Obesity Dataset Analysis

Unveiling Obesity Trends in Latin America

- Backdrop and Context:
- The World Health Organization classifies a BMI ≥ 25 as overweight and a BMI ≥ 30 as obesity in adults. It is projected that by 2030, the proportion of individuals classified as obese or overweight in Latin America and the Caribbean will rise to 81.9%.

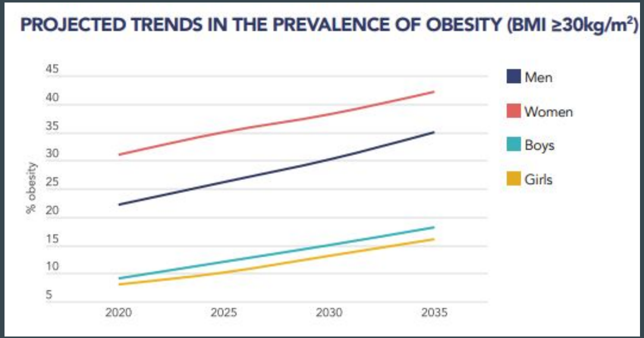
- Research about obesity in Latin America:
- Kain et al (2003) state that several factors such as poor nutrition, socioeconomic causes, and a sedentary lifestyle have contributed to the rise in obesity levels in the region.
- A 2019 study by Jiwani et al. found that obesity prevalence is on the rise in the region, with pronounced increases in rural areas and disadvantaged groups, as well as in affluent, urban populations.
- In a study of adults from 8 Latin American nations, deVicto et al (2023) explored how sedentary time and physical activity affected obesity indicators.
- The effect of genetic factors on obesity in Latin America was studied by Guevara-Ramírez et al (2022).

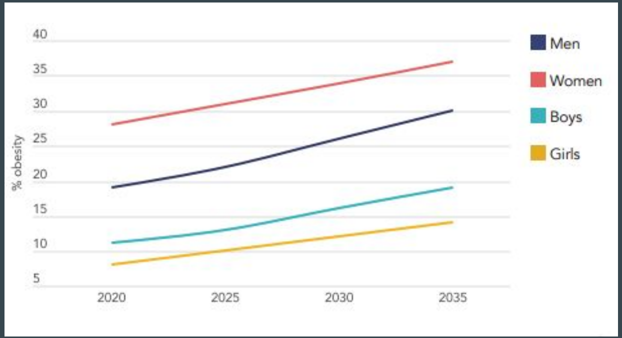
- Research Question:
- What fuels the escalating obesity rates in Latin America, and how can data analytics help discover insights about the contributing factors?

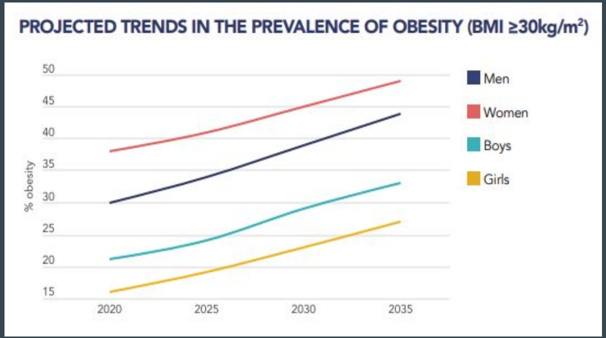
- Dataset Description:
- The dataset included individuals from Mexico, Peru, and Colombia, with diverse age groups and lifestyles, which helped us understand the interplay of various factors.

- Survey Methodology:
- The dataset authors employed a web-based survey and synthetic techniques to gather 17 attributes and 2111 records, offering a holistic view of dietary habits, physical conditions, and socio-demographics.

Prevalence of Obesity







COLOMBIA

23.9%

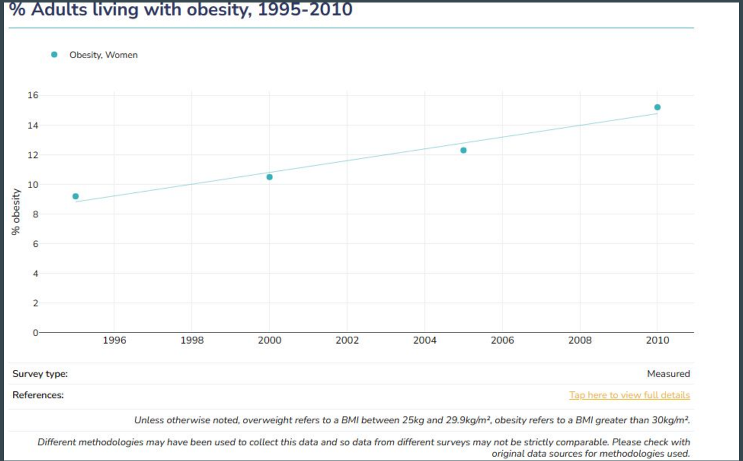
PERU

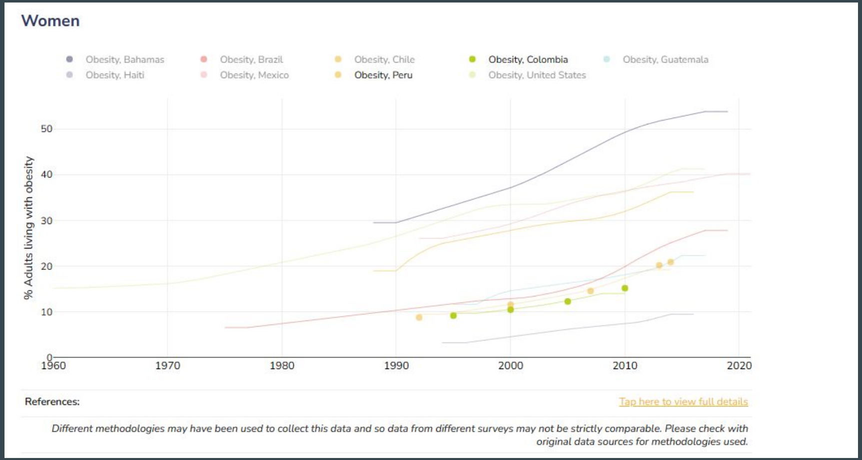
21.1%

MEXICO

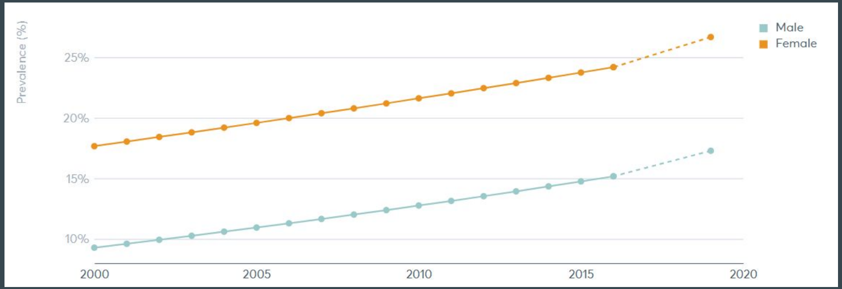
30.6%

COLOMBIA

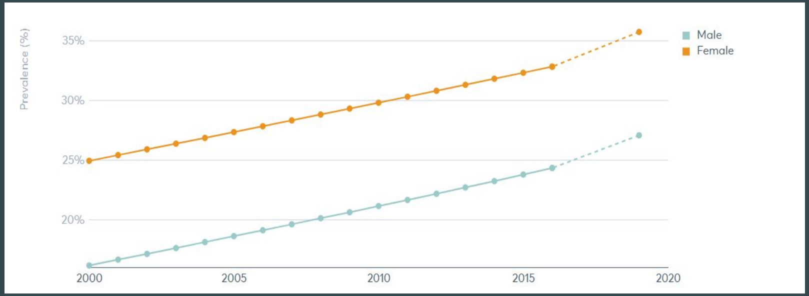




PERU



MEXICO



Variables

Continuous

Variable
Age
Height (in meters)
Weight (in kgs)

Binary

Variable	Categories
Gender	Female/Male
Family history of overweight	Yes/No
Calorie Consumption Monitoring	Yes/No
Frequent high-caloric food	Yes/No
Smoke	Yes/No

Categorical

Variable	Categories
Frequency of consumption of vegetables (FCVC)	Never/ Sometimes/ Always
Number of main meals (NCP)	Between 1 and 2/ Three/ More than three
Consumption of alcohol (CALC)	I do not drink/ Sometimes/Frequently/Always
Consumption of food between meals (CAEC)	No/ Sometimes/ Frequently /Always
Daily water consumption (CH2O)	Less than a liter/Between 1 and 2 L/More than 2 L
Time using technology devices (TUE)	0-2 hours/ 3-5 hours / > 5 hours
Transportation used	Public Transportation/Automobiles/Bike/Motorbike /Walking
Physical activity frequency (FAF)	I do not have/ 1 or 2 days/ 2 or 4 days/4 or 5 days
NObesity (NObeyesdad)	Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III

Data Preparation

Issues with the dataset

- The dataset did not have any missing data in any of the columns.
- However, there were decimal values for many of the variables, even for the categorical ones (probably because a major part of the data was generated synthetically).
- This posed a problem in analyzing the dataset.

How was the data handled?

- Therefore, to simplify the data for the ease of performing analysis, the values with decimals were rounded off to the nearest whole number.
- This was useful in obtaining well defined categories for the categorical variables.
- The new variables obtained were re-coded accordingly with a different name.

Exploratory Data Analysis

- Descriptive statistics to understand the distribution of weight of the sample population
- How many people have a positive family history of overweight?
- What are the different modes of transportation used, and what is their distribution?
- What is the distribution of individuals across different BMI categories?
- What is the mean weight of people who monitor their calorie consumption as compared to those who do not?

Descriptive Statistics

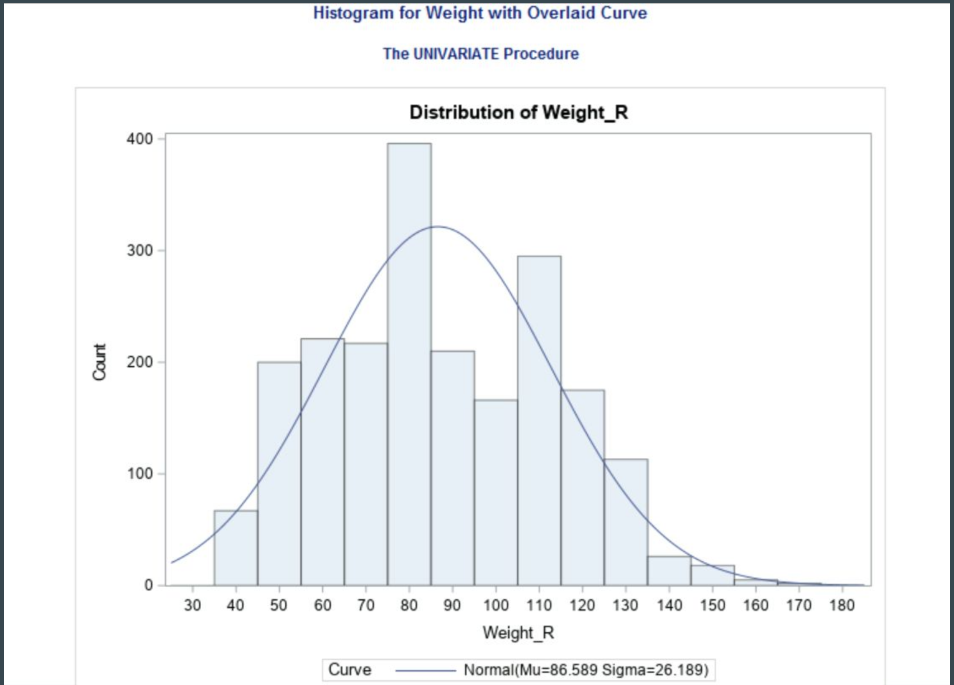
SAS Procedure Used: PROC UNIVARIATE

```
proc univariate data=obesity_data;*dataset to be used for descriptive statistics;
 title 'Descriptive Statistics';
 ods select BasicMeasures;*to create a descriptive statistics table;
 var weight_r; *descriptive statistics for specified variable;
 run;

proc univariate data=obesity_data noprint;
 title 'Histogram for Weight with Overlaid Curve';
 var weight_r;
 histogram / normal (color=blue w=4) vscale=count midpoints=(30 to 180 by 10);
 run;

proc univariate data=obesity_data normal plot; *to request normality tests and plots;
 title 'Descriptive Statistics for Weight';
 var weight_r;
 run;
```

Descriptive Statistics

The UNIVARIATE Procedure
Variable: Weight_R

Basic Statistical Measures			
Location		**Variability**	
Mean	86.58882	Std Deviation	26.18857
Median	83.00000	Variance	685.84128
Mode	80.00000	Range	134.00000
		Interquartile Range	42.00000



Mean	86.58 kgs
Median	83 kgs
Mode	80 kgs

- Since Mean > Median > Mode:
  - Weight is slightly skewed to the right.

Descriptive Statistics for Weight

The UNIVARIATE Procedure
Variable: Weight_R

Moments			
N	2111	Sum Weights	2111
Mean	86.5888205	Sum Observations	182789
Std Deviation	26.1885715	Variance	685.841278
Skewness	0.2558505	Kurtosis	-0.6996963
Uncorrected SS	17274609	Corrected SS	1447125.1
Coeff Variation	30.2447491	Std Error Mean	0.5699906

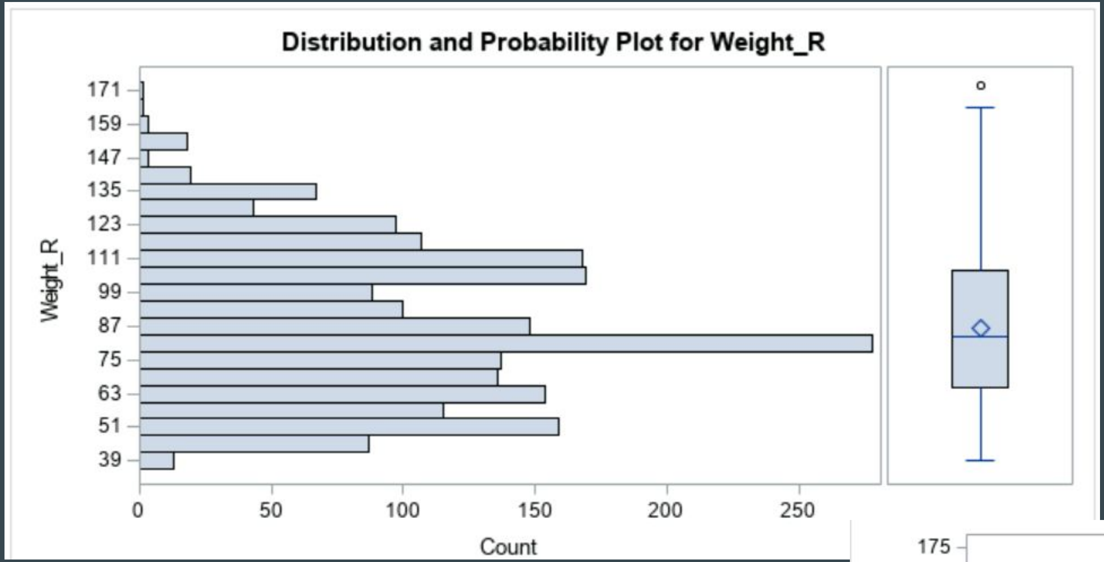
Basic Statistical Measures			
Location		Variability	
Mean	86.58882	Std Deviation	26.18857
Median	83.00000	Variance	685.84128
Mode	80.00000	Range	134.00000
		Interquartile Range	42.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	151.9127	Pr > |t|	<.0001
Sign	M	1055.5	Pr >= |M|	<.0001
Signed Rank	S	1114608	Pr >= |S|	<.0001

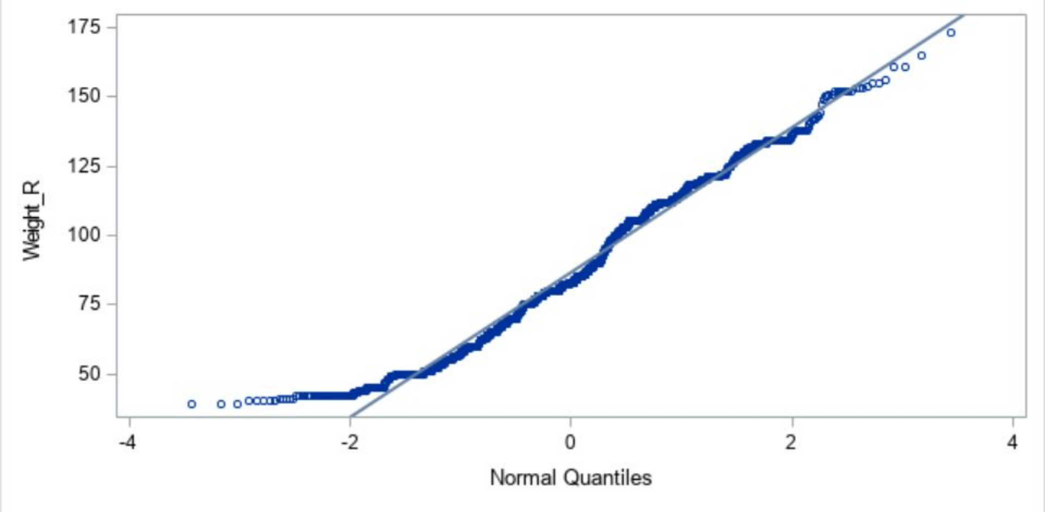
Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.065632	Pr > D	<0.0100
Cramer-von Mises	W-Sq	2.218566	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	13.60518	Pr > A-Sq	<0.0050

All 3 tests for normality have a p value **< 0.05**

Therefore, H0 is rejected and H1 is accepted i.e Weight is not distributed normally.



Distribution and Probability Plot for Weight_R
Weight_R
171
159
147
135
123
111
99
87
75
63
51
39
0
50
100
150
200
250
Count



Weight_R
175
150
125
100
75
50
-4
-2
0
2
4
Normal Quantiles

Descriptive Statistics

SAS Procedure used : PROC FREQ

```
proc freq data=obesity_data order=freq; *data is ordered in descending frequency;
  title 'Frequency Table for Family History of Obesity';
  tables family_history_with_overweight;
  run;
```

Frequency Table for Family History of Obesity

The FREQ Procedure

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

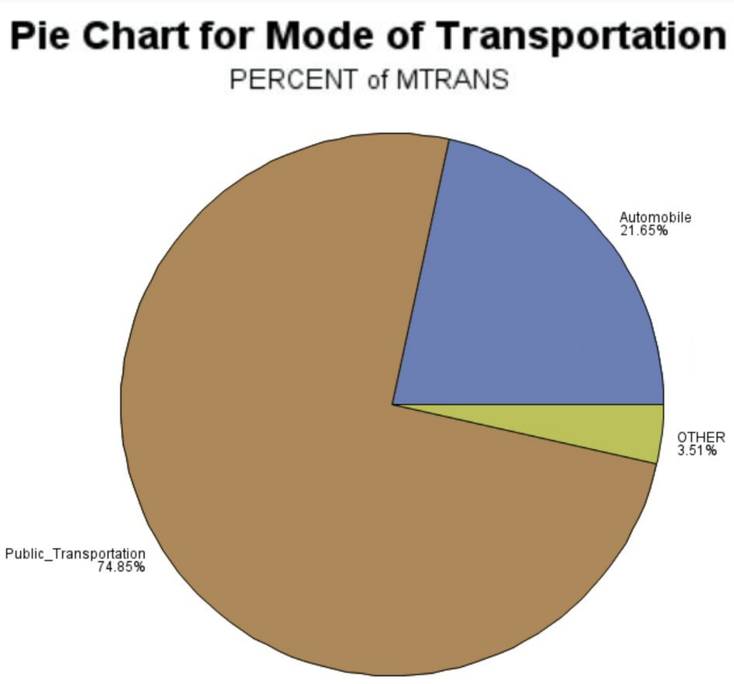
- 81.76 % (1,726 people) had a positive family history.
- 18.24 % (385 people) did not have a family history.

Descriptive Statistics

```
proc gchart data=obesity_data;
 title 'Pie Chart for Mode of Transportation';
 pie mtrans / type=percent; *display only the frequency;
 run;
```

SAS Procedure used :

PROC GCHART

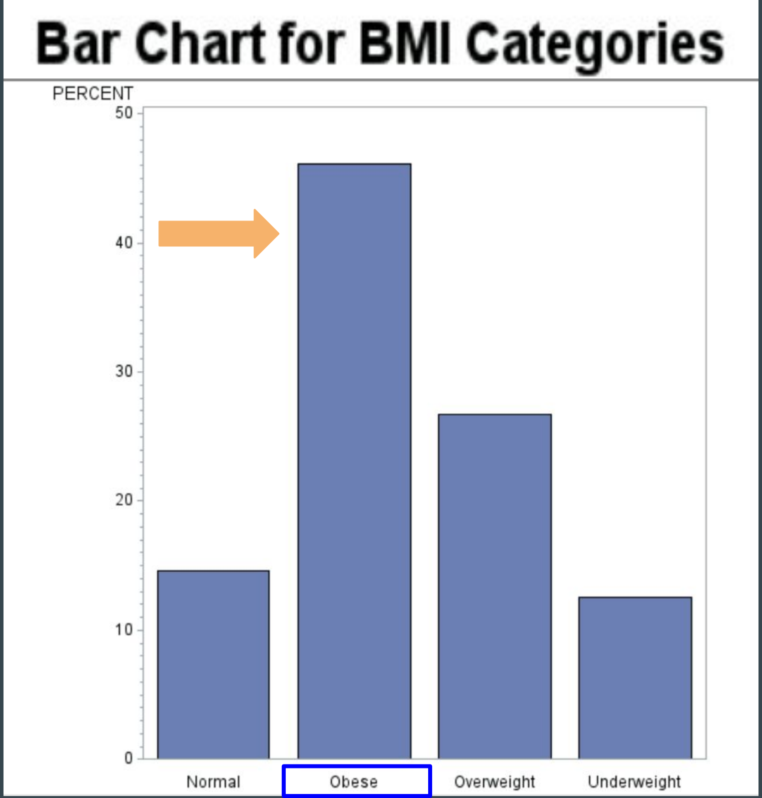


Results:

- 74.85 % people used public transportation
- 21.65 % people used automobiles
- 3.51 % people used other modes of transportation (i.e motorbikes, bikes, and walking)

Descriptive Statistics

SAS Procedure Used : PROC GCHART




```
data obesity_data;
set obesity_data;
if bmi_r < 18.5 then bmi_category = 'Underweight';
else if bmi_r >=18.5 and bmi_r < 25 then bmi_category = 'Normal';
else if bmi_r >=25 and bmi_r < 30 then bmi_category = 'Overweight';
else if bmi_r >=30 then bmi_category = 'Obese';
run;
```

```
proc gchart data=obesity_data;
title 'Bar Chart for BMI Categories';
vbar bmi_category / type=percent;
run;
```

- The majority of the people in the dataset belonged to the ‘Obese’ category (~ 45 %).
- Obese : BMI >= 30

Descriptive Statistics

SAS Procedure Used : PROC SGPLOT , PROC MEANS

```
proc means data=obesity_data order=internal;
  title 'Frequency Table of Weight by Monitoring of Calorie Consumption';
  class scc; /* scc acts as a classifier variable */
  var weight_r;
  run;
```

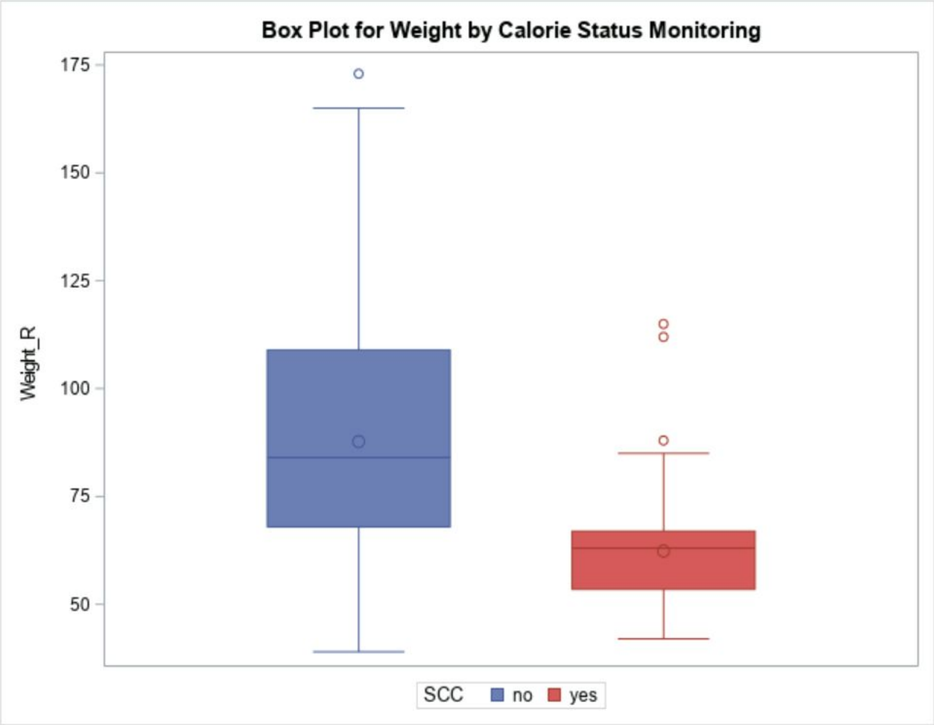
Frequency Table of Weight by Monitoring of Calorie Consumption

The MEANS Procedure

Analysis Variable : Weight_R						
SCC	N Obs	N	Mean	Std Dev	Minimum	Maximum
no	2015	2015	87.7424318	26.0695095	39.0000000	173.0000000
yes	96	96	62.3750000	14.2918821	42.0000000	115.0000000

- People who **did not monitor** their calorie consumption had a **higher mean weight (87.7 kgs)** compared to **those who did (62.3 kgs)**.

```
proc sgplot data=obesity_data;
 title 'Box Plot for Weight by Calorie Status Monitoring';
 vbox weight_r / group=scc; /*grouping weight(continuous) by calorie status monitoring(binary)*/
 run;
```



Hypothesis testing:

- Is there a potential association between consuming food between meals and the level of obesity (BMI category)?

H0 = The variables CAEC and BMI_category are independent of each other

H1 = The variables CAEC and BMI_category are dependent on each other

Chi-square test

```
/* perform a chi square test of independence for caec and bmi category */
proc freq data=obesity_data;
title 'Chi square test';
tables caec * bmi_category / chisq expected;
run;
```

Chi square test

The FREQ Procedure

Frequency
Expected
Percent
Row Pct
Col Pct

Table of CAEC by bmi_category

CAEC	bmi_category				
	Normal	Obese	Overweight	Underweight	Total
Always	35 7.7579 1.66 66.04 11.33	8 24.454 0.38 15.09 0.82	8 14.16 0.38 15.09 1.42	2 6.6281 0.09 3.77 0.76	53 2.51
Frequently	86 35.423 4.07 35.54 27.83	8 111.66 0.38 3.31 0.82	30 64.656 1.42 12.40 5.32	118 30.264 5.59 48.76 44.70	242 11.46
Sometimes	178 258.35 8.43 10.08 57.61	956 814.36 45.29 54.16 98.15	490 471.56 23.21 27.76 86.88	141 220.73 6.68 7.99 53.41	1765 83.61
no	10 7.4652 0.47 19.61 3.24	2 23.531 0.09 3.92 0.21	36 13.626 1.71 70.59 6.38	3 6.378 0.14 5.88 1.14	51 2.42
Total	309 14.64	974 46.14	564 26.72	264 12.51	2111 100.00

SAS Procedure Used : PROC FREQ chisq expected

Key Findings and Interpretation:

Statistics for Table of CAEC by bmi_category

Statistic	DF	Value	Prob
Chi-Square	9	692.2451	<.0001
Likelihood Ratio Chi-Square	9	605.6985	<.0001
Mantel-Haenszel Chi-Square	1	0.1852	0.6670
Phi Coefficient		0.5726	
Contingency Coefficient		0.4969	
Cramer's V		0.3306	

The chi-square test yielded a **highly significant** result **(p < 0.0001**), indicating a strong association between CAEC and BMI_category.

- This means that H1 is accepted :
  - CAEC and BMI_category are dependent on each other
  - There is a relationship between the consumption of food between meals and the level of obesity

Hypothesis Testing:

- Is there a potential association between frequent consumption of high calorie food and family history of overweight?

H0 = Frequent consumption of high calorie food and family history of overweight are independent of each other

H1 = Frequent consumption of high calorie food and family history of overweight are dependent on each other

Chi square test, Odds Ratio, Relative Risk

```
/*sort family history variable in descending order */
proc sort data=obesity_data; by descending family_history_with_overweight descending favc;
run;

/* perform a chi square test of independence for family history and consumption of high calorie food */
proc freq data=obesity_data order=data;
title 'Chi square test of Group Independence';
tables favc * family_history_with_overweight / chisq expected;
run;

/* calculate odds ratio and relative risk */
proc freq data=obesity_data order=data;
title 'OR and RR for Family History of Overweight and Consumption of High Calorie Food';
tables favc * family_history_with_overweight / chisq expected relrisk;
run;
```

SAS Procedure Used :
PROC FREQ , chisq, relrisk

Chi square test of Group Independence

The FREQ Procedure

Frequency
Expected
Percent
Row Pct
Col Pct

Table of FAVC by family_history_with_overweight			
	family_history_with_overweight		
FAVC	yes	no	Total
yes	1580 1525.7 74.85 84.67 91.54	286 340.32 13.55 15.33 74.29	1866 88.39
no	146 200.32 6.92 59.59 8.46	99 44.683 4.69 40.41 25.71	245 11.61
Total	1726 81.76	385 18.24	2111 100.00

Statistics for Table of FAVC by family_history_with_overweight

Statistic	DF	Value	Prob
Chi-Square	1	91.3615	<.0001
Likelihood Ratio Chi-Square	1	76.2402	<.0001
Continuity Adj. Chi-Square	1	89.6872	<.0001
Mantel-Haenszel Chi-Square	1	91.3182	<.0001
Phi Coefficient		0.2080	
Contingency Coefficient		0.2037	
Cramer's V		0.2080	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	1580
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	<.0001
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

OR and RR for Family History of Overweight and Consumption of High Calorie Food

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	3.7460	2.8183	4.9792
Relative Risk (Column 1)	1.4209	1.2794	1.5780
Relative Risk (Column 2)	0.3793	0.3150	0.4567

Sample Size = 2111

Key Findings and Interpretation:

Measure	Value	Interpretation
p-value	**< 0.0001**	Null hypothesis is rejected, which indicates a significant association between the variables.
Odds Ratio	**3.74**	People with a family history of overweight are 3.74 times more likely to consume high calorie foods frequently.
Relative Risk: Column 1	**1.42**	People with a family history of overweight have a higher chance of frequent high-calorie food consumption compared to those without a family history.
Relative Risk: Column 2	**0.37**	Individuals without a family history of overweight have a lower chance of frequent high-calorie food consumption compared to those with a family history.

Hypothesis Testing:

- Is there a significant difference in the average BMI of males and females?

➢ $\mu 1$ = The average BMI of males

➢ $\mu 2$ = The average BMI of females

Hypothesis:

- H0: $\mu 1 = \mu 2$
- H1: $\mu 1$ ~= $\mu 2$

Independent t-test (Two sample t-test)

```
/* normality tests and histogram for groups to be compared */
proc univariate data=obesity_data;
title 'BMI for Males';
var bmi_r;
histogram bmi_r / vscale = count midpoints=(10 to 70 by 5) normal;
where gender = 'male';
run;

proc univariate data=obesity_data;
title 'BMI for Females';
var bmi_r;
histogram bmi_r / vscale = count midpoints=(10 to 70 by 5) normal;
where gender = 'female';
run;

/* Perform a two sample t-test to compare the mean BMI of Males and Females */
proc ttest data=obesity_data plots=none;
title 'Two sample t-test';
class gender;
var bmi_r;
run;
```

BMI for Males

The UNIVARIATE Procedure
Fitted Normal Distribution for BMI_R

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	29.28118
Std Dev	Sigma	6.348229

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.0819148	Pr > D	<0.010
Cramer-von Mises	W-Sq	1.6621401	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	12.7194862	Pr > A-Sq	<0.005

BMI for Females

The UNIVARIATE Procedure
Fitted Normal Distribution for BMI_R

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	30.13557
Std Dev	Sigma	9.408026

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.1185437	Pr > D	<0.010
Cramer-von Mises	W-Sq	3.0663782	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	21.5018861	Pr > A-Sq	<0.005

SAS Procedures used :

- PROC UNIVARIATE
- PROC TTEST

p values for all 3 tests (for both groups) are **< 0.05.**

- Non-normal distribution

Results and Interpretation

The TTEST Procedure

Variable: BMI_R

Gender	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Female		1043	30.1356	9.4080	0.2913	13.1000	50.8000
Male		1068	29.2812	6.3482	0.1943	13.3000	49.5000
Diff (1-2)	Pooled		0.8544	8.0075	0.3486		
Diff (1-2)	Satterthwaite		0.8544		0.3501		

Gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Female		30.1356	29.5639	30.7072	9.4080	9.0209	9.8301
Male		29.2812	28.9000	29.6623	6.3482	6.0900	6.6295
Diff (1-2)	Pooled	0.8544	0.1708	1.5380	8.0075	7.7730	8.2567
Diff (1-2)	Satterthwaite	0.8544	0.1677	1.5411			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2109	2.45	0.0143
Satterthwaite	Unequal	1822.7	2.44	0.0148

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	1042	1067	2.20	<.0001

- The Levene test (equality of variances) shows $p < 0.0001$.
- Therefore, we assume unequal variances (Satterthwaite method).
- There is a statistically significant difference in the average BMI for males and females.
  - 29.28 kg/m2 for males vs 30.13 kg/m2 for females

Linear Correlation

Question: What is the correlation between BMI and Age?

H0 = Age and BMI are not correlated with each other

H1 = Age and BMI are correlated with each other

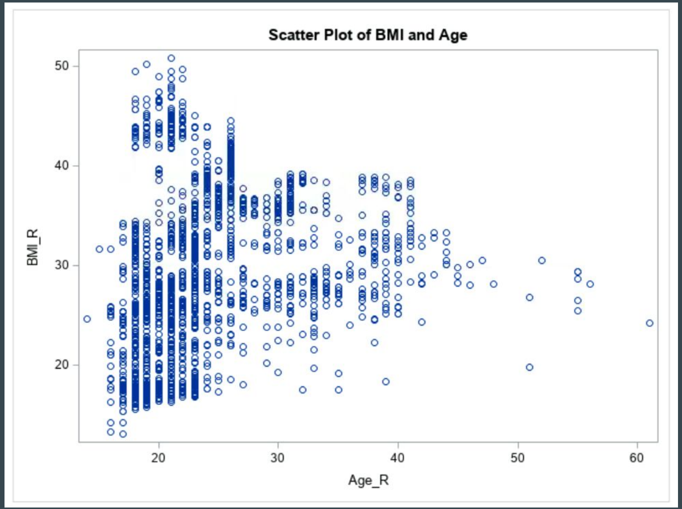
SAS Procedures used :

- PROC SGSCATTER
- PROC CORR

```
/* Scatter plot and linear correlation test for BMI_R and Age_R */

/* create a scatter plot */
proc sgscatter data=obesity_data;
plot BMI_R*Age_R;
title 'Scatter Plot of BMI and Age';
run;
```

```
/* perform linear correlation test for BMI_R and Age_R */
proc corr data=obesity_data pearson /* use Pearson correlation coefficient */;
var bmi_r age_r;
title 'Linear Correlation for BMI and Age';
run;
```

Results and Interpretation



Linear Correlation for BMI and Age

The CORR Procedure

2 Variables:	BMI_R Age_R

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
BMI_R	2111	29.70332	8.01698	62704	13.10000	50.80000
Age_R	2111	24.31596	6.35708	51331	14.00000	61.00000

Pearson Correlation Coefficients, N = 2111
Prob > |r| under H0: Rho=0

	BMI_R	Age_R
BMI_R	1.00000	0.24525 <.0001
Age_R	0.24525 <.0001	1.00000

Parameter	Value	Interpretation
Correlation Coefficient	**0.24**	Weakly positive correlation
p value	**< 0.0001**	Statistically significant

Linear Regression Analysis

- How can the relationship between BMI and height be characterized through linear regression?

SAS Procedure used: PROC REG

```
/* perform linear regression analysis for BMI and Height */
ods graphics off;
proc reg data=obesity_data;
title 'Linear Regression Analysis for BMI_R and Height_R';
model bmi_r = height_r;
run;
ods graphics on;
```

Parameter	Value	Interpretation
R-Square	**0.0169**	~ 1.69 % of variance in BMI is explained by height
p value	**< 0.0001**	Statistically significant

- The model has a **very low** explanatory power

Linear Regression Analysis for BMI_R and Height_R

The REG Procedure
Model: MODEL1
Dependent Variable: BMI_R

Number of Observations Read	2111
Number of Observations Used	2111

Analysis of Variance

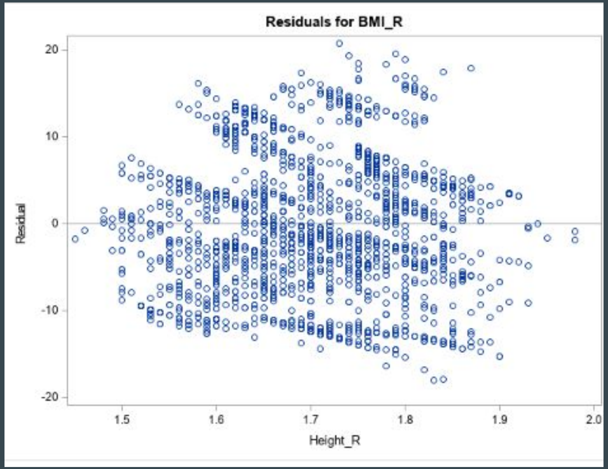
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2290.33765	2290.33765	36.23	<.0001
Error	2109	133323	63.21640		
Corrected Total	2110	135614			

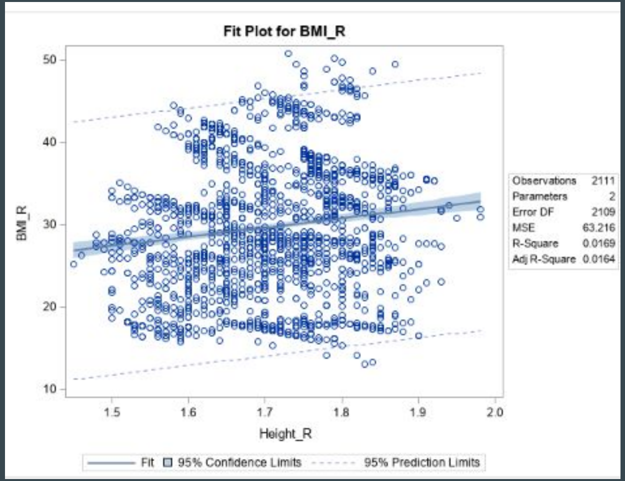
Root MSE	7.95087	R-Square	$\boxed{0.0169}$
Dependent Mean	29.70332	Adj R-Sq	0.0164
Coeff Var	26.76763		

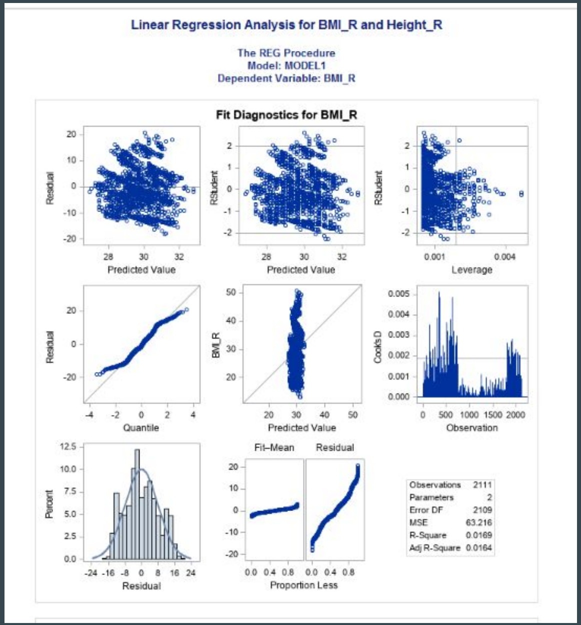
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	10.71567	3.15928	3.39	0.0007
Height_R	1	11.15857	1.85385	6.02	<.0001

Scatter plot and Simple Linear Regression Line







Multiple Regression Analysis

How do factors such as a positive family history, dietary habits, physical activity, and smoking influence the BMI?

SAS Procedure used: PROC GLM

```
/* using proc glm for categorical variables */
ods graphics off;
proc glm data=obesity_data;
title 'Multiple Regression Analysis';
ods select ParameterEstimates;
class family_history_with_overweight (ref = 'yes') favc (ref = 'yes') scc (ref = 'no') faf_r (ref = '0') smoke (ref = 'yes');
model bmi_r = family_history_with_overweight favc scc faf_r smoke / solution;
run;
ods graphics on;
```

Multiple Regression Analysis

The GLM Procedure

Dependent Variable: BMI_R

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	33.07140397	B	1.06074954	31.18	<.0001
family_history_with_ no	-9.04780560	B	0.39810166	-22.73	<.0001
family_history_with_ yes	0.00000000	B	.	.	.
FAVC no	-3.19731602	B	0.48236028	-6.63	<.0001
FAVC yes	0.00000000	B	.	.	.
SCC yes	-2.63304099	B	0.73660023	-3.57	0.0004
SCC no	0.00000000	B	.	.	.
FAF_R 1	-1.01038579	B	0.35340238	-2.86	0.0043
FAF_R 2	-2.04583980	B	0.39865839	-5.13	<.0001
FAF_R 3	-4.20712331	B	0.68071833	-6.18	<.0001
FAF_R 0	0.00000000	B	.	.	.
SMOKE no	-0.14081793	B	1.04257079	-0.14	0.8926
SMOKE yes	0.00000000	B	.	.	.

- Expected BMI for people with a positive family history, frequent high calorie food consumption, non-monitored calorie consumption, no physical activity, and positive smoking habit is 33.07 kg/m2 (Obese)

- Family history, frequent high-calorie food consumption, calorie consumption monitoring, and regular physical activity show a significant association with BMI ($p < 0.05$), while smoking does not ($p = 0.89$).

- This analysis however has its limitations because it does not take into consideration the effect of interactions between the different factors.

Logistic Regression Analysis

Analyzing the impact of age on obesity likelihood (being in the overweight or obese category) using logistic regression

```
data obesity_data;
 set obesity_data;
 if bmi_r < 18.5 then bmi_category = 'Underweight';
 else if bmi_r >=18.5 and bmi_r < 25 then bmi_category = 'Normal';
 else if bmi_r >=25 and bmi_r < 30 then bmi_category = 'Overweight';
 else if bmi_r >=30 then bmi_category = 'Obese';
 run;

data logistic_regression;
 set obesity_data (keep=bmi_category age_r);
 if bmi_category in ('Underweight', 'Normal') then bmi_category_binary=0;
 else bmi_category_binary=1;
 run;

proc logistic data=logistic_regression;
 model bmi_category_binary (event='1') = age_r;
 run;
```

This SAS code conducts a logistic regression analysis for a single variable, Age_R, to predict the likelihood of obesity (bmi_category_binary=1).

Logistic Regression Analysis

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	2470.524	2138.088
SC	2476.179	2149.398
-2 Log L	2468.524	2134.088

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	334.4362	1	<.0001
Score	236.7443	1	<.0001
Wald	204.3220	1	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age_R	1.242	1.206	1.280

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	73.7	Somers' D	0.537
Percent Discordant	20.0	Gamma	0.573
Percent Tied	6.3	Tau-a	0.212
Pairs	881274	c	0.768

The odds ratio for Age_R is 1.242, suggesting that the odds of being in the overweight/obese category increase by approximately 24.2% for each one-unit increase in Age_R.

AIC (Akaike Information Criterion): 2470.524

Summary of Findings:

Weight Distribution:
Weight distribution slightly skewed to the right.
Non-normal distribution observed.

Family History:
81.76% had a positive family history of overweight.

BMI category:
About 45% were in the 'Obese' category.

Transportation:
74.85% used public transportation.

Calorie Monitoring:
Monitoring calorie consumption associated with lower average weight.

Dietary Habits:
Eating between meals significantly associated with BMI categories.
Positive family history linked to frequent high-calorie food consumption.

Gender and BMI:
Statistically significant relationship observed.

Age and BMI:
Weakly positive correlation observed.
Odds of overweight/obese category increased by 24.2% with each one-unit increase in age.

Biometric Characteristics:
Height minimally explained BMI variance.

Associations with BMI:
Family history, frequent high-calorie food consumption, calorie monitoring, and regular physical activity showed significant associations.

Smoking did not show a significant relationship.

Conclusion

- It can be inferred that obesity is influenced by multiple diverse factors, including demographic and physical characteristics, diet, physical activity, and familial history.
- Research and data analytics can identify trends and patterns, and discover correlations between different factors, so as to help in designing and implementing strategies to effectively manage this critical health issue in the region.

Thoughts about improving the project:

- Explore relationships between variables in greater detail by conducting additional tests with varied combinations of variables.
- Investigate alternative methods for cleaning and preparing the dataset to enhance analysis possibilities.

Conclusion

Next steps for the future:

- Conduct more regression analyses to test different types of models to predict obesity levels
- Work with another dataset that is larger and more comprehensive

Positive Takeaways:

- Uncovering insights from real-world datasets and utilizing SAS to address practical challenges

Limitations/Challenges:

- Navigating errors and troubleshooting issues within SAS

References

- Guevara-Ramírez, P., Cadena-Ullauri, S., Ruiz-Pozo, V. A., Tamayo-Trujillo, R., Paz-Cruz, E., Simancas-Racines, D., & Zambrano, A. K. (2022). Genetics, genomics, and diet interactions in obesity in the Latin American environment. *Frontiers in Nutrition, 9.* https://doi.org/10.3389/fnut.2022.1063286
- Holub CK, Elder JP, Arredondo EM, Barquera S, Eisenberg CM, Sánchez Romero LM, Rivera J, Lobelo F, Simoes EJ. Obesity control in Latin American and U.S. Latinos: a systematic review. Am J Prev Med. 2013 May;44(5):529-37. doi: 10.1016/j.amepre.2013.01.023. PMID: 23597819; PMCID: PMC4808744.
- Jiwani, S. S., Carrillo-Larco, R. M., Hernández-Vásquez, A., Barrientos-Gutiérrez, T., Basto-Abreu, A., Gutierrez, L., Irazola, V., Nieto-Martínez, R., Nunes, B. P., Parra, D. C., & Miranda, J. J. (2019). The shift of obesity burden by socioeconomic status between 1998 and 2017 in Latin America and the Caribbean: a cross-sectional series study. *The Lancet. Global health, 7*(12), e1644–e1654. https://doi.org/10.1016/S2214-109X(19)30421-8
- *Obesity rates by country 2023.* Wisevoter. (2023, March 22). https://wisevoter.com/country-rankings/obesity-rates-by-country/
- World Health Organization. (n.d.). *Obesity and overweight.* World Health Organization. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight
- World Obesity Federation, World Obesity Atlas 2023. https://data.worldobesity.org/publications/?cat=19
- https://globalnutritionreport.org/resources/nutrition-profiles/
- *World Obesity Federation Global Obesity Observatory*. (n.d.). World Obesity Federation Global Obesity Observatory. https://data.worldobesity.org/country/
- de Victo, E. R., Fisberg, M., Solé, D., Kovalskys, I., Gómez, G., Rigotti, A., Cortes, L. Y., Yépez-Garcia, M. C., Pareja, R., Herrera-Cuenca, M., Drenowatz, C., Christofaro, D., Araujo, T., Silva, D., & Ferrari, G. (2023). Joint Association between Sedentary Time and Moderate-to-Vigorous Physical Activity with Obesity Risk in Adults from Latin America. *International Journal of Environmental Research and Public Health, 20*(8), 5562. https://doi.org/10.3390/ijerph20085562
- Kain, J., Vio, F., & Albala, C. (2003). Obesity trends and determinant factors in Latin America. *Cadernos de Saúde Pública, 19,* S77-S86.

Appendix 1

Dataset Name : **Estimation of obesity levels based on eating habits and physical condition**

Dataset source :

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

Link to article describing the dataset:

Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico
https://www.sciencedirect.com/science/article/pii/S2352340919306985?via%3Dihub

Appendix 2

Background research and introduction:
- Lizeth Ildefonso

Data Analysis:
- Ketaki Narendra Gharpuray
- Aashi Sethiya

Summary and Conclusion:
- Zy'Ada Hansley

Appendix 3

Topics not covered:

- Importing the dataset in SAS
- Limitations of the study