



# 사회과학 연구자의 R 동행기

---

발표자: 최재성 (성균관대학교 경제대학)

2021-11-19

# 발표 개요

---

대학에서 데이터 분석을 가르치고  
사회과학 분야에서 실증 연구를  
업으로 삼고 있는 발표자는

왜 R에 관심을 가지게 되었고,

무엇을 익혔고,

어떤 일을 해왔을까?

사회과학 연구의 도구로서 R의  
다양한 활용 가능성에 대한 생각을  
나누어본다.

1. 동행의 시작

2. 동행하며 마주한 세상

3. 함께 걷기

4. 한 걸음 더 내딛기

# Chapter 1

## 동행의 시작



Social Science & Medicine

Volume 134, June 2015, Pages 1-11



## Separating boys and girls and increasing weight? Assessing the impacts of single-sex schools through random assignment in Seoul

Jaesung Choi <sup>a, \*</sup> , Hyunjoon Park <sup>b</sup>, Jere R. Behrman <sup>c</sup>

We rely on two datasets for Korean adolescents. For **school-level analysis**, we use a school-level database on health outcomes of middle- and high-school students compiled by the Korean government, as reported by each school in accordance with educational law. The compiled data are publicly available online ([www.schoolinfo.go.kr](http://www.schoolinfo.go.kr)). Starting with elementary school students, the Korean government mandates school-level physical examinations and release of information on a yearly basis. In 2009, physical ex-

Hence, **results for three years (2010, 2011, and 2012 examinations) were available for middle schools while results for two years (2011 and 2012 examinations) were available for high schools** as of the writing of this paper. If a school is coeducational, each item is reported by gender. With regard to the weight-related measure, each school reports school mean body mass indices (BMI) for each grade level by gender.

# 학교알리미 데이터 수집 프로젝트

## 고등 경기고등학교 ★

설립구분: 공립    설립유형: 단설    학교특성: 일반고등학교

설립일자: 1900년 10월 03일

학생수: 1,156명 (남 1,156명, 여 0명)

교원수: 96명 (남 34명, 여 62명)

체육집회공간: 2실

대표번호: 02-3496-7300    팩스: 02-3496-7497

행정실: 070-3496-7310    교무실: 070-3496-7330

홈페이지: <http://kyunggi.sen.hs.kr>

주소: 서울특별시 강남구 영동대로 643

관할교육청: 서울특별시교육청



### 공시정보

2020년

선택

학생현황	교원현황	교육활동	교육여건	예결산현황	학업성취사항
<ul style="list-style-type: none"><li>학교 현황</li><li>입학전형 요강</li><li>학생의 체력 증진에 관한 사항</li></ul>	<ul style="list-style-type: none"><li>성별 학생수</li><li>입학생 현황</li></ul>		<ul style="list-style-type: none"><li>학년별·학급별 학생수</li><li>졸업생의 진로 현황</li></ul>	<ul style="list-style-type: none"><li>전·출입 및 학업중단 학생 수</li><li>장학금 수혜 현황</li></ul>	

# 학교알리미 데이터 수집 프로젝트

## 15-자. 학생의 체력 증진에 관한 사항

공시년월 (1차) 2020년 04월

### ■ 종목별 평균기록 통계[2019.12.31 기준]

구분		심폐지구력			유연성		근력/근지구력			순발력		비만		
급별	학년	성별	왕복오래달리기(회)	오래달리기걸기(분.초)	스텝검사(PE)	앉아윗몸앞으로굽히기(cm)	종합유연성(점)	(무릎대고)팔굽혀펴기(회)	윗몸말아올리기(회)	약력(kg)	50m달리기(초)	제자리멀리뛰기(cm)	BMI(kg/m <sup>2</sup> )	체지방률(%fat)
중	1	남	56.9			7.6		26	100	28.1	9	190	20.9	
중	1	여	43.9			16.7		42	52	24.5	9.2	140	19.6	20.5
중	2	남	63.4	7.51		10.8		41		34.3	8	170	21.1	
중	2	여	36.6			17.1			58.5	24.8	9.3	142.5	20.1	
중	3	남	66.6	8.07		12.7		31.5		38	7.8	207	21.4	
중	3	여	40.7			19.4			80	26.5	9.4	180	20.6	

Ctrl + C  
Ctrl + V




**Autolt**

Autolt v3 is a freeware BASIC-like scripting language designed for automating the Windows GUI and general scripting. It uses a combination of simulated keystrokes, mouse movement and window/control manipulation in order to automate tasks in a way not possible or reliable with other languages.

[Read More...](#)



조력자 찾기  
(C#)

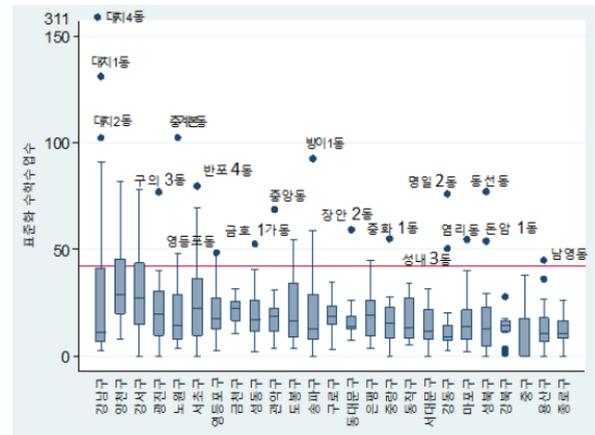
# 문상균, 배한나, & 최재성 (2016)

## “학원정보 공공데이터를 활용한 서울시 사교육 공급에 관한 분석”

학생들의 사교육 이용은 사교육 수요뿐만 아니라 사교육 공급에도 영향을 받는다. 따라서 인구학적 특성과 지역에 따른 교육 기회의 격차를 이해하고, 교육 정책의 개발 및 효율적 집행을 위해서는 사교육 공급 측면을 정확하게 파악하는 작업이 필요하다. 본 연구는 **나이스 학원 민원서비스를 통해 공시되는 서울에 소재한 모든 학원에 대한 정보를 수집(web scraping)하고, 공시 자료에 포함된 개별 학원의 주소와 학원에서 제공하는 수업 정보를 활용하여 서울시 사교육 공급의 공간분포를 살펴본다.** 이 과정에서 지역·과목특성·학교급별 차이에 주목한다. 분석의 결과, 서울시 사교육 공급은 소수의 행정동을 중



〈그림 2〉 서울 행정동별 표준화 수업 수 단계구분도



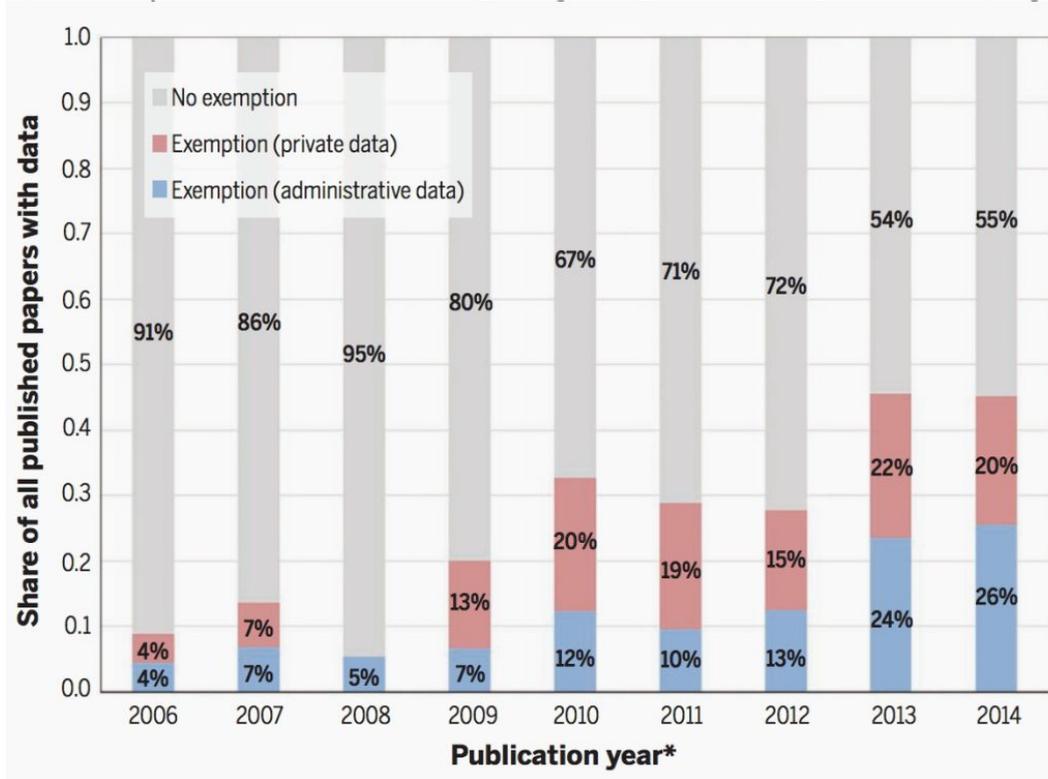
〈그림 3〉 행정구별 행정동의 표준화 수학 수업 수

# Chapter 2

## 동행하며 마주한 세상

# Economics in the Age of Big Data

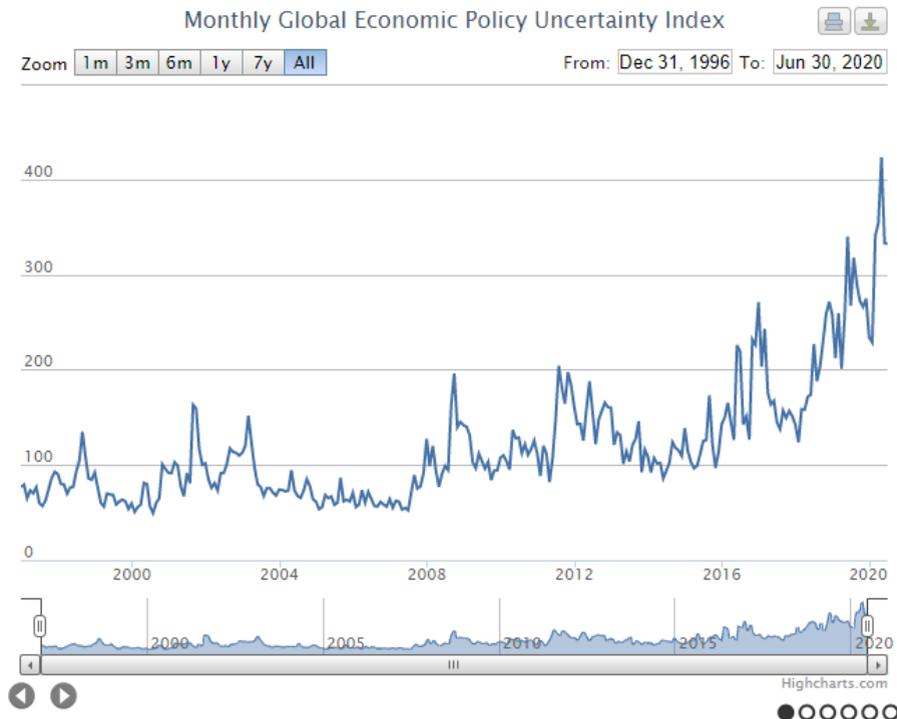
Einav & Levin, "[Economics in the age of big data](#)" (Science, 2014)



- 대규모 행정 데이터와 민간 부문과의 협업을 통한 연구 증가 (Google, Facebook, Amazon)
- 새로운 유형의 자료를 활용한 연구 증가 - 검색 기록, SNS, 스캐너, GPS
- 흥미로운 논문들은 여기서 확인: [Link](#)

# Baker et al. (QJE 2016) - EPU

- Assemble the full texts of 10 leading newspapers to construct a daily index of economic policy uncertainty:  
**Economic Policy Uncertainty Index**
- Important for understanding firm investment decisions and macroeconomic activity



## 경제불확실성 지수(Economic Policy Uncertainty, EPU Index)

실시간으로 생성되는 뉴스기사의 텍스트 데이터를 분석해 경제 흐름을 파악  
매월 경제불확실성 지수 제공

EPU 지수

관련자료

 EPU 지수란?

 월별 지수 추이 (Monthly Index)

2013.01 ~ 2021.07

 EPU 지수   불안 지수 (Anxiety)   위기 지수 (Crisis)

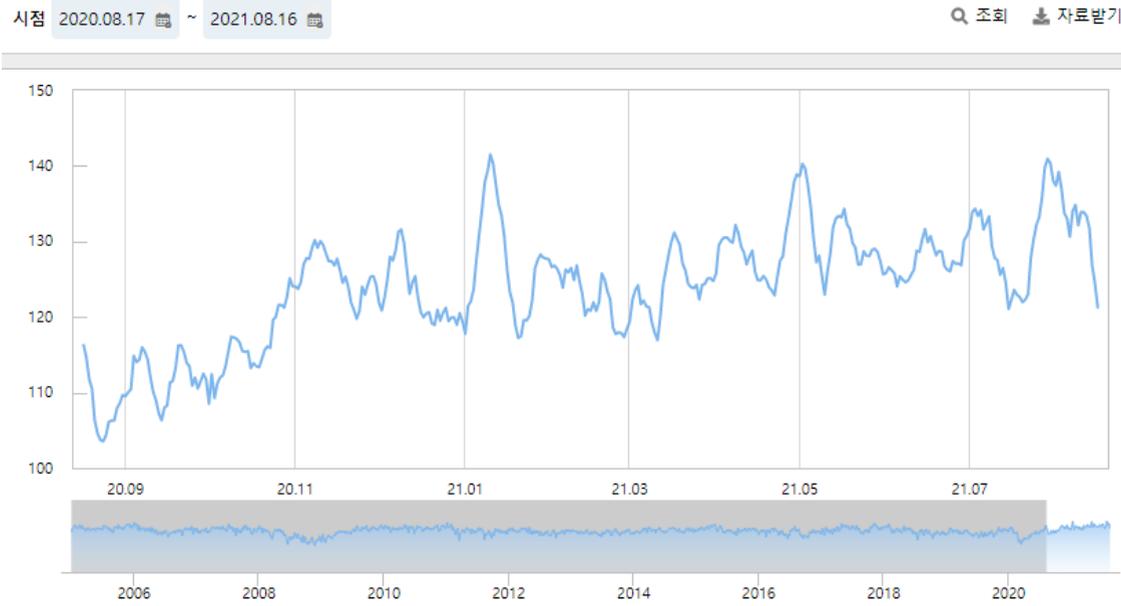


# 뉴스심리지수 동향 by 한국은행

## 뉴스심리지수(NSI, News Sentiment Index)

- 직전 7일간 뉴스기사에 나타난 경제심리를 일 단위로 지수화
- 포털사이트의 경제분야 뉴스기사에서 표본문장을 무작위 추출
- 각 문장에 나타난 경제심리를 기계학습을 통해 긍정, 부정, 중립으로 분류하고 이를 이용해 지수 산출

$$NSI = \frac{\text{기간내 긍정문장 수} - \text{부정문장 수}}{\text{기간내 긍정문장 수} + \text{부정문장 수}} \times 100 + 100$$



## “Polarized embrace: South Korean media coverage of human rights, 1990–2016”

- **Webscraped all articles published in the four newspapers with the term “human rights” (인권)**
  - 조선일보, 중앙일보, 한겨레신문, 경향신문
  - 101,689 편의 기사
- 인권과 관련이 없는 기사 제외하기
  - Inkwon emerged as a part of the name of a famous Korean celebrity, Inkwon Jeon (전인권)
  - Inkwon was a part of another word, such as coupons (할인권)
- **Special symbols and characters**
- **Meta-information** such as author information for newspaper articles and news providers such as AP News
- **Advertisements and additional news lists attached**

# R의 다양한 활용 - 웹데이터 수집과 자동화



파이썬 입문 및 웹 크롤링을 활용한 다양한 자동화 어플리케이션 제작하기

진행률: 16.36% | 기한: 무제한



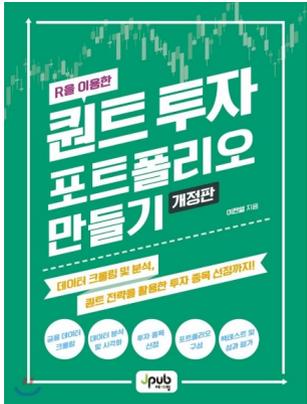
파이썬입문과 크롤링기초 부트캠프 (2021 업데이트) [쉽게! 견고한 자...

진행률: 68.49% | 기한: 무제한



현존 최강 크롤링 기술: Scrapy와 Selenium 정복

진행률: 100.00% | 기한: 무제한



sinlearn

로그인 | 회원가입

NEW 게임 · 프로그래밍 > 프로그래밍 언어

내 업무를 대신 할 파이썬(Python) 웹크롤링 & 자동화 (feat. 주식, 부동산 데이터 / 인스타그램)

61명이 수강하고 있어요.

DeeplingSauce

# Python 데이터 분석 웹 크롤링 업무자동화

Requests Selenium

Web Crawling & Automation  
with Python

올인원 패키지 Online.

6개월치 업무를 하루만에 끝내는  
업무 자동화.

<p><b>만이다</b></p> <p>예제 24개만 배껴쓰면 OK</p>	<p><b>쉽다</b></p> <p>파이썬, 코딩 몰라도 OK</p>	<p><b>유용하다</b></p> <p>업무별 예제 &amp; 매크로 라이브러리로 바로 실무 적용 OK</p>
--	--	---

- 1 데이터 과학 <sup>1</sup>
- 2 데이터
- 3 통계학 <sup>3</sup>
- 4 데이터 사이언스 작업흐름도
- 5 데이터 사이언스 응용
- 6 R의 파이썬 대비 우수성 <sup>5</sup>
- 7 데이터 과학 교육
- 7.1 우수 데이터 과학 교육**
- 7.2 데이터 사이언스 교재
- 7.3 대쉬보드(Dashboard)
- 8 구글 추세로 본 데이터 과학

## 7.1 우수 데이터 과학 교육

- 데이터 과학 입문 ← 데이터 과학 논문 제작
  - 데이터 과학 입문(2019): 봄학기
  - 데이터 과학 입문(2020): 봄학기 ← 대폭 보강 예정 ^^
- 데이터 과학을 위한 소프트웨어 공학(Software Engineering for Data Science) ← 데이터 과학 제품 제작
  - Software Engineering for Data Science (2019): 가을학기
- 문건웅 교수님 Shiny 강의
  - 강사님이 대학교수로 전문 교육인
  - R전문가를 대상으로 R Meetup에서 무료 진행
  - 라이브 코딩으로 3시간 동안 수강생에게 실도움이 되도록 강의를 진행
- 김재광 교수님 유튜브 강의
  - Computer Age Statistical Inference를 한국어로 그것도 동영상으로 무료강의
- 소프트웨어/데이터 카펜트리(Software/Data Carpentry)
  - 과학 컴퓨팅(Scientific Computing)에 대한 기본기 교육
  - 수십년에 걸쳐 전세계 검증된 교육과정
  - Software Carpentry

- R과 동행하는 길에 많은 가르침을 주신 이들
  - Google, Naver, Stack Overflow
  - 이광춘, 유충현, 문건웅, 우석진, 나성호, 이현열, 박찬엽
  - 김영우, 김승욱, 슬기로운통계생활(Issac Lee)
  - 언론사 데이터저널리즘팀(특히, SBS 마부작침 배여운 & 안혜민)
  
- 페이스북 커뮤니티

# Chapter 3

## 함께 걷기



매일 아침 7시,

전날의 주요 뉴스나  
나의 관심사와 관련된 뉴스,

팟캐스트나 Youtube에  
새로 올라온 아이템을,

“맞춤형 뉴스레터”로  
받아볼 수는 없을까?

# 맞춤형 Morning Letter

Morning Letter

1. Article 2. Podcast

Chart A1 (Issued at 2020-08-11 11:51:25)  
랭킹 뉴스 - 경제 분야, 사회 분야

Section	Ranking	Title	Source
경제	1	당판 돈 35억 주식에 '물뿜'한 70대...세금폭탄샷 '머니무브'	한국경제
경제	2	정용진·백중원 '콜라보' 파워 무섭네...위기의 장어도 풀졌다	중앙일보
경제	3	재난지원금 평평 쓰더니 수해복구 쓸 돈 바닥...4차추경 2조 넘을듯	매일경제
경제	4	서초구청장의 기세제안... "재산세 50% 깎아주겠다"	매일경제
경제	5	호이 5000만원 투자한 그 펀드, 1년 만에 2800만원 벌었다	중앙일보
경제	6	부부 공동명의로 2... '취득세 3000만원' 폭탄, "잠이 안 온다"	머니투데이
경제	7	노영민 반포 아파트 팔았나 안 팔았나, 등기 명의는 그대로	중앙일보
경제	8	일본 불매운동 1년...소비재 수입, 맥주 84%↓, 승용차 51%↓	연합뉴스
경제	9	부동산 우울증... 무주택 '좌절' 1주택 '합승' 다주택 '분노'	조선일보
경제	10	한달새 전셋값 2억 치솟자 '뉴물의 대응'...규제 따로 현실 따로	매일경제
경제	11	여자는 '엄마, 나 용돈 줘', 남자는 '이 말'에 보이시피상 당했다	조선일보
경제	12	"요즘 누가 현금 써요?... 그래서 현금만 써 봤습니다(아무이슈)"	서울신문
경제	13	정용진·백중원 손 잡자... 재고 900톤 바다장어 '품질 대란'	아시아경제
경제	14	폭등하는 전셋값 언급하고...대통령 "집값 안정세 시작됐다"	매일경제
경제	15	1주택자도 증세 못피해... 송파 17억 집 보유세 818만→1169만원	동아일보
경제	16	투자의 神 "개미를 조심해! 지금 주식 시장은 카지노판"	조선일보
경제	17	'안산 아내 살해 혐의' 남편에 보형금 100억원 지급되나	서울경제
경제	18	"집 안사면 되지 뭐"... 더 뜨거워진 2030의 '수입차 플렉스'	한국경제

Chart B1  
네이버 뉴스 검색

Date	Search	Title	Source
2020.08.11.	성균관대 경제학과	'최초 공개' 서울대 학생 합격선... 정시 합격선도 등급 대신 '확산점수'로...	에듀동아
2020.08.11.	성균관대 경제학과	국내 기업 10곳 중 4곳 "고용유지 어려움... 정부지원 간절"	이코노믹리뷰
2020.08.10.	성균관대 경제학과	코로나19 쇼크... 기업들 '고용유지' 총력전	문화저널21
2020.08.10.	성균관대 경제학과	공매도 투론회 개최... 규제 개선방향 모색	데일리그라운드
2020.08.10.	성균관대 경제학과	'뜨거운 감자' 공매도 투론회... 참가 신청 1분 만에 마감	뉴스1
2020.08.10.	성균관대 경제학과	기업 10곳 중 4곳 "코로나로 일감 줄어 직원 감원해야할 상황"	청년일보
2020.08.10.	성균관대 경제학과	기업 10곳 중 4곳, 코로나 충격에 고용조정 요구 직면	아소스스타임스
2020.08.10.	성균관대 경제학과	기업 10곳 중 4곳, 코로나로 고용조정 불가	스트레이트뉴스
2020.08.10.	성균관대 경제학과	"한국 기업, 코로나에도 고용 지켜... 정부 지원 필요한 때"	한국금융신문
2020.08.10.	성균관대 경제학과	대한상의 "기업 10곳 중 4곳 고용조정 상황 직면... 지원 필요"	KBS
2020.08.08.	마부작침	[마침] 떨어졌다! 동네의회... 업무추진비 편	SBS
2020.08.07.	마부작침	[마부작침 외전] 동네의회 월다가 내 영혼도 털린 사연	SBS
2020.08.07.	마부작침	[비디오머그] 조각상 발가락 부러뜨린 빈인 CCTV 영상 공개... 아무렇지 않은 척...	SBS
2020.08.07.	마부작침	전 세계 코로나19 확진 1천900만 명... 4일에 10만 명씩 늘어	SBS

- NAVER 뉴스 - [랭킹뉴스](#) - 경제/사회
- 팟빵 관심 채널: "이진우의 손에 잡히는 경제", "세바시", etc.

## 맞춤형 콘텐츠로 모닝 레터를 만들고, 원하는 시간에 레터를 자동 생성하고, 다양한 매체로 공유하기

- 콘텐츠 수집하기: 웹스크래핑/API
- 레터 작성하기: Flex Dashboard (R Markdown)
- R 스크립트 자동 실행 스케줄
- E-mail/SNS/웹 포스팅/Cloud 서비스 공유

**학술 연구를 위한 웹데이터 수집 &  
마케팅 리서치도 결국 동일한 방식**

- R 프로그램 활용 사례 발표
  - Shiny
  - GIS 분석
  - 텍스트마이닝
  - R & SQL
  - 영상 처리, 위성 사진 분석
  - Reticulate 패키지 for Python
  - R 패키지 만들기
  - 병렬처리
- 데이터저널리즘 관련 특강 후 설문 결과: “더 배우고 싶은 것은?”
  - 시각화 – 인터랙티브, 맞춤형 그래프 디테일 설정 (BBC, Economist, etc)
  - 분석 내용이 포함된 글을 쓰고 포스팅하기 (Distill & Substack)
  - PDF 문서에 포함된 정보를 활용하는 분석

- [How Has Labor Demand Been Affected by the COVID-19 Pandemic? Evidence from \*\*Job Ads\*\* in Mexico](#) (Campos et al., 2021)
- [Combining \*\*Satellite Imagery\*\* and \*\*Machine Learning\*\* to \*\*Predict Poverty\*\*](#) (Jean et al., 2016)
- [\*\*Gender Differences\*\* in Recognition for Group Work](#) (Sarsons, 2021)
- [\*\*Nowcasting Gentrification\*\* Using \*\*Airbnb Data\*\*](#) (Jain et al., 2021)
- [Social Media and \*\*Fake News\*\* in The 2016 Election](#) (Allcott & Gentzkow, 2017)
- [Web Scraping \*\*Housing Prices in Real-time\*\*: the Covid-19 Crisis in the UK](#) (Bricongne et al., 2021)

# 사회과학 연구자를 위한 KOSSDA 강의

## 웹데이터를 활용한 빅데이터 분석

웹스크래핑과 API, 텍스트 자료 처리, Markdown의 활용, 시각화 및 자동화, R 실습

### 1. 과정 개요

개요	
워크숍 목표 및 개요	이 워크숍은 R을 사용해서 웹에서 데이터를 수집하고, 이렇게 수집된 다양한 형태의 데이터를 전처리 후 분석하는 능력을 갖추도록 돕는 것을 목표로 한다. 또한 R에서 생성한 분석 결과를 이메일이나 SNS를 통해 공유하는 방법과 자료 수집 및 분석이 반복적으로 수행되도록 자동화하는 방법을 다룬다. 아울러 학술연구에 활용할 수 있는 유용한 시각화 방법과 웹에서 수집한 데이터를 활용한 사회과학 연구들을 소개한다.
참가 대상	웹에서 자료를 수집하고 이를 활용하여 학술연구를 수행하고자 하는 학부생, 대학원생 및 연구자
선수 과목	기본적인 R 프로그램에 대한 이해 - R을 사용해서 기초통계 분석이 가능하고, 조건문, 반복문, 함수를 다룰 수 있어야 합니다.
워크숍 운영방식	1) 사전 제작 동영상 강의 (관련 배경 지식과 함수 소개 및 기초 사례 실습) 2) zoom 실시간 온라인 강의 (심화 사례 실습 및 Q&A)
교재 및 참고문헌	강사의 강의노트가 제공됩니다.  <참고문헌> - 김영우. 2017. Do it! 쉽게 배우는 R 데이터 분석. 이지스퍼블리싱 - 이현열. 2021. R을 이용한 퀀트 투자 포트폴리오 만들기. 제이펍 - Rafael A. Irizarry. 2021. Introduction to Data Science (웹에 공개됨) - Hadley Wickham & Garrett Golemund. 2021. R for Data Science (웹에 공개됨)

- 다양하고 흥미로운 질문을 통해 동료 연구자들의 관심사를 알게 되는 기회
  - Youtube 자막, Google Play Store
- 여러 분야의 학술 연구를 접하고 공유하는 네트워킹

# Chapter 4

## 한 걸음 더 내딛기

# 함께 나누고 싶은 몇 가지 생각

- 'Stata 바라기'였던 이에게 R은 많은 가능성을 보여줌
  - 웹자료 수집과 활용
  - GIS 분석
  - 텍스트마이닝
- 여전히 Stata로 하지 못하는 기능 위주로 R을 활용하지만, R 생태계는 점점 더 매력적
  - Tidyverse
  - Tidymodels
  - Tidytext
  - Rstudio
  - DataCamp



# 기필코! 프로젝트

## • 데이터저널리즘 분야에 대한 꾸준한 관심과 참여



## • 통계청 MDIS 자료와 API를 활용한 대시보드 시각화



- 사회과학 연구를 위한 **데이터 패키지 만들기**
  - 학교알리미, 대학알리미, 각종 공공데이터

## “Python을 사용할 생각은 없나요?”

- 현실적 고민 - Python에 대한 수요가 점점 높아짐
  - 학부생 교양 필수 과목에서 Python을 사용
  - 퀀트응용경제학과 석사 프로그램은 금융권 재직자가 50% 이상
- 연구 측면
  - **“내가 하고 싶은 작업을 쉽게 익혀서 활용할 수 있는 툴은?”**
  - 실증연구는 “질문(도메인 지식) + 데이터 + 방법론 (+ 스토리텔링)”
- 강의 측면
  - **“15주 동안 나는 무엇을 학생들에게 꼭 전달하고 싶은가?”**
  - <노동경제학> 프로그래밍 실습 5~6시간, Stata -> R
  - <Labor Informatics> R 사용 비중 변화, 50% -> 65%+
  - 응용미시경제학 분야 논문은 Stata & R 사용 비중이 압도적

# 사회과학 연구자의 선택? Stata, R or Python?

## [Gendered Language on the Economics Job Market Rumors Forum \(Wu, 2018\)](#)

### Replication data for: Gendered Language

Principal Investigator(s):  Alice H. Wu

Version:  V1

Name 	File Type 	Size 	L
 <a href="#">lasso-linear-pronoun-sample.py</a>	text/x-python	2.9 KB	1
 <a href="#">lasso-logit-full-sample.py</a>	text/x-python	3 KB	1
 <a href="#">lasso-logit-pronoun-sample.py</a>	text/x-python	3.1 KB	1
 <a href="#">gendered_posts.csv</a>	text/plain	145.3 MB	1
 <a href="#">keys_to_X.csv</a>	text/plain	3.8 MB	1
 <a href="#">tables-figures.R</a>	text/plain	2 KB	1

## [Combining Satellite Imagery and Machine Learning to Predict Poverty \(Jean et al., 2016\)](#)

The data and code in this repository allows users to generate figures appearing in the main text of the paper *Combining satellite imagery and machine learning to predict poverty* (except for Figure 2, which is constructed from specific satellite images). Paper figures may differ aesthetically due to post-processing.

Code was written in R 3.2.4 and Python 2.7.

3. Unzip these files so that `**data/input/LSMS**` contains:
  1. UGA\_2011\_UNPS\_v01\_M\_STATA
  2. TZA\_2012\_LSMS\_v01\_M\_STATA\_English\_labels
  3. DATA (formerly NGA\_2012\_LSMS\_v03\_M\_STATA)
  4. MWI\_2013\_IHPS\_v01\_M\_STATA

**“내가 하고 싶은 작업을 쉽게 익혀서 활용할 수 있는 툴은?”  
“어쩌면... 결국은 순서의 문제”**

경청해 주셔서  
감사합니다.

R 사용자회

KO  EA  
컨퍼런스