Project 1

1. [2 marks] The first part of our PPDAC framework is to identify the problem you are addressing with these data. State the question you are trying to answer and let us know what type of question this is in terms of the PPDAC framework. A question statement should be as specific as possible. For example: Do students who regularly get 8 hours of sleep have fewer visits to the health center? This question is an example of an etiologic or causal question.

The question we are trying to answer is: Does the percentage of healthy food stores in a certain radius vary by county and what factors (race, income) contribute to that pattern? This question is an example of a descriptional question.

2. [2 marks] Why is this question interesting or important? You could talk here about how existing data/studies suggest this might be important, how the findings might make an impact, how the findings might be used, or why you are personally interested in this question.

This question is interesting as public health students because to address public health issues such as heart attack, kidney failure, and osteoporosis, we must first make sure that everyone is given the same opportunity to buy groceries and healthy foods from supermarkets that provide these. We hope that our study can provide insight into how health issues related to food may be caused by biased city planning by the government.

3. [2 marks] What is the target population for your project? Why was this target chosen? (i.e., what was your rationale for wanting to answer this question in this specific population?)

Our target population is California residents. We wanted to learn more about how California residents' access to healthy food varies by county, because even within Berkeley we've observed unequal access to healthy food. We were curious whether access to healthy food varies on a larger scale as well.

4. [2 marks] What is the sampling frame used to collect the data you are using? It may be helpful here to read any protocol papers, trial registration records, '.Readme' files or documentation that are associated with your dataset. If you have trouble identifying how the records/individuals were sampled, confirm with your supporting GSI that your dataset will be usable for the purposes of the class. Describe why you think this sampling strategy is appropriate for your question. To what group(s) would you feel comfortable generalizing the findings of your study and why?

From the dataset description, it says: "Dun and Bradstreet geocoded data on the locations of food retailers in California was obtained via an internal data user agreement with the California Department of Public Health, Nutrition Education and Obesity Prevention Branch (NEOPB). Methodology developed by the Centers for Disease Control and Prevention (CDC) was used to calculate the modified retail food environment index, with some modifications." These data show the ratio of healthy food stores to unhealthy food stores on the region, county, census tract, and town level. This strategy is appropriate for our question because it shows the availability of healthy foods to different communities in California. The findings of this study cannot really be generalized because it is looking at geographical location as a main variable in the results.

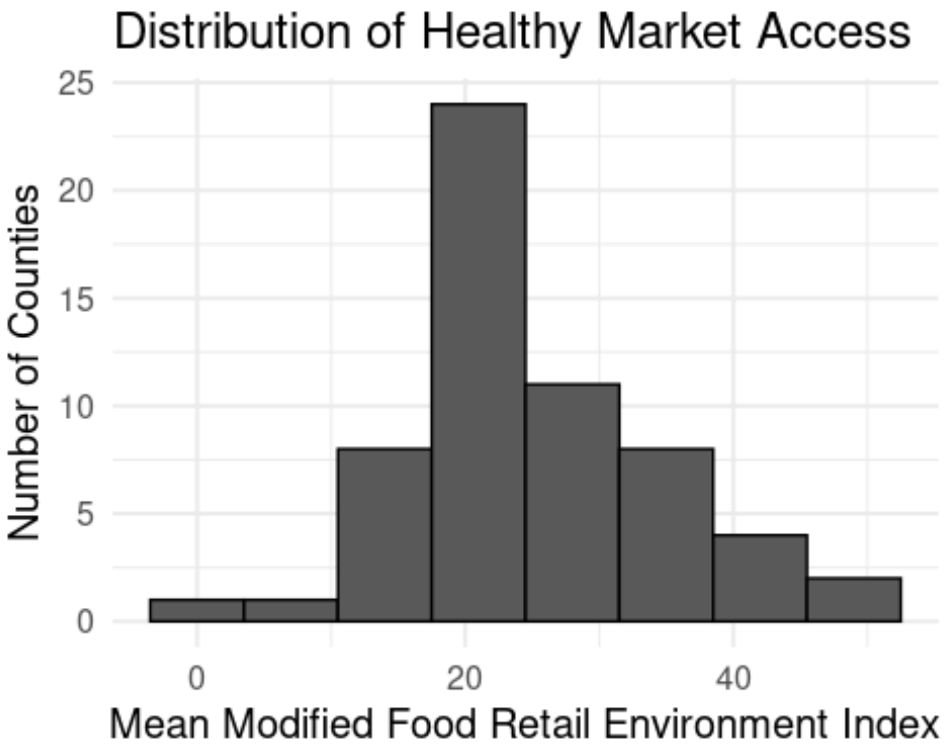
5. [2 marks] Write a brief description (1-4 sentences) of the source and contents of your dataset. Provide a URL to the original data source if applicable. If not (e.g., the data came from your internship), provide 1-2 sentences saying where the data came from. If you completed a web form to access the data and selected a subset, describe these steps (including any options you selected) and the date you accessed the data.

We searched up datasets related to supermarkets and healthy food access on California Health and Human Services because this site was recommended on the dataset criteria document. The dataset can be found at this link: https://data.chhs.ca.gov/dataset/modified-retail-food-environment-index

This dataset has the Modified Retail Food Environment Index, which is the percentage of healthy supermarkets within a certain radius for each region of California during the year of 2017.

```
#6. Write code below to import your data into R. Assign your dataset to an
object. Make sure to include and annotate this code in your submission (you
can use a # to comment out regular text within code chunks to annotate).

library(readxl) #loads functions necessary to read an excel file
library(dplyr) #loads functions necessary to clean up the dataset

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

#loads excel file into R and assigns it to an object and uploads excel file
into project folder
hciretailfoodenvironment75cacoreplct201715nov17 <-
read_excel("project/hciretailfoodenvironment75cacoreplct201715nov17.xlsx")
supermarkets <- hciretailfoodenvironment75cacoreplct201715nov17
#assigns dataset to an object

# Cleaning up the dataset before question 7
supermarkets2 <- select(supermarkets,"county_name","estimate") %>%
  #Selecting specific string variables (county_name and estimate) and
assigning to a new variable
```

```
  filter(estimate != "NA") %>% group_by(county_name) %>%
  #function is filtering out when estimate is equal to NA and grouping rows
by county name
  summarize(mean_estimate = mean(estimate))
  #function is totaling the mean estimate for each county

#7. Write code in R (included in your submission with annotation) to answer
the following questions: i) What are the dimensions of the dataset? ii) What
are the variable names of the variables in your dataset? iii) Print the first
six rows of the dataset

dim(supermarkets2)

## [1] 59  2

#function returns the dimensions of the cleaned dataset
names(supermarkets2)

## [1] "county_name"   "mean_estimate"

#function returns the names of the variables in the cleaned dataset
head(supermarkets2)

## # A tibble: 6 x 2
##   county_name mean_estimate
##   <chr>               <dbl>
## 1 Alameda              20.1
## 2 Alpine                0
## 3 Amador               21.2
## 4 Butte                30.0
## 5 Calaveras            17.0
## 6 Colusa               46.5

#function returns the first six rows of the cleaned dataset

#8. Use the data to demonstrate a data visualization skill we have covered
during Part I of the course. Choose a visualization relevant to your stated
problem. Include your code in your submission. For example, you could
visualize the distribution of our outcome with a histogram, or use a bar
graph to represent the distribution of your exposure variable.

library(ggplot2)
ggplot(data = supermarkets2, aes(x= mean_estimate)) +
         geom_histogram(col="black", binwidth= 7) +
         labs(y= "Number of Counties", x= "Mean Modified Food Retail
Environment Index", title = "Distribution of Healthy Market Access") +
         theme_minimal(base_size = 15)
```



9. [2 marks] Describe the skill that you are demonstrating and interpret your findings. For example, if you have created a histogram, describe the central tendency, shape of the distribution, etc.

We are making a histogram and demonstrating that we can visualize the distribution of modified retail food environment indexes by county in California. The histogram is unimodal, with a peak around 20, and left-skewed, meaning the mean is less than the median.

10. [1 mark] Include your work for Part I.

11. [2 marks] Calculate a marginal probability based on your outcome variable. Provide an equation (using probability notation) that describes this probability. For example, if my outcome variable is height in inches, I might calculate the probability that an individual in the dataset has a height of greater than 60 inches. P(height ≥ 60) =?. This would be a marginal probability. You may need to first add a new variable to your dataset to calculate your probability of interest, such as a binary variable indicating whether height is greater than 60 inches. There is a resource video about how to code such variables that could be helpful!

```
supermarkets2 <- supermarkets2[-c(59),]
supermarkets2$GoodModifiedRetailFoodIndex <-
ifelse(supermarkets2$mean_estimate>16.77, "Yes", "No")
Bad<-supermarkets2 %>% filter(GoodModifiedRetailFoodIndex== "No")
Good<-supermarkets2 %>% filter(GoodModifiedRetailFoodIndex== "Yes")
```

```
count(Bad)/(count(Bad)+count(Good))

##          n
## 1 0.137931
```

#p(index< 16.77) = 8/51 = 0.14

12. [2 marks] Using any two variables in your dataset (or derived variables), calculate a conditional probability. Provide an equation (using probability notation) that describes this probability and then use R to calculate it.

```
supermarkets4<- supermarkets2 %>% filter(county_name %in% c("Alameda",
"Solano", "Los Angeles" , "Santa Clara", "San Francisco", "San Mateo",
"Contra Costa", "San Joaquin", "Sacramento", "Merced" ))
supermarkets4$GoodModifiedRetailFoodIndex <-
ifelse(supermarkets4$mean_estimate>16.77, "Yes", "No")

No <- supermarkets4 %>% filter(GoodModifiedRetailFoodIndex== "No")
#dataset of diverse counties with index less than 16.77
Yes <- supermarkets4 %>% filter(GoodModifiedRetailFoodIndex== "Yes")
#dataset of diverse counties with index greater than 16.77

count(No)/count(Yes)

##      n
## 1 0.25
```

#p(index<16.77 | diverse county) = 0.25

13. [2 marks] Does your dataset contain a continuous variable? If it does, does the distribution of that variable appear to be normal? Justify your answer using a plot. If your data does not contain a continuous variable, give an example related to your dataset of a hypothetical variable that is continuous. That is, imagine what a continuous variable could be in relation to your dataset and topic of interest. For this hypothetical variable, describe what you imagine its shape might be, and how you would check whether or not it is normally distributed.

#There is a continuous variable being the modified retail food environment index which is based on ratios of healthy food stores to food stores in general and other factors that modify it. It is a sliding scale type of measurement rather than a discrete one. #Created a histogram of the modified retail food index across counties. It does not appear to be normal as it skews right.

14. [4 marks] Does your dataset contain a binary variable? If so, does this variable meet the criteria to be considered binomially distributed? If so, describe this variable in terms of n and p. Calculate a probability based on this variable, first write the formula for the probability and then using R to calculate the probability (you do not need to calculate the probability by hand). If your data does not contain a binary variable, you can create one based on an underlying continuous variable or a categorical variable with > 2 levels to answer this question.

#No, our data set does not contain a binary variable, so it does not meet the criteria to be binomially distributed. However, we can create a binary variable by measuring whether or not a county has a good modified retail food index, or having a mean estimate larger than 16.7. The variable does meet the criteria to be binomially distributed because there are a fixed number of observations, each observation is independent of one another and has the same probability of occurring, and each observation is either a success or failure. The variable has a sample size of 58 counties (n=58) with a probability of a county having a bad/low modified retail food index of 13.793% occurring (p=0.13793). By using the formula, pbinom(q= 17, size= 58, prob= .13793), in R, it prints a probability of 0.9994246

```
supermarkets2 <- supermarkets2[-c(59),]

supermarkets2$GoodModifiedRetailFoodIndex <-
ifelse(supermarkets2$mean_estimate>16.77, "Yes", "No")

pbinom(q= 17, size= 58, prob= .13793)

## [1] 0.9994246
```

16. [1 mark] Include parts I and II of your project.

17. [2 marks] Identify a statistical test to apply to your data. This must be a statistical test that we cover in part III of the course. Name the statistical test you have chosen and explain why this is the appropriate test for these data. For example, if I have pre- and post-intervention measurements of morning sleepiness recorded as a quantitative variable, I might choose a paired t test, because the paired t-test is appropriate for continuous outcome data in 2 groups that are inherently related.

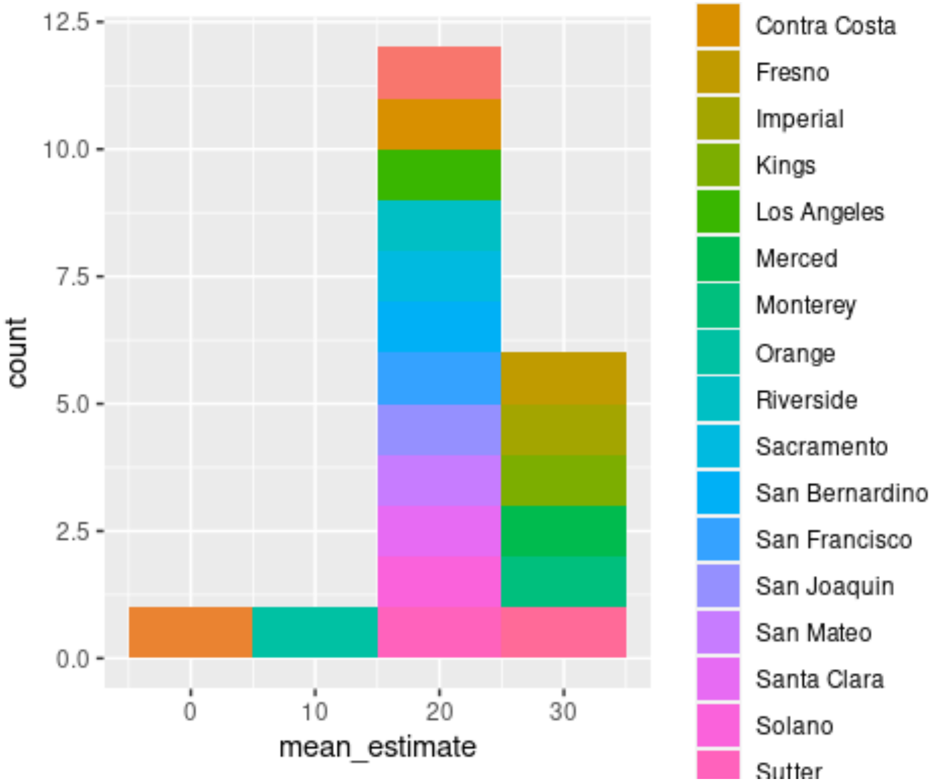
#We would use a two sample t-test to compare the means of the estimates (food modified retail indexes) of the most diverse counties and the means of the estimates of the least diverse counties in California. These groups represent distinct populations of California based on County and the estimates are independent of one another. Thus the two sample t-test is appropriate to measure the outcome data between these two groups.

18. [2 marks] What assumptions are required by the testing method you chose? Are these assumptions met by your data? How did you assess this? For example, one of the assumptions of the t-test is that the data are normally distributed, so you might choose to assess this with a histogram, or a q-q plot.

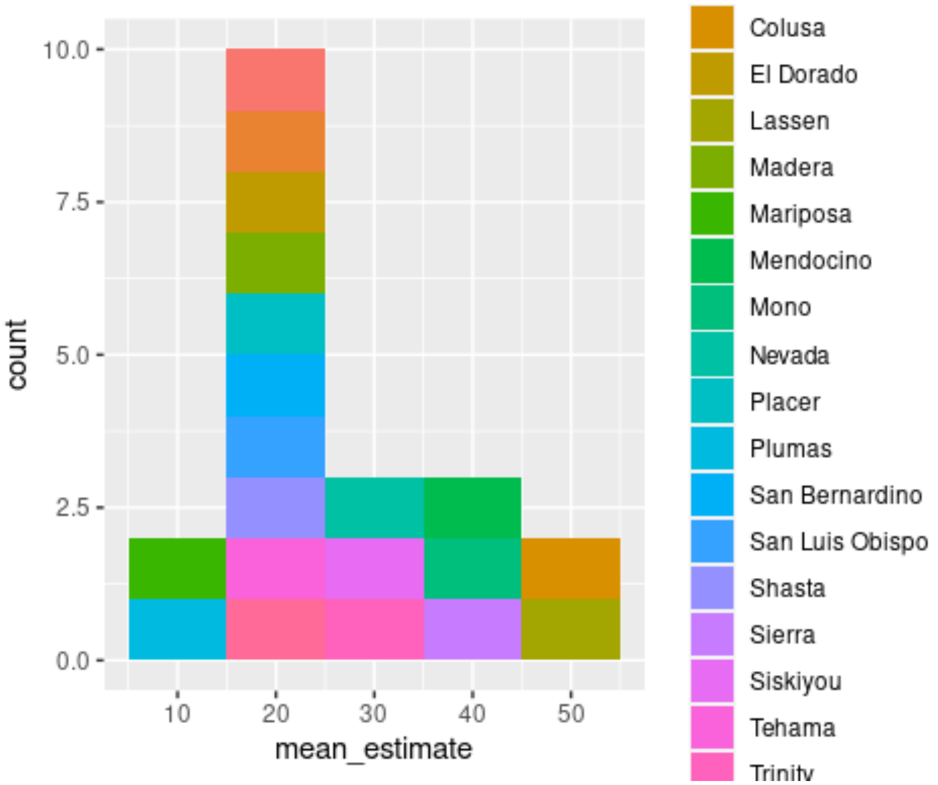
#We have two SRSs representing two distinct populations based on diversity. These samples are independent of one another and measure the same quantitative variable. Both populations are normally distributed, as shown in both histograms, and display similar shapes and have no strong outliers.

```
Most_Diverse_California <- slice(supermarkets2, c(1,48, 19, 43, 38, 41, 39,
7, 34, 24, 27, 33, 36, 30, 10, 2, 16, 13, 57, 51))

Least_Diverse_California <- slice(supermarkets2, c(55, 46, 32, 29, 22, 20, 5,
6, 9, 3, 45, 53, 52, 40, 47, 23, 26, 18, 31, 36))
```

```
ggplot(Most_Diverse_California, aes(x= mean_estimate)) +
  geom_histogram(aes(fill = county_name), binwidth = 10)
```



```
ggplot(Least_Diverse_California, aes(x= mean_estimate)) +
  geom_histogram(aes(fill = county_name), binwidth = 10)
```



19. [2 marks] Clearly state the null and alternative hypotheses for your test.

#Null Hypothesis: There is no significant difference in mean modified retail food index between the most diverse counties and the least diverse counties, other than due to random chance.

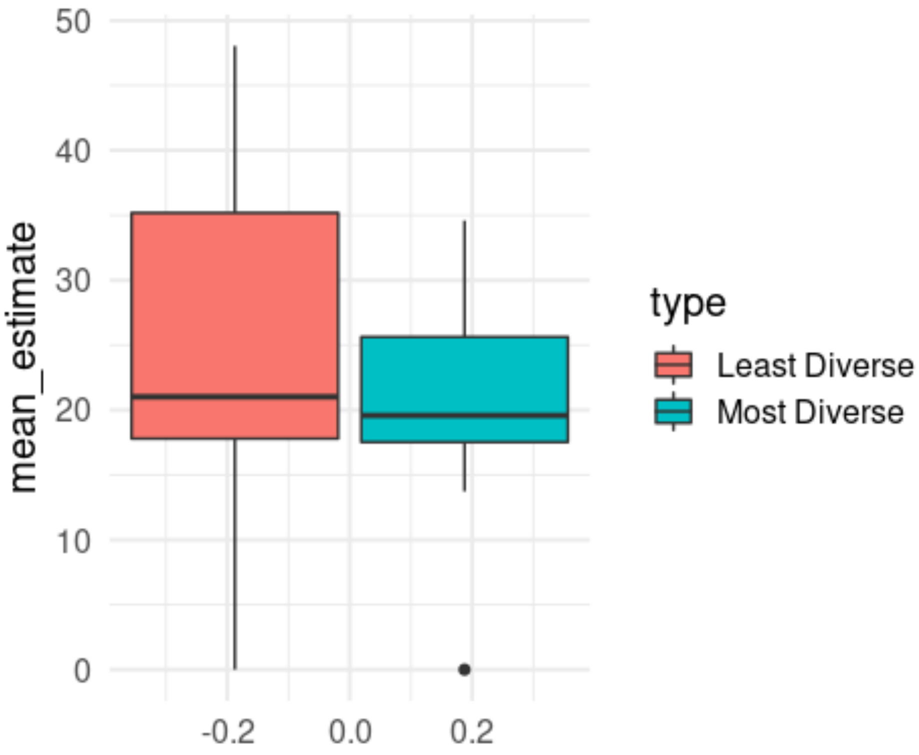
#Alternative Hypothesis: There is a difference in mean modified retail food indexes between the most diverse counties and the least diverse counties in California due to factors other than random chance.

```
#20. [2 marks] Conduct the statistical test. Include the R code you used to
generate your results. Annotate your code to help us follow your reasoning.

Most_Diverse_California_1 <- select(Most_Diverse_California, "mean_estimate")
#We are using the new data set that contains the 20 most diverse counties in
California by slicing them out from a previous data set and then only
including the mean estimate values previously calculated

Least_Diverse_California_1 <- select(Least_Diverse_California,
"mean_estimate")
#We are using the new data set that contains the 20 least diverse counties in
California by slicing them out from a previous data set and then only
including the mean estimate values previously calculated

t.test(Most_Diverse_California_1, Least_Diverse_California_1, alternative =
"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  Most_Diverse_California_1 and Least_Diverse_California_1
## t = -1.4872, df = 32.462, p-value = 0.1466
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.983121   1.710247
## sample estimates:
## mean of x mean of y
##  20.98927  25.62570

#Here we are using the t.test function to conduct a two sample t test between
our two created data sets
```

21. [4 marks] Present your results in a clear summary. This should include both a text summary and a table or figure with appropriate labeling. For example, if your outcome and predictor/exposure variables are both binary, this might be a 2x2 table. If your method was regression, you might present your regression line graphically. Include your code and annotations.

```
Least_Diverse_California_2 <-
c(16.063606,38.888889,11.050864,25.462444,9.739583,23.002568,17.015554,46.476
15,.005158,21.209406,20.821961,25.777778,20.395898,20.279270,33.953340,41.662
603,40.482804,48.055541,19.112513,18.057317)
Most_Diverse_California_2 <-
c(20.13551,16.64193,19.35193,17.19568,18.78477,19.80336,24.99949,16.07761,18.
95512,34.59156,26.61578,17.63823,18.05732,13.73114,26.20311,0.00000,25.44057,
33.34816,28.82003,23.39399)

Diversity_Data <- data.frame(mean_estimate= c(Least_Diverse_California_2,
Most_Diverse_California_2), type= c(rep("Least Diverse", 20), rep("Most
Diverse", 20)))

ggplot(Diversity_Data, aes(y=mean_estimate)) +
  geom_boxplot(aes(fill=type)) +
  theme_minimal(base_size = 15)
```



#Our t-statistic found from the t-test is -1.4872 and our 95% Confidence interval was found to be from -10.983121 and 1.710247. It was also found that our df value or degrees of freedom was 32.462 and our calculated p-value was 0.1466. Furthermore, our estimated mean difference is about -4.643 #We created two box plots for the Least Diverse and Most Diverse datasets anbd we found that values vary more in the Least Diverse data set.

22. [4 marks] Interpret your findings. Include a statement about the evidence, your conclusions, and the generalizability of your findings. Our analyses and conclusions depend on the quality of our study design and the methods of data collection. Any missteps or oversights during the data collection process could potentially change the outcome of what we are trying to find. Consider the methods used to collect the data you analyzed. Was there any potential issue in how the participants were selected/recruited, retained, or assessed that may have impacted the outcome of your analysis/visualization? Were there any potential biases that you might be concerned about? Were there factors that were not measured or considered that you think could be important to the interpretation of these data?

#The results showed that 95 percent of confidence intervals we make will contain the null hypothesis value (mean difference is equal to 0), which means we can be fairly confident in saying that the most diverse and the least diverse counties do not have a difference in modified retail food indexes due to factors other than random chance. Our p-value is 0.1466, meaning it is fairly strong evidence for the null hypothesis.

#The data collection process failed to consider the fact that counties have different sizes and populations. Therefore, the amount of markets within a certain radius can be less helpful in modeling the accessibility of the supermarkets for county residents.

#There is no issue with the way the participants were chosen because the data is from a retail database on all markets, grocery stores, and food vendors in California. These affect all Californians and cover almost all of the places residents of California can get food.

#A potential bias is that the definition of healthy food can be biased in a way that favors traditional American healthy markets or restaurants as the only healthy option. Although they tried to label parent businesses that were on a list of healthy parent businesses as healthy, not all parent businesses have access to this list and cannot put themselves in. Therefore, other diverse markets that are healthy may be labelled as unhealthy. This would lead to a promotion of stereotypically American food choice when trying to promote health systemically. #A factor we should consider is whether there is accessible public transportation to get to these supermarkets because not everyone has access to transportation. Some need supermarkets to be within walking distance while others are able to drive to different locations.

23. [1 mark] Create a statement of contribution. This is now common in journal articles. For example, the American Journal of Epidemiology provides the following instructions to authors: “Authorship credit should be based on criteria developed by the International Committee for Medical Journal Editors (ICMJE): 1) substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data; 2) drafting the article or reviewing it and, if appropriate, revising it critically for important intellectual content; 3) final approval of the version to be published. Authors should meet all conditions. In addition, each author must certify that he or she has participated sufficiently in the work to believe in its overall validity and to take public responsibility for appropriate portions of its content. Author names should be listed in ScholarOne and author contributions should be detailed in the cover letter (e.g., “Author A designed the study and directed its implementation, including quality assurance and control. Author B helped supervise the field activities and designed the study’s analytic strategy. Author C helped conduct the literature review and prepare the Methods and the Discussion sections of the text.”). An example from a recent issue of the BMJ (Woolf, Masters, and Aron BMJ 2021;373:n1343): “Contributors: SHW led the production of this manuscript and had primary responsibility for the composition. He is guarantor. RKM contributed revisions and had primary responsibility for data acquisition and analysis, the modeling results that form the basis for this study, and production of the supplementary 2 material. LYA contributed revisions and had primary responsibility for dealing with the study’s policy implications in the discussion section. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.” For your project, please craft a statement indicating the contributions of each group member. If your group divided the assignment responsibilities by question you may use question numbers to indicate which member had primary responsibility for each question (for example: Member XX had primary responsibility for questions x,x,x. . .).

#Nora had primary responsibility for questions 4, 6, 9, 13, 19, and helped on 17, 18, and 22; helped out with others' questions and contributed ideas for how to go about data analysis/what to include.

#Chris had primary responsibility for questions 7,8,14,18,20,21 and put together the pdfs for submission and contributed ideas for how to go about data analysis. #Christine had primary responsibility for questions… #Our 4th member, Maddie, helped on the first part of the project, most importantly suggesting to get race data as a second variable for our analysis. #All collaborated and provided feedback on questions.