Lifestyle Analytics and Metrics System
A Data Warehousing and Business Intelligence Project

1. Background and Topic Overview

In recent years, there has been a growing emphasis on data-driven health and wellness decision-making. Individuals, fitness professionals, and wellness platforms increasingly rely on data to understand how daily lifestyle choices such as physical activity, diet, and hydration affect overall health outcomes.

This project focuses on the domain of Lifestyle Analytics, which combines behavioral, fitness and nutritional data to analyze patterns that influence health and fitness. From a Data Warehousing and Business Intelligence (DWBI) perspective, lifestyle data presents an ideal use case. It is multi-dimensional, metric-intensive, and requires significant transformation before it can be used effectively for analysis. This project leverages those characteristics to design an end-to-end BI solution that transforms raw lifestyle data into meaningful, decision-ready insights.

2. Purpose of the Analysis

Choosing the Right Lifestyle Strategy is hard. Individuals often struggle to answer practical, goal-oriented questions such as:
Which workouts actually burn the most calories for my fitness goal?
Which diet aligns best with muscle gain or fat loss?
Am I eating too much or burning too little?
Do my age, activity level, and experience change what works best for me?

While lifestyle data is widely available, it is usually fragmented across workouts, nutrition, and health metrics, making it difficult to translate raw data into actionable decisions.

Core Problem Statement

The dashboard addresses the problem of translating complex lifestyle, fitness, and nutrition data into personalized, goal-driven decisions for individuals. It helps users identify the right combination of workouts and diet plans based on their fitness goals, demographics, and behavioral patterns, rather than relying on generic fitness advice.

3. Proposed Target Audience of this project

The analysis and dashboards developed in this project provide value to a wide range of stakeholders. For **individuals and fitness enthusiasts**, the insights enable a clearer understanding of how workout routines, dietary choices, and daily behaviors influence overall health and wellness. **Health and wellness platforms** can leverage similar analytical frameworks to deliver personalized recommendations, enhance user engagement, and improve the effectiveness of their services. Additionally, **fitness trainers and nutrition coaches** can utilize these insights to identify effective workout strategies and dietary plans tailored to different user segments, thereby supporting more informed and data-driven coaching decisions.

4. Key Features Of The Project

A key feature of this project is the use of Databricks Workflows to automate the end-to-end data processing pipeline. Multiple notebooks responsible for data ingestion, cleaning, transformation, and feature engineering are integrated into a single workflow. This workflow is scheduled to run automatically, ensuring that the dataset remains updated without manual intervention. By orchestrating notebook execution in a defined sequence, Databricks Workflows improve reliability, reproducibility, and scalability of the analytics process, aligning with industry-standard data engineering practices.

Another major feature is the Diet Type Drill-Down Dashboard, which enables detailed, interactive analysis of nutritional patterns. Users can begin with a high-level view of overall nutrition trends and then drill through specific diet types such as Vegetarian, Keto, or Vegan. Within each diet type, the dashboard provides insights into nutrient composition, cooking methods, calorie intake, ratings, and preferred fitness goals. This drill-down capability allows users to explore granular details while maintaining contextual continuity, making the dashboard intuitive and effective for comparative analysis and informed decision-making.

5. Dataset And Its Significance

The dataset selected for this project is the **"Life Style Data"** dataset from Kaggle, curated by **Omar Essa**.

Dataset link:
https://www.kaggle.com/datasets/jockeroika/life-style-data

This dataset was chosen for several key reasons:
Rich Multi-Dimensional Structure
The dataset includes demographic, fitness, nutrition, and biometric attributes, making it well-suited for dimensional analysis.

BI-Friendly Metrics
Measures such as calories burned, session duration, heart rate, nutrient intake, and health scores which are extremely beneficial in impactful KPI creation.

ETL Opportunities
The raw dataset requires cleansing, normalization, and metric derivation which is ideal for applying data warehousing and ETL concepts.

6. Dataset Description

The original Kaggle dataset contains simulated yet realistic lifestyle and health-related data at an individual level. Each record represents a user profile with attributes capturing daily behaviors and health indicators.

Key Data Categories:

Demographic Attributes
Age, Gender

Fitness and Activity Attributes

Workout type (Cardio, HIIT, Strength, Yoga), Workout frequency (days per week), Session duration (hours), Calories burned, Calories burned per minute, Target muscle groups, Difficulty level (Beginner, Intermediate, Advanced)

Nutrition and Diet Attributes
Diet type (Balanced, Keto, Low-Carb, Vegan, Vegetarian, Paleo), Meal type, Calories consumed, Macronutrients (proteins, fats, carbohydrates), Sugar, sodium, cholesterol, Cooking method and cooking time, Water intake.

Health and Biometric Attributes
BMI
Resting, average, and maximum heart rate

7. Everything About ETL

ETL Development

As part of this initiative, a refined Extract, Transform, and Load (ETL) process has been developed using Databricks, based on Medallion Architecture principles, to better utilize lifestyle, nutritional, and fitness information. The focus of this ETL is on scalability, quality, and a clean separation of concerns in processing. The entire processing is achieved by using Databricks Workflows.

ETL Processing

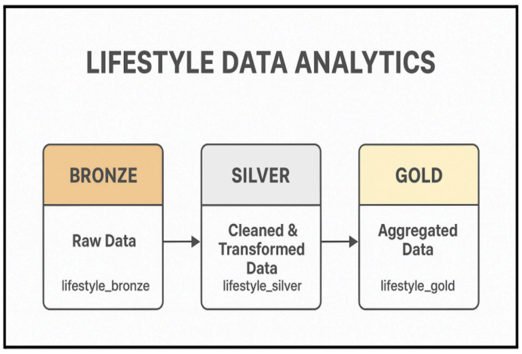
The ETL processing begins with schema creation on Databricks for project asset organization. A raw CSV file with lifestyle, workout, and dietary data is loaded into the system and registered inside the Bronze level. The purpose of this level is to stage data in a format without applying any transformations. Staging this data in this manner incorporates elements of being a reliable source of truth, facilitating auditing, reprocessing, and debugging, which become very important in contemporary data warehousing implementations.

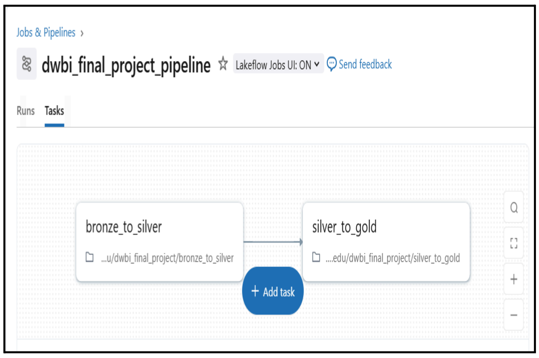
The pipeline further progresses from the Bronze level to the Silver level, where data cleaning, standardization, and processing are accomplished with the application of PySpark. Some new variables have been introduced to improve analysis interpretability. Such variables include categoricals such as **weight_category** formed from BMI, **activity_level** based on workout frequency, **experience_category** deducted from experience level, **fitness_goal** based on caloric balance, and **age_range** obtained from age segmentation. The new variables transform basic numeric columns into relevant business dimensions, making it easier to work with in analysis and generating reports.

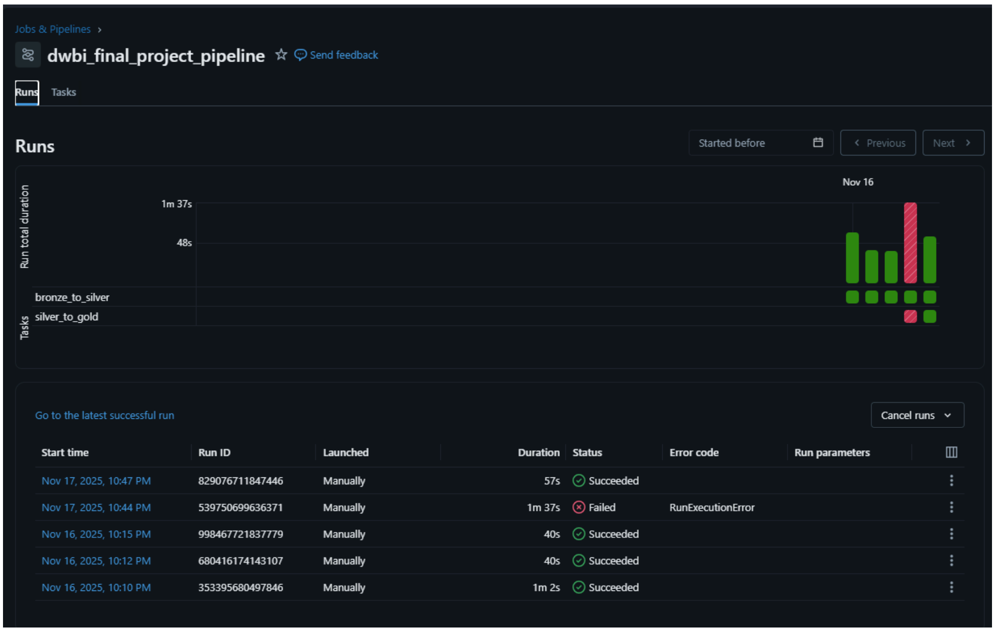
Apart from this, The **silver layer schema** preserves a very tight standardization for all columns in this level. The names of all columns are changed to lower case, spaces are replaced with underscores, and special characters are removed. With this standardization, it is possible to achieve uniformity in schema and simplicity in queries, and it will be compatible with other SQL-based analysis tools and BI solutions. The resulting data is further saved in a **Delta table** in Silver schema with a view established on this table.

The last step of this pipeline is used for transferring the data from the Silver level to the Gold level, which is optimized for reporting purposes. In this level, analytically relevant attributes are chosen, numeric columns are truncated to a standard format, and business-level filtering takes place where all entries with inaccurate information are filtered out, such as people with illogical age entries. The resulting Gold table holds a consumption-ready dataset of all nutritional information, workout performance, physiological characteristics, and derived lifestyle facts in one table. Based on this table, Power BI dashboards are produced.

The **Databricks Workflow** below shows in detail the end-to-end processing of such a pipeline with two dependencies: bronze_to_silver and silver_to_gold. Such a workflow ensures a proper order of execution of these tasks. Moreover, all dependencies will be adhered to in this manner, reflecting a proper standard of processing in this industry or rather in this field. In summary, this ETL stage highlights a good application of Medallion Architecture in a cloud analytical platform. Through continuous refinement of data from the Bronze, Silver, and Gold levels, this project safeguards the quality of information, analytical usability, and scalability of analysis.




Jobs & Pipelines ›
dwbi_final_project_pipeline
Lakeflow Jobs UI: ON
Send feedback
Runs
Tasks
bronze_to_silver
...u/dwbi_final_project/bronze_to_silver
silver_to_gold
....edu/dwbi_final_project/silver_to_gold
+ Add task



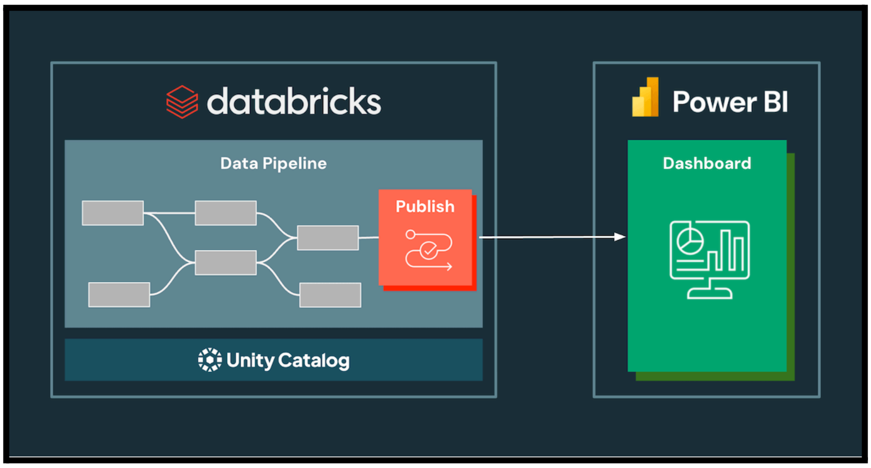
Runs

Start time	Run ID	Launched	Duration	Status	Error code	Run parameters
Nov 17, 2025, 10:47 PM	829076711847446	Manually	57s	✓ Succeeded		
Nov 17, 2025, 10:44 PM	539750699636371	Manually	1m 37s	☒ Failed	RunExecutionError	
Nov 16, 2025, 10:15 PM	998467721837779	Manually	40s	✓ Succeeded		
Nov 16, 2025, 10:12 PM	680416174143107	Manually	40s	✓ Succeeded		
Nov 16, 2025, 10:10 PM	353395680497846	Manually	1m 2s	✓ Succeeded		

8. Establishing Connection Between Databricks and Power BI

To facilitate the creation of analytical models and dashboard solutions, Databricks was thus integrated with another important technology called Power BI. The role of Databricks in this case was to act as a processing and transformation stage where all the relevant fitness and nutrition datasets were cleaned and optimized using PySpark in a large scale manner. After processing the datasets, Power BI was then linked to Databricks.

The connection was established using the Databricks connector available in Power BI. This connector allows Power BI to directly query Databricks SQL warehouses or clusters using secure authentication methods such as personal access tokens. With this integration, Power BI can access datasets already processed in Databricks without requiring replication of datasets, thus maintaining uniformity in both layers.



9. Analytical Approach

The analysis focuses on examining the relationship between lifestyle behaviors and fitness and wellness outcomes using a data-driven approach. Key analytical questions include identifying which workout types result in the highest calorie expenditure, how workout intensity and performance vary across different age groups, and how various diet types influence calorie intake and overall health-related indicators. Additionally, the analysis explores lifestyle patterns associated with specific fitness objectives, such as muscle gain versus fat loss (cutting).

To address these questions, relevant metrics were aggregated and analyzed across multiple dimensions, including age, workout type, diet type, and fitness goals. This approach enables effective comparison, trend identification, and drill-down exploration, allowing stakeholders to uncover meaningful insights and patterns within the data.

10. Additional Metrics/Measures And Their Purpose

In addition to the transformed dataset, a set of derived measures was implemented in Power BI to enrich the analytical layer. These measures were designed to quantify efficiency, nutritional balance, and workout effectiveness, thereby supporting deeper exploratory and comparative analysis.

Calorie Efficiency Ratio

This metric helps compare the total calories burned through workouts to the total calories consumed from food. The Calories Burned:Calories ratio was developed by taking the sum of Calories Burned and dividing it by the sum of Calories, helping to determine how well a particular diet-workout combination supports calorie expenditure.

Diet Efficiency Index (DEI)

Diet Efficiency Index is a measure of quality regarding the macronutrient composition of diets. It applies weighted importance to proteins, carbohydrates, and fats, then normalizes the result by total calorie intake. This helps in the identification of diets that provide better nutritional value per calorie.

Muscle Engagement

This is a count of how many muscle groups were hit in a workout. This was derived by simply counting the target muscle group field and is useful in indicating variety in workouts, but more importantly, overall coverage of muscles.

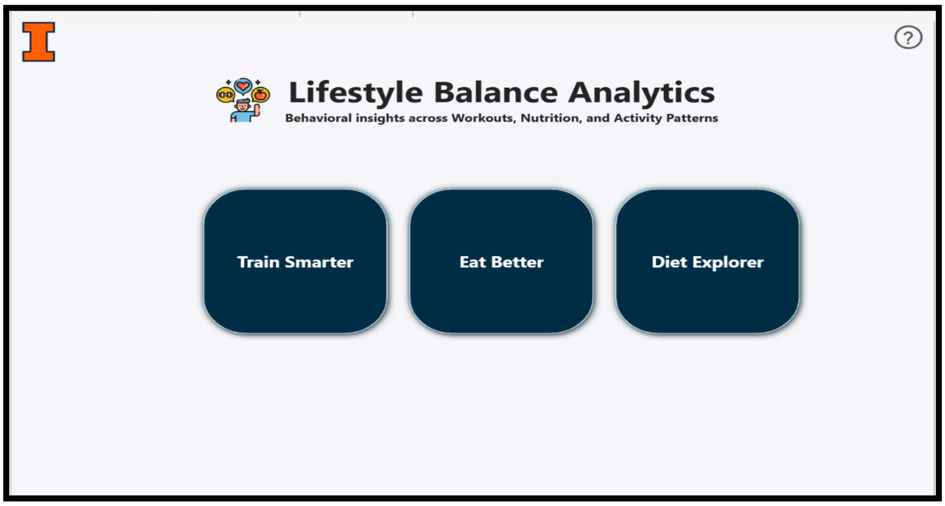
Nutrient Value

This dynamic measure shows the average value of a selected nutrient-sugar, protein, sodium, or fats, dependent upon user interaction through the axis selection. It was created using a SWITCH statement that changes the displayed nutrient based on which axis is selected, thus allowing for interactive comparison of nutrients across the types of diets.

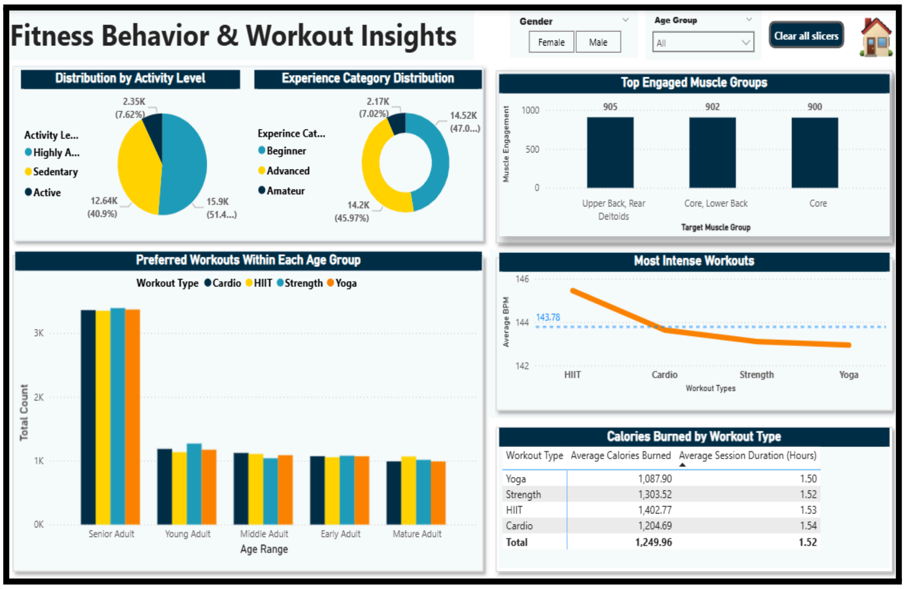
11. Dashboard Overview

Each dashboard is designed to answer specific analytical questions while maintaining a consistent visual language. The Power BI solution consists of four main analytical views:

Home / Navigation Dashboard


Lifestyle Balance Analytics
Behavioral insights across Workouts, Nutrition, and Activity Patterns
Train Smarter
Eat Better
Diet Explorer

Fitness Behavior & Workout Insights



Workout Type	Average Calories Burned	Average Session Duration (Hours)
Yoga	1,087.90	1.50
Strength	1,303.52	1.52
HIIT	1,402.77	1.53
Cardio	1,204.69	1.54
Total	**1,249.96**	**1.52**

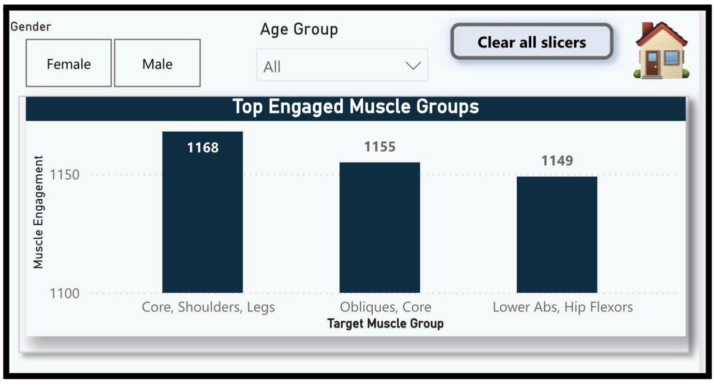
What questions does this dashboard answer?

This dashboard focuses on workout patterns, intensity, and engagement across different user groups. It helps answer questions such as:

- How are users distributed across different activity levels (Sedentary, Active, Highly Active)?
- What is the overall experience level composition of users (Beginner, Amateur, Advanced)?
- Which muscle groups are most frequently targeted during workouts?
- How do workout preferences (Cardio, HIIT, Strength, Yoga) vary across different age groups?
- Which workout types are the most intense, based on average heart rate?
- Which workout types result in the highest calorie burn and longest average session durations?
- How do workout behaviors differ when filtered by gender or age group?

Overall, this dashboard enables stakeholders to understand exercise habits, intensity levels, and workout effectiveness across demographics.

Top Engaged Muscle Groups



Visualization Type: **Bar Chart**, used to compare **engagement levels** across different target muscle groups with a **dynamically changing X axis.**

Metrics Analyzed:
Muscle Engagement Score aggregated by **Target Muscle Group**.
Values represent total engagement across all recorded workout sessions.

What the Visualization Shows:
The most engaged muscle groups are:
Core, Shoulders, Legs
Obliques, Core
Lower Abs, Hip Flexors
These groups show the highest cumulative engagement values.

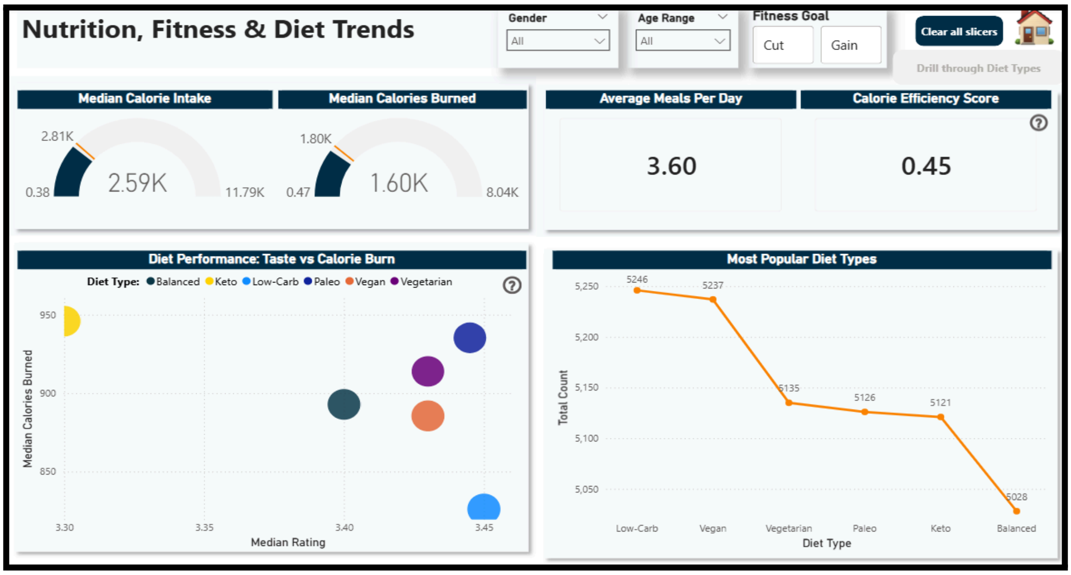
Insights Gained:

- Compound muscle groups involving the **core and lower body** dominate workout engagement.

- This indicates a strong emphasis on **functional and stability-based training** within the population.

- Programs targeting multi-muscle engagement appear more prevalent and effective across workout types.

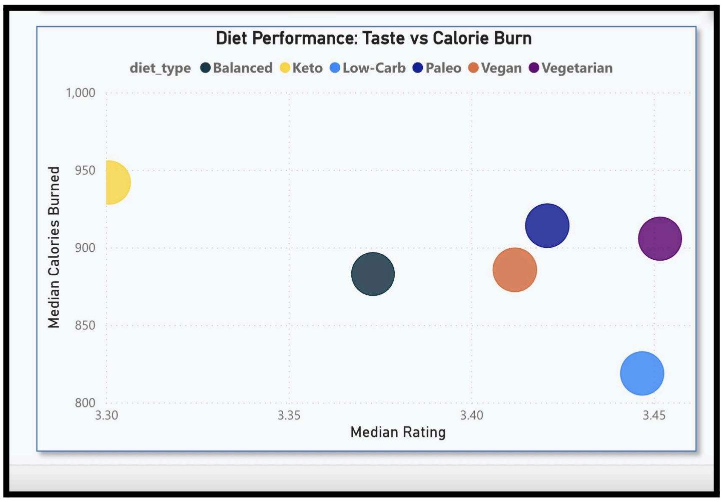
Nutrition, Fitness & Diet Trends



What questions does this dashboard answer?

- What is the typical calorie balance of users?
- How do median calorie intake and median calories burned compare across the population?
- What is the average number of meals per day, and how does it vary by demographic filters?
- Which diet types provide better calorie burn outcomes relative to intake (Calorie Efficiency Score)?
- Which diet types are most popular among users?
- How do different diet types trade off between taste and performance?

Diet Performance: Taste vs Calorie Burn



Visualization Type: Bubble Scatter Plot, used to analyze trade-offs between two continuous metrics, with bubble size representing relative magnitude.

Metrics Analyzed:

X-axis: Median Rating (user satisfaction / taste)
Y-axis: Median Calories Burned
Bubble Size: Relative volume or impact per diet type
Dimension: Diet Type

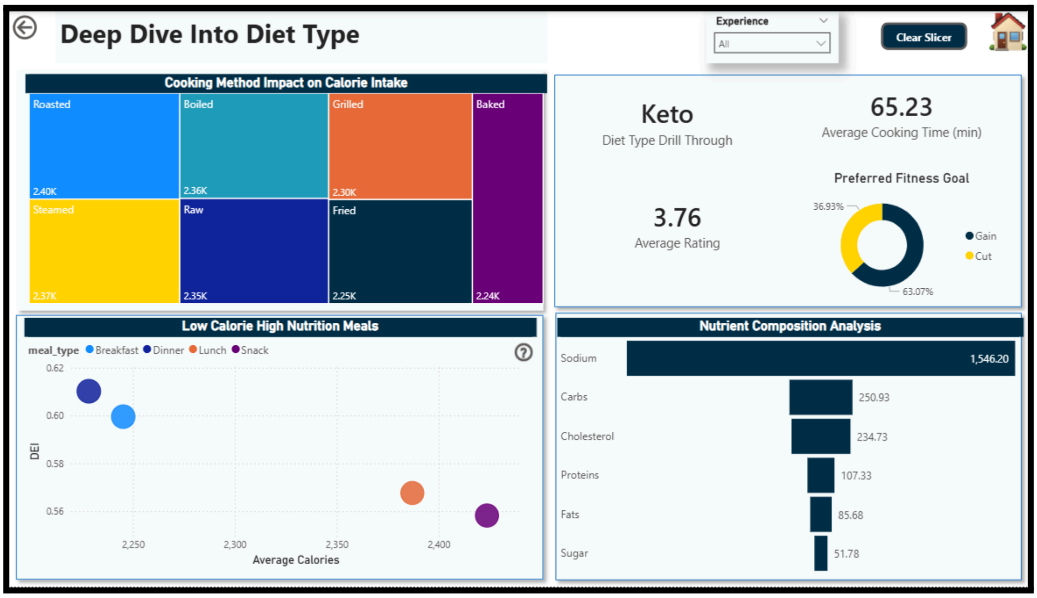
What the Visualization Shows:

- Diets differ significantly in how they balance user satisfaction and calorie burn.
- Some diets achieve higher ratings but lower calorie burn, while others show the opposite pattern.

Insights Gained:

- No single diet dominates both taste and performance, indicating trade-offs between enjoyment and fitness efficiency.
- Balanced and Paleo diets show relatively strong calorie burn with moderate satisfaction.
- This visualization supports informed dietary decision-making by highlighting which diets best align with specific fitness goals.

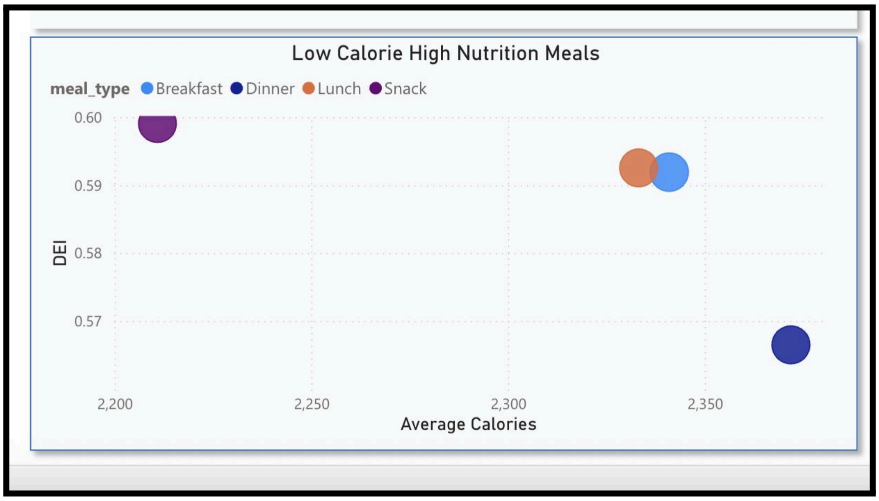
Deep Dive into Diet Type (Drill-Through)



What questions does this dashboard answer?

- How does a specific diet type perform when analyzed in isolation?
- What cooking methods contribute most to calorie intake within this diet?
- Which meals provide the best nutrition at lower calorie levels for this diet?
- What is the nutritional composition of this diet?
- What fitness goals do users following this diet primarily pursue?
- How demanding is this diet in terms of preparation time?

Low-Calorie, High-Nutrition Meals



Visualization Type: **Bubble Scatter Plot**, used to analyze the relationship between calorie intake and nutritional efficiency across meal types..

Metrics Analyzed:
X-axis: Average Calories
Y-axis: Dietary Efficiency Index (DEI)
Bubble Color: Meal Type (Breakfast, Lunch, Dinner, Snack)

What the Visualization Shows:

- Snacks and breakfasts achieve **higher nutritional efficiency at lower calorie levels**.
- Dinner meals tend to have higher calorie values with comparatively lower DEI scores.

Insights Gained:

- **Snacks emerge as the most calorie-efficient meal type**, delivering high nutrition per calorie.
- Breakfast and lunch provide a balanced trade-off between energy intake and nutrition.
- This visualization helps identify **optimal meal types for weight management and nutrition-focused planning**.

Prototype: How a 24-Year-Old Who Wants to Gain Muscle Benefits

A case can be taken, for example, of a 24-year-old person who wants to work on overall health and fitness along with a structured lifestyle approach. The user starts with very little knowledge about how their current health and fitness patterns along with their lifestyle approach affect their health outcomes. The dashboard system works as an analytical tool for self-learning and decision-making.

A user interacts with this system by starting from its homepage, also serving as an interface for navigation. From this point, the user selects his or her “Fitness Behavior and Workout Insights.” They also use age filters to select data for “Young Adulthood,” allowing them to compare data related to this age group.

This step-by-step analysis offers an opportunity to recognize missing elements in his or her current exercise routine, whether it lacks strength training or fails to target certain muscles. While working out or participating in certain strength exercises, they can compare data related to “calories burned,” “intensity,” “activity level,” “patterns,” or “muscle group.”

Once the behaviors related to physical training are grasped, the user moves on to evaluate diet habits from the ‘Nutrition, Fitness, and Diet Trends’ dashboard. In this module, calorie and calorie burn levels are compared by analyzing median calorie data in order to assess the balance that needs to be created by training and dieting for achieving a certain fitness goal related to cutting and muscle gain and/or weight loss and weight management by analyzing diet types based on health metrics and ‘health score and rating.’

For a deeper form of personalization, the user navigates to the Deep Dive into Diet Type dashboard from here. This level of personalization also facilitates the exploration of a diet type of their choice, whether vegetarian or high protein dieting. The user will assess the cooking methods and their influence on calorie needs, analyze the nutritional efficiency of foods being eaten by considering the foods with the highest nutritional efficiency, and view the nutritional composition of foods in aspects such as protein, fats, sodium, and sugar content. By taking this analytical walkthrough process, all of these dashboards ultimately lead the user through exploratory analysis into insight-driven actions. The learner can better understand how to apply optimization strategies for workouts, adjust nutritional consumption, and modify habits accordingly in relation to overall fitness objectives. This prototype showcases how the entire dashboard system serves as a decision support system by providing valuable lifestyle information as data-driven recommendations for young individuals.

Difficulties Faced During the Project

One of the primary challenges encountered during this project was the design and derivation of meaningful analytical metrics. Although the dataset contained a wide range of raw attributes, it did not directly offer high-level measures that could easily translate into actionable insights. As a result, considerable effort was required during the ETL phase to conceptualize, engineer, and validate derived metrics such as calorie efficiency, diet efficiency index, and muscle engagement scores. Developing these measures involved careful judgment to ensure they were both analytically useful and statistically reliable, making metric design a non-trivial component of the transformation pipeline.

Another limitation of the dataset was the absence of geographical attributes, which restricted the scope of spatial analysis. This prevented the use of geographic visualizations such as maps or region-based comparisons. Consequently, the analysis relied more heavily on demographic, behavioral, and categorical dimensions—including age groups, workout types, and diet categories. To compensate, the dashboards were designed with greater emphasis on comparative views and drill-through analysis to preserve analytical depth.

The nature of health and lifestyle data also introduced challenges. Many numerical variables, such as calorie intake, heart rate, and health scores, exhibited relatively small variations across users. However, in health analytics, even minor differences can have meaningful implications. This made visual scaling and normalization particularly sensitive, as poorly chosen scales could either exaggerate trends or obscure important patterns. Careful calibration of axes, aggregation levels, and visualization types was necessary to ensure clarity without misrepresenting the data.

From a data engineering perspective, maintaining schema consistency and standardization throughout the ETL pipeline required additional effort. Inconsistencies in column naming, mixed data formats, and the introduction of derived categorical fields had to be resolved to support accurate BI modeling and avoid semantic confusion within Power BI.

Finally, aligning data warehousing best practices with business interpretability was an ongoing consideration. Transformations and aggregations needed to be technically sound while remaining intuitive for end users. This required iterative refinement of both the semantic model and the dashboards to ensure that insights were not only correct, but also clear, interpretable, and actionable.

Key Findings

- The majority of users falling into **Highly Active** or **sedentary segments shows inconsistency across populations.**
- **HIIT workouts** consistently emerge as the **most intense** and **time-efficient**, delivering the highest average heart rate and calorie burn without longer session durations.
- **Workout preferences vary by age**, with younger users favoring high-intensity workouts (HIIT, Strength) and older users shifting toward more moderate routines.
- **Low-Carb and Vegan diets are the most popular**, while balanced diets show lower adoption, suggesting users prefer structured diet plans.
- No single diet excels in both **user satisfaction (taste)** and **calorie burn**, highlighting clear trade-offs between enjoyment and fitness effectiveness.
- High Calorie Burn with Low User Rating: **Keto**
- Moderate Calorie Burn with High User Rating: **Paleo**

Feedback from Users

We approached the ticket master group which consisted of members in the 24–25 age group. They found the dashboard intuitive and highly relevant to their fitness and lifestyle goals. The clear separation between workout, nutrition, and diet-specific insights made it easy to explore information without feeling overwhelmed. The interactive filters and drill-down features helped them compare diet types and workout patterns effectively, and the KPIs were simple yet informative for everyday decision-making.

However, they suggested the dashboard could be further improved by adding brief descriptions or tooltips explaining certain metrics for first-time users which we later added to the dashboard.Additionally, trend-based visuals showing progress over time could enhance engagement and help users track long-term improvements.

Conclusion

In conclusion, the project successfully demonstrates that an integrated lifestyle dataset can be turned into useful information by using an analytics stack. The dashboards created in this project show how exercise, nutrients, and fitness targets are interconnected, and overall, this solution helps in making decisions and is applicable for a fitness and health-related use case.