"Predictive analysis on Obesity Levels based on Lifestyle Factors: A Machine Learning Classification Analysis"

Instructor: Sun Makosso-Kallyth

Project by: Sweta

Index

Introduction

Obesity is a growing concern worldwide and is associated with various health risks such as heart disease, diabetes, and high blood pressure. By understanding the key determinants of obesity, we can develop targeted interventions to prevent and manage this condition.

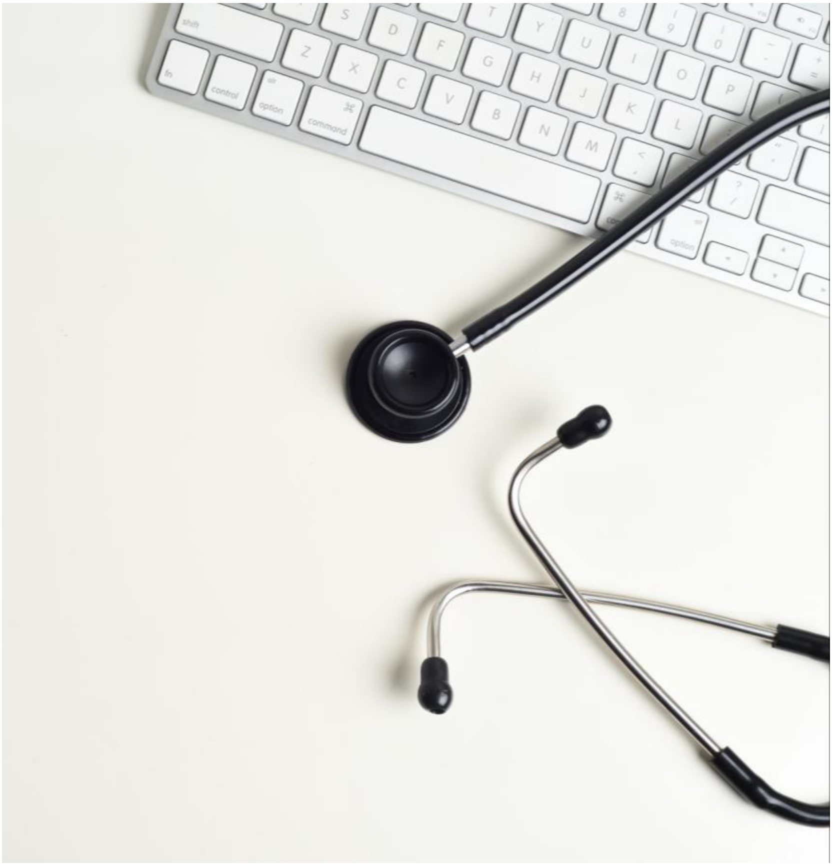
Our dataset includes information on gender, age, height, weight, family history of obesity, high caloric food consumption, vegetable intake, frequency of main meals, snacking habits, smoking habits, water intake, monitoring of calories, physical activity level, screen time, alcohol consumption, mode of transportation used, and obesity level.

Through the analysis, we hope to gain valuable insights into the factors that contribute to obesity and provide recommendations for lifestyle changes and interventions to promote healthier outcomes

Objective and Business Problem

- Predictive model to assess obesity levels based on demographic, lifestyle, and health factors.
- Targeted interventions and personalized health plans to prevent and manage obesity.
- Healthcare providers can use the model to assess obesity risk in patients and provide personalized recommendations.
- Insurance companies can identify high-risk individuals and offer preventive care programs to reduce healthcare costs.
- Public health organizations can develop targeted campaigns to promote healthy lifestyle choices and prevent obesity.
- Individuals can receive personalized recommendations to improve their health and reduce obesity risk.

Overall, stakeholders in the healthcare industry can benefit from insights into factors influencing obesity and enable targeted interventions effectively.



Questions for this analysis

- 1.Can we predict the likelihood of obesity based on variables such as family history with overweight, high caloric food consumption and physical activity level?
- 2. How does gender impact obesity levels in individuals?
- 3. Are there any relationships between eating habits (such as main meals daily, vegetable consumption, and snacking between meals) and obesity levels?
- 4. Can we classify individuals into different obesity levels based on their demographic and lifestyle factors?
- 5. How do factors like smoking, daily water intake, and use of tech devices impact obesity levels?
- 6.Can we predict the risk of obesity based on a combination of variables such as transportation mode used, monitor calories, and physical activity level?

Exploring the Dataset

Observation-2111

No	Variables	Renamed-Variables	Categories/Variable Type	Explanation Of Variables
1	Gender	Gender	Male, Female	indicating the gender of the individual.
2	Age	Age	numeric	representing the age of the individual.
3	Height	Height	numeric	indicating the height of the individual.
4	Weight	Weight	numeric	representing the weight of the individual.
5	family_history_with_overweight	Family_History	no, yes	indicating whether there is a family history of any health conditions.
6	FAVC	High_Caloric_Food	no, yes	indicating whether the individual consumes high-caloric foods.
7	FCVC	Eat_Vegetables	numeric	representing the frequency of vegetable consumption by the individual.
8	NCP	Main_Meals_Daily	numeric	indicating the number of main meals consumed daily by the individual.
9	CAEC	Eat_Between_Meals	Always, Frequently, sometimes, no	indicating whether the individual eats between meals.
10	SMOKE	SMOKE	no, yes	indicating whether the individual smokes.
11	CH2O	Water_Daily	numeric	representing the amount of water consumed daily by the individual.
12	SCC	Monitor_Calories	no, yes	indicating whether the individual monitors their calorie intake.
13	FAF	Physical_Activity	numeric	representing the level of physical activity of the individual.
14	TUE	Tech_Device_Time	numeric	indicating the amount of time spent on tech devices by the individual.
15	CALC	Drink_Alcohol	Always, Frequently, sometimes, no	indicating whether the individual drinks alcohol.
16	MTRANS	Transportation_Used	Automobile,Bike,Motorbike,Public_Transportation,Walking	indicating the mode of transportation used by the individual.
17	NObeyesdad	Obesity_Level	Insufficient_Weight, Normal_Weight, Obesity_Type_I, Obesity_Type_II, Obesity_Type_III, Overweight_Level_I, Overweight_Level_II	representing the level of obesity of the individual.

Dataset: UCI repository

Data Transformation

The FREQ Procedure

Drink_Alcohol	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Always	1	0.05	1	0.05
Frequently	70	3.32	71	3.36
Sometimes	1401	66.37	1472	69.73
no	639	30.27	2111	100.00

Eat_Between_Meals	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Always	53	2.51	53	2.51
Frequently	242	11.46	295	13.97
Sometimes	1765	83.61	2060	97.58
no	51	2.42	2111	100.00

Obesity_Level	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Insufficient_Weight	272	12.88	272	12.88
Normal_Weight	287	13.60	559	26.48
Obesity_Type_I	351	16.63	910	43.11
Obesity_Type_II	297	14.07	1207	57.18
Obesity_Type_III	324	15.35	1531	72.52
Overweight_Level_I	290	13.74	1821	86.26
Overweight_Level_II	290	13.74	2111	100.00

Transportation_Used	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Automobile	457	21.65	457	21.65
Bike	7	0.33	464	21.98
Motorbike	11	0.52	475	22.50
Public_Transportation	1580	74.85	2055	97.35
Walking	56	2.65	2111	100.00

→

The FREQ Procedure

Drink_Alcohol	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Always	71	3.36	71	3.36
Sometimes	1401	66.37	1472	69.73
no	639	30.27	2111	100.00

Eat_Between_Meals	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Always	295	13.97	295	13.97
Sometimes	1816	86.03	2111	100.00

Obesity_Level	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Normal_Weight	559	26.48	559	26.48
Obese	1552	73.52	2111	100.00

Transportation_Used	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Automobile	457	21.65	457	21.65
Motorbike	18	0.85	475	22.50
Public_Transportation	1580	74.85	2055	97.35
Walking	56	2.65	2111	100.00

Obesity_Level	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Normal_Weight	559	26.48	559	26.48
Obese	1552	73.52	2111	100.00

Reducing the categories with less frequencies to nearby related category to reduce quasi gap in data modelling.

Target Variable: Obesity level was reduced to 2 (binary) from multi-class classification to predict obesity and normal weight.

Descriptive Statistics and missing values

- There are no missing values in the dataset
- Categories are reduced of those which had less frequencies

Variable	N Miss
Age	0
Height	0
Weight	0
Eat_Vegetables	0
Main_Meals_Daily	0
Water_Daily	0
Physical_Activity	0
Tech_Device_Time	0

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Age	2111	24.3125999	6.3459683	14.0000000	61.0000000
Height	2111	1.7016774	0.0933048	1.4500000	1.9800000
Weight	2111	86.5860581	26.1911717	39.0000000	173.0000000
Eat_Vegetables	2111	2.4190431	0.5339266	1.0000000	3.0000000
Main_Meals_Daily	2111	2.6856280	0.7780386	1.0000000	4.0000000
Water_Daily	2111	2.0080114	0.6129535	1.0000000	3.0000000
Physical_Activity	2111	1.0102977	0.8505924	0	3.0000000
Tech_Device_Time	2111	0.6578659	0.6089273	0	2.0000000

The FREQ Procedure

Drink_Alcohol	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Always	71	3.36	71	3.36
Sometimes	1401	66.37	1472	69.73
no	639	30.27	2111	100.00

Eat_Between_Meals	Frequency	Percent	Cumulative Frequency	Cumulative Percen
Always	295	13.97	295	13.97
Sometimes	1816	86.03	2111	100.00

Family_History	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	385	18.24	385	18.24
yes	1726	81.76	2111	100.00

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	1043	49.41	1043	49.41
Male	1068	50.59	2111	100.00

High_Caloric_Food	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	245	11.61	245	11.61
yes	1866	88.39	2111	100.00

Monitor_Calories	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	2015	95.45	2015	95.45
yes	96	4.55	2111	100.00

Obesity_Level	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Normal_Weight	559	26.48	559	26.48
Obese	1552	73.52	2111	100.00

SMOKE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	2067	97.92	2067	97.92
yes	44	2.08	2111	100.00

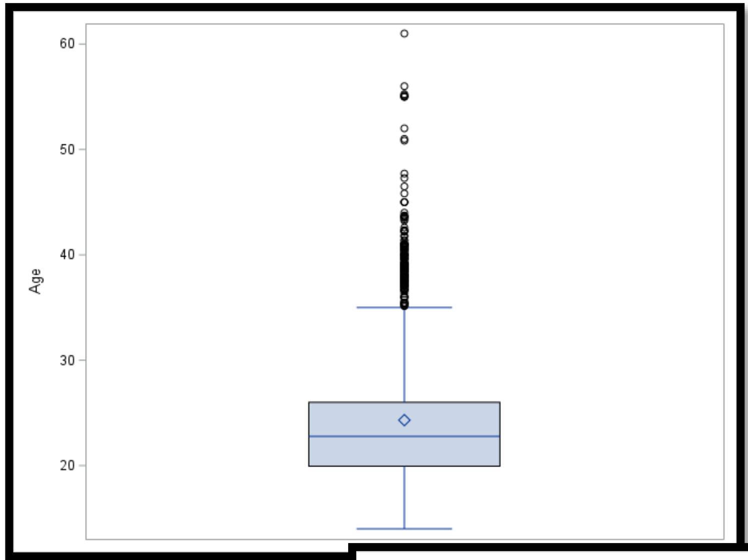
Transportation_Used	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Automobile	457	21.65	457	21.65
Motorbike	18	0.85	475	22.50
Public_Transportation	1580	74.85	2055	97.35
Walking	56	2.65	2111	100.00

The mean values and confidence intervals for the variables suggest that the sample population is relatively young, moderately active, and has a balanced diet.

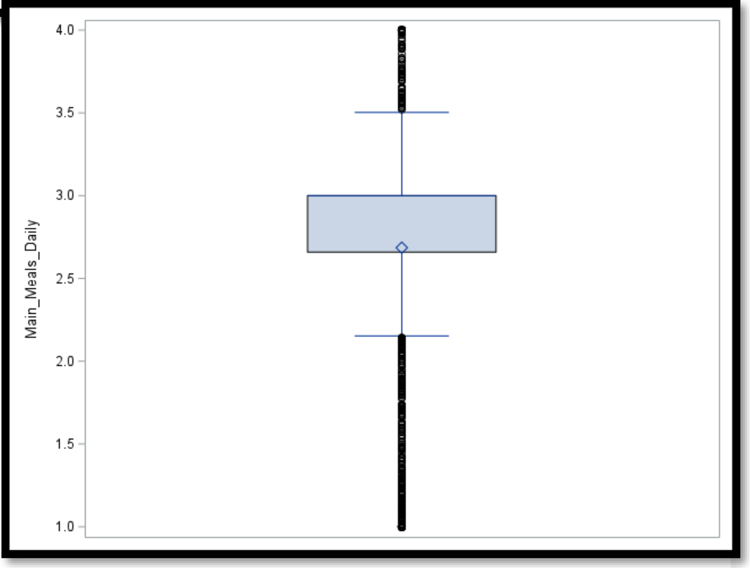
Outliers

I have capped the outliers as age had upper-level extreme outliers and main meals daily had lower-level extreme outliers

Variable	N	Mean	Std Dev	Minimum	Maximum
Age	2111	24.3125999	6.3459683	14.0000000	61.0000000
Eat_Vegetables	2111	2.4190431	0.5339266	1.0000000	3.0000000
Main_Meals_Daily	2111	2.6856280	0.7780386	1.0000000	4.0000000
Water_Daily	2111	2.0080114	0.6129535	1.0000000	3.0000000
Physical_Activity	2111	1.0102977	0.8505924	0	3.0000000
Tech_Device_Time	2111	0.6578659	0.6089273	0	2.0000000
Height	2111	1.7016774	0.0933048	1.4500000	1.9800000
Weight	2111	86.5860581	26.1911717	39.0000000	173.0000000

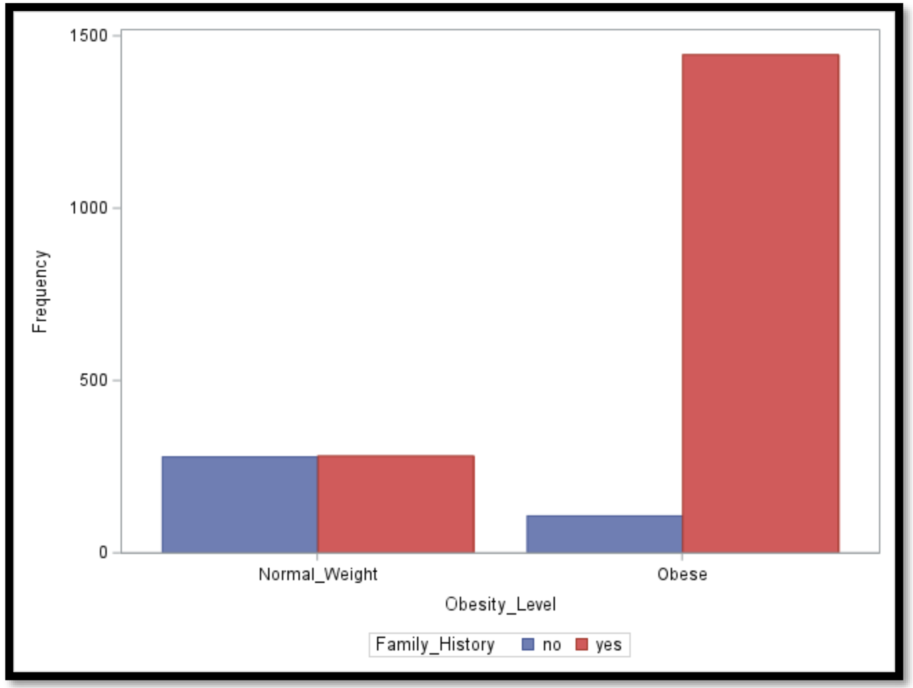


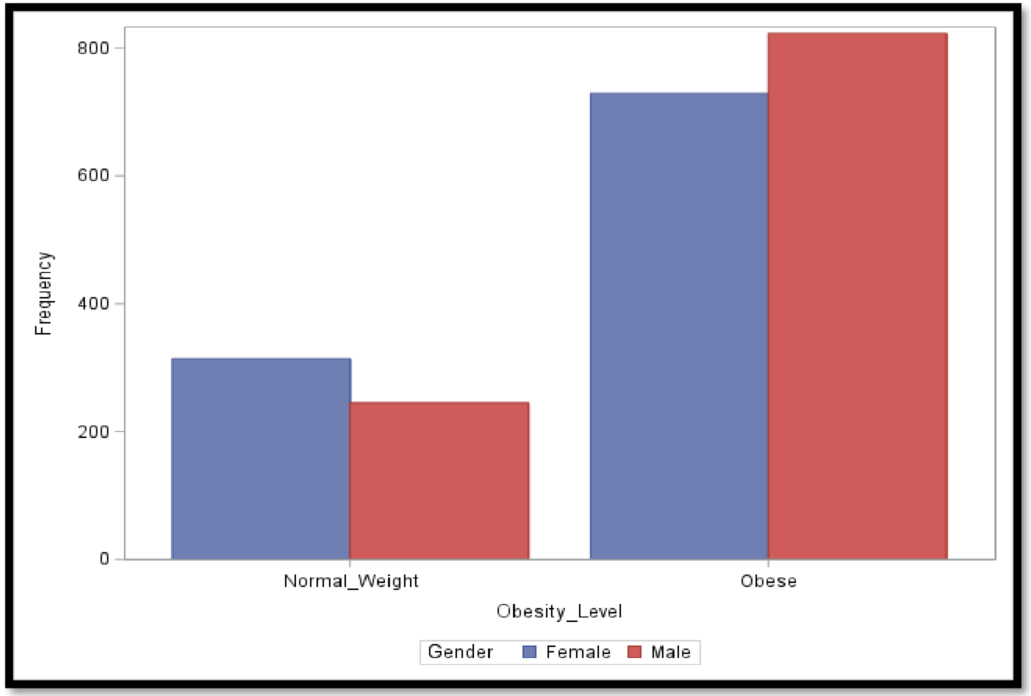
Variable	N	Mean	Std Dev	Minimum	Maximum
Age	2111	23.9103906	5.2776732	14.0000000	35.0806340
Eat_Vegetables	2111	2.4190431	0.5339266	1.0000000	3.0000000
Main_Meals_Daily	2111	2.8354847	0.4010020	2.1465975	3.5120415
Water_Daily	2111	2.0080114	0.6129535	1.0000000	3.0000000
Physical_Activity	2111	1.0102977	0.8505924	0	3.0000000
Tech_Device_Time	2111	0.6578659	0.6089273	0	2.0000000
Height	2111	1.7016756	0.0932995	1.4500000	1.9762325
Weight	2111	86.5849074	26.1874263	39.0000000	170.5707420



Exploratory Data Analysis

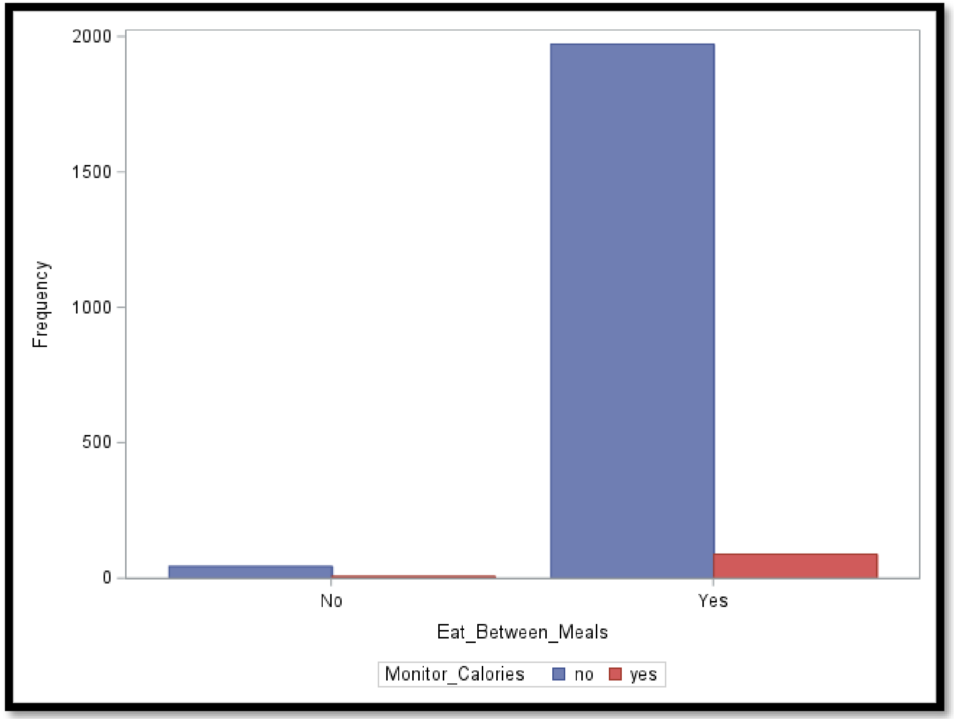
Gender and Family history of obesity

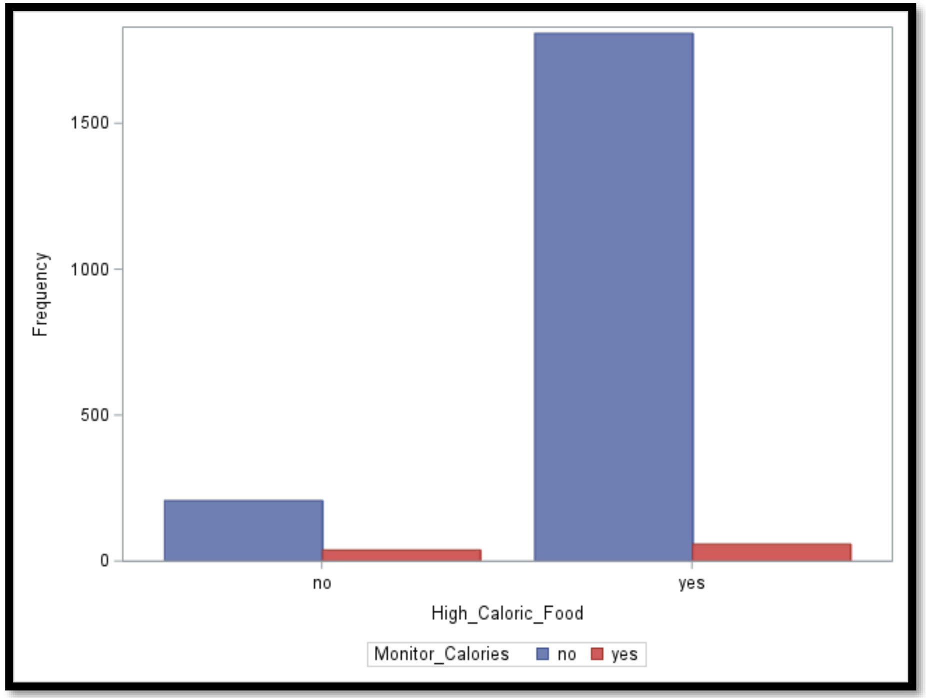




- Individuals having obesity have a family history of obesity
- Male had more obesity compared to female and female had more normal weight than male

" Eating between meals and having high caloric food? Are you monitoring your calorie intake?"





- Those eating between meals does not monitor calories
- Those having high caloric food does not monitor calories

Correlation Matrix

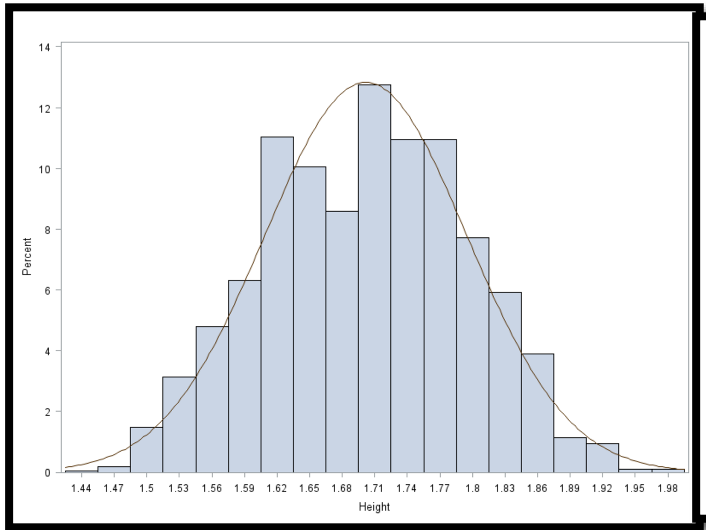
- Age has weak positive correlations with weight, daily water intake, and physical activity, but a moderate negative correlation with tech device time.
- Height has moderate positive correlations with weight, main meals daily, water intake, and physical activity.
- Weight has moderate positive correlations with main meals daily and weak positive correlations with water intake, eating vegetables, and physical activity.
- Eating vegetables has a weak positive correlation with main meals daily and a weak negative correlation with age.
- Main meals daily has weak positive correlations with water intake and physical activity.
- Water intake has weak positive correlations with physical activity and age.
- Physical activity has a weak positive correlation with tech device time.

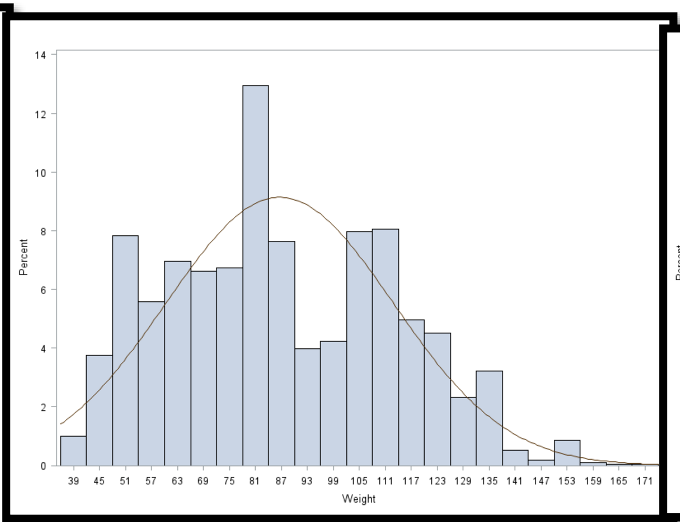
Pearson Correlation Coefficients, N = 2111 Prob > |r| under H0: Rho=0								
	Age	Height	Weight	Eat_Vegetables	Main_Meals_Daily	Water_Daily	Physical_Activity	Tech_Device_Time
Age	1	-0.00234	0.24996	0.03087	-0.09647	-0.04161	-0.15838	-0.29483
Height	-0.00234	1	0.46312	-0.03811	0.20865	0.21336	0.29473	0.0519
Weight	0.24996	0.46312	1	0.21611	0.0548	0.2006	-0.05149	-0.0716
Eat_Vegetables	0.03087	-0.03811	0.21611	1	0.04045	0.06846	0.01994	-0.10113
Main_Meals_Daily	-0.09647	0.20865	0.0548	0.04045	1	0.05864	0.12937	0.03794
Water_Daily	-0.04161	0.21336	0.2006	0.06846	0.05864	1	0.16724	0.01197
Physical_Activity	-0.15838	0.29473	-0.05149	0.01994	0.12937	0.16724	1	0.05856
Tech_Device_Time	-0.29483	0.0519	-0.0716	-0.10113	0.03794	0.01197	0.05856	1

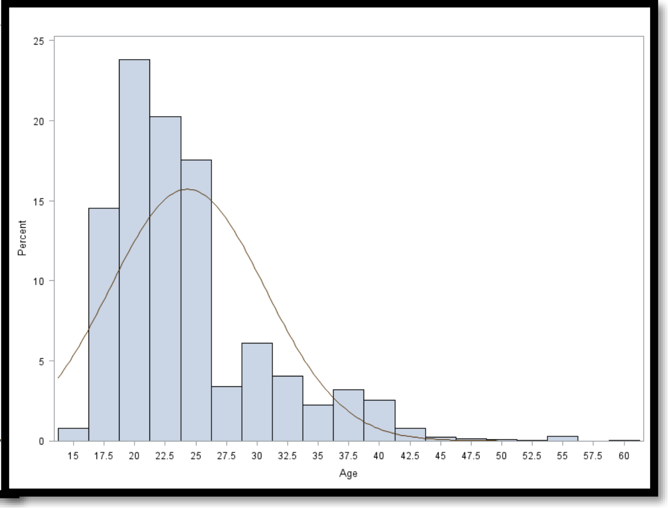
Inferential Analysis

Univariate Analysis

Histogram of height, weight and age







- Age has a mean of 24.31 years and a standard deviation of 6.35, with positive skewness (1.53) and high kurtosis (2.83).

- Weight has a mean of 86.59 units and a standard deviation of 26.19, with slightly positively skewed data (0.26) and negative kurtosis (-0.70).

- Height has a mean of 1.70 units and a standard deviation of 0.09, with nearly normally distributed data (skewness: -0.01, kurtosis: -0.56).

Bivariate analysis

Individuals having high caloric food and monitor calories or not?

(H0)Both variables are independent
(Ha)There is an association between the two variables

- There is a significant association between Monitor_Calories and High_Caloric_Food (Chi-Square = 76.7361, $p < 0.0001$).
- The Cramer's V value of -0.1907 indicates a weak negative association between these two variables.

Those who do not monitor their calories are more likely to consume high-caloric foods compared to those who do monitor their calories.

So we reject the null hypothesis

Frequency Percent Row Pct Col Pct	Table of High_Caloric_Food by Monitor_Calories			
		Monitor_Calories		
	High_Caloric_Food	no	yes	Total
	no	207 9.81 84.49 10.27	38 1.80 15.51 39.58	245 11.61
	yes	1808 85.65 96.89 89.73	58 2.75 3.11 60.42	1866 88.39
	Total	2015 95.45	96 4.55	2111 100.00

Statistics for Table of High_Caloric_Food by Monitor_Calories

Statistic	DF	Value	Prob
Chi-Square	1	76.7361	<.0001
Likelihood Ratio Chi-Square	1	52.7145	<.0001
Continuity Adj. Chi-Square	1	73.9056	<.0001
Mantel-Haenszel Chi-Square	1	76.6997	<.0001
Phi Coefficient		-0.1907	
Contingency Coefficient		0.1873	
Cramer's V		-0.1907	

Is there a significant difference in weight between individuals with and without a family history of obesity?

H0: no different in weight between individuals with or without family history
Ha: there is a significant difference in weight between individuals with or without family history

- mean weight for individuals without a family history of obesity is 59.04,
- mean weight for individuals with a family history is 92.73.

- The t-test results show a significant difference in weight between the two groups

- Therefore, we reject the null hypothesis and conclude that there is a significant difference in weight between individuals with and without a family history of obesity.

The TTEST Procedure

Variable: Weight

Family_History	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
no		385	59.0411	14.1815	0.7228	39.1018	115.0
yes		1726	92.7302	24.2322	0.5833	39.0000	173.0
Diff (1-2)	Pooled		-33.6891	22.7355	1.2814		
Diff (1-2)	Satterthwaite		-33.6891		0.9288		

Family_History	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
no		59.0411	57.6201	60.4622	14.1815	13.2456	15.2609
yes		92.7302	91.5862	93.8742	24.2322	23.4499	25.0688
Diff (1-2)	Pooled	-33.6891	-36.2021	-31.1760	22.7355	22.0696	23.4430
Diff (1-2)	Satterthwaite	-33.6891	-35.5117	-31.8664			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2109	-26.29	<.0001
Satterthwaite	Unequal	956.71	-36.27	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	1725	384	2.92	<.0001

Does the mode of transportation used (Automobile, Public Transportation, Walking) have a significant effect on an individual's weight?

H0: no relationship between the mode of transportation used and weight
Ha: There is a significant relationship between the mode of transportation used and weight.

The ANOVA test shows a significant F value of 11.92 with a p-value less than 0.0001. This indicates that there is a significant difference in weight based on the mode of transportation used.

We reject the null hypothesis and accept the alternative hypothesis that there is a significant relationship between the mode of transportation used and weight.

The ANOVA Procedure

Dependent Variable: Weight

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	16185.671	8092.836	11.92	<.0001
Error	2108	1431226.806	678.950		
Corrected Total	2110	1447412.477			

R-Square	Coeff Var	Root MSE	Weight Mean
0.011182	30.09338	26.05667	86.58606

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Transportation_Used	2	16185.67100	8092.83550	11.92	<.0001

Does eat between meals affect obesity level?

(H0) that Obesity_Level and Eat_Between_Meals are independent.
(Ha)There is an association between the two variables

- There is no significant association between Obesity_Level and Eat_Between_Meals (Chi-Square = 0.0263, p = 0.8711).
- The Cramer's V value of -0.0035 indicates a very weak negative association between these two variables.
- The Fisher's Exact Test does not show a significant p-value, suggesting that there is no relationship between these variables
- Both normal weight individuals and obese individuals are equally likely to consume between meals.
- We accept null that they are independent

Frequency
Percent
Row Pct
Col Pct

Table of Eat_Between_Meals by Obesity_Level

	Obesity_Level		
Eat_Between_Meals	Normal_Weight	Obese	Total
No	13 0.62 25.49 2.33	38 1.80 74.51 2.45	51 2.42
Yes	546 25.86 26.50 97.67	1514 71.72 73.50 97.55	2060 97.58
Total	559 26.48	1552 73.52	2111 100.00

Statistics for Table of Eat_Between_Meals by Obesity_Level

Statistic	DF	Value	Prob
Chi-Square	1	0.0263	0.8711
Likelihood Ratio Chi-Square	1	0.0265	0.8706
Continuity Adj. Chi-Square	1	0.0000	0.9987
Mantel-Haenszel Chi-Square	1	0.0263	0.8712
Phi Coefficient		-0.0035	
Contingency Coefficient		0.0035	
Cramer's V		-0.0035	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	13
Left-sided Pr <= F	0.5090
Right-sided Pr >= F	0.6181
Table Probability (P)	0.1271
Two-sided Pr <= P	1.0000

Multivariate analysis

Is there a difference in the number of main meals consumed daily between individuals with normal weight and obese individuals?

H0: There is no difference in the number of main meals consumed daily between individuals with normal weight and obese individuals.

The parameter estimate of -0.3388 suggests that for every additional main meal consumed daily, the log odds of being obese decreases by 0.3388.

This result was statistically significant ($p < 0.0001$), indicating that a higher frequency of main meals may be associated with a lower likelihood of obesity.

So we reject the null.

The GENMOD Procedure

Model Information	
Data Set	WORK.HEALTH
Distribution	Binomial
Link Function	Logit
Dependent Variable	Obesity_Level

Number of Observations Read	2111
Number of Observations Used	2111
Number of Events	1552
Number of Trials	2111

Class Level Information

Class	Levels	Values
Obesity_Level	2	Obese Normal_Weight

Response Profile

Ordered Value	Obesity_Level	Total Frequency
1	Obese	1552
2	Normal_Weight	559

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Log Likelihood		-1207.4922	
Full Log Likelihood		-1207.4922	
AIC (smaller is better)		2418.9844	
AICC (smaller is better)		2418.9901	
BIC (smaller is better)		2430.2942	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.9476	0.2001	1.5554	2.3399	94.71	<.0001
Main_Meals_Daily	1	-0.3388	0.0696	-0.4753	-0.2023	23.66	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

How does eating vegetables affect the obesity level?

H0: There is no relationship between vegetable consumption and obesity.
Ha: There is a significant relationship between vegetable consumption and obesity.

The parameter estimate of 0.0636 indicates that for every unit increase in vegetable consumption, the log odds of being obese increases by 0.0636. However, this variable was not found to be statistically significant (p = 0.4900), suggesting that vegetable consumption may not be significantly related to obesity in this model.

This means that there is not enough evidence to suggest that vegetable consumption is significantly related to obesity in this model. So we accept null

The GENMOD Procedure

Model Information	
Data Set	WORK.HEALTH
Distribution	Binomial
Link Function	Logit
Dependent Variable	Obesity_Level

Number of Observations Read	2111
Number of Observations Used	2111
Number of Events	1552
Number of Trials	2111

Class Level Information		
Class	Levels	Values
Obesity_Level	2	Obese Normal_Weight

Response Profile		
Ordered Value	Obesity_Level	Total Frequency
1	Obese	1552
2	Normal_Weight	559

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Log Likelihood		-1219.9654	
Full Log Likelihood		-1219.9654	
AIC (smaller is better)		2443.9309	
AICC (smaller is better)		2443.9365	
BIC (smaller is better)		2455.2407	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	0.8676	0.2275	0.4218	1.3134	14.55	0.0001
Eat_Vegetables	1	0.0636	0.0921	-0.1170	0.2441	0.48	0.4900
Scale	0	1.0000	0.0000	1.0000	1.0000		

How does drinking adequate daily water affect obesity level?

H0: there is no association between daily water consumption and obesity
Ha: there is a significant association between daily water consumption and obesity

- The parameter estimate of 0.5400 suggests that for every additional unit of daily water consumption, the log odds of being obese increases by 0.5400.
- This result was statistically significant ($p < 0.0001$), indicating that higher water intake may be associated with a higher likelihood of obesity.
- So null hypothesis can be rejected, as there is a statistically significant association between the two variables.

The GENMOD Procedure

Model Information	
Data Set	WORK.HEALTH
Distribution	Binomial
Link Function	Logit
Dependent Variable	Obesity_Level

Number of Observations Read	2111
Number of Observations Used	2111
Number of Events	1552
Number of Trials	2111

Class Level Information		
Class	Levels	Values
Obesity_Level	2	Obese Normal_Weight

Response Profile		
Ordered Value	Obesity_Level	Total Frequency
1	Obese	1552
2	Normal_Weight	559

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Log Likelihood		-1198.0754	
Full Log Likelihood		-1198.0754	
AIC (smaller is better)		2400.1507	
AICC (smaller is better)		2400.1564	
BIC (smaller is better)		2411.4606	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.0378	0.1649	-0.3609	0.2854	0.05	0.8189
Water_Daily	1	0.5400	0.0821	0.3791	0.7008	43.30	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Is there a relation with amount of time spent on tech device and weight?

H0: There is no relationship between the amount of time spent on a tech device and an individual's weight
Ha: There is a significant relationship between these two variables.

The negative parameter estimate of -3.08 indicated that as the amount of time spent on a tech device increases, the weight of an individual tends to decrease.

There is a statistically significant negative relationship between time spent on tech devices and weight. This suggests that individuals who spend more time on tech devices are likely to have lower weight.

Dependent Variable: Weight

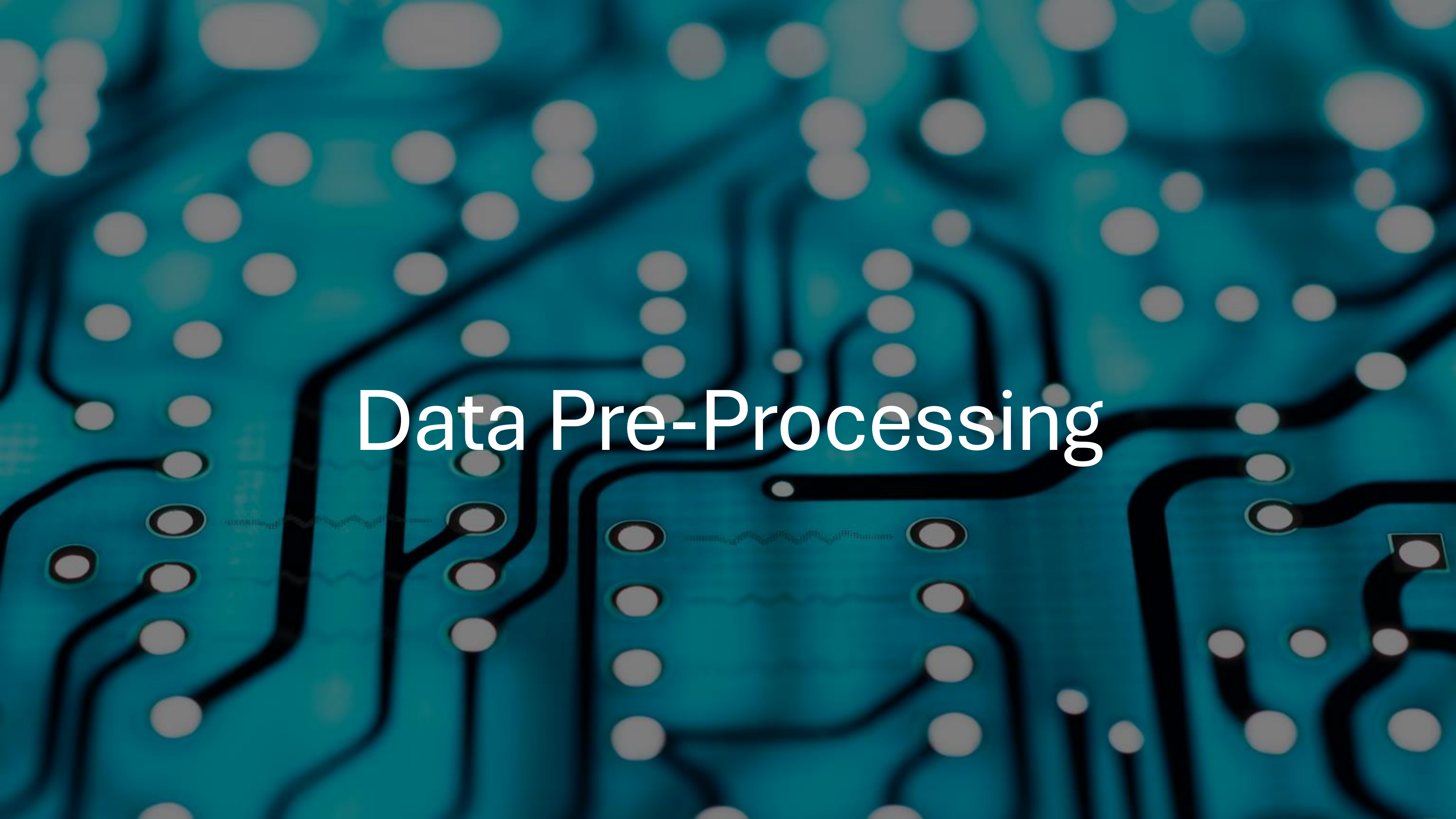
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7412.240	7412.240	10.86	0.0010
Error	2109	1440000.237	682.788		
Corrected Total	2110	1447412.477			

R-Square	Coeff Var	Root MSE	Weight Mean
0.005121	30.17832	26.13022	86.58606

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Tech_Device_Time	1	7412.239963	7412.239963	10.86	0.0010

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Tech_Device_Time	1	7412.239963	7412.239963	10.86	0.0010

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	88.61096700	0.83734307	105.82	<.0001
Tech_Device_Time	-3.07799628	0.93419244	-3.29	0.0010



Data Pre-Processing

Normalization and handling imbalanced data

The FREQ Procedure

Obesity_Level	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Normal_Weight	559	26.48	559	26.48
Obese	1552	73.52	2111	100.00

The SAS System

The FREQ Procedure

Obesity_Level	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Normal_Weight	3354	30.18	3354	30.18
Obese	7760	69.82	11114	100.00

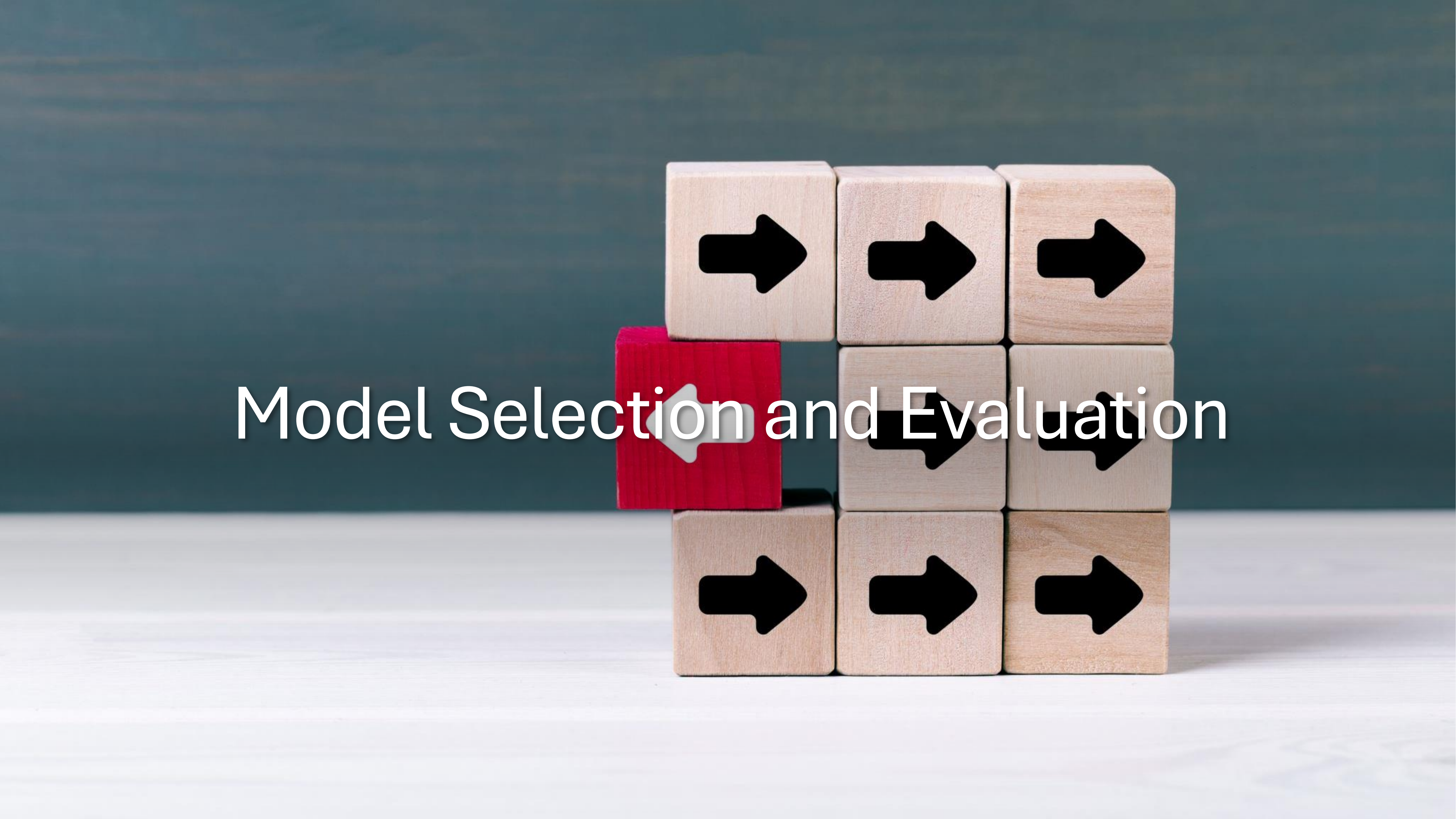
```
data balanced_obesity;
set work.health;
if Obesity_Level = 'Normal_Weight' then output;
do i=1 to 5;
output;
end;
run;
```

Variable	Mean	Std Dev
Age	-0.00	1.00
Eat_Vegetables	-0.00	1.00
Main_Meals_Daily	-0.00	1.00
Water_Daily	-0.00	1.00
Physical_Activity	-0.00	1.00
Tech_Device_Time	-0.00	1.00
Height	0.00	1.00
Weight	0.00	1.00

```
proc standard data=work.health out=work.obesity mean=0 std=1;
var Age Eat_Vegetables Main_Meals_Daily water_daily Physical_Activity  Tech_Device_Time Height weight;
run;
proc means data=work.obesity mean stddev ndec=2;
var Age Eat_Vegetables Main_Meals_Daily water_daily Physical_Activity  Tech_Device_Time height weight;
run;
```

Normalized the numeric data

After balancing the data, there were 7760 individuals classified as obese, representing 69.82% of the sample. The number of individuals classified as having a normal weight increased to 3354, accounting for 30.18% of the sample.



Model Selection and Evaluation

Logistic regression

```
/*train-test split*/
data train test;
set WORK.BALANCED_OBESITY;
if ranuni(0) < 0.8 then output train;
else output test;
run;
```

```
/*model training logistic regression*/
proc logistic data=train descending;
   class Gender Family_History High_Caloric_Food Eat_Between_Meals smoke(ref='yes')
 Monitor_Calories Drink_Alcohol Transportation_Used Obesity_Level;
  model Obesity_Level = Gender Family_History High_Caloric_Food
Eat_Between_Meals smoke Monitor_Calories Drink_Alcohol
   Transportation_Used  weight Age Eat_Vegetables Main_Meals_Daily water_daily
Physical_Activity Tech_Device_Time/aggregate=none
 link=logit selection=backward;
score data=train out=predicted_train;
run;
/*predictions on train set*/
data predicted_train;
 set predicted_train;
if  P_normal<P_obese then predicted_obesity_level = "Obese";
else predicted_obesity_level = "Normal_Weight";
run;
/* Calculate accuracy fopr training set*/
data scored_train;
set predicted_train;
true_obesity_level = obesity_level;
predict_obesity_level = predicted_obesity_level;
run;
data scored_train;
  set scored_train;
  if true_obesity_level = predict_obesity_level then correct_prediction = 1;
  else correct_prediction = 0;
run;

proc means data=scored_train mean;
  var correct_prediction;
run;
/*plor ROC curve on training set*/
ods graphics on;
proc logistic data=predicted_train plots(only)=(roc);
model obesity_level =  Age Eat_Vegetables Main_Meals_Daily water_daily
Physical_Activity  Tech_Device_Time weight;
run;
ods graphics off;
```

Results for Training set

- Logistic regression model used to predict likelihood of being in "Obese" category

- Backward elimination procedure used to remove non-significant predictor variables

- Final model included significant predictors such as Gender, Family History, Eat Between Meals, Smoking status, Monitoring Calories, etc.

- Odds ratio estimates provided information on impact of predictors on likelihood of being Obese

- Model showed good predictive performance with high percent concordant (99.1%) and high Somers' D statistic (0.981)

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-21.5818	0.9832	481.8718	<.0001
Gender	Female	1	1.8192	0.0946	369.8542	<.0001
Family_History	no	1	-0.1347	0.0681	3.9153	0.0478
Eat_Between_Meals	No	1	0.9402	0.1272	54.6168	<.0001
SMOKE	no	1	0.9220	0.1881	24.0226	<.0001
Monitor_Calories	no	1	-0.4343	0.0927	21.9341	<.0001
Drink_Alcohol	no	1	-0.2466	0.0664	13.7972	0.0002
Transportation_Used	Automobile	1	0.3039	0.1398	4.7240	0.0297
Transportation_Used	Public_Transportation	1	1.0254	0.1189	74.3471	<.0001
Weight		1	0.3453	0.0118	861.6731	<.0001
Age		1	0.0817	0.0150	29.4874	<.0001
Eat_Vegetables		1	-0.6592	0.1190	30.6937	<.0001
Main_Meals_Daily		1	-1.0615	0.1529	48.2075	<.0001
Water_Daily		1	0.2558	0.1070	5.7128	0.0168
Physical_Activity		1	-0.6386	0.0721	78.3493	<.0001
Tech_Device_Time		1	0.3827	0.0934	16.7771	<.0001

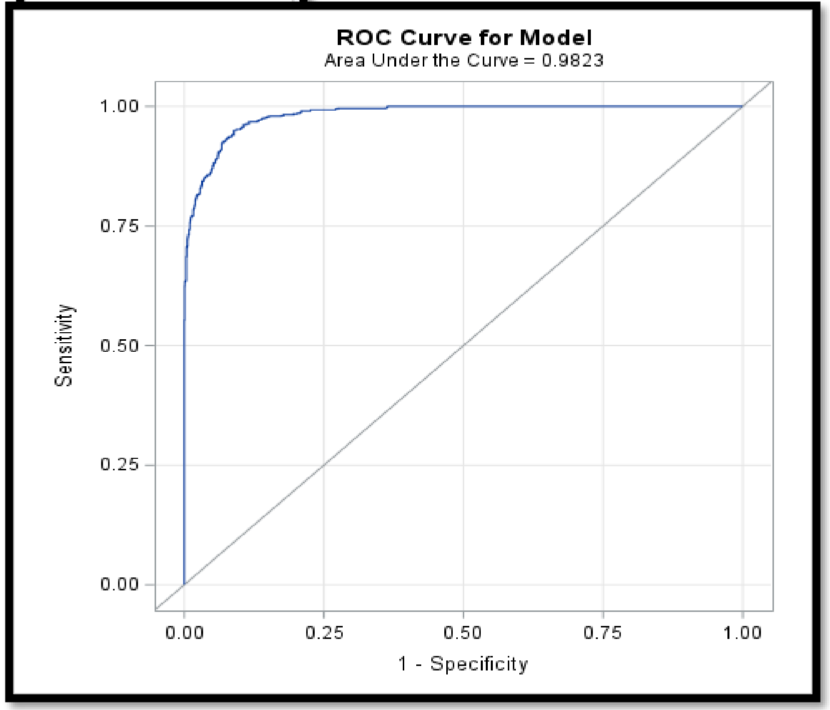
Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Gender Female vs Male	38.028	26.246	55.098
Family_History no vs yes	0.764	0.585	0.997
Eat_Between_Meals No vs Yes	6.556	3.982	10.796
SMOKE no vs yes	6.322	3.024	13.215
Monitor_Calories no vs yes	0.420	0.292	0.603
Drink_Alcohol no vs yes	0.611	0.471	0.792
Transportation_Used Automobile vs Walking	5.120	2.831	9.260
Transportation_Used Public_Transportation vs Walking	10.535	6.156	18.030
Weight	1.412	1.380	1.445
Age	1.085	1.054	1.118
Eat_Vegetables	0.517	0.410	0.653
Main_Meals_Daily	0.346	0.256	0.467
Water_Daily	1.291	1.047	1.593
Physical_Activity	0.528	0.458	0.608
Tech_Device_Time	1.466	1.221	1.761

ROC Curve and Accuracy for train set

- The model has an accuracy of 69.98%
- The AUC of the ROC curve is 0.9823, indicating very good predictive power
- A higher AUC value closer to 1 suggests a better-performing model
- The AUC of 0.9823 suggests high discriminatory power in predicting obesity levels
- The model is able to accurately predict obesity levels based on input variables

Analysis Variable : correct_prediction
Mean
0.6998426



Results on test set

- Logistic regression analysis on testing data to predict likelihood of being in "Obese" category
- Backward elimination procedure used to identify significant predictors of obesity
- Significant predictors in final model:
- Gender, Eat_Between_Meals, Monitor_Calories, Transportation_Used, Weight, Age, Eat_Vegetables, Main_Meals_Daily, Physical_Activity, Tech_Device_Time
- Model had good predictive ability with high percent concordant and high Somers' D value
- Age, Eat_Vegetables, Main_Meals_Daily, Physical_Activity, and Tech_Device_Time were significant predictors of being in "Normal_Weight" category
- Findings provide insights for potential interventions and strategies to reduce obesity risk

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-23.5630	2.0479	132.3921	<.0001
Gender	Female	1	1.7886	0.1823	96.2945	<.0001
Eat_Between_Meals	No	1	1.1342	0.2519	20.2736	<.0001
Monitor_Calories	no	1	-0.8780	0.2056	18.2361	<.0001
Transportation_Used	Automobile	1	0.0982	0.3195	0.0944	0.7586
Transportation_Used	Public_Transportation	1	1.3545	0.2626	26.6032	<.0001
Weight		1	0.3708	0.0246	226.7498	<.0001
Age		1	0.1196	0.0334	12.8554	0.0003
Eat_Vegetables		1	-0.7309	0.2181	11.2277	0.0008
Main_Meals_Daily		1	-0.6194	0.2884	4.6129	0.0317
Physical_Activity		1	-0.5779	0.1440	16.0964	<.0001
Tech_Device_Time		1	0.3969	0.1852	4.5955	0.0321

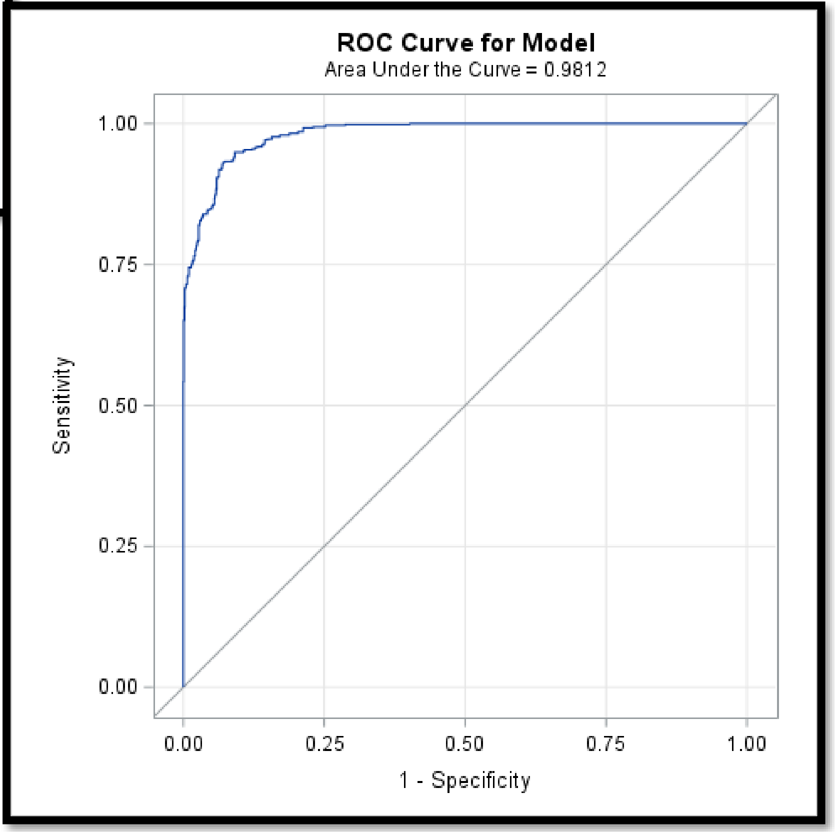
Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Gender Female vs Male	35.771	17.508	73.083
Eat_Between_Meals No vs Yes	9.664	3.600	25.942
Monitor_Calories no vs yes	0.173	0.077	0.387
Transportation_Used Automobile vs Walking	4.716	1.267	17.551
Transportation_Used Public_Transportation vs Walking	16.563	5.193	52.832
Weight	1.449	1.381	1.521
Age	1.127	1.056	1.203
Eat_Vegetables	0.481	0.314	0.738
Main_Meals_Daily	0.538	0.306	0.947
Physical_Activity	0.561	0.423	0.744
Tech_Device_Time	1.487	1.035	2.138

ROC Curve and Accuracy for train set

The accuracy of the model is 0.6917, which means that the model correctly predicts the Obesity level (Normal Weight or Obese) around 69.17% of the time.

Analysis Variable : correct_prediction
Mean
0.6917192



The AUC value of 0.9812 indicates that the model has a high discriminatory power, as it is able to distinguish between the two classes (Normal Weight and Obese) very well.

Conclusion

- Female gender was a strong predictor of obesity levels
- Older age was associated with a higher likelihood of obesity
- Higher weight was a significant predictor of obesity
- Eating fewer vegetables and having fewer main meals daily increased the risk of obesity
- Less physical activity and more time spent on tech devices were linked to obesity
- Not monitoring calories and using public transportation were also risk factors for obesity
- Drinking alcohol was also found to be a significant predictor of obesity levels.

Recommendation

- PROMOTE HEALTHY EATING HABITS
- INCREASE PHYSICAL ACTIVITY
- MONITOR WEIGHT AND CALORIE INTAKE
- PROMOTE ACTIVE TRANSPORTATION
- RAISE AWARENESS THROUGH PUBLIC HEALTH CAMPAIGNS AND EDUCATION PROGRAMS
- IMPLEMENT PERSONALIZED INTERVENTIONS BASED ON INDIVIDUAL LIFESTYLE FACTORS
- TARGET INTERVENTIONS TOWARDS INDIVIDUALS AT RISK OF OBESITY TO PREVENT AND TREAT THE CONDITION

Dataset:
https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition