1

데이터 읽기_쓰기

1. 데이터 구조

분석 데이터 : 행렬 구조

◉ 행 : 개체(모집단, 표본)의 변수(열) 관측치

- 동일 개체는 결코 2개 행에 분리 입력될 수 없음
- 시계열 데이터의 경우 행은 날짜임

◉ 열 : 개체의 관심 특성(변수)

- 동일 변수는 절대 2개 열에 입력되지 않음

2. 데이터 형식

데이터는 숫자와 문자로 이루어져 있음, 이미지도 숫자의 조합임

숫자는 연산 작업이 가능

문자는 개체에 대한 식별 정보임

1) 행렬

R	SAS
as.matrix()	PROC IML
matrix(c(),nrwo=, ncol=)	

◉ 열이 하나인 행렬 : 열 벡터

◉ 행이 하나인 행렬 : 행 벡터

2) 데이터

내부 분석 모듈을 활용하여 분석 가능한 데이터 형태

R	SAS
data.frame()	data
외부데이터 읽어오면 데이터 형식으로 내부 저장	

◉ as.matrix() 함수 : 데이터를 행렬로 변환

3) 변수 속성

R	SAS
as.char(변수명)	format 변수명 $
as.numeric(변수명)	format 변수명 6.3
as.Date	format 변수명 date7.

3. 통계소프트웨어 데이터

모든 통계소프트웨어는 자신들이 다룰 수 있는 형태의 데이터 포맷이 있음

R	SAS
data.frame	확장자 sas7bdat
오브젝트	data

다양한 형식의 외부 데이터(txt, xlsx, csv, db, ...)를 R, SAS 통계소프트웨어를 사용하여 분석하려면 각 소프트웨어 데이터 형식에 맞추어야 함

4. 외부 데이터

모든 통계소프트웨어는 다양한 형식의 외부 데이터를 불러오는_내보내는 함수나 PROC 혹은 메뉴가 있음

측정장비, DB 서버 등에서 데이터를 외부로 내보는 경우 가장 널리 사용되는 형식이 TXT형식(파일 사이즈가 가장 적음)으로 내보내며 행의 관측치는 코마(,), |, 혹은 &으로 구분하여 내보낸다.

TXT 형식의 데이터의 문제는 각 열 변수에 대한 정보를 따로 적어야 한다는 것이다. 이는 매우 번거러운 일이다. 이로 인하여 txt 형식의 데이터 제공 시 반드시 변수 속성 및 이름이 있는 메타 데이터를 제공하거나 “변수명 및 포맷”을 적용한 프로그램을 제공함(통계청 MDIS 참고)

다루기 가장 편리한 형식 : CSV (comma seperate value) 범용 스프레드시트 엑셀에서 모든 데이터의 형식을 불러올 수 있고 이를 CSV 포맷으로 내보낼 수 있음

5. 예제 데이터

SAS sashelp 라이브러리 내의 예제 데이터 baseball을 활용

```
Name,Team,nAtBat,nHits,nHome,nRuns,nRBI,nBB,YrMajor,CrAtBat,CrHits,Cr
"Allanson, Andy",Cleveland,293,66,1,30,29,14,1,293,66,1,30,29,14,Amer
"Ashby, Alan",Houston,315,81,7,24,38,39,14,3449,835,69,321,414,375,Na
"Davis, Alan",Seattle,479,130,18,66,72,76,3,1624,457,63,224,266,263,A
"Dawson, Andre",Montreal,496,141,20,65,78,37,11,5628,1575,225,828,838
```

Name	Team	nAtBat	nHits	nHome
Allanson, Andy	Cleveland	293	66	1
Ashby, Alan	Houston	315	81	7
Davis, Alan	Seattle	479	130	18

문자 : 선수이름, 팀 이름, 소속 리그 등

숫자 :

6. 데이터 읽기

1) CSV 형식

SAS : work 라이브러리 baseball sas 데이터

```
proc import datafile='C:\tmp\baseball.csv'
out=baseball dbms=csv replace;
run;
```

VIEWTABLE: Work.Baseball

	Name	Team	nAtBat	nHits
1	Allanson, Andy	Cleveland	293	66
2	Ashby, Alan	Houston	315	81
3	Davis, Alan	Seattle	479	130
4	Dawson, Andre	Montreal	496	141

R : ds data.frame 오브젝트

```
ds<-read.csv("C:/TMP/
baseball.csv",header=True)
names(ds)
```

```
> names(ds)
 [1] "Name"      "Team"      "nAtBat"    "nHits"     "nHome"
 [6] "nRuns"     "nRBI"      "nBB"       "YrMajor"   "CrAtBat"
[11] "CrHits"    "CrHome"    "CrRuns"    "CrRbi"     "CrBB"
[16] "League"    "Division"  "Position"  "nOuts"     "nAssts"
[21] "nError"    "Salary"    "Div"       "logSalary"
```

2) txt 형식(comma)

SAS : work 라이브러리 baseball1 sas 데이터

```
proc import datafile='C:\tmp\baseball.txt'
out=baseball1 dbms=dlm  replace;
delimiter=',';
run;
```

VIEWTABLE: Work.Baseball1

	Name	Team	nAtBat	nHits	nHome
1	Allanson, Andy	Cleveland	293	66	1
2	Ashby, Alan	Houston	315	81	7
3	Davis, Alan	Seattle	479	130	18
4	Dawson, Andre	Montreal	496	141	20
5	Galarraga, Andres	Montreal	321	87	10

R : ds1 data.frame 오브젝트

```
ds1<-read.table("C:/TMP/
baseball.txt",sep=",",header=T)
names(ds1)
```

```
> names(ds)
 [1] "Name"      "Team"      "nAtBat"    "nHits"     "nHome"
 [6] "nRuns"     "nRBI"      "nBB"       "YrMajor"   "CrAtBat"
[11] "CrHits"    "CrHome"    "CrRuns"    "CrRbi"     "CrBB"
[16] "League"    "Division"  "Position"  "nOuts"     "nAssts"
[21] "nError"    "Salary"    "Div"       "logSalary"
```

7. 날짜 형식 데이터 만들기

```
proc contents data=sashelp.air;
run;
```

변수와 속성 리스트(오름차순)

#	변수	유형	길이	출력형식	레이블
2	AIR	숫자	8		international airline travel (thousands)
1	DATE	숫자	8	MONYY.	

	DATE	international airline travel (thousands)
1	JAN49	112
2	FEB49	118
3	MAR49	132

```
522  data _null_;
523      a="03jan1961"d;
524      put "SAS날짜 숫자 매칭" a;
525      put "SAS날짜 포맷" a=date7.;
526  run;

SAS날짜 숫자 매칭368
SAS날짜 포맷a=03JAN61
```

- 1961년 1월 3일 = 정수 368로 인식함
- date7. 포맷 : 날짜 형식, 7자리

SAS

```
data air; set sashelp.air;
   format date2 date7.; date2=_n_;
   day=weekday(date2); week=week(date2);
   yr=year(date2); mon=month(date2);
qtr=qtr(date2);
run;
```

- weekday() : 요일, 1=일요일, 2=월요일, ..., 7
- week() : 한 해 몇 번째 주인가?
- qtr() : 분기, year() : 해

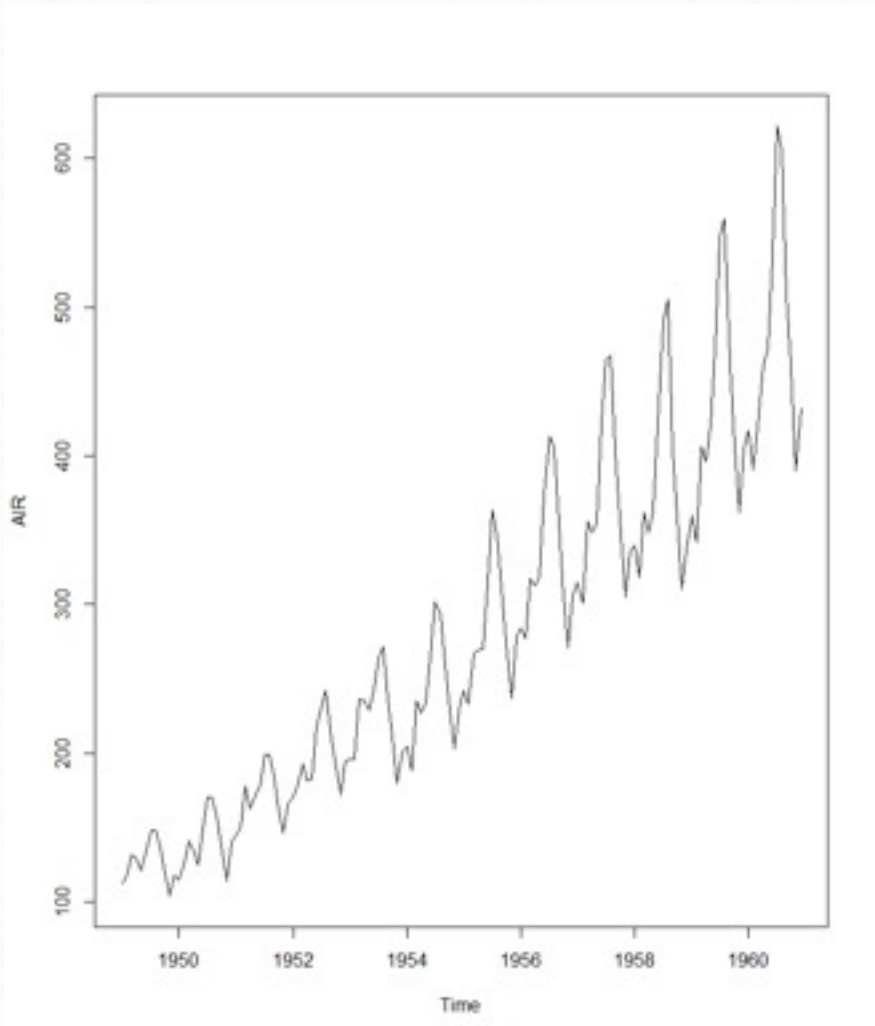
	DATE	international airline travel (thousands)	date2	day	week	yr	mon	qtr
1	JAN49	112	02JAN60	7	0	1960	1	1
2	FEB49	118	03JAN60	1	1	1960	1	1
3	****	132	04JAN60	2	1	1960	1	1
4	APR49	129	05JAN60	3	1	1960	1	1
5	****	121	06JAN60	4	1	1960	1	1
6	JUN49	135	07JAN60	5	1	1960	1	1
7	JUL49	148	08JAN60	6	1	1960	1	1

```
symbol i=join v=dot c=red;
title "승객수(단위: 천명)";
footnote "1949년 1월 ~ 1960년 12월" ;
proc gplot data=sashelp.air;
      plot air*date;
run;
```

- symbol : 산점도 개체 심볼, v(alue), c(olor)
- title : 그래프 타이틀, footnote : 아래 제목

R

```
library(sas7bdat)
air<-read.sas7bdat("air.sas7bdat")[2]
air.ts<-ts(air, start = c(1949,1),
end=c(1960,12), frequency = 12)
plot(air.ts)
day<-format(seq(as.Date('1949-01-01'),
to=as.Date('1960-12-01'),by='1
month'),"%m/%y")
```



8.데이터 내보내기

1) CSV 형식으로 내 보내기

SAS

```
proc export data=ds.baseball   outfile='C:
\tmp\baseball.csv' dbms=csv replace;
 run;
 proc export data=ds.baseball   outfile='C:
\tmp\baseball.xlsx' dbms=xlsx replace;
 run;
```

R

```
write.csv(ds, file='C:/TMP/
baseball_out2.csv')
```

2) txt 형식으로 내보내기

SAS

```
proc export data=ds.baseball   outfile='C:
\tmp\baseball.txt'   dbms=dlm replace;
delimiter=','; run;
```

R

```
write.table(ds, file='C:/TMP/
baseball_out.csv')
```