

## A vibrant, top-down view of a variety of fresh fruits and vegetables, including lemons, tomatoes, avocados, blueberries, and kiwis, arranged in a heart shape on a dark background. The arrangement is composed of several heart-shaped white bowls and individual pieces of produce. The items include: sliced lemons, cherry tomatoes, almond slices, mango chunks, quinoa, garlic, a halved avocado, blueberries, kiwi slices, cucumber sticks, ginger root, and a mix of seeds. The overall composition is colorful and healthy, set against a dark, textured background.

# Variables

- Participant\_ID: identifier for each participant
- Age: age of the participant in years
- Gender: gender of the participant (M/F)
- Current\_Weight (lbs): initial weight of the participant in pounds
- BMR (Calories): Basal Metabolic Rate, the number of calories burned at rest
- Daily\_Calories\_Consumed: average daily caloric intake
- Daily\_Caloric\_Surplus/Deficit: difference between calories consumed and BMR per day
- Weight\_Change (lbs): estimated amount of weight change in pounds
- Duration (weeks): duration of the observation period (1-12 weeks range)
- Physical\_Activity\_Level: participant's level of physical activity (Sedentary, Lightly Active, Moderately Active, or Very Active)
- Sleep\_Quality: self-reported sleep quality (Poor, Fair, Good, or Excellent)
- Stress\_Level: participant's self-reported stress level (1-10 scale)
- Final\_Weight (lbs): weight of the participant at the end of the observation in pounds

Data:

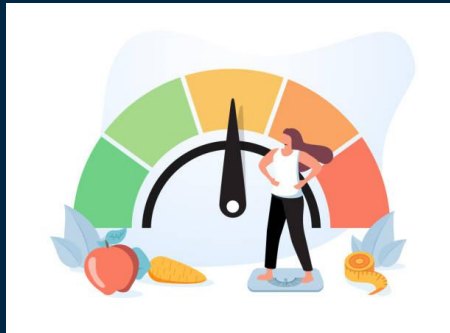
<https://www.kaggle.com/datasets/abdullah0a/comprehensive-weight-change-prediction>

Question 1: What variables are the most influential for predicting the weight change over a specified duration?



# Variables

- Weight change
- Daily caloric surplus/deficit
- Physical activity level
- Age
- BMR
- Current weight

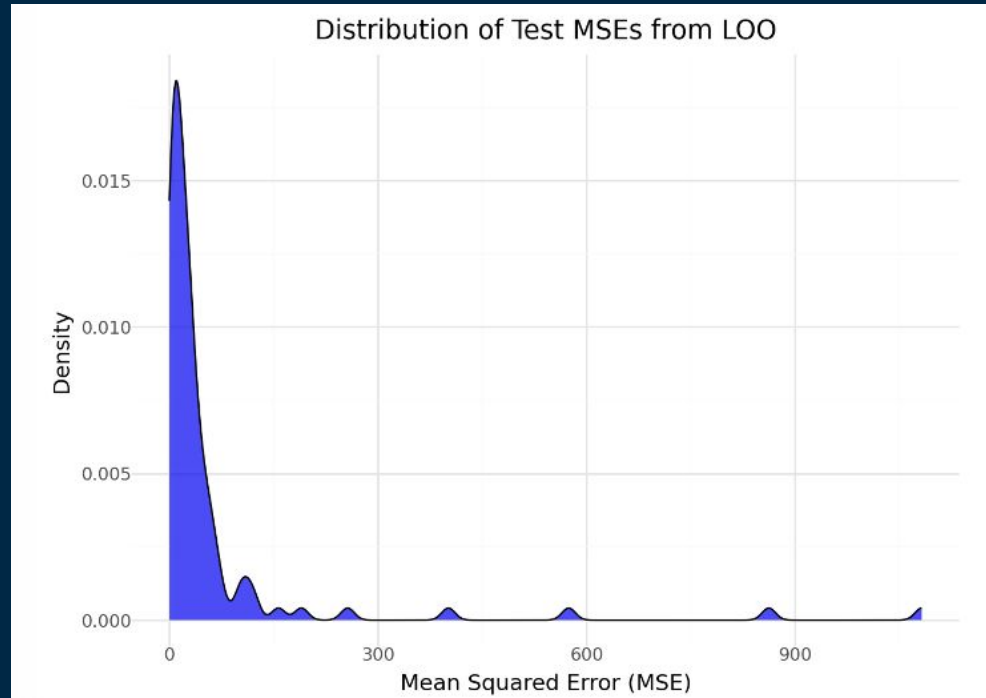


# Methods

- Created a linear regression model using the variables
- Load in dataset and drop any missing values
- Split X (predictors) and y (weight change)
- Use LabelEncoder to make categorical variable numeric for physical activity level
- Train test split and z-scoring with StandardScaler
- Use leave one out validation
- Find coefficients for each feature
- Create ggplot using geom\_density and make scatter plots

# Results

Leave-one-out linear regression model test MSEs:



## Train and test MSE mean:

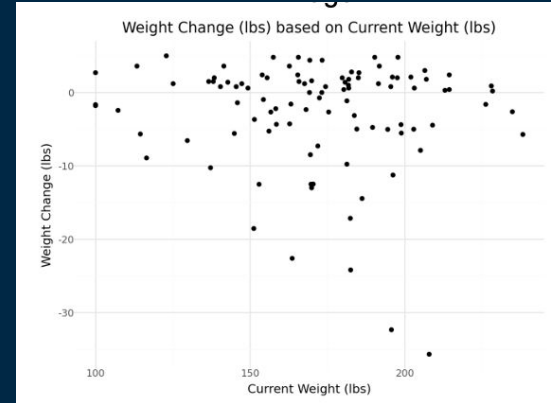
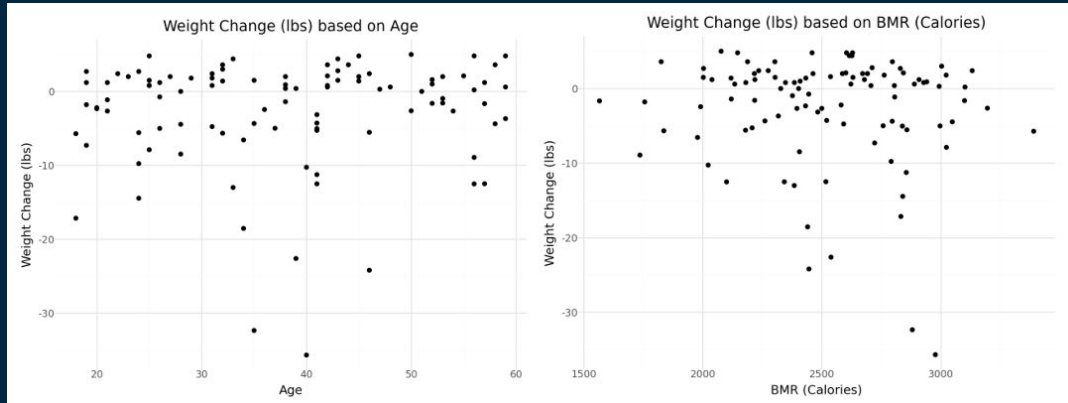
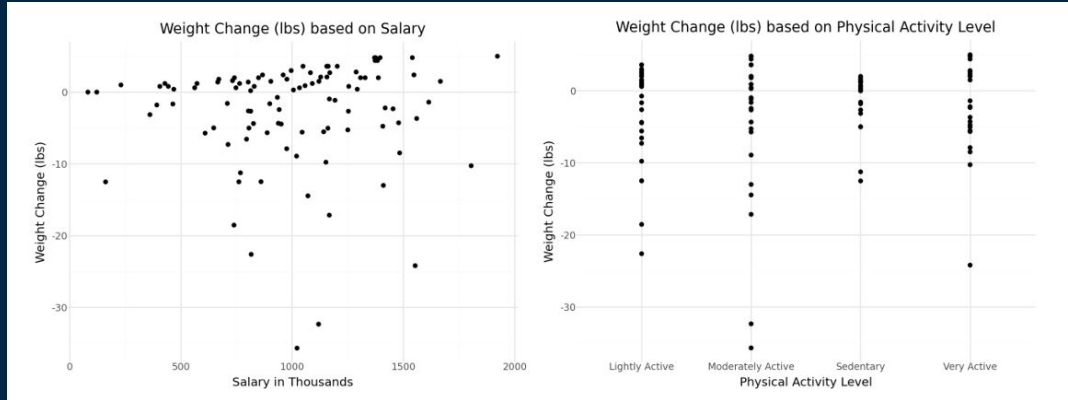
Train MSE : 53.988930808166806

Test MSE : 59.45644167128791

## Coefficient of the features:

Feature	Coefficient
Daily Caloric Surplus/Deficit	0.066899
Physical Activity Level	0.171624
Age	0.326022
BMR (Calories)	-1.084684
Current Weight (lbs)	0.332971

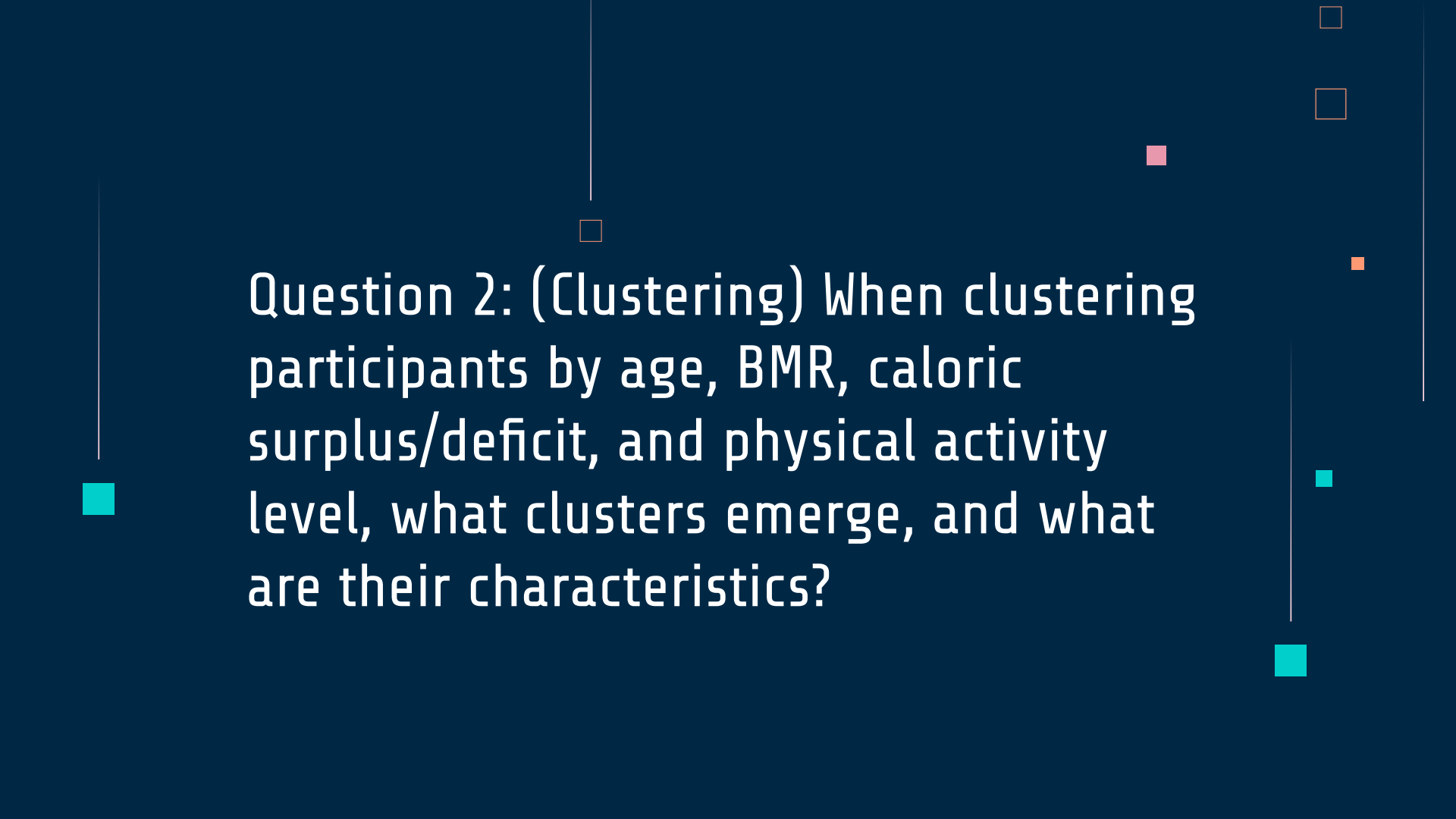
# Scatter plots:





# Discussion

- Linear Regression Model
  - Leave-One-Out (LOO) method
  - Test MSEs were visualized using `geom_density` ggplot
  - Train MSE: 53.99
  - Test MSE: 59.46
- Coefficients
  - Influence of each feature on weight change
  - Most impactful are BMR, current weight, and then age
- Scatter plots
  - Strongest trends in BMR and current weight
  - Other features are weak or show no clear relationship
- BMR and current weight are the most influential for predicting the weight change



Question 2: (Clustering) When clustering participants by age, BMR, caloric surplus/deficit, and physical activity level, what clusters emerge, and what are their characteristics?

# Variables

- Age
- BMR
- Daily caloric surplus/deficit
- Physical activity level

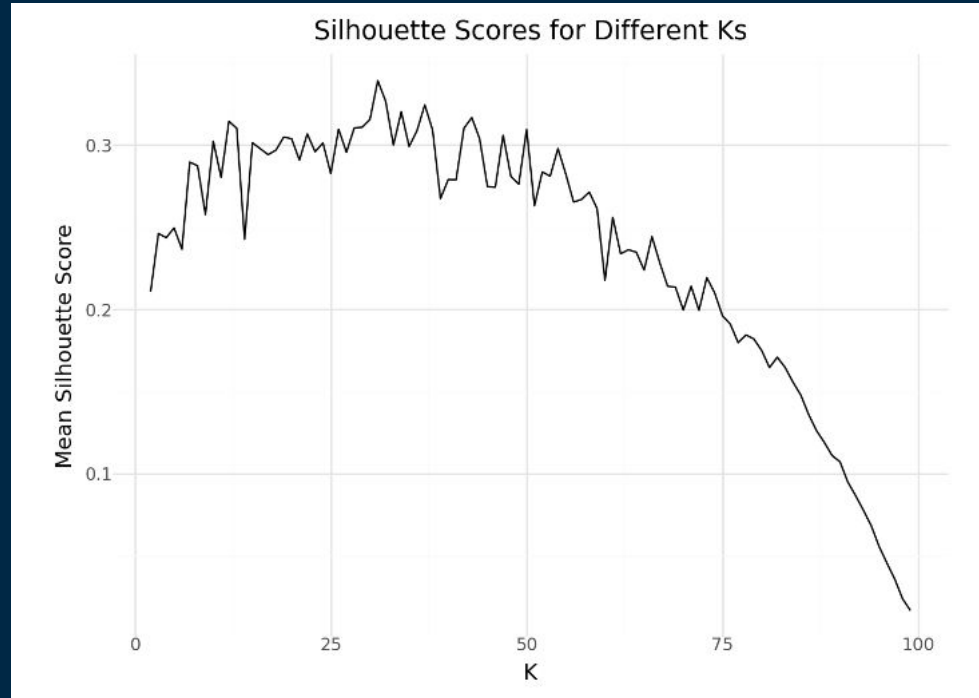


# Methods

- Perform k-means clustering
- Use X to hold the 4 features
- Use LabelEncoder to make categorical variable numeric for physical activity level
- StandardScaler to standardize and create a dictionary
- Create ggplot to show the silhouette scores for the different Ks
- Test the different k-means values to see which performs best
- Find the silhouette score averages

# Results

Silhouette scores for different Ks:



# Scatterplots of the different features:

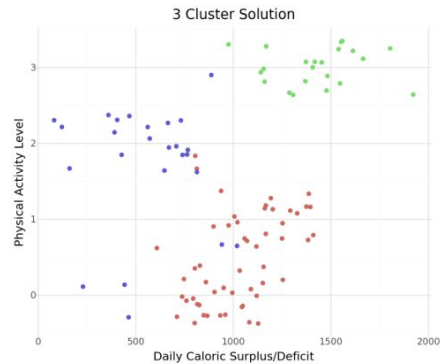


Image 8



Image 9

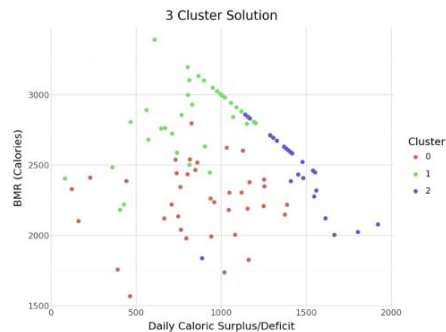


Image 10

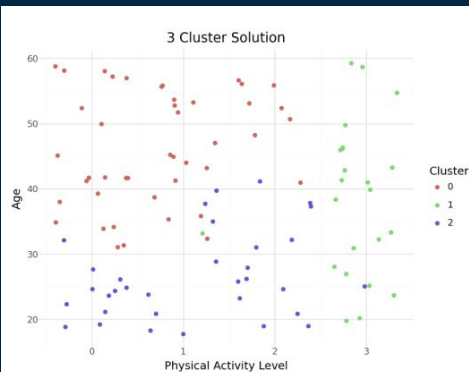


Image 11

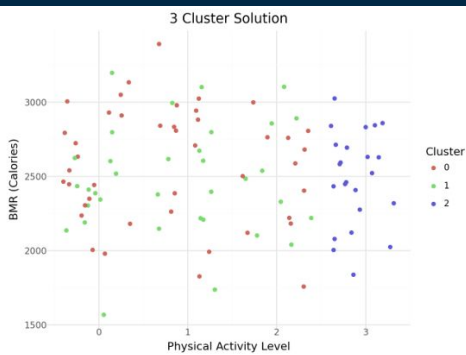


Image 12

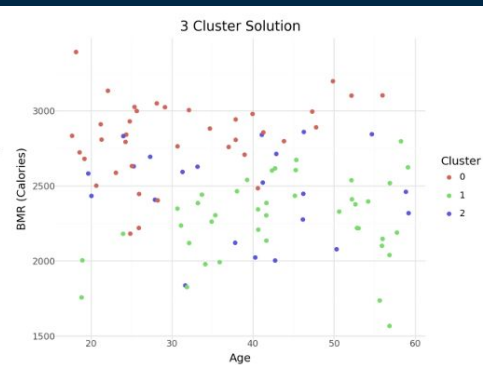
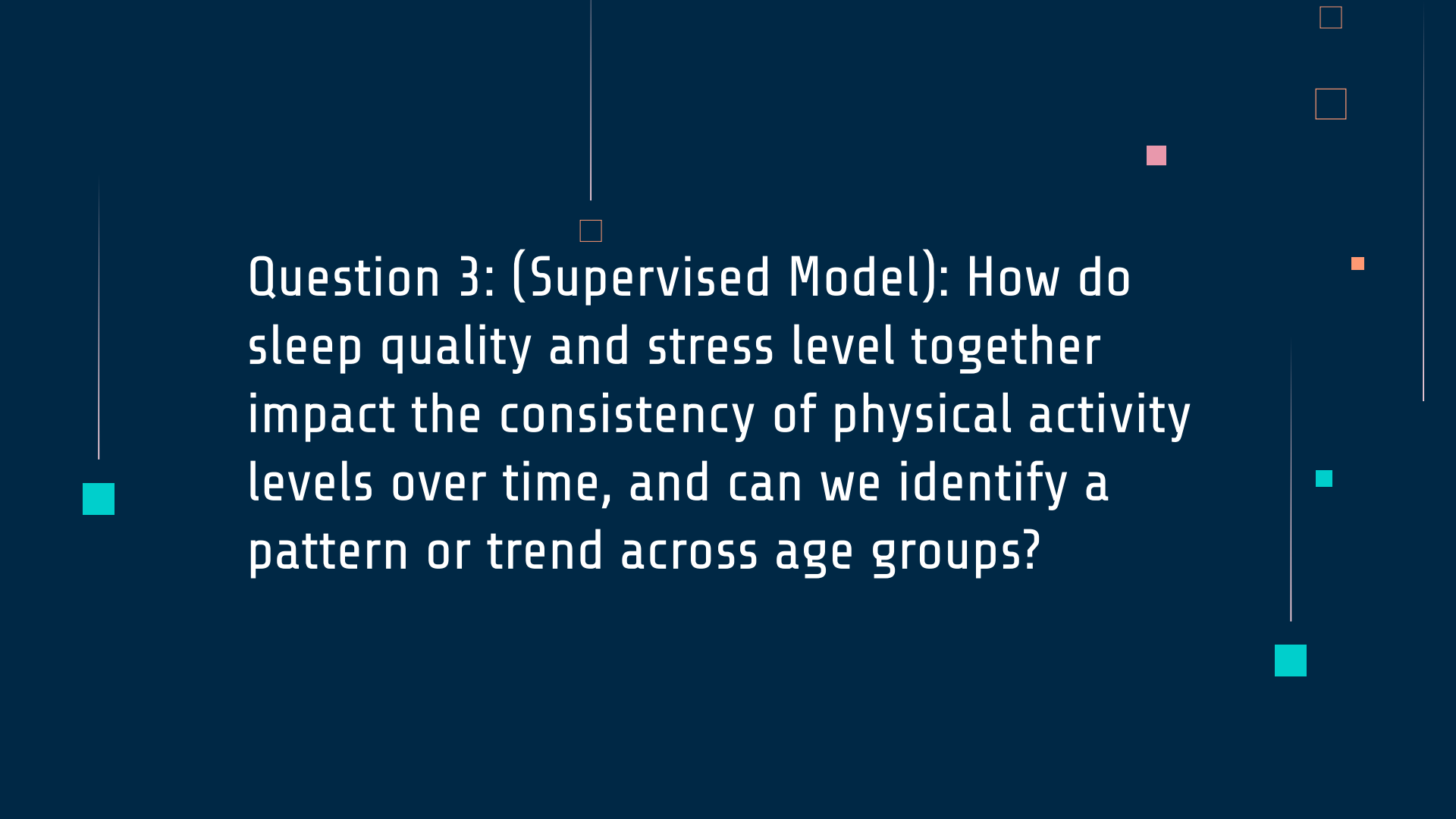


Image 13

# Discussion

- K-Means
  - Chosen for clear assignments, interpretability based on feature average, and assume spherical shape
- Silhouette score
  - Average silhouette score: 0.0167
- Clustering
  - Three clusters selected based on testing
- Observation
  - Best clustering is physical activity level and daily caloric surplus/deficit
  - Silhouette score: 0.2577 (highest among tested features)
  - Better score compared to the rest but still moderate separation and cohesion
  - Desired scores is above 0.5 for high cohesion and separation

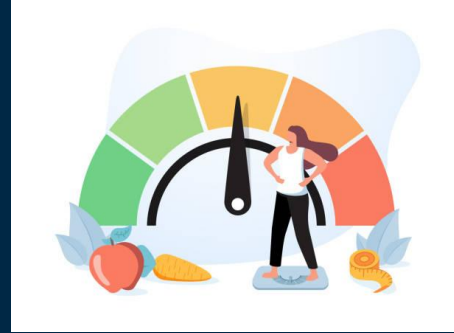


Question 3: (Supervised Model): How do sleep quality and stress level together impact the consistency of physical activity levels over time, and can we identify a pattern or trend across age groups?



# Variables

- Sleep Quality
  - Excellent, good, fair, and poor
- Stress Level
  - Scale of 1-10
- Age
  - Varied
- BMR
- Calories
- Gender
- Target Variable: physical activity level
  - Low, moderate, high

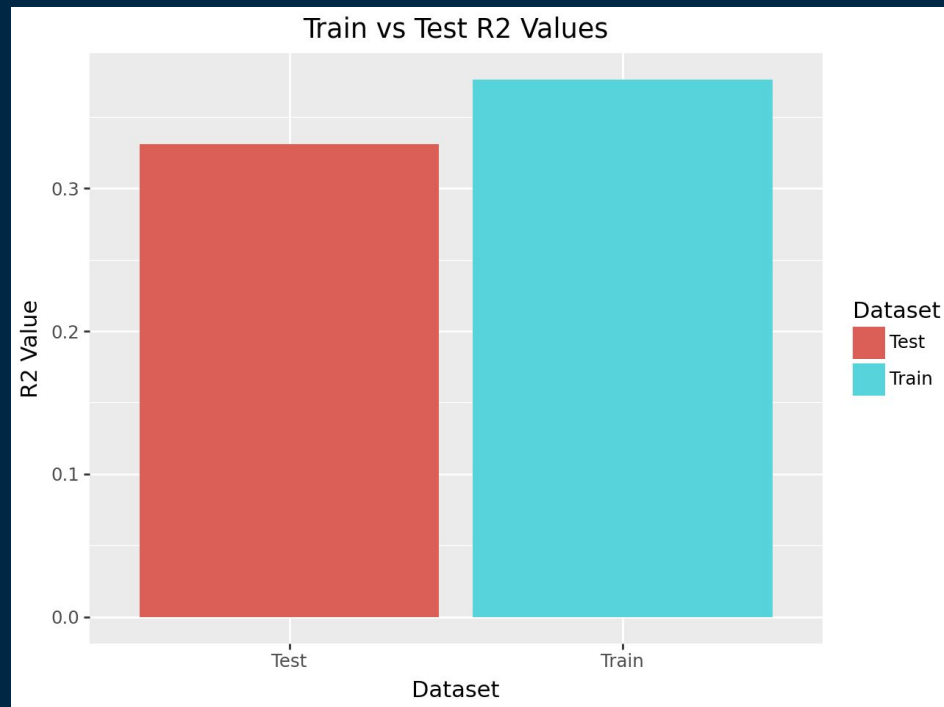


# Methods

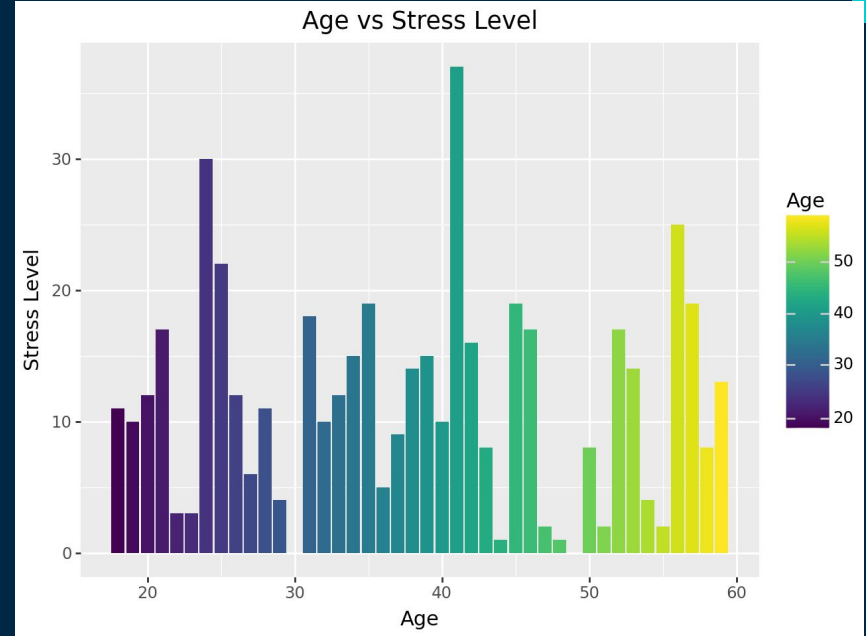
- A linear regression model was built, including interaction terms to explore combined effects such as Sleep Quality × Stress Level.
- The data was split into 80% for training and 20% for testing to evaluate the model's performance.
- Cross-validation was performed to ensure the stability and consistency of the results
- Handling Missing Values:
  - Stress Level: Missing values were replaced with the mean of the available data.
  - Sleep Quality: Rows with missing values were removed since one-hot encoding requires complete categorical data

# Results

```
➡ Train MSE : 3.9776163286126476  
Train MAE : 1.7053998710270981  
Train MAPE: 0.7026716571203618  
Train R2 : 0.37599900717911205  
Test MSE : 4.833966000835138  
Test MAE : 1.9298532232692787  
Test MAPE : 0.6137611239407665  
Test R2 : 0.3311703907526615
```



# Results



Question 4: (Dimensionality Reduction) How does a model using PCA on all continuous variables (BMR, daily caloric surplus/deficit, stress level) compare to a model with the original variables in terms of mean squared error when predicting final weight?

# Variables

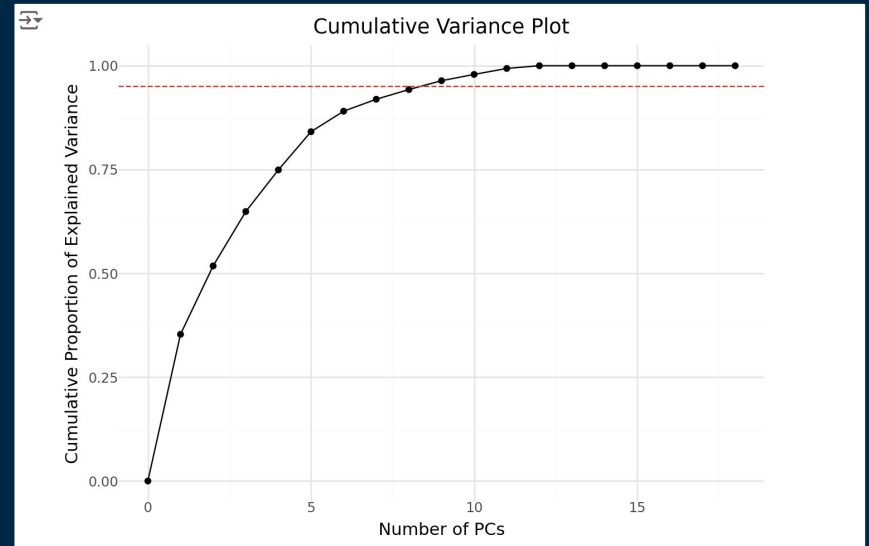
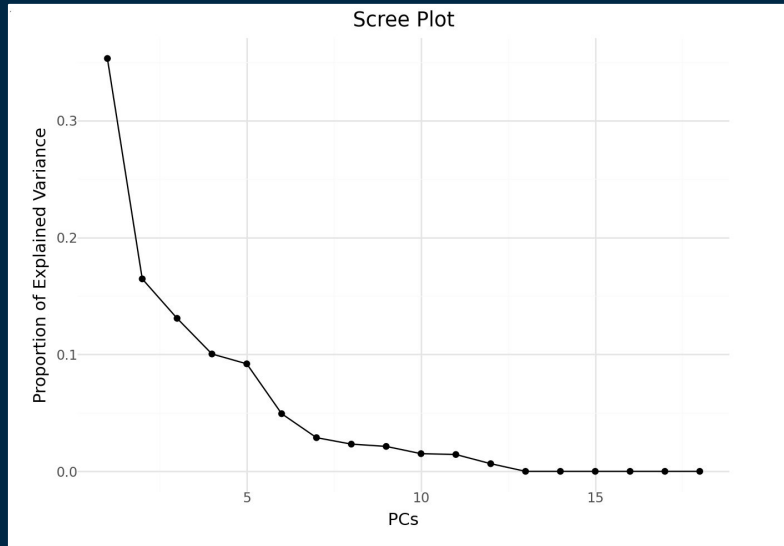
- BMR
  - Continuous variable
- Age
- Physical activity level
- Daily caloric surplus/deficit
  - Continuous variable
- Stress Level
  - 1-10 scale
- Target Variable
  - Final weight (lbs)



# Methods

- A linear regression model was created to compare the performance of two approaches for predicting Final Weight: one using PCA-transformed components and another using the original continuous variables.
- The PCA approach involved transforming the continuous predictors into principal components to reduce dimensionality while retaining the majority of the variance.
- Continuous variables were standardized using StandardScaler to ensure all predictors were on the same scale before applying PCA or fitting the regression models.
- The dataset was split into 80% training and 20% testing for both models.
- A linear regression model using the PCA-transformed data (10 principal components as predictors).

# Results





# Results

