Homework9

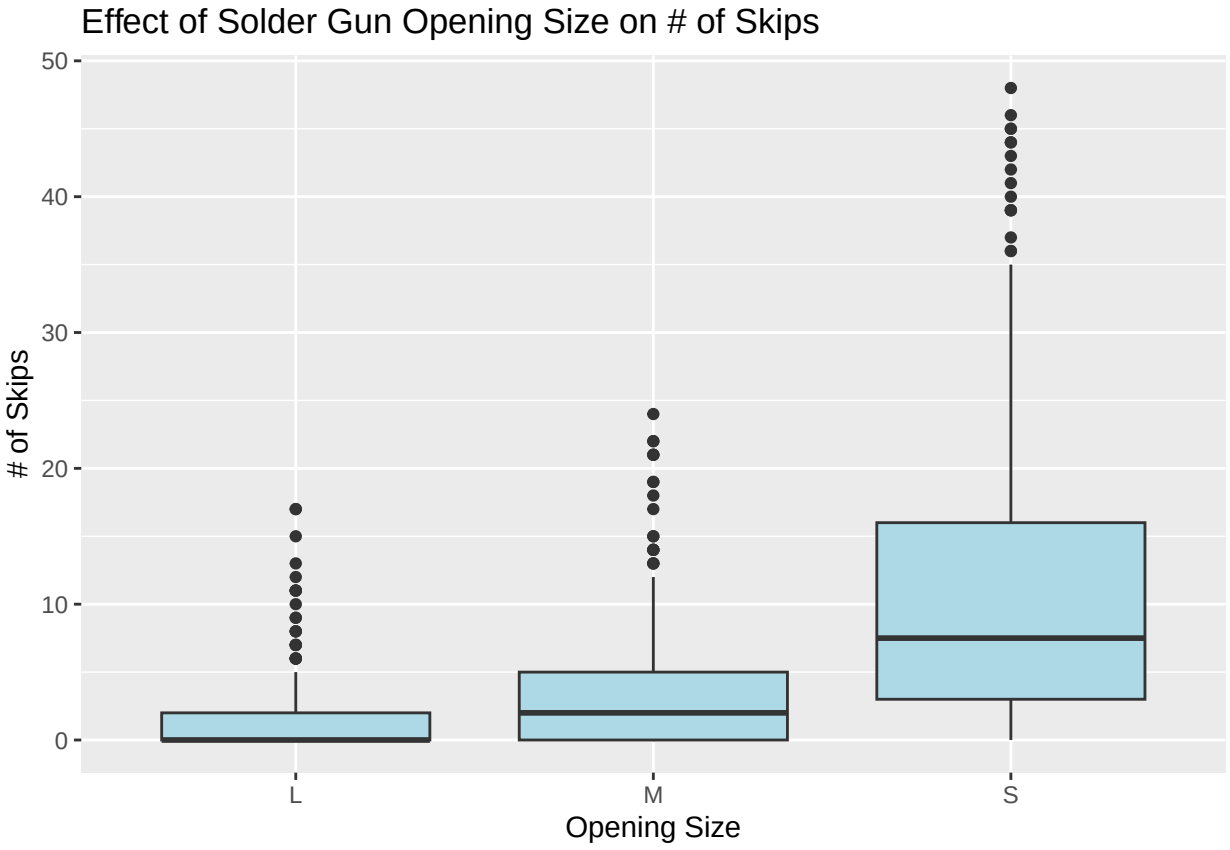
Kavya Malgi

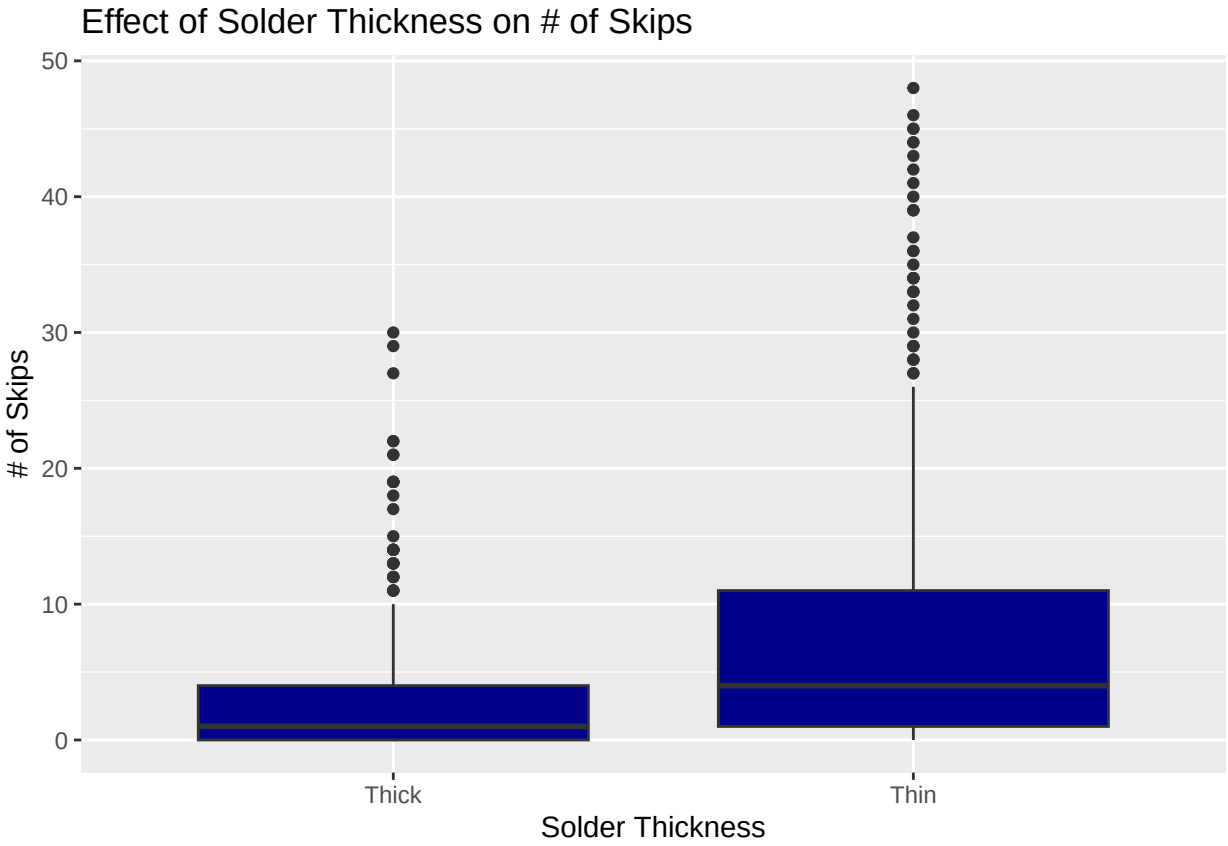
2025-04-20

github link: https://github.com/kavyamalgi/Homework9

Problem 1: Manufacturing flaws in circuit boards

Part A: Make two plots. The first plot should provide evidence that the size of the opening on the solder gun is related to the number of skips. The second should provide evidence that the thickness of the alloy used for soldering is related to the number of skips. Give each plot an informative caption describing what is shown in the plot.





Plot 1: The boxplot depicts the # of solder skips that vary with the solder gun opening size. Boards that are manufactured with smaller openings have fewer skips on avg. meaning better quality.

Plot 2: The boxplot shows that the thinner a solder is, the less amount of skips there are compared to the thicker solder. This indicates taht solder thickness is a factor in the manufacturing quality.

Part B: Build a regression model with skips as the outcome and with the following terms as predictors:

- a main effect for Opening
- a main effect for Solder type
- an interaction between Opening and Solder type

Make a table that shows the estimate and 95% = large-sample confidence interval for each coefficient in your model.

```
## # A tibble: 6 x 7
##   term                estimate std.error statistic  p.value conf.low conf.high
##   <chr>                  <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)            0.393     0.520     0.756 4.50e- 1   -0.628      1.41
## 2 OpeningM               2.41      0.736     3.27  1.11e- 3    0.962      3.85
## 3 OpeningS               5.13      0.736     6.97  6.29e-12    3.68       6.57
## 4 SolderThin             2.28      0.736     3.10  2.01e- 3    0.836      3.72
## 5 OpeningM:SolderThin   -0.740     1.04     -0.711 4.77e- 1   -2.78       1.30
## 6 OpeningS:SolderThin    9.65      1.04      9.28  1.29e-19    7.61      11.7
```

Part C: Interpret each estimated coefficient in your model in no more than 1-2 sentences. A good template here is provided in the course packet, when we fit a model for the video games data that had an interaction in it and interpreted each coefficient in a sentence or two

```
## The baseline # of skips for curcuit boards that were manufactured with a large opening and thick sol

## The main effect for a medium opening is NA skips. This shows the effect of a medium opening.

## The main effect for a small opening is NA skips. This shows the effect of a small opening.

## The main effect for a thin solder is NA skips. This shows the effect of a thin solder in isolation

## The interaction effect for medium opening and thin solder is NA skips.

## The interaction effect for small opening and thin solder is NA skips.
```

Part D: If you had to recommend a combination of Opening size and Solder thickness to AT&T based on this analysis, which one would it be, and why? (Remember, the goal is to minimize the number of skips in the manufacturing process.)

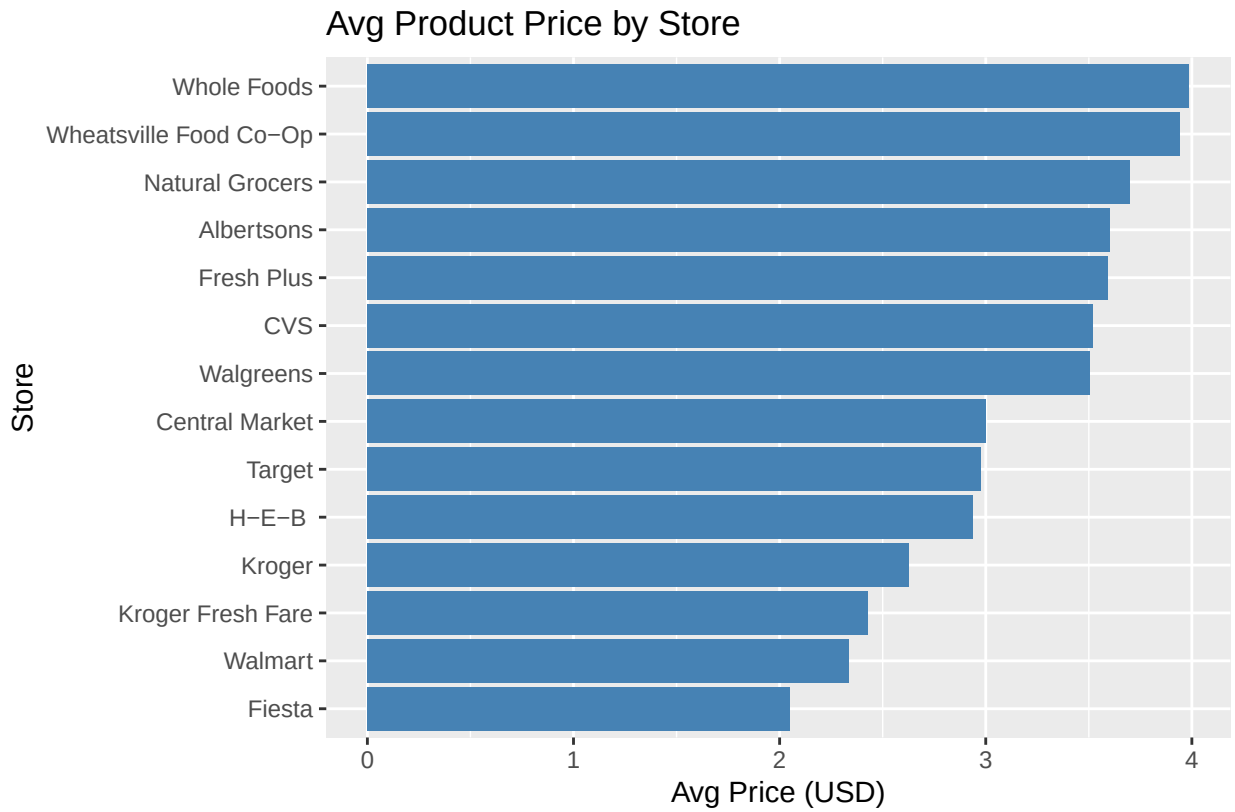
If I had to recommend a combination of Opening size and Solder thickness, based on the regression results, I would say that AT&T should use a small opening and a thin solder.

Using this combination would end up in the lowest predicted # of skips, when discussing the individual effects of Opening size and Solder thickness for both.

The model showed that thinner solder reduces skips compared to thicker solder. Furthermore, small openings reduce skips compared to medium and large openings. Lastly, the interaction term for a small opening and thin solder doesn't change the benefit, and results in the combination to be more beneficial.

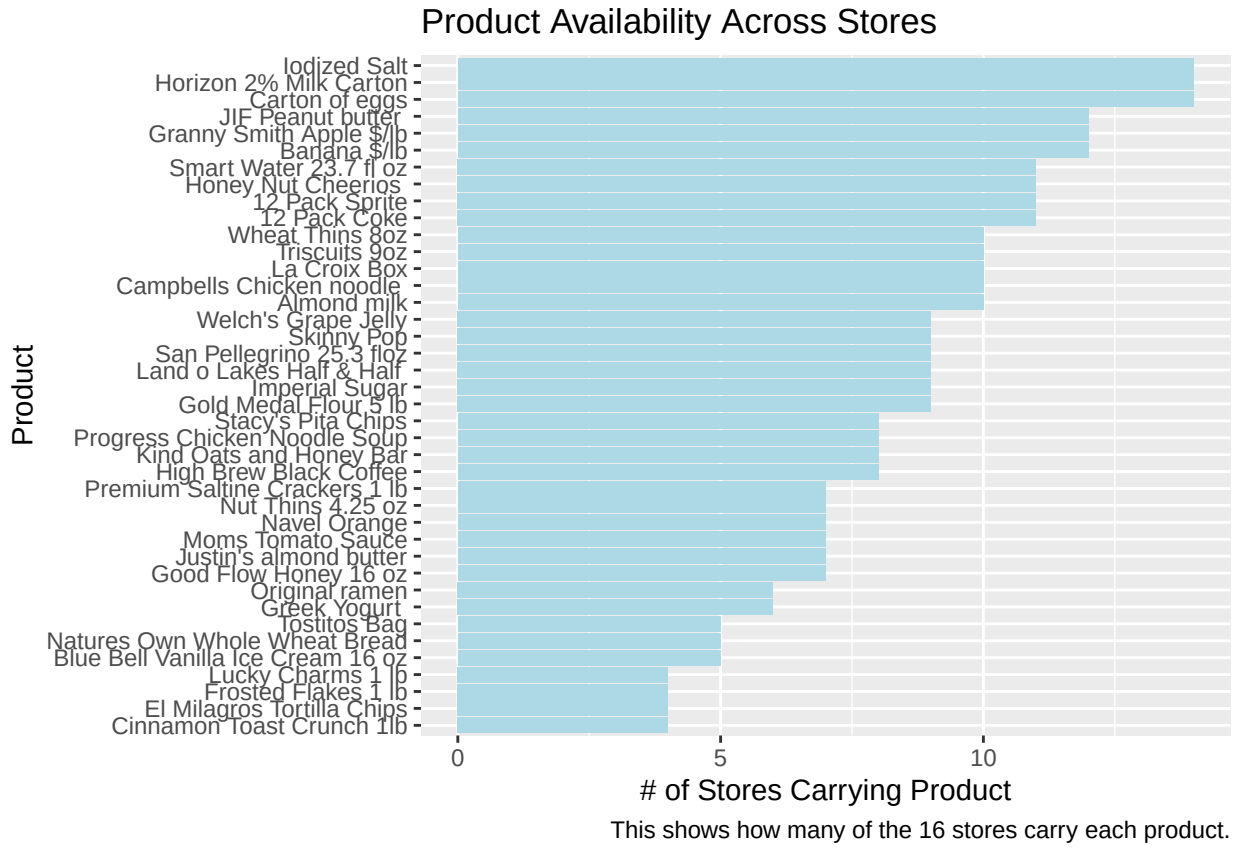
Problem 2: Grocery Store Prices

Part A. What kind of price differences do we see across the different stores? Make a bar graph with Store on the vertical axis and average price of products sold at that store on the horizontal axis. (Remember coord_flip.) Give your plot an informative caption. You'll need to wrangle the data into an appropriate form first before you can make your plot.



The bar graph shows the avg. price for all products that were sold at each store.

Part B: Please make a bar graph with Product on the vertical axis and number of stores selling that product on the horizontal axis. Give your bar graph an informative caption. Again, you'll need to wrangle the data into an appropriate form first before you can make your plot. (For the purposes of this question, you can treat the two HEBs and two Whole Foods as separate stores, which makes the data wrangling easier. You'll know you've gotten this right if your bar graph maxes out at 16 for eggs and milk.)



Part C: Now let's use regression to try to isolate the effects of Type of store versus the actual products being sold. Fit a model for Price versus Product and the Type of store. Fill in the blanks: "Compared with ordinary grocery stores (like Albertsons, HEB, or Krogers), convenience stores charge somewhere between (lower bound) and (upper bound) dollars more for the same product." Use a large-sample confidence interval here, and round your answer to two decimal places, i.e. the nearest penny.

```
## # A tibble: 1 x 2
##   conf.low conf.high
##      <dbl>     <dbl>
## 1     0.41      0.92
```

Compared with ordinary grocery stores (like Albertsons, H-E-B, or Kroger), convenience stores charge between $0.41 and $0.92 more for the same product, on average.

Part D. Now fit a model for Price versus Product and Store. Which two stores seem to charge the lowest prices when comparing the same product? Which two stores seem to charge the highest prices when comparing the same product?

```
## # A tibble: 13 x 4
##    term                          estimate conf.low conf.high
##    <chr>                            <dbl>    <dbl>     <dbl>
##  1 "StoreWalmart"                   -0.99    -1.45     -0.53
##  2 "StoreKroger Fresh Fare"         -0.9     -1.36     -0.44
##  3 "StoreFiesta"                    -0.7     -1.23     -0.17
##  4 "StoreKroger"                    -0.7     -1.16     -0.24
```

```
##  5 "StoreH-E-B "                    -0.65    -0.95     -0.35
##  6 "StoreCentral Market"            -0.57    -0.92     -0.23
##  7 "StoreTarget"                    -0.37    -0.75      0
##  8 "StoreNatural Grocers"           -0.08    -0.47      0.31
##  9 "StoreFresh Plus"                -0.04    -0.35      0.28
## 10 "StoreCVS"                        0.19    -0.17      0.55
## 11 "StoreWalgreens"                  0.22    -0.14      0.57
## 12 "StoreWheatsville Food Co-Op"     0.29    -0.06      0.64
## 13 "StoreWhole Foods"                0.36     0.02      0.71
```

After using the regression model comparing stores for the same product, the 2 stores with the lowest prices are Walmart (estimate = -0.99) and Kroger Fresh Fare (estimate = -0.90). The 2 stores that have the highest prices are Whole Foods (estimate = 0.36) and Wheatsville Food Co-Op (estimate = 0.29). The estimates are relative to the baseline store and describe the avg. difference in prices for the same product after accounting for differences across products, suggesting that the store choice has an important impact on price.

Part E. Central Market is owned by HEB but has a reputation as a fancier grocery store that charges premium prices. But is that because Central Market charges more for the same product? (This is referred to as price discrimination in the marketing world.) Or, on the other hand, is that because Central Market sells different products that are inherently more expensive than those sold at a typical HEB? Let's use your model from Part D to try to disambiguate between two possibilities:

- **Central Market charges more than HEB for the same product.**
- **Central Market charges a similar amount to HEB for the same product.**

Inspect the coefficients from your fitted model. Which of these two possibilities looks right to you? Cite specific numerical evidence from your model. Try to put any difference between HEB and Central Market into the larger context: how big is the HEB/Central Market difference, compared to differences among other stores?

Referencing the regression model from Part D, we can compare the coeffs for HEB and Central Market.

- HEB: -0.65
- Central Market: -0.57

The coefficients above show the difference in price in relation to the baseline store. Central Market is only \$0.08 more expensive than HEB for the same products. The difference is small compared to the differences from other stores. It is safe to conclude that Central Market doesn't significantly charge more than HEB for the same product.

Part F. Finally let's consider the Income variable. To facilitate interpretation, first use mutate to define an Income10K variable that measures income in multiples of \$10,000 (e.g. 1 = \$10,000, 2 = \$20,000, and so on). Then fit a model for Price versus Product and Income10K and use your model to answer these two questions:

- **Based on the sign of the Income10K coefficient, do consumers in poorer ZIP codes seem to pay more or less for the same product, on average? How do you know?**
- **How large is the estimated size of the effect of Income10K on Price?**

```
##
## Call:
## lm(formula = Price ~ Product + Income10K, data = grocery)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9105 -0.4594 -0.0742  0.3881  4.0025
```

```
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               5.616264   0.248809  22.573  < 2e-16
## Product12 Pack Sprite                    -0.018333   0.328221  -0.056 0.955491
## ProductAlmond milk                       -2.113978   0.344698  -6.133 2.55e-09
## ProductBanana $/lb                       -4.908097   0.316362 -15.514  < 2e-16
## ProductBlue Bell Vanilla Ice Cream 16 oz -2.907746   0.429038  -6.777 5.92e-11
## ProductCampbells Chicken noodle          -3.372977   0.344698  -9.785  < 2e-16
## ProductCarton of eggs                    -2.973685   0.307025  -9.685  < 2e-16
## ProductCinnamon Toast Crunch 1lb         -1.195252   0.464561  -2.573 0.010538
## ProductEl Milagros Tortilla Chips        -1.999912   0.464486  -4.306 2.22e-05
## ProductFrosted Flakes 1 lb               -1.450252   0.464561  -3.122 0.001962
##
## (Intercept)                              ***
## Product12 Pack Sprite
## ProductAlmond milk                       ***
## ProductBanana $/lb                       ***
## ProductBlue Bell Vanilla Ice Cream 16 oz ***
## ProductCampbells Chicken noodle          ***
## ProductCarton of eggs                    ***
## ProductCinnamon Toast Crunch 1lb         *
## ProductEl Milagros Tortilla Chips        ***
## ProductFrosted Flakes 1 lb               **
##  [ reached getOption("max.print") -- omitted 31 rows ]
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.804 on 319 degrees of freedom
## Multiple R-squared:  0.8607, Adjusted R-squared:  0.8432
## F-statistic: 49.28 on 40 and 319 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = Price_s ~ Product + Income10K_s, data = standard)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.43328 -0.22624 -0.03655  0.19112  1.97104
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               1.080713   0.114403   9.447  < 2e-16
## Product12 Pack Sprite                    -0.009028   0.161633  -0.056 0.955491
## ProductAlmond milk                       -1.041034   0.169747  -6.133 2.55e-09
## ProductBanana $/lb                       -2.417004   0.155793 -15.514  < 2e-16
## ProductBlue Bell Vanilla Ice Cream 16 oz -1.431927   0.211281  -6.777 5.92e-11
## ProductCampbells Chicken noodle          -1.661031   0.169747  -9.785  < 2e-16
## ProductCarton of eggs                    -1.464398   0.151195  -9.685  < 2e-16
## ProductCinnamon Toast Crunch 1lb         -0.588604   0.228774  -2.573 0.010538
## ProductEl Milagros Tortilla Chips        -0.984862   0.228737  -4.306 2.22e-05
## ProductFrosted Flakes 1 lb               -0.714180   0.228774  -3.122 0.001962
##
## (Intercept)                              ***
```

```
## Product12 Pack Sprite
## ProductAlmond milk                        ***
## ProductBanana $/lb                        ***
## ProductBlue Bell Vanilla Ice Cream 16 oz ***
## ProductCampbells Chicken noodle           ***
## ProductCarton of eggs                     ***
## ProductCinnamon Toast Crunch 1lb          *
## ProductEl Milagros Tortilla Chips         ***
## ProductFrosted Flakes 1 lb                **
##  [ reached getOption("max.print") -- omitted 31 rows ]
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3959 on 319 degrees of freedom
## Multiple R-squared:  0.8607, Adjusted R-squared:  0.8432
## F-statistic: 49.28 on 40 and 319 DF,  p-value: < 2.2e-16
```

After fitting a multiple regression model with Price as the outcome and Product and Income10K as predictors. Based on the sign of the Income10K coefficient, consumers in poorer ZIP codes may more if they have a negative coefficient and consumers in wealthier ZIP codes pay more for the same products if the coefficient is positive.

The coefficient on Income10K was negative (-0.014), indicating that consumers in poorer countries tend to pay more for the same product, on average. Although this is true, the relationship was not statistically significant(p=0.14) indicating that the evidence is weak that the income meaningfully affects price at all.

Furthermore, in order to interpret the size of this effect, I standardized both income and price and ran a second model using z-scores. The coefficient in this model for Income10K_s was -0.032, meaning that A one standard deviation increase in ZIP code income is associated with a 0.03 standard deviation decrease in product price. It is safe to say that income level may have only a minimal effect on the prices consumers pay for the same grocery products.

Problem 3: Redlining

A. ZIP codes with a higher percentage of minority residents tend to have more FAIR policies per 100 housing units. **TRUE**

- Figure A1 shows a clear positive linear relationship and the regression table for model_A shows the coefficient for minority as 0.014 with a p-val < 0.001, and the CI [0.009, 0.018] which doesn't include 0. It is safe to say that as % minority increases, FAIR policy rate increases.

B. The evidence suggests an interaction effect between minority percentage and the age of the housing stock in the way that these two variables are related to the number of FAIR policies in a ZIP code. **FALSE**

- There is no interaction model between minority and housing age is given. The only possible model is model_C which is in regards to minority and fire risk.

C. The relationship between minority percentage and number of FAIR policies per 100 housing units is stronger in high-fire-risk ZIP codes than in low-fire-risk ZIP codes. **FALSE**

- In Figure C1, the figure shows almost parallel slopes for high and low fire risk meaning that there might be similar strength of relationship. In model_C, the interaction term minority:fire_riskLow has a coeff of -0.001 with a p-val of 0.839 and a CI that does include 0.

D. Even without controlling for any other variables, income "explains away" all the association between minority percentage and FAIR policy uptake. **FALSE**

- Referencing to model_D2, income is a control, and the coefficient for minority is still significant (0.01, p = 0.002), therefore the association between minority % and FAIR policies remains strong after controlling for income.

E. Minority percentage and number of FAIR policies are still associated at the ZIP code level, even after controlling for income, fire risk, and housing age. **TRUE**

- Referencing model_E, the coefficient for minority is 0.008, with a p-val of 0.006, and a confidence interval - (0.003, 0.014) which doesn't include 0 and therefore shows a statistically significant association is there after a full adjustment.