Assignment part 1

feedback and coding tips

Introduction

In this assignment, you will work with an individualized dataset derived from the Cleveland heart disease dataset. This dataset contains information on patients with suspected heart disease and includes various demographic, clinical, and diagnostic variables. Your task is to perform a series of analyses to explore the dataset and investigate the relationship between different variables and the presence of heart disease.

Dataset description

The Cleveland heart disease dataset originates from the Cleveland Clinic Foundation and focuses on heart disease diagnosis. It includes data from 303 patients on the following variables:

- `age`: Age in years
- `sex`: Sex (1 = female; 2 = male)
- `cp`: Chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
- `trestbps`: Resting blood pressure (mm Hg at hospital admission)
- `chol`: Serum cholesterol in mg/dl
- `fbs`: Fasting blood sugar > 120 mg/dl
- `restecg`: Resting electrocardiographic results (1 = normal; 2 = ST-T wave abnormality; 3 = left ventricular hypertrophy)
- `thalach`: Maximum heart rate achieved
- `exang`: Exercise-induced angina (1 = no; 2 = yes)
- `oldpeak`: ST depression induced by exercise relative to rest
- `slope`: Slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)
- `ca`: Number of major vessels (0-3) colored by fluoroscopy
- `thal`: Thallium heart scan results (1 = normal; 2 = fixed defect; 3 = reversible defect)
- `target`: Diagnosis of heart disease (1 = heart disease; 2 = no heart disease)

Objectives

In this assignment, you will explore the Cleveland heart disease dataset to answer the following research questions:

- Does maximum heart rate achieved differ across chest pain types?
- Is the presence of heart disease associated with sex or fasting blood sugar?

Steps to complete the assignment

Step 1: Create an individualized dataset

Step 2: Create a baseline characteristics table

- Include all variables in the dataset apart from the outcome variable `target`
  - Summarize demographic variables (e.g., `age`, `sex`)
  - Summarize clinical variables (e.g., `chol`, `trestbps`, `thalach`, `cp`)
- Decide on suitable summary measures for each variable
  - Use appropriate measures for continuous variables (e.g., mean, standard deviation, median, interquartile range)
  - Use frequency counts and percentages for categorical variables
- Present your table clearly
  - Use meaningful labels, headings, and clear formatting

```
library(tableone) # for creating baseline characteristics table
library(knitr) # for rendering tables
library(labelled) # for setting variable labels

library(ggplot2) # for creating plots
library(gridExtra) # for arranging plots

# Customize variable labels
heart_data <- heart_data %>%
  set_variable_labels(
    age = "Age (years)",
    sex = "Sex",
    cp = "Chest Pain Type",
    trestbps = "Resting Blood Pressure (mmHg)",
    chol = "Cholesterol (mg/dL)",
```

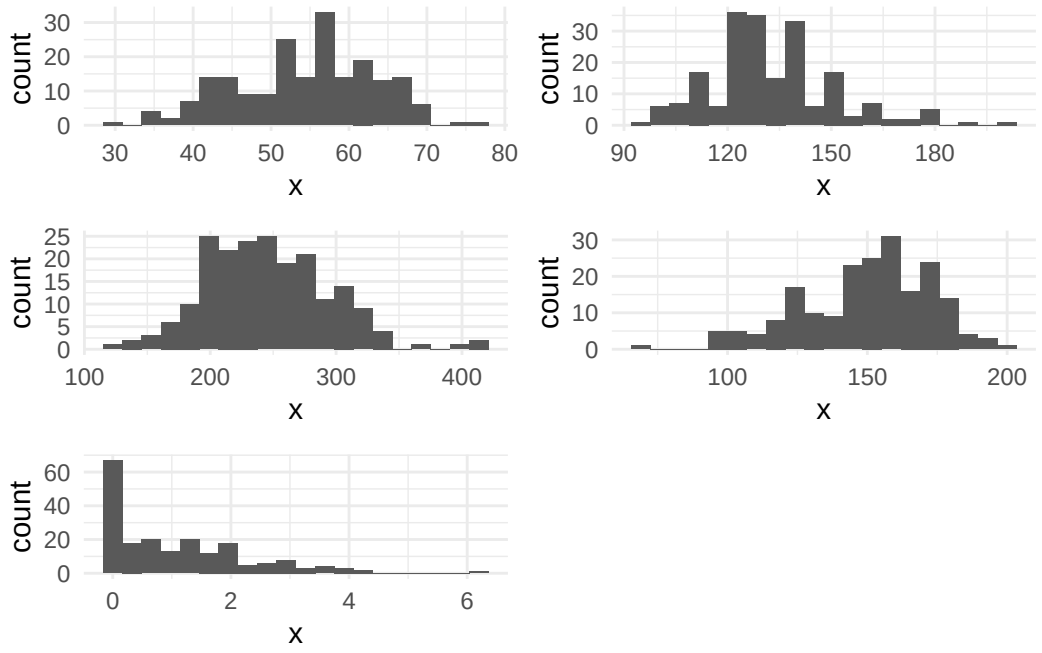
```
    fbs = "Fasting Blood Sugar",
    restecg = "Resting ECG Results",
    thalach = "Max Heart Rate Achieved",
    exang = "Exercise-Induced Angina",
    oldpeak = "ST Depression",
    slope = "Slope of ST Segment",
    ca = "Number of Major Vessels",
    thal = "Thalassemia Type",
    target = "Heart Disease Status"
  )

# Define categorical variables
categorical_vars <- c("sex", "cp", "fbs", "restecg", "exang", "slope", "ca", "thal")
continuous_vars <- c("age", "trestbps", "chol", "thalach", "oldpeak")

# Define all variables for the table
all_vars <- c(continuous_vars, categorical_vars)

# Create the table
table1 <- CreateTableOne(vars = all_vars,
                         data = heart_data,
                         factorVars = categorical_vars)

# Create histograms for continuous variables
histograms <- lapply(heart_data[continuous_vars], function(x) ggplot(heart_data, aes(x = x))
grid.arrange(grobs = histograms, ncol = 2)
```




```
# Render the table with knitr::kable
kable(print(table1, nonnormal = c("oldpeak"), varLabels = TRUE, printToggle = FALSE))
```

	Overall
n	200
Age (years) (mean (SD))	54.43 (9.03)
Resting Blood Pressure (mmHg) (mean (SD))	132.62 (18.33)
Cholesterol (mg/dL) (mean (SD))	245.18 (49.81)
Max Heart Rate Achieved (mean (SD))	150.06 (23.27)
ST Depression (median [IQR])	0.80 [0.00, 1.65]
Sex = Male (%)	134 (67.0)
Chest Pain Type (%)	
Typical angina	16 (8.0)
Atypical angina	38 (19.0)
Non-anginal pain	56 (28.0)
Asymptomatic	90 (45.0)
Fasting Blood Sugar = True (%)	29 (14.5)
Resting ECG Results (%)	
Normal	102 (51.0)
ST-T wave abnormality	2 (1.0)
Left ventricular hypertrophy	96 (48.0)

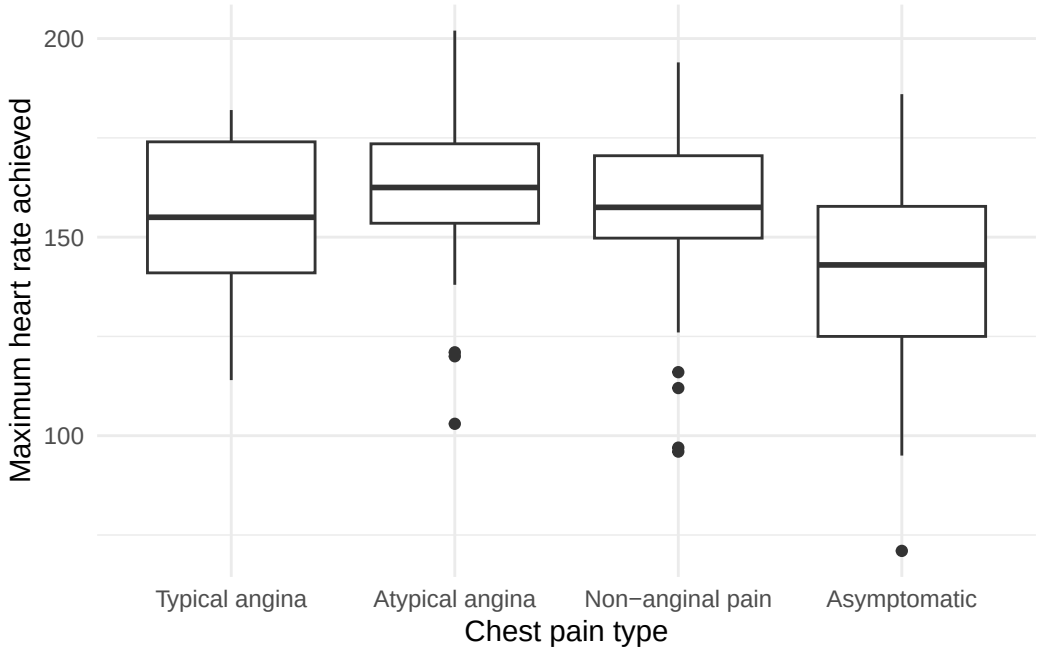
	Overall
Exercise-Induced Angina = Yes (%)	66 (33.0)
Slope of ST Segment (%)	
Upsloping	92 (46.0)
Flat	92 (46.0)
Downsloping	16 (8.0)
Number of Major Vessels (%)	
0	122 (61.6)
1	42 (21.2)
2	22 (11.1)
3	12 (6.1)
Thalassemia Type (%)	
Normal	112 (56.3)
Fixed defect	9 (4.5)
Reversible defect	78 (39.2)

Step 3: Perform the analysis for the first research question

Research Question: *Does maximum heart rate achieved differ across chest pain types?*

- Visualize the data
  - Create a boxplot to compare the distribution of maximum heart rate achieved across the four chest pain types

```
# Create a boxplot to compare the distribution of maximum heart rate achieved across the four
library(ggplot2)
ggplot(heart_data, aes(x = factor(cp), y = thalach)) +
  geom_boxplot() +
  labs(x = "Chest pain type", y = "Maximum heart rate achieved") +
  theme_minimal()
```



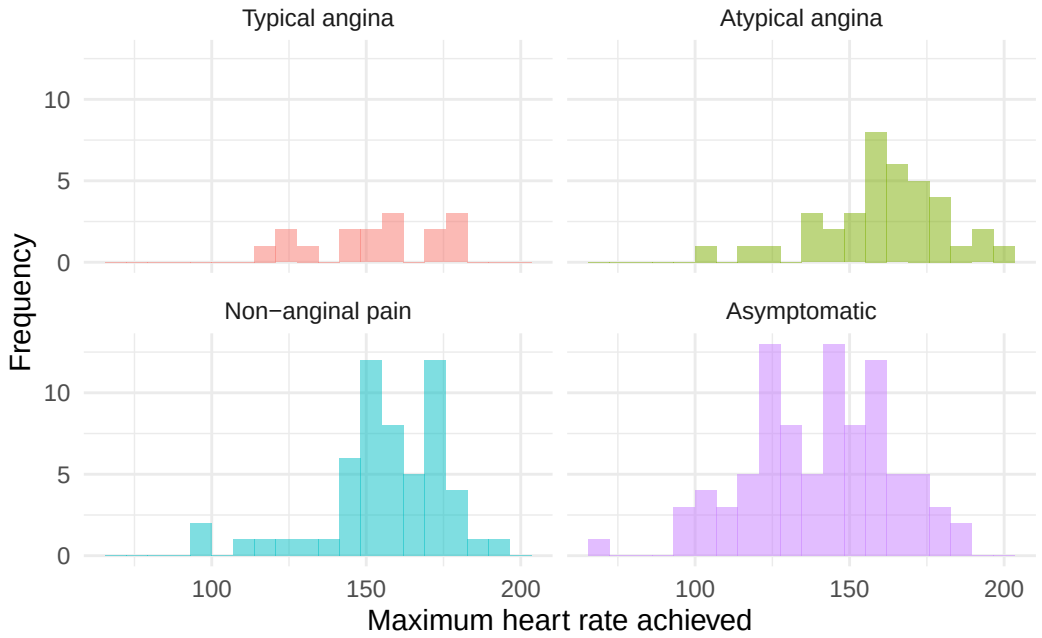
- Calculate the estimated population means and 95% confidence intervals for maximum heart rate achieved for each of the four chest pain types

```
# Calculate the estimated population means and 95% confidence intervals for maximum heart ra
for (i in levels(heart_data$cp)) {
  mean_thalach <- mean(heart_data$thalach[heart_data$cp == i], na.rm = TRUE)
  se_thalach <- sd(heart_data$thalach[heart_data$cp == i], na.rm = TRUE) / sqrt(sum(heart_da
  ci_thalach <- qt(0.975, sum(heart_data$cp == i, na.rm = TRUE) - 1) * se_thalach
  cat(paste("Chest pain type", i, ": Mean = ", round(mean_thalach, 2), ", 95% CI = (", round
}
```

```
Chest pain type Typical angina : Mean =  153.44 , 95% CI = ( 142.05 ,  164.83 )
Chest pain type Atypical angina : Mean =  161.53 , 95% CI = ( 154.77 ,  168.28 )
Chest pain type Non-anginal pain : Mean =  156.57 , 95% CI = ( 151.27 ,  161.87 )
Chest pain type Asymptomatic : Mean =  140.57 , 95% CI = ( 135.7 ,  145.43 )
```

- Select and perform an appropriate test
  - Use one-way ANOVA if normality and equal variances are met
  - Use Kruskal-Wallis test if assumptions are violated
  - Perform post-hoc comparisons using Bonferroni adjusted p-values if significant differences are found

```r
# Create histograms of maximum heart rate achieved for each chest pain type
ggplot(heart_data, aes(x = thalach, fill = factor(cp))) +
  geom_histogram(bins = 20, alpha = 0.5) +
  facet_wrap(~ cp) +
  labs(x = "Maximum heart rate achieved", y = "Frequency", fill = "Chest pain type") +
  theme_minimal() +
  theme(legend.position = "none")
```




```r
# Perform levene's test for homogeneity of variances
library(car)
leveneTest(heart_data$thalach ~ heart_data$cp) # variances are equal
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   3  1.5584 0.2008
      196
```

one-way ANOVA

```r
# Perform one-way ANOVA
anova_thalach <- aov(thalach ~ cp, data = heart_data)
summary(anova_thalach)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
cp            3  15664    5221   11.11 9.13e-07 ***
Residuals   196  92135     470
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Perform post-hoc comparisons using Bonferroni adjusted p-values
pairwise.t.test(heart_data$thalach, heart_data$cp, p.adj = "bonferroni")
```

```
    Pairwise comparisons using t tests with pooled SD

data:  heart_data$thalach and heart_data$cp

                Typical angina Atypical angina Non-anginal pain
Atypical angina  1.00000        -               -
Non-anginal pain 1.00000        1.00000         -
Asymptomatic     0.17911        7.7e-06         0.00014

P value adjustment method: bonferroni
```

```r
# Perform Tukey's HSD post-hoc test
tukey_results <- TukeyHSD(anova_thalach)
print(tukey_results)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = thalach ~ cp, data = heart_data)

$cp
                                      diff        lwr        upr     p adj
Atypical angina-Typical angina     8.088816  -8.654139  24.831770 0.5946651
Non-anginal pain-Typical angina    3.133929 -12.791798  19.059655 0.9566562
Asymptomatic-Typical angina      -12.870833 -28.113421   2.371754 0.1301598
Non-anginal pain-Atypical angina  -4.954887 -16.762582   6.852808 0.6976795
Asymptomatic-Atypical angina     -20.959649 -31.828383 -10.090916 0.0000076
Asymptomatic-Non-anginal pain    -16.004762 -25.566749  -6.442775 0.0001349
```

Kruskal-Wallis test

```
# Perform Kruskal-Wallis test
kruskal.test(heart_data$thalach ~ heart_data$cp)
```

```
    Kruskal-Wallis rank sum test

data:  heart_data$thalach by heart_data$cp
Kruskal-Wallis chi-squared = 29.972, df = 3, p-value = 1.399e-06
```

```
# Perform post-hoc comparisons using pairwise Wilcoxon tests
pairwise.wilcox.test(heart_data$thalach, heart_data$cp, p.adjust.method = "bonferroni")
```

```
    Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  heart_data$thalach and heart_data$cp

                Typical angina Atypical angina Non-anginal pain
Atypical angina  1.00000        -               -
Non-anginal pain 1.00000        1.00000         -
Asymptomatic     0.25664        3.5e-05         0.00011

P value adjustment method: bonferroni
```

```
# Perform post-hoc comparisons using Dunn's test
library(dunn.test)
dunn.test(heart_data$thalach, heart_data$cp, method = "bonferroni", altp = TRUE)
```

```
  Kruskal-Wallis rank sum test

data: x and group
Kruskal-Wallis chi-squared = 29.9725, df = 3, p-value = 0

                           Comparison of x by group
                                 (Bonferroni)
Col Mean-|
Row Mean |   Asymptom   Atypical   Non-angi
---------+---------------------------------
```

```
Atypical |  -4.710798
         |    0.0000*
         |
Non-angi |  -4.154178   0.972094
         |    0.0002*     1.0000
         |
Typical  |  -2.062334   1.180502   0.520346
         |     0.2351     1.0000     1.0000

alpha = 0.05
Reject Ho if p <= alpha
```

Step 4: Perform the analysis for the second research question

Research Question: *Is the presence of heart disease associated with sex or fasting blood sugar?*

- Summarize the data
  - Calculate and report the prevalence of heart disease for each group within sex and fasting blood sugar
  - Include percentages and 95% confidence intervals for each group
- Select and perform an appropriate test
  - Use a Chi-Square test of homogeneity or Fisher's Exact Test, depending on the expected cell counts

Association between heart disease and sex

Descriptives

	Sex (1 = female; 2 = male)			Statistic	Std. Error
Diagnosis of heart disease	Female	Mean		1,74	,045
		95% Confidence Interval for Mean	Lower Bound	1,65	
			Upper Bound	1,83	
		5% Trimmed Mean		1,77	
		Median		2,00	
		Variance		,193	
		Std. Deviation		,440	
		Minimum		1	
		Maximum		2	
		Range		1	
		Interquartile Range		1	
		Skewness		-1,125	,245
		Kurtosis		-,750	,485
	Male	Mean		1,45	,035
		95% Confidence Interval for Mean	Lower Bound	1,38	
			Upper Bound	1,52	
		5% Trimmed Mean		1,44	
		Median		1,00	
		Variance		,248	
		Std. Deviation		,498	
		Minimum		1	
		Maximum		2	
		Range		1	
		Interquartile Range		1	
		Skewness		,216	,169
		Kurtosis		-1,972	,337

Figure 1: 95% CIs using the orginal variable coding (1 = heart disease; 2 = no heart disease)

Descriptives

	Sex (1 = female; 2 = male)			Statistic	Std. Error
hd_recoded	Female	Mean		,2577	,04464
		95% Confidence Interval for Mean	Lower Bound	,1691	
			Upper Bound	,3463	
		5% Trimmed Mean		,2308	
		Median		,0000	
		Variance		,193	
		Std. Deviation		,43966	
		Minimum		,00	
		Maximum		1,00	
		Range		1,00	
		Interquartile Range		1,00	
		Skewness		1,125	,245
		Kurtosis		-,750	,485
	Male	Mean		,5534	,03472
		95% Confidence Interval for Mean	Lower Bound	,4849	
			Upper Bound	,6219	
		5% Trimmed Mean		,5593	
		Median		1,0000	
		Variance		,248	
		Std. Deviation		,49835	
		Minimum		,00	
		Maximum		1,00	
		Range		1,00	
		Interquartile Range		1,00	
		Skewness		-,216	,169
		Kurtosis		-1,972	,337

Figure 2: 95% CIs using the 0/1 recoded variable (1 = heart disease; 0 = no heart disease)

```
table(heart_data$sex, heart_data$target)
```

```
        Heart Disease No Heart Disease
Female             17               49
Male               76               58
```

```r
# Calculate and report the prevalence of heart disease for sex
prop.table(table(heart_data$sex, heart_data$target), margin = 1)
```

```
        Heart Disease No Heart Disease
  Female     0.2575758        0.7424242
  Male       0.5671642        0.4328358
```

```r
# Calculate and report the prevalence of heart disease for sex
prop.table(table(heart_data$sex, heart_data$target))
```

```
        Heart Disease No Heart Disease
  Female         0.085            0.245
  Male           0.380            0.290
```

```r
# Perform a chi-square test of homogeneity
chisq.test(table(heart_data$sex, heart_data$target))
```

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  table(heart_data$sex, heart_data$target)
X-squared = 15.815, df = 1, p-value = 6.985e-05
```

```r
# Retrieve expected cell counts
chisq.test(table(heart_data$sex, heart_data$target))$expected
```

```
        Heart Disease No Heart Disease
  Female         30.69            35.31
  Male           62.31            71.69
```

Association between heart disease and fasting blood sugar

```r
# Construct contingency table
table(heart_data$fbs, heart_data$target)
```

```
       Heart Disease No Heart Disease
  False             81                90
  True              12                17
```

```r
# Calculate and report the prevalence of heart disease for fasting blood sugar
prop.table(table(heart_data$fbs, heart_data$target), margin = 1)
```

```
       Heart Disease No Heart Disease
  False     0.4736842        0.5263158
  True      0.4137931        0.5862069
```

```r
# Perform a chi-square test of homogeneity
chisq.test(table(heart_data$fbs, heart_data$target))
```

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  table(heart_data$fbs, heart_data$target)
X-squared = 0.15729, df = 1, p-value = 0.6917
```

```r
# Retrieve expected cell counts
chisq.test(table(heart_data$fbs, heart_data$target))$expected
```

```
       Heart Disease No Heart Disease
  False        79.515           91.485
  True         13.485           15.515
```