Life Expectancy Predictions

INSY662 - 075

Presented by:

Madeleine Dinh

Alexandra Guion

Juliana Hubacova

Boyang Wan

Table of Contents

Introduction

Objective

Predict life expectancy using lifestyle, demographic, and health factors.

Approach

Analyze nature vs nurture influences based on Minnesota Twin Study insights.

Impact

Provide actionable data for healthcare, government, and insurance sectors.

Minnesota Twin Study Insights

Study Focus

Conducted by Thomas Bouchard (University of Minnesota, 1979). Studied twins raised apart to explore **Nature (genetics) vs. Nurture (environment)**

Key Findings

STAY TUNED!

Relevance to Project

Provides framework for analyzing life expectancy as a function of both genetic and environmental influences.

Analytical Problem

Contextual information

U.S. life expectancy has declined to 76.4 years, the shortest in nearly 2 decades

Nature vs. Nurture Analysis

Explore relationships between genetic predispositions and lifestyle factors affecting longevity

Life Expectancy Analysis and Predictions

Feature selection (identify key variables impacting life expectancy), predictive modeling and model evaluation

Stakeholders

Healthcare providers, government agencies, insurance companies, and general public

A look at our Data

Features

Demographics
Gender
Height (cm)
Weight
Income
Education Level
BMI

Biometrics/Nature
Blood Pressure
Cholesterol Level
Bone Density
Vision Sharpness
Hearing Ability
Cognitive Function
Family History
Blood Glucose Level

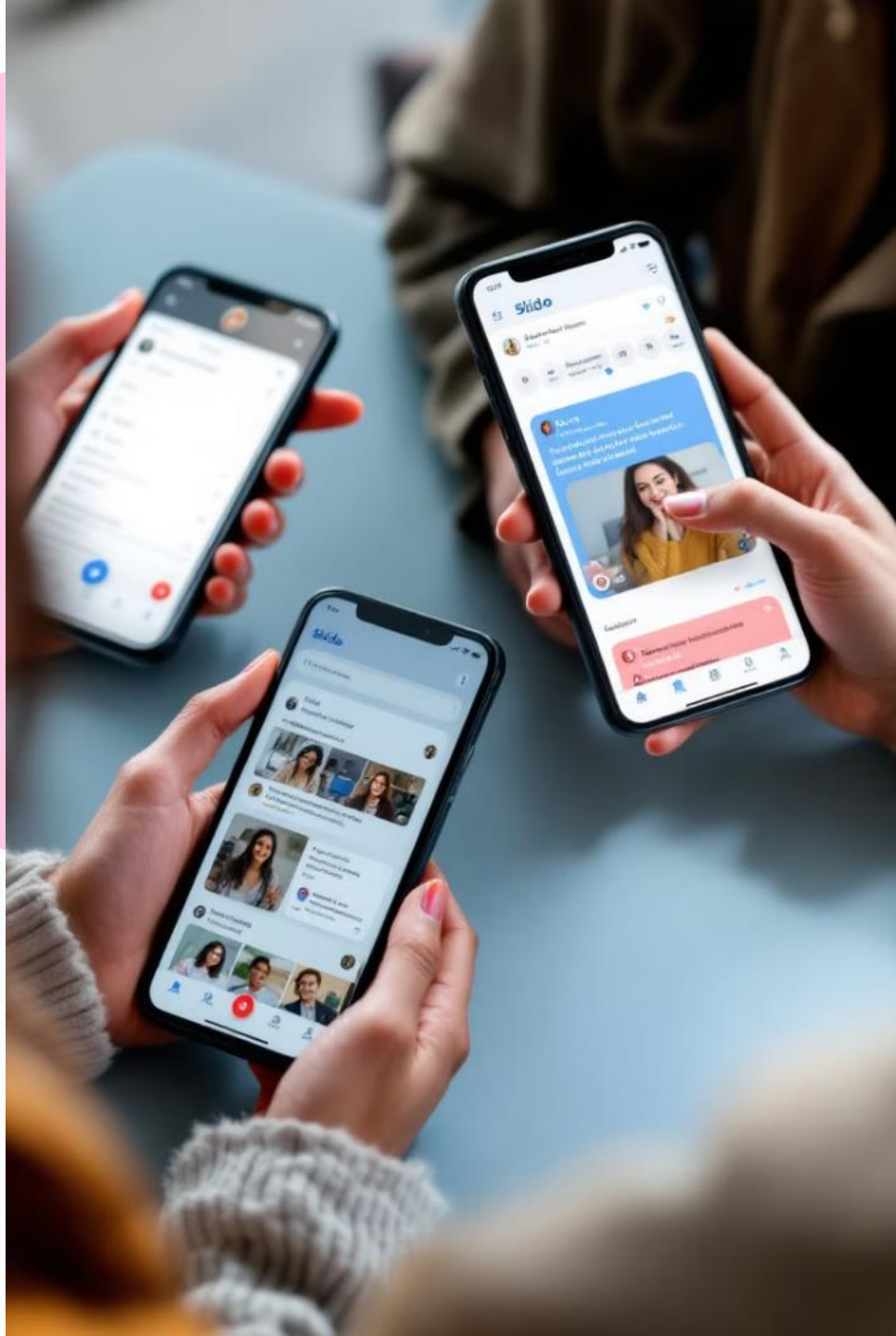
Lifestyle/Nurture
Physical Activity Level
Smoking Status
Alcohol Consumption
Diet
Sun Exposure
Pollution Exposure
Sleep Patterns
Stress Levels
Mental Health Status
Medication Use
Chronic Diseases

Target Variables

Age (years)

Interactive Slido Session

Can you guess which feature is more important?

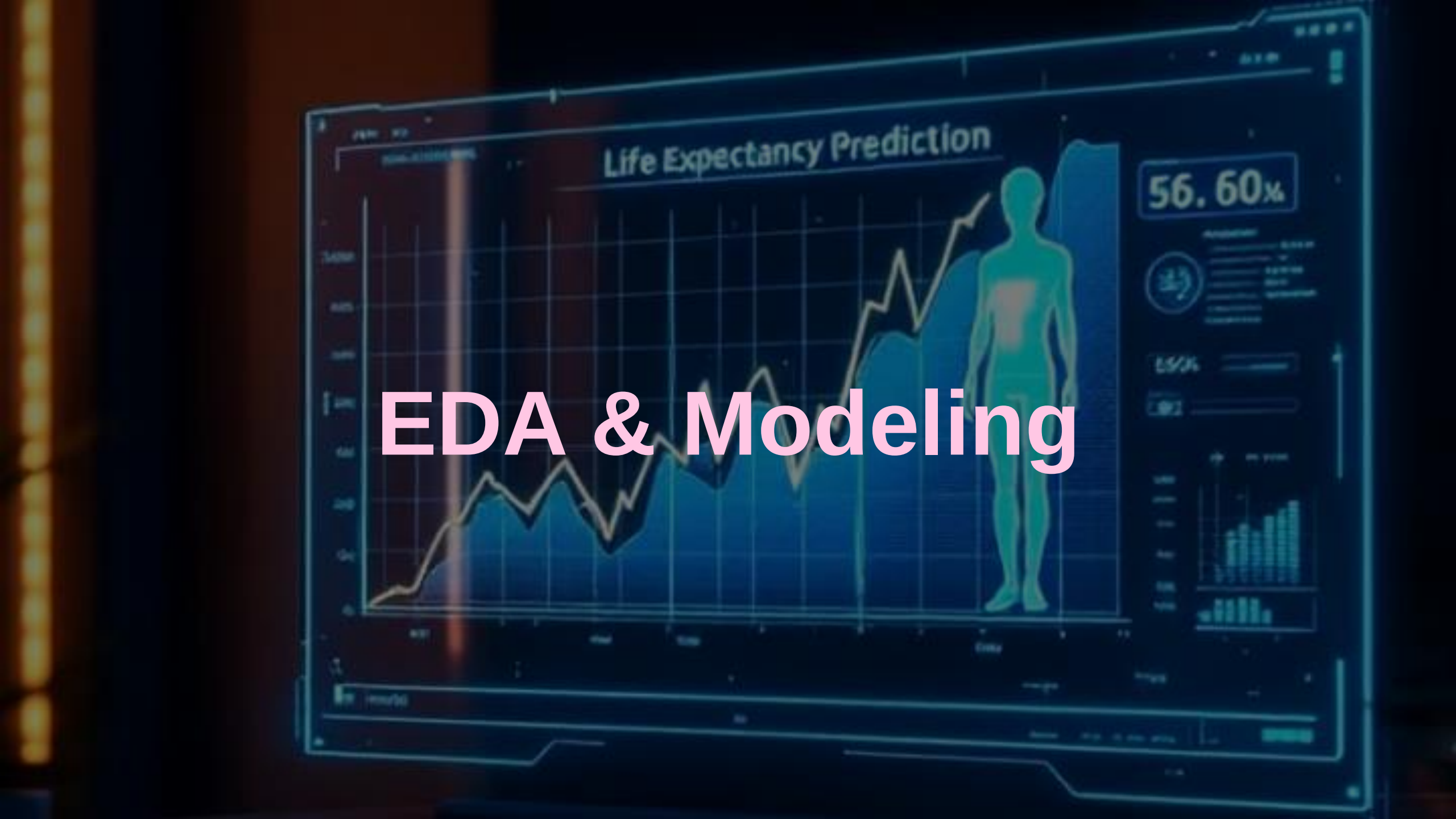


slido

Please download and install the Slido app on all computers you use

What feature has the most influence on life expectancy?

ⓘ Start presenting to display the poll results on this slide.

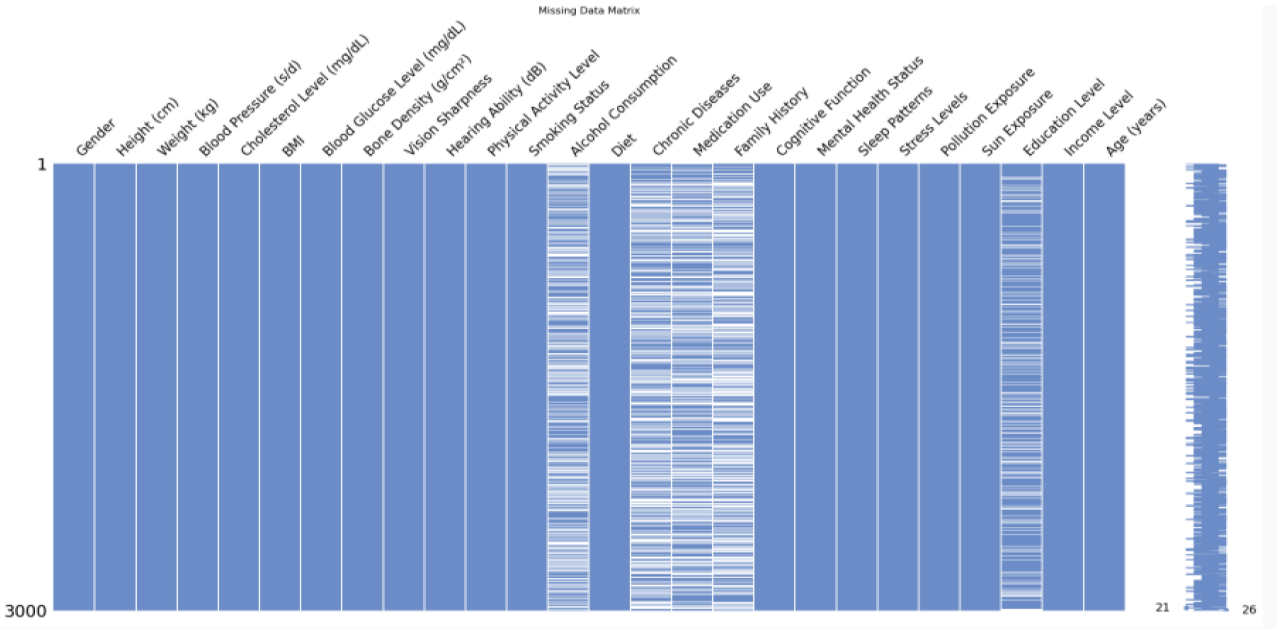


EDA & Modeling

Missing Values

We initially identified some **missing values** in categorical predictors. However, upon closer inspection, we found that these values are not truly missing; they represent **“None.”**

- For example, in the case of alcohol consumption, a blank entry indicates that the individual does not consume alcohol.



Therefore, after encoding all categorical predictors into dummy variables, we added a **“None”** column for each predictor with missing values. This column represents cases where the missing value indicates **“None”** as a distinct category.

```
Alcohol Consumption: ['None' 'Occasional' 'Frequent']
Chronic Diseases: ['None' 'Hypertension' 'Diabetes' 'Heart Disease']
Medication Use: ['None' 'Regular' 'Occasional']
Family History: ['None' 'Heart Disease' 'Hypertension' 'Diabetes']
Education Level: ['None' 'Undergraduate' 'High School' 'Postgraduate']
```

Data Conversion

During initial data exploration, we discovered that the predictor '**Blood Pressure (s/d)'** is stored as an object with string values, even though it is meant to be a numerical variable. Upon closer inspection, we found that 'Blood Pressure (s/d)' is recorded in the format "'Systolic'/'Diastolic'," combining both values into a single string.

```
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 26 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Gender                    3000 non-null   object
 1   Height (cm)               3000 non-null   float64
 2   Weight (kg)               3000 non-null   float64
 3   Blood Pressure (s/d)      3000 non-null   object
 4   Cholesterol Level (mg/dL) 3000 non-null   float64
```

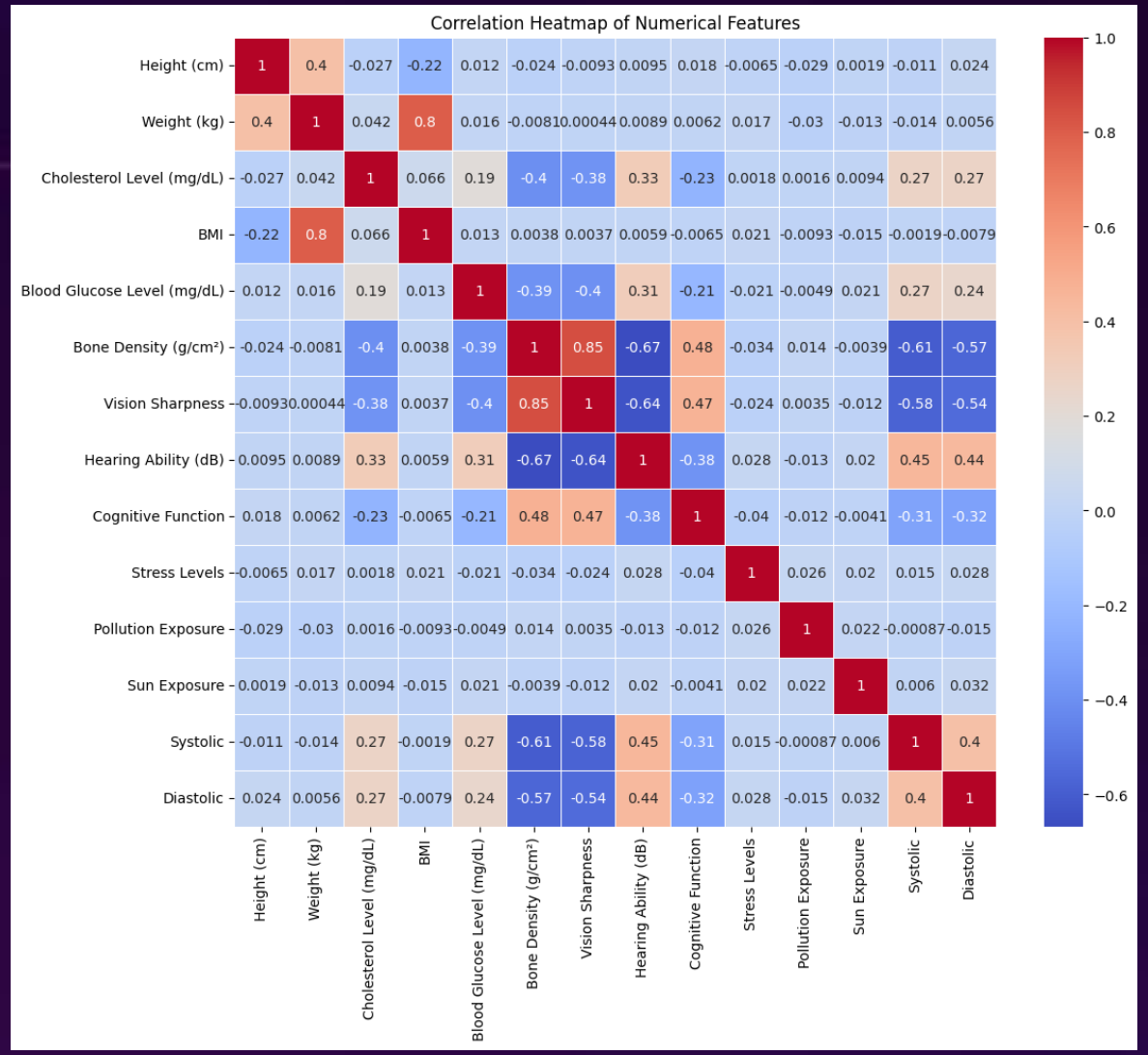
	Gender	Height (cm)	Weight (kg)	Blood Pressure (s/d)
0	Male	171.148359	86.185197	151/109
1	Male	172.946206	79.641937	134/112
2	Female	155.945488	49.167058	160/101

Therefore, we separated 'Blood Pressure (s/d)' into two individual columns: one for 'Systolic' and another for 'Diastolic.' We then set their data types to "int," enabling us to treat them as numerical variables in the model.

Blood Pressure (s/d)
151/109
134/112
160/101
133/94

Systolic	Diastolic
151	109
134	112
160	101
133	94

Multicollinearity of Numerical Variables



	Feature	VIF
0	Height (cm)	606.609744
1	Weight (kg)	250.671063
2	BMI	230.516559
3	Diastolic	141.499350
4	Systolic	138.208360
5	Cholesterol Level (mg/dL)	111.673533
6	Blood Glucose Level (mg/dL)	59.543749
7	Cognitive Function	40.543283
8	Bone Density (g/cm²)	24.083763
9	Vision Sharpness	23.747346
10	Hearing Ability (dB)	22.408642
11	Stress Levels	5.516319
12	Pollution Exposure	4.086307
13	Sun Exposure	3.951976

We used two methods to detect multicollinearity among numerical variables: a **correlation heatmap and Variance Inflation Factor (VIF) scores**. Based on these results, we identified a high correlation between “BMI,” “Height,” and “Weight.” Since BMI is derived from height and weight, we decided to remove “Height” and “Weight” from the dataset.

Other variables also showed some correlation, but without an obvious intuitive relationship. As we plan to conduct feature selection later, we opted to retain these variables for now.

Outliers Detection

We used the IQR method to detect outliers, setting the thresholds as follows:

- **Lower Bound:** Q1 – 1.5 × IQR
- **Upper Bound:** Q3 + 1.5 × IQR

The number of outliers detected for each variable is as follows:

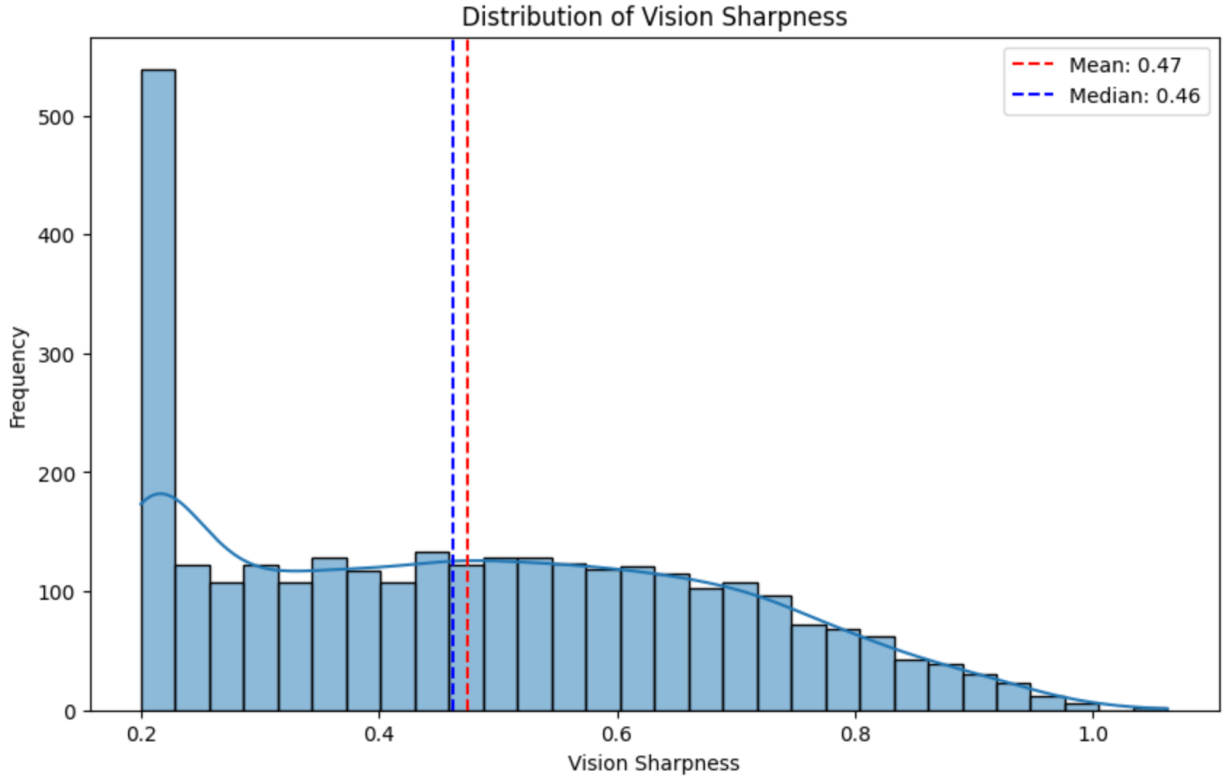
Number of outliers in each column:

	Feature	Number of Outliers
0	Cholesterol Level (mg/dL)	19
1	BMI	16
2	Blood Glucose Level (mg/dL)	13
3	Bone Density (g/cm²)	0
4	Vision Sharpness	0
5	Hearing Ability (dB)	9
6	Cognitive Function	11
7	Stress Levels	0
8	Pollution Exposure	0
9	Sun Exposure	0
10	Age (years)	0
11	Systolic	4
12	Diastolic	6

Given that we have 3,000 observations and a maximum of only 19 outliers per variable, the impact on our model is minimal. Additionally, it's reasonable to expect natural variation in biometric or lifestyle measurements, so these outliers likely represent valid data rather than errors.
Therefore, we chose not to remove any outliers.

Skewness

We plotted the distribution histograms for all numerical predictors and found that the only irregularity was in "**Vision Sharpness**." However, the difference between the mean and median is only 0.1, which is negligible. Therefore, we chose **not to make any adjustments** to address the slight skewness.



Model Results – Linear Regression

5.15

RMSE

Model predictions deviate by about 5.15 years from actual age

0.94

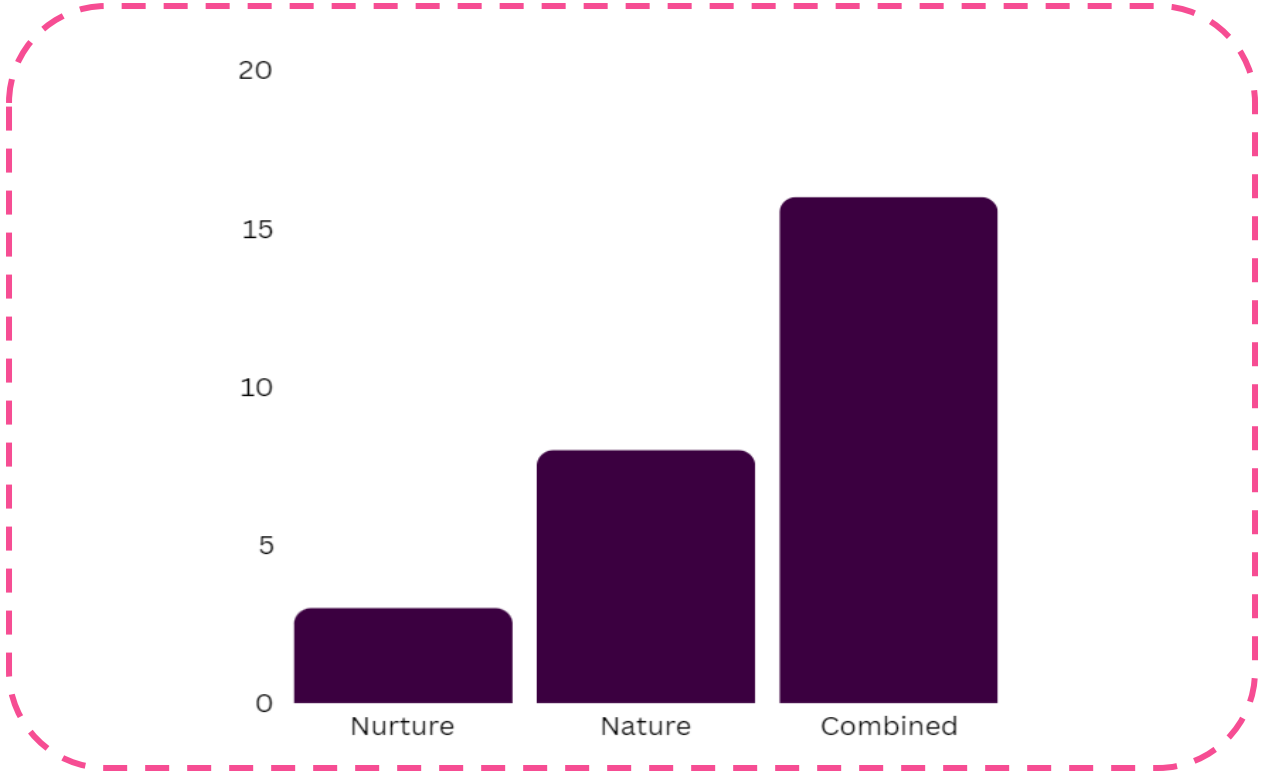
R-Squared

Selected features explain 94% of the variability

```
Decision Tree - R²: 0.8502, RMSE: 7.9398
Random Forest - R²: 0.9281, RMSE: 5.5014
Gradient Boosting - R²: 0.9325, RMSE: 5.3311
```

Results Interpretation

- Features from the 'nature' category are more impactful on predicting life expectancy. **Bone density** was found to be the most important feature.
- Previous studies showed that both nature and nurture are just as important, however.
- Remember the **Twins Study?**



Minnesota Twin Study Insights

Study Focus

Conducted by Thomas Bouchard (University of Minnesota, 1979). Studied twins raised apart to explore **Nature (genetics) vs. Nurture (environment)**

Key Findings

Genetic Influence: identical twins raised apart showed strong similarities in traits like personality, intelligence, health

Environmental impact: **differences in lifestyle** (e.g. diet, exercise) led to **different health outcomes**, showing environmental influence

Relevance to Project

Provides framework for analyzing life expectancy as a function of both genetic and environmental influences.

Stakeholders

Business Impacts:
Insurance Companies

More accurate premiums

Insurers can set premiums that better reflect each client's actual health risks. For example, a healthier client might pay less, while high-risk clients might see higher premiums.

Development of wellness programs

The life expectancy model can provide insights for insurers to create wellness initiatives that encourage healthier lifestyles among policyholders.
For example, some insurance companies already provided tools and resources that support and improve plan member health.

Competitive advantage

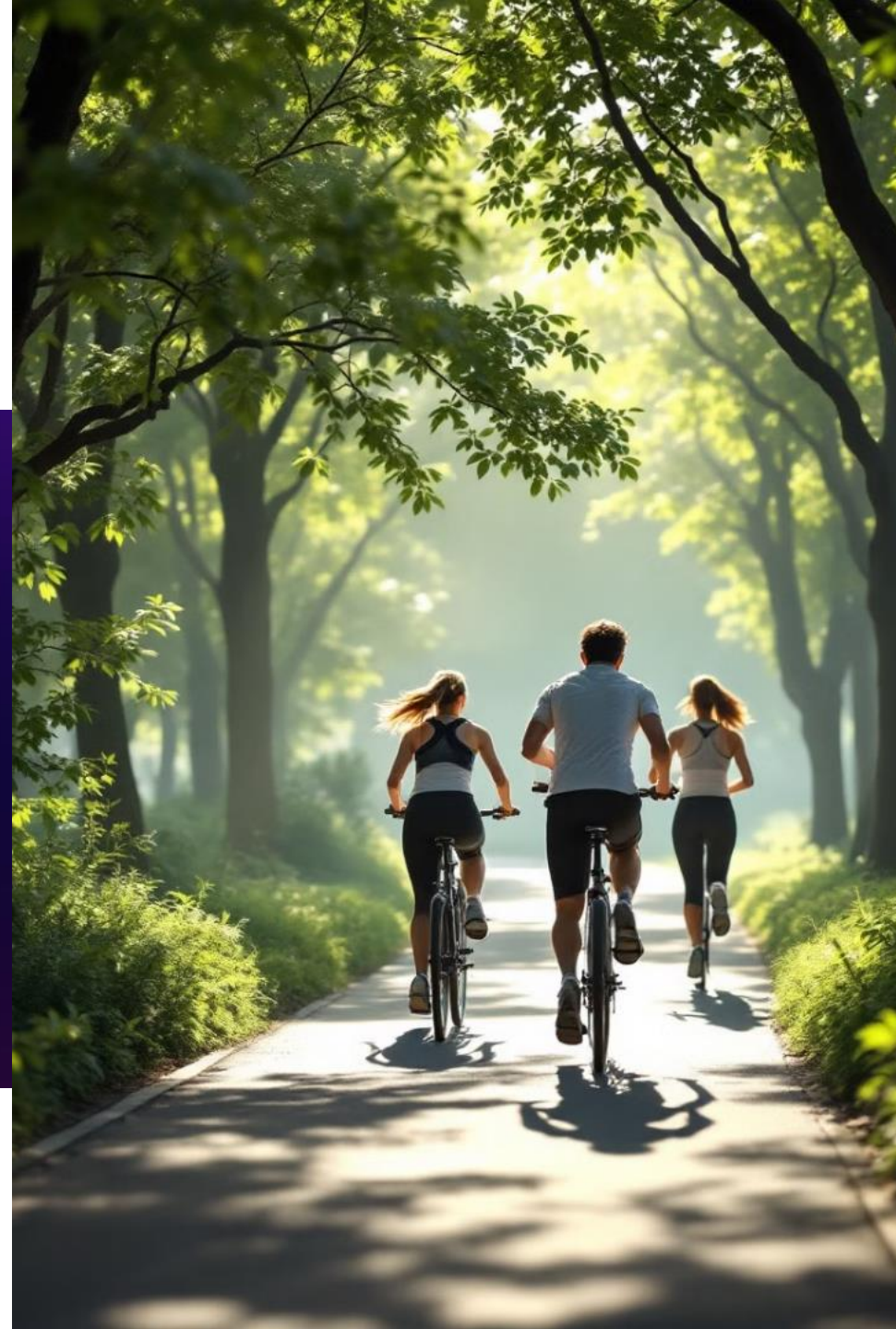
By analyzing lifestyle factors, insurers can offer tailored recommendations or rewards programs. This can make the company stand out from competitors who rely solely on traditional underwriting metrics.

Market differentiation

Younger, health-conscious customers might choose an insurer that rewards healthy behaviors like regular exercise or maintaining a healthy BMI

Business Impacts:
General Public

Increased Health Awareness	People gain clearer understanding of how lifestyle choices impact longevity
Motivation for Change	Seeing direct effects of habits on lifespan encourages positive lifestyle modifications
Preventive Action	Individuals more likely to pursue regular check-ups and adopt healthier habits



Business Impacts: Healthcare Providers

Resource Allocation

Prioritize resources for high-risk patients and promote healthier behaviors.

Patient Education

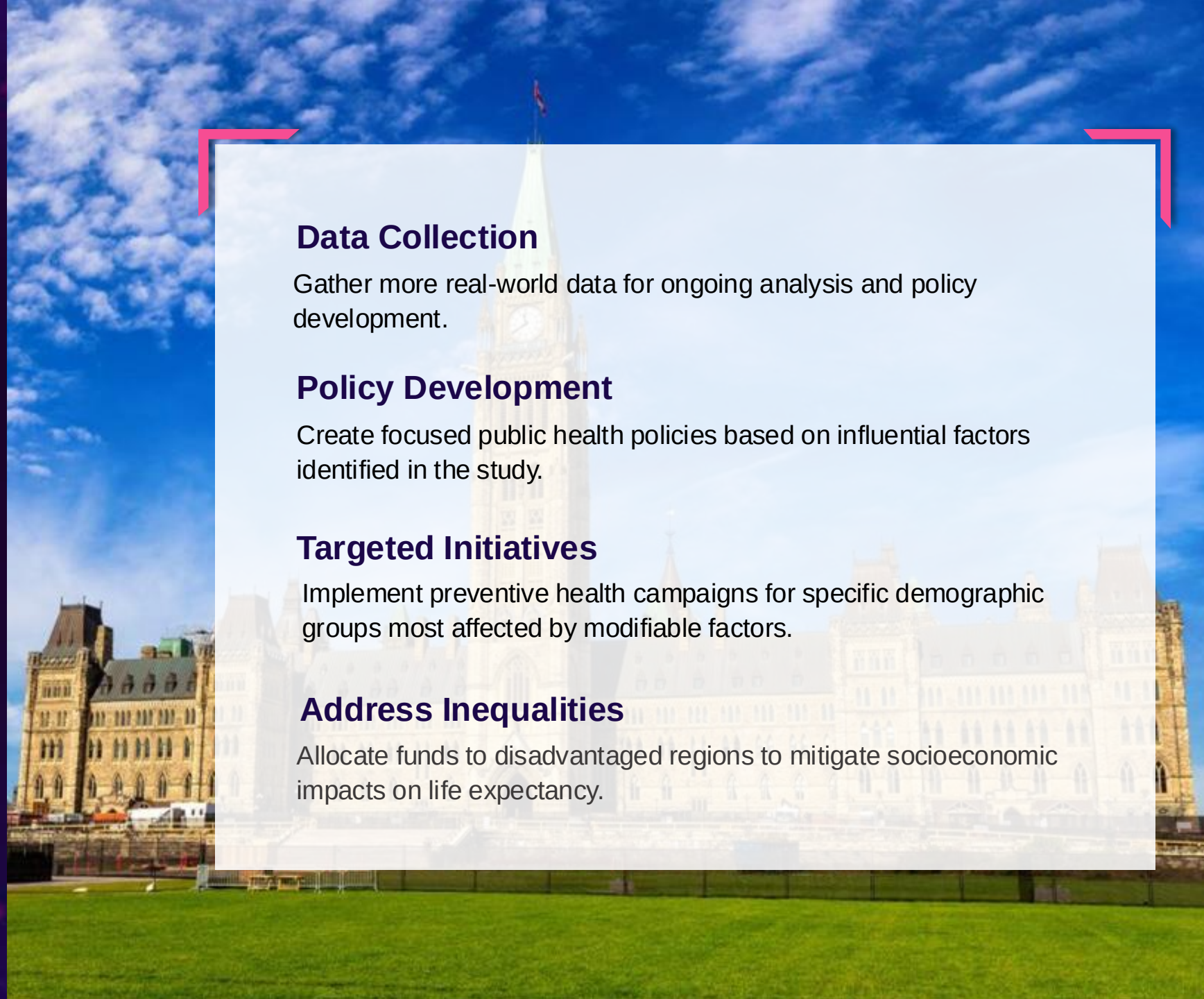
Educate patients about lifestyle changes that could lead to healthier, longer lives

Preventive Programs

Develop programs like fitness incentives or dietary workshops based on identified risk factors.

Business Impacts:

Government Agencies



Data Collection

Gather more real-world data for ongoing analysis and policy development.

Policy Development

Create focused public health policies based on influential factors identified in the study.

Targeted Initiatives

Implement preventive health campaigns for specific demographic groups most affected by modifiable factors.

Address Inequalities

Allocate funds to disadvantaged regions to mitigate socioeconomic impacts on life expectancy.

References

- Nature v nurture: research shows it’s both - UQ News - The University of Queensland, Australia
- Sources of Human Psychological Differences: The Minnesota Study of Twins Reared Apart
- https://www.npr.org/sections/health-shots/2022/12/22/1144864971/american-life-expectancy-is-now-at-its-lowest-in-nearly-two-decades