King Saud University

College of Computer and Information Sciences

Department of Information Technology

جامعة الملك سعود
King Saud University

كلية علوم الحاسب والمعلومات
قسم تقنية المعلومات

Afaq | آفاق

2nd semester 1446

Prepared by:

Team Members	ID
Ghadeer Alnuwaysir	444200420
Maha Alruwais	444200749
Norah Almadhi	44420089
Rana Albridi	444201094
Lamees Alghamdi	444201177

Supervised by:
Dr. Mashael Sultan Aldayel

Table of Contents

Introduction

Saudi Arabia is a tourism destination with immense untapped potential, offering a remarkable variety of breathtaking natural landscapes, historical sites, and vibrant cities. Despite this rich cultural and natural heritage, many unique destinations remain under-explored and under-promoted, making it challenging for visitors to discover the hidden gems of the Kingdom. Additionally, the lack of data and clear insights into the factors that make these destinations attractive hinders their development and promotion.

This study aims to utilize data available on the Flickr platform. By collecting and analyzing these images and their associated metadata, we can highlight the most attractive destinations, examine the seasonal patterns of their popularity, and understand the visual elements that influence audience engagement. This study will contribute to raising awareness of Saudi Arabia's beauty and diversity while supporting the development of tourism in a way that highlights the value of these destinations and promotes their sustainability.

Main Research Question:

Which attractions in Saudi Arabia deserve more attention based on public photo data?

Objectives:

Main Research Question:

Which attractions in Saudi Arabia deserve more attention based on public photo data?

Seasonal Question

How does the popularity of these destinations vary across different seasons?

Time-Specific Question

What time of day are these places more appealing based on photo interactions?

Visual Elements Question

What visual elements (trees, buildings, animals, rivers…) are most frequently present in photos of popular destinations?

Destination-Based Question

Which types of destinations (historical sites, cities, natural landscapes…) are most favored by visitors?

Region-Based Question

Which region of Saudi Arabia has the highest tourism appeal based on public photo data?

Data Description

Data Sources

Flickr[1] is a photo-sharing platform that provides an API to access image metadata, user interactions, and URLs, enabling data extraction for analyzing photography trends and user preferences.

- **Flickr API**: We used three main services:
  - getFavorites: To count photo favorites.
  - getInfo: To retrieve detailed photo metadata.
  - getComments: To collect user comments.
- **OpenAI CLIP**: Used to analyze and classify images by region type and visual elements.
- **OpenCV2**: Applied to distinguish between day and night images based on pixel brightness.
- **Sentiment Analysis (Transformers)**: Used to classify comments as positive, negative, or neutral, with a scoring function to summarize overall sentiment.

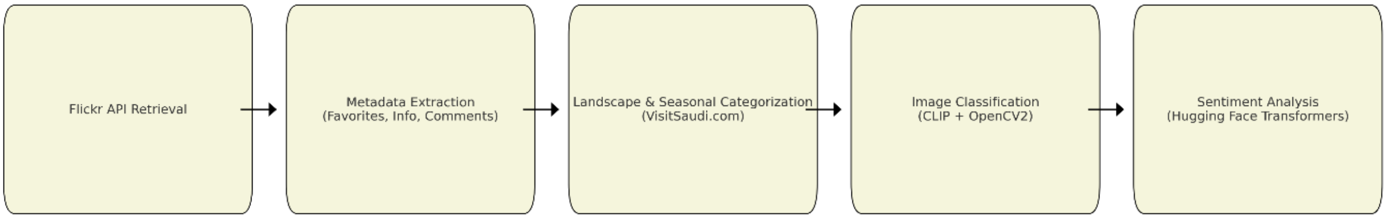
Other Resources:

We used **VisitSaudi[2]**, Saudi Arabia’s official tourism website, to guide the structuring of landscape types, visual elements, and seasonal classifications in our project.

- **Landscape and Attractions**:
  Destinations were categorized into historical, natural, and cultural types based on Visit Saudi’s structure. Key visual elements (e.g., trees, sky, mountains) were identified through analyzing official attraction images.
- **Seasonal Classification**:
  Images were assigned to seasons based on their capture date, following Visit Saudi’s seasonal definitions.

Data Collection:

Data Collection and Processing Workflow



1. Flickr API Retrieval:
The process begins by using Flickr's API to access raw data.

2. Metadata Extraction (Favorites, Info, Comments):
From the API responses, we extract essential metadata such as the number of favorites, detailed photo information (title, description, date uploaded), and all available user comments related to each photo.

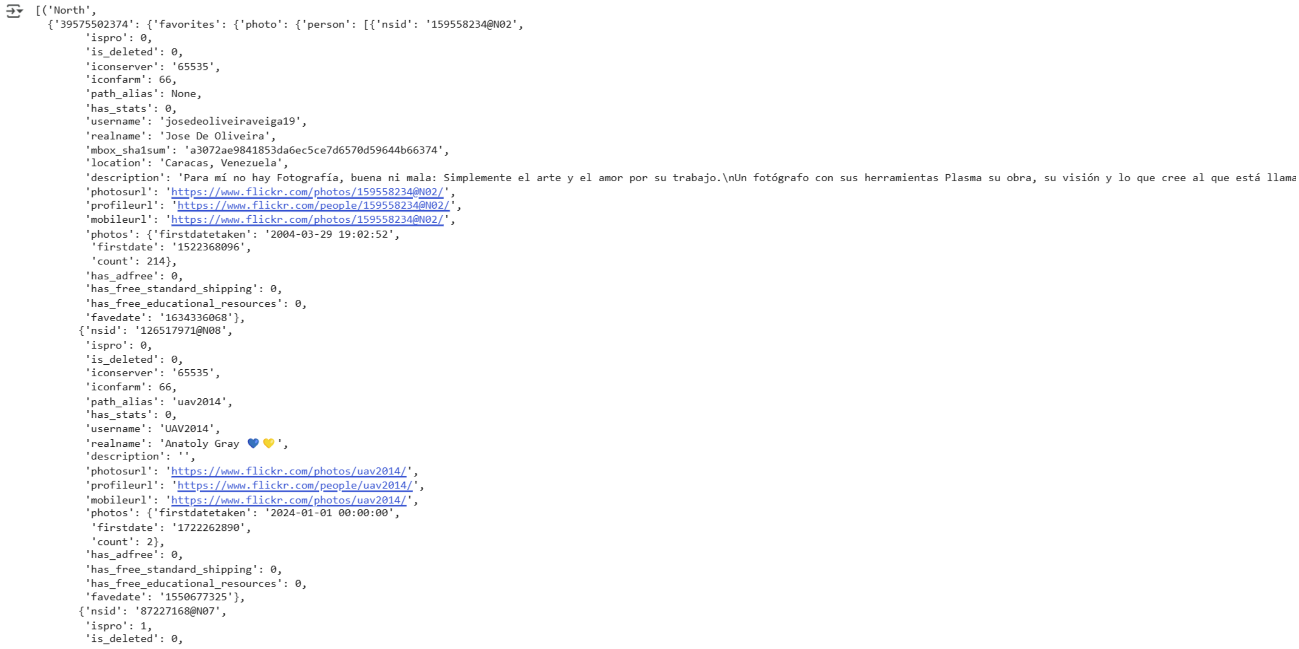
3. Landscape & Seasonal Categorization (VisitSaudi.com):
Using Visit Saudi, we categorize each photo based on its associated landscape type (Coastal, Urban, Historical, Nature, Desert) and assign a season based on the photo's date, aligning it with Saudi Arabia's seasonal periods.

4. Image Classification (CLIP[3] + OpenCV2[4]):
We apply OpenAI's CLIP model to detect visual elements ("sky", "water", "mountains", "buildings", "trees", "cars", "people", "animals", "sand", "grass", "clouds", "sun", "moon", "boats", "roads", "flowers", "statues", "rocks", "mosques", and "snow") within the images. Additionally, OpenCV2 analyzes image brightness to determine whether the photo was taken during the day or at night.

5. Sentiment Analysis (Hugging Face Transformers[5]):
Comments collected from Flickr are analyzed for sentiment using Hugging Face's Transformer models. Each comment is classified as positive, negative, or neutral, and a cumulative sentiment score is calculated for each photo.

Structured and Unstructured data samples

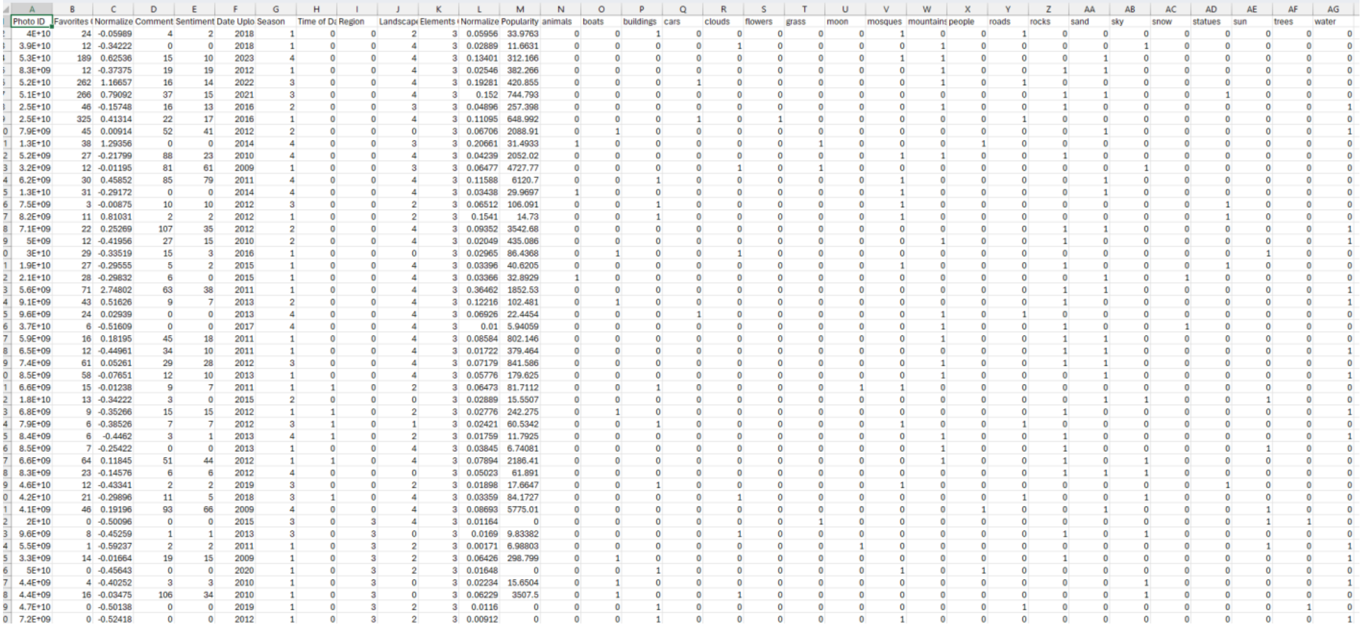
1. Unstructured



```
[('North',
  {'39575502374': {'favorites': {'photo': {'person': [{'nsid': '159558234@N02',
        'ispro': 0,
        'is_deleted': 0,
        'iconserver': '65535',
        'iconfarm': 66,
        'path_alias': None,
        'has_stats': 0,
        'username': 'josedeoliveiraveiga19',
        'realname': 'Jose De Oliveira',
        'mbox_sha1sum': 'a3072ae9841853da6ec5ce7d6570d59644b66374',
        'location': 'Caracas, Venezuela',
        'description': 'Para mí no hay Fotografía, buena ni mala: Simplemente el arte y el amor por su trabajo.\nUn fotógrafo con sus herramientas Plasma su obra, su visión y lo que cree al que está llama
        'photosurl': 'https://www.flickr.com/photos/159558234@N02/',
        'profileurl': 'https://www.flickr.com/people/159558234@N02/',
        'mobileurl': 'https://www.flickr.com/photos/159558234@N02/',
        'photos': {'firstdatetaken': '2004-03-29 19:02:52',
         'firstdate': '1522368096',
         'count': 214},
        'has_adfree': 0,
        'has_free_standard_shipping': 0,
        'has_free_educational_resources': 0,
        'favedate': '1634336068'},
       {'nsid': '126517971@N08',
        'ispro': 0,
        'is_deleted': 0,
        'iconserver': '65535',
        'iconfarm': 66,
        'path_alias': 'uav2014',
        'has_stats': 0,
        'username': 'UAV2014',
        'realname': 'Anatoly Gray 💙💛',
        'description': '',
        'photosurl': 'https://www.flickr.com/photos/uav2014/',
        'profileurl': 'https://www.flickr.com/people/uav2014/',
        'mobileurl': 'https://www.flickr.com/photos/uav2014/',
        'photos': {'firstdatetaken': '2024-01-01 00:00:00',
         'firstdate': '1722262890',
         'count': 2},
        'has_adfree': 0,
        'has_free_standard_shipping': 0,
        'has_free_educational_resources': 0,
        'favedate': '1550677325'},
       {'nsid': '87227168@N07',
        'ispro': 1,
        'is_deleted': 0,
```

The data is in JSON format and includes photo id, favorite photos, and related metadata.

2. Final Structured



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
Photo ID	Favorites (	Normalize	Comment	Sentiment	Date Uplo	Season	Time of Da	Region	Landscape	Elements (	Normalize	Popularity	animals	boats	buildings	cars	clouds	flowers	grass	moon	mosques	mountains	people	roads	rocks	sand	sky	snow	statues	sun	trees	water
4E+10	24	-0.05989	4	2	2018	1	0	0	2	3	0.05956	33.9763	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
3.9E+10	12	-0.34222	0	0	2018	1	0	0	4	3	0.02889	11.6631	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
5.3E+10	189	0.62536	15	10	2023	4	0	0	4	3	0.13401	312.166	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0
8.3E+09	12	-0.37375	19	19	2012	1	0	0	4	3	0.02546	382.266	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0
5.2E+10	262	1.16657	16	14	2022	3	0	0	4	3	0.19281	420.855	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
5.1E+10	266	0.79092	37	15	2021	3	0	0	4	3	0.152	744.793	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0
2.5E+10	46	-0.15748	16	13	2016	2	0	0	3	3	0.04896	257.398	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1
2.5E+10	325	0.41314	22	17	2016	1	0	0	4	3	0.11095	648.992	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
7.9E+09	45	0.00914	52	41	2012	2	0	0	0	3	0.06706	2088.91	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
1.3E+10	38	1.29356	0	0	2014	4	0	0	3	3	0.20661	31.4933	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
5.2E+09	27	-0.21799	88	23	2010	4	0	0	4	3	0.04239	2052.02	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0
3.2E+09	12	-0.01195	81	61	2009	1	0	0	3	3	0.06477	4727.77	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
6.2E+09	30	0.45852	85	79	2011	4	0	0	4	3	0.11588	6120.7	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
1.3E+10	31	-0.29172	0	0	2014	4	0	0	4	3	0.03438	29.9697	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
7.5E+09	3	-0.00875	10	10	2012	3	0	0	2	3	0.06512	106.091	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
8.2E+09	11	0.81031	2	2	2012	1	0	0	2	3	0.1541	14.73	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
7.1E+09	22	0.25269	107	35	2012	2	0	0	4	3	0.09352	3542.68	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1
5E+09	12	-0.41956	27	15	2010	2	0	0	4	3	0.02049	435.086	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1
3E+10	29	-0.33519	15	3	2016	1	0	0	0	3	0.02965	86.4368	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
1.9E+10	27	-0.29555	5	2	2015	1	0	0	4	3	0.03396	40.6205	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0
2.1E+10	28	-0.29832	6	0	2015	1	0	0	4	3	0.03366	32.8929	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
5.6E+09	71	2.74802	63	38	2011	1	0	0	4	3	0.36462	1852.53	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1
9.1E+09	43	0.51626	9	7	2013	2	0	0	4	3	0.12216	102.481	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
9.6E+09	24	0.02939	0	0	2013	4	0	0	4	3	0.06926	22.4454	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
3.7E+10	6	-0.51609	0	0	2017	4	0	0	4	3	0.01	5.94059	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0
5.9E+09	16	0.18195	45	18	2011	1	0	0	4	3	0.08584	802.146	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0
6.5E+09	12	-0.44961	34	10	2011	1	0	0	4	3	0.01722	379.464	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1
7.4E+09	61	0.05261	29	28	2012	3	0	0	4	3	0.07179	841.586	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0
8.5E+09	58	-0.07651	12	10	2013	1	0	0	4	3	0.05776	179.625	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1
6.6E+09	15	-0.01238	9	7	2011	1	1	0	2	3	0.06473	81.7112	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
1.8E+10	13	-0.34222	3	0	2015	2	0	0	0	3	0.02889	15.5507	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0
6.8E+09	9	-0.35266	15	15	2012	1	1	0	2	3	0.02776	242.275	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
7.9E+09	6	-0.38526	7	7	2012	3	1	0	1	3	0.02421	60.5342	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
8.4E+09	6	-0.4462	3	1	2013	4	1	0	2	3	0.01759	11.7925	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1
8.5E+09	7	-0.25422	0	0	2013	1	0	0	4	3	0.03845	6.74081	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0
6.6E+09	64	0.11845	51	44	2012	1	1	0	4	3	0.07894	2186.41	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0
8.3E+09	23	-0.14576	6	6	2012	4	0	0	0	3	0.05023	61.891	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
4.6E+10	12	-0.43341	2	2	2019	3	0	0	2	3	0.01898	17.6647	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
4.2E+10	21	-0.29896	11	5	2018	3	1	0	4	3	0.03359	84.1727	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0
4.1E+09	46	0.19196	93	66	2009	4	0	0	4	3	0.08693	5775.01	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0
2E+10	0	-0.50096	0	0	2015	3	0	3	4	3	0.01164	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0
9.6E+09	8	-0.45259	1	1	2013	3	0	3	0	3	0.0169	9.83382	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0
5.5E+09	1	-0.59237	2	2	2011	1	0	3	2	3	0.00171	6.98803	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1
3.3E+09	14	-0.01664	19	15	2009	1	0	3	2	3	0.06426	298.799	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
5E+10	0	-0.45643	0	0	2020	1	0	3	2	3	0.01648	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
4.4E+09	4	-0.40252	3	3	2010	1	0	3	0	3	0.02234	15.6504	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
4.4E+09	16	-0.03475	106	34	2010	1	0	3	0	3	0.06229	3507.5	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
4.7E+10	0	-0.50138	0	0	2019	1	0	3	2	3	0.0116	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
7.2E+09	0	-0.52418	0	0	2012	1	0	3	2	3	0.00912	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1

Feature Summary:

The final dataset consists of (199 observations, 33 features, and different data types like int64, float64).

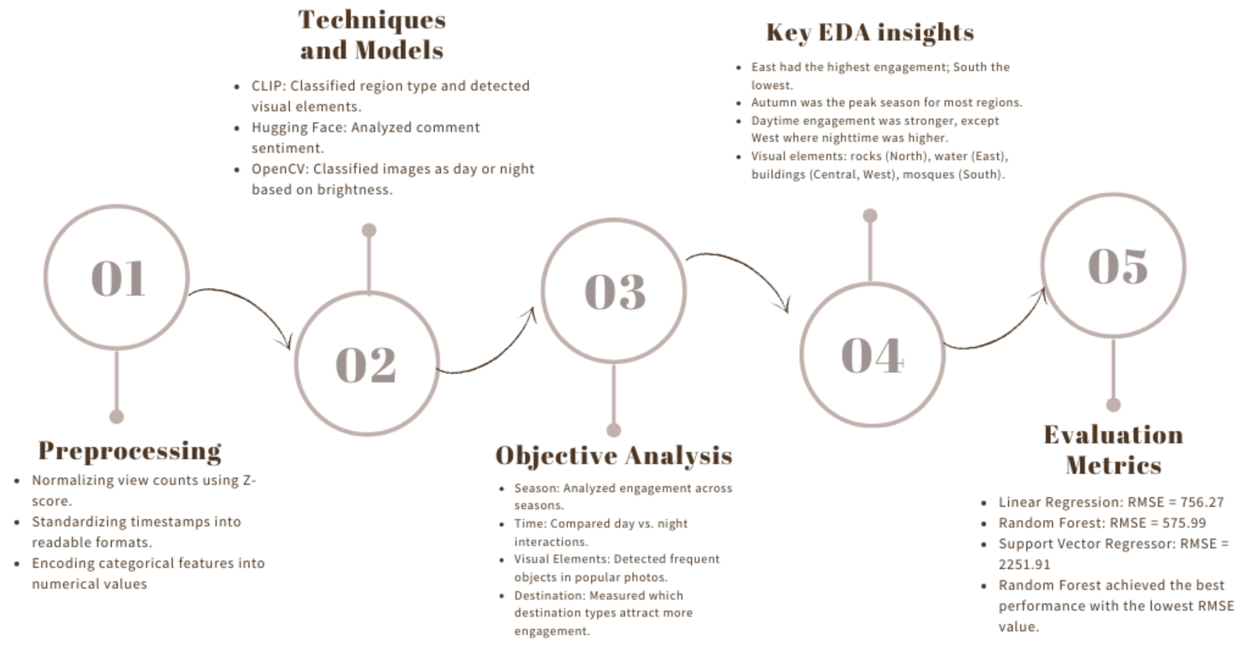
Feature	Data Type	Non-Null Count
Photo ID	int64	199
Favorites Count	int64	199
Normalized View Count	float64	199
Comments Count	int64	199
Sentiment Score	int64	199
Date Uploaded	int64	199
Season	int64	199
Time of Day	int64	199
Region	int64	199
Landscape Type	int64	199
Elements Count	int64	199
Normalized View Count (0-1)	float64	199
Popularity Score	float64	199
animals	int64	199
boats	int64	199
buildings	int64	199
cars	int64	199
clouds	int64	199
flowers	int64	199
grass	int64	199
moon	int64	199
mosques	int64	199
mountains	int64	199
people	int64	199

roads	int64	199
rocks	int64	199
sand	int64	199
sky	int64	199
snow	int64	199
statues	int64	199
sun	int64	199
trees	int64	199
water	int64	199

Method

To tackle our research questions, we adopted a multi-stage methodology involving data preprocessing, feature extraction, exploratory analysis, and model evaluation.

The flowchart below summarizes the entire process — from retrieving and transforming raw photo data to applying AI models and generating insights based on public engagement with tourism images.



Preprocessing

- Normalizing view counts using Z-score.
- Standardizing timestamps into readable formats.
- Encoding categorical features into numerical values

Techniques and Models

- CLIP: Classified region type and detected visual elements.
- Hugging Face: Analyzed comment sentiment.
- OpenCV: Classified images as day or night based on brightness.

Objective Analysis

- Season: Analyzed engagement across seasons.
- Time: Compared day vs. night interactions.
- Visual Elements: Detected frequent objects in popular photos.
- Destination: Measured which destination types attract more engagement.

Key EDA insights

- East had the highest engagement; South the lowest.
- Autumn was the peak season for most regions.
- Daytime engagement was stronger, except West where nighttime was higher.
- Visual elements: rocks (North), water (East), buildings (Central, West), mosques (South).

Evaluation Metrics

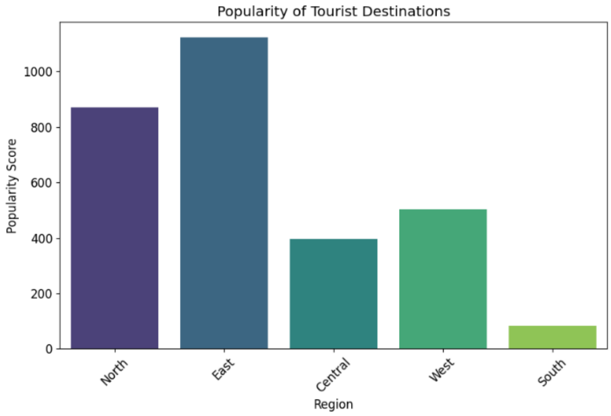
- Linear Regression: RMSE = 756.27
- Random Forest: RMSE = 575.99
- Support Vector Regressor: RMSE = 2251.91
- Random Forest achieved the best performance with the lowest RMSE value.

This visual captures the key techniques, analytical steps, and performance metrics used throughout our study.

Results and Discussions

Our analysis provided key insights across multiple dimensions relevant to tourism engagement in Saudi Arabia, using a dataset of public Flickr images.

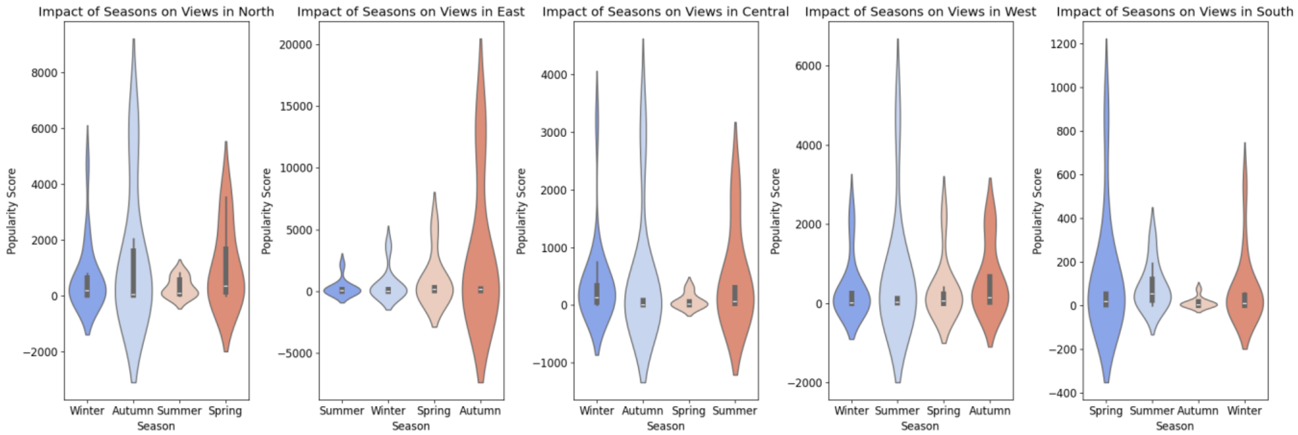
1. **Region Question: Which region of Saudi Arabia has the highest tourism appeal?** We evaluated each region's overall popularity score based on aggregated engagement metrics



The **East region** received the highest total popularity score, followed by the **North**. The **South** consistently showed the lowest engagement.

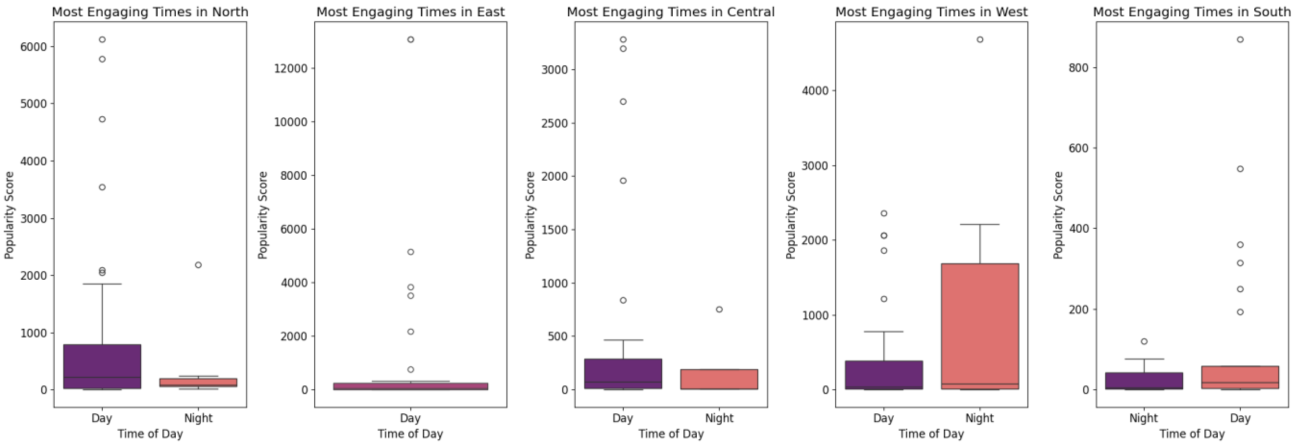
2. **Seasonal Variation: How does popularity change across seasons?**

By categorizing photos based on the season they were taken, we identified patterns:



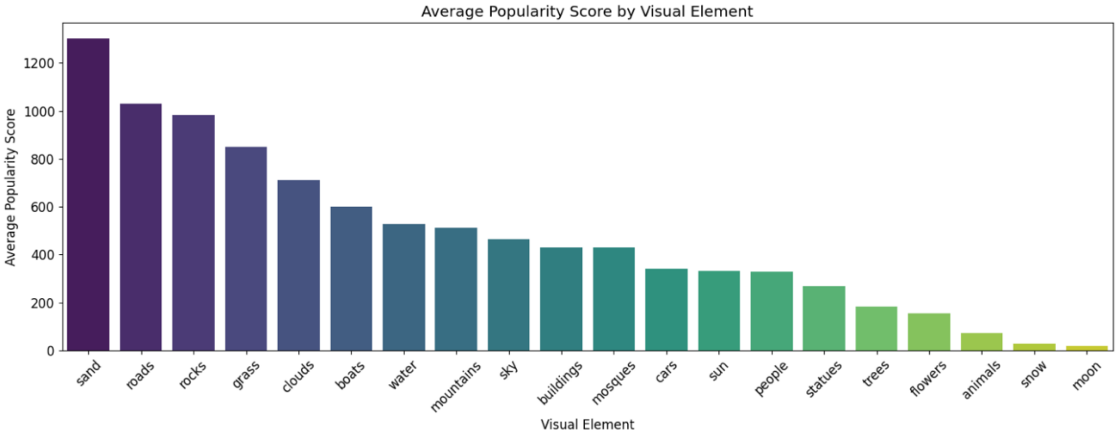
- **Autumn** emerged as the peak season across multiple regions, notably the **East**.
- **Summer and Winter** had lower engagement overall.

3. **Time-Specific Insights: What time of day are destinations more appealing?**
   We split engagement by day vs. night for each region.



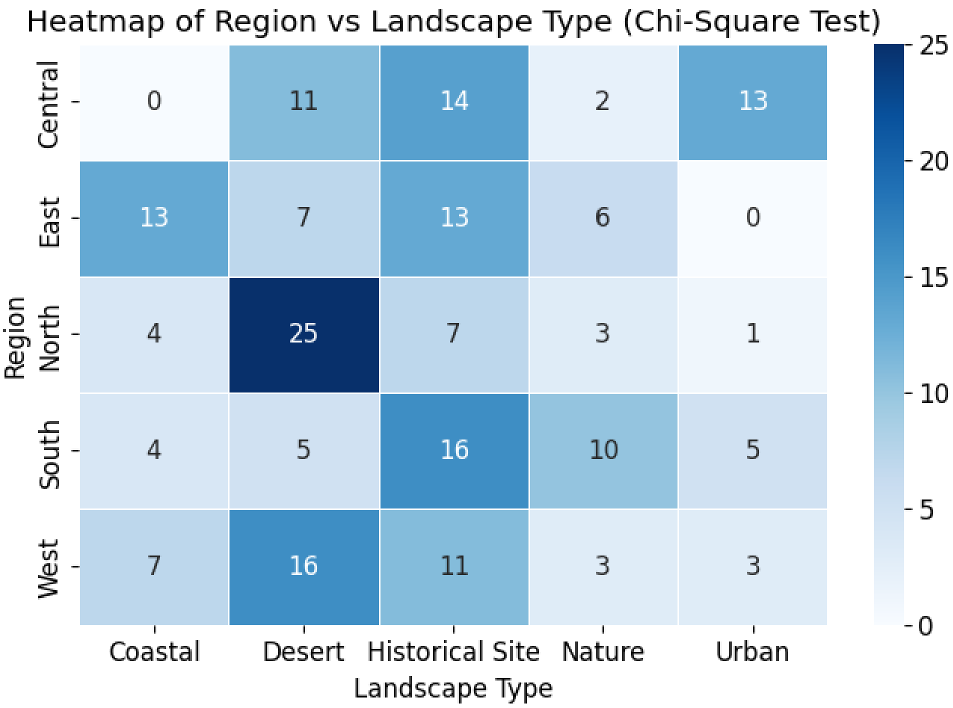
- **Daytime photos** had higher average popularity in most regions, especially the **North** and **West**.
- The **East region shows only daytime values**, likely due to misclassification by the brightness-based model, despite night photos being included.
- The **Central** region was unique, with some preference toward night photos.

4. **Visual Elements: What features appear in popular destination photos?**
   We used the CLIP model to extract dominant visual elements and calculated their average popularity scores to identify which features tend to appear in highly engaging photos.



- Common high-engagement elements included **Sand, Roads,** and **Rocks.**
- Regions varied: the **East** had many photos featuring **coastal** elements, while the **South** had a noticeable presence of **mountains** and **grasslands.**

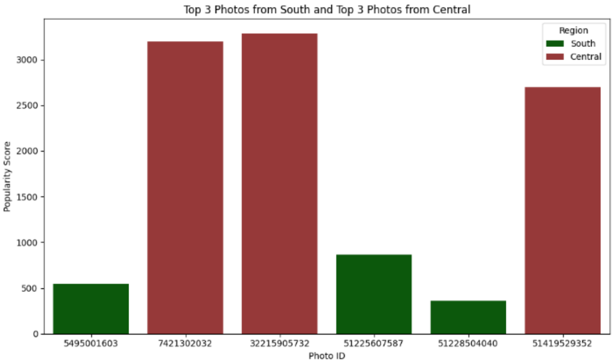
5. **Destination-Based Insights: Which types of destinations are most favored?**
   We analyzed the distribution of landscape types across regions using a Chi-Square heatmap.



- **Deserts** were dominant in the **North**, while historical sites appeared most in the **South** and **Central**.
- The **East** region lacked **urban** classifications, which may reflect either the nature of sampled photos or limitations in the classification model.

6. **Main Research Question: Which attractions in Saudi Arabia deserve more attention?**

Although both the **South** and **Central** regions ranked lowest in overall engagement, a clear disparity was observed in their top-performing attractions. The top 3 photos from the **Central region** significantly outperformed those from the **South**, which had much lower popularity scores across the board.



- This gap highlights that attractions in the **South** are **not only underrepresented** but also **under-engaged**, signaling a need for stronger tourism promotion.
- The results suggest that the **South deserves more attention and visibility**, especially given its cultural and natural potential

Modeling Task and communication:

The primary goal of the modeling was to develop a predictive model for **Popularity Score**, a composite metric designed to quantify how engaging a tourism photo is based on public data. By accurately modeling this score, we aimed to identify the underlying factors that drive photo popularity, providing valuable insights for promoting underrepresented attractions in Saudi Arabia.

1. Modeling Approach

Popularity Score

We defined the **Popularity Score** using a custom formula that emphasizes **engagement efficiency**—how often users chose to interact (like or comment) relative to the number of views. This reflects not just visibility, but **audience interest and sentiment-drivenreaction**.

Formula:

$$Popularity\ Score = \frac{Favorites\ Count + (Comments\ Count \times (Sentiment\ Score + 1))}{1 + Normalized\ View\ Count}$$

This approach allowed us to predict a more meaningful metric than simple favorites or views alone, as it captures both attention and reception quality.

Feature Engineering

Several preprocessing steps we needed to apply in this phase to prepare the data more for machine learning:

- **One-hot encoding** of the *Visual Elements* column (which contained comma-separated text describing objects in each photo) to transform it into numerical features.
- **Timestamp transformation**: *Date Uploaded* was converted to a UNIX timestamp for temporal modeling.

Train/Test Split To evaluate model generalizability, our dataset was split into:

- **80% for training**
- **20% for testing**

This ratio provided a balanced setup that supported reliable performance evaluation while preserving enough data for training.

2. Models Selected

To compare different levels of complexity in the dataset, we selected three distinct regression models, **Linear Regression**, **Random Forest Regressor**, and **Support Vector Regressor (SVR)**. These models were chosen because they align well with the structure and size of our data, which includes both numerical and transformed categorical features like sentiment scores, visual elements, timestamps, and regional classifications.

A. Linear Regression (Baseline Model)

- **Why Chosen:**
  It served as a baseline to help us evaluate whether more complex models significantly outperform a simple, interpretable model.
- **How It Works:**
  Assumes a direct linear relationship between input and the target variable (Popularity Score).

B. Random Forest Regressor

- **Why Chosen:**
  Because of its strong ability to model **non-linear relationships** and work effectively with high-dimensional, mixed-type data like ours.
- **How It Works:**
  Builds an ensemble of decision trees and averages their predictions to reduce overfitting and improve generalization.

C. Support Vector Regressor (SVR)

- **Why Chosen:**
  It's known for performing well on smaller datasets and was selected to test if it could capture subtle, margin-based decision boundaries in our image engagement data.
- **How It Works:**
  Fits a regression line within a defined margin using kernel functions to model non-linear relationships.

3. Evaluation Metrics

To evaluate model performance, we used three common regression metrics:

- **MAE (Mean Absolute Error):** Measures average magnitude of prediction errors.
- **MSE (Mean Squared Error):** Punishes larger errors more heavily by squaring differences.
- **R² Score:** Measures the proportion of variance in the dependent variable predictable from the independent variables.
- **RMSE (Root Mean Squared Error):** Provides error magnitude in the same unit as the target variable, making interpretation more intuitive.

4. Model Performance Results

Model	MAE	MSE	RMSE	R² Score
Linear Regression	236.32	519,952.76	759.75	0.8702
Random Forest Regressor	180	321,855.23	576.70	**0.9288**
Support Vector Regressor	699.35	4,945,423.62	2223.83	-0.0939

Model Comparison and Best Model Justification

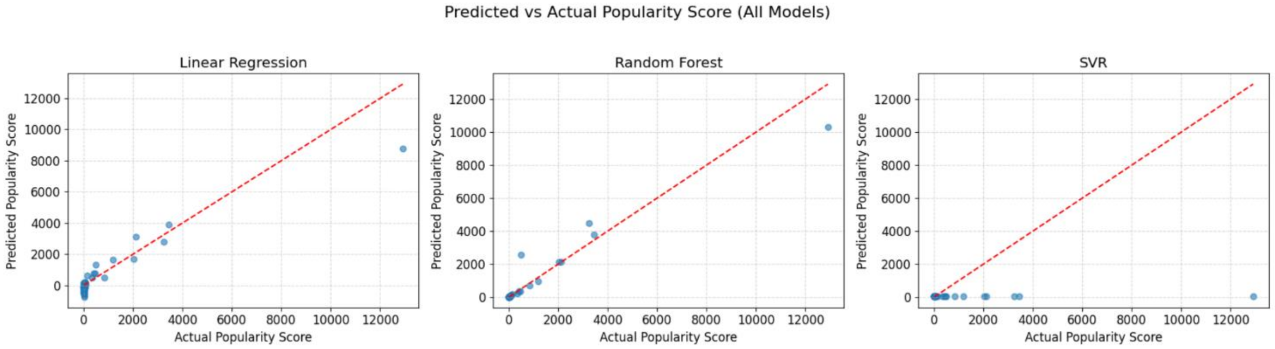
Among the three models tested, the **Random Forest Regressor** emerged as the best-performing model:

- It achieved the **lowest MAE and MSE**, indicating the smallest average and squared error.
- Its **R² score of 0.93** shows that the model explains over 93% of the variance in popularity scores far more than Linear Regression (87%) and vastly better than SVR (which performed worse than a mean-only baseline).
- The significantly lower RMSE of 576.70 compared to 759.75 and 2223.83 further supports its accuracy and robustness against large errors.

These results show that Random Forest successfully captured complex **interactions** in our dataset, which is especially useful considering we're working with diverse input features like sentiment scores, timestamps, and image classifications.

Predicted vs Actual Popularity Scores Across Models

This comparison illustrates each model's ability to predict the Popularity Score based on engagement features. **The Random Forest Regressor (center)** shows the **best** alignment with the ideal prediction line (red dashed), indicating higher accuracy and better generalization. **Linear Regression (left)** captures basic trends but misses non-linear patterns, while **SVR (right)** struggles with underfitting. These results reinforce Random Forest as the most suitable model, confirming that visual elements, sentiment, and region meaningfully contribute to predicting image popularity.



5. Interpretation and Communication of Results

The modeling task demonstrated that photo engagement on Flickr can be predicted using machine learning. More importantly, the predicted Popularity Score offers an interpretable and scalable way to promote tourism attractions in Saudi Arabia.

- **High R^2 values** in Random Forest reflect that our features especially *sentiment, visual elements, and region* have **strong** explanatory power.
- **Low MAE** means predictions are accurate enough to guide tourism promotion strategies.
- The **RMSE**, which remained relatively low in the best-performing model, further confirms the model's reliability by showing minimal large prediction errors.

These predictions, paired with EDA findings, help identify underappreciated regions or seasons that could benefit from more attention in marketing campaigns.

Conclusion and Future Work

Our analysis explored public engagement with Saudi tourist destinations using Flickr photo data, uncovering trends in seasonal popularity, peak visit times, regional appeal, and key visual elements. The Eastern and Northern regions showed the highest levels of popularity, while the Southern region consistently had the lowest engagement, despite its rich cultural heritage and natural beauty. This clearly suggests that the South deserves more focused promotion and visibility to help balance tourism appeal across the Kingdom.

Future work can focus on the following improvements:

- **Expanding data sources** beyond Flickr to include Instagram, TripAdvisor, or Google Maps to improve representativeness.
- **Enhancing image classification accuracy** by fine-tuning CLIP models on tourism-specific datasets, especially to *reduce misclassification of night photos.*
- **Conducting qualitative user studies** to validate AI-driven findings with real tourist feedback and perception.
- **Collaborating with local tourism authorities to** create targeted promotional campaigns for low-engagement areas like the South.

Challenges and recommendations

1. **Limited Accessibility to Comprehensive Data:** Flickr API offers valuable visual data but it is limited by the availability and quality of user-uploaded content, with some regions poorly represented.
   What we recommend is collaborating with local tourism boards or communities to gather more representative content across all regions.
2. **Data Quality and Consistency:** Metadata issues such as irrelevant titles, or incorrect tags affect dataset reliability.
   What we recommend is the removal of irrelevant attributes to ensure the accuracy of the data. Also, use automated scripts to clean and standardize metadata, such as correcting tags to ensure the accuracy of the data.
3. **Seasonal Bias:** Photo availability often varies by season, leading to uneven representation of certain periods.
   What we recommend is to apply statistical techniques to adjust for seasonal overrepresentation and ensure balanced insights.

4. **Subjectivity in Analysis:** User-generated content like titles and comments can be subjective, which may not align with the study's objectives.
   What we recommend is to define clear guidelines for filtering and evaluating content to ensure relevance. Additionally, utilize Natural Language Processing (NLP) tools to analyze text objectively and reduce bias.
5. **High Percentage of Missing Data:** The Flickr API often retrieves location metadata, but a significant portion of this data is incomplete or missing, reducing its usefulness for analysis.
   What we recommend is to remove the "location" attribute. We introduced a new "region" column. Instead of relying on incomplete metadata, we manually categorized the posts into five major regions: Central, West, North, South, and East. This manual assignment allowed us to maintain geographic insights and ensure a more structured and reliable classification for analysis.
6. **Time-Consuming Image Processing:** Processing images, including classification and feature extraction, required significant computational resources.
   What we recommend: Start image processing early to avoid delays and allocate sufficient time for computational tasks.
7. **East Region Had Only Daytime Values:** The dataset for the East region contained only daytime images, likely due to confusion between sunrise and sunset, leading to an imbalance in time-based analysis.
   What we recommend**:** Use a more advanced model that doesn't rely solely on brightness for day/night classification, incorporating contextual features like shadows, artificial lighting, and metadata timestamps.
8. **A Missing Photo Reduced the Dataset from 200 to 199:** One image was lost, and due to computational costs, we couldn't redo the entire processing.
   What we recommend: Implement robust backup mechanisms and checkpoints during data processing to prevent data loss and ensure the ability to recover missing images without restarting the entire workflow.

References

[1] Flickr, **“Flickr,”** *Flickr*, Dec. 04, 2018. https://www.flickr.com/

[2] **“Welcome to Saudi Arabia,”** *www.visitsaudi.com*. https://www.visitsaudi.com/en

[3] **“CLIP: Connecting text and images,”** *Openai.com*, 2021. https://openai.com/index/clip/

[4] OpenCV, **“OpenCV library,”** *Opencv.org*, 2019. https://opencv.org/

[5] **“Transformers,”** *huggingface.co*. https://huggingface.co/docs/transformers/en/index