

2013년도 미국 센서스 데이터
분석을 통해 알아보는

한국인 이민자 라이프스타일

홍익대학교 대학원 산업공학과	박대한
중앙대학교 심리학과	여새바그별
중앙대학교 영어영문학과	이승진
이화여자대학교 사회학과	최윤영
중앙대학교 사회학과	최진혁





2013년도 미국센서스데이터

분석을 통해 알아보는

한국인 이민자 라이프스타일

I . 헬조선

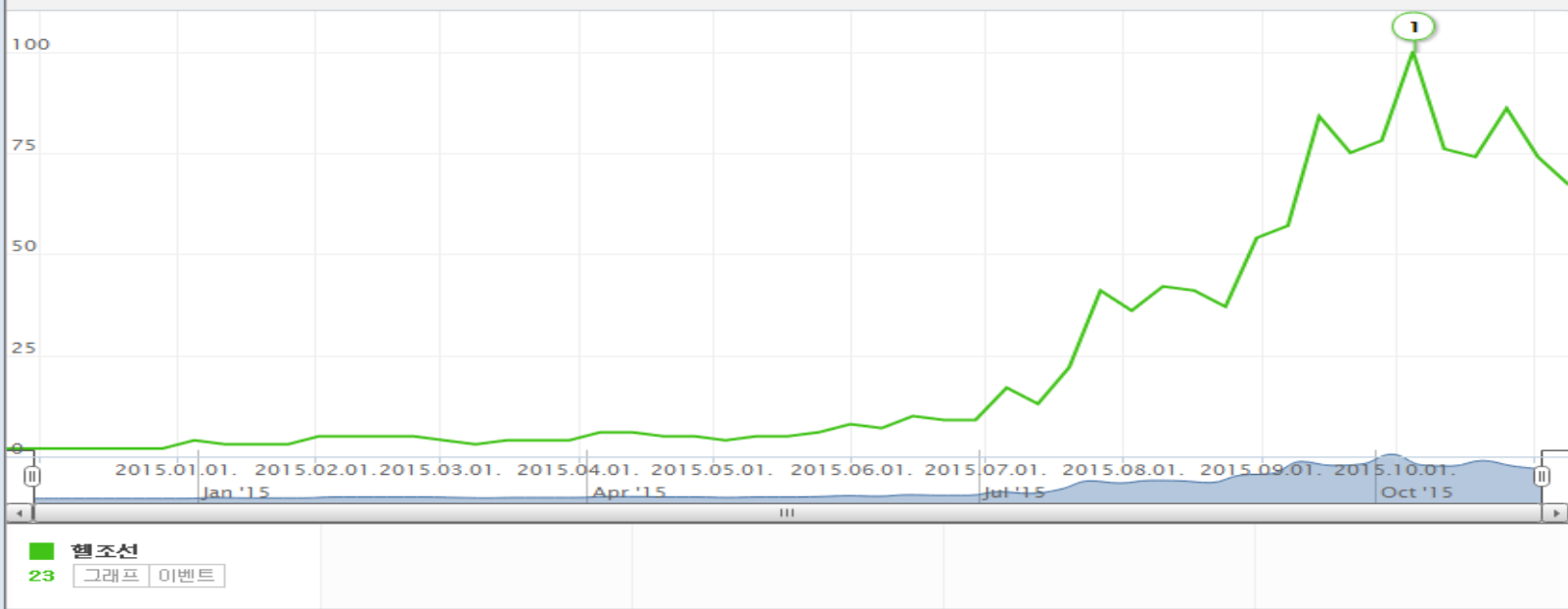
Hell + 조선(朝鮮)의 합성어로,
마치 지옥과도 같은 한국이라는 뜻을 담은 신조어다.

INTRO >>

헬조선이라는 단어는
현재 우리사회를 전반적으로 설명하는 단어로 자리매김하였다

지난 1년간 헬조선 검색어 추이

검색횟수를 주간으로 합산하여 조회 기간 내 최대 검색량을 100으로 나타낸 그래프입니다.



자료: 네이버트렌드 트렌드 검색 (14.11.20~15.11.20, 1년간 검색량)

이러한 헬조선 현상아래,
미국 이민에 대한 사람들의 관심이 높아졌다

점점 늘어나는 미국내 한국인 이민자수

radiokOREA™

미국이민 증가, 3년마다 LA 한곳 생겨난다

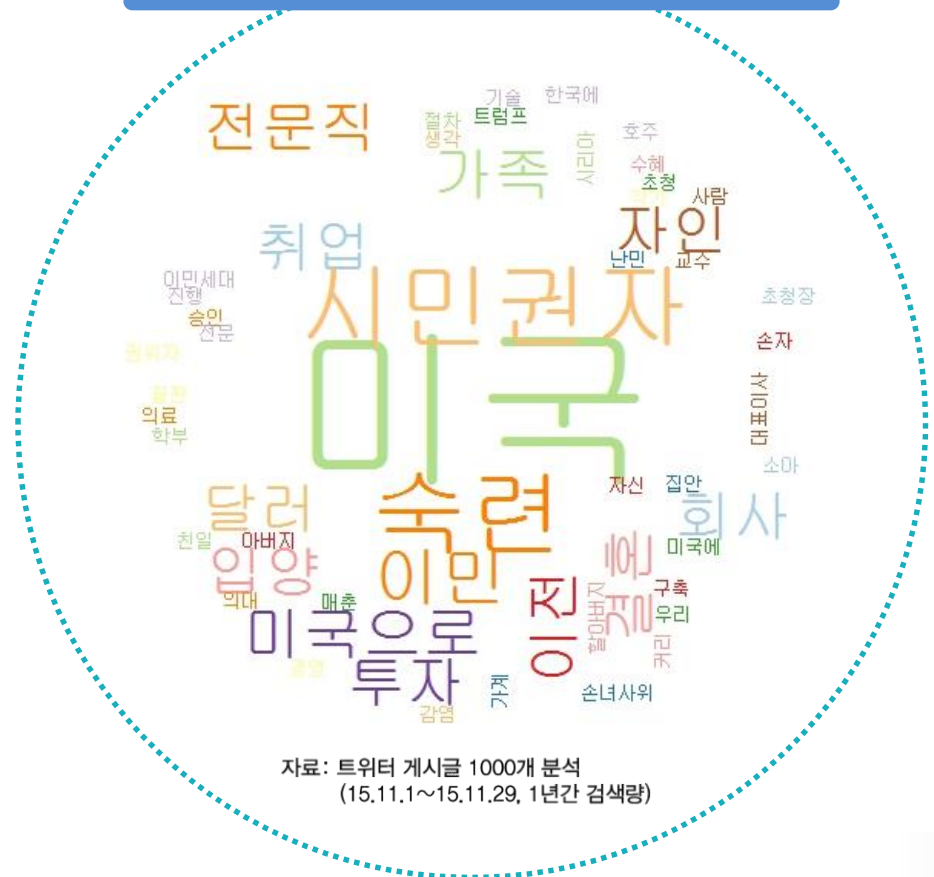
한국일보
THE KOREA TIMES

이민자 인구 연 124만 증가

chosun.com

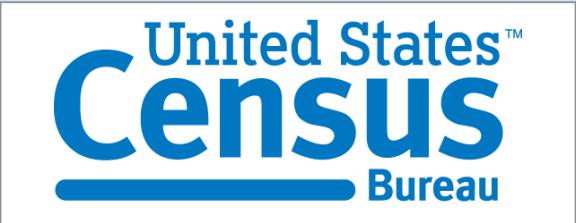
한국인 취업이민 신청 폭증

'미국' / '이민' 키워드 검색 관련 워드 클라우드



2013 미국 센서스 데이터 분석

이미 미국에 정착한 한국인 이민자들이 어떻게 살고 있는지
2013 미국 센서스(인구 총 조사) 데이터를 통해 파악해보자

2013 미국 센서스 데이터	
	
항목	내용
시행 일시	2013
총 조사 인원	약 300만명 (미국인구의 약 1%)
조사항목	283개



II. 근로시간

근로시간 비교 >>

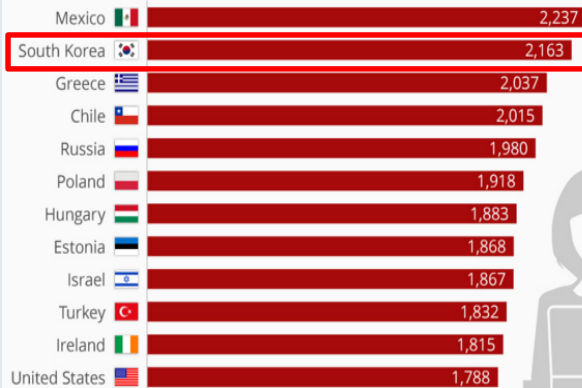
개인이 **일주일 내 근로하는 시간**을 뜻하는 개념으로써,
개인의 삶의 질에 큰 영향을 끼치는 요소.

개인의 삶의 질에 영향을 끼치는 중요한 요소 근로시간

2013 OECD 국가 근로시간 상위 순위

The Countries With The Most Annual Working Hours

Average working hours per year in OECD countries (2013 or latest year)*



출처: 2013, Forbes, "These Countries Have The Most Annual Working Hours"

우리나라는 OECD국가 중 근로시간 2위로 나타남.

2014 OECD 국가 근로시간 상위 순위

OECD 주요국 연간 노동시간

(단위: 시간, 취업자 기준, 한국은 통계청 자료 추산)



자료: OECD, 한국노동사회연구소

출처: 2015, 세계일보, "작년 1인당 근로 2285시간... OECD 국가 중 최고"

우리나라는 OECD국가 중 가장 높은 근로시간을 기록함

각 이민자 집단별 주당 근로시간 비교

비교군 선정

1 독일 이민자



2013 OECD 국가별
근로시간순위가
가장 낮은 나라

2 스웨덴 이민자



전통적으로
사회복지제도가 발달한
대표적인 복지국가

3 이스라엘 이민자



전통적으로 국민성이
근면하다고 여겨지는
이스라엘 이민자 집단

4 멕시코 이민자



2013 OECD 국가별
근로시간순위가
가장 높은 나라

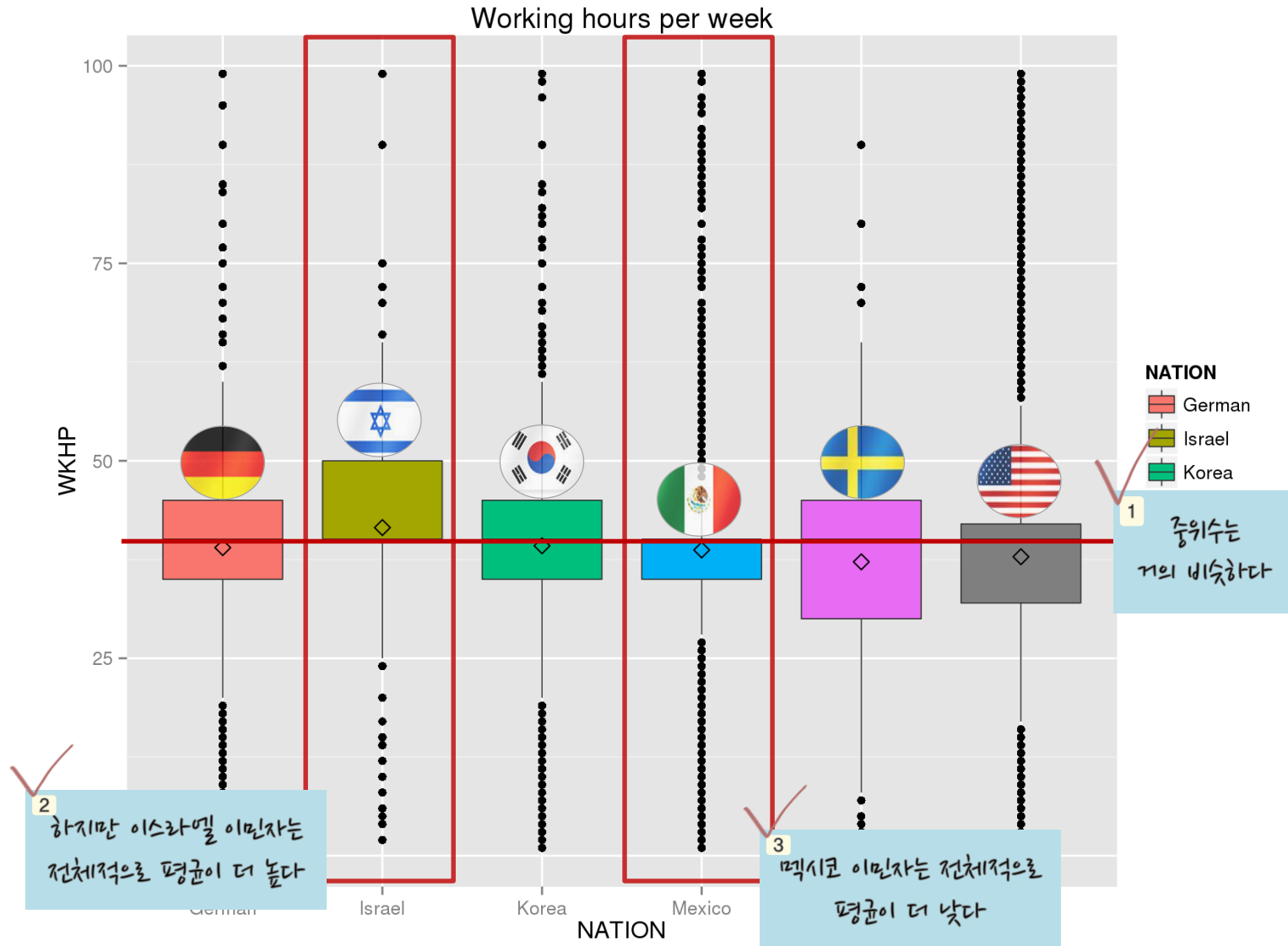
5 나머지 미국인 집단



이들 4개국 이민자를 제외한
나머지 전체 미국인 집단

미국내 각국 이민자 집단

각 이민자 집단별 주당 근로시간 비교



각 이민자 집단별 자영업자 주당 근로시간 비교

각 국가 출신 이민자들의
근로특성 을 쉽게 알 수 있는 집단



자영업자 (Self-employment~)

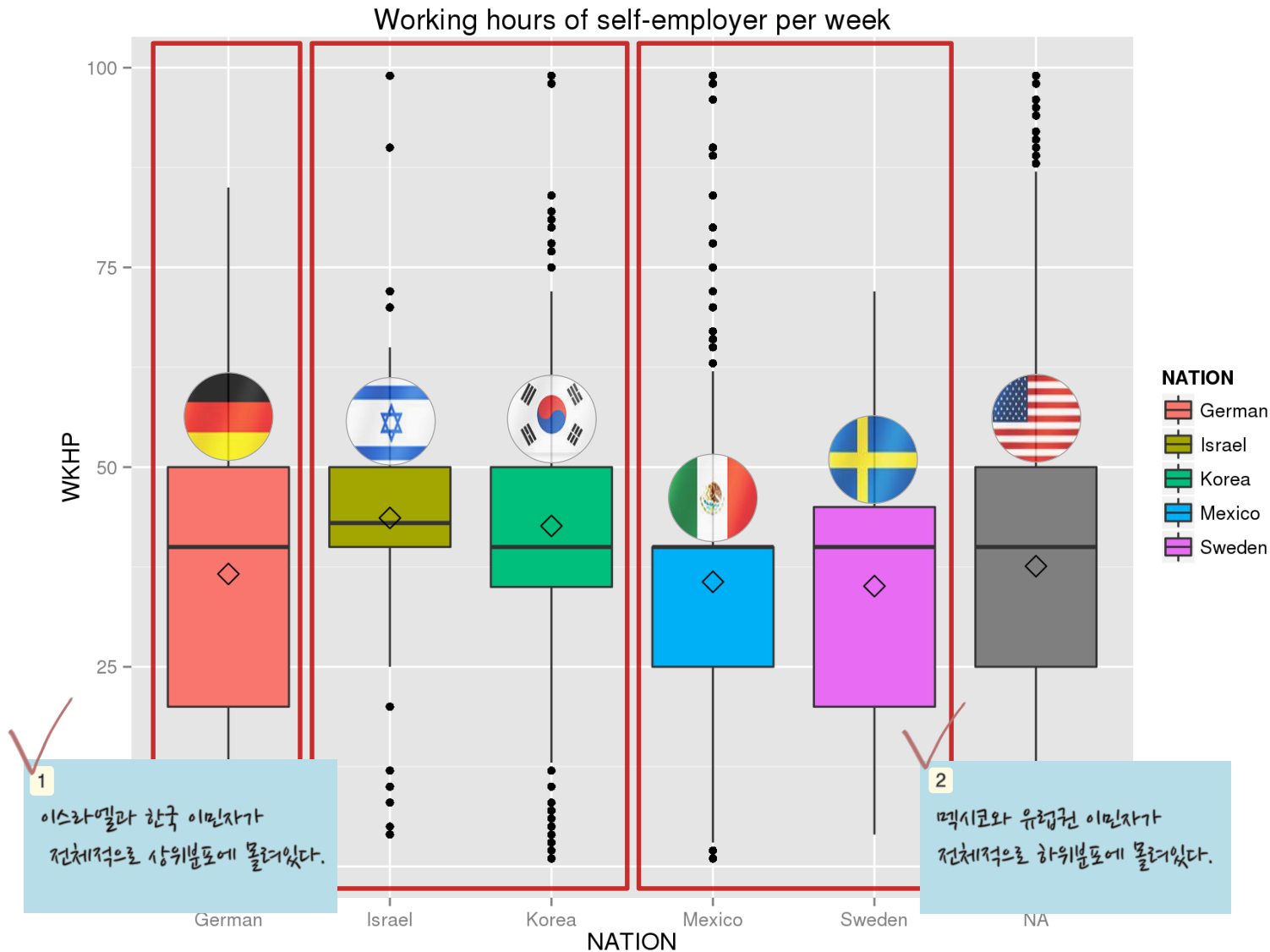
특정 기업 집단이나 조직에
소속되지 않은 근로자

비교적 자신의 근로시간을
탄력적으로 설정할 수 있음



따라서 각 국가별 자영업자의
근로시간을 비교해보았다 ○

각 이민자 집단별 자영업자 주당 근로시간 비교



비교를 통해 알아본 한국인 이민자의 근로시간 특성

국가별 미국 내 이민자 주당근로시간

순위	국가	자국내 근로시간(h)	미국내 근로시간(h)	자국과미국내 근로시간차이	자영업자 근로시간(h)	자국과미국내 자영업자 근로시간차이
1	이스라엘	35,904	41,545	5,641	43,633	7,729
2	한국	39,981	39,238	-0.743	42,635	2,654
3	독일	26,202	39,006	12,804	36,626	10,424
4	멕시코	43,012	38,707	-4,305	35,636	-7,376
5	스웨덴	30,904	37,850	6,946	35,596	4,692

근로환경이 열악한 한국의 상황과
미국 내 한국인 이민자의 근로 환경을 비교해 볼 때
크게 달라지지 않았다○

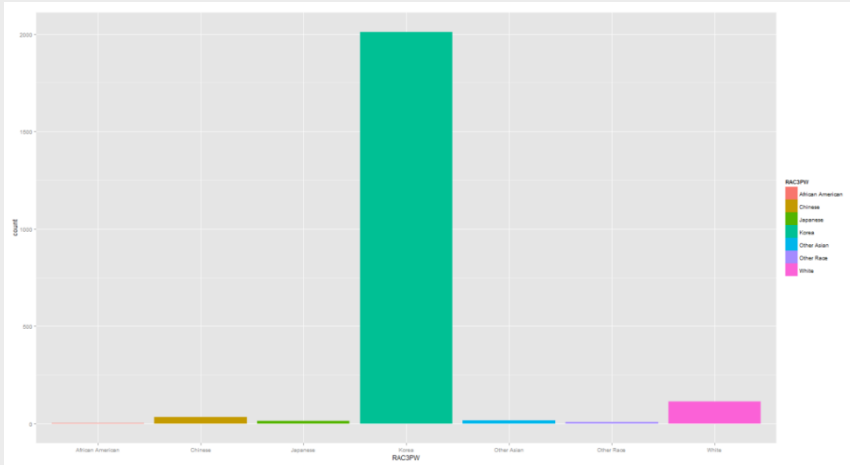
III. 결혼

모델링 >>

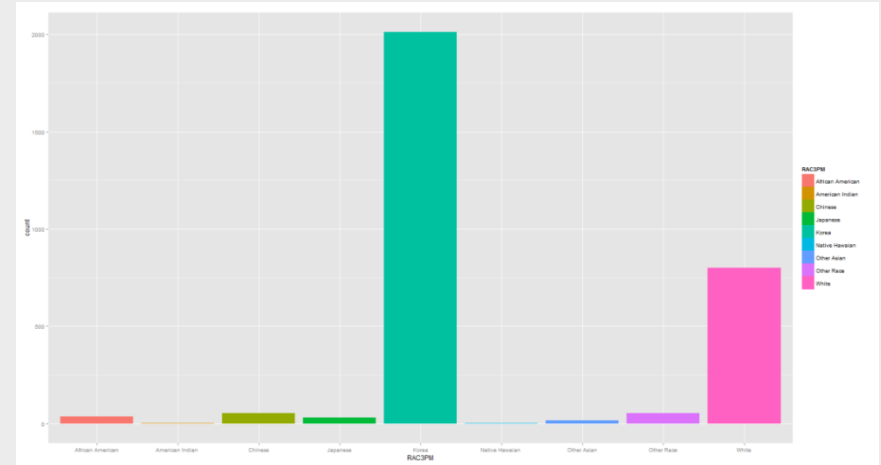
결혼은 개인의 삶에 중요한 부분을 차지하는 요소,
어떤 사람들이 결혼을 했는지 **모델링**을 수행

성별에 따른 한국인 이민자들의 결혼 상대를 살펴보면

결혼 상대자의 인종 [남자]



결혼 상대자의 인종 [여자]



〈배우자의 인종 관련〉

	남자	여자
배우자가 한국인	2,010명 (91.32%)	2,012명 (66.80%)
배우자가 외국인	191명 (8.68%)	1000명(33.20%)
배우자가 한국인/외국인	10:1	2:1

남성의 경우 91%, 여성의 경우 33%의
한국인 이민자가 비한국인과 결혼하였다.

어떠한 사람들이 외국인과 결혼하였는지 알아보자

TARGET 변수 생성

	TAR
배우자가 한국인	0
배우자가 외국인	1

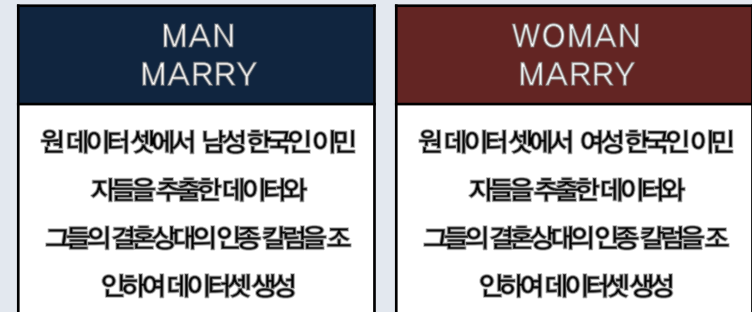
배우자가 한국인인 경우 0,
배우자가 외국인인 경우 1
Target 변수 생성

TAR	MAN MARRY	WOMAN MARRY
0	2010 (91%)	2012 (67%)
1	191 (9%)	1000 (33%)
0:1비율	10:1	2:1

성별에 따른 TAR 변수의 비율 차이가 상당하기에
남성과 여성으로 나누어서 모델링

데이터 셋 분할

데이터셋



성별에 따라 분할된 데이터 셋과
외국인 결혼 여부 변수를 생성하여
분류 모델링을 진행

어떤 분석 기법을 적용할 것인가? : 의사결정나무(Decision Tree)

의사결정나무 분석(Decision Tree)

결과를 이해하고
해석하기 쉬움

Rule 추출 용이

CART Algorithm의
Binary Split을
통한 해석의 단순화



의사결정나무 통한
남녀 각각의 특성에 따른 Rule 개발

Decision Tree – Modeling

[Cost Parameter값을 기준으로 가지치기]

CP table [남자]

	CP	N split	Error	CV Error	CV Error STD
1	0.02792	0	1.00	1.00	0.0691
2	0.02094	6	0.8324	1.00	0.0691
3	0.01570	7	0.8115	0.9424	0.0673
4	0.01308	8	0.7958	0.9947	0.0689
5	0.01047	12	0.7434	1.0052	0.0693
6	0.01	16	0.7015	1.0314	0.0701



X Error와 X 표준편차를
고려하여 CP값 **0.015**
를 기준으로 Pruning

CP table [여자]

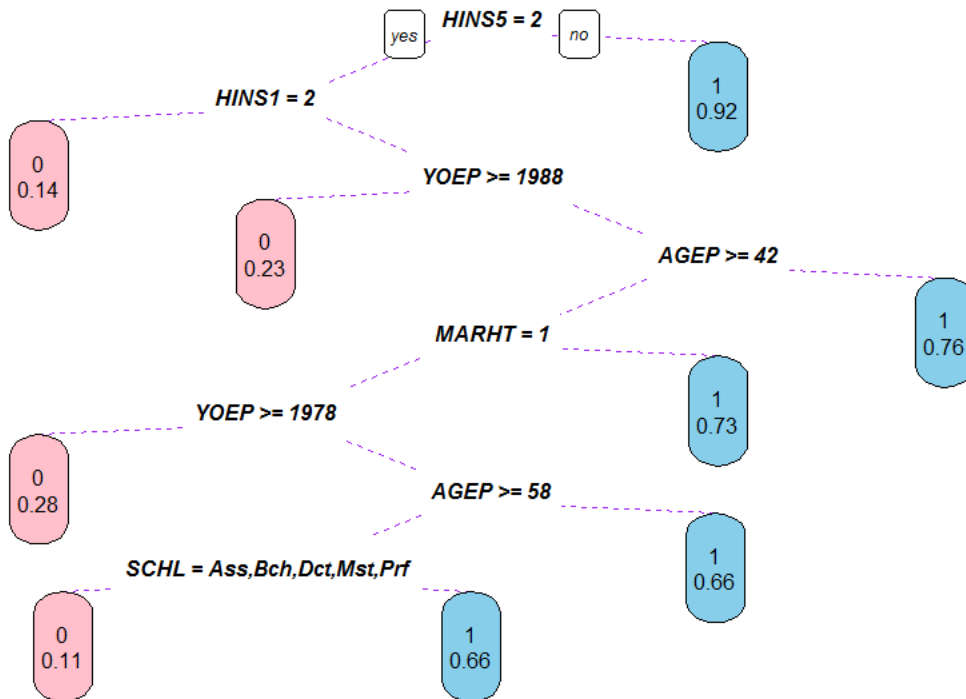
	CP	N split	Error	CV Error	CV Error STD
1	0.2140	0	1.00	1.00	0.02584
2	0.0395	1	0.786	0.786	0.02410
3	0.0215	5	0.620	0.632	0.02234
4	0.0130	7	0.577	0.612	0.02208
5	0.0100	8	0.564	0.607	0.02201



밑으로 내려갈수록 X Error가
줄어들어 디폴트 값인 **0.01**
을 사용하기에 추가적으로
Pruning이 필요없음

외국인과 결혼한 여성 - 분석 결과

CART 알고리즘을 사용한 분류 [여성]








RULE 변수 설명

변수명	설명
HINS5	군인건강보험 수령 여부
HINS1	고용보험 수령 여부
YOEP	이민 온 해
AGEP	나이
MARHT	결혼 횟수
SCHL	학력

Confusion Matrix

실제값/예측값	0	1
0 (한국인과 결혼)	1850명	162명
1 (외국인과 결혼)	402명	598명

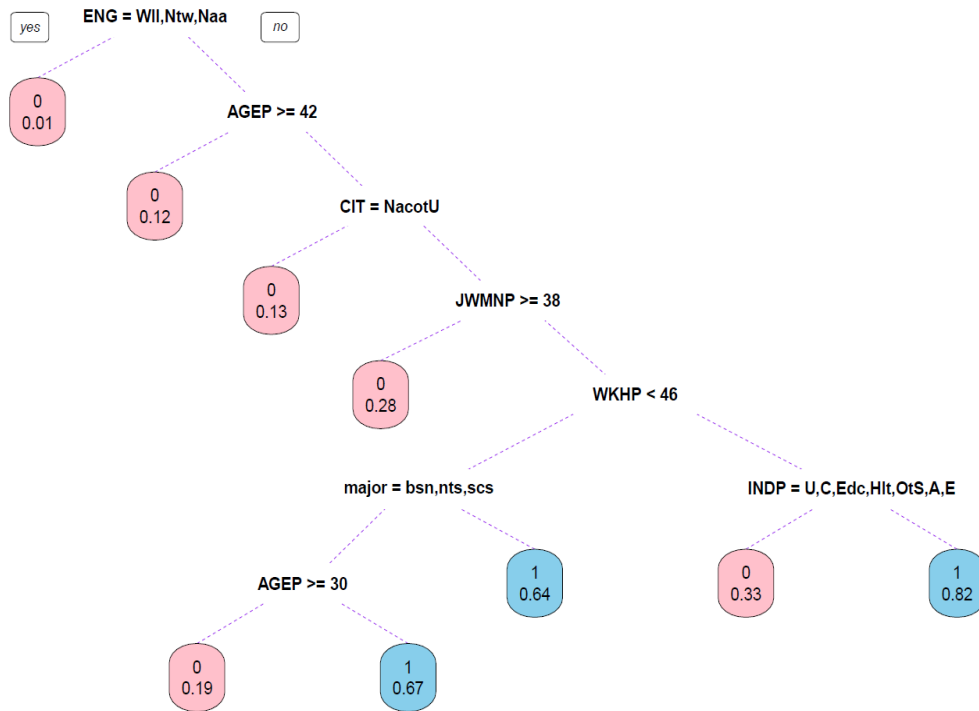
외국인과 결혼한 여성 – RULE

RULE1	RULE2	RULE3	RULE4	RULE5
군인건강보험에 가입된 여성	이민당시 나이가 17세 미만인 여성	88년 이전에 이민 & 42세 이상 & 재혼 여성	이민당시 나이 20대 & 초혼 여성	이민 당시 나이가 30대 이 상 & 초혼 & 최종학력 학사미만 여성
				
Rate : 235/256 [91.79%] Coverage : 235/1000 [23.5%]	Rate : 169/222 [76.12%] Coverage : 169/1000 [16.9%]	Rate : 80/110 [72.72%] Coverage : 80/1000 [8%]	Rate : 87/131 [66.41%] Coverage : 87/1000 [8.7%]	Rate : 27/41 [65.85%] Coverage : 27/1000 [2.7%]

Precision [정밀도]	Recall [민감도]
78.6% (598 / 760)	59.8% (598 / 1000)

외국인과 결혼한 남성 – 분석 결과

CART 알고리즘을 사용한 분류 [남성]



RULE 변수 설명

변수명	설명
ENG	영어 사용 정도
AGEP	나이
CIT	시민권의 지위
JWMNP	직장까지의 소요시간
WKHP	주당 근로시간
INDP	산업군
Major	전공

Confusion Matrix

실제값/예측값	0	1
0 (한국인과 결혼)	1,958명	25명
1 (외국인과 결혼)	127명	64명

외국인과 결혼한 남성 – RULE

공통RULE

1. 영어 말하기 능력이 최상인 경우
2. 어떤 지위로든, 시민권이 있는 경우
3. 집에서 근무지까지의 거리가 38분 이내인 경우

RULE1	RULE2	RULE3
주당 근로시간 >= 46시간 & 광업/제조업/무역업/물류업 산업군 종사자	주당 근로시간 < 46 & 전공 IN 경영학/사회과학/자연과학 & 나이 < 30	주당 근로시간 < 46 & 전공 NOT IN 경영학, 사회과학, 자연과학
Rate : 31/38 [81.58%] Coverage : 31/191 [16.2%]	Rate : 6/9 [66.67%] Coverage : 6/191 [3.14%]	Rate : 27/42 [64.29%] Coverage : 27/191 [14.14%]

Precision [정밀도]

Recall [민감도]

71.9%
(64 / 89)33.5%
(64 / 191)

IV. 시대별 분석

시대별 분석 >>

한국인 이민자들이 미국 사회에 잘 적응하고 있는지를 알기 위해
이민 온 시기별 분석을 수행

아래 주소에서 분석 결과를 확인하실 수 있습니다

Shiny 앱

: <https://daehani.shinyapps.io/BOAZ/>

발표 영상(14분 ~ 23분)

: https://www.youtube.com/watch?v=Tc0m_Yjx_AY

분석 결과

: <https://www.kaggle.com/daehani/d/census/2013-american-community-survey/look-over-korean-immigrants-life-style>

Kaggle : Competitions Prize

kaggle

Host

Competitions

Datasets

Scripts

Jobs

Community ▾

Daehani

Logout

All Forums » USA Census

Search

«
Prev
Topic

Prizes Distribution

Start Watching

3

Thanks to everyone for sharing their code and their thoughts through all these amazing scripts! We found Kagglers' work on the US Census data inspiring and your analyses generated lots of interesting discussions around our lunch table. We've identified the prize winning scripts below, and I'll be reaching out to the authors to arrange shipping.

Kaggle Staff's Favorites



- [Wake Me Up Before You Go Go](#) by [rmnppt](#)
- [Look Over Korean Immigrants Lifestyle](#) by [Daehani](#)
- [How Does Gender Influence Wage?](#) by [si7/sk](#)

Most Upvotes

- [Should I Do a PhD?](#) by [A.M.A](#)
- [The Working Moms](#) by [huili0140](#)
- [The Richest 5%](#) by [Anton Poznyakovskiy](#)

Most Comments

- [Should I Do a PhD?](#) by [A.M.A](#)
- [Making a Map: Easy Example Using Basemap](#) by [Phil Butcher](#)
- [Effects of Month of Birth in Adulthood](#) by [Christie Haskell](#)

#1 | Posted 34 days ago



Anna Montoya
Kaggle Admin

Reply

B *I*



Kaggle : Scripts of the Week

➤ The official blog of
kaggle.com

Search

Categories

DATA SCIENCE NEWS AND
EDITORIALS

KAGGLE NEWS

SCRIPTS

SCRIPTS OF THE WEEK

TUTORIALS AND WINNERS'
INTERVIEWS

Recent Comments

MACHINE LEARNING: LINKS, NEWS AND
RESOURCES (14) | ANGEL "JAVA" LOPEZ
ON BLOG

ON RANDOM FOREST OF 'GIVE ME
SOME CREDIT' SURVEY RESULTS

ASEEM SHARMA ON
ENGINEERING PRACTICES IN
DATA SCIENCE

FAHIM ZADA
ON DEEP LEARNING HOW I DID
IT: MERCK 1ST PLACE INTERVIEW

SAIHTTAM ON HOW MUCH DID IT
RAIN? II: 2ND PLACE, LUIS ANDRE
DUTRA E SILVA

SEE THE SCRIPT ON KAGGLE & SHARE YOUR THOUGHTS IN THE COMMENTS. HOW WOULD YOU BETTER APPROACH
FINDING THE CORRECT SENTIMENTS FOR THE US & UK?

November 27: Look Over Korean Immigrants Lifestyle

Created by: **Daehani** + Seungjin Lee, Byeol Yeo, Jinhyuk Choi, Yoonyoung Choi at Big Data Club BOAZ in Korea

Public Dataset: **USA Census**

Language: RMarkdown

What motivated you to create this script?

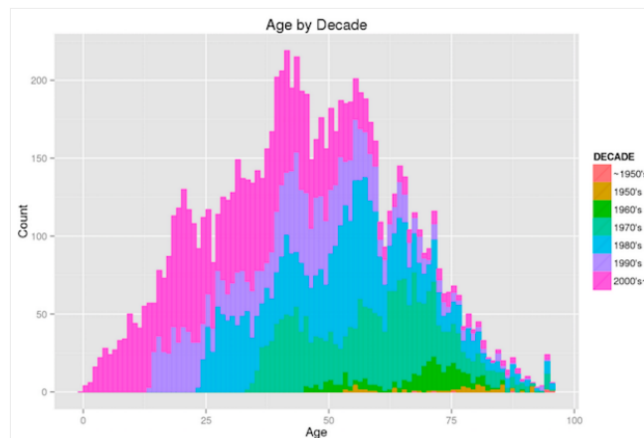
We are students in BOAZ which is a big data students' club in Korea. We found the USA Census dataset on Kaggle and wanted to connect it to Korea somehow. Finally, we had the idea to look at Korean immigrants' history in the USA.

What did you learn from the code/output?

We learned a lot about exploratory analysis in R by creating the script. And, we could better understand Korean immigrants' history and their life style via the Census Data, not from a textbook!

What can other data scientists learn from your script?

We spent almost all our time on how we could visualize the data effectively and deliver a message. There were many consideration about color, transparency, factor position and so on. There are many areas for improvement, but we hope all people enjoyed reading through it! 😊



감사합니다
