Yonsei Data Science Lab
Summer 2025 EDA Project

DATASCIENCE LAB

리뷰-인프라 기반 국내 관광객 유입 요인 분석

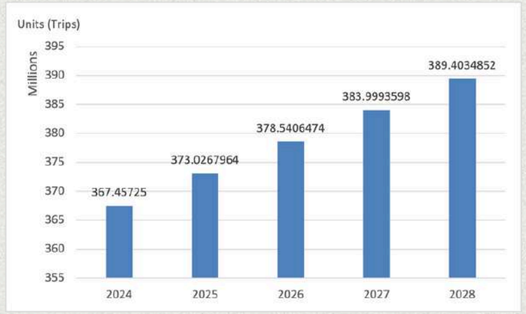
13기 강승우 이진우 | 14기 고서연 신지원

Contents

01 주제 소개

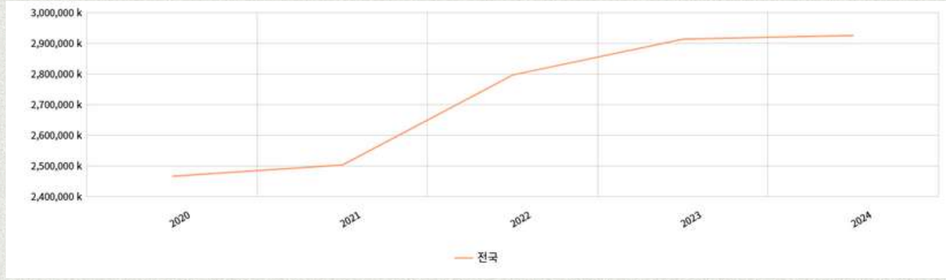
주제 선정 배경

팬데믹 이후,
내국인 중심의 관광 수요 빠르게 회복



- 국내 여행은 2023년 기준 361.8백만 회로 **전년 대비 9.2% 증가**
- 국내 여행 비중이 꾸준히 증가하며 **내국인이 중심이 된 관광 패턴** 정착
- 연도별 방문자수 추세 또한 꾸준히 증가하는 추세

관광 산업은 국가 경제의 성장 동력



- 내/외국인의 전국 관광객 수 꾸준히 증가 추세

항목	2023년 실제치	2034년 전망치
GDP 기여	84.71조 원	130.9조 원
일자리 수	134만 명	180만 명

- 2034년까지 한국 관광산업의 **연간 GDP 기여도는 130.9조 원, 전체 GDP의 약 5%**를 차지할 전망
- 고용 측면에서도 180만 명 이상을 지원하며 **전체 일자리의 6.8%**를 차지하게 될 것으로 예측됨

<u>지속 가능한 관광 활성화를 위해, 방문객 유입을 결정짓는 다양한 요인을 통합적으로 분석하는 것이 중요</u>

프로젝트 목적

관광객 수에 영향을 주는 요인을 분석해 국내 관광지 활성화 전략 수립
관광객 리뷰 및 관광 인프라 데이터를 활용한 통합 분석 수행

주요 분석

리뷰 데이터 기반 감정 분석

관광지별 긍·부정 감정 분포 및 변화 추이 분석

리뷰 데이터 기반 핵심 키워드 추출

관광객들이 중요하게 여기는 요소 파악

관광 인프라와 방문자 수 관계 분석

안내소, 숙박, 축제 등의 인프라 요소가
방문자 수에 미치는 영향 파악

02 데이터 개요

데이터 소개

데이터셋 요약

1. **정량적 데이터: 관광객 수 (2020~2024), 관광안내소/숙박/축제 위치 정보**
2. **정성적 데이터: 네이버 지도에서 크롤링한 관광지 리뷰 (상위/하위 100곳)**

데이터명	주요 컬럼	데이터 출처
기초지자체 방문자수	2020~2024년 연도별 방문자 수	한국관광공사
전국관광안내소표준데이터	관광안내소명, 위치, 위경도	공공데이터포털
전국관광지정보표준데이터	관광지명, 위치, 위경도	공공데이터포털
전국관광펜션업소표준데이터	숙박업소명, 주소, 좌표정보	공공데이터포털
전국문화축제표준데이터	축제명, 위치, 위경도	공공데이터포털
naver_map_reviews_top100	관광지명, 리뷰, 리뷰 수, 방문객 수	네이버 지도 크롤링
naver_map_reviews_bottom100	관광지명, 리뷰, 리뷰 수, 방문객 수	네이버 지도 크롤링
관광지별 방문자수 데이터	관광지명, 2020~2024년 방문자 수	한국관광공사
행정구역 경계 데이터	X	Geoservice Web

데이터 수집 방법

관광객 수 데이터

- 한국관광공사에서 지역별로 다운로드 받은 후 병합

인프라 데이터

- 공공데이터포털에서 다운로드

네이버 리뷰 데이터

- 관광지별 관광객 수 데이터에서 관광지만 추출 후 각 관광지의 리뷰를 최대 100개씩 네이버 지도 리뷰에서 크롤링
- 관광객 수 상위 100개 관광지와 하위 100개 관광지 대상

행정구역 경계 데이터

- Geoservice Web에서 다운로드

데이터 전처리 방법

관광지별 관광객 수 데이터

- 필요한 컬럼만 남긴 후, 결측치 존재 행 제거
- 주소 문자열에서 행정구역(구/군/시) 추출하여 각 관광지별 위치정보 통일

인프라 데이터

- 필요한 컬럼만 남긴 후, 결측치 존재 행 제거, 숙박업소 데이터의 경우 폐업 업소 제거
- 위경도 데이터를 좌표계 EPSG:5179로 통일하기 위해 좌표계 변환 적용
  → GeoDataFrame 생성

네이버 리뷰 데이터

- 추후 임베딩 시 노이즈가 될 수 있는 특수기호 삭제

03 데이터 분석

데이터 분석 과정 요약

1) 감정 분석

온라인 리뷰 텍스트
→ 긍정/부정 감성 분석 평가

2) 키워드 분석

온라인 리뷰 텍스트 임베딩
→ 코사인 유사도 기반 키워드 매핑

3) 관광 인프라 공간 커버리지 분석

상관 분석
다중 회귀 분석 (OLS Regression)

4) 데이터 시각화

지역 간 비교를 위해 4가지 변수 시각화
: 2024년 방문자 수, 관광지당 안내소 평균,
관광지당 숙박업소 평균, 지역 축제 개수

감성 분석 과정

1. 감성 분석 모델 초기화 (총 4개의 모델 사용)

- WhitePeak/BERT-Korean, KoMiniLM, DistilBERT-SST2, nlptown Multilingual

2. 감성 분석

- 관광지 별 리뷰들을 모두 감성 분석 모델에 적용 후 Softmax 확률을 추출
  → 1~0 사이의 연속 분포로 출력

3. 관광지 평균 감성 분석 결과 비교

- Top 100 관광지와 Bottom 100 관광지의 평균 감성 분석 점수 비교

감성 분석 사용 언어모델 비교

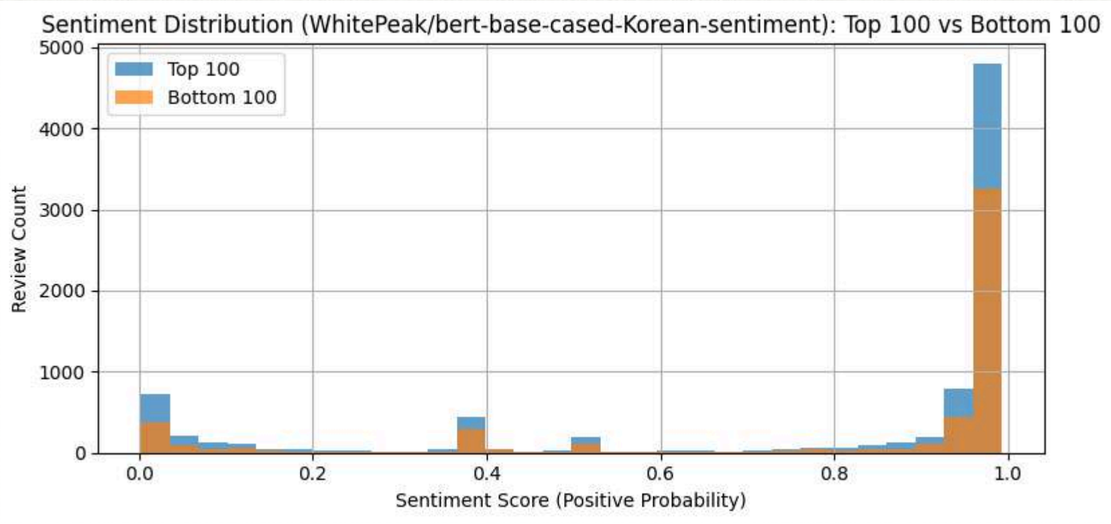
모델별 감성 분석 성능 및 특징

모델	분포 특징	핵심 특징	한글 리뷰 적합도
한글 파인튜닝 모델 (WhitePeak-bert-based-model)	감성 점수가 0 또는 1에 극단 집중 → 중간값 거의 없음	단적인 긍정/부정으로 분류하는 경향이 있어 관광지별 극명한 차이를 보여줌	★★★★☆
영어 파인튜닝 모델 (distilbert)	0.2~1.0 구간에 고르게 분포 → 부드러운 연속 스펙트럼	연속적인 점수 분포로 미세한 감성 차이까지 반영, 실제 긍정/부정 분포도 상위/하위 100에서 명확히 달라짐	★★★☆☆
트위터 특화(cardiffnlp-twitter-robert-based)	대다수 리뷰가 0.05~0.2 낮은 점수에 몰림 → 거의 부정으로 예측	한글 리뷰에서 거의 모두를 부정으로 해석하는 경향이 뚜렷해, 본 실험 목적에는 적합하지 않다.	★☆☆☆☆
멀티랭귀지 BERT(nlptown)	0, 0.25, 0.5, 0.75, 1.0 의 5단계 계단형 분포	멀티랭귀지 BERT(nlptown)는 5단계 등급(0~1)으로 점수를 매겨, 극단적인 분포와 등급화된 분포 특징을 보임	★★☆☆☆

→ 공통적으로, 상위 관광지일수록 긍정 리뷰의 분포가 더 높게 나타나고, 하위 관광지는 부정에 치우치는 경향이 나타난다.

리뷰 감성 분석 (1/2)

사용한 모델 1) WhitePeak/bert-base-cased-Korean-sentiment(한국어 파인튜닝)

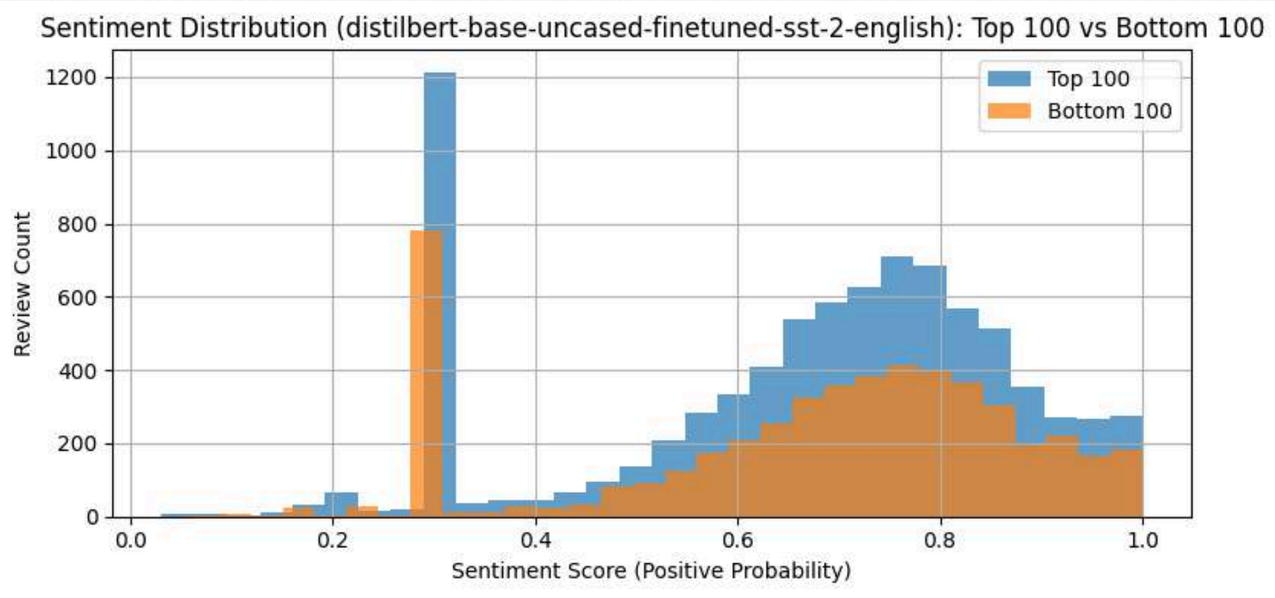


- 리뷰의 긍정·부정 분포를 계산하고 관광지별 평균 감정 점수 분석
- 리뷰의 감성 점수가 극단적으로 0(부정) 또는 1(긍정)에 몰려 있음
- **긍정/ 부정 확률을 명확히 (0 또는 1)로 판별, 관광객이 많은 상위 관광지가 대체로 긍정적인 평가가 많음(1.0 비율이 높음)**

리뷰 예시	긍정 확률(0~1)	해석
작품 관람하기에 참 좋습니다	0.989	매우 긍정적
이 전시 최악입니다	0.002	매우 부정적

리뷰 감성 분석 (2/2)

사용한 모델 2) distilbert-base-uncased-finetuned-sst-2-english (영어 기반 감정분석 모델)



- 감성 점수를 **연속적으로** 잘 분포시킴
- **관광객 수가 많은 상위 100 관광지가 하위 100에 비해 리뷰가 더 긍정적임**을 시각적으로 보여줌

키워드 임베딩 분석 과정

1. 임베딩 모델 초기화 (총 4개의 모델 사용)

- bge-m3-korean, KoSimCSE-roberta, KoSimCSE-bert, KR-SBERT-V40K-klueNLI-augSTS

2. 키워드 임베딩 생성

- 키워드 예시: '자연: 산, 바다, 호수 등 자연 경관', '음식: 맛집, 카페, 음식거리' 등
- 총 18개 키워드 선정 후 임베딩 정확도 향상을 위해 간단한 설명 추가 후 임베딩

3. 리뷰 임베딩 생성

- 모든 리뷰 텍스트를 임베딩하여 벡터로 변환

5. 코사인 유사도 계산 및 결과 출력

- 리뷰 벡터와 키워드 벡터의 유사도 비교 후 Place별로 평균 집계

5. 일정 임계값 이상의 상위 키워드 추출

- 임계값을 0.3-0.5까지 0.05간격으로 바꿔 가며 Place별 상위 키워드 추출

키워드 임베딩 사용 언어모델 비교

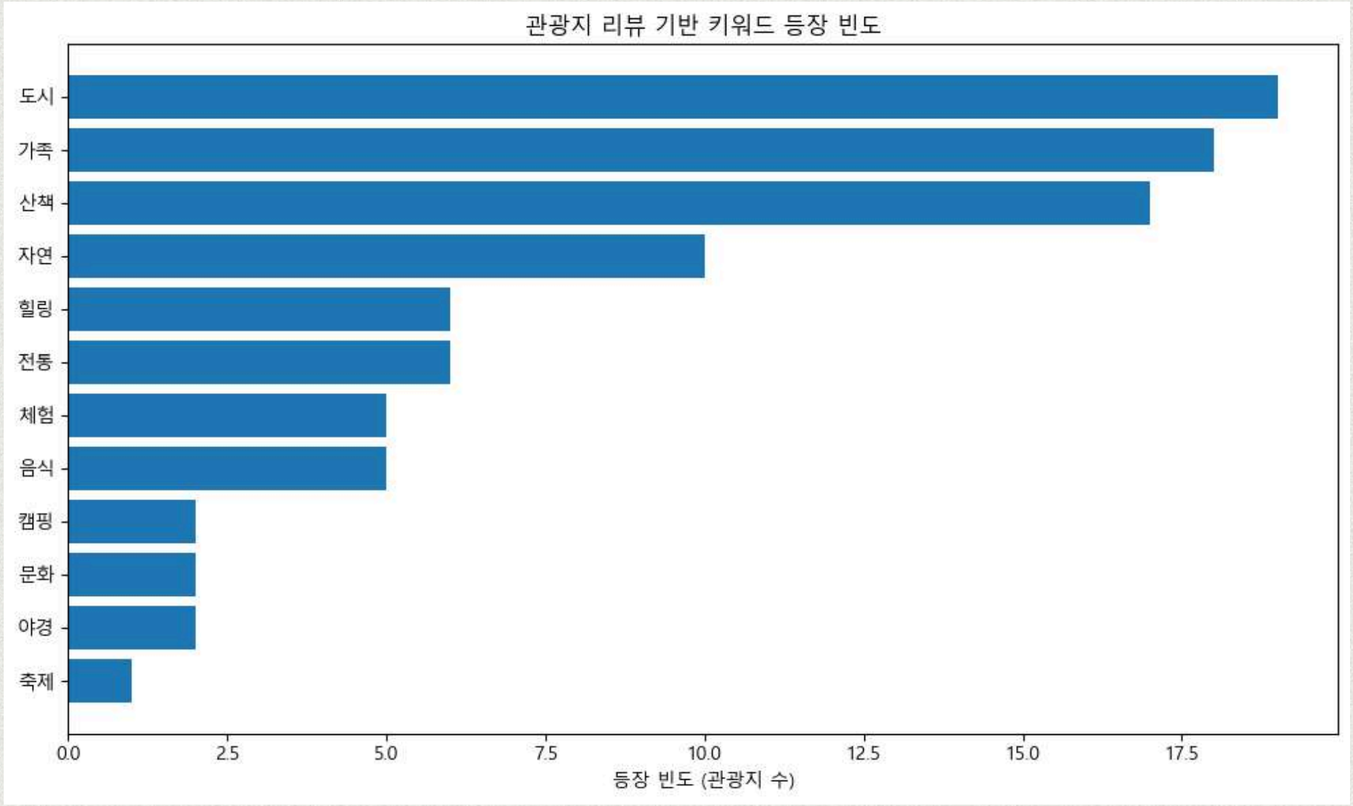
모델별 키워드 임베딩 성능 및 특징

모델	모델 특징	추천 Threshold 및 목적	한글 리뷰 적합도
upskyy/bge-m3-korean	BGE 기반 대규모 한국어 임베딩 모델	@0.30(키워드 양 최대화)	★★★☆☆
BM-K/KoSimCSE-roberta-multitask	RoBERTa 기반 한국어 KoSimCSE 다중 Task 임베딩 모델	@0.35-0.40(키워드 안정성/정확도 균형)	★★★★★
BM-K/KoSimCSE-bert multitask	BERT 기반 한국어 KoSimCSE 다중 Task 임베딩 모델	@0.35-0.40(안정적 결과 확보)	★★★★☆
snunlp/SBERT-V40K-klueNLI-augSTS	한국어 NLI + STS로 강화된 SBERT	@0.35-0.40(균형잡힌 키워드 추출)	★★★☆☆

인기 키워드 분석

최종 사용 모델 - Roberta-multitask @ 0.4

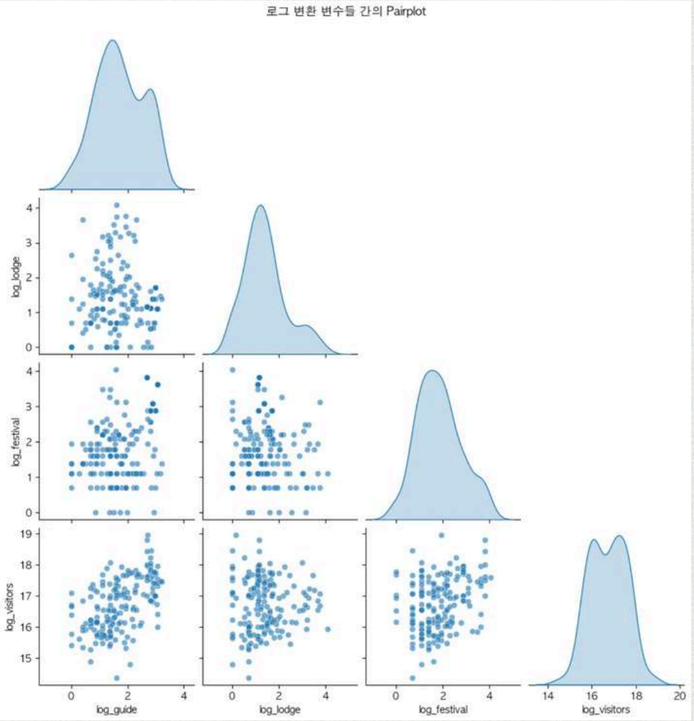
관광지명	대표 키워드
동궁과 월지	야경, 자연
고석정꽃밭	산책, 자연
국립중앙박물관	문화, 전통

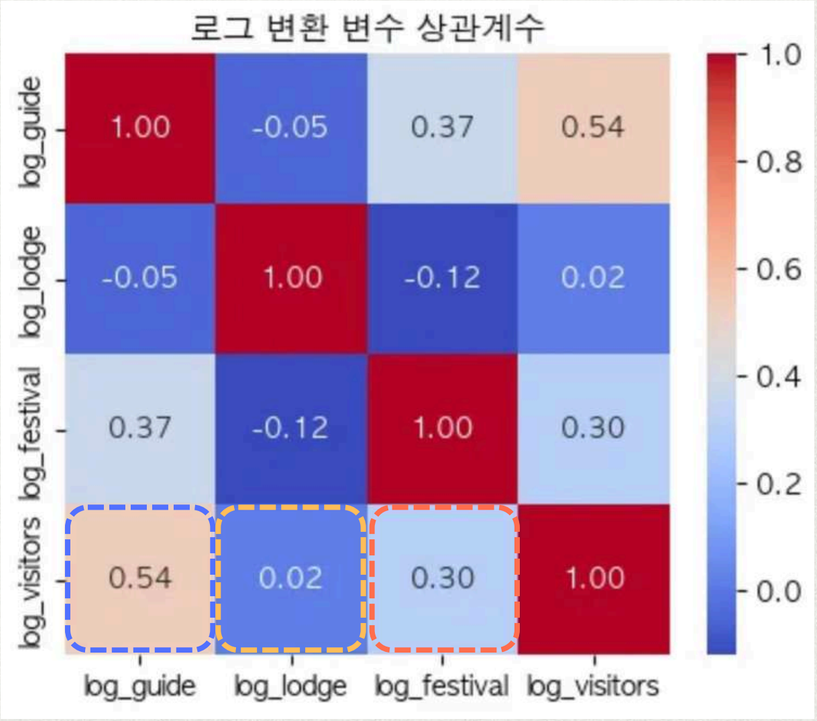


관광 인프라 영향력 분석 (1/2)

상관 분석 결과 (로그 변환 데이터)

- Pairplot 및 상관관계 히트맵 분석 결과





- 결과 해석

안내소(log_guide)와 방문자 수(log_visitors)의 상관계수가 0.54로 가장 높음

축제(log_festival)와 방문자 수 간에도 유의미한 상관(0.30)이 있음

숙박(log_lodge)은 방문자 수와 거의 상관 없음(0.02)

관광 인프라 영향력 분석 (2/2)

다중 회귀 분석 (OLS Regression)

OLS Regression Results

Dep. Variable:	log_visitors	R-squared:	0.302
Model:	OLS	Adj. R-squared:	0.290
Method:	Least Squares	F-statistic:	24.52
Date:	Sun, 27 Jul 2025	Prob (F-statistic):	3.11e-13
Time:	15:37:23	Log-Likelihood:	-187.13
No. Observations:	174	AIC:	382.3
Df Residuals:	170	BIC:	394.9
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	15.5771	0.175	89.125	0.000	15.232	15.922
log_guide	0.5016	0.071	7.094	0.000	0.362	0.641
log_lodge	0.0522	0.061	0.862	0.390	-0.067	0.172
log_festival	0.1171	0.064	1.842	0.067	-0.008	0.243

Omnibus:	0.464	Durbin-Watson:	1.626
Prob(Omnibus):	0.793	Jarque-Bera (JB):	0.578
Skew:	-0.112	Prob(JB):	0.749
Kurtosis:	2.827	Cond. No.	10.8

결과 해석

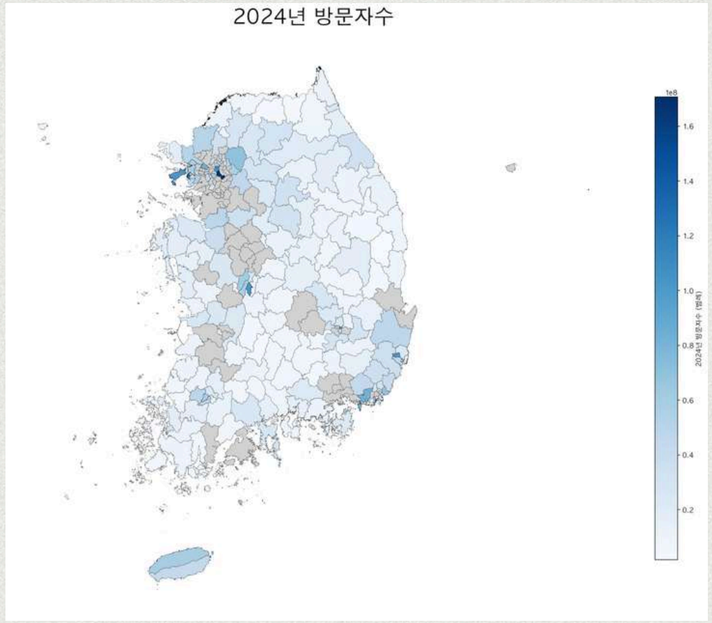
- **관광 안내소 수는 관광객 수에 가장 큰 영향을 미치는 요소 (매우 유의미, p<0.001)**
- 축제는 관광객 수에 긍정적 영향을 미치는 요소 (약한 유의성, p=0.067)
- 숙박업소의 양적 수는 관광객 수와 유의미한 상관성을 보이지 않음 → 질적 개선 필요

데이터 시각화 (1/3)

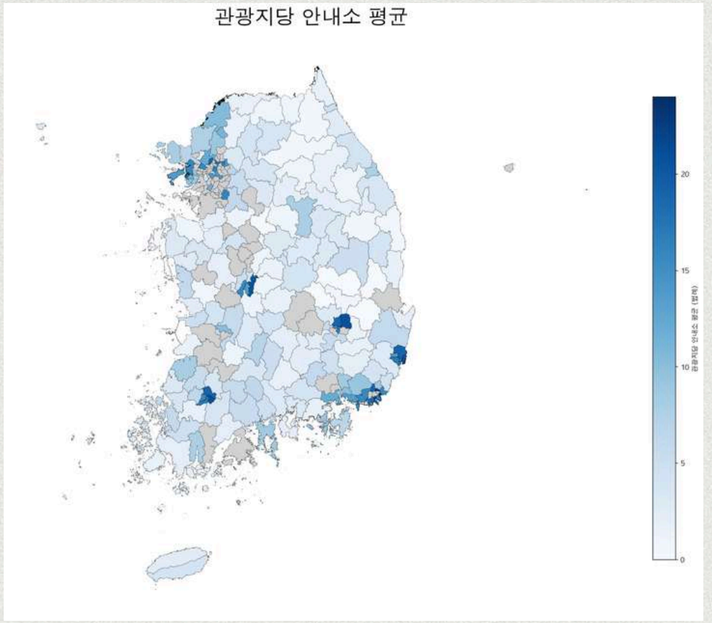
공간 데이터 시각화 (Choropleth Maps)

전국 기초지자체를 기준으로 다음 4가지 변수를 시각화, **지역 간 비교 가능**

2024년 방문자 수



관광지당 안내소 평균

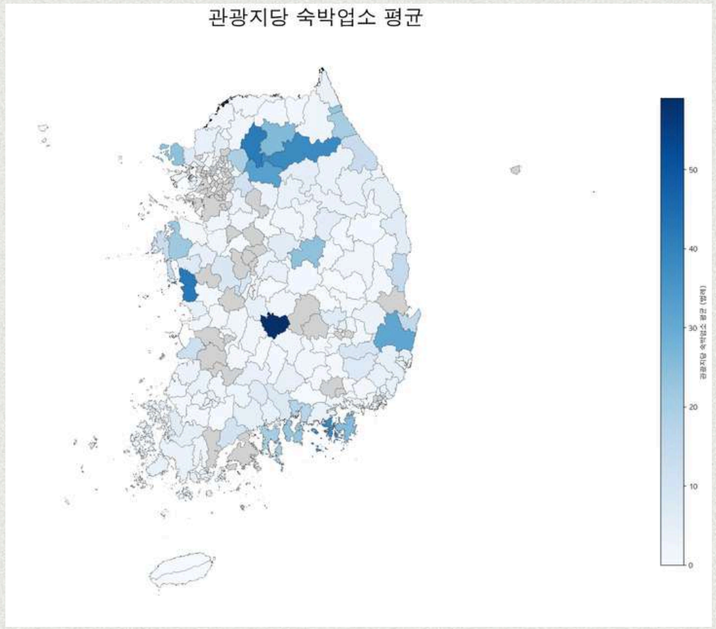


데이터 시각화 (1/3)

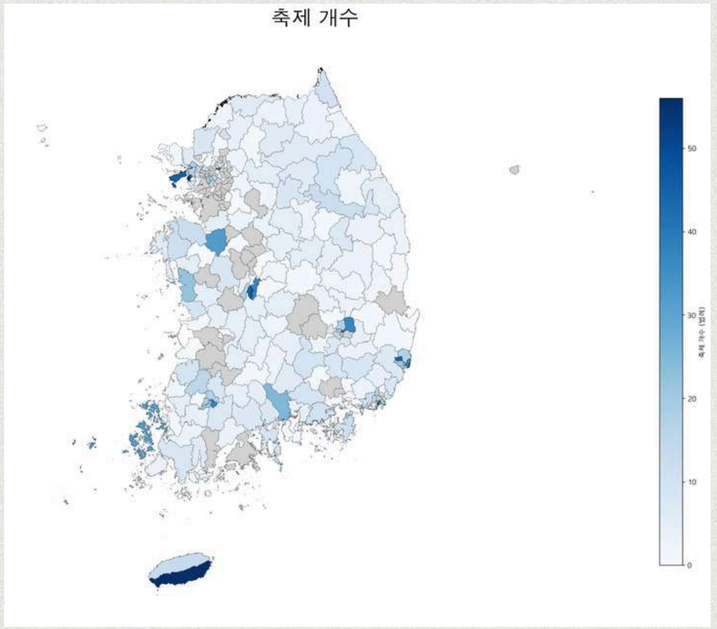
공간 데이터 시각화 (Choropleth Maps)

전국 기초지자체를 기준으로 다음 4가지 변수를 시각화, **지역 간 비교 가능**

관광지당 숙박업소 평균



지역 축제 개수



데이터 시각화 (2/3)

방문자수 상위/ 하위 관광지 키워드 시각화(워드 클라우드)

상위 100곳 관광지 키워드

하위 100곳 관광지 키워드

-> 아이와 체험하는 곳으로 박물관/ 미술관보다 놀이기구와 아쿠아리움 등의 **액티브한 장소가 방문자수 상위 관광지에서 발견 됨**

데이터 시각화 (2/3)

방문자수 상위/ 하위 관광지 키워드 시각화(워드 클라우드)

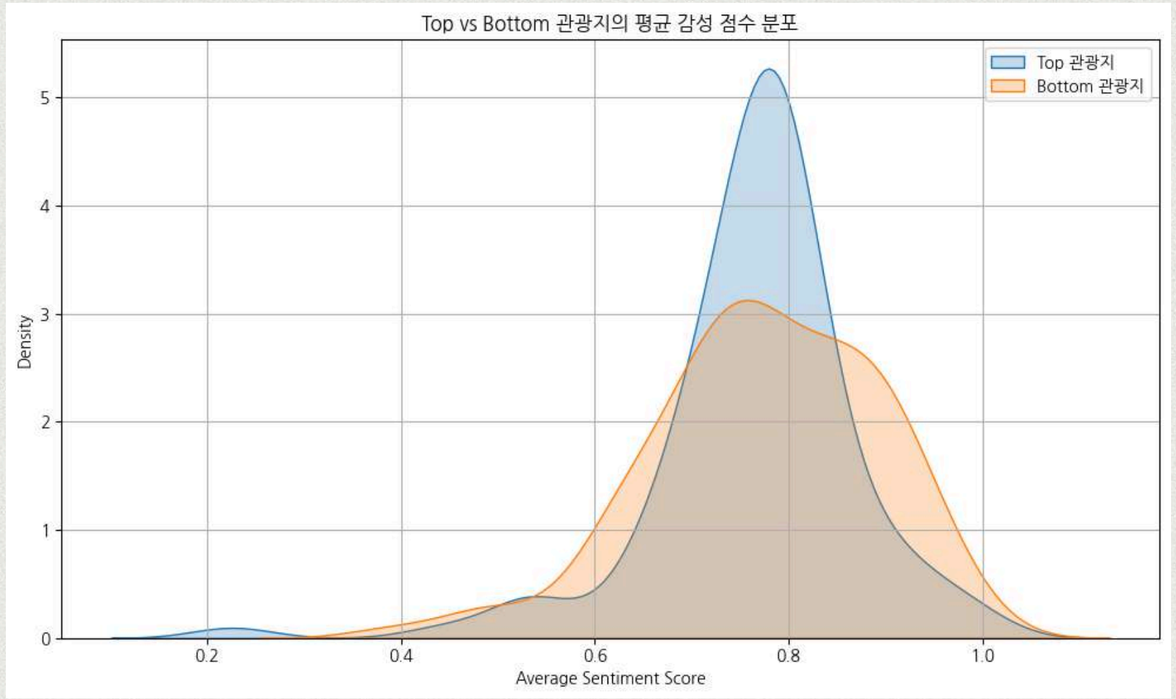
상위 100곳 관광지 키워드

하위 100곳 관광지 키워드

→ 긍정 표현보다, ‘무섭다’, ‘불친절하다’, ‘비싸다’, ‘시끄럽다’ 등의 **부정 표현이 하위 관광지에서 발견 됨.**

데이터 시각화 (3/3)

방문자수 상위/ 하위 관광지의 평균 감성 점수 분포



관광지의 리뷰별 감성점수를 관광지별로 합산 후, 평균값의 분포

[1] 상위 100곳 관광지(파란색)

- 분포 중심이 0.8 부근에 뚜렷하게 몰려 있음
- 곡선이 좁고 뾰족함 → 대다수 관광지에서 평균 감성이 고르게 긍정적임

→ 방문자 수가 많은 관광지는 전반적으로 꾸준히 긍정적인 경험을 제공하는 경향이 있음. 감성 편차가 적고, 기대가 안정적으로 충족됨.

[2] 하위 100곳 관광지(주황색)

- 분포가 넓고 평평, 더 퍼져 있음
- 특히 0.9 이상 영역에서 Top보다 밀도가 높음

→ 방문자 수가 적은 관광지 중 일부는 오히려 긍정적인 평가를 받고 있음, 하지만 **전체적으로는 감성 점수가 이질적으로 분포함 (긍정/부정/중립 모두 존재)**

04 결론 및 제언

주요 인사이트 및 마케팅 제안 (1/2)

[1] 감성 분석 언어 모델 실험 결과

- **관광객 방문 수가 많은 상위 관광지일수록 긍정 리뷰의 분포가 더 높게 나타나고, 하위 관광지는 부정에 치우치는 경향이 뚜렷함**
- 한글 감성분석 모델을 쓸수록 리뷰 감성의 상위/ 하위 관광지 차이가 극명하게 드러남
- 리뷰 키워드 중시/ 정량 평가 점수 위주, 연속 분포 등 산출하고 싶은 데이터 특징에 따라 모델을 다르게 사용할 수 있음

[2] 관광객 유입 요인

- **분석 결과, 긍정적인 리뷰와 안내소 접근성이 관광객 유입에 가장 큰 영향을 주는 요소임을 확인**

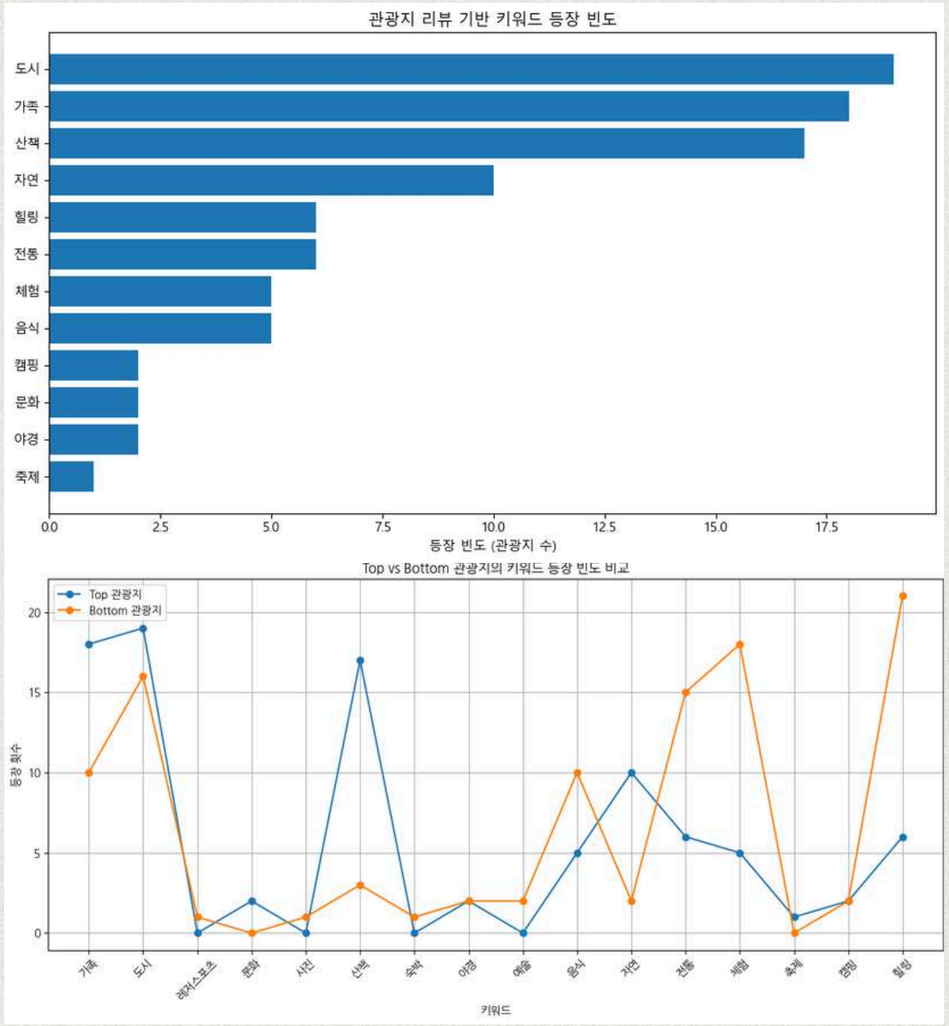
안내소의 추가 설치 및 운영시간 확대는 관광객 유치 및 증가에 유의미한 영향을 줄 것으로 판단됨

주요 인사이트 및 마케팅 제안 (2/2)

[3] 키워드로 보는 여행 트렌드

- ‘도시’, ‘가족’, ‘산책’, 과 관련된 관광지의 리뷰수가 절대적으로 많고, ‘축제’, ‘야경’ 키워드는 상대적으로 낮은 빈도로 언급됨. 축제 혹은 야경 중심 관광지보다, 도시 근처의 가족과 산책 가능한 관광지가 국내 여행 주요 트렌드라는 점을 알 수 있음

- 관광지마다 명확한 키워드가 있으며, 관광지의 테마 마케팅에 응용 가능



주요 인사이트 및 마케팅 제안 (2/2)

[4] 숨겨진 관광지

- 관광객이 많지 않은 관광지 중 리뷰 감성 분석 결과가 0.9 이상인 관광지 리스트

1. 거북이마을
2. 담양 커피농장
3. 박물관 휴르
4. 구포국수체험관
5. 한국등잔박물관

한계점

1. 사용자 특성 혹은 상황 고려 부족

- 리뷰 작성자의 연령대, 여행 목적, 계절적 요인(예: 여름에는 자연, 겨울에는 실내 위주 등) 같은 맥락을 고려하지 않음
- 특정 키워드가 일시적으로 많거나 적을 수 있음 (예: 해당 연도에 짧게 유행한 관광지 혹은 인기 있던 키워드가 반영될 수 있음
- 리뷰 이벤트를 진행한 경우 리뷰 데이터에 노이즈로 작용할 수도 있음

2. 단순화된 해석 가능성

- 리뷰 수가 많다고 해서 반드시 그 주제가 '더 인기' 있다는 해석은 과도한 일반화일 수 있음
- 예: '숙박'은 필수 요소라 별도로 언급하지 않는 사용자도 많을 수 있음 → 리뷰 언급이 적다고 해서 관심이 낮다고 보긴 어려움

감사합니다