Analytical Report

Vehicle Count Forecasting and Parking Optimization for the CalIt2 Building

1. Introduction and Topic Selection

This report presents a comprehensive analysis of vehicle count data for the CalIt2 Building parking facility, culminating in a prescriptive recommendation for optimal lot capacity. The overall goal of this study is to transform raw sensor data into actionable insights to improve parking management, reduce congestion, and optimize resource allocation.

1.1 Topic Justification

The topic, Vehicle Count Forecasting and Parking Optimization, was chosen because it addresses a critical, real-world operational challenge faced by facility managers: balancing supply and demand for a finite resource (parking spaces). The available data, consisting of high-frequency vehicle counts and external event schedules, is perfectly suited for a multi-stage analytical approach that covers all three major types of business analytics: descriptive, predictive, and prescriptive. By successfully applying these methods, the study provides a robust framework for managing dynamic parking demand, particularly during high-traffic periods such as scheduled events.

2. Data Collection and Validation

The foundation of this analysis rests on two primary datasets: vehicle entry/exit counts and a log of scheduled events. The integrity and appropriateness of this data are paramount to the trustworthiness of the final recommendations.

2.1 Data Sources and Quantity

The core data consists of vehicle counts recorded at 5-minute intervals over an extended period. This high-frequency time-series data is ideal for capturing the fine-grained, dynamic nature of parking lot usage. A secondary dataset contains information on scheduled events (e.g., games), including start/end times, attendance, and outcomes. The combination of internal usage data and external influence data (events) provides a rich context for modeling demand. The resulting dataset, with over 10,000 5-minute interval records, is appropriate for robust time-series modeling.

2.2 Data Cleaning and Trustworthiness

To ensure the reliability of the analysis, a rigorous data cleaning process was applied:

1. Temporal Alignment: Separate 'Date' and 'Time' columns were merged into a single datetime index, creating a unified time-series structure.
2. Error Imputation: Sensor errors, recorded as -1 values, were identified and treated as missing data (NaN).
3. Missing Value Handling: A total of 2,903 missing values were imputed using a combination of forward-fill (ffill) and backward-fill (bfill). Crucially, this imputation method was chosen over interpolation to ensure that all vehicle counts remained integers, as fractional vehicle counts are physically impossible. This decision maintains the real-world integrity of the data.
4. Feature Engineering: The event data was merged with the vehicle count data to create two binary indicator columns: event_active (1 if an event was ongoing) and event_day (1 if an event was scheduled for that day).

This systematic cleaning and validation process ensures that the data used for subsequent analysis is accurate, complete, and trustworthy.

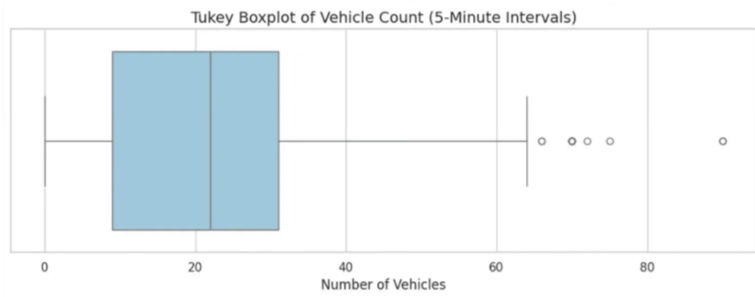
2.3 Outlier Detection and Central Tendency Analysis

To better understand the distribution of vehicle counts and identify unusual traffic patterns, we performed an outlier detection analysis using **Tukey's method**, a widely accepted statistical approach based on the interquartile range (IQR).

Using this method:

- We first calculated the 25th percentile (Q1) and 75th percentile (Q3) of the dataset.
- The interquartile range was computed as IQR = Q3 – Q1.
- Any vehicle count value below Q1 – 1.5×IQR **or above** Q3 + 1.5×IQR was classified as an outlier.

This analysis revealed 12 outliers, all of which corresponded to significant spikes in vehicle activity, most likely during event days or periods of abnormal parking demand.



To summarize the central tendencies of the vehicle traffic:

- **Mean vehicle count** = 21.07
- **Median vehicle count** = 22.00
- **Standard deviation** = 13.02

The close proximity of the mean and median suggests a fairly symmetric distribution, while the standard deviation reflects moderate variability across the time series. These summary statistics, along with outlier detection, helped contextualize typical parking behavior versus exceptional traffic spikes.

```
Mean: 21.07
Median: 22.0
Standard Deviation: 13.02
Lower Bound: -24.00
Upper Bound: 64.00
Number of Outliers Detected: 12
All Outliers Detected (Tukey Method):
                     Count  event_active  event_day
datetime
2005-04-13 22:10:00     66             0          1
2005-07-03 23:00:00     66             0          1
2005-07-03 23:40:00     72             0          1
2005-07-14 08:25:00     90             0          1
2005-07-14 08:30:00     90             0          1
2005-07-14 08:50:00     70             0          1
2005-07-14 08:55:00     70             0          1
2005-07-14 09:00:00     70             0          1
2005-07-14 09:05:00     70             0          1
2005-07-14 09:10:00     70             0          1
2005-07-15 14:40:00     75             0          1
2005-08-26 22:30:00     70             0          1
```

3. Descriptive Analytics and Visualization

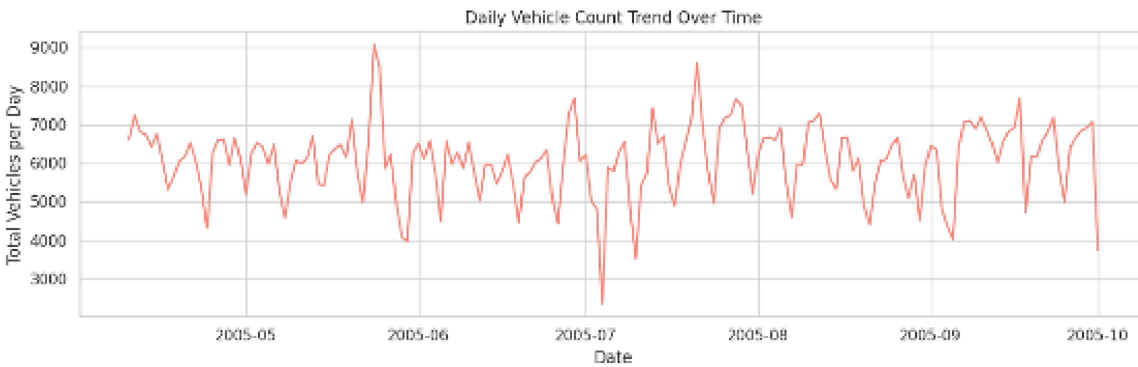
Descriptive analytics was used to summarize the historical data, identify key usage patterns, and provide a foundational understanding of parking demand drivers.

3.1 Key Findings and Interpretations

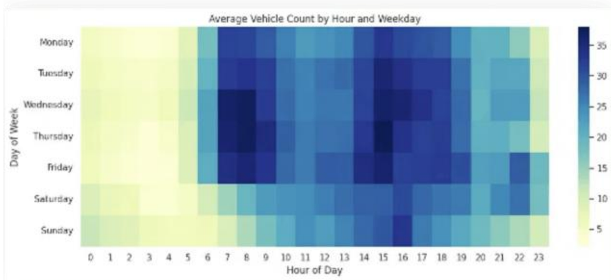
Descriptive Analysis	Visualization Type	Key Finding	Interpretation
Overall Trend	Line Plot (Vehicle Count Over Time)	Clear weekly cycle with dips on weekends and peaks on weekdays. Suddenly, high spikes are present.	Parking demand is primarily driven by Items weekday activities (e.g., work, classes). Spikes strongly suggests that scheduled events significantly increase traffic volume.
Peak UsageExcitation	Heatmap (Hour vs. Weekday)	Busiest times are consistent between 10 AM and 4 PM on weekdays. The peak hour is around 3 PM.	The lot's primary function is to support typical daytime work and academic hours, confirming its role as a facility for employees and visitors rather than overnight or weekend leisure.
Event Impact	Barplot (Event vs. Non-Event Days)	Event days show a much higher average vehicle count than non-event days.	Events are a major factor in parking demand, necessitating specific management strategies these days.
Day Type	Barplot (Weekend vs. Weekday)	Vehicle counts are significantly higher on weekdays.	Maintenance and non-essential activities could be optimally scheduled for weekends when demand is low.

These descriptive insights confirm a strong seasonality and a clear event-driven demand structure, which are critical inputs for the predictive model.

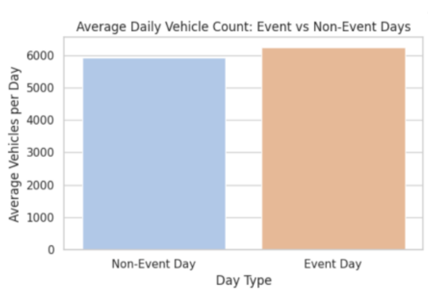
Line Plot (Vehicle Count Over Time)

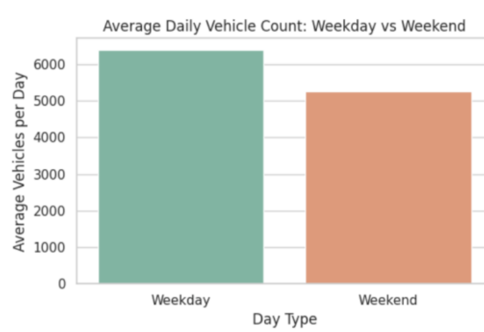


Heatmap (Hour vs. Weekday)



Bar plot (Event vs. Non-Event Days) **Bar plot (Weekend vs. Weekday)**





4. Predictive Analytics

The predictive phase focused on forecasting future vehicle counts to enable proactive resource planning.

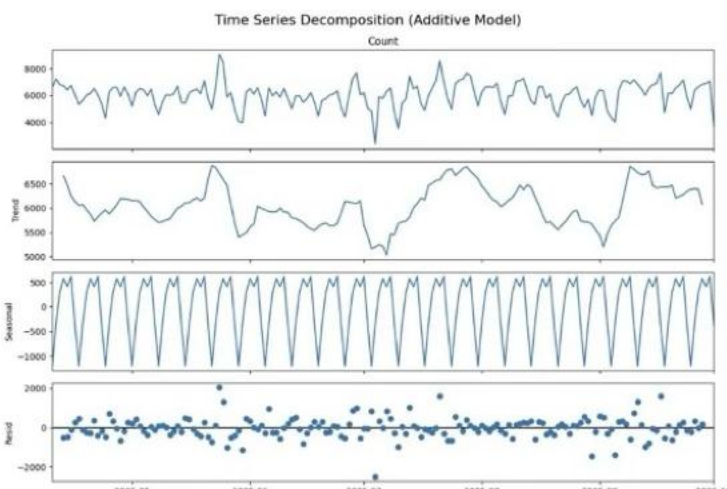
4.1 Methodology: Autoregressive (AR) Model

Given the strong weekly seasonality identified in the descriptive analysis, an Autoregressive (AR) model was selected as the appropriate time-series forecasting technique.

Stationarity: The Augmented Dickey-Fuller (ADF) test indicated that the raw time series was non-stationary. To meet the requirements of the AR model, first-order differencing was applied to the data.

Lag Selection: A lag of 2016 was chosen, corresponding to one full week of 5-minute intervals (24 hours 12 intervals/hour 7 days = 2016). This lag ensures the model captures the dominant weekly seasonality pattern.

Forecast Horizon: The model was used to predict 10,080 future values, representing five weeks of 5-minute interval data.



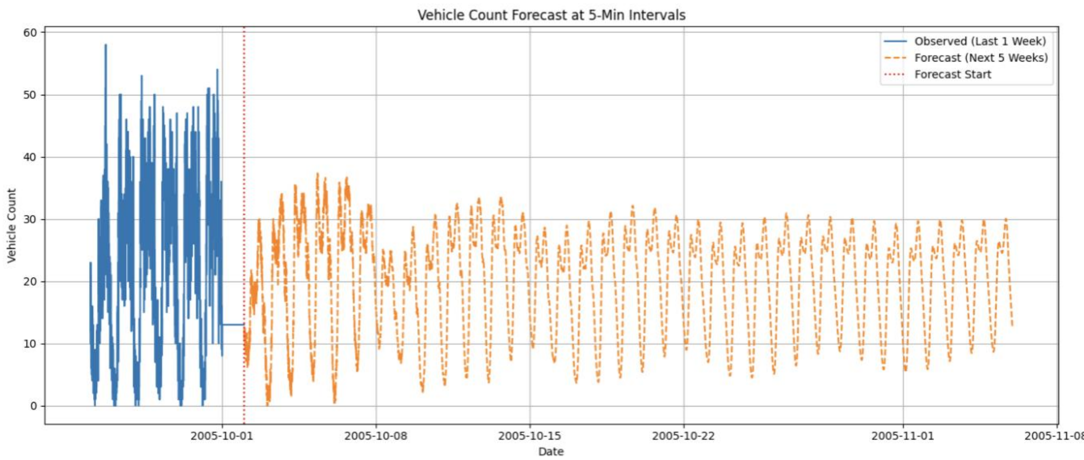
4.2 Forecast Results

The 5-week forecast was aggregated into weekly totals for ease of interpretation:

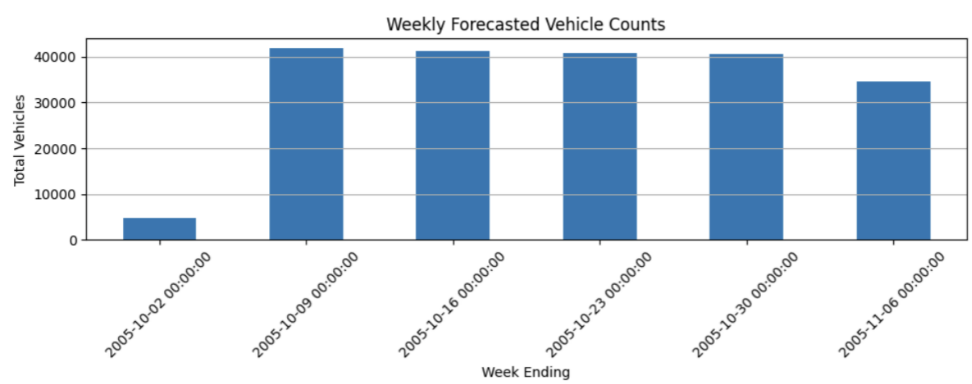
Week Ending	Forecasted Vehicle Count
Oct 02, 2005	4,797
Oct 09, 2005	41,830
Oct 16, 2005	41,305

Oct 23, 2005	40,884
Oct 30, 2005	40,556
Nov 06, 2005	34,546

Forecast of 5-Minute Interval Vehicle Counts




```
Weekly Forecast Summary:
2005-10-02     4797
2005-10-09    41830
2005-10-16    41305
2005-10-23    40884
2005-10-30    40556
2005-11-06    34546
   Freq: W-SUN, dtype: int64
```



Interpretation: The forecast predicts a consistently high demand, averaging over 40,000 vehicles per week, with a slight decreasing trend over the five-week period. The low count for the first week (Oct 02) suggests it may be a partial week or an outlier, while the subsequent weeks establish the baseline demand. This forecast allows management to anticipate future traffic loads and plan staffing, security, and maintenance schedules accordingly.

5. Prescriptive Analytics

The prescriptive phase utilized the insights from the descriptive and predictive models to recommend an optimal operational strategy specifically, the ideal maximum capacity for the parking lot.

5.1 Optimization Goal and Methodology

The objective of the prescriptive model was to determine the lot capacity that minimizes the total cost associated with both overflow and vacancy.

Overflow Cost: Represents lost revenue, customer dissatisfaction, and potential traffic congestion when demand exceeds capacity.

Vacancy Cost: Represents wasted resources and opportunity cost when capacity significantly exceeds demand.

The model tested various capacity scenarios (from 0 to 40 vehicles) and calculated the average overflow and average vacancy for each scenario based on the predicted 5 week; 5-minute vehicle counts. The optimal capacity is the one that results in the smallest absolute difference between the average overflow and the average vacancy.

5.2 Optimal Capacity Recommendation

The initial approximate analysis narrowed the optimal capacity to 20, with a difference of 0.556 between average overflow and average vacancy. A more granular analysis was then performed for capacities between 15 and 25.

Scenario - Capacity (Vehicles)	Avg. Overflow	Avg. Vacancy	Difference (Overflow - Vacancy)
19	4.035	2.479	1.556
20	3.434	2.878	0.556

21	2.858	3.302	-0.444
22	2.306	3.750	-1.444

Recommendation: The analysis identifies the optimal maximum capacity as 21 vehicles. This capacity yields the smallest absolute difference between average overflow and average vacancy (0.444), representing the best balance between meeting demand and avoiding excessive unused space. This recommendation is critical for setting dynamic pricing, controlling entry during peak times, or defining the maximum number of permits to issue.

6. Critical Thinking Suggestions

While the analysis provides a clear and actionable recommendation, a critical review of the methodology reveals several shortcomings and opportunities for future improvement.

6.1 Shortcomings of the Current Model

1. Imputation Method: The use of ffill and bfill for missing values, while preserving integer counts, is a simplistic imputation method. It assumes that the vehicle count remains constant until the next recorded value, which may not accurately reflect real-world fluctuations. A more sophisticated approach, such as seasonal-trend decomposition using LOESS (STL) or a Kalman filter, could provide more accurate imputations.
2. Model Simplicity: The Autoregressive (AR) model is relatively simple. It only uses past values of the vehicle count itself. A more advanced model, such as Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors (SARIMAX), could incorporate the event_active and event_day features directly into the forecast, potentially improving accuracy, especially around event days.
3. Optimization Cost Function: The prescriptive model minimizes the difference between average overflow and average vacancy. This implicitly assumes that the cost of one unit of overflow is equal to the cost of one unit of vacancy. In reality, overflow cost (e.g., traffic fines, customer complaints) is often significantly higher than vacancy cost (e.g., minor opportunity loss).

7. Conclusion

This project successfully demonstrated how a structured analytics framework can be applied to solve a real-world operational challenge parking demand management at the CalIt2 Building. Through a combination of descriptive, predictive, and prescriptive analytics, we extracted actionable insights from raw vehicle count data and formulated strategic recommendations.

Descriptive analytics revealed strong weekly seasonality, significant differences between weekdays and weekends, and noticeable spikes in vehicle counts during scheduled events. These patterns were supported by heatmaps, time series plots, and bar charts.

Predictive analytics involved building an Autoregressive (AR) model with a 2016-lag (equivalent to one week of 5-minute intervals), which forecasted future vehicle traffic for five weeks. The model predicted approximately 40,000 vehicles per week, highlighting the need for robust capacity planning.

Prescriptive analytics used a simulation-based approach to identify the parking lot capacity that minimizes the average difference between overflow and vacancy. The analysis determined that the capacity of 21 vehicles offers the best balance, supporting operational efficiency without underutilizing resources.

8. Recommendations

To enhance the accuracy, adaptability, and real-world usefulness of the parking prescriptive model, the following integrated recommendations are proposed:

1. **Incorporate Weighted Cost Functions**
   Update the prescriptive model to use a refined, weighted cost structure that reflects the higher penalty of parking overflow compared to vacancy. For example:
   Minimize: (3 × Average Overflow) + (1 × Average Vacancy)
   This would push the optimal capacity slightly higher and prioritize customer satisfaction, reduced traffic congestion, and avoidance of lost revenue.
2. **Adopt Dynamic Capacity Management**
   Replace the static optimal capacity (e.g., 21) with a dynamic, time-varying capacity strategy. Using predictive forecasts, capacity recommendations can adjust based on time of day, expected demand peaks (e.g., the 3 PM surge), day of week, or special event activity.
3. **Upgrade Forecasting Methods**
   Improve predictive accuracy by testing advanced forecasting techniques such as SARIMAX, Prophet, or machine-learning models like XGBoost, which can incorporate features such as game schedules, holidays, weather, or promotional events.

4. **Increase Data Granularity**
   Explore collecting data at 1-minute intervals rather than 5-minute intervals to better capture rapid fluctuations in vehicle flow—critical for real-time traffic and capacity management systems.
5. **Improve Imputation Techniques**
   Replace simple forward/backward-fill methods with more robust techniques such as Kalman filtering, STL decomposition, or model-based imputation to better preserve temporal patterns and avoid bias in peak-hour predictions.
6. **Develop a Real-Time Operational Dashboard**
   Implement an interactive dashboard to monitor live vehicle counts, visualize traffic conditions, and generate alerts when occupancy approaches critical thresholds. This would help staff respond proactively and enhance overall operational efficiency.