University of California, Davis

Homework # 3: Working with Craiglist Data

Ana Boeriu
STA 141A
Dr. Nick Ulle
October 29, 2018

I. Introduction

Craigslist is a website that allows people to post classified advertisement for free. These posts can span a variety of subjects such as cars, bikes, jobs, and even housing. For this assignment I will be focusing on apartment rental postings in California. I will be using exploratory data analysis to further analyse any trends and patterns in the data. This report can be useful for anyone looking to rent an apartment in various locations in California. Because this data is created by Craiglist, I need to be aware that there can be multiple errors,which can skew the dataset.

II. Cleaning up the dataset

Before I begin exploring this dataset, I will first begin by observing any outliers or anomalies in this dataset. I will look at the variables state,bedrooms,bathrooms and price. Any outliers or anomalies found in this dataset will be discussed below removed, fixed or replaced with na.

A. States

The variable states has some anomalies. There are nine unique states in the dataset that consisted of California, Connecticut, Florida,Nevada, Ohio, Utah, Virginia, Washington and North Carolina. All states (with the exception of California and Utah), used California craigslist to post about apartment rentals in the respective state. Thus seven of the nine states were removed. California and Utah were the only states that posted about apartment rentals in California thus I kept those in the dataset. Because the purpose of this dataset is to provide information regarding apartment rentals in California, regardless of where the person actually lives, I will be removing all states (except for Utah and California) from the data set. This however does not suggest that I have removed all the errors. There could still be some more errors.

B. Bedrooms

Next, I will also look at the number of bedrooms apartments have. The floorplan of the apartments ranges from studio apartments (0 bedrooms) to seven bedrooms. It is highly unlikely that an apartment has seven bedrooms. Normally a house would be more likely to have seven bedrooms. Upon exploring the variable bedrooms, I noticed that all apartment posts regarding 7 bedrooms are actually cleaning services. Similarly, all five and six bedroom apartments are actually houses. Thus I will be removing all the posts with 5,6 or 7 bedrooms in order to focus more on apartment rentals.

C. Bathrooms

Furthermore I can notice that there are some apartments that have more bathrooms than bedrooms. In that case I have set any apartments where number of bathrooms is larger than number of bedrooms and the apartment has more than one bedrooms to na. This will better reflect the data and prevent any unnecessary removal of potentially useful data.

D.Prices

When assessing the variable prices I have noticed that some posts did not include a hyphen between the possible range of prices. For example, one post in row 15961 (of the original dataset) included $3,4083,742 as the price for a 3 bedroom in Mountain View. It is unlikely that a 3 bedroom would cost 3 million dollars per month. The actual price was a range $3408 - $3742. Thus, for such instances, I have modified prices for those postings to the largest value in the range.

E. Note

While removing any unwanted data and fixing some mistakes, I noticed that Utah is no longer included in the variable state. The data that I removed could have included Utah as a state.

III. Getting familiar with the data

After removing the necessary anomalies and outliers, I will give a brief overview of the data and an overall summary.

A. Rows

The rows correspond to the number of observations or, in this dataset,the number of apartment rental posts in the data. There are a total of 20770 apartment posts.

B. Columns

For each observation recorded there are 20 variables (columns in the dataset). This excludes the variable I created called family friendly which will be discussed later. The variables contain information such as the date that post was published on, the date the post was edited (if at all), if the post was deleted, size of the apartment,number of bedrooms and bathrooms, and cost of apartment any many more.

C. Span of the rental posts

In this data set the apartment posts range from September 8,2018 to October 15,2018. This includes the date posted and also the date deleted.

D. Missing Data

Number of dat	Number of cells with missing data
425,790	31,150

Out of the total number of cells , only 6.82% is missing.

How many listings do not have something missing	How many listings have something missing
2,714	18,056

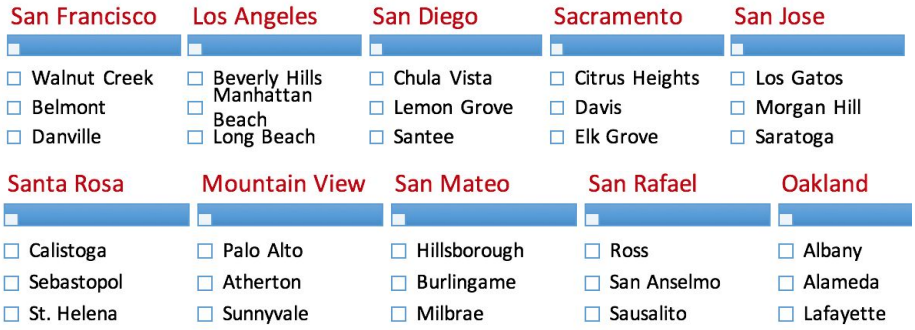
These numbers give a better understanding of how apartmen postings have something missing in the post. For example, maybe county is missing or square feet etc.

IV. Family Friendly Apartments Based on Suburban or City Location

The goal of this part will be to determine which apartments are more family friendly based on a city or suburban location. I will define family friendly as apartments that are larger than one bedroom, allow pets, and also have parking space available. Any apartment posting that satisfies these conditions will be considered family friendly. To be able to categorize these postings I will create a new column (variable) called family friendly that checks for the previously stated conditions. Furthermore, below I will define how I chose my cities and suburbs and plot a graph to show if cities are more family friendly than suburbs or vice versa.

A. Defining a city and a suburb

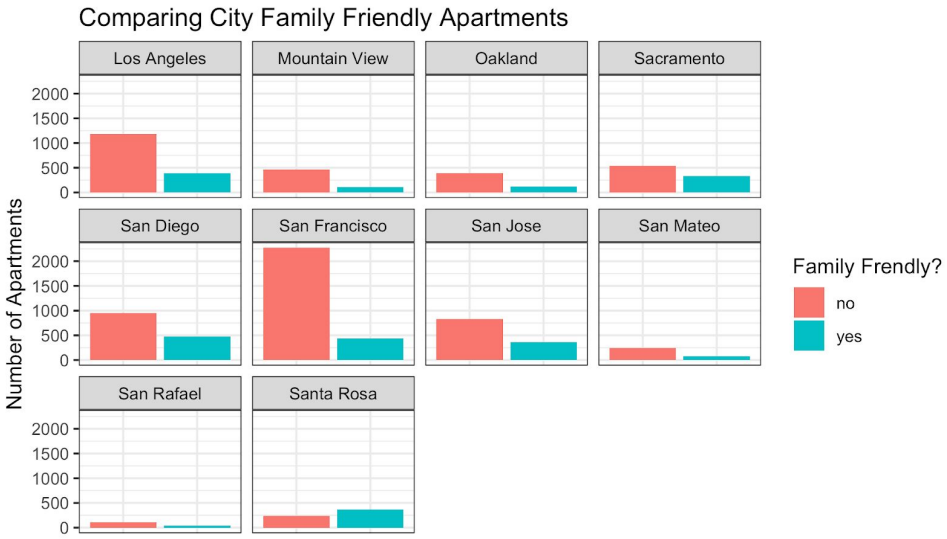
In this section I will focus on defining a city and a suburb. The cities were chosen based on top popularity(how many occurrences were seen throughout the dataset), and well known cities(which ones people know is a city). Similarly the suburbs were chosen based on closest distance to the respective cities by looking at google maps. Below, is a figure depicting the cities(red) and suburbs(below in black). Each city has three suburbs. From here on, unless otherwise stated, when I refer to cities and suburbs I will use the specific locations from the chart shown below.



B. Comparing Family Friendly apartments between cities and suburbs

Here I will look at family friendly apartments in cities versus suburbs based on my definition from above. Recall that I defined family friendly as any apartment larger than one bedroom, allows pets, and also has parking space available.

1.City



This plot shows which cities satisfy the condition of family friendly apartments and also how many apartments, in the respective city, are either family friendly or not family friendly. From the plot we can notice that San Francisco has the most apartments and is the least family friendly city. Santa Rosa is the only city that adhere the condition of family friendlyness. Furthermore, San Rafael has the least amount of apartments.

2. Suburbs

Here we will look at the family friendliness of suburburban apartments.

In this graph we can see which suburban places are family friendly and also which are more popular (have more apartments). The suburbs that have very little data satisfying the condition of family friendly, such as Ross and Saratoga, show that there were very few (if any) apartment rental posts for those locations. Out of the suburbs listed above in the graph, Sunnyvale is the most popular non- family friendly suburb followed by Palo Alto or Chula Vista. Furthermore, Citrus Heights and Elk Grove are the family friendly suburbs.

VI. Effect of Number of Bedrooms and Bathrooms on Price

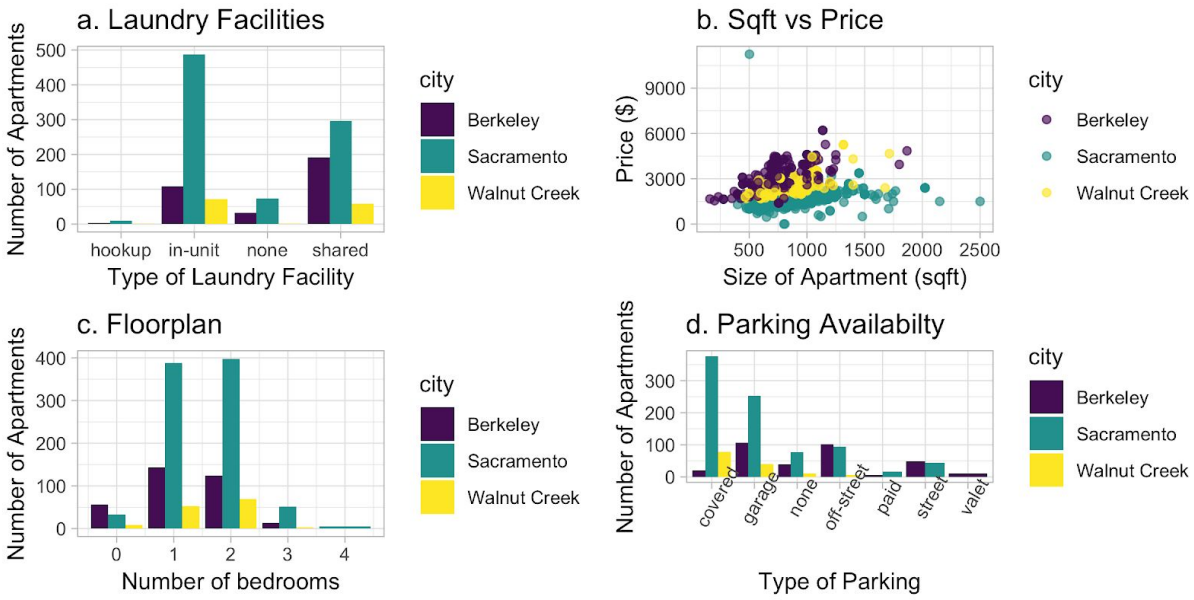
I will be looking at how the number of bedrooms and bathrooms affect the price of an apartment. I will be looking at how the number of bedrooms and bathrooms affect price regardless of a city or suburb

location of the apartment.

In the graph presented, there is a positive trend between price and number of bedrooms. As the number of bedrooms increases, so does the price on average. The highest price for a studio apartment is $11,250 in Sacramento. Furthermore, the average price for a studio apartment is $2,10.78. Similarly the higest price for a two bedroom apartment is $16,500 while the average price for a two bedroom is $2,174.77. For a four bedroom apartment I can notice that the cost of an apartment with three bathrooms is more expensive than an apartment with four bathrooms. This difference in price could also be influenced by other variables such as location. The four bedroom four bathroom apartment is located in Tiburon while the four bedrooms 3 bathrooms is located in San Fransisco. Because of such instances, the data is not consistent enough to determine what affects the cost of renting an apartment.

VII. Similarities of Apartments in Similar Geographical Regions.

In this section I will be looking at how apartments are similar and different in three geographically similar cities. More specifically I will look at Berkeley, Walnut Creek and Sacramento. I will be looking at laundry facilities,floorplan, square footage and parking availability. I choose these cities based on a topographical map provided by Google Maps and also Wikapedia.



a. Number of Apartments with Specified Laundry Facilities

This graph highlights the different types of laundry facilities that each location has. Sacramento has the most apartments with in-unit laundry facilities while Walnut Creek has the least number of in unit washers and dryers. Similarly, out of all the locations, Walnut Creek has the least number apartments with any type of laundry facility. Furthermore all the locations have the most apartments with either in unit or shared laundry facilities.

b. Looking at Price and Size of Apartments

On average, Berkely has the highest prices compared to Sacramento and Walnut Creek. However it is important to point out that overall, Berkely apartments (which are smaller) are more expensive than Sacramento apartments(which are larger). Furthermore, Sacramento has an unusually expensive apartmant that costs about $10,000 ($11,250 to be exact) for 502 square feet. This value should be considered an error in the dataset.

c. Floorplan

For all of these locations, one and two bedroom apartments are the most popular. Sacramento has the highest number of one bedroom and two bedroom apartments and is also the only location that has four bedroom apartments. Furthermore the distribution of Walnut Creek apartments is different than the distribution Sacramento apartments. In other words, there is an increasing trend for Walnut Creek while for Sacramento there is an increase then a decrease.

d. Parking

All the locations offer some kind of parking and have garage parking as one of the most popular parking types. Sacramento has the most apartments with either covered parking of garage parking. Furthermore, Walnut Creek is the only location that has very few apartments with no parking.

VIII. Questions that can be Answered With this Dataset

I will be asking 10 questions in relationship to this dataset that I thought of when working with the dataset. In addition to writing down the questions I will provide an explanation as to who can benefit from the anser. Some questions may be follow up questions from previous analysis done above.

A. **How does the floorplan of the apartment affect square footage between cities and suburbs?**
   - ❑ This question is useful for anyone looking to buy or rent an apartment. If a higher price signifies a larger apartment than someone might choose to get a smaller one.

B. **How does price of the apartment influence parking space availability between cities and suburbs?**
   - ❑ People will want to know what kind of parking is offered between different apartments. Furhtermore, someone may want to know if a more expensive price means a more sophisticated type of parking like valet.

C. **How does pet policy differ between city and suburbs?**
   - ❑ Anyone looking to bring their pets along will want to be informed of apartments pet ploicies especially with cities and suburbs since this could be a determining factor.

D. **How does square footage affect prices in the city?**
   - ❑ Someone who is looking in their pricerange of theor budget will definately need to consider the effect of size. If necessary the person could decide to rent a smaller apartment if the desired apartment is out of his or her price range.

E. **How do Chula Vista and San Fransisco differ in the architecture of apartment rentals?**
   - ❑ Both are big cities and someone may want to know how they are different. Specifically a person may want to know how is an apartment in southern California differnt that an apartment in northern California.

F. **How much is the difference in price between a 2 bedroom 1 bathroom and a 2 bedroom 2 bathroom apartment?**
   - **a.** This question is relevant to see how much prices differ between an apartment with the same bedroom size but different bathroom size. Someone considering a budget may want to choose the less expensive regardless of the number of bathrooms.

G. **How does size of the apartment influence parking space availability?**
   - ❑ People who have multiple cars may want to know what parking is available and if larger sized apartments provide more parking.

H. **Which craigslist tend to be more popular?**
   - ❑ Someone looking for an apartment may want to know where to look for a broader variety of apartments.

I. **What day of the week do people tend to post the most and how often?**
   - ❏ This can help people determine when would be a good day to check for new apartments posts. Furthermore it will also help determine how often apartment rentals are becoming available.

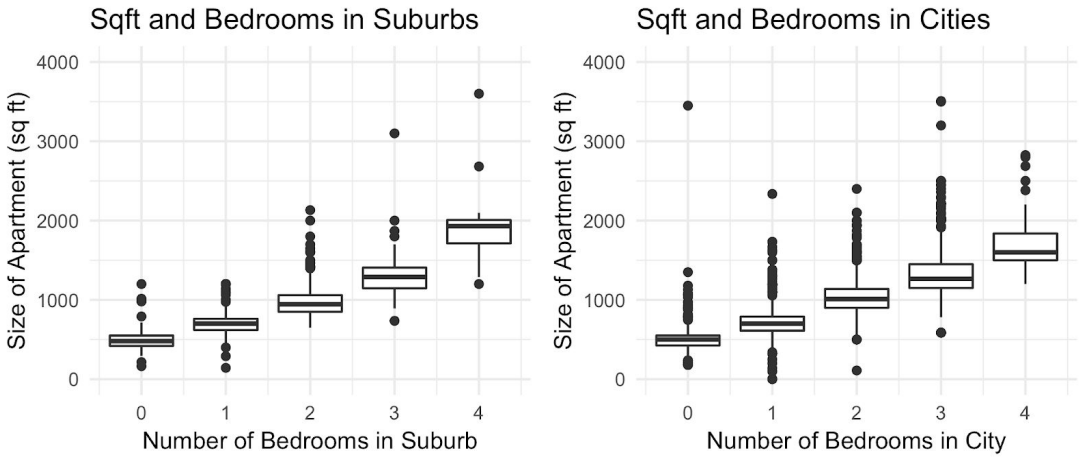
J. **What season has the most apartment posts? Are those posts updated or left alone?**
   - ❏ If posts are updated more often this would suggest that the person renting out the apartment is having a hard time finding someone. Thus by figuring out what time of year and how often posts are updated, people can plan ahead and begin looking for apartments earlier.

VIIII. Answering five of the above questions

Below I will be answering five of the questions I asked above. There is no particular order and each question is stated before I answer it.

A. **How does the floorplan of the apartment affects size of apartments between cities and suburbs?**

I will be looking at the size of the apartment (in sq ft) and number of bedrooms for apartments in cities and suburbs. Because there are only three unusual values that exceed 4,000 square feet (one of which is a one bedroom and the other are 2 bedrooms) I will keep the y axis consistent with the same units and labels. I will limit the y axis to a maximum of 4,000 square feet.

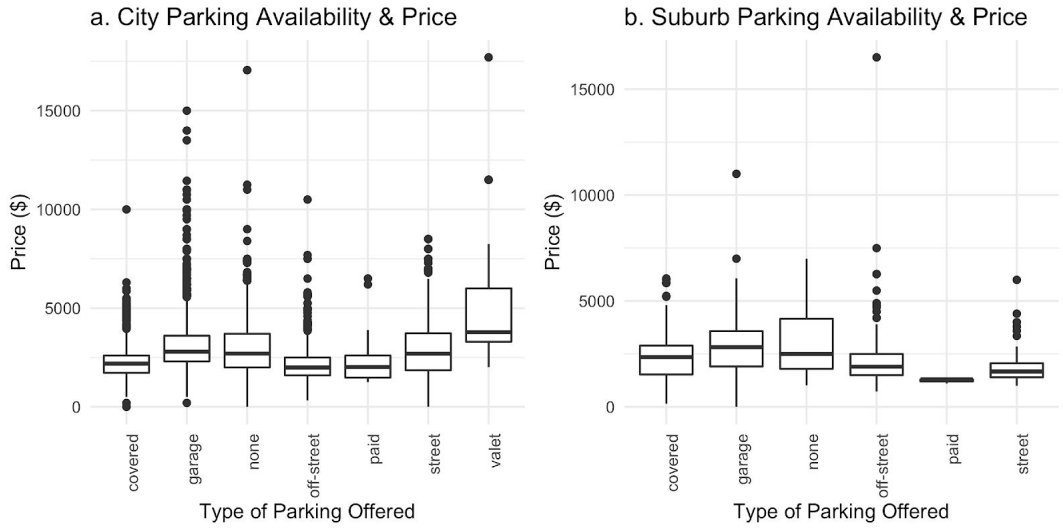


In this graph I grouped all of the chosen cities and suburbs and plotted them in two seperate graphs. The apartments in the suburbs have majority four bedroom apartments. To futher expand, the 75th percentile for a four bedroom apartment (for both a city or suburb location) is larger than any other type of apartment in the suburbs. Both graphs have a positve trend where as the square footage increases, the number of bedrooms also increases. If I only consider the cities, there seems to be an quite a high value for a studio apartment. The studio apartment outlier has a total square fottage of 3,450 in the city of Los Ageles.. Furthermore the highest square footage that an apartment in the suburbs has is 3,600. This 3,600 sq ft apartment has four bedrooms and is located in the suburb of Beverly Hills. The 3,200 square foot

apartment with 3 bedrooms is located in the suburb of Sierra Madre. Although these values are unusual, I would not consider them outliers since they are not outlandish.

B. How does price of the apartment influence parking space availability between cities and suburbs?

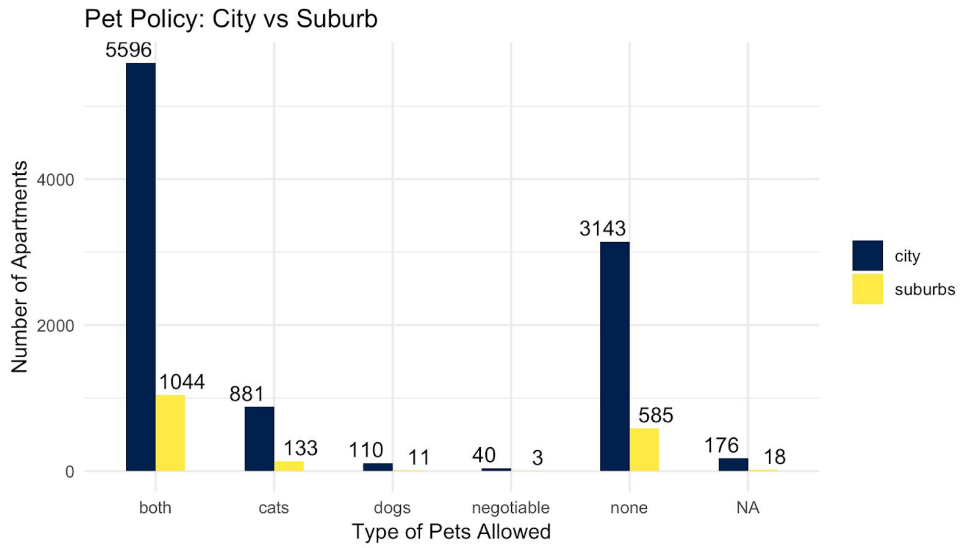
Here I will be examining the effect of price and parking and if a higher price implies that apartment complexes have parking provided.



In these graphs the type of parking corresponds to how expensive an apartmetn rental is on average. For example, valet parking would correspond to luxury apartments. An important note is that city valet parking and suburban non parking are both right skewed and seem to have a large variability compared to any other parking category in the respective locations(city or suburb). Overall city apartments tend to have much more parking than suburban apartments. Furthermore, valet parking in the city has the highest median value. The 75th percentile for valet parking in the city is larger than any other city parking category. On the contrary, valet parking in the suburbs has the second lowest median value.

C. How does pet policy differ between city and suburbs?

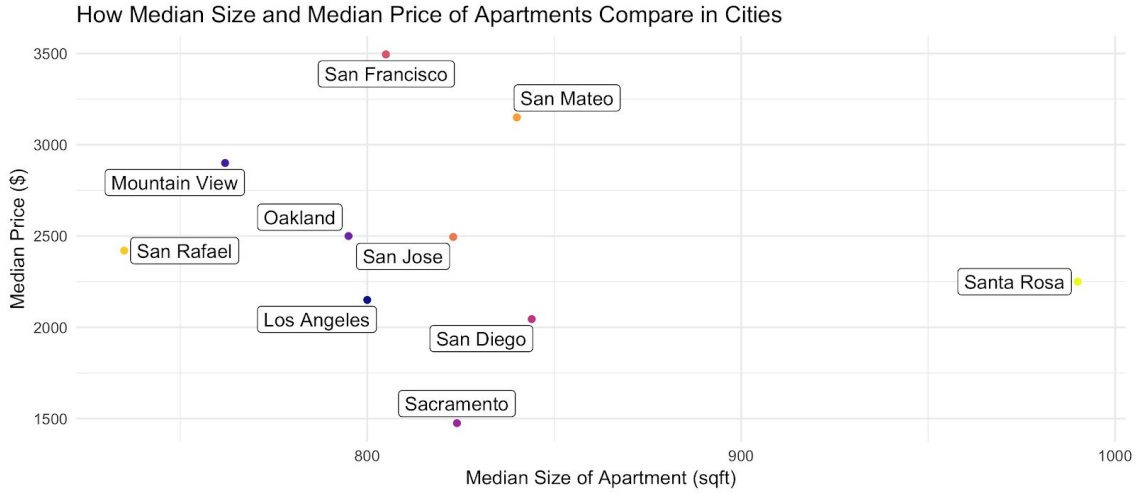
Here we will be looking at how cities and suburbs have different pet policies by grouping cities and suburbs into two groups.



In this graph, apartments in the suburbs tend to have much fewer pet policies compared to city, I can notice that for cities, the ratio of no pets allowed(none) to allowing both dogs and cats is about 50%. This suggests that for every two apartments that do not allow any pets there will be one apartment that allows both dogs and cats. Furthermore, there are much fewer apartments in the suburbs that have a pet policy compared to city apartments. In both categories(city and suburbs) the two most popular pet policy enforcement is allowing both dogs and cats and not allowing any pets.

D. How does square footage affect prices in the city?

In this section we will be comparing size and price of the apartments. Because there are lots of apartment posts for the respective cities, we will be focusing on the median price and median square footage which will produce only 10 points. I will chose to the median because it is affected less by skewness of the data than compared to the mean.

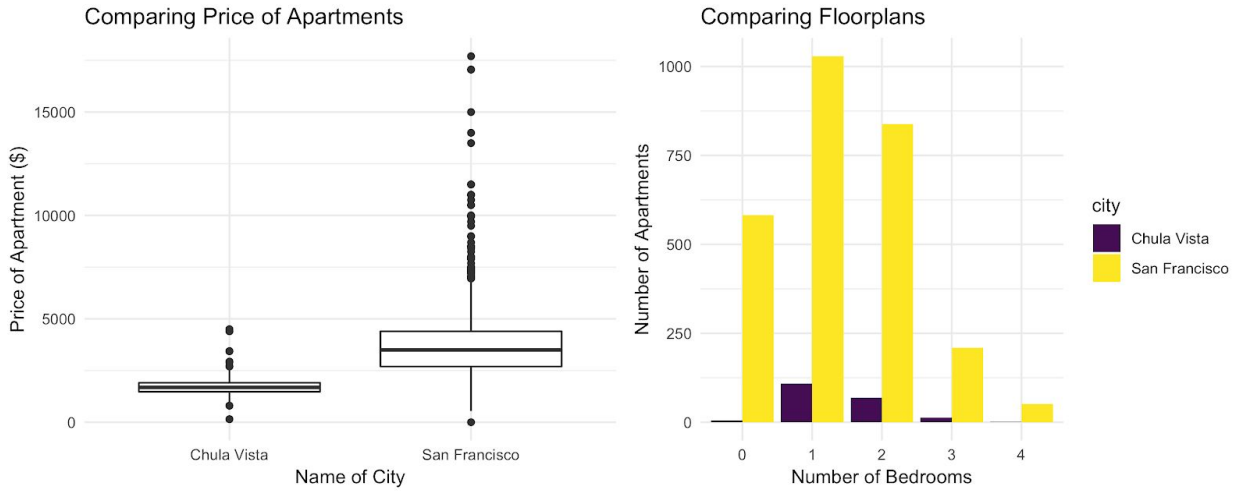


This graph shows that Mountain View and San Rafael are the cities with the most expensive apartments. On the other hand, Santa Rosa has the least expensive apartment prices comparing to the rest of the cities.

Moreover, San Francisco has the highest median price and a smaller median square footage. This suggests that San Fransisco has the most expensive apartments. What is interesting to note is price difference between Sacramento and San Josel. Even though they have very similar sizes of apartments, the prices are quite different. San Jose has a much larger average price, causing apartments in that location to be more expensive than apartmetns in Sacramento.

E. How does Chula Vista and San Fransisco differ in the architecture of apartment rentals?

I will be examining the cost of apartments in Chula Vista versus San Francisco and look for any trends in the data.



In this graph, San Francisco has about 75% more apartments than Chula Vista. Furthermore, Chula Vista has much less variety of prices than San Fransisco. This could be due to the fact that San Fransisco has more apartmetns than Chula Vista. However despite the differences, both locations have the highest number of apartments for one and two bedrooms.

X. Discussing Limitations

Throughout this report I have discussed some errors and anomalies in this dataset. However I do not guarantee that I found all the errors. The way this data set was generated was through Craigslist. Because the public generated this dataset, one should expect errors. Similarly, this dataset focused on apartment rentals in California, which can lead bias which can cause data to be inaccurate and skewed. Overall there seems to be enough observations to support my data, however a more accurate and reliable information could have helped me further provide stronger statistical data analysis.

XI R Appendix

```
##intro
cl_apartments = readRDS("/Users/aboeriu/Documents/UC Davis/Year 3/STA 141A/assignment
3/cl_apartments.rds")
library(ggplot2)
library(viridis)
library(grid)
library(gridExtra)
library(ggpubr)

## clean up part

## look at prices
sort(cl_apartments$price,decreasing = T)
cl_apartments[which.max(cl_apartments$price=="34083742"),]
cl_apartments$price[15961] = 3742
cl_apartments[which.max(cl_apartments$price=="9951095"),]
cl_apartments$price[ 4531]=1092
range(cl_apartments$price,na.rm = T)
##############################################################################
# removing 5-7 bedroom apartments

cl_apartments=cl_apartments[cl_apartments$bedrooms<5,]
table(cl_apartments$bedrooms)

##############################################################################

#remove all states but CA
#some_states= (cl_apartments$state=="CA" & cl_apartments$state=="UT")
cl_apartments=cl_apartments[which(cl_apartments$state=="CA"),]

table(cl_apartments$state)
##############################################################################

#look at the states

table(cl_apartments$state)
which(cl_apartments$state=="CT")
cl_apartments[3657,]
which(cl_apartments$state=="FL")
which(cl_apartments$state=="MD")
which(cl_apartments$state=="NC")
which(cl_apartments$state=="NV")
```

```
which(cl_apartments$state=="OH")
#remove ohio,MD,NC,NV,FL,VA,WA, entries b/c apartments are not in California.
which(cl_apartments$state=="UT")
#remove UT because appt is in UT and they posted on sfcraigslist
which(cl_apartments$state=="VA")
#set entry 8248 to NA b/c appt is in tijuana mexico
which(cl_apartments$state=="WA")
#########################################################################
#look at # of bedrooms
table(cl_apartments$bedrooms)
which(cl_apartments$bedrooms==5)
cl_apartments[439,]
#this is actually a home

#########################################################################
#look at bathrooms
which(cl_apartments$bathrooms==6)
which(cl_apartments$bathrooms==5)
which(cl_apartments$bathrooms==4.5)
which(cl_apartments$bathrooms==4)
which(cl_apartments$bathrooms==3.5)
which(cl_apartments$bathrooms==3)
which(cl_apartments$bathrooms==2.5)

which(cl_apartments$bathrooms==0)
#fixed these
cl_apartments$bathrooms[11412]=2
cl_apartments$bathrooms[11355]=1

table(cl_apartments$bathrooms)
#########################################################################################
#############
#where bath > bedrooms
ind = which(cl_apartments$bedrooms > 0 & cl_apartments$bedrooms < cl_apartments$bathrooms)
cl_apartments$bathrooms[ind] = NA
#########################################################################################
#########################
#check range of posts
range(cl_apartments$date_posted)
range(cl_apartments$date_posted)

#Missing data(mia)
table(is.na(cl_apartments))
```

```
#table(is.na(cl_apartments[cl_apartments$city,]))
any(is.na(cl_apartments[1,]))

table(apply(cl_apartments, 1,FUN = function(x){any(is.na(x))}))
total_obs=31150+425790
percentage= (31150/total_obs)*100
percentage
##############################################################################################
###################
#2.1.2
#unique(cl_apartments$state)
#table(cl_apartments$state)
#which(cl_apartments$state=="CA")
#cl_apartments[which(cl_apartments$state=="CA"),]
#table(cl_apartments$state[is.na(cl_apartments$city)])
#out of these states _ have missing city info
#sum(cl_apartments$state=="CA",na.rm=T)
##############################################################################################
########################
#identifying top 10 major cities
sort(table(cl_apartments$city[cl_apartments$state=="CA"]),decreasing=T)
head(sort(table(cl_apartments$city[cl_apartments$state=="CA"]),decreasing=T),10)
top10_cities<-tail(sort(table(cl_apartments$city[cl_apartments$state=="CA"])),10)
top10_cities
cl_apartments$city%in%names(top10_cities)


#choosing suburbs
#here I chose the cities based off of the top number of cities that
#had the most appt sales, and also preference

#the suburbs I choose based on prefrence how close it was to city using
#google maps

suburbs=c("Walnut Creek","Belmont","Danville",
        "Beverly Hills","Manhattan Beach", "Long Beach",
        "Chula Vista","Lemon Grove","Santee",
        "Citrus Heights","Davis","Elk Grove",
        "Los Gatos","Morgan Hill","Saratoga",
        "Calistoga" ,"Sebastopol","St. Helena",
        "Palo Alto","Atherton","Sunnyvale",
        "Hillsborough","Burlingame","Millbrae",
        "Ross","San Anselmo","Sausalito",
```

```
        "Albany","Alameda","Lafayette")

major_10_cities = c("San Francisco",
                    "Los Angeles",
                    "San Diego",
                    "Sacramento",
                    "San Jose",
                    "Santa Rosa",
                    "Mountain View",
                    "San Mateo",
                    "San Rafael",
                    "Oakland")

#city_and_suburbs = data.frame(major_9_cities, suburbs)

################################################################################################
################
# table of my cities and # of bedrooms

#table(cl_apartments$city=="San Fransisco"[cl_apartments$pets])
my_city=cl_apartments$city %in% major_10_cities
organize_city=cl_apartments$city[my_city]
display_only_my_city=droplevels(organize_city,na.rm=T)
table(cl_apartments$bedrooms[my_city],display_only_my_city)
#table(cl_apartments$state[cl_apartments$bedrooms])
table(cl_apartments$pets[my_city],display_only_my_city)
sum(cl_apartments$bedrooms[my_city],display_only_my_city,na.rm=T)
table(cl_apartments$bedrooms,cl_apartments$pets)

# table of suburbs same as above

my_suburbs=cl_apartments$city %in% suburbs
organize_suburbs=cl_apartments$city[my_suburbs]
display_only_my_suburbs=droplevels(organize_suburbs)
#sum(cl_apartments$bedrooms[my_suburbs],display_only_my_suburbs,na.rm=T)
#sum(cl_apartments$bedrooms[my_suburbs]>3,display_only_my_suburbs,na.rm=T)
#mean(cl_apartments$bedrooms[my_suburbs],display_only_my_suburbs,na.rm=T)

################################################################################################
################

#How floorplan of the apartment affects size (in sq ft) in cities and suburbs
```

```
f1=ggplot(cl_apartments[cl_apartments$city %in% suburbs,])+
  geom_boxplot(aes(x=bedrooms,y=sqft, group=bedrooms))+
  ggtitle("Sqft and Bedrooms in Suburbs")+
  xlab("Number of Bedrooms in Suburb")+
  ylab("Size of Apartment (sq ft)")+
 coord_cartesian(ylim = c(0, 4000))+
  theme_minimal()+
  ggsave("suburban bedrooms and Sqft.png",width=7, height = 4)

#to find highest sq ft in the 4 bed category
which.max(cl_apartments$sqft>3000 & cl_apartments$bedrooms==4)
cl_apartments[2199,]
cl_apartments$sqft[2199]

#to find highest pt in 3 bed category
which.max(cl_apartments$sqft>3000 & cl_apartments$bedrooms==3)
cl_apartments[1898,]
cl_apartments$sqft[1898]

# do the same for city
f2=ggplot(cl_apartments[cl_apartments$city %in% major_10_cities,])+
  geom_boxplot(aes(x=bedrooms,y=sqft, group=bedrooms))+
  ggtitle("Sqft and Bedrooms in Cities")+
  xlab("Number of Bedrooms in City")+
  ylab("Size of Apartment (sq ft)")+
  #coord_cartesian(ylim = c(0, 4000))+
  theme_minimal()+
  ggsave("city beds &sqft.png",height=4,width = 7)
#side by side plot
ggarrange(f1, f2,ncol = 2, nrow = 1)+
  ggsave("sidebyside boxplot.png",height=3,width=7)


#identify studio city ouliers
which.max(cl_apartments$sqft>3000 & cl_apartments$bedrooms==0)
cl_apartments[2664,]
cl_apartments$sqft[2664]
which.max(cl_apartments$sqft>3200 & cl_apartments$bedrooms==3)
cl_apartments[18086,]
##############################################################################################
####################
#plots for num of pets vs city and suburbs
ggplot(cl_apartments[cl_apartments$city %in% suburbs,])+
```

```
  geom_bar(aes(x=pets, fill=city),position="dodge")+
  scale_color_viridis()+
  theme_bw()
# do the same for city

ggplot(cl_apartments[cl_apartments$city %in% major_10_cities,])+
  geom_bar(aes(x=pets, fill=city),position="dodge")+
  scale_color_viridis()+
  theme_bw()+
  ggtitle("Pet Policy in Major Cities")+
  xlab("Pets Allowed")+
  ylab("Number of apartments that allow pet")
##########################################################################################
##################################

##########################################################################################
####################################
#which adds more to rent bed or bath?

#cl_apartments[!is.na(cl_apartments$bathrooms),]

#all places
ggplot(cl_apartments[!is.na(cl_apartments$bathrooms),])+
  geom_point(aes(x=factor(bedrooms),y=price, color=factor(bathrooms)),alpha=0.5)+
  ylab("Price ($)")+
  ylim(0,12500)+
  xlab("Number of Bedrooms")+
  ggtitle("Comparing Apartments: Price, Number of Beds and Bath")+
  labs(colour = "Number of Bathrooms") +
  scale_color_viridis(discrete = T)+
  theme_minimal()+
  ggsave("graph all bed bath price.png", height=4, width=7)

# ggplot(cl_apartments[!is.na(cl_apartments$bathrooms),])+
#   geom_point(aes(x=factor(bedrooms),y=price, color=factor(bathrooms)),alpha=0.5)+
#   ylab("Price ($)")+
#   ylim(0,12500)+
#   xlab("Number of Bedrooms")+
#   ggtitle("Comparing Apartments: Price, Number of Beds and Bath")+
#   labs(colour = "Number of Bathrooms") +
#   scale_color_viridis(discrete = T)+
#   theme_minimal()+
#   ggsave("graph all bed bath price.png", height=4, width=7)
```

```
#find the max price for a studio, and also average.
aggregate(price~bedrooms,data=cl_apartments[!is.na(cl_apartments$bathrooms),],mean)
which.max(cl_apartments$bedrooms==0 &cl_apartments$price>10000)
cl_apartments[5010,]
#find the max price for 2bed and also average.
which.max(cl_apartments$bedrooms==2 &cl_apartments$price>15000)
cl_apartments[8663,]
#find prices for 4 bed 3 bath 4 bed 2 bath
which(cl_apartments$bedrooms==4 & cl_apartments$price>15000 & cl_apartments$bathrooms==3)
cl_apartments[19520,]
which(cl_apartments$bedrooms==4 & cl_apartments$price>15000 & cl_apartments$bathrooms==4)
#could be duplicates
cl_apartments[12343,]
# #suburb
# rent_sub=ggplot(cl_apartments[cl_apartments$city %in% suburbs,])+
#   geom_point(aes(x=bedrooms,y=price, fill=factor(bathrooms)),alpha=0.5)+
#   ylab("Price ($)")+
#   xlab("Bedrooms")+
#   ggtitle("Suburban Apartments")+
#   labs(colour = "Number of Bathrooms") +
#   scale_color_viridis(discrete = T)+
#   theme_minimal()+
#   ggsave("graph suburb bed bath price.png", height=4, width=7)
# city
# rent_city=ggplot(cl_apartments[cl_apartments$city %in% major_10_cities,])+
# geom_point(aes(x=bedrooms,y=price, color=factor(bathrooms),alpha=0.5))+
# ylab("Price ($)")+
# xlab("Bedrooms")+
# ggtitle("Looking ar City Apartments")+
# labs(colour = "Number of Bathrooms") +
# theme(legend.position = "left")+
# scale_color_viridis(discrete = T)+
# theme_minimal()+
# ggsave("graph city bed bath price.png", height=4, width=7)
#
#  ggarrange(rent_all, rent_all1, ncol = 2, nrow = 1)+
#    ggsave("sidebyside rent.png",height=4,width=9)

##############################################################################################
#########################

# plotting barplot for #bedrooms and 10 citites in CA
```

```
#ggplot(cl_apartments[cl_apartments$city %in% c("Sacramento","Mountain View", "San Francisco"),
])+
 # geom_bar(aes(x=bedrooms))+
 # facet_grid(city~.)
#SF has the most studios and all of them have lots of

#ggplot(cl_apartments[cl_apartments$city %in% suburbs, ])+
  #geom_bar(aes(x=bedrooms))+
  #facet_wrap(city~.,ncol = 3)
#Some code
#Make a 'big 10' dataframe
#cl_big10<-cl_apartments[cl_apartments$city %in% major_10_cities, ]
#cl_big10$city<-factor(cl_big10$city,levels=major_10_cities)

#data.frame(City=c("San Francisco","San Francisco","San Francisco"),Suburb=c("Walnut Creek",))

##############################################################################################
#######################
# Family Friendly
#with help from http://www.cookbook-r.com/Graphs/Axes_(ggplot2)/ on removing tickmarks and
# also x axis labels

#family friendly  variable
cl_apartments$family_friendly =(cl_apartments$bedrooms>1 &
                          cl_apartments$pets!='none' &
                          cl_apartments$parking!="none" &
                          !(is.na(cl_apartments$parking) | is.na(cl_apartments$pets)))
#plots
cl_apartments$family_friendly = factor(cl_apartments$family_friendly, labels = c("no", "yes"))

ggplot(cl_apartments[cl_apartments$city %in% major_10_cities, ])+
  geom_bar(aes(x=family_friendly,fill=family_friendly))+
  facet_wrap(city~.)+
  #scale_fill_viridis()+
  theme_bw()+
  scale_fill_discrete(name="Family Frendly?",labels=c("no", "yes"))+
  ggtitle("Comparing City Family Friendly Apartments")+
  ylab("Number of Apartments")+
  theme(axis.title.x=element_blank(), axis.text.x=element_blank(),axis.ticks.x=element_blank() )+
  ggsave("family friend city city.png",height = 4,width=7)


ggplot(cl_apartments[cl_apartments$city %in% suburbs, ])+
```

```
  geom_bar(aes(x=family_friendly,fill=family_friendly))+
  facet_wrap(city~.)+
  theme_bw()+
  scale_fill_discrete(name="Family Frendly?",labels=c("no", "yes"))+
  scale_color_viridis()+
  ggtitle("Comparing Suburban Family Friendly Apartments")+
  ylab("Number of Apartments")+
  theme(axis.title.x=element_blank(), axis.text.x=element_blank(),axis.ticks.x=element_blank())+
  ggsave("family friend suburb.png",height = 7,width=14)

#to check that the graph shows the total # of appt just divided into family friendly.
sum(cl_apartments$city=="Sunnyvale",na.rm=T)

##############################################################################################
################

#How does square footage affect price?
#used http://www.cookbook-r.com/Graphs/Legends_(ggplot2)/ to figure out how to remove legends

 #use aggregate to plot by city
price_per_city=data.frame(aggregate(price~city,data=cl_apartments[cl_apartments$city %in%
major_10_cities,],
        median))
sqft_per_city= data.frame(aggregate(sqft~city,data=cl_apartments[cl_apartments$city %in%
major_10_cities,],
                    median))
avg_price_avg_sqft=data.frame(merge(price_per_city,sqft_per_city))
library(ggrepel)
ggplot(avg_price_avg_sqft)+
  geom_point(aes(x=sqft,y=price,color=city))+
  ggtitle("How Median Size and Median Price of Apartments Compare in Cities")+
  xlab("Median Size of Apartment (sqft)")+
  ylab("Median Price ($)")+
  guides(color=FALSE)+
  geom_label_repel(aes(x=sqft,y=price,label=city),angle = 45)+
  scale_color_viridis_d(option= "plasma")+
  theme_minimal()+
  ggsave("median price,sqft per city.png", height=4, width =9)
##############################################################################
#similarity vs difference btween WC Sac and Berkely

compare= c("Walnut Creek","Sacramento","Berkeley")
g1=ggplot(cl_apartments[cl_apartments$city %in% compare & !is.na(cl_apartments$laundry), ])+
```

```
  #geom_boxplot(aes(y=sqft, x=city))+
  geom_bar(aes(x=laundry,fill=city),position="dodge")+
  #facet_grid(family_friendly~.)+
  theme_light()+
  labs(title="Comparing Size of Apartments in specified Cities")+
  ggtitle('a. Laundry Facilities ')+
  xlab("Type of Laundry Facility")+
  ylab("Number of Apartments")+
  scale_fill_viridis_d()+
  ggsave("similarGeographic area.png",height=2,width=5)


g2=ggplot(cl_apartments[cl_apartments$city %in% compare, ])+
  geom_point(aes(x=sqft,y=price, color=city),alpha=0.7)+
  labs(title="Comparing Price and Size of Apartments")+
  ggtitle('b. Sqft vs Price')+
  xlab("Size of Apartment (sqft)")+
  ylab("Price ($)")+
  scale_color_viridis_d()+
  theme_light()+
  ggsave("sqft $ similarGeographic area.png",height=3,width=5)

# which.max(cl_apartments$city=="Sacramento" & cl_apartments$price>9000)
# cl_apartments["5010",]
# cl_apartments$price[5010]


g3=ggplot(cl_apartments[cl_apartments$city %in% compare, ])+
  geom_bar(aes(x=bedrooms,fill=city),position="dodge")+
  labs(title="Comparing Price and Size of Apartments")+
  ggtitle("c. Floorplan")+
  xlab("Number of bedrooms")+
  ylab("Number of Apartments")+
  scale_fill_viridis_d()+
  theme_light()+
  ggsave("count bed similarGeographic area.png",height=3,width=5)


g4=ggplot(cl_apartments[cl_apartments$city %in% compare & !is.na(cl_apartments$parking), ])+
  geom_bar(aes(x=parking,fill=city),position="dodge")+
  labs(title="Comparing Parking")+
  ggtitle("d. Parking Availabilty")+
  xlab("Type of Parking")+
```

```
  ylab("Number of Apartments")+
  scale_fill_viridis_d()+
  theme_light()+
  theme(axis.text.x = element_text(angle = 60))
  ggsave("count parking similarGeographic area.png",height=3,width=5)

ggarrange(g1, g2, g3,g4, ncol = 2, nrow = 2)+
  ggsave("sidebyside geograph.png",height=4,width=8)


which(cl_apartments$price>9000 & cl_apartments$city=="Sacramento")
cl_apartments[5010,]

#range(cl_apartments$date_posted)
#table(cl_apartments$city=="Berkely" & cl_apartments$deleted)
##############################################################################

#Parking and Price city vs suburb

library(ggplot2)
p1=ggplot(cl_apartments[cl_apartments$city %in% major_10_cities & !is.na(cl_apartments$parking),])+
  geom_boxplot(aes(x=parking, y=price))+
  scale_color_viridis(discrete=T)+
  theme_minimal()+
  ggtitle("a. City Parking Availability & Price")+
  xlab("Type of Parking Offered")+
  ylab("Price ($)")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  ggsave("parking and price.png", height=3, width=7.5)


p2=ggplot(cl_apartments[cl_apartments$city %in% suburbs & !is.na(cl_apartments$parking),])+
    geom_boxplot(aes(x=parking, y=price))+
    scale_color_viridis(discrete=T)+
    theme_minimal()+
    ggtitle("b. Suburb Parking Availability & Price")+
    xlab("Type of Parking Offered")+
    ylab("Price ($)")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
  ggsave("parking and price sub.png", height=3, width=7.5)

ggarrange(p1, p2,ncol = 2, nrow = 1)+
  ggsave("sidebyside parking.png",height=4,width=8)
```

```
##############################################################################
# pet policy group city vs suburb

  cl_apartments$city_or_sub = NA
  cl_apartments$city_or_sub[cl_apartments$city %in% suburbs] = "suburbs"
  cl_apartments$city_or_sub[cl_apartments$city %in% major_10_cities] = "city"
  cl_apartments$city_or_sub = factor(cl_apartments$city_or_sub)

  ggplot(cl_apartments[!is.na(cl_apartments$city_or_sub), ],aes(x=pets, fill=city_or_sub))+
    geom_bar(position="dodge", stat = "count", width=0.5)+
    ggtitle("Pet Policy: City vs Suburb")+
    geom_text(aes(label=..count..),stat="count",position=position_dodge(0.9),vjust=-0.4)+
    xlab("Type of Pets Allowed")+
    ylab("Number of Apartments")+
    guides(fill=guide_legend(title=NULL))+
    scale_fill_viridis_d(option=c("cividis")) +
    theme_minimal()+
    ggsave("pet policy city vs suburb.png", height=4,width=7)

##############################################################################
#SF vs chula vista appt

  SF_vs_chula= c("San Francisco","Chula Vista")
 sf_chul1= ggplot(cl_apartments[cl_apartments$city %in%  SF_vs_danville , ])+
    geom_boxplot(aes(x=city, y=price))+
    ggtitle('Comparing Price of Apartments')+
    xlab("Name of City")+
    ylab("Price of Apartment ($)")+
    theme_minimal()+
    ggsave("sF vs danv $.png",height=4,width=7)

  #make boxplots with # beds

 sf_chul2=ggplot(cl_apartments[cl_apartments$city %in% SF_vs_danville &
!is.na(cl_apartments$bedrooms), ])+
    geom_bar(aes(x=factor(bedrooms),fill=city),position="dodge")+
    scale_fill_viridis_d() +
    ggtitle("Comparing Floorplans")+
    xlab("Number of Bedrooms")+
    ylab("Number of Apartments")+
    theme_minimal()+
```

```
   ggsave("sF vs chula $.png",height=4,width=7)


  ggarrange(sf_chul1,sf_chul2,ncol = 2, nrow = 1)+
    ggsave("sidebyside sf_beds.png",height=4,width=10)
##############################################################################################
# also used these
#http://www.sthda.com/english/articles/32-r-graphics-essentials/132-plot-grouped-data-box-plot-bar-plot-and-more/
#https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html
  #http://www.sthda.com/english/wiki/ggplot2-texts-add-text-annotations-to-a-graph-in-r-software
```