STAT 630: Homework 6

Due: November 4th

```
library(dplyr)
library(ggplot2)
library(gtsummary)

chs <- read.csv(here::here("data/chsData.txt"), sep = "")
```

Part 0: Literature Review

1. Find one or two reputable resources (link them in your document) that explain the relationship between weight and systolic blood pressure. In a few sentences, describe the findings of the article.

Answers will vary. I expect you found that increased weight is associated with an increase in blood pressure.

2. Based on your findings, what do you believe is the relationship between weight and SBP?

As weight increases, SBP increases.

Part 1: Data Cleaning

3. 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, exint0). Make sure to rename the levels of these factors to match the description in the table above.

```
chs <- chs %>%
  rename(sex = gender) %>%
  mutate(clinic = factor(
    clinic, labels = c("Sacremento", "Forsyth",
                       "Washington", "Pittsburgh")),
    season = factor(season,
                         labels = c("Summer", "Fall", "Winter", "Spring")),
    sex = factor(sex, labels = c("Female", "Male")),
    arth = factor(arth, labels = c("None", "Arthritis")),
    diab = factor(diab, labels = c("None", "Borderline", "Diabetes")),
    income = factor(income,
                    labels = c("<=5", "5-8", "8-12", "12-16",
                               "16-24", "24-35", "35-50", ">=50")),
    exint0 = factor(exint0,
                    labels = c("no exercise", "low intensity",
                               "moderate intensity", "high intensity")))
```

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

```r
chs <- chs %>%
  mutate(sbp140 = factor(ifelse(chs$sbp < 140, "<140mmHg", ">=140mmHg")))
```

Part 2: Exploratory Data Analysis

4. Missing Values:

   a. Make a publication-quality table that shows the number of missing rows for each variable in the dataset.

```r
miss <- sapply(chs[,-17], function(x) sum(is.na(x)))
miss <- as.table(miss)

rownames(miss) <- c("Clinic", "Initial Date", "Season", "Sex", "Age",
                    "Weight", "Weight50", "Grade", "Arthritis",
                    "SBP", "Pack Years", "Diabetes", "Income",
                    "Exercise", "Blocks",
                    "Kilocalories")
knitr::kable(miss)
```

Var1	Freq
Clinic	0
Initial Date	0
Season	0
Sex	0
Age	0
Weight	7
Weight50	88
Grade	6
Arthritis	33
SBP	7
Pack Years	67
Diabetes	17
Income	157
Exercise	2
Blocks	25
Kilocalories	4

   b. Based on the table you made in part (a), do you think we will introduce any bias in our study if we remove these missing values? Explain.
   The maximum percentage of missingness is 6% for the income variable. Since all variables have less than 10% missingness, there should not be too much worry about bias.

Note: will accept most reasonable answers here.

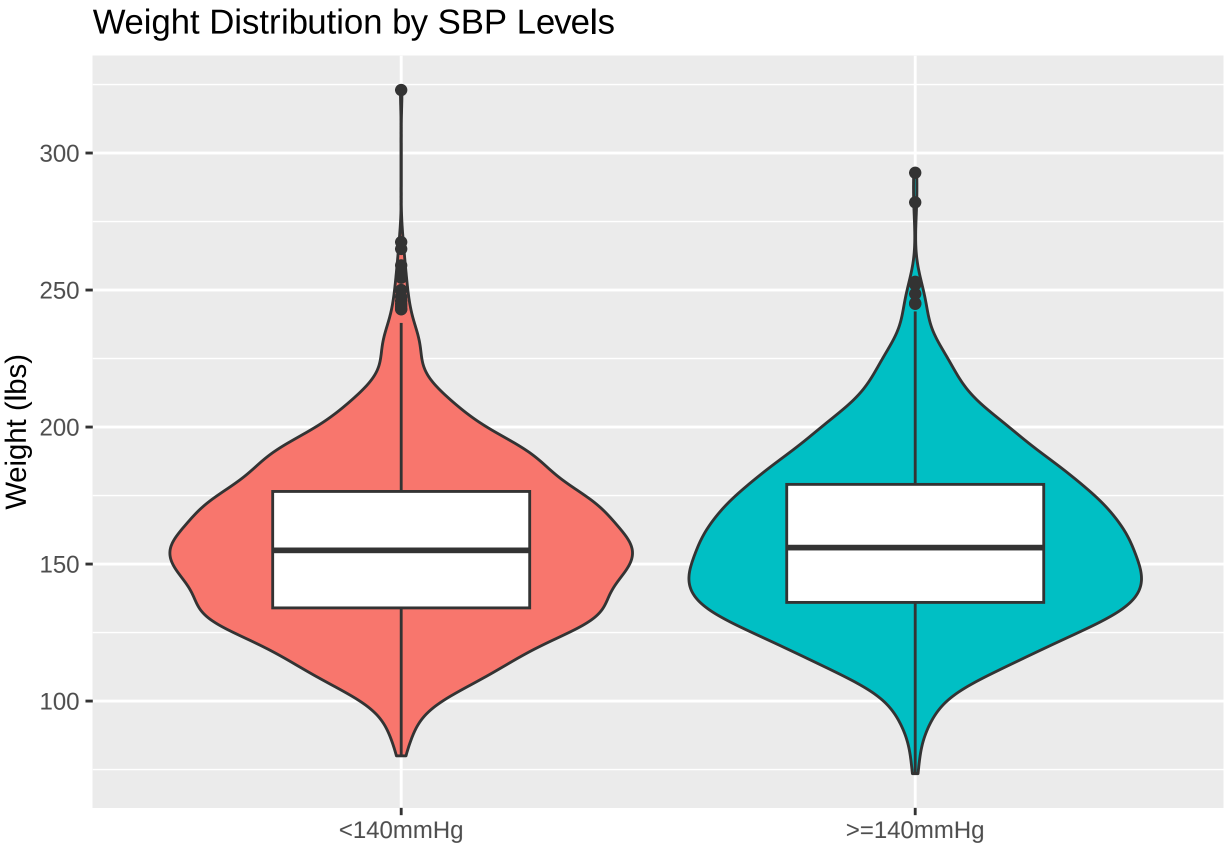
Regardless of your answer to 2) b., remove any missing values.

I later regretted having this component in the homework. I will show both removing and not removing rows with missing data.

```
chs_miss <- na.omit(chs)
```

5. Plot 1: Create a single (meaning only one) appropriate plot to show the relationship between your new high sbp indicator (sbp140) and weight (weight). You may use any plotting functions from any R package. Be sure to include proper x and y-axis labels and a title.

```
chs %>%
  filter(!is.na(sbp140)) %>%
ggplot(aes(x = sbp140, y = weight)) +
  geom_violin(aes(fill = sbp140),
              show.legend = FALSE) +
  geom_boxplot(width = 0.5,
               show.legend = FALSE) +
  labs(x = NULL,
       y = "Weight (lbs)",
       title = "Weight Distribution by SBP Levels")
```



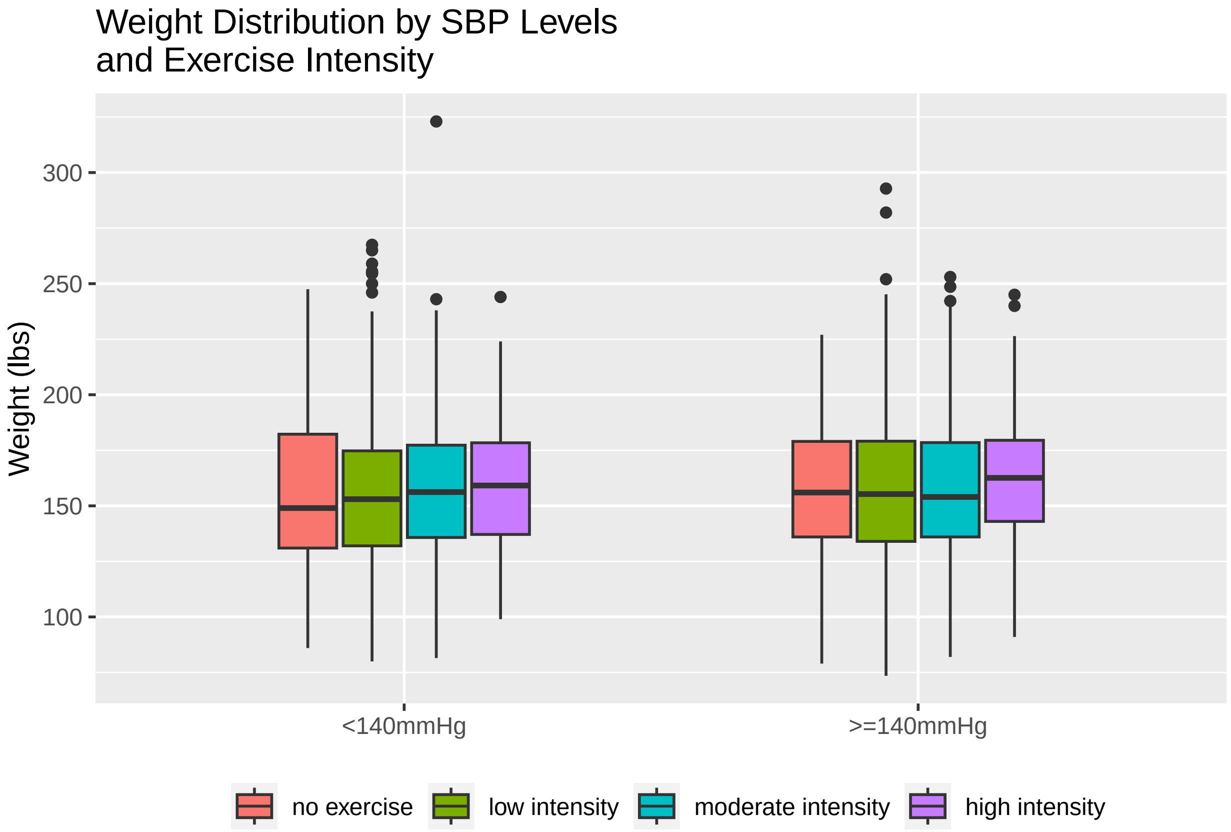
6. List one potential confounder from the dataset and explain how it is both related to sbp140 and weight.

Many of the variables in the dataset are possible confounders. For example,

Exercise intensity could be a confounder. The amount a person exercises is both related to weight (in general (not always), more exercise = lower weight) and to blood pressure (more exercise = lower blood pressure).

7. Plot 2: Create a single appropriate plot to show the relationship between the high sbp indicator (sbp140) and weight (weight), and the confounding variable you chose in Question 4, i.e., your plot should include 3 variables from the dataset. You may use any plotting functions. Be sure to include proper x and y-axis labels and a title.

```r
chs %>%
  filter(!is.na(sbp140), !is.na(exint0)) %>%
  ggplot(aes(x = sbp140, y = weight, fill = exint0)) +
  geom_boxplot(width = 0.5) +
  theme(legend.position = "bottom") +
  labs(x = NULL,
       y = "Weight (lbs)",
       fill = NULL,
       title = "Weight Distribution by SBP Levels \nand Exercise Intensity")
```



8. Descriptive statistics: Create the following table:

	SBP <140 mmHg	**SBP >=140 mmHg**
	mean(sd) or n(%)	**mean(sd) or n(%)**
Sex		
Weight		
Diabetes		
Age		

```r
chs %>%
  select(sbp140, sex, weight, diab, age) %>%
  tbl_summary(by = sbp140,
              statistic = all_continuous() ~ "{mean} ({sd})",
              digits = all_continuous() ~ c(2,2),
              label = list(
                sex = "Sex",
                weight = "Weight",
                diab = "Diabetes",
                age = "Age"
              ))
```

Characteristic	**<140mmHg**, N = 1,517	**>=140mmHg**, N = 916
Sex		
Female	909 (60%)	548 (60%)
Male	608 (40%)	368 (40%)
Weight	156.74 (30.84)	158.71 (31.47)
Unknown	4	2
Diabetes		
None	1,197 (79%)	638 (70%)
Borderline	184 (12%)	133 (15%)
Diabetes	127 (8.4%)	138 (15%)
Unknown	9	7
Age	71.22 (4.75)	72.90 (5.46)

```
chs_miss %>%
  select(sbp140, sex, weight, diab, age) %>%
  tbl_summary(by = sbp140,
              statistic = all_continuous() ~ "{mean} ({sd})",
              digits = all_continuous() ~ c(2,2),
              label = list(
                sex = "Sex",
                weight = "Weight",
                diab = "Diabetes",
                age = "Age"
              ))
```

Characteristic	**<140mmHg**, N = 1,287	**>=140mmHg**, N = 776
Sex		
Female	763 (59%)	455 (59%)
Male	524 (41%)	321 (41%)
Weight	156.84 (30.20)	158.84 (31.34)
Diabetes		
None	1,023 (79%)	548 (71%)
Borderline	159 (12%)	114 (15%)
Diabetes	105 (8.2%)	114 (15%)
Age	71.20 (4.70)	72.83 (5.36)

Part 3: Data Analysis

You may use any R functions for the problems below. You do not need to show formulas or do any of the work "by hand".

9. 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.

1) Hypotheses:

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

2) Confidence Interval
$\hat{p} \pm z^* \sqrt{\hat{p}(1-\hat{p})/n}$

```
# With full dataset
table(chs$sbp140)
```

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

```
n <- 2433
prop.test(916, 2433, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  916 out of 2433, null probability 0.5
## X-squared = 148.46, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3574468 0.3959224
## sample estimates:
##         p
## 0.3764899
```

```
phat <- 916 / n
phat + c(-1,1) * qnorm(0.975) * sqrt(phat*(1-phat)/ n)
```

```
## [1] 0.3572379 0.3957419
```

```
# With na.omit()
prop.test(776, 2063, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  776 out of 2063, null probability 0.5
## X-squared = 126.57, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3554961 0.3972668
## sample estimates:
##         p
## 0.3761512
```

```r
n_miss <- 2063

phat_miss <- 776 / n_miss

phat_miss + c(-1,1) * qnorm(0.975) * sqrt(phat_miss*(1-phat_miss)/ n_miss)
```

```
## [1] 0.3552477 0.3970548
```

CI = (0.3574468, 0.3959224) with prop.test()

CI = (0.3572379, 0.3957419) by hand

CI (na) = (0.3554961, 0.3972668) with prop.test()

CI (na) = (0.3552477, 0.3970548) by hand

3) Decision and Conclusion

Since 0.5 is not within the 95% confidence interval, we have enough evidence that the true proportion of people with high SBP (SBP >= 140 mmHg) is not 0.5.

10. Do people with high sbp ($>140*mmHg*), *onaverage*, *weighmorecomparedtothosewhohavelowsbp*(<$140 mmHg)? Test this by 1. writing the null and alternative hypothesis in symbols, 2. computing the test statistic and p-value, and 3. writing a conclusion in the context of the problem. **Assume conditions are met; you do not need to check. Include any R code used in this analysis.**

1) Hypotheses:

H_0: $\mu_1 - \mu_2 = 0$
H_A: $\mu_1 - \mu_2 > 0$ *where 1 = High SBP and 2 = Low SBP*
OR H_A: $\mu_1 - \mu_2 < 0$ *where 1 = Low SBP and 2 = High SBP*

2) Check conditions: skip
3) Test Statistic and p-value:

```r
grp1 <- chs %>% filter(sbp140 == ">=140mmHg") %>% select(weight)
grp2 <- chs %>% filter(sbp140 == "<140mmHg") %>% select(weight)

high_low <- t.test(grp1, grp2, alternative = "greater")

grp1_miss <- chs_miss %>% filter(sbp140 == ">=140mmHg") %>% select(weight)
grp2_miss <- chs_miss %>% filter(sbp140 == "<140mmHg") %>% select(weight)

high_low_miss <- t.test(grp1_miss, grp2_miss, alternative = "greater")
```

	1 = High, 2 = Low
Test Statistic	1.5037089
p-value	0.0664115
Test Statistic (na)	1.4226069
p-value (na)	0.0775234

Decision and Conclusion:

Fail to reject H_0*: We do not have enough evidence that people with high sbp (*$\geq$ *140 mmHg), on average, weigh more compared to those who have low sbp ($<$140 mmHg).*

11. Use the following code to create a new variable called weight_grp, which groups weight into 3 categories: $<$ 135 lbs, 135-160lbs, and $>$ 160 lbs.

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

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

Now, test whether or not there is a relationship between low/high sbp (sbp140) and weight group (weight_grp). Perform a full 5-step hypothesis test, including checking conditions.

1) Hypotheses:

H_0: *weight group and sbp are independent*
H_A: *weight group and sbp are not independent*

2) Conditions:
1. Independence: Assume participants were randomly selected
2. Large counts:

```
tab <- table(chs$sbp140, chs$weight_grp)
test <- chisq.test(tab)
test$expected
```

```
##
##              < 135 lbs > 160 lbs 135-160 lbs
##   <140mmHg   382.8352  667.1558     467.009
##   >=140mmHg  231.1648  402.8442     281.991
```

```
tab_miss <- table(chs_miss$sbp140, chs_miss$weight_grp)
test_miss <- chisq.test(tab_miss)
test_miss$expected
```

```
##
##              < 135 lbs > 160 lbs 135-160 lbs
##   <140mmHg   323.7775   565.207    398.0155
##   >=140mmHg  195.2225   340.793    239.9845
```

All counts are ≥ 5.

3) Test Statistic:

$\chi^2_2 = 1.460287$
χ^2_2 (na) $= 2.2339452$

4) P-value:
p-value $= 0.4818399$
p-value (na) $= 0.3272691$

5) Decision and Conclusion:
Fail to reject H_0*: We do not have enough evidence that sbp and weight group are not independent.*

Part 4: Discuss the Results

12. Based on the results above, do you think there is evidence that increased weight is associated with increased blood pressure? State yes or no, and explain how you came to this decision using the results of your tests from above.

Both the independent means t-test and the chi-squared test for independence were not significant, meaning that we could not find evidence that those with high sbp weigh more or that there is an association between weight (group) and high sbp, respectively. Thus, there does not seem to be evidence that increased weight is associated with increased blood pressure. To test this directly, we could create simple linear regression model (coming soon!).