

부동산 데이터를 활용한 소상공인 매출등급 예측

TEAM : 신사임당



원윤정

jennie071512@gmail.com



이다경

dakyeong98@naver.com



장희중(팀장)

vicky4150@korea.ac.kr



홍수빈

subinn2584@naver.com

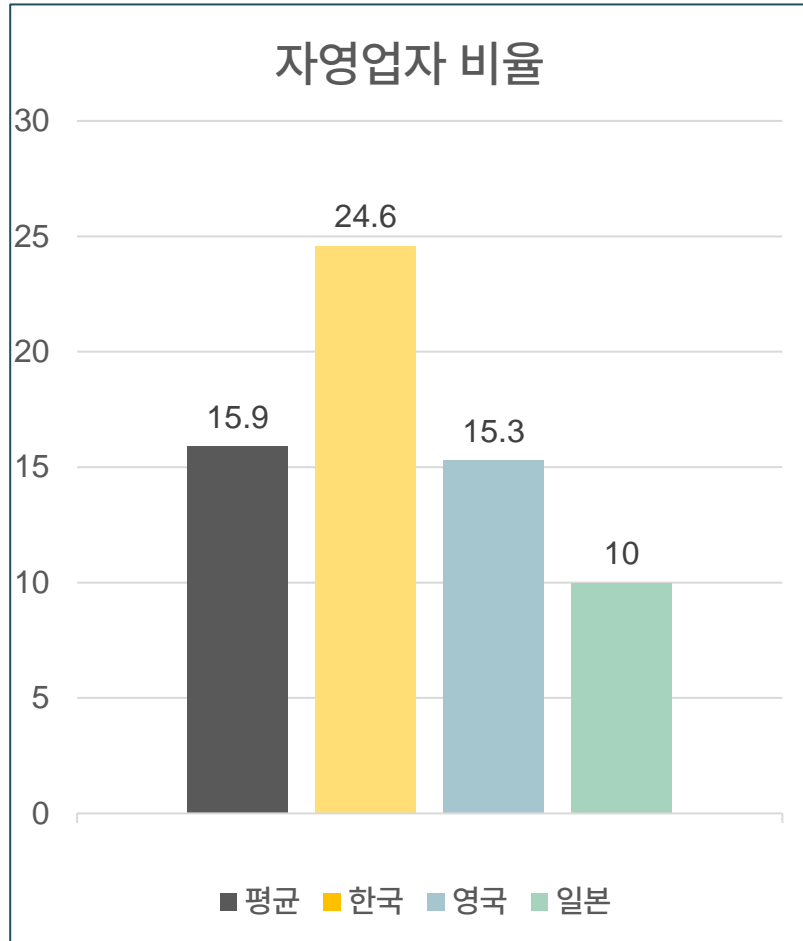
목차

- 1 분석 배경 및 목적
- 2 분석 과정
- 3 분석 결과
- 4 결과 활용 방안

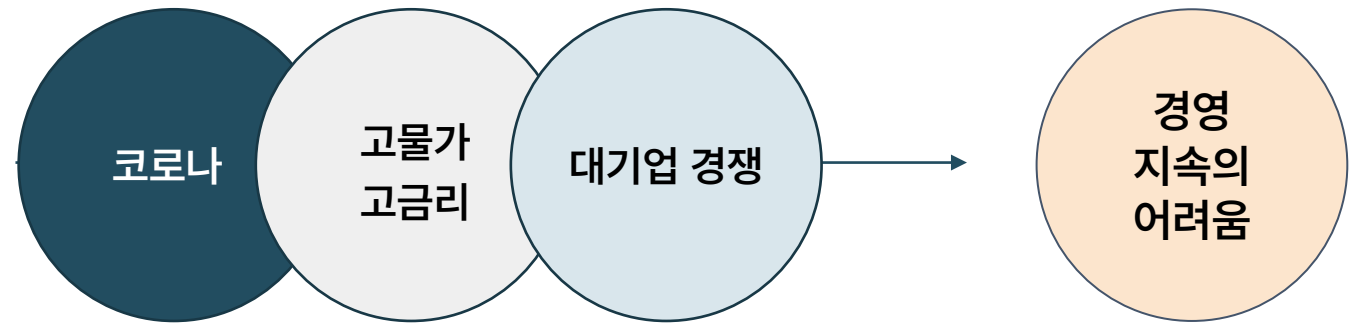


1

분석 배경 및 목적

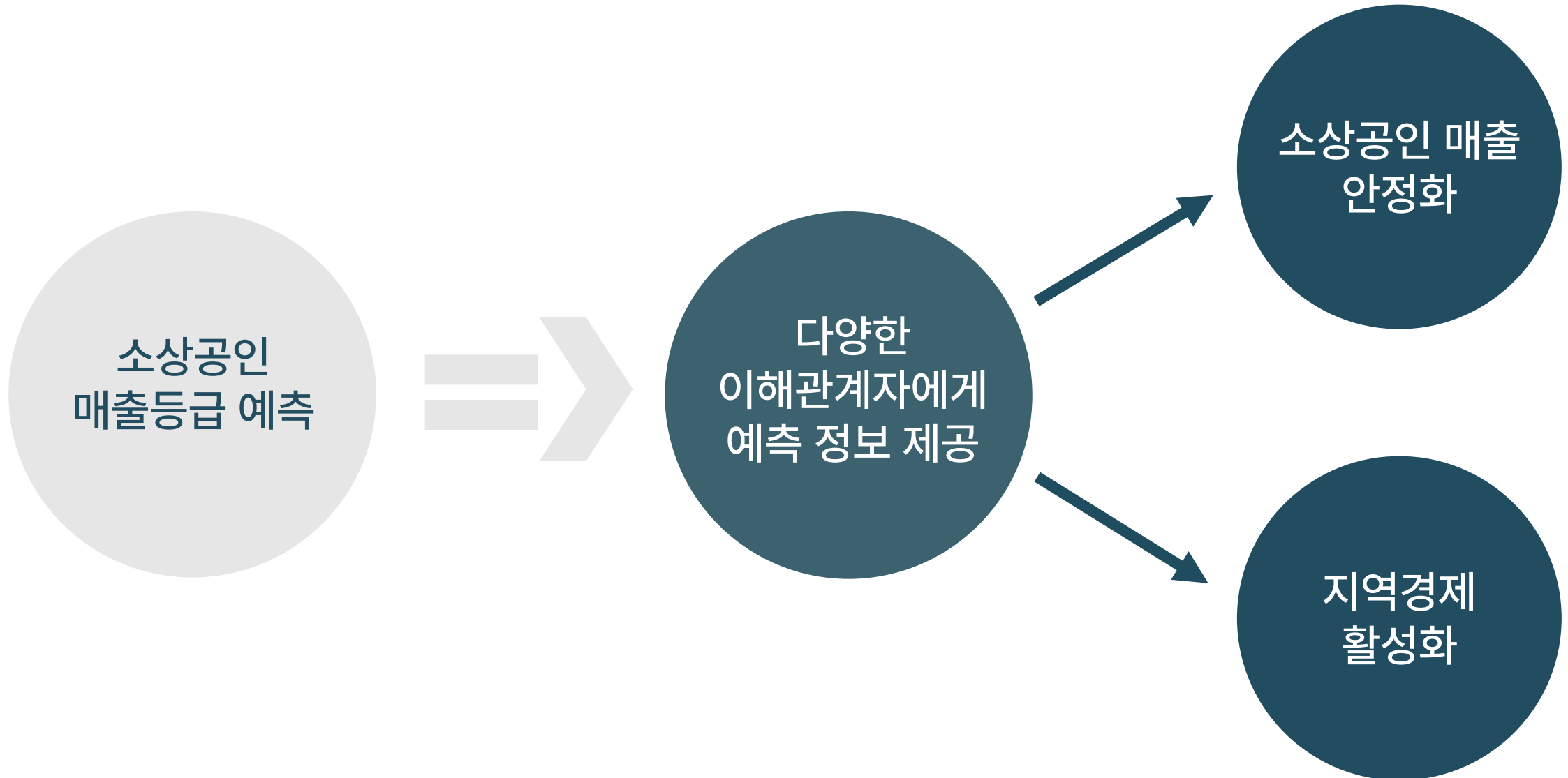


- OECD 회원국 중에서 자영업자 비중이 평균 15.9%보다 약 1.5배 높은 24.6%로 경쟁 치열
- 서울시 소상공인
: 전체 소상공인 사업체수의 21.6% 차지
→ 지역 경제에 큰 영향 끼침



특히 요식업과 같이 진입장벽이 낮아
창업/폐업 비율이 높은 업종에서는 정확한 매출 진단 필요

분석 목적



소상공인의 경쟁력 강화

정확한 매출등급의 예측은 소상공인의 경영에 큰 도움이 되며, 경쟁력을 강화할 수 있음

부동산 투자의 효율성 향상

투자 위치를 결정할 때, 소상공인의 매출등급을 고려하여 효율적인 투자 결정을 내릴 수 있음

지역경제 활성화

지역경제에 중요한 역할을 하는 소상공인에게 필요한 것들을 지원함으로써 지역경제를 활성화 시킬 수 있음

새로운 사업 시작 지원

새롭게 사업을 시작하는 소상공인이 적합한 위치와 업종을 선택할 수 있도록 지원할 수 있음

2

분석과정



>>



>>



>>



- 데이터 가공 및 결측값 처리
- Scaling

- 업종전체.업종 구분 선택
- 모델 튜닝 및 최종 모델 선택

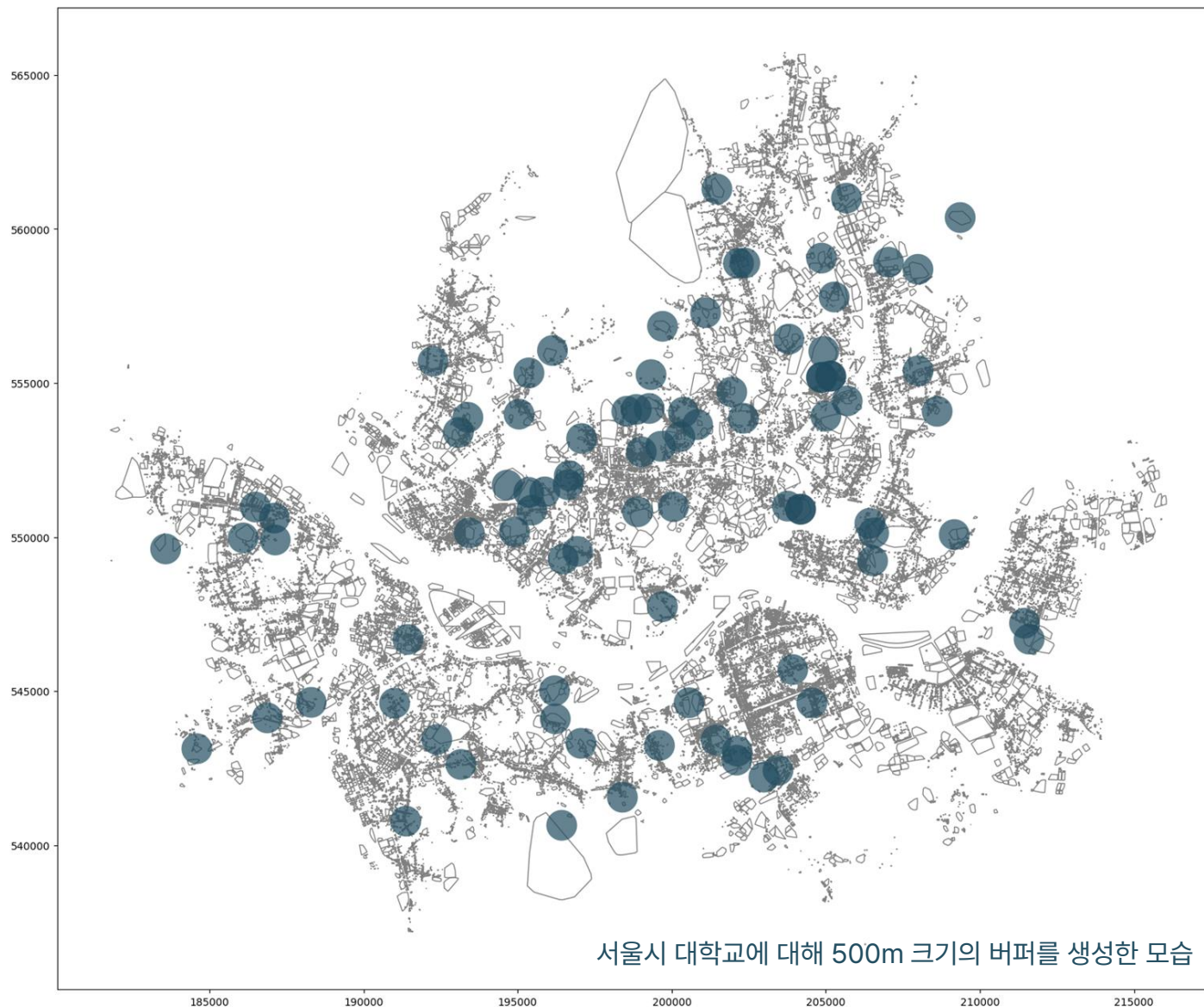
출처	원본 데이터	사용한 변수
서울시 상권분석 서비스 - 지역.상권 별 현황	점포수	전체점포수
		프랜차이즈점포수
		일반점포수
	개폐업수(률)	개업수
		폐업수
		개업률
		폐업률
	인구수	길단위유동인구
		주거인구
		직장인구
	임대시세	전체 임대료
		1층 임대료
		1층 외 임대료

출처	원본 데이터	사용한 변수
부동산 빅데이터 플랫폼	부동산거래량(금액) 대비 유동인구	부동산거래대비유동인구
	젠트리피케이션 위험지수	젠트리피케이션
	공실률 대비 매매가 임대료	공실률대비매매가임대료
서울특별시 빅데이터 캠퍼스	수도권 지하철 공간데이터	지하철개수
	서울시 대학교 공간데이터	대학교여부
스타벅스 홈페이지 크롤링	스타벅스 매장 위치	스타벅스 개수
서울 열린 데이터 광장	서울시 생활물가지수 통계	생활물가지수
	서울교통공사 연도별 일별 시간대별 역별 승하차 인원	출근시간_승차수
		출근시간_하차수
		주말_하차수
행정안전부	행정안전부_지역별(행정동) 성별 연 령별 주민등록 인구수	65세이상
		65세이상_남
		65세이상_여

지하철, 대학교 공간데이터

→ 500m 크기의 버퍼 생성

필지번호별로 포함하고 있는
버퍼의 개수를 count하거나,
버퍼를 포함하고 있는지를 계산



필지번호 - 행정동명 매칭

점포수, 개폐업수(률), 인구수, 임대시세, 65세이상 등의 변수

→ 행정동을 기준으로 집계된 데이터

'소상공인 매출등급 예상 대상 필지'에 수집한 변수들을 추가하기 위해서는 해당 필지에 대한 행정동명 정보가 필요

법정동 코드				
DATA_CRTR_Y M	LT_UNQ_NO	STDG_EMD_C D	INDUSTRY_CD	SLS_GRD
202201	1111010100100030100	11110101	A03	4
202201	1156011100104460000	11560111	A01	5
202201	1156011100104450001	11560111	A03	3
202201	1156011100104390000	11560111	A04	4
202201	1156011100104340000	11560111	A01	5

필지번호 - 행정동명 매칭

QGIS를 이용하여 서울시 필지번호와 행정동명 매칭

			행정동 이름
PNU	SGG_CD	SGG_NM	ADM_NM
1111010100100010000	11110	종로구	청운효자동
1111010100100010001	11110	종로구	청운효자동
1111010100100010002	11110	종로구	청운효자동
1111010100100010003	11110	종로구	청운효자동
1111010100100010004	11110	종로구	청운효자동

수집한 변수 추가

매출등급 예측 모델링을 위해 raw 데이터에 수집한 변수들을 추가함

시점 기준 매칭

생활물가지수

행정동명 기준 매칭

전체점포수, 프랜차이즈점포수, 일반점포수, 개업수, 폐업수, 개업률, 폐업률,
길단위유동인구, 주거인구, 직장인구, 전체 임대료, 1층 임대료, 1층 외 임대료,
65세이상, 65세이상_남, 65세이상_여

필지번호 기준 매칭

부동산거래대비유동인구, 젠트리피케이션, 공실률대비매매가임대료, 지하철개수,
대학교여부, 스타벅스개수, 출근시간_승차수, 출근시간_하차수, 주말_하차수

결측값 처리

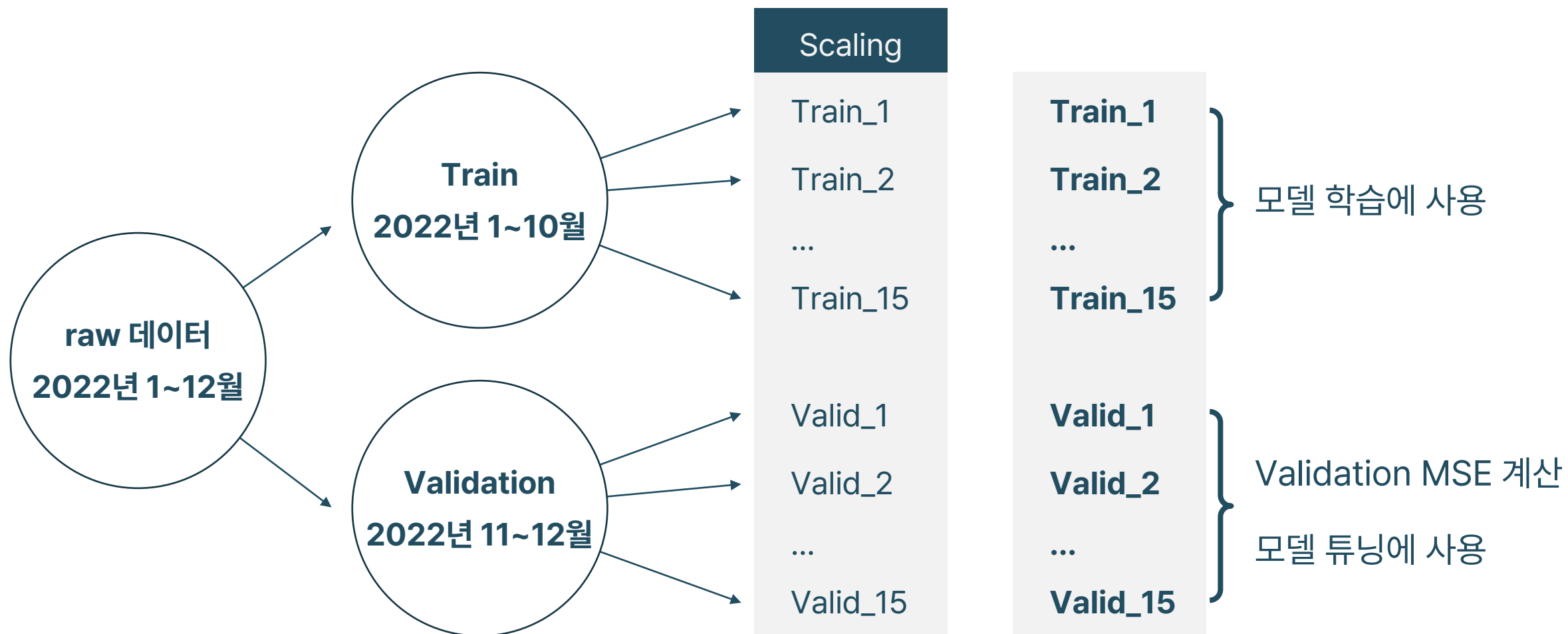
= 법정동 코드

PNU	필지번호 앞 10자리
1111010100100010000	1111010100
1111010100100010001	1111010100
1111010100100010002	1111010100
1111010100100010003	1111010100
1111010100100010004	1111010100

같은 법정동 코드를 갖는

다른 필지들의 **평균값**으로 대체

Train/validation data 분리 및 Scaling



하나의 모델/업종별 모델 결정

업종 전체

KMeans
그룹화A01
A03
그 외 업종

업종별

Random
ForestRandom
Forest

Extra Tree

Random
Forest

Extra Tree

select

업종 전체	군집화	업종별
0.6821	0.6380	0.6323

변수 선택



" 높은 상관관계를 갖는 변수 중 상대적으로 중요도가 낮은 변수를 제거 "

분류모델 vs. 회귀모델 -> 더 작은 mse를 갖는 모델로 select

LT_UNQ_NO	True SLS_GRD	Predicted SLS_GRD
1111010100100030100	4	4
1156011100104460000	5	5
1156011100104450001	3	2
1156011100104390000	4	4
1156011100104340000	5	5

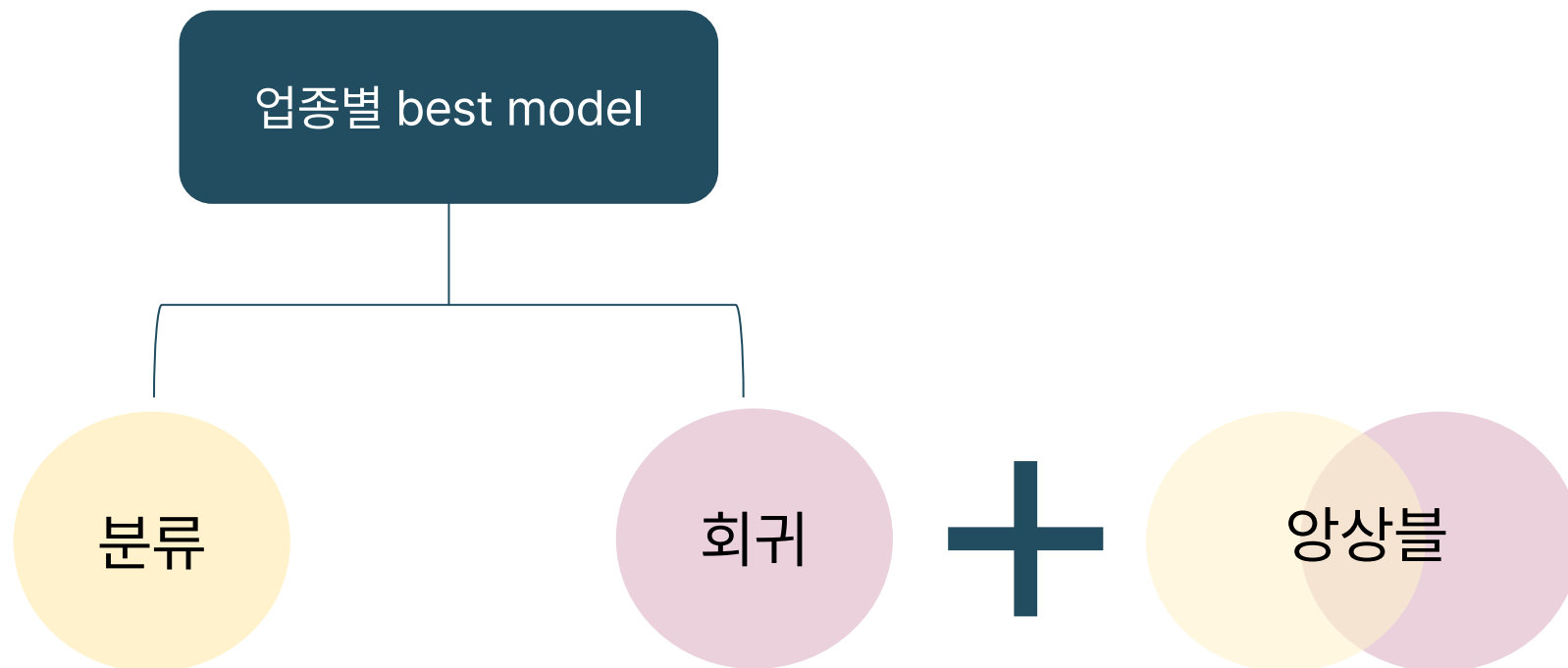
LT_UNQ_NO	True SLS_GRD	round Predicted SLS_GRD	Predicted SLS_GRD
1111010100100030100	4	4	3.5
1156011100104460000	5	5	4.9
1156011100104450001	3	2	2.3
1156011100104390000	4	4	4.1
1156011100104340000	5	5	4.6



반올림

최종 모델 선택

GridsearchCV를 이용하여 파라미터 튜닝 진행 후, 업종별로 가장 낮은 MSE를 갖는 최종 모델 선택



최종 모델 선택

업종	모델 종류	세부 모델
A01	회귀	ExtraTreesRegressor()
A02	분류	ExtraTreesClassifier()
A03	회귀	ExtraTreesRegressor()
A04	앙상블 (분류+회귀)	ExtraTreesClassifier() + ExtraTreesRegressor()
A05	회귀	ExtraTreesRegressor()
A06	앙상블 (분류+회귀)	ExtraTreesClassifier() + ExtraTreesRegressor()
A07	앙상블 (분류+회귀)	ExtraTreesClassifier() + ExtraTreesRegressor()
A08	회귀	ExtraTreesRegressor()
A09	회귀	ExtraTreesRegressor()

업종	모델 종류	세부 모델
A10	회귀	RandomForestRegressor()
A11	회귀	ExtraTreesRegressor()
A12	회귀	LGBMRegressor()
A13	앙상블 (분류+회귀)	ExtraTreesClassifier() + ExtraTreesRegressor()
A14	앙상블 (분류+회귀)	ExtraTreesRegressor()
A15	앙상블 (분류+회귀)	RandomForestClassifier()

업종	튜닝 전	튜닝 후
Total	0.6323	0.5665

3

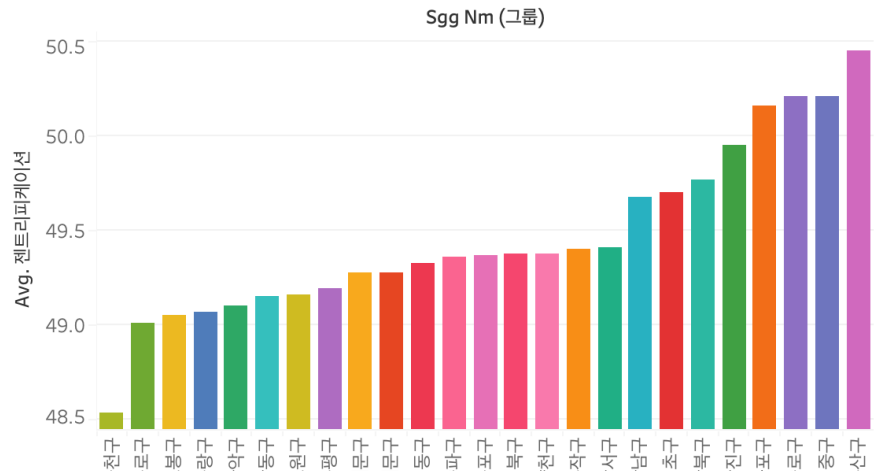
분석 결과

Part 3

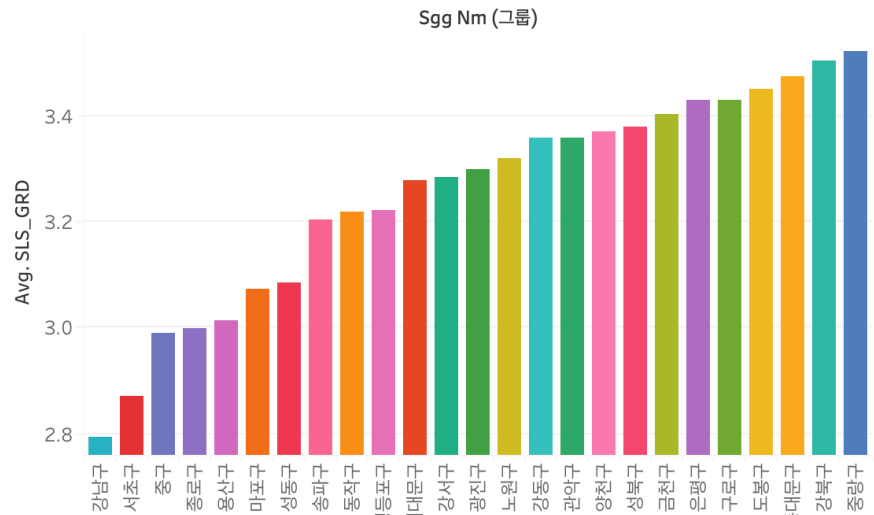
분석 결과

"업종 전체"에 유의미한 변수 추출

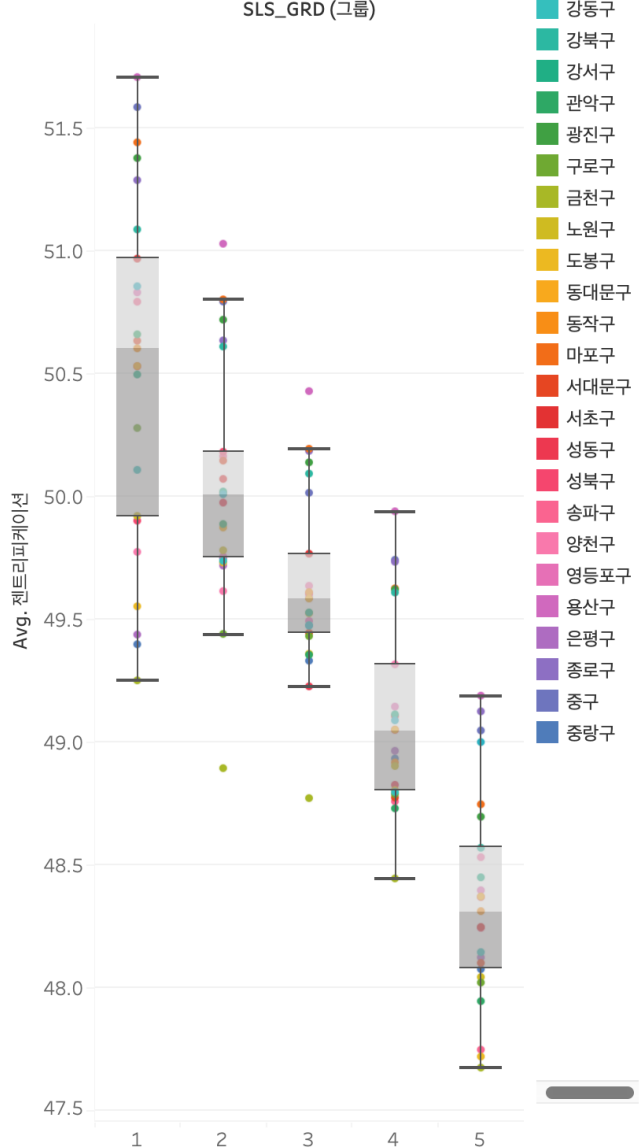
<구별 젠트리피케이션 위험지수 평균>



<구별 등급 평균>

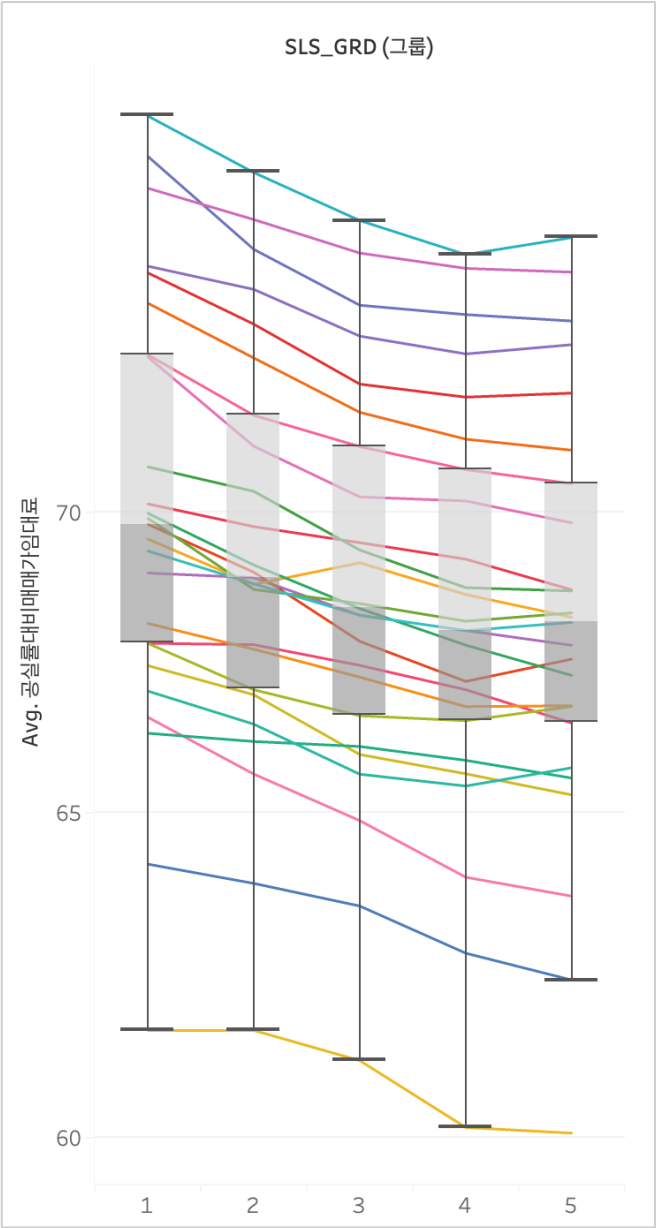
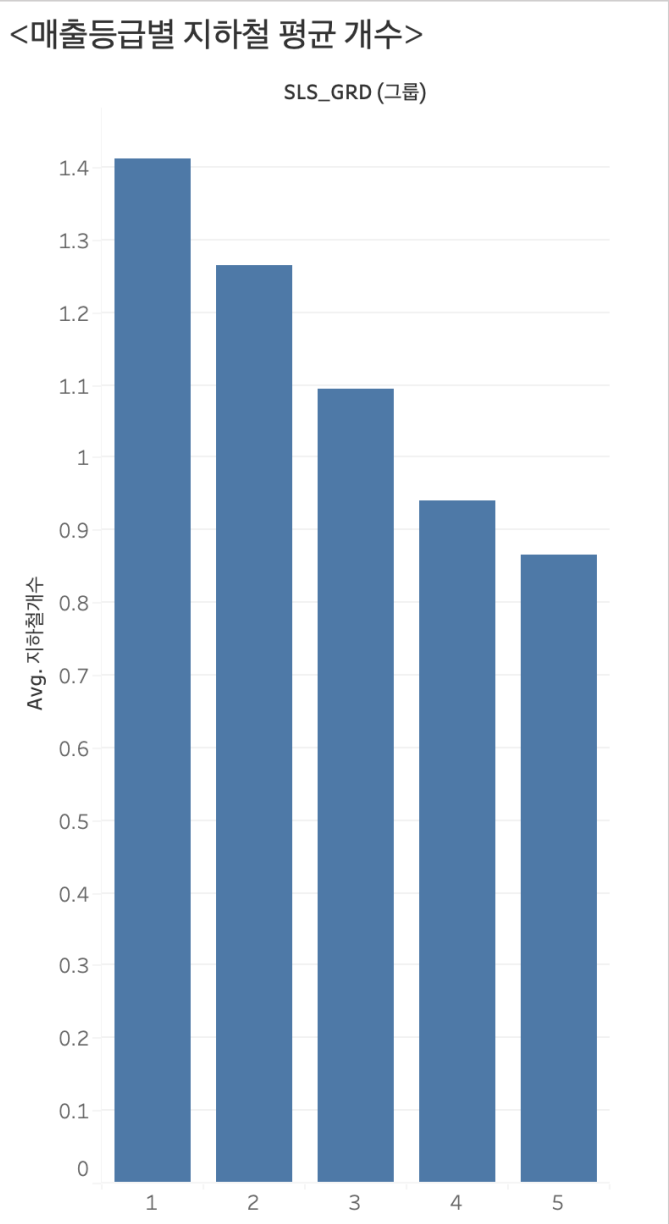


<등급별 젠트리피케이션 위험지수 평균 Boxplot>



젠트리피케이션 ↑
매출등급 ↑

분석 결과 "업종 전체"에 유의미한 변수 추출

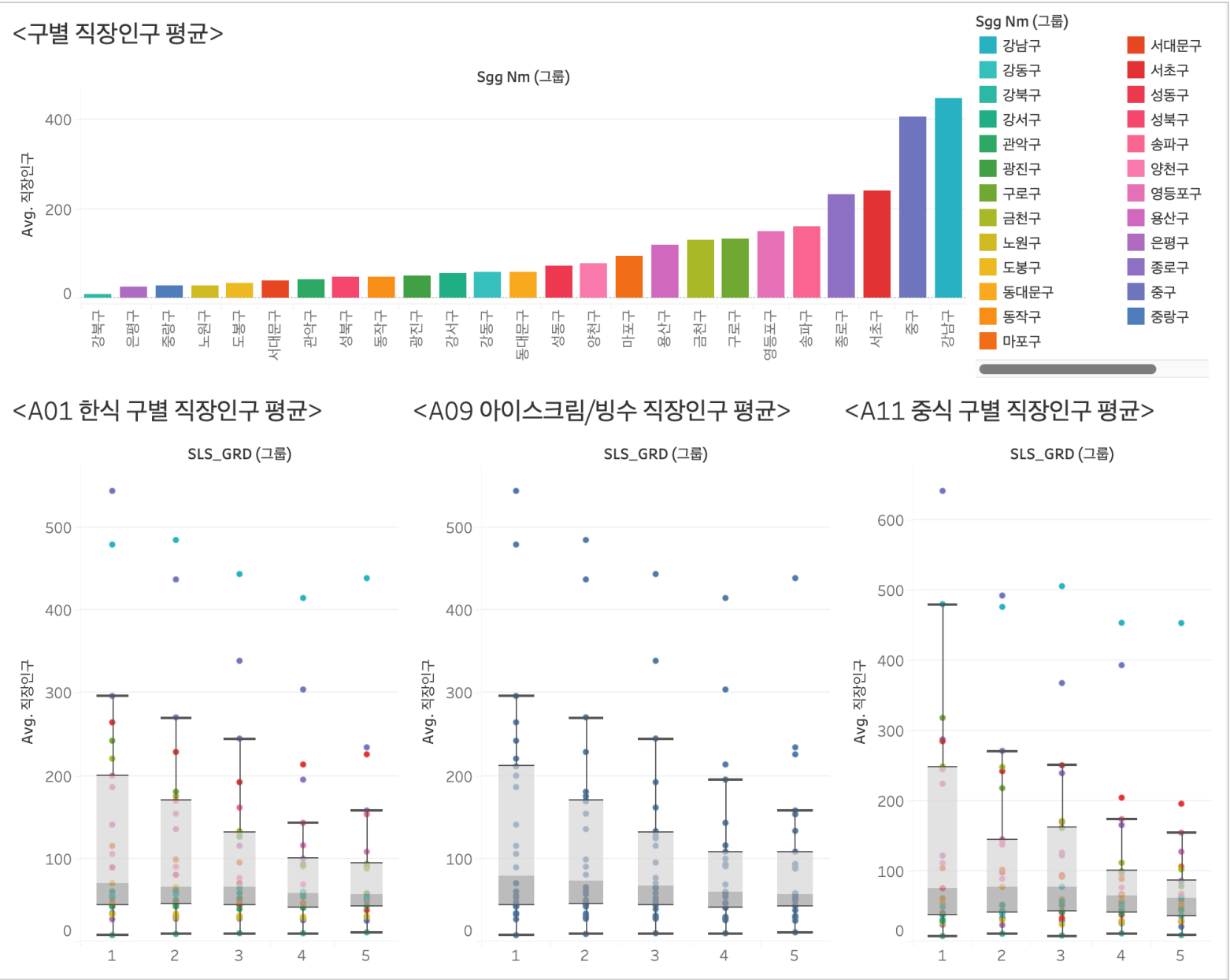


지하철 개수 ↑,
공실률대비매매가임대료 ↑
매출등급 ↑

Part 3

분석 결과

“업종별” 유의미한 변수 추출

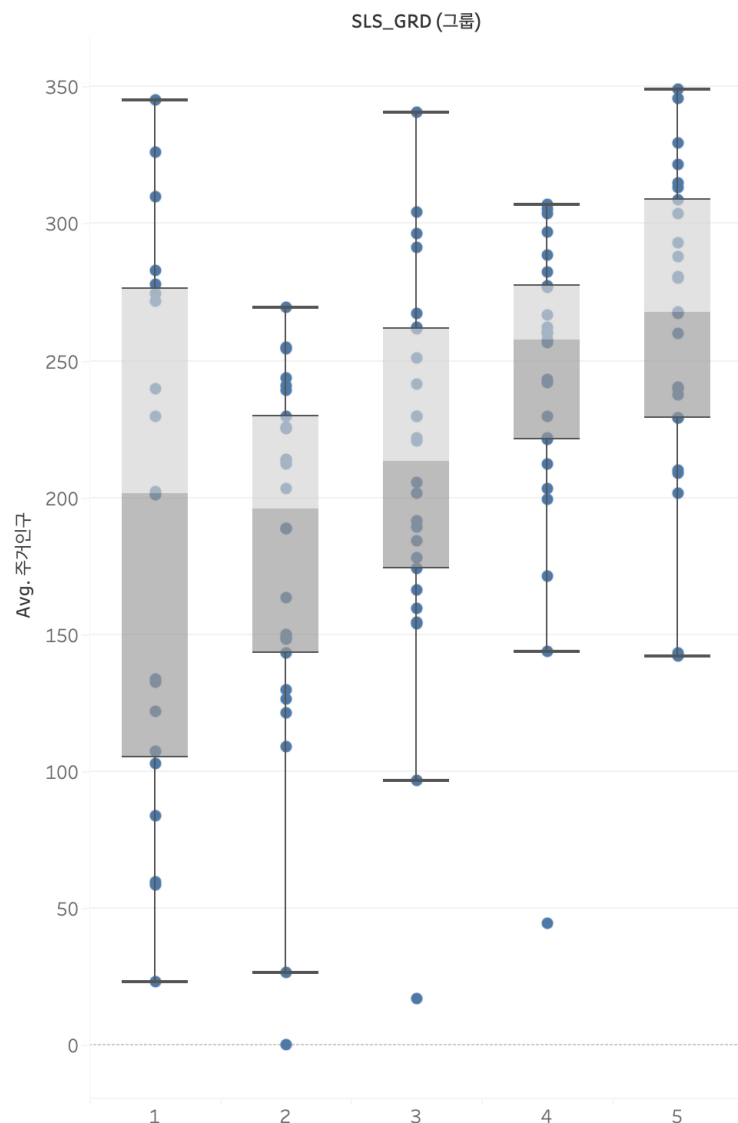


[A01] , [A09] , [A11]

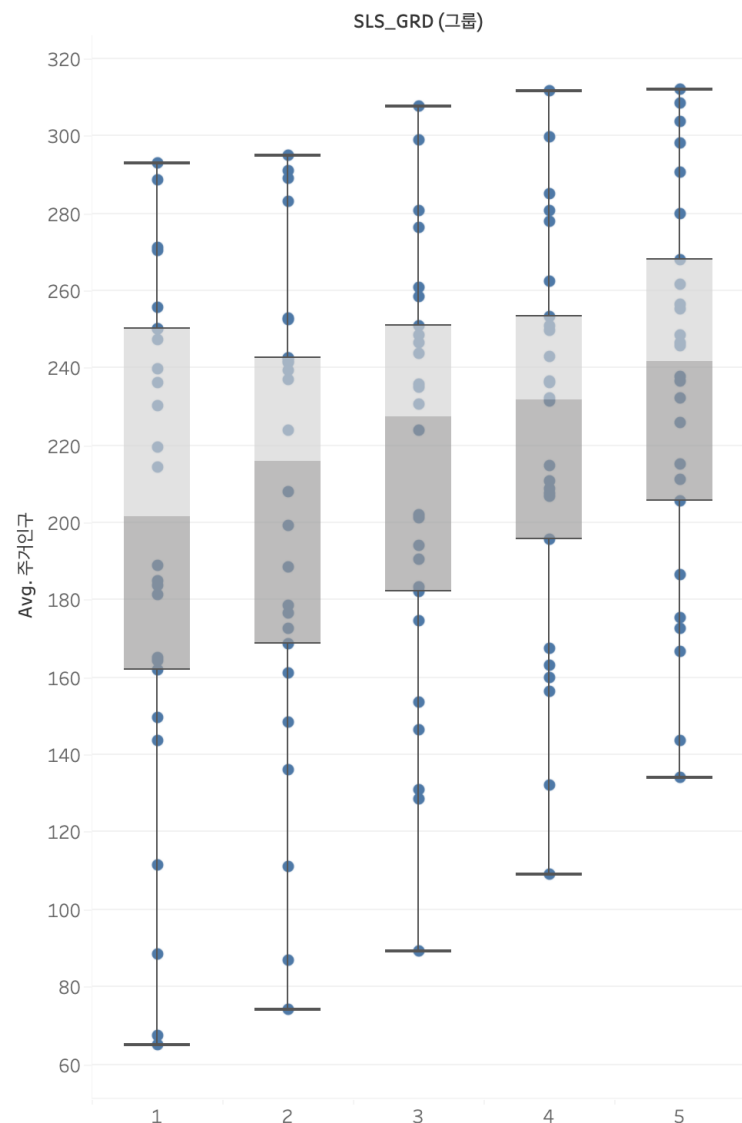
직장인구 ↑
매출등급 ↑

분석 결과 “업종별” 유의미한 변수 추출

<A10 기타외국식 주거인구 평균>



<A15 음료(커피 외) 주거인구 평균>



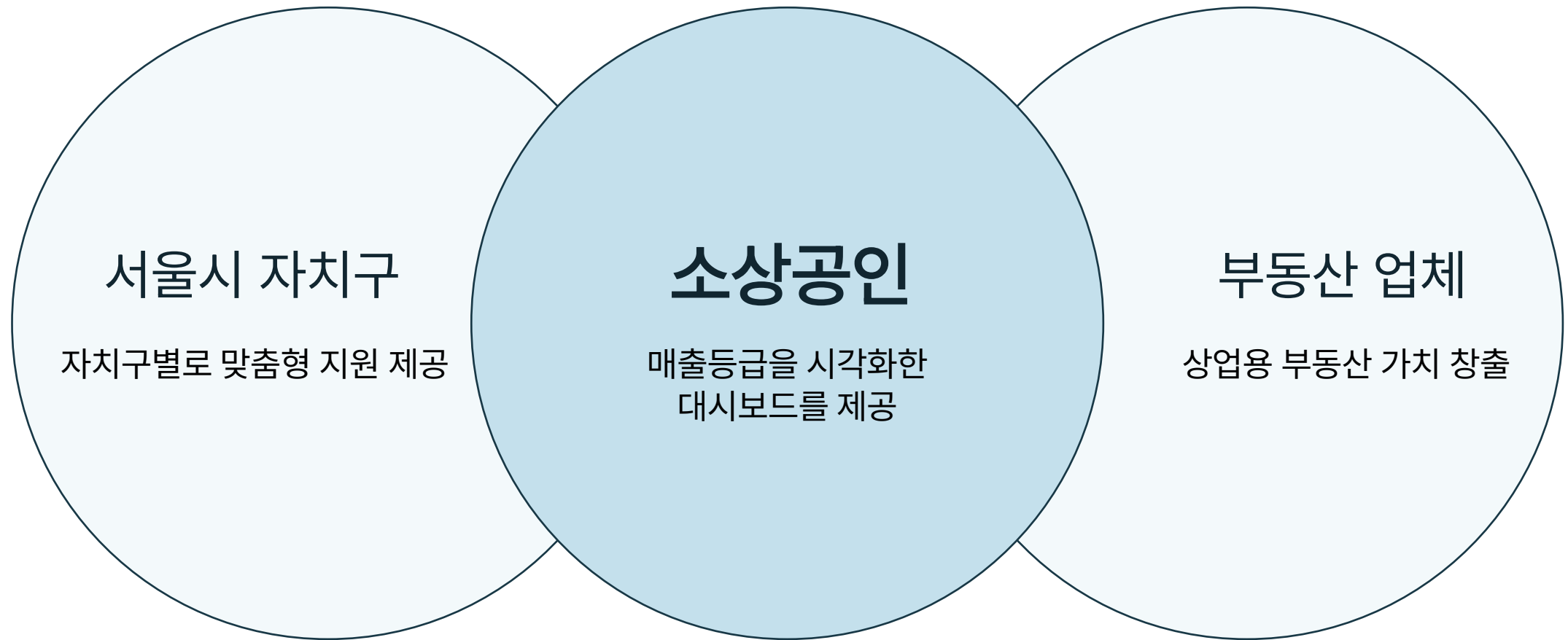
[A10] , [A15]

주거인구 ↑

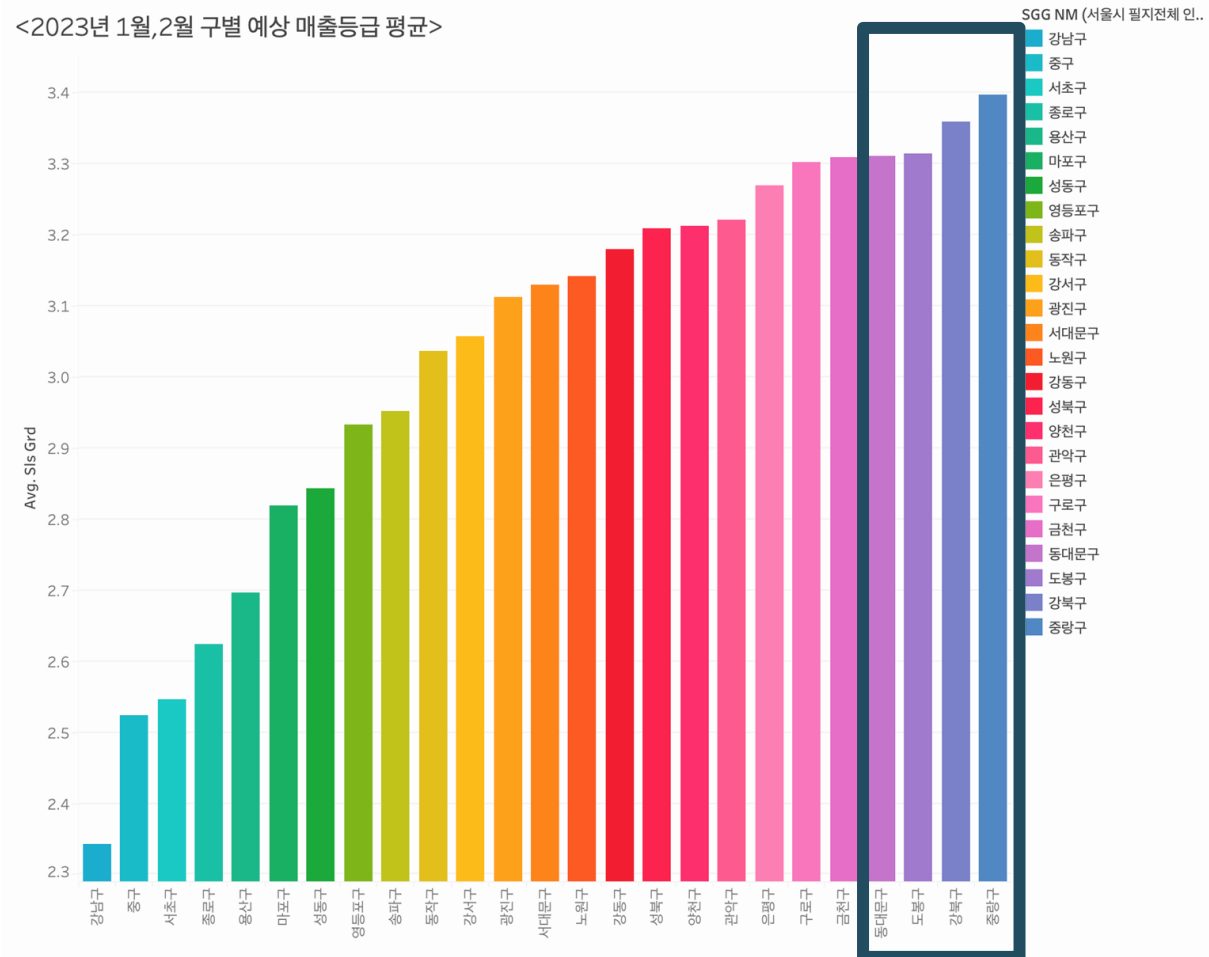
매출등급 ↓

4

결과 활용 방안



매출등급이 낮은 지역구에 대해 적절한 정책 시행



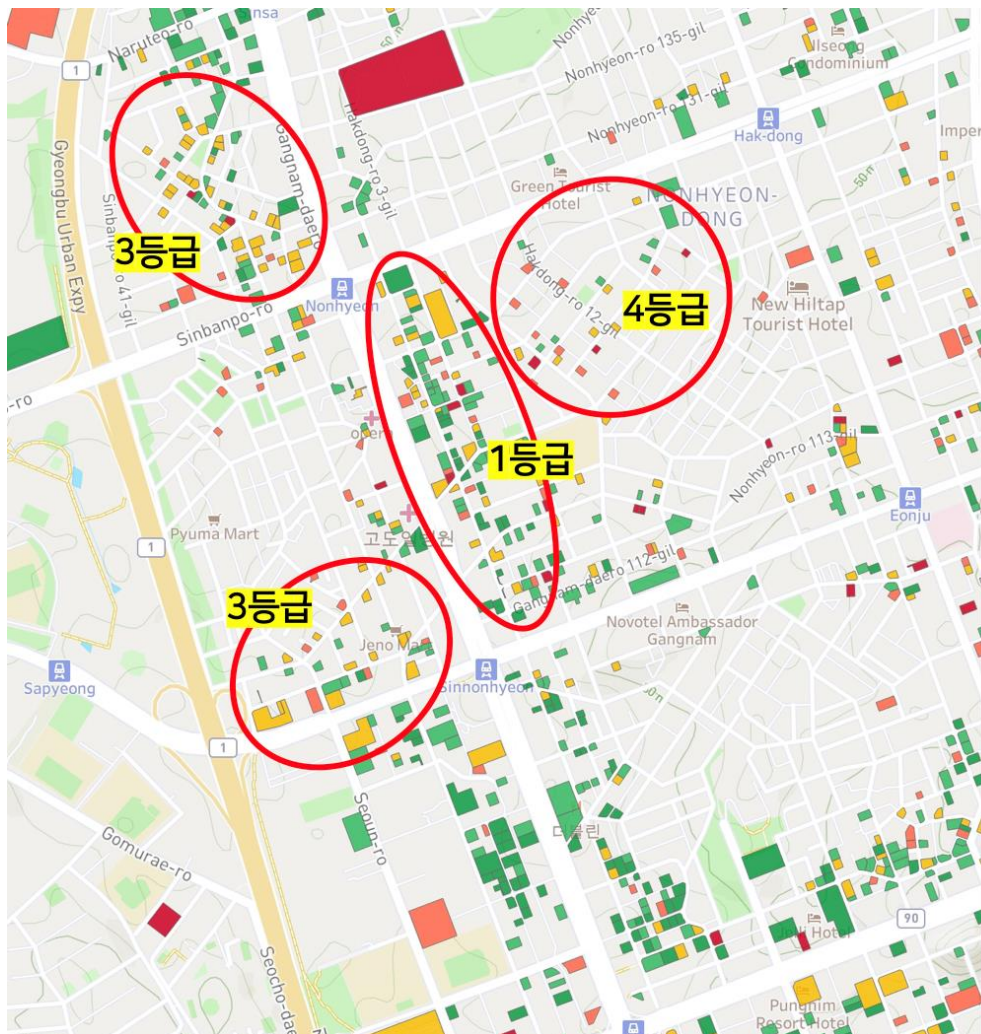
저금리 대출과 자금 지원

적극적인 자금 정책을 시행함으로써 지역 활성화
촉진 및 소상공인의 경제 활동 지원

청년 소상공인 지원

매출등급과 함께 젠트리피케이션 위험지수도 낮은
지역에서 청년 소상공인을 위한 정책을 시행함으로
써 일자리 창출 및 지역의 경제 향상

대시보드를 통해 소상공인의 적절한 의사결정 지원



지역별 매출등급 시각화 대시보드 제공

소상공인들이 사업 시작 전 적절한 점포 위치를
선택 할 수 있도록 지원

업종별 위치 인사이트 제공

특정 업종이 어느 변화가에서 가장 성공적일지에
대한 인사이트를 제공

A14 서양식

Data Crtr Ym (그룹)

- 202208
- 202209
- 202210
- 202211
- 202212
- 202301
- 202302

Sls Grd (그룹)

- 1
- 2
- 3
- 4
- 5

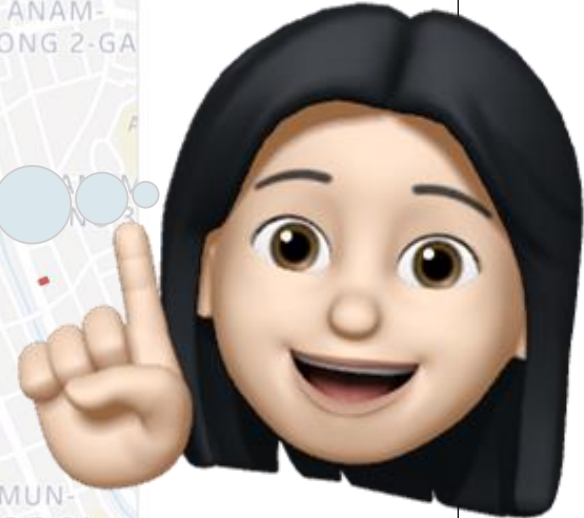
Industry Cd (그룹)

A14



양식집을
차리려고 하는데
A, B중에
어디가 더 좋을까?

난 **A**를 추천할게!
왜냐면 여기 지도를 보면
B보다 A의 양식집
매출등급이 높게
나타나고 있어



23년 1월, 2월 전체 업종 매출등급 예측



필지별 예상 매출등급 기반으로 부동산 투자에 대한 의사결정을 지원함으로써 투자의 효율성 최적화

THANK
YOU!

