LOCATING THE NEXT STARBUCKS IN SEOUL

1팀 범주형 자료분석

고은영 계승환

전연호 오희준

김민구 김주영

TABLE OF CONTENTS

1. 주제선정

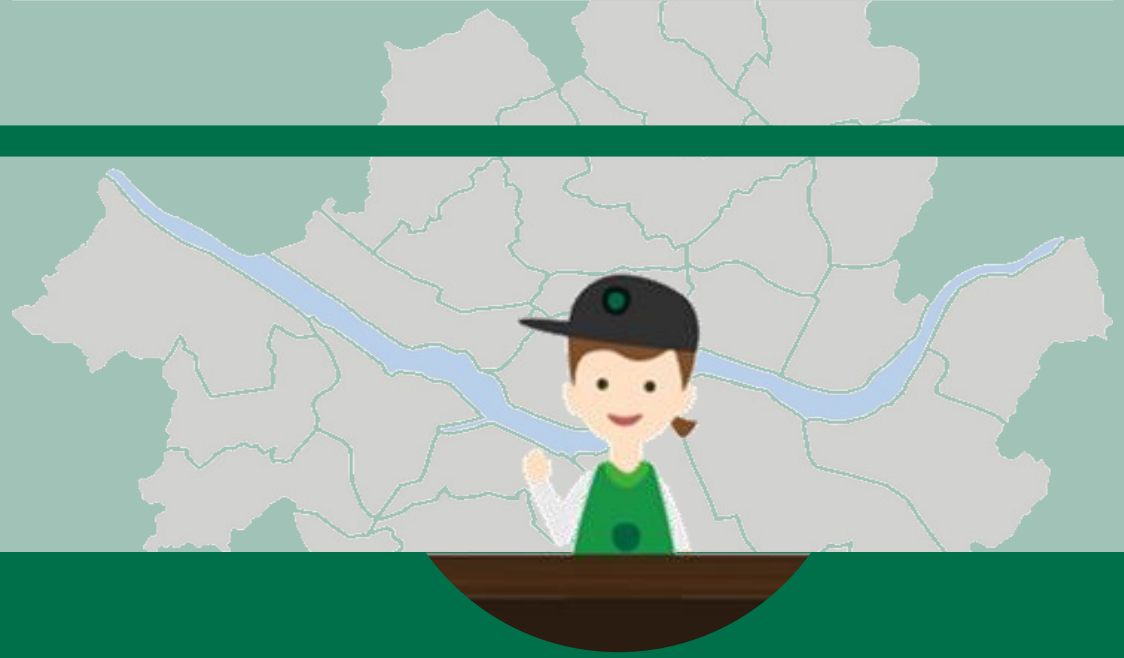
2. 데이터 소개

3. 패턴분석

4. 회귀모형

5. 포아송 회귀 모형

6. 예측 및 결론



1. 주제선정

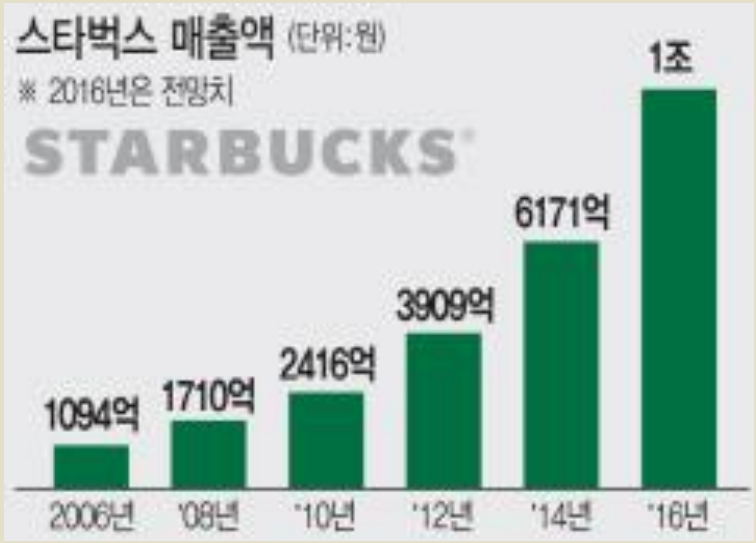
주제선정 이유

스타벅스 관련 통계들

[2016년 1000호점 돌파]

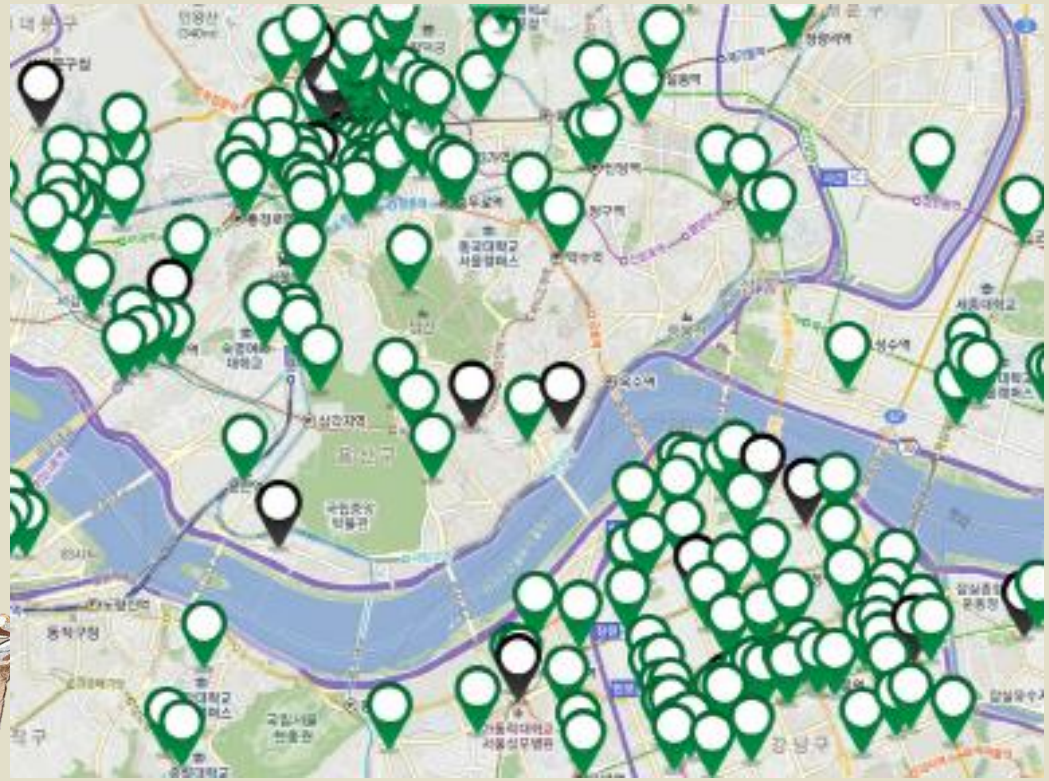
[여성 커피브랜드 선호도 1위]

[2016년 매출액 1조원 달성]



그렇다면 스타벅스는 어디에 분포하고 있나?

주제선정 이유



타 카페 브랜드와는 다른 공격적인 입점 전략

동시에 뛰어난 입지 안목과 '스타벅스 효과'가 주목받고 있다!

왜 스타벅스인가?

1. 정확한 상권분석과 뛰어난 집객 효과

스타벅스에 가면 올리브영이 보인다

스타벅스서 커피 마시고 나오면 눈앞에 올리브영 매장
목 좋은 자리에 입점 겹쳐..2030 타겟 '고수의 안목'

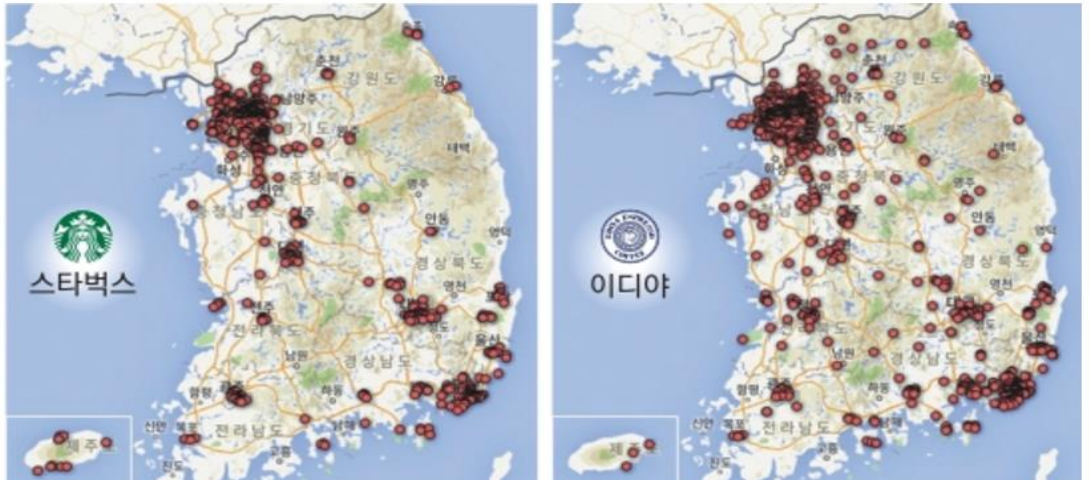
기사입력 : 2017년02월24일 10:32 최종수정 : 2017년02월24일 10:36

<일러스트=홍종현 미술기자> NEWSPIM

이디야의 필승 전략..'스타벅스 옆을 사수하라'

입력시간 | 2015.05.04 03:00 | 함정선 기자 mint@edaily.co.kr

이디야 커피는 모방 입점 전략으로 2015년 기준
매출액 1000억원 돌파, 영업이익이 66% 증가

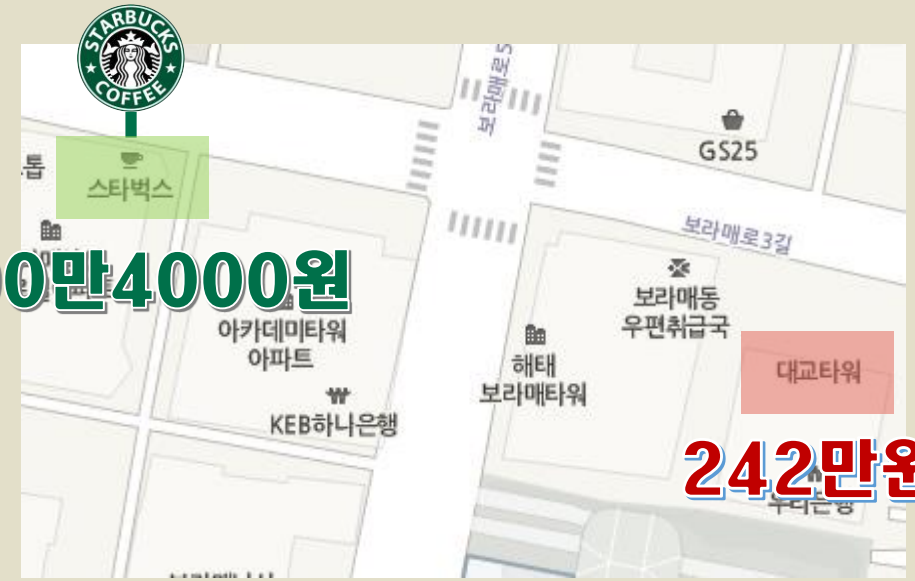


상당히 유사하게 분포한 다는 것을 알 수 있다!

왜 스타벅스인가?

2. 해당 상권 부동산에 막대한 영향

스타벅스 입점 유무에 따라 크게 달라지는 상가의 시가 (m^2당)



서울 동작구 신대방동

부산시 중구 대연동

주제선정

스타벅스의 **입점 전략**을 통계적으로 분석해보고,
다음 매장의 입점 위치를 예측해보자!

주제선정

누구에게 도움이 될까?

타 업체 : 입점 전략 수립 시 참고 지표 제공

TOM N TOM COFFEE

The Coffee Bean & Tea Leaf Est. 1963

Angel in-us Coffee 엔제리너스커피

HOLLYS COFFEE

부동산 투자자 : 부동산 투자 전략에 도움

이번 주차에서는...

행정동 별 특성을 나타낼 수 있는 다양한 변수들을 수집

☞ 중요 변수들을 선별하여 해당 변수들을 이용한 예측 모형을 세우고 그를 통해 **스타벅스 신규 지점 입점 후보군을 결정**!

2. 데이터 소개

데이터 소개

- 기본 정보
- 상주 인구
- 집객 시설
- 지수 지표
- 주변 상권 변수
- 유동/직장 인구

데이터 소개

<기본정보>

dong	code	starbucks	area	lat	long
행정동 명	행정동코드	스타벅스 수	면적(km^2)	위도	경도

<상주인구>

M live10~60	F live10~60	Apt.num	Apt.area	Apt.price
남성 상주인구 10~60대 이상	여성 상주인구 10~60대 이상	아파트 단지 수	아파트 평균 면적	아파트 평균 시가

데이터 소개

<집객시설>

Govern	bank	High	uni	Depart	Market	theater	Subway	Bus
관공서 수	은행 수	고등학교 수	대학교 수	백화점 수	마켓 수	극장 수	지하철 역 수	버스정류장 수

데이터 소개

<지수지표>

Close.rate	Start.index	Overpop	Active	Growth	stable
폐업신고율	신규창업위험도	과밀지수	활성도	성장성	안정성

신규창업위험도	행정구역 내 43개 생활밀착형 업종 기준 신규 창업 시 위험도를 폐업률과 3년 생존율로 결합(0~100으로 환산) 하여 만든 지표 (41/51/60 기준 4가지로 나뉨)
과밀지수	각 행정동의 인구밀도 정도
활성도	해당 행정동의 상권 활성화 정도
성장성	해당 행정동의 상권 성장 가능 정도
안정성	해당 행정동의 상권이 안정된 정도

데이터 소개

<주변 상권 변수>

Cloth.sales	Cloth.month	Cloth.num	Cos.sales	Cos.month	Cos.num
의류점 총 매출액	의류점 평균 영업개월 수	의류점 점포수	화장품점 총 매출액	화장품점 평균 영업개월 수	화장품점 점포 수

<주변 상권 변수>

Drink.sales	Drink.month	Drink.num
카페음료 총 매출액	카페음료 평균 영업개월 수	카페음료 점포 수

<유동/직장인구>

M/F move	Move10~60	M/F job
남/여성 유동인구	10대~ 60대이상 유동인구	남/여성 직장인구

3. 패턴분석

Introduction

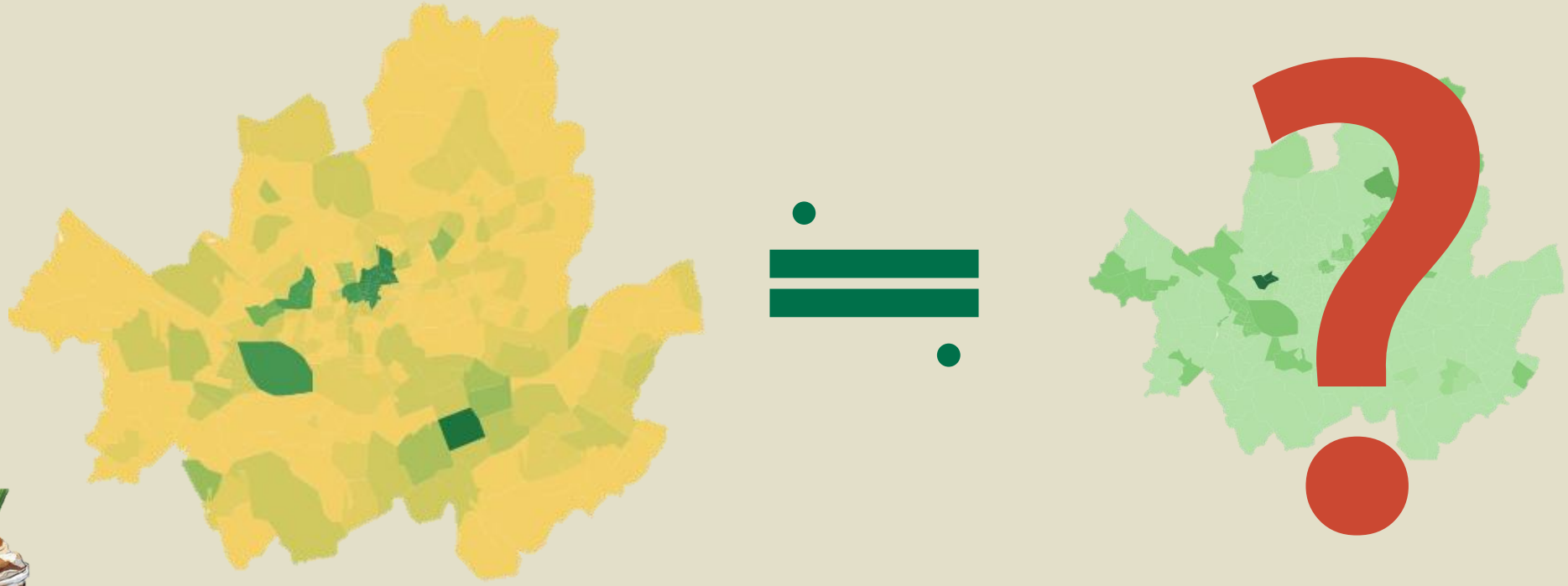
패턴분석?

- 통계적으로 정확한 값을 얻을 수 있는 분석은 아님

  하지만 비슷한 분포 형태를 보이는 변수
  ☞ 스타벅스 입점에 영향을 미치는 변수로 예상 가능!
  ☞ 이후 통계적인 요인분석을 통한 결과와 비교해보자

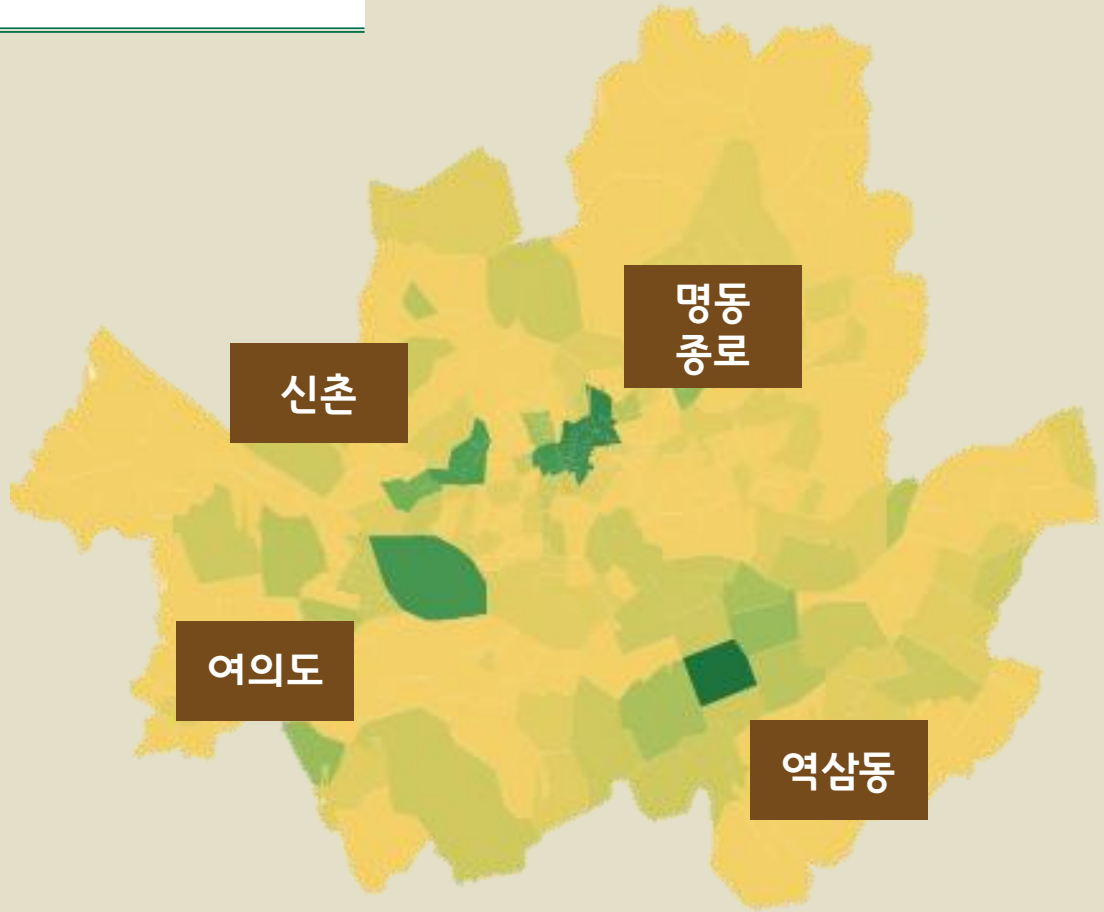
- **변수의 시각화**를 통해 변수에 대한 이해도를 높일 수 있다!

Introduction



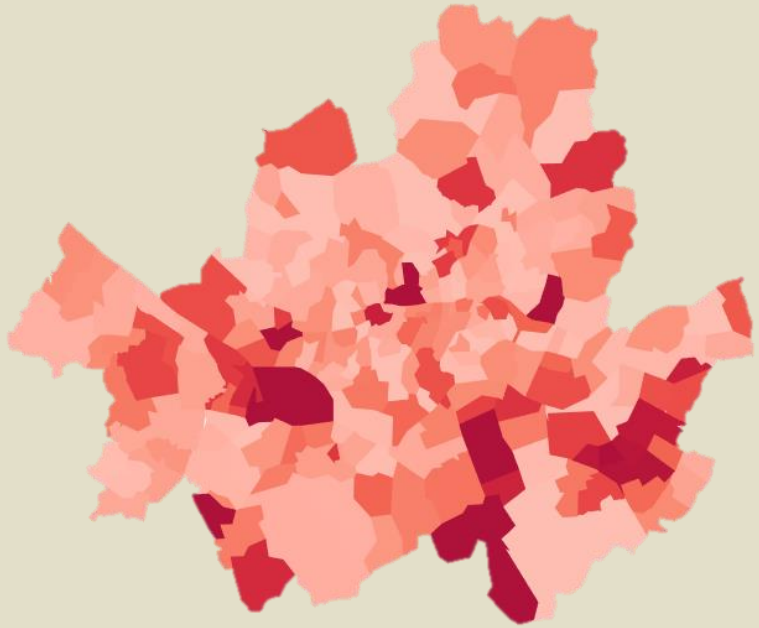
<스타벅스 밀집 지도>

Introduction

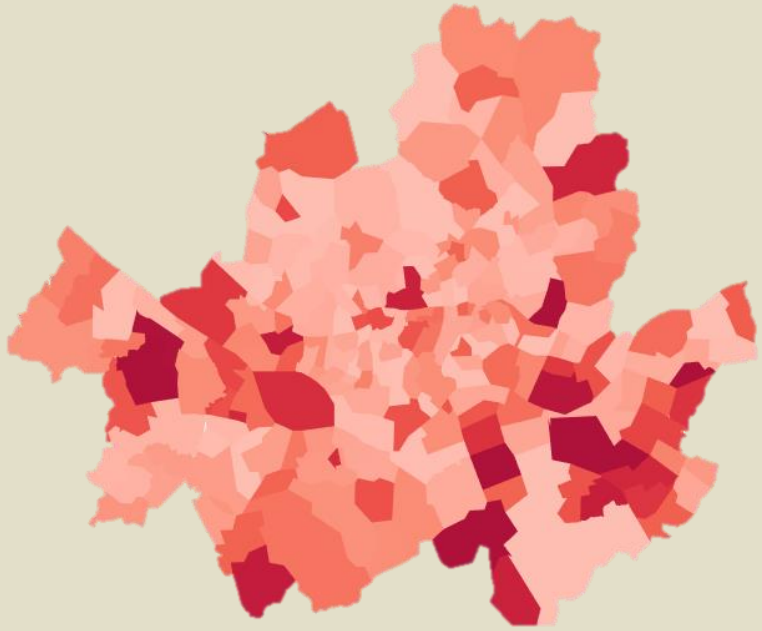


- 상권
- 인구
- 아파트
- 집객 시설
- 주변 상권

인구

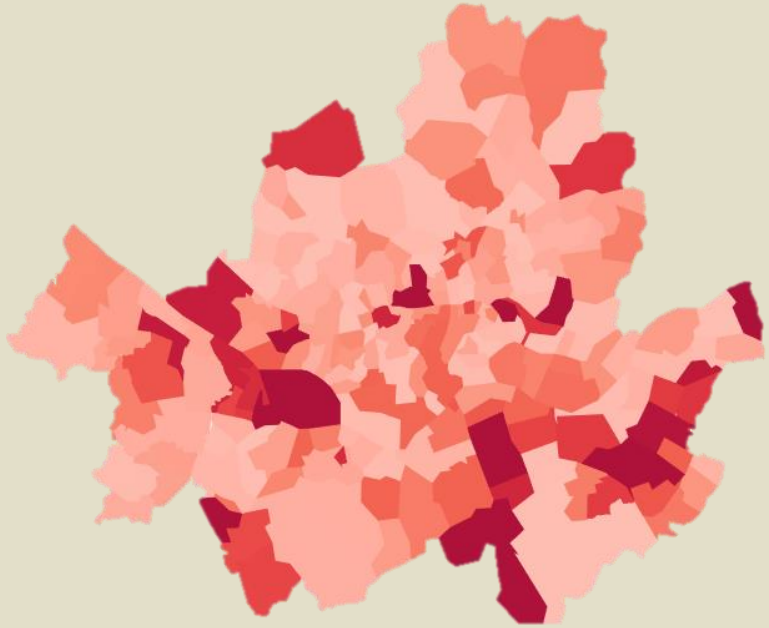


<여성 직장 인구>

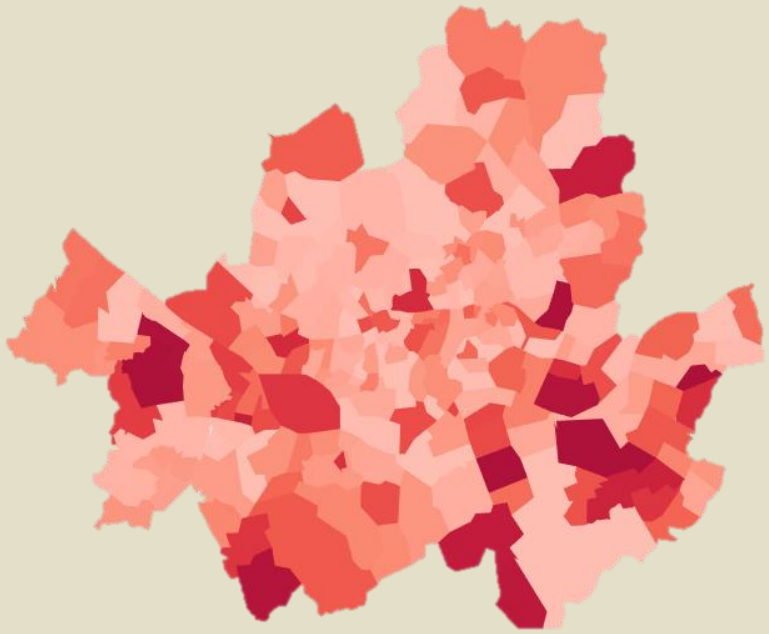


<30대 여성 거주 인구>

인구

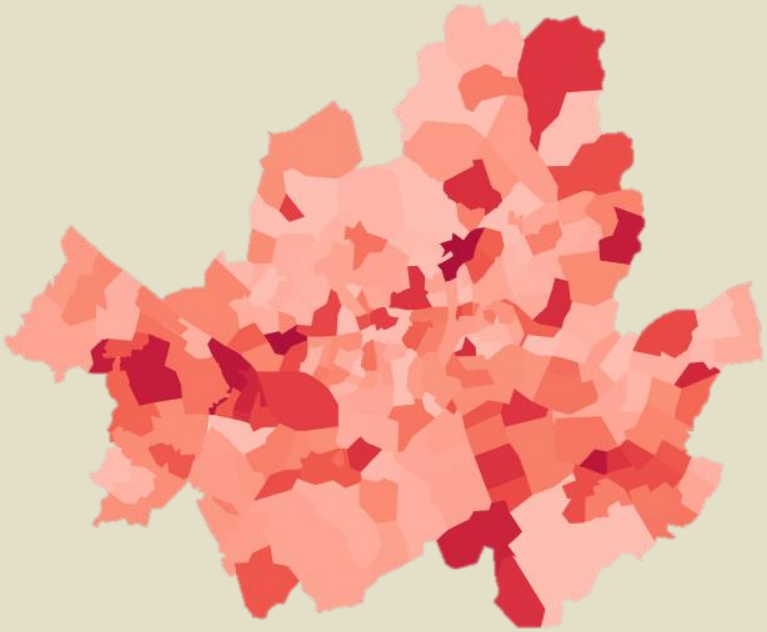


<남성 직장 인구>

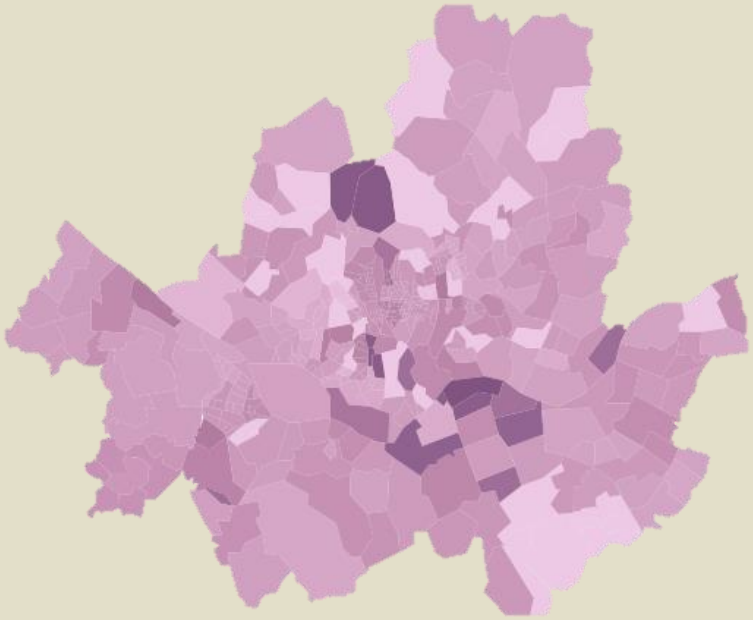


<30대 남성 거주 인구>

인구 및 거주시설

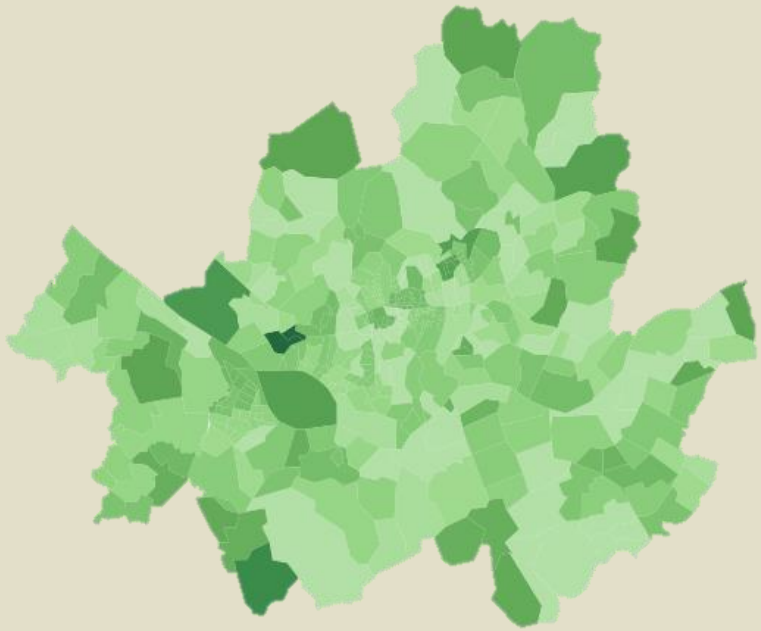


<20대 유동 인구>

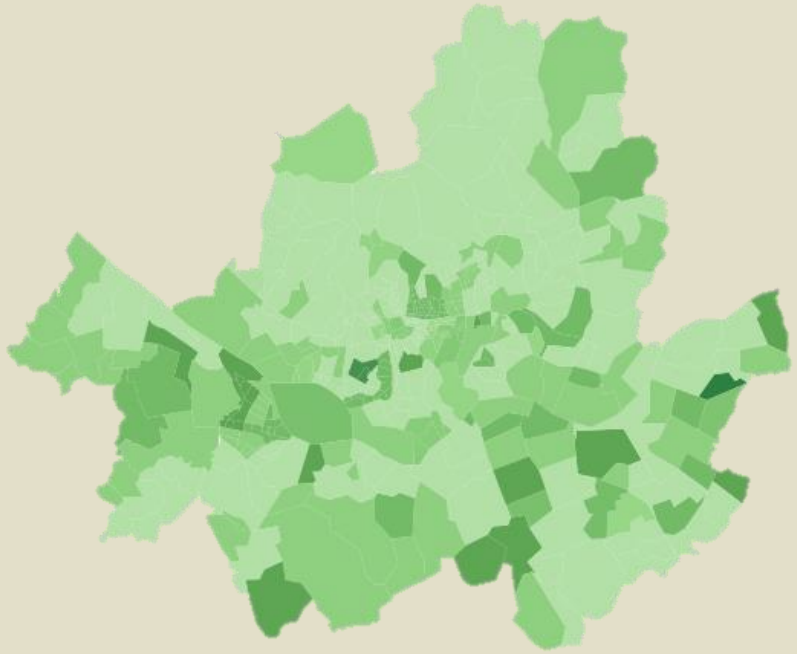


<아파트 면적>
: 스타벅스 분포와 반대의 형태
☞ 스타벅스 입지와 음의 상관관계

집객 시설

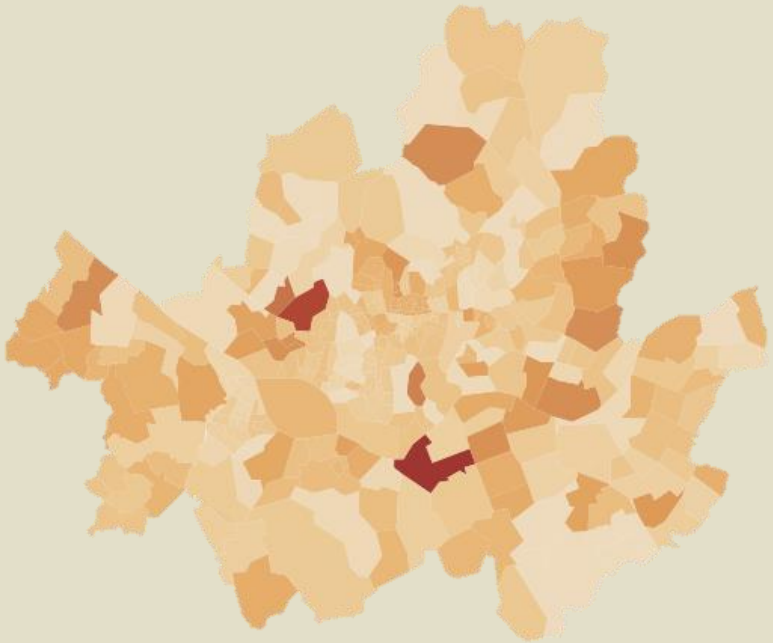


<버스 정류장 수>

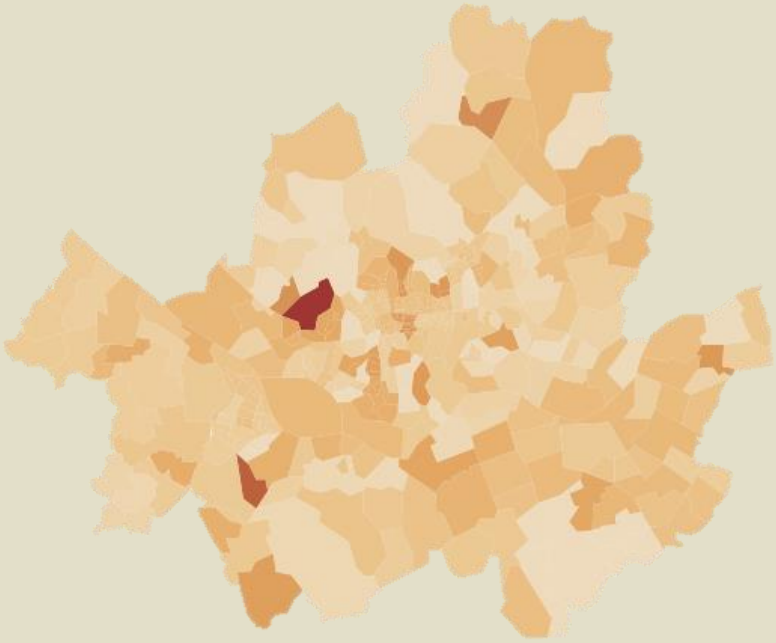


<관공서 수>

주변 상권



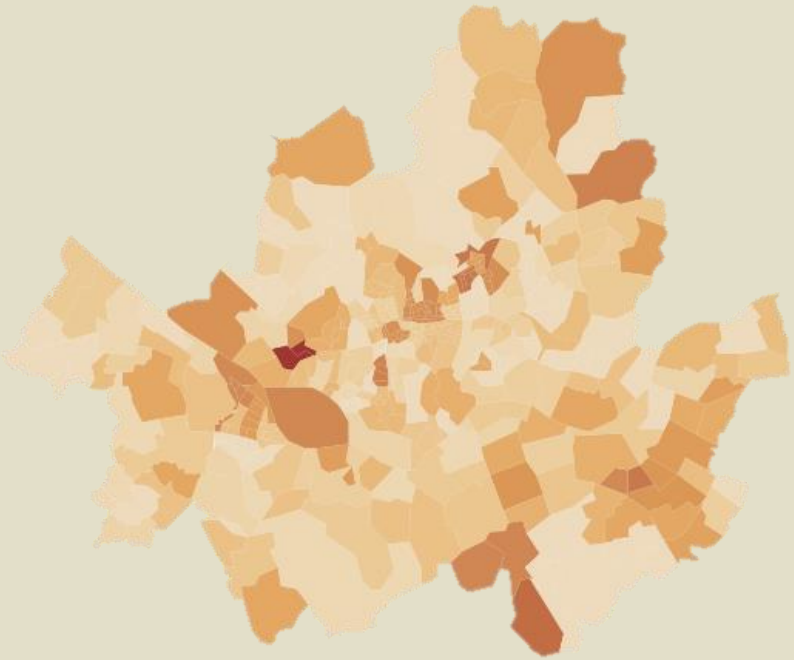
<평균 의류점 영업 개월 수 >



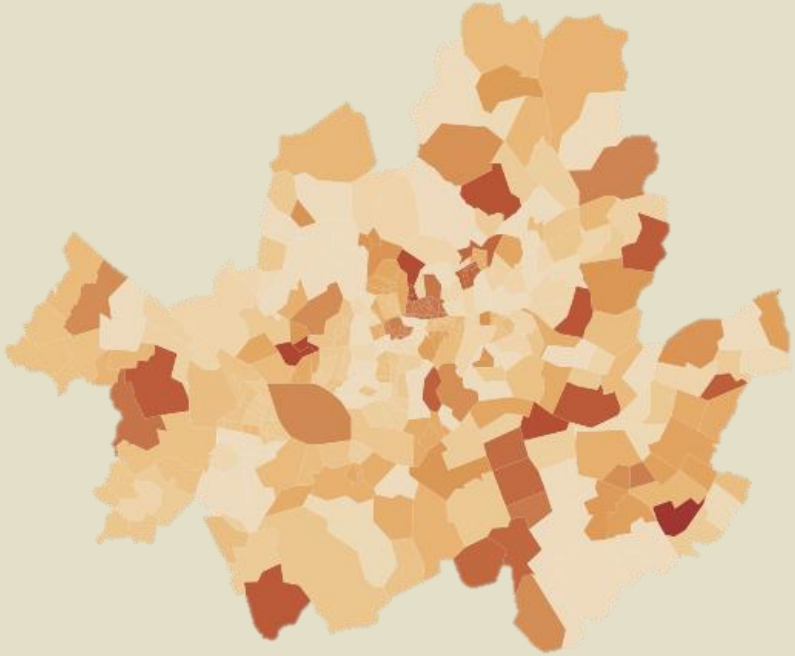
<평균 카페 영업 개월 수 >

: 스타벅스 분포와 반대의 형태 ☞ 스타벅스 입지와 **음의 상관관계!**

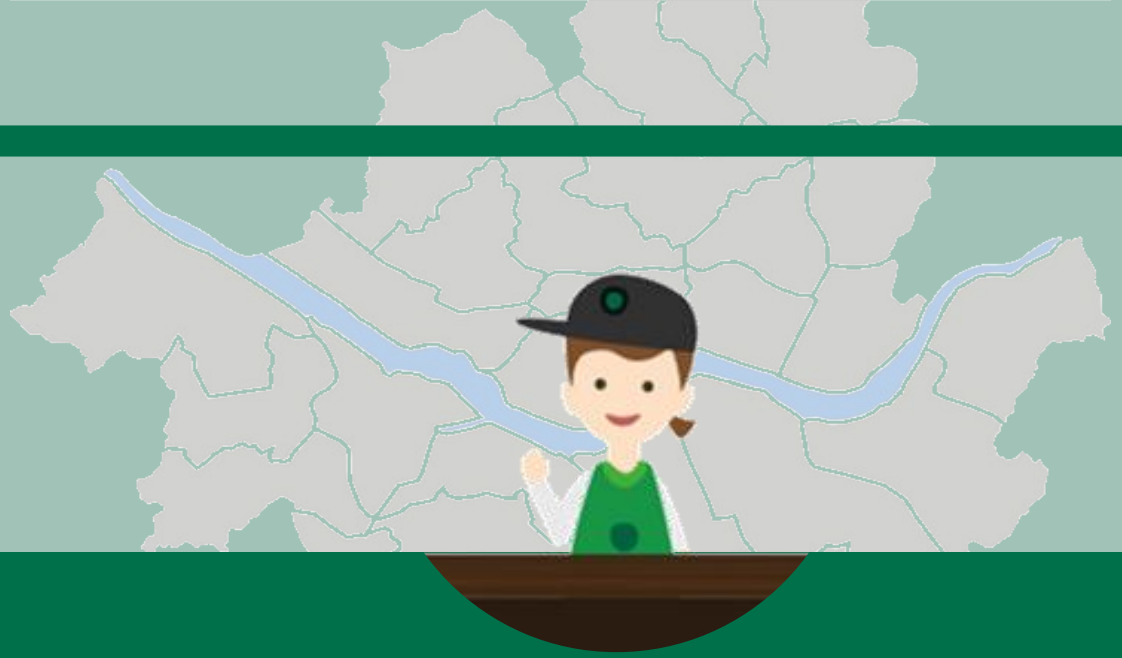
주변 상권



<카페 점포 수 >



<의류점 점포 수>



4. 회귀모형

선형회귀모형

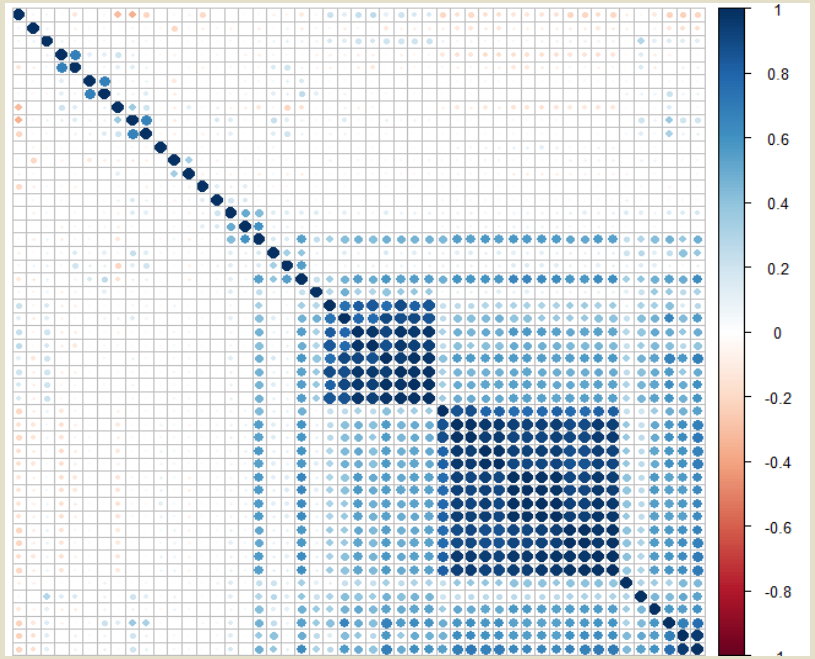
수많은 모형 중, 왜 하필 선형회귀부터?

- ✓ 가장 기본적이고 해석이 쉬운 모형
- ✓ 중요한 변수의 효과 확인 및 변수 선택할 때 좋음
- ✓ 모든 모형은 선형회귀에서 확장! : GLM

☞ 간단한 회귀 모형을 활용, 중요 변수들을 골라보자!

선형회귀모형

예측 변수 간 다중 공선성 확인




```
> sum(vif(data[,-1])>10)
[1] 22
```

예측 변수 간 상관관계 plot 확인

☞ 다중 공선성이 존재하는 것으로 보임

VIF가 10을 넘는 변수 다수 존재

☞ 변수 선택이 필요!

선형회귀모형

다중 공선성 문제 해결을 위해 변수 선택

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.035e+00  8.317e-01   1.244  0.21438
fjob         4.055e-04  9.871e-05   4.107 5.00e-05 ***
flive30      4.717e-03  2.056e-03   2.294  0.02241 *
flive10     -3.399e-03  8.312e-04  -4.089 5.39e-05 ***
mlive30     -3.414e-03  1.761e-03  -1.939  0.05336 .
apt.area     1.392e-02  5.789e-03   2.405  0.01669 *
move20       9.467e-05  2.394e-05   3.955 9.30e-05 ***
bank        -1.438e-01  6.919e-02  -2.078  0.03843 *
start.index -2.341e-02  1.464e-02  -1.599  0.11078
move30       1.905e-04  5.763e-05   3.306  0.00104 **
mmove       -1.100e-04  2.472e-05  -4.450 1.16e-05 ***
drink.month  2.455e-03  1.197e-03   2.050  0.04107 *
cloth.month -1.107e-03  3.765e-04  -2.941  0.00349 **
cloth.num    2.567e-02  1.124e-02   2.283  0.02301 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.494 on 346 degrees of freedom
Multiple R-squared:  0.396,     Adjusted R-squared:  0.3733
F-statistic: 17.45 on 13 and 346 DF,  p-value: < 2.2e-16
```

✓ **Stepwise-method**
- AIC를 기준으로 변수 선택

✓ **Wald Test & R-Squared**
- 해당 모형이 적절해 보임
- 가정들도 만족할까?

선형회귀모형

선형회귀모형 가정들을 확인해보자

```
> shapiro.test(resid)

        Shapiro-Wilk normality test

data:  resid
W = 0.77404, p-value < 2.2e-16

> bptest(fit)

        studentized Breusch-Pagan test

data:  fit
BP = 71.362, df = 13, p-value = 4.502e-10

> dwtest(fit)

        Durbin-Watson test

data:  fit
DW = 1.9586, p-value = 0.3048
alternative hypothesis: true autocorrelation is greater than 0
```

✓ **정규성 검정**

- H_0 : 정규성을 만족한다.

→ 기각: 정규성 만족 X

✓ **등분산성 검정**

- H_0 : 등분산성을 만족한다.

→ 기각: 등분산성 만족 X

✓ **독립성 검정**

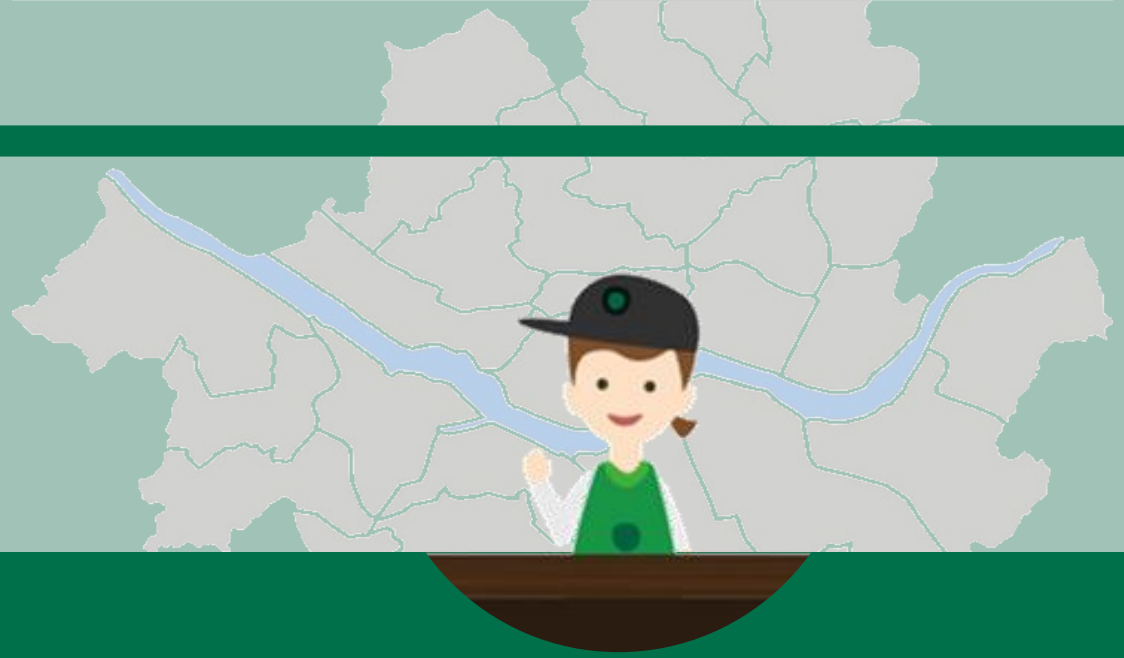
- H_0 : 독립성을 만족한다

→ 기각 X: 독립성 만족

선형회귀모형

가정을 만족하지 않아, 예측에는 사용할 수 있지만
중요 변수들을 추론, 선택하기에는 적절하지 않다

가정이 완화된 다른 모형을 적합 해보자!



5. 포아송 회귀 모형

복습

포아송 회귀 모형이 뭐였지..?

랜덤 성분의 기대값에 대해 포아송 분포를 가정→ 로그 연결함수 사용!

$$\log \mu = \alpha + \beta x$$

$$\mu = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x}$$

Y변수가 도수자료일 경우 사용하는 GLM!

복습

포아송 회귀 모형 예시

예시) 시간의 흐름과 가스 사고 발생 건수의 연관성 비교

연도(X)	가스사고 발생세대 (Y)		총 세대수
	YES	NO	
1988	12	324	336
1989	13	334	347
1990	14	371	385
…	…	…	…
2014	13	597	610
2015	13	603	616
2016	14	604	618

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.564743   0.099661   25.73   <2e-16 ***
Year        0.002240   0.006061    0.37    0.712
---
```

시간의 흐름과 상관없이
12~14 사이로 별 변화가 없어 보인다!

복습

포아송 회귀 모형 예시

예시) 시간의 흐름과 가스 사고 발생률의 연관성 비교

연도(X)	가스사고 발생세대 (Y)		총 세대수
	YES	NO	
1988	12	324	336
1989	13	334	347
1990	14	371	385
…	…	…	…
2014	13	597	610
2015	13	603	616
2016	14	604	618

그러나,
시간의 흐름에 따라 총 세대수는
2배 가까이
크게 변한 것을 알 수 있다!

따라서,

-> $\frac{\text{발생 건수}}{\text{총 세대 수}}$ = 발생률

로 분석하고 싶다!

율자료 포아송회귀

율 자료(Rate Data)에 대한 포아송 회귀 모형

$$\log\frac{\mu}{t} = \log\mu - \log t = \alpha + \beta x$$

cf) $-\log t$: $\log\mu$에 대한 수정항(offset)

어떤 사건이 다른 크기의 지표(index)에 따라 나타날 경우,
율(Rate)에 대한 모형을 세워야 한다!

율자료 포아송회귀

율 자료(Rate Data)에 대한 포아송 회귀 모형

$$\log\frac{\mu}{t} = \log\mu - \log t = \alpha + \beta x$$

cf) $-\log t$: $\log\mu$에 대한 수정항(offset)

$\therefore$ x의 한 단위 증가는 $\frac{\mu}{t}$ 에 대해 e^{β} 배 만큼의 영향을 미침

i) $\beta = 0$: $\frac{\mu}{t}$가 x의 변화에 영향을 받지 않음

ii) $\beta > 0$: x가 증가함에 따라 $\frac{\mu}{t}$도 증가

iii) $\beta < 0$: x가 증가함에 따라 $\frac{\mu}{t}$는 감소

복습

포아송 회귀 모형 예시

예시) 시간의 흐름과 가스 사고 발생율 의 연관성 비교

연도(X)	가스사고 발생세대 (Y)		총 세대수
	YES	NO	
1988	12	324	336
1989	13	334	347
1990	14	371	385
…	…	…	…
2014	13	597	610
2015	13	603	616
2016	14	604	618

```
Call:
glm(formula = Y ~ Year + offset(log(n)) family = poisson(link = "log"),
    data = data2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.68986  -0.21376   0.04114   0.30218   0.70018

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.408443   0.100777 -33.822   <2e-16 ***
Year        -0.015376   0.006152  -2.499   0.0124 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 9.9570  on 28  degrees of freedom
Residual deviance: 3.7207  on 27  degrees of freedom
AIC: 136.55

Number of Fisher Scoring iterations: 4
```

Year 변수가 이번엔 유의해진다!!

포아송 회귀모형

행정동 면적을 offset으로 삼아 포아송 회귀모형

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.512e+00  2.244e-01 -15.652  < 2e-16 ***
move20       8.790e-05  9.499e-06   9.253  < 2e-16 ***
flive40      4.571e-02  3.986e-03  11.469  < 2e-16 ***
bus         -1.708e-01  2.877e-02  -5.936 2.93e-09 ***
cloth.month -2.531e-03  4.260e-04  -5.942 2.81e-09 ***
market      -1.174e+00  1.103e-01 -10.644  < 2e-16 ***
apt.price    4.228e-09  6.065e-10   6.971 3.14e-12 ***
cloth.sales -3.944e-09  8.459e-10  -4.663 3.12e-06 ***
fjob         1.134e-03  2.691e-04   4.213 2.52e-05 ***
govern      -5.759e-01  9.207e-02  -6.255 3.97e-10 ***
mjob        -3.884e-04  1.819e-04  -2.135 0.032721 *
subway      -9.442e-01  1.336e-01  -7.066 1.59e-12 ***
mlive10     -2.406e-02  1.468e-03 -16.383  < 2e-16 ***
mlive40     -2.911e-02  3.389e-03  -8.591  < 2e-16 ***
mlive30      1.170e-02  2.929e-03   3.995 6.48e-05 ***
flive30     -4.266e-03  2.618e-03  -1.630 0.103204
drink.month  6.744e-03  9.674e-04   6.971 3.14e-12 ***
mlive60     -6.001e-03  1.682e-03  -3.567 0.000361 ***
mlive50      6.228e-03  2.846e-03   2.188 0.028669 *
apt.num     -1.462e-02  2.765e-03  -5.287 1.24e-07 ***
cos.num      5.388e-02  1.064e-02   5.062 4.15e-07 ***
high         3.963e+00  9.319e-01   4.253 2.11e-05 ***
drink.num   -7.748e-02  2.205e-02  -3.514 0.000442 ***
uni          1.683e+00  6.174e-01   2.726 0.006420 **
flive20     -5.150e-03  1.899e-03  -2.712 0.006687 **
mlive20      5.651e-03  2.546e-03   2.220 0.026434 *
theater     -1.707e-01  8.039e-02  -2.123 0.033730 *
```

- ✓ **Stepwise-method**
  - AIC를 기준으로 변수 선택
- ✓ **Wald Test**
  - $H_0 : \beta_j = 0$ (선택된 변수들 유의함)
- ✓ **변수를 더 추려낼 수는 없을까?**

포아송 회귀

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.491e+00  2.159e-01 -16.171  < 2e-16 ***
move20       8.124e-05  7.235e-06  11.228  < 2e-16 ***
mlive10     -2.432e-02  1.200e-03 -20.276  < 2e-16 ***
flive40      4.882e-02  3.515e-03  13.889  < 2e-16 ***
bus         -1.660e-01  2.854e-02  -5.817 5.99e-09 ***
mlive40     -3.109e-02  3.141e-03  -9.896  < 2e-16 ***
subway      -1.022e+00  1.355e-01  -7.545 4.53e-14 ***
market      -1.196e+00  1.101e-01 -10.857  < 2e-16 ***
fjob         5.245e-04  8.231e-05   6.372 1.87e-10 ***
cloth.sales -3.367e-09  8.014e-10  -4.201 2.65e-05 ***
mlive30      1.246e-02  2.590e-03   4.812 1.50e-06 ***
cos.num      6.470e-02  9.182e-03   7.047 1.83e-12 ***
flive30     -6.021e-03  2.356e-03  -2.556  0.01060 *
govern      -5.747e-01  9.132e-02  -6.293 3.12e-10 ***
apt.num     -1.546e-02  2.622e-03  -5.897 3.70e-09 ***
drink.month  6.591e-03  9.955e-04   6.621 3.57e-11 ***
apt.price    4.153e-09  5.896e-10   7.043 1.88e-12 ***
cloth.month -2.308e-03  4.059e-04  -5.686 1.30e-08 ***
mlive60     -5.731e-03  1.432e-03  -4.003 6.25e-05 ***
drink.num   -5.820e-02  1.978e-02  -2.943  0.00325 **
high         1.975e+00  4.750e-01   4.157 3.23e-05 ***
mlive50      6.583e-03  2.634e-03   2.499  0.01244 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- ✓ AIC기준으로 만든 모델을 FULL model로 생각!
- ✓ Stepwise-method
  - BIC를 기준으로 다시 변수선택
- ✓ 모든 변수가 유의하다!

포아송 회귀

```
> dwtest(bic.fit)

        Durbin-Watson test

data:  bic.fit
DW = 2.0308, p-value = 0.5676
alternative hypothesis: true autocorrelation is greater than 0
```

```
> dispersiontest(bic.fit,alternative="two.sided")

        Dispersion test

data:  bic.fit
z = 1.5476, p-value = 0.1217
alternative hypothesis: true dispersion is not equal to 1
sample estimates:
dispersion
  23.99911
```

✓ 독립성 검정
- 0.5675 > 0.05

- 독립성 가정을 만족한다!

✓ 과산포 검정
- 0.1217 > 0.05

- 평균과 분산이 같다!
(포아송 분포 사용 가능)

포아송 회귀

선택된 변수들 :

20대 유동인구 수

60대 남성 거주인구 수
50대 남성 거주인구 수
40대 남성 거주인구 수
40대 여성 거주인구 수
30대 남성 거주인구 수
30대 여성 거주인구 수
10대 남성 거주인구 수

버스 정류장 수
지하철 역 수
고등학교 수
관공서 수
아파트 가격
아파트 수
시장 수

여성 직장 인구 수
의류점포 매출액
의류점포 수
화장품 점포 수

카페 평균 영업 개월 수
카페 점포 수

수집한 변수 49개 ☞ 21개로 축약

포아송 회귀

패턴 분석 결과와의 비교

<패턴분석으로 예상한 중요 변수>

- 여성/남성 직장 인구
- 30대 여성/남성 거주 인구
- 20대 유동 인구
- 아파트 면적
- 버스 정류장 수
- 관공서 수
- 의류점 / 카페 영업 개월 수
- 의류점 / 카페 점포 수

<회귀 모형으로 최종 선택한 변수>

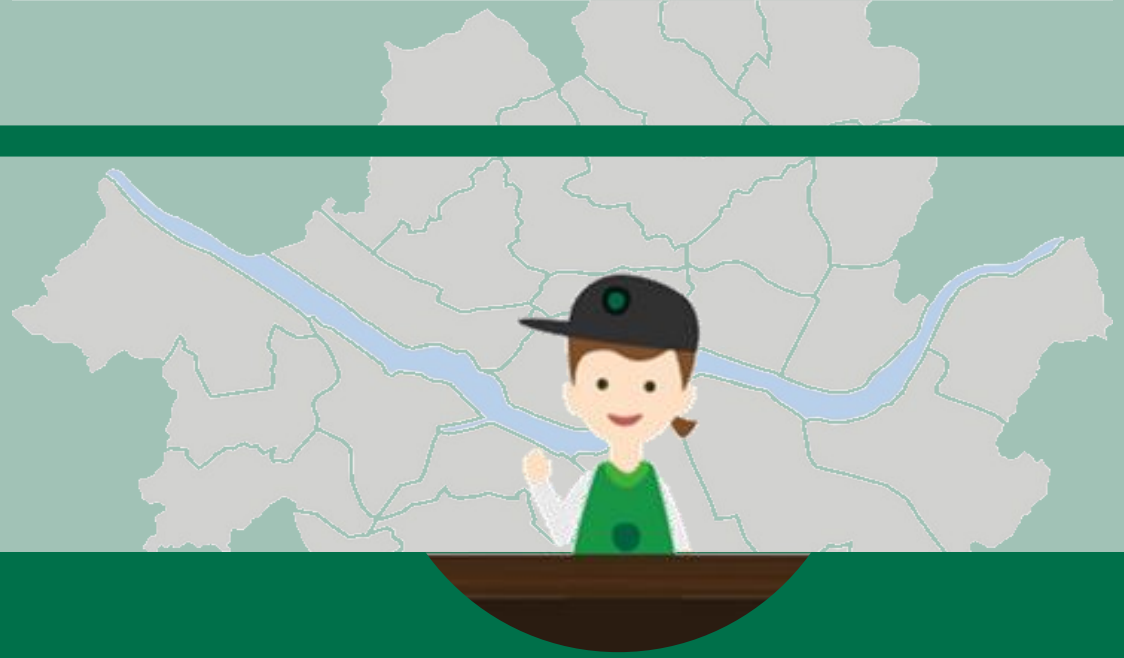
- 여성 직장 인구
- 연령별 여성/남성 거주 인구

…

- 버스 정류장 수
- 지하철 역 수
- 관공서 수

…

- 카페 영업 개월 수
- 의류점 / 카페 / 화장품점 점포 수



6. 예측 및 결론

예측

회귀 모형에서 변수 선택을 시행,
수집한 변수들 중 **다중 공선성이 존재하지 않으면서 유의한 변수들로 축약**

↓

해당 변수들을 활용해
스타벅스가 추가적으로 입점할 가능성이
높은 행정동을 찾아보자!

예측

- 5-fold CV를 통한 포아송 회귀 예측력 확인

```
> mean(cvstat)
[1] 2.269257
```

CV error : 2.2692

☞ 이해하기 쉬우면서 보다 예측력이 좋은 모형은 없을까?

우선, 이해하기 쉬운 tree 모형으로 예측 모형을 세워보자!

Regression tree

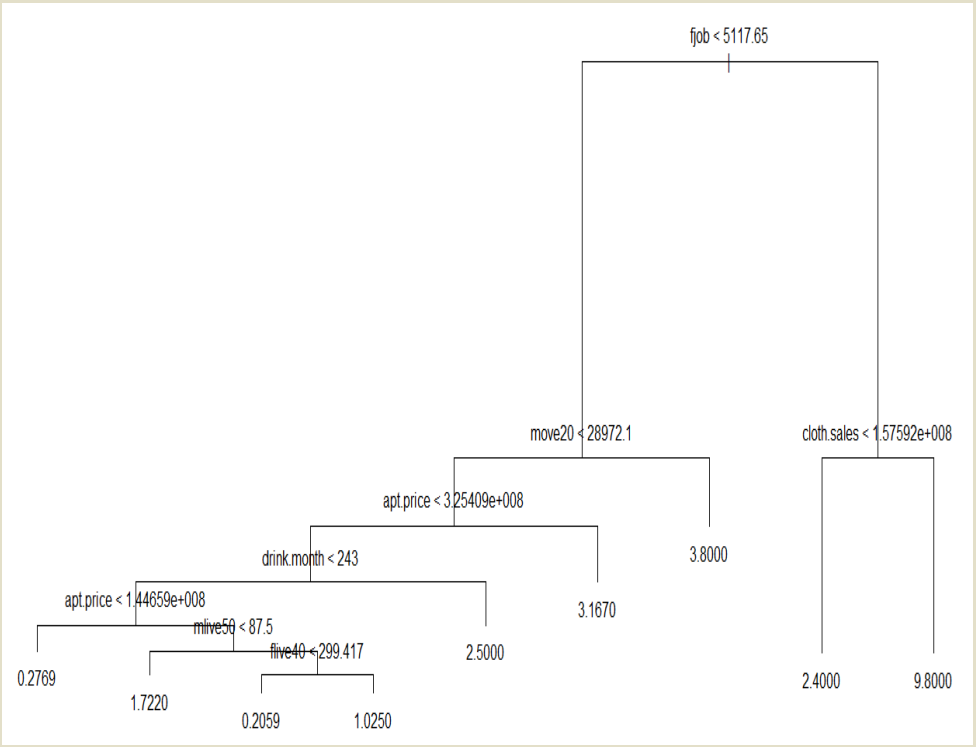
- 사용된 변수들 확인

```
> summary(tree.fit)

Regression tree:
tree(formula = starbucks ~ ., data = sel.dat, subset = train)
Variables actually used in tree construction:
[1] "fjob"       "move20"     "apt.price"
[4] "drink.month" "mlive50"    "flive40"
[7] "cloth.sales"
Number of terminal nodes:  9
Residual mean deviance:  1.983 = 479.8 / 242
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-7.8000 -0.2769 -0.2769  0.0000  0.7231  9.5000
```

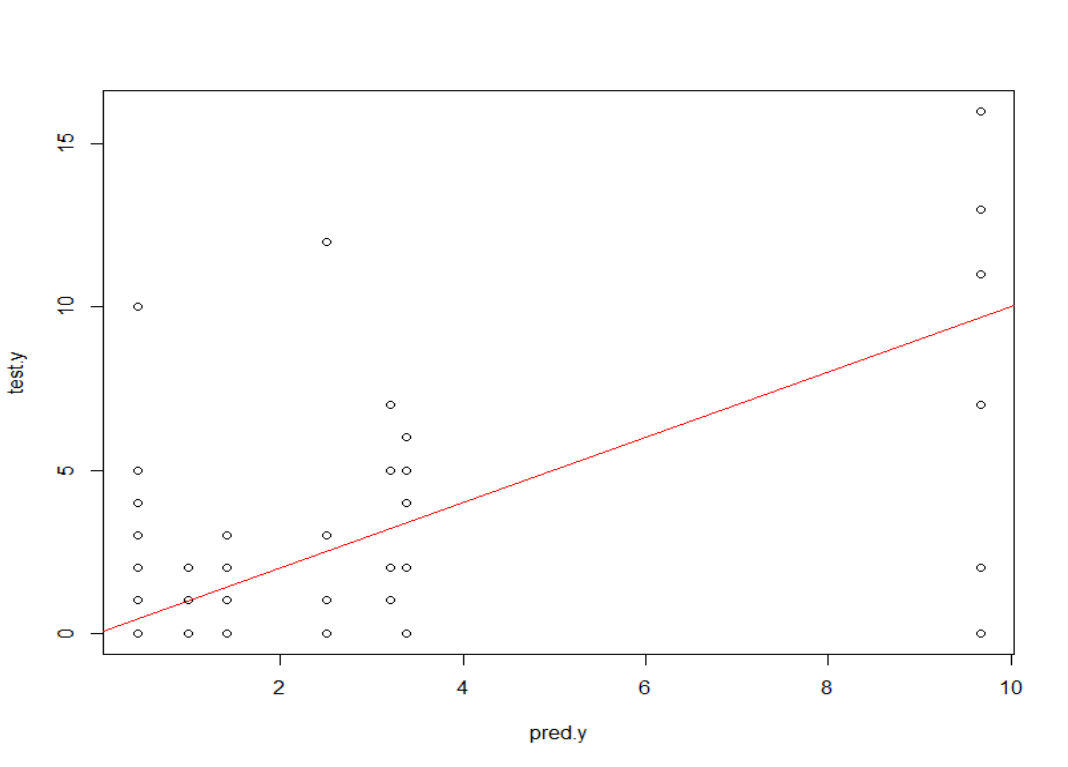
여성 직장인구 수 20대 유동인구 수
아파트 가격 카페 평균 영업개월 수
50대 남 거주인구 수 의류점 매출액
40대 여 거주인구 수
☞ 총 7개 변수 사용

- Tree plot



Regression tree

예측 값과 실제 값의 비교



비교 plot 확인 결과,
예측 오차가 클 것으로 예상

가지치기를 한 후 예측 오차를 비교해보자!

Regression tree

가지치기 : CV를 통한 tree의 가지 수 결정

- 가지 수에 따른 CV error

```
> cv.fit = cv.tree(tree.fit,K=5)
> plot(cv.fit$size,cv.fit$dev,type='b')
```

CV error가 가장 낮아지는 4개로 가지 설정

- 사용된 변수들 확인

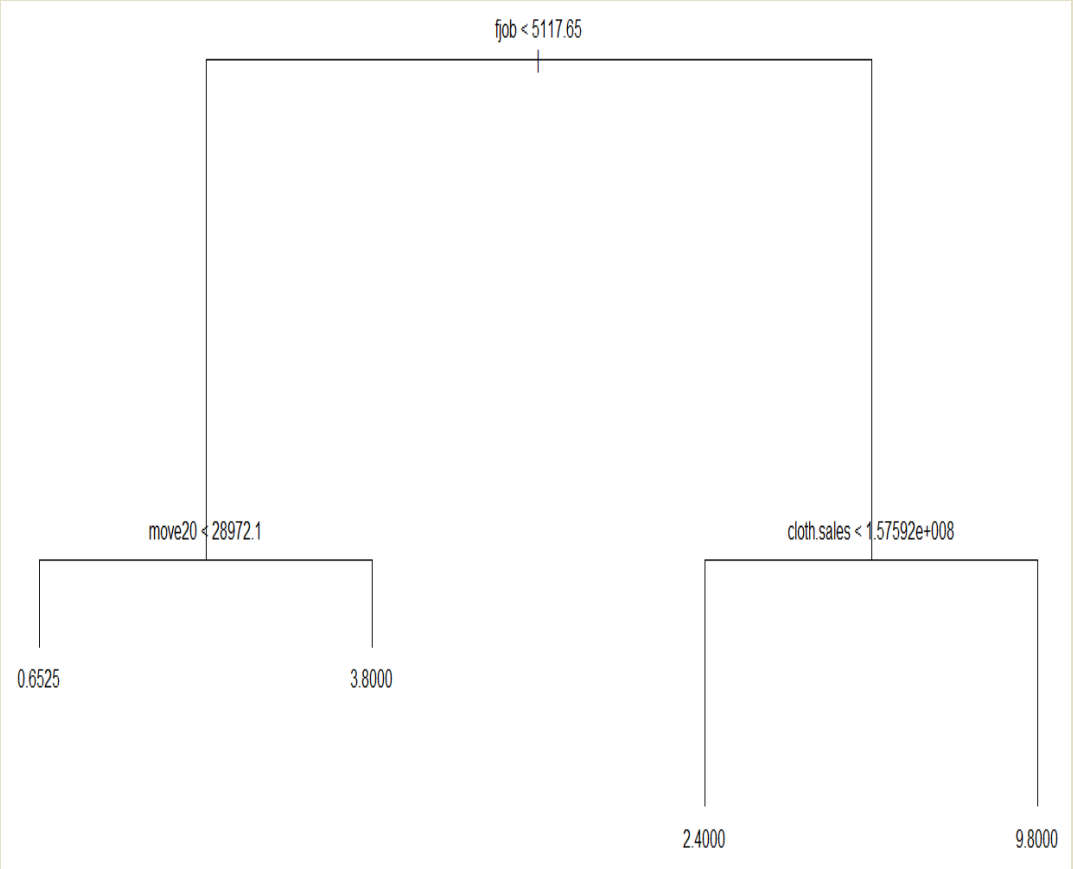
```
> summary(prune.fit)

Regression tree:
snip.tree(tree = tree.fit, nodes = 4L)
Variables actually used in tree construction:
[1] "fjob"        "move20"      "cloth.sales"
Number of terminal nodes:  4
Residual mean deviance:  2.414 = 596.3 / 247
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-7.8000 -0.6525 -0.6525  0.0000  0.3475 11.3500
```

여성 직장인구 수, 20대 유동인구 수, 의류점 매출액
☞ 총 3개의 변수 사용

Regression tree

가지치기한 tree plot 확인 및 해석



- 여성 직장인구 수가 5117명 이하, 20대 유동인구 수가 28000명 이상인 행정동에는 대략 4개의 스타벅스가 입점한다
- 여성 직장인구 수가 5118명 이상, 의류 가게 총 매출액이 1억 5천만원 이상인 행정동에는 대략 9.8개의 스타벅스가 입점한다

Regression tree

가지치기 전후 tree 모형 예측오차 비교

- 5-fold CV error

```
> mean(cvstat)
[1] 1.005798
```

```
> mean(cvstat)
[1] 0.9398217
```

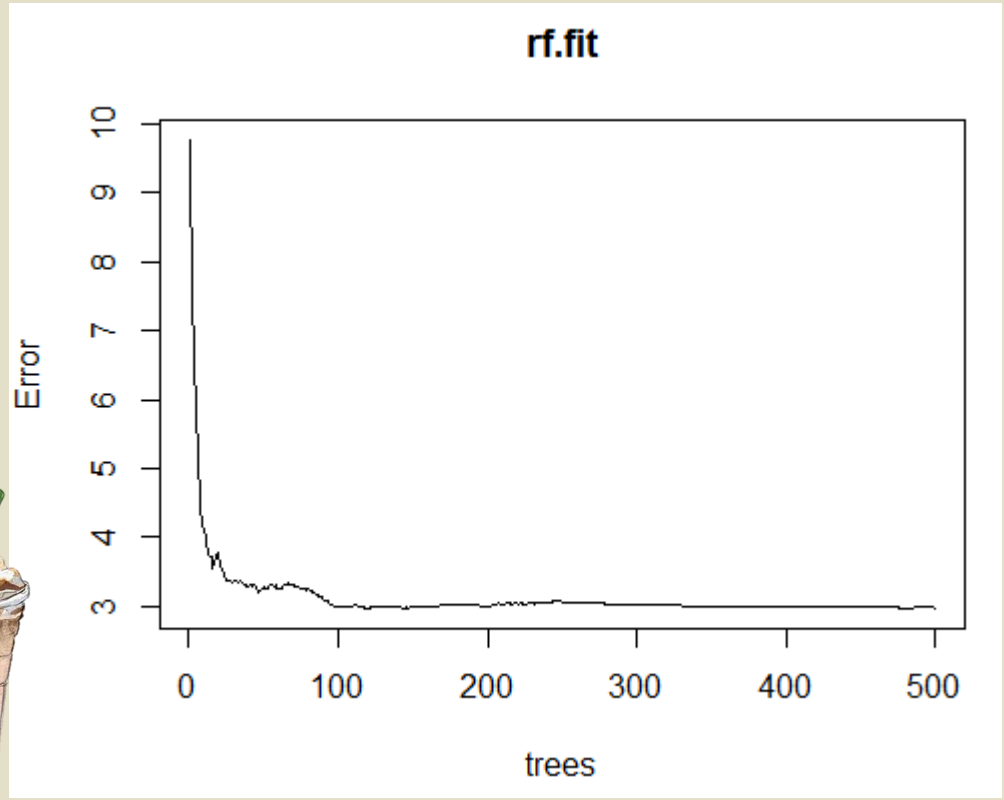
- 가지치기 전 모형의 CV error : 1.0057
- 가지치기 후 모형의 CV error : 0.9398

☞ 예측 오차가 줄어들었다

과적합 방지해 예측 오차 좀 더 줄일 수 있는 random Forest도 사용, 비교해보자!

Random Forest

```
> rf.fit = randomForest(starbucks~.,data=sel.dat,subset=train,ntree=500,mtry=7)
> plot(rf.fit)
```



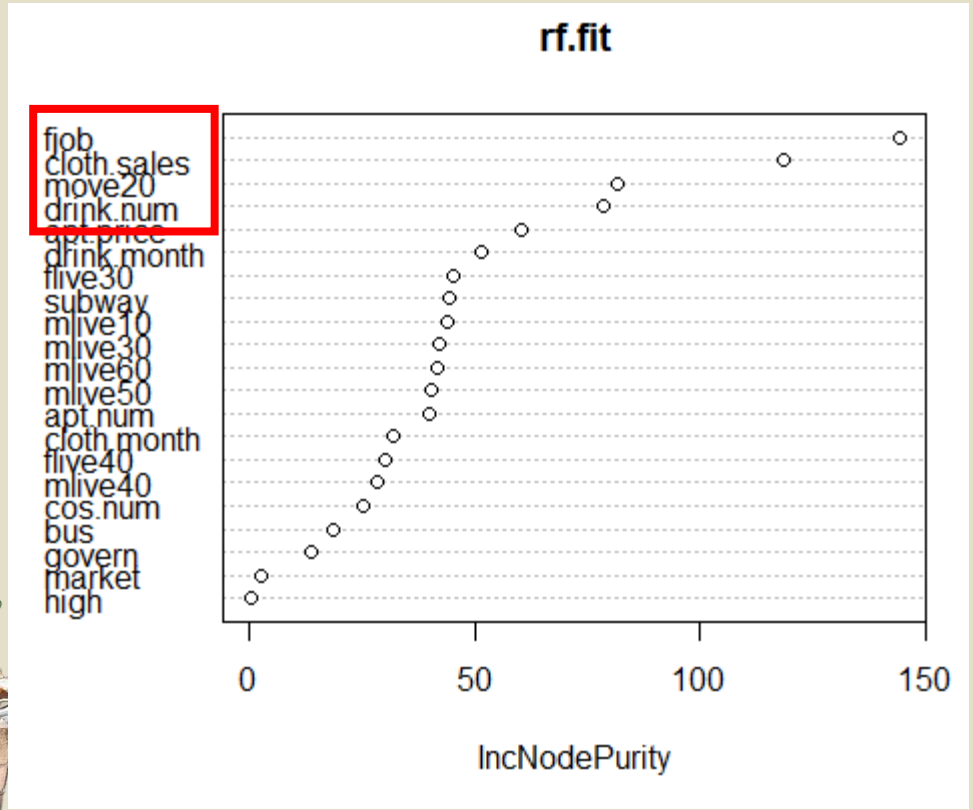
- 500개 정도의 나무 형성하면 오차의 변동 거의 없음

☞ 나무 개수 500개로 충분! (default : 1000)

- (변수 개수)/3 정도의 변수 후보군을 설정하는 것이 이상적

☞ 변수 후보군 (mtry) 7개로 설정

Random Forest



변수 중요도 확인

- Random Forest는 블랙박스 모형이므로 직접적인 변수 효과 해석은 불가능
- 다만 node purity를 기준으로 변수의 중요도 파악 가능

Random Forest에서 중요하게 사용된 변수

☞ 여성 직장인구 수, 의류 가게 매출액, 20대 유동인구 수, 카페 점포 수

Random Forest

예측오차 확인 및 최종 후보군 결정

```
> mean(cvstat)
[1] 0.8819232
```

- Random Forest의 CV error : 0.8819

☞ 앞선 regression tree 모형들보다 예측오차가 적으므로 최종 예측에 random forest 사용!

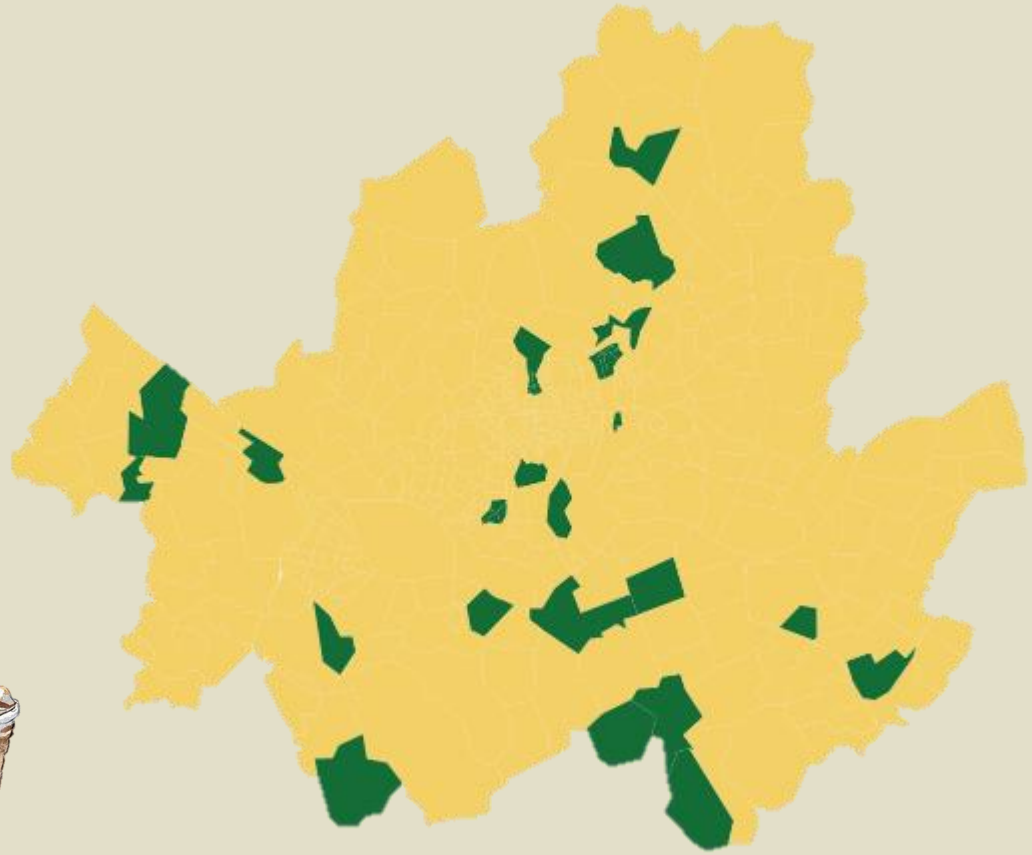
```
> sum(pred.y2 > sel.dat[,"starbucks"]+1)
[1] 31
```

예측 점포 수가 현존 점포 수보다 많은 행정동을 스타벅스 입점 후보군으로 설정

☞ 총 31개의 행정동

랜덤 포레스트

입점 후보군 행정동



- 삼청동
- 신당동
- 가양 제1동

…

- 시흥 제1동
- 잠실 7동

1주차 한계점

스타벅스가 추가로
입점 될 것 같은 행정동만 선별!

☞ 행정동만을 기준으로 하면
지나치게 포괄적

☞ 해당 행정동 어디에?

2주차 목표

선택된 행정동 안에서
스타벅스가 입점한 지역 및 상권
특성을 분석

☞ 최종적으로 어느 지역에
입점할지 예측 지점 결정!

감사합니다