Project Brief: Heart Disease Analysis in the U.S. (2022)

For CareerFoundry Data Analytics Program – Data Immersion: Achievement 6

Project Overview

- **Motivation:** AEDs (Automated External Defibrillators) play a critical role in saving lives during cardiac emergencies. Laws and regulations requiring the presence of AEDs in public gathering places have been enacted in all fifty states in the U.S. There has been a concerning rise in heart disease incidents among individuals, including at younger ages, attributed to increasing stress from work and lifestyle factors. An AED supplier is responsible for 1) optimizing AED allocation based on the latest heart disease statistics, and 2) launching targeted public service announcements to raise awareness about heart health across the U.S.
- **Objective:**
  - To develop an up-to-date allocation plan for AEDs by analyzing the prevalence of heart disease in each state.
  - To identify and prioritize the most influential factors contributing to heart disease for the public service announcements.
- **Scope:** The project covers all public gathering places in each state of the U.S., and it will plan the AED allocation for the upcoming year.

Key Questions

Questions are raised to explore three key aspects regarding the heart disease:

1. What is the rate of suffering from heart disease in the US?
   1.1 Does the rate of suffering differ among states?
   1.2 Which states have the highest rate of suffering?

2. What are the most relevant indicators of heart disease?
   2.1 CDC indicates the three major factors are high blood pressure, high cholesterol, and smoking, is it true when the dataset is currently up to 2022?
   2.2 Are these three factors significant in every U.S. state?
   2.3 Apart from these three factors, what would be the dominant factors of heart disease?

3. What would be the most effective approach for predicting heart disease?
   3.1 Which are suitable models for prediction?
   3.2 Which factors need to be included in models (partially or entirely)?
   3.3 What level of accuracy can be expected?

4. Privacy and Ethics Questions
   4.1 Which law governs privacy when using medical information?
   4.2 Are there any legal considerations when utilizing data, such as gender?

Data source

The dataset originally comes from 2022 annual CDC (Centers for Disease Control and Prevention) survey, including data from over 400,000 adults regarding their health status. It is a major component of the Behavioral Risk Factor Surveillance System (BRFSS).

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data

Data Collection Method:

This is administrative data collected annually through telephone surveys conducted by the BRFSS, focusing on the health status of U.S. residents. This dataset is current up to 2022.

Data Contents:

The dataset contains 40 variables associated with the heart disease, a leading cause of death across various demographics in the U.S., including parameters such as blood pressure, cholesterol, smoking, diabetes status, obesity etc. The original CDC data encompasses approximately 300 variables, with 40 most relevant ones selected in this dataset.

Limitations & Ethics:

Data from the CDC is generally expected to be free from ethical issues. Considering the data collection method, this dataset may contain potential limitations including incomplete responses, the likelihood of undercounting certain groups, and, conversely, overcounting in some cases. Additionally, since the results from telephone interviews are manually recorded, there may be a degree of inaccuracy in the data.

Relevancy:

This data set aligns closely with the project objectives and hypotheses. It provides the insight into health status (via 40 indicators which are most relevant to the heart disease) of over 400,000 adults across every U.S. state, updated to 2022. This information can be used to identify significant indicators affecting the possibility of heart disease.

Reason for Data Selection:

This data set meets the data requirement stated in the A6 project brief. Additionally, with the increasing stress from work and life, individuals' health conditions are deteriorating, leading to more frequent occurrences of heart disease at younger age. Identifying the most relevant factors for heart disease is critical and beneficial for healthcare initiatives.

Data Profile

Variables & Data Types:

Variables	Description	Time-variant/-invarian	Structured/Unstructure	Qualitative/Quantitativ
State	Name of the state	Time-invariant	Structured	Qualitative, Norminal
Sex	Gender	Time-invariant	Structured	Qualitative, Norminal
GeneralHealth	A general health condition	Time-variant	Structured	Qualitative, Ordinal
PhysicalHealthDays	Poor physical health days in past 30 days	Time-variant	Structured	Quantitative, Discrete
MentalHealthDays	Poor mental health days in past 30 days	Time-variant	Structured	Quantitative, Discrete
LastCheckupTime	How long since last visit doctor?	Time-variant	Structured	Qualitative, Ordinal
PhysicalActivities	Any sports in the past month	Time-variant	Structured	Qualitative, Nominal
SleepHours	Average hours of sleep	Time-invariant	Structured	Quantitative, Discrete
RemovedTeeth	How many teeth removed	Time-invariant	Structured	Qualitative, Ordinal
HadHeartAttack	Ever had a heart attack?	Time-invariant	Structured	Qualitative, Nominal
HadAngina	Ever had angina?	Time-invariant	Structured	Qualitative, Nominal
HadStroke	Ever had a stroke?	Time-invariant	Structured	Qualitative, Nominal
HadAsthma	Ever had asthma?	Time-invariant	Structured	Qualitative, Nominal
HadSkinCancer	Ever had skin cancer?	Time-invariant	Structured	Qualitative, Nominal
HadCOPD	Ever had COPD (chronic obstructive pulmonary disease)?	Time-invariant	Structured	Qualitative, Nominal
HadDepressiveDisorder	Ever had depressive disorder?	Time-invariant	Structured	Qualitative, Nominal
HadKidneyDisease	Ever had kidney disease?	Time-invariant	Structured	Qualitative, Nominal
HadArthritis	Ever had arthritis?	Time-invariant	Structured	Qualitative, Nominal
HadDiabetes	Ever had diabetes?	Time-invariant	Structured	Qualitative, Nominal
DeafOrHardOfHearing	Ever deaf or hard of hearing?	Time-invariant	Structured	Qualitative, Nominal
BlindOrVisionDifficulty	Ever blind or vision difficulty?	Time-invariant	Structured	Qualitative, Nominal
DifficultyConcentrating	Ever difficulty concentrating?	Time-invariant	Structured	Qualitative, Nominal
DifficultyWalking	Ever difficulty walking or climbing?	Time-invariant	Structured	Qualitative, Nominal
DifficultyDressingBathing	Ever difficulty dressing or bathing?	Time-invariant	Structured	Qualitative, Nominal
DifficultyErrands	Ever difficulty errands?	Time-invariant	Structured	Qualitative, Nominal
SmokerStatus	Levels of smoker status	Time-invariant	Structured	Qualitative, Ordinal
ECigaretteUsage	Ever use E Cigarette?	Time-invariant	Structured	Qualitative, Ordinal
ChestScan	Ever chest scan?	Time-invariant	Structured	Qualitative, Nominal
RaceEthnicityCategory	Category of race ethnicity	Time-invariant	Structured	Qualitative, Ordinal
AgeCategory	Category of age	Time-invariant	Structured	Qualitative, Ordinal
HeightInMeters	Height in meters	Time-invariant	Structured	Quantitative, Continuous
WeightInKilograms	Weight in kilograms	Time-invariant	Structured	Quantitative, Continuous
BMI	Body mass index	Time-invariant	Structured	Quantitative, Continuous
AlcoholDrinkers	Is a alcohol drinker?	Time-invariant	Structured	Qualitative, Nominal
HIVTesting	Ever HIVTesting ?	Time-invariant	Structured	Qualitative, Nominal
FluVaxLast12	Flu vaccination during last 12 months?	Time-variant	Structured	Qualitative, Nominal
PneumoVaxEver	Ever pneumonia shot?	Time-invariant	Structured	Qualitative, Nominal
TetanusLast10Tdap	Tetanus shot in the past 10 years?	Time-variant	Structured	Qualitative, Ordinal
HighRiskLastYear	Had high risk last year?	Time-variant	Structured	Qualitative, Nominal
CovidPos	Ever Covid positive?	Time-invariant	Structured	Qualitative, Nominal

Data Consistency Check & Basic Statistics:

Variables	Count	Mean	std	Min.	25%	50%	75%	Max
State	246022							
Sex	246022							
GeneralHealth	246022							
PhysicalHealthDays	246022	4.12	8.41	0	0	0	3	30
MentalHealthDays	246022	4.17	8.20	0	0	0	4	30
LastCheckupTime	246022							
PhysicalActivities	246022							
SleepHours	246022	7.02	1.44	1	6	7	8	24
RemovedTeeth	246022							
HadHeartAttack	246022							
HadAngina	246022							
HadStroke	246022							
HadAsthma	246022							
HadSkinCancer	246022							
HadCOPD	246022							
HadDepressiveDisorder	246022							
HadKidneyDisease	246022							
HadArthritis	246022							
HadDiabetes	246022							
DeafOrHardOfHearing	246022							
BlindOrVisionDifficulty	246022							
DifficultyConcentrating	246022							
DifficultyWalking	246022							
DifficultyDressingBathing	246022							
DifficultyErrands	246022							
SmokerStatus	246022							
ECigaretteUsage	246022							
ChestScan	246022							
RaceEthnicityCategory	246022							
AgeCategory	246022							
HeightInMeters	246022	1.71	0.11	0.91	1.63	1.70	1.78	2.41
WeightInKilograms	246022	83.62	21.32	28.12	68.04	81.65	95.25	292.57
BMI	246022	28.67	6.51	12.02	24.27	27.46	31.89	97.65
AlcoholDrinkers	246022							
HIVTesting	246022							
FluVaxLast12	246022							
PneumoVaxEver	246022							
TetanusLast10Tdap	246022							
HighRiskLastYear	246022							
CovidPos	246022							

Deliverables

- Create a narrative to communicate the research findings and insights in the form of a Tableau Storyboard.
- Create a GitHub repository for this project.
- Turn this project into a project case study in the data analyst portfolio.

**This is a fabricated business case/scenario for personal portfolio purpose.*