I.T. Girls
Final Report

May 5th, 2021
Health Status Prediction

SARAH PRICE
RUTH WHITEHOUSE
SHABABA KAMREEN
LILLIAN SHEPPARD

Team Member Introductions

SARAH PRICE

Team Manager
Lead Data Analyst

"I learned a lot about more advanced analytics techniques and algorithms, like PCA. The work along the way was extremely challenging, but the satisfaction of finishing made it worth it."

RUTH WHITEHOUSE

Client Liaison
Lead Data Modeler

"I learned the value of communication and teamwork and what they can do for a project. I liked learning how to analyze and cleanse data with tools commonly used in the industry."

SHABABA KAMREEN

Scrum Master
Co-Visualizer

"I learned about analyzing a huge dataset. I liked working with this team. <3"

LILLIAN SHEPPARD

Project Scribe
Co-Visualizer

"This course was an incredible way to end my college career. I learned so many valuable lessons and concepts not only about Data Science but about the importance of having a strong team dynamic and a creative yet structured mind. We have a lot to take with us to the real world."

Implemented Features

STATISTICAL

- Complete Regressional Analysis for each of our top contributing risk factors (Age, Cholesterol, Diabetes, Hypertension, Physical Inactivity, and Smoking).
- Successful application of PCA to all datasets to narrow down which of the factors have the most impact on Heart Disease.
- Thorough Statistical Analysis and K-Means Clustering for a more in-depth analysis.

VISUAL

- Create accurate and engaging visuals for all statistical tests performed on the data.

FUTURE

Unfortunately, the team was unable to implement an interactive function for our predictive model – this is something we hope another team implements should they pick up where we left off in the future.

KNOWN ISSUES

Due to the nature of our data, we had a reoccurring problem with depicting the results of our analysis accurately. Scatterplots were our graph of choice, but we were advised that future calculations would benefit from a variety of other graphs.

Additionally, the team was not able to create one of the visuals required to be done using D3.js. We encountered errors for which none of our debugging could solve. A more in-depth explanation including screenshots of our error messages can be found directly on our GitHub in a file called TODO.md.

ADDITIONAL REQUIREMENTS

The team was also responsible for creating an accompanying website and poster. Using a template for each is acceptable but editing for the website must be done in HTML.

Databases

LINKS

- https://tinyurl.com/dudch3f5
- https://tinyurl.com/3yxhf5nz
- https://tinyurl.com/4vacexj3
- https://tinyurl.com/bd8t7dyx

PRE-PROCESSING

Combining the Datasets

We collected datasets related to Heart Disease, one for each team member. Three of the datasets came from Kaggle – an online resource for datasets. The fourth was from the National Health Interview Survey. Some rudimentary data cleansing was conducted on each individual dataset separately in preparation for analysis. However, we later decided to combine each dataset (with the exception of the fourth due to formatting discrepancies) based on common risk factors that the datasets shared.

HYPOTHESES

As a group, we determined 5 hypotheses we wanted to test based on the exploration of our individual datasets. Because of challenges with the combination of the datasets, we have 4 hypotheses on the merged dataset (Sarah, Ruth, Lily), and one on the individual (May).

- The higher the cholesterol, the more likely a person is to have heart disease.
- Older patients have a higher likelihood of being heart disease positive.
- There is a direct correlation between patients who have diabetes and those who have heart disease.
- Using hypertension, physical inactivity, and smoking as risk factors, people are at the highest risk for heart disease if they have hypertension.

METHODS FOR ANALYSIS

Regression Analysis

To determine the validity of the hypotheses we have formulated, we ran a linear regression tests between the factors in each hypothesis. We tested this using an R^2 value - this is a number between 0 and 1 that tells you how strongly correlated the factors are. The closer the R^2 value is to 1, the stronger the correlation. For simplicity's sake, we looked at the correlations using a Linear Regression model.

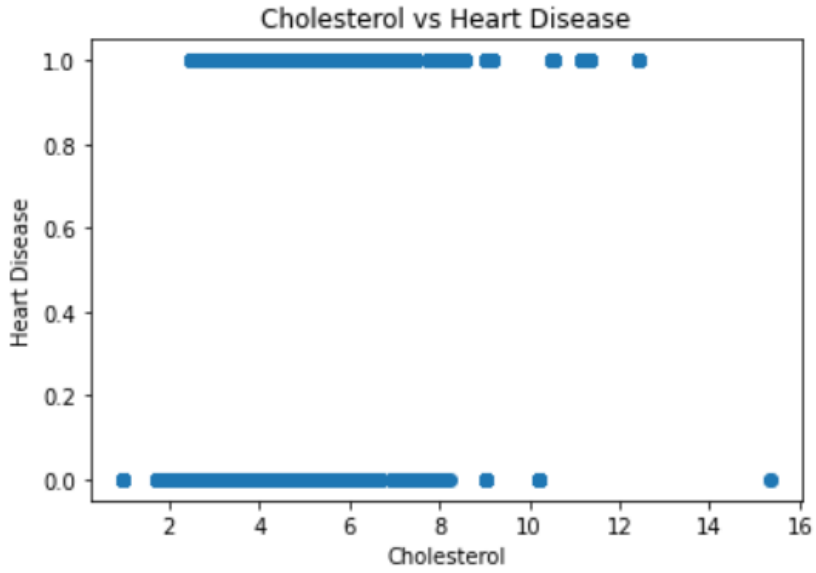
Applying PCA

In order to get a better look at our data, we applied PCA - principal component analysis - to narrow down which of the factors have the most impact on heart disease. For PCA to work, we had to convert all of the string values in the dataset into numerical values.

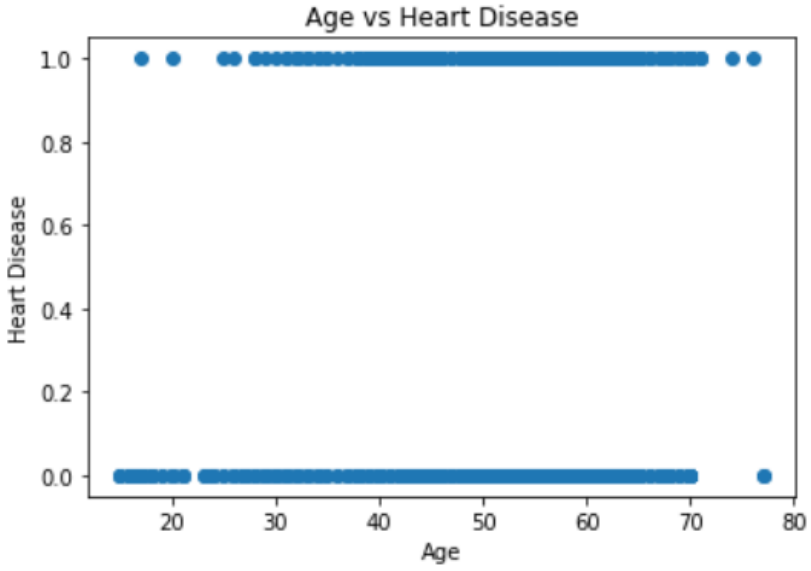
K-Means Clustering

After determining the important factors, we looked at them in more depth. To look at the results of clustering, we applied K-Means clustering to the age column. Because our diagnosis column only contains 2 categories of values, we chose to set up 2 cluster centers, as shown in the below code (n_clusters=2).

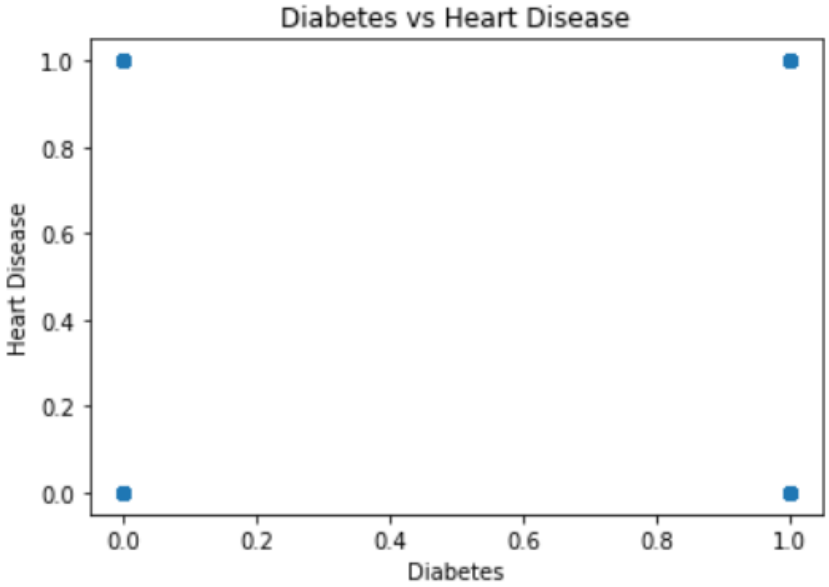
Results



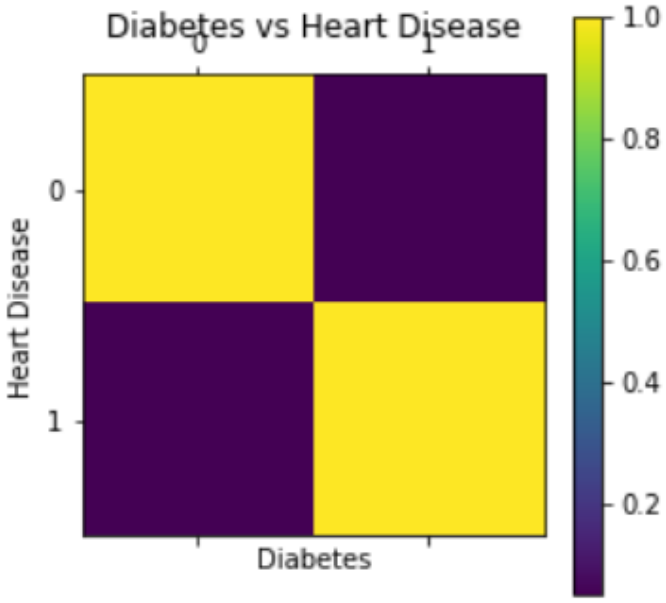
We hypothesized that higher cholesterol put the patient at higher risk for heart disease. The linear regression analysis between cholesterol and heart disease shows an R^2 value of 0.1526. This is not a very strong correlation.



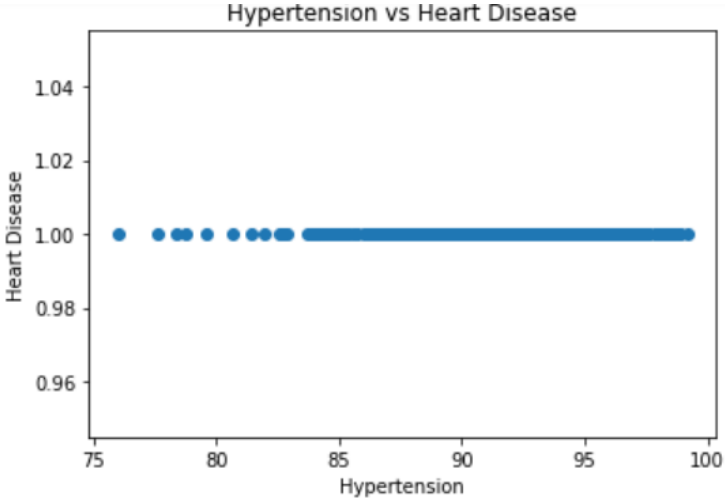
The next hypothesis we wanted to test was whether age impacted heart disease positivity. While the binary diagnosis still makes the graph hard to read, we can clearly see that the age for heart disease positive patients is typically higher than patients without.

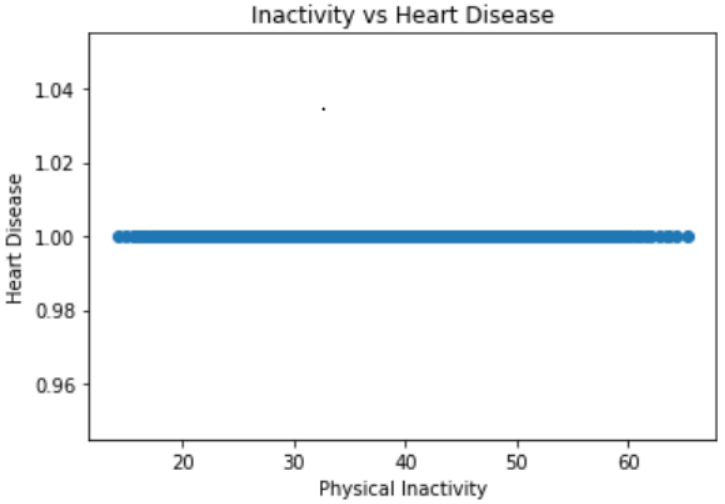


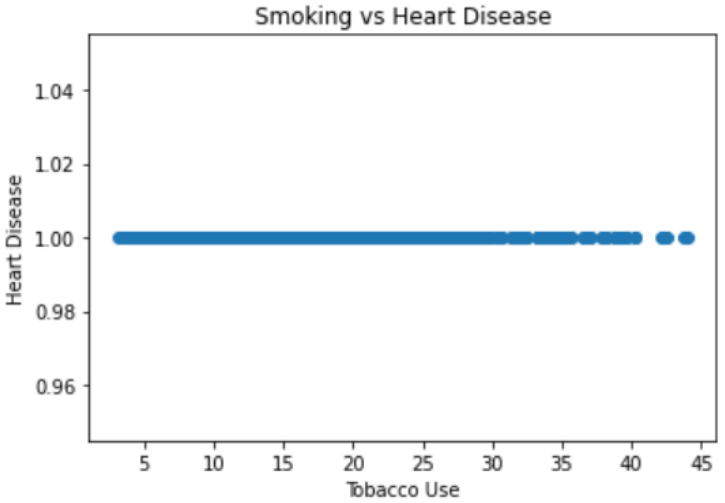
Based on our individual data exploration, we theorize that patients who have diabetes have a higher risk of heart disease. This correlation is also extremely low - the R^2 value is only a 0.0499. Because both factors in this correlation are binary, a scatterplot is not the best way to show the correlation between the two factors. In order to get a better look at the information, we'll try a correlation matrix.



The correlation matrix does not show any more in-depth explanation, but it makes the information easier to see.







This hypothesis was made using the individual dataset that our team is struggling to merge into the other 3 datasets. The factors we would like to compare are the likelihood of cardiovascular disease compared to the risk factors: hypertension, physical inactivity, and smoking. Unfortunately, this does not show us much information about whether this factor contributes to heart disease, since all of these cases were positive.

QUICK LINKS

Screencast Demo