How General Health Index (GHI) performs better to predict health levels*

A comparitive analysis of GHI and BMI against body measurements

Aamishi Avarsekar

December 3, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

BMI is a common indicator for general health but it has its limitations. It does not take into consideration the actual body composition of an individual and other important factors such as sex and age. Through this paper, I would like to propose a new way categorising people based on other body measurements such as the waist-to-hip ratio and wrist and ankle circumferences measurements as they are more indicative of a person's visceral fat, this proving a more accurate representation of general health. I would like to propose a newer formula for General Health Index (GHI) through the BodyM dataset. I plan on using the Isaac Kuzmar dataset that has accurate measurements of body fat to create a Confusion Matrix to see how my model compares in providing more accurate body fat estimations through easlily measurable data such as waist and hip circumference.

- this paper uses 2 datasets to capture the difference in fat assessment for popular methods. BMI is a popular way of estimating health issues in an individual but it fails to capture details like body composition. In its initial design, BMI was designed using only male [] subjects and thus cannot be applied to females. However, due to excess popularity, it is used as a self-assessment to estimate your body composition and health risks. However, due to its inability to capture body composition, a lot of false positives and false negatives are produced. These indivuals often identify health risks [] later in the onset. My aim with this paper is to introduce simple measurements that can be recorded at home and help indivuals recognise when they should seek mdeical attention.

*Code and data are available at: https://github.com/aamishi/ImprovedGeneralHealthIndex/

1.1 BMI Classification:

Table 1: BMI Categories as per the Government of Canada Guidelines

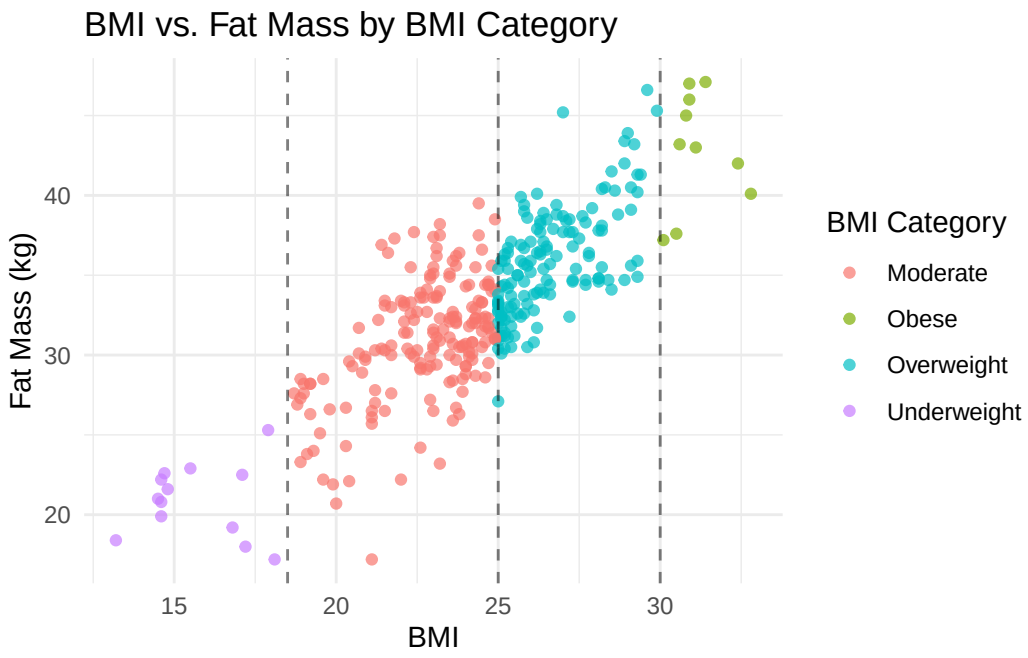
Category	BMI Range
Underweight	Below 18.5
Normal weight	18.5–24.9
Overweight	25.0–29.9
Obese - Class I	30.0–34.9
Obese - Class II	35.0–39.9
Obese - Class III	40.0 and above

1.1.1 The difference between an underweight and normal person:

Let us take a closer look at the BMI levels below 25 for underweight and normal weight adults.

2 Data Section: Part 1: What is the difference between an underweight / athletic person /

2.0.1 All categories



A new way of estimating general body health is to use WHR. This ratio aims to capture the excess fat deposit in a person's abdominal area which is were most people store adipose fat. A high WHR indicates high levels of fat. However, this is not enough. This is simply the most effective way of identyfing the most obvious and dangerous place of fat accumulation.

3 Data

3.1 Overview

3.1.0.1 General Statement

- [is this data observational or sample] The data used for this paper has two different sources and serve two different purpose. The first dataset considered is the BodyM dataset from Amazon Web Services (AWS) [ref]. This dataset serves as the training data for the model to predict the category of fat in a person. This dataset also included silhouettes to capture accurate body compositions of real test subjects. [the purpose of this]. This data was collected to support the estimation of bodily measurements using Machine Learning [reference]. The silhouettes are not being used for the purposes of this paper.
- [is this data observational or sample] The second dataset is the Body Fat Distribution distribution collected for estimating a new body fat measurement technique by Isaac Kuzmar Et Al. [reference]. For the purposes of this paper, this dataset would be refered to as the Body Compositon dataset as it used to calculate fat percentages based on actual body composition. It serves as a comparision for the model's predictions.

3.1.1 BodyM Dataset - AWS Marketplace for Open Data [fix this heading]

The BodyM Dataset, refered to teh Body Measurements dataset throughout this paper, was collected for [Ruiz2022] Ruiz Et Al. in their paper with a focus on underrepresented body types in estimation of fat and its susbquent health risks. The main data captured from this collection was the front and lateral silhouttes of [2000something] test subjects. These silhouttes were then converted to black and white images to be used by their augementation model. The subjects include X male and Y female aged A to B. The body measurements in this dataset were generated through their adversarial body simulator (ABS). The ABS was specifically modeled to capture underrespresented body types. The S3 [reference] package included 3 datasets: training, Test A and Test B. This paper uses the training dataset. The visual images of test subjects were photographed and 3D-scanned by lab technicians. - Tibble of this dataset after cleaning:

```
# measurements_tibble <- as_tibble(cleaned_data_model)
# kable(measurements_tibble, caption = "Example Tibble Output")
```

3.1.1.1 The variables of my use:

For this paper, this dataset is used to estimate the category of body fat that a person carries. To effectively differentiate the different body types, the model considers measurements that could indicate actual body compositions. I chose gender, height and weight and basic predictors for the model. However, according to [reference, WHO], waist-to-hip ratio is more accurate to predict levels of body fat in a person. Based on [reference], the two sexes have different tendencies of excess body fat collection also known as [adipose fat]. In males, excess body fat tends to collect in the abdominal region, whereas in females, fat collects in the hips. This ratio is indicative of fat accumulation and a higher ratio can be used as a precursor indicator for obesity [reference].

To consider body composition along the height of a person, a new variable called height-to-hip ratio, height_hip_ratio, was created. [referernce], something like two people with the same WHR could differ in the fat category that they belong to depending on the height. A tall person with a higher WHR is at a lower risk of developing cardiovascular risks than a shorter person with the same height [reference].

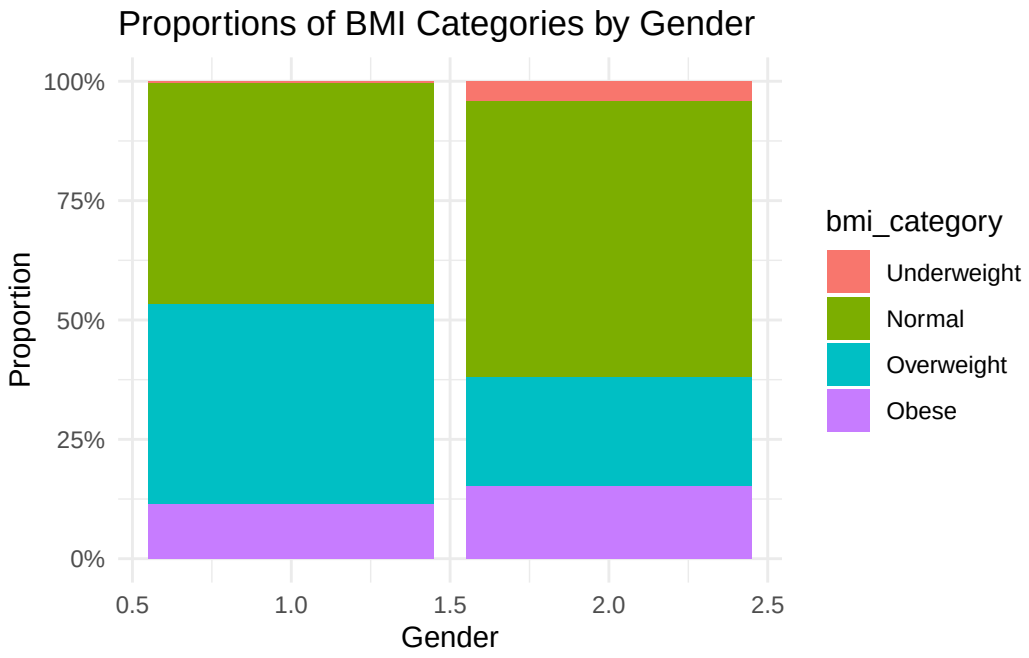
Lastly, to improve the quality of the model, I also selected ankle and wrist circumference measurements. [reference], [this is important for boundary values].

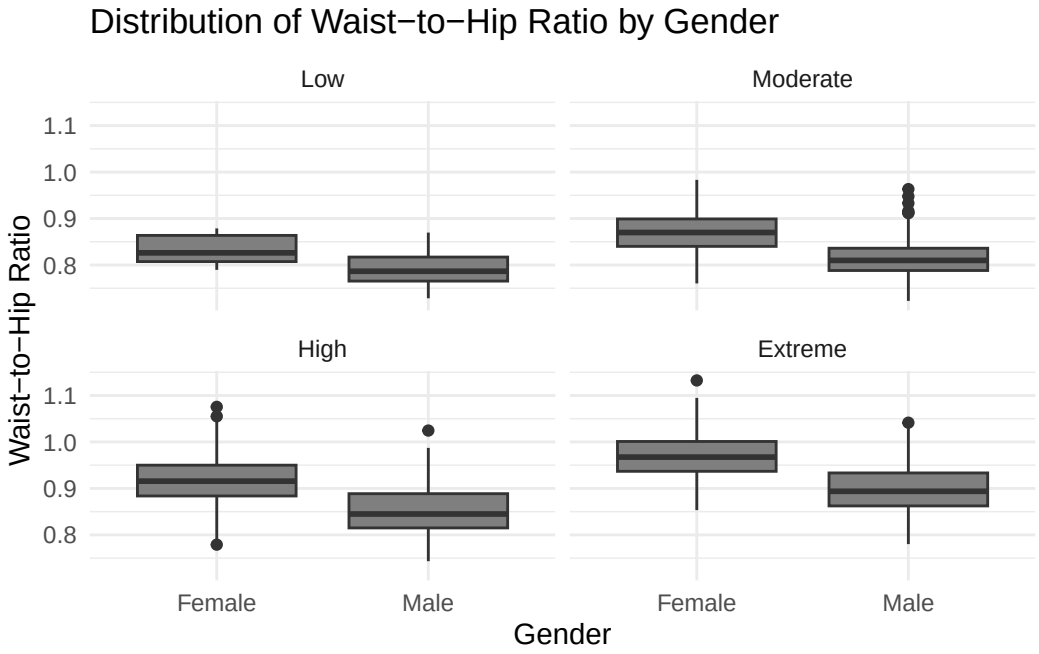
- include a table for WHR and who's statement

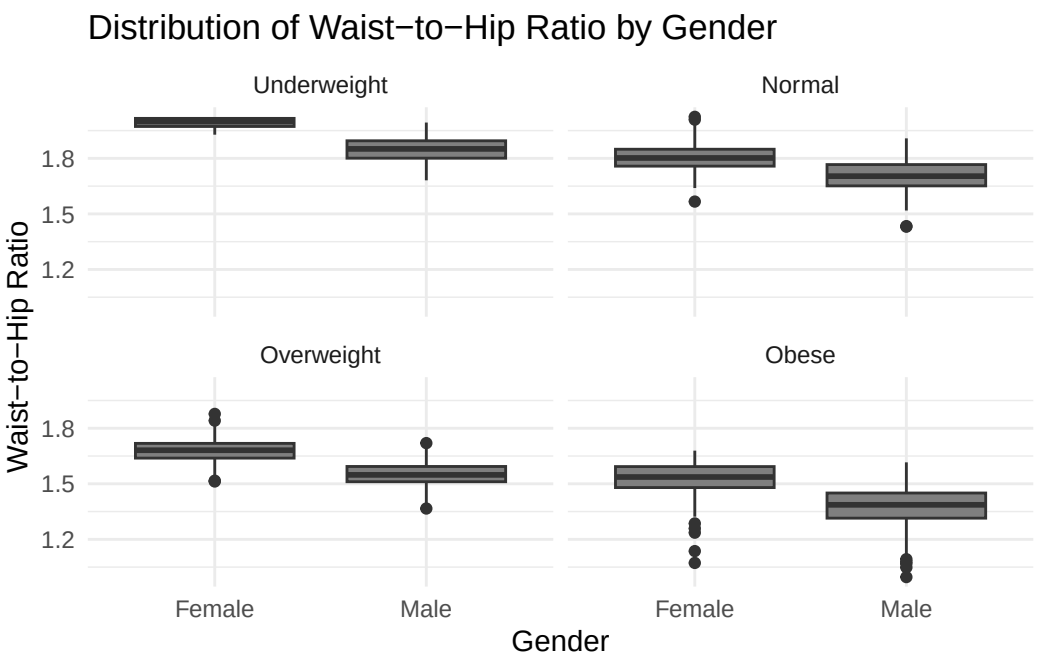
3.1.1.2 The estimand for the model

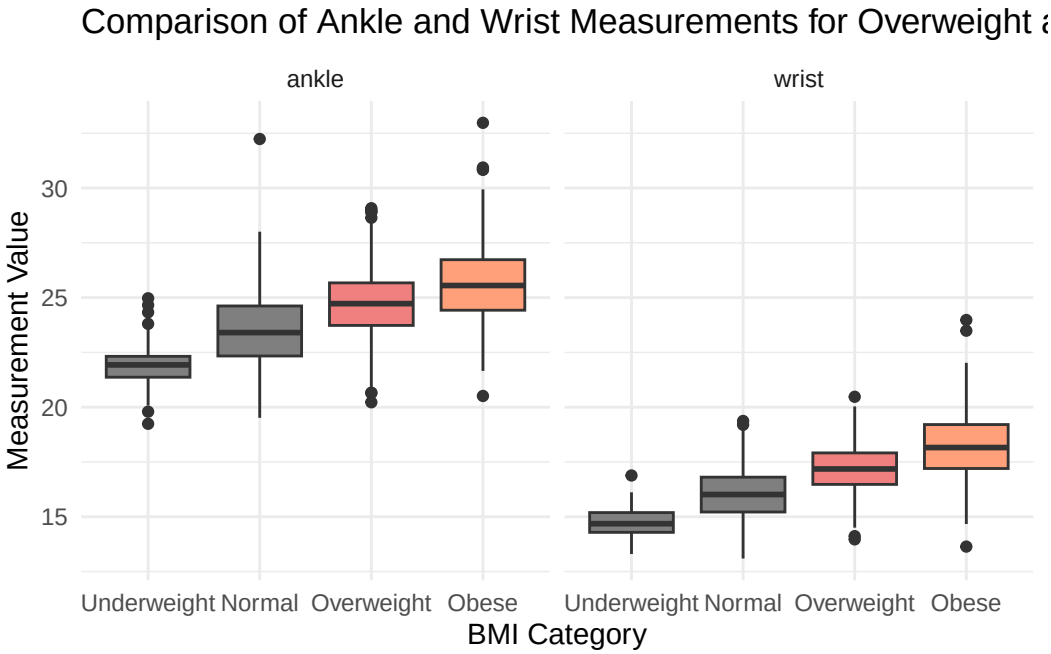
The estimand for the model, fat_percentage_category, is created using WHR as a proxy initially. It is calculated based on the subjects gender and WHR as males and females have different fat accumulation patters. The ratio intervels for these categories was derived from []. The model then predicts a more accurate fat percentage category based on other predictors.

- visualisations.


Proportions of BMI Categories by Gender
100%
75%
50%
25%
0%
Proportion
0.5
1.0
1.5
2.0
2.5
Gender
bmi_category
Underweight
Normal
Overweight
Obese


Distribution of Waist−to−Hip Ratio by Gender
Low
Moderate
High
Extreme
1.1
1.0
0.9
0.8
Waist−to−Hip Ratio
Female
Male
Gender





3.1.1.3 graph what i will be using in my model

3.1.2 Body Composition Data - Isaac Kuzmar Et Al.

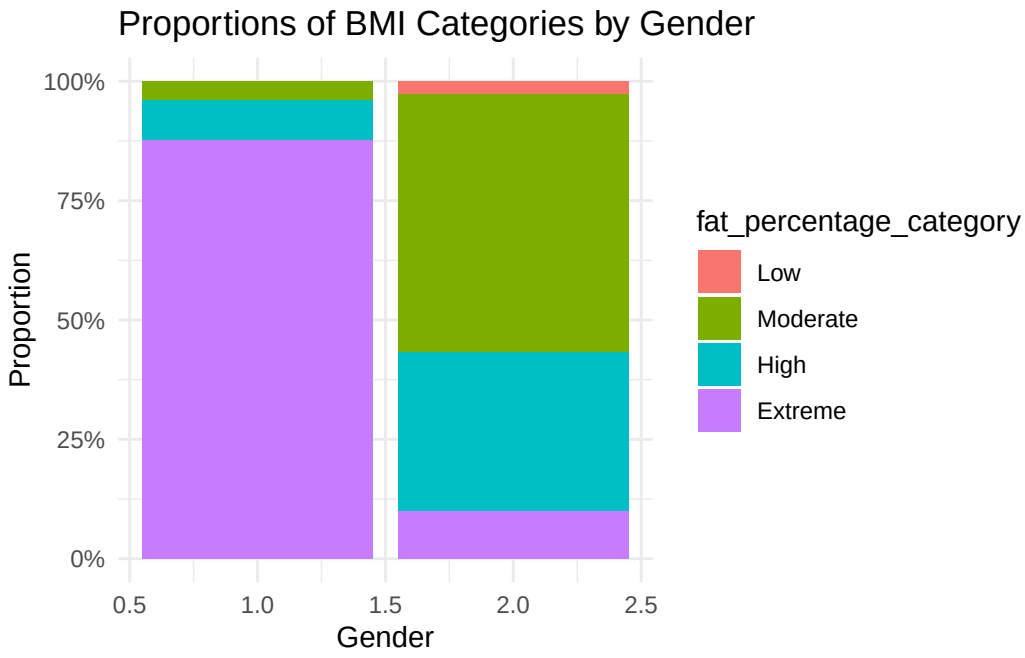
3.1.3 Data Overview

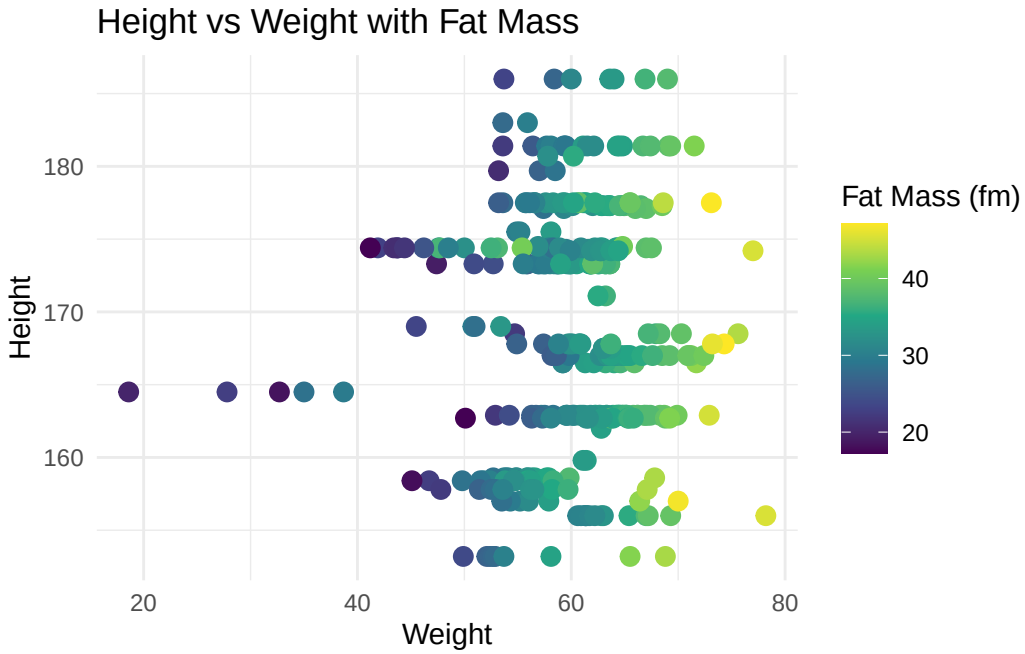
The dataset compiles the body measurements of subjects aged, 18 and 60, and specifically with a desire to lose weight and improve body image. The participants resided in Barranquilla, Colombia and consisted of 234 males and 111 females. Medical exclusions were made while recruiting the subjects, such as pregnant women or people with medical pacemakers. The bodily measurements such as fat mass in kilograms and fat free mas in kilograms were determined using the Tanita MC-780 [reference] body composition analyzer.

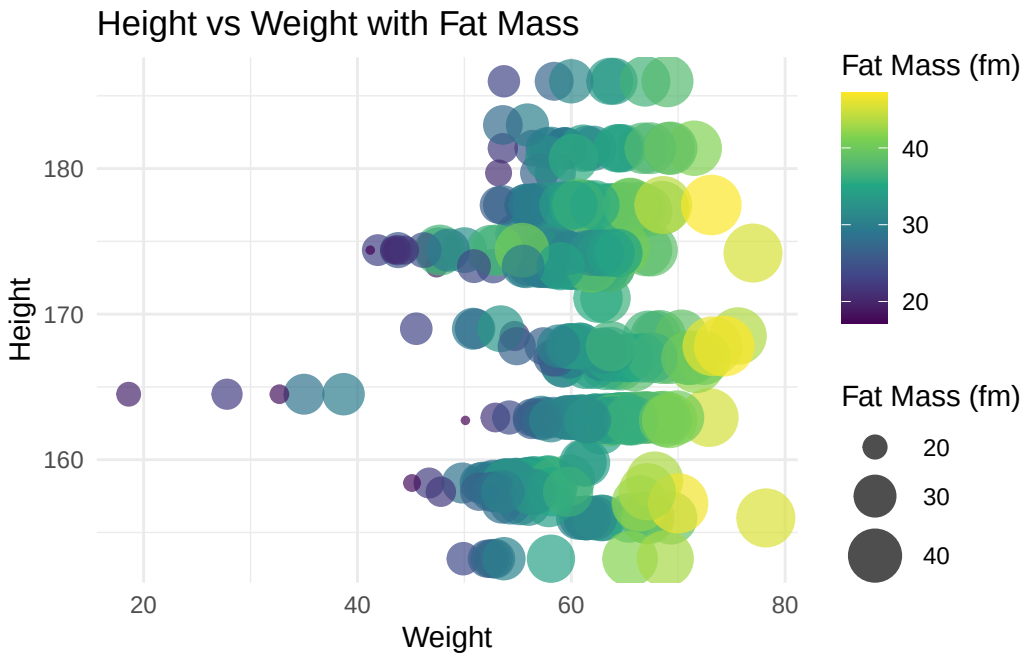
3.1.4 The variables of my use:

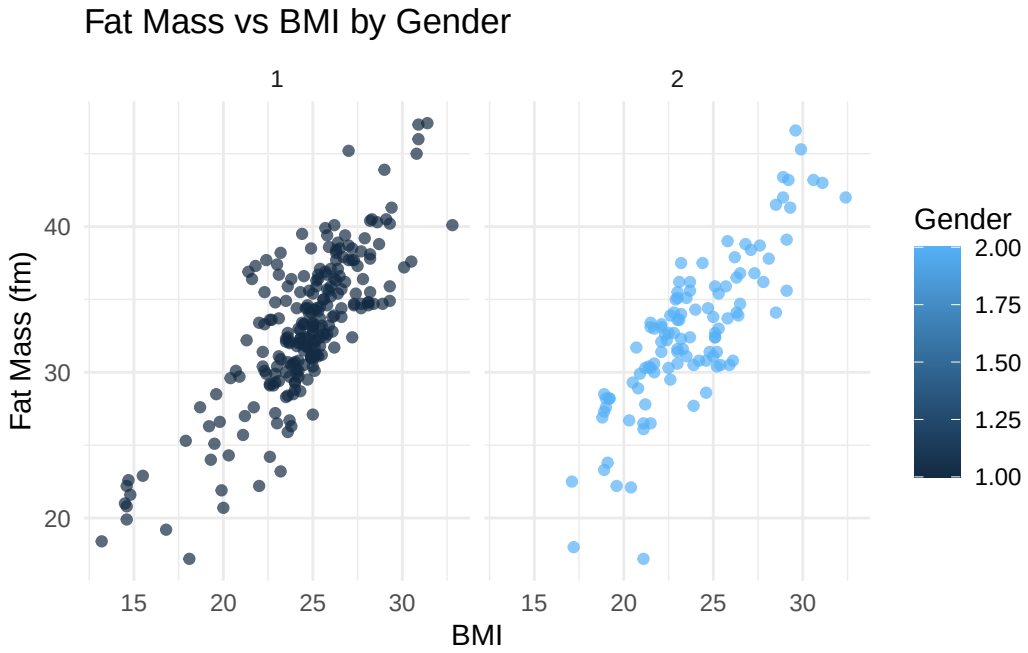
The main variables that I considered fat mass percentage fm. This variable is obtained by dividing the fat mass in kilogram by the total weight in kilograms. The fat mass percentage is considered the main predictor for fat mass categories. As the body composition is more accuarate at measuring the actual amount of excess fat a person carries and can differentiate it from other weights such as fat free muscle mass [reference the variable] #### graph what i will be using in my model

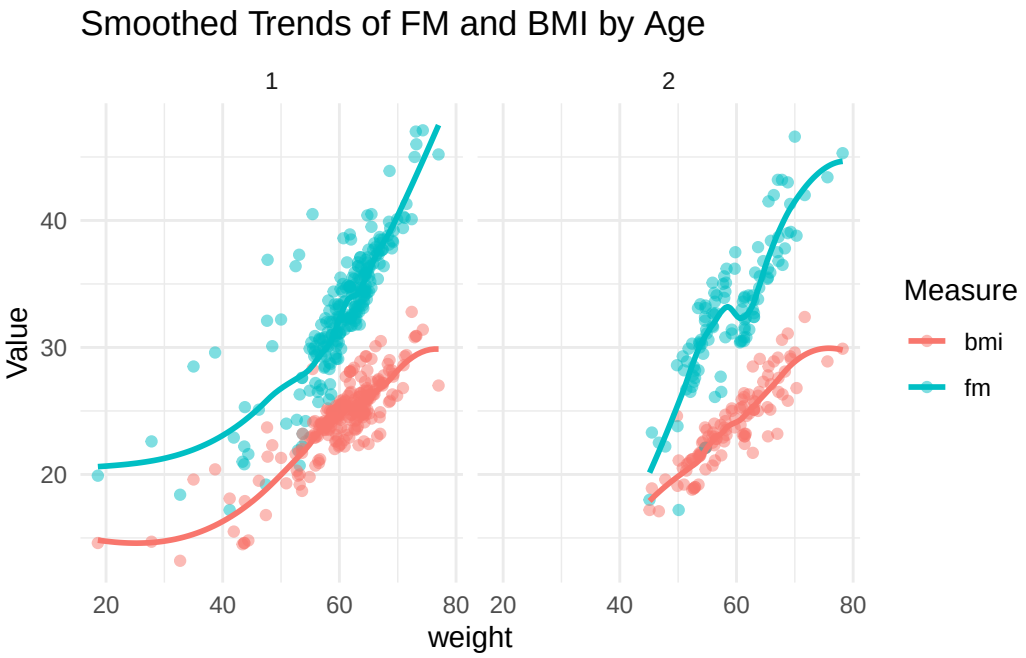
```
# A tibble: 345 x 15
   gender   age height weight   bmi fat_mass_kg    fm ffm_kg bone_mass_kg
   <fct>  <dbl>  <dbl>  <dbl> <dbl>       <dbl> <dbl>  <dbl>        <dbl>
 1 Female    18    156   61.4  23          19.4  31.6   42           2.1
 2 Female    18    156   61.4  23          19.3  31.4   42.1         2.1
 3 Female    18    156   61.3  26.1        18.9  30.8   42.4         2.2
 4 Female    18    156   62.8  21.7        20.7  33     42.1         2.1
 5 Female    18    156   63    25.1        20.6  32.7   42.4         2.2
 6 Female    18    156   65.4  23          23.2  35.5   42.2         2.2
 7 Female    18    156   67    23.2        25.1  37.5   41.9         2.1
 8 Female    19    156   67.2  27.6        26    38.7   41.2         2.1
 9 Female    19    156   69.3  29.1        27.1  39.1   42.2         2.2
10 Female    19    156   78.2  29.9        35.4  45.3   42.8         2.2
# i 335 more rows
# i 6 more variables: muscle_mass_kg <dbl>, bmr_kcal <dbl>, fm_trunk <dbl>,
#   bmi_category <chr>, fat_percentage_category <fct>, height_category <fct>
```


Proportions of BMI Categories by Gender
100%
75%
50%
25%
0%
Proportion
0.5
1.0
1.5
2.0
2.5
Gender
fat_percentage_category
Low
Moderate
High
Extreme


Height vs Weight with Fat Mass
180
170
160
Height
20
40
60
80
Weight
Fat Mass (fm)
40
30
20


Height vs Weight with Fat Mass
Height
180
170
160
20
40
60
80
Weight
Fat Mass (fm)
40
30
20
Fat Mass (fm)
20
30
40


Fat Mass vs BMI by Gender
1
2
Fat Mass (fm)
40
30
20
15
20
25
30
BMI
Gender
2.00
1.75
1.50
1.25
1.00



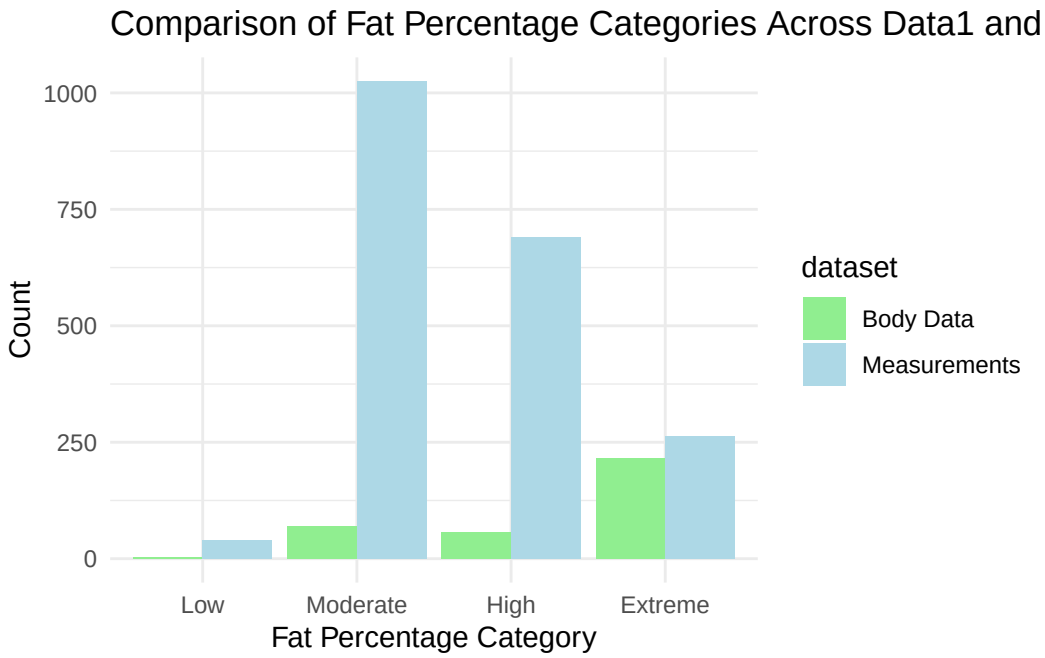
- talk about the crowding for men

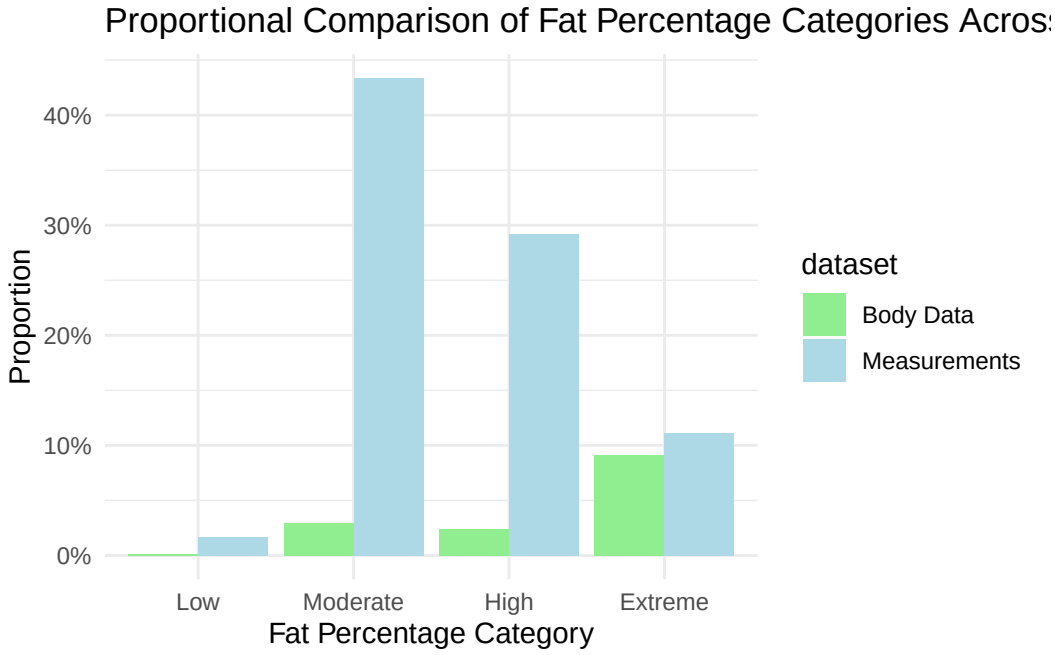
3.2 what original BMI and WHR fail to capture

3.3 How are these data sets important are how they relate to each other

3.3.0.1 desitination

3.3.0.2 well as any relationships between the variables.





Overview text

3.4 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

Talk more about it.

Talk way more about it.

3.5 Predictor variables

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

4 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

- multinomial logistic regression because we are predicting multiple categories

4.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

4.1.1 Model justification

- I wanted to provide a categorisation of fat levels in an indivual based on simple measurements. Using measurements such as height and weight that are used in the formula and WHR and ankle and wrist measurements, I wanted to capture a person's body compositon as much as possible. i expected that there is a positve relationship between a person's excess body fat levels and their WHR. In addition to WHR, the height if a person is also important to capture the distribution of fat, a short person with a high WHR would be at a higher risk of obesity related diseases.
- men are more likely to collect fat in their abdomen area and women in their hip area. To capture nature body shapes and leanness, i also considered….

4.2 Model General Overview:

This paper uses 2 models to model the two datasets. The first model, the Body Measurements Model, would be used to predict fat percetage categories using the measurements of person's body that can be obtained using a measuring tape. The second model, the Body Composition Model, is also designed to access the relation between different bodiliy compositions such as total body weight, excess fat weight, lean muscle mass in kilograms.

I used the packages [] to do []. Both models employ Multilevel Regression [] as the Fat Percentage Category has multiple categories including: essential, athlete, normal, high.

4.2.1 Model 1:

In this model, the dependent variable is fat_percentage_category. The model is set up as follows

$$\log\left(\frac{P(y=j)}{P(y=\text{ref})}\right) = \beta_{0j} + \beta_{1j}\cdot\text{height} + \beta_{2j}\cdot\text{gender} + \beta_{3j}\cdot\text{waist\_hip\_ratio} + \beta_{4j}\cdot\text{height\_hip\_ratio} \tag{7}$$

Where: - (y) is the **fat percentage category**. The categories include low, moderate, high, and extreme. These equate low, moderate, high and extreme levels of body fat. - (P(y = j)) is the probability that the fat percentage falls into category (j). - (P(y = ref)) is the reference category. - (*{0j},* {1j}, …) are the coefficients to be estimated for each category (j).

4.2.2 Model Justification

It is incredibily challenging to differentiate the body composition of a person based soley on their total weight, height and weight only. However, to increase the [goodness] of the classification, this model uses the waist_to_hip ratio, which is especially an important indicator for fat accumualtion across both genders. The model also uses height_hip_ratio to account for different body types. [this sentence might go in the datasection] As height is an important factor to evaluate the collection of fat in latitdunial(how tf do you spell it) places on a person's body, this is an important factor to consider. However, using height simply does not effectvely communicate how a person's body is composed latitudinally. So I chose to employ the height_hip_ratio ratio to provide more classification between body types. Additional features such and ankle and wrist are used to further differentiate between the categories.

4.2.3 Assumptions and limitations

Assumptions: - this model assumes that there is an obvious difference in body measurements and fat storing capacities in males and females, hence gender is an important predictor and also affects the proxy values of fat percentage owing to the calculation based on WHR.

Limitations:

- The dataset does not include the age of the particapants. while a general age range is provided, such as 18-X years old, different age groups could differ in how they are categorised.
- This model would not be able to differentiate between age groups and hence would work poorly if this information is included.
-

4.2.4 Model validation

4.2.5 Model 2:

In this model, the dependent variable is fat_percentage_category. The model is set up as follows

- fat_percentage_category = beta_0 + beta_1() + beta_() + beta_() + beta_()

where

- fatmass:

- weight:
- height_hip_ratio

4.2.6 Model Justification

4.2.7 Assumptions and limitations

Assumptions: - this model assumes that there is an obvious difference in body measurements and fat storing capacities in males and females, hence gender is an important predictor and also affects the proxy values of fat percentage owing to the calculation based on WHR.

Limitations:

- The dataset does not include the age of the particapants. while a general age range is provided, such as 18-X years old, different age groups could differ in how they are categorised.
- This model would not be able to differentiate between age groups and hence would work poorly if this information is included.
- This model is also built on data that was generated by the ABS [] simulator so it may be inherently working with an error.

4.2.8 Model validation:

Model out-of-sample testing:

The dataset provided by BodyM already had a trainining and testing split. I used the Test A to test the accuracy of the model. As the method of collection was similar to the training data, the potential errors in the model could only arise from the fact that it is unseen data.

More elaboare explanation, tables and figures can be found at [].

RSME: I also computed the RSME to evaluate the model's performance as a lower RSME indicates accuracy in predicted vs actual fat percentage categories.

5 Results

5.1 Model Results for Measurements Model

Our results are summarized in Table 2. Our model's results and interpretation:

Table 2: TODO

Table 2: Summary of Measurements Model

	(Intercept)	height	gender	waist_hip_ratio	height_hip_ratio
Moderate	34.34511	0.1632657	-2.997520	18.21029	-37.45804
High	60.03388	0.3063854	-5.101766	34.97975	-74.26296
Extreme	78.18906	0.4569892	-6.262725	54.36190	-113.34181

5.2 Model Results for Body Mass Model

Our results are summarized in Table 3. Our model's results and interpretation:

Table 3: TODO

Table 3: Summary of Measurements Model

	(Intercept)	height	gender	fm
Moderate	-15.87179	0.1811493	-27.88642	1.946447
High	-12.83259	0.0020248	-63.42437	4.850373
Extreme	-14.02425	-0.1279283	-85.85319	6.537356

6 Discussion

6.1 First discussion point

- height is not all, personality is more important ## Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

6.2 Second discussion point

6.3 Third discussion point

6.4 Weaknesses and next steps

- the BodyM dataset does not include age. An older person has lower metabolism [] and a higher chance of collecting fat on their abdomen.

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

we compare the posterior with the prior. This shows...

References

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.