Explore the factors that may be associated with obesity

Xueyan Hu

2024-04-21

# 1. Summary/Abstract

Obesity is a chronic complex disease which may increase the risk of severe health issues including diabetes, cardiovascular disease and even cancers. And the condition of obesity in real life could be connected to individual physical behaviors, such as physical activity frequency, alcohol consumption, age, gender and inherit. The project is to explore the potential factors that might be related to obesity level using machine learning models to a dataset collected from an investigation done in South American countries including Colombia, Peru and Mexico. The data analysis result has shown that age, gender, family history, drinking water, physical activity, and commute transportation are all predictors for obesity level, which is represented by individual BMI. As for the outcome, when linear model is fit to data which has BMI as continuous variable, the performance is better than using obesity level as categorical variable.

# 2. Introduction

## 2.1 General Background Information

[Obesity](https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight) is a chronic complex disease defined by excessive fat deposits that can impair health. Overweight and obesity are predominately a result of a sustained positive energy balance, stemming from a combination of excess dietary energy intake (mainly due to poor eating habits) and reduced energy expenditure (due to lack of physical activity and prolonged sedentary activities). It is a significant risk factor for and contributor to increased morbidity and mortality, mainly by its increasing risk of type 2 diabetes and heart disease, as well as certain types of cancer and chronic diseases, including osteoarthritis, liver and kidney disease, sleep apnea, and depression (Pi-Sunyer, 2009). Also it does affct bone health and reproduction.

In the recent decades, obesity has become a more and more common and severe problem across the world with worldwide obesity rates tripling since 1975, thus reaching pandemic proportions (Moschonis & Trakman, 2023). Worldwide adult obesity has more than doubled since 1990, and adolescent obesity has quadrupled. In 2022, 1 in 8 people in the world were living with obesity. In the US. the situation seems similar with the worldwide trend. In 2022, 22 states had at least 35% of adults with obesity, up from 19 states in 2021. Ten years ago, CDC said, no state had an adult obesity prevalence at or above 35%.

Obesity level is basically defined by Body Mass Index ([BMI](https://www.cdc.gov/obesity/basics/adult-defining.html)). BMI is a person’s weight in kilograms divided by the square of height in meters. A high BMI can indicate high body fatness. BMI less than 18.5 is defined as underweight; BMI range between 18.5-24.9 falls within the healthy weight range. If BMI is from 25 to 29.9, then the person is regarded as overweight. And when someone’s BMI is higher than 30, he/she is at the obesity range.

There is no doubt that obesity is a multifarious disease due to cryogenics environments, psycho-social factors and genetic variants in most cases, but many factors can contribute to excess weight gain including eating patterns, physical activity levels, and sleep routines. Social determinants of health, genetics, and taking certain medications also play a role.

There is a lot of discussion on the Internet and social media about fitness and weight loss, and people assume that physical activity can lose weight as a fact. Also KOLs usually encourage people to drink more water for higher metabolic levels. But I have not seen a clear data to prove it. At the same time, the effect of physical activity to lose weight is also different from person to person.

In addition, according to what I learned from nutrition course, alcohol consumption may be related to weight gain because alcohol metabolism and carbohydrate metabolism both go through the same pathway/cycle in the middle and late stages (Cederbaum, 2012). Ethanol is converted into acetaldehyde by the action of ethanol dehydrogenase, and the acetic acid produced by acetaldehyde dehydrogenase eventually enters the tricarboxylic acid cycle to produce energy, which is consistent with process of the consumption of carbohydrates for energy. If someone is lacking of oxidative enzymes, ethanol is nonoxidatively metabolized by two pathways. A reaction catalyzed by the enzyme fatty acid ethyl ester (FAEE) synthase leads to the formation of molecules known as FAEEs. A reaction with the enzyme phospholipase D (PLD) results in the formation of a phospholipid known as phosphatidyl ethanol (Zakhari, 2006). No matter how calalystic reactions work, finally alcohol will be converted to the reagents for tricarboxylic acid cycle and then energy is released.

## 2.2 Description of data and data source

The data is obtained from a website [UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition) which have collected the survey results of individuals from the countries of Mexico, Peru and Colombia on their eating habits and physical condition, as well as their estimation of obesity levels. The original dataset contains 17 variables and 2111 observations, labeled with the class variable Obesity Level which has 7 categories.

## 2.3 Questions/Hypotheses to be addressed

I would like to figure out if higher frequency of physical activity and drinking more water could keep people away from obesity and whether heredity is an important factor for obesity. Also, I will focus on the relationship between alcohol consumption frequency and obesity level.

# 3. Methods

Data cleaning including generating necessary variable, renaming categories and variables and so on.

First to take a deeper look of the data, I would like to do some descriptive analysis including bar chart, boxplot, scatterplot etc. to indicate the distribution of the data.

Then simple linear fit will be applied to explore the relationship between variables and outcomes (both categorical and numeric), and proper predictors will be pick out from the variables. And then machine learning models will be chosen and fitted to the dataset wieh each performance metrics will be determined and compared for reporting the “best” model among the choices based on reality.

## 3.1 Schematic of workflow

## 3.2 Data acquisition

The dataset presents the results of a survey on the topic of obesity including eating habits and personal physical conditions and answers to behavioral questions.

## 3.3 Data import and cleaning

The detailed raw data and processed data are stored in ‘data’ folder. And the code for data cleaning is in the ‘processing-code’ subfolder under ‘R’ folder.

I made a code book based on the article reporting this dataset (Palechor & Hoz Manotas, 2019) first before doing any cleaning.

Obesity level is a categorical variable defined by BMI range, so BMI as a new variable is created and it could be analyzed as a numeric variable. Also, to make the variable name easy to understand, I changed some abbreviated names to ones with fully spelled key words. Male is replaced by 1 and female by 2 under Gender variable for more convenient further analysis.

## 3.4 Statistical analysis

Bar chart, box plot will show the distribution of obesity level by different variables, as well as the population features of the people under investigation. All figures will be saved in ‘figure’ subfolder under ‘result’ folder and shown in the manuscript.

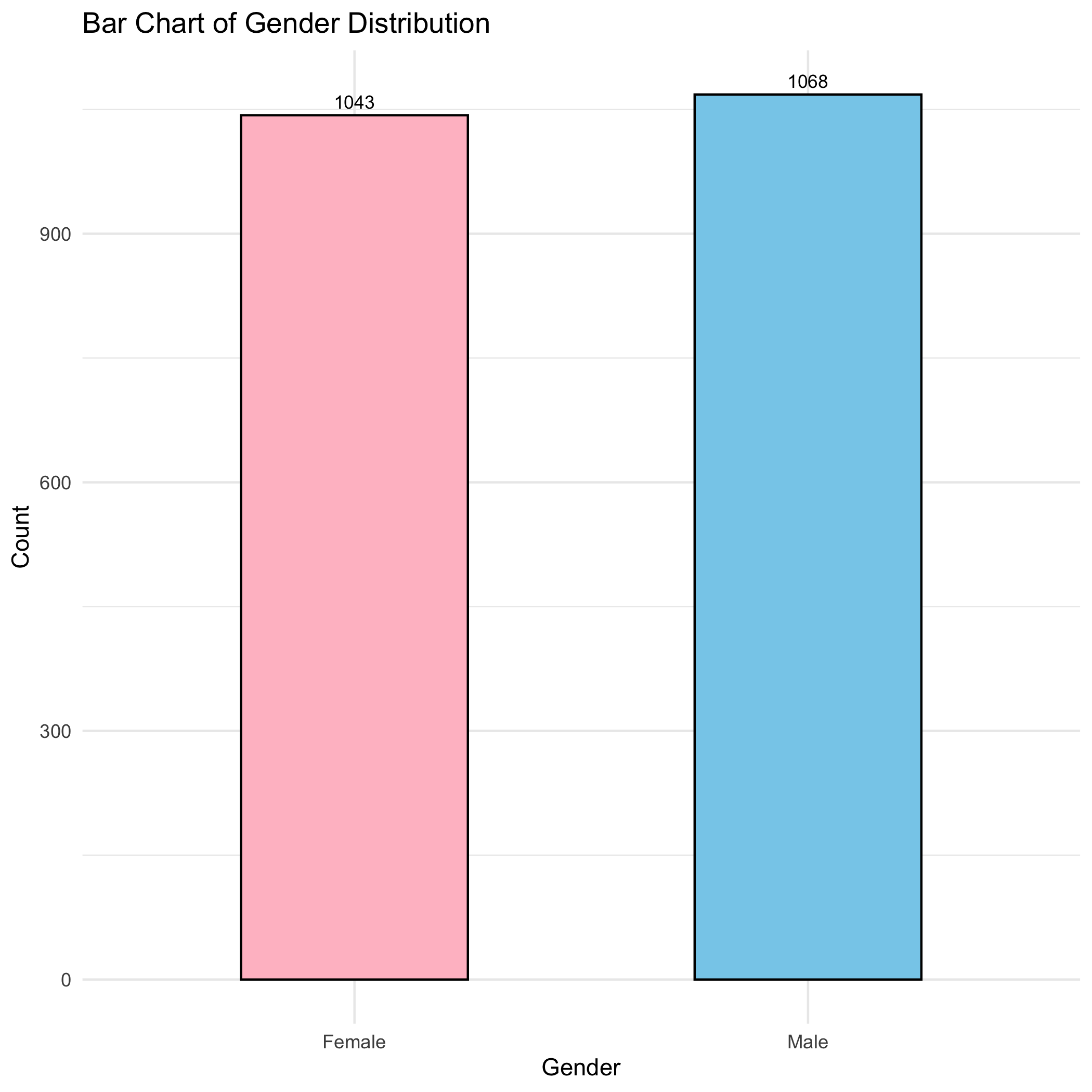
By fitting the data with simple linear model, the result will show whether each independent variable has main effect to the dependent variable and the possible interactions between variables will also be detected. And the summary table is stored in ‘table’ subfolder under ‘result’ folder.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

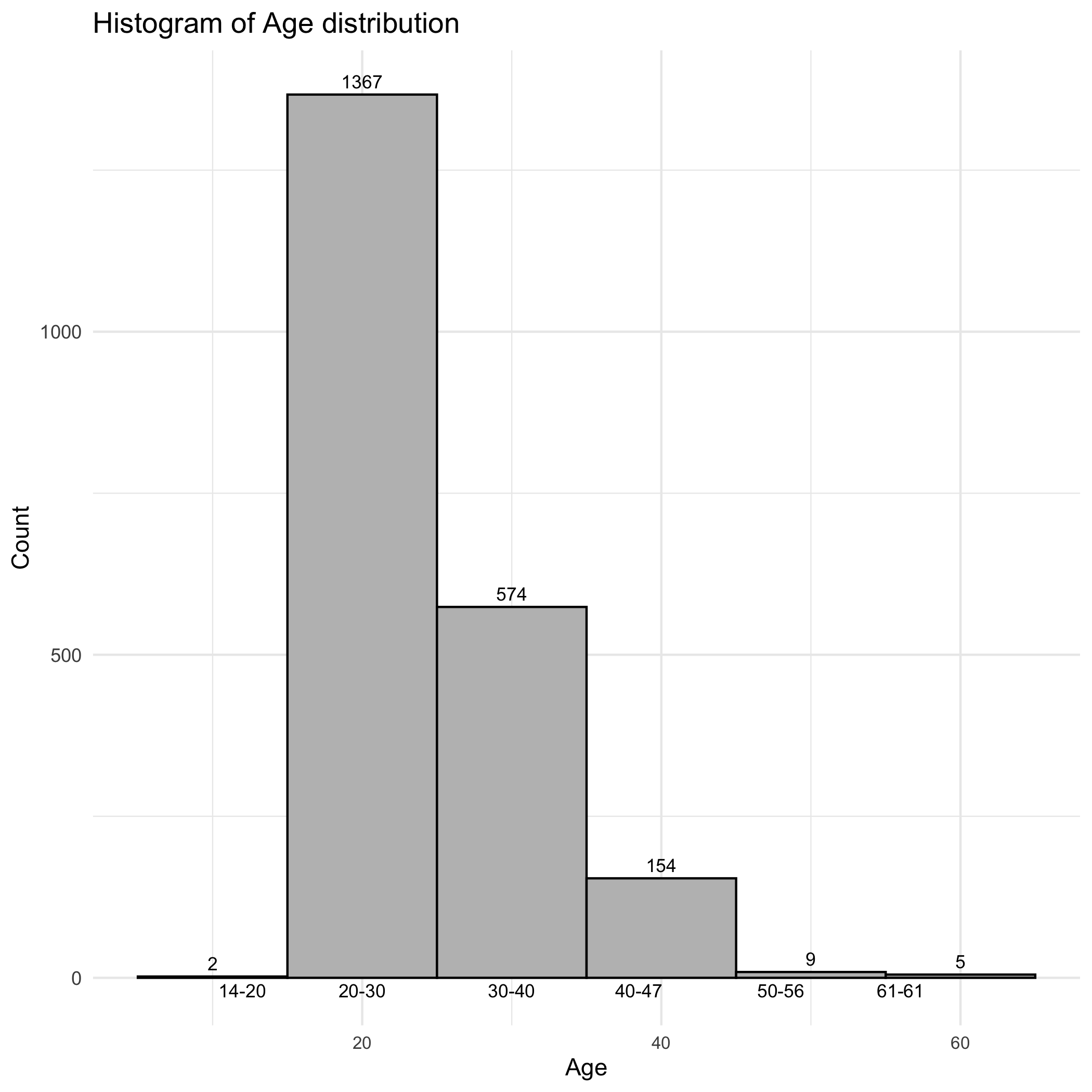
Bar chart shows that the ratio of male and female populaiton is quite close to 1:1, which means genders of the cohort under investigation are balanced.

Figure 1. Gender distribution.



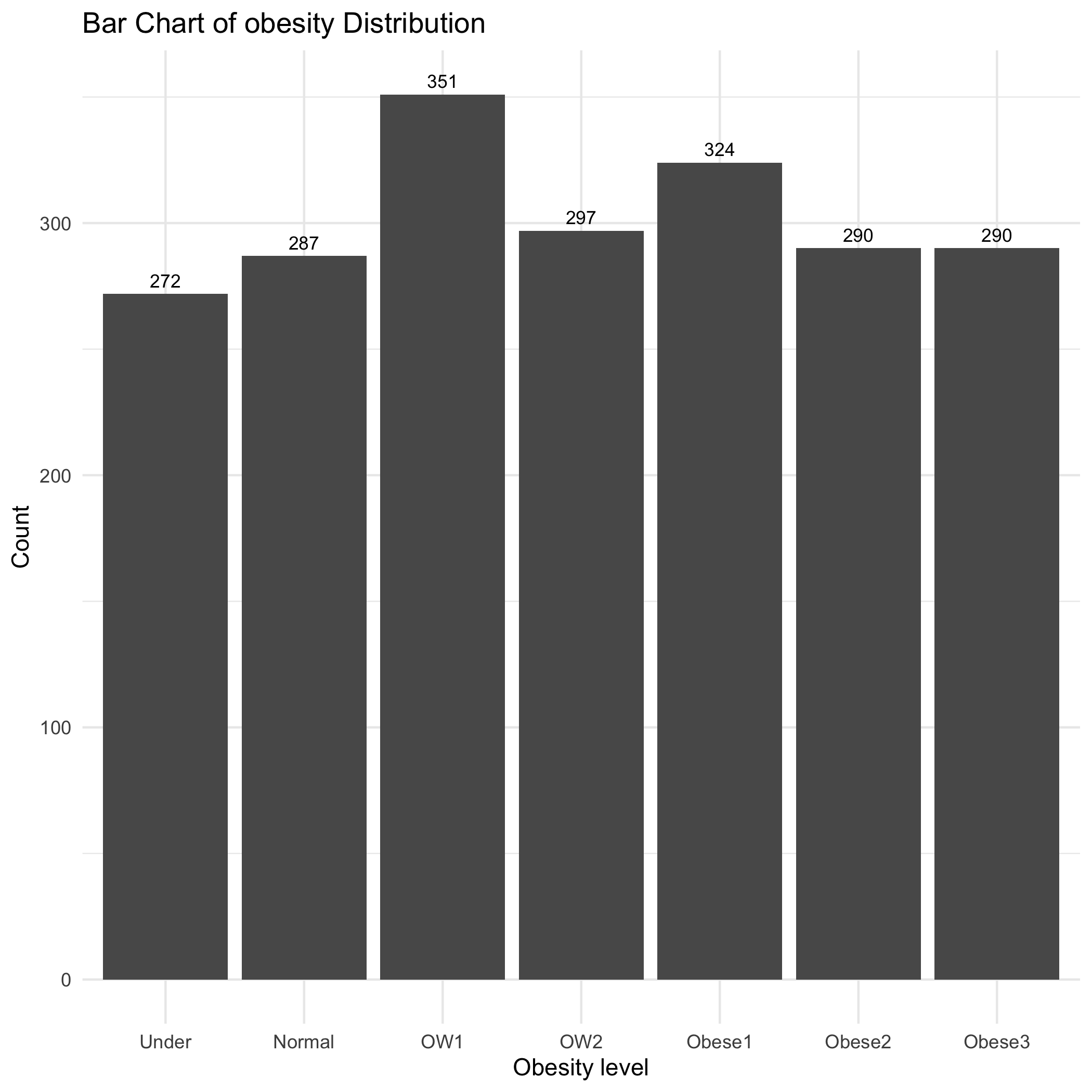
The age distribution figure shows that people aged 20 to 30 account for the largest proportion of the total population which is nearly two thirds of the total population. Then the second highest ranking is 30-40, occupying about a quarter of the total population.

Figure 2. Age distribution



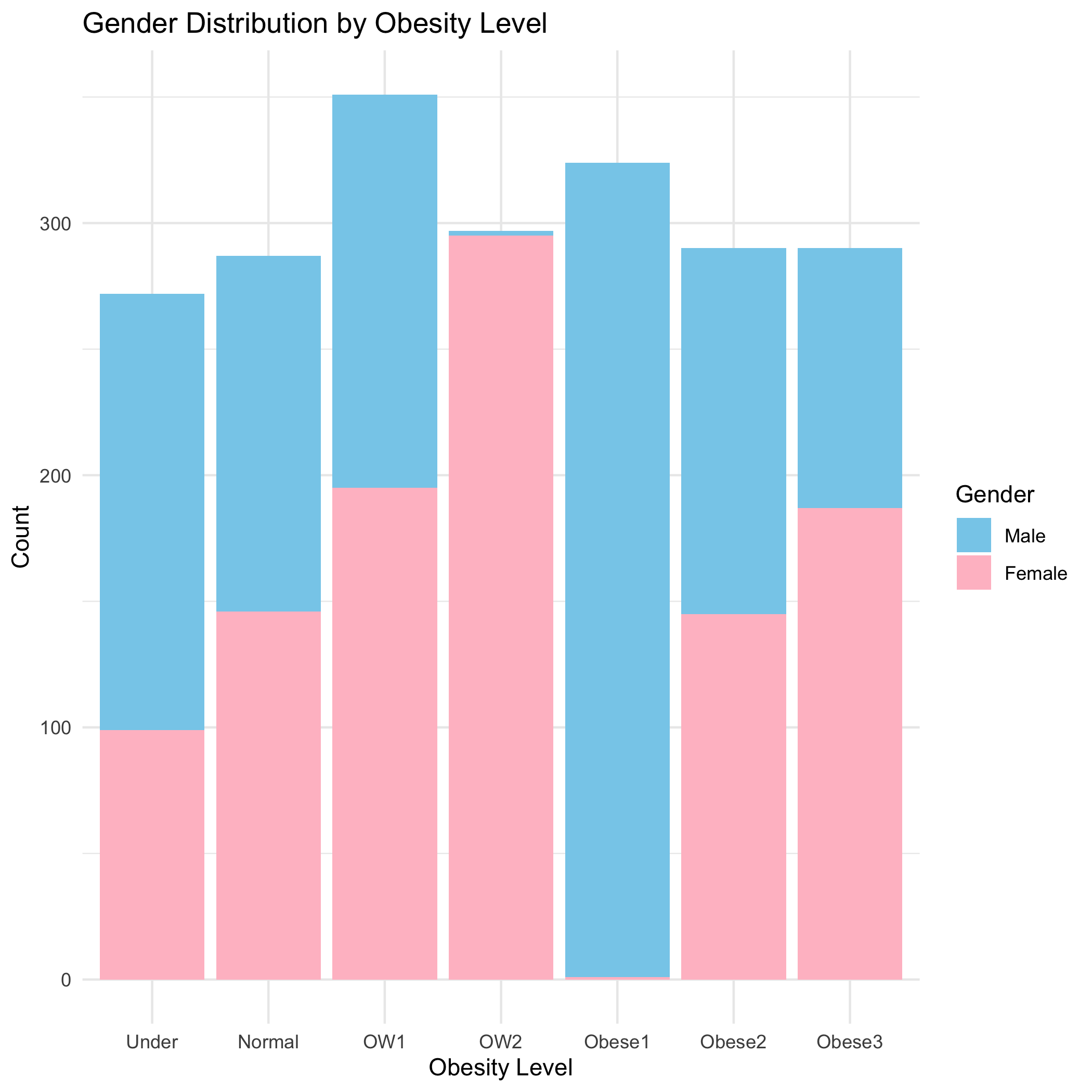
It can be seen that almost three fourths of the people are identified with over weight. The highest population fall into over weight level 1, while the under weight level has the lowest population.

Figure 3. Obesity level distribution.



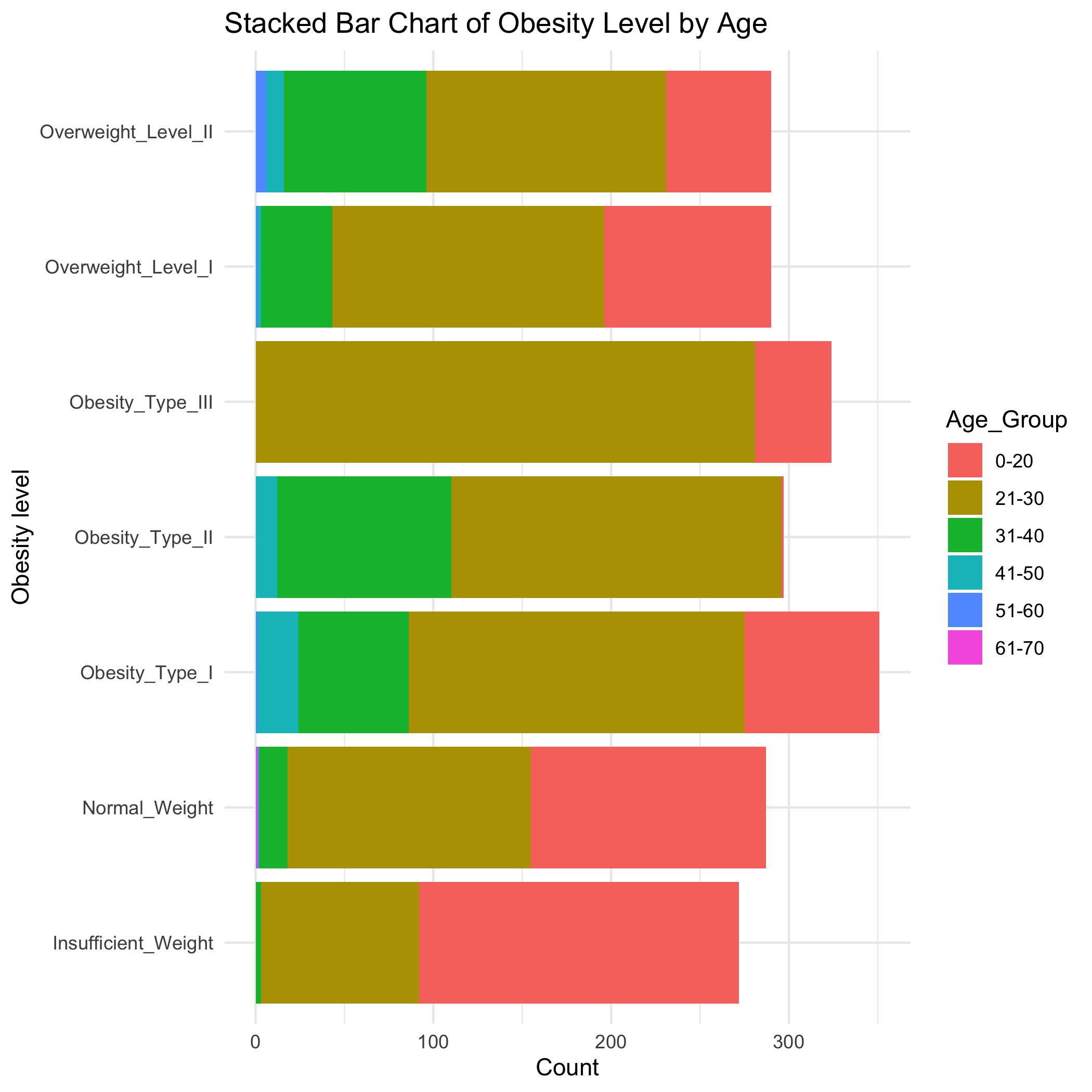
The gender ratio is close to 1:1 but its distribution under each obesity level is different. There is no obvious pattern of obesity levels by gender. The ratios of male and female in other obesity levels are close to 1:1 except for overweight level 2 that women occupy almost all of the population and Obese level 1, most of people under which are male. In addition, around two thirds of underweight people are men.

Figure 4. Gender distribution in each obesity levels.



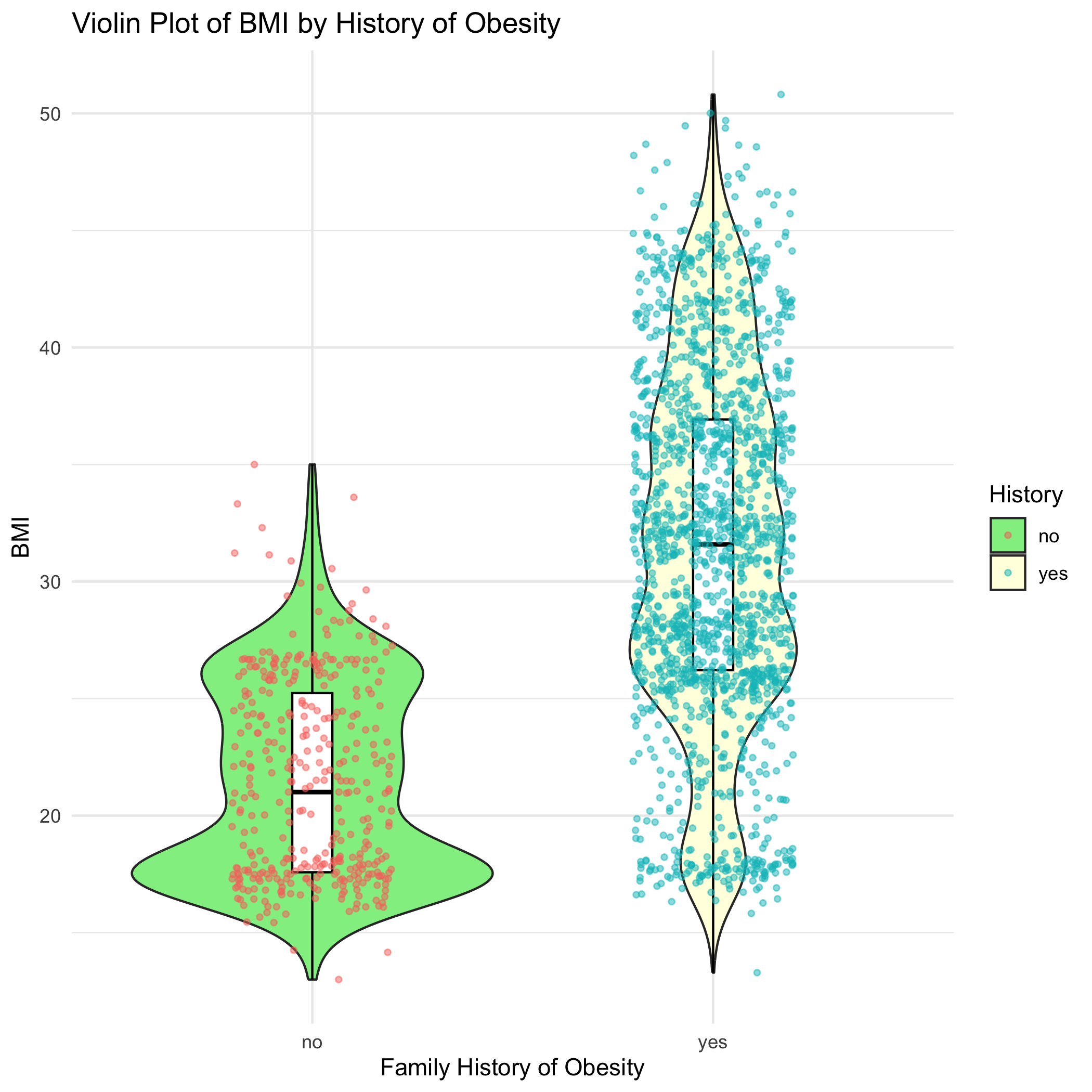
Since people aged between 21 and 30 occupy the largest part of the total population, I could expect that in each level, most of people would fall in the age group of 21-30. And it can be seen from this bar chart. Moreover, more than half of underweight people are from the age group 0-20, it might reflect the problem that health condition and nutrition intake should be regarded seriously by the countries that attend the survey.

Figure 5. Age distribution in each obesity levels.



It is apparent that people with family obesity history may tend to be overweight according to figure 6 below. The median BMI is much higher for people who have family obesity history than the people who don’t. And the data density of no family history group is aggregated at the level under 30. There is an article talking about family obesity history and its impact on children in the family. The article finds childhood obesity is positively correlated to the obesity status of their parents (Moschonis & Trakman, 2023). When both parents were obese, approximately half of the children in Europe were overweight. This highlights the presence of an “obesogenic environment” typically fostered by obese parents, impacting not only themselves but also their children. This environment often involves reduced physical activity, increased sedentary behavior such as screen time, and higher intake of calorie-dense foods, contributing to weight gain across the family. Therefore, family history is not only a genetic effect, lifestyle, environment also plays an important role for offspring obesity.

Figure 6. Violin plot of BMI by family obesity history.



## 4.2 Basic statistical analysis

Since some obvious relationship could be shown in figures, I choose other variables as predictors in exploratory analysis.

First, I was curious that whether there is a negative correlation between transportation type and physical activity frequency, I converted transportation types into numbers indicating the energy consuming levels from 1 (lowest automobile) to 4 (highest walking). Then I used correlation function to calculate the coefficient for these two variable but it was pretty low (0.0059).

Then, I would like to fit linear model for BMI by both water consumption and physical activity. And the result indicates that there is a connection effect between them two, which means that both of them contribution to obesity control.

Also, I use single indicator, alcohol consumption and BMI as an outcome to fit linear model. It shows that if an individual consumes alcohol sometimes than never, his BMI increase 3.98. But the correlations are all negative for level 3 and 4, which means consuming more alcohol may decrease BMI.

Because I still wonder if obesity could be related to gender or age. So I fit simple linear model to the dataset using gender and age as independent variable. The result indicates the interactions between them. It is a factor out of my expectation. And I decide to include this interaction for next machine learning models.

In the last model I select all the variables that I think may have effect on BMI as predictors and BMI as outcome.

Since I think it is quite obvious that family history is an important factor for obesity, and also I have mentioned some reasons how family obesity could influence the kids involved, I choose to skip it for this basic exploratory analysis here. But I will include it in the following part.

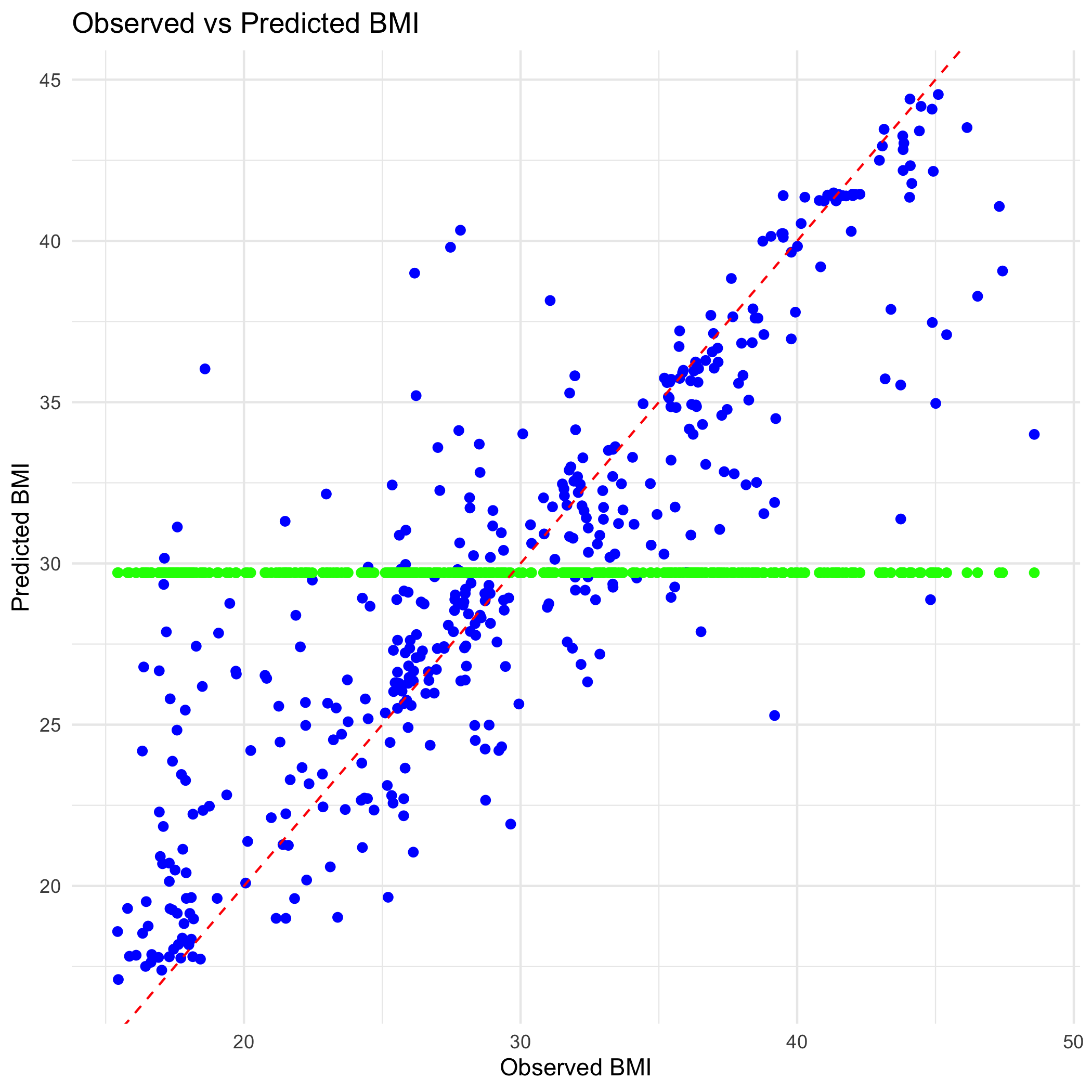
## 4.3 Full analysis

| Model | Predictor | Metric Value | Metric Value on testing data |
| --- | --- | --- | --- |
| Linear null | - | RMSE=8.00 | RMSE=8.04 |
| Linear regression 1 | Age, History, Alcohol, MTRANS | RMSE=0.62 | RMSE=6.61 |
| Linear regression 2 | All | RMSE=6.30 | RMSE=3.99 |
| Linear regression 2 + LASSO model | All | RMSE=7.38  l=0.0011 | RMSE=7.38 |
| Random forest | Gender, Age, History, Water, Alcohol, FAF | RMSE=3.22 | RMSE=3.97 |
| Logistic null | - | Accuracy=0.17 | Accuracy=0.15 |
| Discriminant analysis | All | Accuracy=0.50 | Accuracy=0.48 |
| Multinominal logistic regression | All | Accuracy=0.58 | Accuracy=0.52 |

The above table has listed all of the the models I have chosen to fit the data together with the null models.

When I use BMI as a continuous outcome, linear regression models can be applied to the data. Compared to the linear null model, I think all of the ML models perform better according to RMSE value. Since I think it is reasonable that physical activity is a potential factor that may influence the obesity level, although linear model with all variables and interactions has higher RMSE, I think regression model 2 that picks all of the predictors is better. Then since random forest model selects all predictors and RMSE value is relatively lower among all the linear models, RF model is the choice for linear model.

The figure below shows the prediction condition using random forest model together with the null model on testing data. Both sides of the 45 degree line have outliners and the scatters are not clustered close enough to this line which indicates the prediction is not really well.



As for obesity level as a categorical outcome, logistic models are fit to the data. And I use accuracy as the performance metric for this evaluation. First, both the models I use have better performance than null model. And according to the accuracy, multinominal logistic regression performs better.

Because I have chosen random forest model between linear regression models and multinominal logistic regression between logistic models, I use test data to assess the quality of these two models separately. And the new RMSE on test data for random forest model is 3.97, and it is 3.22 when the model is trained and cross validation is applied. And the new Accuracy is 0.52 when it is used on test data, also lower than 0.58 when the model is trained. The metric value difference is a sign that both of the models are overfitting.

So I try to use test data to assess all of the models in the table above and collect new metric values. For linear regression 1, the RMSE on testing data is 6.61, which is much higher than that on training data. So this model is not doing prediction well. The new RMSE for linear regression 2 is 3.99 lower than 6.30 on training data, which is a sign of good prediction. Besides, LASSO model has the same RMSE on both training data and testing data which is 7.38. It is also a reasonable result for good prediction. Both logistic models have lower accuracy than on test data than training data. Based on all the result of the machine learning I have used on dataset, I think linear regression model with LASSO regularization is a good choice among them.

# 5. Discussion

## 5.1 Summary and Interpretation

I chose a dataset that summarize an investigation result done in South American countries about individual physical behaviors and the classification of obesity level based on the information of height and weight obtained from the investigation. I had some questions to explore and did data analysis on this dataset.

I found simple linear model fitting can answer my questions, which were that alcohol consumption was correlated to obesity, as well as there were interactions between physical activity frequency and water drinking amount. In addition, I found there was nearly no correlation between transportation choice and physical activity, which could answer another question I had. At the same time, outside the scope of the problem, I also discovered the interaction between gender and age in this process.

For machine learning model fitting, I found the best predictors for BMI index based on the value of some performance metrics after trying both linear regression and classification model, which were gender, age, family history, water drinking amount, frequency of physical activity, alcohol consumption frequency and commute transporation. Although the machine learning models that I used for fitting the dataset are different types and their metrics and not be compared directly, the assessment of model quality using testing data could give a clue for the model selection. Considering all the results of performance metrics, I finally chose liner regression model with LASSO regularization with all predictors plus BMI as the outcome. The reason was that RMSE calculated of predicted MBI values on testing data split from cleaned dataset was as same as the result when it was trained. Also, regularization was applied by LASSO model to reduce overfitting at the same time.

## 5.2 Strengths and Limitations

I think my analysis has include two different types of model according to the type of outcome and each type include 2 models, which can provide more information and the result can be shown in different aspects as well as more choices were provided. Also, interactions between predictors were found and included in the model fitting process. Both predictors with and without interactions were all fit and compared with each other. And LASSO model was used for decreasing overfitting.

I did tried to do tuning for random forest model but I failed for running it successfully, so this could be a limitation that I have. And maybe there could be more metrics that I can also calculate but I only chose 1 for each model. Furthermore, chatGPT told me that even if testing result has a sign for well predicting, there still could be other factors for further assessment such as F1-score and so on. But base on my understanding, I don’t know what other methods I could choose from so I just stopped here.

## 5.3 Conclusions

The take-home message for this project is that obesity is correlated to individual physical behaviors. Based on the result of data analysis, age, gender, family history, drinking water, physical activity, and commute transportation are all predictors for obesity level which is represented by BMI index.

# 6. References

Cederbaum, A. (2012). ALCOHOL METABOLISM. *NIH Public Access*, *16*(4), 667685. <https://doi.org/10.1016/j.cld.2012.08.002>

Moschonis, G., & Trakman, G. L. (2023). Overweight and obesity: The interplay of eating habits and physical activity. *Nutrients*, *15*(13). <https://doi.org/10.3390/nu15132896>

Palechor, F. M., & Hoz Manotas, A. de la. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico. *Data in Brief*, *25*, 104344. https://doi.org/<https://doi.org/10.1016/j.dib.2019.104344>

Pi-Sunyer, X. (2009). The Medical Risks of Obesity. *Postgraduate Medicine*, *121*(6), 21–33. <https://doi.org/10.3810/pgm.2009.11.2074>

Zakhari, S. (2006). Overview: How is alcohol metabolized by the body? *Pubmed Central*, *29*(4), 245–255. <https://doi.org/PMCID: PMC6527027>