**AHDS Assessment 2 Report**

**Research Question:- Does and 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 produce 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 explains briefly what each may be useful for. Initially I identified all ‘junk foods’ in the raw data set, then reduced this randomly to a subset of 10 junk foods.

|  |  |  |
| --- | --- | --- |
| Dataset | Column | |
| BMI | “SEQN” – ID for merging | |
| “BMXBMI” – Require this as the dependent variable for the second half of the research question | |
| DIET  These columns are the different Foods to be used as explanatory variables (EVs): | “SEQN” – ID for merging | |
| 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” – ID for merging | |
| “INDHHINC” – Household Income, EV in first study, this is grouped into 13 categories | |
| “RIDAGEMN” – Age in Months, possible confounder | |
| “DMDHHSIZ” – Number in Household, possible confounder | |

Once we’ve formed this dataset, we want to remove erroneous or useless rows of data. The first ‘cleaning’ we do removes any BMI values that could be from mistaken inputs or even if not, they aren’t representative of the general public (as we want our answer to the research question to be). Therefore, we set boundaries at 10 and 80 and trim any individuals with BMI outside of this.

Next, we want to remove any rows in the data which are incomplete – don’t have data for all the columns we have selected. When the data was initially collected, the DIET and DEMO datasets were done using multiple choice questions so we have easy categorical variables to work with, where every individual should score from 1 to 13. If they don’t then we know there is an error, this is normally marked with an input of “77”, “88”, “99” or “.”. We remove all these points by swapping any appearances of these in the data with “NA” and use the “na.omit” function to remove any incomplete rows.

Finally, we add one column to the data called “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:**

Addressing the research question requires us to fit two main regression models. One with Household Income as the EV and jfs as the dependent variable and another with jfs as the EV and BMI as the dependent. We will use these to produce static plots, where the models’ fitted values are imposed on top of the raw data. We hope this, alongside analysis of the model output, will provide evidence to answer our research question.

We will also develop a shiny app, which can be accessed by running the ‘us\_food\_shinapp.R’ through the analysis pipeline. It will include interactive plots where you can observe the association each type of junk food has with household income and allow you to compare them. There are also toggles for regression type and confidence intervals.

**Results:**

Plotting our results we see household income is associated with a very small reduction in jfs.

**Change this plot probs: include all food stuff**

**Think this plot needs big change as doesn’t show much and looks too similar to other plot, also probs should show models without rawdata, think I might fit fp models and see what they look like, maybe a bit much with 10 models we’ll see. I think either all fp or all linear though**

**A graph with a line and dots

Description automatically generated with medium confidence**

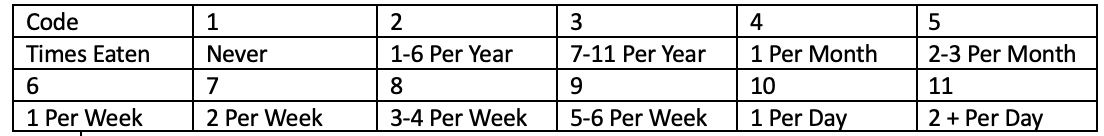
**A graph with a line and dots

Description automatically generated with medium confidence**

two ggplot2 visualisations embedded in text describing your interpretation of them (i.e. what does each plot show us that we did not know before). This section should also include interpretation of the interactive Shiny visualisation that can be viewed separately.

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

* A number on a white background

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