# **E6 Site Analysis**

## **Boutique Fitness Site Selection: Data Collection, Processing, and Modeling**

### **Executive Summary**

This project develops a data-driven framework for evaluating U.S. markets for boutique fitness studios, using OrangeTheory Fitness (OTF) as a proxy. Location data was extracted from the OTF API and enriched with demographic and economic indicators from the U.S. Census Bureau. Drive-time isochrones were generated to define each studio’s trade area, and population-weighted metrics (income, population size, healthcare workforce) were computed. A scoring model was designed to capture market potential, incorporating non-linear effects of household income, population density, and occupational composition. The processed data was imported into SQL for management and then visualized in Tableau, enabling both site-level and city-level insights. The result is an interactive, replicable tool for identifying promising markets and prioritizing growth opportunities in the boutique fitness industry.

## **1. Data Collection**

### 1.1 OrangeTheory Locations

* Accessed the **OrangeTheory Fitness API** to extract all U.S. studio locations.
* Collected data included: id, name, status, address, city, state, postal\_code, latitude, longitude, and phone.
* Exported the results into a CSV for downstream processing.

### 1.2 Demographic Data

* Pulled U.S. Census Bureau **American Community Survey (ACS)** data via the Census API.
* Collected tract-level variables:
  + Total population (B01003\_001E)
  + Median household income (B19013\_001E)
  + Occupation counts (S2401) to capture healthcare-related jobs (practitioners, technologists).

## **2. Data Cleaning & Preparation**

### 2.1 Studio Data

* Removed unnecessary columns (phone, address, etc.) to simplify ingestion.
* Converted latitude/longitude fields to DECIMAL(10,7) to prevent truncation during SQL import.
* Assigned id as the **primary key** for studios.

### 2.2 Demographics

* Linked ACS tract-level data with each studio’s **trade area** using spatial joins (see Section 3).
* Computed percentage of medical professionals
* Fixed scaling errors to ensure percentages were between 0 and 1.

### 2.3 Handling Issues

* Phone numbers caused import errors → dropped.
* Some income values in ACS were **negative (missing codes)**; these were cleaned or replaced with NULL.

## **3. Isochrone Construction**

* Used a routing service (8-minute drive-time polygons) to approximate each studio’s trade area.
* Intersected **Census tracts** with each isochrone to allocate demographic values.
* Calculated **population-weighted averages**
* Exported results as isochrone\_demographics.csv.

## **4. SQL Database Design**

* Created normalized tables:
  + otf\_with\_demographics → studio-level data + demographics.
  + isochrone\_demographics → drive-time trade area metrics.
* id used as **primary key** in otf\_with\_demographics.
* iso\_id (studio id + \_8min) linked back to each studio.

## **5. Modeling Market Potential**

### 5.1 Key Factors

1. **Population Scale** – larger populations increase participation.
2. **Household Income (non-linear)** – participation rises with income but saturates.
   * Anchored at:
     + $60k → ~8% participation
     + $150k → ~25% participation
   * Modeled with a **logistic curve** where coefficients were chosen to hit anchor points.
3. **Healthcare Workforce Concentration** – proxy for health-conscious community.

### 5.2 Scoring Model

* Combined factors into a composite **site score**

## **6. Tableau Visualization**

* Imported SQL outputs into **Tableau** for exploration.
* Built dashboards at both **site-level (studio IDs)** and **city-level (aggregated)**.
* Features included:
  + Map visualizations of trade areas.
  + Bar charts of **top scoring markets**.
  + Filters for **state**, **income range**, and **population size**.
* Handled duplicate city names by allowing **multiple rows per city** (each representing a distinct studio ID).

## **7. Deliverables**

* **SQL Database**: cleaned and normalized datasets.
* **Excel exports**: cross-checked intermediate outputs.
* **Tableau Dashboard**: interactive scoring tool.
* **Documentation (this report)**: full pipeline description.