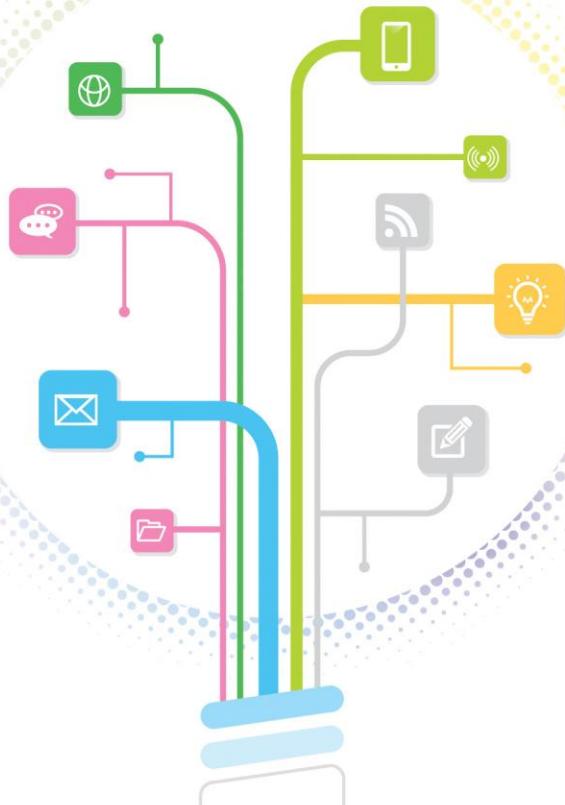


데이터 분석 콘텐츠 활용 매뉴얼



미래창조과학부



한국정보화진흥원



KBiG 빅데이터 전략센터

CONTENTS

Beginning Level 초급과정

I 개요

개요	9
----	---

II 수집

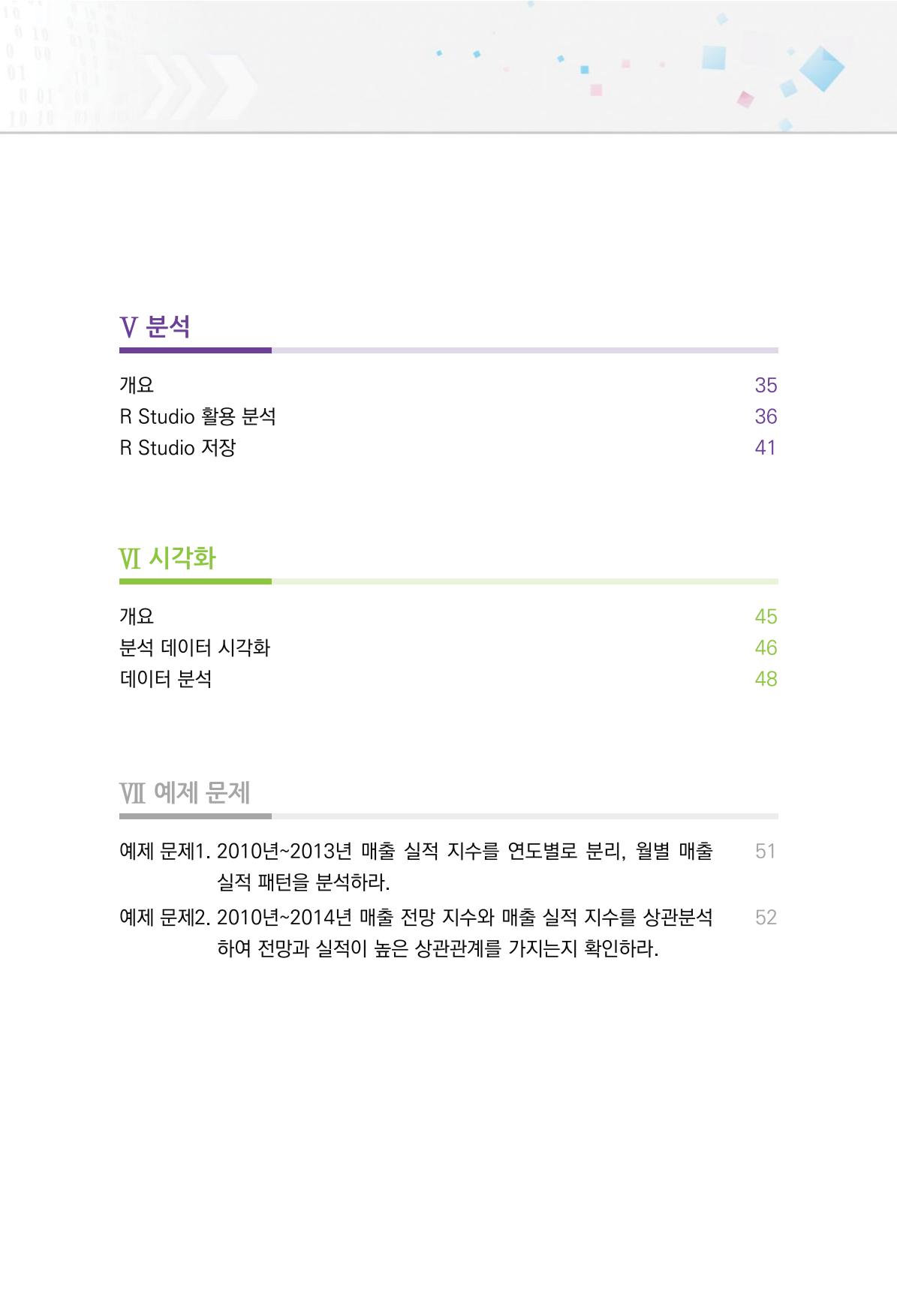
개요	13
수집 데이터	14
데이터 수집	15
데이터 작업 영역 이동 스크립트	18

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	41

VI 시각화

개요	45
분석 데이터 시각화	46
데이터 분석	48

VII 예제 문제

예제 문제1. 2010년~2013년 매출 실적 지수를 연도별로 분리, 월별 매출 실적 패턴을 분석하라.	51
예제 문제2. 2010년~2014년 매출 전망 지수와 매출 실적 지수를 상관분석 하여 전망과 실적이 높은 상관관계를 가지는지 확인하라.	52

CONTENTS

Intermediate Level **중급과정**

I 개요

개요	57
----	----

II 수집

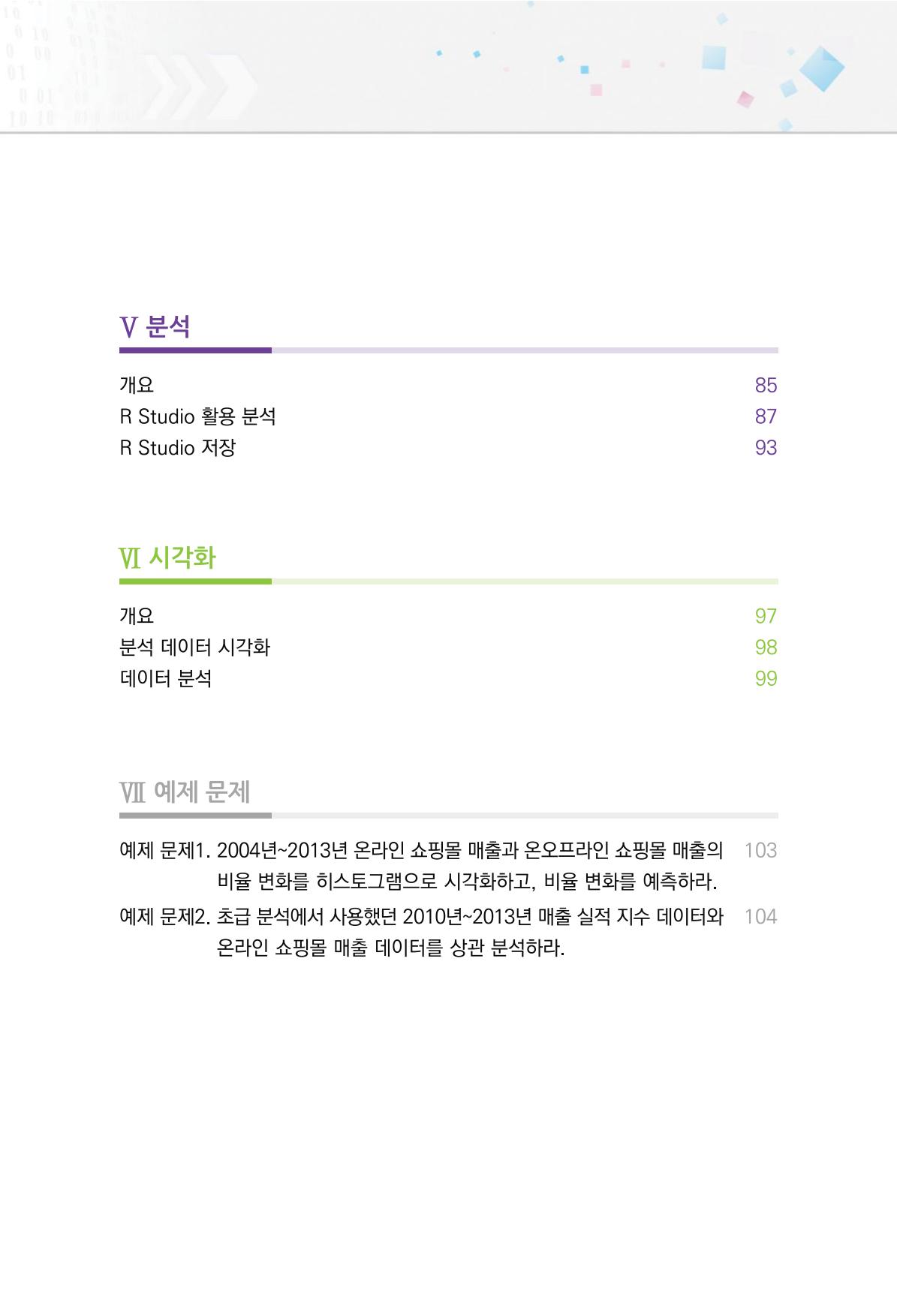
개요	61
수집 데이터	62
데이터 수집	66
데이터 작업 영역 이동 스크립트	69

III 가공

개요	73
데이터 가공 R 스크립트	77

IV 저장

개요	81
R Studio 활용 저장	82



V 분석

개요	85
R Studio 활용 분석	87
R Studio 저장	93

VI 시각화

개요	97
분석 데이터 시각화	98
데이터 분석	99

VII 예제 문제

예제 문제1. 2004년~2013년 온라인 쇼핑몰 매출과 온오프라인 쇼핑몰 매출의 비율 변화를 히스토그램으로 시각화하고, 비율 변화를 예측하라.	103
예제 문제2. 초급 분석에서 사용했던 2010년~2013년 매출 실적 지수 데이터와 온라인 쇼핑몰 매출 데이터를 상관 분석하라.	104



패션 

Beginning Level

초급과정







I 개요

개요

9

8

I

개요



개요

국가 통계 포털(<http://kosis.kr/>)로부터 추출한 2010년~2012년 시장 경기 동향 지수 통계 데이터를 바탕으로 연간 의류 매출 전망 지수와 실적 지수에 대해 관계비율분석을 실행하여 전망지수대비 실적지수 비율을 계산하고, 2013년 매출 전망 지수를 바탕으로 관계 비율 분석법을 응용한 단순 예측 방법을 사용하여 2013년의 매출실적지수를 예측한다. 이후, 2013년에 실제 발표된 2013년 매출실적지수와 비교하여 예측 방법을 검증한다. 이러한 방법으로 매출 전망지수가 발표되면, 실적지수를 예측하여 매출 목표 설정에 참고 자료로 활용한다.



> 활용 데이터

- **fashion_sales_2010_2012.csv :**

2010년~2012년 의류/패션 분야 연간 의류 매출 전망 지수 데이터

- **fashion_prospect_2010_2012.csv :**

2010년~2012년 의류/패션 분야 연간 의류 매출 실적 지수 데이터

- **fashion_sales_2013.csv :**

2013년 의류/패션 분야 연간 의류 매출 전망 지수 데이터

- **fashion_prospect_2013.csv :**

2013년 의류/패션 분야 연간 의류 매출 실적 지수 데이터

▶ 선행학습

- 리눅스 – 파일시스템 구조, 쉘 명령어, 쉘 스크립트 실행 방법
- R 프로그래밍 언어(문법, 패키지 추가 설치 방법)
- R 차트 – 설정 방법, 멀티 차트 표현 방법

▶ 요구사항

- 수집된 2010년~2012년 패션 부문 매출 전망 지수와 매출 실적 지수를 비교 분석하여, 전망과 실적의 비율을 계산하고, 이를 바탕으로 2013년의 매출 전망 지수로부터 매출 실적 지수를 예측하라.

▶ 분석 절차

- 수집된 2010년~2012년 매출 전망 지수와 실적 지수를 로드한다.
- 시계열 분석에 용이한 데이터 형태로 변화하기 위해 R Studio 의 “zoo”라이브러리 객체로 변환한다.
- 매출 전망 지수와 실적지수를 시계열 차트로 시각화하여 전망과 실적의 차이가 일정한 패턴을 보이는지 패턴 분석한다.
- 관계비율 분석법을 적용하여 3년간 월별 매출 전망 지수 대비 실적 지수 비율을 계산한다.
- 2013년 매출 전망 지수와 실적 지수를 로드한다.
- 매출 전망 지수가 매출 실적 지수에 선행하여 발표되는 점에 착안, 2013년 매출 전망 지수에 평균 매출 전망 대비 실적 지수 비율을 곱하여 2013년도 실적지수를 관계 비율 분석법을 응용한 단순 예측 기법으로 예측한다.
 - 수식 : 2013년 월별 매출실적지수 =
2013년 월별 매출전망지수 * 과거 3년간 월별 매출전망대비매출실적 비율
- 실제 2013년도 실적 지수와 단순 예측 방법으로 예측한 2013년도 실적 예측치를 그래프로 시각화하여, 관계비율 분석법을 응용한 단순 예측 방법의 적합성을 검증한다.



1

2

II 수집

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



수집

> 개요

패션 데이터는 소상공인진흥공단에서 국가통계포털(<http://kosis.kr/>)을 통해 발표하는 2010년~2013년 시장경기동향조사로부터 패션 부문의 데이터를 수집하여 분석에 용이하게 편집하여 제공한다.

> 수집 방법

- 데이터 제공 :** 패션/의류 쇼핑몰 매출 통계 데이터는 국가 통계 포털 (<http://kosis.kr/>)로부터 제공해 주는 데이터를 OpenAPI, 자료수집기(Crawler)를 통하여 데이터를 수집하였고, 실습용 자료는 빅데이터 분석 활용센터에 접속하여 패션 데이터 셋을 다운로드 할 수 있도록 원시데이터를 제공하고 있다.

The screenshot shows the KOSIS homepage with a search bar and navigation links. The main content area displays a chart titled '고용률(14.11)' with a value of '60.8%' and a small illustration of a person. Below the chart, there's a section for '주제별통계' (Topic-wise statistics) with various icons and links. To the right, there's a sidebar with a cartoon character and text about KOSIS services. At the bottom, there are buttons for '등록사각화 콘텐츠' (Registered Content), '공지사항' (Announcements), '보도자료' (Press Releases), '최근수록자료' (Recently Collected Materials), and other navigation links.

> 수집 데이터

> 매출 전망 지수 데이터(fashion_prospect_2010_2011.csv)

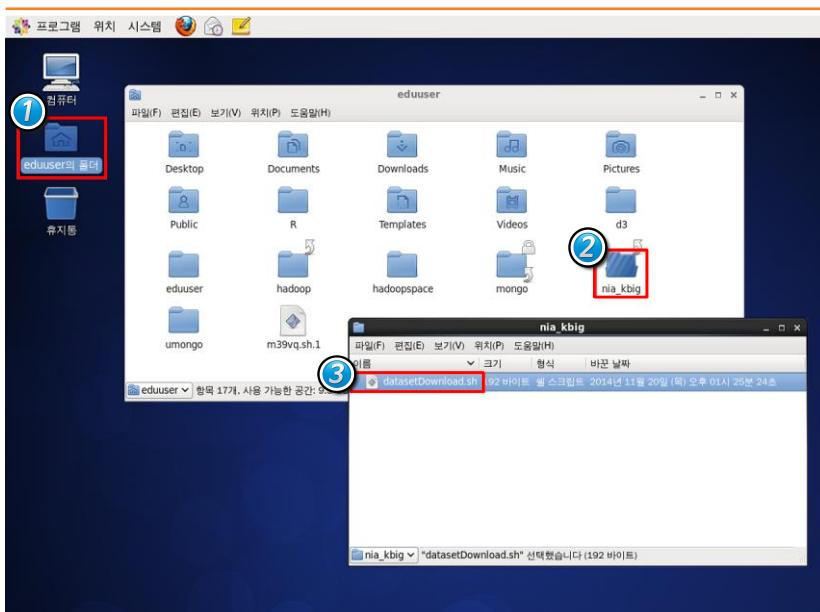
DATE	SALES_PROSPECT
2010-01-01	61.2
2010-02-01	65.6
2010-03-01	122.4
2010-04-01	126
2010-05-01	116
2010-06-01	87.5
2010-07-01	67.4
2010-08-01	59.2
2010-09-01	125.6
2010-10-01	132.4
2010-11-01	118.2
2010-12-01	99.1
2011-01-01	65.7
2011-02-01	67.1
2011-03-01	118.6
2011-04-01	114.3
2011-05-01	105.1
2011-06-01	90.2
2011-07-01	76.3
2011-08-01	54.2
2011-09-01	103.4
2011-10-01	112.9
2011-11-01	101.1
2011-12-01	87.6
2012-01-01	72.6
2012-02-01	46
2012-03-01	113.8
2012-04-01	102.6
2012-05-01	106.8
2012-06-01	82.3
2012-07-01	72.4
2012-08-01	42.3
2012-09-01	99.2
2012-10-01	114.1
2012-11-01	100.1
2012-12-01	88.1

- 매출 실적 데이터도 같은 형태로 제공하고 있다.

▶ 데이터 수집(datasetDownload.sh)

- 데이터 저장소에서 서버 로컬로 패션 데이터 셋을 복사해 온다.
 - **fashion_sales_2010_2012.csv** :
2010년~2012년 의류/패션 분야 연간 의류 매출 전망 지수 데이터
 - **fashion_prospect_2010_2012.csv** :
2010년~2012년 의류/패션 분야 연간 의류 매출 실적 지수 데이터
 - **fashion_sales_2013.csv** :
2013년 의류/패션 분야 연간 의류 매출 전망 지수 데이터
 - **fashion_prospect_2013.csv** :
2013년 의류/패션 분야 연간 의류 매출 실적 지수 데이터

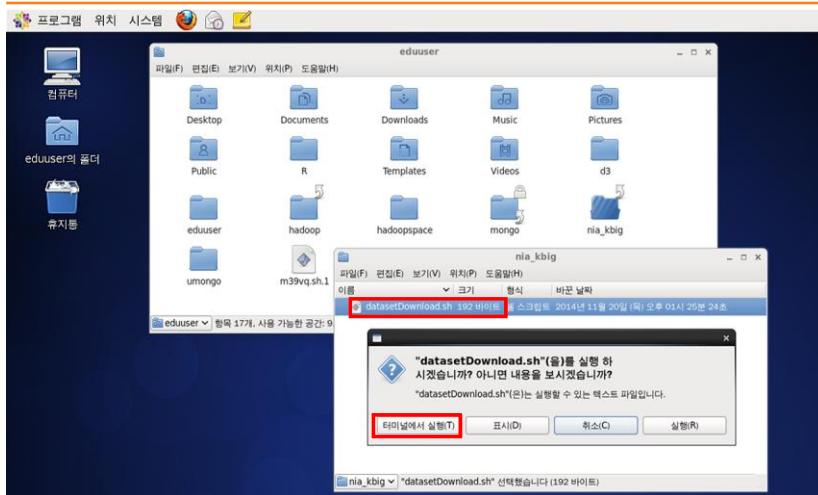
▶ 실습코드 디렉토리로 이동



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

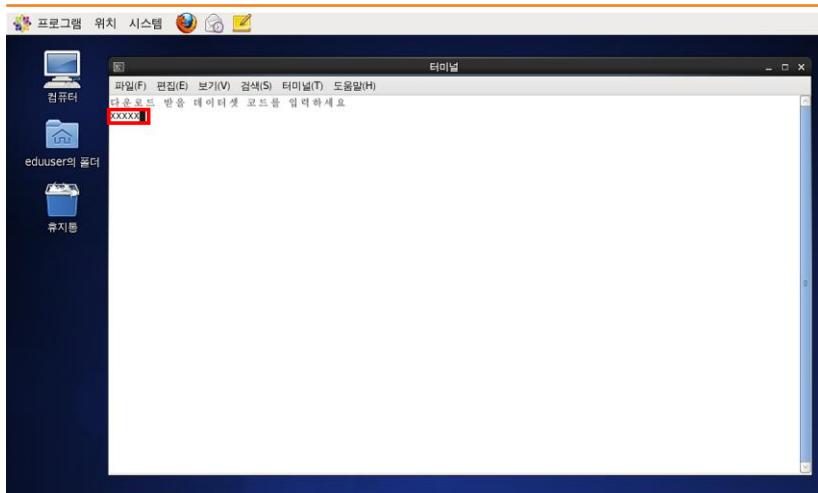
▶ 레파지토리에서 데이터 수집

datasetDownload.sh (원시데이터를 로컬서버로 복사하는 스크립트)



- '터미널에서 실행' 버튼을 클릭한다.

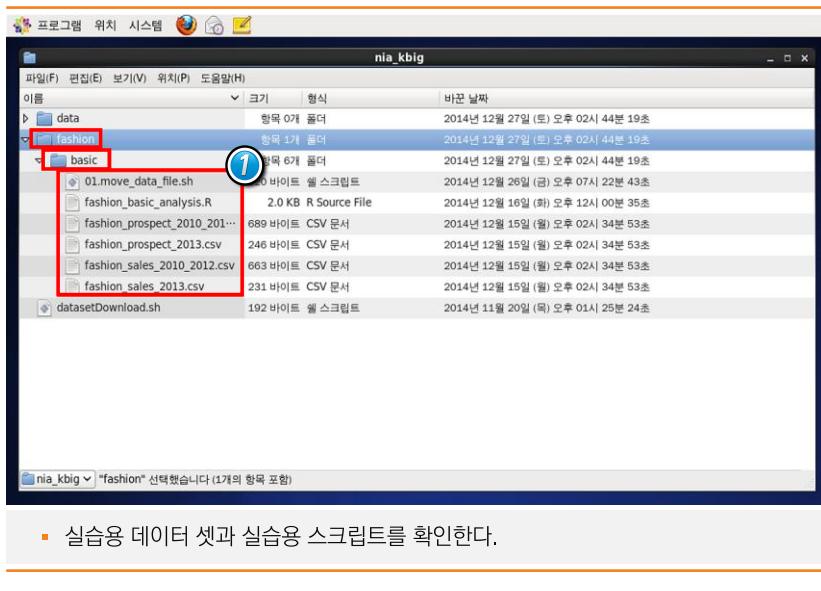
▶ 데이터셋 코드 입력



- 다운로드 받을 데이터셋 코드를 입력 후 엔터

II. 수집

▶ 데이터셋과 실습용 쉘 스크립트



▶ ① 데이터 및 스크립트

▪ 01.move_data_file.sh :

로컬로 수집해온 데이터를 작업 영역 Data 폴더로 자료를 이동하는 스크립트

▪ fashion_basic_analysis.R :

패션 데이터 분석용 R 스크립트.

▪ fashion_sales_2010_2012.csv :

2010년~2012년 의류/패션 분야 연간 의류 매출 전망 지수 데이터

▪ fashion_prospect_2010_2012.csv :

2010년~2012년 의류/패션 분야 연간 의류 매출 실적 지수 데이터

▪ fashion_sales_2013.csv :

2013년 의류/패션 분야 연간 의류 매출 전망 지수 데이터

▪ fashion_prospect_2013.csv :

2013년 의류/패션 분야 연간 의류 매출 실적 지수 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. # 패션 데이터 파일 정의
03. TARGET_FASHION=/home/eduuser/nia_kbig/fashion/basic/*.csv
04. # 작업 디렉토리 정의
05. LOCAL_DIR=/home/eduuser/nia_kbig/data/
06. mv $TARGET_FASHION $LOCAL_DIR
07.

```



- 데이터 작업 영역 이동 스크립트 소스(01.move_data_file.sh)
- 라인 03 : 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 시계열 분석을 위한 수집된 데이터 셋을 작업 영역 Data 폴더로 자료를 이동

- “./01.move data file.sh”를 입력하여 준비된 패션 데이터를 이동시키다.

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 패션 데이터의 가공은, 전처리 단계에서 수집된 4년간(2010년~2013년)의 의류 분야 매출 전망, 실적 지수 데이터를 R Studio에서 로드하여 시계열 분석에 유용한 zoo 라이브러리 객체 형태로 변환하도록 한다.

> 가공 방법

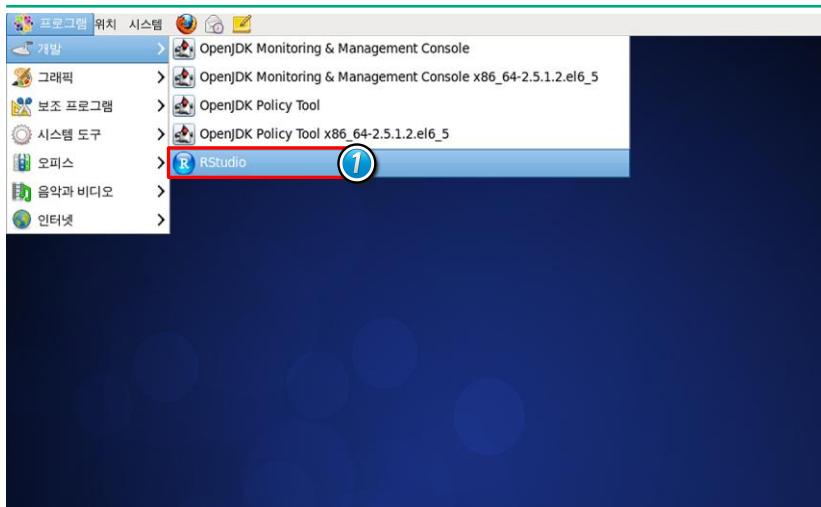
- **분석 도구 실행** : 가공 분석을 위해, 프로그래밍 도구인 R Studio를 실행한다.
- **데이터 로드** : 2010년~2014년 의류 분야 매출 전망 데이터, 매출 실적 데이터를 R Studio에서 각각 읽어들인다.
- **데이터 변환** : R Studio에는 시계열 분석을 위한 여러 가지 라이브러리가 존재한다. 이 중 가장 일반적으로 사용하는 “zoo”라이브러리를 활용하여 데이터를 “zoo” 객체로 변환한다.



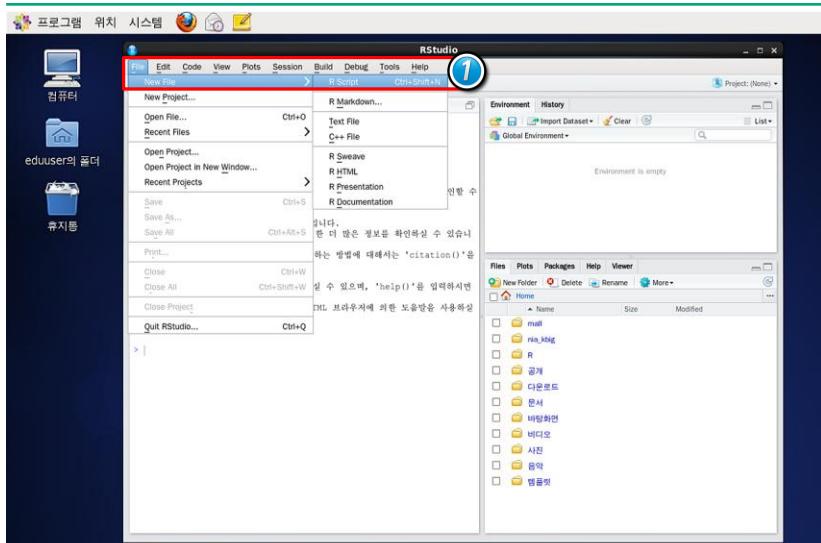
용어 정리

- **zoo 라이브러리** : R을 활용한 시계열 분석에 매우 보편적으로 사용되는 패키지 라이브러리이다. 내부적으로 Index/Date/Time 을 키로 가지는 여러 항목의 시계열 데이터를 처리하기 위한 매트릭스 형태의 자료구조를 지니고 있으며, 시계열 항목(컬럼)간의 연산에 관련된 유용한 함수들을 내포하고 있다.

▶ 데이터 가공

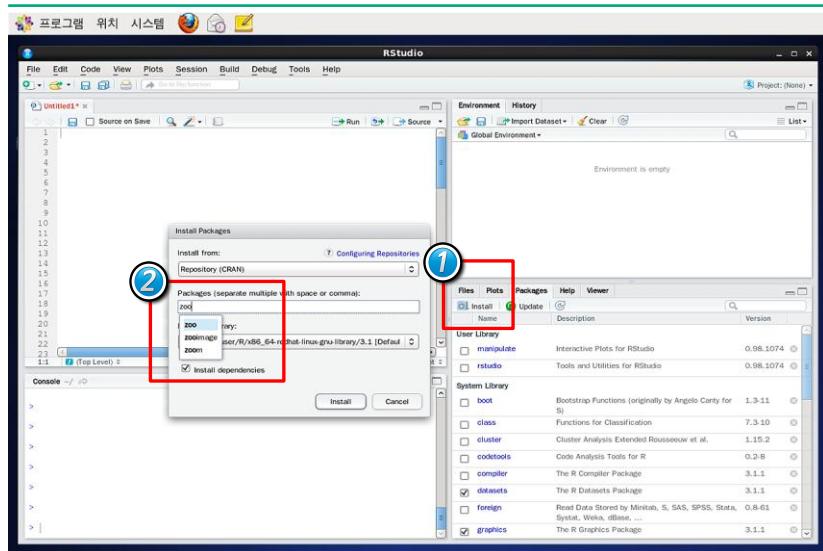


- ① 왼쪽 상단의 [“프로그램” 클릭] > [“개발” 클릭] > [“RStudio” 클릭]으로 분석 도구인 R Studio를 실행한다.



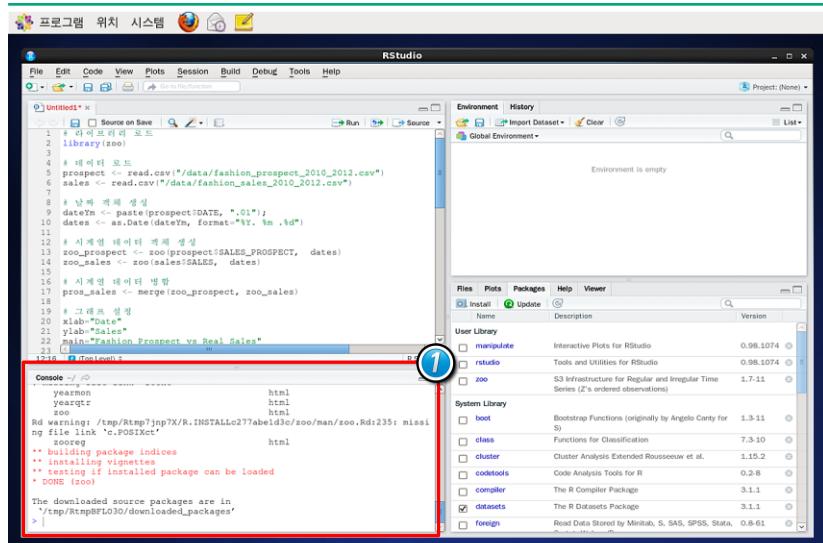
- ② 패션 데이터의 분석 및 가공을 위해 프로그램 작업 파일 (“New File” 클릭) “R_Script” 클릭)을 선택한다.

III. 가공



3. R에서 시계열 분석을 위해 일반적으로 많이 사용되는 “zoo” 라이브러리를 설치한다.

① 패키지 탭에 install 버튼을 누른 후 ② zoo 를 입력하여 라이브러리를 선택한다.



4. ① “zoo” 라이브러리가 설치되는 모습을 확인할 수 있다.

```

1 # 라이브러리 로드
2 library(zoo)
3
4 # 2010년 ~ 2012년 데이터 로드
5 prospect <- read.csv("home/edususer/nia_kbgi/data/fashion_prospect_2010")
6 sales <- read.csv("home/edususer/nia_kbgi/data/fashion_sales_2010_2012")
7
8 # 날짜 객체 생성
9 dateYM <- paste(prospect$DATE, ".01");
10 dates <- as.Date(dateYM, format = "%Y. %m .%d")
11
12 # 시계열 데이터 객체 생성
13 zoo_prospect <- zoo(prospect$SALES_PROSPECT, dates)
14 zoo_sales <- zoo(sales$SALES, dates)
15
16 # 시계열 데이터 병합
17 pros_sales <- merge(zoo_prospect, zoo_sales)
18
19 # 2013년에 대한 시계열 대# 2013년 매출 전망 및 실적 데이터 로드
20 prospect_2013 <- read.csv("data/fashion_prospect_2013.csv")
21 sales_2013 <- read.csv("data/fashion_sales_2013.csv")
22
23 dateYM_2013 <- paste(sales_2013$DATE, ".01");
24 dates_2013 <- as.Date(dateYM_2013, format = "%Y. %m .%d")
25
26 zoo_prospect_2013 <- zoo(prospect_2013$SALES, dates)
27 zoo_sales_2013 <- zoo(sales_2013$SALES, dates)
28
29
30
31
32
33
34
35
36
37
  
```

The screenshot shows the RStudio interface with the code editor containing R script for data loading and manipulation. The environment pane shows variables like `prospect` and `sales` with their respective data frames. The global environment pane lists packages like `zoo` and `rstatistic`.

5. 2010년~2012년 매출 전망 데이터와 매출 실적 데이터를 각각 로드하고, 시계열 분석이 가능한 “zoo” 객체로 변환하도록 R 스크립트를 작성한다.

```

1 # 라이브러리 로드
2 library(zoo)
3
4 # 2010년 ~ 2012년 데이터 로드
5 prospect <- read.csv("home/edususer/nia_kbgi/data/fashion_prospect_2010")
6 sales <- read.csv("home/edususer/nia_kbgi/data/fashion_sales_2010_2012")
7
8 # 날짜 객체 생성
9 dateYM <- paste(prospect$DATE, ".01");
10 dates <- as.Date(dateYM, format = "%Y. %m .%d")
11
12 # 시계열 데이터 객체 생성
13 zoo_prospect <- zoo(prospect$SALES_PROSPECT, dates)
14 zoo_sales <- zoo(sales$SALES, dates)
15
16 # 시계열 데이터 병합
17 pros_sales <- merge(zoo_prospect, zoo_sales)
18
19 # 2013년에 대한 시계열 대# 2013년 매출 전망 및 실적 데이터 로드
20 prospect_2013 <- read.csv("home/edususer/nia_kbgi/data/fashion_prospect_2013")
21 sales_2013 <- read.csv("home/edususer/nia_kbgi/data/fashion_sales_2013")
22
23 dateYM_2013 <- paste(sales_2013$DATE, ".01");
24 dates_2013 <- as.Date(dateYM_2013, format = "%Y. %m .%d")
25
26 zoo_prospect_2013 <- zoo(prospect_2013$SALES, dates)
27 zoo_sales_2013 <- zoo(sales_2013$SALES, dates)
28
29
30
31
32
33
34
35
36
37
  
```

This screenshot is similar to the previous one but includes numbered circles: circle 1 points to the first few lines of the R code, and circle 2 points to the last few lines where the 2013 data is being processed.

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

III. 가공

▶ 데이터 가공 R 스크립트

```
01. # 라이브러리 로드
02. library(zoo)
03.
04. # 2010년 ~ 2012년 데이터 로드
05. prospect <- read.csv("/home/eduuser/nia_kbig/data/fashion_prospect_2010_
    ↴ 2012.csv")
06. sales <- read.csv("/home/eduuser/nia_kbig/data/fashion_sales_2010_2012.csv")
07.
08. # 날짜 객체 생성
09. dateYm <- paste(prospect$DATE, ".01");
10. dates <- as.Date(dateYm, format="%Y. %m .%d")
11.
12. # 시계열 데이터 객체 생성
13. zoo_prospect <- zoo(prospect$SALES_PROSPECT, dates)
14. zoo_sales <- zoo(sales$SALES, dates)
15.
16. # 시계열 데이터 병합
17. pros_sales <- merge(zoo_prospect, zoo_sales)
18.
19. # 2013년에 대한 시계열 데이터# 2013년 매출 전망 및 실적 데이터 로드
20. prospect_2013 <-
    ↴ read.csv("/home/eduuser/nia_kbig/data/fashion_prospect_2013.csv")
21. sales_2013 <- read.csv("/home/eduuser/nia_kbig/data/fashion_sales_2013.csv")
22. # 2013년 시계열 데이터 생성
23. dateYm_2013 <- paste(sales_2013$DATE, ".01");
24. dates_2013 <- as.Date(dateYm_2013, format="%Y. %m .%d")
25.
26. zoo_prospect_2013 <- zoo(prospect_2013$SALES, dates)
27. zoo_sales_2013 <- zoo(sales_2013$SALES, dates)
28.
```



- 데이터 가공 R 스크립트
- 라인 05~06 : 2010~2012년 매출 전망 지수 데이터와 매출 실적 데이터파일을 읽어들여 R 데이터 객체(prospect, sales)로 저장하는 라인이다.
- 라인 09~10 : zoo 라이브러리 객체를 만들기 위해 날짜 객체를 생성하는 라인이다.
- 라인 13~17 : zoo 함수를 활용하여 5,6 라인에서 만든 R 데이터 객체와 9, 10 라인에서 생성한 날짜 객체로부터 zoo 라이브러리 객체(zoo_prospect, zoo_sales)를 생성한 후, merge 함수를 활용하여 통합 객체(pros_sales)를 생성하는 라인이다.
- 라인 20~21 : 2013년 매출 전망 지수 데이터와 매출 실적 데이터파일을 읽어들여 R 데이터 객체 (prospect , sales)로 저장하는 라인이다.
- 라인 23~24 : zoo 라이브러리 객체를 만들기 위해 날짜 객체를 생성하는 라인이다.
- 라인 26~27 : zoo 함수를 활용하여 5,6 라인에서 만든 R 데이터 객체와 9, 10 라인에서 생성한 날짜 객체로부터 zoo 라이브러리 객체(zoo_prospect_2013, zoo_sales_2013)를 생성하는 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

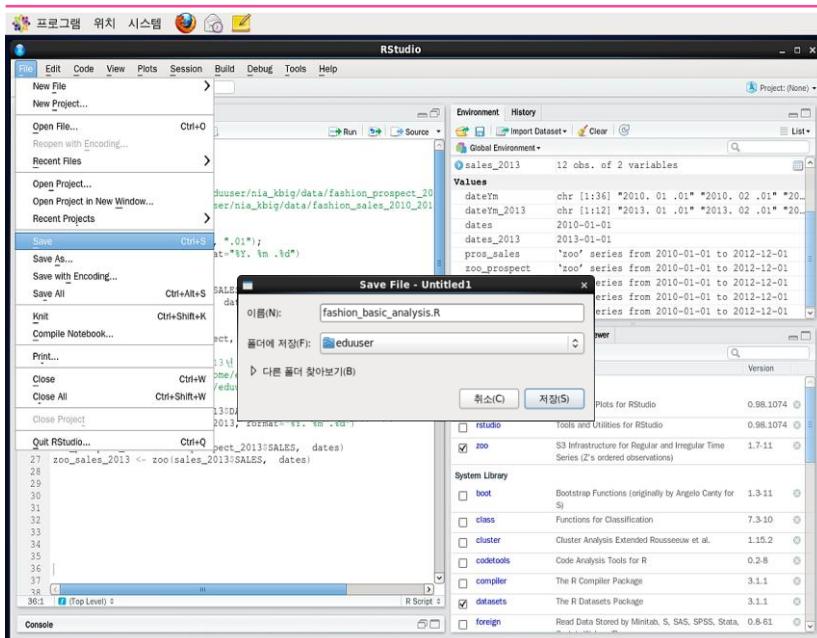
R Studio를 활용하여 데이터 로드 > 가공 > 분석 > 시각화 단계를 한번에 실행하므로, 별도의 저장 과정은 생략한다. 가공된 데이터는 메모리상에 존재하며, 지금까지 작성한 분석 프로그램 소스를 저장한다.

> 저장 방법

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

R Studio 활용 저장

데이터 저장



1. 패션 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/fashion_basic_analysis.r”로 저장한다.

W





V 분석

개요	35
R Studio 활용 분석	36
R Studio 저장	41

V

분석

> 개요

패션 데이터의 분석은 R Studio에 내장된 zoo 라이브러리 객체 및 그래프 기능을 활용하여 시계열 분석을 통한 패턴 분석을 한다. 2010년~2012년 매출 전망 지수와 실적 지수 데이터를 비교 분석하여 전망 대비 실적 비율을 계산하고, 2013년 매출 전망 지수를 기반으로 관계 비율 분석법을 응용한 단순 예측 기법으로 2013년 매출 실적을 예측한다.

> 데이터 분석 방법

- 가공 단계에서 zoo 라이브러리 객체로 변환한 2010년 ~ 2012년 매출 전망 지수와 실적 지수 데이터를 시계열 분석을 통한 패턴 분석을 하기위해 시각화 하여 데이터의 패턴을 파악한다.
- 파악한 패턴을 통해 2013년 매출 실적 유추 방법으로 관계 비율 분석법을 응용한 단순 예측 기법을 선택한다.
- 가공 단계에서 변환한 2013년 매출 전망 데이터를 사용하여 관계 비율 분석 법을 응용한 단순 예측 기법으로 2013년 매출 실적을 예측 분석을 한다.
- **관계 비율 분석법을 응용한 단순 예측 기법 수식 :**
2013년 월별 매출실적지수 = 2013년 월별 매출전망지수
* 과거 3년간 월별 매출 전망 대비 매출 실적 비율
- 유추한 2013년 매출 실적과 실제 2013년 매출 실적 지수를 비교/검증한다.

> R Studio 활용 분석

> 데이터 불러오기

zoo_prospect (가공 단계에서 생성한 3년 간의 매출 전망 지수)

```

16 # 시계열 데이터 병합
17 pros_sales <- merge(zoo_prospect, zoo_sales)
18
19 # 2013년에 대한 시계열 대비 2013년 매출 전망 및 실제 데이터로드
20 sales_2013 <- read.csv("/home/edouser/nia_kbgi/data/fashion_sales_2013.csv")
21 sales_2013 <- read.csv("/home/edouser/nia_kbgi/data/fashion_sales_2013.csv")
22 # 2013년 시계열 데이터 생성
23 dateYm_2013 <- as.Date(sales_2013$DATE, "%Y-%m-%d")
24 dates_2013 <- as.Date(dateYm_2013, format="%Y-%m-%d")
25
26 zoo_prospect_2013 <- zoo(prospect_2013$SALES, dates)
27 zoo_sales_2013 <- zoo(sales_2013$SALES, dates)
28
29
30 #zoo_prospect
31
32
33
34
35
36
37
32.1 [Top Level] >

```

Console

```

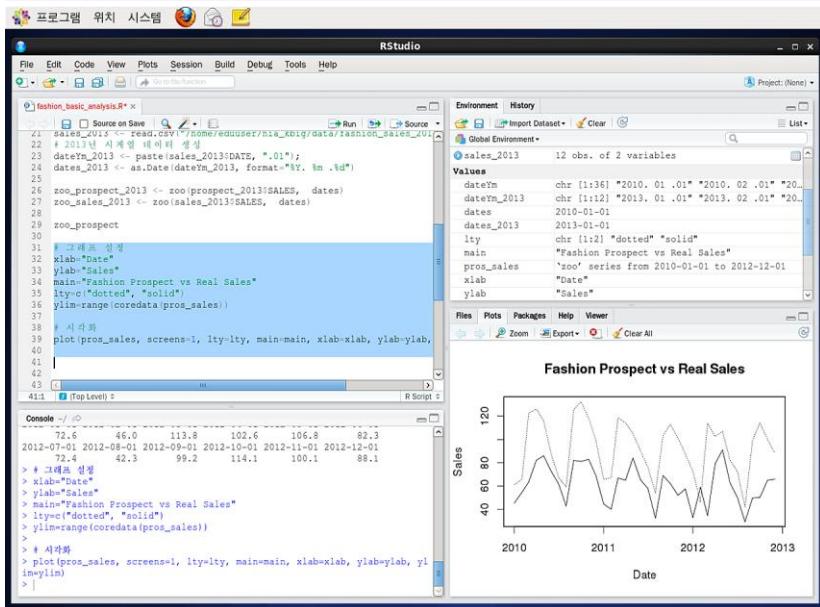
> zoo_prospect
> zoo_prospect
2010-01-01 2010-02-01 2010-03-01 2010-04-01 2010-05-01 2010-06-01
61.2      65.6     122.4    126.0    116.0     87.5
2010-07-01 2010-08-01 2010-09-01 2010-10-01 2010-11-01 2010-12-01
67.4      59.2     125.6    132.4    118.2     99.1
2011-01-01 2011-02-01 2011-03-01 2011-04-01 2011-05-01 2011-06-01
65.7      67.1     118.6    114.3    105.1     90.2
2011-07-01 2011-08-01 2011-09-01 2011-10-01 2011-11-01 2011-12-01
76.3      54.2     103.4    112.9    101.1     87.6
2012-01-01 2012-02-01 2012-03-01 2012-04-01 2012-05-01 2012-06-01
72.6      46.0     113.8    102.6    106.8     82.3
2012-07-01 2012-08-01 2012-09-01 2012-10-01 2012-11-01 2012-12-01
72.4      42.3     99.2     114.1    100.1     88.1
> |

```

- ① 가공한 데이터가 잘 들어가 있는지 확인하기 위해 “zoo_prospect”를 입력하고 위와 같이 블럭을 선택한 후, Ctrl+Enter를 입력하면, ②와 같이 데이터를 확인할 수 있다. 다른 변수도 마찬가지 방법으로 확인할 수 있다.

➤ 데이터 분석

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



- 매출 전망과 실적이 어떠한 패턴으로 나타나는지 파악하기 위해 가공 단계에서 통합한 2010년~2012년 매출 전망 및 실적 데이터를 plot 함수를 활용하여 시각화해 본다.
- #주) 그래프 모양을 보면, 전망과 실적이 거의 같은 패턴을 보이며, 비교적 일정한 비율로 차이가 남을 알 수 있다. 따라서, [매출실적예측 = 매출 전망 * (전망 대비 실적 비율)]로 가정하여 2013년 매출 실적을 예측해 보기로 한다.

```

1. # 그래프 설정
2. xlab="Date"
3. ylab="Sales"
4. main="Fashion Prospect vs Real Sales"
5. lty=c("dotted", "solid")
6. ylim=range(coredata(pros_sales))
7.
8. # 2010~2012 전망지수 대비 실적지수 시각화
9. plot(pros_sales, screens=1, lty=lty, main=main, xlab=xlab, ylab=ylab, ylim=ylim)

```



- 패션 분석 및 시각화 R 스크립트 소스(fashion_basic_analysis.R)
- 라인 01~06 : 시각화를 하기 위해 그래프 설정값들을 지정하는 라인이다.
- 라인 09 : plot 함수를 사용하여 2010~2012 전망지수 대비 실적지수를 시각화하는 라인이다.

I. 개요

II. 수집

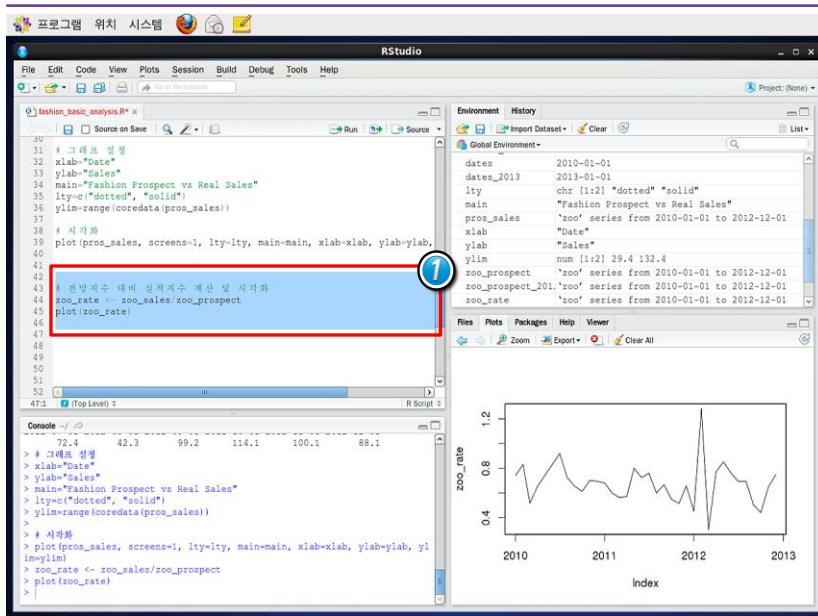
III. 가공

IV. 저장

V. 분석

VI. 시각화

V. 분석

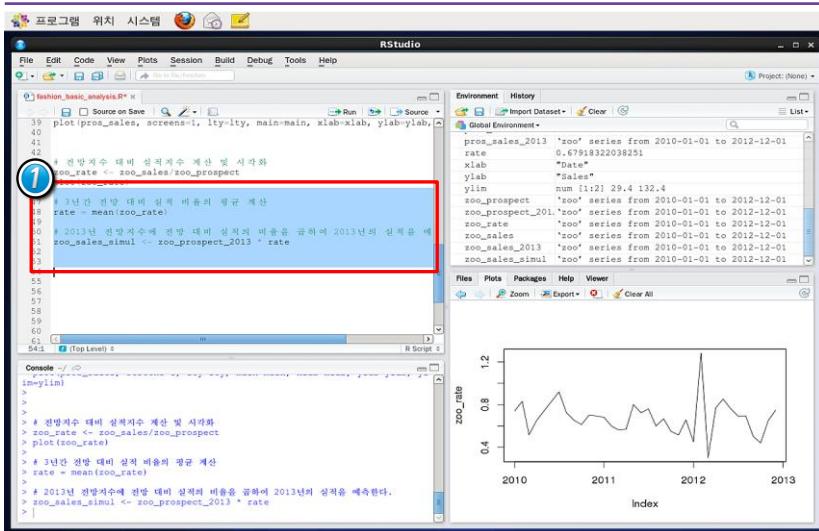


- ① 매출 전망 대비 실적의 비율 차이에도 월별로 패턴이 존재하는지 확인하기 위하여 계산 후에 그래프로 시각화해 본다.
- #주) 시각화 결과, 전망의 오차에 대해서는 12개월 단위나 분기 단위의 특별한 패턴은 보이지 않는다. 따라서, 월별로 패턴을 적용하기 보다 평균 비율을 활용하여 예측하기로 한다.

```
1. # 전망지수 대비 실적지수 계산 및 시각화
2. zoo_rate <- zoo_sales/zoo_prospect
3. plot(zoo_rate)
```



- 패션 분석 및 시각화 R 스크립트 소스(fashion_basic_analysis.R)
- 라인 02 : 월별 전망지수 대비 실적 지수 비율을 계산하여 객체(zoo_rate)로 저장하는 라인이다.
- 라인 03 : plot 함수를 사용하여 월별 전망지수 대비 실적 지수 비율을 그래프로 시각화하는 라인이다.



3. ① 3년간의 매출 전망 대비 대비 실적 데이터의 평균값을 구하고, 2013년 매출 전망 데이터에 계산된 평균값을 곱하여 2013년 매출 실적을 예측하여 객체에 저장한다.

```

1. # 3년간 전망 대비 실적 비율의 평균 계산
2. rate = mean(zoo_rate)
3. # 2013년 판매 전망 및 실적 데이터 로드
4. prospect_2013 <- read.csv("/data/fashion_prospect_2013.csv")
5. sales_2013 <- read.csv("/data/fashion_sales_2013.csv")
6. # 2013년에 대한 시계열 데이터 생성
7. dateYm_2013 <- paste(sales_2013$DATE, ".01");
8. dates_2013 <- as.Date(dateYm_2013, format="%Y. %m .%d")
9. zoo_prospect_2013 <- zoo(prospect_2013$SALES, dates)
10. zoo_sales_2013 <- zoo(sales_2013$SALES, dates)
11. # 2013년 전망지수에 전망 대비 실적의 비율을 곱하여 2013년의 실적을 예측한다.
12. zoo_sales_simul <- zoo_prospect_2013 * rate

```



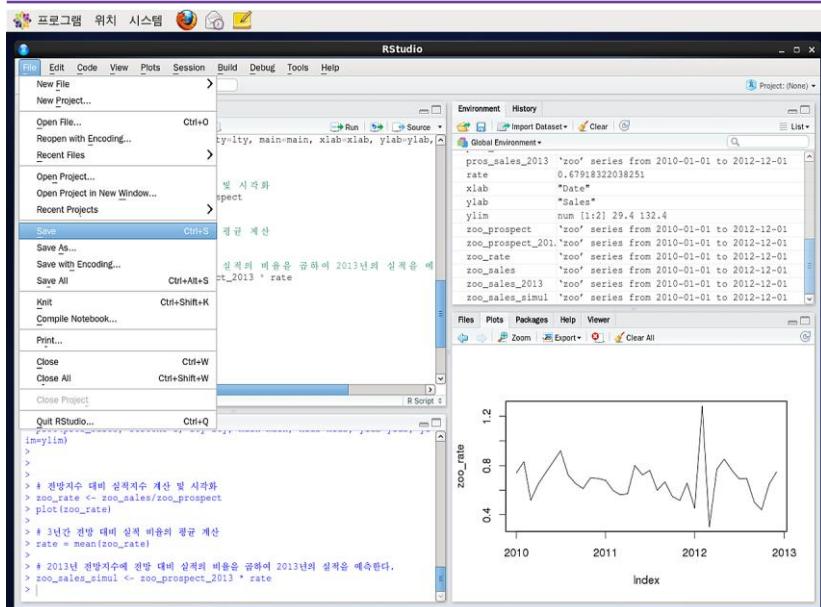
• 패션 분석 및 시각화 R 스크립트 소스(fashion_basic_analysis.R)

- 라인 02 : 3년간 월별 매출 전망 대비 실적 비율의 평균값을 계산하여 변수(rate)에 저장하는 라인이다.
- 라인 4~10 : 2013년 판매 전망 및 실적 데이터를 로드한 후, 각각 zoo 라이브러리 시계열 객체(zoo_prospect_2013, zoo_sales_2013)로 저장하는 라인이다.
- 라인 12 : 2013년 전망지수에 라인2에서 계산한 '평균 매출전망 대비 실적 비율'을 곱하여 2013년의 실적을 예측하여 객체(zoo_sales_simul)로 저장하는 라인이다.

➤ R Studio 저장

➤ 분석 결과 저장

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



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

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화



1

2

VI 시각화

개요	45
분석 데이터 시각화	46
데이터 분석	48

VI

시각화

> 개요

패션 데이터의 분석 과정에서는 시계열 분석 기법을 통해 그래프 시각화를 하여 데이터의 흐름을 분석하고, 분석 결과에 따라서 이후 필요한 분석 방법을 선택해 가는 과정을 실행했다. 따라서 패션 데이터 시각화 과정에서는 최종 결과물인 2013년 매출 실적 지수 예측 데이터와 2013년 실제 매출 실적 지수 데이터에 대한 단순 예측 방법의 적합성을 검증을 해본다.

> 시각화 방법

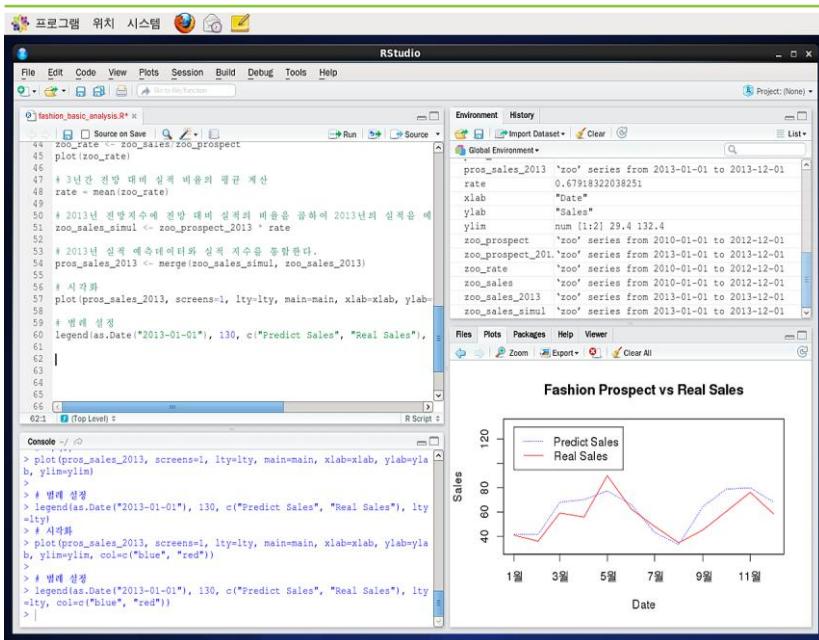
- 분석 과정에서 분석한 2013년 매출 실적 지수에 대한 예상 데이터와, 가공 과정에서 생성한 2013년 실제 매출 실적 지수 데이터를 하나의 시계열 객체로 통합한다.
- 통합한 시계열 데이터를 시각화하여, 2013년 매출실적에 대한 예측치의 정합성을 판단한다.

> 시각화 과정



▶ 분석 데이터 시각화

▶ 데이터 시각화



- 2013년 매출 전망 지수에 분석 과정에서 계산한 2013년 매출 실적 지수 예측 데이터와 가공 과정에서 생성한 실제 2013년 매출 실적 지수를 하나의 객체로 통합한 후 꺠은선 차트로 시각화한다.

```

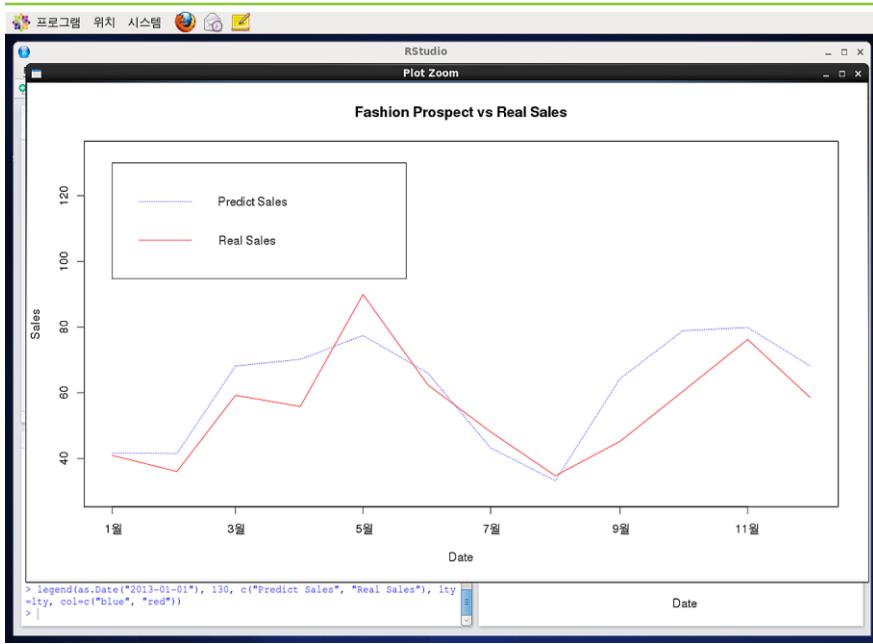
01. # 2013년 실적 예측데이터와 실적 지수를 통합한다.
02. pros_sales_2013 <- merge(zoo_sales_simul, zoo_sales_2013)
03.
04. # 시각화
05. plot(pros_sales_2013, screens=1, lty=lty, main=main, xlab=xlab, ylab=ylab,
06.       ↪ ylim=ylim, col=c("blue", "red"))
07.
08. # 범례 설정
    legend(as.Date("2013-01-01"), 130, c("Predict-Sales", "Real Sales"), lty=lty,
    ↪ col=c("blue", "red"))
  
```

VI. 시각화



- 46페이지 패션 분석 및 시각화 R 스크립트 소스(fashion_basic_analysis.R)
- 라인 02 : 2013년 매출 실적에 대한 예측 데이터와, 실제 수집된 2013년 매출 실적 지수 데이터를 통합하여 하나의 객체(pros_sales_2013)로 저장하는 라인이다.
- 라인 05 : plot 함수를 사용하여 2013년 매출 실적에 대한 예측 데이터와, 실제 수집된 2013년 매출 실적 지수 데이터를 그래프로 시각화하는 라인이다.
- 라인 08 : legend함수를 사용하여 시각화한 그래프에 범례를 설정하고 표시하는 라인이다.

▶ 데이터 분석



- 최종 결과 그래프에는, 분석 과정에서 계산하여 예측한 2013년 실적 지수가 블루색 점선으로, 실제 2013년 실적 지수가 빨간색 실선으로 표시되어 있다.
- 분석 과정에서 전망치와 실적치가 거의 비슷한 패턴으로 움직이되, 연월에 상관없이 항상 실적치가 전망치에 비해 일정 비율 정도 낮은 패턴을 보이는 점에 착안, 관계 비율분석법을 응용하여 2013년 매출 전망 지수를 토대로 예측한 매출 예상 그래프가 실제 매출 실적 그래프와 거의 비슷한 레벨에서 움직이는 것을 볼 수 있다.
- 이를 통해, 매출 전망 지수는 매출 실적의 변화를 예상하기 위한 지수로 사용하기에는 부족함이 있지만, 일정한 비율로 높게 나타나는 경향이 강하므로, 이를 감안하여 전망치와 실적치 사이에 관계비율분석을 응용한 단순 예측기법으로 경기를 좀 더 정확하게 판단하여 매출 목표치 설정에 활용한다.
- 이러한 방법으로 매출 전망지수가 발표되면, 실적지수를 예측하여 매출 목표 설정에 참고 자료로 활용한다.



VII 예제문제

예제 문제1

51

예제 문제2

52

예 / 제 / 문 / 제

예제 1

2010년~2013년 매출 실적 지수를 연도별로 분리,
월별 매출 실적 패턴을 분석하라.

- 4년간의 월별 매출 실적 지수를 분석하여, 매출 실적 지수를 하나의 그래프로 시각화하고 매출 지수의 월별 변화를 분석하라.

- 2010년~2012년 매출 실적 지수를 로드한다. 2013년 매출 실적 지수를 로드한다.
R Studio를 사용하여 4년간의 정보를 하나로 통합하고, 4년간 1월~12월 매출 실적 지수를 꺾은선 그래프로 표현한다.
- 매출실적지수가 높은 계절과 낮은 계절을 구분하고, 그 원인에 대해 추론해 본다.

예제 2

2010년~2014년 매출 전망 지수와 매출 실적 지수를 상관 분석하여 전망과 실적이 높은 상관관계를 가지는지 확인 하라.

- 2010년~2014년 4년간의 매출 전망 지수 데이터와 매출 실적 지수 데이터를 시계열 데이터로 변환하고, 회귀 분석을 통해 상관관계를 시각화하라.

- 2010년~2014년 매출 전망 지수 데이터를 로드한다.
- 2010년~2014년 매출 실적 지수 데이터를 로드한다.
- 로드한 4년간의 매출 전망 및 실적 지수를 시계열 객체로 변환하고, 하나의 객체로 병합한다.
- 상관계수를 계산하고, 선형회귀분석을 통해 상관관계를 시각화한다.



패션 

Intermediate Level

중급과정







I 개요

개요

57

56

I

개요

> 개요

국가 통계 포털(<http://kosis.kr/>)로부터 수집하여 제공하는 과거 10년간 (2004년~2013년) 패션, 의류 분야 온라인/오프라인 쇼핑몰 매출 통계 조사 자료로부터 쇼핑몰 매출 데이터를 시계열 분석을 위한 성분 분석 하여 계절 패턴과 추세를 추출하고, 이를 바탕으로 R Studio에서 제공 하는 예측분석 패키지인 Forecast 라이브러리를 활용하여 ARIMA 모형을 이용한 예측 분석을 실행, 패션 분야 온라인 쇼핑몰의 미래 매출 곡선을 예측한다.

> 활용 데이터

- **fashion_mall_sales.csv** : 2004~2013년(10년간) 온라인/오프라인 패션/의류 쇼핑몰 월간 매출 통계 데이터

> 선행학습

- **리눅스** – 파일시스템 구조, 쉘 명령어, 쉘 스크립트 실행 방법
- **R 프로그래밍 언어** – 기본 지식(문법, 패키지 추가 설치 방법)
- **통계** – 시계열 성분 분석(decompose), 예측분석 (ARIMA 분석)
- **R 차트** – 설정 방법, 멀티 차트 표현 방법

▶ 요구사항

- 수집된 2004년~2013년 패션 부문 온라인/오프라인 쇼핑몰의 매출 통계를 시각화하고, 계절에 따른 효과, 매출 수준의 장기적 추세 등을 파악하라.
- 10년간의 시계열 데이터에 대해 성분 분석과 예측 분석을 통해 미래 매출 곡선을 계산하고 시각화하라.

▶ 분석 절차

- 수집된 2004년~2013년 온라인/오프라인 쇼핑몰 데이터를 로드한다.
- 예측 분석에 사용할 시계열 데이터 제어를 위해 zoo 라이브러리 객체로 변환 한다.
- 지난 10년간의 데이터를 예측 분석을 위해 시각화하고, 계절별로 반복적인 패턴이 있는지 패턴 분석을 하고, 장기적인 추세가 어떠한지 시각화 그래프로 확인한다.
- 시계열 데이터의 성분 분석(decomposition) 기법을 통해 눈으로 확인한 매출 그래프의 패턴을 추세(trend), 계절 패턴(seasonal) 등의 성분으로 나누어 시각화한다.
- R에서 제공하는 예측 분석 라이브러리(forecast)를 활용하여 미래 매출 곡선을 계산하고 시각화한다.



용어 정리

- **Forecast 라이브러리** : R을 활용한 예측 분석에 사용되는 매우 강력한 패키지 라이브러리이다. 특히, 자동화된 ARIMA 분석 함수를 제공한다. 입, 출력에 사용되는 시계열 데이터 객체로는 Zoo 라이브러리 객체를 사용한다.
- **zoo 라이브러리** : R을 활용한 시계열 분석에 매우 보편적으로 사용되는 패키지 라이브러리이다. 내부적으로 Index/Date/Time 을 키로 가지는 여러 항목의 시계열 데이터를 처리하기 위한 매트릭스 형태의 자료구조를 지니고 있으며, 시계열 항목(컬럼)간의 연산에 관련된 유용한 함수들을 내포하고 있다.
- **ARIMA 모형** : 시계열 데이터의 t 시점의 데이터가 과거 시점($t-1, t-2, \dots$)의 데이터에 의해 설명이 가능하다는 가정하에, 과거 관측치를 활용하여 미래 시점의 데이터를 예측하기 위해 만들어진 가장 일반적인 예측 분석 모형. 자기회귀누적이동평균 모형이라고도 부른다.



II 수집

개요	61
수집 데이터	62
데이터 수집	66
데이터 작업 영역 이동 스크립트	69



수집

> 개요

패션 데이터는 국가통계포털(<http://kosis.kr/>)로부터 10년간의 패션/의류 부문 쇼핑몰 매출 통계 데이터를 분석에 용이한 데이터만 추출하여 제공한다.

> 수집 방법

- 데이터 제공 :** 패션/의류 쇼핑몰 매출 통계 데이터는 국가 통계 포털 (<http://kosis.kr/>)로부터 제공해 주는 데이터를 OpenAPI, 자료수집기(Crawler)를 통하여 데이터를 수집하였고, 실습용 자료는 빅데이터 분석 활용센터에 접속하여 패션 데이터 셋을 다운로드 할 수 있도록 원시데이터를 제공하고 있다.

The screenshot shows the KOSIS website interface. At the top, there's a search bar and navigation links for '국가통계', '지역통계', '국제·북한통계', '인구통계', '온라인간행물', '서비스현황안내', and '통계설명자료'. Below the header, there's a large graphic with the text '고용률(14.11) 60.8%' and some icons. To the right of this graphic is a table titled 'KOSIS 100대 지표' with data for '고용률(14.11)' and other metrics like '고용증가율(14.11)', '실업률(14.11)', and '1인당 국민소득(14.11)'. Further down, there are sections for '주제별통계' (including icons for '인구 가구', '환경', '교통·경보신', '생활·금융', '도소매·서비스', '무역·국경수지', '문화·언어', and '보건·사회 복지') and '팝업존' (with a cartoon illustration of a person holding a book). At the bottom, there are buttons for '동계사각화 컨텐츠', '공지사항', '보도자료', '최근수록자료', '도너메일리', and '더보기'.

> 수집 데이터

> 매출 통계 데이터 데이터(fashion_mall_sales.csv)

DATE	TOTAL	ONLINE	ONOFFLINE
2004-01-01	69058	38783	30275
2004-02-01	67260	39383	27876
2004-03-01	77778	46681	31097
2004-04-01	77469	48089	29380
2004-05-01	75828	48703	27125
2004-06-01	75669	49193	26475
2004-07-01	73539	48827	24712
2004-08-01	60178	38432	21746
2004-09-01	73244	50178	23066
2004-10-01	86063	61065	24997
2004-11-01	95525	67727	27798
2004-12-01	102193	73825	28368
2005-01-01	96239	71803	24436
2005-02-01	89804	67921	21883
2005-03-01	109672	84737	24935
2005-04-01	116257	92111	24146
2005-05-01	123367	99006	24361
2005-06-01	118720	96696	22025
2005-07-01	120507	99595	20913
2005-08-01	109680	89600	20081
2005-09-01	139793	115220	24573
2005-10-01	162566	135856	26710
2005-11-01	188369	158870	29500
2005-12-01	208125	175748	32377
2006-01-01	176442	147654	28788
2006-02-01	164393	138768	25625
2006-03-01	193179	163598	29581
2006-04-01	185118	155996	29122
2006-05-01	195751	166291	29460
2006-06-01	186968	161982	24986
2006-07-01	185166	161425	23741
2006-08-01	160598	138495	22103

II. 수집

DATE	TOTAL	ONLINE	ONOFFLINE
2006-09-01	215900	188102	27798
2006-10-01	206956	178317	28639
2006-11-01	252721	217843	34878
2006-12-01	248466	214596	33870
2007-01-01	212763	179269	33494
2007-02-01	195627	167341	28285
2007-03-01	231646	198950	32696
2007-04-01	221702	188228	33474
2007-05-01	231393	196394	34999
2007-06-01	212746	180891	31855
2007-07-01	207483	178186	29297
2007-08-01	171964	145084	26880
2007-09-01	211801	183258	28543
2007-10-01	259305	224895	34410
2007-11-01	284249	242637	41612
2007-12-01	273312	233447	39866
2008-01-01	241626	202642	38984
2008-02-01	219018	182407	36611
2008-03-01	265293	227103	38190
2008-04-01	256013	216653	39360
2008-05-01	246924	214619	32305
2008-06-01	239908	201697	38211
2008-07-01	239573	203508	36065
2008-08-01	183841	153648	30193
2008-09-01	227938	190776	37162
2008-10-01	284052	239149	44903
2008-11-01	290192	244045	46146
2008-12-01	301214	251866	49348
2009-01-01	251027	206929	44097
2009-02-01	256222	209083	47139
2009-03-01	291521	238835	52686
2009-04-01	279999	224390	55609

DATE	TOTAL	ONLINE	ONOFFLINE
2009-04-01	275648	223602	52047
2009-06-01	277030	225041	51989
2009-07-01	281820	227769	54051
2009-08-01	226406	176243	50163
2009-09-01	283416	221749	61667
2009-10-01	326280	254746	71534
2009-11-01	379250	294447	84803
2009-12-01	395250	301739	93511
2010-01-01	331177	248272	82905
2010-02-01	285447	217260	68187
2010-03-01	355410	272982	82428
2010-04-01	349609	260371	89238
2010-05-01	354680	273738	80942
2010-06-01	349578	268987	80591
2010-07-01	336712	258548	78163
2010-08-01	271149	196176	74973
2010-09-01	313018	230681	82337
2010-10-01	400180	298848	101332
2010-11-01	440222	328122	112099
2010-12-01	460929	338130	122799
2011-01-01	400399	279387	121012
2011-02-01	308180	213754	94427
2011-03-01	417261	299428	117833
2011-04-01	390652	276389	114263
2011-05-01	395116	283425	111691
2011-06-01	396658	286686	109973
2011-07-01	367094	262897	104197
2011-08-01	339105	241793	97312
2011-09-01	386835	275863	110972
2011-10-01	441619	311235	130384
2011-11-01	499686	360360	139327
2011-12-01	526684	374357	152327

II. 수집

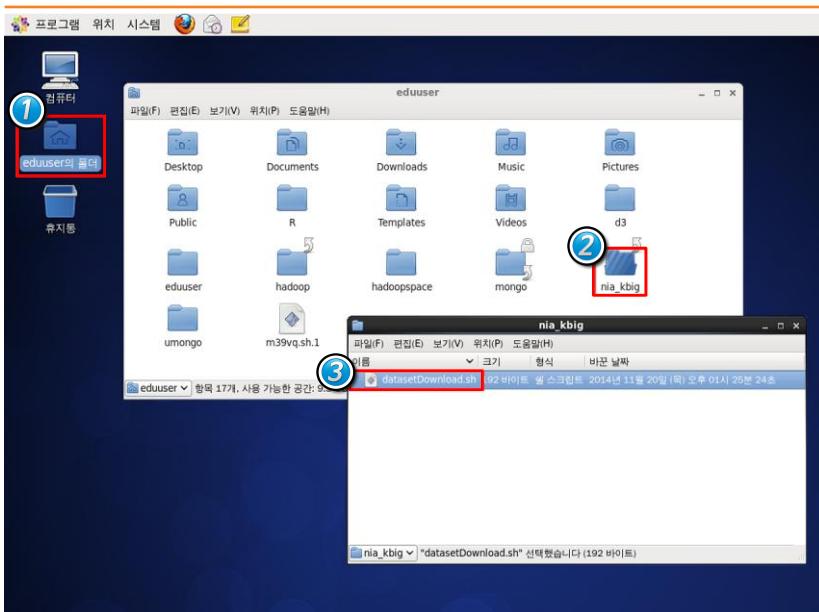
DATE	TOTAL	ONLINE	ONOFFLINE
2012-01-01	446926	310652	136273
2012-02-01	432352	291623	140729
2012-03-01	435784	299377	136406
2012-04-01	430001	303216	126786
2012-05-01	446954	319830	127124
2012-06-01	429320	313258	116061
2012-07-01	437810	313749	124061
2012-08-01	369835	251964	117870
2012-09-01	430818	306293	124525
2012-10-01	535913	382328	153585
2012-11-01	608285	433476	174809
2012-12-01	605526	423304	182222
2013-01-01	498249	340897	157352
2013-02-01	417431	283151	134280
2013-03-01	517305	372221	145084
2013-04-01	509526	363974	145552
2013-05-01	527358	377832	149526
2013-06-01	481583	344823	136760
2013-07-01	502784	358863	143921
2013-08-01	412928	274296	138631
2013-09-01	476736	325002	151734
2013-10-01	584295	395589	188707
2013-11-01	664780	455824	208956
2013-12-01	687679	469437	218242

- 10년간(2004년~2013년)의 온라인/오프라인 쇼핑몰의 매출 통계를 활용할 수 있다.

> 데이터 수집(datasetDownload.sh)

- 데이터 저장소에서 서버 로컬로 패션 데이터 셋을 복사해 온다.
fashion_mall_sales.csv : 2004~2013년(10년간) 온라인/오프라인 패션/
 의류 쇼핑몰 월간 매출 통계 데이터

> 실습코드 디렉토리로 이동

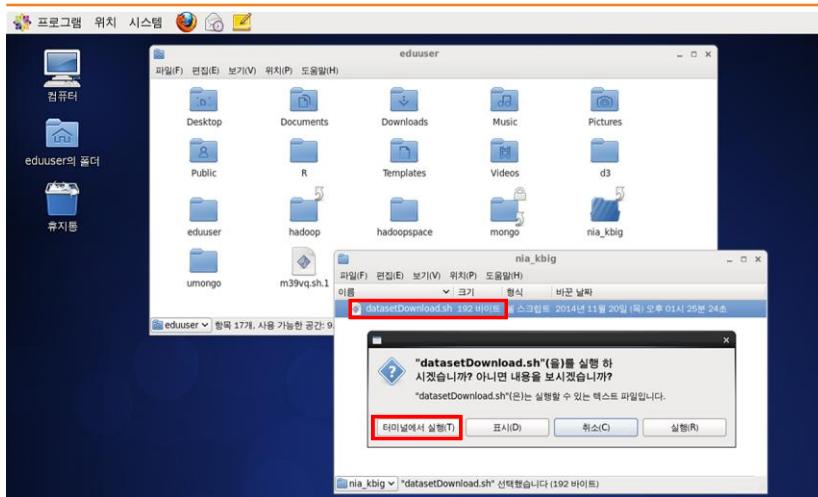


- 로그인 후 바탕화면에서 eduuser 폴더를 오픈한다.
- nia_kbig 폴더를 오픈한다.
- datasetDownload.sh를 더블클릭하여 실행한다.

II. 수집

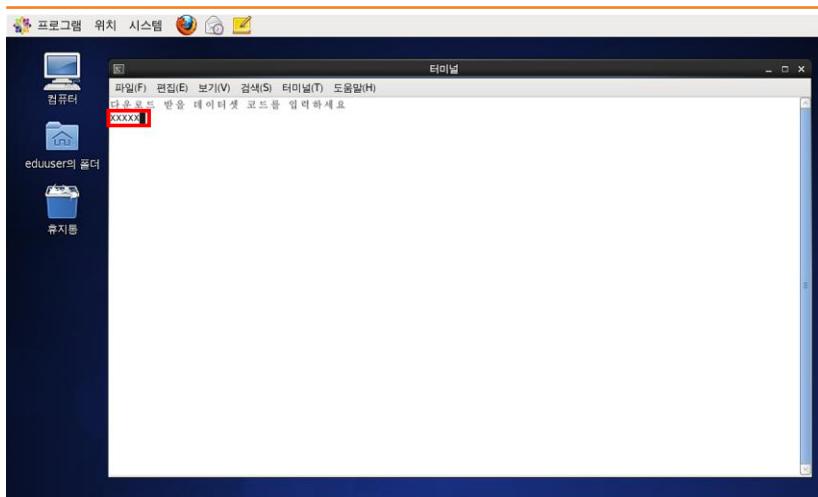
▶ 레파지토리에서 데이터 수집

datasetDownload.sh (원시데이터로 컬서버로 복사)



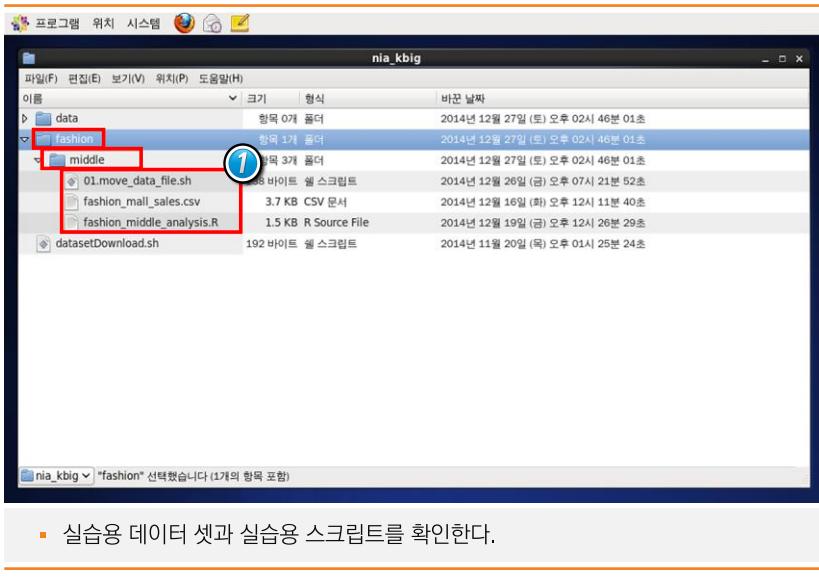
- '터미널에서 실행' 버튼을 클릭한다.

▶ 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 :

로컬로 수집해온 데이터를 작업 영역 Data 폴더로 자료를 이동하는 스크립트

▪ fashion_middle_analysis.R :

패션 데이터 분석용 R 스크립트

▪ fashion_mall_sales.csv : 2004~2013년(10년간) 온라인/오프라인 패션/

의류 쇼핑몰 월간 매출 통계 데이터

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 (작업 영역 폴더로 원시데이터 이동)

```
01.#!/bin/bash  
02. # 패션 데이터 정의  
03. TARGET_FASHION=/home/eduuser/nia_kbig/fashion/middle/*.csv  
04. # 작업 디렉토리 정의  
05. LOCAL_DIR=/home/eduuser/nia_kbig/data/  
06. mv $TARGET_FASHION $LOCAL_DIR  
07.
```



- 데이터 작업 영역 이동 스크립트 소스(01.move_data_file.sh)
- 라인 03 : 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.

▶ 수집 데이터 셋 작업 영역 폴더 이동

- R Studio에서 시계열 분석/패턴분석/예측 분석을 위한 수집된 데이터 셋을 작업 영역 Data 폴더로 자료를 이동

```

eduuser@localhost middle] $ ll
합계 8
-rwxr-xr-x. 1 eduuser eduuser 216 2014-12-14 08:46 01.move_data_file.sh
-rw-r--r--. 1 eduuser eduuser 3751 2014-12-16 12:11 fashion_mall_sales.csv
[eduuser@localhost middle] $ ./01.move_data_file.sh

```

- “./01.move_data_file.sh”를 입력하여 준비된 패션 데이터를 이동시킨다.





III 가공

개요

73

데이터 가공 R 스크립트

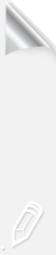
77



가공

> 개요

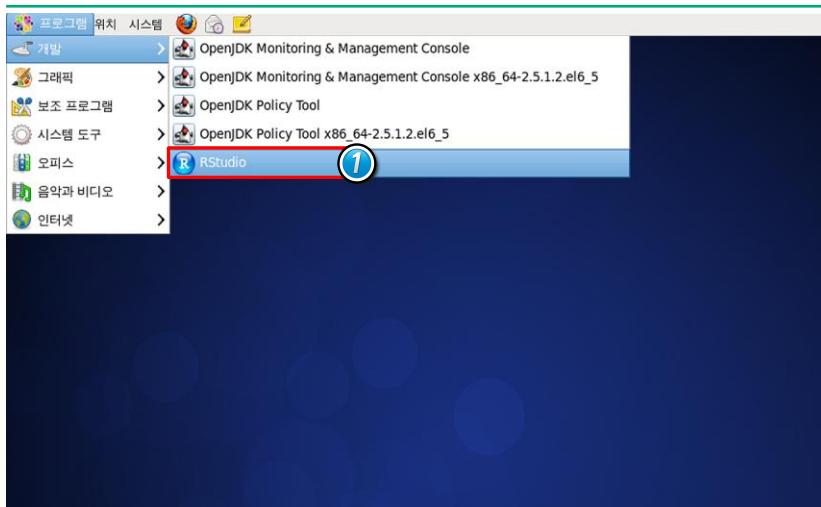
작업 영역 폴더에 복사한 패션 데이터의 가공은, 수집된 10년간(2004년 ~2013년)의 패션/의류 쇼핑몰 분야 매출 통계 데이터를 R Studio에서 로드하여 시계열 데이터 제어를 위해 zoo 라이브러리 객체 형태로 변환하도록 한다.



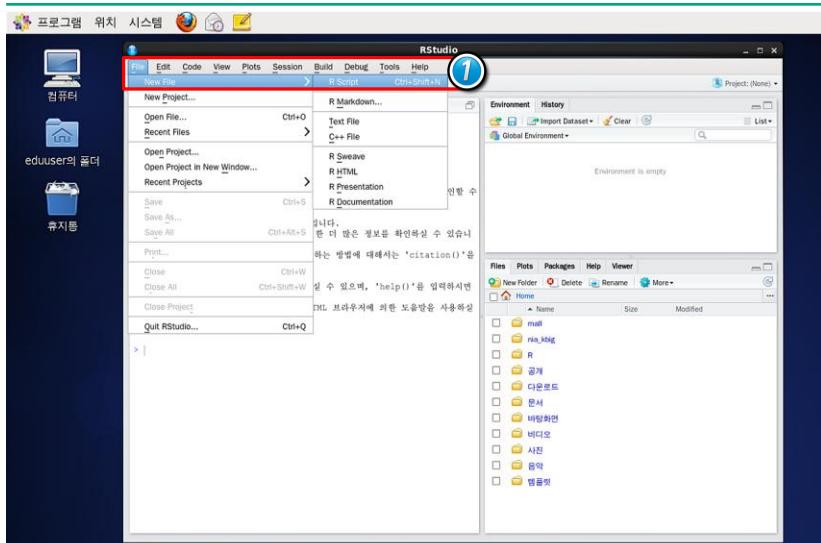
> 가공 방법

- **분석 도구 실행** : 가공 분석을 위해, 프로그래밍 도구인 R Studio를 실행한다.
- **데이터 로드** : 2004년~2013년 의류 분야 매출 통계 데이터를 R Studio에서 읽어들인다.
- **데이터 변환** : R Studio에는 시계열 분석을 위한 여러가지 라이브러리가 존재한다. 이 중 가장 일반적으로 사용하는 “zoo”라이브러리를 활용하여 데이터를 “zoo” 객체로 변환한다.

▶ 데이터 가공

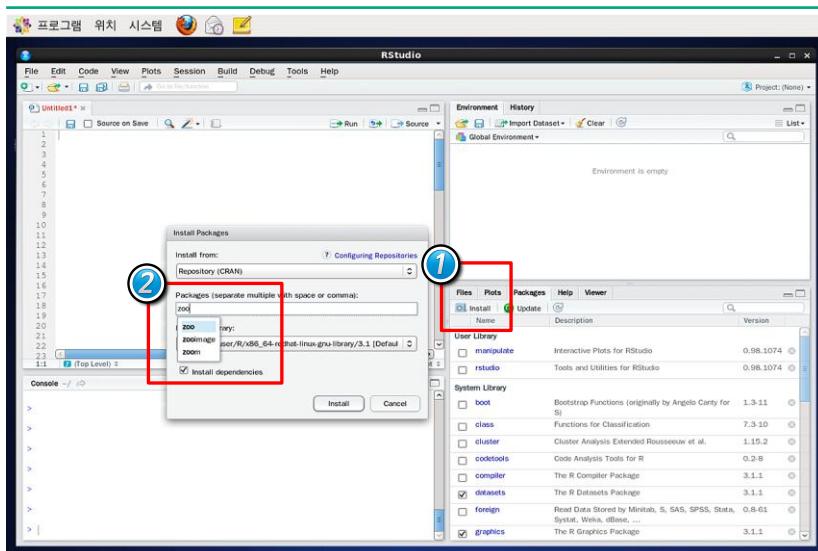


- ① 왼쪽 상단의 [“프로그램” 클릭] > [개발] 클릭 > [“RStudio” 클릭]으로 분석 도구인 R Studio를 실행한다.



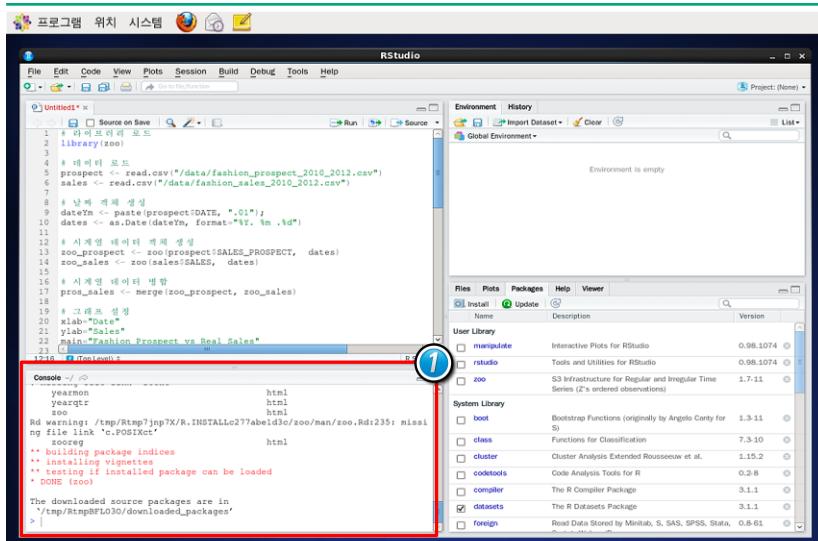
- ② 패션 데이터의 분석 및 가공을 위해 프로그램 작업 파일 (“New File” 클릭) “R_Script” 클릭)을 선택한다.

III. 가공



3. R 에서 시계열 분석을 위해 일반적으로 많이 사용되는 “zoo” 라이브러리를 설치한다.

① 패키지 탭에 install 버튼을 누른 후 ② zoo 를 입력하여 라이브러리를 선택한다.



4. ① “zoo” 라이브러리가 설치되는 모습을 확인할 수 있다.

The screenshot shows the RStudio interface. In the top-left pane, there is an 'Untitled.R' script with the following R code:

```

1 library(zoo)
2
3 # 데이터 로드
4 sales <- read.csv("/home/eduuser/nia_kbkg/data/fashion_mall_sales.csv")
5
6 # 날짜 객체 생성
7 dates <- as.Date(sales$DATE, format="%Y-%m-%d")
8
9 sales <- sales[, !(colnames(sales) %in% c("DATE"))]
10
11 # 시계열 데이터 객체 생성
12 zoo_sales <- zoo(sales, dates)
13
14
15
16
17
18
19
20
21
22
23

```

In the top-right pane, the 'Global Environment' tab is selected, showing the variables defined in the script:

- Data**: sales (120 obs. of 3 variables)
- Values**: dates (2004-01-01), zoo_sales ('zoo' series from 2004-01-01 to 2013-12-01)

The bottom pane is the 'Console'.

5. 2004년~2013년 매출 통계 데이터를 로드하고, 시계열 데이터 제어가 용이한 zoo 라이브러리 객체로 변환하도록 R 스크립트를 작성한다.

This screenshot is similar to the previous one, but with two annotations:

- Annotation ①: A red circle with the number 1 points to the first few lines of the R script where the 'zoo' library is loaded and the 'sales' data is read.
- Annotation ②: A red circle with the number 2 points to the 'zoo_sales' variable in the 'Values' section of the Global Environment pane, indicating that the data has been successfully converted into a zoo object.

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

III. 가공

▶ 데이터 가공 R 스크립트

```
01. library(zoo)
02.
03. # 데이터 로드
04. sales <- read.csv("/home/eduuser/nia_kbig/data/fashion_mall_sales.csv",
05. ↪ header=T)
06.
07. # 날짜 객체 생성
08. dates <- as.Date(sales$DATE, format="%Y-%m-%d")
09.
10. sales <- sales[, !(colnames(sales) %in% c("DATE"))]
11.
12. # 시계열 데이터 객체 생성
13. zoo_sales <- zoo(sales, dates)
```



- 패션 분석 및 시각화 R 스크립트 소스(fashion_basic_analysis.R)
- 라인 04 : 2004년~2013년 온/오프라인 패션 쇼핑몰 매출 통계 데이터 파일(fasion_mall_sales.csv)을 읽어들여 R 데이터 객체(sales)로 저장하는 라인이다.
- 라인 07 : zoo라이브러리 객체를 만들기 위해 날짜 객체를 생성하는 라인이다.
- 라인 09 : 4라인에서 읽어들인 R 데이터 객체에서 날짜 컬럼을 제거하는 라인이다.
- 라인 12 : zoo함수를 활용하여 9 라인에서 가공한 R 데이터 객체와 7 라인에서 생성한 날짜 객체로부터 zoo라이브러리 객체(zoo_sales)를 생성하는 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

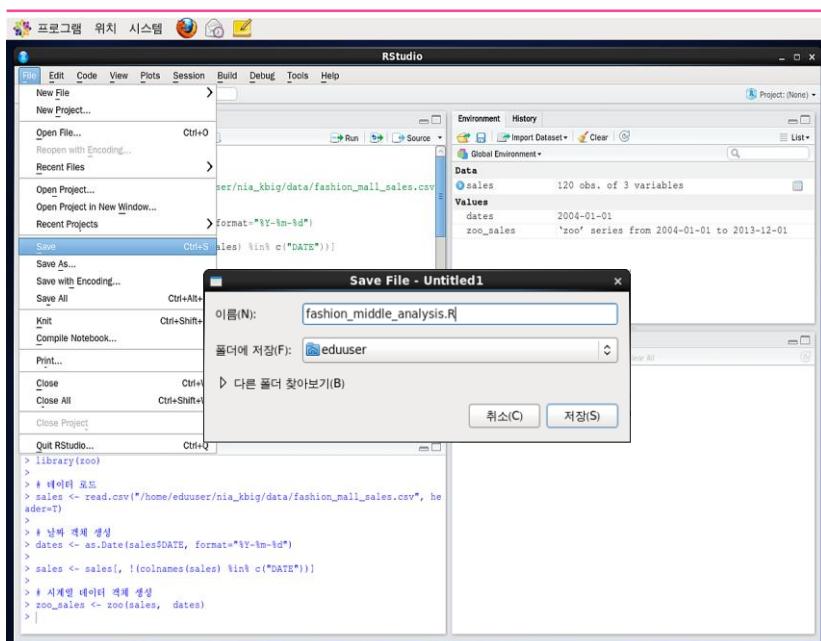
R Studio를 활용하여 데이터 로드 > 가공 > 분석 > 시각화 단계를 한번에 실행하므로, 별도의 저장 과정은 생략한다. 가공된 데이터는 메모리상에 존재하며, 지금까지 작성한 분석 프로그램 소스를 저장한다.

> 저장 방법

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

▶ R Studio 활용 저장

▶ 데이터 저장



1. 패션 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/fashion_middle_analysis.R”로 저장한다.

W





V 분석

개요	85
R Studio 활용 분석	87
R Studio 저장	93

V 분석

> 개요

패션 데이터 분석은 R Studio에 내장된 zoo 라이브러리 객체 및 그래프 기능을 활용한다. 2004년~2013년 패션/의류 분야 온/오프라인 쇼핑몰 매출 통계 데이터에 대해 시계열 성분 분석(decomposition)으로 트렌드 와 계절 패턴을 분석하고, forecast 라이브러리를 활용하여 ARIMA 예측 분석 기법으로 미래의 온/오프라인 쇼핑몰 매출 곡선을 예측한다.

> 데이터 분석 방법

- 가공 단계에서 시계열 분석 객체로 변환한 2004년 ~ 2013년 매출 통계 데이터를 시각화하여 데이터의 패턴을 개괄적으로 파악한다.
- 시계열 성분 분석을 통해 쇼핑몰 매출의 추세(trend), 계절적 반복 패턴 (seasonal) 등 성분을 추출하여 데이터 패턴을 정확하게 파악한다.
- 예측 분석 기법인 ARIMA 분석을 적용하여 미래의 매출 곡선을 계산한다.



• ARIMA 모델

- 현실세계의 경제가 과거의 지식과 경험에 기초한 행동을 하는 사람들에 의해 움직이고 있다는 사실에 기초로 하여 미래를 예측하기 위해 고안한 모델이다.
- 확실한 추세를 가지며, 계절적 요인 등에 의해 반복적인 패턴이 드러나는 데이터의 경우에 있어 미래 예측에는 ARIMA 모델은 매우 적합하다.
- 그 외의 시계열 데이터 모델에는 시간 영역보다는 반복되는 주파수(frequency)에 초점을 두는 푸리에 분석 등이 있다.

• forecast 패키지

- ARIMA 모델을 쉽게 적용하기 위해 forecast 패키지에서 제공되는 auto.arima 함수는, ARIMA 분석시에 가장 어려운 부분이기도 한 모형 차수 값을 자동으로 찾아내어 ARIMA 모델을 적용시켜 주는 함수이다.
- R Studio에서 forecast 패키지 설치가 정상적으로 되지 않는 경우
<http://cran.r-project.org/src/contrib/Archive/RcppArmadillo/>
위 URL에서 운영체제에 적합한 버전을 다운로드하여 설치한 후, forecast 패키지의 설치를 다시 시도하면 정상적으로 설치된다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화

➤ R Studio 활용 분석

➤ 데이터 불러오기

zoo_sales (가공 단계에서 생성한 10년간의 매출 통계 데이터)

```

library(zoo)
# 데이터 로드
sales <- read.csv("/home/eduuser/nia_kbig/data/fashion_mall_sales.csv")
# 날짜 객체 생성
dates <- as.Date(sales$DATE, format="%Y-%m-%d")
sales <- sales[, !(colnames(sales) %in% c("DATE"))]
# 시계열 데이터 객체 생성
zoo_sales <- zoo(sales, dates)
# 데이터 확인
zoo_sales

```

Console

```

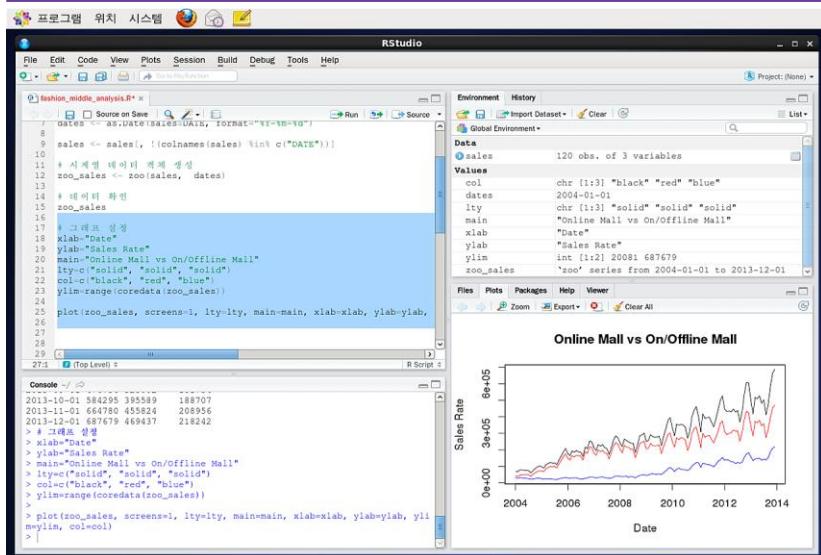
2012-12-01 605526 423304 18222
2013-01-01 498249 340897 157352
2013-02-01 417654 301515 134286
2013-03-01 527395 373811 145944
2013-04-01 509526 363974 145552
2013-05-01 527358 377832 149526
2013-06-01 481583 344823 136760
2013-07-01 509584 348863 147361
2013-08-01 502293 278002 138691
2013-09-01 476734 325002 151734
2013-10-01 584295 395589 188707
2013-11-01 664780 455824 208956
2013-12-01 687679 469437 218242
> |

```

- ❶ 가공한 데이터가 잘 들어가 있는지 확인하기 위해 “zoo_sales”를 입력하고 위와 같이 블럭을 선택한 후, Ctrl+Enter 를 입력하면, ❷ 와 같이 데이터를 확인할 수 있다.

▶ 데이터 분석

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



1. 패션/의류 쇼핑몰의 매출 데이터가 어떠한 패턴으로 나타나는지 파악하기 위해 가공 단계에서 통합한 2004년~2013년 매출 통계 데이터를 plot 함수를 활용하여 시각화해 본다.

- #주) 그래프 모양을 보면, 1년(12개월) 단위로 같은 모양을 보이면서 우상향하는 추세를 그리고 있는 것을 파악할 수 있다. (즉, '추세'와 '반복 패턴'을 가진다.)

```

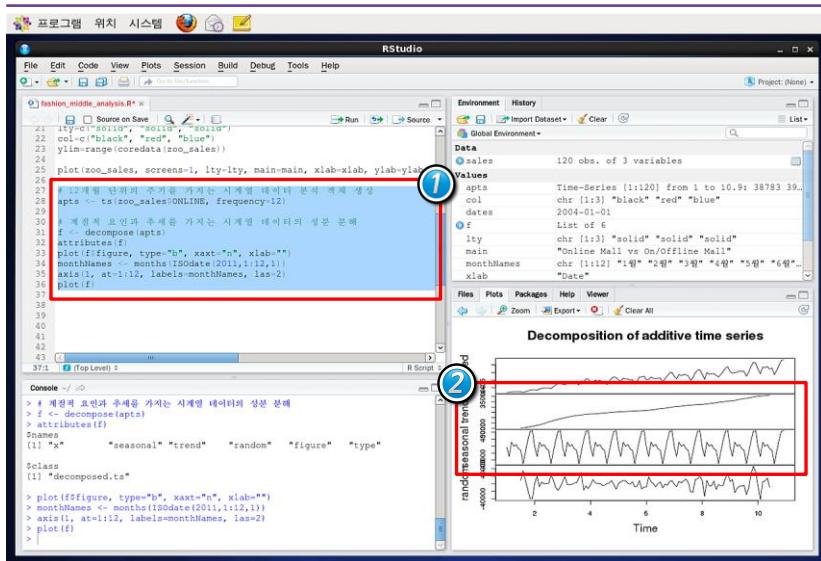
1. # 그래프 설정
2. xlab="Date"
3. ylab="Sales Rate"
4. main="Online Mall vs On/Offline Mall"
5. lty=c("solid", "solid", "solid")
6. col=c("black", "red", "blue")
7. ylim=range(coredata(zoo_sales))
8. plot(zoo_sales, screens=1, lty=lty, main=main, xlab=xlab, ylab=ylab, ylim=ylim, col=col)

```



- 패션 분석 및 시각화 R 스크립트 소스(fashion_basic_analysis.R)
- 라인 01~07 : 시각화를 하기 위해 그래프 설정값들을 지정하는 라인이다.
- 라인 08 : plot 함수를 사용하여 2010~2012 전망지수 대비 실적지수를 시각화하는 라인이다.

V. 분석



2. ①② 그래프를 통해 시각적으로 파악한 추세(trend)와 반복 패턴(seasonal)을 성분 분석 기법을 통해 추출하여 정확하게 파악한다.

- #주) 시각화 결과, 전망의 오차에 대해서는 12개월 단위나 분기 단위의 특별한 패턴은 보이지 않는다. 따라서, 월별로 패턴을 적용하기 보다 평균 비율을 활용하여 예측 하기로 한다.

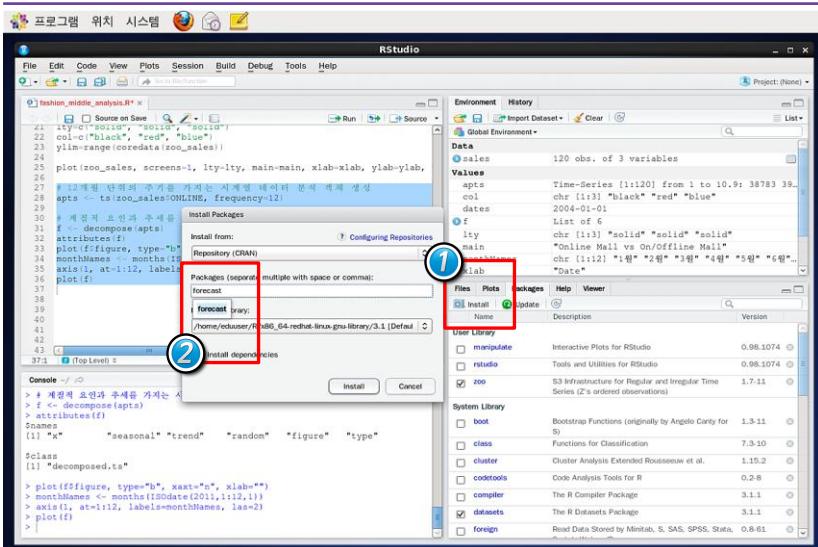
```

1. # 12개월 단위의 주기를 가지는 시계열 데이터 분석 객체 생성
2. apts <- ts(zoo_sales$ONLINE, frequency=12)
3. # 계절적 요인과 추세를 가지는 시계열 데이터의 성분 분해
4. f <- decompose(apts)
5. attributes(f)
6. plot(f$figure, type="b", xaxt="n", xlab="")
7. monthNames <- months(ISOdate(2011,1:12,1))
8. axis(1, at=1:12, labels=monthNames, las=2)
9. plot(f)

```

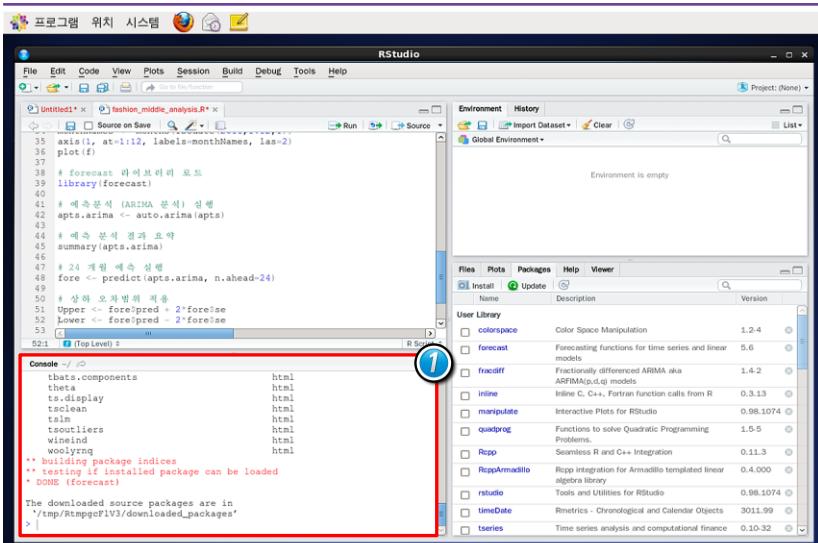


- 패션 분석 및 시각화 R 스크립트 소스(fashion_basic_analysis.R)
- 라인 02 : ts함수를 사용하여 12개월 단위의 주기를 가지는 시계열 데이터를 기공하여 객체(apts)로 저장하는 라인이다.
- 라인 04~05 : decompose 함수를 사용하여 시계열 데이터를 추세(trend), 계절패턴(seasonal), 무작위 성분(random) 등 세가지 성분으로 분해하는 라인이다.
- 라인 07~09 : 성분별로 분해한 데이터를 plot 함수를 사용하여 멀티 그래프로 시각화하는 라인이다.



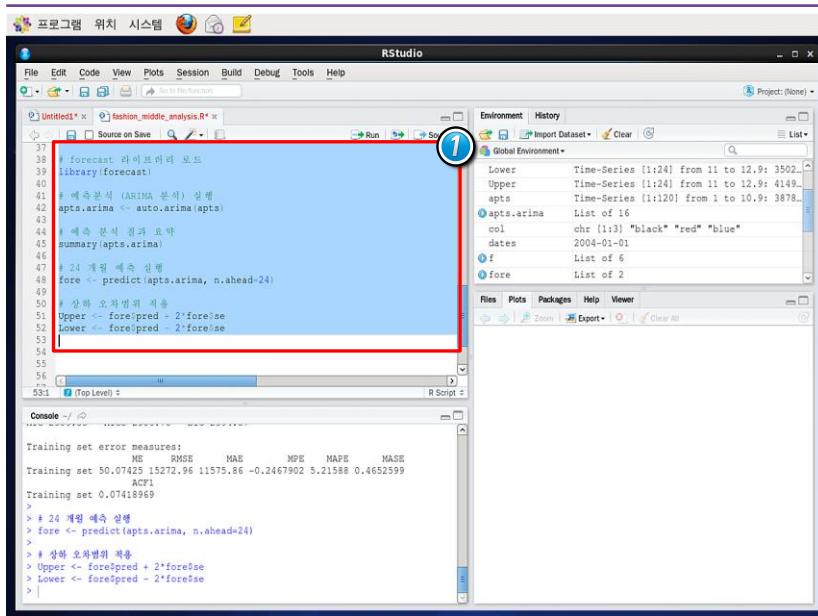
3. R에서 제공하는 예측분석 라이브러리인 “forecast” 라이브러리를 설치한다.

① 패키지 탭에 install 버튼을 누른 후 ② forecast를 입력하여 라이브러리를 선택한다.



4. ① “forecast” 라이브러리가 설치되는 모습을 확인할 수 있다.

V. 분석



5. ① R “forecast” 라이브러리에서 제공하는 ARIMA 분석 함수를 통해 예측 분석을 실행하고, predict 함수를 통해 미래 24개월의 매출 곡선을 예측한다. ±오차 범위를 적용하여 Upper/Lower 매출 곡선도 계산한다.

```
1. library(forecast)
2.
3. # 예측 분석(ARIMA 분석) 실행
4. apts <- ts(zoo_sales$ONLINE, frequency=12)
5. apts.arima <- auto.arima(apts)
6.
7. summary(apts.arima)
8.
9. # 24개월 예측 실행
10. fore <- predict(apts.arima, n.ahead=24)
11.
12. # 상하 오차 범위 적용
13. Upper <- fore$pred + 2*fore$se
14. Lower <- fore$pred - 2*fore$se
```



- 패션 분석 및 시각화 R 스크립트 소스(fashion_basic_analysis.R)
- 라인 01 : 예측 분석을 실행하기 위한 라인이다.
- 라인 04 : ts 함수를 사용하여 시계열 분석용 객체(apts)로 변환하는 라인이다.
- 라인 05 : auto.arima 함수를 사용하여 ARIMA 분석을 실행하여 결과를 apts.arima 객체로 저장하는 라인이다.
- 라인 07 : ARIMA 분석의 결과를 요약하여 출력해 보는 라인이다.
- 라인 10 : predict 함수를 사용하여 예측분석을 실행한 결과를 객체(fore)로 저장하는 라인이다.
- 라인 13~14 : 예측 분석 곡선의 상하 오차 범위 곡선을 계산하여 각각의 객체(Upper, Lower)로 저장하는 라인이다.

I. 개요

II. 수집

III. 가공

IV. 저장

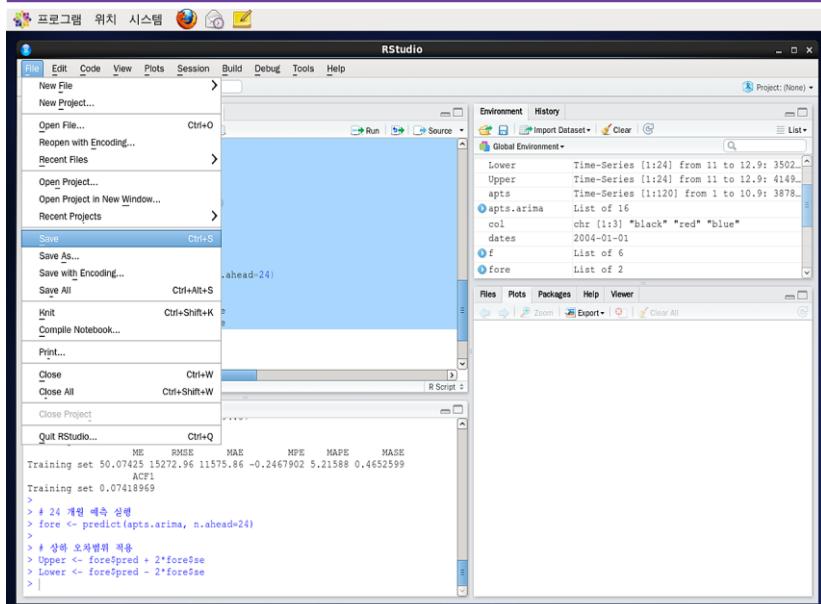
V. 분석

VI. 시각화

➤ 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 left pane displays the R script editor with the following R code:

```
ME RMSE MAE MPE MAPE MASE
Training set 50.07425 15272.96 11575.86 -0.2467902 5.21588 0.4652599
ACF1
Training set 0.07418969
> | 24 예측 신생
> fore <- predict(amps.arima, n.ahead=24)
> |
> # 상하 오차범위 적용
> Upper <- fore$pred + 2*fore$se
> Lower <- fore$pred - 2*fore$se
> |
```

The right pane shows the Global Environment, listing variables such as Lower, Upper, apts, col, dates, f, and fore, along with their types and values.

1. “File/Save”를 클릭하여 지금까지 패션 데이터를 분석하기 위해 작성한 프로그램을 저장한다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화



1

2



VI 시각화

개요	97
분석 데이터 시각화	98
데이터 분석	99

VI

시각화

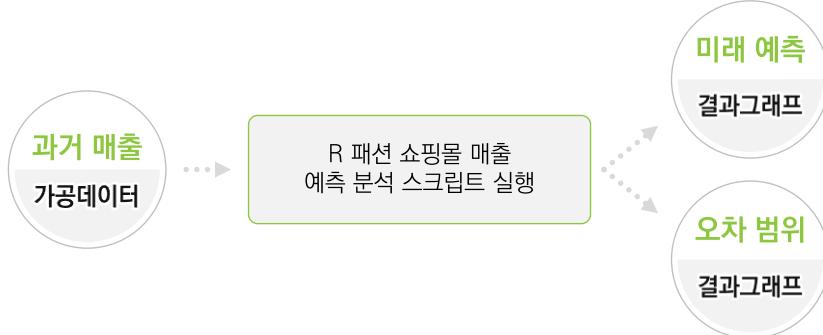
> 개요

패션 데이터의 분석 과정에서는 가공한 시계열 데이터를 시각화하여 데이터의 흐름을 눈으로 관측하고, 성분 분석 기법을 통해 눈으로 관측한 결과를 정확하게 시각화하였다. 이를 바탕으로 ARIMA 모형으로 예측 분석을 실행하여 미래 매출 곡선을 계산하였다. 예측된 매출 곡선에는 오차 범위가 존재하므로, 오차 범위와 함께 미래 매출 곡선을 시각화 한다.

> 시각화 방법

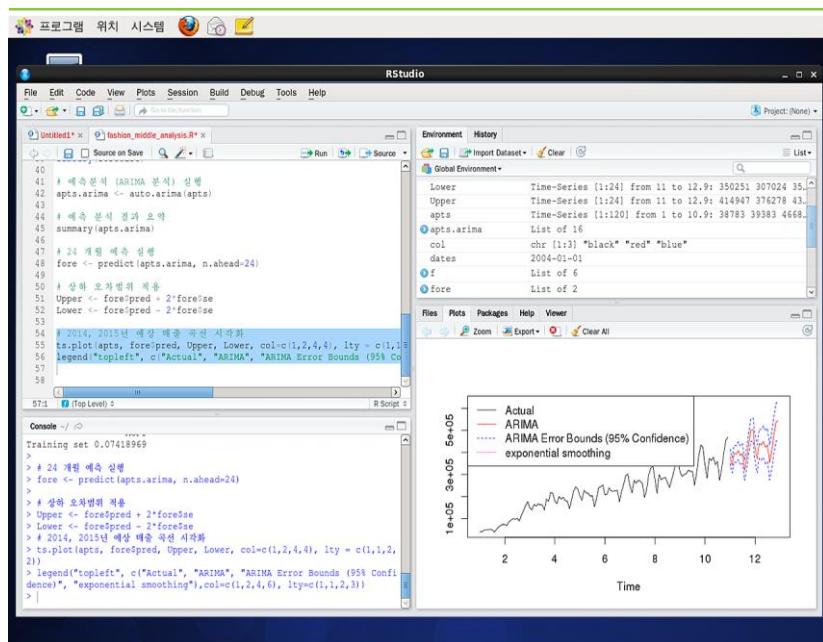
- 분석 과정에서 분석한 2년간(2014, 2015년)의 미래 매출 곡선을 꺾은선 그래프로 시각화한다.
- 이때, 오차 범위 곡선을 다른 색상, 다른 타입의 선으로 함께 표시한다.

> 시각화 과정



▶ 분석 데이터 시각화

▶ 데이터 시각화



1. 2004년~2013년 매출 통계 데이터에 대한 예측 분석을 통한 2014~2015년 예상
매출 곡선을 오차 범위를 포함하여 시각화한다.

```

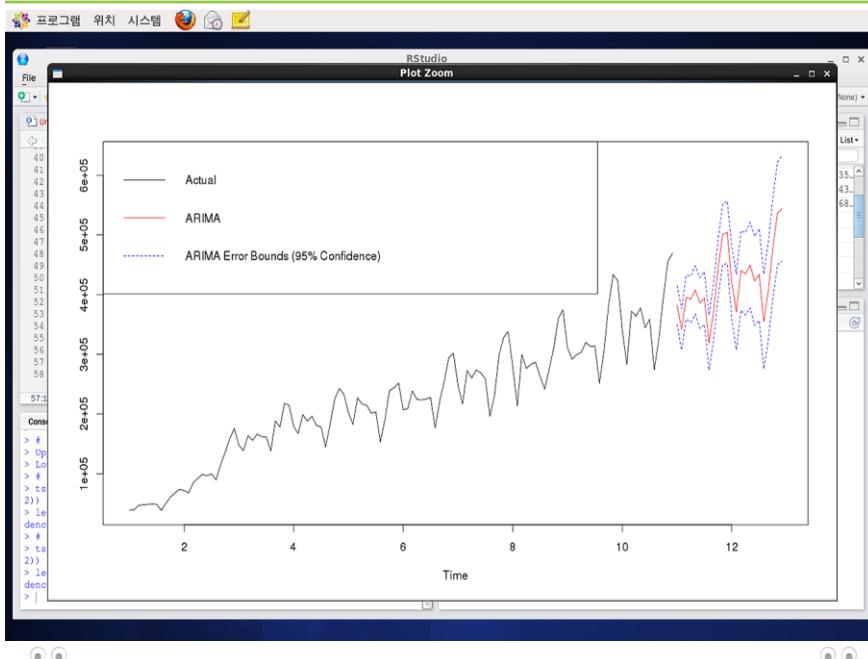
01. # 시각화
02. ts.plot(aps, fore$pred, Upper, Lower, col=c(1,2,4,4), lty = c(1,1,2,2))
03. # 범례 설정
04. legend("topleft", c("Actual", "ARIMA", "ARIMA Error Bounds (95% Confidence)", "exponential smoothing"), col=c(1,2,4,6), lty=c(1,1,2,3))

```



- 패션 분석 및 시각화 R 스크립트 소스(fashion_basic_analysis.R)
- 라인 02 : ts.plot 함수를 활용하여 2004년~2013년의 실제 매출 실적 곡선에 2014~2015년
매출 실적 곡선을 통합하여 그래프로 시각화하는 라인이다.
- 라인 04 : legend 함수를 사용하여 그래프에 범례를 표시하는 라인이다.

> 데이터 분석



- 최종 결과 그래프에는, 분석 과정에서 계산하여 예측한 2014~2015년 매출 예상 그래프는 붉은색 실선으로, 상하 오차 범위 곡선은 푸른색 점선으로 표현되어 있다.
 - 매출 예상 그래프를 보면, 상승추세(trend)와 반복패턴(seasonal)이 과거 10년간의 데이터에 비추어 봄도 거의 유사한 곡선을 그리며 예측이 정상적으로 이루어졌음을 알 수 있다.
 - 2014년~2015년 온라인 쇼핑몰 매출은 신뢰구간 95% 안에서 푸른색 상하 오차 범위 곡선을 벗어나지 않을 것임을 나타낸다.
 - 이처럼, 반복 패턴을 가지며 방향성을 가지는 데이터는 R 의 “forecast” 패키지를 활용하여 간단한 ARIMA 분석 만으로도 미래 예측 곡선을 만들어 낼 수 있다.
 - 이렇게 예측한 매출 예상 곡선은, 제조 계획, 마케팅 전략 수립 등 경영상의 결정에 도움이 되는 지표로 다양하게 활용된다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화



VII 예제문제

예제 문제1

103

예제 문제2

104

예 / 제 / 문 / 제

예제 1

2004년~2013년 온라인 쇼핑몰 매출과 온오프라인 쇼핑몰 매출의 비율 변화를 히스토그램으로 시각화하고, 비율 변화를 예측하라.

- 10년간의 온라인, 온오프라인 매장의 월별 매출 데이터를 분석하여, 매출 비율의 변화를 시각화하고, 이를 바탕으로 미래 매출 비율 변화를 예측하여 시각화하라.
 - 2004년~2013년 온라인 쇼핑몰, 온오프라인 쇼핑몰의 매출 데이터를 로드한 후, 매출 비율 그래프를 작성한다.
 - 10년간의 매출 비율 그래프로부터 성분 분석을 실행한다.
 - 미래 매출 비율의 변화를 예측하라.

예제 2

초급 분석에서 사용했던 2010년~2013년 매출 실적 지수 데이터와 온라인 쇼핑몰 매출 데이터를 상관 분석하라.

- 2010년~2013년 3년간의 패션/의류 매출 실적 지수 데이터와, 패션/의류 온라인 쇼핑몰의 실제 매출 통계 데이터를 시계열 데이터로 변환하고, 회귀 분석을 통해 상관관계를 시각화하라.
 - 2010년~2013년 매출 실적 지수 데이터를 로드한다.
 - 2010년~2013년 온라인 의류 쇼핑몰 매출 통계 데이터를 로드한다.
 - 로드한 3년간의 매출 자수 및 온라인 쇼핑몰 매출 통계 데이터를 시계열 객체로 변환하고, 하나의 객체로 병합한다.
 - 상관계수를 계산하고, 선형회귀분석을 통해 상관관계를 시각화한다.

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

