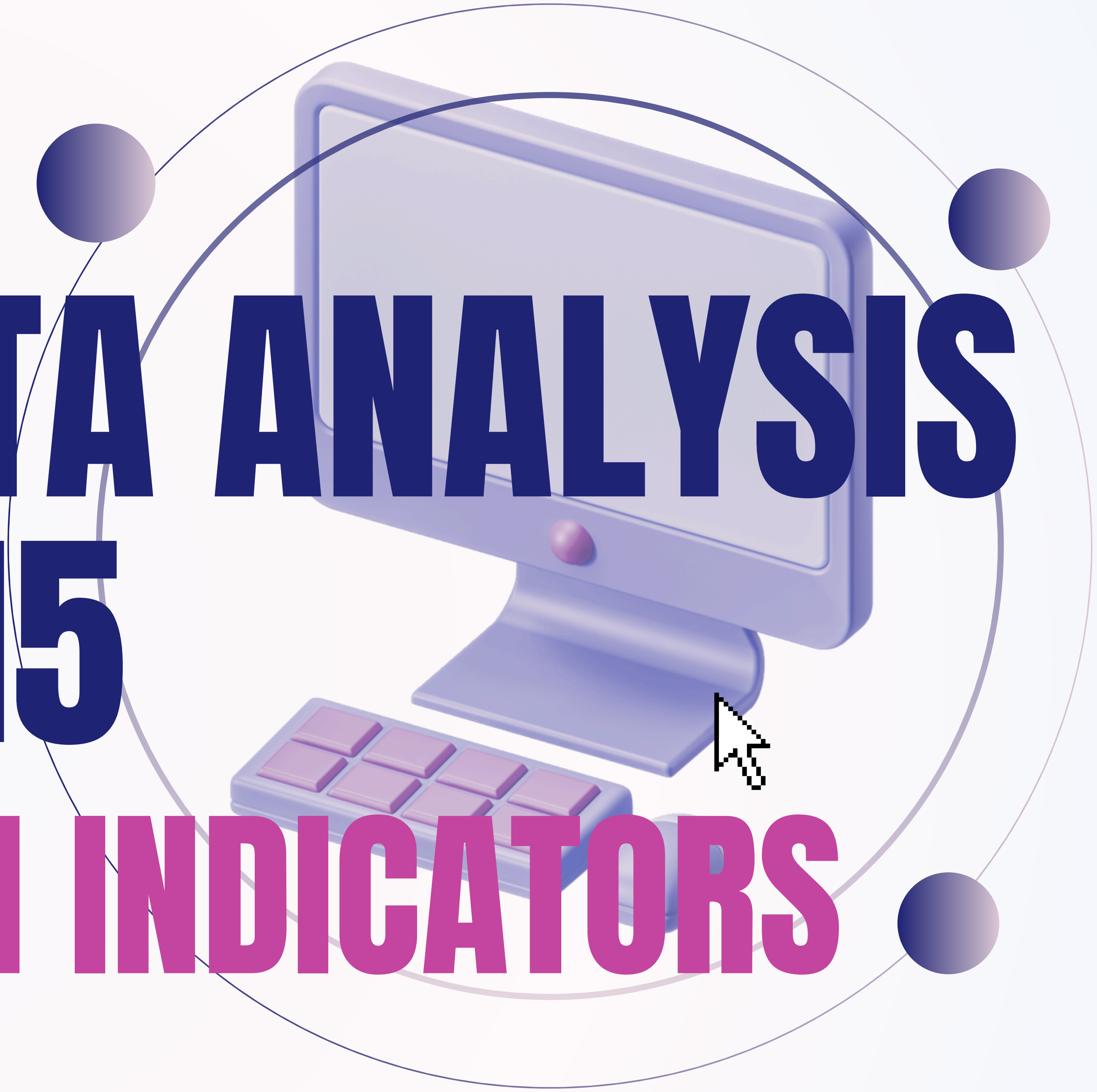
universidade
de aveiro

Statistics Laboratories Unit Course

Professor Luís M. Silva

EXPLORATORY DATA ANALYSIS OF THE BRFSS 2015

HEART DISEASE HEALTH INDICATORS DATASET



Afonso D. Carreira, Marta F. Carvalho, Rita S. Marques, Tomás V. Geraldes
Department of Medical Sciences, University of Aveiro

SUMMARY

1

INTRODUCTION

2

OBJECTIVES

3

DATA SET LOADING AND UNDERSTANDING

4

DATA AND GRAPHICAL ANALYSIS

5

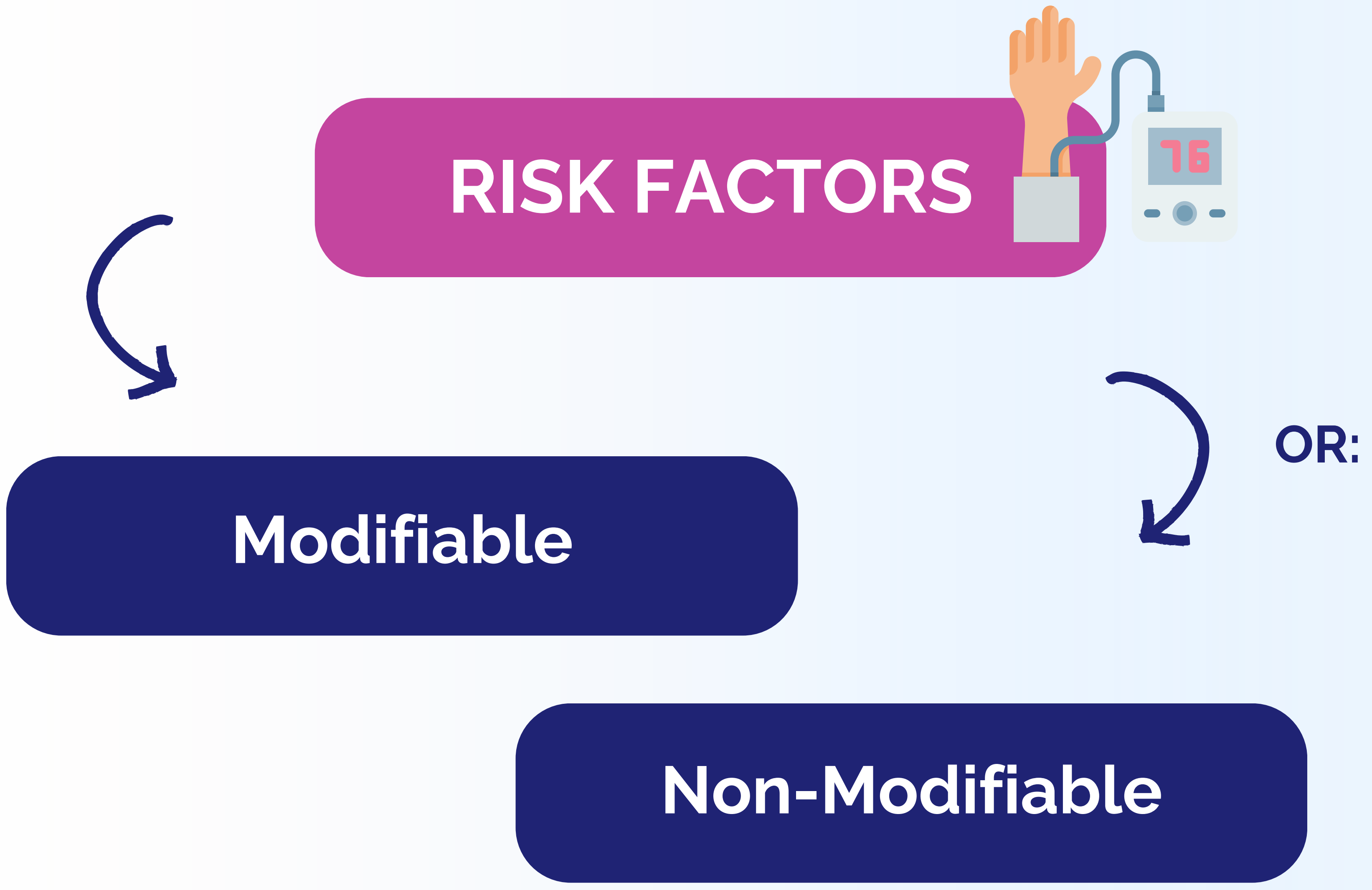
KEY POINTS

INTRODUCTION

Cardiovascular disease (CVD) encompasses a range of conditions affecting the heart and blood vessels, including high blood pressure, atherosclerosis, heart failure, strokes, arrhythmias, and valvular heart disease.

WORLD HEALTH ORGANIZATION (WHO):

Leading cause of death globally, claiming approximately 17.9 million lives annually.


RISK FACTORS

OR:

Modifiable

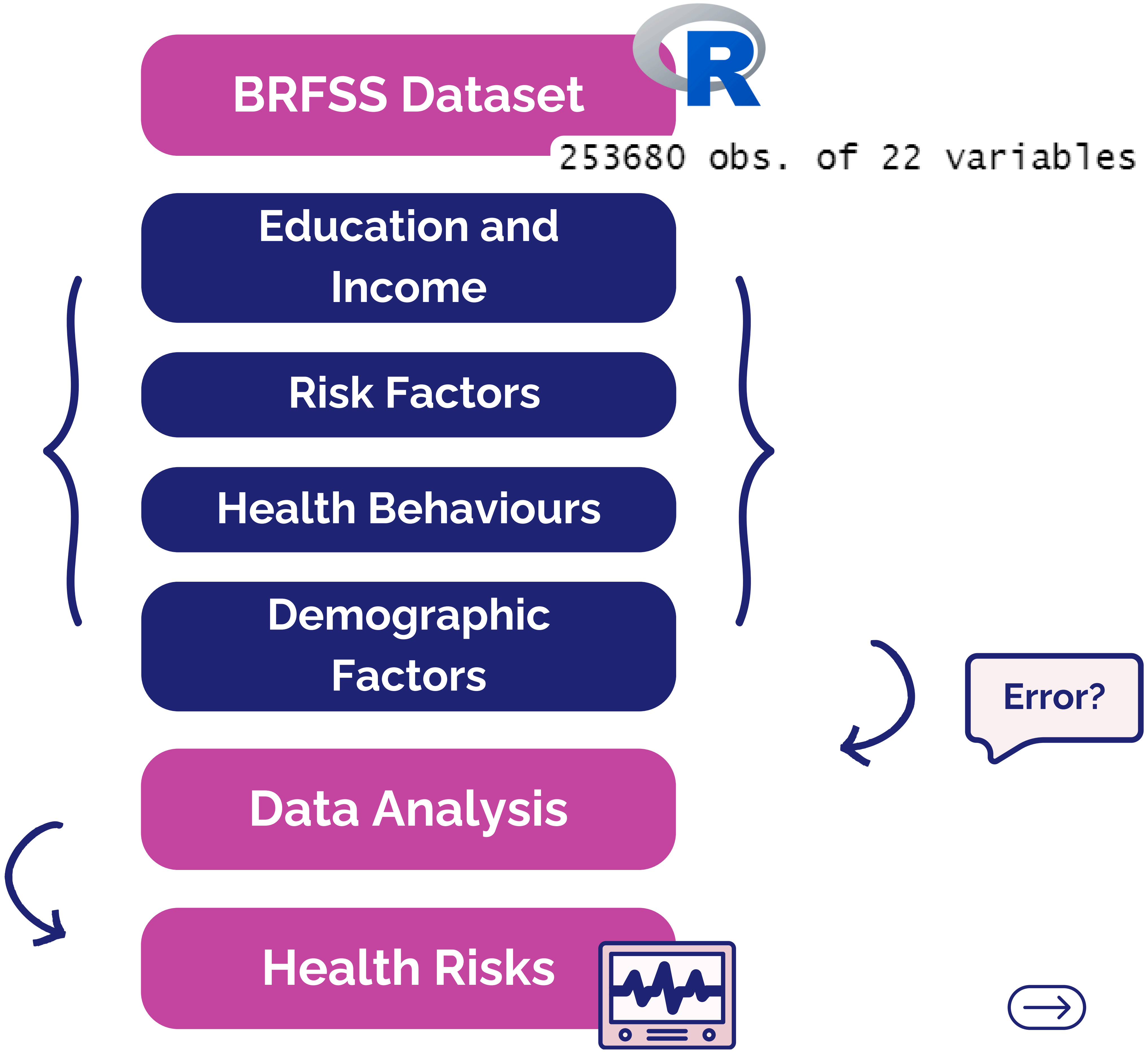
Non-Modifiable

Heart disease risk factors include uncontrollable elements (**age, sex, genetics**) and modifiable ones (**diet, smoking, stress**), where managing the latter can lower CVD risk.

BEHAVIOURAL RISK FACTOR SURVEILLANCE SYSTEM

BRFSS™
Behavioral Risk Factor Surveillance System

The BRFSS is a **U.S. health survey** that **collects state-level data** on citizens' behaviours and conditions. That precious information **can be used to assess various diseases' risk**, such as heart disease risk.



BRFSS Dataset

```
253680 obs. of 22 variables
```

Education and Income

Risk Factors

Health Behaviours

Demographic Factors

Error?

Data Analysis

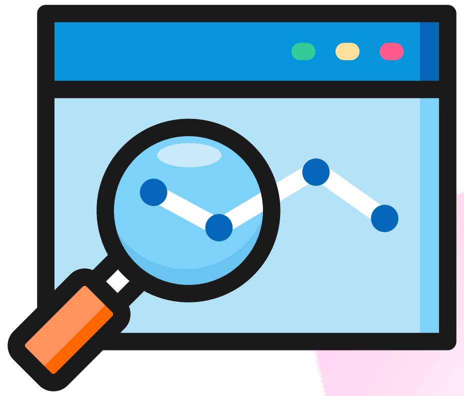
Health Risks

OBJECTIVES

- Analysis of **Indirect** and **Direct Risk Factors** in Cardiovascular Disease
- Exploration of **Health Patterns** and **Behaviours** Related to Cardiovascular Risk
- Assess **potential associations** that the factors under study might imply **in relation to CVDs**
- Development of **Predictive Models** for Cardiovascular Disease
- Discussion and Proposals for **Preventative Strategies**

Age, sex, income and education

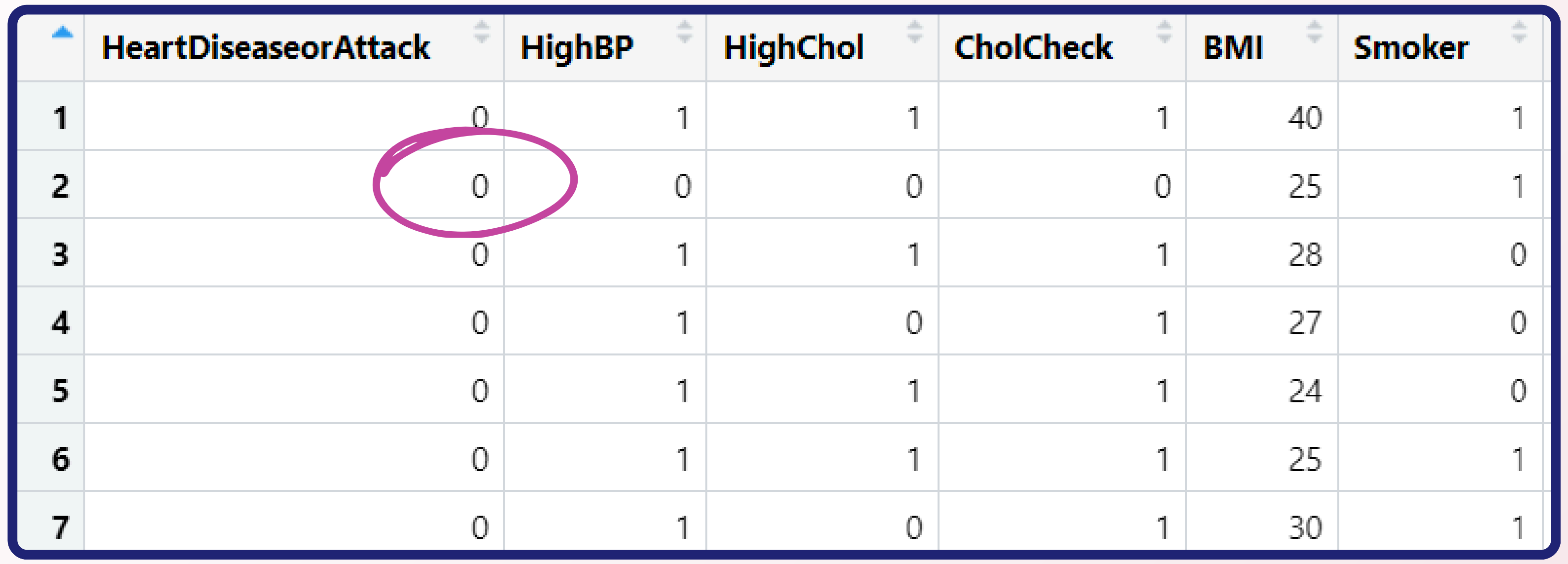
Hypertension, cholesterol levels, diabetes, smoking and physical inactivity (...)



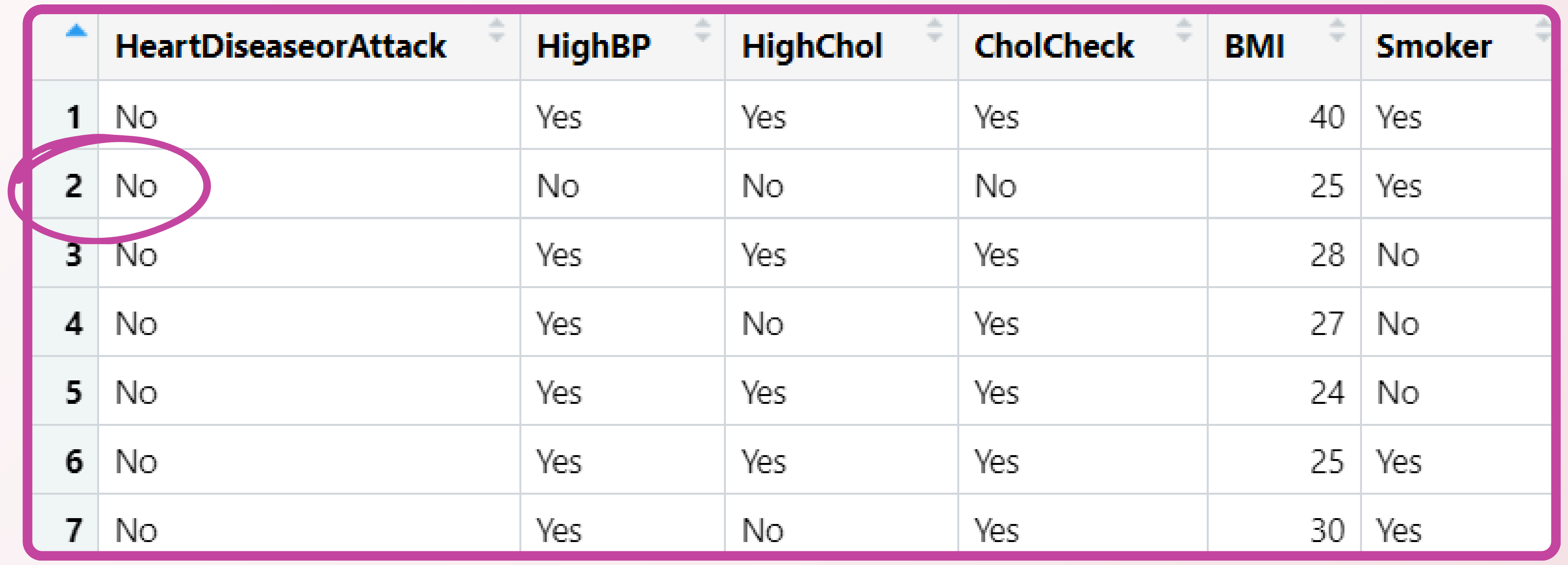
DATA SET LOADING AND UNDERSTANDING

The data, provided in **CSV format**, was imported into R as a **data frame**. Each column is **numeric**, with most responses categorised as **binary variables**.

Data processing converted numeric columns to **factors** for accurate interpretation. The final dataset includes **253,680 records** and **22 columns**, reflecting the sample size and responses.



	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker
1	0	1	1	1	40	1
2	0	0	0	0	25	1
3	0	1	1	1	28	0
4	0	1	0	1	27	0
5	0	1	1	1	24	0
6	0	1	1	1	25	1
7	0	1	0	1	30	1



	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker
1	No	Yes	Yes	Yes	40	Yes
2	No	No	No	No	25	Yes
3	No	Yes	Yes	Yes	28	No
4	No	Yes	No	Yes	27	No
5	No	Yes	Yes	Yes	24	No
6	No	Yes	Yes	Yes	25	Yes
7	No	Yes	No	Yes	30	Yes

DATA ANALYSIS

To assess data distribution, patterns, correlations, and potential anomalies.

Variables were briefly compared to factors, using **multivariate charts** in ggplot2.

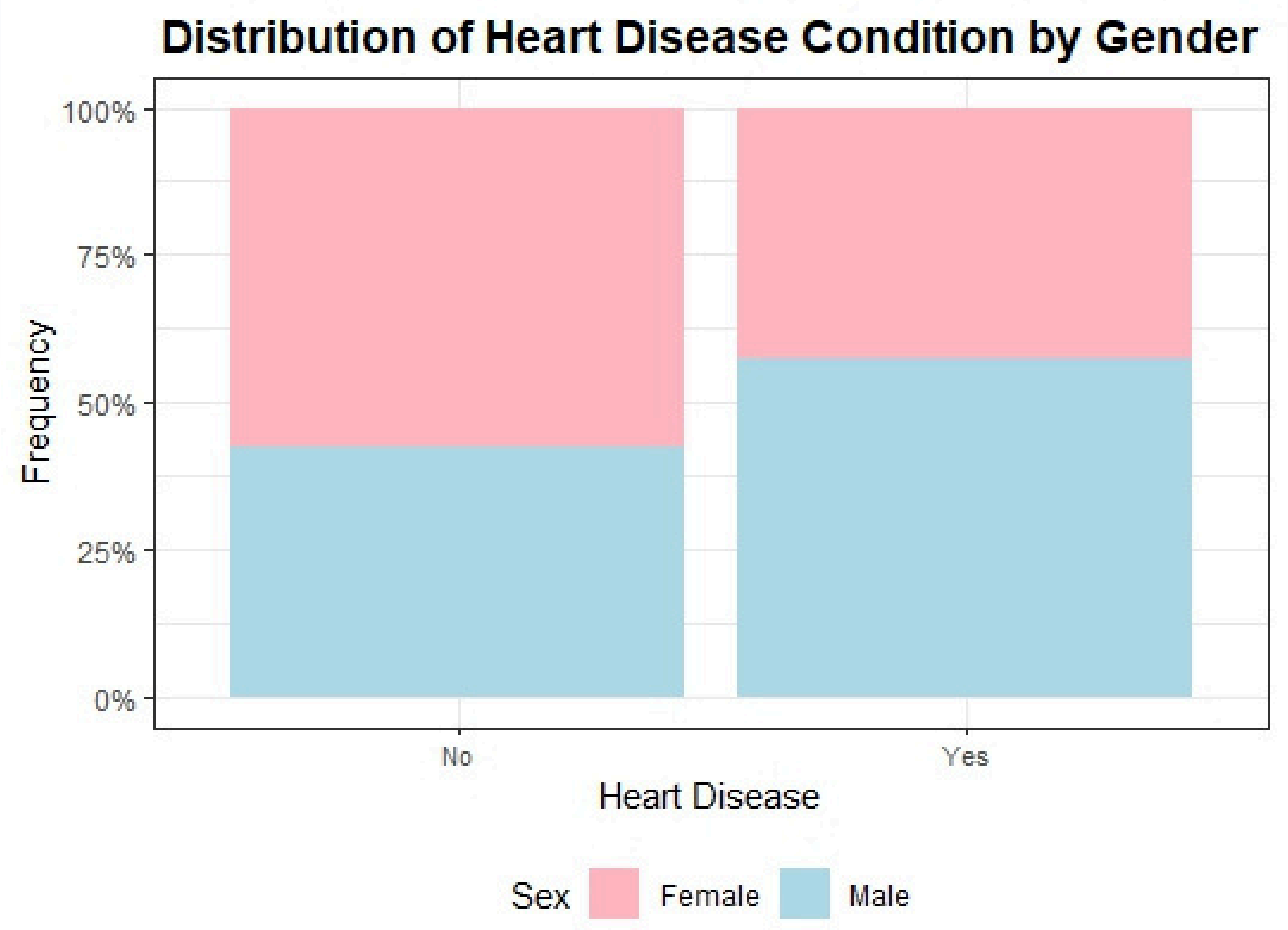
Binomial logistic regression was employed to predict heart disease risk from survey responses, producing **probability rankings** for each individual's likelihood of cardiovascular disease.

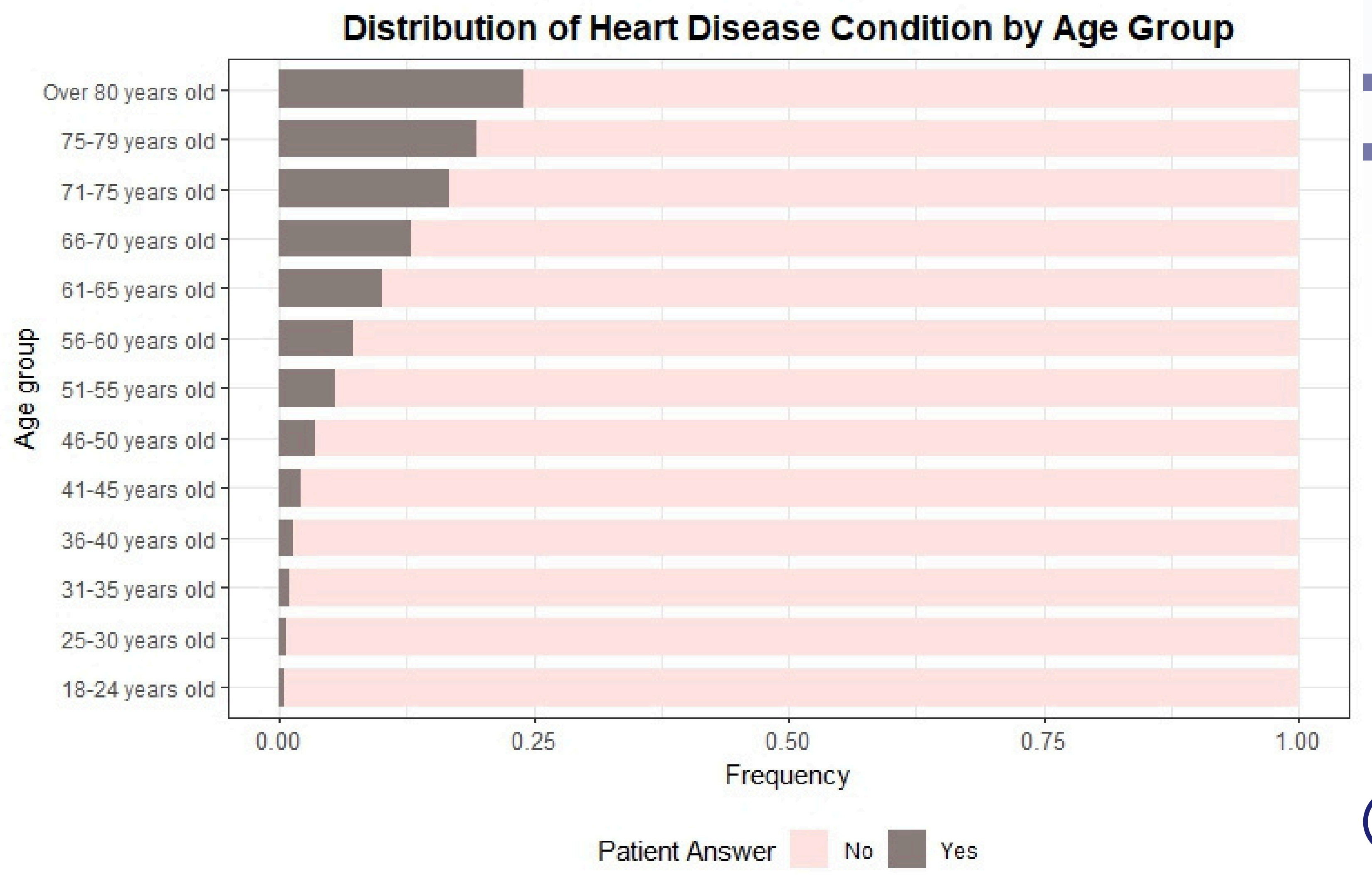
The dataset was split into a **70% training set** and a **30% test set** for robust model validation.

Model fit was assessed using a **likelihood ratio test**.
Chi-squared ANOVA test rejected the null hypothesis.

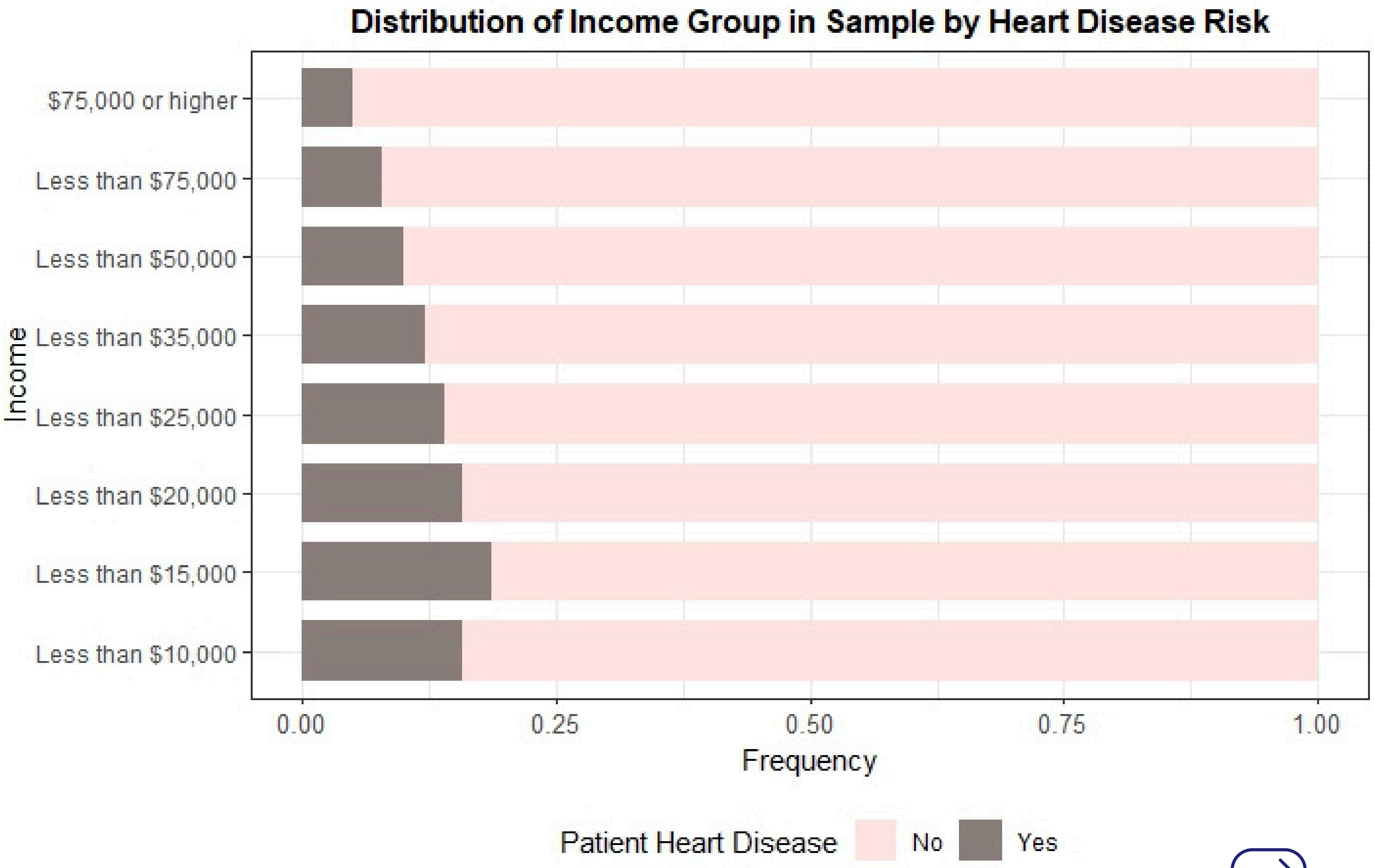
GRAPHICAL ANALYSIS

Age and **gender** are key factors in analysing the **prevalence of cardiovascular disease**.

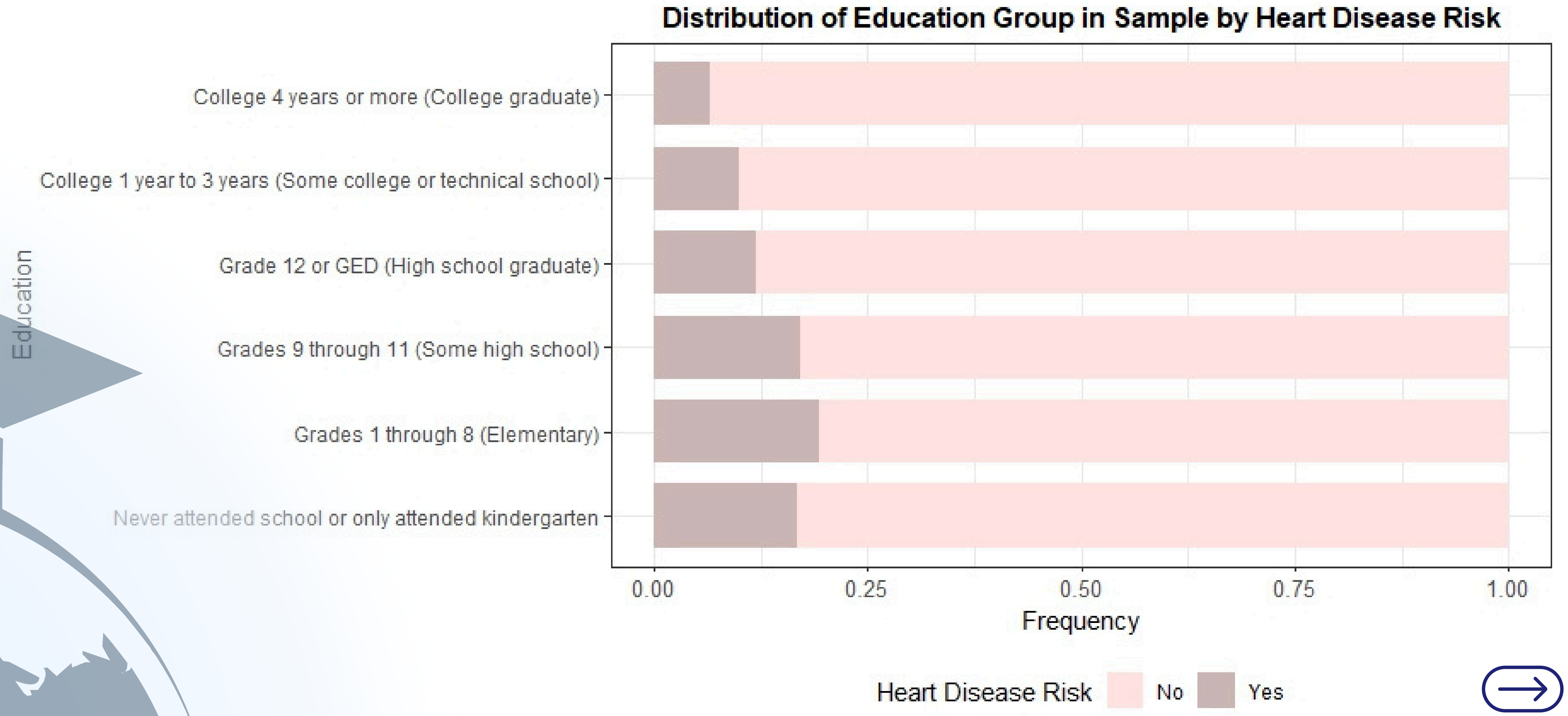




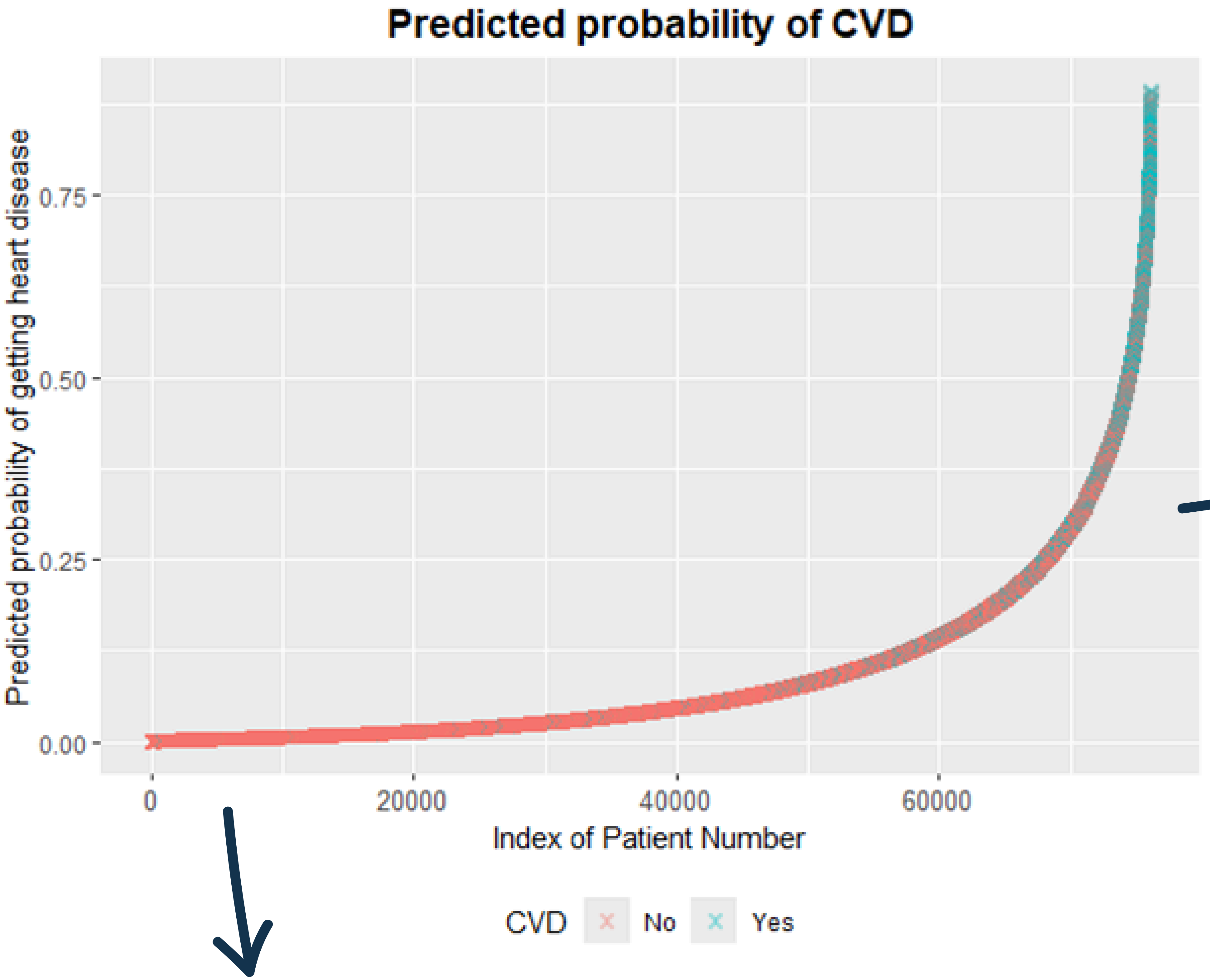
This graphic suggests that **more income** might have some association with **better access to health care and preventive services**, healthier lifestyle choices **versus lower income.**



Individuals with **higher education levels** are more likely to engage in **preventive healthcare practices**, including **regular check-ups** and **early detection screenings** for cardiovascular conditions.



Distribution of Education Group in Sample by Heart Disease Risk



High probability - positive diagnosis for CVD

False negatives?

Low specificity (0.1256):
the test correctly identifies only a small proportion of cases that do not actually have the disease

Low negative predictive value (0.5416):
just over half of individuals with a predicted negative result are actually disease-free

Low probability - negative diagnosis for CVD

KEY POINTS:

- The **Behavioural Risk Factor Surveillance System** (BRFSS) is a respectable tool for **analysis of the relationship between various lifestyle factors** and **the likelihood of developing heart diseases**.

- In our Exploratory Data Analysis (EDA), we conclude that **education** and **income** are modifiable risk factors that can be correlated with Cardiovascular Heart Disease (CVD) alongside the unmodifiable risk factors **gender** and **age**.

- We were able to **predict the probability of heart disease** in patients having into account their answers to the questionnaire, allowing to **develop predictive heart disease models**.