DSDHT FALL 2017
Carlos Sathler (cssathler@gmail.com)

Module 2 – Tableau Exercise

Introduction

In this report we analyze 2 datasets with heart-disease diagnostic information. The first data set is from the Hungarian Institute of Cardiology and the second is from the V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. The datasets are available here: http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/. The purpose of our analysis is to verify if common, well held beliefs about heart disease hold in the datasets; on the Hungarian dataset we look for correlation between heart disease and gender, age, cholesterol and high blood pressure levels; on the Cleveland dataset we look for correlation between heart disease, chest pain and heart rate during exercise. This analysis was conducted entirely in Tableau.

Dataset Description

A complete description of the 2 datasets used can be found here: http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names.

Below is a summary of the raw data we imported into Tableau, and what fields were used in our analysis.

Field Number	Description	Comments
1. #3 (age)	3 age: age in years	Used in analysis of the Hungarian dataset.
2. #4 (sex)	4 sex: sex (1 = male; 0 = female)	Used in analysis of both the Hungarian and the Cleveland dataset.
3. #9 (cp)	9 cp: chest pain type	Used in the analysis of the Cleveland dataset. Values: 1-Typical Angina; 2-Atypical Angina; 3-Non-anginal Pain; 4-Asymptomatic.
4. #10 (trestbps)	10 trestbps: resting blood pressure (in mm Hg on admission to the hospital)	Used in analysis of the Hungarian dataset.
5. #12 (chol)	12 chol: serum cholestoral in mg/dl	Used in analysis of the Hungarian dataset.
6. #16 (fbs)	16 fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)	Not used in our analysis.
7. #19 (restecg)	19 restecg: resting electrocardiographic results	Not used in our analysis.
8. #32 (thalach)	32 thalach: maximum heart rate achieved	Used in analysis of Cleveland dataset.
9. #38 (exang)	38 exang: exercise induced angina (1 = yes; 0 = no)	Used in analysis of Cleveland dataset.
10. #40 (oldpeak)	40 oldpeak = ST depression induced by exercise relative to rest	Not used in our analysis.
11. #41 (slope)	41 slope: the slope of the peak exercise ST segment	Not used in our analysis.
12. #44 (ca)	44 ca: number of major vessels (0-3) colored by flourosopy	Not used in our analysis.
13. #51 (thal)	51 thal: 3 = normal; 6 = fixed defect; 7 = reversable defect	Not used in our analysis.
14. #58 (num) (the predicted attribute)	58 num: diagnosis of heart disease (angiographic disease status)	Used in both analysis. In the Hungarian dataset: 0-less than 50% diameter narrowing; 1-more than 50% diameter narrowing. In the Cleveland dataset: values from 0 to 4 where 0 represents absence of disease.

Hungarian Dataset Analysis

On the Hungarian dataset we looked for correlation between heart disease and gender, age, cholesterol and high blood pressure levels. It is a common held belief that heart disease is more common in older individuals, males, and that high cholesterol and high blood pressure are risk factors. Can we find confirmation of this in our dataset?

Pre-processing

The following pre-processing steps were performed:

1. Changed numeric categorical fields to dimensions with meaningful labels, e.g., sex=0 became Gender=Male
2. Binned cholesterol and blood pressure variables

DSDHT FALL 2017
Carlos Sathler (cssathler@gmail.com)

Analysis

Plot HeartDiseaseByGender

- Clearly more males are included in the study, and we see there are more healthy as well as sick males.
- The *proportion* of sick males vs. females is clearly higher than the *proportion* of healthy males vs. females.

Plot HeartDiseaseByGender

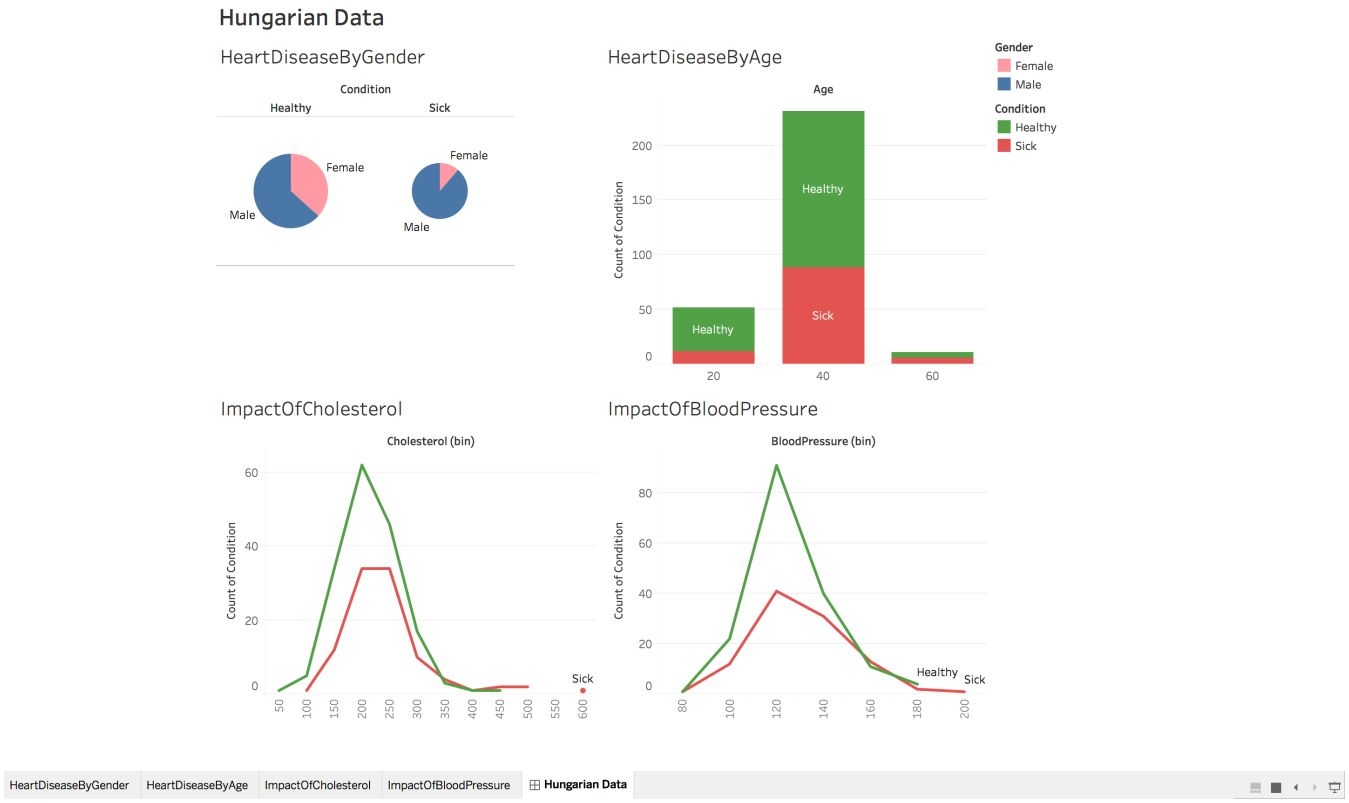
- Most individuals fall in the 40 to 60 age range.
- The *proportion* of sick vs. healthy individuals increases as the age increases.
- At the 60 and over age range approximately 50% of the individuals observed are sick.

Plot ImpactOfCholesterol

- The plot shows a higher number of individuals around the 200 ml/dl cholesterol level.
- The slope of the increase of healthy individuals is higher than the slope of the increase of sick individuals up to the 200 ml/dl cholesterol level; beyond that level the slope of the decrease of healthy individuals is steeper than the decrease in number of sick individuals.
- At high levels of cholesterol all observed individuals are sick.

Plot ImpactOfBloodPressure

- The plot shows a higher number of individuals around the 120 blood pressure level.
- The slope of the increase of healthy individuals is higher than the slope of the increase of sick individuals up to the 120 blood pressure level; beyond that level the slope of the decrease of healthy individuals is much steeper than the decrease in number of sick individuals.
- At high levels of blood pressure all observed individuals are sick.



DSDHT FALL 2017
Carlos Sathler (cssathler@gmail.com)

Conclusion – Hungarian Dataset

The analysis of the Hungarian dataset seems to confirm the common belief that the risk of heart disease is higher for males and that it increases with age, high cholesterol and blood pressure levels. This conclusion is possible only if we assume that sample used in this study is representative of the Hungarian population at large.

Cleveland Dataset Analysis

On the Cleveland dataset we looked for correlation between heart disease and chest pain and maximum heart rate during exercise. We also looked for variations due to gender. It is a common held belief that symptoms of heart disease, namely heart pain, will flare up during exercise and that sick individuals are not as capable of enduring intense exercise. Can our data confirm these commonly held beliefs?

Pre-processing

The following pre-processing steps were performed:

1. Changed numeric categorical fields to dimensions with meaningful labels, e.g., exercise induced angina pain=0 became ExerciseAngina=Yes
2. Binned maximum heart rate during exercise and blood pressure variables

Analysis

Plot ExercisePainAngina

- The number of sick vs. healthy individuals is about the same.
- More than 50% of the sick individuals experience exercise induced angina pain while less than 20% of healthy individuals experience exercised induced angina pain.

Plot MaxHeartRate

- The proportion of healthy vs. sick individuals increases significantly with max heart rate achieved during exercise.
- If an individual's max heart rate is less than 140 there is a 50% chance that individual is sick.
- The higher the max heart rate, the less likely the individual is sick.
- If an individual's max heart rate is 180 or higher, the chance is small to none the individual is sick.

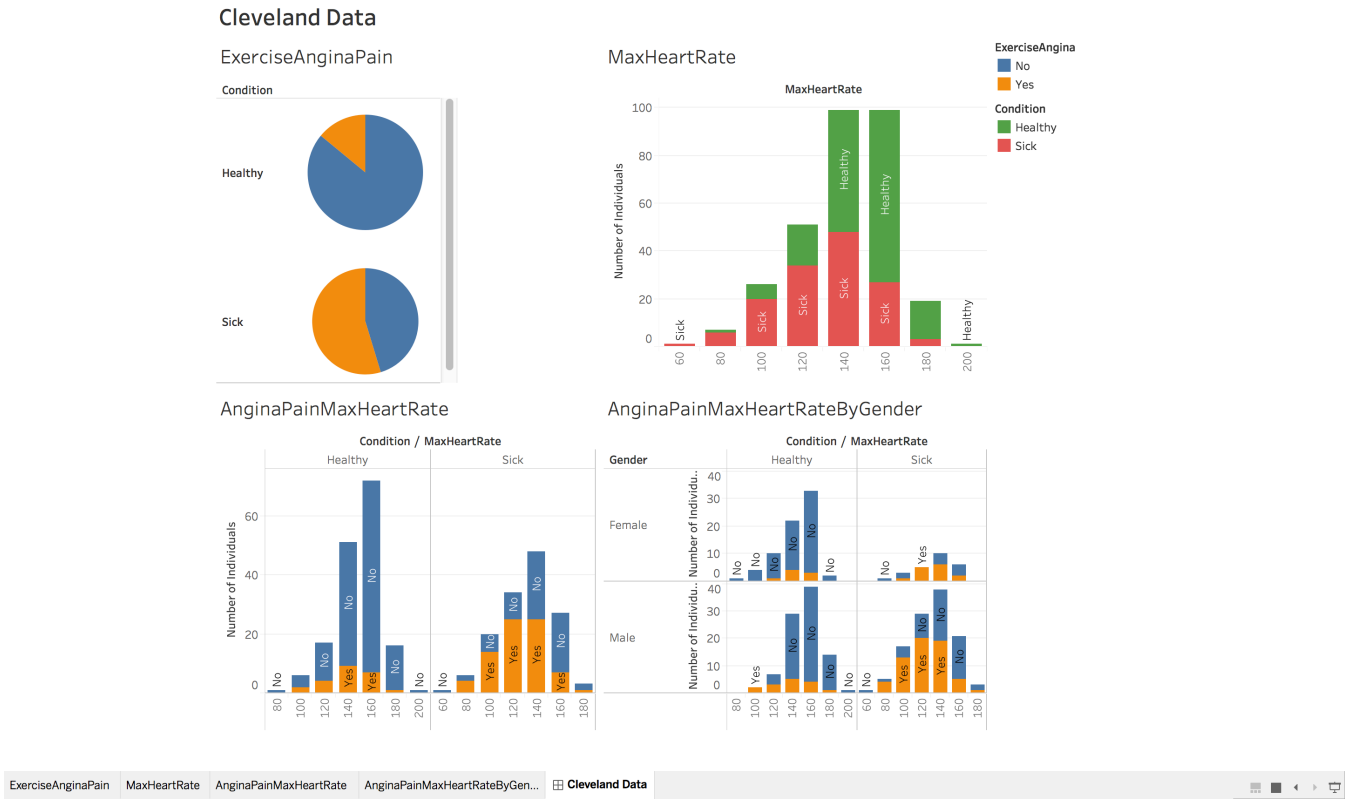
Plot AnginaPainMaxHeartRate

- Exercise induced angina pain could be a good indicator of sickness for max heart rate under 140; but a higher false positive rate should be expected for lower max heart rate values.
- The proportion of false negatives increases significantly for max heart rates above 140.

Plot AnginaPainMaxHeartRateByGender

- Males achieve a higher max heart rate in general.
- In a general sense, the observations made for AnginaPainMaxHeartRate plot apply to this plot, except that 140 max rate was a good threshold for the entire population but this plot shows separate threshold need to be used depending on the individual's gender.

DSDHT FALL 2017
Carlos Sathler (cssathler@gmail.com)



Conclusion – Cleveland Dataset

The analysis of the Cleveland dataset seems to confirm the common belief that the risk of heart disease is higher on individuals who experience heart pain during exercise. Additionally, the data shows evidence that healthy individuals tend to be able to endure more intense exercise. Our analysis therefore confirms commonly held beliefs about the correlation between heart disease and chest pain, as well as heart disease and ability to perform intense exercise. These conclusions apply to males and females alike.