İstanbul
Bilgi University

# IE 421

# Data Science Term Project - Proposal

# Data and The City

**By**

Mehmet Fatih Çetinkaya 121205031

Mehmet Tolga Çakan 121203029

Demet İrem Yılmaz 121203095

Sıla Kahya 123203018

Eylül Balcı 122203020

https://github.com/BILGI-IE-421/ie421-2025-2026-1-termproject-data-and-the-city

## Research Questions

**1st Question:**

To what extent has the gender participation gap converged towards parity from 1896 to 2024, and, quantitatively, what is the deviation from a perfect 50/50 split in the sport disciplines of the Paris 2024 Games?

**Description:**

In this project, we investigate the longitudinal evolution of gender representation in the Olympic Games. While general trends suggest an increase in female participation, we aim to quantify the exact rate of convergence over 120 years. Specifically, we will use the Paris 2024 dataset the first Olympics to explicitly target full gender parity to measure statistical deviations across different sport disciplines. This descriptive analysis provides the foundational "data story" of our project, visualizing how the Olympics have transformed from a male dominated event into a balanced global system.

---

**2nd Question:**

How accurately can a Multiple Linear Regression model, trained on historical delegation sizes and performance metrics, predict the total medal counts for the unseen Paris 2024 Games, and what is the margin of error (RMSE) when applied to top performing nations?

**Description:**

We aim to model national performance not as a random occurrence but as a predictable outcome of input variables. We will train a Multiple Linear Regression model using historical data from the modern Olympic era (e.g., 1960–2016) to establish the relationship between a country's delegation size (resource allocation) and its medal success. Uniquely, we will use the Paris 2024 results as an "unseen" validation set to test our model's predictive power and calculate the Root Mean Square Error (RMSE), thereby evaluating the model's real world generalizability.

---

**3rd Question:**

To what extent do biometric features (Age, Height, Weight) serve as significant predictive indicators for medal success, and how effectively can a Classification model differentiate between medalists and non medalists in high physicality sports?

**Description:**

Moving from the macro (country) level to the micro (athlete) level, this question addresses the optimization of human performance. We will implement a Binary Classification Model (e.g., Logistic Regression) to determine if physical attributes are statistically significant predictors of winning a medal. By treating the outcome as a binary variable (Medal vs. No Medal), we aim to uncover hidden patterns in biometric data and assess the model's accuracy in distinguishing elite performers from the general participant pool. In practice, we plan to focus on high physicality sports (e.g., Athletics and Swimming) to obtain a more homogeneous and comparable athlete subset.

---

## Datasets

**1st & 3rd Questions:**

120 Years of Olympic History: Athletes and Results

Link:
https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results

Description:

This is our primary historical dataset containing over 270,000 athlete-event records from 1896 to 2016. It provides the essential baseline for analyzing the evolution of the gender gap (Q1) and serves as the training ground for our Classification model (Q3) by providing labeled data on athlete age, height, weight, sex, sport, and medal outcomes.

In practice, for some of the analyses and models we will focus on the more recent decades (e.g., post-2000) due to data completeness for biometric attributes.

---

**1st & 2nd Questions:**

Paris 2024 Olympic Summer Games

Link: https://www.kaggle.com/datasets/piterfm/paris-2024-olympic-summer-games

Description:

This dataset serves as our critical “Validation Set.” For Q1, it provides the target data to measure whether the 50/50 gender parity goal was achieved at the Paris 2024 Games and how this parity breaks down across different sports. For Q2, it acts as the “unseen future” to test the accuracy of our Regression model, allowing us to compare our predicted medal counts against the actual 2024 results for major medal winning nations.

---

**2nd Question:**

Tokyo 2020 Olympics

Link: https://www.kaggle.com/datasets/arjunprasadsarkhel/2021-olympics-in-tokyo

Description:

We use the Tokyo 2020 dataset as an intermediate test set to calibrate our Regression model before applying it to Paris 2024. It helps us understand recent trends in delegation sizes and medal distributions, ensuring that our model is tuned to the competitive landscape of modern Olympics rather than only historical patterns.