CSE:587 Data Intensive Computing
Project-phase-1

REPORT (Team No: 25)

Team Members

Student Name	Student UB Number
Karthik Sharma Madugula	50611293
Santosh Kota	50593968
Harshita Sherla	50593920
Riya Agarwal	50609491

Question 1

Explain the potential of your project to contribute to your problem domain. Discuss why this contribution is crucial.

This project will give personalized dietary insights by analyzing users' calorie intake. It will give recommendations to help improve eating habits and promote a better healthy lifestyle. This is crucial because good dietary habits are essential to preventing diseases like diabetes, obesity, and so on. Personalized recommendations based on food intake will help people to make healthier choices.

Question 2

1st Hypothesis

Hypothesis: Carbohydrate intake correlates with calorie intake more than fat or protein intake.
Objective: Deduce the contribution of total carbs, proteins, and fats to total intake.
Significance: Identifying the contribution of carbohydrates to total calorie intake.

2nd Hypothesis

Hypothesis: Male users consume more calories than female users on average.
Objective: Test if there is a significant difference in calorie intake between genders.
Significance: This will help in customizing recommendations by gender.

3rd Hypothesis

Hypothesis: Age influences BMR and thus affects total calorie requirements.
Objective: Determine how BMR changes with age and its relationship with calorie intake.
Significance: Understanding how age influences BMR and calorie requirements is important for creating personalized nutrition plans.

4th Hypothesis

Hypothesis: People with higher BMR consume more protein.
Objective: Test whether individuals with higher BMR have a diet rich in protein.

Significance: Understanding the relationship between higher BMR and protein intake can inform personalized dietary recommendations for improved metabolism and health outcomes.

5th Hypothesis

Hypothesis: Females tend to have lower BMR compared to males.
Objective: To compare the BMR of males and females to identify any significant differences based on gender.
Significance: Understanding the relationship between gender and BMR is essential for providing personalized health recommendations and assessing metabolic differences between males and females.

6th Hypothesis

Hypothesis: Higher Physical Exercise is Associated with Higher Calorie Intake.
Objective: Test whether physical exercise is associated with higher calorie intake.
Significance: Understanding the relationship between physical exercise and calorie intake is crucial for optimizing nutrition and energy balance.

7th Hypothesis

Hypothesis: Physical exercise frequency is positively correlated with daily protein intake among individuals aged 15-30.
Objective: To investigate the relationship between the frequency of physical exercise and the frequency and daily protein intake of individuals 15-20.
Significance: To suggest specific dietary recommendations and ensure that individuals meet their nutritional needs for muscle recovery and overall health.

8th Hypothesis

Hypothesis: Body Mass Index affects Basal Metabolic Rate.
Objective: To look at the relationship between BMI and BMR in the general population and see if people with a higher BMI have a higher basal metabolic rate, independent of age.
Significance: Understanding this association allows us to adjust nutrient intake for better health outcomes.

9th Hypothesis

Hypothesis: Eating more meals each day leads to higher BMI.
Objective: To determine whether eating more frequently throughout the day is associated with a higher BMI.
Significance: Understanding if meal frequency influences body weight can lead to healthier eating habits for weight management.

Question 3

The source of the data that we acquired is Kaggle.
Source: https://www.kaggle.com/datasets/abdussamad123/user-daily-nutritional-intake

Dataset Overview

The dataset contains information on the daily nutritional intake of approximately 2,000 users. It consists of survey data collected from residents and peers.
Dimensions: 2,182 rows and 11 columns
Features:

- Gender: Male (0) or Female (1) - Indicates the subject's gender.

- Age: Ranges from 15 to 75 years - Represents the subject's age, rounded down to the nearest integer.
- Daily Meal Frequency: Ranges from 2 to 4 - Indicates the average number of daily meals consumed by the subject.
- Physical Exercise: Ranges from 0 to 4 - Represent the level of daily exercise, with 0 indicating no exercise and 4 indicating extremely heavy exercise.
- Height: Ranges from 122 cm to 188 cm - The subject's height was recorded during the survey.
- Weight: Ranges from 35 kg to 150 kg - The subject's weight was recorded during the survey.
- BMR: Ranges from 862 to 2,410 kcal - The Basal Metabolic Rate is calculated based on age, height, and weight.
- Carbs: Ranges from 129 g to 461 g - The total daily carbohydrate intake of the subject.
- Proteins: Ranges from 51 g to 184 g - The total daily protein intake of the subject.
- Fats: Ranges from 34 g to 123 g - The total daily fat intake of the subject.

Question 4: Data Cleaning Steps

1. Identified missing values by defining potential representations of missing data and used `df.isin()` to check for their presence.
2. Detected duplicate rows in the dataset using `df.duplicated()` and calculated the number of duplicates.
3. Removed duplicate rows from the DataFrame using `df.drop_duplicates()` and updated the dataset.
4. Added new columns, BMI using the formula: $\text{BMI} = \frac{\text{Weight}}{(\text{Height in meters})^2}$.
5. Replaced numeric values in the 'Physical exercise' column with descriptive labels for better interpretability.
6. Binned the 'Age' column into specific age groups using `pd.cut()` to categorize ages into defined ranges.
7. Checked the final structure of the cleaned dataset with `df.info()` to confirm successful data cleaning.

Question 5

- Riya Agarwal (50609491)

1st Hypothesis
Hypothesis: Carbohydrate intake correlates with calorie intake more than fat or protein intake.
Objective: Deduce the contribution of total carbs, proteins, and fats to total intake.
EDA operations:

- **Visualization:** Created a pair plot comparing Carbs, Proteins, Fats, and Calories to visualize their relationships. Calculated and visualized the correlation matrix between these features using a heatmap. Created a stacked bar plot showing the average macronutrient contribution to total calories.
- **Statistical Test:** Calculated the correlation coefficients between each macronutrient and total Calories. Calculated the percentage contribution of each macronutrient to total calories. Displayed the average percentage contribution of each macronutrient to total calories.
- Interpreted the results, identifying which macronutrient has the highest correlation with total calorie intake.

This analysis will help visualize and quantify the relationships between carbohydrates, proteins, fats, and total calorie intake. It will provide insights into which macronutrient contributes most significantly to overall calorie consumption and how strongly each macronutrient correlates with total calories.

2nd Hypothesis

- Riya Agarwal (50609491)

Hypothesis: Male users consume more calories than female users on average.
Objective: Test if there is a significant difference in calorie intake between genders.
EDA Operations:

- **Visualization:** Created a box plot comparing calorie intake by gender using seaborn.
- **Statistical Test:** Performed a t-test to test for differences in mean calorie intake between males and females. Calculated the t-statistic and p-value from the t-test and also the mean calorie intake for each gender. Showcased if the difference is statistically significant based on the p-value.

This analysis will help visualize the distribution of calorie intake for both genders and statistically assess whether there is a significant difference in average calorie intake.

3rd Hypothesis

- Harshita Sherla(50593920)

Hypothesis: Age influences BMR and thus affects total calorie requirements.
Objective: To investigate the relationship between age, BMR, and calorie intake.
EDA Operations:

- **Visualization:** Created scatter plots to visualize the relationship between age and BMR as well as age and total calorie intake. Used regression lines to assess trends.
- **Statistical Test:** Performed correlation analysis to quantify the relationship between age, BMR, and calorie intake. Conducted regression analysis to determine the extent to which age predicts BMR and calorie intake.

This analysis will help in understanding the relationship between age and BMR, as well as its effect on total calorie intake, enabling personalized dietary recommendations.

4th Hypothesis

- Harshita Sherla(50593920)

Hypothesis: People with higher BMR consume more protein.
Objective: Test whether individuals with higher BMR have a diet rich in protein.
EDA Operations:

- **Visualization:** Created scatter plots comparing BMR and protein intake, including regression lines to analyze the trend.
- **Statistical Test:** Conducted correlation analysis between BMR and protein intake to identify if a significant relationship exists.

This analysis will provide insights into the protein consumption patterns of individuals with varying BMRs and guide dietary recommendations for protein intake.

5th Hypothesis

- Karthik Sharma Madugula (50611293)

Hypothesis: Females tend to have lower BMR compared to males.
Objective: To compare the BMR of males and females to identify any significant differences based on gender.
EDA Operations:

- **Visualization:** Created box plots for BMR by gender to visualize any disparities.
- **Statistical Test:** Conducted a t-test to compare mean BMR values between males and females, calculating the t-statistic and p-value.

This analysis will highlight any significant differences in BMR between genders, which can inform personalized health recommendations.

6th Hypothesis

- Karthik Sharma Madugula (50611293)

Hypothesis: Higher Physical Exercise is Associated with Higher Calorie Intake.
Objective: Test whether physical exercise is associated with higher calorie intake.
EDA Operations:

- **Visualization:** Created scatter plots to visualize the relationship between physical exercise and calorie intake.
- **Statistical Test:** Performed correlation analysis to quantify the association between physical exercise and calorie intake.

This analysis will provide insights into how physical activity levels may influence overall calorie intake, which can inform nutritional recommendations.

7th Hypothesis

- Santosh Kota (50593968)

Hypothesis: Physical exercise frequency is positively correlated with daily protein intake among individuals aged 15-30.
Objective: To investigate the relationship between the frequency of physical exercise and the frequency and daily protein intake of individuals aged 15-20.
EDA Operations:

- **Visualization:** Created scatter plots to examine the relationship between physical exercise frequency and daily protein intake.
- **Statistical Test:** Conducted correlation analysis to assess the strength and direction of the relationship between these two variables.

This analysis will shed light on the dietary habits of younger individuals who engage in physical exercise, helping to make informed dietary recommendations.

8th Hypothesis

- Santosh Kota (50593968)

Hypothesis: Body Mass Index affects Basal Metabolic Rate.
Objective: To look at the relationship between BMI and BMR in the general population.
EDA Operations:

- **Visualization:** Created scatter plots comparing BMI and BMR.
- **Statistical Test:** Conducted regression analysis to determine the extent to which BMI predicts BMR.

This analysis will help us understand the relationship between BMI and BMR, which is important for personalized nutrition plans.

9th Hypothesis

- Karthik Sharma Madugula (50611293)

Hypothesis: Eating more meals each day leads to higher BMI.
Objective: To determine whether eating more frequently throughout the day is associated with a higher BMI.
EDA Operations:

- **Visualization:** Created box plots to visualize the relationship between daily meal frequency and BMI.
- **Statistical Test:** Conducted correlation analysis to assess the relationship between meal frequency and BMI.

This analysis will help determine if meal frequency influences body weight, which can inform dietary recommendations for weight management.

Data Cleaning

Check for the null:

values when we work on the data analysis one of the main constraints is data set having null values so in the initial step itself we makesure that thare are no null values.

Removing duplicates:

when the dataset contains duplicate values the we wil get skewed results, incorrect model predictions and it will increase the size of data set unnecessarily

Transformation data

To ensure analysis, we addressed inconsistencies in the dataset by transforming the data. This allowed us to conduct a more reliable hypothesis analysis based on cons.

Added new columns

We added a new column to calculate the BMI, which enhances the efficiency and accuracy of our data analysis. This additional metric provides deeper insights into the dataset.

Handled Categorical Data:

We changed some of the categorical data so that we are able to perform the operation on the dataset.

In [9]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from scipy.stats import pearsonr

df = pd.read_csv("user_nutritional_data.csv")
df.info()
#Identifing missing values by defining potential representations of missi
missing_values = ["NA", np.nan, " ", None]
missing= df.isin(missing_values)
missing.head()

#Removed duplicate rows from the DataFrame using df.drop_duplicates() and
duplicate_rows = df[df.duplicated()]
len(duplicate_rows)

df = df.drop_duplicates()
```

```
#Added new columns, BMI using the formula: BMI = Weight / (Height in mete
df["BMI"] = df["Weight"] / ((df["Height"]/100) ** 2)

#Replaced numeric values in the 'Physical exercise' column with descripti
df['Physical exercise'] = df['Physical exercise'].replace({
    0: 'None',
    1: 'Low',
    2: 'Moderate',
    3: 'High',
    4: 'Very High'
})

#Binned the 'Age' column into specific age groups using pd.cut() to categ
bins = [0, 18, 30, 45, 60, 100]
labels = ['0-18', '19-30', '31-45', '46-60', '60+']
df['Age Group'] = pd.cut(df['Age'], bins=bins, labels=labels, right=False

df.info()
df.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2182 entries, 0 to 2181
Data columns (total 11 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   Gender                 2182 non-null   int64  
 1   Age                    2182 non-null   int64  
 2   Daily meals frequency  2182 non-null   int64  
 3   Physical exercise      2182 non-null   int64  
 4   Height                 2182 non-null   int64  
 5   Weight                 2182 non-null   float64
 6   BMR                    2182 non-null   float64
 7   Carbs                  2182 non-null   float64
 8   Proteins               2182 non-null   float64
 9   Fats                   2182 non-null   float64
 10  Calories               2182 non-null   float64
dtypes: float64(6), int64(5)
memory usage: 187.6 KB
<class 'pandas.core.frame.DataFrame'>
Index: 2098 entries, 0 to 2181
Data columns (total 13 columns):
 #   Column                 Non-Null Count  Dtype   
---  ------                 --------------  -----   
 0   Gender                 2098 non-null   int64   
 1   Age                    2098 non-null   int64   
 2   Daily meals frequency  2098 non-null   int64   
 3   Physical exercise      2098 non-null   object  
 4   Height                 2098 non-null   int64   
 5   Weight                 2098 non-null   float64 
 6   BMR                    2098 non-null   float64 
 7   Carbs                  2098 non-null   float64 
 8   Proteins               2098 non-null   float64 
 9   Fats                   2098 non-null   float64 
 10  Calories               2098 non-null   float64 
 11  BMI                    2098 non-null   float64 
 12  Age Group              2098 non-null   category
dtypes: category(1), float64(7), int64(4), object(1)
memory usage: 215.3+ KB
```

Out[9]:

	Gender	Age	Daily meals frequency	Physical exercise	Height	Weight	BMR	Carbs	Proteins
0	0	29	3	None	165	101.0	1901.25	285.188	114.075
1	1	25	3	Very High	165	53.0	1275.25	302.872	121.149
2	0	23	2	None	170	70.0	1652.50	247.875	99.150
3	0	22	3	None	168	112.0	2065.00	309.750	123.900
4	0	19	3	Moderate	175	67.0	1673.75	324.289	129.716

Hypothesis-1:

Carbohydrate intake correlates with calorie intake more thea fat or protien intake.

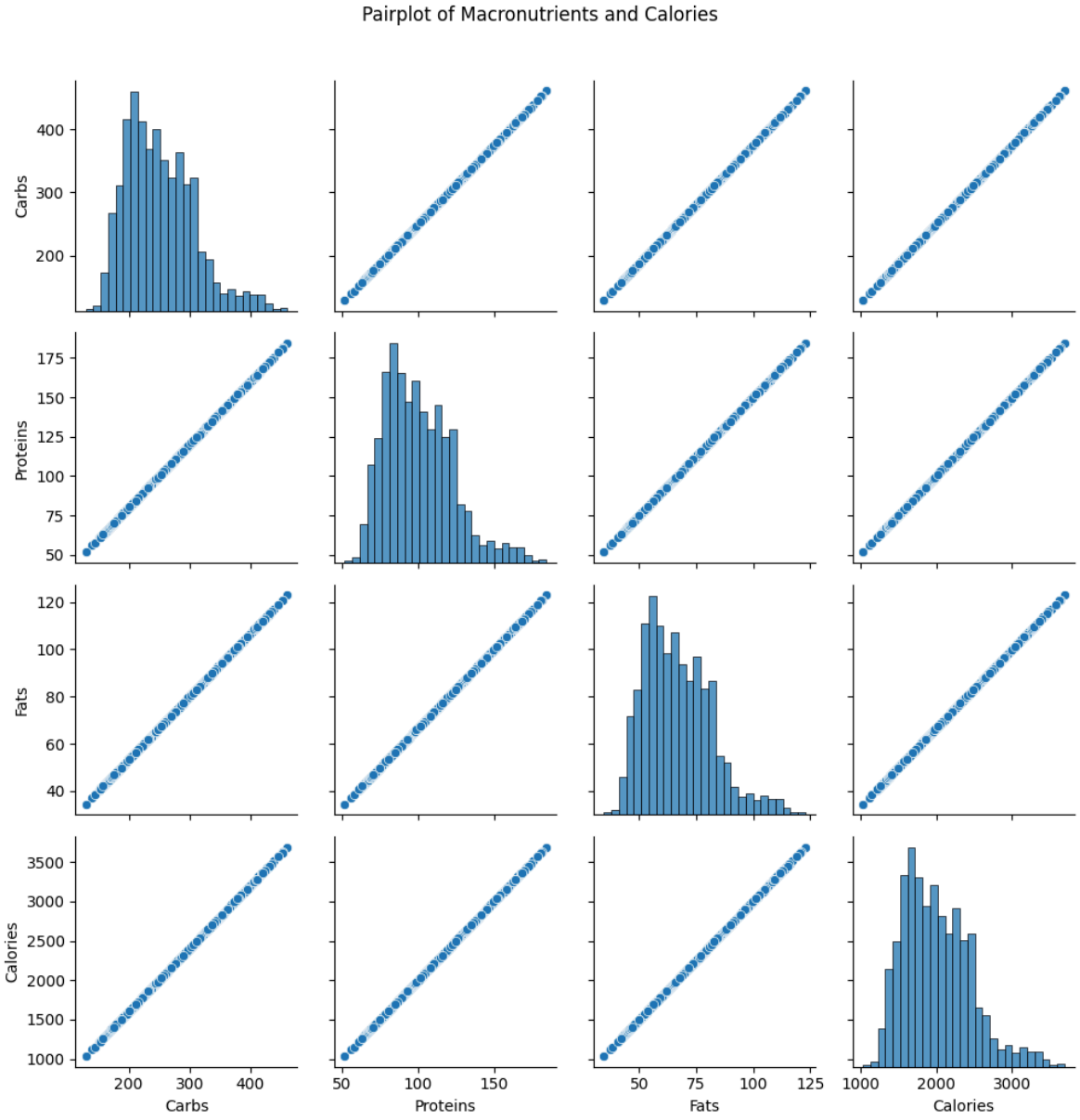
Objective: Deduce the contribution of total carbs proteins and fats to total intake

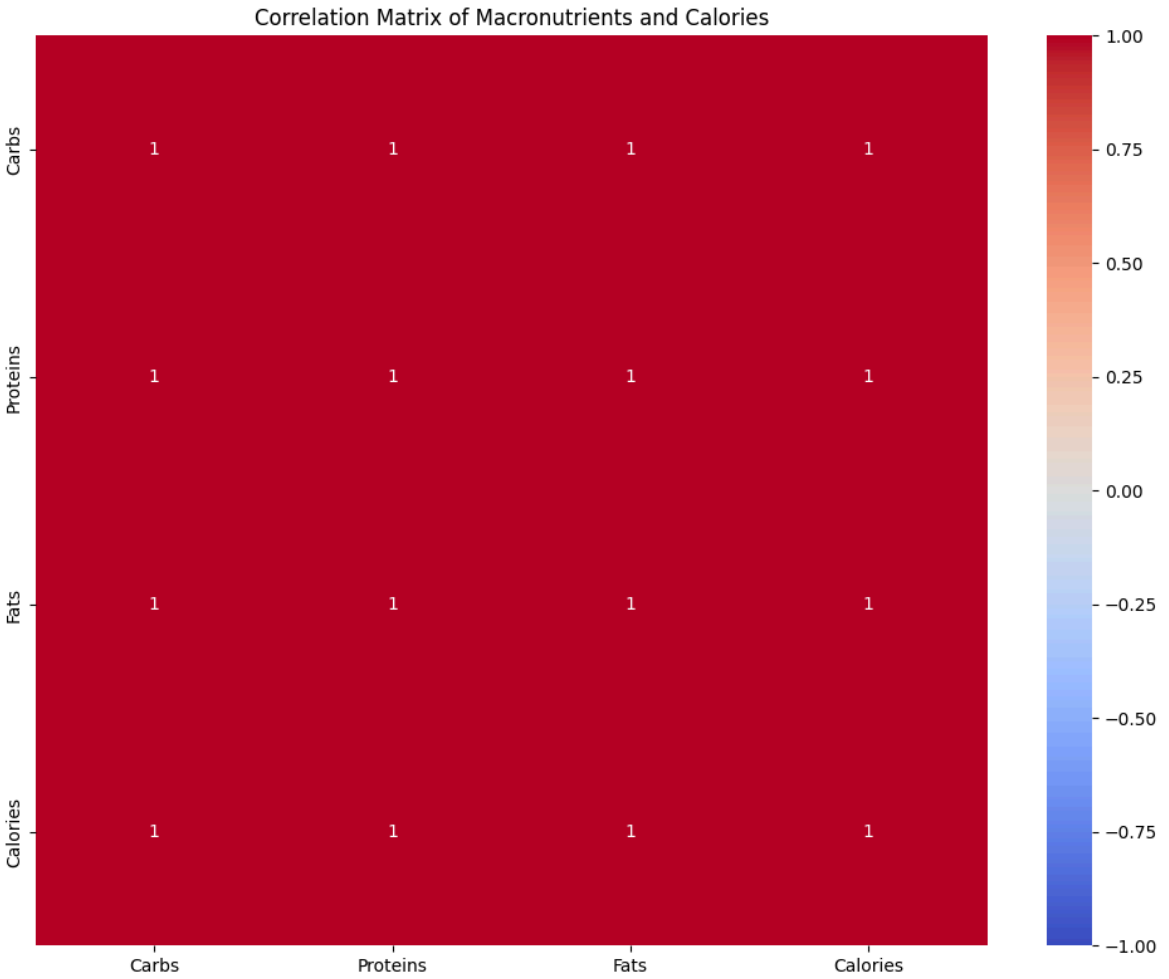
Riya Agarwal (50609491)

In [10]:
```
plt.figure(figsize=(12, 10))
sns.pairplot(df[['Carbs', 'Proteins', 'Fats', 'Calories']])
plt.suptitle('Pairplot of Macronutrients and Calories', y=1.02)
plt.tight_layout()
plt.show()

correlation_matrix = df[['Carbs', 'Proteins', 'Fats', 'Calories']].corr()

plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', vmin=-1, vma
plt.title('Correlation Matrix of Macronutrients and Calories')
plt.tight_layout()
plt.show()

correlations_with_calories = correlation_matrix['Calories'].sort_values(a
print("Correlation coefficients with Calories:")
print(correlations_with_calories)

df['Carbs_Calories'] = df['Carbs'] * 4
df['Proteins_Calories'] = df['Proteins'] * 4
df['Fats_Calories'] = df['Fats'] * 9

df['Carbs_Percentage'] = df['Carbs_Calories'] / df['Calories'] * 100
df['Proteins_Percentage'] = df['Proteins_Calories'] / df['Calories'] * 10
df['Fats_Percentage'] = df['Fats_Calories'] / df['Calories'] * 100

average_percentages = df[['Carbs_Percentage', 'Proteins_Percentage', 'Fat
print("\nAverage percentage contribution to total calories:")
print(average_percentages)

plt.figure(figsize=(10, 6))
average_percentages.plot(kind='bar', stacked=True)
plt.title('Average Macronutrient Contribution to Total Calories')
```

```
plt.xlabel('Macronutrients')
plt.ylabel('Percentage Contribution')
plt.legend(title='Macronutrients', bbox_to_anchor=(1.05, 1), loc='upper l
plt.tight_layout()
plt.show()

print("\nInterpretation:")
highest_correlation = correlations_with_calories.index[1]  # Index 0 is C
print(f"The macronutrient with the highest correlation to total calorie i
print(f"This suggests that {highest_correlation} intake has the strongest
```

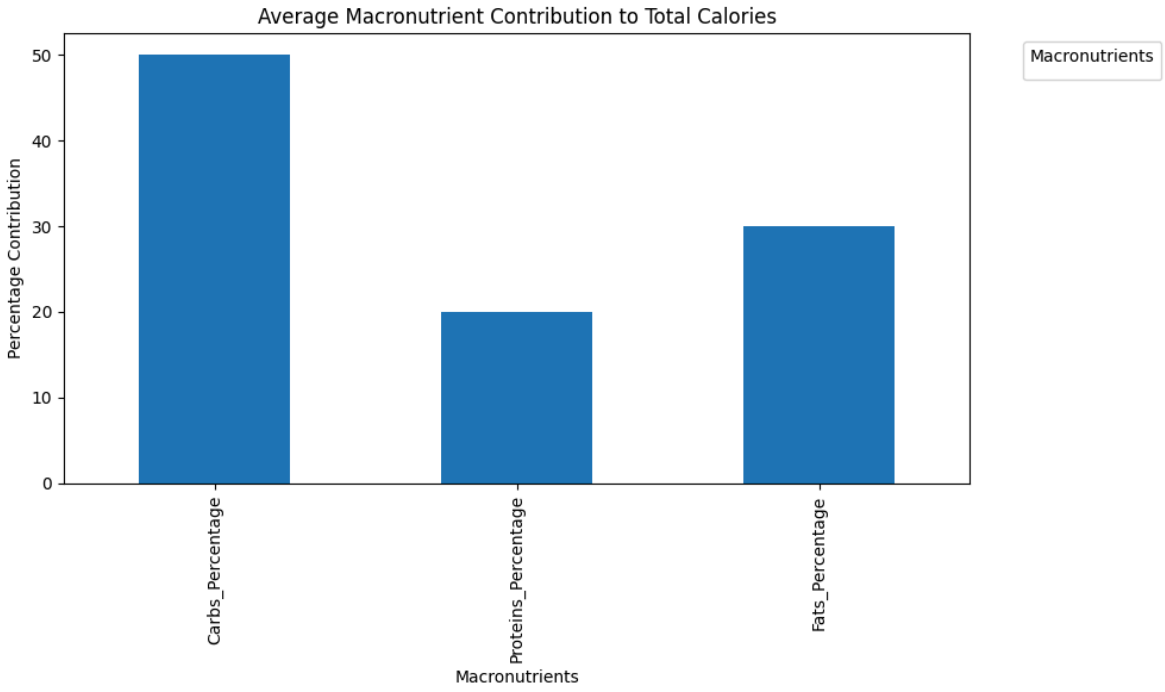
```
<Figure size 1200x1000 with 0 Axes>
```






```
Correlation coefficients with Calories:
Calories    1.0
Carbs       1.0
Proteins    1.0
Fats        1.0
Name: Calories, dtype: float64

Average percentage contribution to total calories:
Carbs_Percentage       50.000004
Proteins_Percentage    20.000000
Fats_Percentage        29.999997
dtype: float64
```

```
/var/folders/dh/v1sv1c0j4x36j10zr8nh0xlh0000gn/T/ipykernel_35196/413787755
0.py:36: UserWarning: No artists with labels found to put in legend.  Note
that artists whose label start with an underscore are ignored when legend
() is called with no argument.
  plt.legend(title='Macronutrients', bbox_to_anchor=(1.05, 1), loc='upper
left')
```




```
Interpretation:
The macronutrient with the highest correlation to total calorie intake is:
Carbs
This suggests that Carbs intake has the strongest relationship with overal
l calorie consumption.
```

Hypothesis - 2:

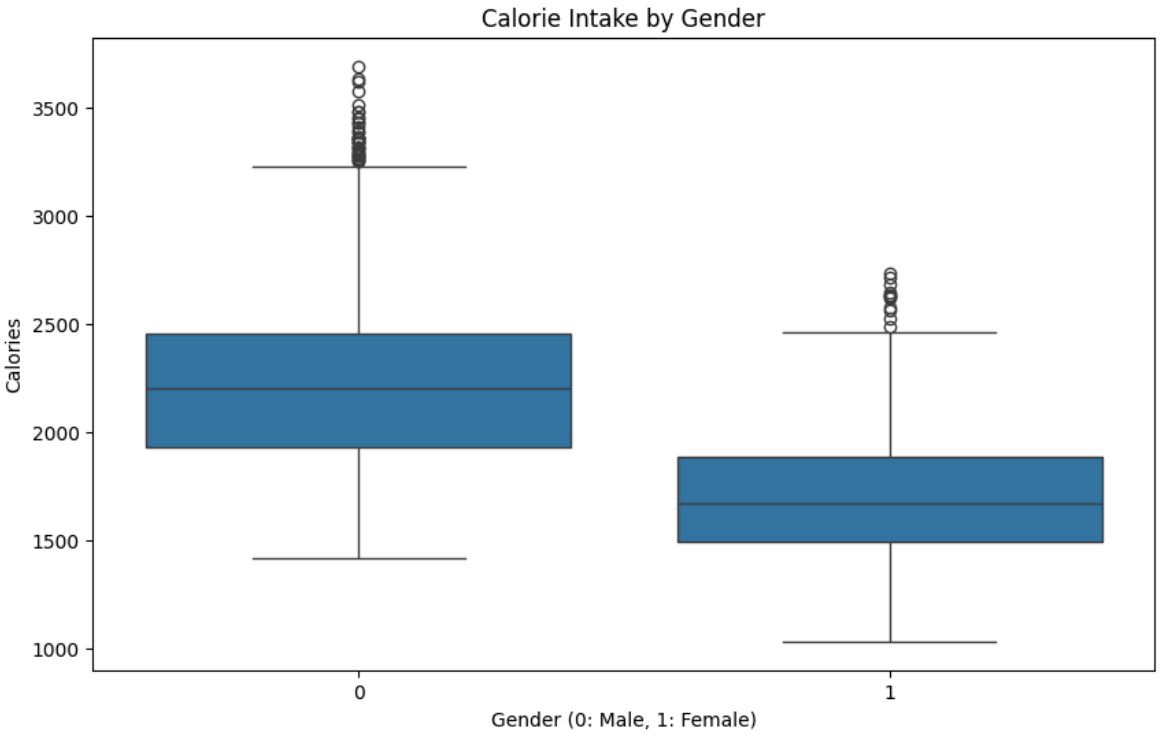
Male users consume more calories than female users on average.

Objective: Test if there is a significant difference in calorie intake between genders.

Riya Agarwal (50609491)

In [11]:
```
plt.figure(figsize=(10, 6))
sns.boxplot(x='Gender', y='Calories', data=df)
plt.title('Calorie Intake by Gender')
plt.xlabel('Gender (0: Male, 1: Female)')
plt.ylabel('Calories')
plt.show()

male_calories = df[df['Gender'] == 0]['Calories']
female_calories = df[df['Gender'] == 1]['Calories']

t_statistic, p_value = stats.ttest_ind(male_calories, female_calories)

print(f"T-statistic: {t_statistic}")
print(f"P-value: {p_value}")

male_mean = male_calories.mean()
female_mean = female_calories.mean()
print(f"Mean calorie intake for males: {male_mean:.2f}")
print(f"Mean calorie intake for females: {female_mean:.2f}")

alpha = 0.05
```

```
if p_value < alpha:
    print("The difference in calorie intake between genders is statistica
else:
    print("There is no statistically significant difference in calorie in
```




```
T-statistic: 31.346863206616725
P-value: 3.2266758500920674e-177
Mean calorie intake for males: 2225.04
Mean calorie intake for females: 1710.20
The difference in calorie intake between genders is statistically signific
ant.
```

Hypothesis-3:

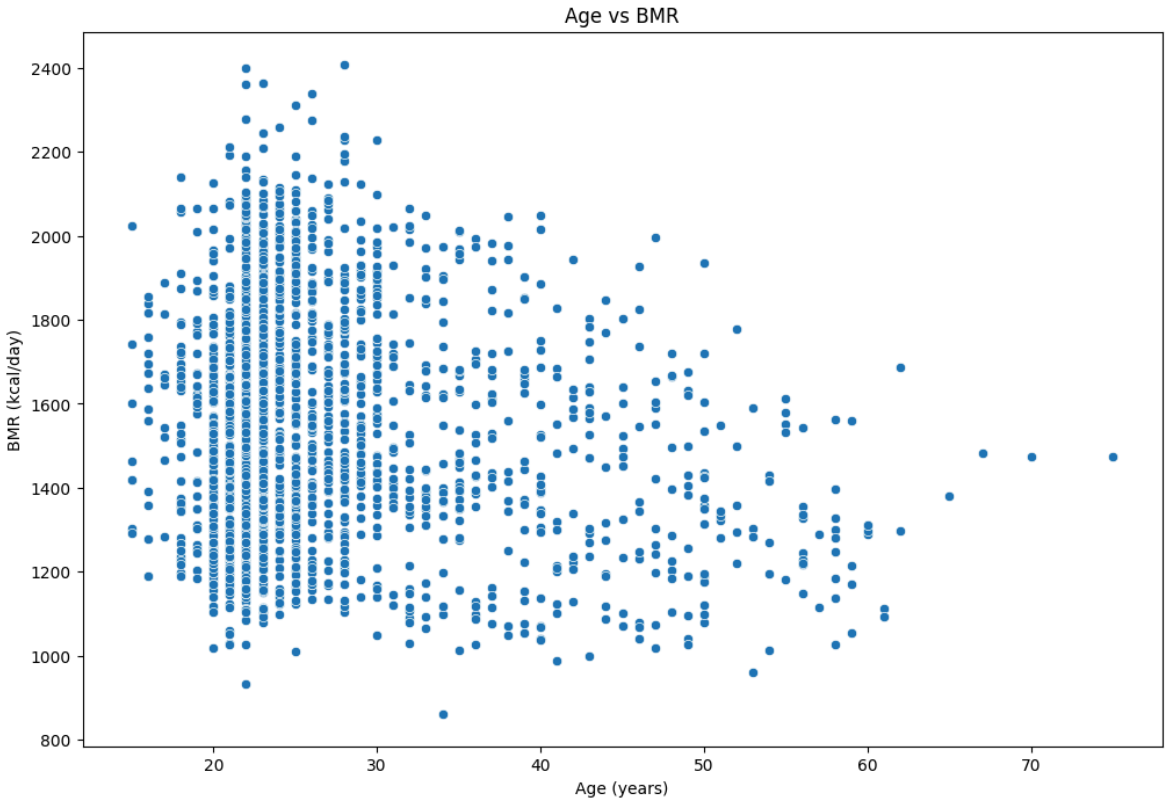
Age influences BMR and thus affects total calorie requirements.

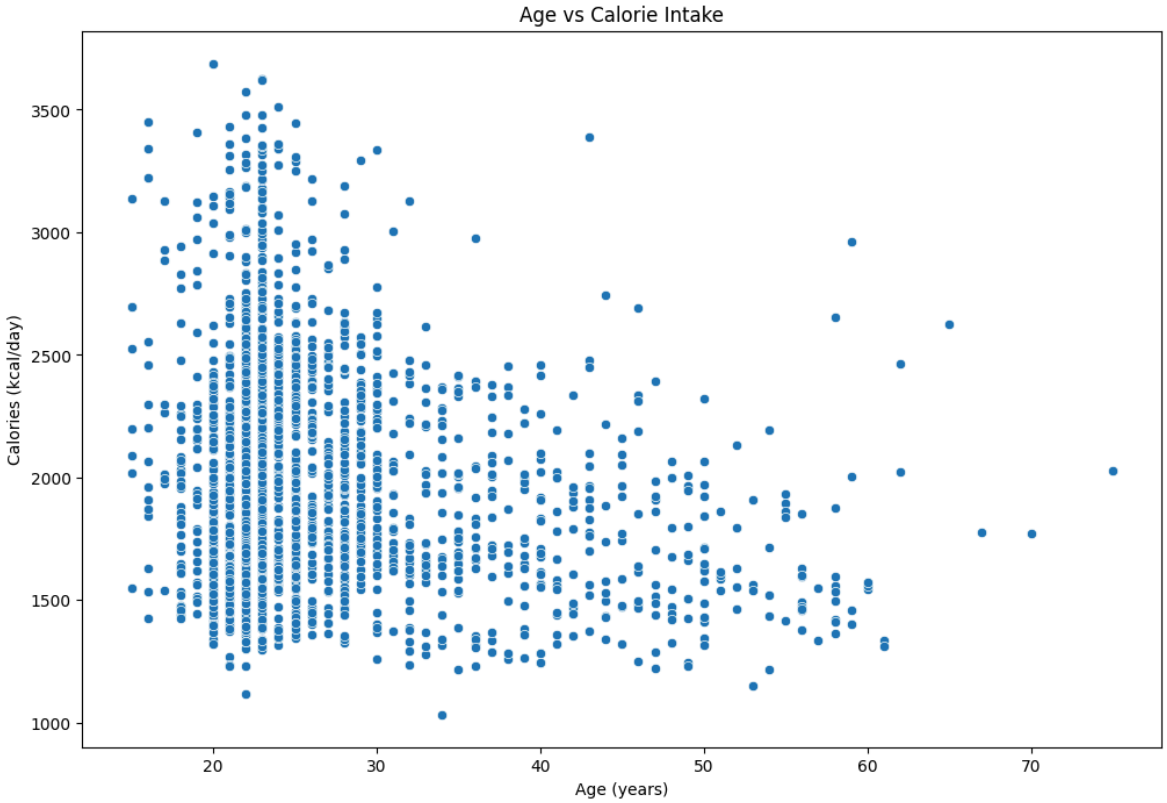
Objective: Determine how BMR changes with age and its relationship with calorie intake.

Harshita Sherla(50593920)

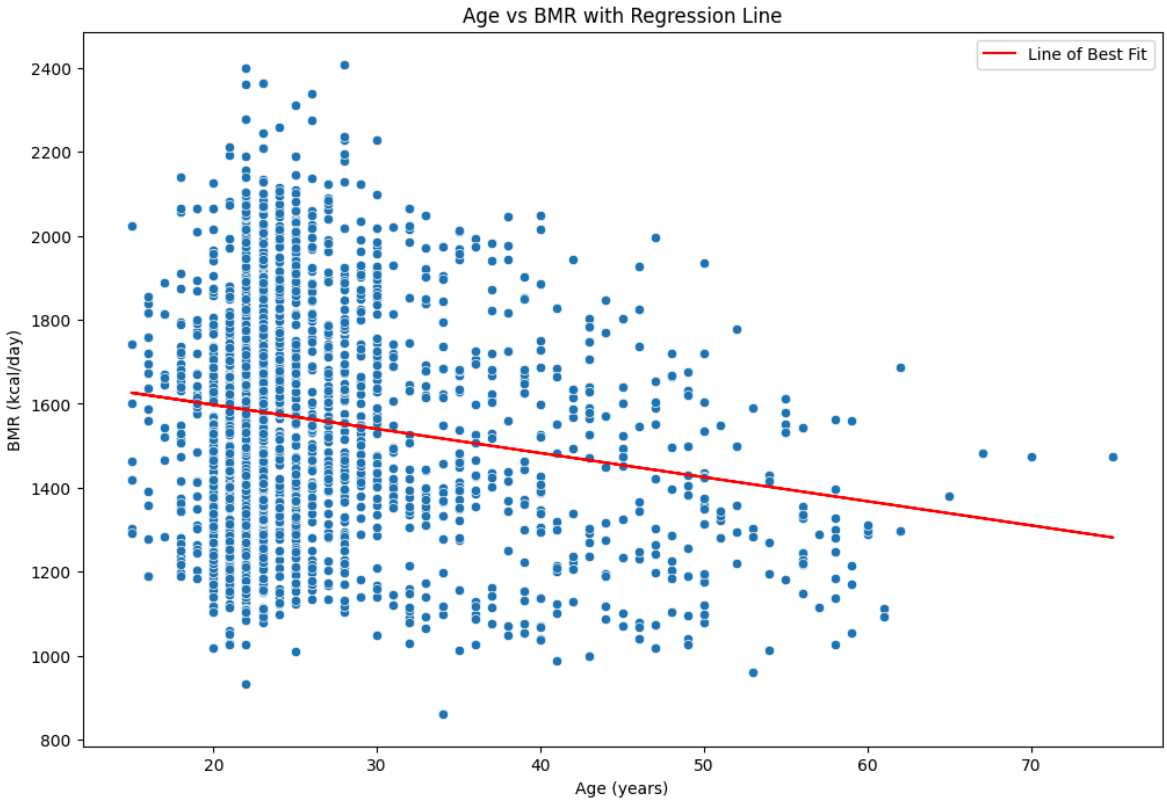
In [12]:
```
plt.figure(figsize=(12, 8))
sns.scatterplot(x='Age', y='BMR', data=df)
plt.title('Age vs BMR')
plt.xlabel('Age (years)')
plt.ylabel('BMR (kcal/day)')
plt.show()

plt.figure(figsize=(12, 8))
sns.scatterplot(x='Age', y='Calories', data=df)
plt.title('Age vs Calorie Intake')
plt.xlabel('Age (years)')
plt.ylabel('Calories (kcal/day)')
plt.show()
```

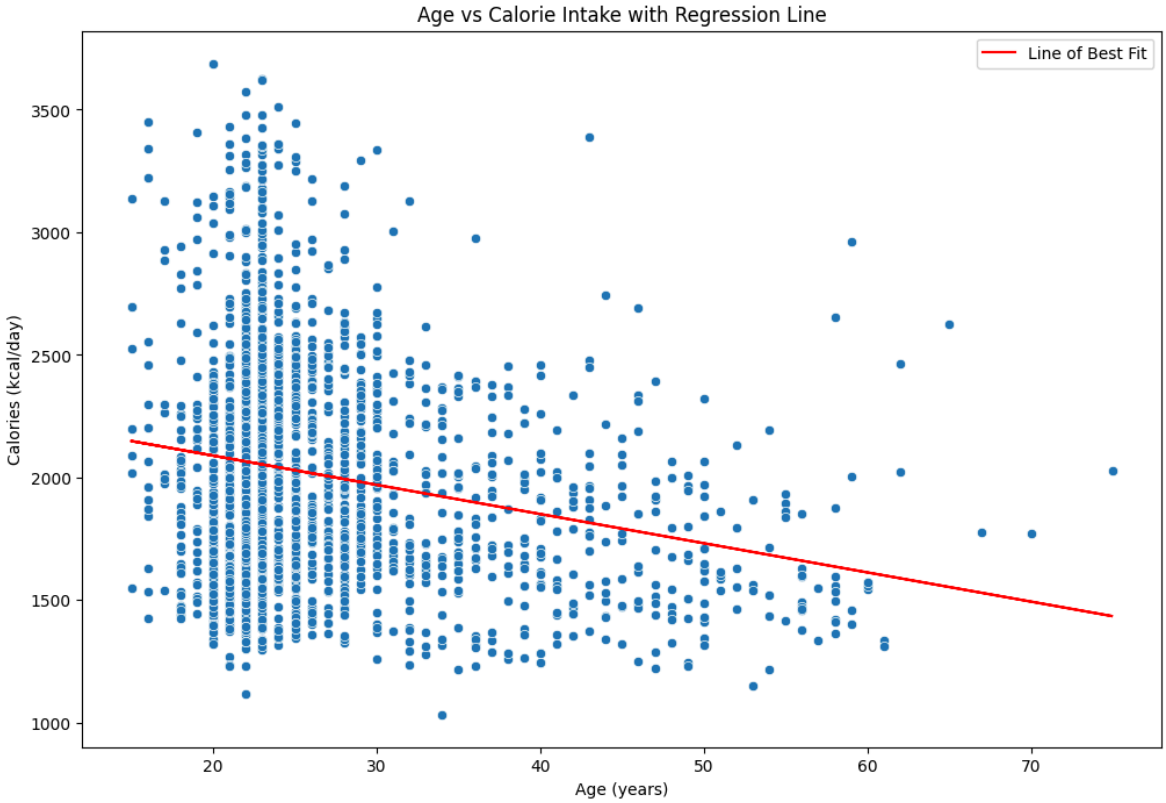
```
correlation_bmr_age, p_value_bmr_age = stats.pearsonr(df['Age'], df['BMR'
correlation_calories_age, p_value_calories_age = stats.pearsonr(df['Age']

print(f"Correlation between Age and BMR: {correlation_bmr_age:.4f} (p-val
print(f"Correlation between Age and Calorie Intake: {correlation_calories

slope_bmr, intercept_bmr, r_value_bmr, p_value_bmr, std_err_bmr = stats.l

line_bmr = slope_bmr * df['Age'] + intercept_bmr
plt.figure(figsize=(12, 8))
sns.scatterplot(x='Age', y='BMR', data=df)
plt.plot(df['Age'], line_bmr, color='red', label='Line of Best Fit')
plt.title('Age vs BMR with Regression Line')
plt.xlabel('Age (years)')
plt.ylabel('BMR (kcal/day)')
plt.legend()
plt.show()

slope_calories, intercept_calories, r_value_calories, p_value_calories, s

line_calories = slope_calories * df['Age'] + intercept_calories
plt.figure(figsize=(12, 8))
sns.scatterplot(x='Age', y='Calories', data=df)
plt.plot(df['Age'], line_calories, color='red', label='Line of Best Fit')
plt.title('Age vs Calorie Intake with Regression Line')
plt.xlabel('Age (years)')
plt.ylabel('Calories (kcal/day)')
plt.legend()
plt.show()

print("\nInterpretation:")
if p_value_bmr < 0.05:
    print("There is a statistically significant relationship between Age
else:
    print("There is no statistically significant relationship between Age

if p_value_calories < 0.05:
    print("There is a statistically significant relationship between Age
else:
    print("There is no statistically significant relationship between Age
```






```
Correlation between Age and BMR: -0.1810 (p-value: 0.0000)
Correlation between Age and Calorie Intake: -0.2258 (p-value: 0.0000)
```






```
Interpretation:
There is a statistically significant relationship between Age and BMR.
There is a statistically significant relationship between Age and Calorie
Intake.
```

Hypothesis-4:

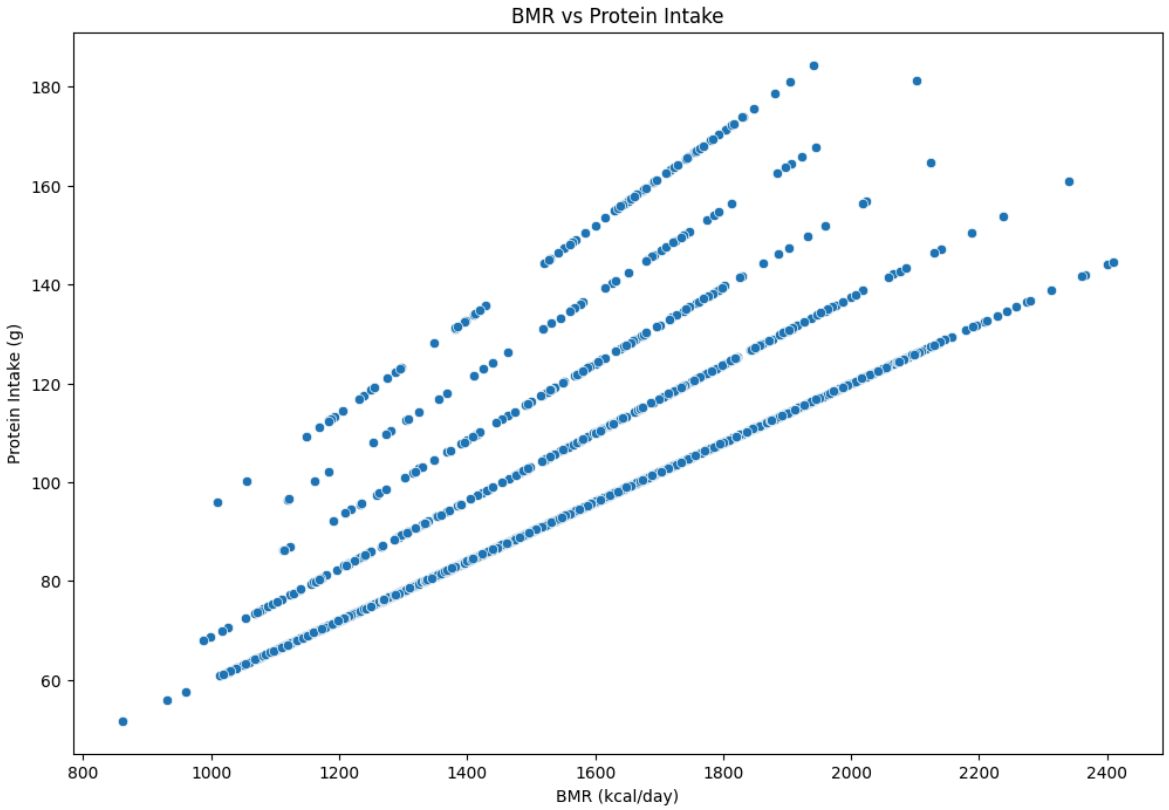
People with higher BMR consume more proteins.

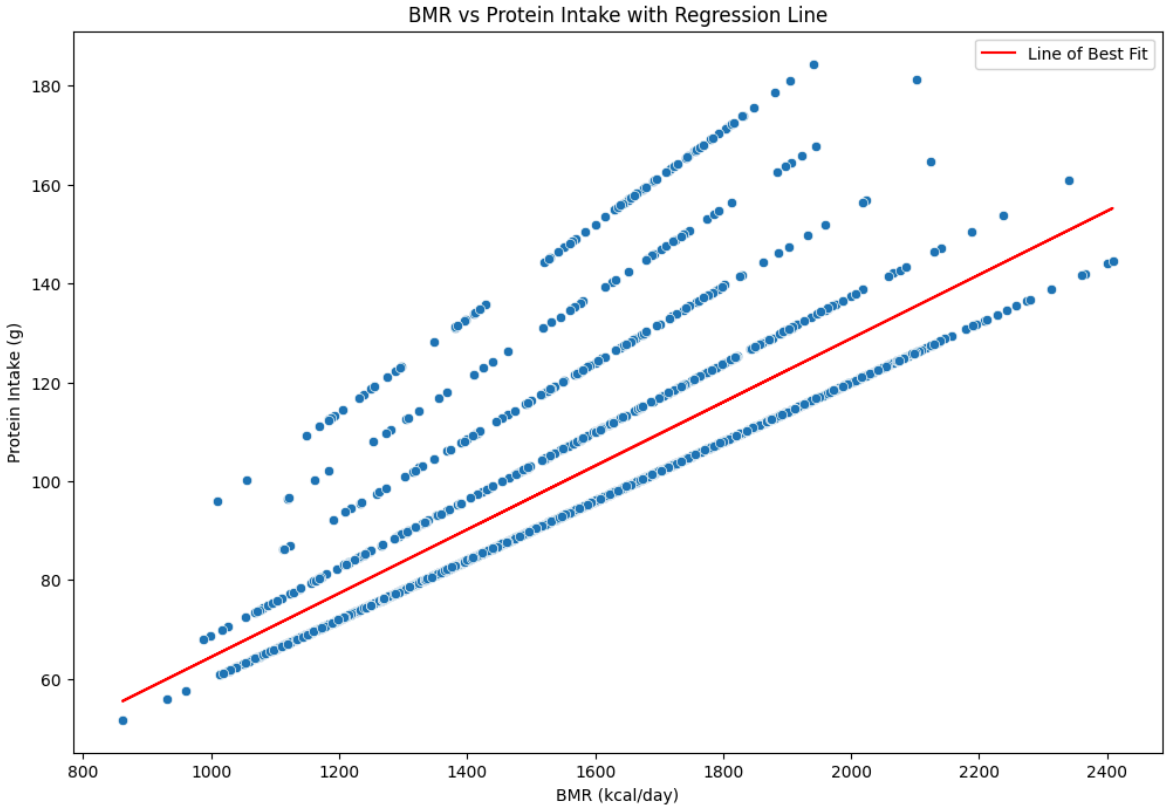
Objective: Test whether individuals with higher BMR have a diet richer in proteins.

Harshita Sherla(50593920)

In [5]:
```
plt.figure(figsize=(12, 8))
sns.scatterplot(x='BMR', y='Proteins', data=df)
plt.title('BMR vs Protein Intake')
plt.xlabel('BMR (kcal/day)')
plt.ylabel('Protein Intake (g)')
plt.show()

slope, intercept, r_value, p_value, std_err = stats.linregress(df['BMR'],

line = slope * df['BMR'] + intercept

plt.figure(figsize=(12, 8))
sns.scatterplot(x='BMR', y='Proteins', data=df)
plt.plot(df['BMR'], line, color='red', label='Line of Best Fit')
plt.title('BMR vs Protein Intake with Regression Line')
plt.xlabel('BMR (kcal/day)')
plt.ylabel('Protein Intake (g)')
plt.legend()
plt.show()

print(f"Slope: {slope:.4f}")
print(f"Intercept: {intercept:.4f}")
print(f"R-squared: {r_value**2:.4f}")
print(f"P-value: {p_value:.4e}")

correlation_coefficient = df['BMR'].corr(df['Proteins'])
print(f"Correlation coefficient: {correlation_coefficient:.4f}")

mean_bmr = df['BMR'].mean()
mean_protein = df['Proteins'].mean()
high_bmr = df[df['BMR'] > mean_bmr]
low_bmr = df[df['BMR'] <= mean_bmr]

print(f"\nMean BMR: {mean_bmr:.2f} kcal/day")
print(f"Mean Protein Intake: {mean_protein:.2f} g")
print(f"Mean Protein Intake for High BMR group: {high_bmr['Proteins'].mea
print(f"Mean Protein Intake for Low BMR group: {low_bmr['Proteins'].mean(

t_statistic, t_p_value = stats.ttest_ind(high_bmr['Proteins'], low_bmr['P
print(f"\nt-test p-value: {t_p_value:.4e}")

print("\nInterpretation:")
if p_value < 0.05:
    print("There is a statistically significant relationship between BMR
else:
    print("There is no statistically significant relationship between BMR

if r_value**2 > 0.5:
    print("The relationship between BMR and protein intake is strong.")
elif r_value**2 > 0.3:
    print("The relationship between BMR and protein intake is moderate.")
else:
    print("The relationship between BMR and protein intake is weak.")

if slope > 0:
    print("As BMR increases, protein intake tends to increase.")
```

```
else:
    print("As BMR increases, protein intake tends to decrease.")

if t_p_value < 0.05:
    print("There is a significant difference in protein intake between hi
else:
    print("There is no significant difference in protein intake between h
```





```
Slope: 0.0645
Intercept: -0.0559
R-squared: 0.6032
P-value: 0.0000e+00
Correlation coefficient: 0.7766

Mean BMR: 1558.23 kcal/day
Mean Protein Intake: 100.39 g
Mean Protein Intake for High BMR group: 116.43 g
Mean Protein Intake for Low BMR group: 84.98 g

t-test p-value: 2.9853e-307

Interpretation:
There is a statistically significant relationship between BMR and protein
intake.
The relationship between BMR and protein intake is strong.
As BMR increases, protein intake tends to increase.
There is a significant difference in protein intake between high and low B
MR groups.
```

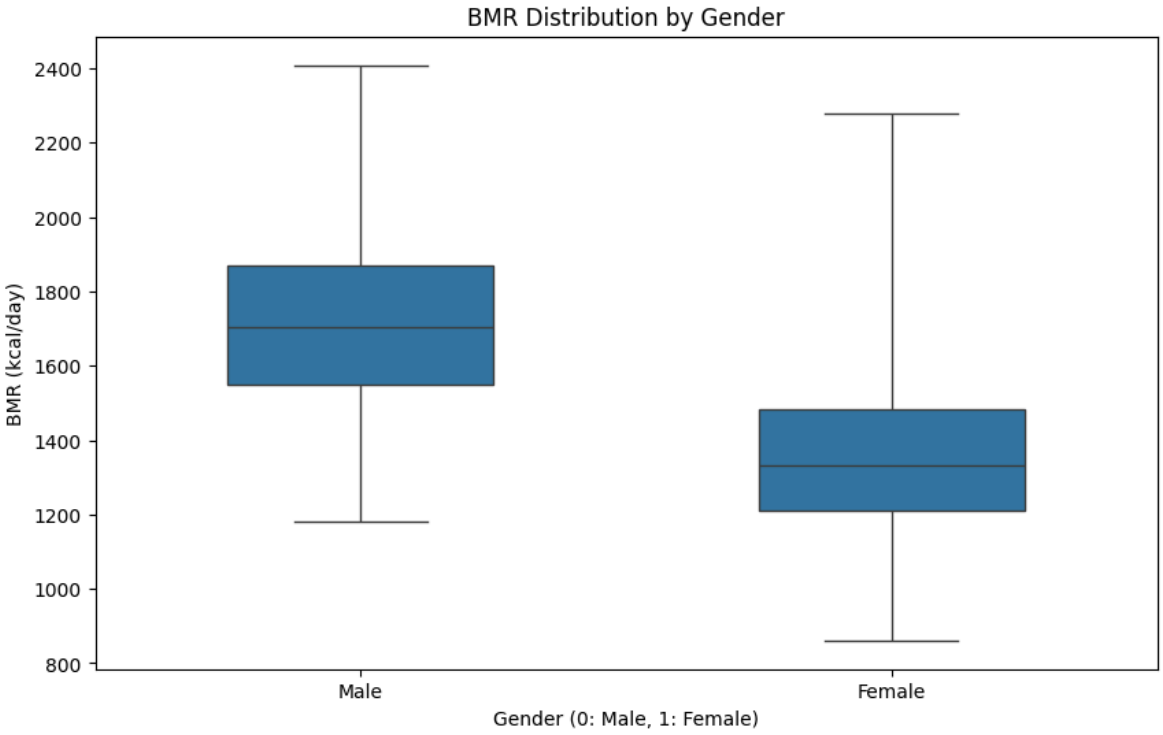
Hypothesis-5:

Females tend to have lower BMR compared to males

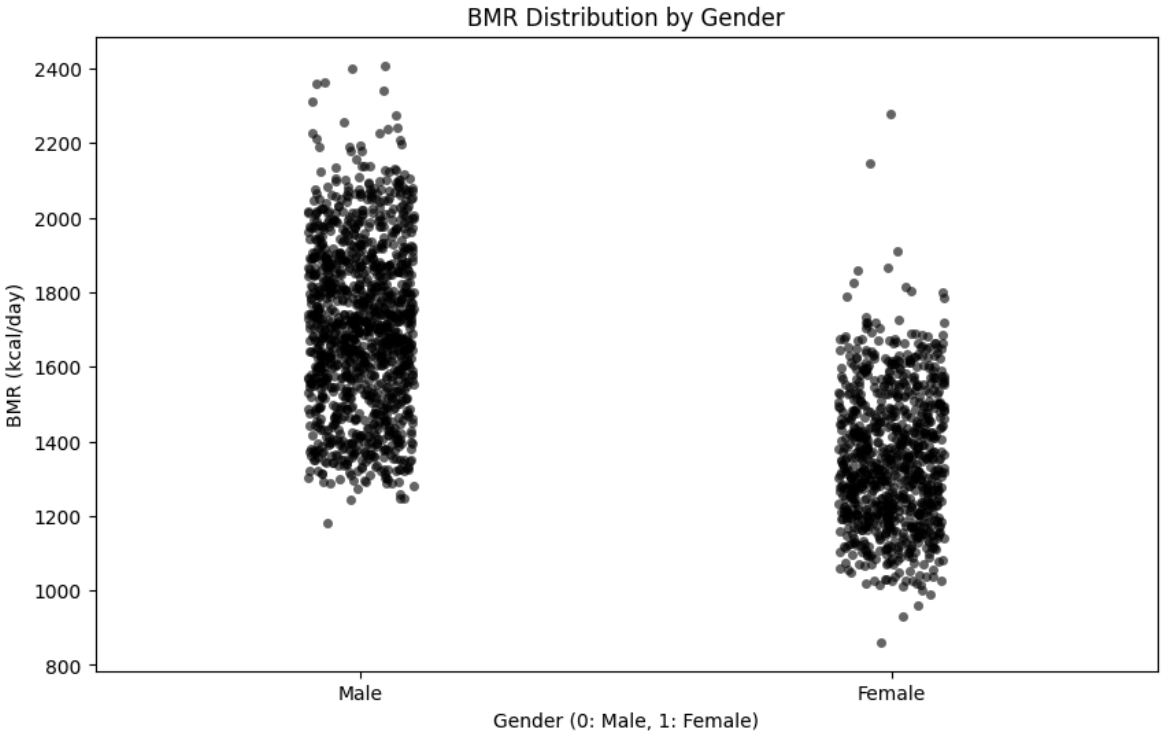
Objective: To compare the BMR of males and females to identify any significant differences based on gender.

Karthik Sharma Madugula (50611293)

In [27]:
```
plt.figure(figsize=(10, 6))
sns.boxplot(x='Gender', y='BMR', data=df, whis=[0, 100], width=0.5, flier

plt.title('BMR Distribution by Gender')
plt.xlabel('Gender (0: Male, 1: Female)')
plt.ylabel('BMR (kcal/day)')
plt.xticks([0, 1], ['Male', 'Female'])
plt.show()
```



In [26]:
```
plt.figure(figsize=(10, 6))
sns.stripplot(x='Gender', y='BMR', data=df, color='black', alpha=0.6, jit
plt.title('BMR Distribution by Gender')
plt.xlabel('Gender (0: Male, 1: Female)')
plt.ylabel('BMR (kcal/day)')
plt.xticks([0, 1], ['Male', 'Female'])
plt.show()
```



Hypothesis-6:

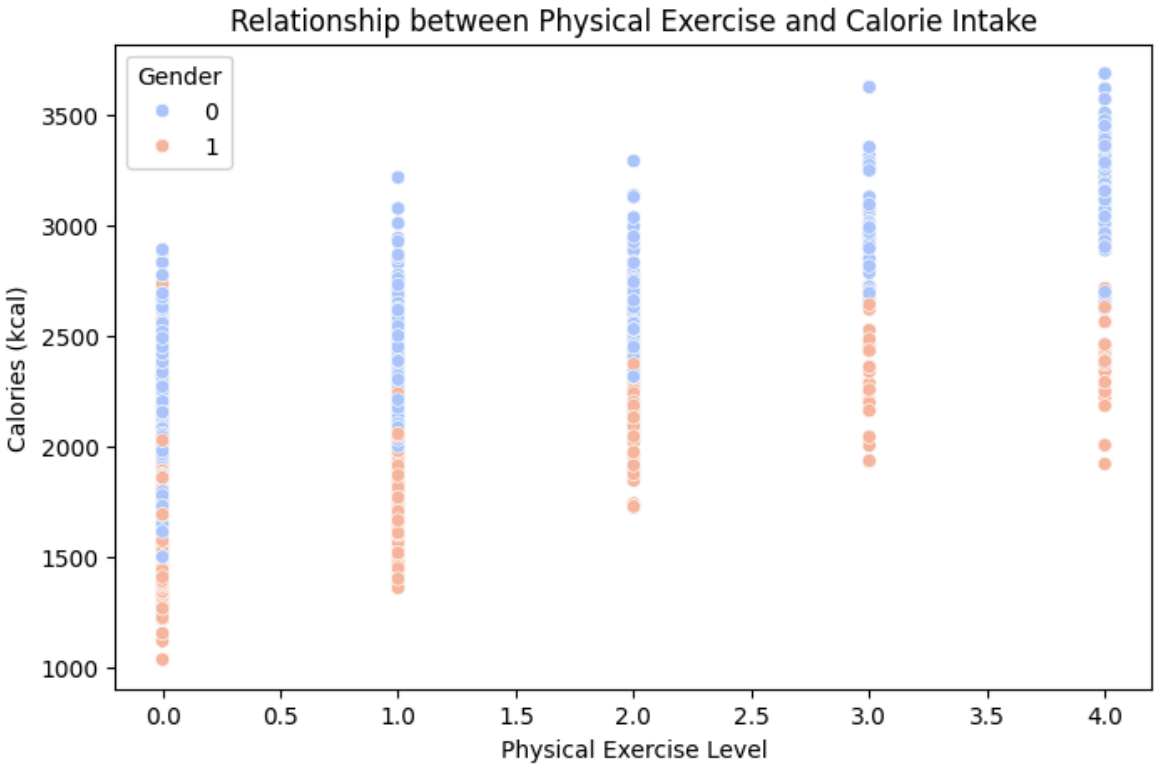
Higher Physical Exercise is Associated with Higher Calorie Intake

Objective: Test whether physical exercise is associated with higher calorie intake

Karthik Sharma Madugula (50611293)

In [24]:
```
exercise_numeric = df['Physical exercise'].map({
    'None': 0,
    'Low': 1,
    'Moderate': 2,
    'High': 3,
    'Very High': 4
}).astype(int)

correlation_exercise_calories = exercise_numeric.corr(df['Calories'])
print(f'Correlation between physical exercise and calorie intake: {correl

plt.figure(figsize=(8, 5))
sns.scatterplot(x=exercise_numeric, y=df['Calories'], hue=df['Gender'], p
plt.title('Relationship between Physical Exercise and Calorie Intake')
plt.xlabel('Physical Exercise Level')
plt.ylabel('Calories (kcal)')
plt.show()
```

```
Correlation between physical exercise and calorie intake: 0.6409
```



Hypothesis-7:

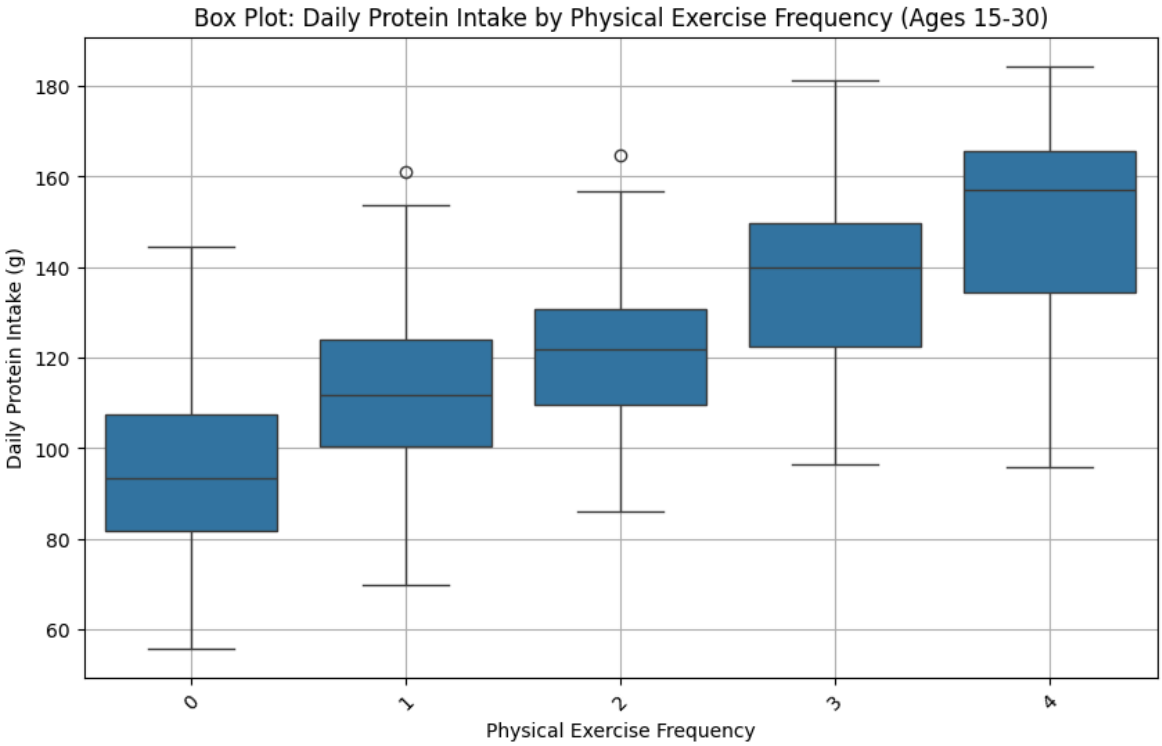
Physical exercise frequency is positively correlated with daily protein intake among individuals aged 15-30

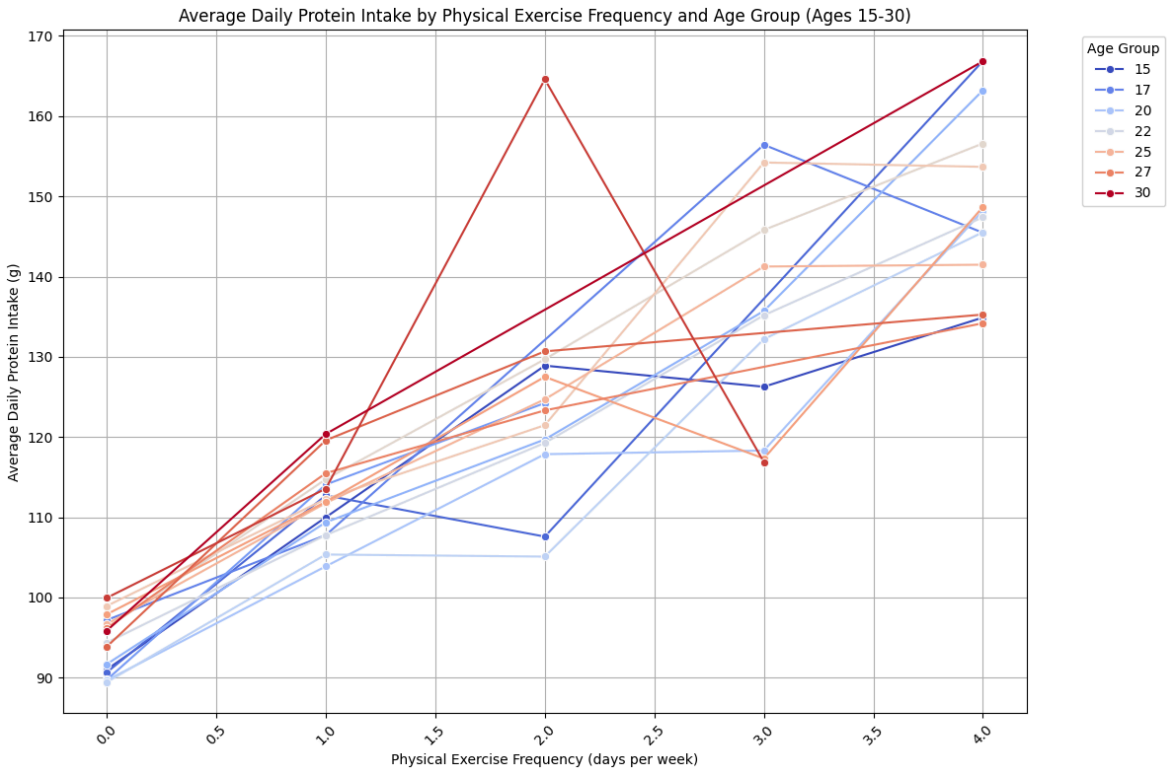
Objective: To investigate the relationship between the frequency of physical exercise and the frequency and daily protein intake of individuals 15-20

Santosh Kota (50593968)

In [22]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# creating a data frame with age between 15 - 30
df_age_15_30 = df[(df['Age'] >= 15) & (df['Age'] <= 30)].copy()

exercise_mapping = {
    'None': 0,
    'Low': 1,
    'Moderate': 2,
    'High': 3,
    'Very High': 4
}

df_age_15_30['Physical exercise numeric'] = df_age_15_30['Physical exerci

correlation_protein_exercise = df_age_15_30['Physical exercise numeric'].
print(f'Correlation between physical exercise frequency and daily protein

plt.figure(figsize=(10, 6))
sns.boxplot(x='Physical exercise numeric', y='Proteins', data=df_age_15_3
plt.title('Box Plot: Daily Protein Intake by Physical Exercise Frequency
plt.xlabel('Physical Exercise Frequency')
plt.ylabel('Daily Protein Intake (g)')
plt.grid(True)
plt.xticks(rotation=45)
plt.show()

avg_protein_by_exercise_age = df_age_15_30.groupby(['Physical exercise nu

plt.figure(figsize=(12, 8))
sns.lineplot(data=avg_protein_by_exercise_age,
             x='Physical exercise numeric',
             y='Proteins',
             hue='Age',
             marker='o',
             palette='coolwarm')

plt.title('Average Daily Protein Intake by Physical Exercise Frequency an
plt.xlabel('Physical Exercise Frequency (days per week)')
plt.ylabel('Average Daily Protein Intake (g)')
plt.grid(True)
plt.legend(title='Age Group', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

```
Correlation between physical exercise frequency and daily protein intake:
0.6588
```





The correlation of 0.7451 reflects a strong relationship, indicating that individuals who engage in higher levels of physical exercise are likely to consume higher amounts of protein.

Hypothesis-8:

Body Mass Index affects Basal Metabolic rate

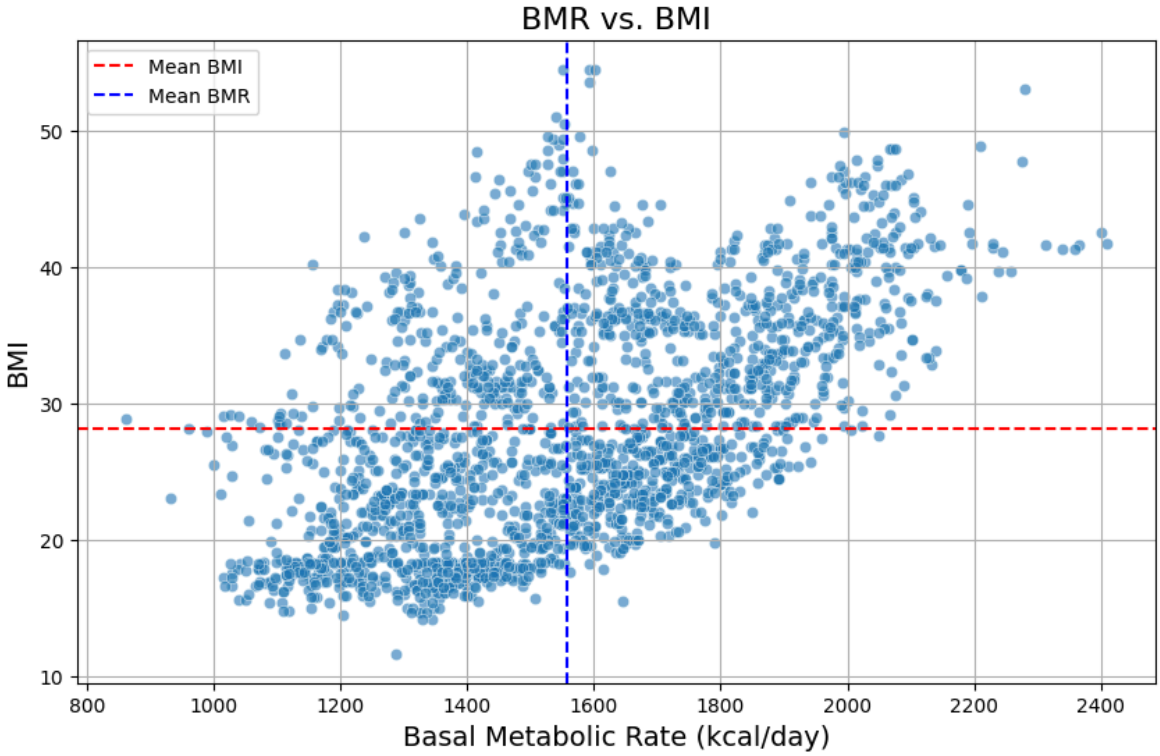
Objective: To look at the relationship between BMI and BMR in the general population and see if people with a higher BMI have a higher basal metabolic rate, independent

of age.

Santosh Kota(50593968)

In [18]:
```
from scipy.stats import pearsonr

correlation, p_value = pearsonr(df['BMR'], df['BMI'])

print(f'Correlation between BMR and BMI: {correlation:.2f}')
print(f'p-value: {p_value:.4f}')

if p_value < 0.05:
    print("We have a statistically significant correlation.")
else:
    print("We don't have a statistically significant correlation.")
plt.figure(figsize=(10, 6))
sns.scatterplot(x='BMR', y='BMI', data=df, alpha=0.6)
plt.title('BMR vs. BMI', fontsize=16)
plt.xlabel('Basal Metabolic Rate (kcal/day)', fontsize=14)
plt.ylabel('BMI', fontsize=14)
plt.axhline(y=df['BMI'].mean(), color='red', linestyle='--', label='Mean
plt.axvline(x=df['BMR'].mean(), color='blue', linestyle='--', label='Mean
plt.legend()
plt.grid()
plt.show()
```

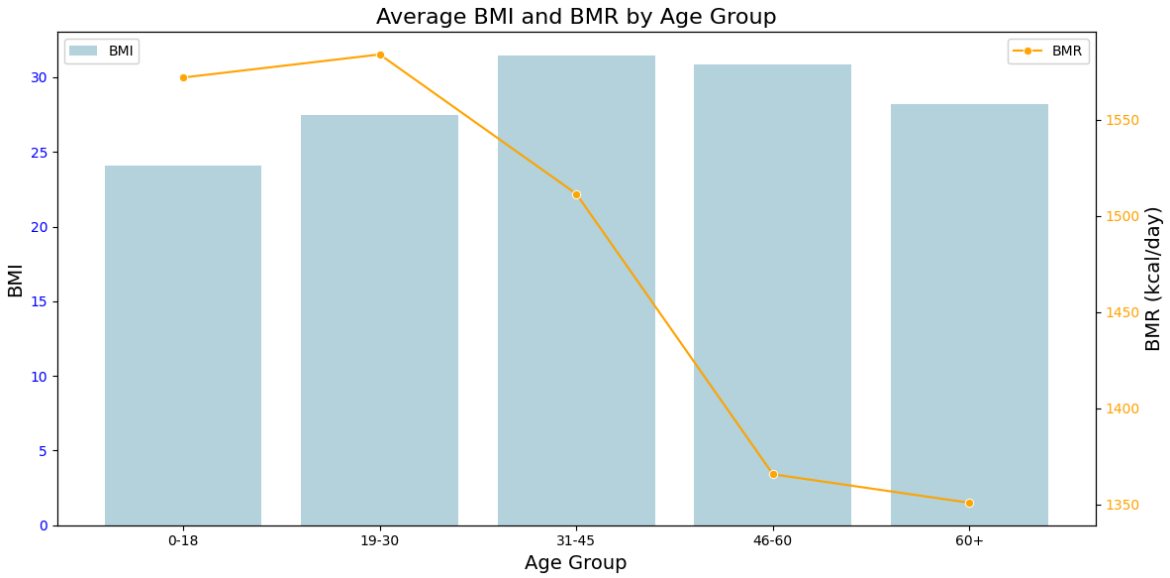
```
Correlation between BMR and BMI: 0.55
p-value: 0.0000
We have a statistically significant correlation.
```



From the plot we can see a positive correlation. so we can say, individuals with a higher BMI are prone to have a higher metabolic rate.

In [19]:
```
age_group_summary = df.groupby('Age Group', observed=True).agg({'BMI': 'm

fig, ax1 = plt.subplots(figsize=(12, 6))
```

```
sns.barplot(x='Age Group', y='BMI', data=age_group_summary, ax=ax1, color
ax1.set_ylabel('BMI', fontsize=14)
ax1.set_xlabel('Age Group', fontsize=14)
ax1.tick_params(axis='y', labelcolor='blue')
ax1.set_title('Average BMI and BMR by Age Group', fontsize=16)
ax1.legend(loc='upper left')

ax2 = ax1.twinx()
sns.lineplot(x='Age Group', y='BMR', data=age_group_summary, ax=ax2, colo
ax2.set_ylabel('BMR (kcal/day)', fontsize=14)
ax2.tick_params(axis='y', labelcolor='orange')

plt.xticks(rotation=45)
plt.tight_layout()

plt.show()
```



Based on this relationship between age groups, BMR, and BMI, we may conclude that middle-aged persons may have a high BMI while maintaining a steady or high BMR, whereas older adults often suffer a drop in BMR and BMI, due to various lifestyle and dietary factors.

Hypothesis-9:

Eating more meals each day leads to higher BMI

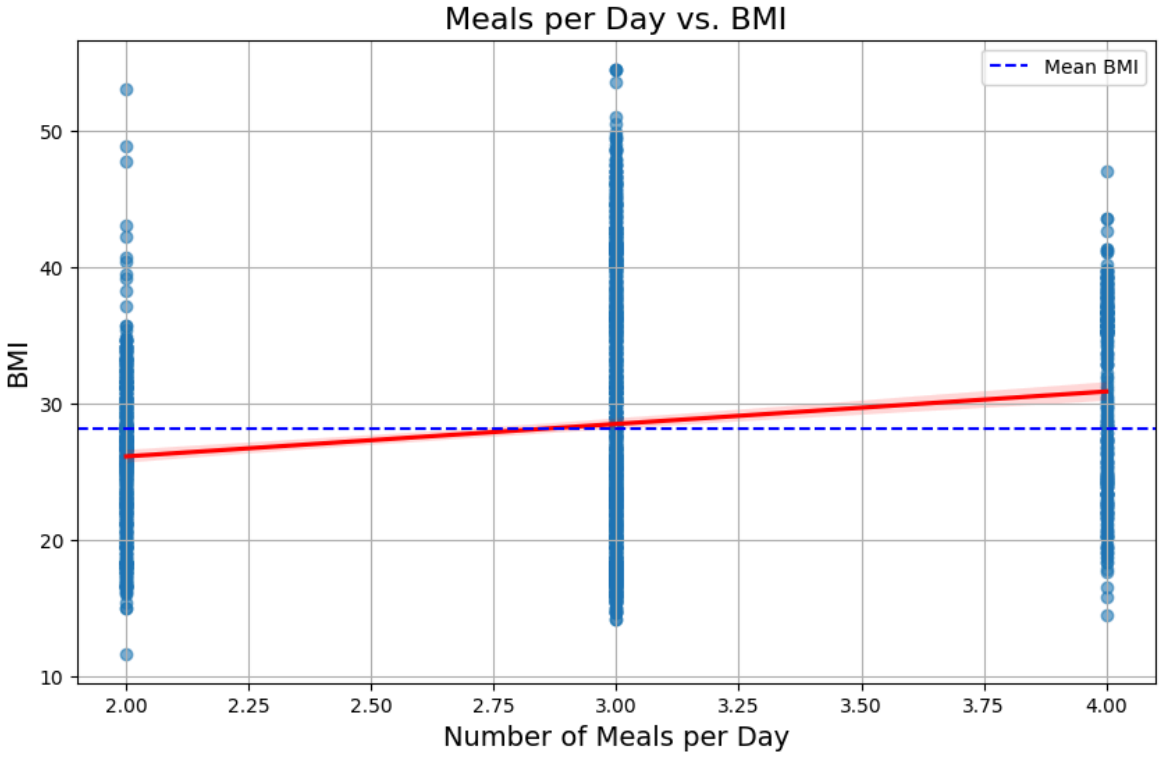
Objective: To determine whether eating more frequently throughout the day is associated with a higher BMI.

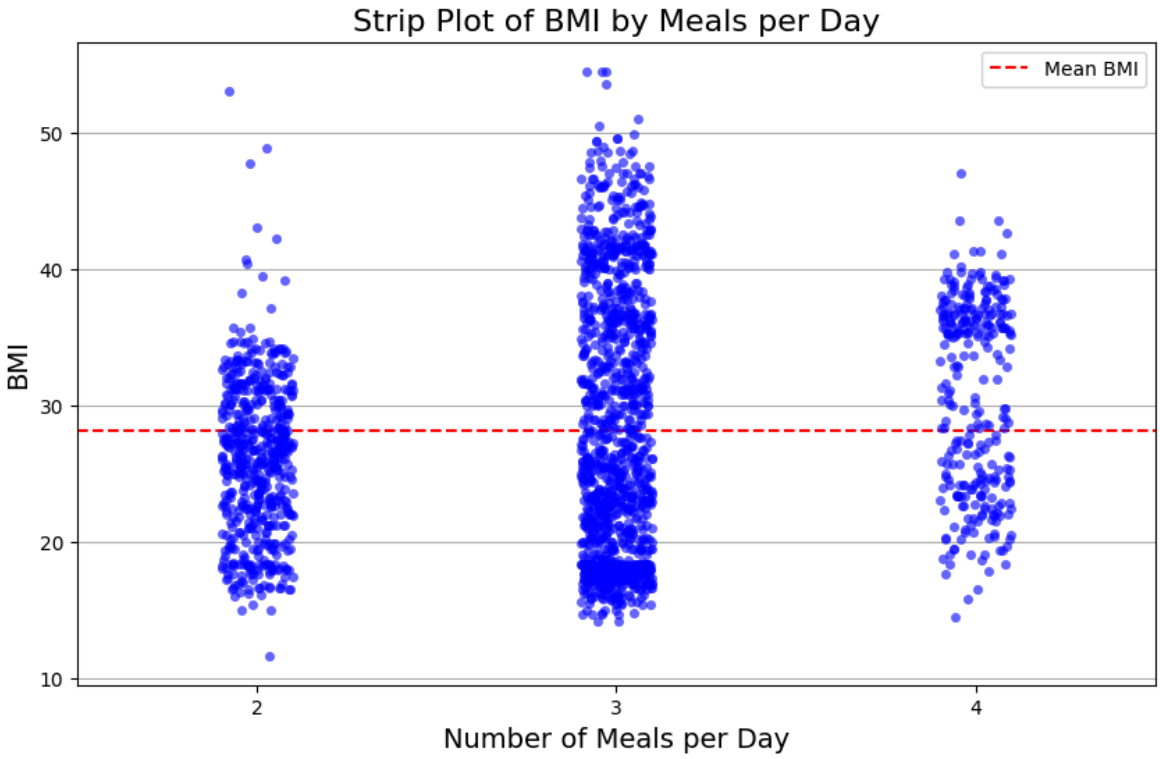
Karthik Sharma Madugula (50611293)

In [20]:
```
correlation, p_value = pearsonr(df['Daily meals frequency'], df['BMI'])

print(f'Correlation between Number of Meals per Day and BMI: {correlation
print(f'p-value: {p_value:.4f}')

if p_value < 0.05:
    print("we have a statistically significant correlation.")
```

```
else:
    print("we don't have a statistically significant correlation.")
plt.figure(figsize=(10, 6))
sns.regplot(x='Daily meals frequency', y='BMI', data=df, scatter_kws={'al
plt.title('Meals per Day vs. BMI', fontsize=16)
plt.xlabel('Number of Meals per Day', fontsize=14)
plt.ylabel('BMI', fontsize=14)
plt.axhline(y=df['BMI'].mean(), color='blue', linestyle='--', label='Mean
plt.grid()
plt.legend()
plt.show()

plt.figure(figsize=(10, 6))
sns.stripplot(x='Daily meals frequency', y='BMI', data=df, color='blue',
plt.title('Strip Plot of BMI by Meals per Day', fontsize=16)
plt.xlabel('Number of Meals per Day', fontsize=14)
plt.ylabel('BMI', fontsize=14)
plt.axhline(y=df['BMI'].mean(), color='red', linestyle='--', label='Mean
plt.legend()
plt.grid(axis='y')
plt.show()
```

```
Correlation between Number of Meals per Day and BMI: 0.18
p-value: 0.0000
we have a statistically significant correlation.
```





Based on the scatter plot, as the number of meals increases, there is a slight increase in BMI as well,

In [ ]: