Explore the factors that may be associated with obesity

Xueyan Hu

2024-03-28

# 1. Summary/Abstract

In this project, I have found……

# 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. Obesity can lead to increased risk of type 2 diabetes and heart disease, it can affect bone health and reproduction, it increases the risk of certain cancers.

In the recent decades, obesity has become a more and more common and severe problem across the world.

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.

There is no doubt that obesity is a multifactorial disease due to obesogenic environments, psycho-social factors and genetic variants in most cases, but personal behavior could also has correlation with the result of overweight and obesity.

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 ususally 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](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6527027/) and carbohydrate metabolism both go through the same pathway/cycle in the middle and late stages.

## 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.

## 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).

Prediction will be made.

## 3.1 Schematic of workflow

## 3.2 Data acquisition

The dataset presents the results of a survey on the topic of obesity including 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.

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.

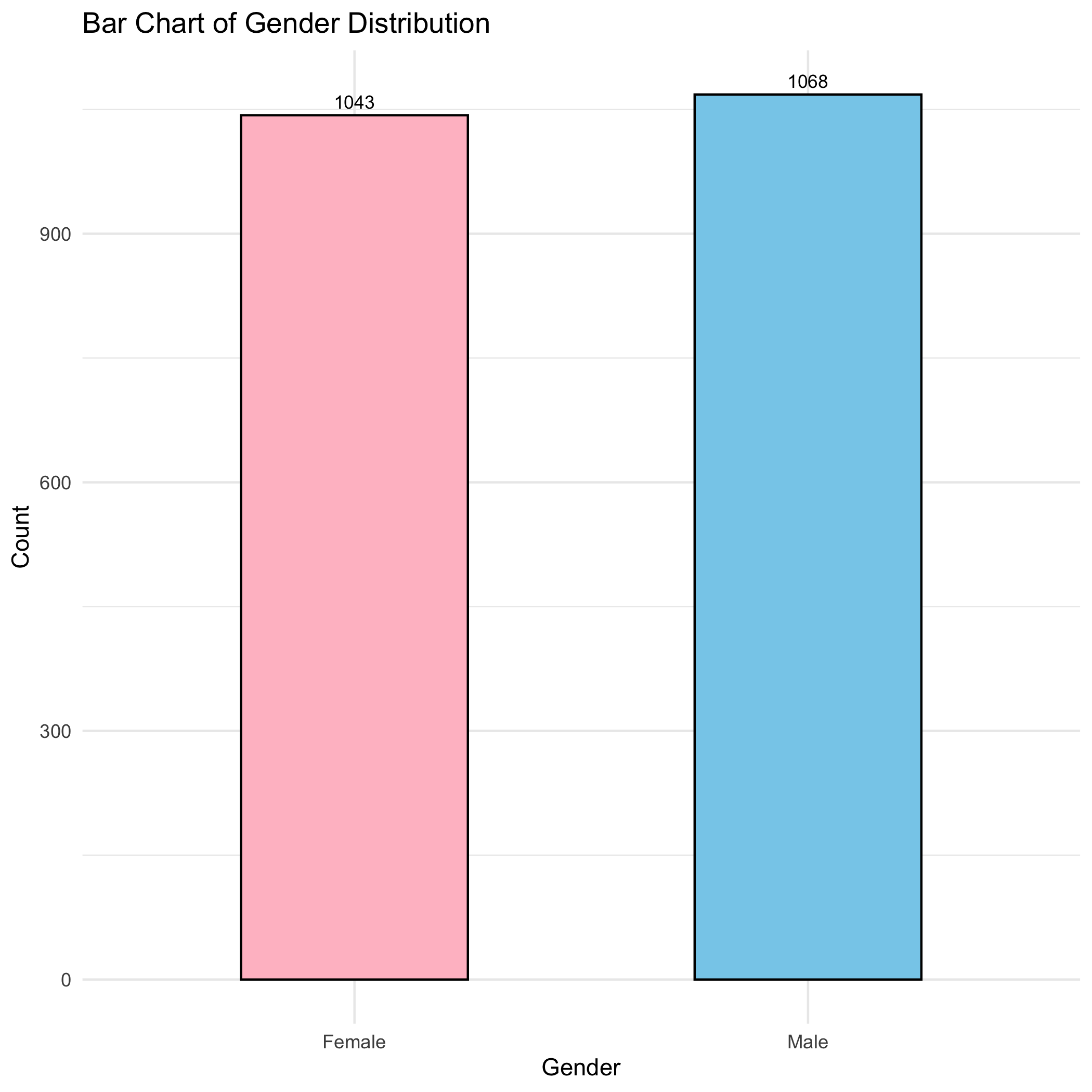
By fitting the data with simple linear model, the result will show whether each independent variable has main effect to the dependent variable through the P value. And the summary table is stored in ‘table’ subfolder under ‘result’ folder.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

Bar chart shows that the genders of the cohort under investigation are balanced.

Figure 1. Gender distribution.



It can be seen that almost two thirds of the people are identified with over weight.

Figure 2. Age distribution

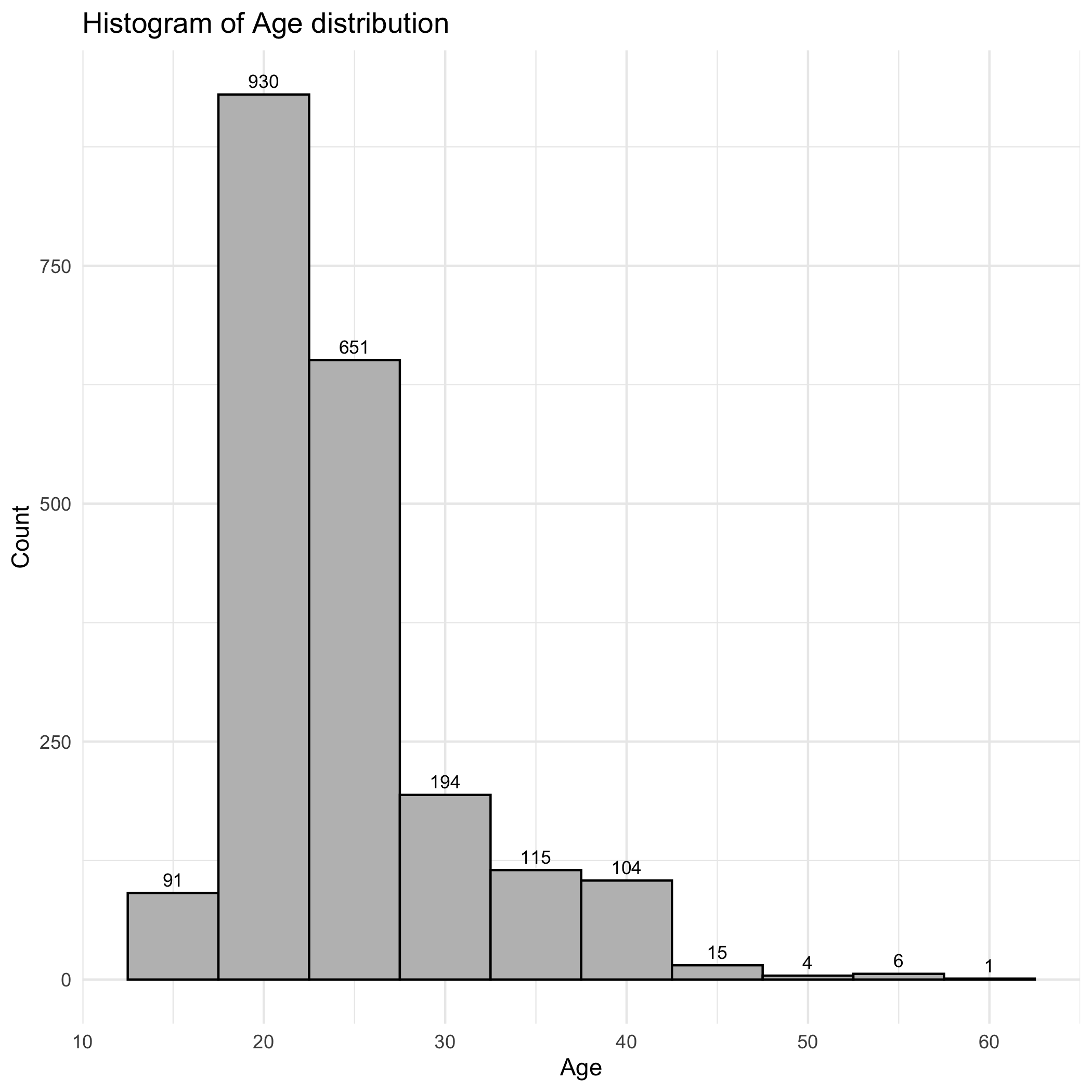
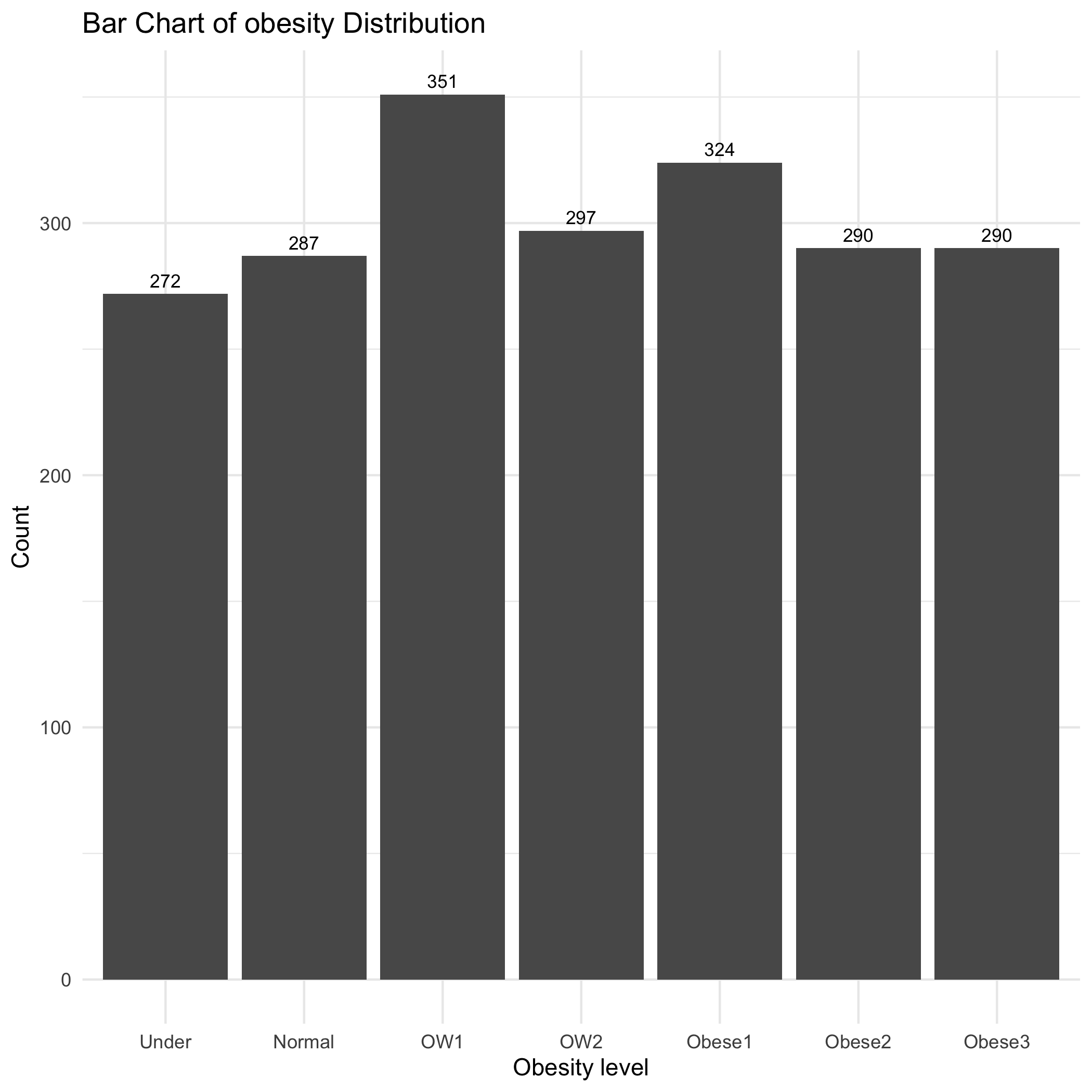
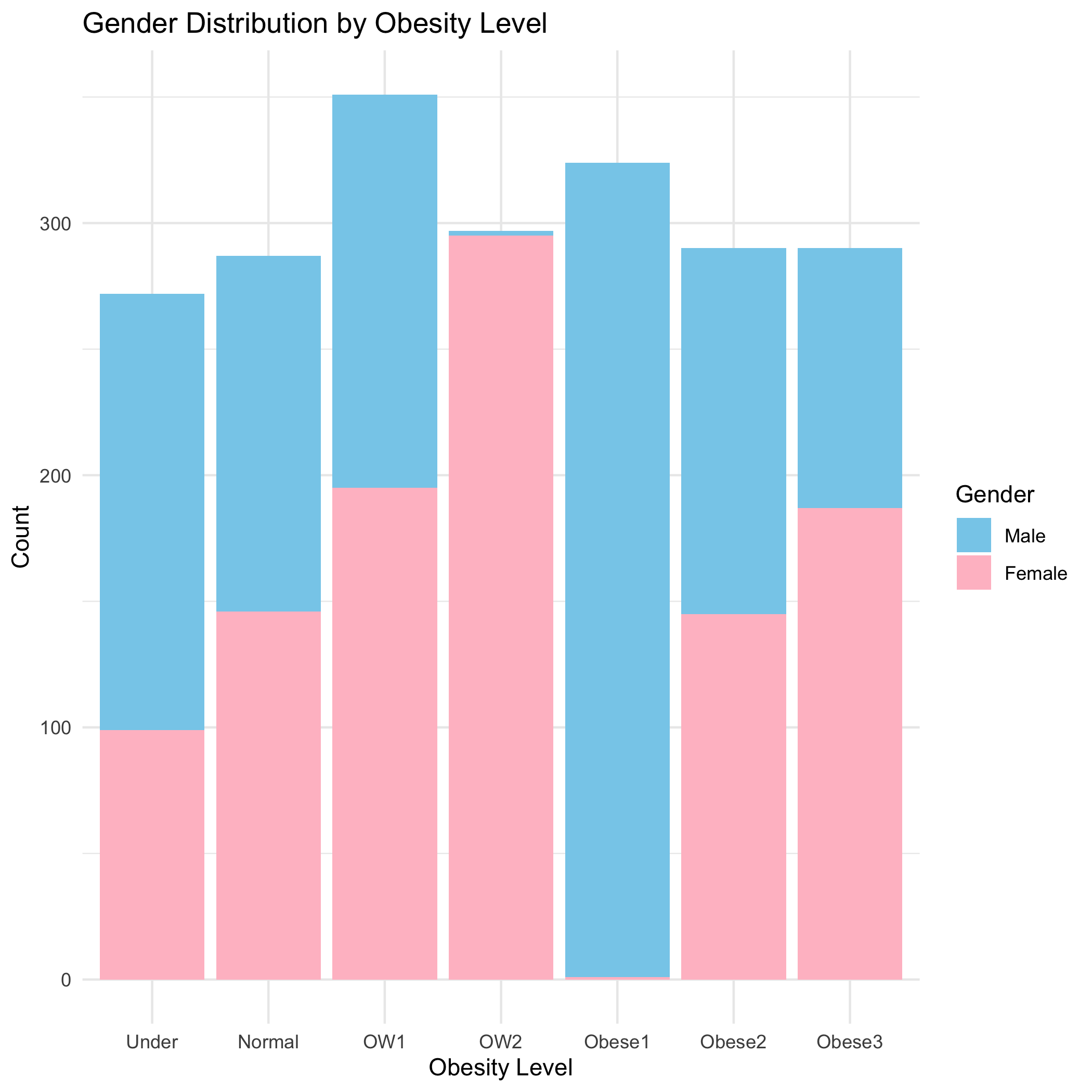


Figure 3. Obesity level distribution.



There is no obvious pattern of obesity levels by gender.

Figure 4. Gender distribution in each obesity levels.



It is apparent that people with family obesity history may tend to be over weight according to figure 4 below.

Figure 5. Age distribution in each obesity levels.

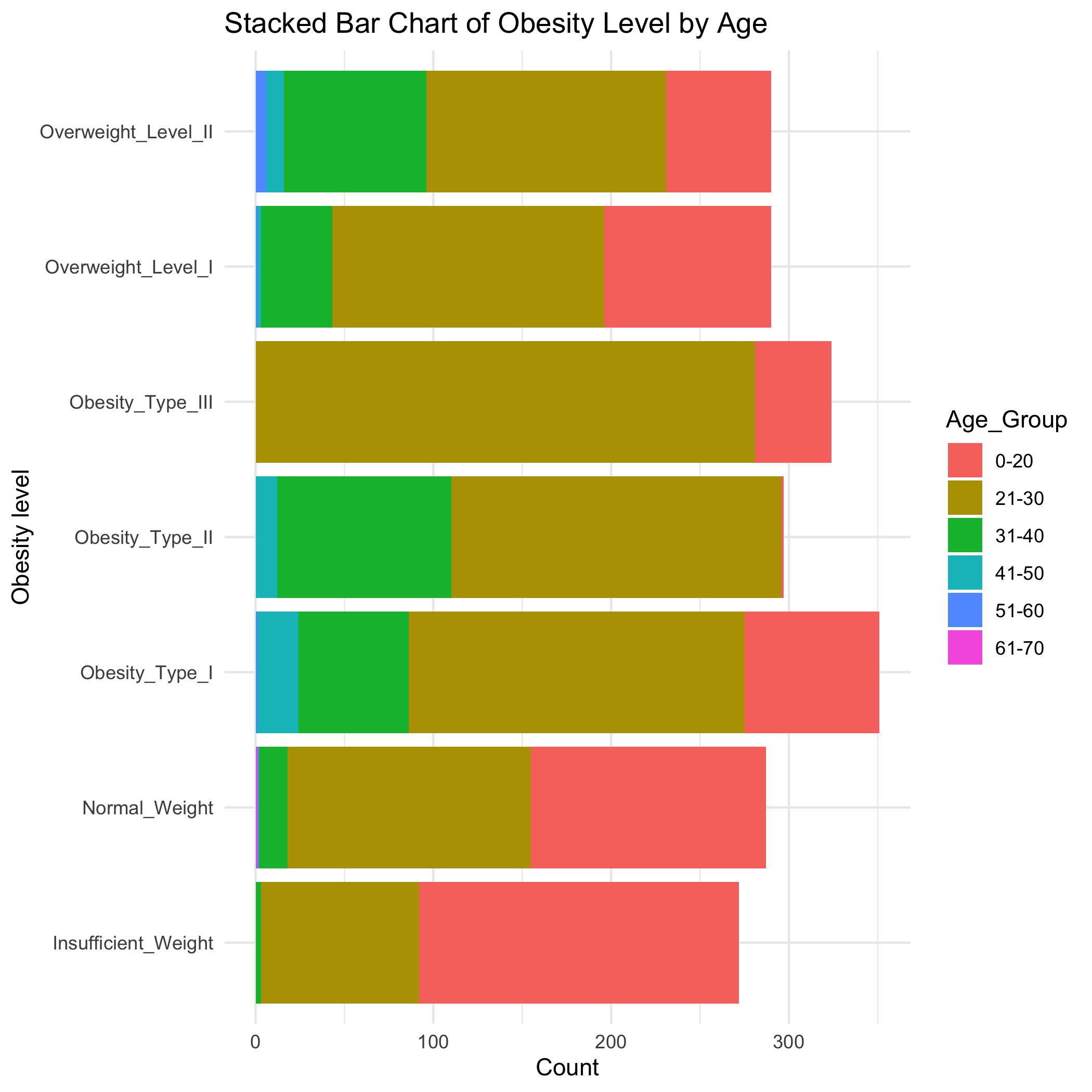
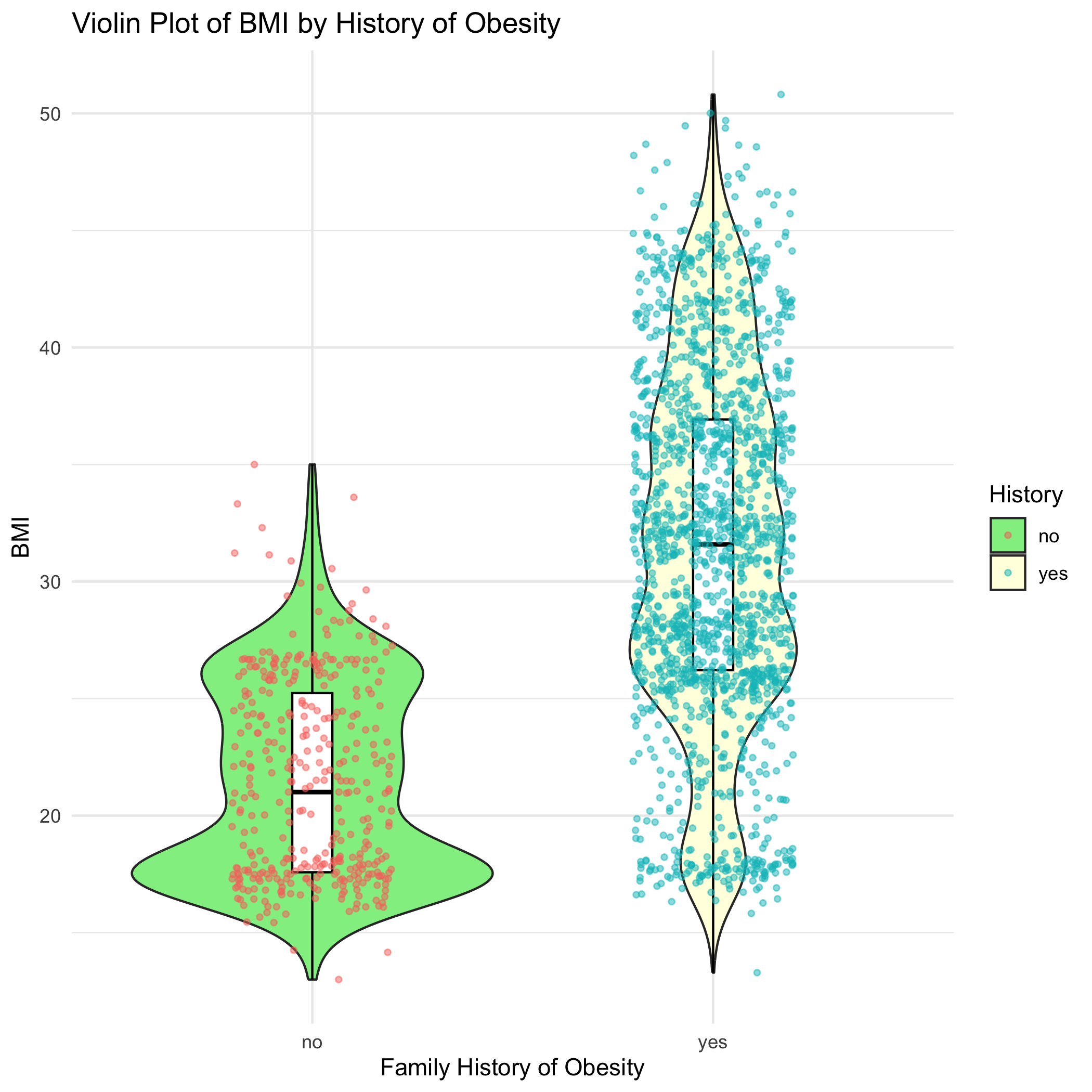


Figure 6. Violin plot of BMI by family obesity history.



## 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 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.

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

## 4.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

# 5. Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

**Haven’t updated yet but will do it later!!!**

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (Leek & Peng, 2015) discusses types of analyses.

These papers (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, Shen, & Handel, 2020) are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like, I just used the generic word references.bib but giving it a more descriptive name is probably better.

# 6. References

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*(6228), 1314–1315. <https://doi.org/10.1126/science.aaa6146>

McKay, B., Ebell, M., Billings, W. Z., Dale, A. P., Shen, Y., & Handel, A. (2020). Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery. *Open Forum Infectious Diseases*, *7*(11), ofaa494. <https://doi.org/10.1093/ofid/ofaa494>

McKay, B., Ebell, M., Dale, A. P., Shen, Y., & Handel, A. (2020). Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients. *Proceedings. Biological Sciences*, *287*(1927), 20200496. <https://doi.org/10.1098/rspb.2020.0496>