Table of Contents

[Executive Summary 2](#_Toc178152595)

[Introduction 2](#_Toc178152596)

[Justification for the Case Study 3](#_Toc178152597)

[Data Importation and Preparation 3](#_Toc178152598)

[Exploratory Data Visualization 4](#_Toc178152599)

[Model Training and Testing 6](#_Toc178152600)

[Conclusion 7](#_Toc178152601)

[References 8](#_Toc178152602)

# Executive Summary

Data analysis is increasingly becoming important in the society and being used in various fields such as in decision making, making personalized recommendations, ensuring operational efficiency, helping organizations gain a competitive advantage and in innovation and development. This project aimed to leverage on this fast-growing field to estimate obesity levels of people in Mexico, Peru and Colombia based on their eating habits, physical conditions and other features which are to be discussed in the coming sections. The analysis focused of utilizing the R programming language particularly the ggplot library to gain an understanding of the data through exploratory data analysis and visualizations. The results of the analysis showed that

# Introduction

Obesity is not just a cosmetic concern affecting millions of people across the globe but a serious medical health threat that affects people of different ages and genders. According to the world health organization (2022) 1 in every 8 people suffers from obesity with the number of adults with the condition doubling since 1990. Research has been conducted on the issue of obesity previously with studies showing that the is a significant correlation between the diets taken by people, frequency of the diets, Caloric composition of the consumed foods, sugar consumption, healthcare plans among others (Hwalla & Jaafar, 2020; Safaei et al., 2021; Wiley, Wakefield and Silver, 2020). The conclusions as to what causes obesity and people becoming overweight is thus a widely controversial issue which necessitates further studies and utilization of data analytics models to uncover. It is also important to consider that the levels of obesity are different and caused by varying variables in different regions therefore making the results obtained in this analysis project to be tied to the three countries where the data was collected in Mexico, Peru, and Colombia.

# Justification for the Case Study

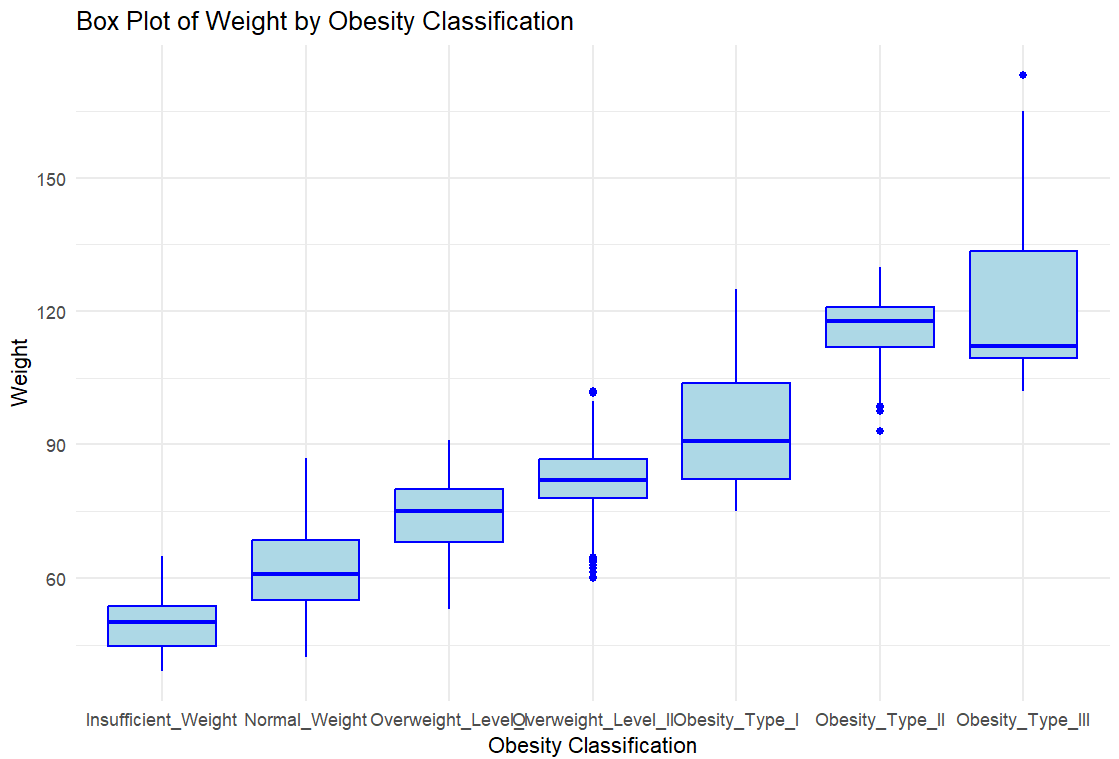
The rationale for selecting an obesity dataset in the three South American countries was justified by several reasons. Studies have consistently showed that obesity is becoming a major issue in Latin America with it being linked to various non communicable diseases such as heart attack, various forms of cancers, diabetes, high cholesterol, sleep problems, and liver diseases (Halpern et al., 2020). It is therefore imperative for an understanding to be created as to what are the major causes of the problem so that proactive measures are taken and the burden on the medical institutions in Mexico, Peru, and Colombia is eliminated. The second rationale for the selection of the dataset was its columns, over 2000 which are statistically enough to gain meaningful conclusion on how the features relate to the control variable. Moreover, this dataset is clean with no missing variables and has all variables that are known today. The case study dataset also has a diverse data represented in the variables making it fun for creation of catching and attractive visualizations. Through taking a serious issue like healthcare of the Latin American people the research not only confirms or invalidates previous findings on the issue but provides a hands-on conclusion on what can be done to eliminate obesity.

# Data Importation and Preparation

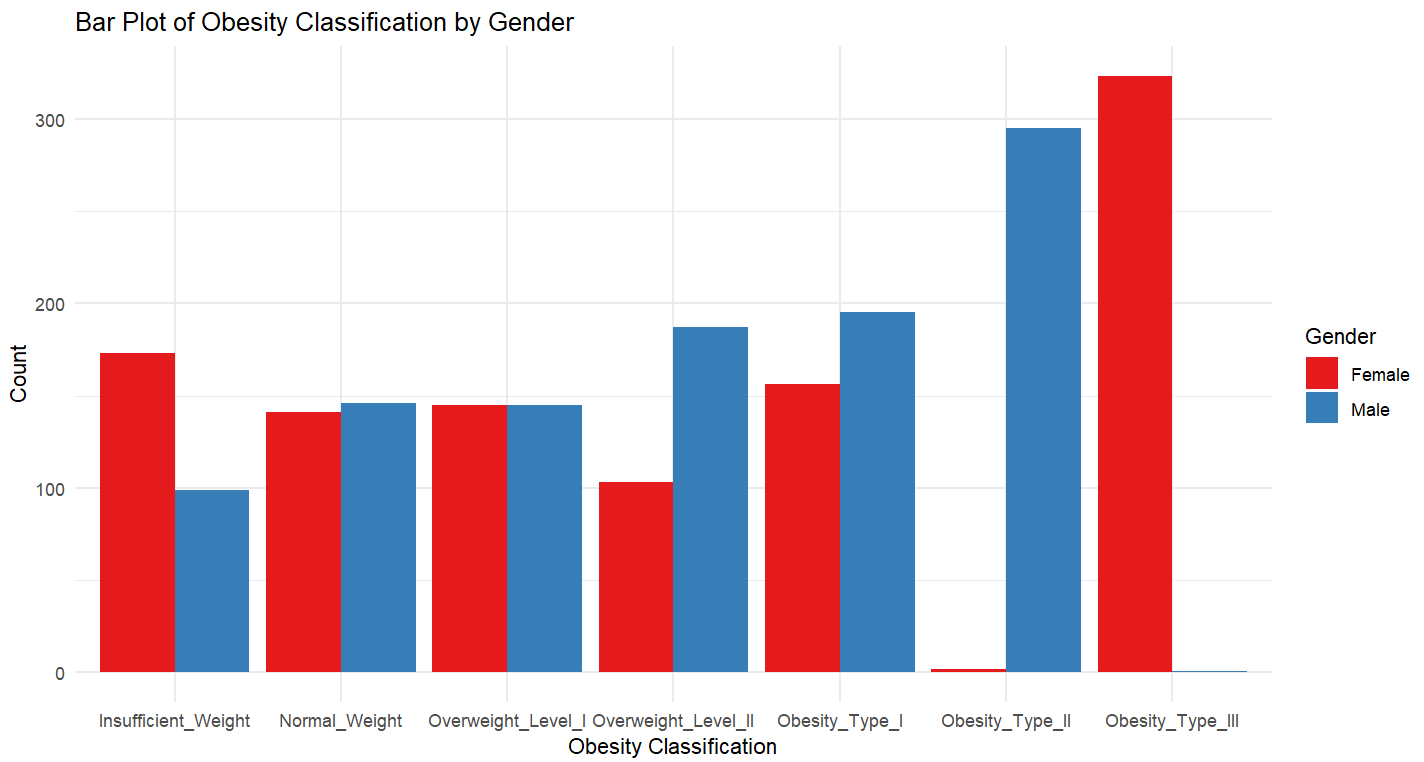
The first step done was to know the primary working directory in Rstudio and storing the downloaded dataset named ObesityData there for easier manipulation. Next was reading the dataset into a data frame named obesity and since the dataset had column headers this was set into true. For ease during the data manipulation stage the feature with the title family\_history\_with\_overweight had its title changed to FHO and the change verified. The dataset was clean and had no missing values has it had been stated in its summary from the provider as results showed that there were no null values. A summary of the obesity dataset showed that the data consisted of 2111 records, the ages of the responder’s ranges from 14-61 years, height from 1.45 to 1.98m, weight 39 to 173 kg. Most of the columns in the dataset were found to be categorical and there were no missing values. For easier and accurate analysis and visualization the columns were crosschecked to ensure that they had the correct datatypes and no changes were required.

# Exploratory Data Visualization

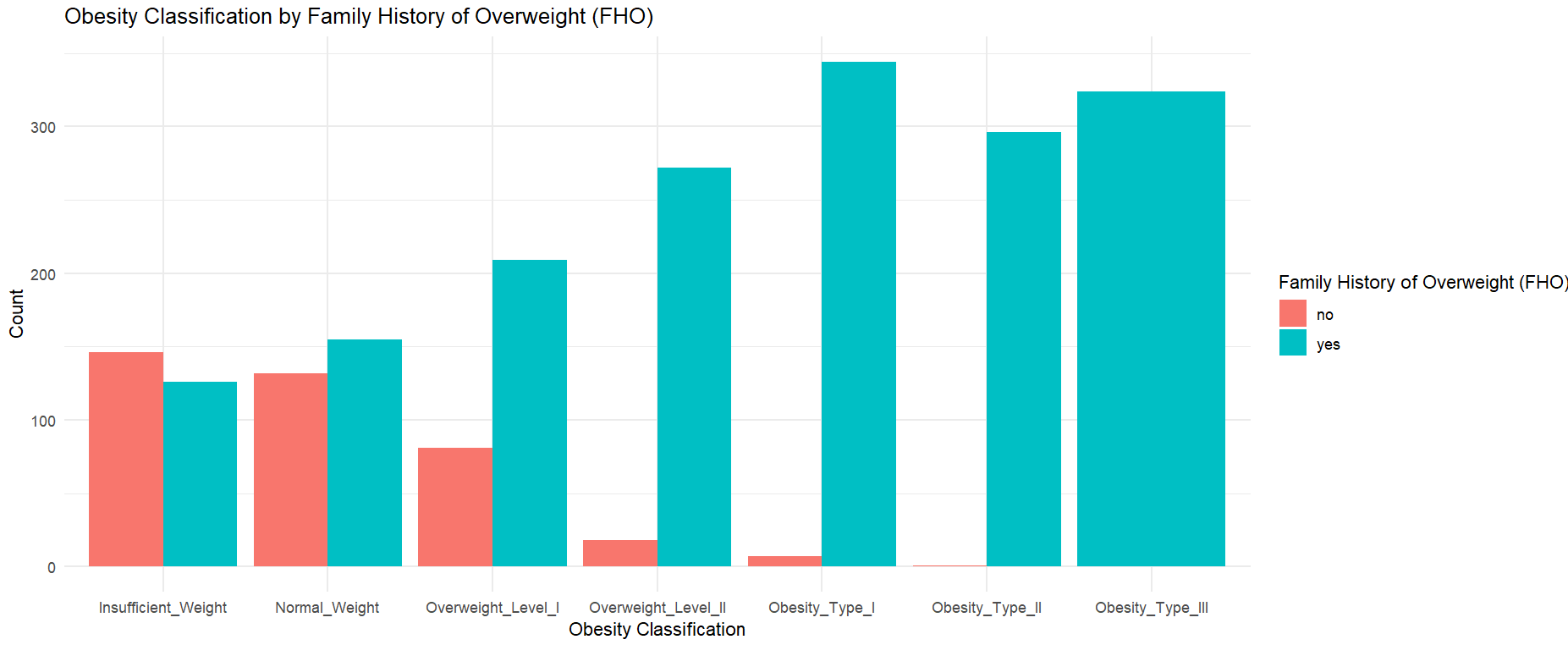
Once the data was ready for analysis, the ggplot2 an effective and popular library for R was installed and leaded. The first step was to visualize how various features of the dataset related to the obesity classification of the participants. Obesity in people is measured based on the body mass index which is the weight and height of the participants. As such it was found that weight is a significant contributor to the condition as shown below.



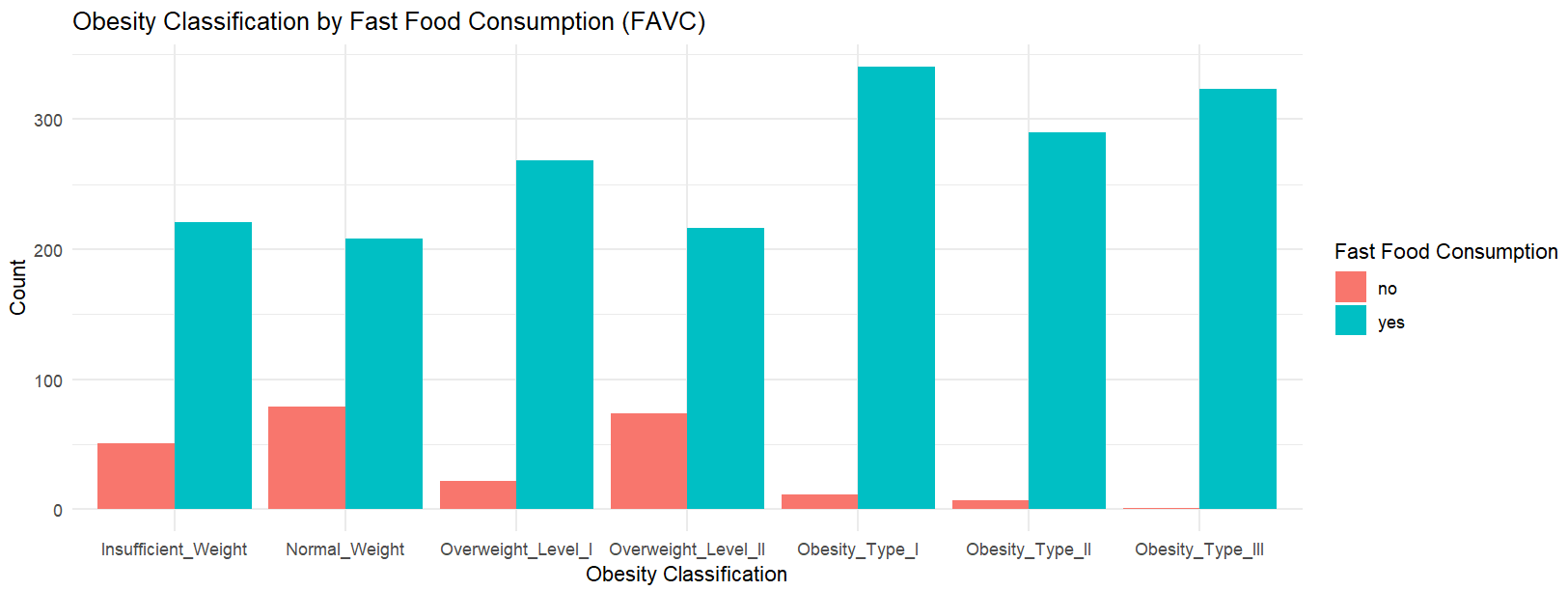
Gender was also found to have impacts of the obesity classification for an individual especially for the type II and III obesities where significant differences were observed between men and women as shown in the figure below.



The results showed that coming from a family with an history of having obesity classifications was a strong predictor for being overweight or being diagnosed with obesity type I, II or III. It was noted that participants with obesity type all came from a family with obesity type 3 as shown in the figure below.



It was established that people who do not consume fast foods were more likely to have normal weights than people who consumed fast foods. There were very few people who claimed that they did not consume fast foods that had obesity types I, II and III as shown in the figure below.



# Model Training and Testing

To train and test a machine learning model to predict the type of obesity classification (NObeyesdad) based on the predictor features in your dataset, R using the caret package was used. A decision tree model was used as the machine learning model of choice. The data was encoded into categorical variables to convert them into factors. The created model had an accuracy of 97.38%, which means it correctly classified 97.38% of the cases. This is quite high, indicating strong model performance.

# Conclusion

The analysis of obesity levels in Mexico, Peru, and Colombia based on dietary habits, physical conditions, and other relevant factors has yielded significant insights into the predictors of obesity and the effectiveness of data-driven models in identifying obesity types. By leveraging the power of R programming and the ggplot2 library, the project effectively visualized the relationships between various factors and obesity classifications, revealing key contributors such as weight, gender, family history, and fast-food consumption.

The exploratory data analysis confirmed that obesity is influenced by a combination of genetic, lifestyle, and dietary factors, with family history, gender, and fast-food consumption standing out as strong predictors. These findings align with existing research and highlight the importance of targeted interventions in addressing obesity within these populations.

The machine learning model, trained using a decision tree algorithm, achieved an impressive accuracy of 97.38%, demonstrating the potential of predictive analytics in healthcare. This high accuracy underscores the model's capability to correctly classify individuals into their respective obesity categories based on the input features.

Overall, this project not only corroborates previous studies on the factors contributing to obesity but also provides a robust predictive model that can be used to inform healthcare strategies in Latin America. The results emphasize the need for continued research and data-driven approaches to tackle obesity, ultimately aiming to reduce the burden on healthcare systems and improve the quality of life for individuals in these regions.

# References

Halpern, B. *et al.* (2020) ‘Obesity and Covid‐19 in Latin America: A tragedy of two pandemics—official document of the Latin American Federation of Obesity Societies’, *Obesity Reviews*, 22(3). doi:10.1111/obr.13165.

Hwalla, N. and Jaafar, Z. (2020) ‘Dietary management of obesity: A review of the evidence’, *Diagnostics*, 11(1), p. 24. doi:10.3390/diagnostics11010024.

World Health Organization (2022) *Obesity and overweight.* Available at <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight#:~:text=In%202022%2C%201%20in%208,million%20were%20living%20with%20obesity>. Accessed: 11 may 2020.

Willey, J., Wakefield, M. and Silver, H.J. (2020) ‘Exploring the diets of adults with obesity and type II diabetes from nine diverse countries: Dietary intakes, patterns, and quality’, *Nutrients*, 12(7), pp. 1–16. doi:10.3390/nu12072027.