Tiffanie F. Stone

12-06-2019

590A – R Data Management

**Methods and Results**

**Introduction and Objectives**

Publicly available USDA data on National Food Consumption was analyzed to find trends and differences across years food category and income level. The information collected will be used as a baseline for a larger study of US food systems at a city scale.

The objectives were as follows:

1. The US national trends for food consumption by food category (fruit, vegetable, grain, oil, meat, dairy) annually from 1994 - 2007.

2. The difference between low- and high-income food consumption when compared to the average.

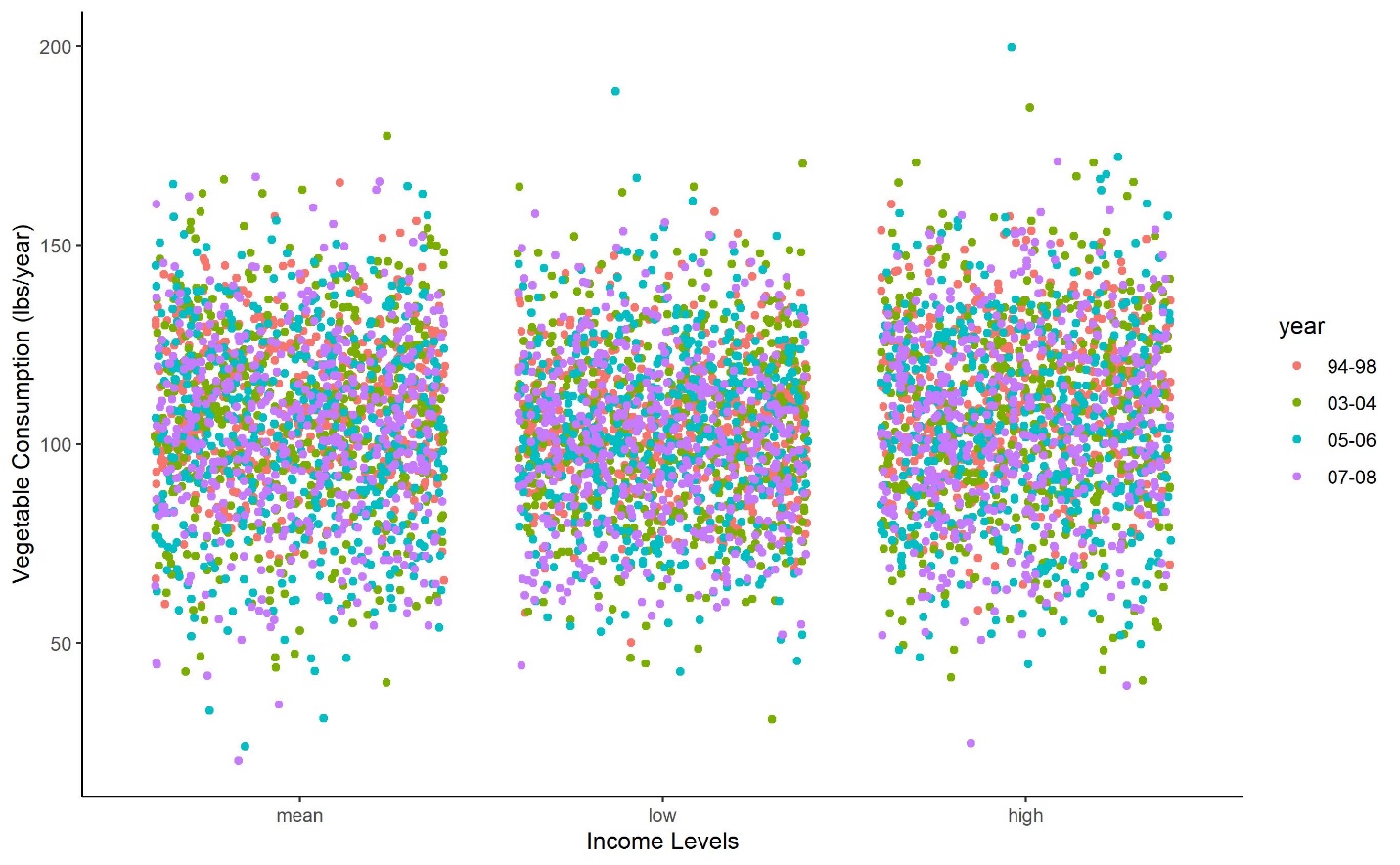
3. A simulated data set for vegetable consumption was statistically analyzed for significant differences across years and income levels.

**Methods**

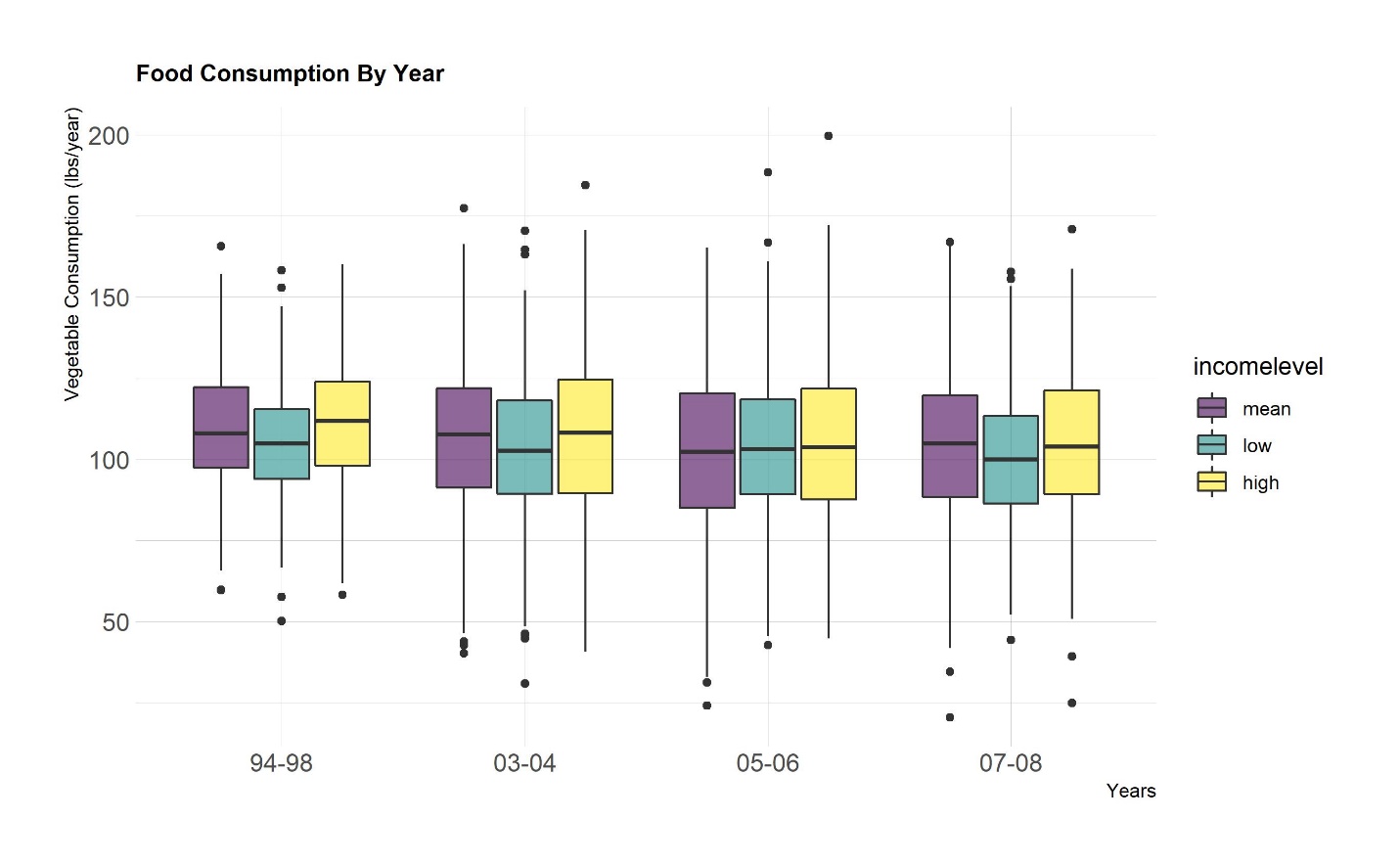
Given the lack of raw survey data, statistical analysis was conducted on a simulated data set for vegetable consumption in the US. R was used to generate the simulated data and to analyze the data collected (R Core Team, 2013). Tidyverse packages were used as the primary coding method (Wickham, 2017). A linear model (lm) was used because the data appeared to follow a normal distribution of error based on the resid\_panel ran on the linear model selected. The resid\_panel showed a linear relationship which favored linear models as opposed to glm or glmer. Three linear model approaches were considered, two using p-values, one using AIC values. The Fit-Full Model was accepted without interaction between the two predictors because the interaction was not significant. Both the Fit-Full and the Traditional Hypothesis Testing led to the same resulting model.

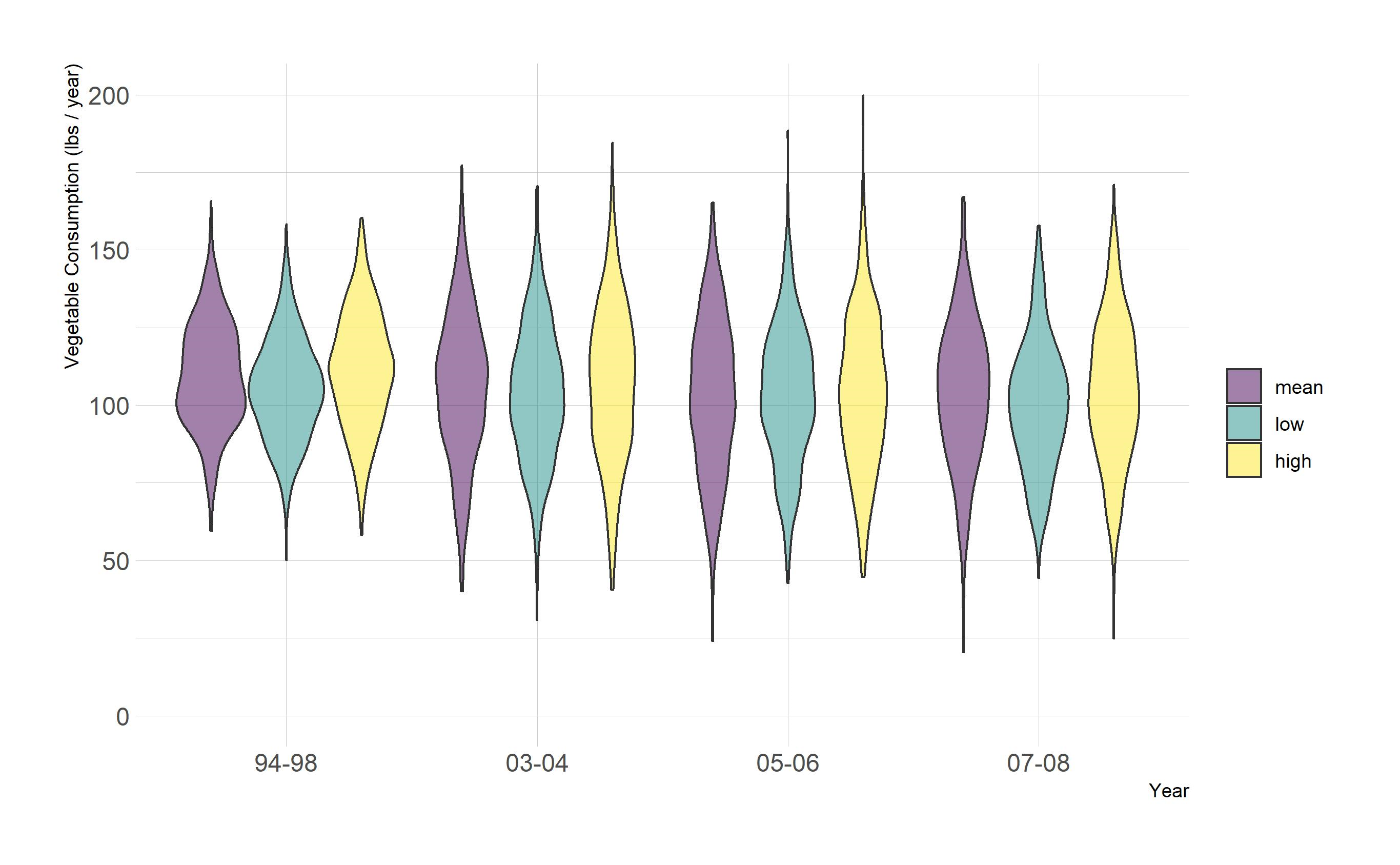
The response was lbs food category per year. Predictors were income levels (low, high and mean) and year. No random effects were used. Interactions between predictors were not significant and were not utilized in the final linear model. The equation for the linear model is: vegsimmod2a <- lm(Simulated Vegetable in lbs/year ~ incomelevel+year, data=Vegetable Consumption Simulated Dataset). P-values were used as the method of inference. Ggplot2 was used to generate all graphics (Wickham, 2016).

**Results**

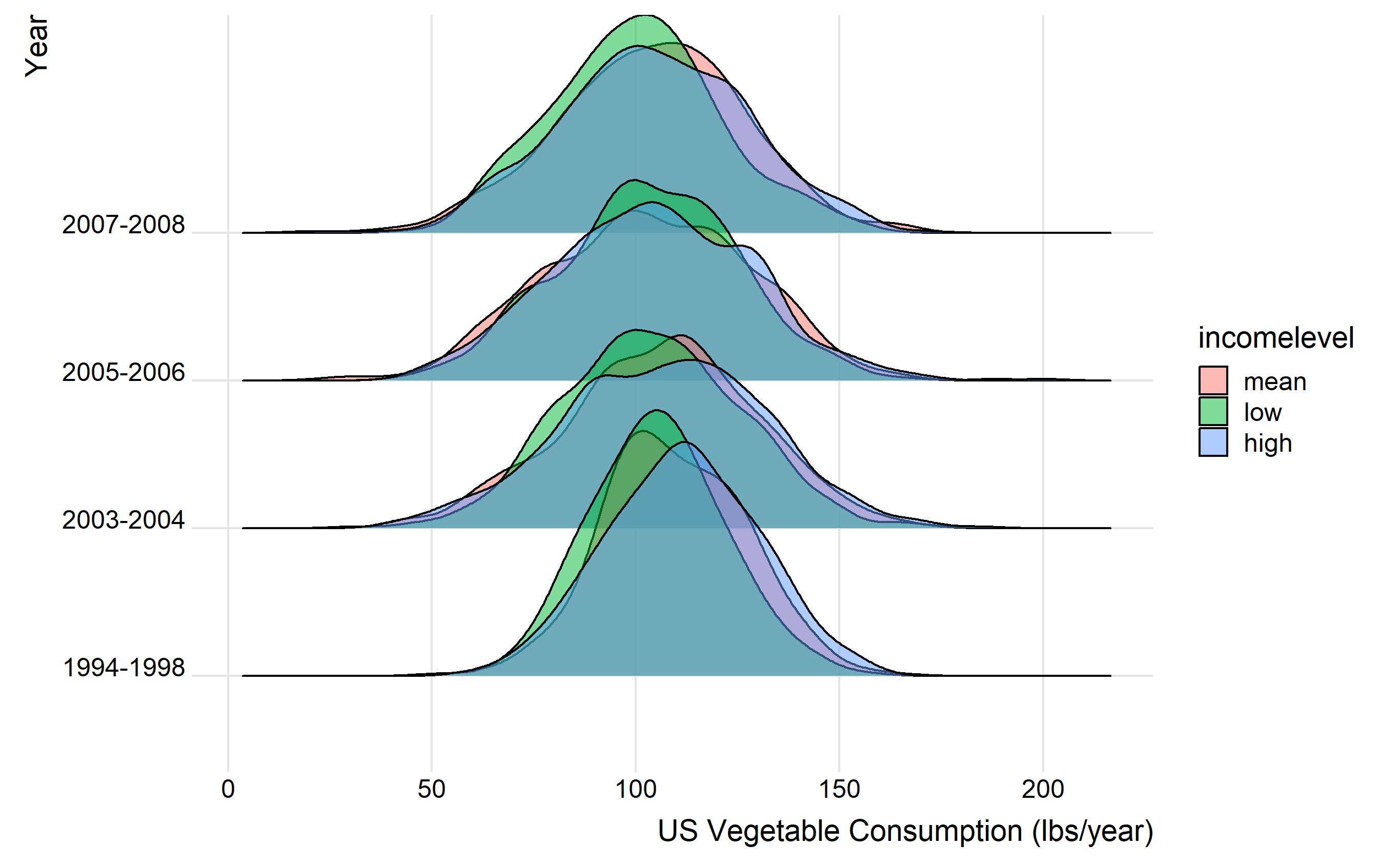


Vegetable consumption does not vary

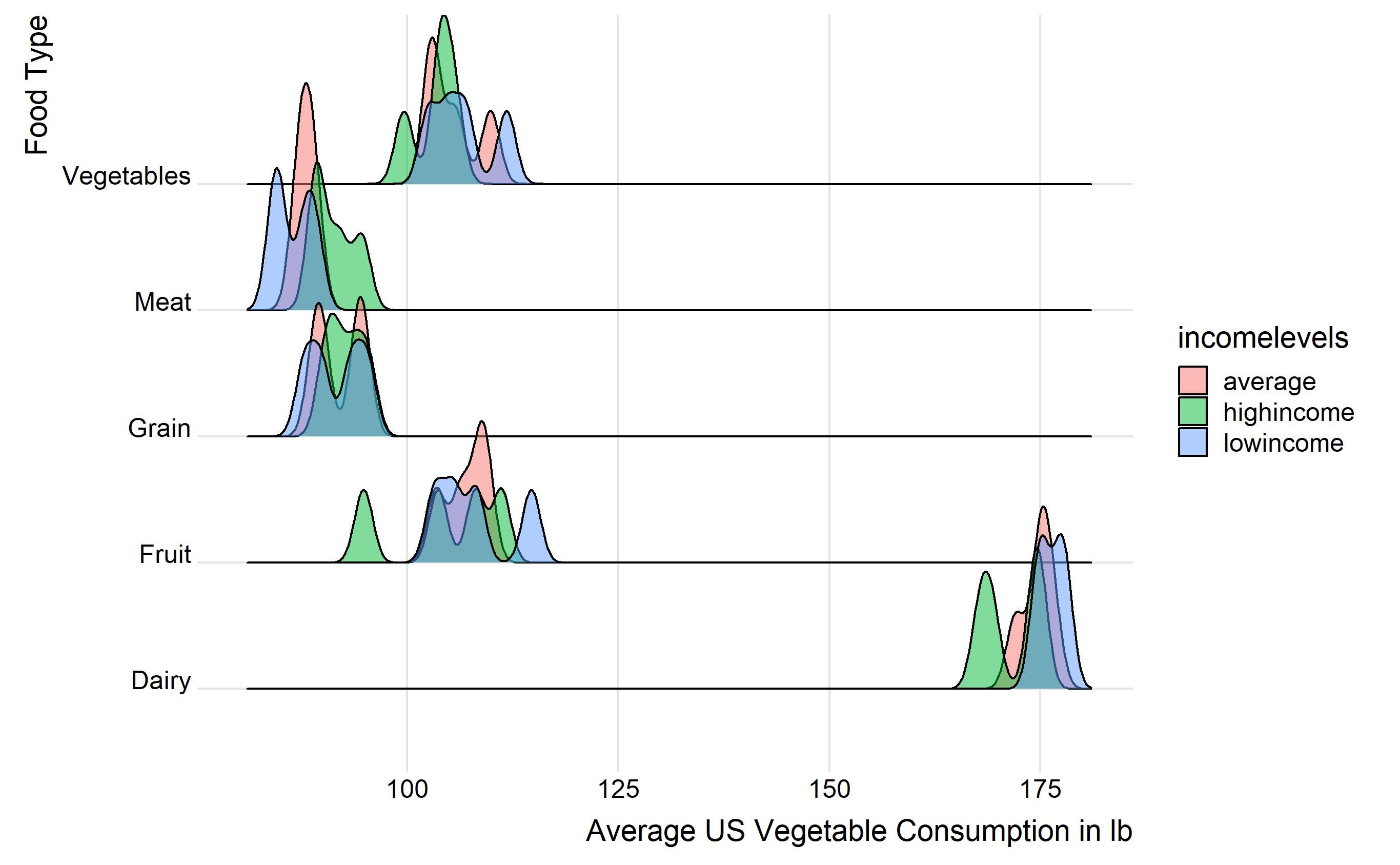


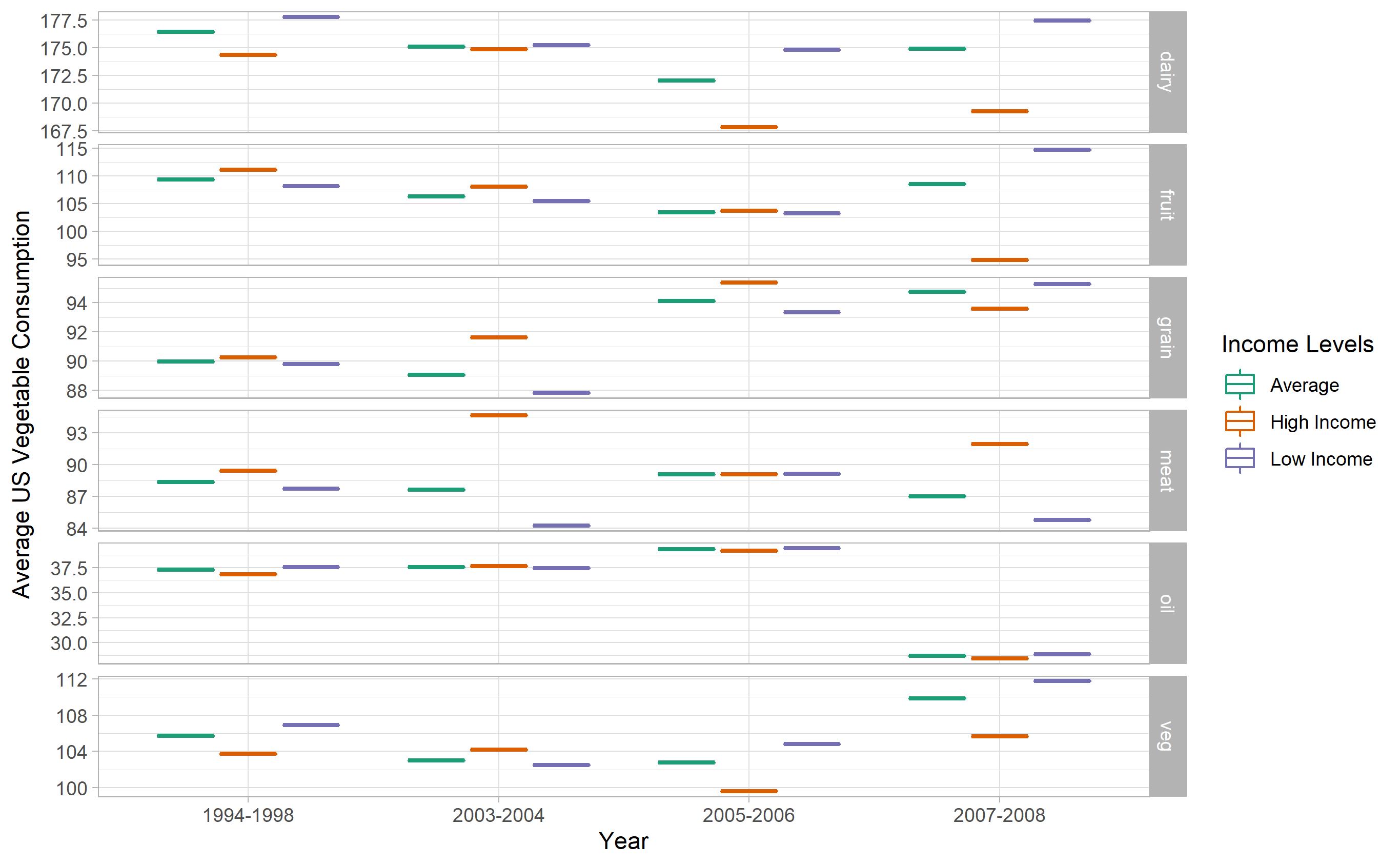


There are significant differences in vegetable consumption from year to year, due to the large sample size. A trend is not visible but might come to light if more years were tracked. There are small variations in the in the mean



The mean across years is slightly above 100 lbs of vegetables per year. The low-income samples eat less vegetables on average, whereas the high income eat slightly more vegetables per year, though differences are not statistically significant.





These graphs compare low, average and high-income food consumption, across food types and years. The income levels appear to be grouped very closely, there is not a wide variation. High income is slightly higher for meat, and low income is slightly higher for vegetables and dairy most years. This is a surprising result. Expected lower amounts of fruit and vegetable consumption for low income.

**Sources**

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Wickham, Hadley. (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wickham, Hadley (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse