Variables Affecting Obesity in Colombian, Peruvian, and Mexican Populations

Aishwarya Kotha, Nikita Mathur, Sofia Qian, Vivian Bui, Yijie Lian

01

Problem Introduction & Motivation

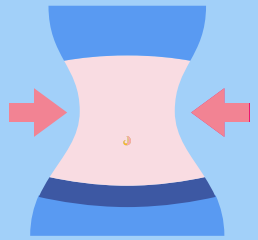
Problem motivation

- Obesity is a growing problem worldwide, and it is important to examine the obesity levels of people to determine how to best remedy this global issue.
- The World Health Organization states that the fundamental cause of obesity is an energy imbalance between calories consumed and calories expended.
- We will examine how people's eating and behavioral habits will affect their obesity levels as determined by mass body index in this study.
- Our study focuses on data collected from Colombia, Peru, and Mexico.

Problem introduction: Classification problem

1.

Given a set of features X, our goal is to determine the class that each data point belongs to



2.

Our problem: Whether a specific set of obesity risk factors and physical health conditions can determine if a person is obese or not.

02
About our data

Our Data

Table 1
Questions of the survey used for initial recollection of information.

Questions	Possible Answers
¿What is your gender?	• Female • Male
¿what is your age?	Numeric value
¿what is your height?	Numeric value in meters
¿what is your weight?	Numeric value in kilograms
¿Has a family member suffered or suffers from overweight?	• Yes • No
¿Do you eat high caloric food frequently?	• Yes • No
¿Do you usually eat vegetables in your meals?	• Never • Sometimes • Always
¿How many main meals do you have daily?	• Between 1 y 2 • Three • More than three
¿Do you eat any food between meals?	• No • Sometimes • Frequently • Always
¿Do you smoke?	• Yes • No
¿How much water do you drink daily?	• Less than a liter • Between 1 and 2 L • More than 2 L
¿Do you monitor the calories you eat daily?	• Yes • No
¿How often do you have physical activity?	• I do not have • 1 or 2 days • 2 or 4 days • 4 or 5 days
¿How much time do you use technological devices such as cell phone, videogames, television, computer and others?	• 0–2 hours • 3–5 hours • More than 5 hours
¿how often do you drink alcohol?	• I do not drink • Sometimes • Frequently • Always
¿Which transportation do you usually use?	• Automobile • Motorbike • Bike • Public Transportation • Walking

- **Imported Data**
- **Learned about the origin of our data –**
  - Collected from populations in Colombia, Peru, and Mexico
  - Used a web based survey with unbiased questions to collect data
  - Questions with multiple answer choices numbered from 0 or 1

Our Data

- **2111 observations, 485 of which is collected**
  - **Why use synthetic data?**
    - Imbalance in classification categories
  - **How was the data generated?**
    - Weka
    - SMOTE
    - 77% of the data is synthetic

- **Cleaned the data keeping synthetically created data issues in mind**
  - Rounded while accounting for whether answers start at 0 or 1

BEFORE

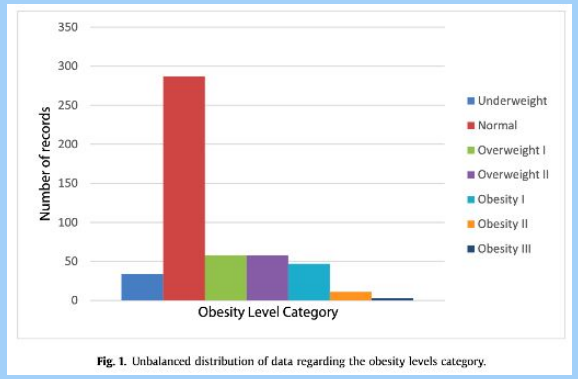


Fig. 1. Unbalanced distribution of data regarding the obesity levels category.

AFTER

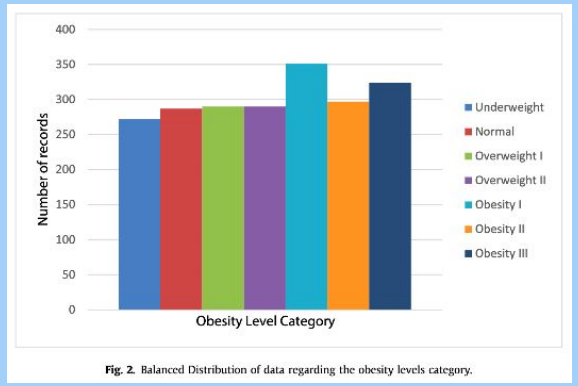


Fig. 2. Balanced Distribution of data regarding the obesity levels category.

Our Data

	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRAN
0	Female	21.0	1.62	64.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	0.0	1.0	no	Public_Transportati
1	Female	21.0	1.52	56.0	yes	no	3.0	3.0	Sometimes	yes	3.0	yes	3.0	0.0	Sometimes	Public_Transportati
2	Male	23.0	1.80	77.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	2.0	1.0	Frequently	Public_Transportati
3	Male	27.0	1.80	87.0	no	no	3.0	3.0	Sometimes	no	2.0	no	2.0	0.0	Frequently	Walki
4	Male	22.0	1.78	89.8	no	no	2.0	1.0	Sometimes	no	2.0	no	0.0	0.0	Sometimes	Public_Transportati

1

Obesity Risk Factors:

- FAVC: frequent consumption of high caloric food
- FCVC: frequency of consumption of vegetables (1, 2, 3)
- NCP: number of main meals (1, 2, 3, 4)
- CAEC: consumption of food between meals
- CH2O: consumption of water daily (1, 2, 3)
- CALC: consumption of alcohol

2

Physical Health and Condition:

- SCC: calories consumption monitoring
- FAF: physical activity frequency (0, 1, 2, 3)
- TUE: time using technology devices (0, 1, 2)
- MTRANS: transportation used

3

Obesity Variables (Levels)

- Insufficient Weight
- Normal Weight
- Overweight Level I
- Overweight Level II
- Obesity Type I
- Obesity Type II
- Obesity Type III

Obesity groupings were determined by values using the following equation: Mass Body Index = Weight/(height*height), which were then compared with information from the WHO and Mexican Normativity.

About Obesity and BMI

How is obesity defined?

Having excessive body fat that increases the risk of health problems

What is BMI?

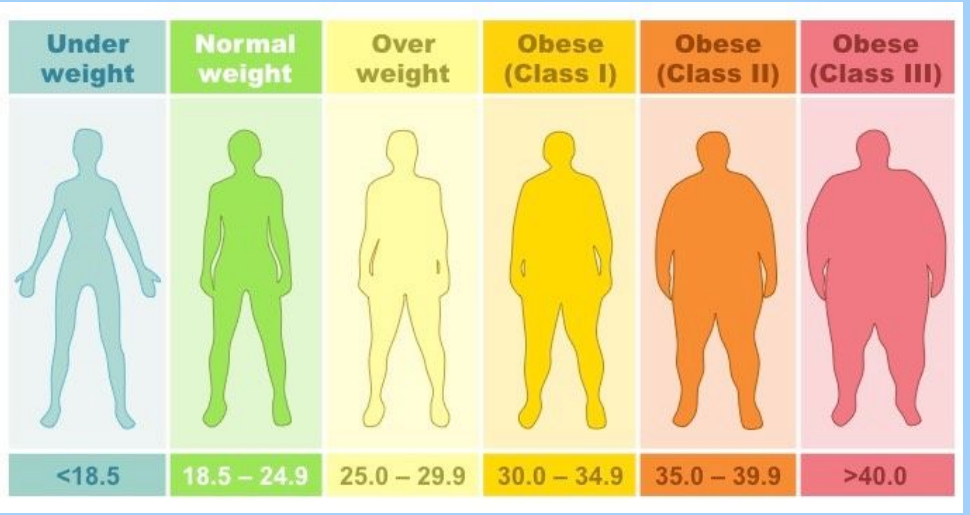
BMI is Body Mass Index, calculated by dividing mass (kg) by height squared (m^2)

CDC BMI Ranges

Underweight range → <18.5
Healthy Weight range → 18.5-24.9
Overweight range → 25.0-29.9
Obese range → >30.0

Obesity Classes/Types

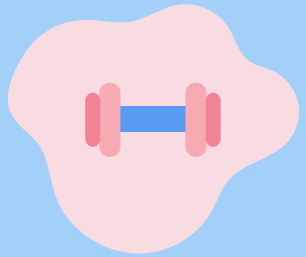
Obese Class 1 → 30.0-34.9
Obese Class 2 → 35.0-39.9
Obese Class 3 → >40.0



03

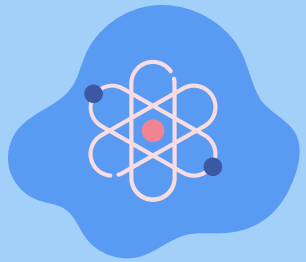
Exploratory Analysis

Exploratory Analysis



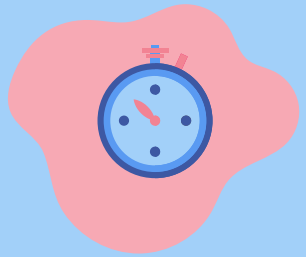
1.

Compared BMI with personal and frequency variables and visualized their relationships



2.

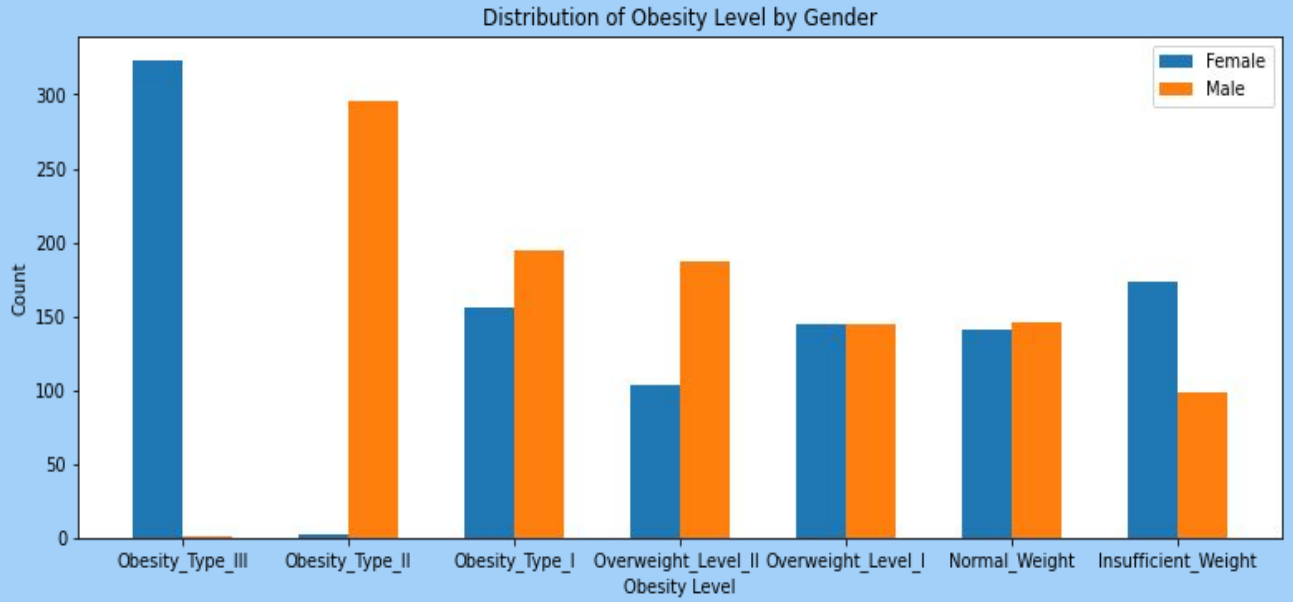
Compared BMI with **Family History**, **Frequency of Vegetable Consumption,** and **Frequency of Physical Activity** and visualized their relationships



3.

Conducted regressions on their relationships

Exploratory Analysis

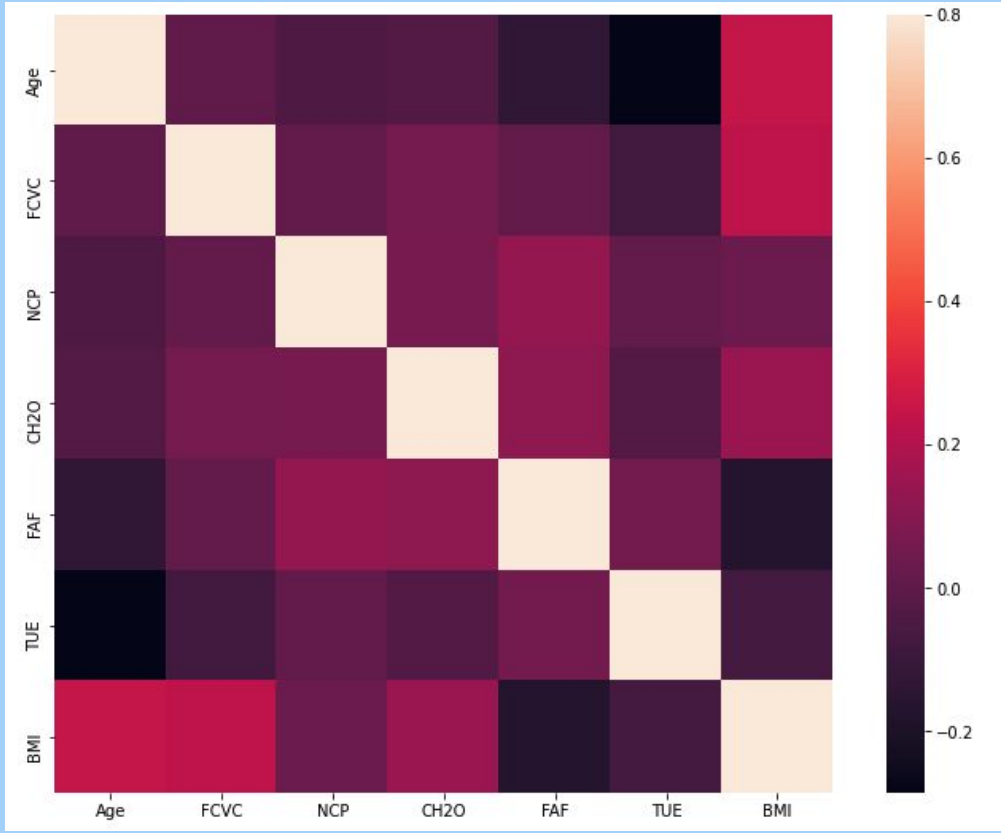


Distribution of obesity level by gender –

Data shows that maximum female respondents fell under Obesity Type III and maximum male respondents fell under Obesity Type II

Exploratory Analysis

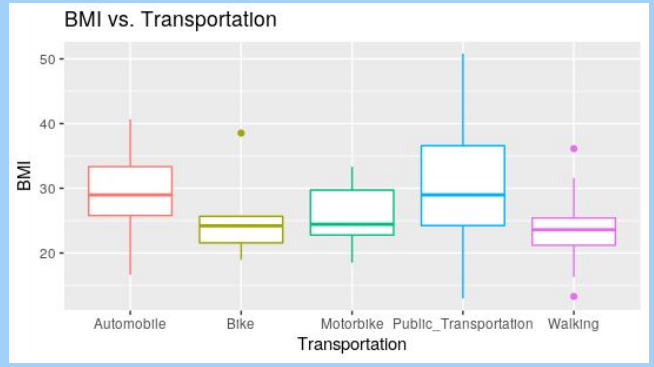
Heat Map to determine variables with strong correlations

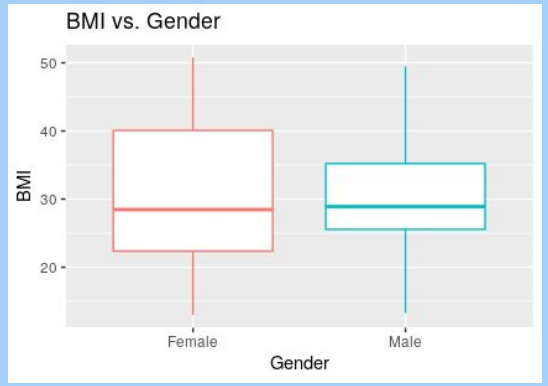


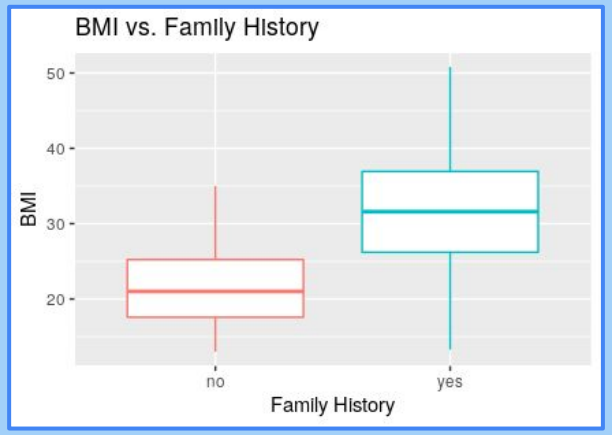
Identified strong correlations –

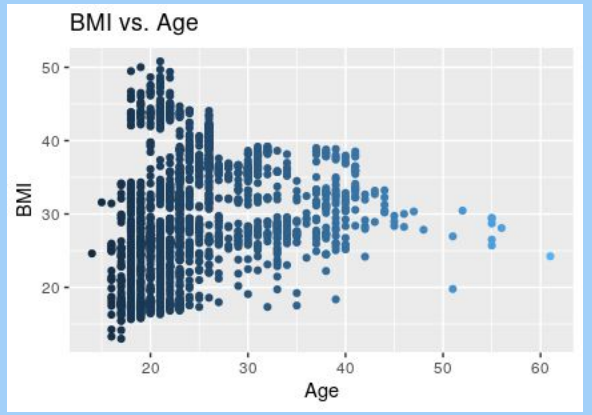
- FAVC - Frequency of Vegetable Consumption
- FAF - Frequency of Physical Activity
- Family History - Family History of Obesity

Personal Variables vs BMI

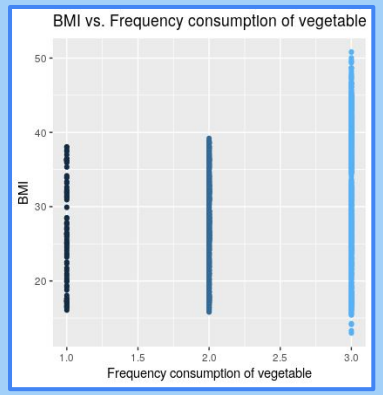


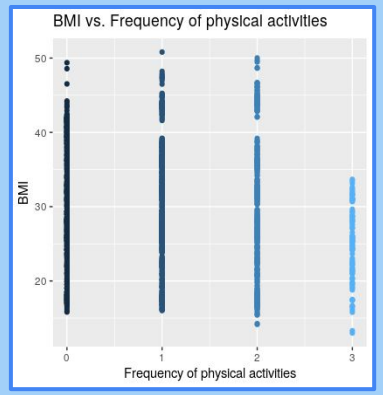


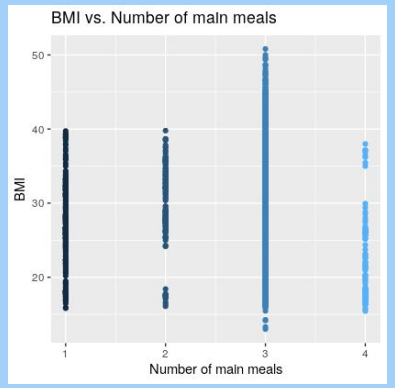


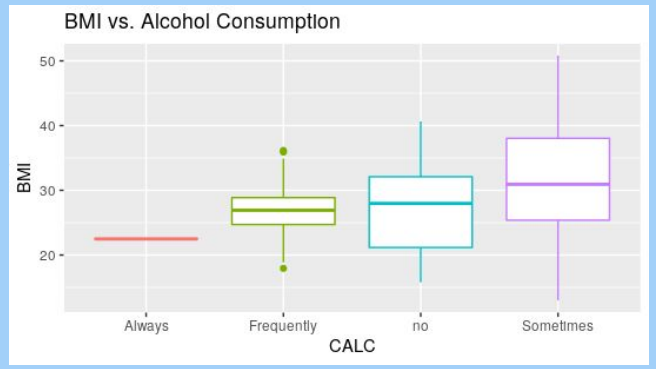


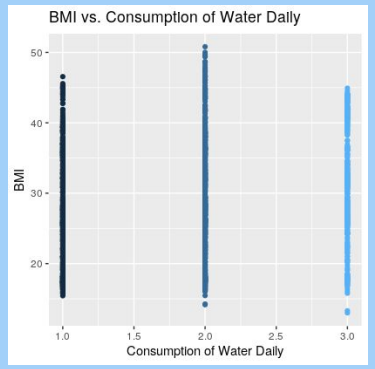
Frequency Variables vs BMI

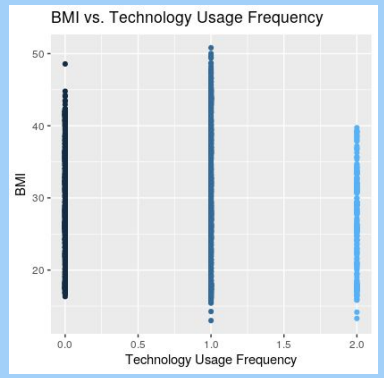


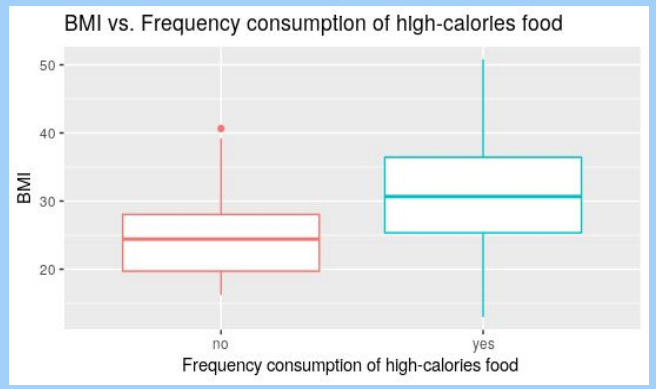




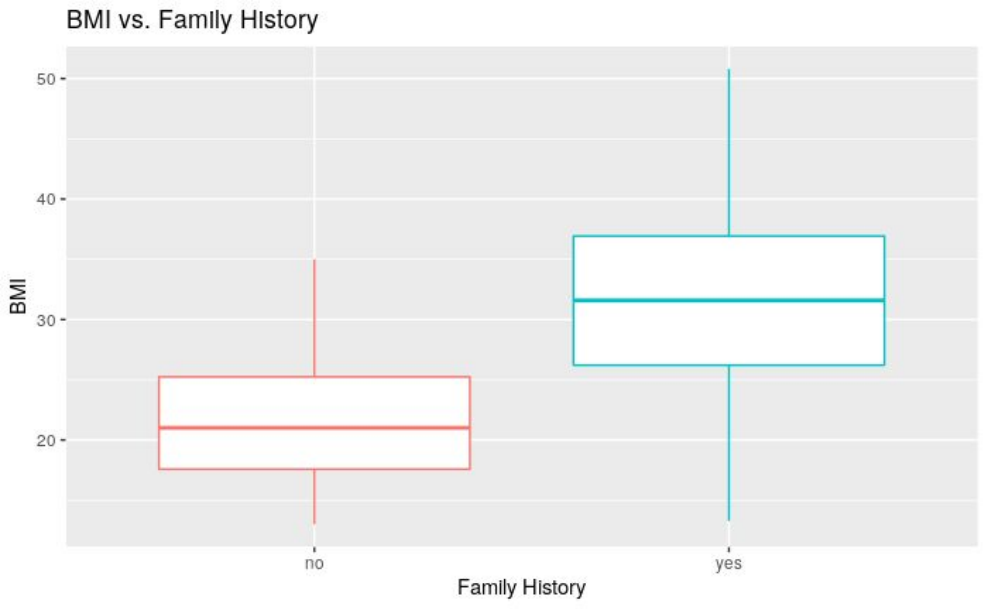








BMI vs Family History

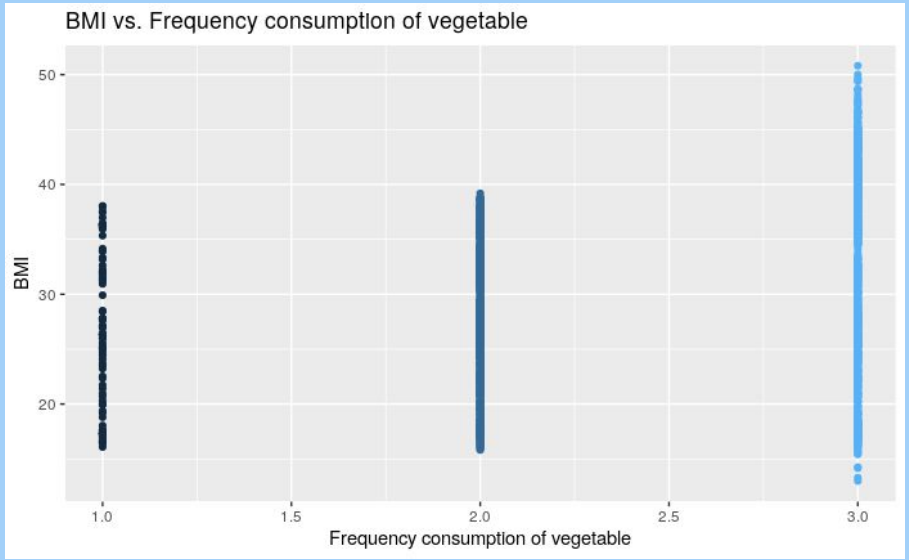


Dep. Variable:	BMI	**R-squared:**	0.234
Model:	OLS	**Adj. R-squared:**	0.233
Method:	Least Squares	**F-statistic:**	643.5
Date:	Tue, 30 Nov 2021	**Prob (F-statistic):**	4.03e-124
Time:	20:47:46	**Log-Likelihood:**	-7106.5
No. Observations:	2111	**AIC:**	1.422e+04
Df Residuals:	2109	**BIC:**	1.423e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	21.5005	0.357	60.144	0.000	20.799	22.202
C(family_history_with_overweight)[T.yes]	10.0287	0.395	25.367	0.000	9.253	10.804

Omnibus:	27.897	**Durbin-Watson:**	0.477
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	17.020
Skew:	0.006	**Prob(JB):**	0.000201
Kurtosis:	2.560	**Cond. No.**	4.48

BMI vs FCVC

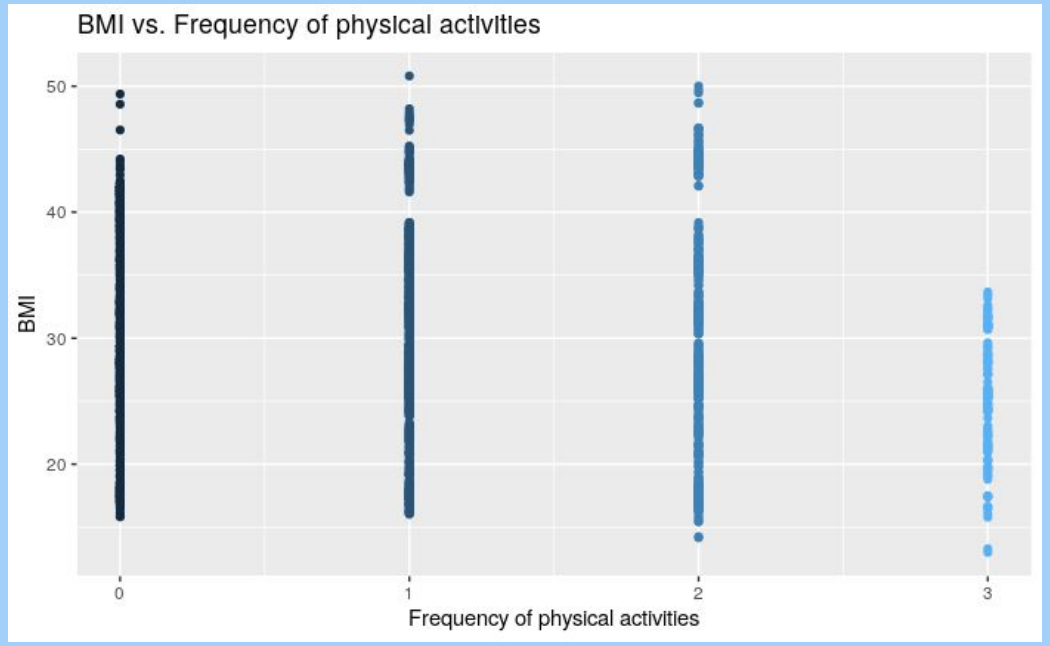


Dep. Variable:	BMI	**R-squared:**	0.061
Model:	OLS	**Adj. R-squared:**	0.060
Method:	Least Squares	**F-statistic:**	136.0
Date:	Tue, 30 Nov 2021	**Prob (F-statistic):**	1.72e-30
Time:	20:47:46	**Log-Likelihood:**	-7321.6
No. Observations:	2111	**AIC:**	1.465e+04
Df Residuals:	2109	**BIC:**	1.466e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	24.2604	0.496	48.892	0.000	23.287	25.233
C(FAVC)[T.yes]	6.1540	0.528	11.660	0.000	5.119	7.189

Omnibus:	124.261	**Durbin-Watson:**	0.263
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	47.169
Skew:	0.050	**Prob(JB):**	5.72e-11
Kurtosis:	2.275	**Cond. No.**	5.71

BMI vs Frequency of physical activities



Dep. Variable:	BMI	**R-squared:**	0.032
Model:	OLS	**Adj. R-squared:**	0.031
Method:	Least Squares	**F-statistic:**	23.45
Date:	Tue, 30 Nov 2021	**Prob (F-statistic):**	6.32e-15
Time:	20:47:47	**Log-Likelihood:**	-7352.9
No. Observations:	2111	**AIC:**	1.471e+04
Df Residuals:	2107	**BIC:**	1.474e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	31.0088	0.294	105.504	0.000	30.432	31.585
C(FAF)[T.1.0]	-1.0723	0.408	-2.628	0.009	-1.873	-0.272
C(FAF)[T.2.0]	-2.5024	0.460	-5.438	0.000	-3.405	-1.600
C(FAF)[T.3.0]	-5.7915	0.780	-7.421	0.000	-7.322	-4.261

Omnibus:	115.984	**Durbin-Watson:**	0.224
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	51.159
Skew:	0.155	**Prob(JB):**	7.78e-12
Kurtosis:	2.303	**Cond. No.**	5.28

Regression on Selected Variables

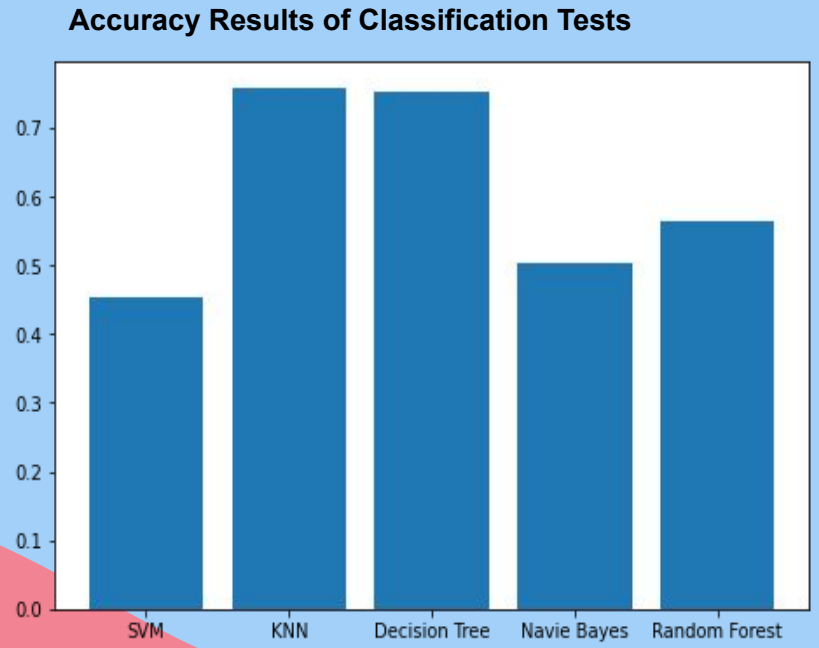
Dep. Variable:	BMI	**R-squared:**	0.275
Model:	OLS	**Adj. R-squared:**	0.273
Method:	Least Squares	**F-statistic:**	159.4
Date:	Wed, 01 Dec 2021	**Prob (F-statistic):**	5.85e-144
Time:	11:17:12	**Log-Likelihood:**	-7048.6
No. Observations:	2111	**AIC:**	1.411e+04
Df Residuals:	2105	**BIC:**	1.414e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	20.1732	0.551	36.628	0.000	19.093	21.253
C(family_history_with_overweight)[T.1]	9.2656	0.394	23.515	0.000	8.493	10.038
C(FAVC)[T.1]	3.4552	0.477	7.243	0.000	2.520	4.391
C(FAF)[T.1.0]	-1.0223	0.354	-2.891	0.004	-1.716	-0.329
C(FAF)[T.2.0]	-2.0528	0.399	-5.141	0.000	-2.836	-1.270
C(FAF)[T.3.0]	-4.3431	0.680	-6.384	0.000	-5.677	-3.009

- Family history has higher coefficient compared to other variables (9.266)

03
Classification Tests

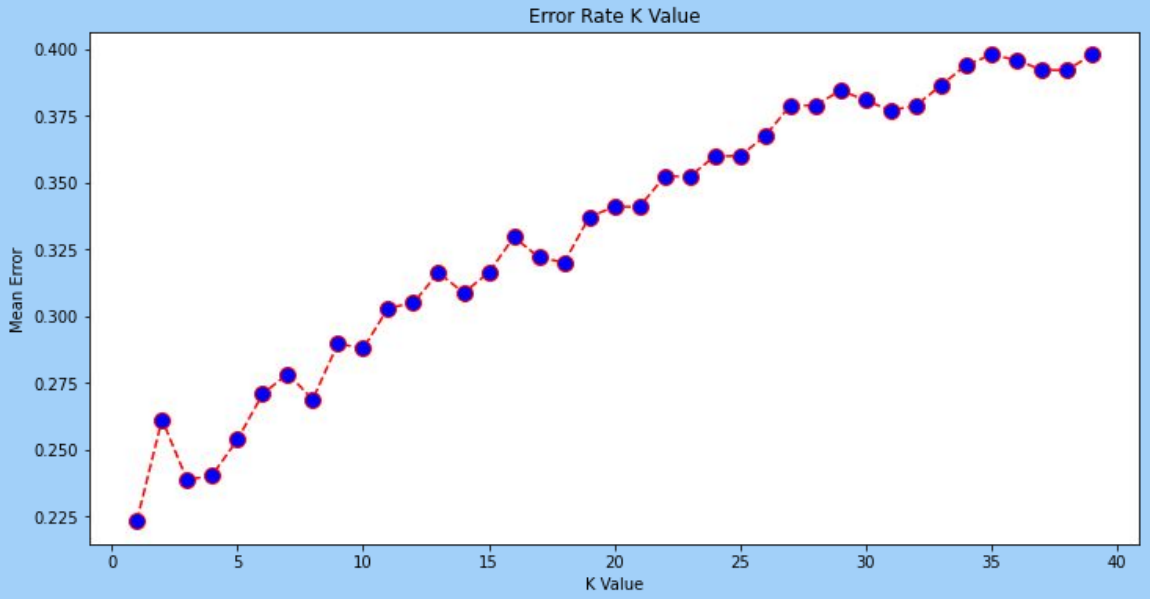
Classification Tests



- **Conducted classification tests**
  - SVM (Accuracy: 0.454)
  - KNN (Accuracy: 0.759)
  - Decision Tree (Accuracy: 0.752)
  - Naive Bayes (Accuracy: 0.504)
  - Random Forest (Accuracy: 0.564)

- **KNN has the highest accuracy score**

KNN: K-Nearest Neighbours



Optimal K value?

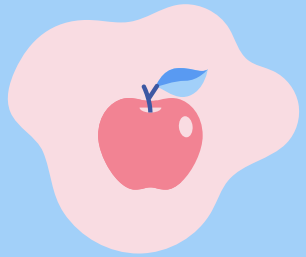
- Derived a plot between error rate and K values ranging from 0-40

- K value = 3, shows minimum error rate
- In the future: Pick either 4 or 7 based on the purpose of the study

Limitations

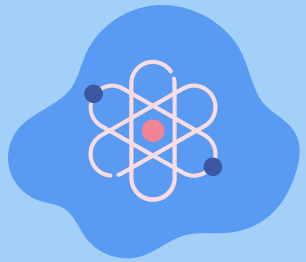
- KNN limitations → Lazy algorithm and expensive, not good at picking up edge cases
- 77% of data is synthetic, might not be applicable for the populations data was collected from
- Survey was only available for 30 days
- Response Bias, volunteer bias
- Randomization issue
- Lack of external validity

Conclusion



1.

KNN model helped produce the highest accuracy.



2.

On KNN model, our classification was about 70% correct to produce the obesity level

3.

Among the various factors, family history had the highest coefficient and R^2 values

Key Takeaways & Next Steps

- **Key Takeaways**
  - Process of Cleaning Data
  - Learned how to train data and compare classification models

- **Want to collect more data that more accurately represents the populations in question**

- **Test the KNN model we created**

- **Study pattern in each category**

Questions?