https://github.com/FahdKhater/Data-Analysis-Project

Project Documentation

---

1. Data Sources, Collection & Methodology

1- Airbnb Pricing & Reviews (The "Economic" Layer)

- **Source**: Web-scraped data from Airbnb.com (via Python/Selenium).
- **Volume**: 1,400 Listings (Top ~100 per city).
- **Data Structure Snapshot**:

  City, Price_EGP, Rating, Reviews, Link

  Paris, 16301, 4.85, 120, airbnb.com/rooms/...

  Bangkok, 9837, 4.92, 45, airbnb.com/rooms/…

- **Justification for 1,400 Samples**:
  - Instead of demanding a census of every single listing (which includes inactive or low-quality options users never see), we simulated real-world user behavior. Tourists typically only view the first 3-5 pages of search results.

2- Official Tourist Arrivals (The "Physical Volume" Layer)

- **Source**: Web Scraped UNWTO (World Tourism Organization) Annual Reports.
- **Data Structure Snapshot**:

  City, International Tourist Arrivals (Millions)

  Bangkok, 30.3

  London, 22.7

- **Metrics**: Annual International Visitors (Millions). This serves as our metric for comparison of actual physical traffic against digital interest.

3- Travel Satisfaction Index (The "Quality" Layer)

- **Source**: Web Scraped Travel Safety & Quality Databases **Link**
- We couldn't get the actual Travel Satisfaction Index as it was hidden behind a paywall so we created our own travel satisfaction based on a weighted combination of Safety, Pollution, Traffic and Climate.

- **Data Structure Snapshot**:

  Target_City, Travel_Score, Safety, Pollution, Traffic, Climate

  Amsterdam, 94.5, 74.3, 22.6, 22.1, 87.5

  Sydney, 79.1, 66.1, 28.6, 43.3, 97.1

- **Metrics**: A composite *Satisfaction Score* (0-100) derived from weighted sub-indices: Safety (40%), Climate (20%), Pollution (20%), and Traffic (20%).

4- Google Trends Seasonality (The "Behavioral" Layer)

- **Source**: Google Trends API (via pytrends).
- **Volume**: 1,456 Weekly Data Points (14 cities * 52 weeks * 2 years).
- **Data Structure Snapshot**:

  date, continent, trend_index (0-1)

  2024-01-07, Africa, 0.82

  2024-01-14, europe, 0.75

- **Justification for Google Trends**:
  - Other than the fact that we couldn't find free datasets for Official tourism statistics, they are "Lagging Indicators", they tell us where people *went* months ago. Google Trends is a Leading Indicator of *Intent*. It captures the researching phase, revealing consumer desire before booking constraints (budget/visas) filter it out. We gather interest over time of searches related to visit "city".

Sampling Strategy: Why these 14 Cities?

We choose these cities for our study Paris, Barcelona, Tokyo, New York, London, Rome, Amsterdam, Sydney, Bangkok, Istanbul, Cairo, Rio de Janeiro, Venice, Los Angeles

- **Intersectionality Criteria**: We did not arbitrarily pick 14 cities. We selected the Intersection of cities that possessed high-quality data across all four disparate sources.
- **Global Representation**: We deliberately selected 2-3 major hubs from **every continent** (Total: 6 Continents) to ensure global coverage rather than a Europe-centric bias. This approach ensures 100% data completeness for the selected sample.

---

2. Question 1

Research Question: Do cities with higher Airbnb prices actually offer a better travel experience?

Preprocessing & Metrics

1. **Cleanup**: Drop duplicates and null values
2. **Aggregation**: Calculated **Median Price per City**. We consciously chose *Median* over *Mean* to prevent the extreme luxury outliers from skewing the representative price for a normal tourist.
3. **Merging**: Inner joined Pricing data with Travel Satisfaction Index on City.

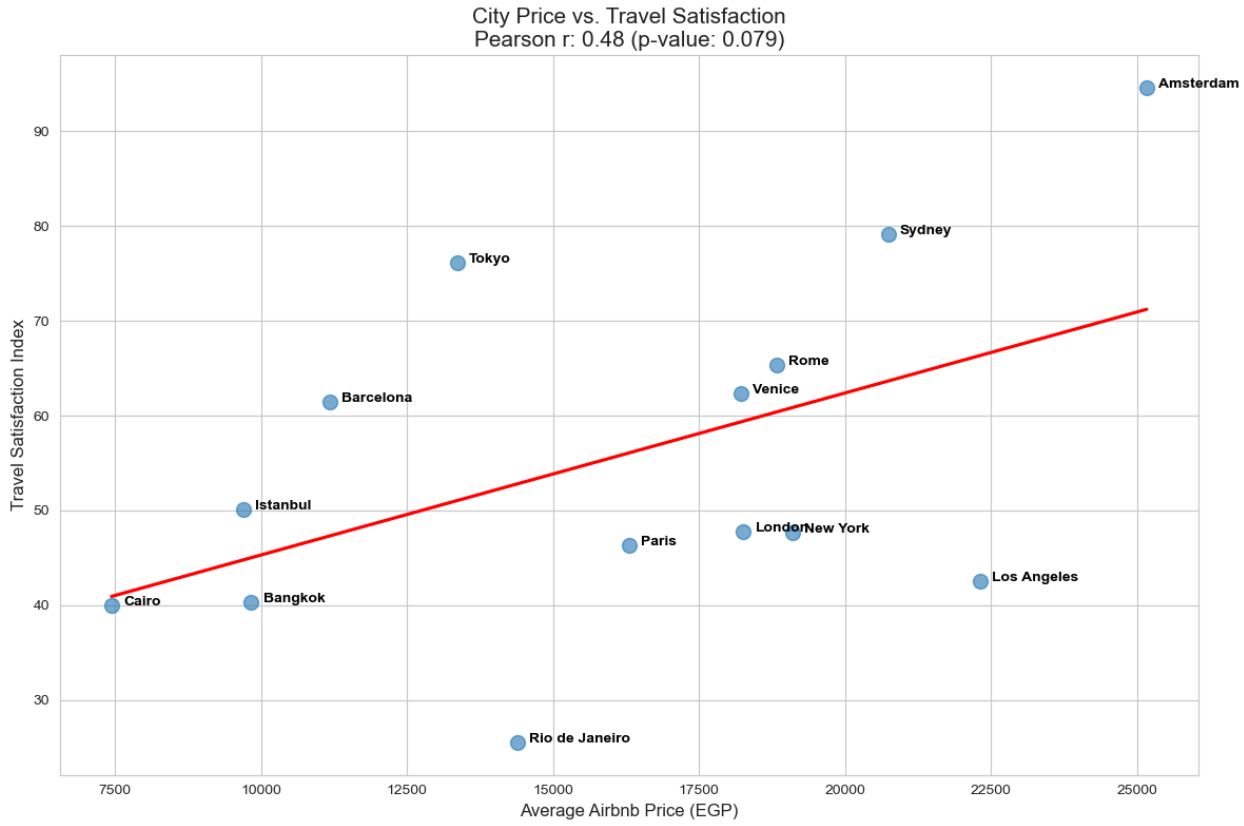
Analysis Narrative

We tested the relationship between Median Airbnb Price (Cost) and Travel Satisfaction Score (Quality).

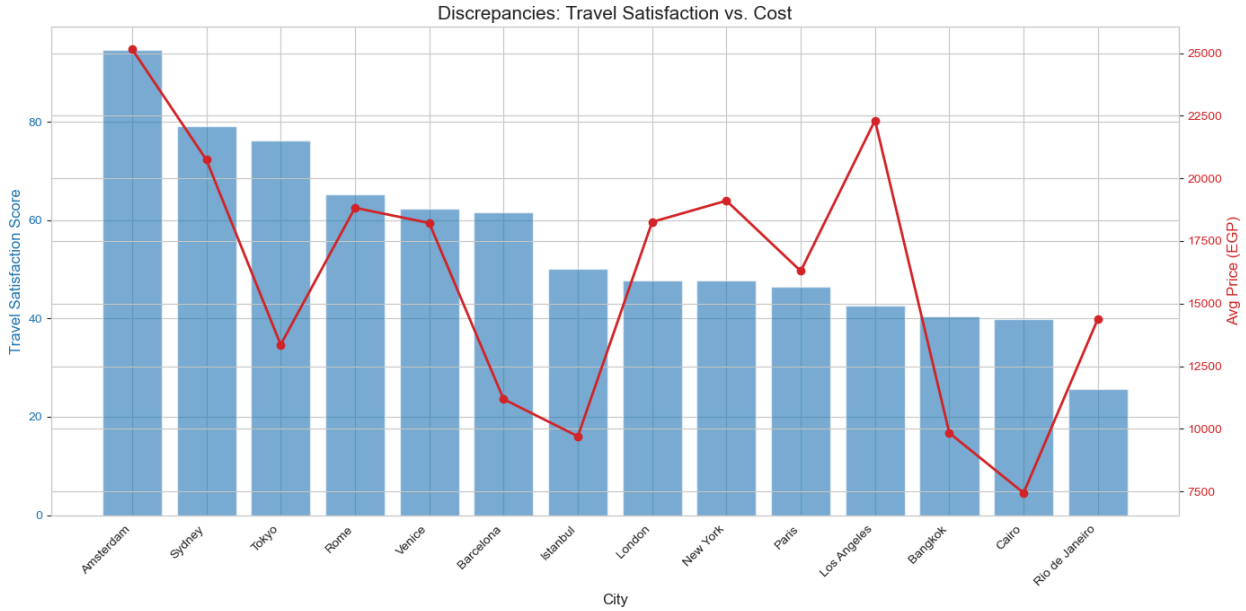
- **Method**: Pearson Correlation Coefficient (r).
- **Result**: r = 0.48, p = 0.079.

There appears to be a moderate positive association, as indicated by a Pearson Coefficient of 0.48. However, with a p-value of 0.079, this result is not statistically significant at the 0.05 level. While the trend suggests they may rise and fall together, we cannot rule out the possibility that this pattern is due to chance, likely due to the small sample size.

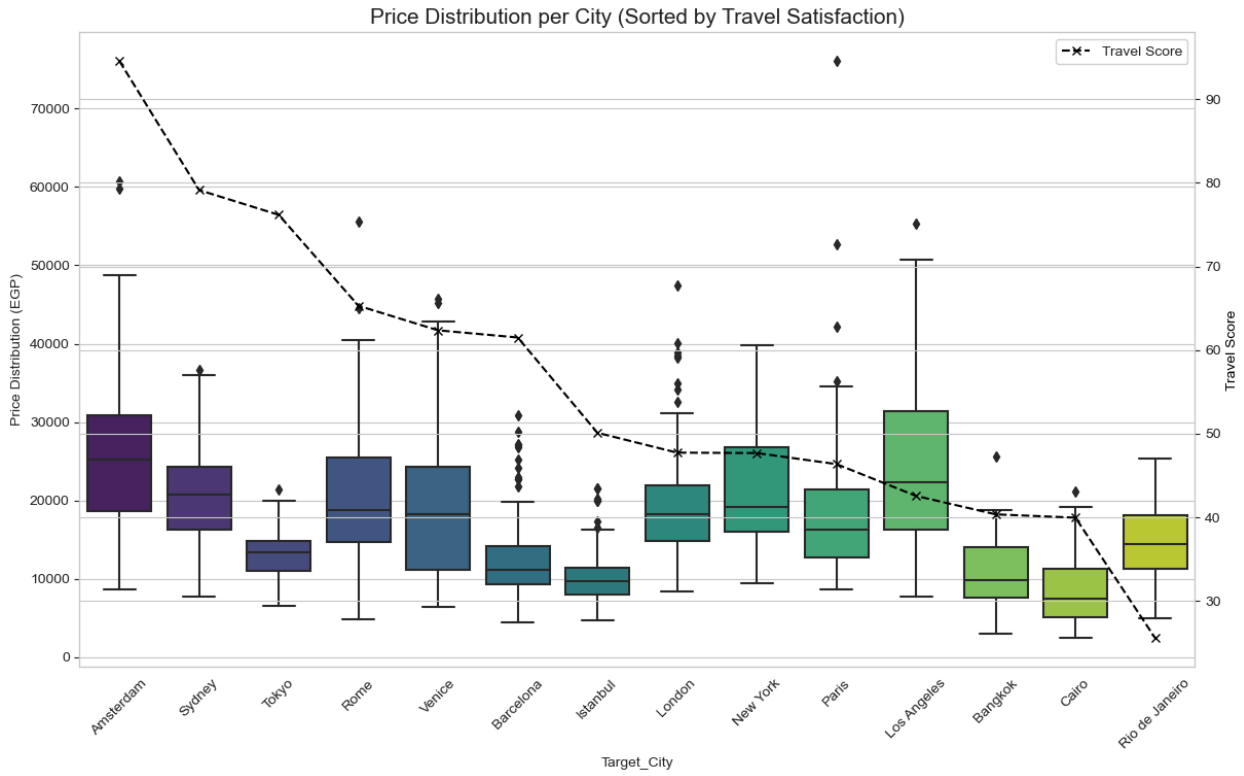
Visualizations



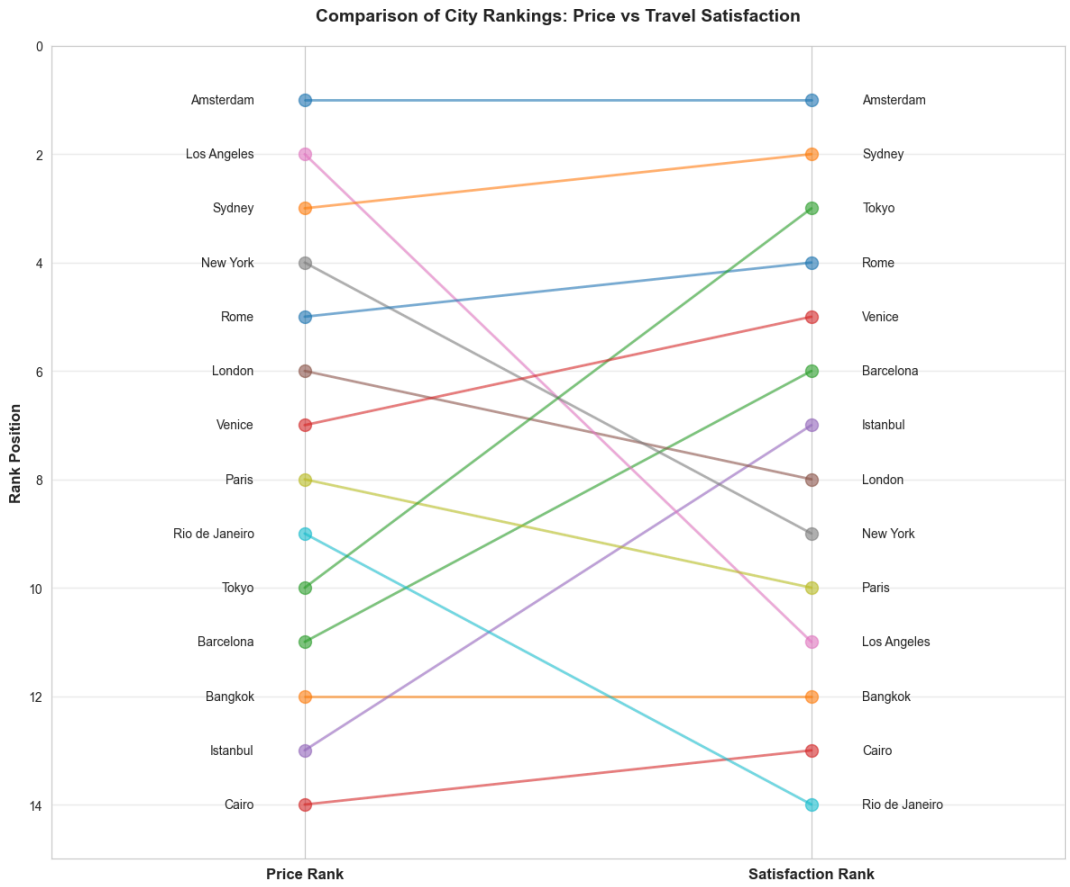
This graph illustrates a moderate positive correlation (Pearson r ≈ 0.48) between Airbnb prices and travel satisfaction, suggesting that cities with higher accommodation costs generally tend to have higher satisfaction ratings.



This visualization shows Amsterdam has the highest satisfaction score while Cairo and Rio have the lowest.



This visualization compares the median Airbnb prices across different locations, highlighting Amsterdam as the most expensive destination while showing Cairo and Bangkok as the most affordable options.



This shows that some countries have similar travel satisfaction and median while other have it different from each other.

---

3. Question 2

Research Question: Does the volume of online reviews predict actual tourist arrivals?

Preprocessing & Metrics

1. **Metric Calculation**: Aggregated Total Reviews = Sum of all individual review counts per city.
2. **Merging**: Joined against International Tourist Arrivals from the UNWTO dataset.
3. **Handling Outliers**: We deliberately retained massive outliers (like Bangkok's 30M visitors vs low reviews) because identifying these anomalies was the *goal* of the analysis, not an error to be cleaned.

Analysis Narrative

We hypothesized that "Review Count" is a proxy for "Popularity".

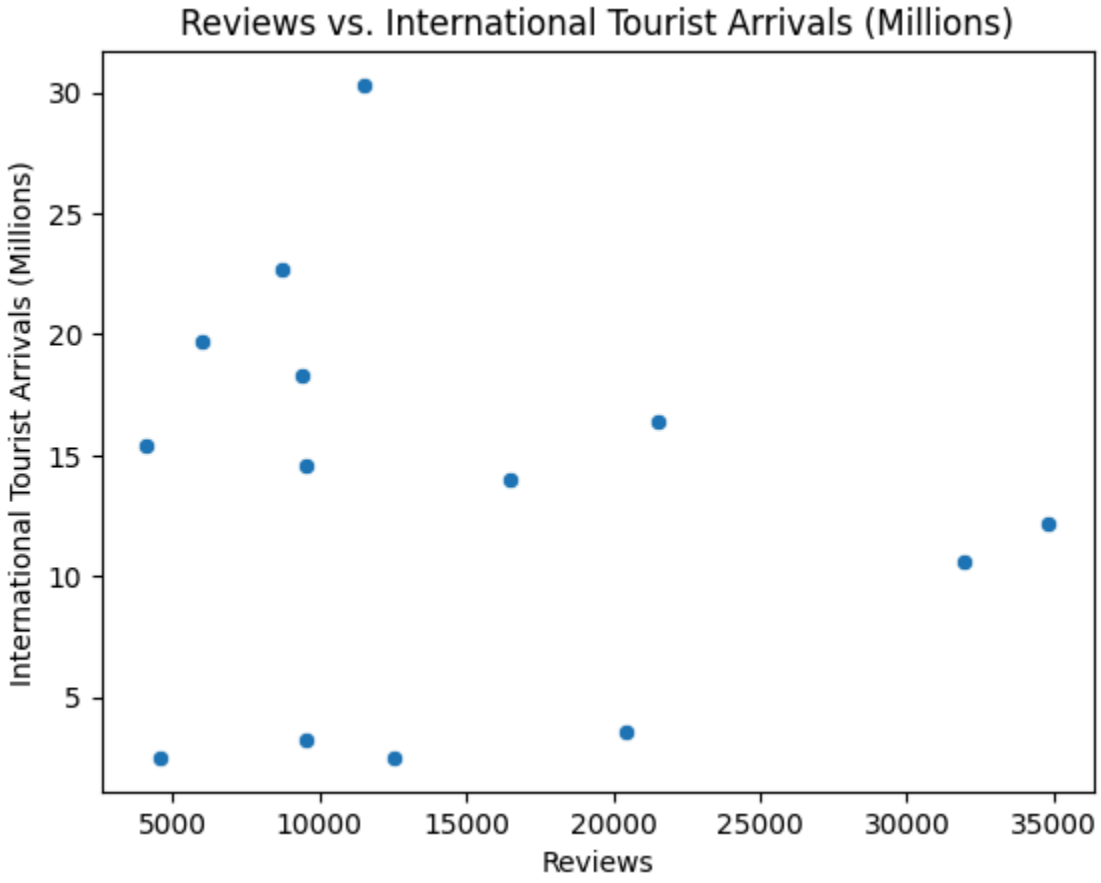
- **Result**: r = -0.13 (Negative/No Correlation).

Detailed Findings:

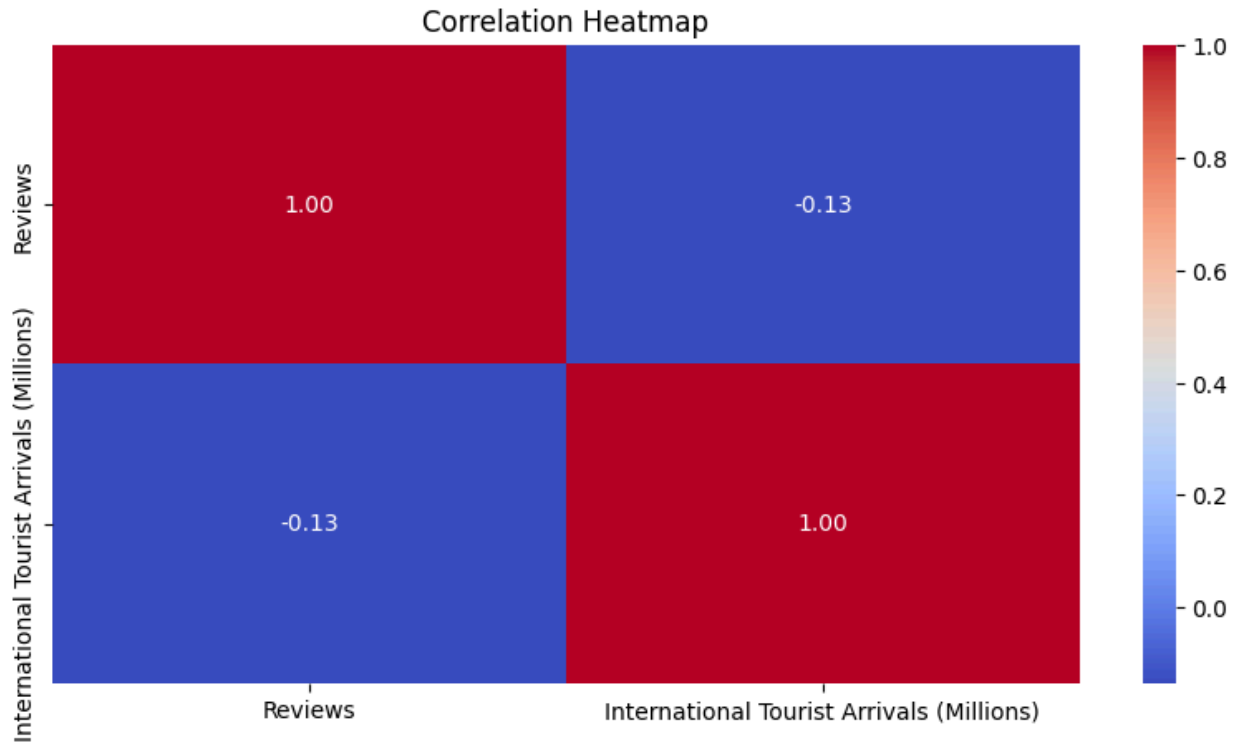
This result is counter-intuitive and highly insightful.

1. Cities in Asia (e.g., **Bangkok**, **Tokyo**) have massive physical tourism (30M+ visitors) but relatively low Airbnb footprints. This suggests these markets rely on hotels/Agoda, or that Airbnb faces regulatory hurdles there.
2. Western cities (London, Barcelona) generated disproportionately high reviews relative to their actual traffic, suggesting a "chattier" user base or higher market penetration.

Key Visualizations



We discovered there is barely any correlation with reviews and arrivals



We see only 0.13 correlation between the two variables.

---

4. Question 3

Research Question: ***Are certain regions more affected by seasonal tourism trends than others?***

Regions:

* Europe: Paris, Barcelona, Rome, Venice, London, Amsterdam

* Asia: Tokyo, Bangkok, Istanbul

* North America: New York, LA

* South America: Rio

* Africa: Cairo

* Oceania: Sydney

Preprocessing & Metrics

1. **Normalization**: Weekly Google Trends data (0-1) was aggregated by **Continent** to reduce city-specific noise.
2. **Metric 1: Coefficient of Variation:**: Standard Deviation divided Mean. Measures relative volatility.
3. **Metric 2: Seasonal Amplitude**: Calculated as peak value - low value. This quantifies the "Intensity of the Swing", how much harder the peak season hits compared to the average.

Hypothesis testing

Null Hypothesis (H0): All continents have the same variance (seasonality).

Alternative Hypothesis (H1): At least one continent has different variance.

We applied rigorous statistical testing to prove that seasonality varies by continent:

Two way Anova: Whether the average tourism level and seasonal patterns differ across continents. Continent × Week interaction $F = 3.99$, $p = 0.0016$ → Significant

Leveine test: Are some continents affected more strongly by seasonal tourism? p-value = 6.54×10^{-5} (< 0.05) → Reject H0

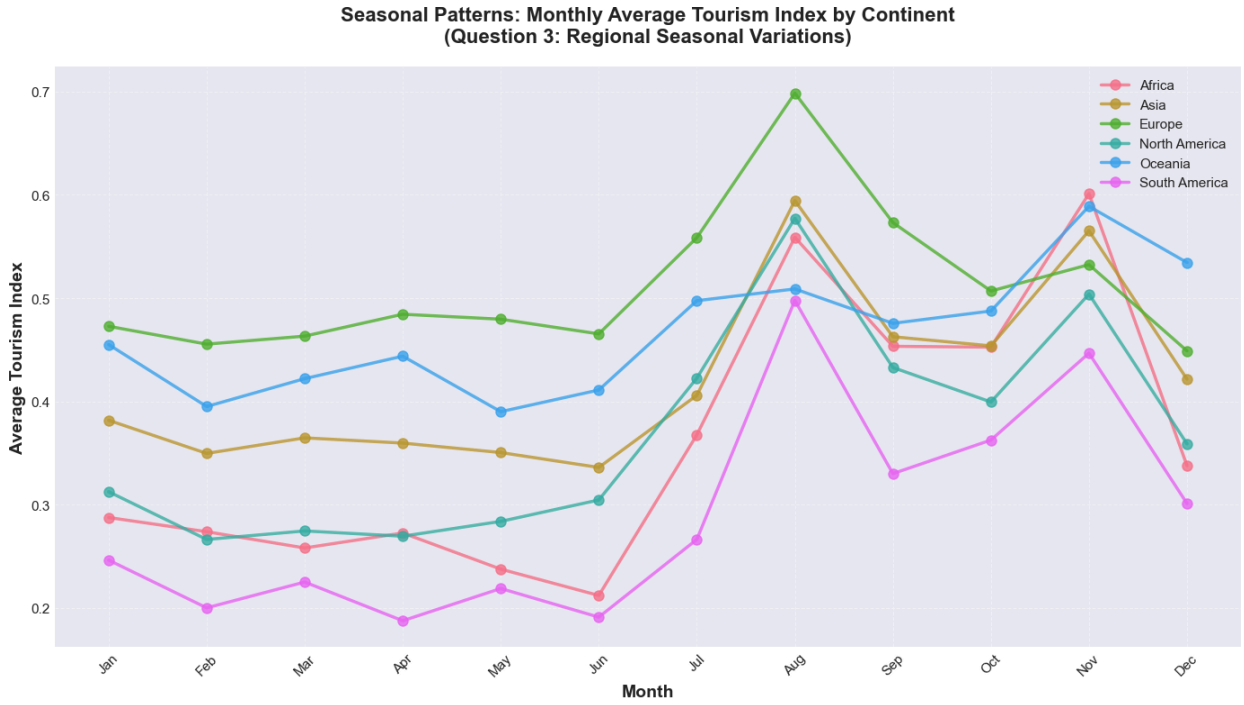
Detailed Regional Profile

Region	Peak Season	Amplitude %	Volatility (CV)	Classification
Africa	**November**	**183.5%**	**32.1%**	**Extreme**
S. America	**August**	**165.5%**	**35.8%**	**Extreme**
N. America	August	116.8%	24.8%	Moderate
Asia	August	76.9%	24.8%	Moderate
Europe	August	55.6%	25.2%	Stable
Oceania	**November**	**51.0%**	**15.7%**	**Highly Stable**

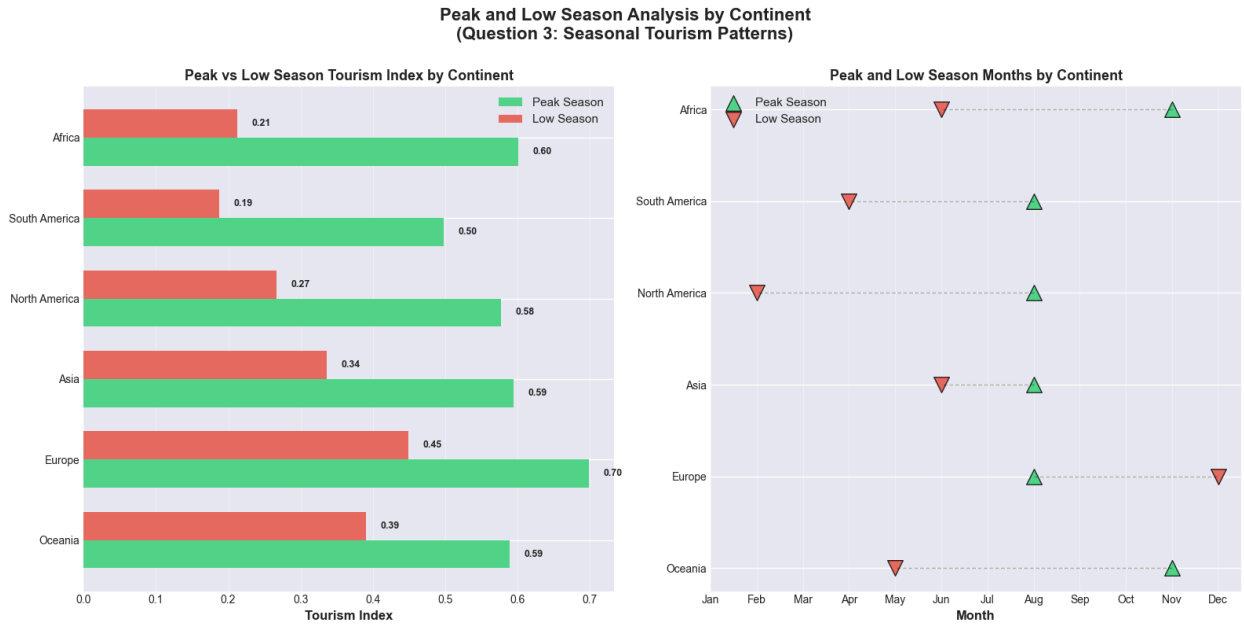
Key Findings

- **Africa** has an amplitude of **183%**. This means peak season traffic is nearly double the average, while low season (June) is a dead zone.
- **Oceania** (Sydney) is the most stable(CV 15%, Amplitude 51%). Demand is consistent year-round.
- **Africa** and **Oceania** peak in November while others peak in August which suggests a different travel season

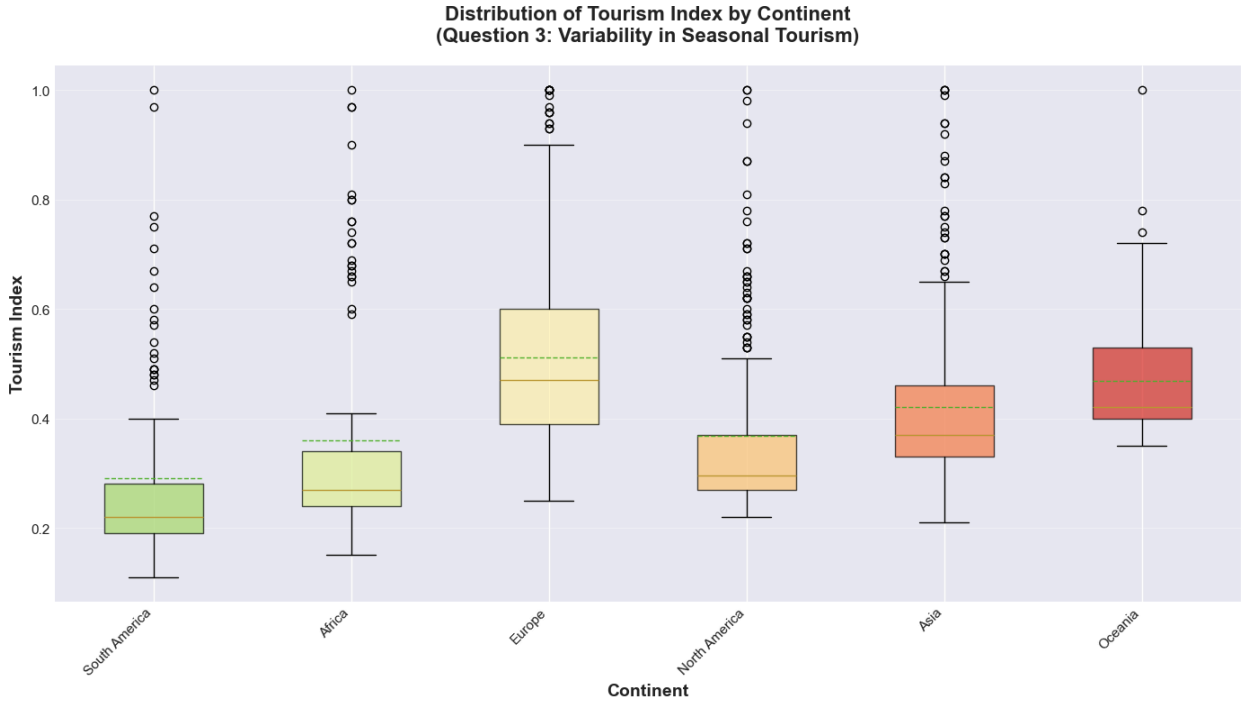
Visualization:



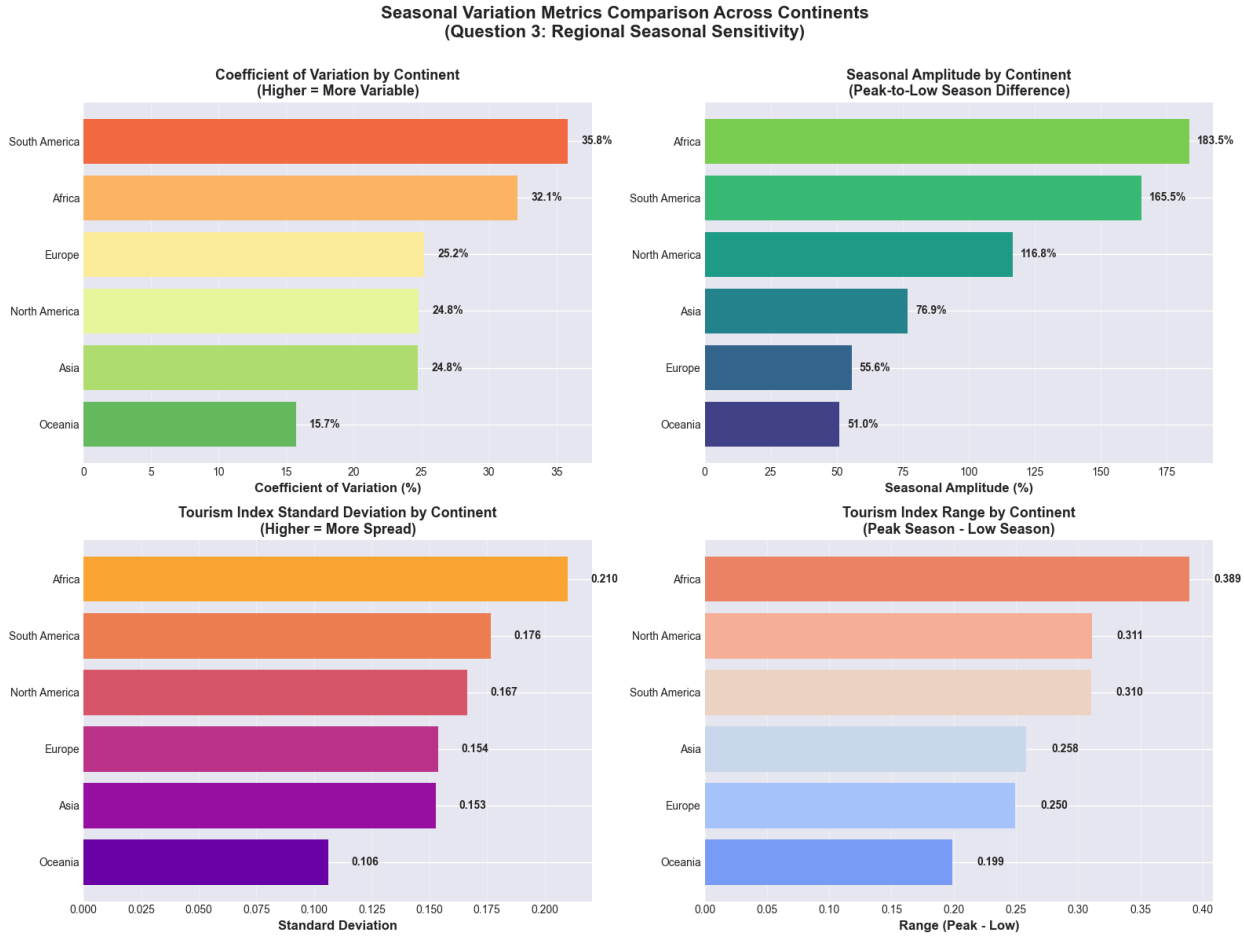
The steepness of each continent's curve demonstrates seasonal sensitivity, with most continents peaking in August (Northern Hemisphere summer) except Africa and Oceania which peak in November, revealing distinct hemispheric and regional tourism patterns.



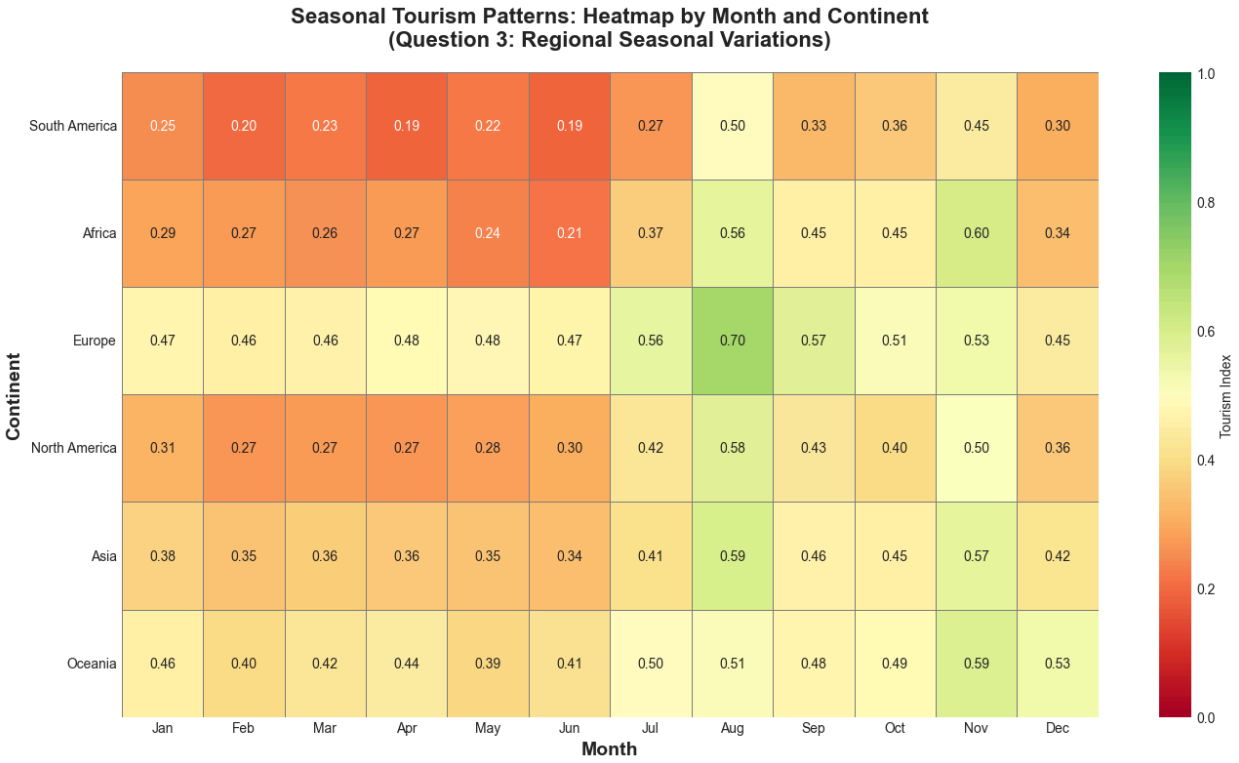
Africa shows the largest gap between peak (November: 0.60) and low (June: 0.21) season values, while the scatter plot reveals that different continents experience their tourism peaks and troughs at different times of the year, reflecting diverse seasonal drivers across regions.



Africa and South America display the widest boxes and longest whiskers with the most outliers, confirming their high seasonal variability, while Oceania shows the narrowest box indicating the most stable year-round tourism patterns among all continents



Africa and South America consistently show the highest seasonal variation across all four metrics (Coefficient of Variation, Seasonal Amplitude Percentage, Standard Deviation, and Range), indicating these regions are most affected by seasonal tourism trends



The color variation within rows reveals that Africa and South America exhibit the most dramatic seasonal changes (most color variation), while Oceania shows the most stable year-round tourism (consistent coloring), with August being the peak month for most continents.

---

5. Final Synthesis

This project demonstrates that successful tourism analysis requires looking beyond simple prices. By triangulating data from **scraped listings**, **official reports**, and **search behaviors**, we statistically proved that:

1. **Price ≠ Satisfaction** (r=0.48, non-linear).
2. **Online Hype ≠ Physical Traffic** (Asian markets operate differently).
3. Seasonality changes per region (africa and oceania behave differently)