CS439: Intro to Data Science Final Project Report
Spring 2025

Decoded Gains: A Data-Driven Approach to Understanding Protein and Training Efficiency

By:
Derick Vega
Darren Bonjour
Burak Serik

Professor Chatevurdi

May 9, 2025

Github Repository: https://github.com/derickv12/git-pull-git-swole.git

Introduction:

For unsupervised learning projects:

What question are you trying to answer

Our basic intent was to check if we can apply unsupervised learning techniques that would end up producing valuable clusters of muscle-building outcomes on nutritional and training intervention variables. Instead of forecasting results, a more interesting question to ask is if different types of interventions exist in the data, and if some variable combinations like protein intake and frequency arise to actually yield higher gains in lean body mass (LBM). More specifically, we wanted to get an idea of what a theoretically highly efficient intervention would look like from both an outcome and nutritional load perspective.

How do you plan to answer it?

First, we clean a real world clinical trials dataset on resistance training interventions.To better assess the quality of each intervention, we will create a new metric called LBM gain per gram of protein intake to measure efficiency. The data is then filtered to include only variables that factor in the relationship we are trying to observe, such as training duration, frequency, energy intake, protein intake, and lean body mass (LBM) change. Based on these features, KMeans clustering will be applied to the study, using the silhouette score to decide on the optimal cluster number. For visualization purposes and to assess cluster isolation, we will conduct Principal Component Analysis (PCA) to reduce the dimensionality of the data by interpreting box plots and summaries for each cluster. These patterns will be analyzed to determine which intervention profiles are seemingly most efficient. This analysis will deepen our understanding in regards to what training and nutrition combinations work best for muscle growth, providing insights that can be leveraged for real world applications such as sports and health.

How does this approach relate to the lectures/papers we discussed?

Our project uses several of the core tools and techniques explored throughout the course. We performed data cleaning, transformation, and preprocessing using Python, Pandas, and NumPy skills essential for any data cleaning project. We used KMeans for clustering and silhouette scores to evaluate clustering quality.

We also reduced the dimension of the data with Principal Component Analysis (PCA) to reduce complexity for better interpretation of clusters. Finally, we chose plots that communicate trends clearly and avoid misleading impressions. The whole workflow is a real world implementation of what we were taught in class about building data-pipelines, i.e., turning raw inputs into useful insights. Most importantly, this gave us a demonstration of how to bring all those tools in one go towards the exploration of a problem, the creation of new metrics, and the drawing of useful conclusions.

Motivation:

Why is your project important?

This project is important because it uses actual clinical data about how various resistance training methods influence the efficiency of muscle gain and how protein intake contributes to it. Our work transcends all that by harnessing clustering to fish out patterns that people may not immediately see in the usual statistical summaries. It provides a data set for weighing intervention quality, both onto research and real-life applications.

Why are you excited about it?

This project resonated with each of us because health, exercise, and muscle-building are things we all cared very deeply about, from just someday competing, to enhancing simple day-to-day movement. Each one of us has personal training goals, with a few of us training regularly for strength, aesthetics, performance, or health. Some of us also dove into research, for instance, conducting personal research, reading extensively and also doing a lot of private research on muscle hypertrophy and nutrition long before the course started. The added thrill for us was merging this real-world passion with the technical tools imparted to computer science. It gave us the ability to work with real life unsupervised learning and data analysis pertaining to a subject we truly care about and draw conclusions that could lead to smarter training or nutrition decisions. The most rewarding part was seeing how CS knowledge can actually apply to real-world applications rather than being reserved solely for theoretical academic work.

What are some existing questions in the area?

Even after years of research, there’s still a lack of consensus on what truly defines optimal conditions for muscle growth. A major point of debate is protein intake, how much is actually needed to support hypertrophy. Even the commonly recommended grams of protein per lbs of body weight vary widely depending on the source, which is why it's typically presented as an approximate range rather than a fixed number. Furthermore, training frequency, duration, total calorie intake, rest, and other variables all interact with each other in complex, non-linear ways. Once we add genetics into the equation, differences in metabolism, muscle fiber composition, and recovery speed, it becomes very evident that no single approach exists that works for everyone. This variability is why projects like ours matter. Since accounting for every individual factor isn’t feasible, clustering helps us uncover broader patterns in the data and draw useful, general insights about what kinds of interventions tend to be more efficient for muscle gain across a range of body types and study conditions.

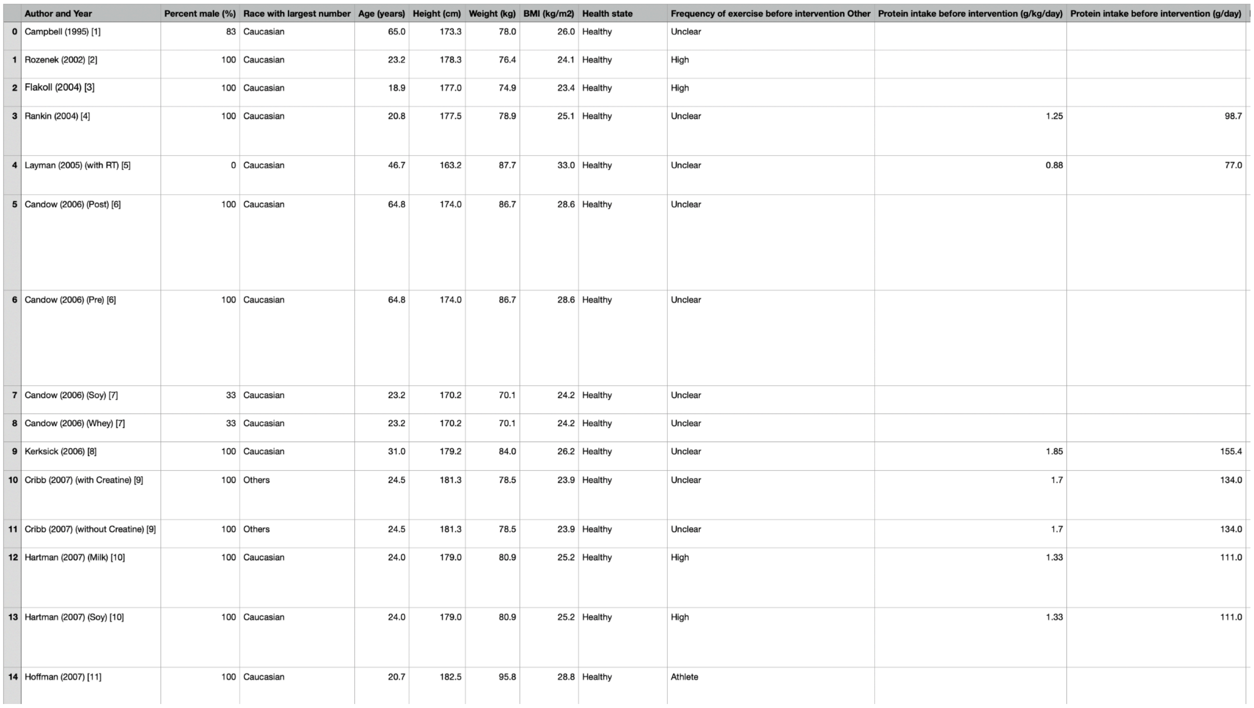
Are there any prior related works? Provide a brief summary.

Absolutely, this topic has been explored for decades, with a large and growing body of research focused on the relationship between resistance training, nutrition, and muscle hypertrophy. One of the earliest and most influential figures in the field was Arthur Jones, the founder of Nautilus. Through controlled experiments in tightly regulated training environments, he emphasized the importance of eccentric (negative) training in stimulating muscle growth, an insight that has since been validated and built upon in modern exercise science today. His findings not only reshaped training principles but also led to the creation of some of the most effective, biomechanics-driven gym equipment that is still widely used today. Since then, research has expanded across countless variables such as protein intake ranges, training frequency, volume, and recovery. However, many open questions remain. Our project adds to this evolving conversation by using modern clustering techniques to surface patterns in intervention efficiency across a diverse set of studies.

Method:

What dataset did you use?

We used a dataset called *“Protein Intake and Muscle Mass Data PMC7727026”* from Kaggle. It’s based on a 2020 meta-analysis by Tagawa et al. that looked at how protein intake affects muscle growth. The person who uploaded the dataset pulled the numbers from the study’s tables and graphs, but only included the studies that used resistance training. Each row in the dataset is a single group from a study—either experimental or control—so we could look at the results more closely. The person also added new columns to make the data easier to work with, like total protein intake and lean body mass change.



	Author and Year	Percent male (%)	Race with largest number	Age (years)	Height (cm)	Weight (kg)	BMI (kg/m2)	Health state	Frequency of exercise before intervention Other	Protein intake before intervention (g/kg/day)	Protein intake before intervention (g/day)
0	Campbell (1995) [1]	83	Caucasian	65.0	173.3	78.0	26.0	Healthy	Unclear		
1	Rozenek (2002) [2]	100	Caucasian	23.2	178.3	76.4	24.1	Healthy	High		
2	Flakoll (2004) [3]	100	Caucasian	18.9	177.0	74.9	23.4	Healthy	High		
3	Rankin (2004) [4]	100	Caucasian	20.8	177.5	78.9	25.1	Healthy	Unclear	1.25	98.7
4	Layman (2005) (with RT) [5]	0	Caucasian	46.7	163.2	87.7	33.0	Healthy	Unclear	0.88	77.0
5	Candow (2006) (Post) [6]	100	Caucasian	64.8	174.0	86.7	28.6	Healthy	Unclear		
6	Candow (2006) (Pre) [6]	100	Caucasian	64.8	174.0	86.7	28.6	Healthy	Unclear		
7	Candow (2006) (Soy) [7]	33	Caucasian	23.2	170.2	70.1	24.2	Healthy	Unclear		
8	Candow (2006) (Whey) [7]	33	Caucasian	23.2	170.2	70.1	24.2	Healthy	Unclear		
9	Kerksick (2006) [8]	100	Caucasian	31.0	179.2	84.0	26.2	Healthy	Unclear	1.85	155.4
10	Cribb (2007) (with Creatine) [9]	100	Others	24.5	181.3	78.5	23.9	Healthy	Unclear	1.7	134.0
11	Cribb (2007) (without Creatine) [9]	100	Others	24.5	181.3	78.5	23.9	Healthy	Unclear	1.7	134.0
12	Hartman (2007) (Milk) [10]	100	Caucasian	24.0	179.0	80.9	25.2	Healthy	High	1.33	111.0
13	Hartman (2007) (Soy) [10]	100	Caucasian	24.0	179.0	80.9	25.2	Healthy	High	1.33	111.0
14	Hoffman (2007) [11]	100	Caucasian	20.7	182.5	95.8	28.8	Healthy	Athlete		

Figure 1: Snippet of raw dataset for context (some columns not seen in image)

What form does this data have? Is it images, raw text, tabular, etc? What are the Features?

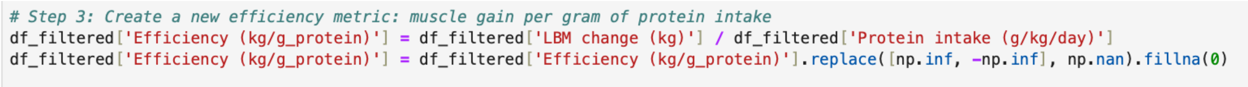
The data is in tabular format and presented as a CSV file. Each row corresponds to a study group and each column corresponds to a feature. Some primary features include protein intake (g/kg/day), energy intake, training frequency, intervention length, and change in lean body mass. Demographic features such as age, weight, height, and BMI also exist, along with whether the subject is in the experimental or the control group. Additional columns were generated to calculate these things: LBM gained per week, per unit of protein, or relative to body weight. A subset of these features were used for the analysis, focusing on those most relevant to training and nutrition.

For unsupervised learning projects (Methodology):

What analysis did you do?
What would be your implementation steps?
How will you evaluate your method?
How will you test and measure success?

We performed unsupervised learning to find patterns in the different training and nutrition setups for ideal and efficient muscle growth. Along the process, we cleaned the dataset and chose key variables such as: Duration (weeks), Frequency (times/week), Energy intake (kcal/kg/day), Protein intake (g/kg/day), and an outcome variable-LBM change (kg). A notable observation from a biological perspective is a new feature called

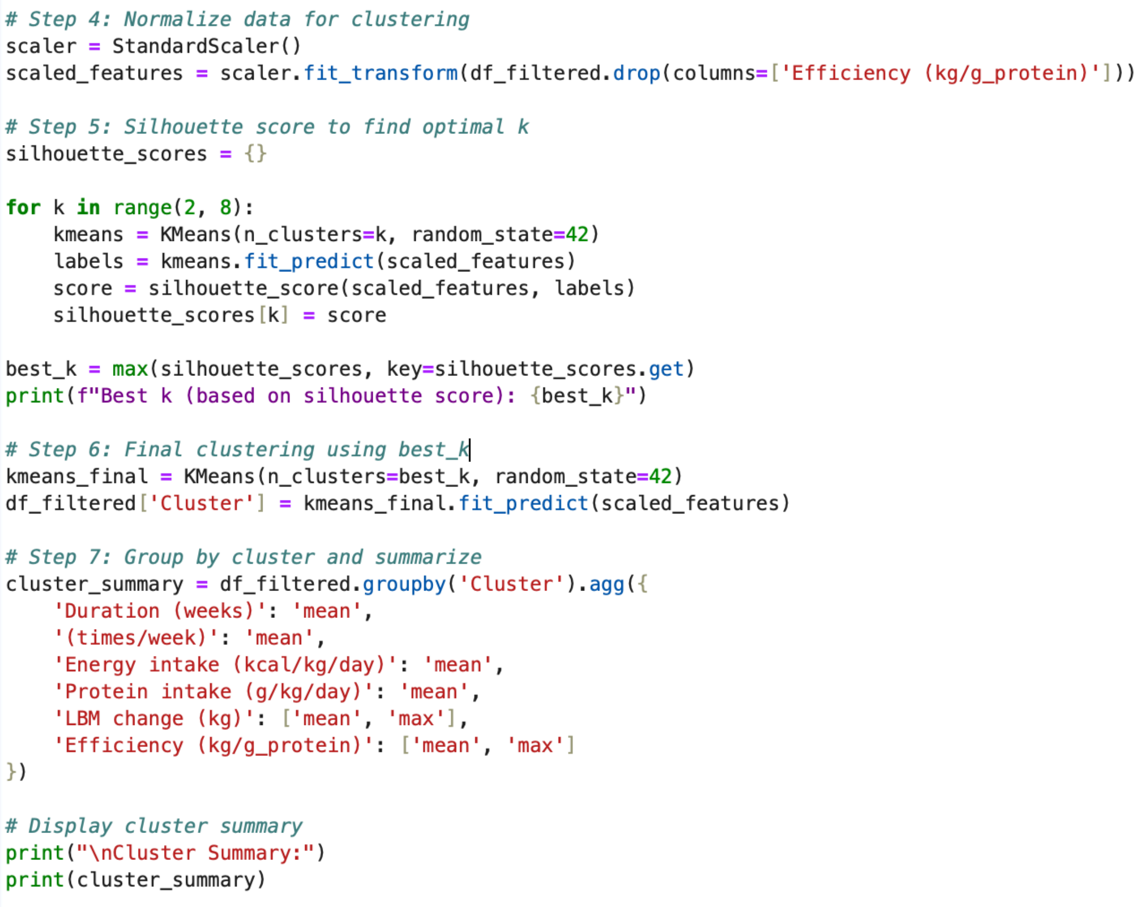
Efficiency, calculated as LBM change per unit of protein, i.e., how efficiently each group turned protein into muscle.



```
# Step 3: Create a new efficiency metric: muscle gain per gram of protein intake
df_filtered['Efficiency (kg/g_protein)'] = df_filtered['LBM change (kg)'] / df_filtered['Protein intake (g/kg/day)']
df_filtered['Efficiency (kg/g_protein)'] = df_filtered['Efficiency (kg/g_protein)'].replace([np.inf, -np.inf], np.nan).fillna(0)
```

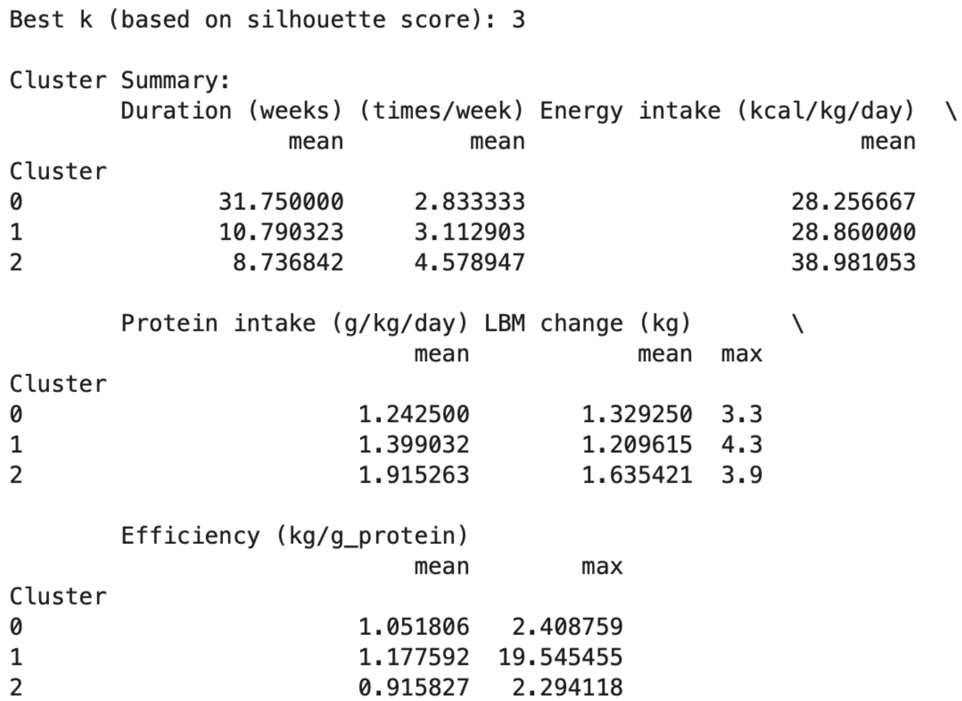
Figure 2: Code snippet calculating efficiency feature column

After dropping rows with missing data and scaling the selected features with StandardScaler, we used the KMeans clustering feature to group similar interventions. We tested multiple values for k and evaluated them using the silhouette score, which measured how well the data points fit within their assigned clusters versus others. The silhouette scores proved that k=3 had the highest value, which is what we used for the final clustering output.



```
# Step 4: Normalize data for clustering
scaler = StandardScaler()
scaled_features = scaler.fit_transform(df_filtered.drop(columns=['Efficiency (kg/g_protein)']))

# Step 5: Silhouette score to find optimal k
silhouette_scores = {}

for k in range(2, 8):
    kmeans = KMeans(n_clusters=k, random_state=42)
    labels = kmeans.fit_predict(scaled_features)
    score = silhouette_score(scaled_features, labels)
    silhouette_scores[k] = score

best_k = max(silhouette_scores, key=silhouette_scores.get)
print(f"Best k (based on silhouette score): {best_k}")

# Step 6: Final clustering using best_k
kmeans_final = KMeans(n_clusters=best_k, random_state=42)
df_filtered['Cluster'] = kmeans_final.fit_predict(scaled_features)

# Step 7: Group by cluster and summarize
cluster_summary = df_filtered.groupby('Cluster').agg({
    'Duration (weeks)': 'mean',
    '(times/week)': 'mean',
    'Energy intake (kcal/kg/day)': 'mean',
    'Protein intake (g/kg/day)': 'mean',
    'LBM change (kg)': ['mean', 'max'],
    'Efficiency (kg/g_protein)': ['mean', 'max']
})

# Display cluster summary
print("\nCluster Summary:")
print(cluster_summary)
```

Figure 3: Code showing silhouette score calculation



```
Best k (based on silhouette score): 3

Cluster Summary:
        Duration (weeks) (times/week) Energy intake (kcal/kg/day)  \
                    mean         mean                        mean
Cluster
0              31.750000     2.833333                   28.256667
1              10.790323     3.112903                   28.860000
2               8.736842     4.578947                   38.981053

        Protein intake (g/kg/day) LBM change (kg)       \
                             mean            mean  max
Cluster
0                        1.242500        1.329250  3.3
1                        1.399032        1.209615  4.3
2                        1.915263        1.635421  3.9

        Efficiency (kg/g_protein)
                             mean        max
Cluster
0                        1.051806   2.408759
1                        1.177592  19.545455
2                        0.915827   2.294118
```

Figure 4: Cluster summary output

To further validate our claims on whether the clusters were visually meaningful or not, we used Principal Component Analysis (PCA) on the data to reduce the dimensions and plotted the clusters in 2D. PCA achieves this by taking numerous related features (e.g., protein intake, duration, and LBM change) and reducing them to a handful of new variables and principal components that preserve the most important patterns in the data. Instead of plotting the initial five features separately, we reduced them to two new axes (PC1 and PC2) that represent the most significant directions of variance for all the groups:

- PC1 can be thought of as a composite of input-intensive factors like protein and energy intake.
- PC2 tends to capture outcomes or efficiency, such as how much muscle was gained relative to input levels.

Each dot on the scatter plot represents one group from a study. It's placed where it stands on these two new composite x and y dimensions. From the scatterplot, we observe that the points that are near each other had similar overall patterns in the original training and nutrition variables, although the original numbers differed somewhat.

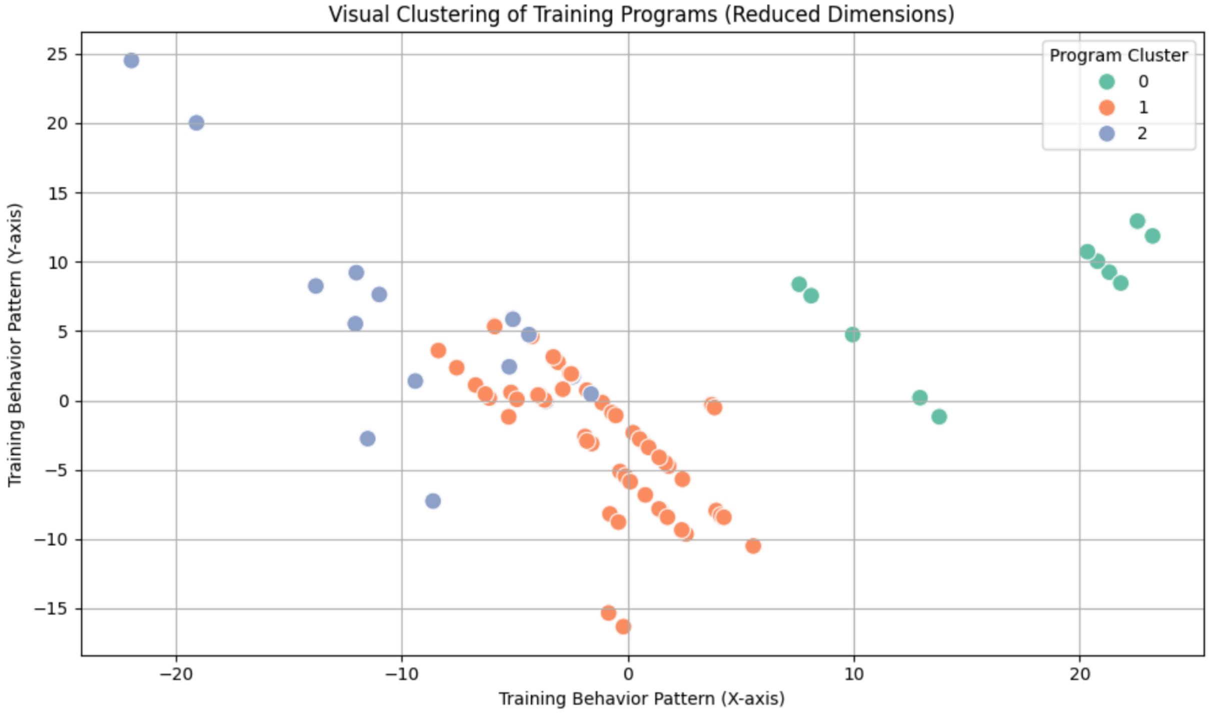


Figure 5: Clustering Scatter plot

While the graph doesn't quantify the components, it helps reveal group relationships visually, but enables us to understand group relationships. If the clusters were still discrete after PCA, it would indicate that our unsupervised model was finding useful structure. Also, it allowed us to visually identify outliers, like the top left point of Figure 5, which aligned to very low input values.

Protein intake and LBM change were then compared within each cluster using box plots. This provided us with a clearer view of what every group did. It also assisted in establishing the fact that the clusters were not random divisions, but rather distinguished variables based on varying intervention profiles.

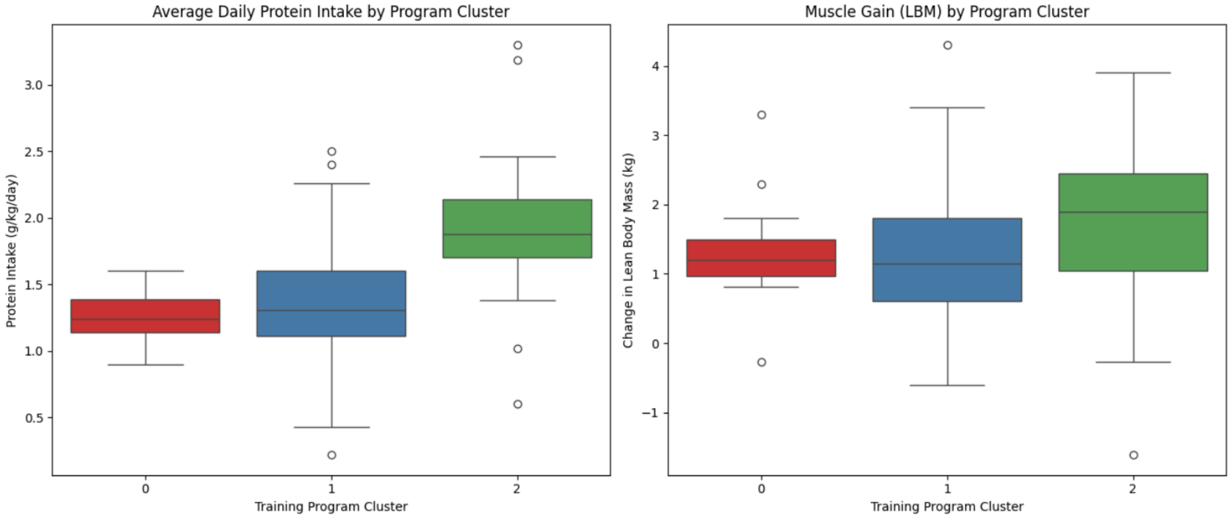


Figure 6: Boxplots showing Protein Intake and LBM Change by cluster

We have included success tests in every part of the process. We measured silhouette scores in order to identify how well-clustering was working, applied PCA plots to manually check whether groupings were valid, and contrasted efficiency as well as uptake level among groups in an attempt to validate the output being produced. We considered the process successful if it uncovered clear, meaningful clusters tied to training and nutrition variables, without needing labeled data.

Results:

For unsupervised learning projects:
What results did your analysis show? Visualize them if possible
What new questions do these results raise, and how can they be addressed by further analysis?
Repeat as necessary

Our cluster analysis revealed three unique profiles of training and nutrition interventions, each with different levels of lean body mass (LBM) change and efficiency. The most intriguing finding was in Cluster 2, which consistently reported significant muscle gain with a moderate protein intake and regular training frequency. It was the most efficient group with better results for less input. This challenged the common assumption that more protein always leads to more growth. On the other hand, Cluster 0 included groups with longer training durations but without proportional muscle gains. This shows that training for longer durations does not necessarily equate to more results if intake or program design is not maximised.

Lastly, Cluster 1 showed frequent training, but low muscle gains which implied inefficiency or even underfeeding. It recorded the lowest mean gain in LBM, and hence it was the worst performing cluster overall.

As expected, the results naturally raised questions about why certain clusters were better or worse performing than others. Specifically, why did Cluster 2 produce the largest lean body mass gains on minimal input, and why did Cluster 1 underperform on regular training? These questions prompted us to investigate further into the CSV dataset and find out what variables the groups of each cluster shared. By identifying patterns such as consistent protein intake ranges, energy balance, and duration thresholds, we had a better sense of what kinds of intervention combinations actually work best.

We also identified one clear outlier in the scatterplot from group 2, a group that gained a lot of muscle with very low intake and short training duration. It stood apart from all other groups in PCA space, prompting us to trace it back through the data and ask why this group behaved so differently. Was it a measurement anomaly, or did it reflect something physiological like newbie gains? Nonetheless, it didn't significantly change the broader trends observed across the rest of the dataset.

Overall, the results suggest that clustering can significantly differentiate between intervention types and identify which groups of inputs relate to enhanced performance and efficiency of LBM gains.

Discussion:

What outcome did you expect from your results?
How did your actual results differ from your expected results?

We went into this project with an open mindset, just letting the data show us whatever it had to say about muscle growth. That said, our initial assumptions matched common beliefs among gym-goers that protein intake, how often you train, and how long you train all play a big part in hypertrophy, and that getting about 1g of protein per kg of body weight is usually considered enough. But the results actually told us a different story. The most efficient group in our clustering analysis gained more muscle when they were taking in somewhere between 1.5 to 2.0g/kg/day of protein and training frequently. That matched up with a hunch we had from the beginning — that going above the basic protein recommendation can lead to better muscle growth. It was unexpectedly reassuring to see the data support this idea so clearly.

If your final report differs from your proposed project, discuss the differences, why you made certain changes, and the bottlenecks that prevented you from proceeding with the proposed project.

Our final report is significantly different from our original proposal. This project idea was actually our original idea in the initial brainstorming phase, but we were having trouble finding a usable dataset. We had also considered using wearable technology to track our own data, web scraping websites like MyFitnessPal, or using public APIs. Yet we soon came to understand that although these sites hold heaps of useful information, much of it is personal and therefore inaccessible, and its use would be a breach of user privacy. That made the project considerably more difficult to undertake. As a result, we moved to a new idea that was all about improving YouTube's Explore algorithm specifically towards young fitness influencers, but it wasn't one we were as excited about. Once you told us we could make our own dataset, we returned to our original idea, thinking we could just invent sample numbers to practice the skills. Luckily, we eventually came across a real-world dataset on Kaggle that lined up almost exactly with what we had in mind from the start. After that, we just dove in and got to work.
One useful thing about the PCA scatter plot was that it helped us spot outliers—like that one point in the upper left of Figure 5, which stood out due to its extremely low input values.After grouping the data, we used box plots to compare protein intake and lean body mass changes across the different clusters, which gave us a better sense of how each group performed.