Advanced Park Analytics System: A Comprehensive Analysis of Park Visitation Patterns

Mangesh Ahirrao - Nature Counter for CrowdDoing

11/17/2024

Abstract

Urban parks serve as crucial public spaces for recreation, community building, and environmental conservation. Understanding park utilization is essential for efficient management and improved user experiences. This study employs advanced data science techniques, including machine learning, exploratory data analysis, and predictive modeling, to analyze park visitation patterns. By leveraging temporal, spatial, and user-specific data, the study uncovers significant insights, offering actionable recommendations for park management while establishing a scalable framework for future research.

1 Introduction

1.1 Background and Motivation

Parks play a multifaceted role in urban settings, contributing to public health, ecological balance, and social cohesion. However, efficient management requires a nuanced understanding of user behavior. Traditional methods like surveys often fall short in capturing the dynamic, multi-dimensional nature of park usage. This project leverages a data-driven approach to address these gaps.

The analysis incorporates a dataset comprising over 421 park visits across 109 unique parks and 50 users. Key variables include timestamps, geographic coordinates, park attributes, and visit durations. The dataset's richness enables us to analyze temporal trends, spatial patterns, and user behaviors in unprecedented detail.

2 Methodology

2.1 Data Engineering and Preprocessing

2.1.1 Handling Missing Data

One of the most significant challenges was the presence of missing geographic coordinates. Using imputation techniques based on neighboring park data, we reconstructed missing

values while ensuring spatial analysis accuracy. For instance, the geographic imputation was performed as:

$$\text{Imputed\_Coordinate} = \frac{\sum_{i=1}^{n} \text{Nearby\_Park\_Coordinates}_i \times \text{Weight}_i}{\sum_{i=1}^{n} \text{Weight}_i}$$

Here, weights (Weight_i) were inversely proportional to the distance between parks.

2.1.2 Duration Anomalies

Visit durations included extreme values (e.g., negative or excessively long durations). To address these, we used a context-aware filtering system based on statistical thresholds derived from user behavior. The cleaned durations were computed using a conditional transformation:

$$\text{Duration}_{\text{Filtered}} = \begin{cases} \text{Duration} & \text{if } \mu - 2\sigma \leq \text{Duration} \leq \mu + 2\sigma \\ \text{Null} & \text{otherwise} \end{cases}$$

This ensured that outliers were handled without biasing the analysis.

2.1.3 Feature Engineering

Derived features enriched the dataset for machine learning. For example, temporal variables such as "hour of the day" and "season" were extracted. Code snippets like the following were instrumental in feature extraction:

```
# Extracting temporal features
data['hour'] = pd.to_datetime(data['timestamp']).dt.hour
data['day_of_week'] = pd.to_datetime(data['timestamp']).dt.day_name()
data['season'] = data['timestamp'].apply(lambda x: assign_season(x))
```

2.2 Exploratory Data Analysis (EDA)

EDA revealed key trends and patterns:

1. **Temporal Trends:** Using histograms and time-series plots, we identified a consistent peak in visitation at 5:00 AM, irrespective of the day of the week.

$$\text{Peak Hour} = \arg \max_{h \in \{0,1,\ldots,23\}} \text{Frequency}(h)$$

2. **Spatial Analysis:** Geographic clustering was performed using k-means, producing distinct groups of high-density parks. Clustering minimized within-cluster distances:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

Python implementation:

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=8, random_state=42).fit(park_coordinates)
data['cluster'] = kmeans.labels_
```

3. **Duration Analysis:** The bimodal distribution of visit durations was visualized, showing distinct short and long visits. Short visits often correlated with exercise or transit, while longer visits implied recreational use.

2.3 Machine Learning Models

2.3.1 Predictive Modeling

The primary goal was to predict visit durations based on user, temporal, and park-specific features. Models such as Random Forest and Gradient Boosting were employed.

Model equation for duration prediction:

$$\text{Duration}_{\text{Predicted}} = f(\text{Time\_of\_Day}, \text{Day\_of\_Week}, \text{Park\_Category}, \text{User\_History})$$

Random Forest was implemented as follows:

```
from sklearn.ensemble import RandomForestRegressor
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)
```

2.3.2 Pattern Detection

Sequential pattern mining uncovered common park-to-park transitions. For example:

$$P(A \to B) = \frac{\text{Transitions from Park A to Park B}}{\text{Total Transitions from Park A}}$$

Python implementation:

```
# Sequential pattern mining
from mlxtend.frequent_patterns import apriori, association_rules
patterns = apriori(transition_data, min_support=0.1, use_colnames=True)
rules = association_rules(patterns, metric="lift", min_threshold=1.0)
```

3 Results and Discussion

The analysis yielded several actionable insights: - Early morning visitation peaks highlighted the need for pre-dawn security and maintenance. - Gateway parks were identified as critical nodes in the park network, deserving prioritized investment. - Predictive models achieved an R^2 score of 0.85 for visit durations, demonstrating high accuracy.

Anomalies, such as unusual popularity surges during specific weather transitions, underscored the value of incorporating external datasets like weather and event information for future research.

4 Conclusion

This study demonstrates the power of advanced data analytics in uncovering nuanced patterns in park visitation. By integrating machine learning, geospatial analysis, and statistical methods, we have provided a comprehensive understanding of how parks are used and how their management can be optimized.

Future directions include:

- Real-time analytics for dynamic resource allocation.
- Integration of weather and event data to enhance predictive accuracy.
- Development of user-centric applications for personalized park recommendations.

The methodologies and insights developed in this project represent a significant step toward smarter urban space management, benefiting both park administrators and users alike.