Factors influencing heart attack likelihood among youngsters in India

Scarongella Benedicta, Troiano Antonio *January 2025*

1) Introduction

In recent years, India has witnessed a troubling increase in heart attacks among younger individuals. Traditionally associated with men in their late 50s, heart attacks have now become more prevalent among those in their 20s to 40s. A heart attack occurs when blood flow to a part of the heart is blocked, causing damage to the heart muscle. This is often due to the buildup of fatty deposits in the arteries, and if not promptly treated, it can lead to severe consequences or even death. While heart attacks were once considered primarily a concern for older adults, the post-pandemic era has revealed alarming trends: younger generations in India are becoming more susceptible to cardiovascular diseases. Medical experts have expressed concerns about this shift, emphasizing the urgency of understanding its underlying causes. Our research focuses on identifying the actual factors contributing to this concerning increase. We aim to determine whether modifications in lifestyle and behavior could mitigate these risks.

2) Dataset

The dataset we selected was published on “Kaggle" under an MIT license and recorded data among the years 2019-2023. It contains data on factors influencing heart attack likelihood among young individuals in India (ages 18-35). It captures demographic, lifestyle, medical, and clinical test information. We removed the following columns from the given dataset, as we deemed these factors less relevant to our study: “Exercise-induced Angina”, “Region”, and “Blood Pressure”. Below there is a schematic description of the factors considered in the dataset. Note that some column names have been modified by us and are provided in parentheses:

Target Variable:

- Heart Attack Likelihood (HeartAttack): A binary outcome variable indicating whether or not the individual is at risk of a heart attack (Yes/No).

Demographics:

- Age: Age of the individual, ranging from 18 to 35 years.
- Gender: The sex of the individual (Male, Female, Other)
- Urban/Rural (UR): Whether the individual resides in an urban or rural area.
- Socio-economic Status (SES): Classification of the individual’s socioeconomic status (Low, Middle, High).

Lifestyle Factors:

- Smoking Consumption (Smoking): Frequency of smoking consumption (Never, Occasionally, Regularly).
- Alcohol Consumption (Alcohol): Frequency of alcohol consumption (Never, Occasionally, Regularly).
- Dietary Preferences (Diet): The type of diet followed by the individual (Vegetarian, Non-Vegetarian, Vegan).
- Physical Activity (PhysicalA): The level of physical activity of the individual (Sedentary, Moderate, High).
- Screen Time(Screen): The number of hours spent on screens per day.
- Sleep Duration(Sleep): The average number of hours the individual sleeps per night.

Medical History:

- Family History of Heart Disease(FamilyHistory): If the individual has a family history of heart disease (Yes/No).
- Diabetes: Whether the individual suffers from diabetes (Yes/No).
- Hypertension: Whether the individual suffers from hypertension (Yes/No).
- Cholesterol Levels (Cholesterol): The total cholesterol levels in mg/dL.
- Body Mass Index (BMI): A measure of the individual's body fat based on weight and height.
- Stress Levels (Stress) : Categorization of stress levels (Low, Medium, High).

Clinical and Test Results:

- Resting Heart Rate (RHR):The individual’s heart rate when at rest, measured in beats per minute.

- ECG: Whether the individual's electrocardiogram results were normal or abnormal.
- Chest Pain Type: Classification of chest pain (Typical, Atypical, etc.).
- Maximum Heart Rate Achieved (MHRA): The maximum heart rate recorded during stress or exercise.
- Blood Oxygen Levels (BloodOxygen): The percentage of oxygen saturation in the blood.
- Triglyceride Levels (Triglyceride): The amount of fat present in the blood, measured in mg/dL.

3) Exploring the dataset

To ensure data quality and integrity, we began by conducting a preliminary check in R to confirm the absence of missing values in the dataset. This verification revealed that no missing entries were present, providing a robust foundation for further analysis. Next, we examined the summary statistics of our dataset. This step highlighted that all variables were of comparable magnitude and fell within a relatively small range, minimizing concerns about scale or potential outliers distorting the analysis. For univariate analysis, we explored each variable individually: for continuous variables, we employed histograms to estimate their distribution and compared them to a standard normal, while for categorical variables we used bar plots to visualize the frequency of each category, providing insights into the relative prevalence of different groups within the dataset. We then shifted our focus to bivariate analysis, examining the relationship between each factor and heart attack risk: for categorical variables, bar plots were used to evaluate the proportion of individuals with and without heart attacks across different categories, whereas for continuous variables, box plots were employed to compare the distribution of these variables between individuals affected and unaffected by heart attacks. [3] Surprisingly, no single variable exhibited a dominant trend or clear association with heart attack risk in isolation. This absence of distinct patterns suggests a potential interaction among multiple factors, highlighting the need for a more complex modeling approach to uncover the relationships driving heart attack risk.

4) Logistic Regression

Since our objective was to predict the likelihood of a heart attack, which is a categorical variable (it can assume values "Yes" or "No"), we employed a logistic regression framework. Logistic regression is well-suited for this classification because it models the log-odds of the probability of an event as a linear combination of predictors. To evaluate the individual contribution of different covariates, we first conducted Likelihood Ratio Tests (LRTs) on each of them separately. The LRT compares two nested models: one that includes the predictor of interest and one that excludes it. Formally, it tests the null hypothesis $\mathbf{H_0: \beta = 0}$ (the coefficient of our interest predictor is zero, hence it has no effect) against the alternative one $\mathbf{H_1: \beta \neq 0.}$ If the p-value is sufficiently low (< 0.05), we reject H_0 and conclude that the predictor significantly influences the likelihood of a heart attack risk.

Consistent with our initial exploratory plots, most covariates did not yield statistically significant p-values when tested individually, suggesting that none of these factors alone adequately explains heart attack likelihood. Only the variables ***"Stress"*** and ***"UR"*** showed sufficient evidence to reject H_0 in isolation, but we suspected these single predictors could not fully account for a phenomenon as complex as heart attack risk. We constructed a comprehensive model incorporating most of the predictors and performed a multivariable logistic regression to evaluate their collective impact, but once again only very few variables emerged to be significant. Based on these observations, we hypothesized that interactions between multiple variables might be more impactful than standalone effects. Since considering the interactions among all variables would have produced an extremely complex model, we assessed on our own several combinations, and then selected those we deemed most interrelated for inclusion. Specifically, we examined the interactions between:

- ***Gender* × *Diet* × *Physical Activity***
- ***Hypertension* × *Age***
- ***Hypertension* × *Diabetes***
- ***Hypertension* × *Diabetes* × *Chest Pain***
- ***Age* × *Gender* × *Diet***

Indeed, from the final logistic regression [1], we observe that several main effects and interaction terms achieve conventional significance ($p < 0.05$), namely:

- *PhysicalAModerate*
- *GenderMale:PhysicalAModerate*
- *DietVegetarian:PhysicalASedentary*
- *DiabetesYes:ChestPainAtypical*
- *DiabetesYes:ChestPainTypical*
- *GenderMale:DietVegetarian:PhysicalAModerate*
- *GenderMale:DietVegetarian:PhysicalASedentary*
- *HypertensionYes:DiabetesYes:ChestPainTypical*

These findings confirm that heart attack likelihood is strongly influenced by **combinations** of covariates, rather than by any single factor in isolation.

5) LASSO

Building on the interactions discovered in our initial logistic regression model, we recognized that the model, despite yielding significant predictors, remained too large and thus susceptible to overfitting, a condition in which the model fits the current dataset well but struggles to generalize to unseen data. To address this, we introduced LASSO (Least Absolute Shrinkage and Selection Operator), an L1-penalized regression technique that inherently selects relevant variables by shrinking the coefficients of less influential predictors toward zero. Technically, LASSO adds a penalty term to the standard logistic regression objective function, where λ controls the amount of shrinkage.

Coefficients considered not important for the prediction are driven all the way to zero, removing those predictors from the model. This automatic feature selection helps to prevent multicollinearity, reduce variance, and produce a simpler but more robust model. Even if there exists an optimal λ from a theoretical perspective, in practice, we relied on a 10-fold cross-validation to find the optimal value (the penalty strength) that balances model complexity and predictive accuracy. After fitting the LASSO model at the best λ, we inspected the non-zero coefficients to identify those variables and interactions contributing most strongly to heart attack likelihood. The remaining terms include [2]:

- **GenderMale**
- **GenderFemale:DietNon-Vegetarian:PhysicalAHigh**
- **GenderMale:DietVegetarian:PhysicalAHigh**
- **GenderFemale:DietVegetarian:PhysicalASedentary**
- **HypertensionYes:DiabetesYes:ChestPainAtypical**
- **HypertensionNo:DiabetesYes:ChestPainTypical**

We then refit a reduced logistic model containing only these non-zero predictors, effectively yielding a parsimonious model that preserves the most informative main effects and interactions. This final step ensures that our model is less prone to overfitting and remains interpretable, focusing on the predictors that genuinely matter.

6) AIC, BIC, LikeLihood Ratio Test

After deriving our final (reduced) model from LASSO, we evaluated its performance against the initial (full) model using AIC, BIC, and a Likelihood Ratio Test (LRT).

The full model yielded an AIC of 10162.54 and a BIC of 10530.27, whereas the reduced model produced lower values (AIC = 10101.58, BIC = 10152.05), suggesting a better balance between model fit and complexity. We then conducted an LRT to compare the deviance of the two nested models, which yielded a high p-value (around 0.9793), indicating that the improvement in fit by the full model's additional parameters was not statistically significant [4]. Taken together, these results justify preferring the simpler final model, as it adequately explains the data without overfitting or unnecessary complexity.

7) Conclusions

The dataset we chose reflects the complex and multifaceted nature of cardiovascular health, particularly among young individuals in India. Our findings highlight that heart attack risk cannot be adequately explained by isolated variables, as it emerges from interactions among lifestyle,

demographic, and medical factors. Through our logistic regression and LASSO reduction, we discovered combinations of factors that significantly contribute to heart attack likelihood, either amplifying or reducing this risk.

Notably, some coefficients in our analysis were negative, indicating protective factors that decrease the likelihood of heart attacks. These findings suggest that specific healthy lifestyle changes, such as engaging in moderate physical activity or adopting certain dietary habits, could play a vital role in mitigating cardiovascular risk. These insights emphasize the importance of promoting positive lifestyle interventions as part of public health initiatives.

Nonetheless, our study has certain limitations that warrant consideration. The findings are based on a dataset reflecting young individuals in India, and may not generalize to older populations or those in different regions. Potential biases in self-reported variables, as well as the simplifications inherent in LASSO modeling may have also impacted the results. Additionally, the cross-sectional nature of the data limits causal interpretations; future research incorporating longitudinal data, a more diverse sample, and additional variables could provide a more comprehensive understanding of cardiovascular risk factors. In conclusion, our study underlines the critical need for a holistic approach in addressing cardiovascular health. By identifying and addressing high-risk combinations of factors, public health initiatives and clinical interventions can be better tailored to reduce the alarming trend of heart attacks in younger populations in India.

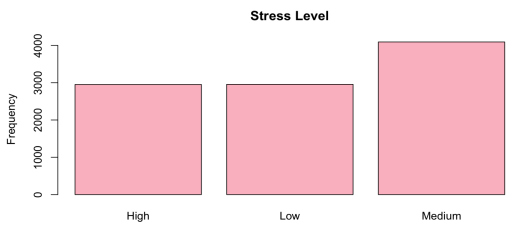
1)

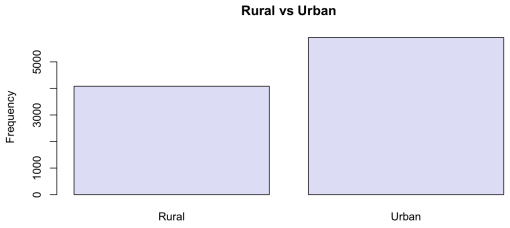
```
Call:
glm(formula = HeartAttack2 ~ Gender + Diet + PhysicalA + Age +
    Hypertension + Diabetes + ChestPain + Gender * Diet * PhysicalA +
    Hypertension * Age + Hypertension * Diabetes + Hypertension *
    Diabetes * ChestPain + Age * Gender * Diet, family = binomial,
    data = data_set)

Coefficients: (1 not defined because of singularities)
                                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                                    -1.816e+00  3.367e-01  -5.392 6.96e-08 ***
GenderMale                                      6.760e-01  4.419e-01   1.530  0.12610
GenderOther                                     1.292e+00  1.737e+00   0.744  0.45716
DietVegan                                       8.057e-01  7.766e-01   1.037  0.29952
DietVegetarian                                  6.289e-01  4.685e-01   1.342  0.17951
PhysicalAModerate                               4.071e-01  1.972e-01   2.064  0.03898 *
PhysicalASedentary                              3.751e-01  1.945e-01   1.929  0.05374 .
Age                                             2.811e-03  1.041e-02   0.270  0.78713
HypertensionYes                                 2.172e-01  3.252e-01   0.668  0.50434
DiabetesYes                                    -2.365e-01  1.549e-01  -1.526  0.12695
ChestPainAtypical                              -1.056e-01  9.115e-02  -1.159  0.24654
ChestPainNon-anginal                           -8.397e-05  8.972e-02  -0.001  0.99925
ChestPainTypical                               -1.003e-01  9.022e-02  -1.112  0.26614
GenderMale:DietVegan                           -1.080e+00  1.046e+00  -1.032  0.30216
GenderOther:DietVegan                          -5.307e+00  2.995e+00  -1.772  0.07642 .
GenderMale:DietVegetarian                      -9.544e-01  6.554e-01  -1.456  0.14533
GenderOther:DietVegetarian                     -2.612e+00  2.536e+00  -1.030  0.30297
GenderMale:PhysicalAModerate                   -5.280e-01  2.674e-01  -1.975  0.04827 *
GenderOther:PhysicalAModerate                  -4.929e-02  1.162e+00  -0.042  0.96618
GenderMale:PhysicalASedentary                  -4.550e-01  2.620e-01  -1.736  0.08249 .
GenderOther:PhysicalASedentary                  5.731e-01  1.134e+00   0.505  0.61324
DietVegan:PhysicalAModerate                    -3.578e-01  4.772e-01  -0.750  0.45338
DietVegetarian:PhysicalAModerate               -3.417e-01  2.710e-01  -1.261  0.20733
DietVegan:PhysicalASedentary                   -2.213e-01  4.663e-01  -0.475  0.63509
DietVegetarian:PhysicalASedentary              -5.597e-01  2.685e-01  -2.085  0.03710 *
Age:HypertensionYes                            -1.358e-02  1.131e-02  -1.200  0.23019
HypertensionYes:DiabetesYes                     4.158e-01  2.881e-01   1.443  0.14889
HypertensionYes:ChestPainAtypical               1.233e-01  1.876e-01   0.657  0.51123
HypertensionYes:ChestPainNon-anginal            1.240e-02  1.838e-01   0.067  0.94622
HypertensionYes:ChestPainTypical                3.325e-01  1.812e-01   1.835  0.06644 .
DietVegetarian:PhysicalASedentary              -5.597e-01  2.685e-01  -2.085  0.03710 *
Age:HypertensionYes                            -1.358e-02  1.131e-02  -1.200  0.23019
HypertensionYes:DiabetesYes                     4.158e-01  2.881e-01   1.443  0.14889
HypertensionYes:ChestPainAtypical               1.233e-01  1.876e-01   0.657  0.51123
HypertensionYes:ChestPainNon-anginal            1.240e-02  1.838e-01   0.067  0.94622
HypertensionYes:ChestPainTypical                3.325e-01  1.812e-01   1.835  0.06644 .
DiabetesYes:ChestPainAtypical                   4.771e-01  2.071e-01   2.304  0.02124 *
DiabetesYes:ChestPainNon-anginal                1.345e-01  2.114e-01   0.636  0.52453
DiabetesYes:ChestPainTypical                    4.811e-01  2.064e-01   2.331  0.01976 *
GenderMale:Age                                 -3.436e-03  1.386e-02  -0.248  0.80422
GenderOther:Age                                -5.990e-02  5.271e-02  -1.136  0.25583
DietVegan:Age                                  -1.763e-02  2.439e-02  -0.723  0.46975
DietVegetarian:Age                             -7.424e-03  1.499e-02  -0.495  0.62049
GenderMale:DietVegan:PhysicalAModerate          4.074e-01  6.127e-01   0.665  0.50605
GenderOther:DietVegan:PhysicalAModerate         9.378e-01  1.187e+00   0.790  0.42940
GenderMale:DietVegetarian:PhysicalAModerate     9.266e-01  3.903e-01   2.374  0.01760 *
GenderOther:DietVegetarian:PhysicalAModerate    5.919e-01  1.630e+00   0.363  0.71643
GenderMale:DietVegan:PhysicalASedentary        -7.339e-02  6.055e-01  -0.121  0.90353
GenderOther:DietVegan:PhysicalASedentary               NA         NA      NA       NA
GenderMale:DietVegetarian:PhysicalASedentary    1.116e+00  3.846e-01   2.902  0.00371 **
GenderOther:DietVegetarian:PhysicalASedentary   4.194e-01  1.605e+00   0.261  0.79377
HypertensionYes:DiabetesYes:ChestPainAtypical  -2.153e-01  3.990e-01  -0.540  0.58950
HypertensionYes:DiabetesYes:ChestPainNon-anginal -5.531e-01  4.235e-01  -1.306  0.19148
HypertensionYes:DiabetesYes:ChestPainTypical   -9.631e-01  4.110e-01  -2.343  0.01912 *
GenderMale:DietVegan:Age                        3.143e-02  3.336e-02   0.942  0.34620
GenderOther:DietVegan:Age                       1.847e-01  1.017e-01   1.816  0.06940 .
GenderMale:DietVegetarian:Age                  -9.769e-04  2.074e-02  -0.047  0.96244
GenderOther:DietVegetarian:Age                  7.638e-02  7.591e-02   1.006  0.31427
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10113  on 9999  degrees of freedom
Residual deviance: 10061  on 9949  degrees of freedom
AIC: 10163

Number of Fisher Scoring iterations: 4
```

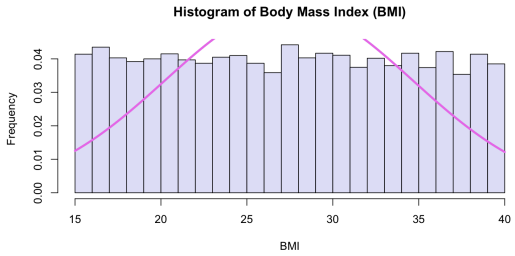
2)

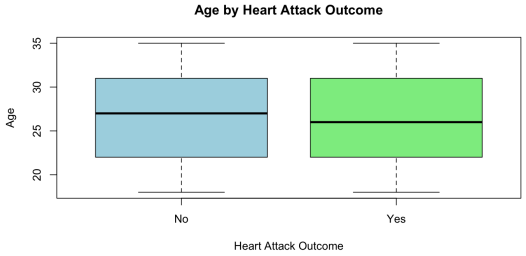
```
> coef(lasso_model)
63 x 1 sparse Matrix of class "dgCMatrix"
                                                        s0
(Intercept)                                    -1.36981107
GenderMale                                      0.03172172
GenderOther                                     .
DietVegan                                       .
DietVegetarian                                  .
PhysicalAModerate                               .
PhysicalASedentary                              .
Age                                             .
HypertensionYes                                 .
DiabetesYes                                     .
ChestPainAtypical                               .
ChestPainNon-anginal                            .
ChestPainTypical                                .
Age:HypertensionYes                             .
HypertensionYes:DiabetesYes                     .
GenderFemale:DietNon-Vegetarian:PhysicalAHigh  -0.13772904
GenderMale:DietNon-Vegetarian:PhysicalAHigh     .
GenderOther:DietNon-Vegetarian:PhysicalAHigh    .
GenderFemale:DietVegan:PhysicalAHigh            .
GenderMale:DietVegan:PhysicalAHigh              .
GenderOther:DietVegan:PhysicalAHigh             .
GenderFemale:DietVegetarian:PhysicalAHigh       .
GenderMale:DietVegetarian:PhysicalAHigh        -0.13170891
GenderOther:DietVegetarian:PhysicalAHigh        .
GenderFemale:DietNon-Vegetarian:PhysicalAModerate  .
GenderMale:DietNon-Vegetarian:PhysicalAModerate    .
GenderOther:DietNon-Vegetarian:PhysicalAModerate   .
GenderFemale:DietVegan:PhysicalAModerate        .
GenderMale:DietVegan:PhysicalAModerate          .
GenderOther:DietVegan:PhysicalAModerate         .
GenderFemale:DietVegetarian:PhysicalAModerate   .
GenderMale:DietVegetarian:PhysicalAModerate     .
GenderOther:DietVegetarian:PhysicalAModerate    .
GenderFemale:DietNon-Vegetarian:PhysicalASedentary .
GenderMale:DietNon-Vegetarian:PhysicalASedentary   .
GenderOther:DietNon-Vegetarian:PhysicalASedentary  .
GenderFemale:DietVegan:PhysicalASedentary       .
GenderMale:DietVegan:PhysicalASedentary         .
GenderOther:DietVegan:PhysicalASedentary        .
GenderFemale:DietVegetarian:PhysicalASedentary -0.04436716
GenderMale:DietVegetarian:PhysicalASedentary    .
GenderOther:DietVegetarian:PhysicalASedentary   .
HypertensionNo:DiabetesNo:ChestPainAtypical     .
HypertensionYes:DiabetesNo:ChestPainAtypical    .
HypertensionNo:DiabetesYes:ChestPainAtypical    .
HypertensionYes:DiabetesYes:ChestPainAtypical   0.06095277
HypertensionNo:DiabetesNo:ChestPainNon-anginal  .
HypertensionYes:DiabetesNo:ChestPainNon-anginal .
HypertensionNo:DiabetesYes:ChestPainNon-anginal .
HypertensionYes:DiabetesYes:ChestPainNon-anginal .
HypertensionNo:DiabetesNo:ChestPainTypical      .
HypertensionYes:DiabetesNo:ChestPainTypical     .
HypertensionNo:DiabetesYes:ChestPainTypical     0.01758345
HypertensionYes:DiabetesYes:ChestPainTypical    .
GenderFemale:DietNon-Vegetarian:Age             .
GenderMale:DietNon-Vegetarian:Age               .
GenderOther:DietNon-Vegetarian:Age              .
GenderFemale:DietVegan:Age                      .
GenderMale:DietVegan:Age                        .
GenderOther:DietVegan:Age                       .
GenderFemale:DietVegetarian:Age                 .
GenderMale:DietVegetarian:Age                   .
GenderOther:DietVegetarian:Age                  .
```

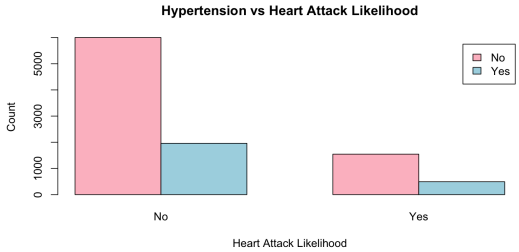
3)

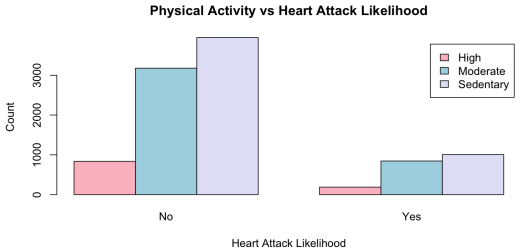


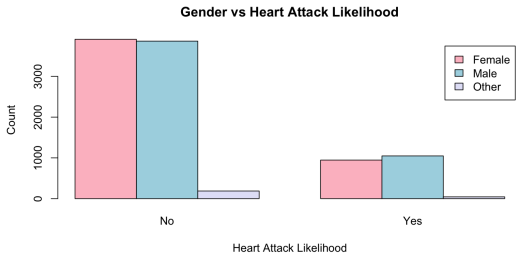


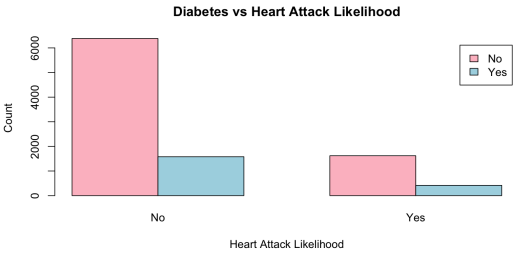




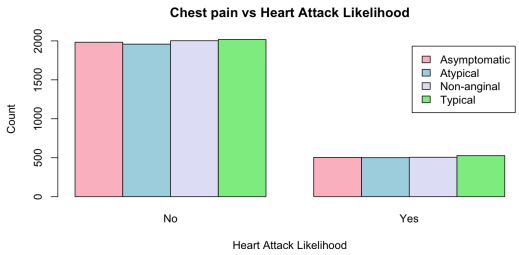








4)




```
Likelihood ratio test

Model 1: HeartAttack2 ~ GenderMale +
GenderFemale.DietNon.Vegetarian.PhysicalAHigh +
    GenderMale.DietVegetarian.PhysicalAHigh +
GenderFemale.DietVegetarian.PhysicalASedentary +
    HypertensionYes.DiabetesYes.ChestPainAtypical +
HypertensionNo.DiabetesYes.ChestPainTypical
Model 2: HeartAttack2 ~ Gender + Diet + PhysicalA + Age + Hypertension +
    Diabetes + ChestPain + Gender:Diet:PhysicalA + Hypertension:Age +
    Hypertension:Diabetes + Hypertension:Diabetes:ChestPain +
    Age:Gender:Diet
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   7 -5043.8
2  51 -5030.3 44 27.032     0.9793
```

8) Bibliography

- https://www.kaggle.com/datasets/ankushpanday1/heart-attack-in-youth-of-india
- https://even.in/blog/young-and-heartbroken-indians-experiencing-heart-attacks-10-years-earlier-than-the-west/#:~:text=In%20recent%20years%2C%20India%20has,40%2D69%20year%20age%20group
- https://www.geetanjalihospital.co.in/blogs/view/heart-attack-in-younger-age-how-to-avoid#:~:text=Medical%20news%20statistics%20show%20around,attack%20at%20a%20younger%20age