데이터 마이닝을 이용한 콜택시 수요 예측 모형

데이터마이닝 1조

김단아 김준회 이재호 이현준

Index .

01 주제 선정 배경

- ✓ 택시 수요자와 공급자의 불균형 문제
- ✓ 서울시 '대기식 콜택시' 제도를 데이터 마이닝을 이용하여 구별로 예측하고자 함

01 주제 선정 배경

날씨

요일

지역
특성

휴일

02 변수의 이해

<데이터 수집 과정>

- ✓ SKT 빅데이터 허브에서 16년(1월1일~12월31일) 콜택시 발신 자료 다운로드
- ✓ 기상자료 개방포털에서 16년 서울의 일별 기상 데이터 다운로드
- ✓ 서울 열린 데이터 광장에서 16년 인구, 지리 정보 등 다운로드
- ✓ 구글링을 활용하여 직접 자료 입력

02 변수의 이해

Call, 시간 변수

- ✓ Call: 시간별 콜택시 통화 건수
- ✓ Date: 관측치의 날짜
- ✓ DayofWeek: 요일
- ✓ Is_holiday: 휴일 여부
- ✓ Time: 시간대를 morning / daytime / night로 나눔

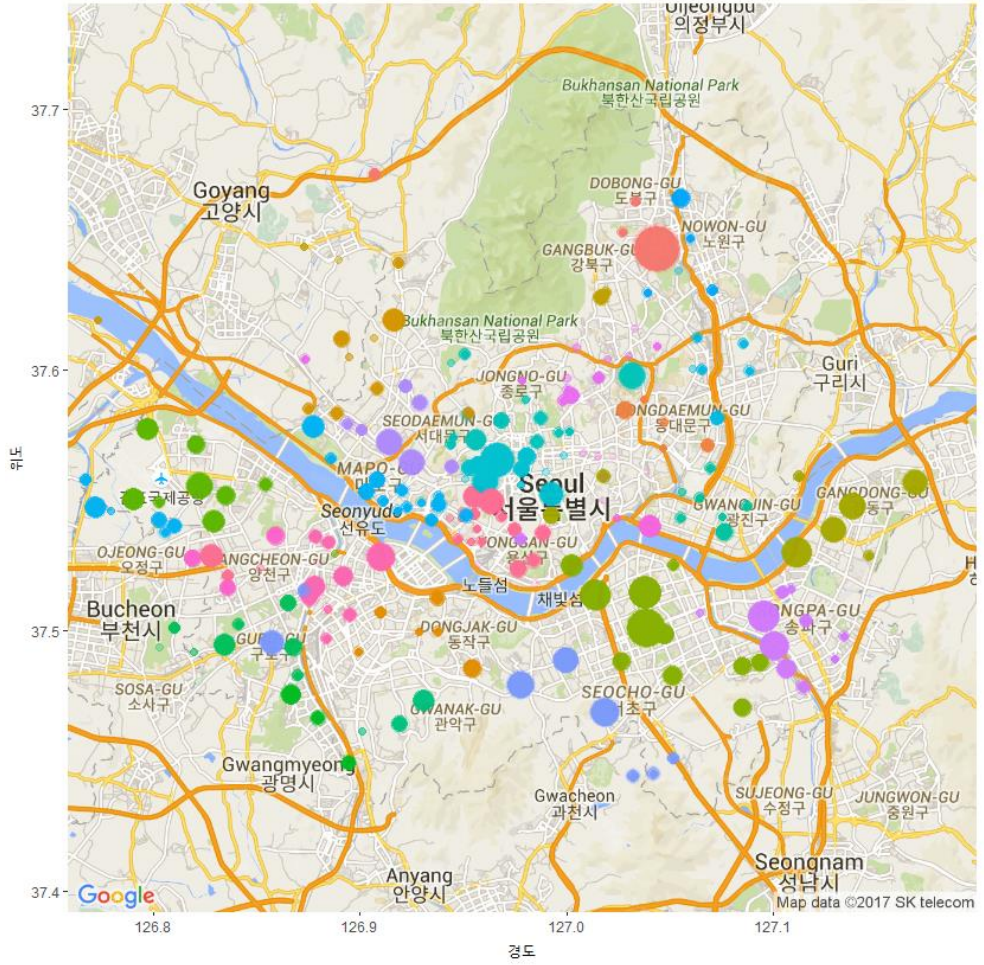
날씨 변수

- ✓ Avg_temp: 일평균 기온
- ✓ Min_temp: 최저 기온
- ✓ Max_temp: 최고 기온
- ✓ Day_rainfall: 일 강수량
- ✓ Avg_wind: 일평균 풍속
- ✓ Avg_dust: 일평균 미세먼지량
- ✓ Avg_humid: 일평균 습도

지역 변수

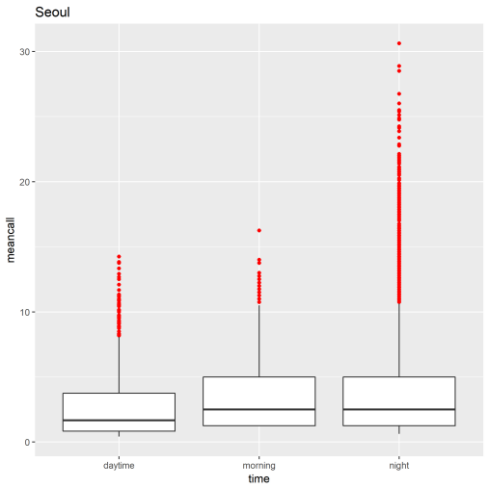
- ✓ Goo: 구(區) 이름
- ✓ Lat/Lon: 위도/경도
- ✓ Population: 주민등록 인구수
- ✓ Univ: 대학 수
- ✓ Tour: 관광 명소 수
- ✓ Worker: 사업체 종사자 수
- ✓ Pub: 유흥업소 수
- ✓ Hotel: 호텔 수

02 변수의 이해 – 데이터 수집 과정

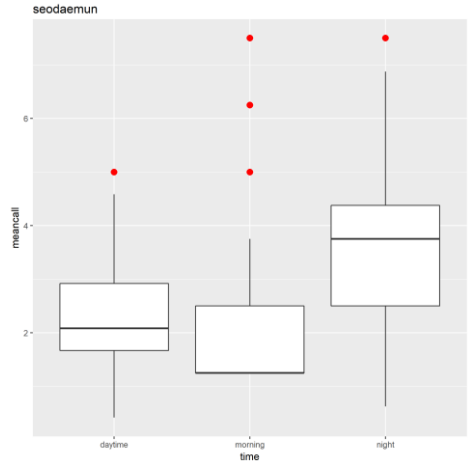


00시 ~ 23시 (1시간 간격) Skt 이용자의

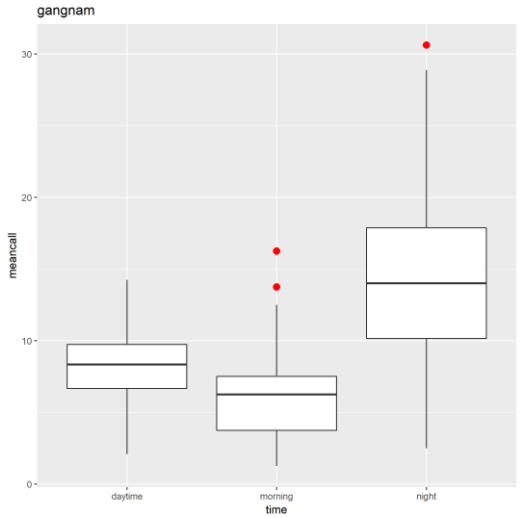
콜택시 발신 지역구별로 로드맵 시각화



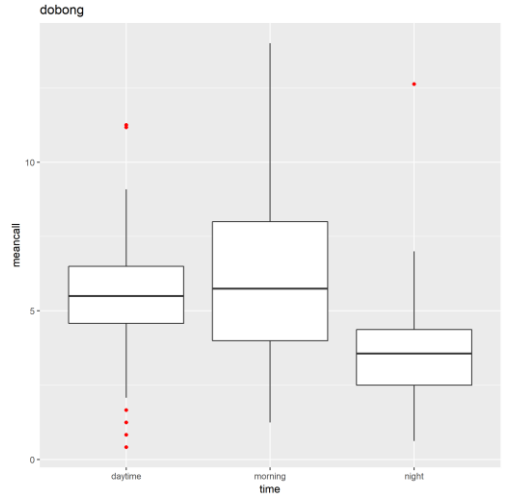
서울시 전체



서대문구



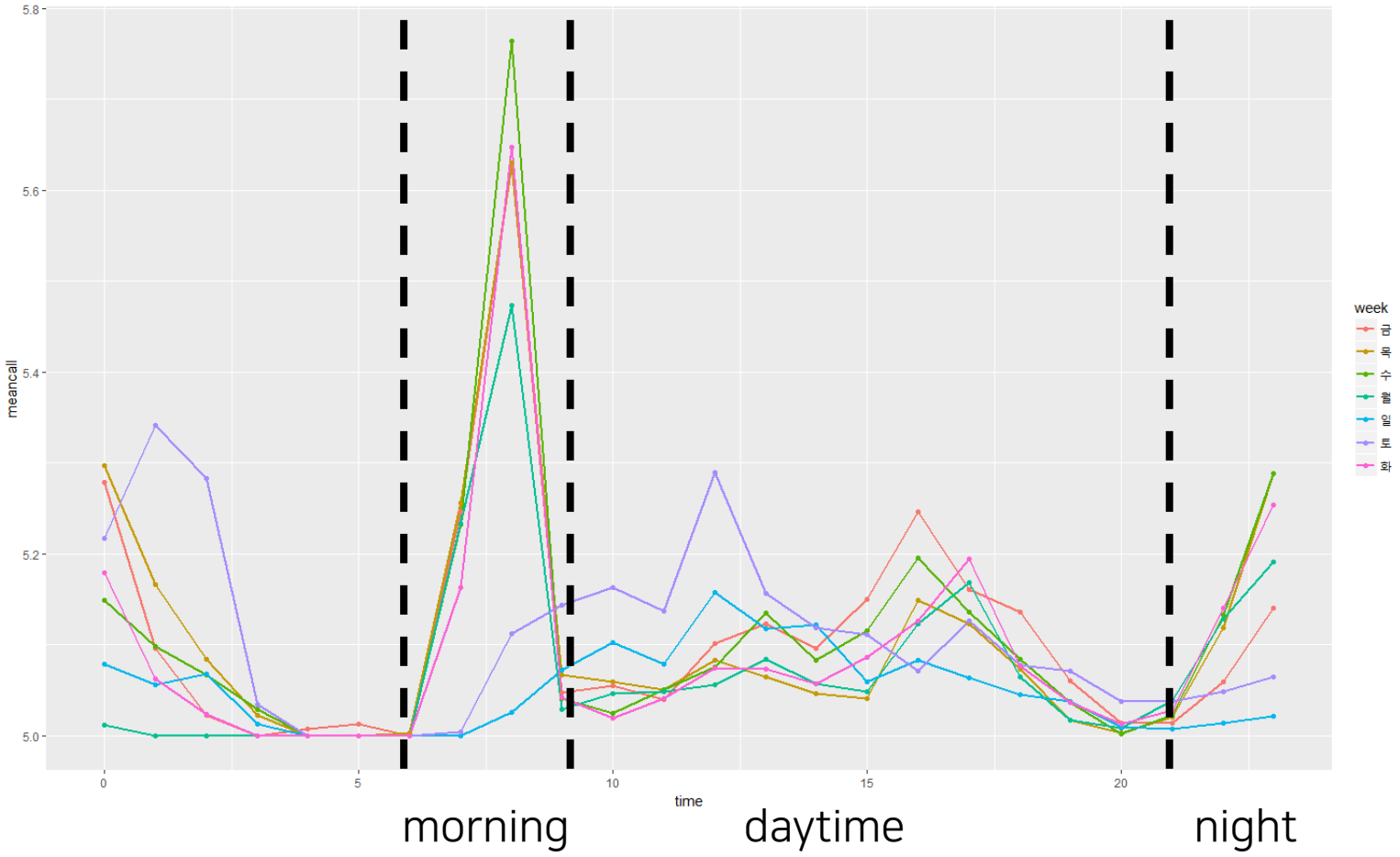
강남구



도봉구

02 변수의 이해 – 데이터 수집 과정

요일 · 시간 별 평균 콜 수 그래프



Time 범주형 변수 생성

- ✓ morning : 06시~09시(4시간)
- ✓ daytime : 10시~21시(12시간)
- ✓ night : 22시~05시(8시간)

각 시간대별 길이가 다르므로

시간당 call 수 (meancall) 이용

02 변수의 이해 – 데이터 수집 과정

‘taxicall2016’ 데이터의 모습

date	day	is_h	time	goo	lat	lon	popu	hotel	univ	tour	work	pub	avg_t	min_t	max_t	day_t	avg_w	avg_c	avg_h	mean	call
2016-01-01	fri	1	morning	gangnam	37.496	127.07	572140	60	0	14	711278	626	1.2	-3.3	4	0	1.6	58	73	8.75	35
2016-01-01	fri	1	morning	gangdong	37.549	127.15	448471	4	0	2	143061	359	1.2	-3.3	4	0	1.6	58	73	2.5	10
2016-01-01	fri	1	morning	gangbuk	37.647	127.01	330704	5	0	1	69787	333	1.2	-3.3	4	0	1.6	58	73	1.25	5
2016-01-01	fri	1	morning	geumchur	37.46	126.9	254654	3	0	1	223058	144	1.2	-3.3	4	0	1.6	58	73	1.25	5
2016-01-01	fri	1	morning	dobong	37.666	127.03	350272	1	1	1	68669	131	1.2	-3.3	4	0	1.6	58	73	5	20
2016-01-01	fri	1	morning	seocho(gu	37.477	127.04	451477	11	1	9	439963	321	1.2	-3.3	4	0	1.6	58	73	1.25	5
2016-01-01	fri	1	morning	songpa(gu	37.505	127.11	664946	14	1	6	302517	460	1.2	-3.3	4	0	1.6	58	73	5	20
2016-01-01	fri	1	morning	yeongdeu	37.521	126.91	406779	21	0	8	362524	483	1.2	-3.3	4	0	1.6	58	73	1.25	5
2016-01-01	fri	1	morning	yongsan	37.531	126.98	245102	11	1	7	133446	163	1.2	-3.3	4	0	1.6	58	73	2.5	10
2016-01-01	fri	1	morning	eunpyung	37.618	126.92	495937	6	1	1	87693	383	1.2	-3.3	4	0	1.6	58	73	1.25	5
2016-01-01	fri	1	morning	jongro	37.599	126.99	161922	37	4	59	269106	367	1.2	-3.3	4	0	1.6	58	73	1.25	5
2016-01-01	fri	1	morning	jung	37.558	126.99	134409	80	1	43	423808	453	1.2	-3.3	4	0	1.6	58	73	1.25	5
2016-01-01	fri	1	night	gangnam	37.496	127.07	572140	60	0	14	711278	626	1.2	-3.3	4	0	1.6	58	73	14.375	115
2016-01-01	fri	1	night	gangdong	37.549	127.15	448471	4	0	2	143061	359	1.2	-3.3	4	0	1.6	58	73	8.125	65
2016-01-01	fri	1	night	gangbuk	37.647	127.01	330704	5	0	1	69787	333	1.2	-3.3	4	0	1.6	58	73	0.625	5
2016-01-01	fri	1	night	gangseo	37.566	126.82	602104	23	1	1	199289	281	1.2	-3.3	4	0	1.6	58	73	1.875	15
2016-01-01	fri	1	night	gwanak	37.465	126.94	525607	5	1	3	119180	325	1.2	-3.3	4	0	1.6	58	73	1.25	10
2016-01-01	fri	1	night	guro(gu)	37.495	126.86	449600	5	3	1	210506	240	1.2	-3.3	4	0	1.6	58	73	3.75	30
2016-01-01	fri	1	night	geumchur	37.46	126.9	254654	3	0	1	223058	144	1.2	-3.3	4	0	1.6	58	73	0.625	5
2016-01-01	fri	1	night	nowon	37.655	127.08	571212	1	6	1	114736	141	1.2	-3.3	4	0	1.6	58	73	1.25	10
2016-01-01	fri	1	night	dobong	37.666	127.03	350272	1	1	1	68669	131	1.2	-3.3	4	0	1.6	58	73	6.875	55
2016-01-01	fri	1	night	dongdaen	37.584	127.05	370312	14	4	5	143858	308	1.2	-3.3	4	0	1.6	58	73	0.625	5

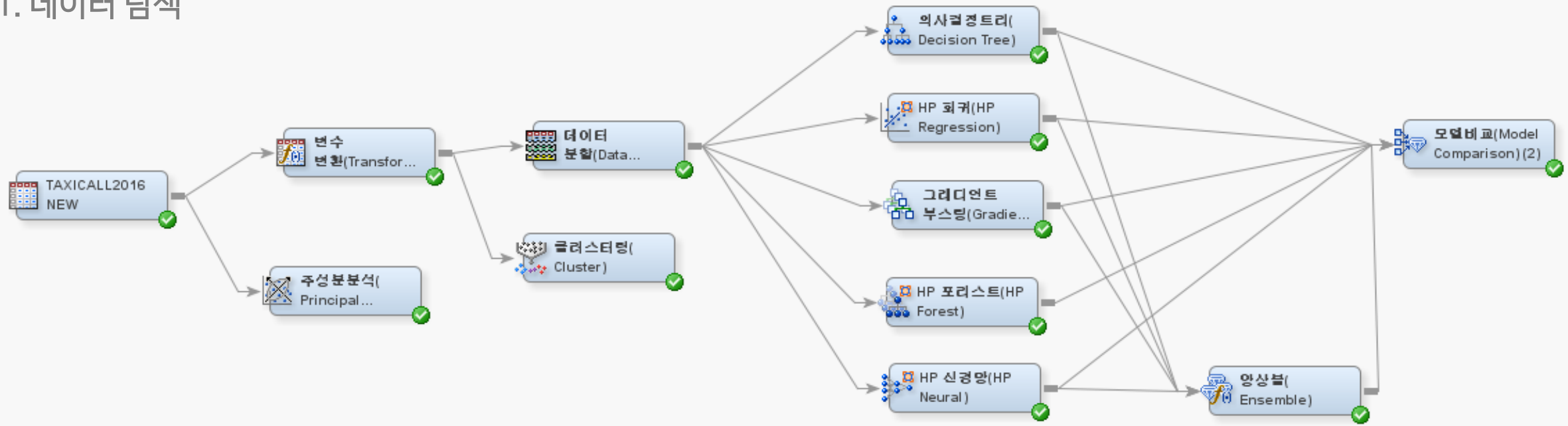
- ✓ Variable 22개
- ✓ Observation 19869개

03 데이터 탐색

SAS E-miner 다이어그램

2. 데이터 모델링

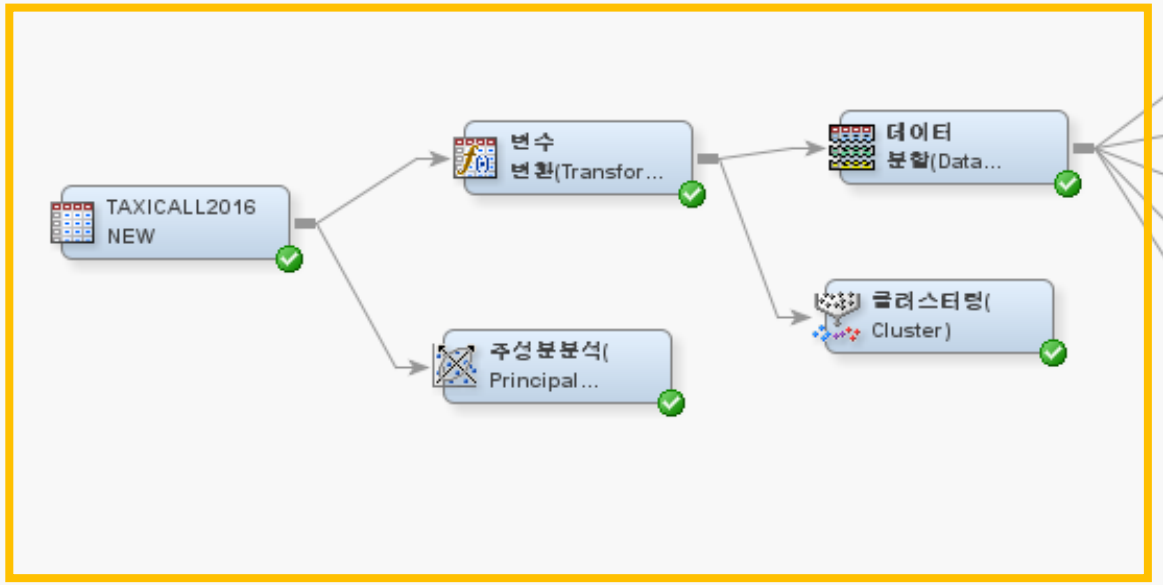
1. 데이터 탐색



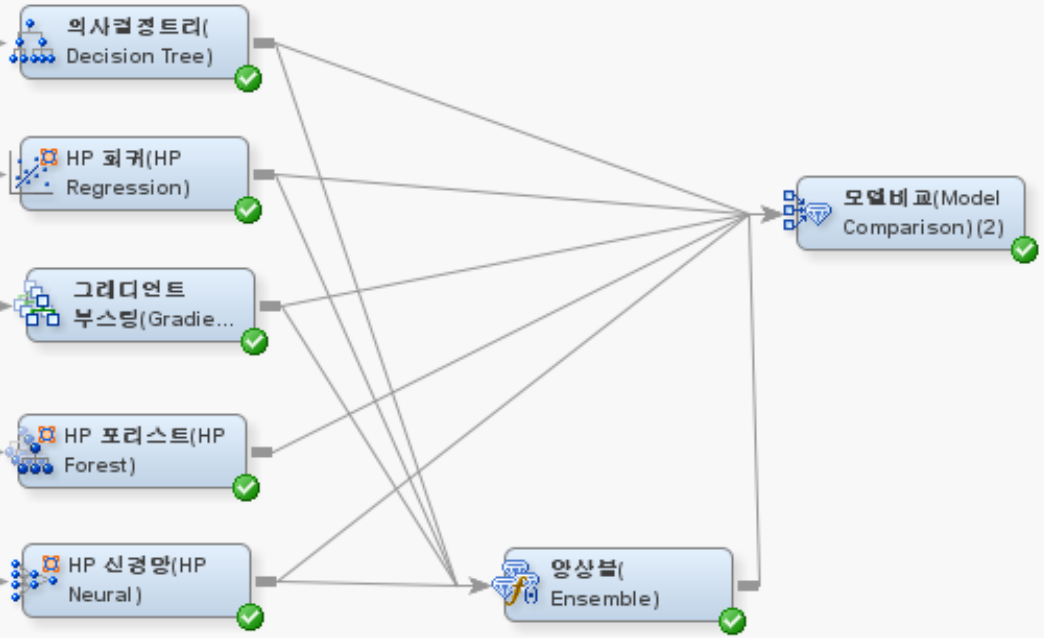
03 데이터 탐색

SAS E-miner 다이어그램

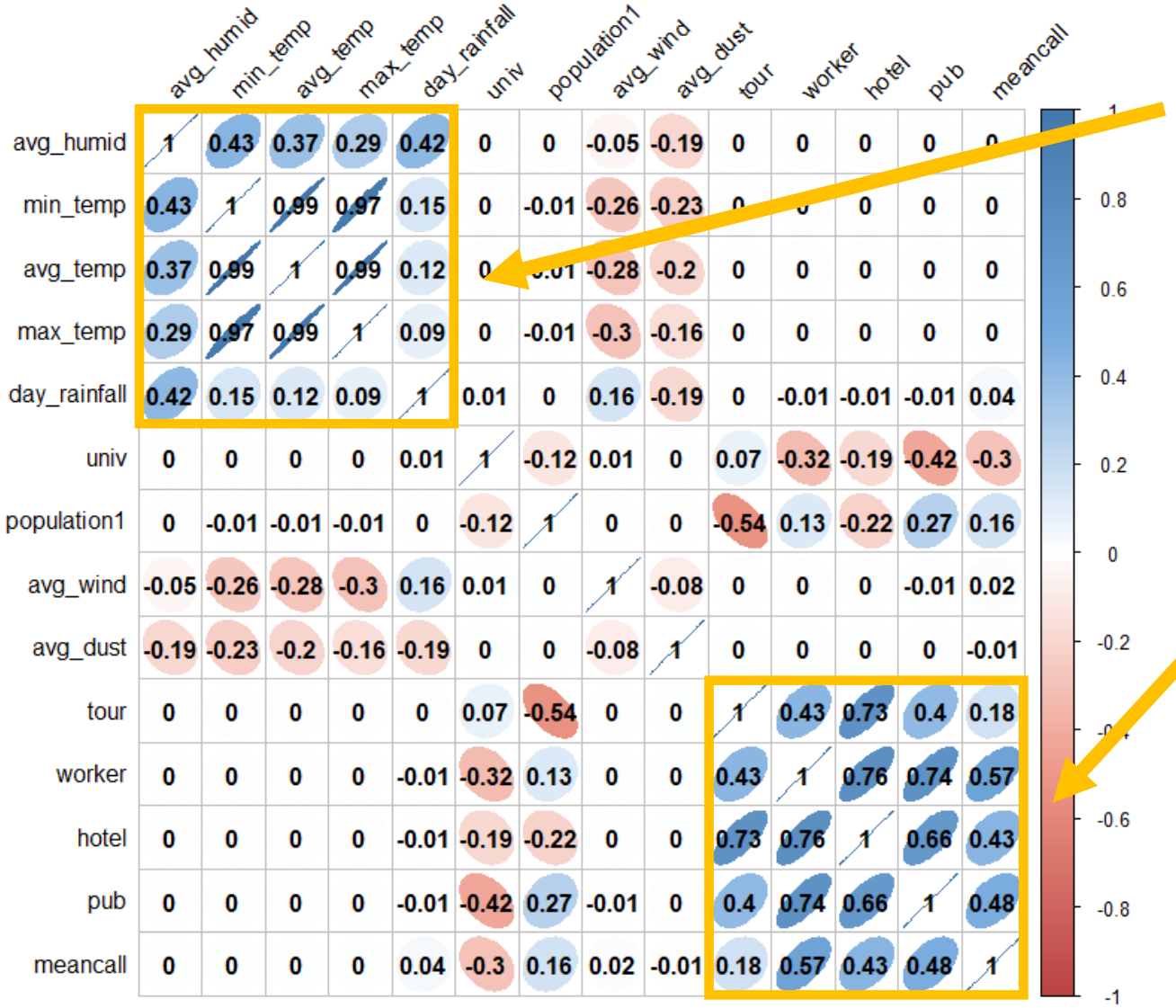
1. 데이터 탐색



2. 데이터 모델링



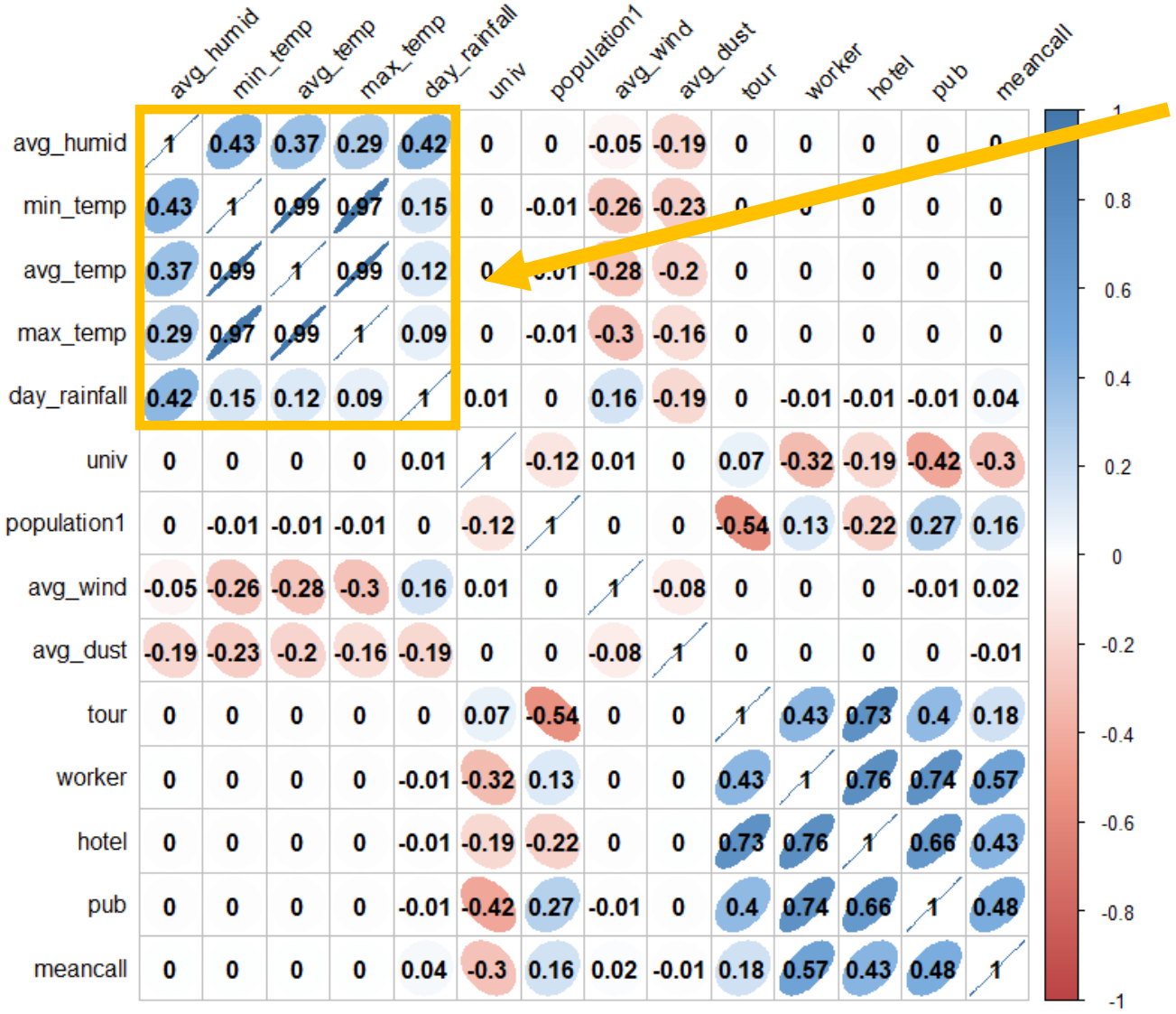
03 데이터 탐색 – Clustering



날씨 변수

지역 특성 변수

03 데이터 탐색 – Clustering



	avg_humid	min_temp	avg_temp	max_temp	day_rainfall	univ	population1	avg_wind	avg_dust	tour	worker	hotel	pub	meancall
avg_humid	1	0.43	0.37	0.29	0.42	0	0	-0.05	-0.19	0	0	0	0	0
min_temp	0.43	1	0.99	0.97	0.15	0	-0.01	-0.26	-0.23	0	0	0	0	0
avg_temp	0.37	0.99	1	0.99	0.12	0	-0.01	-0.28	-0.2	0	0	0	0	0
max_temp	0.29	0.97	0.99	1	0.09	0	-0.01	-0.3	-0.16	0	0	0	0	0
day_rainfall	0.42	0.15	0.12	0.09	1	0.01	0	0.16	-0.19	0	-0.01	-0.01	-0.01	0.04
univ	0	0	0	0	0.01	1	-0.12	0.01	0	0.07	-0.32	-0.19	-0.42	-0.3
population1	0	-0.01	-0.01	-0.01	0	-0.12	1	0	0	-0.54	0.13	-0.22	0.27	0.16
avg_wind	-0.05	-0.26	-0.28	-0.3	0.16	0.01	0	1	-0.08	0	0	0	-0.01	0.02
avg_dust	-0.19	-0.23	-0.2	-0.16	-0.19	0	0	-0.08	1	0	0	0	0	-0.01
tour	0	0	0	0	0	0.07	-0.54	0	0	1	0.43	0.73	0.4	0.18
worker	0	0	0	0	-0.01	-0.32	0.13	0	0	0.43	1	0.76	0.74	0.57
hotel	0	0	0	0	-0.01	-0.19	-0.22	0	0	0.73	0.76	1	0.66	0.43
pub	0	0	0	0	-0.01	-0.42	0.27	-0.01	0	0.4	0.74	0.66	1	0.48
meancall	0	0	0	0	0.04	-0.3	0.16	0.02	-0.01	0.18	0.57	0.43	0.48	1

1 0.8 0.6 0.4 0.2 0 -0.2 -0.4 -0.6 -0.8 -1

날씨 변수

1. 날씨변수들의 corr을 해결해보자.

(1) log(avg_wind)

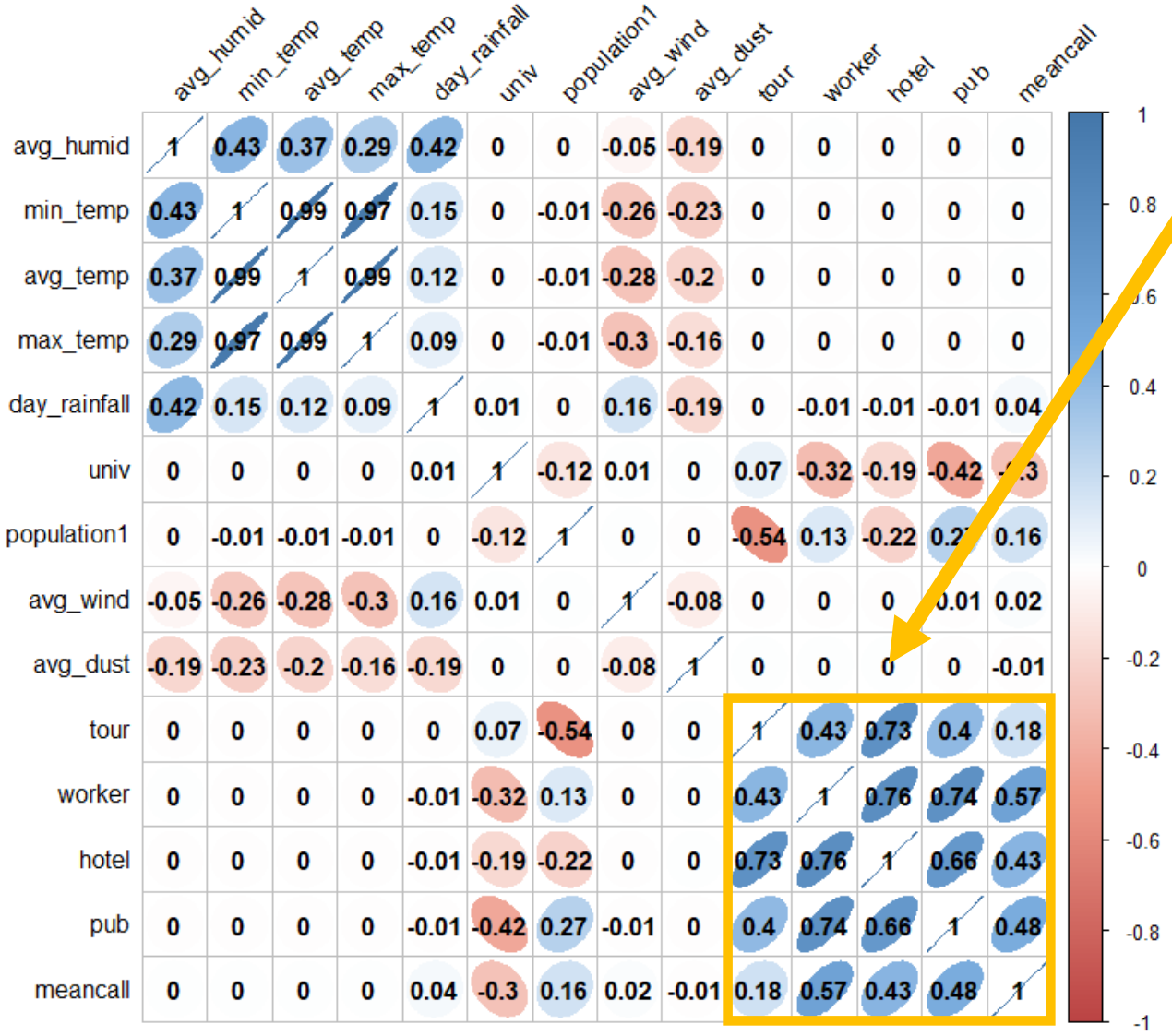
(2) log(avg_humid)

(3) log(avg_dust)

(4) avg_temp

(5) difftemp = max_temp - min_temp

03 데이터 탐색 – Clustering



	avg_humid	min_temp	avg_temp	max_temp	day_rainfall	univ	population1	avg_wind	avg_dust	tour	worker	hotel	pub	meancall
avg_humid	1	0.43	0.37	0.29	0.42	0	0	-0.05	-0.19	0	0	0	0	0
min_temp	0.43	1	0.99	0.97	0.15	0	-0.01	-0.26	-0.23	0	0	0	0	0
avg_temp	0.37	0.99	1	0.99	0.12	0	-0.01	-0.28	-0.2	0	0	0	0	0
max_temp	0.29	0.97	0.99	1	0.09	0	-0.01	-0.3	-0.16	0	0	0	0	0
day_rainfall	0.42	0.15	0.12	0.09	1	0.01	0	0.16	-0.19	0	-0.01	-0.01	-0.01	0.04
univ	0	0	0	0	0.01	1	-0.12	0.01	0	0.07	-0.32	-0.19	-0.42	-0.3
population1	0	-0.01	-0.01	-0.01	0	-0.12	1	0	0	-0.54	0.13	-0.22	0.2	0.16
avg_wind	-0.05	-0.26	-0.28	-0.3	0.16	0.01	0	1	-0.08	0	0	0	0.01	0.02
avg_dust	-0.19	-0.23	-0.2	-0.16	-0.19	0	0	-0.08	1	0	0	0	0	-0.01
tour	0	0	0	0	0	0.07	-0.54	0	0	1	0.43	0.73	0.4	0.18
worker	0	0	0	0	-0.01	-0.32	0.13	0	0	0.43	1	0.76	0.74	0.57
hotel	0	0	0	0	-0.01	-0.19	-0.22	0	0	0.73	0.76	1	0.66	0.43
pub	0	0	0	0	-0.01	-0.42	0.27	-0.01	0	0.4	0.74	0.66	1	0.48
meancall	0	0	0	0	0.04	-0.3	0.16	0.02	-0.01	0.18	0.57	0.43	0.48	1

지역 특성 변수

1. hotel, tour, worker, pub의 corr을 해결해보자.

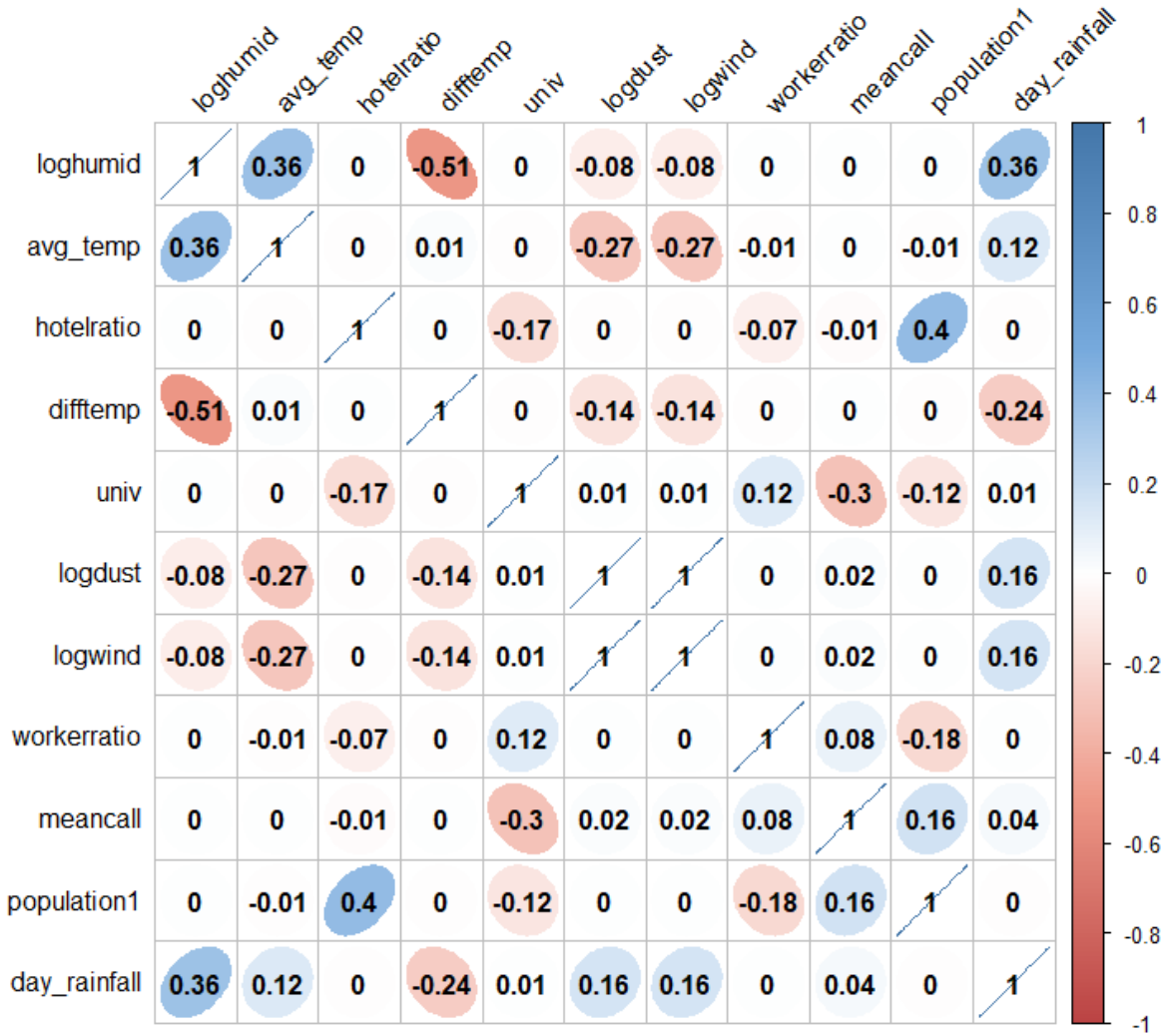
(1) hotel ~ tour corr. = 0.73

=> tour ratio = log(hotel/tour)

(2) worker ~ pub corr. = 0.74

=> pubratio = log(pub/worker)

03 데이터 탐색 – Clustering



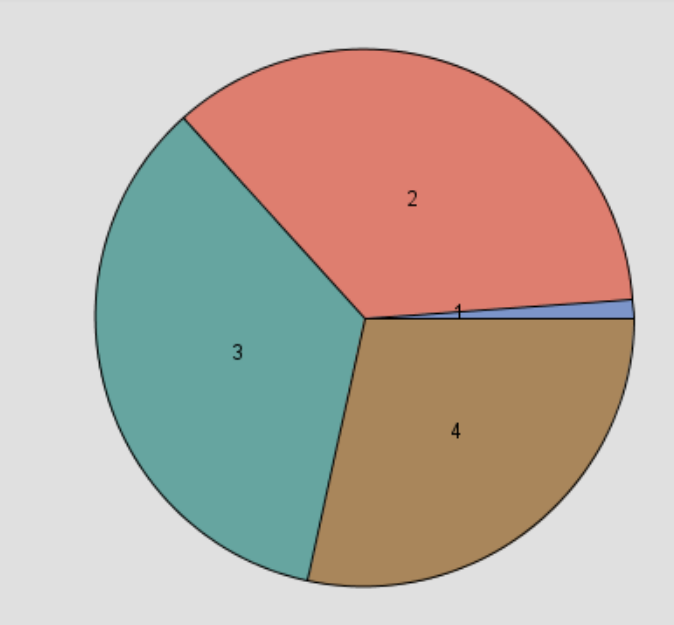
변수 변환 과정을 통해

- ✓ Correlation을 낮춰줌
- ✓ 분산을 안정화시켜줌
- ✓ 상대적 지표로 바꾸어줌

-> Clustering 가능

03 데이터 탐색 – Clustering

변수 변환 후 clustering 결과

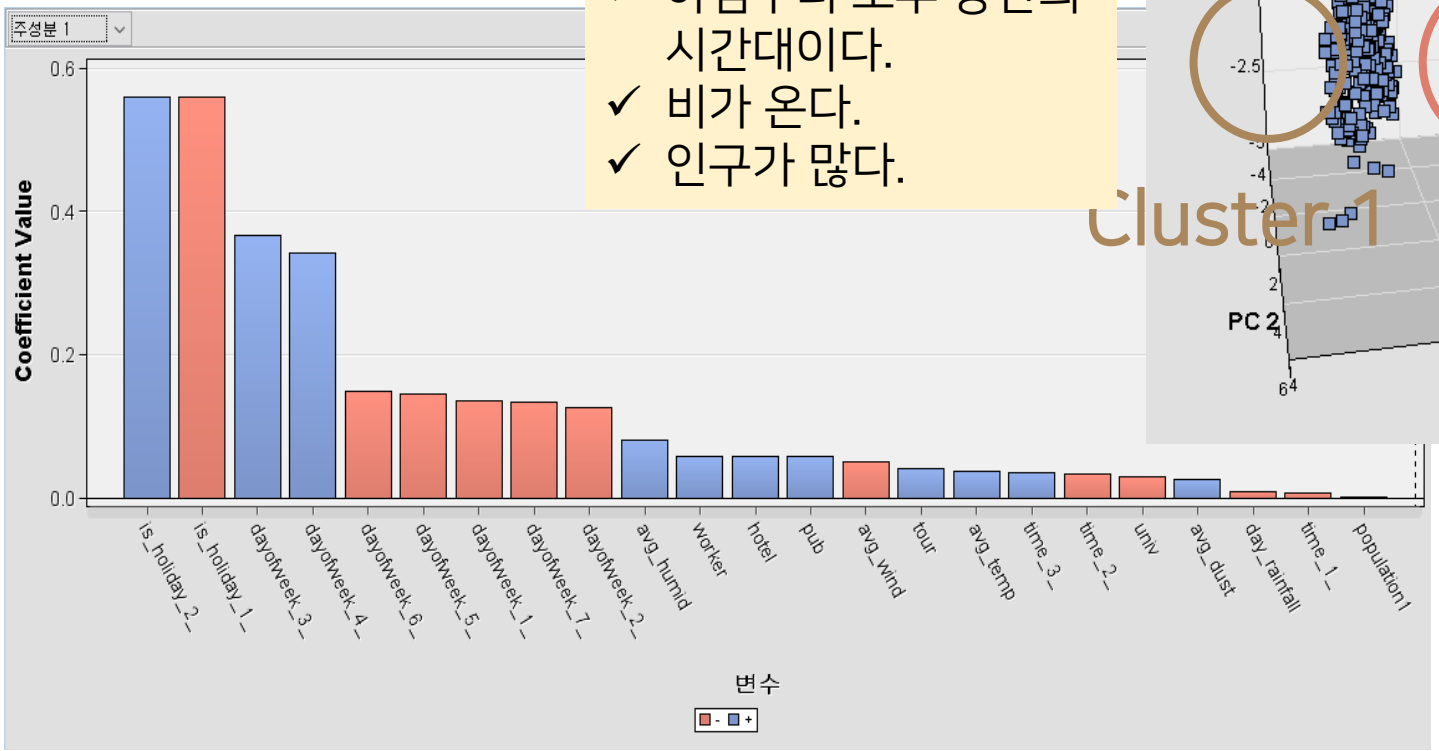


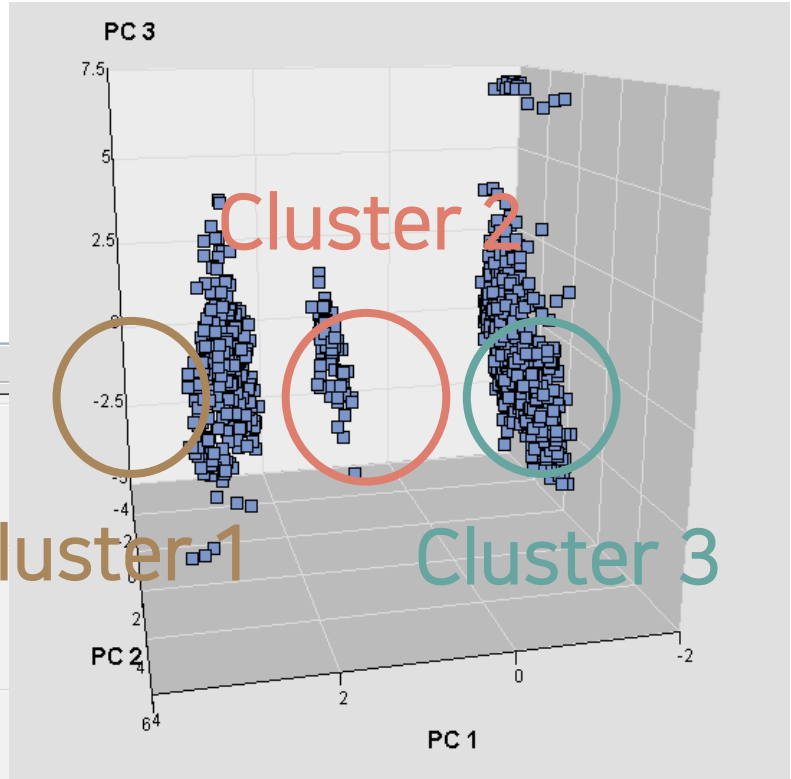
✓ 3개의 군집으로
묶이는 것을 확인

03 데이터 탐색 – PCA

PCA 결과

PC1이 작다는 것은..
- ✓ Holiday이 아니다.
- ✓ 주중이다.
- ✓ 바람이 많이 분다.
- ✓ 산업종사자수가 많다.
- ✓ 아침부터 오후 동안의 시간대이다.
- ✓ 비가 온다.
- ✓ 인구가 많다.





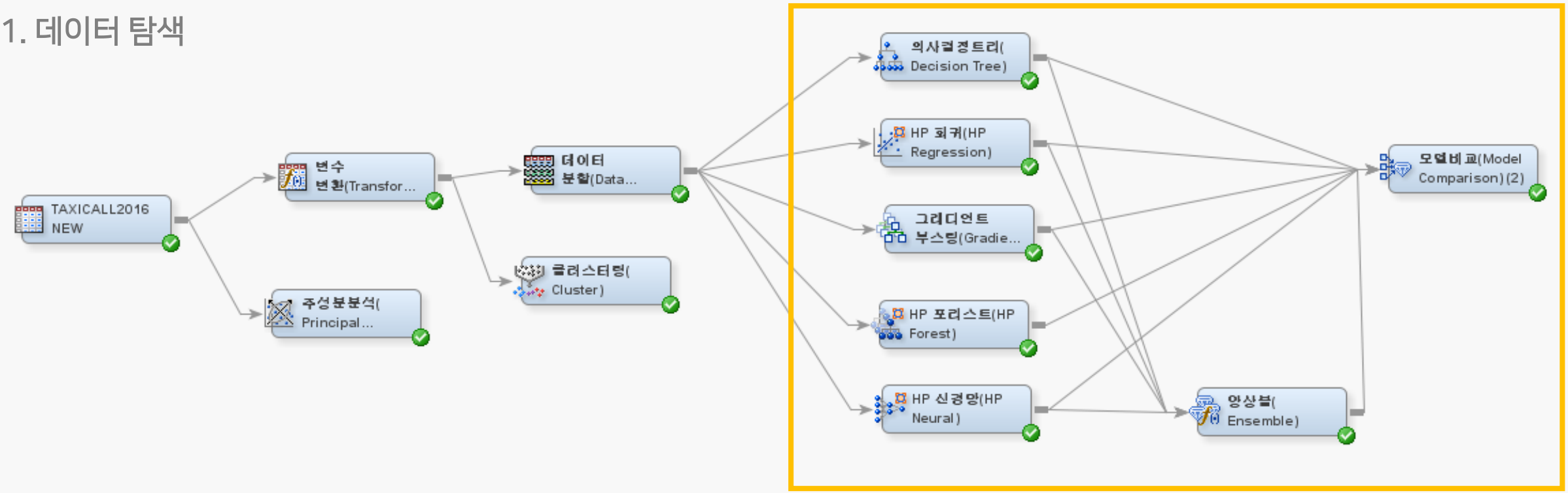
PC1이 크다는 것은..
- ✓ Holiday이다.
- ✓ 주말이다.
- ✓ 습도가 높다.
- ✓ 산업종사자수가 많다.
- ✓ 호텔이 많다.
- ✓ 술집이 많다.
- ✓ 관광지가 많다.
- ✓ 평균 기온이 낮다.
- ✓ 밤시간대이다.
- ✓ 미세먼지가 많다.

04 데이터 분석

SAS E-miner 다이어그램

1. 데이터 탐색

2. 데이터 모델링



신경망분석 WHY?

1) 콜 택시 수요 예측이 목표
2) Y변수가 연속형
3) X변수 간의 다중공선성 문제 해결
4) 모델 비교 - 적합도 높음

모델 비교

선택된 모델	선행 노드	모델 노드	모델 설명	타겟 변수	타겟 레이블	선택 기준: Valid: Average Squared Error	T o F s
Y	HPNNA	HPNNA	HP 신경망(...	meancall		2.189639	
	HPDMForest	HPDMForest	HP 포리스...	meancall		2.599922	
	Ensmbl	Ensmbl	앙상블(Ens...	meancall		3.475714	
	Boost	Boost	그래디언트 ...	meancall		4.000833	
	Tree	Tree	의사결정트...	meancall		4.464011	
	HPReg	HPReg	HP 회귀(HP...	meancall		7.319991	

* 앙상블 =>트리, 그레이디언트 부스팅, NN, 회귀분석

신경망 심층분석

1) 변수가 -1~1 사이 값을 가져야 적합도 높음
→ z-score 이용 표준화

2) 적절한 수의 변수 필요
→ 변수 선택 과정 필요

- 신경망 분석 옵션

- Train : Validation = 60 : 40
- '구'로 층화 추출
- NN : 2 히든 레이어, 18 뉴런

변수 선택 과정

1) 단계별 회귀분석 – 유의한 변수

2) Random Forest - 중요도가 높은 변수

3) 위 결과를 조합한 주관적 선택

04 데이터 분석 – 신경망분석 변수 선택

1) 단계별 회귀분석 – 6개 변수

The HPREG Procedure

Selection Summary

Step	Effect Entered	Number Effects In	p Value
0	Intercept	1	1.0000
1	univ	2	<.0001
2	time	3	<.0001
3	population1	4	<.0001
4	dayofweek	5	<.0001
5	is_holiday	6	<.0001
6	day_rainfall	7	<.0001

선택된 모델	선행 노드	모델 노드	모델 설명	타겟 변수	타겟 레이블	선택 기준: Valid: Average Squared Error	Tr ot Fr s
Y	HPNNA	HPNNA	HP 신경망...	meancall		2.097189	
	HPDMForest	HPDMForest	HP 포리스...	meancall		2.599922	
	Ensmbl	Ensmbl	앙상블(Ens...	meancall		3.440145	
	Boost	Boost	그래디언트 ...	meancall		4.000833	
	Tree	Tree	의사결정트...	meancall		4.464011	
	HPReg	HPReg	HP 회귀(HP...	meancall		7.319991	

2) Random Forest 모형 – 10개 변수

변수 이름	분리 규칙 개수	분석: 평균 제곱오차	분석: 절대 오차	OOB: 평균 제곱오차 ▼	OOB: 절대 오차
worker	990	1.660803	0.267056	1.64501	0.265920
pub	772	0.944079	0.155426	0.93806	0.154191
time	1564	0.667926	0.105780	0.67062	0.105826
tour	582	0.568098	0.114043	0.55862	0.112020
population1	727	0.553515	0.126780	0.54903	0.125982
univ	639	0.441165	0.104629	0.44082	0.103209
hotel	440	0.369846	0.080762	0.36900	0.080244
dayofweek	861	0.143762	0.024219	0.12691	0.019667
s_holiday	651	0.060919	0.013538	0.05912	0.012621
avg_temp	239	0.011763	0.002277	0.00297	0.000419
avg_humid	138	0.002829	0.000746	-0.00130	-0.000229
difftemp	117	0.001503	0.000553	-0.00168	-0.000327
avg_wind	146	0.003260	0.000836	-0.00201	-0.000367
avg_dust	166	0.004220	0.000949	-0.00267	-0.000340
day_rainfall	202	0.009360	0.001410	-0.00451	-0.000456

선택된 모델	선행 노드	모델 노드	모델 설명	타겟 변수	타겟 레이블	선택 기준: Valid: Average Squared Error
Y	HPNNA	HPNNA	HP 신경망(...	meancall		1.860429
	HPDMForest	HPDMForest	HP 포리스...	meancall		2.679397
	Ensmbl	Ensmbl	앙상블(Ens...	meancall		2.743341
	Boost	Boost	그래디언트 ...	meancall		2.857267
	Tree	Tree	의사결정트...	meancall		3.91761
	HPReg	HPReg	HP 회귀(HP...	meancall		5.7569

3) 주관적 선택 – 9개 변수

선택이유

✓ 2개의 분석 결과 나온 변수의 교집합
-> Dayofweek, Time, Is_holiday, Rainfall, population, univ, Hotel, Tour, Worker, Pub

✓ 앞의 상관분석에서 Hotel ~ Tour의 상관계수 높음
-> Hotel 변수 제거

3) 종합적 선택 – 9개 변수

Dayofweek
Time
Is_holiday
Rainfall
Population
Univ
Tour
Worker
Pub

선택된 모델	선행 노드	모델 노드	모델 설명	타겟 변수	타겟 레이블	선택 기준: Valid: Average Squared Error
/	HPNNA	HPNNA	HP 신경망(..	meancall		1.864292
	HPDMForest	HPDMForest	HP 포리스...	meancall		2.400784
	Ensmbl	Ensmbl	앙상블(Ens...	meancall		2.865891
	Boost	Boost	그래디언트 ...	meancall		2.895238
	Tree	Tree	의사결정트...	meancall		3.91761
	HPReg	HPReg	HP 회귀(HP...	meancall		7.319991

04 데이터 분석 – 최종 선택

변수 최종 선택 – 종합적 변수 선택 NN모델

Error 0.004 가량 상승하였지만,

✓ 변수 1개를 제거하는 것에 비해 미미한 변화

✓ 모델의 overfitting 문제 예방

✓ 데이터 수집비용 절약 차원에서 적은 변수가 유리

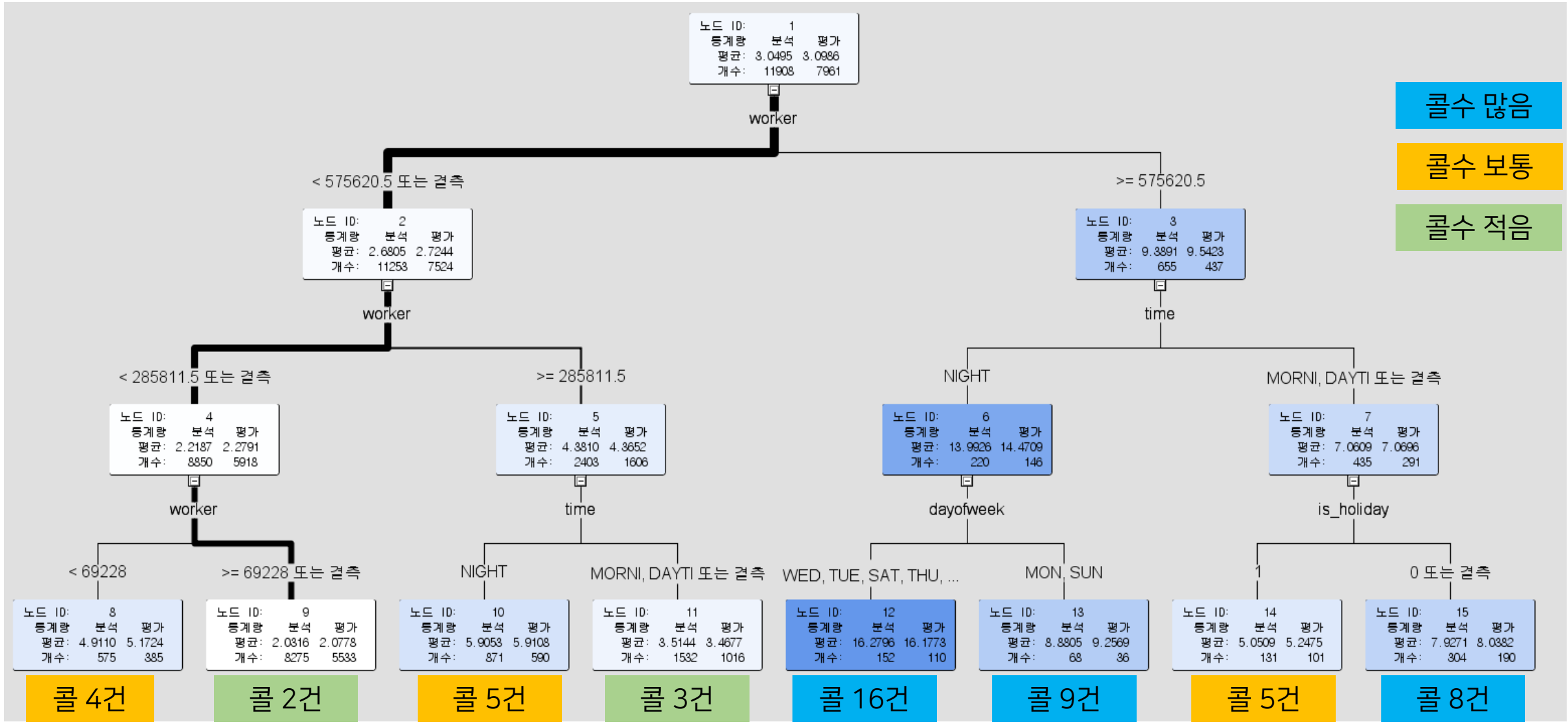
최종 모형으로 **9개**의 변수 사용

*Dayofweek, Time, Is_holiday, Rainfall, Tour, Worker, Pub, Population, Univ

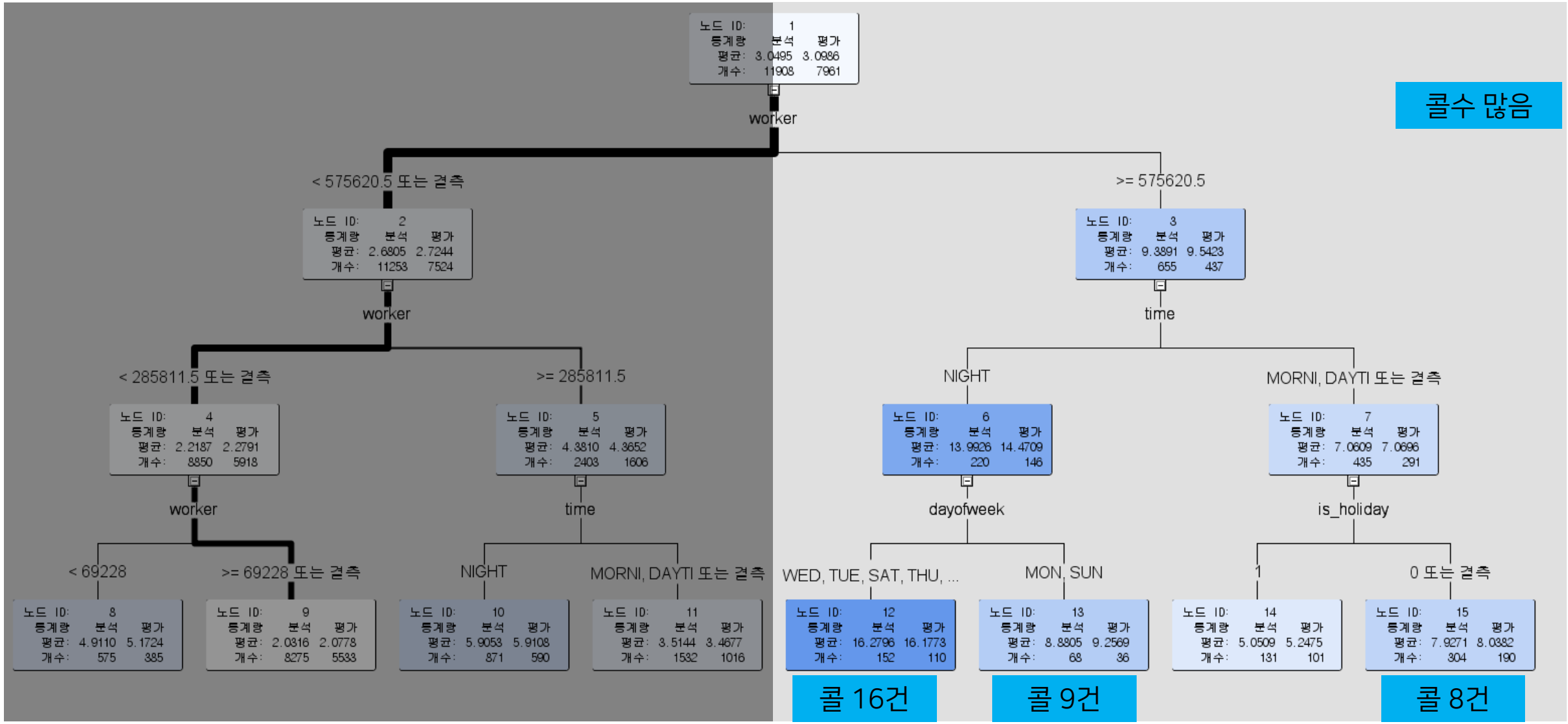
04 데이터 분석 – 최종 선택

관측치 수	id	date △	dayofweek	time	goo	meancall	Predicted: meancall	Residual: meancall	population1
177,0	13,0	2016, 1, 1	fri	night	gangnam	14,375	13,652898822172428	0,7221011778275717	572140,0
10932,0	56,0	2016, 1, 1	fri	dayti	jung	1,666666667	1,8428216110811841	-0,17615494408118404	134409,0
12451,0	33,0	2016, 1, 1	fri	night	jongro	1,25	2,6332396118290213	-1,3832396118290213	161922,0
15454,0	39,0	2016, 1, 1	fri	dayti	gwangji	0,833333333	0,8462558652716141	-0,012922532271614107	372104,0
6679,0	37,0	2016, 1, 1	fri	dayti	gangseo	3,333333333	3,5782842953620357	-0,24495096236203562	602104,0
3414,0	35,0	2016, 1, 1	fri	dayti	gangnam	5,0	6,501757372712156	-1,5017573727121558	572140,0
18309,0	41,0	2016, 1, 1	fri	dayti	nowon	0,416666667	0,8715041939042979	-0,4548375269042979	571212,0
18310,0	43,0	2016, 1, 1	fri	dayti	dongdae	0,416666667	0,7223907667941454	-0,3057240997941454	370312,0
12448,0	17,0	2016, 1, 1	fri	night	gwanak	1,25	1,1725965186057445	0,07740348139425546	525607,0
364,0	28,0	2016, 1, 1	fri	night	songpa(	11,25	5,275957273205596	5,974042726794404	664946,0
926,0	1,0	2016, 1, 1	fri	morni	gangnam	8,75	4,350157284455906	4,399842715544094	572140,0
18898,0	51,0	2016, 1, 1	fri	dayti	yangchu	0,416666667	0,9419913401478244	-0,5253246731478244	481845,0
12450,0	24,0	2016, 1, 1	fri	night	seodaem	1,25	2,9021115425877317	-1,6521115425877317	325871,0
12449,0	20,0	2016, 1, 1	fri	night	nowon	1,25	1,248012663861982	0,0019873361380180476	571212,0
16656,0	27,0	2016, 1, 1	fri	night	sungbuk	0,625	1,254676110778287	-0,6296761107782869	461617,0
11590,0	49,0	2016, 1, 1	fri	dayti	sungbuk	1,25	1,0547910763430024	0,1952089236569976	461617,0
16655,0	22,0	2016, 1, 1	fri	night	dongdae	0,625	1,1141799644448627	-0,48917996444486267	370312,0
5676,0	23,0	2016, 1, 1	fri	night	mapo(gu	3,75	3,9700302534941248	-0,22003025349412475	390887,0
16661,0	78,0	2016, 1, 2	sat	night	mapo(gu	0,625	5,258873820105114	-4,633873820105114	390887,0
16659,0	76,0	2016, 1, 2	sat	night	dongdae	0,625	1,0390661076169576	-0,4140661076169576	370312,0
3930,0	58,0	2016, 1, 2	sat	morni	gangnam	5,0	4,420751523220632	0,5792484767793677	572140,0
5678,0	75,0	2016, 1, 2	sat	night	dobong	3,75	4,098949657483216	-0,3489496574832156	350272,0
5679,0	81,0	2016, 1, 2	sat	night	songpa(	3,75	6,436798460762872	-2,686798460762872	664946,0
16658,0	73,0	2016, 1, 2	sat	night	geumchu	0,625	1,5506274681589893	-0,9256274681589893	254654,0

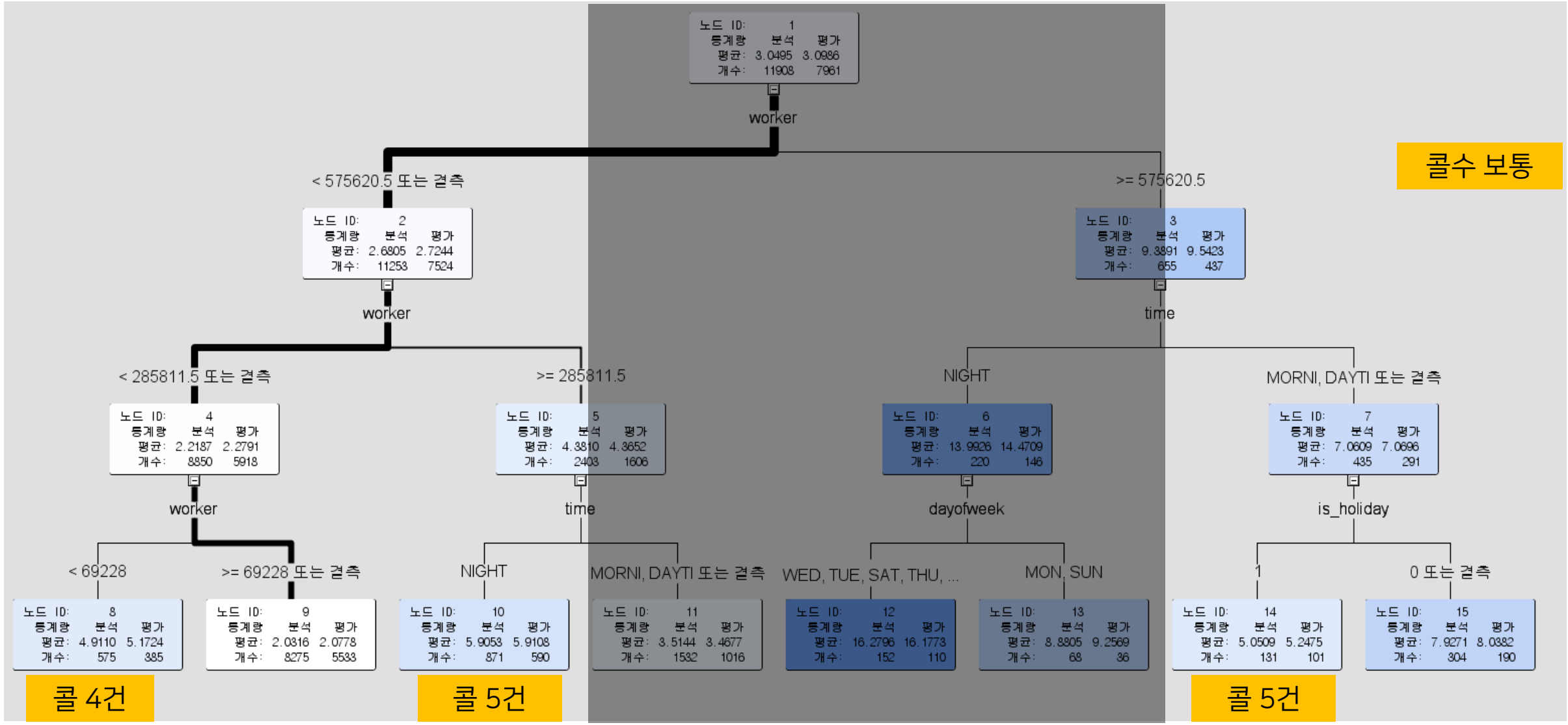
05 모형 해석 – Decision tree



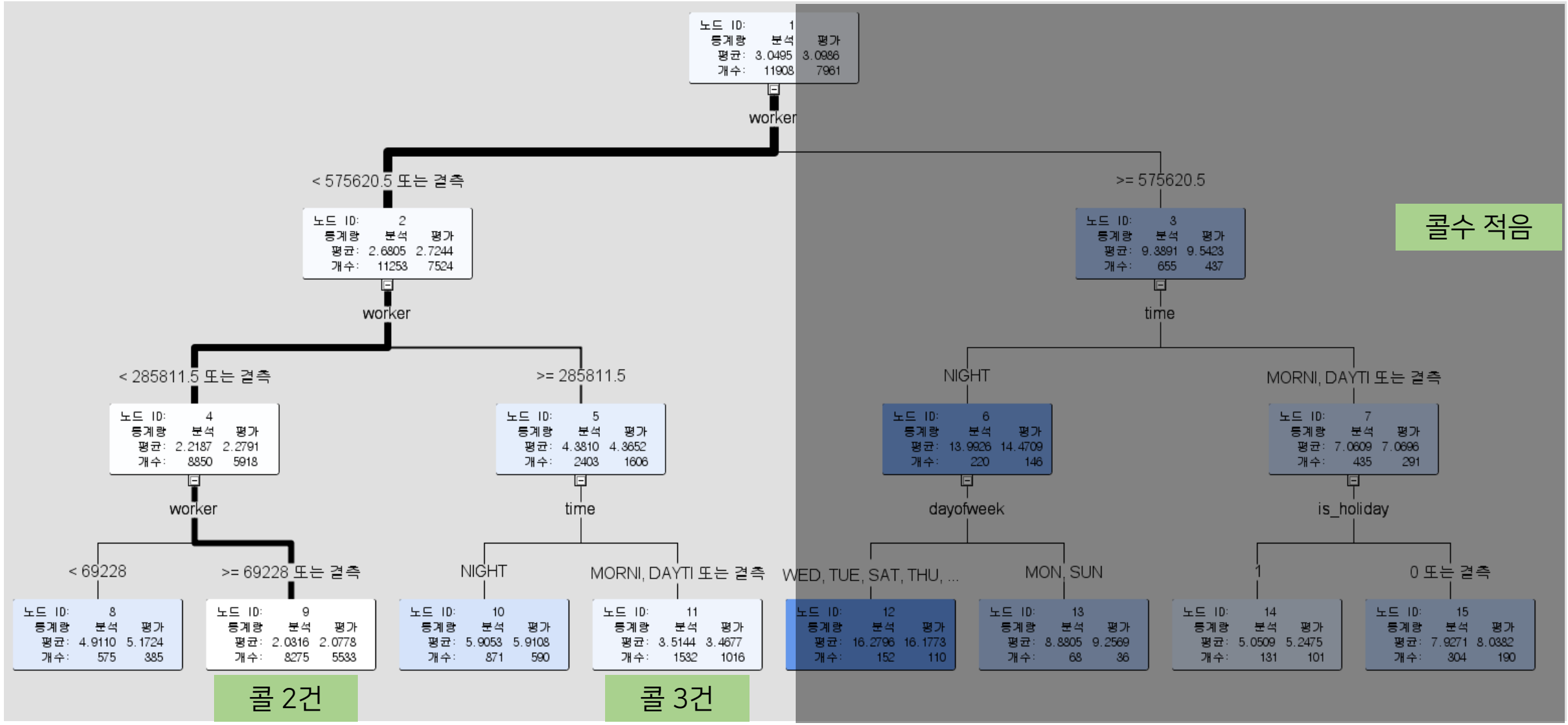
05 모형 해석 – Decision tree



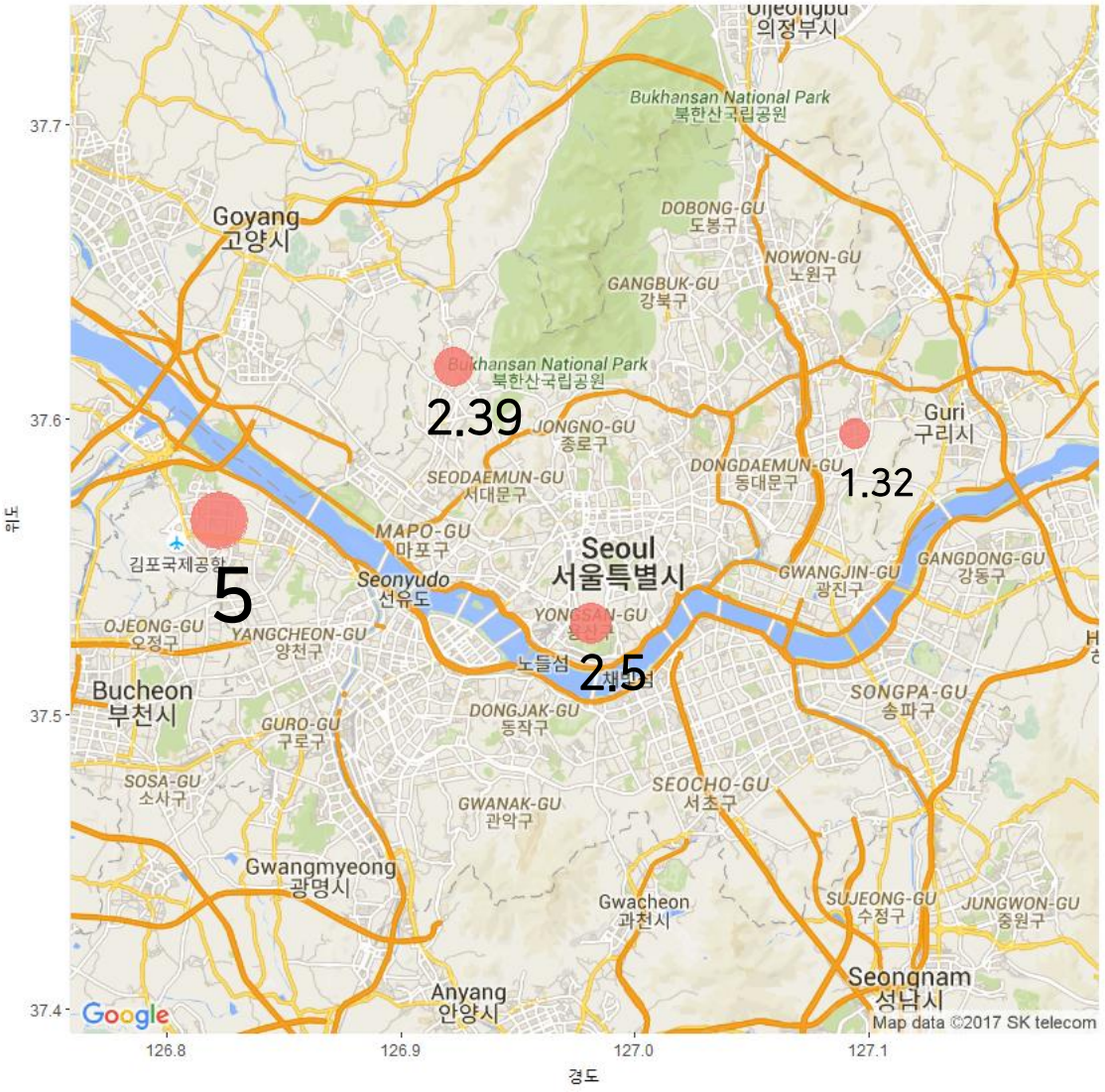
05 모형 해석 – Decision tree



05 모형 해석 – Decision tree



05 모형 검증 – test set 발췌



2016년 7월 1일 금요일 morning

1일 강우량 108.5mm

goo	worker	pub	hotel	tour	univ	Predicted Y	Y	Residual
gangseo	199289	281	23	1	1	5.03	6.25	1.22
yongsan	133446	163	11	7	1	2.51	2.50	-0.01
eunpyun	87693	383	6	1	1	2.39	2.50	0.11
jungran	99241	134	3	1	0	1.32	1.25	-0.07

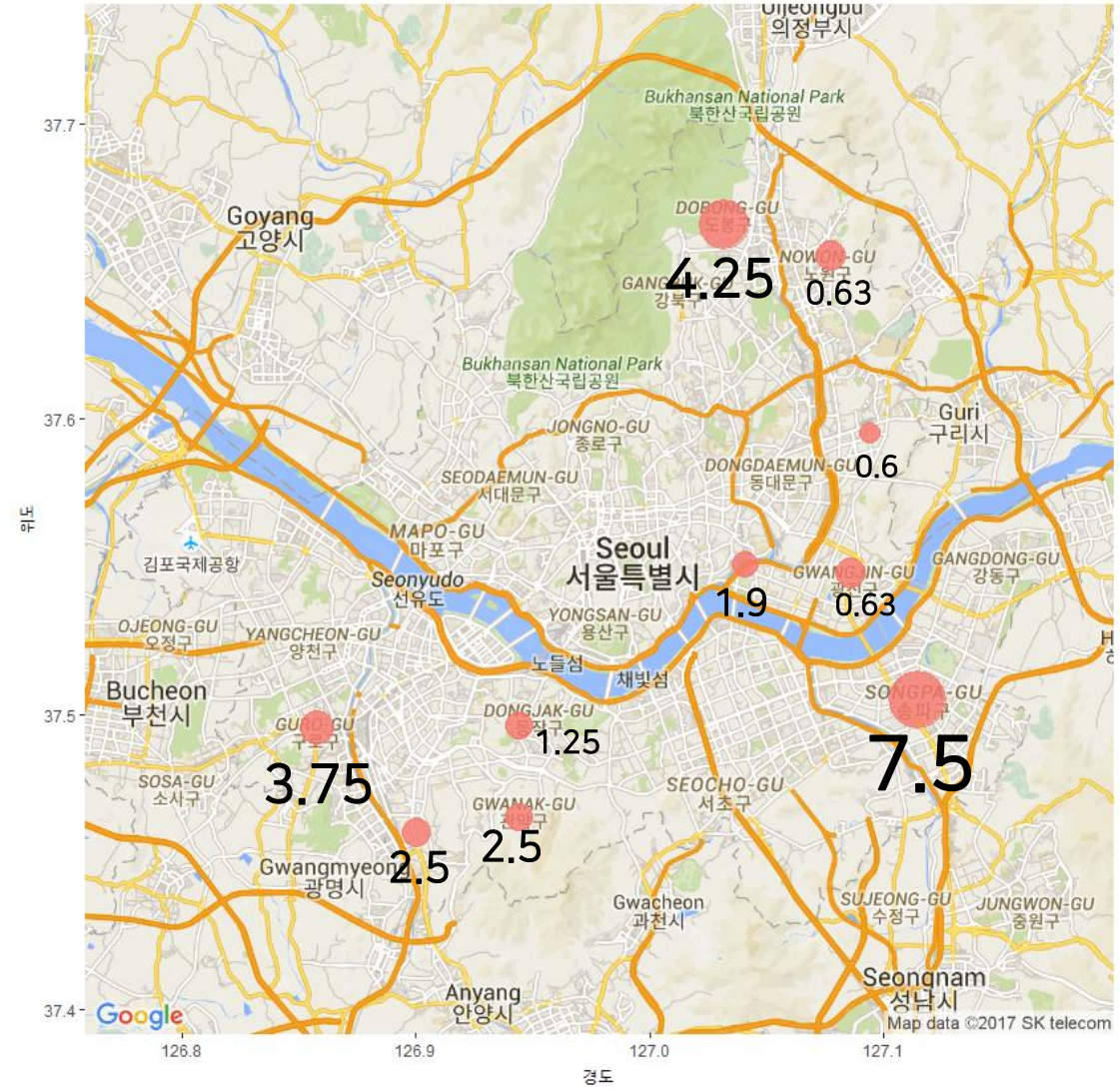
05 모형 검증 - test set 발췌

2016년 7월 1일 금요일 daytime

1일 강우량 108.5mm

goo	worker	pub	hotel	tour	univ	Predicted Y	Y	Residual
gangnam	711278	626	60	14	0	11.91	14.25	2.34
nowon	114736	141	1	1	6	1.26	0.83	-0.43
seodaem	111615	128	5	3	6	3.03	2.08	-0.94
sungdon	162019	74	7	1	1	1.32	0.42	-0.90
sungbuk	113893	87	5	1	7	1.52	0.83	-0.69
yongsan	133446	163	11	7	1	4.45	3.33	-1.12
eunpyun	87693	383	6	1	1	2.59	2.50	-0.09

05 모형 검증 - test set 발췌



2016년 7월 1일 금요일 night

1일 강우량 108.5mm

goo	worker	pub	hotel	tour	univ	Predicted Y	Y	Residual
gwanak	119180	325	5	3	1	1.39	2.50	1.11
gwangji	123689	108	6	3	3	1.69	0.63	-1.06
guro	210506	240	5	1	3	2.09	3.75	1.66
geumchu	223058	144	3	1	0	1.47	2.50	1.03
nowon	114736	141	1	1	6	1.58	0.63	-0.96
dobong	68669	131	1	1	1	4.79	4.25	-0.54
dongjak	103915	158	3	1	3	1.31	1.25	-0.06
sungdon	162019	74	7	1	1	1.20	1.88	0.68
songpa(	302517	460	14	6	1	6.20	7.50	1.30
jungran	99241	134	3	1	0	0.76	0.63	-0.14

06 결론 및 한계점

결론

✓ **수집할 데이터 비용 절약**

가장 유의미한 변수 9개: Dayofweek, Time, Is_holiday, Day_rainfall, Tour, Worker, Pub, Population, Univ

=>의사결정트리 해석 – 9개 변수 中 종사자 수 가장 유의

=>온도 x

✓ **모델 활용을 통한 사업비 절약**

변수 입력 ->콜 택시 수요 예측 -> 적재적소 콜 택시 배차

06 결론 및 한계점

한계점

✓ **교통정보 통계(일반택시 승하차 수 혹은 유동인구) 반영 필요**

=> 더 정확한 콜 택시 수요 예측

✓ **공휴일 고유 특성에 따른 예측오차**

ex) 강남구, 설날 -> 서울 유출 인구 多

✓ **지역 특성을 반영하는 변수의 한계**

ex) 도봉구 창동 -> 장애인 및 노인시설 집중

감사합니다.

데이터마이닝 1조

김단아 김준회 이재호 이현준