A2: Speculation Dataset Exploration

The chosen dataset contains Boston residential sales data from 2000 to 2023 and abundant information about the seller, buyer, price, location, kind of building, if the sale was a flip, etc. Before this assignment, I knew very little about speculation, and after the background reading, I was fascinated with the practice because flipping is a hobby I know many of my friends' parents participate in. That becomes one of my initial questions, along with other curiosities in flipping strategies I was excited to explore as I traversed the rich dataset.

Initial Questions

1. Are individual or small investors contributing to the negative impacts of speculation? Are wealthy individuals that buy and flip houses part of the issue, or is there impact negligible?
    1. **Motivation:** Many people I know personally are doing this.
2. How do buyers choose property to invest in? Do they choose residences in their geographic area? Are certain traits of investors correlated with certain traits of residences? For instance, institutional investors are more likely to invest in houses of a particularly wealthy or historical zip code.
    1. **Motivation:** This could help us understand how investors choose their property, and thus create policy that targets harmful speculation practices.
3. Which sales make the most profit? What types of investors make them? What kinds of homes? How much flip time?
    1. **Motivation:** This information could help us identify factors that make speculation enticing, and provide insight into how to reduce the incentive.

Phase 1: Exploration

Checking Spatial Data
[locations.png]

I first began by exploring the spatial data, mapping out the latitude and longitudes. I saw a few incorrect values at (0, 0), so I filtered those out. I then found a strange cluster in Maine, Northwest of Boston, all with the street, Union Park. I checked their zip code, and confirmed that they were meant to be placed at Union Park St in Boston. I fixed those points by remapping their latitude and longitudes to that of Union Park St, Boston. Otherwise, I confirmed that the spatial data looked clean, including zip codes, street names, and occasionally, addresses.

Search for Individual Sellers
It initially confused me that there wasn't a category for individual sellers, i.e., regular people just selling their houses. There was information about whether each seller was a bank, government entity, LLC, trust, etc., but none indicated the kind of seller I had expected to see.

[usage-indiv??.png]
[style-indiv??.png]
I attempted to test a proxy for this group by plotting home usage and style against sellers that were not any of these large entities. I was searching for some kind of signal that the non-entities were individuals, such as them selling single family homes disproportionately more than entities.

[entity-vs-nonentity-total]
Since non-entities have many more sales in general, I was not confident in labeling non-entities as these regular individuals. Upon further research, I found that wealthy individuals frequently sell through LLCs and trusts for liability protection, tax planning, etc., and learned that searching for this "individual" did not align with the nature of this data, which records how sellers' legal form rather than social.

Big Buyers Buy from Big Sellers
I then performed a sanity check that "big" buyers should buy from "big" sellers. Specifically, attributes that make buyers "big," such as being institutional investors, owning many properties, having high holding value, etc. bought from similarly "big" sellers.

[buyer-seller-type-heatmap.png]
The heatmap that visualized frequency of sales between all types of buyers and sellers exhibited a clear diagonal, confirming that big buyers and big sellers are strongly correlated.

[big-buyers-sellers-big-price.png]
I then wondered what it means for a sale to be "big," and if big buyers and sellers make big purchases. I decided to use the price of the sale as my "bigness" proxy, and by plotting price across both buyer and seller types in a line chart, I verified that the average price of homes by buyers and sellers of the same type are very similar as well. Big buyers make big purchases from big sellers.

Time Trends

With over 20 years of data, I created basic exploratory visualizations of key variables over time.

[price-over-time.png]
First, I plotted average price and flipped profit over time to view large temporal trends in the housing price points. They both seemed to be steadily increasing, although they appeared to have substantial volatility. I found that averaging real-estate sales was ineffective due to being skewed by extreme high-value outliers. I switched to medians to more accurately reflect the

typical property in the market. After switching, I found a much smoother upward trend in price and price difference over time.

I also experimented with spatial animations over time to get a sense of how sales behavior moved around Boston over time.
[cash-animation]
I visualized where cash sales seemed to be densest over time, and saw a peak around 2012 (which I confirmed with a basic line chart), but not explicit movement to and from a particular area.

[percent-investor-animation]
Next, I explored which zip codes investors were targeting throughout time. By coloring zip codes by the number of purchases made by investors in the sales of that area, I could easily grasp that investors have always been drawn to Back Bay but are steadily expanding their reach across Boston.

Price Distribution
[price-distribution.png]
Lastly, I looked at a logarithmic distribution of price across types of buyers for a high-level overview. Institutional buyers have a very wide range of sales, with the highest max by orders of magnitude. However, their median sale is a bit lower but comparable to other buyer types. Small and non-investors are the most concentrated groups, with most of their sales falling between around $300,000 and $2,000,000.

Phase 2: Answering Questions

Q1: Are individual or small investors contributing to the negative impacts of speculation?

Since my initial exploration found that legal identity doesn’t map cleanly onto economic behavior, the relevant distinction is not whether an investor is an individual or an incorporated entity, but whether they operate at a small scale. For this reason, I used the “Small Investor” category to capture small-scale actors regardless of legal structure.

[flip-frequency-buyers.png]
I first checked if small buyers were disproportionately associated with short-term flips, an indication of rapid turnover. If small investors drove rapid turnover activity, they would be raising bidding pressure and amplifying price cycles. I found that although small investors perform the most short-term flips, the relationship is proportionate since small investors perform more flips overall. However, it is clear that small investors play a very large role in the flipping scheme.

[small-over-time.png]

Since my last visualization was an aggregation of all data over the 20 years, I checked the investor composition across time to see if small investors were a growing force in flipping. Mapping the percentage that small investors made of buyer investors over time, however, it appeared as if small investors have consistently made up the largest part of the investor composition, and their share has remained relatively stagnant over time.

[year-built-flips.png]
Lastly, I investigated if small investors were disproportionately targeting historic residences. Older housing is often cheaper because they're located in lower-income or transitional neighborhoods, so targeting older housing can amplify affordability pressure if paired with rapid resale and high markups. Conversely, it seems like medium investors target the oldest residences, while small investors tend to purchase newer homes.

Small-scale investors don't appear to disproportionately drive short-term speculative flipping. Their market share has remained relatively stable over time, and they are more active in newer housing stock rather than older distressed properties. The evidence does not strongly support a narrative that "regular wealthy individuals" are increasing speculative pressure in the Boston residential market.

Q2: How do buyers choose property to invest in?

[buyer-zips-map.png] [buyer-zips-heat-no-non.png]
My initial instinct was to find zip codes or regions that attracted certain types of investors. After looking at the zip codes with the most sales and how buyer types were distributed across them, I couldn't find a clear correlation between any buyer type and zip code. All groups seemed evenly spread between the top 6 zip codes with most sales. Since spatial quantities were difficult to interpret, I created a heat map excluding non-investors due to scale dominance and to focus on investors, and found no abnormal associations between ZIP code and buyer type.

[riviera.png]
I wondered if the previous zip code visuals lacked nuance because they were aggregated over the past 20 years, so I studied the same heatmap but for each year rather than overall. At first, I observed the same pattern of smaller investors buying more in all zip codes, except for one or two spikes in institutional investment in certain years. However, once I looked more closely at these spikes in institutional investment, I found that they were caused by units of a large building being sold as separate purchases. Since the colors are mapped by count and splitting the data by year led to much lower values, these large purchases heavily skewed the data.

[buyer-use-map.png] [buyer-use-heat.png]
Since it didn't seem that investor type was related to zip code, I instead looked at my next best guess for a factor that investors heavily considered, residence usage. With the spatial map, I noticed that a few usages did not follow the previous trend: non-investors largely avoided them rather than being the most prominent type of investor. In the heatmap, I excluded all usages that followed the pattern of "smaller buyer buys more of the residence" to focus on the atypical

cases. The next most common pattern seemed to be that small investors were the prominent buyer. General office spaces and retail stores, however, were bought out most by large investors. This could suggest that some property types require expertise, capital, or risk tolerance beyond typical households, and small investors are not uniformly driving all market activity. Their dominance is limited to certain property uses.

Q3: Which sales make the most profit?

I expected that bigger investors would make higher profit, so I was more interested in who makes more profit off of the flip relative to the original price. To check the quality of the profit data, I created basic graphs of price diff %. I found many abnormally large values, and after investigating their last and new sale prices, concluded that many percent differences were calculated incorrectly or exhibited divide by 0 errors because of missing prices. I filtered out the rows with missing prices, and recreated the percent price difference field by calculating (price - last sale price / last sale price) * 100. There were still some extraordinarily high percent differences, and I enjoyed researching the buildings that exhibited these extreme flips. The highest percent difference was a shocking 1 million percent for 660 Beacon St by an institutional buyer, which turned out to be the building attached to the historic Citgo sign!

Since I was focusing on profit off of flips, I also investigated the flip information. The flipped horizon column seemed corrupted, with many negative values. I chose to use the month horizon attribute instead, which had a clean domain of 0 - 24 months.

[flipped-distribution-horizons.png]
A comprehensive box and whisker plot allowed me to quickly understand how percentage flip profit was higher for bigger buyers. Coloring by the amount of months it took for each flip, there was no immediate relation between flip horizon and profit. I then wondered, though, if bigger buyers took longer or smaller time to flip, and if that had any connection to their profit. This led me to my next visual.

[flip-composition.png]
It seems like flip time is unrelated to buyer type and profit, since institutional investors actually have low flip times, potentially because they have the resources to complete their flips faster. Medium investors also seem to have shorter flip times, but the rest have higher ones. There doesn’t seem to be consistency between flip time and profit. I figure that the flipping strategies that maximize profit differ for each buyer type and flip time. I chose to split out those variables in the next visualization.

[flip-profitability.png]
My final visualization attempts to capture potential strategies of buyers according to their investor type, flip term, and percentage of their purchase in mortgage. The combination of these factors allow us to analyze the profitability of buyer portfolios with varying risk tolerance, capital, and experience. Profit percentage is placed on the y-axis to capture investment performance, while mortgage share on the x-axis proxies for leverage and financial risk. Arranging the scatter

within a grid of buyer investor type and flip term isolates differences in scale and holding strategy, allowing comparison of how leverage relates to profitability across investor groups and time horizons. Immediately, we can take away some really fascinating insights! For long-term flips of any investor type, highest profitability follows lower mortgage percentages and thus favors lower leverage risk. For non-investors, however, many are able to gain profit with medium and high mortgage percentages regardless of the flip term.

Lessons Learned

Many visuals and perspectives can be used to answer one question. One visual is never conclusive. I was surprised to find how many combinations of variables, format, filters, etc. could be utilized to explore one seemingly simple question. It took a large number of those to feel content with an investigation. Each visual required a few more simple charts to confirm certain traits about each dimension of data, and prompted a dozen more follow-up questions.

Bigger and aggregate data doesn't mean better. We're used to seeing nice strong trends and expressive, intuitive visuals, but when exploring raw data, it's very rare that you'll see what you expect by just plotting one variable on y and another on x. Achieving a clear trend often comes with some good intuition on the visualizer's part, extensive research, clever choice of chart and colors, and many layers of filtering, cleaning. Smaller, more filtered data also doesn't mean better. After some initial explorations of such a large dataset, it felt like many layers of filtering and scoping down would bring out important nuances. In many cases, however, it led to muddying by outliers and unclean data.