
2025/12/3

서울 관광 재단/ 식당 품질 데이터 분석

2025170958 산업경영공학부 박준우

목차

01

데이터 수집 및 전처리

2p ~ 4p

02

데이터 통계 및 변수 분석

5p ~ 7p

03

식당별 특징 분석

8p ~ 10p

04

추가 분석 및 분석결과 요약

11p ~ 12p

05

결론 및 한계점

13p ~ 15p

데이터 분석 목적

모범식당이 가지는 변수간의 상관관계 분석
실생활과 밀접한 영향을 미치는, 음식점이라는 주제로 데이터 분석
을 진행하고 싶었음

음식점의 평가에 미치는 요소들을 분석

데이터 수집

OpenAPI 호출 URL

<https://api.odcloud.kr/api/15098044/v1/uddi:79ad0666-857c-4137-b048-fab6e9be866b>

데이터 타입: JSON

수집 도구: Python

```
# "returnType": "JSON", # 기본이 JSON이라 생략해도 됨
}

res = requests.get(url, params=params)

print("Status code:", res.status_code)
print("Raw text (앞부분):", res.text[:300])

# 혹시 에러메시지 확인용
if res.status_code != 200:
    raise SystemExit("API 호출 실패")

data = res.json()

# data 구조 확인
print(data.keys()) # dict: page, perPage, totalCount, currentCount, matchCount, data

df = pd.DataFrame(data["data"])
print(df.head())
print("행 개수:", len(df))
```

... Status code: 200
Raw text (앞부분): {"currentCount":1000,"data":[{"(RTI)지수":"4.2138394","네이버 평점":"4.40","수용태세지수":"0.370","식당(1D

	(RTI)지수	네이버 평점	수용태세지수	식당 (ID)	식당명	씨트립 평점	어워드정보설명	온라인화진행여부	인기도	#
0	4.2138394	4.40	0.370	10012	명동정	None	모범음식점 (2015)	N	0.33	
1	4.4236489	4.45	0.330	10014	봉추찜닭	4.7	모범음식점 (2019)	Y	0.51	
2	None	None	0.082	10015	금계찜닭	None	None	Y	None	
3	3.9222200	4.54	0.244	10016	명동한우방	None	None	N	0.09	
4	4.4000694	4.33	0.307	10018	삼호복집	None	모범음식점 (2004)	Y	0.33	

	지역명	지점명	트립어드바이저 평점
0	중구	None	None
1	중구	명동중앙우체국점	None
2	중구	명동점	None
3	중구	명동본점	None
4	서대문구	None	4.0

행 개수: 1000

데이터 전처리

결측치 제거: 리뷰수, 가격대 등 일부 컬럼 보정

컬럼명 정리: 공백 제거, 영어 변수명으로 통일(MATPLOTLIB을 사용하면 한글이 깨질 수 있기 때문.)

데이터 타입 변환

가격대 → FLOAT

리뷰수/품질지수 → INT

불필요 컬럼 제거

```
df_raw.columns = df_raw.columns.str.strip()

rename_map = {
    "식당명": "name",
    "어워드정보": "award",
    "(RTI)지수": "rti_score",
    "RTI지수": "rti_score",
    "온라인화진행여부": "online_yn",
    "수용태세지수": "preparedness_score",
    "인기도": "popularity",
    "트립어드바이저평점": "tripadvisor_score",
    "씨트립평점": "ctrip_score",
    "네이버평점": "naver_score",
    "지역명": "region",
    "시군구명": "region",
    "콘텐츠ID": "content_id",
    "식당ID": "content_id",
}

df = df_raw.rename(columns=rename_map)

print("rename 적용 후 컬럼:")
print(df.columns.tolist())

... rename 적용 후 컬럼:
['rti_score', 'naver_score', 'preparedness_score', '식당(ID)', 'name', 'ctrip_score', '어워드정보설명', 'online_yn', 'popula
```

데이터 구조 및 변수 소개

변수 확인

1. 식당 - ID 형태로 이뤄져 있음
2. 식당명
3. 지점명(프랜차이즈일 경우 존재)
4. 지역명(서울)
5. 어워드 정보
6. (RTI)지수(실수형 자료)
7. 온라인화지수(BOOL VALUE)
8. 수용태세지수(실수형 자료)
9. 트립어워드평점(공란이 많음 → 분석에 용이 X)
10. 씨트립평점(공란이 많음 → 분석에 용이 X)
11. 네이버평점(비교적 공란이 적음 → 분석에 용이)

1. 핵심 품질 지표 - 식당의 여러 상태를 나타냄

RTI_SCORE
PREPAREDNESS_SCORE
POPULARITY

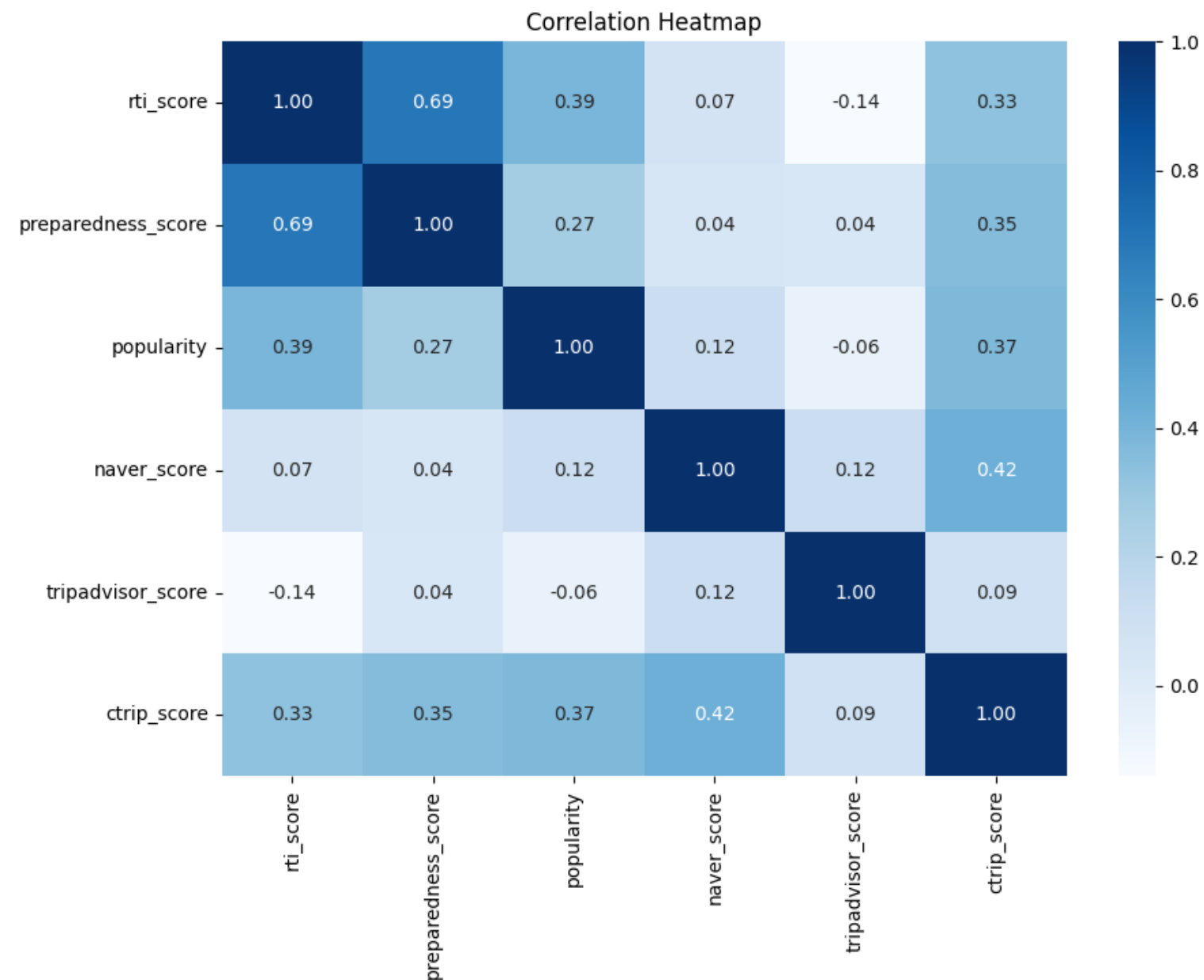
2. 평점

NAVER_SCORE
TRIPADVISOR_SCORE
CTRIP_SCORE

3. 범주형 변수

REGION
ONLINE_YN
어워드정보설명

데이터 분석 (1)



히트맵을 사용해 변수간 상관관계를 파악해 봤다.

전체적으로 품질 관련 지표(RTI_SCORE, PREPAREDNESS_SCORE)간에는 중간 정도의 양의 상관관계가 존재하지만, 외부 플랫폼 평점(NAVER, TRIPADVISOR, CTRIP)과의 상관은 약하거나 거의 없는 것으로 나타남.

RTI_SCORE ↔ PREPAREDNESS_SCORE : 0.69 (강한 양의 상관)

“RTI 종합품질지수”와 “수용태세지수”는 의미적으로 가까운 변수라 두 점수가 함께 높아지는 경향.

즉 서비스 품질이 좋은 식당은 전반적인 준비·응대 수준도 함께 좋다.

RTI_SCORE ↔ 외부 평점간의 상관관계

NAVER_SCORE: 0.07 (거의 무관)

TRIPADVISOR_SCORE: -0.14 (사실상 무관)

CTRIP_SCORE: 0.33 (약한 양의 상관)

식당 평점에 영향을 미치는 변수 분석

온라인 서비스 제공 여부에 따른 품질차이

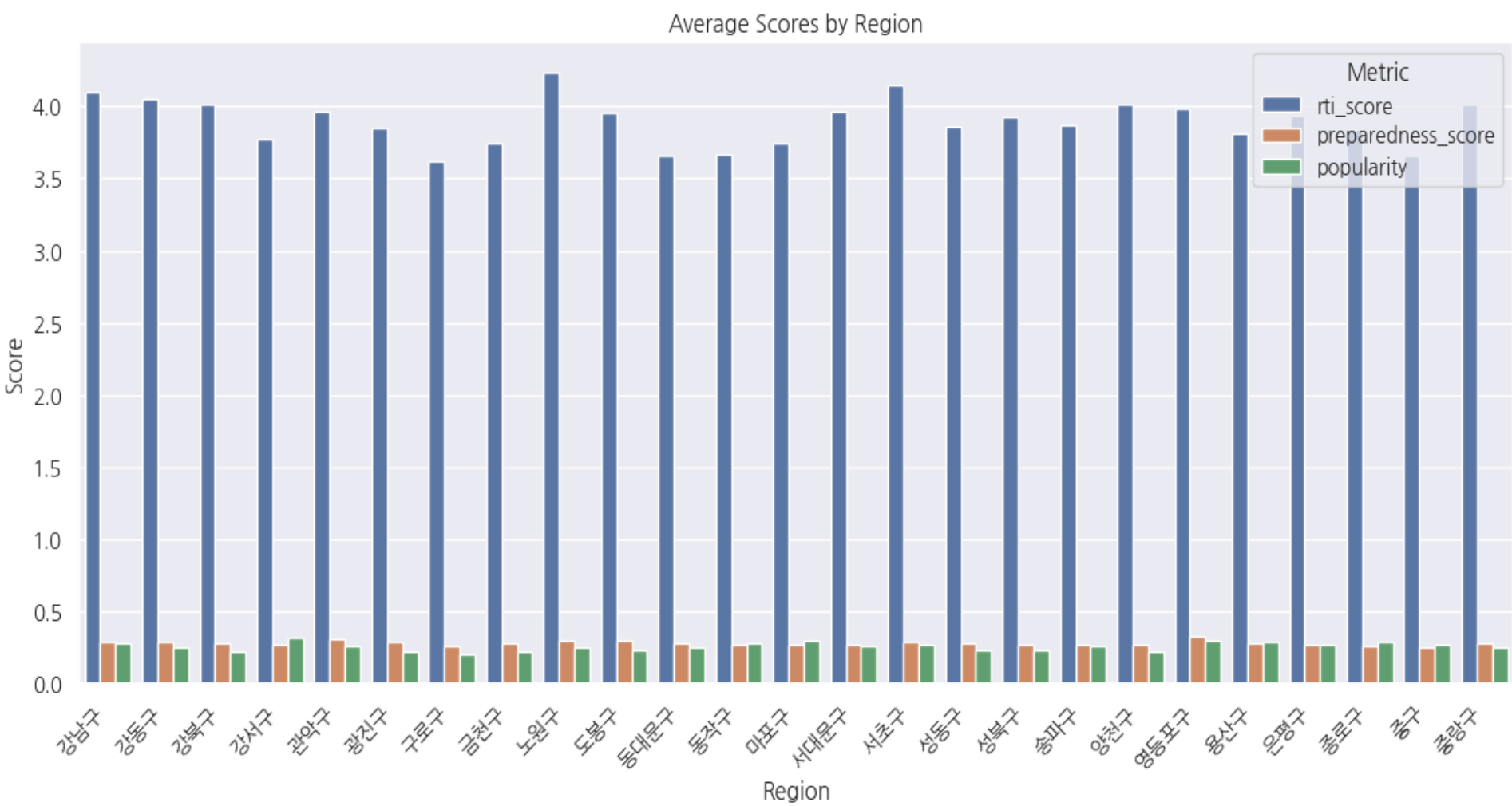
상위 이상치(OUTLIERS)가 많음 → 온라인 서비스가 없어도 일부 식당은 인기도가 매우 높은 경우가 존재한다는 의미.
중앙값이 더 높음 → 전반적으로 온라인 서비스가 있는 식당의 인기도가 더 높다고 해석 가능.

전체적으로 더 촘촘한 분포 → 인기도가 일정 수준 안에서 안정적으로 모여 있음.

하단에 이상치 존재 → 온라인 서비스가 있어도 인기도가 매우 낮은 식당도 존재한다는 의미.



지역에 따른 품질 차이
• 유의미한 평점 차이는 찾아 볼 수 없었음.



품질 수준(상·중·하)별 식당 특징 분석

RTI·수용태세·인기도를 합쳐서 종합 품질 점수(QUALITY_MEAN) 만들기

이 점수를 기준으로 상·중·하 3개 그룹으로 나누기



품질 수준이 높을수록 온라인 서비스 이용률이 다소 증가한다.

HIGH 그룹 Y 비율이 가장 높고

LOW 그룹 Y 비율이 가장 낮음

☆ 하지만 어느 그룹이든 Y보다 N이 훨씬 많다.

즉, 품질과 관계없이 온라인 서비스 이용이 전반적으로 낮게 나타남.

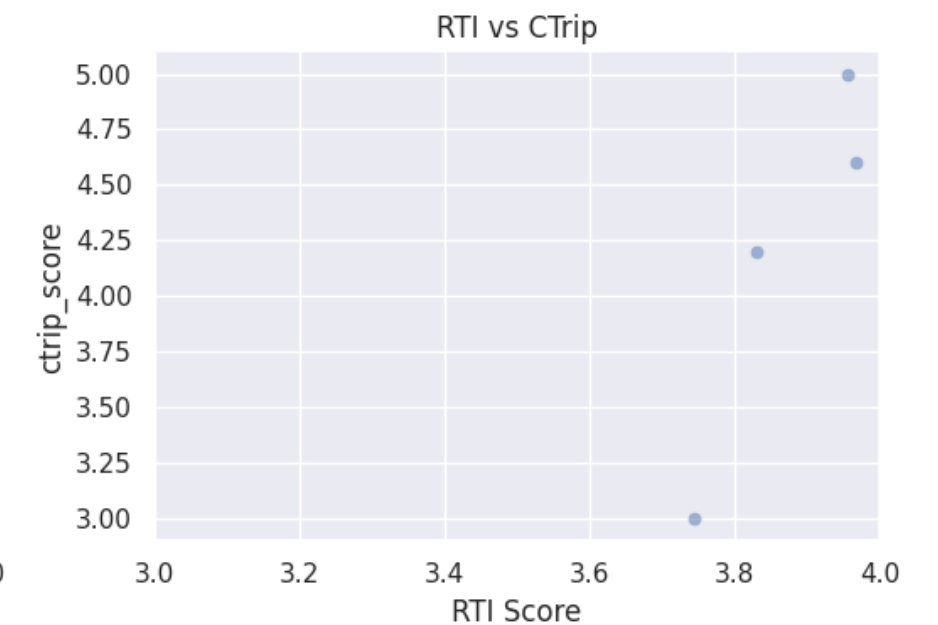
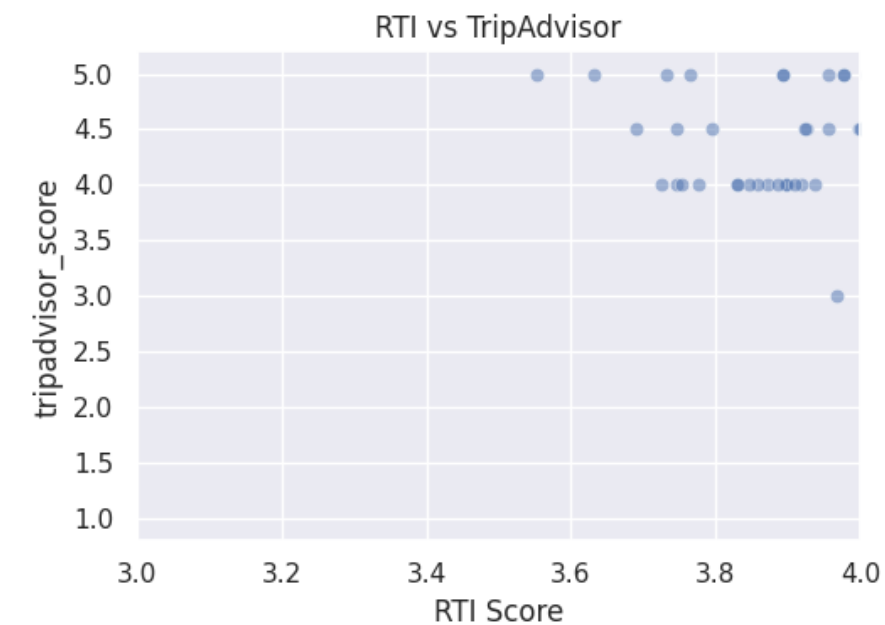
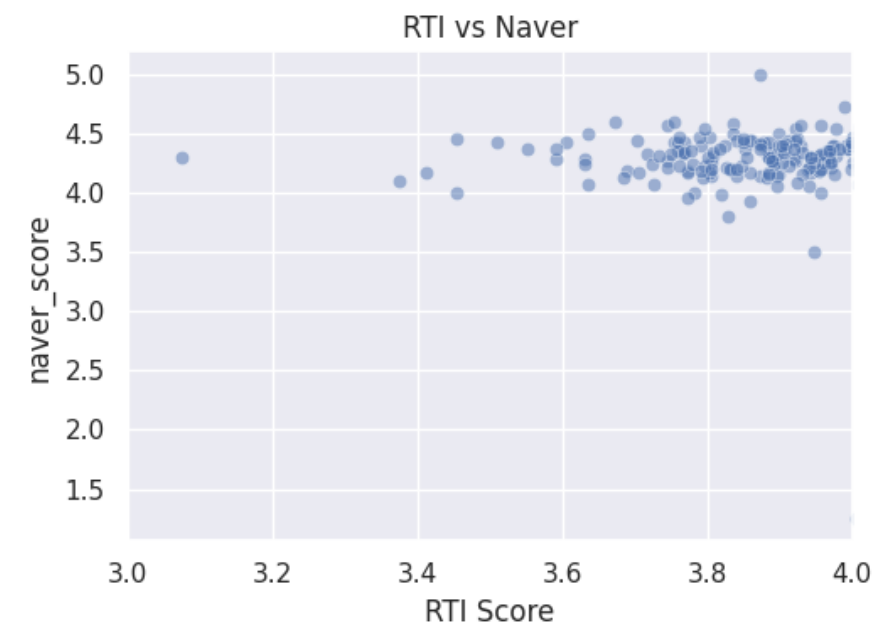
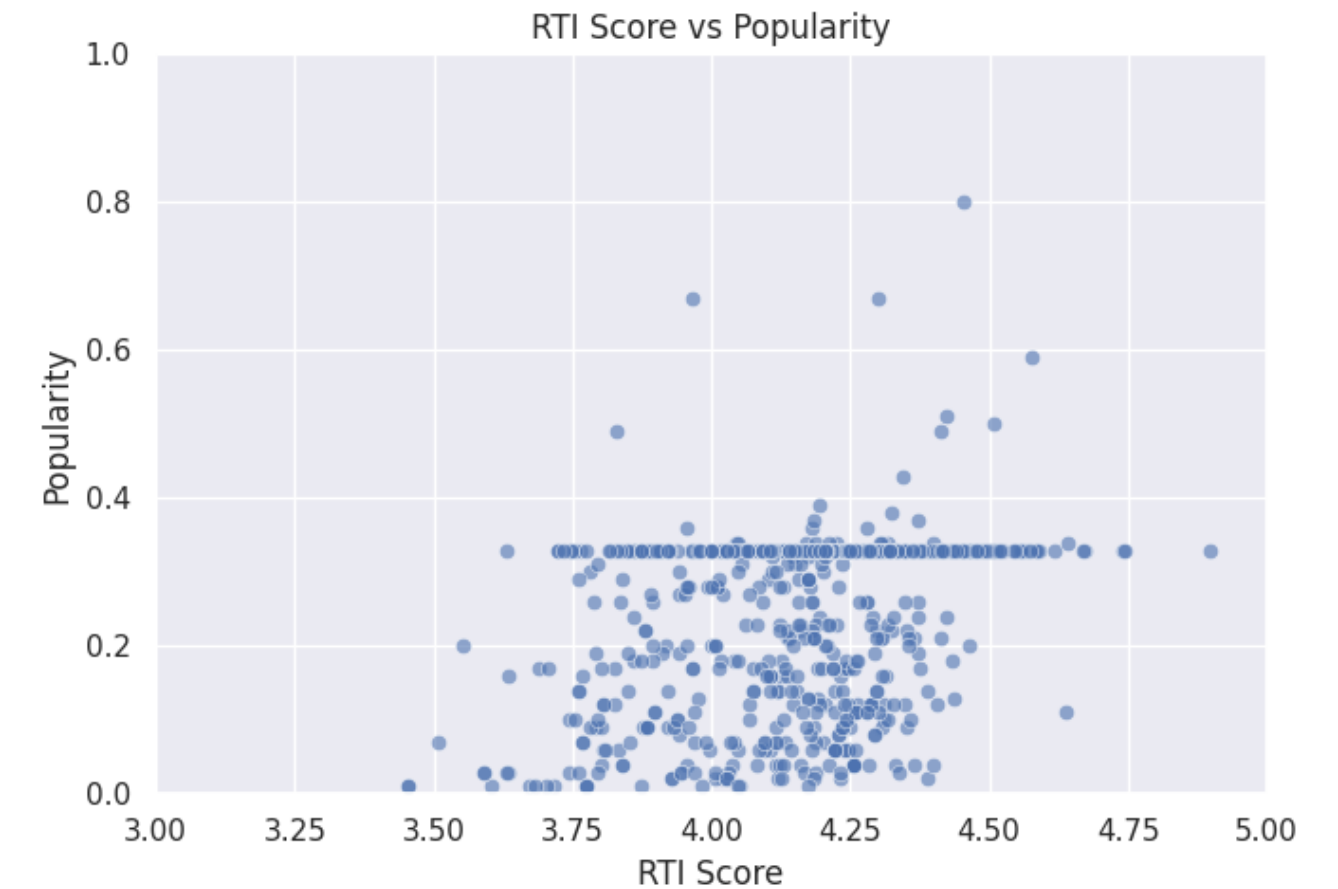
식당 데이터 분석 - 추가 인사이트

RTI VS 인기도 산점도(SCATTER PLOT)

→ 약한 양의 상관관계를 보인다. 그러나 이상치가 많음.(인기와 평점이 꼭 비례하는것은 아님을 시각적으로 보여줌)

RTI VS 외부 평점 산점도 (네이버/트립어드바이저/CTRIP)

→ 외부 평점과 품질 지표의 상관이 약함을 시각적으로 확인



상위 10% 식당의 특징 분석

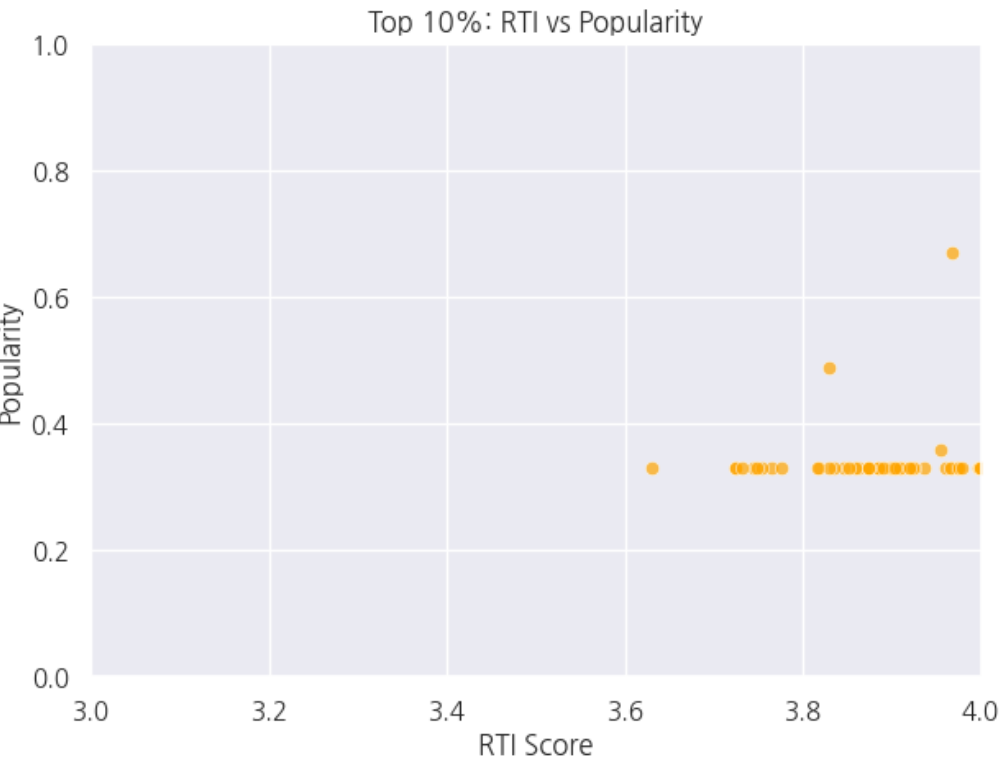
전체 평균:
RTI_SCORE 3.91
PREPAREDNESS_SCORE 0.28
POPULARITY 0.26
DTYPE: FLOAT64

상위 10% 평균:
RTI_SCORE 4.25
PREPAREDNESS_SCORE 0.31
POPULARITY 0.33
DTYPE: FLOAT64

상위 판단의 기준은 인기로 결정.

가설 설정 : 상위 10% 식당의 평균 준비도, RTI 스코어는 전체 평균보다 높을 것이다.

반면 상위 10%에서도 RTI 스코어와 인기는 큰 상관관계가 없는 것으로 분석되었다.



상위 10% 식당 분석(2)

지역에 따라, 상위 10% 식당의 수를 비교해봄. 단 '수'를 이용해 판단할경우, 절대적인 음식점 수가 많은 쪽이 통계적 이점을 가져감으로 비율로 판단.

강남구, 송파구, 중구, 종로구, 용산구, 서초구.... 순으로 상위 10% 식당이 많이 분포해 있었다.



region	
강남구	15.9
송파구	8.6
중구	6.8
강서구	6.5
종로구	6.1
용산구	5.4
서초구	5.4
서대문구	4.8
성동구	3.9
관악구	3.9
은평구	3.9
영등포구	3.8
마포구	3.4
광진구	3.2
노원구	2.7
동작구	2.5
중랑구	2.3
구로구	1.8
강동구	1.6

결론

식당 품질은 지역적 특성과 서비스 운영 방식에 따라 차이를 보인다

온라인 서비스 제공 여부는 품질과 직접적인 관련이 없다

외부 리뷰 기반 평점(네이버·트립어드바이저·CTRIP)은 공식 품질지표와 무관하다

인기도(노출·방문량)는 품질보다 지역·상권력에 더 큰 영향을 받는다

품질·인기도·평점은 서로 다른 차원의 지표다

어려웠던/한계점

1. OPEN API를 불러오는 과정에서의 어려움

- OPEN API의 서버 상태가 좋지 않아 중간에 OPEN API를 불러 올 수 없었던 이슈가 있었음. OPEN API에 대한 이해가 부족해 키를 불러오는 것에 약간의 어려움이 있었음

2. MATPLOTLIB 모듈 사용시의 어려움

X축과 Y축 범위 설정, 그리고 한글로 표현된 변수들을 깨짐 없이 표현하는 방법에 대해 어려움이 있었음. → 영어로 표현할 수 있는 변수는 모두 전환했으나 중구, 서초구 같은 지명은 전환하기 힘들. → 맑은 고딕체를 코랩에 다운 받아 해결.

3. 자료의 아쉬움

자료에 포함된 평점 데이터 중, 네이버 평점 데이터는 좋았지만 그 밖 해외 사이트에 등록되어 있는 식당 평점 데이터는 NONE 값이 너무 많아 유의미한 통계 결과를 내기 매우 좋지 않았음.

마지막 페이지(끝)

- **Thank you for Listening**

Q&A
