Heart Disease Data Analysis - Cleveland

Shrey Patel
Data 607

Contents

- Introduction
    - Data set
    - Task
- Getting data ready for analysis
    - Handling missing data
    - Type conversion
- Exploratory analysis
- Uncertainties

Introduction

Cleveland data set (Processed)

- 14 attributes: age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, num
- The target attribute, num, takes 5 values (5 levels of prevalence), the higher the value, the higher the chance of heart disease.
- New variable, outcome, takes two values, 0 or 1, representing the absence or the presence of heart disease, respectively.

Task

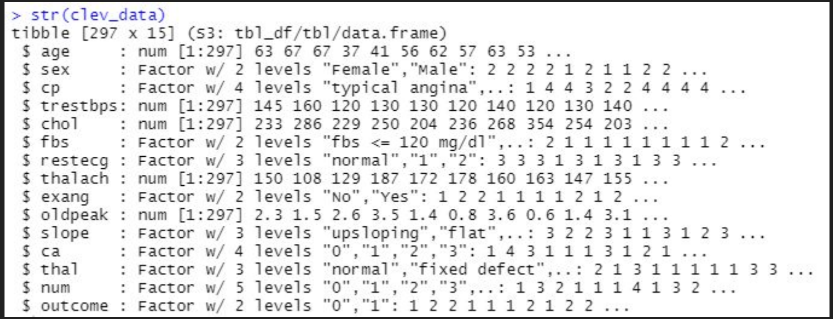
Using exploratory analysis, discovering and understanding the prevalence of heart diseases amongst females/males of different age groups in Cleveland. Moreover, how fasting blood sugar (fbs) levels may affect the chances of having a heart disease?

Getting data ready for analysis

Handling Missing Data:

- Four data points with missing data in ‘ca’, and two data points with missing data in ‘thal’
- Rows with missing data removed from the data set. [6/303 < 2% of the total observations]

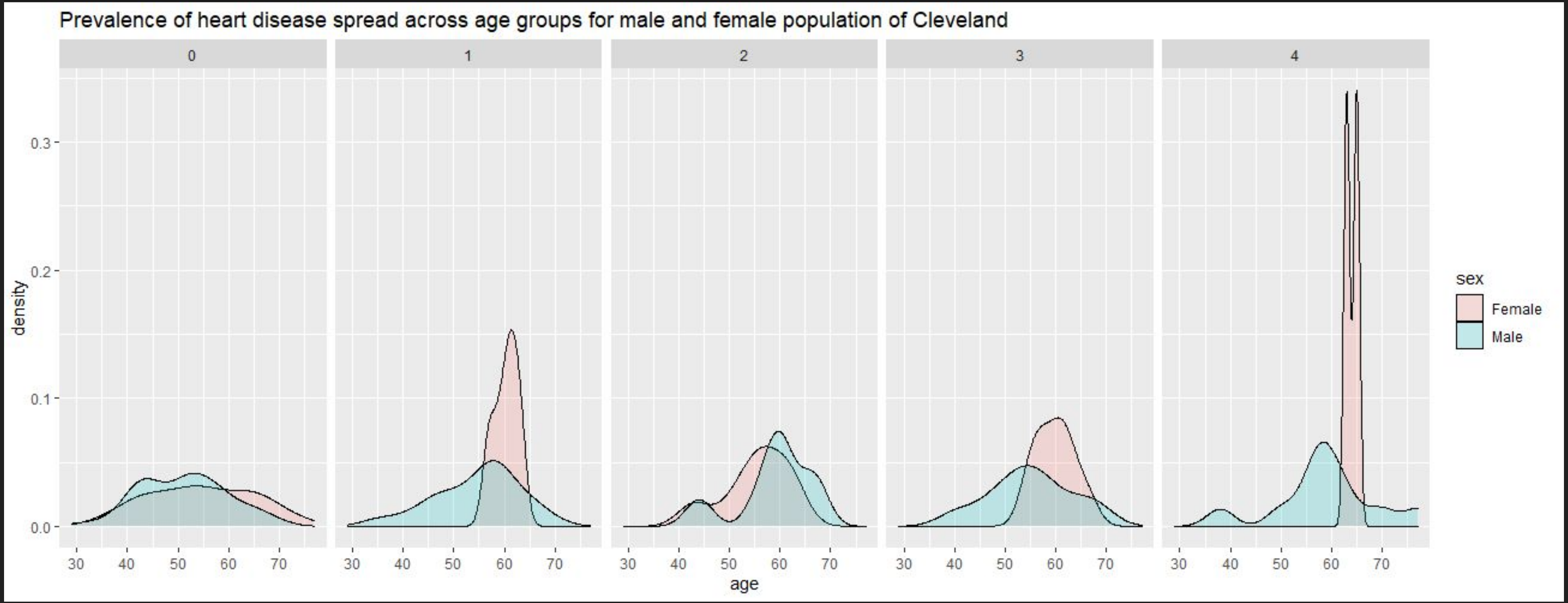
Type conversion: Converted the categorical variables from numeric to factors, with redefined levels.



```
> str(clev_data)
tibble [297 x 15] (S3: tbl_df/tbl/data.frame)
 $ age     : num [1:297] 63 67 67 37 41 56 62 57 63 53 ...
 $ sex     : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 2 1 1 2 2 ...
 $ cp      : Factor w/ 4 levels "typical angina",..: 1 4 4 3 2 2 4 4 4 4 ...
 $ trestbps: num [1:297] 145 160 120 130 130 120 140 120 130 140 ...
 $ chol    : num [1:297] 233 286 229 250 204 236 268 354 254 203 ...
 $ fbs     : Factor w/ 2 levels "fbs <= 120 mg/dl",..: 2 1 1 1 1 1 1 1 1 2 ...
 $ restecg : Factor w/ 3 levels "normal","1","2": 3 3 3 1 3 1 3 1 3 3 ...
 $ thalach : num [1:297] 150 108 129 187 172 178 160 163 147 155 ...
 $ exang   : Factor w/ 2 levels "No","Yes": 1 2 2 1 1 1 1 2 1 2 ...
 $ oldpeak : num [1:297] 2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
 $ slope   : Factor w/ 3 levels "upsloping","flat",..: 3 2 2 3 1 1 3 1 2 3 ...
 $ ca      : Factor w/ 4 levels "0","1","2","3": 1 4 3 1 1 1 3 1 2 1 ...
 $ thal    : Factor w/ 3 levels "normal","fixed defect",..: 2 1 3 1 1 1 1 1 3 3 ...
 $ num     : Factor w/ 5 levels "0","1","2","3",..: 1 3 2 1 1 1 4 1 3 2 ...
 $ outcome : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 2 1 2 2 ...
```

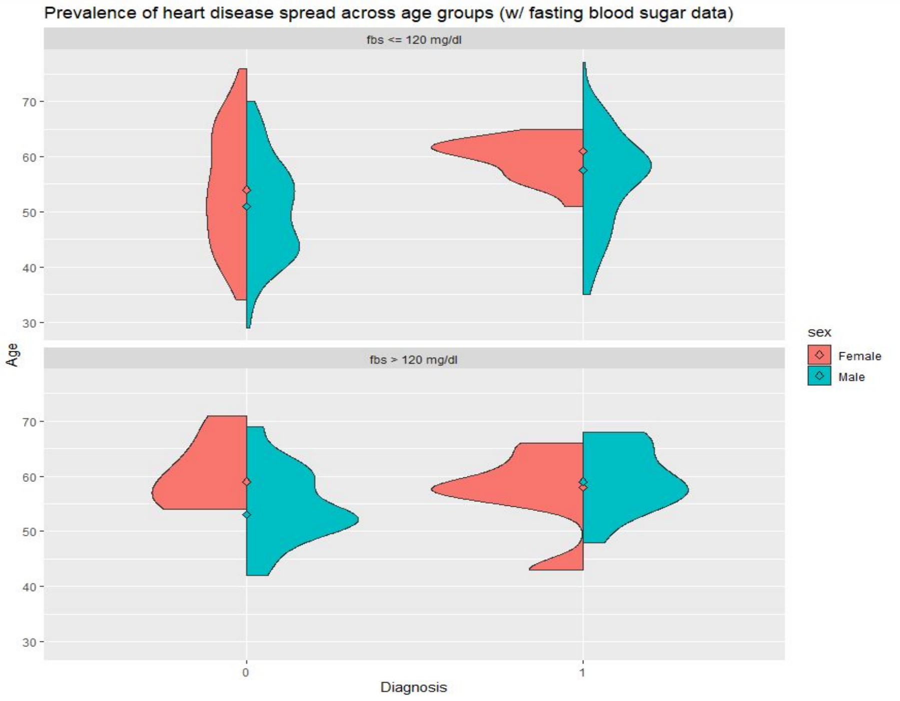
Exploratory Analysis

Chances of heart disease based on age and gender:



Exploratory Analysis

Chances of heart disease w.r.t age, gender, and fasting blood sugar (fbs) levels:



- Females b/w 55 and 65 have higher chance of being diagnosed with a heart disease, regardless of the fbs.
- Same goes for men b/w 55 and 65
- Regardless of the gender, people with abnormal fbs -> higher prevalence of a heart disease
- Healthy population is almost uniformly distributed between age of 45-65, for men, and 40-55 for women.
- There is one particular information this plot does **not** convey - discussed under uncertainties.

Uncertainties

However, our deductions accompany some empirical uncertainties.

- Are the readings reliable? *[Direct uncertainties]*
  - Outliers in the fasting blood sugar levels may denote the subject might have indeed ate something before health checkup.
  - Some of the variables like age, might have data-entry errors
  - Misdiagnosis by health professionals is also a factor that must be accounted for when discussing variables like cp (chest pain), as the medical procedure to diagnose chest pain usually consists of communication between the physician and the patient over a period of time (days/weeks).
- Is there enough data? *[Indirect uncertainty]*
  - The data set does not have any data points that represent females of age less than 43 with abnormal blood sugar levels.

Thank you!