Segment Analysis of AI-Powered Local Tourism Recommender

Prepared By: Vishal Dangiwala & Deepali Malik

GitHub Link: AI Powered Local Tourism Recommender

Problem Statement

Tourists exploring destinations in India often face challenges in discovering personalized travel recommendations that align with their interests, budget, and expectations. Existing platforms like Google Maps and TripAdvisor provide generic suggestions but fail to adapt dynamically to individual preferences, popularity trends, or customer sentiments reflected in reviews.

To bridge this gap, our project focuses on tourism segmentation based on key travel attributes. By analyzing geographic location, significance of places, and Google reviews, we aim to uncover patterns that influence tourist preferences and behaviors.

Our approach involves pairwise clustering using DBSCAN, K-Means, and Hierarchical Clustering to segment data in three key areas:

- **Geographic Clustering**: Identifying tourist hotspots based on location and popularity.
- **Significance-Based Clustering**: Understanding visit patterns based on place type, entrance fee, and visit duration.
- **Review-Based Clustering** : Analyzing how ratings, review counts, and entrance fees impact customer experiences.

By leveraging these insights, our study aims to assist businesses, tourism boards, and travelers in making data-driven decisions, leading to smarter travel recommendations and optimized tourism strategies.

Fermi Estimation

To construct an effective tourism segmentation model, it is crucial to estimate the usefulness of selected attributes. Fermi estimation helps assess whether our chosen segmentation approach is practical and relevant for identifying meaningful clusters in the dataset.

I. Identifying the Most Effective Segmentation Attributes

Tourism trends are influenced by multiple factors such as location, attraction significance, and customer reviews. While many attributes can impact travel decisions, our goal is to focus on those that provide the most actionable insights while ensuring data availability and computational efficiency.

After evaluating different factors, we selected the following three key attributes for segmentation:

- **Geographic Location**: Helps identify popular tourist zones and cities.
- **Significance of Place**: Determines how different types of attractions influence visit duration and cost.
- **Google Reviews**: Captures public perception and satisfaction levels.

These attributes are quantifiable, relevant, and practical, making them effective for clustering and segmentation.

II. Justifying Attribute Selection

To validate our selection, we assess how these attributes contribute to meaningful segmentation using real-world approximations:

1. **Geographic Location:** *Understanding Regional Tourism Trends*
    - Certain **zones** or **cities** receive more tourist attention due to accessibility, infrastructure, or cultural importance.
    - Popular destinations often have higher **review counts** and better **ratings**, while lesser-known places may lack visibility.

    Justification: Helps identify hotspots and optimize tourism marketing strategies.

2. **Significance of Place:** *Categorizing Attractions for Better Insights*
    - Attractions differ by type: **historical, religious, entertainment, natural, or cultural.**
    - Historical sites may have **higher entrance fees**, while religious sites might be free but require longer **visit durations**.

    Justification: Affects trip planning, pricing strategies, and visitor engagement.

3. **Google Reviews: Measuring Customer Perception**
    - **Higher-rated** places often attract more tourists, while low-rated places may indicate poor experiences.
    - **Review count** vs. **visit duration** helps determine if longer visits correlate with better reviews.

    Justification: Helps businesses improve services and provides insights into visitor satisfaction.

III. Why These Attributes Over Others?

Alternative factors like **distance from airport, permission for camera, or best time to visit** were considered but found to be:

- Inconsistent in available datasets.
- Less impactful in influencing segmentation compared to chosen attributes.

Our selection balances regional, financial, and behavioral insights, ensuring efficient clustering and practical dataset usability.

Data Sources

To implement **tourism segmentation**, we needed a dataset that effectively captures key attributes like **Geography, Significance, and Google Reviews**. After evaluating multiple datasets, we selected and refined one that aligns with our objectives.

- Source: Travel Dataset: Guide to India's Must See Places (Kaggle)
- **Reason for Selection:** This dataset provides rich insights into tourist **attractions, reviews, and visit patterns**, making it suitable for clustering analysis.

By structuring the dataset with clear attributes and meaningful categories, we ensure that our clustering approach effectively reflects real-world tourism trends and provides actionable insights.

Data Analysis

Preprocessing is essential to clean, transform, and prepare the data for meaningful analysis. Below are the key preprocessing steps along with their descriptions for your report:

1. Handling Missing Values

If any dataset entries contain missing values, they need to be handled. If too many values are missing in a column and it is irrelevant, then the column would be dropped to avoid bias.

2. Handling Duplicate Values

Duplicate values in a dataset can lead to skewed analysis and incorrect results. It's important to identify and handle them properly. If duplicates are found, they would be removed to ensure the dataset accurately reflects unique data points.

3. Feature Selection

Feature selection was performed to create distinct segmentation approaches, resulting in six datasets focusing on different clustering aspects:

A. Geographic-Based Clustering:

- **Zone, Number of Google Reviews:** Analyzing popularity by region.
- **City, Google Review Rating:** Identifying cities with high or low review ratings.

B. Significance-Based Clustering:

- **Significance, Entrance Fee:** Comparing costs across different significance types.
- **Significance, Visit Duration:** Estimating the time required for visits to different types of places.

C. **Google Review-Based Clustering:**

- **Review Rating, Entrance Fee:** Examining whether expensive places receive better reviews.

- **Review Count, Visit Duration:** Investigating if longer visits correlate with more reviews.

These feature sets provide insights into geographic trends, significance-based factors, and Google review influences, aiding in targeted clustering for segmentation analysis.

4. **Feature Encoding**

   To convert categorical variables into numerical values for clustering and analysis, Label Encoding was applied. This method assigns a unique integer to each category in the feature. Specifically:

   Label Encoding was used to convert categorical variables (e.g., city names, zone types, significance types) into numerical labels, making them suitable for clustering algorithms and further analysis.

   This approach ensures that the model can interpret categorical data while maintaining the inherent order or distinctness of the categories.

5. **Feature Scaling**

   To ensure that all features contribute equally to the clustering process, Feature Scaling was performed using Standard Scaler. This technique standardizes the features by transforming them to have a mean of 0 and a standard deviation of 1.

Machine Learning Techniques Used

To analyze and segment the Indian EV market, we used clustering techniques that grouped customers based on similarities in their preferences, behaviors, and demographics. The following three clustering algorithms were chosen:

1. **K-Means Clustering**

   Used to classify customers into well-defined market segments based on their purchase intent, environmental concerns, and income levels.

   Limitations:

   - Struggles with non-spherical clusters and noisy data**.**
   - Requires specifying the number of clusters (K) in advance.

2. **Hierarchical Clustering**

   Applied to observe the hierarchical relationship between segments and validate the number of clusters in the dataset.

   Limitations:

   - Computationally expensive for large datasets.
   - Sensitive to noise and outliers.

3. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**

   Utilized to detect outliers in consumer behavior and segment irregular patterns in adoption trends.

   Limitations:

   - Choosing the correct ε (epsilon) value is difficult.
   - Not suitable for datasets with varying densities in clusters.

Visualization of the Results

1. **Bar Plot**

   The bar plot shows the number of customers in each cluster, helping assess which segments are the largest.

2. **Pie Chart**

   A circular statistical graphic that represents data as proportional slices, showing the relative distribution of categories.

3. **Histogram**

   A bar graph that represents the frequency distribution of numerical data by grouping values into intervals**.**

Results

After segmentation, we derived key insights from different segmentation strategies. The visualizations helped us understand patterns in **geographic popularity, significance-based segmentation, and Google reviews**. Below are the findings:

I. Geographic-Based Clustering:

- **Zone vs. Google Reviews**: The **Northern** region is highly popular, while others have low engagement, indicating potential areas for tourism promotion.
- **City vs. Review Rating**: Certain cities consistently receive **higher ratings** as compared to others, helping identify top travel destinations.

II. Significance-Based Clustering:

- **Significance vs. Entrance Fee**: **Historical** sites tend to have higher entry fees, while religious sites are often free or low-cost.
- **Significance vs. Visit Duration**: **Entertainment and nature-based places** require **longer visit durations**, whereas religious sites have shorter visits.

III. Google Review-Based Clustering:

- **Review Rating vs. Entrance Fee**: In most of the cases, **expensive** places get **good reviews**,
- **Review Count vs. Visit Duration**: **Longer visits** correlate with **more reviews**, showing visitor engagement levels.

Conclusion

Through clustering algorithms and visual analysis, we uncovered valuable patterns in **geographic popularity, significance-based segmentation, and Google reviews**. The findings highlight key trends in visitor behavior, regional engagement, and pricing impact on reviews. These insights can help **businesses, tourism boards, and policymakers**:

- Improve customer targeting based on location and preferences.
- Optimize pricing strategies for different types of attractions.
- Enhance marketing strategies for different types of attractions.
- Personalize recommendations to improve visitor experience.

By leveraging these insights, we can make data-driven decisions to improve tourism planning and enhance customer satisfaction