

Phase 2 - Logbook: Detailed Overview

Date of Collection: February 1-6, 2025

Data Collection Methods Overview:

- 1. TripAdvisor and Booking.com (Feb 1 and Feb 4, 2025):** Utilized Apify for web scraping, focusing on hotel reviews across Saudi Arabia. Targeted data included customer reviews, ratings, and accommodation specifics.
- 2. Ministry of Tourism Open Data Portal (Feb 2, 2025):** Direct download of tourist statistics including overnight stays and spending patterns.
- 3. Survey Distribution (Feb 4-6, 2025):** Implemented through Google Forms, targeting both domestic and international tourists' preferences and satisfaction levels.

Exploratory Data Analysis (EDA)

Primary Data

Surveys:

The objective here is to conduct an in-depth analysis on the collected survey responses to uncover patterns, trends, and anomalies.

Tools & Libraries used:

- **Python Libraries:** Pandas, Matplotlib, Seaborn
- **Visualization Tools:** Matplotlib, Seaborn

Statistical Summaries:

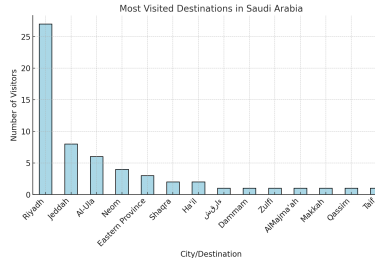
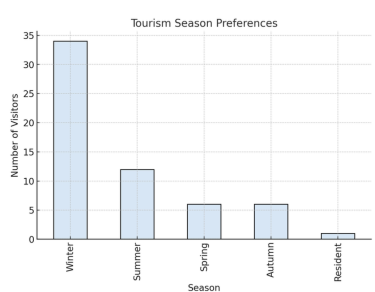
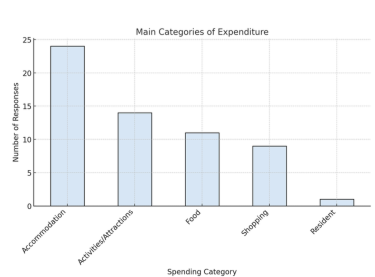
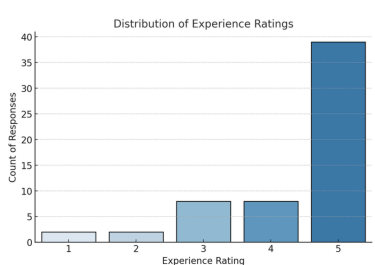
- Generated statistical summaries for numerical variables
- Analyzed experience ratings and event influence scores
- Counted unique values for categorical variables (e.g., most visited destinations, spending categories).
- Identified trends in seasonal visits and expenditure patterns.

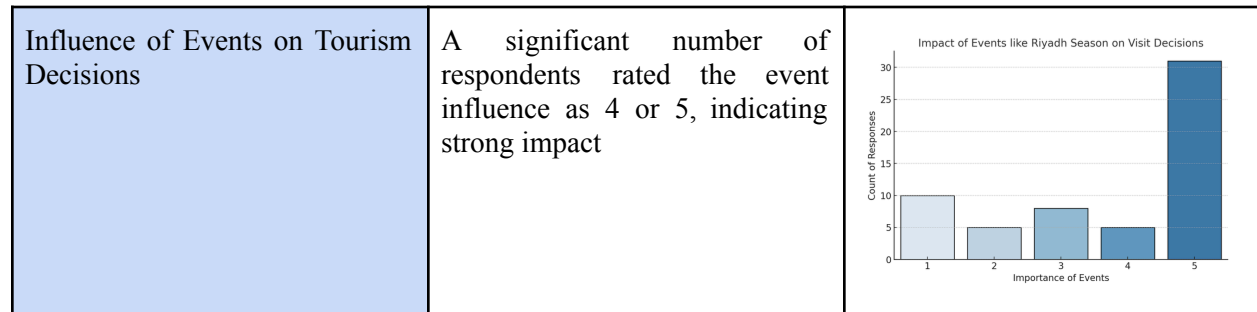
Metric	Experience Rating	Event Influence
Mean	4.36	3,71
Min	1,00	1.00
Max	5.00	5.00
Std Dev	1.06	1.57

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.

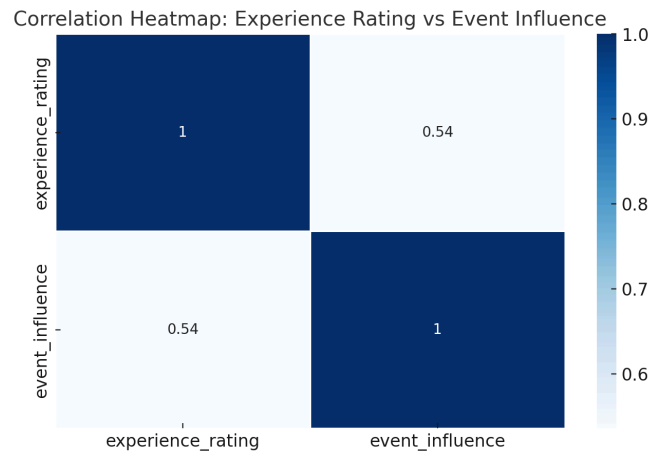
Visual Analysis:

Category	Finding	Visualization																												
Most Visited Destinations	Riyadh was the most visited destination, followed by Dammam	 <table border="1"><caption>Most Visited Destinations in Saudi Arabia</caption><thead><tr><th>City/Destination</th><th>Number of Visitors</th></tr></thead><tbody><tr><td>Riyadh</td><td>25</td></tr><tr><td>Jeddah</td><td>8</td></tr><tr><td>Al-Ula</td><td>6</td></tr><tr><td>Hail</td><td>4</td></tr><tr><td>Eastern Province</td><td>3</td></tr><tr><td>Shagra</td><td>2</td></tr><tr><td>Ha'il</td><td>2</td></tr><tr><td>Dammam</td><td>1</td></tr><tr><td>Zulfi</td><td>1</td></tr><tr><td>Al-Madina</td><td>1</td></tr><tr><td>Makkah</td><td>1</td></tr><tr><td>Qassim</td><td>1</td></tr><tr><td>Tabuk</td><td>1</td></tr></tbody></table>	City/Destination	Number of Visitors	Riyadh	25	Jeddah	8	Al-Ula	6	Hail	4	Eastern Province	3	Shagra	2	Ha'il	2	Dammam	1	Zulfi	1	Al-Madina	1	Makkah	1	Qassim	1	Tabuk	1
City/Destination	Number of Visitors																													
Riyadh	25																													
Jeddah	8																													
Al-Ula	6																													
Hail	4																													
Eastern Province	3																													
Shagra	2																													
Ha'il	2																													
Dammam	1																													
Zulfi	1																													
Al-Madina	1																													
Makkah	1																													
Qassim	1																													
Tabuk	1																													
Seasonal Tourism Preferences	Winter was the most preferred season for visiting Saudi Arabia, while summer had the lowest number of visitors	 <table border="1"><caption>Tourism Season Preferences</caption><thead><tr><th>Season</th><th>Number of Visitors</th></tr></thead><tbody><tr><td>Winter</td><td>35</td></tr><tr><td>Summer</td><td>12</td></tr><tr><td>Spring</td><td>6</td></tr><tr><td>Autumn</td><td>6</td></tr><tr><td>Resident</td><td>1</td></tr></tbody></table>	Season	Number of Visitors	Winter	35	Summer	12	Spring	6	Autumn	6	Resident	1																
Season	Number of Visitors																													
Winter	35																													
Summer	12																													
Spring	6																													
Autumn	6																													
Resident	1																													
Spending Categories	Most respondents spent the highest on Food, Attractions, and Accommodation	 <table border="1"><caption>Main Categories of Expenditure</caption><thead><tr><th>Spending Category</th><th>Number of Responses</th></tr></thead><tbody><tr><td>Accommodation</td><td>24</td></tr><tr><td>Attractions</td><td>14</td></tr><tr><td>Food</td><td>11</td></tr><tr><td>Shopping</td><td>9</td></tr><tr><td>Resident</td><td>1</td></tr></tbody></table>	Spending Category	Number of Responses	Accommodation	24	Attractions	14	Food	11	Shopping	9	Resident	1																
Spending Category	Number of Responses																													
Accommodation	24																													
Attractions	14																													
Food	11																													
Shopping	9																													
Resident	1																													
Experience Ratings Distribution	Majority of visitors rated their experience 4 or 5 stars, confirming overall satisfaction	 <table border="1"><caption>Distribution of Experience Ratings</caption><thead><tr><th>Experience Rating</th><th>Count of Responses</th></tr></thead><tbody><tr><td>1</td><td>2</td></tr><tr><td>2</td><td>2</td></tr><tr><td>3</td><td>8</td></tr><tr><td>4</td><td>8</td></tr><tr><td>5</td><td>38</td></tr></tbody></table>	Experience Rating	Count of Responses	1	2	2	2	3	8	4	8	5	38																
Experience Rating	Count of Responses																													
1	2																													
2	2																													
3	8																													
4	8																													
5	38																													



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.

X:

The objective here is to conduct an in-depth analysis on the primary dataset (X data) to uncover patterns, trends, and anomalies in social media discussions related to tourism in Saudi Arabia.

Tools & Libraries used:

- **Python Libraries:** Pandas, Matplotlib, Seaborn
- **Visualization Tools:** Matplotlib, Seaborn

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:

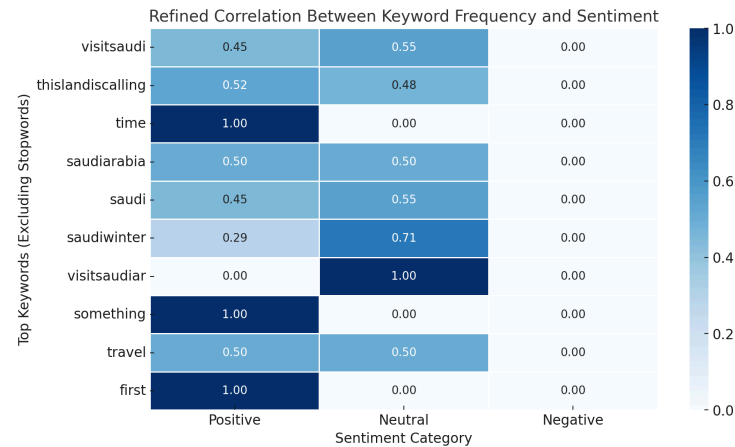
- The majority of tweets were posted between **January 22** and **January 29**, 2025
- Most tweets were posted around midday (**average time: 12:43 PM**)
- Tweets were more frequent on **Wednesdays** and **Thursdays**.

Visual Analysis:

Category	Finding	Visualization																				
Distribution of Tweets Over Days	The highest volume of tweets was posted between January 23 and January 26.	<table><caption>Distribution of Tweets Over Days</caption><thead><tr><th>Day of the Month</th><th>Number of Tweets</th></tr></thead><tbody><tr><td>22</td><td>5</td></tr><tr><td>23</td><td>11</td></tr><tr><td>24</td><td>12</td></tr><tr><td>25</td><td>7</td></tr><tr><td>26</td><td>7</td></tr><tr><td>27</td><td>2</td></tr><tr><td>28</td><td>6</td></tr><tr><td>29</td><td>5</td></tr></tbody></table>	Day of the Month	Number of Tweets	22	5	23	11	24	12	25	7	26	7	27	2	28	6	29	5		
Day of the Month	Number of Tweets																					
22	5																					
23	11																					
24	12																					
25	7																					
26	7																					
27	2																					
28	6																					
29	5																					
Distribution of Tweets by Hour	Tweets peaked around morning and midday hours.	<table><caption>Distribution of Tweets by Hour</caption><thead><tr><th>Hour of the Day</th><th>Number of Tweets</th></tr></thead><tbody><tr><td>2.5</td><td>1</td></tr><tr><td>5.0</td><td>3</td></tr><tr><td>7.5</td><td>9</td></tr><tr><td>10.0</td><td>6</td></tr><tr><td>12.5</td><td>2</td></tr><tr><td>15.0</td><td>8</td></tr><tr><td>17.5</td><td>3</td></tr><tr><td>20.0</td><td>4</td></tr><tr><td>22.5</td><td>2</td></tr></tbody></table>	Hour of the Day	Number of Tweets	2.5	1	5.0	3	7.5	9	10.0	6	12.5	2	15.0	8	17.5	3	20.0	4	22.5	2
Hour of the Day	Number of Tweets																					
2.5	1																					
5.0	3																					
7.5	9																					
10.0	6																					
12.5	2																					
15.0	8																					
17.5	3																					
20.0	4																					
22.5	2																					
Distribution of Tweets by Weekday	Most tweets were posted on Wednesdays and Thursdays, with fewer on Mondays.	<table><caption>Distribution of Tweets by Weekday</caption><thead><tr><th>Day of the Week (0=Monday, 6=Sunday)</th><th>Number of Tweets</th></tr></thead><tbody><tr><td>0</td><td>2</td></tr><tr><td>1</td><td>6</td></tr><tr><td>2</td><td>10</td></tr><tr><td>3</td><td>11</td></tr><tr><td>4</td><td>12</td></tr><tr><td>5</td><td>5</td></tr><tr><td>6</td><td>7</td></tr></tbody></table>	Day of the Week (0=Monday, 6=Sunday)	Number of Tweets	0	2	1	6	2	10	3	11	4	12	5	5	6	7				
Day of the Week (0=Monday, 6=Sunday)	Number of Tweets																					
0	2																					
1	6																					
2	10																					
3	11																					
4	12																					
5	5																					
6	7																					
Keyword Frequency Analysis	Keywords such as "visit saudi," "travel," "adventure," "Riyadh," "Jeddah," "AlUla," and "tourism" appeared most often.	<table><caption>Distribution of Tweets by Weekday</caption><thead><tr><th>Day of the Week (0=Monday, 6=Sunday)</th><th>Number of Tweets</th></tr></thead><tbody><tr><td>0</td><td>2</td></tr><tr><td>1</td><td>6</td></tr><tr><td>2</td><td>10</td></tr><tr><td>3</td><td>11</td></tr><tr><td>4</td><td>12</td></tr><tr><td>5</td><td>5</td></tr><tr><td>6</td><td>7</td></tr></tbody></table>	Day of the Week (0=Monday, 6=Sunday)	Number of Tweets	0	2	1	6	2	10	3	11	4	12	5	5	6	7				
Day of the Week (0=Monday, 6=Sunday)	Number of Tweets																					
0	2																					
1	6																					
2	10																					
3	11																					
4	12																					
5	5																					
6	7																					
Sentiment Analysis of Tweets	<p>Majority of posts were Neutral</p> <p>Very few Negative posts</p> <p>Significant number of posts has a positive sentiment</p>																					

Correlation Analysis:

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



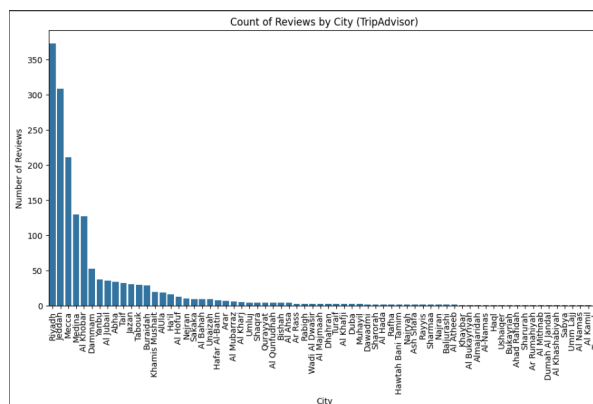
Findings:

- 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

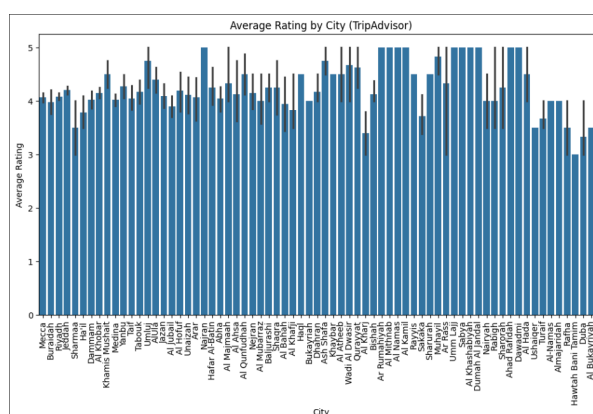
Exploratory Data Analysis (EDA)

Secondary Data

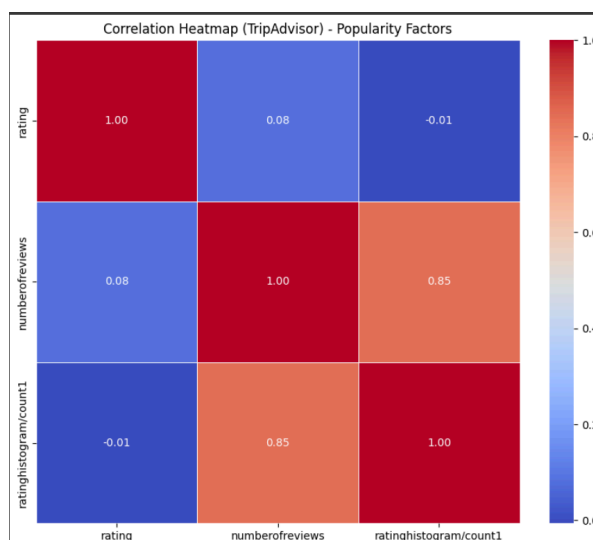
This will show the number of reviews for each city.



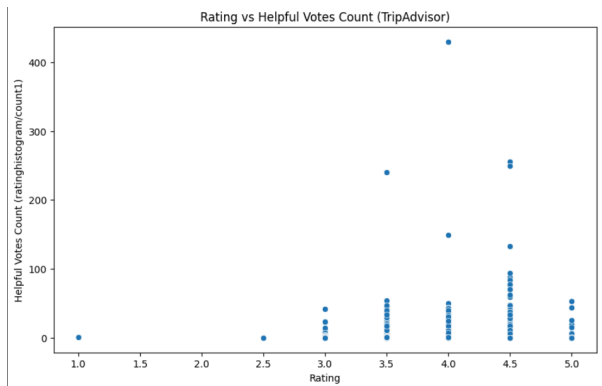
This will show the average rating for each city.



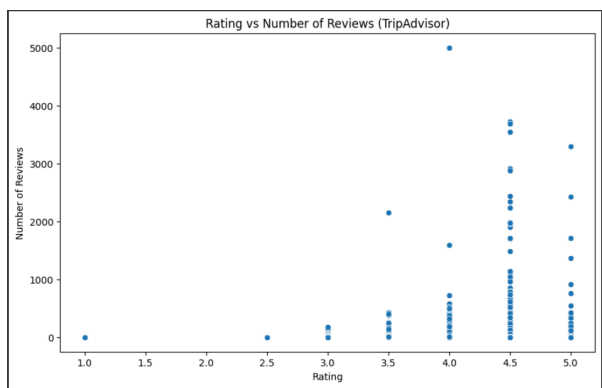
This will show the correlation between features like rating, number of reviews, and rating histogram counts.



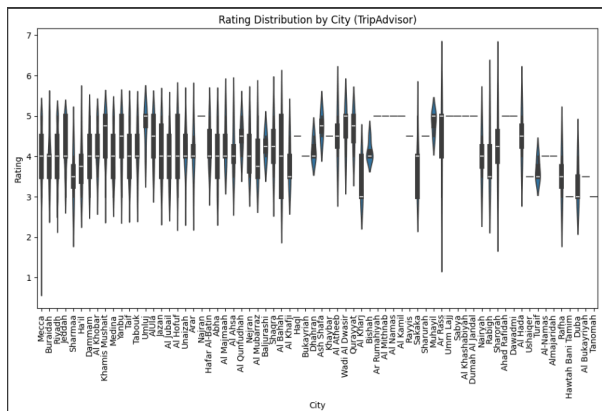
This plot will help visualize the relationship between hotel ratings and the number of reviews in the first rating category (e.g., 1-star or 2-star reviews). It shows whether higher-rated hotels tend to have fewer low-star reviews, or if there's any other pattern in the distribution of ratings and low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



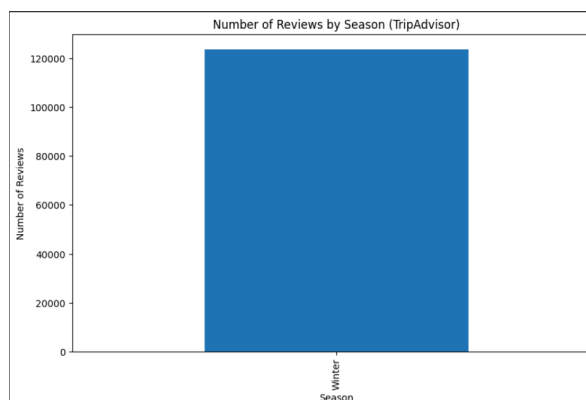
This plot will show how rating correlates with the number of reviews.



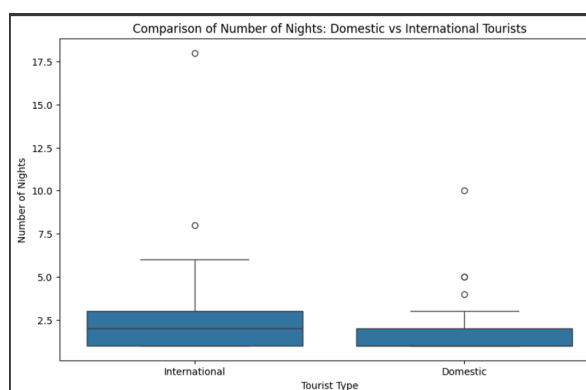
This plot will help visualize the distribution of ratings by city, showing the spread and any outliers.



A bar plot will show how the number of reviews changes by season, helping us identify peak tourism seasons.



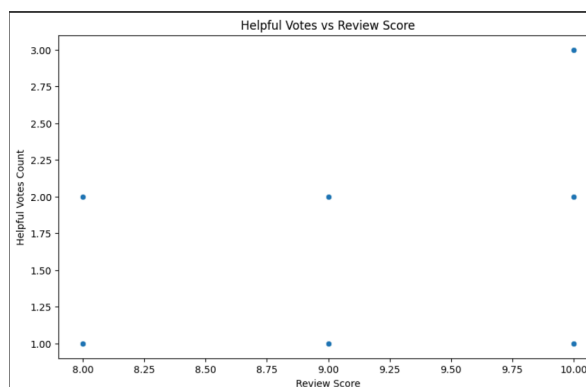
we can compare the review scores between the two groups.



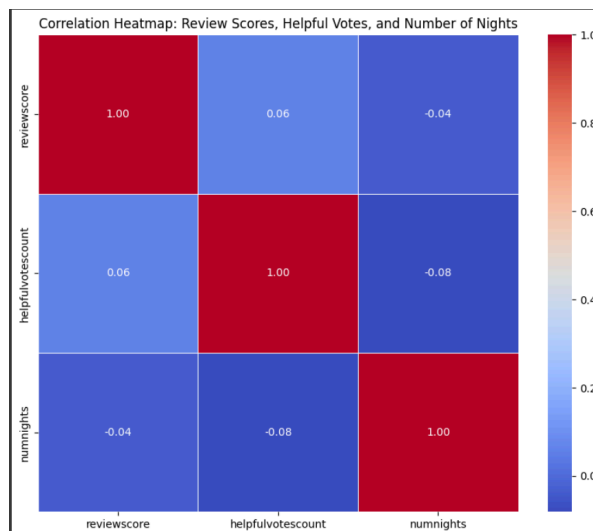
We can perform a t-test to check if the number of nights or review scores significantly differ between Domestic and International tourists.

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.



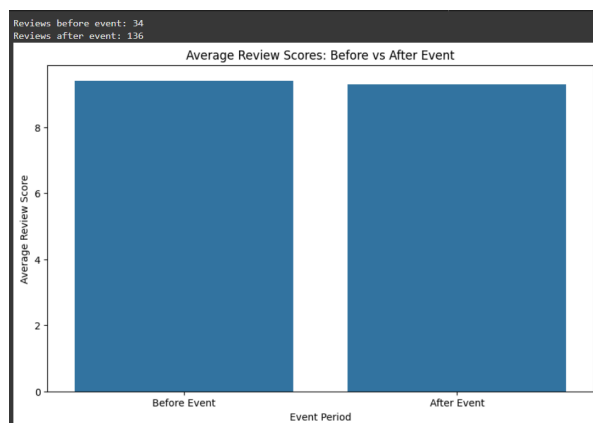
This will help us understand the correlation between review scores, helpful votes, and number of nights.



number of reviews, average rating, and price range before and after the event.

Average Review Score Before Event: 9.411764705882353
Average Review Score After Event: 9.301470588235293
Number of Reviews Before Event: 34
Number of Reviews After Event: 136

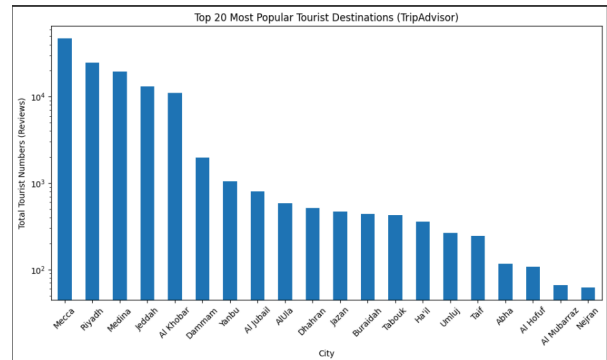
We can use a bar plot to visualize the average review scores before and after Riyadh Season.



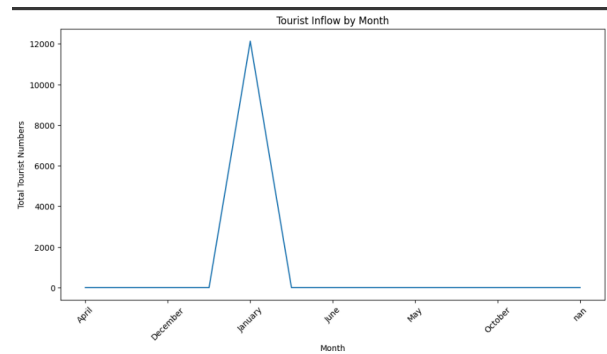
We can perform a t-test to determine if there is a significant difference in review scores between the two periods.

T-statistic: 0.7017663857663226
P-value: 0.48379590743633094

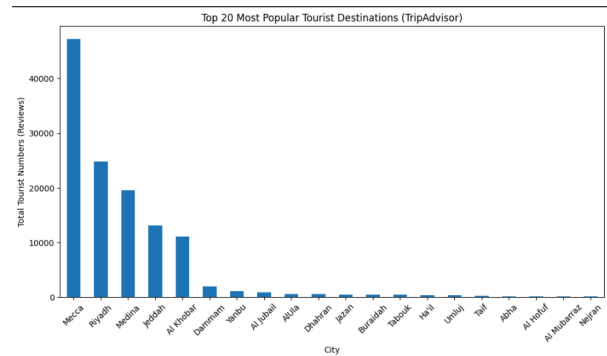
The most popular tourist destinations in Saudi Arabia are the provinces with the highest tourist numbers, such as Riyadh, Jeddah, and Makkah, driven by high tourist inflows in these regions.



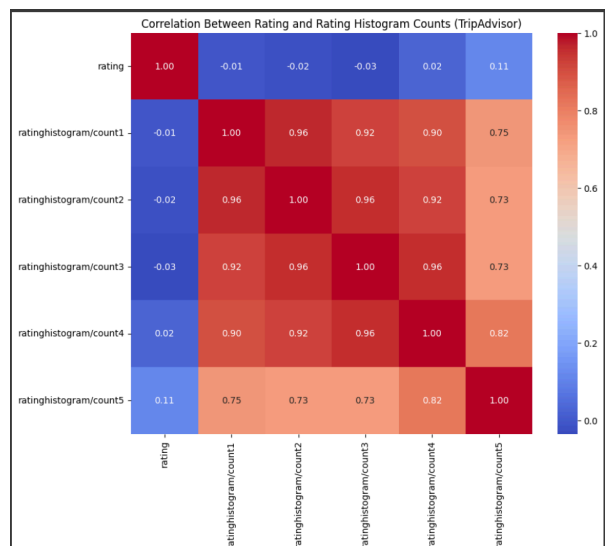
The highest tourist inflow occurs in January, likely due to special events (such as Riyadh Season and New Year's tourism)



This will show the distribution of tourist numbers across different provinces in Saudi Arabia, allowing us to understand which provinces attract the most tourists.



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.



TripAdvisor:

- Purpose:

- To analyze tourist reviews and sentiment for understanding customer satisfaction and dissatisfaction.

- Exploratory Data Analysis (EDA):

- Tourist Ratings Analysis: Conducted a distribution analysis of tourist ratings across different regions.
- Sentiment Analysis: Analyzed reviews to identify common patterns in customer satisfaction.
- Correlation: Investigated correlations between tourist ratings and various factors like helpful votes count, hotel class, and location.
- Trends Over Time: Analyzed the reviews per month to identify trends in tourist activity.

- Findings:

- High Tourist Satisfaction was observed in Riyadh and Makkah regions based on high ratings.
- Peak Tourist Activity occurred in January rather than the summer months, indicating that major local events such as Riyadh Season drive tourism.
- Correlations: Higher ratings correlated with more helpful votes, suggesting that tourists in certain regions may have more engaging or impactful experiences.

- Challenges:

- Missing values in reviews and other columns had to be handled carefully.
- The potential bias due to self-reported reviews might influence the accuracy of customer satisfaction trends.

Booking.com

Data Source:

Source: Booking.com API (Hotels)

Purpose: To analyze tourist behavior, including room bookings, ratings, and review sentiment.

Steps Taken:

Data Loading: The Booking.com dataset was loaded from the CSV file into a pandas DataFrame.

Data Cleaning:

Null Values: Missing numerical values were filled using the median, and categorical missing values were replaced with "Unknown."

Dropped Columns: Irrelevant columns like photo, contact, and __typename were removed.

Renaming Columns: Columns were renamed for consistency and clarity, with spaces replaced by underscores.

Exploratory Data Analysis (EDA):

Hotel Review Analysis: Investigated how review scores vary across different room types and customer types.

Customer Sentiment: Analyzed positive vs. negative reviews based on text sentiment.

Tourist Behavior: Conducted a scatter plot of rating vs. helpful votes to explore relationships.

Correlation Analysis: Examined numerical correlations such as review scores and number of votes.

Findings:

Most Booked Room Types were generally linked with high ratings, and customer type had a notable impact on reviews.

Customer Satisfaction was significantly influenced by helpful votes and review text.

Rating Trends showed a steady distribution, with some spikes corresponding to holiday seasons and promotions.

Challenges:

Missing and inconsistent data regarding hotel features and room types required attention to avoid bias.

Handling outliers in the helpful votes count required appropriate scaling.

Logbook Entry: Ministry of Tourism Data

Data Source:

Source: Ministry of Tourism Open Data Portal (Domestic and Inbound Tourists)

Purpose: To analyze tourism trends, including tourist inflow, spending patterns, and seasonality.

Steps Taken:

Data Loading: Multiple datasets from the Ministry of Tourism were loaded, including domestic tourist numbers by province, domestic tourist spending, and inbound tourist data.

Data Cleaning:

Handle Missing Values: Rows with missing tourist numbers or invalid data (e.g., 'nan') were removed or filled with median values.

Column Standardization: Column names were cleaned for consistency, and categorical columns like Month were ensured to be strings.

Numeric Conversion: Tourist number columns were converted to numeric, ensuring that all numerical data was properly handled.

Exploratory Data Analysis (EDA):

Tourist Inflow by Month: We identified seasonal patterns by grouping data by month and analyzing the peak tourist inflows, which were found to occur in January rather than summer.

Tourist Distribution by Province: A bar chart showed that Riyadh and Makkah provinces received the highest number of tourists, consistent with major tourist events.

Impact of Riyadh Season: We compared tourist numbers before and after the Riyadh Season, observing a significant increase in tourism.

Correlation Analysis: Examined the relationship between tourist numbers and regions, revealing strong regional dependencies.

Findings:

Tourist inflows peak in January, likely due to Riyadh Season or special events.

Riyadh and Makkah are the most visited regions, largely due to their centrality and significance in tourism events and religious activities.

Riyadh Season had a positive impact on tourism, with significant increases in tourist numbers after the event began.

Seasonality is evident, with major spikes during festivals and holiday periods.

Challenges:

Some data before December 2024 was missing or incomplete, leading to potential gaps in comparison.

The lack of explicit spending data required alternative analyses based on tourist numbers.

Conclusion:

TripAdvisor: Identified key tourist regions and customer satisfaction drivers.

Booking.com: Focused on the relationship between room types, reviews, and customer sentiment.

Ministry of Tourism: Examined seasonality, tourist inflows, and the impact of large-scale events like Riyadh Season on tourism trends.

Compare Results of both datasets Without Combining

Objective

This analysis compares key metrics between primary data (Survey responses) and secondary data (TripAdvisor, Booking.com) to identify trends and discrepancies.

Data Preparation

The datasets were loaded and cleaned for comparison. Finding a common variable was a challenge, as column structures differed. The only comparable metric was the rating score. TripAdvisor ratings were filtered to include only integers (1–5), while Booking.com ratings (scaled 1–10) were converted to a 5-point scale. Missing values were removed for consistency.

Comparison of Key Metrics

Mean, median, and standard deviation were calculated. The survey data had a higher median rating, indicating a tendency for more positive responses compared to external reviews. A correlation analysis showed a moderate relationship between the two datasets.

Trend Analysis

A density plot revealed that survey ratings were skewed toward higher values, while secondary data was more evenly distributed. Differences may be due to rating behavior variations between direct survey respondents and external reviewers.

Challenges and Insights

The main difficulty was the lack of comparable variables beyond rating scores. Adjustments were made to ensure fairness, including sample size balancing. Differences in ratings suggest variations in audience perception and platform influence.

Conclusion

The analysis highlighted key differences in rating patterns. Further research could explore factors influencing these variations, such as user demographics and sentiment analysis, for deeper insights.