

Gym Members – Analysis Report

| Scope | |
|-------------|---|
| Assumptions | For the sake of interpretability, the dataset is used under the assumption that it is complete and gathered from a single gym location. |
| Objective | <p>The objective of this data analysis is to provide gym owners with useful data on customer behavior and preferences.</p> <p>Customers can be segmented according to age, gender, experience level and BMI. An average customer profile is created for each of the customer segments. These profiles can be used to target gym services better for each segment.</p> |
| Hypothesis | This analysis operates under the hypothesis that key characteristics essential to customer satisfaction can be identified within different customer segments. |

| Transformations | |
|------------------------|---|
| Medallion Architecture | <p>To enhance reusability and align with the general principles of medallion architecture, the dataset is processed in the following stages:</p> <ul style="list-style-type: none">• Bronze: The raw dataset is ingested, and a schema is applied.• Silver: The validated dataset undergoes filtering, and unnecessary dimensions are removed.• Gold: The production-ready data is used for analysis and reporting. |
| Schema | <p>The following changes are applied to the raw data.</p> <p><u>Renaming columns</u></p> <p>"Age": "age", "Gender": "gender", "Weight (kg)": "weight_kg", "Height (m)": "height_m", "Max_BPM": "max_bpm", "Avg_BPM": "avg_bpm", "Resting_BPM": "resting_bpm", "Session_Duration (hours)": "session_duration_hours",</p> |

```
"Calories_Burned": "calories_burned",  
"Workout_Type": "workout_type",  
"Fat_Percentage": "fat_percentage",  
"Water_Intake (liters)": "water_intake_l",  
"Workout_Frequency (days/week)": "workout_frequency_days_per_week",  
"Experience_Level": "experience_level",  
"BMI": "bmi"
```

Enforcing Datatypes

```
"age": "Int64",  
"gender": "string",  
"weight_kg": "Float64",  
"height_m": "Float64",  
"max_bpm": "Int64",  
"avg_bpm": "Int64",  
"resting_bpm": "Int64",  
"session_duration_hours": "Float64",  
"calories_burned": "Float64",  
"workout_type": "string",  
"fat_percentage": "Float64",  
"water_intake_l": "Float64",  
"workout_frequency_days_per_week": "Int64",  
"experience_level": "Int64",  
"bmi": "Float64"
```

Interpreting Measurement Types

```
"age": "continuous",  
"gender": "nominal",  
"weight_kg": "continuous",  
"height_m": "continuous",  
"max_bpm": "discrete",  
"avg_bpm": "discrete",  
"resting_bpm": "discrete",  
"session_duration_hours": "continuous",  
"calories_burned": "continuous",  
"workout_type": "categorical",  
"fat_percentage": "continuous",  
"water_intake_l": "continuous",  
"workout_frequency_days_per_week": "ratio",  
"experience_level": "ordinal",  
"bmi": "continuous"
```

Augmented Dimensions**bmi_encoded**

BMI can be utilized to classify customers based on their health status. To have BMI in an interpretable format, a new dimension is created with BMI transformed according to following classification:

| BMI | Classification |
|-------------|----------------|
| < 18,5 | Underweight |
| 18,6 – 24,9 | Normal |
| 25 – 29,9 | Overweight |
| 30 – 39,9 | Obesity |
| 40 - | Morbid obesity |

Source of classification:

<https://www.ibihealthcare.com/bariatric/bmi-of-morbid-obesity/>

This new dimension “bmi_encoded” is added to the dataset schema as a categorical value.

fat_percentage_encoded

BMI is a useful indicator of health for the majority of people, but in some cases muscle mass gets classified as obesity. To avoid this misclassification, a new dimension is created with fat percentage transformed according to following classification:

| Gender | Age | Fat percentage | Classification |
|--------|---------|----------------|----------------|
| Female | 18 – 39 | 0 – 21,9 | Underfat |
| | | 22 – 33,9 | Healthy |
| | | 34 – 39,9 | Overfat |
| | | 40 – | Obese |
| Female | 40 – 59 | 0 – 23,9 | Underfat |
| | | 24 – 34,9 | Healthy |
| | | 35 – 40,9 | Overfat |
| | | 41 – | Obese |
| Male | 18 – 39 | 0 – 7,9 | Underfat |
| | | 8 – 20,9 | Healthy |
| | | 21 – 25,9 | Overfat |
| | | 26 – | Obese |
| Male | 40 – 59 | 0 – 10,9 | Underfat |
| | | 11 – 22,9 | Healthy |
| | | 23 – 27,9 | Overfat |
| | | 28 – | Obese |

Source of classification:

<https://tanita.eu/blog/healthy-body-fat-percentage>

| Removing Unnecessary Dimensions | <p>This new dimension “fat_percentage_encoded” is added to the dataset schema as a categorical value.</p> <p>experience_level_encoded</p> <p>To enhance readability of experience level dimension, it is encoded accordingly:</p> <table data-bbox="384 517 796 663"><tr><th>Experience level</th><th>Encoding</th></tr><tr><td>1</td><td>beginner</td></tr><tr><td>2</td><td>regular</td></tr><tr><td>3</td><td>expert</td></tr></table> <p>This new dimension “experience_level_encoded” is added to the dataset schema as a categorical value.</p> <p>hours_spent_at_gym_weekly</p> <p>Time spent at gym weekly can be calculated multiplication of dimensions workout_frequency_days_per_week and session_duration_hours.</p> <p>This new dimension “hours_spent_at_gym_weekly” is added to the dataset schema as a continuous value.</p> <p>age_encoded</p> <p>To ease the analysis, a new dimension ‘age_encoded’ is added with brackets: 18 – 29, 30 – 39, 40 – 49 and 50 – 59.</p> <p>session_duration_hours_encoded</p> <p>To ease the analysis, a new dimension ‘session_duration_hours_encoded’ is added with quartiles as brackets.</p> <p>hours_spent_at_gym_weekly_encoded</p> <p>To ease the analysis, a new dimension ‘hours_spent_at_gym_weekly_encoded’ is added with quartiles as brackets.</p> <p>These dimensions are considered unnecessary for the objectives of this analysis and are dropped from the production data:</p> <ul style="list-style-type: none">• weight_kg• height_m• max_bpm• avg_bpm• resting_bpm• calories_burned | Experience level | Encoding | 1 | beginner | 2 | regular | 3 | expert |
|--|---|------------------|----------|---|----------|---|---------|---|--------|
| Experience level | Encoding | | | | | | | | |
| 1 | beginner | | | | | | | | |
| 2 | regular | | | | | | | | |
| 3 | expert | | | | | | | | |

| | |
|--|---|
| | <ul style="list-style-type: none"> • water_intake_l • bmi • fat_percentage |
|--|---|

| Recognizing attributes of significance | |
|--|--|
| Methodology | Dataset is explored to recognize attributes of significance with file 'GymMember_CustomerAnalysis_statistical_filtering.ipynb'. |
| Intuition | Intuitively differences should appear in combinations of age, gender, experience and BMI. |
| Average Customer | The profile of an average customer is created by filtering out the mean, median and mode values of the whole dataset. |
| Single Dimension Deviations | <p>Distribution of each dimension is observed to recognize values that might be useful for customer segmentation purposes.</p> <p>The following observations are made:</p> <ul style="list-style-type: none"> • 'workout_frequency_days_per_week' has the least amount of observations on members working out 2 days (20.25%) or 5 days (10.48%) per week. • 'bmi_encoded' and 'fat_percentage_encoded' has the least amount of observations on members whoa are underweight or obese. • 'experience_level_encoded' has the least amount of observations on 'expert' members (19.63%). |
| Filtered Dimension Observations | <p>The following observations are made when filtering dimensions into subgroups by age:</p> <ul style="list-style-type: none"> • Largest number of beginners (41.56%) is in the age group 50-59 • Largest number of experts (22.04%) is in the age group 40-49 • Largest difference in the most popular and the least popular workout type is in the age group 30-39 with cardio (31.1%) and yoga (19.62%) • Smallest number (16.73%) of 2 days per week members is in the age group 40-49 • Largest number (23.44%) of 2 days per week members is in the age group 30-39 • Largest number (25.71%) of members in the quartile with the longest session period is in the age group 40-49 |

- Largest number (27.27%) of members in the quartile with the shortest session period is in the age group 30-39
- Largest number of members with obesity (20.25%) or morbid obesity (4.55%) based on BMI, is in the age group 50-59
- Smallest number of members with obesity (10.29%) and largest number with underweight (21.81%) based on BMI, is in the age group 40-49

The following observations are made when filtering dimensions into subgroups by **gender**:

- Based on BMI, morbid obesity is more prevalent (6.75%) among male members (0% among females)
- Based on BMI, obesity is more prevalent (26.59%) among male members (5.24% among females)
- Based on BMI, normal weight was more prevalent (47.6%) among female members (28.97% among males)
- Based on fat percentage, obesity is more prevalent (25.44%) among male members (0% among females)
- Based on fat percentage, being overfat is more prevalent (39.05%) among male members (3.91% among females)
- Healthy fat percentage is more prevalent (76.52%) among female members (34.53% among males)
- Based on fat percentage, being underfat is more prevalent (19.57%) among female members (1.18% among males)

The following observations are made when filtering dimensions into subgroups by **experience level**:

- The popularity of workout types among beginners (from most to least popular) is 1. Cardio 2. Strength 3. HIIT 4. Yoga
- The popularity of workout types among experts (from most to least popular) is 1. Yoga 2. HIIT 3. Strength 4. Cardio
- The popularity of workout types among regulars (from most to least popular) is 1. Strength 2. Cardio 3. Yoga 4. HIIT
- The most popular workout frequencies by experience level are beginner 2 days per week, regular 4 days per week, expert 5 days per week\
- The most popular session durations by experience level are beginner 0.5-1.04 hours per session, regular 1.26-1.46 hours per session, expert 1.46-2.00 hours per session
- Based on fat percentage the most similar subgroups are beginners and regulars, both consisting of healthy, overfat or obese members