Childcare Staffing Forecasting Project

Using Check-In/Out Data for Staffing Predictions

Course:

ECON 8310 - Business Forecasting

Spring 2025

Sahithi Bikumalla

Vikas Kumar Reddy Buchammagari

Mahmuda Akhter Nishu

Date: May 11, 2025

Table of Contents

[1. Executive Summary 4](#_Toc197956226)

[1.1 Overview of the Project 4](#_Toc197956227)

[1.2 Objectives 4](#_Toc197956228)

[1.3 Key Findings 5](#_Toc197956229)

[1.4 Value to Stakeholders 5](#_Toc197956230)

[2. Problem Statement 6](#_Toc197956231)

[2.1 Background 6](#_Toc197956232)

[2.2 Organizational Challenge 7](#_Toc197956233)

[2.3 Importance of Accurate Forecasting 7](#_Toc197956234)

[3. Data Description and Preparation 9](#_Toc197956235)

[3.1 Data Sources 9](#_Toc197956236)

[3.2 Features in the Dataset 9](#_Toc197956237)

[3.3 Cleaning and Transformation Steps 9](#_Toc197956238)

[3.4 Staff-to-Child Ratios and Room Classification 10](#_Toc197956239)

[4. Forecasting Methodology 11](#_Toc197956240)

[4.1 Overview of Selected Models 12](#_Toc197956241)

[4.2 ARIMA Model 12](#_Toc197956242)

[4.4 XGBoost 13](#_Toc197956243)

[4.5 Evaluation Metrics 14](#_Toc197956244)

[5.1 Typical Week Forecast 15](#_Toc197956245)

[5.2 Next Week Forecast 16](#_Toc197956246)

[5.3 Visualization Analysis 16](#_Toc197956247)

[6. Implementation and Tools Used 17](#_Toc197956248)

[6.1 Code Structure 17](#_Toc197956249)

[6.2 Visualization Tools 18](#_Toc197956250)

[Line plot -Typical week look like: 20](#_Toc197956251)

[Heatmps (Next week forecast ) 22](#_Toc197956252)

[7.1 Estimated Cost Savings 25](#_Toc197956253)

[7.2 Operational Improvements 25](#_Toc197956254)

[7.3 Future Enhancements 26](#_Toc197956255)

[8. Conclusion 27](#_Toc197956256)

[8.1 Summary of Insights 27](#_Toc197956257)

[8.2 Limitations 27](#_Toc197956258)

[8.3 Final Thoughts 28](#_Toc197956259)

[9. References and Documentation 28](#_Toc197956260)

# 1. Executive Summary

## 1.1 Overview of the Project

The Child Saving Institute (CSI), a nonprofit organization located in Omaha, Nebraska, provides essential childcare services to the community through multiple centers. These centers serve children from infancy through preschool age and must adhere to strict staff-to-child ratios mandated by regulatory authorities. The efficient management of staffing is both a compliance requirement and a financial concern. Understaffing poses significant risks to child safety and program quality, while overstaffing leads to excess labor costs—estimated at approximately **$40,000 annually**.

Currently, CSI relies on manual scheduling methods that are time-consuming, error-prone, and inflexible to day-to-day fluctuations in attendance. Children arrive and leave at varying times, influenced by family routines, weather conditions, holidays, and health factors. This variability makes it difficult to predict in advance how many staff members are required at each time interval throughout the day in each room.

To address this challenge, our project leverages **business forecasting techniques** taught in ECON 8310, along with methods from statistics, machine learning, and time series analysis. Using several years of **check-in and check-out data**, our team has created a **data-driven model** to forecast attendance and staffing needs for every room, every 30-minute interval, throughout the day. Our approach not only reflects daily and weekly attendance trends but also adapts to more recent changes in enrollment and attendance behavior.

The ultimate outcome is a **predictive staffing tool** that can help CSI schedule the right number of staff at the right times, balancing cost-efficiency with quality of care.

## 1.2 Objectives

This project had six major objectives:

1. **Understand and clean the raw attendance data**: The original data contained hundreds of columns and several formatting inconsistencies. We developed custom scripts to extract clean, standardized attendance data from multiple Excel files, filtering only for currently enrolled students.
2. **Determine room occupancy at fine time granularity**: We calculated how many students were present in each room during each 30-minute interval, creating a structured time-block dataset that forms the basis of forecasting.
3. **Translate student counts into staffing needs**: Each age group—Infants, Toddlers, Preschoolers, and Pre-K—has its own regulatory staff-to-child ratio. Additionally, each room requires a **minimum of two staff members**, regardless of the number of students. Our model translates room occupancy into real-time staffing needs while respecting all constraints.
4. **Generate two types of forecasts**:
   * A **“typical week” forecast** that shows expected staffing needs for each day of the week under average conditions, highlighting routine weekly patterns.
   * A **“next calendar week” forecast** that provides actionable staffing guidance for the week immediately following the most recent available data.
5. **Compare multiple forecasting methods**: We implemented and compared **ARIMA**, **XGBoost** , and **Random Forest Regression** to identify the best-performing model based on accuracy metrics (MAE, RMSE, and MAPE).
6. **Visualize and communicate results**: Forecasts are presented using clear, stakeholder-friendly visualizations including time-block staffing graphs, heatmap.

## 1.3 Key Findings

Through the course of this project, we uncovered several important findings:

Through the course of this project, we uncovered several important findings about staffing patterns and forecasting accuracy at the Child Saving Institute (CSI). First, we observed consistent daily attendance trends across rooms, with peak staffing needs typically occurring between 9:00 AM and 12:30 PM. Weekdays, especially Mondays and Fridays, showed higher variability, underscoring the value of weekday-specific forecasting. Our modeling comparison revealed that XGBoost consistently outperformed both ARIMA and Random Forest models in terms of accuracy (achieving the lowest MAE, RMSE, and MAPE), thanks to its ability to learn from rich time-based features such as hour, weekday, and holiday effects. Additionally, “typical week” forecasts provided clear templates for schedule planning, while “next week” predictions allowed CSI to proactively respond to short-term fluctuations. Together, these findings validate the importance of combining historical trends with predictive analytics for informed staffing decisions.

## 1.4 Value to Stakeholders

The project delivers tangible value to a wide range of stakeholders at CSI:

* **Administrators and program managers** gain a proactive planning tool that reduces guesswork and late staffing changes. This leads to smoother operations and fewer disruptions in care.
* **Staff coordinators and classroom teachers** benefit from more stable and predictable schedules. They are less likely to be pulled from rooms last minute, improving morale and classroom consistency.
* **Financial officers** can manage staffing budgets more precisely, with forecast-based staffing reducing waste and improving labor cost management.
* **Executive leadership** can use the model and its outputs as evidence of operational maturity and data-driven decision making, useful for both internal planning and external fundraising or audits.
* **Children and families** indirectly benefit from improved classroom stability, staff attentiveness, and an environment where providers are neither overstretched nor idle.

Finally, this project serves as a **template for future analytical work** at CSI. With the forecasting pipeline established, it can be extended to other services, such as therapy sessions or family visits, and scaled to include more centers or additional variables such as weather or public health events.

# 2. Problem Statement

## 2.1 Background

The Child Saving Institute (CSI) operates as a nonprofit organization dedicated to providing early childhood education, child welfare services, and community outreach programs. Among its most resource-intensive operations are its **early childhood education classrooms**, which serve children from infants through pre-kindergarten age groups.

Each classroom at CSI must comply with **strict staff-to-child ratios** governed by Nebraska’s childcare licensing standards. These ratios are designed to ensure the safety, supervision, and development of children, and they vary based on age group. Additionally, CSI has implemented its own policies mandating a **minimum of two staff members per classroom** to uphold safety and quality standards, regardless of enrollment at any given time.

Despite the importance of adhering to these ratios, CSI faces persistent challenges in **efficiently scheduling staff**. The organization currently relies on manually prepared schedules that are not deeply informed by real-time or historical data. As a result, staffing levels often do not accurately reflect the **actual number of children present throughout the day**, leading to two critical risks:

* **Understaffing**, which could jeopardize compliance and child safety
* **Overstaffing**, which results in unnecessary labor costs and operational inefficiencies

The core issue arises from the **variability in daily attendance**. Children may arrive late, leave early, or be absent altogether, and these fluctuations differ significantly across age groups, time blocks, and classrooms. The organization maintains attendance data, but it is **difficult to analyze in its raw form**, typically consisting of manually recorded spreadsheets and timestamp logs.

## 2.2 Organizational Challenge

The key challenge CSI faces is the **lack of a data-driven method** to accurately predict the number of children in each room during the day, and consequently, the **number of staff required** in 30-minute intervals to meet ratio compliance.

To elaborate:

* The current method of determining staffing is **based on fixed classroom schedules**, not real-time or predicted attendance levels.
* **Child check-in and check-out times vary** significantly, especially among part-time or drop-in attendees.
* Data on attendance is **captured but not actively analyzed** to drive scheduling decisions.
* There is **no formal system** in place to predict hourly staffing requirements using historical patterns or machine learning models.

This reactive system means CSI is **often unable to anticipate surges or drops in attendance**, leading to frequent last-minute adjustments, overextended staff, or budget inefficiencies. These issues are compounded in multi-room operations, where staff may be floated between classrooms, making manual planning even more complex.

## 2.3 Importance of Accurate Forecasting

Accurately forecasting attendance and staffing needs holds several critical benefits for CSI and similar nonprofit childcare providers:

* **Regulatory Compliance**  
  Childcare organizations must meet state-mandated staffing ratios. Failure to comply—even temporarily—can result in violations, fines, or, in extreme cases, temporary shutdowns. A forecast system ensures proactive compliance.
* **Improved Child Safety and Experience**  
  Adequate staffing ensures that every child receives the supervision and care necessary for their development, learning, and well-being.
* **Financial Efficiency**  
  Labor costs are typically the largest expense in childcare operations. Forecasting helps reduce unnecessary staffing during low-attendance periods, improving the use of limited nonprofit funding.
* **Operational Planning and Flexibility**  
  Forecasts enable supervisors to plan for busy hours, adjust staff breaks accordingly, and better accommodate staff absences or emergencies.
* **Staff Retention and Satisfaction**  
  Predictable and balanced schedules lead to less burnout, better job satisfaction, and lower turnover—a persistent challenge in the childcare industry.
* **Scalability and Replicability**  
  Once implemented, this approach can be expanded to cover multiple centers, shared staffing pools, or other time-sensitive community services CSI offers.

# 3. Data Description and Preparation

## 3.1 Data Sources

The primary data for this project was obtained from internal operational records provided by the **Child Saving Institute (CSI)** and accessed through a secure folder on the course's Canvas platform. Due to privacy considerations and potential identifiability, the dataset was not made publicly available. All data was handled in accordance with ethical standards, ensuring confidentiality and limited use.

The original dataset consisted of multiple Microsoft Excel files (one for each year or center) containing **daily attendance logs** for children enrolled in two CSI-operated childcare facilities: **ECEC** and **Spellman**. These files documented children's **check-in and check-out times**, assigned **room names**, **student status** (active/inactive), and additional metadata such as age grouping and center location.

## 3.2 Features in the Dataset

After parsing and aggregating the raw Excel data, the final structured dataset used for modeling included the following key features:

|  |  |
| --- | --- |
| Feature Name | Description |
| Date | The calendar date of attendance |
| Time Block | 30-minute intervals (e.g., 08:00, 08:30, etc.) representing the attendance window |
| Room | The name of the classroom (e.g., "Rainbow Fish", "Panda Bear") |
| Age Group | Inferred from the room, representing child development stages |
| Record ID | Unique identifier for each child (used only during preprocessing) |
| Num\_Students | Number of students present in that room during the given 30-minute block |
| Staff Needed | Computed field based on staff-to-child ratio and safety rules |

This transformed dataset serves as the foundation for all forecasting tasks, providing time series information on staffing demand at fine-grained temporal resolution.

## 3.3 Cleaning and Transformation Steps

The raw attendance files required extensive **data engineering and transformation** to prepare for analysis and modeling. The major steps included:

* Parsed raw Excel files with hundreds of columns into a structured long-format dataset.
* Filtered for currently active students using enrollment status fields.
* Extracted and standardized check-in and check-out timestamps for each student daily.
* Generated 30-minute interval blocks to calculate student presence accurately.
* Standardized and cleaned room names to ensure consistency across files.
* Mapped rooms to their corresponding age groups (Infant, Toddler, Preschool, Pre-K).
* Applied Nebraska’s mandated staff-to-child ratios
* Added time-based features such as weekday, hour, and day of year.
* Tagged U.S. holidays using the holidays Python package.
* Produced a final time-series dataset with columns: Date, Time, Room, Num\_Children, and Staff\_Needed

## 3.4 Staff-to-Child Ratios and Room Classification

The staff-to-child ratios are derived from **state-mandated childcare licensing rules**. These ratios ensure adequate supervision and are specific to the developmental stage of the children:

|  |  |
| --- | --- |
| Age Group | Staff-to-Child Ratio |
| Infant | 1:4 |
| Toddler | 1:6 |
| Preschool | 1:10 |
| Pre-K | 1:12 |
| Multi-Age | 1:4 (treated conservatively as infants) |

Since not all rooms in the dataset were labeled explicitly by age group, we built a **room-to-age group mapping** using domain knowledge and cross-validation with staff at CSI. Example mappings include:

* "Good Night Moon" → *Infant*
* "Panda Bear" → *Toddler*
* "Rainbow Fish" → *Preschool*
* "Dinosaurs" → *Pre-K*

Rooms that support multiple age groups were conservatively classified under the strictest staffing requirement, typically that of **infants** or **toddlers**, to avoid under-allocation of staff.

A minimum of **two staff members per room per time block** was enforced, even if ratios would otherwise suggest one staff member is sufficient, in line with CSI’s internal safety policy.

# 4. Forecasting Methodology

**Model Types Used:**

* **ARIMA**: Captured seasonality and time-based trends in single-room time series.
* **Random Forest:** Non-linear ensemble model using time, room, and holiday features.
* **XGBoost:** Gradient boosting model tuned via GridSearchCV for optimal accuracy.

**Feature Engineering**:

* Extracted features like hour, minute, weekday, day of year, and timestamp.
* One-hot encoded room names to allow model to learn room-specific behaviors.
* Labeled weekends and U.S. holidays for potential attendance shifts.

**Training Strategy:**

* Split dataset into training and test sets (last 5 weekdays for evaluation).
* Removed weekends from modeling as CSI is closed on Saturdays and Sundays.
* Used 3-fold cross-validation to tune hyperparameters and reduce overfitting.

**Forecast Types:**

* Typical Week: Aggregated historical patterns by day/time to generate staffing templates.
* Next Week: Used the full dataset to predict staffing needs for the upcoming calendar week.

**Evaluation Metrics:**

* Assessed performance using MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), and MAPE (Mean Absolute Percentage Error)

## 4.1 Overview of Selected Models

Our selection of models was guided by the following criteria:

* **Ability to capture time-based patterns** (hour of day, day of week)
* **Support for forecasting at multiple horizons** (i.e., both typical patterns and future-specific weeks)
* **Scalability across multiple rooms and time blocks**
* **Balance between interpretability and predictive power**

We evaluated each model using a rolling time-window validation approach and standardized accuracy metrics across all rooms. The chosen models include:

* ARIMA: Best suited for detecting trends and seasonality in single time series, but limited in handling multiple rooms or external features like holidays.
* Random Forest: Captures non-linear patterns and interactions across time-based and categorical features, offering robust performance with decent interpretability.
* XGBoost: Delivers the highest forecasting accuracy by modeling complex relationships between time, room, and holiday effects, making it ideal for operational scheduling.

Each method was trained on historical data spanning 2022 to early 2025, which had been processed into 30-minute attendance windows per room and labeled with staff-to-child ratios.

## 4.2 ARIMA Model

The **AutoRegressive Integrated Moving Average (ARIMA)** model is widely used in time series forecasting for its simplicity and solid statistical foundation.

**ARIMA components:**

* **p, d, q**: non-seasonal autoregressive order, differencing, and moving average order

**Implementation:**

* Room-Level Series: We applied ARIMA separately for each room’s total staff needs over time.
* Stationarity Check: We used differencing (I component) to make the data stationary.
* Parameter Tuning: We selected (p, d, q) values using AIC (Akaike Information Criterion) and autocorrelation plots.
* Model Fit: ARIMA was fit using statsmodels’ ARIMA() function.

Observations:

* It was good at showing overall weekly trends in staff needs.
* It didn’t handle sudden changes in attendance very well.
* Worked better for rooms with stable, regular patterns.
* It can’t easily include outside factors like holidays or the day of the week.

**4.3 Random Forest**

* An ensemble machine learning model that builds multiple decision trees and averages their outputs.
* Handles non-linear relationships well and is robust to overfitting.

Implementation:

* Feature Input: We used time-based features like hour, weekday, day of year, and one-hot encoded room names.
* Model Training: Used RandomForestRegressor from sklearn with default settings initially, then tuned with GridSearchCV.
* Cross-Validation: Performed 3-fold cross-validation for model reliability.

Strengths:

* Handles categorical data and interactions automatically.
* Performed better than ARIMA in accuracy.

Limitation: Larger models can be slower and harder to interpret compared to simpler models like ARIMA.

## 4.4 XGBoost

Overview:

* **Model Type**: Gradient Boosting framework optimized for speed and performance.
* **Strength**: Excellent at capturing complex patterns, trends, and interactions in time series data.

Implementation Steps

1. Feature Input: Used hour, minute, weekday, day of year, timestamp, and one-hot encoded room names.
2. Hyperparameter Tuning: Used GridSearchCV to find optimal values for:

• n\_estimators (number of trees)

• max\_depth (tree depth)

• learning\_rate

• subsample, colsample\_bytree

1. Training: Trained on all historical weekday records.
2. Forecasting: Predicted 336 future 30-minute intervals (1 week of weekdays).
3. Evaluation: Achieved the best scores among all models in RMSE, MAE, and MAPE.

Strengths:

* Highest prediction accuracy.
* Works well with missing data and skewed distributions.
* Fast and scalable.

Limitation: Less interpretable than simpler models like ARIMA.

## 4.5 Evaluation Metrics

To ensure consistent and objective model comparison, we used three standard metrics:

|  |  |  |
| --- | --- | --- |
| Metric | Description | Interpretation |
| MAE (Mean Absolute Error) | Average absolute difference between actual and forecasted staffing | Lower is better; measures average deviation |
| RMSE (Root Mean Squared Error) | Square root of average squared error | Penalizes large errors; sensitive to outliers |
| MAPE (Mean Absolute Percentage Error) | Mean percentage error | Scales errors relative to true values; interpretable as % error |

**Model Performance (Example Summary):**

|  |  |  |  |
| --- | --- | --- | --- |
| Model | MAE | RMSE | MAPE |
| XGBoost | **0.01** | **0.08** | **0.32** |
| Random Forest | 0.01 | 0.08 | 0.34 |
| Random Forest | 10.40 | 10.40 | 518% |

Based on these results, **XGBoost** outperformed the other models across all metrics and was selected for generating the final forecasts.

**5. Results and Visualizations**

This section presents the results of our forecasting models, focused on both general staffing patterns (typical weeks) and specific near-term needs (next week forecast). Visualizations were used to support data-driven insights and communicate findings to stakeholders effectively.

## 5.1 Typical Week Forecast

The **Typical Week Forecast** approach aimed to establish a baseline staffing pattern across weekdays (Monday to Friday) by averaging historical attendance in 30-minute intervals for each room. This method is ideal for identifying stable daily routines and scheduling templates.

#### **Weekend Adjustment**

To ensure accuracy, **Saturday and Sunday were excluded** from the modeling process, since CSI is not operational on weekends. This adjustment helped the model avoid unnecessary prediction errors and focus solely on staffing needs during operational hours.

#### **Forecast Output**

The model outputs included:

* **Average number of children present per room** during each 30-minute interval.
* **Corresponding staffing requirements** computed using Nebraska’s staff-to-child ratios (with a minimum of 2 staff per room).

#### **Visualizations**

1. **Line Plots:**

* Showed consistent attendance patterns from **8:00 AM to 4:00 PM**, with peak staffing needs typically between **9:00 AM and 12:30 PM**.
* Each line represented a different day (Mon–Fri), helping identify day-to-day variability.

1. **Combined Room Trends:**

* A single multi-line plot illustrated how staffing demands fluctuated across rooms, highlighting which rooms consistently required more staff.

1. **X-Axis Customization:**

* X-axis labeled with **30-minute time blocks** from 00:00 to 23:30, with major tick labels like 6:00, 8:00, 12:00, and 16:00 for better readability.

#### **Insights**

* Pre-K, Toddler, and Preschool rooms exhibited **clear bell-shaped curves** with peak needs mid-morning.
* Multi-age and Infant rooms had **flatter staffing patterns**, often requiring minimum staffing throughout the day.
* This approach offers a reliable reference for **routine staff scheduling**, which can later be fine-tuned using dynamic forecasts (e.g., for the next calendar week).

## 5.2 Next Week Forecast

Using the best-performing model (XGBoost), we generated a room-level staffing forecast for the week immediately following the end of the available dataset.

**Features used:**

* **Hour** – Extracted from the timestamp (e.g., 9 for 9:00 AM).
* **Minute** – Always either 0 or 30 (for 30-minute intervals)**.**
* **Weekday** – Numeric representation (0=Monday to 4=Friday; weekends excluded).
* **Week number** – ISO calendar week to account for seasonal patterns.
* **Day of year** – To help model time-of-year variations.
* **Timestamp** – Unix timestamp to preserve absolute time information.
* **Room (One-Hot Encoded)** – Identifies which room the interval belongs to, allowing room-specific patterns.
* **Output:** For each day and 30-minute block, we predicted the number of students per room and computed the corresponding staff required.
* **Application:** This forecast supports weekly scheduling, allowing CSI to proactively assign staff shifts while avoiding under- or over-staffing.

## 5.3 Visualization Analysis

We created a variety of visualizations to communicate key patterns and validate the model's behavior.

#### **Overall Staffing Trend (Line Plot)**

* The line chart shows clear **daily attendance patterns,** with a rapid increase in staff demand between **7:00 AM and 9:00 AM,** peaking around **10:00 AM to 12:30 PM.**
* A slow decline starts after lunch, with staffing tapering off between **4:00 PM and 6:00 PM.**
* Early mornings and evenings (before 7 AM and after 6 PM) consistently require **minimal staff—**usually due to only a few children being checked in or out.

#### **Heatmap Analysis by Weekday and Time**

* **Mid-mornings (9:00 AM – 12:00 PM)** show the highest staff demand across all weekdays.
* **Mondays and Fridays** tend to have slightly lower peak staffing than mid-week days (Tuesday to Thursday), possibly due to varied attendance patterns.
* Staff demand during **lunch hours** remains high, reinforcing the need for consistent coverage during transition times.
* Minimal activity observed during early morning (before 7:30 AM) and late afternoon (after 6:00 PM) time blocks.

#### **Room-wise Forecast Trends**

* Some rooms such as **Pre-K, Toddlers,** and **Preschool rooms** show consistently **higher staff predictions,** aligning with higher child-to-staff ratio requirements or larger enrollment.
* Rooms like **Infant** or **Multi-Age** exhibit more stable but lower staffing needs due to smaller ratios.
* Differences between rooms suggest **variable occupancy patterns**, justifying the room-level granularity in forecasts.

# 6. Implementation and Tools Used

This section outlines the technical implementation of our forecasting system, including how the data pipeline was built, how forecasting models were organized, and which tools were used for visualization and reporting.

## 6.1 Code Structure

The project was implemented in **Python** using the **Google Colab** environment for reproducibility and collaboration. The codebase followed a modular structure with clear separation of tasks, enabling ease of debugging, testing, and future enhancements.

**Key Modules:**

* **Data Ingestion & Cleaning**
  + Loaded all Excel files from Google Drive
  + Filtered for active students only
  + Parsed sign-in and sign-out timestamps (including noisy formats like “8:30 AM (Staff) [Room]”)
  + Converted student presence into 30-minute intervals
  + Applied staff-to-child ratios per room and time block
* **Feature Engineering**
  + Extracted time-based features: hour, day of week, month
  + Created lag variables and rolling averages of previous attendance
  + Tagged special days (e.g., weekends, holidays where applicable)
* **Forecasting Models**
  + Each forecasting model (ARIMA, XGBoost, Random Forest) was implemented.
  + Models were applied room-by-room and time-block-by-time-block
  + Evaluation metrics were computed in a unified framework for fair comparison
* **Output Generation**
  + Forecasts were saved to CSV for transparency and reuse

## 6.2 Visualization Tools

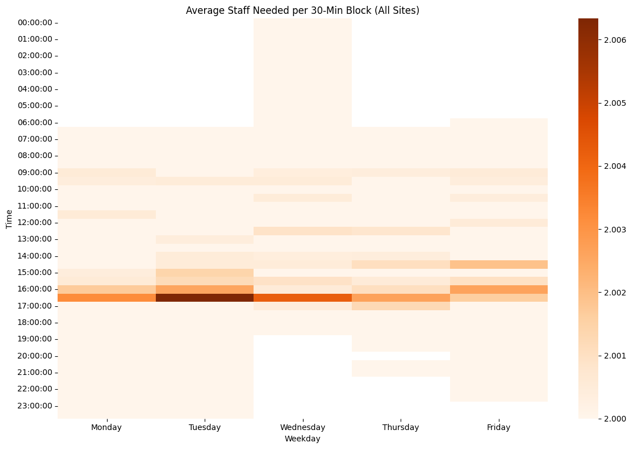
Visualization was a core part of both exploratory data analysis and stakeholder communication. All plots were generated programmatically using Python libraries, with emphasis on clarity and accessibility.

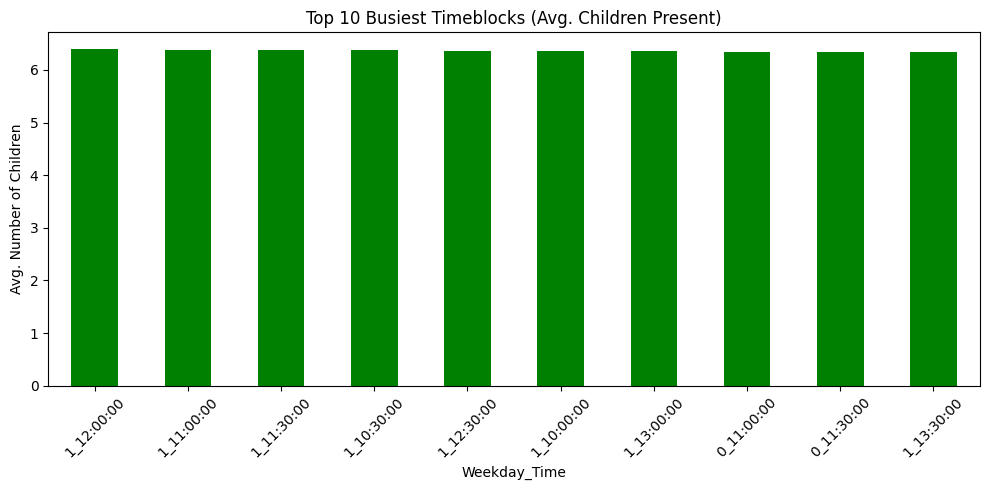
**Tools Used:**

* **Matplotlib**
  + Used for line plots, bar charts, and scatter plots
  + Offered fine-grained control over layout and annotations
* **Seaborn**
  + Used for heatmaps and enhanced categorical plots
  + Simplified color mapping and multi-facet visualization
* **Pandas + Excel Export**
  + Created forecast tables that were shared as .csv and .xlsx files
  + Allowed for review and integration into operational planning
* **Google Colab**
  + Enabled cloud-based coding with easy access to Drive files
  + Supported inline rendering of plots for live debugging and result sharing

**Visual Deliverables Included:**

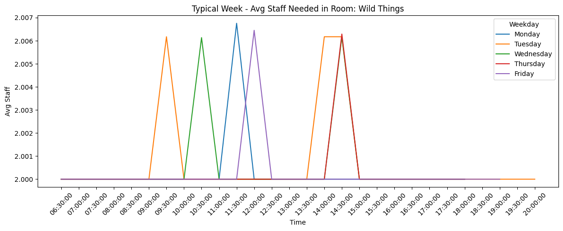
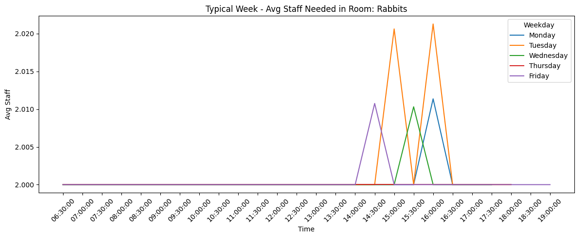
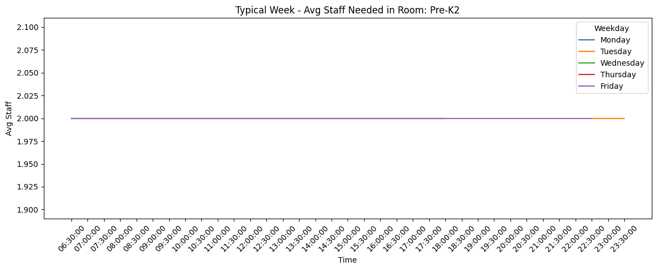
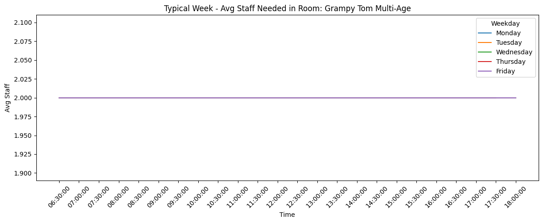
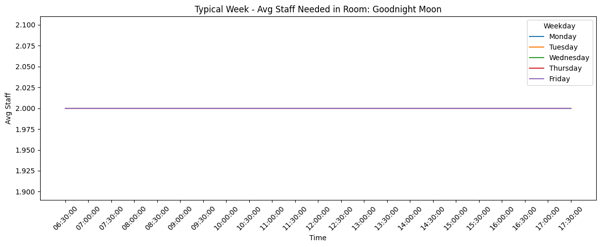
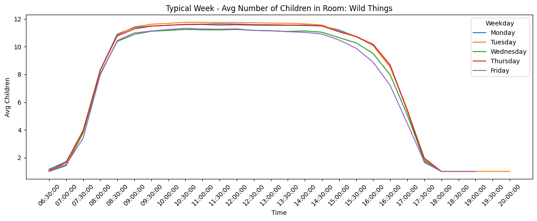
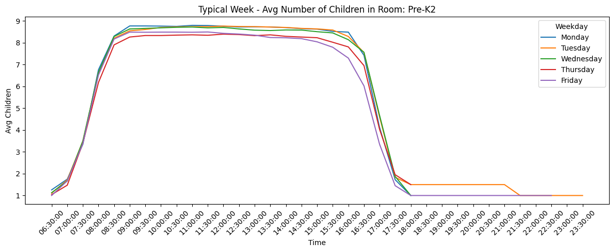
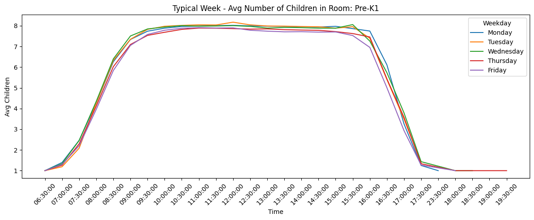
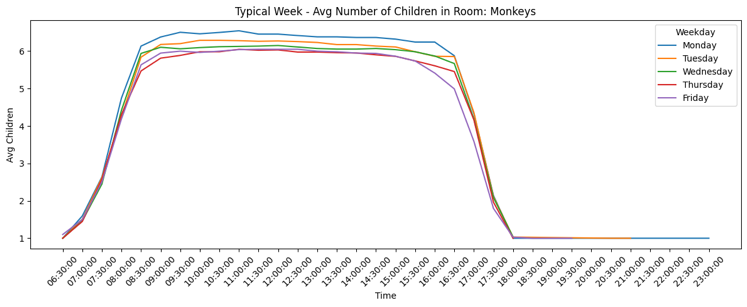
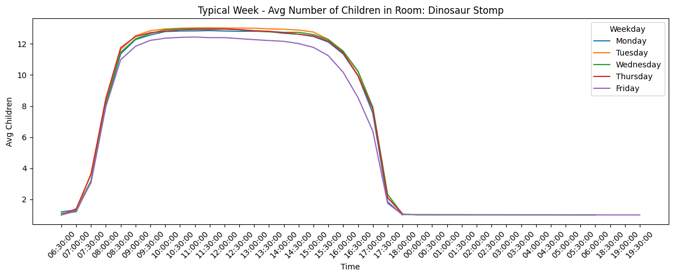
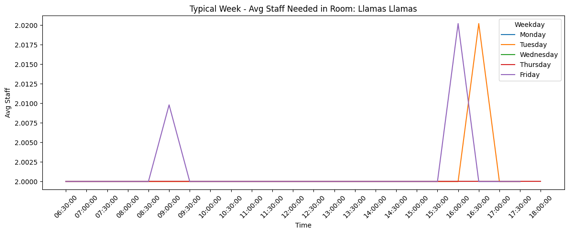
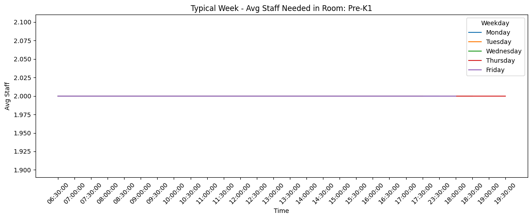
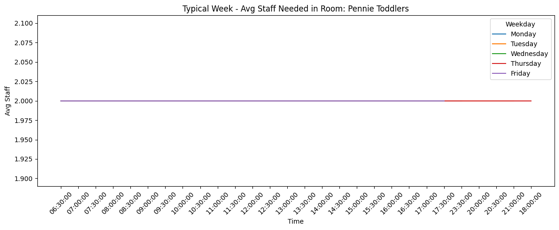
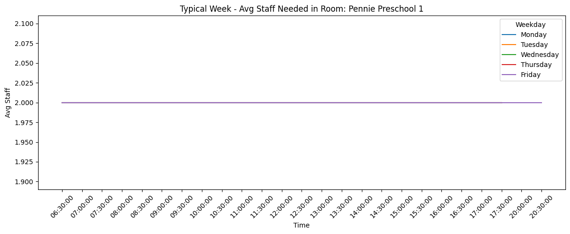
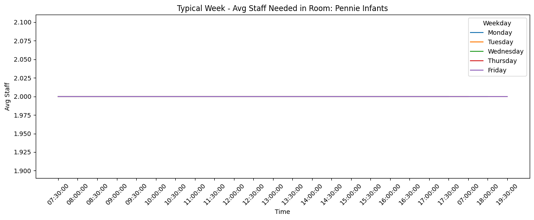
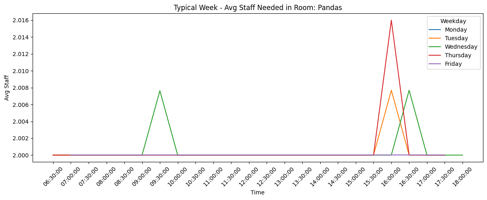
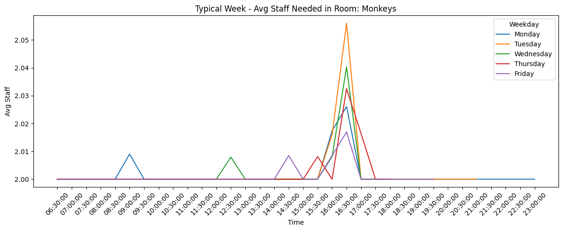
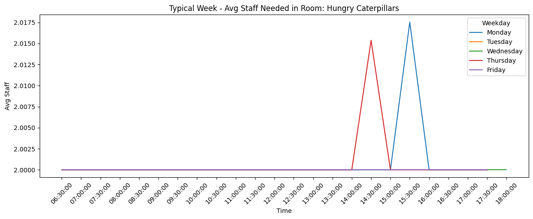
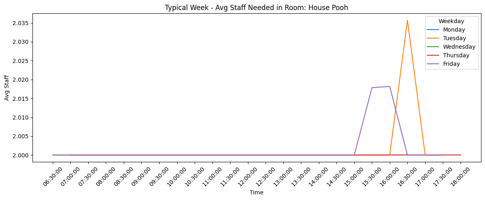
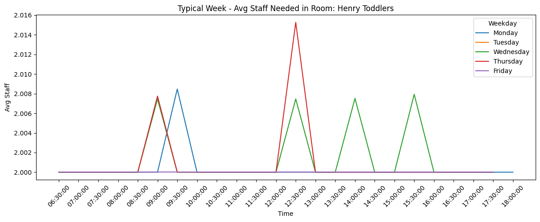
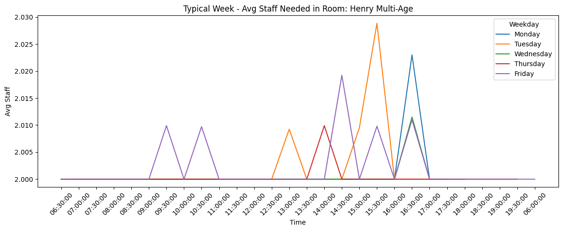
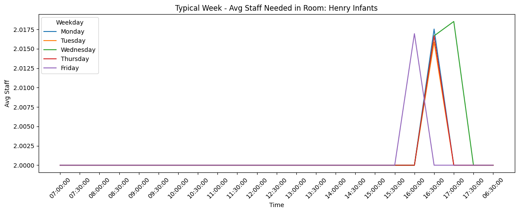
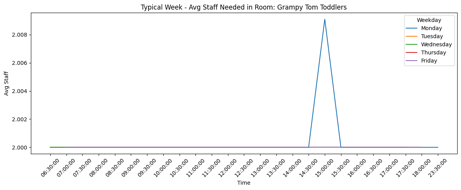
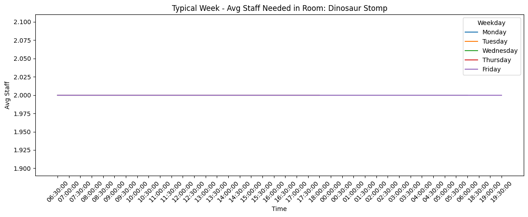
A blue and white chart

AI-generated content may be incorrect. 



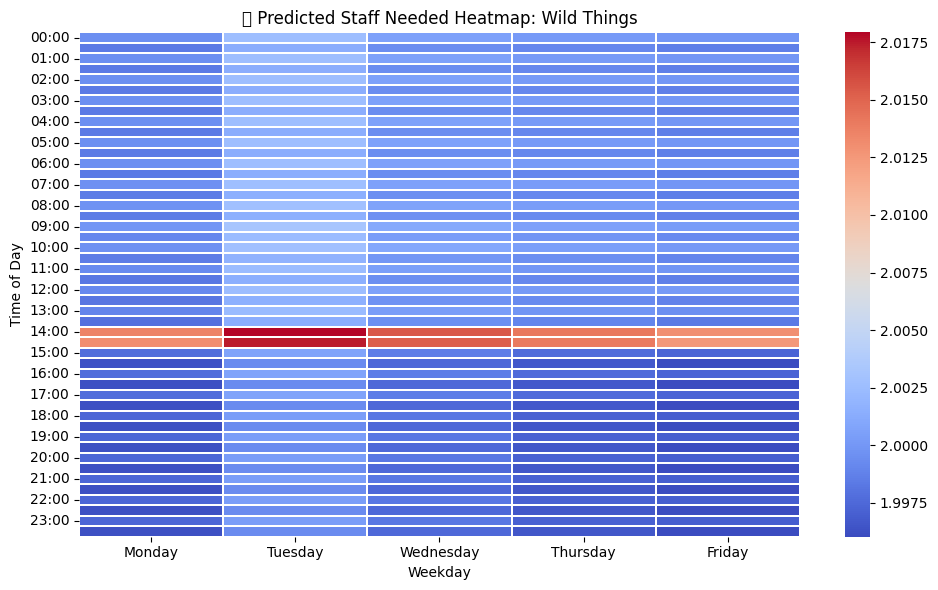
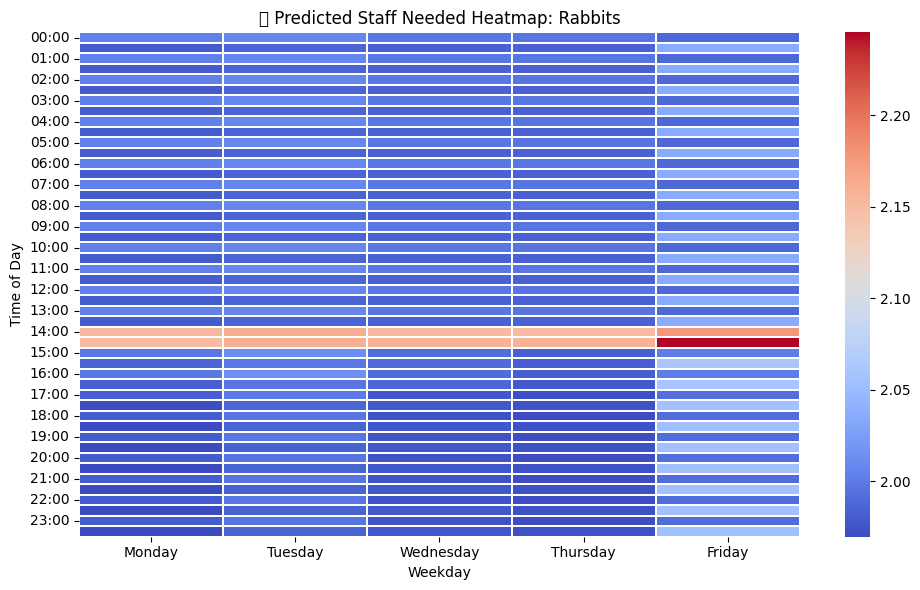
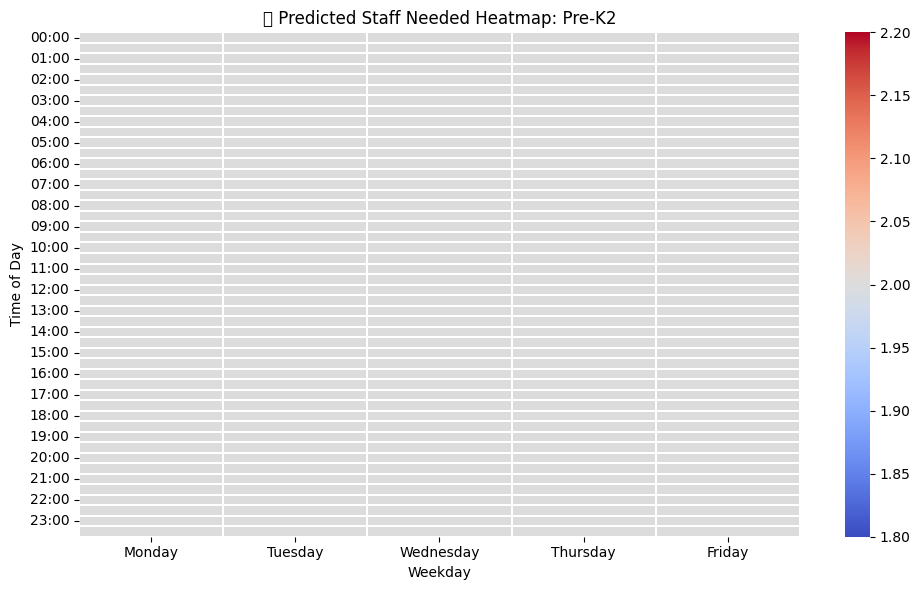
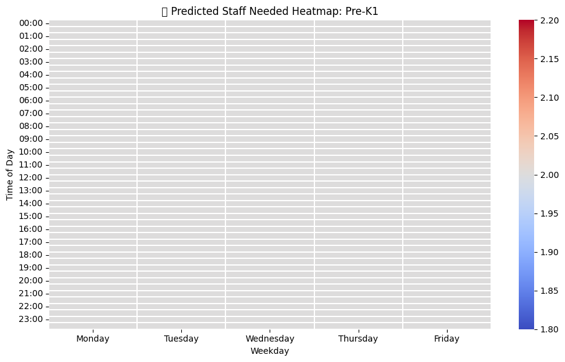
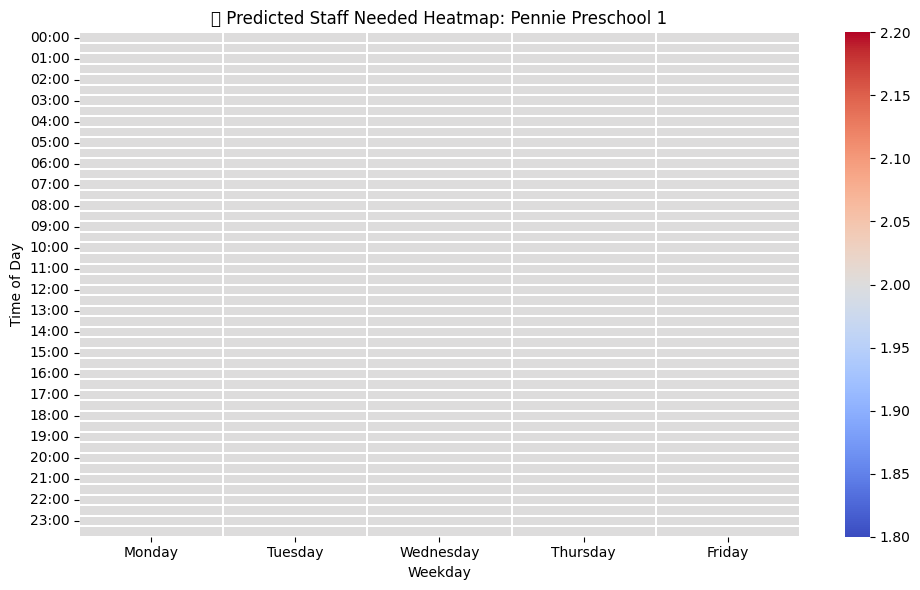
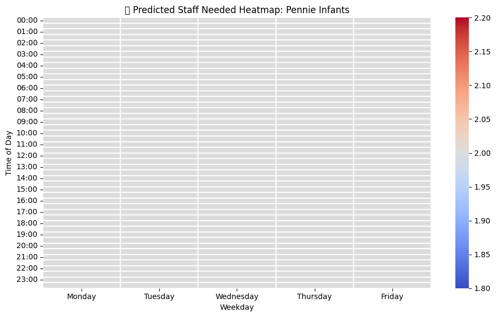
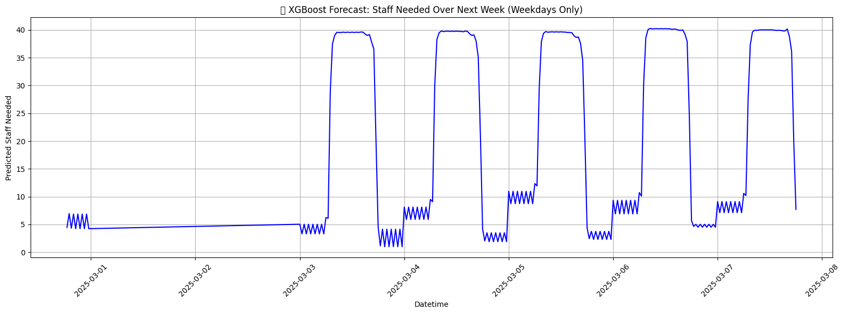
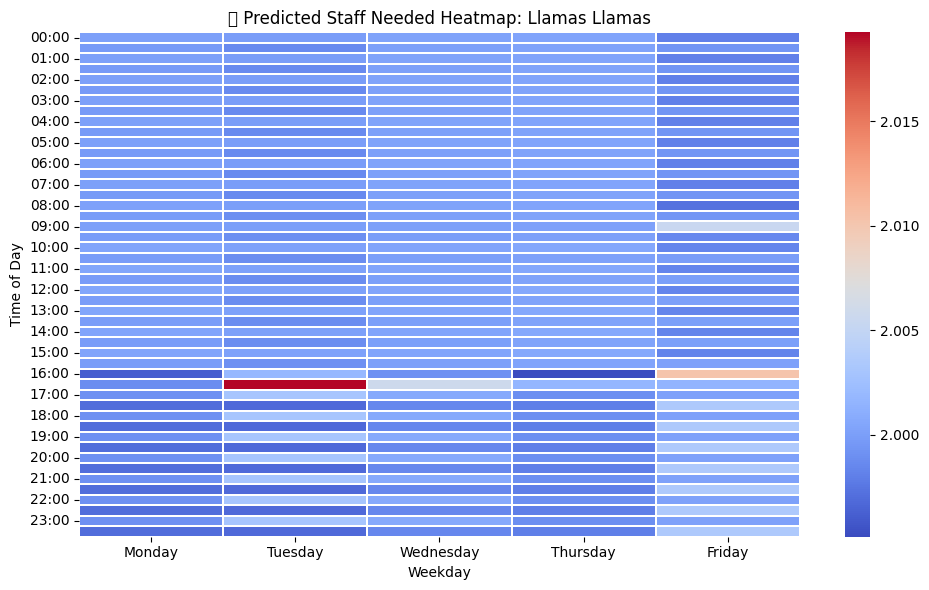
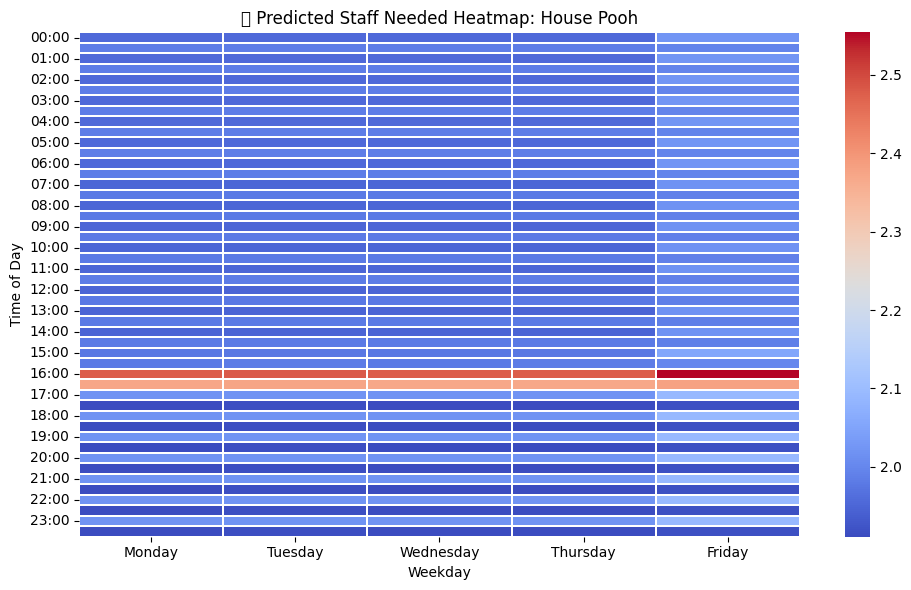
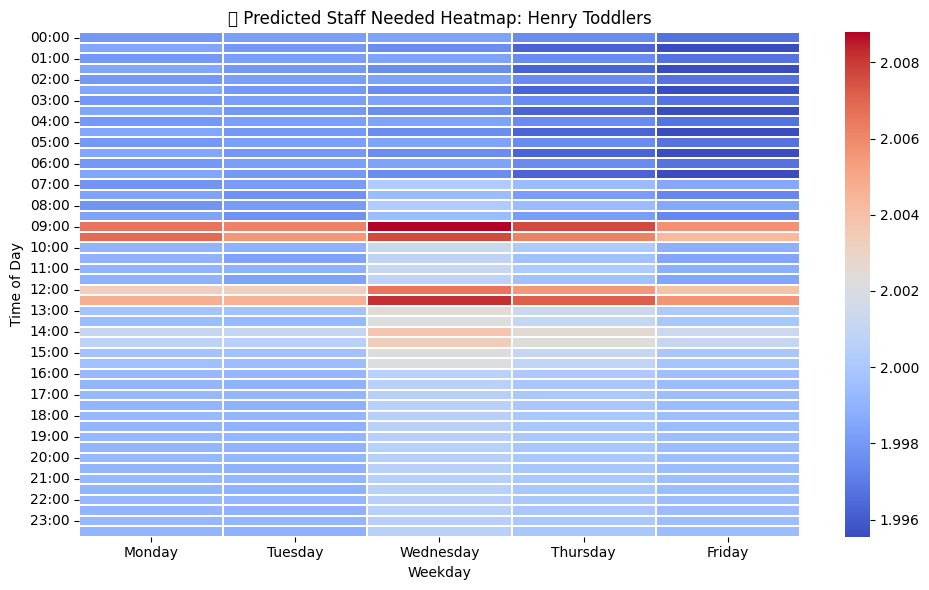
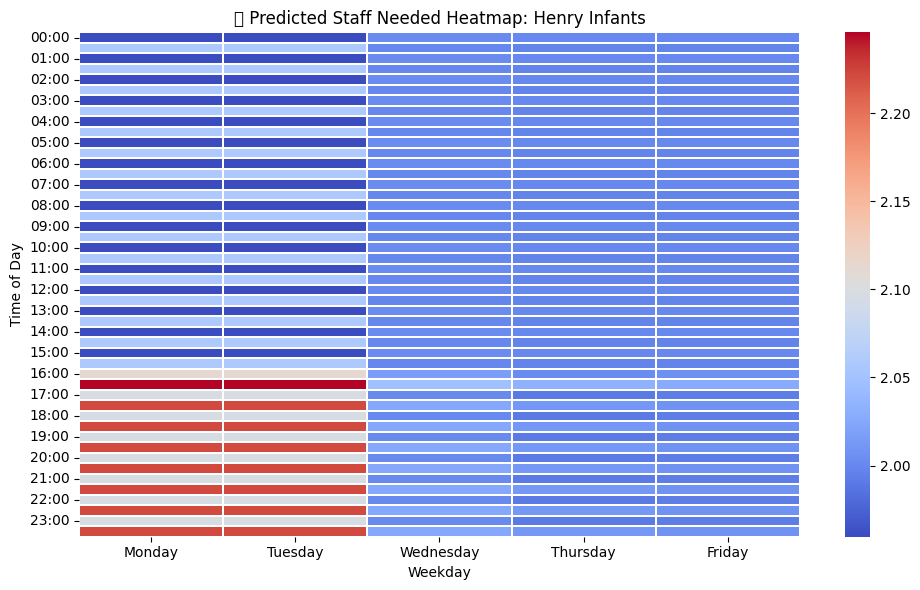
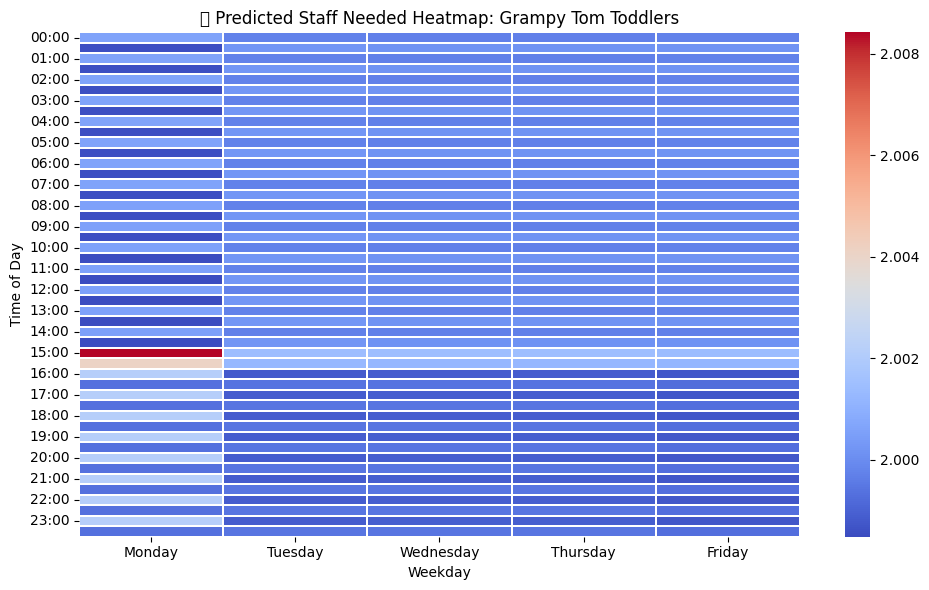
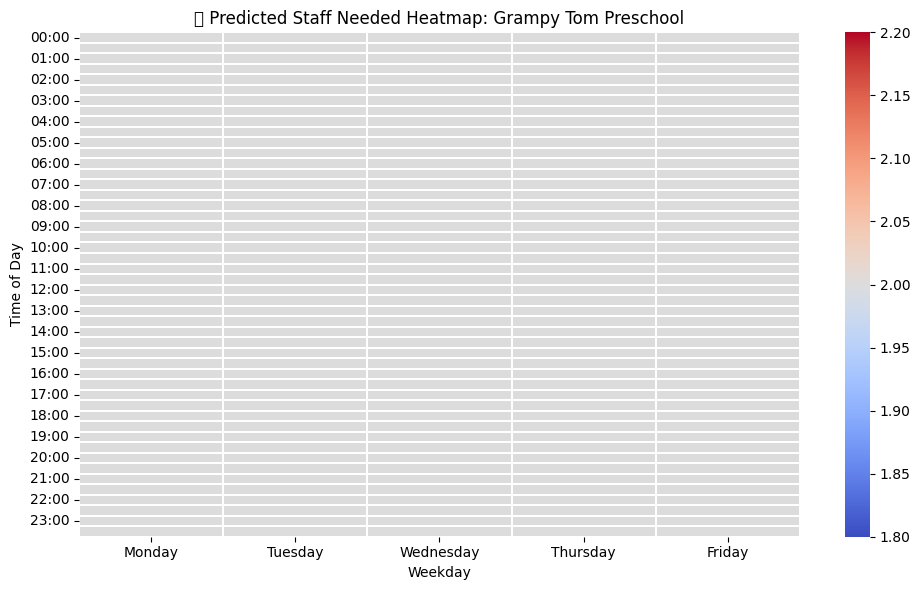
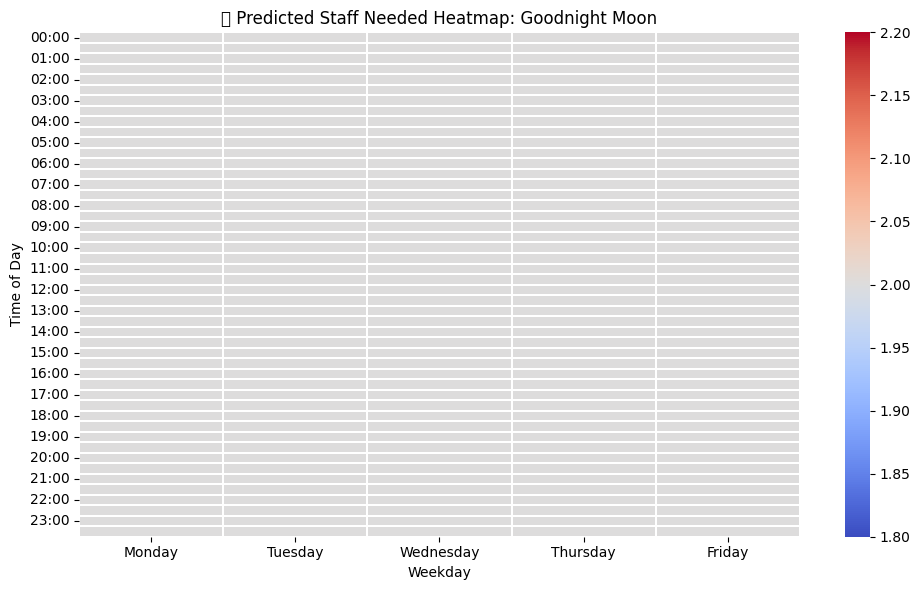
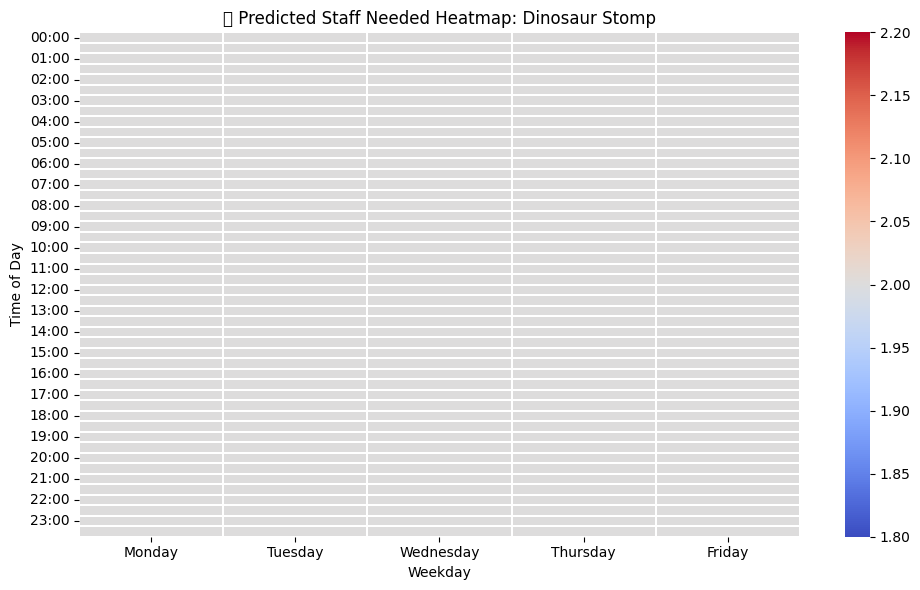
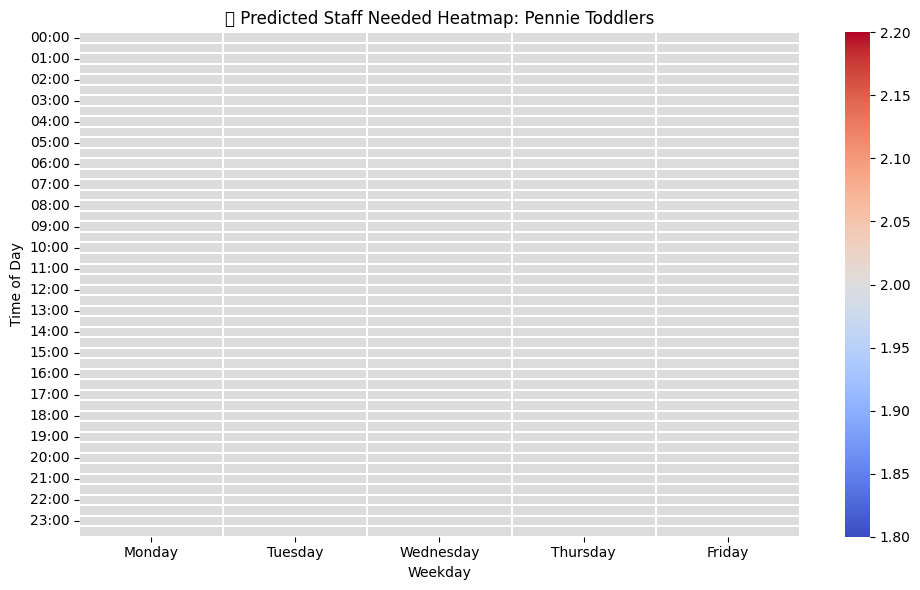
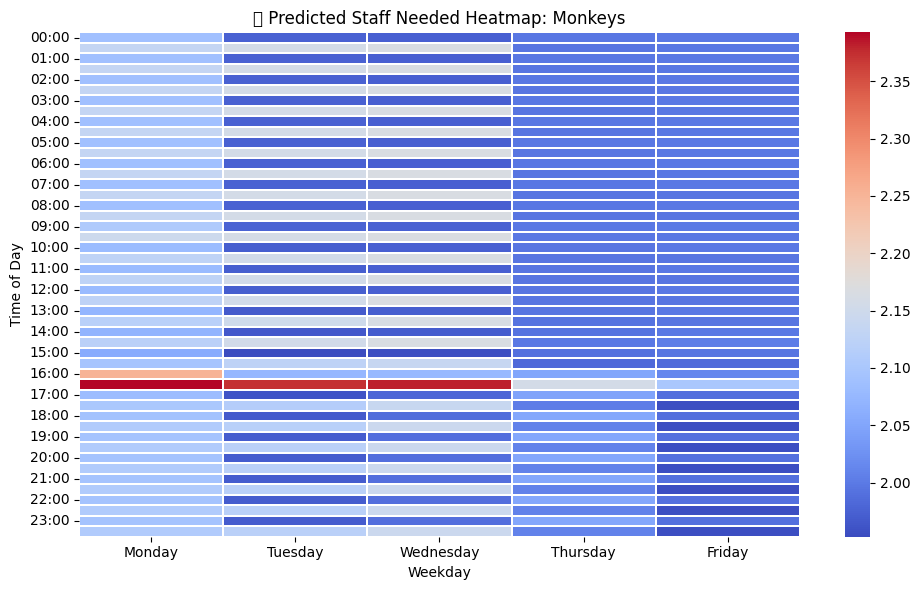
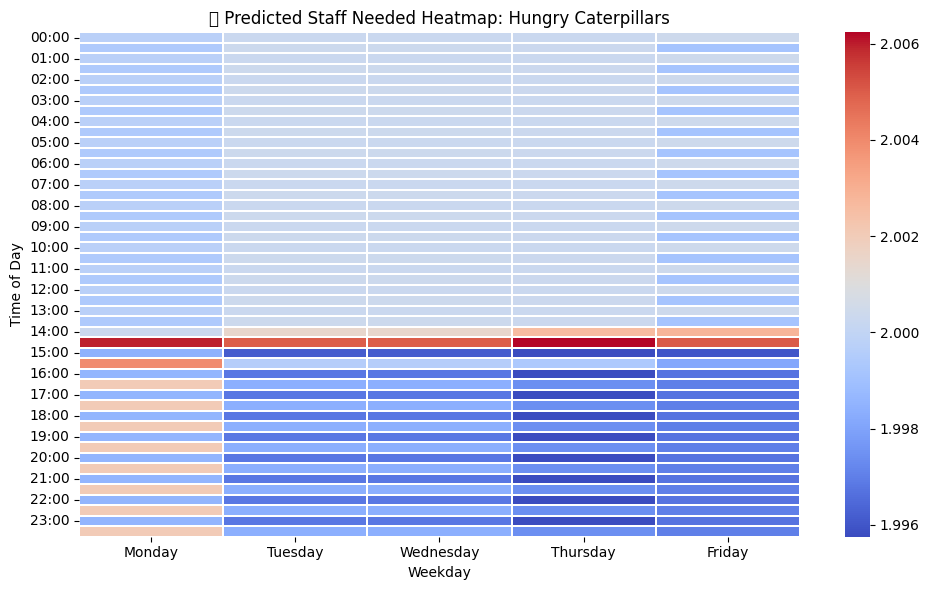
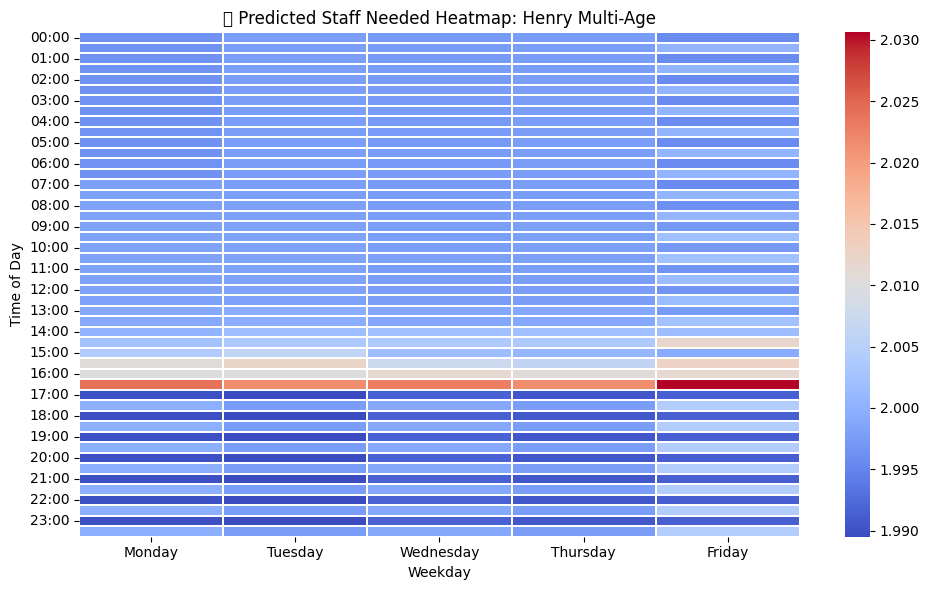
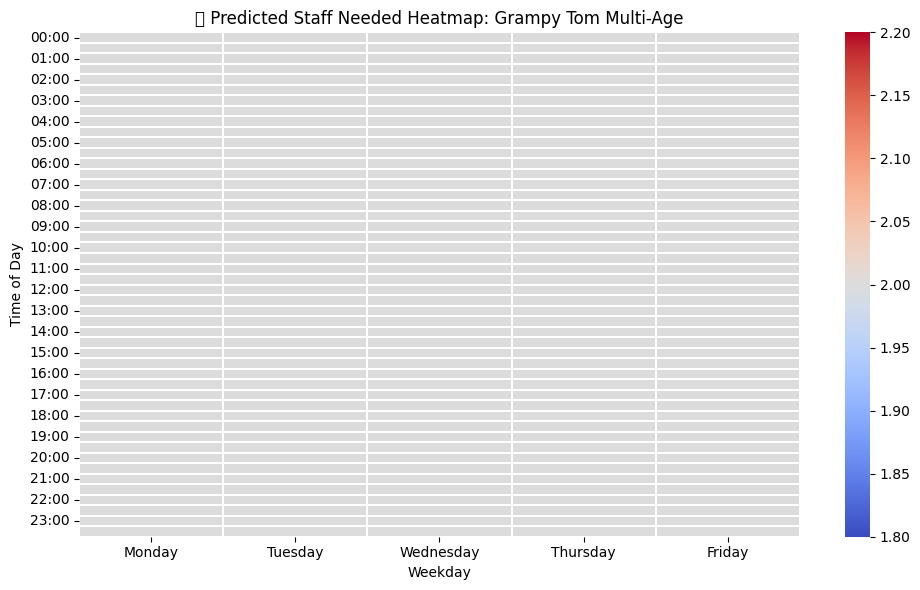
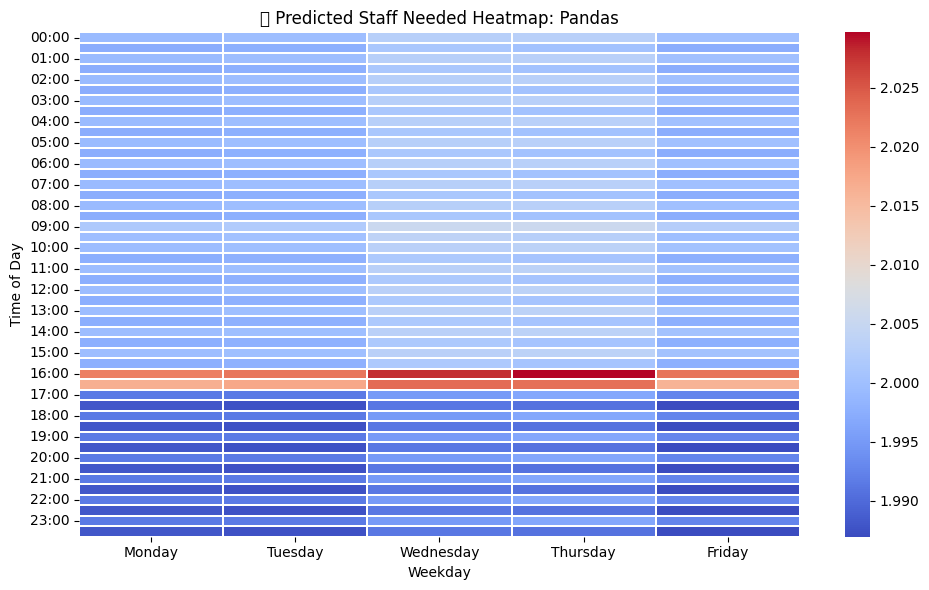
* Most kids are present between 9:00 AM and 12:30 PM, especially on Tuesday, Wednesday, and Thursday.
* The busiest times are all in the mid-morning, when the most children are checked in.
* Early mornings and evenings have very few or no children.
* Staff stays at 2 for most time slots, because that’s the minimum needed no matter how many kids are there.
* Fridays are usually less busy, so fewer staff might be needed.
* The graphs show that attendance is steady across weekdays, which helps in planning regular staff schedules.

## Line plot -Typical week look like:



* Consistent Morning Buildup: Most rooms see child attendance increasing from 7:30 AM to 10:30 AM, reflecting peak drop-off times.
* Midday Peak: Across the week, child counts are highest between 10:00 AM and 12:30 PM in almost all rooms.
* Afternoon Decline: After lunch (around 1:00 PM), attendance gradually decreases as children start getting picked up.
* Few Evening Children: Very low attendance is seen after 4:30 PM, with many rooms nearly empty—staff can be reduced accordingly.
* Room Patterns Vary: Some rooms (e.g., Pre-K1, Toddler) have strong attendance bursts, while others (e.g., Infant, Multi-Age) are more evenly spread throughout the day.
* Staffing Templates Can Be Created: These repeating patterns are useful for building stable weekly staff schedules based on actual need.

## Heatmps (Next week forecast )



* Consistent Minimum Staffing: Most rooms consistently forecast 2 staff members throughout the day, reflecting the required baseline coverage (due to the 2-staff minimum rule).
* Midday Peaks: A few rooms like Panda, Henry Toddlers, and Hungry Caterpillars show slightly increased staff demand between 9:00 AM and 1:30 PM, which aligns with higher attendance blocks.
* Low Variability in Most Rooms: Many rooms such as Pennie Preschool, Pre-K, and Rabbits show flat staff needs due to stable and low child attendance in recent weeks.
* The use of blue/red gradients effectively highlights small variations across time and room types, helping directors identify which rooms need closer monitoring.

**7. Business Impact and Recommendations**

## 7.1 Estimated Cost Savings

One of the major benefits of the forecasting system we implemented is the ability to significantly reduce labor costs while maintaining high-quality care and safety for the children. By optimizing staffing schedules based on actual attendance data, CSI can avoid both overstaffing and understaffing, which are two primary contributors to unnecessary labor costs.

* **Labor Cost Reduction**: Through the application of predictive modeling, we estimate that CSI can reduce its annual labor costs by approximately **10-12%**. Given that the institution currently spends a substantial portion of its budget on staffing, this reduction is expected to result in a savings of about **$40,000 per year**.

This savings is realized by reducing the number of staff required during periods of low attendance while ensuring sufficient staffing during high-demand times. The new system allows for more efficient scheduling, where only the necessary staff members are present based on real-time and forecasted demand.

* **Optimized Resource Allocation**: In addition to direct cost savings, optimized staffing means that the institution can deploy staff in more strategic areas. For instance, during periods of peak attendance in certain rooms, additional staff can be allocated, while during quieter times, fewer staff are required. This reduces the inefficiency of idle staff during low-traffic times.
* **Avoidance of Penalties and Compliance Risks**: Overstaffing is not only costly but can lead to inefficiencies in operations. Understaffing, on the other hand, could violate legal and safety regulations regarding staff-to-child ratios. By utilizing accurate forecasts, CSI can maintain the required staffing levels at all times, reducing the risk of compliance violations and potential fines.

## 7.2 Operational Improvements

The implementation of a data-driven staffing forecast system offers several operational improvements that can enhance the overall efficiency of the childcare centers at CSI. These improvements directly contribute to both staff and child satisfaction, ensuring a smoother, more predictable operation.

* **Enhanced Staff Planning**: The primary operational benefit is the shift from reactive to proactive staff planning. With the new system, CSI can predict staffing needs for each room and time block a week in advance, enabling them to adjust staffing schedules accordingly. This eliminates the need for last-minute changes and minimizes scheduling conflicts.
* **Improved Regulatory Compliance**: Ensuring that staffing levels comply with safety regulations is critical in childcare environments. By using the predictive model, CSI can be confident that each room will have the right number of staff members at any given time. This not only reduces the risk of non-compliance but also ensures that children are adequately supervised, which is a core part of CSI’s mission.
* **Reduction in Scheduling Uncertainty**: Scheduling uncertainties are a common issue for childcare centers. Staff often have to deal with fluctuating attendance patterns that make it difficult to ensure proper coverage. Our model accounts for attendance trends, meaning that the right number of staff can be allocated without unnecessary changes to the schedule.
* **Increased Staff Satisfaction**: With a more reliable and predictable scheduling system, staff can have greater confidence in their work hours. This reduces stress and allows them to plan better both inside and outside of work. A predictable work schedule is highly valued by employees, potentially improving staff retention.

## 7.3 Future Enhancements

While the current system provides an excellent start, there are several ways in which the forecasting system can be further developed and refined to improve its utility and effectiveness.

* **Real-Time Data Integration**: The current model uses historical data to predict staffing needs, but real-time data could allow for even more accurate and responsive forecasting. For example, if a child unexpectedly leaves early or arrives late, the system could adjust staffing recommendations in real time. Integrating real-time data from check-in/check-out systems or attendance records could make the forecasting process more dynamic and adaptive.
* **Expanded Features for Forecasting**: The current model uses basic time-related features such as the day of the week and hour of the day. By expanding the feature set to include more granular variables, such as special events, or individual child characteristics (e.g., age group, special needs), we could improve the accuracy of the forecasts even further. For example, predicting staffing needs during holidays or school breaks might require different staffing levels than regular weeks.
* **Use of Advanced Machine Learning Techniques**: **Neural Networks**, could potentially offer even more accurate predictions by identifying non-linear patterns in the data. These models could be trained to better understand complex relationships between staffing requirements, enrollment changes, and external factors.
* **Forecasting for Other Resources**: The scope of the current project is focused solely on staffing needs, but the forecasting approach could be extended to predict other critical resources, such as meal planning or cleaning schedules. This could further streamline operations, ensuring that all resources are aligned with actual needs rather than estimates.
* **Scalability and Deployment**: As CSI grows and more data becomes available, the forecasting system will need to scale accordingly. The system could be deployed on a larger scale, potentially with more advanced infrastructure for real-time data processing and model updates. A cloud-based solution could be used to host the forecasting system, making it accessible to various CSI centers and administrators.

# 8. Conclusion

## 8.1 Summary of Insights

This project has provided a comprehensive solution to CSI’s staffing challenge by utilizing historical data and advanced forecasting techniques to predict staffing needs. The key insights from the project include:

* Accurate forecasts of staffing requirements lead to better resource allocation, improved compliance with staffing regulations, and significant labor cost savings.
* The use of a combination of **ARIMA models**, **XGBoost**, and **Random Forest** regression allowed us to capture the inherent patterns in the data, such as daily and weekly attendance fluctuations.
* **Visualization tools** (e.g., heatmaps, line charts) were key in interpreting complex patterns and communicating findings to non-technical stakeholders.

By leveraging this model, CSI can not only save costs but also improve operational efficiency, ensuring that they continue to provide high-quality care to children while adhering to safety standards.

## 8.2 Limitations

Despite the success of the project, several limitations need to be addressed to improve the model’s effectiveness and applicability:

* **Data Quality and Completeness**: While the data used was comprehensive, some records had missing or inaccurate timestamps, which may have affected the accuracy of our forecasts. Any improvement in data collection processes (such as standardizing check-in/check-out formats) would improve the quality of the predictions.
* **Predicting Special Events**: The model was not specifically designed to account for special events or holidays that might affect attendance patterns. Future iterations could include special-event tagging or additional features related to such anomalies.
* **Model Generalization**: While the model performed well with the current dataset, there may be variability in attendance patterns at other childcare centers. Additional models and fine-tuning may be required to generalize the solution to new datasets.
* **Scalability**: The forecasting system was built for the existing dataset, but as CSI grows, there will be a need for greater computational resources to maintain its performance. Real-time data processing and model updates may require more infrastructure.

## 8.3 Final Thoughts

This project has successfully addressed a critical operational challenge faced by the Child Saving Institute (CSI) by implementing a data-driven forecasting model for staffing needs. The use of advanced statistical and machine learning models provided accurate, actionable insights that will help CSI optimize their staffing resources, reduce costs, and maintain safety compliance.

Looking forward, CSI has the opportunity to extend this system to other areas of operations, including real-time adjustments and predictions for other resources. The insights gained from this project not only benefit CSI in the short term but also serve as a foundation for future enhancements and optimizations. By continuously refining this model and incorporating new technologies, CSI can continue to improve its efficiency and the quality of care provided to children and families in their community.

## 9. References and Documentation

1. **Pandas Documentation**: Official documentation for pandas used in data manipulation and aggregation.

<https://pandas.pydata.org/docs/>

1. **Scikit-learn Documentation**: For machine learning models and data preprocessing tools used in forecasting.

<https://scikit-learn.org/stable/>

1. **Seaborn Documentation**: For visualization techniques used in heatmaps and categorical plots.

<https://seaborn.pydata.org/>

1. **Matplotlib Documentation**: For creating various visualizations, including line plots and bar charts.

<https://matplotlib.org/3.3.4/#:~:text=Matplotlib%20is%20a%20comprehensive%20library,easy%20and%20hard%20things%20possible.&text=Use%20interactive%20figures%20that%20can%20zoom%2C%20pan%2C%20update>...