Exploratory Data Analysis (EDA) Report

Project: Dynamic Pricing Strategies for Fitness Classes based on Demand, Time, and Location

Author : Knnani Vidita

Objective: To identify patterns in booking behavior, attendance, and pricing that justify a dynamic revenue optimization model.

1. Data Cleaning & Integration (from Component 2)

Before analysis, we performed a rigorous cleaning and merging process to ensure a "Single Source of Truth":

- **Dataset Merging:** Successfully merged the April–May and June 2018 datasets using pd.merge() without data loss.
- **Quality Assurance:** Identified and handled **overbooking anomalies** (where bookings exceeded capacity).
    - Handled negative pricing outliers and standardized currency to INR.
    - Validated that no missing values remained in core columns (Price, Number Booked, Capacity).
- **Feature Engineering:** Created the Occupancy_Rate metric and extracted temporal features (Hour, Day, Month) to enable demand-side analysis.Extracted ActivityName from ActivitySiteID and handled missing values.

2. Temporal Demand Discovery

The EDA revealed a highly predictable, time-sensitive demand structure:

- **The "Double Peak" Phenomenon:** Attendance surges significantly between **06:00–09:00** and **17:00–20:00**.
- **Weekend vs. Weekday:** Mid-week (Tuesday/Wednesday) shows higher professional-class engagement, while weekends show a more distributed, lower-intensity booking pattern.

3. Price-Occupancy Analysis

We examined how current static pricing interacts with actual gym usage:

- **Revenue Inefficiency:** We found many classes with **100% occupancy** priced exactly the same as classes with **10% occupancy**.
- **Site Variance:** Demand patterns vary by location, suggesting that a "Site-Specific" factor must be included in the dynamic pricing algorithm.

4. Activity Performance Benchmarks

- **Core Drivers:** Yoga and Strength training consistently achieve higher occupancy rates (>75%) regardless of the time slot.
- **Inventory Opportunity:** Specialized classes during mid-day "slumps" represent the greatest opportunity for discount-driven occupancy growth.

5. Final Conclusion & Path to Modeling

Based on the EDA and the cleaned data from Component 2, the following steps were justified:

1. **Forecasting:** Use **Prophet** to capture the "Double Peak" seasonality discovered in EDA.
2. **Elasticity:** Quantify the price sensitivity of the "High Performance" activities to set surge limits.
3. **Strategy:** Transition from a "Flat-Rate" to a "Yield-Centric" model to capture the value shown in the peak demand periods.

Appendix: Statistical & Empirical Evidence

Evidence 1: Data Integration and Column Extraction

	ActivitySiteID	ActivityDescription	BookingEndDateTime (Month / Day / Year)	BookingStartTime	MaxBookees	Number Booked	Price (INR)
0	BRP	20:20:20 10-11 Am	01-May-18	10:00:00	35	22	499.0
1	BRP	20:20:20 10-11 Am	03-Apr-18	10:00:00	35	32	499.0
2	BRP	20:20:20 10-11 Am	05-Jun-18	10:00:00	35	25	499.0
3	BRP	20:20:20 10-11 Am	08-May-18	10:00:00	35	30	499.0
4	BRP	20:20:20 10-11 Am	10-Apr-18	10:00:00	35	26	499.0
...	...	...	...	...	...	...	...
3284	TSC	Zumba - 6-7pm	21-Jun-18	18:00:00	50	12	1299.0
3285	TSC	Zumba - 6-7pm	26-Apr-18	18:00:00	50	17	1299.0
3286	TSC	Zumba - 6-7pm	31-May-18	18:00:00	50	13	1299.0
3287	TSC	Zumba 6.00 - 7.00pm	12-Apr-18	18:00:00	35	14	1299.0
3288	TSC	Zumba 6.00 - 7.00pm	19-Apr-18	18:00:00	35	8	1299.0

3289 rows × 7 columns

	ActivitySiteID	ActivityName	BookingEndDateTime (Month / Day / Year)	BookingStartTime	MaxBookees	Number Booked	OccupancyRate	Price (INR)
0	BRP	20-20-20	2018-05-01	10:00:00	35	22	0.63	499.0
1	BRP	20-20-20	2018-04-03	10:00:00	35	32	0.91	499.0
2	BRP	20-20-20	2018-06-05	10:00:00	35	25	0.71	499.0
3	BRP	20-20-20	2018-05-08	10:00:00	35	30	0.86	499.0
4	BRP	20-20-20	2018-04-10	10:00:00	35	26	0.74	499.0

Table A1

Evidence 2: Data Cleaning Results

```
Final Missing Value Check:
ActivitySiteID                          0
ActivityName                            0
BookingEndDateTime (Month / Day / Year)    0
BookingStartTime                        0
MaxBookees                              0
Number Booked                           0
OccupancyRate                           0
Price (INR)                             0
dtype: int64
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3289 entries, 0 to 3288
Data columns (total 8 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   ActivitySiteID                          3289 non-null   object
 1   ActivityName                            3289 non-null   object
 2   BookingEndDateTime (Month / Day / Year)  3289 non-null   datetime64[ns]
 3   BookingStartTime                        3289 non-null   object
 4   MaxBookees                              3289 non-null   int64
 5   Number Booked                           3289 non-null   int64
 6   OccupancyRate                           3289 non-null   float64
 7   Price (INR)                             3289 non-null   float64
dtypes: datetime64[ns](1), float64(2), int64(2), object(3)
memory usage: 205.7+ KB
```

Table A2

Evidence 3: The "Double Peak" Demand Curve

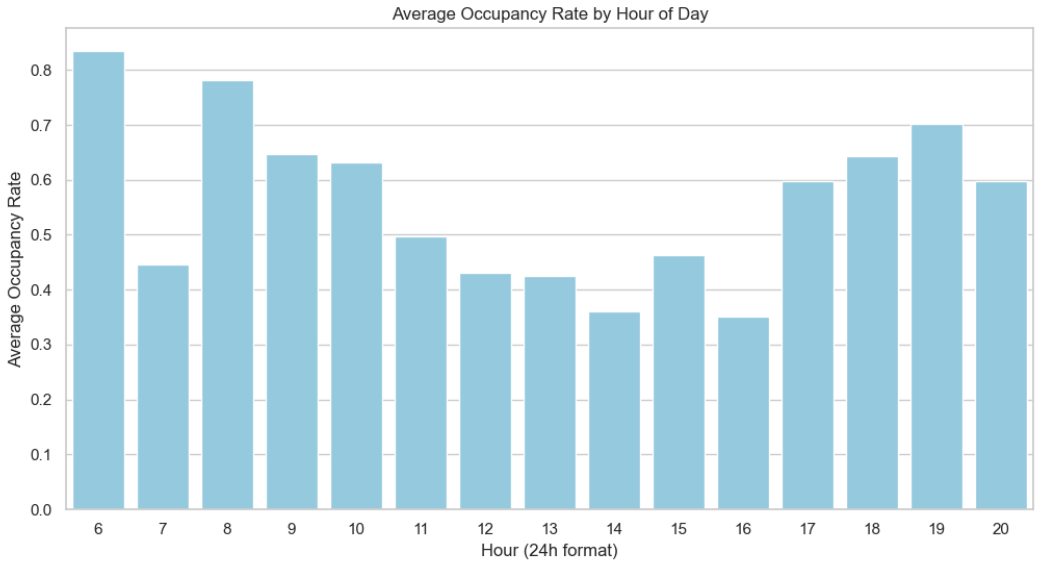

Figure A1

Evidence 4: Day-of-Week Attendance Variance

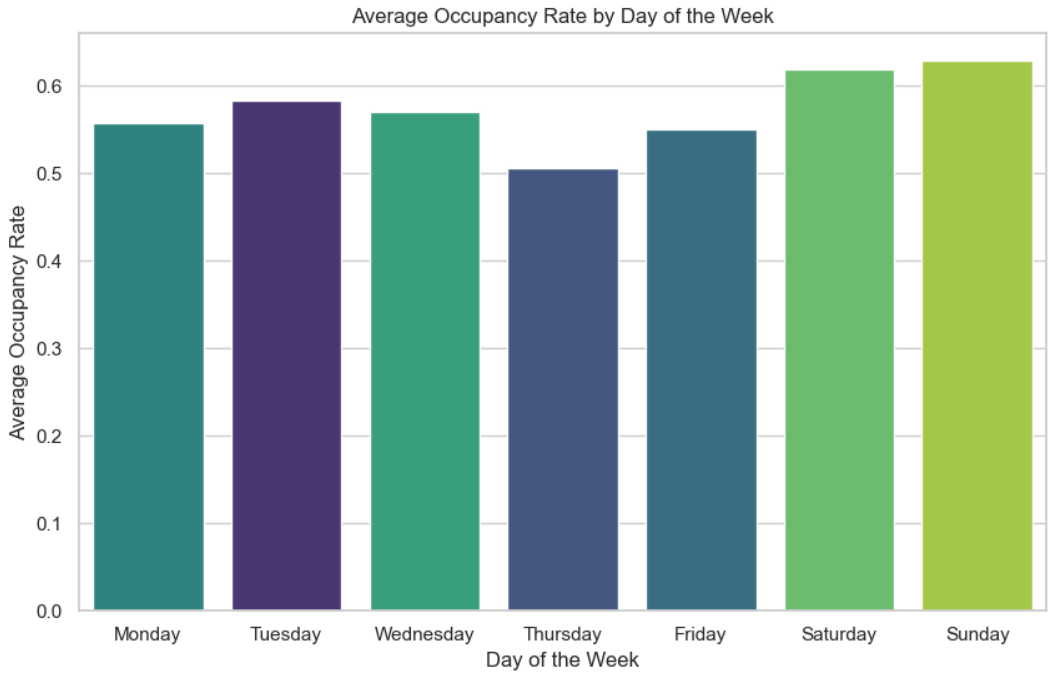

Figure A2

Evidence 5: Activity-Wise Occupancy Benchmarks

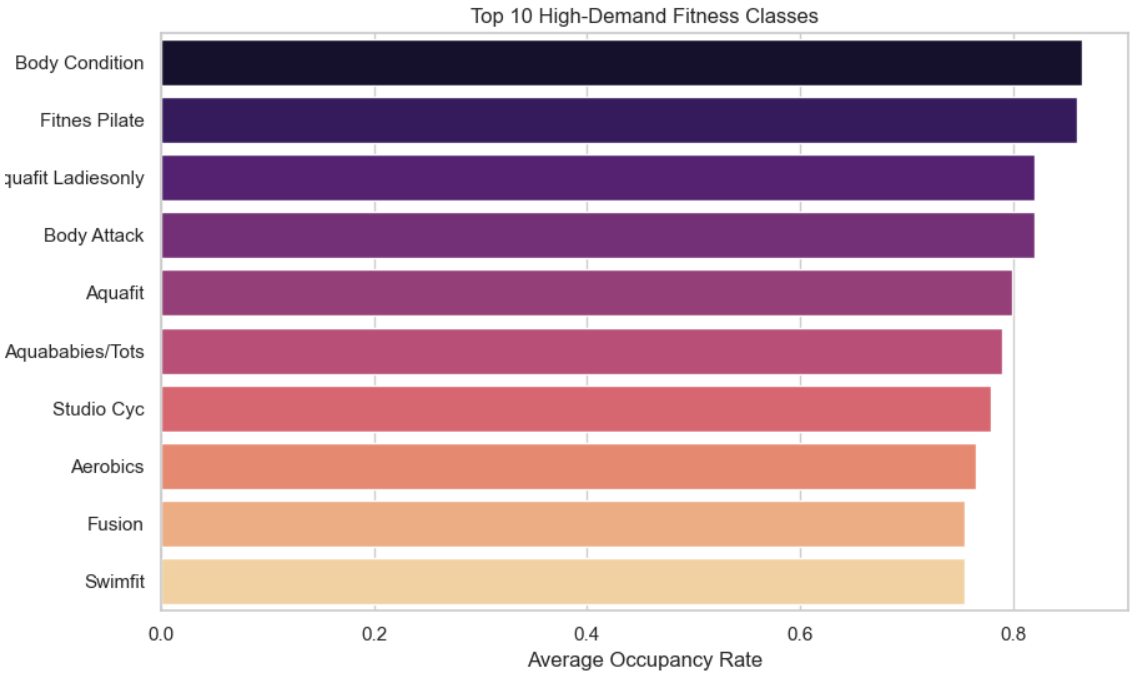

Figure A3

Evidence 6: Average Occupancy Rate by Location

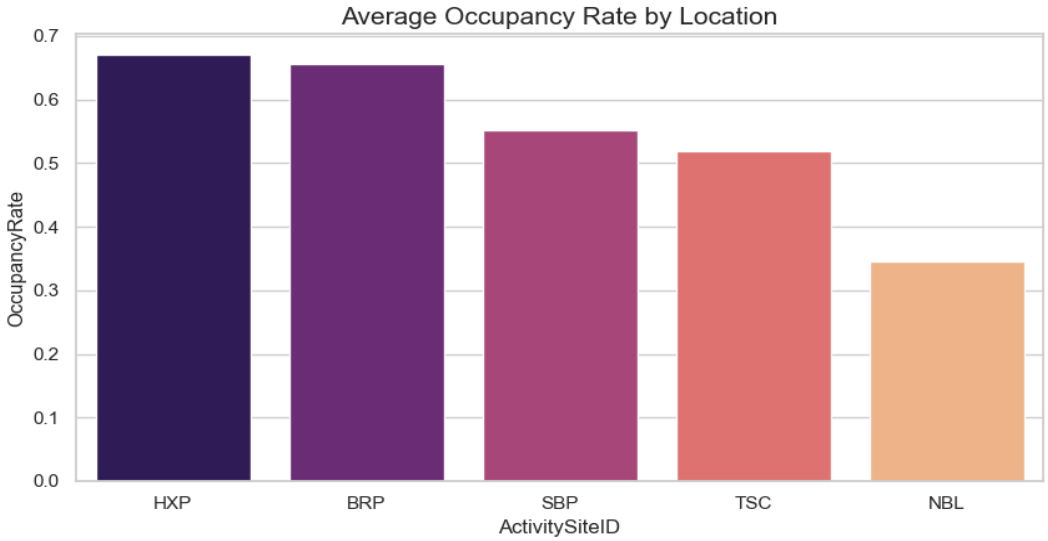

Figure A4

Evidence 7: Anomalies Detection

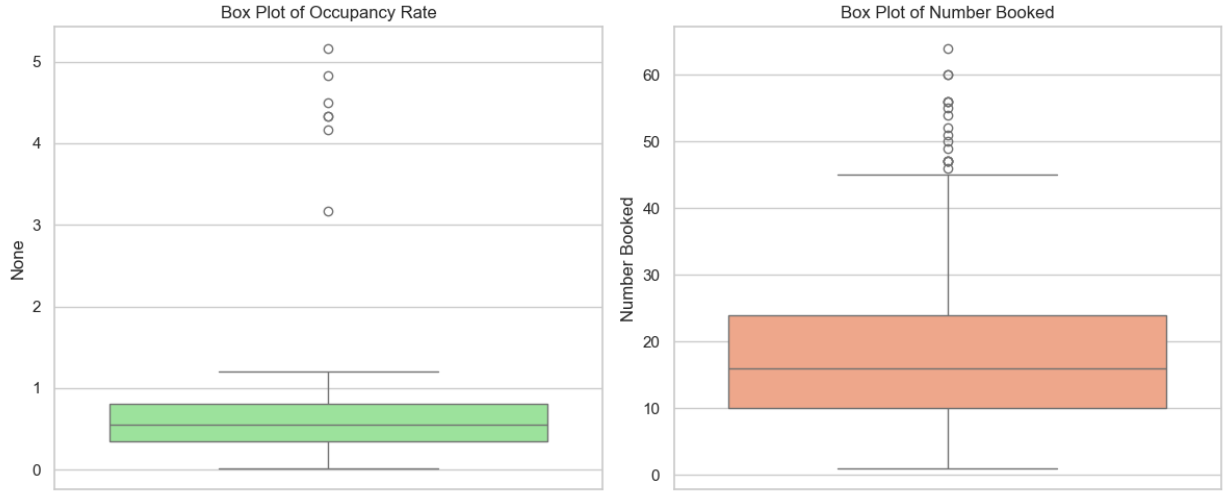

Figure A5