**From:** Daniel Posthumus

**To:** CDTFA

**CC:** Ryan Cummings and Neale Mahoney

**Date:** 12/20/2024

**Subject:** Hypermarket Event Study Data Build

**Summary**

This memo provides the background to the data that we used to construct the write-up, “Hypermarket and Low-Priced Competition Event Study Analysis”. Specifically, it covers the following:

1. I distinguish between a location-based and an owner-based Station ID variable.
2. I create an inclusion rule for our dataset.
3. I summarize the different types of events we can study:
   * I compile statistics of firms undergoing each change.
   * I map the relative frequency of these events in the Bay Area as an example of geographic heterogeneity.
4. Then, I focus on constructing our event study data:
   * I create geographical circles capturing firms within a 3-mile radius, 3-10 mile donut, and 10-20 donut of stations undergoing events.
   * I create a transition matrix to understand what shares of brand changes match to certain combinations of store brands.
5. Finally, I conclude by analyzing hypermarkets:
   * I explore different ways to define hypermarkets, or “branded bargain stations”, either by 1) a predefined list of hypermarket store brands or 2) price relative to the unbranded average.

The takeaways are the following:

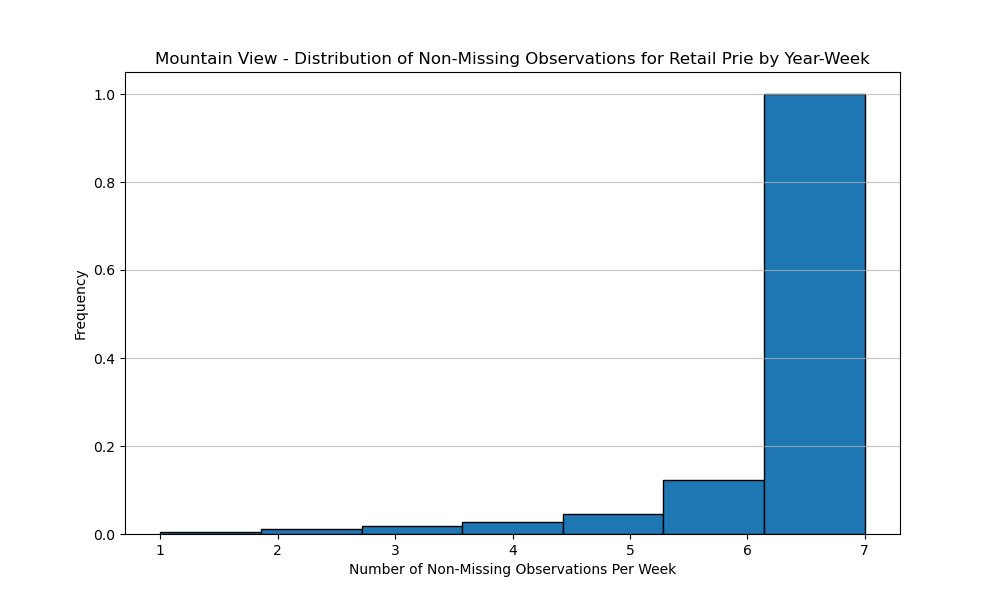
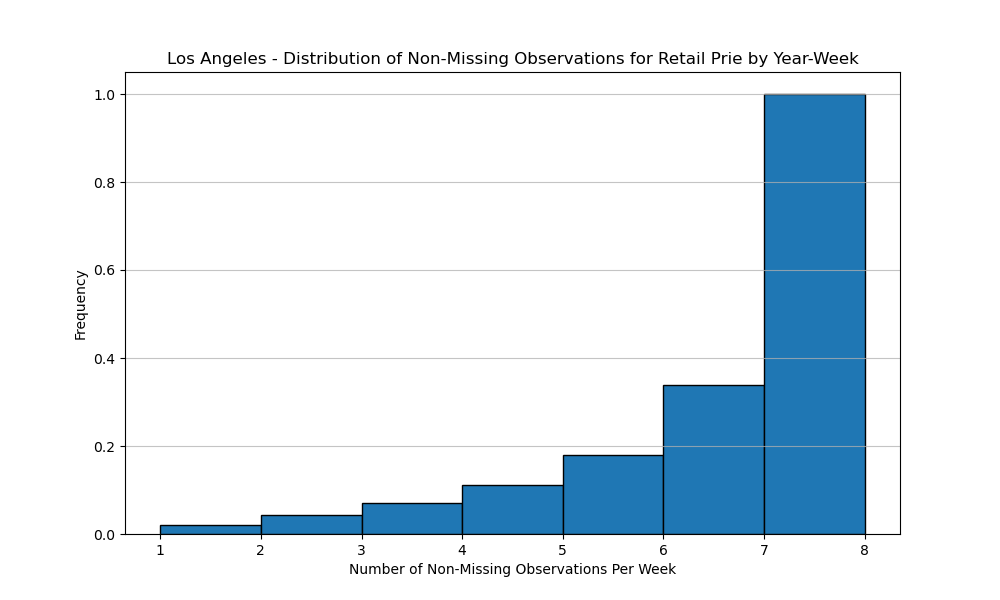
* The station ID pre-included in OPIS only considers location; thus, if there is a change of ownership/brand at a station while its location remains constant, OPIS’s Fuellocation ID variable remains unchanged. To capture changes in brand, I create a new ID unique to location / brand combinations.
* After some analysis of the data, I only include stations that, for a given week, have at least five non-missing observations.
  + The attrition rate resulting from this inclusion rule is heterogeneous across geography (i.e., Mountain View’s attrition rate is lower than Los Angeles’).
* On average, exits in the data occur nearly two years earlier than entrances.
* Exiting firms have margins greater than the average station, while entering firms have margins lower than the average station.
  + Unbranded stations are much more common among *both* entering and exiting firms than either firms undergoing brand changes as well as the average station.
  + This is a puzzle; intuitively, we’d expect firms to exit because of lower margins/profits. However, margin alone gives an incomplete picture without quantity; perhaps the stations didn’t sell enough gasoline to recoup fixed costs. More analysis into the characteristics of exiting/entering firms is needed.
* Based on price dynamics around events in Los Angeles:
  + Firms undergoing brand changes have higher prices than nearby stations *before* the change, and lower prices than nearby stations *after* the change.
  + After entering from nothing, firms have lower prices than nearby stations.
  + Between (approximately) 52-20 weeks *before* exiting to nothing, firms have lower prices than nearby stations; then, around 20 weeks before exiting, firms have approximately the same prices as nearby stations before exceeding nearby stations’ prices immediately before exiting.

**What ‘Fuellocation ID’ Identifies**

Fuellocation ID identifies the *Location* of a gas station–denoted by its Latitude/Longitude coordinates. It does *not* uniquely identify the ownership/brand + location of a gas station. **If a gas station changes ownership but remains in the same physical location**, **its Fuellocation ID variable** **won’t change.**

To capture changes in ownership, I crafted a new ID variable, ‘Unique ID’ which combines information from ‘Fuellocation ID’ and ‘Store Brand’. The biggest difference between the two ID variables are one-month spikes in the ‘Unique ID’ count; these occur because this new ID ‘double counts’ stations that change ownership within a month.[[1]](#footnote-1)

**Defining The Inclusion Rule**

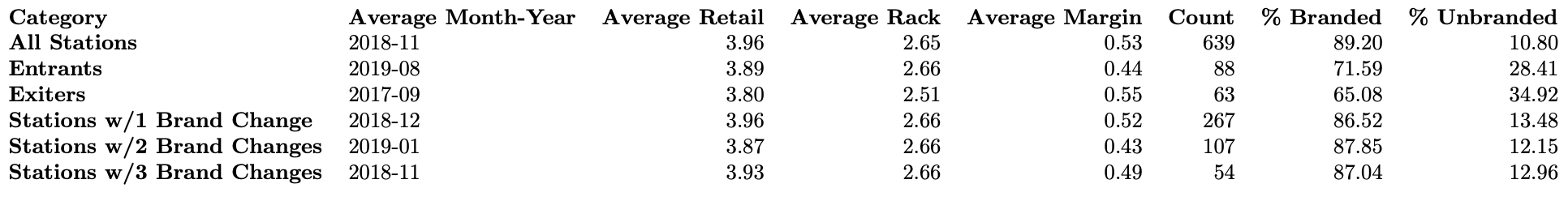
Here is a histogram showing the cumulative relative frequency for the number of non-missing observations per *week* for Los Angeles and Mountain View. To interpret the graphs, for example in LA, less than 20% of all year-weeks with observed gas station data have 5 or fewer non-missing observations. Therefore, I set the inclusion rule to the requirement that a gas station, for a given week of data, contain at least 5 non-missing observations of price. The rest of the data works with data filtered as such.

**Different Types of Exits/Entries**

There are 3 possible events that can occur to a gas station:

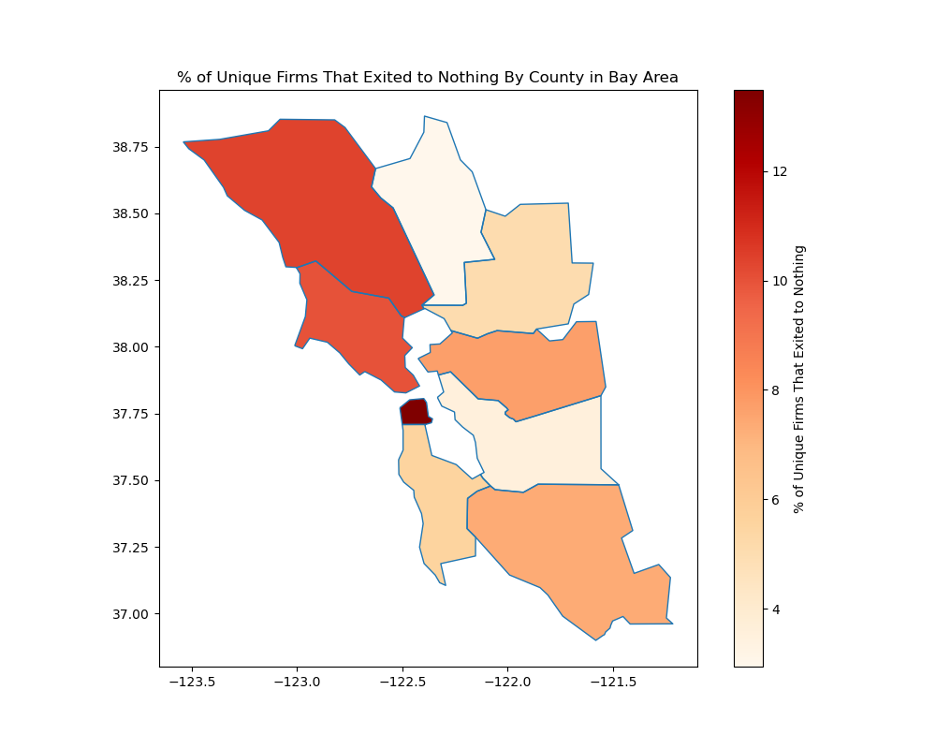
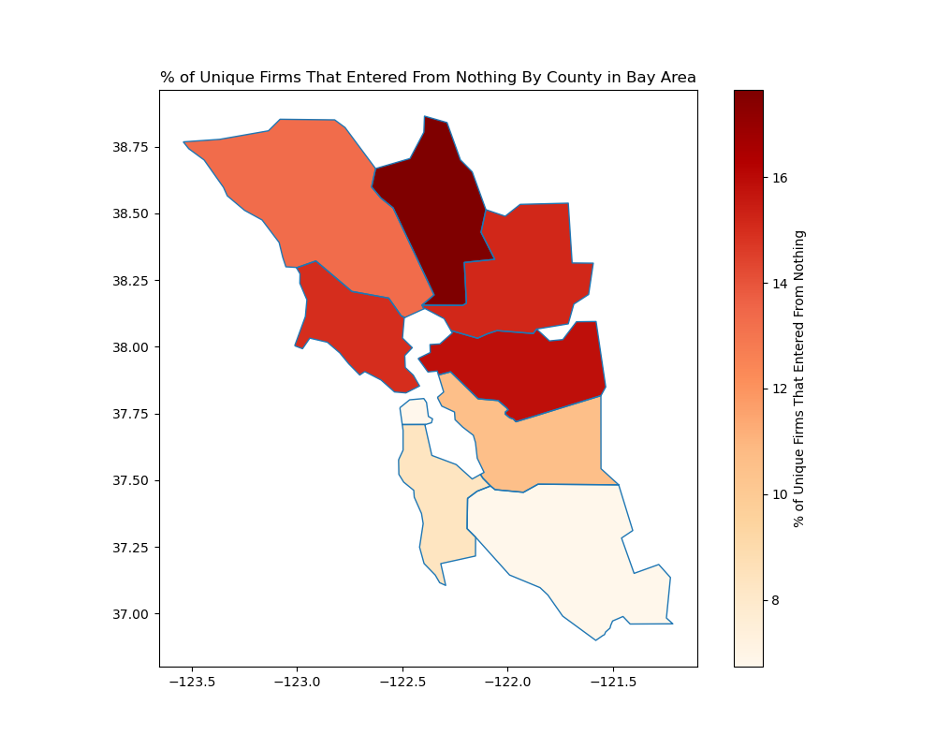
1. Entry from nothing
   1. This occurs when there is a completely new location; i.e., if a station appears at a location at which there was previously not a gas station.
2. Exit to nothing
   1. This occurs when there is no new gas station, under different ownership, that appears after the final observed instance of a gas station.
3. Ownership/brand change
   1. This occurs when ‘station id’ changes but ‘Fuellocation ID’ doesn’t; i.e., the brand of the store changes while the physical location doesn’t.

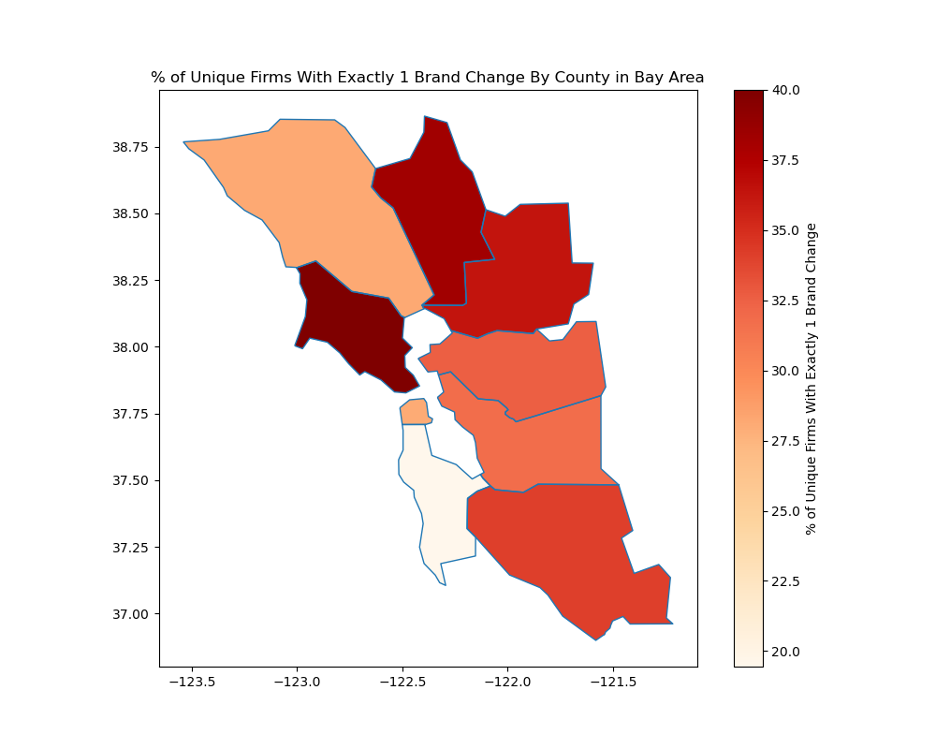
**Summary Statistics Table**

The below table is a summary statistics table, including *only* Los Angeles:

**Descriptive Maps**

To get a sense of geographical heterogeneity in the relative frequency of events, I put together the following maps for the Bay Area, which captures the % of unique gas stations undergoing a particular event on the county level.

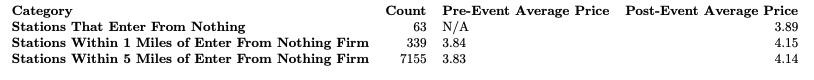
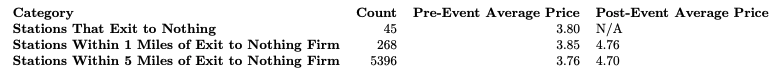
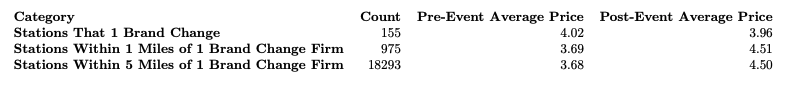




**Event Study Dataset Creation**

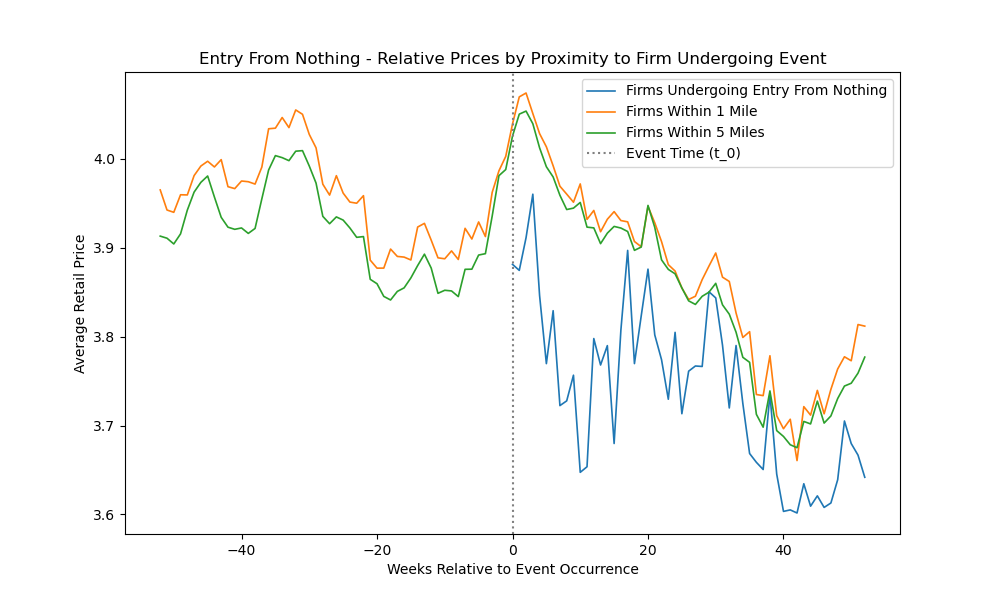
I’ve successfully created code that, for each firm, pulls all stations that are within *x* miles of the station undergoing the event. For now, I am focusing on stations within a 1- and 5-mile radius of the station that is directly undergoing the event, where there station undergoing the event is the center.

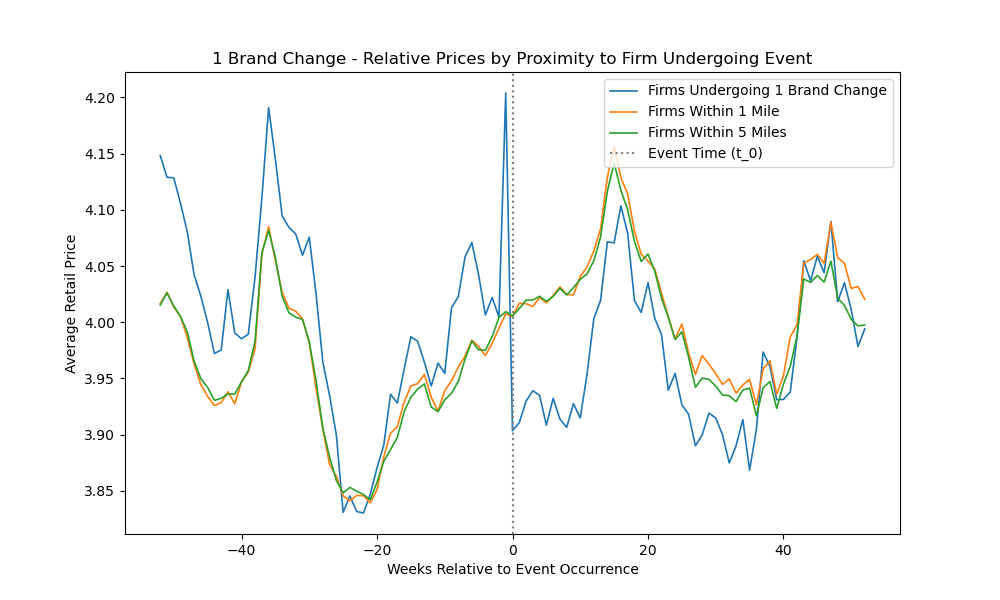
Here is a series of tables capturing summary statistics for firms directly undergoing events, and those within these donuts by event specifically for Los Angeles. I focus on Los Angeles because of the computation expense of running this analysis; we are working to get access to Stanford’s high-power computer cluster and then we will be able to run this code for statewide data.

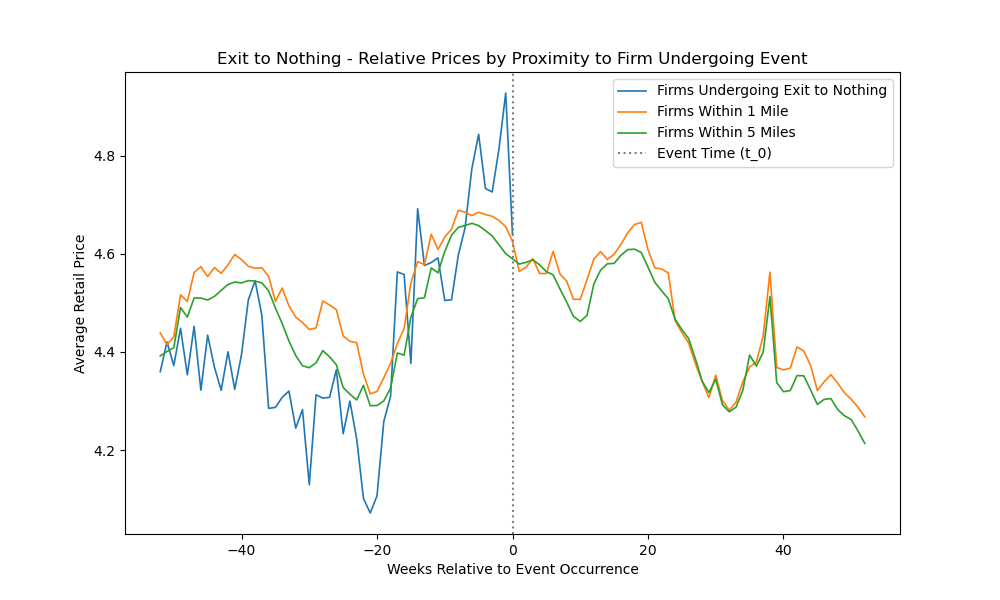
****

**Price Behavior Around Events**

These tables merely present *static* summary statistics; they give us no indication of the *dynamic* pricing behavior of firms undergoing events or the firms within their donuts around events. For this, I plot the average prices for firms undergoing events and those within a 1-mile or 5-mile radius, by relative week (relative to the event occurring).

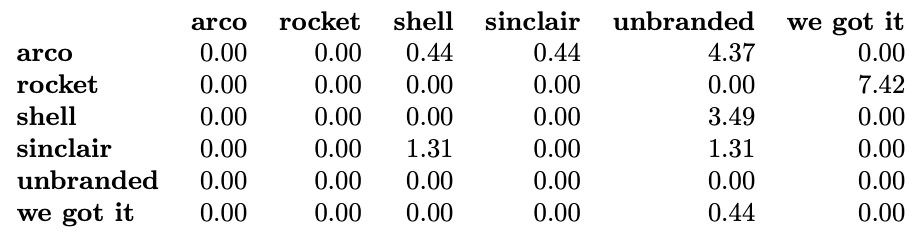
The figures below suggest that: 1) entering firms have lower prices than those near them; 2) firms with brand changes typically have higher prices before the brand change, lower prices immediately after the brand change that converge with prices of stations near them; and 3) firms that exit to nothing have lower prices until about 20 weeks prior to exiting, at which their prices appear very similar to those near them, and then immediately before exiting they look to have higher prices.****

****

****

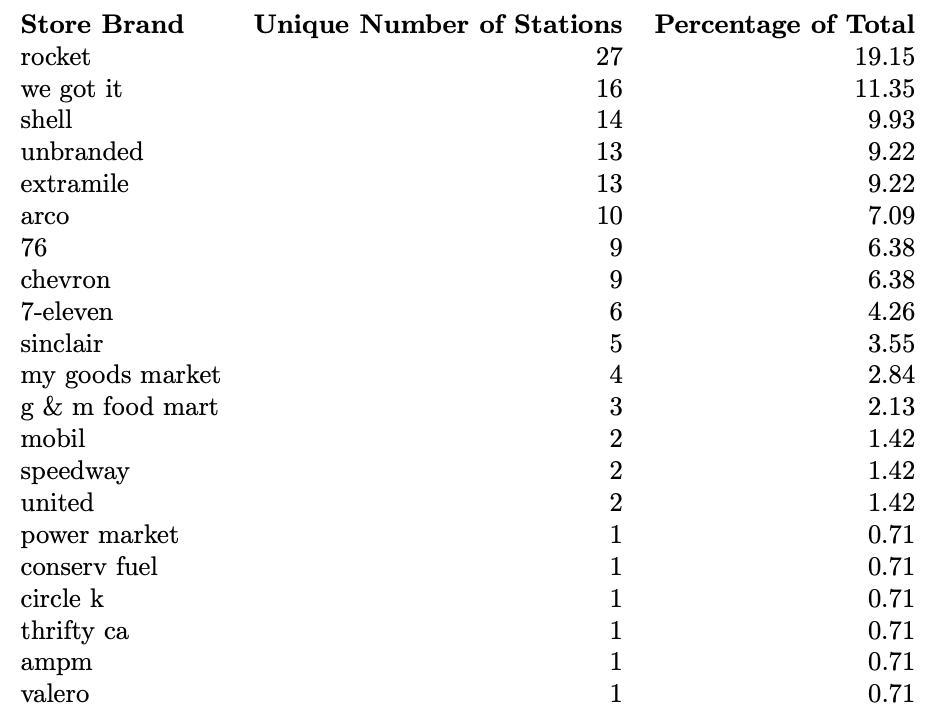
**Transition Matrix**

Below, I’ve created a transition matrix, which captures the relative frequency of brand changes by pre- and post-change brands, to the 5 most common brands, + ‘unbranded’ with at least one brand change for Los Angeles. For example, 0.44% of all brand changes in Los Angeles were stations switching from Arco to Shell. 4.37% of brand changes in Los Angeles were stations switching from Arco to Unbranded.



Note the large shares of stations transition *to* being unbranded, while among the top 5 brands, there is not a single brand change where a station transitions *from* being unbranded to being branded.

Then, here is a table with the relative frequency of each brands among gas stations in Los Angeles:

****

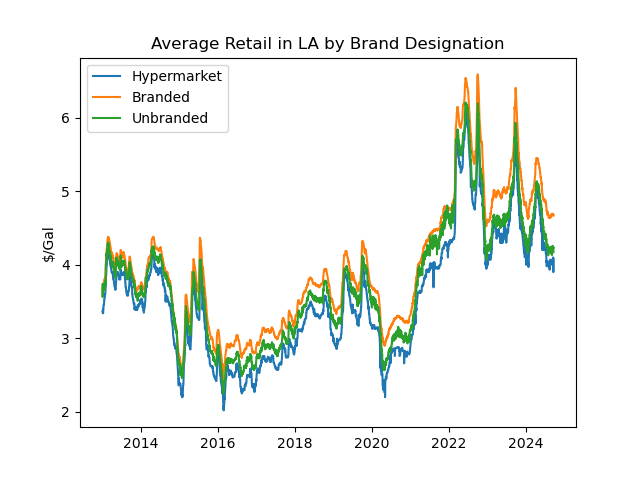
**Hypermarket/Branded Bargain Stations Definition**

The CEC/CDTFA working definition of a hypermarket is a store/brand who primarily sells something other than gasoline, i.e. large retailers that sell branded gasoline at very low prices relative to *both* branded and unbranded gasoline.

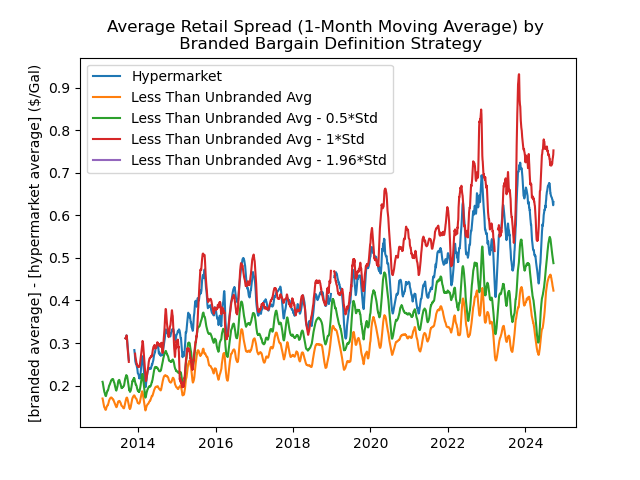
Looking at all brands within Los Angeles, here are the following that under the above definition I would consider a hypermarket (with the number of stations belonging to each brand in parentheses):

* Costco (2 stations)
* Food 4 Less (1 station)
* Ralphs / Kroger (1 station)

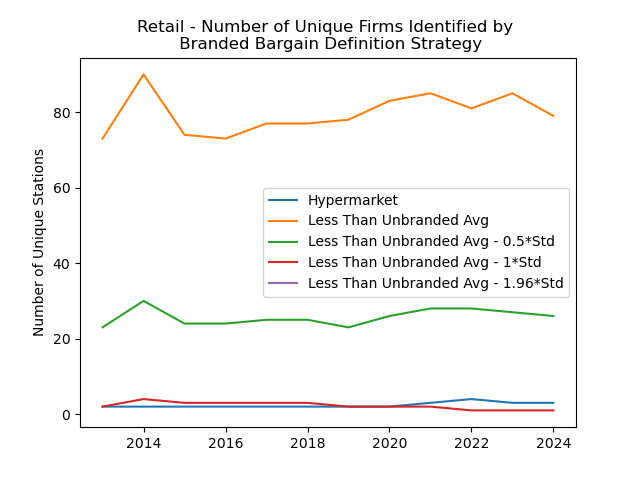
To check the price characteristics of these stations, I have plotted below the average retail prices over time for 1) these hypermarkets, 2) all branded gasoline stations, and 3) all unbranded gasoline stations in Los Angeles:



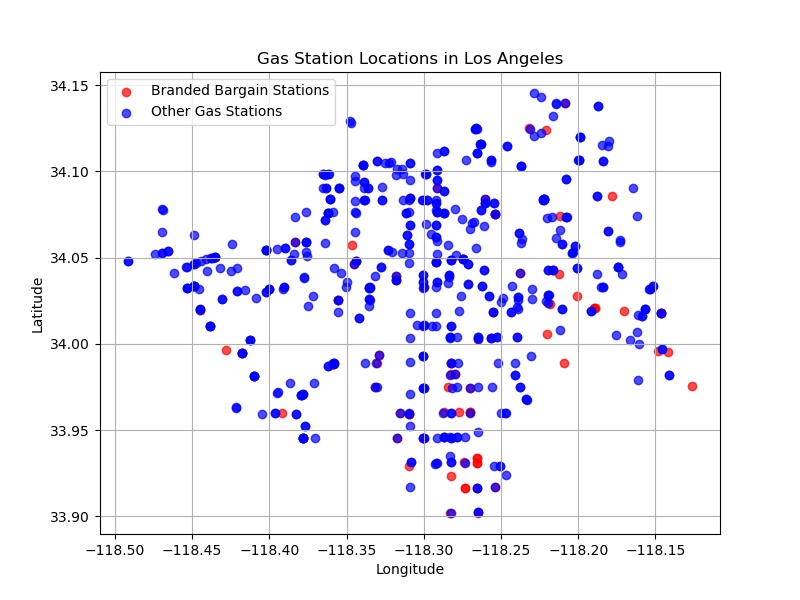
The only problem with moving forward with this designation is that it leaves with an incredibly small sample of hypermarkets to analyze, specifically 4 stations. Thus, another approach is to define “branded bargain stations” by their relationship to the average unbranded price.[[2]](#footnote-2) Below I plot the average price spread [(branded average) - (subsample average)], by different restrictions on prices:



This expanded definition also gives us a much larger sample to work with; below I’ve plotted the share of all the gas stations in Los Angeles that would satisfy each of the requirements describe above:

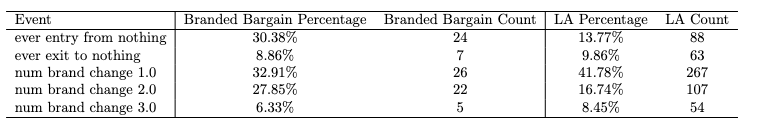


We can visualize the geographical distribution of these hypermarkets in comparison to all gas stations with the following scatterplot (in the following scatterplot, a hypermarket is defined using the 0.5\*standard deviation requirement explained above).



This scatterplot suggests to me that there is a bias to the geographic distribution of the “Branded Bargain Stations”; this may be because we’re comparing these stations to the city-wide average, so they’re concentrated in lower-priced parts of the city. One way to refine and make more precise this inclusion requirement is to compare stations’ prices only to those within 5 miles; for example, we could classify as a “branded bargain station” any gas station that sells at least 0.5 standard deviations below the average retail price for all unbranded stations within 5 miles. We are open to suggestions as to how to define the “branded bargain stations”.

Finally, we can check if there is sufficient variation in event occurrence among these branded bargain stations (using the 0.5 standard deviations below the unbranded average criterion):



This suggests that stations entering from nothing are over-represented among the branded bargain stations; this makes a great deal of intuitive sense, since, as we’ve seen in the price dynamics graphs, it appears entering stations try to undercut competitors’ prices.

1. Consider a station owned by Shell that changes ownership halfway through the month to Texaco and has six observations under Shell’s ownership and 12 observations under Texaco’s. Previously, using only location this would be identified as only one unique station for that month. Now, using ownership and location, this is identified as two unique stations for that month. [↑](#footnote-ref-1)
2. Our preferred requirement is that a branded gasoline station sell at least 0.5 standard deviations below the unbranded average for all of Los Angeles on at least 50% of days in the data. I also plot the average price spreads for 1) stations that sell below the unbranded average for all of Los Angeles on at least 50% of days in the data and 2) stations that sell at least 1.96 standard deviations below the unbranded average on at least 50% of days in the data. [↑](#footnote-ref-2)