Impact of Seasonal Changes on Gym Usage

Kritthika Shanmugan, Keerthi Gajjela, Sathiyashivani Sathish Kumar
Team: 03, DS522 -Data Acquisition and Analytics

shanmugamkritthika@cityuniversity.edu
gajjelakeerthi@cityuniversity.edu
sathishkumarsathiya@cityuniversity.edu

Abstract

We reported the impact of seasonal changes on gym usage patterns to understand how factors such as temperature, daylight duration, holidays and seasonal activities influence gym attendance. By analyzing different datasets in the Kaggle, we tried to find the fluctuations in member attendance across different times of the year. In this, we are working on Python for data acquisition and Power BI for visualization. We do process like collecting different dataset, data cleaning, with the chosen data, through statistical analysis and data visualization, we plan to identify trends in user behavior and determine if specific seasons correlate with increased or decreased gym activity. From these details, this can help fitness facility management strategies, allowing for better resource allocation, tailored membership promotions, and optimized staff scheduling during peak and off-peak seasons. This project says how seasonal factors influence exercise habits, supports strategic planning for gyms and adds to broader health and wellness initiatives.
Keywords: Power BI, Data cleaning, Seasonal Variability, Gym Attendance Patterns, Exercise Habits, Behavioral Analytics.

1. INTRODUCTION

Seasonal changes and holidays have a significant impact on people's lifestyles, including their gym attendance patterns. Factors like temperature, daylight hours, holidays, and seasonal activities affect motivation and availability for exercise. Understanding these patterns is essential for gym owners and managers to improve their operational efficiency and member satisfaction. As we work on this project, we look at how these seasonal changes affect the use of the gym by evaluating data from the gym membership and overcrowding datasets. We will apply concepts learned in the readings of Python through the acquisition and preprocessing-data cleaning and wrangling-to make sure that the datasets will be appropriate and valid for analysis. These steps are important in finding inconsistencies, missing values, or formatting issues so that subsequently accurate insights can be allowed.

We do use of statistical analysis, as we learned from these readings to bring out the trends and correlations, followed by visualization through Power BI to communicate our findings in an easy-to-understand manner. Charts will visually indicate such patterns as high attendance during colder months or low attendance during summer vacations, thereby drawing actionable inferences. This project presents relevant datasets, such as the Gym Membership Dataset from Kaggle and a manually collected dataset on gym crowdedness. The combination of these resources develops a better understanding of the trends of the year in gym usage.

The study will therefore help gym managers with information for better staffing decisions, resource allocation, and promotion strategies. The gyms will, by using seasonal and holiday patterns, be able to align operational optimizations to help meet the demands of their members for a broader improvement in user experience in their health and wellness initiatives. This will be followed by a step-by-step analysis of how external factors influence behavior in the gym, using strong preprocessing and analysis techniques to make this research valuable and impactful.

2. LITERATURE REVIEW

Garriga, Sempere-Rubio, Molina-Prados, and Faubel (2022) explore the seasonal effects on PA levels through a systematic review from 18 countries. Seasonal variation in PA is detected, reporting higher levels in summer than in winter.

This trend was also seen across populations with different chronic diseases, such as COPD and heart failure. The authors point out that in the analytical process, 26 studies using objective tools (like pedometers and accelerometers) and subjective tools (like questionnaires) underlined the precision of objective tools in catching the seasonal variation that subjective instruments often underestimate. The research underlines the importance of seasonality in configuring PA patterns and how this has implications for public health strategies. Some proposed interventions are light PA promotion during winter to balance increased sedentary behavior. The findings emphasize the importance of taking seasonality into account when designing PA promotion programs and clinical research aimed at ensuring year-round activity. This review provides some important information to the policymakers and health professionals in developing tailored strategies to overcome seasonal barriers to physical activity.

Buchowski et al. (2009) investigate seasonal variation in physical activity (PA) among women who reside in metropolitan Nashville, TN, and surrounding counties. In this study, 57 healthy women aged between 20–54 years with a range of body mass indices were engaged in measuring their PA for 7 continuous days across three seasons of the year using accelerometers. In addition, EE and PA intensity was measured as METs by indirect calorimetry. Results suggest that PA is significantly lower in winter compared with summer-131 ± 45 vs. 144 ± 54 × 10^3 counts/day; P = .025-and spring/fall, P = .027. Activity on weekends was more severely reduced in winter; PA counts were 22,652 lower (P = .008). Sedentary time was increased by 35 minutes/day in winter (P =.007), while light and moderate-to-vigorous activities were reduced by 29 minutes/day (P =.018) and 6 minutes/day (P =.051), respectively. These findings indicate a strong seasonal effect on PA, particularly in weekends, which is of critical importance in the development of tailored programs of PA that can maintain activity throughout the year.

Rand et al. (2020) systematically reviewed interventions that were aimed at improving attendance at health and fitness venues. It considered BCTs and included 14 studies covering 20 interventions. The review revealed a general difficulty in sustaining attendance, with it frequently declining after the initiation of membership. Most interventions had either a trivial or small effect on attendance, while four interventions revealed medium to large effect sizes (Cohen's d = 0.60–1.45). The number of BCTs used varied, but the most common was "Prompts/Cues." Interventions reporting larger effects also used "Problem solving," "Pros and cons," and "Goal setting (behavior)" combined with reviewing behavior goals. Nonetheless, the authors point to the general lack of evidence regarding the actual effective BCTs and encourage more detailed research in trying to identify the active ingredients of effective interventions. This review underscores the potential of tailored strategies to boost attendance, benefiting public health by promoting sustained physical activity through improved engagement with fitness venues(12889_2020_Article_9898).

McKinney, W. (2022) provides comprehensive guidance on data wrangling and analysis using Python. He discusses the cleaning, transformation, and visualization of data, making this highly relevant to projects that use large datasets. There are practical examples in the book using Pandas and NumPy; thus, these provide tools for efficient data pre-processing. Such methodologies are important in ensuring that datasets, while being analyzed, are well-structured and reliable. In this study, cleaning and merging of datasets were done using Python-based techniques to yield accurate insights into the trends of use of gyms. Given McKinney's focus on practical applications, it is a good resource for the implementation of statistical and analytical techniques in real-world scenarios.

3. DATA

We use for the efficient analysis of seasonal variation and holidays impact on the attendance of the gym, we implement two rich and detailed datasets from Kaggle. These datasets are Gym Membership Dataset and Crowdedness at Campus Gym Dataset. Together, these provide a sound foundation for ascertaining patterns in usage, preference of members, and extra factors responsible for trends in attendance. The Gym Dataset consists of 1,000 entries across 17 columns, including demographics of members, types of membership, and usage patterns. The Crowd Dataset comprises 62,184 entries across 11 columns, including the crowd sizes and times for attendance and external factors such as weather conditions, holidays, or weekends, to help examine attendance patterns.

Gym Membership Dataset

We elaborated on the dataset and found that it includes demographic and behavioral information of the people who are into gym membership. It includes age, gender, type of membership, and visits per week. The main attributes include visit_per_week and days_per_week, which will help in understanding the number of regular visitors along with the frequency of peak usage. These would help with understanding the number of regular visitors and how often peak usage would occur. Some additional information on average check-in and check-out times is also provided, as well as extra service usage concerning personal training and a sauna. These trends will provide insight into member group behavior, comparing premium and standard memberships and identifying the high or low-attendance periods, thus contributing to optimized gym management.

```
GYM DATASET
   id  gender    birthday  Age abonoment_type  visit_per_week  \
0   1  Female  1997-04-18   27        Premium               4
1   2  Female  1977-09-18   47       Standard               3
2   3    Male  1983-03-30   41        Premium               1
3   4    Male  1980-04-12   44        Premium               3
4   5    Male  1980-09-10   44       Standard               2

        days_per_week  attend_group_lesson            fav_group_lesson  \
0  Mon, Sat, Tue, Wed                 True  Kickboxen, BodyPump, Zumba
1       Mon, Sat, Wed                False                         NaN
2                 Sat                 True                       XCore
3       Sat, Tue, Wed                False                         NaN
4            Thu, Wed                 True       Running, Yoga, Zumba

  avg_time_check_in avg_time_check_out  avg_time_in_gym  drink_abo  \
0          19:31:00           21:27:00              116      False
1          19:31:00           20:19:00               48      False
2          08:29:00           10:32:00              123       True
3          09:54:00           11:33:00               99       True
4          08:29:00           09:19:00               50      False

            fav_drink  personal_training name_personal_trainer  uses_sauna
0                 NaN              False                   NaN        True
1                 NaN               True               Chantal       False
2  berry_boost, lemon               True                  Mike       False
3       passion_fruit               True                  Mike        True
4                 NaN               True                  Mike       False
```

Crowd Dataset

The Crowd dataset contains temporal data reflecting the attendance to a gym, which depends on factors related to weather conditions, holidays, and academic schedules. Some of the key fields include number_people, date, and timestamp, which are relevant for the counting of crowd size at every hour of the day. Other external influences include is_weekend, is_holiday, is_start_of_semester, and is_during_semester, with which one can analyze attendance patterns related to weekends, holidays, and academic schedules. Temperature provides further insights into how weather conditions affect gym attendance. Combining this dataset with the Gym Membership dataset will provide a full view of what determines the use of the gym over time.

```
CROWD DATASET
   number_people                       date  timestamp  day_of_week  \
0             37  2015-08-14 17:00:11-07:00      61211            4
1             45  2015-08-14 17:20:14-07:00      62414            4
2             40  2015-08-14 17:30:15-07:00      63015            4
3             44  2015-08-14 17:40:16-07:00      63616            4
4             45  2015-08-14 17:50:17-07:00      64217            4

   is_weekend  is_holiday  temperature  is_start_of_semester  \
0           0           0        71.76                     0
1           0           0        71.76                     0
2           0           0        71.76                     0
3           0           0        71.76                     0
4           0           0        71.76                     0

   is_during_semester  month  hour
0                   0      8    17
1                   0      8    17
2                   0      8    17
3                   0      8    17
4                   0      8    17
```

4. DATA CLEANING AND PREPROCESSING

Merging Data

Using the two datasets, we merged the data using a right join on the Gym Membership and Crowdedness at Campus Gym datasets. This will be the reason for choosing a right join because we would want to preserve all the data from the crowd dataset, as it contained timestamped information on when people go to the gym and external factors such as temperature and holidays and how big the crowds are. The demographic and usage information related to individual members was obtained from the Gym Membership dataset. This was done to ensure that both datasets would align correctly by adding an id column to the Crowdedness at Campus Gym Dataset using the index, thereby creating a common identifier to merge on. The result was merged on the id column with the Crowdedness at Campus Gym Dataset.

Data Cleaning in the Gym Membership Dataset

Handling Missing Values

For this, we replaced the missing values in the categorical columns, such as gender, with a default value ('Unknown'). For numerical columns such as age, missing values were filled with the mean of that column.

Transforming Data

The days_per_week column in the Gym Membership Dataset contained a list of days on which a member usually visits the gym. We created the visit_per_week column by counting the number of days listed in the days_per_week column. In visit_per_week, missing values were imputed with the average value calculated from the available records.

Data Cleaning in the Crowdedness Dataset

Date Conversion and Temperature Clipping
The date column was converted to the proper datetime format, allowing the facilitation of time-based analysis. This transformation will ensure that chronological data is handled correctly. The temperature column contained values over an enormous range. We have clipped the temperature values to lie between 60°F and 70°F because extremely high or low temperatures outside this range were unlikely to influence gym attendance in the context of the current project.

Filtering Data
Since the capacity of this gym is limited, any entry with the value of number_people over 50 was excluded, because at any one time, the maximum occupancy in that facility is 50 members.

```
RangeIndex: 51581 entries, 0 to 51580
Data columns (total 13 columns):
```

The combined dataset that results from all these integrations contains 51,581 records with 13 columns, incorporating member demographics, attendance frequencies, and temporal and seasonal trends. The comprehensive data, cleaned up and saved as cleaned_combined_gymdata.csv, provides a robust backbone upon which seasonal and holiday impacts can be analyzed effectively for actionable insights to improve gym operations.

```
   id   age  gender  visit_per_week                       date  number_people  \
0   1  27.0  Female               4  2015-08-14 17:00:11-07:00             37
1   2  47.0  Female               3  2015-08-14 17:20:14-07:00             45
2   3  41.0    Male               1  2015-08-14 17:30:15-07:00             40
3   4  44.0    Male               3  2015-08-14 17:40:16-07:00             44
4   5  44.0    Male               2  2015-08-14 17:50:17-07:00             45

   temperature  is_weekend  is_holiday  day_of_week  month  season  \
0         70.0           0           0            4      8  Summer
1         70.0           0           0            4      8  Summer
2         70.0           0           0            4      8  Summer
3         70.0           0           0            4      8  Summer
4         70.0           0           0            4      8  Summer

   weekend_visits
0               0
1               0
2               0
3               0
4               0
```

5. DATA WRANGLING

I would like to explain about data wrangling which also known as cleaning or pre-processing of data, is an activity that transforms and cleans our raw data into a more suitable analysis form. It includes tasks on missing values handling, transformation of variables, merging datasets, creating new features, etc. We used data wrangling to ensure that the information is accurate, complete, and well-structured for further analysis.

We found data wrangling is an important step in preparing the gym attendance data for analysis in this project. The raw data consists of several datasets that are inconsistent, with some values missing and some variables needing transformation to become insightful. For example, one key step we used here, to create a season column by categorizing the month in which people visit the gym.

```
Season trends
 0       Summer
1        Summer
2        Summer
3        Summer
4        Summer
          ...
51576    Spring
51577    Spring
51578    Spring
51579    Spring
51580    Spring
Name: season, Length: 51581, dtype: object
```

In this idea we achieved by mapping each month to one of the seasons-Winter, Spring, Summer, or Fall-using a function and applying that to the dataset. We also had planned to use wrangling step in creating the weekend_visits column. It computes attendance over weekends by multiplying the is_weekend column with the number_people column; thus, allowing the dataset to reflect the pattern of attendance specifically for weekends.

```
Weekend_visits
 0        0
1        0
2        0
3        0
4        0
        ..
51576    23
51577    21
51578    25
51579    18
51580    23
Name: weekend_visits, Length: 51581, dtype: int64
```

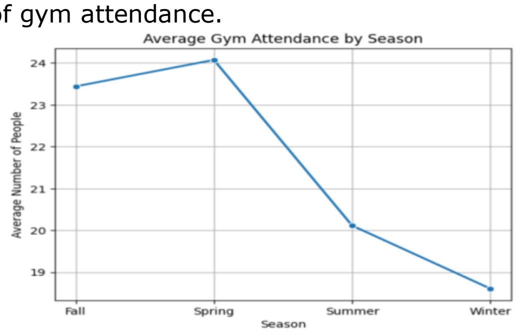
These are wrangling techniques that helped us in cleaning and reshaping the data to extract meaningful patterns such as seasonal variations and weekend attendance. Data wrangling converts the dataset into an analyzable form, so the dataset is ready for detailed statistical analysis and visualization, which is the backbone of the insights to be generated in this project. If proper data wrangling were not performed, the raw data would not be appropriate for analysis or decision-making.

6. RESULTS

Several major steps to analyze the gym attendance trends in the use of different techniques:

Average gym attendance by season

Calculating the average attendance per season from this dataset using a grouping of a newly created column "Season" and the application of a grouping and aggregation technique, determines the mean of the number_people values per season. This is important because one gets to see the comparative situation across the four seasons of gym attendance.

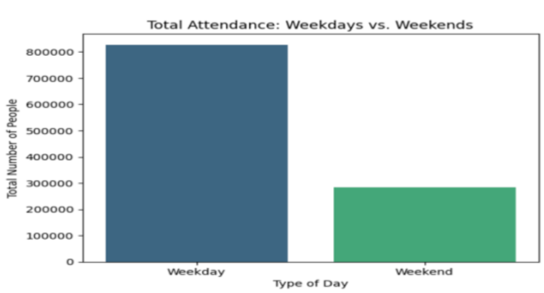


We created a line plot visualizing average gym attendance by season. we used sns.lineplot to plot the data with seasons on the x-axis and average attendance on the y-axis, adding markers for each data point.

Total Attendance: Weekdays vs. Weekends

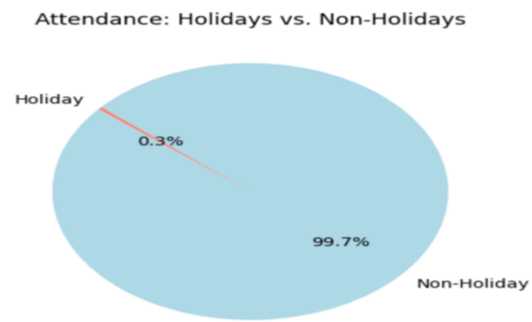
Next, the weekend vs. weekday analysis is performed through filtering based on the is_weekend column. The dataset is split into two groups: weekends and weekdays.

The filtering and summing technique is used to find the total number of attendees for each group by summing the number_people values. It gives insights into how gym attendance differs between weekends and weekdays. Here, we displayed the total attendance for a weekday versus a weekend, using a bar plot. We created a bar plot of data using sns.barplot.



Attendance: Holidays vs. Non-Holidays

We did the analysis of attendance on holidays, the same method is followed. It filters the data based on the is_holiday column to get only the visits that were during holidays. It then again uses the filtering and summing technique to find the total attendance during holidays by summing the number_people for all the holiday visits. This gives insight into how holidays impact the attendance at the gym.

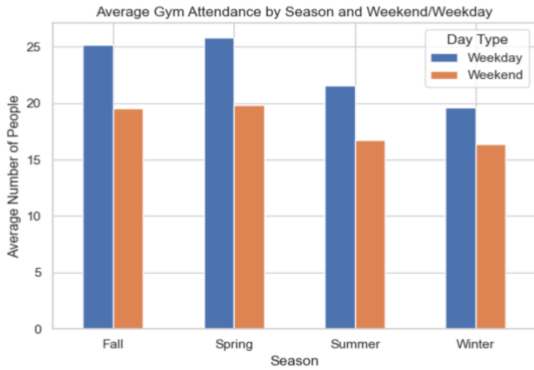


Finally, the processed dataset we saved using the exporting data technique. This modified dataset, with added columns such as season and weekend_visits, is saved as a CSV file named processed_gym_attendance.csv. We planned to make the data prepared for further analysis and visualization in, say, Power BI and get it ready for in-depth reporting and decision-making. Each of these steps employs techniques that will help in transforming raw data into actionable insights to comprehend the attendance trends of the gym across various time cycles.

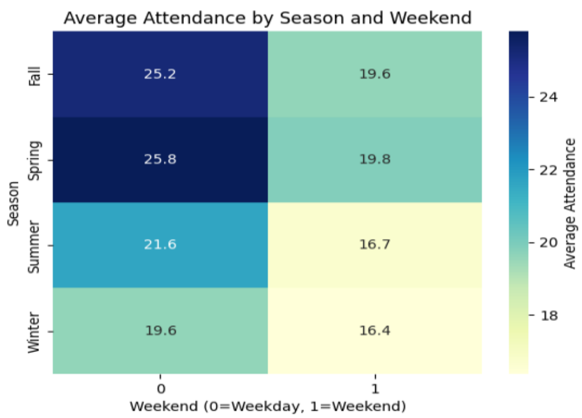
Pivot Table

A pivot table is a data summarization tool used in data analysis by reorganizing and aggregating data into a compact table to facilitate comparison more easily. Here, we used it to calculate the average number of people attending the gym based on seasons and whether the day falls on a weekend or not. The pivot table hereby groups data by season and differentiates between weekend and weekday attendance. The output shows the average attendance for both weekdays and weekends for each season: Winter, Spring, Summer, and Fall.

In our project, we used two different types of visualizations to present the pivot table data. we used a bar chart to display the same data, comparing weekday and weekend attendance across different seasons. The bar chart allows us to visually compare the attendance numbers for each season, highlighting differences between weekends and weekdays.



we used a heatmap to show the average attendance by season and weekend/weekday type.



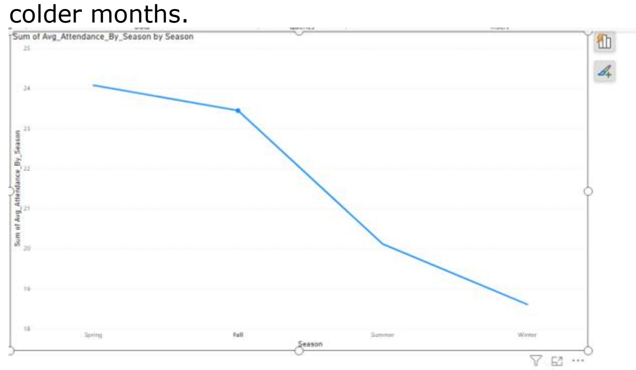
POWER BI Visualization

Power BI was important to visualize the pattern and insights derived from our dataset, through which we did several dynamic and interactive dashboards, highlighting the communication of findings effectively.

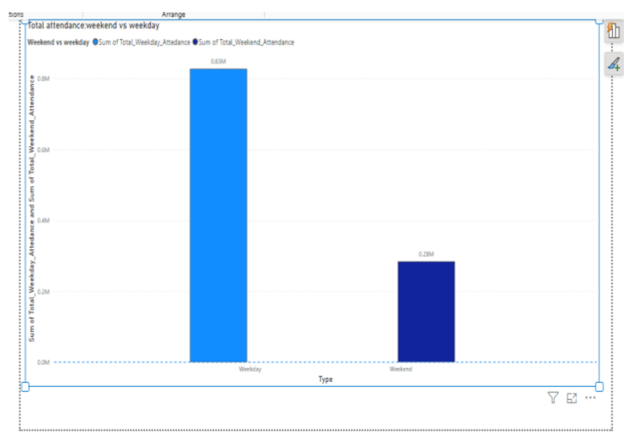
We performed differentials of the two data sets were merged and cleaned using Power BI for easier creation of insightful visualizations. We established Power BI in the relationship between tables with the removal of redundant irrelevant data. The major transformation that took place included filling in missing values, erasing duplicates, and the harmonization of columns against their consistency.

The resulting dataset was then visualized through Power BI dashboards, which were used to show trends, correlations, and actionable insights. We used line charts to display time-based trends and bar graphs to highlight categorical distributions. In addition, we utilized custom DAX queries in calculating the aggregated metrics, such as average attendance, seasonal trends, and user engagement rate, which enhanced our analysis.

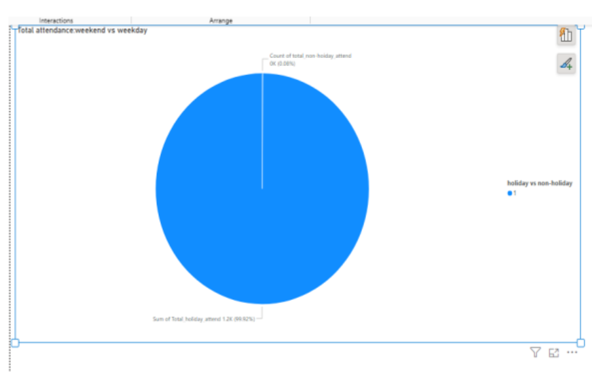
All project deliverables, including the Power BI file (.pbix), screenshots of the dashboards, DAX queries, Python scripts used for pre-processing, and outputs, are hosted on GitHub for transparency and reproducibility. Visualizations used are line charts, bar graphs, and heatmaps, which displayed the trends in gym attendance: seasonality, weekdays vs. weekends, and holidays vs. non-holidays. For example, a line chart indicated the presence of a seasonal pattern: more people attended in warmer than in colder months.



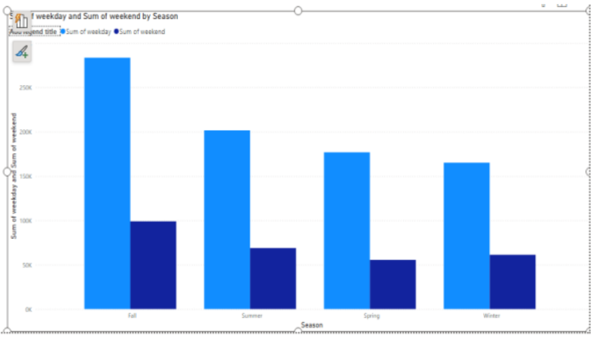
The bar graphs, showing the difference in attendance during weekends and holidays, helped the management to plan accordingly.



Holiday vs non-holiday



Pivot Table for all seasons based on weekdays and weekends



7. CONCLUSION

We finally focused on uncovering the pattern of people who attended the gym; most notably, seasonal and holidays, along with weekend breaks. The screenshot data showed low occupancy in colder months such as winter and high activities during summer, while weekends were the only time when the majority noticed low attendance. We did essential parts of cleaning and merging data included handling missing values, feature engineering, and finding out an important insight.

Mainly, we did visualization with Python and Power BI on attendance dynamics enabled managers to make better staffing, promotions, and facility preparations during peak periods. The following analysis underlines how weather and holidays are important determinants of demand for this gym.

However, we found some limitations: the trends may vary by location and type of gym, and the datasets from Kaggle might not fully represent diverse gym users across demographics. Future research could incorporate variables such as gym capacity and engagement strategies to gain deeper insights into attendance patterns and their role in promoting health and wellness.

8. WORKLOAD ASSIGNMENT

Shivani Sathish Kumar- Shivani focuses on data collection, preprocessing, and analysis, ensuring the dataset is clean and ready for exploration. She takes responsibility for identifying inconsistencies, handling missing values, and organizing data into a usable format.

Kritthika Shanmugam - Kritthika specializes in data analysis and visualization, identifying patterns and trends related to gym usage, particularly seasonal variations. Using tools such as Python and Power BI, she will create meaningful visual representations to simplify complex data. Her visualizations will be instrumental in highlighting key insights, making the results accessible and easy to understand.

Keerthi Gajjela - Keerthi takes the lead in report writing and presentation preparation, ensuring all project findings are compiled into a clear and concise report.

7. REFERENCES

Garriga, A., Sempere-Rubio, N., Molina-Prados, M. J., & Faubel, R. (2021). Impact of seasonality on physical activity: A systematic review. *International Journal of Environmental Research and Public Health*, *19*(1), 2. https://doi.org/10.3390/ijerph19010002

Buchowski, M., Choi, L., Majchrzak, K., Acra, S., Mathews, C., & Chen, K. (2009). Seasonal changes in amount and patterns of physical activity in women. *Journal of Physical Activity and Health*, *6*(2), 252–261. https://doi.org/10.1123/jpah.6.2.252

Rand, M., Norman, P., & Goyder, E. (2020). A systematic review of interventions to increase attendance at health and fitness venues: identifying key behaviour change techniques. *BMC public health*, *20*(1), 1874. https://doi.org/10.1186/s12889-020-09898-6

McKinney, W. (2022). *Python for data analysis: Data wrangling with Pandas, NumPy, and Jupyter* (3rd ed.). O'Reilly Media Inc.

Adam, T. (n.d.). *Gym membership dataset: Exploring membership trends: A synthetic dataset for EDA and machine learning* [Data set]. Kaggle. Retrieved October 14, 2024, from https://www.kaggle.com/datasets/ka66ledata/gym-membership-dataset/data

Rose, N. (n.d.). *Crowdedness at the campus gym: Number of attendees every 10 minutes from the last year in the gym* [Data set]. Kaggle. Retrieved October 14, 2024, from https://www.kaggle.com/datasets/nsrose7224/crowdedness-at-the-campus-gym

Kritthika. (2024). *Kritthika-DS522_GYM_USAGE_ANALYSIS*. GitHub. https://github.com/Kritthika/Kritthika-DS522_GYM_USAGE_ANALYSIS