



# **PANDAS DATA ANALYSIS** **PROJECT**



## **GYM & DIET RECOMMENDATION**

Prepared by:

AAKASH V DEVAN

# INTRODUCTION

The dataset used in this project focuses on analyzing various aspects of health and fitness, including lifestyle habits, medical conditions, and demographic information. This dataset provides a comprehensive view of individuals' health-related metrics, helping to uncover patterns and insights into how different factors influence physical fitness and health conditions like hypertension and diabetes.

## ABOUT DATASET

Dataset Includes:

- ❖ Rows : 14590
- ❖ Columns : 13

➤ **df.dtypes:**

ID	int64
Sex	object
Age	int64
Height	float64
Weight	float64
Hypertension	object
Diabetes	object
BMI	float64
Level	object
Fitness Goal	object
Fitness Type	object
Exercises	object
Diet	object

➤ **df.shape:**  
(14589, 13)

## Key Features:

### Demographics:

- **Age:** Age of individuals, which allows grouping into age categories for trend analysis.
- **Sex:** Gender information, enabling comparisons of health and fitness goals between males and females.
- **Height:** Height of the Person
- **Weight:** Weight of the person

### Health Metrics:

- **BMI (Body Mass Index):** A crucial indicator of body fat based on height and weight, which is central to understanding health risks.
- **Level:** Categorized based on the BMI
- **Hypertension:** Information on whether an individual has hypertension (Yes/No).
- **Diabetes:** Information on whether an individual has diabetes (Yes/No).

### Lifestyle and Fitness:

- **Fitness Goal:** Describes the primary objective of an individual, such as weight loss, muscle gain, or endurance improvement.
- **Exercises:** Lists recommended or followed exercises, providing insights into popular fitness routines.
- **Fitness Type:** Specifies the fitness category (e.g., cardio, strength training, or flexibility).
- **Diet:** Diet plan Recommendation

# OBJECTIVES



The primary objective of this project is to analyze the provided health and fitness dataset to extract meaningful insights, identify patterns, and generate actionable recommendations. The project aims to investigate the relationship between Body Mass Index (BMI) and health conditions like hypertension and diabetes, identifying critical BMI thresholds that indicate higher risks. It also seeks to examine how fitness goals vary across different demographics, such as age, gender, and health conditions, to highlight common objectives for various groups. Furthermore, the project explores the types and frequencies of exercises recommended or followed, identifying popular activities and their alignment with specific goals or health conditions.

- A better understanding of how lifestyle and health are interlinked.
- Enhanced strategies for promoting fitness and preventing chronic conditions.

# DATA CLEANING

## ➤ `df.isna().sum()`

ID	0
Sex	0
Age	0
Height	0
Weight	0
Hypertension	0
Diabetes	0
BMI	0
Level	0
Fitness Goal	0
Fitness Type	0
Exercises	0
Diet	0

There is no NULL Values

## ➤ `df.duplicated().sum()`

0

There is no Duplicates

## Drop Unnecessary Rows & Columns

From Diet Column, found that some rows contain '*Diet*' are values

➤ `df['Diet'].value_counts()`

Vegetables: (Garlic, Mushroom, Green Papper, Icebetg Lettuce); Protein Intake: (Baru Nuts, Beech Nuts, Hemp Seeds, Cheese Spandwich); Juice: (Apple Juice, Mango juice,and Beetroot juice)	5038
Vegetables: (Broccoli, Carrots, Spinach, Lettuce, Onion); Protein Intake: (Cheese, Cattoge cheese, Skim Milk, Law fat Milk, and Baru Nuts); Juice: (Fruit Juice, Aloe vera juice, Cold-pressed juice, and Watermelon juice)	2507
Vegetables: (Garlic, Roma Tomatoes, Capers, Green Papper, and Iceberg Lettuce); Protein Intake: (Cheese Sandwich, Baru Nuts, Beech Nuts, Squash Seeds, Mixed Teff, peanut butter, and jelly sandwich); Juice: (Apple juice, beetroot juice, and mango juice)	1688
Vegetables: (Mixed greens, cherry tomatoes, cucumbers, bell peppers, carrots, celery, bell peppers);Protein Intake: (Chicken, fish, tofu, or legumes); Juice : (Green juice,kale, spinach, cucumber, celery, and apple)	1100
Vegetables: (Tomatoes, Garlic, leafy greens, broccoli, carrots, and bell peppers); Protein Intake: (poultry, fish, tofu, legumes, and low-fat dairy products); Juice: (Apple juice, beetroot juice and mango juice)	844
Vegetables: (Garlic, Roma Tomatoes, Capers and Iceberg Lettuce); Protein Intake: (Cheese Standwish, Baru Nuts, Beech Nuts, Squash Seeds, and Mixed Teff); Juice: (Apple juice, beetroot juice and mango juice)	844
Vegetables: (Garlic, mushroon, green papper and water chestnut);Protein Intake: ( Baru Nuts, Beech Nuts, and black walnut); Juice : (Apple juice, Mango, and Beetroot Juice)	844
Vegetables: (Garlic, mushroon, green papper);Protein Intake: ( Baru Nuts, Beech Nuts, and Hemp Seeds); Juice : (Apple juice, Mango, and Beetroot Juice)	844
Vegetables: (Carrots, Sweet Potato, and Lettuce); Protein Intake: (Red meats, poultry, fish, eggs, dairy products, legumes, and nuts); Juice: (Fruit juice, watermelon juice, carrot juice, apple juice and mango juice)	422
Vegetables: (Carrots, Sweet Potato, Lettuce); Protein Intake: (Red meats, poultry, fish, eggs, dairy products, legumes, and nuts); Juice: (Fruit juice, watermelon juice, carrot juice, apple juice and mango juice)	422
Diet	36

There are 36 rows just contains ‘*Diet*’ as Values

So Remove all rows which is equal to ‘*Diet*’

➤ `df=df[df['Diet']!='Diet']`

Vegetables: (Garlic, Mushroom, Green Papper, Icebetg Lettuce); Protein Intake: (Baru Nuts, Beech Nuts, Hemp Seeds, Cheese Spandwich); Juice: (Apple Juice, Mango juice,and Beetroot juice)	5038
Vegetables: (Broccoli, Carrots, Spinach, Lettuce, Onion); Protein Intake: (Cheese, Cattoge cheese, Skim Milk, Law fat Milk, and Baru Nuts); Juice: (Fruit Juice, Aloe vera juice, Cold-pressed juice, and Watermelon juice)	2507
Vegetables: (Garlic, Roma Tomatoes, Capers, Green Papper, and Iceberg Lettuce); Protein Intake: (Cheese Sandwich, Baru Nuts, Beech Nuts, Squash Seeds, Mixed Teff, peanut butter, and jelly sandwich); Juice: (Apple juice, beetroot juice, and mango juice)	1688
Vegetables: (Mixed greens, cherry tomatoes, cucumbers, bell peppers, carrots, celery, bell peppers);Protein Intake: (Chicken, fish, tofu, or legumes); Juice : (Green juice,kale, spinach, cucumber, celery, and apple)	1100
Vegetables: (Tomatoes, Garlic, leafy greens, broccoli, carrots, and bell peppers); Protein Intake: (poultry, fish, tofu, legumes, and low-fat dairy products); Juice: (Apple juice, beetroot juice and mango juice)	844
Vegetables: (Garlic, Roma Tomatoes, Capers and Iceberg Lettuce); Protein Intake: (Cheese Standwish, Baru Nuts, Beech Nuts, Squash Seeds, and Mixed Teff); Juice: (Apple juice, beetroot juice and mango juice)	844
Vegetables: (Garlic, mushroon, green papper and water chestnut);Protein Intake: ( Baru Nuts, Beech Nuts, and black walnut); Juice : (Apple juice, Mango, and Beetroot Juice)	844
Vegetables: (Garlic, mushroon, green papper);Protein Intake: ( Baru Nuts, Beech Nuts, and Hemp Seeds); Juice : (Apple juice, Mango, and Beetroot Juice)	844
Vegetables: (Carrots, Sweet Potato, and Lettuce); Protein Intake: (Red meats, poultry, fish, eggs, dairy products, legumes, and nuts); Juice: (Fruit juice, watermelon juice, carrot juice, apple juice and mango juice)	422
Vegetables: (Carrots, Sweet Potato, Lettuce); Protein Intake: (Red meats, poultry, fish, eggs, dairy products, legumes, and nuts); Juice: (Fruit juice, watermelon juice, carrot juice, apple juice and mango juice)	422

➤ `df.drop(columns='ID',inplace=True)`

The ID Column is not Useful for my analysis. So remove that Column





## FEATURE EXTRACTION

In my Dataset, The Diet column contains 3 category of diet plans:

['Vegetables', 'Protein Intake', 'Juice ']

➤ `df['Diet'].value_counts().head(1)`

	count
Vegetables: (Garlic, Mushroom, Green Papper, Icebetg Lettuce); Protein Intake: (Baru Nuts, Beech Nuts, Hemp Seeds, Cheese Spandwich); Juice: (Apple Juice, Mango juice,and Beetroot juice)	5038

dtype: int64

Using Feature Extraction divide the Diet Column into 3 Columns

['Vegetables', 'Protein Intake', 'Juice ']

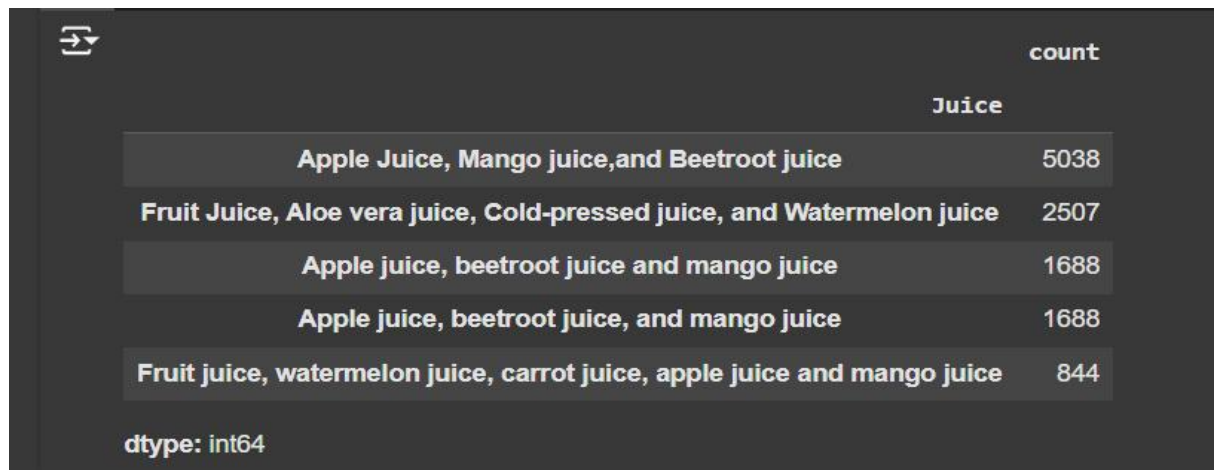
- `df['Vegetables'] = df['Diet'].str.extract(r'Vegetables:\s*((.*?))')`
- `df['Protein_Intake'] = df['Diet'].str.extract(r'Protein Intake:\s*((.*?))')`
- `df['Juice'] = df['Diet'].str.extract(r'Juice:\s*((.*?))')`

Then Remove the Diet Column

➤ **df.drop(columns='Diet',inplace=True)**

There are some rows which has same meaning but the rows are different

➤ **df['Juice'].value\_counts()**



Juice	count
Apple Juice, Mango juice,and Beetroot juice	5038
Fruit Juice, Aloe vera juice, Cold-pressed juice, and Watermelon juice	2507
Apple juice, beetroot juice and mango juice	1688
Apple juice, beetroot juice, and mango juice	1688
Fruit juice, watermelon juice, carrot juice, apple juice and mango juice	844

dtype: int64

The 3<sup>rd</sup> and 4<sup>th</sup> rows are same values but it treated as 2 rows because of an extra comma(,) occurred.

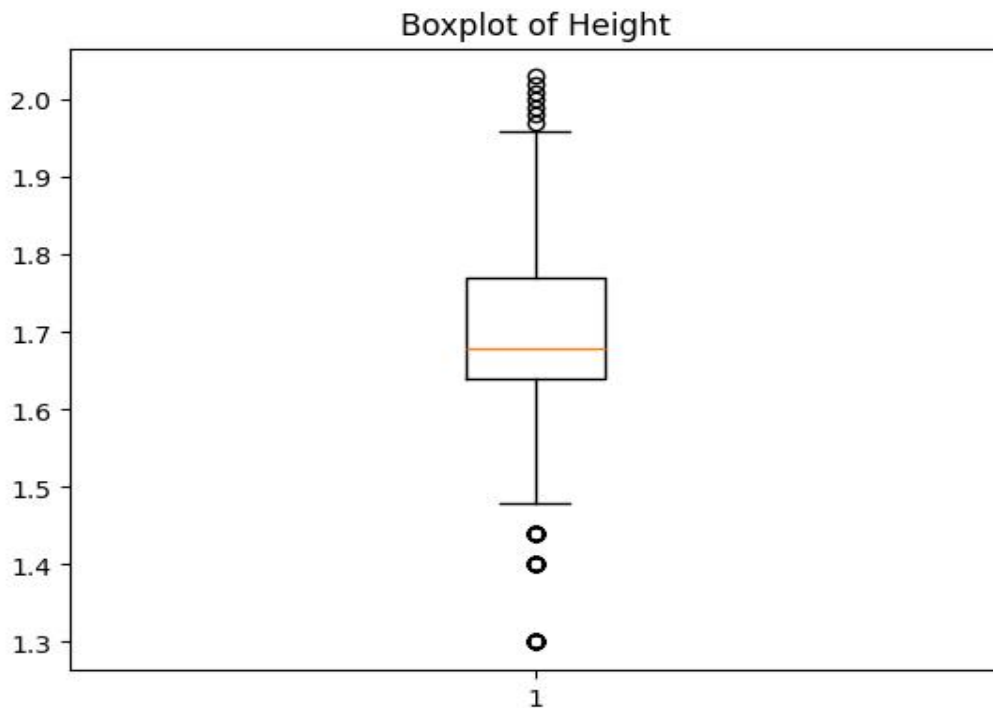
***‘Vegetables & Juice’*** columns are face these issues. To solve this problem use Replace function.

- **df['Vegetables']=df['Vegetables'].replace('Carrots, Sweet Potato, and Lettuce','Carrots, Sweet Potato, Lettuce')**
- **df['Juice']=df['Juice'].replace('Apple juice, beetroot juice, and mango juice','Apple juice, beetroot juice and mango juice')**



# OUTLIER DETECTION

```
➤ plt.boxplot(df['Height'])  
    plt.title("Boxplot of Height")  
    plt.show()
```



To remove Outliers:

```
➤ q1=df['Height'].quantile(0.25)  
    q3=df['Height'].quantile(0.75)  
    iqr=q3-q1  
    min_range=q1-1.5*iqr  
    max_range=q3+1.5*iqr  
    max_range  
    df=df[(df['Height']<max_range)&(df['Height']>=min_range)]
```

To Reset Index:

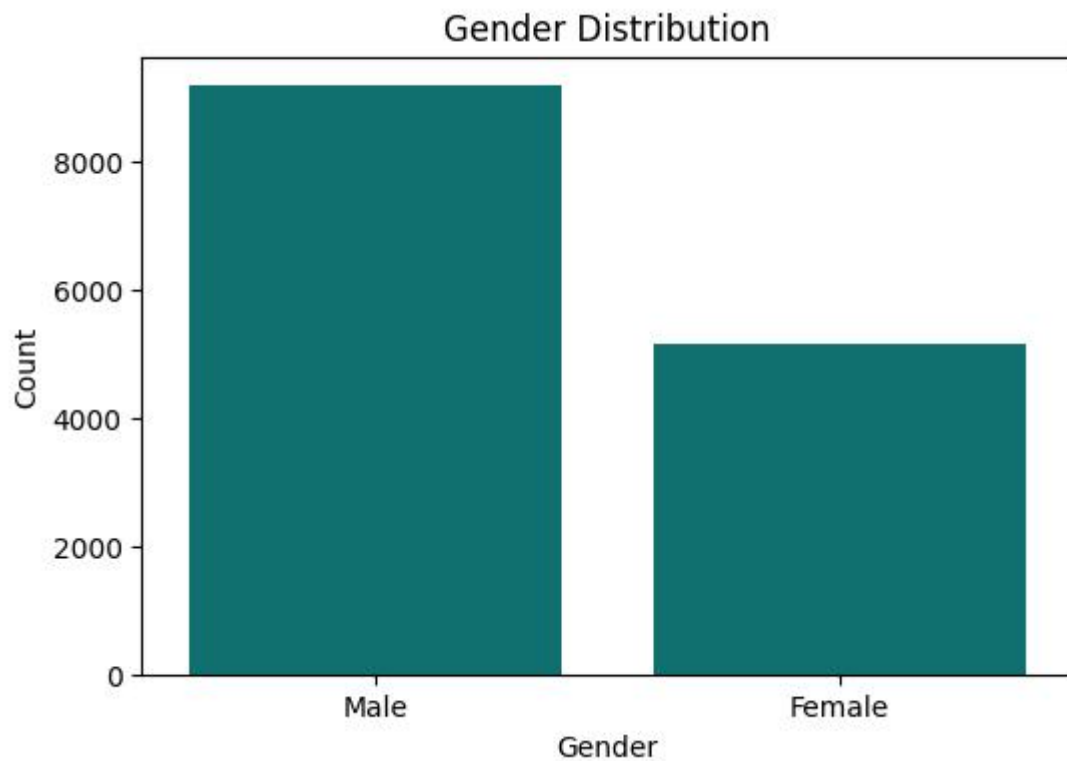
```
➤ df.reset_index(inplace=True,drop=True)
```

# DATA ANALYSIS

## Gender Distribution:

➤ `gender_distribution=df['Sex'].value_counts()`  
`gender_distribution`

SEX	COUNT
Male	9164
Female	5147



## Key Insights:

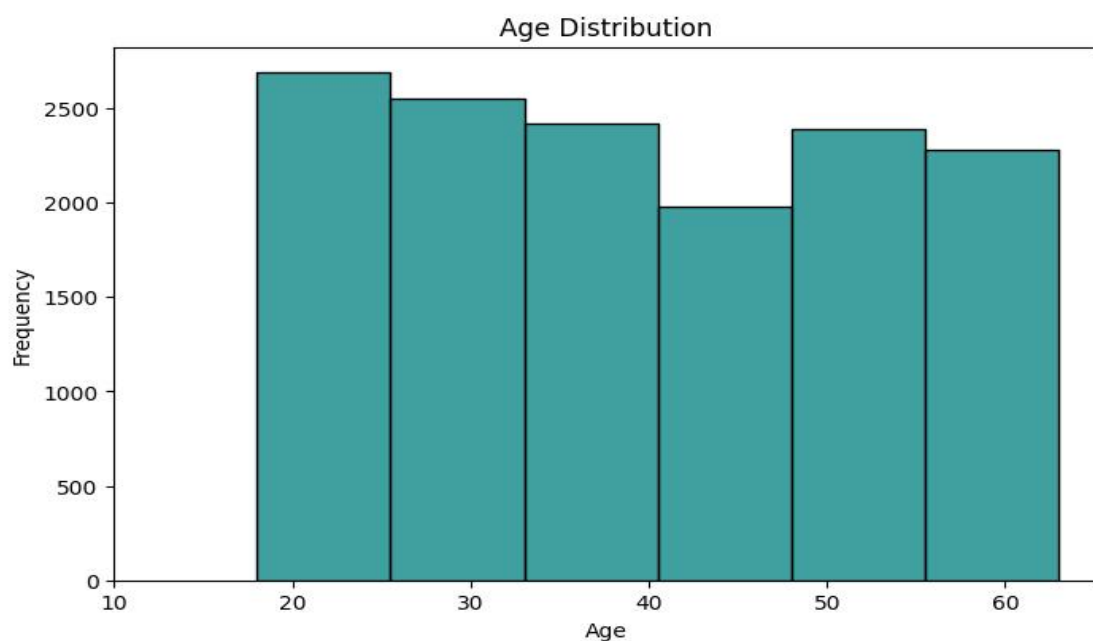
- There are 9164 Males and 5147 Females are Doing Workouts
- Males are doing workouts more

## Age Distribution:

➤ `age_distribution=df['Age'].describe()`

`age_distribution`

	AGE
<b>count</b>	<b>14311.00</b>
<b>mean</b>	<b>39.54084</b>
<b>std</b>	<b>13.28198</b>
<b>min</b>	<b>18.00000</b>
<b>25%</b>	<b>28.00000</b>
<b>50%</b>	<b>39.00000</b>
<b>75%</b>	<b>51.00000</b>
<b>max</b>	<b>63.00000</b>



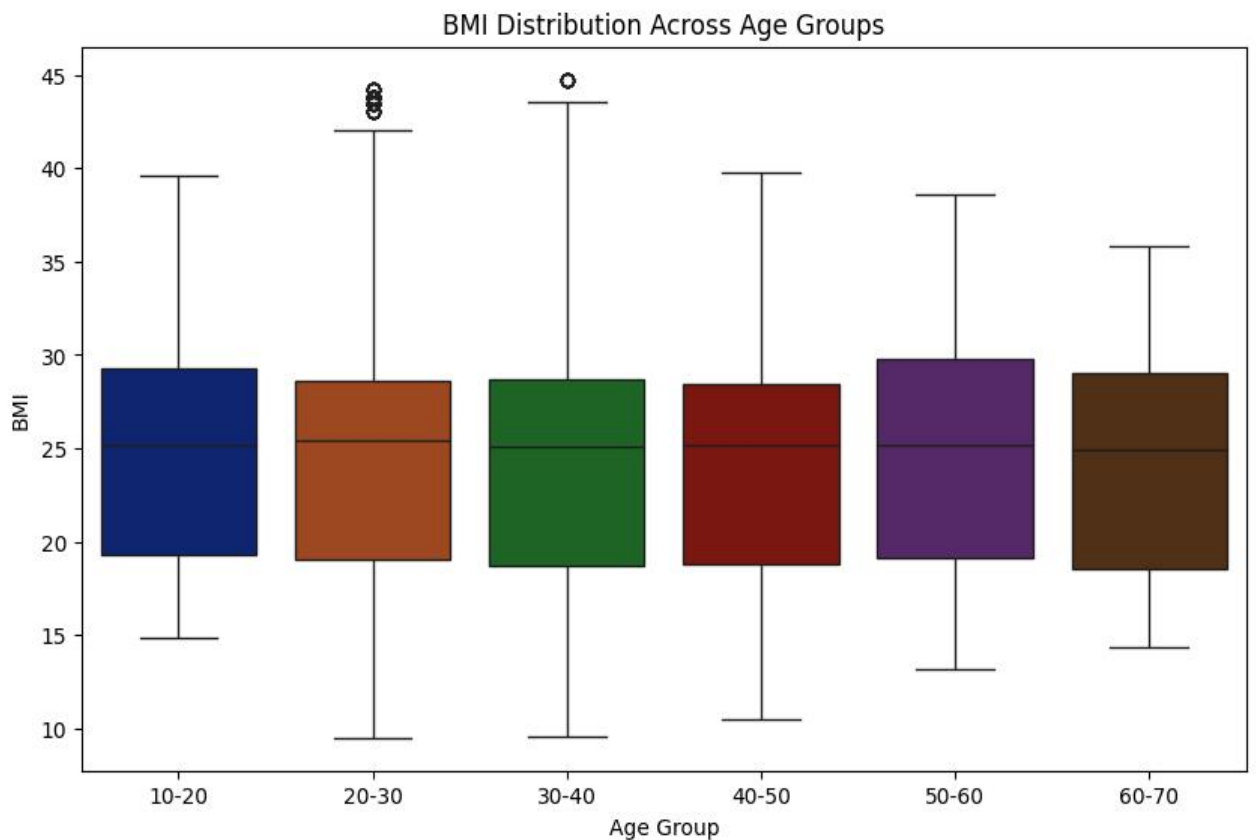
## Key Insights:

- The Age group are in between 18-63
- From Age group of 18-25 was doing more workout and second was 25-35
- Around Age group 40-48 was doing less workout

## BMI Distribution Across Age Groups:

```
➤ age_groups=pd.cut(df['Age'],bins=[10,20,30,40,50,60,70],labels=["10-20","20-30","30-40","40-50","50-60","60-70"])  
df['Age Group']=age_groups  
bmi_by_age_group=df.groupby('Age Group')['BMI'].describe()  
bmi_by_age_group
```

Age Group	count	mean	std	min	25%	50%	75%	max
10-20	997.0	24.572	6.25	14.88	19.27	25.15	29.32	39.56
20-30	3520.0	24.277	6.52	9.52	19.05	25.39	28.60	44.20
30-40	3140.0	24.026	6.48	9.62	18.71	25.10	28.72	44.73
40-50	2893.0	24.017	6.26	10.49	18.79	25.18	28.41	39.76
50-60	2948.0	24.261	5.98	13.15	19.16	25.15	29.78	38.59
60-70	813.0	24.007	6.11	14.35	18.56	24.93	29.00	35.86



### Key Insights:

- 20-30 has the highest count
- If the BMI is in between 18.5 - 24.9 then we considered as Normal. From this understood that most of them have not normal BMI

### Proportion of Hypertension and Diabetes :

```
➤ hypertension_counts=df['Hypertension'].value_counts()
diabetes_counts=df['Diabetes'].value_counts()
print("\nHypertension Counts:\n", hypertension_counts)
print("\nDiabetes Counts:\n", diabetes_counts)
```

Hypertension Counts:

Hypertension

No 7694

Yes 6617

Name: count, dtype: int64

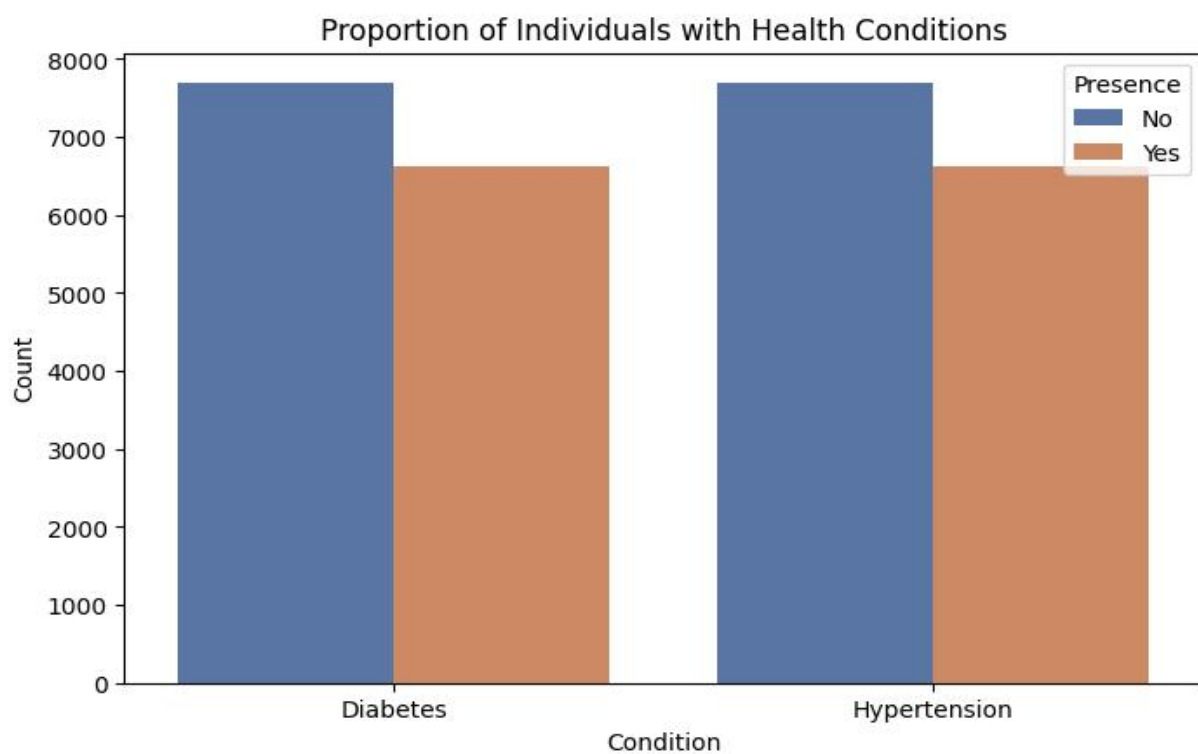
Diabetes Counts:

Diabetes

No 7694

Yes 6617

Name: count, dtype: int64



### Key Insights:

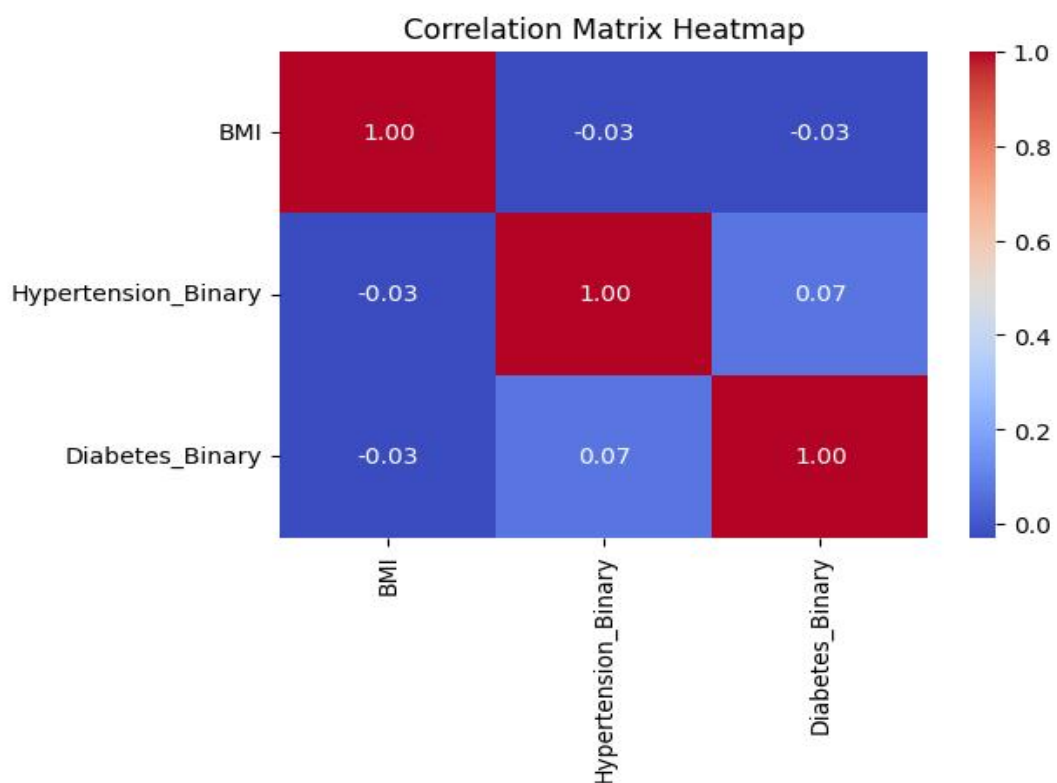
- Persons with no Diabetes and Hypertension are more
- But above 80% have Diabetes and Hypertension



## Correlation Between BMI and Health Conditions:

```
➤ df['Hypertension_Binary']=df['Hypertension'].apply(lambda x: 1 if x
== "Yes" else 0)
df['Diabetes_Binary']=df['Diabetes'].apply(lambda x: 1 if
x == "Yes" else 0)
correlation_matrix=df[['BMI', 'Hypertension_Binary',
'Diabetes_Binary']].corr()
correlation_matrix
```

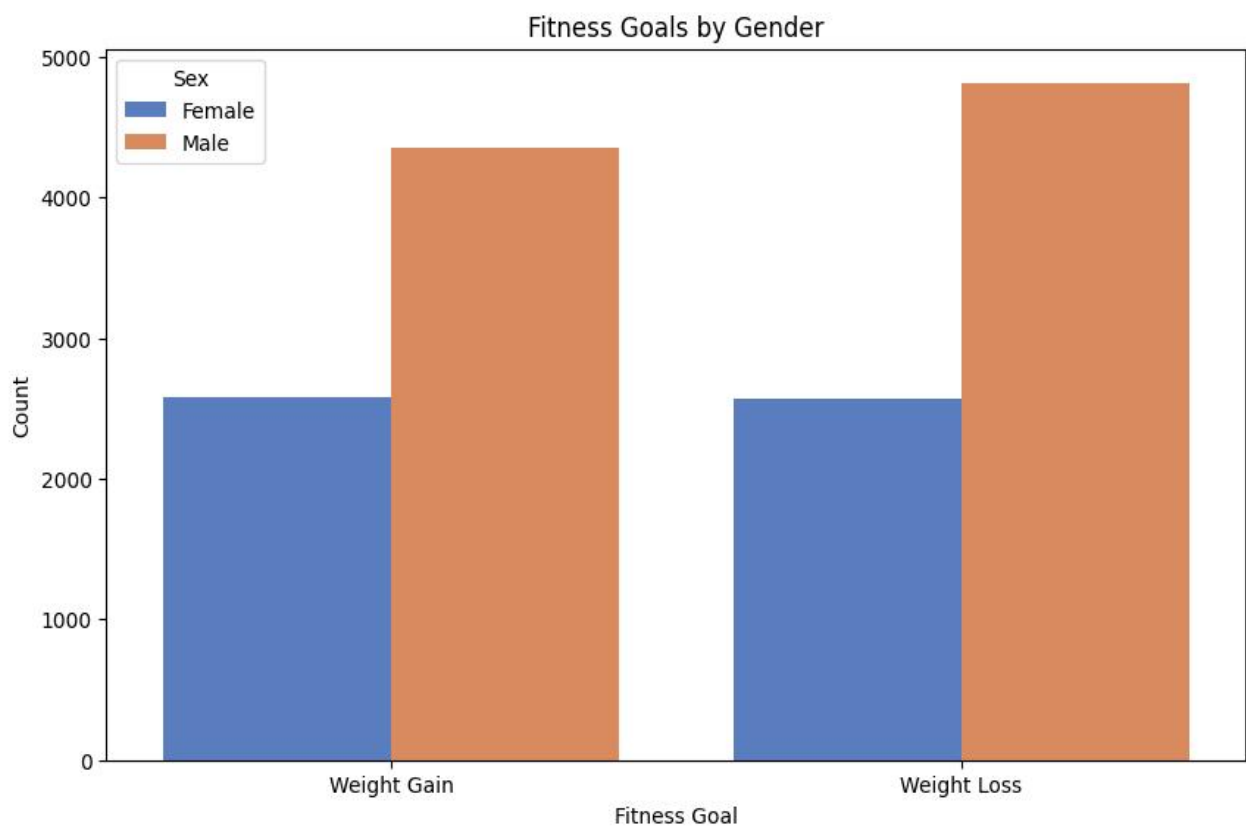
	BMI	Hypertension	Diabetes
BMI	1.0000	-0.0278	-0.0275
Hypertension	-0.0278	1.0000	0.0698
Diabetes	-0.0275	0.0698	1.0000



## Fitness Goals by Gender:

```
➤ fitness_goals_by_gender=df.groupby(['Sex','Fitness  
Goal']).size().reset_index(name='Count')  
fitness_goals_by_gender
```

	Sex	Fitness Goal	Count
0	Female	Weight Gain	2578
1	Female	Weight Loss	2571
2	Male	Weight Gain	4349
3	Male	Weight Loss	4815



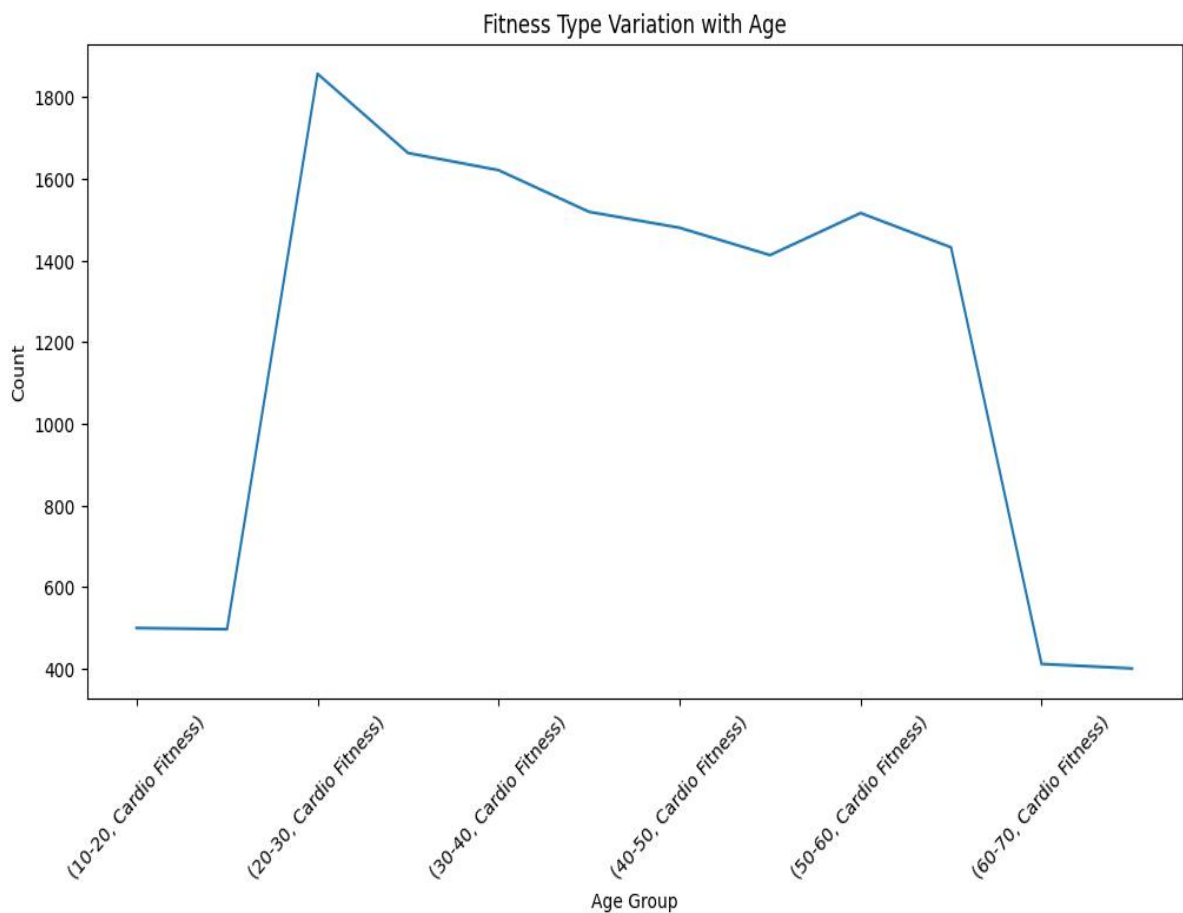
## Key Insights:

- In Males more are try to loss weight
- Females are approximately equal in both gain and loss. But more are try to gain weight

## Fitness Type Variation with Age:

```
➤ fitness_type_by_age=df.groupby(['Age Group','Fitness  
Type']).size().reset_index(name='Count')  
fitness_type_by_age
```

	Age Group	Fitness Type	Count
0	10-20	Cardio Fitness	500
1	10-20	Muscular Fitness	497
2	20-30	Cardio Fitness	1857
3	20-30	Muscular Fitness	1663
4	30-40	Cardio Fitness	1621
5	30-40	Muscular Fitness	1519
6	40-50	Cardio Fitness	1480
7	40-50	Muscular Fitness	1413
8	50-60	Cardio Fitness	1516
9	50-60	Muscular Fitness	1432
10	60-70	Cardio Fitness	412
11	60-70	Muscular Fitness	401



### Key Insights:

- Most of them are doing Cardio Fitness and they are in the age group of 20-30
- 60-70 Age groups do less Muscular Fitness Workouts.
- Also they are less is Cardio Fitness

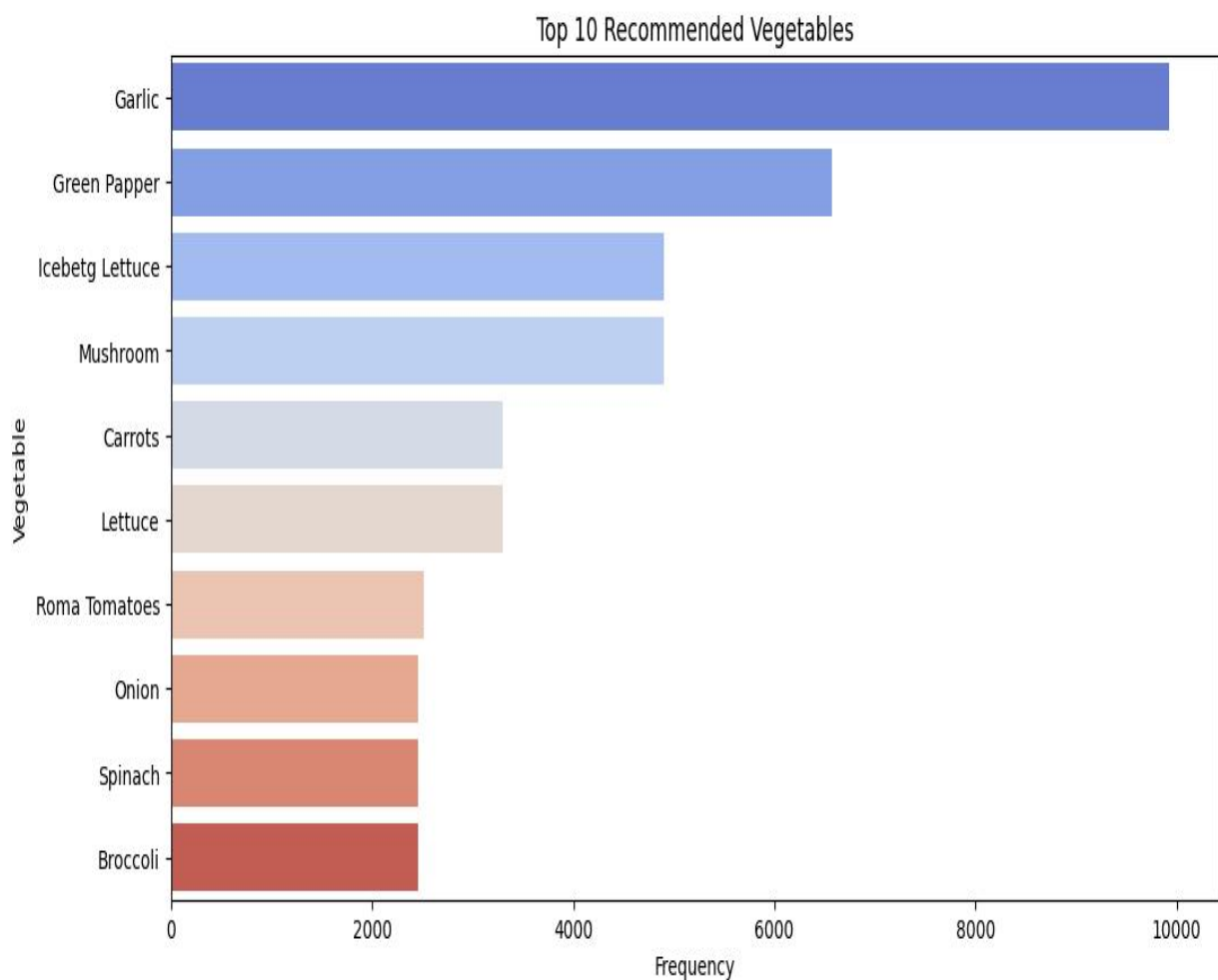
### Most Recommended Vegetables and Proteins:

```
➤ vegetable_recommendations = Counter(", ".join(df['Vegetables']).split(", 
"))
vegetable_freq_table=
pd.DataFrame.from_dict(vegetable_recommendations,
orient="index").reset_index()
vegetable_freq_table.columns=["Vegetable","Frequency"]
```

```
vegetable_freq_table.sort_values(by="Frequency",ascending=False,inplace=True)
```

```
vegetable_freq_table.reset_index(inplace=True,drop=True)
```

```
vegetable_freq_table
```



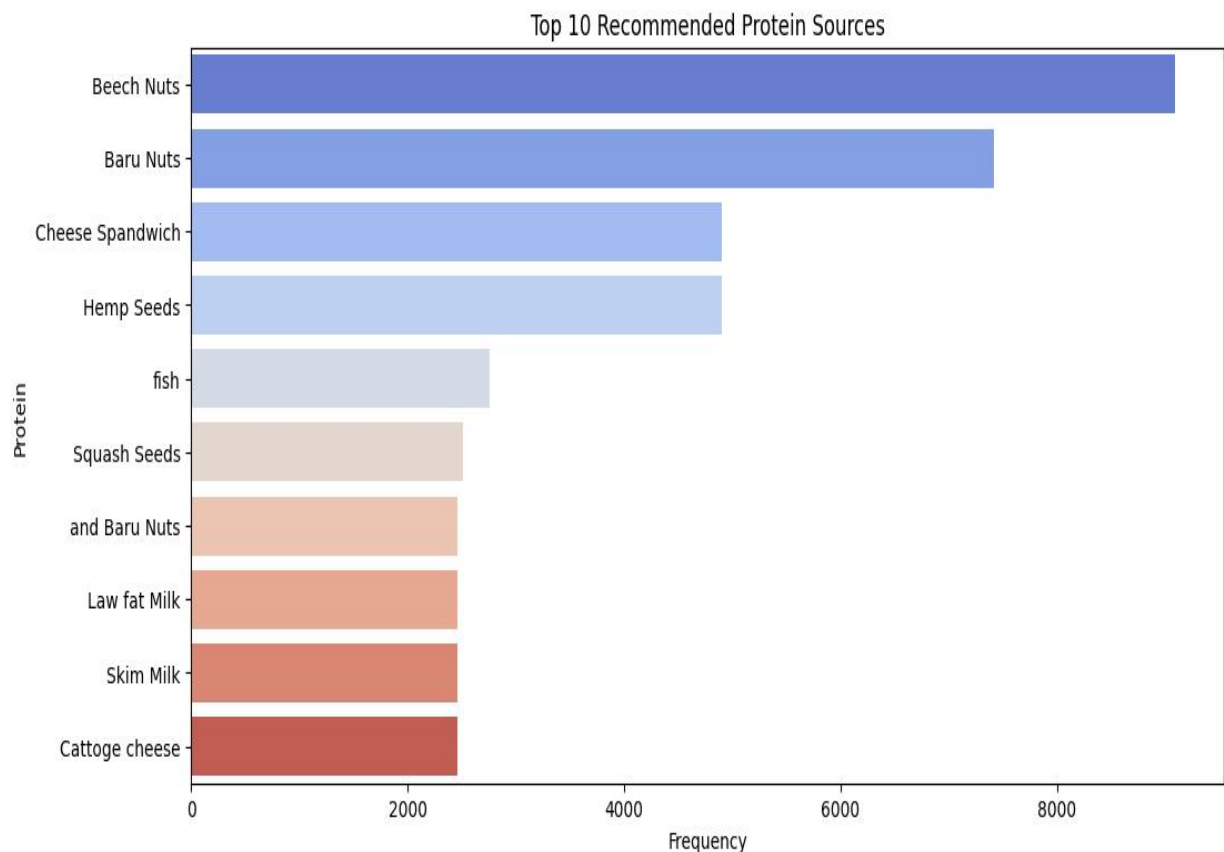
### **Key Insights:**

- Garlic is most used vegetable
- Green Papper ,Icebetg Lettuce and Mushroom also Recommended

```

➤ protein_recommendations = Counter(",
    ".join(df['Protein_Intake']).split(", "))
    protein_freq_table =
pd.DataFrame.from_dict(protein_recommendations,
    orient="index").reset_index()
    protein_freq_table.columns = ["Protein", "Frequency"]
    protein_freq_table.sort_values(by="Frequency", ascending=False,
    inplace=True)
    protein_freq_table.reset_index(inplace=True, drop=True)
    protein_freq_table

```

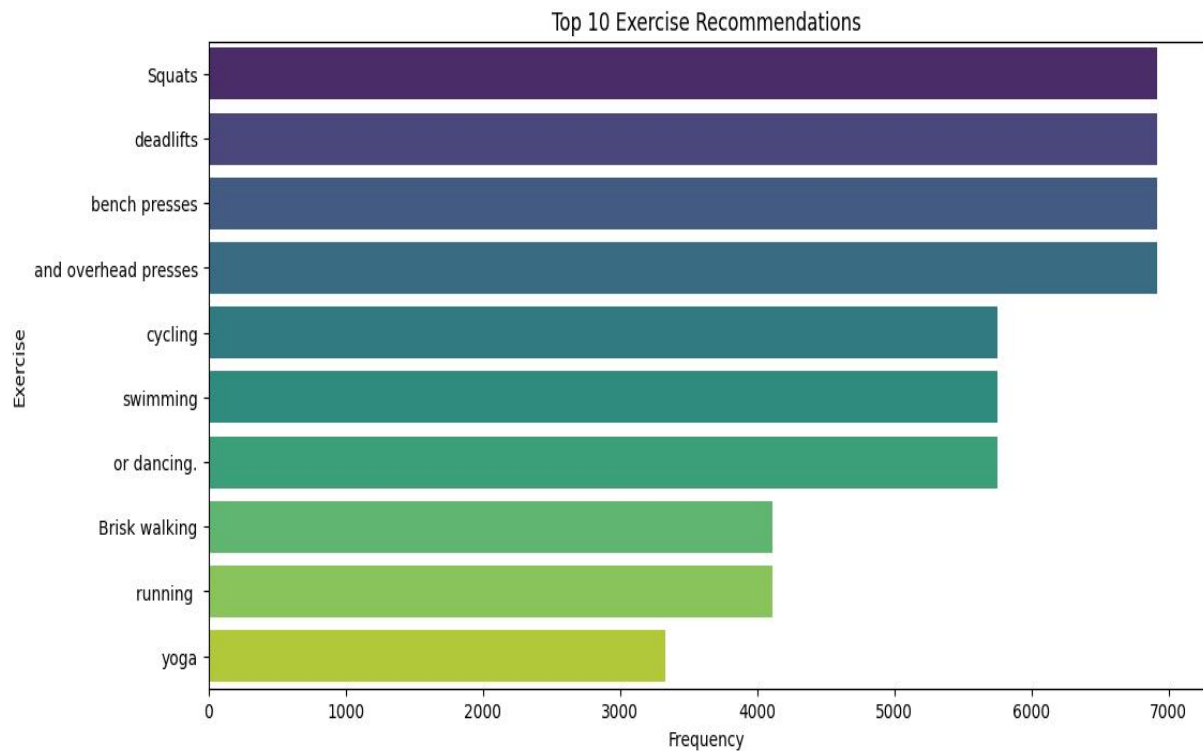


### Key Insights:

- Beech Nuts and Baru Nuts are mostly use proteins
- Cattoge cheese was used less



## TOP 10 WORKOUTS :

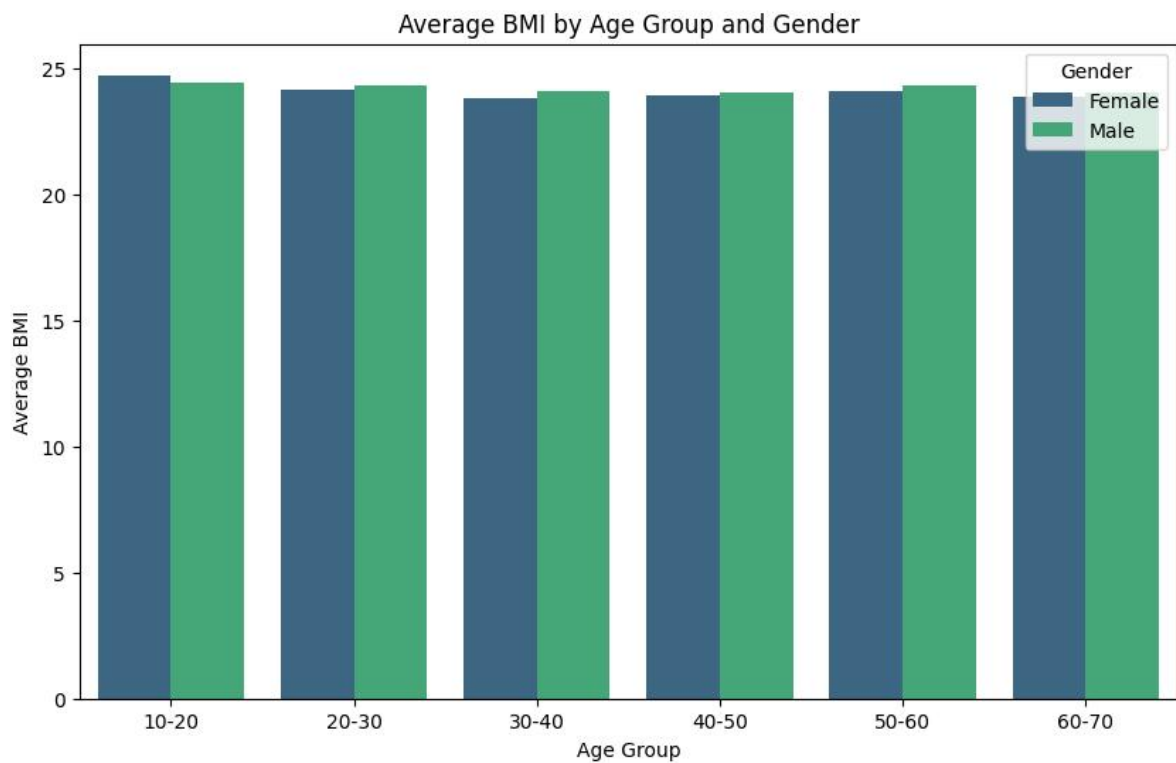


### Key Insights:

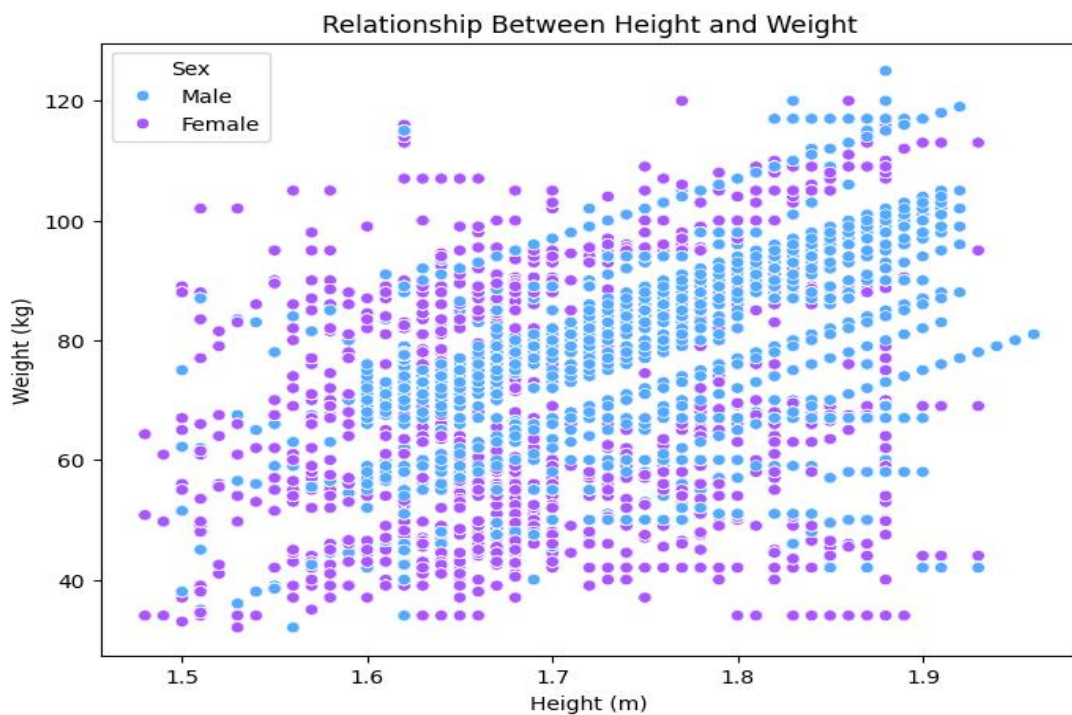
- Squats,Deadlifts,Bench Presses,overhead presses are Highly recommended Workouts
- Yoga has less demand

### Average BMI by Age Group and Gender:

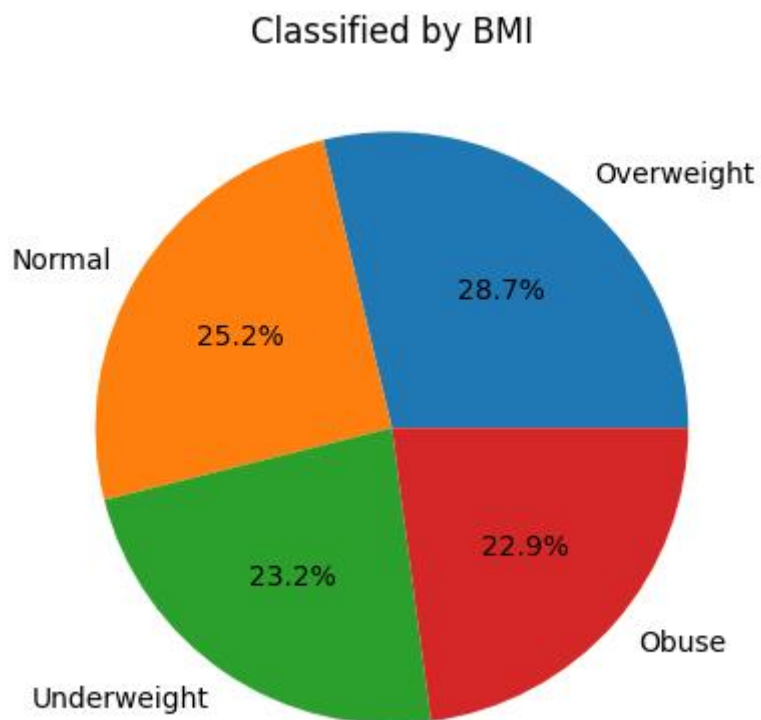
```
➤ average_bmi_table=df.groupby(['Age  
Group','Sex'])['BMI'].mean().reset_index(name='Count')  
average_bmi_table
```



## Relationship Between Weight and Height:



### Pie-Plot of BMI Level:



### Key Insights:

- Overweight are more
- 25% are Normal BMI

# CONCLUSION

This analysis provides valuable insights into the relationship between health, fitness, and demographic factors based on the given dataset. The study highlights the importance of monitoring BMI and its strong association with health conditions like hypertension and diabetes. By categorizing individuals into age groups and fitness goals, it was possible to observe trends in exercise preferences and health conditions across different demographics.

The findings also emphasize the need for targeted fitness programs and lifestyle interventions, especially for groups at higher risk of chronic conditions. For example, age groups with higher BMIs may benefit from tailored exercise recommendations to lower their risk of hypertension and diabetes. Additionally, differences in fitness goals between genders and age groups suggest that health professionals should consider these factors when designing fitness plans.

- Doing Workout Mostly in the Age Group of 20-30
- Males are more number on doing workout
- Most of them not have the normal BMI
- Squats, Deadlifts, Bench Presses, overhead presses are Highly recommended Workouts