

R 프로그래밍

(9주차)

2016. 04. 30(토)

장운호

(ADP 002-0004)

목차

※ 지난 주 복습

I. 외부 데이터

II. 데이터 읽고 저장하기

III. 데이터 정렬

IV. 날짜형 데이터 처리

※ R 데이터 구조 요약

같은 종류의
데이터 타입을 가진
벡터를
수용하는
Data Type

배열(array)
교재 3장

매트릭스(matrix)
교재 3장

N차원데이터
수용가능

행(row)과
열(column)로
이루어진
2차원 데이터

다른 종류의
데이터 타입을 가진
벡터들을
결합시킬 수 있는
Data Type

리스트(list)
교재 4장

데이터프레임(dataframe)
교재 5장

[1] [[1]]	[2] [[2]]	[3] [[3]]	[4] [[4]]
장운호	46	등산	관악산
		볼링	북한산
			청계산

이름	나이	취미	비고
장운호	25	등산	관악산
홍길동	60	볼링	A클럽
김철수	45	볼링	B클럽

범주형 데이터
벡터들을
효율적으로 표현하는
Data Type

요인(factor)
교재 6장

남	0
여	1

이름
(level)

알파벳순 정수
배정 (default)



I. 외부 데이터

1. 외부 데이터의 종류

외부 데이터는 크게 텍스트 데이터와 바이너리 데이터로 구분할 수 있음.

텍스트
(Text)

데이터

- CSV (Comma Separated Value)
- TSV (Tab Separated Value)
- HTML (Hyper Text MarkUp Language)
- XML (Extensible MarkUp Language)
- SVG (Scalable Vector Graphics)

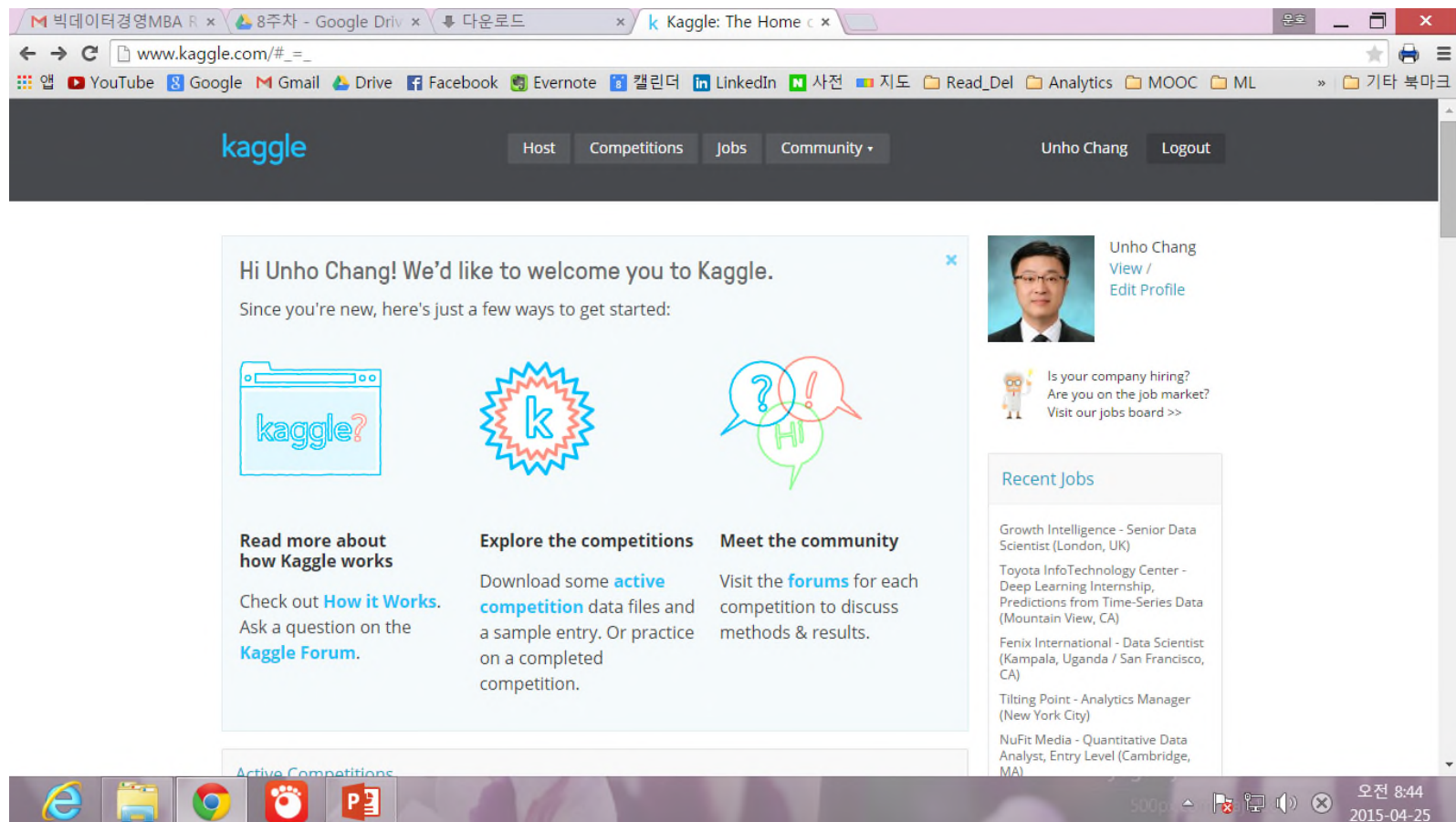
바이너리
(Binary)

데이터

- docx (MS Word 문서 파일)
- xlsx (MS Excel 문서 파일)
- pptx (MS Powerpoint 문서 파일)
- psd (포토샵 그림 저장 파일)
- RData (R Object 저장 파일)


2. 외부 데이터 Source

매우 다양한 곳에서 데이터를 구할 수 있으나, kaggle 등에 유용한 데이터가 많음.




3. 실습용 외부데이터 소개

www.capitalbikeshare.com/trip-history-data



에 대 한 자주 묻는 질문(FAQ) 연락처 스토어 [회원 로그인](#) [가입](#)

보도 자료 키트 홈 어떻게 작동 역 지도 가격 자전거로 탐험 안전 파트너 뉴스 시스템 데이터




여행 기록 데이터

여행 기록 데이터의 개요

임대 시스템 내에서 발생 하면 우리의 소프트웨어는 여행에 대 한 기본 데이터를 수집 합니다. 그 데이터는 우리의 시스템에서 내보낸 고 분석 이나 연구의 다양 한 유형에 대해 사용할 수 있습니다. 이 데이터를 사용할 수 있도록 하여 그 분석 이나

[가입](#)
하루, 3 일, 달, 또는 년에 대 한.





II. 데이터 읽고 저장하기

1. read.csv 함수

read.csv : CSV 파일을 data.frame Type으로 읽어 들임.

문법

```
read.csv (  
  file, # 파일명(문자열로 표시)  
  header=TRUE, #파일의 첫행의 헤더 처리 여부  
  na.strings="NA",  
  stringsAsFactors=TRUE,  
  nrows=      #읽어들일 행의 숫자를 미리 지정가능  
  fileEncoding="CP949", #windows에서 생성된 파일을 (from)  
  encoding="UTF-8"      #Linux에서 읽어들이때(to)  
)
```

2. save / load 함수

객체를 바이너리 형태로 저장하고, 저장된 바이너리 파일을 다시 읽어 들이는 함수

- 객체를 그대로 저장하기 때문에, 저장전의 상태가 완벽하게 보관됨.

문법

save(

R-Object,...,

file="파일명" # 작업디렉토리 외부에 있는 디렉토리에 save하고자 경우는

Full 경로명을 정확히 입력해야 제대로 save됨.

)

load(

"파일명" # 작업디렉토리 외부에 있는 데이터를 load하고자 경우는

Full 경로명을 정확히 입력해야 제대로 load됨.

)

3. write.csv 함수

write.csv : R 데이터 객체를 CSV 파일로 저장해 줌.

문법

write.csv (

객체명, # 메모리에 올라가 있는 데이터 객체명

file="파일명", # 저장하고자 하는 파일명을 문자열로 지정

fileEncoding="CP949" or "UTF-8" #저장하고자하는 문자열의 인코딩

row.names=**TRUE**

)

4. 객체 관리 함수

현재 메모리 상에 올라가 있는 데이터를 확인하고,
필요시 객체를 메모리 상에서 지움으로써 메모리 공간을 늘리는 함수

ls() #현재 메모리에 올라가 있는 모든 객체(objects)들을
List-up 해서 파일명으로된 벡터를 반환 해줌

rm(객체명) 또는
rm(list=c("객체명",.....,"객체명"))

#접근가능한 객체들 중에서 객체명으로 지정한 객체를
지워주는(remove) 함수



Ⅲ. 데이터 정렬

1. sort / rank

숫자나 문자로 된 벡터를 알파벳순의 오름차순으로 정렬하는 함수

```
> age <- c(25, 60, 45, 19, 48, 27)
```

```
> sort(age)
[1] 19 25 27 45 48 60
```

```
> sort(age, decreasing = T)
[1] 60 48 45 27 25 19
```

```
> rank(age)
[1] 2 6 4 1 5 3      # 가장 낮은 수치에 1을 부여함.
```

2. order

정렬된 이후의 자리번호를 벡터로 리턴해 주는 함수로, 여러 열(column)으로 이루어진 매트릭스나 데이터 프레임 전체를 정렬하는 용도로 자주 사용됨.

```
> age <- c(25, 60, 45, 19, 48, 27)
```

```
> order(age)
[1] 4 1 6 3 5 2
```

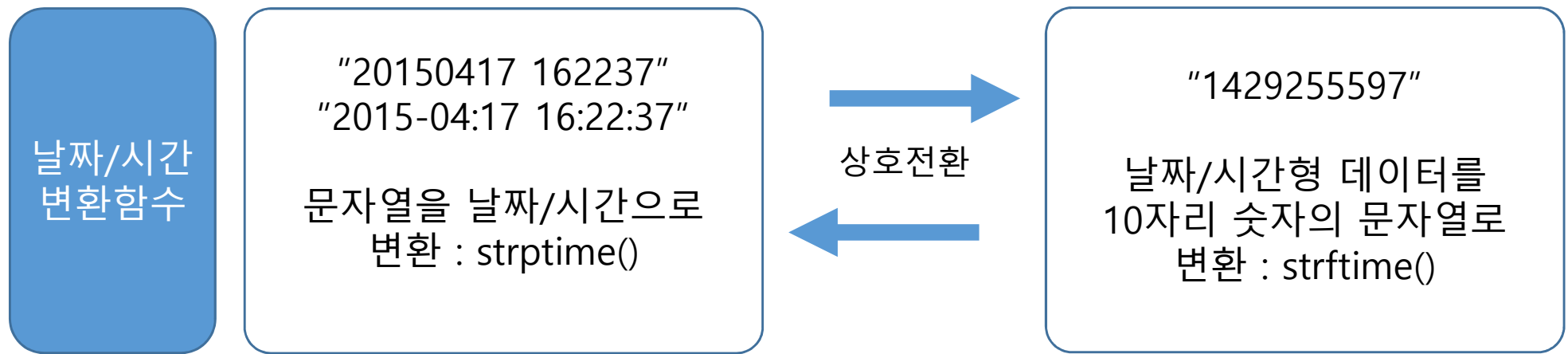
```
> order(age, decreasing = T)
[1] 2 5 3 6 1 4
```



IV. 날짜형 데이터 처리

1. 날짜형 데이터의 타입 구분

- 1) POSIXct : 1970년 1월 1일 0시 0분 0초 부터의 경과 시간을 초단위로 세어서 현재시간 표시
- 2) POSIXlt : 시, 분, 초, 요일, 매월초하루부터의 경과 날짜수, 매년 초하루부터의 경과 날짜수, 1900년 이후 경과된 년수, 썸머타임여부, 등을 list 데이터형태로 변환하고 표시



2. 날짜형 데이터 처리

수집된 텍스트내의 날짜 표현 문자열을 날짜/시간형 데이터로 형변환이 가능함.

- 포맷 리터럴(literal) 지정을 통해 원하는 형태 즉, 주/월/시간대 등으로 변환하여 활용이 가능함.

Strptime

내부구성

(POSIXlt)

\$sec	\$min	\$hour	\$mday	\$mon	\$year
[1] 37	[1] 26	[1] 16	[1] 17	[1] 3 ¹⁾	[1] 115
\$wday	\$yday	\$isdst	\$zone	\$gmtoff	
[1] 5 ²⁾	[1] 106	[1] 0	[1] "KST"	[1] NA	

주: 1) 1월(0), 2월(1), 3월(2), 4월(3), 5월(4), 6월(5), 7월(6), ... , 11월(10), 12월(11)

2) 일(0), 월(1), 화(2), 수(3), 목(4), 금(5), 토(6)

날짜형
포맷
지정
리터럴

리터럴	의미
%Y	연도를 4자리 숫자로 표시
%m	월을 2자리 이하 숫자로 표시
%d	날짜를 1부터 31의 숫자로 표시
%H	시간을 0부터 23의 숫자로 표시
%M	분을 0부터 59의 숫자로 표시
%S	초를 0부터 59의 숫자로 표시

리터럴	의미
%j	당해년도 몇번째 날짜(1~366)로 표시
%W	월요일 기준 당해년도 주차(00~56) 표시
%w	요일을 정수(0~6, 일요일 0)로 표시
%U	일요일 기준 당해년도 주차(00~56) 표시
%u	요일을 정수(1~7, 일요일 1)로 표시
%p	해당 타임존에 맞는 오전/오후 표시

3. 타임존 지정

날짜를 수치데이터로 바꾸어 프로그램적으로 처리하기 위해서는
시간카운트의 origin과 time zone을 별도의 Argument로 지정해야 함.

- 이 경우 origin과 time zone은 함수에서 요구는 형태와 약간만 달라도, 에러가 발생하는 바, 이에 주의해야 함.

Origin
(시간계산
기준일자)

1970년 1월 1일.

※ Default Format 반드시 준수 필요 : "1970-01-01"

Time zone
(세계 표준시와
의
차이 반영)

위키피디아의

[List of tz database time zones](#)에 나와 있는 TZ명

한국 : "Asia/Seoul"

미국 워싱턴 : "America/Dawson"

일본 : "Asia/Tokyo" 등

※ GMT : [Africa/Abidjan](#)과 동일

End of Document.

감사합니다.