서울시 골목상권 월 매출 예측하기

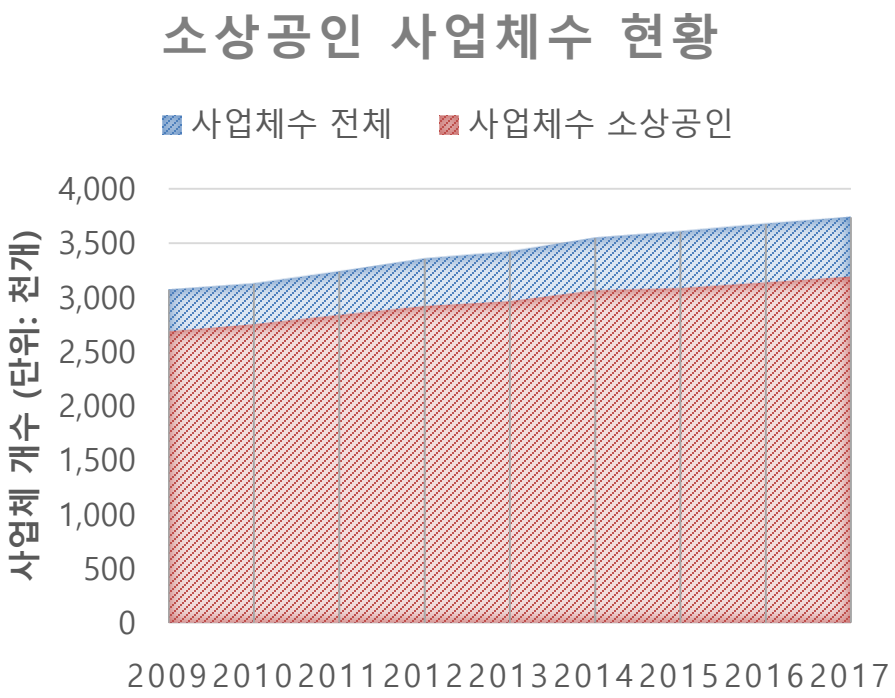
김건우, 박동재, 장상현, 장우빈, 허은정

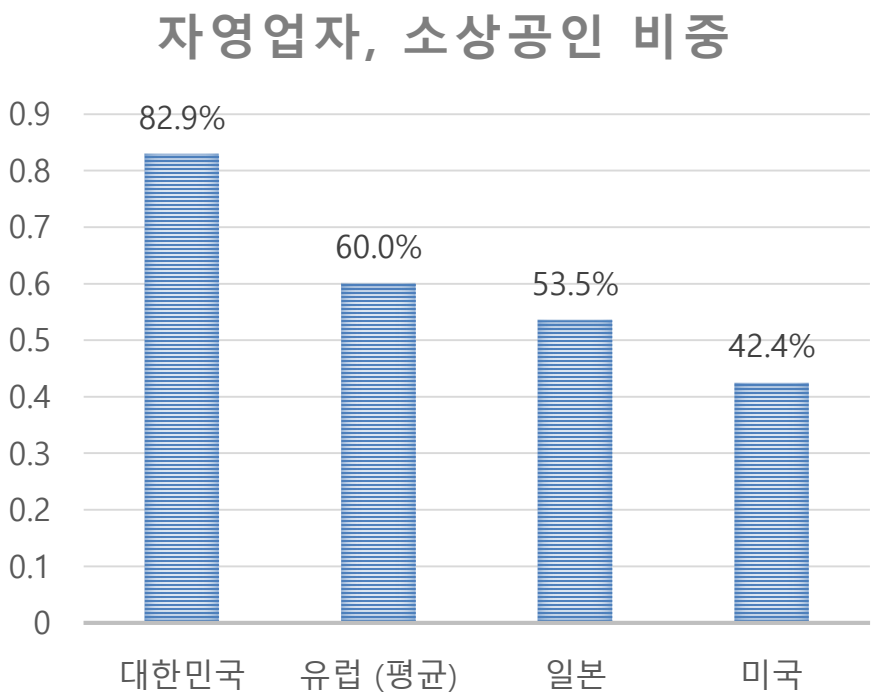
5조

연구 배경

소상공인 사업체 수 비율
전체 사업체 수의 85.3% 차지
(출처: 통계청)

자영업자, 소상공인 비중이
해외 선진국보다 훨씬 높음
(출처: OECD)





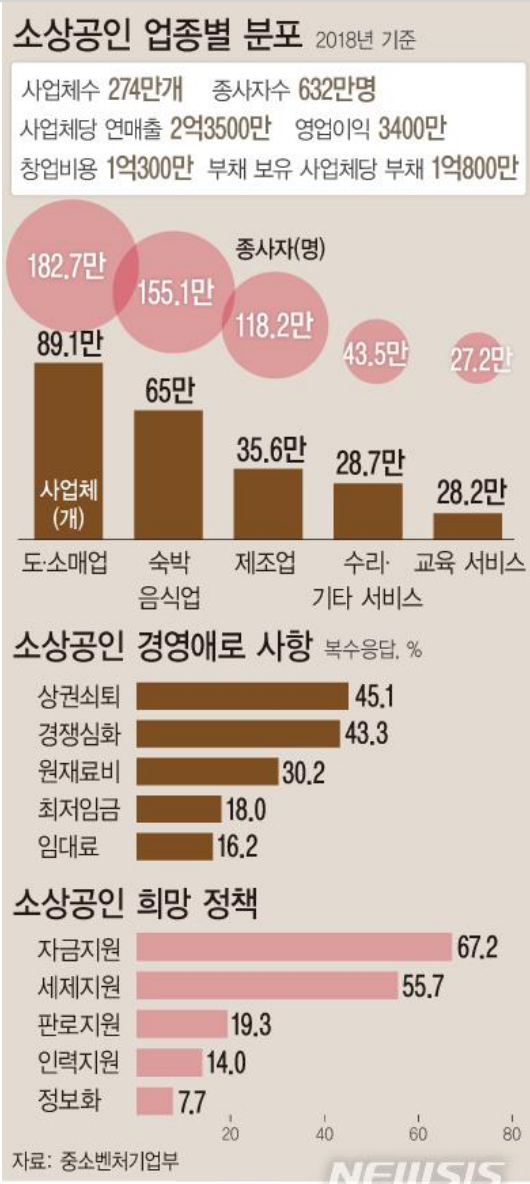
연구 배경

소상공인 절반 5년 내 망한다...10명중 9명 "최저임금 못 견뎌"

창업후 데스밸리 못건너도 정부 재기책 이용률은 11% 그쳐

최저임금 문제에 대해서는 59% "직원 줄였다"...업계 조사

(서울=뉴스1) 최동현 기자 | 2019-05-30 12:11 송고



연구 배경

"소상공인 절반 5년 내 망한다" (뉴스1 2019. 5. 30)

"소상공인, 2곳 중 1곳은 '빚'…평균 부채 1억8100만원" (뉴시스 2019. 12. 27)

□ 정부와 지자체의 지원정책 O but 개별 소상공인 대상 X

□ 소상공인은 진입장벽이 낮은 골목상권을 중심으로 창업활동

□ 발달상권에 비해 골목상권의 생존율은 현저히 낮음

소상공인 지원 정책에 필요

신규 소상공인 창업의 지표

골목상권 분석 필요

선행 연구 분석

선행연구	독립변수	종속변수	분석방법
빅데이터 분석을 통한 서울시 골목상권 분석 (2017)	• 아파트 (면적·가격별 세대 수), • 유동인구 (나이, 시간대, 성별, 요일) • 주거인구 (나이, 성별) • 직장인구 (나이, 성별) • 매출 (요일, 성별, 나이) • 수입 및 지출 (식품, 교육 등) • 점포	골목상권의 전체 매출액	~~다중회귀분석~~ 클러스터링
지리가중회귀분석을 이용한 고객 특성별 골목상권 매출액 영향 연구 (2018)	• 고객특성(성별, 연령대) • 입지특성 (상권 내 종사자수, 사업체 창업률, 지하철역 및 버스정류장과의 거리) • 구조특성 (건축물밀도, 골목상권 면적)	골목상권의 전체 매출액	OLS GWR (지리가중회귀분석)
서울시 골목상권 매출액에 영향을 미치는 요인에 관한 연구(2019)	• 상권특성 (연령, 시간대, 업종수, 상권면적, 유동인구) • 도시공간구조특성 (도시, 부도심 용도지역, 발달상권 인접 더미변수), • 배후지역 특성 (배후지역 아파트 가구수, 비아파트 가구수, 배후지역 월평균 소득금액의 로그값, 대형상업시설 영향;더미변수)	골목상권의 전체 매출액	다중회귀분석

선행 연구 분석

선행연구	문제 정의	결론	한계
빅데이터 분석을 통한 서울시 골목상권 분석 (2017)	• 업종별로 성격이 비슷한, 지역구로 구성된 상권을 정의하고 대표적인 특성을 파악	• 업종에 따라 다르게 형성되는 서울시 구 단위 군집의 구성요소를 확인 후, 각 군집별 특징 및 매출상관요인을 분석 확인	• 다중공선성으로 인해 회귀분석은 실패 • 활용 데이터의 양적인 한계
지리가중회귀분석을 이용한 고객 특성별 골목상권 매출액 영향 연구 (2018)	• 고객특성이 골목상권 매출액에 미치는 영향 분석	• 방법론적인 면에서 지리가중회귀분석이 OLS 회귀분석보다 더 우수한 것을 확인 • 골목상권별로 매출액 영향 요인을 식별	• 고객특성 중 성별과 연령에 한정된 분석만 수행했다는 점에서 한계
서울시 골목상권 매출액에 영향을 미치는 요인에 관한 연구(2019)	• 서울시 골목상권 매출액에 영향을 미치는 요인을 상권특성, 배후지역 특성, 공간구조 특성 등으로 구분하여 규명	• 골목상권이 지리적 입지 여건에 따라 다른 특성을 보이는 것을 확인 • 매출상관요인을 분석 확인	• 개별 점포별 특성이 아닌 상권을 큰 단위로 분석 • 업종 고려 X • 상권분석이 구체화되지 못함

우리의 목표

□ 선행 연구논문의 한계점 극복, 더 나은 예측력 모델 제안

- 지리적 특성 변수 + 업종별 특성 포함 범주형 데이터 추가
- 군집화(Clustering) 및 세분화(Segmentation)

  → **다수준(multi-level) 분석을 통해 예측력 강화**

□ 2015~2019, 5년간 축적된 데이터셋 활용

- 10개의 데이터 테이블, 1144개 컬럼, 약 39만개의 데이터

  → **기존 연구논문의 데이터 양적 한계를 극복**

□ **더 높은 R-squared 값을 갖는 모델 도출**

분석 대상 – 서울시 골목상권

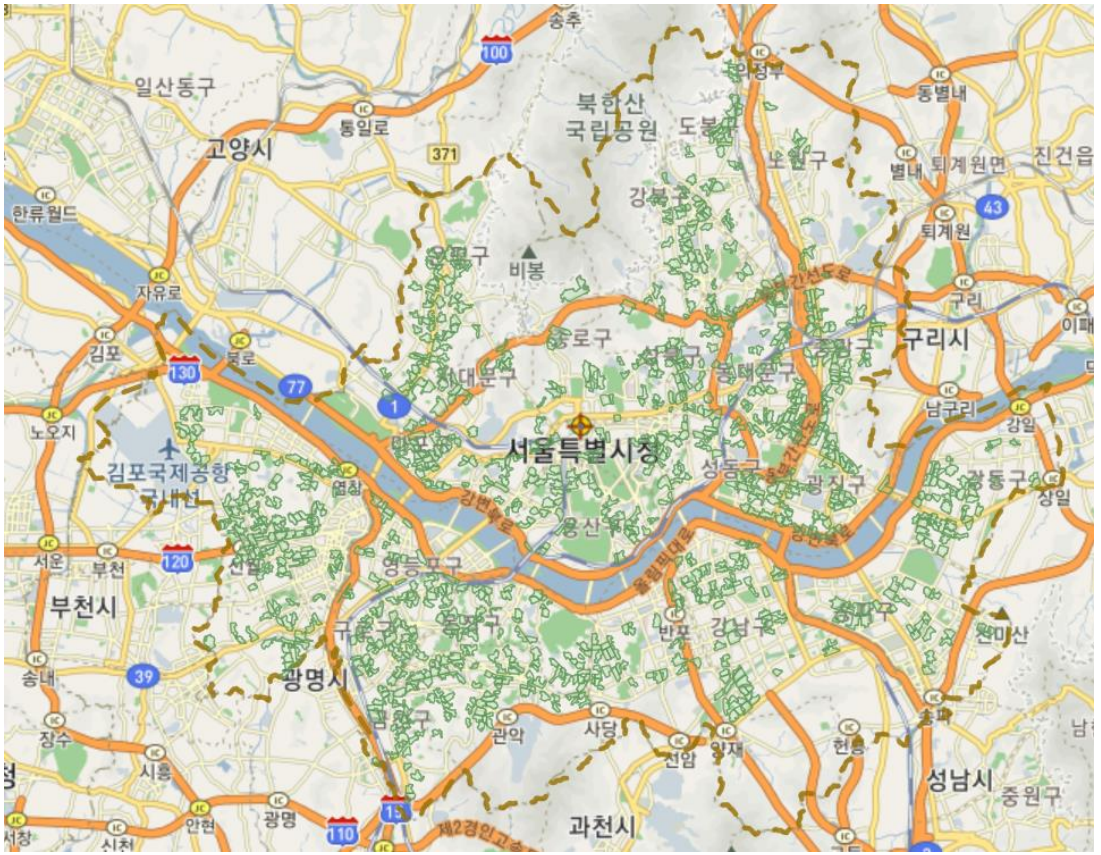
골목상권의 사전적 정의

골목점포 정의

① 생활밀접업종을 포함한 점포
② 발달상권에 포함되지 않는 점포
③ 배후지가 주거밀집 지역에 포함되는 점포
④ 전통시장에 포함되지 않는 점포
⑤ 길에 위치한 점포

골목상권 정의

① 일정 점포 수 이상의 상권
② 골목점포의 밀집도가 높은 상권



전처리 방법


서울 열린데이터광장
9개의 데이터
약 90만개의
매출 데이터
크롤링
병합 후, 연도, 분기, 상권코드, 업종코드로
Grouping
상권코드가 Category embedding 된
각 상권별 업종 단위의 데이터
결측치 제거/보정
변수 분석 및 선택

전처리 방법 / EDA 과정

종속 변수	
Distribution	특이사항 없음
Outliers	이상치 확인
Missing Data	해당사항 없음

독립 변수	
Distribution	특이사항 없음
Outliers	Robust Scaler
Missing Data	0으로 대체
Correlation	수치형 독립변수엔 다중공선성을 유발하는 선형 종속이 없음

독립 변수와 종속 변수	
Scatter plot	특이사항 없음
Linear or Non-linear	특이사항 없음
Correlation	특이사항 없음

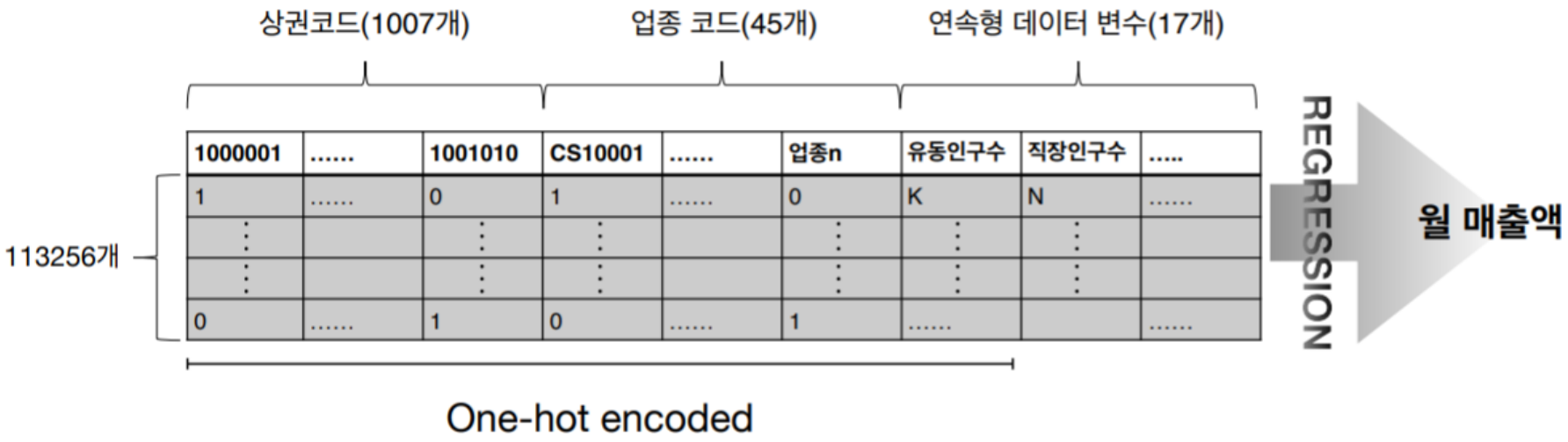
전처리 방법 / EDA

독립변수

변수명	설명
sales_female	여성의 매출액 / 총 매출액 (50m Cell 단위)
sales_2030s	전체 매출액 대비 20-30대가 지출한 매출액의 비율
sales_06_11	전체 매출액 대비 06시- 11시의 매출액의 비율
sales_11_14	전체 매출액 대비 11시- 14시의 매출액의 비율
sales_14_17	전체 매출액 대비 14시- 17시의 매출액의 비율
sales_17_21	전체 매출액 대비 17시- 21시의 매출액의 비율
sales_21_24	전체 매출액 대비 21시- 24시의 매출액의 비율
sales_weekday	전체 매출액 대비 주중 매출액의 비율
s_work_female	총 직장인구 대비 여성 직장인 인구의 비율 (50m Cell 단위)
s_float_male	총 유동인구 대비 남성 유동인구의 비율 (50m Cell 단위)
s_float_female	총 유동인구 대비 여성 유동인구의 비율 (50m Cell 단위)
b_facil_total	총 집객시설 수
s_store_no_of_store	사업자등록번호 기반 서울시 소재 사업체 수
s_store_no_of_opening	개업 점포수 (개업 신고 사업자)
s_store_no_of_closing	폐업 점포수 (폐업 신고 사업자)
b_income_avg_monthly_inc	국민건강보험공단의 건강보험료 납부 20분위를 기준소득월액으로 환산하여 주거지 기반으로 소득분위(10분위)를 산출한 배후지역의 월별 평균소득 정보
b_apt_avg_price	서울시 공간정보담당관에서 제공된 아파트 DB기반으로 산출한 아파트 평균시가를 1평당 가격으로 산출

전처리 방법 / EDA

종속변수	sales_total	3개 카드사의 카드승인금액 + 서울시 자체 보정 45개 생활밀집업종 매출액



LR 분석) 골목상권 1007개 + 업종 45개 + 수치형 데이터 17개

OLS Regression Results

Dep. Variable:	sales_total	R-squared:	0.443
Model:	OLS	Adj. R-squared:	0.436
Method:	Least Squares	F-statistic:	66.62
Date:	Thu, 30 Jul 2020	Prob (F-statistic):	0.00
Time:	17:50:22	Log-Likelihood:	-2.0276e+06
No. Observations:	90612	AIC:	4.057e+06
Df Residuals:	89544	BIC:	4.067e+06
Df Model:	1067		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
s_store_no_of_store	1.348e+08	1.27e+06	106.360	0.000	1.32e+08	1.37e+08
s_store_no_of_opening	2.019e+07	4.76e+06	4.238	0.000	1.09e+07	2.95e+07
s_store_no_of_closing	5.373e+07	5.01e+06	10.714	0.000	4.39e+07	6.36e+07
s_work_female	4.625e+04	4241.472	10.905	0.000	3.79e+04	5.46e+04
s_float_male	581.9175	90.342	6.441	0.000	404.847	758.988
s_float_female	-102.5411	94.624	-1.084	0.279	-288.003	82.921
b_facil_total	-1.747e+06	2.78e+05	-6.284	0.000	-2.29e+06	-1.2e+06
b_apt_avg_price	0.4037	0.274	1.473	0.141	-0.133	0.941
b_income_avg_monthly_inc	560.2975	32.085	17.463	0.000	497.411	623.184
sales_weekday	-6.113e+08	3.96e+07	-15.427	0.000	-6.89e+08	-5.34e+08
sales_female	6.653e+07	2.99e+07	2.226	0.026		
sales_2030s	-7.684e+07	2.89e+07	-2.661	0.008		
sales_06_11	3.779e+08	9.29e+07	4.067	0.000		
sales_11_14	1.408e+08	7.87e+07	1.789	0.074		
sales_14_17	2.207e+08	7.92e+07	2.786	0.005		
sales_17_21	3.074e+08	7.48e+07	4.111	0.000		
sales_21_24	2.832e+08	9.7e+07	2.921	0.003		
district_1000001	-1.012e+08	1.38e+08	-0.733	0.464		
district_1000002	1.084e+07	3.53e+08	0.031	0.976		
district_1000003	5.798e+08	1.32e+08	4.406	0.000		
district_1000004	1.076e+08	1.43e+08	0.751	0.453		
district_1000005	-5.478e+08	1.7e+08	-3.214	0.001		
district_1000006	1.718e+08	1.68e+08	1.020	0.308		
district_1000007	6.052e+08	1.85e+08	3.269	0.001	2.42e+08	9.68e+08
district_1000008	4.558e+08	2.23e+08	2.046	0.041	1.92e+07	8.92e+08
district_1000009	2.393e+08	1.96e+08	1.219	0.223	-1.45e+08	6.24e+08
district_1000010	-4.548e+08	1.29e+08	-3.533	0.000	-7.07e+08	-2.02e+08

Omnibus:	116172.839	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	85264505.736
Skew:	6.589	Prob(JB):	0.00
Kurtosis:	152.700	Cond. No.	2.25e+15

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.25e+15. This might indicate that there are strong multicollinearity or other numerical problems.

- ☐ R-squared: 0.443
- ☐ 다중공선성 존재
  → 모델을 명확히 설명하기 어렵다
- ☐ 1007개 데이터를 모두 사용하니 데이터의 차원이 과도하게 높아짐
- ☐ 1007개를 줄여보기로 결정

LR 분석) 시군구 25개 + 업종 45개 + 수치형 데이터 17개

OLS Regression Results

Dep. Variable:	sales_total	R-squared:	0.297
Model:	OLS	Adj. R-squared:	0.296
Method:	Least Squares	F-statistic:	931.4
Date:	Thu, 30 Jul 2020	Prob (F-statistic):	0.00
Time:	18:20:27	Log-Likelihood:	-2.0380e+06
No. Observations:	90612	AIC:	4.076e+06
Df Residuals:	90570	BIC:	4.076e+06
Df Model:	41		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
s_store_no_of_store	4.909e+08	4.81e+06	102.005	0.000	4.81e+08	5e+08
s_store_no_of_opening	4.06e+07	5.12e+06	7.934	0.000	3.06e+07	5.06e+07
s_store_no_of_closing	9.224e+07	5.4e+06	17.073	0.000	8.16e+07	1.03e+08
s_work_female	1.805e+07	2.13e+06	8.463	0.000	1.39e+07	2.22e+07
s_float_male	1.804e+08	1.77e+07	10.183	0.000	1.46e+08	2.15e+08

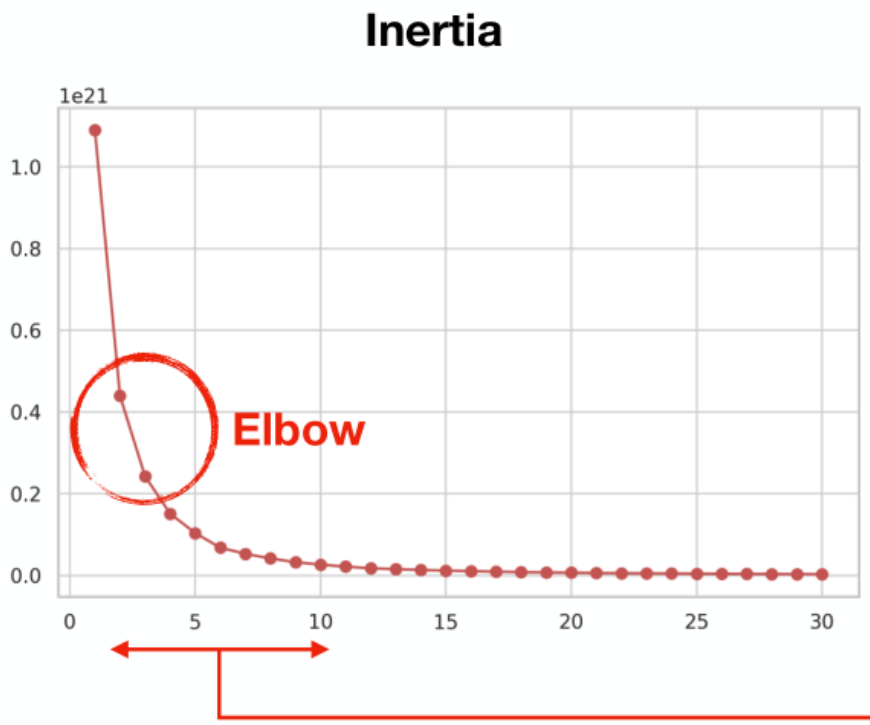
Omnibus:	113870.598	Durbin-Watson:	1.993
Prob(Omnibus):	0.000	Jarque-Bera (JB):	63648986.018
Skew:	6.434	Prob(JB):	0.00
Kurtosis:	132.201	Cond. No.	30.9

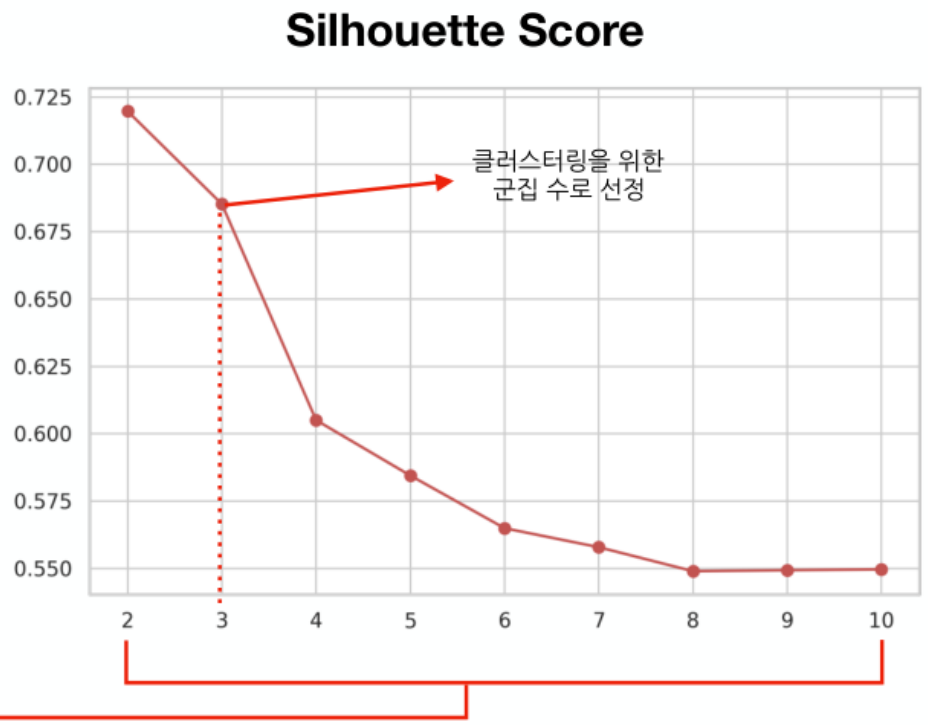
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

- 1007개의 골목상권을 지리적 특성 '시군구'로 통합
- 다중공선성은 제거됨
- R-squared: 0.297
  → 너무 낮은 값이 도출됨

LR 분석) K-means 3개 + 업종 45개 + 수치형 데이터 17개

이너시아 밸류, 실루엣 계수를 이용한 적정 군집수 판단
N = 3



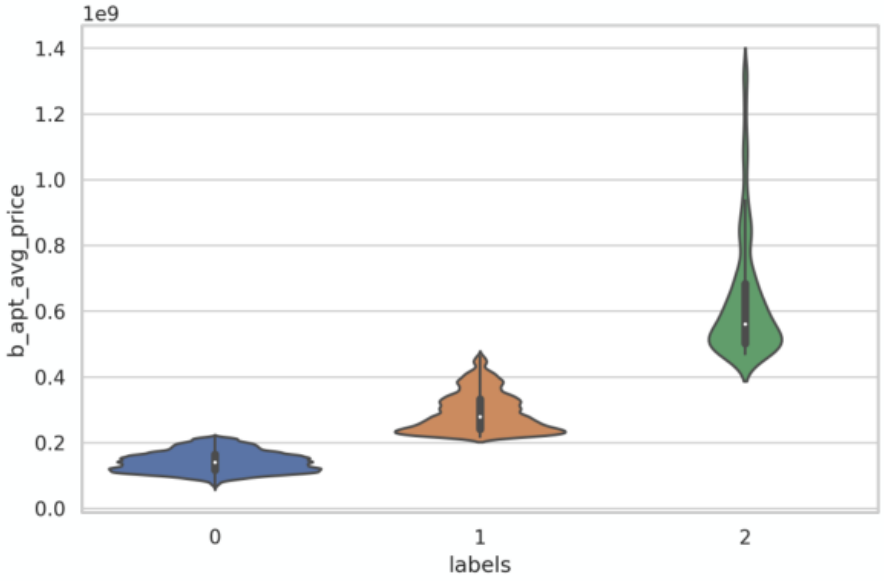


LR 분석) K-means 3개 + 업종 45개 + 수치형 데이터 17개

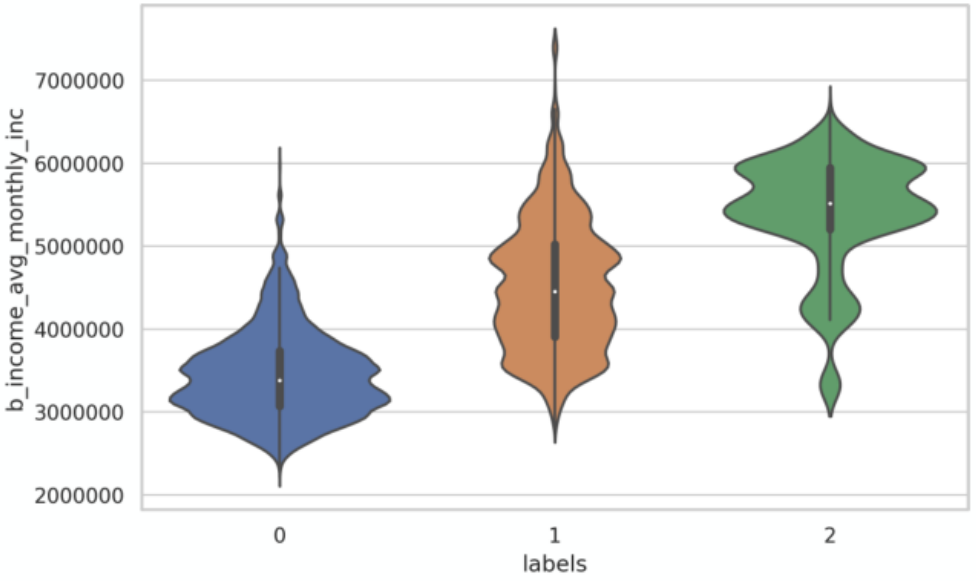
N = 3

군집 별 특성을 파악

배후지 아파트 평균시가



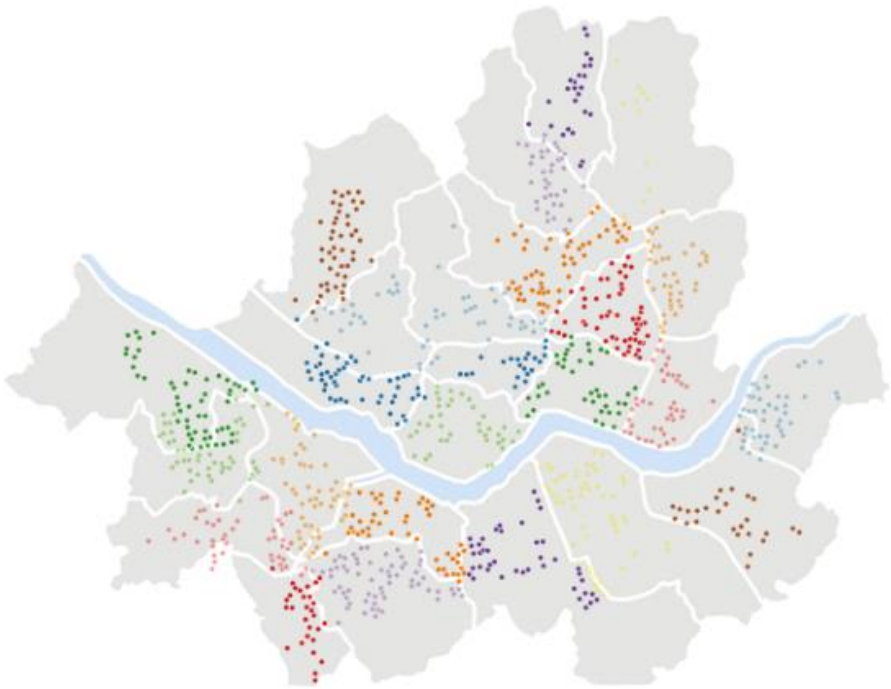
배후지 월 평균소득



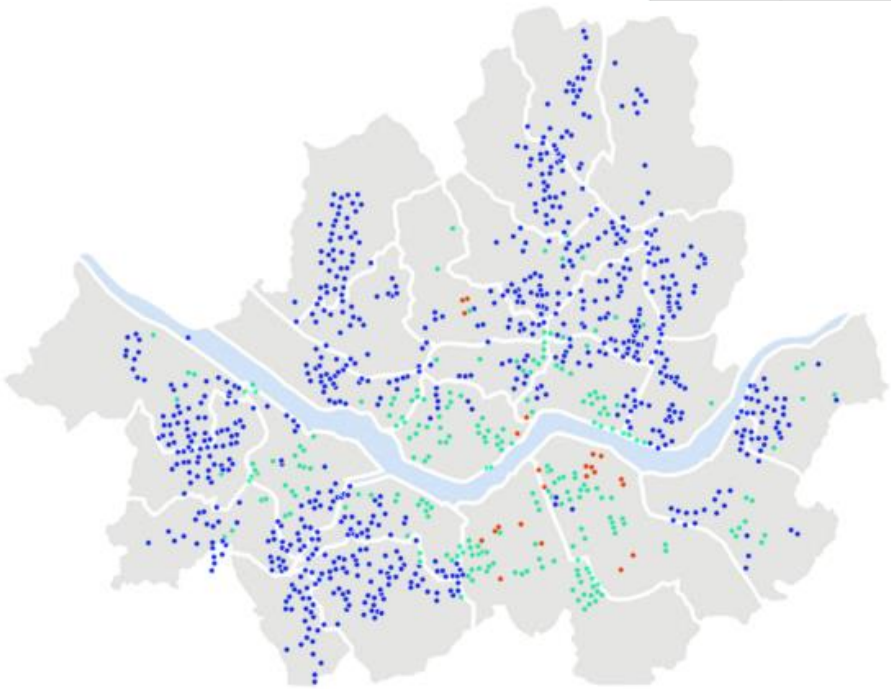
LR 분석) K-means 3개 + 업종 45개 + 수치형 데이터 17개

월 평균 소득 및 아파트 가격이 높은 지역들이
군집 2에 분포되어 있음을 확인 가능

군집	데이터 수
0	87703
1	23514
2	2048



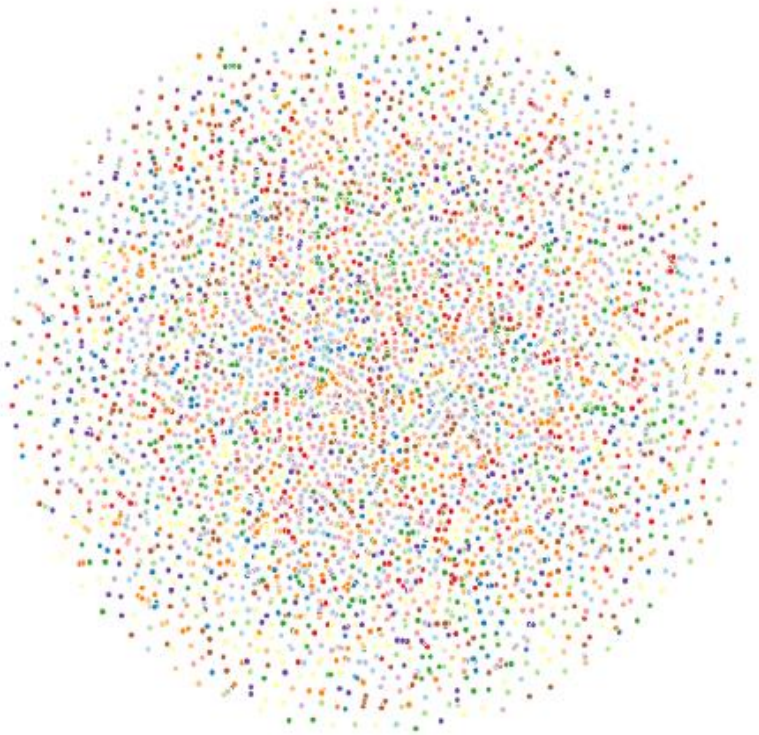
Geographical labeling (시군구)



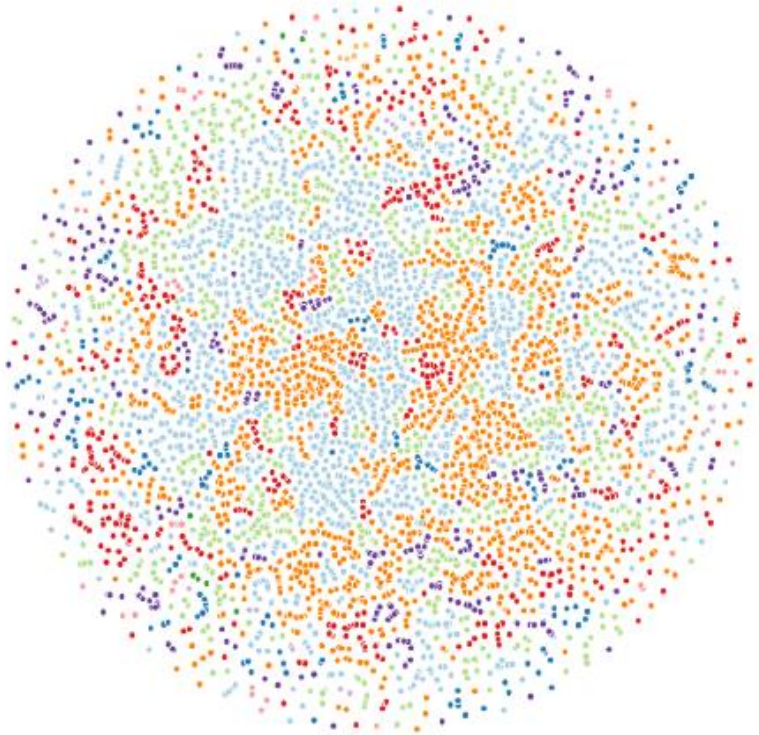
Data-driven labeling (Clustering)

LR 분석) K-means 3개 + 업종 45개 + 수치형 데이터 17개

클러스터 군집수를 10개로 늘려 시각화한 결과
시군구 대비 유의미한 군집을 형성, 비슷한 속성끼리 묶임



Geographical labeling (시군구)



Data-driven labeling (Clustering)

LR 분석) K-means 3개 + 업종 45개 + 수치형 데이터 17개

OLS Regression Results

Dep. Variable:	sales_total	R-squared:	0.406
Model:	OLS	Adj. R-squared:	0.405
Method:	Least Squares	F-statistic:	981.4
Date:	Thu, 30 Jul 2020	Prob (F-statistic):	0.00
Time:	18:28:02	Log-Likelihood:	-2.0305e+06
No. Observations:	90612	AIC:	4.061e+06
Df Residuals:	90548	BIC:	4.062e+06
Df Model:	63		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
s_store_no_of_store	5.384e+08	5.03e+06	107.016	0.000	5.29e+08	5.48e+08
s_store_no_of_opening	3.293e+07	4.83e+06	6.820	0.000	2.35e+07	4.24e+07
s_store_no_of_closing	6.456e+07	5.09e+06	12.681	0.000	5.46e+07	7.45e+07
s_work_female	1.909e+07	1.91e+06	9.977	0.000	1.53e+07	2.28e+07
s_float_male	2.422e+08	1.59e+07	15.236	0.000	2.11e+08	2.73e+08
s_float_female	-8.173e+07	1.64e+07	-4.971	0.000	-1.14e+08	-4.95e+07
b_facil_total	-7.283e+06	5.7e+06				
b_apt_avg_price	4.737e+07	8.48e+06				
b_income_avg_monthly_inc	1.164e+08	7.64e+06				
sales_weekday	-1.082e+08	7.19e+06				
sales_female	1.289e+07	8.7e+06				
sales_2030s	3.922e+07	8.23e+06				
sales_06_11	2.925e+07	1.02e+07				
sales_11_14	1.811e+07	1.43e+07				
sales_14_17	2.469e+07	1.55e+07				
sales_17_21	5.015e+07	1.43e+07				
sales_21_24	6.869e+07	1.95e+07				
labels_0	3.682e+08	1.07e+07				
labels_1	3.44e+08	1.79e+07				
labels_2	5.322e+08	5.55e+07				

Omnibus:	117588.613	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	88690169.806
Skew:	6.743	Prob(JB):	0.00
Kurtosis:	155.673	Cond. No.	2.25e+15

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.34e-25. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

- 다중공선성이 다시 나타남
- R-squared: 0.406
- 데이터의 설명력이 부족

LR 분석) t-sne 군집화 + 업종 + 수치형 데이터

OLS Regression Results

Dep. Variable:	sales_total	R-squared:	0.406
Model:	OLS	Adj. R-squared:	0.405
Method:	Least Squares	F-statistic:	980.9
Date:	Thu, 30 Jul 2020	Prob (F-statistic):	0.00
Time:	18:34:02	Log-Likelihood:	-2.0305e+06
No. Observations:	90612	AIC:	4.061e+06
Df Residuals:	90548	BIC:	4.062e+06
Df Model:	63		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
s_store_no_of_store	5.378e+08	5.03e+06	106.895	0.000	5.28e+08	5.48e+08
s_store_no_of_opening	3.315e+07	4.83e+06	6.866	0.000	2.37e+07	4.26e+07
s_store_no_of_closing	6.503e+07	5.09e+06	12.774	0.000	5.51e+07	7.5e+07
s_work_female	1.863e+07	1.91e+06	9.745	0.000	1.49e+07	2.24e+07
s_float_male	2.303e+08	1.58e+07	14.603	0.000	1.99e+08	2.61e+08

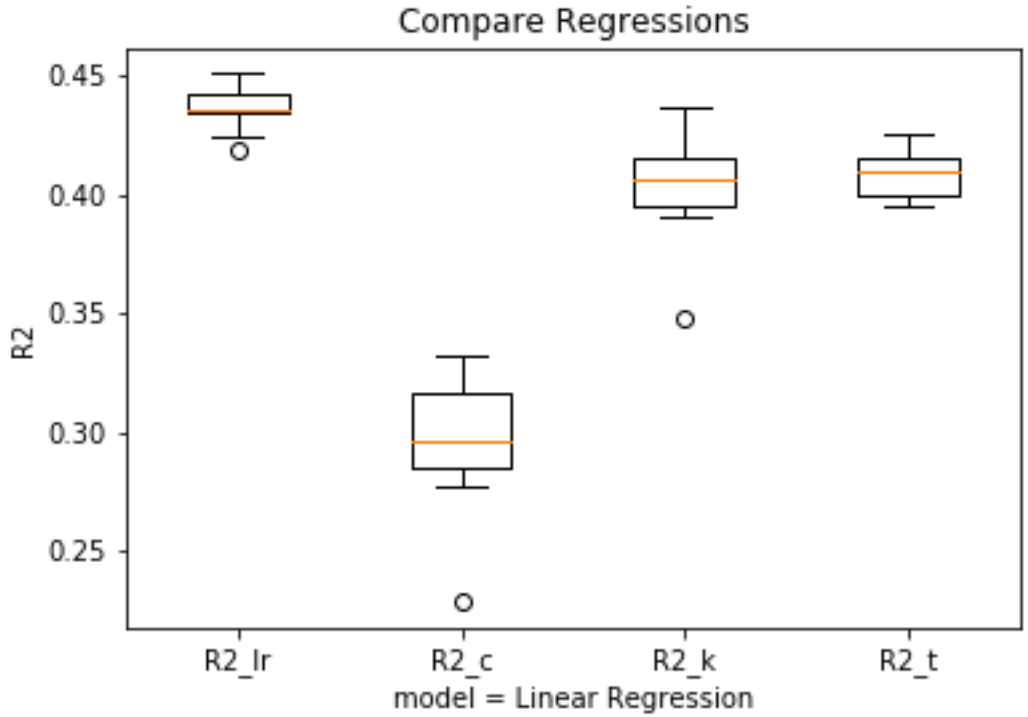
Omnibus:	117558.789	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	88574307.755
Skew:	6.739	Prob(JB):	0.00
Kurtosis:	155.573	Cond. No.	739.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

- ☐ 다중공선성은 사라짐
- ☐ R-squared: 0.406
- ☐ 여전히 데이터의 설명력이 부족

선형 회귀 모델의 문제점

- 다중공선성 문제
- 낮은 R-squared 값이 지속됨 **(0.297~0.443)**
- 각 모델을 10번씩 돌려 나온 평균 값 그래프



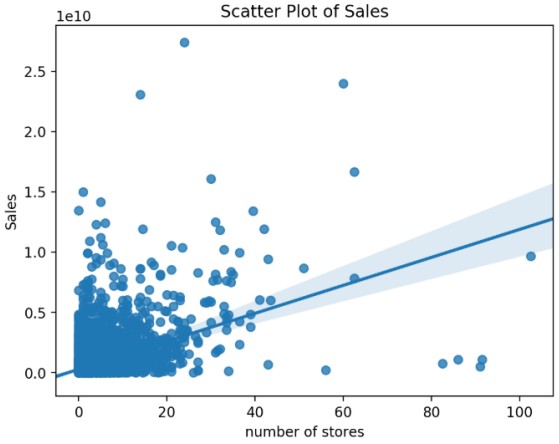
선형 회귀 모델은 분석에 적절하지 않다

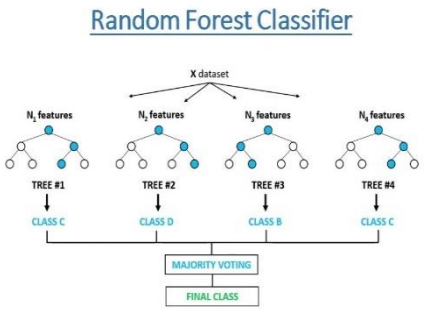
비선형 회귀 모델 비교

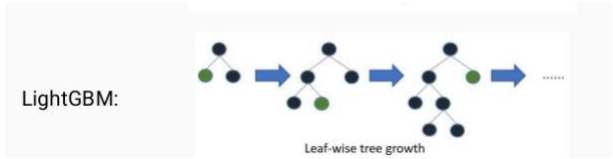
분석 내용

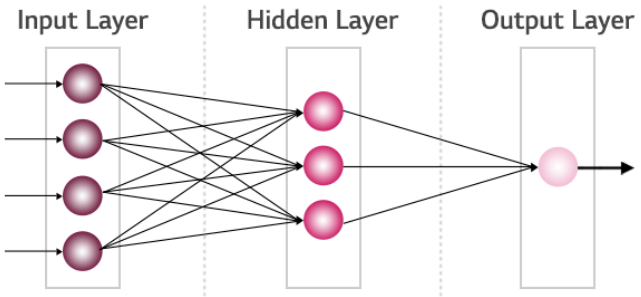
상권코드와 시군구, K-means, t-sne를 데이터에 적용해 가장 성능이 좋은 모델을 검증
R-squared 값이 가장 높고 RMSE 오차값이 가장 낮은 회귀 모델을 찾아보았음

[Linear Regression] vs [Random Forest] vs [Light GBM] vs [Neural Network]

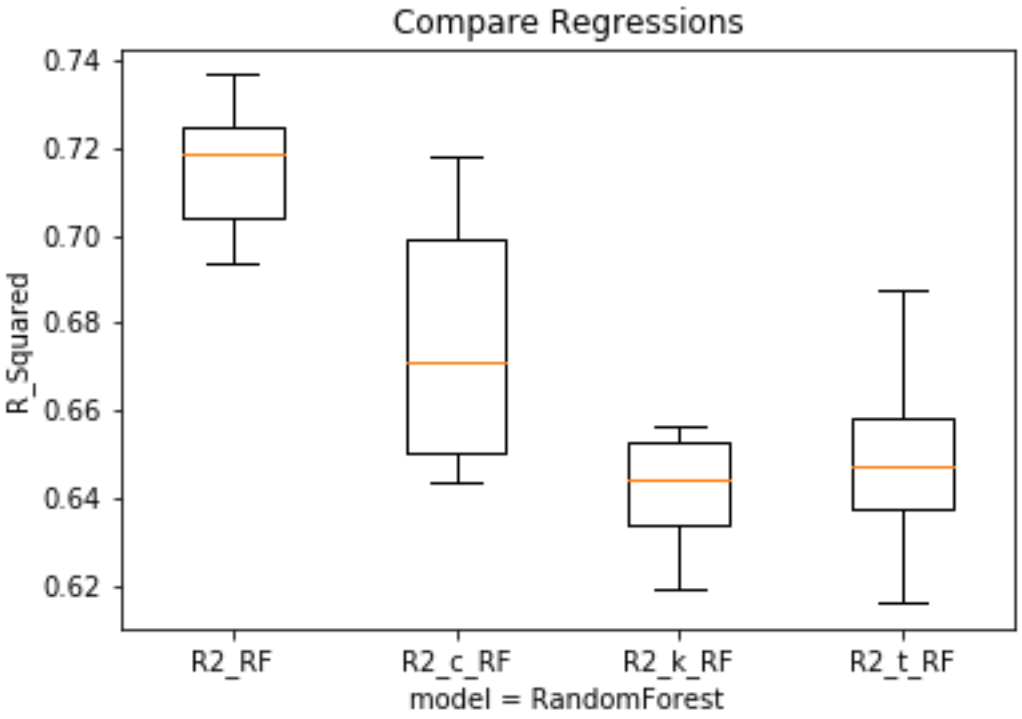








Random Forest 분석



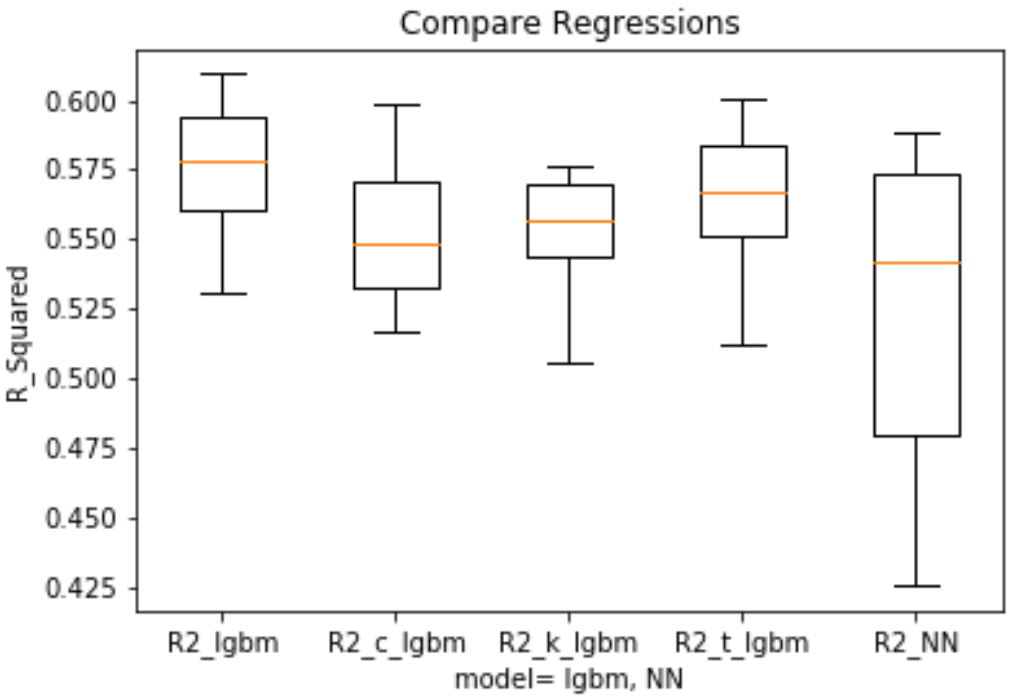
R-squared 값

최소 0.669
최대 0.728
평균 0.714

R2_RF = 1007개의 상권 데이터셋
R2_c_RF = 25개 업종 데이터셋
R2_k_RF = K-means 3개
R2_t_RF = t-sne 군집화

0.7을 넘는 매우 우수한 결과값이 도출됨

Light GBM / Neural Network 분석



LGBM/NN 모델의 R-squared 값

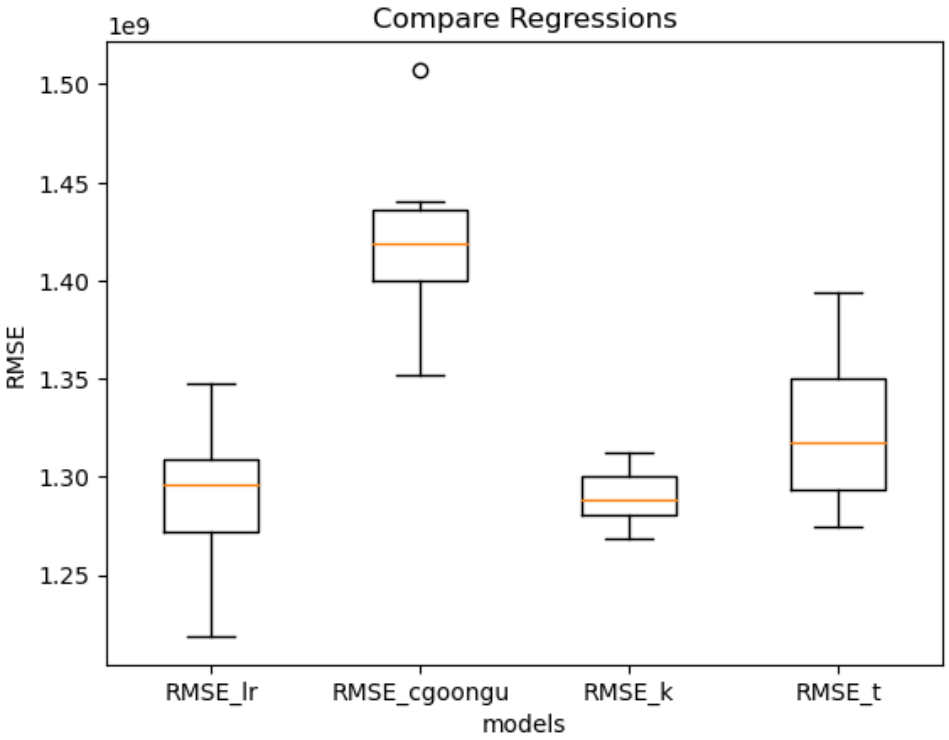
최소 0.515
최대 0.565
평균 0.551

R2_lgbm = 1007개 상권 데이터셋
R2_c_lgbm = 25개 업종 데이터셋
R2_k_lgbm = K-means 3개
R2_t_lgbm = t-sne 군집화
R2_NN = 1007개 + 인공신경망

선형 회귀모델보다는 결과값이 우수하나, Random Forest 보다는 낮음

Root Mean Squared Error 수치

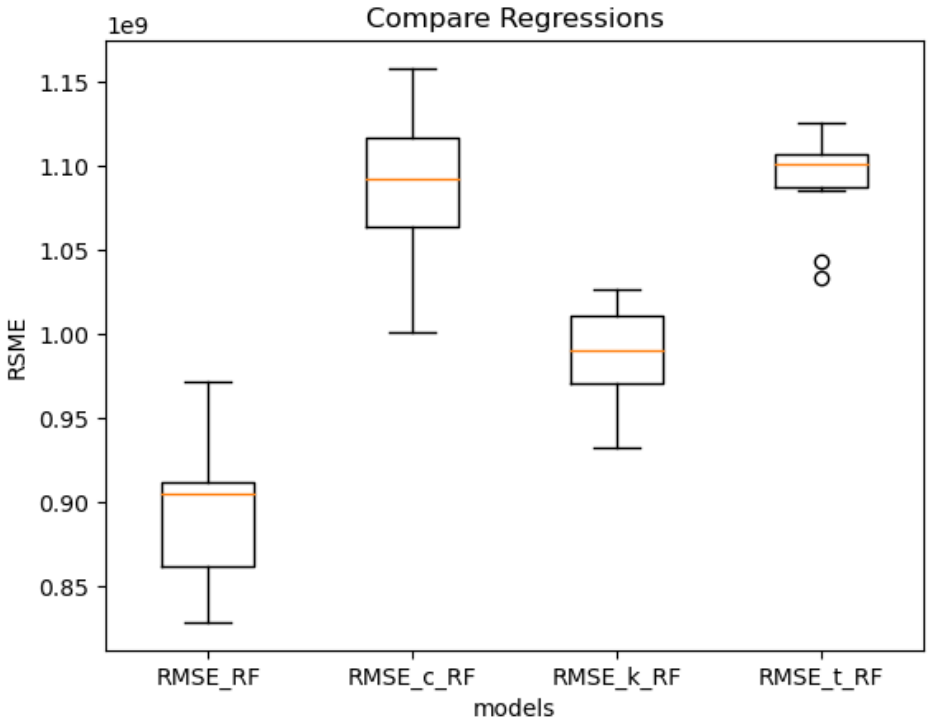
1) Linear regression



전반적으로 12억이 넘는 수치

Root Mean Squared Error 수치

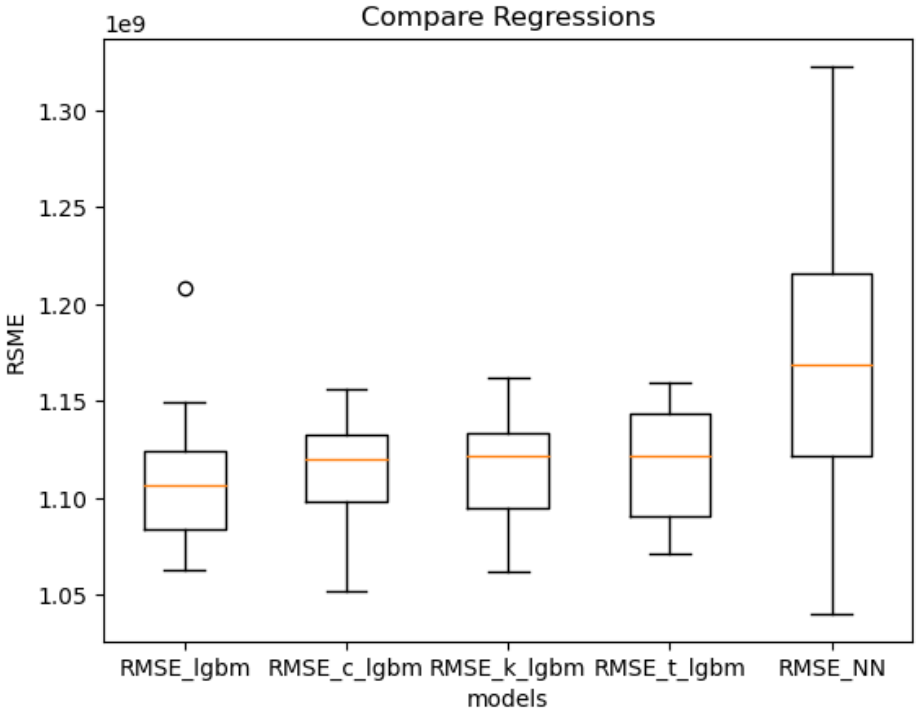
2) Random Forest



선형회귀 모델보다 낮은 오차값
약 8억~11억
가장 낮은 RMSE 값은 8억대

Root Mean Squared Error 수치

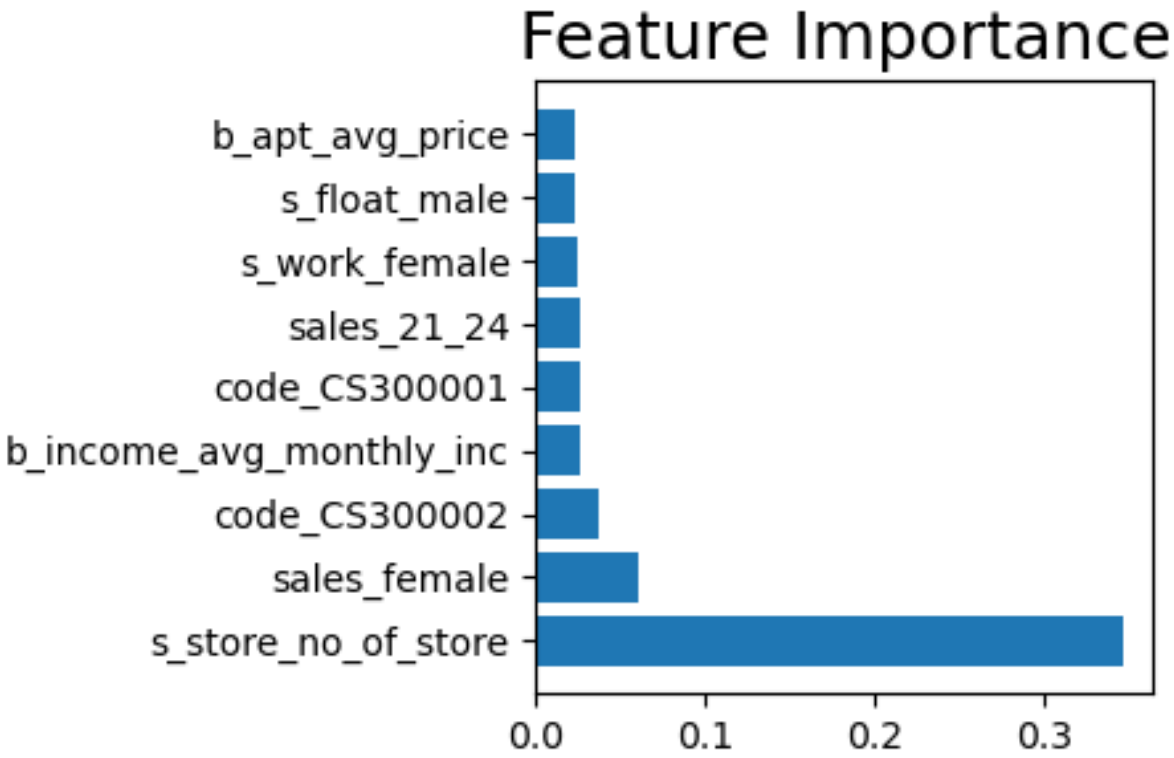
3) Light GBM / Neural Network



Light GBM은 11억대
Neural Network는 10억~13억대

예측률 0.7, 오차값 8억
Random Forest 분석이 가장 유의미한 결과 도출

인사이트 도출



1) 주변에 매장 수가 많을수록 매출액 증가 경향
2) 여성 매출액이 많을수록 매출액 증가 경향
3) 저녁 9시 ~ 다음날 0시 사이 시간대 매출액 증가 경향
4) 편의점, 슈퍼마켓 업종의 강세
5) 여성 직장인 인구, 남성 유동 인구가 많을수록 매출액 증가 경향

한계점

- 인공신경망 모델링이 생각처럼 잘 되지 않은 것

- K-means 이외의 클러스터링 방법을 시도하지 못한 것
  클러스터링 방법을 더 많이 익혀,
  다중공선성이 나타나지 않게끔 하고 싶다.


Thank
you!