EDA_Expedia

July 14, 2025

1 Expedia Consumer Analysis

1.0.1 Intro

This research project focuses on analyzing traveler behavior patterns, crafting data-driven insights, and developing effective marketing strategies that will enhance how Expedia's travel partners engage with customers throughout their decision-making process. By understanding the nuanced ways sponsored travel content influences consumer choices at different stages of the travel planning journey, we can optimize content strategy, improve conversion rates, and create more meaningful connections between travelers and travel providers.

The insights generated from this analysis will enable Expedia's partners to deliver more targeted, relevant, and effective sponsored content that resonates with consumers' evolving needs and preferences in an increasingly competitive marketplace.

I have conducted a Consumer insights survey and gained over a 1000 responses, with these responses I hope to be able to identify key demographics and their social media habits.

1.0.2 Executive Summary

In an increasingly crowded digital marketplace, Expedia must understand how travelers behave — not just where they go, but how they **plan, engage, and decide**.

This project explores: - Who Expedia's travelers are - How they interact with **sponsored content** - Which platforms shape their booking behavior - How Expedia can tailor **marketing strategies** to each group

2 Data Import and Cleaning

[24]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.decomposition import PCA
from sklearn.metrics import silhouette_score
import warnings
from IPython.display import display, HTML
```

```python
warnings.filterwarnings('ignore')

# Load the dataset
df = pd.read_csv('Expedia_Consumer_Survey_Responses.csv', encoding='utf-8')

# Basic info about the dataset
print("Dataset shape:", df.shape)
print("\
Column names:")
for i, col in enumerate(df.columns):
    print(f"{i+1}. {col}")
```

```
Dataset shape: (1069, 24)
Column names:
1. Timestamp
2. gender
3. What is your age group?
4. How often do you plan and book leisure travel (vacations, weekend trips,
etc.)?
5. How much do you spend on a typical trip?
6. What type of trips do you most enjoy? (Select up to 3)
7. How often do you use Social media platforms (Instagram, Facebook, TikTok)
such as Facebook, Twitter, TikTok, Instagram, Youtube?
8. Which Social media platforms (Instagram, Facebook, TikTok) Platform do you
use the most?
9. Have you ever come across a sponsored travel post or ad while browsing Social
media platforms (Instagram, Facebook, TikTok)?
10. If Yes: Which platform do you recall seeing sponsored travel content on most
often?
11. When you see sponsored travel content, how likely are you to take the
following actions?
12. Sponsored travel content helps me discover destinations or experiences I
might not have considered otherwise?
13. Which sources do you typically use FIRST when starting to research a new
travel destination?
14. Which type of travel content do you find most helpful when making booking
decisions?
15. When researching travel destinations or accommodations, how often do you
notice that some content is labeled as 'sponsored,' 'ad,' or 'promoted'?
16. How much do you agree with this statement: 'I trust user-generated content
(reviews, photos from other travelers) more than professionally produced travel
content.'
17. How long does your typical travel planning process take from initial
research to final booking?
18. On which device do you most often research and book travel?
19. How often do you seek opinions from friends, family, or online communities
before making a travel booking decision?
20. Which new travel content features would you find most helpful when planning
```

a trip?
21. Think about a recent travel booking you made. Describe how different types of content (photos, reviews, ads, recommendations) influenced your final decision. What made certain content more convincing or trustworthy to you?
22. Based on your experience with travel content and booking websites, what changes would make you more likely to trust and engage with travel recommendations or promotions? What frustrates you most about current travel content?
23. Name
24. Email

[26]:

```
# Hidden css: CSS for the grid layout
css = """
<style>
.grid-container {
    display: grid;
    grid-template-columns: 1fr 1fr;
    gap: 20px;
    margin-bottom: 20px;
}
.table-container {
    border: 1px solid #e0e0e0;
    border-radius: 5px;
    padding: 10px;
    background-color: white;
}
.demographic-table {
    width: 100%;
    border-collapse: collapse;
}
.demographic-table th, .demographic-table td {
    padding: 8px;
    text-align: left;
    border-bottom: 1px solid #ddd;
}
.demographic-table th {
    background-color: #f2f2f2;
}
h3 {
    margin-top: 0;
    color: #333;
}
</style>
"""

# CSS for the grid layout (enhanced for crosstabs)
css = """
```

```
<style>
.grid-container {
    display: grid;
    grid-template-columns: 1fr 1fr;
    gap: 20px;
    margin-bottom: 20px;
}
.table-container {
    border: 1px solid #e0e0e0;
    border-radius: 5px;
    padding: 10px;
    background-color: white;
}
.crosstab-table {
    width: 100%;
    border-collapse: collapse;
    margin-top: 10px;
}
.crosstab-table th, .crosstab-table td {
    padding: 8px;
    text-align: center;
    border: 1px solid #ddd;
}
.crosstab-table th {
    background-color: #f2f2f2;
    font-weight: bold;
    word-break: break-word;
    white-space: normal;
    max-width: 100px;
}
.crosstab-table tr:hover {
    background-color: #f5f5f5;
}
h3 {
    margin-top: 0;
    color: #333;
    border-bottom: 1px solid #eee;
    padding-bottom: 5px;
}
.pct-table {
    width: 100%;
    margin-top: 15px;
}
</style>
"""
```

[28]:

```
# Cleaning up column names for easier analysis
df.columns = [
    'timestamp', 'gender', 'age_group', 'travel_frequency', 'spending_amount',
    'trip_types', 'social_media_usage', 'primary_social_platform',
    'seen_sponsored_ads', 'sponsored_platform', 'sponsored_action_likelihood',
    'sponsored_discovery_agreement', 'research_sources', 'helpful_content_type',
    'notice_sponsored_frequency', 'trust_ugc_vs_professional',␣
  ↪'planning_duration',
    'research_device', 'seek_opinions_frequency', 'desired_features',
    'booking_influence_description', 'trust_engagement_suggestions', 'name',␣
  ↪'email'
]

# Remove unnecessary columns
columns_to_remove = ['timestamp', 'name', 'email',␣
  ↪'booking_influence_description', 'trust_engagement_suggestions' ]
for col in columns_to_remove:
    if col in df.columns:
        df = df.drop(col, axis=1)

print("Cleaned dataset shape:", df.shape)
print("Remaining columns:", len(df.columns))

# Check for missing values
missing_counts = df.isnull().sum()
print("\
Missing values per column:")
for col, count in missing_counts.items():
    if count > 0:
        print(col + ":", count)


# Display basic info and first few rows
print("Dataset Overview:")
print(f"Total responses: {len(df)}")
print(f"Columns: {len(df.columns)}")
print("\
First 5 rows:")
df.head()
```

```
Cleaned dataset shape: (1069, 19)
Remaining columns: 19
Missing values per column:
sponsored_platform: 144
Dataset Overview:
Total responses: 1069
Columns: 19
```

```
First 5 rows:
```

```
[28]:    gender age_group         travel_frequency  spending_amount  \
      0    Male        60  Less than once per year  more than $2500
      1  Female        60  Less than once per year  more than $2500
      2  Female        60  Less than once per year  more than $2500
      3  Female        60  Less than once per year  more than $2500
      4    Male        60  Less than once per year  more than $2500

                                                trip_types social_media_usage  \
      0  Visiting friends or family, Cultural / histori…              Never
      1  Visiting friends or family, City breaks, Welln…              Never
      2  Adventure / outdoor activities, Luxury travel,…              Never
      3       Budget travel, Beach holidays, Luxury travel              Never
      4  Luxury travel, Cultural / historical experienc…              Never

        primary_social_platform seen_sponsored_ads sponsored_platform  \
      0             None At All                 No                NaN
      1             None At All                 No                NaN
      2             None At All                 No                NaN
      3             None At All                 No                NaN
      4             None At All                 No                NaN

        sponsored_action_likelihood sponsored_discovery_agreement  \
      0                      Ignore                         Agree
      1                      Ignore                         Agree
      2                      Ignore                         Agree
      3                      Ignore                         Agree
      4                      Ignore                         Agree

                           research_sources         helpful_content_type  \
      0  Friends and family recommendations  Price comparisons and deals
      1  Friends and family recommendations  Price comparisons and deals
      2  Friends and family recommendations  Price comparisons and deals
      3  Friends and family recommendations  Price comparisons and deals
      4  Friends and family recommendations  Price comparisons and deals

        notice_sponsored_frequency trust_ugc_vs_professional planning_duration  \
      0              Always notice                  Disagree         1-3 weeks
      1              Always notice                  Disagree         1-3 weeks
      2              Always notice                  Disagree         1-3 weeks
      3              Always notice                  Disagree         1-3 weeks
      4              Always notice                  Disagree         1-3 weeks

        research_device seek_opinions_frequency  \
      0          Laptop                   Often
      1          Laptop                   Often
```

```
2          Laptop                     Often
3          Laptop                     Often
4          Laptop                     Often

                        desired_features
0  Budget calculators or cost breakdowns
1  Budget calculators or cost breakdowns
2  Budget calculators or cost breakdowns
3  Budget calculators or cost breakdowns
4  Budget calculators or cost breakdowns
```

3 Demographic Analysis

Before diving into clustering analysis and marketing strategy development, it's essential to understand the fundamental characteristics of our dataset through demographic analysis. This exploratory step provides crucial insights that will inform our clustering approach and validate our eventual marketing segmentation strategy.

3.1 Key Demographic Dimensions

We will examine the dataset across four critical demographic dimensions:

3.1.1 Gender Distribution

- Understanding gender representation in our travel customer base
- Identifying potential gender-based preferences or behaviors

3.1.2 Age Groups

- Analyzing generational differences in travel behavior and preferences
- Identifying age-specific marketing opportunities

3.1.3 Trip Spending Patterns

- Examining the distribution of spending levels across customers
- Understanding budget preferences and willingness to pay
- Identifying potential revenue opportunities and price sensitivity

3.1.4 Travel Frequency

- Analyzing how often customers travel throughout the year
- Identifying frequent vs. occasional travelers for targeted strategies

[31]:

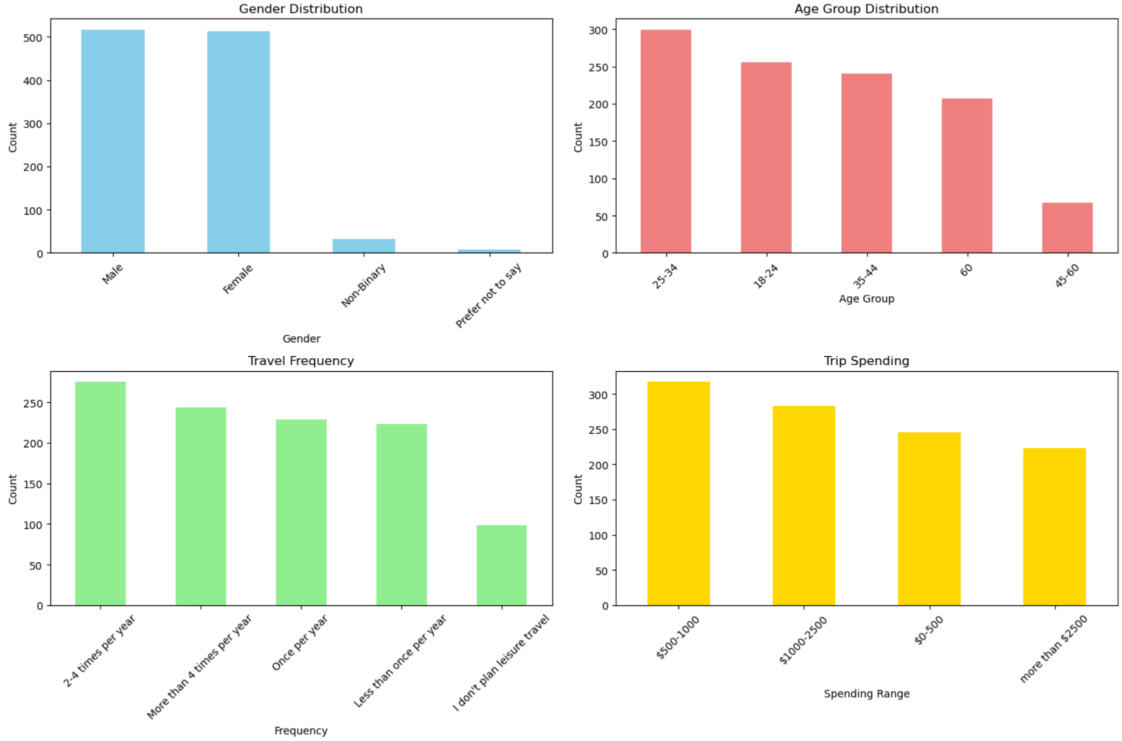
```
def create_demographic_table(df, column_name, display_name):
    counts = df[column_name].value_counts()
    percentages = round(df[column_name].value_counts(normalize=True) * 100, 1)

    result = pd.DataFrame({
        'Category': counts.index,
        'Count': counts.values,
        'Percentage (%)': percentages.values
    })

    # Format the table as HTML with CSS
    table_html = result.to_html(index=False, classes='demographic-table')
    return f"""
    <div class="table-container">
        <h3>{display_name}</h3>
        {table_html}
    </div>
    """

# Create the tables
tables_html = f"""
<div class="grid-container">
    {create_demographic_table(df, 'gender', 'Gender Distribution')}
    {create_demographic_table(df, 'age_group', 'Age Group Distribution')}
    {create_demographic_table(df, 'travel_frequency', 'Travel Frequency␣
  ↪Distribution')}
    {create_demographic_table(df, 'spending_amount', 'Trip Spending')}
</div>
"""

# Display everything
display(HTML(css + tables_html))
```

```
<IPython.core.display.HTML object>
```

[33]:

```
# Create visualizations for key demographic trends
fig, axes = plt.subplots(2, 2, figsize=(15, 10))

# Gender distribution
df['gender'].value_counts().plot(kind='bar', ax=axes[0,0], color='skyblue')
axes[0,0].set_title('Gender Distribution')
axes[0,0].set_xlabel('Gender')
axes[0,0].set_ylabel('Count')
axes[0,0].tick_params(axis='x', rotation=45)

# Age group distribution
df['age_group'].value_counts().plot(kind='bar', ax=axes[0,1],␣
  ↪color='lightcoral')
axes[0,1].set_title('Age Group Distribution')
axes[0,1].set_xlabel('Age Group')
axes[0,1].set_ylabel('Count')
axes[0,1].tick_params(axis='x', rotation=45)
```

```
# Travel frequency
df['travel_frequency'].value_counts().plot(kind='bar', ax=axes[1,0],␣
  ↪color='lightgreen')
axes[1,0].set_title('Travel Frequency')
axes[1,0].set_xlabel('Frequency')
axes[1,0].set_ylabel('Count')
axes[1,0].tick_params(axis='x', rotation=45)

# Trip spending
df['spending_amount'].value_counts().plot(kind='bar', ax=axes[1,1],␣
  ↪color='gold')
axes[1,1].set_title('Trip Spending')
axes[1,1].set_xlabel('Spending Range')
axes[1,1].set_ylabel('Count')
axes[1,1].tick_params(axis='x', rotation=45)

plt.tight_layout()
plt.show()

print("Demographics summary complete.")
```



```
Demographics summary complete.
```

3.2 Demographic Insights

- The sample is very well balanced with nearly equal representation of males (48.3%) and females (48.0%), plus a small but meaningful representation of non-binary respondents (3.1%).
- The largest segments are 25-34 year-olds (28.0%) and 18-24 year-olds (23.9%), indicating a younger-skewed sample that's highly relevant for digital marketing strategies.
- Most respondents are active travelers - 48.5% travel 2+ times per year, with only 9.2% not planning leisure travel at all.
- The spending is fairly distributed across ranges, with the largest group spending 500-1000 per trip (29.7%), followed by 1000-2500 (26.5%).

3.3 Social media behavior analysis

[37]:

```
tables_html = f"""
<div class="grid-container">
    {create_demographic_table(df, 'social_media_usage', 'Social Media Usage')}
    {create_demographic_table(df, 'primary_social_platform', 'Primary␣
  ↪Platform')}
    {create_demographic_table(df, 'seen_sponsored_ads', ' Seen Sponsored Ads ')}
    {create_demographic_table(df, 'sponsored_action_likelihood', 'Action to␣
  ↪Sponsored Content')}
</div>
"""

# Display everything
display(HTML(css + tables_html))
```

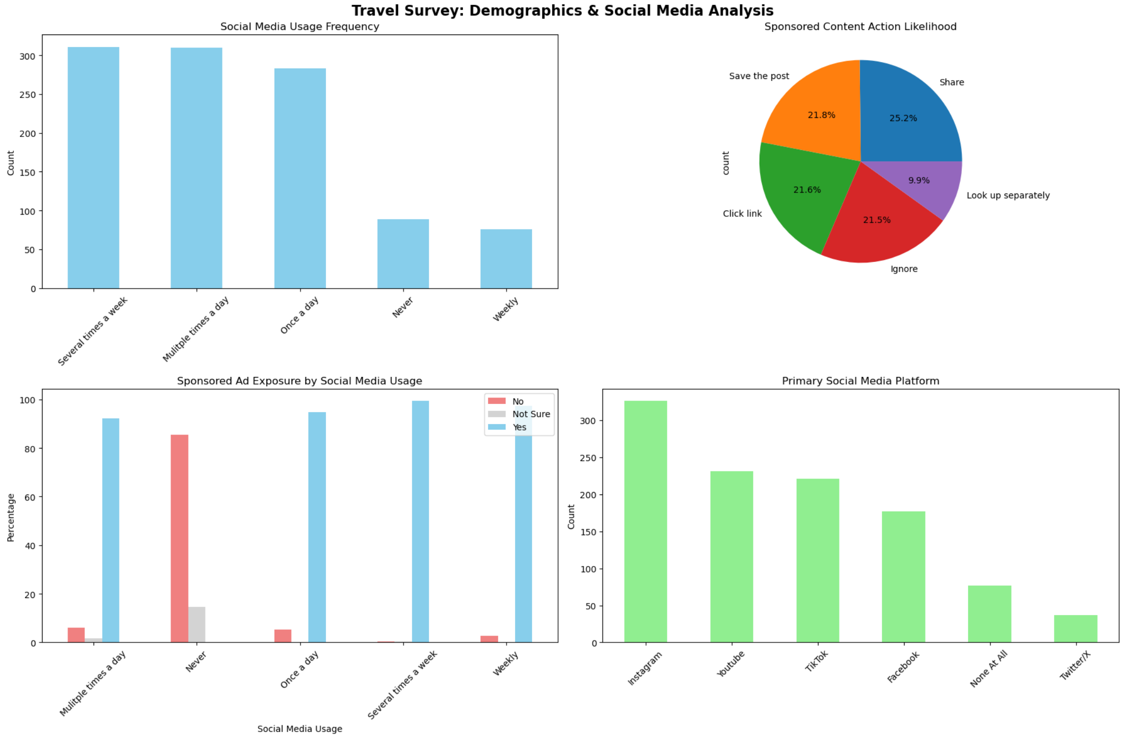
```
<IPython.core.display.HTML object>
```

3.3.1 Cross Tabular Analysis

[40]:

```
# Create visualizations for key demographics and social media trends
fig, axes = plt.subplots(2, 2, figsize=(18, 12))
fig.suptitle('Travel Survey: Demographics & Social Media Analysis',␣
  ↪fontsize=16, fontweight='bold')


# Social media usage
df['social_media_usage'].value_counts().plot(kind='bar', ax=axes[0,0],␣
  ↪color='skyblue')
axes[0,0].set_title('Social Media Usage Frequency')
axes[0,0].set_xlabel('')
axes[0,0].set_ylabel('Count')
axes[0,0].tick_params(axis='x', rotation=45)

# Action to Sponsored Content
df['sponsored_action_likelihood'].value_counts().plot(kind='pie', ax=axes[0,1],␣
  ↪autopct='%1.1f%%')
```

```
axes[0,1].set_title('Sponsored Content Action Likelihood')

# Primary platform
df['primary_social_platform'].value_counts().plot(kind='bar', ax=axes[1,1],␣
 ↪color='lightgreen')
axes[1,1].set_title('Primary Social Media Platform')
axes[1,1].set_xlabel('')
axes[1,1].set_ylabel('Count')
axes[1,1].tick_params(axis='x', rotation=45)

# Social media usage vs sponsored ads
colors = ['lightcoral', 'lightgray', 'skyblue']
social_sponsored_pct = pd.crosstab(df['social_media_usage'],␣
 ↪df['seen_sponsored_ads'], normalize='index') * 100
social_sponsored_pct.plot(kind='bar', ax=axes[1,0], color=colors)
axes[1,0].set_title('Sponsored Ad Exposure by Social Media Usage')
axes[1,0].set_xlabel('Social Media Usage')
axes[1,0].set_ylabel('Percentage')
axes[1,0].tick_params(axis='x', rotation=45)
axes[1,0].legend(['No', 'Not Sure', 'Yes'])


plt.tight_layout()
plt.show()
```



[42]:

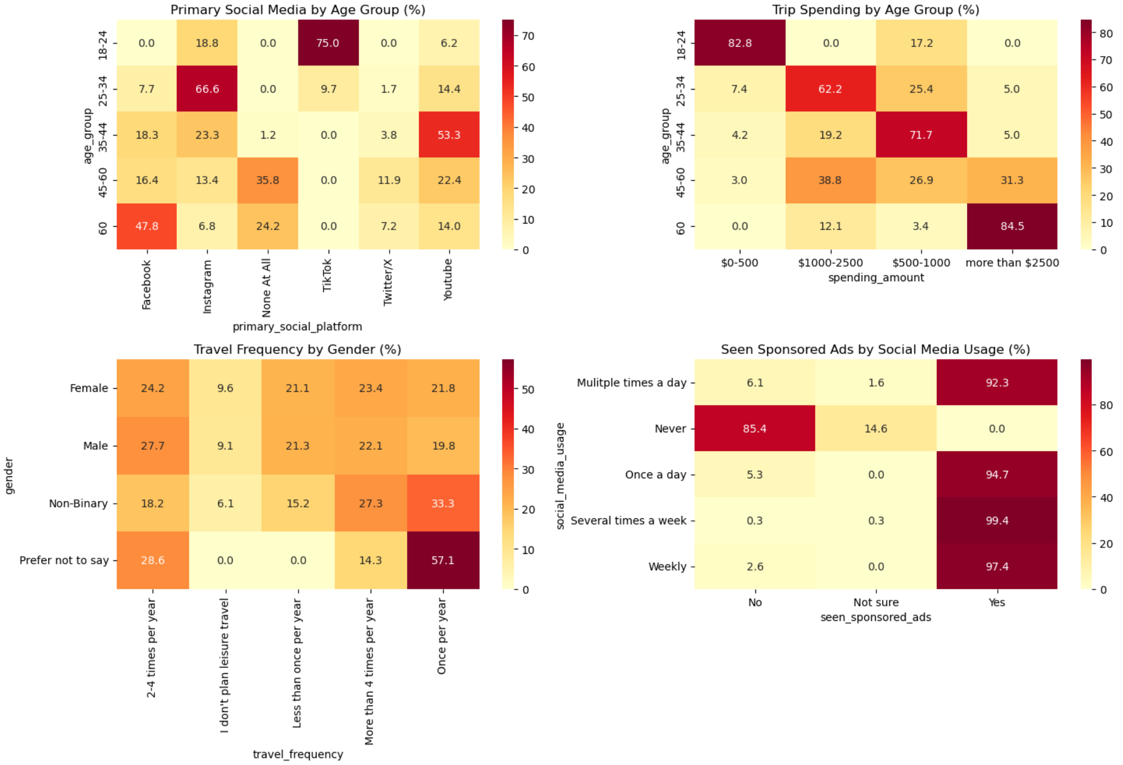
```
def create_crosstab_table(df, row_var, col_var, title):
    crosstab = pd.crosstab(df[row_var], df[col_var])
    row_pct = crosstab.div(crosstab.sum(axis=1), axis=0).round(2) * 100
    table_html = crosstab.to_html(classes='crosstab-table')
    pct_html = row_pct.to_html(classes='pct-table')

    return f"""
    <div class="table-container">
        <h3>{title}</h3>
        {table_html}
        <!-- Uncomment for percentages -->
        <!-- <h4>Row Percentages</h4> -->
        <!-- {pct_html} -->
    </div>
    """
# Create the cross-tabulation tables
tables_html = f"""
<h2 style="color: #2c3e50; margin-bottom: 20px;">Cross-Tabulation Analysis</h2>
<div class="grid-container">
    {create_crosstab_table(df, 'age_group', 'spending_amount', 'Age Group vs␣
 ↪Trip Spending')}
    {create_crosstab_table(df, 'age_group', 'primary_social_platform', 'Age␣
 ↪Group vs Social Media Platform')}
    {create_crosstab_table(df, 'gender', 'travel_frequency', 'Gender vs Travel␣
 ↪Frequency')}
    {create_crosstab_table(df, 'social_media_usage', 'seen_sponsored_ads',␣
 ↪'Social Media Usage vs Sponsored Ads')}

</div>
"""
display(HTML(css + tables_html))
```

```
<IPython.core.display.HTML object>
```

[44]:

```
#  visualization for social media and cross-tabulation insights
fig, axes = plt.subplots(2, 2, figsize=(15, 10))



# Age vs Social Media
age_social_pct = pd.crosstab(df['age_group'], df['primary_social_platform'],␣
 ↪normalize='index') * 100
sns.heatmap(age_social_pct, annot=True, fmt='.1f', ax=axes[0,0], cmap='YlOrRd')
axes[0,0].set_title('Primary Social Media by Age Group (%)')

# Age vs spending heatmap
```

```
age_spending_pct = pd.crosstab(df['age_group'], df['spending_amount'],␣
 ↪normalize='index') * 100
sns.heatmap(age_spending_pct, annot=True, fmt='.1f', ax=axes[0,1],␣
 ↪cmap='YlOrRd')
axes[0,1].set_title('Trip Spending by Age Group (%)')

# Gender vs Travel Frequency
gender_travel_pct = pd.crosstab(df['gender'], df['travel_frequency'],␣
 ↪normalize='index') * 100
sns.heatmap(gender_travel_pct, annot=True, fmt='.1f', ax=axes[1,0],␣
 ↪cmap='YlOrRd')
axes[1,0].set_title('Travel Frequency by Gender (%)')

# Gender vs Travel Frequency
social_sponsored_pct = pd.crosstab(df['social_media_usage'],␣
 ↪df['seen_sponsored_ads'], normalize='index') * 100
sns.heatmap(social_sponsored_pct, annot=True, fmt='.1f', ax=axes[1,1],␣
 ↪cmap='YlOrRd')
axes[1,1].set_title('Seen Sponsored Ads by Social Media Usage (%)')

plt.tight_layout()
plt.show()
```



3.4 # CROSS-SEGMENT INSIGHTS

3.5 Major Hot Spots & Implications

3.5.1 Primary Social Media by Age Group:

- **Hot spot**: 18-24 age group on TikTok (75.0%)
  Implication: Gen Z heavily dominates TikTok - prime target for youth marketing
- **Hot spot**: 25-34 on Instagram (66.6%)
  Implication: Millennials are Instagram's core - visual content marketing opportunity
- **Hot spot**: 35-44 on YouTube (53.3%)
  Implication: Older millennials prefer long-form video content
- **Hot spot**: 60+ on Facebook (47.6%)
  Implication: Older demographics still loyal to traditional social platforms

3.5.2 Trip Spending by Age Group:

- **Hot spot**: 18-24 in lowest spending (0-500: 82.8%)
  Implication: Young adults are budget-conscious travelers - target with deals
- **Hot spot**: 25-34 in mid-range (1000-2500: 62.2%)
  Implication: Peak earning millennials can afford moderate luxury
- **Hot spot**: 35-44 in higher spending (500-1000: 71.7%)
  Implication: Middle-aged travelers prioritize quality experiences
- **Hot spot**: 60+ in premium spending (2500+: 84.5%)
  Implication: Older travelers have highest disposable income for luxury travel

3.5.3 Seen Sponsored Ads by Social Media Usage:

- **Hot spot**: Heavy users seeing ads (94.8-100%)
  Implication: Algorithm targeting is extremely effective for frequent users
- **Hot spot**: Non-users NOT seeing ads (93.6% see none)
  Implication: Clear digital divide - offline marketing needed for non-users

3.5.4 Travel Frequency by Gender:

- **Relatively balanced** across all categories
  Implication: Gender isn't a strong predictor of travel frequency - focus on other demographics

3.6 Key Takeaways

1. **Age drives platform choice** - target different ages on their preferred platforms
2. **Spending power increases with age** - luxury travel marketing should focus on 60+ demographic
3. **Social media advertising works** - heavy users are saturated with ads
4. **Young travelers need budget options** - while older travelers will pay premium

# Booking habits

[48]:

```
tables_html = f"""
<div class="grid-container">
    {create_demographic_table(df, 'research_sources', 'Research Sources')}
    {create_demographic_table(df,  'helpful_content_type',  'Helpful Content')}
    {create_demographic_table(df, 'planning_duration', ' Planning Duration  ')}
    {create_demographic_table(df,     'research_device',     'Research Device')}
    {create_demographic_table(df,    'desired_features',     'Desired␣
 ↪Features')}

</div>
"""

# Display everything
display(HTML(css + tables_html))
```

```
<IPython.core.display.HTML object>
```

4 Expedia Consumer Survey - Booking Analysis

4.1 Survey Overview and Methodology

This analysis examines consumer behavior and preferences in travel research and booking through five key dimensions. Understanding these patterns is crucial for developing effective digital marketing strategies and improving customer experience in the travel industry.

4.2 Key Research Findings

4.2.1 1. Research Sources Analysis

Key Insights: - **Search engines remain dominant** (21.3%), confirming the importance of SEO and SEM strategies - **Social media platforms are nearly equal** (18.6%), highlighting the critical role of social media marketing - **Traditional sources still matter**: Travel booking websites (17.7%) and word-of-mouth (16.5%) remain significant - **Mobile apps underperform** (6.7%), suggesting opportunity for mobile engagement improvement

4.2.2 2. Helpful Content Preferences

Key Insights: - **Visual content dominates**: Video (32.5%) and photos (27.9%) account for 60.4% of preferences - **Price sensitivity is high**: 20.5% prioritize deals and comparisons - **User-generated content matters**: Reviews from travelers (14.5%) provide social proof - **Traditional text content underperforms**: Detailed descriptions only 1.0%

4.2.3 3. Planning Duration Patterns

Key Insights: - **Short-term planners dominate**: 56.9% plan 1-3 weeks ahead - **Spontaneous travel exists**: 21.3% plan within a week (including same-day) - **Long-term planners are minority**: Only 20.9% plan more than a month ahead - **Marketing timing is crucial**: Most campaigns should target 1-3 week planning window

4.2.4 4. Research Device Preferences

Key Insights: - **Mobile-first approach essential**: 46.6% primarily use smartphones - **Desktop/ Laptop still relevant**: 27.1% use Laptop/ Desktop computers - **Tablet market signif-

icant: 23.6% use tablets for research - **Multi-device usage minimal**: Only 1.3% use devices equally - **Responsive design critical**: Need optimization across smartphone, laptop, and tablet

4.2.5 5. Desired Features for Future Development

Key Insights: - **AI personalization in high demand**: 42.6% want AI itinerary suggestions - **Budget consciousness remains**: 16.1% want cost calculation tools - **Immersive technologies gaining traction**: 24.8% want 360° tours and AR previews - **Discovery features valued**: 6.5% want curated hidden gems

4.3 Strategic Implications

4.3.1 Content Strategy

- Prioritize video content creation (32.5% preference)
- Invest in high-quality destination photography

4.3.2 Platform Strategy

- Maintain strong SEO/SEM presence (21.3% use search engines)
- Develop comprehensive social media marketing (18.6% use social platforms)
- Optimize mobile experience (46.6% use smartphones)
-

4.3.3 Timing Strategy

- Target customers 1-3 weeks before travel (56.9% planning window)
- Develop last-minute deal campaigns (21.3% plan within a week)

4.3.4 Innovation Opportunities

- Develop AI-powered itinerary tools (42.6% demand)
- Create comprehensive budget calculators (16.1% demand)
- Invest in 360° video and AR technologies (24.8% combined demand)

5 Expedia Traveler Segmentation

5.0.1 Objective

Use unsupervised machine learning to identify meaningful clusters of Expedia survey respondents based on travel behavior, demographics, and social media usage.

5.1 Clustering Analysis

5.1.1 Discovering Natural Segments

Not all travelers behave the same — but are there **underlying patterns** we can trust?

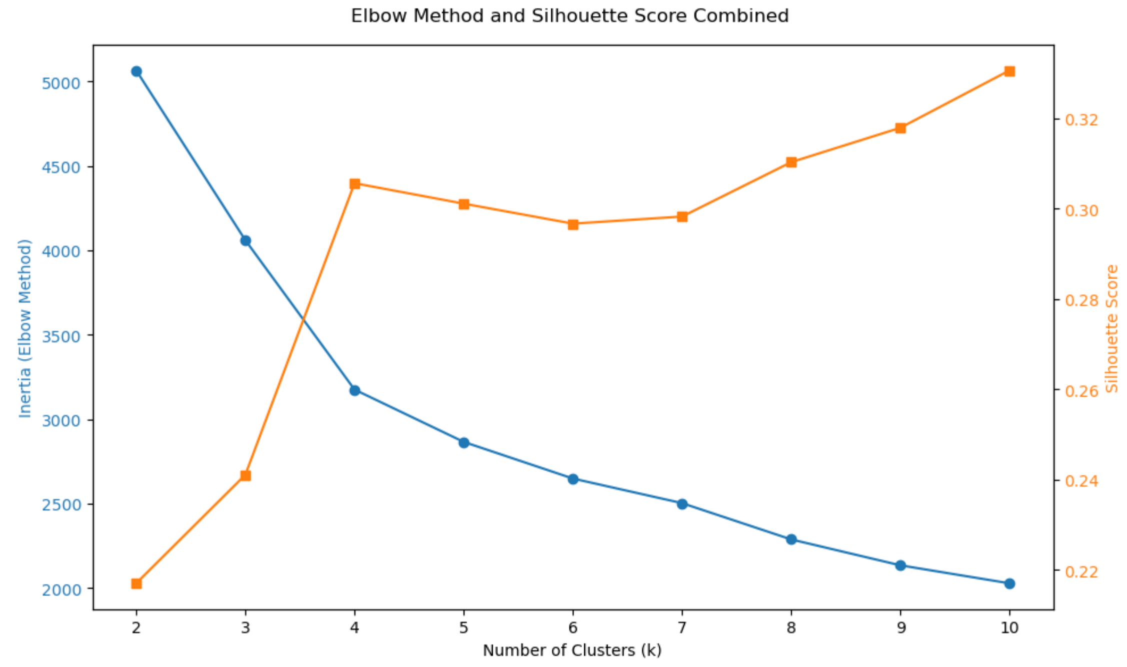
We used **unsupervised machine learning (K-Means clustering)** and dimensionality reduction (PCA) to explore the natural structure of traveler responses, focussing on the relation between Demographics and Social Media Patterns to know which groups to target and the best way to reach them.

The results suggest **4 strong clusters**. Here's how we know: - The **elbow method** shows diminishing returns beyond 4 clusters - The **silhouette score** confirms meaningful separation at k=4

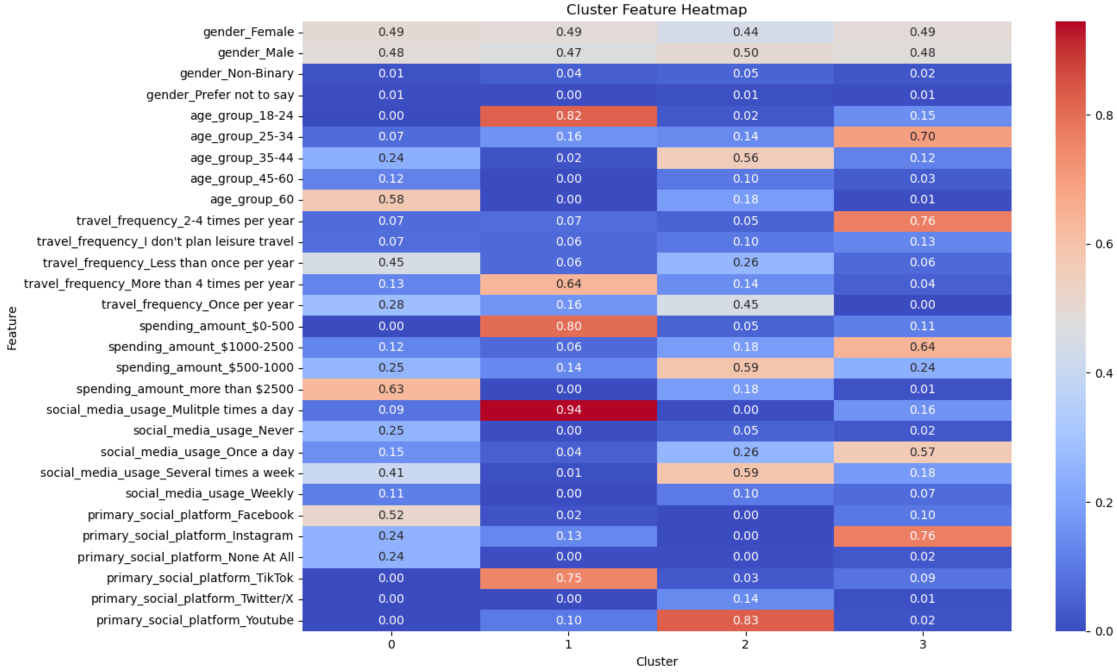
[52]:

```
## 2. Select Relevant Features
features = [
    'gender', 'age_group', 'travel_frequency', 'spending_amount',
    'social_media_usage', 'primary_social_platform'
]
df_cluster = df[features].copy()

## 3. Encode Categorical Features
le = LabelEncoder()
for col in df_cluster.columns:
    df_cluster[col] = le.fit_transform(df_cluster[col])

## 4. Scale Features
scaler = StandardScaler()
df_scaled = scaler.fit_transform(df_cluster)


range_n_clusters = list(range(2, 11))
inertia = []
silhouette_scores = []

# Recompute KMeans for each k and collect metrics
for n_clusters in range_n_clusters:
    kmeans = KMeans(n_clusters=n_clusters, random_state=42)
    cluster_labels = kmeans.fit_predict(df_scaled)
    inertia.append(kmeans.inertia_)
    silhouette_scores.append(silhouette_score(df_scaled, cluster_labels))

# Plot combined elbow and silhouette score
fig, ax1 = plt.subplots(figsize=(10, 6))

color = 'tab:blue'
ax1.set_xlabel('Number of Clusters (k)')
ax1.set_ylabel('Inertia (Elbow Method)', color=color)
ax1.plot(range_n_clusters, inertia, marker='o', color=color, label='Inertia')
ax1.tick_params(axis='y', labelcolor=color)

ax2 = ax1.twinx()
color = 'tab:orange'
```

```
ax2.set_ylabel('Silhouette Score', color=color)
ax2.plot(range_n_clusters, silhouette_scores, marker='s', color=color,␣
 ↪label='Silhouette Score')
ax2.tick_params(axis='y', labelcolor=color)

fig.suptitle('Elbow Method and Silhouette Score Combined')
fig.tight_layout()
plt.show()


## 6. Fit KMeans with Optimal K (e.g., 4)

optimal_k = 4
kmeans = KMeans(n_clusters=optimal_k, random_state=42)
df['cluster'] = kmeans.fit_predict(df_scaled)
```



[54]:
```
## 9. : Visual Heatmap of Encoded Features

encoded_features = pd.get_dummies(df[features])
cluster_means = encoded_features.groupby(df['cluster']).mean()

plt.figure(figsize=(14, 8))
sns.heatmap(cluster_means.T, cmap='coolwarm', annot=True, fmt='.2f')
plt.title('Cluster Feature Heatmap')
plt.xlabel('Cluster')
plt.ylabel('Feature')
plt.tight_layout()
plt.show()
```

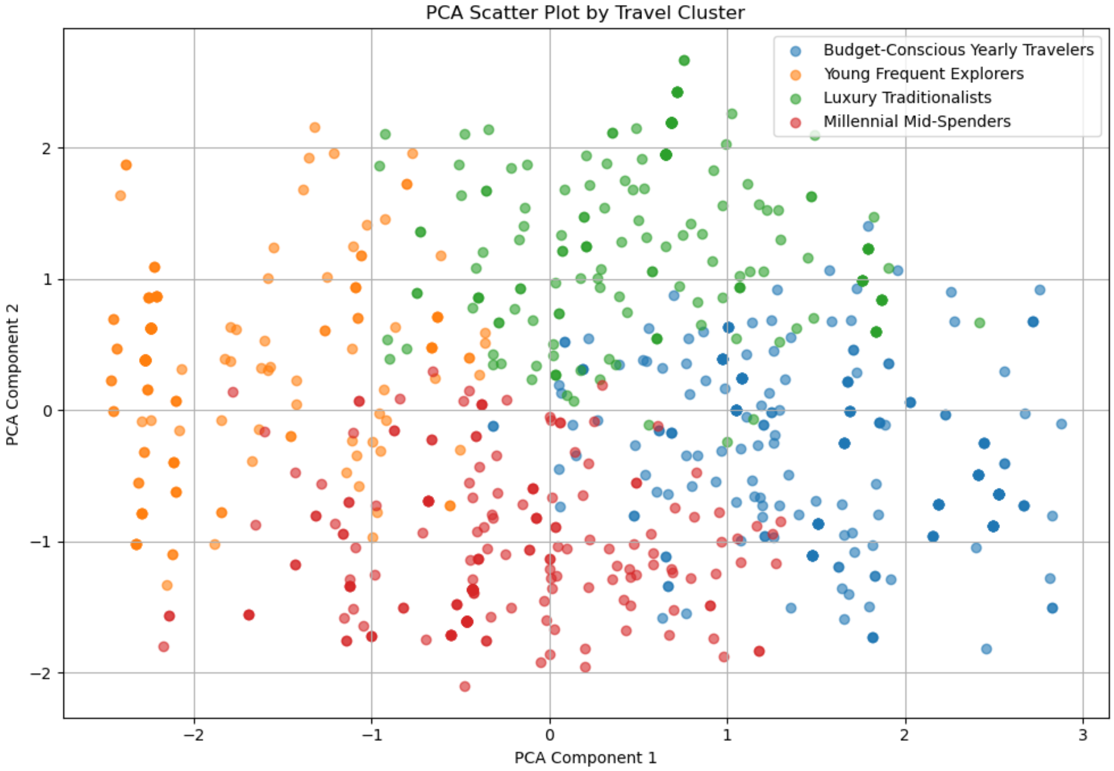


[56]:
```
# Group by cluster and compute the mode (most common response) for each column
cluster_characteristics = df.groupby('cluster')[features].agg(lambda x: x.
 ↪mode()[0])

# Add count of respondents per cluster
cluster_characteristics['count'] = df['cluster'].value_counts().sort_index()
cluster_characteristics
```

[56]:
```
         gender age_group           travel_frequency  spending_amount  \
cluster
0        Female        60     Less than once per year  more than $2500
1        Female     18-24  More than 4 times per year           $0-500
2          Male     35-44               Once per year        $500-1000
3        Female     25-34          2-4 times per year       $1000-2500

           social_media_usage primary_social_platform  count
cluster
0        Several times a week                Facebook    279
1        Mulitple times a day                  TikTok    251
2        Several times a week                 Youtube    243
3                  Once a day               Instagram    296
```

[58]:

```python
# Reduce dimensions to 2 for visualization
pca = PCA(n_components=2)
components = pca.fit_transform(df_scaled)

# Map cluster labels to names
cluster_names = {
    0: "Budget-Conscious Yearly Travelers",
    1: "Young Frequent Explorers",
    2: "Luxury Traditionalists",
    3: "Millennial Mid-Spenders"
}
df['cluster_name'] = df['cluster'].map(cluster_names)

# Plot PCA scatter with cluster names
plt.figure(figsize=(10, 7))
for label, name in cluster_names.items():
    plt.scatter(
        components[df['cluster'] == label, 0],
        components[df['cluster'] == label, 1],
        label=name,
        alpha=0.6
    )

plt.xlabel("PCA Component 1")
plt.ylabel("PCA Component 2")
plt.title("PCA Scatter Plot by Travel Cluster")
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.show()
```



5.2 Cluster Profiles & Marketing Recommendations

Based on our consumer survey data, we conducted a cluster analysis using behavioral features such as age, gender, spending, travel frequency, and social media use. We identified four distinct traveler segments, each with unique habits and content preferences.

This segmentation enables Expedia and its partners to deliver more targeted, relevant, and engaging travel content, increasing conversion rates and strengthening customer loyalty.

5.2.1 Cluster 0: Budget-Conscious Yearly Travelers

- **Age**: 35–44
- **Gender**: Mostly Male
- **Travel Frequency**: Once a year
- **Spending**: $500–$1000
- **Platform**: YouTube
- **Social Media Usage**: Several times a week
- **Count**: ~24% of users

Behavior Insight They are careful planners who rely on reviews and long-form content. Often book for family or practical purposes.

Marketing Strategy

- Partner with YouTubers focused on affordable travel guides and destination overviews.

- Emphasize **value-driven packages** and **trip bundling (flight + hotel)**.
- Offer tools like **price trackers, loyalty discounts**, or **early booking incentives**.
- Use **email retargeting** during peak holiday periods.

5.2.2 Cluster 1: Young Frequent Explorers

- **Age**: 18–24
- **Gender**: Mostly Female
- **Travel Frequency**: 4+ times per year
- **Spending**: Under $500
- **Platform**: TikTok
- **Social Media Usage**: Multiple times daily
- **Count**: ~26%

Behavior Insight Adventurous, highly social, and always on the lookout for a good deal. Travel is lifestyle—not just leisure.

Marketing Strategy

- Prioritize **short-form, viral video content** with trending audio and hashtags.
- Use influencers to showcase **budget travel hacks**, **group hostel stays**, or **impromptu weekend trips**.
- Highlight **flexible payment options** and **mobile-first experiences**.
- Offer curated guides for solo travel, digital nomads, and festival destinations.

5.2.3 Cluster 2: Luxury Traditionalists

- **Age**: 60+
- **Gender**: Mostly Female
- **Travel Frequency**: Rarely
- **Spending**: Over $2500
- **Platform**: Facebook
- **Social Media Usage**: Weekly or less
- **Count**: ~28%

Behavior Insight Prefer fewer trips, but when they travel, they want comfort, safety, and premium experiences. Trust and clarity matter.

Marketing Strategy

- Feature **high-end curated packages** (cruises, heritage tours, luxury resorts).
- Invest in **Facebook carousel ads** and newsletters that highlight **testimonials, ratings, and guarantees**.
- Prioritize **customer service messaging**—VIP treatment, refund policy, accessibility.
- Use **longer planning windows** and target during off-peak seasons.

5.2.4 Cluster 3: Millennial Mid-Spenders

- **Age**: 25–34
- **Gender**: Mostly Female
- **Travel Frequency**: 2–4 times per year
- **Spending**: $1000–$2500
- **Platform**: Instagram
- **Social Media Usage**: Daily
- **Count**: ~29%

Behavior Insight Enjoy balance: they value aesthetics and new experiences but have some spending power. Travel = self-care + social currency.

Marketing Strategy

- Use **Instagram stories + reels** with location tags, experiences, and influencer takeovers.
- Focus on **romantic getaways, wellness retreats**, and **experiential stays**.
- Showcase **eco-conscious choices**, **hidden gems**, and **photogenic places**.
- Offer **bundled upgrades** like spa access, breakfast included, or exclusive tours.

5.3 Final Thoughts

This behavioral segmentation can power Expedia's marketing strategy in several ways:

- Better **ad targeting and platform selection**
- Personalized **content strategy** by audience type
- More effective **conversion tracking** along the customer journey
- Improved value for **sponsored content partners**

5.4 Conclusion & Strategic Implications

This analysis reveals clear behavioral clusters among Expedia's digitally active travelers. By using K-Means clustering on key demographic and behavioral variables (age, gender, travel frequency, spending, and social media usage), we identified **four distinct traveler personas**. Each group demonstrates unique preferences in content type, platform engagement, and booking behaviors.

Key Takeaways: - **Young Frequent Explorers** (18–24, TikTok-centric) respond to short-form content, budget deals, and influencer travel hacks. - **Millennial Mid-Spenders** (25–34, Instagram) balance value and experience, prioritizing wellness, aesthetics, and social sharing. - **Budget-Conscious Yearly Travelers** (35–44, YouTube) plan deliberately and prefer reviews and price-sensitive tools. - **Luxury Traditionalists** (60+, Facebook) book less often but spend more, valuing premium experiences, trust, and simplicity.

Strategic Value: These insights offer a clear framework for tailoring Expedia's **sponsored content strategy**: - Match platforms and formats to each cluster's digital behavior - Personalize messaging around trip frequency and price sensitivity - Align influencer partnerships and visuals to segment-specific values

Operational Application: - Enhance ad targeting using cluster-informed personas - Guide media buys across channels like TikTok, Instagram, YouTube, and Facebook - Inform UI/UX improvements on Expedia's booking funnel (e.g., curated content, platform-specific landing pages)

This segmentation empowers Expedia and its partners to **move beyond generic campaigns**, and toward **micro-targeted engagement** that resonates with evolving traveler needs in a competitive, content-driven landscape.

[ ]: