Homework 10

Problem 1: Redlining

The term "redlining" dates to the late 1960s, when it was first used by community activists to describe the practice wherein banks would "draw a red line on a map" to mark an area where they simply refused to grant loans. In areas that banks were allegedly redlining, people would find it difficult to get access to credit, regardless of circumstance or merit.

Through originally coined in the context of bank loans, the term "redlining" has since acquired a broader meaning, and can apply to any denial of service based on ethically dubious grounds—race, gender, sexual orientation, and so forth. No financial institution openly admits to redlining, of course. It's always a question of whether such practices are happening behind closed doors, with the denial of service then being rationalized in public using other, more socially acceptable reasons.

We will now consider a famous example of possible redlining in the insurance industry, from Chicago in the 1970s. To measure access to the private home insurance market, we will use a proxy: the number of policies issued by a federal program called FAIR, which offers insurance to those who cannot find coverage from private firms. Homeowners are required to carry insurance as a condition for receiving a mortgage, so if someone wants to buy a house, they must get insurance somewhere, whether through the private market or through FAIR. So the logic here is that the more FAIR policies we see per capita in a ZIP code, the less the residents in that ZIP code are able to turn to the private market.

However, private insurance companies can have perfectly reasonable, non-discriminatory grounds for denying insurance to people. Some areas of town, for example, are more subject to fire because of nearby industry, mix of building materials, or prevailing wind patterns. Other areas have mainly older houses that pose greater liability risks. In individual cases, it is therefore hard to tell whether an insurance company's decision to deny coverage was the result of redlining, or a non-race-based actuarial calculation. Nonetheless, we can try to adjust for these ZIP-code differences and look for city-wide patterns that might constitute evidence of discrimination.

A second issue with this proxy is that it is difficult to tell whether people must turn to a FAIR policy because they are denied coverage outright or because they cannot afford the private coverage offered. Hence we also have to adjust for differences in income among different areas that might explain differences in FAIR policies, even in the absence of racism.

In `redline.csv`, you are given data on 40 different ZIP codes in Chicago. (All we have is the ZIP-code-level data, rather than data for individual homeowners.) For each ZIP code, you have the following information:

- `policies`: new FAIR plan policies and renewals per 100 housing units in that ZIP code. Remember that more FAIR policies means that residents in that ZIP code are buying private insurance at lower per-capita rates. It's our proxy for access to the private market, where more policies implies less access.
- `minority`: percentage of residents in that ZIP code who, on the last US census, self-identified as a member of racial/ethnic minority (i.e. other than non-Hispanic white).
- `fire`: fires per 100 housing units in that ZIP code
- `age`: percent of housing units in that ZIP code built before WWII
- `income`: median family income in thousands of dollars

Use a linear regression model to assess whether there is an association between the number of FAIR policies and the racial/ethic composition of a ZIP code, adjusting for the `fire`, `age`, and `income` variables. Write a short report summarizing your analysis and interpreting results, formatting your write-up in the following sections:

1) Question: What question are you trying to answer?
2) Approach: What approach/statistical tool did you use to answer the question?
3) Results: What evidence/results did your approach provide to answer the question? (E.g. any numbers, tables, figures as appropriate.)

4) Conclusion: What is your conclusion about your question? Provide a written interpretation of your results, understandable to stakeholders who might plausibly take an interest in this data set.

Provide confidence intervals where appropriate.

Problem 2: Grocery store prices

In this question, you will look at data on prices in Texas grocery stores in order to examine the role of "place" and brand in grocery pricing. There are at least two big issues here. One issue is the value of what marketers call "place." Customers choose where they are going to shop for a lot of reasons: merchandise assortment, convenience, atmosphere, sales assistance, additional services, and many other reasons. If those things are right, retailers can charge a higher price than competitors and customers still feel like they are getting the best deal. Whole Foods, for example, has a reputation as "Whole Paycheck": a store that charges more for the same products, presumably because of the atmosphere and experience it offers its customers. (Whether this reputation is based in fact is something you'll assess below.)

A second issue is related to the policy problem of food deserts: urban (often low-income) areas that are under-served by grocery stores selling affordable or good-quality fresh food. In such areas, the best close option for groceries might be a convenience store like CVS or Walgreen's, rather than a full-service grocery store. If these convenience stores charge more for the same product – perhaps due to lack of scale – then some of society's most under-privileged citizens might end up paying more for food than everyone else.

Description of data

To look at these issues, pricing data on 40 different products was collected from 16 different stores in Houston, Austin, and Fort Worth. The 40 products were chosen to cover a broad spectrum of widely available items. They came in 6 broad categories, as follows.

- Cereals: Honey Nut Cheerios, Lucky Charms, Cinnamon Toast Crunch, Frosted Flakes
- Drinks: 12 pack Coke, 12 pack Sprite, Almond Milk, High Brew Black Coffee, La Croix Box, Land o Lake Half & Half, Smart Water 23.7 fl oz
- Staples: Horizon 2% Milk, Carton of 12 Eggs, Gold Medal Flour 5 lb, Imperial Sugar, Iodized Salt
- Fruits: Banana 1 lb, Granny Smith Apple 1 lb, Navel Orange 1 lb
- Snacks: Blue Bell Vanilla Ice Cream 16 oz, Campbell's Chicken Noodle Soup, El Milagros Tortilla Chips, Kind Oats and Honey Granola Bar, Nature's Own Whole Wheat Bread, Nut Thins 4.25 oz, Saltine Crackers 1 lb, Progresso Chicken Noodle Soup, Skinny Pop, Stacy's Pita Chips, Tostitos Bag, Wheat Thins 8 oz
- Spread/Sauce: Good Flow Honey 16 oz, JIF Peanut Butter, Justin's Almond Butter, Mom's Tomato Sauce, Welch's Grape Jelly

Each row in the data set corresponds to a *single product in a single store.* The variables in the data set are as follows:

- Product: one of 40 products listed above (e.g. 12 pack Coke, Campbell's Chicken Noodle Soup, etc.). Not all stores carry all products (see Part B below).
- Category: one of six product categories listed above (cereal, drink, staple, snack, spread/sauce, fruit).
- Price: in dollars
- Type: which type of store? "Convenience" means CVS/Walgreens. "High-end" means a stereotypically fancy store like Whole Foods or Central Market. "Natural" means a store focusing on natural/organic products: here, Wheatsville Co-op and Natural Grocers. "Small format" is an ordinary grocery store in a physically smaller location: here, Kroger Fresh Fare in downtown Houston, and Fresh Plus in West Campus. Everything else that's a full-sized grocery store—Albertsons, Fiesta, Kroger, HEB, Target, and Wal Mart—falls in the "Grocery" category.
- Store: which store? There are 14 store names and 16 stores in the data set: HEB Houston and HEB Austin are both called HEB; likewise, Whole Foods Houston and Whole Foods Austin are both called Whole Foods. That's why you see two rows for HEB and Whole Foods for each product.
- Income: average income in the ZIP code where the store is located, in dollars.

The questions

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.

Part B. Your plot in Part A shows differences in average price across stores. But a big issue with this plot is that not all stores sell all products. Some products (eggs, milk) are sold in all stores. But others aren't. For example, Walgreen's wasn't selling Greek yogurt in that store on that day; and you're about as likely to find Lucky Charms at Wheatsville Co-op or Whole Foods as you are to find an actual leprechaun. Therefore, if you want to make an apples-to-apples comparison of prices, you can't simply ask which grocery store charged the highest average price across all products.

To illustrate this fact, 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.

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?

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?

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`? To answer this question: fill in the blank: "A one-standard deviation increase in the income of a ZIP code seems to be associated with a (**blank**) standard-deviation change in the price that consumers in that ZIP code expect to pay for the same product." Remember standardized coefficients from the textbook.