Homework 6

Ly Nguyen

2023-11-01

The Cardiovascular Health Study (CHS) is a population-based, longitudinal study of coronary heart disease and stroke in adults aged 65 years and older. Study participants were recruited in 1989-1990 from four communities: Forsyth County, NC; Sacramento County, CA; Washington County, MD; and Pittsburgh, PA. The data for this study consists of the subset of participants recruited in the first wave of recruitment who were 'healthy,' that is, had no history of heart or circulation disease, no restriction of daily activities by illness, and no medications that would indicate heart disease. A large number of variables were determined for each study participant at baseline, that is, at the time of recruitment. The baseline examination consisted of a home interview and a clinic examination. During the home interview, information was collected on prior medical history, medical usage, and physical activity. Information was also obtained regarding the presence of impairments in physical functioning. The clinic examination included a fasting blood draw and seated blood pressure measurements.

```
chsData <- read.table("~/Documents/Personal Docs_East Bay/STAT 630/HW6/chsData.txt", quote="\"", commen
View(chsData)
library(dplyr)
library(ggplot2)
library(openintro)
library(knitr)
```

Part 1: Data Cleaning

```
1. Using any of the methods we have learned in class, clean the dataset by:
    a. Changing categorical variables to factors (this includes clinic, season, arth, diab, income, exi
```

```
# Data Cleaning
chsData <- chsData %>%
  mutate(clinic = factor(clinic,
                         labels = c("Sacramento", "Forsyth", "Washington", "Pittsburgh")),
         season = factor(season,
                         labels = c("summer", "fall", "winter", "spring")),
         arth = factor(arth,
                       labels = c("none", "arthritis")),
         diab = factor(diab,
                       labels = c("none", "borderline", "diabetes")),
         income = factor(income,
                         labels = c("< 5k", "5k-8k", "8k-12k", "12k-16k", "16k-24k", "24k-35k", "35k-50
         exint0 = factor(exint0,
                         labels = c("no exercise", "low intensity", "moderate intensity", "high intensi
)


summary(chsData)
```

```
##         clinic       initdate         season        gender
```

```
##  Sacramento:678   Min.   :148529   summer:627   Min.   :0.0000
##  Forsyth   :619   1st Qu.:148623   fall  :647   1st Qu.:0.0000
##  Washington:550   Median :148708   winter:613   Median :0.0000
##  Pittsburgh:593   Mean   :148712   spring:553   Mean   :0.4012
##                   3rd Qu.:148799                3rd Qu.:1.0000
##                   Max.   :148883                Max.   :1.0000
##
##       age            weight         weight50         grade
##  Min.   :65.00   Min.   : 73.5   Min.   : 80.0   Min.   : 0.00
##  1st Qu.:68.00   1st Qu.:134.5   1st Qu.:130.0   1st Qu.:12.00
##  Median :71.00   Median :155.6   Median :149.0   Median :12.00
##  Mean   :71.86   Mean   :157.5   Mean   :150.8   Mean   :14.32
##  3rd Qu.:75.00   3rd Qu.:178.0   3rd Qu.:170.0   3rd Qu.:19.00
##  Max.   :95.00   Max.   :323.0   Max.   :290.0   Max.   :21.00
##                  NA's   :7       NA's   :88      NA's   :6
##         arth           sbp             pkyrs                diab
##  none     :1270   Min.   : 79.0   Min.   :  0.00   none      :1836
##  arthritis:1137   1st Qu.:120.0   1st Qu.:  0.00   borderline: 322
##  NA's     :  33   Median :133.0   Median :  0.60   diabetes  : 265
##                   Mean   :134.8   Mean   : 16.65   NA's      :  17
##                   3rd Qu.:148.0   3rd Qu.: 27.00
##                   Max.   :227.0   Max.   :204.00
##                   NA's   :7       NA's   :67
##      income                    exint0         block0           kcal0
##  16k-24k:476   no exercise       : 134   Min.   :  0.00   Min.   :    0.0
##  24k-35k:415   low intensity     :1152   1st Qu.: 10.00   1st Qu.:  213.8
##  > 50k  :372   moderate intensity: 867   Median : 24.00   Median :  735.0
##  12k-16k:328   high intensity    : 285   Mean   : 45.75   Mean   : 1341.2
##  35k-50k:259   NA's              :   2   3rd Qu.: 60.00   3rd Qu.: 1768.1
##  (Other):433                             Max.   :300.00   Max.   :14160.0
##  NA's   :157                             NA's   :25       NA's   :4
```

b. Making a new variable called sbp140 that is a binary indicator of whether a person's sbp is >= 140 o

```
# Making a new binary variable "sbp140" that is a  indicator of whether a person's sbp is >= 140 or < 1
# 1st way: (I just wanna try on this way)
# chsData$sbp140 <- numeric(length(chsData$sbp))
# for (i in 1:nrow(chsData)) {
#   if (!is.na(chsData$sbp[i])) {
#     if (chsData$sbp[i] >= 140) {
#       chsData$sbp140[i] <- 1
#     } else {
#       chsData$sbp140[i] <- 0
#     }
#   } else {
#     is.na(chsData$sbp[i]) <- NA
#   }
# }

# 2nd way:
chsData <- chsData %>%
  mutate(sbp140 = case_when(sbp >= 140 ~ ">= 140",
                            sbp >= 0 & sbp < 140 ~ "< 140",
                            TRUE ~ NA)) %>%
  mutate(sbp140 = factor(sbp140))
```

```r
summary(chsData)
```

```
##         clinic       initdate         season        gender
##  Sacramento:678   Min.   :148529   summer:627   Min.   :0.0000
##  Forsyth   :619   1st Qu.:148623   fall  :647   1st Qu.:0.0000
##  Washington:550   Median :148708   winter:613   Median :0.0000
##  Pittsburgh:593   Mean   :148712   spring:553   Mean   :0.4012
##                   3rd Qu.:148799                3rd Qu.:1.0000
##                   Max.   :148883                Max.   :1.0000
##
##       age            weight         weight50         grade
##  Min.   :65.00   Min.   : 73.5   Min.   : 80.0   Min.   : 0.00
##  1st Qu.:68.00   1st Qu.:134.5   1st Qu.:130.0   1st Qu.:12.00
##  Median :71.00   Median :155.6   Median :149.0   Median :12.00
##  Mean   :71.86   Mean   :157.5   Mean   :150.8   Mean   :14.32
##  3rd Qu.:75.00   3rd Qu.:178.0   3rd Qu.:170.0   3rd Qu.:19.00
##  Max.   :95.00   Max.   :323.0   Max.   :290.0   Max.   :21.00
##                  NA's   :7       NA's   :88      NA's   :6
##         arth           sbp             pkyrs                diab
##  none     :1270   Min.   : 79.0   Min.   :  0.00   none      :1836
##  arthritis:1137   1st Qu.:120.0   1st Qu.:  0.00   borderline: 322
##  NA's     :  33   Median :133.0   Median :  0.60   diabetes  : 265
##                   Mean   :134.8   Mean   : 16.65   NA's      :  17
##                   3rd Qu.:148.0   3rd Qu.: 27.00
##                   Max.   :227.0   Max.   :204.00
##                   NA's   :7       NA's   :67
##      income                    exint0         block0           kcal0
##  16k-24k:476   no exercise       : 134   Min.   :  0.00   Min.   :    0.0
##  24k-35k:415   low intensity     :1152   1st Qu.: 10.00   1st Qu.:  213.8
##  > 50k  :372   moderate intensity: 867   Median : 24.00   Median :  735.0
##  12k-16k:328   high intensity    : 285   Mean   : 45.75   Mean   : 1341.2
##  35k-50k:259   NA's              :   2   3rd Qu.: 60.00   3rd Qu.: 1768.1
##  (Other):433                             Max.   :300.00   Max.   :14160.0
##  NA's   :157                             NA's   :25       NA's   :4
##     sbp140
##  < 140 :1517
##  >= 140: 916
##  NA's  :   7
##
##
##
##
```

Part 2: Exploratory Data Analysis

2. Missing Values:
    a. Make a publication-quality table that shows the number of missing rows for each variable in the d

```r
missing_values <- colSums(is.na(chsData))

summary_table <- data.frame(
  Variable = names(missing_values),
  Missing_Count = missing_values
```

```
)

summary_table <- summary_table[order(summary_table$Missing_Count),]
summary_table
```

```
##          Variable Missing_Count
## clinic     clinic             0
## initdate initdate             0
## season     season             0
## gender     gender             0
## age           age             0
## exint0     exint0             2
## kcal0       kcal0             4
## grade       grade             6
## weight     weight             7
## sbp           sbp             7
## sbp140     sbp140             7
## diab         diab            17
## block0     block0            25
## arth         arth            33
## pkyrs       pkyrs            67
## weight50 weight50            88
## income     income           157
```

```
kable(summary_table)
```

	Variable	Missing_Count
clinic	clinic	0
initdate	initdate	0
season	season	0
gender	gender	0
age	age	0
exint0	exint0	2
kcal0	kcal0	4
grade	grade	6
weight	weight	7
sbp	sbp	7
sbp140	sbp140	7
diab	diab	17
block0	block0	25
arth	arth	33
pkyrs	pkyrs	67
weight50	weight50	88
income	income	157

b. Based on the table you made in part (a), do you think we will introduce any bias in our study if w

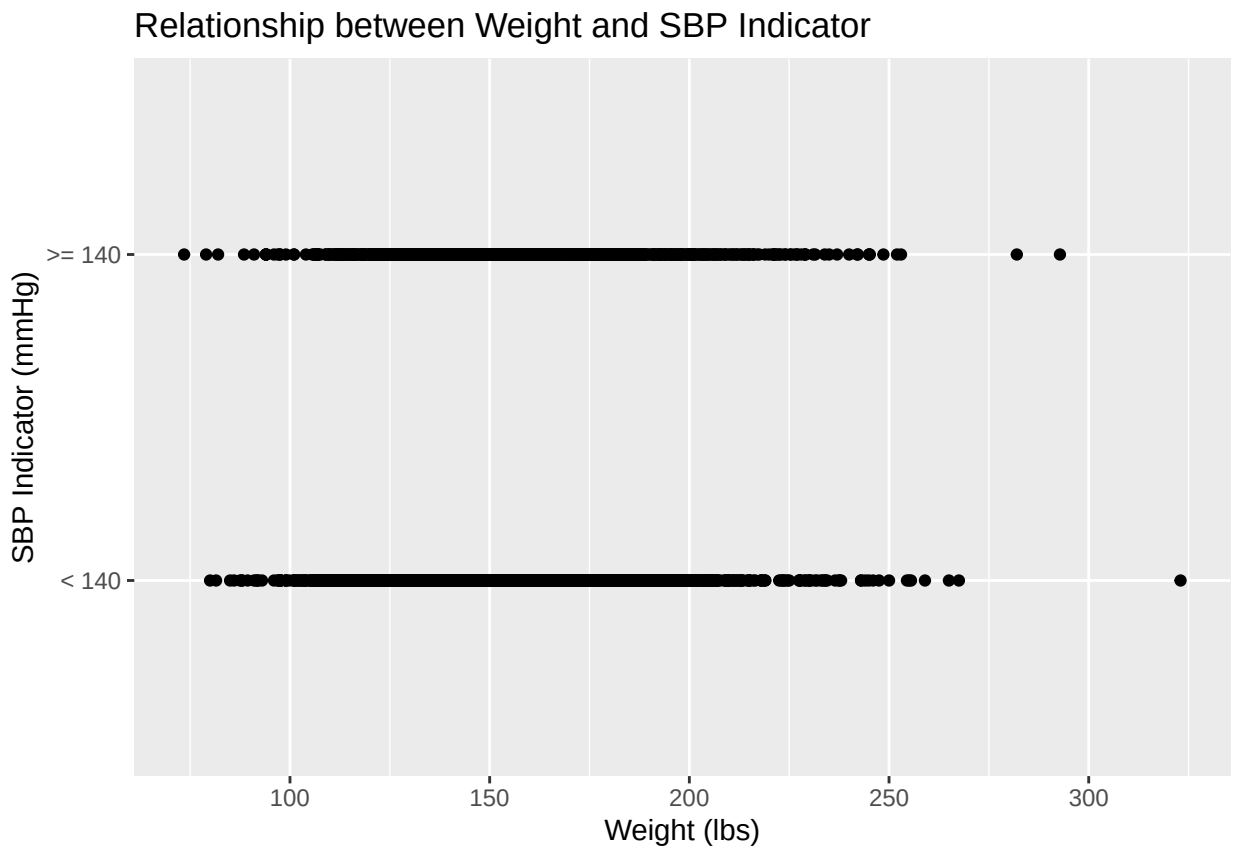
It can be seen that the missing values are not in some certain comunities's data. For some variables: exint0, kcal0, grade, weight, sbp, diab the number of their missing values is pretty small compared to 2440 observations. Therefore, removing their missing values may not cause bias because they are randomly missed and in small counts, so they are not representative of the whole population. However, for some variables with relatively higher missing counts: block0, arth, pkyrs, weight50 and income their missing values are from 25 to 157, they are likely to to introduce bias if we remove them.

Regardless of your answer to 2) b., remove any missing values.

3. Plot 1: Create a single (meaning only one) appropriate plot to show the relationship between your n

```r
chsData_filtered <- chsData %>%
  filter(!is.na(sbp140) & !is.na(weight) & !is.na(exint0))

# Create the plot
ggplot(chsData_filtered, aes(x = weight, y = sbp140)) +
  geom_point() +
  labs(x = "Weight (lbs)", y = "SBP Indicator (mmHg)", title = "Relationship between Weight and SBP Ind
```



4. List one potential confounder from the dataset and explain how it is both related to sbp140 and weig

exint0 : Baseline measure of exercise intensity could be a potential confounder.

- exint0 relates to sbp140: People who engage in regular exercise may have blood pressure reduction on average due to the positive effects of exercise on cardiovascular health.

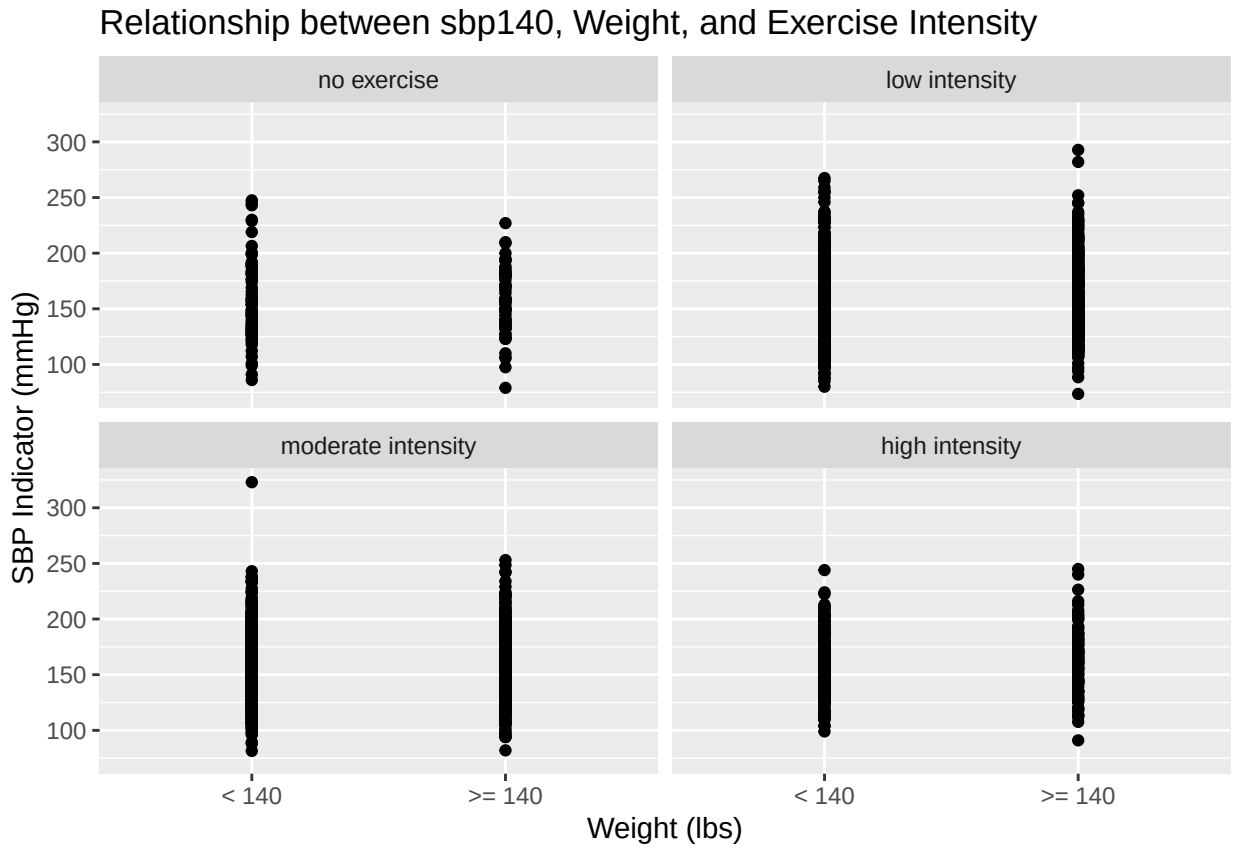
Reference: All about exercise and blood pressure - Kendall Reagan Nutrition Center. (2023, February 17). Kendall Reagan Nutrition Center. https://www.chhs.colostate.edu/krnc/monthly-blog/all-about-exercise-and-blood-pressure/#:~:text=Regular%20aerobic%20exercise%20results%20in,24%20hours%20after%20the%20activity!

- exint0 relates to weight: generally, if people do exercise regularly, they can gain weight loss or maintenance. The higher levels of exercise intensity the higher weight losses. So exint0 can affect weight.

Reference: Kerr, M. (2022, January 19). Exercise and weight loss. Healthline. https://www.healthline.com/health/exercise-and-weight-loss

5. Plot 2: Create a single appropriate plot to show the relationship between the high sbp indicator (sbp

```
# Create the plot to show the relationship between the high sbp indicator (sbp140) and weight (weight)
ggplot(chsData_filtered, aes(x = sbp140, y = weight)) +
  geom_point() +
  facet_wrap(~ exint0) +
  labs(title = "Relationship between sbp140, Weight, and Exercise Intensity",
       y = "SBP Indicator (mmHg)", x = "Weight (lbs)")
```



6. Descriptive statistics: Create the following table:

```
chsData$gender_f <- factor(chsData$gender,
                           labels = c("Female", "Male"))

# table(chsData$gender_f) / nrow(chsData)
table(chsData$gender_f, chsData$sbp140)
```

```
##
##          < 140 >= 140
##   Female   909    548
##   Male     608    368
```

```
props <- round(prop.table(table(chsData$gender_f, chsData$sbp140)),2)
props
```

```
##
##          < 140 >= 140
##   Female  0.37   0.23
```

```
##   Male    0.25   0.15
```

```
#Mean of weight for individuals with SBP < 140
mean_wt_lower_140 <- round(mean(chsData$weight[chsData$sbp < 140] , na.rm = TRUE),2)
mean_wt_lower_140
```

```
## [1] 156.74
```

```
#Standard deviation of weight for individuals with SBP < 140
sd_wt_lower_140 <- round(sd(chsData$weight[chsData$sbp < 140], na.rm = TRUE),2)
sd_wt_lower_140
```

```
## [1] 30.84
```

```
#Mean of weight for individuals with SBP >= 140
mean_wt_higher_140 <- round(mean(chsData$weight[chsData$sbp >= 140] , na.rm = TRUE),2)
mean_wt_higher_140
```

```
## [1] 158.71
```

```
#Standard deviation of weight for individuals with SBP >= 140
sd_wt_higher_140 <- round(sd(chsData$weight[chsData$sbp >=140], na.rm = TRUE),2)
sd_wt_higher_140
```

```
## [1] 31.47
```

```
#Number and % of Diabetes according to sbp categories
table(chsData$diab, chsData$sbp140)
```

```
##
##              < 140 >= 140
##   none        1197    638
##   borderline   184    133
##   diabetes     127    138
```

```
props_1 <- round(prop.table(table(chsData$diab, chsData$sbp140)),2)
props_1
```

```
##
##              < 140 >= 140
##   none        0.50   0.26
##   borderline  0.08   0.06
##   diabetes    0.05   0.06
```

```
#Mean of age for individuals with SBP < 140
mean_age_lower_140 <- round(mean(chsData$age[chsData$sbp < 140] , na.rm = TRUE),2)
mean_age_lower_140
```

```
## [1] 71.22
```

```
#Standard deviation of age for individuals with SBP < 140
sd_age_lower_140 <- round(sd(chsData$age[chsData$sbp < 140], na.rm = TRUE),2)
sd_age_lower_140
```

```
## [1] 4.75
```

```
#Mean of age for individuals with SBP >= 140
mean_age_higher_140 <- round(mean(chsData$age[chsData$sbp >= 140] , na.rm = TRUE),2)
mean_age_higher_140
```

```
## [1] 72.9
```

```r
#Standard deviation of age for individuals with SBP >= 140
sd_age_higher_140 <- round(sd(chsData$age[chsData$sbp >=140], na.rm = TRUE),2)
sd_age_higher_140
```

```
## [1] 5.46
```

Variable	SBP < 140 mmHg	SBP >= 140 mmHg
	mean(sd) or n (%)	mean(sd) or n (%)
Sex	Female: 909 (37%)	Female: 548 (23%)
	Male: 680 (25%)	Male: 368 (15%)
Weight	mean: 156.74 (sd: 30.84)	mean: 158.71 (sd: 31.47)
Diabetes	None: 1197 (50%)	None: 638 (26%)
	Borderline: 184 (8%)	Borderline: 133 (6%)
	Diabetes: 127 (5%)	Diabetes: 138 (6%)
Age	mean: 71.22 (sd: 4.75)	mean: 72.9 (sd: 5.46)

Part 3: Data Analysis

7. First, we want to know if the proportion of those with high sbp (i.e., sbp > 140 mmHg) is different from 50%. Test this by 1. writing the null and alternative hypothesis in symbols, 2. computing a 95% confidence interval, and 3. making a decision and concluding in the context of the problem. Assume conditions are met; you do not need to check. Include any R code used in this analysis.

Step 1: Hypotheses:

H_0: $p = 0.5$ H_A: $p \neq 0.5$

Step 2:

Computing a 95% confidence interval.

```r
table(chsData$sbp140)
```

```
##
##  < 140 >= 140
##   1517    916
```

```r
addmargins(table(chsData$sbp140))
```

```
##
##  < 140 >= 140    Sum
##   1517    916   2433
```

```r
table(chsData$sbp140) / nrow(chsData)
```

```
##
##     < 140    >= 140
## 0.6217213 0.3754098
```

```r
num_total <- 2440

prop.test(x = 916, n = 2440, p = 0.50, alternative = "two.sided")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  916 out of 2440, null probability 0.5
## X-squared = 151, df = 1, p-value < 2.2e-16
```

```
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3562039 0.3950116
## sample estimates:
##         p
## 0.3754098
```

Step 3: Making a decision and concluding in the context of the problem

Decision: Fail to reject H_0. . Conclusion: We do not have enough evidence to conclude that the proportion of those with high sbp (i.e., sbp $>$ 140 mmHg) is different from 50%.

`8. Do people with high sbp (>140 mmHg), on average, weigh more compared to those who have low sbp (<140`

1) Write the hypotheses.

$H_0 : \mu_{\geq 140} = \mu_{<140}$

$H_A : \mu_{\geq 140} > \mu_{<140}$

2) Computing the test statistic

```
xbar1 <- mean_wt_higher_140
xbar1
```

```
## [1] 158.71
```

```
xbar2 <- mean_wt_higher_140
xbar2
```

```
## [1] 158.71
```

```
s1 <- sd_wt_higher_140
s2 <- sd_wt_lower_140

n1 <- length(chsData$sbp140)
n2 <- length(chsData$sbp140)

stat <- xbar1 - xbar2
null_value <- 0
se <- sqrt(s1^2/n1 + s2^2/n2)

df <- min(c(n1, n2)) - 1


t_stat <- (stat - null_value) / se
t_stat
```

```
## [1] 0
```

```
# sbp_higher_140 <- chsData %>%
#   filter(sbp140 == ">= 140", !is.na(weight)) %>%
#   select(weight) %>%
#   pull()
#
# sbp_lower_140 <- chsData %>%
#   filter(sbp140 == "< 140", !is.na(weight)) %>%
#   select(weight) %>%
#   pull()
#
# t.test <- t.test(sbp_higher_140, sbp_lower_140, alternative = "greater",
```

```
#         conf.level = 0.95)
# t.test
```

3) Calculate p-value

```
p_val <- pt(t_stat, df = df, lower.tail = FALSE)
p_val
```

```
## [1] 0.5
```

4) Decision and conclusion in context:

Decision: Fail to reject H_0

Conclusion: We do not have evidence to conclude that people with high sbp (>140 mmHg), on average, weigh more compared to those who have low sbp (<140 mmHg).

```
9. Use the following code to create a new variable called weight_grp, which groups weight into 3 catego
```

```
chsData <- chsData %>%
  mutate(weight_grp = case_when(weight < 135 ~ "< 135 lbs",
                                weight >= 135 & weight <= 160 ~ "135-160 lbs",
                                TRUE ~ "> 160 lbs")) %>%
    mutate(weight_grp = factor(weight_grp))
```

Step 1: Hypotheses

H_0: There is no relationship between low/high systolic blood pressure (sbp140) and weight group (weight_grp)

H_A: There is a relationship between low/high systolic blood pressure (sbp140) and weight group (weight_grp)

Step 2: Check conditions

Independence: Assuming the participants researched were randomly selected.

Expected counts: 2440 observations > 5 (checked)

Step 3: Calculate p-value using R function

```
data <- table(chsData$sbp140, chsData$weight_grp)
data
```

```
##
##          < 135 lbs > 160 lbs 135-160 lbs
##   < 140        395       657         465
##   >= 140       219       413         284
```

```
test <- chisq.test(data)
test
```

```
##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 1.4603, df = 2, p-value = 0.4818
```

Step 4: Decision

p-value > 0.05. So, we fail to reject null hypotheses.

Step 5: Conclusion in context:

We do not have enough evidence to conclude that low/high systolic blood pressure (sbp140) and weight group (weight_grp) have a relationship.

Part 4: Discuss the Results

10. Based on the results above, do you think there is evidence that increased weight is associated with

As the results of question 8 and question 9 we can see that we do not have evidence to conclude that people with high sbp (>140 mmHg), on average, weigh more compared to those who have low sbp (<140 mmHg), additionally, we also do not have enough evidence to conclude that low/high systolic blood pressure (sbp140) and weight group (weight_grp) have a relationship. Since weight gained is unlikely to introduce higher sbp, besides blood pressure has no relationship with weight, so there is not enough evidence to conclude that increased weight is associated with increased blood pressure.