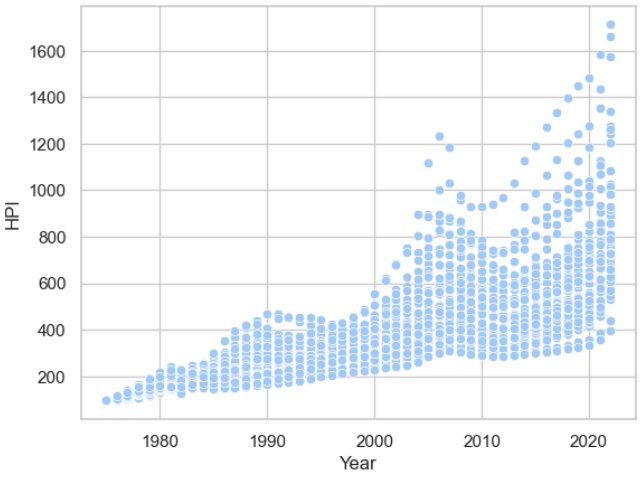
Julia Kirk
INFO 2201
Final Write-Up

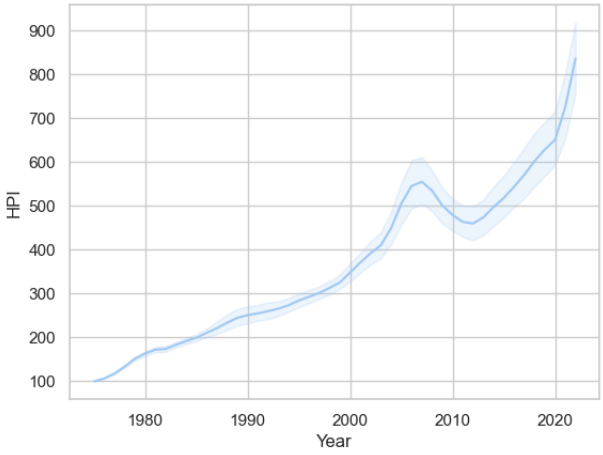
California Housing Prices

I originally started with the research question, "What trends exist in housing prices and occupancy rates within the United States?" However, as I dove into the project, I realized the scope was too large. So, I chose to focus on California Housing. My new research question is: How do California housing prices compare to pricing in other states, and how do major cities differ within California?

I started by looking at HPI (Housing Price Index) data from the Federal Housing Finance Agency. This measures the movement of single-family home prices and the average change in resale and refinancing prices over time. I created a pivot table of HPI by year for each state. This visualization showed me that there is a lot of variation between states. The HPI for each state started the same in 1975 as 100 for each state, but as the years went on, there was a lot of differentiation between the states.

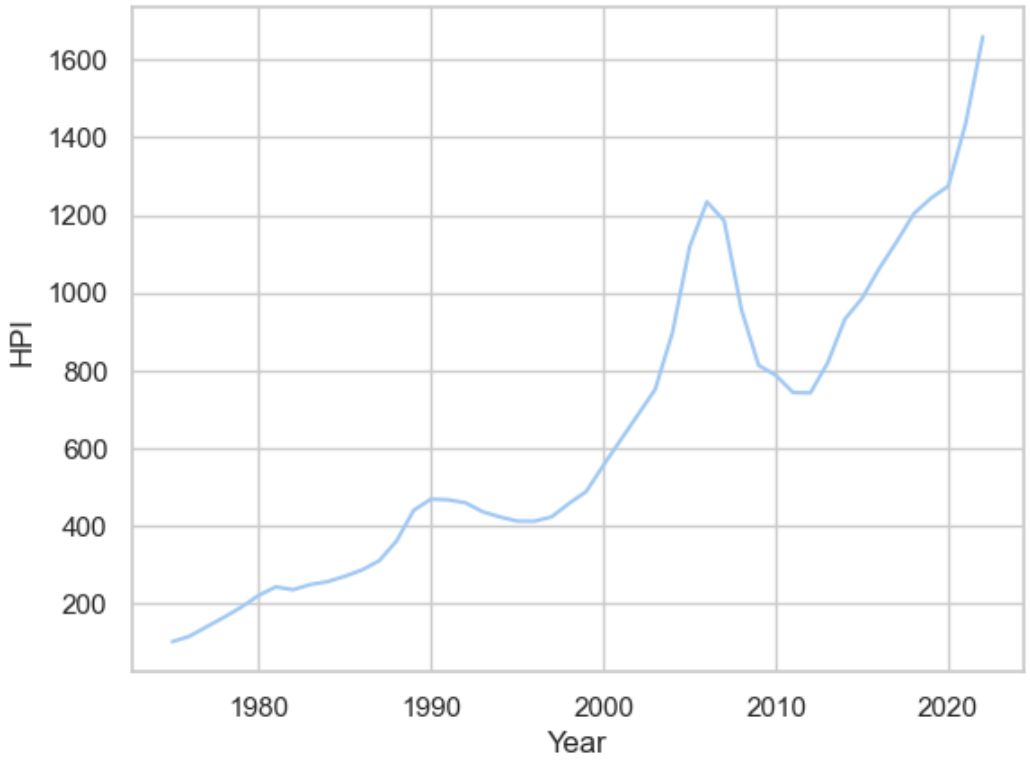
To show this difference better, I made a line plot showing the average HPI and then a scatterplot showing each state as a dot. Both of these graphs show that most states had their housing prices slowly grow, but the scatter plot showed some outliers on the top side of HPI. From there, I wanted to know what these states were.





I sorted by maximum HPI and looked at the top states. These were Washington DC, California, Washington, Hawaii, and Massachusetts. I also looked at the minimum and maximum

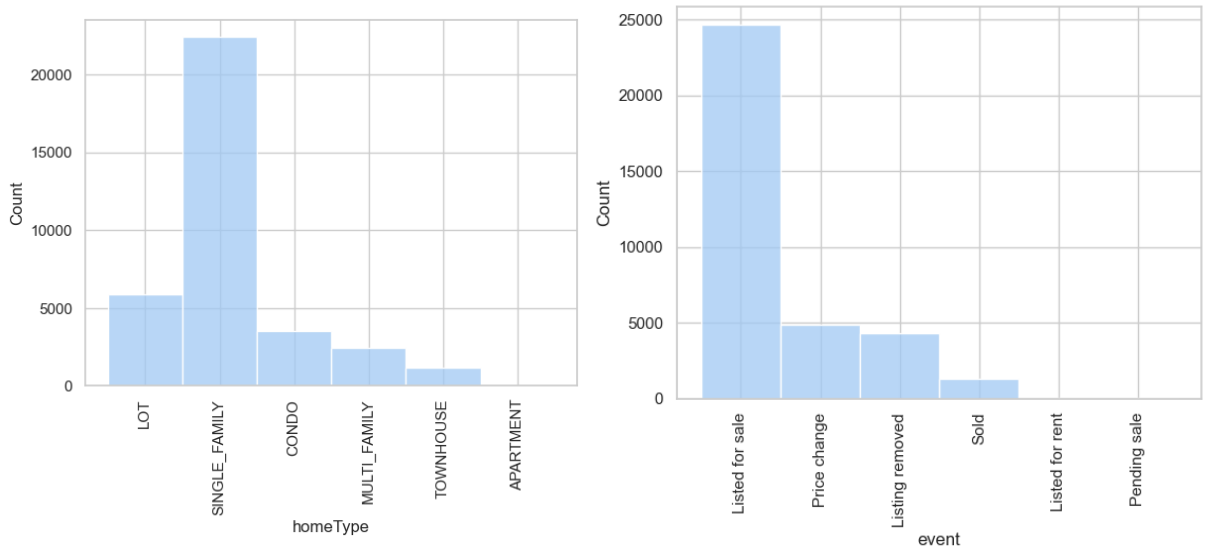
annual change in HPI for each state and realized that there isn't much difference between the maximum and minimum changes per state.



I chose to focus on California for the remainder of my project because it had a significant change in pricing over the last 50 years and has a large population with several major cities to look at pricing within and compare. It is also a state that is very well known for having a high cost of living within its major cities, but there is also a large amount of the state that is farmland as well as protected nature, and I thought that it would be an interesting case study to look at. This line plot of California HPI shows that there was rapid growth and decrease over time within California's housing market. By looking at the data visually, I can see that California's housing market was hit a lot harder than the rest of the country by the recession but also increased significantly faster in the years leading up to and after the 2008 recession. The 2022 data shows that the HPI for California 1660 is almost double the National Average of 835.5.

The following source I used was a CSV of listing data from January to June of 2021. This dataset was very useful in looking at specific property listings and price breakdowns and what features properties include. I started by looking at what types of properties were on the market during this timeframe. I noticed that most of the properties for sale were single-family homes with lots, townhomes, and multi-family homes, all with similar amounts. What stood out to me was the lack of apartments. There was a category for it, but none were listed. I wasn't sure if this

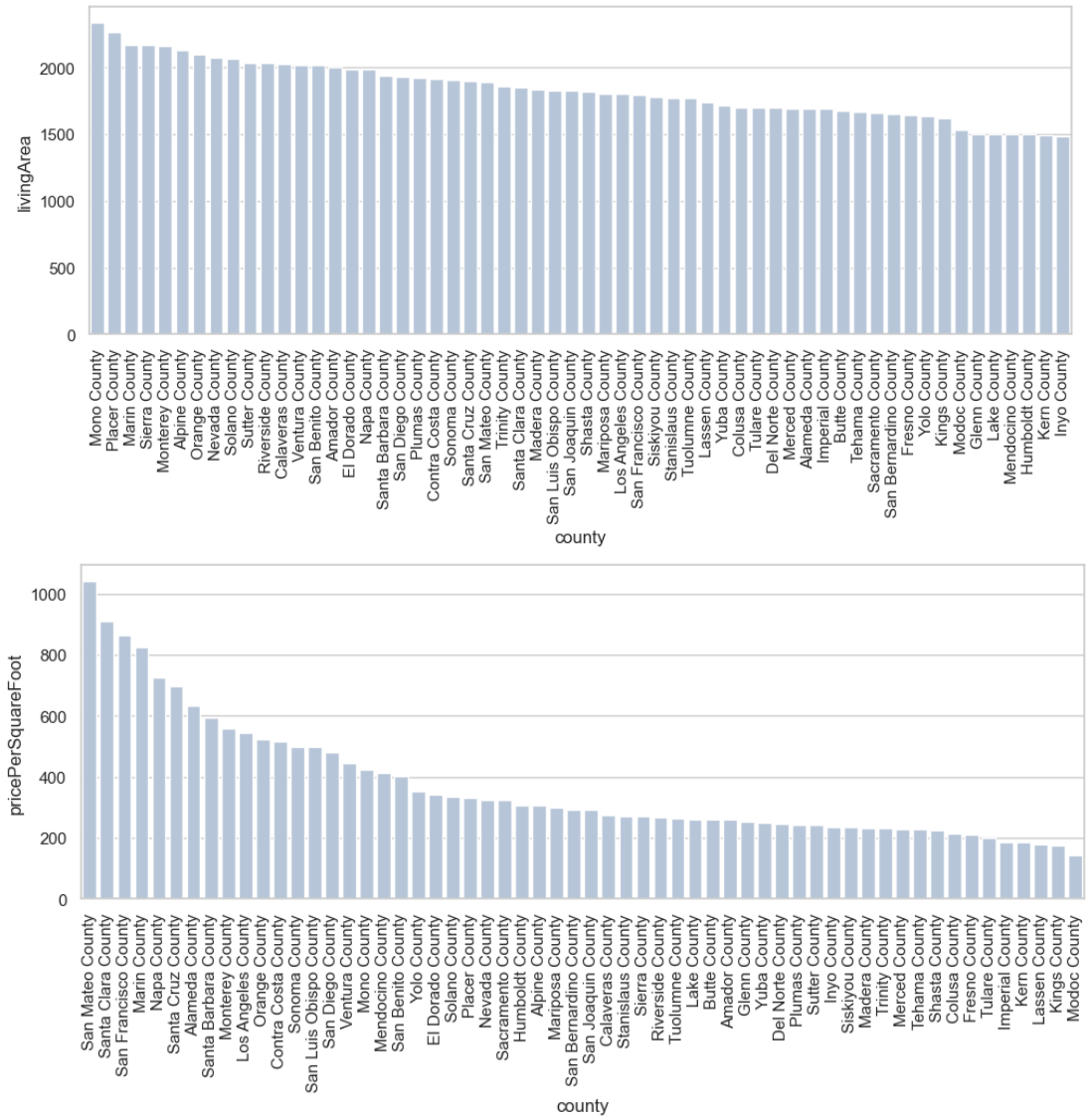
had to do with how sales of apartments are listed or if it was an issue with the data collection method.



Most of this data set, as well as the HPI data, mostly revolved around single-family homes. I decided to create a new dataframe of single-family properties to look more into features and pricing on these properties.

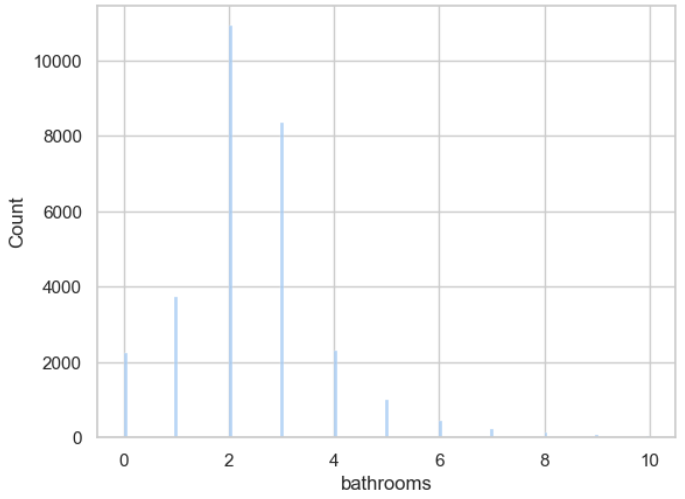
I took the mean and median price of all lot types by county and was interested in seeing where the most valuable properties were. Not surprisingly, the top four median property value counties were in the San Francisco Bay area. The rest of the top 10 counties are coastal or well-known wine regions.

I looked at the price per square foot by county. The counties with the highest cost per square foot were similar to the highest-price counties and were mostly in Northern California and the San Francisco Bay Area. I then looked at the mean and median living area for each county. The clear pattern with the top living area counties is that the majority of larger houses reside in rural northern California, around the Yosemite/ Tahoe area, or coastal areas without a major city. The bar charts I created helped to show that there is some difference in median living area between counties, but there is a major difference in price per square foot. Both of the graphs helped to further endorse the assumptions I had about housing costs in the Bay Area. The cost is very high for the amount of space you get.

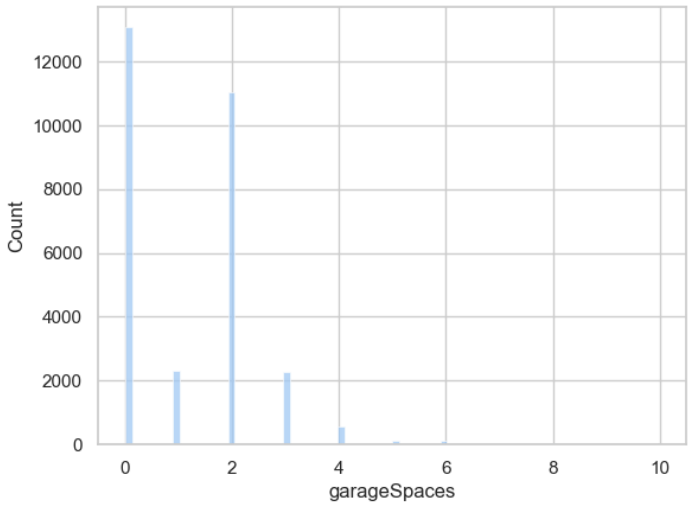


From there, I was very interested in seeing when properties were built and seeing if that was a factor that had anything to do with pricing. To do this, I created a dataframe of all lot types except for lots since this would skew the data since these still needed to be built and wouldn't help me to see what houses were for sale during the time frame. I looked at both the mean, and the median, and one thing that really stood out was there was a large difference between the two, so I looked at the minimum and maximum values and noticed there were quite a few zeros as the minimum for missing data which would explain why the means were significantly lower than the medians. I chose to sort by median because of this discrepancy and noticed that the San Francisco market had the lowest median, which made sense because of the large number of Victorian houses in the city.

I was interested in what features housing in California has and what the distribution of bedrooms, bathrooms, and parking is within the state. To do this, I created bar charts showing the number of bedrooms, bathrooms, and parking spaces. These are all factors I assumed to be pretty common across the state, with differences in parking being the main difference between rural and urban areas. These are all very standard items when looking at listings, so I assumed they were mostly accurate since they are common items to filter online.

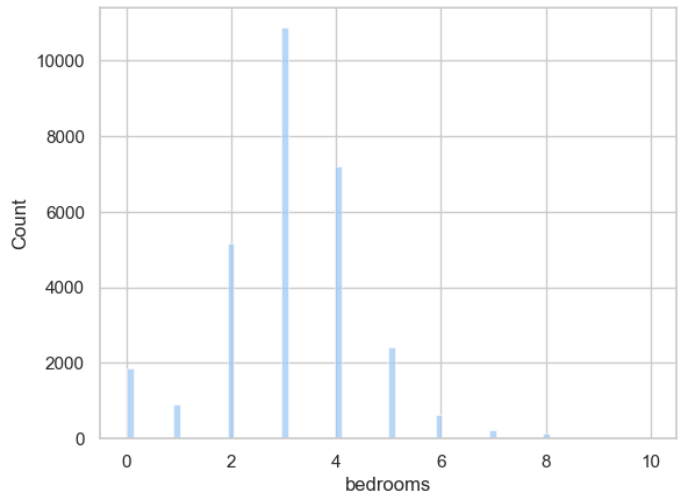


I had to change the x-axis to be smaller to be able to actually view the plots and see differences well because there were several hotels or mislabeled listings. I couldn't discern any information on the counts of any of these variables.



The bathroom count was something I found interesting. Most living spaces had between one and four bathrooms, but there were over 2000 properties that were listed as having zero bathrooms, which I found odd. Without access to all the listing data, I assumed they were just mislabeled.

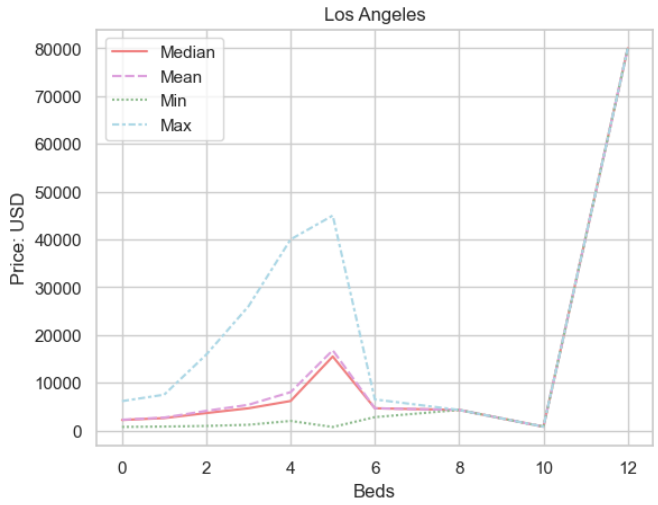
Garage spaces were mostly either zero or two, which makes sense considering the climate as well as what percentage of the state lives within cities without designated parking spaces.

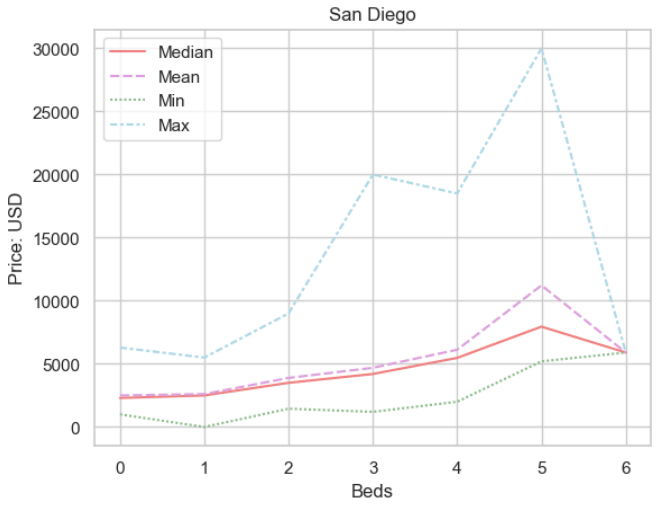


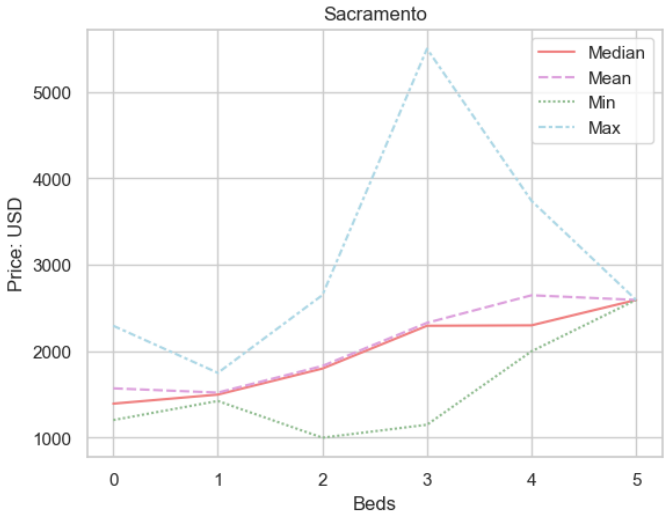
Bedrooms were mostly within the range of 2 - 4 bedrooms per unit, which makes a lot of sense. I found it interesting that there was more than one bedroom. The prevalence of studio apartments makes sense for the number of cities within the state.

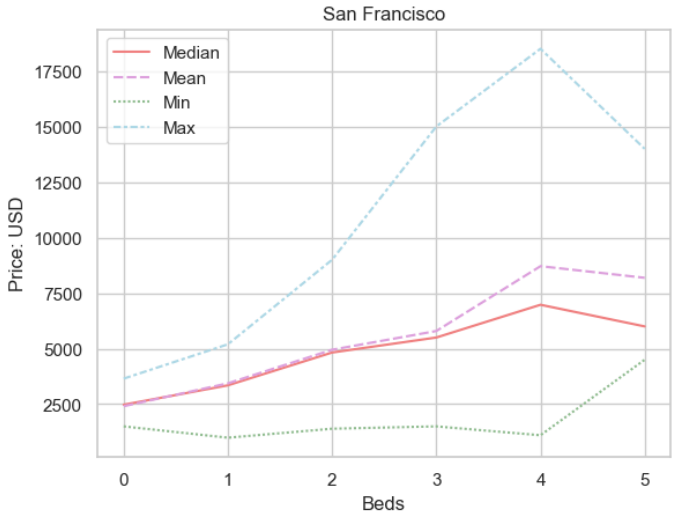
From there, I moved on to my next data source, the Mashvisor API, which shows real estate property data by city. I chose to focus on long term rental rates rather than AirBNB data since I felt it is more representative of the cost of living and housing trends in a city. I chose to focus on 4 California cities: Los Angeles, San Francisco, San Diego, and Sacramento. I chose these cities because there is location diversity as well as diversity within the industries that make up the local economies.

I created dictionaries with information on maximum, minimum, mean, and median rent prices for each bedroom amount per city. I then made them into pandas dataframes to be able to easily manipulate the data and create visualizations. I made line graphs to show the trends in rental costs by city.



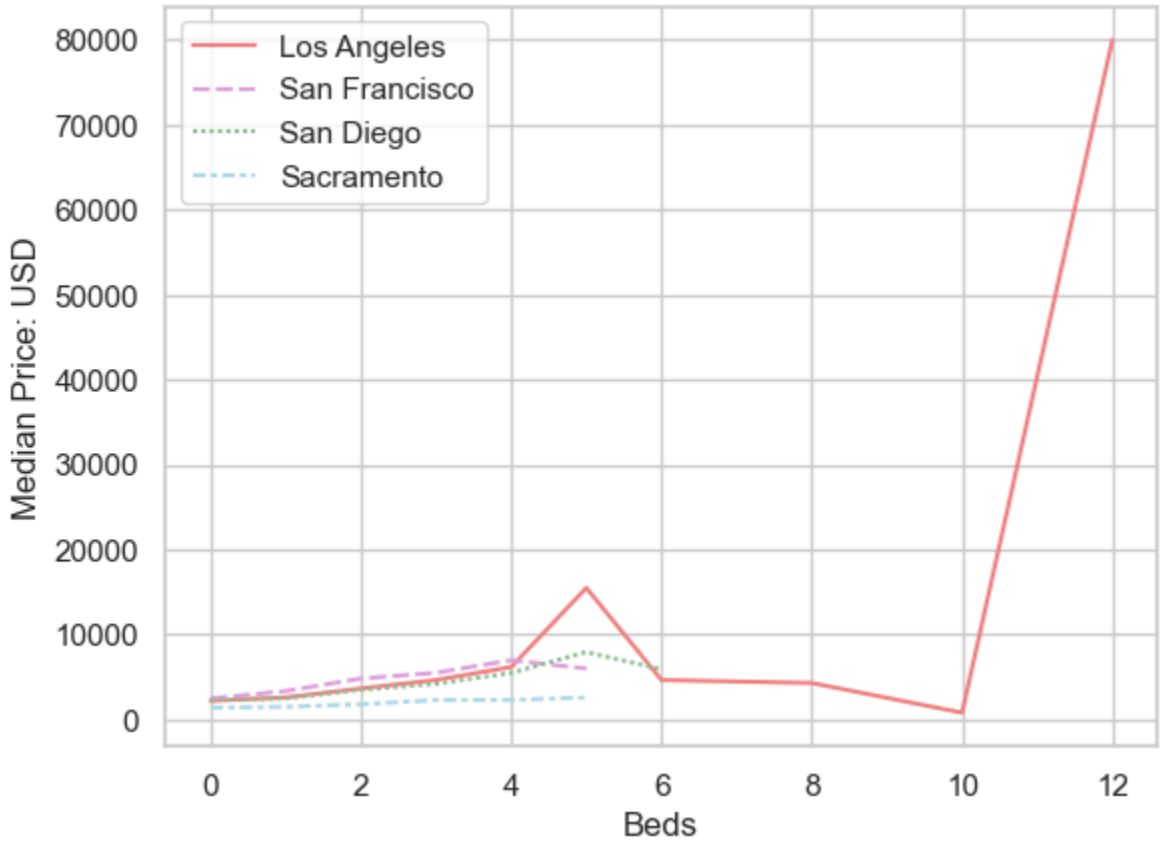






By looking at these visualizations, it helped to show that the distribution of price per bedroom is very similar from city to city, but the starting point for rent is different. Los Angeles

also has much larger houses available for rent during the time period than any other city. The maximum rent also shocked me. Since San Francisco was the most expensive city for buying, I expected them to also have higher rental rates, but the maximum rent was lower than both Los Angeles and San Diego. I wanted to compare all four cities to each other, so I graphed the median rent for each city on one graph to view the differences.



This showed that Los Angeles, San Francisco, and San Diego all had pretty similar rental rates for the number of bedrooms, and Sacramento had significantly lower rental rates than the other cities.

My overall takeaway from this project is that more desirable areas, such as cities and coastal areas, tend to have higher living costs. The square footage of housing remains pretty consistent across the state of California, but the price you pay for that area varies from area to area. Some shocking takeaways for me were about the prices of rural mountainous areas; they weren't as expensive as I had expected. I would like to be able to compare the prices in Tahoe and Big Bear areas to Colorado ski towns and see how the pricing compares.