**Does an individual’s household income affect the amount of different Junk Foods they eat? And how does the total amount of Junk Food eaten affect an individual’s BMI?**

**Data Management:**

To investigate this question, we first want to extract the relevant data from the large, raw, .csv files we have been provided. To avoid unnecessary processing, we should first choose the columns we require from each of the datasets, and also preserve the “SEQN” column for data linking. Then use the merge() function to link the data together. The table below shows which columns we require from each dataset. I initially identified all ‘Junk Foods’ in the raw data set, then reduced this randomly to a subset of 10 junk foods.

|  |  |  |
| --- | --- | --- |
| BMI | “SEQN” – ID for merging | |
| “BMXBMI” – This is BMI for part 2 of the study | |
| DIET  Foods to be used as explanatory variables (EVs): | “SEQN” | |
| Fries-“FFQ0047” | Spareribs-“FFQ0081” |
| Pancakes-“FFQ0059” | Fried Fish-“FFQ0094” |
| Mac and Cheese-“FFQ0061” | Cake-“FFQ0114” |
| Peanut Butter-“FFQ0068” | Cookies-“FFQ0115” |
| Hamburgers-“FFQ0075” | Chocolate Candy-“FFQ0120” |
| DEMO | “SEQN” | |
| “INDHHINC” – Household Income grouped into 13 categories, EV in first study | |

Next, we want to remove erroneous or useless rows of data. The first ‘cleaning’ we do limits BMI to a valid range, outside of which the data may be erroneous or unrepresentative. Therefore, we set boundaries at 10 and 80 and remove any individuals in the study with BMI outside of this.

The next task is to remove individuals with incomplete data for our chosen relevant categories. The foods are scored on a scale of 1 to 13 (see appendices) and errors are marked with a score of “77”, “88”, “99” or “.”. To remove these points we swap any appearances of these in the data with “NA” and use the “na.omit” function to remove the incomplete rows.

Finally, we add one column to the data labelled “jfs” (Junk-Food-Score), which sums the scores from the food variables to indicate the total junk food an individual eats. Doing this produces a column with range 9 to 94, median 36 and an IQR from 30 to 43.

Our data is now ready to produce plots.

**Data Visualisation Approach:**

To address the research question, we aim to produce one plot demonstrating the associations between different foods (and jfs) with Household income and another for the association between jfs as the EV and BMI as dependent variable. The first will show a series of fitted linear models to highlight the relationships. The second will present the raw data – in boxplot form – with a model fitted on top of it. These, alongside statistical measures, can hopefully answer our research question.

To support the first study, we also produce a shiny app. It can be accessed by running the ‘us\_food\_shinapp.R’ in the ‘visualisation’ directory – instructions in ‘readme.md’. This allows you to view the raw data for each food type along with the fitted models, and for comparisons between any three of the food models. There are also toggles for confidence intervals and for linear or categorical regression.

**Results:**

A graph of colored lines

Description automatically generated

Our first plot consolidates the main findings from the shiny app (CIs removed for clarity). The plot shows how overall there is a general decrease in the amount of junk food eaten as household income increases: see the dotted ‘Junk- Food-Score’ line. When testing the relationship we see reasonable evidence for this trend with a p-value of 0.021, more detail in the shiny app.

Note, Peanut Butter, Cookies and Chocolate Candy show the opposite relationship.

**A graph with a line of black and white squares

Description automatically generated with medium confidence**The second plot investigates how BMI is associated with consuming junk food (measured by JFS). Before the study it was assumed BMI would increase with JFS, however our study found the opposite. The plot includes both a categorical and linear fitted line on top of the raw data (they’re hard to distinguish) and both indicate than an individual who consumes more junk food will likely have lower BMI.

The outputs from our linear model were a beta\_1 parameter of -0.475 (-0.545, -0.405) with an associated p-value of <0.001 indicating there is very strong evidence this association is present in our data.

An explanation for the results we have observed is that people with higher BMI are forced to eat less junk food because they are trying to lose weight.

Condfounding? – adjusted, kind of

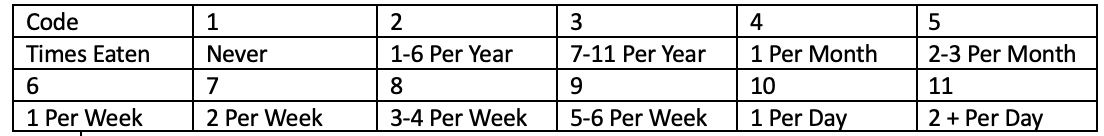
We adjust for this using a weighted average of a stratified model. Weights are 1/std^2. Doesn’t make much difference

**Red line = stratified, black = adjusted in reg model and use subset means**

Can explain process tho, think the plot is acc quite nice now

**§: Appendices: Explanation for Categorical Variables**

* A number on a white background

  Description automatically generatedHousehold Income Categories:
* ****General Food Categories