환경범죄학에 기반한 위험 지역 요인 분석

Contents

01. 프로젝트 분석 배경

I. 범죄분석을 하는 이유
II. 우리의 분석 목표

02. 시스템 아키텍처

I. 시스템 구성도 및 데이터의 흐름(Logic)
II. 사용기술

03. 데이터 수집 및 정제

I. 데이터 선정기준과 수집방법
II. 크롤링에 사용된 기술
III. 하둡, 하이브 and 정제

04. 데이터 분석

I. Data Aanalysis – Used Data
II. Improved Methods
III. Model Evaluation
IV. Depth Analysis
V. 기대효과 및 활용방안

05. 웹 시각화

I. Web을 통한 데이터 시각화
II. 웹 구성에 사용된 응용 프로그램
III. 지도 시스템의 차별성

PART 1

프로젝트 배경

PART 1

분석배경 - Current address of Crime

PART 1

분석배경 – What is Crime ?

"범죄자와 피해자가 동시에 **특정 장소**에서 벌이는 역동적 이벤트"

PART 1 분석배경 – 선행연구 결과, 분석목표 설정

특정 지역이 전체 범죄의 일정 비율을 차지

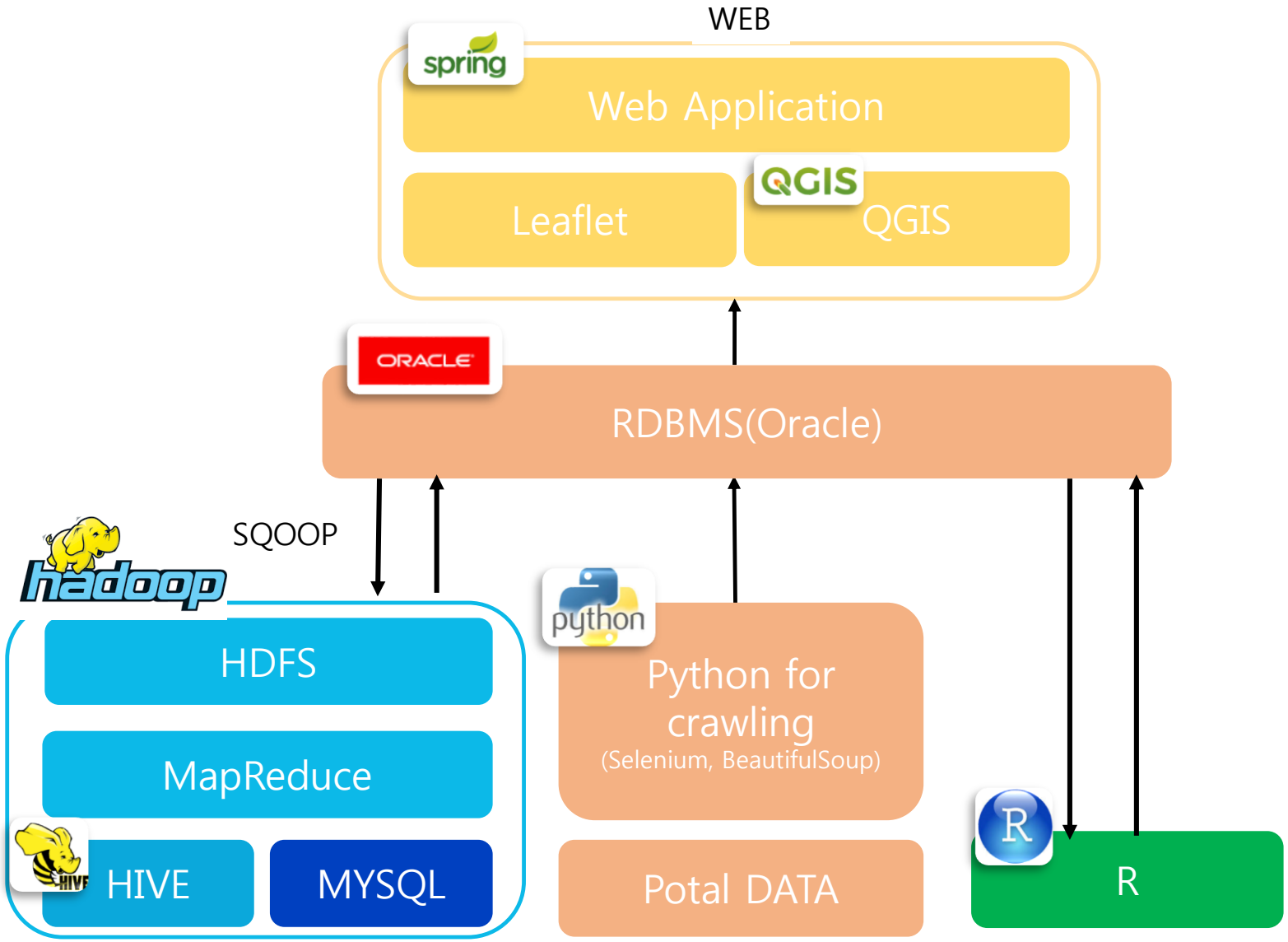
시애틀에서 14년 간의 범죄 데이터를 분석 →	범죄의 절반이 전체 도시의 4.5%에서 발생
미국 미니애폴리스 911 신고전화를 분석 →	도시 전체 주소의 3.5%에서 신고전화의 50%가 접수

PART 2

시스템 아키텍처

PART 2

System Architecture & Data Flow



PART 3

데이터 수집 및 정제

PART 3

Data Collection

05

지역별 특성 요인

GRDP, 시군구별 면적/인구,
경찰출동시간, 지역안전등급,
지역별 스트레스 인지율 등

01

지역별 교육 시설

유아교육시설,
학교, 학원시설 등

3,000,000 rows ↑

04

지역별 범죄 요인

살인, 절도, 폭행,
성폭행, 강도 등 빈도수

02

지역별 주거 요소

원룸, 아파트,
단독 주택 등 생활시설

03

지역별 시설 요인

편의점, 주유소, 병원,
음식점, 카페,
CCTV 등

PART 3

데이터 수집 - Web Crawling

```
with open('keyword_sec.txt', "r",encoding="utf-8") as f:
    for i in f.read().split("\n"):
        list_search.append(i)
    f.close()
```

전국 행정구역을 나눈 'keyword_sec.txt' 파일을 읽기 전용으로 불러와서 검색에 필요한 지역키워드로 활용

```
fcsv = open('motel_second.csv' , 'w', encoding="utf-8", newline='')
wr = csv.writer(fcsv)
```

크롤링을 통한 결과를 'motel_second.csv' 파일에 출력

```
driver.get('https://map.naver.com/')
```

네이버 지도를 검색사이트로 파싱

```
search.send_keys(list_search[i]+"편의점")

button_search = driver.find_element_by_xpath('//*
[@id="header"]/div[1]/fieldset/button').click()
time.sleep(0.5)
cnt=0
```

키워드를 입력하고 검색버튼을 클릭하여 데이터를 수집하기 위해 웹에 데이터를 표출한다

```
try:
    page = driver.find_element_by_xpath('//*
    [@id="panel"]/div[2]/div[1]/div[2]/div[2]/div/div/a[%d]' % k).click()
    time.sleep(0.5)
```

다음페이지로 이동하기 위한 버튼클릭을 수행한다

```
list_store = list_all.find_element_by_xpath('//*
[@id="panel"]/div[2]/div[1]/div[2]/div[2]/ul/li[%d]' % L)

result_data = []

store_name = list_store.find_element_by_tag_name('dt')


store_addr = list_store.find_element_by_tag_name('dd')


first = list_search[i].split(" ",1)
if len(first) == 2:
    if not first[1] in store_addr.text:
        flag1 = 0
        break
else:
    if not first[0] in store_addr.text:
        flag1 = 0
        break
```

원하는 데이터를 수집하기 위해 중복제거를 수행한다.

기본 환경셋팅

검색화면 출력

데이터 수집

PART 3

데이터 정제 - HIVE 성능 향상

USING LIKE FUNCTION

```
hadoop@debi: ~
hive> create table cctv_count as
    > select a.c_zone_concated as zone, sum(cctvl_num)as sum from crimetable a l
eft outer join (select cctvl_add, cctvl_num from cctvlocation where length(cctvl
_add) >3) b on (b.cctvl_add like concat('%',a.c_zone_concated,'%')) group by a.c
_zone_concated;
Warning: Map Join MAPJOIN[17][bigTable=?] in task 'Stage-2:MAPRED' is a cross pr
oduct
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the futu
re versions. Consider using a different execution engine (i.e. spark, tez) or us
ing Hive 1.X releases.
Query ID = hadoop_20180826150500_435b5d70-2b27-485d-817f-e8a6a9d33726
Total jobs = 1

MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 183.57 sec   HDFS Read: 42644
 HDFS Write: 7226 SUCCESS
Total MapReduce CPU Time Spent: 3 minutes 3 seconds 570 msec
OK
Time taken: 100.202 seconds
```

TEZ ENGINE | ORC FORMAT | SNAPPY COMPRESS

```
hive> create table orc_cctv_count
    >  row format delimited
    >  fields terminated by ','
    >  lines terminated by '\n'
    >  stored as orc
    >  tblproperties("orc.compress"="SNAPPY")
    >  as
    >  select a.c_zone_concated as location, sum(cctvl_num) as cnt from crimet
    > left outer join
    > (select zone, cctvl_num from orc_cctvlocation where length(zone) >3) b
    > on (b.zone = a.c_zone_concated)
    > group by a.c_zone_concated;
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0
Map 3 .......... container     SUCCEEDED      1          1        0        0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 6.15 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://localhost:9000/user/hive/warehouse/orc_cctv_co
OK
Time taken: 12.915 seconds
```

연산시간 88% 감소

PART 3

Data Handling With HIVE

지표 축소	이상값 제거	데이터 집계	데이터 통합	데이터 단위화
EX) 한식, 중식 등	결측값, 폐업 등	250개 시군구 기준 집계	EX) 학교, 원룸	인구 만명당 상가 수
↓	↓	↓	↓	↓
98개 -> 18개	약 300만건 -> 260만건	9종 항목	9종 데이터 통합	31종 데이터 단위화

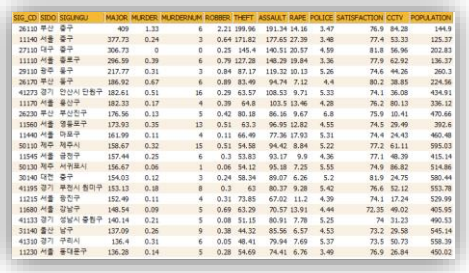
PART 4

데이터 분석

PART 4

Data Analysis – Used Data

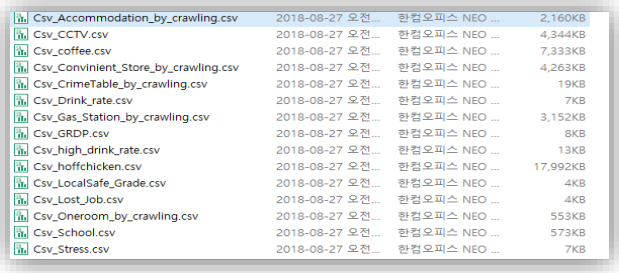
종속변수(Subordination variable)



SIG_CD	SIDO	SIGUNGU	MAJOR	MURDER	MURDERNUM	ROBBER	THEFT	ASSAULT	RAPE	POLICE	SATISFACTION	CCTV	POPULATION
26110	부산	중구	409	1.33	6	2.21	199.96	191.34	14.16	3.47	76.9	84.28	144.9
11140	서울	중구	377.73	0.24	3	0.64	171.82	177.65	27.39	3.48	77.4	53.33	125.37
27110	대구	중구	306.73	0	0	0.25	145.4	140.51	20.57	4.59	81.8	56.96	202.83
11110	서울	종로구	296.59	0.39	6	0.79	127.28	148.29	19.84	3.36	77.9	62.92	136.37
29110	광주	동구	217.77	0.31	3	0.84	87.17	119.32	10.13	5.26	74.6	44.26	260.3
26170	부산	동구	186.92	0.67	6	0.89	83.49	94.74	7.12	4.4	80.2	38.85	224.56
41273	경기	안산시 단원구	182.61	0.51	16	0.29	63.57	108.53	9.71	5.33	74.1	36.08	434.91
11170	서울	용산구	182.33	0.17	4	0.39	64.8	103.5	13.46	4.28	76.2	80.13	336.12
26230	부산	부산진구	176.56	0.13	5	0.42	80.18	86.16	9.67	6.8	75.9	10.41	470.66
11560	서울	영등포구	173.93	0.35	13	0.51	63.3	96.95	12.82	4.55	74.5	29.49	392.6
11440	서울	마포구	161.99	0.11	4	0.11	66.49	77.36	17.93	5.31	74.4	24.43	460.48
50110	제주	제주시	158.67	0.32	15	0.51	54.58	94.42	8.84	5.22	77.2	61.11	595.03
11545	서울	금천구	157.44	0.25	6	0.3	53.83	93.17	9.9	4.36	77.1	48.39	415.14
50130	제주	서귀포시	156.67	0.06	1	0.06	54.12	95.18	7.25	5.55	74.9	86.82	514.86
30140	대전	중구	154.03	0.12	3	0.24	58.34	89.07	6.26	5.2	81.9	24.75	580.44
41195	경기	부천시 원미구	153.13	0.18	8	0.3	63	80.37	9.28	5.42	76.6	52.12	553.78
11215	서울	광진구	152.49	0.11	4	0.31	73.85	67.02	11.2	4.39	74.1	17.24	529.99
11680	서울	강남구	148.54	0.09	5	0.69	63.29	70.57	13.91	4.44	72.35	49.02	405.95
41133	경기	성남시 중원구	140.14	0.21	5	0.08	51.15	80.91	7.78	5.25	74	31.23	490.53
31140	울산	남구	137.09	0.26	9	0.38	44.32	85.56	6.57	4.53	73.2	29.58	545.14
41310	경기	구리시	136.4	0.31	6	0.05	48.41	79.94	7.69	5.37	73.5	50.73	558.39
11230	서울	동대문구	136.28	0.14	5	0.28	54.69	74.41	6.76	3.49	76.9	26.84	450.02

5대 강력범죄

독립변수(Independent variable)



Csv_Accommodation_by_crawling.csv	2018-08-27 오전...	한컴오피스 NEO ...	2,160KB
Csv_CCTV.csv	2018-08-27 오전...	한컴오피스 NEO ...	4,344KB
Csv_coffee.csv	2018-08-27 오전...	한컴오피스 NEO ...	7,333KB
Csv_Convinient_Store_by_crawling.csv	2018-08-27 오전...	한컴오피스 NEO ...	4,263KB
Csv_CrimeTable_by_crawling.csv	2018-08-27 오전...	한컴오피스 NEO ...	19KB
Csv_Drink_rate.csv	2018-08-27 오전...	한컴오피스 NEO ...	7KB
Csv_Gas_Station_by_crawling.csv	2018-08-27 오전...	한컴오피스 NEO ...	3,152KB
Csv_GRDP.csv	2018-08-27 오전...	한컴오피스 NEO ...	8KB
Csv_high_drink_rate.csv	2018-08-27 오전...	한컴오피스 NEO ...	13KB
Csv_hoffchicken.csv	2018-08-27 오전...	한컴오피스 NEO ...	17,992KB
Csv_LocalSafe_Grade.csv	2018-08-27 오전...	한컴오피스 NEO ...	4KB
Csv_Lost_Job.csv	2018-08-27 오전...	한컴오피스 NEO ...	4KB
Csv_Oneroom_by_crawling.csv	2018-08-27 오전...	한컴오피스 NEO ...	553KB
Csv_School.csv	2018-08-27 오전...	한컴오피스 NEO ...	573KB
Csv_Stress.csv	2018-08-27 오전...	한컴오피스 NEO ...	7KB

지역별 요인 (38개)

Multiple Linear Regression Analsys

↓

Over Fitting

↓

Improved Methods

PART 4

Improved Methods

1. Subset selection

- 선택적으로 변수들을 골라내서 least square를 하는 방법

1. Forward
2. Backward
3. Stepwise

2. Shrinkage

- 모든 변수를 적합하며, 계수들을 0으로 수렴하거나 제한을 두는 방법

1. Ridge
2. Lasso
3. Elasticnet

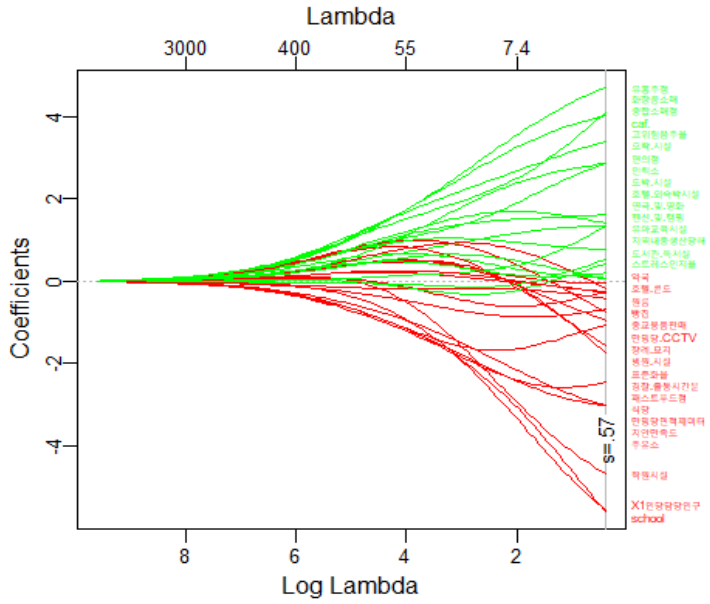
3. Dimension Reduction

- 변수 자체를 변환하여 적합하는 방법

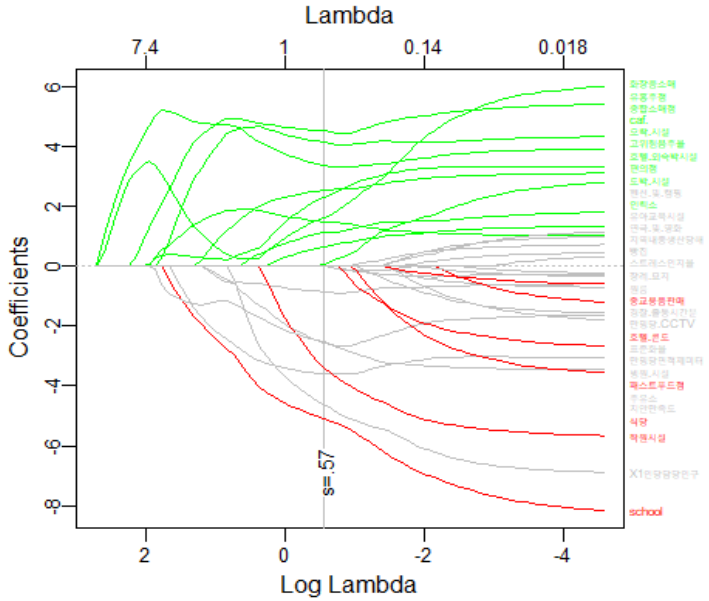
PART 4

Shrinkage Methods

Ridge -> 계수가 0에 가까워짐 (모형복잡)
Lasso -> 계수가 0으로 수렴(모형간결, 해석력↑)



Ridge



Lasso

↗(hyper parameter)
-> K-fold Cross Validation

가장 작은 CV ERROR를 보인
↗를 선정해 모형적합

PART 4

Model Evaluation

○ RMSE

	linear	forward	backward	stepwise	ridge	lasso	elasticnet
살인	0.237	0.151	0.153	0.153	0.179	0.179	0.179
강도	0.273	0.155	0.159	0.16	0.213	0.213	0.213
절도	23.728	8.654	8.803	9.793	11.091	11.091	10.536
폭력	36.77	10.863	11.045	11.874	13	13	13.342
성폭력	4.919	1.378	1.404	1.418	1.673	1.673	1.697

○ Selected Factor

	Stepwise	Lasso	Elasticnet
살인	9개	2개	5개
강도	10개	2개	5개
절도	13개	22개	28개
폭력	13개	17개	23개
성폭력	15개	16개	24개

PART 4

Depth Analysis - 폭력

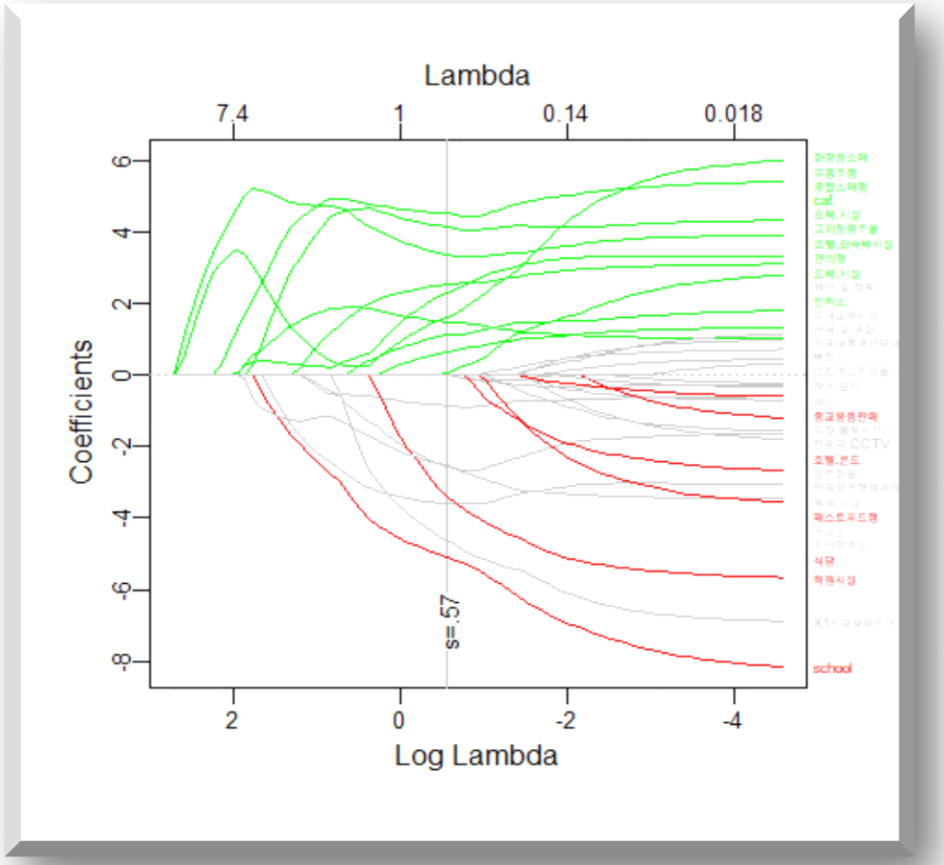
b0	T	D	M2	M3	M10	M11	M12	M15	M16	M24	M26	M27	M29	E1	E2	G	H
56.39349	-0.87173	-2.57429	2.30742	0.63825	4.49791	1.46202	-0.00745	4.11849	-3.61106	1.62747	0.04133	3.34738	1.10882	-3.41105	-5.0955	-2.54475	2.53077

비례 요인

M2(오락 시설)	M24(화장품 소매점)
M10(유흥주점)	M26(호텔 외 숙박시설)
M3(도박 시설)	M27(카페)
M11(인력소)	M29(편의점)
M15(종합소매점)	H(고위험 음주율)

반비례 요인

T(경찰 출동시간(분))	G(만명당 면적 제곱미터)
D(치안만족도)	M12(장례/묘지)
M16(주유소)	E1(학원시설)
E2(학교)	



PART 4

기대효과 및 활용방안

유흥지점 多 경찰 小

SIG_CD	시도	시군구	만명당 폭력	만명당 유흥주점
26110	부산광역시	중구	191.34	90.35414417
50130	제주특별자치도	서귀포시	95.18	49.85018946
42210	강원도	속초시	85.5	49.3449891
46110	전라남도	목포시	83.03	49.30612381

부산시 중구 **폭력 1위**

CCTV, 경찰인력 추가 배치 필요!!!

카페 多 편의점 多 종합소매점 多

⟶ 유동인구 多

시도	시군구	만명당 절도	만명당 폭력	만명당 카페	만명당 편의점	만명당 종합소매점
대구광역시	중구	145.4	140.51	0.00790206	0.000714888	0.018484949
부산광역시	중구	199.96	191.34	0.006941843	0.00063909	0.022059634
서울특별시	중구	171.82	177.65	0.006794938	0.001712355	0.038566914
서울특별시	종로구	127.28	148.29	0.005968264	0.001033921	0.019830239

대구시 중구 **폭력 4위, 절도 3위**

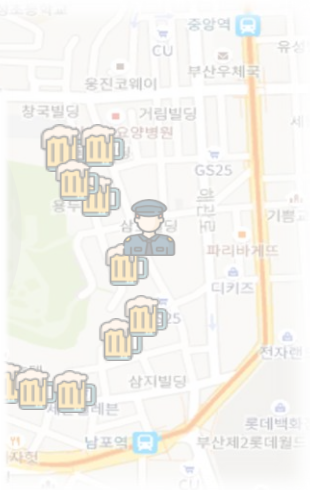
PART 4

기대효과 및 활용방안

SIG_CD	시도
26110	부산광역시
50130	제주특별자
42210	강원도
46110	전라남도

LA, 캘리포니아
총기범죄가 많이 일어난 지역에
경찰인력 추가 배치

총기범죄 발생률
22.59% 감소



폭력 1위

영국
"Trafford" 범죄 핫스팟 단속

절도 발생률
26% 감소

카페 多

"실제 세부지역 데이터로 범죄 예측 및 예방 가능"

시도	시군구	만명당				
대구광역시	중구					
부산광역시	중구	199.96	191.34	0.006941843	0.00063909	0.022059634
서울특별시	중구	171.82	177.65	0.006794938	0.001712355	0.038566914
서울특별시	종로구	127.28	148.29	0.005968264	0.001033921	0.019830239

대구시 중구 폭력 4위, 절도 3위

PART 5

웹 시각화

PART 5

웹 구성도

01 Main.jsp

- ✓ 범죄분석을 하는 이유
- ✓ 범죄사회학 관련 배경지식
- ✓ 범죄 회귀분석 사례
- ✓ 빅데이터를 활용한 범죄 예방 관련 기사

02 Chart.jsp

- ✓ 지역별 강력 5대범죄 차트 시각화
- ✓ 범죄유형별 요인 분석

03 Map.jsp

- ✓ 사용자의편의와 가독성을 높이기 위한 지도 서비스와 지역별 범죄정보 제공

Map - QGIS