

Final Project DS5110

Import Data

```
setwd("D:\\Northeastern\\Spring 2024\\DS5110\\DS5110---Heart-Disease-Analysis\\")

# UCI data
cleveland <- read.csv("uci\\processed.cleveland.data")
hungarian <- read.csv("uci\\processed.hungarian.data")
va <- read.csv("uci\\processed.va.data")
switzerland <- read.csv("uci\\processed.switzerland.data")
```

Clean Data

```
# tidy data
dataLists <- list(cleveland, hungarian, va, switzerland)
columnNames <- c("age", "sex", "cp", "trestbps", "chol", "fbs", "restecg",
                 "thalach", "exang", "oldpeak", "slope", "ca", "thal",
                 "have_heart_disease")

# Rename function
renameColumns <- function(df) {
  names(df) <- columnNames
  return(df)
}

cleveland <- renameColumns(cleveland)
hungarian <- renameColumns(hungarian)
va <- renameColumns(va)
switzerland <- renameColumns(switzerland)

# merge all data frame into 1
uci <- rbind(cleveland, hungarian, va, switzerland)

uci <- data.frame(uci)

# convert "?" into N/A and then remove N/A values
for (col_name in names(uci)) {
  uci[[col_name]][uci[[col_name]] == "?"] <- NA
}

uci <- uci %>%
  mutate(sex = case_when(sex == 0 ~ "female",
                        sex == 1 ~ "males")) %>%
  mutate(cp = case_when(cp == 1 ~ "typical angina",
                        cp == 2 ~ "atypical angina",
                        cp == 3 ~ "non-anginal pain",
                        cp == 4 ~ "asymptomatic")) %>%
```

```

mutate(fbs = case_when(fbs == "0" ~ "true",
                        fbs == "1" ~ "false")) %>%
mutate(restecg = case_when(restecg == 0 ~ "normal",
                           restecg == 1 ~ "ST-T wave abnormality",
                           restecg == 2 ~ "left ventricular hypertrophy")) %>%
mutate(exang = case_when(exang == 0 ~ "no",
                         exang == 1 ~ "yes")) %>%
mutate(slope = case_when(slope == "1" ~ "upsloping",
                         slope == "2" ~ "flat",
                         slope == "3" ~ "downsloping")) %>%
mutate(thal = case_when(thal %in% c("3.0") ~ "normal",
                        thal %in% c("6.0") ~ "fixed defect",
                        thal %in% c("7.0", "7") ~ "reversible defect")) %>%
mutate(have_heart_disease = case_when(have_heart_disease == 0 ~ "no",
                                       have_heart_disease %in% c(1, 2, 3, 4) ~
                                       "yes"))

# mutate to numeric
uci$trestbps <- as.numeric(uci$trestbps)
uci$chol <- as.numeric(uci$chol)
uci$thalach <- as.numeric(uci$thalach)
uci$oldpeak <- as.numeric(uci$oldpeak)
uci$ca <- as.numeric(uci$ca)

str(uci)

```

```

## 'data.frame':   916 obs. of  14 variables:
## $ age          : num  67 67 37 41 56 62 57 63 53 57 ...
## $ sex          : chr   "males" "males" "males" "female" ...
## $ cp          : chr   "asymptomatic" "asymptomatic" "non-anginal pain" "atypical angina" ...
## $ trestbps     : num   160 120 130 130 120 140 120 130 140 140 ...
## $ chol         : num   286 229 250 204 236 268 354 254 203 192 ...
## $ fbs         : chr   "true" "true" "true" "true" ...
## $ restecg     : chr   "left ventricular hypertrophy" "left ventricular hypertrophy" "normal" ...
## $ thalach     : num   108 129 187 172 178 160 163 147 155 148 ...
## $ exang       : chr   "yes" "yes" "no" "no" ...
## $ oldpeak     : num    1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 0.4 ...
## $ slope       : chr   "flat" "flat" "downsloping" "upsloping" ...
## $ ca         : num    3 2 0 0 0 2 0 1 0 0 ...
## $ thal       : chr   "normal" "reversible defect" "normal" "normal" ...
## $ have_heart_disease: chr   "yes" "yes" "no" "no" ...

```

```
unique(uci$chol)
```

```

## [1] 286 229 250 204 236 268 354 254 203 192 294 256 263 199 168 239 275 266
## [19] 211 283 284 224 206 219 340 226 247 167 230 335 234 233 177 276 353 243
## [37] 225 302 212 330 175 417 197 198 290 253 172 273 213 305 216 304 188 282
## [55] 185 232 326 231 269 267 248 360 258 308 245 270 208 264 321 274 325 235
## [73] 257 164 141 252 255 201 222 260 182 303 265 309 307 249 186 341 183 407
## [91] 217 288 220 209 227 261 174 281 221 205 240 289 318 298 564 246 322 299
## [109] 300 293 277 214 207 223 160 394 184 315 409 244 195 196 126 313 259 200
## [127] 262 215 228 193 271 210 327 149 295 306 178 237 218 242 319 166 180 311
## [145] 278 342 169 187 157 176 241 131 NA 161 173 194 297 292 339 147 291 358
## [163] 412 238 163 280 202 328 129 190 179 272 100 468 320 312 171 365 344 85
## [181] 347 251 287 156 117 466 338 529 392 329 355 603 404 518 285 279 388 336

```

```
## [199] 491 331 393    0 153 316 458 384 349 142 181 310 170 369 165 337 333 139
## [217] 385
```

```
# filter the NA values
for (col in names(uci)) {
  if (is.numeric(uci[[col]])) {
    mean_val <- round(mean(uci[[col]], na.rm = TRUE))
    uci[[col]][is.na(uci[[col]])] <- mean_val
  } else {
    mode_val <- names(sort(table(uci[[col]]), decreasing = TRUE))[1]
    uci[[col]][is.na(uci[[col]])] <- mode_val
  }
}
```

```
# get unique values
uniqueValues <- sapply(uci, unique)
uniqueValues
```

```
## $age
## [1] 67 37 41 56 62 57 63 53 44 52 48 54 49 64 58 60 50 66 43 40 69 59 42 55 61
## [26] 65 71 51 46 45 39 68 47 34 35 29 70 77 38 74 76 30 31 32 33 36 72 75 73
##
## $sex
## [1] "males" "female"
##
## $cp
## [1] "asymptomatic" "non-anginal pain" "atypical angina" "typical angina"
##
## $trestbps
## [1] 160 120 130 140 172 150 110 132 117 135 112 105 124 125 142 128 145 170 155
## [20] 104 180 138 108 134 122 115 118 100 200 94 165 102 152 101 126 174 148 178
## [39] 158 192 129 144 123 136 146 106 156 154 114 164 98 190 113 92 116 0 96
## [58] 127 95 80 185
##
## $chol
## [1] 286 229 250 204 236 268 354 254 203 192 294 256 263 199 168 239 275 266
## [19] 211 283 284 224 206 219 340 226 247 167 230 335 234 233 177 276 353 243
## [37] 225 302 212 330 175 417 197 198 290 253 172 273 213 305 216 304 188 282
## [55] 185 232 326 231 269 267 248 360 258 308 245 270 208 264 321 274 325 235
## [73] 257 164 141 252 255 201 222 260 182 303 265 309 307 249 186 341 183 407
## [91] 217 288 220 209 227 261 174 281 221 205 240 289 318 298 564 246 322 299
## [109] 300 293 277 214 207 223 160 394 184 315 409 244 195 196 126 313 259 200
## [127] 262 215 228 193 271 210 327 149 295 306 178 237 218 242 319 166 180 311
## [145] 278 342 169 187 157 176 241 131 161 173 194 297 292 339 147 291 358 412
## [163] 238 163 280 202 328 129 190 179 272 100 468 320 312 171 365 344 85 347
## [181] 251 287 156 117 466 338 529 392 329 355 603 404 518 285 279 388 336 491
## [199] 331 393 0 153 316 458 384 349 142 181 310 170 369 165 337 333 139 385
##
## $fbs
## [1] "true" "false"
##
## $restecg
## [1] "left ventricular hypertrophy" "normal"
## [3] "ST-T wave abnormality"
##
```

```
## $thalach
## [1] 108 129 187 172 178 160 163 147 155 148 153 142 173 162 174 168 139 171
## [19] 144 132 158 114 151 161 179 120 112 137 157 169 165 123 128 152 140 188
## [37] 109 125 131 170 113 99 177 141 180 111 143 182 150 156 115 149 145 146
## [55] 175 186 185 159 130 190 136 97 127 154 133 126 202 103 166 164 184 124
## [73] 122 96 138 88 105 194 195 106 167 95 192 117 121 116 71 118 181 134
## [91] 90 98 176 135 110 100 87 102 92 91 82 119 94 86 84 80 107 69
## [109] 73 93 104 60 83 63 70 77 72 78 67
##
## $exang
## [1] "yes" "no"
##
## $oldpeak
## [1] 1.5 2.6 3.5 1.4 0.8 3.6 0.6 3.1 0.4 1.3 0.0 0.5 1.6 1.0 1.2
## [16] 0.2 1.8 3.2 2.4 2.0 2.5 2.2 2.8 3.0 3.4 6.2 4.0 5.6 2.9 0.1
## [31] 2.1 1.9 4.2 0.9 1.1 3.8 0.7 0.3 2.3 4.4 5.0 -0.5 1.7 -1.1 -1.5
## [46] -0.1 -2.6 -0.7 -2.0 -1.0 -0.8 -0.9 3.7
##
## $slope
## [1] "flat" "downsloping" "upsloping"
##
## $ca
## [1] 3 2 0 1
##
## $thal
## [1] "normal" "reversible defect" "fixed defect"
##
## $have_heart_disease
## [1] "yes" "no"
```

Export to CSV

```
folder_path <- "./cleaned-data/" # Change this to your desired folder path

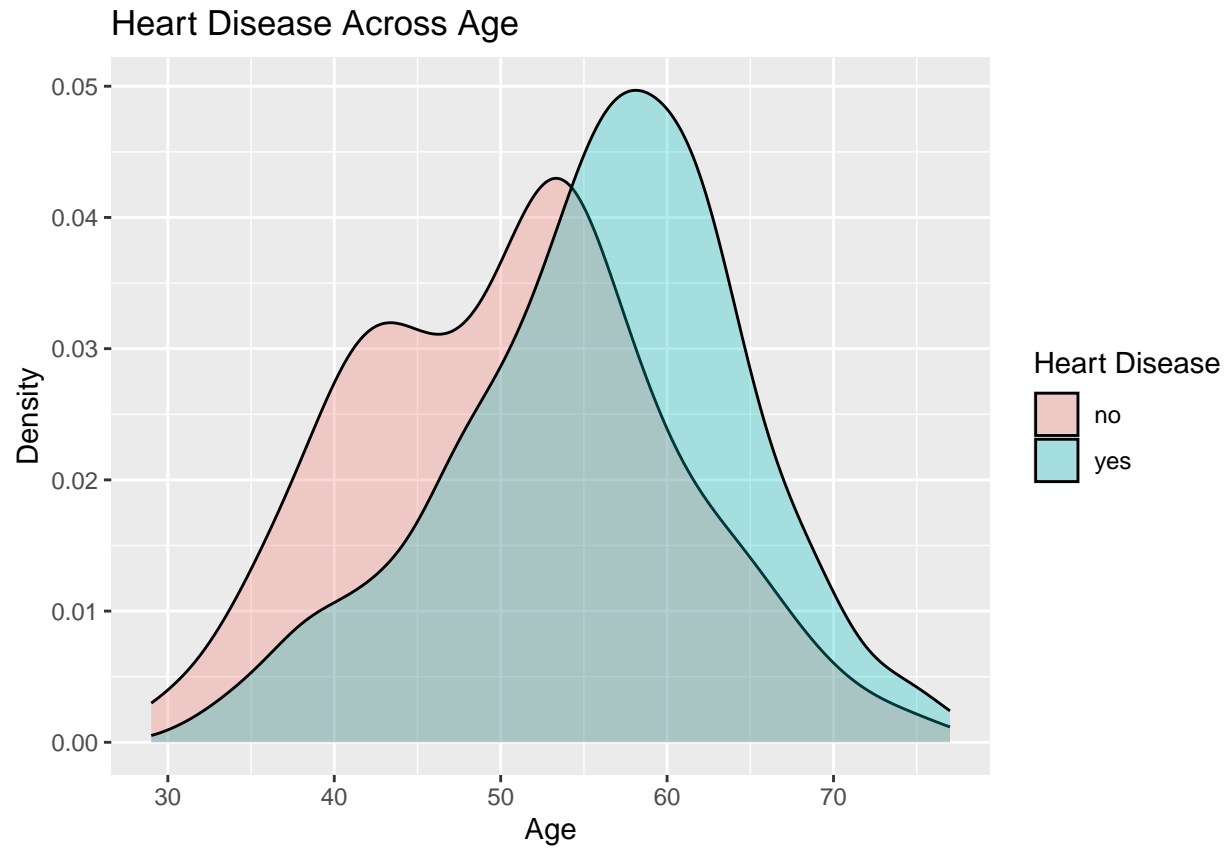
# Create the folder if it doesn't already exist
if (!dir.exists(folder_path)) {
  dir.create(folder_path)
}

file_path <- file.path(folder_path, "cleaned-uci.csv")

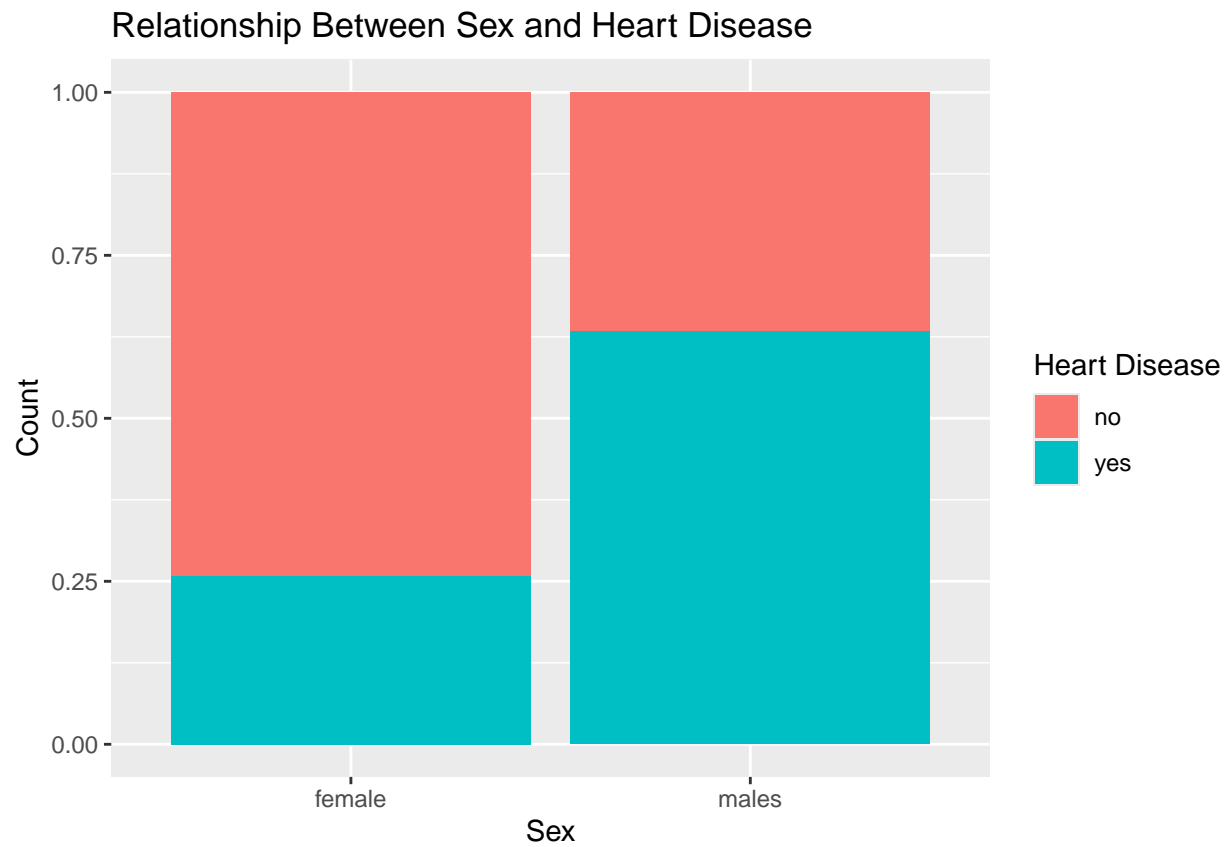
# Export the dataframe to a CSV file
write.csv(uci, file = file_path, row.names = FALSE)
```

Visualization

```
ggplot(uci, aes(x = age, fill = as.factor(have_heart_disease))) +
  geom_density(alpha = 0.3) +
  labs(x = "Age", y = "Density", fill = "Heart Disease") +
  ggtitle("Heart Disease Across Age")
```



```
ggplot(uci, aes(x = sex, fill = as.factor(have_heart_disease))) +  
  geom_bar(position = "fill") +  
  labs(x = "Sex", y = "Count", fill = "Heart Disease") +  
  ggtitle("Relationship Between Sex and Heart Disease")
```



```
ggplot(uci, aes(x = cp, fill = as.factor(have_heart_disease))) +  
  geom_bar(position = "dodge") +  
  labs(x = "Chest Pain Type", y = "Count", fill = "Heart Disease") +  
  ggtitle("Relationship Between Chest Pain Type and Heart Disease")
```

