Health Aware Obesity Data Analysis Report

02-2024| By: Emmanuel Jacob

Table of Contents

Executive Summary

The exploratory data analysis (EDA) conducted on the dataset collected by HealthAware provides valuable insights into the factors influencing the health and lifestyles of young individuals. The analysis focused on descriptive statistics, demographic characteristics, weight, and body mass index (BMI), dietary habits, physical activity, alcohol and smoking habits, and transportation choices.

Key findings from the analysis include:

- The dataset represents a balanced gender distribution with a majority of individuals below the age of 30.
- A substantial portion of individuals in the dataset are overweight or obese, with 46% classified as obese.
- Individuals with a family history of overweight or obesity display a significantly higher likelihood of being overweight themselves.
- An overwhelming 88% of individuals report frequent consumption of high-caloric foods, contributing to weight gain and poor health outcomes.
- Regular physical activity is associated with lower BMI, while sedentary behaviours, especially excessive technology device usage, are correlated to higher BMI.
- Smoking habits and family history of obesity play significant roles in influencing BMI.
- Transportation choices impact BMI, with individuals using active transportation showing healthier weights compared to those relying on automobiles or public transit.

Based on these findings, recommendations for HealthAware's campaign for promoting healthier lifestyles among young people include implementing nutritional education programs, promoting regular physical activity, addressing sedentary behaviours, creating supportive environments for healthy living initiatives, and developing targeted initiatives for high-risk groups.

Introduction

I analysed HealthAware‘s dataset to promote healthier lifestyles among young people through evidence-based interventions and policy initiatives. In alignment with this mission, an exploratory data analysis (EDA) was performed to gain insights into the factors influencing health-related behaviours and outcomes among young individuals.

The dataset under analysis provides comprehensive information on demographic characteristics, weight and weight status, dietary habits, physical activity levels, sedentary behaviours, alcohol and smoking habits, and transportation choices. By analysing descriptive statistics and conducting in-depth analyses, this report aims to identify key trends, patterns, and associations within the dataset to inform targeted interventions and strategies for promoting healthier lifestyles.

Data Description

The dataset contains information collected by HealthAware for their campaign on promoting healthier lifestyles among young people. It includes data from a sample of individuals, capturing various demographic and health-related factors that may influence lifestyle choices and overall health outcomes.

Number of Observations:

The dataset consists of 2111 individual records.

Number of Variables:

There are 17 variables included in the dataset and I have added 3 more.

Gender:

- Indicates the gender of the individual (e.g., Male, Female).
- Gender may influence lifestyle choices and health behaviours, such as dietary preferences and physical activity levels.

Age:

- Represents the age of the individual in years.
- Age is a key demographic factor that can impact health outcomes. It helps in understanding health-related trends across different age groups.

Height (m):

- The height of the individual in meters.
- Height is a measure of physical stature and can be correlated with body mass index (BMI) calculations to assess weight status and overall health.

Weight (Kg):

- The weight of the individual in kilograms.
- Weight is an important indicator of overall health and can be used in combination with height to calculate BMI and assess weight status.

Family History with Overweight:

- Indicates whether the individual has a family history of overweight or obesity (e.g., Yes, No).

- Family history can be a significant risk factor for overweight and obesity, as genetics play a role in predisposition to certain health conditions.

Frequent consumption of high-caloric food (FAVC):

- Reflects the frequency of consumption of high-calorie foods.
- High-caloric food consumption is associated with weight gain and increased risk of obesity and related health conditions.

Frequency of consumption of vegetables (FCVC):

- Indicates how often the individual consumes vegetables.
- Vegetable consumption is a key component of a healthy diet, providing essential nutrients and fibre that contribute to overall health and well-being.

Number of main meals (NCP):

- Represents the number of main meals consumed per day.
- Meal frequency may impact metabolic rate and energy balance, influencing weight management and overall health outcomes.

Consumption of food between meals (CAEC):

- Indicates the frequency of food consumption between main meals.
- Snacking behaviour can affect overall calorie intake and nutritional balance, influencing weight status and metabolic health.

Consumption of water daily (CH20):

- Reflects the amount of water consumed by the individual daily.
- hydration is essential for maintaining optimal health and supporting various bodily functions, including metabolism and digestion.

Consumption of alcohol (CALC):

- Indicates whether the individual consumes alcohol (e.g., Yes, No).
- Alcohol consumption can impact health outcomes, including weight management, liver health, and risk of chronic diseases.

Calorie consumption monitoring (SCC):

- Reflects whether the individual monitors their calorie consumption (e.g., Yes, No).
- Calorie monitoring can promote awareness of dietary habits and support weight management goals by tracking energy intake.

Physical activity frequency (FAF):

- Represents the frequency of physical activity per week.
- Regular physical activity is associated with numerous health benefits, including weight management, cardiovascular health, and mental well-being.

Time using technology devices (TUE):

- Indicates the amount of time spent using technology devices daily.
- Excessive screen time may contribute to sedentary behaviour and negative health outcomes, such as obesity and poor sleep quality.

Transportation used (MTRANS):

- Reflects the mode of transportation typically used by the individual (e.g., Walking, Car, Public transport).
- Transportation choices can influence physical activity levels and overall energy expenditure, impacting weight management and cardiovascular health.

Ethics and Compliance

It is important that personal data is collected and processed lawfully, with individuals provided with clear and transparent information regarding the purposes of data processing.

strict access controls and data minimization practices to ensure that only authorised personnel have access to personal data and that data is only used for the specific purposes for which it was collected. Additionally, keeping a comprehensive record of data processing activities to demonstrate compliance with DPA requirements.

In safeguarding the confidentiality, integrity, and availability of personal data, implementing strong measures to keep data safe and secure. This includes encryption of sensitive data both in transit and in storage, regular audits to identify and mitigate potential vulnerabilities, and the implementation of access controls to restrict unauthorised access to data. Furthermore, providing ongoing staff training on data security best practices will ensure that all employees understand their responsibilities in maintaining data security.

Methodology

Methods

Descriptive statistics

Descriptive statistics were utilised to summarise and describe the characteristics of the dataset. This included measures such as mean, median, mode, standard deviation, and range for numerical variables, as well as frequency distributions for categorical variables.

Descriptive statistics were used to gain an initial understanding of the dataset's characteristics and distribution of variables, providing valuable insights into the population under study.

Comparative analysis

Comparative analysis techniques were utilised to compare different groups within the dataset. This included comparing demographic groups (e.g., gender, and age) and examining differences in health-related factors and behaviours.

Comparative analysis techniques were used to test differences between demographic groups and identify potential differences in health related.

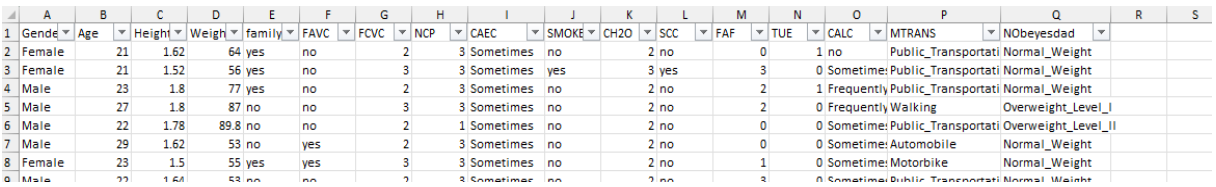
Data Preprocessing

Before conducting the analysis, several data preprocessing steps were implemented to ensure data quality and prepare the dataset for analysis:

Missing values were identified and handled appropriately. In this dataset, I mainly deleted it since it was missing rows.

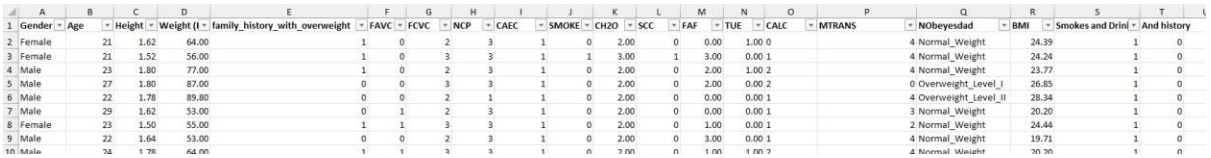
Data transformation techniques, such as normalisation or standardisation, were applied to ensure that variables were on a similar scale and met the requirement for statistical analysis.

Categorical variables were encoded into a numerical format using techniques such as one-hot encoding or label encoding to facilitate analysis with statistical algorithms.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Gende	Age	Height	Weigh	family	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad		
2	Female	21	1.62	64	yes	no	2	3	Sometimes	no	2	no	0	1	no	Public_Transportati	Normal_Weight		
3	Female	21	1.52	56	yes	no	3	3	Sometimes	yes	3	yes	3	0	Sometime:	Public_Transportati	Normal_Weight		
4	Male	23	1.8	77	yes	no	2	3	Sometimes	no	2	no	2	1	Frequently	Public_Transportati	Normal_Weight		
5	Male	27	1.8	87	no	no	3	3	Sometimes	no	2	no	2	0	Frequently	Walking	Overweight_Level_I		
6	Male	22	1.78	89.8	no	no	2	1	Sometimes	no	2	no	0	0	Sometime:	Public_Transportati	Overweight_Level_II		
7	Male	29	1.62	53	no	yes	2	3	Sometimes	no	2	no	0	0	Sometime:	Automobile	Normal_Weight		
8	Female	23	1.5	55	yes	yes	3	3	Sometimes	no	2	no	1	0	Sometime:	Motorbike	Normal_Weight		
9	Male	22	1.64	53	no	no	2	3	Sometimes	no	2	no	3	0	Sometime:	Public_Transportati	Normal_Weight		

Before



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Gender	Age	Height	Weight (k	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad	BMI	Smokes and Drini	And history
2	Female	21	1.62	64.00	1	0	2	3	1	0	2.00	0	0.00	1.00	0	4	Normal_Weight	24.39	1	0
3	Female	21	1.52	56.00	1	0	3	3	1	1	3.00	1	3.00	0.00	1	4	Normal_Weight	24.24	1	0
4	Male	23	1.80	77.00	1	0	2	3	1	0	2.00	0	2.00	1.00	2	4	Normal_Weight	23.77	1	0
5	Male	27	1.80	87.00	0	0	3	3	1	0	2.00	0	2.00	0.00	2	0	Overweight_Level_I	26.85	1	0
6	Male	22	1.78	89.80	0	0	2	1	1	0	2.00	0	0.00	0.00	1	4	Overweight_Level_II	28.34	1	0
7	Male	29	1.62	53.00	0	1	2	3	1	0	2.00	0	0.00	0.00	1	3	Normal_Weight	20.20	1	0
8	Female	23	1.50	55.00	1	1	3	3	1	0	2.00	0	1.00	0.00	1	2	Normal_Weight	24.44	1	0
9	Male	22	1.64	53.00	0	0	2	3	1	0	2.00	0	3.00	0.00	1	4	Normal_Weight	19.71	1	0
10	Male	24	1.78	64.00	1	1	3	3	1	0	2.00	0	1.00	1.00	2	4	Normal_Weight	20.20	1	0

After

Added additional columns such as BMI, individuals who smoke and consume alcohol, and individuals who smoke and consume alcohol with a family history of obesity.

Hypothesis

Null Hypothesis: There is no relationship between transportation choices and BMI among young people.

Alternate Hypothesis: Individuals who use active transportation methods (e.g., walking or cycling) have lower BMI compared to those who use passive transportation methods (e.g., car or public transport).

Exploratory Data Analysis

Descriptive Statistics

Descriptive Statistics	*BMI*	Weight (Kg)	*Age*	*Time using technology*	*Physical activity frequency*
Mean	29.70	86.59	24.31	0.66	1.01
Standard Error	0.17	0.57	0.14	0.01	0.02
Median	28.72	83.00	22.78	0.63	1.00
Mode	26.67	80.00	18.00	0.00	0.00
Standard Deviation	8.01	26.19	6.35	0.61	0.85
Sample Variance	64.18	685.98	40.27	0.37	0.72
Kurtosis	-0.81	-0.70	2.83	-0.55	-0.62
Skewness	0.15	0.26	1.53	0.62	0.50
Range	37.81	134.00	47.00	2.00	3.00
Minimum	13.00	39.00	14.00	0.00	0.00
Maximum	50.81	173.00	61.00	2.00	3.00
Sum	62697.04	182783.17	51323.90	1388.75	2132.74
Count	2111.00	2111.00	2111.00	2111.00	2111.00
Coefficient of variation	26.97	30.25	26.10	92.56	84.19

BMI (Body Mass Index)

- The average BMI is 29.70, indicating that the average individuals in the dataset are overweight.
- The median BMI is slightly lower than the average, suggesting a slightly larger number of individuals with lower BMI.
- The range of BMI is 37.81 which is relatively wide, indicating variability in weight status within the sample.

Weight (Kg)

- The average weight is 86.59 kg, with a standard deviation of 26.19 kg, indicating considerable variability in weights.
- The median weight is lower than the average, suggesting that more individuals are below average with potential outliers on the higher end.
- The range of weight is 134.00 kg, suggesting variability in weight status within the Sample.

Age

- The average age of individuals in the dataset is 24.31 years, with a standard deviation of 6.35 years, indicating relatively low variability in age compared to weight and BMI.
- The distribution of age appears to be on the lower end, with a median age of 22.78 years, lower than the average age.
- The age range is 47 years suggests that the sample includes individuals across a wide age range.

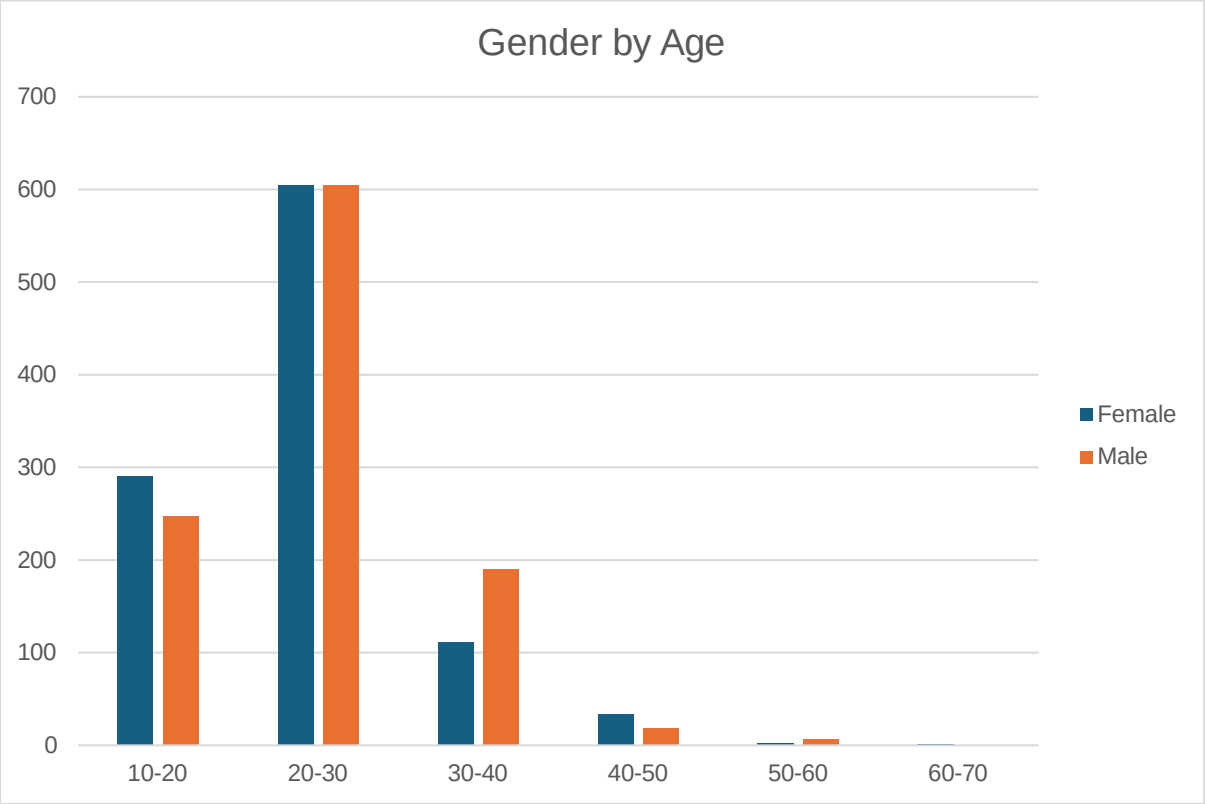
Time using technology.

- The average time spent using technology devices daily is 0.66 hours, indicating that, on average, individuals spend less than an hour using technology devices per day.
- The distribution of time using technology appears to be positively biased, with a mode of 0.00 hours, suggesting that most of the individuals spend minimal time using technology devices.
- The range of time using technology is 2 hours indicating variability in technology usage habits within the sample.

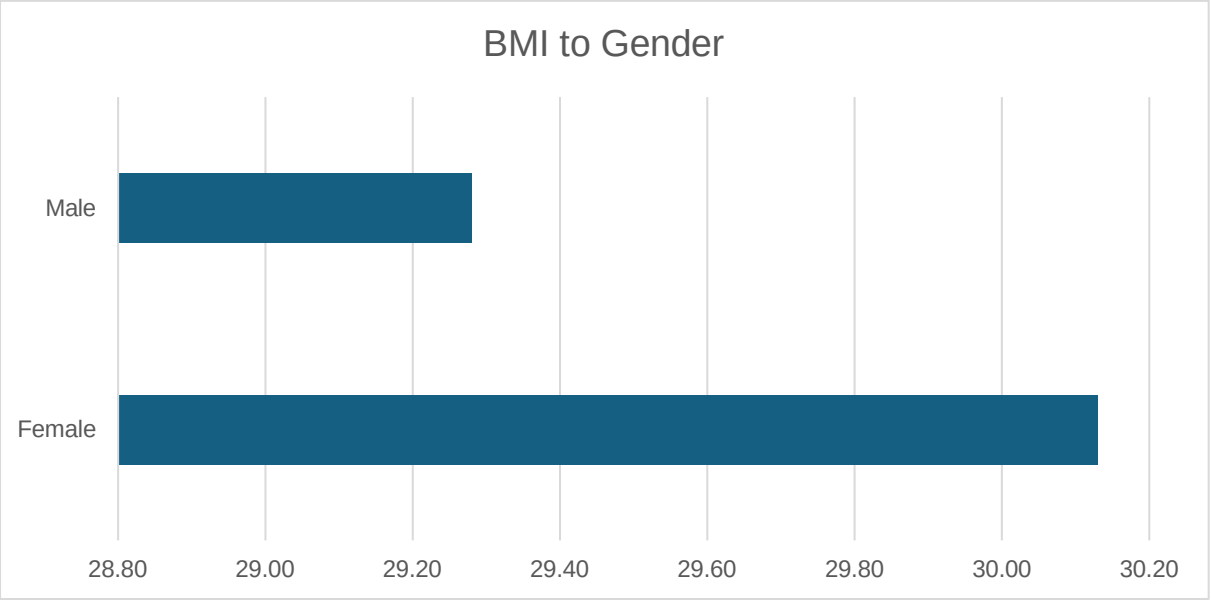
Physical activity frequency

- The average frequency of physical activity per week is 1.01, indicating that, on average, individuals engage in physical activity slightly more than once a week.
- With a mode of 0.00, suggests that many individuals are sedentary.
- The range of physical activity frequency is 3.

Demographic Analysis

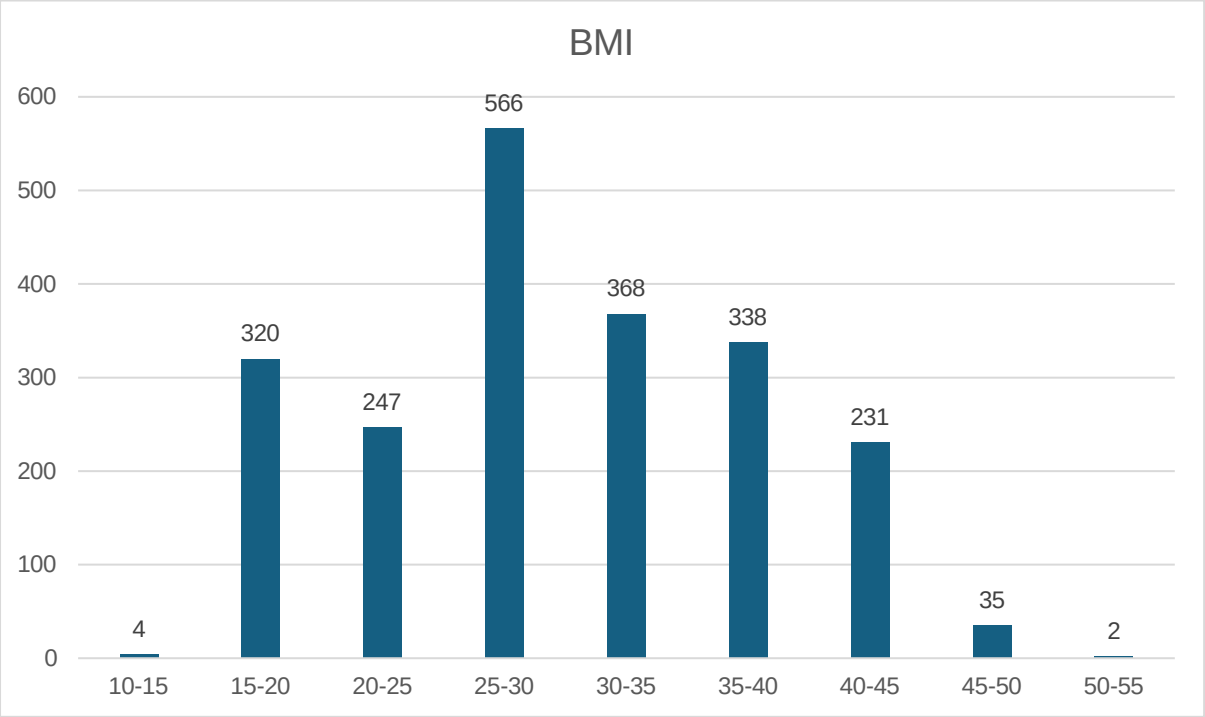


The column chart shows the age and gender distribution in the dataset. As you can see the gender distribution is close to even, with most of the individuals being below 30.



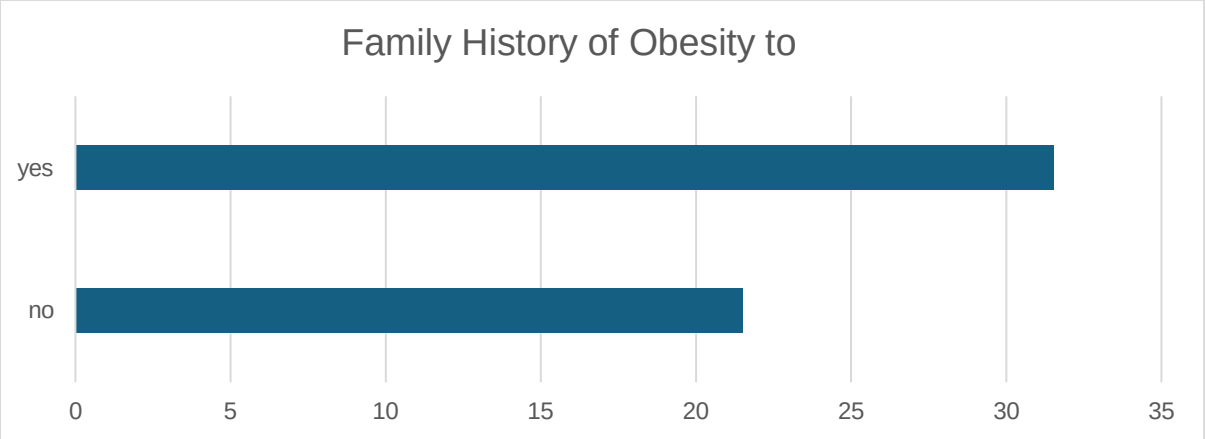
Using the gender distribution to BMI I can see that females on average have a higher BMI compared to males.

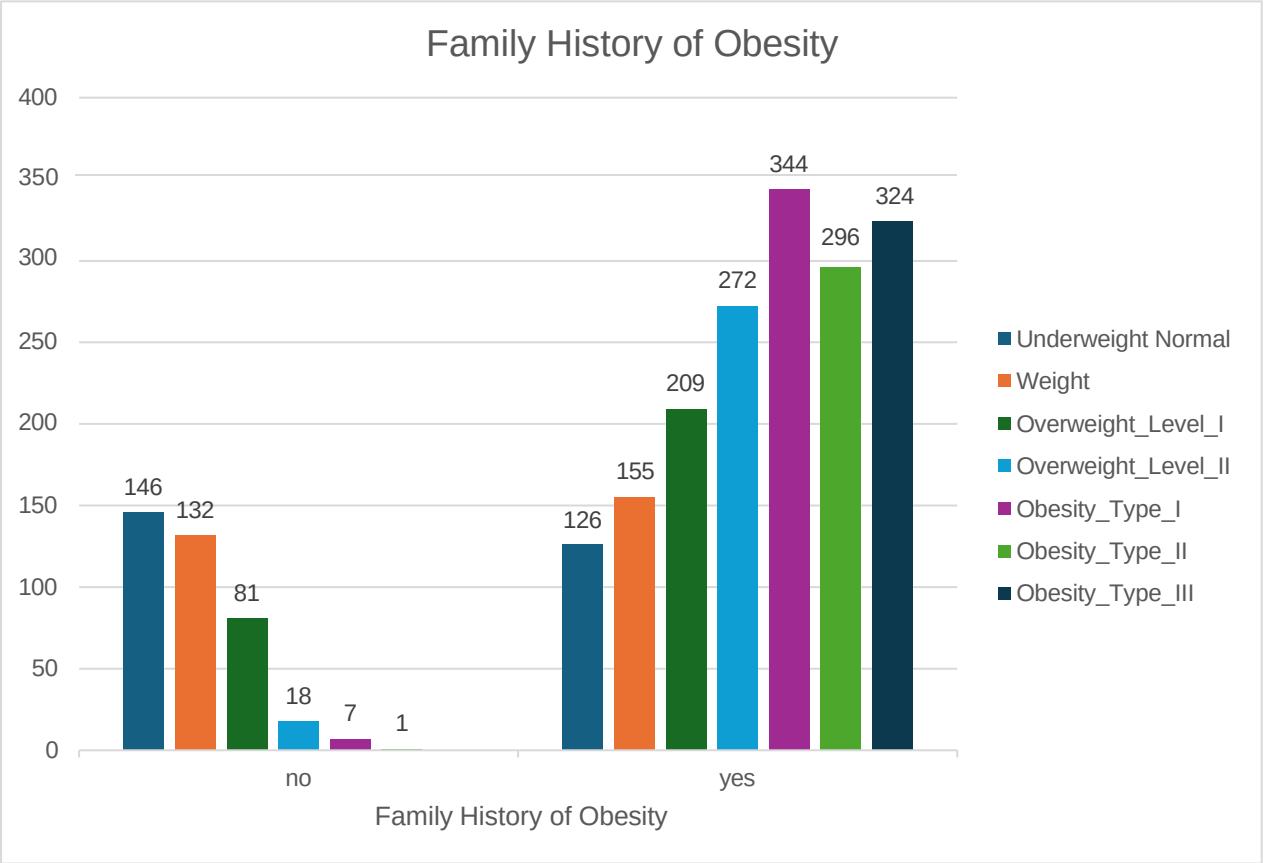
Weight and Body Composition



The distribution of BMI indicates that 27% of the individuals fall within the normal range, 27% are overweight, and 46% are obese.

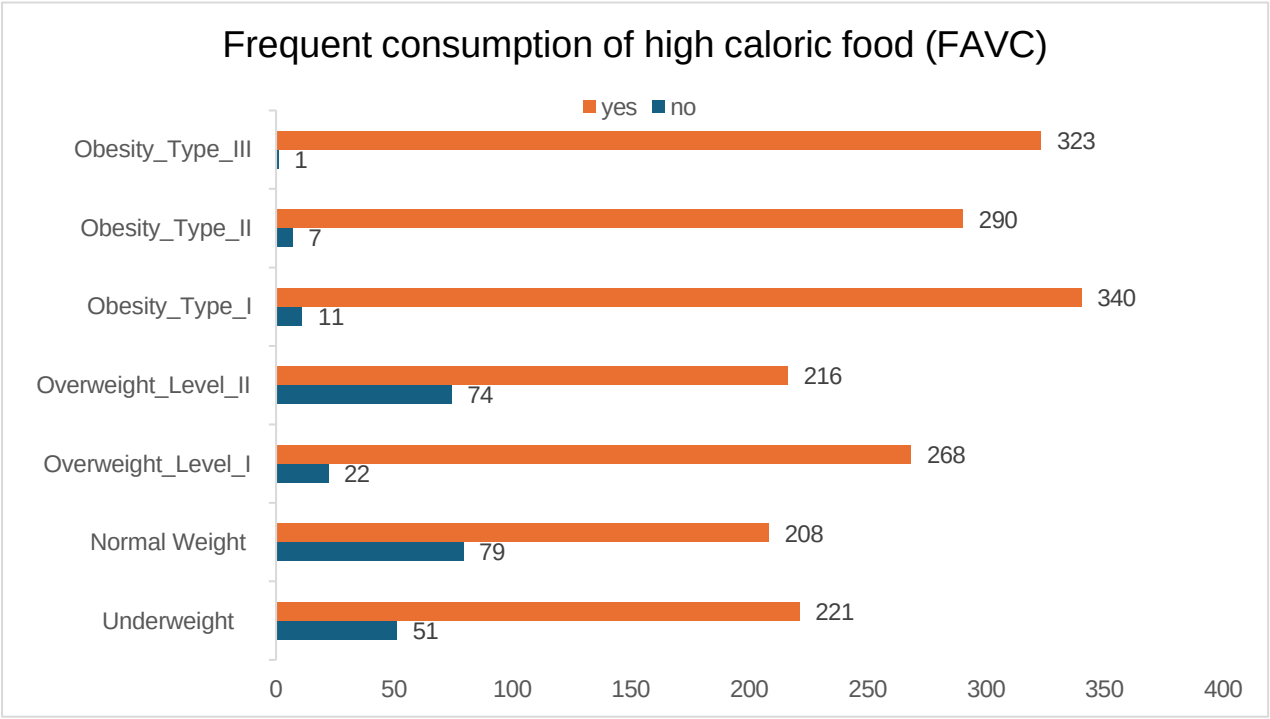
Family History of Obesity



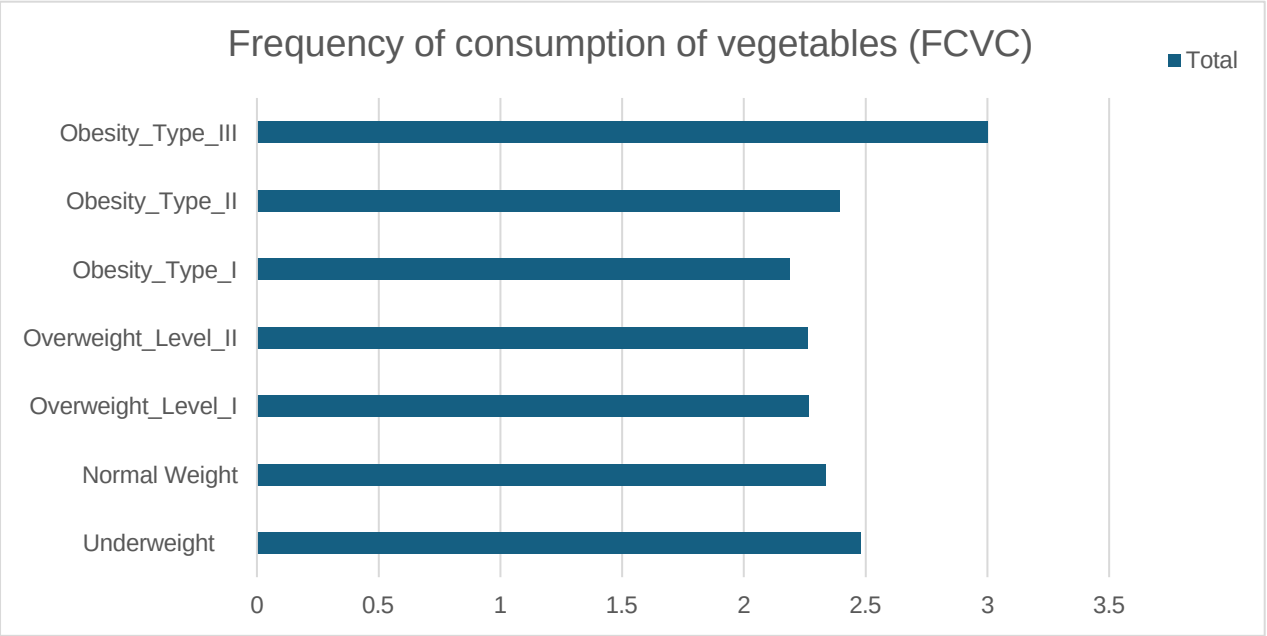


Individuals with a family history of overweight are 68% more likely to be overweight or obese compared to those without.

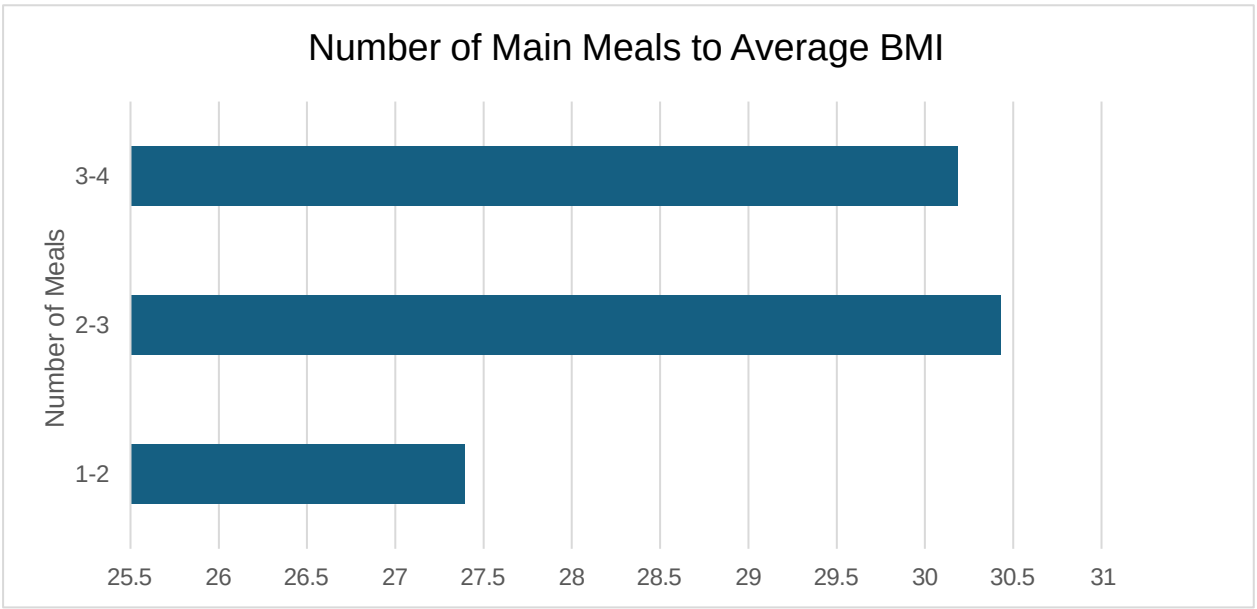
Dietary Habits Analysis



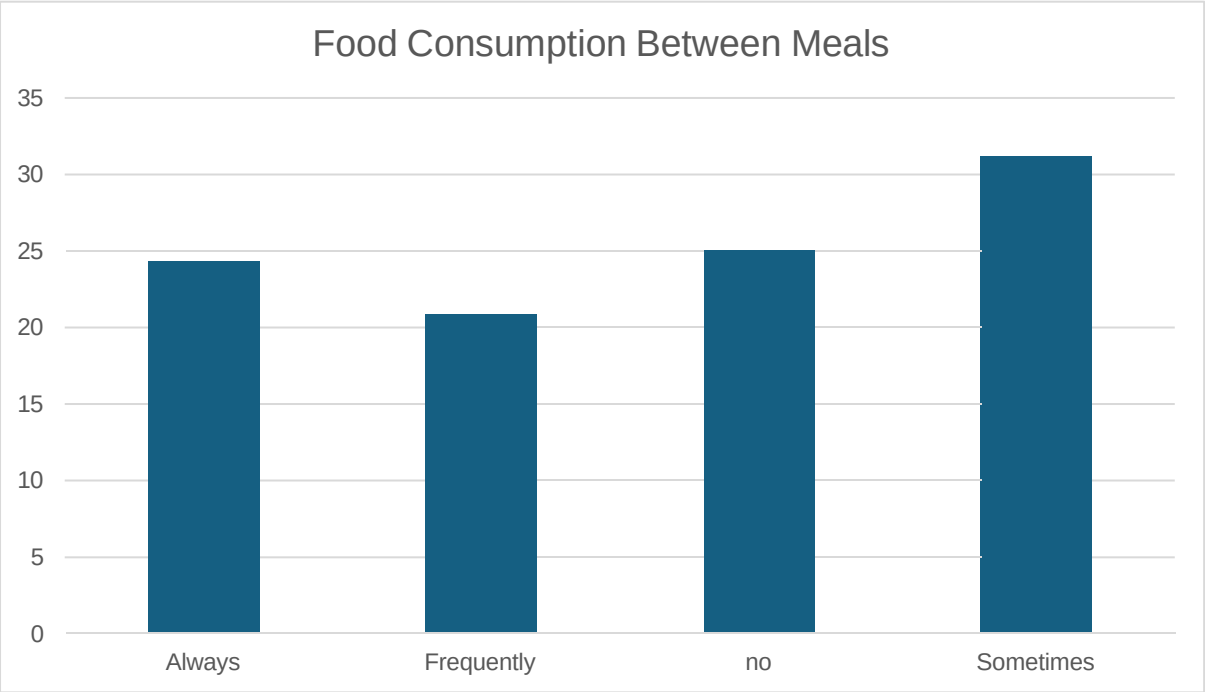
88% of individuals reported frequent consumption of high-caloric foods, which may contribute to weight gain and poor health outcomes. There is a positive correlation between frequent consumption of high-caloric foods and weight status.



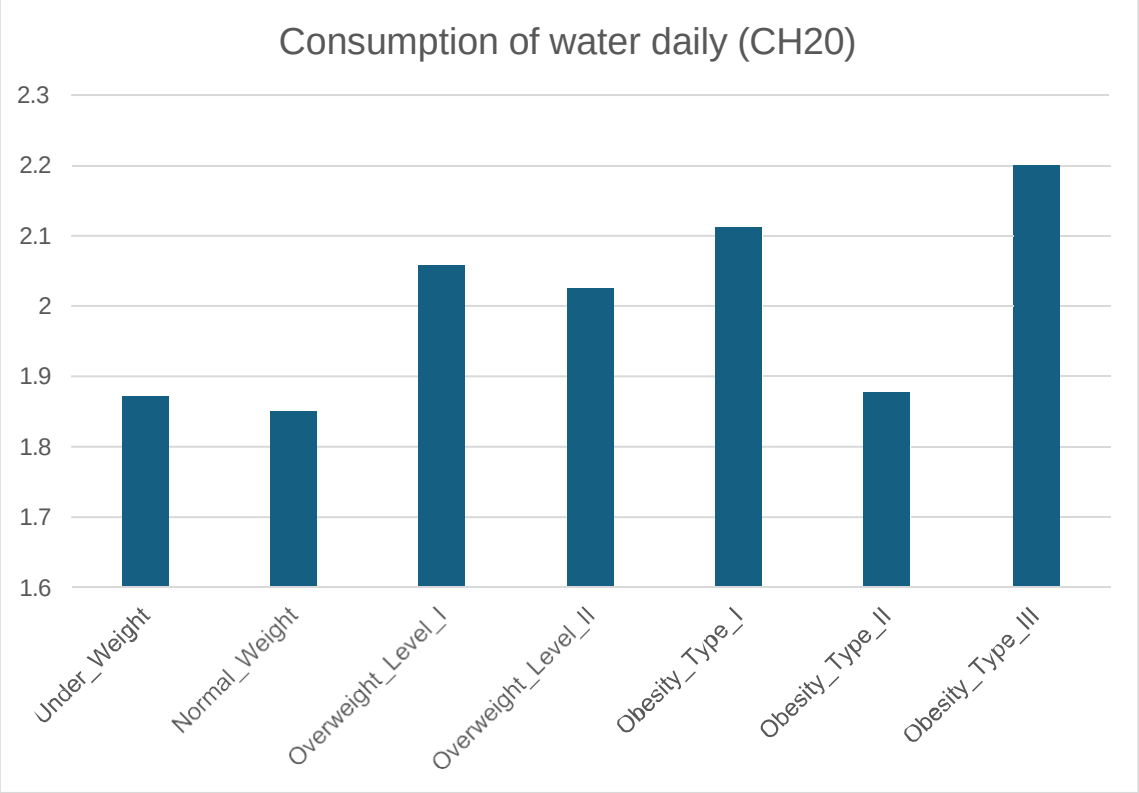
There is no correlation between the frequency of consumption of vegetables and weight status. Further data is needed on the types of vegetables such as starchy or fibrous. For the graph above I predict that most vegetables consumed are starchy such as potatoes.



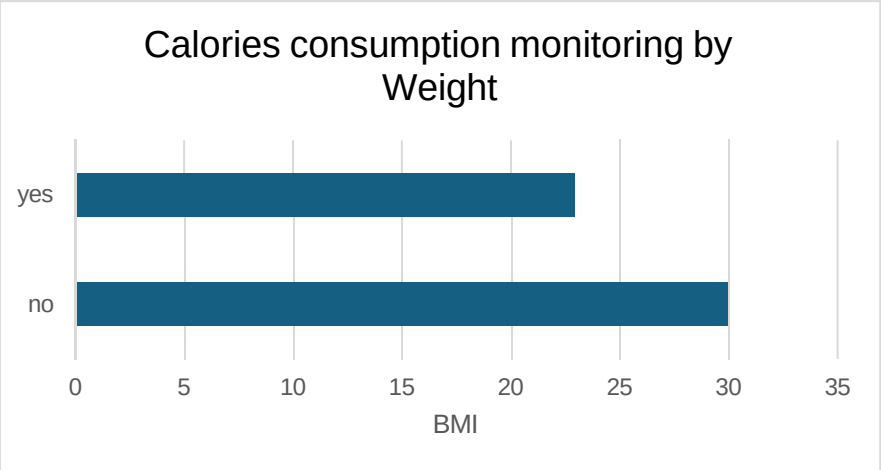
68% of the individuals have 3 to 4 meals daily. Which may contribute to weight gain and poor health outcomes. Individuals who eat 1 to 2 meals daily have a significantly lower BMI compared to 2+ meals. Further data on the types of diets individuals have can lead to more accurate insights.



69% of individuals consume food sometimes between meals. Which may contribute to weight gain and poor health outcomes. 43% of Individuals in normal or below consume food frequently or always between meals compared to 3% in overweight or above. This could mean frequent snacking can help maintain a healthy BMI. Again, further data into the types of snacks can provide more insights.

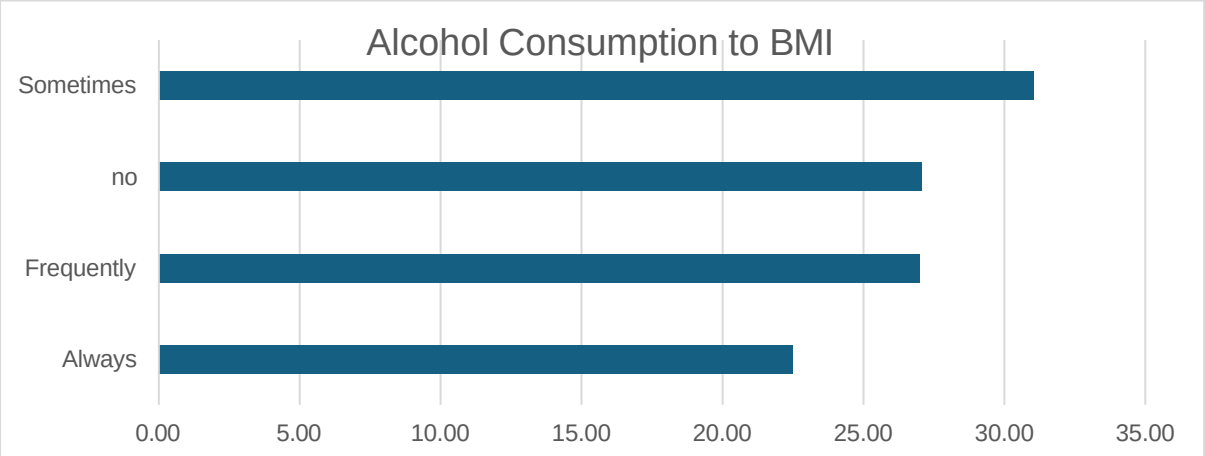


The average consumption of water is higher in overweight and obese individuals compared to normal and underweight individuals.

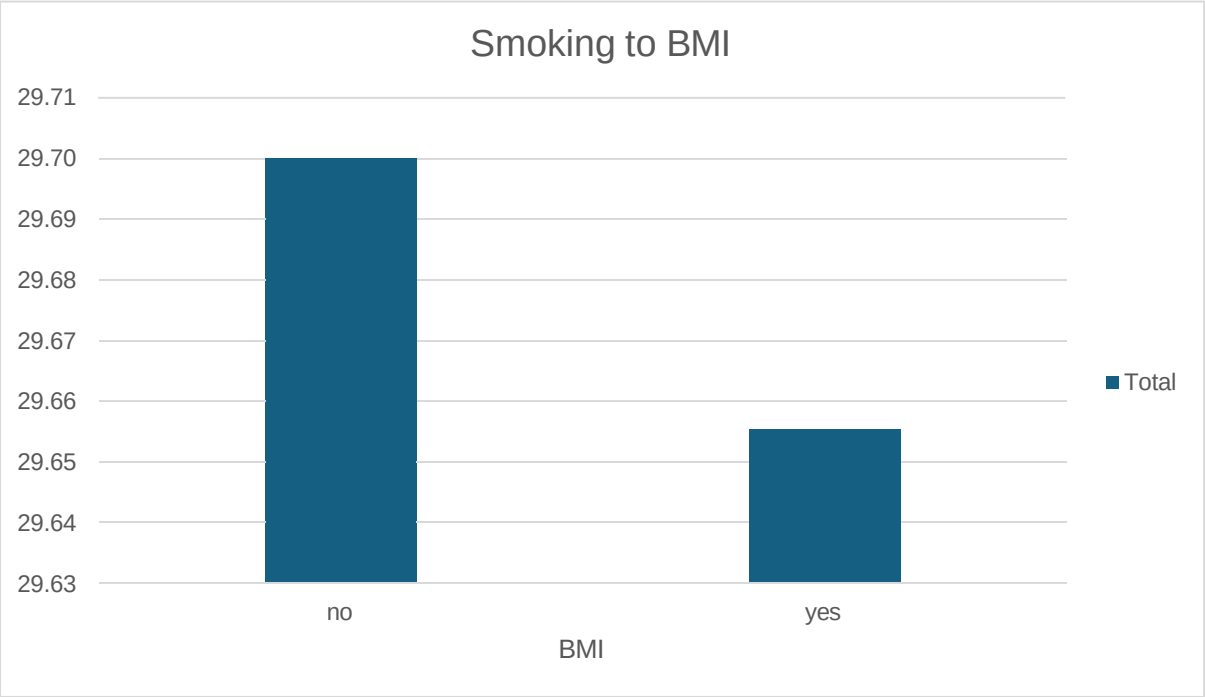


The average BMI of an individual who monitors their calorie consumption is 7 points lower than an individual who doesn't.

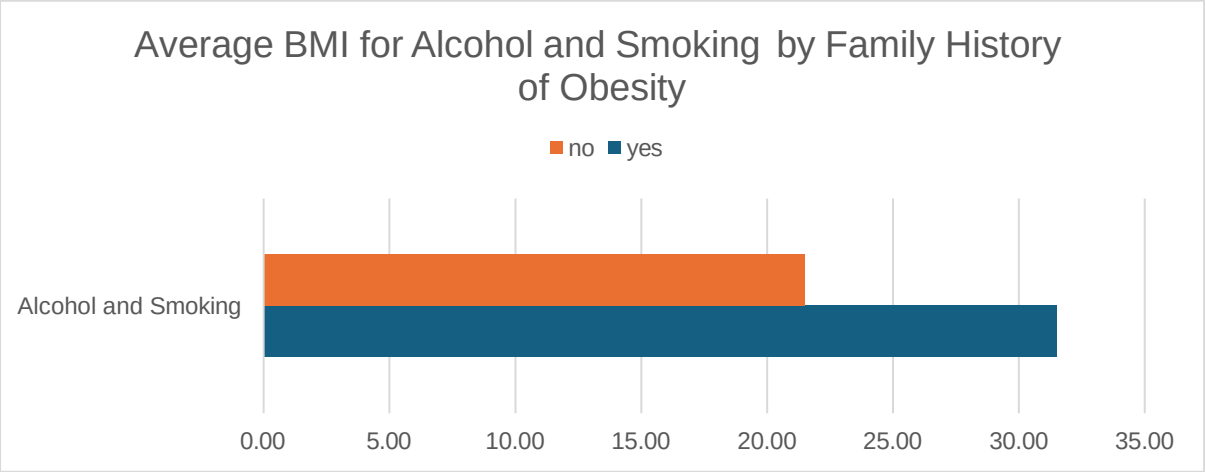
Alcohol and Smoking



There is no significant correlation between average BMI and Alcohol consumption. Data on the type if alcohol individuals consume the most can maybe provide better insights.

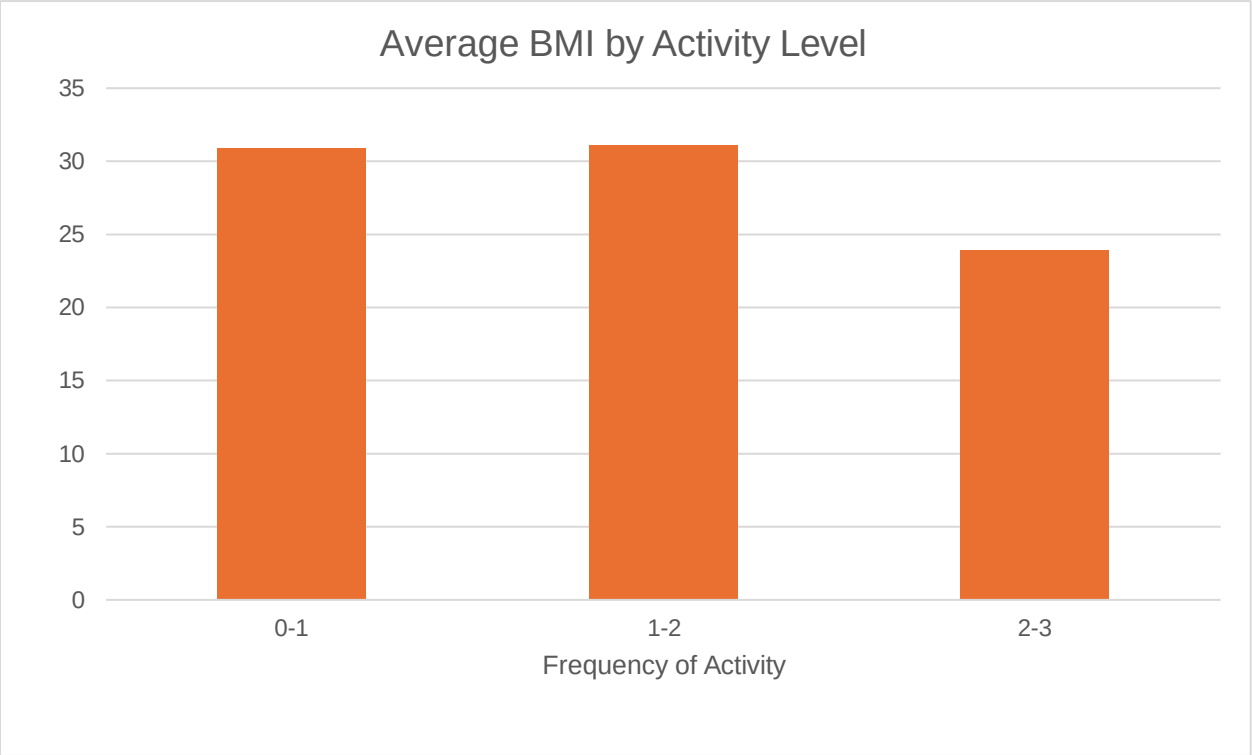


There is a slight correlation between smoking and BMI. This might be because smoking can suppress hunger. Data on other hunger suppressants such as coffee can prove to be insightful.

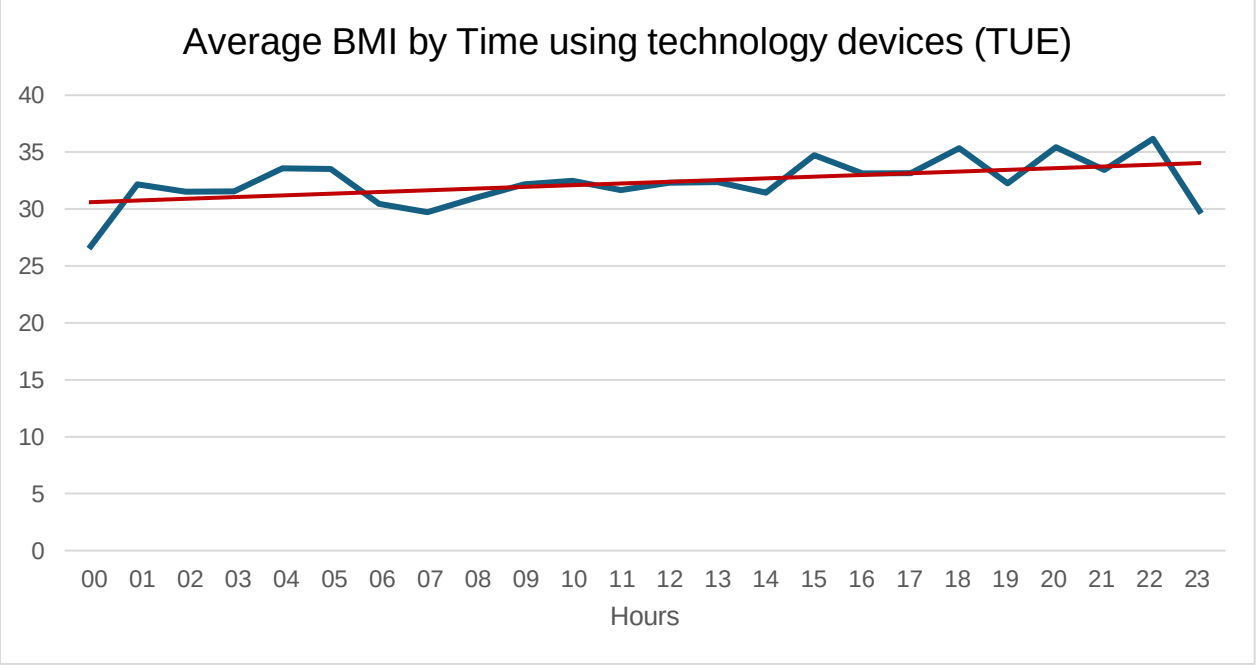


Individuals who smoke and consume alcohol are more likely to have higher BMI. However, individuals with a family history of obesity are significantly more likely to be affected.

Physical Activity and Sedentary Behaviours Analysis

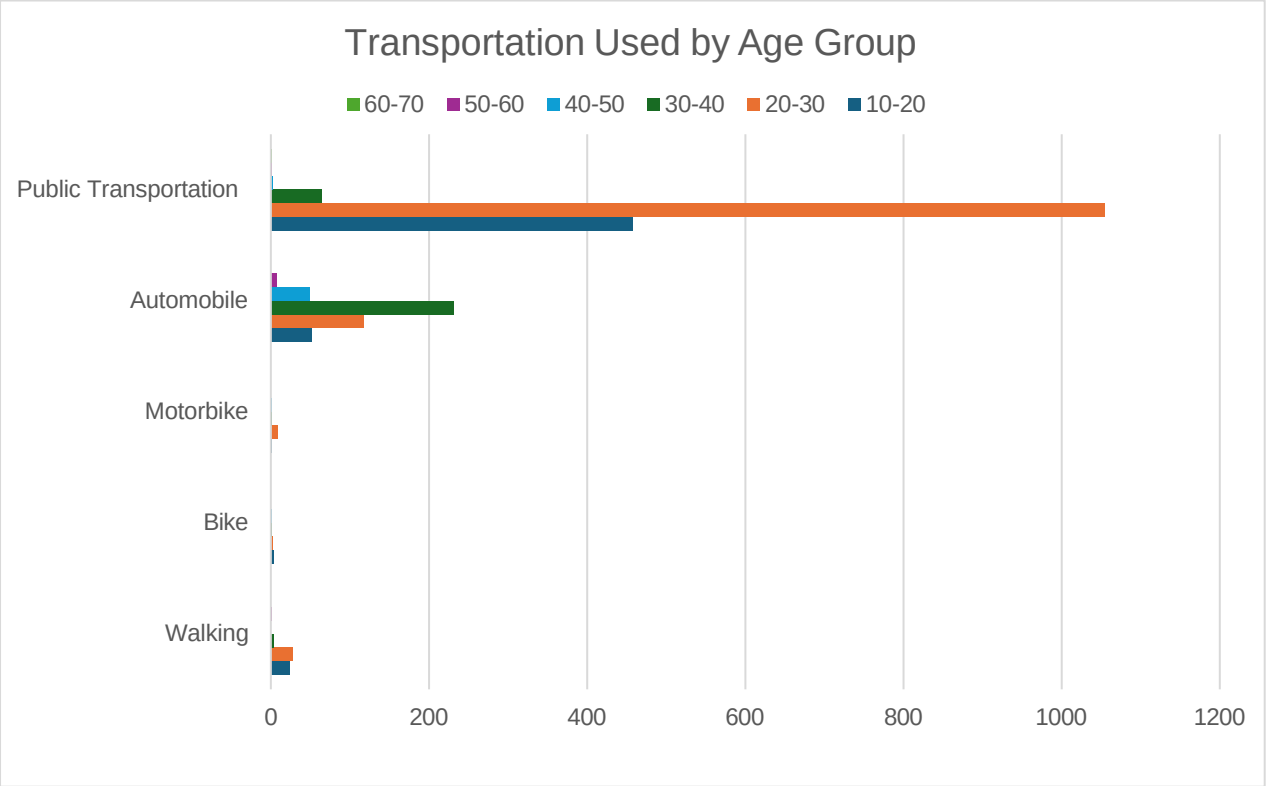


Individuals who engage in regular physical activity are more likely to have a normal BMI. There is no difference in BMI between stationary individuals and individuals who are active 1 to 2 times a week.

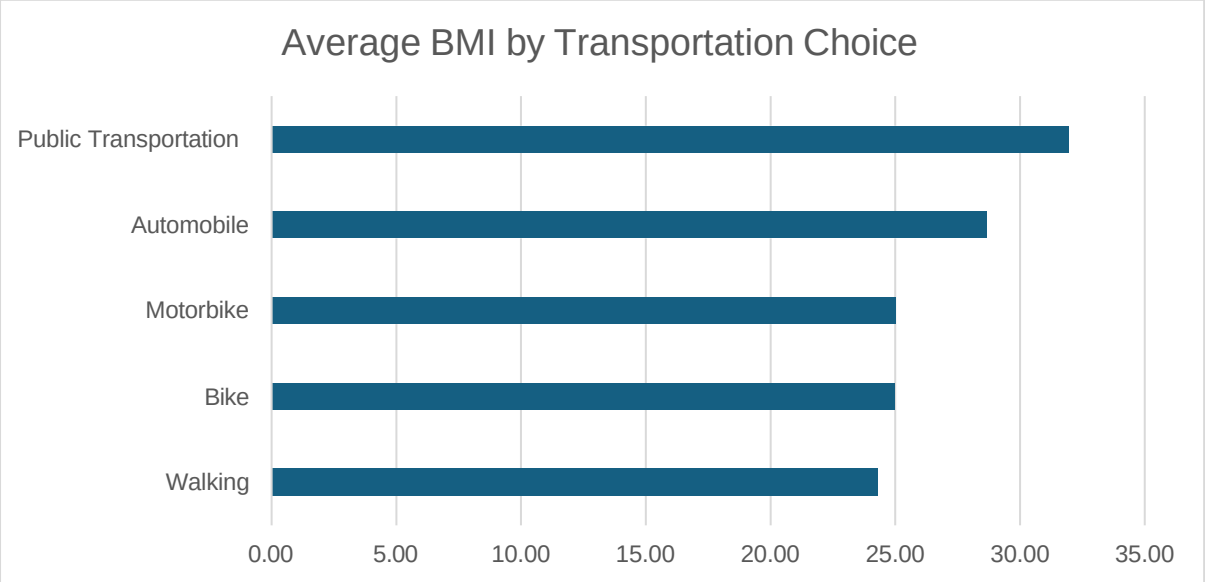


There is a slight positive correlation between time spent on devices and average BMI.

Transportation Analysis



75% of individuals primarily use public transportation, which may impact their overall physical activity levels.



Individuals who use active transportation on average have a BMI of 24.5. Individuals who utilise automobiles have an average BMI of 28.66. finally, individuals who use public transportation have the highest average BMI of 31.94. The only passive transportation that has a normal BMI is motorbikes.

Conclusion

Demographically, the dataset represents a balanced gender distribution, with a majority of individuals below the age of 30. Females had a higher average BMI compared to males, indicating potential gender-based disparities in the dataset.

When investigating weight and BMI, the analysis reveals that a substantial number of individuals fall within the overweight and obese categories, with 46% of the individuals classified as obese. Moreover, individuals with a family history of overweight or obesity display a significantly higher likelihood of being overweight themselves.

Dietary habits emerge as a focal point of concern, with a staggering 88% of individuals reporting frequent consumption of high-caloric foods. This dietary pattern likely contributes to weight gain, as supported by the positive correlation between high-caloric food intake and BMI. Conversely, there's insufficient evidence to establish a significant relationship between vegetable consumption frequency and weight status, requiring further investigation into the types and quantities of vegetables consumed.

Individuals engaging in regular physical activity tend to display lower BMI, highlighting the importance of active living. Individuals showing sedentary behaviours, especially those who spend too much time on technology devices are more like to be overweight.

When analysing the role of alcohol consumption and smoking habits on BMI. While no substantial correlation is observed between alcohol intake and BMI, smoking appears to have a slight influence, possibly through appetite suppression. However, individuals with a family history of obesity are notably prone to higher BMI, irrespective of their alcohol and smoking behaviours.

Transportation choices, play a key role in affecting health outcomes. The reliance on public transportation, coupled with a lack of active transportation options, may contribute to higher BMI among individuals. Those utilising active modes of transportation, such as walking or cycling, demonstrate healthier weight compared to individuals relying on automobiles or public transit.

In summary, the exploratory analysis highlights the genetic predispositions, dietary habits, lifestyle choices, and environmental factors in shaping a healthy lifestyle. These insights help guide targeted interventions and policy initiatives aimed at fostering healthier lifestyles and reducing the increasing public health crisis of obesity and related illnesses among young populations.

Recommendations

Based on the findings from the exploratory data analysis, I would recommend:

- Implement nutritional education programs aimed at increasing awareness of healthy dietary habits and the harmful effects of consuming high-caloric foods.
- Develop initiatives to promote regular physical activity among young people, highlighting the importance of incorporating exercise into daily routines.
- Create opportunities for physical activity, such as community sports programs, recreational facilities, and active transportation infrastructure (e.g., bike lanes, walking paths).
- Implement behavioural change strategies to address sedentary behaviours, such as reducing screen time and encouraging breaks for physical activity during prolonged periods of technology use.
- Partner with local organisations, schools, and healthcare providers to create a supportive environment for healthy living initiatives.
- Support initiatives to create safe and walkable communities, including urban planning strategies that prioritize pedestrian-friendly infrastructure and active transportation options.
- Develop targeted initiatives for high-risk groups, such as individuals with a family history of obesity or those exhibiting sedentary behaviours, to address specific risk factors and promote preventive measures.