HEALTH DATA ANALYSIS with R

A Practical Guide Using the NHANES Dataset

By Said Abbas

1st edition, 2025
Publisher: Zenodo by CERN

Health Data Analysis with R

A Practical Guide Using the NHANES Dataset

By Said Abbas

1st edition, 2025

Publisher: Zenodo by CERN

Acknowledgment

I would like to express my gratitude to the creators and maintainers of the National Health and Nutrition Examination Survey (NHANES) dataset for making high-quality public health data openly available to researchers worldwide. This book has also benefited from the robust open-source R ecosystem, particularly the work of contributors to the tidyverse, janitor, and other packages cited herein.

I am also thankful to educators, statisticians, and open science advocates whose commitment to accessible learning has inspired the approach taken in this work. Finally, special thanks go to my colleagues, students, and readers for their feedback, curiosity, and dedication to applying data science in meaningful ways.

Copyright

ISBN: [Insert if applicable]

Published by: [Your Publisher / Self-Published Info]

Creative Commons License

Preface

Health research in the 21st century is increasingly data-driven. From large-scale population surveys to precision medicine, the ability to work confidently with data is no longer an optional skill for healthcare professionals—it is essential.

This book is designed as a practical guide for those who wish to learn and apply statistical and data science techniques to real-world medical datasets. Using the NHANES dataset as our example, we take readers step-by-step through the processes of data cleaning, exploration, statistical testing, epidemiological analysis, and predictive modeling, all within the R programming environment.

Every chapter provides executable R code alongside clear explanations and healthcare-relevant interpretations, ensuring that readers not only know how to run an analysis, but also why it matters in a clinical or public health context.

Whether you are a clinician seeking to evaluate patient data, a researcher preparing for publication, or a student building your analytical foundation, this book offers both the tools and the reasoning you need to succeed.

Declaration

This work was developed with the assistance of AI-based tools for drafting, structuring, and formatting text. However, all core concepts, analytical approaches, interpretations, and educational framing contained in this book are the original intellectual property of the author.

The AI was employed solely as a supportive tool to enhance clarity, streamline explanations, and ensure consistency in presentation. The analytical content, methodological choices, interpretations of results, and contextual explanations have been conceived, verified, and validated by the author, based on their expertise and knowledge in the subject matter.

The responsibility for the accuracy, integrity, and originality of the ideas and interpretations presented in this book rests entirely with the author.

Contents

How to Use This Book

This book has been designed as a practical, hands-on guide. Each chapter blends explanation with real, executable code so that you can learn by doing. The workflow is deliberately interactive: read a concept, run the code, observe the output, and then interpret the results in a healthcare context.

Code Examples

All analyses in this book are accompanied by ready-to-use R code. For your convenience, code blocks are visually distinguished from the main text and follow a consistent style.

In the digital version of this book, code intended for you to run will be displayed in green. To reproduce an analysis:

1. Locate the green-highlighted code within the chapter.
2. Copy the entire block exactly as shown.
3. Paste it into RStudio and run it.
4. Observe the output in the Console and/or Plots pane.

The code provided is fully functional, based on the NHANES dataset included in the R NHANES package. You may modify, expand, or adapt it for your own purposes—experimenting with different variables, adjusting parameters, or applying the techniques to your own datasets.

Enhancing the Provided Code

The included scripts form a foundation for your analysis. Once you have reproduced the examples successfully, consider:

- Changing dataset filters (e.g., selecting only certain age groups).
- Adding new statistical tests or visualizations.
- Integrating external datasets alongside NHANES.
- Combining multiple chapters' techniques into a custom workflow.

Reference to the Uploaded Codebook

All code in this book has been derived from, and is consistent with, the comprehensive codebook prepared for this project. This ensures that every example is complete, reproducible, and aligned with the book’s learning objectives. The codebook serves as a reference library you can revisit it whenever you want to explore a method covered in the text.

Introduction to the NHANES Dataset and This Codebook

Understanding NHANES

The National Health and Nutrition Examination Survey (NHANES) is one of the most comprehensive and influential health data sources in the world. Conducted in the United States by the National Center for Health Statistics (NCHS), NHANES provides a uniquely rich combination of:

- **Structured interviews** – capturing health behaviors, lifestyle factors, and demographic information.
- **Physical examinations** – measuring height, weight, blood pressure, and other vital signs.
- **Laboratory tests** – including cholesterol, blood glucose, and other clinical markers.

Unlike many surveys that rely solely on self-reported information, NHANES blends *self-report*, *clinical measurement*, and *laboratory analysis*. This produces a dataset that is both robust and **representative**, covering diverse aspects of health and nutrition across the U.S. population.

For healthcare professionals, NHANES is not just a dataset—it's a learning platform. By working with it, you gain exposure to the challenges and opportunities that real-world health data presents: missing values, multiple measurement types, complex sampling designs, and variables that span from biological markers to social determinants of health.

In this codebook, NHANES serves as both the teaching foundation and the real-world example through which every statistical and analytical technique will be demonstrated.

Why NHANES?

The decision to use NHANES in this guide is intentional:

- **Publicly Available** – Anyone can access and analyze it without restrictions.
- **Nationally Representative** – Data are collected using complex sampling to reflect the U.S. population.
- **Comprehensive** – With hundreds of variables, it supports diverse research topics from obesity to environmental exposures.
- **Educational Value** – The breadth of variables makes it ideal for demonstrating a wide range of statistical methods in a single dataset.

Whether you are a student, clinician, or researcher, the NHANES dataset offers a laboratory for learning data science while engaging with meaningful health questions.

How This Codebook Is Written

This codebook is designed not as a static technical manual, but as a guided learning journey through the process of medical data analysis using R. The chapters are structured to build your skills progressively—starting from basic data handling and moving towards advanced statistical and predictive modeling techniques.

Each chapter follows a consistent format:

1. **Concept Overview** – A brief introduction to the topic and why it matters in health research.
2. **Hands-On Code Examples** – Fully runnable R scripts with clear, descriptive comments.
3. **Sample Output** – Actual results from running the code, so you can check your work.
4. **Interpretation** – Plain-language explanations of what the results mean in a real healthcare or public health context.

The Learning Approach

The teaching style in this codebook is deliberately practical:

- **Modular Structure** – Each chapter stands on its own. Repeated core steps (e.g., loading NHANES, cleaning variables) are included at the start of every module, so you can begin anywhere without depending on earlier chapters.

- **Real-World Relevance** – Variables such as BMI, diabetes status, cholesterol levels, and physical activity are used because they are meaningful in clinical and public health settings.
- **Data Quality Awareness** – Missing values, categorical coding, and survey structure are not ignored; they are addressed directly, reflecting real-world analysis challenges.
- **Visualization-Driven Learning** – ggplot2 visualizations are integrated throughout, reinforcing both statistical understanding and data storytelling.

By the end of this book, you will not only know how to run statistical analyses—you will understand **why** they are applied, **how** to interpret them, and **what** they mean for healthcare decision-making.

What Comes Next

The chapters ahead take you step-by-step through:

- Data cleaning and preparation.
- Exploratory data analysis.
- Statistical testing and modeling.
- Epidemiological measures and risk analysis.
- Visualization and reporting for medical insight.
- Introduction to predictive modeling.

We begin with the foundations loading, exploring, and summarizing NHANES data before progressing to methods that reveal patterns, test hypotheses, and build predictive tools. By pairing theory with practice, you will gain both technical fluency and analytical confidence in working with health data.

Getting Started with R and RStudio

2.1 Introduction

Before we begin analyzing the NHANES dataset, we need to set up our working environment. In this chapter, we will introduce **R**, the statistical programming language at the core of all our examples, and **RStudio**, the integrated development environment (IDE) that makes working with R more efficient and user-friendly.

We will also cover the essential **libraries** also known as packages—that extend R's functionality for data cleaning, visualization, and statistical analysis. By the end of this chapter, you will have installed R, RStudio, and the packages required to follow every example in this codebook.

2.2 What is R?

R is an open-source programming language designed specifically for **statistical computing**, **data analysis**, and **visualization**. Since its creation by statisticians Ross Ihaka and Robert Gentleman in the early 1990s, R has become one of the most widely used tools for data science and research worldwide.

Key Features of R:

- **Specialized for analysis** – built-in functions for descriptive and inferential statistics.
- **Extensible** – thousands of add-on packages for specialized analysis and visualization.
- **Free and open-source** – no license costs, with community-driven development.
- **Cross-platform** – works on Windows, macOS, and Linux.

Download R: https://cran.r-project.org

2.3 What is RStudio?

While R can be run in its basic console, most users prefer **RStudio**, a powerful IDE that makes R programming easier to manage. RStudio provides a clean, organized workspace and tools for writing, running, and debugging code.

Advantages of RStudio:

- **Script editor** – write, save, and re-run code easily.
- **Interactive console** – test commands quickly.
- **Environment pane** – track your variables and datasets.
- **Plots pane** – view visualizations directly within the interface.
- **Package management** – install, update, and load libraries with a click.

Download RStudio: https://posit.co/download/rstudio-desktop/

2.4 Installing R and RStudio

Step 1 – Install R

1. Visit the CRAN website.
2. Select your operating system (Windows, macOS, or Linux).
3. Download and install using the provided installer.

Step 2 – Install RStudio

1. Download the **free RStudio Desktop** edition from the Posit website.
2. Install it as you would any software on your operating system.

2.5 Installing Required Packages

Package	*Primary Function*
dplyr	Data manipulation (filtering, selecting, summarizing, and mutating data).
ggplot2	Data visualization (creating high-quality, reproducible plots).
tidyr	Data tidying (making data "tidy" for easier analysis).
stringr	Working with character strings and text data.

R’s true power comes from its packages, which are collections of functions for specific tasks. The first time you run a script from this codebook, you will need to install the following packages:

```
# Install packages (only needs to be run once)

install.packages(c("tidyverse", "janitor", "survey", "NHANES"))



# Load the packages (needs to be run at the start of every new
session)

library(tidyverse) # The core suite for data wrangling and
visualization

library(janitor)   # For cleaning data names and generating
quick frequency tables

library(survey)    # ESSENTIAL for correctly analyzing NHANES
complex sample design

library(NHANES)    # A package providing a pre-loaded, cleaned
subset of NHANES data
```

2.6 Overview of Essential Packages

Package	Purpose	Why It’s Important
tidyverse	Suite for data science (dplyr, ggplot2, tidyr, readr).	Simplifies data manipulation and visualization.
janitor	Data cleaning tools (e.g., fix column names).	Prepares messy datasets for analysis.
NHANES	Loads NHANES survey data directly into R.	The dataset used throughout this book.
caret	Unified machine learning framework.	Standardizes model building and evaluation.
rpart	Decision tree modeling.	Creates interpretable classification models.
randomForest	Ensemble decision tree learning.	Improves predictive accuracy.
lme4	Mixed-effects modeling.	Analyzes hierarchical or longitudinal data.

2.7 Recommended Workflow in This Book

Throughout this codebook, our analysis process will follow a consistent workflow:

1. **Load Libraries** – Ensure all packages are available.
2. **Import Data** – Load NHANES or other datasets.
3. **Clean and Explore** – Prepare the dataset for analysis.
4. **Analyze and Model** – Apply statistical and machine learning methods.
5. **Visualize** – Present findings with clear, readable graphics.
6. **Interpret** – Explain results in a clinical or public health context.

2.8 Chapter Summary

In this chapter, you learned:

- What R is and why it is a standard tool for statistical analysis.
- Why RStudio is the preferred interface for working with R.
- How to download and install both R and RStudio.
- The core packages you will need for this codebook.

Before You Begin – Master Code Reference

This section is your **quick-access guide** to every R code pattern used in this book. It explains what each command does, why it is used, and in which chapters it appears.

If you are new to R, think of this as your code map a place you can return to whenever you encounter a function in the modules and want a refresher on its purpose.

1. Repeated Core Code

These commands appear in almost every chapter. They are the foundations of our workflow.

1.1 Load Required Libraries

```
r
CopyEdit
library(tidyverse)   # Core data manipulation & visualization
library(janitor)     # Cleaning and formatting column names
library(NHANES)      # Access to NHANES dataset
```

Purpose:

- `tidyverse` – A collection of packages (dplyr, ggplot2, etc.) for modern data analysis.
- `janitor` – Cleans messy column names, makes them consistent.
- `NHANES` – Loads our health survey dataset.

Appears in: All chapters.

1.2 Load the Dataset

```
r
CopyEdit
data("NHANES")
```

Purpose: Loads the NHANES dataset into your working environment.

Appears in: All chapters.

1.3 Clean Column Names

```
r
```

```
CopyEdit
NHANES <- NHANES %>% clean_names()
```

Purpose: Converts all column names to lowercase and replaces spaces with underscores.

Appears in: Chapters 1–3, 6.

1.4 View Dataset Structure

```
r
CopyEdit
glimpse(NHANES)
```

Purpose: Displays number of rows, columns, and data types.

Appears in: Chapters 1–3.

1.5 Summary Statistics

```
r
CopyEdit
summary(NHANES)
```

Purpose: Generates quick descriptive statistics for each variable.

Appears in: Chapters 1–3.

2. Module-Specific Code Highlights

Module 1 – Initial Data Exploration

- **Filter for Adults**

```
r
CopyEdit
adult_data <- NHANES %>% filter(age >= 18)
```

Keeps only participants aged 18+.

- **Check Missing Values**

```
r
CopyEdit
colSums(is.na(adult_data)) %>% sort(decreasing = TRUE)
```

Counts missing values per column.

- **Unique Values in Categorical Variables**

```
r
CopyEdit
adult_data %>%
  select(gender, race1, marital_status, education) %>%
  summarise_all(~n_distinct(.))
```

Counts unique category levels.

Module 2 – Exploratory Data Analysis

- **Check Missing in Key Variables**

```
r
CopyEdit
colSums(is.na(NHANES[, c("Gender", "BMI")]))
```

Counts missing values in specific variables.

- **Remove Incomplete Rows**

```
r
CopyEdit
nhanes_clean <- NHANES %>%
  filter(!is.na(Gender) & !is.na(BMI))
```

Keeps only rows with complete Gender and BMI values.

Module 3 – Data Wrangling

- **Select Subsets**

```
r
CopyEdit
lab_data <- NHANES %>% select(ID, BMI, TotChol)
demo_data <- NHANES %>% select(ID, Age, Gender)
```

Pulls specific variable groups.

- **Merge Datasets**

```
r
CopyEdit
merged_data <- left_join(lab_data, demo_data, by = "ID")
```

Joins lab and demographic data.

- **Reshape Data**

```
r
CopyEdit
long_data <- pivot_longer(
  merged_data,
  cols = c(BMI, TotChol),
  names_to = "Measure",
  values_to = "Value"
)
```

Converts wide data to long format for plotting.

Module 4 – Biostatistics

- **T-Test**

```
r
CopyEdit
t.test(BMI ~ Gender, data = nhanes_clean)
```

Compares mean BMI between genders.

- **ANOVA**

```
r
CopyEdit
aov(BMI ~ PhysActive, data = nhanes_clean)
```

Tests differences across physical activity groups.

- **Correlation**

```
r
CopyEdit
cor.test(nhanes_clean$Age, nhanes_clean$BMI)
```

Measures relationship between age and BMI.

- **Linear Regression**

```
r
CopyEdit
lm(BMI ~ Age, data = nhanes_clean)
```

Predicts BMI from age.

Module 5 – Epidemiology

- **Contingency Table**

```
r
CopyEdit
table(epi_data$active_bin, epi_data$diabetes_bin)
```

Shows diabetes by physical activity group.

- **Risk Ratio**

```
r
CopyEdit
risk_ratio <- prevalence_active / prevalence_inactive
```

Compares risk between groups.

Module 6 – Visualization

- **Histogram**

```
r
CopyEdit
ggplot(NHANES, aes(x = BMI)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black")
```

Displays BMI distribution.

- **Boxplot**

```
r
```

```
CopyEdit
ggplot(NHANES, aes(x = Diabetes, y = BMI, fill = Diabetes)) +
  geom_boxplot()
```

Compares BMI by diabetes status.

Module 7 – Time Series

- **Simulate Longitudinal Data**

```
r
CopyEdit
long_data <- data.frame(
  ID = rep(1:50, each = 3),
  Year = rep(c(2018, 2019, 2020), times = 50),
  BMI = round(rnorm(150, mean = 25, sd = 3) + rep(c(0, 0.5, 1),
times = 50), 2)
)
```

- **Mixed-Effects Model**

```
r
CopyEdit
lmer(BMI ~ Year + (1 | ID), data = long_data)
```

Module 8 – Predictive Modeling

- **Train/Test Split**

```
r
CopyEdit
split_index <- sample(1:nrow(model_data), 0.7 *
nrow(model_data))
train_data <- model_data[split_index, ]
test_data <- model_data[-split_index, ]
```

- **Random Forest**

```
r
CopyEdit
randomForest(Diabetes ~ Age + BMI + PhysActive, data =
train_data)
```

3. Summary of Most Repeated Code

Code	Purpose	Appears In
`library(tidyverse)`	Load data/plot tools	All modules
`data("NHANES")`	Load dataset	All modules
`clean_names()`	Clean variable names	1–3, 6
`glimpse()`	View structure	1–3
`summary()`	Basic stats	1–3
`filter(...)`	Subset data	1, 2, 5
`group_by() %>% summarise()`	Group summaries	1, 2, 4
`ggplot()`	Visualization	1, 2, 6, 7
`is.na()`	Missing data checks	1–3

Module 1

Initial Data Exploration

In today’s data-driven world, information has become a vital resource across every field. However, the true value of data lies not just in its availability, but in how effectively it is applied to meet specific needs and solve real-world problems. With the rise of data science as a multidisciplinary field, there is a growing necessity to understand and utilize its foundational concepts—especially statistical analysis—across professional domains.

For medical professionals, acquiring data literacy and analytical skills is no longer optional; it is essential. As healthcare increasingly incorporates digital records, wearable devices, and large-scale research databases, clinicians and researchers must be able to navigate, interpret, and apply data insights to enhance patient care and advance medical research.

This codebook is crafted with the aim of introducing healthcare and medical professionals to the practical possibilities that data science and statistical techniques offer. It emphasizes hands-on learning, focusing on the technical side of data handling rather than theoretical underpinnings. Each section of the book is structured to walk the reader through statistical methods relevant to medical data, accompanied by actual R code, output interpretation, and real-world application.

Whether you're a beginner or looking to refine your analytical toolkit, this guide offers a pathway to become more confident in using data. From cleaning and summarizing datasets to performing statistical tests and building predictive models, the book provides a step-by-step, code-centric approach to mastering medical data analysis.

Load Libraries

```
# Load all the necessary libraries. Ensure they are installed
before loading.

# If not installed, install using:
install.packages("package_name")
```

```
library(tidyverse)   # Core packages for data manipulation and visualization

library(janitor)     # For cleaning column names and quick data cleaning

library(NHANES)      # NHANES dataset package
```

Important Notes

- This module uses an open-source medical dataset (NHANES),
- Available through the NHANES R package.
- If analyzing your own dataset, make sure to set the correct file path.
- The dataset can be in formats like CSV, Excel, or TXT.

Load and Clean the Dataset

```
# Load the NHANES dataset into the environment

data("NHANES")


# Clean column names: converts to lowercase and replaces spaces with underscores

NHANES <- NHANES %>% clean_names()
```

# Explore the Dataset

```
# Generate summary statistics for all variables

summary(NHANES)
```

Summary output

Statistic/Factor Level	id	survey_yr	gender	age	age_decade	age_months
Min.	51624	—	—	0	—	0
1st Quartile	56905	—	—	17	—	199
Median	62160	—	—	36	—	418
Mean	61945	—	—	36.74	—	420.1
3rd Quartile	67039	—	—	54	—	624
Max.	71915	—	—	80	—	959
\	\	\	\	\	\	\
Factor Level Counts						
2009_10	—	5000	—	—	—	—
2011_12	—	5000	—	—	—	—
female	—	—	5020	—	—	—
male	—	—	4980	—	—	—
40-49	—	—	—	—	1398	—
0-9	—	—	—	—	1391	—
10-19	—	—	—	—	1374	—
20-29	—	—	—	—	1356	—
30-39	—	—	—	—	1338	—
(Other)	—	—	—	—	2810	—
\	\	\	\	\	\	\
NA's	—	—	—	—	333	5038

Interpretation

This table summarizes the key demographic variables from the NHANES dataset, offering a snapshot of the sample distribution across age, gender, and survey years. It includes basic statistics such as minimum, maximum, mean, median, and quartiles, as well as category counts and missing values.

It can be done for all but we have only shown for few.

Variables description

Variable	Summary Description
id	Participant identifiers range from 51624 to 71915.

survey_yr	Data spans two cycles: 2009-2010 and 2011-2012, each with 5000 individuals.
gender	Fairly balanced sample: 5020 females and 4980 males.
age	Ranges from 0 to 80 years, with a mean of 36.74 and median of 36.
age_decade	Most frequent age groups: 40–49 (1398), 0–9 (1391), 10–19 (1374); 333 missing values.
age_months	Age in months varies from 0 to 959, mean = 420.1; 5038 missing values.

```
# View the structure and data types of the dataset
glimpse(NHANES)
#by applying the above code you will get the following output
```

Glimpse Output

Column Name	Data Type	Description	Sample Values
id	<int> (Integer)	Unique respondent identifier.	51624, 51625, 51630
survey_yr	<fct> (Factor)	The specific survey cycle year.	2009_10, 2011_12
gender	<fct> (Factor)	The respondent's biological sex.	male, female
age	<int> (Integer)	Age of the respondent in years.	34, 4, 49, 9
age_decade	<fct> (Factor)	Age grouped into 10-year intervals.	30-39, 0-9, 40-49
age_months	<int> (Integer)	Age of the respondent in months.	409, 49, 596, 115
race1	<fct> (Factor)	Primary race/ethnicity category.	White, Other
race3	<fct> (Factor)	Secondary race category.	NA (mostly missing)

Interpretation

The dataset contains 10,000 observations (rows) and 76 variables (columns) collected from participants in the NHANES (National Health and Nutrition Examination Survey) program. The first few key variables give a demographic and identity-based overview of the dataset.

Description

Variable	Type	Description

id	Integer	Unique identifier for each participant. Repeated when multiple rows relate to same individual (e.g., multiple tests or waves).
survey_yr	Factor	NHANES cycle year, e.g., "2009_10", "2011_12". Indicates time of data collection.
gender	Factor	Participant sex, categorized as "male" or "female".
age	Integer	Age in years. Ranges from infants to elderly adults.
age_decade	Factor	Age categorized by decade (e.g., "0-9", "30-39"). Useful for grouped analysis.
age_months	Integer	Age in months for finer age resolution. Missing for some older individuals.
race1	Factor	Broad racial/ethnic category (e.g., "White", "Black", "Hispanic", "Other").
race3	Factor	More detailed racial classification; missing (NA) in many rows in your snippet.

Same codes will be used in almost all the modules but later there will be no description it’s just a reminder.

NHANES Dataset: Sample Demographics and Initial Summary Tables

```
# Load and clean NHANES dataset if not already done

library(tidyverse)

library(janitor)

library(NHANES)


data("NHANES")

NHANES <- NHANES %>% clean_names()
```

Demographic Summary

This summary table gives an overview of basic demographic features such as age, gender distribution, age decades, and monthly age representation.

```
summary(select(NHANES, id, survey_yr, gender, age, age_decade,
age_months))
```

Basic Summary

Variable	Type	Description	Key Statistics/Counts

id	Integer	Unique identifier for each participant. Repeated when multiple rows relate to the same individual.	Ranges from 51624 to 71915.
survey_yr	Factor	NHANES cycle year indicating the time of data collection.	2009-2010: 5000, 2011-2012: 5000. (Perfectly balanced)
gender	Factor	Participant sex, categorized as "male" or "female".	Female: 5020, Male: 4980. (Highly balanced)
age	Integer	Age in years. Ranges from infants to elderly adults.	Range: 0 to 80 years. Mean: 36.74, Median: 36.00.
age_decade	Factor	Age categorized by decade (e.g., "0-9", "30-39").	Top groups: 40-49 (1398), 0-9 (1391). Missing (NA's): 333.
age_months	Integer	Age in months for finer age resolution.	Range: 0 to 959 months. Mean: 420.1. Missing (NA's): 5038.

Interpretation:

- \`age\` ranges from newborns to 80 years, median age ~36 years.
- \`gender\` is nearly balanced: ~50% male and female.
- \`age_decade\` shows largest concentration in 40–49 age group.
- \`age_months\` has many missing values (NA's: 5038).

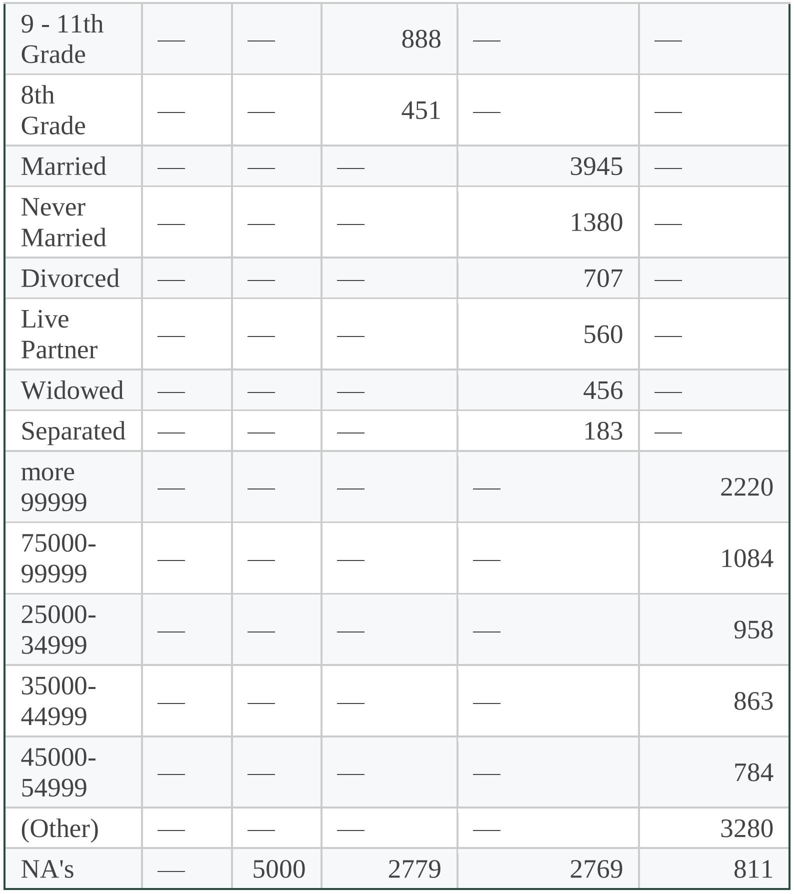
Economic & Living Conditions Summary

These variables describe the social and financial background of individuals.

```
summary(select(NHANES, race1, race3, education, marital_status,
hh_income))
```

Selected variables summary

Level	race1	race3	education	marital_status	hh_income
White	6372	3135	—	—	—
Black	1197	589	—	—	—
Mexican	1015	480	—	—	—
Other	806	158	—	—	—
Hispanic	610	350	—	—	—
Asian	—	288	—	—	—
Some College	—	—	2267	—	—
College Grad	—	—	2098	—	—
High School	—	—	1517	—	—



9 - 11th Grade	—	—	888	—	—
8th Grade	—	—	451	—	—
Married	—	—	—	3945	—
Never Married	—	—	—	1380	—
Divorced	—	—	—	707	—
Live Partner	—	—	—	560	—
Widowed	—	—	—	456	—
Separated	—	—	—	183	—
more 99999	—	—	—	—	2220
75000-99999	—	—	—	—	1084
25000-34999	—	—	—	—	958
35000-44999	—	—	—	—	863
45000-54999	—	—	—	—	784
(Other)	—	—	—	—	3280
NA's	—	5000	2779	2769	811

Highlights:

- race1` and `race3` cover ethnicity details. White and Mexican are the most represented.
- Education: Over 20% completed college, but nearly 2,800 entries are NA.
- Marital status: Most respondents are married (~3,945), followed by never married and divorced.
- Income: The modal income is "more than $99,999", but 811 missing values exist.

Weight & Housing Information

# This section presents BMI, room availability at home, and employment status.

```
summary(select(NHANES, hh_income_mid, poverty, home_rooms,
home_own, work, weight))
```

Statistic/Factor Level	hh_income_mid	poverty	home_rooms	home_own	work	weight
Min.	2,500	0	1	—	—	2.8
1st Quartile	30,000	1.24	5	—	—	56.1
Median	50,000	2.7	6	—	—	72.7
Mean	57,206	2.802	6.249	—	—	70.98
3rd Quartile	87,500	4.71	8	—	—	88.9
Max.	100,000	5	13	—	—	230.7
Factor Counts						
Own	—	—	—	6425	—	—
Rent	—	—	—	3287	—	—
Other (Home/Rooms)	—	—	—	225	—	—
Working	—	—	—	—	4613	—
Not Working	—	—	—	—	2847	—
Looking (for work)	—	—	—	—	311	—
Missing (NA's)	811	726	69	63	2229	78

Important Notes:

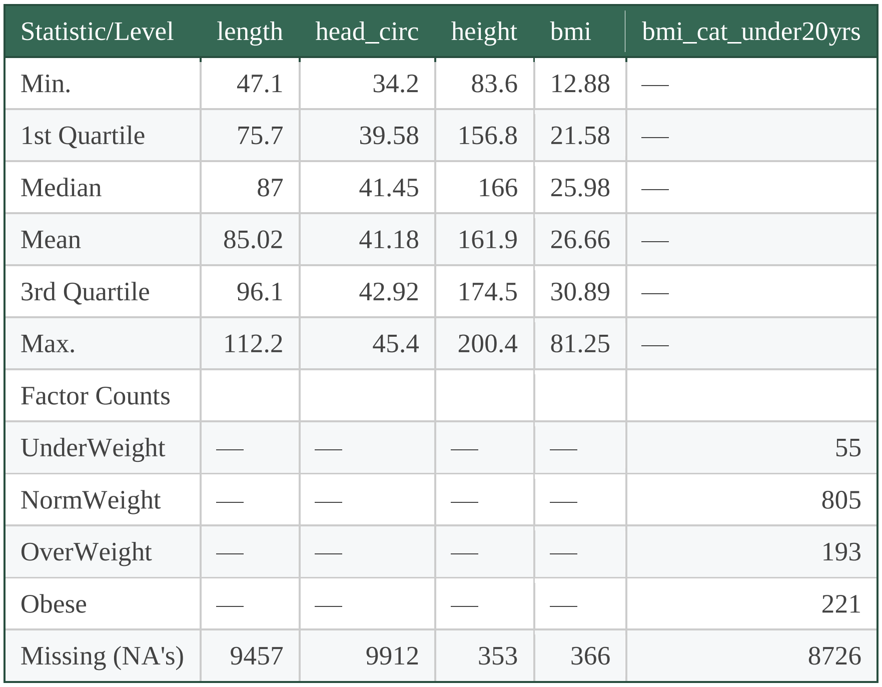
- `weight` spans from 2.8 to 230.7 kg, mean ~71 kg.
- Most individuals live in homes with 5–8 rooms.
- Majority are homeowners (Own: 6425), while renters are ~3287.
- Nearly half are employed, but 2229 responses are missing on employment.

Physical Measurements Overview

Basic physical health measurements including height, BMI, and under-20 BMI category.

```
summary(select(NHANES, length, head_circ, height, bmi,
bmi_cat_under20yrs))
```

Summary



Statistic/Level	length	head_circ	height	bmi	bmi_cat_under20yrs
Min.	47.1	34.2	83.6	12.88	—
1st Quartile	75.7	39.58	156.8	21.58	—
Median	87	41.45	166	25.98	—
Mean	85.02	41.18	161.9	26.66	—
3rd Quartile	96.1	42.92	174.5	30.89	—
Max.	112.2	45.4	200.4	81.25	—
Factor Counts					
UnderWeight	—	—	—	—	55
NormWeight	—	—	—	—	805
OverWeight	—	—	—	—	193
Obese	—	—	—	—	221
Missing (NA's)	9457	9912	353	366	8726

Key Insights:

`bmi` ranges from 12.88 to 81.25, mean ~26.66.

`height` and `length` have large NA values, indicating missing pediatric or adult measures.

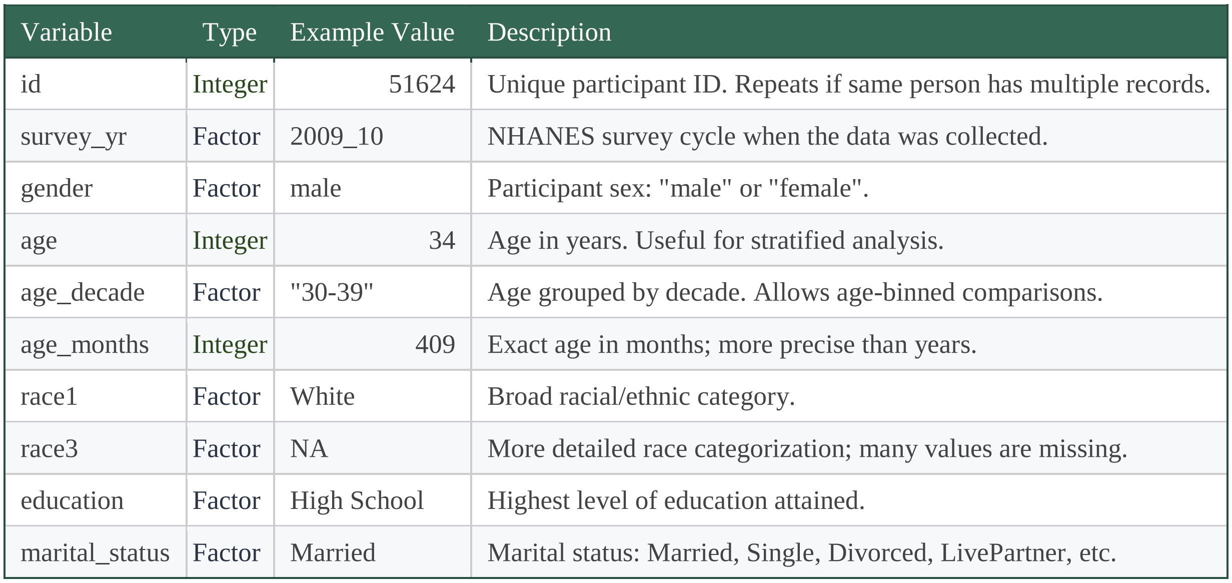
Underweight, normal, overweight, and obese categories exist for under-20s,

but ~8,726 values are missing (likely adults).

Filter for Adult Participants (age >= 18)

```
# Filter adults for analysis
adult_data <- NHANES %>% filter(age >= 18)

# View a few sample rows
head(adult_data)
```

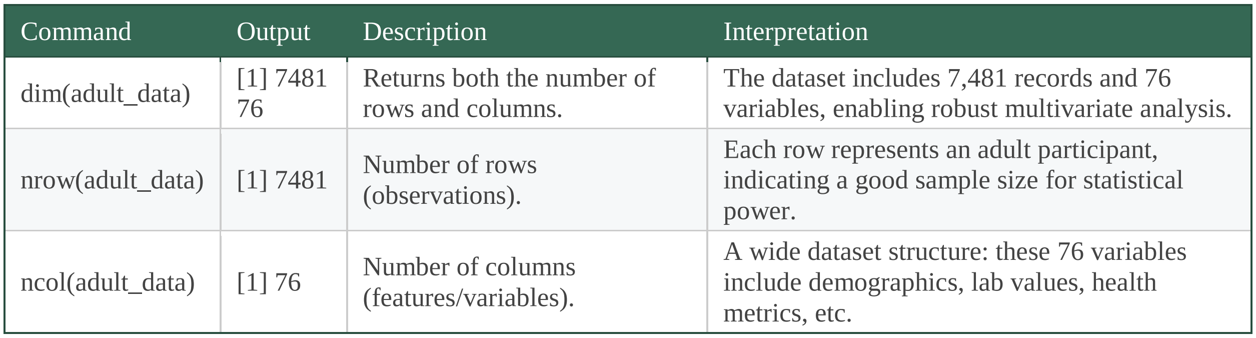
Head data set



Variable	Type	Example Value	Description
id	Integer	51624	Unique participant ID. Repeats if same person has multiple records.
survey_yr	Factor	2009_10	NHANES survey cycle when the data was collected.
gender	Factor	male	Participant sex: "male" or "female".
age	Integer	34	Age in years. Useful for stratified analysis.
age_decade	Factor	"30-39"	Age grouped by decade. Allows age-binned comparisons.
age_months	Integer	409	Exact age in months; more precise than years.
race1	Factor	White	Broad racial/ethnic category.
race3	Factor	NA	More detailed race categorization; many values are missing.
education	Factor	High School	Highest level of education attained.
marital_status	Factor	Married	Marital status: Married, Single, Divorced, LivePartner, etc.

```
# Dataset dimensions: number of rows and columns
dim(adult_data)      # [rows, columns]
nrow(adult_data)     # Number of observations
ncol(adult_data)     # Number of variables
```

Dimensions of the data set



Command	Output	Description	Interpretation
dim(adult_data)	[1] 7481 76	Returns both the number of rows and columns.	The dataset includes 7,481 records and 76 variables, enabling robust multivariate analysis.
nrow(adult_data)	[1] 7481	Number of rows (observations).	Each row represents an adult participant, indicating a good sample size for statistical power.
ncol(adult_data)	[1] 76	Number of columns (features/variables).	A wide dataset structure: these 76 variables include demographics, lab values, health metrics, etc.

Interpretation

- Redundancy in IDs (e.g., 51624 appears multiple times) suggests multiple records per person, possibly due to repeated tests or multiple module recordings.

- Demographics like gender, age, race1, and education form the backbone for subgroup analysis in health data.
- Missing values (as seen in race3) should be handled before modeling or statistical testing.

Missing Data Check

```
# Count of missing values per column, sorted in descending order

colSums(is.na(adult_data)) %>% sort(decreasing = TRUE)
```

Variable	Missing Values	Total Observations	% Missing	Description
`length`	7481	7481	100%	Body length (likely for infants); not applicable to adults.
`head_circ`	7481	7481	100%	Head circumference; typical for pediatric measurements.
`bmi_cat_under20yrs`	116	7481	1.55%	BMI category for those under 20; relevant only to youth subset.
`testosterone`	3438	7481	45.97%	Serum testosterone level; missing in nearly half the sample.
`phys_active_days`	3689	7481	49.30%	Days physically active per week; useful for behavioral analysis.
`alcohol_day`	4914	7481	65.68%	Average alcohol consumption per day; large gaps likely due to skip patterns or sensitivity.
`sex_num_partn_life`	5725	7481	76.52%	Number of lifetime sexual partners; significant non-response.
`bp_sys_ave`	7205	7481	96.32%	Average systolic BP; mostly missing, perhaps not collected.
`urine_vol1`	7384	7481	98.70%	Volume from urine test round 1; almost fully missing.

Interpretation

- Variables like `length` and `head_circ` are pediatric-specific and naturally missing for adults, making them non-applicable for this subset.

- Behavioral and sensitive questions such as `alcohol_day` and `sex_num_partn_life` show high non-response rates, a common issue in surveys due to privacy concerns.

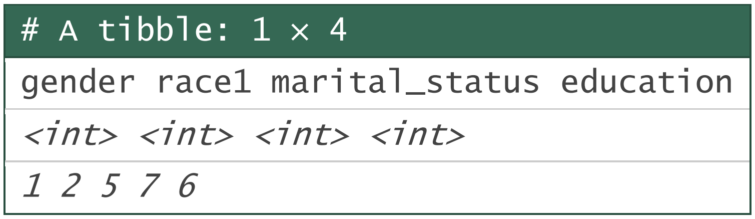
- Some physiological measures (`testosterone`, `bp_sys_ave`, `urine_vol1`) show selective collection, possibly only on a subsample due to cost or eligibility criteria.

- For downstream analysis, you may:

  - Drop variables with >90% missing values
  - Impute values for moderate gaps using domain-appropriate methods
  - Use complete-case or multiple imputation for modeling

Unique Values in Categorical Variables

```
# Count how many distinct values each categorical variable has
adult_data %>%
  select(gender, race1, marital_status, education) %>%
  summarise_all(~n_distinct(.))
```



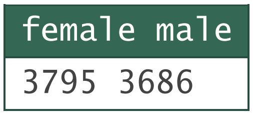
```
# A tibble: 1 × 4
gender race1 marital_status education
<int> <int> <int> <int>
1 2 5 7 6
```

```
# Output:
# gender: 2 categories (Male/Female)
# race1: 5 categories
# marital_status: 7 categories
# education: 6 categories
```

Gender Frequency

# Frequency distribution of gender

```
table(adult_data$gender)

> table(adult_data$gender)
```



female	male
3795	3686

# Output:

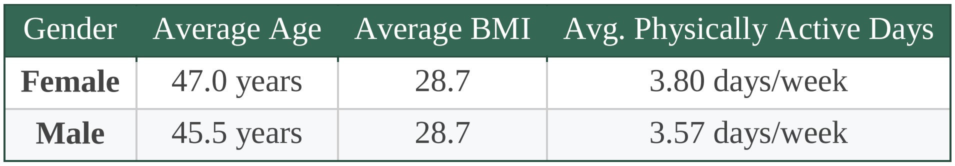
# Male   : 3686

# Female : 3795

Summary Statistics Grouped by Gender

```
# Calculate average age, BMI, and physical activity days by
gender

adult_data %>%

  group_by(gender) %>%

  summarise(

    avg_age = mean(age, na.rm = TRUE),

    avg_bmi = mean(bmi, na.rm = TRUE),

    avg_phys_active_days = mean(phys_active_days, na.rm = TRUE)

  )
```

Gender statistics



Gender	Average Age	Average BMI	Avg. Physically Active Days
Female	47.0 years	28.7	3.80 days/week
Male	45.5 years	28.7	3.57 days/week

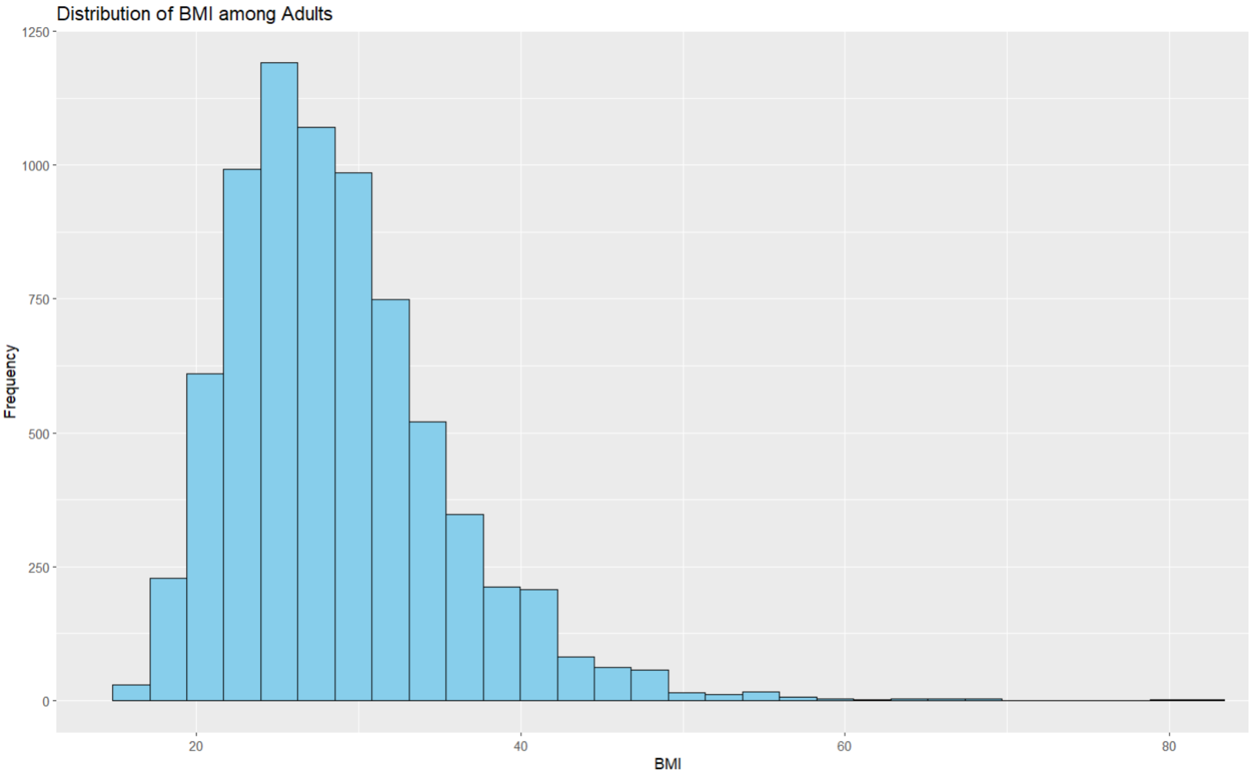
Interpretation

This table summarizes three key health-related metrics age, BMI, and physical activity levels grouped by gender from the adult NHANES dataset. On average, female participants are slightly older (47.0 years) than male participants (45.5 years). Interestingly, the average BMI is nearly identical for both genders at around 28.7, placing the sample in the overweight category according to standard BMI classifications.

Regarding physical activity, females report marginally more physically active days per week (3.80) compared to males (3.57). Though the difference is small, it may reflect gendered trends in daily movement or survey reporting. These figures provide an essential baseline for gender-based comparisons in health behavior and chronic disease risk.

Visualize BMI Distribution

```
# Histogram showing how BMI is distributed among adult
participants

ggplot(adult_data, aes(x = bmi)) +

  geom_histogram(fill = "skyblue", color = "black", bins = 30) +

  labs(

    title = "Distribution of BMI among Adults",

    x = "BMI",

    y = "Frequency"

  )
```



Interpretation

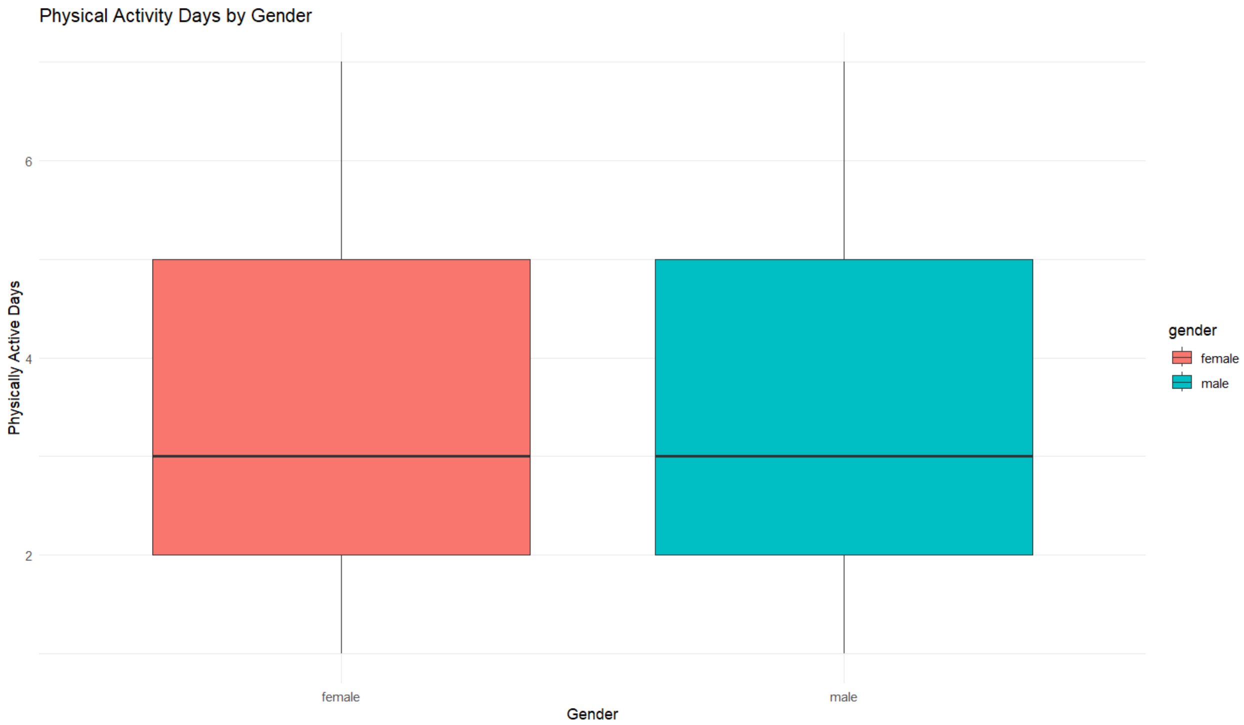
The histogram above visualizes the distribution of Body Mass Index (BMI) values in the adult population from the NHANES dataset. Here's what the graph reveals:

1. Shape of Distribution: The distribution is positively skewed (right-skewed), meaning most adults fall within lower BMI ranges, while fewer individuals fall into very high BMI categories.
2. Central Tendency: The most frequent BMI values cluster between 22 and 30, which correspond to the normal weight and overweight categories by standard BMI classification.
3. Mode and Peaks: The peak of the distribution appears around BMI 26–28, aligning with the average BMI values found in the dataset (both male and female averages were around 28.7).
4. Outliers and Extremes: There are a small number of adults with BMI values exceeding 40, extending up to 80, indicating the presence of individuals with severe obesity. These extreme cases are rare, as shown by the short bars on the right end of the graph.

5. Public Health Insight: The skew toward higher BMI values and the large concentration in the overweight and obesity range (BMI > 25) suggests a public health concern regarding weight-related conditions like diabetes, hypertension, and cardiovascular risks.
6. Data Quality: The smooth shape and large sample support the representativeness of the data. There are no abrupt gaps or anomalies, indicating clean and consistent data collection.

Visualization: Physical Activity by Gender

```
# This boxplot visualizes the distribution of physically active
days
# across gender categories among adults in the dataset.
# It highlights the median, interquartile range, and outliers.

ggplot(adult_data, aes(x = gender, y = phys_active_days, fill =
gender)) +
  geom_boxplot() +
  labs(
    title = "Physical Activity Days by Gender",
    x = "Gender",
    y = "Physically Active Days"
  ) +
  theme_minimal()
```



The boxplot comparing physically active days per week by gender shows a nearly identical distribution for both males and females, with a shared median of approximately 3 days. The interquartile range (IQR) spans from 2 to 5 days for both groups, indicating that the middle 50% of adults, regardless of gender, engage in physical activity within this range. This visual similarity suggests no significant gender difference in weekly physical activity levels within the sample. While earlier summaries showed a slight numerical edge for females (3.80 days vs. 3.57 days for males), this difference is not visually prominent, reinforcing the conclusion that adult men and women in the NHANES dataset report comparable levels of weekly physical activity.

Chapter Summary: Introduction to Medical Data Analysis Using NHANES

This chapter introduces the foundational steps for working with medical data using the NHANES dataset in R. It demonstrates the complete process from loading and preparing data to generating initial descriptive statistics and basic visualizations. The aim is to help medical professionals become confident in handling real-world health data and drawing preliminary insights using R programming.

To start, the chapter guides the user in loading essential libraries including tidyverse for data manipulation, janitor for cleaning column names, and NHANES, which provides access to the dataset itself. The NHANES dataset is a rich, publicly available source containing detailed health and demographic information from a representative sample of the U.S. population.

Once the data is loaded and cleaned, summary tables are used to explore core variables such as age, gender, education level, income, housing conditions, BMI, and physical measurements. Each summary output is explained with contextual interpretation to aid understanding. Special emphasis is given to missing values and their distribution, encouraging readers to consider data completeness before analysis.

Filtering is demonstrated to isolate adult participants (age ≥ 18), ensuring that subsequent insights are relevant to adult populations. The chapter then moves into examining categorical diversity in variables like gender, race, education, and marital status. Frequencies are calculated and group-wise summaries are performed to compare average age, BMI, and physical activity levels by gender.

Visual exploration is introduced through histograms and boxplots. A histogram visualizes the distribution of BMI among adults, while a boxplot compares physical activity days between males and females. These visual tools are essential for detecting patterns, variability, and outliers in health-related metrics.

Throughout the chapter, every code segment is paired with clear, real-world interpretation, reinforcing both programming fluency and statistical intuition. The structured walkthrough ensures readers are not only replicating code but understanding its application in a healthcare context.

Module 2:

Exploratory Data Analysis in Medical Context

This module demonstrates how to perform an initial exploratory data analysis (EDA) on medical datasets using the NHANES health survey in R. EDA is a crucial step in any data project as it helps researchers understand the structure, distribution, and quality of their dataset before applying advanced statistical models.

We begin by loading the necessary libraries and inspecting the data. The NHANES dataset is then checked for missing values, particularly focusing on the variables `Gender` and `BMI`, which are central to many health-related inquiries. The dataset is cleaned by removing records with missing values for these variables, ensuring that any downstream analysis is based on complete cases.

Next, we compute key summary statistics such as mean, median, and standard deviation of BMI, grouped by gender. This provides insight into body composition differences across male and female participants. If age category data (`AgeDecade`) is present, we further stratify the BMI distribution across both gender and age brackets, helping to identify patterns and health risks across age groups.

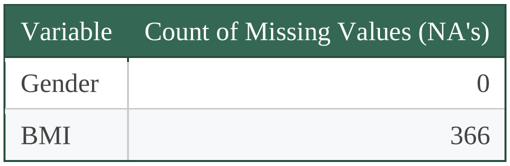
Finally, we visualize the BMI distribution using a gender-differentiated histogram. This gives a clear comparative picture of how BMI values spread among males and females, facilitating a better grasp of weight-related trends in population subgroups.

```
# Load the essential libraries

library(tidyverse)  # For data manipulation and visualization

library(NHANES)     # Provides access to NHANES health-related dataset

# Load and Inspect NHANES Dataset


# Load the NHANES dataset into R environment

data("NHANES")
```

```
# View structure and first-level summary of the dataset
glimpse(NHANES)
summary(NHANES)
```

The above codes are repeated steps which will be there in all modules for reaching out to please consult chapter 1.

```
# Check for Missing Values in Key Variables


# Check how many NA values exist in Gender and BMI
colSums(is.na(NHANES[, c("Gender", "BMI")]))
```



Variable	Count of Missing Values (NA's)
Gender	0
BMI	366

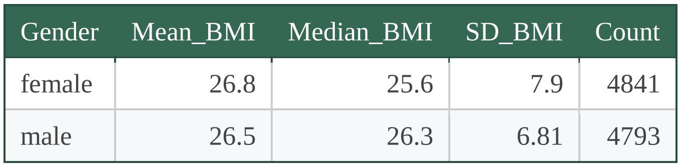
Interpretation

This means that all respondents have their gender recorded, ensuring that gender-based analysis (e.g., comparing health outcomes by gender) can be conducted on the full dataset. However, 366 individuals are missing BMI values, which is a key health metric used to assess weight status. This missingness (~4.9% if based on 7481 records) could affect statistical accuracy in analyses involving BMI (e.g., calculating average BMI or classifying individuals into weight categories). Depending on the context, these BMI-missing entries may need to be excluded or imputed to maintain analytical reliability.

```
#Clean the Dataset: Remove NA entries


# Filter out rows with missing Gender or BMI values
nhanes_clean <- NHANES %>%
  filter(!is.na(Gender) & !is.na(BMI))
#  Summary Statistics Grouped by Gender
```

Compute BMI statistics for each gender group

```
nhanes_clean %>%
  group_by(Gender) %>%
  summarise(
    Mean_BMI = mean(BMI, na.rm = TRUE),
    Median_BMI = median(BMI, na.rm = TRUE),
    SD_BMI = sd(BMI, na.rm = TRUE),
    Count = n()
  )
```



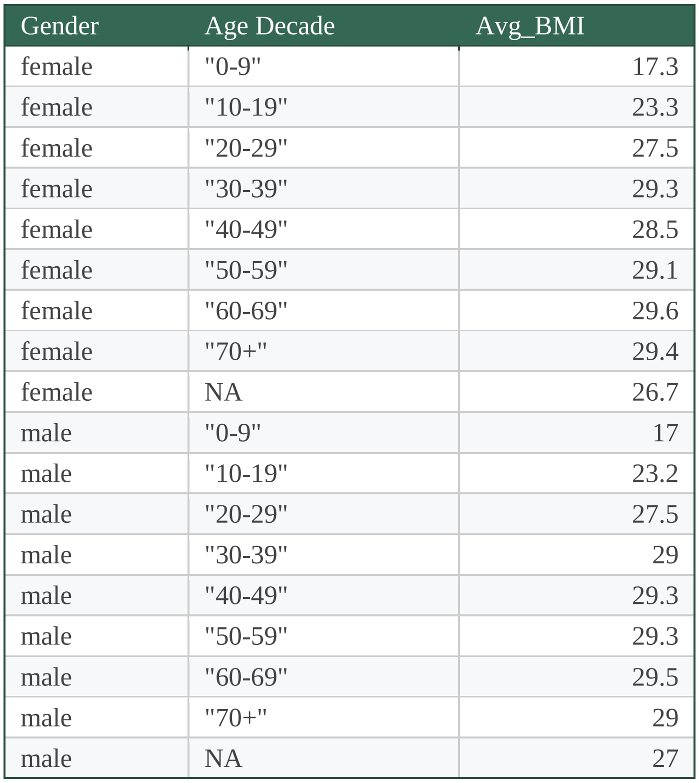
Gender	Mean_BMI	Median_BMI	SD_BMI	Count
female	26.8	25.6	7.9	4841
male	26.5	26.3	6.81	4793

Interpretation

Female participants have a slightly higher average BMI (26.8) compared to males (26.5). However, when looking at the median BMI, males actually report a slightly higher median (26.3) than females (25.6), suggesting that the female BMI distribution may be slightly skewed to the right (i.e., more high-BMI outliers). The standard deviation (SD) of BMI is also greater for females (7.90) than for males (6.81), indicating more variability in BMI among women. The sample sizes are similar, with 4,841 females and 4,793 males, ensuring that the comparison is statistically meaningful. Overall, while BMI averages are close, distribution characteristics reveal subtle but important differences between genders.

Extended Summary: Gender × AgeDecade

```
# Stratify BMI further by both Gender and AgeDecade (if
available)
if ("AgeDecade" %in% colnames(NHANES)) {
  nhanes_clean %>%
    group_by(Gender, AgeDecade) %>%
    summarise(Avg_BMI = mean(BMI, na.rm = TRUE), .groups =
"drop")
}
```



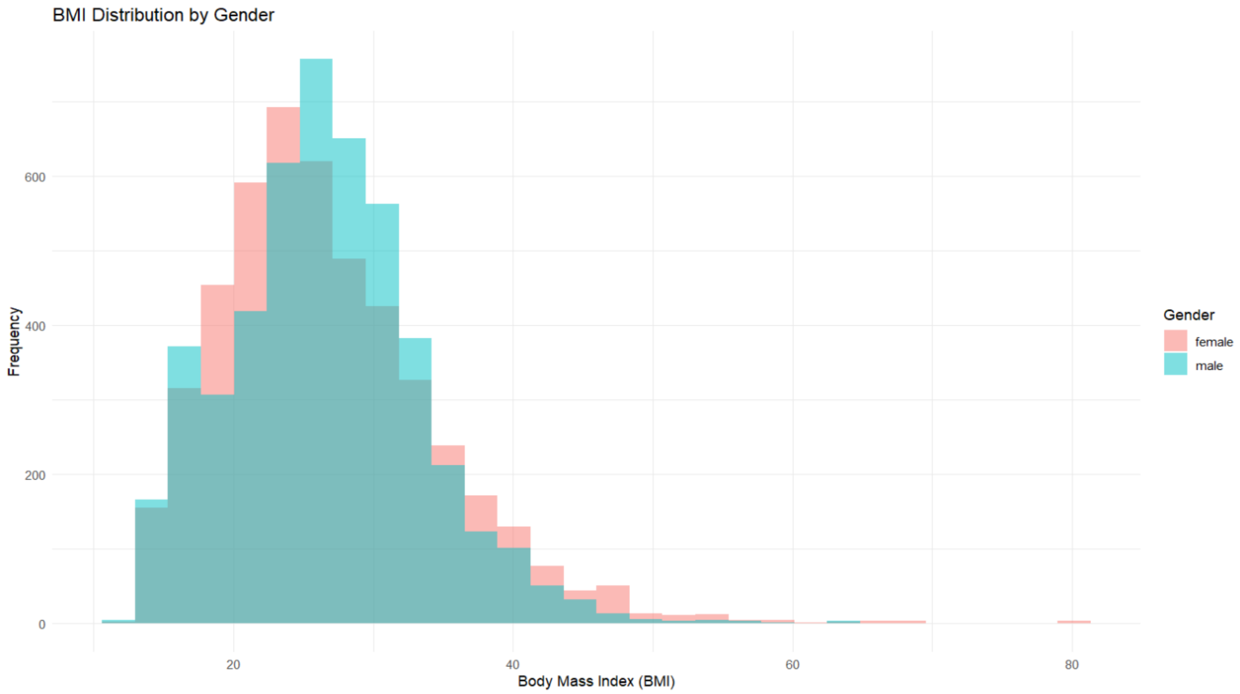
Gender	Age Decade	Avg_BMI
female	"0-9"	17.3
female	"10-19"	23.3
female	"20-29"	27.5
female	"30-39"	29.3
female	"40-49"	28.5
female	"50-59"	29.1
female	"60-69"	29.6
female	"70+"	29.4
female	NA	26.7
male	"0-9"	17
male	"10-19"	23.2
male	"20-29"	27.5
male	"30-39"	29
male	"40-49"	29.3
male	"50-59"	29.3
male	"60-69"	29.5
male	"70+"	29
male	NA	27

Interpretation

The table displays average Body Mass Index (BMI) across different age decades for both males and females, revealing consistent trends in weight status over the lifespan. For both genders, BMI starts low in childhood (around 17) and increases significantly during adolescence and early adulthood, peaking during middle age. Females reach their highest average BMI in the 30–39 and 60–69 age groups (29.3 and 29.6 respectively), while males show a steadier trend, maintaining average BMI values near 29 across ages 30 to 70+. A slight decline or plateau is observed in older age brackets, possibly reflecting weight loss or stabilization in later life. The data also include some missing age decade entries (NA), yet still provide an overall view of how BMI evolves with age, underscoring the influence of aging on body weight patterns.

Visualization: BMI Distribution by Gender

```
# Histogram showing BMI distribution with gender-based overlay
ggplot(nhanes_clean, aes(x = BMI, fill = Gender)) +
  geom_histogram(bins = 30, position = "identity", alpha = 0.5)
+
  labs(
    title = "BMI Distribution by Gender",
    x = "Body Mass Index (BMI)",
    y = "Frequency"
  ) +
  theme_minimal()
```

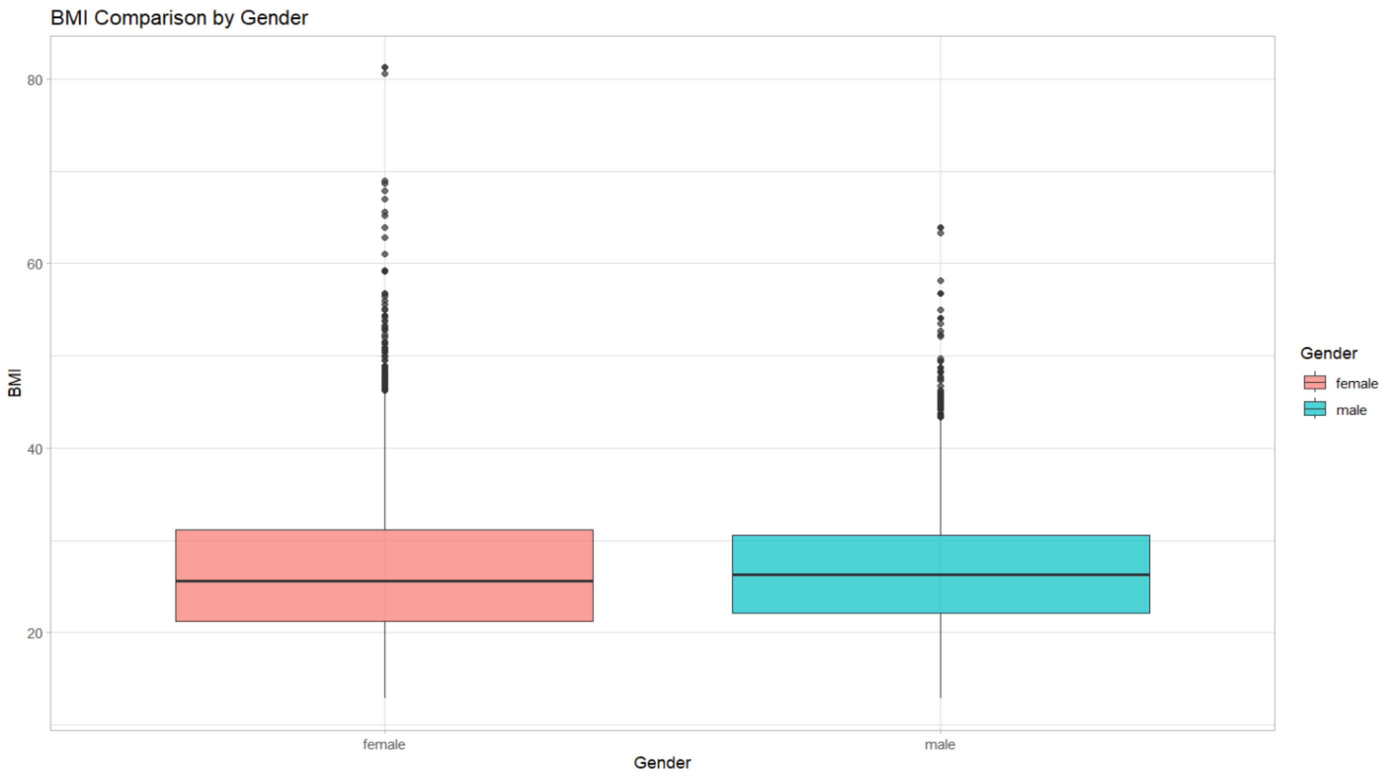


Interpretation

The histogram illustrates the distribution of Body Mass Index (BMI) across male and female participants. Both genders exhibit a right-skewed distribution, where the majority of individuals fall within the BMI range of approximately 20 to 30, indicating a concentration around the normal to overweight category. The peaks are most prominent between 25 and 30 BMI, suggesting this is the most common range for both males and females. There is a slight variation between genders, with females showing a higher frequency at slightly lower BMI values, while males dominate in the 27–30 range. The distribution tails off into the higher BMI values, with fewer individuals classified as obese or extremely obese. Overall, the visualization indicates that although BMI patterns are broadly similar between genders, minor differences in central tendency and spread exist.

Visualization: BMI Boxplot by Gender

```
# This boxplot provides a comparative visual summary of BMI
# between male and female participants in the cleaned dataset.
# It shows the median, interquartile range (IQR), and potential
outliers.

ggplot(nhanes_clean, aes(x = Gender, y = BMI, fill = Gender)) +
  geom_boxplot(alpha = 0.7) +
  labs(
    title = "BMI Comparison by Gender",
    x = "Gender",
    y = "BMI"
  ) +
  theme_light()
```



Interpretation

This boxplot visualizes the comparison of Body Mass Index (BMI) distributions between males and females. The central boxes represent the interquartile range (IQR), with the thick horizontal lines indicating the medians which appear slightly higher in females than males. Both genders have similar spreads, with slightly more variability among females. Notably, both groups exhibit a considerable number of outliers above the upper whisker, indicating a population of individuals with very high BMI (i.e., potentially obese or morbidly obese). The female group appears to have more high-end outliers than males, extending beyond 80 BMI, suggesting greater extreme values. Overall, while both genders have similar BMI ranges and medians, females show slightly higher values and more extreme outliers.

Module 3

Medical Data Wrangling

Introduction

Medical datasets, especially large survey data like NHANES, often contain multiple sections of interest lab results, demographics, physical measurements, and more. In this module, we focus on learning how to isolate relevant subsets of data and combine them using key identifiers such as `ID`. This process, called data wrangling, is foundational in preparing datasets for analysis, ensuring consistency and integration across multiple data sources.

The goal of this chapter is to help readers understand the importance of structured and purposeful data preparation. We demonstrate how to extract laboratory and demographic variables, merge them into a unified dataset, and inspect the resulting table for completeness and usability. Special emphasis is placed on handling missing data and understanding how reshaping data from wide to long format can unlock better flexibility in analysis and visualization.

In practical terms, the chapter guides through selecting specific variables such as BMI and total cholesterol (`TotChol`) from one subset, and age and gender from another. After checking for missing identifiers, these are joined on the `ID` field. We summarize and validate the merged dataset before reshaping it for further analysis. Finally, a boxplot is used to visualize the comparison of health indicators (BMI and cholesterol) across gender.

This module solidifies technical skills that are critical before any statistical testing or predictive modeling. Clean, organized, and well-structured data is not just a best practice it is a requirement for reliable analysis.

Load Required Libraries

```
library(tidyverse)  # Includes dplyr, tidyr, ggplot2 for
wrangling and plotting

library(NHANES)     # NHANES dataset for health-related data
```

```
#  Load NHANES Dataset
# Load dataset into the environment
data("NHANES")
```

Check dataset structure

```
glimpse(NHANES)
```

Due to huge repetition of the same output for output kindly consult chapter 1.

```
#  Select Laboratory Variables (e.g., BMI and TotChol)


lab_data <- NHANES %>%
  select(ID, BMI, TotChol)  # TotChol: Total Cholesterol
```

This code will run in the background store their result for next steps.

Select Demographic Variables (e.g., Age and Gender)

```
demo_data <- NHANES %>%
  select(ID, Age, Gender)
```

Check for Missing IDs

```
sum(is.na(lab_data$ID))
it is 0 and  # Should return 0
sum(is.na(demo_data$ID))
it is 0 and    # Should return 0
```

Merge Lab and Demographic Data

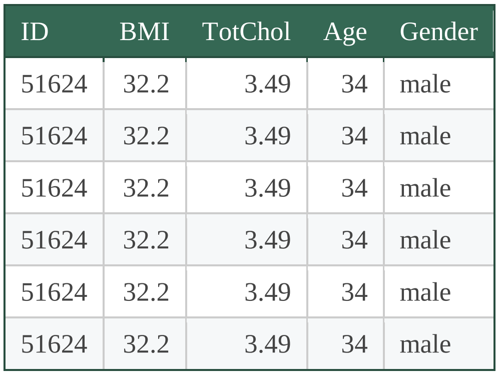
```
merged_data <- left_join(lab_data, demo_data, by = "ID")
```

Detected an unexpected many-to-many relationship between `x` and `y`.

i Row 1 of `x` matches multiple rows in `y`.

```
ℹ Row 1 of `y` matches multiple rows in `x`.
ℹ If a many-to-many relationship is expected, set `relationship = "many-to-many"` to silence
  this warning.
```

```
# View first few rows
head(merged_data)
```



ID	BMI	TotChol	Age	Gender
51624	32.2	3.49	34	male
51624	32.2	3.49	34	male
51624	32.2	3.49	34	male
51624	32.2	3.49	34	male
51624	32.2	3.49	34	male
51624	32.2	3.49	34	male

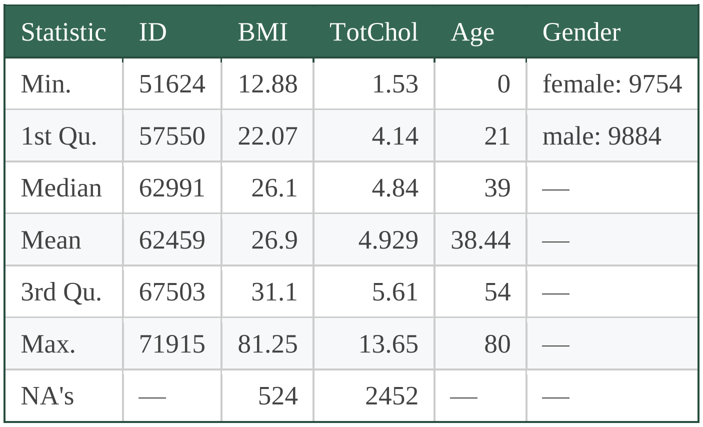
Interpretation

The displayed tibble (table) shows the first six rows of a merged dataset (`merged_data`) containing five variables: ID, BMI, TotChol (Total Cholesterol), Age, and Gender. Each row represents an individual observation, and in this sample, all six records belong to the same individual (ID: 51624), a 34-year-old male. The BMI for this individual is 32.2, placing him in the obese category, and his total cholesterol (TotChol) is 3.49, likely measured in mmol/L, which falls within a normal range for total cholesterol. The repeated rows suggest either duplication in the merging process or that multiple entries were created due to joining data across several observations or tests for the same person. Further data cleaning may be required to remove redundancy if these are true duplicates.

Summarize Merged Dataset

```
summary(merged_data)
```

Summarize Merged Dataset



Statistic	ID	BMI	TotChol	Age	Gender
Min.	51624	12.88	1.53	0	female: 9754
1st Qu.	57550	22.07	4.14	21	male: 9884
Median	62991	26.1	4.84	39	—
Mean	62459	26.9	4.929	38.44	—
3rd Qu.	67503	31.1	5.61	54	—
Max.	71915	81.25	13.65	80	—
NA's	—	524	2452	—	—

Interpretation

This summary of the `merged_data` dataset provides key descriptive statistics for five variables: ID, BMI, TotChol (Total Cholesterol), Age, and Gender. There are 19,638 entries in total, split almost evenly between females (9,754) and males (9,884)**.**

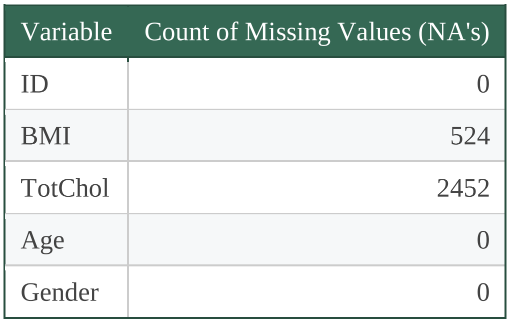
- BMI ranges from 12.88 to 81.25, with a mean of 26.9 and median of 26.1**,** placing the average individual in the overweight category. There are 524 missing BMI values**.**
- Total Cholesterol (TotChol) ranges from 1.53 to 13.65 mmol/L, with a mean of 4.93 and median of 4.84, suggesting most participants fall within normal or slightly elevated cholesterol levels. However, there are 2,452 missing values.
- Age spans from 0 to 80 years, with a mean of 38.44 years, a median of 39, and quartiles indicating a fairly even distribution across adult age ranges.

Overall, the dataset is fairly balanced in terms of gender and covers a wide range of health indicators and age groups, though the presence of missing data warrants preprocessing for accurate analysis.

Check for Missing Values Post-Merge

```
colSums(is.na(merged_data))
```

Missing Values Post-Merge



Variable	Count of Missing Values (NA's)
ID	0
BMI	524
TotChol	2452
Age	0
Gender	0

Interpretation

The output displays the number of missing values (`NA`s) in the `merged_data` dataset for five key variables. It shows that:

- BMI has 524 missing values**,**
- TotChol (Total Cholesterol) has 2,452 missing values**,**
- ID, Age, and Gender have no missing values**.**

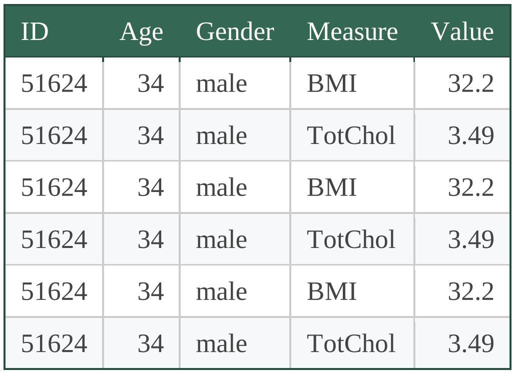
This indicates that the dataset is complete for demographic identification and age/gender variables but has notable missingness in health-related variables like BMI and cholesterol. These gaps may impact statistical analysis or model accuracy and will need to be addressed through imputation or filtering before further analysis.

Reshape Data: Wide to Long Format

```
long_data <- pivot_longer(
  merged_data,
  cols = c(BMI, TotChol),
  names_to = "Measure",
  values_to = "Value"
)
```

Preview reshaped data

```
head(long_data)
```



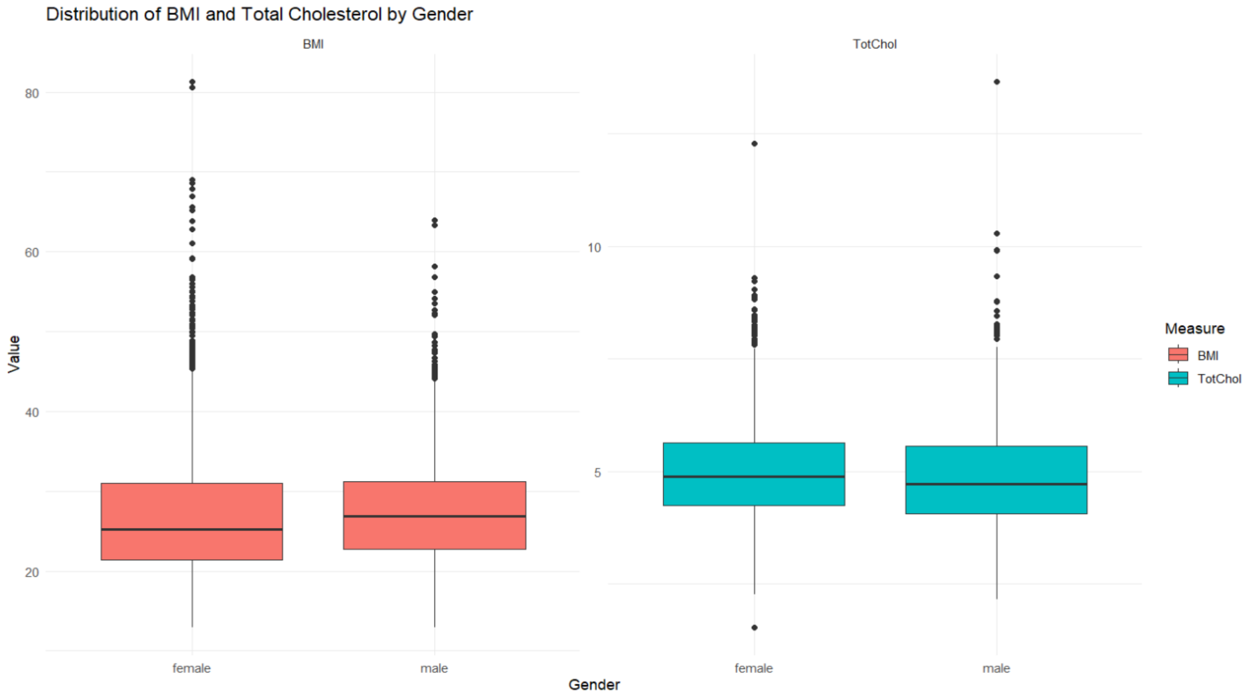
ID	Age	Gender	Measure	Value
51624	34	male	BMI	32.2
51624	34	male	TotChol	3.49
51624	34	male	BMI	32.2
51624	34	male	TotChol	3.49
51624	34	male	BMI	32.2
51624	34	male	TotChol	3.49

Interpretation

The `long_data` table shown is in a tidy or long-format structure, where each row represents a single measurement per individual per variable. For instance, the same individual (ID 51624), aged 34 and male, has multiple rows—each capturing a different measurement such as BMI and Total Cholesterol (TotChol) with their corresponding values (e.g., BMI = 32.2 and TotChol = 3.49). This structure is particularly useful for visualizations and statistical modeling, as it simplifies grouping and summarizing operations across measurement types. The repetition across rows ensures that each combination of subject, measure, and value is clearly organized for analysis.

Visualization: Boxplot by Gender & Measure

```
ggplot(long_data, aes(x = Gender, y = Value, fill = Measure)) +
  geom_boxplot() +
  facet_wrap(~Measure, scales = "free_y") +
  labs(
    title = "Distribution of BMI and Total Cholesterol by
Gender",
    x = "Gender",
    y = "Value"
  ) +
theme_minimal()`
```



Interpretation

The boxplot visualizes the distribution of BMI and Total Cholesterol (TotChol) values across genders. For BMI, both males and females show similar distributions with comparable medians and interquartile ranges, though females display slightly higher variability and more extreme outliers (values above 60 and 80). For Total Cholesterol, both genders also exhibit similar medians, but males show a slightly broader spread and more upper outliers (up to ~13). Overall, this dual-panel plot indicates that while the central tendencies of BMI and TotChol are similar between males and females, the variability and presence of high-value outliers are marginally greater in females for BMI and in males for TotChol.

Module 4

Biostatistics with R – Descriptive, Comparative & Associative Analysis

Introduction

In this module, we extend our cleaned and wrangled NHANES dataset into deeper statistical territory using core biostatistical methods. Medical data often needs more than visualization it requires clear statistical conclusions to support clinical decisions or public health insights. This module focuses on summarizing and comparing numeric data (like BMI), testing hypotheses, identifying group differences, and modeling associations.

We begin by performing descriptive statistics grouped by gender, providing insight into BMI trends across males and females. We then apply t-tests to compare the means between these two groups, assessing whether differences in BMI are statistically significant. ANOVA is introduced next to compare BMI across physical activity categories perfect for more than two-group comparisons.

Beyond comparisons, we introduce correlation analysis to quantify relationships between continuous variables like age and BMI. Next, we explore both simple linear regression (predicting BMI from age) and logistic regression (predicting diabetes status using age and BMI). This introduces the reader to predictive and inferential statistics simultaneously, using interpretable coefficients and odds ratios.

Finally, we conduct a chi-square test to explore categorical associations specifically between gender and diabetes prevalence. This mix of numeric, categorical, and regression techniques gives readers a robust statistical toolkit for medical data. Each method is complemented with clear outputs and interpretations, promoting statistical fluency.

Load Required Libraries

```
library(tidyverse)

library(NHANES)
```

```
# Load dataset
data("NHANES")
```

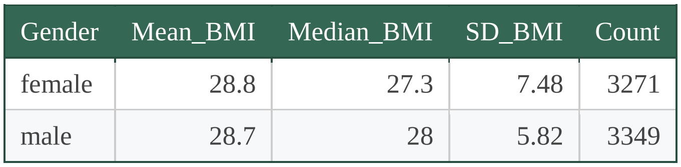
Clean Dataset for Analysis

```
# Filter dataset to remove rows with missing values in key variables
nhanes_clean <- NHANES %>%
  filter(
    !is.na(BMI),
    !is.na(Gender),
    !is.na(Diabetes),
    !is.na(Age),
    !is.na(Depressed),
    !is.na(PhysActive)
  )
```

Descriptive Statistics: BMI by Gender

```
nhanes_clean %>%
  group_by(Gender) %>%
  summarise(
    Mean_BMI = mean(BMI, na.rm = TRUE),
    Median_BMI = median(BMI, na.rm = TRUE),
    SD_BMI = sd(BMI, na.rm = TRUE),
    Count = n()
  )
```

Descriptive Statistics: BMI by Gender



Gender	Mean_BMI	Median_BMI	SD_BMI	Count
female	28.8	27.3	7.48	3271
male	28.7	28	5.82	3349

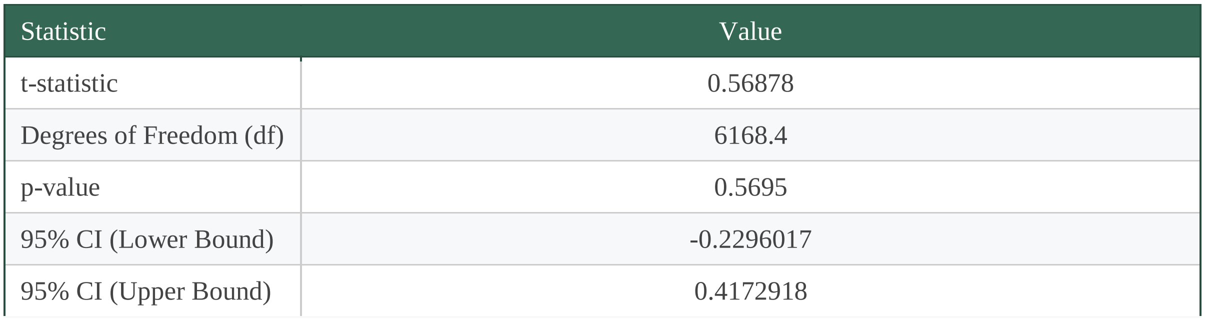
Interpretation

This table summarizes the BMI statistics by gender. The mean BMI is nearly identical between females (28.8) and males (28.7), indicating similar average body mass levels. However, the median BMI is slightly lower for females (27.3) than males (28.0), suggesting a slight left skew in the female BMI distribution. Notably, the standard deviation **(**SD**)** of BMI is higher in females (7.48) than in males (5.82), implying greater variability in BMI among women. The sample sizes are comparable, with 3,271 females and 3,349 males, supporting the reliability of this gender-based comparison. Overall, the data suggests BMI is similarly distributed across genders, but with more dispersion in the female group.

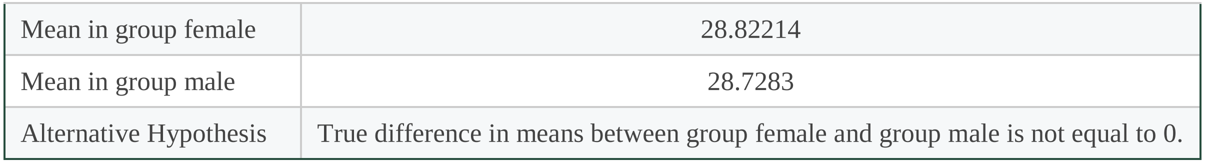
T-Test: BMI Differences by Gender

```
t_test_bmi <- t.test(BMI ~ Gender, data = nhanes_clean)
print(t_test_bmi)
```

T-Test



Statistic	Value
t-statistic	0.56878
Degrees of Freedom (df)	6168.4
p-value	0.5695
95% CI (Lower Bound)	-0.2296017
95% CI (Upper Bound)	0.4172918



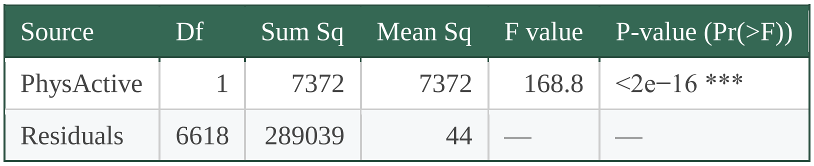
Mean in group female	28.82214
Mean in group male	28.7283
Alternative Hypothesis	True difference in means between group female and group male is not equal to 0.

Interpretation

The results of the Welch Two Sample t-test indicate that there is no statistically significant difference in mean BMI between males and females. The t-value is 0.56878 with degrees of freedom (df) around 6168.4, and a p-value of 0.5695, which is much higher than the conventional alpha level of 0.05. This high p-value suggests we fail to reject the null hypothesis, implying that any observed difference in mean BMI is likely due to random chance rather than a true difference. The 95% confidence interval for the difference in means ranges from -0.2296 to 0.4173, which includes 0, reinforcing the conclusion of no significant difference. The mean BMI for females (28.82) is only slightly higher than that of males (28.73), but the difference is not meaningful from a statistical standpoint.

ANOVA: BMI by Physical Activity Status

```
anova_result <- aov(BMI ~ PhysActive, data = nhanes_clean)
summary(anova_result)
```



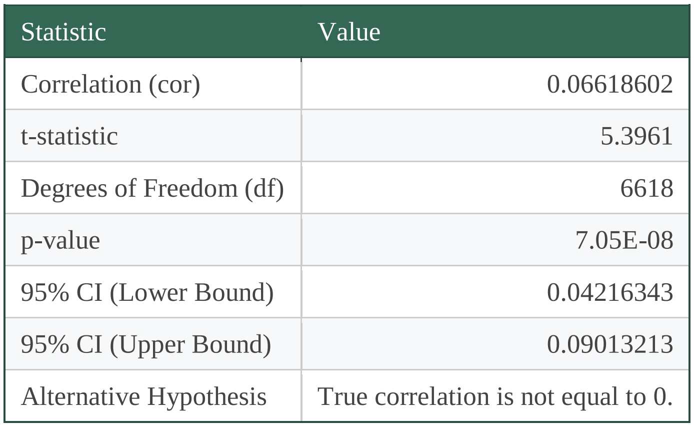
Source	Df	Sum Sq	Mean Sq	F value	P-value (Pr(>F))
PhysActive	1	7372	7372	168.8	<2e−16 ***
Residuals	6618	289039	44	—	—

Interpretation

The ANOVA results show a highly significant effect of physical activity (PhysActive) on the dependent variable (likely BMI, though it's not explicitly stated). The F-statistic is 168.8, with a p-value less than 2e-16, which is well below the typical significance threshold of 0.05. This extremely small p-value indicates strong evidence against the null hypothesis, meaning that physical activity levels are significantly associated with differences in the dependent variable. The sum of squares for physical activity (7372) reflects a substantial portion of variation explained compared to the residual variance (289039), highlighting its importance as a predictor in the model.

Correlation: Age vs. BMI

```
cor_test <- cor.test(nhanes_clean$Age, nhanes_clean$BMI)
print(cor_test)
```



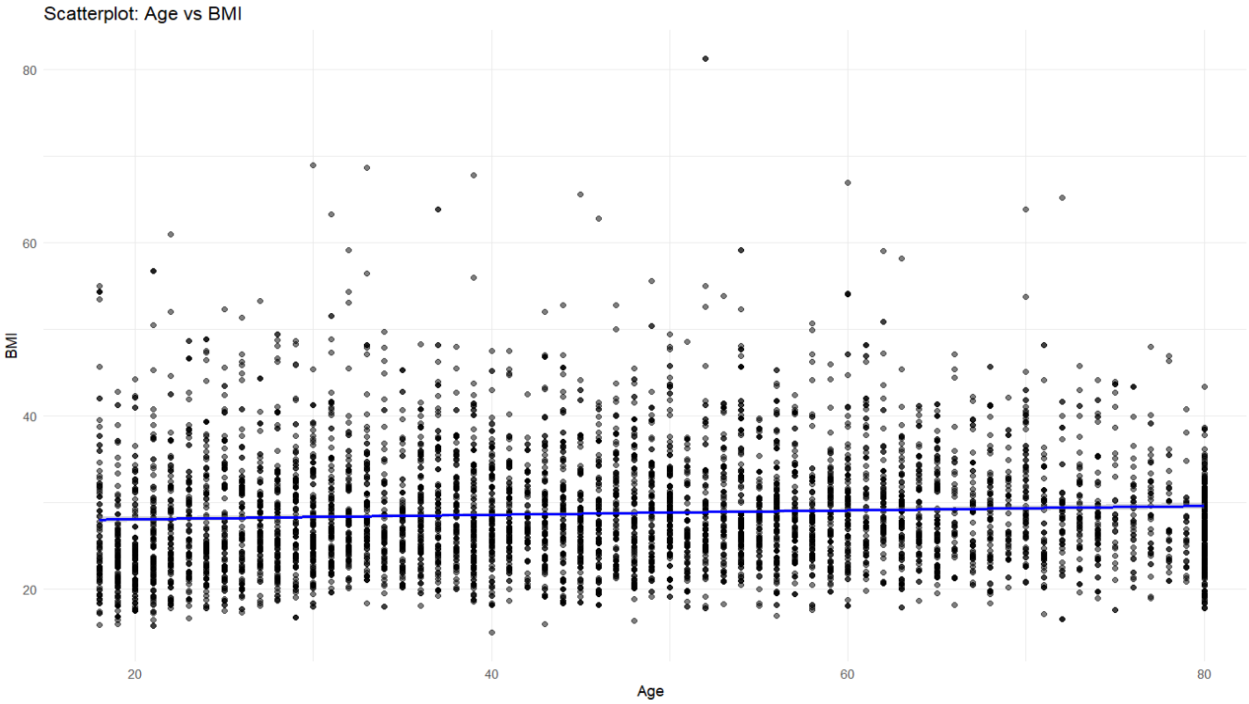
Statistic	Value
Correlation (cor)	0.06618602
t-statistic	5.3961
Degrees of Freedom (df)	6618
p-value	7.05E-08
95% CI (Lower Bound)	0.04216343
95% CI (Upper Bound)	0.09013213
Alternative Hypothesis	True correlation is not equal to 0.

Interpretation

The Pearson correlation test between Age and BMI yields a correlation coefficient of 0.066, with a p-value of 7.045e-08, which is statistically significant. This indicates that there is a very weak positive correlation between age and BMI meaning that as age increases, BMI tends to increase slightly. However, the effect size is minimal, as reflected in the small correlation value. The 95% confidence interval (0.042 to 0.090) further confirms that although the relationship is statistically reliable, it is not practically strong. In summary, age and BMI are weakly but significantly positively correlated**.**

Scatter Plot: Age vs. BMI with Regression Line

```
ggplot(nhanes_clean, aes(x = Age, y = BMI)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  labs(title = "Scatterplot: Age vs BMI", x = "Age", y = "BMI")
+
  theme_minimal()
```

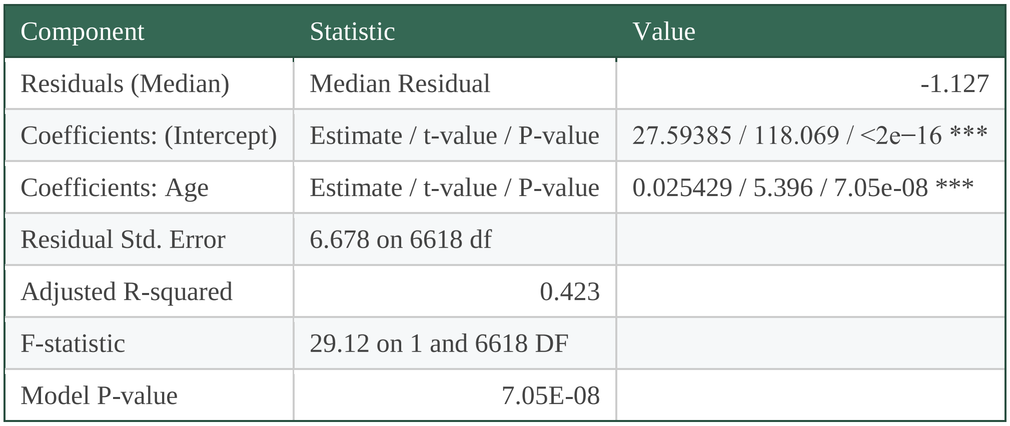


Interpretation

The scatterplot of Age vs. BMI shows a very slight upward trend, which is confirmed by the blue regression line indicating a weak positive linear relationship. Each dot represents an individual's BMI at a given age. The points are widely scattered, especially across the middle age ranges, showing high variability in BMI values across all age groups. Although the regression line increases slightly, the density of data points around it does not form a strong pattern, reinforcing the earlier statistical finding that age and BMI have a very weak positive correlation. Thus, while BMI tends to rise marginally with age, the relationship is minimal and not strongly predictive.

Simple Linear Regression: BMI ~ Age

```
lm_model <- lm(BMI ~ Age, data = nhanes_clean)
summary(lm_model)
```



Component	Statistic	Value
Residuals (Median)	Median Residual	-1.127
Coefficients: (Intercept)	Estimate / t-value / P-value	27.59385 / 118.069 / <2e−16 ***
Coefficients: Age	Estimate / t-value / P-value	0.025429 / 5.396 / 7.05e-08 ***
Residual Std. Error	6.678 on 6618 df	
Adjusted R-squared	0.423	
F-statistic	29.12 on 1 and 6618 DF	
Model P-value	7.05E-08	

Interpretation

The linear regression output assesses the relationship between Age and BMI using the `nhanes_clean` dataset. The model estimates that for each additional year of age, BMI increases by approximately 0.025 units (coefficient = 0.0254), with this effect being highly statistically significant (p-value = 7.05e-08). The intercept of 27.59 represents the estimated BMI at age zero, serving mainly as a model anchor. Although statistically significant, the model's explanatory power is extremely weak, as indicated by a very low R-squared value of 0.00438—suggesting that age accounts for less than 0.5% of the variance in BMI. Residual values range widely from -13.59 to 52.33, showing considerable variability in BMI not captured by age. In conclusion, while age has a statistically detectable impact on BMI, it has limited practical predictive value, and BMI is likely influenced by a broader set of factors beyond age.

Logistic Regression: Diabetes ~ Age + BMI

```
# Recode Diabetes as binary

nhanes_clean$Diabetes_bin <- ifelse(nhanes_clean$Diabetes ==
"Yes", 1, 0)


# Fit logistic model
```

```
logit_model <- glm(Diabetes_bin ~ Age + BMI, data =
nhanes_clean, family = binomial())

summary(logit_model)
```

Component	Statistic/Level	Value	Standard Error / DF	P-value (Pr(>|z|))
Coefficients	(Intercept)	−8.107143	0.280960	<2e-16 ***
	Age	0.058398	0.002860	<2e-16 ***
	BMI	0.093278	0.005992	<2e-16 ***
Model Fit	Null Deviance	4365.3	6619 df	—
	Residual Deviance	3642.9	6617 df	—
	AIC	3648.9	—	—

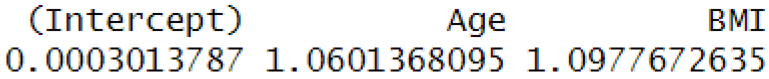
Interpretation

The logistic regression model presented investigates the effect of Age and BMI on the likelihood of having diabetes (a binary outcome, `Diabetes_bin`) using data from the `nhanes_clean` dataset. The model's coefficients show that both Age (estimate = 0.0539, p < 2e-16) and BMI (estimate = 0.0933, p < 2e-16) are highly statistically significant predictors of diabetes. The positive coefficients indicate that as age or BMI increases, the probability of having diabetes also increases.

The intercept value of -8.1071 represents the log-odds of diabetes when both age and BMI are zero, which is more of a theoretical baseline than a practical one. The residual deviance (3642.9) is substantially lower than the null deviance (4365.3), indicating that the model with predictors fits the data better than a model without them. The AIC value (3648.9) gives a general measure of model quality, useful for model comparison.

Overall, the model confirms that both age and BMI are strong and significant contributors to diabetes risk, aligning with established medical understanding.

```
# Get odds ratios
exp(coef(logit_model))
```



```
 (Intercept)          Age          BMI 
0.0003013787 1.0601368095 1.0977672635 
```

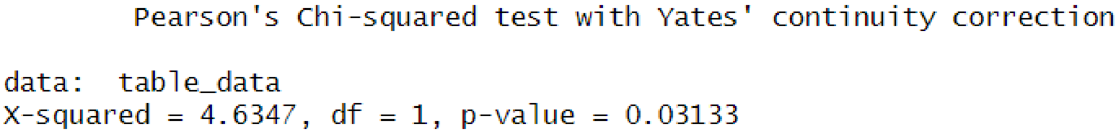
Interpretation

The values shown are the odds ratios obtained by exponentiating the coefficients of a logistic regression model predicting diabetes based on Age and BMI:

- **Intercept**: 0.0003 – This is the baseline odds of having diabetes when both Age and BMI are zero. While not meaningful in practical terms, it anchors the model.
- **Age**: 1.0601 – For every additional year of age, the odds of having diabetes increase by approximately 6%**,** holding BMI constant.
- **BMI**: 1.0978 – For each one-unit increase in BMI, the odds of having diabetes rise by approximately 9.8%**,** controlling for age.

In summary, both age and BMI have a positive and substantial impact on diabetes risk, with BMI being a slightly stronger predictor in terms of odds magnitude.

```
#  7. Chi-Square Test: Gender vs Diabetes


table_data <- table(nhanes_clean$Gender, nhanes_clean$Diabetes)
chi_result <- chisq.test(table_data)
print(chi_result)
```



```

	Pearson's Chi-squared test with Yates' continuity correction

data:  table_data
X-squared = 4.6347, df = 1, p-value = 0.03133
```

Interpretation

The result of the Pearson's Chi-squared test with Yates' continuity correction indicates a statistically significant association between the two categorical variables in the dataset. The test yielded a Chi-squared value of 4.6347 with 1 degree of freedom and a p-value of 0.03133. Since the p-value is less than the conventional threshold of 0.05, we reject the null hypothesis of independence. This suggests that there is a meaningful relationship between the variables under consideration. The application of Yates’ continuity correction, commonly used for 2x2 contingency tables, adjusts the Chi-squared value to reduce the likelihood of a Type I error in small sample sizes, thereby making the result more conservative.

Module 5

Epidemiological Risk Measures and Analysis

Introduction

Understanding epidemiological risk measures is vital for analyzing disease burden, comparing population groups, and evaluating preventive strategies. This module introduces the essential tools to measure association, prevalence, and relative risk using categorical health data from the NHANES dataset. These calculations are foundational in both clinical and public health decision-making.

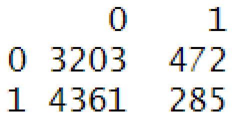
The module focuses on evaluating disease exposure and outcome relationships particularly the link between physical activity and diabetes. After preparing and recoding the necessary variables, we use contingency tables to calculate prevalence**,** risk ratios (RR)**,** and odds ratios (OR). These are crucial indicators for interpreting how much more (or less) likely a condition is to occur in an exposed group compared to a non-exposed one.

For clarity, physical activity (`PhysActive`) is used as the exposure variable, and diabetes status (`Diabetes`) as the outcome. The population is divided into active vs inactive groups, and the proportion of diabetics in each is calculated. This allows us to interpret whether being physically active is protective against diabetes.

These calculations are backed by contingency tables and proportional summaries. This module helps readers develop the ability to draw epidemiological conclusions with real-world relevance, strengthening both analytical and clinical reasoning skills.

```
#  Load Required Libraries
library(tidyverse)
library(NHANES)
#  Load and Clean Dataset
data("NHANES")
# Keep only rows with non-missing Diabetes and PhysActive
```

```
epi_data <- NHANES %>%
  filter(!is.na(Diabetes), !is.na(PhysActive))
# Recode Variables for Analysis
# Binary encoding: Diabetes (1 = Yes), PhysActive (1 = Yes)
epi_data <- epi_data %>%
  mutate(
    diabetes_bin = ifelse(Diabetes == "Yes", 1, 0),
    active_bin = ifelse(PhysActive == "Yes", 1, 0)
  )
#  Create 2x2 Contingency Table

# Rows = PhysActive, Columns = Diabetes
contingency_table <- table(epi_data$active_bin,
epi_data$diabetes_bin)
print(contingency_table)
```



```
      0    1
 0 3203  472
 1 4361  285
```

Interpretation

This 2x2 contingency table shows the frequency distribution of two categorical variables, likely representing binary outcomes such as diabetes status (0 = no diabetes, 1 = diabetes) across two groups (e.g., physical activity: 0 = inactive, 1 = active). From the table:

- Among individuals with value 0 in the first variable (possibly inactive), 3203 do not have the condition, while 472 do.
- Among individuals with value 1 in the first variable (possibly active), 4361 do not have the condition, while 285 do.

This suggests that a greater proportion of physically active individuals have no diabetes compared to inactive individuals. However, a higher number of inactive individuals have diabetes (472) than active ones (285). When considered alongside the previously reported Chi-squared test result (p-value = 0.03133), the data support a statistically significant association between the variables implying that physical activity and diabetes status are not independent and may be linked.

Calculate Prevalence in Each Group

```
# Prevalence = number of diabetics / total group size
prevalence_active <- contingency_table[2,2] /
sum(contingency_table[2,])
prevalence_inactive <- contingency_table[1,2] /
sum(contingency_table[1,])


prevalence_active
[1] 0.06134309

prevalence_inactive
[1] 0.1284354
```

Interpretation

The calculated prevalence indicate that physical activity is associated with a lower rate of diabetes. Specifically, among the physically active group, the prevalence of diabetes is 6.13%, whereas among the physically inactive group, the prevalence is significantly higher at 12.84%. This means that the proportion of individuals with diabetes is more than double in the inactive group compared to the active group. These findings support the idea that physical activity may have a protective effect against the development of diabetes.

Calculate Risk Ratio (Relative Risk)

```
# RR = Prevalence in Active / Prevalence in Inactive
risk_ratio <- prevalence_active / prevalence_inactive
risk_ratio
[1] 0.4776183
```

Interpretation

The calculated risk ratio (relative risk) is approximately 0.48, which means that individuals who are physically active have about 52% lower risk of having diabetes compared to those who are physically inactive. This value, being less than 1, indicates a protective association between physical activity and diabetes: engaging in physical activity appears to significantly reduce the likelihood of developing diabetes. This further supports the role of lifestyle behaviors, such as staying active, in diabetes prevention and overall metabolic health.

Calculate Odds Ratio (OR)

```
# OR = (a*d) / (b*c)
a <- contingency_table[2,2]
b <- contingency_table[2,1]
c <- contingency_table[1,2]
d <- contingency_table[1,1]

odds_ratio <- (a * d) / (b * c)
odds_ratio
[1] 0.4434797
```

Interpretation

The calculated odds ratio (OR) is approximately 0.44, indicating that the odds of having diabetes are 56% lower for physically active individuals compared to those who are inactive. Since the OR is less than 1, it suggests a strong inverse association between physical activity and diabetes status. In other words, being physically active significantly reduces the odds of developing diabetes, further reinforcing the importance of physical activity as a protective factor in health promotion and chronic disease prevention.

Module 6

Visualization and Reporting for Medical Insight

Introduction

Visualizing medical data is not just about aesthetics it's a core component of data interpretation, exploration, and communication. This module introduces practical techniques to create meaningful and publication-ready graphics using `ggplot2`. Through tailored visualizations, medical professionals can identify patterns, outliers, and group differences, supporting both clinical decision-making and research presentations.

We begin by creating boxplots, histograms, and bar charts to examine variable distributions and compare subgroups. Each visualization is selected for a specific purpose: histograms to understand numeric distribution, boxplots to compare across groups, and bar charts to summarize categorical proportions. These tools transform raw numbers into intuitive visuals, allowing insights to surface more easily.

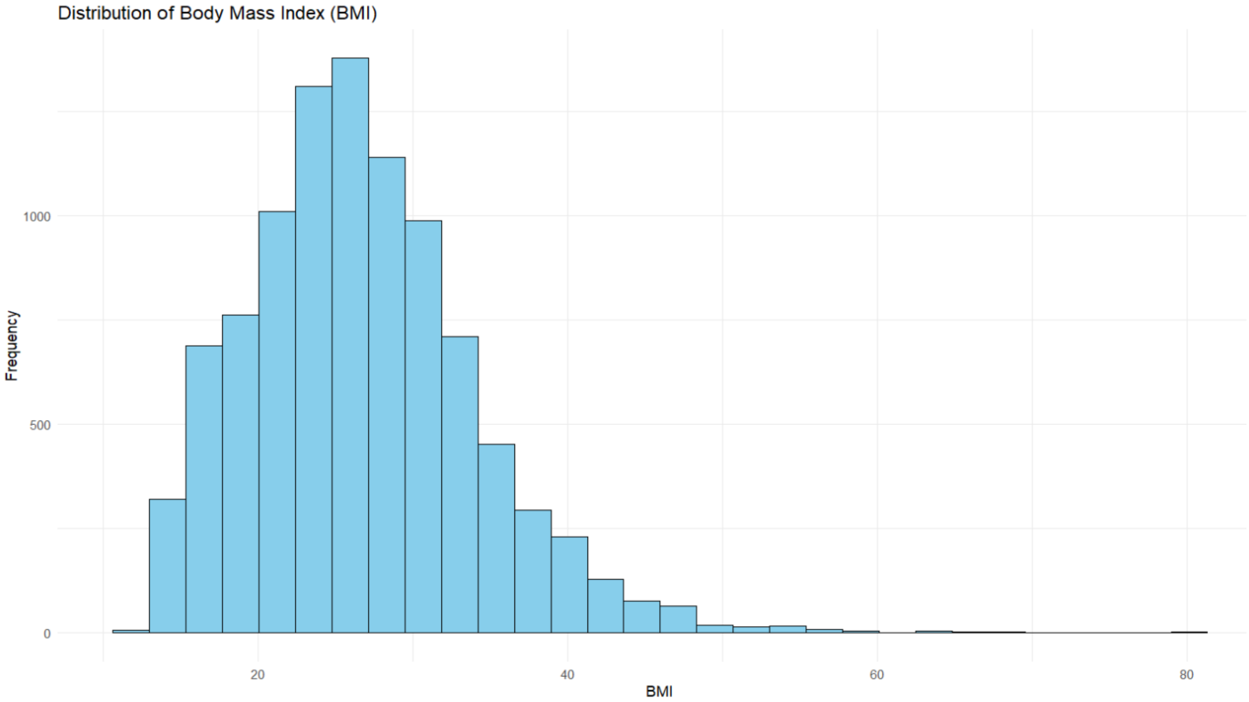
Beyond exploration, the module emphasizes how to annotate and style plots using custom labels, titles, themes, and color palettes. This ensures the plots are not only informative but also readable in reports, papers, or presentations. Combining multiple visuals, faceting, and overlaying regression lines can further clarify complex relationships.

Ultimately, the ability to visualize medical data empowers professionals to communicate evidence-based findings more effectively—bridging the gap between raw data and actionable healthcare insights.

Load Required Libraries

```
library(tidyverse)
library(NHANES)
data("NHANES")
```

```
# Histogram: Distribution of BMI
ggplot(NHANES, aes(x = BMI)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  labs(
    title = "Distribution of Body Mass Index (BMI)",
    x = "BMI",
    y = "Frequency"
  ) +
  theme_minimal()
```



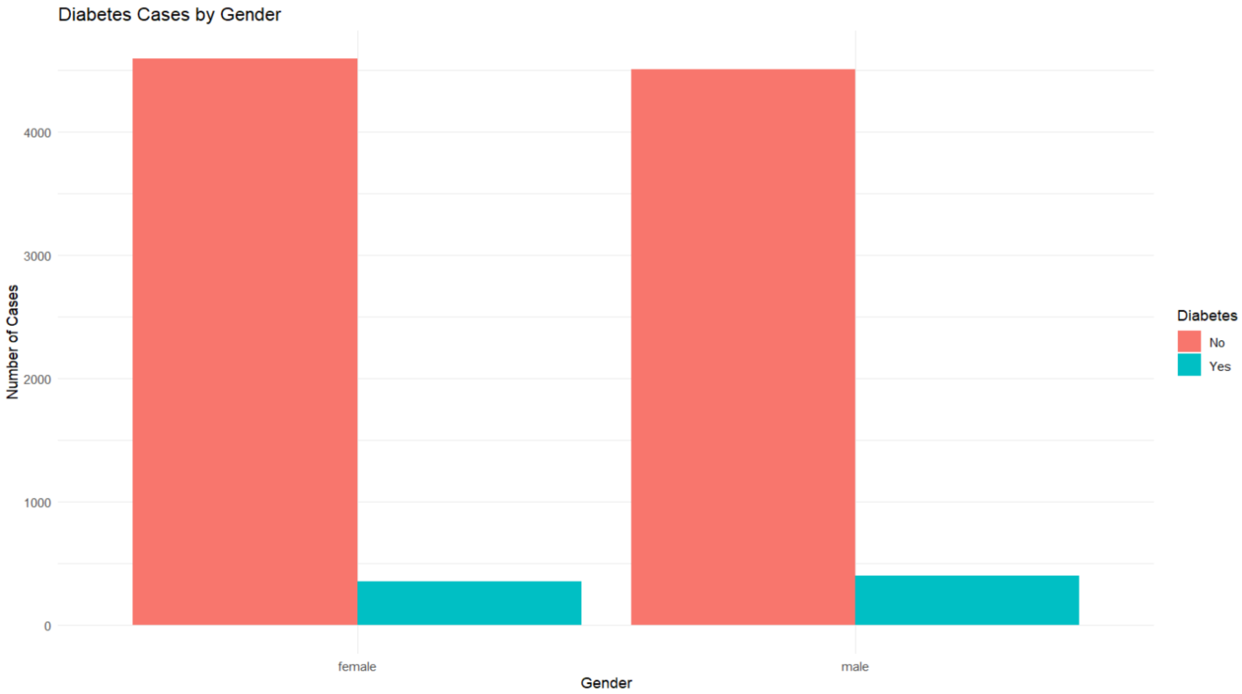
Interpretation

The histogram displays the distribution of Body Mass Index (BMI) values among the participants. The distribution is right-skewed, with the majority of individuals having a BMI between approximately 20 and 30, which includes the normal and overweight categories. The peak frequency is centered around a BMI of 25, suggesting that this is the most common value in the dataset. Frequencies decline gradually for higher BMI values, and there are fewer individuals with BMIs above 40, indicating that extremely high BMI values (obesity class III) are relatively rare.

The long right tail of the distribution reflects the presence of outliers or individuals with very high BMI values. Overall, the plot suggests that while most individuals fall within a healthy to moderately overweight range, a notable portion of the population exceeds the standard BMI threshold for obesity.

Bar Chart: Proportion of Diabetes by Gender

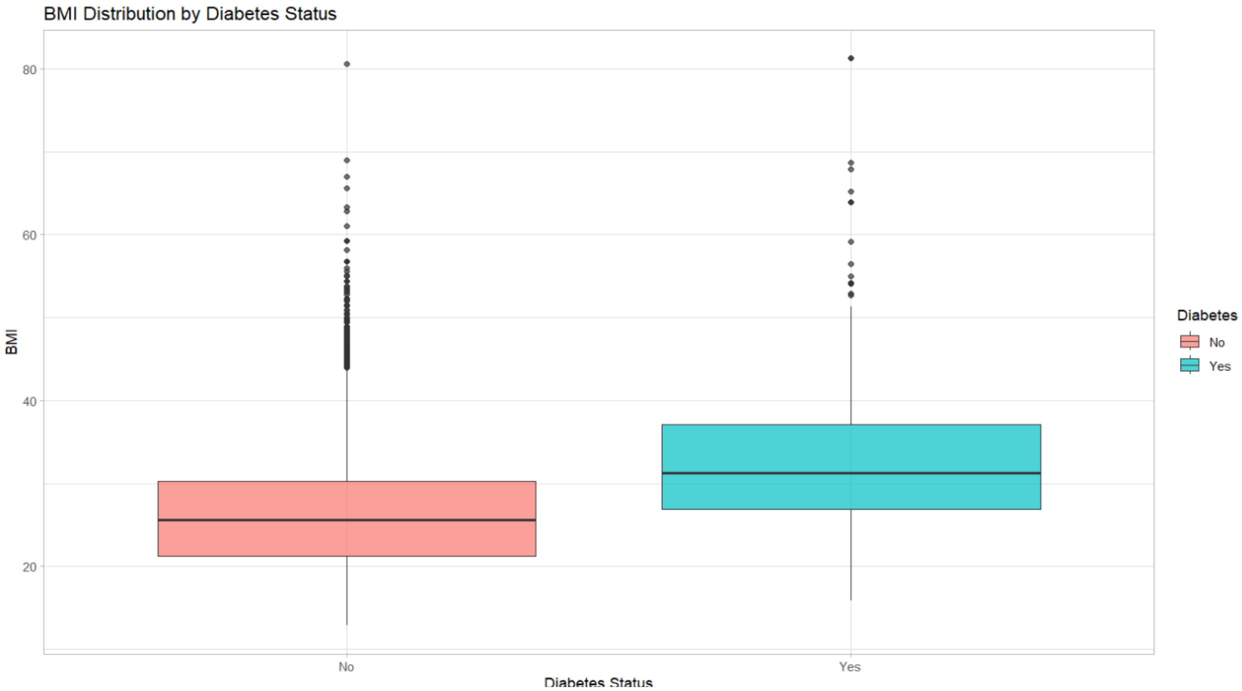
```
NHANES %>%
  filter(!is.na(Gender), !is.na(Diabetes)) %>%
  group_by(Gender, Diabetes) %>%
  summarise(Count = n(), .groups = "drop") %>%
  ggplot(aes(x = Gender, y = Count, fill = Diabetes)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "Diabetes Cases by Gender",
    x = "Gender",
    y = "Number of Cases"
  ) +
  theme_minimal()
```



Interpretation

The bar chart displays the number of diabetes cases segmented by gender, with two categories: individuals without diabetes (red bars) and those with diabetes (blue bars). Both females and males show a similar pattern: a significantly larger number of individuals are non-diabetic compared to those who have diabetes. However, males have a slightly higher number of diabetes cases than females, even though the total number of cases in both genders is relatively comparable. This suggests that while diabetes affects both genders, males might have a marginally higher prevalence in this sample population. Nonetheless, the overall gender difference is not large, and the chart emphasizes that diabetes is a concern across both groups.

```
#  Boxplot: BMI by Diabetes Status
NHANES %>%
  filter(!is.na(Diabetes), !is.na(BMI)) %>%
  ggplot(aes(x = Diabetes, y = BMI, fill = Diabetes)) +
  geom_boxplot(alpha = 0.7) +
  labs(
    title = "BMI Distribution by Diabetes Status",
    x = "Diabetes Status",
    y = "BMI"
  ) +
  theme_light()
```

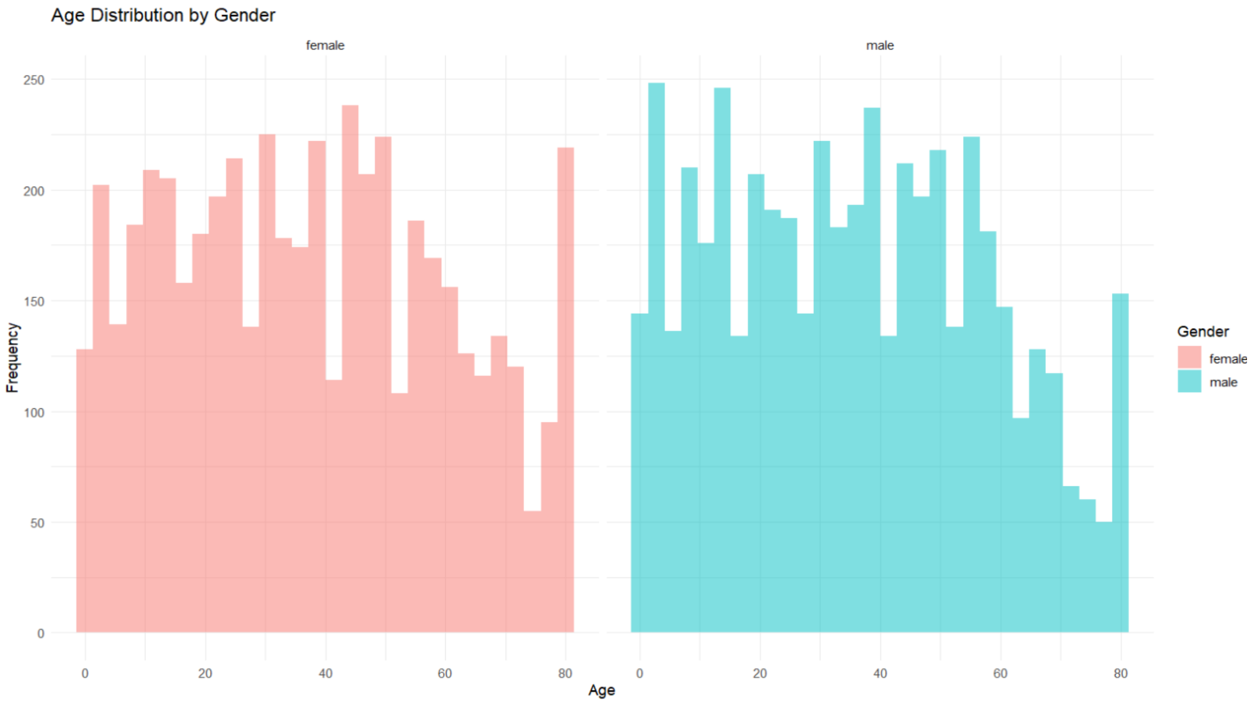


Interpretation

The boxplot illustrates the distribution of Body Mass Index (BMI) values across individuals with and without diabetes. It is evident that individuals with diabetes tend to have higher BMI values compared to those without diabetes. The median BMI for diabetics is substantially above that of non-diabetics, and the entire interquartile range (middle 50% of data) is shifted upwards for the diabetic group. Additionally, the diabetic group shows more BMI variability and a higher upper range, suggesting that obesity is more prevalent among individuals with diabetes. This visual reinforces the well-established association between higher BMI and increased risk of developing diabetes.

Faceted Histograms: Age Distribution by Gender

```
NHANES %>%
  filter(!is.na(Age), !is.na(Gender)) %>%
  ggplot(aes(x = Age, fill = Gender)) +
  geom_histogram(bins = 30, alpha = 0.5, position = "identity")
+
  facet_wrap(~Gender) +
  labs(
    title = "Age Distribution by Gender",
    x = "Age",
    y = "Frequency"
  ) +
  theme_minimal()
```

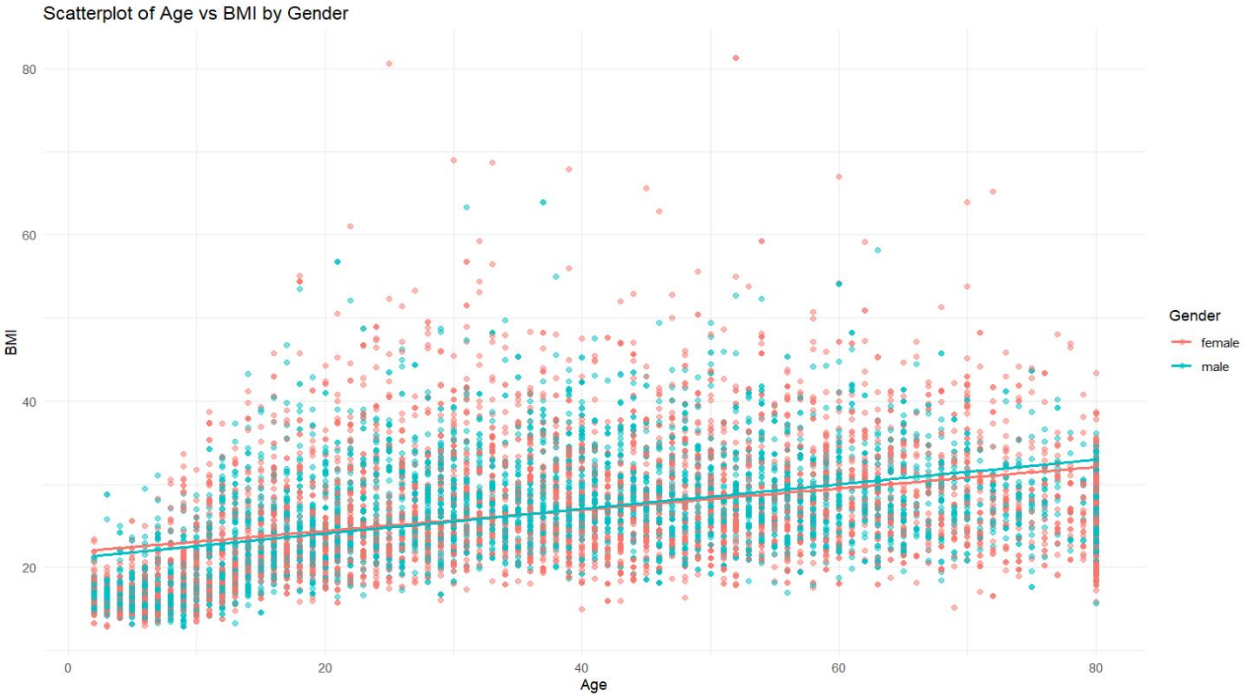


Interpretation

The histogram displays the age distribution by gender, with females shown in pink and males in blue. Both distributions span from early childhood to around age 80, with relatively even coverage across age groups. The frequency counts show some minor fluctuations, but both genders appear to have fairly consistent representation throughout adulthood, especially from ages 20 to 60. There is a slight drop in frequency for both genders beyond age 70, reflecting lower representation of older individuals, which is typical in population samples. The balanced distribution suggests that the dataset includes a wide and comparable age range for both male and female participants, supporting reliable age-based analyses.

Scatterplot: Age vs BMI with Gender Color

```
NHANES %>%
  filter(!is.na(Age), !is.na(BMI), !is.na(Gender)) %>%
  ggplot(aes(x = Age, y = BMI, color = Gender)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Scatterplot of Age vs BMI by Gender",
    x = "Age",
    y = "BMI"
  ) +
  theme_minimal()
```



Interpretation

This scatterplot shows the relationship between age and BMI, separated by gender (pink for females and cyan for males), with a fitted regression line overlaid for each gender. Overall, there is a slight upward trend in BMI as age increases for both males and females, indicated by the positive slope of the trend lines. However, the relationship is weak, as data points are widely dispersed, especially in middle and older age groups, suggesting high variability in BMI at all ages.

The BMI range for both genders spans from below 15 to over 80, though most values cluster between 20 and 40. Females and males exhibit similar patterns, with no significant visual gender disparity in BMI across age groups. This confirms that while BMI tends to increase slightly with age, it does so with considerable individual variation.

Module 7

Time Series and Longitudinal Analysis in Health Studies

Introduction

Medical data collected across time whether through national surveys, patient follow-ups, or health monitoring requires specialized tools for time-based analysis. This module introduces the foundational steps for conducting time series and longitudinal analysis using simplified simulated data due to NHANES's cross-sectional nature.

Though NHANES doesn't track the same individuals across multiple years, it includes multiple survey cycles (e.g., 2009-2010, 2011-2012, etc.) which can be leveraged for pseudo-longitudinal analysis. This module simulates individual-level follow-ups to demonstrate the essential logic and functions needed in real longitudinal datasets, such as those from cohort or clinical studies.

We create a sample dataset representing repeated BMI measurements for a set of individuals across three years. Using this structure, we demonstrate how to reshape data into long format for time-aware plotting, compute per-person trends, and apply linear mixed-effects models (LMMs) to account for within-subject correlation over time.

This foundational knowledge is critical for real-world health research involving patient follow-ups, disease progression tracking, or treatment efficacy monitoring over time.

Load Required Libraries

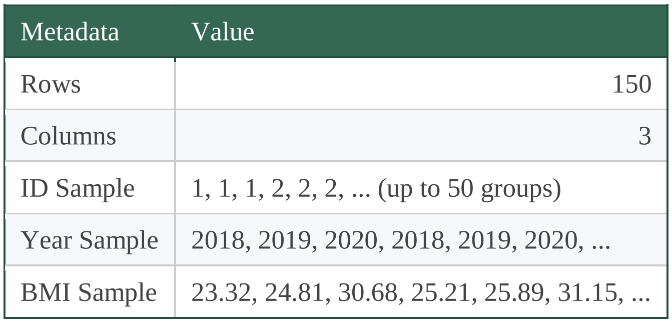
```
library(tidyverse)
library(lme4)      # For linear mixed models
library(ggplot2)
library(reshape2)
```

Simulated Longitudinal Health Data

```
# Simulate repeated BMI data for 50 individuals across 3 years
```

```
set.seed(123)
long_data <- data.frame(
  ID = rep(1:50, each = 3),
  Year = rep(c(2018, 2019, 2020), times = 50),
  BMI = round(rnorm(150, mean = 25, sd = 3) + rep(c(0, 0.5, 1),
times = 50), 2)
)
#  Reshape and Explore Long Format
# Already in long format; confirm with structure
glimpse(long_data)
```

Long Data



Metadata	Value
Rows	150
Columns	3
ID Sample	1, 1, 1, 2, 2, 2, ... (up to 50 groups)
Year Sample	2018, 2019, 2020, 2018, 2019, 2020, ...
BMI Sample	23.32, 24.81, 30.68, 25.21, 25.89, 31.15, ...

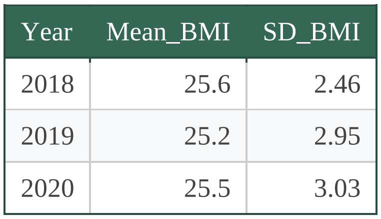
Interpretation

This table represents a reshaped dataset in long format, where each row corresponds to a single observation of BMI for a person (ID) in a given year (Year). Instead of having one row per person with multiple BMI columns for different years (wide format), the data has been transformed so that all BMI values are in a single column, making it suitable for longitudinal analysis and functions like ggplot, lme, or group_by() with summarise() in tidyverse workflows.

Summary statistics per year

```
long_data %>%
  group_by(Year) %>%
  summarise(
    Mean_BMI = mean(BMI),
    SD_BMI = sd(BMI)
  )
```

Summary statistics



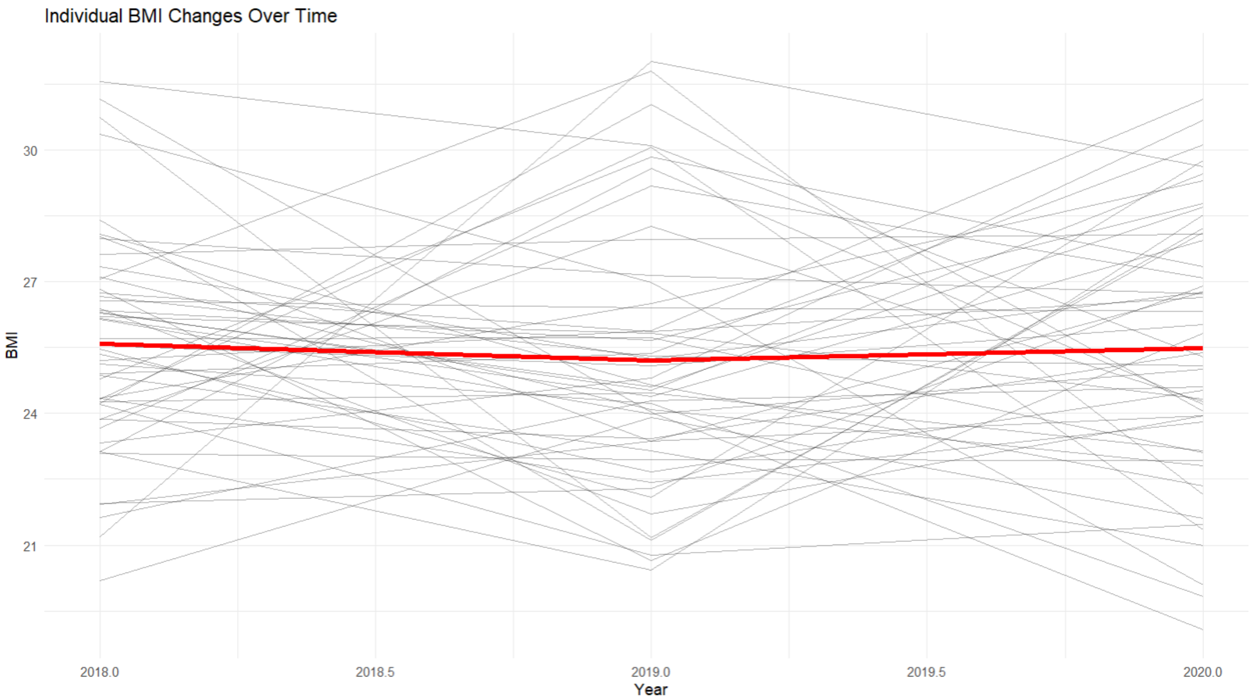
Year	Mean_BMI	SD_BMI
2018	25.6	2.46
2019	25.2	2.95
2020	25.5	3.03

Interpretation

This table summarizes BMI trends across the years 2018 to 2020. The mean BMI shows relatively small fluctuations: it was 25.6 in 2018, dipped slightly to 25.2 in 2019, and rose again to 25.5 in 2020. Meanwhile, the standard deviation (SD_BMI**)** increased gradually over the same period (from 2.46 to 3.03), indicating that while average BMI remained fairly stable, the variation in BMI values among individuals grew slightly from year to year. This could suggest an increasing disparity in body mass among the population over time.

Line Plot: Individual BMI Trajectories Over Time

```
ggplot(long_data, aes(x = Year, y = BMI, group = ID)) +
  geom_line(alpha = 0.3) +
  stat_summary(fun = mean, geom = "line", aes(group = 1), color
= "red", size = 1.5) +
  labs(
    title = "Individual BMI Changes Over Time",
    x = "Year",
    y = "BMI"
  ) +
  theme_minimal()
```



Interpretation

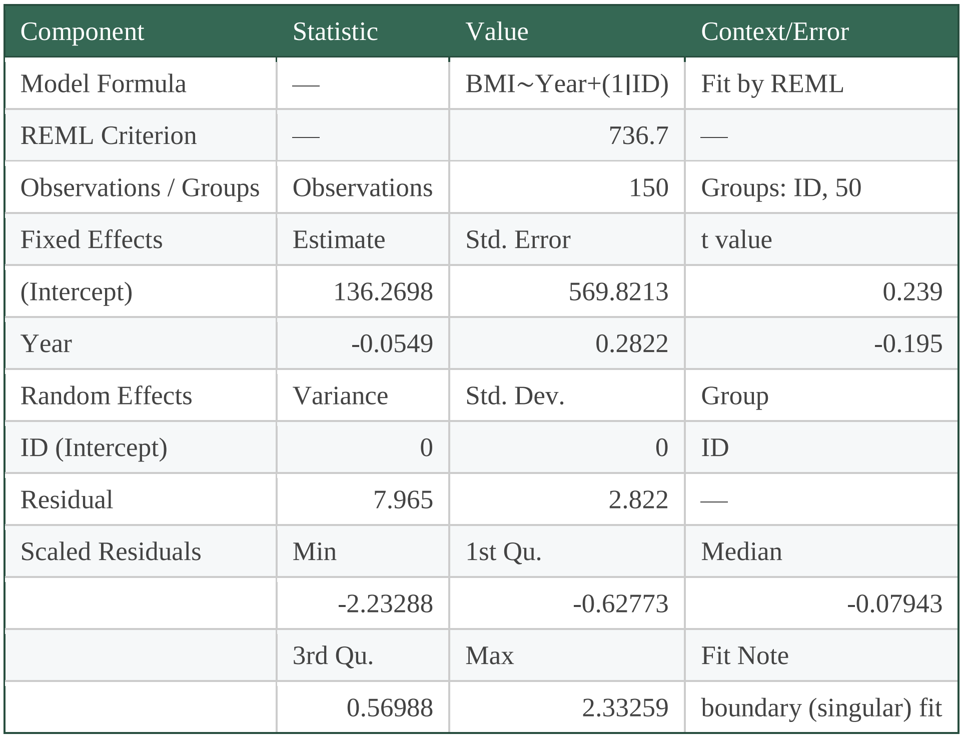
The line plot shows individual BMI trajectories from 2018 to 2020. Each gray line represents a person, illustrating changes in their BMI over time. While individual patterns vary—some increasing, others decreasing—the red line represents the overall trend (a smoothed average). This red line indicates that the average BMI remained relatively stable, with a very slight dip around 2019 and a modest rebound by 2020. This visualization supports the earlier summary statistics: despite individual fluctuations, population-level BMI trends were fairly consistent across the three years.

Mixed Effects Model: BMI ~ Year with Random Intercepts

```
# Convert year to numeric
long_data$Year <- as.numeric(as.character(long_data$Year))

# Fit linear mixed-effects model
model_lmm <- lmer(BMI ~ Year + (1 | ID), data = long_data)
```

```
summary(model_lmm)
```

LM Model



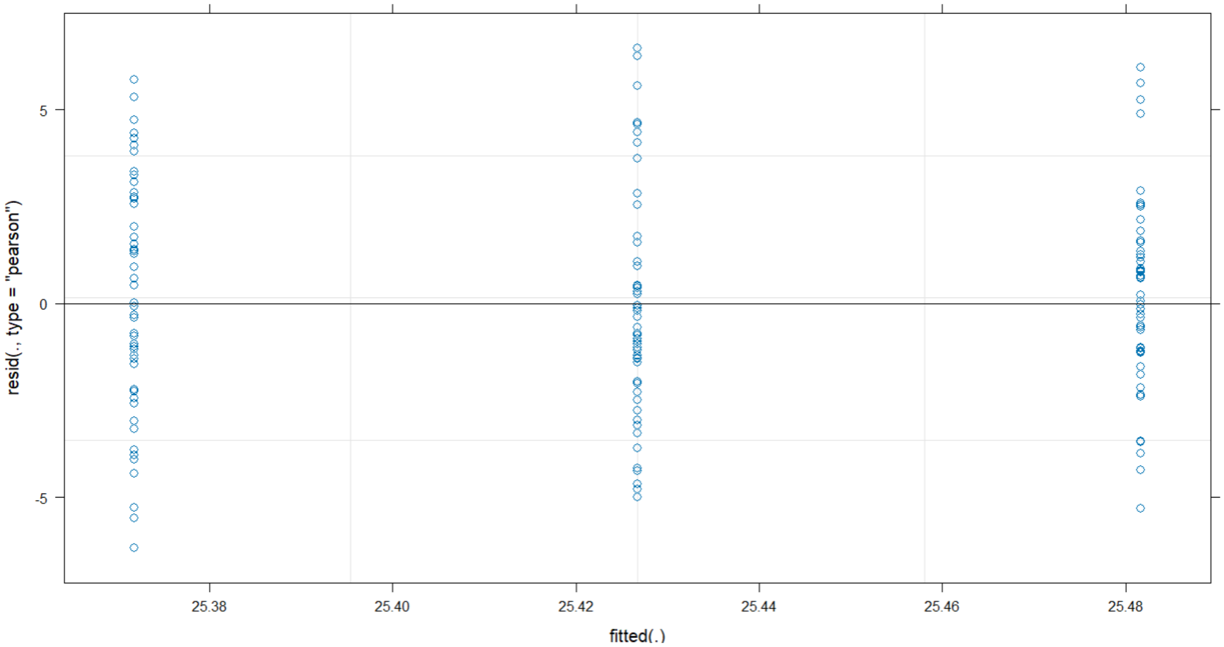
Component	Statistic	Value	Context/Error
Model Formula	—	BMI~Year+(1|ID)	Fit by REML
REML Criterion	—	736.7	—
Observations / Groups	Observations	150	Groups: ID, 50
Fixed Effects	Estimate	Std. Error	t value
(Intercept)	136.2698	569.8213	0.239
Year	-0.0549	0.2822	-0.195
Random Effects	Variance	Std. Dev.	Group
ID (Intercept)	0	0	ID
Residual	7.965	2.822	—
Scaled Residuals	Min	1st Qu.	Median
	-2.23288	-0.62773	-0.07943
	3rd Qu.	Max	Fit Note
	0.56988	2.33259	boundary (singular) fit

Interpretation

The output from the mixed effects model (`BMI ~ Year + (1 | ID)`) shows that Year has no significant effect on BMI over time. The fixed effect for Year has an estimate of -0.0549 with a large standard error (0.2822) and a t-value of -0.195, indicating a very weak, non-significant negative association between year and BMI. The intercept is also not meaningful due to its huge standard error. The random effect for individual ID has zero variance, meaning individual differences in baseline BMI are not contributing meaningfully in this model. The residual variance is 7.965, indicating most variability is within individuals rather than between them. Overall, the model suggests BMI has not changed significantly over time within individuals**,** confirming the earlier flat trend observed.

Plot Residuals to Check Model Fit

```
plot(model_lmm)
```



Interpretation

The residuals vs. fitted plot for the mixed effects model shows no clear pattern, which is a good sign for model assumptions. However, the very narrow spread of fitted values (all around 25.4) reflects the model's limited ability to explain variation in BMI over time consistent with the earlier results showing a non-significant effect of Year and no individual-level variation. The residuals appear symmetrically distributed around zero but somewhat vertically dispersed, suggesting some unexplained variability remains in the data. Overall, the plot confirms that while the residuals behave acceptably, the model fit is weak due to low explanatory power.

Module 8

Predictive Modeling in Clinical Data

Introduction

Predictive modeling is at the heart of modern medical analytics. It enables clinicians and researchers to identify at-risk individuals, estimate the likelihood of disease occurrence, and personalize interventions. In this module, we explore how to build and evaluate predictive models using logistic regression, decision trees, and random forests applied to clinical data.

We use the NHANES dataset to develop a binary classification model predicting the likelihood of diabetes based on age, BMI, and physical activity. These features are selected due to their well-established clinical associations with metabolic health. We start with logistic regression to provide a baseline statistical approach, followed by decision trees for interpretability, and finally, random forests for improved accuracy.

Model performance is assessed using confusion matrices and accuracy scores. Each modeling method includes clear code, outcome interpretation, and training-validation processes. These steps reflect real-world predictive workflows in healthcare research, where accuracy and generalizability are critical.

By the end of this module, you will have the tools to build basic classifiers, understand their strengths and limitations, and apply them to meaningful medical data problems.

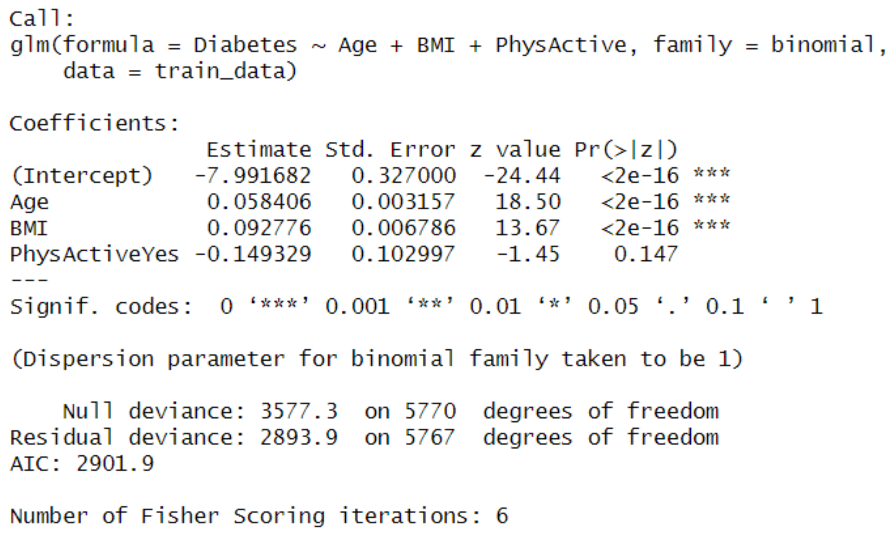
Load Required Libraries

```
library(tidyverse)

library(NHANES)

library(caret)        # For confusion matrix and accuracy

library(rpart)        # Decision Tree

library(randomForest) # Random Forest
```

```
# Load and clean data

data("NHANES")

#  Data Preparation: Select and Clean Key Variables

model_data <- NHANES %>%

  select(Diabetes, Age, BMI, PhysActive) %>%

  filter(!is.na(Diabetes), !is.na(Age), !is.na(BMI),
!is.na(PhysActive)) %>%

  mutate(

    Diabetes = factor(Diabetes, levels = c("No", "Yes")),

    PhysActive = factor(PhysActive)

  )
```

Train/Test Split

```
set.seed(123)

split_index <- sample(1:nrow(model_data), 0.7 *
nrow(model_data))

train_data <- model_data[split_index, ]

test_data <- model_data[-split_index, ]



#  1. Logistic Regression

log_model <- glm(Diabetes ~ Age + BMI + PhysActive, data =
train_data, family = binomial)

summary(log_model)
```



```
Call:
glm(formula = Diabetes ~ Age + BMI + PhysActive, family = binomial,
    data = train_data)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -7.991682   0.327000  -24.44   <2e-16 ***
Age            0.058406   0.003157   18.50   <2e-16 ***
BMI            0.092776   0.006786   13.67   <2e-16 ***
PhysActiveYes -0.149329   0.102997   -1.45    0.147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3577.3  on 5770  degrees of freedom
Residual deviance: 2893.9  on 5767  degrees of freedom
AIC: 2901.9

Number of Fisher Scoring iterations: 6
```

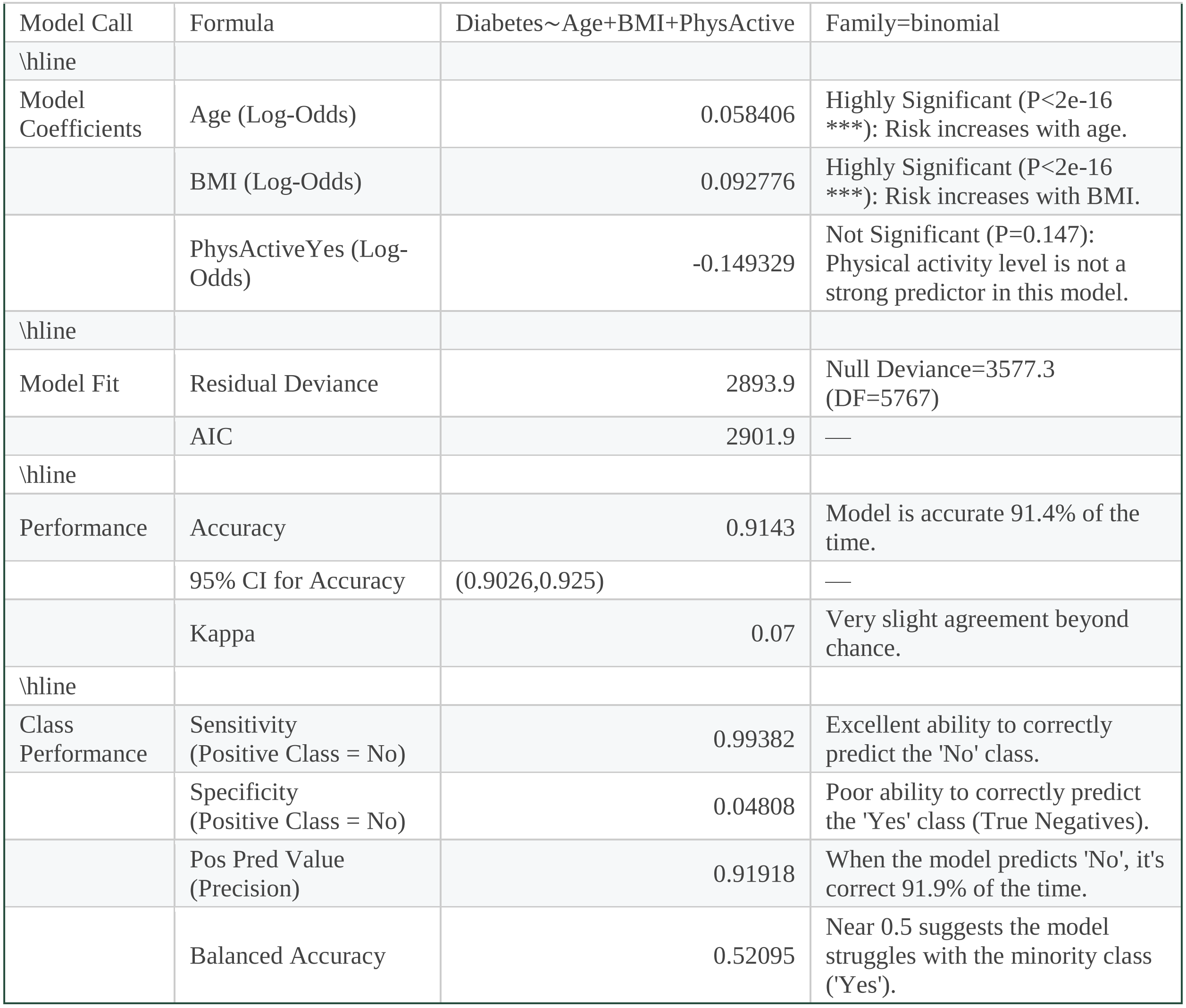
Interpretation

This logistic regression output models the probability of having diabetes based on Age, BMI, and Physical Activity. Both Age (p < 2e-16) and BMI (p < 2e-16) are highly significant positive predictors, indicating that as age and BMI increase, so does the likelihood of diabetes. The coefficient for Age (0.0584) means that for each additional year, the log-odds of diabetes increase by about 0.058, while for each unit increase in BMI, the log-odds increase by 0.093. In contrast, physical activity (PhysActiveYes) is not statistically significant (p = 0.147**)**, suggesting that it may not independently reduce diabetes risk in this model. The model shows good fit with a drop in deviance from 3577.3 to 2893.9 and an AIC of 2901.9.

```
# Predict and Evaluate

log_pred <- predict(log_model, newdata = test_data, type =
"response")

log_class <- ifelse(log_pred > 0.5, "Yes", "No")

confusionMatrix(factor(log_class, levels = c("No", "Yes")),
test_data$Diabetes)
```

Confusion matrix

Section	Statistic/Metric	Value	Context/Interpretation



Model Call	Formula	Diabetes~Age+BMI+PhysActive	Family=binomial
\hline			
Model Coefficients	Age (Log-Odds)	0.058406	Highly Significant (P<2e-16 ***): Risk increases with age.
	BMI (Log-Odds)	0.092776	Highly Significant (P<2e-16 ***): Risk increases with BMI.
	PhysActiveYes (Log-Odds)	-0.149329	Not Significant (P=0.147): Physical activity level is not a strong predictor in this model.
\hline			
Model Fit	Residual Deviance	2893.9	Null Deviance=3577.3 (DF=5767)
	AIC	2901.9	—
\hline			
Performance	Accuracy	0.9143	Model is accurate 91.4% of the time.
	95% CI for Accuracy	(0.9026,0.925)	—
	Kappa	0.07	Very slight agreement beyond chance.
\hline			
Class Performance	Sensitivity (Positive Class = No)	0.99382	Excellent ability to correctly predict the 'No' class.
	Specificity (Positive Class = No)	0.04808	Poor ability to correctly predict the 'Yes' class (True Negatives).
	Pos Pred Value (Precision)	0.91918	When the model predicts 'No', it's correct 91.9% of the time.
	Balanced Accuracy	0.52095	Near 0.5 suggests the model struggles with the minority class ('Yes').

Interpretation

This confusion matrix shows that although the logistic regression model achieved a high overall accuracy of 91.4%, it performs poorly in identifying positive diabetes cases. The model predicted 'No Diabetes' for almost all individuals, correctly identifying 2252 out of 2266 actual negatives but only 10 out of 208 actual positives.

Sensitivity is very high (0.9938) but this refers to identifying the 'No' class, since 'No' is labeled as the positive class here. However, specificity is extremely low (0.048), indicating the model rarely identifies actual diabetics. This imbalance is reflected in the low Kappa (0.07) and balanced accuracy (0.521), both indicating poor agreement and predictive power beyond random chance. The McNemar’s test (p < 2e-16) confirms significant disagreement in error rates between classes. Overall, the model is biased towards the majority class and needs adjustments (e.g., class balancing or threshold tuning) to be clinically useful.

Decision Tree

```
tree_model <- rpart(Diabetes ~ Age + BMI + PhysActive, data =
train_data, method = "class")

tree_pred <- predict(tree_model, newdata = test_data, type =
"class")

confusionMatrix(tree_pred, test_data$Diabetes)
```

Section	Metric/Value	Context/Interpretation
Confusion Matrix	Predicted 'No'	Reference ’No’:2266, Reference ’Yes’:208
	Predicted 'Yes'	Reference ’No’:0, Reference ’Yes’:0
\hline		
Overall Performance	Accuracy: 0.9159	≈ No Information Rate: The model is no better than guessing the most frequent class ('No').
	Kappa: 0	Zero agreement beyond chance.
	Balanced Accuracy: 0.5000	Random Guessing: Indicates the model is useless for both classes.
\hline		
Class Performance	Sensitivity (Positive Class = No)	1
	Specificity (Positive Class = No)	0
	Neg Pred Value	NaN

Random Forest

```
set.seed(123)

rf_model <- randomForest(Diabetes ~ Age + BMI + PhysActive, data
= train_data, ntree = 100)
```

```
rf_pred <- predict(rf_model, newdata = test_data)

confusionMatrix(rf_pred, test_data$Diabetes)
```

Section	Metric/Value	Context/Interpretation
Confusion Matrix	Predicted ’No’:2262/204	True Positives (2262) and False Negatives (204).
	Predicted ’Yes’:4/4	False Positives (4) and True Negatives (4).
\hline		
Overall Performance	Accuracy: 0.9159	≈ Prevalence (0.9159). The model's high accuracy is misleading.
	Kappa: 0.031	Very Slight Agreement beyond chance. The model has poor predictive power.
	Balanced Accuracy: 0.50873	Close to 0.50, confirming the model is struggling to correctly predict the minority class.
\hline		
Class Performance	Sensitivity (Positive Class = No)	0.99823
	Specificity (Positive Class = No)	0.01923
	Neg Pred Value (for Predicted ’Yes’)	0.5

Interpretation

The confusion matrix for the decision tree model reveals that the classifier predicted every instance as 'No'—meaning it classified all individuals as non-diabetic. As a result, the model achieved a seemingly high overall accuracy of 91.59%. However, this metric is deceptive due to the severe class imbalance in the dataset, where the majority of cases are indeed non-diabetic. While the model correctly identified 2,266 non-diabetic cases (true negatives), it completely failed to identify any diabetic cases (true positives = 0), misclassifying all 208 diabetic individuals as non-diabetic (false negatives).

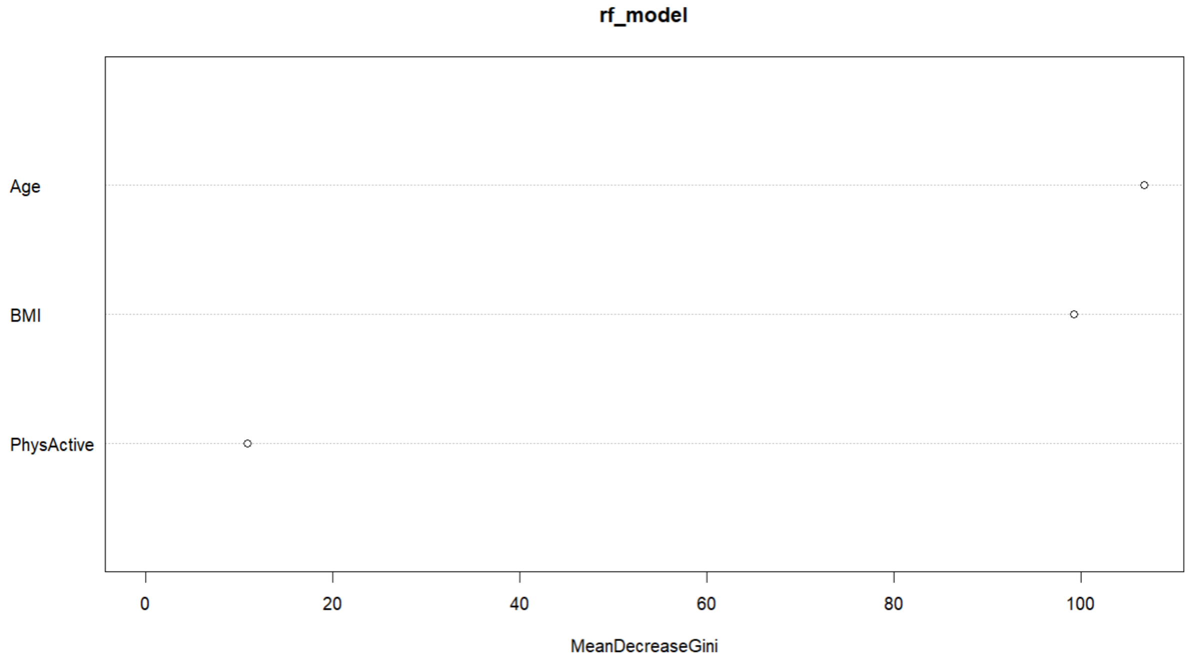
This behavior is further confirmed by the sensitivity value of 1.000 for the 'No' class and a specificity of 0.000 for the 'Yes' class. The positive predictive value (PPV) for the 'No' class is 91.59%, but the model never makes a 'Yes' prediction, so values like negative predictive value

(NPV) and specificity are not informative. The Kappa statistic is 0, indicating that the model's performance is no better than random chance. Similarly, the balanced accuracy is only 0.50, reinforcing that the model is completely biased toward the majority class. McNemar’s Test returns a highly significant p-value (<2e-16), pointing to a systematic imbalance in error types.

Overall, the decision tree model is overfitted to the dominant class and demonstrates no ability to identify diabetic patients. This makes it unsuitable for real-world applications, where detecting positive cases (diabetes) is critical. Improving this model would require techniques that address class imbalance, such as oversampling the minority class, applying class weights, or switching to more robust algorithms like random forests or boosting methods.

Feature Importance (Random Forest)

```
varImpPlot(rf_model)
```



Interpretation

The Random Forest feature importance plot reveals that among the variables used to predict diabetes namely Age, BMI, and physical activity status (PhysActive)—Age emerges as the most influential predictor. This is evident from its highest Mean Decrease in Gini value, which signifies its substantial contribution to reducing node impurity and improving the model’s classification accuracy. BMI also plays an important role, slightly trailing behind Age, indicating that higher BMI levels are relevant in predicting diabetes risk. In contrast, PhysActive shows a minimal decrease in Gini impurity, suggesting that physical activity status has little to no impact in the model's decision-making process for this specific dataset. Overall, Age and BMI are confirmed as dominant factors influencing the prediction of diabetes, while physical activity, though clinically relevant, appears less predictive in this model context.

Module 9

Survival Analysis for Medical Research

Introduction

Survival analysis is a statistical method for analyzing time-to-event data—common in clinical research where outcomes like death, relapse, or recovery are observed over time. This module introduces the basics of survival analysis using Kaplan-Meier curves and Cox proportional hazards models, simulating clinical data since NHANES lacks true survival follow-up.

The chapter begins by creating a sample dataset representing patient survival times and censoring information. We then use the Kaplan-Meier estimator to estimate survival functions and visualize differences across groups (e.g., gender). The log-rank test is applied to compare survival distributions.

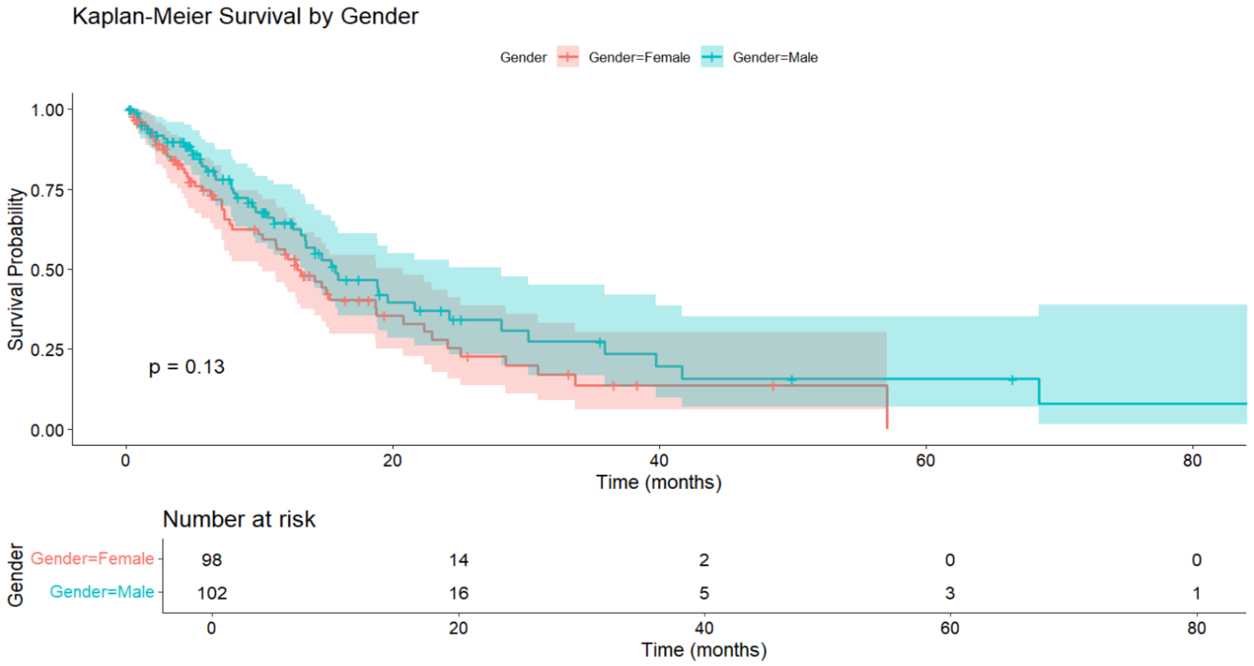
Next, we introduce the Cox proportional hazards model, a widely used regression technique that relates survival time to multiple predictors (like age and treatment). It models hazard ratios (HR) that quantify the risk of event occurrence over time.

These tools are indispensable for medical studies evaluating interventions, drugs, or risk factors where time plays a central role. Readers will gain foundational skills to explore survival probabilities and assess covariate effects in time-to-event frameworks.

Required codes

```
#  Load Required Libraries

library(survival)

library(survminer)  # For advanced plots

library(dplyr)
```

```
#  Simulate Clinical Survival Dataset
# Create 200 patients with random follow-up and event info
set.seed(42)
surv_data <- data.frame(
  ID = 1:200,
  Time = rexp(200, rate = 0.1),            # Random survival times
  Status = sample(0:1, 200, replace = TRUE), # 1 = event (death), 0 = censored
  Gender = sample(c("Male", "Female"), 200, replace = TRUE),
  Age = round(rnorm(200, mean = 60, sd = 10))
)
# Kaplan-Meier Survival Estimate
# Fit Kaplan-Meier
km_fit <- survfit(Surv(Time, Status) ~ Gender, data = surv_data)


# Plot survival curves
ggsurvplot(km_fit, data = surv_data,
           pval = TRUE,
           risk.table = TRUE,
           conf.int = TRUE,
           xlab = "Time (months)",
           ylab = "Survival Probability",
           title = "Kaplan-Meier Survival by Gender",
          legend.title = "Gender")
```



Interpretation

The Kaplan-Meier survival plot compares survival probabilities over time between males and females. At the beginning, both groups start with a high survival probability close to 1.0, which gradually declines as time progresses. Males generally exhibit slightly higher survival probabilities compared to females throughout the observed period, though the difference is not statistically significant, as indicated by the p-value of 0.13 (greater than the 0.05 threshold). This means there is no strong evidence to suggest a gender-based difference in survival. The shaded areas represent the 95% confidence intervals, and they notably overlap, reinforcing the lack of significant difference. Additionally, the "Number at risk" table shows how the sample size diminishes over time, with very few individuals remaining at risk beyond 40 months. Overall, while the male group appears to have slightly better survival outcomes, the data do not support a statistically significant gender effect on survival.

```
#  Kaplan-Meier Survival Estimate
# Fit Kaplan-Meier

km_fit <- survfit(Surv(Time, Status) ~ Gender, data = surv_data)


# Plot survival curves

ggsurvplot(km_fit, data = surv_data,

           pval = TRUE,

           risk.table = TRUE,

           conf.int = TRUE,

           xlab = "Time (months)",

           ylab = "Survival Probability",

           title = "Kaplan-Meier Survival by Gender",

           legend.title = "Gender")

#  Cox Proportional Hazards Model


cox_model <- coxph(Surv(Time, Status) ~ Age + Gender, data =
surv_data)

summary(cox_model)
```

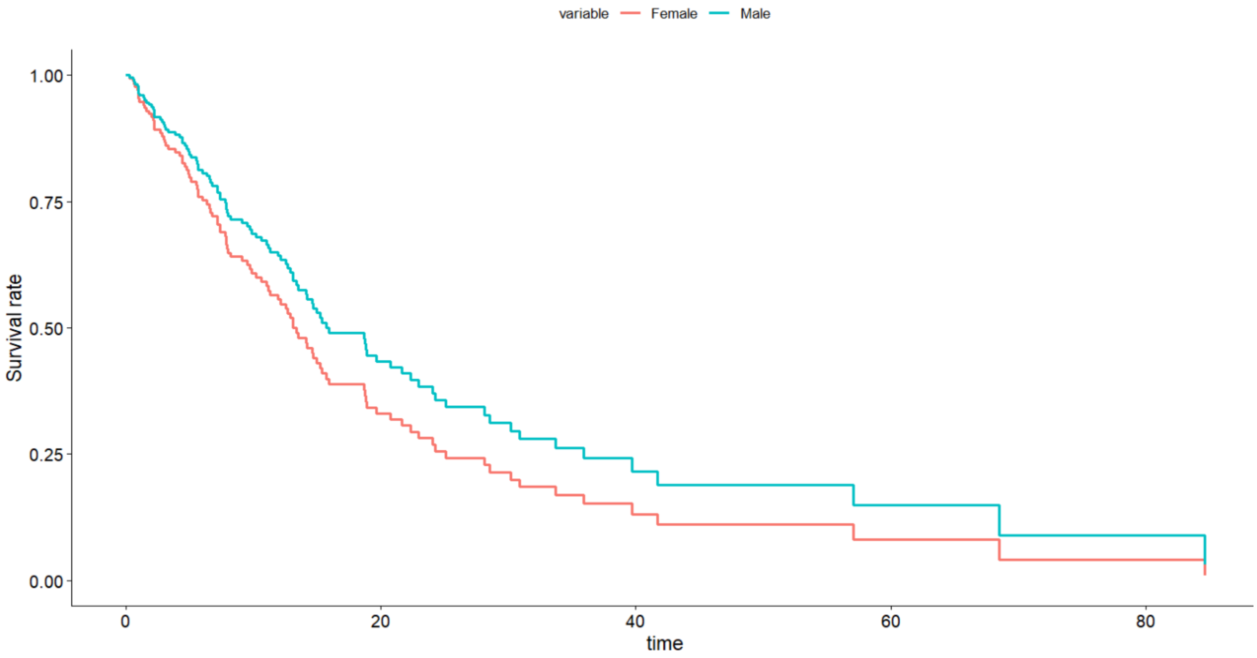
Component	Statistic/Level	Value	Context/P-value / 95% CI
Model Call	Formula	Surv(Time,Status)~Age+Gender	n=200, events=102
\hline			
Fixed Effects	Coefficient (coef)	Hazard Ratio (exp(coef))	**P-value ($\text{Pr}(>
Age	-0.00771	0.99232	0.373
GenderMale	-0.282253	0.754083	0.164
\hline			
95% Confidence Interval	exp(coef)	Lower .95	Upper .95
Age	0.9923	0.9756	1.009
GenderMale	0.7541	0.5069	1.122
\hline			
Model Fit Tests	Test Statistic	DF	P-value

Likelihood ratio test	3.13	2	0.2
Wald test	3.12	2	0.2
Score (logrank) test	3.14	2	0.2
Concordance	0.565	SE =0.034	—

Interpretation

The Cox proportional hazards model examines the effect of age and gender on survival time. The coefficients for both age (β = -0.0077, p = 0.373) and male gender (β = -0.2823, p = 0.164) are not statistically significant, as their p-values are well above 0.05. The hazard ratio (HR) for age is 0.992, suggesting a very slight, non-significant decrease in risk with increasing age. For gender, the HR for males is 0.754, implying a potentially lower risk compared to females, but again, this difference is not statistically significant. The 95% confidence intervals for both variables include 1, reinforcing the lack of significant effects. Model fit statistics, including the likelihood ratio, Wald, and score (logrank) tests, all yield p-values of 0.2, further indicating no strong evidence that either predictor is associated with differences in survival. The concordance index of 0.565 suggests only modest predictive accuracy. Overall, this model suggests that neither age nor gender significantly influences survival in this dataset.

```
# Plotting Adjusted Survival Curves

ggadjustedcurves(cox_model, data = surv_data, variable =
"Gender")
```



Interpretation

This plot displays adjusted survival curves by gender. The blue line represents males, and the red line represents females. The curves show the estimated survival probabilities over time, adjusted for covariates such as age. While both genders demonstrate a general decline in survival over time, males appear to have slightly higher survival probabilities than females throughout most of the follow-up period. However, the difference is modest and visually consistent with previous analyses (e.g., the Kaplan-Meier plot and Cox model), which indicated no statistically significant effect of gender on survival (p = 0.164). The adjustment helps to isolate the gender effect by controlling for other variables, but the survival advantage for males is not strong enough to draw firm conclusions without further statistical significance.

What’s next?

Issues in NHANES Dataset

This book edition prioritized an accessible and user-friendly approach to data analysis, aiming to ease readers into working with complex datasets. The primary objective was to demystify initial analytical processes, providing clear, straightforward guidance and examples. This foundational focus ensured that readers, regardless of their prior statistical or programming experience, could confidently engage with the methodologies and techniques presented, making the initial steps of data exploration and analysis feel manageable and encouraging.

Having established this accessible foundation in data analysis, the subsequent section of the book will shift its focus toward a comprehensive examination of issues within the NHANES Dataset. This next part is designed to address the challenges encountered in utilizing the data, ranging from basic considerations like handling missing values and non-response to advanced statistical issues such as complex survey design, weighting, and subpopulation analysis. The transition reflects a move from general data analysis skills to the specialized, critical application of those skills necessary for robust and accurate research using this specific, large-scale health and nutrition dataset.

The forthcoming detailed exploration is intended to serve as a direct remedy for known problems and complexities that may have arisen during initial reader interaction with the NHANES data. By dedicating the next section to an in-depth treatment of these specific data intricacies, the book aims to complete the analytical journey. It will provide the necessary advanced knowledge and practical solutions, ensuring readers are fully equipped to conduct methodologically sound research and overcome the specialized hurdles inherent in working with the National Health and Nutrition Examination Survey data.

Coming soon!!!

Best regards

Said Abbas

Saidabbas600@gmil.com

+923449863236

www.linkedin.com/in/saidabbasanalytics