TEAM 6
HEALTH DATA

Samriddhi Matharu, Shaswat Sharma,
Tehkhum Sultanali, Erica Xue

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Dataset

- Kaggle: Indicators of Heart Disease (CDC BRFSS 2022)
- 246,022 U.S. adults • 50 health, behavior, & demographic variables

Why It Matters

- In the U.S., heart disease kills someone every **34** seconds (CDC, 2024)

Strongest Takeaway

Across 246,000 adults, we found that **age is the most dominant predictor** of heart-attack risk, **BMI alone explains almost nothing**, and **meaningful risk only emerges when health and behavioral factors are layered together.** Patterns in sleep, mental/physical health, and lifestyle form clear clusters that reveal who is truly at elevated risk, showing that **multi-factor profiles, not single metrics, provide the most actionable insight.**

PROBLEM CONTEXT

Why It Matters

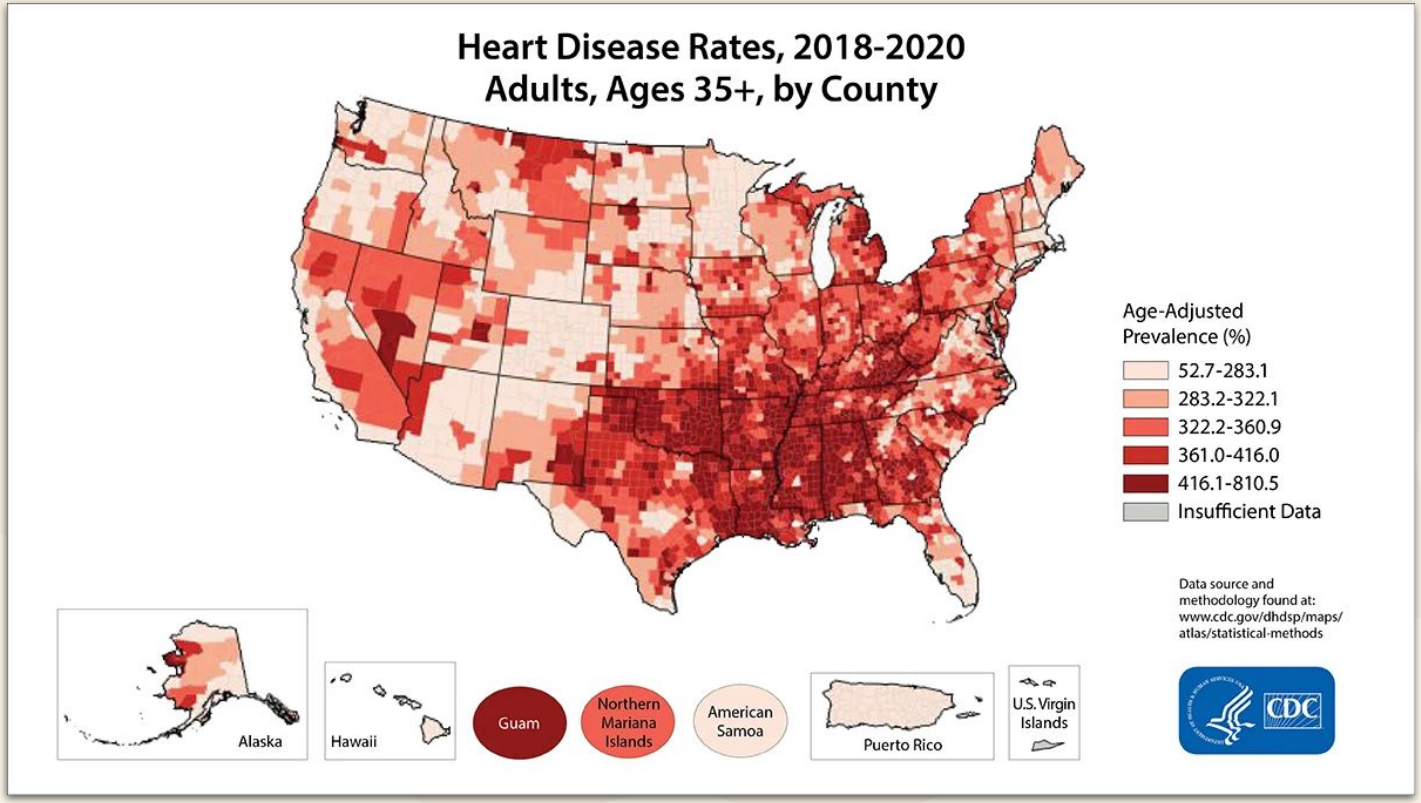
- Heart disease is the #1 cause of death across most U.S. groups
- Heart disease cost about $417.9 billion from 2020-2021 (CDC, 2024)

Who Cares

- Healthcare, public health agencies, policymakers
- Everyone, because heart disease impacts people across all demographics

What This Supports

- Targeted prevention & resource allocation


Heart Disease Rates, 2018-2020
Adults, Ages 35+, by County
Age-Adjusted Prevalence (%)
52.7-283.1
283.2-322.1
322.2-360.9
361.0-416.0
416.1-810.5
Insufficient Data
Data source and methodology found at: www.cdc.gov/dhdsp/maps/atlas/statistical-methods
Alaska
Hawaii
Guam
Northern Mariana Islands
American Samoa
Puerto Rico
U.S. Virgin Islands
CDC

OUR DATA

Source & Scope

- Drawn from CDC BRFSS 2022: conduct 400,000+ adult interviews annually across all 50 states, D.C., and 3 territories
- Kaggle provides **a cleaned subset focused on personal key indicators of heart disease**

Main Variables Used

- **Health indicators:** BMI, SleepHours, PhysicalHealthDays, MentalHealthDays
- **Cardiovascular outcomes:** HadHeartAttack, HadAngina, HadStroke
- **Demographics:** AgeCategory, Sex, RaceEthnicityCategory
- **Behavioral factors:** SmokerStatus, AlcoholDrinkers

Engineered features: BMI category, sleep buckets, *high_bmi_or_poor_health*, composite risk

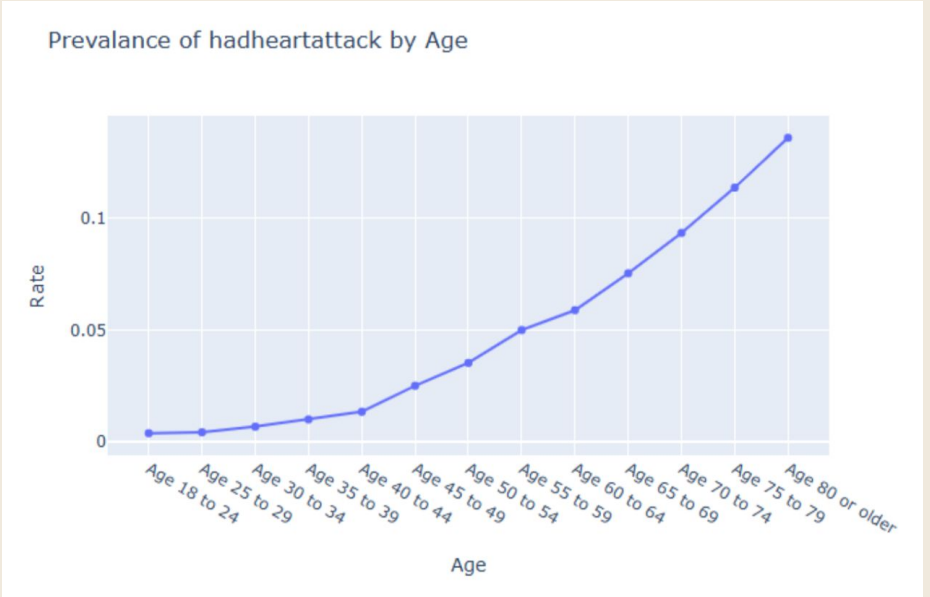
Early Data Characteristics

- **0 missing values** in the cleaned file
- Age reported as **categorical range**, not exact years
- Heart-attack cases are **sparse** (~5.5%)
- Self-reported measures → **potential bias/noise**

OUR DATA

Key Variable Relationships

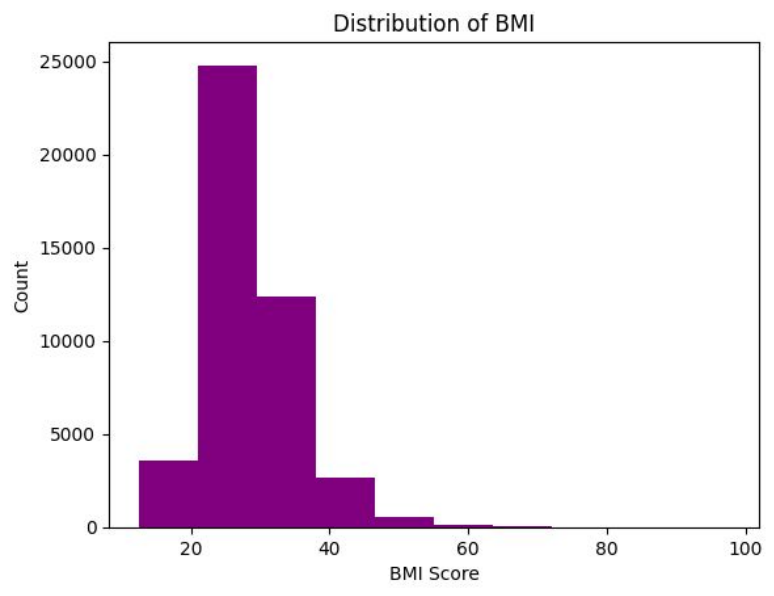
- BMI alone shows almost **no predictive value** ($R^2 \approx 0.001$)
- Risk increases when BMI **combines with age or poor health/mental health days**
- **Males show higher rates** in overweight/obese BMI groups

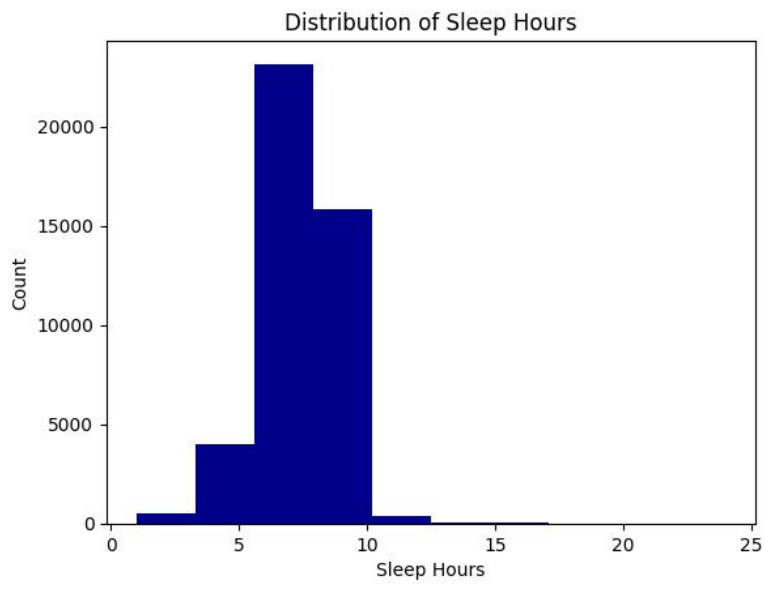


Early Exploration and Data Preparation

- Cleaned Dataset (Standardized yes/no to 1/0, normalized categories)
- Explored distributions of core variables
- Identified Initial health patterns
- Notificed skewed variables and some early trends:
  - BMI is right skewed
  - Sleep hours cluster around 6-8 hours
  - Smoking and e-cigarette usage remains common across all ages groups
  - Heart attack prevalence is low, but rises sharply with age.

Early Exploration and Data Preparation

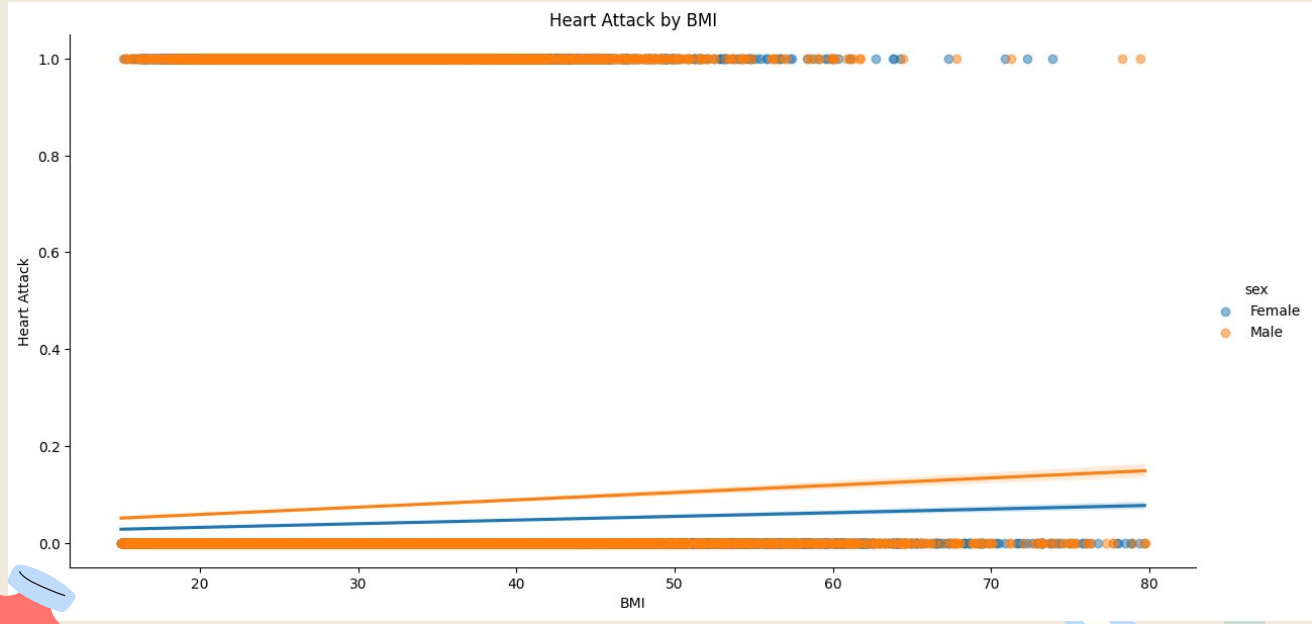




Intermediate Visualizations & Insights

BMI alone is not predictive

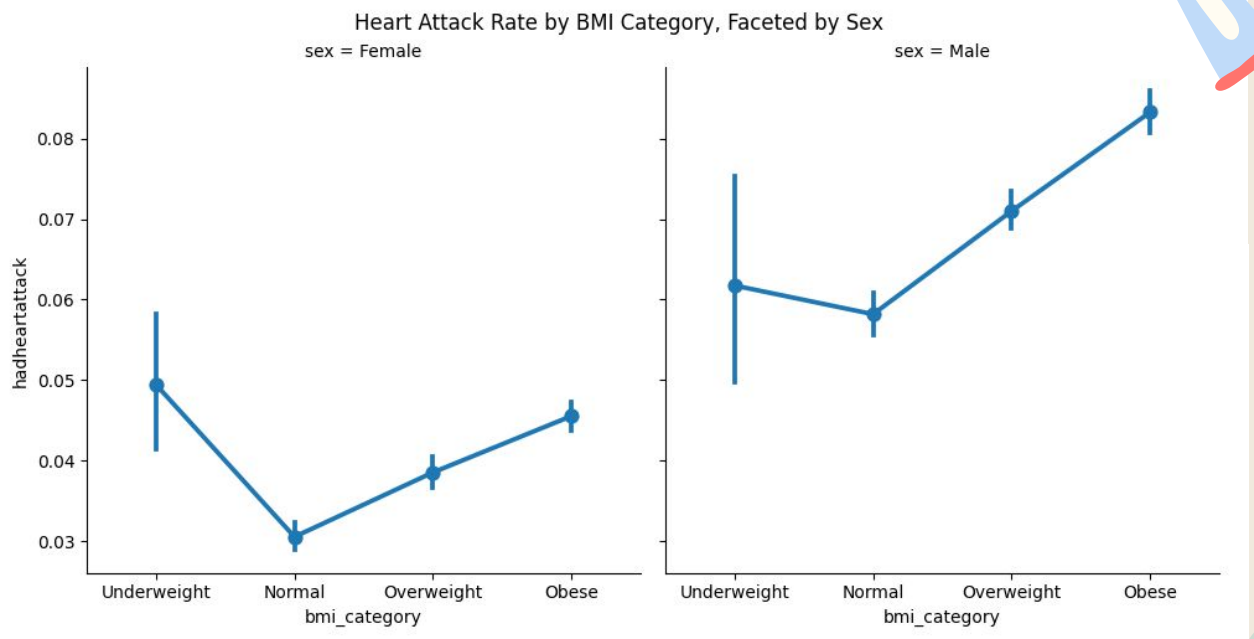
- Logistic regression: slope = 0.0011, R^2 = 0.001



Intermediate Visualizations & Insights

Differences in Sex matter more than BMI alone

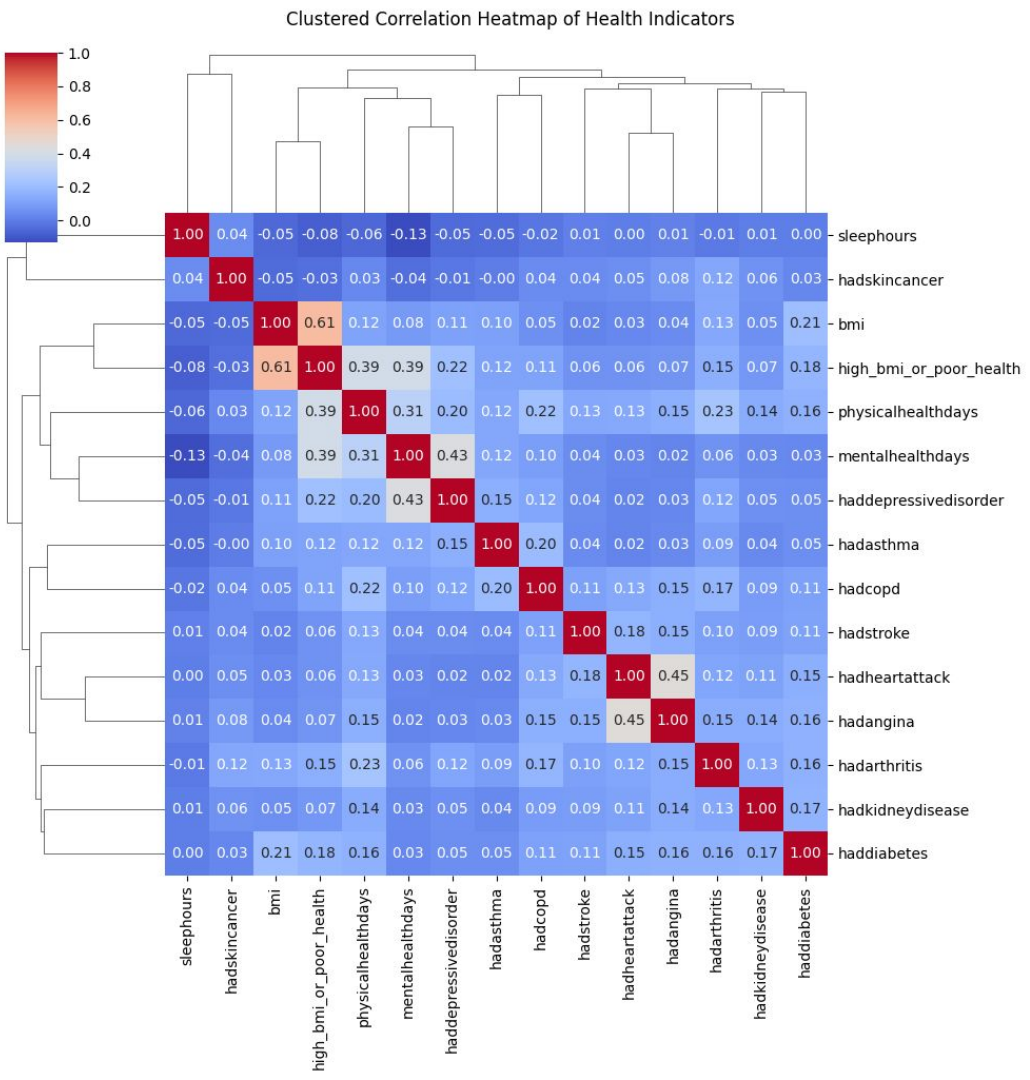
- Females: Have a slight upward trend in heart attacks across all BMI categories
- Males: Remain flat at a lower BMI but sharply rises at the overweight/obese categories.



Intermediate Visualizations & Insights

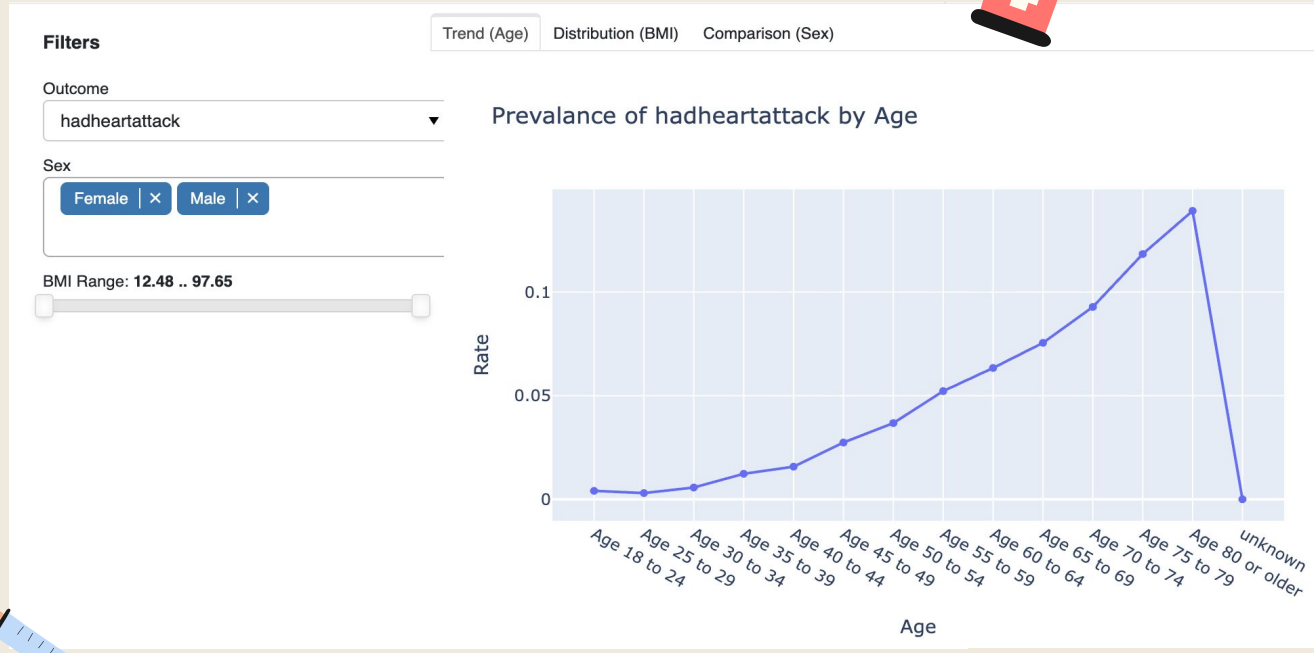
A Correlation Heatmap showed weak single variable relationships

- Heart disease is Multi factored



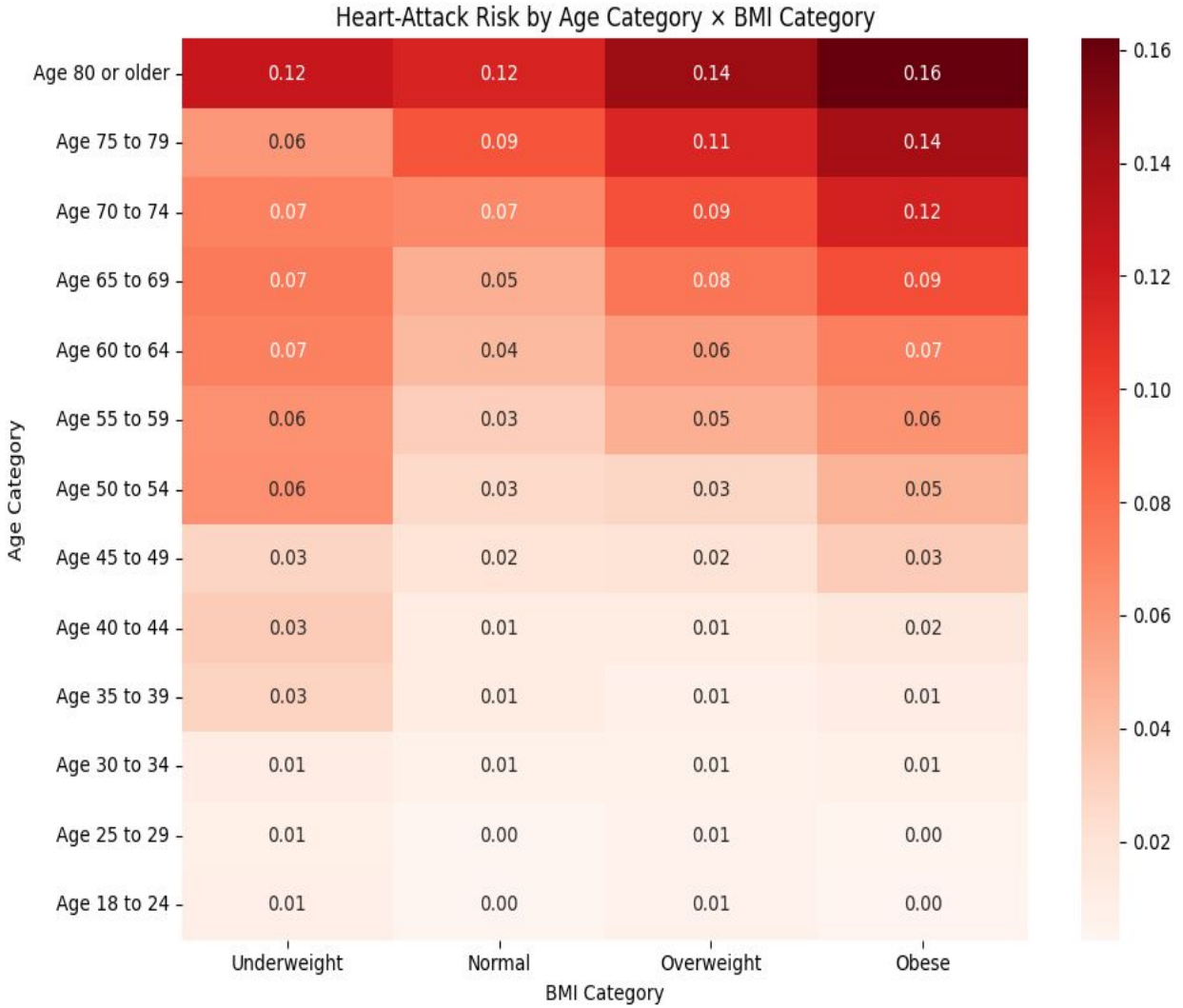
Interactive Dashboard

- Built with **Panel + Plotly**
- Features:
  - Select Outcome (heartattack, stroke, diabetes)
  - Filter by Sex
  - Adjust BMI range
  - Dynamic tabs (age trends, BMI distributions, sex comparisons)
- Helped uncover demographic patterns instantly



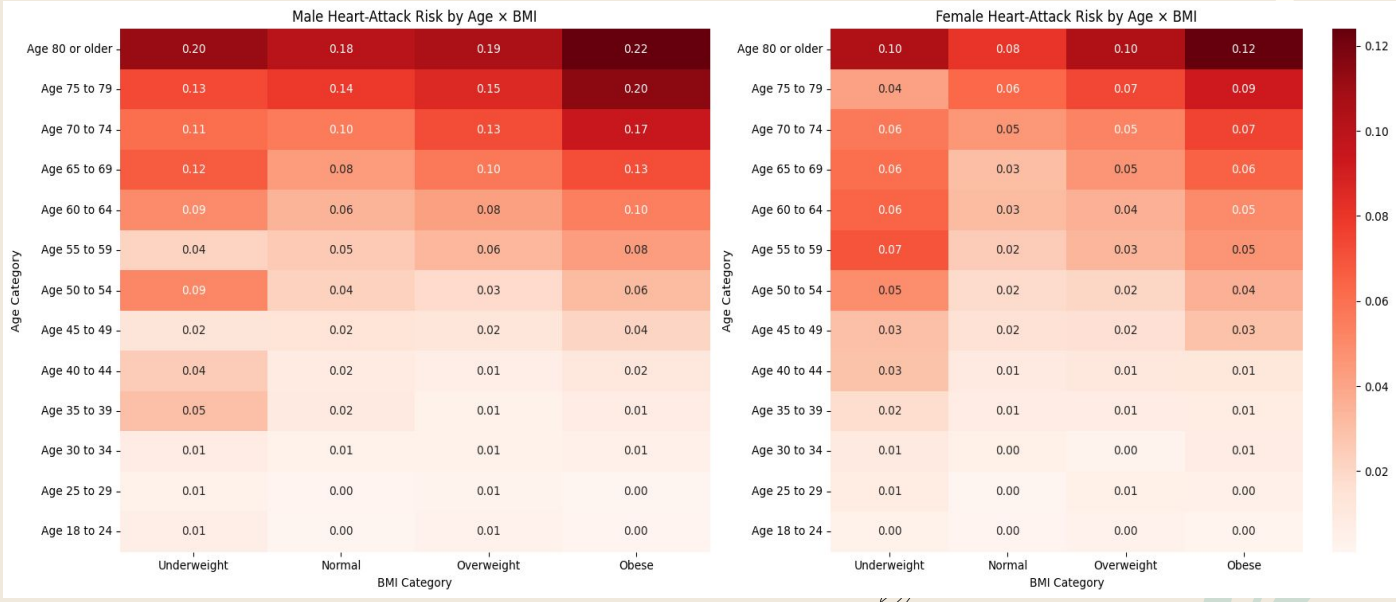
Heart Attack Risk by Age x BMI Category

- Age is the dominant predictor of heart attack risk
- BMI adds a secondary increase within each age group
- Highest Prevalence are Obese adults over the age of 65



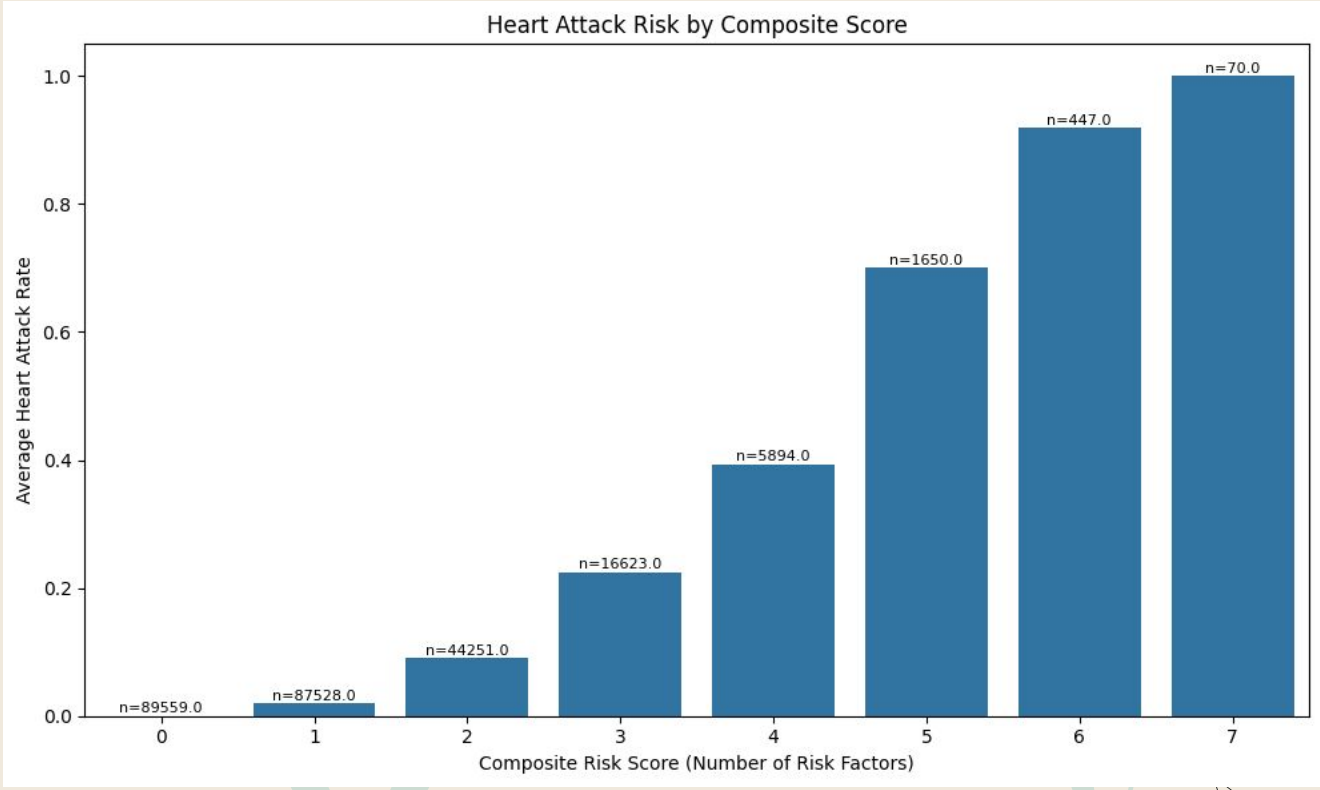
Sex Differences across Age x BMI

- Men show higher risk than woman across nearly every age x BMI combination.
- Female risk grows more steadily, while male risk jumps sharply with age and obesity
- Supports targeted clinical outreach, especially for older men.



Composite Risk Score Visualization

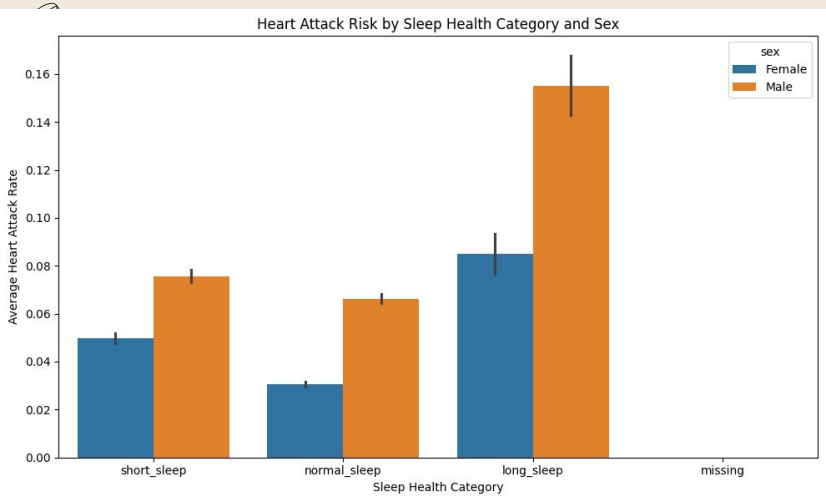
- Combining risk factors produces a multiplicative increase in prevalence.
- Risk rises slowly for 0 to 3 factors, then spikes steeply at 5+ factors.



INSIGHTS & RECOMMENDATIONS

- Age is the dominant predictor of heart-attack prevalence.
  - Risk jumps significantly after age 55.
  - Healthy-lifestyle interventions should target mid-life adults before risk accelerates.
- Single metrics (like BMI) are insufficient.
  - BMI alone explains almost none of the variance.
  - Composite risk scores should replace single-factor analysis.
- Behavioral clustering matters.
  - Poor sleep, poor mental health, poor physical health, and high BMI cluster strongly.
  - Holistic wellness programs could yield high ROI.

INSIGHTS & RECOMMENDATIONS



- Men consistently show higher prevalence across BMI and age categories.
  - Male-targeted preventive outreach may reduce cardiac events.
- Sleep is an underrated risk indicator.
  - Long sleepers show substantially higher heart-attack rates, followed by short sleepers.
- Combined risk factors produce multiplicative (not additive) effects.
  - Composite risk score visualization shows steep, escalating risk.
  - Organizations should track risk layering, not isolated variables.

LIMITATIONS & NEXT STEPS

Limitations

- Self-reported data may contain bias (height, weight, behavior)
- Data does not have clinical measurements like cholesterol, blood pressure, blood glucose
- Heart attack outcome is self-reported and binary

Next Steps

- Create a polished web application of our findings
- Build a machine-learning classifier (logistic regression or random forest)
- Introduce geographical mapping by state
- Explore causal inference (e.g., DoWhy)
- Incorporate time-series trends using multiple survey years

THANK YOU!