Class: IST 687

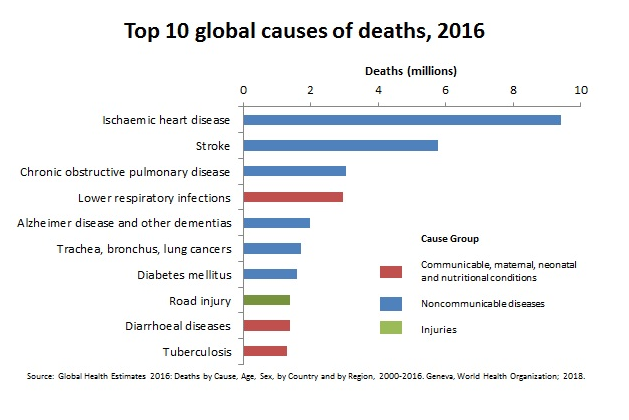
Kaye Burnet, Tiffanie Mac Donald, Andrew Lemon

Final Project

09/23/2019

**Background & Scope**

[According to the World Health Organization](https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death), cardiovascular disease is the number one cause of death worldwide. Given its significance, it is important to understand the risk factors for heart health problems to help develop creative and effective strategies to combat cardiovascular disease. The purpose of this project is to identify trends in heart health data to predict cardiovascular health status and the likelihood of certain cardiovascular events.



**Data Set**

The [Heart Disease UCI dataset](https://archive.ics.uci.edu/ml/datasets/Heart+Disease) contains 76 attributes and 303 usable observations. It includes data from patients at Hungarian Institute of Cardiology in Budapest; University Hospital in Zurich, Switzerland; University Hospital in Basel, Switzerland; V.A. Medical Center in Long Beach; and Cleveland Clinic Foundation in Cleveland. The data was collected over several years in the 1980s.

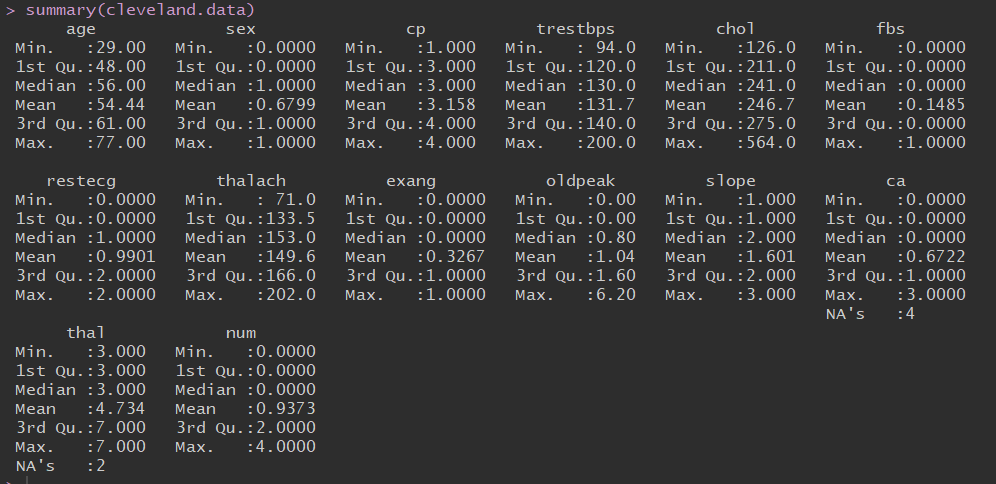
Of the data that was originally collected for the study, only 303 out of 457 records are usable for analysis, either due to missing data or record corruption. Of the 303 usable records, some contain missing data in one or more columns, but due to the limited size of the data set, this analysis will endeavor to include as many records as possible. This could mean substituting average values for missing data.

Of the 76 variables originally collected for each record, our analysis will only include 14. This is partially due to corrupted files, and partially due to relevance. Since our aim is to study heart health, we only included variables relevant to that topic.

We will be using a pre-cleaned data set provided by the original researchers, instead of the corrupted files. The researchers separated the processed data by location where the data was collected: Cleveland, Hungary, Switzerland, & the V.A. hospital in Long Beach.

The Heart Disease UCI dataset was chosen because it is freely available, robust, and time-tested. It is popular with students of data science due to its completeness and quality, which will allow us to reference other research in our analysis for comparison and insight.

**Initial Evaluation of Data**



The process began by reading in the data using read.csv(). To improve our understanding of the data, the columns were labeled, "age" in years, "sex" is gender, "cp" is chest pain type, "trestbps" is resting blood pressure, "chol" is serum cholesterol in mg/dl, "fbs" is fasting blood sugar, "restecg" is resting electrocardiographic results, "thalach" is maximum hear rate acheived,"exang" is exercise induced angina (1 = yes, 0 = no), "oldpeak" is depression induced by exercise relative to rest, "slope" is the slope of the peak exercise segment, "ca" is the number of major vessels colored by flourosopy, "thal" is a categorical variable that evaluates the presence of Thalassemia, a blood defect where 3 = normal; 6 = fixed defect; 7 = reversable defect. "Num" is our predicted value or target signified by a 1 being the patient does have heart disease and a 0 meaning the patient does NOT have heart disease.

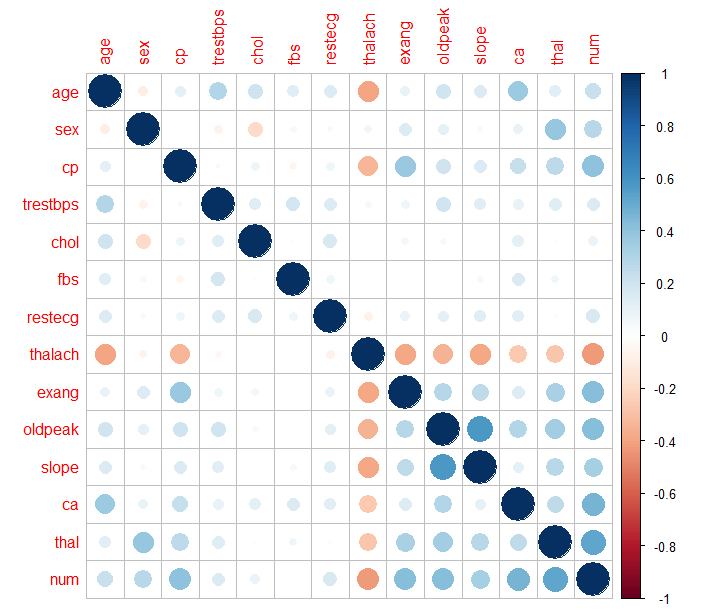
The cleveland dataset was chosen to eliminate the challenges that presented themselves when all the datasets were combined.

**Data Wrangling**

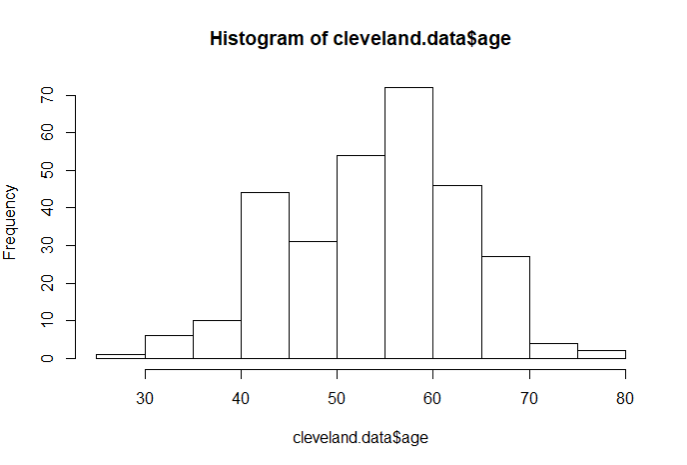
There were six null values present in the data. Since this was an insignificant portion of the data, the null values were removed to make analysis more meaningful.

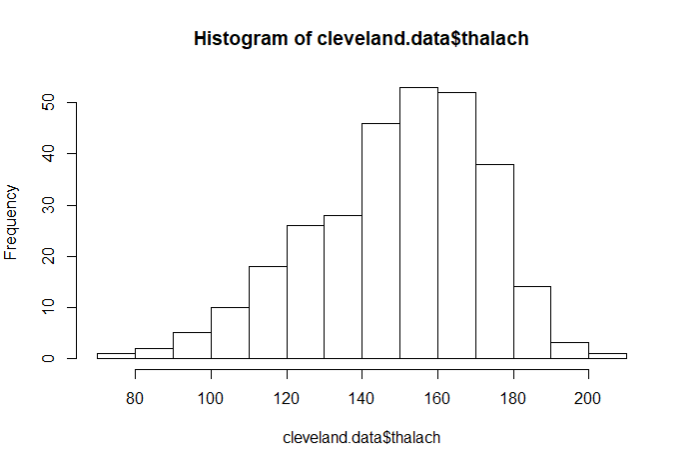
**Determining Important Variables with Correlation Matrix**

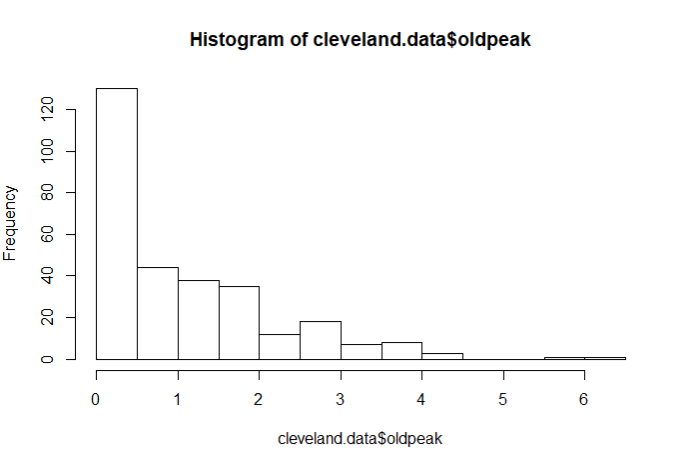
According to the correlation matrix, trestbps, chol, fbs and restecg are the least correlated variables with num, the variable we would like to predict. Therefore, we will drop these from our analysis.

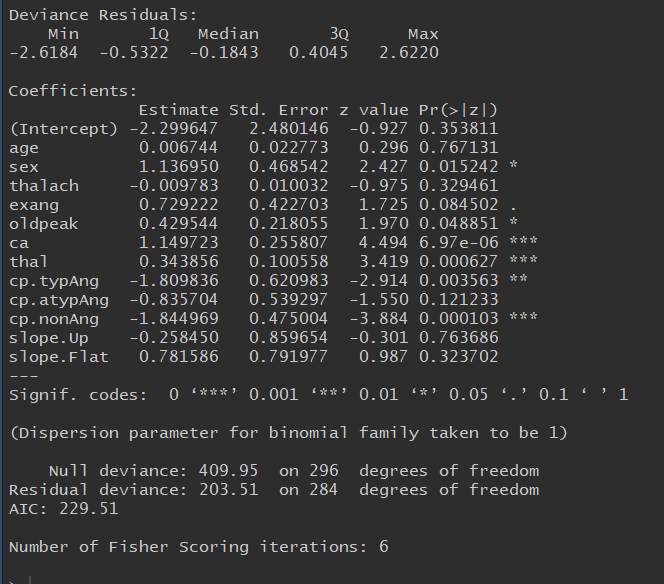


**Descriptive Look at Variables**

We created histograms of each categorical x-variable. As shown below, Age was normally distributed.

Thalach was slightly left-skewed, but approximately normally distributed. Oldpeak was significantly right-skewed.  


**Dummy Variables**  
  
 Two of our x-variables, CP and Slope, were categorical variables with more than two values. To include them in the analysis, we encoded them as dummy variables. In keeping with statistical best practices, we created *k-1* dummy variables for *k* initial variable categories.  


**Prediction Using Multiple Logistic Linear Regression**  
 

When interrupting the data above, one can see some surprising results. Interesting enough, the age of individuals in this study had no significant impact on their likelihood of heart disease. As one might expect, typical anginal pain and non-anginal pain are negatively associated with the presence of heart disease. The number of vessels colored by fluoroscopy (ca) has a strong positive association with the presence of heart disease, while being male has a weak but positive correlation. The presence of Thalassemia, a blood defect, had a strong positive association with the presence of heart disease.  
  
  
**Conclusions**

Through our analysis, we were able to gain further insight into the causes and risk factors for heart disease, as well as eliminate some potential risk factors that were determined to be insignificant. This kind of analysis can help the medical community better treat heart disease and focus future research.  
  
**R Code**

#-------------------DATA INGESTION AND SUMMARY STATISTICS-------------------------------

cleveland.data <- read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data",header=FALSE,sep=",",na.strings = '?')

names(cleveland.data) <- c( "age", "sex", "cp", "trestbps", "chol","fbs", "restecg",

"thalach","exang", "oldpeak","slope", "ca", "thal", "num")

summary(cleveland.data)

str(cleveland.data)

#-----------------------------LOAD PACKAGES---------------------------------

install.packages("corrplot")

install.packages("tidyverse")

install.packages("ggcorrplot")

install.packages("caret")

library('corrplot')

library('tidyverse')

library('ggcorrplot')

library('ggplot2')

library('plyr')

library('caret')

#-----------------------------DATA CLEAN UP---------------------------------

#Any value for "num" greater than zero indicates the presence of heart disease.

cleveland.data$num[cleveland.data$num>0]<-1

cleveland.data$num

head(cleveland.data)

sum(is.na(cleveland.data))

cleveland.data <- na.omit(cleveland.data)

#str(heart.data)

#-----------------------------CORRELATION ANALYSIS--------------------------

# Displaying the coralation matrix

corr <- cor(cleveland.data)

# Visualize the correlation matrix

corrplot(corr)

corr<-cor(cleveland.data)

head(corr)

head(cleveland.data)

names(cleveland.data)

cleveland.data<-cleveland.data[]

#According to the Correlation matrix, trestbps, chol, fbs and restecg are the least correlated variables with

#num, the variable we would like to predict. Therefore, we will drop these from our analysis.

#----------------------------CHECK REGRESSION ASSUMPTIONS----------------------

#First, we check to make sure the x-variables are normally distributed

hist(cleveland.data$age) #normally distributed

hist(cleveland.data$sex) #1 for male, 0 for female

hist(cleveland.data$cp) #1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic

hist(cleveland.data$thalach) #Slightly left-skewed (negatively skewed) but approximately normal

hist(cleveland.data$exang) #Exercise induced angina (1 = yes; 0 = no)

hist(cleveland.data$oldpeak) # Heavily right-skewed. ST depression induced by exercise relative to rest

hist(cleveland.data$slope) # The slope of the peak exercise ST segment (1: upsloping,2: flat,3: downsloping)

#---------------------------CREATE DUMMY VARIABLES---------------------------

#Create dummy variables for CP:

cleveland.data$cp.typAng<-ifelse(cleveland.data$cp==1,1,0)

cleveland.data$cp.atypAng<-ifelse(cleveland.data$cp==2,1,0)

cleveland.data$cp.nonAng<-ifelse(cleveland.data$cp==3,1,0)

cleveland.data$cp.asymp<-ifelse(cleveland.data$cp==4,1,0)

#Create dummy variables for Slope:

cleveland.data$slope.Up<-ifelse(cleveland.data$slope==1,1,0)

cleveland.data$slope.Flat<-ifelse(cleveland.data$slope==2,1,0)

cleveland.data$slope.Down<-ifelse(cleveland.data$slope==3,1,0)

length(cleveland.data$slope.Down)

#--------------------------ELIMATE VARIABLES--------------------------------

cleveland.data = subset(cleveland.data, select = c(-cp,-slope,-trestbps,-chol,-fbs,-restecg,-cp.asymp,-slope.Down))

cleveland.data=subset(cleveland.data, select = c(-cp.asymp,-slope.Down))

names(cleveland.data)

#----------------------------MULTIPLE LINEAR REGRESSION----------------------

logistic <- glm(num ~ ., data=cleveland.data, family="binomial")

summary(logistic)