

# 7 관광 활용 매뉴얼

데이터 분석 콘텐츠



미래창조과학부



한국정보화진흥원



# CONTENTS

Beginning Level 초급과정

## I 개요

개요	9
----	---

## II 수집

개요	13
수집 데이터	14
데이터 수집	15
데이터 작업 영역 이동 스크립트	18

## III 가공

개요	23
데이터 가공 R 스크립트	27

## IV 저장

개요	31
R Studio 활용 저장	32



## V 분석

---

개요	35
R Studio 활용 분석	36
R Studio 저장	39

## VI 시각화

---

개요	43
분석 데이터 시각화	45
데이터 분석	49

## VII 예제 문제

---

예제 문제1. 내국인 관광객 수와 외국인 관광객의 월간 방문객 수를 연도별로 비교 분석하라.	53
예제 문제2. 내국인 관광객 수와 외국인 관광객 수 사이의 상관관계를 분석하라.	54

# CONTENTS

Intermediate Level **중급과정**

## I 개요

개요	59
----	----

## II 수집

개요	63
수집 데이터	64
데이터 수집	66
데이터 작업 영역 이동 스크립트	69

## III 가공

개요	73
데이터 가공 R 스크립트	76

## IV 저장

개요	81
R Studio 활용 저장	82

## V 분석

---

개요	85
R Studio 활용 분석	87
R Studio 저장	90

## VI 시각화

---

개요	93
데이터 분석	95

## VII 예제 문제

---

예제 문제1. 3년간 월별로 취합된 내외국인 관광객 수 및 주요 관광지 관광 수입 데이터에 쉽게 구할 수 있는 월간 통계 데이터를 매쉬업 하여 다중 회귀 분석하라.	99
예제 문제2. 2년간 월별 교통수단별 관광객 수와 항공기나 선박 사고 관련 뉴스 데이터 발생 빈도를 매쉬업하여 사고와 교통 수단 선택의 연관성을 분석하라.	100



관광 

Beginning Level

초급과정







## I 개요

개요

9

8

# I

## 개요

### > 개요

제주시 통합자료센터(<http://jejudb.jejesi.go.kr/>)로부터 2010년 ~ 2012년 제주도 내외국인 입도 관광객 수 및 주요 관광지 관광 수입 통계 데이터를 바탕으로 내외국인 유입 관광객 수와 실제 관광 수입의 연관성을 분석하여 제주도 관광수입의 증대를 위한 전략 수립의 기초 자료를 마련한다.

### > 활용 데이터

- **jeju\_2010.csv :**  
2010년 제주도 내외국인 입도 관광객수 및 주요 관광지 관광수입 데이터
- **jeju\_2011.csv :**  
2011년 제주도 내외국인 입도 관광객수 및 주요 관광지 관광수입 데이터
- **jeju\_2012.csv :**  
2012년 제주도 내외국인 입도 관광객수 및 주요 관광지 관광수입 데이터

### > 선행학습

- **리눅스** – 파일시스템 구조, 쉘 명령어, 쉘 스크립트 실행 방법
- **R 프로그래밍 언어** – 기본 지식(문법, 패키지 추가 설치 방법)
- **통계** – 상관관계(Correlation), 회귀분석(regression)
- **R 쳐트** – 구성 방법

## > 요구사항

- 수집된 2010년~2012년 제주도 관광 데이터 셋을 분석에 필요한 저장 공간으로 이동시키고, 내외국인 입도 정보와 관광 수입 간의 상관관계를 분석하여 관광객 유입이 관광수입으로 얼마나 이어지는지 분석하라.

## > 분석 절차

- 수집된 2010년~2012년 제주도 관광 데이터 셋을 로드한다.
- 연도별로 각각 수집된 월별 데이터를 통합한다.
- 내국인 입도 관광객 수와 주요 관광지 관광수입의 상관관계를 구한다.
- 외국인 입도 관광객 수와 주요 관광지 관광수입의 상관관계를 구한다.
- 내국인과 외국인 관광객 중 어느 쪽이 관광지 관광수입으로 연결되는지 판단한다.
- 3년간 내국인과 외국인 관광객 합계를 구한다.
- 내외국인 관광객 수의 비중을 시각화하고 관광수입과의 연관도를 함께 비교하여 의미 있는 결론을 도출한다.



- 상관분석(Correlation Analysis)** : 확률론과 통계학에서 두 변수 간에 어떤 선형적 관계를 갖고 있는지를 분석하는 방법이다. 두 변수는 서로 독립적인 관계로부터 서로 상관된 관계일 수 있으며 이때 두 변수 간의 관계의 강도를 상관관계(Correlation, Correlation coefficient)라 한다.
- 상관계수(Correlation coefficient)** : 상관계수( $r$ )는 두 개의 변수 간의 선형 관계의 방향과 강도를 나타내는 변수로,  $-1 \leq r \leq 1$  의 범위를 갖는다.  
 $-1$ 이나  $1$ 에 가까울수록 강한 상관관계를 가지며,  $0$ 에 가까울수록 상관관계가 적다.  
 양수이면 정의 상관관계, 음수이면 역의 상관관계가 있음을 나타낸다.



1

2

## II 수집

개요	13
수집 데이터	14
데이터 수집	15
데이터 작업 영역 이동 스크립트	18



# 수집

## > 개요

관광 데이터는 제주시 통합자료센터(<http://jejudb.jejesi.go.kr/>)에서 매년 발표하는 종합 통계 연보 데이터로부터 최근 3년 (2010년~2012) 연간의 데이터 중 분석에 필요한 정보(내 외국인 관광객 및 관광 수입 정보)를 추출하여 분석에 용이하게 편집하여 제공한다.

## > 수집 방법

- **데이터 제공 :** 관광 데이터는 제주시 통합자료센터(<http://jejudb.jejesi.go.kr/>)로부터 수동으로 데이터를 수집할 수 있으며, 실습용 자료는 빅데이터 분석 활용센터에 접속하여 수집 데이터 셋을 다운로드할 수 있도록 원시데이터를 제공하고 있다.

The screenshot shows the homepage of the Jeju City Integrated Database Center (<http://jejudb.jejesi.go.kr/>). The main content area features a photograph of a library or bookstore interior. To the right, there is a search bar labeled "DATA SEARCH" and a sidebar with various links and icons related to tourism data analysis, such as "도서검색" (Book Search), "기록물검색" (Record Material Search), "도서대출신청" (Book Loan Application), and "통계 Q&A". The top navigation bar includes links for "도서정보", "비전지기획률", "시군통합마인저", "통계보는제주시", "별관마당", and "관련사이트".

## > 수집 데이터

- 2010년 제주도 내외국인 관광객 수 및 관광수입 데이터(jeju\_2010.csv)

IN_VISITOR	OUT_VISITOR	INCOME
480778	30197	222548
464067	45445	235283
514987	43923	237449
658352	70029	292744
667689	77881	315586
544390	69631	268960
590093	78914	319334
686331	95752	379581
518338	75337	280745
634212	86257	318276
569616	54819	275501
472448	48815	240707

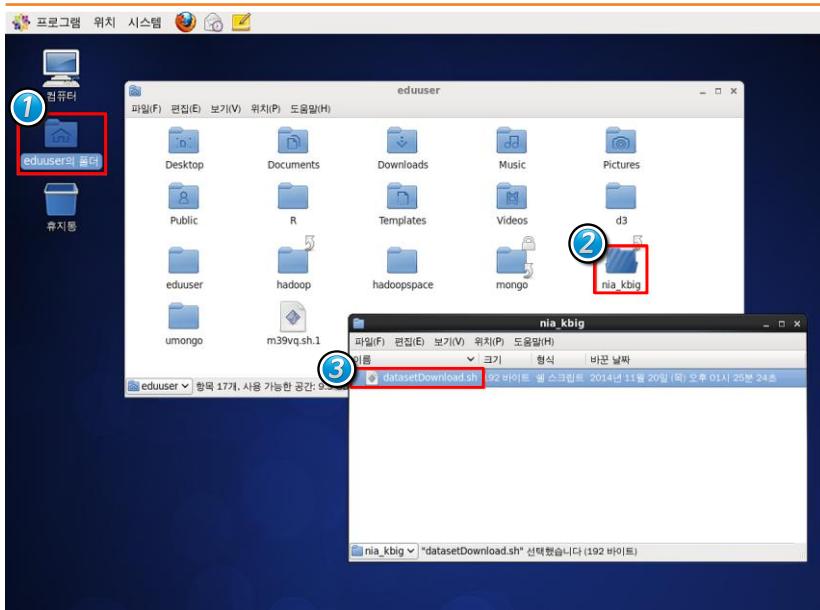
순서대로 2010년 1월~12월 데이터를 표현하고 있다.

- IN\_VISITOR : 내국인 관광객 수(명)
- OUT\_VISITOR : 외국인 관광객 수(명)
- INCOME : 주요 관광지 관광수입(백만 원)

## ▶ 데이터 수집(datasetDownload.sh)

- 데이터 저장소에서 서버 로컬로 관광 데이터 셋을 복사해 온다.
  - **jeju\_2010.csv :**  
2010년 제주도 내외국인 입도 관광객수 및 주요 관광지 관광수입 데이터
  - **jeju\_2011.csv :**  
2011년 제주도 내외국인 입도 관광객수 및 주요 관광지 관광수입 데이터
  - **jeju\_2012.csv :**  
2012년 제주도 내외국인 입도 관광객수 및 주요 관광지 관광수입 데이터

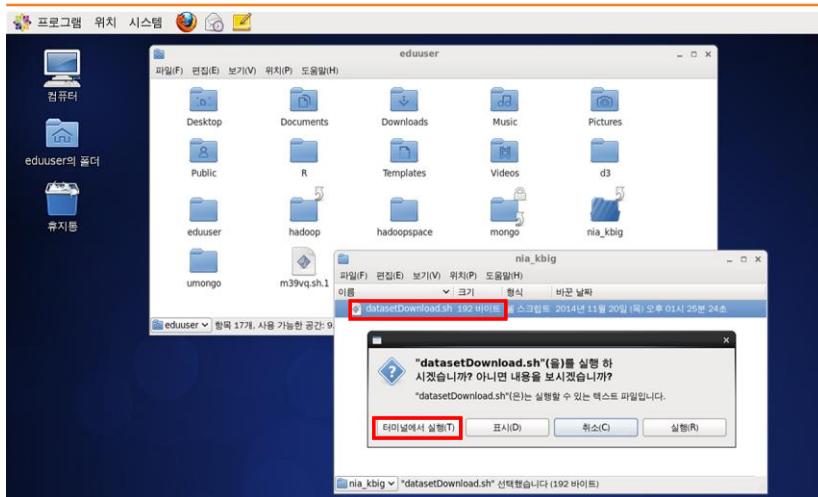
## ▶ 실습코드 디렉토리로 이동



- ① 로그인 후 바탕화면에서 eduuser 폴더를 오픈한다.
- ② nia\_kbig 폴더를 오픈한다.
- ③ datasetDownload.sh를 더블클릭하여 실행한다.

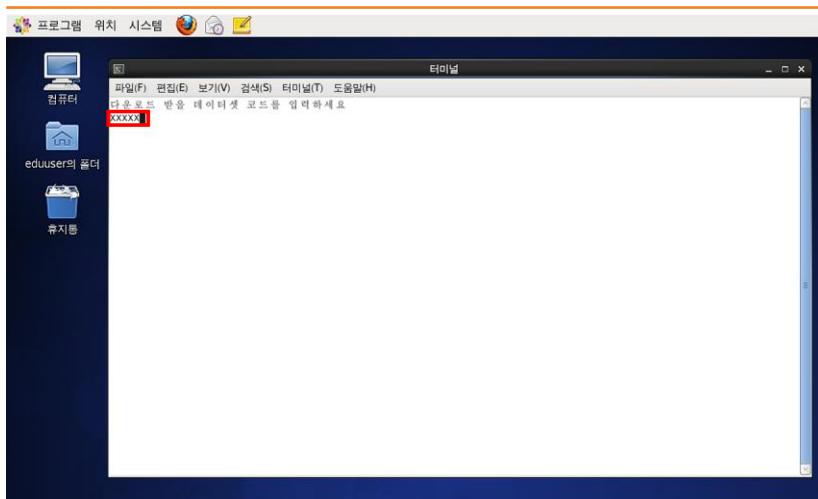
## ▶ 레파지토리에서 데이터 수집

### datasetDownload.sh (원시데이터로 컬서버로 복사)



- '터미널에서 실행' 버튼을 클릭한다.

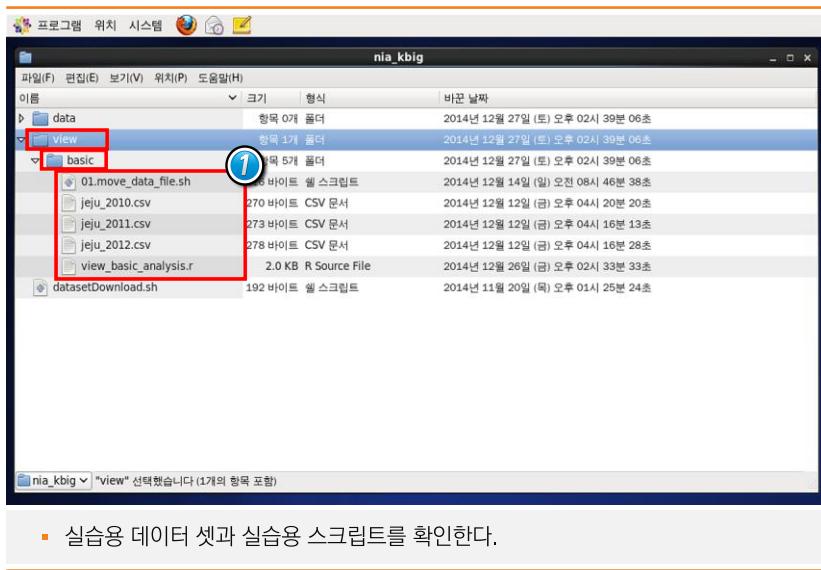
## ▶ 데이터셋 코드 입력



- 다운로드 받을 데이터셋 코드를 입력 후 엔터

## II. 수집

### ▶ 데이터셋과 실습용 쉘 스크립트



- 실습용 데이터셋과 실습용 스크립트를 확인한다.

### ▶ ① 데이터 및 스크립트

#### ▪ 01.move\_data\_file.sh :

로컬로 수집해온 데이터를 작업 영역 Data 폴더로 자료를 이동하는 스크립트

#### ▪ view\_basic\_analysis.R :

관광 데이터 분석용 R 스크립트

#### ▪ jeju\_2010.csv :

2010년 제주도 내외국인 입도 관광객수 및 주요 관광지 관광수입 데이터

#### ▪ jeju\_2011.csv :

2011년 제주도 내외국인 입도 관광객수 및 주요 관광지 관광수입 데이터

#### ▪ jeju\_2012.csv :

2012년 제주도 내외국인 입도 관광객수 및 주요 관광지 관광수입 데이터

## > 데이터 작업 영역 이동 스크립트(01.move\_data\_file.sh)

### > 데이터 이동 스크립트

- 로컬로 수집해온 데이터를 작업 영역 Data 폴더로 자료를 이동하는 스크립트

#### 01.move\_data\_file.sh

```

01.#!/bin/bash
02.TARGET_JEJU=/home/eduuser/nia_kbig/view/basic/jeju_*.csv
03.
04.# 작업 디렉토리 정의
05.LOCAL_DIR=/home/eduuser/nia_kbig/data/
06.mv $TARGET_JEJU $LOCAL_DIR
07.

```



- 데이터 작업 영역 이동 스크립트 소스(01.move\_data\_file.sh)
- 라인 02 : 다운로드 받은 원시데이터 파일들의 위치(path)를 변수(TARGET\_FASHION)로 지정하는 라인이다.
- 라인 05 : 작업영역 디렉토리의 위치(path)를 변수(LOCAL\_DIR)로 지정하는 라인이다.
- 라인 06 : mv 명령어를 사용하여 다운로드 받은 원시데이터 파일들을 작업영역 디렉토리로 이동시키는 라인이다.

## II. 수집

## ▶ 수집 데이터 셋 작업 영역 폴더 이동

- R Studio에서 상관분석을 위한 수집된 데이터 셋을 작업 영역 Data 폴더로 자료를 이동

```
eduuser@cm04 ~]$ ll
total 20
-rwxr-xr-x 1 eduuser eduuser 216 2014-12-14 08:46 01.move_data_file.sh
-rw-r--r-- 1 eduuser eduuser 283 2014-12-04 22:22 01.move_data_file.sh~
-rw-r--r-- 1 eduuser eduuser 270 2014-12-12 16:20 jeju_2010.csv
-rw-r--r-- 1 eduuser eduuser 273 2014-12-12 16:16 jeju_2011.csv
-rw-r--r-- 1 eduuser eduuser 278 2014-12-12 16:16 jeju_2012.csv
[eduuser@cm04 basic]$ ./01.move_data_file.sh
```

5. “./01.move data file.sh”를 입력하여 준비된 관광 데이터를 이동시킨다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화





### III 가공

개요

23

데이터 가공 R 스크립트

27



## 가공

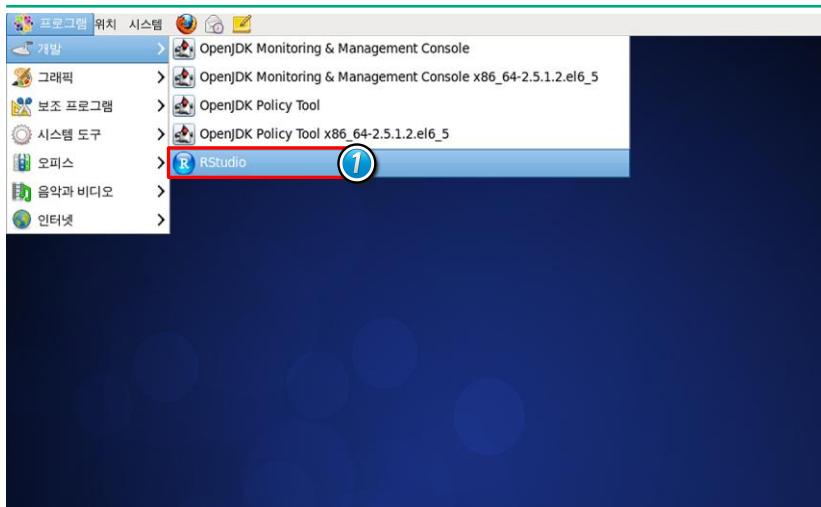
### > 개요

작업 영역 폴더에 복사한 관광 데이터의 가공은, 연도별로 수집된 3년간(2010년~2012년)의 데이터를 읽어들여 하나의 데이터로 통합하고, 3년간의 데이터를 매시업하여 시계열 분석을 위한 데이터로 만들어 상관분석(연관도 분석)에 적합한 형태로 변환하도록 한다.

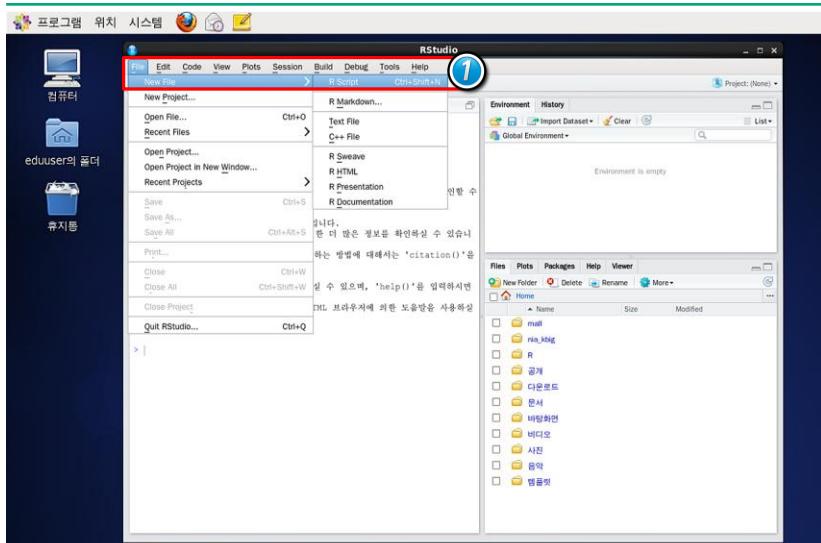
### > 가공 방법

- **분석 도구 실행** : 가공 분석을 위해, 프로그래밍 도구인 R을 실행한다.
- **데이터 로드** : 3년간의 데이터가 각각의 파일로 저장되어 있으므로, R Studio에서 각각의 데이터를 읽어들인다.
- **데이터 통합** : R Studio는 동일한 형태의 데이터를 통합할 수 있는 함수를 제공한다. 이를 활용하여 각각 읽어들인 3년간의 데이터를 하나로 통합한다.

## ▶ 데이터 가공

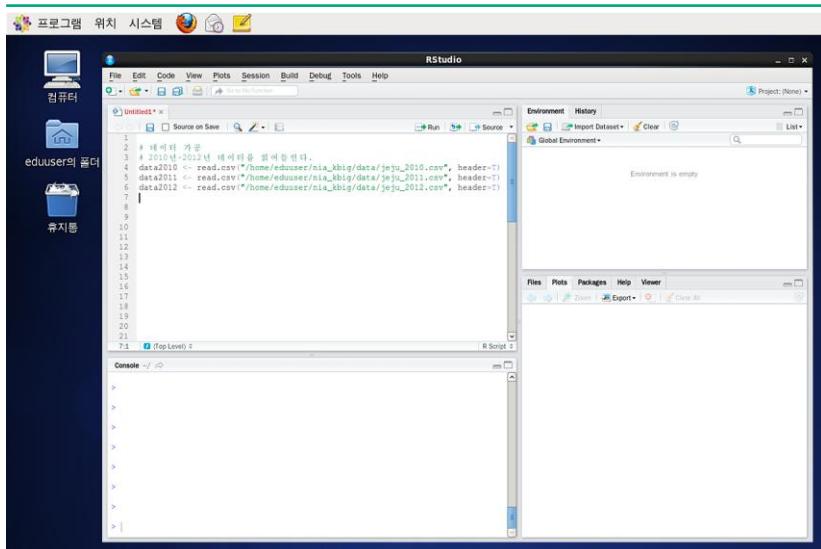


- ① 왼쪽 상단의 [“프로그램” 클릭] > [“개발” 클릭] > [“RStudio” 클릭]으로 분석 도구인 R Studio를 실행한다.

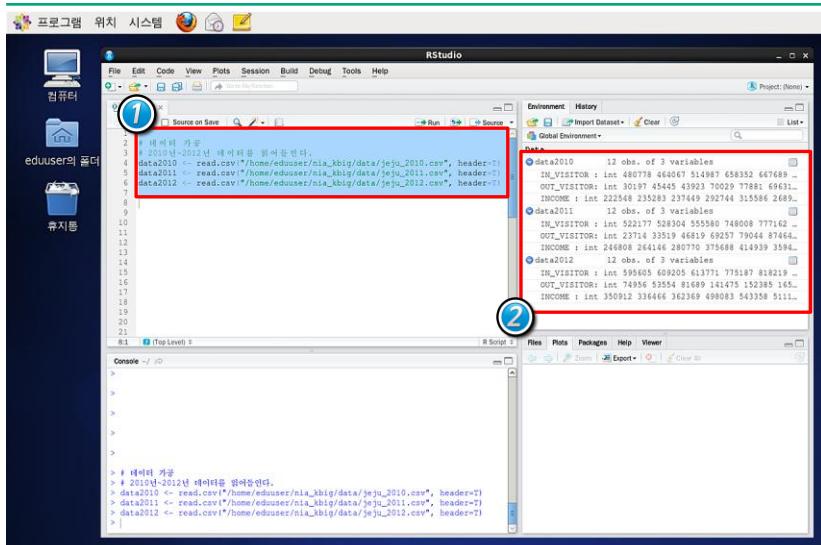


- ② 관광 데이터의 분석 및 가공을 위해 프로그램 작업 파일 ( “New File” 클릭 ) > [“R\_Script” 클릭]을 선택한다.

### III. 가공



3. 2010년~2012년 제주도 관광 데이터를 각각 로드하도록 R 스크립트를 작성한다.



4. ① 현재까지 작성한 스크립트 코드를 선택하여 Ctrl+Enter를 입력하면, ②와 같이 데이터가 로드된 것을 볼 수 있다.(코드의 부분 실행은 R스크립트만의 장점이다.)

The screenshot shows the RStudio interface. In the top-left corner, there's a desktop icon for 'eduserver의 폴드' (eduserver's folder) containing icons for '컴퓨터' (Computer), 'eduserver 폴더' (eduserver folder), and '휴지통' (Recycle Bin). The main window displays R code in the 'Source' tab and its execution results in the 'Console' tab.

```

1 # 데이터 가공
2 # 2010년~2012년 데이터를 읽어들린다.
3 # 2010년 데이터는 read.csv("home/eduserver/nia_kbig/data/jeju_2010.csv", header=T)
4 data2010 <- read.csv("home/eduserver/nia_kbig/data/jeju_2010.csv", header=T)
5 data2011 <- read.csv("home/eduserver/nia_kbig/data/jeju_2011.csv", header=T)
6 data2012 <- read.csv("home/eduserver/nia_kbig/data/jeju_2012.csv", header=T)
7
8 # 2010년~2012년 데이터를 통합하여 하나의 데이터로 만든다.
9 inputData <- rbind(data2010, data2011,
10 inputData <- rbind(inputData, data2012)
11
12
13
14
15
16
17
18
19
20
21

```

In the 'Environment' pane, two datasets are listed:

- data2010**: 12 obs. of 3 variables
- data2011**: 12 obs. of 3 variables
- data2012**: 12 obs. of 3 variables
- inputData**: 36 obs. of 3 variables

The 'inputData' dataset includes columns: TEL\_VISITOR (int 480778 444047 514987 658352 657659...), OUT\_VISITOR (int 30197 45445 43923 70229 77881 69431...), and INCOME (int 222548 235283 237449 292744 315586 2689...).

5. ① 각각 로드한 2010년~2012년 제주도 관광 데이터를 통합하도록 코드를 작성하고 부분 실행하면, ② 와 같이 inputData로 통합된 것을 확인할 수 있다.  
 ▪ #주) rbind 함수는 동일한 형태의 데이터를 리니어하게 연결하여 하나의 데이터로 통합하는 함수이다.

### III. 가공

#### ▶ 데이터 가공 R 스크립트

```
01. #데이터 가공  
02. # 2010년 ~ 2012년 데이터를 읽어들인다.  
03. data2010 <- read.csv("/home/eduuser/nia_kbig/data/jeju_2010.csv", header=T)  
04. data2011 <- read.csv("/home/eduuser/nia_kbig/data/jeju_2011.csv", header=T)  
05. data2012 <- read.csv("/home/eduuser/nia_kbig/data/jeju_2012.csv", header=T)  
06.  
07. #2010년 ~ 2012년 데이터를 통합한다.  
08. inputData <- rbind(data2010, data2011)  
09. inputData <- rbind(inputData, data2012)
```



##### • 데이터 가공 R 스크립트 소스

- 라인 03~05 : 2010~2012년 제주도 관광 데이터파일(jeju\_2010.csv, jeju\_2011.csv, jeju\_2012.csv)을 읽어들여 R 데이터 객체(data2010, data2011, data2012)로 저장하는 라인이다.
- 라인 08~09 : rbind 함수를 활용하여 3~5라인에서 읽어들인 연도별 R 데이터 객체들(data\_2010, data2011, data2012)을 하나의 R 데이터 객체(inputData)로 통합하는 라인이다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화



## IV 저 장

개요

31

R Studio 활용 저장

32

# IV

## 저장

### > 개요

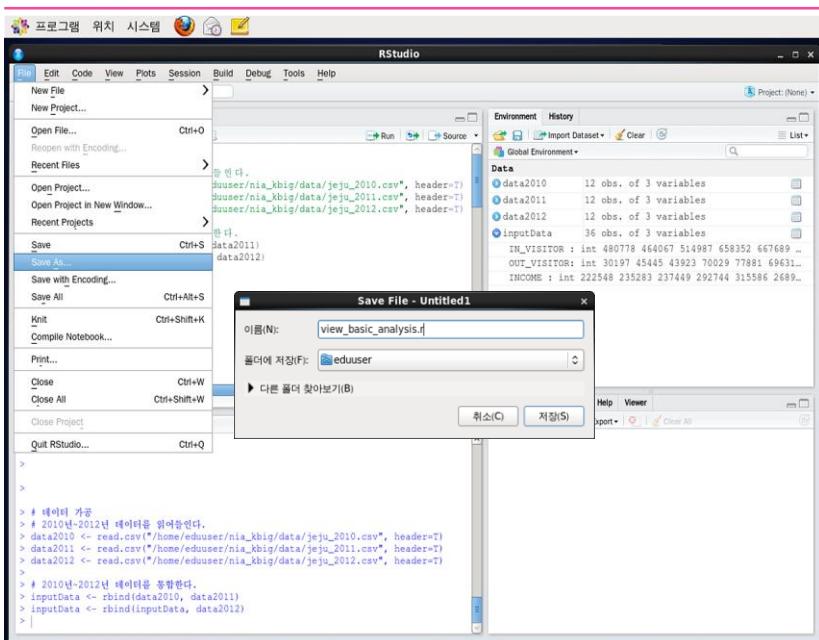
R Studio를 활용하여 데이터 로드 > 가공 > 분석 > 시각화 단계를 한꺼번에 실행하므로, 별도의 저장 과정은 생략한다. 가공된 데이터는 메모리상에 존재하며, 지금까지 작성된 분석 프로그램 소스를 저장한다.

### > 저장 방법

- **가공된 데이터 메모리 저장** : 관광 데이터 분석을 위해 가공한 데이터는 R Studio 메모리상에 저장된다.
- **소스 저장** : 작성 중인 제주도 관광 데이터 분석 프로그램을 저장한다.

## > R Studio 활용 저장

### > 데이터 저장



1. 관광 데이터 분석을 위해 작성 중인 프로그램 소스를 저장한다

- #주) 작성 중인 프로그램 소스를 저장하는 방법은 메뉴의 “File” -> “Save”를 이용하거나 도구상자의 저장 아이콘을 이용한다. 저장시 저장 위치 및 파일명은 “/home/eduuser/nia\_kbig/view\_basic\_analysis.r”로 저장한다.

# W





## V 분석

개요	35
R Studio 활용 분석	36
R Studio 저장	39

# V 분석

## > 개요

관광 데이터 분석은 R Studio에 내장된 상관관계 분석 함수를 활용하여 연관성이 있을 법한 내외국인 관광객 수와 관광수입의 상관계수를 계산하고, 이를 시각화하여 내국인과 외국인 중 어느 쪽이 관광지 관광수입과 연관성이 높은지 비교한다.

## > 데이터 분석 방법

- **상관 분석** : 두 개 이상의 시계열 데이터 사이의 상관관계를 계산하는 통계 기법을 사용한다.
- **상관계수 계산** : 내국인 방문객과 관광수입, 외국인 방문객과 관광수입 간의 상관 계수를 계산한다.(R Studio 의 cor 함수 활용)
- **상관분석** : R Studio 의 lm 함수를 활용하여 상관 분석을 한다.

## > R Studio 활용 분석

### > 데이터 불러오기

inputData (가공 단계에서 통합한 관광 데이터)

The screenshot shows the RStudio interface with the following details:

- Environment:** Shows objects `data2010`, `data2011`, `data2012` and `inputData`.
- Data View:** Shows the structure of `inputData` with variables `IN\_VISITOR`, `OUT\_VISITOR`, and `INCOME`.
- Console:** Shows the command `inputData` being run (circled 1) and its resulting output (circled 2).
- Output:** The output in the Console shows 36 rows of data starting with row 22.

```

1 # 데이터 가공
2 # 2010년~2012년 데이터를 읽어들인다.
3 # 파일은 각각 'jeju_2010.csv', 'jeju_2011.csv', 'jeju_2012.csv'이다.
4 data2010 <- read.csv("/home/eduser/nia_kbigs/data/jeju_2010.csv", header=T)
5 data2011 <- read.csv("/home/eduser/nia_kbigs/data/jeju_2011.csv", header=T)
6 data2012 <- read.csv("/home/eduser/nia_kbigs/data/jeju_2012.csv", header=T)
7
8 # 2010년~2012년 데이터를 통합한다.
9 inputData <- rbind(data2010, data2011)
10 inputData <- rbind(inputData, data2012)
11
12 # 가공 대상자를 확인한다.
13 inputData
14
15
16
17
18
19
20 # 대국민 관광객수와 관광 수입의 상관관계를 구한다.
21
22 729359 128903 471956
23 621228 94085 369573
24 556939 91186 352568
25 595605 74956 350912
26 609205 53554 336466
27 613771 81689 362369
28 713187 16034 451383
29 838219 152385 542158
30 682740 165576 511132
31 663594 219538 584248
32 697843 224623 606998
33 591470 16034 451383
34 750220 168834 511132
35 661053 114209 408060
36 551591 104356 337860
  
```

- ① 가공한 데이터가 잘 들어가 있는지 확인하기 위해 “inputData”를 입력하고 위와 같이 블록을 선택한 후, Ctrl+Enter를 입력하면, ② 와 같이 데이터를 확인할 수 있다.

## ▶ 데이터 분석

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

The screenshot shows the RStudio interface. The left pane displays the R script 'view\_basic\_analysis.R' with the following code:

```

1 # 데이터 가공
2 # 2010년~2012년 데이터를 읽어들인다.
3 data2010 <- read.csv("home/edususer/nia_kbkg/data/jeju_2010.csv", header=T)
4 data2011 <- read.csv("home/edususer/nia_kbkg/data/jeju_2011.csv", header=T)
5 data2012 <- read.csv("home/edususer/nia_kbkg/data/jeju_2012.csv", header=T)
6
7 # 2010년~2012년 데이터를 통합한다.
8 inputData <- rbind(data2010, data2011)
9 inputData <- rbind(inputData, data2012)
10
11 # 가공 데이터를 확인 한다.
12 inputData
13
14 # 내국인 관광객수와 관광수입의 상관관계를 구한다.
15 in_cor <- cor(inputData$IN_VISITOR, inputData$INCOME)
16 out_cor <- cor(inputData$OUT_VISITOR, inputData$INCOME)
17
18
19
20
21
22
23
24
25 595605 74056 350912
26 609205 53554 336466
27 613771 81689 362369
28 775187 141475 490803
29 818219 152385 543358
30 691440 113255 337242
31 663594 219538 584248
32 697843 224623 605998
33 591470 160204 453063
34 750026 188834 537363
35 660153 113255 337240
36 551591 141475 337263
> # 내국인 관광객수와 관광수입의 상관관계를 구한다.
> in_cor <- cor(inputData$IN_VISITOR, inputData$INCOME)
> out_cor <- cor(inputData$OUT_VISITOR, inputData$INCOME)
>

```

The right pane shows the Global Environment, listing datasets and variables:

- data2010: 12 obs. of 3 variables
- data2011: 12 obs. of 3 variables
- data2012: 12 obs. of 3 variables
- inputData: 36 obs. of 3 variables
- in\_cor: 0.791893463902311
- out\_cor: 0.947812259459879

- cor 함수를 활용하여 내국인 방문객과 관광수입, 외국인 방문객과 관광수입 간의 상관계수를 계산하여 각각 변수에 저장한다.

```

01. #내외국인 관광객수와 관광수입의 상관관계를 구한다.
02. in_cor = cor(inputData$IN_VISITOR, inputData$INCOME)
03. out_cor = cor(inputData$OUT_VISITOR, inputData$INCOME)
04.

```



- 관광 분석 및 시각화 R 스크립트 소스(view\_basic\_analysis.R)
- 라인 02 : 내국인 관광객수와 관광수입 사이의 상관관계를 계산하여 변수(in\_cor)로 저장하는 라인이다.
- 라인 03 : 외국인 관광객수와 관광수입 사이의 상관관계를 계산하여 변수(out\_cor)로 저장하는 라인이다.

```

13 inputData
14
15 # 내국인 관광객수와 관광수입의 상관분석을 실행한다.
16 in_cor = cor(inputData$IN_VISITOR, inputData$INCOME)
17 out_cor = cor(inputData$OUT_VISITOR, inputData$INCOME)
18
19 # 내국인 관광객과 관광수입의 상관분석 실행
20 in_r <- lm(inputData$INCOME, inputData$IN_VISITOR)
21
22 # 외국인 관광객과 관광수입의 상관분석 실행
23 out_r <- lm(inputData$INCOME, inputData$OUT_VISITOR)
24
25
26
27
28
29
30
31
32
33

```

Console

```

30 682740 165576 511132
31 663594 219538 584248
32 697843 224623 605998
33 591470 160388 453063
34 759226 209834 537533
35 641053 114209 408060
36 551591 104356 337263
> # 내국인 관광객수와 관광수입의 상관분석을 구한다.
> in_cor = cor(inputData$IN_VISITOR, inputData$INCOME)
> out_cor = cor(inputData$OUT_VISITOR, inputData$INCOME)
> in_r <- lm(inputData$INCOME, inputData$IN_VISITOR)
>
> # 외국인 관광객과 관광수입의 상관분석 실행
> out_r <- lm(inputData$INCOME, inputData$OUT_VISITOR)
> |

```

2. ① lm 함수를 활용하여 내국인 방문객과 관광수입, 외국인 방문객과 관광수입 간의 상관 분석을 실행한다.( lm 함수는 R Studio에 내장되어 있는 상관분석 함수이다.)

```

01. #내국인 관광객수와 관광수입의 상관분석을 실행한다.
02. in_r = lm(inputData$INCOME, inputData$IN_VISITOR)
03. #외국인 관광객수와 관광수입의 상관분석을 실행한다.
04. out_r = lm(inputData$INCOME, inputData$OUT_VISITOR)

```

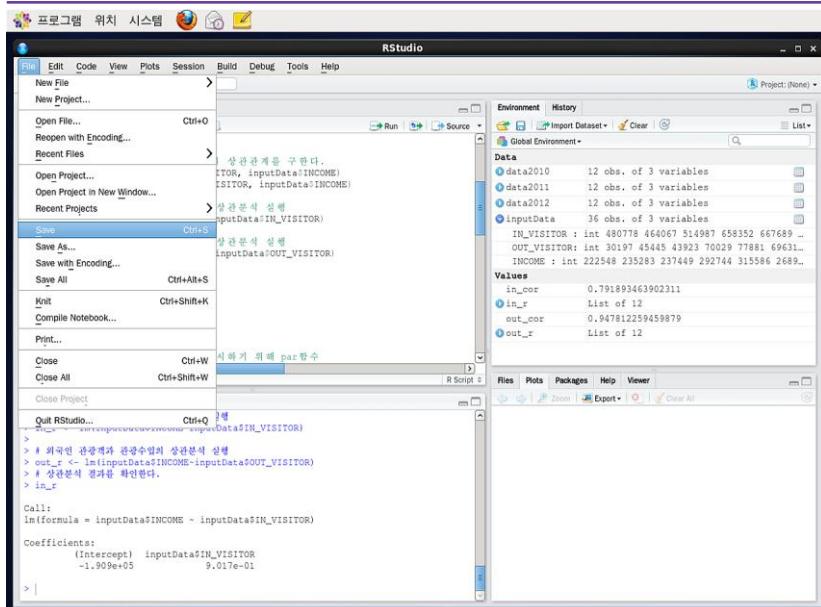


- 관광 분석 및 시각화 R 스크립트 소스(view\_basic\_analysis.R)
- 라인 02 : lm 함수를 사용하여 내국인 관광객수와 관광수입을 상관 분석하여 객체(in\_r)로 저장하는 라인이다.
- 라인 04 : lm 함수를 사용하여 외국인 관광객수와 관광수입을 상관 분석하여 객체(out\_r)로 저장하는 라인이다.

## ➤ R Studio 저장

### ➤ 분석 결과 저장

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



The screenshot shows the RStudio interface. The code editor pane contains the following R script:

```

# 관광 분석 실행
inputData<-INCOME
IN_VISITOR<-VISITOR
inputData$INCOME<-INCOME
inputData$OUT_VISITOR<-OUT_VISITOR
inputData$IN_VISITOR<-IN_VISITOR
inputData$OUT_VISITOR<-OUT_VISITOR
# 외국인 관광객과 관광수입의 상관분석 실행
out_r <- lm(inputData$INCOME~inputData$OUT_VISITOR)
# 상관분석 결과를 확인한다.
in_r

```

The environment pane shows the following variables:

Type	Name	Description
Data	data2010	12 obs. of 3 variables
Data	data2011	12 obs. of 3 variables
Data	data2012	12 obs. of 3 variables
Data	inputData	36 obs. of 3 variables
Values	in_r	0.791893463902311
Values	in_r\$cor	List of 12
Values	out_r	0.947812259459879
Values	out_r\$cor	List of 12

3. “File/Save”를 클릭하여 지금까지 관광 데이터를 분석하기 위해 작성한 프로그램을 저장한다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화



1

2



## VI 시각화

개요	43
분석 데이터 시각화	45
데이터 분석	49

# VI

## 시각화

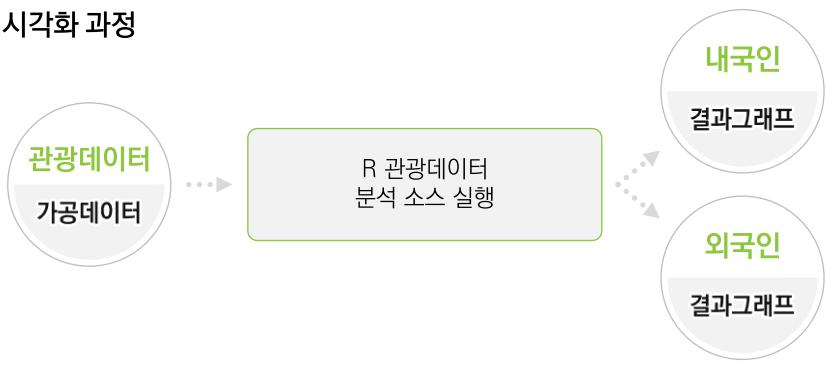
### > 개요

관광 데이터의 분석 과정에서 상관 분석의 결과는 다양한 방법으로 시각화하여 분석할 수 있으며, 이를 통해 데이터의 변화 및 분포를 해석하고 데이터에 대한 분석 효과를 적용할 수 있다. 관광 데이터 셋 분석에서는 내/외국인 방문객 수와 관광수입과의 상관계수를 계산하고 단순 선형 회귀 분석으로 계산한 결과를 화면에 표시하여 비교 분석한다.

### > 시각화 방법

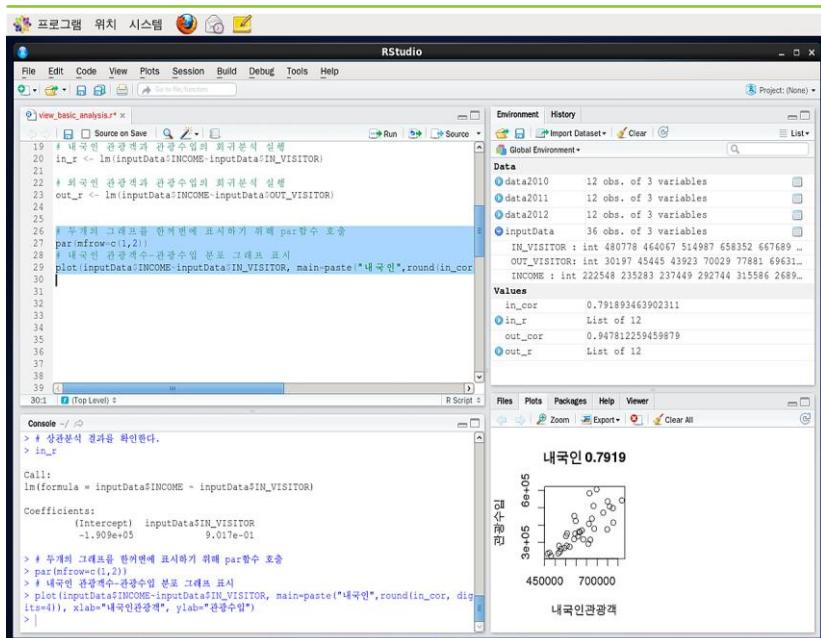
- 내/외국인 관광객 수를 가로 축으로, 관광 수입을 세로축으로 하는 분포 그래프를 작성한다.
- 회귀분석의 결과를 활용하여 분포 그래프 위에 선형 회귀 분석 함수를 표현하여 시각화한다.
- 시각화된 내/외국인의 그래프를 비교 분석한다.

## ▶ 시각화 과정



## ▶ 분석 데이터 시각화

### ▶ 데이터 시각화



1. 내국인에 대한 분석 결과를 표현하기 위해 스크립트를 작성하고 실행한다.

- `par` 함수를 사용하여 가로로 두 개의 그래프를 한꺼번에 표시하도록 설정한다.
- `plot` 함수를 사용하여 내국인 관광객 수–관광수입의 분포 그래프를 표시한다.

```

01. #두개의 그래프를 한꺼번에 표시하기 위해 par 함수 호출
02. par(mfrow=c(1,2))
03. #내국인 관광객수–관광수입 분포 그래프 표시
04. plot(inputData$INCOME~inputData$IN_VISITOR, main=paste("내국인",
   ↵ round(in_cor, digits=4)), xlab="내국인관광객", ylab="관광수입")

```



- 관광 분석 및 시각화 R 스크립트 소스(`view_basic_analysis.R`)
- 라인 02 : `par` 함수를 사용하여 가로 1개, 세로 2개의 그래프를 그리도록 설정하는 라인이다.
- 라인 04 : `plot` 함수를 사용하여 내국인 관광객수와 관광수입의 분포를 그래프로 시각화하는 라인이다.

The screenshot shows the RStudio interface with the following details:

- File**, **Edit**, **Code**, **View**, **Plots**, **Session**, **Build**, **Debug**, **Tools**, **Help** menu items.
- RStudio** title bar.
- Project: (None)** in the top right.
- Environment** and **History** tabs in the top right.
- Source on Save** button in the top left.
- view\_basic\_analysis.r** file open in the script editor.
- Code Editor** pane showing R code:

```
1 # 내국인 관광객과 관광수입의 회귀분석 실행
2 in_r <- lm(inputData$INCOME~inputData$IN_VISITOR)
3
4 # 외국인 관광객과 관광수입의 회귀분석 실행
5 out_r <- lm(inputData$INCOME~inputData$OUT_VISITOR)
6
7 # 두개의 그레프를 한꺼번에 표시하기 위해 par함수 호출
8 par(mfrow=c(1,2))
9 # 내국인 관광객수-관광수입 분포 그레프 표시
10 plot(inputData$INCOME~inputData$IN_VISITOR, main=paste("내국인",round(in_r$cor
11
12 # 선형 회귀분석 표현
13 abline(in_r)
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
```

- Console** tab active.
- Call:** lm(formula = inputData\$INCOME ~ inputData\$IN\_VISITOR)
- Coefficients:**

	(Intercept)	inputData\$IN_VISITOR
	9.017e-01	-1.909e+05

- Plot:** A scatter plot titled "내국인 관광객수-관광수입 분포" showing a positive linear relationship between IN\_VISITOR (X-axis, 450,000 to 700,000) and INCOME (Y-axis, 6e+05 to 8e+05).
- Environment** pane:

  - data2010: 12 obs. of 3 variables
  - data2011: 12 obs. of 3 variables
  - data2012: 12 obs. of 3 variables
  - inputData: 36 obs. of 3 variables
    - IN\_VISITOR: int 480778 460607 514987 658352 667689 ...
    - OUT\_VISITOR: int 30197 45445 43923 70029 77881 69631 ...
    - INCOME : int 222548 235283 237449 292744 315586 2689 ...

- Values** pane:

	inCor	in_r	outCor	out_r
inCor	0.79189346390231			
in_r		List of 12		
outCor			0.947812259459879	
out_r				List of 12

- Plots** tab active.
- Zoom**, **Export**, **Clear All** buttons in the bottom right.

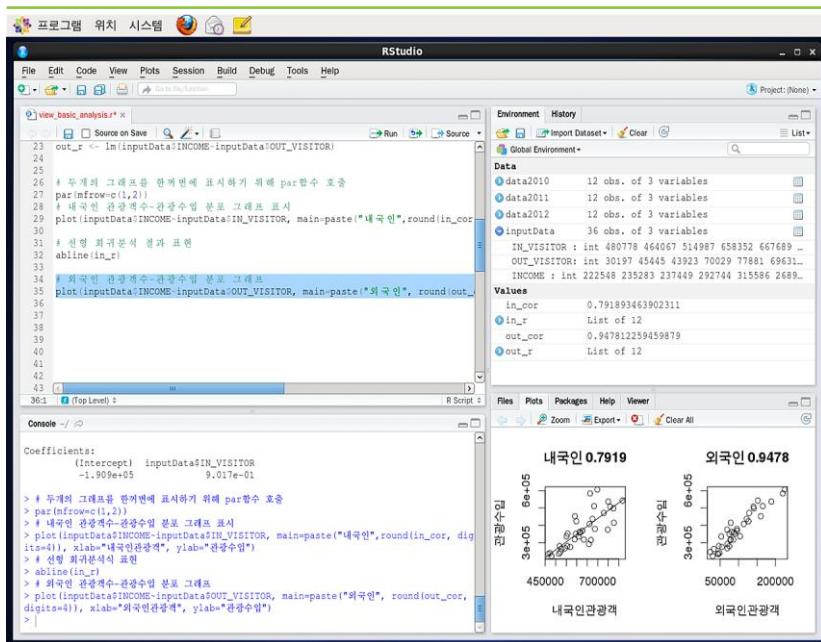
- abline 함수를 사용하여 분석 단계에서 계산해 둔 내국인 관광객과 관광수입과의 상관관계식(in\_r)을 분포 그래프 위에 표현한다.

- 01. #선형 회귀분석 결과 표현
  - 02. abline(in\_r)



- 관광 분석 및 시각화 R 스크립트 소스(`view_basic_analysis.R`)
  - 라인 02 : `abline` 함수를 사용하여 선형 회귀 분석 결과를 분포 그래프 위에 시각화하는 라인이다.

## VI. 시각화



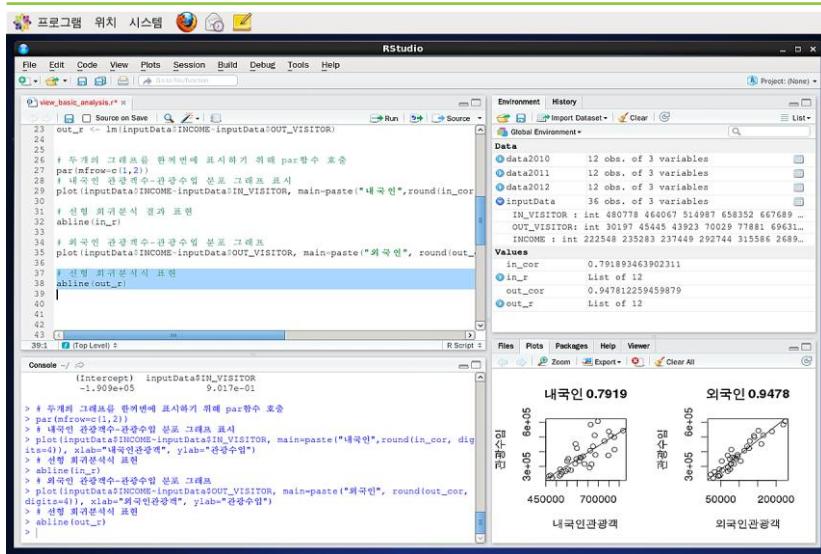
2. 외국인에 대한 분석 결과를 표현하기 위한 스크립트를 작성하고 실행한다.

- par 함수를 사용하여 가로로 두 개의 그래프를 한꺼번에 표시하도록 설정한다.
- plot 함수를 사용하여 내국인 관광객 수-관광수입의 분포 그래프를 표시한다.

01. #외국인 관광객수-관광수입 분포 그래프 표시
02. plot(inputData\$INCOME~inputData\$OUT\_VISITOR, main=paste("외국인",  
    ↳ round(out\_cor, digits=4)), xlab="외국인관광객", ylab="관광수입")



- 관광 분석 및 시각화 R 스크립트 소스(view\_basic\_analysis.R)
- 라인 02 : lplot 함수를 사용하여 외국인 관광객수와 관광수입의 분포를 그래프로 시각화하는 라인이다.



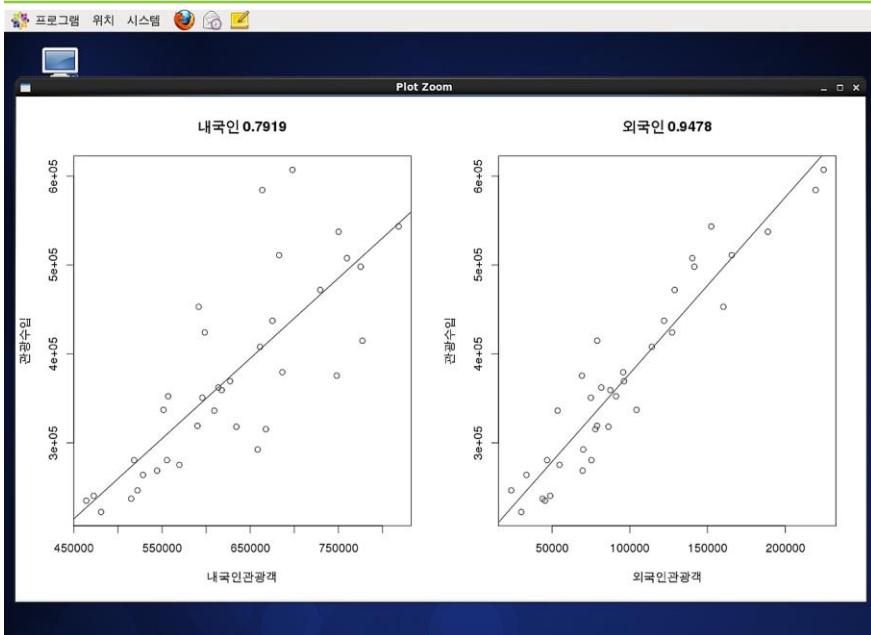
- `abline` 함수를 사용하여 분석 단계에서 계산해 둔 외국인 관광객과 관광수입과의 상관관계식(`out_r`)을 분포 그래프 위에 표현한다.
- 분석 결과 그래프를 Zoom 하여 표시하면 위와 같다.

01. #선형 회귀분석 결과 표현
02. `abline(out_r)`



- 관광 분석 및 시각화 R 스크립트 소스(`view_basic_analysis.R`)
- 라인 02 : `lplot` 함수를 사용하여 외국인 관광객수와 관광수입의 분포를 그래프로 시각화하는 라인이다.

## ▶ 데이터 분석



- **상관 계수 분석** : 내국인 관광객 수와 관광지 수입의 상관계수는 0.7919로 정의 관계에 있고, 외국인 관광객 수와 관광지 수입의 상관계수는 0.94778로 역시 정의 관계에 있다.
- 관광객 수가 늘어나게 되면 관광지 수입이 늘어나는 것은 자명 한 일이다. 그러나 내국인 관광객의 수가 압도적으로 많음에도 불구하고 관광지의 관광수입과의 상관 관계는 외국인 관광객의 숫자와 더 연관이 깊음을 알 수 있다.
- 두 개의 상관 그래프에서도 알 수 있듯이 외국인 그래프 쪽의 기울기가 더 높은 것을 알 수 있다.
- **이를 통해**, 총 관광객 수는 내국인이 더 많지만, 유료 관광지의 방문 비율은 외국인이 훨씬 높음을 알 수 있다. 따라서, 제주도의 관광지 수입을 빠르게 높이기 위해서는 외국인의 관광객 유입을 늘리는 것이 유리하며, 관광객 수가 압도적으로 많은 내국인 관광객들을 대상으로는 관광지에 대한 대대적인 홍보와 프로모션을 통해 유료 관광지 유입률을 늘리는 전략을 취해야 함을 알 수 있다.

- **상관분석(Correlation Analysis)** : 확률론과 통계학에서 두 변수간에 어떤 선형적 관계를 갖고 있는지를 분석하는 방법이다. 두변수는 서로 독립적인 관계로부터 서로 상관된 관계일 수 있으며 이때 두 변수간의 관계의 강도를 상관관계(Correlation, Correlation coefficient)라 한다.
- **상관계수(Correlation coefficient)** : 상관계수( $r$ )는 두개의 변수 간의 선형관계의 방향과 강도를 나타내는 변수로,  $-1 \leq r \leq 1$  의 범위를 갖는다.
  - 1이나 1에 가까울수록 강한 상관 관계를 가지며, 0에 가까울수록 상관관계가 적다.
  - 양수이면 정의 상관관계, 음수이면 역의 상관관계가 있음을 나타낸다.



## VII 예제문제

예제 문제1

53

예제 문제2

54

# 예 / 제 / 문 / 제

## 예제 1

내국인 관광객 수와 외국인 관광객의 월간 방문객 수를 연도별로 비교 분석하라.

- 제주도 관광 데이터에서 월간 방문 객수를 꺾은선 차트로 표현하고, 내국인과 외국인의 방문 패턴을 비교 분석하라.

- 연도별로 저장된 관광 데이터를 로드한다.
- R 스크립트를 활용하여 3년간 1월~12월 방문객 수를 꺾은선 그래프로 표현한다.
- 내국인과 외국인의 방문 성수기, 비성수기를 구분하고, 그 원인에 대해 추론해 본다.

## 예제 2

### 내국인 관광객 수와 외국인 관광객 수 사이의 상관관계를 분석하라.

- 2010년~2012년 3년간의 관광 데이터를 기반으로 내국인 관광객과 외국인 관광객 수 사이의 상관계수를 구하고, 회귀분석 결과를 시각화하라.

- 연도별로 저장된 데이터를 로드한다.
- R Studio를 활용하여 3년간의 데이터를 통합 가공한다.
- 내국인 관광객 수와 외국인 관광객 수 사이의 상관계수를 구한다.
- 내국인 관광객 수와 외국인 관광객 수를 회귀분석하고, 결과를 시각화한다.
- 내/외국인 관광객 수 사이에 어떤 관계를 갖는지 분석하고, 원인을 유추해 본다.



관광 

Intermediate Level

중급과정







## I 개요

개요

59

58

# I

## 개요



### 개요

제주시 통합자료센터(<http://jejudb.jejesi.go.kr/>)로부터 2010년 ~ 2012년 제주도 내외국인 입도 관광객 수 및 주요 관광지 관광 수입 데이터와 교통수단별 제주도 관광객 수를 바탕으로 내국인 유입 관광객 수, 외국인 관광객 수, 항공편 이용 관광객 수, 선편 이용 관광객 수에 대하여 다중 회귀 분석을 통해 네 가지 데이터 각 항목 사이의 상관관계를 분석하여, 제주도 주요 관광지의 관광 수입이 내국인 방문객과 외국인 방문객 중 어느 쪽의 영향을 더 많이 받는지, 항공편 이용 관광객과 선편 이용 관광객 중 어느 쪽의 영향을 많이 받는지 등을 분석한다.



### > 활용 데이터

- **transport\_2010.csv :**  
2010년 교통수단별(항공편, 선편) 제주도 관광객수 데이터
- **transport\_2011.csv :**  
2011년 교통수단별(항공편, 선편) 제주도 관광객수 데이터
- **transport\_2012.csv :**  
2012년 교통수단별(항공편, 선편) 제주도 관광객수 데이터
- **jeju\_2010.csv :**  
2010년 제주도 내외국인 입도 관광객수 및 주요 관광지 관광수입 데이터
- **jeju\_2011.csv :**  
2011년 제주도 내외국인 입도 관광객수 및 주요 관광지 관광수입 데이터
- **jeju\_2012.csv :**  
2012년 제주도 내외국인 입도 관광객수 및 주요 관광지 관광수입 데이터

## > 선행학습

- 리눅스 - 파일시스템 구조, 쉘 명령어, 쉘 스크립트 실행 방법
- R 프로그래밍 언어 - 기본 지식(문법, 패키지 추가 설치 방법)
- 통계 - 상관관계(Correlation), 다중회귀분석(multi regression)
- R 차트 - 구성 방법

## > 요구사항

- 수집된 2010년~2012년 제주도 관광 데이터 셋을 분석에 필요한 저장 공간으로 이동시키고, 내외국인 관광객 및 교통수단별 방문객 수 사이에 존재하는 상관관계를 모두 파악하라.

## > 분석 절차

- 수집된 2010년~2012년 제주도 관광 데이터 셋을 로드한다.
- 연도별로 각각 수집된 내외국인 관광객 데이터를 통합한다.
- 연도별로 각각 수집된 교통수단별 관광객 데이터를 통합한다.
- 통합된 내외국인 관광객 정보와 교통수단별 관광객 데이터를 모두 통합한다.
- 다중 회귀분석을 실행한다.
- 다중 회귀분석 결과를 시각화한다.
- 시각화한 결과를 보고, 각 항목들 간의 상관계수를 비교하여, 내국인 관광객과 외국인 관광객 중 어느 쪽이 주요 관광지 관광 수입에 영향을 미치는지, 항공편 방문객과 선편 방문객 중 어느 편이 관광 수입과 연관되어 있는지 판단한다.



## II 수집

개요	63
수집 데이터	64
데이터 수집	66
데이터 작업 영역 이동 스크립트	69



# 수집

## > 개요

관광 데이터는 제주시 통합자료센터(<http://jejudb.jejesi.go.kr/>)에서 매년 발표하는 종합 통계 연보 데이터로부터 최근 3년 (2010년~2012) 연간의 데이터 및 교통수단별 관광객 데이터 중 분석에 필요한 정보를 추출하여 분석에 용이하게 편집하여 제공한다.

## > 수집 방법

- **데이터 제공 :** 관광 데이터는 제주시 통합자료센터(<http://jejudb.jejesi.go.kr/>)로부터 수동으로 데이터를 수집할 수 있으며, 실습용 자료는 빅데이터 분석 활용센터에 접속하여 수집 데이터 셋을 다운로드할 수 있도록 원시데이터를 제공하고 있다.

The screenshot shows the homepage of the Jeju City Integrated Data Center (<http://jejudb.jejesi.go.kr/>). The main content area features a large image of a library or bookstore. To the right, there are several search and information boxes:

- 도서검색**: A search box for books with a blue book icon.
- 기록물검색**: A search box for historical documents with a blue document icon.
- 도서대출신청**: A button to apply for book lending with a blue book icon.
- 통합 Q&A**: A button for general questions and answers with a blue Q&A icon.

Below these are sections for **도서대출신청**, **기록물검색**, **제주시민**, **제주시민**, and **제주시민**. At the bottom, there is a footer with copyright information and a link to the search results page.

## > 수집 데이터

- 2010년 제주도 내외국인 관광객 수 및 관광수입 데이터(jeju\_2010.csv)

IN_VISITOR	OUT_VISITOR	INCOME
480778	30197	222548
464067	45445	235283
514987	43923	237449
658352	70029	292744
667689	77881	315586
544390	69631	268960
590093	78914	319334
686331	95752	379581
518338	75337	280745
634212	86257	318276
569616	54819	275501
472448	48815	240707

순서대로 2010년 1월~12월 데이터를 표현하고 있다.

- IN\_VISITOR : 내국인 관광객 수(명)
- OUT\_VISITOR : 외국인 관광객 수(명)
- INCOME : 관광수입(백만 원)

## II. 수집

### ➤ 2010년 제주도 교통수단별 관광객 수 데이터(transport\_2010.csv)

DATE	AIR	SHIP
201001	451220	59755
201002	466066	43446
201003	503338	55572
201004	631555	96826
201005	644118	101452
201006	546866	67155
201007	576129	92878
201008	660888	121195
201009	517101	76574
201010	622093	98376
201011	552150	72285
201012	476434	44829

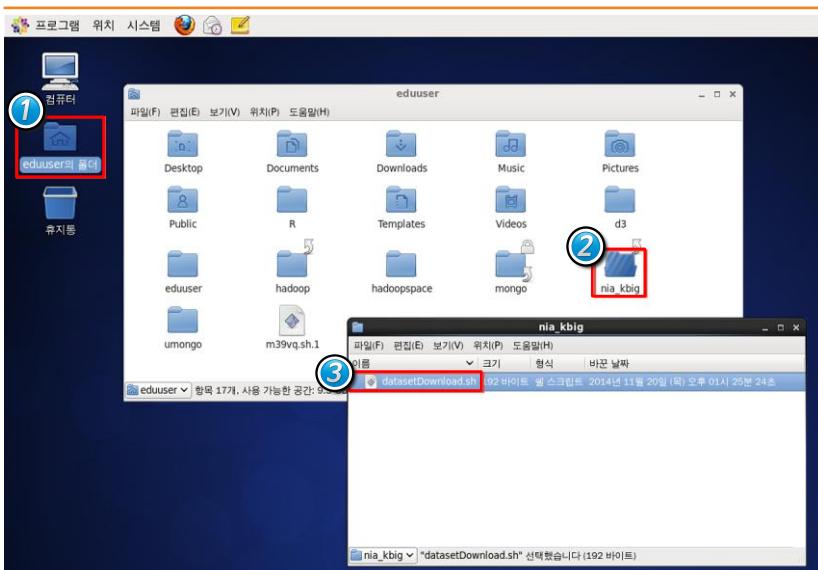
▪ AIR : 항공편 이용 관광객 수(명)

▪ SHIP : 선편 이용 관광객 수(명)

## > 데이터 수집(datasetDownload.sh)

- 데이터 저장소에서 서버 로컬로 관광 데이터 셋을 복사해 온다.
  - **transport\_2010.csv** : 2010년 교통수단별(항공편, 선편) 제주도 관광객수 데이터
  - **transport\_2011.csv** : 2011년 교통수단별(항공편, 선편) 제주도 관광객수 데이터
  - **transport\_2012.csv** : 2012년 교통수단별(항공편, 선편) 제주도 관광객수 데이터
  - **jeju\_2010.csv** :  
2010년 제주도 내외국인 입도 관광객수 및 주요 관광지 관광수입 데이터
  - **jeju\_2011.csv** :  
2011년 제주도 내외국인 입도 관광객수 및 주요 관광지 관광수입 데이터
  - **jeju\_2012.csv** :  
2012년 제주도 내외국인 입도 관광객수 및 주요 관광지 관광수입 데이터

## > 실습코드 디렉토리로 이동

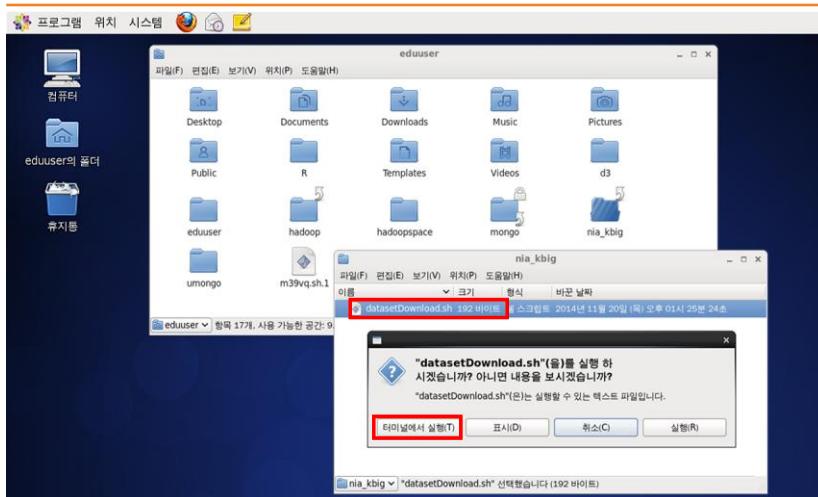


- ① 로그인 후 바탕화면에서 eduuser 폴더를 오픈한다.
- ② nia\_kbig 폴더를 오픈한다.
- ③ datasetDownload.sh를 더블클릭하여 실행한다.

## II. 수집

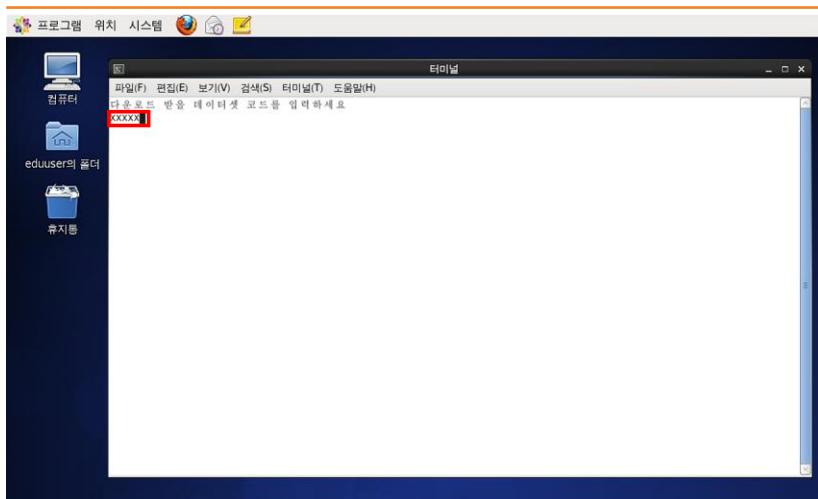
### ▶ 레파지토리에서 데이터 수집

datasetDownload.sh (원시데이터로 컬서버로 복사)



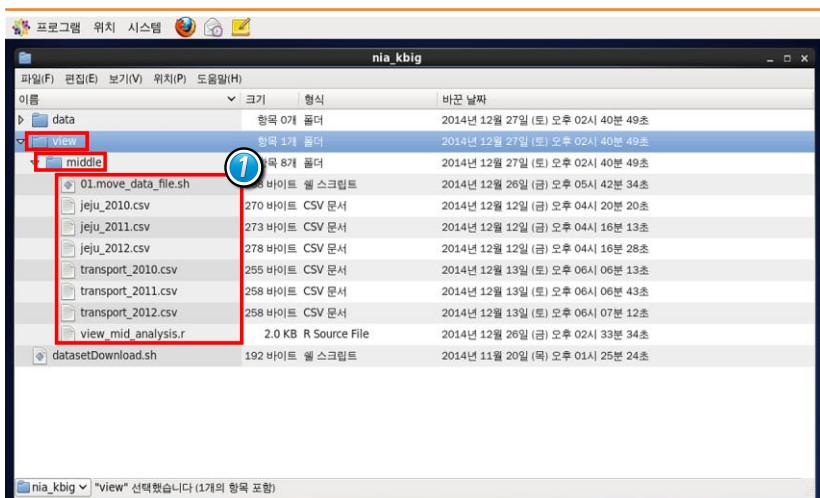
- '터미널에서 실행' 버튼을 클릭한다.

### ▶ 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 :**  
로컬로 수집해온 데이터를 작업 영역 Data 폴더로 자료를 이동하는 스크립트
- **view\_mid\_analysis.R :**  
관광 데이터 분석용 R 스크립트.
- **transport\_2010.csv :**  
2010년 교통수단별(항공편, 선편) 제주도 관광객수 데이터
- **transport\_2011.csv :**  
2011년 교통수단별(항공편, 선편) 제주도 관광객수 데이터
- **transport\_2012.csv :**  
2012년 교통수단별(항공편, 선편) 제주도 관광객수 데이터
- **jeju\_2010.csv :**  
2010년 제주도 내외국인 입도 관광객수 및 주요 관광지 관광수입 데이터
- **jeju\_2011.csv :**  
2011년 제주도 내외국인 입도 관광객수 및 주요 관광지 관광수입 데이터
- **jeju\_2012.csv :**  
2012년 제주도 내외국인 입도 관광객수 및 주요 관광지 관광수입 데이터

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화

## II. 수집

### ▶ 데이터 작업 영역 이동 스크립트(01.move\_data\_file.sh)

#### ▶ 데이터 이동 스크립트

- 로컬로 수집해온 데이터를 작업 영역 Data 폴더로 자료를 이동하는 스크립트

#### 01.move\_data\_file.sh

```
01. #!/bin/bash  
02. TARGET_JEJU=/home/eduuser/nia_kbig/view/middle/j*.csv  
03. # 작업 디렉토리 정의  
04. LOCAL_DIR=/home/eduuser/nia_kbig/data/  
05. mv $TARGET_JEJU $LOCAL_DIR
```



#### 부연설명

- 데이터 작업 영역 이동 스크립트 소스(01.move\_data\_file.sh)
- 라인 02 : 다운로드 받은 원시데이터 파일들의 위치(path)를 변수(TARGET\_JEJU)로 지정하는 라인이다.
- 라인 04 : 작업영역 디렉토리의 위치(path)를 변수(LOCAL\_DIR)로 지정하는 라인이다.
- 라인 05 : mv 명령어를 사용하여 다운로드 받은 원시데이터 파일들을 작업영역 디렉토리로 이동시키는 라인이다.

## ▶ 수집 데이터 셋 작업 영역 폴더 이동

- R Studio에서 상관분석 및 다중 회귀분석을 위한 수집된 데이터 셋을 작업 영역 Data 폴더로 자료를 이동

```

eduuser@localhost:~/nia_kbig/view/middle
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
mall      view_mid_analysis.r  문서   사진
nia_kbig  공개      바탕화면 음악
| eduuser@localhost ~]$ cd nia_kbig/
| eduuser@localhost nia_kbig]$ ls
00.data_download.sh  data  dist  mall  view
| eduuser@localhost nia_kbig]$ ll
합계 16
-rwxr-xr-x  1 edusuuser edusuuser 0 2014-12-04 17:30 00.data_download.sh
drwxrwxr-x  2 edusuuser edusuuser 4096 2014-12-13 18:08 data
drwxrwxr-x  4 edusuuser edusuuser 4096 2014-12-09 20:45 dist
drwxrwxr-x  4 edusuuser edusuuser 4096 2014-12-04 22:20 mall
drwxrwxr-x  4 edusuuser edusuuser 4096 2014-12-15 08:46 view
| eduuser@localhost nia_kbig]$ ll
합계 32
-rwxr-xr-x  1 edusuuser edusuuser 213 2014-12-15 08:47 01.move_data_file.sh
-rw-r--r--  1 edusuuser edusuuser 216 2014-12-14 08:46 01.move_data_file.sh~
-rw-r--r--  1 edusuuser edusuuser 270 2014-12-12 16:20 jeju_2010.csv
-rw-r--r--  1 edusuuser edusuuser 273 2014-12-12 16:16 jeju_2011.csv
-rw-r--r--  1 edusuuser edusuuser 278 2014-12-12 16:16 jeju_2012.csv
-rw-r--r--  1 edusuuser edusuuser 255 2014-12-13 18:06 transport_2010.csv
-rw-r--r--  1 edusuuser edusuuser 258 2014-12-13 18:06 transport_2011.csv
-rw-r--r--  1 edusuuser edusuuser 258 2014-12-13 18:07 transport_2012.csv
| eduuser@localhost middle]$ ./01.move_data_file.sh

```

5. ① ②에서와 같이 ./01.move\_data\_file.sh를 입력하여 준비된 관광 데이터를 이동시킨다.





개요

73

데이터 가공 R 스크립트

76



## 가공

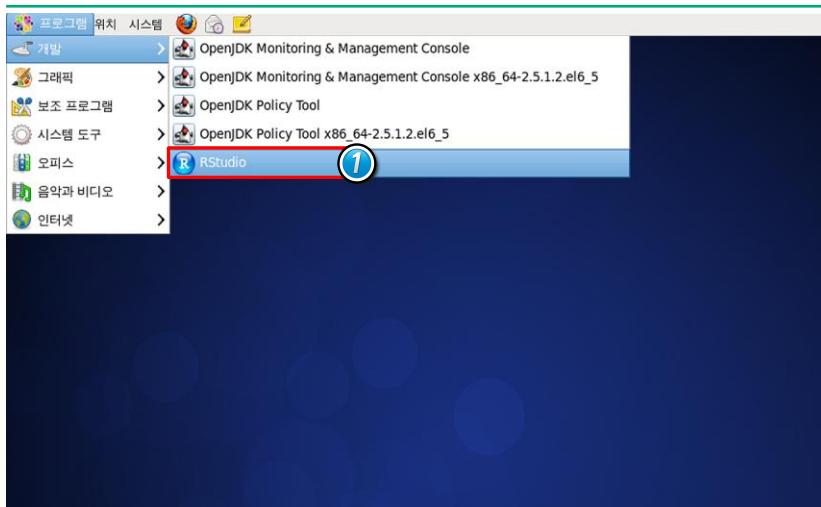
### ▶ 개요

작업 영역 폴더에 복사한 관광 데이터의 가공은, 연도별로 수집된 3년간(2010년~2012년)의 내외국인 제주도 관광객 데이터와 교통수단별(항공편, 선편) 관광객 데이터를 각각 로드하여 하나의 데이터로 통합하여 상관분석(연관도 분석) 및 다중 회귀분석에 적합한 형태로 가공한다.

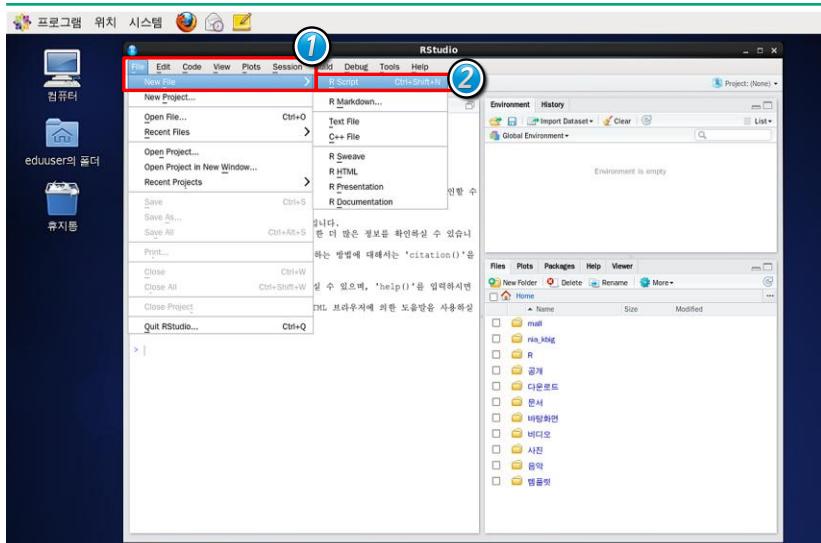
### ▶ 가공 방법

- **분석 도구 실행** : 가공 분석을 위해, 프로그래밍 도구인 R Studio를 실행한다.
- **데이터 로드** : 3년간의 내외국인 제주도 관광객 데이터와 교통수단별(항공편, 선편) 관광객 데이터를 읽어들인다.
- **데이터 통합** : R Studio는 동일한 형태의 데이터를 통합할 수 있는 함수를 제공한다. 이를 활용하여 각각 읽어들인 3년간의 데이터를 하나로 통합한다.

## ▶ 데이터 가공



- ① 원쪽 상단의 [“프로그램” 클릭] > [개발] 클릭 > [“RStudio” 클릭]으로 분석 도구인 R Studio를 실행한다.



- ① ② 관광 데이터의 분석 및 가공을 위해 프로그램 작업 파일 (“New File” 클릭) > “R\_Script” 클릭)을 선택한다.

### III. 가공

```
Untitled.R
1 # 내외국인 관광객 방문객 데이터 모드
2 data_2010 <- read.csv("data/jeju_2010.csv")
3 data_2011 <- read.csv("data/jeju_2011.csv")
4 data_2012 <- read.csv("data/jeju_2012.csv")
5
6 # 교통수단별 관광객 데이터 모드
7 transport_2010 <- read.csv("data/transport_2010.csv", header=TRUE)
8 transport_2011 <- read.csv("data/transport_2011.csv", header=TRUE)
9 transport_2012 <- read.csv("data/transport_2012.csv", header=TRUE)
10
11
12
13
14
15
16
17
```

3. 2010년~2012년 내외국인 제주도 관광 데이터 및 교통수단별 관광객 데이터를 각각 로드하도록 R 스크립트를 작성한다.

```
Untitled.R
1 # 내외국인 관광객 데이터 모드
2 data_2010 <- read.csv("/home/eduser/nia_kb1g/data/jeju_2010.csv")
3 data_2011 <- read.csv("/home/eduser/nia_kb1g/data/jeju_2011.csv")
4 data_2012 <- read.csv("/home/eduser/nia_kb1g/data/jeju_2012.csv")
5
6 # 교통수단별 관광객 데이터 모드
7 transport_2010 <- read.csv("/home/eduser/nia_kb1g/data/transport_2010.csv",
8 transport_2011 <- read.csv("/home/eduser/nia_kb1g/data/transport_2011.csv",
9 transport_2012 <- read.csv("/home/eduser/nia_kb1g/data/transport_2012.csv",
10
11
12
13
14
15
16
17
```

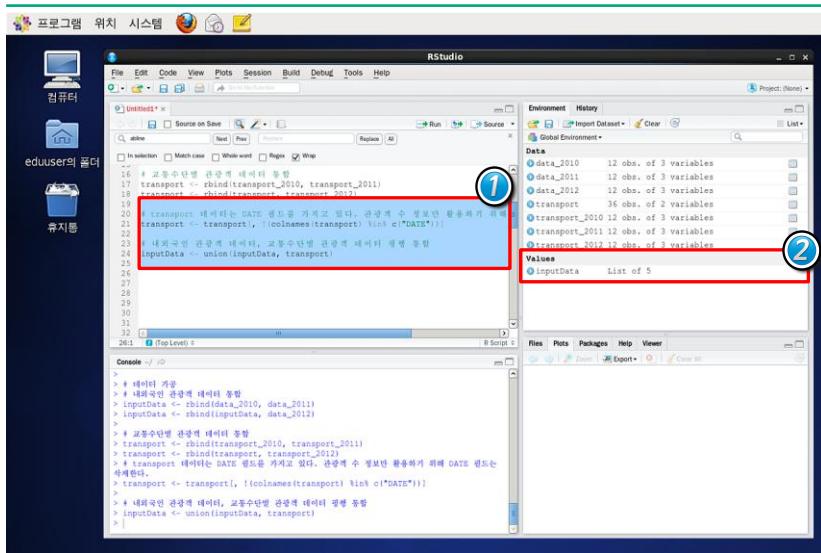
4. ① 현재까지 작성한 스크립트 코드를 선택하여 Ctrl+Enter 를 입력하면, ②와 같이 데이터가 로드된 것을 볼 수 있다.(코드의 부분 실행은 R 스크립트만의 장점이다.)

## ▶ 데이터 가공 R 스크립트

```

File Edit Code View Plots Session Build Debug Tools Help
RStudio
Source Editor * Data
File Edit Code View Plots Session Build Debug Tools Help
Source Editor * Data
Source On Save Test Run Replace All
In selection Match case Whole word Regular Wrap
7 transport_2010 <- read.csv(*"/home/eduuser/nia_kbig/data/transport_2010.csv")
8 inputData <- rbind(data_2010, transport_2010)
9 transport_2011 <- read.csv(*"/home/eduuser/nia_kbig/data/transport_2011.csv")
10 transport_2012 <- read.csv(*"/home/eduuser/nia_kbig/data/transport_2012.csv")
11 # 내외국인 관광객 데이터 통합
12 inputData <- rbind(data_2010, data_2011)
13 inputData <- rbind(inputData, data_2012)
14 #
15 # 교통수단별 관광객 데이터 통합
16 transport <- rbind(transport_2010, transport_2011)
17 transport <- rbind(transport, transport_2012)
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
279
280
281
282
283
284
285
286
287
288
289
289
290
291
292
293
294
295
296
297
298
299
299
300
301
302
303
304
305
306
307
308
309
309
310
311
312
313
314
315
316
317
318
319
319
320
321
322
323
324
325
326
327
328
329
329
330
331
332
333
334
335
336
337
338
339
339
340
341
342
343
344
345
346
347
348
349
349
350
351
352
353
354
355
356
357
358
359
359
360
361
362
363
364
365
366
367
368
369
369
370
371
372
373
374
375
376
377
378
379
379
380
381
382
383
384
385
386
387
388
389
389
390
391
392
393
394
395
396
397
398
399
399
400
401
402
403
404
405
406
407
408
409
409
410
411
412
413
414
415
416
417
418
419
419
420
421
422
423
424
425
426
427
428
429
429
430
431
432
433
434
435
436
437
438
439
439
440
441
442
443
444
445
446
447
448
449
449
450
451
452
453
454
455
456
457
458
459
459
460
461
462
463
464
465
466
467
468
469
469
470
471
472
473
474
475
476
477
478
479
479
480
481
482
483
484
485
486
487
488
489
489
490
491
492
493
494
495
496
497
498
499
499
500
501
502
503
504
505
506
507
508
509
509
510
511
512
513
514
515
516
517
518
519
519
520
521
522
523
524
525
526
527
528
529
529
530
531
532
533
534
535
536
537
538
539
539
540
541
542
543
544
545
546
547
548
549
549
550
551
552
553
554
555
556
557
558
559
559
560
561
562
563
564
565
566
567
568
569
569
570
571
572
573
574
575
576
577
578
579
579
580
581
582
583
584
585
586
587
588
589
589
590
591
592
593
594
595
596
597
598
599
599
600
601
602
603
604
605
606
607
608
609
609
610
611
612
613
614
615
616
617
618
619
619
620
621
622
623
624
625
626
627
628
629
629
630
631
632
633
634
635
636
637
638
639
639
640
641
642
643
644
645
646
647
648
649
649
650
651
652
653
654
655
656
657
658
659
659
660
661
662
663
664
665
666
667
668
669
669
670
671
672
673
674
675
676
677
678
679
679
680
681
682
683
684
685
686
687
688
689
689
690
691
692
693
694
695
696
697
697
698
699
700
701
702
703
704
705
706
707
708
709
709
710
711
712
713
714
715
716
717
718
719
719
720
721
722
723
724
725
726
727
728
729
729
730
731
732
733
734
735
736
737
738
739
739
740
741
742
743
744
745
746
747
748
749
749
750
751
752
753
754
755
756
757
758
759
759
760
761
762
763
764
765
766
767
768
769
769
770
771
772
773
774
775
776
777
778
779
779
780
781
782
783
784
785
786
787
788
789
789
790
791
792
793
794
795
796
797
797
798
799
800
801
802
803
804
805
806
807
808
809
809
810
811
812
813
814
815
816
817
818
819
819
820
821
822
823
824
825
826
827
828
829
829
830
831
832
833
834
835
836
837
838
839
839
840
841
842
843
844
845
846
847
848
849
849
850
851
852
853
854
855
856
857
858
859
859
860
861
862
863
864
865
866
867
868
869
869
870
871
872
873
874
875
876
877
878
879
879
880
881
882
883
884
885
886
887
888
888
889
889
890
891
892
893
894
895
896
897
897
898
899
900
901
902
903
904
905
906
907
908
909
909
910
911
912
913
914
915
916
917
917
918
919
920
921
922
923
924
925
926
927
928
929
929
930
931
932
933
934
935
936
937
938
939
939
940
941
942
943
944
945
946
947
947
948
949
950
951
952
953
954
955
956
957
958
959
959
960
961
962
963
964
965
966
967
967
968
969
970
971
972
973
974
975
976
977
978
979
979
980
981
982
983
984
985
986
987
987
988
989
989
990
991
992
993
994
995
996
997
998
999
999
1000
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1088
1089
1089
1090
1091
1092
1093
1094
1095
1096
1096
1097
1098
1099
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1188
1189
1189
1190
1191
1192
1193
1194
1195
1196
1197
1197
1198
1199
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1248
1249
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1278
1279
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1288
1289
1289
1290
1291
1292
1293
1294
1295
1296
1297
1297
1298
1299
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1348
1349
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1378
1379
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1388
1389
1389
1390
1391
1392
1393
1394
1395
1396
1397
1397
1398
1399
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1418
1419
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1428
1429
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1438
1439
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1448
1449
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1468
1469
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1478
1479
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1488
1489
1489
1490
1491
1492
1493
1494
1495
1496
1497
1497
1498
1499
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1518
1519
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1528
1529
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1538
1539
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1548
1549
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1568
1569
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1578
1579
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1588
1589
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1598
1599
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1618
1619
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1628
1629
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1638
1639
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1648
1649
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1668
1669
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1678
1679
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1688
1689
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1698
1699
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1718
1719
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1728
1729
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1738
1739
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1748
1749
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1768
1769
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1778
1779
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1788
1789
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1798
1799
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1818
1819
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1828
1829
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1838
1839
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1848
1849
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1868
1869
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1878
1879
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1888
1889
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1898
1899
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1918
1919
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1928
1929
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1938
1939
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1948
1949
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1968
1969
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1978
1979
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1988
1989
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2018
2019
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2048
2049
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2068
2069
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2078
2079
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2088
2089
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2098
2099
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2148
2149
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2178
2179
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2188
2189
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2198
2199
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2248
2249
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2278
2279
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2288
2289
2289
2290
2291
2292
2293
2294
2295
2296
2297
2297
2298
2299
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2309
2310
2311
2312
231
```

### III. 가공



6. ① 교통수단별 데이터는 DATE 필드를 가지고 있는데, 다중 회귀 분석을 하기 위해 해당 필드는 삭제한 후에 inputData 와 transport 데이터를 패러렐하게 통합한다.  
② 통합된 데이터를 확인할 수 있다.
- #주) union 함수는 동일한 여러 컬럼의 데이터를 패러렐하게 합쳐 주는 기능을 수행하는 함수이다.

```
01. # 교통수단별 관광객 데이터는 DATE 필드를 가지고 있다. 관광객 수 정보만 활용하기  
02. ↪ 위해 DATE 필드는 삭제한다.  
03. transport <- transport[, !(colnames(transport) %in% c("DATE"))]  
04.  
05. # 내외국인 관광객 데이터와 교통수단별 관광객 데이터 통합  
06. inputData <- union(inputData, transport)  
07.
```



- 관광 분석 및 시각화 R 스크립트 소스([view\\_middle\\_analysis.R](#))
- 라인 03 : 교통수단별 관광객 데이터에서 DATE 컬럼을 삭제하는 라인이다.
- 라인 06 : union 함수를 사용하여 내외국인 관광객 데이터와 교통수단별 관광객 데이터를 통합하는 라인이다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화



## IV 저 장

개요	81
R Studio 활용 저장	82

# IV

## 저장

### > 개요

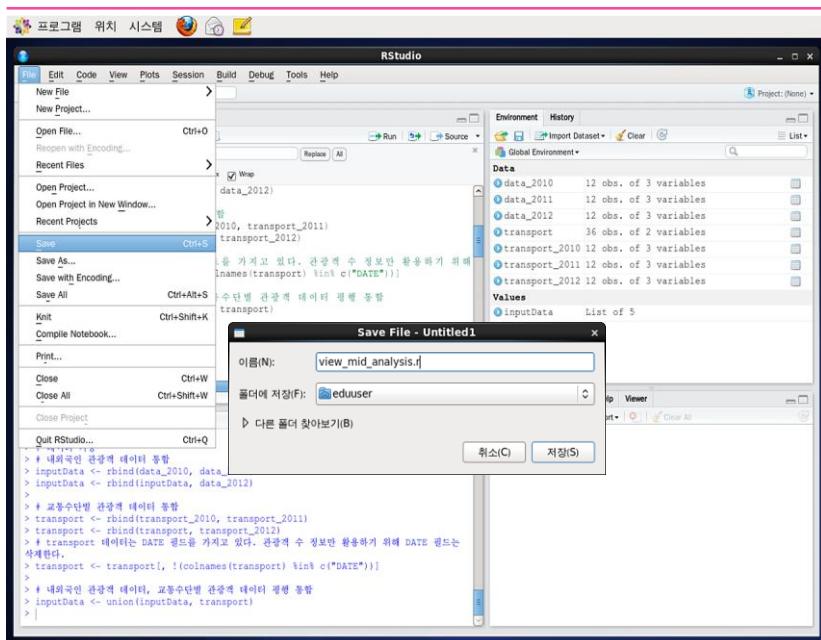
R Studio를 활용하여 데이터 로드 > 가공 > 분석 > 시각화 단계를 한꺼번에 실행하므로, 별도의 저장 과정은 생략한다. 가공된 데이터는 메모리상에 존재하며, 지금까지 작성된 분석 프로그램 소스를 저장한다.

### > 저장 방법

- **가공된 데이터 메모리 저장** : 관광 데이터 분석을 위해 가공한 데이터는 R Studio 메모리상에 저장된다.
- **소스 저장** : 작성 중인 제주도 관광 데이터 분석 프로그램을 저장한다.

## > R Studio 활용 저장

### > 데이터 저장



1. 관광 데이터 분석을 위해 작성 중인 프로그램 소스를 저장한다
- #주) 작성 중인 프로그램 소스를 저장하는 방법은 메뉴의 "File" > "Save"를 이용하거나 도구상자의 저장 아이콘을 이용한다. 저장시 저장 위치 및 파일명은 "/home/eduuser/nia\_kbig/view\_mid\_analysis.r"로 저장한다.

# W





## V 분석

개요	85
R Studio 활용 분석	87
R Studio 저장	90

# V 분석

## > 개요

관광 데이터 분석은 R Studio에 내장된 상관관계 분석 기법을 활용하여 내외국인 관광객 수와 외국인 관광객 수, 항공편 이용 관광객 수와 선편 이용 관광객 수, 관광수입 간에 다중 회귀 분석을 활용하여 관광객의 내외국인 구분과 교통수단에 따라서 관광 수입에 영향을 많이 주는 팩터를 찾아낸다.

## > 데이터 분석 방법

- **다중 회귀 분석** : 두 개 이상의 시계열 데이터 사이의 상관관계를 계산하는 통계 기법을 사용한다.
- **상관계수 계산** : 내외국인 관광객과 관광수입, 교통수단별 관광객 간의 상관계수를 계산한다.(R Studio의 cor 함수 활용)
- **다중회귀분석 함수** : R Studio의 pair 함수를 활용하여 다중 상관 분석을 한다.



- **상관분석(Correlation Analysis)** : 확률론과 통계학에서 두 변수 간에 어떤 선형적 관계를 갖고 있는지를 분석하는 방법이다. 두 변수는 서로 독립적인 관계로부터 서로 상관된 관계일 수 있으며 이때 두 변수 간의 관계의 강도를 상관관계(Correlation, Correlation coefficient)라 한다.
- **상관계수(Correlation coefficient)** : 상관계수( $r$ )는 두개의 변수 간의 선형관계의 방향과 강도를 나타내는 변수로,  $-1 \leq r \leq 1$  의 범위를 갖는다.
  - 1이나 1에 가까울수록 강한 상관 관계를 가지며, 0에 가까울수록 상관관계가 적다.
  - 양수이면 정의 상관관계, 음수이면 역의 상관관계가 있음을 나타낸다.
- **회귀분석(regression analysis)** : 둘 또는 그 이상의 변수 사이의 관계 특히 변수 사이의 인과 관계를 분석하는 추측 통계의 한 분야이다. 회귀분석은 특정 변수값의 변화와 다른 변수값의 변화가 가지는 수학적 선형의 함수식을 파악함으로써 상호관계를 추론하게 되는데 추정된 함수식을 회귀식이라고 한다. 이러한 회귀식을 통하여 특정 변수(독립변수 또는 설명변수라고 함)의 변화가 다른 변수(종속변수라고 함)의 변화와 어떤 관련성이 있는지 관련이 있다면 어느 변수의 변화가 원인이 되고 어느 변수의 변화가 결과적인 현상인지 등에 관한 사항을 분석할 수 있다.
- **다중회귀분석(Multiple Regression Analysis)** : 두 변수 이상의 독립변수(영향변수, 원인변수)들이 종속변수(결과변수)에 어떠한 영향을 미치는지를 알기 위한 분석 기법이다. 예를 들면, 골프 용품에 대한 전체 만족도에 골프용품의 타구감, 방향성, 명성 등의 변수들이 어떠한 영향을 미치는지를 알기 위해서는 다중회귀분석을 하여야 한다. 따라서 독립변수들이 종속변수에 미치는 상대적 영향력을 알아볼 수 있으며, 이러한 독립변수 값들의 변화에 따라 종속변수 값이 어떻게 변화하는지를 예측할 수 있다. 독립변수와 종속변수에 쓰인 변수들의 측정척도는 등간척도(interval scale)나 비율척도(ratio scale)의 메트릭(metric) 자료이어야 하나, 독립변수가 명목척도일 경우에는 0과 1을 사용한 더미변수(dummy variables)를 이용하여 분석할 수 있다.

I. 개요

II. 수집

III. 가공

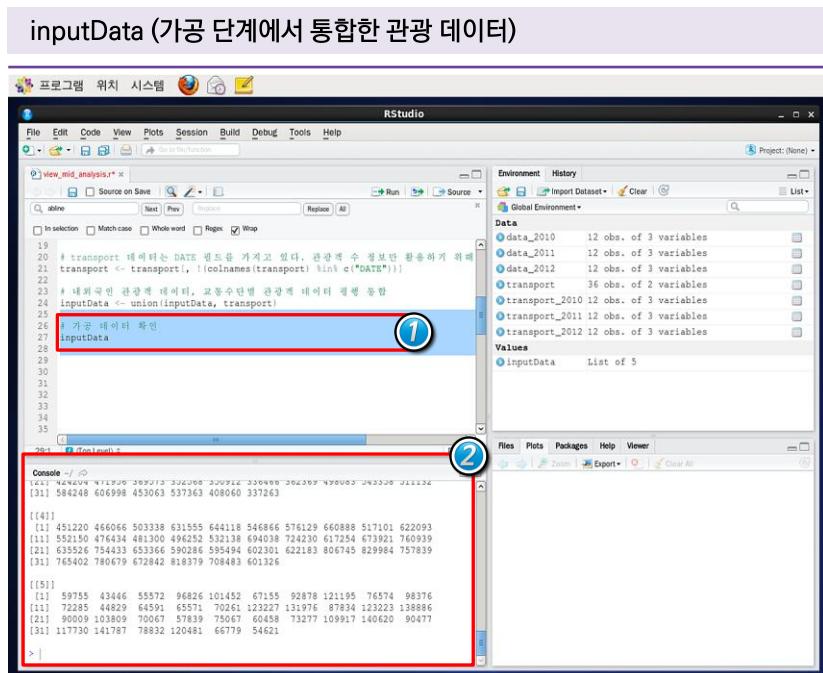
IV. 저장

V. 분석

VI. 시각화

## > R Studio 활용 분석

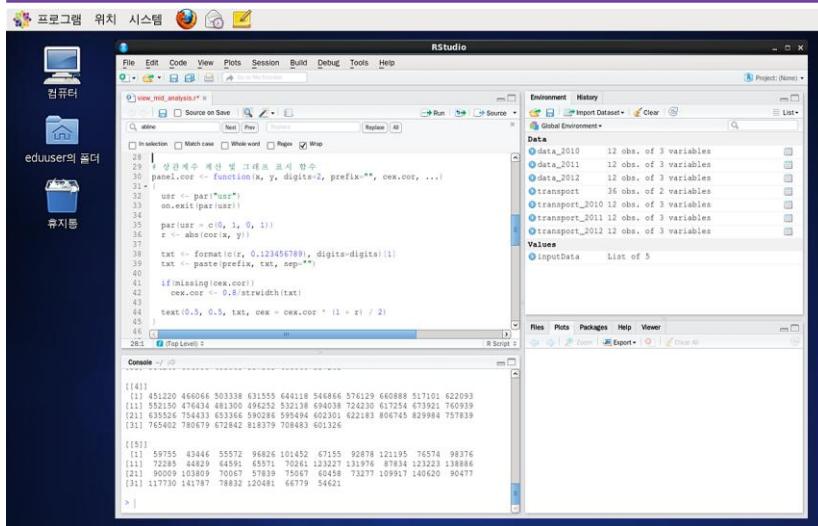
## > 데이터 불러오기



- ① 가공한 데이터가 잘 들어가 있는지 확인하기 위해 “`inputData`”를 입력하고 위와 같이 블록을 선택한 후, `Ctrl+Enter`을 입력하면, ② 와 같이 데이터를 확인할 수 있다.

## ▶ 데이터 분석

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



- 두 개의 벡터(배열) 데이터를 입력받아 cor 함수를 사용하여 그래프 패널에 표현하는 panel.cor 함수를 정의한다.

```

01. # 1:1 상관계수 계산 및 그래프 표시 함수
02. panel.cor <- function(x, y, digits=2, prefix="", cex.cor, ...)
03. {
04.   usr <- par("usr")
05.   on.exit(par(usr))

06.

07.   par usr = c(0, 1, 0, 1)
08.   r <- abs(cor(x, y))

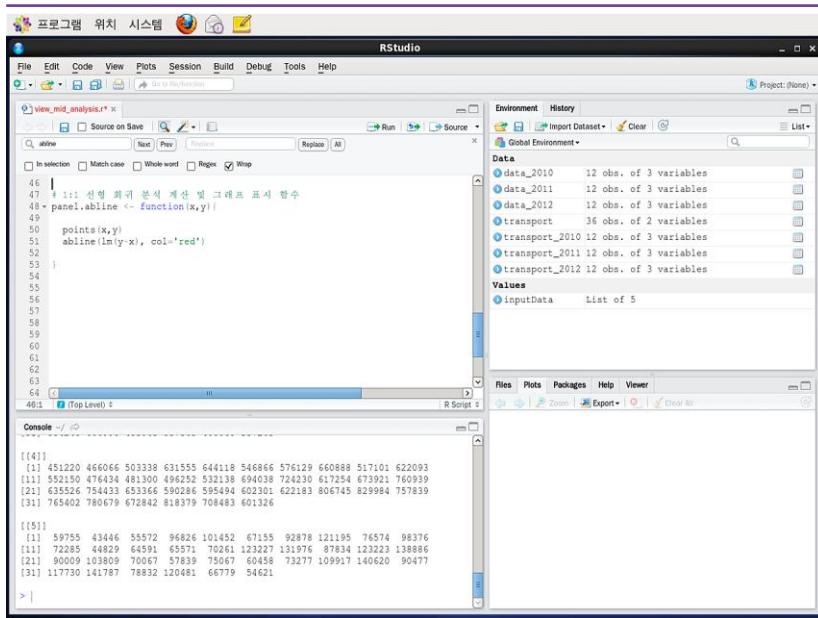
09.

10.   txt <- format(c(r, 0.123456789), digits=digits)[1]
11.   txt <- paste(prefix, txt, sep="")
12.

13.   if(missing(cex.cor))
14.     cex.cor <- 0.8/strwidth(txt)
15.
16.   text(0.5, 0.5, txt, cex = cex.cor * (1 + r) / 2)
17. }

```

## V. 분석



2. 두 개의 벡터(배열) 데이터를 입력받아 lm 함수 및 abline 함수를 사용하여 그래프 패널에 표현하는 panel.abline 함수를 정의한다.

```
01. # 1:1 선형 회귀 분석 계산 및 그래프 표시 함수
02. panel.abline <- function(x,y){
03.   points(x,y)
04.   abline(lm(y~x), col='red')
05. }
```

## > R Studio 저장

### > 분석 결과 저장

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

The screenshot shows the RStudio IDE interface. The menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Tools, Help. The code editor shows R script code. The environment pane shows global variables like data\_2010 through data\_2012, transport, and transport\_2010-2012. The console pane at the bottom shows the execution of R code, including matrix operations and data frame creation.

```
[1:1] 451220 466066 503338 631555 644118 546866 576129 660888 517101 622093
[1:1] 552150 476434 481300 496252 532130 694038 724230 617254 673921 760939
[2:1] 635526 754433 653366 590286 595494 602301 622183 806745 829984 757839
[3:1] 765402 780679 672842 818379 708483 601326
> | [15]
[1:1] 59755 43446 55572 96826 101452 67155 92878 121195 76574 98376
[1:1] 72285 44829 64591 65571 70261 123227 131976 87834 123223 138886
[2:1] 90009 103809 70067 57839 75067 60458 73277 109917 140620 90477
[3:1] 117730 141787 78832 120481 66779 54621
> |
```

- “File/Save”를 클릭하여 지금까지 관광 데이터를 분석하기 위해 작성한 프로그램을 저장한다.

- #주) 최종 분석 및 시각화를 위한 작업은 ‘시각화’ 과정에서 설명한다.



1

2



## VI 시각화

개요	93
데이터 분석	95

# VI

## 시각화

### > 개요

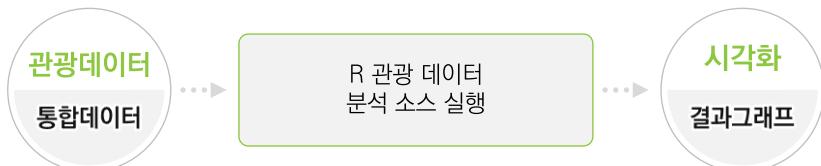
관광 데이터의 분석 과정에서는 내외국인 관광객 수와 주요 관광지 관광수입, 교통수단별 관광객 수 등 여러 데이터들 간의 상관관계를 한 번에 분석하여 나온 결과를 정확하게 시각화한다. 관광객들의 행동 패턴이나 유료관광지 방문 패턴 등을 유추할 수 있다. 다중회귀분석은 오픈오피스 등 다른 프로그램으로도 어렵지 않게 가능하지만, R Studio를 활용하면 분석과 함께 시각화까지 일괄적으로 진행할 수 있다.



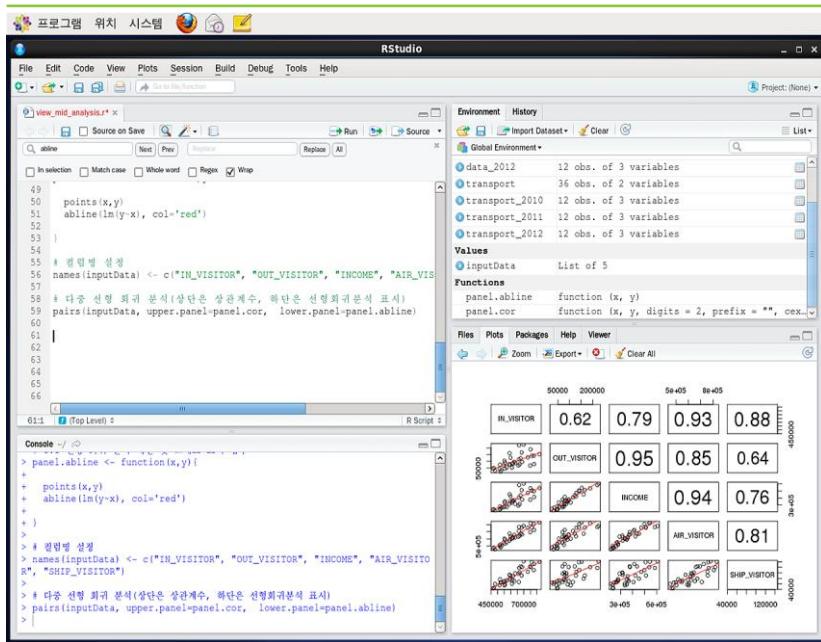
### > 시각화 방법

- 가공 단계에서 생성한 통합 멀티 컬럼 데이터를 활용한다.
- 분석 단계에서 작성한 panel.cor 함수와 panel.abline 함수를 사용하여 다중회귀 분석 및 시각화를 진행한다.
- 시각화된 다중회귀분석 그래프를 보고 결론을 도출한다.

### > 데이터 변환



## > 데이터 시각화



1. R Studio에서 제공하는 다중회귀분석 함수인 pair 함수를 호출하여 최종 분석 및 시각화를 실행한다.

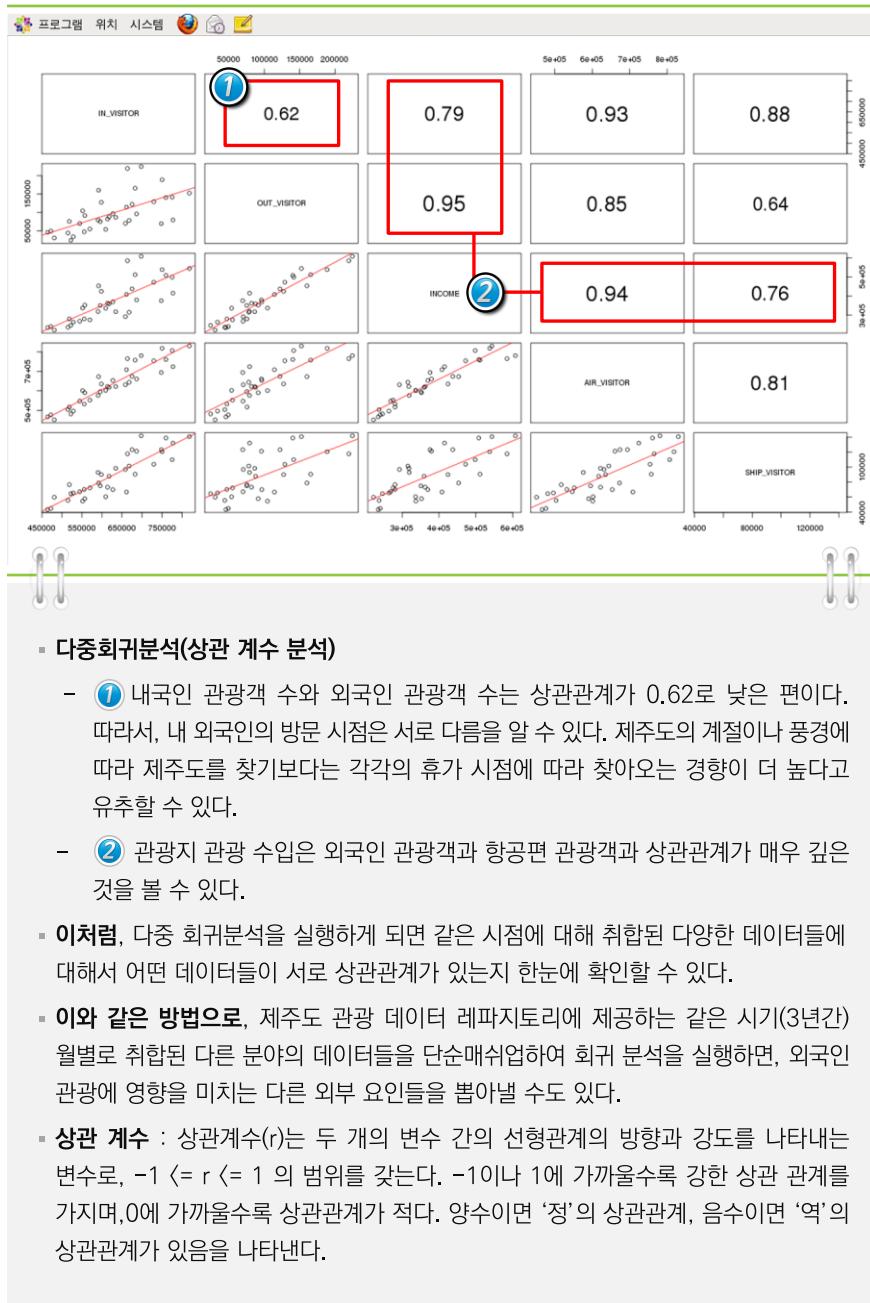
- upper.panel=panel.cor : 다중회귀분석 그래프의 우상단에 상관계수를 표시한다.
- lower.panel=panel.abline : 다중회귀분석 그래프의 좌하단에 선형회귀분석 그래프를 표시한다.

```

01. # 컬럼명 설정
02. names(inputData) <- c("IN_VISITOR", "OUT_VISITOR", "INCOME",
03. ↪ "AIR_VISITOR", "SHIP_VISITOR")
04. # 다중 선형 회귀 분석(상단은 상관계수, 하단은 선형회귀분석 표시)
05. pairs(inputData, upper.panel=panel.cor, lower.panel=panel.abline)

```

## ▶ 데이터 분석



### ■ 다중회귀분석(상관 계수 분석)

- ① 내국인 관광객 수와 외국인 관광객 수는 상관관계가 0.62로 낮은 편이다. 따라서, 내 외국인의 방문 시점은 서로 다른 것을 알 수 있다. 제주도의 계절이나 풍경에 따라 제주도를 찾기보다는 각각의 휴가 시점에 따라 찾아오는 경향이 더 높다고 유추할 수 있다.
- ② 관광지 관광 수입은 외국인 관광객과 항공편 관광객과 상관관계가 매우 깊은 것을 볼 수 있다.
- **이처럼**, 다중 회귀분석을 실행하게 되면 같은 시점에 대해 취합된 다양한 데이터들에 대해서 어떤 데이터들이 서로 상관관계가 있는지 한눈에 확인할 수 있다.
- **이와 같은 방법으로**, 제주도 관광 데이터 레파지토리에 제공하는 같은 시기(3년간) 월별로 취합된 다른 분야의 데이터들을 단순매쉬업하여 회귀 분석을 실행하면, 외국인 관광에 영향을 미치는 다른 외부 요인들을 뽑아낼 수도 있다.
- **상관 계수** : 상관계수( $r$ )는 두 개의 변수 간의 선형관계의 방향과 강도를 나타내는 변수로,  $-1 < r < 1$  의 범위를 갖는다. -1이나 1에 가까울수록 강한 상관 관계를 가지며, 0에 가까울수록 상관관계가 적다. 양수이면 '정'의 상관관계, 음수이면 '역'의 상관관계가 있음을 나타낸다.



## 용어정리

- **상관계수(Correlation coefficient)** : 상관계수( $r$ )는 두 개의 변수 간의 선형관계의 방향과 강도를 나타내는 변수로,  $-1 \leq r \leq 1$  의 범위를 갖는다.  
-1이나 1에 가까울수록 강한 상관 관계를 가지며, 0에 가까울수록 상관관계가 적다.  
양수이면 정의 상관관계, 음수이면 역의 상관관계가 있음을 나타낸다.
- **다중회귀분석(Multiple Regression Analysis)** : 두 변수 이상의 독립변수(영향변수, 원인변수)들이 종속변수(결과변수)에 어떠한 영향을 미치는지를 알기 위한 분석기법이다. 예를 들면, 골프 용품에 대한 전체만족도에 골프용품의 타구감, 방향성, 명성 등의 변수들이 어떠한 영향을 미치는지를 알기 위해서는 다중회귀분석을 하여야 한다. 따라서 독립변수들이 종속변수에 미치는 상대적 영향력을 알아볼 수 있으며, 이러한 독립변수 값들의 변화에 따라 종속변수 값이 어떻게 변화하는지를 예측할 수 있다. 독립변수와 종속변수에 쓰인 변수들의 측정척도는 등간척도(interval scale)나 비율척도(ratio scale)의 메트릭(metric) 자료이어야 하나, 독립변수가 명목척도일 경우에는 0과 1을 사용한 더미변수(dummy variables)를 이용하여 분석할 수 있다.

I.개요

II.수집

III.기공

IV.저장

V.분석

VI.시각화



## VII 예제문제

예제 문제1

99

예제 문제2

100

# 예 / 제 / 문 / 제

## 예제 1

3년간 월별로 취합된 내외국인 관광객 수 및 주요 관광지  
관광수입 데이터에 쉽게 구할 수 있는 월간 통계 데이터를  
매쉬업하여 다중 회귀 분석하라.

- 제주도 관광 데이터에 외부 데이터를 함께 다중 회귀분석하여 내국인 관광객과  
외국인 관광객 수에 영향을 미치는 외부 팩터를 한가지 이상 찾아낸다.

- 관광객 수에 영향이 있을 법한 월간 통계 데이터들을 수집한다.
- 연도별로 저장된 제주도 관광 데이터를 로드한다.
- 수집한 데이터들과 제주도 관광 데이터를 통합한다.
- R Studio의 pair 함수를 활용하여 다중 회귀분석 및 시각화 한 후 상관관계가  
0.9 이상인 데이터를 추출한다.

## 예제 2

2년간의 월별 교통수단별 관광객 수와 항공기나 선박 사고 관련 뉴스 데이터 발생 빈도를 매쉬업하여 사고와 교통수단 선택의 연관성을 분석하라.

- 2011년~2012년 2년간의 교통수단별 관광객수를 시각화 하고, 유입 패턴을 기초 분석하고, 같은 기간 항공기, 선박 사고 관련 뉴스 발생 빈도 데이터를 회귀 분석하여, 사고 발생에 따라 관광객이 교통수단을 선택하는 데에 어느 정도 영향을 끼치는지 분석하라.

- 연도별로 저장된 데이터를 로드한다.
- R Studio를 활용하여 2년간의 교통수단별 관광객 수 데이터를 통합 가공하고, 꺾은선 차트로 시각화한다.
- 소셜 분석 과정에서 제공되는 2011년~2012년 2년간의 뉴스 데이터에서 항공기 사고, 선박 사고 관련 뉴스의 월별 발생 빈도를 각각 추출한다.
- 교통수단별 관광객 수 데이터와 항공기 사고, 선박 사고 관련 뉴스 발생 빈도 데이터를 통합, 다중 회귀분석을 실행하여 시각화 한 후 상관관계가 어느 정도 있는지 확인한다.

## **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](http://www.nia.or.kr)

