Sta 141a

Rohan Malhotra
Statistics Department
University of California, Davis
October 2018

As discussed in the previous assignment, Craigslist is a website, in which one can easily post advertisements on what product they are selling. The last assignment consisted with the analysis of the Craigslist data set, which consisted of apartment rental listings in California. This assignment deals, with looking at the spatial characteristics of the rental market and looking at the demographic features. The demographic features were obtained from the US Census Bureau, which provides a lot of data sets such as spatial data sets as well. The data set used, is demographic information on cities of California from 2010.

1 Explanatory Data Analysis

1.1 Merging the data

Since the Craigslist data and Census data are two different data sets, they had to be merged into one whole data set. Before merging the data, the Craigslist data was cleaned in the same way as it was in assignment 3, for further information on how the Craigslist data was cleaned refer to section 5. The data was merged by matching the cities in the Census data set to the cities listed in the Craigslist data set. To completely do this keywords such as “cdp”, “city”, and “California” had to be removed from the description of the cities in the Census data set so it can perfectly match up with the cities in the Craigslist data set.

1.2 Rows and Columns

The merged data set consists of 12,644 rows (apartment listings) and 394 columns (features). Some of the features of the data set include advertisement there are 20 variables/ pieces of information that was given. These columns have information on the apartment (from the craigslist data) such as: latitude, longitude, place, city, bedrooms, bathrooms, pet policy, garage information, laundry information, parking information, when the post was posted and updated, and the sqft. Additionally, these columns include some features of demographics (from Census data) such as: information on race, age, gender, housing tenure, and housing by type.

1.3 Missing features

Looking at the merged data, there are some expected missing features such as pet policies, garage options and parking options which make sense as not every apartment may have mentioned those. Next, looking at the demographics part of the data set, it is too large to examine every feature, but looking through the data set some common errors I did notice were that for some cities there was missing data, which was denoted by an (X).

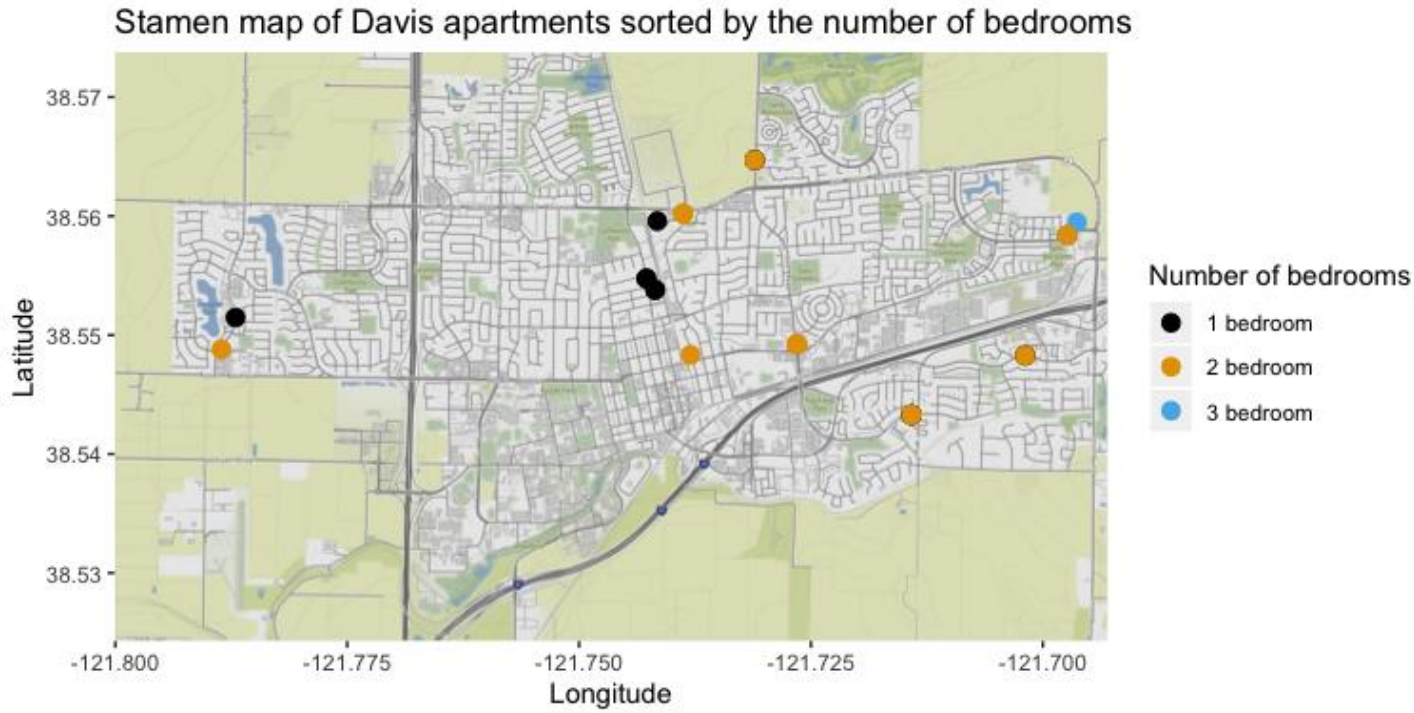
2 Spatial characteristics of the city of Davis, CA

2.1 Setting up

Using GGmaps in R, I could get a better look at the features of Davis. To do this, I first sub-setted the data to only the apartment listings happening in Davis. From this I got a total of 39 apartment listings. The one limitation that I will encounter, is the low amount of sample size. Due to a small sample size my interpretations, may not be completely accurate. Though there are, 39 observations the maps below will not have 39 points because some apartments have the same latitude and longitude coordinates as the others.

2.2 Number of bedrooms in apartments in Davis

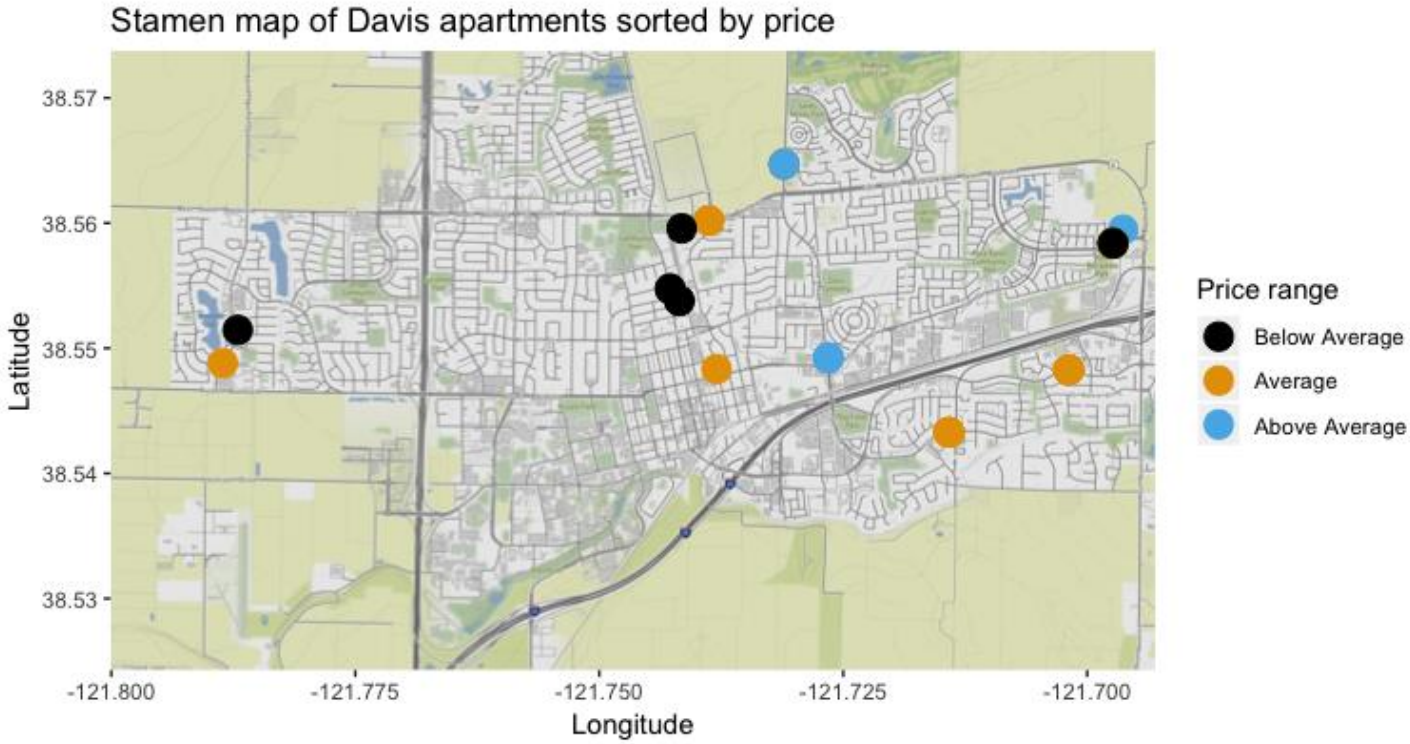
The first, feature that was analyzed was bedrooms in apartments in Davis. I was interested in examining if there is a relationship among the number of bedrooms in apartments and areas of Davis. This can be examined to through the graph below



It is seen from the figure above that one bedroom apartments in Davis tend to be in the Western and Central parts of Davis. 2 bedroom apartments are the most common and are mostly in the Southern and Central parts of Davis. Finally, 3 bedroom apartments are the least commons and are typically in Eastern Davis.

2.3 Variation of apartment prices throughout Davis

The next feature that was examined was price, to see if there are any patterns between the price of apartments and the area of Davis. To do this I first looked at the five-number summary of the prices of the apartments in Davis. In order, to make the graph easier to understand, I divided up the price into three categories: Below average (defined as anything lower than the first quartile, < $1,407), above average (defined as anything above the third quartile, > $1,839) and average (defined as anything between the 1st and 3rd quartiles, $1,407 < x < $1,839). The variation of prices among areas of Davis can be seen on the graph on the next page.



From the figure, above, it is seen that Davis has many apartments that are below average, which are primarily found in Central Davis. Next, as expected there are many houses in the average price range, and are dispersed throughout Davis, as seen above they can be found in Western, Central, and Southern Davis. Finally, above average apartments in Davis are found in the Northern, Central, and Eastern parts of Davis.

2.4 Laundry options in apartments throughout Davis

Finally, the last feature that was looked at was the laundry options apartments offer. The goal is to see if there is some pattern or relationship among certain laundry options in apartments throughout different areas of Davis. This relationship can be seen through the figure below.

From the map above it is seen, apartments in Davis, tend to offer a lot of shared laundry options which is spread out throughout Davis. This makes sense because since Davis is a college town, many apartments have shared options throughout all areas of Davis. On the other hand, in unit laundry is often seen in the central and southern parts of Davis, further away from the campus. Finally, hook up laundry is seen in the very eastern part in Davis, and do not seem to be common.

2.5 Conclusion

In conclusion, for every feature there is a similar pattern seen when, as the apartment moves further and further away from the UC Davis campus. For, bedrooms as apartments listings move away from the campus the number of bedrooms in apartments tend to increase and the highest value of three bedrooms occurs. Similarly, with price, the price increases and the apartment listing with above average price is the furthest away from campus. Finally, with laundry options, the hook-up laundry is only seen in the apartment listing furthest away from campus. I conclude, that the higher standard apartments in Davis tend to be furthest away from campus. However, as said before my conclusion is not completely accurate as there are not enough sample points to support it.

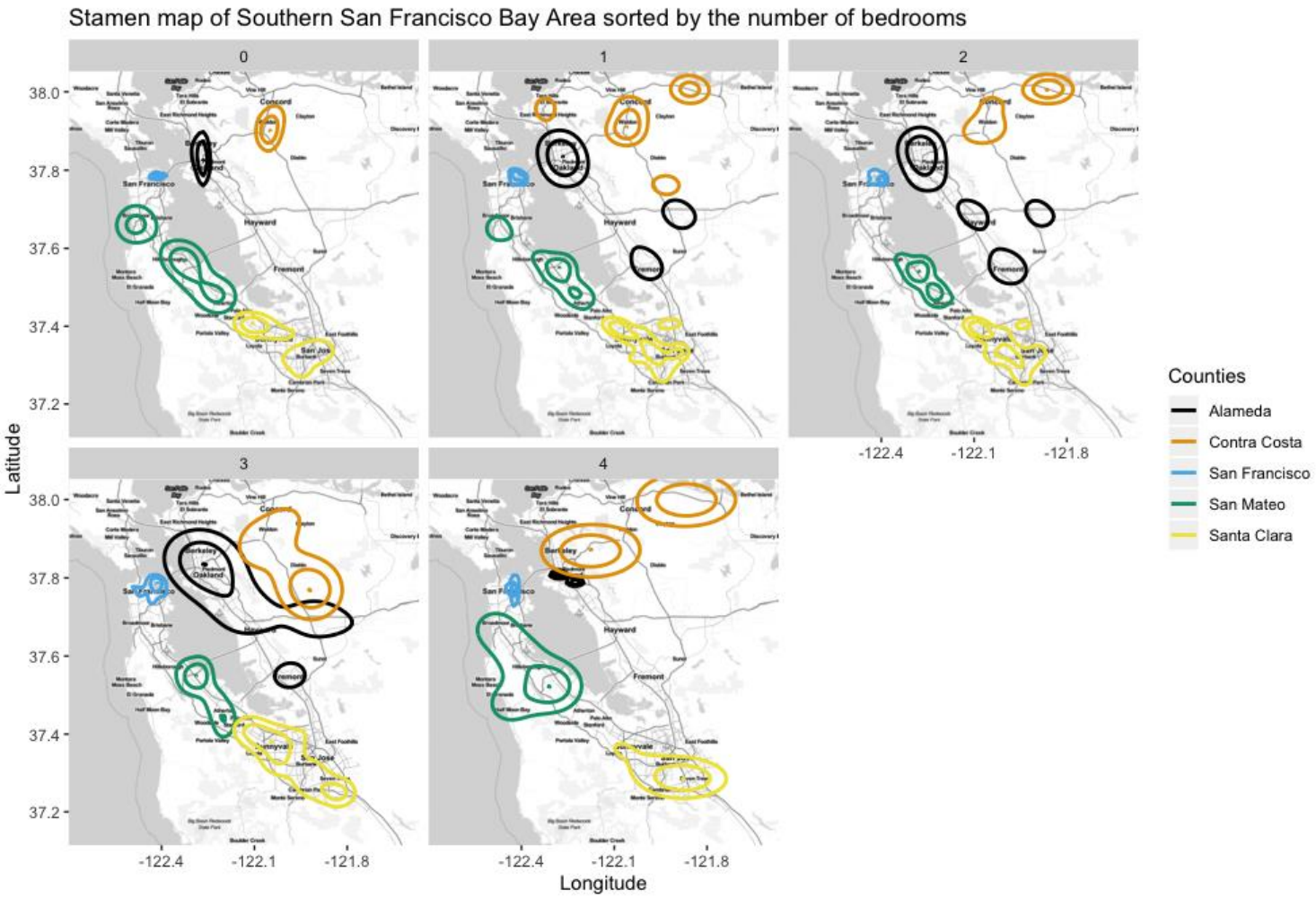
3 Spatial characteristics of the Southern San Francisco Bay Area

3.1 Setting up

After, looking at Davis, I wanted to explore a bigger city so I looked at the Southern San Francisco Bay Area. To subset the data I only looked at apartment listings with the counties of San Francisco, San Mateo, Santa Clara, Alameda, and Contra Costa. For all these counties, there are a total of 6,277 apartment listings. To create the map, I used -122.7 degrees' longitude and 37.1137 degrees' latitude to -121.5685 longitude and 38.052 degrees' latitude.

3.2 Number of bedrooms in apartments throughout the Southern SF Bay Area

As with Davis, with San Francisco I wanted to examine the relationship among the number of bedrooms in apartments and certain areas of San Francisco. I only looked at bedrooms 1-4 because there were to little of sample of sample points for apartments with 5 bedrooms. This is shown on the next page.



The figure above shows the distribution of apartments with 1-4 bedrooms, throughout the different counties of Southern Bay Area. From this is seen that, apartments with one bedroom are most common in the San Mateo and Santa Clara counties. Two bedroom apartments are almost equally dispersed throughout Southern Bay Area, but it seems that Santa Clara and Contra Costa tend to have a little more two bedroom apartments than the other counties. Three bedroom apartments, are clearly seen mainly in Alameda, Contra Costa and Santa Clara. Finally, four bedroom apartments are mainly in Contra Costa and San Mateo.

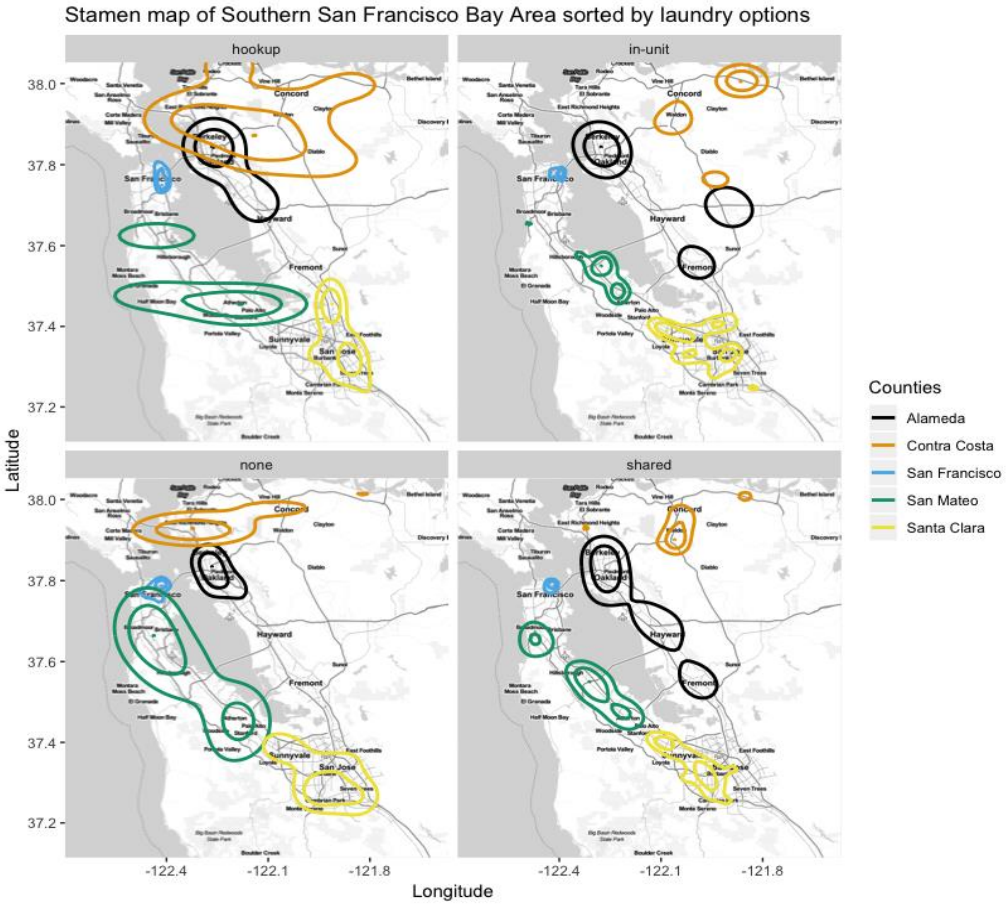
3.3 Variation of apartment prices throughout Southern San Francisco Bay Area

As many know San Francisco is now an expensive place to live in California. Therefore, I wanted to see if there is an pattern among price and areas of Southern San Francisco. To do this the price was divided into the same three categories as seen with Davis in 2.3, using the five-number summary. Below average is considered below the first quartile (< $2,300), above average is anything above the third quartile, (> $3,495) and average was considered anything between the first and third quartile.

From the figure, above it can be seen, the cheapest places in the Southern Bay Area are mainly Alameda and San Mateo. Additionally, most apartments in every county have an average price, but some counties such as San Mateo and Santa Clara offer many apartments with an above average price.

3.4 Laundry options in apartments throughout the Southern Bay Area

After personally, visiting San Francisco a couple of times I noticed that apartments tend to look small and jointed together for the most part. For that reason, I wanted to explore the laundry options throughout the Southern part of the Bay Area and see if there was any relationship with certain counties. Below is a figure that can help demonstrate this.



Looking at the figure on the left, for the hook-up laundry option, it is in all counties, but Contra Costa tend to have more apartments with that option. In-unit laundry is primarily seen in both the San Mateo and Santa Clara counties. Finally, apartments in the San Mateo tend to offer no laundry at all more than any other laundry options. Finally, it seems apartments with shared laundry are almost equally disbursed throughout all the counties.

3.5 **Conclusion**

In Conclusion, it is seen that the more luxury apartments in Southern San Francisco can be found in mainly the Santa Clara and San Mateo county. When compared to every other county, they seemed to have more apartments with four bedrooms, more apartments with an above average price and more apartments with in – unit laundry option. I can be sure that my conclusions from looking at the Southern Bay Area are more accurate than my conclusions after looking at the same features in Davis, because there were more sample points for the Southern Bay Area. On the other hand, the Bay Area data consisted of too many data points, so I used the density plots to avoid the over plotting and make the graphs easier to interpret by the readers.

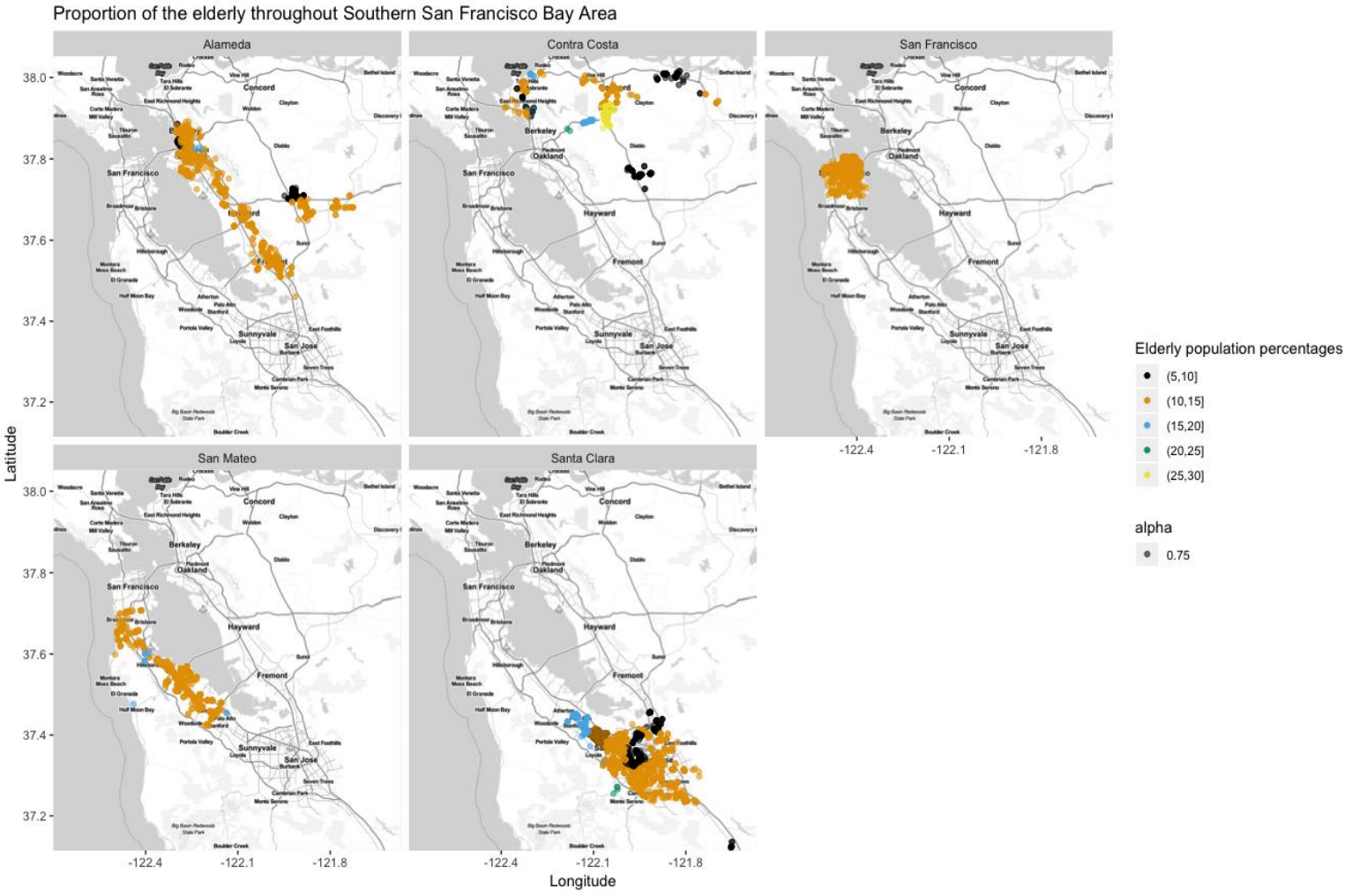
4. Elderly population in the Southern San Francisco Bay

4.1 Setting up

To look at this, as done in section 3, the data was sub setted to only include the Alameda, Contra Costa, San Francisco, San Mateo and Santa Clara counties. Next, to get the most equal representation, I looked at the proportion of people with an age of 65 years or higher.

4.2 Elderly population

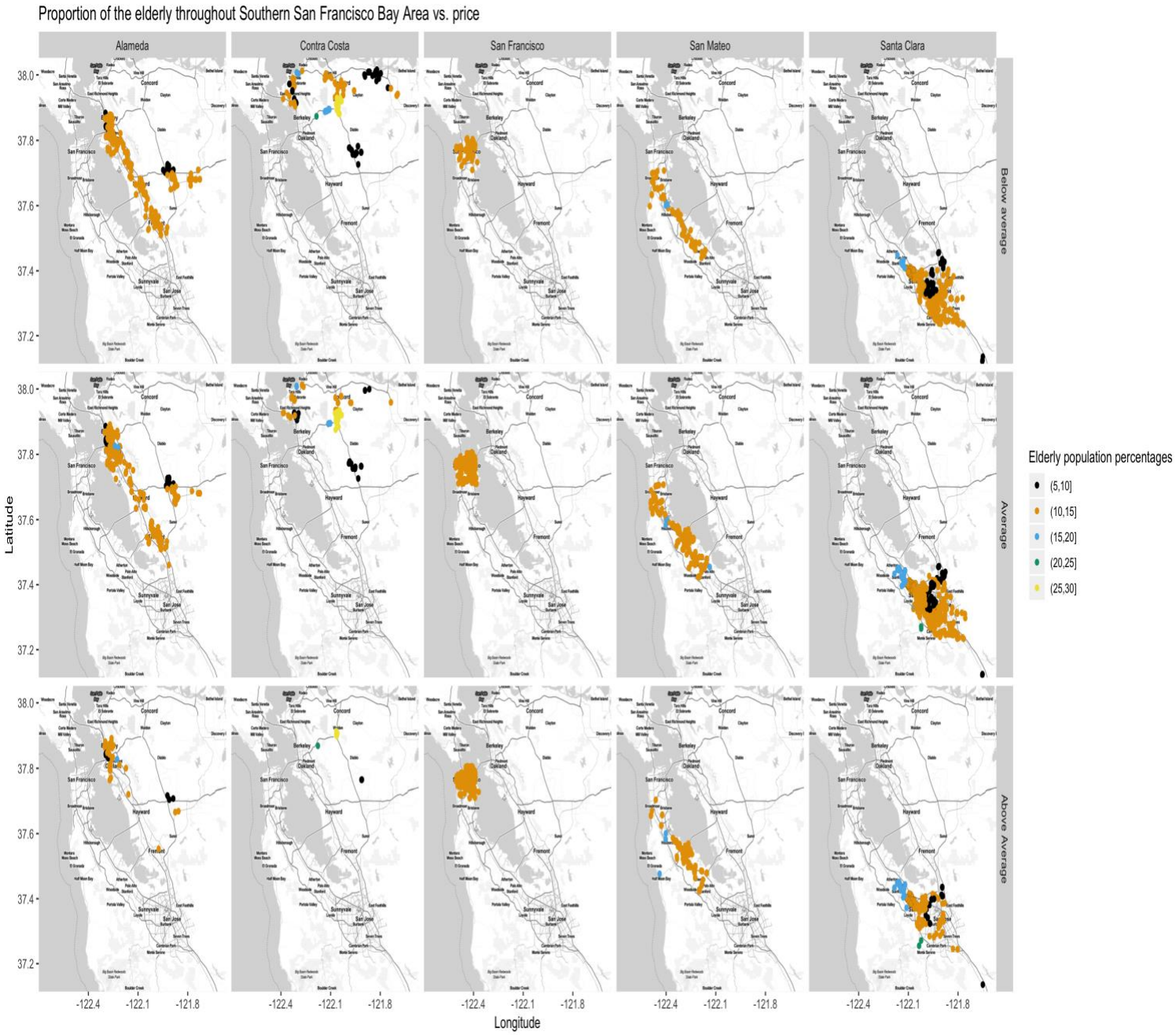
The graph below shows the proportion of elderly population, based on the city in which the apartment is listed in.



Looking at the proportions, it seems all counties mostly have an elderly population of 10 to 15 percent. Furthermore, Contra Costa has the most cities with a relatively high elderly population, as they have the highest number of cities with a 25 to 30 percent elderly population. On the other hand, Santa Clara seems to have the most cities with a relatively low elderly population, as they have the have the highest number of cities with an elderly population of 5 to 10 percent.

4.3 Relationship among elderly population and rental market

I divided the price into three categories, and these categories were below average, above average, and average. Those price ranges are defined the same way they are in section 3.3 so please refer to that. Below is a figure that shows all the counties through the price ranges.



From the figure, it can be seen that San Francisco tries to help the elderly, as there are many apartments priced below average in cities with mainly a 10 to 15 percent elderly population. Next, there are not many apartments that were priced above average, but Santa Clara had more apartments with an above average price than any other county, due to it having the most cities with the lowest elderly population. With the average price apartments, they seem to have primarily between a 10 to 15 percent elderly population, except for apartment listings in the Contra Costa county.

4.4 Conclusion

In conclusion, I would say that Southern San Francisco Bay Area is very friendly towards the elderly. All the counties have a majority of 10 to 15 percent elderly population throughout their cities. With price, it is seen that, primarily price for apartment rentals are average or below average, but of course there are some apartment rentals above average. I would say out of all those counties, the best place for the elderly would be the Contra Costa county as it has more cities with a higher percentage of elderly than the other counties. Also it has the most apartment listings with a below average price and the least apartment listings with an above average price. Finally, Santa Clara is the least elderly friendly county. This is because it has the most apartment listings with an above average price, but also because it has the most cities with a very low percentage of elderly people.

5 Cleaning of the Craigslist data set

The raw data consisted of 21948 posts and 20 variables. The data had a lot of limitations that I could resolve, but some limitations I could not resolve. Below will show what I did to clean the data and resolve as many limitations as I could.

5.1 Missing Values

First, I looked at the missing values in the data set to see if I noticed any patterns. Some patterns I noticed when I explored the missing data is that all the posts that were missing bathrooms also were missing the bedrooms, meaning there were the same number of missing features for bedrooms and bathrooms. This same relationship applies to the latitude and longitude variables also the state and county variables. The only two features that were not missing were the deleted and craigslist variables.

5.2 Outliers

Next, I looked at the quantitative variables to see if there were any outliers. Looking at bedrooms, When I looked at the posts with 5 or more bedrooms, they were not apartments, but rather houses so I had to remove them. Next, looking at bathrooms, there were some apartments with 0 bathrooms, looking at those posts, no posts mentioned there were 0 bathrooms, but rather it did not have anything about bathrooms in the text so I deleted all observations with 0 bathrooms. Then there were some apartments listed under 80 sqft and above 190000 sqft. I deemed these observations as outliers and deleted them. Finally, looking at price I noticed many posts with prices below 200. After looking at those posts, it turned out those posts were either the wrong rate and were charging for apartments weekly, the posts did not relate to apartment rentals, or were services for apartments.

5.3 Wrong Information

I also looked at the quantitative data, to remove some of the falsely reported data. I noticed that some had states reported other than California, but this is data only for California so I removed them. I also noticed some texts that were not apartments, but those were removed when I edited the price variable as described above. Finally, there were some duplicate posts and text which I also deleted.

5.4 Unavoidable limitations

The data set did have some limitations which I could not resolve. One is that some posts were houses and not apartment. In the study, I counted townhouses as apartments, but there were still some listings that were for houses which I could not avoid, but it was too minimal to affect the data. Next, is the number of observations. In my analysis, I looked at many questions which involved looking at specific cities. Due to the massive difference in observations among cities I feel like my answers cannot be concrete as there are not enough sample points.