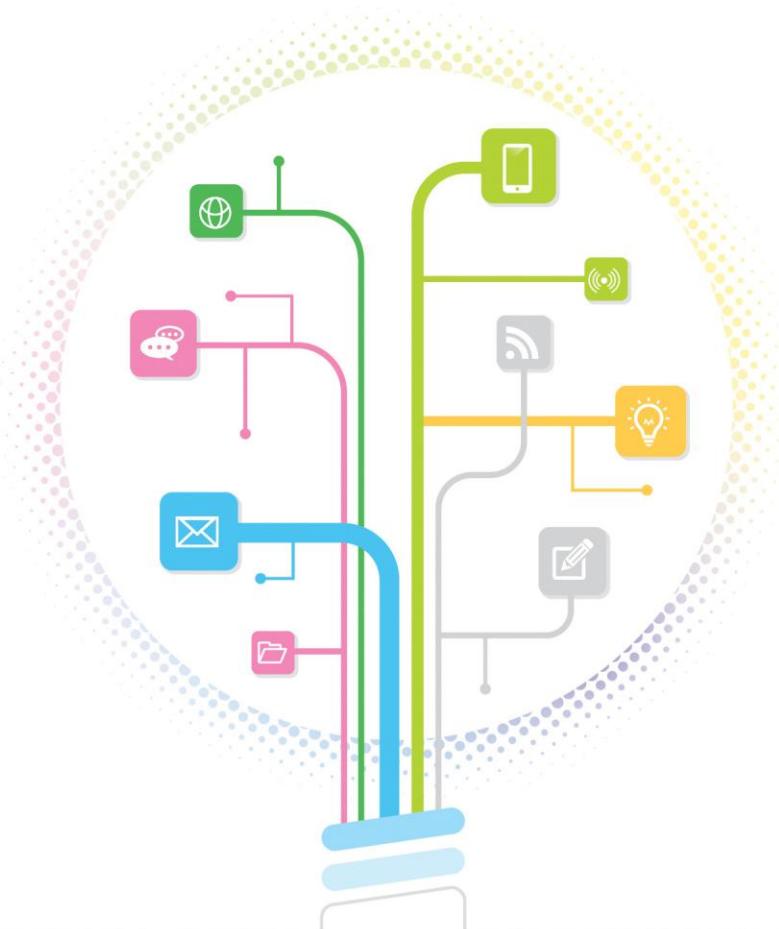


데이터 분석 콘텐츠 10 글로벌 활용 매뉴얼



미래창조과학부



한국정보화진흥원



CONTENTS

Beginning Level 초급과정

I 개요

개요	9
----	---

II 수집

개요	13
수집 데이터	14
데이터 수집	16
데이터 작업 영역 이동 스크립트	19

III 가공

개요	23
데이터 가공 R 스크립트	28

IV 저장

개요	31
R Studio 활용 저장	32



V 분석

개요	37
R Studio 활용 분석	38

VI 시각화

개요	43
분석 데이터 시각화	45
데이터 분석	46

VII 예제 문제

예제 문제1. 헬스케어 데이터에서 미국내 인종별 서비스 신뢰도를 분석하라.	49
예제 문제2. 미국의 헬스케어 서비스에 대한 성별 신뢰도를 분석하라.	50

CONTENTS

Intermediate Level **중급과정**

I 개요

개요	55
----	----

II 수집

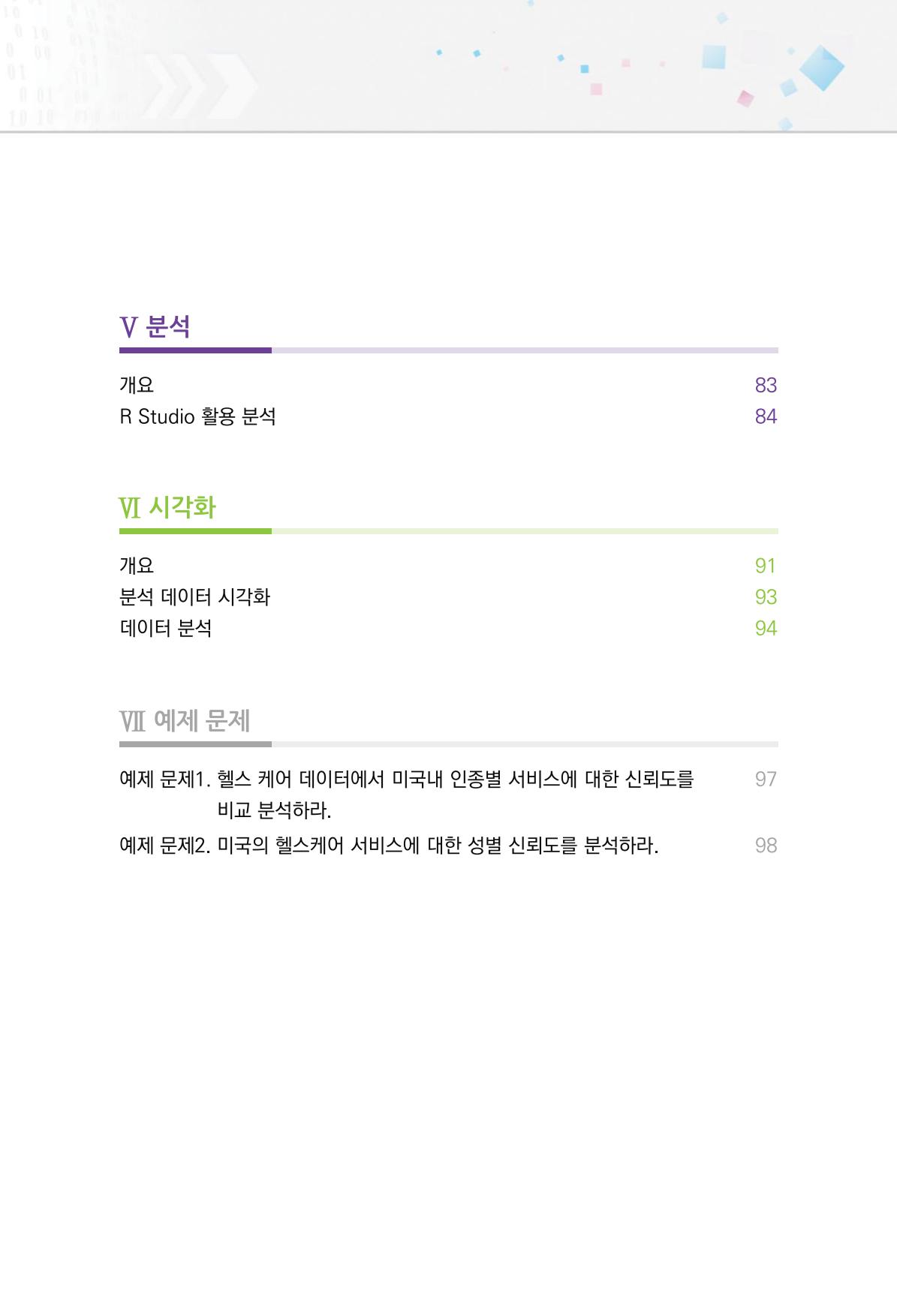
개요	59
수집 데이터	60
데이터 수집	62
데이터 작업 영역 이동 스크립트	65

III 가공

개요	69
데이터 가공 R 스크립트	74

IV 저장

개요	77
R Studio 활용 저장	78



V 분석

개요	83
R Studio 활용 분석	84

VI 시각화

개요	91
분석 데이터 시각화	93
데이터 분석	94

VII 예제 문제

예제 문제1. 헬스 케어 데이터에서 미국내 인종별 서비스에 대한 신뢰도를 비교 분석하라.	97
예제 문제2. 미국의 헬스케어 서비스에 대한 성별 신뢰도를 분석하라.	98



글로벌 

Beginning Level

초급과정







I 개요

개요

9

I

개요

▶ 개요

글로벌 데이터는 공공과 연구 데이터로 구분되며, 세계적으로 헬스케어와 빅데이터 분석의 관심도가 높아지고 있기 때문에 미국의 헬스케어 연구 센터의 데이터를 이용한다. 해당 데이터에 포함되어 있는 데이터의 일부를 추출하는 방법과 추출된 데이터를 이용하여 2007년 이후의 연령별 헬스케어 우선 대상자에 대한 신뢰도를 알아보고자 한다. 이를 통해 국내 헬스케어 서비스에 대한 연령별 신뢰도와 비교하여 헬스케어 서비스를 재점검할 수 있다.

▶ 활용 데이터

- LHI Access To Health Services Data 2014 : 미국 헬스케어 센터(LHIA)

▶ 선행학습

- 리눅스 – 파일시스템 구조, 쉘 명령어, 쉘 스크립트 실행 방법
- R 프로그래밍 언어 – 파일 불러오기, 라이브러리 등록, 데이터 함수(프레임, 테이블), 그래프 함수, 제어문(함수 호출, 외부 함수, 함수 정의) 사용방법
- 데이터 구조 – CSV, JSON 데이터 구조, Text 파일 저장 구조 이해
- R 차트 – 내부 차트(막대, 바, 원 등), 외부 차트(막대, 클라우드, 3D, D3 차트) 사용방법

▶ 요구사항

- 엑셀 데이터에서 분석에 필요한 연령별 데이터만을 추출하여 관리한다.
- 신뢰도의 증감에 대한 연령별 시계열 분석을 실시하라.
- 신뢰도의 증감에 대한 평균 증가율을 구하고, 평균 증가율보다 높은 연령과 신뢰도 구간을 찾아라

▶ 분석 절차

- 수집된 글로벌 헬스케어 데이터 셋을 분석 저장소로 복사한다.
- 엑셀 데이터의 구조상 각각의 시트별로 데이터를 불러온다.
- 각각의 시트에서 연령대별 증감분포 분석에 필요한 연도별 신뢰 구간 분석 데이터를 추출한다.
- 추출된 데이터는 연령대별로 구분하여 연도에 따른 증가/감소 형태를 분석한다.
- 출력된 그래프를 통해 연령대별 집중도가 가장 높은 세대를 구분하고, 이를 통해 헬스케어 서비스의 관심도가 높아지는 형태를 분석한다.



- **헬스케어(Healthcare)** : 질병을 치료하고 관리하는 것 외에 의학, 치의학, 약리학, 임상병리 진단학, 간호학 등 건강과 관련된 전문가들이 사람의 건강을 유지하도록 하는 행위



II 수집

개요	13
수집 데이터	14
데이터 수집	16
데이터 작업 영역 이동 스크립트	19



수집

> 개요

글로벌 데이터는 공공데이터를 기준으로 다양한 데이터가 공개되고 있으며, 공개된 데이터는 연도별 기준에 따라 가공, 정제가 되어 있는 데이터들이 존재한다. 그렇기 때문에 수집을 위한 데이터 선정이 중요하고, 이를 수집하는 방법에 따라 접근하는 것이 중요하다. 그래서 글로벌 데이터는 각 국가나 연구기관이 제공하는 API 또는 기업의 수집기를 통해 수집할 수 있으며, 수집된 글로벌 데이터를 통해 사회적 이슈나 비즈니스 분석 등과 같은 다양한 사회적 분석 모델에 적용할 수 있다.

> 수집 방법

- **API 데이터 수집** : 글로벌 데이터의 수집은 API를 이용하여 일정량의 데이터를 정기적으로 수집할 수 있으며, 각 국가별, 연구기관별 데이터 제공 정책에 따라 수집 데이터의 범위를 제한하고 있다.
- **데이터 제공** : 글로벌 데이터는 OpenAPI, 자료수집기 (Crawler), 데이터 구매 등으로 데이터를 수집 할 수 있으며, 실습용 자료는 빅데이터 분석활용 센터에 접속하여 글로벌 헬스케어 데이터 셋을 다운로드 받을 수 있도록 원시 데이터를 제공하고 있다.



- **비정형 데이터(Unstructured Data)** : 일정한 규격이나 형태를 지닌 숫자 데이터와 달리 그림, 영상, 문서 등과 같이 서로 다른 형태의 구조화되지 않은 데이터

> 수집 데이터

> 미국 헬스케어 센터 데이터

LHI Topic: Access to Health Services			
Objective: AHS-1.1			
Increase the proportion of persons with medical insurance			
Population	Year 2008 Percent	Year 2008 Standard Error	Year 2008 Lower 95% Confidence Interval
Target: 100			
Increase the proportion of persons with medical insurance			
Total	83.2	0.286	82.6
Sex			
Male	81.7	0.326	81
Female	84.6	0.321	84
Race/Ethnicity			
American Indian or Alaska Native only	71.6	3.3	65.1
Asian only	86.1	0.938	84.2
Native Hawaiian or Other Pacific Islander only	77.4	5.845	66
Black or African American only	82	0.567	80.9
White only	83.3	0.33	82.7
2 or more races	84.2	1.651	81
Hispanic or Latino	66.7	0.724	65.3
Not Hispanic or Latino	86.5	0.282	85.9
Black or African American only, not Hispanic or Latino	82.1	0.576	81
White only, not Hispanic or Latino	87.5	0.328	86.9
Age group			

> 글로벌 헬스케어의 데이터 구조

- 엑셀 데이터 : 미국의 헬스케어 센터에서 제공하는 데이터는 엑셀 기반(2010 버전)으로 되어 있으며, 다양한 데이터가 묶여 있는 데이터이다. 버전상의 문제로 2010 이하 버전의 엑셀 문서로 변환이 필요하다.

II. 수집

➤ 글로벌 헬스케어 데이터의 예

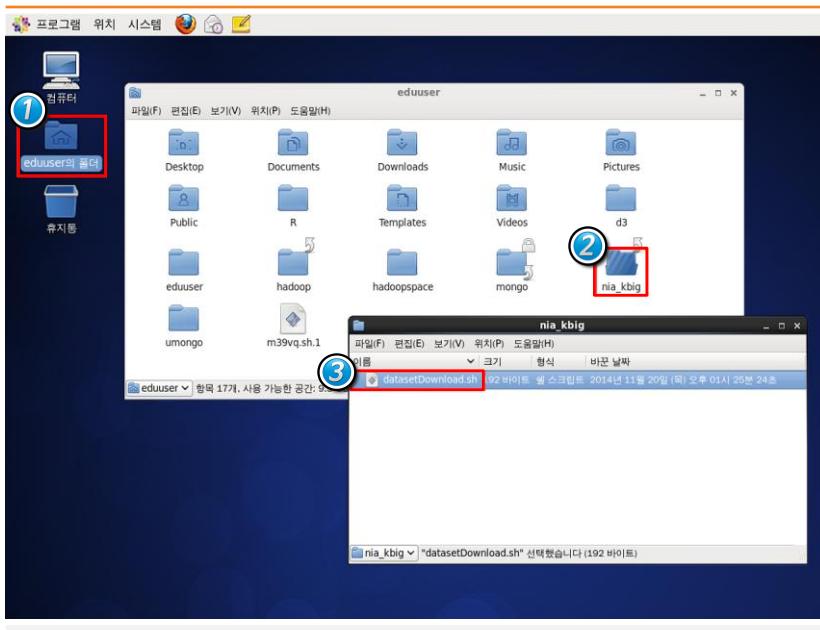
		Year 2008 Percent	Year 2008 Standard Error
1	LHI.Topic...Access.to.Health.Services	NA.	NA..1
2	Objective: AHS-1.1	NA	NA
3	Increase the proportion of persons with medical insurance	NA	NA
4	Population		
5	Target: 100	NA	NA
6	Increase the proportion of persons with medical insurance	NA	NA
7	Total	83.2	0.286
8	Sex	NA	NA
9	Male	81.7	0.326
10	Female	84.6	0.321
11	Race/Ethnicity	NA	NA
12	American Indian or Alaska Native only	71.6	3.3
13	Asian only	86.1	0.938
14	Native Hawaiian or Other Pacific Islander only	77.4	5.845
15	Black or African American only	82	0.567
16	White only	83.3	0.33
17	2 or more races	84.2	1.651
18	Hispanic or Latino	66.7	0.724
19	Not Hispanic or Latino	86.5	0.282
20	Black or African American only, not Hispanic or Latino	82.1	0.576
	White only, not Hispanic or Latino	87.5	0.328

- R에서 읽어 들인 헬스케어 데이터의 예임

> 데이터 수집

- 데이터 저장소에서 서버 로컬로 데이터 셋을 복사해 온다.
 - LHI Access To Health Services Data 2014.xls : 소셜 데이터

> 실습코드 디렉토리로 이동

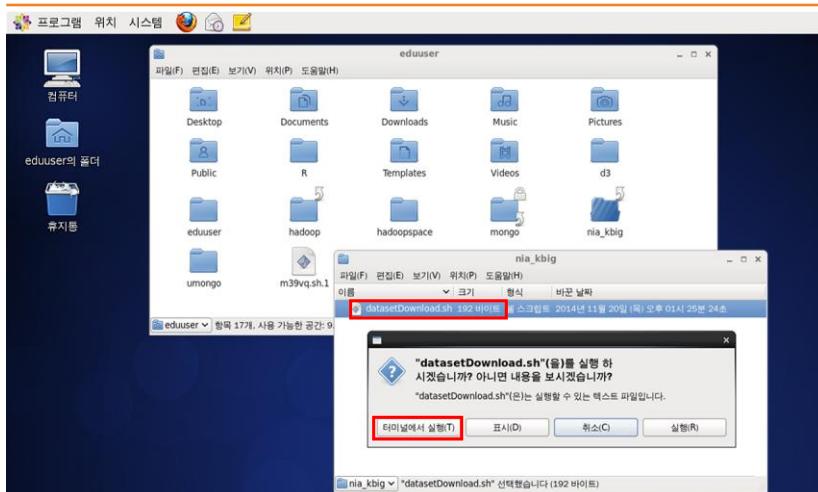


- ① 로그인 후 바탕화면에서 eduuser 폴더를 오픈한다.
- ② nia_kbig 폴더를 오픈한다.
- ③ datasetDownload.sh를 더블클릭하여 실행한다.

II. 수집

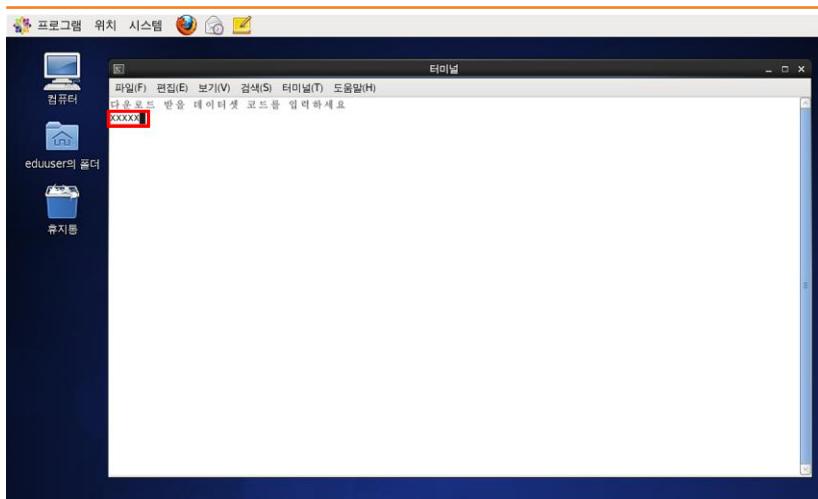
▶ 레파지토리에서 데이터 수집

datasetDownload.sh (원시데이터로 컬서버로 복사)



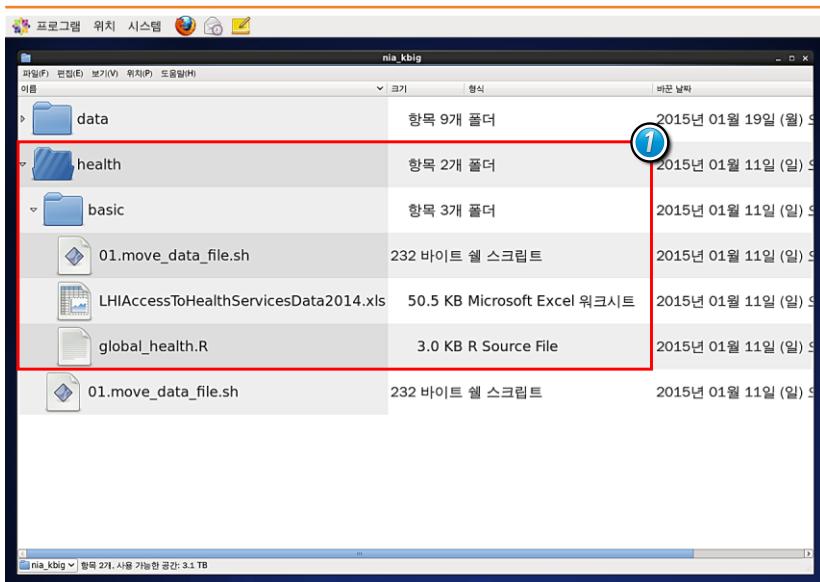
- '터미널에서 실행' 버튼을 클릭한다.

▶ 데이터셋 코드 입력



- 다운로드 받을 데이터셋 코드를 입력 후 엔터

▶ 데이터셋과 실습용 쉘 스크립트



- 실습용 데이터셋과 실습용 스크립트를 확인한다.

▶ ① 데이터 및 스크립트

- **01.move_data_file.sh** : 작업 영역 Data 폴더로 자료 이동하는 스크립트
- **global_health.R** : R 분석 스크립트
- **LHIAccessToHealthServicesData2014.xls** : 글로벌 헬스케어 데이터

II. 수집

▶ 데이터 작업 영역 이동 스크립트(01.move_data_file.sh)

▶ 데이터 작업 공간으로 이동

- 로컬로 수집해 온 데이터를 작업 영역 Data 폴더로 자료를 이동하는 스크립트

01.move_data_file.sh

```
01.#!/bin/bash  
02. #Social Data file define  
03. TARGET_HEALTH=/home/eduuser/nia_kbig/health/basic/LHIA*.xls  
04.  
05. # 작업 디렉토리 정의  
06. LOCAL_DIR=/home/eduuser/nia_kbig/data/  
07. mv $TARGET_HEALTH $LOCAL_DIR  
08.
```



- 분석 원시 데이터 이동 스크립트 소스(01.move_data_file.sh)
- 라인 01~03 : 이동시킬 데이터 파일과 파일의 위치를 “TARGET_HEALTH”로 정의하며, 기호 “#”은 주석을 의미한다.
- 라인 05~06 : 데이터를 이동시킬 위치 정보를 “LOCAL_DIR”이라는 이름으로 기록한다.
- 라인 07 : mv 명령을 이용하여 분석할 데이터를 소셜 폴더에서 분석 폴더로 이동시킨다.

▶ 수집 데이터 셋 작업 영역 폴더 이동

- 로컬로 수집해 온 데이터를 작업 영역 Data 폴더로 자료를 이동하는 스크립트

The screenshot shows a terminal window with the following content:

```
[eduuser@cm01 basic]$ ll
合계 60
-rwxr-xr-x 1 eduuser eduuser 232 2015-01-11 22:12 01.move_data_file.sh
-rw-r--r-- 1 eduuser eduuser 51712 2015-01-11 21:50 LHIAccessToHealthServicesD
ata2014.xls
-rw-r--r-- 1 eduuser eduuser 3050 2015-01-11 23:32 global_health.R
[eduuser@cm01 basic]$ ./01.move_data_file.sh
```

The command `./01.move_data_file.sh` is highlighted with a red rectangle.

로컬에 원시 데이터를 작업 영역 폴더로 이동시킨다.

`./ 01.move_data_file.sh` 입력 후 엔터





III 가공

개요	23
데이터 가공 R 스크립트	28



가공

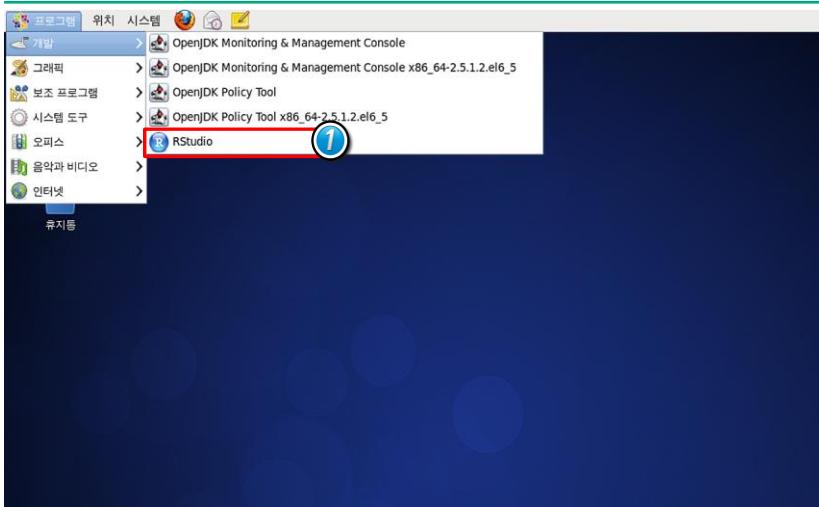
▶ 개요

글로벌 헬스케어 데이터는 다양한 분석 데이터와 함께 원시 데이터를 포함하고 있다. 그렇기 때문에 분석에 필요한 데이터를 추출하는 것이 중요하다. 특히, 엑셀 데이터는 여러 개의 시트를 포함하고 있기 때문에 원하는 데이터가 포함되어 있는 시트를 가져오는 것이 중요하다. 분석을 위해서는 엑셀 시트별로 데이터를 가져오고, 서로 다른 데이터를 추출하는 가공단계가 중요하다.

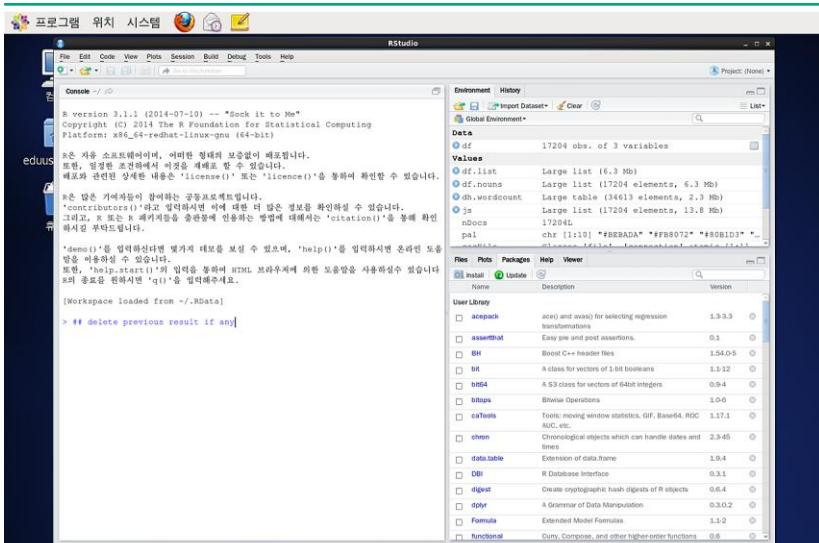
▶ 가공 방법

- **데이터 파일 재저장** : 글로벌 헬스케어 데이터는 MS 엑셀의 2010 버전을 사용하고 있기 때문에 이를 재 구성하기 위한 파일 버전 교체가 필요하다. 제공되는 데이터는 이미 MS 엑셀 2010 버전에서 MS엑셀 2007 버전으로 파일 버전을 변환한 파일이다.
- **데이터 가공 준비** : R에서 데이터 가공을 위한 라이브러리 리스트를 확인하고, 해당 라이브러리를 설치한다.
- **엑셀 시트별 불러오기** : 엑셀은 여러 개의 시트로 구성되어 다양한 데이터를 포함하고 있다. 그렇기 때문에 R에서 이를 불러오기 위해서는 XLSX라는 엑셀 라이브러리가 필요하다.
- **가공 분석을 위해**, 프로그래밍 도구인 R을 실행한다. R은 10,000 줄 이상의 데이터 처리 제약이 있기 때문에 대용량의 글로벌 데이터 처리를 위해서는 Map Reduce와 결합하여 처리한다.

▶ 데이터 가공



- ① 왼쪽 상단의 [“프로그램” 클릭 → “개발” 클릭 → “RStudio” 클릭]으로 분석 도구인 R을 실행한다.

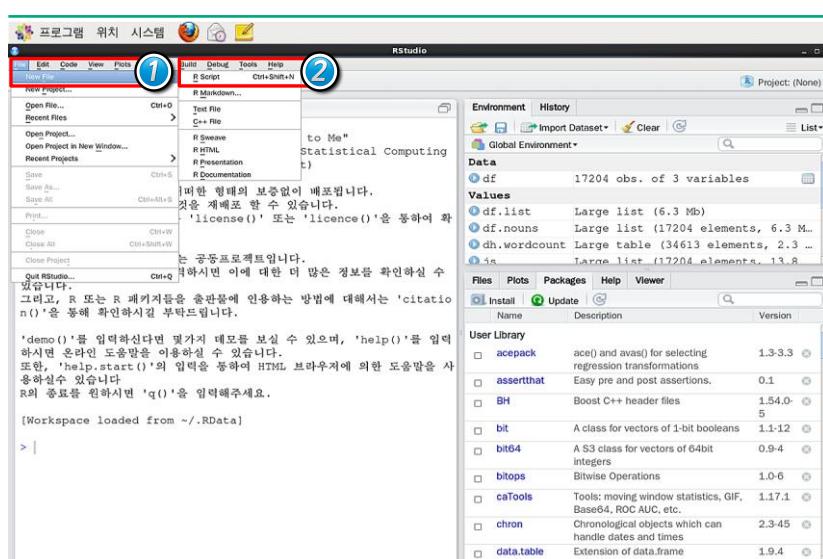


2. 글로벌 데이터 분석을 위한 샘플 데이터에서 분석에 필요한 시트와 데이터 정보를 확인한다.

III. 가공

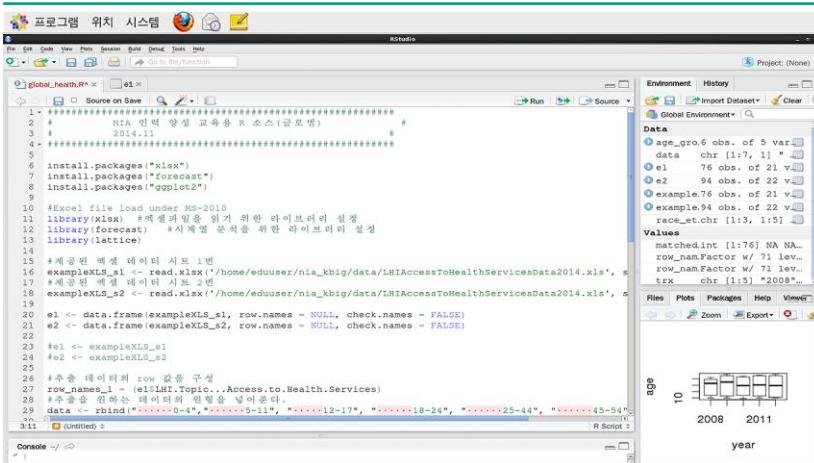
Z3	Age group			
24	<18	91	0.402	90.2
25	0-4	92.6	0.578	91.5
26	5-11	91.1	0.485	90.1
27	12-17	89.4	0.541	88.4
28	18-44	75.6	0.422	74.8
29	18-24	71	0.789	69.5
30	25-44	77.2	0.443	76.4
31	45-64	86.4	0.345	85.8
32	45-54	85.1	0.457	84.2
33	55-64	88.2	0.442	87.3
34	Educational attainment (25 years and over)			
35	< High school	56.9	0.922	55.1
36	High school	78.5	0.532	77.5
37	Some college	83.3	0.582	82.2
38	Associates degree	87	0.637	85.7
39	4-year college degree	91.6	0.398	90.8
40	Advanced degree	95.5	0.393	94.8
41	Family income (percent poverty threshold)			
42	AHS-1,1 / AHS-3			
43	준비			

- 엑셀 파일은 앞의 그림에서와 같이, 하나 이상의 시트를 가지고 있으며, 각각의 시트는 다양한 데이터들을 포함하고 있다.



The screenshot shows the RStudio interface. The 'File' menu is open, with the 'R Script' option highlighted. The main workspace shows some R code and data frames. The 'User Library' pane on the right lists several packages like acepack, assertthat, BH, bit, bit64, bitops, caTools, chron, and data.table.

- ① ② 소셜 트위터 데이터에서 JSON 데이터 인식과 분리, 한글 데이터 분석에 필요한 가공을 위해 프로그램 작업 파일("New File" 클릭 > "R_Script" 클릭)을 선택한다.



4. 분석에 필요한 라이브러리 파일을 설치하기 위해, 필요한 라이브러리를 작성하고 실행한다.

- #주) R 프로그램 분석을 위한 사전 라이브러리 설치는 install.package("라이브러리 이름")으로 설치하거나 오른쪽 하단의 패널을 이용하여 설치한다. 설치할 패키지 리스트는 아래와 같다. 작성된 줄의 끝에서 “Ctrl+ Enter”를 입력하여 실행한다.

```

01. #라이브러리 리스트
02. install.packages("xlsx")      #엑셀 파일 처리를 위한 라이브러리
03. install.packages("forecast")  #시계열 분석 처리를 위한 라이브러리
04. install.packages("ggplot2")   #그래프 출력 처리를 위한 라이브러리

```

III. 가공

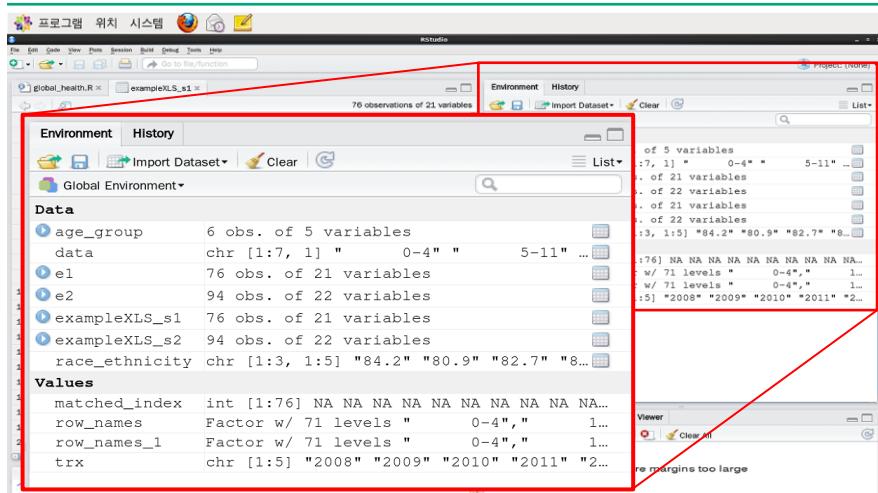
```
global_health.R
6 install.packages("xlsx")
7 install.packages("forecast")
8 install.packages("ggplot2")
9
10 #Excel file load under MS-2010
11 library(xlsx)      #엑셀파일을 읽기 위한 라이브러리 설정
12 library(forecast)  #시계열 분석을 위한 라이브러리 설정
13 library(lattice)   #그래프 처리를 위한 라이브러리 설정
14
15 #제공된
16 example(global_health.R)
17 #제공된
18 example(global_health.R)
19 |   6 install.packages("xlsx")
20 |   7 install.packages("forecast")
21 |   8 install.packages("ggplot2")
22 |
23 |   e1 <- 10 #Excel file load under MS-2010
24 |   e2 <- 11 library(xlsx)      #엑셀파일을 읽기 위한 라이브러리 설정
25 |   #주출 12 library(forecast)  #시계열 분석을 위한 라이브러리 설정
26 |   row.names 13 library(lattice)   #그래프 처리를 위한 라이브러리 설정
27 |   #주출은 14
28
29 data <- rbind(e1$Topic...Access.to.Health.Services)
30 row.names = (e1$LH1.Topic...Access.to.Health.Services)
31 matched_index = (match(row.names, data)) #NA를 제외한 행번호 확인
32 data.frame(matched_index) #매칭된 인덱스 번호를 확인 한다.
33
34 #확인된 행번호의 내용을 추출하여 하나의 데이터 구조로 통합, 또한 일정한 순서로 필요한 열별 데이터를
19:1 | R Script
```

5. 설치된 라이브러리를 프로그램에 이용하기 위해 “library(‘이름’)”을 이용하여 불러온다.

- #주) 설치된 라이브러리들을 불러오는 리스트는 아래와 같다.

```
01. #Excel file load under MS-2010
02. library(xlsx)      #엑셀파일을 읽기 위한 라이브러리 설정
03. library(forecast)  #시계열 분석을 위한 라이브러리 설정
04. library(lattice)   #그래프 처리를 위한 라이브러리 설정
```

▶ 데이터 가공 R 스크립트(global_health.R)



- 분석에 필요한 엑셀의 시트 정보를 확인하고, 시트를 불러온다.
- 불러온 데이터는 각각 데이터 프레임으로 구조화하여 관리한다.

```

01. #제공된 엑셀 데이터 시트 1번
02. exampleXLS_s1 <- read.xlsx('/home/eduuser/nia_kbig/data/
  ↪ LHIAccessToHealthServicesData2014.xls', sheetIndex=1)
03. #제공된 엑셀 데이터 시트 2번
04. exampleXLS_s2 <- read.xlsx('/home/eduuser/nia_kbig/data/
  ↪ LHIAccessToHealthServicesData2014.xls', sheetIndex=2)
05.
06. e1 <- data.frame(exampleXLS_s1, row.names = NULL, check.names = FALSE)
07. e2 <- data.frame(exampleXLS_s2, row.names = NULL, check.names = FALSE)

```



- 분석 원시 데이터 이동 스크립트 소스(global_health.R)
- 라인 01~04 : 분석을 위한 엑셀 데이터 파일의 시트별로 분리하여 불러오기 위해 시트 번호를 표시하고 각각의 시트 정보를 읽어들인다.
- 라인 06 : 읽어들인 시트 1번의 데이터를 분석을 위해 데이터 프레임 구조로 변경한다.
- 라인 07 : 읽어들인 시트 2번의 데이터를 분석을 위해 데이터 프레임 구조로 변경한다.



IV 저 장

개요

31

R Studio 활용 저장

32

> 개요

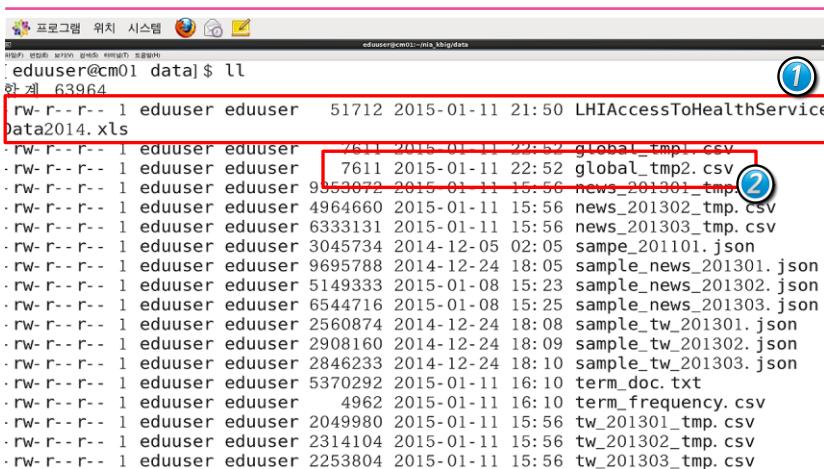
글로벌 헬스케어 데이터는 엑셀의 시트별로 분리하여 읽어 들인 데이터를 임시로 저장하고, 계속적인 분석을 위해 사용할 수 있다. 그렇기 때문에 저장은 임시 파일의 의미를 가지고 있는 형태의 파일명과 엑셀의 데이터 구조를 유지하는 테이블(Table) 구조로 저장한다. 또한 필요할 경우에는 추가적으로 NoSQL이나 관계형 데이터베이스에 별도로 저장하여 관리 할 수 있다.

> 저장 방법

- **가공된 데이터 임시 저장** : 글로벌 헬스케어 엑셀 데이터를 각각의 시트별로 분리하여 임시 파일로 저장한다.
- **저장 파일의 구조 정의** : 다양한 데이터를 포함하고 있기 때문에 Table 구조로 데이터를 저장한다.
- **소스 저장** : 작성 중인 글로벌 헬스케어 분석 프로그램을 저장한다.

> R Studio 활용 저장(global_health.R)

> 데이터 저장



```
eduuser@cm01 data]$ ll
한계 63964
rw-r--r-- 1 eduuser eduuser 51712 2015-01-11 21:50 LHIAccessToHealthService
data2014.xls
rw-r--r-- 1 eduuser eduuser 7611 2015-01-11 22:52 global_tmp1.csv
rw-r--r-- 1 eduuser eduuser 9853072 2015-01-11 15:56 news_201301_tmp.csv
rw-r--r-- 1 eduuser eduuser 4964660 2015-01-11 15:56 news_201302_tmp.csv
rw-r--r-- 1 eduuser eduuser 6333131 2015-01-11 15:56 news_201303_tmp.csv
rw-r--r-- 1 eduuser eduuser 3045734 2014-12-05 02:05 sample_201101.json
rw-r--r-- 1 eduuser eduuser 9695788 2014-12-24 18:05 sample_news_201301.json
rw-r--r-- 1 eduuser eduuser 5149333 2015-01-08 15:23 sample_news_201302.json
rw-r--r-- 1 eduuser eduuser 6544716 2015-01-08 15:25 sample_news_201303.json
rw-r--r-- 1 eduuser eduuser 2560874 2014-12-24 18:08 sample_tw_201301.json
rw-r--r-- 1 eduuser eduuser 2908160 2014-12-24 18:09 sample_tw_201302.json
rw-r--r-- 1 eduuser eduuser 2846233 2014-12-24 18:10 sample_tw_201303.json
rw-r--r-- 1 eduuser eduuser 5370292 2015-01-11 16:10 term.doc.txt
rw-r--r-- 1 eduuser eduuser 4962 2015-01-11 16:10 term_frequency.csv
rw-r--r-- 1 eduuser eduuser 2049980 2015-01-11 15:56 tw_201301_tmp.csv
rw-r--r-- 1 eduuser eduuser 2314104 2015-01-11 15:56 tw_201302_tmp.csv
rw-r--r-- 1 eduuser eduuser 2253804 2015-01-11 15:56 tw_201303_tmp.csv
```

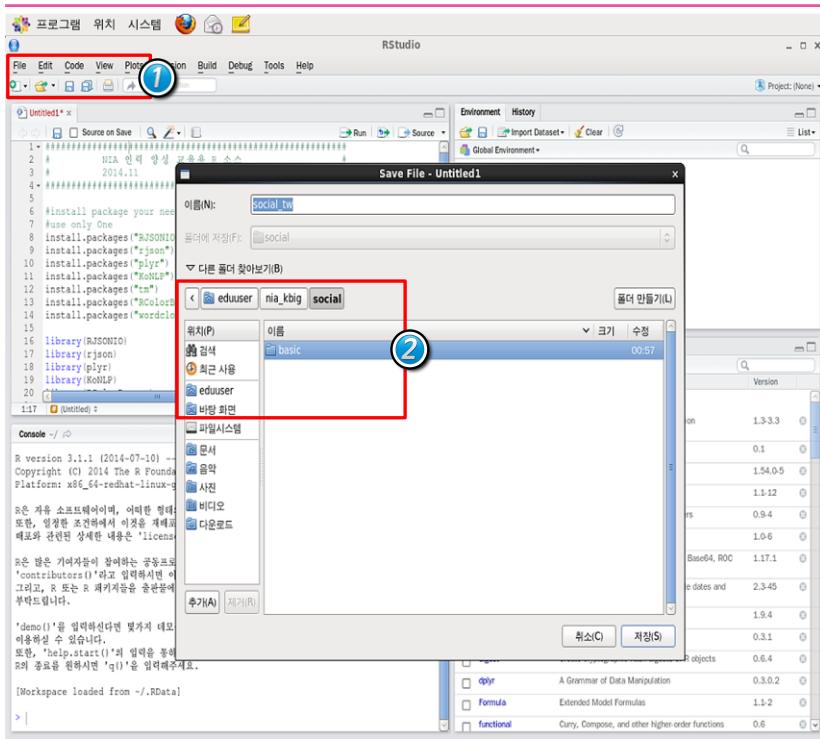
1. 안정적인 분석을 위해, 가공된 데이터를 파일로 저장한다. 저장된 파일은 다른 분석 프로그램으로도 사용할 수 있도록 “CSV” 파일로 저장한다.
- #주) 저장된 파일은 실행시켜 놓은 터미널을 통해 확인할 수 있다. ①은 저장된 파일과 위치에 해당하는 파일이고, ②는 R에서 실행된 CSV로, 임시 저장되는 명령어의 실행 정보이다.

```
01. #각 시트에 대한 데이터를 별도로 분리 저장
02. write.table(e1, file="/home/eduuser/nia_kbig/data/global_tmp1.csv",
   ↵ append=FALSE, quote=FALSE, sep = "", row.names=FALSE)
03. write.table(e2, file="/home/eduuser/nia_kbig/data/global_tmp2.csv",
   ↵ append=FALSE, quote=FALSE, sep = "", row.names=FALSE)
```



- 분석 원시 데이터 이동 스크립트 소스(global_health.R)
- 라인 02 : 데이터 프레임으로 가공된 시트 1번의 데이터를 테이블 구조로 변경하여 CSV로 저장한다.
- 라인 03 : 데이터 프레임으로 가공된 시트 2번의 데이터를 테이블 구조로 변경하여 CSV로 저장한다.

IV. 저장



2. 글로벌 헬스케어 분석을 위해 작성 중인 프로그램 소스를 저장한다.

- #주) 작성 중인 프로그램 소스를 저장하는 방법은 메뉴의 “File” → “Save”를 이용하거나 도구상자의 저장 아이콘을 이용한다. 저장 시 저장 위치는 eduuser라는 폴더를 선택하여 하위 폴더를 따라 저장할 위치를 이동하여 최종적으로 “basic”을 선택한다. 파일명은 본인이 작성한다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화

W



V 분석

개요

37

R Studio 활용 분석

38

V 분석

> 개요

글로벌 데이터는 일차적으로 통계 또는 분석된 데이터를 제공하는 경우가 많기 때문에 해당 데이터를 어떻게 분리하고, 선별하는지가 매우 중요하다. 또한 분리과정에서 데이터의 형식(숫자형, 문자형, 논리형 등)에 따라 분석의 가능여부가 결정되기 때문에 원하는 데이터의 분리와 해당 데이터의 분석 상태에 대한 이해가 필요하다. 글로벌 헬스케어 데이터는 분석된 데이터를 가지고 나이에 따른 신뢰 상관 관계를 분석한다.

> 데이터 분석 방법

- **상관 분석** : 두 개 이상의 시계열 데이터 사이의 상관관계를 계산하는 통계 기법을 사용한다.
- **추이 분석** : 특정 구간에 대한 상태와 변화를 추적하고, 예상 변화를 추론하는 휴리스틱 분석 기법을 사용한다.

▶ R Studio 활용 분석

▶ 데이터 불러오기

글로벌 헬스케어 데이터

	Population	Year 2008 Percent	Year 2008 Standard Error
Target: 100			
Increase the proportion of persons with medical insurance			
Total			
Sex			
Male			
Female			
Race/Ethnicity			
American Indian or Alaska Native only			
Asian only			
Native Hawaiian or Other Pacific Islander only			
Black or African American only			
White only			
2 or more races			
Hispanic or Latino		66.7	0.724
Not Hispanic or Latino		86.5	0.282
Black or African American only, not Hispanic or Latino		82.1	0.576
White only, not Hispanic or Latino		87.5	0.328
Age group			

- Data 저장 폴더를 클릭하여 복사된 글로벌 헬스케어 데이터를 오픈오피스로 불러들일 수 있다. 해당 데이터의 내용과 활용을 위한 데이터의 이해를 쉽게 진행할 수 있다.

▶ 데이터 분석

- #주) 앞의 작성 중인 R 프로그램 소스에 이어서 작업한다. 작업 내용은 아래와 같다.

The screenshot shows the RStudio interface. The top panel displays the global_health.R script with several lines of R code. The bottom panel shows the R console with the following command history:

```

Console ~/nia_kbig/health/basic/ ↵
w.names=FALSE)
> write.table(e1, file="/home/eduuser/nia_kbig/data/global_tmp2.csv", append=FALSE, quote=FALSE, sep ="", ro
w.names=FALSE)
> write.table(e1, file="/home/eduuser/nia_kbig/data/global_tmpl.csv", append=FALSE, quote=FALSE, sep ="", ro
w.names=FALSE)
> |

```

- 불러들인 엑셀의 시트에서 분석에 필요한 데이터만을 추출하는 작업을 수행한다.
이를 통해, 시트 내에 섞여 있는 데이터 중에서 연령대 데이터를 추출할 수 있다.

```

01. #추출 데이터의 row 값을 구성
02. row_names_1 = (e1$LHI.Topic...Access.to.Health.Services)
03. #추출을 원하는 데이터의 원형을 넣어준다.
04. data <- rbind(" 0-4", " 5-11", " 12-17", " 18-24", " 25-44",
   ↴ " 45-54", " 55-64")
05. row_names = (e1$LHI.Topic...Access.to.Health.Services)
06. matched_index = (match(row_names, data)) #NA를 제외한 행번호 확인한다.
07. data.frame(matched_index) #매칭된 인덱스 번호를 확인한다.

```



- 분석 원시 데이터 이동 스크립트 소스(global_health.R)
- 라인 02 : 분석할 시트 1번의 데이터에서 분석을 위한 Row의 이름 정보를 추출한다.
- 라인 04 : 추출한 Row 이름과 분석할 연령별로 분류된 정보를 찾기 위한 연령 그룹 정보를 문자로 입력한다.
- 라인 05 ~ 06 : 재구성된 Row 이름을 추출하고, 분석을 위해 입력된 이름과 매칭하여 위치 정보를 확인한다.
- 라인 07 : 매칭된 인덱스는 원본 데이터에서 분석 데이터에 대한 인덱스로 프레임으로 변환하여 관리한다.



- 분석 원시 데이터 이동 스크립트 소스(global_health.R)
- 라인 03 : 매칭된 인덱스 정보를 이용하여 분석을 위한 컬럼 위치 정보에 따라 데이터를 추출하여 그룹으로 분류한다. 동일 패턴으로 그룹을 형성하고 있기 때문에 이 패턴을 적용하여 데이터를 분리한다.
- 라인 07 : 추가 분석을 위한 인종별 데이터를 추출하는 과정이며, 연령대로 추출하는 과정과 동일한 패턴으로 인종별 데이터를 추출할 수 있다.
- 라인 11 ~ 15 : 추출된 정보를 재분배하여 분석을 위한 정보를 연령별로 각각 분류하여 배치한다.

2. 매핑된 데이터에서 분석에 필요한 데이터 범위를 선정하고, 데이터를 추출한다.

```

01. #확인된 행번호의 내용을 추출하여 하나의 데이터 구조로 통합, 또한 일정한 순서로 필요한 컬럼 데이터 추출
02. #4번부터 최종 21번 컬럼까지 중에서 4개씩 동일한 패턴으로 데이터 반복
03. age_group <- data.frame(rbind(e1[23, seq(4,21,4)], e1[24, seq(4,21,4)],
  ↪ e1[27, seq(4,21,4)], e1[28, seq(4,21,4)], e1[30, seq(4,21,4)],
  ↪ e1[31, seq(4,21,4)]), row.names = NULL, check.names = FALSE)
04.
05.
06. #미국내 거주하는 인종별 분류 정보
07. race_ethnicity <- as.matrix(rbind(e1[12, seq(4,21,4)], e1[14, seq(4,21,4)],
  ↪ e1[15, seq(4,21,4)]), row.names = NULL, check.names = FALSE)
08.
09.
10. # convert factor to numeric for convenience
11. age_group$NA..2 <- as.numeric(age_group$NA..2)
12. age_group$NA..6 <- as.numeric(age_group$NA..6)
13. age_group$NA..10 <- as.numeric(age_group$NA..10)
14. age_group$NA..14 <- as.numeric(age_group$NA..14)
15. age_group$NA..18 <- as.numeric(age_group$NA..18)
16.

```



1

2



VI 시각화

개요	43
분석 데이터 시각화	45
데이터 분석	46

VI

시각화

> 개요

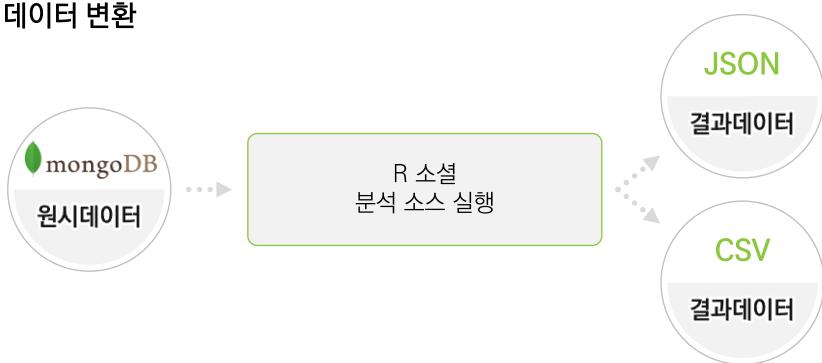
분석 결과는 다양한 방법으로 시각화하여 분석할 수 있으며, 이를 통해 데이터의 변화 및 분포를 해석하고, 데이터에 대한 분석 효과를 국내의 산업 또는 서비스 분야에 적용할 수 있다. 이를 위해 이미 분석되었거나 분석한 데이터의 상태와 추이에 대한 시각화 분석이 중요하며, 변화된 정보의 해석이 중요하다. 따라서 다양한 데이터에서 분석에 필요한 데이터만을 추출하고, 이를 시각화하는 것은 고도의 데이터 해석 능력이 요구된다.



> 시각화 방법 및 활용 기술

- **분석 결과에 대한 그래프 설정** : 추이분석에 유리한 시각화 도구는 박스 그래프 또는 라인 그래프로 출력하는 것이다.
- **저장된 포맷에 맞는 그래프 도구 설정** : 저장된 데이터를 이용하여 시각화할 수 있는 도구도 다양하기 때문에 이를 위해 시각화 도구를 선택한다.
- **추이해석** : 시각화된 박스 그래프의 추이 간격과 구간 이동 상태에 대한 의미를 해석하고, 미래 예측을 위한 추이 변화를 추론한다.

▶ 데이터 변환



> 분석 데이터 시각화(global_health.R)

> 데이터 시각화

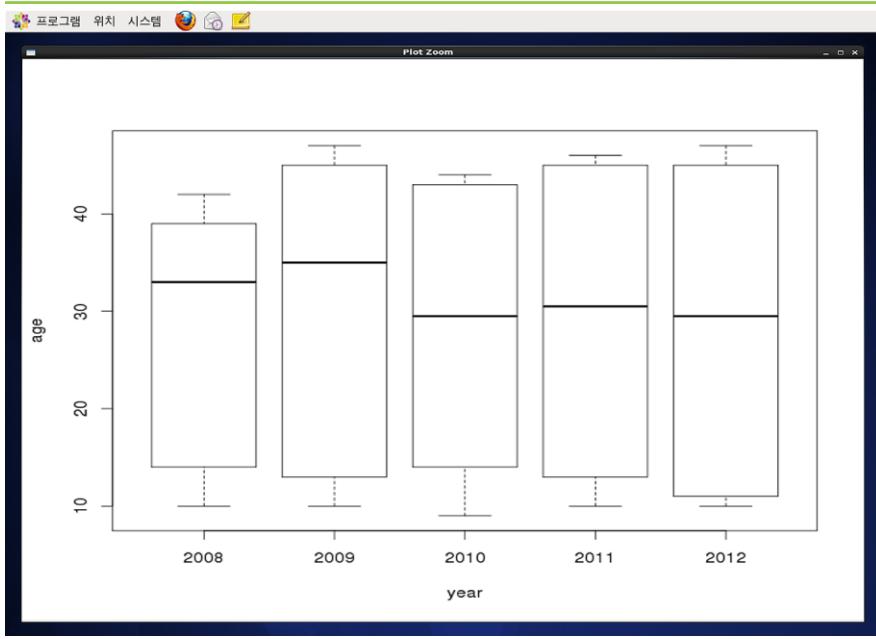
1. 시각화를 위한 데이터 범위와 그래프를 선정한다.
 - #주) 그래프는 박스 그래프로 출력하여 수치 데이터의 연도별 변화와 구간 변화에 대한 추이와 추론을 할 수 있다.

```
01. #데이터를 시계열 자료 형식으로 변환  
02. trx <- c("2008", "2009", "2010", "2011", "2012")  
03. #헤더 정보 삭제, 헤더는 el  
04. names(age_group) <- trx  
05. boxplot(age_group[, 1:5], xlab = "year", ylab = "age")
```



- 분석 원시 데이터 이동 스크립트 소스(global_health.R)
- 라인 02 : 시각화를 위한 그래프의 X축 변량을 표시하기 위해 각각의 변량 정보를 입력한다.
- 라인 04 : X축 변량 정보를 분류된 연령 정보의 헤더 정보로 표현하기 위해 매핑한다.
- 라인 05 : 박스 그래프를 이용하여 전체 연령에 대한 연도별 구간 변화율을 표현하기 위한 분류 데이터 정보와 X축, Y축의 이름을 입력한다.

▶ 데이터 분석



- **주요 분석 :** 미국의 헬스케어 시장이 점차 확대되고 있지만 2010년을 기점으로 헬스케어에 대한 연령별 이해도와 체험에 따른 변화가 거의 없는 형태로 발전되고 있다.
- 미국의 헬스케어 신뢰도는 나이가 많은 계층인 40대 이후에서 가장 높으며, 2010년 이후에 30대 이전의 나이에도 건강에 대한 관심과 서비스의 관심이 증가하고 있음을 알 수 있다.
- **국내 연계 추론 분석 :** 국내의 헬스케어 서비스는 아직까지 소비층이 활성화되어 있지 않지만 헬스케어 서비스에 대한 인지도와 신뢰도는 미국과 같이 2008년의 신뢰도 형태를 취하고 있다. 특히, 국내의 고령화가 빠르게 진행되면서 건강관리와 유지를 위한 헬스케어 서비스의 관심과 신뢰도는 급격하게 증가할 것으로 판단된다.



VII 예제문제

예제 문제1

49

예제 문제2

50

예 / 제 / 문 / 제

예제 1

헬스케어 데이터에서 미국내 인종별 서비스 신뢰도를 분석하라.

- 미국의 헬스케어 서비스에 대한 신뢰도를 측정한 데이터를 이용하여, 미국내 인종별 서비스에 대한 신뢰도의 추이를 분석하고, 박스 그래프로 시각화하라.

- 미국내 헬스케어 데이터에서 인종별 신뢰도 데이터를 추출하고, 원본 데이터와 분리한다.
- 추출된 데이터에서 연도별 분석된 데이터에 대한 패턴을 확인하고 필요 데이터를 연도별에서 추출한다.
- 인종별로 추출된 연도 데이터를 데이터 프레임으로 구조화시키고 저장한다.
- 데이터 추이 분석을 위해 박스 그래프를 이용하여 시각화한다.

예제 2

미국의 헬스케어 서비스에 대한 성별 신뢰도를 분석하라.

- 미국의 헬스 케어 서비스에 대한 신뢰도를 측정한 데이터를 이용하여, 미국 내 거주하는 성별 서비스에 대한 신뢰도의 추이를 분석하고, 박스 그래프로 시각화 하라.
 - 미국내 헬스케어 데이터에서 성별 신뢰도 데이터를 추출하고, 원본 데이터와 분리한다.
 - 추출된 데이터에서 연도별 분석된 데이터에 대한 패턴을 확인하고 필요 데이터를 연도별에서 추출한다.
 - 성별로 추출된 연도 데이터를 데이터 프레임으로 구조화시키고 저장한다.
 - 데이터 추이 분석을 위해 박스 그래프를 이용하여 시각화한다.



글로벌



Intermediate Level

중급과정







개요

55

I

개요

> 개요

글로벌 데이터는 공공과 연구 데이터로 구분되며, 세계적으로 헬스케어와 빅데이터 분석의 관심도가 높아지고 있기 때문에 미국의 헬스케어 연구 센터의 데이터를 이용한다. 해당 데이터에 포함되어 있는 데이터의 일부를 추출하는 방법과 추출된 데이터를 이용하여 2007년 이후의 연령별 헬스케어 우선 대상자에 대한 신뢰도를 알아보고자 한다. 이를 통해 국내 헬스케어 서비스에 대한 연령별 신뢰도와 비교하여 헬스케어 서비스를 재점검할 수 있다.

> 활용 데이터

- LHI Access To Health Services Data 2014 : 미국 헬스케어 센터(LHIA)

> 선행학습

- 리눅스 – 파일시스템 구조, 쉘 명령어, 쉘 스크립트 실행 방법
- R 프로그래밍 언어 – 파일 불러오기, 라이브러리 등록, 데이터 함수(프레임, 테이블), 그래프 함수, 제어문(함수 호출, 외부 함수, 함수 정의) 사용방법
- 데이터 구조 – CSV, JSON 데이터 구조, Text 파일 저장 구조 이해
- R 차트 – 내부 차트(막대, 바, 원 등), 외부 차트(막대, 클라우드, 3D, D3 차트) 사용방법

▶ 요구사항

- 엑셀 데이터에서 분석에 필요한 연령별 데이터만을 각각의 시트별로 추출하여 관리한다.
- 신뢰도의 증감에 대한 연령별 시계열 분석을 실시하라.
- 신뢰도의 증감에 대한 평균 증가율을 구하고, 평균 증가율보다 높은 연령과 신뢰도 구간을 찾아라. 또한 시트별로 신뢰도 구간을 함께 시각화하여 비교 한다.

▶ 분석 절차

- 수집된 글로벌 헬스 케어 데이터 셋을 분석 저장소로 복사한다.
- 엑셀 데이터의 구조상 각각의 시트별로 데이터를 불러온다.
- 각각의 시트에서 연령대별로 증감 분포 분석에 필요한 연도별 신뢰 구간 분석 데이터를 추출한다.
- 추출된 데이터는 연령대별로 분류하여 연도별 신뢰 구간에 따른 증가/감소 형태를 분류하고, 분석한다.
- 출력된 그래프를 통해 연령대별 집중도가 가장 높은 세대를 구분하고, 이를 통해 헬스 케어 서비스의 관심도가 높아지는 형태를 분석한다.



- **헬스케어(Healthcare)** : 질병을 치료하고 관리하는 것 외에 의학, 치의학, 약리학, 임상병리 진단학, 간호학 등 건강과 관련된 전문가들이 사람의 건강을 유지하도록 하는 행위



II 수집

개요	59
수집 데이터	60
데이터 수집	62
데이터 작업 영역 이동 스크립트	65



수집

> 개요

글로벌 데이터는 공공데이터를 기준으로 다양한 데이터가 공개되고 있으며, 공개된 데이터는 연도별 기준에 따라 가공, 정제가 되어 있는 데이터들이 존재한다. 그렇기 때문에 수집을 위한 데이터 선정이 중요하고, 이를 수집하는 방법에 따라 접근하는 것이 중요하다. 그래서 글로벌 데이터는 각 국가나 연구기관이 제공하는 API 또는 기업의 수집기를 통해 수집할 수 있으며, 수집된 글로벌 데이터를 통해 사회적 이슈나 비즈니스 분석 등과 같은 다양한 사회적 분석 모델에 적용할 수 있다.

> 수집 방법

- **API 데이터 수집** : 글로벌 데이터의 수집은 API를 이용하여 일정량의 데이터를 정기적으로 수집할 수 있으며, 각 국가별, 연구기관별 데이터 제공 정책에 따라 수집 데이터의 범위를 제한하고 있다.
- **데이터 제공** : 글로벌 데이터는 OpenAPI, 자료수집기 (Crawler), 데이터 구매 등으로 데이터를 수집 할 수 있으며, 실습용 자료는 빅데이터 분석활용 센터에 접속하여 글로벌 헬스케어 데이터 셋을 다운로드 받을 수 있도록 원시 데이터를 제공하고 있다.



용어정리

- **비정형 데이터(Unstructured Data)** : 일정한 규격이나 형태를 지닌 숫자 데이터와 달리 그림, 영상, 문서 등과 같이 서로 다른 형태의 구조화되지 않은 데이터

> 수집 데이터

> 미국 헬스케어 센터 데이터

LHI Topic: Access to Health Services
Objective: AHS-3
Percent of persons with a usual primary care provider

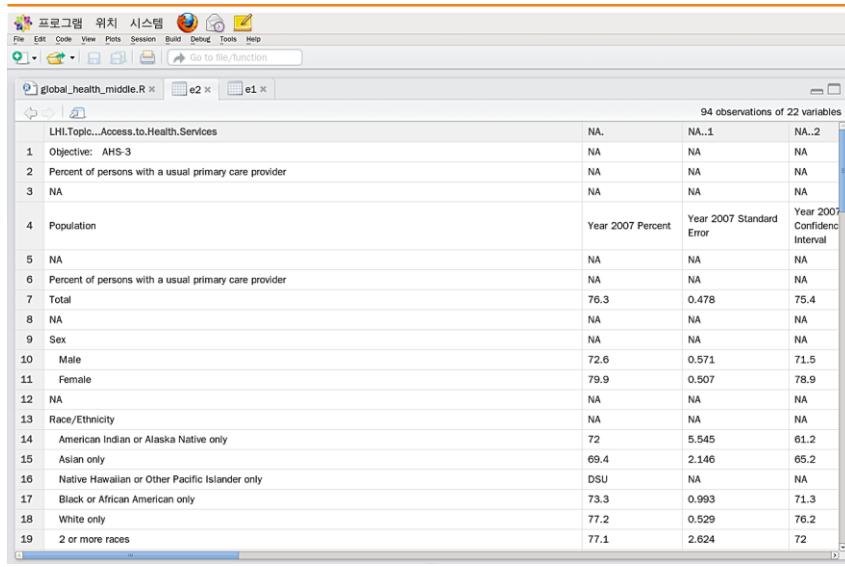
	Population	Year 2007 Percent	Year 2007 Standard Error	Year 2007 Lower 95% Confidence Interval	Year 2007 Upper 95% Confidence Interval
7	Percent of persons with a usual primary care provider				
8	Total	76.3	0.478	75.4	
10	Sex				
11	Male	72.6	0.571	71.5	
12	Female	79.9	0.507	78.9	
14	Race/Ethnicity				
15	American Indian or Alaska Native only	72.0	5.545	61.2	
16	Asian only	69.4	2.146	65.2	
17	Native Hawaiian or Other Pacific Islander only	DSU			

> 글로벌 헬스케어의 데이터 구조

- 엑셀 데이터 : 미국의 헬스케어 센터에서 제공하는 데이터는 엑셀 기반(2010 버전)으로 되어 있으며, 다양한 데이터가 묶여 있는 데이터이다. 버전상의 문제로 2010 이하 버전의 엑셀 문서로 변환이 필요하다. 이 데이터는 AHS-1.1과 AHS-3으로 두 개의 시트를 가지고 있다.

II. 수집

▶ 글로벌 헬스케어 데이터의 예



The screenshot shows the RStudio interface with a data frame titled "LHI.Topic...Access.to.Health.Services". The data frame contains 94 observations across 22 variables. The variables include "Objective: AHS-3", "Percent of persons with a usual primary care provider", "NA", "Population", "Year 2007 Percent", "Year 2007 Standard Error", and "Year 2007 Confidence Interval". The data shows various percentages for different demographic groups, such as 76.3% for Total population and 72.6% for Male.

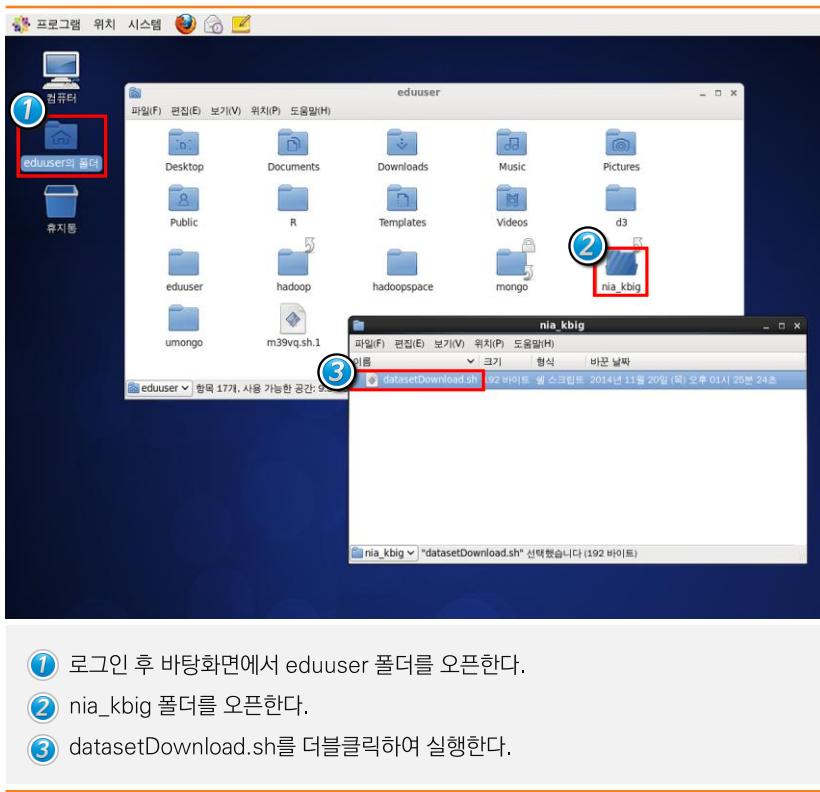
	LHI.Topic...Access.to.Health.Services	NA..1	NA..2
1	Objective: AHS-3	NA	NA
2	Percent of persons with a usual primary care provider	NA	NA
3	NA	NA	NA
4	Population	Year 2007 Percent	Year 2007 Standard Error
5	NA	NA	NA
6	Percent of persons with a usual primary care provider	NA	NA
7	Total	76.3	0.478
8	NA	NA	NA
9	Sex	NA	NA
10	Male	72.6	0.571
11	Female	79.9	0.507
12	NA	NA	NA
13	Race/Ethnicity	NA	NA
14	American Indian or Alaska Native only	72	5.545
15	Asian only	69.4	2.146
16	Native Hawaiian or Other Pacific Islander only	DSU	NA
17	Black or African American only	73.3	0.993
18	White only	77.2	0.529
19	2 or more races	77.1	2.624

- R에서 읽어 들인 헬스케어 데이터 중 AHS-3의 예임

> 데이터 수집

- 데이터 저장소에서 서버 로컬로 데이터 셋을 복사해 온다.
 - LHI Access To Health Services Data 2014.xls : 소셜 데이터

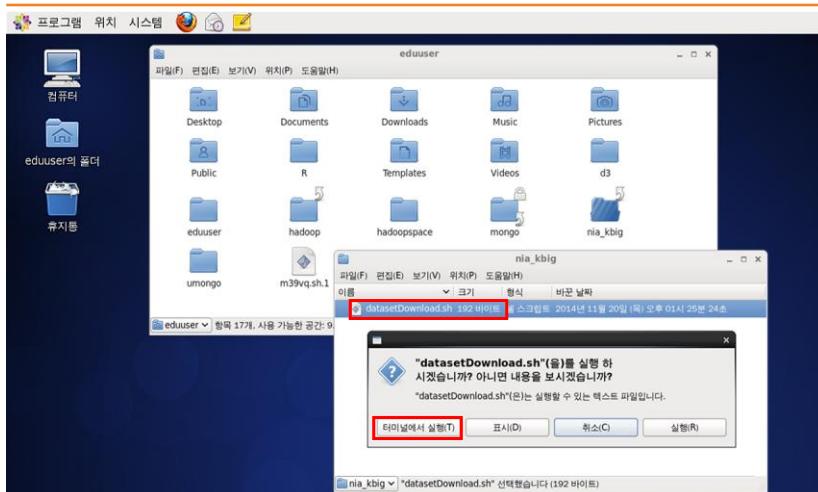
> 실습코드 디렉토리로 이동



II. 수집

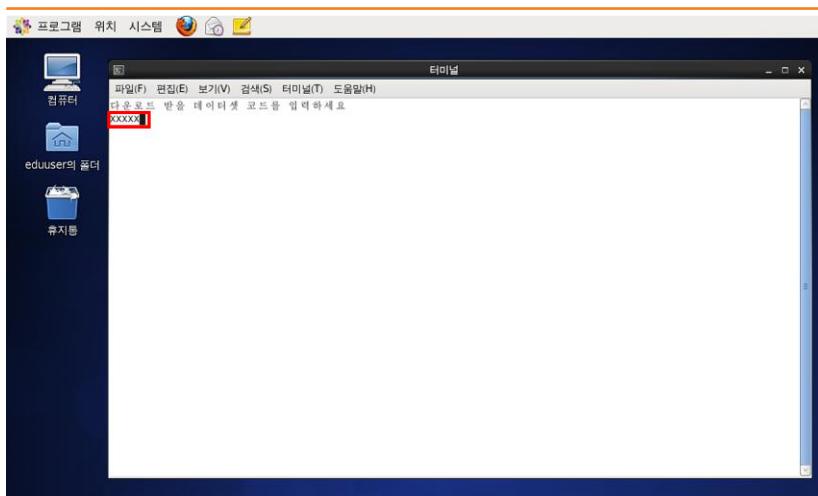
▶ 레파지토리에서 데이터 수집

datasetDownload.sh (원시데이터로 컬서버로 복사)



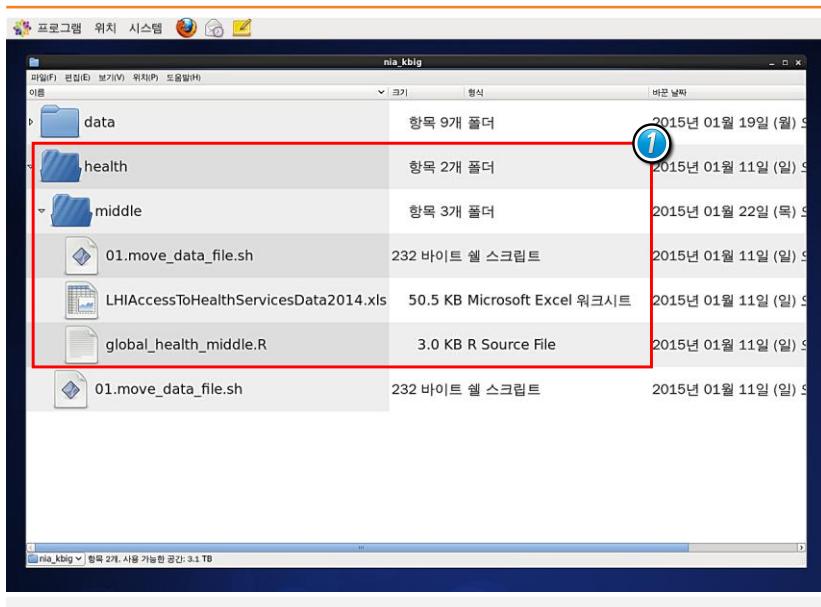
- '터미널에서 실행' 버튼을 클릭한다.

▶ 데이터셋 코드 입력



- 다운로드 받은 데이터셋 코드를 입력 후 엔터

▶ 데이터셋과 실습용 쉘 스크립트



- 실습용 데이터셋과 실습용 스크립트를 확인한다.

▶ ① 데이터 및 스크립트

- **01.move_data_file.sh** : 작업 영역 Data 폴더로 자료 이동하는 스크립트
- **global_health_middle.R** : R 분석 스크립트
- **LHIAccessToHealthServicesData2014.xls** : 글로벌 헬스케어 데이터

II. 수집

▶ 데이터 작업 영역 이동 스크립트(01.move_data_file.sh)

▶ 데이터 작업 공간으로 이동

- 로컬로 수집해 온 데이터를 작업 영역 Data 폴더로 자료를 이동하는 스크립트

01.move_data_file.sh

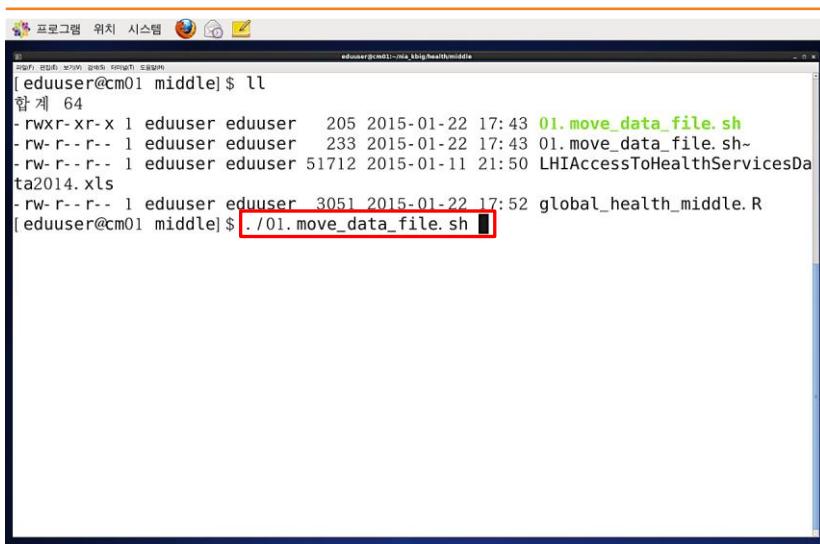
```
01. #!/bin/bash
02. #Social Data file define
03. TARGET_HEALTH=/home/eduuser/nia_kbig/health/middle/LHIA*.xls
04.
05. # 작업 디렉토리 정의
06. LOCAL_DIR=/home/eduuser/nia_kbig/data/
07. mv $TARGET_HEALTH $LOCAL_DIR
08.
```



- 분석 원시 데이터 이동 스크립트 소스(01.move_data_file.sh)
- 라인 01~03 : 이동시킬 데이터 파일과 파일의 위치를 “TARGET_HEALTH”로 정의하며, 기호 “#”은 주석을 의미한다.
- 라인 05~06 : 데이터를 이동시킬 위치 정보를 “LOCAL_DIR”이라는 이름으로 기록한다.
- 라인 07 : mv 명령을 이용하여 분석할 데이터를 소설 폴더에서 분석 폴더로 이동시킨다.

▶ 수집 데이터 셋 작업 영역 폴더 이동

- 로컬로 수집해 온 데이터를 작업 영역 Data 폴더로 자료를 이동하는 스크립트



A screenshot of a Linux terminal window titled 'eduuser@cm01:~/nca_kbig/health/middle'. The window shows a file listing with the command 'll'. One file, '01.move_data_file.sh', is highlighted in green. The user then runs the command '. /01.move_data_file.sh' at the prompt. The terminal window has a dark blue header bar and a light gray body.

```
[eduuser@cm01 middle]$ ll
합계 64
-rwxr-xr-x 1 eduuser eduuser 205 2015-01-22 17:43 01.move_data_file.sh
-rw-r--r-- 1 eduuser eduuser 233 2015-01-22 17:43 01.move_data_file.sh~
-rw-r--r-- 1 eduuser eduuser 51712 2015-01-11 21:50 LHIAccessToHealthServicesData2014.xls
-rw-r--r-- 1 eduuser eduuser 3051 2015-01-22 17:52 global_health_middle.R
[eduuser@cm01 middle]$ ./01.move_data_file.sh
```

로컬에 원시 데이터를 작업 영역 폴더로 이동시킨다.

. /01.move_data_file.sh 입력 후 엔터





III 가공

개요

69

데이터 가공 R 스크립트

74



가공

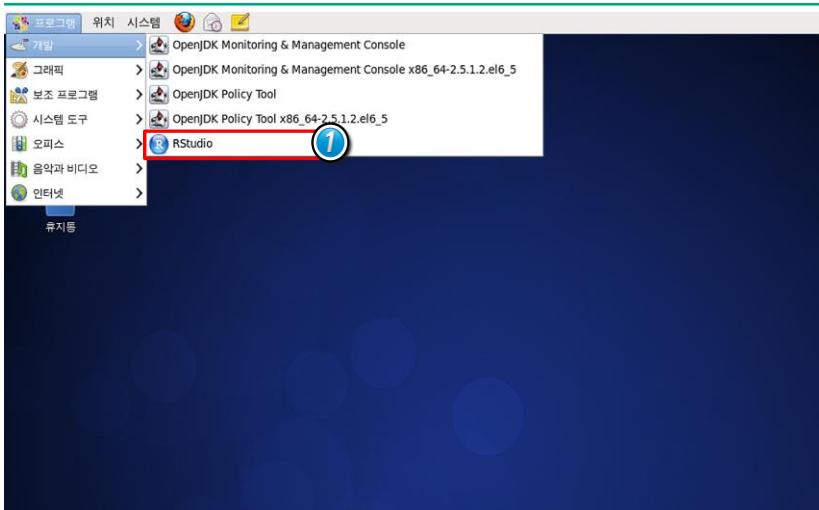
▶ 개요

글로벌 헬스케어 데이터는 다양한 분석 데이터와 함께 원시 데이터를 포함하고 있다. 그렇기 때문에 분석에 필요한 데이터를 추출하는 것이 중요하다. 특히, 엑셀 데이터는 여러 개의 시트를 포함하고 있기 때문에 원하는 데이터가 포함되어 있는 시트를 가져오는 것이 중요하다. 분석을 위해서는 엑셀 시트별로 데이터를 가져오고, 서로 다른 데이터를 추출하는 가공단계가 중요하다.

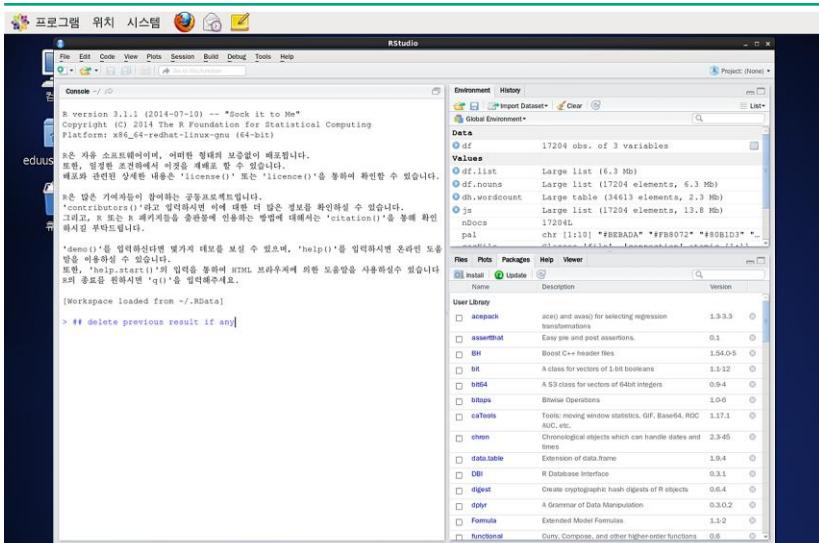
▶ 가공 방법

- **데이터 파일 재저장** : 글로벌 헬스케어 데이터는 MS 엑셀의 2010 버전을 사용하고 있기 때문에 이를 재 구성하기 위한 파일 버전 교체가 필요하다. 제공되는 데이터는 이미 MS 엑셀 2010 버전에서 MS엑셀 2007 버전으로 파일 버전을 변환한 파일이다.
- **데이터 가공 준비** : R에서 데이터 가공을 위한 라이브러리 리스트를 확인하고, 해당 라이브러리를 설치한다.
- **엑셀 시트별 불러오기** : 엑셀은 여러 개의 시트로 구성되어 다양한 데이터를 포함하고 있다. 그렇기 때문에 R에서 이를 불러오기 위해서는 XLSX라는 엑셀 라이브러리가 필요하다.
- **가공 분석을 위해**, 프로그래밍 도구인 R을 실행한다. R은 10,000 줄 이상의 데이터 처리 제약이 있기 때문에 대용량의 글로벌 데이터 처리를 위해서는 Map Reduce와 결합하여 처리한다.

▶ 데이터 가공



- ① 왼쪽 상단의 [“프로그램” 클릭 → “개발” 클릭 → “RStudio” 클릭]으로 분석 도구인 R을 실행한다.



2. 글로벌 데이터 분석을 위한 샘플 데이터에서 분석에 필요한 시트와 데이터 정보를 확인한다.

III. 가공

Z3	Age group			
24	<18	91	0.402	90.2
25	0-4	92.6	0.578	91.5
26	5-11	91.1	0.485	90.1
27	12-17	89.4	0.541	88.4
28	18-44	75.6	0.422	74.8
29	18-24	71	0.789	69.5
30	25-44	77.2	0.443	76.4
31	45-64	86.4	0.345	85.8
32	45-54	85.1	0.457	84.2
33	55-64	88.2	0.442	87.3
34	Educational attainment (25 years and over)			
35	< High school	56.9	0.922	55.1
36	High school	78.5	0.532	77.5
37	Some college	83.3	0.582	82.2
38	Associates degree	87	0.637	85.7
39	4-year college degree	91.6	0.398	90.8
40	Advanced degree	95.5	0.393	94.8
41	Family income (percent poverty threshold)			
42	AHS-1,1 / AHS-3			
43	준비			

- 엑셀 파일은 앞의 그림에서와 같이, 하나 이상의 시트를 가지고 있으며, 각각의 시트는 다양한 데이터들을 포함하고 있다.

The screenshot shows the RStudio interface. The menu bar at the top includes 'File' (with 'New File...' circled in blue), 'Edit', 'View', 'Plots', 'Tools', 'Help', and 'R Script' (with 'Ctrl+Shift+N' circled in blue). The 'File' menu has options like 'Open File...', 'Recent Projects', 'Save', 'Print...', 'Close All', and 'Quit RStudio...'. The 'R Script' option is highlighted. The main workspace shows a data frame with 17204 observations and 3 variables. The 'User Library' pane on the right lists various R packages such as 'acepack', 'assertthat', 'BH', 'bit', 'bit64', 'bitops', 'caTools', 'chron', and 'data.table'.

```

'demo()'를 입력하신다면 몇 가지 테스트를 보실 수 있으며, 'help()'를 입력하신다면 온라인 도움말을 이용하실 수 있습니다.  

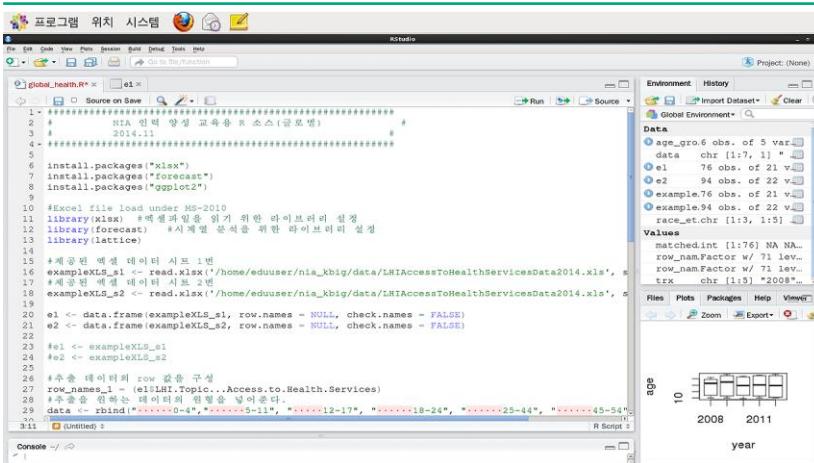
또한, 'help.start()'의 입력을 통하여 HTML 브라우저에 의한 도움말을 사용하실 수 있습니다.  

R의 종료를 원하시면 'q()'을 입력해주세요.  

[Workspace loaded from ~/.RData]
>

```

- ① ② 소셜 트위터 데이터에서 JSON 데이터 인식과 분리, 한글 데이터 분석에 필요한 가공을 위해 프로그램 작업 파일("New File" 클릭 > "R_Script" 클릭)을 선택한다.



4. 분석에 필요한 라이브러리 파일을 설치하기 위해, 필요한 라이브러리를 작성하고 실행한다.

- #주) R 프로그램 분석을 위한 사전 라이브러리 설치는 install.package("라이브러리 이름")으로 설치하거나 오른쪽 하단의 패널을 이용하여 설치한다. 설치할 패키지 리스트는 아래와 같다. 작성된 줄의 끝에서 “Ctrl+ Enter”를 입력하여 실행한다.

```

01. #라이브러리 리스트
02. install.packages("xlsx")      #엑셀 파일 처리를 위한 라이브러리
03. install.packages("forecast")  #시계열 분석 처리를 위한 라이브러리
04. install.packages("ggplot2")   #그래프 출력 처리를 위한 라이브러리

```

III. 가공

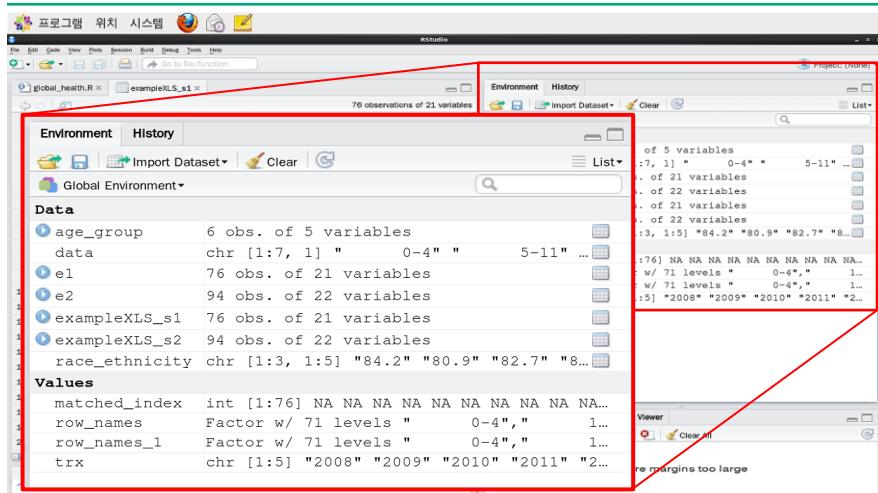
```
6 install.packages("xlsx")
7 install.packages("forecast")
8 install.packages("ggplot2")
9
10 #Excel file load under MS-2010
11 library(xlsx)      #엑셀파일을 읽기 위한 라이브러리 설정
12 library(forecast)  #시계열 분석을 위한 라이브러리 설정
13 library(lattice)   #그래프 처리를 위한 라이브러리 설정
14
15 #제공된
16 example(global_health.R)
17 #제공된
18 example(global_health.R)
19 |       6 install.packages("xlsx")
20 |       7 install.packages("forecast")
21 |       8 install.packages("ggplot2")
22 |
23 |       10 #Excel file load under MS-2010
24 |       11 library(xlsx)      #엑셀파일을 읽기 위한 라이브러리 설정
25 |       12 library(forecast)  #시계열 분석을 위한 라이브러리 설정
26 |       13 library(lattice)   #그래프 처리를 위한 라이브러리 설정
27 row.names = 14
28 #주출
29 data <- rbind(...
30 row.names = (e1$LH1.Topic...Access.to.Health.Services)
31 matched_index = (match(row.names, data)) #NA를 제외한 행번호 확인
32 data.frame(matched_index) #매칭된 인덱스 번호를 확인 한다.
33
34 #확인된 행번호의 내용을 추출하여 하나의 데이터 구조로 통합, 또한 일정한 순서로 필요한 열별 데이터를
19:1 | (Untitled) : R Script
```

5. 설치된 라이브러리를 프로그램에 이용하기 위해 “library(‘이름’)”을 이용하여 불러온다.

- #주) 설치된 라이브러리들을 불러오는 리스트는 아래와 같다.

```
1. #Excel file load under MS-2010
2. library(xlsx)      #엑셀파일을 읽기 위한 라이브러리 설정
3. library(forecast)  #시계열 분석을 위한 라이브러리 설정
4. library(lattice)   #그래프 처리를 위한 라이브러리 설정
```

▶ 데이터 가공 R 스크립트(global_health_middle.R)



- 분석에 필요한 엑셀의 시트 정보를 확인하고, 시트를 불러온다.
- 불러온 데이터는 각각 데이터 프레임으로 구조화하여 관리한다.

```

01. #제공된 엑셀 데이터 시트 1번
02. exampleXLS_s1 <- read.xlsx('/home/eduuser/nia_kbig/data/
  ↳ LHIAccessToHealthServicesData2014.xls', sheetIndex=1)
03. #제공된 엑셀 데이터 시트 2번
04. exampleXLS_s2 <- read.xlsx('/home/eduuser/nia_kbig/data/
  ↳ LHIAccessToHealthServicesData2014.xls', sheetIndex=2)
05.
06. e1 <- data.frame(exampleXLS_s1, row.names = NULL, check.names = FALSE)
07. e2 <- data.frame(exampleXLS_s2, row.names = NULL, check.names = FALSE)

```



- 분석 원시 데이터 이동 스크립트 소스(global_health.R)
- 라인 01~04 : 분석을 위한 엑셀 데이터 파일의 시트별로 분리하여 불러오기 위해 시트 번호를 표시하고 각각의 시트 정보를 읽어들인다.
- 라인 06 : 읽어들인 시트 1번의 데이터를 분석을 위해 데이터 프레임 구조로 변경한다.
- 라인 07 : 읽어들인 시트 2번의 데이터를 분석을 위해 데이터 프레임 구조로 변경한다.



IV 저 장

개요

77

R Studio 활용 저장

78

> 개요

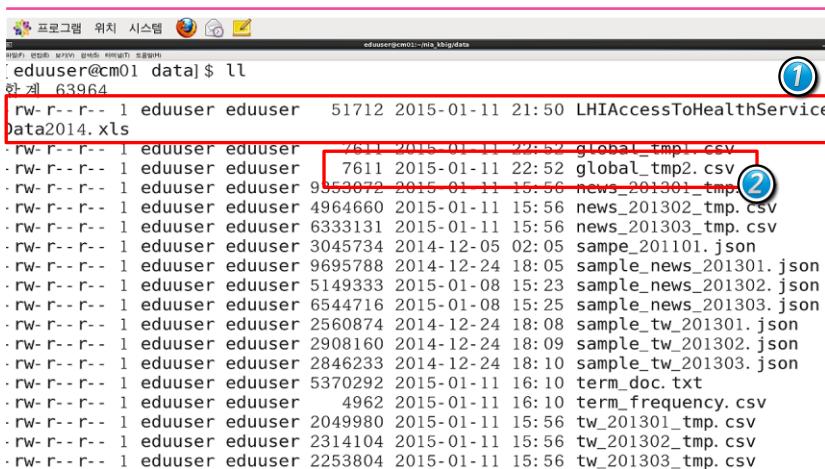
글로벌 헬스케어 데이터는 엑셀의 시트별로 분리하여 읽어 들인 데이터를 임시로 저장하고, 계속적인 분석을 위해 사용할 수 있다. 그렇기 때문에 저장은 임시 파일의 의미를 가지고 있는 형태의 파일명과 엑셀의 데이터 구조를 유지하는 테이블(Table) 구조로 저장한다. 또한 필요할 경우에는 추가적으로 NoSQL이나 관계형 데이터베이스에 별도로 저장하여 관리 할 수 있다.

> 저장 방법

- **가공된 데이터 임시 저장** : 글로벌 헬스케어 엑셀 데이터를 각각의 시트별로 분리하여 임시 파일로 저장한다.
- **저장 파일의 구조 정의** : 다양한 데이터를 포함하고 있기 때문에 Table 구조로 데이터를 저장한다.
- **소스 저장** : 작성 중인 글로벌 헬스케어 분석 프로그램을 저장한다.

> R Studio 활용 저장(global_health_middle.R)

> 데이터 저장



```
eduuser@cm01 data]$ ll
한계 63964
rw-r--r-- 1 eduuser eduuser 51712 2015-01-11 21:50 LHIAccessToHealthService
data2014.xls
rw-r--r-- 1 eduuser eduuser 7611 2015-01-11 22:52 global_tmp1.csv
rw-r--r-- 1 eduuser eduuser 9853072 2015-01-11 15:56 news_201301_tmp.csv
rw-r--r-- 1 eduuser eduuser 4964660 2015-01-11 15:56 news_201302_tmp.csv
rw-r--r-- 1 eduuser eduuser 6333131 2015-01-11 15:56 news_201303_tmp.csv
rw-r--r-- 1 eduuser eduuser 3045734 2014-12-05 02:05 sample_201101.json
rw-r--r-- 1 eduuser eduuser 9695788 2014-12-24 18:05 sample_news_201301.json
rw-r--r-- 1 eduuser eduuser 5149333 2015-01-08 15:23 sample_news_201302.json
rw-r--r-- 1 eduuser eduuser 6544716 2015-01-08 15:25 sample_news_201303.json
rw-r--r-- 1 eduuser eduuser 2560874 2014-12-24 18:08 sample_tw_201301.json
rw-r--r-- 1 eduuser eduuser 2908160 2014-12-24 18:09 sample_tw_201302.json
rw-r--r-- 1 eduuser eduuser 2846233 2014-12-24 18:10 sample_tw_201303.json
rw-r--r-- 1 eduuser eduuser 5370292 2015-01-11 16:10 term.doc.txt
rw-r--r-- 1 eduuser eduuser 4962 2015-01-11 16:10 term_frequency.csv
rw-r--r-- 1 eduuser eduuser 2049980 2015-01-11 15:56 tw_201301_tmp.csv
rw-r--r-- 1 eduuser eduuser 2314104 2015-01-11 15:56 tw_201302_tmp.csv
rw-r--r-- 1 eduuser eduuser 2253804 2015-01-11 15:56 tw_201303_tmp.csv
```

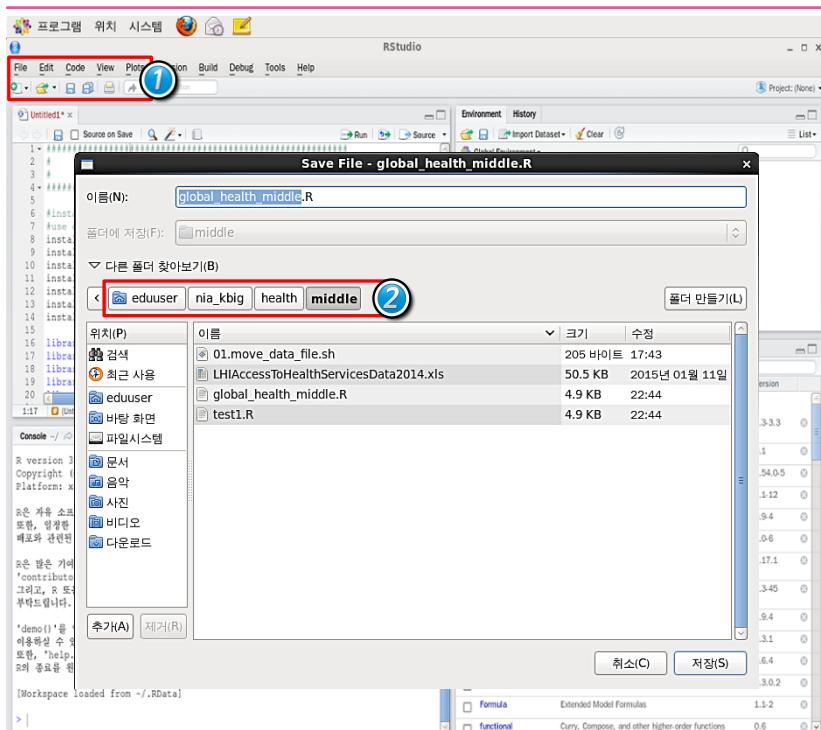
1. 안정적인 분석을 위해, 가공된 데이터를 파일로 저장한다. 저장된 파일은 다른 분석 프로그램으로도 사용할 수 있도록 “CSV” 파일로 저장한다.
- #주) 저장된 파일은 실행시켜 놓은 터미널을 통해 확인할 수 있다. ①은 저장된 파일과 위치에 해당하는 파일이고, ②는 R에서 실행된 CSV로, 임시 저장되는 명령어의 실행 정보이다.

```
01. #각 시트에 대한 데이터를 별도로 분리 저장
02. write.table(e1, file="/home/eduuser/nia_kbig/data/global_tmp1.csv",
   ↪ append=FALSE, quote=FALSE, sep = "", row.names=FALSE)
03. write.table(e2, file="/home/eduuser/nia_kbig/data/global_tmp2.csv",
   ↪ append=FALSE, quote=FALSE, sep = "", row.names=FALSE)
```



- 분석 원시 데이터 이동 스크립트 소스(global_health.R)
- 라인 02 : 데이터 프레임으로 가공된 시트 1번의 데이터를 테이블 구조로 변경하여 CSV로 저장한다.
- 라인 03 : 데이터 프레임으로 가공된 시트 2번의 데이터를 테이블 구조로 변경하여 CSV로 저장한다.

IV. 저장



2. 글로벌 헬스케어 분석을 위해 작성 중인 프로그램 소스를 저장한다.

- #주) 작성 중인 프로그램 소스를 저장하는 방법은 메뉴의 “File” → “Save”를 이용하거나 도구상자의 저장 아이콘을 이용한다. 저장 시 저장 위치는 eduuser라는 폴더를 선택하여 하위 폴더를 따라 저장할 위치를 이동하여 최종적으로 “middle”을 선택한다. 파일명은 본인이 작성한다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화

W





V 분석

개요	83
R Studio 활용 분석	84

V 분석

➤ 개요

글로벌 데이터는 일차적으로 통계 또는 분석된 데이터를 제공하는 경우가 많기 때문에 해당 데이터를 어떻게 분리하고, 선별하는지가 매우 중요하다. 또한 분리과정에서 데이터의 형식(숫자형, 문자형, 논리형 등)에 따라 분석의 가능여부가 결정되기 때문에 원하는 데이터의 분리와 해당 데이터의 분석 상태에 대한 이해가 필요하다. 글로벌 헬스케어 데이터는 분석된 데이터를 가지고 나이에 따른 신뢰 상관 관계를 분석한다.

➤ 데이터 분석 방법

- **상관 분석** : 두 개 이상의 시계열 데이터 사이의 상관관계를 계산하는 통계 기법을 사용한다.
- **추이 분석** : 특정 구간에 대한 상태와 변화를 추적하고, 예상 변화를 추론하는 휴리스틱 분석 기법을 사용한다.

▶ R Studio 활용 분석

▶ 데이터 불러오기

글로벌 헬스케어 데이터

	Population	Year 2008 Percent	Year 2008 Standard Error
Target: 100			
Increase the proportion of persons with medical insurance			
Total			
Sex			
Male			
Female			
Race/Ethnicity			
American Indian or Alaska Native only			
Asian only			
Native Hawaiian or Other Pacific Islander only			
Black or African American only			
White only			
2 or more races			
Hispanic or Latino		66.7	0.724
Not Hispanic or Latino		86.5	0.282
Black or African American only, not Hispanic or Latino		82.1	0.576
White only, not Hispanic or Latino		87.5	0.328
Age group			

- Data 저장 폴더를 클릭하여 복사된 글로벌 헬스케어 데이터를 오픈오피스로 불러들일 수 있다. 해당 데이터의 내용과 활용을 위한 데이터의 이해를 쉽게 진행할 수 있다.

▶ 데이터 분석

- #주) 앞의 작성 중인 R 프로그램 소스에 이어서 작업한다. 작업 내용은 아래와 같다.

```

global_health.R* social_txting_Lv2.R*
25 write.table(el, file="/home/eduuser/nia_kbig/data/global_tmp2.csv", append=FALSE, quote=FALSE, sep ="",
26 |
27 #주출 데이터의 row 값을 구성
28 row_names_1 = (el$LHI.Topic...Access.to.Health.Services)
29 #주출을 원하는 데이터의 원형을 넣어준다.
30 data <- rbind("0-4","5-11", "12-17", "18-24", "25-44", "45-54",
31 row.names = (el$LHI.Topic...Access.to.Health.Services)
32 matched_index = (match(row_names, data)) #NA를 제외한 행번호 확인
33 data.frame(matched_index) #매칭된 인덱스 번호를 확인한다.
34
35 #확인된 행번호의 내용을 주출하여 하나의 데이터 구조로 통합, 또한 일정한 순서로 필요한 컬럼 데이터 주출
36 #4번부터 최종 21번 컬럼까지 중에서 4개씩 풀일한 패턴으로 데이터 반복
37 #age_group <- as.matrix(rbind(c(el[23, seq(4,21,4)],c(el[24, seq(4,21,4)],c(el[27, seq(4,21,4)],
38 #                           c(el[28, seq(4,21,4)],c(el[30, seq(4,21,4)],c(el[31, seq(4,21,4)))]),
39
40 age_group <- data.frame(rbind(el[23, seq(4,21,4)], el[24, seq(4,21,4)],el[27, seq(4,21,4)],
41                             el[28, seq(4,21,4)],el[30, seq(4,21,4)],el[31, seq(4,21,4)]),
42                             row.names = NULL, check.names = FALSE)
43 #미국내 거주하는 인종별 분류 정보
44 race_ethnicity <- as.matrix(rbind(el[12, seq(4,21,4)],el[14, seq(4,21,4)],el[15, seq(4,21,4)] ),
45                             row.names = NULL, check.names = FALSE)
46
47
28.1 (Untitled) R Script
Console ~/nia_kbig/health/basic/
w.names=FALSE)
> write.table(el, file="/home/eduuser/nia_kbig/data/global_tmp2.csv", append=FALSE, quote=FALSE, sep ="", ro
w.names=FALSE)
> write.table(el, file="/home/eduuser/nia_kbig/data/global_tmpl.csv", append=FALSE, quote=FALSE, sep ="", ro
w.names=FALSE)
> |

```

- 불러들인 엑셀의 시트에서 분석에 필요한 데이터만을 추출하는 작업을 수행한다.
이를 통해, 시트 내에 섞여 있는 데이터 중에서 연령대 데이터를 추출할 수 있다.

```

01. #AHS-1.1 시트에서 연령별 추출 데이터의 row 값을 구성
02. #추출을 원하는 데이터의 원형을 넣어준다.
03. data_s1<- rbind("0-4", "5-11", "12-17", "18-24", "25-44", |
                    "45-54", "55-64")
04. row_names_s1 = (el$LHI.Topic...Access.to.Health.Services)
05. matched_index_s1 = (match(row_names, data)) #NA를 제외한 행번호 확인한다.
06. data.frame(matched_index_s1) #매칭된 인덱스 번호를 확인한다.

```



- 분석 원시 데이터 이동 스크립트 소스(global_health_middle.R)
- 라인 03 : 시트 1인 AHS-1.1의 데이터에서 추출한 Row 이름과 분석할 연령별로 분류된 정보를 찾기 위한 연령 그룹 정보를 문자로 입력한다.
- 라인 04 ~ 05 : 재구성된 Row 이름을 추출하고, 분석을 위해 입력된 이름과 매칭하여 위치 정보를 확인한다.
- 라인 06 : 매칭된 인덱스는 원본 데이터에서 분석 데이터에 대한 인덱스로, 프레임으로 변환하여 관리한다.

2. 매핑된 데이터에서 분석에 필요한 데이터 범위를 선정하고, 데이터를 추출한다.

```

01. #확인된 행번호의 내용을 추출하여 하나의 데이터 구조로 통합, 또한 일정한 순서로
    ↳ 필요한 컬럼 데이터 추출
02. #4번부터 최종 21번 컬럼까지 중에서 4개씩 동일한 패턴으로 데이터 반복
age_group_s1 <- data.frame(rbind(e1[23, seq(4,21,4)], e1[24, seq(4,21,4)],
03.   ↳ e1[27, seq(4,21,4)], e1[28, seq(4,21,4)], e1[30, seq(4,21,4)], e1[31,
      seq(4,21,4)]), row.names = NULL, check.names = FALSE)
04.
05. #미국내 거주하는 인종별 분류 정보
race_ethnicity <- as.matrix(rbind(e1[12, seq(4,21,4)], e1[14, seq(4,21,4)],
06.   ↳ e1[15, seq(4,21,4)]), row.names = NULL, check.names = FALSE)
07.
08. # convert factor to numeric for convenience
10. age_group$NA..2 <- as.numeric(age_group$NA..2)
11. age_group$NA..6 <- as.numeric(age_group$NA..6)
12. age_group$NA..10 <- as.numeric(age_group$NA..10)
13. age_group$NA..14 <- as.numeric(age_group$NA..14)
14. age_group$NA..18 <- as.numeric(age_group$NA..18)
15. #AHS-3 시트에서 연령별 추출 데이터의 row 값을 구성
16. #추출을 원하는 데이터의 원형을 넣어준다.
17. data_s2 <- rbind(" 0-4", " 5-11", " 12-17", " 18-24", " 25-44",
    ↳ " 45-54", " 55-64", " 65-74", " 75-84")
18. row.names_s2 = (e2$LHI.Topic...Access.to.Health.Services)

```



부연설명

- 분석 원시 데이터 이동 스크립트 소스(global_health_middle.R)
- 라인 03 : 매칭된 인덱스 정보를 이용하여 분석을 위한 컬럼 위치 정보에 따라 데이터를 추출하여 그룹으로 분류한다. 동일 패턴으로 그룹을 형성하고 있기 때문에 이 패턴을 적용하여 데이터를 분리한다.
- 라인 06 : 추가 분석을 위한 인종별 데이터를 추출하는 과정이며, 연령대로 추출하는 과정과 동일한 패턴으로 인종별 데이터를 추출할 수 있다.
- 라인 10 ~ 14 : 추출된 정보를 재분배하여 분석을 위한 정보를 연령별로 각각 분류하여 배치한다.
- 라인 17 : 시트 2인 AHS-3의 데이터에서 추출한 Row 이름과 분석할 연령별로 분류된 정보를 찾기 위한 연령 그룹 정보를 문자로 입력한다.
- 라인 18 : 재구성된 Row 이름을 추출한다.

V. 분석

```
01. #NA를 제외한 행번호 확인
02. matched_index_s2 = (match(row.names_s2, data_s2))
03. data.frame(matched_index_s2) #매칭된 인덱스 번호를 확인한다.
04. #확인된 행번호의 내용을 추출하여 하나의 데이터 구조로 통합, 또한 일정한 순서로
  ↪ 필요한 컬럼 데이터 추출
05. #4번부터 최종 21번 컬럼까지 중에서 4개씩 동일한 패턴으로 데이터 반복
  age_group_s2 <- data.frame(rbind(e2[27, seq(4,21,4)], e2[28, seq(4,21,4)],
  ↪ e2[29, seq(4,21,4)], e2[31, seq(4,21,4)], e2[32, seq(4,21,4)],
  06.   e2[34, seq(4,21,4)], e2[35, seq(4,21,4)], e2[37, seq(4,21,4)],
  ↪ e2[38, seq(4,21,4)], e2[39, seq(4,21,4)], row.names = NULL,
  ↪ check.names = FALSE)
07.
08. #미국내 거주하는 인종별 분류 정보
09. race_ethnicity_s2 <- as.matrix(rbind(e2[14, seq(4,21,4)], e2[15, seq(4,21,4)],
  ↪ e2[16, seq(4,21,4)], e2[17, seq(4,21,4)], e2[18, seq(4,21,4)]),
  ↪ row.names = NULL, check.names = FALSE)
10.
11.
12. # 분산 분석을 위한 데이터의 숫자 처리
13. age_group_s2$NA..2 <- as.numeric(age_group_s2$NA..2)
14. age_group_s2$NA..6 <- as.numeric(age_group_s2$NA..6)
15. age_group_s2$NA..10 <- as.numeric(age_group_s2$NA..10)
16. age_group_s2$NA..14 <- as.numeric(age_group_s2$NA..14)
17. age_group_s2$NA..18 <- as.numeric(age_group_s2$NA..18)
```



- 분석 원시 데이터 이동 **스크립트 소스(global_health_health.R)**
- 라인 02 : 분석을 위해 입력된 이름과 매칭하여 위치정보를 확인한다.
- 라인 03 : 매칭된 인덱스는 원본 데이터에서 분석 데이터에 대한 인덱스로, 프레임으로 변환하여 관리한다.
- 라인 06 : 매칭된 인덱스 정보를 이용하여 분석을 위한 컬럼 위치 정보에 따라 데이터를 추출하여 그룹으로 분류한다. 동일 패턴으로 그룹을 형성하고 있기 때문에 이 패턴을 적용하여 데이터를 분리한다.
- 라인 10 : 추가 분석을 위한 인종별 데이터를 추출하는 과정이며, 연령대로 추출하는 과정과 동일한 패턴으로 인종별 데이터를 추출할 수 있다.
- 라인 13 ~ 17 : 추출된 정보를 재분배하여 분석을 위한 정보를 연령별로 각각 분류하여 배치한다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화



1

2



VI 시각화

개요	91
분석 데이터 시각화	93
데이터 분석	94

VI

시각화



개요

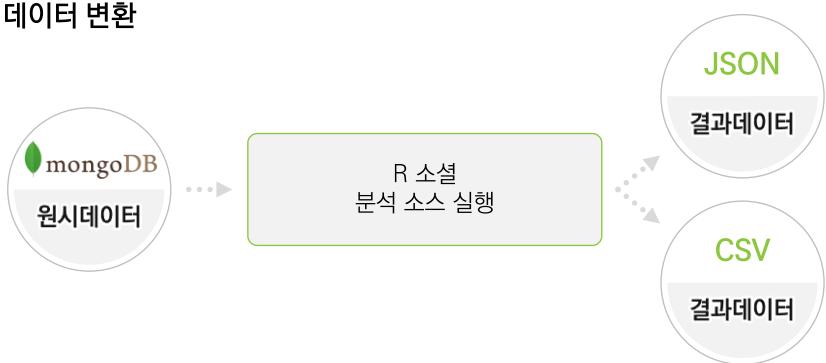
분석 결과는 다양한 방법으로 시각화하여 분석할 수 있으며, 이를 통해 데이터의 변화 및 분포를 해석하고, 데이터에 대한 분석 효과를 국내의 산업 또는 서비스 분야에 적용할 수 있다. 이를 위해 이미 분석되었거나 분석한 데이터의 상태와 추이에 대한 시각화 분석이 중요하며, 변화된 정보의 해석이 중요하다. 따라서 다양한 데이터에서 분석에 필요한 데이터만을 추출하고, 이를 시각화하는 것은 고도의 데이터 해석 능력이 요구된다.



> 시각화 방법 및 활용 기술

- **분석 결과에 대한 그래프 설정** : 추이분석에 유리한 시각화 도구는 박스 그래프 또는 라인 그래프로 출력하는 것이다.
- **저장된 포맷에 맞는 그래프 도구 설정** : 저장된 데이터를 이용하여 시각화할 수 있는 도구도 다양하기 때문에 이를 위해 시각화 도구를 선택한다.
- **추이해석** : 시각화된 박스 그래프의 추이 간격과 구간 이동 상태에 대한 의미를 해석하고, 미래 예측을 위한 추이 변화를 추론한다.

▶ 데이터 변환



I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화

> 분석 데이터 시각화

> 데이터 시각화

1. 시각화를 위한 데이터 범위와 그래프를 선정한다.
- #주) 그래프는 박스 그래프로 출력하여 수치 데이터의 연도별 변화와 구간 변화에 대한 추이와 추론을 할 수 있다.

```

01. #데이터를 시계열 자료 형식으로 변환
02. trx_s1 <- c("2008", "2009", "2010", "2011", "2012")
03. trx_s2 <- c("2007", "2008", "2009", "2010", "2011")
04. #헤더 정보 삭제, 헤더는 e1, e2
05. names(age_group_s1) <- trx_s1
06. names(age_group_s2) <- trx_s2
07.
08. #시각화 출력
09. par(mfrow = c(2,1)) #시각화 레이아웃을 표현할 구조 정의로, 2행, 1열을 표시
10. boxplot(age_group_s1[, 1:5] ,main = "AHS-1.1", xlab = "year", ylab ="age")
11. boxplot(age_group_s2[, 1:5] ,main = "AHS-3", xlab = "year", ylab ="age")

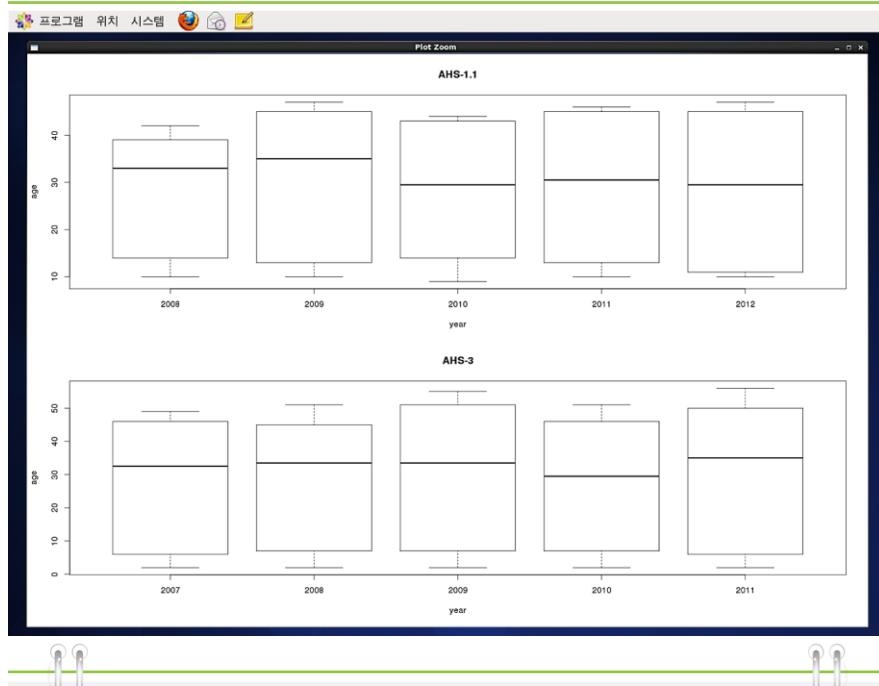
```



부연설명

- 분석 원시 데이터 이동 스크립트 소스(global_health_middle.R)
- 라인 02~03 : 시각화를 위한 그래프의 X축 변량을 표시하기 위해, 시트 1번과 시트 2번에 대한 각각의 변량 정보를 입력한다.
- 라인 05~06 : X축 변량 정보를 분류된 연령 정보의 헤더 정보로 표현하기 위해 매핑한다.
- 라인 09 : 두 개의 데이터에 대한 결과를 비교하기 위해, 시각화하기 위한 레이아웃 배치를 정의한다. 이때 두개의 그래프를 출력하기 위해서는 R 프로그램이 최대화된 상태로 유지되어야 한다.
- 라인 10~11 : 박스 그래프를 이용하여 시트1번과 시트2번의 전체 연령에 대한 연도별 구간 변화율을 표현하기 위한 분류 데이터 정보와 X축, Y축의 이름을 입력한다.

▶ 데이터 분석



- **주요 분석** : 미국의 헬스케어 시장이 점차 확대되고 있지만 2010년을 기점으로 헬스케어에 대한 연령별 이해도와 체험에 따른 변화가 거의 없는 형태로 발전되고 있다.
- AHS-1.1과 AHS-3의 데이터 주요 결과를 살펴보면, 미국의 헬스케어 신뢰도는 30대 초반이 가장 높은 신뢰도를 유지하면서 헬스케어 서비스를 이용하고 있다. 특히, 2010년은 헬스케어 서비스에 대한 기대 신뢰도의 연령층이 확대되어 낮아졌지만, 2011년 AHS-3 헬스케어 서비스에 대한 연령 신뢰도는 고연령층에서 높은 신뢰도를 얻고 있음을 알 수 있다.
- **국내 연계 추론 분석** : 국내의 헬스케어 서비스는 아직까지 스마트 폰을 즐겨 사용하는 연령층이 소비층으로 활성화되어 있고, 고연령층의 이용도와 신뢰도는 미약하다. 하지만 스마트 제품을 사용하는 연령층이 확대되고 활용도와 서비스의 편리성이 높아지면서 헬스케어 서비스에 대한 관심과 신뢰도가 높아지고 있으며 이를 통한 서비스 활용화가 급격하게 증가할 것으로 판단된다.



VII 예제문제

예제 문제1

97

예제 문제2

98

예 / 제 / 문 / 제

예제 1

헬스 케어 데이터에서 국내 인종별 서비스에 대한 신뢰도를 비교 분석하라.

- 미국 내 인종별 헬스케어 서비스에 대한 신뢰도의 추이를 분석하기 위한 데이터 추출과 박스 그래프로 시각화하여 비교하라.
 - 미국 내 헬스케어 데이터 AHS-1.1과 AHS-3에서 인종별 신뢰도 데이터를 추출하고, 원본 데이터와 분리한다.
 - 추출된 데이터에서 연도별 분석된 데이터에 대한 패턴을 확인하고 필요 데이터를 연도별에서 추출한다.
 - 인종별로 추출된 연도 데이터를 데이터 프레임으로 구조화시키고 저장한다.
 - 데이터 추이 분석을 위해 박스 그래프 또는 추이선을 이용하여 시각화하고, 그 결과를 비교 분석하라.

예제 2

미국의 헬스케어 서비스에 대한 성별 신뢰도를 분석하라.

- 미국내 성별 헬스케어 서비스에 대한 신뢰도의 추이를 분석하고, 박스 그래프로 시각화 하라.

- 미국내 헬스케어 데이터 AHS-1.1과 AHS-3에서 성별 신뢰도 데이터를 추출하고, 원본 데이터와 분리한다.
- 추출된 데이터에서 연도별 분석된 데이터에 대한 패턴을 확인하고 필요 데이터를 연도별에서 추출한다.
- 성별로 추출된 연도 데이터를 데이터 프레임으로 구조화시키고 저장한다.
- 데이터 추이 분석을 위해 박스 그래프 또는 추이선을 이용하여 시각화하고, 그 결과를 비교 분석하라.

데이터 분석 콘텐츠 활용 매뉴얼

2014년 12월 인쇄

2015년 1월 발행

발 행 처 한국정보화진흥원 빅데이터전략센터

집 필 신신애, 김성현, 박재원, 김현태, 김지홍, 정다운,
이승하, 신은비

주 소 서울시 중구 청계천로 14

연 락 처 (02) 2131-0114

인 쇄 HNJ Printing

〈비매품〉

[데 이 터 분 석 콘 텐 츠]
활용 매뉴얼

NIA  한국정보화진흥원

(100-775) 서울시 종구 청계천로 14 한국정보화진흥원
TEL 02-2131-0114 FAX 02-2131-0109
www.nia.or.kr

