Man Pan
Student ID: 914656278
Email: manpan@ucdavis.edu

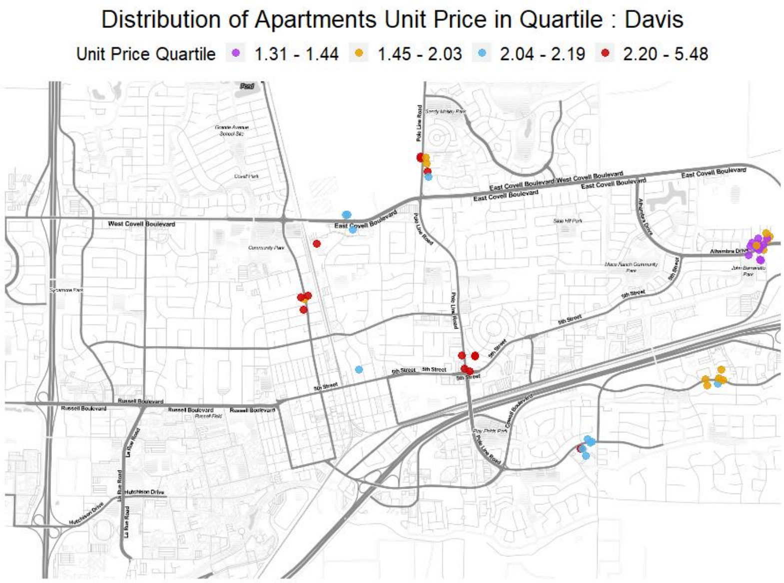
1. Description

The report has three parts. First one is apartments accommodation in Davis; second one is apartments accommodation in Southern Bay Area; last one is to analyze rental market in cities have the oldest population in Southern Bay Area.
In order to analyze those questions, I should deal with errors and outliers in the data set firstly. I find some errors in price: two posts' price are 30000000 and 9900000, which are wrong (cannot match the price in 'text' column). In the 'text', the prices are 3408, 995 separately. Also, from the box plot, I think some outliers would cause some negative influences to my conclusion. Thus, I limit the price from 200 to 17700, the size from 100 to 5000. Then, I select apartments in Davis(64 posts) and apartment in Southern Bay Area (10904 posts) separately.

2. Davis Apartments Accommodation

2.1 Relation of Apartments Price and location

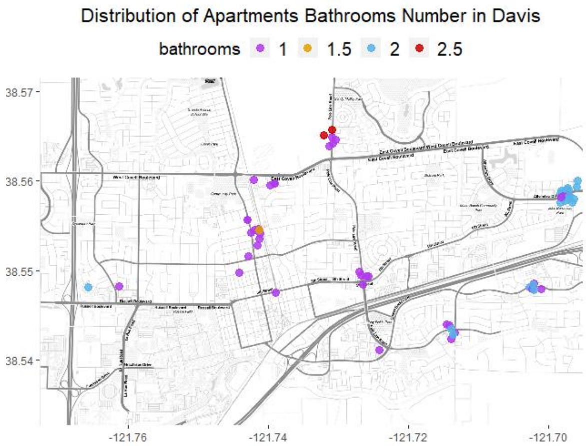
From the data set, I find apartments rent range from 721 to 2700 in Davis. But each accommodation will effect the rent, I think I need to take the size of apartments into account. I create a new variable named unit price (price/size) as the standard to find high rent apartments. Cause the unit price is a numeric variable, I prefer to split them into groups by quartile to help me analyze. By quartile, I split apartments into four groups : 1.31 - 1.44 unit price, 1.45 - 2.03 unit price, 2.04 - 2.19 unit price, 2.20 - 5.48 unit price.

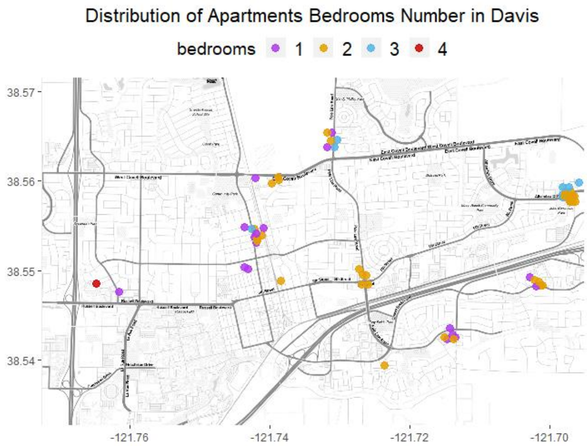


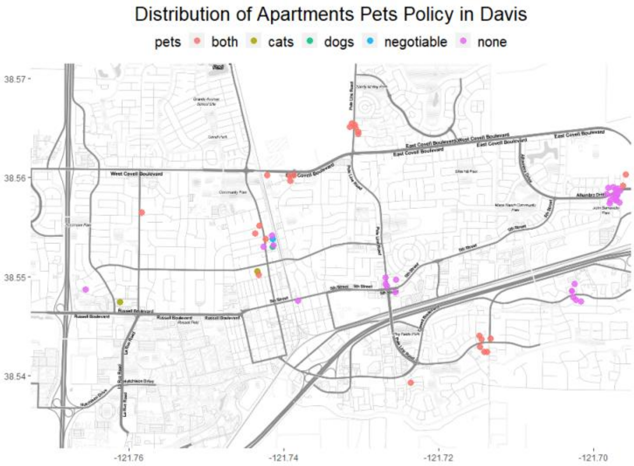
As the picture shows above: red color means high rent apartments, while purple means low rent apartments. From the map, we know the bottom left location is nearby campus. And I find the red points are near to campus, purple points are concentrated in one location that are far away from campus. For low rent apartments, they are all nearby Alhambra Drive that are away from campus and downtown. For relative low rent apartments , they are concentrate on two places that are relative near to downtown compared to low rent apartments. For high and relative high rent apartments, they are nearby campus and in the middle of downtown, some of them are near to Safeway, which are so convenient for people life. Thus, we can conclude that high rent apartments are more likely to near to Campus, downtown and market.

2.2 Apartments are more likely to family-friendly when they are far away from downtown.

I am interested in the family friendly apartments. I define family friendly as 2 or more bedrooms, 2 or more bathrooms, hookup or in-unit laundry, covered and garage parking and pets allowed or negotiable apartments.



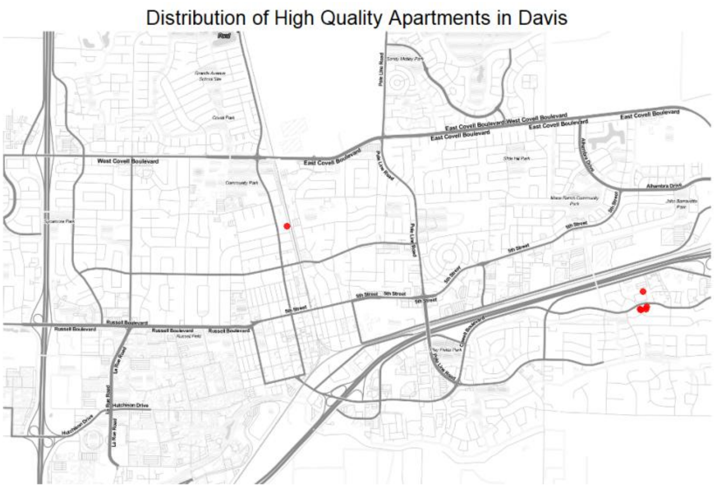




From the pictures above: We can find that when we focus on the location that are far away from the downtown, apartments are more likely to have 2 bathrooms , 2 or 3 bedrooms, hookup or shared laundry, covered/ off-street/garage parking, none/both pets. Except the pets policy, I can find some rules: the number of bedrooms and bathrooms will increase when apartments are far away from downtown. And apartments would provide laundry (shared or hookup) and parking (covered or garage). Although there are some apartments are off-street policy, I search more about the parking policy in that area, I find those apartments are near to main route. That's why they are not allow street parking. But they have enough place to park near to their apartments. Thus, we can conclude that apartments are more likely to family-friendly when they are far away from downtown.

2.3 Location and Unit Price of High-Quality Apartments

I am interested in high-quality apartments. Firstly, I should define the definition of high quality apartments : en-suit, allow parking, have laundry. For en-suit apartments, I select apartments that the number of bathrooms are equal or larger than that of bedrooms. Then, I select covered, street, garage and paid parking, in-unit, hookup and shared laundry. In Davis, there are just 6 high quality apartments among 64 Davis apartments.



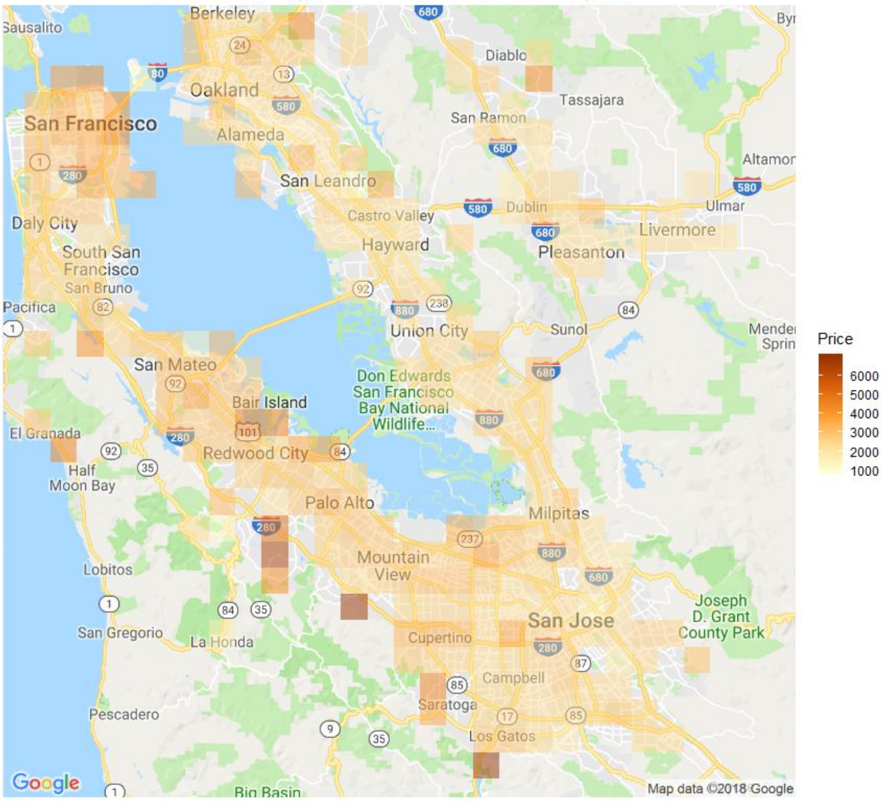
As the picture shows above: the six apartments are located in two place, and just one near to campus and downtown. Five apartment are located in the Edge Apartments. Then I find the six apartments' unit price are 5.48, 2.06,1.82,2.03,1.82,1.82. The highest unit price apartment is near to downtown. The rest five apartments' unit price are near to the median of all Davis rent, indicating the rent are not too high. Thus, I can find the Edge Apartments are suitable for students. Because apartments there are have high quality but relative low rent.

3. Southern Bay Area Apartments Accommodation

3.1 Relation of Apartments Price and location

Similar to the question in 2.1, I want to explore the relationship of Apartments price and location. But in the area of Southern Bay Area, there are so many different cities which have different rent market. I get southern bay area data set from 6 counties: San Francisco, Alameda, Contra Costa, Santa Clara, San Mateo. They contain 67 cities totally. I find that the range of price in southern bay area is much wider than that of Davis. That's because some big cities like San Francisco, San Jose have high rent which will enlarge the highest rent in southern bay area. The opposite is also true.

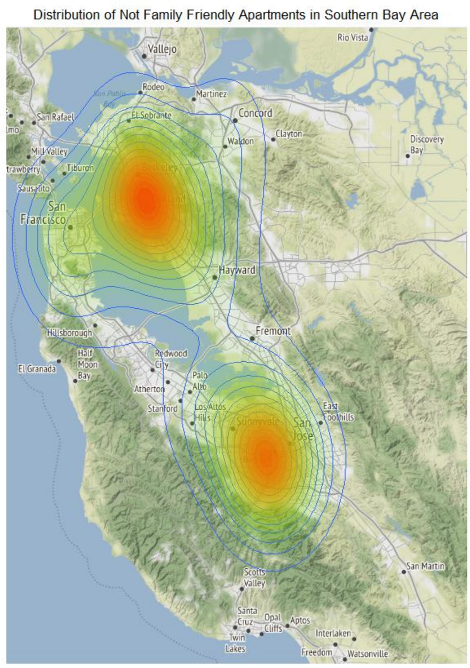
Distribution of Apartments Price in Southern Bay Area

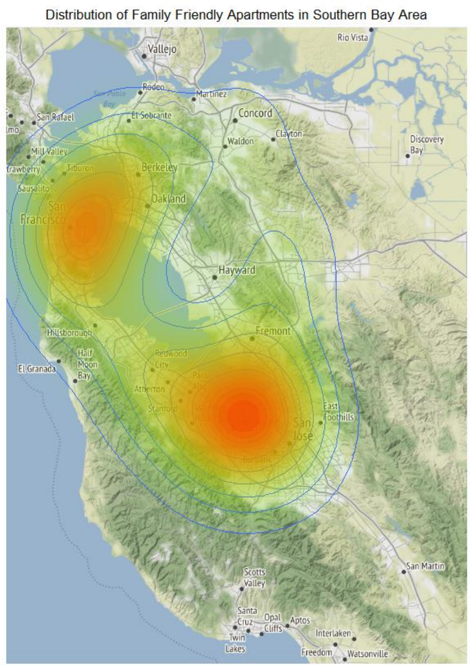


As the heat map shows above: I find high price apartments (darker color) are concentrated around San Francisco, Berkeley, Oakland, Diablo, San Leandro, San Mateo, Bair Island, EL Granada, Mountain Vies, Saratoga , San Jose and Los Gatos. Those high rent apartments are located in big cities which have larger population and many big companies or near to famous campus and seaside. Thus we can make a similar conclusion that the rent price will increase when they are near to campus or prosperous commerce and numerous residents area.

3.2 Are apartments more likely to family-friendly when they are far away from Major Cities ?

Considering the large number of posts in southern bay area, I choose 2D density plot to clarify the location of different apartments. Considering different cities, I change my definition of Family Friendly Apartment a little. I define family friendly as 3 or more bedrooms, 2.5 or more bathrooms, hookup or in-unit laundry, covered, paid and garage parking and pets allowed or negotiable apartments. Then I define apartments are not family friendly as: 0 -2 bedrooms, 0-2 bathrooms, none pets, none and shared laundry, street, off-street or none parking. Also, I limit the rent in the range of 1th quartile to 3rd quartile, because we all want to live in a family friendly and low rent house. I use density 2D plot to see the location of family friendly and not family friendly apartments.





As the picture above shows: for not family friendly apartments, they are mainly located in Oakland/Berkeley and Burbank, I know Oakland is the largest city and the county seat of Alameda County. Following the 1906 San Francisco earthquake, many San Francisco citizens moved to Oakland, enlarging the city's population, increasing its housing stock and improving its infrastructure. And Burbank only a few miles northeast of Hollywood, numerous media and entertainment companies are headquartered or have significant production facilities. Thus, I think large population and Prosperous commence lead to more not family friendly apartments. For family friendly apartments, they are mainly located in San Francisco and Sunnyvale, I know

Sunnyvale is a city located in Santa Clara County, its population is not on the top of US, but it near to San Jose, near to the high tech area Silicon Valley. Thus, it is a good place to live because it has a relative lower rent but a relative convenient and comfortable life. Although San Francisco also has several family friendly apartments, I think that is because we have a large number of samples belongs to San Francisco. The number of each kind of apartments maybe larger than that of other regions.
Thus, I think apartments are more likely to be family friendly when they are far away from the major cities. Finally, I think I can recommend families to live around Sunnyvale.

3.3 Location and Unit Price of High-Quality Apartments in Southern Bay Area

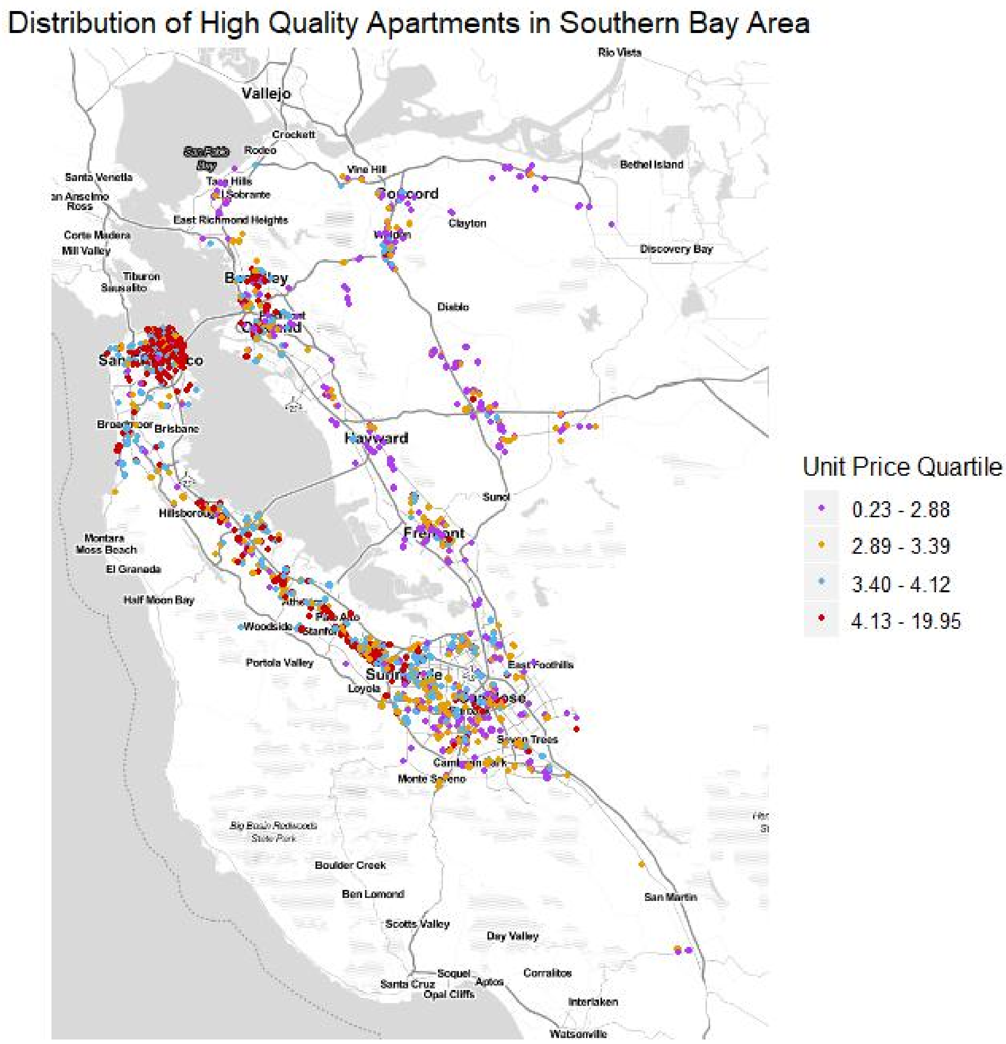
Similar to the definition of high quality apartments in Davis, I use the same criteria to select high quality apartments in Southern Bay Area. For en-suit apartments, I select apartments that the number of bathrooms are equal or larger than that of bedrooms. Then, I select covered, street, garage and paid parking, in-unit, hookup and shared laundry. In Southern Bay Area, there are 4283 high quality apartments among 10904 apartments. Almost half of apartments are high quality, so living environment is relatively good. Specially, I am also interested in the percentage of high quality apartments in each county.

County	Total Apartments Number	High-Quality Apartments Number	Percentage
Alameda	1942	707	36.41%
Contra Costa	854	356	41.69%
San Francisco	2894	564	19.49%
San Mateo	1621	733	45.22%
Santa Clara	3593	1923	53.52%

As the table shows above: I find Santa Clara has the largest percentage of high quality apartments (53.52%), while San Francisco has the lowest percentage of high quality apartments(19.94%).

Then, I want to explore the location of high quality apartments by unit price quartile. Similar to the analysis in Davis Apartment, I create a new variable named unit price (price/size) as the standard to find high rent apartments. Cause the unit price is a numeric variable, I prefer to split them into groups by quartile to help me analyze. By quartile, I split apartments into four groups : 2.23 - 2.88 unit price, 2.89 - 3.39 unit price, 3.40 - 4.12 unit price, 4.13 - 19.95 unit price.

As the picture shows below: the high rent apartments are mainly located in: San Francisco, Berkeley, Oakland, Sunnyvale, Stanford, San Jose, Brisbane, Broadmoor and Hillsborough. Then, the low rent apartments are mainly located in Cambrian Park, Seven Trees, East Foothills, Fremont, Hayward, Concord, Waldon, Clayton, Vine Hill, Tara Hills, El Sobrante and East Richmond Heights. I find San Francisco has the lowest percentage of high quality apartments and highest rent unit price. But Santa Clara has the highest percentage of high quality apartments and relative lower unit price rent. Thus, I think Santa Clara is more suitable for people to live, compared to San Francisco. Also, I find the area relatively away form big cities, like SF, Berkeley, San Jose or Stanford, apartments are more likely to have lower unit rent price, higher quality level. That may be because big cities not only have large population, but they also have many big companies to improve city GDP, lead to homes lack, high consumption and salary level. Thus, I don't suggest people live in big cities.



4. Cities have the oldest populations in the Southern San Francisco Bay Area

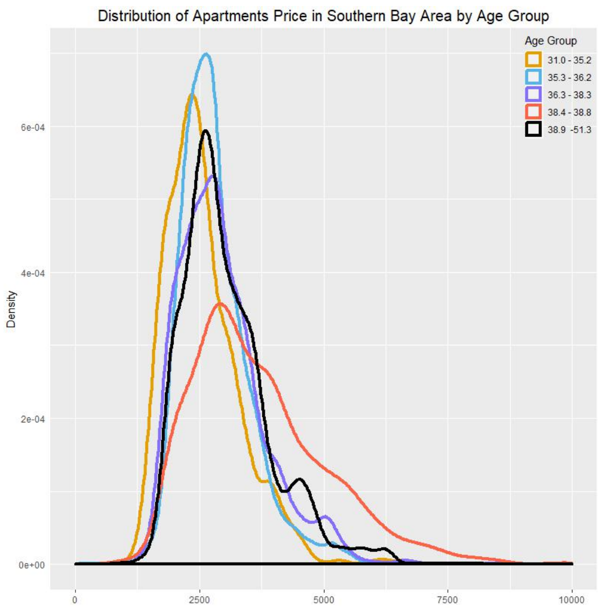
4.1 Cities have the oldest populations

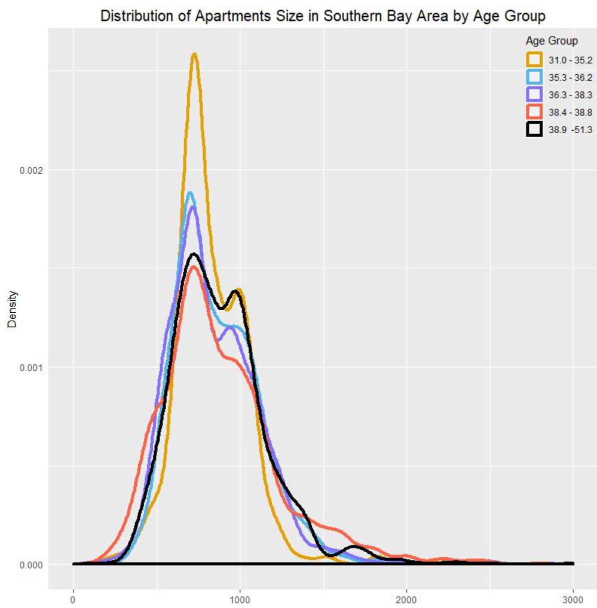
Firstly, I define the oldest populations as people who are 65 years old and over. I need select special column form census data, and merge it to apartments data set.

Because different cities have different population, it will be more meaningful to use the percent of oldest population to find which city has more oldest populations. Form the output, I select top 5 cities that have the oldest populations. They are Portola Valley, Walnut Creek, Los Altos Hills, Atherton, Hillsborough. Their percent of oldest populations are 26.9%, 26.6%, 23.2%, 22.6%, 20.5% separately.

4.2 Relation of Population Age and Rental Market

In this question, I select the median age percent as criteria to split age group. Each group contains even number of apartments, and each group indicates different level of age population. As the picture shows below: I find for apartments' price, the 2nd group (35.3% - 36.2% old population) has an obvious difference to the 4th group (38.4% - 38.8% old population). Apartments in relative old population group tend to have a lower rent. But they tend to have a smaller size apartments. Form 4.1, I find the cities have more old population are small, that's why apartments in those areas have a relative small size and lower rent. Also, I have to mention, the size and rent effect each other, they may be confounding variables. Then, I calculate the proportion of pet allowed, garage parking, in-unit laundry in each age group. I find apartments in high age group have a high probability to have both pet allowed policy, garage parking and in-unit laundry. Thus, I conclude that apartments in old population cities are more likely have a high quality.





5. Limitation

Sample size: I find there are big difference in sample size for cities. For example, San Francisco has 2893 posts, while Manteca has 1 post. The huge difference will lead to some wrong conclusions. Also, for small sample, we can not apply statistical analysis in them. Davis apartments sample size is not big enough. However, cities in large sample always present an obvious influence no matter how limits we do for apartments.

Confounding variables: there are some confounding variables in data set that influence both the dependent variable and independent variable causing a spurious association. That also may lead to some wrong conclusions.

6. Conclusion

From this report, I find, in Davis, apartments would have a high rent when they near to campus, downtown and market; apartments are family friendly if they are far away from downtown; and the Edge Apartments have high quality and relative low rent.

In southern Bay Area, the rent price will increase when they are near to campus or prosperous commerce and numerous residents area; Sunnyvale is a recommend place to live for families; apartments in big cities have a lower quality compared to smaller cities.

Portola Valley, Walnut Creek, Los Altos Hills, Atherton, Hillsborough have the oldest population in southern Bay Area. Apartments in more old population cities tend to have a high quality.

Appendix

```
apartments = readRDS("cl_apartments.rds")
Davis = subset(apartments, city == "Davis")
census = read.csv('2010_census_data/DEC_10_SF1_SF1DP1_with_ann.csv',header = T)
# correct errors
apartments$price[apartments$price >= 30000000] = 3408
apartments$price[apartments$price >= 9900000] = 995

# remove outliers
library(data.table)
outlierReplace = function(dataframe, cols, rows, newValue = NA) {
  if (any(rows)) {
    set(dataframe, rows, cols, newValue)
  }
}
outlierReplace(apartments, "sqft",
               which(apartments$sqft > 5000), NA)
outlierReplace(apartments, "sqft",
               which(apartments$sqft < 100), NA)
outlierReplace(apartments, "price",
               which(apartments$price < 200), NA)

library(ggplot2)
library(ggmap)
library(RColorBrewer)
library(gridExtra)
qmplot(longitude, latitude, data = Davis, colour = I('red'), size = I(3), darken = .1)
#   ggmap davis
bbox = c(
  -121.772613, 38.532675, # bottom left
  -121.695120, 38.571458 # top right
)
m = get_stamenmap(bbox, zoom = 15,maptype = "toner-lite") # zoom default = 10

#   # distribution of pre price
D = Davis[,c('longitude','latitude','price','sqft')]
D$pre = D$price / D$sqft
D = D[,c('longitude','latitude','pre')]

D$quartile = as.integer(cut(D$pre, quantile(D$pre, probs=0:4/4,na.rm = T), include.lowest=TRUE))
D$quartile = as.factor(D$quartile)
D = na.omit(D)
```

```
# get pre-price quartile
quantile(Davis_per_price$pre, c(0,0.25, 0.5, 0.75,1), type = 1)
# 1.309689 1.444328 2.025140 2.189286 5.482517

ggmap(m) + geom_jitter(aes(x = longitude, y = latitude, colour = quartile),data = D,height = 0.001, width =
0.001,size = 3,alpha = 0.85) +
  labs(title = "Distribution of Apartments Unit Price in Quartile : Davis", x='',y='')+
guides(colour=guide_legend(title="Unit Price Quartile")) +
  scale_colour_manual(labels = c("1.31 - 1.44", "1.45 - 2.03","2.04 - 2.19", "2.20 - 5.48"),
                      values=c('darkorchid2','#E69F00','#56B4E9','red3'))+
  theme(legend.position="top",plot.title = element_text(hjust = 0.5,size = 20),
        legend.text = element_text(size=15),legend.title = element_text(size=15),
        axis.text = element_blank(),axis.ticks = element_blank())

Davis_bed = Davis[,c('longitude','latitude','bedrooms')]
Davis_bed$bedrooms = as.factor(Davis_bed$bedrooms)
Davis_bed = na.omit(Davis_bed)
p3 = ggmap(m) + geom_jitter(aes(x = longitude, y = latitude, colour = bedrooms),data = Davis_bed,
                            height = 0.001, width = 0.001,size = 3,alpha = 0.85) +
  labs(title = "Distribution of Apartments Bedrooms Number in Davis", x='',y='')+
  theme(legend.position="top",plot.title = element_text(hjust = 0.5,size = 16),
        legend.text = element_text(size=15),legend.title = element_text(size=14))+
  scale_colour_manual(values=c('darkorchid2','#E69F00','#56B4E9','red3'))


Davis_bath = Davis[,c('longitude','latitude','bathrooms')]
Davis_bath$bathrooms = as.factor(Davis_bath$bathrooms)
Davis_bath = na.omit(Davis_bath)
p4 = ggmap(m) + geom_jitter(aes(x = longitude, y = latitude, colour = bathrooms),data = Davis_bath,
                            height = 0.001, width = 0.001,size = 3,alpha = 0.85) +
  labs(title = "Distribution of Apartments Bathrooms Number in Davis", x='',y='')+
  theme(legend.position="top",plot.title = element_text(hjust = 0.5,size = 16),
        legend.text = element_text(size=15),legend.title = element_text(size=14))+
  scale_colour_manual(values=c('darkorchid2','#E69F00','#56B4E9','red3'))


Davis_parking = Davis[,c('longitude','latitude','parking')]
Davis_parking$parking = as.factor(Davis_parking$parking)
Davis_parking = na.omit(Davis_parking)
p5  = ggmap(m) + geom_jitter(aes(x = longitude, y = latitude, colour = parking),data = Davis_parking,
                             height = 0.001, width = 0.001,size = 3,alpha = 0.85) +
  labs(title = "Distribution of Apartments Parking Policy in Davis", x='',y='')+
  theme(legend.position="top",plot.title = element_text(hjust = 0.5,size = 20),
        legend.text = element_text(size=15),legend.title = element_text(size=15))
```

```
Davis_laundry = Davis[,c('longitude','latitude','laundry')]
Davis_laundry$laundry = as.factor(Davis_laundry$laundry)
Davis_laundry = na.omit(Davis_laundry)
p6 = ggmap(m) + geom_jitter(aes(x = longitude, y = latitude, colour = laundry),data = Davis_laundry,
                                height = 0.001, width = 0.001,size = 3,alpha = 0.85) +
  labs(title = "Distribution of Apartments Laundry in Davis", x='',y='')+
  theme(legend.position="top",plot.title = element_text(hjust = 0.5,size = 20),
        legend.text = element_text(size=15),legend.title = element_text(size=15))

Davis_pets = Davis[,c('longitude','latitude','pets')]
Davis_pets$pets = as.factor(Davis_pets$pets)
Davis_pets = na.omit(Davis_pets)
p7 = ggmap(m) + geom_jitter(aes(x = longitude, y = latitude, colour = pets),data = Davis_pets,
                                height = 0.001, width = 0.001,size = 3,alpha = 0.85) +
  labs(title = "Distribution of Apartments Pets Policy in Davis", x='',y='')+
  theme(legend.position="top",plot.title = element_text(hjust = 0.5,size = 20),
        legend.text = element_text(size=15),legend.title = element_text(size=15))

grid.arrange(p4,p3,nrow = 1)
grid.arrange(p6,p5,nrow = 1)
p7

# high quality Apartments
D_suit = Davis[which(Davis$bathrooms >= Davis$bedrooms),]
D_suit = D_suit[D_suit[,'parking']%in%c('covered','street','garage','paid'),]
D_suit = D_suit[D_suit[,'laundry']%in%c('in-unit','shared','hookup'),]
D_suit$pre = D_suit$price / D_suit$sqft

ggmap(m) + geom_jitter(aes(x = longitude, y = latitude),data = D_suit,
                           height = 0.001, width = 0.001,size = 3,alpha = 0.85,color = 'red') +
  labs(title = "Distribution of High Quality Apartments in Davis", x='',y='')+
  theme(legend.position="top",plot.title = element_text(hjust = 0.5,size = 20),
        legend.text = element_text(size=15),legend.title = element_text(size=15),
        axis.text = element_blank(),axis.ticks = element_blank())

############################################################ Southern Bay Area
range(SBA$latitude)
range(SBA$longitude)

bbox = c(
  -122.60, 36.90, # bottom left
  -121.45, 38.16) # top right
```

```
p = get_stamenmap(bbox)
ggmap(p)
SBA = apartments[apartments[,'county']%in%
                     c('San Francisco', 'San Mateo', 'Santa Clara', 'Alameda','Contra Costa'),]
SBA = na.omit(SBA)
SBA$pre = SBA$price / SBA$sqft
SBA$quartile = as.integer(cut(SBA$pre, quantile(SBA$pre, probs=0:4/4), include.lowest=TRUE))
SBA$quartile = as.factor(SBA$quartile)

# family friendly
S_ff = SBA[SBA[,'bedrooms']%in%c('3','4','5','6'),]
S_ff = S_ff[S_ff[,'bathrooms']%in%c('2.5','3','3.5','4','4.5','5','5.5'),]
S_ff = S_ff[S_ff[,'pets']%in%c('both','cats','dogs','negotiable'),]
S_ff = S_ff[S_ff[,'parking']%in%c('covered','garage','paid'),]
S_ff = S_ff[S_ff[,'laundry']%in%c('in-unit','hookup'),]
S_ff = S_ff[S_ff[,'quartile']%in%c('1','2'),]
S_ff = na.omit(S_ff)

g2 = ggmap(p, extent = "panel", maprange=FALSE) +
  geom_density2d(data = S_ff, aes(x = longitude, y = latitude)) +
  stat_density2d(data = S_ff, aes(x = longitude, y = latitude, fill = ..level.., alpha = ..level..),
                 size = 0.01, bins = 30, geom = 'polygon') +
  scale_fill_gradient(low = "green", high = "red") +
  scale_alpha(range = c(0.00, 0.25), guide = FALSE) +
  theme(legend.position = "none", axis.title = element_blank(),plot.title = element_text(hjust = 0.5,size = 
13),
        axis.text = element_blank(),axis.ticks = element_blank()) +
  labs(title = 'Distribution of Family Friendly Apartments in Southern Bay Area')


# not friendly
S_ff_no = SBA[SBA[,'bedrooms']%in%c('0','1','2'),]
S_ff_no = S_ff_no[S_ff_no[,'bathrooms']%in%c('0','1','1.5','2'),]
S_ff_no = S_ff_no[S_ff_no[,'pets']%in%c('none'),]
S_ff_no = S_ff_no[S_ff_no[,'parking']%in%c('none','off-street','street'),]
S_ff_no = S_ff_no[S_ff_no[,'laundry']%in%c('none','shared'),]
S_ff_no = S_ff_no[S_ff_no[,'quartile']%in%c('1','2'),]

S_ff_no = na.omit(S_ff_no)

g1 = ggmap(p, extent = "panel", maprange=FALSE) +
  geom_density2d(data = S_ff_no, aes(x = longitude, y = latitude)) +
  stat_density2d(data = S_ff_no, aes(x = longitude, y = latitude, fill = ..level.., alpha = ..level..),
                 size = 0.01, bins = 30, geom = 'polygon') +
```

```
  scale_fill_gradient(low = "green", high = "red") +
  scale_alpha(range = c(0.00, 0.25), guide = FALSE) +
  theme(legend.position = "none", axis.title = element_blank(),plot.title = element_text(hjust = 0.5,size =
13),
        axis.text = element_blank(),axis.ticks = element_blank()) +
  labs(title = 'Distribution of Not Family Friendly Apartments in Southern Bay Area')

grid.arrange(g1,g2,nrow = 1)

# southern bay   price/size
SBA_sqft = SBA[,c('longitude','latitude','sqft','price','county')]
SBA_sqft = na.omit(SBA_sqft)
SBA_sqft$pre = SBA_sqft$price / SBA_sqft$sqft
SBA_sqft$quartile = as.integer(cut(SBA_sqft$pre, quantile(SBA_sqft$pre, probs=0:4/4),
include.lowest=TRUE))
SBA_sqft$quartile = as.factor(SBA_sqft$quartile)

# heat map
lon = c(-122.90,-121.25)
lat = c(36.90,38.16)
YlOrBr <- c("#FFFFD4", "#FED98E", "#FE9929", "#D95F0E", "#993404")
MyMap <- get_map(location = c(lon = mean(lon), lat = mean(lat)),
                 source = "google", maptype = "roadmap", crop = FALSE)
ggmap(MyMap) %+% SBA_price +
  aes(x = longitude, y = latitude, z = price) +
  stat_summary_2d(fun = median, alpha = 0.5) +
  scale_fill_gradientn(name = "Price",colours = YlOrBr, space = "Lab") +
  labs(title ='Distribution of Apartments Price in Southern Bay Area',x = "", y = "") +
  coord_map()+ theme(axis.text = element_blank(),axis.ticks = element_blank(),
                     plot.title = element_text(hjust = 0.5,size = 15))

# high quality apartments in S Bay area
SBA$pre = SBA$price / SBA$sqft
SBA$quartile = as.integer(cut(SBA$pre, quantile(SBA$pre, probs=0:4/4,na.rm =T),
include.lowest=TRUE))
SBA$quartile = as.factor(SBA$quartile)
quantile(SBA$pre, c(0,0.25, 0.5, 0.75,1), type = 1,na.rm = T)
# 0.2320205   2.8853914   3.3913043   4.1176471 19.9500000

S_suit = SBA[which(SBA$bathrooms >= SBA$bedrooms),]
S_suit = S_suit[S_suit[,'parking']%in%c('covered','street','garage','paid'),]
S_suit = S_suit[S_suit[,'laundry']%in%c('in-unit','shared','hookup'),]
S_suit = na.omit(S_suit)
```

```
ggmap(p) + geom_jitter(aes(x = longitude, y = latitude, colour = quartile),data = S_suit,
                                    height = 0.001, width = 0.001,size = 1) +
   labs(title = "Distribution of High Quality Apartments in Southern Bay Area", x='',y='')+
   theme(legend.position="right",plot.title = element_text(hjust = 0.5,size = 14),
            legend.text = element_text(size=10),legend.title = element_text(size=12),
            axis.text = element_blank(),axis.ticks = element_blank())+
guides(colour=guide_legend(title="Unit Price Quartile"))+
   scale_colour_manual(labels = c("0.23 - 2.88", "2.89 - 3.39","3.40 - 4.12", "4.13 - 19.95"),
                                   values=c('darkorchid2','#E69F00','#56B4E9','red3'))

table(factor(S_suit$county))
table(factor(SBA$county))

################################### oldest population
levels(SBA$city)
census = read.csv('2010_census_data/DEC_10_SF1_SF1DP1_with_ann.csv',header = T)
census = census[-1,]
library(stringr)
census[, 3] = str_remove_all(census[, 3], ' (town|city|CDP|CDP \\(.+ County\\)), California')

old = census[,c('GEO.display.label','HD02_S025','HD01_S020')]
colnames(old)[colnames(old)=="GEO.display.label"] <- "city"
colnames(old)[colnames(old)=="HD01_S020"] <- "Total population percent median age"
colnames(old)[colnames(old)=="HD02_S025"] <- "Total population Percent 65+"

SBA_2 =
SBA[,c('latitude','longitude','price','sqft','bedrooms','bathrooms','pets','laundry','parking','city','county','pre','quartile')]
M = merge(old,SBA_2,by = 'city')

M_1 = M[,c(1,2,3)]
M_1 = unique(M_1)
M_2 = M[,c(1,3)]
M_2 = unique(M_2)
M_2$`Total population percent median age` = as.character(M_2$`Total population percent median age` )
M_2$`Total population percent median age` = as.numeric(M_2$`Total population percent median age` )
M_2$quartile = as.integer(cut(M_2$`Total population percent median age`, quantile(M_2$`Total population percent median age`, probs=0:4/4,na.rm =T), include.lowest=TRUE))
M_2$quartile = as.factor(M_2$quartile)
quantile(M_2$`Total population percent median age` , c(0, 0.5, 1), type = 1,na.rm = T)
# 31.0 36.1 38.7 42.6 51.3

M$`Total population percent median age` = as.character(M$`Total population percent median age` )
M$`Total population percent median age` = as.numeric(M$`Total population percent median age` )
```

```
group = rep(0, nrow(M))
M = cbind(M, group)
M = na.omit(M)
breaks = quantile(M$`Total population percent median age`, probs = seq(0, 1, 0.2), name = FALSE)

for(i in 1:nrow(M)){
  for(j in 1:5){
    if(M$`Total population percent median age`[i] >= breaks[j] & M$`Total population percent median
age`[i] <= breaks[j+1]){M$group[i] = j}
  }
}

M$group = as.factor(M$group)
h1 = ggplot(M,aes(x=price, color = group)) + geom_density(alpha = 0.3,size = 1.5) +
  labs(title = "Distribution of Apartments Price in Southern Bay Area by Age Group") +
  theme(plot.title = element_text(hjust = 0.5,size = 15))+
  labs(y="Density")+
  labs(x=" ") + scale_color_manual(labels = c("31.0 - 36.5", "36.6 - 51.3"),
                                   values=c("#E69F00", "#56B4E9"))+
  theme(legend.background = element_blank(),legend.justification=c(1,1), legend.position=c(1, 1),
        axis.ticks.y = element_blank()) +
  guides(color=guide_legend(title="Age Group")) + xlim(0,10000)


h2 = ggplot(M,aes(x=sqft, color = group)) + geom_density(alpha = 0.3,size = 1.5) +
  labs(title = "Distribution of Apartments Size in Southern Bay Area by Age Group") +
  theme(plot.title = element_text(hjust = 0.5,size = 15))+
  labs(y="Density")+
  labs(x=" ") + scale_color_manual(labels = c("31.0 - 36.5", "36.6 - 51.3"),
                                   values=c("#E69F00", "#56B4E9"))+
  theme(legend.background = element_blank(),legend.justification=c(1,1), legend.position=c(1, 1),
        axis.ticks.y = element_blank()) +
  guides(color=guide_legend(title="Age Group")) + xlim(0,3000)

grid.arrange(h1,h2,nrow = 1)

h3 = ggplot(M, aes(x = as.factor(bedrooms) , fill = group)) +
  geom_bar(position = "dodge")+labs(title = "Number of Bedrooms by Age Group", x = "Number of
Bedrooms") +
  theme(axis.ticks.x = element_blank())+theme(plot.title=element_text(hjust=0.5))+
  scale_fill_manual(labels = c("31.0 - 35.2", "35.3 - 36.2",'36.3 - 38.3','38.4 - 38.8','38.9   -51.3'),
                    values=c("#E69F00", "#56B4E9",'lightslateblue','tomato1','black'))+
```

```
  theme(legend.background = element_blank(),legend.justification=c(1,1), legend.position=c(1, 1))+
  geom_text(stat='count', aes(label=..count..), vjust=-0.2,position = position_dodge(width = 1))

h4 = ggplot(M, aes(x = as.factor(pets) , fill = group)) +
  geom_bar(position = "dodge")+labs(title = "Pets Allowed by Age Group", x = "Pets Allowed") +
  theme(axis.ticks.x = element_blank())+theme(plot.title=element_text(hjust=0.5))+
  scale_fill_manual(labels = c("31.0 - 35.2", "35.3 - 36.2",'36.3 - 38.3','38.4 - 38.8','38.9   -51.3'),
                    values=c("#E69F00", "#56B4E9",'lightslateblue','tomato1','black'))+
  theme(legend.background = element_blank(),legend.justification=c(1,1), legend.position=c(1, 1))+
  geom_text(stat='count', aes(label=..count..), vjust=-0.2,position = position_dodge(width = 1))

h5 = ggplot(M, aes(x = as.factor(parking) , fill = group)) +
  geom_bar(position = "dodge")+labs(title = "Parking Policy by Age Group", x = "Parking Policy") +
  theme(axis.ticks.x = element_blank())+theme(plot.title=element_text(hjust=0.5))+
  scale_fill_manual(labels = c("31.0 - 35.2", "35.3 - 36.2",'36.3 - 38.3','38.4 - 38.8','38.9   -51.3'),
                    values=c("#E69F00", "#56B4E9",'lightslateblue','tomato1','black'))+
  theme(legend.background = element_blank(),legend.justification=c(1,1), legend.position=c(1, 1))+
  geom_text(stat='count', aes(label=..count..), vjust=-0.2, position = position_dodge(width = 1))
h6 = ggplot(M, aes(x = as.factor(laundry) , fill = group)) +
  geom_bar(position = "dodge")+labs(title = "Laundry Situation by Age Group", x = "Laundry Situation") +
  theme(axis.ticks.x = element_blank())+theme(plot.title=element_text(hjust=0.5))+
  scale_fill_manual(labels = c("31.0 - 35.2", "35.3 - 36.2",'36.3 - 38.3','38.4 - 38.8','38.9   -51.3'),
                    values=c("#E69F00", "#56B4E9",'lightslateblue','tomato1','black'))+
  theme(legend.background = element_blank(),legend.justification=c(1,1), legend.position=c(1, 1))+
  geom_text(stat='count', aes(label=..count..), vjust=-0.2, position = position_dodge(width = 1))

grid.arrange(h5,h6,h3,h4,nrow = 2)
by_group = split(M,M$group)
table(by_group[[1]]$pets)
table(by_group[[1]]$laundry)
table(by_group[[1]]$parking)
table(by_group[[2]]$pets)
table(by_group[[2]]$laundry)
table(by_group[[2]]$parking)
table(by_group[[3]]$pets)
table(by_group[[3]]$laundry)
table(by_group[[3]]$parking)
table(by_group[[4]]$pets)
table(by_group[[4]]$laundry)
table(by_group[[4]]$parking)
table(by_group[[5]]$pets)
table(by_group[[5]]$laundry)
table(by_group[[5]]$parking)
```