EXPLORING BMI FACTORS

Abstract

This project explores the potential of unconventional factors for Body Mass Index (BMI) prediction using predictive analytics. Due to limitations in primary data collection (less than 50 responses), the project pivoted to a secondary dataset obtained through Kaggle.

The core objective was to develop a model that leverages less-explored factors to predict BMI. The project employed a multifaceted approach to data analysis, which includes:

1. **Predictive Modeling**: Developing a predictive model to forecast BMI based on the dataset.
2. **Hypothesis Testing**: Performing hypothesis tests to understand the relationships within the data.
3. **Clustering Analysis**: Grouping variables to identify patterns and using these clusters for further analysis.
4. **Principal Component Analysis (PCA)**: Reducing dimensionality and highlighting the most significant variables.
5. **Scott-Knott Analysis**: Classifying the data into distinct groups based on statistical significance.
6. **Factor Analysis**: Identifying underlying factors that influence BMI.

By focusing on unconventional factors and employing a comprehensive data analysis approach, this project presents a unique contribution to the field of BMI prediction. This approach offers the potential to improve prediction accuracy by incorporating a wider range of variables that may reflect an individual's overall lifestyle and health habits.

DATASET:

Dependent variable:

BMI (Body Mass Index) serves as our dependent variable, meaning it's influenced by other factors. BMI is calculated from an individual's height and weight and categorizes them as underweight, normal weight, overweight, or obese. The project considers changes in physical activity level, sleep duration, and dietary habits to affect BMI.

Independent variables:

- Physical activity level: This represents how active each person is, ranging from sedentary (low activity) to highly active. An individual's activity level can impact their BMI and overall health.
- Sleep duration: The amount of sleep each person gets on average per night. Sleep duration can influence weight management and metabolism, thereby affecting BMI.
- Soft drinks: The frequency of consuming sugary beverages. High consumption of soft drinks can contribute to weight gain and obesity.
- Fatty/oily foods: How often individuals consume foods high in fats and oils. Regular consumption of fatty foods can affect BMI and overall health.
- Sugary foods: The frequency of consuming foods high in sugar. Regular consumption of sugary foods can contribute to weight gain and influence BMI.

e	Gender	Height	Weight	BMI	ObesityCa	PhysicalAc	Sleep Dura	Soft drinks	fatty/oily f	Sugary Foods
56	Male	173.5753	71.98205	23.8918	Normal we	4	8	0	2	3
69	Male	164.1273	89.95926	33.3952	Obese	2	4	3	5	4
46	Female	168.0722	72.93063	25.8177	Overweigh	4	5	5	5	4
32	Male	168.4596	84.88691	29.9122	Overweigh	3	4	3	4	4
60	Male	183.5686	69.03895	20.4879	Normal we	3	7	0	2	2
25	Female	166.4056	61.14587	22.0816	Normal we	4	10	0	2	3
78	Male	183.5663	92.20852	27.3643	Overweigh	3	5	4	5	5
38	Male	142.8751	59.35975	29.079	Overweigh	1	6	5	5	5
56	Male	183.4786	75.15767	22.3256	Normal we	4	10	0	2	2
75	Male	182.9741	81.53346	24.3532	Normal we	2	8	0	3	2
36	Male	179.0225	82.62239	25.78	Overweigh	4	6	4	3	3
40	Female	149.8808	52.51836	23.3786	Normal we	1	8	1	3	3
28	Male	180.1889	85.77926	26.4196	Overweigh	1	4	4	3	4
28	Male	169.4988	55.31567	19.2537	Normal we	1	9	1	3	3
41	Male	144.7066	82.16055	39.2362	Obese	1	6	3	5	4
70	Male	182.9818	78.0276	23.3041	Normal we	1	10	1	2	2
53	Male	184.4417	82.27946	24.1865	Normal we	2	10	2	2	3
57	Female	150.9549	51.92495	22.7867	Normal we	3	7	2	3	2
41	Male	171.7542	78.81895	26.7187	Overweigh	1	5	3	3	4
20	Female	183.8859	86.28391	25.5172	Overweigh	1	5	4	4	4
39	Female	182.3016	72.90094	21.93573	Normal we	4	7	1	3	3
70	Male	178.3606	57.72075	18.14404	Underweig	2	9	1	2	0
19	Male	143.5275	83.79621	40.67751	Obese	2	6	3	5	4
41	Male	177.2954	71.70879	22.81273	Normal we	3	9	1	3	3
61	Male	167.819	44.88334	15.93687	Underweig	2	9	0	2	1
47	Female	179.9586	56.55716	17.46395	Underweig	4	10	0	1	0

TEST:

DEFINE: The script consists of two parts, each performing a hypothesis test:

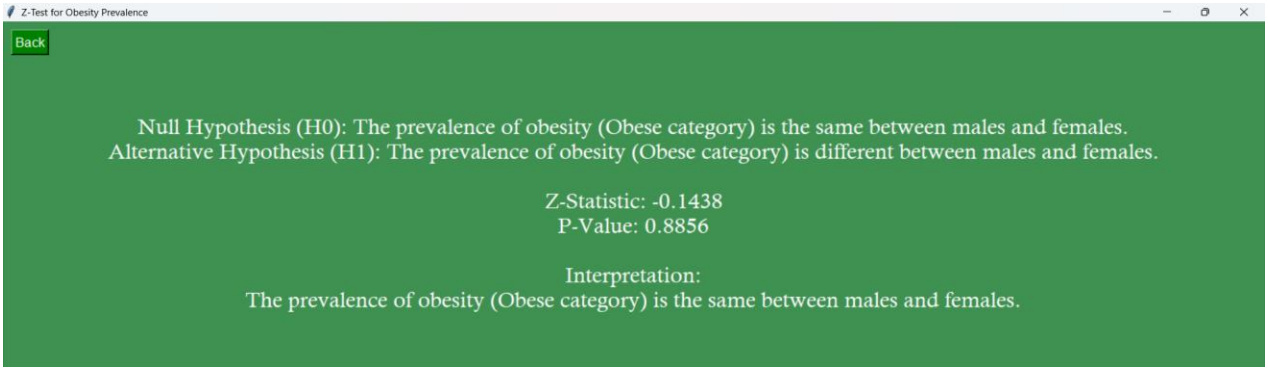
1. **Z-Test for Obesity Prevalence:** This part calculates the prevalence of obesity (Obese category) among males and females using a z-test and determines if there's a significant difference in prevalence between genders.

2. **ANOVA Test for BMI:** This part performs an ANOVA (Analysis of Variance) test to examine if there's a significant difference in mean BMI among individuals with different levels of various factors like sleep duration, physical activity, and consumption of sugary foods, fatty/oily foods, and soft drinks.

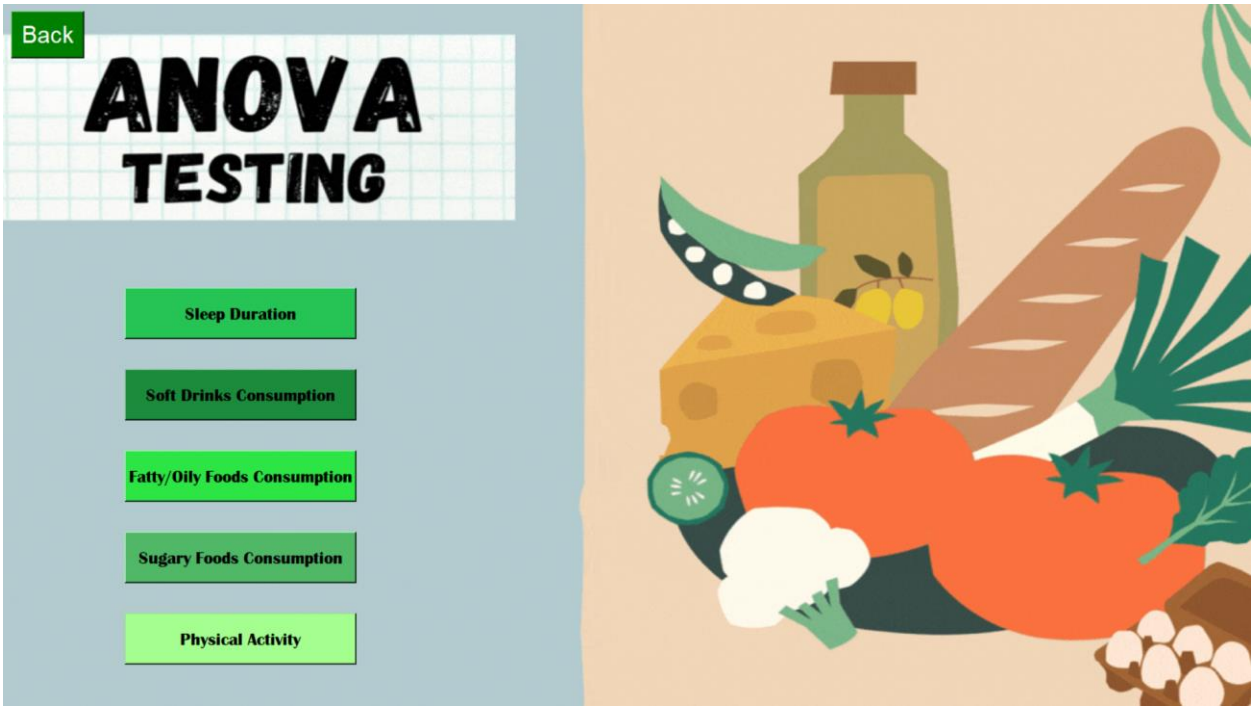
What we have done:

1. **Z-Test for Obesity Prevalence:**
    - Read data from a CSV file containing obesity data.
    - Filtered the data for individuals classified as "Obese".
    - Counted the number of obese individuals by gender.
    - Counted the total number of males and females.
    - Performed a two-sample proportion z-test to compare the prevalence of obesity between males and females.
2. **ANOVA Test for BMI:**
    - Read data from a CSV file containing obesity data.
    - Performed an ANOVA test to compare the mean BMI among individuals with different levels of sleep duration, physical activity, and consumption of sugary foods, fatty/oily foods, and soft drinks.

Screenshot of Output:




Statistical Data
TESTING
Z-Test
ANOVA


Back
ANOVA
TESTING
Sleep Duration
Soft Drinks Consumption
Fatty/Oily Foods Consumption
Sugary Foods Consumption
Physical Activity

ANOVA Test for Sleep Duration

Null Hypothesis (H0): There is no significant difference in mean BMI among individuals with different levels of Sleep Duration.
Alternative Hypothesis (H1): There is a significant difference in mean BMI among individuals with different levels of Sleep Duration.

F-Statistic: 274.3370100060918

P-Value: 2.600765105673935 6e-79

Interpretation:
The prevalence of obesity (Obese category) is different among individuals with different levels of Sleep Duration.

ANOVA Test for Soft drinks

Null Hypothesis (H0): There is no significant difference in mean BMI among individuals with different levels of Soft drinks.
Alternative Hypothesis (H1): There is a significant difference in mean BMI among individuals with different levels of Soft drinks.

F-Statistic: 389.3321882226379

P-Value: 1.1632577974400965e-103

Interpretation:
The prevalence of obesity (Obese category) is different among individuals with different levels of Soft drinks.

ANOVA Test for fatty/oily foods

Null Hypothesis (H0): There is no significant difference in mean BMI among individuals with different levels of fatty/oily foods.
Alternative Hypothesis (H1): There is a significant difference in mean BMI among individuals with different levels of fatty/oily foods.

F-Statistic: 466.4933933081206

P-Value: 6.241693472245509e-124

Interpretation:
The prevalence of obesity (Obese category) is different among individuals with different levels of fatty/oily foods.

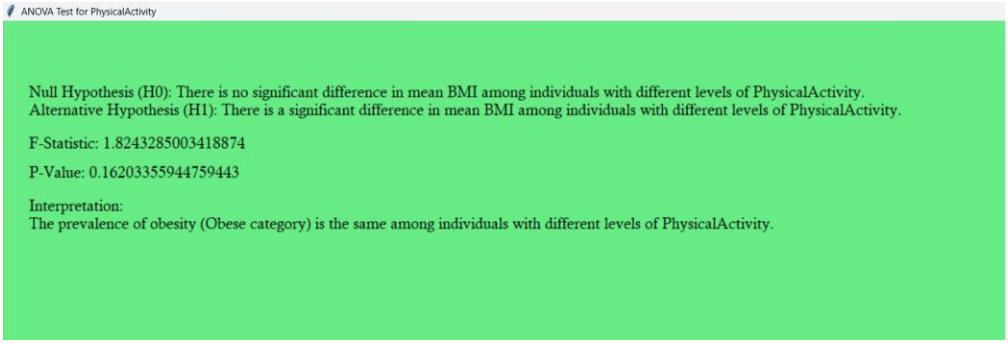
ANOVA Test for Sugary Foods

Null Hypothesis (H0): There is no significant difference in mean BMI among individuals with different levels of Sugary Foods.
Alternative Hypothesis (H1): There is a significant difference in mean BMI among individuals with different levels of Sugary Foods.

F-Statistic: 368.25797014662675

P-Value: 1.0283057644721565e-108

Interpretation:
The prevalence of obesity (Obese category) is different among individuals with different levels of Sugary Foods.



INFERENCE:

The scripts provide a user-friendly way to perform hypothesis tests on obesity prevalence and BMI using graphical user interfaces. Users can input their data and quickly analyze the prevalence of obesity among different genders and the variation in BMI among individuals with different characteristics. These tools can be useful for researchers, healthcare professionals, and anyone interested in understanding obesity-related trends and factors.

CLUSTERING & PCA:

DEFINE:

The script performs clustering analysis on obesity-related data using the KMeans algorithm. It identifies clusters of individuals based on features like BMI, physical activity, sleep duration, and consumption of soft drinks, fatty/oily foods, and sugary foods.Utilized PCA to reduce the dimensionality of the dataset while retaining variance.

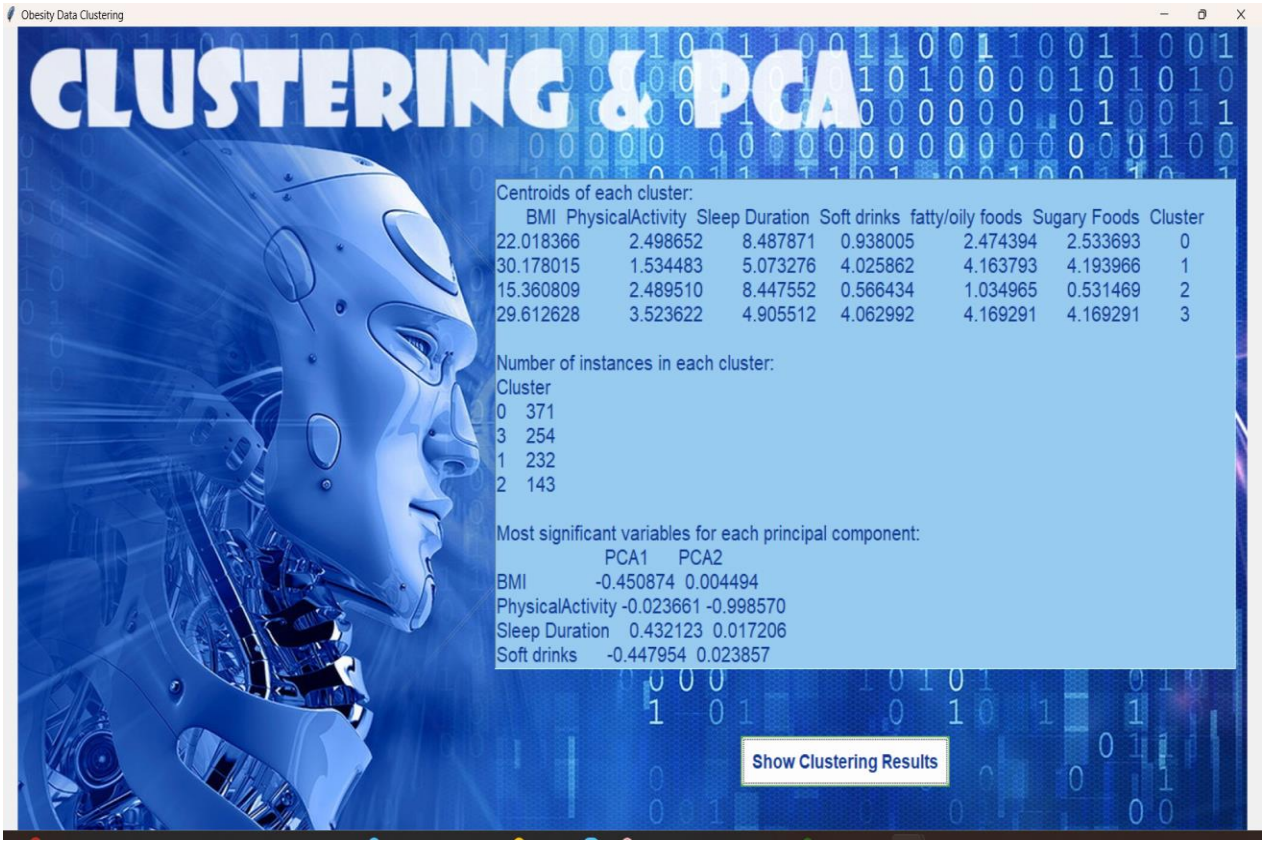
What we have done:

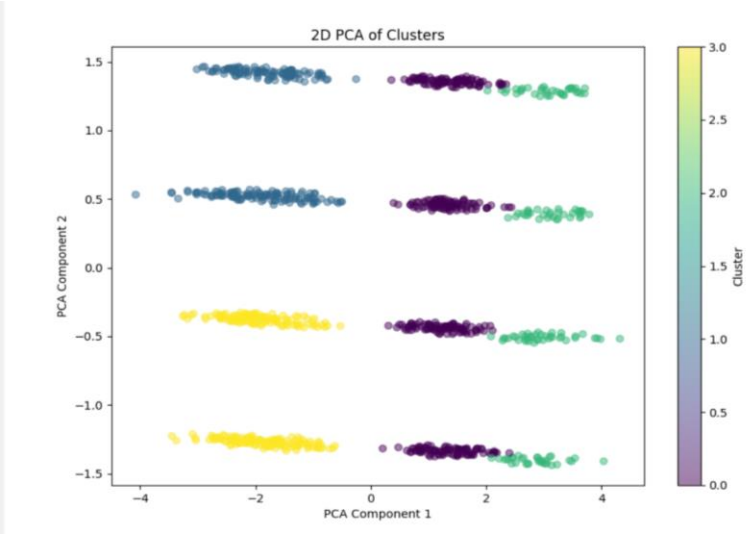
1. **Data Preprocessing:**
    - Loaded the data from a CSV file.
    - Selected relevant features for clustering.
    - Standardized the features to have a mean of 0 & a standard deviation of 1
2. **Clustering:**
    - Applied the KMeans algorithm with 4 clusters to the standardized features.
    - Assigned each data point to a cluster based on its similarity to cluster centroids.

- Calculated centroids for each cluster.

3. **Visualization:**
    - Performed PCA (Principal Component Analysis) to reduce the dimensionality of the data to 2 components.
    - Plotted the data points in a 2D space using PCA components.
    - Colored the data points based on their assigned clusters.
    - Displayed cluster centroids and summary statistics in a GUI window.
    - Provided a button to show a popup window with the PCA plot.

Screenshot of Output:

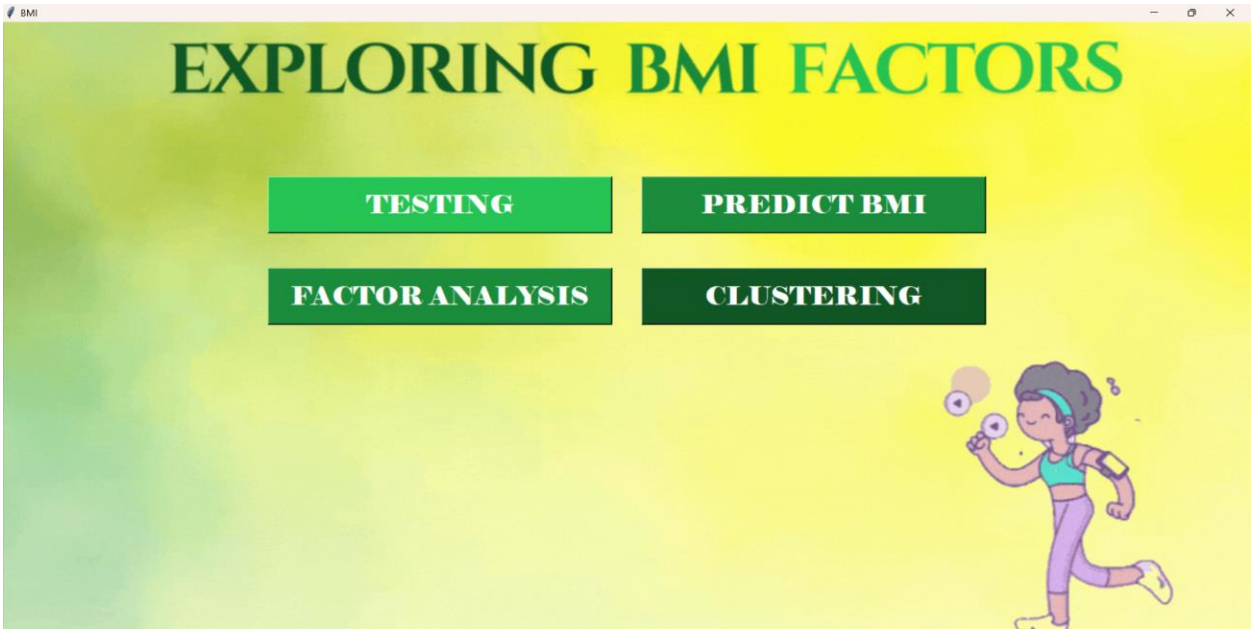


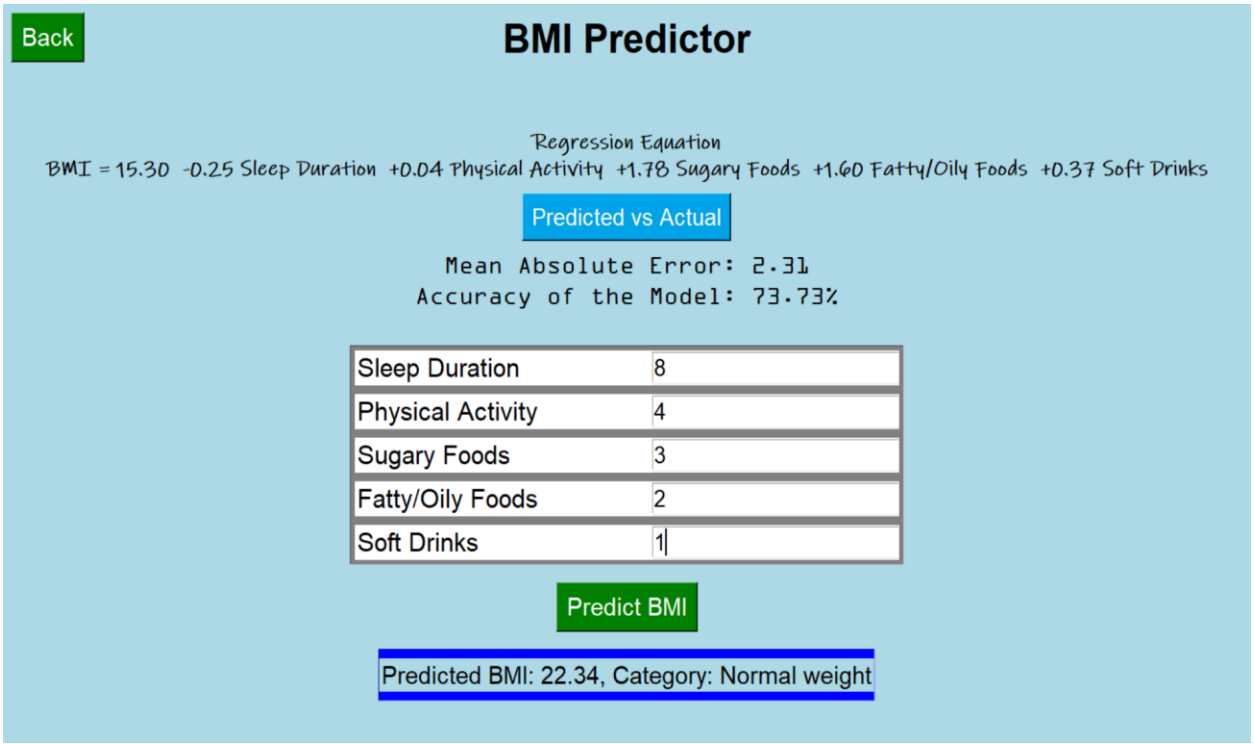


Inference:

In this analysis, PCA was performed within the context of clusters to gain deeper insights into the variability and underlying structure of the data within each cluster. By applying PCA within clusters, we aimed to explore the relationships between features specific to each group, facilitating a more focused interpretation of cluster composition. This approach not only helps in identifying the principal components driving variance within individual clusters but also aids in visualizing the distribution of data points in a reduced-dimensional space, colored by cluster membership. By leveraging PCA within clusters, we can uncover nuanced patterns and relationships that may not be apparent when analyzing the entire dataset collectively, thus providing valuable insights for understanding the heterogeneity of obesity-related factors within distinct subgroups.

Gui Screenshots:






Factor Analysis
SHOW SCREE PLOT
SHOW FACTOR LOADINGS
SHOW INTERPRETATION

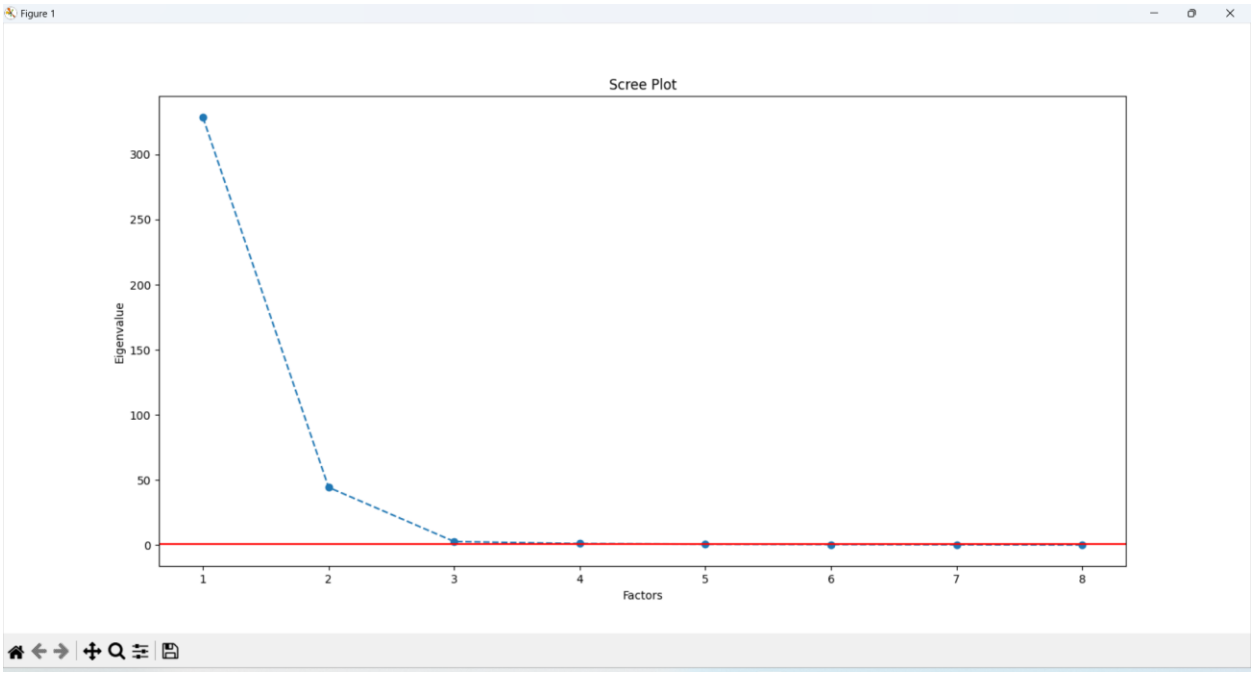


Figure 1
Scree Plot
Eigenvalue
300
250
200
150
100
50
0
1
2
3
4
5
6
7
8
Factors


```
Interpretation:
Number of factors selected using Kaiser criterion: 4


Factor1 Loadings:
Sugary Foods         0.901425
BMI                  0.885514
fatty/oily foods     0.861100
Soft drinks          0.788626
PhysicalActivity     0.001628
Gender              -0.010862
Age                 -0.056291
Sleep Duration      -0.718988
Factor2 Loadings:
Gender               0.978847
Sleep Duration       0.015579
Age                  0.015441
fatty/oily foods     0.014791
BMI                  0.003633
Soft drinks         -0.033409
PhysicalActivity    -0.038796
Sugary Foods        -0.043471
Factor3 Loadings:
Soft drinks          0.431893
```