MISM 6210

Information Visuals and Dashboards for Business
Spring 2023
Final Project

Identifying Key Indicators for Heart Disease

An approach to identify and mitigate key indicators

By:
Dhanush Srinath
Dharmaja Patel
Akhil Khajuria

Table of Contents

Introduction

Heart diseases are one of the leading causes of death worldwide and a major public health concern in the United States. According to the American Heart Association, cardiovascular disease (CVD) is responsible for approximately 836,546 deaths per year, which accounts for one in every three deaths in the United States. Heart Disease is still the leading cause of deaths in the United States. Despite the numerous advancements in medical research and treatment options, heart disease continues to pose a significant threat to public health. Therefore, it is important to identify the key indicators of heart disease and develop effective strategies for prevention and management.

The motivation for studying the key indicators of heart disease in the United States is driven by the significant impact it has on the population's health and well-being. Understanding the factors that contribute to the onset and progression of heart disease can help healthcare professionals and policymakers develop effective prevention and management strategies. Additionally, by identifying the key indicators of heart disease, individuals can take proactive measures to reduce their risk factors and improve their overall health.

Previous research has investigated the key indicators of heart disease in the United States, providing valuable insights into the problem. One study conducted by the Centers for Disease Control and Prevention (CDC) identified hypertension, high cholesterol, and smoking as the three major risk factors for heart disease. Another study conducted by the American Heart Association identified additional risk factors, including diabetes, obesity, physical inactivity, and poor diet. These studies have contributed to the understanding of the problem, however, despite these findings, many individuals still struggle to manage their risk factors for heart diseases.

Target Audience

Our research and analysis is directed to insurance companies who can use this for better prediction of heart disease and prevention.

Dataset Description

For our visualization project on identifying key indicators of heart disease in the United States, we sourced our data from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) Behavioral Health Indicators. This dataset was selected as it contains comprehensive information on various health behaviors, risk factors, and chronic conditions, including heart disease, that affect the US population. However, the dataset was very large, messy, and had many missing values, which required extensive data standardization and cleaning before we could perform visualizations.

Data Cleaning and Standardization:

The dataset we obtained from BRFSS contained 330 columns, making it challenging to perform feature selection or parameter tuning for our visualization project. Additionally, the dataset was collected through a survey of individual respondents, leading to several data quality issues such as missing values and inconsistent responses. Therefore, we performed extensive data standardization to clean the dataset and extract only the columns that were necessary for visualizing key indicators of heart disease.

One of the major steps we took for example, was to standardize the column DIABETE3, which had values like "I don't know" or "Maybe" that were not useful for analysis. We extracted only the binary variables for this column, indicating whether an individual had diabetes or not. This step allowed us to reduce the number of columns and eliminate non-informative data.

Due to the large number of columns in the original dataset, we decided to manually choose the columns we felt were necessary for visualizing key indicators of heart disease. We selected 20 columns (based on research from NIH https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2922828/) that were relevant to heart disease risk factors such as high blood pressure, cholesterol, smoking status, and diabetes status. These columns included variables such as age, sex, body mass index, hypertension, and tobacco use. By selecting these key variables, we were able to focus our visualizations on the most important factors that contribute to heart disease.

Data Analysis and Methodologies:

After cleaning and standardizing the BRFSS Behavioral Health Indicators dataset, we performed logistic regression and random forest models to identify key indicators of heart disease in the United States. We also performed cluster analysis to gain further insights into the factors that contribute to heart disease. We performed predictive modeling to analyze and predict the probability of an individual having a heart disease

Logistic Regression:

We used heart_disease as the target variable and all other factors as independent variables to perform logistic regression on the cleaned dataset. Our model achieved an accuracy score of 0.906 and a precision score of 0.538. This result may be due to the skewness of the data, as only 10% of the respondents in the dataset had a positive flag value for heart disease. An accuracy of this value indicates that the model correctly predicts whether or not someone has a heart disease 90.6% of the time. Precision here is 53% which means that out of all the respondents, the model identified as having a heart disease, only 53%of the people had it.

Random Forest:

We also attempted to perform a random forest model on the dataset to identify key indicators of heart disease. However, the model was time-consuming even with tuning and boosting measures. Therefore, we decided to stick with the logistic regression model and the precision score of 0.53.

Cluster Analysis:

In addition to the regression models, we also performed cluster analysis to gain further insights into the factors that contribute to heart disease. We created three clusters based on the key indicators of heart disease, and the resulting clusters provided satisfactory results.
Based on the output, the clustering algorithm has created three clusters (labels 0, 1, and 2).

Cluster 0 (label 0) contains 100,236 observations with the following characteristics:

- Low prevalence of heart disease, high blood pressure, and high cholesterol
- High physical activity and consumption of fruits and vegetables
- Low alcohol consumption and low healthcare costs
- Average general health and mental health
- Low difficulty walking
- Mostly male, with a mean age of 50-54 and a high income

Cluster 1 (label 1) contains 120,763 observations with the following characteristics:

- High prevalence of heart disease, high blood pressure, and high cholesterol
- High prevalence of diabetes and smoking
- Low physical activity and low consumption of fruits and vegetables
- Low alcohol consumption and low healthcare costs
- Poor general health and mental health
- High difficulty walking
- Mostly female, with a mean age of 60-65 and a low income

Cluster 2 (label 2) contains 69,001 observations with the following characteristics:

- Moderate prevalence of heart disease, high blood pressure, and high cholesterol
- Moderate physical activity and consumption of fruits and vegetables
- Moderate alcohol consumption and healthcare costs
- Average general health and mental health
- Low difficulty walking
- Mostly female, with a mean age of 50-54 and a high income

Despite the insights gained from our logistic regression and cluster analysis, heart disease remains a prevalent issue in the United States and the findings of our visualization project showed that high blood pressure, cholesterol, smoking status, and diabetes status are key indicators of heart disease in the United States. However, it is important to note that these findings are not groundbreaking, as they have been identified in previous research as well.

Data Visualizations

The visualization tools that were used to gain further insights after creating a predictive model on heart disease indicators was *Tableau*. After creating more than 15 visualizations, 4 of them were found resourceful and insightful for indicating and explaining heart disease indicators directed towards insurance companies; as indicated through *Figure1,2,3 & 4*.

From *Figure 1 - Heart Disease Demographics*, we found the Top States prevalent with Heart Diseases in the United States to be;
1) Kansas
2) Nebraska
3) Minnesota
4) Washington
5) Texas

These States were also high in the percentage of people affected by Mental Health Conditions. Interestingly, 34% of the people diagnosed with Mental Health Conditions were diagnosed with a heart disease (*refer Figure 2*). This indicates that Mental Health is an important factor in the prevalence of heart diseases.

Also, from *Figure - 1,* we see that Females are at a higher risk of developing a heart disease than men ; with 13,833 women diagnosed with a heart disease versus 10,322 men.

This trend is also seen in heart disease by age; between males and females. From *Figure 1,* we can see that Risk of Heart Diseases increases with age after 45 years, peaking between 65-69 years for both men and women. Interestingly, the gap between men and women at risk of heart diseases increases between the ages 55 - 74 years; with women more prone to heart diseases. This is due to the fact that women have smaller arteries than men and hence coronary artery diseases develop differently and diffusely in women.

From *Figure 3 - Heart Disease Indicators*, we identified the most prominent indicators of heart diseases, from the data set. The prominent indicators are;
1) High Blood Pressure
2) High Cholesterol
3) Smoking
4) Diabetes

Additionally, Mental Health is also a prominent indicator, as seen from *Figure 2*.

For insurance companies, additional candidate profiling has also been done in *Tableau,* to selectively profile candidates and decide insurance premiums based on the above Demographic factors and indicators, as shown in *Profile 1* and *Profile 2.* The Heart Disease Risk is calculated State wise as well as compared with the National Average in order to account for differences in states and take risk values that are largest in either case.

The visualizations and inferences presented, although meaningful, are the same as outlined by different doctors and medical guidelines. With heart diseases on an increasing trend, this poses a threat to insurance companies, who spend on average $73,000 per heart disease patient , not considering heart transplants that cost over $200,000.Hence other measures must be adopted by insurance companies in order to lead individuals towards a healthy lifestyle which inturn would reduce their risks to developing a heart disease.

Recommendations

Insurance companies can significantly reduce their costs by providing incentives for preventing heart disease. On average, insurance companies spend $73,000 per person with heart disease, which is a substantial amount of money. However, even a small decrease in the risk of heart disease, by as little as 5%, can result in a significant cost reduction of approximately $3,650 per person.

By incentivizing individuals to take proactive measures to prevent heart disease, insurance companies can avoid the high costs associated with future consultations, tests, and treatments. It's a win-win situation for both insurance companies and individuals, as individuals can lead healthier lives and reduce their healthcare costs while insurance companies can save money and improve their bottom line.

Providing incentives for preventing heart disease is a smart investment for insurance companies that can lead to significant cost savings in the long run. By encouraging individuals to take care of their health, insurance companies can create a healthier population while reducing healthcare costs and improving their profitability.

Behavioral Economics For Heart Disease

Behavioral economics can be a useful framework for understanding why incentives for preventing heart disease can be effective. One of the main principles of behavioral economics is that people are often driven by immediate rewards and tend to undervalue future benefits. Which means that individuals may not always prioritize their long-term health over short-term pleasures, even if they are aware of the potential health risks.

By providing incentives for preventive measures, insurance companies can help counteract this, and encourage individuals to take proactive steps towards a healthier lifestyle. For example, offering rewards for meeting specific health goals, such as completing regular check-ups, quitting smoking, or increasing physical activity, can help individuals focus on the long-term benefits of healthy behavior and motivate them to stick to their health plans.

Behavioral economics also emphasizes the importance of framing and context in decision-making. Insurance companies can leverage this principle by framing incentives in a way that appeals to individuals' goals and aspirations. For instance, instead of simply offering a monetary reward for completing a health goal, insurance companies can emphasize the broader benefits of preventive care, such as feeling better and enjoying a higher quality of life.

Incentive Plan

The "Invest in Your Health, Reap the Rewards" incentive plan is an innovative approach to promote healthy living and prevent chronic diseases. The program is designed to encourage individuals to prioritize their health and make positive lifestyle changes by offering financial incentives for achieving their health goals, based on behavioral economics.

Incentive Program Details:

1) Sign up for the program by taking a free health checkup: By completing a health screening, individuals can identify any underlying health issues and get an initial assessment of their overall health. As a reward for signing up, participants receive $50.

2) Choose up to two goals per month: Based on their health screening, participants can select up to two health goals per month that targets specific indicators;

   Goals:
   - Clock 8,000 steps per day (Targets : BMI, Cholesterol, Blood Pressure)
   - No extra sugar (Targets : Diabetes)
   - Reduce 5lbs (Targets : BMI, Cholesterol)
   - Smoke free month (Targets : Smokers)
   - Join a stress management group (Targets : Mental health, Blood Pressure)

3) Work towards achieving goals: Participants will be encouraged to work towards achieving their health goals through support groups (virtual / in-person) for at least 1 month to form a habit and continue for 90 days to make a permanent lifestyle change.

4) Rewards for achieving goals: After completing each goal, participants receive rewards based on the level of difficulty and duration. For example, completing a 1 month challenge earns a $100 reward, completing a 2-month challenge earns a smartwatch, and completing a 3-month challenge earns $300 plus a lower insurance premium.

5) Leaving the program mid-way however, will result in retraction of the initial $50 , hence the participants who leave won't gain anything.

The rewards and levels are as indicated in *Figure 4*.

Conclusion:

Overall, our project highlights the importance of promoting a healthy lifestyle and incentivizing healthy behaviors in reducing the risk of heart disease. Health insurance companies can play a vital role in this by offering resources, support, and incentives to their customers. By doing so, we can not only improve the health outcomes of individuals but also reduce the long-term costs associated with heart disease.

In conclusion, our visualization project has provided some useful insights into the key indicators of heart diseases in the United States, and it is crucial to take action towards promoting a healthy lifestyle and incentivizing healthy behaviors to reduce the burden of heart disease. By promoting a healthy lifestyle and incentivizing healthy behaviors, we can encourage individuals to make positive lifestyle changes and improve their overall health. This can lead to a reduction in the long-term costs associated with heart disease, such as medical expenses and lost productivity. Health insurance companies can play a crucial role in this effort by offering incentives based on the risk factors identified in our visualization project.

Limitations

The initial dataset had about 72 million values, however after cleaning, the dataset had only about 230,000 values. Hence, we can say that the data is skewed. Another limitation is that the available dataset is from 2015, and hence trends in heart disease might have changed. However, the recommendations would still be valid, regardless of the year.

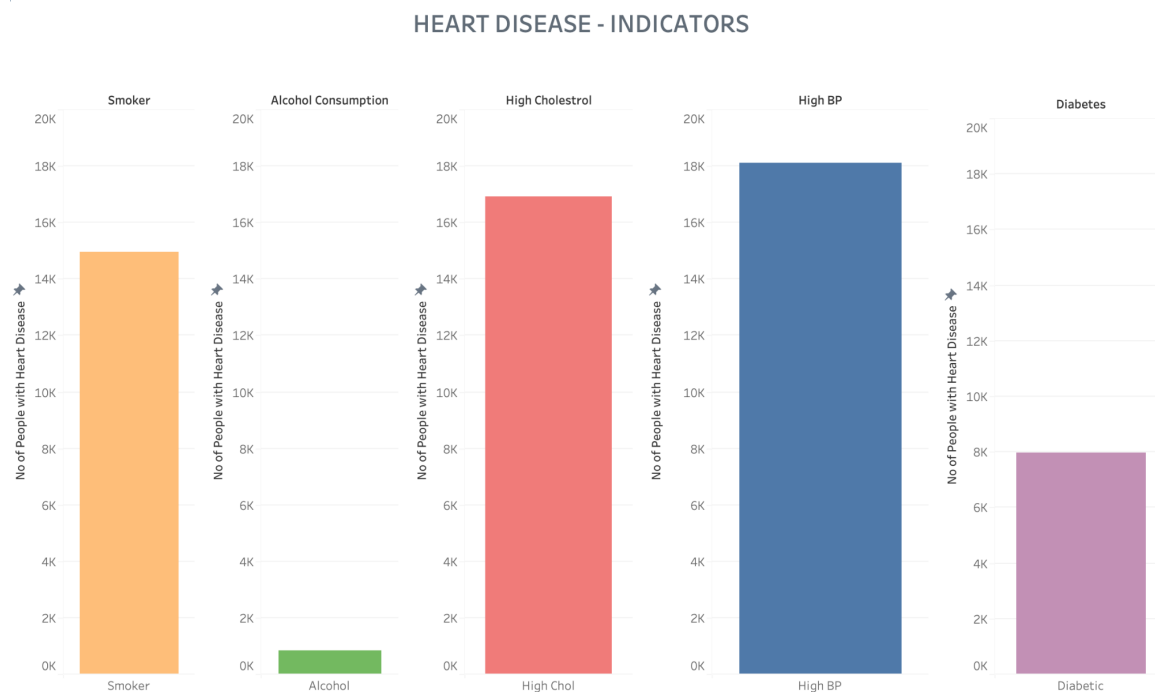As for the incentive plan, not everyone can be motivated through financial incentives.
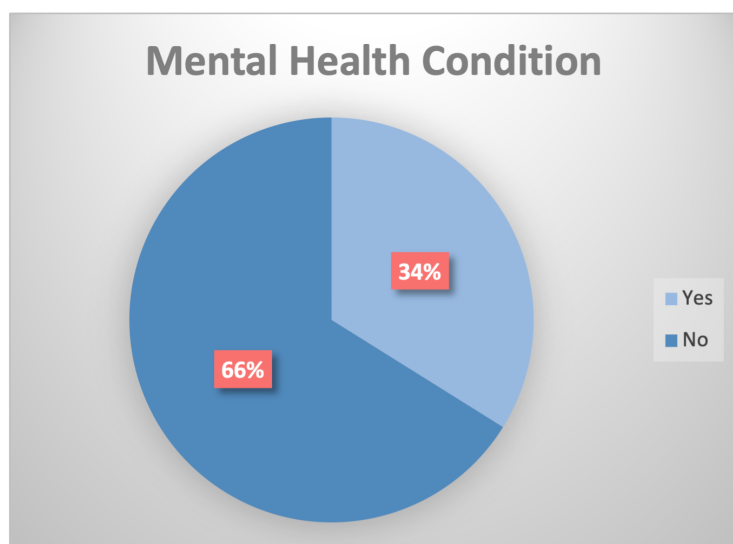
Appendix

HEART DISEASE - INDICATORS

Figure 1 : Heart Disease - Indicators

Figure 2 : Mental Health Conditions and Heart Disease

HEART DISEASE - DEMOGRAPHICS

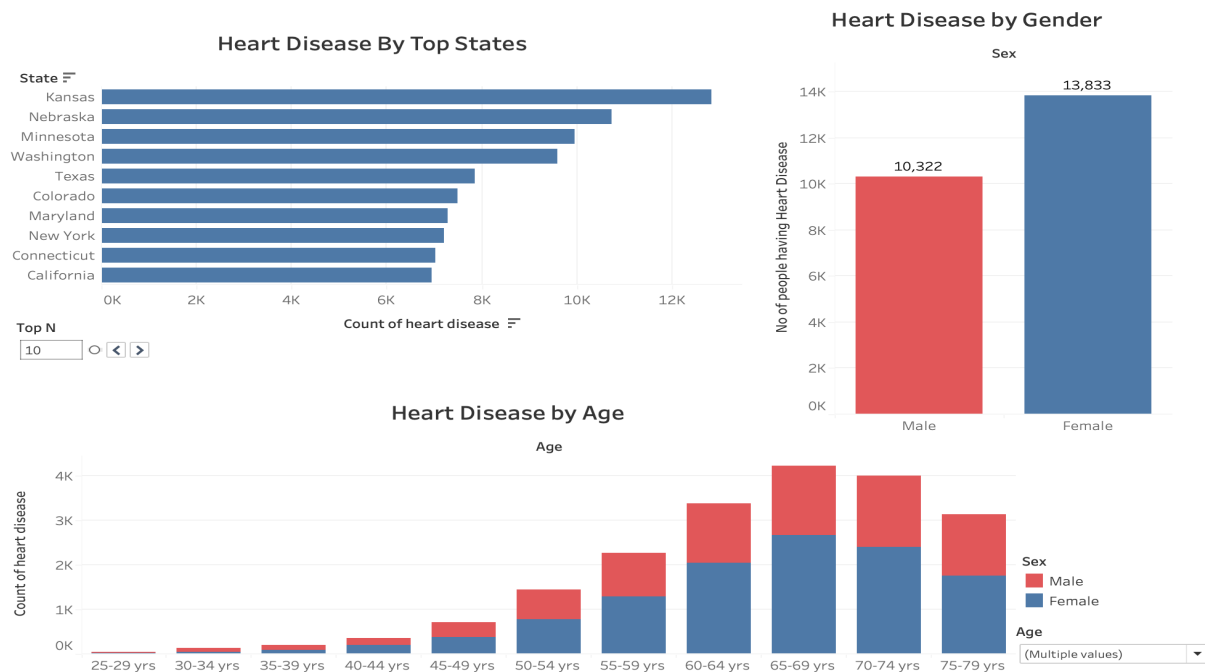### Heart Disease By Top States

Heart Disease by Gender

Heart Disease by Age

Figure 3 : Heart Disease Demographics

HEART DISEASE APPLICANT PROFILING

Heart Disease Risk - State

							Heart Disease	
Sex	Age	State	B.P	Cholestrol	Diabetes	Smoker	Can Have %	Can't Have %
Female	18-24 yrs	Massachusetts	Normal	Normal	No Diabetes	Non Smoker	1.72%	98.28%

Heart Disease Risk - National Average

						Heart Disease	
Sex	Age	B.P	Cholestrol	Diabetes	Smoker	Can Have %	Can't Have %
Female	18-24 yrs	Normal	Normal	No Diabetes	Non Smoker	0.42%	99.58%

Profile 1 : Insurance Applicant Profiling

HEART DISEASE APPLICANT PROFILING

Heart Disease Risk - State

Sex	Age	State	B.P	Cholestrol	Diabetes	Smoker	Heart Disease	
							Can Have %	Can't Have %
Male	60-64 yrs	Minnesota	High	Normal	Diabetic	Smoker	18.18%	81.82%

Heart Disease Risk - National Average

Sex	Age	B.P	Cholestrol	Diabetes	Smoker	Heart Disease	
						Can Have %	Can't Have %
Male	60-64 yrs	High	Normal	Diabetic	Smoker	15.52%	84.48%

Profile 2 : Insurance Applicant Profiling

Indicator	Risk Assessment %
NO INDICATORS ❤️❤️❤️❤️	1.84
SMOKER ❤️❤️❤️	4.44
HIGH CHOLESTROL ❤️❤️❤️	4.92
HIGH B.P ❤️❤️❤️	7.55
SMOKER + CHOLESTROL ❤️❤️	9.97
SMOKER + B.P ❤️❤️	12.22
CHOLESTROL + B.P ❤️❤️	16.28
SMOKER + CHOLESTROL + B.P ❤️	25.58

Figure 4 : Indicators - Risk Assessment Percentage

References

1) National Library of Medicine : Health Insurance and Cardiovascular Disease Risk FactorsBrooks, E. L., Preis, S. R., Hwang, S.-J., Murabito, J. M., Benjamin, E. J., Kelly-Hayes, M., Sorlie, P., & Levy, D. (2010, August). *Health Insurance and Cardiovascular Disease Risk Factors*. The American journal of medicine. Retrieved April 20, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2913281/

2) Centers for Disease Control and Prevention. (2022, July 26). *CDC - BRFSS Annual Survey Data*. Centers for Disease Control and Prevention. Retrieved April 20, 2023, from https://www.cdc.gov/brfss/annual_data/annual_data.htm

3) *Leveraging behavioral economics to improve heart failure care and ...* (n.d.). Retrieved April 20, 2023, from https://www.ahajournals.org/doi/full/10.1161/circulationaha.117.028380

4) Cancer vs. heart disease: Which costs insurers the most money? (n.d.). Retrieved April 20, 2023, from https://www.pacificprime.com/resources/news/cancer-vs.-heart-disease-which-costs-insurers-the-most-money/

5) Hare, A. J., Patel, M. S., Volpp, K., & Adusumalli, S. (2021, October 1). *The role of behavioral economics in improving cardiovascular health behaviors and outcomes - current cardiology reports*. SpringerLink. Retrieved April 20, 2023, from https://link.springer.com/article/10.1007/s11886-021-01584-2

6) Centers for Disease Control and Prevention. (2022, October 14). *Heart disease facts*. Centers for Disease Control and Prevention. Retrieved April 20, 2023, from https://www.cdc.gov/heartdisease/facts.htm#:~:text=Heart%20disease%20is%20the%20leading,groups%20in%20the%20United%20States.&text=One%20person%20dies%20every%2034,United%20States%20from%20cardiovascular%20disease.&text=About%20697%2C000%20people%20in%20the,1%20in%20every%205%20deaths.

7) Hare, A. J., Patel, M. S., Volpp, K., & Adusumalli, S. (2021, October 1). *The role of behavioral economics in improving cardiovascular health behaviors and outcomes - current cardiology reports*. SpringerLink. Retrieved April 20, 2023, from https://link.springer.com/article/10.1007/s11886-021-01584-2