Craigslist Apartment Data

Analysis and Discussion

By Jake Newman

Jake Newman
Mr. Ulle
STS98

California's Craigslist Apartment Rental Market Report

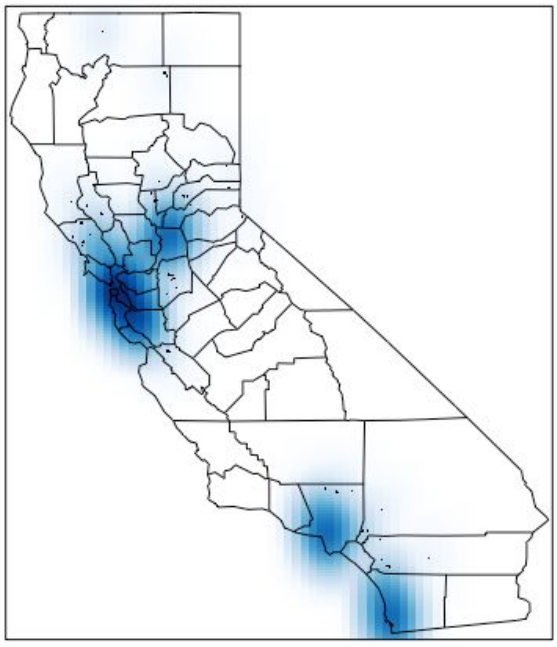
Craigslist is a free online database hosting local classifieds across a wide range of subjects. This service was founded in the San Francisco Bay Area but has since become an international platform, retaining its community moderated philosophy throughout. The dataset analyzed for this report focuses on apartment rental postings in California, encompassing just a small subset of the website's entire contents. Due to Craigslist's inclusive and open culture, there are few strict guidelines about how posts must be formatted, and the dataset certainly reflects this. Various included features are inaccurate for many observations, due to either inadvertent algorithmic malfunction or human error, both accidental and deliberate. In addition, there are a number of included posts that are entirely unrelated to the subject of apartment rentals, the primary topic of this report. Furthermore, a number of posts contain identical information, further distorting the original dataset. Such anomalies are discussed further in response to Question 1. Within the original dataset provided, there are 18,084 observations of apartment rental postings (N = 18,084). However, as previously mentioned, this figure includes a number of invalid, spam, or duplicate posts that do not contribute any meaningful information to this report. To varying degrees of success, each of such glaring anomalies were eliminated when possible.

1. Identify 3 anomalies in the data set. Each should be based on a different variable or set of variables. Why is each anomaly unusual? How it might affect your analysis of the data.

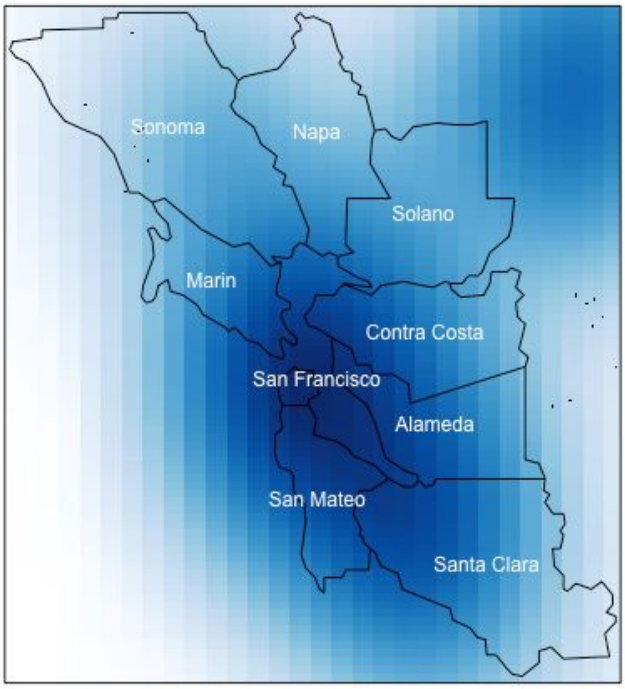
In terms of price, there exist a high number of posts falsely advertising large homes and other non-related services listed for nominal amounts of money, $1 or similar. One such posting describes a "historical mansion film location" in Beverly Hills available for a special event or photoshoot, listed at $1. These anomalies are easily recognizable as irrelevant to and invalid for this report's purpose of analyzing genuine Craigslist apartment rental opportunities. The described outliers mostly contain extremely small price values, distorting the data's spread and erroneously lowering measures of central tendency. When valid price information was contained elsewhere in the post, price values were corrected to reflect this. However, this solution was not always an option. In an effort to minimize remaining price distortion, rent information for posts seeking very low dollar amounts (less than $150) were invalidated, set to NA. This approach is effective in removing a high proportion of these spam postings from affecting price analysis. Nevertheless, the risk for removing valid data does exist as a potential drawback of this method, but due to the relatively small number of observations in this price range to begin with, omitting this specific group altogether does not dramatically impact the data's overall price distribution.

Plotting each observation's listed geographic coordinates on a map reveals how there are a significant amount of posts listed outside of physical land boundaries. According to their latitude and longitude positions, many apartment listings advertise a nonsensical location in open waters, such as the San Francisco Bay, shown in the maps below. For reference, the two maps below effectively show the same data, the second is simply a zoomed in plot of the SF Bay Area to further emphasize how common the described outliers are in this dataset. Clearly, a large number of latitude and longitude coordinates are not entirely accurate. As such, these variables cannot be relied upon for determining the intended geographic location of posts. The "shp_place" and "shp_county" variables seem to offer more reasonable insight into the intended geographic position for each observation and were used instead of latitude and longitude coordinates for most of the analysis contained in this report. This approach effectively reduces potential distortion caused by coordinate inaccuracies in the original dataset. However, the "shp_place" and "shp_county" variables do not offer the same level of detail and specificity as the latitude and longitude coordinates but are informative enough for meaningful analysis. Furthermore, since the "shp_place" and "shp_county" variables were generated based on the provided latitude and longitude coordinates, they are only accurate when the given coordinates are as well. However, for this report, the "shp_" variables provide enough coherent information about the intended geographic position of each post, smoothing out some of the irregularities present in their more detailed form.

Geographic Distribution of Craigslist Rental Postings



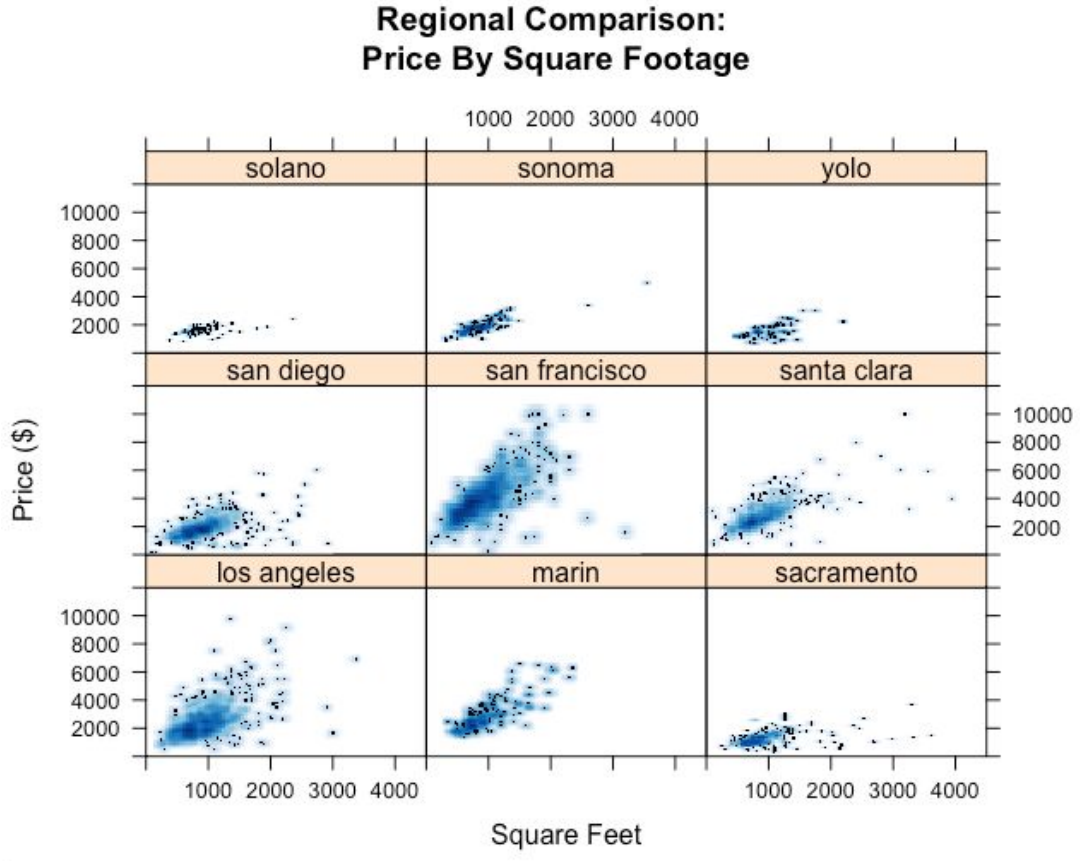
Geographic Distribution
(Zoomed In)



Looking at the written body text included in each post reveals how some of the observations contained in this dataset do not relate to apartment listings at all - instead advertising unrelated services. Examples include the provision of moving assistance and others. These are conceptually irrelevant to this report and would absolutely distort our analyzed data if not removed, likely invaliding our conclusions about apartment rental postings. As such, using a combination of the `grepl()` text search function and subsetting, any post containing a common variation of the word "moving" was labeled as spam and excluded from analysis. This approach is very effective in removing all posts containing the specific string sequences defined in this process. However, there are a number of synonymous ways to communicate the same irrelevant post, some of which may remain undetected. As such, the total absence of these spam posts can not be guaranteed. Nevertheless, a high proportion of these anomalies do get removed in this manner. Another drawback to this approach is that since the spam marking process relies on text string matching, there is a very real possibility that a valid post becomes marked as spam just for containing some variation of the word "moving". However, due to the relatively small number of posts containing this string term, erroneously removing a few valid points should not have a drastic impact on underlying trends.

2. Is there a relationship between apartment size and price? Be careful to account for the effects of other variables such as geographical area. Discuss whether variables not present in the data set could make it difficult to see a relationship.

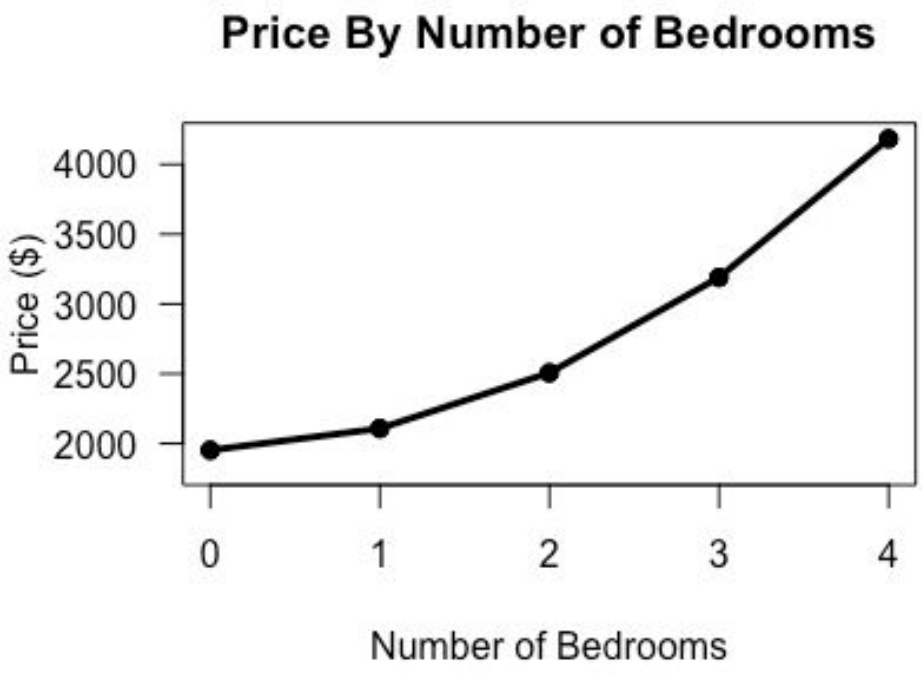
Price generally exhibits a direct correlation with apartment size in square feet, regardless of geographic area. In other terms, rental prices typically increase as square footage increases. Nevertheless, outliers do exist, and the extent of this relationship varies dramatically between different cities. For example, shown in the plots below, San Francisco demonstrates a highly pronounced and directly correlated relationship between price and size in square feet, the steepest of those included below. Furthermore, SF also has the highest occurrence of expensive outliers for smaller units. In stark contrast, Sacramento prices increase much less dramatically when square footage increases. Sacramento also seems to have more large houses at relatively low price points than SF. Variables not present in the dataset certainly may obstruct a full understanding of certain aspects of this relationship. For example, the information about the overall condition of the unit, or how many previous tenants have lived there is not included. These confounding variables would almost certainly affect the rent price. Another piece of useful information could be whether the advertised unit has a view of the ocean, usually considered a desirable feature. Such information could possibly help describe why prices in coastal cities like San Francisco seem to increase at a faster rate than inland regions like Sacramento.



3. Approximately how much does rent increase for an additional bedroom? Does this amount depend on the total number of bedrooms? Does this amount include the cost of an additional bathroom? How can you tell?

Posts advertising more than four bedrooms or bathrooms were excluded from the following analysis due to their having insufficient observations, many of which are likely to be invalid. On a separate note, listings with partial bathrooms were rounded up to the next integer value using the `ceiling()` function for conceptual and visual simplicity when discussing the number of bathrooms present.

Of analyzed posts, the average price increase for an additional bedroom was $557. Shown in the graph below, this amount varies significantly depending on and in direct correlation to the total number of bedrooms. For example, the price difference between a 3 and 4 bedroom unit is $993, more than six times greater than that of a studio and one bedroom unit, a difference of just $155. Clearly, the difference in rent for additional bedrooms does depend on the total number of bedrooms and accelerates at a non-linear rate.

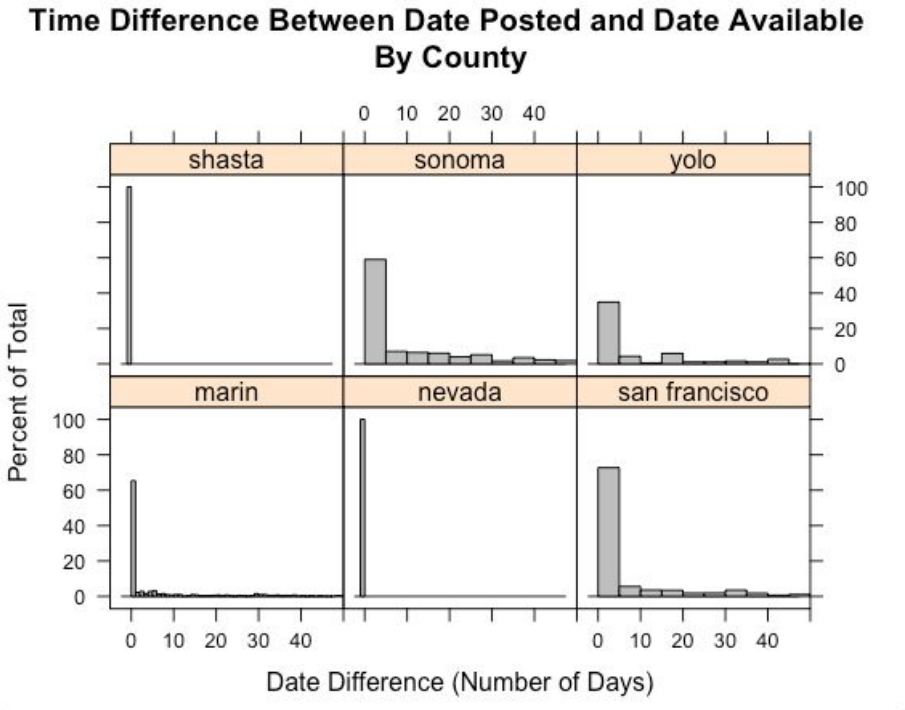


The previously described cost does not include that of a new bathroom. A distribution of price by number of bedrooms is shown in the next figure below, while also accounting for number of bathrooms. The graphed distribution reveals how the average price increases for an additional bedroom does **not** include the cost of an additional bathroom, since the number of high-priced outliers dramatically increases with each additional bathroom, and by more than the average price jump of $557 for additional bedrooms. In sum, additional bathrooms are more expensive than additional bedrooms.

4. How far in advance of an apartment becoming available for rent is a post typically made? Explore the distribution of this time span and discuss whether it varies by region.

In the following analysis, the posting date of each listing was subtracted from the date on which the unit became available, using the `as.Date()` function. Posts returning a negative time difference indicate that the unit became available before being advertised on Craigslist. In other words, posts with a negative time difference were available from the day of its advertisement on Craigslist. As such, for conceptual and graphic simplicity, posts that returned a negative value were reassigned a time difference of zero days. Using the `quantiles()` function reveals that 75% of all posts were created less than two weeks before the unit became available. More interesting, perhaps, is that at least 50% of all observed posts were created when the unit was already available. This suggests that suppliers within the Craigslist rental market are more focused on filling existing vacant units than long term planning.

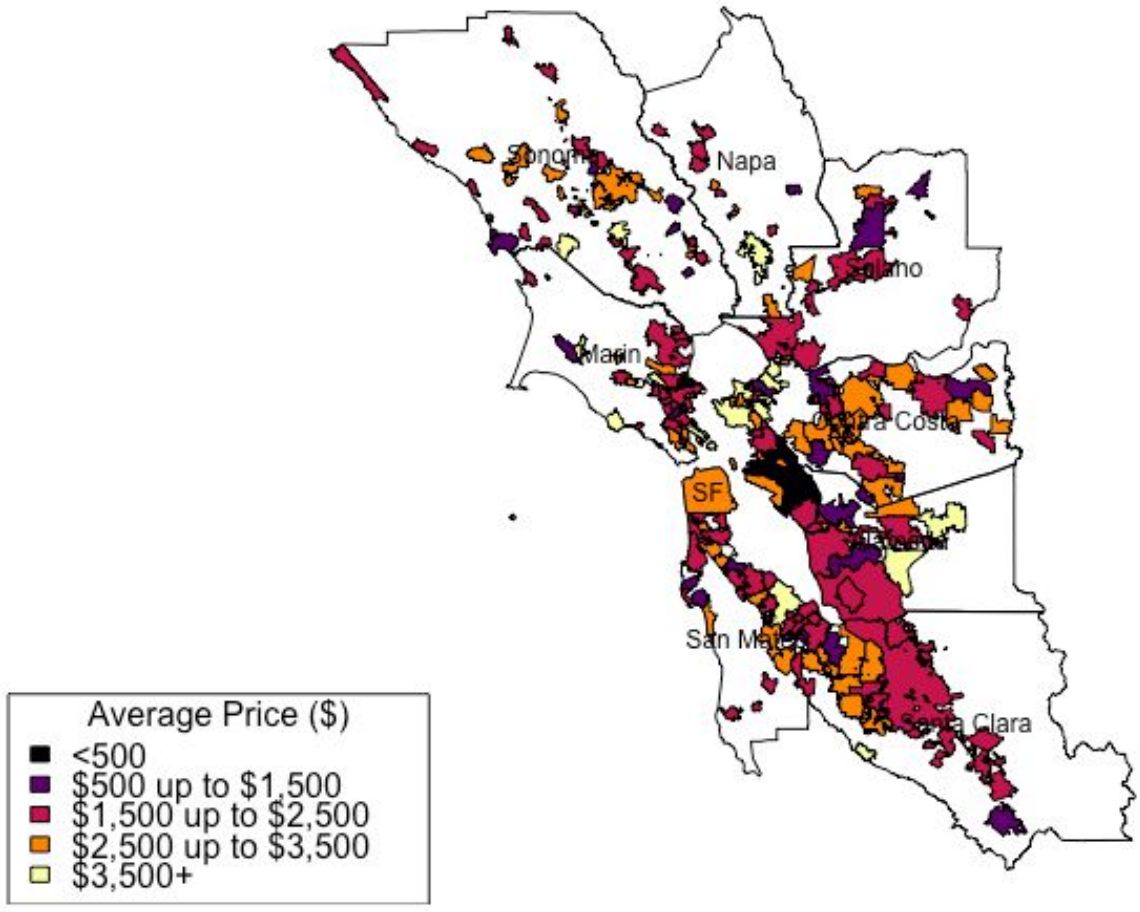
The distribution of how far in advance posts were made before the unit became available for rent does vary significantly depending on geographic region. Shown in the side-by-side histograms below, regions like Shasta and Nevada have a very narrow spread, almost entirely concentrated at very low time differences. Other regions like San Francisco, for example, show a relatively wider, more diverse spread of time differences but remain concentrated towards lower values.



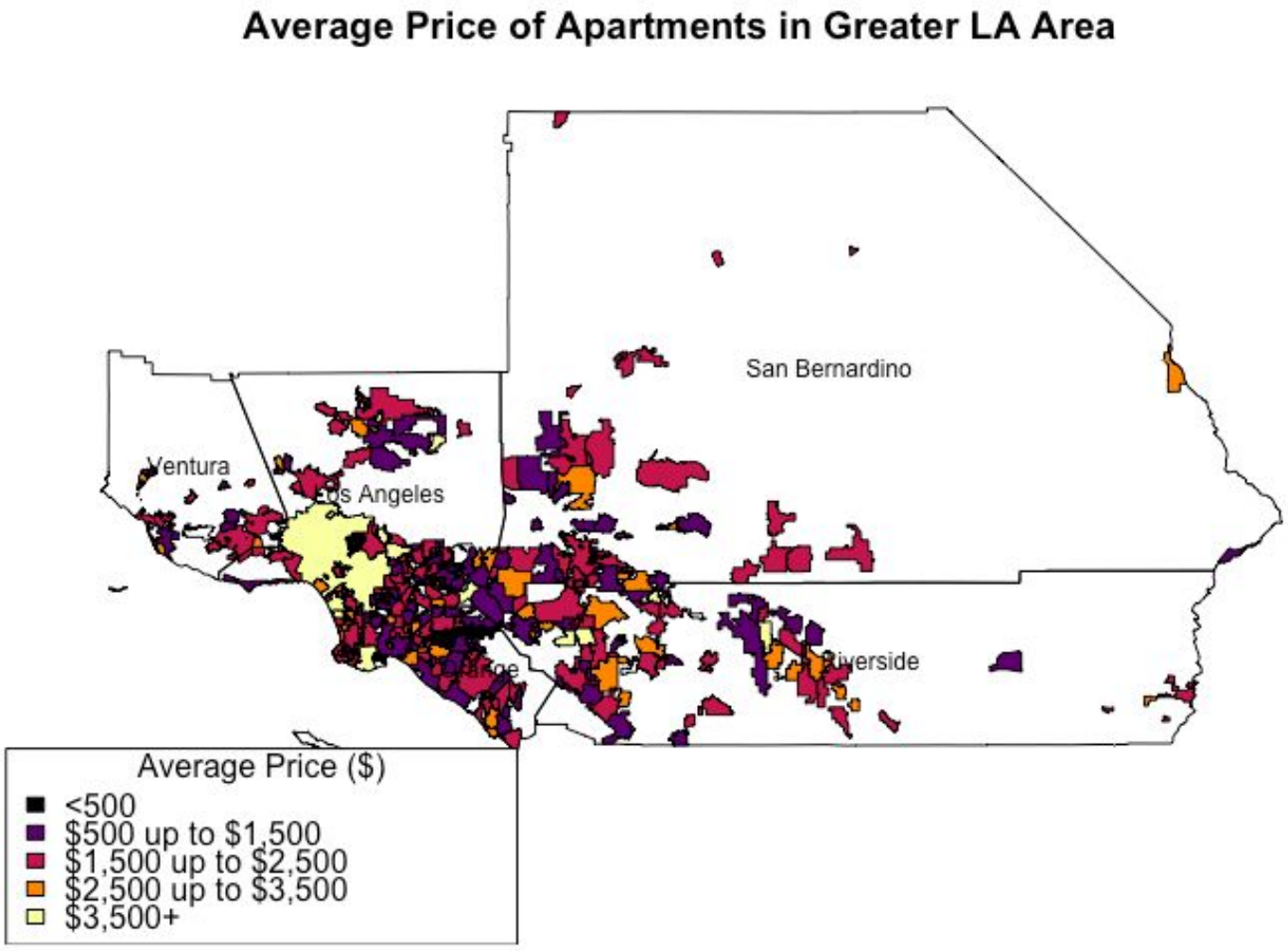
5. Do apartments in similar geographical areas tend to have similar prices? If so, are there any exceptions? If not, do you see any other patterns? Use maps to justify your answer.

Shown in the maps below, apartments in similar geographical areas do tend to have similar prices. The following analysis focuses primarily on the San Francisco Bay Area and Greater Los Angeles region, selected for their density of posts. Plotting the average prices of apartments by city / named rural area using the "shp_place" variable effectively reveals geographic price trends. As predicted, San Francisco has among the highest average rent in the SF Bay Area region. However, certain slivers of San Mateo and Santa Clara, in the heart of Silicon Valley, are even more expensive than SF itself. Places within Marin, Sonoma, and Napa are similarly high priced - not a surprising result. Oakland is a glaring exception. Situated in midst of a generally expensive area, Oakland has the lowest average rent around, perhaps as a result of its historically high crime rates, presence of nearby industrial operations, and other undesirable factors. According to FBI statistics from 2013, Oakland's violent crime rate is more than twice that of SF, a possible explanation for the extreme price differential.

Average Price of Apartments in San Francisco Bay Area

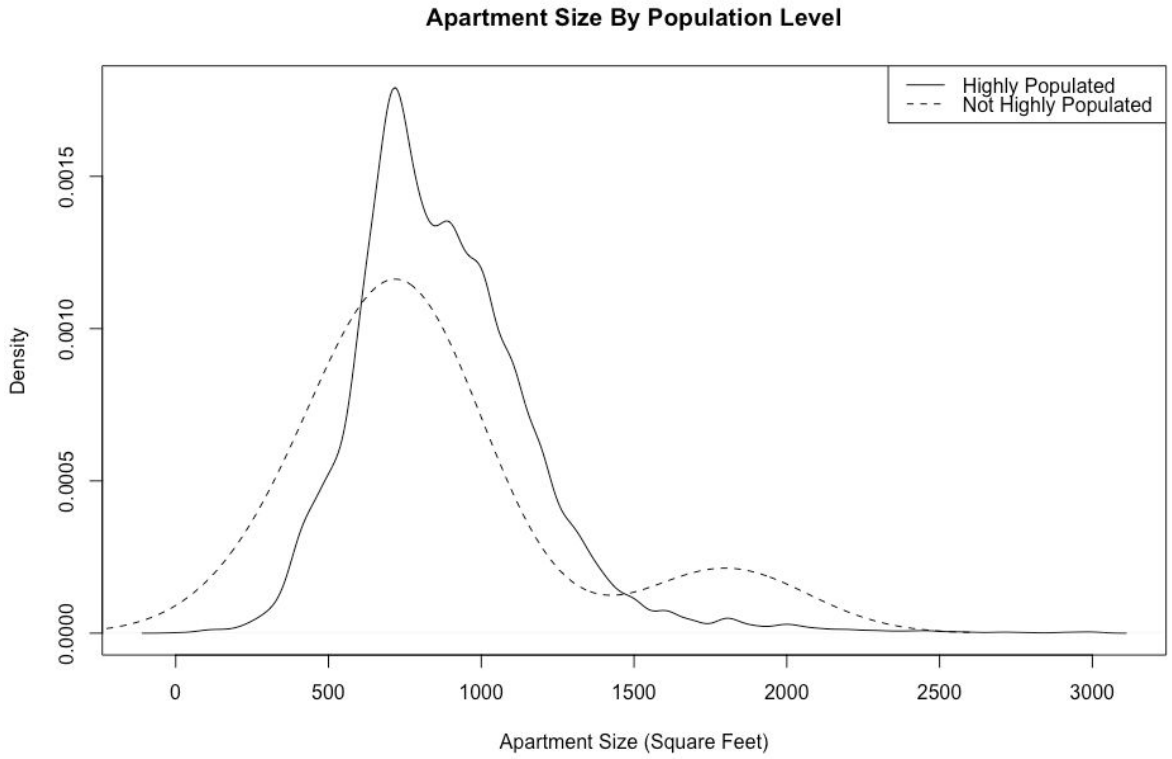


The same trend of nearby regions having similar prices is seen in the Greater Los Angeles region. A notable observation of the Greater LA area is that low priced units are far more common than in the SF Bay Area. Nevertheless, outliers and exceptions do exist, most apparent in the extremely high priced Southeast region of Los Angeles County. This specific area is notoriously expensive, luxurious, and exclusive in terms of real estate - including Beverly Hills, Hollywood, and Santa Monica. As such, it is hardly a surprise that this region is a glaring exception to the price trends of the surrounding area, much of which is very low-priced in comparison.



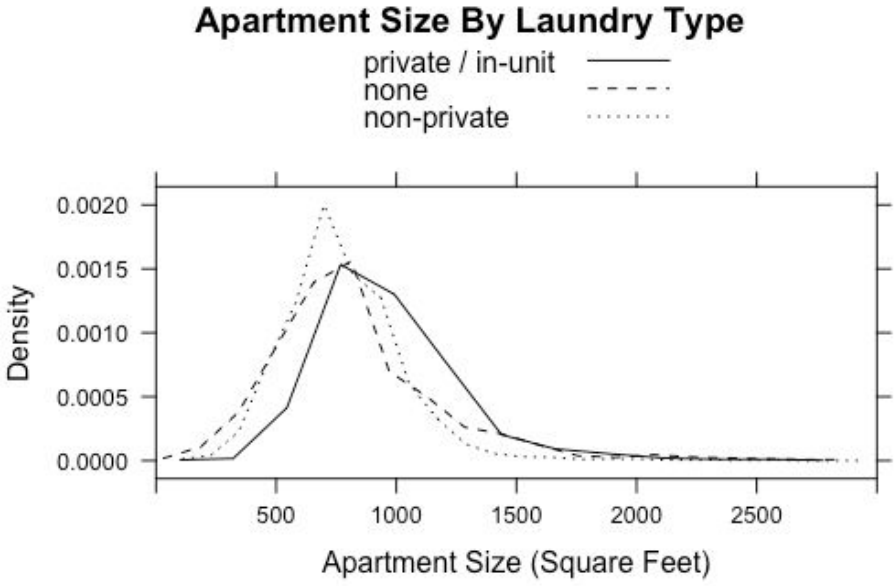
6. Do highly-populated areas tend to have smaller apartments?

Highly populated areas do tend to have smaller apartments. For this analysis, regional categories were defined in accordance with official statistics from the US Census Bureau and the California Department of Finance. Highly populated counties are defined as (in no specific order) Los Angeles, San Diego, Santa Clara, San Francisco, Fresno, and Sacramento. Low population counties are defined as (in no specific order) Del Norte, Siskiyou, Modoc, Humboldt, Trinity, Shasta, Lassen, Plumas, Butte, and Glenn. Shown in the density plot below, nearly all apartments in highly populated counties are smaller than 1,500 square feet in size. Furthermore, in highly populated counties, apartments smaller than 1,000 square feet in size are far more common than those larger than 1,000 square feet. Similarly, in less populated counties, the majority of apartments are less than 1,500 square feet. However, in contrast, there are more units larger than 1,500 square feet in less populated counties than in highly populated counties, shown in the graph below by the dotted line rising above the solid line at x = 1,500 through x = 2,500.



7. Are larger units more likely to have a certain kind of laundry setup? If so, do smaller units share the same setup?

Laundry setup type does seem to be related to an apartment's size in square feet, shown in the grouped density plot below. The original dataset provides 5 unique classifications of laundry types: "[appliance] hookup", "in-unit", "paid", "shared", and "none". For visual simplicity, these were reformatted into 3 categories: private / in-unit (comprised of hookup and in-unit setups), non-private (comprised of shared / paid setups), and no laundry setup available. One drawback of this approach, however, is that certain details about whether the actual laundry appliances are provided is lost. Nevertheless, more than enough information remains to draw meaningful conclusions. Units less than 400 square feet are most likely to have no laundry setup at all. Units between 500 and 800 square feet are most likely to have a non-private laundry setup, likely situated outside of the immediate unit itself. This finding seems very reasonable, since many small apartments simply do not have enough space or the necessary built-in structures to house other types of laundry setups. In contrast, apartments larger than 1,000 square feet are most likely to have a private laundry setup. Generally, large units are more likely to have private laundry setups available, while small units are more likely to have a non-private setup. However, it is important to remember that these estimates are based on somewhat unreliable input data, since the original "laundry" variable is algorithmically defined by a computer based on each post's included text description. Nevertheless, these findings could be interesting and of use to businesses offering non-conventional laundry services in helping them to identify target consumer demographics, primarily those without easy access to an in-unit laundry setup at home.



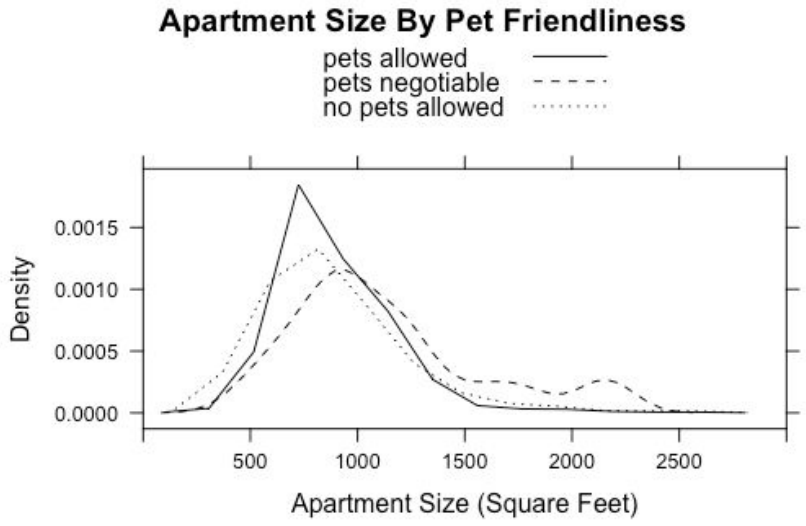
8. Is the availability of parking related to the price of an apartment?

Originally, the dataset provided 7 unique variants of parking, including “covered”, “garage”, “none”, “off-street”, “paid”, “street”, and “valet”. For this analysis, the included parking information was reformatted for conceptual and graphic simplicity. Using the `levels()` function, included parking information was streamlined into three categories: free parking, paid parking, and no parking. This process emphasizes the general relationship between rent prices and whether or not parking is provided. However, a drawback to this approach is that some detail is lost in regard to the specific parking type provided. Nevertheless, sufficient information remains to draw a conclusion, that rent price is related to parking availability.

Shown in the density plot below, inexpensive apartments with a rent price of less than $1,000 were most likely to not have a dedicated parking space of any kind. In contrast, apartments with a rent price between $1,500 and $2,500 were most likely to include free parking, whether it be in a covered space, a garage, or a dedicated off-street space. Surprisingly, the most expensive units, with a rent price of more than $3,000, are most likely to have some form of paid parking available, including valet service, instead of free parking. In summary, the availability of parking is clearly related to the rental price of an apartment. Low-priced units typically do not include parking, while more expensive units typically do include some form of parking. However, it is important to remember that these estimates are based on somewhat unreliable input data, since the original “parking” variable is algorithmically defined by a computer. Nevertheless, these results are interesting because they could easily prove helpful for novice renters in finding the best housing type to suit their budget, especially if they have a car. A ride-sharing service provider like Uber could also find these results useful in identifying likely customers, particularly those living in low priced units with no access to a dedicated parking space at home.

9. Is the size of a listed unit related to its pet-friendliness, as defined by the landlord's willingness to allow pets in the apartment?

The size of an apartment does seem to be related to its pet-friendliness. Originally, the dataset provided 5 unique levels for the landlord's tolerance of pets, specified by "dogs", "cats", "both", "negotiable", and "none". In the following analysis, the included pet tolerance information was reformatted for conceptual and graphic simplicity. Again, using the `levels()` function, included pet tolerance information was streamlined into three categories: pets allowed, pets negotiable, and no pets allowed. This specific procedure helps to reveal and emphasize general pet friendliness trends by apartment size. However, a potential drawback to this approach is that it eliminates some detail about the landlord's preference for dogs or cats. Nevertheless, enough information remains to draw a meaningful conclusion about whether apartment size is related to if pets are allowed in the unit, regardless of the specific animal type. Shown in the density plot below, small units under 500 square feet in size are most likely to not allow pets of any kind. In contrast, larger units between 500 and 1,000 square feet in size are most likely to allow pets of some kind. Units larger than 1,000 square feet in size are most likely to be "negotiable" towards pets, meaning that further discussion with the landlord would be required for pet approval. However, it is important to remember that these findings are based on somewhat unreliable input data, since the original "pets" variable is algorithmically defined by a computer. In the end, these results are still interesting because they suggest that landlords of bigger units tend to be more selective than landlords of units ranging from 550 to 1,000 square feet. Pet owners or similar would, obviously, also find this information useful during apartment hunting, informing them to stay away from small units, as it would not be likely for them to find a suitable apartment.



Sources Cited:

-FBI Crime Statistics:
https://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2013/crime-in-the-u.s.-2013/tables/table-8/table-8-state-cuts/table_8_offenses_known_to_law_enforcement_california_by_city_2013.xls//

-Bay Area Crime Infographic (Based on FBI Data):
https://infogr.am/bay-area-crime-rates-2013

-Craigslist: Frequently Asked Questions:
https://www.craigslist.org/about/help/faq

- US Census Bureau: 2010 Census Data

- California Department of Finance: "Census 2010: Table 3A — Total Population: 2010”