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 <- na.omit(uci)

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") ~ "reversable 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':    298 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" "reversable defect" "normal" "normal" ...
##  $ have_heart_disease: chr  "yes" "yes" "no" "no" ...
##  - attr(*, "na.action")= 'omit' Named int [1:618] 87 166 192 266 287 302 303 304 305 306 ...
##   ..- attr(*, "names")= chr [1:618] "87" "166" "192" "266" ...
```

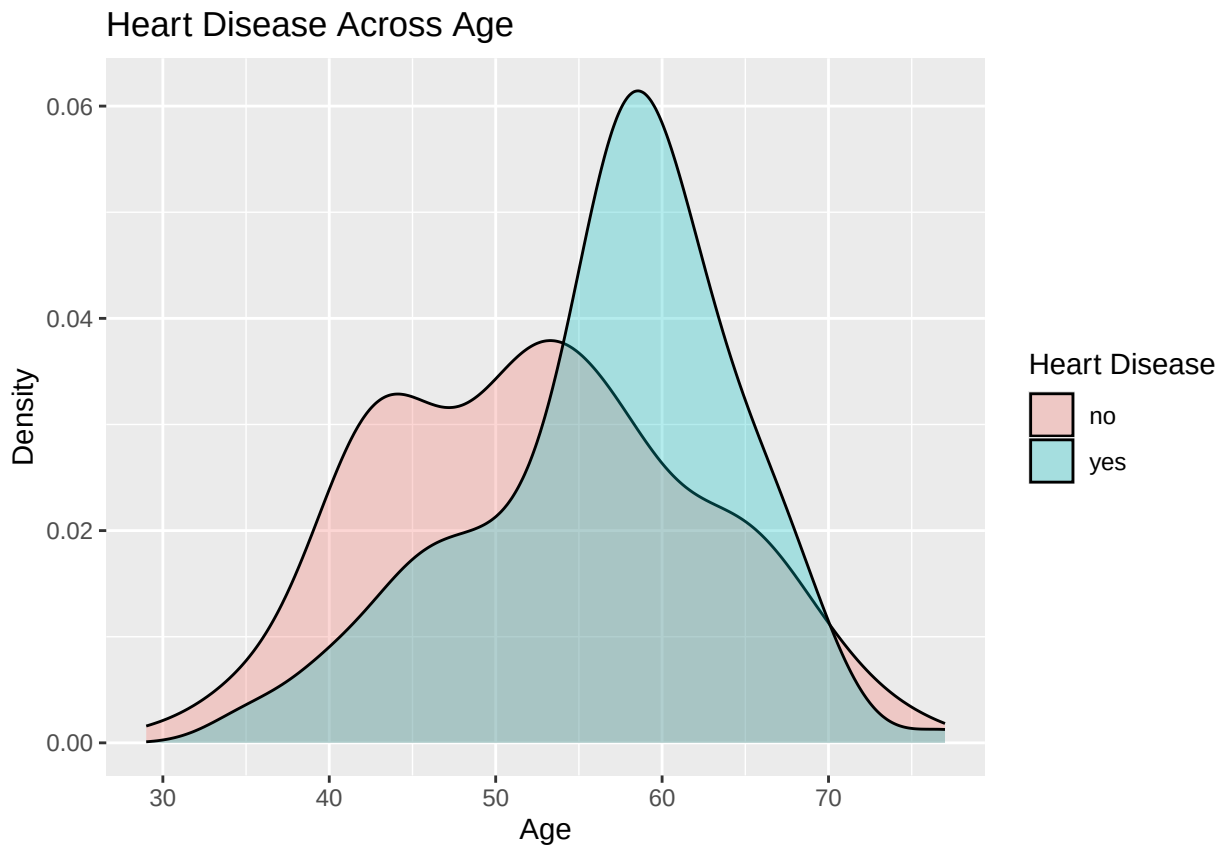
Export to CSV

```
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)
}
```

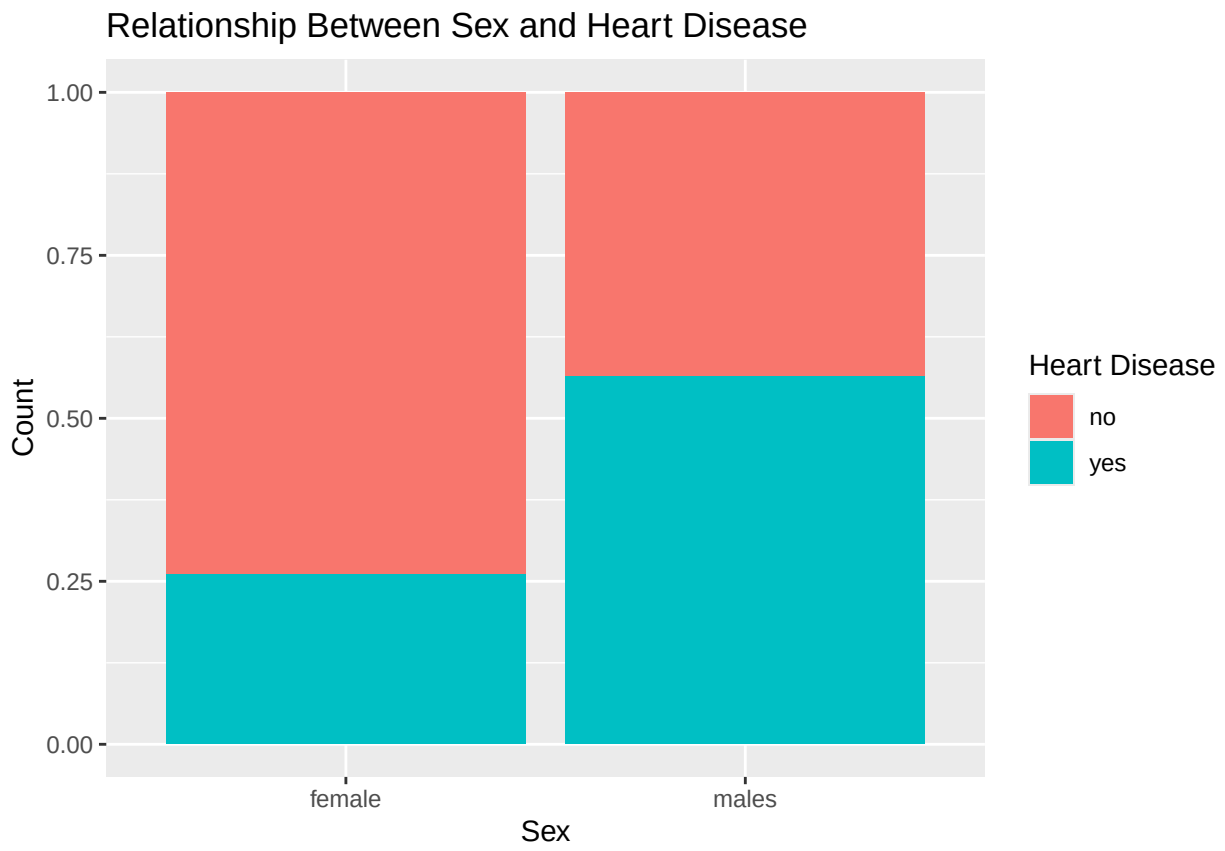
```r
file_path <- file.path(folder_path, "cleaned-uci.csv")

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

Visualization

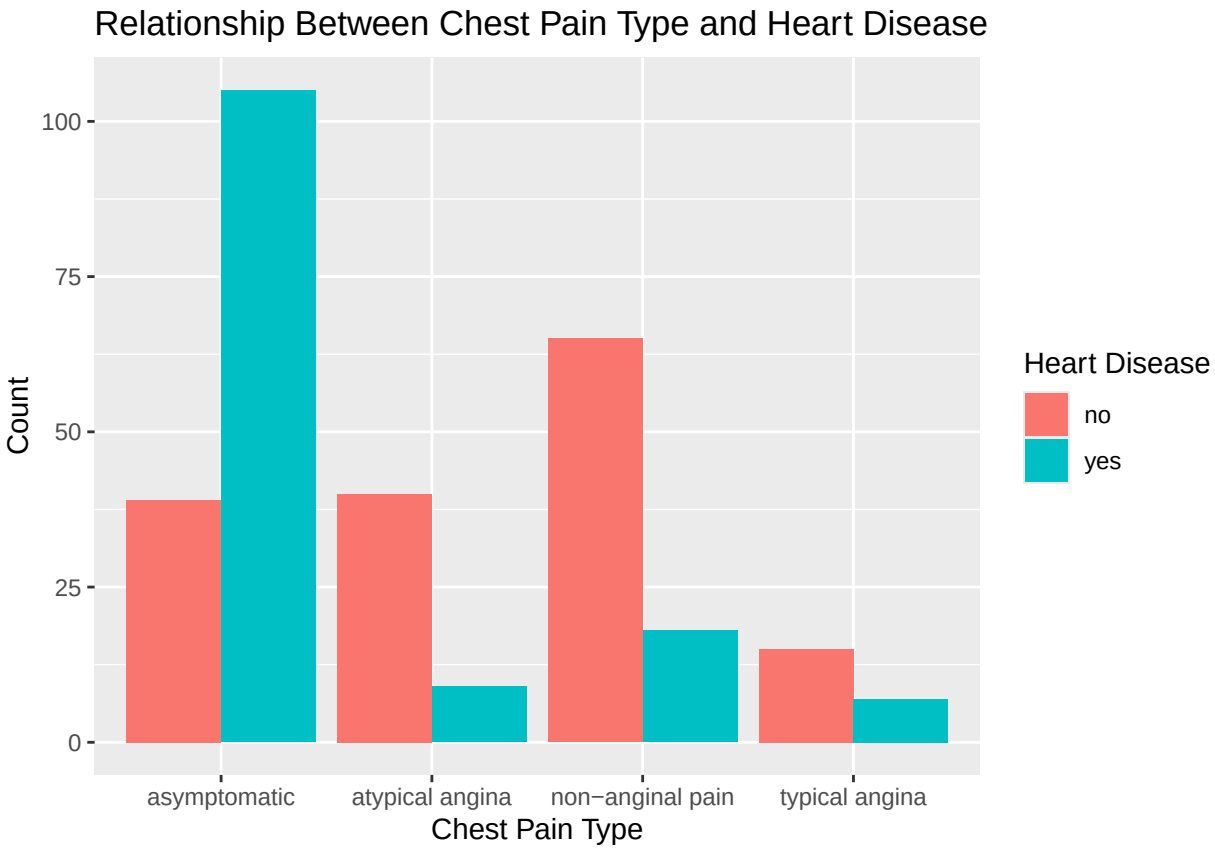
```r
ggplot(uci, aes(x = age, fill = as.factor(have_heart_disease))) +
  geom_density(alpha = 0.3) +
  labs(x = "Age", y = "Density", fill = "Heart Disease") +
  ggtitle("Heart Disease Across Age")
```




```r
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")
```


Relationship Between Chest Pain Type and Heart Disease
Count
100
75
50
25
0
asymptomatic
atypical angina
non−anginal pain
typical angina
Chest Pain Type
Heart Disease
no
yes