2024.11.6.

숙박업소 데이터 수집 분석 및 관련 시각화 대시보드 구축

김도영 김동희 김신웅 김학성 최범준 현승현

Team6 육아일기

CONTENTS

프로젝트 개요

프로젝트 개요

- 지역별, 유형별, 날짜별로 분류된 숙박업소 데이터를 수집, 저장, 분석 및 시각화를 통해 전체적인 파이프라인 및 대시보드를 구축하고자 함.

주제 선정 이유

- 소비자들이 특정 지역, 날짜, 유형에 대한 숙박 트렌드 변화를 사전에 파악하며 소비자들에게 분석된 정보를 제공하여 전략적 대응에 도움을 줄 수 있음.
- 많은 국내 사용자들에게 친숙한 '여기어때' 앱을 활용하여 데이터를 분석하는 것은 국내 숙박업계 트렌드를 효과적으로 파악할 수 있음.

활용기술 및 프로젝트 아키텍처

Language

Python, SQL

Crawling

EventBridge, Lambda, Selenium

Data Pipeline

Boto3, Parquet, S3, Snowflake

BI

Preset

Collaboration

Github, Slack, ZEP, Notion

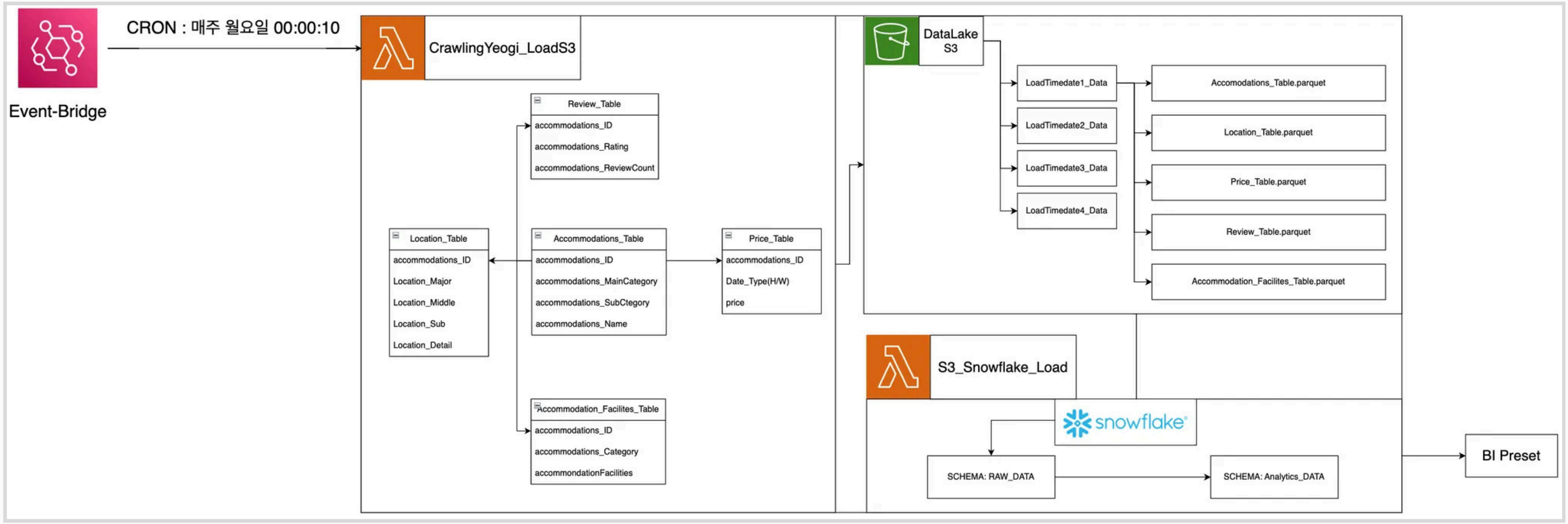









활용기술 및 프로젝트 아키텍처



프로젝트 절차

데이터 수집(Extract)

Selenium 등을 이용한 웹 스크래핑을 통해
다양한 종류의 숙박업소 데이터를 수집

데이터 적재(Load)

변환한 데이터들을 AWS S3
Data Lake에 적재

데이터 변환(Transform)

저장 및 분석이 용이한 Parquet 파일로
변환하여 데이터의 품질 향상

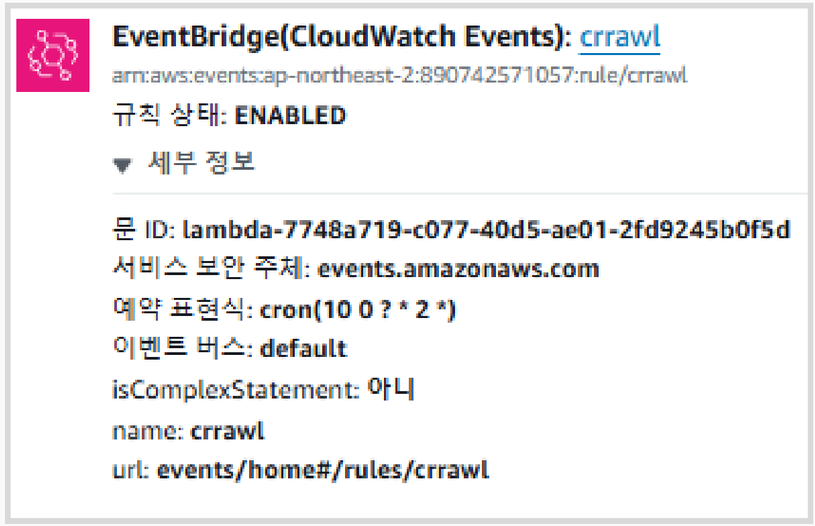
데이터 분석 및 시각화

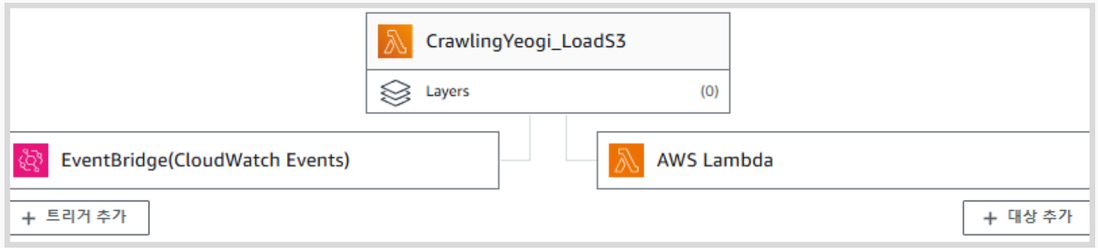
Snowflake으로 데이터 로드 후 분석
분석 결과를 Preset으로 시각화

프로젝트 절차

01 데이터 수집 — 02 데이터 변환 — 03 데이터 적재 — 04 분석 및 시각화

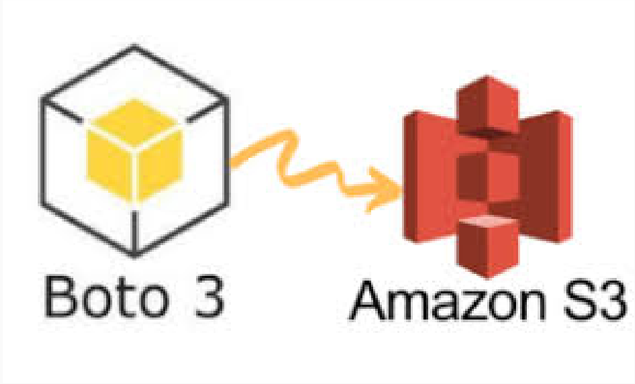
- 데이터 소스에서 Raw data를 수집하며, AWS EventBridge를 활용한 자동화된 이벤트 기반 주기로 AWS Lambda가 크롤링 작업을 수행
- Selenium으로 동적 컨텐츠를 로드하여, 특정 HTML 요소의 데이터를 추출하여 리스트로 저장





프로젝트 절차

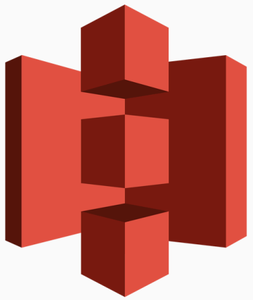
01 데이터 수집 — 02 데이터 변환 — **03 데이터 적재** — 04 분석 및 시각화



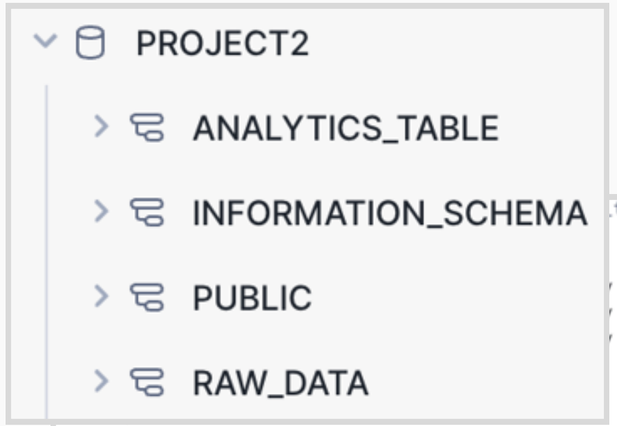
- 변환된 데이터는 Boto3를 활용해 AWS S3에 업로드
- 변환된 Parquet 파일은 S3에 업로드하여 효율적인 쿼리 및 분석을 준비

프로젝트 절차

01 데이터 수집 — 02 데이터 변환 — **03 데이터 적재** — 04 분석 및 시각화



- S3로부터 분석에 필요한 parquet 데이터를 Snowflake에 연동 및 업로드 수행
- Snowflake 분석 및 시각화 작업을 위한 데이터 웨어하우스로 활용



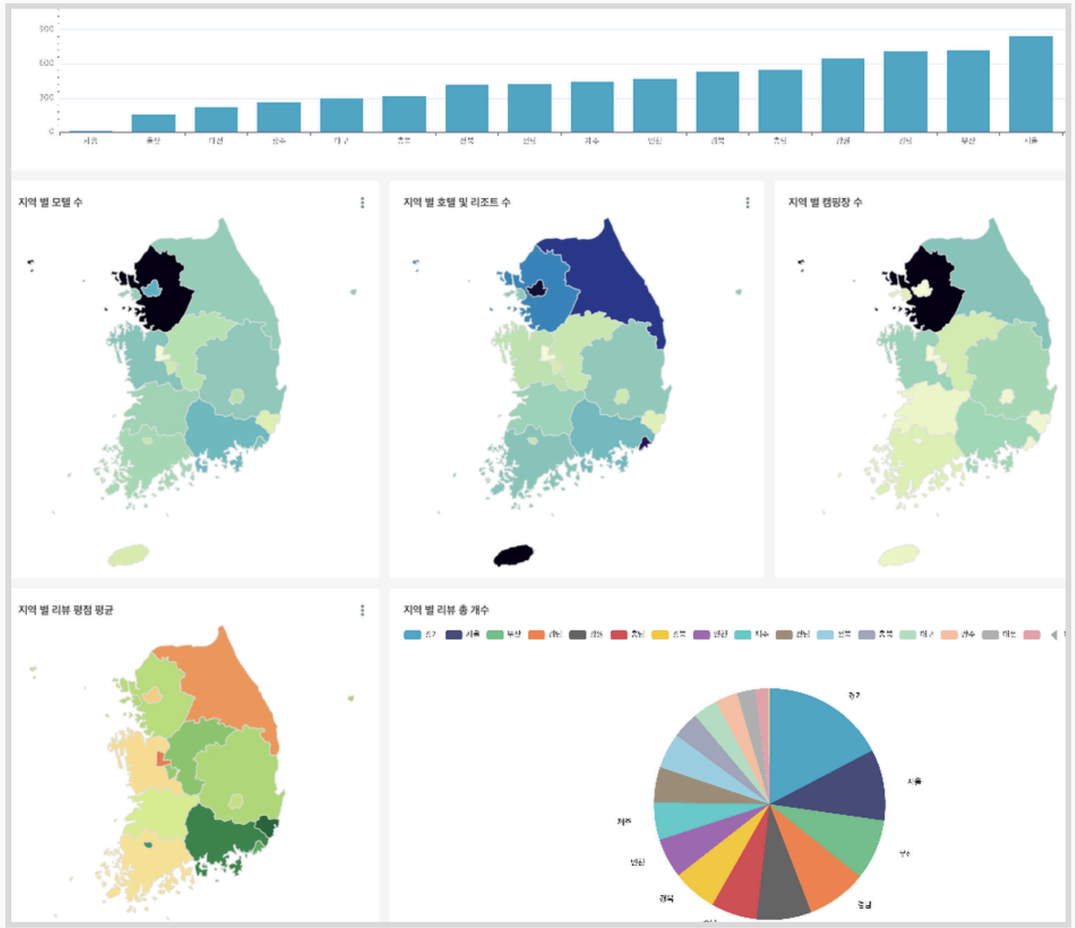

```
table.loc_type_counts AS

= 'Motel' THEN 1 ELSE 0 END) AS "모텔",
= 'Hotel/Resort' THEN 1 ELSE 0 END) AS "호텔/리조트",
= 'Camping' THEN 1 ELSE 0 END) AS "캠핑"
```

```
    SELECT DISTINCT
        accommodation_id,
        accommodation_location_middle,
        CASE
            WHEN accommodation_location_major IN ('전북', '전북특별자치도') THEN '전북'
            WHEN accommodation_location_major IN ('강원', '강원특별자치도', '강원도') THEN '강원'
            WHEN accommodation_location_major IN ('제주도', '제주특별자치도') THEN '제주도'
            WHEN accommodation_location_major IN ('세종', '세종특별자치시') THEN '세종'
            WHEN accommodation_location_major IN ('경상북도', '경북') THEN '경북'
            ELSE accommodation_location_major
        END AS accommodation_location_major
    FROM project2.raw_data.accommodation_location
) l ON a.accommodation_id = l.accommodation_id
GROUP BY
    l.accommodation_location_major
ORDER BY
    l.accommodation_location_major;
```

프로젝트 절차

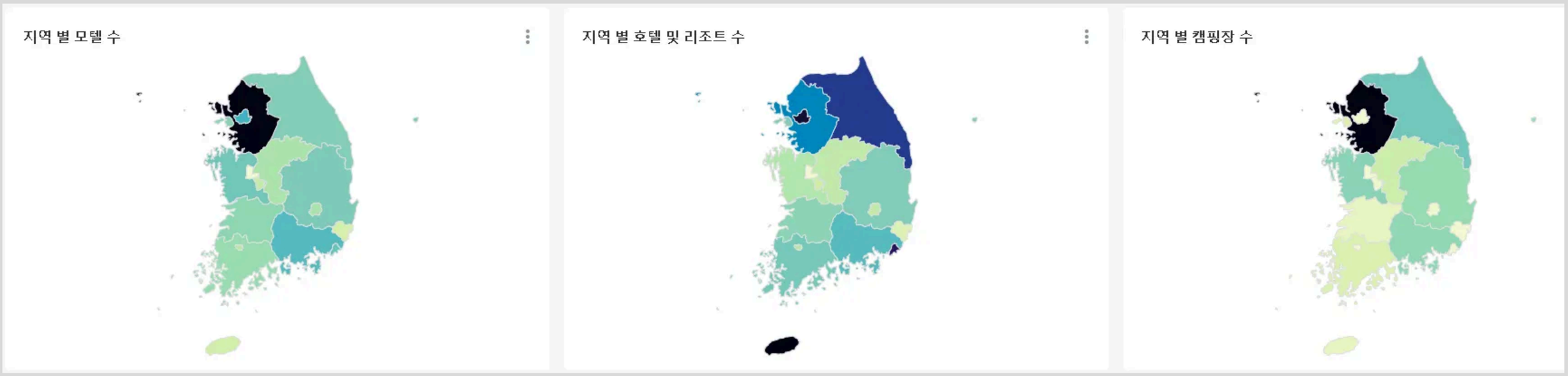
01 데이터 수집 — 02 데이터 변환 — 03 데이터 적재 — 04 분석 및 시각화



- 데이터를 분석하고 결과를 시각적으로 표현
- Preset을 활용하여 데이터 시각화 대시보드 및 차트 생성

프로젝트 주요 결과

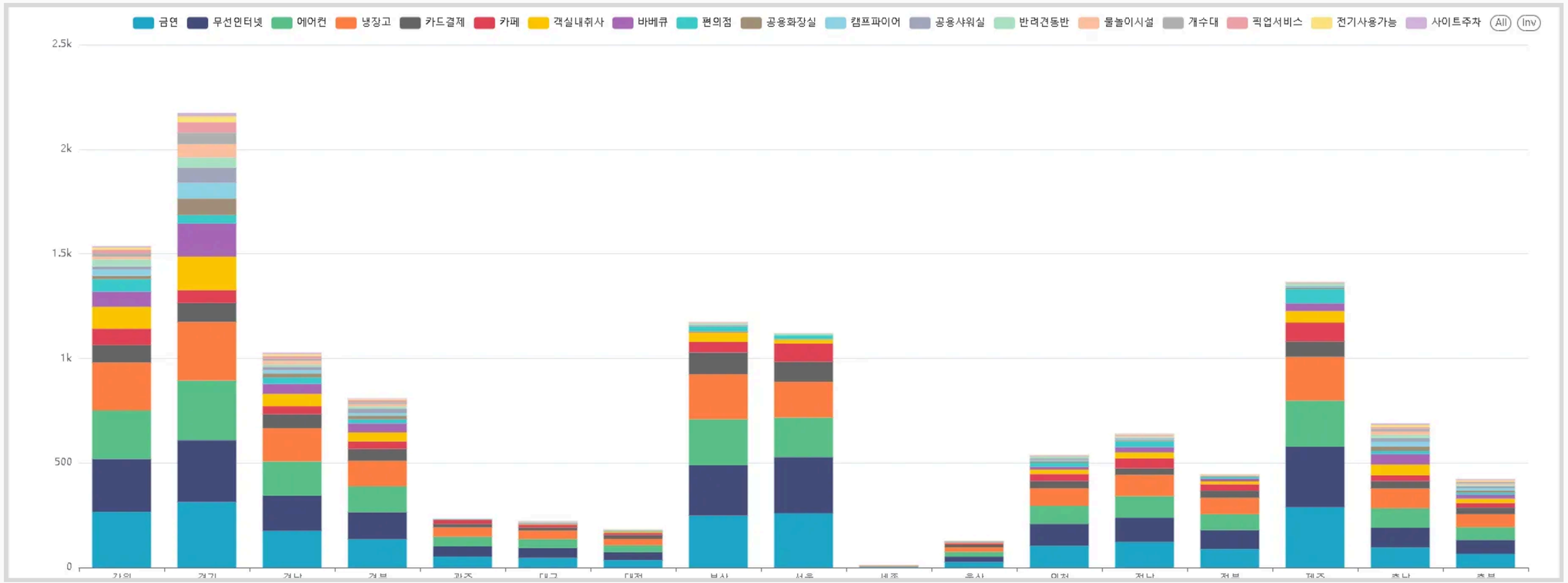
| 지역별 분석



[모텔 – 경기, 호텔/리조트 – 제주/서울/부산, 캠핑 – 경기] 지역에 가장 많음

프로젝트 주요 결과

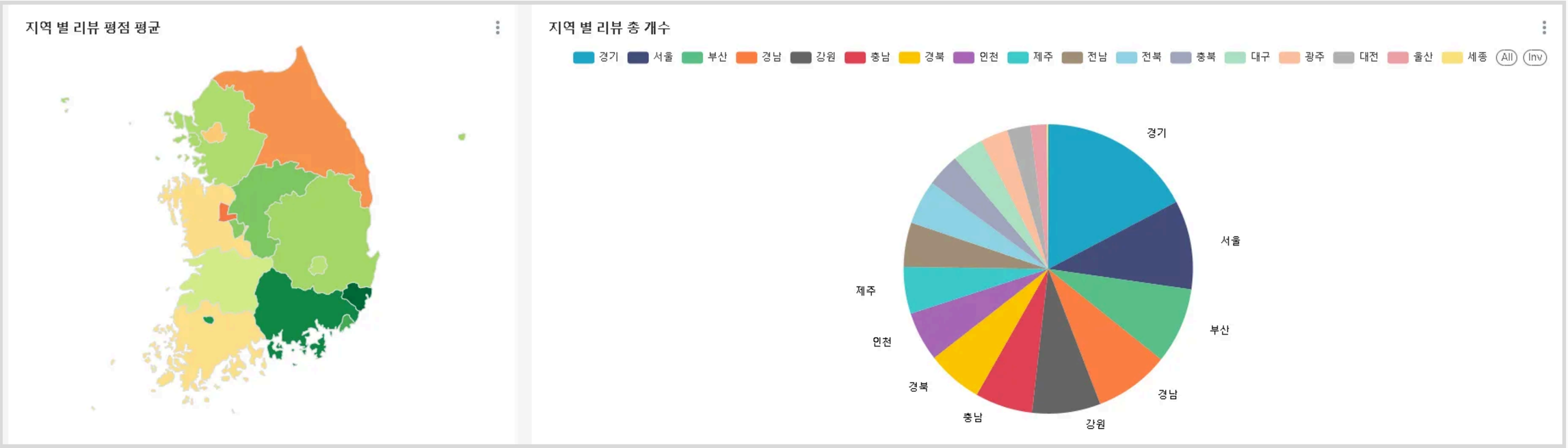
| 지역별 분석



전체적으로 금연, 무선인터넷, 에어컨, 냉장고의 비율이 높음
경기, 강원, 제주, 부산, 서울의 부대시설이 가장 많음

프로젝트 주요 결과

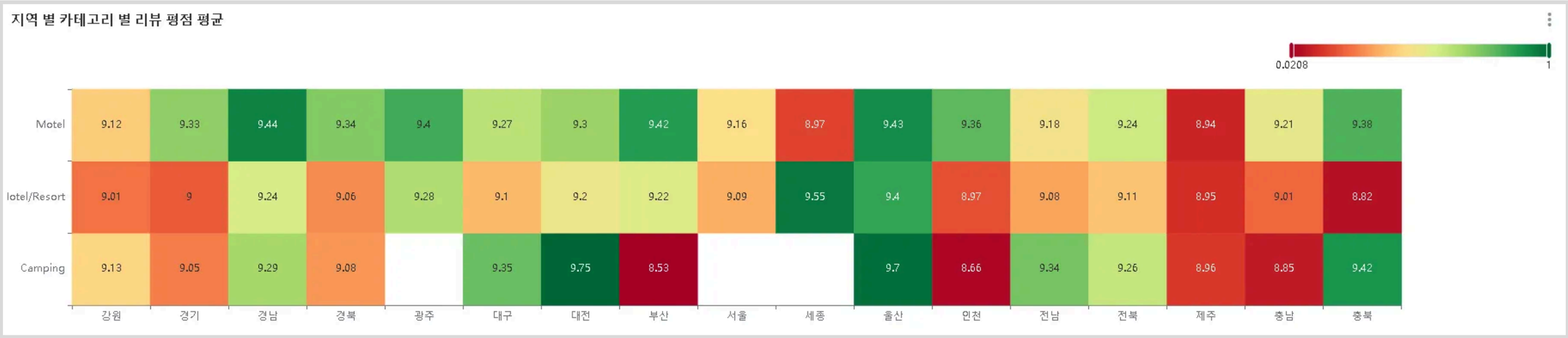
| 지역별 분석



동남부 지방(부산, 경남, 경북)의 리뷰 평점이 높음
리뷰의 개수는 경기, 서울, 부산, 경남, 강원 등 순으로 많음

프로젝트 주요 결과

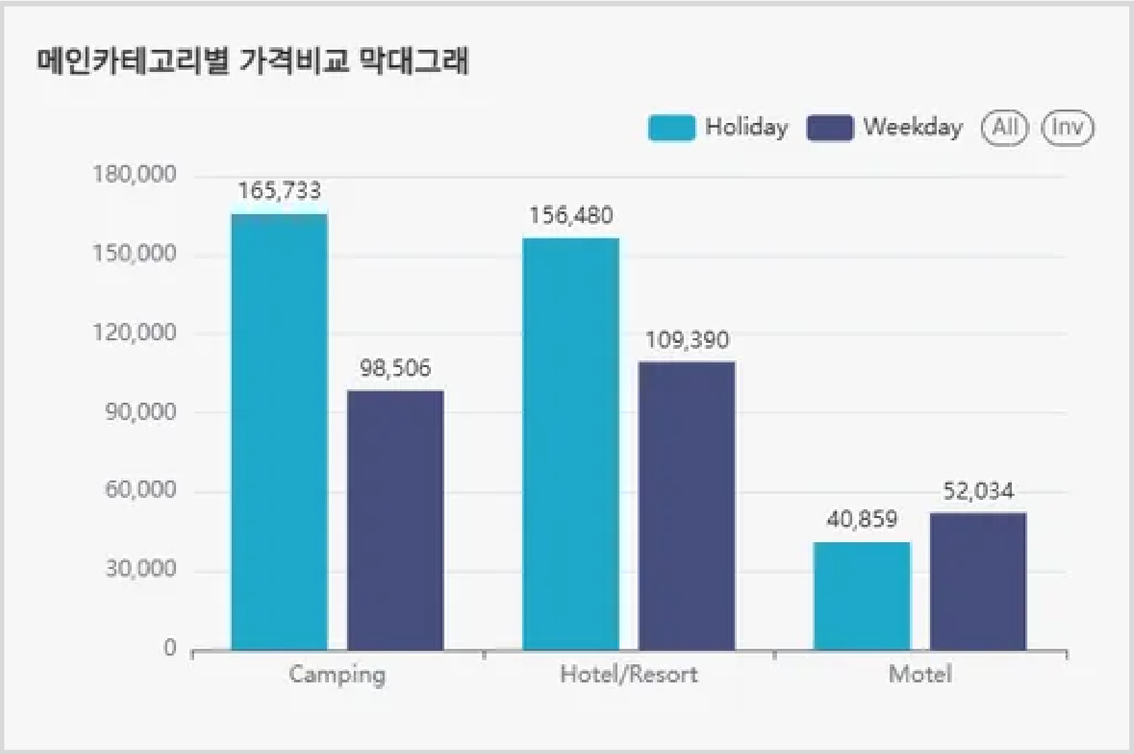
| 지역별 분석



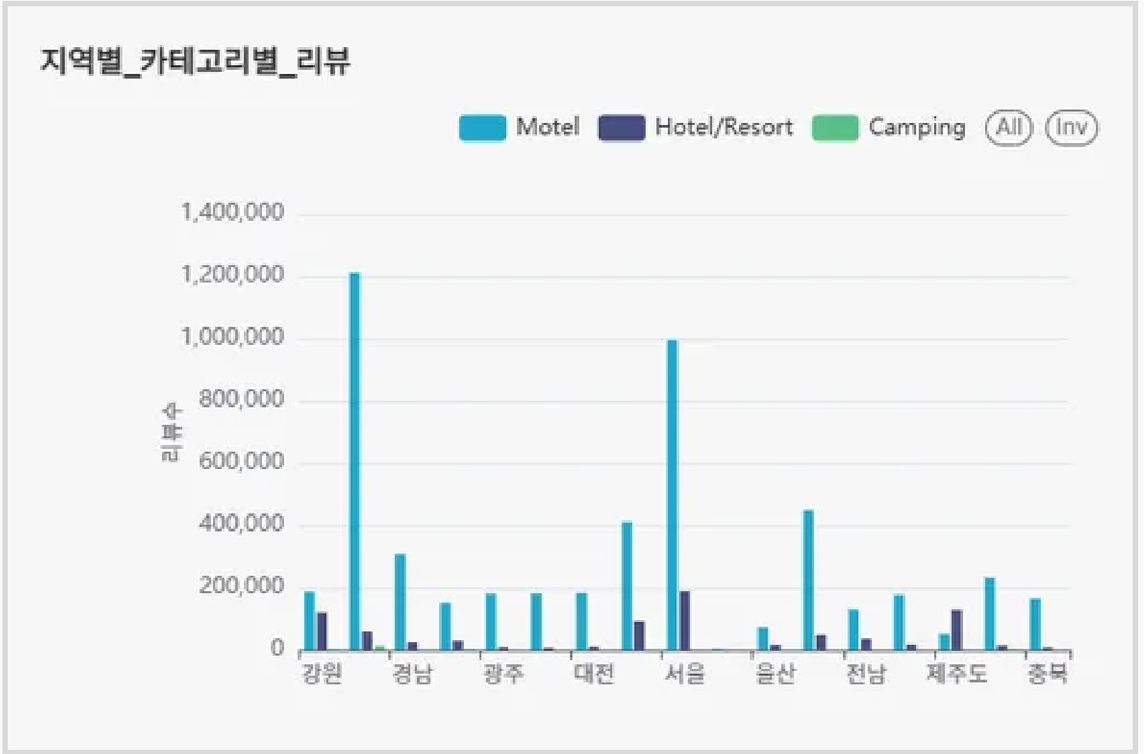
평점이 좋은 숙소는 카테고리에 따라
모텔 – 경남/부산/울산, 호텔/리조트 – 세종/울산, 캠핑장 – 대전.울산임을 알 수 있음

프로젝트 주요 결과

| 메인 카테고리별 분석



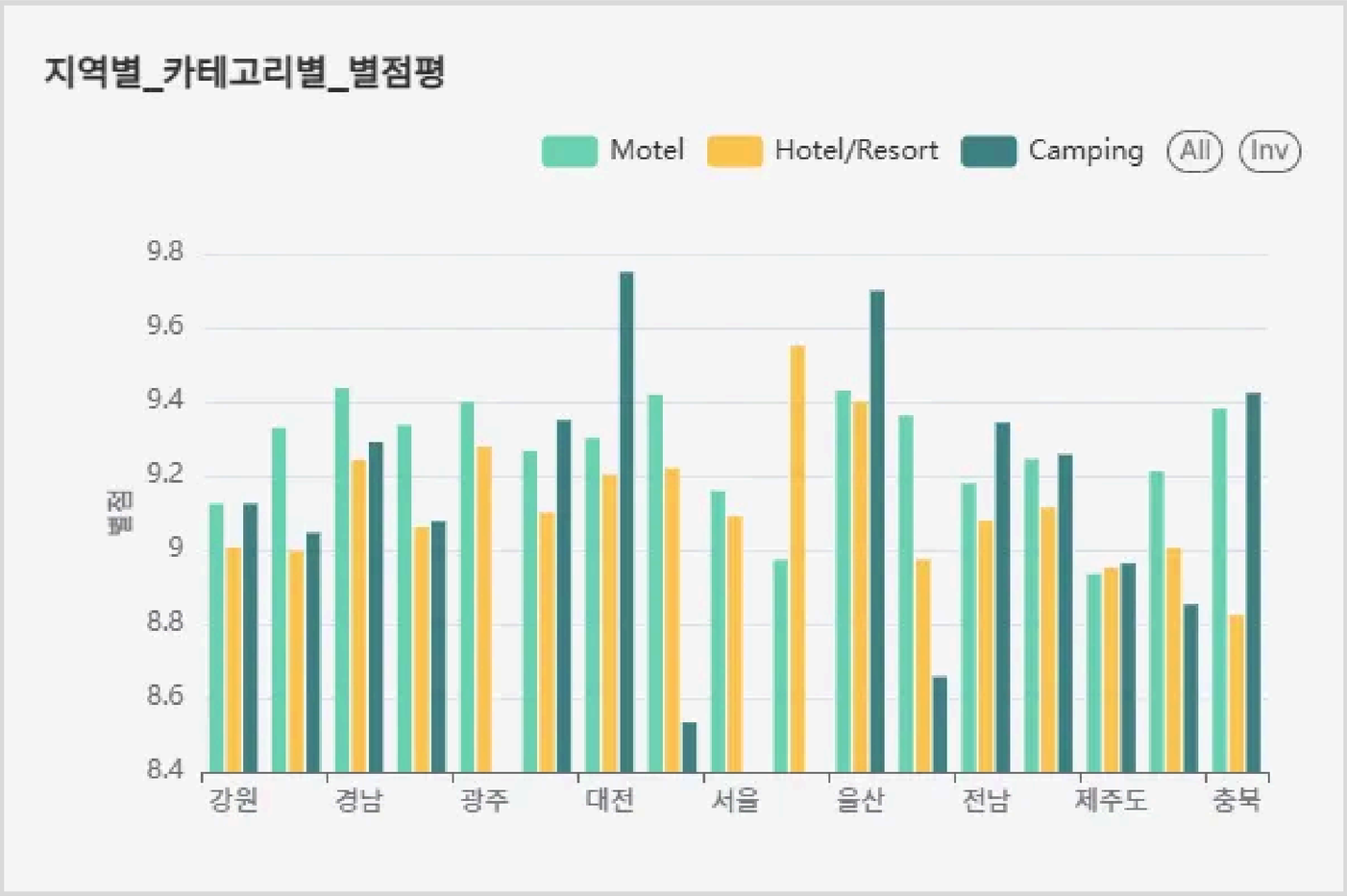
캠핑의 경우 주말/주중 간의 가격 편차가 가장 큼
모텔의 경우는 주중이 주말보다 평균 가격이 더 비쌈



리뷰수의 경우, '경기도 – 모텔'이 가장 많음

프로젝트 주요 결과

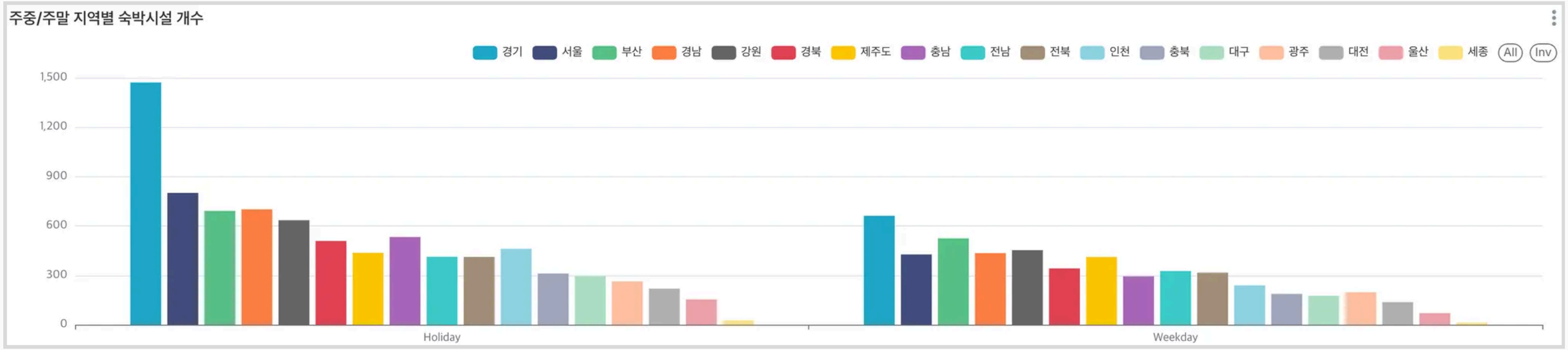
| 메인 카테고리별 분석



평점은 대부분 9점에 가깝게
분포되어 있으나,
‘제주도’의 경우에는 어느 카테고리에서도
9점 이하의 평점을 보임

프로젝트 주요 결과

| 주중/주말별 분석



17개 지역 전반적으로 주중보다 주말에 예약 가능함을 알 수 있음

회고 및 개선점

회고

- 1차 프로젝트에서 진행했던 데이터 스크래핑에서 자동화할 수 있는 EventBridge와 Lambda를 추가 활용하며 스크래핑을 자동화할 수 있었고, 추후에 실시간으로 변하는 데이터 수집에 활용할 수 있는 기술 또한 적용해보고 싶다.
- 분석 목적에 맞는 데이터 수집과, 이를 활용할 수 있는 데이터 정제 과정이 중요하다는 것을 깨달았다. Preset 대시보드를 구축하는 과정을 통해 다양한 분석 결과를 직관적으로 확인할 수 있었다.
- 현업에서 사용하는 AWS의 다양한 기술들과 Snowflake, Preset 등을 다뤄볼 수 있는 시간이어서 좋았다.

개선점

- 웹 사이트의 구조가 변경되거나 보안 기능이 강화될 경우 스크래핑 방식에 대한 조정이 필요할 수 있다.
- 스크래핑한 데이터 중에서 간혹 누락된 정보로 인해 품질 유지가 어려울 가능성이 있어, 데이터의 신뢰성 확보를 위한 기술의 도입이 필요할 수 있다.
- 오픈소스인 Apache Airflow 등을 사용한다면 비용 부담이 적으며 더 유연하고 복잡한 워크플로우를 관리할 수 있다.
- EventBridge와 Lambda를 활용한 주기적인 업데이트를 통해 최신성을 확보하였으나, 예상치 못한 숙소 변동이 있을 수 있기에 Apache Kafka와 같은 실시간성을 유지할 수 있는 기술을 도입할 수 있다.

2024. 11. 6.

감사합니다

Team6 육아일기