To Save a Heart

Henry Oluka, Paul Parkin, and Ryan Woodward

Abstract—Sudden cardiac arrests are among the leading causes of death each year. To improve the health outcomes of heart disease patients, signs leading up to a potential cardiac arrest need to be presented in such a way that they are easy to understand. By performing data analysis we have found correlative evidence between certain factors such as age, heart defects, exercise induced angina, and resting heart rates. An effective visualisation would visualize these factors and their relationships in a manner that improves readability for average citizens. Our project goal is to determine indicators that can be monitored and used to determine when health complications (as a result of heart disease) arise, and possibly contribute to further research and discoveries on this topic. This study promises to provide Physicians with another tool for quickly visualizing diagnostic test data when caring for their patients. We believe that such a visualisation can aid a physician's judgment and decision-making especially in cardiac arrest cases and potentially improve health care delivery and patient health outcomes.

Index Terms—Health, heart disease, heart and hypertension, biology

1 Introduction

About 2.4 million Canadian adults aged 20 years and above are living with heart disease; the second leading cause of death in Canada [1]. Non-fatal heart diseases can still lead to hospitalization, cardiac arrests and other complications where heart disease is a prominent underlying factor in cardiac arrests. This makes heart diseases a threat to Canadians and a financial burden to the healthcare system in Canada.

1.1 Purpose

This project aims to determine patterns in heart diseases that can help physicians effectively care for patients presenting patterns of a fatal complication of a heart disease well in advance of the event.

To achieve this, we asked the question: can selective and associative data be interactively visualized to model the sequence of interactive activities of a physician during patient diagnosis?

1.2 Approach

The best practices of modern DevOps (which is continuous integration and continuous delivery) were used to manage the project. The phases of our approach being: Phase 1 (Requirements Collection & Analysis), Phase 2 (Design), Phase 3 (Development - Continuous integration and Continuous delivery), Phase 4 (Final Presentation and Report).

Data analysis was conducted using MS Excel and RStudio, while our preliminary visualizations were created with D3, and Microsoft Bi. The final visualization was created using Tableau, and Sisense.

1.3 Main Findings

We had anticipated that a visualization might provide clues as to whether or not a patient was suffering from heart disease. However, our visualization was unable to provide any human-in-the-middle insight. Our findings were primarily minor and did not reveal anything significant to the problem. The visualizations presented in this project are unsuited to providing an answer to the proposed research question.

2 Motivation, Data, and Data questions

The data was selected for personal reasons due to a team member's loss of a family member. We intended to determine patterns of fatal heart attacks through analysis and visualization of research data from 303 patients with and without heart disease.

2.1 Data

The first dataset (cleveland.data) can be downloaded publicly from https://archive.ics.uci.edu/ml/datasets/Heart+Disease. The data was collected by the Cleveland Clinic Foundation. There are 282 data cases and 76 data dimensions. The data dimensions describes patients by id, ccf, age, sex, painloc, painexer, relrest, pncaden, cp, trestbps, htn, chol, smoke, cigs, years, fbs, dm, famhist, restecg, ekgmo, ekgday, ekgyr, dig, prop, nitr, pro, diuretic, proto, thaldur, thaltime, met, thalach, thalrest, tpeakbps, tpeakbpd, dummy, trestbpd, exang, xhypo, oldpeak, slope, rldv5, rldv5e, ca, restckm, exerckm, restef, restwm, exeref, exerwm, thal, thalsev, thalpul, earlobe, cmo, cday, cyr, num, lmt, ladprox, laddist, diag, cxmain, ramus, om1, om2, rcaprox, rcadist, lvx1, lvx2, lvx3, lvx4, lvf, cathef, junk, name. The data was cleaned and transferred to a csv file using Java.

The second dataset can be downloaded publicly as a csv file from https://www.kaggle.com/ronitf/heart-disease-uci. This was collected from two cardiology institutes and two universities. There are 303 data cases and 14 data dimensions. The data dimensions describe patients by age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, and target.

2.2 Data Questions

Through the analysis and visualization of the research data, we intend to answer three data questions.

2.2.1 Are there emerging patterns for patients who experience cardiac issues at a younger age?

This question will require the grouping of the values for the age data dimension into 4 categories, one of which will be the youngest age category. The youngest age category will be visualized in comparison to the other categories to identify any patterns or relationship with the remaining data dimensions.

2.2.2 Are there warning signs that a cardiac arrest is imminent given the pain described and the underlying symptoms?

This question requires analysing the dataset for data dimension values that indicate signs of a cardiac arrest. The question requires th/e "human-in-the-middle" in making connections between the pain and the symptoms described by a cardiac arrest patient, and the data

- *Ryan Woodward is a UVic Computer Science, Software Engineering undergraduate student. Email: rnw@uvic.ca.*
- *Henry Oluka is a UVic Computer Science, Health Information Science and Statistics undergraduate student. Email: holuka@uvic.ca.*
- *Paul Parkin is a UVic Computer Science undergraduate student. Email: paulparkin@uvic.ca.*

dimension values that correspond to a cardiac arrest patient's health status.

2.2.3 Is there a threshold at which a patient needs invasive surgery, or can surgery be avoided?

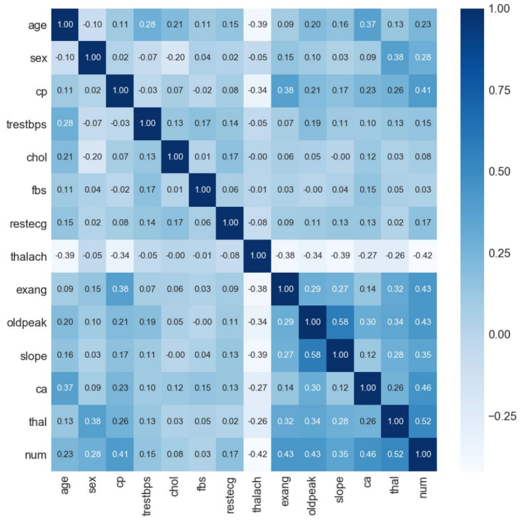
This question requires the "human-in-the-middle" to make decisive judgements on what benchmarks should be ascribed to each data dimension. These benchmarks will further group values for a given dimension into categories (mild, worse, and severe) depending on the severity of the health status the value represents. This should allow the identification of thresholds (per data dimension) that indicate a patient needs invasive surgery.

2.3 Informal Interviews

We interviewed two members on staff at Vancouver General Hospital; Radiologist Dr Nathan Plaa and CT Technician Andrew Moro. Both indicated that a visual chart could help with the complexities of cardiac diagnostics. However, Dr Plaa warned that for the most part our dataset contained information that already required clinical analysis, at which point a decision would be made then and there on whether or not an individual needed more attention based upon technical expertise.

3 Related work

Jae Duk Seo has performed some basic data exploration using the python programming language interfacing with pandas DataFrame and using NumPy algorithms [2]. Using data exploration, Seo has been able to separate categorical and quantitative values. He has also been able to provide a variance/covariance matrix (See Figure 1) which observes correlative relationships between the attributes [4]. Visualizations are mostly histograms, box plots, and bar plots [2].



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
age	1.00	-0.10	0.11	0.28	0.21	0.11	0.15	-0.39	0.09	0.20	0.16	0.37	0.13	0.23
sex	-0.10	1.00	0.02	-0.07	-0.20	0.04	0.02	-0.05	0.15	0.10	0.03	0.09	0.38	0.28
cp	0.11	0.02	1.00	-0.03	0.07	-0.02	0.08	-0.34	0.38	0.21	0.17	0.23	0.26	0.41
trestbps	0.28	-0.07	-0.03	1.00	0.13	0.17	0.14	-0.05	0.07	0.19	0.11	0.10	0.13	0.15
chol	0.21	-0.20	0.07	0.13	1.00	0.01	0.17	-0.00	0.06	0.05	-0.00	0.12	0.03	0.08
fbs	0.11	0.04	-0.02	0.17	0.01	1.00	0.06	-0.01	0.03	-0.00	0.04	0.15	0.05	0.03
restecg	0.15	0.02	0.08	0.14	0.17	0.06	1.00	-0.08	0.09	0.11	0.13	0.13	0.02	0.17
thalach	-0.39	-0.05	-0.34	-0.05	-0.00	-0.01	-0.08	1.00	-0.38	-0.34	-0.39	-0.27	-0.26	-0.42
exang	0.09	0.15	0.38	0.07	0.06	0.03	0.09	-0.38	1.00	0.29	0.27	0.14	0.32	0.43
oldpeak	0.20	0.10	0.21	0.19	0.05	-0.00	0.11	-0.34	0.29	1.00	0.58	0.30	0.34	0.43
slope	0.16	0.03	0.17	0.11	-0.00	0.04	0.13	-0.39	0.27	0.58	1.00	0.12	0.28	0.35
ca	0.37	0.09	0.23	0.10	0.12	0.15	0.13	-0.27	0.14	0.30	0.12	1.00	0.26	0.46
thal	0.13	0.38	0.26	0.13	0.03	0.05	0.02	-0.26	0.32	0.34	0.28	0.26	1.00	0.52
num	0.23	0.28	0.41	0.15	0.08	0.03	0.17	-0.42	0.43	0.43	0.35	0.46	0.52	1.00

Figure 1: Correlation Matrix

The prevalence of blood pressure status in US adults has also been used as a preliminary visualization [3]. The stacked bar graph (See Figure 2) provides a simple tool to view the percentage of US adults that fall into the three blood pressure categories: Normotension, Prehypertension, and Hypertension. It provides an easy look into blood pressure, but doesn't have enough factors relating to preventing heart attacks.

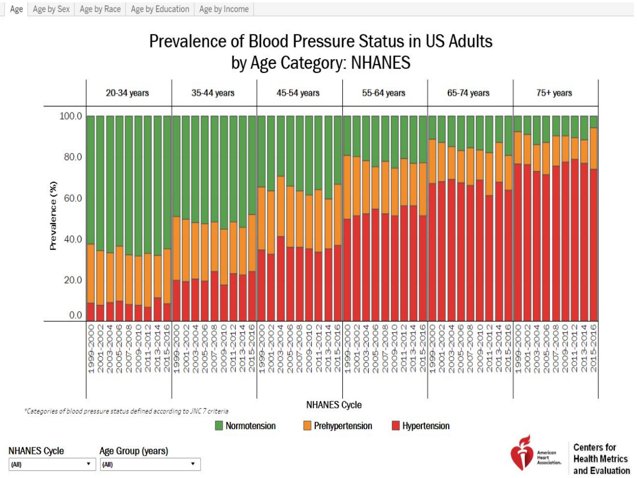
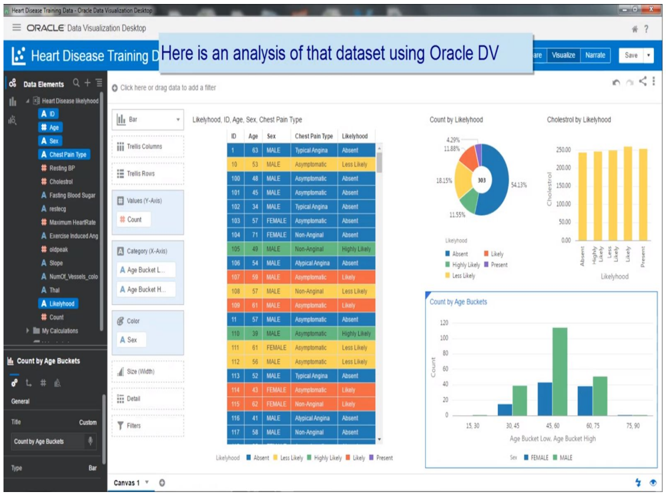


Figure 2: Disease vs Base Heart Rate

Machine learning algorithms with Oracle Data Visualization are used to predict heart disease in patients using a multi-classification neural net model [5]. The network (See Figure 3) uses attributes such as cholesterol, blood pressure and chest pain type to predict the likelihood of heart disease in patients. Likelihood is classed as absent, less likely, likely, highly likely, and present.



Here is an analysis of that dataset using Oracle DV

Likelyhood, ID, Age, Sex, Chest Pain Type

ID	Age	Sex	Chest Pain Type	Likelyhood
1	63	MALE	Typical Angina	Absent
10	53	MALE	Asymptomatic	Less Likely
100	48	MALE	Asymptomatic	Absent
101	45	MALE	Asymptomatic	Absent
102	34	MALE	Typical Angina	Absent
103	57	FEMALE	Asymptomatic	Absent
104	71	FEMALE	Non-Anginal	Absent
105	49	MALE	Non-Anginal	Highly Likely
106	54	MALE	Alypical Angina	Absent
107	59	MALE	Asymptomatic	Likely
108	57	MALE	Non-Anginal	Less Likely
109	61	MALE	Asymptomatic	Likely
11	57	MALE	Asymptomatic	Absent
110	39	MALE	Asymptomatic	Highly Likely
111	61	FEMALE	Asymptomatic	Less Likely
112	56	MALE	Asymptomatic	Less Likely
113	52	MALE	Typical Angina	Absent
114	43	FEMALE	Asymptomatic	Likely
115	62	FEMALE	Non-Anginal	Likely
116	41	MALE	Alypical Angina	Absent
117	58	MALE	Non-Anginal	Absent

Figure 3: Analysis using Oracle Data Visualization

3.1 Alternative Solutions

The existing visualizations related to our project can be expanded upon in many ways. By examining the correlation matrix it provides us with insight into whether the variables can be considered associative. The data analysis also provides information to categorize the data appropriately as quantitative, nominal, and ordinal. This analysis lends itself to a visualization that focuses on position, color saturation, and color hue. The length of certain data dimensions such as heart rate vs cholesterol will have to be normalized in order to maintain integrity of the data.

Alternative visualizations will likely make use of the length function in order to reach a length threshold. This is to say that if the length of these data dimensions were stacked they would exceed the threshold which in turn would indicate that the patient is at a certain risk level. The risk level thresholds could in turn be expanded upon to indicate correlative actions to be taken by the patient or physician.

4 Design Justification

4.1 Polar Charts as opposed to Scatterplots

In an attempt to represent as many of the data dimensions as possible, we determined to use interactive polar charts. Bar graphs and scatter plots were determined to be too heavily clustered. A large problem was the complexity of the data. As an example, there could be two individuals with identical data dimensions, but one would have heart disease and the other would not. When implementing a polar chart, we could see the differences based on the length of each of the polar coordinates.

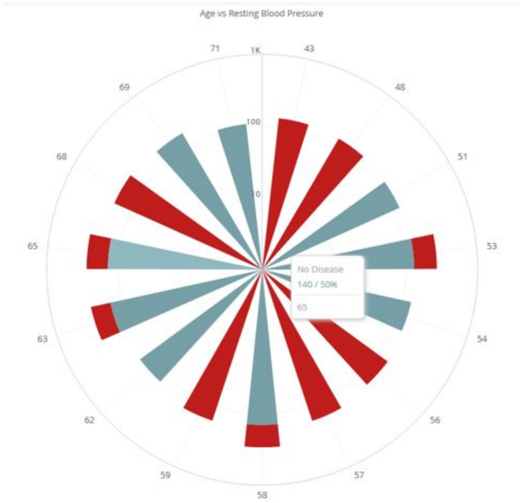


Figure 4: Average Resting Blood Pressure Polar Chart

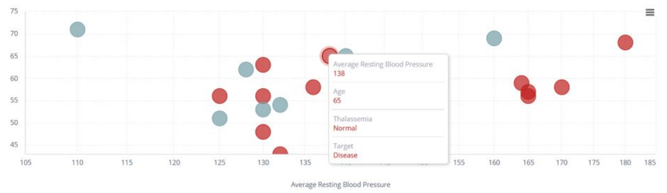


Figure 5: Average Resting Blood Pressure Scatterplot Disease

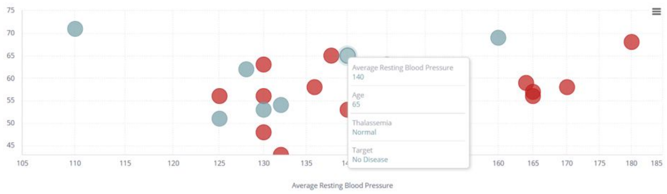


Figure 6: Average Resting Blood Pressure Scatterplot No Disease

4.2 Interactive Build Visualization

Since several data cases had near identical dimensions except for our target, this presented a large problem in visualizing this. We could not use shapes, colors, or size to differentiate these. As a result, it was determined we could eliminate a lot of noise if we did not have to show data that was not pertinent at the time. Also, the numerical display on the edge of a polar chart can only indicate one thing being measured at a time. Any other information simply gets bundled into the pole. Therefore, to be able to remove the confusion relating to the length of the poles, we felt that using a few measurements per polar chart had a greater positive effect.

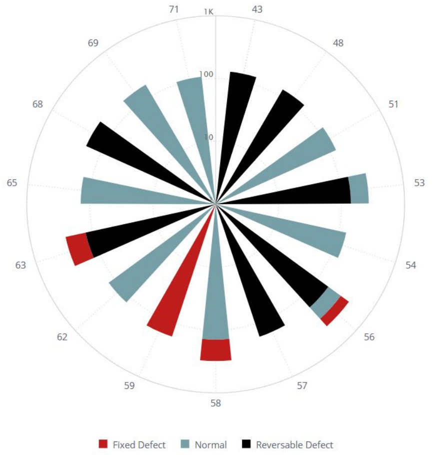


Figure 7: Confusing Measurements on Polar Chart

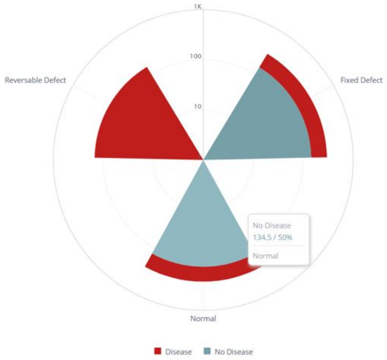


Figure 8: Concise Measurements to Assist User in Decision Making

4.3 Layout

The visual aesthetic to extend polar charts was based upon an interactive choice that the user would have to make for each polar chart. Each arm of the polar chart corresponds to a choice that a user needs to make. Once a choice was made, that arm would extend into another polar chart indicating more interactivity for the user. The existing choices would shrink into the background and the user's focus would shift to the polar chart in front of them.

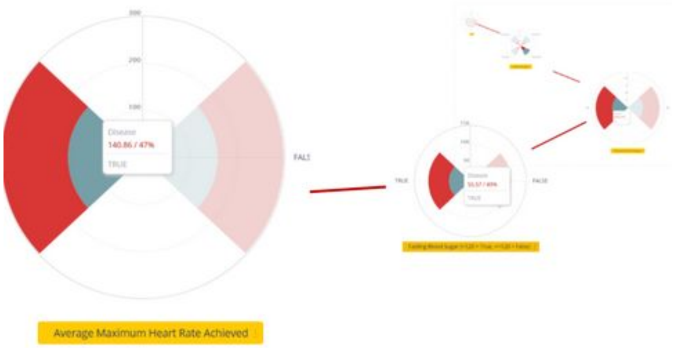


Figure 9: User Flow through Visualization Choices

When interacting with the visualization, the user can answer questions about data that relates to them. The resulting information will display what their current probability of heart disease is based on comparing their answers with the data.

4.4 Alternatives Visulaizations Considered

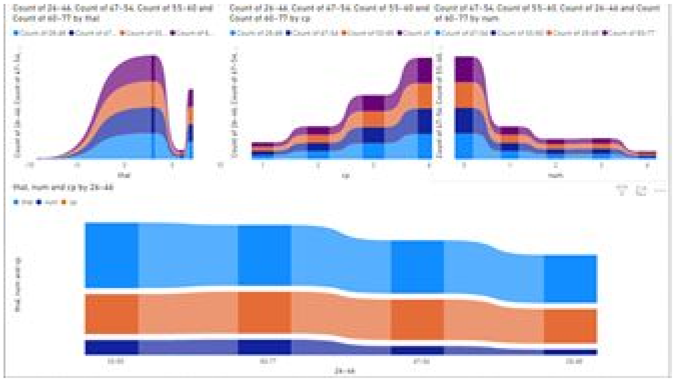


Figure 10: Ribbon chart for Age category vs. Heart disease attribute

A ribbon chart was created using Microsoft PowerBi. It demonstrates a correlation between chest pain, thalassemia, heart disease, and age. Age categories are shown as colour ribbons (See Figure 10). The number of test subjects (per age category) that scored a given value within each heart disease attribute (from left to right: thal, cp, num) is shown. It shows that as Age decreases, the number of test subjects who are observed to have a given heart disease attribute (thal, cp, num) decreases.

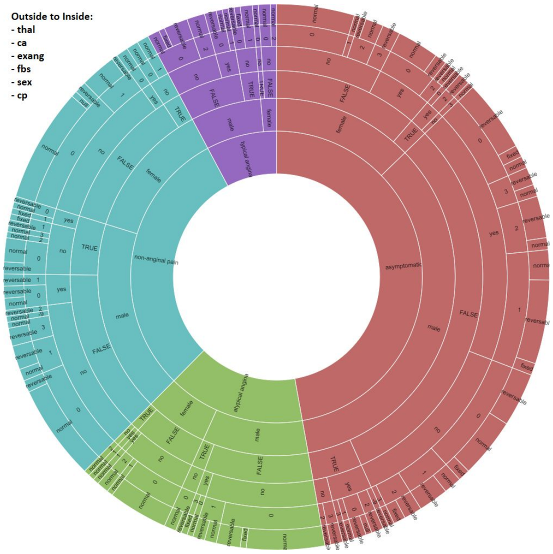


Figure 11: RAWGraphs SunBurst Visualization

The sunburst visualization was considered as a method for a user to trace their data in order to come to a conclusion as to whether or not they had heart disease. This visualization provides interaction as well and will change the hue of surrounding data points as you traverse through the rings. This conflicted with the design philosophy that complexity needed to be reduced in order to provide a better experience for the user. The sunburst visualization doesn't provide opportunity for data discovery. The data is too clustered , and there are no visual marks to indicate whether there could be a potential pattern.

It was determined that polar charts would be less crowded and a more effective visualization model.

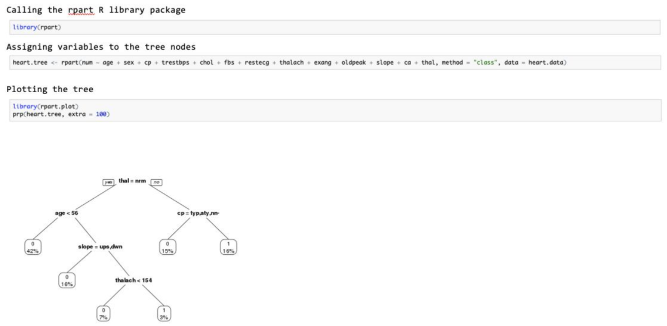
5 Implementation

The implementations are based on the brainstorming of several hand drawn sketches. Then data analysis was conducted on the downloaded data to determine the feasibility of these sketches using RStudio. We plotted 14 heart disease attributes that were commonly referenced by other papers into a Classification Tree and determined that 3 attributes: thal, cp and age were good predictors of the predicted attribute num (Figure 5). These 3 attributes were then used in some of the visualization models with fewer than 10 attributes (10 being few enough for separation of visual marks).

As a group we informally evaluated the models based on which had interactivity which we consider the core component in our visualization. Our final visualization was coded using Tableau, and Sisense.

```
##       age             sex               cp           trestbps
##  Min.   :29.0   female: 97   typical     : 23   Min.   : 94
##  1st Qu.:48.0   male  :206   atypical    : 50   1st Qu.:120
##  Median :56.0                non-anginal : 86   Median :130
##  Mean   :54.4                asymptomatic:144   Mean   :132
##  3rd Qu.:61.0                                   3rd Qu.:140
##  Max.   :77.0                                   Max.   :200
##      chol          fbs            restecg          thalach      exang
##  Min.   :126   false:258   normal      :151   Min.   : 71   no :204
##  1st Qu.:211   true : 45   stt         :  4   1st Qu.:134   yes: 99
##  Median :241               hypertrophy:148   Median :153
##  Mean   :247                                 Mean   :150
##  3rd Qu.:275                                 3rd Qu.:166
##  Max.   :564                                 Max.   :202
##     oldpeak              slope        ca              thal       num
##  Min.   :0.00   upsloping  :142   0.0 :176   normal    :166   0:164
##  1st Qu.:0.00   flat       :140   1.0 : 65   fixed     : 18   1: 55
##  Median :0.80   downsloping: 21   2.0 : 38   reversable:117   2: 36
##  Mean   :1.04                     3.0 : 20   NA's      :  2   3: 35
##  3rd Qu.:1.60                     NA's:  4                    4: 13
##  Max.   :6.20
```

Figure 12: R Output Summary of the 14 Heart Disease Attributes.



Calling the rpart R library package

```
library(rpart)
```

Assigning variables to the tree nodes

```
heart.tree <- rpart(num ~ age + sex + cp + trestbps + chol + fbs + restecg + thalach + exang + oldpeak + slope + ca + thal, method = "class", data = heart.data)
```

Plotting the tree

```
library(rpart.plot)
prp(heart.tree, extra = 100)
```

Figure 13: A Snapshot of the Data Analysis Conducted using R Studio

5.1 Challenges

Technical challenges arose as there were multiple software changes to determine how our visualization was to be created. At first we were using Microsoft PowerBi, however the visualizations it provided didn't promote much discovery of the data. It did have a radial option, but it was missing interactivity and the radial option didn't provide an answer to the probability of a patient having heart disease. So, we shifted to D3 to try to implement our design. But the complexity of learning the software was too high as there were difficulties getting aspects of the code to run properly. There were a lot of unforeseen errors that required debugging. So as such, it was determined that other software should be investigated through a major shift from D3 to Tableau and Sisense. Tableau was able to provide preliminary discovery of the data. However, it was determined that Sisense has the best toolset to handle radial calculations.

Another major challenge was the limited amount of data cases. There were several instances where data cases had similar values for dimensions but they had a different target. As a result when performing analysis on heart disease, an individual could provide all of their information through our interactive visualization and still receive a result stating they have a 50% chance of having heart disease.

6 Data Insights

The 3 attributes, thal, chest pain and age are good covariates for predicting the value of num which is the angiographic disease status of a heart disease patient (Figure 12). Figure 10 indicates that younger age corresponds with lower observed heart disease symptoms. Test subjects within the age category of 55-60 were shown to have more heart disease symptoms (Figure 10). It was determined that thalassemia fixed defect patients are almost guaranteed to have heart disease. But this is no surprise as invasive heart procedures such as fixing a defect normally do not occur for healthy hearts.

We were not anticipating how there were few patterns discovered. It was assumed that patterns would emerge based upon key correlative values as discovered in the correlation matrix. Data values such as thalassemia and chest pain type, or even exercise induced angina were assumed to point drastically to a rise in the probability that a patient could have heart disease. Unfortunately there were data cases where a patient would have a blood disorder, chest pains caused by exercise, and normal anginal chest pains and it was still determined that the patient did not have heart disease.

7 Research Findings

The design process unfortunately did not present findings towards our research question. It did not reduce the technical expertise needed to diagnose heart disease. The results of the interactive visualization mostly led to users having 48% - 52% probability of having heart disease based on all data dimensions they entered. This is not a good result to give confidence to a patient nor a doctor as to whether or not a potential patient has heart disease without doing any invasive procedures.

When Dr Nathan Plaa was shown the results, his statements echoed his original ones. In order to have the results needed to use the interactive visualization, a patient would have needed to have already visited a physician. A patient could not diagnose their chest pain as atypical, nor could a patient determine their own cholesterol levels without going to a lab. A doctor wouldn’t need this visualization to determine whether or not a patient was showing signs of heart disease. Since the patient needs the doctor for these test results, the doctor would tell them the risk for heart disease without a need for a visualization.

He did comment on future work stating that if more data cases were collected then it could present better information.

7.1 Evaluation of Design

We had multiple designs that were elegant but we made our decision on which design to pick by considering how robust and customizable the tool used to create it was. We settled on using Sisense to build the polar charts for the interactive portion of the visualization. The charts were intended to simplify the process of diagnosing heart disease and to remove as much confusion as possible for the user. However, the removal of most of that information may have removed the potential for visualization discovery.

8 Discussion

8.1 Strengths and Weaknesses

A strength of our implementation was the ease of use and the removal of complexity for the user. Since our original plan was to mimic a physician’s diagnostic process through an interactive visualization, we feel we have accomplished that. A user was able to click on sections and the visualization would display a probability of them having heart disease.

The design of the visualization was a weakness in itself. If the interactivity of the visualization only leads to the discovery of a single answer, then there isn’t enough information revealed by it. The visualization itself is very unappealing as well. The intention to remove all of the complexity removed much of what could have made the visualization appealing. As a result the final image seems incomplete.

8.2 Relation to Literature

Our findings have been quite similar to basic medical data exploration [2]. There was a correlation found between the data points. This correlation begins to degrade once more than one comparison is used however. For example there is a strong correlation between thalassemia and heart disease, as well as a type of chest pain and heart disease. But what if the chest pain that an individual has is typical anginal, and the thalassemia was a fixed defective valve. We saw that correlation breaks down entirely as there were multiple times when the visualization would show you have a 50/50 chance of having heart disease. This could be the result of the small amount of data cases though, as some of these could be considered as outliers.

8.3 Future Work

The problem of diagnosing and predicting heart disease lends itself more so to machine learning [5] [8] and medical tests [6]. There are advancements being made in visualizing arteries and organs such as HemoVis which creates a 2D visualization of arteries performing better than a 3D visualization (52% increase) [7]. Future work will likely be using a trained neural net to create visualizations predicting heart disease with patient medical information/history to display to physicians. This way physicians can make decisions based on the results of the neural network along with technical expertise/experience and medical tests.

There are ways these visualizations can be improved upon. Several of the data cases being nominal, specifically chest pain and thalassemia, could potentially be weighted in order to have a more accurate measurable value. This would mean that these could then be quantitative and the chances of having one could potentially lead to an easier visual in terms of pattern discovery. The dataset also only had a binary target of heart disease or no heart disease. This itself could be modified into having multiple answers according to the severity of the disease.

References

[1] Government of Canada. Heart disease in Canada (2017). https://www.canada.ca/en/public-health/services/publications/diseases-conditions/heart-disease-canada.html Accessed March 2, 2020.

[2] Basic Medical Data Exploration / Visualization - Heart Diseases (2018) https://towardsdatascience.com/basic-medical-data-exploration-visualization-heart-diseases-6ab12bc0a8b7 Accessed February 26, 2020.

[3] Vise, Sharlotte. “Prevalence of Blood Pressure Status in US Adults by Age Category: NHANES.” *Health Metrics*, 18 June 2019, healthmetrics.heart.org/prevalence-blood-pressure-status-us-adults-age-category-nhanes/.

[4] Mohiteud. “Heart Diseases Visualization.” *Kaggle*, Kaggle, 27 Mar. 2019, www.kaggle.com/mohiteud/heart-diseases-visualization.

[5] *How strong are our hearts? Oracle DV ML helps with the answer.* (2017, November 16). Oracle Underground Bi and Dataviz. http://oracledataviz.blogspot.com/2017/11/how-strong-are-ours-hearts-oracle-dv.html.

[6] *Heart Disease.* (2018, March 22). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/heart-disease/diagnosis-treatment/drc-20353124

[7] *To diagnose heart disease, visualization experts recommend a simpler approach.* (2011, October 27). Harvard John A. Paulson School of Engineering and Applied Sciences. https://www.seas.harvard.edu/news/2011/10/diagnose-heart-disease-visualization-experts-recommend-simpler-approach

[8] Smiley, S. (2020, January 11). *Diagnostic for Heart Disease with Machine Learning*. Medium. https://towardsdatascience.com/diagnostic-for-heart-disease-with-machine-learning-81b064a3c1dd