Analysis of Sales in Alley Market in Seoul

서울시 골목상권 매출액 분석

- 클러스터링, 모델에 따른 설명력 변화

김건우 박동재 장상현 장우빈 허은정

Analysis of Sales in Alley Market in Seoul

개요

주제	서울시 골목상권 매출액 분석
활용 데이터	년도, 상권, 업종별 매출액 (서울시 45개 생활밀집업종), 시간, 성별, 직장인구, 유동인구, 집객시설 수, 사업체 수, 개폐업률, 배후지 평균 소득, 아파트 평균 시가 (서울 열린데이터광장)
문제 정의	상권, 업종 코드 범주화 (클러스터링) 타겟 변수인 매출액 예측 (회귀 모델링)
분석 과정	분석 배경 -> 데이터 전처리 & 탐색 -> 클러스터링 -> 모델링 -> 결론
분석 환경	Python 언어 사용, Jupyter Notebook 환경에서 작업
활용 패키지	numpy, pandas 연산, 데이터 조작 statsmodels, scikit-learn 모델링 matplotlib.pyplot, seaborn 시각화

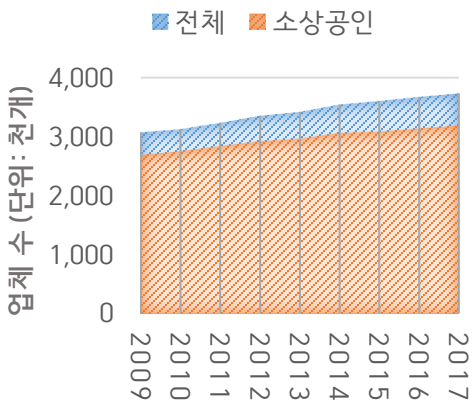
분석 배경

<환경 분석>

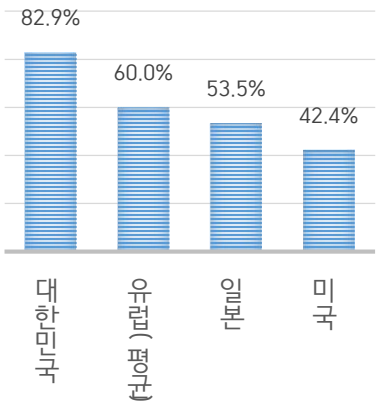
전체 사업체 대비 소상공인 비율 85.3%
(자료 1. 출처: 통계청)

자영업자, 소상공인 비중이
해외 선진국보다 훨씬 높음
(자료 2. 출처: OECD)

(자료 1)
소상공인 사업체수 현황



(자료 2)
자영업자, 소상공인 비중



"소상공인 절반 5년 내 망한다" (뉴스1 2019. 5. 30)
"소상공인, 2곳 중 1곳은 '빚'… 평균 부채 1억8100만원" (뉴시스 2019. 12. 27)

- 소상공인은 진입장벽이 낮은 골목상권 창업 중심
- 발달상권에 비해 골목상권의 낮은 생존율
- 개별 소상공인 대상 맞춤형 지원 정책 부재
- 신규 소상공인 창업 지표 부재

⇨ 골목상권 매출액 분석 필요

분석 배경

<선행 연구>

선행연구	문제 정의	결론	한계
빅데이터 분석을 통한 서울시 골목상권 분석 (2017)	업종, 지역구 기준으로 상권 정의, 대표 특성 파악(클러스터링, 회귀)	• 업종, 구별 군집 구성요소 확인 • 군집별 특징, 매출상관요인 분석	• 다중공선성 발생 • 데이터의 양적인 한계
GWR을 이용한 고객 특성별 골목상권 매출액 영향 연구 (2018)	골목상권 매출액 결정 고객 특성 분석 (OLS, GWR-지리가중회귀)	• GWR이 OLS 회귀분석보다 우수 • 골목상권별 매출액 영향 요인 식별	• 성별과 연령에 한정된 분석
서울시 골목상권 매출액에 영향을 미치는 요인에 관한 연구 (2019)	서울시 골목상권 매출액 결정 요인을 상권, 배후지, 공간구조 등으로 구분하여 규명 (회귀)	• 골목상권이 지리적 입지여건에 따라 다른 특성 • 매출상관요인 분석	• 점포 단위가 아닌 상권 단위 분석 • 업종 고려 X • 상권분석이 구체화되지 못함

<목표 설정>

1) 5년(2015~2019) 간 축적된 데이터셋 활용

10개 테이블, 1144개 컬럼, 약 39만 건의 데이터

→ **기존 연구의 데이터 양적 한계 극복**

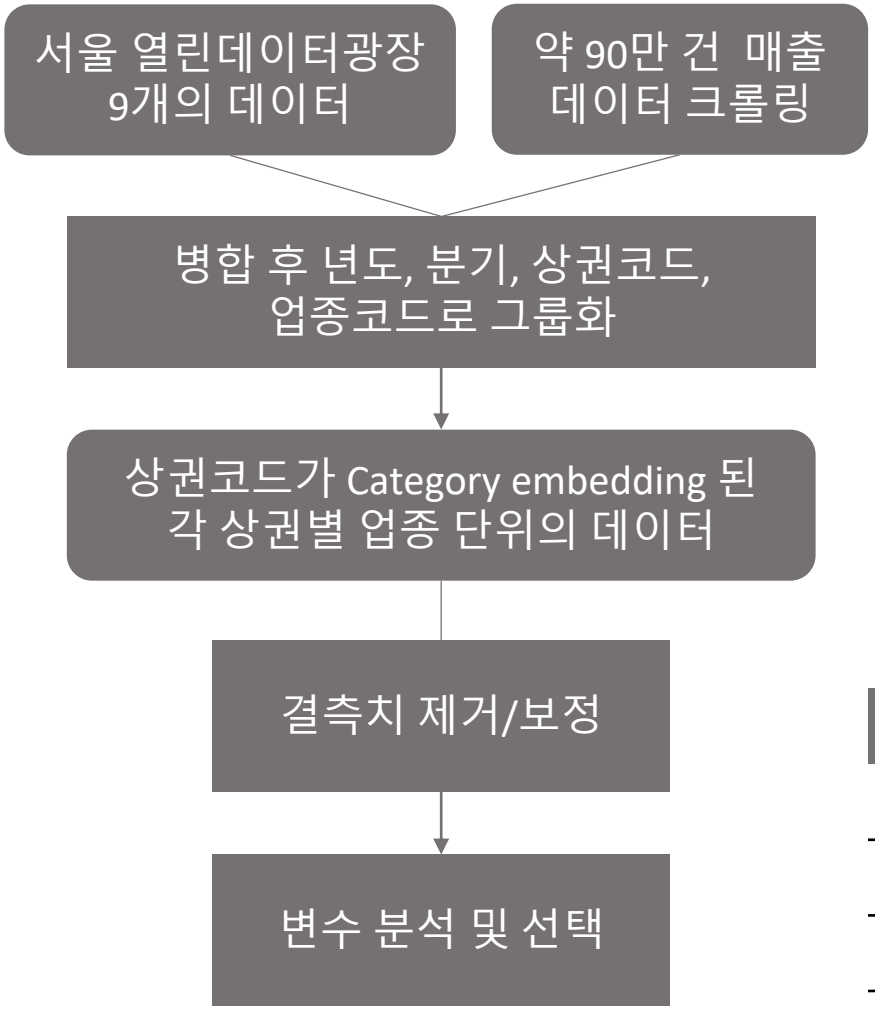
2) 선행 연구 한계 극복, 발전된 모델 제안

→ **지리 변수 + 업종 변수 추가**

→ **클러스터링을 통해 다중공선성 제거, 예측력 강화**

3) **더 높은 R-squared 값을 갖는 회귀 모델 도출**

데이터 전처리 & 탐색



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

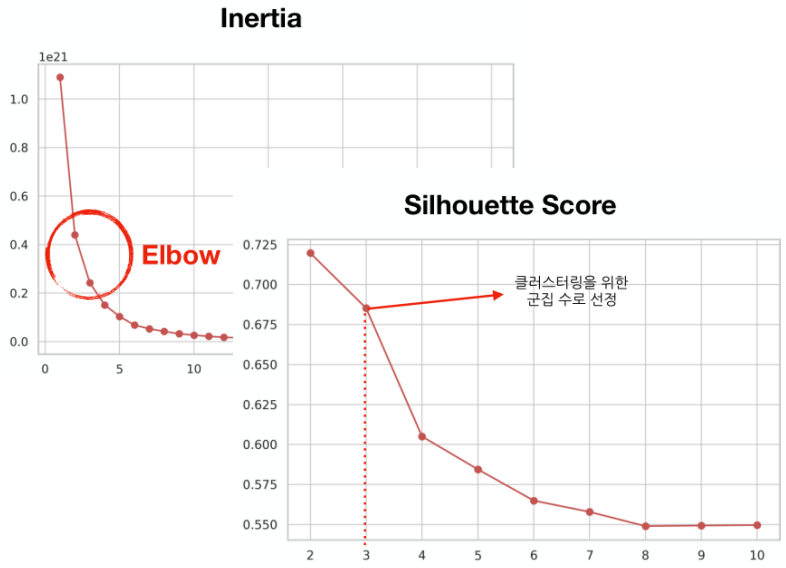
독립 변수와 종속 변수	
Scatter plot	특이사항 없음
Linearity	특이사항 없음
Correlation	특이사항 없음

독립 변수	
Distribution	특이사항 없음
Outliers	Robust Scaling
Missing Data	0으로 대체
Correlation	수치형 독립변수에는 다중공선성 없음

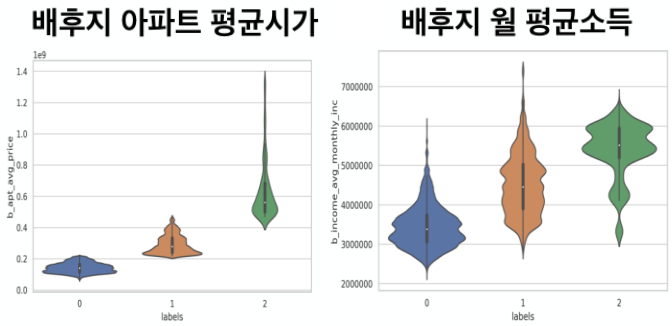
클러스터링 + Baseline 모델링

독립 변수	R2	다중공선성	설명
상권코드(1007개) 더미 변수 포함	0.443	존재	모델 해석에 어려움
시군구코드(25개) 더미 변수 포함	0.297	제거	설명력 낮음
수치형 변수 k-means 클러스터링	0.406	존재	설명력 개선 x
수치형 변수 k-means + t-sne 차원 축소	0.406	제거	설명력 개선 x

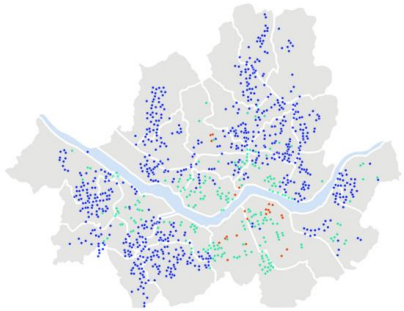
이너시아 밸류, 실루엣 계수를 이용한 적정 군집수 판단
N = 3



k-means로 구한 3개 군집 별 특성 파악



월 평균 소득 및 아파트 가격에 따른 차이



군집	데이터 수
0	87703
1	23514
2	2048

여러 가지 모델 결과 비교

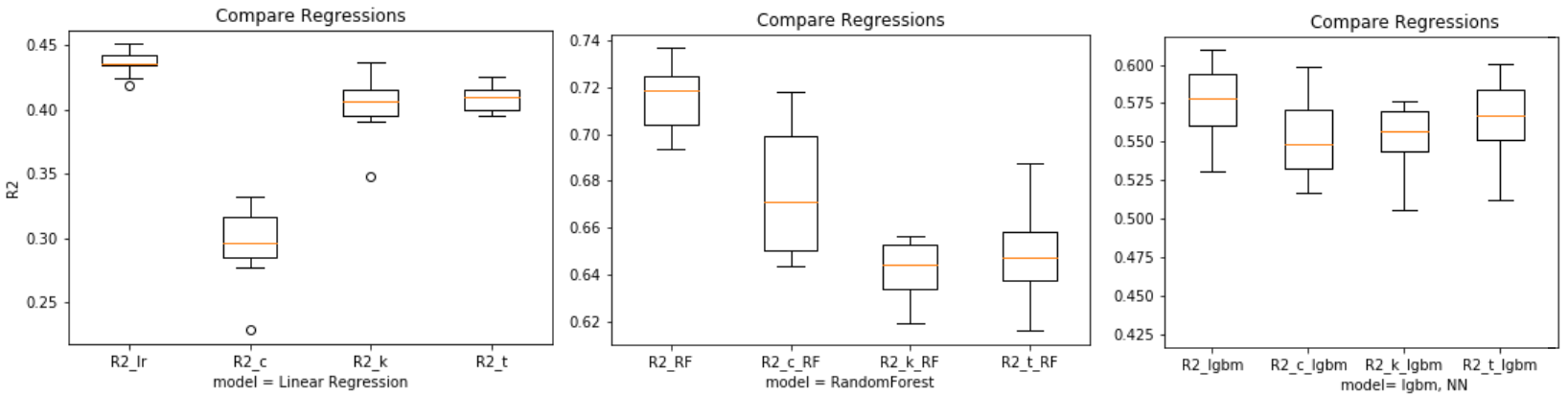
모델	활용 패키지	Min R2	Max R2	Mean R2
Linear Regression	statsmodels	0.297	0.443	0.388
Decision Tree (Random Forest)	scikit-learn	0.669	0.728	0.714
Decision Tree (Boosting)	lightgbm	0.515	0.565	0.551

Linear Regression, Decision Tree (Random Forest, Boosting) 세 가지 모델 별

상권코드, 시군구코드, k-means, t-sne로 생성한 더미 변수를

학습 데이터에 포함시킨 네 가지 결과 (총 3*4=12 case)

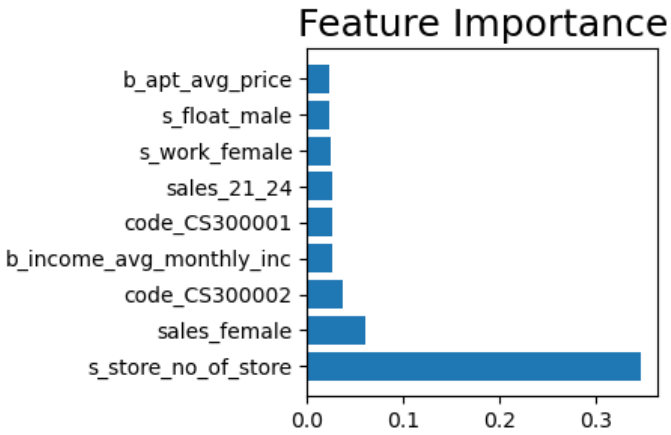
샘플링을 통해 case 별로 각각 10회 학습한 R-sqruared 값의 Box-plot



결론

<매출액 결정 주요 변수>

1) 주변 매장 수 ⇨ 양의
2) 여성 매출액 상관관계
3) 저녁 9시 ~ 다음날 0시
4) 편의점, 슈퍼마켓 업종
5) 여성 직장인, 남성 유동 인구



<한계>

1) 클러스터링에 K-means 알고리즘만을 적용
   -> DBSCAN 등의 추가 방법론 탐색 필요
2) 머신러닝 모델의 성능 개선 미흡
   -> Hyperparameter 조정하여 개선 필요

- 끝 -