# Sta 141a Assignment 3

Rohan Malhotra
Statistics Department
University of California, Davis
October 2018

Craigslist is a website in which users can advertise their products for free. The products can be almost anything, and today is used across 570 cities around the world. The posts are divided by categories which helps users narrow down what exactly they are looking for. This data set consists of apartment rentals in California. The audience for these apartment postings can range anywhere from college students all the way to adults looking for a home. This data was messy, which required cleaning. This report will only refer to the cleaned data, but to learn more about the raw data and what was removed please refer to section 5. The analysis was done using R and the GGplot package.

## 1 Explanatory Data Analysis

### 1.1 Rows and Columns

There are 12,664 rows meaning there are 12,664 posts within the data set. Since it was cleaned there is no duplicate posts included in the data set. Additionally, there are 20 columns within the data set as well, meaning for each apartment advertisement there are 20 variables/ pieces of information that was given. These columns include information 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. All these variables help find patterns and trends within the data set.

### 1.2 Missing features

The variables within the data that have some missing features are: date_updated, pets, city, garage, city_text, laundry, parking, and place. The variables that are missing include make sense as some of them might have just been forgotten in the post or were repeated. For instance, the city variables have some missing features because that information was included in the place variable and vice versa.

### 1.3 Span of the data set

The data set consists of apartment listings posted from September 9$^{th}$, 2018 to October 15, 2018. Furthermore, these listings were updated between October 8$^{th}$ and October 15$^{th}$. Next, looking at the range of cities in the data set. The data set spans from 39 degrees latitude and -123 degrees longitude to 32 degrees latitude and -116 degrees longitude. In other words, the data set includes some, but not all cities from Ukiah all the way to San Diego.

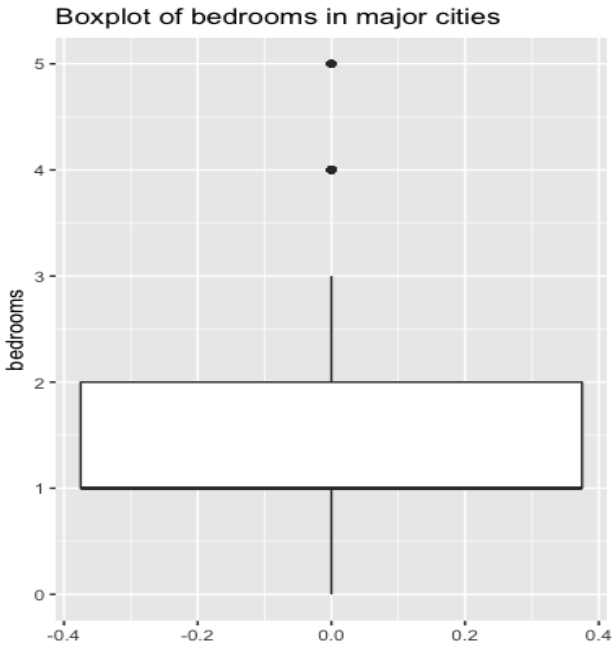
## 2 Apartments in suburbs vs. in major cities

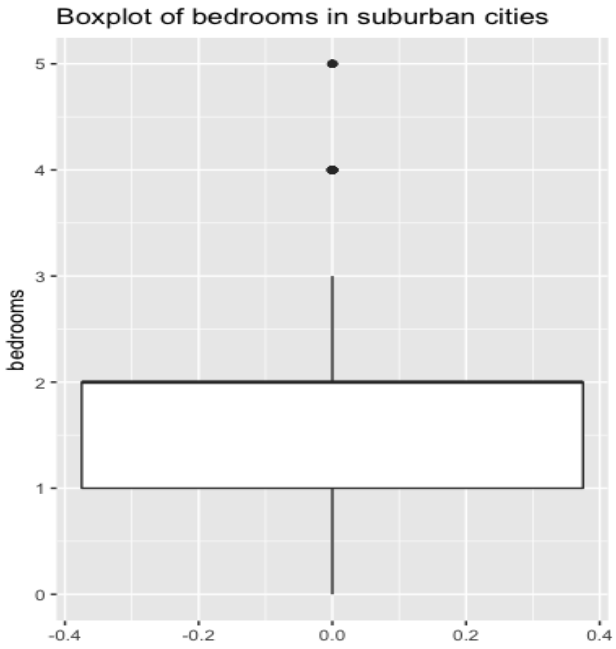
### 2.1 Definitions of suburbs and major cities

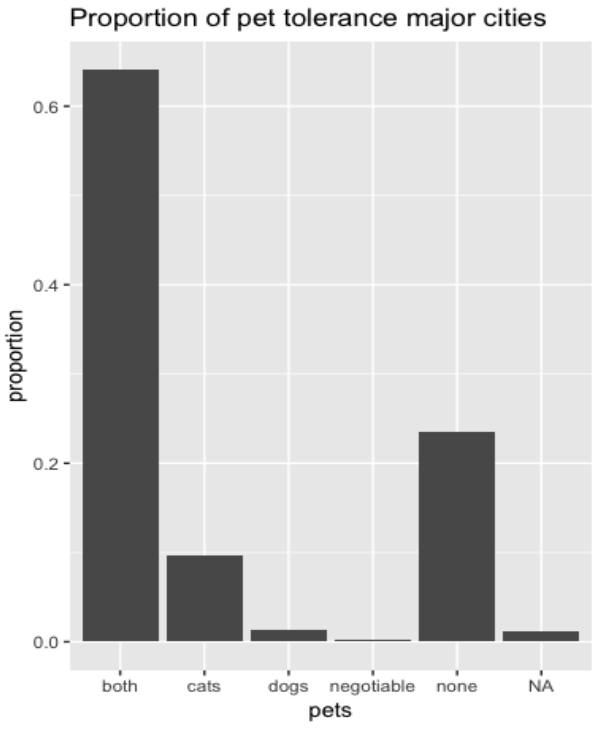
One aspect I was interested in was in looking how apartments in suburbs differ from apartments in major cities. To do this I had to split the data set into the two categories. I defined major cities by looking at the top five cities with the most apartment postings. These cities came out to be Sacramento, Los Angeles, San Jose, San Diego and San Francisco. Suburban cities were all the cities in the data set excluding the major cities listed above.

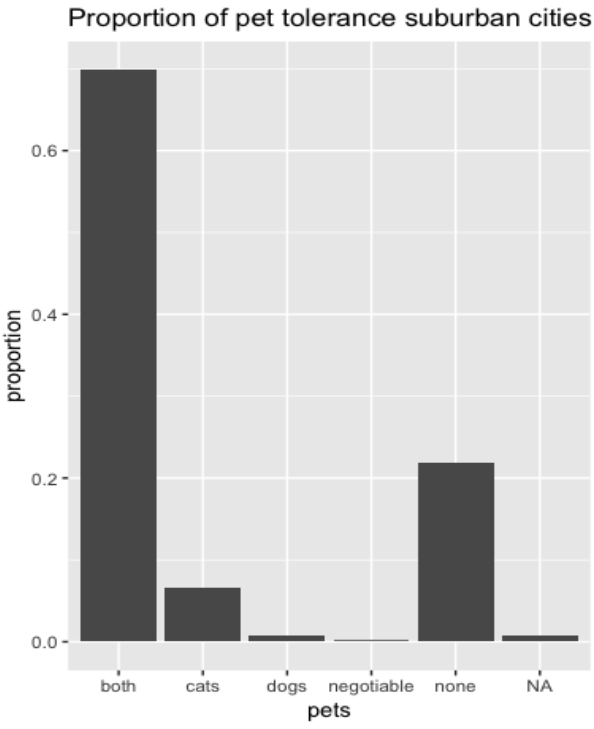
### 2.2 Which is more family friendly?

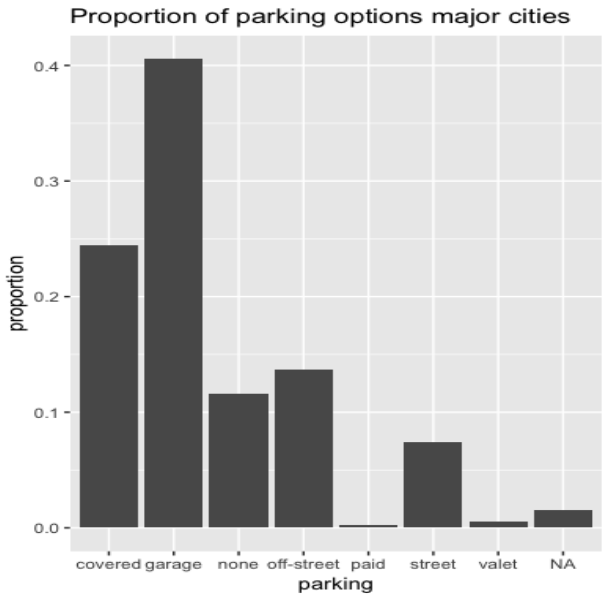
One major thing to look at is for a family looking for an apartment, is there any difference between looking at apartments in major cities and suburban cities. My hypothesis is that I believe apartments tend to try to accommodate to all family needs so there would not be that big of a difference. To analyze this I looked at variables such as bedrooms, pet policy and parking options for both suburban cities and major cities. This data can be summarized within the graphs below.

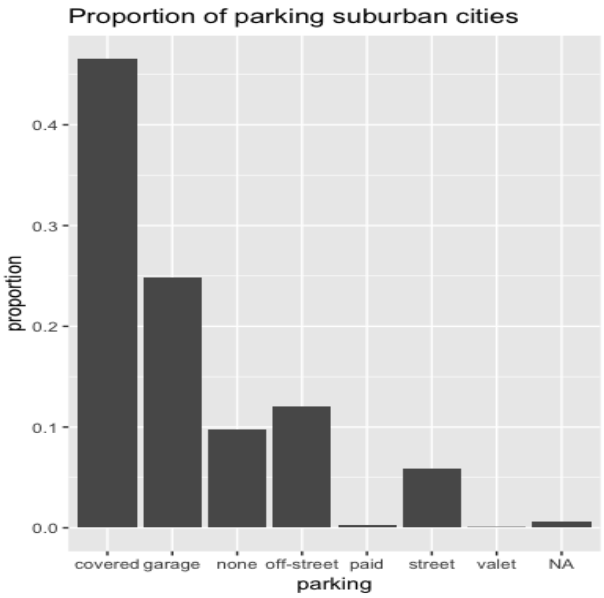




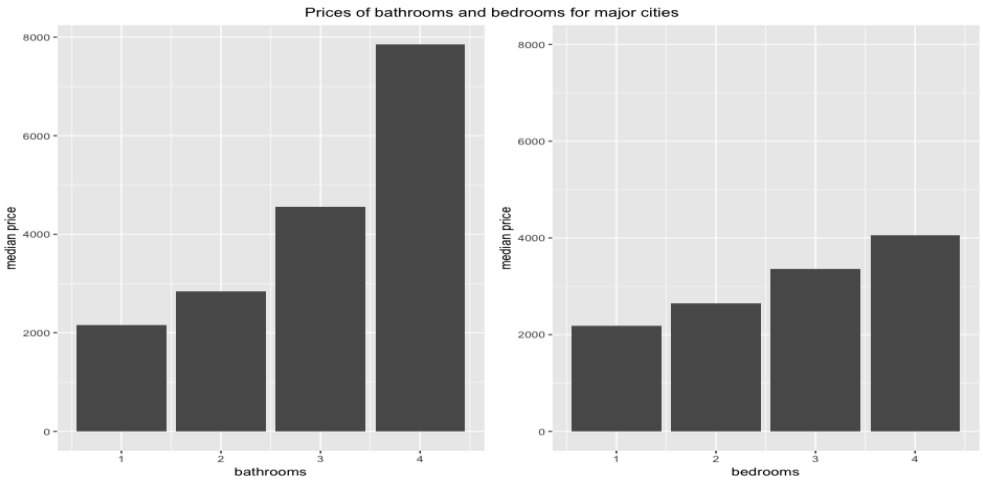



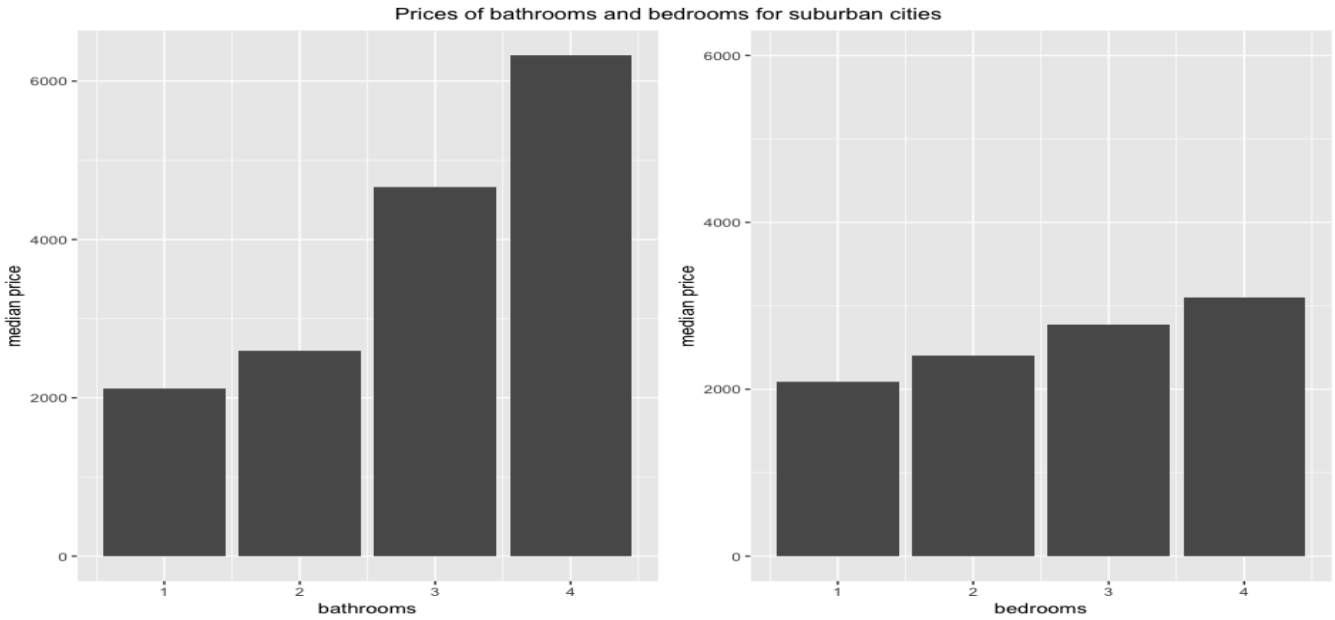

## 2.21 Conclusion

In conclusion, after looking at bedrooms, pet policy, and parking options among the suburban and major cities, I found that though they do have some differences not one is more likely to be family friendly than the other. First, with bedrooms though the median is higher for suburban cities by one bedroom, but looking at the box plot they have IQR and whiskers, suggesting that though they have different medians they are not that much different from each other. Next, looking at the pet policy, both have similar proportions of policy, but suburban cities have a slightly more higher percentage of allowing both cats and dogs. Finally, with parking options major cities offer more of garage parking in contrast with suburban cities which offers covered parking. Both parking options are good. Therefore, A family can look at apartments either in suburban or major cities as they both pretty much a have an equal chance of being family friendly.

## 2.3 Bedrooms versus bathrooms

Price is what most people look at when deciding if the apartment is worth it or not. I wanted to determine which factor adds more to rent, bathrooms or bedrooms. To do this I looked at the median price of bedrooms and bathrooms for both suburban and major cities. By doing this I will also be able to get an overall trend. My hypothesis is that an increase in bedrooms will have more of an impact than an increase in bathrooms, as I feel bedrooms is more a necessity than bathrooms. To get a closer look let's look at the graphs below.






### 2.31 Conclusion

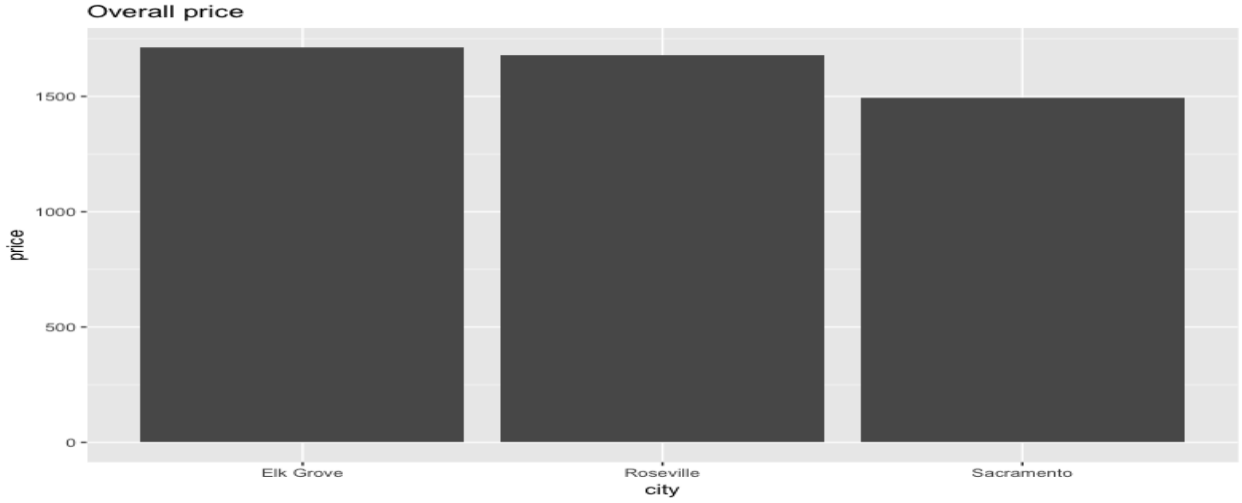
After looking at the graphs, it is apparent that my hypothesis was incorrect. First, looking at the price, the price when looking at apartments just based off the number of bathrooms is higher than the price when looking at apartments just based off the number of bedrooms. Looking at both major and suburban cities, increasing bathrooms increases price more than when increasing the number of bedrooms. Looking at bathrooms, both suburban cities and major cities have around the same amount of increase in price when increasing the number of bathrooms. Foe bedrooms, it is seen that the increase of price when increasing the number of bedrooms is slightly higher in major cities than suburban cities. Looking at both graphs, I can conclude the overall trend is that extra bedrooms add more to rent as the number of bedrooms increases.

## 3 Cities by region

### 3.1 Apartments in similar geographical areas

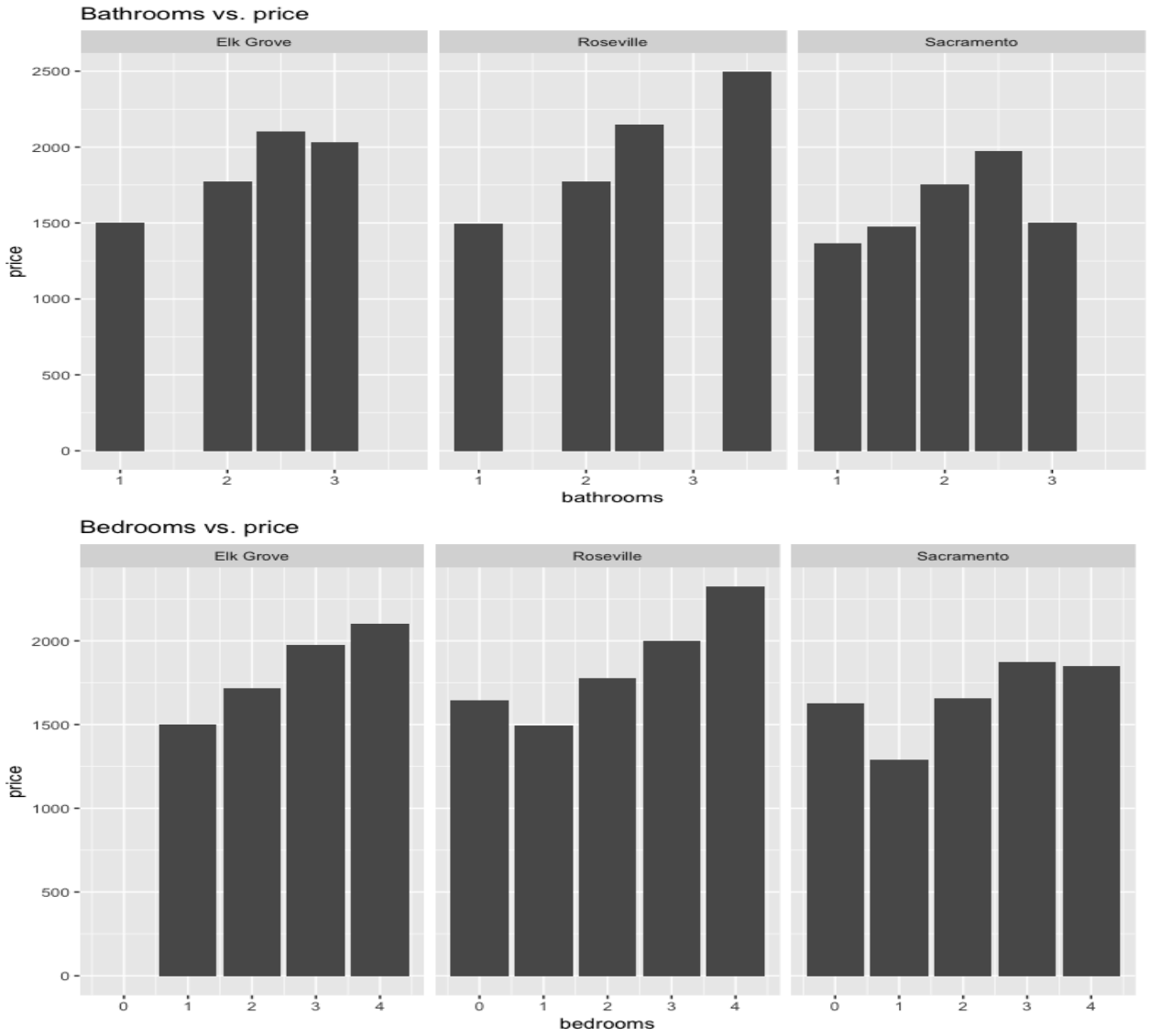
Though apartment listings are separated by city I was interested in finding out if there is a pattern or relation among cities in similar geographical regions. To do this I picked compared variables among Sacramento, Elk grove and Roseville. I looked at the overall median price for apartments in each city, median of prices for bedrooms and bathrooms for each city, the sqft for each city and finally the parking options.

### 3.11 Overall prices



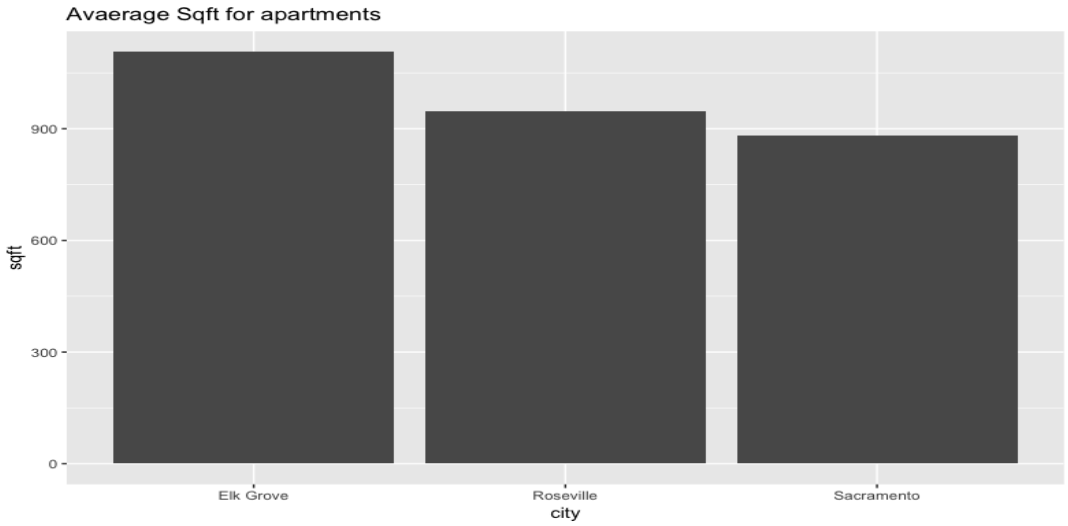
Looking at the overall prices by cities, it is seen that Roseville and Elk Grove have roughly the same median price whereas, Sacramento is slightly cheaper. Being from the Sacramento area, I know that areas such as Roseville and Elk Grove have a higher standard so it makes sense why their price is higher than Sacramento's.

### 3.12 Prices by bedroom and bathrooms




Looking at bathrooms, there is a common trend among all the cities which is that as bathroom increases so does the price. Also, all three apartments have a similar price when looking at apartments purely based on bathrooms. Again, it is seen Sacramento has slightly lower prices which was concluded from part 3.11. Next, bedrooms in Elk Grove and Roseville there is a clear increase in price as bedrooms increase. In Sacramento, though the price increases by adding bedrooms it is a minimal increase suggesting, increasing the number of rooms in Sacramento won't have as a big of an impact on price in contrast to Roseville and Elk Grove.

### 3.13 Sqft




Finally, looking at the average sqft for apartments in each city it is seen that Elk Grove tend to have the biggest apartments in terms of size. As seen from previous comparisons, Sacramento is ranked third and well below both Roseville and Sacramento.

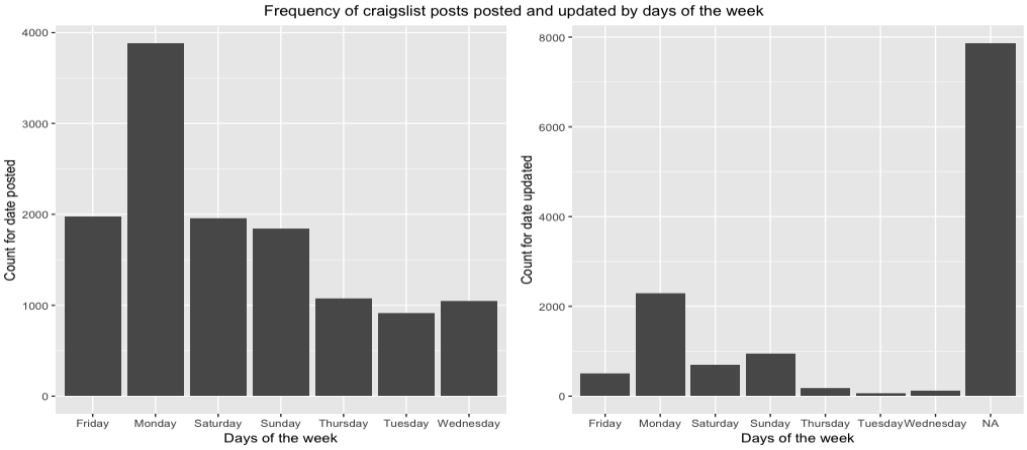
### 3.14 Conclusion

After looking at all these quantitative variables among Sacramento, Roseville, and Elk Grove I could conclude three things. One, since Roseville and Elk Grove tend to have a higher standard of living, by this I mean that those cities are more of a rich area to live in, they will have higher prices and bigger apartments than Sacramento. Next, for families looking to decide between the three the most affordable place to go to is Sacramento, but if one wants an apartment with more space they would want to look for apartments in Roseville or Elk Grove. Finally, I feel geographical regions have some effect on apartments, but it mainly depends on the standard of the city as Elk Grove and Roseville compared to each other more because they have similar standards.

## 4 Other trends noted

### 4.1 Best day to check craigslist to get first dibs on an apartment

One thing I was interested in is to see what day is the best day to check Craigslist to get first dibs on an apartment. I think the best day would probably be Sunday, as it is a weekend where most people get their errands done. The graph bellows how many posts are posted and updated by day.

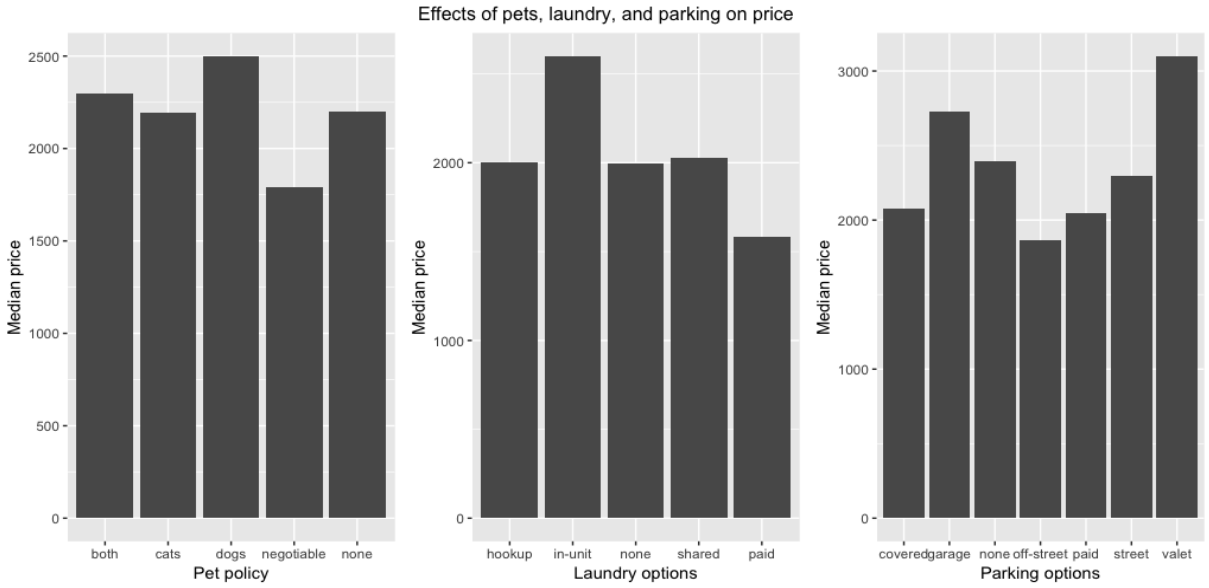


### 4.12 Conclusion

From the graph above it is seen most Craigslist posts are posted and updated on Mondays which is against my hypothesis. Therefore, for people in search of apartments the best day to get first dibs on an apartment is to check the website on Mondays.

### 4.2 Do qualitative features affect the price

Do features such as pet policies, laundry options, and parking options have a big effect on price. To see this, the graph below shows the median price for all these features separated by their respective categories. My hypothesis is that pet policies and laundry options won't affect price and if they do, it will be a minimal amount and parking will have a big impact on price.

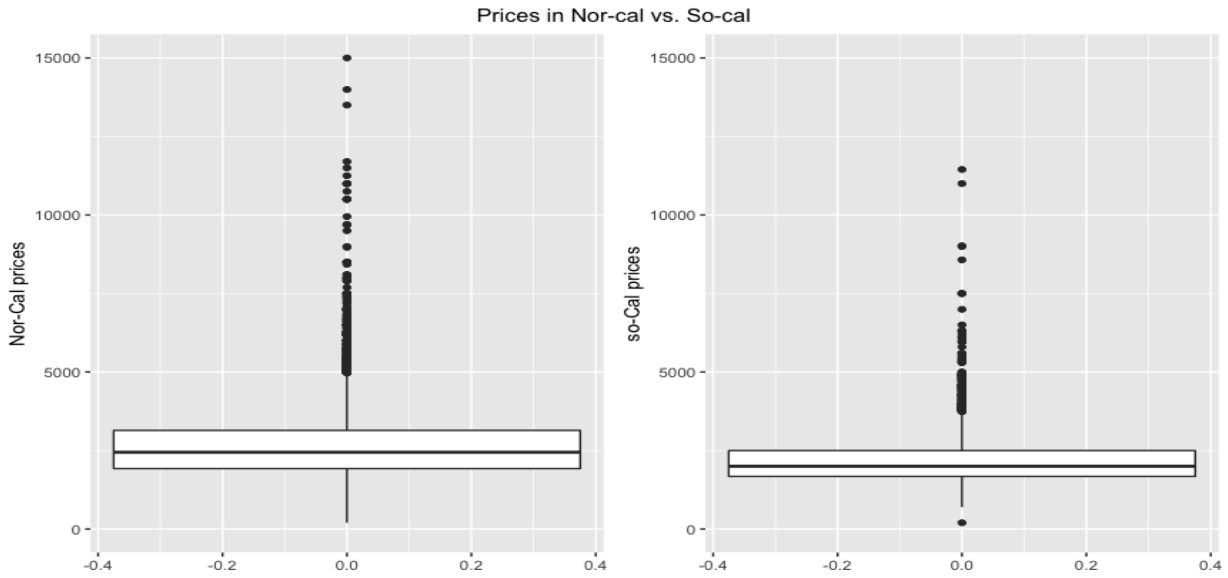



### 4.21 Conclusion

First, looking at pets, the prices among the different categories differ from each other, but not by too much. The only one category that has a drastic difference is the negotiable category, but the prices for being able to have both pets versus the prices for having none pets is slightly higher, therefore this is matches my hypothesis. Therefore, I conclude pet policies slightly affect the price for apartments. Next, looking at the laundry options, it seems that prices are the same among categories, but for in-unit laundry the apartments costs way more. Therefore, I believe that in-unit laundry increases the price, but another category does not influence price as much, this conclusion contradicts my hypothesis. Finally, looking at the garage options all options vary, but it is seen for valet parking people must pay a premium on their apartment rentals. Therefore, parking options do influence the price of an apartment rental.

### 4.3 Northern California vs. Southern California

Do the prices for apartment rentals in Northern California differ from apartment rentals in Southern California? My hypothesis is that cities in Southern California will probably be more expensive to live in because it is more populated, and has a lot of attractions.

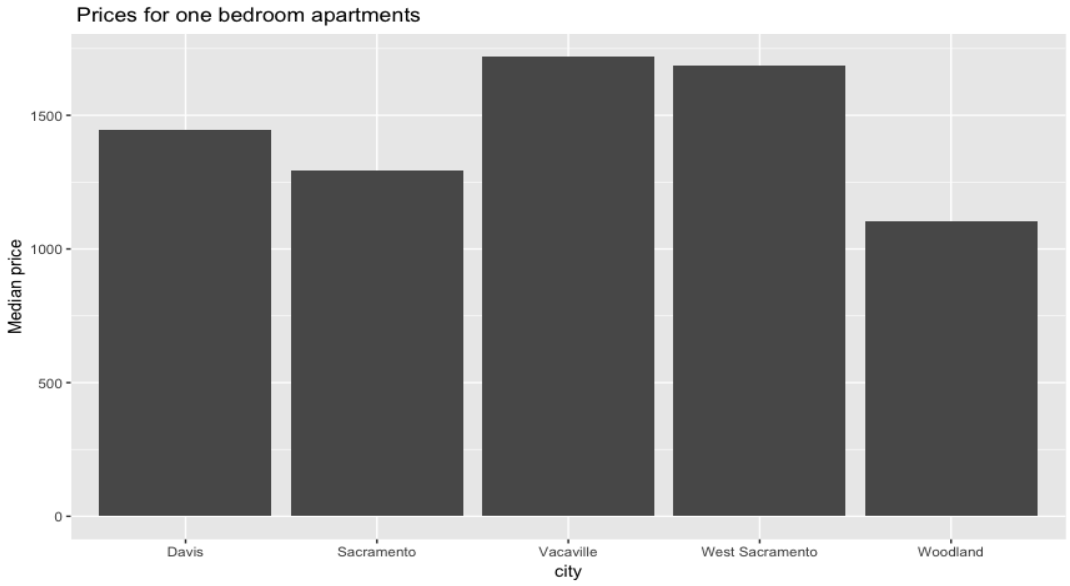



### 4.31 Conclusion

From the boxplots above it is seen that apartment rentals in Northern California are higher than apartments in Southern California. This contradicts my hypothesis, but I believe that this is due to Southern California being so populated that many people may be migrating to Northern California.

### 4.4 Commuting UC Davis Students

Students who transfer to UC Davis tend to commute. I am one of those students therefore, I was interested in knowing if I commute where is the cheapest place for me to find a one bedroom apartment that is still a good commutable distance. To examine this, I looked at the rent for one bedroom apartments in Sacramento, West Sacramento, Davis, Woodland, and Vacaville. My hypothesis is that Sacramento would probably be the cheapest place to find a one bedroom. Below shows the graph of median prices for these cities.

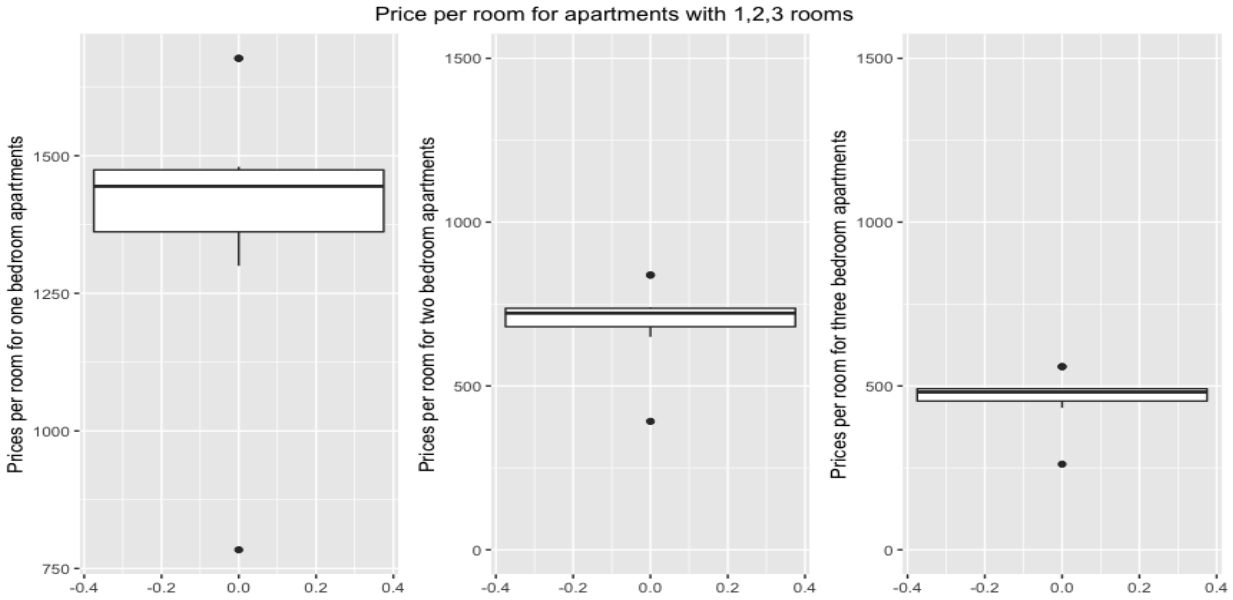



### 4.41 Conclusion

From the graph above it is seen that Woodland is the cheapest place to find a one bedroom apartment. This contradicts my theory of the city being Sacramento. This is great news for students as Woodland is not too far away from Davis and cost of living in Woodland is significantly cheaper than Davis.

### 4.5 Housemates or no housemates

The final question I examined is it is cheaper for students to live with housemates or not. To examine this, I looked at the median price for apartments with 1,2,and 3 bedrooms and then could get the price per room for each one of those apartments. I believe having housemates would not make a difference as since price goes when bedrooms go up, a person will probably end up paying the same for as they would for a one bedroom apartment. The graph below shows median price for a room for apartments with 1,2, and 3 bedrooms.




### 4.51 Conclusion

In conclusion, it is obviously seen above, though there are some outliers, overall a price for a room in a 3 bedroom apartment is significantly lower than a room in one bedroom apartment. Therefore, this contradicts my hypothesis and to get the best price for a room a student should try to find housemates.

## 5 Conclusion

In Conclusion, this data set can answer many questions on apartment rentals in California. I could use this data set to find some trends and patterns I would not have known or guessed. Though I did answer a lot of questions within the data, other questions that can be answered for people who also decide to explore the data include:

- Is there a linear relation among price and sqft? (This question is helpful to determine if square feet does have an effect on price, and will be useful information for those looking for apartments.)
- When people look for apartments does sqft increase price more or does bedrooms? (This question can help to see what increases price more bedrooms or more overall space, this will help users to sort out apartments, whether they want apartments that's more spacious, or if they have a lot of members and need more rooms)
- Are coastal cities more expensive to live in than inland cities? (This is just a question to determine how geographical location may influence price and will help people migrating to California decide if they should live towards coastal cities or not)

- Is there some sort of relation among population and price of apartments? (This is an interesting question to help sort out which type of cities are probable to have a lower rent)
- Are certain regions in California are likely to be more family friendly than others? (This question will help separate to see if some regions are primarily for families or are they mainly for single people, this can help families narrow down which region of California to look at apartment rentals in)

## 6 Raw data

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.

### 6.1 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.

### 6.2 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.

### 6.3 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.

### 6.4 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. Also, I had to look at specific parts of cities for example, in 4.5 I only looked at apartments with 1,2,or 3 bedrooms because Woodland did not have any listings with 4 apartments. I feel like this situation could have been resolved if the span of the dates of the posts was bigger so I could have more observations to analyze. All in all, through these limitations I tried to work around as many limitations as I can to get the best analysis possible

# CODE:

```r
# Assignment 3

# Setting up -------------------
craigslist = readRDS("cl_apartments.rds")
library(ggplot2)
library(ggrepel)
library(ggridges)
library(gridExtra)



# 1 and 5: Raw Data Features/analysis ----------------

# Update the data
craigslist$price[craigslist$price == 34083742] = 3408
craigslist$price[craigslist$price == 9951095 ] = 995

dim(craigslist_raw)
craigslist_raw = readRDS("cl_apartments.rds")
craigslist_raw$price[craigslist$price == 34083742] = 3408
craigslist_raw$price[craigslist$price == 9951095 ] = 995


#  Missing values
# Features with no missing values
craigslist_isna_raw = sapply(craigslist_raw, is.na)
total_isna_raw = apply(craigslist_isna_raw,2,sum)
no_na_raw = craigslist_raw[total_isna_raw == 0]
names(no_na_raw)

# Looking at the features that do have missing features
missing_features_raw = craigslist_raw[total_isna_raw == nrow(craigslist_raw)]
names(missing_features_raw)
sort(total_isna_raw)

# Exploring patterns
no_na_bathrooms = subset(craigslist_raw, !is.na(bedrooms) & !is.na(bathrooms))
nrow(craigslist_raw) - nrow(no_na_bathrooms) # Confirms theory that for all the posts that had na for bedrooms also had na for bathrooms

no_na_lat = subset(craigslist_raw, !is.na(latitude) & !is.na(longitude))
nrow(craigslist_raw) - nrow(no_na_lat) # Confirms theory that for all posts that had na for longitude also had na for latitude

no_na_state = subset(craigslist_raw, !is.na(state) & !is.na(county))
nrow(craigslist_raw) - nrow(no_na_state) # Confirms theory that for all posts that had na for state had it for county



# Outliers (Numerical Data)

# Bathrooms
table(craigslist_raw$bathrooms) # Does not make sense for it to have 0 bathrooms

# Bedrooms
table(craigslist_raw$bedrooms)  # There are 6 and 7 bedrooms, looking at those posts it was seen they were houses
alot_beds = subset(craigslist_raw, bedrooms == 6 |bedrooms == 7 )
alot_beds$title # Looking at the 6/7 bedrooms they are either services or houses, but even for 4/5 bedrooms i noticed some posts are houses not apartments
alot_beds$text

# Sq.ft
table(craigslist_raw$sqft) # Anything under 80 square feet is false/wrong data  as one cannot live under 80 sqft
sort(craigslist_raw$sqft, decreasing = TRUE) # 200000 Seems to lare and is an outlier
craigslist_raw[craigslist_raw$sqft > 9000,] # Looking at the post it is an obvious error

# Price
table(craigslist$price) # Many prices that do not make sense for apartments 0-150


low.price = subset(craigslist, price < 200)
low.price$title
# Looking at posts with price less than 200, I noticed that all of them are either posts for services(such as moving trucks)
# Or have a rate of paying weekly and not monthly, Or have a short term rent.



# Errors in the data set (Categorical values)

# States
table(craigslist$state) # Not all states are in California
```

```r
# 1 and 5: Raw Data Features/analysis
# Subsetting data to get most accurate data ---------------

# Final Subsetting
craigslist = subset(craigslist, price >= 200) # Accounts for price outliers/ textss that are not apartment related
craigslist = subset(craigslist, state == "CA") # Gives us data thats only in CA
craigslist = subset(craigslist, sqft >= 80 & sqft < 199999) # Gets rid of outlier sqft.
craigslist = subset(craigslist, bathrooms > 0) # Gets rid of apartments that did not mention bathroom
craigslist = subset(craigslist, bedrooms < 6)
# Removes duplicate texts and titles
craigslist = craigslist[!duplicated(craigslist$title),]
craigslist = craigslist[!duplicated(craigslist$text),]

# Overview of the updated dataset -----------

# Rows and Columns
dim(craigslist) # 12,664 observations and 20 variables
names(craigslist)

# Missing values
craigslist_isna = sapply(craigslist, is.na)
total_isna = apply(craigslist_isna,2,sum)

# Span of dates
head(sort(craigslist$date_posted)) # From September 9th
tail(sort(craigslist$date_posted)) # To October 15th

head(sort(craigslist$date_updated)) # From October 8th
tail(sort(craigslist$date_updated)) # To October 15th

# Span of Longitude/latitude
max(craigslist$latitude)
min(craigslist$latitude)

max(craigslist$longitude)
min(craigslist$longitude)
```

```r
# 2A -------------------------


# Major cities: defined by the cities with the most apartments
major_names = tail(sort(table(craigslist$city)), 5) # Major cities: Sac, LA, SJ, SD, SF

# REFRENCE TO PATRICK ON SHOWING THIS METHOD OF SUBSET

# Subset data with major cities
major_cities = craigslist[craigslist$city %in% names(major_names),]

# Subset data with Suburban cities
suburban_cities = craigslist[!(craigslist$city %in% names(major_names)),]

# Comparison between Suburban and Major cities, which is more family-friendly?

# Room comparison:

major_bed = ggplot(major_cities, aes(y = bedrooms)) + geom_boxplot() + labs(title = "Boxplot of bedrooms in major cities")

sub_bed = ggplot(suburban_cities, aes(y = bedrooms)) + geom_boxplot() + labs(title = "Boxplot of bedrooms in suburban cities")

grid.arrange(major_bed, sub_bed, ncol = 2)

mean(major_cities$bedrooms, na.rm = TRUE)
mean(suburban_cities$bedrooms, na.rm = TRUE)
# Overall not a big difference among the two

# Pets Comparison:

pets_major = ggplot(major_cities, aes(x = pets, y = stat(prop), group = 1)) + geom_bar() + labs(title = "Proportion of pet tolerance major cities", y = "proportion")
pets_sub = ggplot(suburban_cities, aes(x = pets, y = stat(prop), group = 1)) + geom_bar() + labs(title = "Proportion of pet tolerance suburban cities", y = "proportion")

grid.arrange(pets_major, pets_sub, ncol = 2)

# Parking Comparison:
parking_major = ggplot(major_cities, aes(x = parking, y = stat(prop), group = 1)) + geom_bar() + labs(title = "Proportion of parking options major cities", y = "proportion")
parking_sub = ggplot(suburban_cities, aes(x = parking, y = stat(prop), group = 1)) + geom_bar() + labs(title = "Proportion of parking suburban cities", y = "proportion")

grid.arrange(parking_major, parking_sub, ncol = 2)
```

```r
# 2B ----------------------

# Major city bathrooms vs. bedroom effect on price

# Subset bathrooms and bedroom so they have the same domains
major_city_bathroom = subset(major_cities, bathrooms == 1 | bathrooms == 2 | bathrooms == 3 | bathrooms == 4)
major_city_bedroom = subset(major_cities,  bedrooms == 1 | bedrooms == 2 | bedrooms == 3 | bedrooms == 4 )

# Plot bedrooms and bathrooms vs. price
p = ggplot(major_city_bathroom, aes(x = bathrooms, y = price)) + stat_summary(fun.y = median, geom = "bar") + ylab("median price")
q = ggplot(major_city_bedroom, aes(x = bedrooms, y = price)) + stat_summary(fun.y = median, geom = "bar") + ylab("median price") + ylim(0,8000)

grid.arrange(p,q,ncol = 2, top = "Prices of bathrooms and bedrooms for major cities")


# Suburban city bathrooms vs. bedroom effect on price

# Subset bathrooms and bedroom so they have the same domains
sub_city_bathroom = subset(suburban_cities, bathrooms == 1 | bathrooms == 2 | bathrooms == 3 | bathrooms == 4)
sub_city_bedroom = subset(suburban_cities,  bedrooms == 1 | bedrooms == 2 | bedrooms == 3 | bedrooms == 4 )

# Plot bedrooms and bathrooms vs. price
pp = ggplot(sub_city_bathroom, aes(x = bathrooms, y = price)) + stat_summary(fun.y = median, geom = "bar") + ylab("median price")
qq = ggplot(sub_city_bedroom, aes(x = bedrooms, y = price)) + stat_summary(fun.y = median, geom = "bar")  + ylab("median price") + ylim(0,6000)

grid.arrange(pp,qq,ncol = 2, top = "Prices of bathrooms and bedrooms for suburban cities")


# 2C ----------------------------

# Chose the sacramento area
my_cities = subset(craigslist, city == "Roseville" | city == "Sacramento" | city == "Elk Grove")


overall_price = ggplot(my_cities, aes(x = city, y = price)) + stat_summary(fun.y = median, geom = "bar")  + labs(title = "Overall price")
bed_vs_price = ggplot(my_cities, aes(x = bedrooms, y = price)) + stat_summary(fun.y = median, geom = "bar") + facet_grid(~ city) + labs(title = "Bedrooms vs. price")
bath_vs_price = ggplot(my_cities, aes(x = bathrooms, y = price)) + stat_summary(fun.y = median, geom = "bar") + facet_grid(~ city) + labs(title = "Bathrooms vs. price")
sqft_apt = ggplot(my_cities, aes(x = city, y = sqft)) + stat_summary(fun.y = mean, geom = "bar")  + labs(title = "Avaerage Sqft for apartments")

# Would be too much if i included this too.
pet_comparison = ggplot(my_cities, aes(x = pets, y = stat(prop), group = 1)) + geom_bar() + facet_grid(~ city) + labs(title = "Pets")
parking_comparison = ggplot(my_cities, aes(x = parking, y = stat(prop), group = 1)) + geom_bar() + facet_grid(~ city) + labs(title = "Parking")
laundry_comparison = ggplot(my_cities, aes(x = parking, y = stat(prop), group = 1)) + geom_bar() + facet_grid(~ city) + labs(title = "Laundry")
```

```r
#4A ------------------------

# For people, looking for apartments, which day is the best to check craigslist listings?
# Hypothesis: I would think it would be monday because its the start of the week?

# Get Plot of frequencies of posted and updated by days of the week

# REFERENCE PIAZZA POST 165 ON WEEKDDAYS()
updated = ggplot(craigslist, aes(x = weekdays(date_updated))) + geom_bar() + ylab("Count for date updated") + xlab("Days of the week")
posted = ggplot(craigslist, aes(x = weekdays(date_posted))) + geom_bar() + ylab("Count for date posted") + xlab("Days of the week")

grid.arrange(posted, updated, ncol = 2, top = "Frequency of craigslist posts posted and updated by days of the week")

#4B ----------------------

# Do qualitative features such as pets, laundry, parking have an affect on price?

# Get median price for each category of each variable
pets_price = aggregate(price ~ pets, craigslist, median)
laundry_price = aggregate(price ~ laundry, craigslist, median)
parking_price = aggregate(price ~ parking, craigslist, median)

p1 = ggplot(pets_price, aes(x = pets, y = price)) + geom_bar(stat = "identity") + xlab("Pet policy") + ylab("Median price")
p2 = ggplot(laundry_price, aes(x = laundry, y = price)) + geom_bar(stat = "identity") + xlab("Laundry options") + ylab("Median price")
p3 = ggplot(parking_price, aes(x = parking, y = price)) + geom_bar(stat = "identity") + xlab("Parking options") + ylab("Median price")
grid.arrange(p1, p2, p3, ncol = 3, top = "Effects of pets, laundry, and parking on price")


#4C ---------------------------

# Question: Is it more expensive to rent an apartment in Nor-cal or Southern California
# Define Nor cal as everything above Fresno and so cal everything below it, Fresno is that middle place
# Google search showed fresno has a 36.7 degree la

# Subset cities by Northern California and Southern California
Nor_cal = subset(craigslist, latitude > 36.7)
So_cal = subset(craigslist, latitude < 36.7)

Nor_cal_prices = ggplot(Nor_cal, aes(y = price)) + geom_boxplot() + ylab("Nor-Cal prices") + ylim(0,15000)
So_cal_prices = ggplot(So_cal, aes(y = price)) + geom_boxplot() + ylab("so-Cal prices") + ylim(0,15000)

grid.arrange(Nor_cal_prices, So_cal_prices, ncol = 2, top = "Prices in Nor-cal vs. So-cal")
```

```r
#4D ---------------------------

# Question: For UC Davis students who are able to commute, which city has the most affordable  apartments?
# Easy/Neigboring places for commuting: Sacramento,Davis, Woodland, West Sacramento and vacaville
# Looked at one bedroom apartments as most students have a roomates

# Subset cities close to davis, plot median price

neighbor_cities = subset(craigslist, city == "Sacramento" | city == "West Sacramento" | city == "Davis" | city == "Woodland" | city == "Vacaville")
one_bed = subset(neigbor_cities, bedrooms == 1)
one_bed_gg = ggplot(one_bed, aes(x = city, y = price)) + stat_summary(fun.y = median, geom = "bar")  + labs(title = " Prices for one bedroom apartments") + ylab("Median price") +
  labs(label = "2 bedroom apartments for each city vs. price")


#4E-------------
# Question: For students renting apartments in davis, does the amount of rooms in an apartment affect the price they will pay for their room only?
# Will look at 1-3 rooms as there is no information for davis apartments with 4 bedrooms

# Subset data by bedroom
davis_postings_1 = subset(craigslist, city == "Davis" & bedrooms == 1)
davis_postings_2 = subset(craigslist, city == "Davis" & bedrooms == 2)
davis_postings_3 = subset(craigslist, city == "Davis" & bedrooms == 3)


d1 = ggplot(davis_postings_1, aes(y = price)) + geom_boxplot() + ylab("Prices per room for one bedroom apartments")
d2 = ggplot(davis_postings_1, aes(y = price/2)) + geom_boxplot() + ylim(0, 1500) + ylab("Prices per room for two bedroom apartments")
d3 = ggplot(davis_postings_1, aes(y = price/3)) + geom_boxplot() + ylim(0, 1500) + ylab("Prices per room for three bedroom apartments")



grid.arrange(d1,d2,d3,ncol = 3, top = "Price per room for apartments with 1,2,3 rooms")
```