Data analysis with R (PUHR11103) - Assignment

B248593

Word count: 2473 words

Data input

First the needed libraries are load into R / Rstudio with the `p_load` method found within the `pacman` package (installed previously), which installs packages first if they are not already installed. The Assignment instructions paper is loaded, as it contains the codebook section, which is needed to understand the dataset struture.

The data is accessable through the Learn portal of the University of Edingburgh. The dataset itself is named after a town in Massachusetts USA, which was founded in the 17th century (Framingham History Center, 2024) and had a population of around 72,000 people by 2020 with a median age of 39.1 years. Most of them have a health insurance, only 5,3% of the population are uninsured (United State Census Bureau, 2024). The city is located on the east coast of the US, and is part of the Greater Boston area and surrounded by some lakes as the map shows.

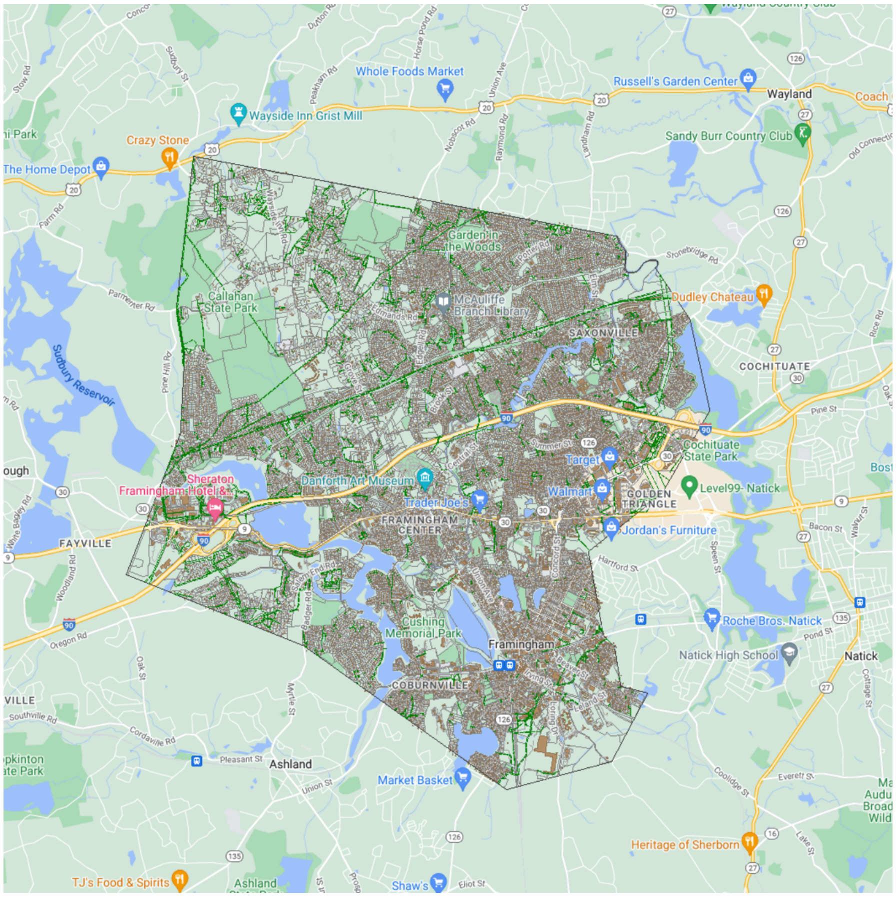


Figure 1: Framingham, Massachusetts (Source: MapGeo)

To open the dataset `framingham_2024.csv` for import into RStudio it is read from the `raw_data` directory using the `pacman::p_load` method which installs a missing package if it is missing.

```r
library(pacman)
pacman::p_load(tidyverse, ggplot2, epitools, finalfit, broom, gt, here)

# Opening assessment instructions with Codebook section
browseURL(here("meta_data", "Assignment_instructions_2024.pdf"))

# Printing out questions / answers about the assignment
readLines(here("meta_data", "Questions.txt"))
```

```r
# Listing files in `raw_data` directory
list.files("./raw_data/")

# Importing the dataset
framingham <- read_csv(here("raw_data", "framingham_2024.csv"))
```

The age distribution of the population (Link: Data Census) of Framingham shows not a pyramid shape, as the two largest age groups are in the 35 - 39 and 40 - 44 for men and 30 - 34 and 35 - 39 for women. National Heart, Lung, and Blood Institute (2024) describes the used data from the Framingham Heart Study (FHS). They state, it contains people from 3 different generations (original participants, children of them and their grandchildren). The study started in 1948 and is still ongoing and shows risk factors for coronary heart disease (CHD).

Data checking

The data set has a total of observations (rows) 4434 and 15 variables (columns). As the `str` method shows, all variables are stored as numerical values of the type `double`.

```r
# Checking the dataset
View(framingham)
dim(framingham)
str(framingham)

# Showing total columns count
ncol(framingham)

# Showing total rows count
nrow(framingham)

# Showing columns with missing values
missing <- colSums(is.na(framingham))

# Just showing columns with missing values
col_missing <- missing[missing > 0]

# Writing column names with value in brackets
col_missing_names <- paste0(names(col_missing), " (", col_missing, ")")
col_comma_separated <- paste(col_missing_names, collapse = ", ")
```

Reading the codebook section of the assignment instructions, the following variables are included in the dataset:

- **sex** -> *Sex*: **1 = male, 2 = female**
- **age** -> *Age* of the participant at exam (**years**)
- **currentSmoker** *Current smoker*: **1 = Yes, 0 = No**
- **cigsPerDay** -> *Number* of cigarettes that the respondent smoked on average in one day
- **BPMeds** -> Use of *Anti-hypertensive medication* at exam: **1 = Yes, 0 = No**
- **prevalentStroke** -> Respondent had a *stroke* previously: **1 = Yes, 0 = No**
- **prevalentHyp** -> *Prevalent Hypertensive*. Subject was defined as hypertensive if treated or if second exam at which mean systolic was >=140 mmHg or mean Diastolic >=90 mmHg: **1 = Yes, 0 = No**
- **diabetes** -> Diabetic according to criteria of first exam treated or first exam with casual glucose of 200 mg/dL or more: **1 = Yes, 0 = No**
- **totChol** -> *Serum total cholesterol* level (**mg/dL**)
- **sysBP** -> *Systolic blood pressure* (mean of last two of three measurements) (**mm Hg**)

- **diaBP** -> *Diastolic blood pressure* (mean of last two of three measurements) (**mm Hg**)
- **BMI** -> *BMI* (**kg/m^2**)
- **heartRate** -> *Heart rate* (Ventricular rate) (**BPM**)
- **glucose** -> *Casual serum glucose* level (**mg/dL**)
- **anyCHD** -> *Outcome: Occurrence of any Coronary Heart Disease* (Angina Pectoris, Myocardial infarction (Hospitalized and silent or unrecognized), Coronary Insufficiency (Unstable Angina), or Fatal Coronary Heart Disease: **1 = Yes, 0 = No**

The columns (variables) cigsPerDay (32), BPMeds (61), totChol (52), BMI (19), heartRate (1), glucose (397) show missing values, the amount of missing values are listed in the brackets.

Data preparation

As the codebook section in the instruction paper mentioned the raw data needs to be recoded. The variables `sex`, `currentSmoker`, `BPMeds`, `prevalentHyp`, `diabetes` and `anyCHD` are transformed to factors. The levels of the `sex` variable are changed to `male` and `female`, the levels of all other variables are changed to `No` and `Yes`.

```r
# Transforming 7 categorical variables into factors
framingham_df <- framingham %>%
  mutate(
    sex = factor(sex, labels = c("male", "female")),
    currentSmoker = factor(currentSmoker, labels = c("No", "Yes")),
    BPMeds = factor(BPMeds, labels = c("No", "Yes")),
    prevalentStroke = factor(prevalentStroke, labels = c("No", "Yes")),
    prevalentHyp = factor(prevalentHyp, labels = c("No", "Yes")),
    diabetes = factor(diabetes, labels = c("No", "Yes")),
    anyCHD = factor(anyCHD, labels = c("No", "Yes"))
  )

# Viewing the dataframe
View(framingham_df)

# Cleaining up the environment
rm(framingham, missing, col_missing, col_missing_names)
```

No further processing is needed. The processed data is stored in the variable `framingham_df`.

Exploratory Data Analysis

1.

a.

First I created a contingency table to see the distribution of the `anyCHD` variable. The distribution of the dependent variable (anyCHD) is shown in the left column of the table. The smoking status (currentSmoker), which is the independent variable, is shown in the top row of the table. The table shows the number of participants in each combination of the two variables.

A barplot will be used to visualize the distribution of the `currentSmoker` variable in relation to the occurrence of any Coronary Heart Disease as it displays the frequency distribution of this categorical variable.

```r
# Contingency table
framingham_df %>%
  select(anyCHD, currentSmoker) %>%
  table()
```

```
##       currentSmoker
## anyCHD   No  Yes
##    No  1630 1564
##    Yes  623  617
```

Both groups (smoking and non smoking) are nearly the same size.

b.

To observe the relationship of age (2 categories) and anyCHD are shown in a contingency table. The occurence is shown by column of the age (numerical) where the first row shows the not existing CHD, the second exisiting CHD.

A barplot will be used to visualize the distribution of the `currentSmoker` variable in relation to the occurrence of any Coronary Heart Disease as it displays the frequency distribution of this categorical variable.

```r
# Creating crosstab
crosstab_age <- framingham_df %>%
  select(age, anyCHD)

# Contingency table
crosstab_age %>%
  table()

# Viewing crosstab
View(crosstab_age)
```

```r
# Range of age
crosstab_age %>%
  select(age) %>%
  range()
```

```
## [1] 32 70
```

```r
# Mean and Median of the age variable
crosstab_age %>%
  summarise(
    age.mean = mean(age),
    age.median = median(age)
  )
```

```
## # A tibble: 1 x 2
##   age.mean age.median
##      <dbl>      <dbl>
## 1     49.9         49
```

```r
# Summarising of age in relation to anyCHD
crosstab_age %>%
  summary_factorlist(
    dependent = "anyCHD", explanatory = "age",
    column = FALSE, total_col = TRUE
  )
```

```
##  label    levels         No        Yes      Total
##    age Mean (SD) 48.9 (8.5) 52.7 (8.5) 49.9 (8.7)
```

The summarisation shows that the mean age of individuals with any coronary heart disease is 52.7, which is higher than the mean age of individuals without any coronary heart disease (48.9). This suggests that individuals with any coronary heart disease are generally older than those without any coronary heart disease.

```
# Recoding age into 2 level factor (40-59, 60-79)
crosstab_age_cat <- framingham_df %>%
  filter(age >= 40) %>%
  mutate(age_cat = cut(age,
    breaks = c(39, 59, 79),
    labels = c("40-59", "60-79")
  )) %>%
  select(anyCHD, age_cat)

# Contingency table
crosstab_age_cat %>%
  table()
```

```
##        age_cat
## anyCHD 40-59 60-79
##    No   2274   452
##    Yes   817   332
```

```
# Viewing crosstab
View(crosstab_age_cat)
```

That created the `crosstab_age_cat` table containing the age in two category groups.

2.

a.

Checking the data for a relation of systolic blood pressure (sysBP) and smoking.

```
# Creating crosstab
crosstab_sysBP <- framingham_df %>%
  select(currentSmoker, sysBP)

# Contingency table
crosstab_sysBP %>%
  table()

# Viewing crosstab
View(crosstab_sysBP)
```

```
# Range of sysBP
crosstab_sysBP %>%
  select(sysBP) %>%
  range()
```

```
## [1]  83.5 295.0
```

```
# Mean and Median of the sysBP variable
crosstab_sysBP %>%
  select(sysBP) %>%
  summarise(
    sysBP.mean = mean(sysBP),
    sysBP.median = median(sysBP)
  )
```

```
## # A tibble: 1 x 2
##   sysBP.mean sysBP.median
##        <dbl>        <dbl>
## 1       133.          129
```

```
# Summarising of smoking status in relation to sysBP
crosstab_sysBP %>%
  summary_factorlist(
    dependent = "sysBP", explanatory = "currentSmoker",
    column = FALSE, total_col = TRUE
  )
```

```
##          label levels      unit        value       Total
##   currentSmoker     No Mean (sd) 135.9 (23.5) 2253 (50.8)
##                    Yes Mean (sd) 129.8 (20.9) 2181 (49.2)
```

The range of the systolic blood pressure is between 83 and 295 mm Hg. The mean 133 mm Hg, and the median 129 mm Hg. For smokers the mean is 130 mm Hg and for non-smokers 136 mm Hg.

b.

Investigating relations between prevalence of Hypertension and smoking.

```
# Creating crosstab
crosstab_preHyp <- framingham_df %>%
  select(currentSmoker, prevalentHyp)

# Contingency table
crosstab_preHyp %>%
  table()

# Viewing crosstab
View(crosstab_preHyp)
```

```
# Summarising of smoking status in relation to sysBP
crosstab_preHyp %>%
  summary_factorlist(
    dependent = "prevalentHyp", explanatory = "currentSmoker",
    column = FALSE, total_col = TRUE
  )
```

```
##          label levels          No        Yes      Total
##   currentSmoker     No 1415 (62.8) 838 (37.2) 2253 (100)
##                    Yes 1589 (72.9) 592 (27.1) 2181 (100)
```

The summary statistics show that the prevalence of hypertension is higher in people who do not smoke (37.2%) than in those who do (27.1%).

3.

As an automated approach to create the tables which is utilized, and no manual data checking was conducted, the `drop_na` method is used to remove missing values from the dataset.

Developing the analysis plan

1.

a.

I plan to answer the question about the risk of heart disease with a riskratio test. The null hypothesis is that the risk of heart disease (binary outcome of “Yes” or “No”) is the same for smokers and non-smokers, which is a binary / categorical variable. The alternative hypothesis is that the risk of heart disease is different for smokers and non-smokers. I will use a significance level of 0.05.

b.

For answering the question of association of CHD and age a wilcox test will be used. The null hypothesis is that the risk of heart disease is not dependent on age. Therefore the alternative hypothesis is it appears differently with changing age groups. Again a significance level of 0.05 will be used.

2.

a.

In this case several tests will be used. The null hypothesis is that the mean systolic blood pressure is the same for smokers and non-smokers. The alternative hypothesis is that the mean systolic blood pressure is different for smokers and non-smokers. A significance level of 0.05 is used.

b.

A Fisher excat test will be used to compare the prevalence of hypertension between smokers and non-smokers. The null hypothesis is that the prevalence of hypertension is the same for smokers and non-smokers. The alternative hypothesis is that the prevalence of hypertension is different for smokers and non-smokers. The same significance level as for the other analyses is used (0.05).

3

As the amount of data categories is complex, a automated approach was chosen to create the tables which shows associations with a significance level of 0.01. The null hypothesis is always that the risk factor is not associated with the occurrence of CHD. The alternative hypothesis is that it is.

Investigating the assumptions

1.

a.

In this case, the assumption of normal distribution is not relevant, as it is the distribution of a discrete variable. The riskratio test will be used.

```r
# Checking assumption
subset_framingham <- framingham_df %>%
  drop_na(currentSmoker) %>%
  sample_n(1000)

# Distribution of currentSmoker
p0 <- subset_framingham %>%
  ggplot(aes(x = currentSmoker)) +
  geom_bar(aes(y = ..prop.., fill = anyCHD)) +
  labs(title = "Figure A1: Current Smoker",
       subtitle = "in relation to the occurrence of any Coronary Heart Disease",
       x = "Current Smoker", y = "Proportion",
       caption = "Data source: Framingham Heart Study")
p0
```

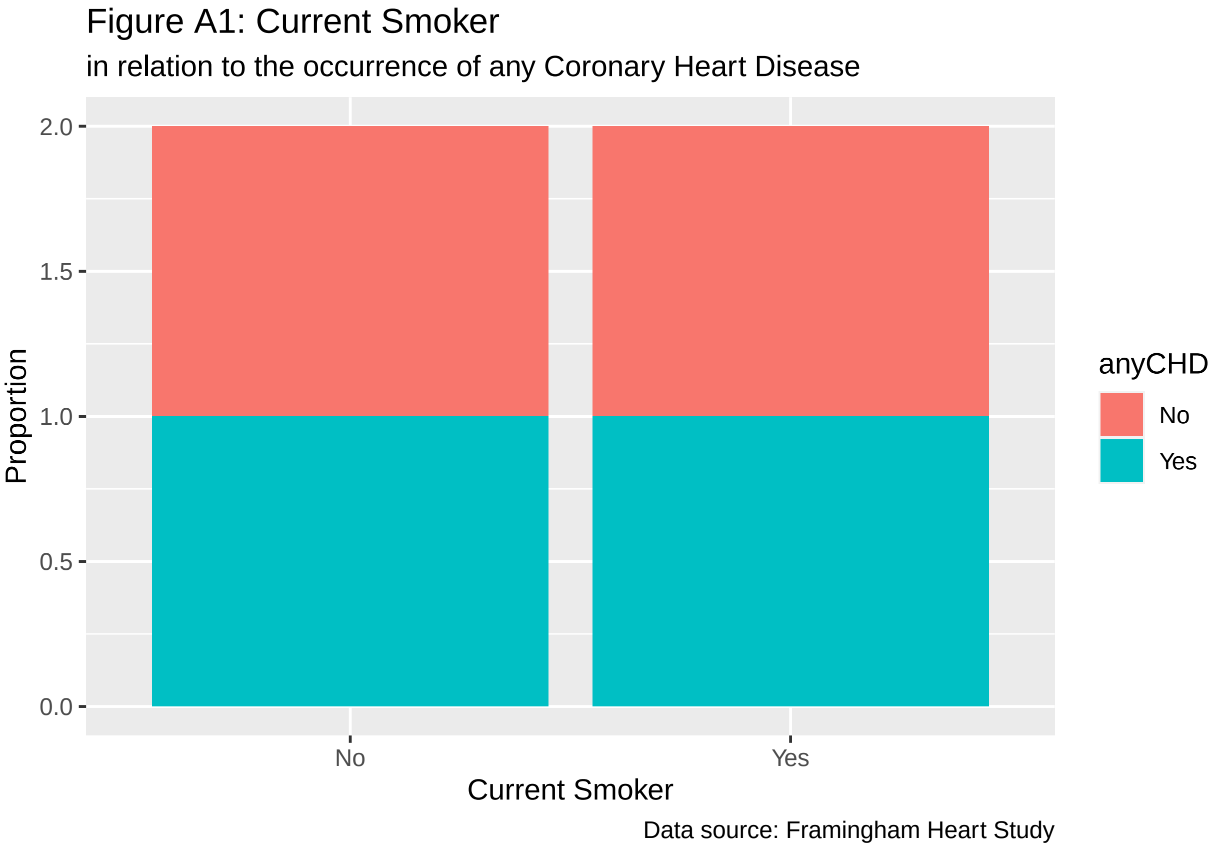


Figure A1: Current Smoker

in relation to the occurrence of any Coronary Heart Disease

Data source: Framingham Heart Study

```
# Saving the plot
ggsave(p0, file = here("figures/", "Plot_A1.png"))

# Cleaning up the environment
rm(subset_framingham)
```

b.

The assumption of normal distribution for age will be tested with a histogram and a Shapiro-Wilk test.

```
# Creating sample dataset
subset_age <-
  crosstab_age %>%
  drop_na(age) %>%
  sample_n(1000)

# Checking distribution of age (also for normality)
p1 <- subset_age %>%
  ggplot(aes(x = age)) +
  geom_histogram(aes(y = after_stat(density))) +
  geom_density(aes(y = after_stat(density))) +
  labs(title = "Figure A2: Age",
       subtitle = "Distribution",
       x = "Age", y = "Density",
       caption = "Data source: Framingham Heart Study")
p1
```

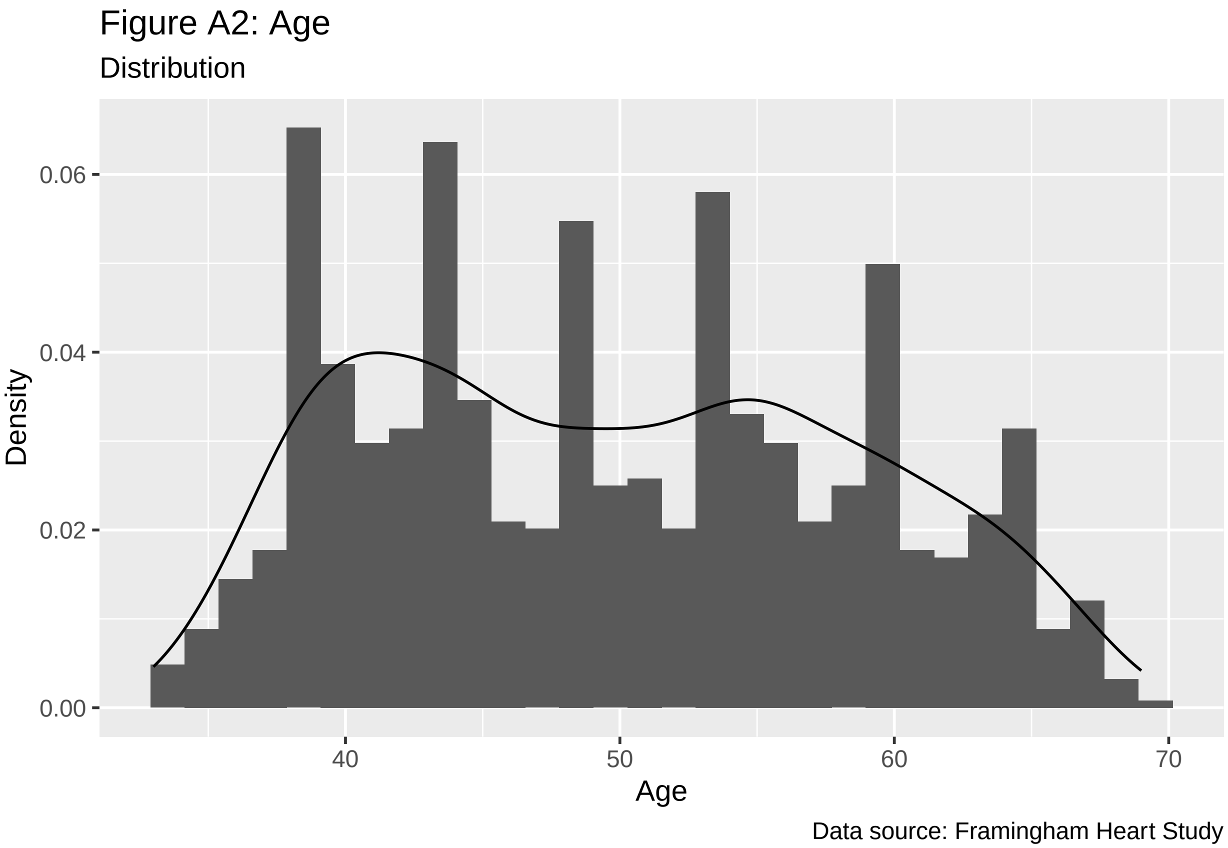


Figure A2: Age

Distribution

Data source: Framingham Heart Study

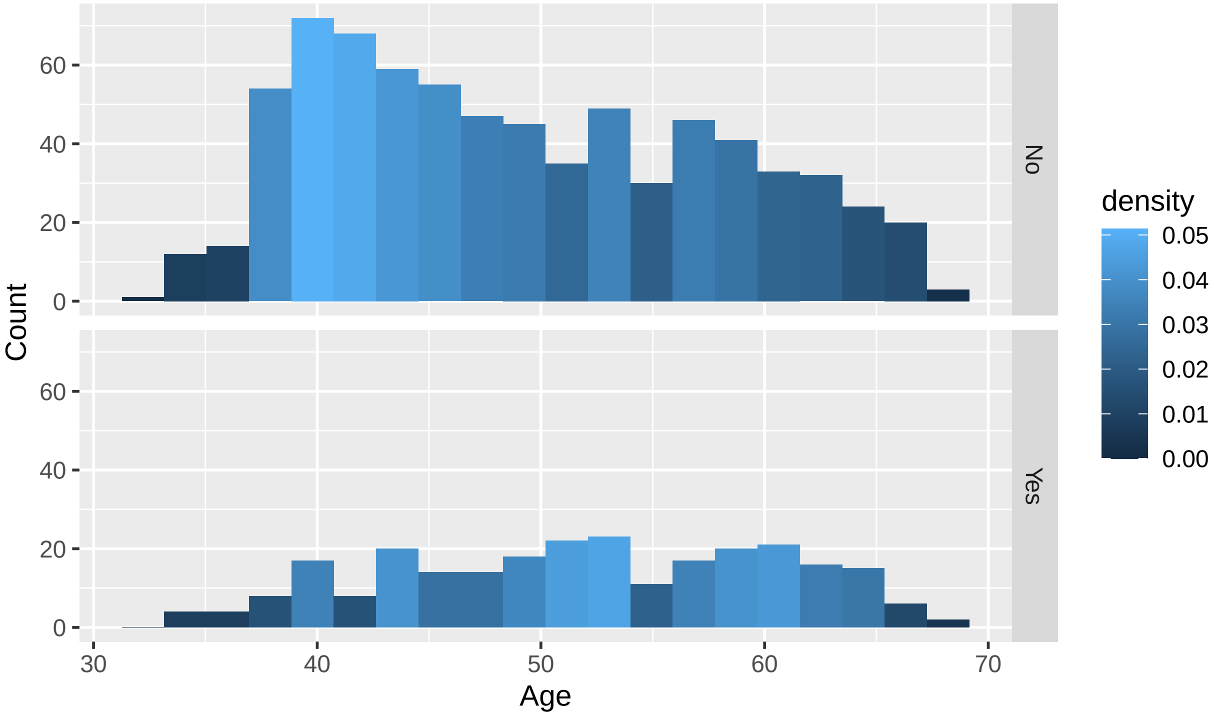
```
# Saving the plot
ggsave(p1, file = here("figures/", "Plot_A2.png"))

# Shapiro-Wilk test
subset_age %>%
  pull(age) %>%
  shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.96304, p-value = 3.142e-15
```

```
# Showing histogram with of age in relation to `anyCHD`
p2 <- subset_age %>%
  ggplot(aes(x = age)) +
  geom_histogram(aes(y = after_stat(count), fill = ..density..), bins = 20) +
  labs(title = "Figure A3: Age distribution",
       subtitle = "in relation to the occurrence of any coronary heart disease",
       x = "Age", y = "Count",
       caption = "Data source: Framingham Heart Study") +
  facet_grid(rows = vars(anyCHD))
p2
```

Figure A3: Age distribution

in relation to the occurrence of any coronary heart disease



Data source: Framingham Heart Study

```r
# Saving the plot
ggsave(p2, file = here("figures/", "Plot_A3.png"))
```

Trusting the appearance of the histograms, showing rise on the left of the graph, and the Shapiro-Wilk test with a p-value of less than 0.05, we assume that the age variable is not normally distributed. Considering that, the use of parametric tests like t-tests is not recommended.

```r
# Checking age_cat
p3 <- crosstab_age_cat %>%
  mutate(anyCHD = fct_rev(anyCHD)) %>%
  ggplot(aes(x = age_cat, fill = anyCHD)) +
  geom_bar(position = "fill") +
  labs(title = "Figure A4: Age categories",
       subtitle = "in relation to the occurrence of any coronary heart disease",
       x = "Age category", y = "Proportion",
       caption = "Data source: Framingham Heart Study") +
  theme_minimal()
p3
```

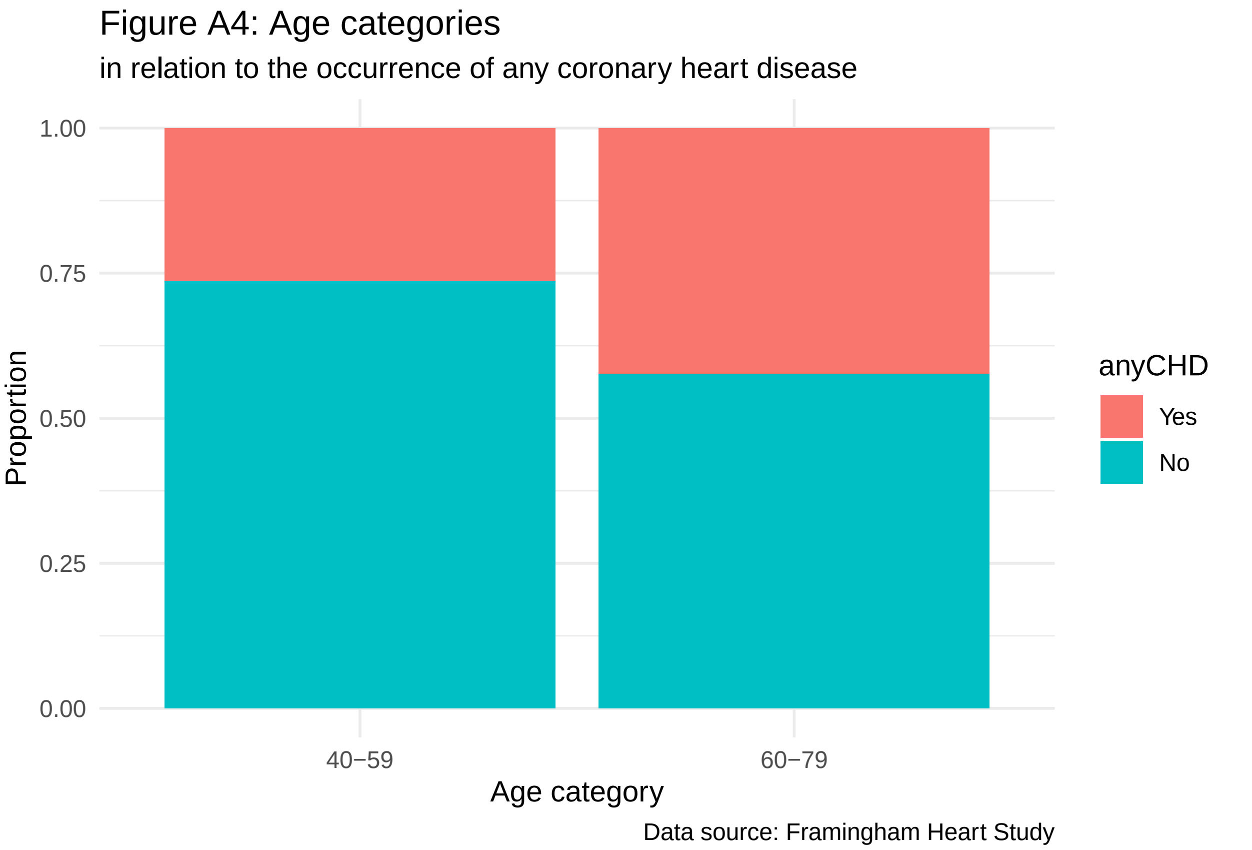


Figure A4: Age categories

in relation to the occurrence of any coronary heart disease

Data source: Framingham Heart Study

```r
# Saving the plot
ggsave(p3, file = here("figures/", "Plot_A4.png"))
```

The barplot shows that the proportion of individuals with any coronary heart or without in the two age categories. It is showing that the proportion of individuals with CHD is higher in the 60-79 age category than in the 40-59 age category.

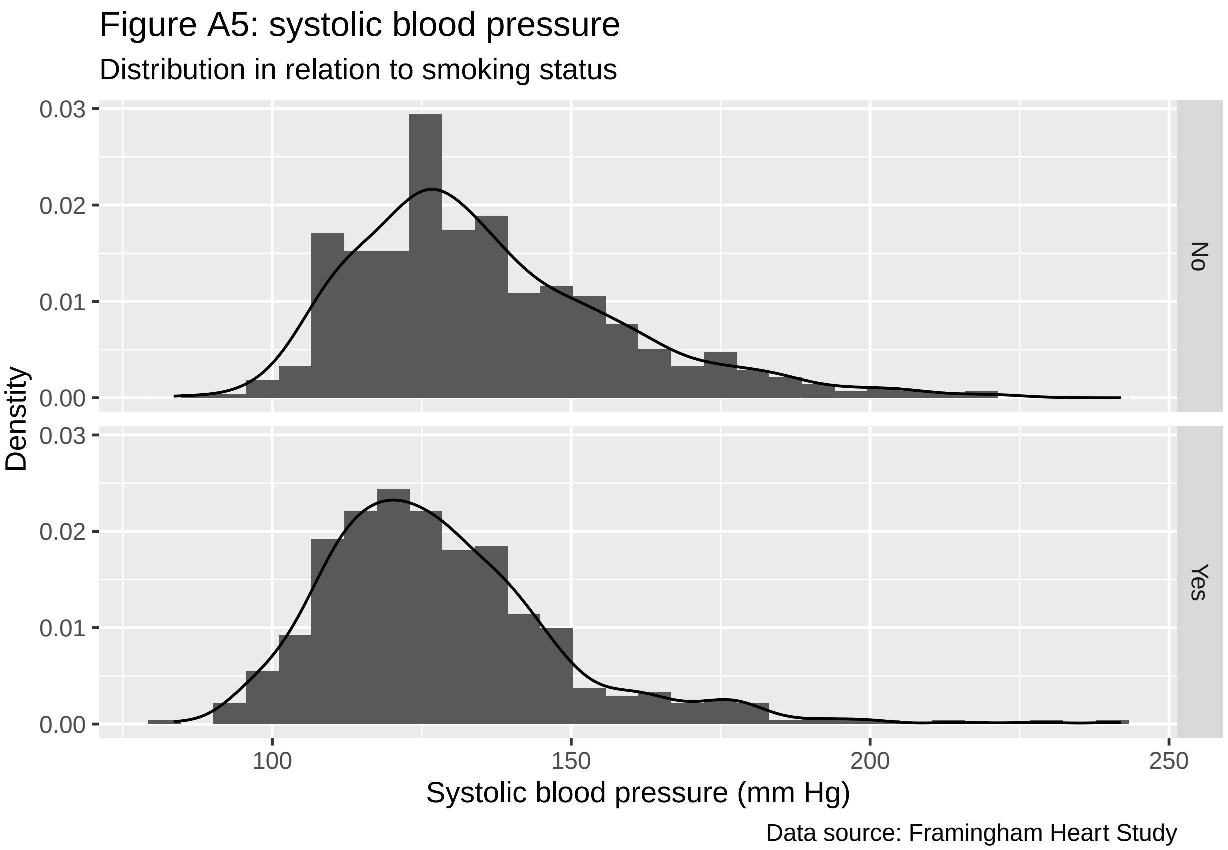
2.

a.

The assumption of normal distribution for systolic blood pressure will be tested.

```r
# Creating sampple dataset
subset_sysBP <- crosstab_sysBP %>%
  drop_na(sysBP) %>%
  sample_n(1000)

# Checking distribution of sysBP (also for normality)
p4 <- subset_sysBP %>%
  ggplot(aes(x = sysBP)) +
  geom_histogram(aes(y = after_stat(density))) +
  geom_density(aes(y = after_stat(density))) +
  labs(title = "Figure A5: systolic blood pressure",
       subtitle = "Distribution in relation to smoking status",
       x = "Systolic blood pressure (mm Hg)", y = "Denstity",
       caption = "Data source: Framingham Heart Study") +
  facet_grid(rows = vars(currentSmoker))
```

```
p4
```



```
# Saving the plot
ggsave(p4, file = here("figures/", "Plot_A5.png"))

# Performing Shapiro-Wilk test
subset_sysBP %>%
  pull(sysBP) %>%
  shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.9325, p-value < 2.2e-16
```

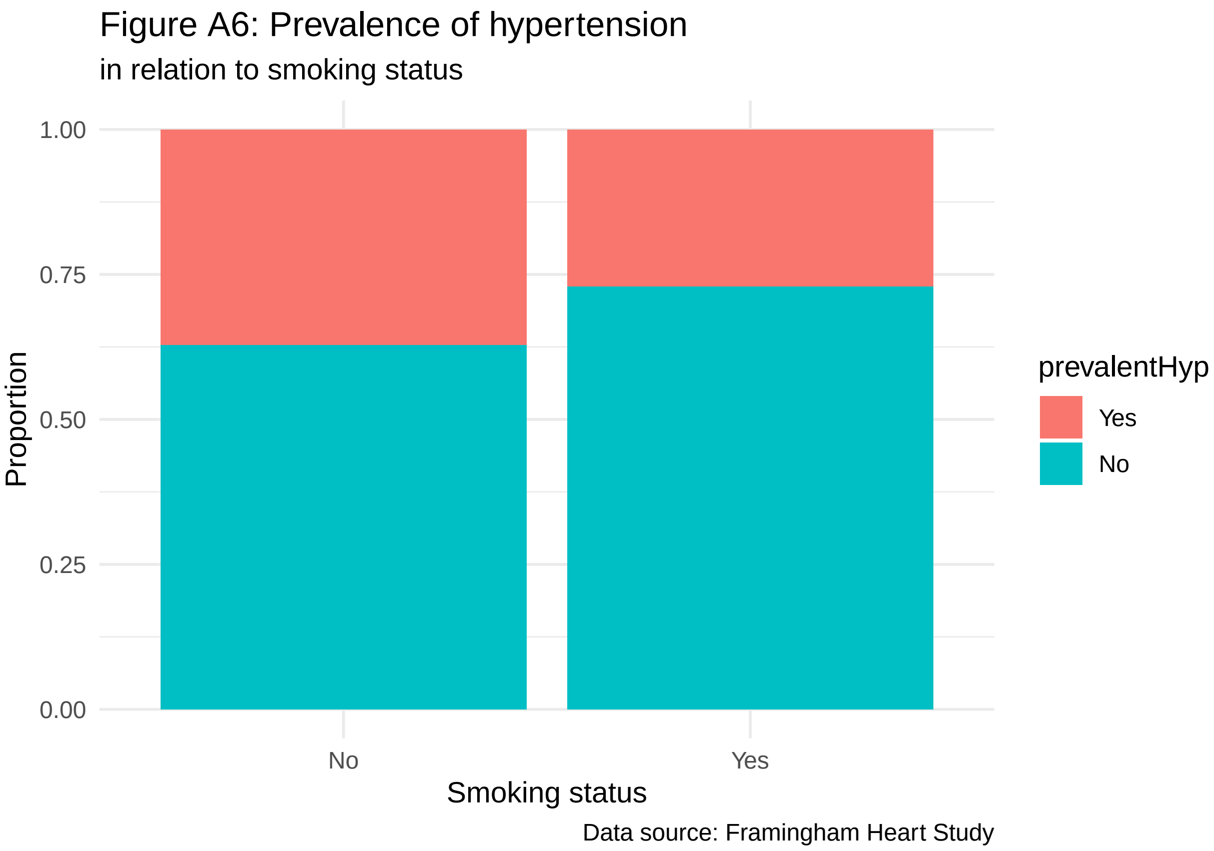
The p-value is very small, suggesting that the systolic blood pressure variable is not normally distributed. This means that we cannot use a parametric test like the t-test.

b.

Here the assumption of normal distribution is not relevant, as it is the distribution of a discrete variable. The Fisher excat test will be used.

```
# Checking distribution of prevalentHyp
p5 <- crosstab_preHyp %>%
  mutate(prevalentHyp = fct_rev(prevalentHyp)) %>%
  ggplot(aes(x = currentSmoker, fill = prevalentHyp)) +
  geom_bar(position = "fill") +
```

```r
  labs(title = "Figure A6: Prevalence of hypertension",
       subtitle = "in relation to smoking status",
       x = "Smoking status", y = "Proportion",
       caption = "Data source: Framingham Heart Study") +
  theme_minimal()
p5
```



```r
# Save the plot
ggsave(p5, file = here("figures/", "Plot_A6.png"))
```

3.

The assumption of normal distribution for continuous variables is tested through the Shapiro-Wilk test. The assumption of normal distribution for categorical variables is not relevant.

Carry out the analysis

1.

a.

The riskratio test is used to compare the risk of heart disease between smokers and non-smokers. The test is performed using the `epitab` function from the `epitools` package. The result is a risk ratio and a 95% confidence interval.

```r
# Calculating relative risk
framingham_df %>%
  select(currentSmoker, anyCHD) %>%
  table() %>%
  epitab(., method = "riskratio", rev = "columns")
```

```
## $tab
##               anyCHD
## currentSmoker Yes        p0   No        p1 riskratio     lower     upper
##           No  623 0.2765202 1630 0.7234798 1.0000000        NA        NA
##           Yes 617 0.2828978 1564 0.7171022 0.9911849 0.9554725 1.028232
##               anyCHD
## currentSmoker   p.value
##           No         NA
##           Yes 0.6395784
##
## $measure
## [1] "wald"
##
## $conf.level
## [1] 0.95
##
## $pvalue
## [1] "fisher.exact"
```

b.

The Wilcoxon rank sum test is used to compare the median age of individuals with and without any coronary heart disease. Also the t-.test is used to compare the mean age that were calculated before.

```
# Performing t-test (is this parametric test valid?
# as age is NOT normally distributed!!!)
crosstab_age %>%
  t.test(age ~ anyCHD, data = .)
```

```
##
##  Welch Two Sample t-test
##
## data:  age by anyCHD
## t = -13.309, df = 2244.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -4.358304 -3.238902
## sample estimates:
##  mean in group No mean in group Yes
##          48.86349          52.66210
```

```
# Performing wilcox-test (non-parametric test, as age is NOT normally distributed)
# age in relation to anyCHD
crosstab_age %>%
  wilcox.test(age ~ anyCHD, data = ., conf.int = TRUE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  age by anyCHD
## W = 1483418, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -4.999981 -3.000064
## sample estimates:
## difference in location
##              -4.000079
```

```r
# Cleaning up the environment
rm(crosstab_age)
```

Now the riskratio is calculated for the two age categories (40 - 59 and 60 - 79), as the oddsratio.

```r
# Relative risk
crosstab_age_cat %>%
  mutate(age_cat = fct_relevel(age_cat, "60-79")) %>%
  table() %>%
  epitab(., method = "riskratio", rev = "columns")
```

```
## $tab
##       age_cat
## anyCHD 40-59        p0 60-79        p1 riskratio    lower    upper      p.value
##    No   2274 0.8341893   452 0.1658107  1.000000       NA       NA           NA
##    Yes   817 0.7110531   332 0.2889469  1.742631 1.539772 1.972216 1.851439e-17
##
## $measure
## [1] "wald"
##
## $conf.level
## [1] 0.95
##
## $pvalue
## [1] "fisher.exact"
```

```r
# Odds ratios
crosstab_age_cat %>%
  mutate(age_cat = fct_relevel(age_cat, "60-79")) %>%
  table() %>%
  epitab(., method = "oddsratio", rev = "columns")
```

```
## $tab
##       age_cat
## anyCHD 40-59        p0 60-79        p1 oddsratio    lower    upper      p.value
##    No   2274 0.7356842   452 0.5765306   1.00000       NA       NA           NA
##    Yes   817 0.2643158   332 0.4234694   2.04441 1.737491 2.405546 1.851439e-17
##
## $measure
## [1] "wald"
##
## $conf.level
## [1] 0.95
##
## $pvalue
## [1] "fisher.exact"
```

```r
# Cleaning up the environment
rm(crosstab_age, crosstab_age_cat, subset_age)
```

2.

a.

Several tests are conducted. First the Chi-square test is used. As the distribution of the systolic blood pressure is not normally distributed, the Wilcoxon rank sum was used to confirm these results. Finally the linear regression is used to verify the outcome.

```
# Performing Chi-squared test
crosstab_sysBP %>%
  table() %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 292.54, df = 236, p-value = 0.007151
```

```
# Performing Wilcoxon rank sum test (non-parametric test,
# as `sysBP` is NOT normally distributed)
crosstab_sysBP %>%
  wilcox.test(sysBP ~ currentSmoker, data = ., conf.int = TRUE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  sysBP by currentSmoker
## W = 2840855, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  4.000047 6.499952
## sample estimates:
## difference in location
##               5.000043
```

```
# Lineare Regression
crosstab_sysBP %>%
  lm(sysBP ~ currentSmoker, data = .) %>%
  summary()
```

```
##
## Call:
## lm(formula = sysBP ~ currentSmoker, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.368 -15.368  -3.868  11.146 159.132
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      135.8680     0.4682 290.221   <2e-16 ***
## currentSmokerYes  -6.0181     0.6675  -9.016   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.22 on 4432 degrees of freedom
## Multiple R-squared:  0.01801,    Adjusted R-squared:  0.01779
## F-statistic: 81.28 on 1 and 4432 DF,  p-value: < 2.2e-16
```

```
# Cleaning up the environment
rm(crosstab_sysBP, subset_sysBP)
```

b.

The Fisher exact test is used to compare the prevalence of hypertension between smokers and non-smokers.

```
# Performing Fisher exact test
crosstab_preHyp %>%
  table() %>%
  fisher.test()
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  .
## p-value = 9.074e-13
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.5526987 0.7159182
## sample estimates:
## odds ratio
##  0.6291657
```

```
# Cleaning up the environment
rm(crosstab_preHyp)
```

3.

The automated approach is used to create the tables which shows associations with a significance level of 0.01.

```
# Looping through the risk factors with Chi-square test
# total cholesterol, smoking status, agecategory, systolic blood pressure,
# BMI, and prevalence of hypertension
risk_factors <- list("totChol", "currentSmoker", "age_cat", "sysBP", "BMI",
                     "prevalentHyp")

# Creating empty list to store results
significance_tb <- list()

# Looping through the risk factors
for (risk_factor in risk_factors) {
  # Looping through both sexes and both gender
  for (sex_var in list("female", "male", c("female", "male"))) {
    # Creating crosstab for each risk factor
    crosstab <- framingham_df %>%
      mutate(age_cat = cut(age,
        breaks = c(39, 59, 79),
        labels = c("40-59", "60-79")
      )) %>%
      drop_na(all_of(risk_factor)) %>%
      select(sex, all_of(risk_factor), anyCHD)
    # Setting Chi-square test for categorical variables
    if (class(crosstab[[risk_factor]]) == "factor") {
      test_type <- "chisq.test"
    }

    # Tests for continuous variables
    else { # nolint
      # Testing for normal distribution
      norm_dist <- crosstab %>%
        pull(all_of(risk_factor)) %>%
        shapiro.test()

      # T-test for continuous variables normally distributed
      if (norm_dist$p.value > 0.01) {
        test_type <- "t.test"
      }

      # Wilcoxon test for continuous variables, not normally distributed
      else { # nolint
        test_type <- "wilcox.test"
      }
    }

    # Performing the test
    result <- suppressWarnings({
      crosstab %>%
        filter(sex %in% sex_var) %>%
        select(anyCHD, all_of(risk_factor)) %>%
        table() %>%
        get(test_type)()
    })

    # Cleaning up the environment
    rm(crosstab)

    # Outputting the result
    cat(paste0("Risk factor: ", risk_factor, ", Gender: ",
               paste(sex_var, collapse = " & ")))

    # Setting significance level to 0.01
    if (result$p.value < 0.01) {
      significance_tb <- append(significance_tb, list(c(risk_factor,
                                  paste(sex_var, collapse = " & "), "Yes"
                                )))
      cat(" -> Significant risk to attain a heart disease (", test_type, ")\n")
    } else {
      significance_tb <- append(significance_tb, list(c(risk_factor,
                                  paste(sex_var, collapse = " & "), "No"
                                )))
      cat(" -> No significant risk to attain a heart disease()", test_type, ")\n")
    }
  }
}
```

```
## Risk factor: totChol, Gender: female -> Significant risk to attain a heart disease ( wilcox.test )
## Risk factor: totChol, Gender: male -> Significant risk to attain a heart disease ( wilcox.test )
## Risk factor: totChol, Gender: female & male -> Significant risk to attain a heart disease ( wilcox.te
## Risk factor: currentSmoker, Gender: female -> Significant risk to attain a heart disease ( chisq.test
## Risk factor: currentSmoker, Gender: male -> No significant risk to attain a heart disease() chisq.tes
## Risk factor: currentSmoker, Gender: female & male -> No significant risk to attain a heart disease()
## Risk factor: age_cat, Gender: female -> Significant risk to attain a heart disease ( chisq.test )
## Risk factor: age_cat, Gender: male -> Significant risk to attain a heart disease ( chisq.test )
## Risk factor: age_cat, Gender: female & male -> Significant risk to attain a heart disease ( chisq.tes
## Risk factor: sysBP, Gender: female -> Significant risk to attain a heart disease ( wilcox.test )
## Risk factor: sysBP, Gender: male -> Significant risk to attain a heart disease ( wilcox.test )
## Risk factor: sysBP, Gender: female & male -> Significant risk to attain a heart disease ( wilcox.test
## Risk factor: BMI, Gender: female -> Significant risk to attain a heart disease ( wilcox.test )
## Risk factor: BMI, Gender: male -> Significant risk to attain a heart disease ( wilcox.test )
## Risk factor: BMI, Gender: female & male -> Significant risk to attain a heart disease ( wilcox.test )
## Risk factor: prevalentHyp, Gender: female -> Significant risk to attain a heart disease ( chisq.test
## Risk factor: prevalentHyp, Gender: male -> Significant risk to attain a heart disease ( chisq.test )
## Risk factor: prevalentHyp, Gender: female & male -> Significant risk to attain a heart disease ( chis
```

```r
# Cleaning up the environment
rm(framingham_df, risk_factors, risk_factor, sex_var, test_type, result)

# Viewing the list
View(significance_tb)

# Transforming the list into a dataframe
significance_tb_df <- significance_tb %>%
  as.data.frame()

# Giving the columns numbers (1, 2, 3 etc.)
colnames(significance_tb_df) <- c(1 : length(significance_tb_df)) # nolint

# Giving the rows names
rownames(significance_tb_df) <- c("RiskFactor", "Sex", "Significant")

# Viewing the dataframe
View(significance_tb_df)

# Transforming the dataframe into tibble
significance_tb_tidied <- significance_tb_df %>%
  t() %>%
  as_tibble()

# Viewing the tibble
View(significance_tb_tidied)

# Creating table A1
t1 <- significance_tb_tidied %>%
  group_by(Significant) %>%
  mutate(Group = Significant) %>%
  gt() %>%
  tab_header(title = "Table A1: Risk factors",
             subtitle = "in relation to occurrence of coronary heart disease") %>%
  tab_source_note(source_note = md("*Data source: Framingham Heart Study*")) %>%
  cols_hide(columns = Group) %>%
  cols_label(Group = md("*Significance*"), RiskFactor = md("**Risk Factor**"),
             Sex = md("**Gender**")) %>%
  cols_move(columns = c(RiskFactor, Sex), after = Group) %>%
  cols_align(align = "left", columns = everything()) %>%
  tab_spanner(label = md("*grouped by Significance*"), columns = RiskFactor) %>%
  tab_style(
    style = list(cell_fill(color = "#bd5c5c")),
    locations = cells_body(rows = Significant == "Yes")
  ) %>%
  tab_style(
    style = list(cell_fill(color = "#4b8b4b")),
    locations = cells_body(rows = Significant == "No")
  )
t1
```

Table A1: Risk factors

in relation to occurrence of coronary heart disease

grouped by Significance	
Risk Factor	**Gender**
Yes	
totChol	female
totChol	male
totChol	female & male
currentSmoker	female
age_cat	female
age_cat	male
age_cat	female & male
sysBP	female
sysBP	male
sysBP	female & male
BMI	female
BMI	male
BMI	female & male
prevalentHyp	female
prevalentHyp	male
prevalentHyp	female & male
No	
currentSmoker	male
currentSmoker	female & male

Data source: Framingham Heart Study

```r
gtsave(t1, filename = "Table_A1.pdf", path = here("tables/"))

# Creating table A2
t2 <- significance_tb_tidied %>%
  group_by(Sex) %>%
  mutate(Group = Sex) %>%
  gt() %>%
  tab_header(title = "Table A2: Risk factors",
             subtitle = "in relation to occurrence of coronary heart disease") %>%
  tab_source_note(source_note = md("*Data source: Framingham Heart Study*")) %>%
  cols_hide(columns = Group) %>%
  cols_label(Significant = md("**Significance**"), RiskFactor = md("**Risk Factor**")) %>%
  cols_move(columns = c(RiskFactor, Sex), after = Group) %>%
  cols_align(align = "left", columns = everything()) %>%
  tab_spanner(label = md("*grouped by Gender*"), columns = Significant) %>%
  tab_style(
    style = cell_fill(color = "#bd5c5c"),
    locations = cells_body(rows = Significant == "Yes")
  ) %>%
  tab_style(
    style = cell_fill(color = "#4b8b4b"),
    locations = cells_body(rows = Significant == "No")
  )
t2
```

Table A2: Risk factors

in relation to occurrence of coronary heart disease

grouped by Gender	
Significance	**Risk Factor**
female	
Yes	totChol
Yes	currentSmoker
Yes	age_cat
Yes	sysBP
Yes	BMI
Yes	prevalentHyp
male	
Yes	totChol
No	currentSmoker
Yes	age_cat
Yes	sysBP
Yes	BMI
Yes	prevalentHyp
female & male	
Yes	totChol
No	currentSmoker
Yes	age_cat
Yes	sysBP
Yes	BMI
Yes	prevalentHyp

Data source: Framingham Heart Study

```
gtsave(t2, filename = "Table_A2.pdf", path = here("tables/"))

# Cleaning up the environment
rm(significance_tb, significance_tb_df, significance_tb_tidied, t1, t2)
```

Interpretation

1.

a.

The relative risk of occurrence of any coronary heart disease given the smoking attribute show a slightly decreases risk factor if the individual is a smoker. With a lower and upper confidence interval of 0.955 and 1.028, which includes 1, there is no significant difference in the risk of occurrence of any coronary heart. Also the p-value of 0.64 is greater than 0.05, suggesting no significant relationship between the variables. This was already expected by the contingency table and the barplot which didn't show any differences in both groups.

b.

The Wilcoxon rank sum test results show a significant difference in the median age of individuals with and without any coronary heart disease. With a small p-value less than 0.05, we see a significant difference in the median age of individuals with and without any coronary heart disease. The confidence interval of -5 and -3 (0 not included) further supports this conclusion, as there seems to be a difference of around 4 years that persons with the disease are older, than these without the disease. After that the relative risk for the two age categories was calculated. Here a significant increase by 74,26% in the risk for the 60-79 age category can be seen. With a lower and upper confidence interval of 1.54 and 1.97, which does not include 1, there is a significant difference in the risk of occurrence of any coronary heart. Also the p-value is less than 0.05, suggesting a significant relationship. The odds ratio of occurrence of any coronary heart disease given the age category attribute indicates an increase in the chance by two if the individual is in the 60-79 age category. With a lower and upper confidence interval of 1.74 and 2.4 (which not included), there is a significant difference in opportunity of the occurrence of any coronary heart. Significance is expected with a p-value of less than 0.05.

2.

a.

The p-value of 0.007 is less than 0.05, suggesting a significant relationship between the variables through the Chi-squared test. That means it can be concluded that the mean systolic blood pressure of individuals who are smokers and those who are not is significantly different. As of non normal distribution of the systolic blood pressure, the Wilcoxon rank sum test was used to confirm these results. It produced a pmuch smaller p-value, giving a stronger proof of a significant difference. The confidence interval of 4 and 6.5 further supports this conclusion, as this range is not including 0 and is narrow. There seems to be a difference of 5 mmHg, meaning that non-smokers have a 5 mm Hg higher systolic blood pressure than non-smokers. The linear regression result show a significant difference in the mean systolic blood pressure of individuals who are smokers and those who are not. With a very small p-value, there is significance. The coefficient of -6 further supports this conclusion, as it indicates that smokers have a lower systolic blood pressure by 6 mmHg.

b.

The p-value is less than 0.05, suggesting a significant relationship between the variables. This means that the odds of having hypertension are significantly different between individuals who smoke and those who do not. The odds ratio of 0.63 indicates that the odds of having hypertension are 37% lower for individuals who smoke compared to those who do not smoke. The 95% confidence interval of 0.55 and 0.71 further supports this conclusion, as it does not include 1, indicating a significant difference.

3.

Smoking is surprisingly not a significant risk factor for the occurrence of any coronary heart disease for male persons of the Framingham Heart Study. All other risk factors show a significant impact on the risk of attaining any coronary heart disease.

Conclusion

Smoking is not a significant risk factor for the occurrence of any coronary heart. That is unexpected as the literature suggests that smoking is a significant risk for diffrent diseases like coronary heart disease (Centers for Disease Control and Prevention, 2021).

On the other hand, the Framingham Heart Study data shows significant impact of all other variables on the risk of attaining any coronary heart disease.

References

Centers for Disease Control and Prevention, 2021. Health effects of cigarette smoking. Centers for Disease Control; Prevention, (Accessed: 18.02.2024).

Framingham History Center, 2024. Mr. Danforth's farms becomes the town of framingham. Framingham History Center, (Accessed: 18.02.2024).

National Heart, Lung, and Blood Institute, 2024. Framingham heart study. National Heart, Lung,; Blood Institute, (Accessed: 18.02.2024).

United State Census Bureau, 2024. Framingham city, massachusetts. United State Census Bureau, (Accessed: 18.02.2024).

Session Info

Debug variable and output of used libraries.

```r
# Knitted with RStudio (2023.09.1+494 "Desert Sunflower" for macOS)
# For word counting the code and it's output will be neglected,
# setting the `debug_var` variable to `TRUE`.
sessionInfo()
```

```
## R version 4.3.2 (2023-10-31)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Sonoma 14.3.1
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapack.dylib;  LAPACK ve
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: Europe/Berlin
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] here_1.0.1        gt_0.10.1         broom_1.0.5       finalfit_1.0.7
##  [5] epitools_0.5-10.1 lubridate_1.9.3   forcats_1.0.0     stringr_1.5.1
##  [9] dplyr_1.1.4       purrr_1.0.2       readr_2.1.5       tidyr_1.3.1
## [13] tibble_3.2.1      ggplot2_3.4.4     tidyverse_2.0.0   pacman_0.5.1
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.2.0  farver_2.1.1      fastmap_1.1.1     webshot2_0.1.1
##  [5] promises_1.2.1    digest_0.6.34     rpart_4.1.21      timechange_0.3.0
##  [9] lifecycle_1.0.4   processx_3.8.3    survival_3.5-7    magrittr_2.0.3
## [13] compiler_4.3.2    rlang_1.1.3       sass_0.4.8        tools_4.3.2
## [17] utf8_1.2.4        yaml_2.3.8        knitr_1.45        labeling_0.4.3
## [21] bit_4.0.5         xml2_1.3.6        websocket_1.4.1   withr_3.0.0
## [25] nnet_7.3-19       grid_4.3.2        fansi_1.0.6       jomo_2.7-6
## [29] colorspace_2.1-0  mice_3.16.0       scales_1.3.0      iterators_1.0.14
## [33] MASS_7.3-60       cli_3.6.2         rmarkdown_2.25    crayon_1.5.2
## [37] ragg_1.2.7        generics_0.1.3    rstudioapi_0.15.0 tzdb_0.4.0
## [41] commonmark_1.9.1  chromote_0.2.0    minqa_1.2.6       splines_4.3.2
## [45] parallel_4.3.2    vctrs_0.6.5       boot_1.3-28.1     glmnet_4.1-8
## [49] Matrix_1.6-1.1    jsonlite_1.8.8    hms_1.1.3         bit64_4.0.5
## [53] mitml_0.4-5       systemfonts_1.0.5 foreach_1.5.2     glue_1.7.0
## [57] nloptr_2.0.3      pan_1.9           ps_1.7.6          codetools_0.2-19
## [61] stringi_1.8.3     shape_1.4.6       gtable_0.3.4      later_1.3.2
## [65] lme4_1.1-35.1     munsell_0.5.0     pillar_1.9.0      htmltools_0.5.7
## [69] R6_2.5.1          textshaping_0.3.7 rprojroot_2.0.4   vroom_1.6.5
## [73] evaluate_0.23     lattice_0.21-9    markdown_1.12     highr_0.10
## [77] backports_1.4.1   Rcpp_1.0.12       nlme_3.1-163      xfun_0.42
## [81] fs_1.6.3          pkgconfig_2.0.3
```