HW 6

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, i.e., 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.

The aim of this analysis is to investigate the association between the response, systolic blood pressure greater than 140 mmHg, and the predictor of interest weight.

Variable	Description
`clinic`	1 = Sacramento, 2 = Forsyth, 3 = Washington, 4 = Pittsburgh
`initdate`	Date of recruitment (in days, measured from an arbitrary starting date)
`season`	Season at baseline, 1 = summer, 2 = fall, 3 = winter, 4 = spring
`gender`	0 = Female, 1 = Male
`age`	Age at baseline (in years)
`weight`	Weight at baseline (in lbs)
`weight50`	Recalled weight at 50 years old (in lbs)
`grade`	Years of education
`arth`	1 = arthritis, 0 = none (at baseline)
`sbp`	Baseline systolic blood pressure (in mmHg)
`pkyrs`	Pack years of smoking history (Number of years as smoker $\times$ number of packs/day $\times$ 365
`diab`	Diabetes at baseline, 1 = none, 2 = borderline, 3 = diabetes
`income`	Household income (in 1k dollars): 1 = $\leq$ 5k, 2 = 5k-8k, 3 = 8k-12k, 4 = 12k-16k, 5 = 16k-24k, 6 = 24k-35k, 7 = 35k-50k, 8 = $\geq$ 50k
`exint0`	Baseline measure of exercise intensity, 0 = no exercise, 1 = low intensity, 2 = moderate intensity, 3 = high intensity
`block0`	Baseline measure of blocks walked in last 2 weeks (at about 12 blocks/mile)
`kcal0`	Baseline measure of estimated kilocalories expended in exercise activity in past 2 weeks

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

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

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.
    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.

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

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

8. Descriptive statistics: Create the following table:

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

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.
10. Do people with high sbp (≥140 mmHg), on average, weigh more compared to those who have low sbp (<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.
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. *Be careful with copy/paste. Sometimes the quotes do not paste well in R.*

```
Unset
chs <- chs %>%
  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))
```

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.

Part 4: Discuss the Results

12. Based on the results above, did you find evidence for or against your hypothesized relationship from Question 2? State yes or no, and explain how you came to this decision using the results of your tests from above in complete sentences.