Written Report

Catherine Le

Introduction

Data Set Background

This dataset is officially called the "Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System" (BRFSS). BRFSS is an on-going, state-based telephone survey conducted by the CDC and state health departments. The data includes self-reported information on nutrition related information and risk factors for all 50 U.S. states. More specifically, it includes state-specific data on adult diet, physical activity, and weight status. This data set also includes information on the survey participants' demographics (e.g., age, ethnicity, income). In this data set, there are 88,629 rows and 33 columns.

Formulated Question

What is the correlation between physical activity levels (e.g., reported exercise frequency) and diet habits (e.g., daily consumption of vegetables)?

The intention of this final project is to further explore the stereotype that people who eat healthier also exercise often. Unfortunately, this data was collected through self-reported surveys, which resulted in numerous instances of missing demographic information. Consequently, I avoided attempting to investigate a question related to demographics due to the limitations of this data set.

Methods

Where I Acquired the Data Set

I exported the data set from the Center for Disease Control and Prevention (CDC) website. Here is the link to access the data: https://data.cdc.gov/Nutrition-Physical-Activity-and-Obesity/Nutrition-Physical-Activity-and-Obesity-Behavioral/hn4x-zwk7

Cleaning the Data

Removing Insufficient Sample Size Data

In this data set, there is a variable titled Data_Value_Footnote. In this column, there were multiple cases where it said "data not available because sample size is insufficient." I removed these rows because all the corresponding rows had NA for the data values. Unfortunately, I do not have enough contextual information to impute the average or mean. Furthermore, the states that had missing values seemed random, so there will should be no bias from removing these values. Therefore, I believed it was necessary to remove these values to avoid skewing my results.

Checking for Missing Values

There is a column in the data set titled Data_Value. This column shows percentages of survey respondents who answered "yes" or "no" to a question. The information in this column is integral to answering my formulated question. Therefore, it is important to determine if there are any missing values in this column.

```
Number of missing values is 0
```

Based on this output, there are no missing values in the Data_Value column.

Checking for Negative Values or Values Greater Than 100

Due to the nature of the Data_Value column, there should not be any negative percentages or percentages over 100. Therefore, I am checking to see if there are any implausible values in this column.

```
Number of negative values is 0
```

Based on this output, there are no negative values in the Data_Value Column.

```
Number of values greater than 100 is 0
```

Based on this output, there are no values greater than 100 in the Data_Value Column.

Wrangling the Data

In order to explore my formulated question further, I filtered out the data for exercise-related questions and vegetable-consumption related questions. I filtered the data set into a subset titled exercise_data. This subset contains rows where the survey question is "Percent of adults who engage in no leisure-time physical activity." I did the same for vegetable-consumption related questions. Here, I filtered the data set for rows that contained "Percent of adults who report consuming vegetables less than one time daily."

I then grouped both new data sets by the variables YearStart (i.e., the year the survey started) and LocationDesc (i.e., the state the survey was given in). Then, using summarize(), I found the mean for the Data_Value column of each group. By averaging and filtering the datasets, I aim to facilitate the creation of more comprehensible plots using a more concise subset of the data.

I then added a column to both aggregated data sets titled DataType. When I later merge the two datasets, this allows me to identify the rows that correspond with either the exercise survey question or the vegetable consumption survey question.

In order to gain a better picture of what my filtered data set looks like at this point, below is a table of the aggregated exercise data.

```
# A tibble: 6 x 4
# Groups:   YearStart [1]
  DataType YearStart LocationDesc Data_Value
  <chr>        <int> <chr>             <dbl>
1 Exercise      2011 Alabama            32.7
2 Exercise      2011 Alaska             23.9
3 Exercise      2011 Arizona            24.5
4 Exercise      2011 Arkansas           30.6
5 Exercise      2011 California         19.6
6 Exercise      2011 Colorado           17.3
```

Similarly, below is a table of the aggregated vegetable consumption data.

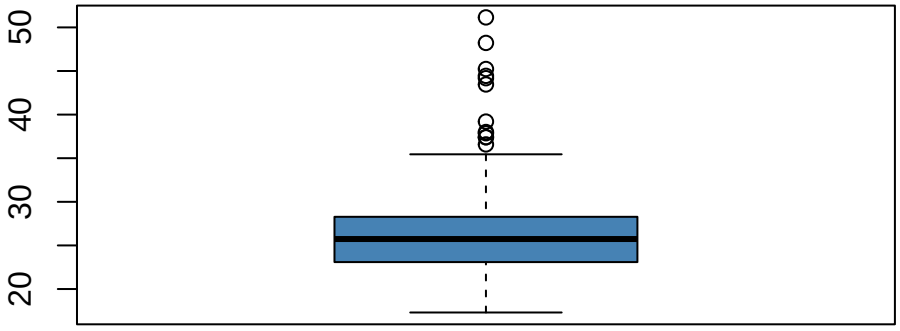
```
# A tibble: 6 x 4
# Groups:   YearStart [1]
  DataType  YearStart LocationDesc Data_Value
  <chr>         <int> <chr>             <dbl>
1 Vegetable      2017 Alabama            19.3
2 Vegetable      2017 Alaska             21.4
```

```
3 Vegetable      2017 Arizona            21.4
4 Vegetable      2017 Arkansas           21.0
5 Vegetable      2017 California         22.0
6 Vegetable      2017 Colorado           19.9
```

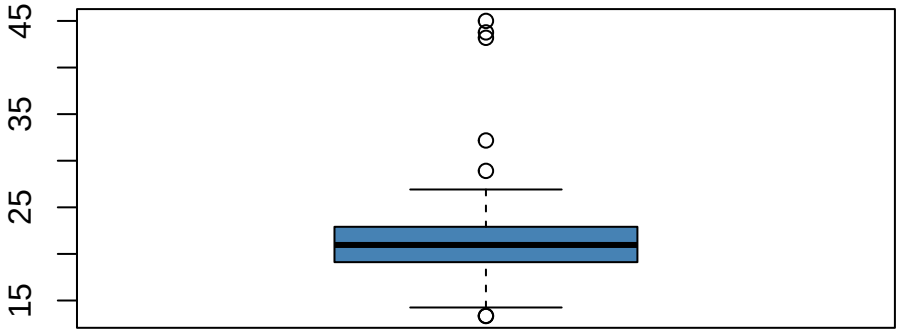
Lastly, I merged the aggregated exercise data set and the aggregated vegetable data set into a final data set I could later make plots with.

Data Exploration

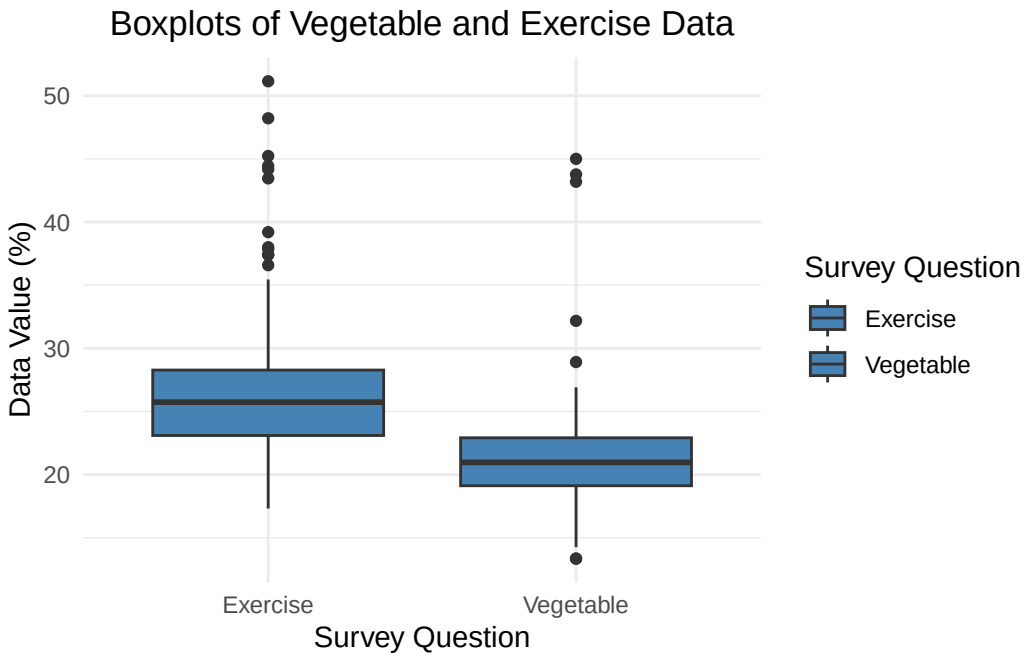
Below is a box-and-whisker plot illustrating the aggregated exercise data. This visualization provides a clearer understanding of the outliers present in the dataset, particularly those exceeding 36 percent.



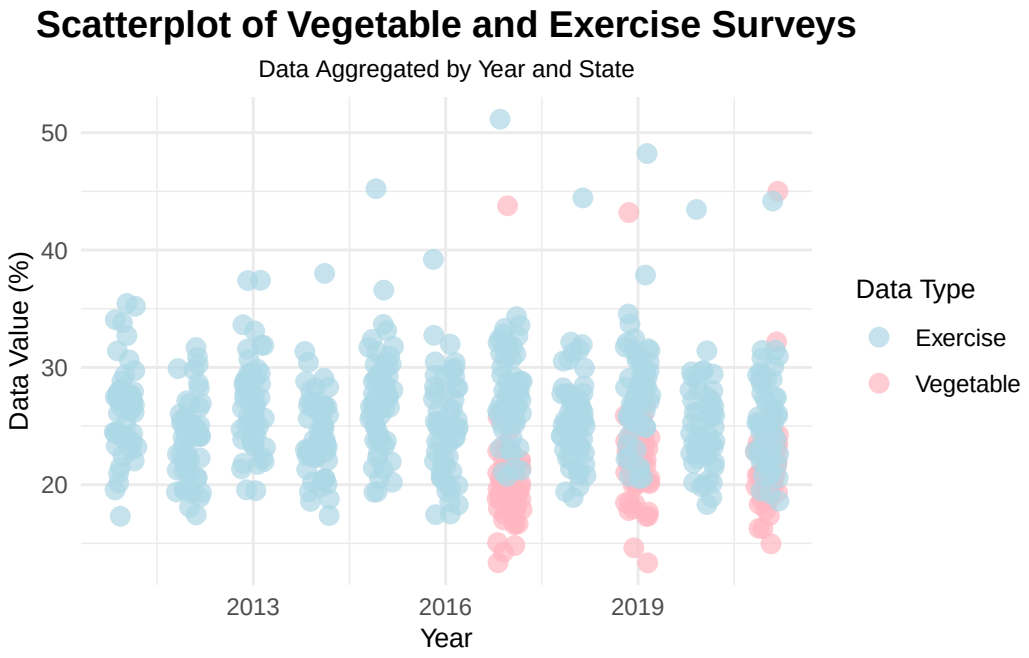
Below is a box-and-whisker plot illustrating the aggregated vegetable consumption data. This visualization provides a clearer understanding of the outliers present in the data set, particularly those exceeding 27 percent and below 15 percent.



Below, I have created a plot with the two boxplots placed next to each other. This allows for easier comparison between the two survey questions. This plot allows me to visually see if there are eyebrow-raising discrepancies between the median of the two data sets, which there are none.



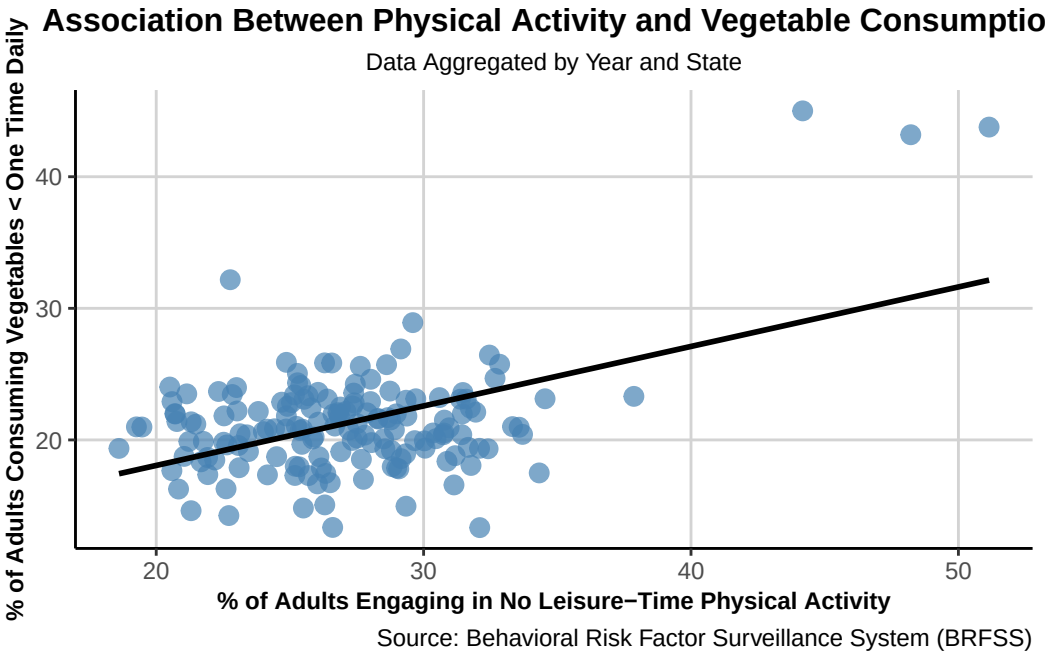
The last tool I used for data exploration was to create a scatterplot where there are different colors for the DataType (vegetable survey question or exercise survey question) variable. Due to the large amount of data points, I jittered the data points and made them transparent to be easier to view. The intention of this plot is to see if there are any noticeable patterns between the two data sets. This plot illustrated that the vegetable survey questions were only conducted after 2017. In contrast, the survey questions about physical activity levels were conducted from 2012 to 2021.



Results

The scatterplot belows illustrates the association between the percentage of adults who engage in no leisure time physical activity and percentage of adults consuming vegetables less than one time daily. The data was aggregated by StartYear (i.e., the year the survey started) and LocationDesc (i.e., the state the survey was conducted in). In this plot, the black line is a linear regression line. In this context, the linear regression line has a positive slope.

```
`geom_smooth()` using formula = 'y ~ x'
```



Conclusion and Summary

In conclusion, the scatterplot in the Results section shows a positive slope for the linear regression line between physical activity levels (e.g., reported exercise frequency) and diet habits (e.g., daily consumption of vegetables). This shows that as the average percentage of adults consuming vegetables less than one time daily increases, the average percentage of adults engaging in no leisure-time activity also increases. Within the context of this dataset, there is a positive association between the vegetable-consumption survey data and the exercise survey data.

However it is important to note that while there is a positive correlation between the two survey data sets, it does not imply causation. It just suggests that the two variables move together. This does not necessarily mean that one variable causes the other. Furthermore, each survey had different target demographics which may have influenced this observed pattern. However, because there were numerous occasions where the self-reported surveys had missing demographic information, this is not something that I could have investigated further.

Lastly, this data from the BRFSS is self-reported information on nutrition and physical activity level. Self-reporting does have limitations. Survey respondents might be consciously or unconsiously influenced by "social desirability." Participants may feel the need to give a more "socially perferable" answer to the survey questions regarding vegetable consumption and diet habits (especially because these surveys were conducted over the telephone). This may introduce a level of bias into this data set and affect the observed pattern.