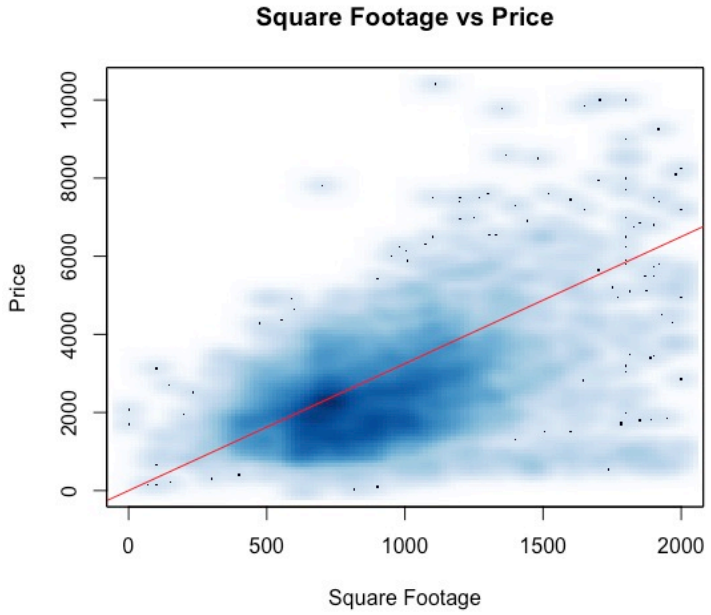
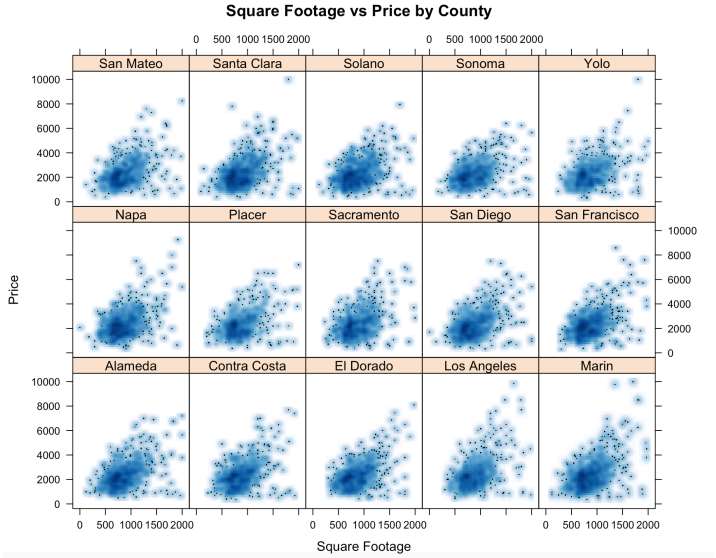
Mia Bajic
Assignment 4 Report

1. **First anomaly:** Many of the postings have duplicate posts of the same advertisements, based on the "title" variable. This misrepresents how many apartments were listed in their respective locations and how common their characteristics are, such as price, square footage, number of bedrooms, and number of bathrooms. This will clearly skew the results and give certain postings more weight compared to other postings. This can be due to the poster updating information about their posting, thereby creating duplicates. In order to clean this up, we have to create an extra copy of our dataset, order it by the date posted, and subset it and use the not operator on the duplicated function so that it gets all of the rows that are not duplicated and assigns it to the copy of our dataset. This removes duplicates but still doesn't tell us whether they were a result of spamming or updating postings.
   Second anomaly: There are postings with prices lower than $300 that do not appear to be apartment listings. Making these postings NA based on the price variable makes sure they don't create an illusion of low prices of apartments. Any posting with a price lower than $300 becomes an NA, and no longer skews results. This is done by creating a copy of our new dataset's price category and subsetting it so that any price lower than 300 becomes an NA.
   Third anomaly: In addition, there appears to be multiple postings of a moving service, eviction solutions, and other postings that do not qualify as apartment listings. If we use the grepl function to search the text of the postings for "moving" we can see which postings are not apartment listings. Then we create a variable for moving and set its True values to NA.
2. As the plot below shows, on average, as apartment size (in terms of square footage) increases, apartment price tends to increase as well. Most apartments listed are in the range of 400 to 1400 square feet and $1000 to $6000 in price.

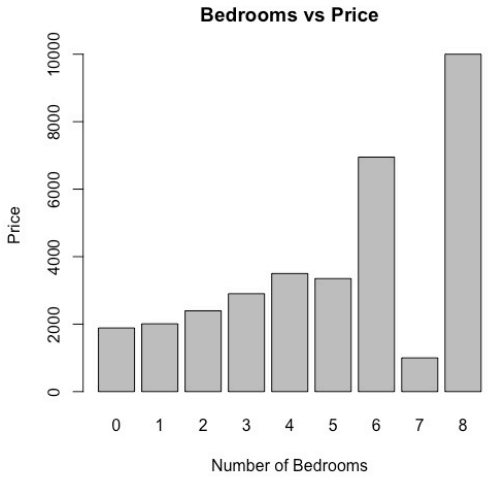


In order to take into account geographical area, a breakdown by county reveals that, most of the time, an increase in square footage results in an increase in price, but the degree to which that happens varies depending on location. Clustering is different for different counties, and the price range varies across counties. The general shape of the scatter is similar for every county, which is to be expected, but there are slight variations in each.

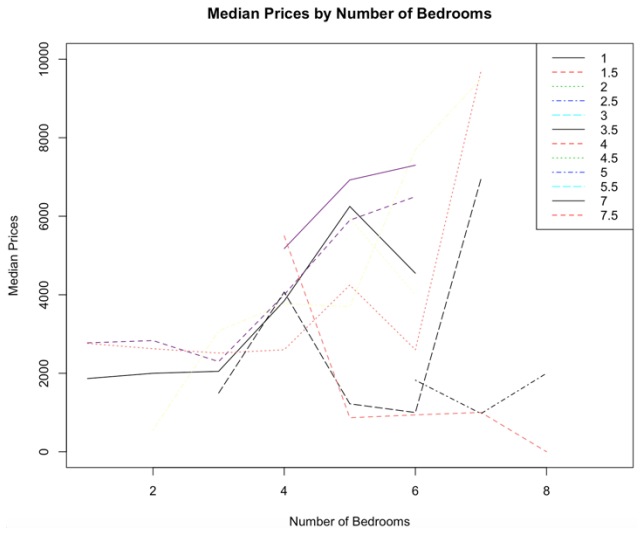


A potentially influential factor not included in the dataset is population in these locations. If there is high population, a low square footage could still result in a high price, and vice versa. In addition, average income per county can make it so that apartments of the same size have different prices in different counties.

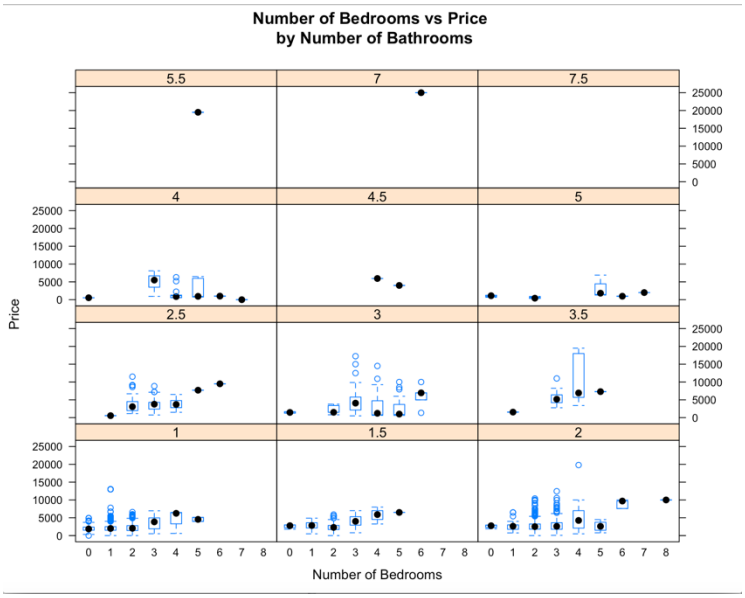
3. The amount by which rent increases for an additional bedroom depends on the total number of bedrooms. Moving from a studio apartment to a one bedroom, rent increases on average by approximately $123. From a one-bedroom to a two-bedroom apartment, rent increases on average by $385. From a two-bedroom to a three-bedroom apartment, rent increases by approx. $507. From a three-bedroom apartment to a four-bedroom apartment, rent increases by $598. It appears that the greater the number of bedrooms, the greater the increase in rent to add one more bedroom. The trend doesn't appear to be linear.



We know the amount by which price increases for an additional bedroom is different depending on the total number of bedrooms. For example, when moving from a studio to a one-bedroom, the rent increases by a smaller amount than when moving from a five-bedroom to a six-bedroom apartment. We can see this trend depicted below in relation to number of bathrooms as well. Part of the increase in rent includes an additional bathroom, but not always. The differing slopes for different bathroom categories proves that a bathroom is not added every single time a bedroom is added.

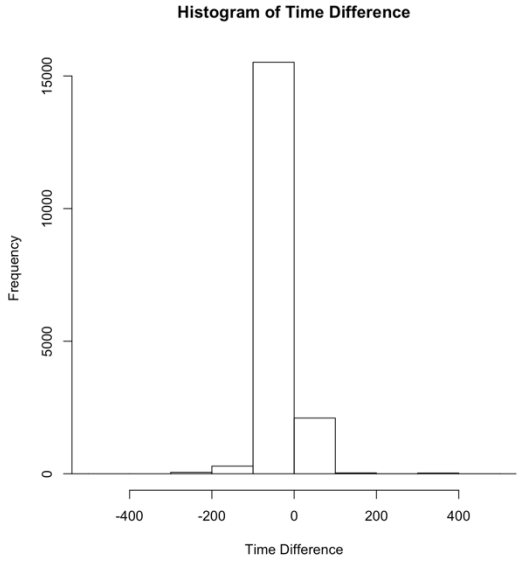


In the side by side histogram below, we can see that studios and one-bedroom apartments only ever have one bathroom, and apartments with 2-6 bedrooms enjoy an array of number of bathrooms. Five-bedroom apartments have the most "Normal"-mimicking distribution of all of the histograms. 2.5 bathrooms is a popular occurrence in apartments with 2-6 bedrooms.

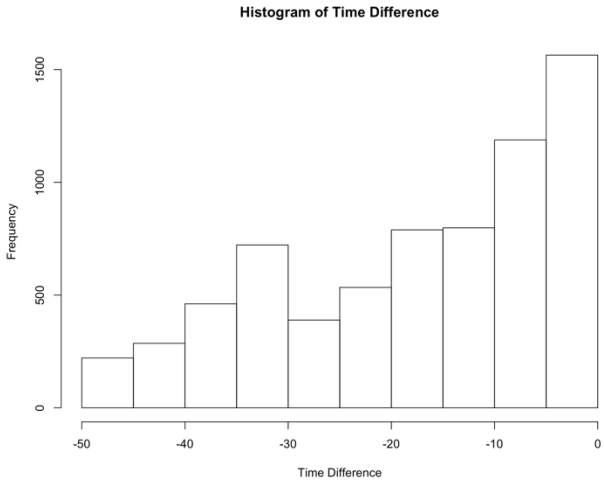


4. A post is typically made 8 days before the apartment becomes available. The mean is 8 days but the median is zero. The histogram of time differences shows that there are,

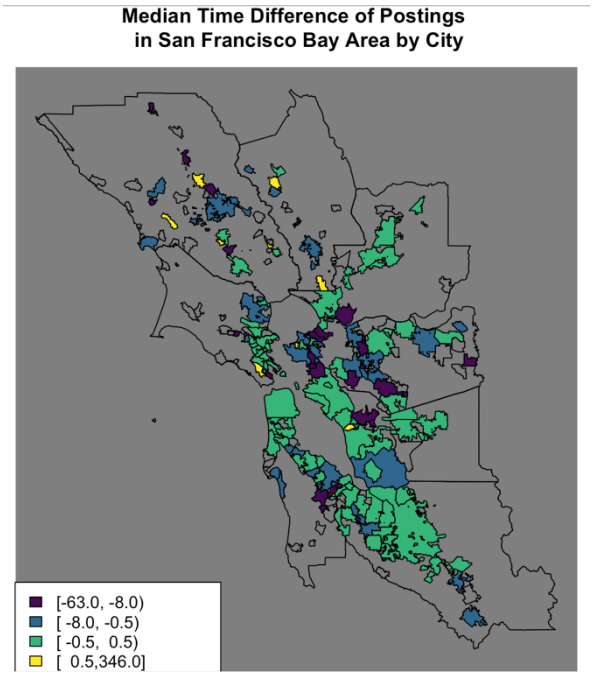
however, posts that are made after the apartment has become available. The overwhelming majority, though, are made before the apartment becomes available. The values are negative because date available was subtracted from date posted.



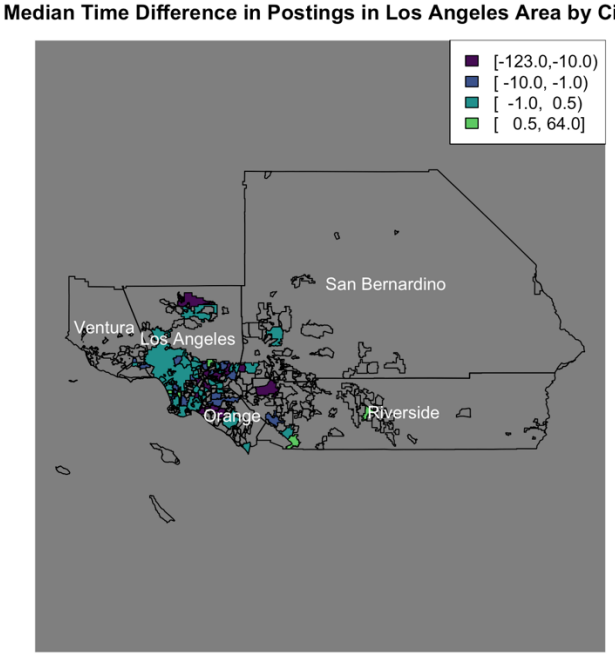
In the range of most common time difference (50 days prior to availability to immediate availability), there is a trend of smaller time difference being more common. As time difference increases, the frequency decreases. However, it is important to look at this by region as well.



The median time difference of posting in San Francisco Bay Area map below shows that a bit under a day in advance is the most popular posting time for the Bay Area. Up to 7 days in advance is the second most popular, and is scattered evenly through the Bay Area. There are a few posts made after the availability date in north Bay Area.

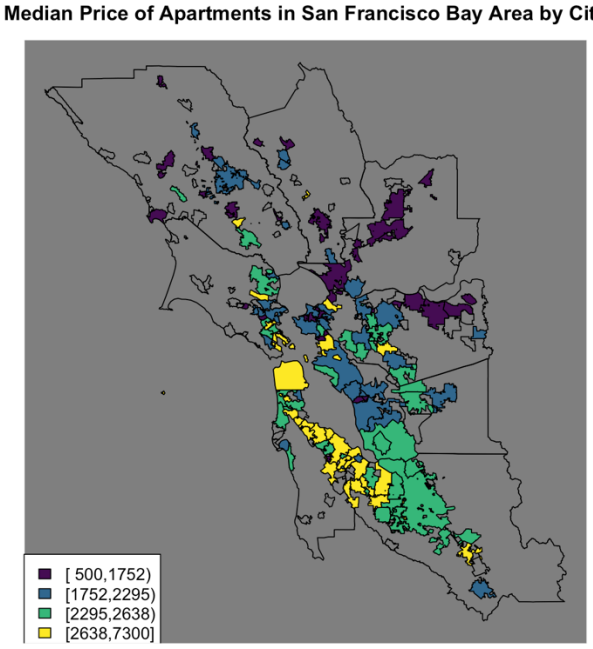


In the median time difference in postings in Los Angeles Area map below, we see that one day in advance to half a day after availability is the most common in the Los Angeles area. There are some postings in the other ranges of availability but they are not nearly as common.

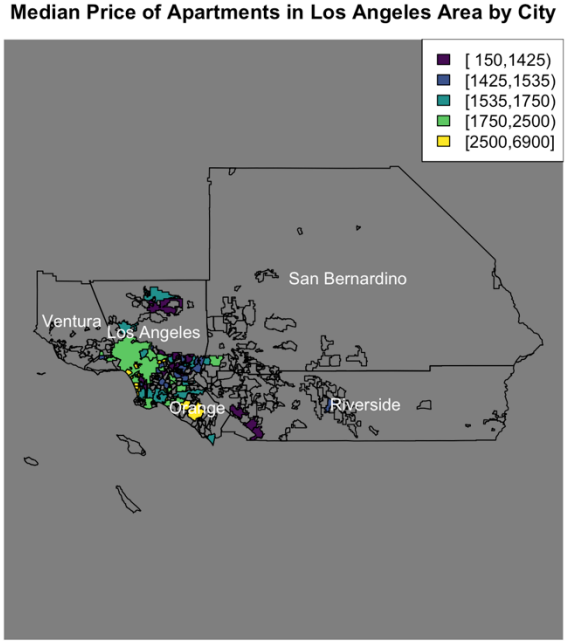


5. To answer this question, I looked at the San Francisco Bay Area and Greater Los Angeles Area. It appears that apartments in similar geographical areas do tend to have similar prices. Apartments in San Mateo tend to be in the same highest price range, apartments in Santa Clara tend be in the same second highest price range. Solano apartments are in the

lowest price range, and the middle and north Bay Area have a mix, but the colors group over multiple cities.



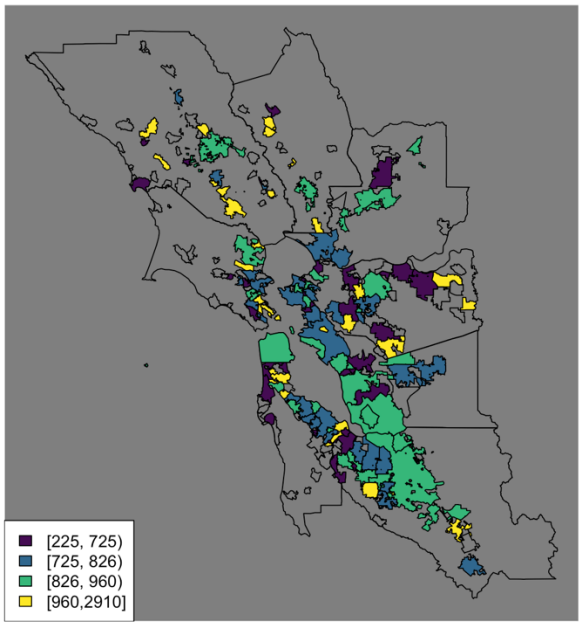
The same applies to the Los Angeles area, where Los Angeles apartments are mostly in the second highest price range. Within each county, certain price ranges group across cities. Clustering of the colors shows us that apartments in similar geographical areas tend to have similar prices.



6. The cities with the highest populations in California are Los Angeles, San Diego, San Jose, San Francisco, Fresno, Sacramento, Long Beach, Oakland, Bakersfield, Anaheim, Santa Ana, Riverside, and Stockton. By referencing outside population maps of

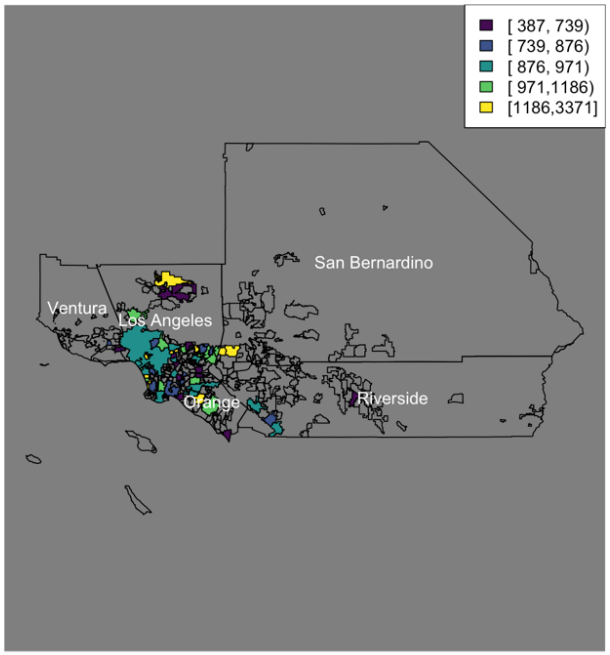
California, the peninsula of the SF Bay Area has high population, and its square footage range is somewhere in the middle compared to all of the SF Bay Area.

Median Square Footage of Apartments in San Francisco Bay Area by City

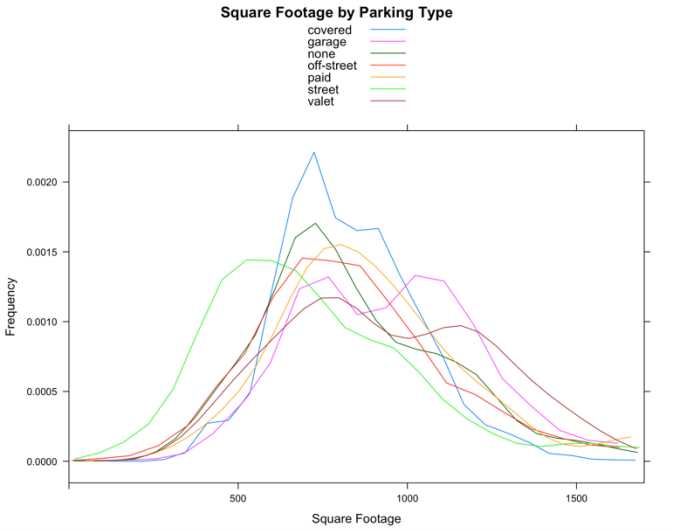


As for the LA area, the area depicted below has higher population than the SF peninsula and higher square footage. It seems that high population does not always correlate with low median square footage for apartments. This isn't too surprising since Los Angeles has a different city layout than San Francisco, which is more packed together.

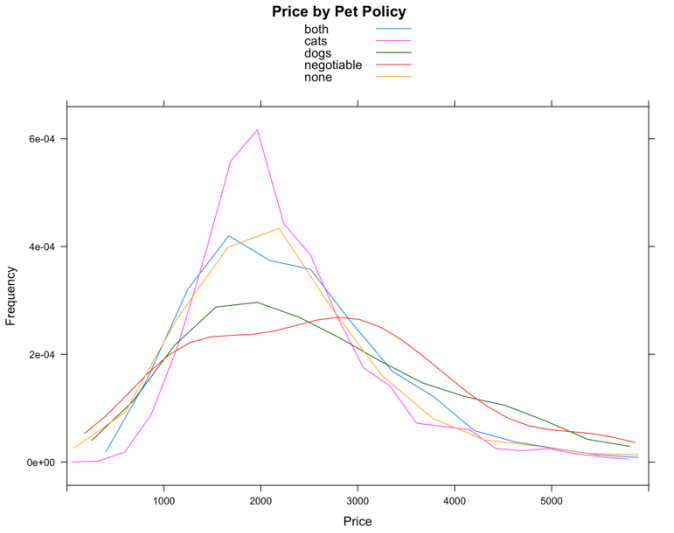
Median Square Footage of Apartments in Los Angeles Area by City



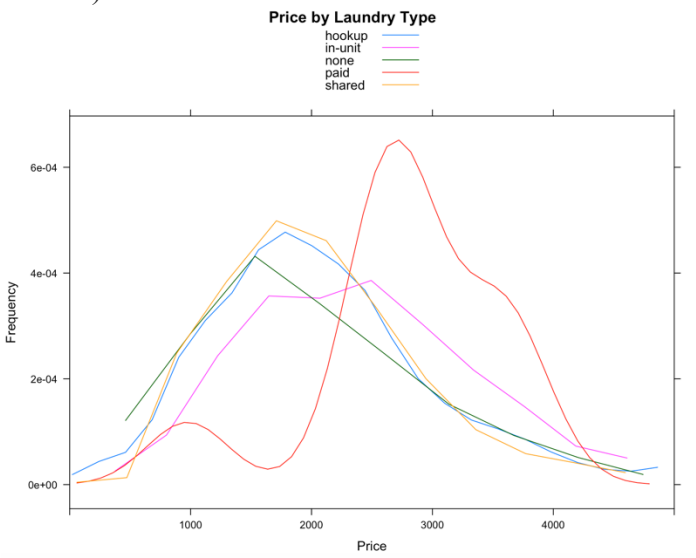
7. **Question 1: Do smaller apartments tend to have less reserved parking?** This question is particularly interesting to companies like Uber and Lyft. If smaller apartments are associated with unreserved parking ("none", "street") then areas with smaller apartments are the best to target for advertising for their services. This is because parking availability can be a deciding factor in car ownership. Since the "street" and "none" categories group mainly in apartments under 1000 square feet, it appears that this could be a good relationship to go off of for these companies. It is unclear, however, which of these categories overlap and are reserved or not.



Question 2: Are more expensive apartments more likely to allow pets? This question is interesting to people who are considering getting a pet. Not only are there the costs associated with pet ownership, but could pet ownership also result in having to find more expensive housing to accommodate for this decision? It appears that negotiable pet policy is nearly evenly spread out across apartments priced $1000 to $3500. Cats only policy is the most popular and peaks at $2000. Both cats and dogs policy and no pets allowed policy are equally popular and cover a similar range of prices. Because of this, it seems that price changes are not a huge factor to consider when deciding to get a pet.



Question 3: Do more expensive apartments have more favorable laundry accommodation? This question is useful for people looking to move into an apartment and would like to know the likelihoods of different laundry accommodations based on their price range. Paid laundry is most common in (expensive) apartments, costing $2500 to $4000. Hook-up and shared are equally likely and are grouped in lower price ranges. No laundry peaks at $1500 and steeply declines thereafter. In-unit is well spaced out in middle range prices. Therefore, it is clear that more expensive apartments have provided laundry (paid and in-unit).



Reference for Question 6: California population map
https://en.wikipedia.org/wiki/California#/media/File:California_population_map.png