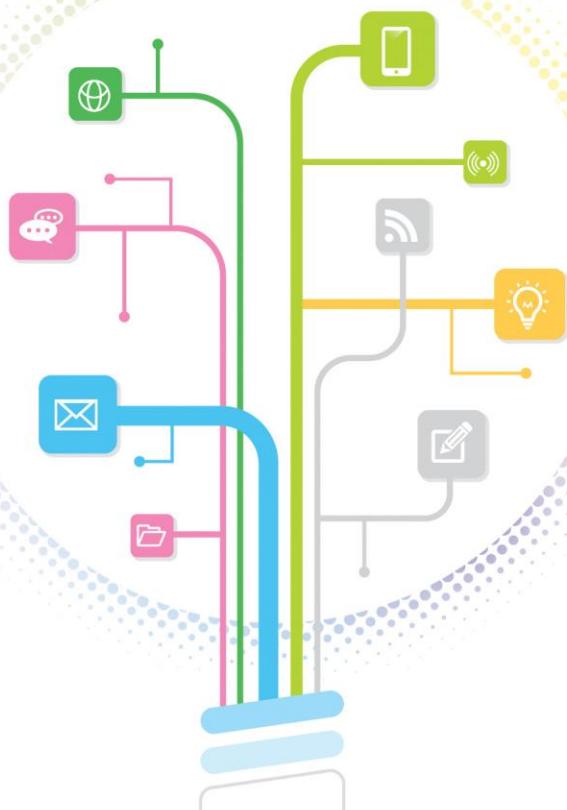


데이터 분석 콘텐츠 활용 매뉴얼



미래창조과학부



한국정보화진흥원



KBIG
빅데이터
전략센터

CONTENTS

Beginning Level 초급과정

I 개요

개요	9
----	---

II 수집

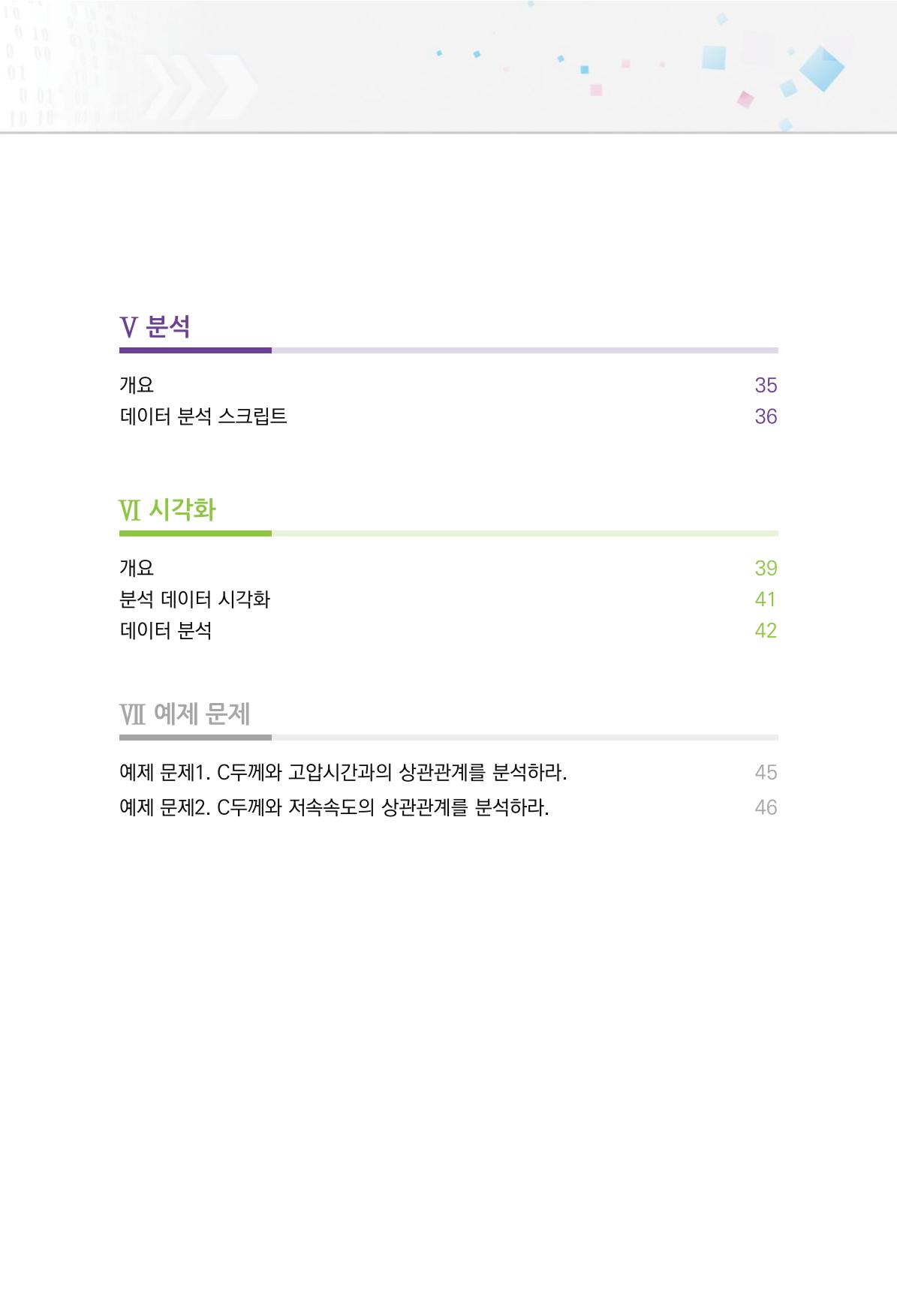
개요	13
교육용 데이터 샘플	14
데이터 수집	15
데이터 작업 영역 이동 스크립트	18

III 가공

개요	23
데이터 가공 스크립트	25

IV 저장

개요	31
----	----



V 분석

개요	35
데이터 분석 스크립트	36

VI 시각화

개요	39
분석 데이터 시각화	41
데이터 분석	42

VII 예제 문제

예제 문제1. C두께와 고암시간과의 상관관계를 분석하라.	45
예제 문제2. C두께와 저속속도의 상관관계를 분석하라.	46

CONTENTS

Intermediate Level 

I 개요

개요	51
----	----

II 수집

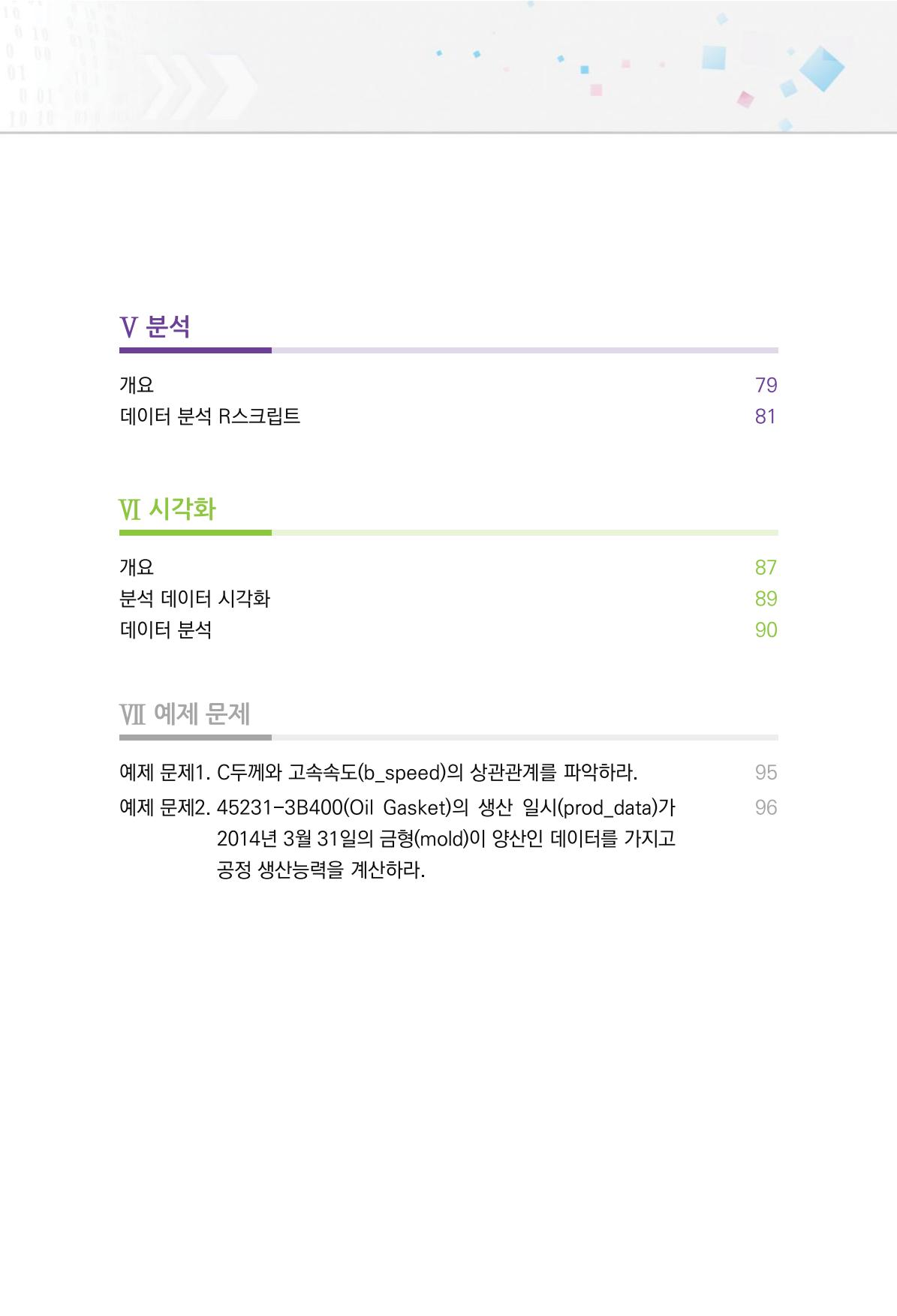
개요	55
교육용 데이터 샘플	56
데이터 수집	57
데이터 작업 영역 이동 스크립트	60

III 가공

개요	65
데이터 가공 스크립트	67

IV 저장

개요	75
----	----



V 분석

개요	79
데이터 분석 R스크립트	81

VI 시각화

개요	87
분석 데이터 시각화	89
데이터 분석	90

VII 예제 문제

예제 문제1. C두께와 고속속도(b_speed)의 상관관계를 파악하라.	95
예제 문제2. 45231-3B400(Oil Gasket)의 생산 일시(prod_data)가 2014년 3월 31일의 금형(mold)이 양산인 데이터를 가지고 공정 생산능력을 계산하라.	96



제조 

Beginning Level

초급과정







I 개요

개요

9

8

I

개요

> 개요

자동차 부품 연구원에서 제공해 준 자동차 부품의 생산 데이터를 바탕으로 제품번호가 ‘45231-3B400(Oil Gasket)’인 양산 데이터를 추출하여 측정값의 C두께(Oil gasket의 탕구 두께)의 값의 히스토그램을 작성하여 C두께의 중앙값, 최댓값, 최솟값을 시각화하여 제품 측정 데이터의 분포 구간을 알아 보고자 한다. 이러한 방법으로 분포 구간에 대한 분석을 통하여 제품에 대한 불량품과 불량률을 알아볼 수 있다.

> 활용 데이터

- **autoparts.csv** : 자동차부품 데이터 셋(2014.3~4)

> 선행학습

- **R 통계** – 히스토그램, csv 파일 Import, Plot 기초
- **자바스크립트** – 객체(내장객체, 브라우저객체), 속성, 변수, 연산자(연산자 우선순위), 제어문, 함수(내장함수, 함수정의) 사용법

> 요구사항

- 자동차 부품 생산 데이터 중 C두께의 히스토그램을 작성하여 C두께의 분포를 형태를 파악한다.
- 측정된 C두께의 최솟값, 중앙값, 최댓값을 구하라.

> 분석 절차

- 수집된 자동차 부품 데이터 셋을 로드한다.
- 수집된 자동차 부품 정보에서 제품번호가 '45231-3B400(Oil Gasket)'인 것을 데이터를 추출한다.
- 제품 생산 일시가 2014-03-31일 데이터만을 추출한다.
- 금형(mold)이 양산, 공정(prod)이 생산인 데이터만을 추출한다.
- 추출한 데이터를 csv 파일로 저장하고 R Studio로 가공된 csv를 불러온다.
- hist() 함수를 사용하여 히스토그램 차트를 출력을 한다.
- 출력 된 히스토그램 차트를 보고 분포 구간에 대한 분석을 통하여 제품에 대한 불량품과 불량률을 알아볼 수 있다.



- **히스토그램** : 데이터가 존재하는 범위를 몇 개의 구간으로 나누어서 각 구간에 들어가는 데이터의 발생 빈도수를 체크하여 막대그래프로 작성한 그림으로 데이터의 분포의 형태를 쉽게 파악하기 위한 용도로 사용한다.
 - 도수: 각각의 구간에 속하는 데이터의 수
 - 구간의 폭: 기둥의 굵기
 - 구간의 경계치: 기둥과 기둥의 경계를 이루는 값
 - 구간의 중심치: 기둥의 중앙에 해당하는 값
- **C두께** : Oil gasket의 탕구두께의 측정값이다.
- **탕구두께** : 별사 슬리브와 다이 공동 사이에 발생되는 둥근 모양의 주입 잔여물
- **hist()** 함수 : 데이터 세트를 통하여 히스토그램을 출력하는 함수



II 수집

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



수집

> 개요

제조 데이터는 자동차 부품연구원에서 제공한 자동차 부품 측정 데이터는 생산라인에서 운영 중인 기계에 대한 파라미터 설정값에 대한 데이터이다. 수집 데이터는 품질 기준 및 대상 제품의 품질관리에 적용되어 제품의 불량률 및 생산능력을 파악을 할 수 있다. 또한 불량 제품이 발생되는 기계의 설정값에 대한 분석으로 생산 수율 및 품질 향상을 위한 예측 모델을 만들 수 있다.



> 수집 방법

- **데이터 제공** : 자동차 부품 측정 데이터는 자동차 부품연구원에서 제공하는 데이터를 OpenAPI, 자료수집기(Crawler)를 통하여 데이터를 수집하였고, 실습용 자료는 빅데이터 분석 활용센터에 접속하여 자동차 부품 데이터 셋을 다운로드 할 수 있도록 원시데이터를 제공하고 있다.



용어정리

- **품질관리(Quality Control)** : 수요자의 요구에 맞는 품질의 제품을 경제적으로 만들어 내기 위한 모든 수단의 체계.

▶ 교육용 데이터 샘플

▶ 자동차 부품데이터(autoparts.csv)

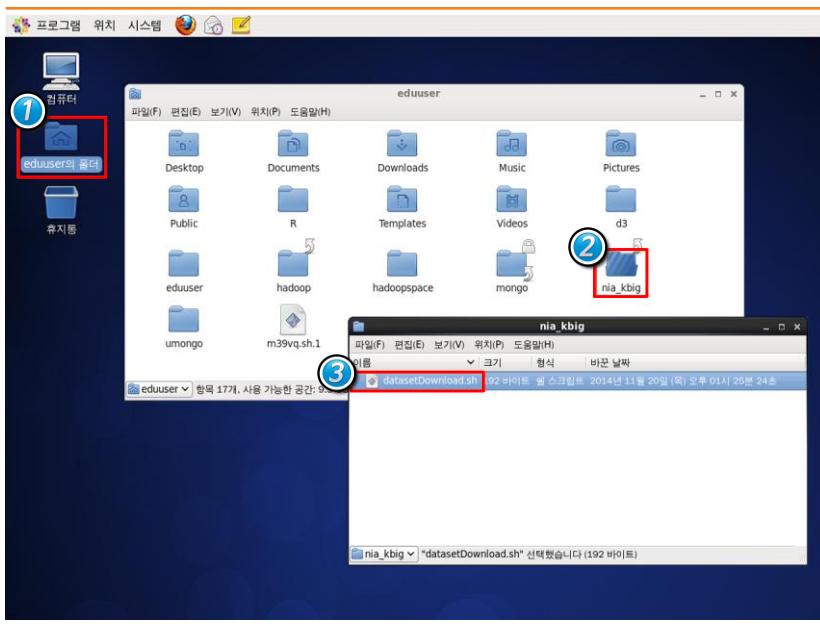
생산일시	제품 번호	제품명	차 수	금형	타	S-NO	고정 시간(sec)	A속도 (m/s)	B속도 (m/s)	이격 (mm)	S이격 (mm)	조간율 (%)	실압력 (MPa)	하중 시간(sec)	고압 시간(ms)	C두께 (mm)
2014-04-01 오전 8:29:12	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 339	80.9	0.778	1.5	171.3	725	82	73.8	20.2	71	25.2
2014-04-01 오전 8:27:51	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 338	80.9	0.782	1.513	173.2	725.6	82	73.8	20.2	72	22.7
2014-04-01 오전 8:26:29	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 337	80.9	0.788	1.511	172.6	725.4	81	73.9	20.2	76	23.5
2014-04-01 오전 8:25:08	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 336	80.9	0.779	1.508	172.4	726.1	82	73.9	20.2	72	23
2014-04-01 오전 8:23:48	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 335	80.9	0.776	1.506	173.9	725.9	82	73.8	20.2	73	21.7
2014-04-01 오전 8:22:27	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 334	80.9	0.779	1.487	174.8	725.1	82	73.8	20.2	73	21.6
2014-04-01 오전 8:21:05	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 333	80.9	0.789	1.509	171.3	726.1	81	73.7	20.2	76	24.1
2014-04-01 오전 8:19:46	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 332	80.9	0.78	1.502	172.2	724.8	82	73.9	20.2	73	24.5
2014-04-01 오전 8:18:26	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 331	80.9	0.77	1.51	171.7	724.9	82	73.8	20.2	65	24.9
2014-04-01 오전 8:17:05	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 330	80.9	0.765	1.502	173.9	724.5	81	73.7	20.2	75	23.1
2014-04-01 오전 8:15:43	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 329	80.9	0.779	1.5	173.8	725.3	81	73.8	20.2	73	22.4
2014-04-01 오전 8:14:23	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 328	80.9	0.79	1.489	172.4	726.3	82	73.8	20.2	73	22.8
2014-04-01 오전 8:13:02	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 327	80.9	0.774	1.492	172.2	725.2	82	73.8	20.2	70	24.1
2014-04-01 오전 8:11:40	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 326	80.9	0.795	1.505	173	727.1	82	73.9	20.2	70	21.4

II. 수집

▶ 데이터 수집(datasetDownload.sh)

- 데이터 저장소에서 서버 로컬로 일반 자동차부품 생산 데이터를 복사해 온다.
 - **autoparts.csv** : 자동차부품 생산 데이터

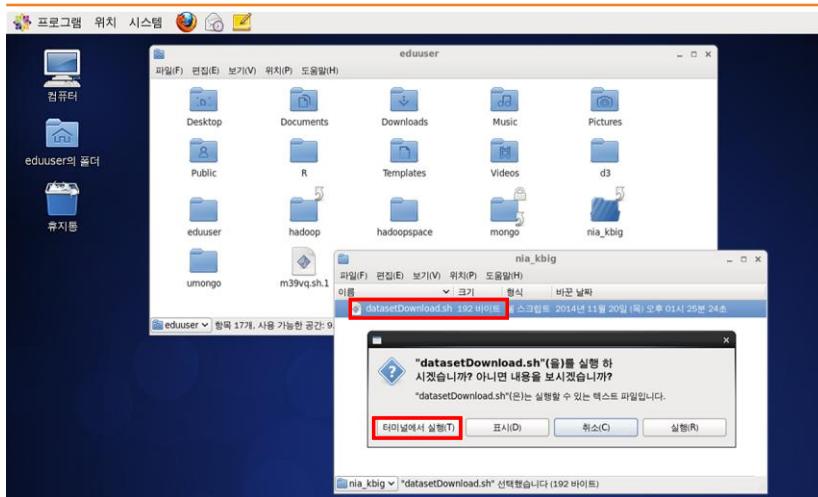
▶ 실습코드 디렉토리로 이동



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

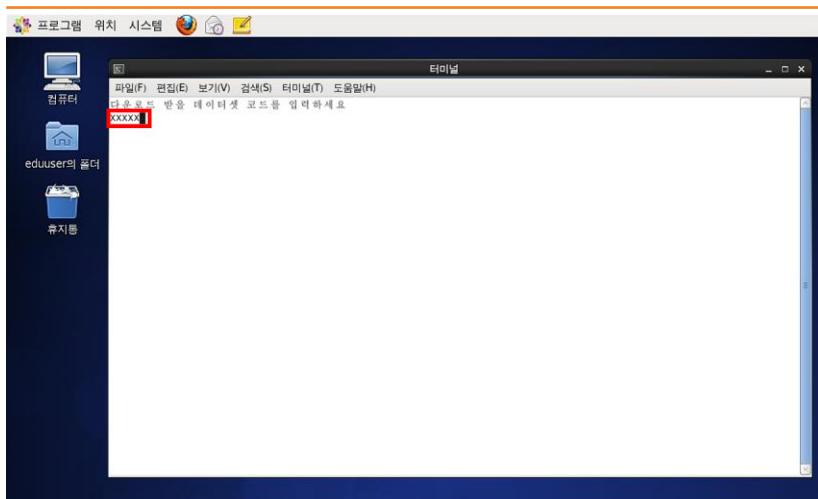
▶ 레파지토리에서 데이터 수집

datasetDownload.sh (원시데이터로 컬서버로 복사)



- '터미널에서 실행' 버튼을 클릭한다.

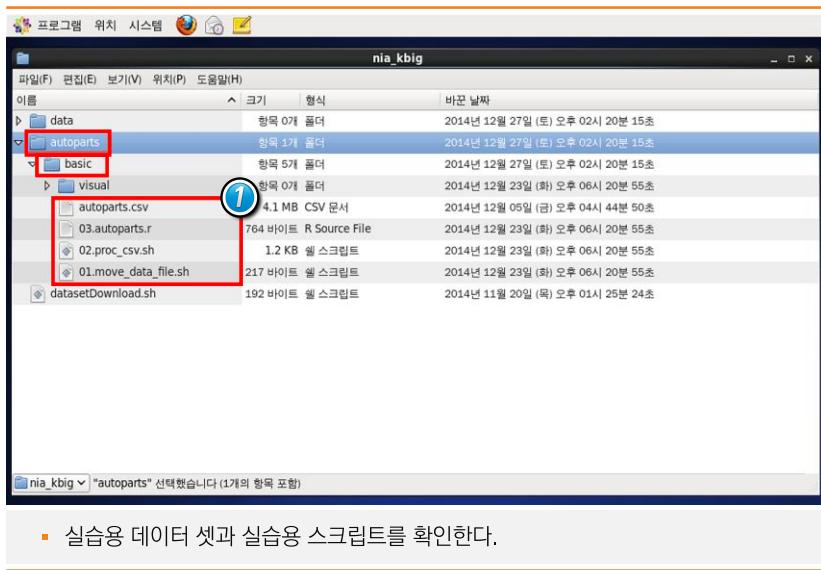
▶ 데이터셋 코드 입력



- 다운로드 받을 데이터셋 코드를 입력 후 엔터

II. 수집

▶ 데이터셋과 실습용 쉘 스크립트



▶ ① 데이터 및 스크립트

▪ 01.move_data_file.sh :

작업영역 Data 폴더로 자료 이동하는 스크립트

▪ 02.proc_csv.sh :

원시데이터에서 분석할 대상을 추출하여 저장하는 스크립트

▪ 03.autoparts.r :

R분석 스크립트

▪ datasetDownload :

레파지토리에서 분석데이터와 실습용 스크립트를 다운로드 스크립트

▪ autoparts.csv :

자동차부품 생산 데이터

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

▶ 데이터 작업 공간으로 이동

- 로컬로 수집해온 데이터를 작업 영역 Data 폴더로 자료를 이동하는 스크립트

01.move_data_file.sh (작업영역 폴더로 원시데이터 이동)

```

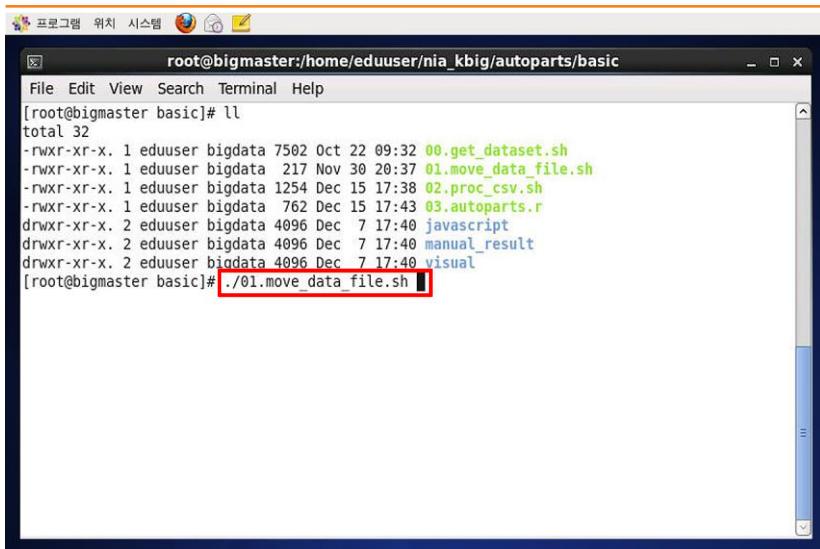
01.#!/bin/bash
02.
03. # 복사 대상 파일 정의
04. # 자동차 부품 데이터
05. TARGET_AUTOPARTS=/home/eduuser/nia_kbig/autoparts/basic/autoparts.csv
06. # 작업영역 디렉토리 정의
07. LOCAL_DIR=/home/eduuser/nia_kbig/data/
08.
09. # 데이터저장소에서 받은 자료를 작업영역인 /home/eduuser/nia_kbig/data/
10. ↛ 폴더로 자료 이동
11. mv $TARGET_AUTOPARTS $LOCAL_DIR
  
```



- 데이터 작업 영역 이동 스크립트 소스(01.move_data_file.sh)
- 라인 05 : 다운로드한 원시데이터 autoparts.csv 파일을 설정하는 라인이다.
- 라인 07 : 작업폴더를 설정하는 라인이다.
- 라인 10 : 작업폴더로 다운로드한 원시데이터를 이동하는 라인이다.

II. 수집

▶ 수집 데이터 셋 작업 영역 폴더 이동



A screenshot of a terminal window titled "root@bigmaster:/home/eduuser/nia_kbig/autoparts/basic". The window shows a file listing with the command "ll" and a command being typed: "./01.move_data_file.sh". The terminal interface includes a menu bar (File, Edit, View, Search, Terminal, Help) and a toolbar with icons for program, location, system, and search.

```
[root@bigmaster basic]# ll
total 32
-rwxr-xr-x. 1 eduuser bigdata 7502 Oct 22 09:32 00.get_dataset.sh
-rwxr-xr-x. 1 eduuser bigdata 217 Nov 30 20:37 01.move_data_file.sh
-rwxr-xr-x. 1 eduuser bigdata 1254 Dec 15 17:38 02.proc_csv.sh
-rwxr-xr-x. 1 eduuser bigdata 762 Dec 15 17:43 03.autoparts.r
drwxr-xr-x. 2 eduuser bigdata 4096 Dec 7 17:40 javascript
drwxr-xr-x. 2 eduuser bigdata 4096 Dec 7 17:40 manual_result
drwxr-xr-x. 2 eduuser bigdata 4096 Dec 7 17:40 visual
[root@bigmaster basic]# ./01.move_data_file.sh
```

- 로컬에 원시데이터를 작업 영역 폴더로 이동 (/home/eduuser/nia_kbig/data/) 시킨다. ./01.move_data_file.sh 입력 후 엔터

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화







가공

> 개요

작업 영역 폴더에 복사한 자동차 부품 생산데이터 중에서 제품에 대한 품질분석을 하기 위해서 금형 상태가 양산이며, 공정이 생산인 데이터를 추출하여 생산된 제품에 대한 불량률과 생산제품의 산포의 분포를 파악하기 위해서 실제 양산된 제품을 선정한다. C두께의 히스토그램을 구하기 위해서 C두께 값만 추출하여 저장을 한다.

> 가공 방법

- 가공 대상은 45231-3B400(Oil Gasket) 추출 데이터 중 생산일시(prod_data)가 2014-03-31이고, 금형(mold)이 양산, 공정(prod)이 생산 제품만을 추출하여 autoparts.r_data.csv로 저장한다.

> 가공 데이터셋

prod_date	prod_no	prod_name	d e g r e e	mold	prod	s_ no	fix_ time (sec)	a_ speed (m/s)	b_ speed (m/s)	seper ation (mm)	s_sep eration (mm)	rate_ term s(%)	mp a	load_ time (sec)	highp ressu re_t ime (ms)	c_ thick ness (mm)
2014-03-07 오후 10:16:16	45231 - 3B400	Oil Gasket	1	양산	생산	35 00 27	80.9	0.669	1.668	182.6	714.2	85	75.3	19.2	76	27
2014-03-07 오후 10:14:54	45231 - 3B400	Oil Gasket	1	양산	생산	35 00 26	81.2	0.647	1.666	184.3	713.3	86	75.3	19.2	80	26.2
2014-03-07 오후 10:13:33	45231 - 3B400	Oil Gasket	1	양산	생산	35 00 25	81.1	0.662	1.661	184.2	713.7	86	75.2	19.2	83	25.9
2014-03-07 오후 10:12:12	45231 - 3B400	Oil Gasket	1	양산	생산	35 00 24	81.1	0.657	1.644	183.3	714	86	75.2	19.2	71	26.5

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화

2014-03-07 오후 10:10:51	45231 - 3B400	Oil Gasket	1	양산	생산	35 00 23	81.1	0.681	1.64	182	715.3	86	75.3	19.2	77	26.5
2014-03-07 오후 10:09:29	45231 - 3B400	Oil Gasket	1	양산	생산	35 00 22	81.4	0.658	1.649	182.8	713.1	87	75.3	19.2	77	27.9
2014-03-07 오후 10:08:09	45231 - 3B400	Oil Gasket	1	양산	생산	35 00 21	81.6	0.647	1.662	184.6	712.8	86	75.2	19.2	75	26.4
2014-03-07 오후 10:06:48	45231 - 3B400	Oil Gasket	1	양산	생산	35 00 20	81.6	0.643	1.644	183	712.9	86	75.2	19.2	78	27.9
2014-03-07 오후 10:05:27	45231 - 3B400	Oil Gasket	1	양산	생산	35 00 19	80.8	0.676	1.647	181.8	713.9	86	75.2	19.2	81	28.1
2014-03-07 오후 10:04:05	45231 - 3B400	Oil Gasket	1	양산	생산	35 00 18	80.9	0.658	1.645	182.7	713.3	85	75.2	19.2	77	27.8
2014-03-07 오후 10:02:44	45231 - 3B400	Oil Gasket	1	양산	생산	35 00 17	81.4	0.663	1.649	183.7	713.2	85	75.2	19.2	78	26.9
2014-03-07 오후 10:01:23	45231 - 3B400	Oil Gasket	1	양산	생산	35 00 16	81.4	0.645	1.656	184.3	713.1	85	75.2	19.2	80	26.4
2014-03-07 오후 10:00:01	45231 - 3B400	Oil Gasket	1	양산	생산	35 00 15	81.1	0.656	1.641	183	713.3	86	75.2	19.2	74	27.5
2014-03-07 오후 9:58:41	45231 - 3B400	Oil Gasket	1	양산	생산	35 00 14	80.8	0.654	1.642	183.1	713.1	86	75.2	19.2	78	27.6

III. 가공

▶ 데이터 가공 스크립트(02.proc_csv.sh)

- 셀 스크립트를 이용하여 생산일시(prod_date):2014-03-31, 금형(mold) : 양산, 공정(prod) : 생산, 특정제품(45231-3B400) 데이터만을 추출한다. 추출한 데이터는 autoparts.r_data.csv로 저장한다.

02.proc_csv.sh (원시데이터에서 분석할 대상을 추출 하여 저장)

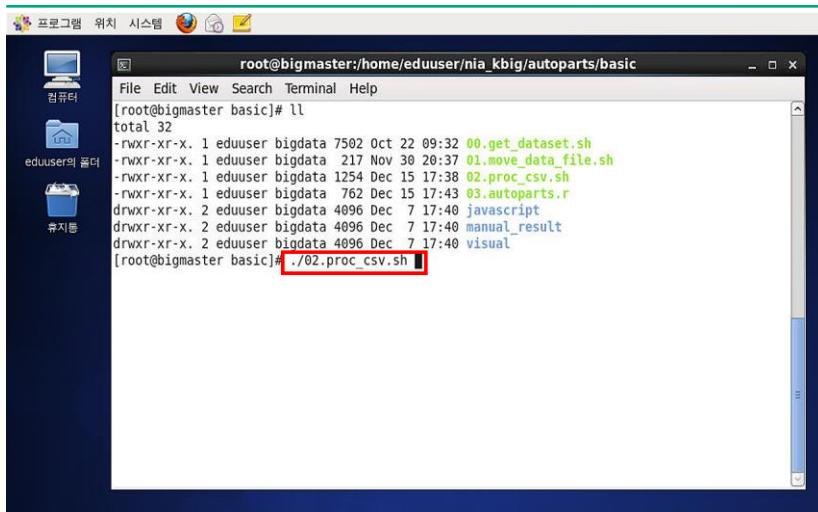
```
01. #!/bin/bash
02.
03. # 입력 CSV 파일 지정
04. INPUT_FILE='/home/eduuser/nia_kbig/data/autoparts.csv'
05. # 출력결과 CSV 파일 지정
06. RESULT_FILE='/home/eduuser/nia_kbig/data/autoparts.r_data.csv'
07.
08. # HEADER컬럼 출력
09. echo "c_thickness" > $RESULT_FILE
10.
11. # ','를 구분자로 해서 파일을 읽어들인다.
12. IFS=','
13. while read prod_date prod_no prod_name degree mold prod s_no fix_time a_s
   ↵peed b_speed seperation s_seperation rate_terms mpa load_time highpres
   ↵ure_time c_thickness
14. do
15. # 금형이 '양산' 인것을 대상으로 한다.
16. if [ $mold != '양산' ]; then
17.     continue;
18. fi
19. # 특정 제품만을 대상으로 한다.
20. if [ $prod_no != '45231-3B400' ]; then
21.     continue;
22. fi
23. # c두께가 데이터범주이내의 것을 대상으로 한다. 21.0~30.0 까지
```



- 데이터 가공 스크립트 소스(02.proc_csv.sh)
- 라인 04~06 : 가공 대상인 자동차 부품 생산 데이터(autoparts.csv) 지정을 하고, 가공데이터를 autoparts.r_data.csv 파일로 저장하는 라인이다.
- 라인 09 : 가공데이터 파일을 생성시 상단에 Header 정보를 출력하는 라인이다.
- 라인 12~22 : 자동차부품생산 데이터 파일을 1라인씩 읽어서 금형(mold)가 양산이고, 제품인 '45231-3B400' 인 제품만을 대상으로 데이터를 분리하는 라인이다.
- 라인 24~34 : 자동차 부품 생산 데이터(autoparts.csv) 중에 C 두께가 30 이상과 21 미만의 데이터를 걸러내는 라인이다.
- 라인 36 : 가공된 데이터를 파일로 저장하는 라인이다.

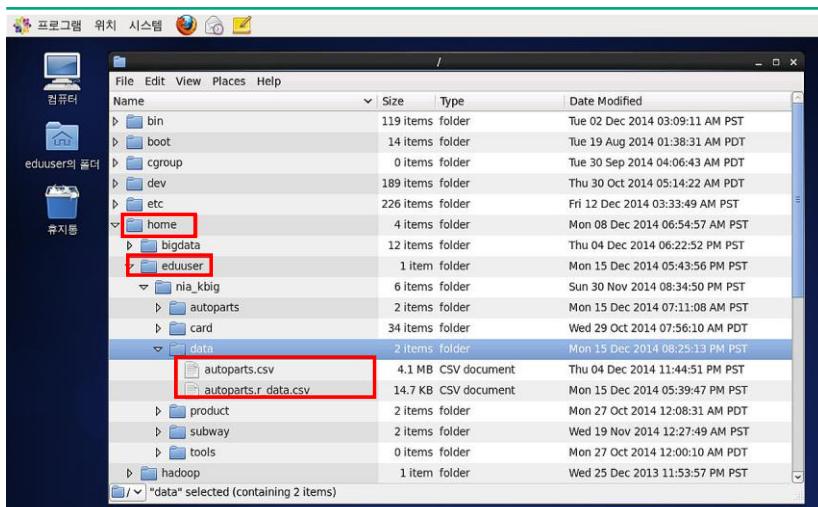
III. 가공

▶ 원시데이터에서 분석 대상 데이터 가공



- 원시 데이터 셋에서 분석할 데이터를 가공하여 autoparts.r_data.csv 파일을 생성한다.
.02.proc_csv.sh 입력 후 엔터

▶ 가공 데이터 작업 영역 폴더에 생성



- /home/eduuser/nia_kbig/data 폴더에 autoparts.r_data.csv 파일이 생성된다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화



IV 저 장

개요

31

IV

저장

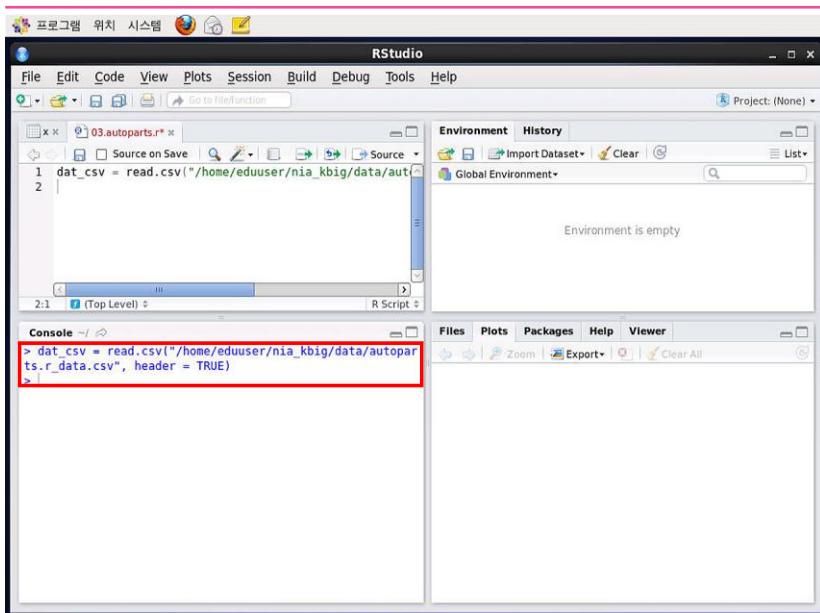
> 개요

자동차 부품 생산 데이터에서 분석 대상인 45231-3B400(Oil Gasket) 제품 중에서 생산일시(prod_data)가 2014-03-31이고 금형(mold)이 양산, 공정(prod)이 생산인 데이터를 추출하여 파일로 저장을 하였다. R Studio를 활용하여 가공 데이터를 불러와서 메모리에 저장을 시킨다. R에서는 텍스트, CSV, 엑셀 데이터 등 다양한 형태의 데이터를 읽어 올 수 있다.

> 저장 방법

- proc_csv.sh 셀에서 저장한 autoparts.r_data.csv 파일을 R Studio를 활용하여 가공된 데이터를 메모리상에 탑재하여 분석할 수 있게 한다.
- **가공데이터** : /home/eduuser/nia_kbig/data/autoparts.r_data.csv
- autoparts<-read.csv(/home/eduuser/nia_kbig/data/auto parts.r_data.csv)

▶ R Studio에 분석 데이터 메모리에 저장



- 가공된 데이터를 R Studio에서 메모리에 분석 데이터를 저장한다.

W



V 분석

개요

35

데이터 분석 스크립트

36

V 분석

> 개요

제조 데이터의 분석은 R Studio에서 가공된 데이터를 불러와서 R 스크립트를 실행을 하여 C두께(탕구 두께)에 대한 히스토그램 차트를 출력한다. R에서 히스토그램을 그리는 함수는 `hist()` 함수이고 여러 가지 옵션을 설정할 수 있다. 분포 구간에 대한 분석을 통하여 제품에 대한 불량품과 불량률을 찾을 수 있다.

> 분석 방법

- R로 `autoparts.r_data.csv` 파일을 로드해서 C두께 히스토그램을 그려서 측정값의 분포를 살펴보고, C두께의 최솟값, 중앙값, 최댓값을 구하여 차트를 생성한다.
- `hist()` 함수를 사용하여 히스토그램을 출력한다.
- C두께의 최솟값, 중앙값, 최댓값을 구한다.

> 가공 데이터 샘플(`autoparts.r_data.csv`)

- `c_thickness`

25.6	26.2	26.5	27.8
25.9	26.2	26.9	27.9
25.9	26.4	27	27.9
25.9	26.4	27.5	28.1
26	26.5	27.6	...

> 데이터 분석 스크립트(03.autoparts.r)

> R 분석 스크립트

- 분석 결과(CSV) 데이터 파일을 R Studio를 이용하여 로드 후 히스토그램 차트와 바 plot 차트로 비교분석한다.

03.autoparts.r (R분석스크립트)

```

01. dat_csv = read.csv("/home/eduuser/nia_kbig/data/autoparts.r_data.csv", hea
    ↪ der = TRUE)
02. par(mfrow=c(1,2))
03. #히스토그램구하기
04. hist(dat_csv$c_thickness, main="C두께 히스토그램", breaks=40 , xlim=c(20,30)
    ↪ , xlab="C두께")
05. #최대,최소,평균값 구하기
06. c_thick_mean<-mean(dat_csv$c_thickness)
07. c_thick_max<-max(dat_csv$c_thickness)
08. c_thick_min<-min(dat_csv$c_thickness)
09. #데이터 프레임 만들기
10. x<-data.frame(최대값=c(c_thick_max), 중간값=c(c_thick_mean), 최소값=c(c_thick_
    ↪ min))
11. #바차트 만들기
12. title<-"C두께"
13. bp <- barplot(as.matrix(x), beside=TRUE , main=title , col=c("green") , ylim=c(0,32))
14. # bar에 라벨을 붙인다.
15. text(x=bp, y=c(x[1],x[2],x[3]), labels=c(x[1],x[2],x[3])) , pos=1 , cex=1.0)

```



• R분석 스크립트 소스(03.autoparts.r)

- 라인 01 : 가공된 자동차부품생산 데이터(autoparts.r_data.csv) 파일을 로드하는 라인이다.
- 라인 04 : hist() 함수를 사용하여 히스토그램 차트를 설정하는 라인이다.
- 라인 06~08 : 최대,최소,평균값을 구하는 라인이다.
- 라인 10 : 데이터 프레임을 만들어 최대값, 중간값, 최소값을 지정하는 라인이다.
- 라인 12~15 : barplot() 함수를 이용하여 바차트 만들기 위한 설정을 하는 라인이다.



1

2



VI 시각화

개요	39
분석 데이터 시각화	41
데이터 분석	42

VI

시각화

> 개요

자동차 부품 생산 데이터의 분석과정에서 가공한 데이터를 가지고 히스토그램 차트를 출력하는 R 스크립트(03.autoparts.r)를 R Studio로 불러와서 실행한다. 히스토그램 결과 차트를 보고 데이터의 분포 양상, 데이터의 산포의 크기, 데이터의 중심이 어디에 위치해 있는지 데이터의 전체적인 모습을 파악한다. R Studio를 활용하면 분석과 함께 시각화 까지 일괄적으로 진행할 수 있다.

> 시각화 방법 및 활용기술

- 분석 결과(CSV) 데이터 파일을 R Studio를 이용하여 로드 후 히스토그램 차트와 바 plot 차트로 시각화한다.
- 히스토그램 차트로 측정값의 분포와 산포를 확인한다.
- 바 plot 차트로 관리범위 내의 C두께의 최댓값, 중간값, 최솟값을 확인한다.

> 데이터 변환

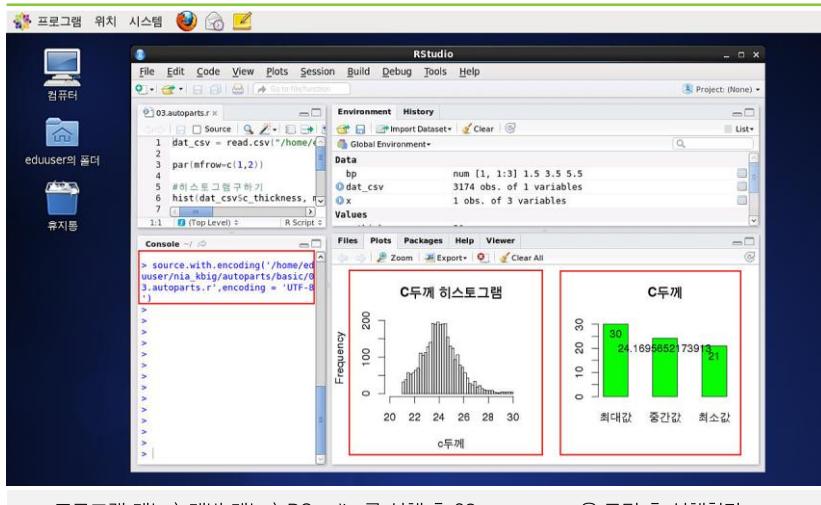


> 시각화 과정

- R Studio를 실행하여 CSV 파일을 로드한다.
- 분석 스크립트를 R 스크립트를 만들어서 실행한다.
- R Studio로 autoparts.r 파일을 로드하여 실행한다.

▶ 분석 데이터 시각화

▶ R Studio 실행



- 프로그램 메뉴 > 개발 메뉴 > RStudio 를 실행 후 03.autoparts.r을 로딩 후 실행한다.

▶ 히스토그램 해석



▪ **일반형** – 일반적으로 나타나는 모양의 뜻수는 중심근처에 가장 많고 중심에서 멀어짐에 따라 서서히 적어지면서 거의 좌우대칭을 이룬다. 공정이 안정된 경우에 나타난다.



▪ **이빠진형** – 계급의 폭을 정수배로 했는지 또는 측정법, 데이터의 맷음법에 버릇이 있는 경우에 나타나는 모양이다.



▪ **비뚤어진형** – 이론적으로 또는 규격치 등으로 하한이 억제되어 있고 어떤값 이하의 값을 취하지 않는 경우에 나타나는 모양이다.



▪ **절벽형** – 규격이하의 것을 전수 선별하여 제거했을 경우, 측정의 속임수, 검사, 측정미스, 오차등에 의해 나타나는 모양이다. 전수선별에 의해 이러한 모양이 되었을 때는 공정능력을 높이든가, 규격의 재검토가 필요하다.



▪ **고원형** – 평균치가 다소 다른 몇 개의 분포가 혼합한 경우에 나타나는 모양이다.



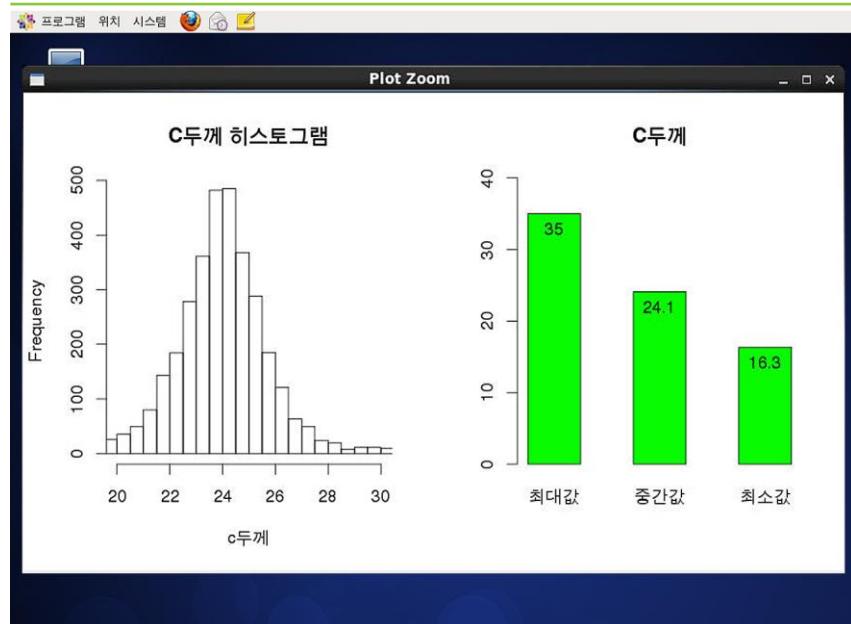
▪ **쌍봉우리형** – 평균치가 다른 2개의 분포가 혼합되어 있는 경우에 나타나는 모양. 예를 들면 두 대의 기계사이, 두 개의 원료의 사이 등 어떤 두 집단의 차이로 발생 가능성이 높다.



▪ **낙도형** – 공정이상, 측정의 오류, 다른 공정의 데이터가 들어있는 등 다른 분포에서 데이터가 약간 혼입한 경우 나타나는 모양이다.

▶ 데이터 분석

▶ 제품번호 45231-3B400(Oil Gasket)인 양산 제품 히스토그램 분석



- 히스토그램 분포가 약간 좌측으로 치우치는 정규분포 현상을 보인다.
- 스펙 범위가 24 ± 3 범위이기 때문에 21~27nm 사이를 유효값으로 처리한다.
- 21 미만과 27.1 이상은 불량품으로 볼 수 있다.
- 최댓값은 35, 중앙값은 24.1, 최솟값은 16.3으로 나왔다.



VII 예제문제

예제 문제1

45

예제 문제2

46

예 / 제 / 문 / 제

예제 1

C두께와 고압시간(hightpressure_time)과의 상관관계를 분석하라.

- C두께(탕구두께)와 제품을 찍을 때의 압력 시간이 두께와의 영향을 미치는지 파악하라.

- 45231-3B400의 생산일시(prod_date)가 2014-03-31, 금형(mold)0 양산이 데이터를 추출한다.
- R Studio에 추출된 데이터를 로딩한다.
- C두께와 압력 시간의 회귀분석을 한다.
- C두께와 압력 시간에 대한 선형회귀분석 그래프를 출력한다.

예제 2

C두께와 저속속도(a_speed)의 상관관계를 분석하라.

- C두께(탕구두께)와 저속속도(a_speed)가 제품을 생산하는데 영향 요인이 있는지 산점도를 출력하여 연관관계를 분석하라.

- 45231-3B400의 생산일시(prod_date)가 2014-03-31, 금형(mold)이 양산이 데이터를 추출한다.
- R Studio에 추출된 데이터를 로딩한다.
- C두께(탕구두께)와 저속속도(a_speed)의 데이터를 가지고 산점도를 출력한다.



제조 

Intermediate Level

중급과정







I 개요

개요

51

I

개요

> 개요

자동차 부품 연구원에서 제공해 준 자동차 부품의 생산 데이터를 바탕으로 자동차 부품 생산 데이터의 C두께(탕구 두께) 데이터를 대상으로 시간대별 측정값을 샘플링하여 Xbar-R 관리도를 구현하여 관리 범위 내의 벗어나는 구간을 찾고, 산점도 차트를 이용하여 C두께에 영향이 있는 설정값을 분석해 보고자 한다.

> 활용 데이터

- **autoparts.csv** : 자동차부품 데이터 셋(2014.3~4)

> 선행학습

- **R 통계** – qcc 패키지, Xbar, R 관리도, Plot차트, csv Import
- **리눅스** – 파일시스템 구조, 쉘 명령어, 쉘 스크립트 실행 방법



- **qcc 패키지** : 품질관리의 관리도를 그리기 위한 통계적 품질관리 패키지
- **Xbar** : 표본의 평균
- **R (Range)** : 최댓값–최솟값으로 0에 가까울수록 좋음.
- **XBar-R관리도** : 품질관리의 한 수법으로, 평균치의 변화를 관리하는 xbar 관리와 편차의 변화를 관리하는 R관리도를 조합한 것으로 제품 생산공정에서 일정기간 마다 관리도에 기입해서 관리 상태를 파악하고, 관리 한계를 벗어난 때는 비정상으로 판단해서 조치한다.
- **산점도(scatter plot)** : 두 개 이상 변수의 동시 분포에서 각 개체를 점으로 표시한 그림으로 두 변수의 관계를 시각적으로 검토할 때 유용하며, 변수들 사이의 관계를 왜곡 시키는 특이점을 확인하는 경우에 유용하다.

▶ 요구사항

- 자동차 부품 정보에서 제품번호가 ‘45231-3B400(Oil Gasket)’인 데이터 중 ‘양산’ 데이터를 추출하여 Xbar-R 관리도를 작성하고 C두께와 연계하여 설정된 파라미터들의 측정값과의 산점도를 그래프로 출력하여 연관성을 파악하라.

▶ 분석 절차

- 수집된 자동차 부품 데이터 세트을 로드한다.
- 제공된 자동차 부품 정보에서 양산 및 제품번호가 ‘45231-3B400 45231-3B400(Oil Gasket)’인 것을 추출한다.
- 제품 생산 일시가 2014-03-31일 데이터 만을 추출한다.
- 금형(mold)이 ‘양산’, 공정(prod)이 ‘생산’ 인 데이터 만을 추출한다.
- 2014-03-31 일자의 시간대 별 샘플 5개를 추출하여 24시간 기준으로 데이터를 분류 저장한다.
- 샘플링 데이터를 가지고 Xbar 및 R 관리도를 작성하고 Xbar 관리범위 (24 ± 2)를 벗어나는 불량제품 발생 시간을 찾는다.
- C두께와 다른 항목 간의 연관성 분석을 위해서 산점도를 출력하여 분석한다.



II 수집

개요	55
교육용 데이터 샘플	56
데이터 수집	57
데이터 작업 영역 이동 스크립트	60



수집

▶ 개요

제조 데이터는 자동차 부품연구원에서 제공한 자동차 부품 측정 데이터는 생산라인에서 운영중인 기계에 대한 파라미터 설정값에 대한 데이터이다. 수집데이터는 품질 기준 및 대상 제품의 품질관리에 적용되어 제품의 불량율 및 생산능력을 파악할 수 있다. 또한 불량 제품이 발생되는 기계의 설정값에 대한 분석으로 생산 수율 및 품질 향상을 위한 예측 모델을 만들 수 있다.

▶ 수집 방법

- **데이터 제공** : 자동차 부품 측정 데이터는 자동차 부품연구원에서 제공하는 데이터를 OpenAPI, 자료수집기(Crawler)를 통하여 데이터를 수집하였고, 실습용 자료는 빅데이터 분석 활용센터에 접속하여 자동차 부품 데이터 셋을 다운로드 할 수 있도록 원시데이터를 제공하고 있다.

▶ 교육용 데이터 샘플

▶ 자동차 부품데이터(autoparts.csv)

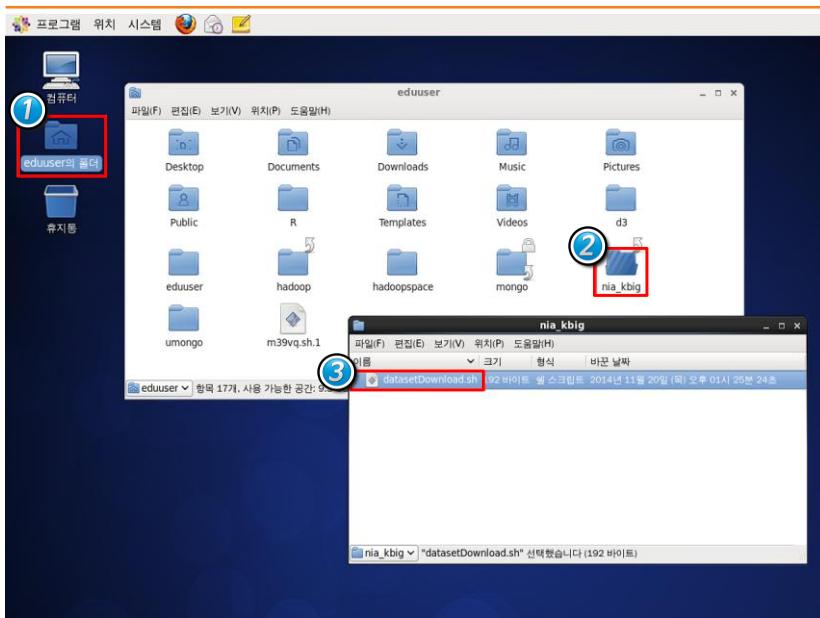
생산일시	제품 번호	제품명	차 수	금형	타	S-NO	고정 시간(sec)	A속도 (m/s)	B속도 (m/s)	이격 (mm)	S이격 (mm)	조간율 (%)	실압력 (MPa)	하중 시간(sec)	고압 시간(ms)	C두께 (mm)
2014-04-01 오전 8:29:12	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 339	80.9	0.778	1.5	171.3	725	82	73.8	20.2	71	25.2
2014-04-01 오전 8:27:51	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 338	80.9	0.782	1.513	173.2	725.6	82	73.8	20.2	72	22.7
2014-04-01 오전 8:26:29	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 337	80.9	0.788	1.511	172.6	725.4	81	73.9	20.2	76	23.5
2014-04-01 오전 8:25:08	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 336	80.9	0.779	1.508	172.4	726.1	82	73.9	20.2	72	23
2014-04-01 오전 8:23:48	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 335	80.9	0.776	1.506	173.9	725.9	82	73.8	20.2	73	21.7
2014-04-01 오전 8:22:27	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 334	80.9	0.779	1.487	174.8	725.1	82	73.8	20.2	73	21.6
2014-04-01 오전 8:21:05	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 333	80.9	0.789	1.509	171.3	726.1	81	73.7	20.2	76	24.1
2014-04-01 오전 8:19:46	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 332	80.9	0.78	1.502	172.2	724.8	82	73.9	20.2	73	24.5
2014-04-01 오전 8:18:26	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 331	80.9	0.77	1.51	171.7	724.9	82	73.8	20.2	65	24.9
2014-04-01 오전 8:17:05	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 330	80.9	0.765	1.502	173.9	724.5	81	73.7	20.2	75	23.1
2014-04-01 오전 8:15:43	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 329	80.9	0.779	1.5	173.8	725.3	81	73.8	20.2	73	22.4
2014-04-01 오전 8:14:23	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 328	80.9	0.79	1.489	172.4	726.3	82	73.8	20.2	73	22.8
2014-04-01 오전 8:13:02	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 327	80.9	0.774	1.492	172.2	725.2	82	73.8	20.2	70	24.1
2014-04-01 오전 8:11:40	90784 - 76001	Oil Gasket	1	승인 대기	생산	369 326	80.9	0.795	1.505	173	727.1	82	73.9	20.2	70	21.4

II. 수집

▶ 데이터 수집(datasetDownload.sh)

- 데이터 저장소에서 서버 로컬로 일반 자동차부품 생산 데이터를 복사해 온다.
 - autoparts.csv** : 자동차부품 생산 데이터

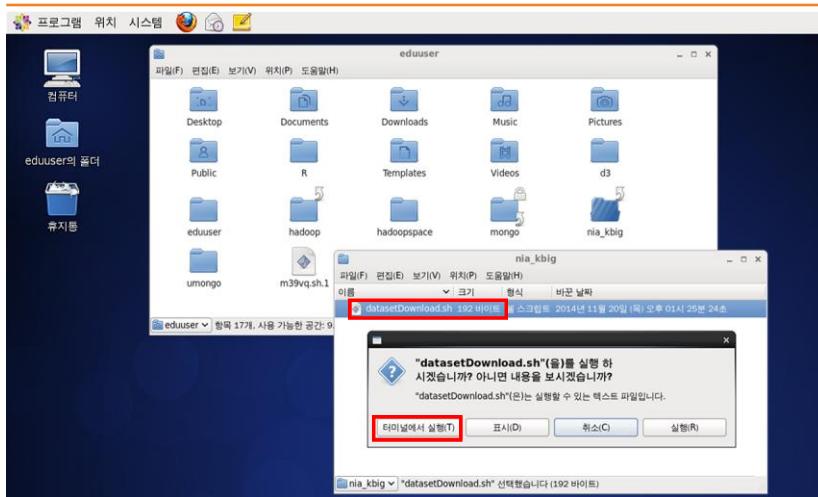
▶ 실습코드 디렉토리로 이동



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

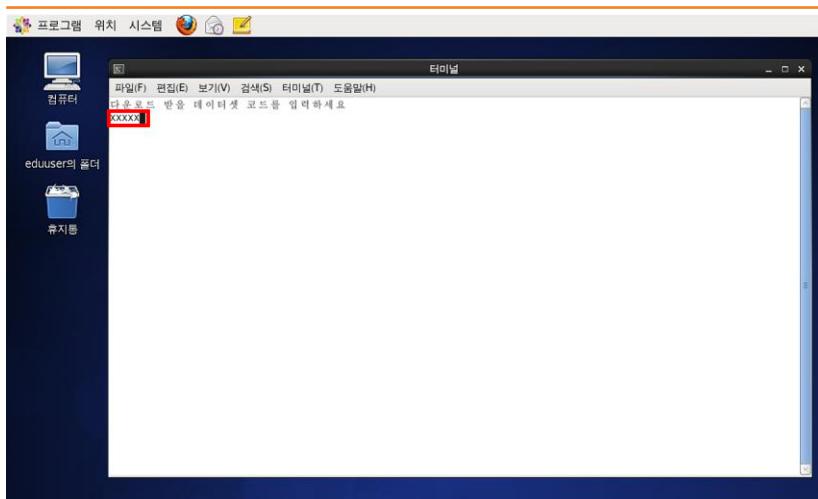
▶ 레파지토리에서 데이터 수집

datasetDownload.sh (원시데이터로 컬서버로 복사)



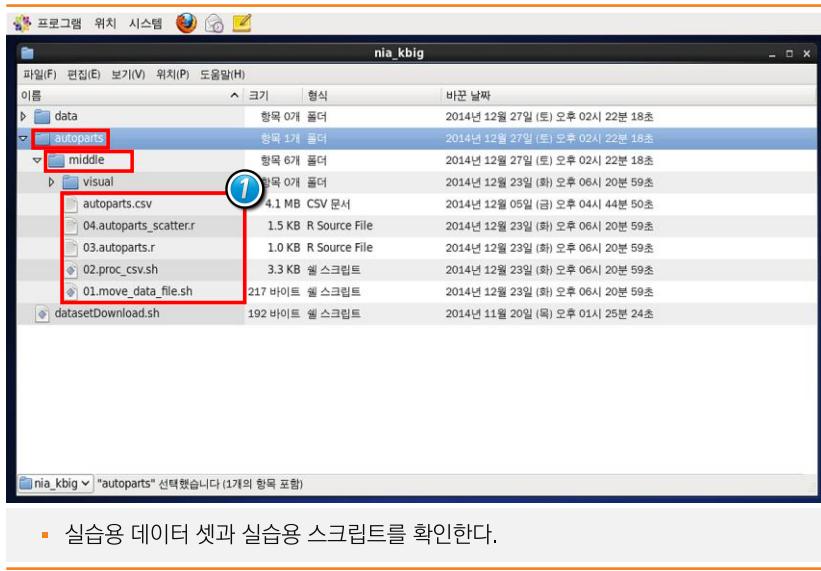
- ▶ '터미널에서 실행' 버튼을 클릭한다.

▶ 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 :

작업영역 Data 폴더로 자료 이동하는 스크립트

■ 02.proc_csv.sh :

원시데이터에서 분석할 대상을 추출하여 저장하는 스크립트

■ 03.autoparts.r :

Xbar-R 관리도 출력 R분석 스크립트

■ 04.autoparts_scatter.r :

산점도 출력 R분석 스크립트

■ datasetDownload :

레파지토리에서 분석데이터와 실습용 스크립트를 다운로드하는 스크립트

■ autoparts.csv :

자동차부품 생산 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 (작업영역 폴더로 원시데이터 이동)

```

01.#!/bin/bash
02. # 복사 대상 파일 정의
03. #자동차 부품 데이터
04. TARGET_AUTOPARTS=/home/eduuser/nia_kbig/autoparts/middle/autoparts
05. ↳ .csv
06. # 작업영역 디렉토리 정의
07. LOCAL_DIR=/home/eduuser/nia_kbig/data/
08. mv $TARGET_AUTOPARTS $LOCAL_DIR

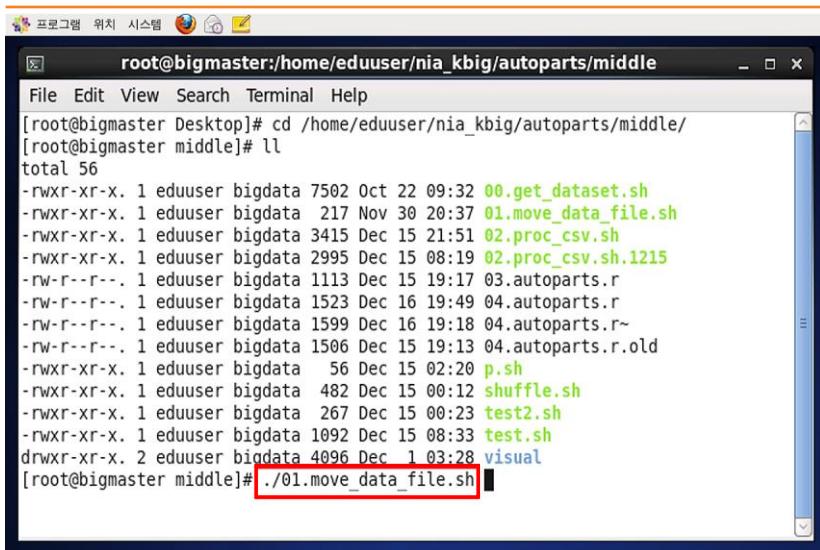
```



- 데이터 작업 영역 이동 스크립트 소스(01.move_data_file.sh)
- 라인 04 : 다운로드한 원시데이터 autoparts.csv 파일을 설정하는 라인이다.
- 라인 06 : 작업영역 폴더를 지정하는 라인이다.
- 라인 07 : 작업 폴더로 다운로드한 원시데이터를 이동하는 라인이다.

II. 수집

▶ 수집 데이터 셋 작업 영역 폴더 이동



A screenshot of a terminal window titled "root@bigmaster:/home/eduuser/nia_kbig/autoparts/middle". The window shows a file listing with the command "ll" and then executes the command "./01.move_data_file.sh". The terminal interface includes a menu bar (File, Edit, View, Search, Terminal, Help) and a toolbar with icons for program, location, system, and help.

```
[root@bigmaster Desktop]# cd /home/eduuser/nia_kbig/autoparts/middle/  
[root@bigmaster middle]# ll  
total 56  
-rwxr-xr-x. 1 eduuser bigdata 7502 Oct 22 09:32 00.get_dataset.sh  
-rwxr-xr-x. 1 eduuser bigdata 217 Nov 30 20:37 01.move_data_file.sh  
-rwxr-xr-x. 1 eduuser bigdata 3415 Dec 15 21:51 02.proc_csv.sh  
-rwxr-xr-x. 1 eduuser bigdata 2995 Dec 15 08:19 02.proc_csv.sh.1215  
-rw-r--r--. 1 eduuser bigdata 1113 Dec 15 19:17 03.autoparts.r  
-rw-r--r--. 1 eduuser bigdata 1523 Dec 16 19:49 04.autoparts.r  
-rw-r--r--. 1 eduuser bigdata 1599 Dec 16 19:18 04.autoparts.r~  
-rw-r--r--. 1 eduuser bigdata 1506 Dec 15 19:13 04.autoparts.r.old  
-rwxr-xr-x. 1 eduuser bigdata 56 Dec 15 02:20 p.sh  
-rwxr-xr-x. 1 eduuser bigdata 482 Dec 15 00:12 shuffle.sh  
-rwxr-xr-x. 1 eduuser bigdata 267 Dec 15 00:23 test2.sh  
-rwxr-xr-x. 1 eduuser bigdata 1092 Dec 15 08:33 test.sh  
drwxr-xr-x. 2 eduuser bigdata 4096 Dec 1 03:28 visual  
[root@bigmaster middle]# ./01.move_data_file.sh
```

- 로컬에 원시데이터를 작업 영역 폴더로 이동 (/home/eduuser/nia_kbig/data/) 시킨다.
`./01.move_data_file.sh` 입력 후 엔터

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화







가공

> 개요

작업 영역 폴더에 복사한 자동차 부품 생산 데이터 중에서 분석 대상인 45231-3B400(Oil Gasket) 제품 중 생산 일시(prod_data)가 2014-03-31이고, 금형(mold)이 양산, 공정(prod)이 생산인 데이터를 추출하고 시간대 별 샘플 데이터 임의 5개씩 데이터를 선정하여 저장한다. C두께와 산점도를 출력하기 위해서 저속속도(a_speed), 고속속도(b_speed), 조건율(rate_terms), 하중시간(load_time), 고압시간(high_pressuer_time), 탕구두께(C두께)의 데이터만을 추출하여 csv 파일로 저장한다.

> 가공 방법

- 자동차 부품 데이터 중 45231-3B400(Oil Gasket) 데이터만을 추출하여 저장한다.
- 분석 대상은 45231-3B400(Oil Gasket) 추출 데이터 중 생산일시(prod_data)가 2014-03-31이고, 금형(mold)이 양산, 공정(prod)이 생산 제품만을 추출하여 시간별 5개 샘플링 데이터는 autoparts.r_sampling.csv 에 저장하고, 산점도 출력할 데이터는 autoparts.r_scatter.csv 파일에 데이터를 저장한다.

> 가공 데이터셋

x1	x2	x3	x4	x5
20.3	24	23.6	25	23.5
25	22	20.8	21.2	22.8
21.6	24.8	22.7	25.2	25.5
22.6	25.8	23.2	25.5	22.3
24.7	24.9	27	26.2	26.9
24.1	24.1	19.4	25.3	23.6
23.5	18.8	20	20.2	26.1
22.8	25.5	27	23.8	22.8
23.8	25.2	24.7	24.4	26.3
23.3	24.7	22.8	25.2	23
25	25.9	21.8	25.7	24.3
23.8	24.5	23	26.4	25.5
22.7	23.9	23.9	24.6	24.1
26.1	25.6	24.7	25.1	25.7
24.3	24.2	24.5	23.4	21.8
24.5	22	23.4	23.5	23.1
20.1	33.6	23.1	31.5	24.3
35.7	22.6	25.3	25.9	24.7
23.1	25.2	23.6	25.2	25.2
23.4	23.2	24.6	24.6	23.5
22.5	23.4	24.3	25	23.3
34	23.1	25.4	24	24.4
25.3	21.5	22.5	23.8	23.6
24.5	25.5	21.3	24.4	21.4

III. 가공

▶ 데이터 가공 스크립트(02.proc_csv.sh)

- 셀 스크립트를 이용하여 양산 및 생산 제품 (45231-3B400) 데이터만을 추출 한다. xbar-r 관리도를 분석할 데이터는 autoparts.r_sampling.csv, 산점도 차트를 위한 autuparts.r_scatter.csv로 저장한다.

02.proc_csv.sh (원시데이터에서 분석할 대상을 추출하여 저장)

```
01.#!/bin/bash
02.#
03.#
04.#
05.#
06.#
07.#
08.#
09.#
10.echo "1) reading/processing..."
11.#
12.echo "" > $TEMP_FILE
13.#
14.echo "a_speed,b_speed,rate_terms,load_time,highpressure_time,c_thickness"
15.#
16.IFS=':'
17.while read prod_date prod_no prod_name degree mold prod s_no fix_time a_
18.    speed b_speed seperation s_seperation rate_terms mpa load_time highpr
19.    essure_time c_thickness
20.do
21.#
22.if [ $mold != '양산' ]; then
23.    continue;
24.f
25.#
26.if [ $prod_no != '45231-3B400' ]; then
27.    continue;
```

```

26.     fi
27.     # 일자가 2014-03-31데이터를 대상으로 한다.
28.     new_prod_date=`echo $prod_date | tr -d " -"`
29.     prod_hour=`echo ${prod_date:11:${#prod_date}-1} | rev | cut -c 7- | rev `
30.     new_prod_date=${new_prod_date:0:8}
31.     is_valid=0
32.     if [ $new_prod_date=="20140331" ]; then
33.         is_valid=1
34.     fi
35.     if [ $is_valid -eq 0 ]; then
36.         continue;
37.     fi
38.     c_thickness=$(echo $c_thickness | sed -e 's/\r//g')
39.
40.     # 필드를 샘플링 데이터 임시파일에 출력한다.
41.     echo "$prod_hour,$c_thickness" >> $TEMP_FILE
42.     # 산점도 처리용 CSV로 출력한다.
43.     echo "$a_speed,$b_speed,$rate_terms,$load_time,$highpressure_time,$c
44.     ↪_thickness" >> $SCATTER_FILE
45.
46. done < $INPUT_FILE
# 샘플링 데이터 파일을 랜덤하게 섞어준다.

```



• 데이터 가공 스크립트 소스(02.proc_csv.sh)

- **라인 03** : 가공 대상인 자동차부품생산 데이터(autoparts.csv) 지정하는 라인이다.
- **라인 05** : 샘플링 데이터 임시 저장파일(autoparts.tmp)을 지정하는 라인이다.
- **라인 07** : 산점도 분석을 위한 결과파일(autoparts.r_scatter.csv)을 지정하는 라인이다.
- **라인 09** : 샘플링 데이터 파일(autoparts_r_sampling.csv)을 지정하는 라인이다.
- **라인 14** : 산점도용 분석용 파일의 헤더 정보를 지정하는 라인이다.
- **라인 16~26** : 원시데이터를 1라인씩 읽어서 금형(mold)이 양산이고 제품이 '45231-3B400'인 데이터만 추출하여 저장하는 라인이다.
- **라인 28~37** : 일자가 2014-03-31 데이터를 대상으로 지정하는 라인이다.
- **라인 41~44** : 샘플링 데이터를 임시 파일에 저장하는 라인이다.

III. 가공

```
47. SHUFFLE_FILE='/home/eduuser/nia_kbig/data/autoparts.shuffle'
48. echo "" > $SHUFFLE_FILE
49. awk 'BEGIN{srand() }
50. { lines[++d]="$0 "
51. END{
52.     while (1){
53.         if (e==d) {break}
54.         RANDOM = int(1 + rand() * d)
55.         if ( RANDOM in lines ){
56.             print lines[RANDOM]
57.             delete lines[RANDOM]
58.             ++e
59.         }
60.     }
61. }' $TEMP_FILE >$SHUFFLE_FILE
62. echo "2) shuffling..."
63. # 샘플링 아이템 임시저장 파일
64. RESULT_TMP='/home/eduuser/nia_kbig/data/autoparts.r_data2.tmp'
65. function makeMatrix() {
66.     new_data="";
67.     match_string="$1."
68.     `cat $SHUFFLE_FILE | grep "$match_string" | head -5 > $RESULT_TMP`'
69.     c=0
70.     IFS=' '
71.     while read prod_h c_thickness
72.     do
73.         c_thickness=$(echo $c_thickness | sed -e 's/\r//g')
74.         if [ $c -eq 0 ]; then
75.             new_data="${c_thickness}"
76.         else
77.             new_data="${new_data},${c_thickness}"
78.         fi
79.         c=`expr $c + 1`
80.     done < $RESULT_TMP
```

```

81. IFS=''''
82. echo $new_data >> $SAMPLING_FILE
83. }
84. # 샘플링 임시 파일에서 시간대별로 시간대별로 랜덤 5개의 아이템을 샘플링 파일에
85. ↪ 출력한다.
86. arr=('오전 12','오전 1','오전 2','오전 3','오전 4','오전 5','오전 6','오전 7','오전 8','오전
87. ↪ 9','오전 10','오전 11','오후 12','오후 1','오후 2','오후 3','오후 4','오후 5','오후 6','
88. ↪ 오후 7','오후 8','오후 9','오후 10','오후 11')
89. echo "x1,x2,x3,x4,x5"> $SAMPLING_FILE
90. IFS=','
91. for prod_hour in ${arr[@]}
92. do
      makeMatrix "$prod_hour"
done
echo "3) completed"

```

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

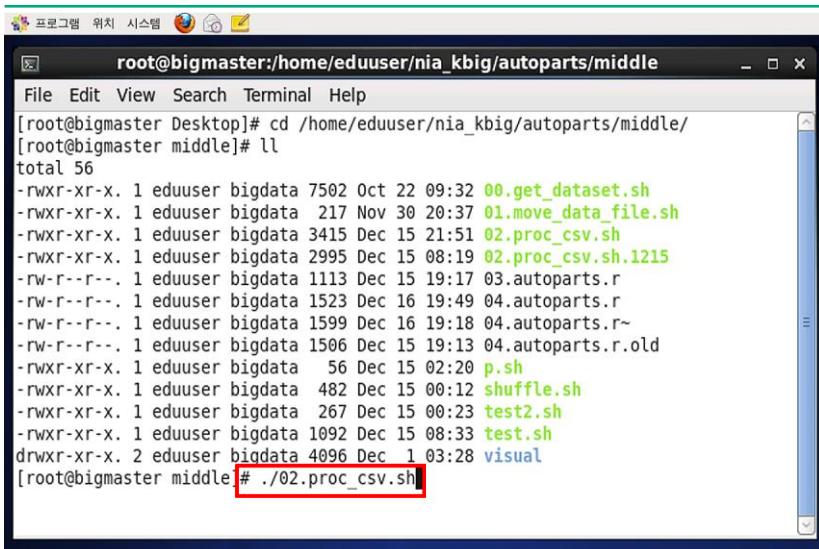
VI. 시각화



- 67페이지 데이터 가공 스크립트 소스(02.proc_csv.sh)
- 라인 47~60 : 샘플링 데이터를 랜덤하게 섞어 저장하는 라인이다.
- 라인 65~83 : 매트릭스 만들 함수를 정의하는 라인이다.
- 라인 85~86 : 시간대별 샘플링 데이터를 5행렬 기준으로 작성하는 라인이다.
- 라인 87~91 : 시간대별 5개의 데이터를 배열로 지정하여 출력하고 24시간 기준으로 출력하는 라인이다.

III. 가공

▶ 원시데이터에서 분석 대상 데이터 가공



The screenshot shows a terminal window titled "root@bigmaster:/home/eduuser/nia_kbig/autoparts/middle". The window contains the following command-line session:

```
[root@bigmaster Desktop]# cd /home/eduuser/nia_kbig/autoparts/middle/  
[root@bigmaster middle]# ll  
total 56  
-rwxr-xr-x. 1 eduuser bigdata 7502 Oct 22 09:32 00.get_dataset.sh  
-rwxr-xr-x. 1 eduuser bigdata 217 Nov 30 20:37 01.move_data_file.sh  
-rwxr-xr-x. 1 eduuser bigdata 3415 Dec 15 21:51 02.proc_csv.sh.1215  
-rw-r--r--. 1 eduuser bigdata 2995 Dec 15 08:19 02.proc_csv.sh.1215  
-rw-r--r--. 1 eduuser bigdata 1113 Dec 15 19:17 03.autoparts.r  
-rw-r--r--. 1 eduuser bigdata 1523 Dec 16 19:49 04.autoparts.r  
-rw-r--r--. 1 eduuser bigdata 1599 Dec 16 19:18 04.autoparts.r~  
-rw-r--r--. 1 eduuser bigdata 1506 Dec 15 19:13 04.autoparts.r.old  
-rwxr-xr-x. 1 eduuser bigdata 56 Dec 15 02:20 p.sh  
-rwxr-xr-x. 1 eduuser bigdata 482 Dec 15 00:12 shuffle.sh  
-rwxr-xr-x. 1 eduuser bigdata 267 Dec 15 00:23 test2.sh  
-rwxr-xr-x. 1 eduuser bigdata 1092 Dec 15 08:33 test.sh  
drwxr-xr-x. 2 eduuser bigdata 4096 Dec 1 03:28 visual  
[root@bigmaster middle]# ./02.proc_csv.sh
```

- 원시 데이터 셋에서 분석할 데이터를 가공하여 autoparts.r_sampling.csv, autoparts.r_scatter.csv 파일을 생성한다.
.02.proc_csv.sh 입력 후 엔터

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화



IV 저 장

개요

75

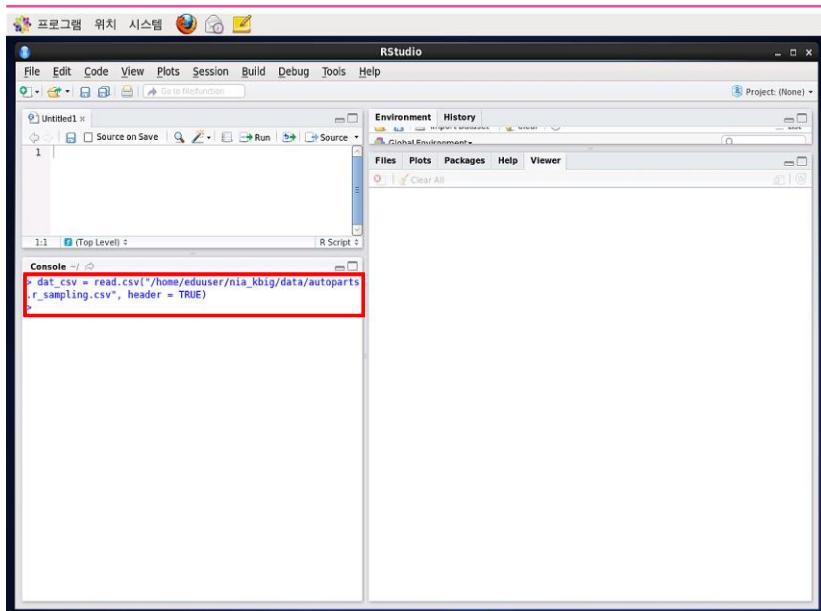
> 개요

자동차 부품 생산 데이터에서 분석 대상인 45231-3B400 (Oil Gasket) 제품 중에서 생산일시(prod_data)가 2014-03-31이고, 금형(mold)이 양산, 공정(prod)이 생산인 데이터 중에서 분석하기 위해서 가공된 XBar-R 관리도 데이터 파일 (autoparts.r_sampling.csv)과 산점도(autuparts.r_scatter.csv) 출력용 데이터 파일을 R Studio 프로그램을 실행하여 메모리에 저장을 시킨다. R에서는 텍스트, csv, 엑셀 데이터 등 다양한 형태의 데이터를 읽어 올 수 있다.

> 저장 방법

- proc_csv.sh 셀에서 저장한 autoparts.r_sampling.csv 파일을 R Studio를 활용하여 가공된 데이터를 메모리상에 탐재하여 분석할 수 있게 한다.
- **XBar-R 관리도 데이터 :**
/home/eduuser/nia_kbig/data/autoparts.r_sampling.csv
- **산점도 데이터 :**
/home/eduuser/nia_kbig/data/ autuparts.r_scatter.csv

▶ R studio에 분석 데이터 메모리에 저장



- 가공된 데이터를 R Studio에서 메모리에 분석 데이터를 저장한다.

W





V 분석

개요

79

데이터 분석 R스크립트

81

V

분석

▶ 개요

제조 데이터의 분석은 R Studio에서 가공된 Xbar-R 관리도 데이터, 산점도 데이터를 불러와서 R qcc 패키지(통계적 품질 관리 패키지)를 이용하여 Xbar-R 관리도를 출력을 한다. 산점도는 plot() 함수를 사용하여 C두께(당구 두께)와 다른 파라미터 값들 과의 산점도 차트를 출력하여 시간대 별 불량률을 찾을 수 있다.

▶ 분석 방법

- 자동차 부품 데이터 중 45231-3B400(Oil Gasket) 제품의 양산 데이터를 시간별로 5개씩 샘플링 하여 R의 qcc 패키지를 이용하여 Xbar-R 관리도를 출력하고 다른 파라미터 데이터와의 상관관계를 산점도로 출력하여 연관성을 분석한다.

▶ 가공 데이터

- 시간별 샘플 데이터 (autoparts.r_sampling.csv)

x1,x2,x3,x4,x5		x1,x2,x3,x4,x5
20.3,24.23.6,25,23.5		22.7,23.9,23.9,24.6,24.1
25.22,20.8,21.2,22.8		26.1,25.6,24.7,25.1,25.7
21.6,24.8,22.7,25.2,25.5		24.3,24.2,24.5,23.4,21.8
22.6,25.8,23.2,25.5,22.3		24.5,22,23.4,23.5,23.1
24.7,24.9,27,26.2,26.9		20.1,33.6,23.1,31.5,24.3

24.1,24.1,19.4,25.3,23.6	35.7,22.6,25.3,25.9,24.7
23.5,18.8,20,20.2,26.1	23.1,25.2,23.6,25.2,25.2
22.8,25.5,27,23.8,22.8	23.4,23.2,24.6,24.6,23.5
23.8,25.2,24.7,24.4,26.3	22.5,23.4,24.3,25,23.3
23.3,24.7,22.8,25.2,23	34,23.1,25.4,24,24.4
25,25.9,21.8,25.7,24.3	25.3,21.5,22.5,23.8,23.6
23.8,24.5,23,26.4,25.5	24.5,25.5,21.3,24.4,21.4

I. 개요

■ 산점도 샘플 데이터 (autoparts.r_scatter.csv)

a_speed,b_speed,rate_terms,load_time,highpressure_time,c_thickness	
0.669,1.668,85,19.2,76,27	0.664,1.649,85,19.2,76,26.2
0.647,1.666,86,19.2,80,26.2	0.669,1.648,85,19.2,75,25.9
0.662,1.661,86,19.2,83,25.9	0.672,1.628,85,19.1,75,25.9
0.657,1.644,86,19.2,71,26.5	0.668,1.633,85,19.2,78,25.6
0.681,1.64,86,19.2,77,26.5	0.66,1.662,86,19.2,75,25.5
0.658,1.649,87,19.2,77,27.9	0.651,1.655,86,19.2,72,25
0.647,1.662,86,19.2,75,26.4	0.672,1.651,86,19.1,74,26.1
0.643,1.644,86,19.2,78,27.9	0.648,1.632,86,19.2,80,26.5
0.676,1.647,86,19.2,81,28.1	0.656,1.652,86,19.2,76,28
0.658,1.645,85,19.2,77,27.8	0.658,1.632,86,19.2,79,27.3
0.663,1.649,85,19.2,78,26.9	0.665,1.629,85,19.1,71,27.4
0.645,1.656,85,19.2,80,26.4	0.659,1.648,86,19.2,79,26.5
0.656,1.641,86,19.2,74,27.5	0.657,1.626,86,19.2,67,26.6
0.654,1.642,86,19.2,78,27.6	0.645,1.626,84,19.2,70,27.1
0.648,1.66,86,19.2,78,26	0.649,1.633,85,19.2,68,25.6

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화

▶ 데이터 분석 R스크립트

▶ Xbar-R관리도 R스크립트(03.autoparts.r)

- 시간별 샘플링한 데이터를 qcc 패키지를 이용하여 Xbar-R 관리도를 출력한다.
- 화면을 2분할하여 각 화면 단위로 Xbar, R 관리도를 출력한다.
- 분석 스크립트를 R 스크립트로 작성하여 실행을 한다.
- C두께와 다른 파라미터와의 산점도를 출력하여 연관성을 분석한다.

03.autoparts.r

```

01. =====
02. # qcc 패키지를 설치한다.
03. =====
04. install.packages("qcc")
05. library(qcc)
06. =====
07. # 샘플링 csv 파일을 로드한다.
08. =====
09. dat_csv = read.csv("/home/eduuser/nia_kbig/data/autoparts.r_sampling.csv", hea
  ↴ der = TRUE)
10. =====
11. # 2x1 화면으로 PLOT 표시
12. =====
13. par(mfrow=c(2,1))
14. =====
15. q1 = qcc(dat_csv, type = "xbar", title="XBar")
16. summary(q1)
17. q2 = qcc(dat_csv, type = "R", title="R관리도")
18. summary(q2)
19. =====
20. # Xbar , Rchart를 plot이용해서 분할된 영역으로
21. # 맞추어준다.
22. =====
23. plot(q1, title="X Bar", add.stats=TRUE, restore.par=FALSE)
24. abline(h=22,col="red")
25. abline(h=26,col="red")
26. plot(q2, title="R관리도", add.stats=TRUE, restore.par=FALSE)
27. abline(h=1.5,col="red")
28. abline(h=7.5,col="red")

```



• R스크립트(03.autoparts.r)

- 라인 04~05 : qcc 패키지를 설치를 하고 라이브러리 로드하는 라인이다.
- 라인 09 : 샘플링 데이터(autoparts.r_sampling.csv) 파일을 읽어 로드하는 라인이다.
- 라인 15~18 : qcc 패키지의 xbar 차트와 R차트를 지정하는 라인이다.
- 라인 23~25 : xbar 차트를 출력을 하고 관리하한과 관리상한선을 차트위에 같이 출력하는 라인이다.
- 라인 26~28 : R차트를 출력을 하고 관리하한과 관리상한선을 차트위에 같이 출력하는 라인이다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화

➤ 산점도 출력 R스크립트(04.autoparts_scatter.r)

04.autoparts_scatter.r

```

01. =====
02. # csv 파일을 로드한다.
03. =====
04. dat_csv = read.csv("/home/eduuser/nia_kbig/data/autoparts.r_scatter.csv", heade
    ↪ r = TRUE , encoding="UTF-8")
05. =====
06. # 3x2 화면으로 PLOT 표시
07. =====
08. par(mfrow=c(3,2))
09. =====
10. # a_speed & c_thickness
11. =====
12. ac<-plot(dat_csv$c_thickness , dat_csv$a_speed, main="a_speed & c_thicknes", xl
    ↪ im=c(0,70) , ylab="a_speed", xlab="c_thicknes")
13. =====
14. # b_speed & c_thickness
15. =====
16. bc<-plot(dat_csv$c_thickness , dat_csv$b_speed , main="b_speed & c_thicknes",
    ↪ xlim=c(0,70), ylab="b_speed", xlab="c_thicknes")
17. =====
18. # rate_term & c_thickness
19. =====
20. rc<-plot(dat_csv$c_thickness , dat_csv$rate_term , main="rate_term & c_thicknes"
    ↪ , xlim=c(0,70), ylab="rate_term", xlab="c_thicknes")
21. =====
22. # load_time & c_thickness
23. =====
24. lc<-plot(dat_csv$c_thickness , dat_csv$load_time , main="load_time & c_thicknes
    ↪ " , xlim=c(0,70), ylab="load_time", xlab="c_thicknes")
25. =====
26. # highpressure_time & c_thickness
27. =====
28. hc<-plot(dat_csv$c_thickness , dat_csv$highpressure_time , main="highpressure_
    ↪ time & c_thicknes" , xlim=c(0,70), ylab="highpressure_time", xlab="c_thicknes")
29. =====

```



- 산점도 출력 R스크립트(04.autoparts_scatter.r)
- 라인 04 : 산점도 데이터 파일(autoparts.r_scatter.csv)을 로드하는 라인이다.
- 라인 12 : C두께와 저속속도(a_speed)와의 산점도를 출력하는 라인이다.
- 라인 16 : C두께와 고속속도(b_speed)와의 산점도를 출력하는 라인이다.
- 라인 20 : C두께와 조건율(rate_term)과의 산점도를 출력하는 라인이다.
- 라인 24 : C두께와 하중시간(load_time)과의 산점도를 출력하는 라인이다.
- 라인 28 : C두께와 고압시간(hightpressure_time)과의 산점도를 출력하는 라인이다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화



1

2



VI 시각화

개요	87
분석 데이터 시각화	89
데이터 분석	90

VI

시각화

> 개요

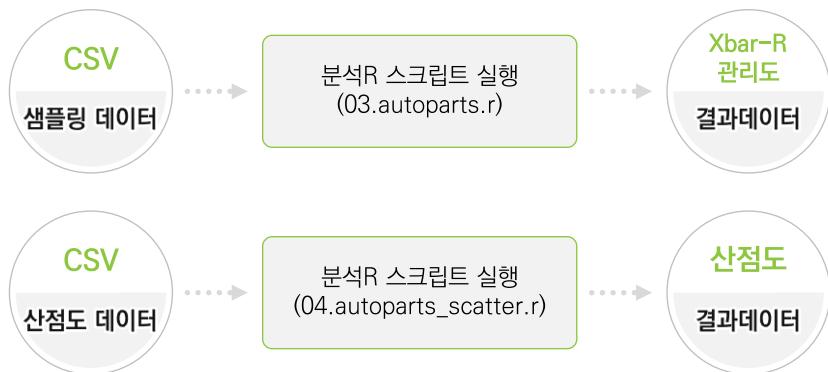
자동차 부품 생산 데이터의 분석과정에서 가공한 데이터를 가지고 Xbar 관리범위(24 ± 2)를 벗어나는 아웃트(관리상한,하한선에서 뛰어나간 점)가 발생되었던 시간대를 찾아보다. 또한 R 관리범위(4.5 ± 3) 안에 들어오는 데이터에서 중심선에서 련(중심선의 한쪽에 점이 모이는 상태, 점의 수를 련의 길이라 함)이 6점 이상을 찍는지 여부를 파악하여 불량이 예상되는 시간대를 파악해본다. R Studio를 활용하면 분석과 함께 시각화까지 일괄적으로 진행할 수 있다.



> 시각화 방법 및 활용기술

- R qcc 패키지를 이용하여 시간별 샘플링한 데이터를 Xbar-R 관리도를 출력을 한다.
- qcc 패키지의 type0이 xbar 와 R을 구분하여 출력한다.
- plot 함수로 C두께와 다른 항목 간의 산점도를 출력한다.

▶ 데이터 변환

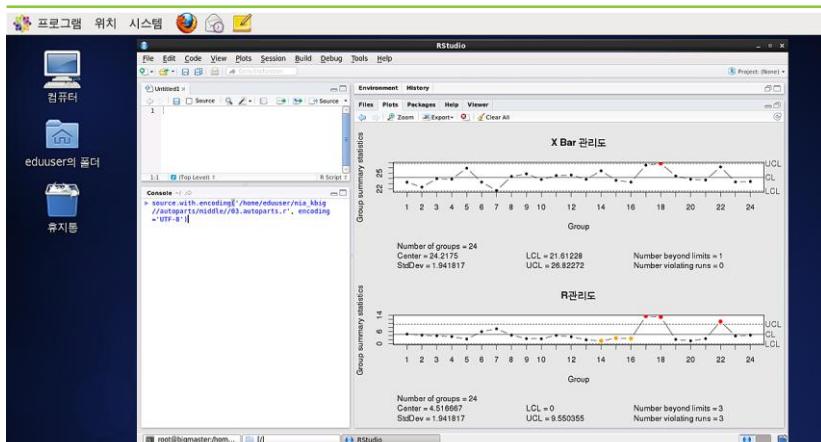


▶ 시각화 과정

- R Studio로 Xbar-R 관리도 작성한 03.autoparts.r 인 R 스크립트를 실행한다.
- R Studio로 산도점 출력은 04.autoparts_scatter.r 스크립트를 실행한다.

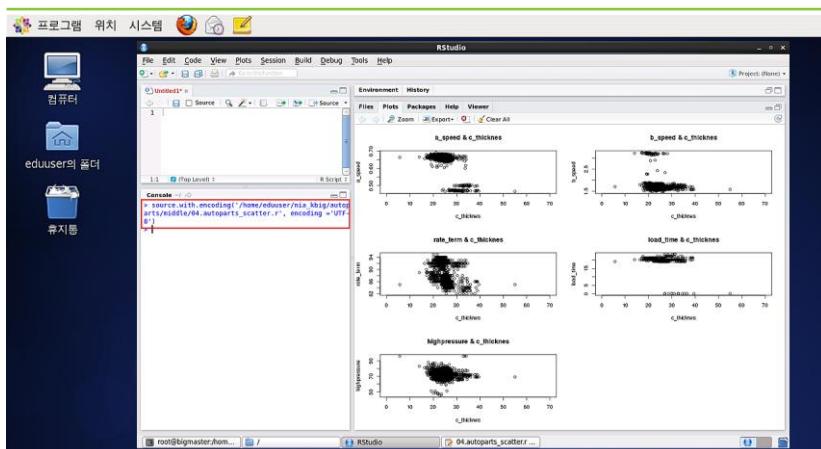
> 분석 데이터 시각화

> R Studio 실행(Xbar-R 관리도 출력)



- 03.autoparts.r 스크립트를 로드 한 후 실행을 하여 Xbar-R 관리도를 출력한다.
- source.with.encoding('/home/eduuser/nia_kbig/autoparts/middle/03.autoparts.r', encoding='UTF-8')

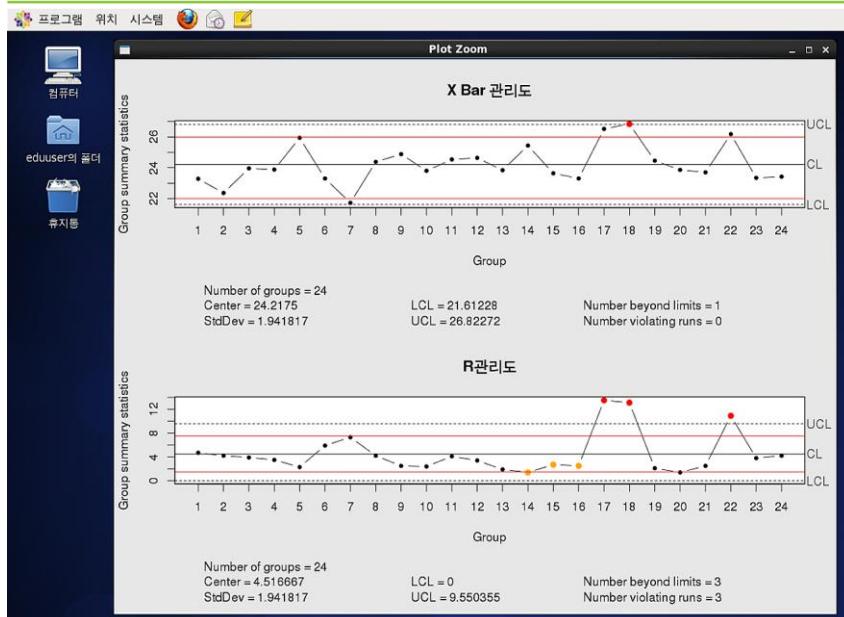
> R Studio 실행(산점도 출력)



- 04.autoparts_scatter.r 스크립트를 로드 한 후 실행을 하여 산점도를 출력한다.
- source.with.encoding('/home/eduuser/nia_kbig/autoparts/middle/04.autoparts_scatter.r', encoding='UTF-8')

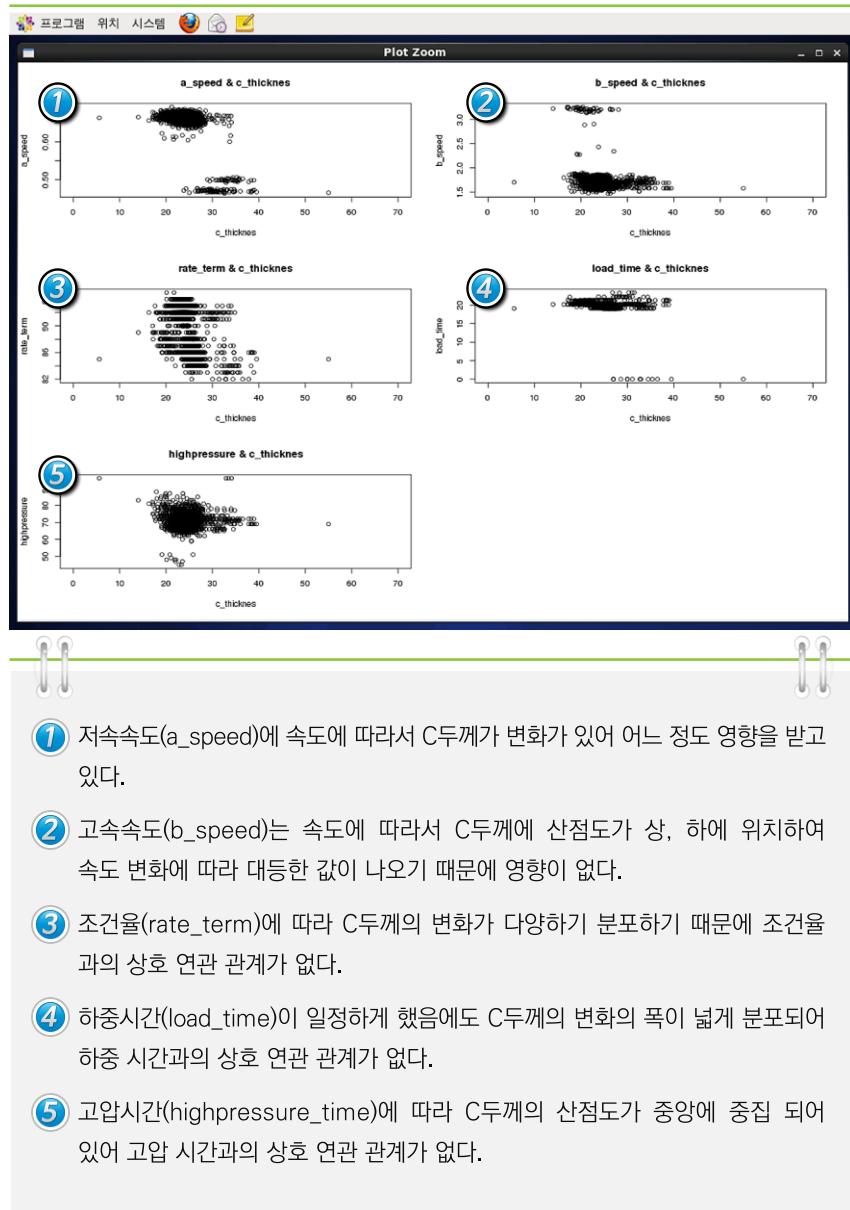
> 데이터 분석

> 시간별 샘플링 데이터 Xbar-R 관리도



- Xbar 출력 결과 관리범위 24 ± 2 를 벗어난 지점은 5시, 7시, 17시, 18시, 22시 부분에 발생이 되었다.
- 규격 상한을 벗어난 시점은 18시 부분에 발생이 되어 제품에 불량이 발생이 되었다.
- R 관리도 출력 결과 관리범위 4.5 ± 3 범위 안에서 14시, 15시, 16시에 Run 3점을 찍고 있어 공정 상태의 문제점이 예상되며 17시, 18시, 22시에 관리범위를 벗어나 불량제품이 생산되었다.

> C두께와 파라미터와의 산점도 출력



I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화



VII 예제문제

예제 문제1

95

예제 문제2

96

예 / 제 / 문 / 제

예제 1

C두께와 고속속도(b_speed)의 상관관계를 파악하라.

- C두께(탕구두께)와 고속속도(b_speed) 파라미터 설정값에 대한 연관성이 있는지를 파악하라.

- 45231-3B400(Oil Gasket)의 생산일시(prod_date)가 2014-03-31, 금형(mold)이 양산이 데이터를 추출한다.
- R Studio에 추출된 데이터를 로딩한다.
- C두께와 고속속도(b_speed)의 산점도를 출력한다.
- C두께와 고속속도(b_speed)의 회귀분석을 한다.
- C두께와 고속속도(b_speed)에 대한 선형회귀분석 그래프를 출력한다.

예제 2

45231-3B400(Oil Gasket)의 생산 일시(prod_data)가 2014년 3월 31일의 금형(mold)이 양산인 데이터를 가지고 공정 생산능력을 계산하라.

- C두께(탕구두께)의 관리범위 24 ± 2 의 범위 밖에 있는 제품을 불량으로 처리하여 해당 제품의 공정능력을 계산하라.

- 45231-3B400(Oil Gasket)의 생산일시(prod_date)가 2014-03-31이고, 금형(mold)이 양산인 데이터를 추출한다.
- R Studio에 추출된 데이터를 로딩 한다.
- 시간별 샘플링 5개씩 취해서 24시간 테이블을 만든다.
1시간 측정 샘플 5개의 시료를 24시간 기준으로 작성 한다.
- qcc 패키지를 사용하여 C두께의 Xbar를 구한다.
- process.capability()를 이용하여 공정능력을 구한다.

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

