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

IT 362:
Principles of Data Science
Course Project

2nd semester, 2024/25

Tourism in Saudi Arabia

Group#:	**8**	
Section#:	**66849**	
Group Members:	**Name:**	**ID:**
	Najla Almazyad	**444200948**
	Jood Alkhrashi	**444203007**
	Ghala Musallam	**444200807**
	Reuof Alanazi	**444200528**

Supervised by: Dr. Reem Fahad Alqifari

King Saud University
College of Computer and Information Sciences
Department of Information Technology

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

Phase 1

1. Introduction

Saudi Arabia is emerging as a global tourism destination under its Vision 2030 initiative, aiming to diversify its economy and promote cultural heritage, natural landscapes, and large-scale projects such as NEOM, Al-Ula, and Riyadh Season. These efforts have significantly increased tourist inflow, making it essential to analyze tourism patterns, seasonal trends, and visitor preferences to optimize resources and enhance the overall tourism experience.

A data-driven approach can help identify peak seasons, popular destinations, and key factors influencing tourist satisfaction, enabling stakeholders to develop effective tourism strategies. By leveraging multiple data sources, this study will provide insights into visitor behaviors and the economic impact of tourism initiatives.

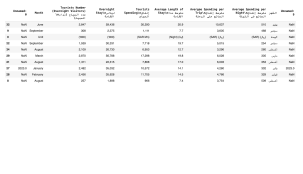
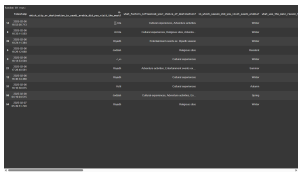
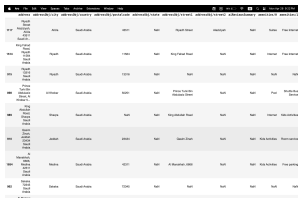
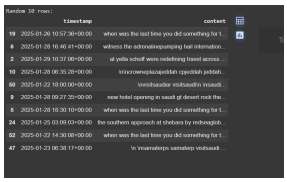
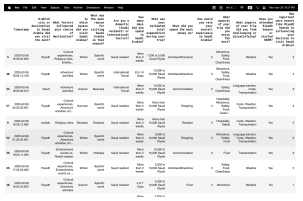
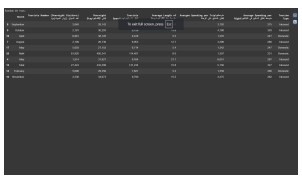
2. Objectives

Our project focuses on answering the following key questions regarding Saudi Arabia's tourism industry:

- What are the most popular tourist destinations in Saudi Arabia?
- Which seasons attract the highest number of visitors, and why?
- How do spending patterns differ between domestic and international tourists?
- What factors contribute most to tourist satisfaction or dissatisfaction?
- How have large-scale projects like Riyadh Season and NEOM influenced visitor trends?

3. Data Description

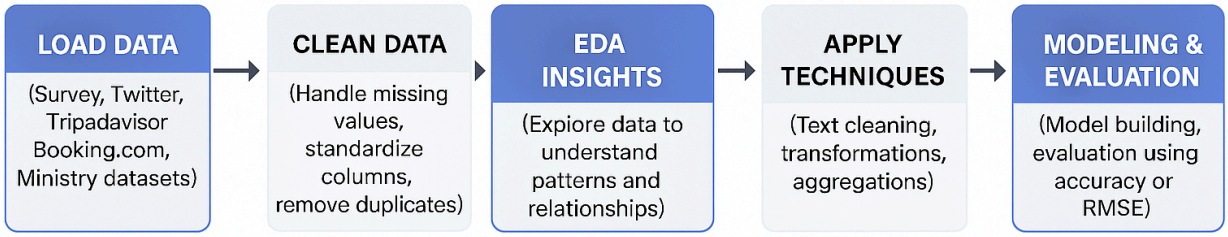
To conduct this study, we collected data from **multiple sources** to cover both **quantitative statistics** and **qualitative feedback** from tourists.

Source	Description	Collection Methods	Unstructured Data SS	Structured Data SS
Ministry of Tourism Open Data Portal (data.gov.sa)	Official statistics on visitor numbers, spending, and destinations.	Public data download (data.gov.sa)		
TripAdvisor API	Tourist reviews and hotel data	Web scraping using Apify		
Surveys	Tourist experiences and preferences	Self-distributed online (Feb 4–6, 2025)		
Booking.com	Accommodation reviews and scores	Web scraping using Apify		
Twitter (X)	Tourist posts, sentiments, and keyword analysis related to tourism.	Web scraping and manual collection from public tweets (January 22–29, 2025)		

3.1. Feature Summary Table

Source	# of observations	# of features	Samples
Ministry of Tourism Open Data Portal (data.gov.sa)	292	323	• Numerical • Nominal
TripAdvisor API	1,656	510	• Nominal • Ordinal
Surveys	53	13	• Nominal • Ordinal
Booking.com	170	79	• Nominal • Ordinal
Twitter (X)	53	6	• Nominal • Temporal

4. Methods



To analyze Saudi Arabia's tourism trends, we followed a structured **data collection, processing, and analysis approach**.

4.1. Data Collection Methods

Utilize both types of data which are *structured* (numeric, categorical) and *unstructured* (textual reviews, etc…) to ensure comprehensive analysis.

Incorporate the data from the multiple resources (Ministry of Tourism, TripAdvisor, Booking.com, and surveys) to capture a wide range of tourist interactions and opinions.

4.2. Data Preprocessing

After collecting the data, the next step is **cleaning and organizing it** to make it usable for analysis. The following steps will be taken:

Data Cleaning:
Address missing values, remove duplicates, and standardize entries to ensure consistency across datasets.

Data Structuring:
Transform unstructured textual data into analyzable formats using natural language processing, enabling sentiment analysis and thematic categorization.

Data Aggregation & Transformation:
Combine data from different sources to create a unified view that allows comparative and trend analysis.

4.3. Analytical Techniques for Answering Research Questions

Research Question	Data Source	Analysis Technique
What are the most popular tourist destinations in Saudi Arabia, and why?	Ministry of Tourism Open Data Portal, Surveys	Frequency analysis to identify the most visited destinations and correlational analysis to explore factors contributing to their popularity.
Which seasons experience the highest tourist inflow?	Ministry of Tourism Open Data Portal	Time-series analysis on visitor data to detect patterns and peaks in tourist arrivals across different seasons.
How do spending patterns differ between domestic and international tourists?	Ministry of Tourism Open Data Portal, Surveys	Statistical analysis to compare spending patterns from different tourist types.
What factors contribute most to tourist satisfaction and dissatisfaction?	TripAdvisor API (hotels), Booking.com, Surveys	Sentiment analysis on hotel reviews and regression analysis on survey data to assess factors influencing satisfaction.
What impact have large-scale projects like Riyadh Season and NEOM had on tourism trends?	Ministry of Tourism Open Data Portal, TripAdvisor API	Comparative analysis to evaluate tourism trends before and after project launches.

By applying these methods, we aim to **extract meaningful insights** from the collected data.

4.4. EDA

4.4.1. Primary Data

Survey:

Statistical Summaries:

- Generated statistical summaries for numerical variables
- Analyzed experience ratings and event influence scores
- Counted unique values for categorical variables (e.g., most visited destinations, spending categories).
- Identified trends in seasonal visits and expenditure patterns.

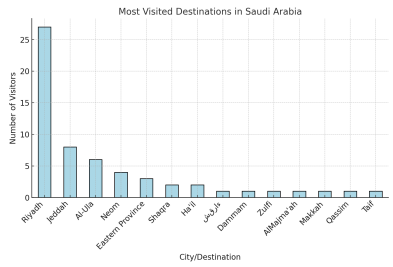
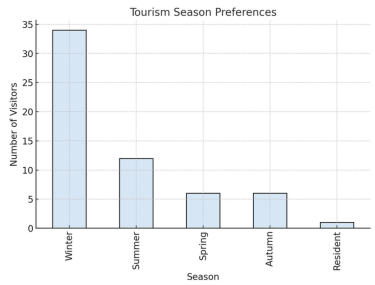
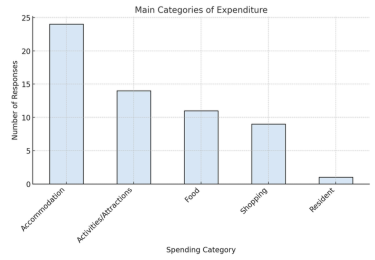
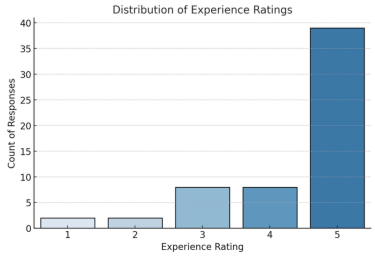
Metric	Experience Rating	Event Influence
Mean	4.36	3,71
Min	1,00	1.00
Max	5.00	5.00
Std Dev	1.06	1.57

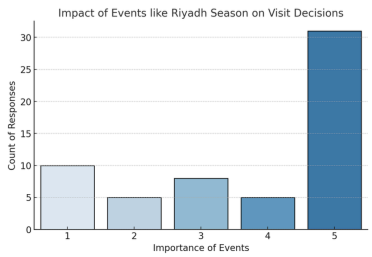
Findings:

- The majority of visitors rated their experience highly (4-5 stars).
- Events played a moderate to strong role in travel decisions, with many respondents rating event influence as 4 or 5.
- A small portion of visitors gave low experience ratings, suggesting some dissatisfaction.

King Saud University
College of Computer and Information Sciences
Department of Information Technology

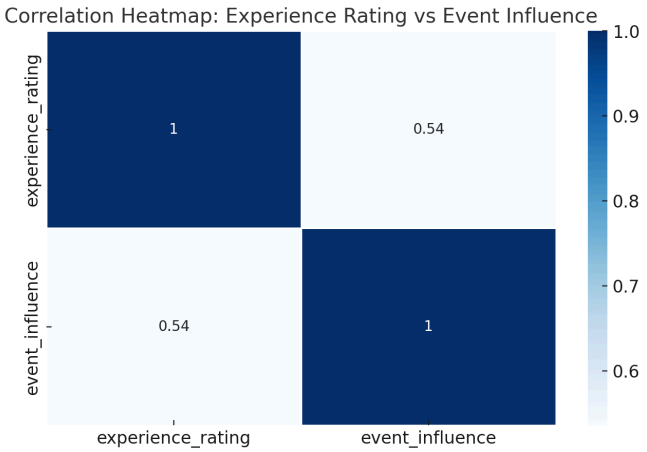
Visual Analysis:

Category	Finding	Visualization
Most Visited Destinations	Riyadh was the most visited destination, followed by Dammam	
Seasonal Tourism Preferences	Winter was the most preferred season for visiting Saudi Arabia, while summer had the lowest number of visitors	
Spending Categories	Most respondents spent the highest on Food, Attractions, and Accommodation	
Experience Ratings Distribution	Majority of visitors rated their experience 4 or 5 stars, confirming overall satisfaction	

Influence of Events on Tourism Decisions	A significant number of respondents rated the event influence as 4 or 5, indicating strong impact	

Correlation Analysis:

To explore the correlation between *experience ratings* and *event influence scores* we generated a heatmap



We found a moderate positive correlation, indicating that while events do impact experience ratings, other factors also play a role.

Experience Rating	**Event influence**	**Correlation = 0.54**

The correlation of 0.54 suggests a moderate relationship, meaning events contribute to experience satisfaction but are not the only factor. Other influences, such as hospitality, costs, or transportation, could also affect overall experience ratings.

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

X “twitter”:

Statistical Summaries:

- Generated summary statistics for tweet timestamps
- Analyzed tweet posting trends over different days and hours

Metric	Year	Month	Day	Hour	Weekday
Count	53	53	53	53	53
Mean	2025	1	25	12.43	3.28
Min	2025	1	25	3	0
Max	2025	1	22	43	6

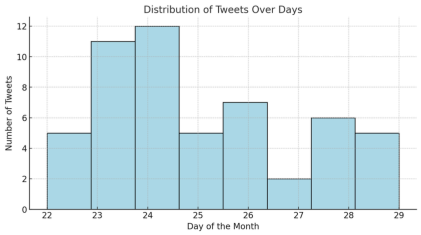
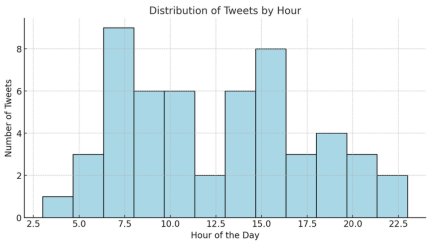
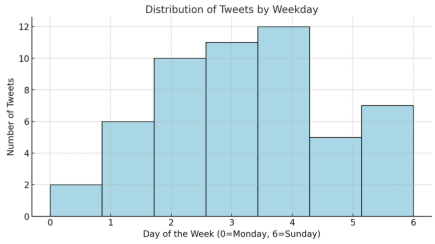
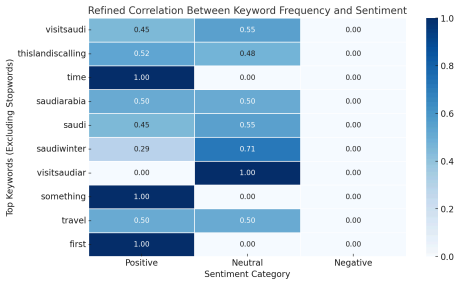
Findings:

- The majority of tweets were posted between **January 22** and **January 29**, 2025
- Most tweets were posted around midday (**average time: 12:43 PM**)
- Tweets were more frequent on **Wednesdays** and **Thursdays**.

King Saud University
College of Computer and Information Sciences
Department of Information Technology

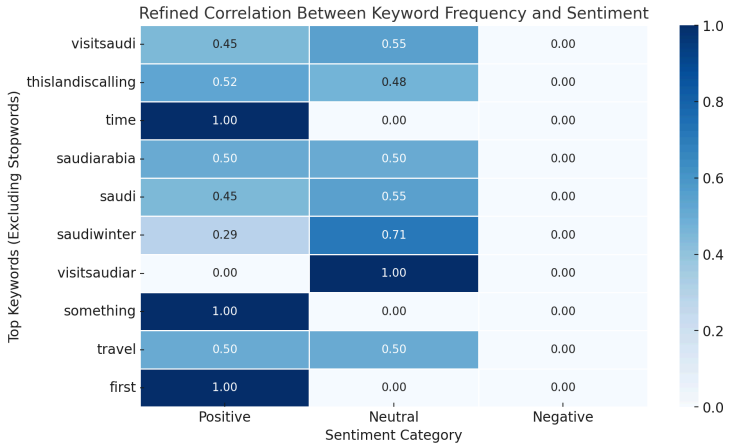
جامعة الملك سعود
King Saud University
كلية علوم الحاسب والمعلومات
قسم تقنية المعلومات

Visual Analysis:

Category	Finding	Visualization
Distribution of Tweets Over Days	The highest volume of tweets was posted between January 23 and January 26.	
Distribution of Tweets by Hour	Tweets peaked around morning and midday hours.	
Distribution of Tweets by Weekday	Most tweets were posted on Wednesdays and Thursdays, with fewer on Mondays.	
Keyword Frequency Analysis	Keywords such as "visit saudi," "travel," "adventure," "Riyadh," "Jeddah," "AlUla," and "tourism" appeared most often.	
Sentiment Analysis of Tweets	Majority of posts were Neutral Very few Negative posts Significant number of posts has a positive sentiment	

Correlation Analysis:

After filtering out stopwords, we analyzed the relationship between key tourism-related words and sentiment categories.

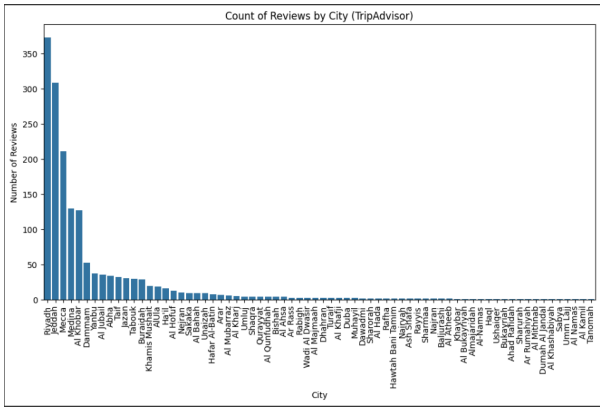
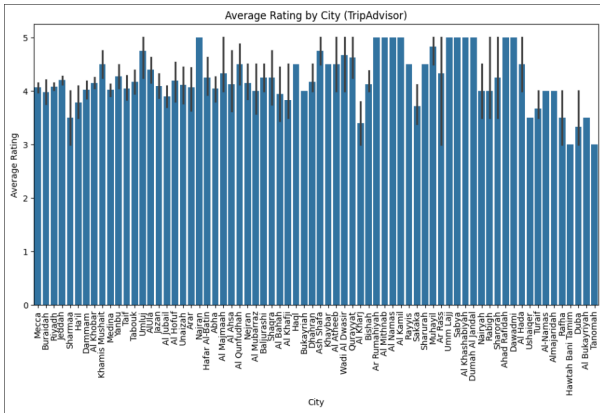
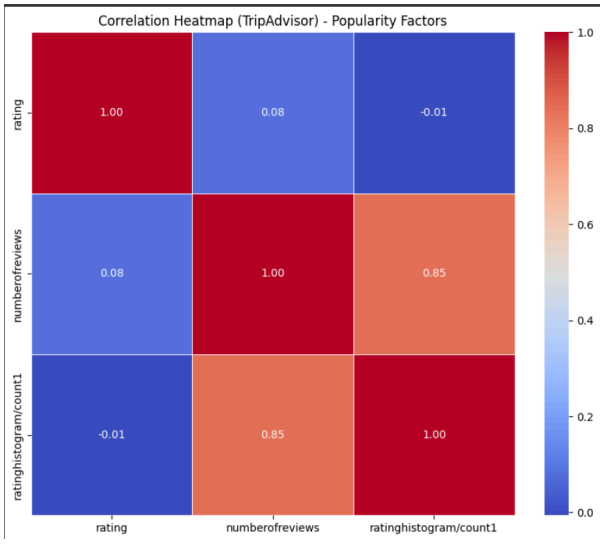


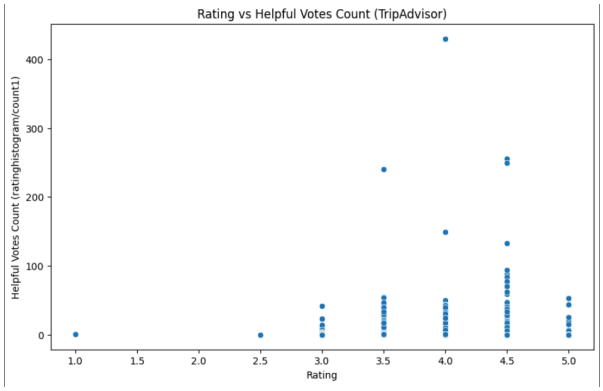
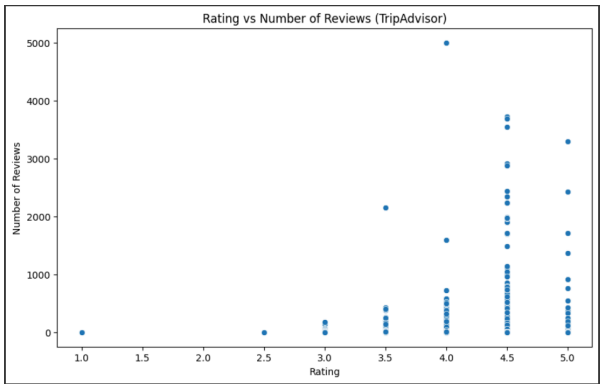
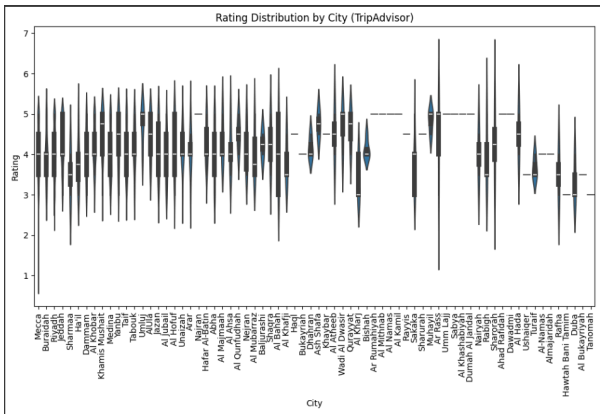
Findings:

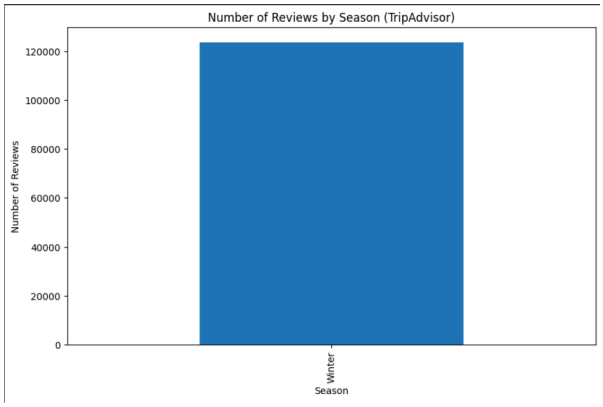
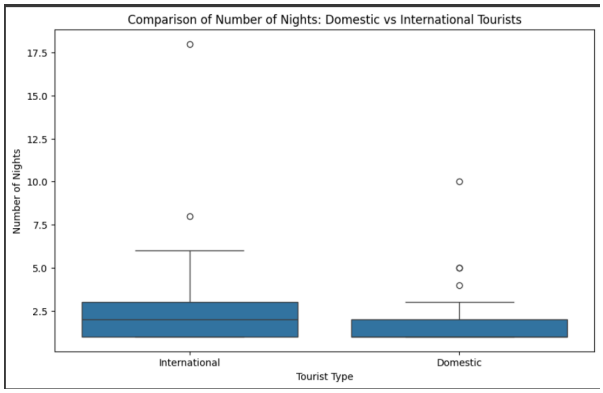
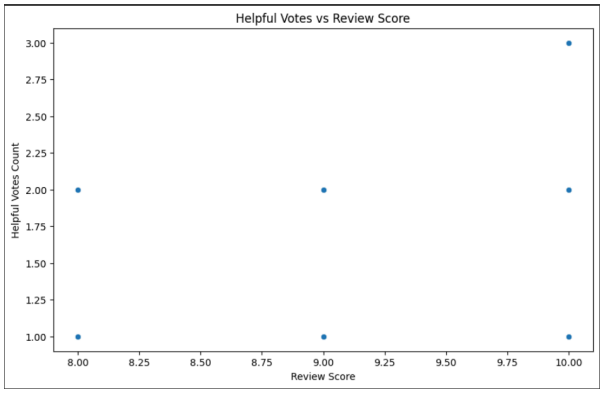
- Some words like "visit saudi" and "thislandiscalling" appear more in positive and neutral tweets, indicating they are promotional
- No major negative words were found in the top 10 most frequent terms
- "Saudi Winter" appears more in Neutral tweets, suggesting informational use rather than emotional sentiment
- "Travel" is evenly split between Positive and Neutral, showing general discussion about tourism.
- Few strong Negative correlations, indicating that tourism-related terms in this dataset are not strongly linked to complaints

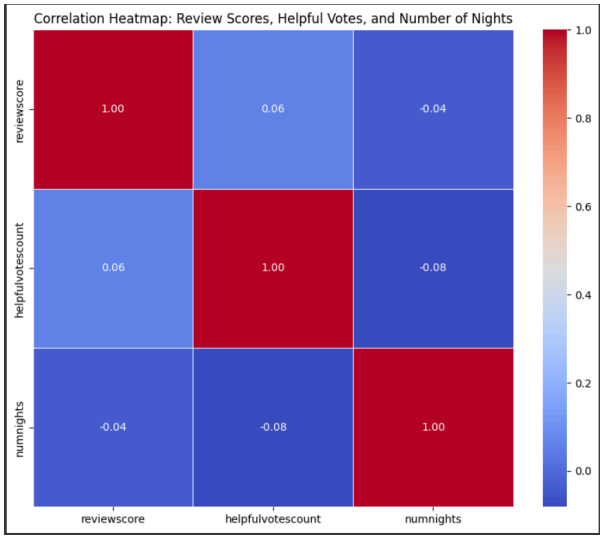
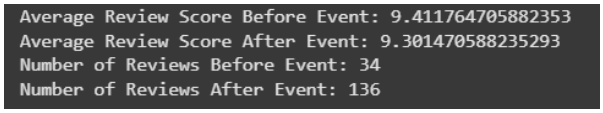
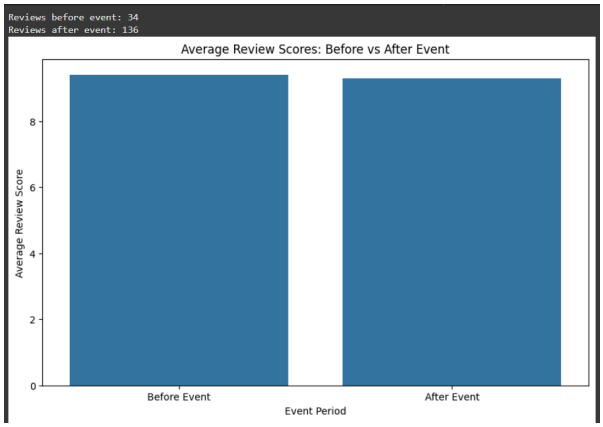
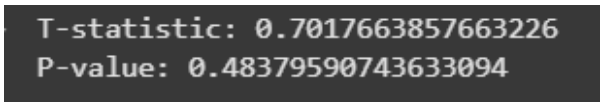
4.4.2. Secondary Data

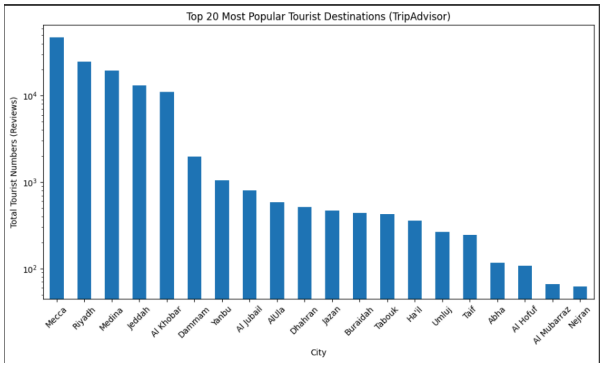
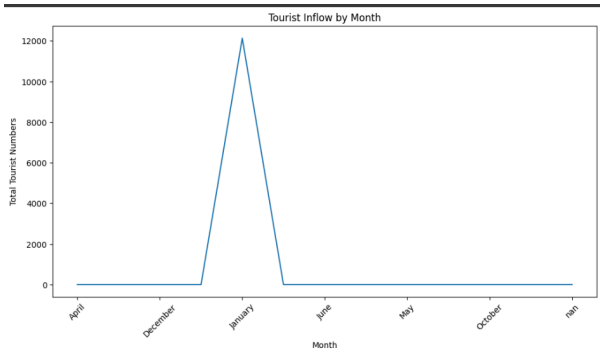
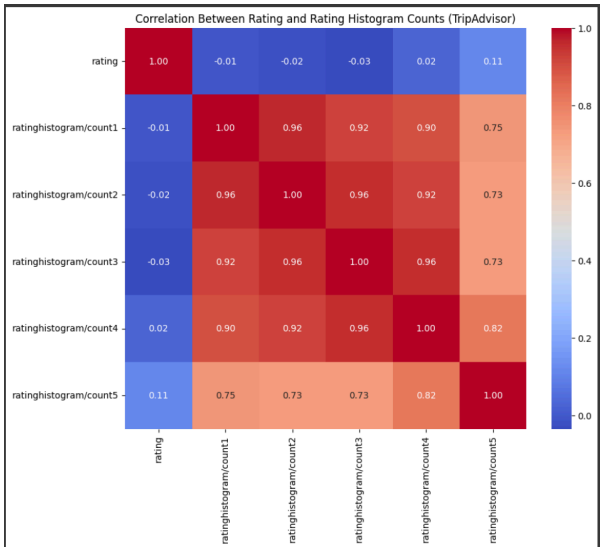
King Saud University
College of Computer and Information Sciences
Department of Information Technology

This will show the number of reviews for each city.	
This will show the average rating for each city.	
This will show the correlation between features like rating, number of reviews, and rating histogram counts.	

This plot will help visualize the relationship between hotel ratings and the number of reviews in the first rating category (e.g., 1-star or 2-star reviews). It shows whether higher-rated hotels tend to have fewer low-star reviews, or if there's any other pattern in the distribution of ratings and low reviews.	
This plot will show how rating correlates with the number of reviews.	
This plot will help visualize the distribution of ratings by city, showing the spread and any outliers.	

A bar plot will show how the number of reviews changes by season, helping us identify peak tourism seasons.	
we can compare the review scores between the two groups.	
We can perform a t-test to check if the number of nights or review scores significantly differ between Domestic and International tourists.	T-statistic for number of nights: -2.4988643195960405, P-value: 0.013419802232587714 T-statistic for review scores: 1.1524452784107362, P-value: 0.25077535813973245
This plot will help visualize if there's a relationship between the helpful votes and review scores.	

This will help us understand the correlation between review scores, helpful votes, and number of nights.	
number of reviews, average rating, and price range before and after the event.	
We can use a bar plot to visualize the average review scores before and after Riyadh Season.	
We can perform a t-test to determine if there is a significant difference in review scores between the two periods.	

The most popular tourist destinations in Saudi Arabia are the provinces with the highest tourist numbers, such as Riyadh, Jeddah, and Makkah, driven by high tourist inflows in these regions.	
The highest tourist inflow occurs in January, likely due to special events (such as Riyadh Season and New Year's tourism)	
This correlation matrix will show how tourist numbers in different provinces relate to each other. A strong correlation between provinces would suggest similar patterns in tourist arrivals, which could indicate shared tourism factors or seasonal trends.	

1.2.1 Metadata Review

Metadata Review for TripAdvisor, Booking.com, and the Ministry of Tourism:

- **Source:**

- TripAdvisor: Data was gathered via web scraping or through the API from TripAdvisor, focusing on user reviews, hotel information, and ratings of hotels and tourist attractions.
- Booking.com: The data was similarly obtained through web scraping or the API from Booking.com, containing user reviews, hotel rankings, prices, and check-in/check-out dates.
- Ministry of Tourism: Data was downloaded directly from the Ministry of Tourism's Open Data Portal, providing tourist statistics, including visitor counts, overnight stays, and tourist spending patterns in Saudi Arabia.

- **Date Collected:**
  - TripAdvisor: Data was collected from February 1 to 4, 2025.
  - Booking.com: Data was collected from February 1 to 4, 2025.
  - Ministry of Tourism: Data was downloaded from the Open Data Portal on February 2, 2025.

- **Collection Method:**
  - TripAdvisor: Data was collected using web scraping techniques via Apify or the TripAdvisor API, which provides access to reviews and hotel-related information.
  - Booking.com: Data was similarly collected via Apify for web scraping or through the Booking.com API.
  - Ministry of Tourism: Data was obtained directly from the Ministry of Tourism Open Data Portal, which offers public access to tourism-related datasets.

1.2.1 Bias Awareness

Bias Awareness for TripAdvisor, Booking.com, and the Ministry of Tourism:

1. Geographical Bias:
   a. **TripAdvisor:**
   - The dataset overrepresents popular tourist destinations in Saudi Arabia, especially cities like Riyadh, Makkah, Medina, and Jeddah. These cities receive the bulk of tourist activity, leading to a skewed representation of tourist experiences.

- Areas with lower tourism traffic are underrepresented, resulting in a geographical bias.

b. Booking.com:

- Similar to TripAdvisor, tourist-heavy cities like Riyadh and Jeddah dominate the dataset. Hotels in high-traffic tourist areas are overrepresented, causing geographical bias.
- The dataset is also skewed toward higher-end accommodations, which attracts a wealthier or business-oriented clientele, further contributing to geographical bias.

c. Ministry of Tourism:

- Tourism data from the Ministry of Tourism is more focused on key urban areas and popular regions. Data from government-targeted tourism programs like NEOM or Riyadh Season can create a bias toward more developed regions with greater tourism infrastructure.

2. Temporal Bias:

a. TripAdvisor:

- The dataset reflects reviews from tourists during peak tourism seasons, such as Ramadan, Hajj, and Riyadh Season, which leads to seasonal bias.
- Post-pandemic effects are evident, with changes in tourism behavior affecting the review data.

b. Booking.com:

- Similar to TripAdvisor, seasonal events like Riyadh Season and Ramadan skew the dataset by generating a surge in reviews during these periods. This introduces temporal bias.
- Promotions and special events might lead to increased reviews during specific periods, affecting the overall tourism trends.

c. Ministry of Tourism:

- Data from the Ministry of Tourism is influenced by seasonal fluctuations in tourism, particularly during high-tourism periods like Hajj, Ramadan, and other festivals.

- Government-driven initiatives like the NEOM project may lead to noticeable spikes in tourist numbers, especially if focused around specific periods.

3. User Bias:

a. TripAdvisor:

- The data is skewed toward extreme reviews, where tourists who had very positive or very negative experiences are more likely to leave feedback.
- Frequent travelers or those staying in luxury accommodations are overrepresented in the reviews, which skews the overall review quality toward more experienced travelers and wealthier demographics.

b. Booking.com:

- Frequent users of Booking.com are overrepresented, especially those staying in higher-end hotels or luxury properties. This results in a dataset with higher ratings and a focus on wealthier tourist demographics.
- International tourists using Booking.com are likely overrepresented, while domestic tourists might be underrepresented.

c. Ministry of Tourism:

- The Ministry of Tourism dataset may overrepresent wealthier tourists or those who stay in more expensive hotels.
- Domestic tourists may be overrepresented in the data, especially during events like Ramadan or Hajj, while international tourism could be underrepresented.

5. Results and Discussions

Comparison of Primary and Secondary Data

The comparison between survey ratings (primary data) and external review ratings (secondary data) revealed both alignments and discrepancies.

Alignments:

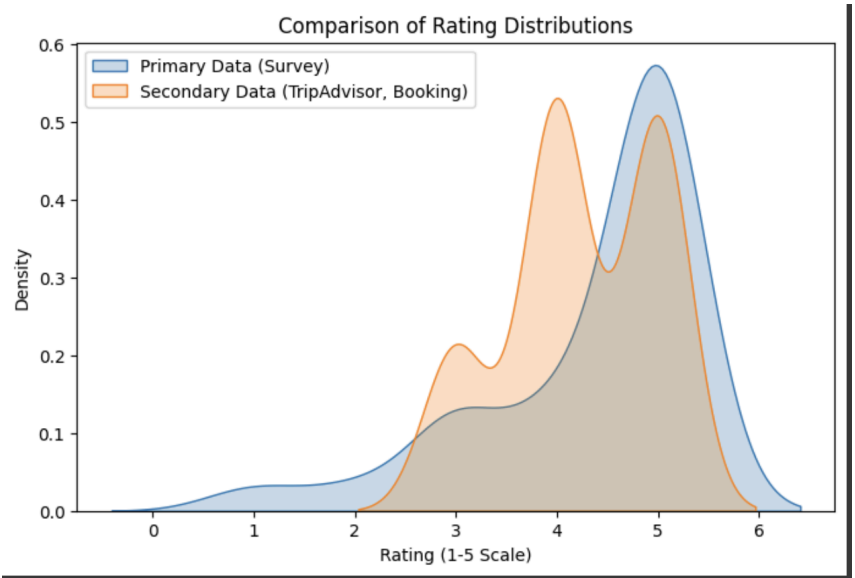
- The average ratings from both sources were similar, indicating a generally positive experience across respondents.
- The overall trend showed that most ratings were concentrated between 4 and 5, suggesting that both datasets reflect high levels of satisfaction.
- A correlation analysis indicated a moderate relationship between the two sources, implying some consistency in rating patterns.

Discrepancies:

- The median rating in the survey data was 5, while it was 4 in the secondary data, suggesting survey respondents were more inclined to give perfect ratings.
- The standard deviation in the survey data was higher, meaning responses varied more compared to the more consistent ratings in external sources.
- The density plot comparison showed that survey ratings were more skewed toward higher values, while TripAdvisor and Booking.com ratings were more evenly distributed.

Visualization:

To illustrate these differences, a comparison table was created displaying key metrics (mean, median, and standard deviation) side by side. Additionally, a density plot was generated to compare rating distributions, highlighting the tendency for higher ratings in the survey responses.



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

```
Differences in Key Metrics:
            mean  median
Rating  0.118644     1.0

Primary Sample Size: 59
Secondary Sample Size: 59
```

```
Key Metrics Comparison Table:
Metric  Primary  Secondary
  Mean 4.355932   4.237288
Median 5.000000   4.000000
   Std 1.062896   0.727276
```

Findings:

1. **Survey respondents tend to give higher ratings than external reviewers.**
   The median rating in the survey data was 5, while in secondary data, it was 4, indicating that survey participants were more likely to give perfect scores
2. **External review ratings are more consistent.**
   The standard deviation in survey data was higher, suggesting a wider range of responses, whereas external ratings were more stable.
3. **The correlation between the two datasets is moderate.**
   While there is some alignment in rating patterns, differences suggest that survey respondents and external reviewers have slightly different perspectives.
4. **Survey data is skewed toward higher ratings.**
   A density plot comparison showed that survey ratings were more concentrated at 4 and 5, while secondary data had a more even distribution.
5. **Differences may be due to rating context or platform effects.**
   Survey respondents may be more engaged or have a different rating mindset compared to users on review platforms, who might be more critical.

These findings suggest that while both datasets capture positive sentiment, differences in rating behavior should be considered when interpreting results.

5.1. Summary of New Insights and Hypotheses

The EDA revealed clear differences in rating patterns between survey respondents and external reviewers. The survey data consistently showed higher ratings and greater variability, while secondary data had a lower median and more consistent ratings.

New Insights

1. Survey respondents may exhibit response bias.
   The tendency for survey participants to give higher ratings could be due to social desirability bias or a more engaged audience.

2. External platforms may encourage more balanced ratings.
   Since external reviews are often influenced by previous ratings, visibility of past reviews, or competitive ranking systems, users may self-adjust their ratings to align with expectations.
3. Different user motivations affect ratings.
   Survey respondents may rate based on overall satisfaction, while external reviewers may rate based on specific experiences (e.g., service, cleanliness, location).

4. The variance in survey responses suggests a diverse respondent base.
   The higher standard deviation in the survey ratings implies that respondents may have different experiences or expectations compared to those who leave reviews on external platforms.

New Hypotheses for Further Investigation

H1: Survey respondents are more likely to provide higher ratings due to engagement bias or lack of critical reviews compared to public review platforms.
H2: Users on external platforms tend to be more critical and objective due to the influence of visible past reviews and comparisons with other destinations or services.
H3: The higher variability in survey responses may be linked to different demographic factors, such as first-time visitors vs. repeat visitors.
H4: Cultural differences in rating behavior may influence the discrepancies between survey and external reviews.

These insights and hypotheses highlight the need for further segmentation and contextual analysis to better understand how different factors influence user ratings.

Colab Notebooks:

- DS - Python Notebook.ipynb
- Structured Data Samples.ipynb
- Raw Data Samples.ipynb
- EDA

6. Challenges & Recommendations

6.1. Challenges Faced

- Some platforms, like TripAdvisor, require approval for API access, which can delay data retrieval due to the limited API access
- It was difficult to get a diverse range of responses from tourists as it was hard to reach them.

6.2. Recommendations

- Use automated API requests to streamline data collection.
- Applying advanced filtering techniques to remove irrelevant social media posts.
- Increase survey distribution to improve data diversity.

7. Conclusion

This project provided a comprehensive analysis of tourism trends in Saudi Arabia by integrating primary and secondary data sources. Key findings revealed that Riyadh, Jeddah, and Makkah are the most popular destinations, winter is the peak tourism season, and spending patterns vary significantly between domestic and international tourists. Sentiment analysis showed overall positive tourist experiences, particularly during major events like Riyadh Season.

The results offer valuable insights for policymakers and stakeholders aiming to enhance tourism strategies and support the country's Vision 2030 goals

8. Future Work

Future research could expand by collecting a larger and more diverse set of survey responses to capture broader visitor demographics. Additionally, applying advanced machine learning models for predictive

analysis could provide deeper forecasts of tourism trends. Incorporating real-time social media analytics and external economic factors would further enrich the insights and help refine tourism development initiatives.