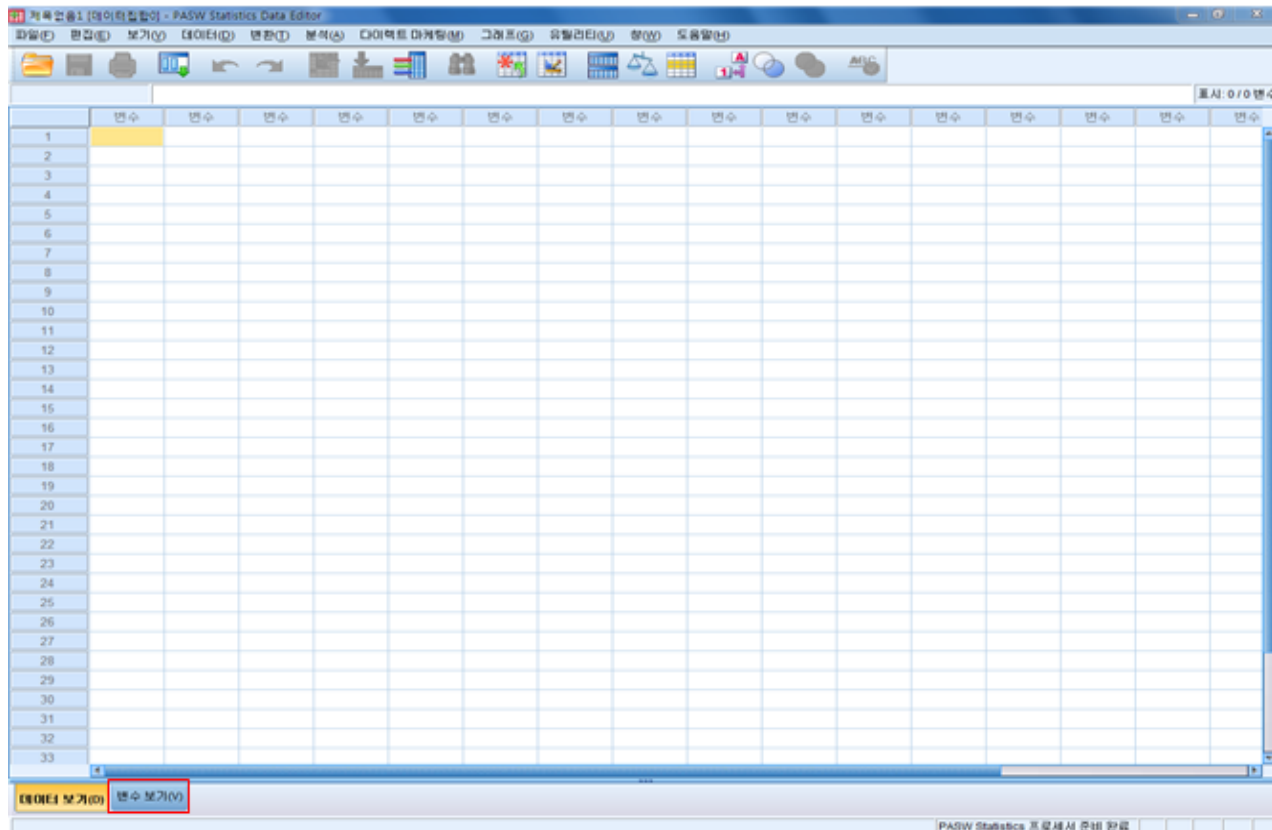


나. SPSS의 특징

- 1 비즈니스 사용자나 분석가 또는 통계프로그래머에게 적합한 프로그램
- 2 사용이 간편하여 비전문가일지라도 단기간에 사용법 습득 가능
- 3 사용자가 속한 기관에 따라 교육기관, 의학연구기관, 공공기관, 병원, 일반기관 등으로 분류된 프로그램 제공
- 4 분석 수준에 따라 3가지 제품 (Standard, Professional, Premium)으로 제공하여 사용자의 편의 도모

3. 빅데이터 분석도구 및 기술 : SPSS

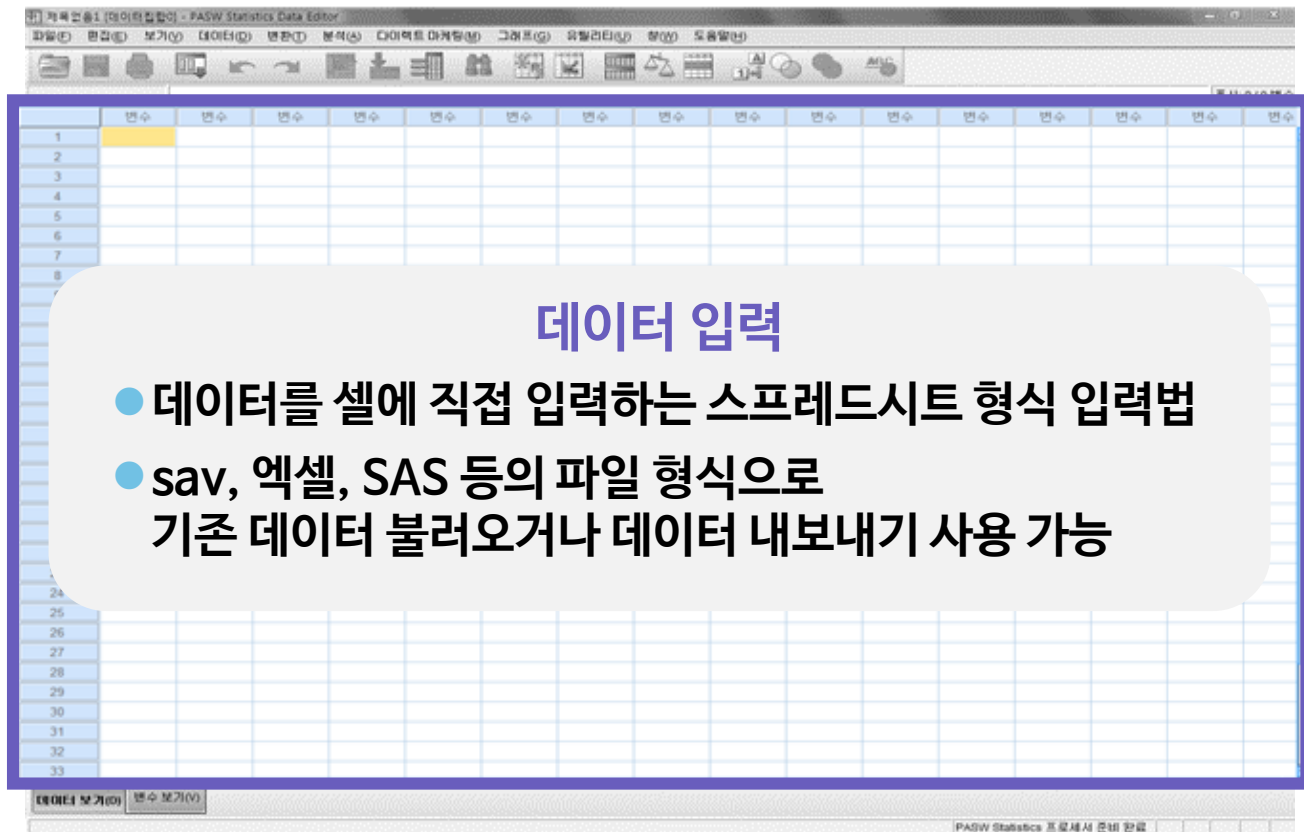
다. SPSS의 데이터 입력과 분석



3. 빅데이터 분석도구 및 기술 : SPSS



다. SPSS의 데이터 입력과 분석

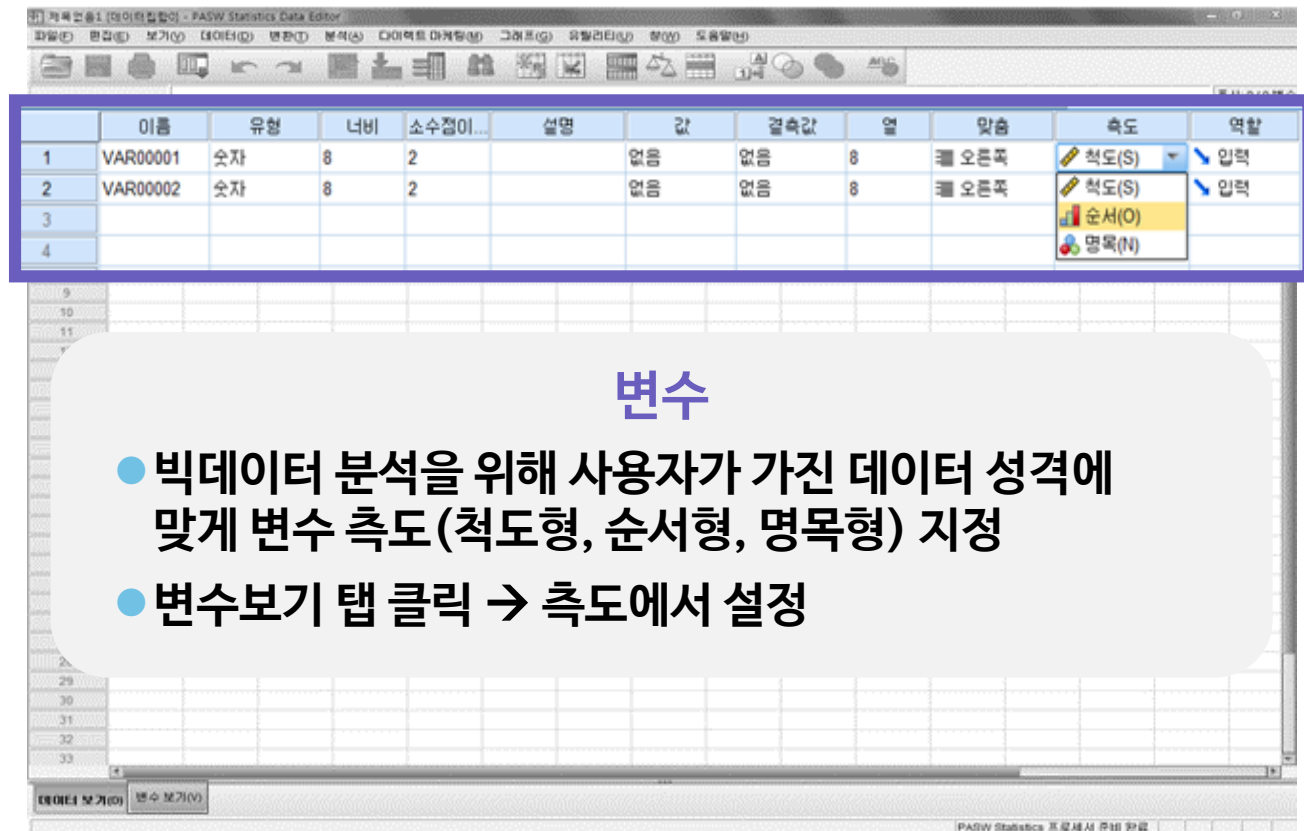


데이터 입력

- 데이터를 셀에 직접 입력하는 스프레드시트 형식 입력법
- sav, 엑셀, SAS 등의 파일 형식으로
기존 데이터 불러오거나 데이터 내보내기 사용 가능

3. 빅데이터 분석도구 및 기술 : SPSS

다. SPSS의 데이터 입력과 분석



	이름	유형	너비	소수점이...	설명	길이	폭	정확도	정확도	정확도	정확도	정확도
1	VAR00001	숫자	8	2		0.00	0.00	8	0.00	0.00	0.00	0.00
2	VAR00002	숫자	8	2		0.00	0.00	8	0.00	0.00	0.00	0.00
3												
4												

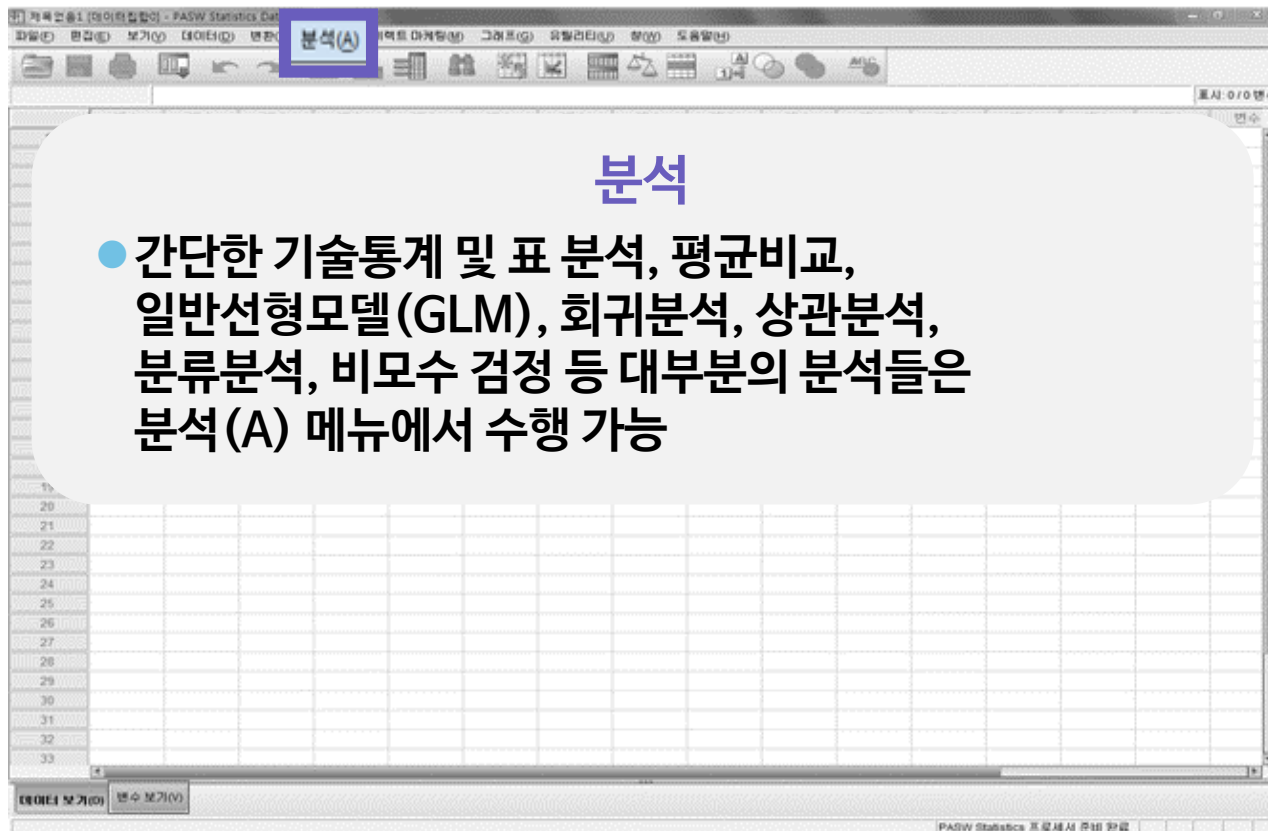
변수

- 빅데이터 분석을 위해 사용자가 가진 데이터 성격에 맞게 변수 측도(척도형, 순서형, 명목형) 지정
- 변수보기 탭 클릭 → 측도에서 설정

3. 빅데이터 분석도구 및 기술 : SPSS



다. SPSS의 데이터 입력과 분석



분석

- 간단한 기술통계 및 표 분석, 평균비교, 일반선형모델(GLM), 회귀분석, 상관분석, 분류분석, 비모수 검정 등 대부분의 분석들은 분석(A) 메뉴에서 수행 가능

3. 빅데이터 분석도구 및 기술 : SPSS



라. SPSS의 데이터 분석 예제

Iris flower data set

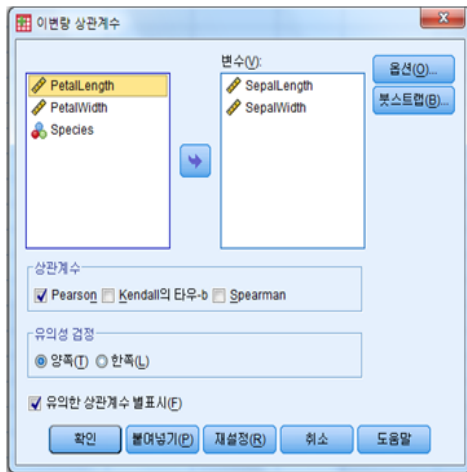
데이터

	이름	유형	너비	소수점이...	피 제	값	강화강	정	범위	측도	분석
1	Sepal.Length	숫자	8	2		0.0	0.0	8	0.0 ~ 7.7	최도(S)	인
2	Sepal.Width	숫자	8	2		0.0	0.0	8	0.0 ~ 4.4	최도(S)	인
3	Petal.Length	숫자	8	2		0.0	0.0	8	0.0 ~ 6.9	최도(S)	인
4	Petal.Width	숫자	8	2		0.0	0.0	8	0.0 ~ 1.9	최도(S)	인
5	Species	숫자	8	2		0.0	0.0	8	0.0 ~ 2.0	범(N)	인

3. 빅데이터 분석도구 및 기술 : SPSS

라. SPSS의 데이터 분석 예제

Iris flower data set



상관계수

		Sepal.Length	Sepal.Width
Sepal.Length	Pearson 상관계수	1	.743**
	유의확률 (양쪽)		.000
	N	50	50
Sepal.Width	Pearson 상관계수	.743**	1
	유의확률 (양쪽)	.000	
	N	50	50

** 상관계수는 0.01 수준(양쪽)에서 유의합니다.

꽃받침의 가로와 세로 길이의
상관계수는 0.743,
즉 양의 상관관계

04

빅데이터 분석도구 및 기술 : SAS

BIG

DATA

4. 빅데이터 분석도구 및 기술 : SAS



가. SAS



고가의 라이선스가 필요한 프로그램

4. 빅데이터 분석도구 및 기술 : SAS



가. SAS



- 공인되어 있는 대부분의 통계분석을 포괄하여 수행 가능
- 매우 정밀한 결과 제공
- 보고서 작성과 그래픽 작업도 가능
- 현재 SAS 9.3 최신버전 제공

4. 빅데이터 분석도구 및 기술 : SAS



나. SAS의 사용

DATA STEP

데이터 입력 및
편집을 위한 단계

- » 데이터의 입력
- » 데이터의 오류 판단 및 수정
- » 데이터의 샘플링 및 병합 등

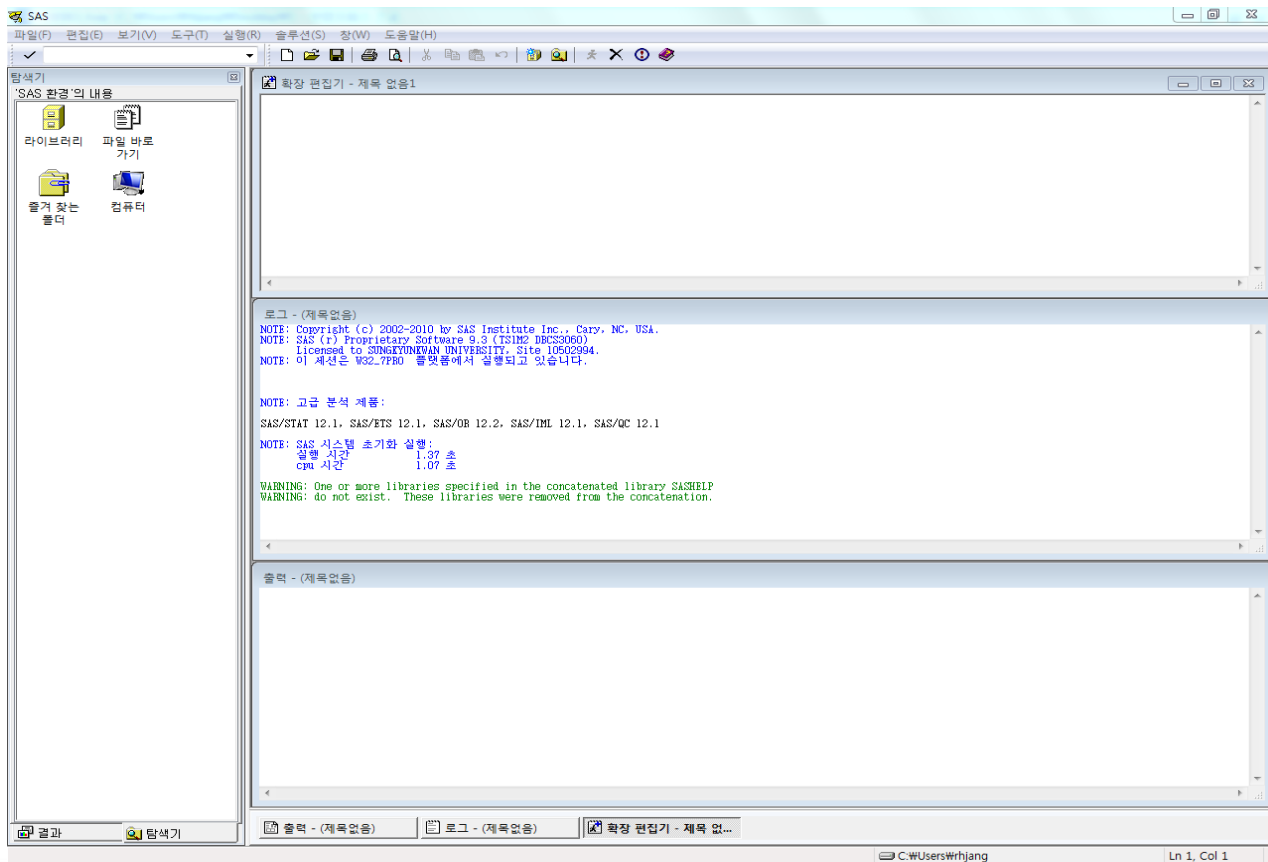
PROC STEP

본격적인
데이터 분석을 위한 단계

- » 데이터의 출력 · 정렬 · 요약
- » 여러 분석 기법을 이용한 통계 분석 수행

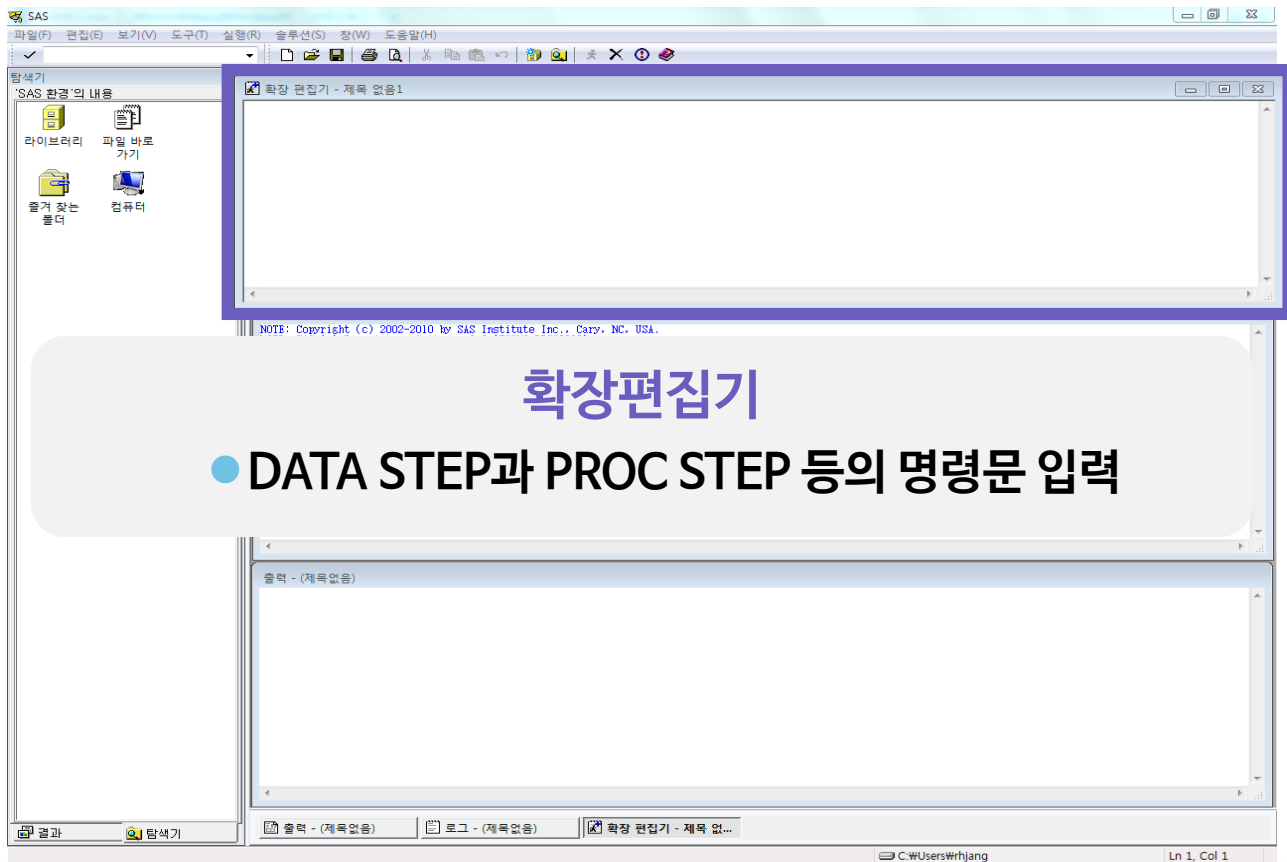
4. 빅데이터 분석도구 및 기술 : SAS

나. SAS의 사용



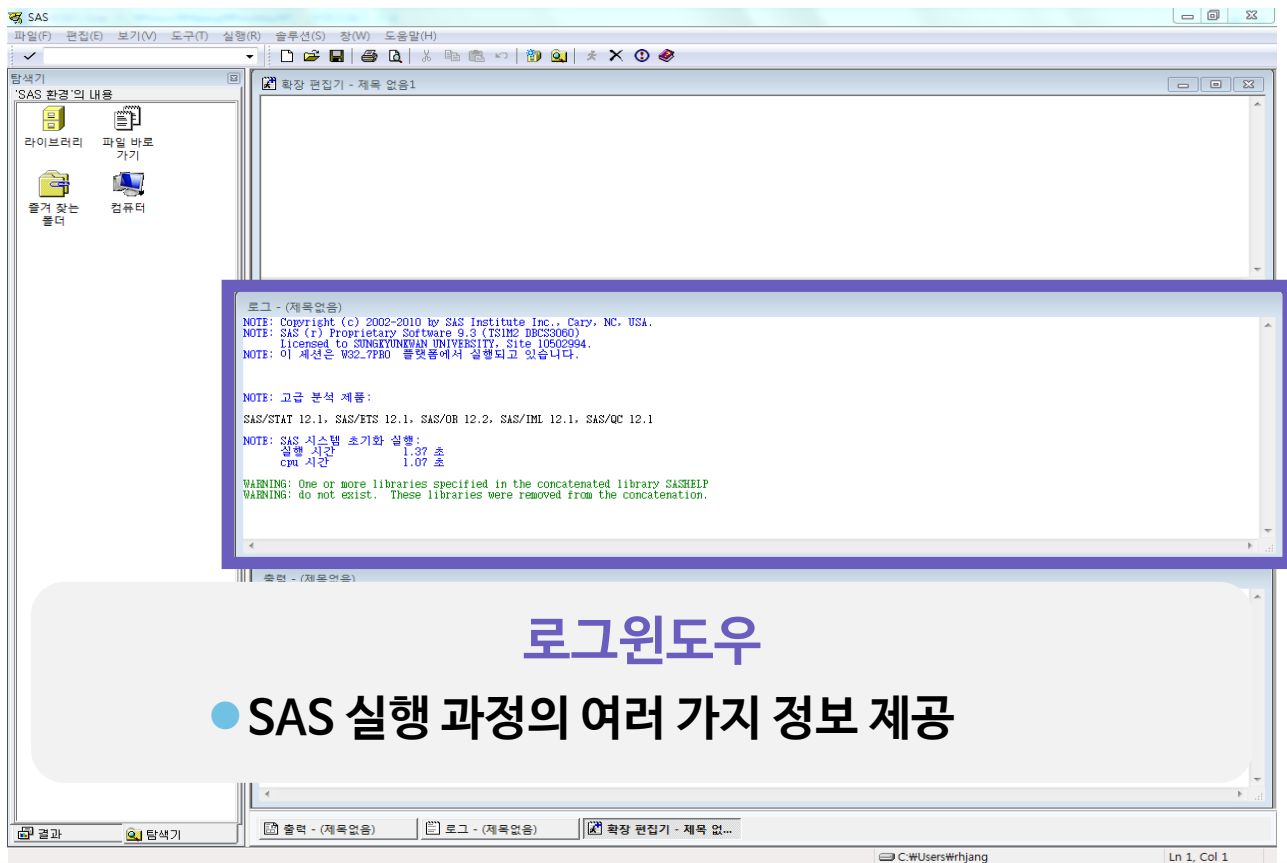
4. 빅데이터 분석도구 및 기술 : SAS

나. SAS의 사용



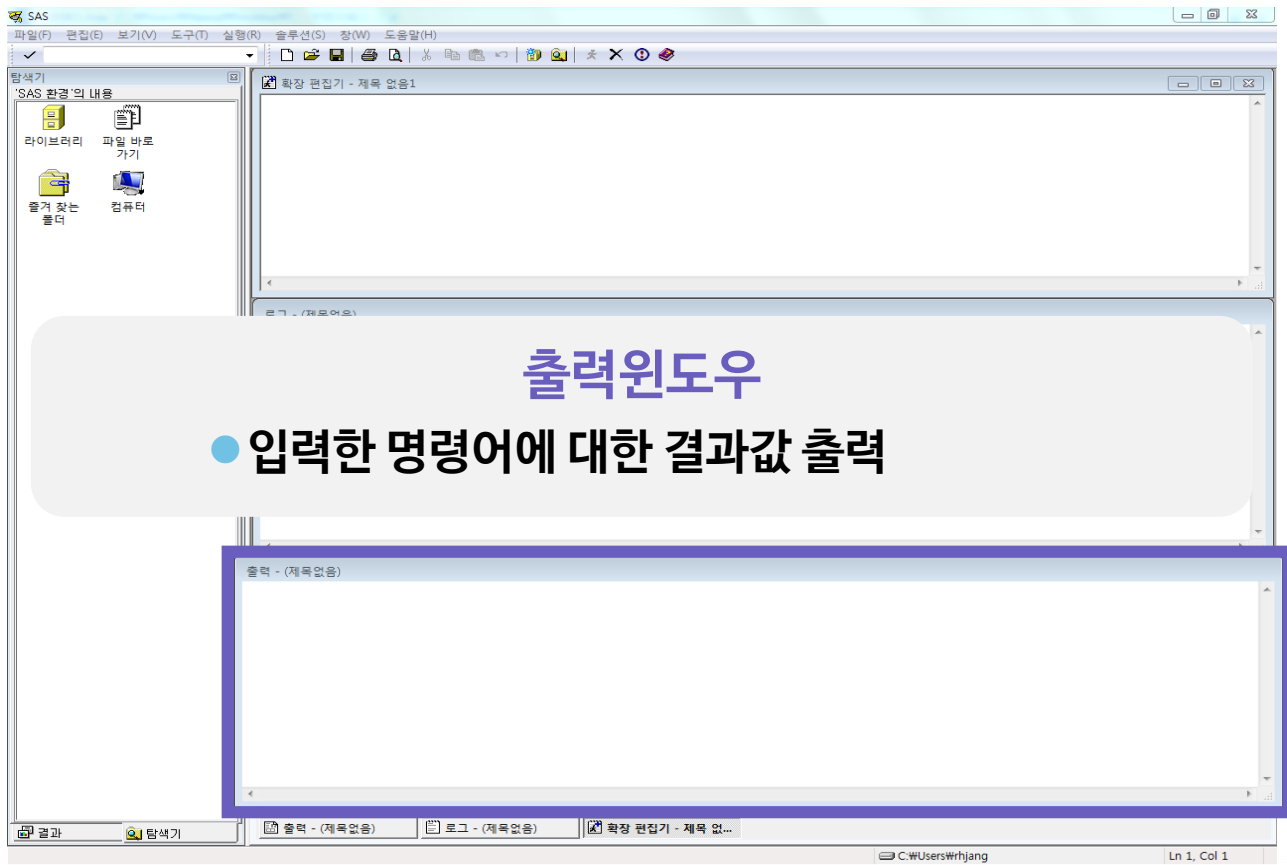
4. 빅데이터 분석도구 및 기술 : SAS

나. SAS의 사용



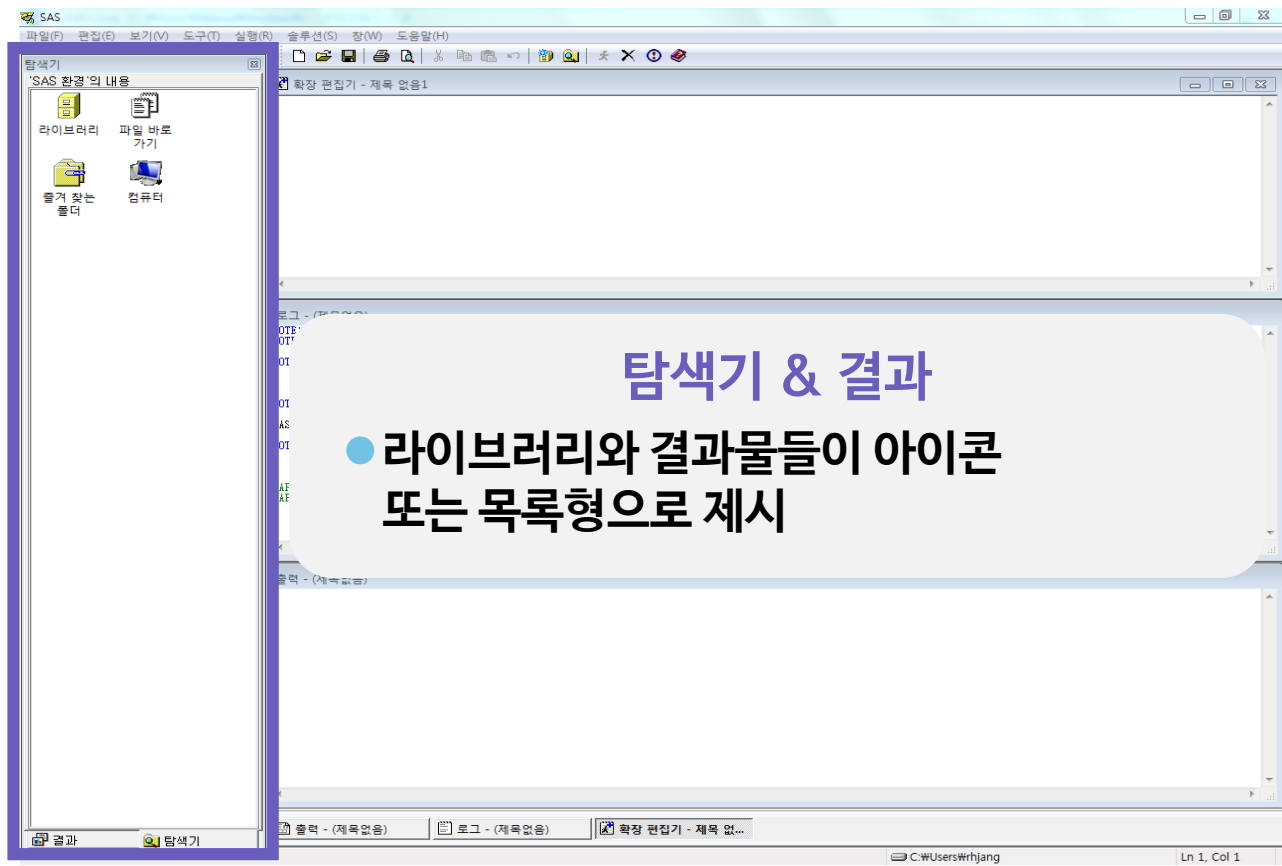
4. 빅데이터 분석도구 및 기술 : SAS

나. SAS의 사용



4. 빅데이터 분석도구 및 기술 : SAS

나. SAS의 사용



4. 빅데이터 분석도구 및 기술 : SAS



다. SAS의 데이터 입력과 분석

- SAS의 데이터 입력은 DATA STEP과 PROC STEP 모두 가능

DATA STEP

» input, cards문을 통해
직접 입력

PROC STEP

» proc import, infile 등의
명령문을 통해 가능

» infile의 경우 csv, xlsx 등
다양한 형식의 파일을 불러
올 수 있어 자주 이용

4. 빅데이터 분석도구 및 기술 : SAS



다. SAS의 데이터 입력과 분석

- SAS의 데이터 분석은 PROC STEP에서만 가능

» 각 통계 기법은 대응하는 프로그램의 명령문이 있고,
다양한 옵션문을 갖기 때문에 분석 정보 다양



4. 빅데이터 분석도구 및 기술 : SAS



다. SAS의 데이터 입력과 분석

- SAS의 데이터 분석은 PROC STEP에서만 가능
 - » 각 통계 기법은 대응하는 프로그램의 명령문이 있고, 다양한 옵션문을 갖기 때문에 분석 정보 다양
 - » SAS Help 창을 통해 프로그램 명령문과 옵션문에 대한 정보 획득 가능

프로그램 명령문을 입력 후,
전부 실행할 경우



실행버튼 또는 'F8',
[실행탭 - 실행(S)] 클릭

프로그램 명령문의
특정 부분만 실행할 경우



해당 명령문 블록 지정 후,
실행버튼 또는 'F8' 클릭

4. 빅데이터 분석도구 및 기술 : SAS



라. SAS의 데이터 분석 예제

Iris flower data set

④ 확장편집기에 입력한 데이터와 명령문

```
data iris;  
input sepal_length sepal_width petal_length petal_width;  
cards;  
5.1          3.5          1.4          0.2  
(중략)  
5            3.3          1.4          0.2  
;  
run;  
proc corr data=iris;  
var sepal_length sepal_width;  
run;
```


4. 빅데이터 분석도구 및 기술 : SAS



라. SAS의 데이터 분석 예제

Iris flower data set

👍 결과 뷰어에 제시되는 결과값

SAS 시스템

CORR 프로시저

2 개의 변수: sepal_length sepal_width

단순 통계량						
변수	N	평균	표준편차	합	최솟값	최댓값
sepal_length	50	5.00600	0.35249	250.30000	4.30000	5.80000
sepal_width	50	3.42800	0.37906	171.40000	2.30000	4.40000

피어슨 상관 계수, N = 50 H0: Rho=0 가정하에서 Prob > r		
	sepal_length	sepal_width
sepal_length	1.00000	0.74255 <.0001
sepal_width	0.74255 <.0001	1.00000

꽃받침의 가로와 세로 길이의
상관계수는 0.7425,
즉 양의 상관관계

05

빅데이터 분석도구 및 기술 : R

BIG

DATA

5. 빅데이터 분석도구 및 기술 : R



가. R



데이터 분석을 위한 통계분석 기법과 알고리즘,
시각화 기능을 지원하는 **오픈 소프트웨어 환경**

5. 빅데이터 분석도구 및 기술 : R



가. R



- 54메가바이트의 매우 작은 프로그램 용량
- 누구나 새로운 통계 데이터 분석 기법 업로드 및 다운
- 빅데이터와 관련된 분석(Big Data Analytics)을 위한 툴(Tools)로 주목
- 현재 3.3.2 최신버전 제공

5. 빅데이터 분석도구 및 기술 : R



나. R의 특징

1 오픈소스 기반의 무료 소프트웨어

5. 빅데이터 분석도구 및 기술 : R



나. R의 특징

- 1 오픈소스 기반의 무료 소프트웨어
- 2 텍스트, 엑셀, DBMS 등 다양한 종류의 정형 · 비정형 데이터를 이용할 수 있는 포괄적인 통계 플랫폼

5. 빅데이터 분석도구 및 기술 : R



나. R의 특징

- 1 오픈소스 기반의 무료 소프트웨어
- 2 텍스트, 엑셀, DBMS 등 다양한 종류의 정형 · 비정형 데이터를 이용할 수 있는 포괄적인 통계 플랫폼
- 3 윈도우, 유닉스, 리눅스, 맥OS 등 다양한 플랫폼에서 작동 가능한 멀티 운영환경 지원

5. 빅데이터 분석도구 및 기술 : R



나. R의 특징

- 1 오픈소스 기반의 무료 소프트웨어
- 2 텍스트, 엑셀, DBMS 등 다양한 종류의 정형 · 비정형 데이터를 이용할 수 있는 포괄적인 통계 플랫폼
- 3 윈도우, 유닉스, 리눅스, 맥OS 등 다양한 플랫폼에서 작동 가능한 멀티 운영환경 지원
- 4 대규모 데이터에서 분석결과를 직관적으로 이해할 수 있도록 시각화 기능 지원

5. 빅데이터 분석도구 및 기술 : R



나. R의 특징

- 1 오픈소스 기반의 무료 소프트웨어
- 2 텍스트, 엑셀, DBMS 등 다양한 종류의 정형 · 비정형 데이터를 이용할 수 있는 포괄적인 통계 플랫폼
- 3 윈도우, 유닉스, 리눅스, 맥OS 등 다양한 플랫폼에서 작동 가능한 멀티 운영환경 지원
- 4 대규모 데이터에서 분석결과를 직관적으로 이해할 수 있도록 시각화 기능 지원
- 5 유사 데이터에 대한 분석 작업을 기존 스크립트를 재사용하면서 처리할 수 있는 작업 재현성 제공

5. 빅데이터 분석도구 및 기술 : R



나. R의 특징

- 1 오픈소스 기반의 무료 소프트웨어
- 2 텍스트, 엑셀, DBMS 등 다양한 종류의 정형 · 비정형 데이터를 이용할 수 있는 포괄적인 통계 플랫폼
- 3 윈도우, 유닉스, 리눅스, 맥OS 등 다양한 플랫폼에서 작동 가능한 멀티 운영환경 지원
- 4 대규모 데이터에서 분석결과를 직관적으로 이해할 수 있도록 시각화 기능 지원
- 5 유사 데이터에 대한 분석 작업을 기존 스크립트를 재사용하면서 처리할 수 있는 작업 재현성 제공
- 6 최신 통계분석 및 마이닝 기능을 가진 패키지 및 샘플이 지속적으로 업데이트됨으로서 전세계적 커뮤니티 생태계 형성

5. 빅데이터 분석도구 및 기술 : R



다. R의 보완점



5. 빅데이터 분석도구 및 기술 : R



라. R 관련 IDE(Integrated Development Environment) 툴 : R Studio

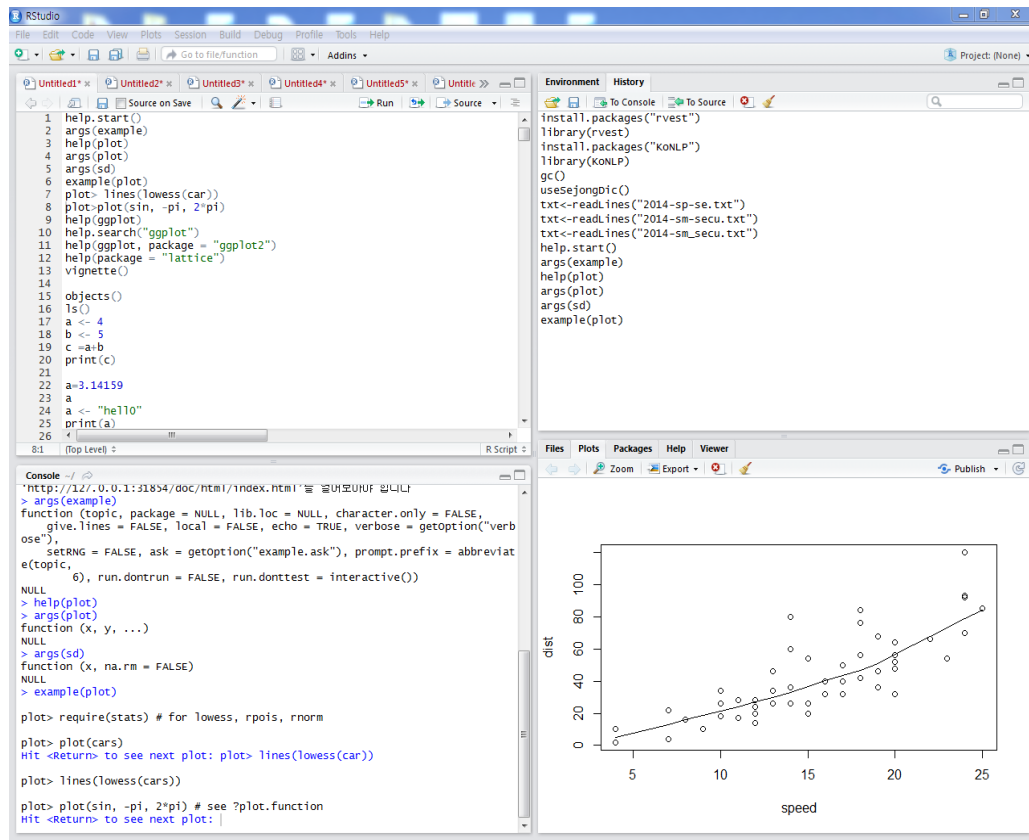


R을 가지고 통계 및 데이터 분석을 할 수 있는
도구 중 **가장** 일반적으로 사용되는 툴

5. 빅데이터 분석도구 및 기술 : R



라. R 관련 IDE(Integrated Development Environment) 툴 : R Studio



5. 빅데이터 분석도구 및 기술 : R



라. R 관련 IDE(Integrated Development Environment) 툴 : R Studio

1

에디터, 콘솔, 명령어 히스토리, 시각화, 파일 탐색,
패키지 관리 등을 한 화면에 제시

5. 빅데이터 분석도구 및 기술 : R



라. R 관련 IDE(Integrated Development Environment) 툴 : R Studio

- 1 에디터, 콘솔, 명령어 히스토리, 시각화, 파일 탐색, 패키지 관리 등을 한 화면에 제시
- 2 쉬운 파일 관리와 소스 코드 관리 시스템과의 연계 가능

5. 빅데이터 분석도구 및 기술 : R



라. R 관련 IDE(Integrated Development Environment) 툴 : R Studio

- 1 에디터, 콘솔, 명령어 히스토리, 시각화, 파일 탐색, 패키지 관리 등을 한 화면에 제시
- 2 쉬운 파일 관리와 소스 코드 관리 시스템과의 연계 가능
- 3 빌트인 데이터 뷰어 내장, 플로팅 히스토리, R help 결합, Sweave, knitr의 통합

5. 빅데이터 분석도구 및 기술 : R



라. R 관련 IDE(Integrated Development Environment) 툴 : R Studio

- 1 에디터, 콘솔, 명령어 히스토리, 시각화, 파일 탐색, 패키지 관리 등을 한 화면에 제시
- 2 쉬운 파일 관리와 소스 코드 관리 시스템과의 연계 가능
- 3 빌트인 데이터 뷰어 내장, 플로팅 히스토리, R help 결합, Sweave, knitr의 통합
- 4 R Markdown을 내장하여 문서와 코드의 결합을 쉽게 하고, 재현성 있는 분석 가능

5. 빅데이터 분석도구 및 기술 : R



라. R 관련 IDE(Integrated Development Environment) 툴 : R Studio

- 1 에디터, 콘솔, 명령어 히스토리, 시각화, 파일 탐색, 패키지 관리 등을 한 화면에 제시
- 2 쉬운 파일 관리와 소스 코드 관리 시스템과의 연계 가능
- 3 빌트인 데이터 뷰어 내장, 플로팅 히스토리, R help 결합, Sweave, knitr의 통합
- 4 R Markdown을 내장하여 문서와 코드의 결합을 쉽게 하고, 재현성 있는 분석 가능
- 5 패키지 빌드 자동화와 Rcpp 편집 환경 제공

5. 빅데이터 분석도구 및 기술 : R



라. R 관련 IDE(Integrated Development Environment) 툴 : R Studio

- 1 에디터, 콘솔, 명령어 히스토리, 시각화, 파일 탐색, 패키지 관리 등을 한 화면에 제시
- 2 쉬운 파일 관리와 소스 코드 관리 시스템과의 연계 가능
- 3 빌트인 데이터 뷰어 내장, 플로팅 히스토리, R help 결합, Sweave, knitr의 통합
- 4 R Markdown을 내장하여 문서와 코드의 결합을 쉽게 하고, 재현성 있는 분석 가능
- 5 패키지 빌드 자동화와 Rcpp 편집 환경 제공
- 6 리눅스, 맥, 윈도우 등 멀티 플랫폼 지원

06

빅데이터 분석도구 및 기술 : Python

BIG

DATA

6. 빅데이터 분석도구 및 기술 : Python



가. Python



동적 타이핑 (Dynamic typing) 범용 프로그래밍 언어

6. 빅데이터 분석도구 및 기술 : Python



가. Python



- 다양한 플랫폼에서 사용 가능
- 라이브러리(모듈) 풍부
- 순수한 프로그램 언어 기능 외의 다른 언어로 쓰여진 모듈들을 연결하는 풀언어(Glue language)로 자주 이용
- 많은 상용 응용 프로그램에서 스크립트 언어로 활용

6. 빅데이터 분석도구 및 기술 : Python



나. Python의 사용

```
1  # -*- coding: utf-8 -*-
2
3  # Define your item pipelines here
4  #
5  # Don't forget to add your pipeline to the ITEM_PIPELINES setting
6  # See: http://doc.scrapy.org/en/latest/topics/item-pipeline.html
7
8  import csv
9
10 class RTPipeline(object):
11
12     def __init__(self):
13         self.csvwriter = csv.writer(open("rt_movies_new.csv", "w"))
14         self.csvwriter.writerow(["title", "score", "genres", "consensus"])
15
16     def process_item(self, item, spider):
17         row = []
18         row.append(item["title"])
19         row.append(item["score"])
20         row.append(' '.join(item["genres"]))
21         row.append(item["consensus"])
22         self.csvwriter.writerow(row)
23         return item
24
```