공공 빅데이터 표준분석모델 매뉴얼

어린이 안전 및 교통사고 원인분석

국민의 삶을 바꾸는 빅데이터

행정안전부 NIA 한국정보화진흥원

목 차

1. 과제 개요

1.1 필요성 및 목표

□ 2014년도부터 교통사고 사망자는 감소하는 추세이나, 2015년 어린이 교통사고 사망자는 전년대비 25% 급증

- 전체 어린이 교통사고 사망자 중 '보행 중 사망'이 63%로 전체 교통사고 사망자 중 보행 중 사망 비율(38.8%)을 훨씬 상회
- 어린이 교통사고 사망자는 등교 시(오전 8시~10시) 16.4%, 하교 시(오후 2시~6시) 62.5% 발생하는 것으로 집계

□ 안전사고 사고유형 중에서는 추락 • 넘어짐 • 미끄러짐이 전체 안전사고의 31.6%를 차지

□ 어린이 보행자 교통/안전사고 감소 및 선제적 어린이 교통안전 방안수립 기반마련과 실제 사고 데이터와 국민 참여형 데이터의 연계분석을 통해 어린이 교통안전 방안 수립 지원 및 향후 이러한 분석결과를 지속적으로 반영할 수 있는 모델 개발 필요

1.2 과제설명

□ 어린이, 교사, 학부모 등 민간협력을 통한 국민 참여 데이터 분석을 바탕으로 어린이 주요 이동경로 수집 및 관련분석 수행

□ 어린이·학부모·교사의 안전의식 고취 및 사고 예방을 위한 교통안전지도 및 '안전/안심 공공서비스' 교육방안 자료 제시

□ 어린이 교통사고 예방을 위한 경찰인력 배치 등 대응방안 수립 지원하고 사고원인분석을 통한 대인(운전자 교육)/대물(시설물 점검) 개선방안 도출을 위한 참

고 자료 제시

□ 어린이 주요 이동경로 파악 및 어린이 교통/안전사고 통계분석 및 Hot Spot 분석 수행을 통한 분석결과 시각화 자료 제시

1.3 전제조건 및 제약사항

□ 차후 각 시도로 확산 시 “어린이 교통안전” 과제에서 수행한 각 지점별 교통/안전사고에 대한 상세 분석결과를 도출할 수 있도록 국민 참여형 데이터 수집을 위한 웹 기반 설문사이트를 제공하여 전국단위로 확산 및 활용

□ 각 시도에서 어린이 교통/안전사고 파악 및 예방을 위해 본 과제의 분석을 활용할 시, 설문사이트를 통해 수집한 참여형데이터 및 제공한 분석 방법론을 기반으로 각 시도별 상황에 맞는 커스터마이징 작업이 필요할 수 있음

1.4 활용데이터

□ 공공데이터 : 교통안전시설물 데이터, CCTV 데이터, 놀이시설 데이터, 119 구급차량출동시스템 데이터, 교통사고 데이터, 초등학교/학원 위치 시스템 데이터, 인구 데이터, 실폭도로 데이터

표 1. 어린이 교통안전 활용 공공데이터

제공기관	데이터 한글명	데이터설명(용도)	데이터 유형
경기도, 이천시	교통안전시설물 데이터 CCTV 데이터 놀이시설 데이터	이천시 시설물 설치현황 데이터 수집 및 교통사고 원인분석	xlsx
경기콘텐츠진흥원	119 구급차량 출동 시스템	119구급차량출동시스템 데이터 수집 및 안전사고 원인분석	xlsx

제공기관	데이터 한글명	데이터설명(용도)	데이터 유형
	데이터		
도로교통공단	교통사고 데이터	경기도, 이천시 교통사고 데이터 수집 및 교통사고 원인분석	xlsx
경기교육청 (경기도 협의)	초등학교/학원 위치 시스템 데이터	초등학교/학원 소재지 데이터 수집 및 안전/교통사고 원인분석	xlsx
통계청	인구 데이터	인구 데이터 수집 및 안전/교통사고 원인분석	shp
국가공간정보포탈	실폭도로 데이터	격자생성 데이터 수집 및 안전/교통사고 원인분석	xlsx

□ 참여형데이터 : 환경분석 데이터, 경로분석 데이터, 행동패턴분석 데이터

표 2. 어린이 교통안전 활용 참여형데이터

구분	초등학생	성인
시기	• 2016.11.07 ~ 2016.12.16 (총 1866건 수집)	• 2016.11.07 ~ 2016.12.16 (총 464건 수집)
대상	• 이천시 내 초등학교 재학생	• 이천시 내 초등학교 교사, 학부모, 관할경찰서 경찰관
지역	• 이천시 - 면적 : 461.36㎢ - 인구 : 20만 4935명 (2015년 기준)	• 이천시 - 면적 : 461.36㎢ - 인구 : 20만 4935명 (2015년 기준)

구분	초등학생	성인
방법	• 웹 기반 구조화된 수집방식 이용 • 경기도 교육청 및 각 학교 협조 하에 수업시간 중 테블릿 PC를 통한 조사 시행	• 교사, 경찰관: 근무지 내 조사 • 학부모: 가정통신문 발송
표본 추출 방법	• 전수조사 자율 참여	• 임의표본 추출
표본 크기	• 최대 12,000명 이내	• 협조 학부모, 교사, 경찰관 (수백명 예정)

1.5 분석기법

□ 탐색적 자료 분석(Exploratory Data Analysis : EDA) 기법을 활용하여 기술통계량을 바탕으로 데이터의 특성을 파악하고 데이터 간 구조적 관계를 파악

□ Map Matching 기법을 이용하여 수집한 데이터의 Point 좌표를 격자 및 도로 링크 등 공간단위에 매핑

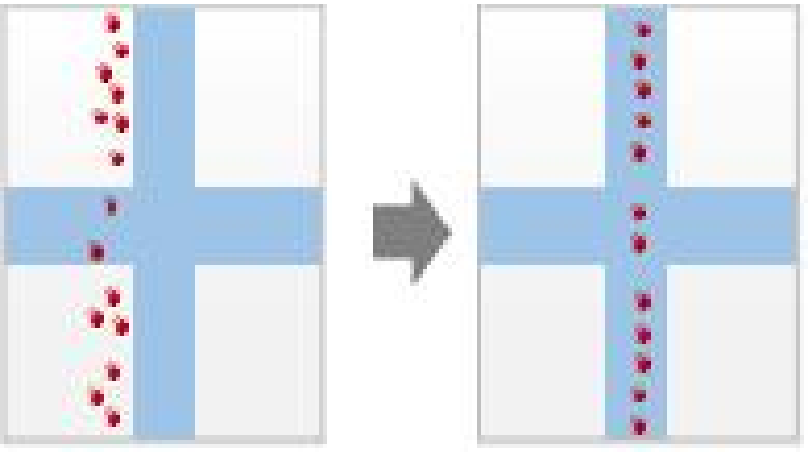


그림 1. Map Matching 기법 예시

□ 지오코팅(Geocoding) 기법을 통해 데이터의 주소 및 위치 정보를 좌표 정보로 변환하고 서로 다른 좌표계는 세계측지계(WGS84)로 변환

□ 공공데이터 및 참여형데이터의 Hot-Spot 분석을 통해 Point 객체 집합의 밀도와 폴리곤 정보를 이용, 지역 및 지점별 Hot/Cool Spot 분석결과 제공

그림 2. Hot Spot 분석 예시

□ 텍스트 마이닝 분석을 통해 교통사고의 개요 등 비정형 텍스트 데이터로부터 중요한 의미를 가지는 핵심어를 추출

1.6 분석결과 활용

□ 분석결과의 기관별 활용(학교, 경찰서, 지자체 시설안전담당 등)

- 학교 및 교사: 녹색 어머니회 지도 지점, 등하교 안전지도 지점 식별, 기타 교통안전 지도 관련 참조에 활용
- 경찰서: 교차로 등 교통지도 지점 식별, 등하교시 교통지도 지점 식별, 학교지점 우범지역 지점 식별에 활용
- 지방자치단체 시설담당: 도로교통공단에서 시설개선 우선지점 식별 시 교통사고 지점 외 기타참고자료로 활용

□ 교통 및 안전사고 사각지역 파악

- 실제 교통사고 데이터와 참여형데이터의 연계 분석을 통해 잠재위험도가 높은 교통 및 안전사고 사각지역 도출
- 도로교통공단 및 교통안전공단에서 실시하는 교통안전 취약지점 식별 및 어린이 안전대책 수립 시, 해당지역 맞춤형 교통위험지점 기초자료로 활용 가능
- 어린이 주요 이동경로지만 학교 반경 300m 이내에 포함되지 않은 집중관리 고려지역 도출하고 지역 맞춤형 효과적인 교통안전 대책 수립 지원

□ 교통사고 위험도 및 잠재위험도 활용

- 교통사고 위험도와 잠재위험도의 산점도 분석을 통해 학교별 우발적 사고군 및 잠재적 사고위험군 파악
- 잠재적 사고위험군에 대한 인적, 물적 자원 투입을 위한 기초자료로 활용 가능

2. 분석환경 설정

2.1 개요

□ 오픈소스 분석 소프트웨어 R을 사용하여 EDA 분석 수행

□ R은 통계 계산과 그래픽을 위한 프로그래밍 언어이자 소프트웨어 환경으로, MS Windows, Mac OS X, Linux 등 다양한 OS에서 사용 가능

□ R은 주로 빅데이터 분석을 목적으로 사용되고 있으며, 5000개가 넘는 패키지(일종의 애플리케이션)들이 다양한 기능을 지원하고 있음

□ R을 보다 쉽게 사용하기 위해 GUI 환경을 제공해주는 오픈소스 소프트웨어인 R-studio을 함께 설치하여 사용

2.2 R

□ R은 프로그래밍 언어이기 때문에 패키지가 미리 프로그램 해 놓은 절차에 제한 받지 않으며, 새로운 방법을 프로그램하기가 상대적으로 쉬움

□ 데이터 분석은 원래 대화식으로, 한 번에 하나의 프로세스를 수행하며 분석하는 동안 보이는 것에 기초해 변경이 가능함

□ 상업용 소프트웨어 S-Plus와 거의 호환 가능함

□ SAS가 일반영역에서 널리 통용되는 통계패키지라면, R은 통계학 연구자에게 가장 인기 있고 Finance와 Bioinformatics에 특히 활용도가 높음

□ 오픈소스 소프트웨어이므로 사용자들이 수많은 새로운 함수를 공유하고, 자유롭게 배포할 수 있어 사용에 제한이 없음

2.3 R Studio

□ R Studio는 R을 사용하기 쉽게 GUI(Graphical User Interface) 환경을 제공해주는 오픈소스 소프트웨어

□ R Studio는 코드 직접실행, 구문강조, 괄호 자동입력지원, 명령어 완성, 다양한 단축키, 데이터 보기 및 가져오기, 그래픽 조작, 프로젝트 관리, 버전 관리 등의 다양한 기능을 제공

□ 에디터, 콘솔, 명령어 히스토리, 시각화, 파일탐색, 패키지 관리 등을 한 화면에서 보여줌

□ 프로젝트의 관점으로 파일 관리를 해주며, 소스코드 관리 시스템과 연계할 수 있음

□ 빌트인 데이터 뷰어 내장, 플로팅 히스토리, R help 결합

□ R Markdown 내장으로 문서와 코드를 결합할 수 있게 하고, 재현성 있는 분석을 가능하게 함

□ 리눅스, 맥, 윈도우 등 멀티 플랫폼 지원

3. 데이터 준비

3.1 개요

□ 참여형데이터 구축 개요

표 3. 데이터 구축 개요

정보 수집	• 등•하굣길 등 어린이 주요 이동경로 수집 • 특정지점 및 경로 위험도•위험요인 수집 • 기존데이터 수록정보 외의 정보 수집 - 사고가 날 뻔한 경험이 있는 지점의 위치 및 상세 상황정보 - 위험인식지역 위치 및 위험사유정보, 상황별 행동유형 수집
분석 다양화	• 기존데이터(교통사고 데이터 등)와 연계 분석 가능 - 세부요건 별 통계 및 분포패턴 차이 비교•융합 분석
결과 및 모형 활용	• 정책수립관련 근거자료 제공 - 관련기관(경찰청, 교육청, 소방본부 등)의 정책 대안마련에 필요한 기초 자료 제공 - 어린이 및 학부모 대상 안심정보 제공에 필요한 자료 제공

3.2 데이터 준비

□ 참여형데이터 구축 수단

- 웹 기반 수집 방식으로 수행 (종이설문보다 광범위한 대상의 조사 및 취합 용이)
- 지도기반 위험지역 및 위험사유를 참여자가 표시하는 것을 기본으로 함
- 이천시 소재 초등학교 어린이 대상 현장 면접조사 시행
- 조사 면접원 15명 투입하여 5개 초등학교 56개 학급 현장 전수 조사
- 현장 면접 조사로 어린이 참여율 3.98%(506명)에서 14.65%(1,866명)로 증가

표 4. 데이터 구축 수단

	포털사이트 협조	웹사이트 구축	(기존)종이수집
공통 전제	• 커뮤니티 맵핑 (Community Mapping)기법 - 지도기반 위험지역을 참여자가 표시하는 것을 기본으로 함 - 표시한 지역에 대한 키워드 및 특이사항을 참여자가 입력 		
장점	• 다수를 대상으로 신속한 조사 가능 • 코딩시간 단축 • 응답률 경향 등의 실시간 모니터링 가능	• 광범위한 대상의 조사 및 취합 용이 • 응답률 경향 등의 실시간 모니터링 가능	• 문항이해도 향상 가능 • 수집대상의 식별 용이
단점	• 문항이 많고 복잡할 경우 응답률 저하 • 불특정다수에 의한 대상군의 신뢰도 문제	• 코딩시간 최소화 관건 • 조사대상 및 조사일시 협의문제 • 문항이해도 저하관련 보완 필요	• 조사대상 및 조사일시 협의문제 • 취합 및 데이터화 하는데 상당한 시간 소요 • 시스템화의 한계 발생

3.3 데이터 전처리

□ Step 1. 데이터 필터링 기준

- 평활화: 한 필드 내 유사 범주 있을 경우 새로운 범주로 통합
- 집계일반화: 통합필드 및 범주유사성 높은 데이터 그룹은 통합하며, 필드마다 분석 대상이 되는 범주만 필터링해 분석
- 정규화: 필드마다 상이한 데이터 표현 방식을 하나의 표준 방식으로 통일

- 결측치: 숫자형 필드의 결측치는 0 또는 평균값으로 대체하고, 문자형 필드의 결측치는 '무응답' 혹은 '없음' 범주로 별도 분류

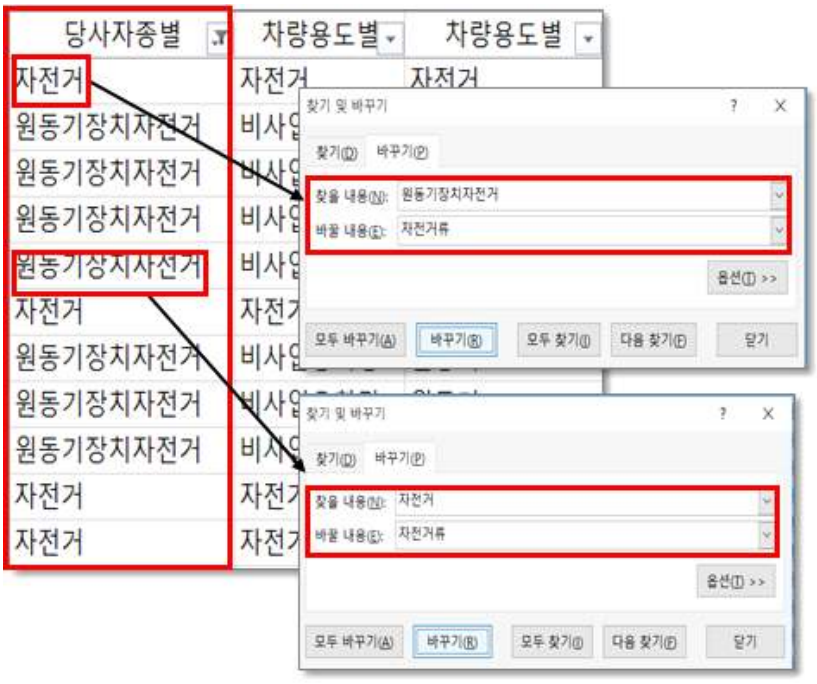


그림 3. 필드 내 범주 평활화

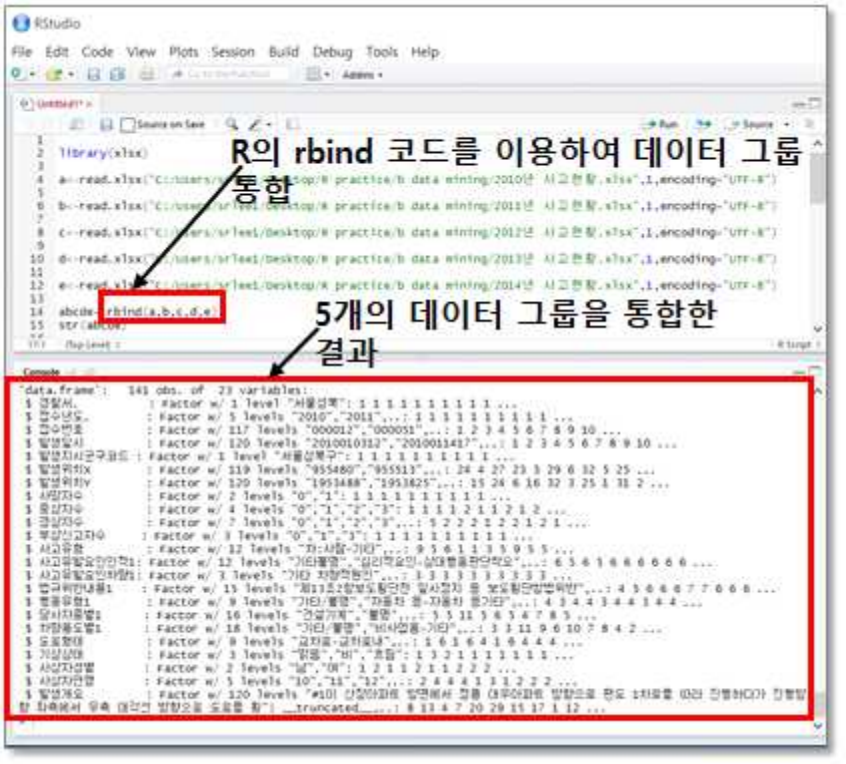


그림 4. 데이터 그룹 집계일반화

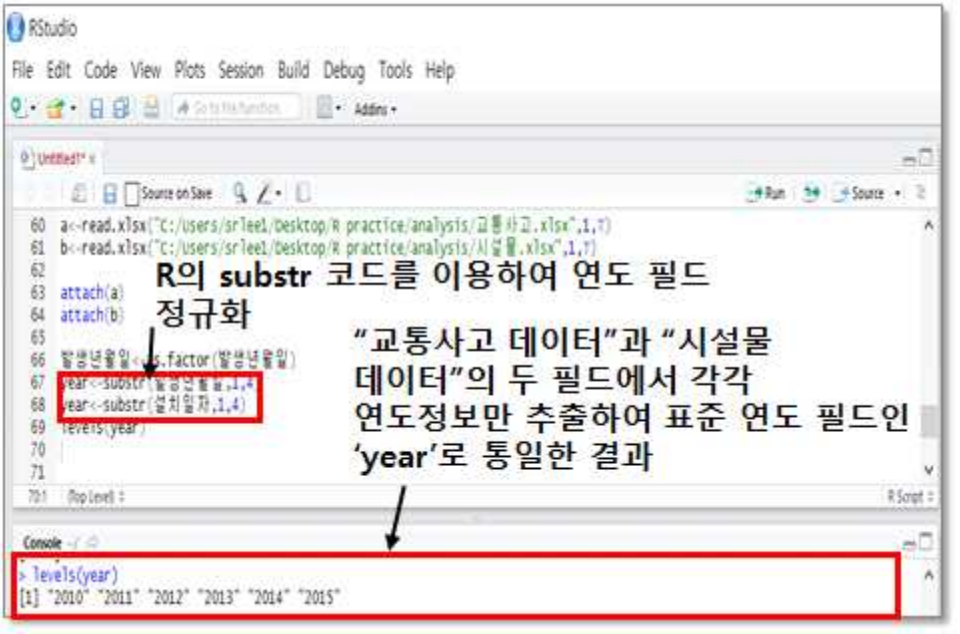


그림 5. 필드 내 범주 정규화

□ Step 2. 데이터 정형화 방안

- 구조화된 단어 데이터 테이블을 이용한 텍스트마이닝 기법으로 텍스트 필드 정형화 (유사어, 불용어 처리)
- 주소 및 위치 데이터를 지오코딩(Geocoding)하여 좌표 정보로 변환하고 서로 다른 좌표계는 세계측지계(WGS84)로 변환

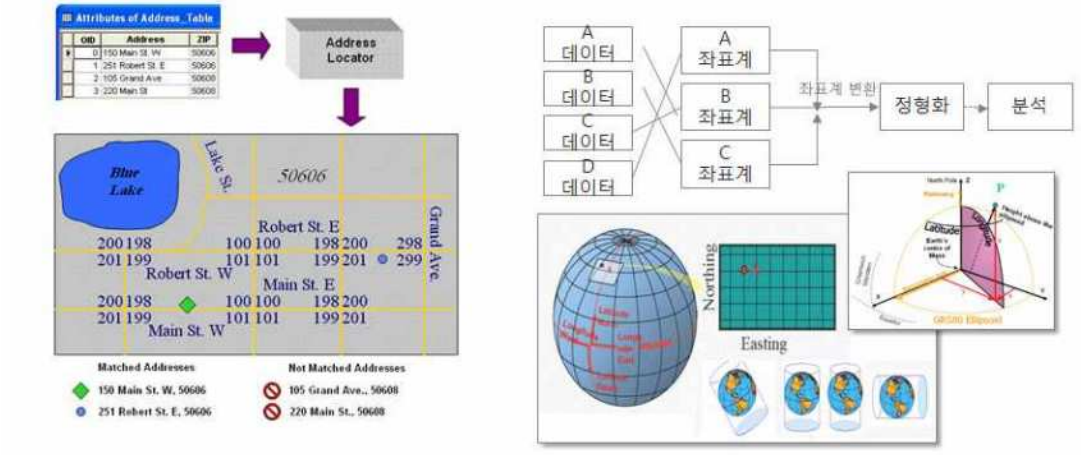
<구조화된(structured) 단어데이터 테이블>

		단어			
		x_1	x_2	....	x_j
문서	x_1	x_{11}	x_{12}	....	x_{1j}
	x_2	x_{21}	x_{22}	....	x_{2j}
	. . . .	. . . .	. . . .		
	x_L	x_1	x_2		x_{ij}

*TF(Term Frequency) = x_{ij}
{문서 i에서 단어 j의 출현빈도}

- 사고관련 키워드 추출
- 단일 키워드 또는 키워드+키워드 형태로 해당 사고 건 추출
- 이와 같은 방법으로 '교통사고 데이터'의 '발생개요' 필드, '119구급차량출동시스템 데이터'의 '평가소견주호소' 필드의 텍스트 데이터 정형화

그림 6. 텍스트 데이터 정형화 방안



- Address Locator를 활용하여 주소 필드를 좌표기반 필드로 지오코딩
- QGIS 이용, 좌표기반 데이터 간 서로 다른 좌표계를 동일 좌표계(WGS84)로 변환

그림 7. 지오코딩 및 세계측지계(WGS84) 변환

□ Step 3. 공간처리 알고리즘

- 참여형데이터의 공간처리를 위해 데이터 수집, 공간정보화 및 변환, 공간매핑 및 연산, 분석결과 시각화 진행

그림 8. 참여형데이터 공간처리 방안

4. 분석 결과(시각화)

4.1 시각화 자료 디렉토리 구성

□ 본 과제의 시각화 자료는 QGIS 프로젝트 파일 기반으로 구성되어 있으며 디렉토리 구조는 아래와 같다.

- 디렉토리 구조는 ZIP 파일을 풀어서 C 드라이브 루트에 “이천시_어린이교통_분석결과_시각화” 폴더에 있는 “공통데이터셋” 폴더와 “학교별 분석결과” 폴더를 넣어야 된다.

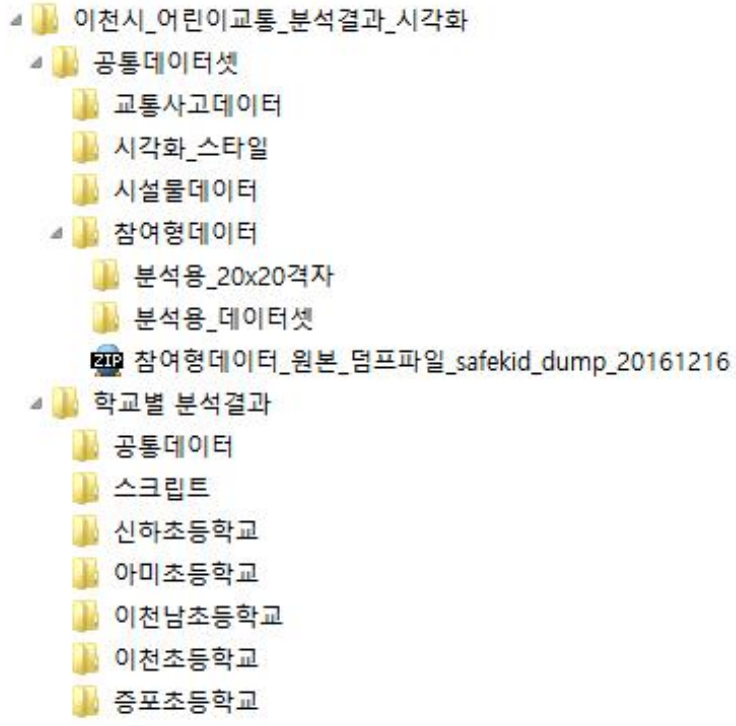


그림 9. QGIS 디렉토리 구조

□ 시각화 파일은 “학교별 분석결과” 폴더 아래의 각 학교별 폴더에 있으며, 확장자는 .qgs(QGIS 프로젝트 파일) 이다.

4.2 시각화 뷰어[QGIS] 설치 및 환경 설정

□ 시각화 툴은 오픈소스 GIS 툴인 QGIS 이며, 버전은 2.18.0 버전을 사용하여야 한다.

□ QGIS 다운로드 방법

- QGIS 는 http://www.qgis.org 에서 다운받을 수 있다.
- QGIS는 "QGIS 독립 설치관리자 버전 2.18 버전 32bit 혹은 64bit를 다운받으면 된다.
- 다운받은 설치 파일을 더블클릭하여 설치를 하면 되며, 모든 선택사항은 기본으로 하여 설치를 하면 된다.

4.3 시각화 조회 방법

□ QGIS를 설치 후 프로그램 메뉴에서 QGIS Desktop 2.18.0을 선택하여 QGIS를 실행한다.

□ 시각화 파일을 열기에 앞서 지도를 플러그인하기 위해 플러그인 > 플러그인 관리 및 설치를 선택하여 아래와 같은 창에서 'TMS'를 검색한다. TMS for Korea를 선택하고 플러그인 설치를 눌러 지도를 설치한다.

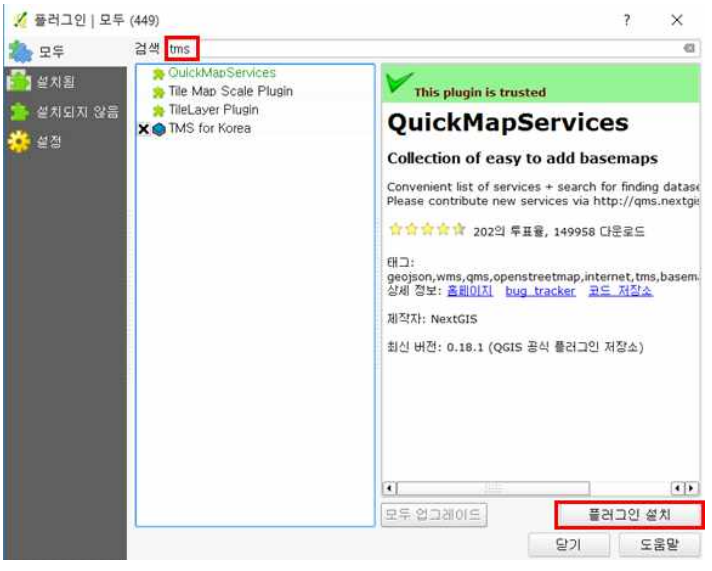


그림 10. 플러그인 설치 창

□ 다음으로 프로젝트 > 열기 메뉴을 선택하여 아래와 같이 QGIS 프로젝트 파일을 선택한다.

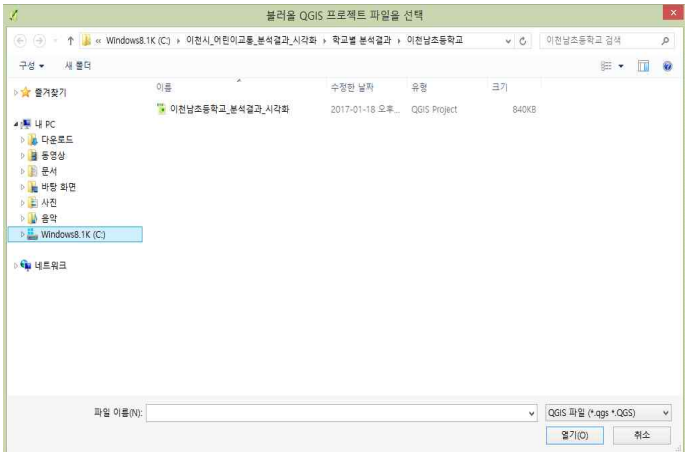


그림 11. QGIS 프로젝트 파일 선택 창

□ 위 프로젝트 파일(.qgs)을 선택하고 웹 > TMS for Korea > Daum Maps > Daum Street를 선택하여 다음 지도를 띄우면 아래와 같이 시각화 데이터가 로드된다.

- 좌측 "Layer Panel"에 데이터 목록이 나타나며, 체크 박스를 On/Off 하여 보고자 하는 레이어만을 볼 수가 있다. 이렇게 관련된 레이어만을 On 하고 중첩하여 스크린샷을 찍어서 주제별 지도 시각화 자료를 만들 수 있다.

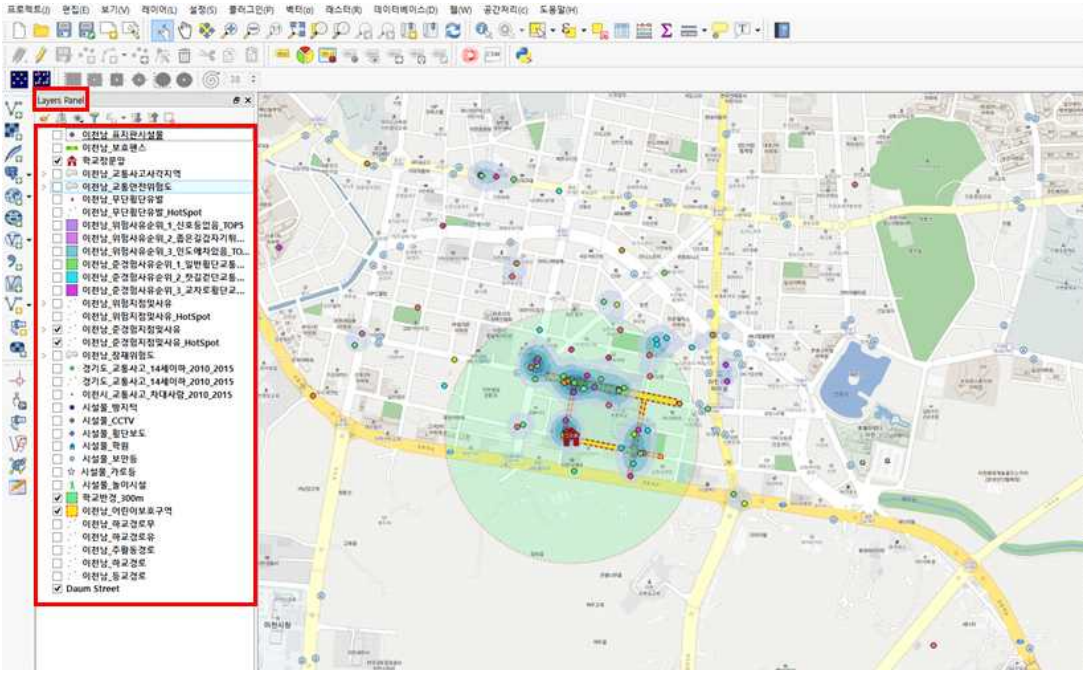


그림 12. 시각화 데이터 로드 예시

5. 공통기반 활용 방안

□ 분석모델 표준화 내용

- 전국의 지자체가 활용할 수 있도록 ① 참여형데이터 수집 및 분석모델, ② 격자단위 등급화 및 히트맵 형식의 Hot-Spot 도출모델, ③ 교통안전위험도 도출모델을 표준화 대상으로 선정함

표 5. 분석모델 표준화 설명

표준화 대상여부	표준화 대상 분석모델	표준화 내용	비고
○	교통 및 안전사고 분석모델	• 참여형데이터 수집 및 분석모델 - 웹 개발 소스 - 설문항목 구성 - 설문결과 반영·분석 모델 • 격자단위 등급화 및 히트맵(heat map) 형식의 Hot-Spot 도출모델 - 격자(100x100, 20x20) 매핑 모델 - 시각화 요소별 Hot-Spot 도출 모델 • 교통안전위험도 도출모델 - 공공데이터, 참여형 데이터 기반 교통안전위험도 산출 모델	• 활용 확산을 위한 일반적 의미의 표준분석모델 개념에 해당

□ 표준분석모델 서비스 방안

- “혜안”에서 표준분석모델의 분석패키지(CD)를 다운로드 받아 사용자 로컬PC에 Q-GIS 등 분석환경을 구축하고, 사용자 매뉴얼에서 제시하는 표준 방법론에 따라 분석을 수행할 수 있도록 지원함

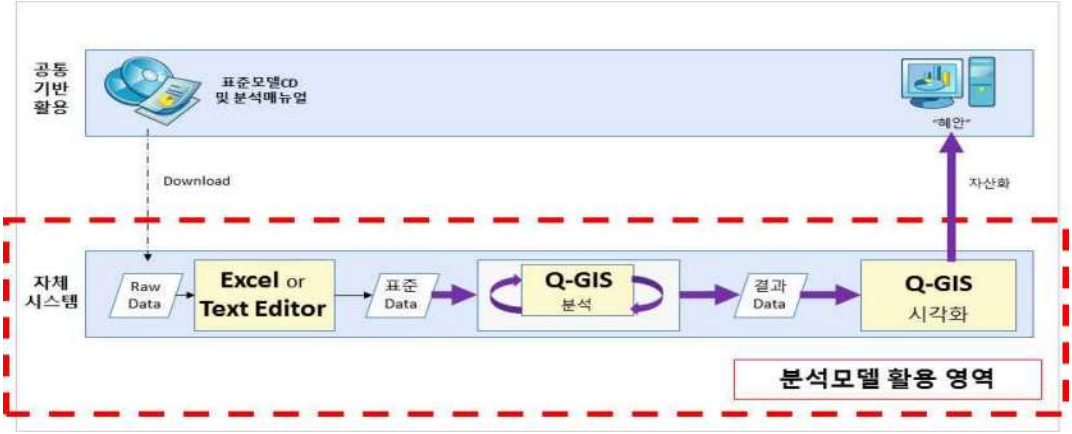


그림 13. 공통기반 적용방안

[부록]

1. 응용프로그램 설치방법

1.1 R 설치방법

□ Step 1. 설치파일 실행

• ‘응용프로그램 설치파일’ 폴더에서 컴퓨터 운영체제에 따라 R 설치파일을 실행

□ Step 2. 설치 언어 선택

• ‘한국어’를 선택하고, ‘확인’ 버튼을 클릭

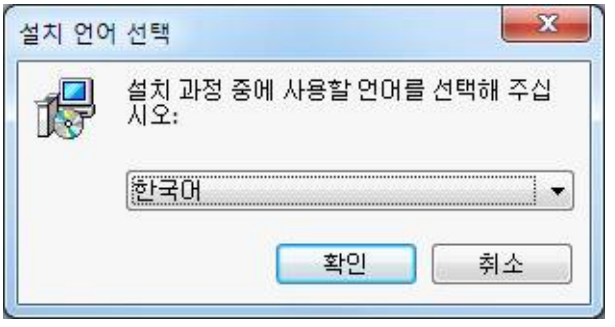


그림 14. 설치 언어 선택

☐ Step 3. 설치 시작 페이지에서 '다음' 버튼을 클릭

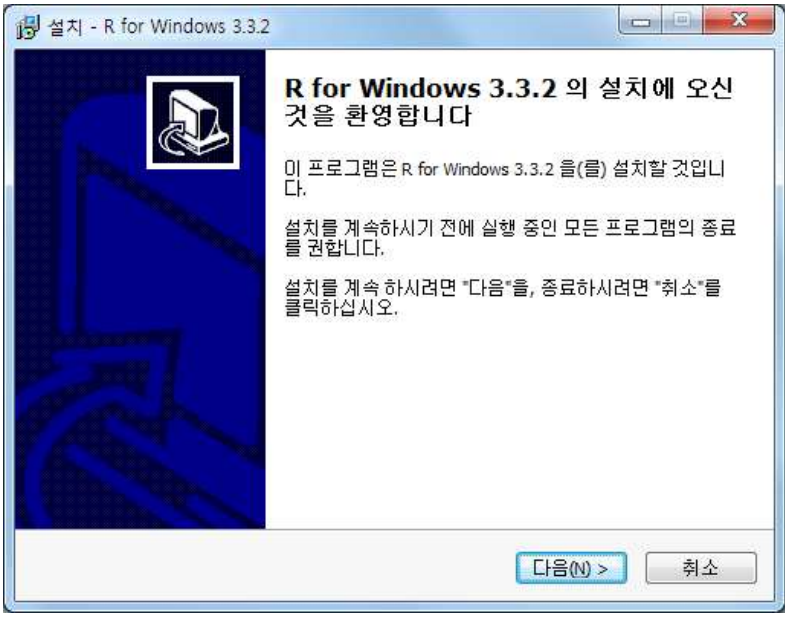


그림 15. R 설치시작페이지

☐ Step 4. 라이센스 페이지에서 '다음' 버튼을 클릭

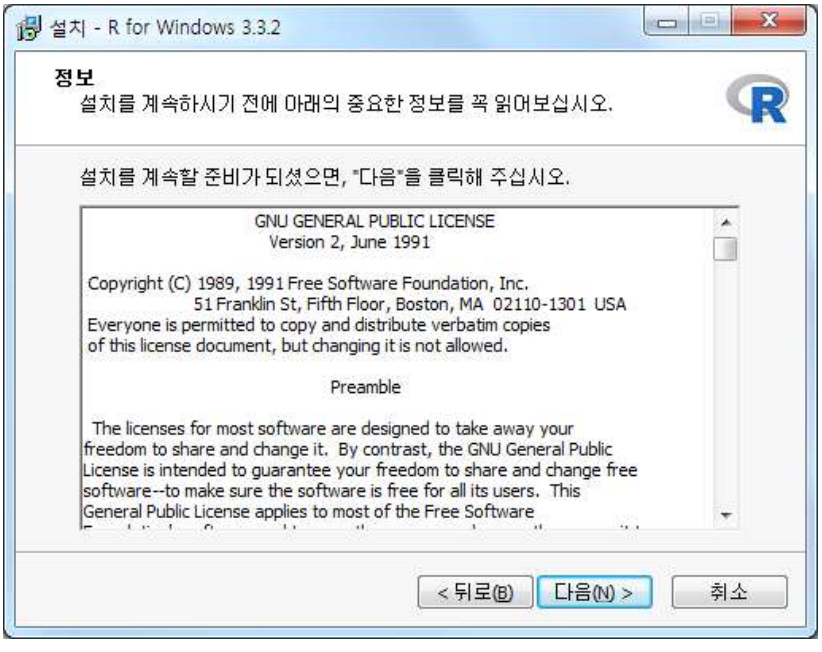


그림 16. R 라이센스 페이지

부 록

□ Step 5. 설치 디렉토리를 설정하고 '다음' 버튼을 클릭

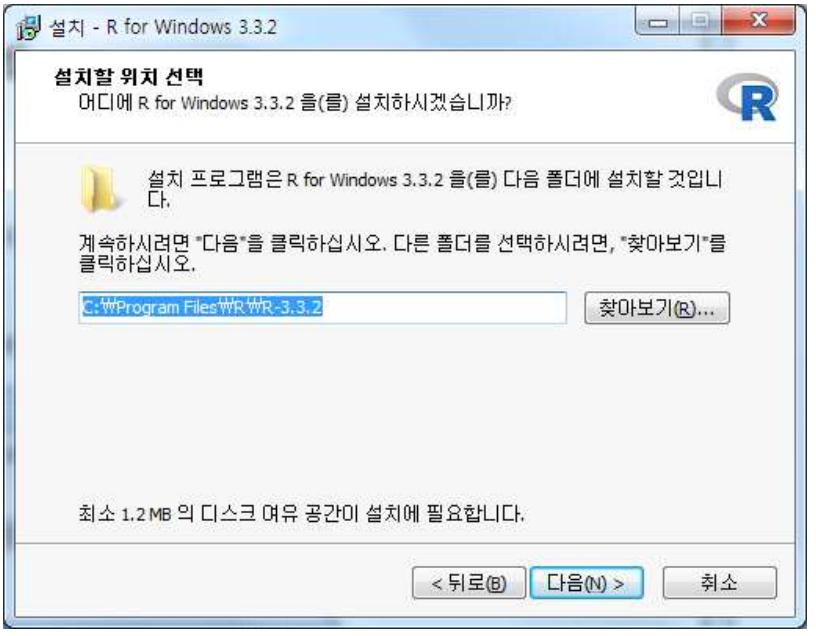


그림 17. R 설치 디렉토리 선택

□ Step 6. 구성 요소 설치 페이지에서 '다음' 버튼을 클릭

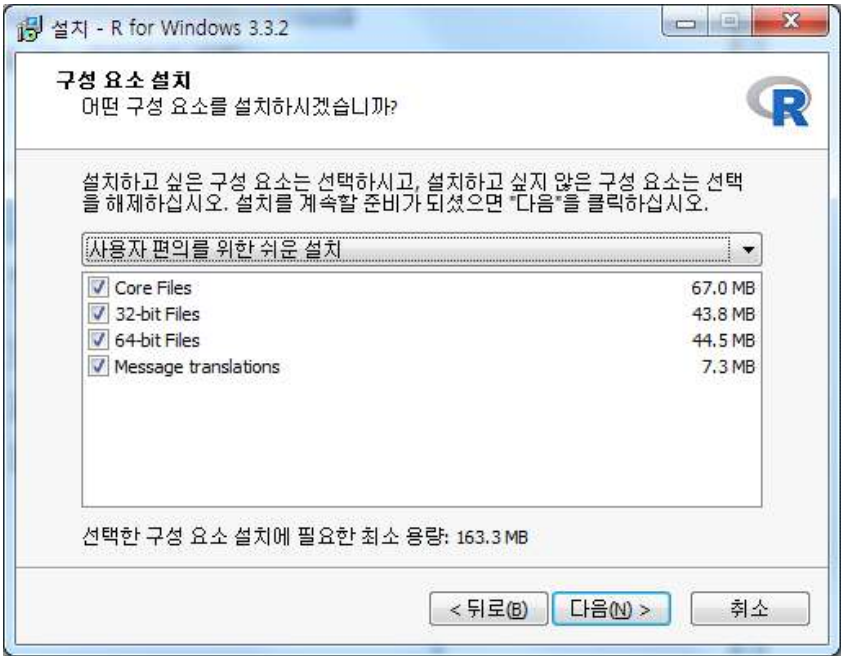


그림 18. R 구성요소 설치 페이지

부록

☐ Step 7. 스타트업 옵션 페이지에서 '다음' 버튼을 클릭

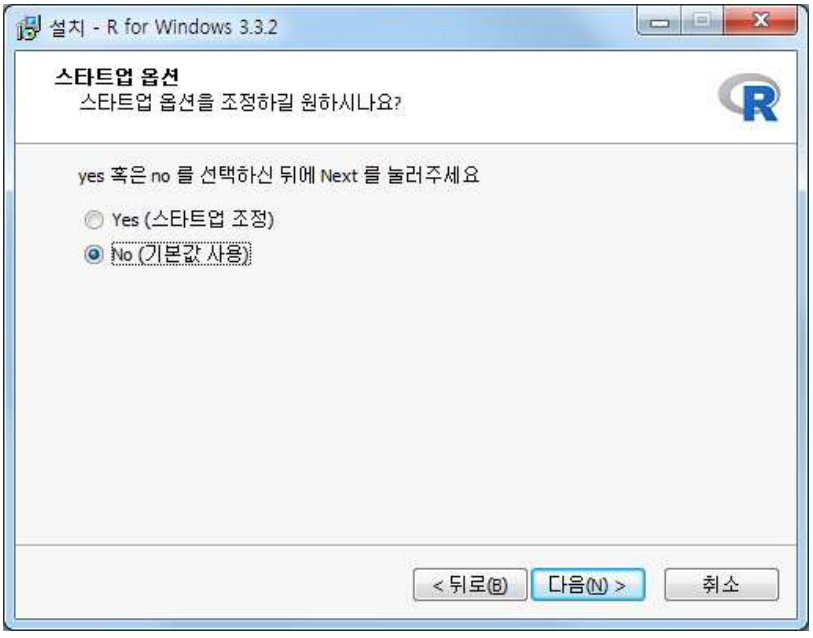


그림 19. R 스타트업 옵션 페이지

☐ Step 8. 추가사항 적용 페이지에서 '다음' 버튼을 클릭

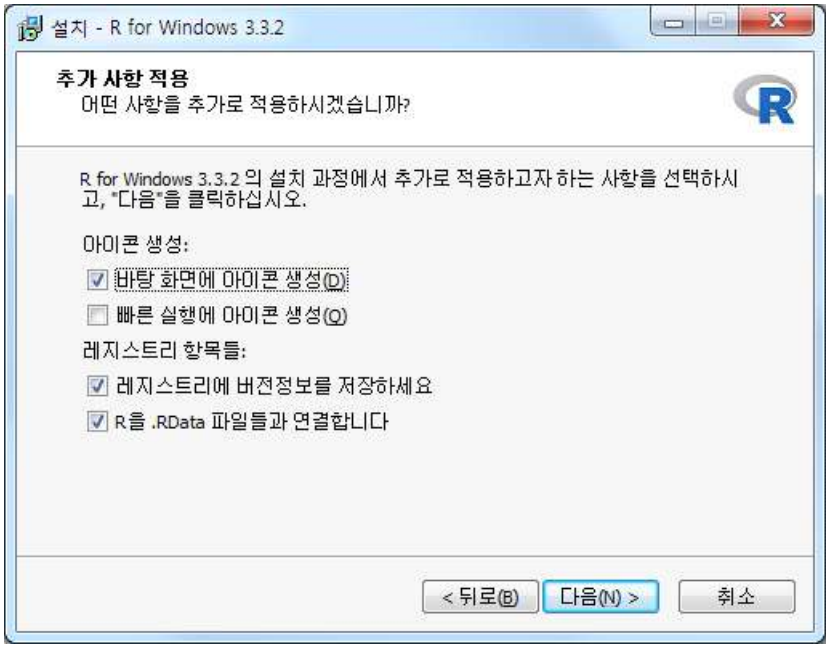


그림 20. R 추가 사항 적용 페이지

□ Step 9. R 설치 진행 후, ‘R 설치 완료’ 창이 뜨면 ‘마침’ 버튼을 클릭하여 설치 완료함

□ Step 10. 설치된 R을 실행시키면 아래 그림과 같이 ‘RGui (R graphic user interface)’ 메인 창 안에 명령어 입력을 기다리는 ‘R Consol’창이 들어있는 형태로 시작되며, 이는 설치가 정상적으로 완료됨

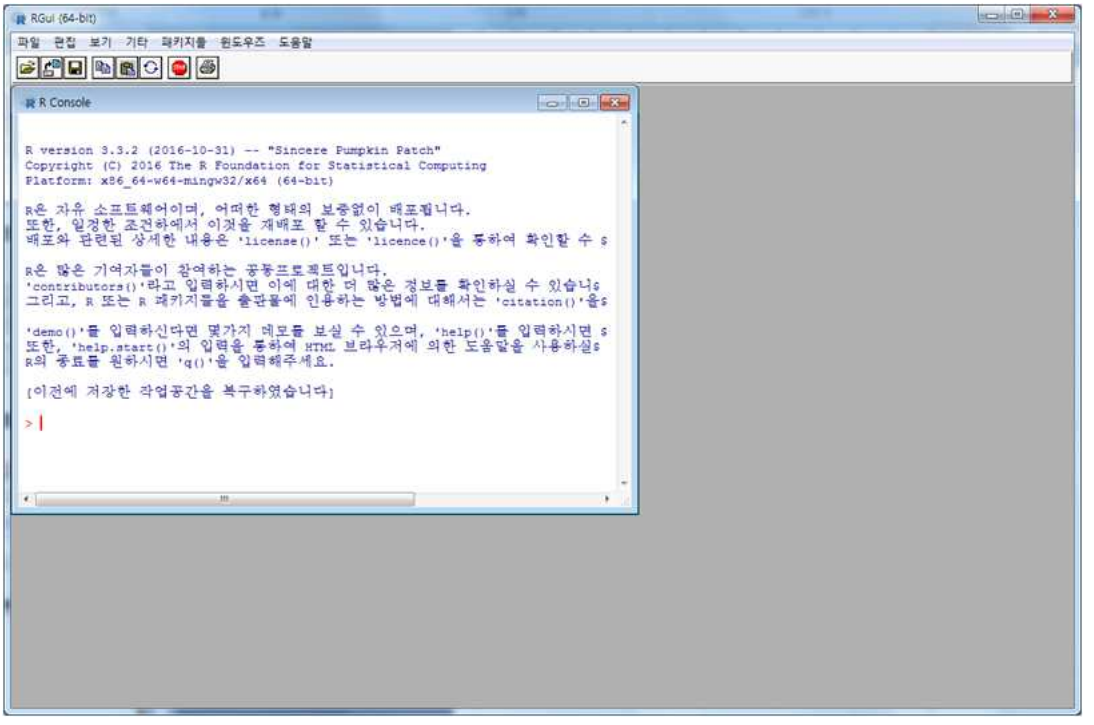


그림 21. R GUI 시작화면

부 록

1.2 R Studio 설치방법

☐ Step 1. 설치파일 실행

- '응용프로그램 설치파일' 폴더에서 컴퓨터 운영체제에 따라 R Studio 설치파일을 실행하고, '다음' 버튼 클릭

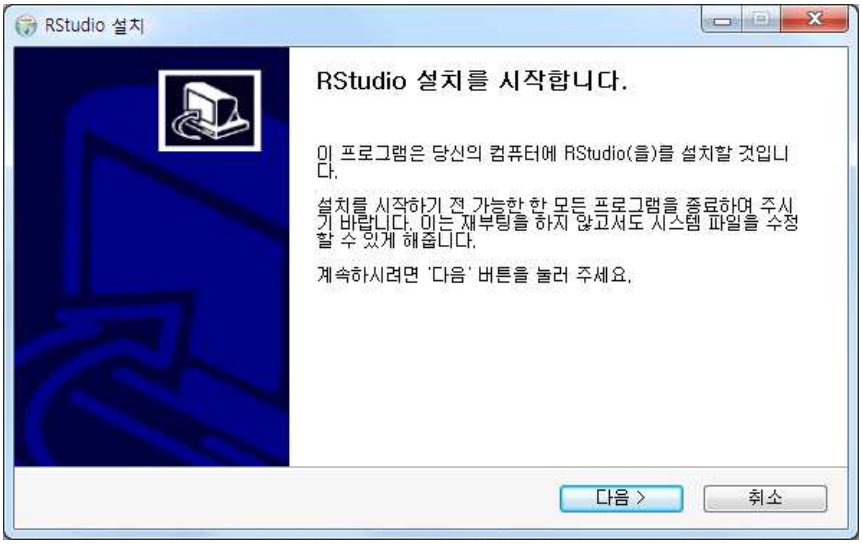


그림 22. R Studio 설치시작페이지

☐ Step 2. 설치파일 실행

- R Studio를 설치 할 폴더를 설정하고 '다음' 버튼 클릭

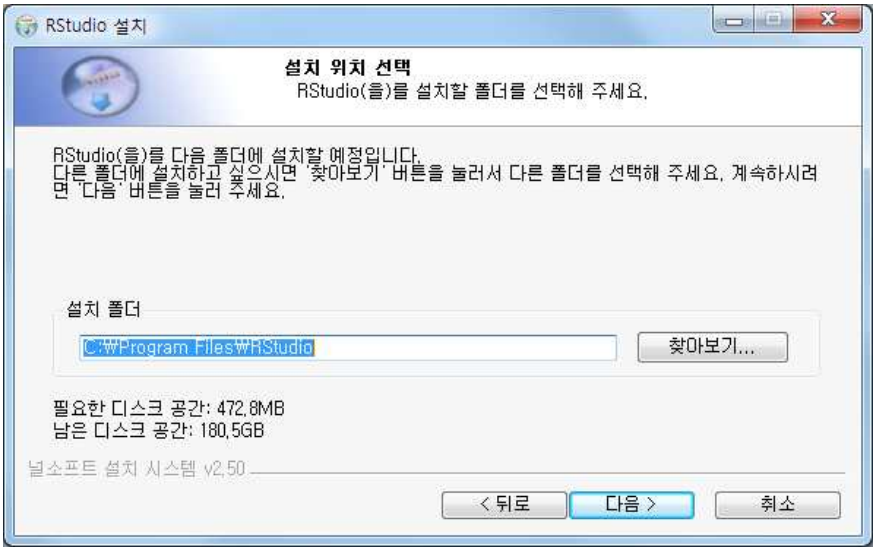


그림 23. R Studio 설치 위치 선택 페이지

☐ Step 3. 시작 메뉴 폴더 설정 후, 설치

- '설치' 버튼을 클릭하여 R studio 설치 진행 후, 'RStudio 설치 완료' 창이 뜨면 '마침' 버튼을 클릭하여 설치 완료함

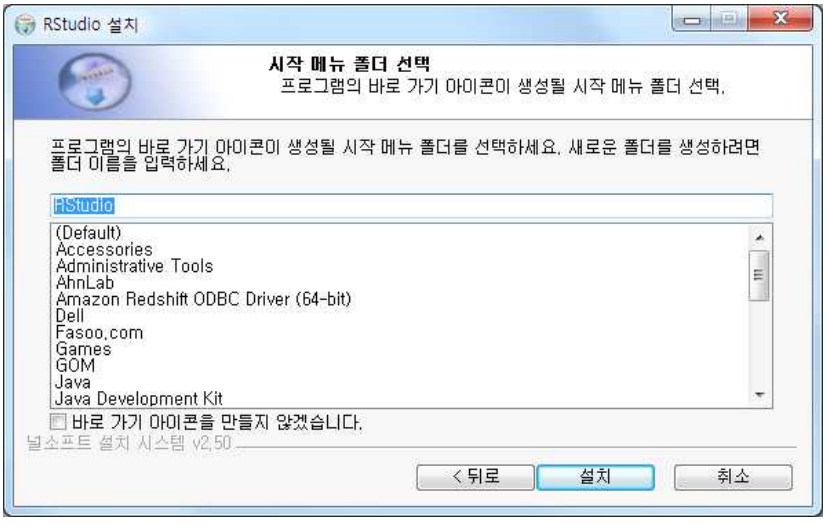


그림 24. R Studio 시작 메뉴 폴더 선택 페이지

☐ Step 4. R Studio 초기화면 확인

- 설치된 R Studio를 실행시키면 아래 그림과 같이 R Studio 초기 화면이 뜨며, 이는 설치가 정상적으로 완료됨

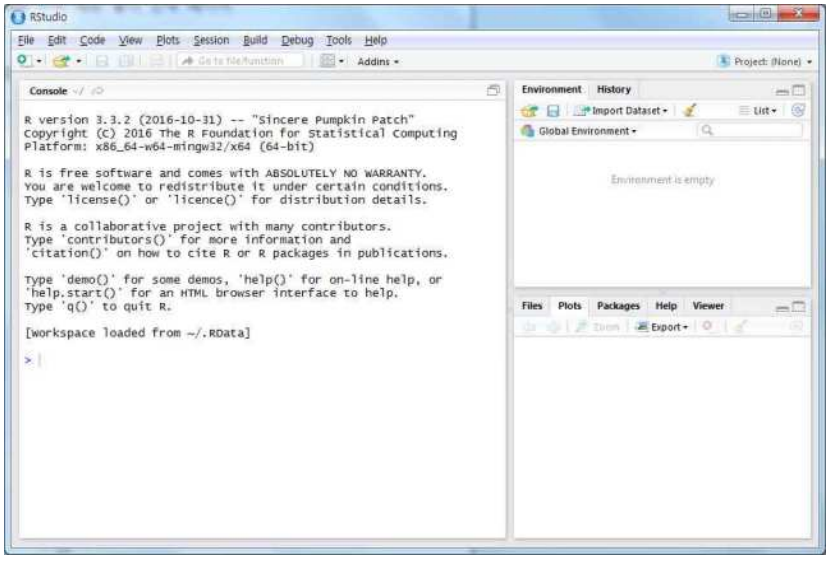


그림 25. R Studio 초기화면

2. R 소스 코드 예시

```
## 교통사고 위험도 ##

# 8) 전체 합산 모델(To the 60%)-------------
rm(list=ls())
source("20161221_icheon_utils_utf8.R")
jpo <- read.csv("grid100_jpo_a10.csv", stringsAsFactors=F,
                    fileEncoding="CP949")
ich <- read.csv("grid100_ich_a10.csv", stringsAsFactors=F,
                    fileEncoding="CP949")
ichs <- read.csv("grid100_ichs_a10.csv", stringsAsFactors=F,
                     fileEncoding="CP949")
ami <- read.csv("grid100_ami_a10.csv", stringsAsFactors=F,
                    fileEncoding="CP949")
sinha <- read.csv("grid100_sinha_a10.csv", stringsAsFactors=F,
                      fileEncoding="CP949")

head(jdata)

jdata$risk_kids <- (jdata$교통_사망*12 + jdata$교통_중상 * 6 +
                      (jdata$교통_경상 + jdata$교통_부상) * 3)/(jdata$교통_사망
+ jdata$교통_중상 +
                                                          jdata$교통_경상      +
jdata$교통_부상)
head(jdata)



for (i in 1:dim(jdata)[2]) {
  idxchk <- which(is.na(jdata[,i]))
  jdata[idxchk,i] <- 0
}
head(jdata)
length(which(jdata$risk_kids > 0))
length(which(jdata$risk_idx > 0))
dim(jdata)

unique(jdata$risk_idx)
```

부 록

```
unique(jdata$risk_kids)
head(jdata)
# save(jdata, file="./정제data/합산_5개_초등학교_g100_데이터명_jdata.RData")
dim(jdata)

library(car)

# (1) 어린이 데이터 위주로 모델 Testing------
head(jdata)
lmfit <- lm(risk_idx ~ 초등학교 + 학원 + 횡단보도 + 방지턱 + cctv +
              교통_성남 + 교통_성여 + 놀이시설 +
              안전_응급 + 안전_준응 + 안전_잠재 +
              + age_0_7_ma + age_0_7_wo + age_8_13_m +
              age_8_13_w + age_14_16_ + age_14_1_1 + age_17_19_ +
age_17_1_1 +
              age_20_29_ + age_20_2_1 + age_30_39_ + age_30_3_1 +
              age_40_49_ + age_40_4_1 + age_50_59_ + age_50_5_1 +
              age_60_69_ + age_60_6_1 + age_70_79_ + age_70_7_1 +
              age_80_89_ + age_80_8_1 + age_90_99_ + age_90_9_1 +
              road_area + road_yn + sum_age_0_ + 참_위험성 +
              참_등교위 + 참_하교위 + 참_활동위 + 참_준경초 +
              참_준경성 + 참_무단유 + 등교경로 + 하교경로 + 활동경로 ,
data=jdata)
summary(lmfit)

durbinWatsonTest(lmfit) # 귀무가설 : 독립적이다. 기각 불가. 고로 독립적이다.
crPlots(lmfit) # 선형성 검증 플롯 : 변수의 갯수만큼 생성되기 때문에 주의해야 한
다.
                # 추후 테스트를 통한 검증 필요
ncvTest(lmfit) # 귀무가설 : 등분산성이다. 기각. 고로 등분산성 가정에 위배된다.
par(mfrow=c(1,1))
spreadLevelPlot(lmfit)

cho_lm_plot(lmfit, f_name="1221_이천초등학교_5개_합계")

steplm <- step(lmfit)

summary(steplm)
```

```
# (2) 사망사고여부 격자 범주형 변수 반영------
head(jdata)

dum <- jdata$교통_사망 + jdata$교통_성사
sum(dum > 0)
sum(jdata$교통_사망 > 0)
sum(jdata$교통_성사 > 0)
jdata$death_yn <- ifelse(dum > 0,1,0)

# save(jdata, file="./정제data/합산_5개_초등학교_g100_데이터명_jdata.RData")

lmfit <- lm(risk_idx ~ 초등학교 + 학원 + 횡단보도 + 방지턱 + cctv +
              교통_성남 + 교통_성여 + 놀이시설 +
              안전_응급 + 안전_준응 + 안전_잠재 + death_yn +
              + age_0_7_ma + age_0_7_wo + age_8_13_m +
              age_8_13_w + age_14_16_ + age_14_1_1 + age_17_19_ +
age_17_1_1 +
              age_20_29_ + age_20_2_1 + age_30_39_ + age_30_3_1 +
              age_40_49_ + age_40_4_1 + age_50_59_ + age_50_5_1 +
              age_60_69_ + age_60_6_1 + age_70_79_ + age_70_7_1 +
              age_80_89_ + age_80_8_1 + age_90_99_ + age_90_9_1 +
              road_area + road_yn + sum_age_0_ + 참_위험성 +
              참_등교위 + 참_하교위 + 참_활동위 + 참_준경초 +
              참_준경성 + 참_무단유 + 등교경로 + 하교경로 + 활동경로 ,
data=jdata)
summary(lmfit)

steplm <- step(lmfit)

summary(steplm)

cho_lm_plot(lmfit, f_name="20161221_사망여부추가_02")


# (3) 설문데이터 정규화(Normalizing)------
head(jdata)
names(jdata)[73:82]
```

```
zdum <- names(jdata)[73:82]
for (z in seq_along(zdum)) {
  assign(paste0("z_",zdum[z]),scale(jdata[,zdum[z]]))
}
# save(jdata, file="./정제data/합산_5개_초등학교_g100_데이터명_jdata.RData")

jdata_z <- cbind(jdata,z_참_등교위, z_참_무단유, z_참_위험성, z_참_준경성,
                 z_참_준경초, z_참_하교위, z_참_활동위, z_등교경로, z_하교경로,
z_활동경로)
head(jdata_z)

lmfit <- lm(risk_idx ~ 초등학교 + 학원 + 횡단보도 + 방지턱 + cctv +
              교통_성남 + 교통_성여 + 놀이시설 +
              안전_응급 + 안전_준응 + 안전_잠재 + death_yn +
              + age_0_7_ma + age_0_7_wo + age_8_13_m +
              age_8_13_w + age_14_16_ + age_14_1_1 + age_17_19_ +
age_17_1_1 +
              age_20_29_ + age_20_2_1 + age_30_39_ + age_30_3_1 +
              age_40_49_ + age_40_4_1 + age_50_59_ + age_50_5_1 +
              age_60_69_ + age_60_6_1 + age_70_79_ + age_70_7_1 +
              age_80_89_ + age_80_8_1 + age_90_99_ + age_90_9_1 +
              road_area + road_yn + sum_age_0_ + z_참_위험성 +
              z_참_등교위 + z_참_하교위 + z_참_활동위 + z_참_준경초 +
              z_참_준경성 + z_참_무단유 + z_등교경로 + z_하교경로 +
              z_활동경로 , data=jdata_z)
summary(lmfit)

steplm <- step(lmfit)

summary(steplm) # 결과에 큰 차이가 없음.

multicol_chk <- cho_lm_plot(lmfit, f_name="20161221_사망여부추가_02")
multicol_chk
head(jdata_z)
lmfit_col <- lm(risk_idx ~ 학원 + 횡단보도 + 방지턱 +
              교통_성남 + 교통_성여 + 놀이시설 +
              안전_응급 + 안전_준응 + 안전_잠재 + death_yn +
              + age_0_7_ma + age_0_7_wo +
```

```
                age_8_13_w + age_14_16_ + age_14_1_1 + age_17_1_1 +
                age_20_29_ + age_20_2_1 + age_30_39_ + age_30_3_1 +
                age_40_49_ + age_70_79_ + age_70_7_1 +
                age_80_89_ + age_80_8_1 + age_90_99_ + age_90_9_1 +
                road_area + road_yn + sum_age_0_ + z_참_위험성 +
                z_참_준경성 + z_등교경로 + z_하교경로, data=jdata_z)
summary(lmfit_col)
cho_lm_plot(lmfit_col, f_name="20161221_사망여부추가_참여형표준화_01")


# (번외) 잠재위험도 산점도 체크----------

jpo10 <- read.csv("grid100_jpo_a19.csv", stringsAsFactors=F,
                  fileEncoding="CP949")
notzero <- which(is.na(jpo10$잠재위험도))
jpo10$잠재위험도[notzero] <- 0
ich10 <- read.csv("grid100_ich_a19.csv", stringsAsFactors=F,
                  fileEncoding="CP949")
notzero <- which(is.na(ich10$잠재위험도))
ich10$잠재위험도[notzero] <- 0
ichs10 <- read.csv("grid100_ichs_a19.csv", stringsAsFactors=F,
                   fileEncoding="CP949")
notzero <- which(is.na(ichs10$잠재위험도))
ichs10$잠재위험도[notzero] <- 0
ami10 <- read.csv("grid100_ami_a19.csv", stringsAsFactors=F,
                  fileEncoding="CP949")
notzero <- which(is.na(ami10$잠재위험도))
ami10$잠재위험도[notzero] <- 0
sinha10 <- read.csv("grid100_sinha_a19.csv", stringsAsFactors=F,
                    fileEncoding="CP949")
notzero <- which(is.na(sinha10$잠재위험도))
sinha10$잠재위험도[notzero] <- 0

library(ggplot2)
school_plt <- ggplot(jpo10, aes(x=잠재위험도, y=risk_idx))
school_plt + geom_point()  + geom_jitter() +
  labs(list(title ="증포초등학교", x = "잠재위험도", y = "교통사고위험도"))
```

```
school_plt <- ggplot(ami10, aes(x=잠재위험도, y=risk_idx))
school_plt + geom_point()  + geom_jitter() +
  labs(list(title ="아미초등학교", x = "잠재위험도", y = "교통사고위험도"))

school_plt <- ggplot(ich10, aes(x=잠재위험도, y=risk_idx))
school_plt + geom_point()  + geom_jitter() +
  labs(list(title ="이천초등학교", x = "잠재위험도", y = "교통사고위험도"))

school_plt <- ggplot(ichs10, aes(x=잠재위험도, y=risk_idx))
school_plt + geom_point()  + geom_jitter() +
  labs(list(title ="이천남 초등학교", x = "잠재위험도", y = "교통사고위험도"))

school_plt <- ggplot(sinha10, aes(x=잠재위험도, y=risk_idx))
school_plt + geom_point()  + geom_jitter() +
  labs(list(title ="신하초등학교", x = "잠재위험도", y = "교통사고위험도"))

icheon_aggr10 <- rbind(jpo10, ich10, ichs10, ami10, sinha10)  #(jpo10, ich10,
ichs10, ami10, sinha10)
idxwhich <- which(is.na(icheon_aggr10$잠재위험도))
icheon_aggr10$잠재위험도[idxwhich] <- 0
notzero <- which(icheon_aggr10$risk_idx > 0)
zero_cham <- which(icheon_aggr10$잠재위험도 == 0)
scatter_data <- icheon_aggr10
dim(icheon_aggr10)

splt <- ggplot(scatter_data, aes(x=잠재위험도, y=risk_idx))
splt + geom_point() + geom_jitter() +
  labs(list(title ="이천시 5개 초등학교 합계", x = "잠재위험도", y = "교통사고위험도
"))




icheon_aggr10 <- rbind(jpo10, ich10, ichs10, ami10, sinha10)  #(jpo10, ich10,
ichs10, ami10, sinha10)
idxwhich <- which(is.na(icheon_aggr10$잠재위험도))
icheon_aggr10$잠재위험도[idx_which] <- 0
notzero <- which(icheon_aggr10$risk_idx > 0)
zero_cham <- which(icheon_aggr10$잠재위험도 == 0)
```

```
scatter_data <- icheon_aggr10[-zero_cham,]
zero_ich <- which(scatter_data$risk_idx == 0)
scatter_data <- scatter_data[-zero_ich,]
dim(scatter_data)

splt <- ggplot(scatter_data, aes(x=잠재위험도, y=risk_idx))
splt + geom_point() + geom_smooth(method="lm")+ geom_jitter() +
  labs(list(title ="5개 초등학교 합계", x = "잠재위험도", y = "교통사고위험도"))

dummy <- icheon_aggr10[,c("risk_idx","잠재위험도")]
dim(dummy)
head(dummy)
dummy$grp <- ifelse(dummy$잠재위험도 >= 0 & dummy$잠재위험도 < 1,
"grp_0",
                ifelse(dummy$잠재위험도 >= 1 & dummy$잠재위험도 < 2,
"grp_1",
                       ifelse(dummy$잠재위험도 >= 2 & dummy$잠재위험도 < 3,
"grp_2",
                              ifelse(dummy$잠재위험도 >= 3 & dummy$잠재위험
도 < 4, "grp_3","grp_4"))))
dummy$grp2 <- ifelse(dummy$risk_idx == 0, 0,1)
head(dummy)
library(dplyr)
smry <- dummy %>%
  group_by(grp, grp2) %>%
  summarize(N=n())
smry
library(reshape2)
smry_cast <- dcast(smry, grp~grp2, value.var="N", fun.aggregate=sum)
smry_cast
names(smry_cast) <- c("grp","None","exist")
plt <- ggplot(smry_cast, aes(x=grp, y=exist))
plt + geom_bar(stat="identity")+
  labs(list(title ="잠재위험도 구간별 교통 사고건수", x = "잠재위험도 그룹", y = "교
통사고건수"))
```

```
## 분석 모델 다중공선성 검증 ##

cho_lm_plot <- function(fit, wide=512, high=512, f_name) {#}
  library(car)
  par(mfrow=c(2,2))
  print(plot(fit))
  png(paste0("./IMG/",f_name,".png"), width=wide, height=high)
  par(mfrow=c(2,2))
  print(plot(fit))
  par(mfrow=c(1,1))
  dev.off()
  par(mfrow=c(1,1))

  steplm <- fit
  par(mfrow=c(1,3))
  print(qqPlot(steplm, main="QQ Plot", simulate=T, id.method="identity"))
  library(MASS)
  sresid <- studres(steplm)
  print(hist(sresid, freq=FALSE, main="표준화 잔차 분포"))
  xfit<-seq(min(sresid),max(sresid),length=40)
  yfit<-dnorm(xfit)
  print(lines(xfit, yfit))
  vif(steplm)# variance inflation factors
  vif_sqrt_chk <- sqrt(vif(steplm)) > 2 # problem?
  vif_10_chk <- vif(steplm) > 10
  print(plot(vif(steplm), xaxt="n", ylim=c(0,12), xlab="", ylab="분산확장지수(VIF)",
             main="다중 공선성 검정", type="l"))
  print(axis(1, at=1:length(names(vif(steplm))), labels=names(vif(steplm)), las=2))
  print(abline(h=10, lwd=2, col="red"))

  png(paste0("./IMG/",f_name,"_다중공선성",".png"), width=wide*2, height=high)
  par(mfrow=c(1,3))
  print(qqPlot(steplm, main="QQ Plot"))
  library(MASS)
  sresid <- studres(steplm)
  print(hist(sresid, freq=FALSE, main="표준화 잔차 분포"))
  xfit<-seq(min(sresid),max(sresid),length=40)
  yfit<-dnorm(xfit)
```

```
  print(lines(xfit, yfit))
  vif(steplm)# variance inflation factors
  vif_sqrt_chk <- sqrt(vif(steplm)) > 2 # problem?
  vif_10_chk <- vif(steplm) > 10
  print(plot(vif(steplm), xaxt="n", ylim=c(0,12), xlab="", ylab="분산확장지수(VIF)",
             main="다중 공선성 검정", type="l"))
  print(axis(1, at=1:length(names(vif(steplm))), labels=names(vif(steplm)), las=2))
  print(abline(h=10, lwd=2, col="red"))
  dev.off()
  par(mfrow=c(1,1))

  return(list(vif_sqrt_chk, vif_10_chk))
}



## 분석모델 - 공공데이터 및 참여형데이터 반영 ##

# rm(list=ls())
# getwd()
# setwd("/exthome/extA/selfRule/")
###
# 2. 이천시 초등학교 인근 100m*100m 격자 데이터 (Workshop_잠정최종모델)
#    - 경기도 이천시 사고 사상자 데이터중에서 해당 초등학교별 인근의
#      어린이 및 성인 사고 관련부분을
#      합산하여 활용함.
#

source("20161221_icheon_utils_utf8.R")



# 1) 증포초---------
# 1) 데이터 일괄 로딩
# 참여형 데이터를 포함한 모든 컬럼을 적용한 DB를 활용

jpo <- read.csv("grid100_jpo_a10.csv", stringsAsFactors=F,
                fileEncoding="CP949")
```

```
ich <- read.csv("grid100_ich_a10.csv", stringsAsFactors=F,
                fileEncoding="CP949")
ichs <- read.csv("grid100_ichs_a10.csv", stringsAsFactors=F,
                 fileEncoding="CP949")
ami <- read.csv("grid100_ami_a10.csv", stringsAsFactors=F,
                fileEncoding="CP949")
sinha <- read.csv("grid100_sinha_a10.csv", stringsAsFactors=F,
                  fileEncoding="CP949")

# 코드 재사용 활용성 제고를 위한 데이터 변수명 통일
jdata <- jpo

for (i in 1:dim(jdata)[2]) {
  idxchk <- which(is.na(jdata[,i]))
  jdata[idxchk,i] <- 0
}
head(jdata)

unique(jdata$risk_idx)
head(jdata)
# save(jdata, file="./정제data/증포초_g100_데이터명_jdata.RData")
dim(jdata)
sum(jdata$risk_idx > 0 )
sum(jdata$risk_idx_kids > 0 )

dum <- jdata$교통_사망 + jdata$교통_성사
sum(dum > 0)
sum(jdata$교통_사망 > 0)
sum(jdata$교통_성사 > 0)
jdata$death_yn <- ifelse(dum > 0,1,0)
head(jdata)

library(car)

# 참여형 데이터 반영모델 (어린이+성인 교통사고 모두 위험도 계산에 포함)
head(jdata)
lmfit <- lm(risk_idx ~ 초등학교 + 학원 + 횡단보도 + 방지턱 + cctv +
              교통_성남 + 교통_성여 +
```

```
              안전_응급 + 안전_준응 + 안전_잠재 + death_yn +
              놀이시설 + + age_0_7_ma + age_0_7_wo + age_8_13_m +
              age_8_13_w + age_14_16_ + age_14_1_1 + age_17_19_ +
age_17_1_1 +
              age_20_29_ + age_20_2_1 + age_30_39_ + age_30_3_1 +
              age_40_49_ + age_40_4_1 + age_50_59_ + age_50_5_1 +
              age_60_69_ + age_60_6_1 + age_70_79_ + age_70_7_1 +
              age_80_89_ + age_80_8_1 + age_90_99_ + age_90_9_1 +
              road_area + road_yn + sum_age_0_ + 참_위험성 +
              참_등교위 + 참_하교위 + 참_활동위 + 참_준경초 +
              참_준경성 + 참_무단유 + 등교경로 + 하교경로 + 활동경로 ,
data=jdata)
summary(lmfit)

steplm1 <- step(lmfit)

summary(steplm1)

cho_lm_plot(steplm1, f_name="증포초등학교")

lmfit2 <- lm(risk_idx ~ 횡단보도 +
               교통_성남 + 교통_성여 +
               death_yn +
               age_70_7_1 +
               age_80_89_ +
               road_area + 참_위험성 + 참_준경초
             , data=jdata)
summary(lmfit2)

steplm1_2 <- step(lmfit2)

summary(steplm1_2)

cho_lm_plot(steplm1_2, f_name="증포초등학교_공선성제거")

# [[2]]
# 횡단보도     방지턱  교통_성남  교통_성여   놀이시설  안전_잠재   death_yn
# FALSE       FALSE       FALSE       FALSE       FALSE       FALSE
```

```
FALSE
# age_80_89_ age_90_99_  road_area  참_위험성  참_준경초
# FALSE      FALSE       FALSE      FALSE      FALSE

# 2) 아미초---------
# 1) 데이터 로딩
# 참여형 데이터를 포함한 모든 컬럼을 적용한 DB를 활용

jdata <- ami

for (i in 1:dim(jdata)[2]) {
  idxchk <- which(is.na(jdata[,i]))
  jdata[idxchk,i] <- 0
}
head(jdata)

unique(jdata$risk_idx)
head(jdata)
# save(jdata, file="./정제data/아미초_g100_데이터명_jdata.RData")
dim(jdata)
sum(jdata$risk_idx > 0 )
sum(jdata$risk_idx_kids > 0 )

dum <- jdata$교통_사망 + jdata$교통_성사
sum(dum > 0)
sum(jdata$교통_사망 > 0)
sum(jdata$교통_성사 > 0)
jdata$death_yn <- ifelse(dum > 0,1,0)
head(jdata)

library(car)

# 참여형 데이터 반영모델 (어린이+성인 교통사고 모두 위험도 계산에 포함)
head(jdata)
lmfit <- lm(risk_idx ~ 초등학교 + 학원 + 횡단보도 + 방지턱 + cctv +
              교통_성남 + 교통_성여 +
              안전_응급 + 안전_준응 + 안전_잠재 + death_yn +
              놀이시설 + + age_0_7_ma + age_0_7_wo + age_8_13_m +
```

```
              age_8_13_w  +  age_14_16_  +  age_14_1_1  +  age_17_19_  +
age_17_1_1 +
              age_20_29_ + age_20_2_1 + age_30_39_ + age_30_3_1 +
              age_40_49_ + age_40_4_1 + age_50_59_ + age_50_5_1 +
              age_60_69_ + age_60_6_1 + age_70_79_ + age_70_7_1 +
              age_80_89_ + age_80_8_1 + age_90_99_ + age_90_9_1 +
              road_area + road_yn + sum_age_0_ + 참_위험성 +
              참_등교위 + 참_하교위 + 참_활동위 + 참_준경초 +
              참_준경성  +  참_무단유  +  등교경로  +  하교경로  +  활동경로  ,
data=jdata)
summary(lmfit)

steplm2 <- step(lmfit)

summary(steplm2)

cho_lm_plot(steplm2, f_name="아미초등학교")

lmfit2 <- lm(risk_idx ~
               교통_성여 +
               안전_응급 + 안전_준응 + death_yn +
               참_하교위 + 참_준경초
             , data=jdata)
summary(lmfit2)

steplm2_2 <- step(lmfit2)

summary(steplm2_2)

cho_lm_plot(steplm2_2, f_name="아미초등학교_공선성제거")


# 3) 이천초---------
# 1) 데이터 로딩
# 참여형 데이터를 포함한 모든 컬럼을 적용한 DB를 활용


jdata <- ich
```

```
for (i in 1:dim(jdata)[2]) {
  idxchk <- which(is.na(jdata[,i]))
  jdata[idxchk,i] <- 0
}
head(jdata)

unique(jdata$risk_idx)
head(jdata)
# save(jdata, file="./정제data/이천초_g100_데이터명_jdata.RData")
dim(jdata)
sum(jdata$risk_idx > 0 )
sum(jdata$risk_idx_kids > 0 )

dum <- jdata$교통_사망 + jdata$교통_성사
sum(dum > 0)
sum(jdata$교통_사망 > 0)
sum(jdata$교통_성사 > 0)
jdata$death_yn <- ifelse(dum > 0,1,0)
head(jdata)

library(car)

# 참여형 데이터 반영모델 (어린이+성인 교통사고 모두 위험도 계산에 포함)
head(jdata)
lmfit <- lm(risk_idx ~ 초등학교 + 학원 + 횡단보도 + 방지턱 + cctv +
              교통_성남 + 교통_성여 +
              안전_응급 + 안전_준응 + 안전_잠재 + death_yn +
              놀이시설 + + age_0_7_ma + age_0_7_wo + age_8_13_m +
              age_8_13_w + age_14_16_ + age_14_1_1 + age_17_19_ +
age_17_1_1 +
              age_20_29_ + age_20_2_1 + age_30_39_ + age_30_3_1 +
              age_40_49_ + age_40_4_1 + age_50_59_ + age_50_5_1 +
              age_60_69_ + age_60_6_1 + age_70_79_ + age_70_7_1 +
              age_80_89_ + age_80_8_1 + age_90_99_ + age_90_9_1 +
              road_area + road_yn + sum_age_0_ + 참_위험성 +
              참_등교위 + 참_하교위 + 참_활동위 + 참_준경초 +
              참_준경성 + 참_무단유 + 등교경로 + 하교경로 + 활동경로 ,
```

```
data=jdata)
summary(lmfit)

steplm3 <- step(lmfit)

summary(steplm3)

cho_lm_plot(steplm3, f_name="이천초등학교")

lmfit2 <- lm(risk_idx ~ 학원 +
               교통_성남 + 교통_성여 + 놀이시설 + 안전_잠재 + death_yn +
               age_80_89_ + age_80_8_1 + age_90_9_1 +
               road_area + road_yn + sum_age_0_ + 참_위험성 +
               참_등교위 + 등교경로
             , data=jdata)
summary(lmfit2)

steplm3_2 <- step(lmfit2)

summary(steplm3_2)

cho_lm_plot(steplm3_2, f_name="이천초등학교_공선성제거")

# 4) 이천남초---------
# 1) 데이터 로딩
# 참여형 데이터를 포함한 모든 컬럼을 적용한 DB를 활용

jdata <- ichs

for (i in 1:dim(jdata)[2]) {
  idxchk <- which(is.na(jdata[,i]))
  jdata[idxchk,i] <- 0
}
head(jdata)

unique(jdata$risk_idx)
head(jdata)
# save(jdata, file="./정제data/이천남초_g100_데이터명_jdata.RData")
```

```
dim(jdata)
sum(jdata$risk_idx > 0 )
sum(jdata$risk_idx_kids > 0 )

dum <- jdata$교통_사망 + jdata$교통_성사
sum(dum > 0)
sum(jdata$교통_사망 > 0)
sum(jdata$교통_성사 > 0)
jdata$death_yn <- ifelse(dum > 0,1,0)
head(jdata)

library(car)

# 참여형 데이터 반영모델 (어린이+성인 교통사고 모두 위험도 계산에 포함)
head(jdata)
lmfit <- lm(risk_idx ~ 초등학교 + 학원 + 횡단보도 + 방지턱 + cctv +
              교통_성남 + 교통_성여 +
              안전_응급 + 안전_준응 + 안전_잠재 + death_yn +
              놀이시설 + + age_0_7_ma + age_0_7_wo + age_8_13_m +
              age_8_13_w + age_14_16_ + age_14_1_1 + age_17_19_ +
age_17_1_1 +
              age_20_29_ + age_20_2_1 + age_30_39_ + age_30_3_1 +
              age_40_49_ + age_40_4_1 + age_50_59_ + age_50_5_1 +
              age_60_69_ + age_60_6_1 + age_70_79_ + age_70_7_1 +
              age_80_89_ + age_80_8_1 + age_90_99_ + age_90_9_1 +
              road_area + road_yn + sum_age_0_ + 참_위험성 +
              참_등교위 + 참_하교위 + 참_활동위 + 참_준경초 +
              참_준경성 + 참_무단유 + 등교경로 + 하교경로 + 활동경로 ,
data=jdata)
summary(lmfit)

steplm4 <- step(lmfit)

summary(steplm4)

cho_lm_plot(steplm4, f_name="이천남초등학교")

lmfit2 <- lm(risk_idx ~ 학원 + 방지턱 +
```

```
                교통_성남 + 교통_성여 +
                death_yn +
                road_area + road_yn + sum_age_0_ + 참_위험성 +
                참_등교위 + 참_하교위 +
                참_준경성
              , data=jdata)
summary(lmfit2)

steplm4_2 <- step(lmfit2)

summary(steplm4_2)

cho_lm_plot(steplm4_2, f_name="이천남초등학교_공선성제거")

# 5) 신하초---------
# 1) 데이터 로딩
# 참여형 데이터를 포함한 모든 컬럼을 적용한 DB를 활용

jdata <- sinha

for (i in 1:dim(jdata)[2]) {
  idxchk <- which(is.na(jdata[,i]))
  jdata[idxchk,i] <- 0
}
head(jdata)

unique(jdata$risk_idx)
head(jdata)
# save(jdata, file="./정제data/신하초_g100_데이터명_jdata.RData")
dim(jdata)
sum(jdata$risk_idx > 0 )
sum(jdata$risk_idx_kids > 0 )

dum <- jdata$교통_사망 + jdata$교통_성사
sum(dum > 0)
sum(jdata$교통_사망 > 0)
sum(jdata$교통_성사 > 0)
jdata$death_yn <- ifelse(dum > 0,1,0)
```

```
head(jdata)

library(car)

# 참여형 데이터 반영모델 (어린이+성인 교통사고 모두 위험도 계산에 포함)
head(jdata)
head(jdata)
jdata$ln_risk_idx <- log(jdata$risk_idx + 1)
lmfit <- lm(ln_risk_idx ~ 초등학교 + 학원 + 횡단보도 + 방지턱 + cctv +
              교통_성남 + 교통_성여 +
              안전_응급 + 안전_준응 + 안전_잠재 + death_yn +
              놀이시설 + age_0_7_ma + age_0_7_wo + age_8_13_m +
              age_8_13_w  +  age_14_16_  +  age_14_1_1  +  age_17_19_  +
age_17_1_1 +
              age_20_29_ + age_20_2_1 + age_30_39_ + age_30_3_1 +
              age_40_49_ + age_40_4_1 + age_50_59_ + age_50_5_1 +
              age_60_69_ + age_60_6_1 + age_70_79_ + age_70_7_1 +
              age_80_89_ + age_80_8_1 + age_90_99_ + age_90_9_1 +
              road_area + road_yn + sum_age_0_ + 참_위험성 +
              참_등교위 + 참_하교위 + 참_활동위 + 참_준경초 +
              참_준경성 + 참_무단유 + 등교경로 + 하교경로 + 활동경로
            , data=jdata)
summary(lmfit)

steplm5 <- step(lmfit)

summary(steplm5)

cho_lm_plot(steplm5, f_name="신하초등학교")

lmfit2 <- lm(risk_idx ~
               교통_성남 + 놀이시설 +
               안전_준응 + death_yn +
               age_80_89_ + age_80_8_1 +
               road_area + road_yn + sum_age_0_ +
               참_준경초
             , data=jdata)
summary(lmfit2)
```

```
steplm5_2 <- step(lmfit2)

summary(steplm5_2)

cho_lm_plot(steplm5_2, f_name="신하초등학교_공선성제거")

# 7) 전체 합산 모델----------

five_elementary <- rbind(jpo,ami,ich,ichs,sinha)

jdata <- five_elementary
dim(jdata)

for (i in 1:dim(jdata)[2]) {
  idxchk <- which(is.na(jdata[,i]))
  jdata[idxchk,i] <- 0
}
head(jdata)
dim(jdata)

unique(jdata$risk_idx)
head(jdata)
# save(jdata, file="./정제data/합산_5개_초등학교_g100_데이터명_jdata.RData")
dim(jdata)
sum(jdata$risk_idx > 0 )

dum <- jdata$교통_사망 + jdata$교통_성사
sum(dum > 0)
sum(jdata$교통_사망 > 0)
sum(jdata$교통_성사 > 0)
jdata$death_yn <- ifelse(dum > 0,1,0)
unique(jdata$road_yn)
head(jdata)

library(car)

# 참여형 데이터 반영모델 (어린이+성인 교통사고 모두 위험도 계산에 포함)
```

```
head(jdata)

lmfit <- lm(risk_idx ~ 초등학교 + 학원 + 횡단보도 + 방지턱 + cctv +
              교통_성남 + 교통_성여 + 놀이시설 + death_yn +
              age_0_7_ma + age_0_7_wo + age_8_13_m +
              age_8_13_w + age_14_16_ + age_14_1_1 + age_17_19_ +
age_17_1_1 +
              age_20_29_ + age_20_2_1 + age_30_39_ + age_30_3_1 +
              age_40_49_ + age_40_4_1 + age_50_59_ + age_50_5_1 +
              age_60_69_ + age_60_6_1 + age_70_79_ + age_70_7_1 +
              age_80_89_ + age_80_8_1 + age_90_99_ + age_90_9_1 +
              road_area + road_yn + sum_age_0_ + 참_위험성 +
              참_등교위 + 참_하교위 + 참_활동위 + 참_준경초 +
              참_준경성 + 참_무단유 + 등교경로 + 하교경로 + 활동경로 ,
data=jdata)
summary(lmfit)

steplmA <- step(lmfit)

summary(steplmA)

cho_lm_plot(steplmA, f_name="초등_5개_전체")

lmfit2 <- lm(risk_idx ~
              교통_성남 + 안전_준응 + death_yn +
              age_20_2_1 + age_30_39_ +
              age_80_8_1 +
              road_area + 참_준경초, data=jdata)
summary(lmfit2)

steplmA_2 <- step(lmfit2)

summary(steplmA_2)

cho_lm_plot(steplmA_2, f_name="초등_5개_전체_공선성제거")

summary(steplm1_2)
cho_lm_plot(steplm1_2, f_name="증포초_공선성제거")
```

```
summary(steplm2_2)
cho_lm_plot(steplm1_2, f_name="아미초_공선성제거")
summary(steplm3_2)
cho_lm_plot(steplm3_2, f_name="이천초_공선성제거")
summary(steplm4_2)
cho_lm_plot(steplm4_2, f_name="이천남초_공선성제거")
summary(steplm5_2)
cho_lm_plot(steplm5_2, f_name="신하초_공선성제거")
summary(steplmA_2)
cho_lm_plot(steplmA_2, f_name="초등_5개_전체_공선성제거")

# outlier 제거--------
outlier_all_elementary <- outlierTest(steplm, n.max = 40)
str(outlier_all_elementary)
del_idx <- attr(outlier_all_elementary$bonf.p,"names")
mean(jdata[del_idx,"risk_idx"])
mean(jdata[,"risk_idx"])
length(del_idx)

jdata_fixed <- jdata[-as.numeric(del_idx),]

# outlier 제거-----
lmfit_fixed <- lm(risk_idx ~ 초등학교 + 학원 + 횡단보도 + 방지턱 + cctv +
                    교통_성남 + 교통_성여 + 놀이시설 + death_yn +
                    age_0_7_ma + age_0_7_wo + age_8_13_m +
                    age_8_13_w + age_14_16_ + age_14_1_1 + age_17_19_ +
age_17_1_1 +
                    age_20_29_ + age_20_2_1 + age_30_39_ + age_30_3_1 +
                    age_40_49_ + age_40_4_1 + age_50_59_ + age_50_5_1 +
                    age_60_69_ + age_60_6_1 + age_70_79_ + age_70_7_1 +
                    age_80_89_ + age_80_8_1 + age_90_99_ + age_90_9_1 +
                    road_area + road_yn + sum_age_0_ + 참_위험성 +
                    참_등교위 + 참_하교위 + 참_활동위 + 참_준경초 +
                    참_준경성 + 참_무단유 + 등교경로 + 하교경로 + 활동경로
              , data=jdata_fixed)
summary(lmfit_fixed)

steplm_fixed <- step(lmfit_fixed)
```

```
summary(steplm_fixed)

anova(steplm_fixed)

cho_lm_plot(steplm_fixed, f_name="이상치제거_5개_학교_전체")

lmfit_fixed2 <- lm(risk_idx ~ 초등학교 + 학원 + 횡단보도 + 방지턱 + cctv +
                     교통_성남 + 교통_성여 + 놀이시설 + death_yn +
                     age_0_7_ma + age_0_7_wo +
                     age_14_16_ + age_14_1_1 + age_17_19_ + age_17_1_1 +
                     age_20_29_ + age_30_3_1 +
                     age_40_49_ + age_40_4_1 + age_50_59_ +
                     age_60_6_1 + age_70_79_ + age_70_7_1 +
                     age_80_89_ + age_80_8_1 + age_90_99_ + age_90_9_1 +
                     road_area + road_yn + sum_age_0_+
                     등교경로 + 하교경로
                   , data=jdata_fixed)
summary(lmfit_fixed2)

steplm_fixed2 <- step(lmfit_fixed2)

summary(steplm_fixed2)

anova(steplm_fixed)




## 이천초등학교 EDA 예시 ##

icheon<-read.csv("d:/00-gg/참여형기초통계_학교별/이천초등학교.csv")

icheon<-icheon[,c(5:8)]
library(dplyr)
icheon_popu_tbl <- tbl_df(icheon)

#학년별 성별
```

```
icheon_popu_tbl_g1 <- group_by(icheon_popu_tbl,p0102,p0103)
icheon_popu_tbl_s1 <- summarise(icheon_popu_tbl_g1, count = n())
icheon_popu_df1<-as.data.frame(icheontb2y2c)
write.csv(icheon_popu_df1,"d:/00-gg/참여형기초통계_학교별/output/이천초_학년별성별.csv")

#학년별 가구원 수
icheon_popu_tbl_g2 <- group_by(icheon_popu_tbl,p0102,p0104)
icheon_popu_tbl_s2 <- summarise(icheon_popu_tbl_g2, count = n())
icheon_popu_df2<-as.data.frame(icheon_popu_tbl_s2)
write.csv(icheontb2y32df,"d:/00-gg/참여형기초통계_학교별/output/이천남초_학년별성별.csv")

#학년별조부모동거여부
icheon_popu_tbl_g3<- group_by(icheon_popu_tbl,p0102,p0105)
icheon_popu_tbl_s3 <- summarise(icheon_popu_tbl_g3, count = n())
icheon_popu_df3<-as.data.frame(icheon_popu_tbl_s3)
write.csv(ic_y4coun2_df,"d:/00-gg/참여형기초통계_학교별/이천초조부모동거여부.csv")

# 어린이 행동특성 질문
icheon_behav<-icheon[,c('p0105','k0501','k0502','k0503','k0504','k0505','k0506',
'k0507','k0508','k0509','k0510','k0511','k0512','k0513')]
icheon_behav_tbl <- tbl_df(icheon_behav)

#조부모와 동거여부별 행동 특성 질문1에 대한 응답
icheon_behav_tbl_g1<- group_by(icheon_behav_tbl,p0105,k0501)
icheon_behav_tbl_s1 <- summarise(icheon_behav_tbl_g1, count = n())
(icheon_behav_tbl_s1)

#조부모와 동거여부별 행동 특성 질문2에 대한 응답
icheon_behav_tbl_g2<- group_by(icheon_behav_tbl,p0105,k0502)
icheon_behav_tbl_s2 <- summarise(icheon_behav_tbl_g2, count = n())
(icheon_behav_tbl_s2)

#조부모와 동거여부별 행동 특성 질문3에 대한 응답
icheon_behav_tbl_g3<- group_by(icheon_behav_tbl,p0105,k0503)
icheon_behav_tbl_s3 <- summarise(icheon_behav_tbl_g3, count = n())
```

```
(icheon_behav_tbl_s3)

#조부모와 동거여부별 행동 특성 질문5에 대한 응답
icheon_behav_tbl_g5<- group_by(icheon_behav_tbl,p0105,k0505)
icheon_behav_tbl_s5 <- summarise(icheon_behav_tbl_g5, count = n())
(icheon_behav_tbl_s5)

#조부모와 동거여부별 행동 특성 질문6에 대한 응답
icheon_behav_tbl_g6<- group_by(icheon_behav_tbl,p0105,k0506)
icheon_behav_tbl_s6 <- summarise(icheon_behav_tbl_g6, count = n())
(icheon_behav_tbl_s6)

#조부모와 동거여부별 행동 특성 질문7에 대한 응답
icheon_behav_tbl_g7<- group_by(icheon_behav_tbl,p0105,k0507)
icheon_behav_tbl_s7 <- summarise(icheon_behav_tbl_g7, count = n())
(icheon_behav_tbl_s7)

#조부모와 동거여부별 행동 특성 질문8에 대한 응답
icheon_behav_tbl_g8<- group_by(icheon_behav_tbl,p0105,k0508)
icheon_behav_tbl_s8 <- summarise(icheon_behav_tbl_g8, count = n())
(icheon_behav_tbl_s8)

#조부모와 동거여부별 행동 특성 질문9에 대한 응답
icheon_behav_tbl_g9<- group_by(icheon_behav_tbl,p0105,k0509)
icheon_behav_tbl_s9 <- summarise(icheon_behav_tbl_g9, count = n())
(icheon_behav_tbl_s9)

#조부모와 동거여부별 행동 특성 질문10에 대한 응답
icheon_behav_tbl_g10<- group_by(icheon_behav_tbl,p0105,k05010)
icheon_behav_tbl_s10 <- summarise(icheon_behav_tbl_g10, count = n())
(icheon_behav_tbl_s10)


#조부모와 동거여부별 행동 특성 질문11에 대한 응답
icheon_behav_tbl_g11<- group_by(icheon_behav_tbl,p0105,k05011)
icheon_behav_tbl_s11 <- summarise(icheon_behav_tbl_g11, count = n())
(icheon_behav_tbl_s11)
```

```
#조부모와 동거여부별 행동 특성 질문12에 대한 응답
icheon_behav_tbl_g12<- group_by(icheon_behav_tbl,p0105,k05012)
icheon_behav_tbl_s12 <- summarise(icheon_behav_tbl_g12, count = n())
(icheon_behav_tbl_s12)

#조부모와 동거여부별 행동 특성 질문13에 대한 응답
icheon_behav_tbl_g13<- group_by(icheon_behav_tbl,p0105,k05013)
icheon_behav_tbl_s13 <- summarise(icheon_behav_tbl_g13, count = n())
(icheon_behav_tbl_s13)

### 이동 수단별 순위
table(icheon$k0203)

### 이동 수단별 하교 시간대 분포
icheon_mov<-icheon[,c('k0202','k0203')]

library(dplyr)

icheon_mov_tbl <- tbl_df(icheon_mov)

icheon_mov_tbl_g <- group_by(icheon_mov_tbl,k0203,k0202)

icheon_mov_tbl_g_s <- summarise(icheon_mov_tbl_g, count = n())

icheon_mov_df<-as.data.frame(icheon_mov_tbl_g_s)

write.csv(ic_nam_actg2df,"d:/00-gg/참여형기초통계_학교별/output/이천_이동수단
별시간대분포.csv")

## 준경험 내역
icheon_jun<-read.csv("d:/00-gg/참여형기초통계_학교별/준경험_이천초.csv")
table(icheon_jun$u0101)

##하교시 인지하는 위험 사유
icheon_inji<-read.csv("d:/00-gg/참여형기초통계_학교별/인지하는위험도와위험사유
_이천남초.csv")
table(icheon_inji$cause)
```

```
##하교시 인지하는 이동수단별 인지하는 위험사유
library(dplyr)
icheon_inji_tbl <- tbl_df(icheon_inji)
icheon_inji_tbl_g <- group_by(icheon_inji_tbl,이동수단_k0203,cause)
icheon_inji_tbl_s <- summarise(icheon_inji_tbl_g, count = n())
icheon_inji_df <-as.data.frame(icheon_inji_tbl_s)
  write.csv(icheon_inji_df,"d:/00-gg/참여형기초통계_학교별/output/이천남_이동수단과이유.csv")
```

공공 빅데이터
표준분석모델 매뉴얼 어린이 안전 및 교통사고 원인분석

초판 1쇄 인쇄 : 2018년 1월
초판 1쇄 발행 : 2018년 1월
발행처 : 행정안전부 · 한국정보화진흥원
펴낸곳 : 행정안전부 · 한국정보화진흥원

행정안전부
서울특별시 종로구 세종대로 209
http://www.mois.go.kr

NIA 한국정보화진흥원
대구광역시 동구 첨단로 53
http://www.nia.or.kr
Tel. (053)230-1564

공공 빅데이터 표준분석모델 매뉴얼

어린이 안전 및 교통사고 원인분석

공공 빅데이터?

정부와 공공기관이 갖고있는 공공데이터와 민간의 다양한 데이터를 융합하고, 분석하여 국민의 복지와 안전, 생활에 필요한 각종 공공 서비스를 국민의 요구와 기대에 맞추어 개선하고, 새로 수립하는데 도움을 줍니다.

혁신의 기반 빅데이터!

흩어져 있으면 그냥 자료, 모아서 분석하면 미래를 예측할 수 있는 빅데이터!
국민을 위한 최선의 정책과 효율적 서비스 수립에 도움을 주며, 미래 예측을 통한 선제적 대응이 가능해집니다.

표준분석모델이란?

중앙부처 및 지자체에서 개별적으로 분석 진행한 과제들을 기초로, 행정업무에 지속적으로 활용이 가능한 분야를 선정하였습니다. 분야별 우수사례를 기반으로 분석 기법과 분석신뢰도를 높이고, 데이터 수집 및 사용법을 표준화 하여, 실무담당자가 비교적 손쉽게 분석을 진행할 수 있도록 표준화한 분석모델 입니다.

행정안전부
서울특별시 종로구 세종대로 209
http://www.mois.go.kr

NIA 한국정보화진흥원
대구광역시 동구 첨단로 53
http://www.nia.or.kr

비매품/무료
14000
9 788984 832909
ISBN 978-89-8483-290-9
ISBN 978-89-8483-286-2 (세트)