Diet Analysis Report

Introduction

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

The data is from Kaggle, titled “Diet Analysis, Predict The Weight” by M. Abdullah. The dataset includes 100 participants' data covering demographics about their diet, physical activity, and lifestyle. It aims to use the information gathered to analyze the trend in how the variables impact weight fluctuations.

The variables include:
- 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

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

Methods

I created a linear regression model to determine which variables contribute to weight change prediction. The variables that were used are weight change (continuous) and the predictors are daily caloric surplus/deficit (continuous), physical activity level (categorical), age (continuous), BMR (continuous), and current weight (continuous). The dataset is loaded and any missing value is dropped. Then X and y are split, with X as

the predictors and y as the weight change. LabelEncoder from Sklearn is used to make the categorical variables numeric. Then, the train test split and z-scoring of the predictors with StandardScaler are performed. During z-scoring, it will scale the numerical variables. Using the leave-one-out validation, the training and testing MSEs are printed. The coefficients for each feature are printed as well. A ggplot using geom_density shows the distribution of test MSEs from leave one out. Lastly, scatter plots were created to visualize how the different predictors work with the weight change.

Results

Result of leave-one-out from linear regression model test MSEs:

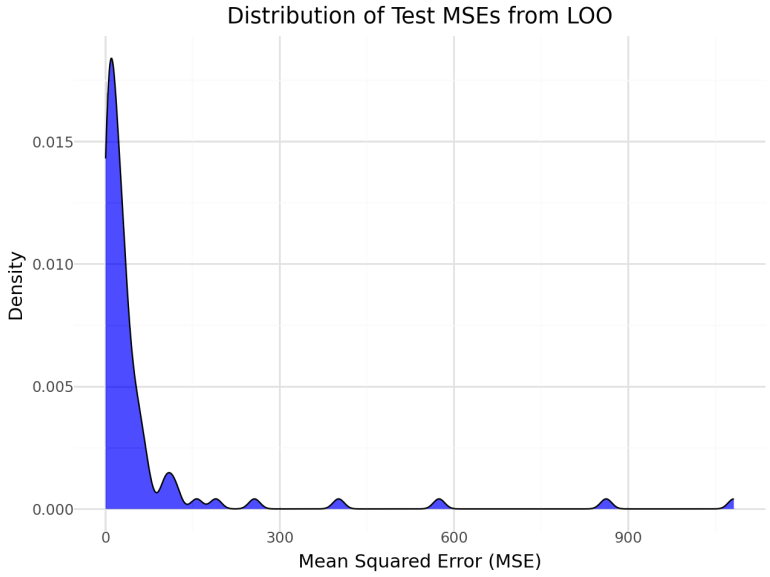


Image 1

Train and test MSE mean:

```
Train MSE : 53.988930808166806
Test MSE  : 59.45644167128791
```

Coefficient of the features:

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

Scatter plots:

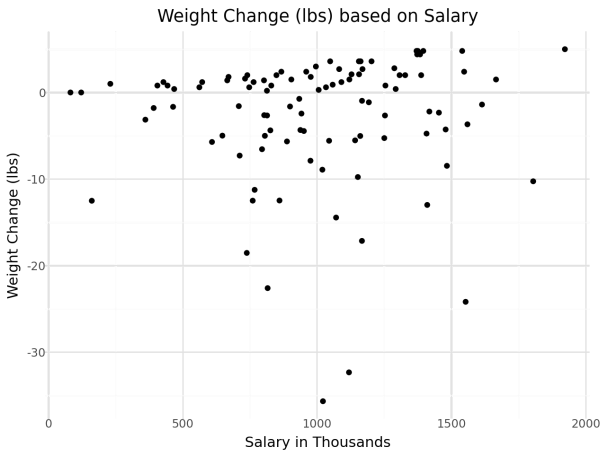


Image 2

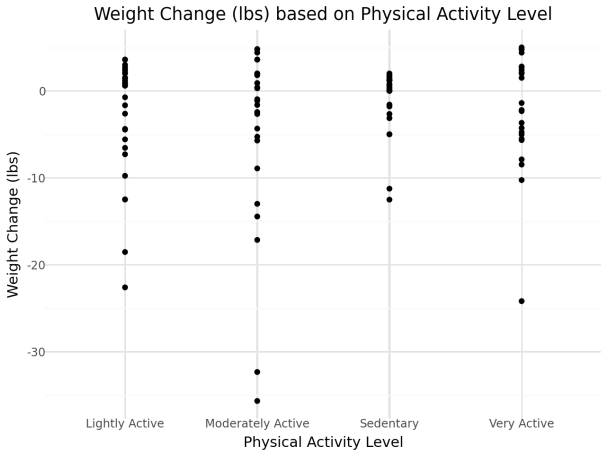


Image 3

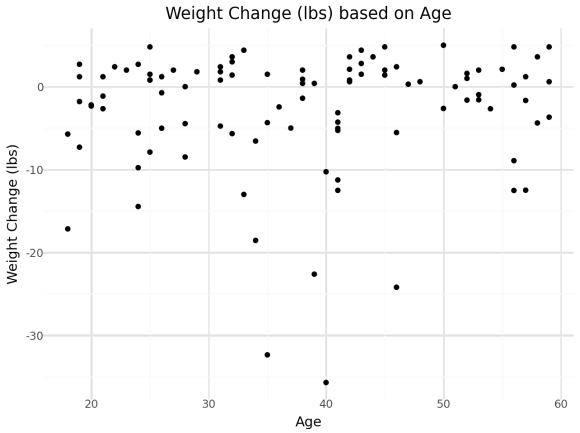


Image 4

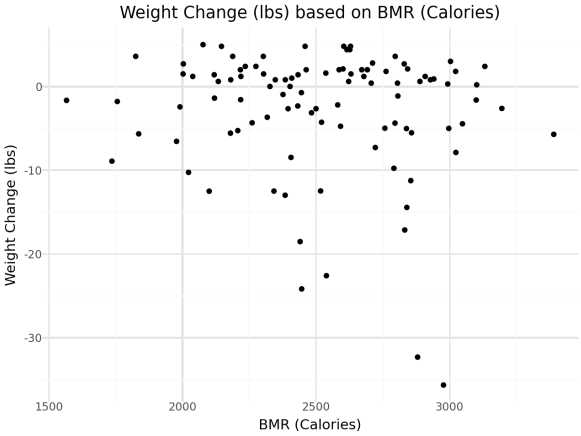


Image 5

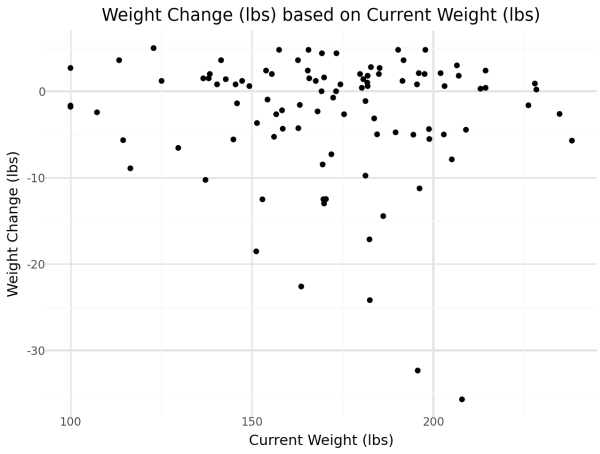


Image 6

The scatter plots show the weight change based on daily caloric surplus/deficit, physical activity level, age, BMR (calories), and current weight (lbs).

Discussion

The linear regression model uses weight change as the dependent variable and the others as the predictors. The leave-one-out method analyzes the model's performance by leaving one row out at a time and measuring the predicted accuracy. Image 1 shows the distribution of test MSEs from LOO using geom_density ggplot. The train MSE is consistent across the splits, given the average of about 53.99, which means the model fits the training data well. The MSE test shows that some splits have very low errors, close to 0, whereas some have extremely high errors, over 300 MSE. The average test MSE is 59.46, higher than the train MSE, suggesting that some are overfitted.

The coefficients represent the influence of each feature on the dependent variable, weight change. The larger values indicate greater influence. A positive coefficient increases weight change as its value increases. On the other hand, the negative coefficient decreases weight change as its value increases. Since the predictors were also standardized using z-scoring, the coefficients are comparable. To answer the question, based on the coefficients BMR is the most influential, followed by current weight and age.

In the scatter plots, BMR and current weight change in images 5 and 6 show the strongest trends with weight change. Image 5 appears to have a weak negative correlation, with individuals with higher BMR values having smaller weight changes. Image 6 shows individuals with heavier weights have slightly larger weight changes. The other features like salary, physical activity, and age show weak or no clear relationship with weight change. The trend aligns with the coefficient analysis, where BMR and current weight had strong influences.

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

Methods

Using the dataset, I selected 4 features to make clusters using k-means. The features are age (continuous), BMR (continuous), daily caloric surplus/deficit (continuous), and physical activity level (categorical). I used X to hold the data features and then made the physical activity level categorical to numeric. Then, I used StandardScaler to standardize and created a dictionary metric to see how k-means performs in the dataset range. I used ggplot to show the silhouette scores for different Ks and found the silhouette score averages. The highest point is around 37 so I tested it for the individual

models. However, I tested different k-means and used 3 for the different features. The clusters are shown for the 6 comparisons, and a scatter plot is created for each.

Results

Silhouette scores for different Ks:

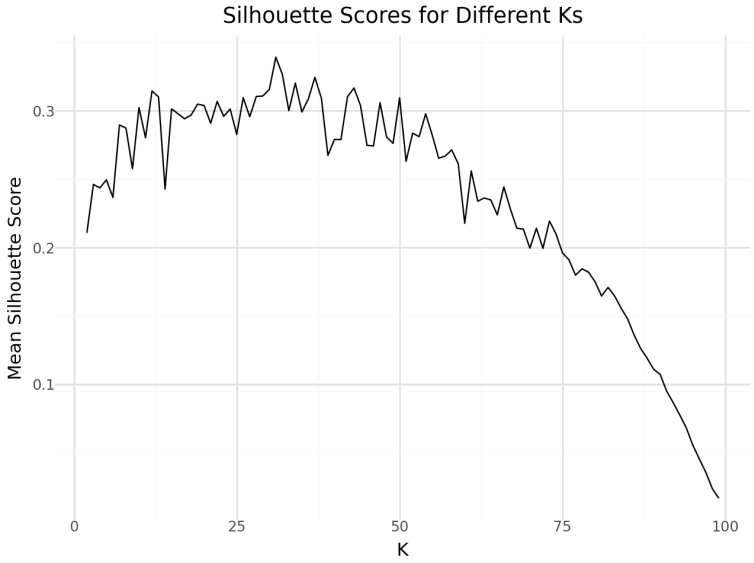


Image 7

Scatterplots of the different features:

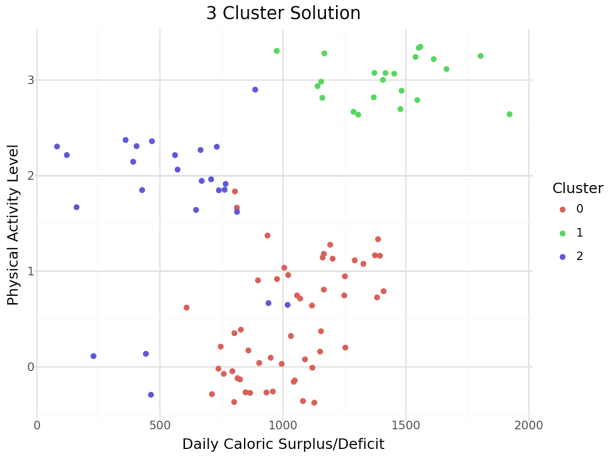


Image 8

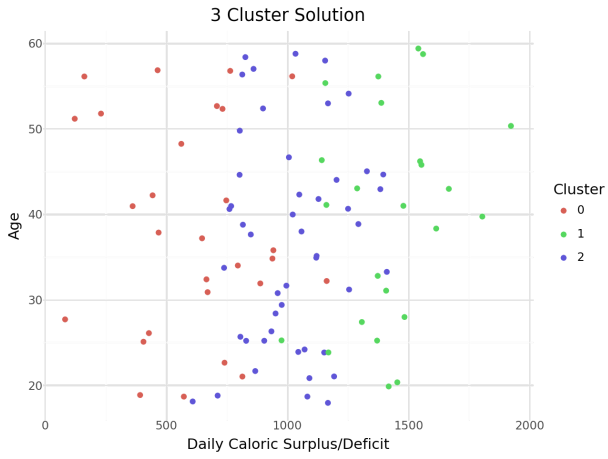


Image 9

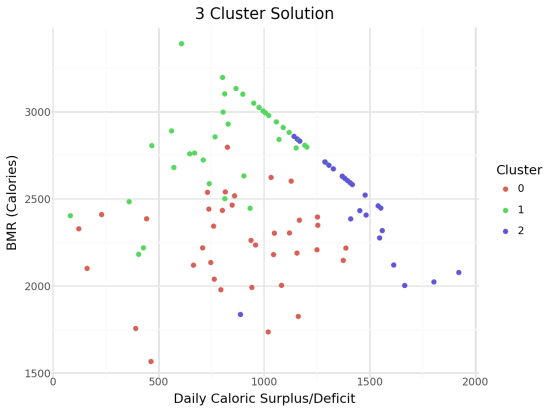


Image 10

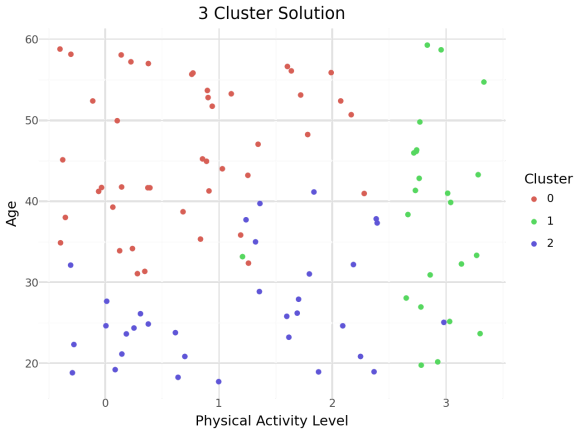


Image 11

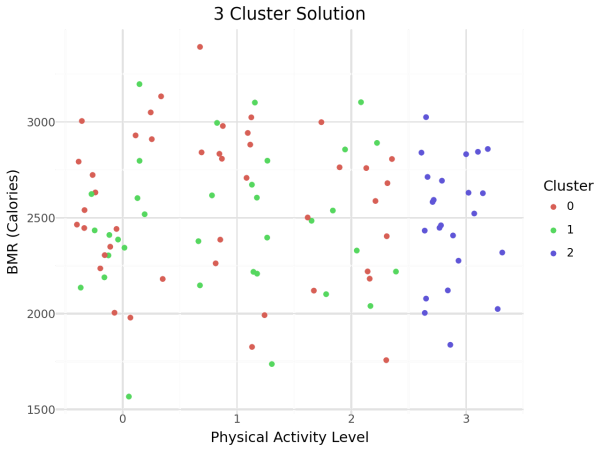


Image 12

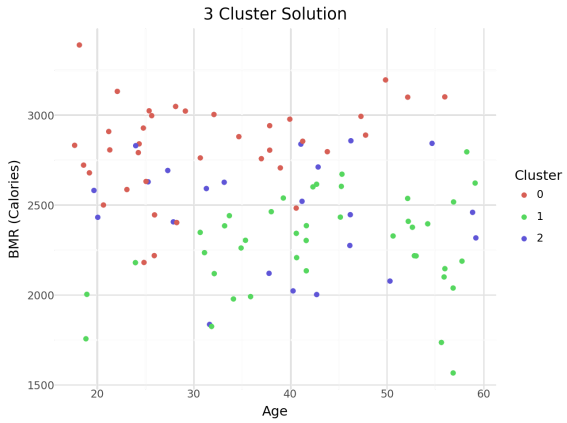


Image 13

Discussion

Clusters are formed based on age, BMR, daily caloric surplus/deficit, and physical activity level. The clustering process will show the patterns among individuals based on the features. K-means is chosen for the clustering method because it provides clear clustering assignments and interprets clusters based on the averages of the features. It also assumes clusters are spherical, which works for features like age, BMR, caloric surplus/deficit, and physical activity level, because they are likely to form distinct groups. For example younger vs. older individuals or physically active vs. inactive individuals. The silhouette score measures how cohesive and separated the clusters are. A +1 score means the clusters are perfectly separated, 0 means overlapping, and -1 means poorly formed. The silhouette score average is 0.0167, which is not performing poorly. The silhouette analysis plot indicates that the clustering solution becomes less impactful as clusters (K) increase. After testing the cluster solutions with different k-means values, I used three cluster solutions as they work with higher silhouette scores.

The scatter plots show the clustering results and represent three clusters based on age, BMR, daily caloric surplus/deficit, and physical activity level. Each cluster represents a group of participants with similar characteristics. None of the clusters shows high separation between the other clusters and high cohesion in each cluster, however the three cluster solutions for physical active level and daily caloric surplus/deficit have high separation as seen in image 8. Its silhouette score is 0.2577, the highest compared to the other features, showing that they are more separated and cohesive. However, 0.2577 does not show high cohesion and separation, as a good silhouette score is greater than 0.5.

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?

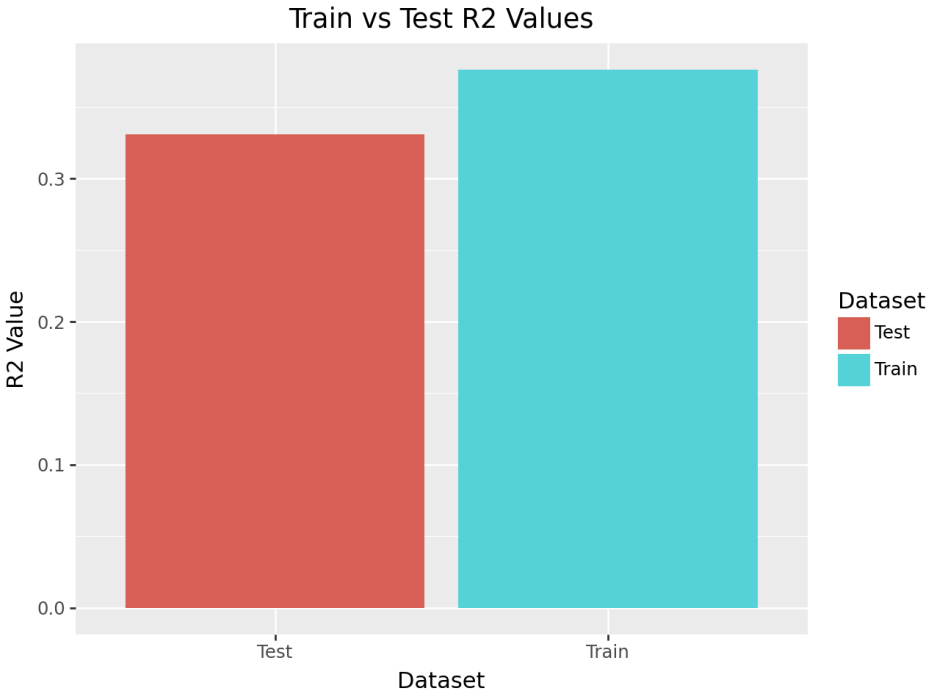
Methods

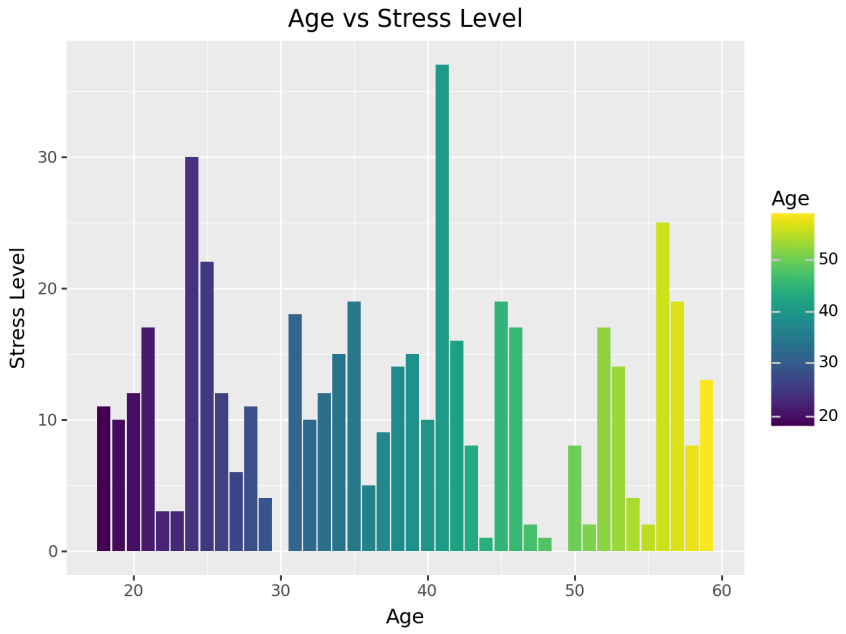
To explore the impact of Sleep Quality and Stress Levels on the consistency of Physical Activity Levels, a linear regression model was used. The predictors included Sleep Quality, Stress Level, and Age, with Physical Activity Level (categorical: Low, Moderate, High) as the target variable. The dataset was preprocessed by encoding physical activity level as an ordinal variable (Low = 1, Moderate = 2, High = 3). Sleep quality was one-hot encoded to represent its categories (Excellent, Good, Fair, Poor). Continuous predictors, such as Age and Stress Level, were standardized using z-score normalization to ensure comparability.

Missing values for stress levels were imputed using the mean, and rows with missing sleep quality scores were dropped from the dataset. The dataset was split into training and testing sets (80/20), ensuring generalization of results. Interaction terms between predictors were added to test for combined effects of stress and sleep quality on physical activity. Model evaluation included calculating R^2 for accuracy and inspecting cross-validation stability. Data visualization techniques, such as bar plots, were used to further examine relationships between variables.

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
```





Discussion

A supervised regression model was employed to analyze how sleep quality and stress level together influenced physical activity levels. The data was split into training and testing sets, and the model's performance was evaluated using R^2 values. The train R^2 was 0.375, while the test R^2 was 0.33, indicating that the model performed poorly overall. The slightly higher train R^2 compared to the test R^2 suggests mild overfitting, where the model captured more variance in the training data than it could generalize to unseen data. However, the difference between the train and test R^2 values was not substantial, meaning the overfitting was not severe. Nevertheless, these results demonstrate that the model struggled to explain the variability in physical activity levels.

The R^2 plot illustrated the gap between training and testing performance. It shows that the model has similar R^2 values for the Train (~0.37) and Test (~0.33) datasets, indicating it generalizes reasonably well without significant overfitting. The Age vs. Stress Level graph revealed a clear trimodal distribution of stress levels across age groups. Stress peaked at three distinct life stages: early adulthood, midlife, and later life. These peaks are likely associated with significant life transitions or challenges. For example, early adulthood may be marked by educational or career pressures, midlife by family and professional responsibilities, and later life by health and financial concerns. These stress peaks can intersect with other factors, potentially leading to outcomes such as weight gain, decreased physical activity, or other health-related issues. The trimodal distribution highlights how stress varies across different stages of life, providing valuable context for understanding its broader impacts on physical activity and health.

The "Sleep Quality vs. Stress Level" graph shows a clear relationship between stress levels and sleep quality. Individuals with Poor Sleep Quality report the highest stress levels, as indicated by the tallest bar, while those with Excellent Sleep Quality have the lowest stress levels. This pattern suggests that as sleep quality declines, stress levels increase significantly. Moderate categories like Fair and Good Sleep Quality display intermediate stress levels, highlighting a gradient relationship between these variables. The graph emphasizes the importance of maintaining good sleep quality to manage stress levels, as poor sleep may contribute to heightened stress and its associated health risks.

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?

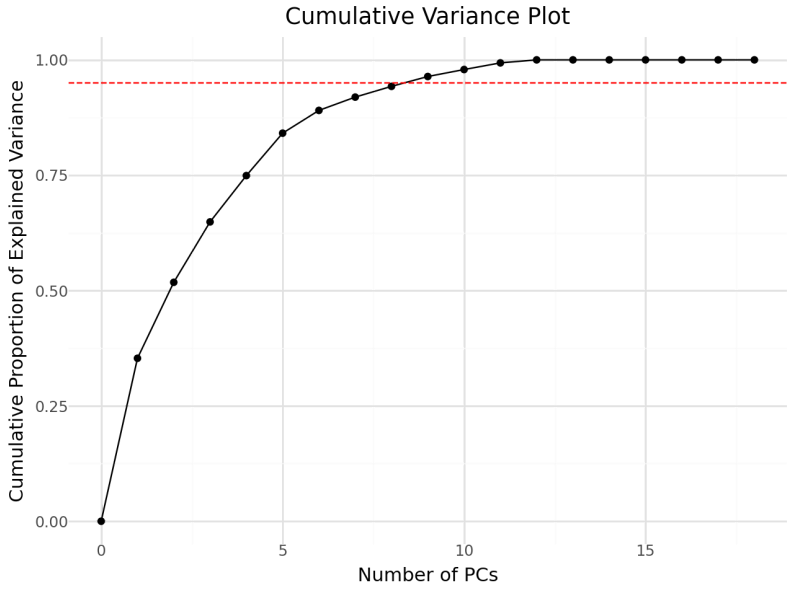
Methods

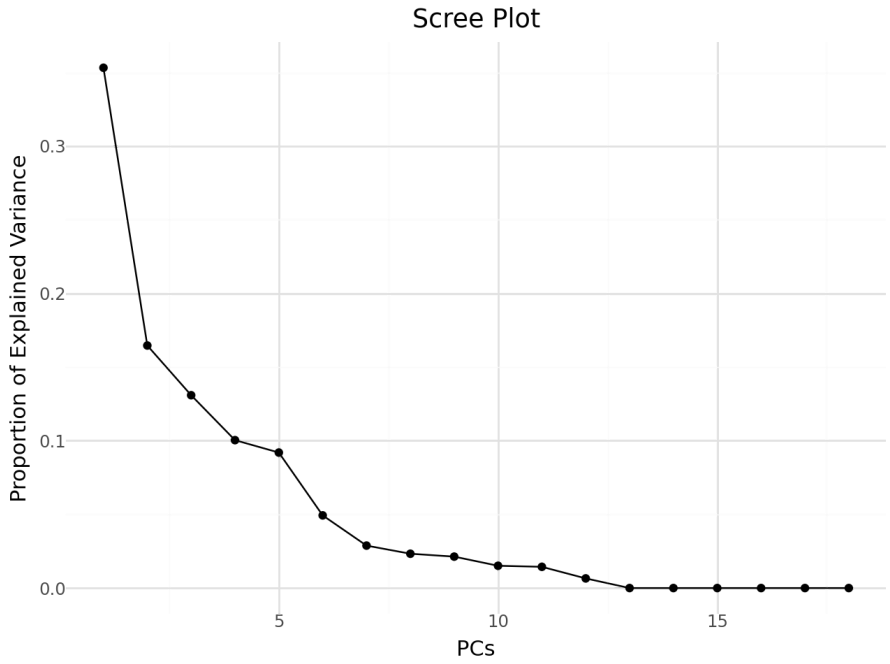
A linear regression model was created to compare the performance of two approaches for predicting Final Weight: one using PCA and another using the original continuous variables. The predictors included BMR, Daily Caloric Surplus/Deficit, and Stress Level. The PCA approach involved transforming these continuous predictors into principal components, reducing dimensionality while retaining most of the variance.

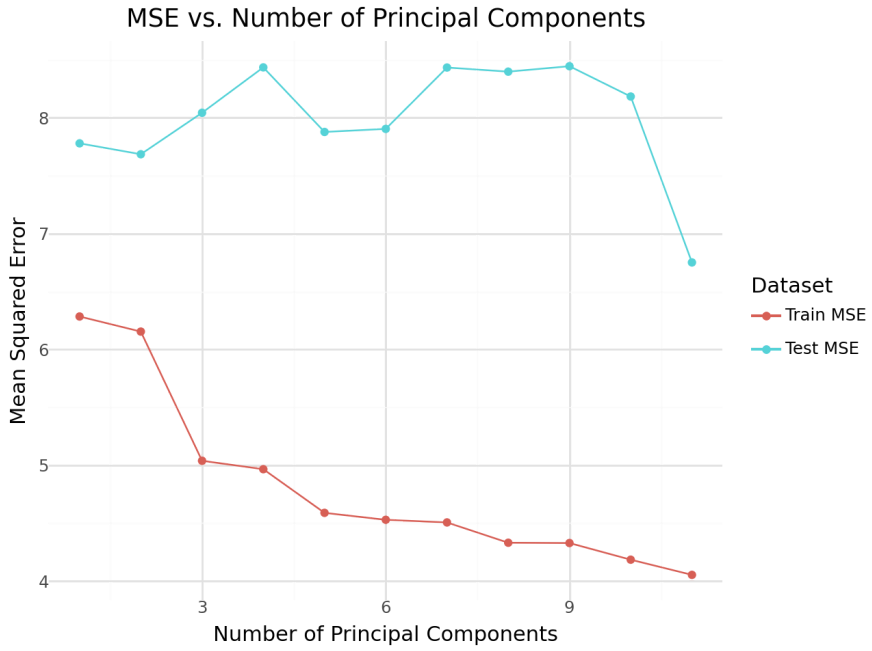
The dataset was preprocessed by standardizing continuous variables with StandardScaler to ensure all features were on a comparable scale. Using the scree plot, it was determined that the first 10 principal components explained over 90% of the variance. These 10 components were used as predictors in the PCA model. For the model with original variables, all predictors were retained without dimensionality reduction.

The dataset was split into training and testing sets (80/20), and models were evaluated using the MSE on the test set. Visualization tools, such as the scree plot and cumulative variance plot, were employed to justify the selection of principal components. The coefficients of each predictor in the original variables model were printed for comparison. The performance of both models was assessed and compared based on their respective MSE values.

Results


Cumulative Variance Plot
Cumulative Proportion of Explained Variance
1.00
0.75
0.50
0.25
0.00
0
5
10
15
Number of PCs


Scree Plot
Proportion of Explained Variance
0.3
0.2
0.1
0.0
5
10
15
PCs



Discussion

A PCA pipeline was implemented to reduce the dimensionality of the continuous variables. The scree plot shows the proportion of variance explained by each principal component, with the first few PCs accounting for the majority of the variance. The steep drop in the scree plot after the first three components indicates that these components capture the most significant patterns in the data, with diminishing returns from including additional components. The cumulative variance plot complements this by illustrating the cumulative proportion of explained variance as more PCs are added. By the 10th PC, over 90% of the variance in the dataset is retained, which guided the decision to use the first 10 components in the PCA model. These plots demonstrate that PCA effectively condenses the information from the original variables into fewer components, balancing dimensionality reduction and variance retention.

Using the PCA-transformed data, a linear regression model was trained to predict Final Weight. The model achieved a test MSE of 3.85, indicating that it effectively captured key patterns in the data while reducing redundancy among predictors.

The second model retained all original continuous variables without dimensionality reduction. While this approach allowed the model to use the full complexity of the dataset, it also increased the risk of multicollinearity, where highly correlated predictors could distort the regression coefficients. Despite this, the model achieved a slightly lower test MSE of 3.75, performing marginally better than the PCA model.

The scree and cumulative variance plots show that the PCA model does a good job of simplifying the data by reducing the number of variables while still keeping most of the important information. However, the small difference in test MSE (3.85 for PCA vs. 3.75 for the original variables) suggests that the original variables kept some details that the PCA model might have missed. This means that while PCA makes the dataset simpler and reduces issues like multicollinearity, it might lose some accuracy if the original features contain complex patterns.

The PCA model is especially useful when the goal is to make the data easier to work with or to avoid problems caused by highly related variables. By turning correlated variables into independent components, PCA helps create a more stable model. On the other hand, the model using original variables might be better for tasks where achieving the highest possible accuracy is more important, as it keeps all the original information intact.

The "MSE vs. Number of Principal Components" graph shows how the Mean Squared Error (MSE) changes with the number of principal components in the PCA-based model for predicting final weight. The training MSE (red line) decreases steadily as more components are added, stabilizing around 9–10 components, indicating that increasing the number of components improves the model's fit by capturing more variance in the data. The test MSE (blue line) fluctuates when fewer components are included, peaking with too few components due to the loss of critical variance, but decreases and stabilizes around 9-10 components. This shows that using 9-10 principal components balances variance retention and overfitting, as these components explain over 90% of the data’s variance based on the cumulative variance plot.

This graph is significant to the analysis because it supports the decision to use 10 components for the PCA model, minimizing the test MSE while ensuring the train MSE is reasonably low. It also highlights that including fewer components risks underfitting, while adding more components beyond 10 offers negligible improvement. For the analysis question, this shows the effectiveness of PCA in reducing dimensionality while retaining predictive power, demonstrating that the PCA model performs efficiently when tuned to the optimal number of components. This insight strengthens the comparison to the original variables model by showing how PCA balances simplicity and accuracy in predictive tasks.