Heart Health: Statistical Analysis Uncovers Most Significant Risk Factors of Coronary Heart Disease

Hersh Nanda[1,2]

[1]BASIS Chandler, Chandler, Arizona, USA

[2]Research Assistant, College of Medicine, University of Arizona, Tucson, Arizona, USA

C: (480) 276-6878 E: hersh.nanda@gmail.com

1. Introduction

Heart disease or cardiovascular disease – are two words that a patient prays not to hear during diagnosis. According to the Centers for Disease Control and Prevention (CDC[1]), cardiovascular disease (CVD) is the leading cause of death for men and women across most racial and ethnic groups in the U.S.[1] CVD is the grim reaper of diseases, accounting for approximately 655,000 deaths annually, or 25% of all deaths in America[2] (pre-COVID-19 pandemic). Besides the toll on human life, the financial impact is equally devastating. US annual cost of heart disease is $219 Billion[3] including cost of health services, medicines, and lost productivity.

The CDC reports that high blood pressure, high cholesterol, and smoking are key risk factors for heart disease and that roughly half (~47%) of Americans have at least one of these risk factors. Additionally, the CDC claims the following risk factors can also put people at higher risk for heart disease: diabetes, obesity, unhealthy diet, physical inactivity, and excessive alcohol use.

This case study reports the results of a research project to gather insights into heart disease lifestyle and risk factors using statistical analysis that is generally used to improve product and service quality in companies. This research aimed to answer the following: what does the

[1] Centers for Disease Control and Prevention. Underlying Cause of Death, 1999–2018. CDC WONDER Online Database. Atlanta, GA: Centers for Disease Control and Prevention; 2018.

[2] Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, et al. Heart disease and stroke statistics—2020 update: a report from the American Heart Association . Circulation. 2020;141(9):e139–e596.

[3] Fryar CD, Chen T-C, Li X. Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999–2010 pdf icon[PDF-494K]. NCHS data brief, no. 103. Hyattsville, MD: National Center for Health Statistics; 2012.

statistical analysis of CDC data from all US states reveal? What known facts does it affirm? What new risk factors does it uncover? What does it reveal about combined effects of these risk factors? Ultimately, the analysis revealed compelling new information which can be applied to extend human life – and it all stems from your actions.

This study analyzed data on 15 different lifestyle factors: High blood pressure, high cholesterol, diabetes, physical inactivity, obesity, smoking, fast food intake, exercise, fruit intake, vegetable intake, air pollution, chronic drinking, insufficient sleep, frequent mental distress, and median household income. Data was obtained from the CDC (minimum 100,000 patients per state), the Environmental Protection Agency (EPA), and the Stanford University Department of Computer Science (which sourced from 23andMe - a genomics and biotech company).

1. Research Methodology

Multiple statistical tests were run on the collected data, including correlation and multiple linear regression analysis to analyze the association between lifestyle factors and heart disease. Visualizations and graphs were also generated to display a visual representation of the data relationships.

4.1 Data visualization

It is essential to create a plot or graph to obtain a visual understanding of how each variable relates to each other. Scatter plots, matrix plots, and surface plots are three forms of visual data representation which were incorporated into analysis. Scatter plots display the correlation between two variables, while matrix plots display the correlations between more than two variables. A matrix plot is an array of individual scatterplots, and it is used to assess the relationship among several pairs of variables at once. 3D plots or surface plots were also generated to visualize the combined association of two predictors with one response variable.

1.1.1 Scatter Plots

Figure 1 is a scatter plot that shows the correlation between smoking rates and the percentage of population with heart disease in US states. The graph shows that there is a strong positive correlation between the two variables, meaning that as smoking rates increase, heart disease rates

are increasing simultaneously. Figure 2 shows that there is a strong positive correlation between percentage of population with heart disease and percentage of population with high blood pressure (BP) in US states. Figure 3 shows that there is a strong negative correlation between percentage of population with high BP and the mean number of exercise sessions per week per person in US states. This means that as people exercise more, the percentage of population with high BP decreases.

Figure 1: Scatterplot of percentage of adults who are smokers versus the percentage of adults who are diagnosed with heart disease

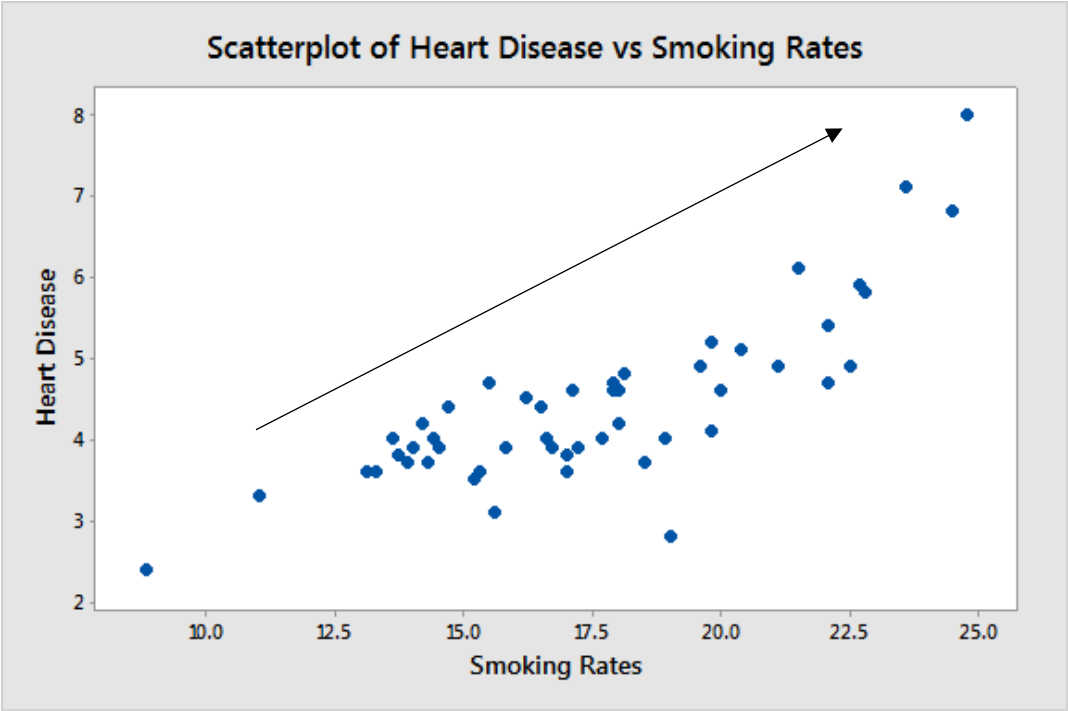


Figure 2: Scatterplot of percentage of adults who have high blood pressure versus the percentage of adults who are diagnosed with heart disease.

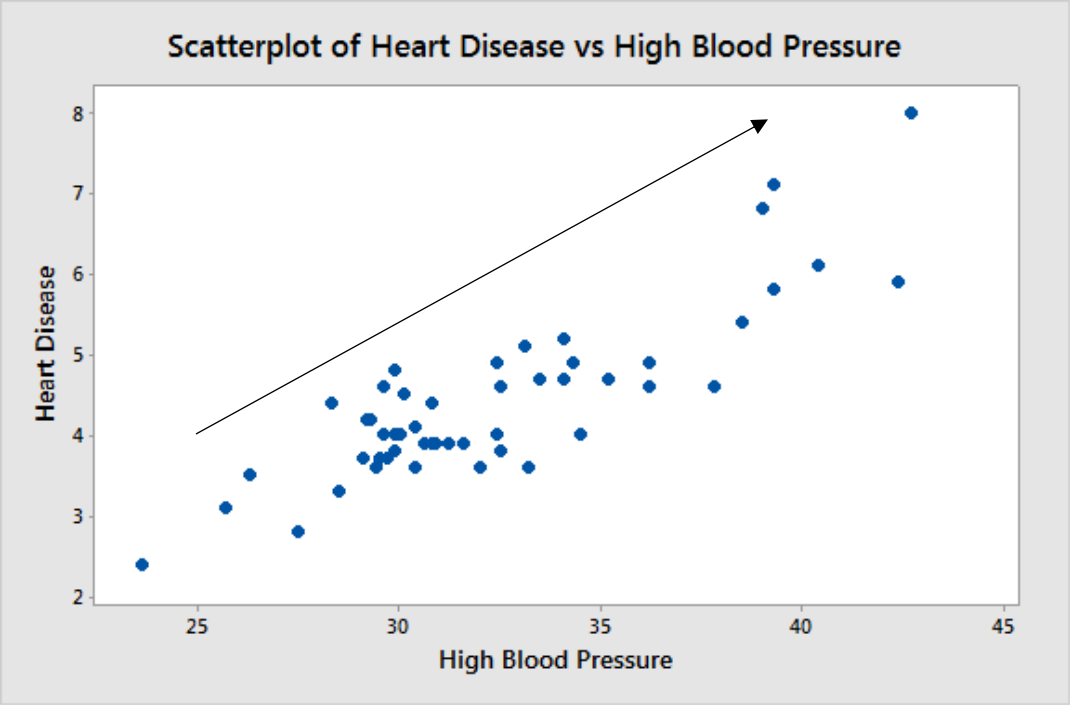
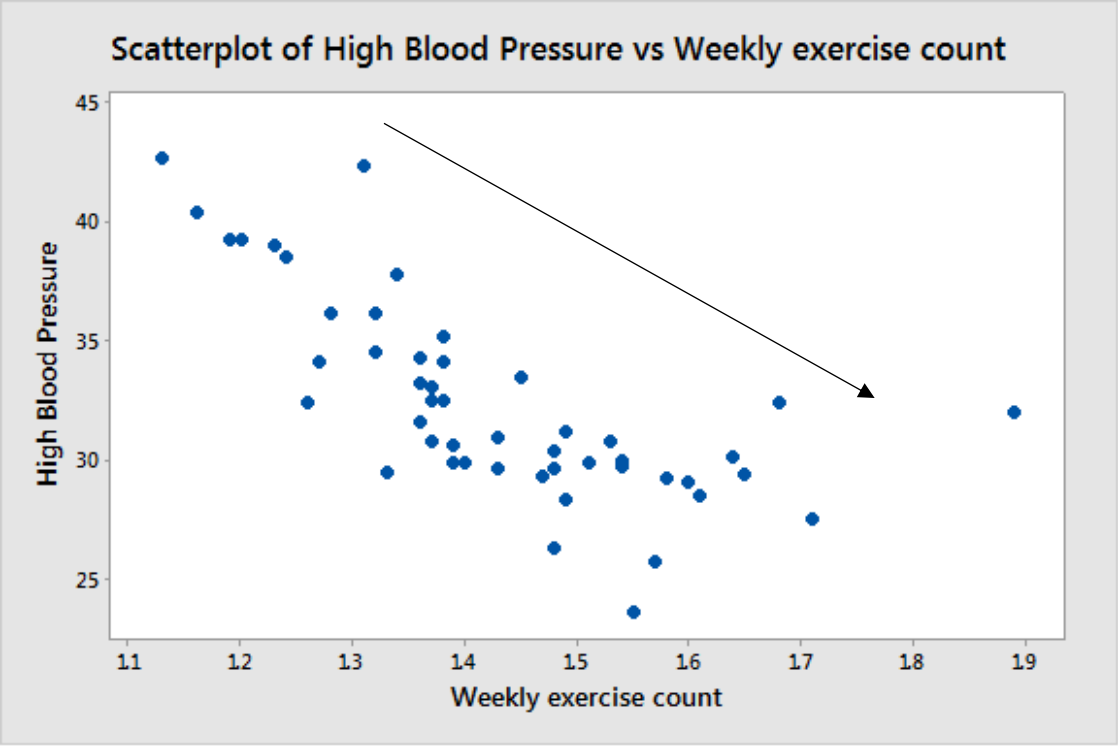


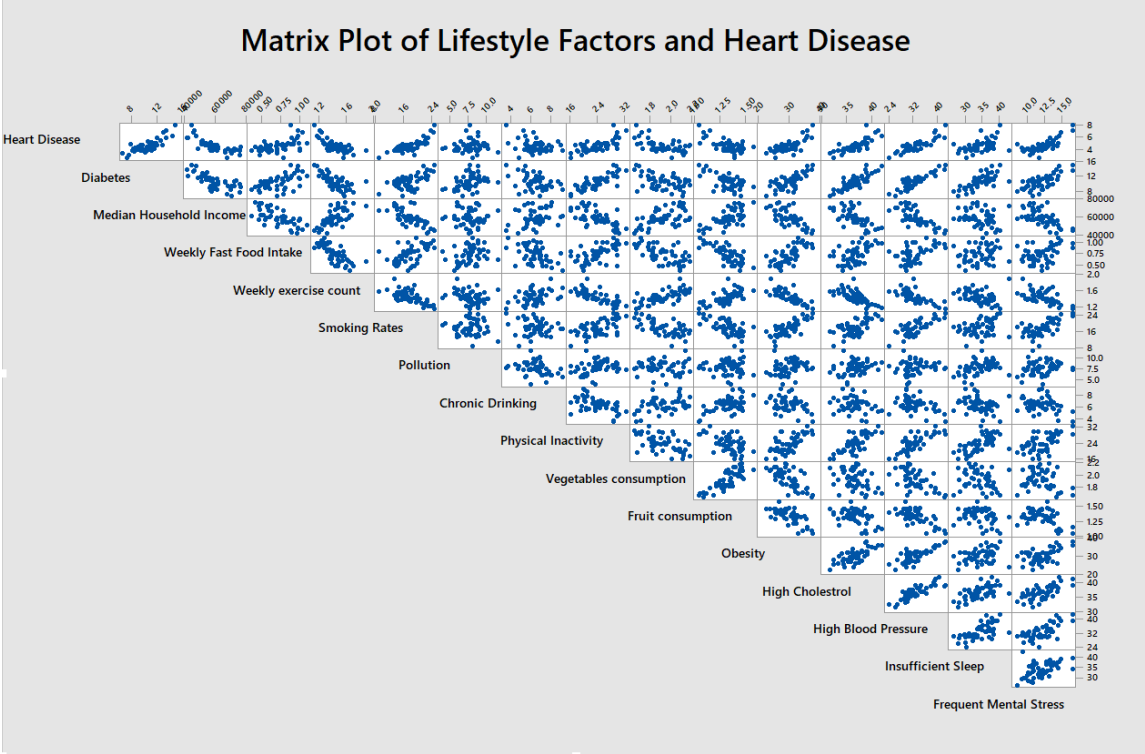
Figure 3: Scatterplot of percentage of adults who have high blood pressure versus the mean number of exercise sessions per week.



1.1.2 Matrix Plot

Figure 4 is a matrix plot that displays the correlation between potential risk factors and heart disease, as well as the correlation among the factors themselves. This can be statistically confirmed by running a Pearson's Correlation test as shown in figures 7a - 7d.

Figure 4: Matrix plot of possible risk factors and heart disease rates



1.1.3 Surface Plot

Figure 5 is a surface plot that shows the correlation between percentage of population with heart disease versus the percentage of population that are smokers and the percentage of population with high cholesterol. This plot shows that as smoking and high cholesterol rates increase, heart disease rates increase simultaneously. Figure 6 shows the correlation between percentage of population with heart disease versus the mean number of servings of vegetables consumed per day and the mean number of exercise sessions per week. This plot shows that as vegetable intake and exercise rates decrease, heart disease rates increase simultaneously.

Figure 5: Surface plot of percentage of adults who are smokers and percentage of adults who have high cholesterol versus the percentage of adults who are diagnosed with heart disease.

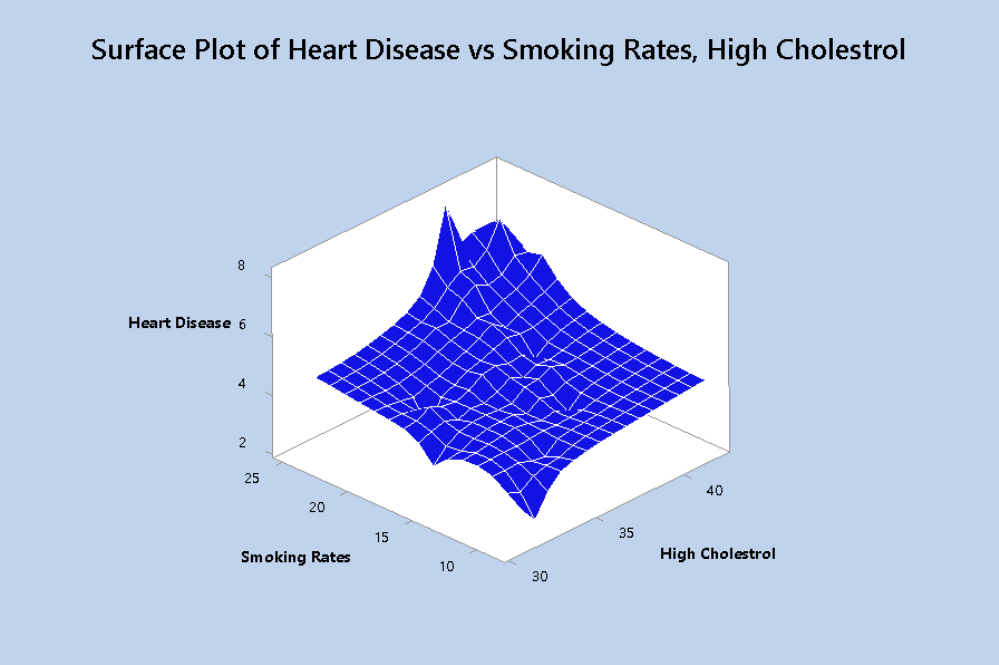
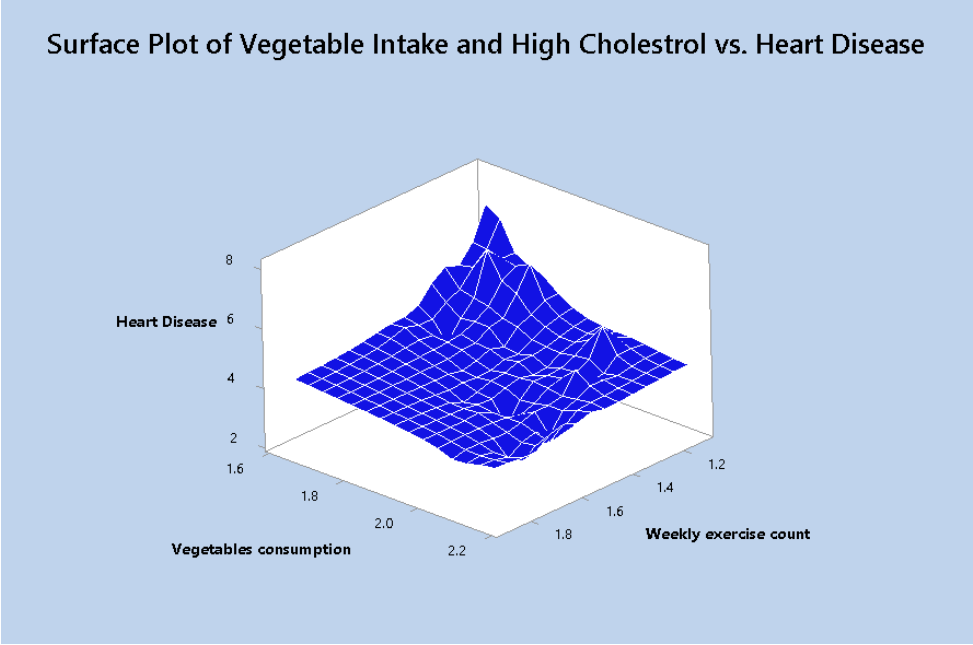


Figure 6: Surface Plot of mean number of vegetables consumed per day by adults and percentage of adults who have high cholesterol versus the percentage of adults who are diagnosed with heart disease



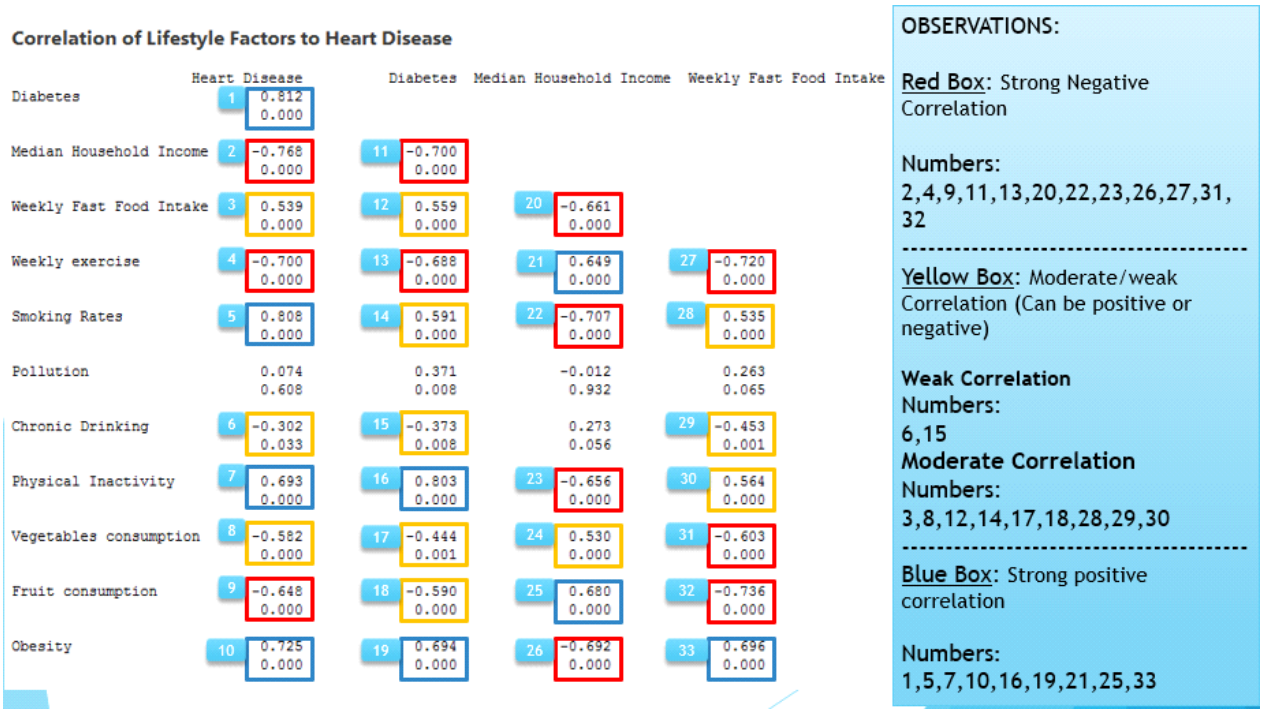
1.2 Correlation Analysis

After the graphs were generated (figures 1-6), a correlation test was performed (figures 7a-7d). Figures 7a-7d display the results from the correlation test for risk factors and rate of heart disease in United States. A correlation test is a statistical test that shows if there is a linear relationship between two variables. The correlation test provided two numerical values: Pearson's Correlation Coefficient, and a number below it called the P-value.

The Pearson's Correlation Coefficient, indicates the strength and direction of correlation. The Pearson's Correlation Coefficient is a numerical value that ranges from -1 to +1. A positive correlation coefficient indicates a positive correlation between the two variables, while a negative correlation coefficient indicates a negative correlation. If the P-value shown from the results was greater than .05, then the Pearson's Correlation Coefficient would have been ignored, because we would be unable to conclude with 95% confidence that there is a statistically significant correlation between the two variables.

The P-value, or probability value, relates to how much risk one is willing to take in during analysis – or conversely – how much confidence one desires in the statistical results. Typically, the P-value is set at 0.05, indicating only 5% risk, or 95% confidence in results. In this case, a 95% confidence level was established to determine whether there is statistical proof of correlation between two variables (note: correlation *does not* imply cause-and-effect relationship). For example, summer ice cream sales and soft drink sales may be highly correlated, but that could be due to the external factor of increasing temperatures. Furthermore, the correlation between these two variables is not due to direct cause and effect, but rather a common factor that is causing variation in both variables.

Figure 7a: Results from the correlation test



OBSERVATIONS:

Red Box: Strong Negative Correlation

Numbers:
2,4,9,11,13,20,22,23,26,27,31, 32

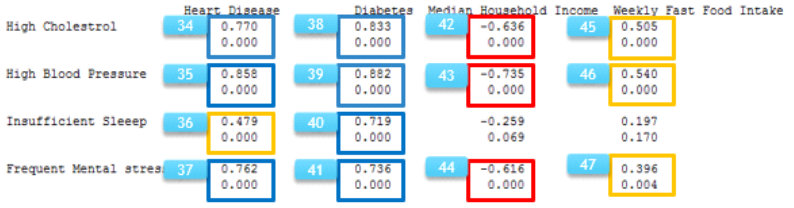
Yellow Box: Moderate/weak Correlation (Can be positive or negative)

Weak Correlation
Numbers:
6,15
Moderate Correlation
Numbers:
3,8,12,14,17,18,28,29,30

Blue Box: Strong positive correlation

Numbers:
1,5,7,10,16,19,21,25,33

Figure 7b: Results from the correlation test



OBSERVATIONS:

Red Box: Strong Negative Correlation

Numbers:
42,43,44

Yellow Box: Moderate/weak Correlation (Can be positive or negative)

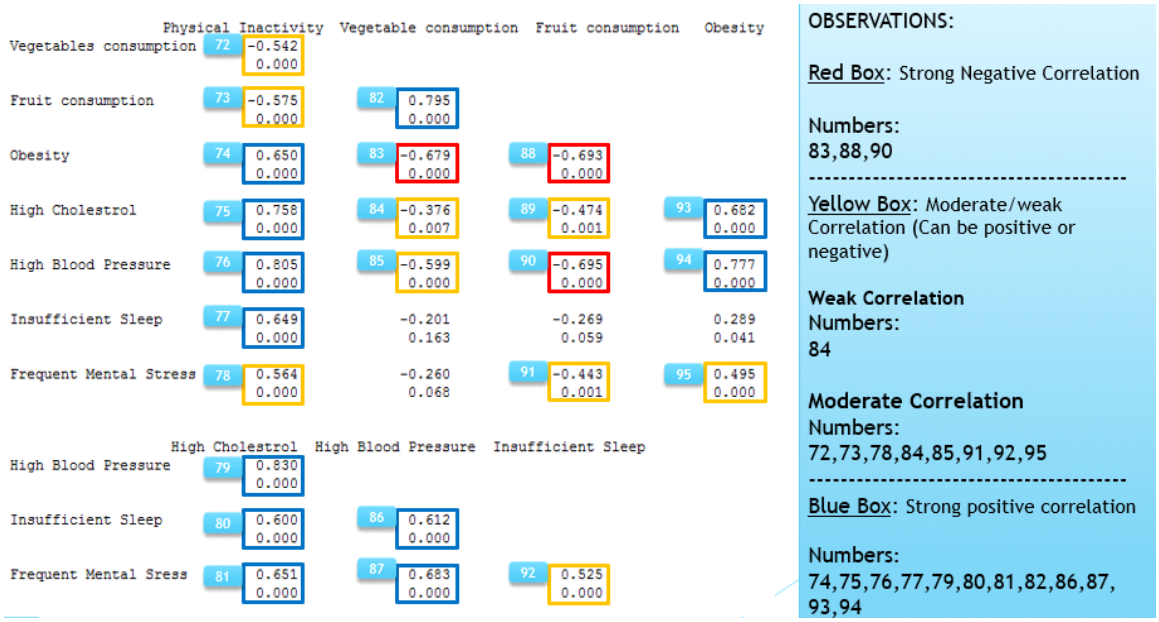
Weak Correlation:
Numbers:
47
Moderate Correlation:
Numbers:
36,45,46

Blue Box: Strong positive correlation

Numbers:
34,35,37,38,39,40,41

Figure 7c: Results from the correlation test

Figure 7d: Results from the correlation test



4.2.1 Correlation Results

The analysis results revealed a total 81 interesting correlations. The correlations between heart disease rates and the possible risk factors are displayed in the first column of figures 7a and 7b.

Likewise, the correlations among the risk factors are shown in the other columns. The blue boxes represent strong positive correlations, the yellow boxes represent moderate positive or moderate negative correlations, and the red boxes represent strong negative correlations.

In the main findings (correlations between heart disease and possible risk factors) it was observed that there is a strong positive correlation between heart disease and diabetes, smoking, physical inactivity, obesity, high cholesterol, high blood pressure, and stress. These all may be likely be very harmful in causing heart disease. A moderately positive correlation was observed between heart disease and fast-food intake, as well as insufficient sleep (these both may likely be harmful in causing heart disease). A moderately negative correlation was observed between heart disease and chronic drinking (this indicates that drinking may have a beneficial effect in preventing heart disease). A strong negative correlation was observed between heart disease and median household income, fruit and vegetable consumption. This indicates as household income increases, the rate of heart disease decreases. A plausible explanation may be that families with higher household income are able to afford a healthier lifestyle. Likewise, fruit and vegetable intake may be beneficial for the heart.

4.3 Multiple Linear Regression Analysis

The final step of this study was to run a multiple linear regression test. This is a statistical test that shows if the variation in one variable is associated with the variation in another variable. In other words, this statistical test is used to estimate how a response variable changes as the independent variables change. During multiple linear regression analysis, there were two important columns that needed to be observed: The VIF and the P-value (see figures 8a and 8b).

Figure 8a: Multiple Linear Regression – Variance Inflation Factor reduction

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.28	3.53	0.36	0.719	
Diabetes	0.223	0.119	1.86	0.071	12.31
Median Household Income	-0.000011	0.000015	-0.74	0.467	4.35
Weekly Fast Food Intake	-0.191	0.692	-0.28	0.784	4.29
Weekly exercise count	0.007	0.982	0.01	0.994	5.56
Smoking Rates	0.0879	0.0470	1.87	0.070	6.79
Pollution	-0.0603	0.0597	-1.01	0.320	2.11
Excessive Drinking	-0.0095	0.0433	-0.22	0.828	3.88
Physical Inactivity	-0.0223	0.0329	-0.68	0.503	4.72
Vegetables consumption	-1.70	1.06	-1.61	0.118	6.24
Fruit consumption	0.92	1.46	0.63	0.531	7.90
Obesity	-0.0263	0.0433	-0.61	0.547	6.11
High Cholestrol	0.0753	0.0618	1.22	0.231	4.97
High Blood Pressure	0.0208	0.0584	0.36	0.724	14.87
Insufficient Sleep	-0.0211	0.0367	-0.57	0.570	4.40
Frequent Mental Stress	0.1118	0.0728	1.54	0.134	4.46

Figure 8a: The variables were removed from highest to lowest VIF until all the VIFs were below 5.

Figure 8b: Multiple Linear Regression – Variance Inflation Factor reduction

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.17	3.05	0.71	0.481	
Median Household Income	-0.000023	0.000014	-1.71	0.095	3.62
Weekly Fast Food Intake	-0.503	0.579	-0.87	0.390	2.90
Weekly exercise count	-0.473	0.894	-0.53	0.600	4.46
Smoking Rates	0.0678	0.0362	1.87	0.069	3.89
Pollution	-0.0487	0.0534	-0.91	0.368	1.63
Excessive Drinking	-0.0094	0.0327	-0.29	0.776	2.14
Physical Inactivity	-0.0177	0.0332	-0.53	0.597	4.66
Vegetables consumption	-1.364	0.718	-1.90	0.065	2.76
High Cholestrol	0.1224	0.0535	2.29	0.028	3.61
Insufficient Sleep	0.0236	0.0307	0.77	0.447	2.97
Frequent Mental Stress	0.1505	0.0666	2.26	0.030	3.61

Figure 8b: Displays remaining variables once all the VIFs were below 5.

4.3.1 Variance Inflation Factor

The VIF (Variance Inflation Factor) is a numerical value that indicates the extent to which multicollinearity (correlation among predictors) is present in a regression analysis. Correlation between the independent variables may lead to inaccurate results because the individual effect of each predictor on the outcome (in this case, heart disease) cannot be individually established. If the VIF of a predictor is greater than 5, then the variable must be removed, and the test is re-run (the predictors are removed from highest to lowest VIF, one factor at a time, until all VIFs are below 5).

4.3.2 P-value

Traditionally, the p-values for each predictor must be below .05 for us to say with 95% confidence that there is an association present between the variables. If the p-value for a predictor is greater than .05, then the variable must be removed, and the test must be re-run (the predictors are removed from highest to lowest p-value, again one factor at a time, until all p-values are below .05). When observing the relationship between two variables, two hypotheses must be considered: The alternative and null hypothesis. If an association is present between two variables, then the alternative hypothesis is true. An alternative hypothesis is the hypothesis that a variation of one variable will imply the variation of another. On the contrary, a null hypothesis is the hypothesis that there is no significant difference between the variation of two specific variables. If there is no statistically significant association between two variables, the null hypothesis is true.

Figure 9: Results of Multiple Linear Regression

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.442586	83.62%	82.16%	77.21%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-1.58	1.74	-0.91	0.370	
Smoking Rates	0.0861	0.0297	2.90	0.006	2.75
Vegetables consumption	-1.423	0.553	-2.58	0.013	1.71
High Cholestrol	0.1360	0.0389	3.50	0.001	2.00
Frequent Mental Stress	0.1891	0.0500	3.78	0.000	2.14

All the factors were removed from highest to lowest p-value until all the p-values were below 0.05

4.3.4 Regression Results

The mathematical model (see figure 10) predicts that for every one percent increase in adults with high cholesterol, the rate of heart disease will *increase* by .136 percent. For every one percent increase in adults with frequent mental stress, the rate of heart disease will *increase* by .189 percent. For every one percent increase in frequent smokers, the rate of heart disease will *increase* by .086 percent. And finally, for one portion increase in the mean number of vegetables consumed per day by adults, the rate of heart disease will *decrease* by 1.423 percent. The statistical validity of these conclusions is extremely high because the Coefficient of Determination [R2 (adjusted)] is 82.16%, and the Standard Error (S) is only .44% (very low). These values suggest that this model explains more than 82% of the variation in rate of heart disease, and it does so with an error of less than 1%.

Mathematical Model: Heart Disease = -1.58 + 0.0861*(Smoking Rates) – 1.423*(Vegetable Consumption) + 0.1360*(High Cholesterol) + 0.1891*(Frequent Mental Stress)

2. Conclusions and Future Research

The results from the regression analysis revealed that the four most significant factors that are associated with heart disease risk are lack of vegetable intake, smoking, high cholesterol, and mental stress. This means that only a few factors are significant, although these factors can also depend on other factors which indirectly influence heart disease.

The findings from this project can be considered by state governments, physicians, and citizens themselves to reduce national and global rates of heart disease. For example, governments can launch public awareness campaigns in order to inform citizens about the detrimental and beneficial effects of different factors that affect heart disease risk. If heart disease rates decrease, then government spending may be diverted from healthcare to other priorities.This project used mathematical modeling to assess the impact of specific factors on the rates of heart disease per US state, which has never been done in past research studies.

3. Additional References

*all figures included in this paper were generated during analysis; none were taken from external sources

[1] JR Neyer et al. Prevalence of Heart Disease--- United States, 2005. Government Printing Office (GPO). Washington, DC 20402-9371. 2005

[2] CDC. Heart Disease Facts. 2017

[3] Investopedia. “What Country Spends the Most on Healthcare?”. https://www.investopedia.com/ask/answers/020915/what-country-spends-most-healthcare.asp

[4] Ginter E. Vegetarian Diets, Chronic Diseases and Longevity. Bratisl Lek Listy. 2008;109(10):463-6. Review. PMID:19166134

[5] Lap Tai Le and Joan Sabaté· PubMed-NCBI-Health Effect of Vegan Diets. Published online 2014 May 27. doi: 10.3390/nu6062131

[6] Appel LJ[1] . The Effects of Dietary Factors on Blood Pressure. Cardiol Clin. 2017 May;35(2):197-212. doi: 10.1016/j.ccl.2016.12.002.

[7] Robert A. Koeth, et al. Intestinal microbiota metabolism of *L*-carnitine, a nutrient in red meat, promotes atherosclerosis. Published online 2013 Apr 7. doi: 10.1038/nm.3145

[8] CDC. Coronary Artery Disease. 2015

[9] CDC. Heart Disease Risk Factors. 2015

[10] Am J Cardiol. Effects of Plant-Based Diets on Plasma Lipids. 2009 Oct 1;104(7):947-56. doi: 10.1016/j.amjcard.2009.05.032.

[12] Mayo Clinic. Coronary Artery Disease. 2018

[13] NIH, Coronary Heart Disease Risk Factors. https://www.nhlbi.nih.gov/health-topics/coronary-heart-disease-risk-factors