



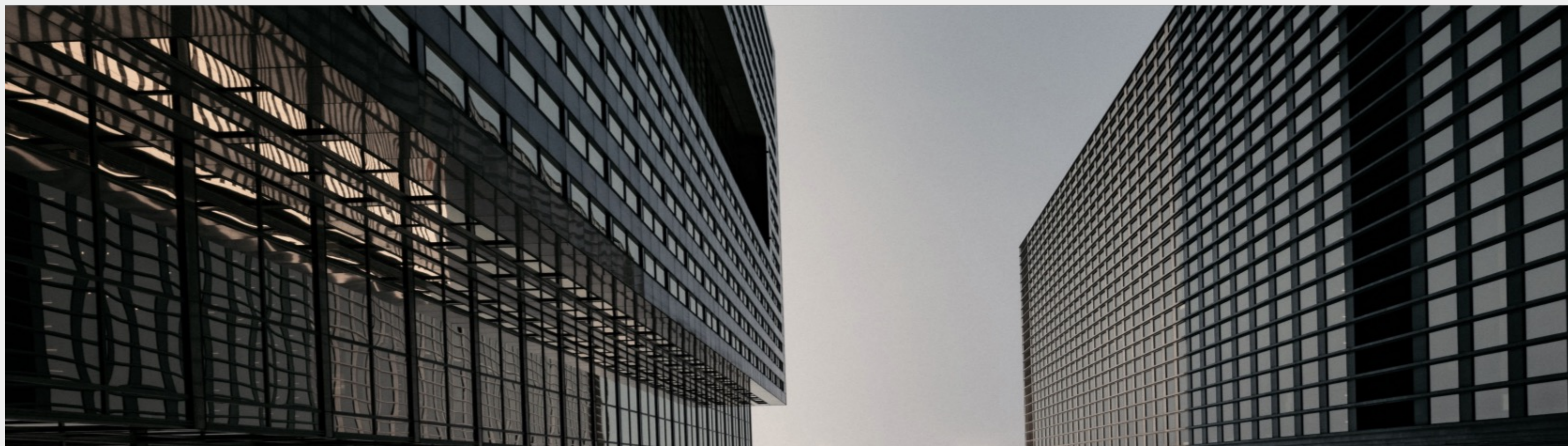
고려대학교
KOREA UNIVERSITY



CCP 13회 기획주제

지역 소득 수준에 영향 받는 업종 탐색

에코노미팀(박무성, 전지우, 배은지, 원윤정)





목차

01 팀 소개

팀명, 팀원 소개

02 연구 소개

연구 목적

데이터 소개

데이터 시각화 및 전처리

03 연구 과정

분석 방법 소개

회귀분석

비모수 검정

04 연구 결과

연구 결과

활용 방안






팀 소개

에코노미(Economy)

데이터 사이언스 학회(KUBIG) 학회원들로 구성된 프로젝트 팀

 팀장

박무성

사회학과 16

데이터 추출,

통계 분석,

기업 컨택



전지우

통계학과 18

데이터 추출,

통계 분석,

시각화



배은지

심리학과 18

데이터 추출,

통계 분석,

가설 검정



원윤정

수학과 19

데이터 추출,

통계 분석,

피피티 제작 및 발표





연구 목적

연구 주제

: 지역 소득 수준에 영향 받는 업종 탐색

주제 선정 이유 및 필요성

자영업자



지역의 소득수준은 매출에 영향을 미치지만
자영업자 스스로 분석할 수 없는 영역임

정부



소상공인 상권분석 지원 서비스의
정확성을 높일 수 있음

기업



데이터 판매 및 사적 상권분석
서비스의 정확도를 높여 이윤 창출 가능



데이터 소개

서울, 경기의 구별 카드사 지역소득수준 데이터 & 구별 매출 데이터 (소분류 업종 코드별)

데이터 선정 이유

REASON 1

부동산 가격 폭등, 건강보험료의 상/하위 가중치 차이



카드사 데이터 사용

REASON 2

우편번호 기준으로 소득분위 산정 후 시각화 결과,
구별 차이가 더 뚜렷



'구'별 소득수준 데이터 사용



데이터 설명

서울,경기의 구별 카드사 지역소득수준 데이터 & 구별 매출 데이터 (소분류 업종 코드별)

기준년월	신우편번호	출생년도	연령대코드	성별코드	개인소득금액
202011	0490*	1976	045	F	3,220
202011	0490*	1976	045	F	4,497
202011	0490*	1975	045	M	3,832

<지역별,연령별,성별, 소득구간 및 가구/개인 소득금액 통계자료>

데이터 설명 : 지역별 연령, 성별, 소득10분위수,
가구추청소득, 개인추정소득으로 통계화한 자료

➡ **지역 소득수준 데이터 추출** (21년 11월 기준)

집계기간 : 1개월

집계차원 : 기준년월, 지역별, 연령, 성별

집계항목 : 소득10분위수, 가구추정소득, 개인추정소득 (단위: 만원)

세분화 기준 : 소득10분위수(01.고소득~10.저소득)

기준년월	신우편번호	서비스소분 류업종코드	매출금액	매출건수	가맹점수
202009	2253*	LS0701	120,000	2	1
202009	0539*	ED0101	518,400	2	1
202009	5004*	SL0204	1,570,000	1,180	1

<지역별 업종별 매출상세 내역 및 유효/개·폐업 가맹점 수>

데이터 설명 : 전국 지역별(우편번호) 업종별 매출상세 내역,
유효/개·폐업 가맹점수

➡ **구별 매출 데이터 추출** (2020~2021년 2년치)

집계기간 : 1개월

집계차원 : 카드결제년월, 지역별(우편번호), 가맹점 업종별,
성별, 연령별, 시간대별, 요일별

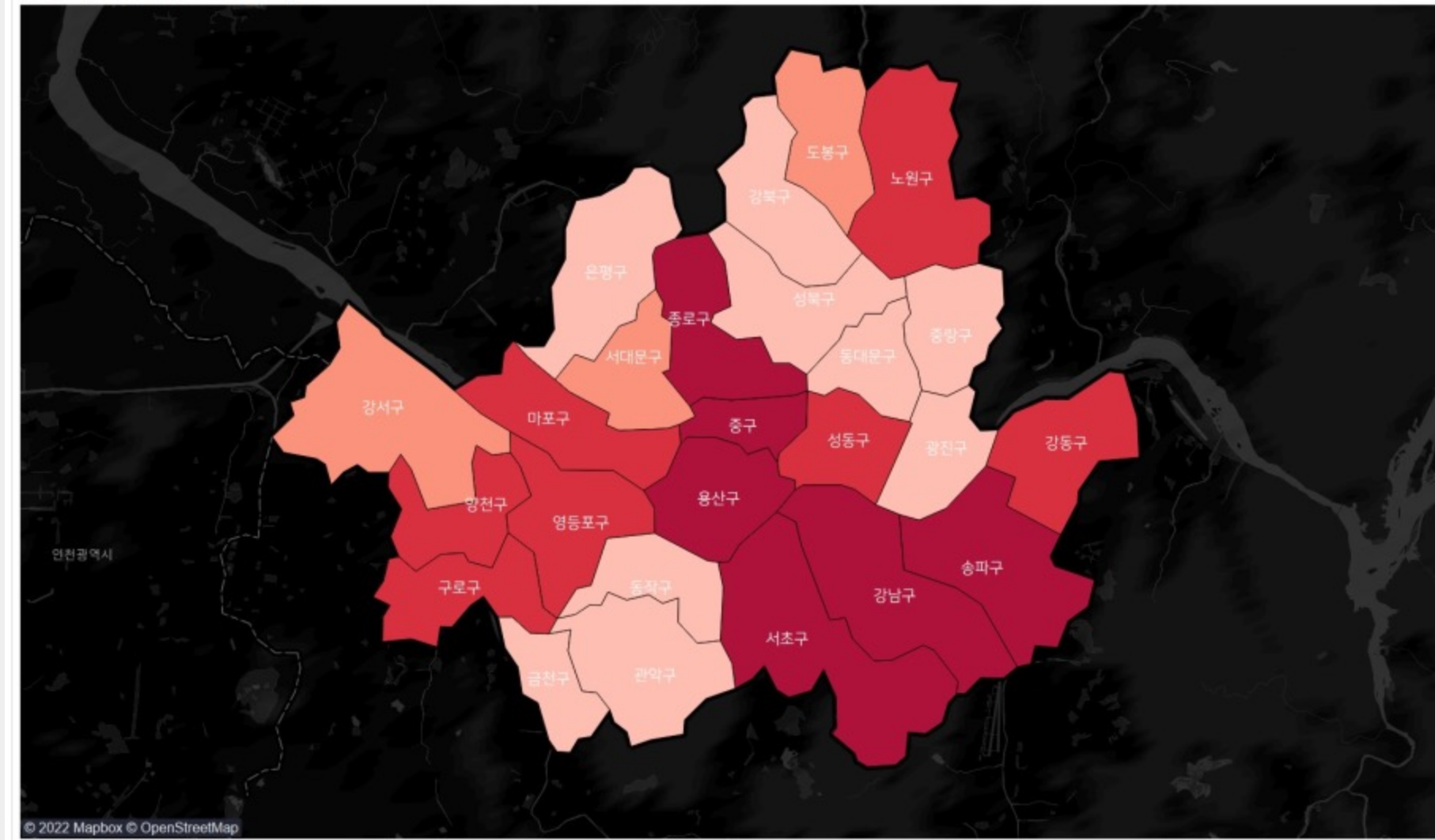
집계항목 : 매출금액, 유효 가맹점수, 개업 가맹점수,
폐업 가맹점수 (단위 : 원, 건)

세분화 기준 : 연령대(10,20,30,40,50,60대) x 성별, 시간대, 요일

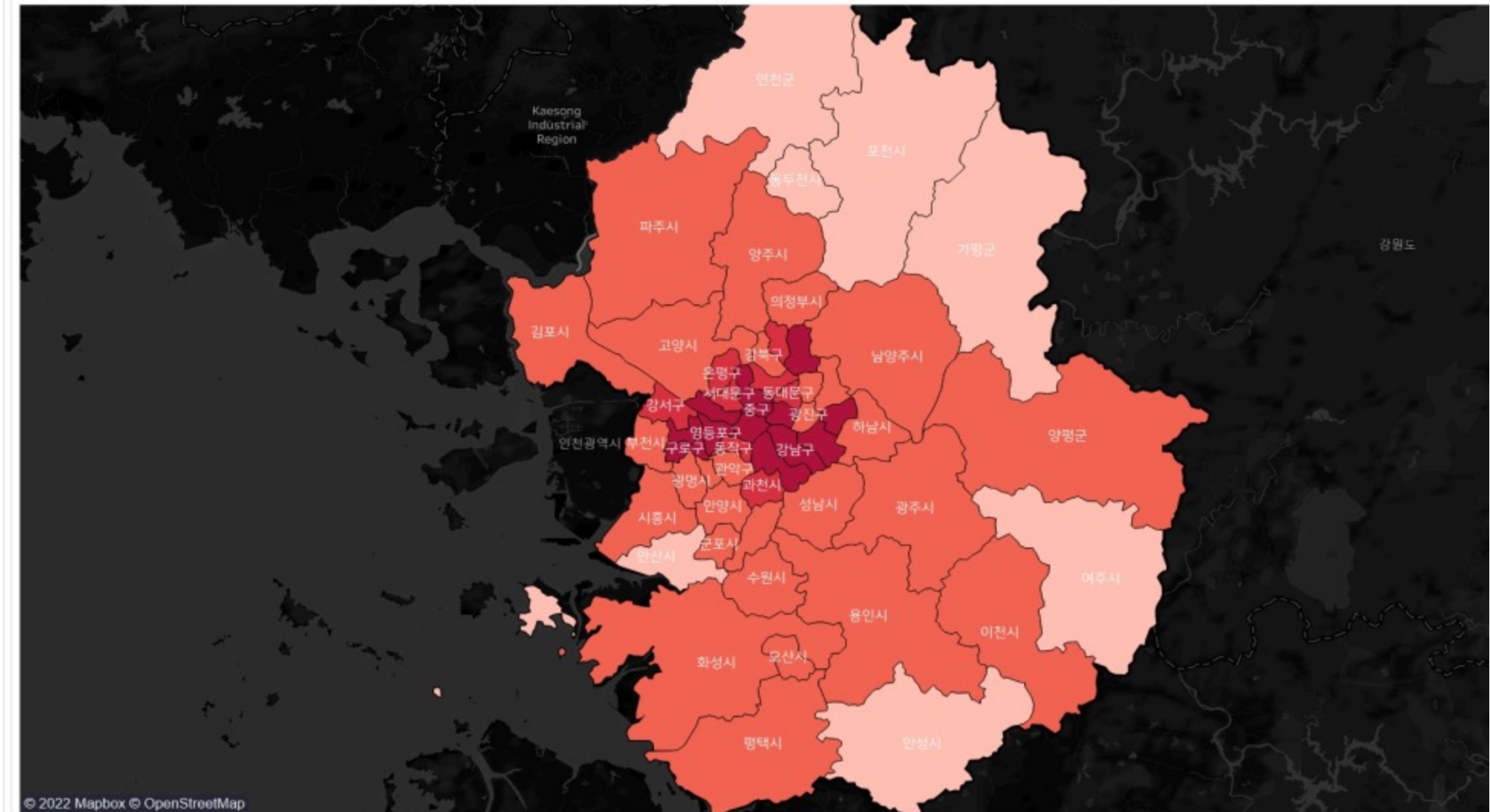


연소득 분포 시각화

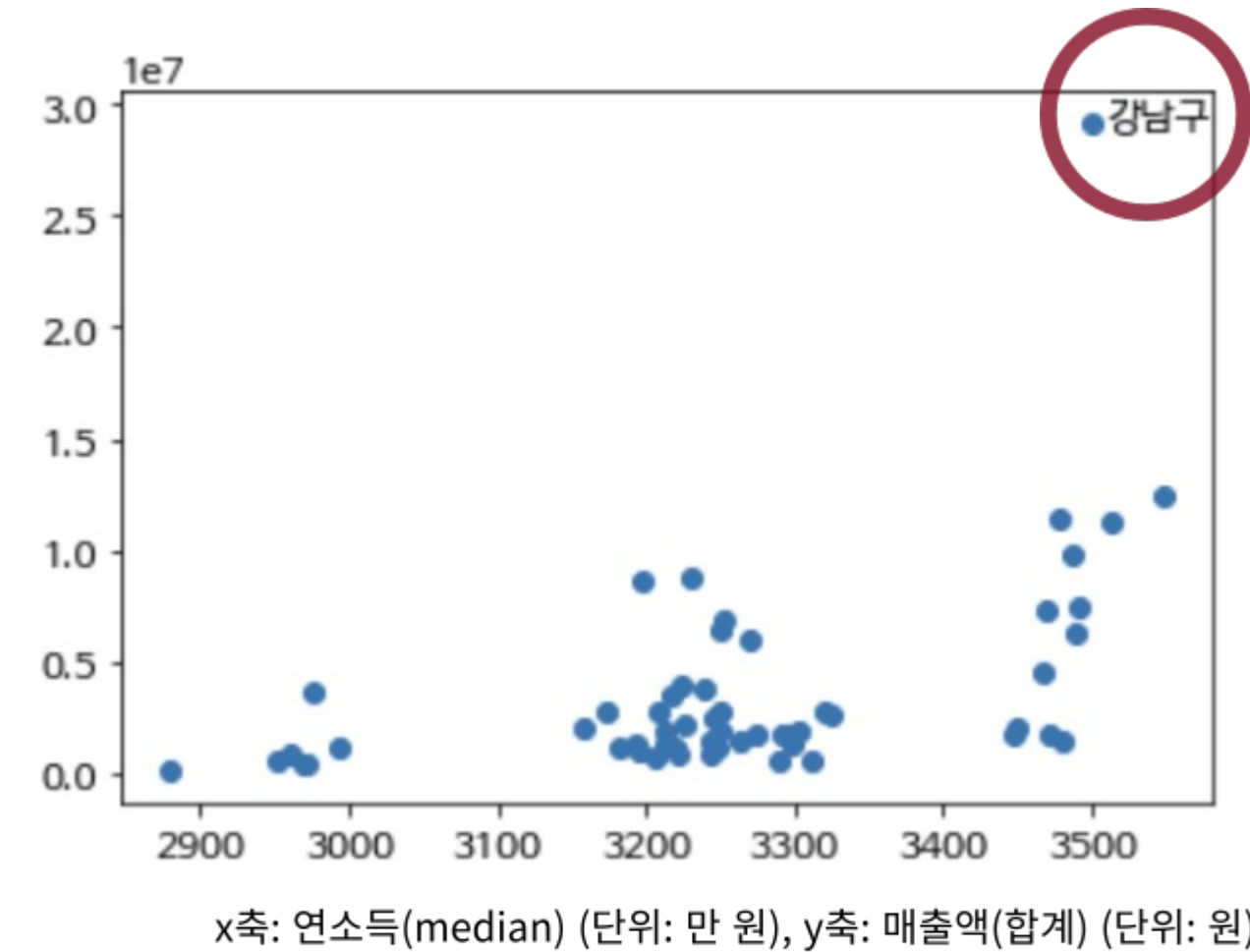
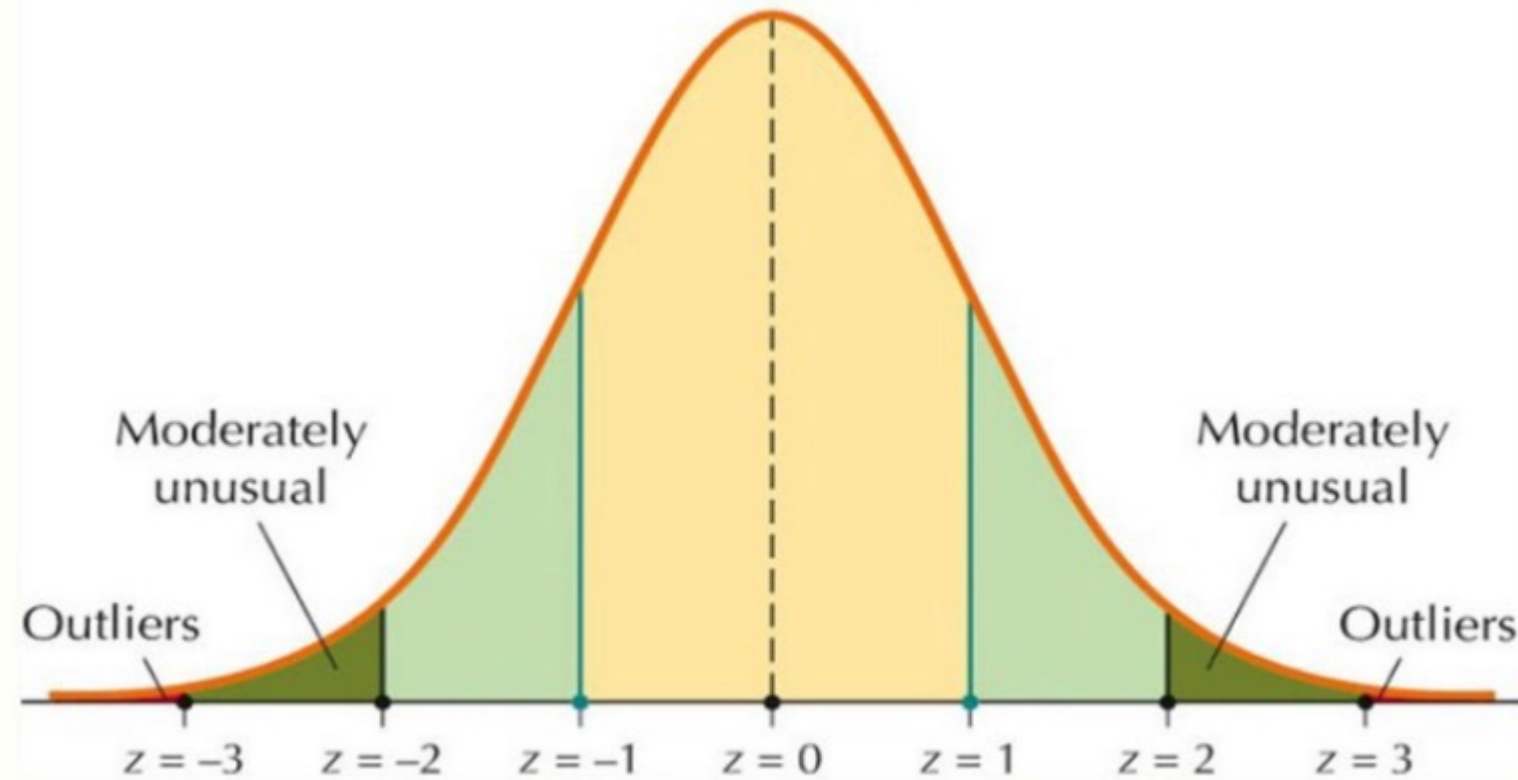
서울특별시 개인소득금액 분포



개인소득금액 분포



데이터 전처리



'경양식/레스토랑' 산점도에서 나타난 강남구 outlier 예시

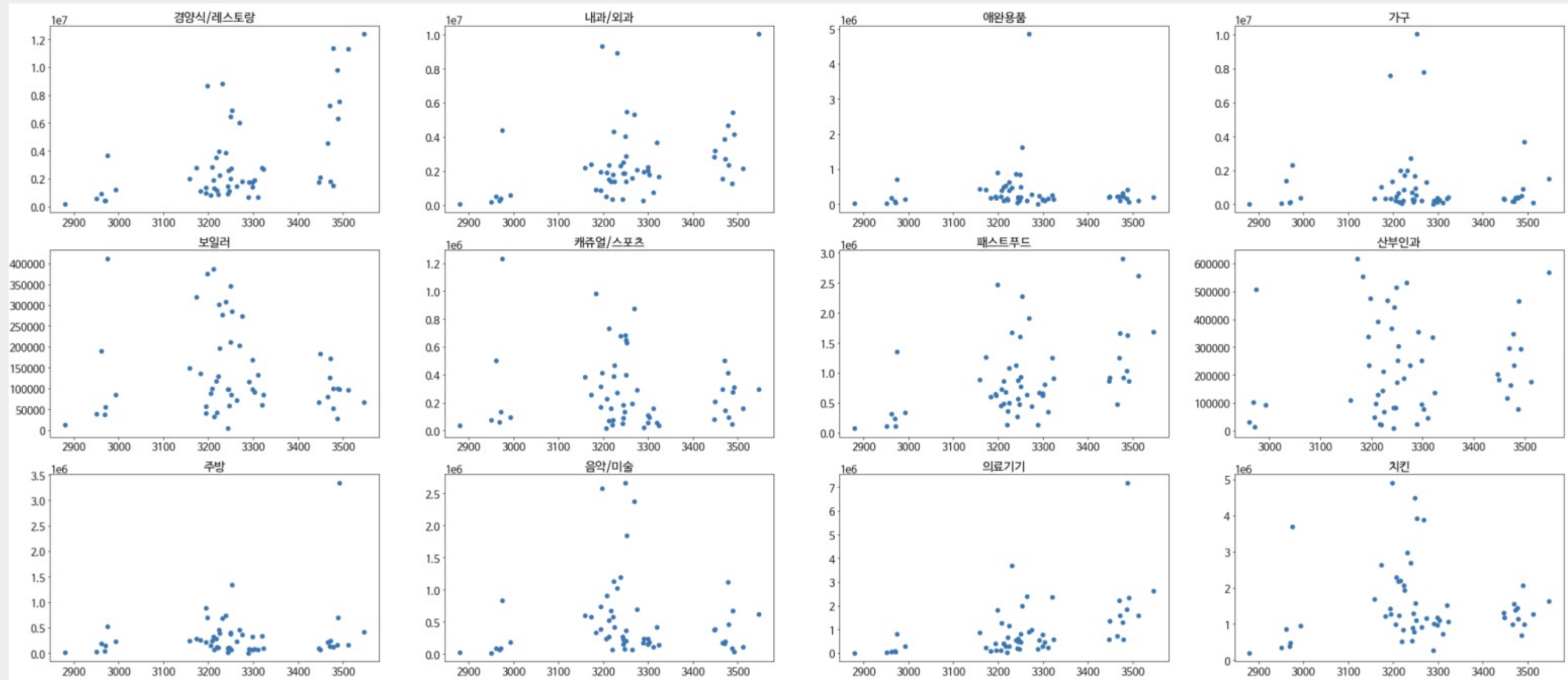
지역 소득수준 데이터를 '3표준편차' 기준으로 outlier 제거
 ➡ outlier 제거 후 소득수준 median 값을 대푯값으로

거의 모든 업종에 대해 '항상' 높은 매출을 나타내는 이상치
 ➡ 강남구 outlier 제거



업종별 산점도 확인

122개 서울(강남구 제외)/경기 지역별 연소득(median)에 따른 매출액(합계) 산점도



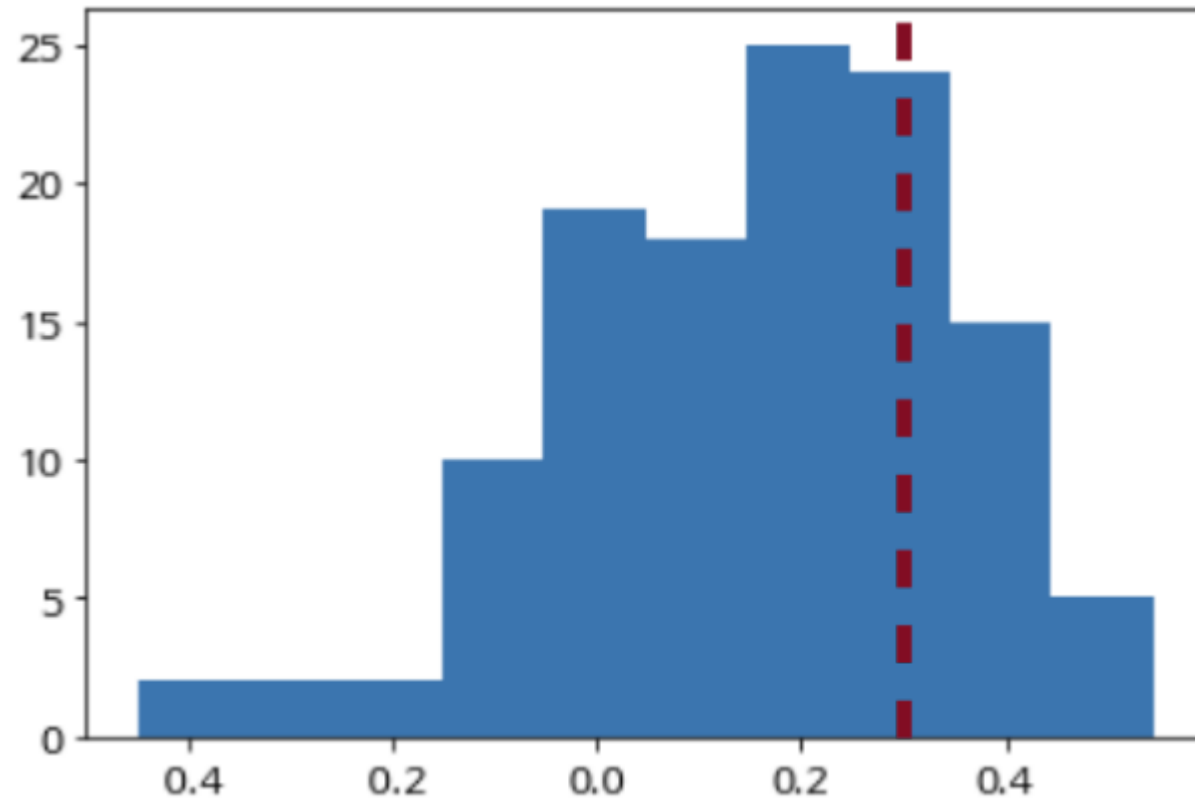
x축: 연소득(median) (단위: 만 원), y축: 매출액(합계) (단위: 원)



업종별 상관계수

업종별 매출액(합계)과 서울(강남구 제외)/경기 지역별 연소득(median)의 상관계수

상관계수 +0.3 이상의 업종 37개 추출



'경양식/레스토랑', '화장품', '의료기기', '패스트푸드',
'휴대폰', '스낵', '피부/비뇨기과', '남성의류', '관광', '주차장' 등

	업종명	상관계수
0	경양식/레스토랑	0.545296
1	화장품	0.500887
2	의료기기	0.485391
3	패스트푸드	0.469093
4	휴대폰	0.459941
...
32	건강보조식품	0.319494
33	중식당	0.305515
34	민속주점	0.305487
35	약국	0.303887
36	제과점	0.303435



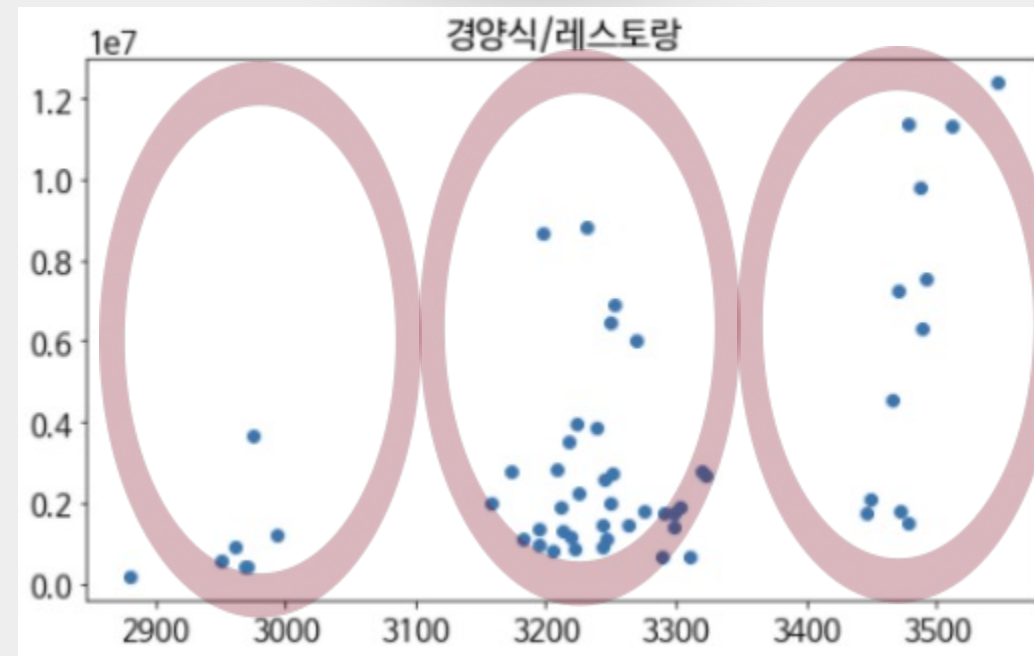
분석 방법

개별 연소득을
연속형 변수로



회귀분석

소득수준(연소득)에 따른
매출액 사이의
상관계수 0.3 이상
업종 37개



저/중/고소득
범주형 변수로



비모수 검정
(ANOVA 검정 代)



회귀분석

X : 지역별 연소득 중앙값

Y : 매출액 합계

회귀식: $Y = \beta_0 + \beta_1 X + \epsilon$

$H_0: \beta_1 = 0$ (지역별 연소득 중앙값은 매출액 합계에 영향을 주는 유의한 변수가 아니다.)

$H_1: \beta_1 \neq 0$ (지역별 연소득 중앙값은 매출액 합계에 영향을 주는 유의한 변수이다.)

	업종명	상관계수	F_pvalue	beta1_hat	beta1_pvalue	adj_R_squared
0	경양식/레스토랑	0.545296	0.000017	10906.270605	0.000017	0.284091
1	화장품	0.500887	0.000098	2200.894071	0.000098	0.236754
...
35	약국	0.303887	0.024098	6357.101351	0.024098	0.075222
36	제과점	0.303435	0.024320	2436.193204	0.024320	0.074942

<단순 회귀분석 결과>

1. 37개 모든 업종의 p-value가 0.03 미만으로
유의하게 나타남
2. R_squared 값은 0.28에서 0.07까지 다양함



회귀분석 가정(오차의 정규성, 등분산성) 만족하지 않는 업종 제외하기

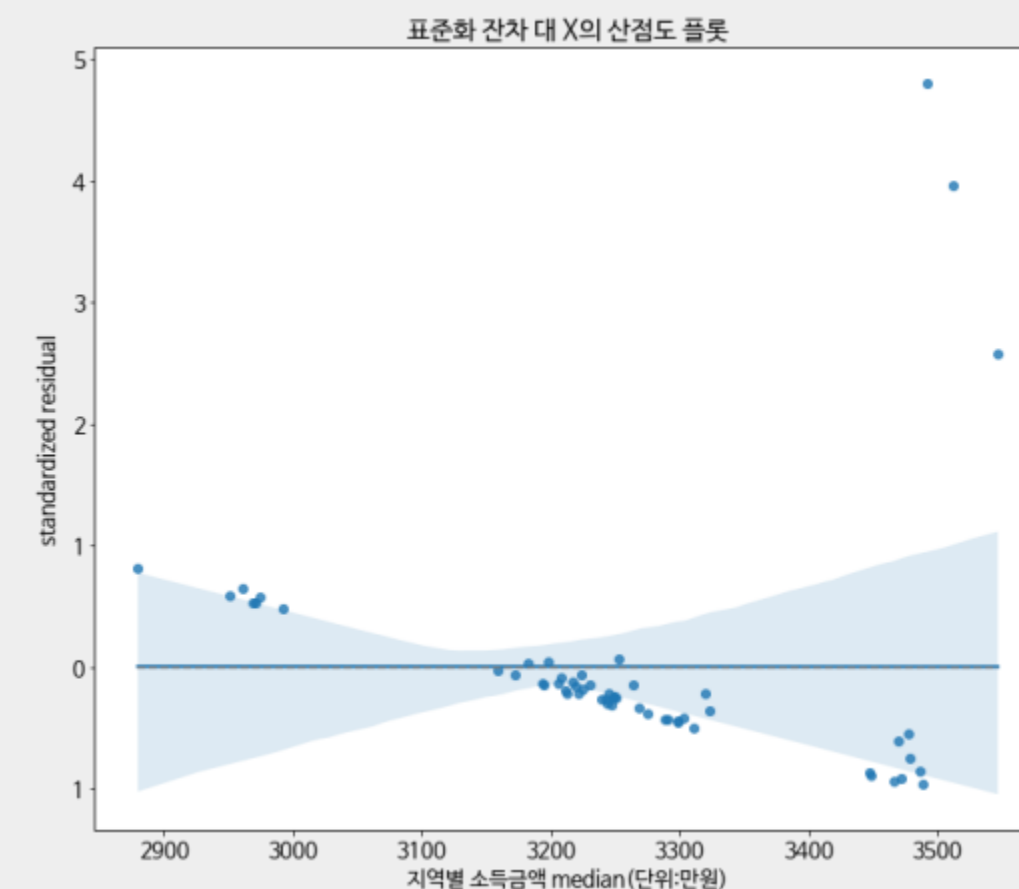
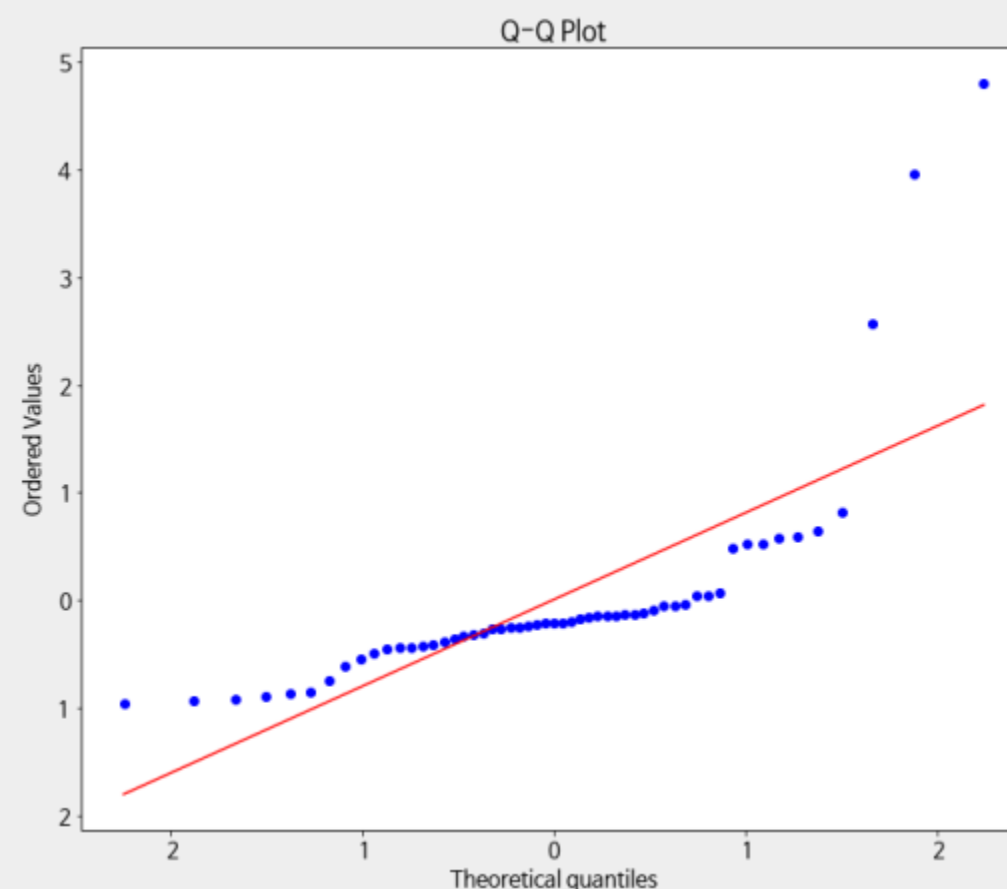
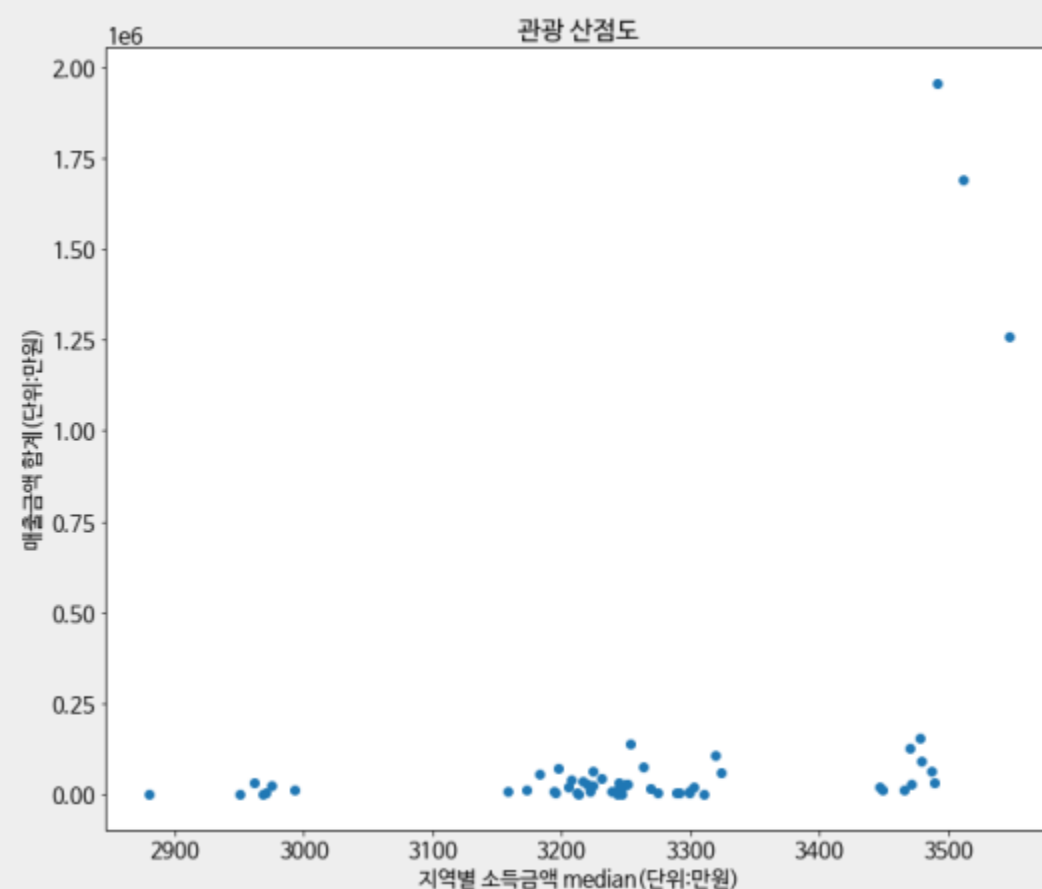
회귀가정을 위반한 업종 (ex. 관광, 잡화점)

해당 업종 리스트 (17개)

: 관광, 주차장, 휘트니스센터, 오토바이판매/수리, 잡화점, 내과/외과, 정형/성형외과, 한의원, 피부관리, 사진관, 치과, 커피/음료, 안과, 민속주점, 약국, 건강보조식품, 제과점

➡ Q-Q plot, 표준화 잔차 대 X의 산점도 분포 확인 결과 오차의 정규성, 등분산성 위반

➡ 해당 업종에 대해서는 회귀분석을 통해서 지역 소득이 매출금액에 영향을 주는지 알 수 없음



회귀분석 가정(오차의 정규성, 등분산성) 만족하는 업종

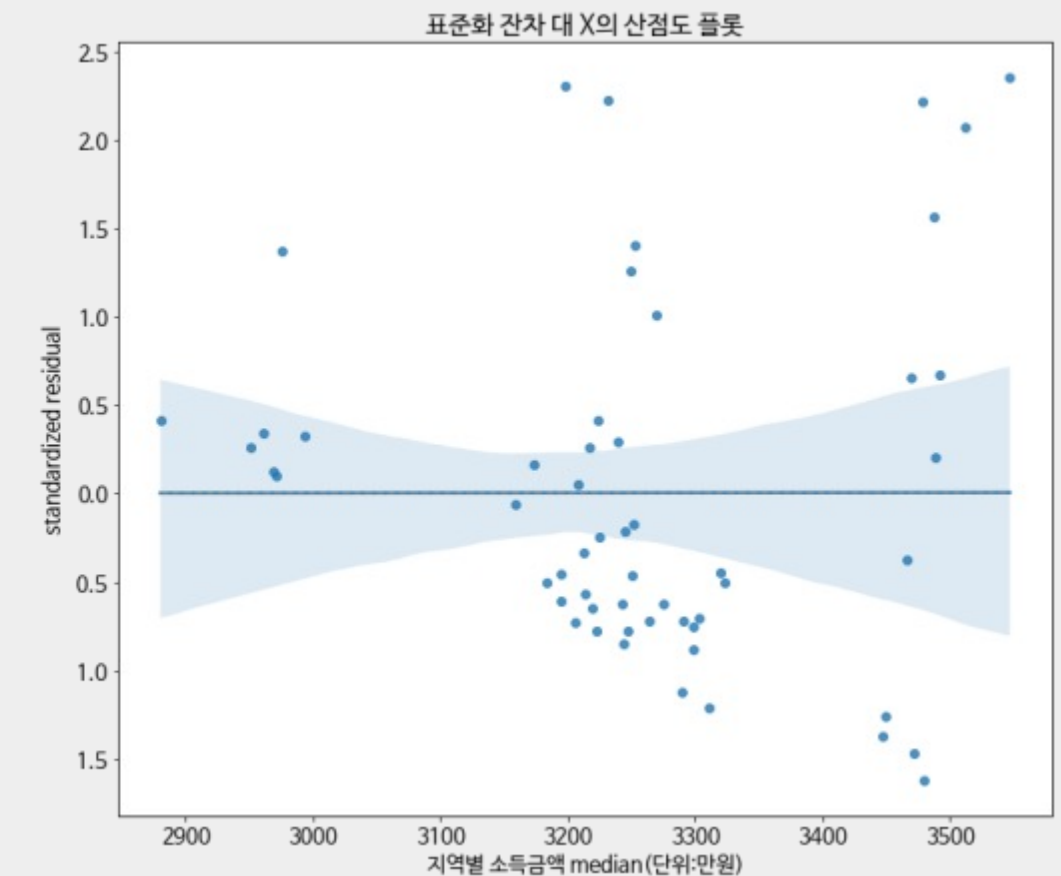
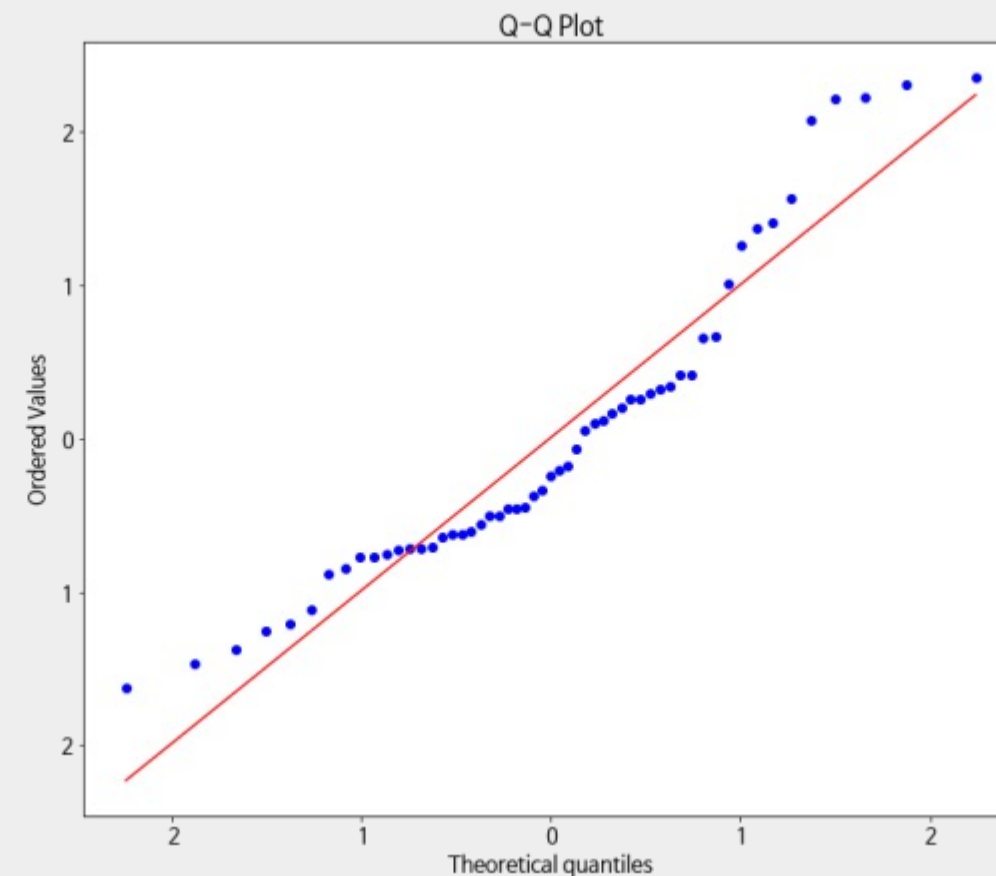
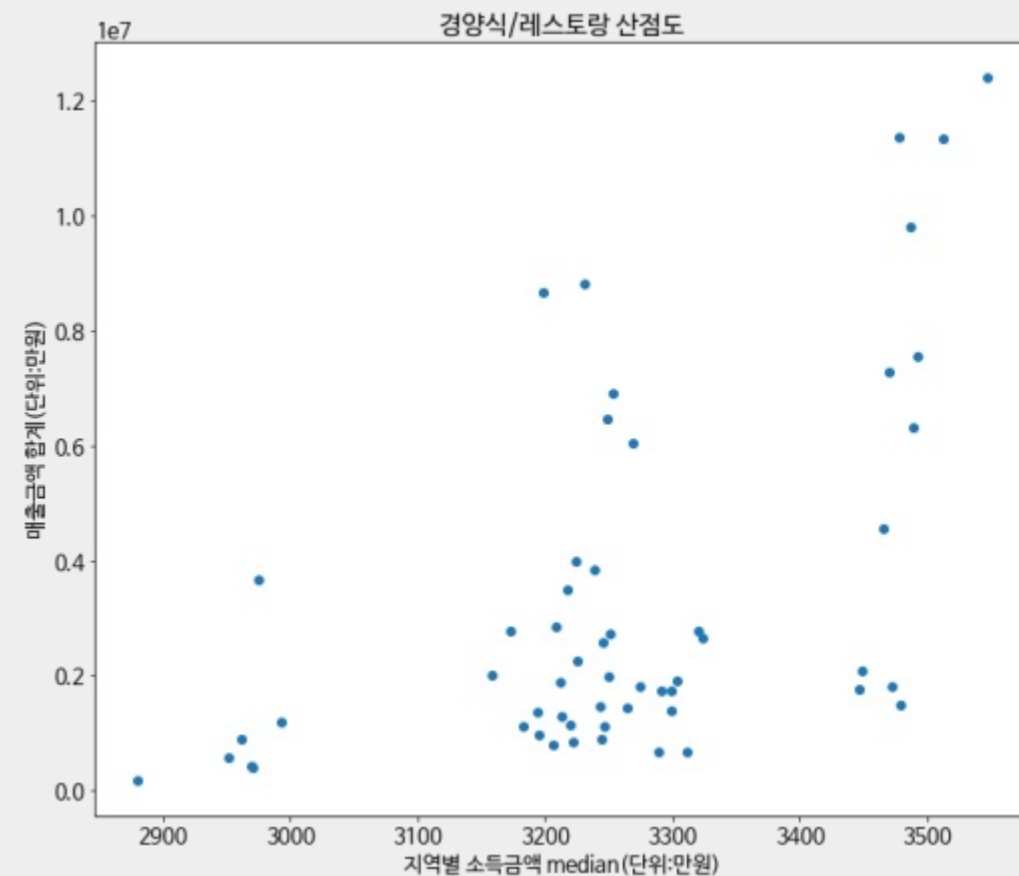
회귀가정을 만족하면서 p-value가 0.03미만으로 유의한 업종 (ex. 경양식/레스토랑, 화장품)

해당 업종 리스트 (20개)

: 경양식/레스토랑, 화장품, 의료기기, 패스트푸드, 휴대폰, 스낵, 피부/비뇨기과, 남성의류, BAR, 초밥/전통일식, 인력사무소, 일반의류, 액세서리, 횃집, 백반/한정식, 김밥/라면/도시락, 종합의류, 컴퓨터/주변기기, 곱창/양구이, 중식당

➡ 유의 수준 0.03에서 귀무가설(H0) 기각

➡ 해당 업종에 대해 지역별 소득수준은 매출금액에 영향을 주는 변수



ANOVA 분산분석

(2000, 3000) : 저소득 지역 그룹 (0 그룹)

(3000, 3400) : 중소득 지역 그룹 (1 그룹)

(3400, 4000) : 고소득 지역 그룹 (2 그룹)

ANOVA 분산분석 전제조건 검정

1. 정규성 검정

귀무가설(H_0) : 정규분포를 따른다.

대립가설(H_1) : 정규분포를 따르지 않는다.

2. 등분산성 검정

귀무가설(H_0) : 집단 간 분산이 동일하다.

대립가설(H_1) : 집단 간 분산이 동일하지 않다.

3. 독립성 검정

자료 수집이 random sampling된 자료라면 만족한다.

결과

업종명	저소득	정규성	p_value	중위소득	정규성	p_value	고소득	정규성	p_value	등분산성	p_value
경양식/레스토랑			0.004756			5.068488e-06			0.127641		0.001162
화장품			0.000451			1.776192e-03			0.041237		0.001338
의료기기			0.002059			2.544351e-06			0.000701		0.124701
패스트푸드			0.001477			4.634273e-04			0.090523		0.317405
휴대폰			0.011039			2.475943e-07			0.931354		0.389595

업종별로 모든 소득구간 그룹이 정규성을 따르는 경우는 없음
(적어도 한 그룹이 정규성 위배)

정규성 가정에 위배되므로 ANOVA를 진행할 수 없음

➡ 비모수검정 진행



비모수검정

Kruskal-Wallis 검정

Kruskal-Wallis Rank Sum Test

: 집단 중 하나라도 정규성 가정이 깨질 때 사용하는 검정방식

가정

1. 독립적인 표본
2. 측정값은 최소 순서형 변수

가설

귀무가설(H_0) : 세 그룹 간 매출금액(median)에 유의한 차이가 없다.

대립가설(H_1) : 세 그룹 간 매출금액(median)에 유의한 차이가 있다.

결과

모든 업종에서 p-value 값이 0.05(0.03으로도 만족)보다 작으므로 귀무가설을 기각

➡ 저소득 지역, 중소득 지역, 고소득 지역 간에
매출금액 크기에 통계적으로 유의미한 차이가 있다.

업종명	크루스칼 검정 p-value
경양식/레스토랑	0.000173
화장품	0.000184
의료기기	0.000053
패스트푸드	0.000395
휴대폰	0.000155



비모수 사후검정 결과

Bonferroni 방법

Bonferroni correction : 집단들을 각각 짝지어서 Wilcoxon test 진행

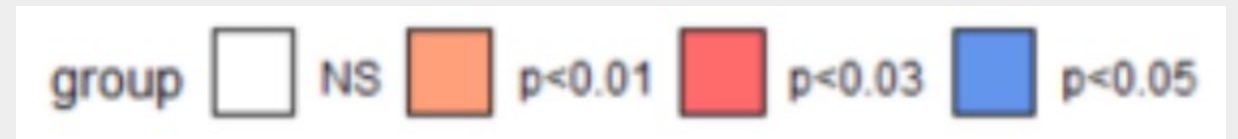
귀무가설(H0) : 두 그룹 간 차이가 없다.

대립가설(H1) : 두 그룹 간 차이가 있다.

결과



- 귀무가설(H0) : 두 지역 그룹 간에 유의미한 차이가 없다.
- 대립가설(H1) : 두 지역 그룹 간에 유의미한 차이가 있다.



※ x축, y축 각각 오른쪽, 위쪽 방향으로
저소득 지역, 중소득 지역, 고소득 지역 순서



연구결과 및 활용 방안

회귀분석

회귀가정을 만족하면서 p-value가 0.03 미만으로 유의한 업종 20개

: 경양식/레스토랑, 화장품, 의료기기, 패스트푸드, 휴대폰, 스낵,
피부/비뇨기과, 남성의류, BAR, 초밥/전통일식, 인력사무소,
일반의류, 액세서리, 횃집, 백반/한정식, 김밥/라면/도시락,
종합의류, 컴퓨터/주변기기, 곱창/양구이, 중식당

➡ 해당 업종은 지역별 소득이 높아질수록 매출금액이 높아진다.

비모수 검정

(2000, 3000) : 저소득 지역 그룹 (0 그룹)

(3000, 3400) : 중소득 지역 그룹 (1 그룹)

(3400, 4000) : 고소득 지역 그룹 (2 그룹)

각 업종별로 모든 그룹이 매출합계 측면에서 서로 유의미함

➡ p-value 기준을 0.03을 볼 때 상관관계가 높은 업종은 대부분
(저소득-중소득), (중소득-고소득), (저소득-고소득) 그룹 간 차이가
모두 유의하게 나타난다.

활용 방안

- 개인이 아닌 지역의 소득수준을 고려한 연구로, 경제 탄력성에 따른 마케팅 활용 가능
- 사회학, 행정학, 경영학 등 다양한 분야의 연구로 확장 가능

