Report 3

Jie Gu
SID:913953707

Part I: Overview

Craigslist is a website that allows people to post classified advertisements for free. These posts span a variety of subjects, including job postings, housing rentals, and item sales. In this report, I will examine a data set based on recent Craigslist posts for apartment rentals in California.

In this data set, each observation is a single post of rental information, and there are 21948 observations. For each post, there are 20 features: 13 original features and 7 features that have been extracted from the text. Original features are title, text, latitude longitude, city_text, date_posted, date_updated, price, deleted, sqft, bedrooms, and bathrooms. Features that have been extracted from the text are pets, laundry, parking, craigslist, place, city, state and county.

The data set spans from 2018-09-08 17:09:33 to 2018-10-15 15:10:22. From the angle of latitude, the data set spans from 28.3112 to 47.68064; from the angle of longitude, the data set spans from -123.3406 to -72.92216. Since we only consider California, I delete posts with states that are not California. Then, in this data set, the southernmost city is San Diego; the northernmost city is Dunsmuir; the westernmost city is Ukiah; the easternmost city is Westmorland. In addition, since there are many posts with the same title or text and other features are highly similar, these posts can be considered as duplicated post. After deleting duplicated post, the three cities with the most posts are San Francisco, Los Angeles and San Diego; many cities like Vernon, Temple City and Truckee only have 1 post.

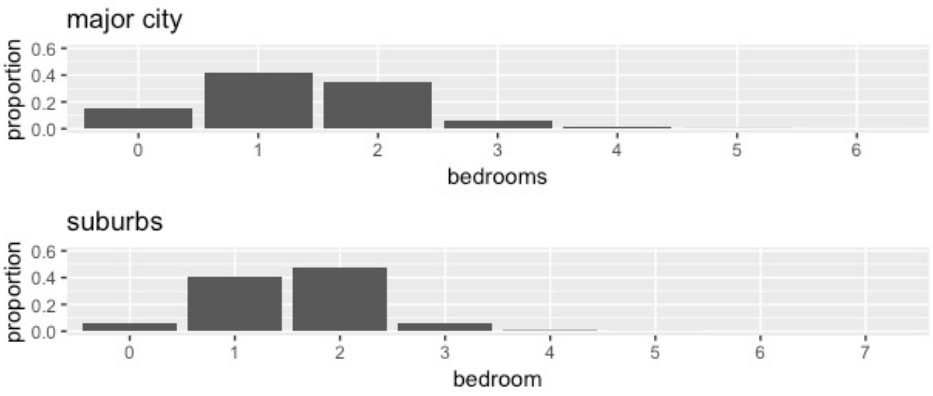
Part II: Solving some basic problems

Question1: Are apartments in suburbs more likely to be family-friendly than apartments in major cities?

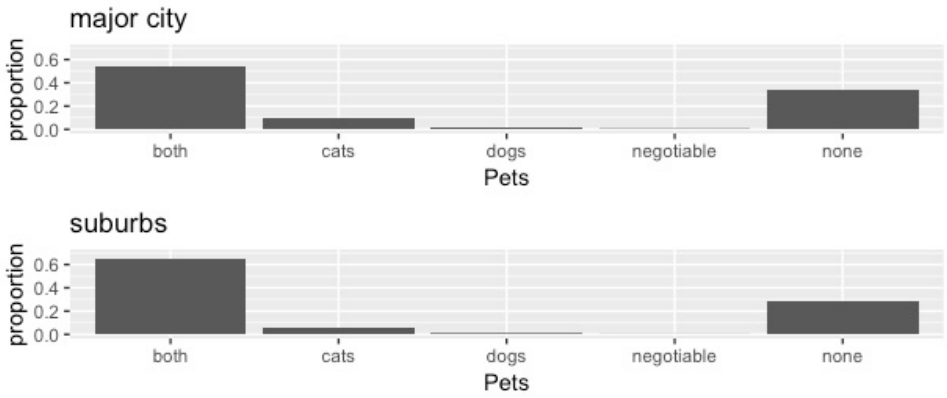
For this question, we should define suburbs and major cities at first. I define the 8 largest cities in CA ranked by population as major cities, and the rest cities as

suburbs. The following is a link to the population of each city: https://en.wikipedia.org/wiki/List_of_largest_California_cities_by_population

Now, let's compare the bar plot of the proportion of apartments with different number of bedrooms for major city and suburbs.



From the bar plot, we can see that the proportion of apartments with 2 or more than 2 bedrooms in suburbs is higher than that in major cities. In major cities, there are many studio apartments, but this proportion is really low in suburbs. Also, there some apartments with 7 bedrooms in suburbs, although the proportion is really low that we cannot see that in the bar plot. Hence, from the aspect of more bedrooms, apartments in suburbs are more family-friendly. Next, let take a look at the pet policy.

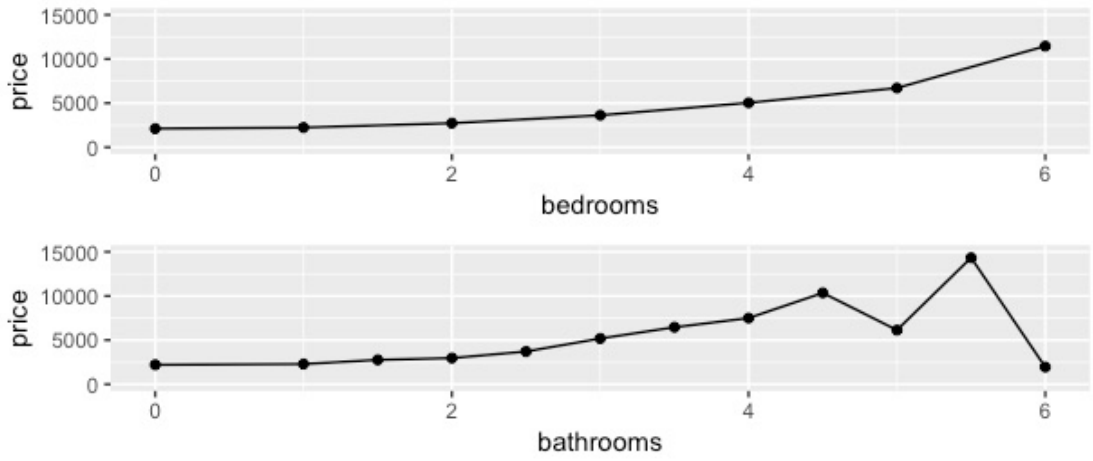


In this bar plot, we can see the proportion of apartments that accept both dogs and cats is higher in suburbs; the proportion of apartments that do not accept any pets is higher in major cities. Therefore, from the aspect of pet policy, apartments in suburbs are more family-friendly. Since the apartments in suburbs have more bedrooms and have a

better pet policy, we can conclude that the apartments in suburbs are more likely to be family-friendly. Now let's move to the next question.

Question2: Which adds more to rent: extra bedrooms, or extra bathrooms?

To solve this problem, we can compare the mean price of apartments with different number of bedrooms and bathrooms, but before that, we need to fix some problems on price. When we observe the price, there are some extremely large price: $9951095 and $34083742. After looking at the post with these two price, we can find that the price is not actually that high. From the text of the post with price $34083742, we can see the price range of apartment is $3408 ~ $3742. As for the post with price $9951095, it actually says that the price for apartments with 1 bedroom is $995 and the price for t apartments with 2 bedrooms is $1095. Therefore, I correct the price $9951095 as $995 and the price $34083724 as $3408. Additionally, there are many price smaller than $100. After looking at those posts, I find that most of them are repairing information and advertisements like improving bad credit. So I just delete them. Now, let's compare the mean price of apartments with different number of bedrooms and bathrooms with plot.



From this plot, we can see that before 5 bedrooms and 5 bathrooms, the mean price of apartments increases faster with the increase of bathroom than with the increase of bedrooms. Therefore, for the apartments with less than 5 bedrooms and 5 bathrooms, extra bathroom adds more to rent. However, the effect change as the number of

bedrooms or bathrooms goes up. More precisely, from 4.5 bathrooms to 5 bathrooms, the mean price decreases; from 5 bathrooms to 6 bathrooms, the mean price increases sharply and then decreases sharply. However, the mean price keeps increasing as the number of bedrooms goes up. Therefore, for the apartments with equal to or more than 5 bedrooms and 5 bathrooms, the situation is hard to tell. Next, let's see the next question.

Question3: Do apartments in similar geographical areas tend to be similar?

For this question, I choose San Francisco, San Jose and Oakland as three similar geographical areas, and I will compare the apartments in these cities from the aspect of the number of bedrooms, price and size. The following is a bar plot about the proportion of apartments with different number of bedrooms in these three cities.

From this bar plot, the similarity is that the proportions of apartments with 1 bedroom and 2 bedrooms are highest in all three cities; and the proportions of apartments with equal to or more than 4 bedrooms are quite low in all three cities. The difference is that the proportion of studio apartment is relatively high in San Francisco than in other two cities. Now before moving to the sqft and price, we should fix some problems in sqft.

When looking at the sqft, I find an extremely large value: 200000, and some extremely small values: 1 and 3. Perhaps these values are errors in data set, so I delete them. Then we can look at the dot plot of the relation between sqft and price in three

cities.

From the dot plot, the similarity in sqft is that the size of most apartments are from 0

square feet to 2500 square feet in all three cities; the difference is that there are also some apartments with a size more than 2500 in San Francisco, which almost do not exist in other two cities. There is also a difference in price: the price of most apartments in Oakland and San Jose varies from $0 to $5000; however, there are many apartments with a price higher than $5000 in San Francisco. Moreover, the 'slope' of relation in San Francisco is larger than that in other two cities. This means that the rent price per square feet of San Francisco is higher than that of other two cities.

After solving three basic questions, let's look at more questions that can be answered by this data set.

Part III: 10 More Questions can be answered

1. Will the parking influence the price of apartments? How does it influence that? This question is meaningful for people who have a car. People who have a requirement on parking (for example, someone want a garage.) can know whether they will pay more, and how much he should pay more or less.

2. Will the laundry influence the price of apartment? How does it influence that? For people who have a specific requirement on laundry, they can know how much they should pay more or less. For people who do not have a specific requirement on laundry, they can know which kind of apartments based on condition of laundry they should choose to save money.

3. Will the apartments with 0,1, 2, or 3 bedrooms be different on the laundry? and how?
This is meaningful for people who have a specific requirement on laundry and can choose living by themselves or living with other people. For example, a person wants a laundry in unit, then he can decide whether he should live with other people by overserving that apartments with how many bedrooms are more like to have that.

4.Does there exist area (by latitude and longitude) in Los Angeles that has many apartments with a relatively low price?
People who want to rent an apartment a low price in LA can look at the map and then search in the area having many apartments with a low price.

5. How does the price of apartments vary in north California and south California?
This question can help people decide whether they should go to north California or south California based on their requirements on the price of apartments.

6. Are apartments in San Francisco and Los Angeles different in rent market (from price and sqft)?
This question can help students who do not know whether they should study in LA or SF learn more information on the price and size of apartments so that they can make a better decision.

7. If a person wants to rent a studio apartment (0 bedrooms) or apartment with one bedroom, which area (by longitude and latitude) should he go?
For a person who are accustomed to live alone and will come to CA, he can know which part of CA has more studio apartments or 1-bedroom apartments. It helps him decide which part of CA he can go.

8. Will the apartments with more bedrooms allow pets more easily?
By solving this problem, people who want to have a pet can decide whether they should live alone (0 or 1 bedroom) or live with other people (2 or more bedrooms).
9.Do apartments have a stricter policy on dogs or cats?

It can help people decide whether they should adopt a dog or a cat based on the pet policy if they will rent an apartment.

10. For counties with more famous colleges (rank top 100), will the price be higher?

It is meaningful for students in famous colleges to learn more about the price of renting an apartment. Also, for students with a car, they can decide whether they should live in nearby county or the county their colleges locate at based on price.

Part IV: Solving some of these problems

In this part, I will solve some problems proposed in part III.

Q1. Will the parking influence the price of apartments? How does it influence that? we can use a bar plot to observe mean price for each parking policy.

Hypothesis: The mean price for off-street will be lowest and the mean price for valet will be highest.

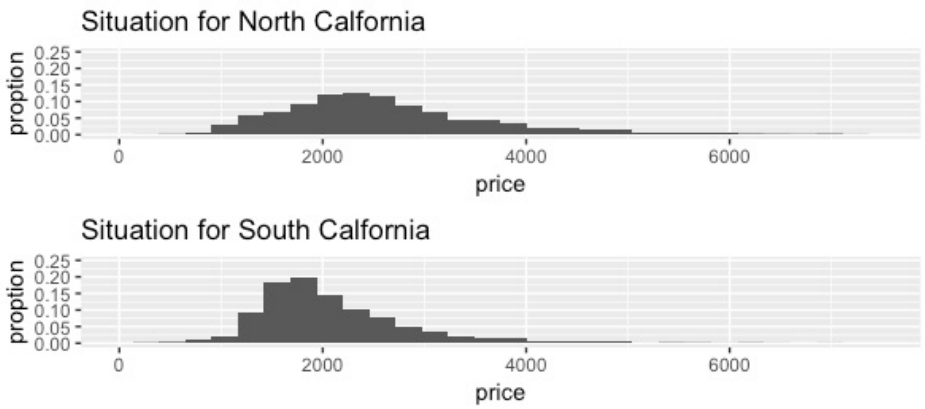
In this plot, covered and paid are different type of information on parking policy, so we ignore these two variables when comparing other variables. We can see the mean prices for off-street and street are the lowest two, and actually they are the same type. The mean price of garage is higher than off-street and street. Also, the mean price for valet parking is much higher than other types. The reason may be that usually only luxury apartments have valet parking. As for the 'none' variable, when I look at the posts, I find that 'none' means the text or title does not mention parking information, so we can neglect it here.

Q2. Will the apartments with 0,1, 2, or 3 bedrooms be different on laundry? and how?

Hypothesis: I guess apartments with more bedrooms are more likely to have a laundry in unit since the use ratio of in-unit laundry is higher for apartments with more bedrooms

In this bar plot, we only compare in-unit and shared laundry since paid is a different type of information, none means the posts do not mention, and hookup means the washing machine and dryer have been hooked up but we do not know whether they are in unit. We can see that in apartments with 0 and 1 bedrooms, the ratio of shared laundry is higher than that of in-unit laundry; however, for apartments with 2 or 3 bedrooms, the ratio of in-unit laundry is higher than that of shared laundry. Therefore, apartments with more bedrooms are more likely to have an in-unit laundry.

Q3. How does the price of apartments vary in north California and south California?
Hypothesis: I think the apartments with a higher price take more proportion in south California since Los Angeles has many posts and the price of apartments is really high there in common sense.
For this problem, we can use a histogram of price to see the proportion for each price interval.

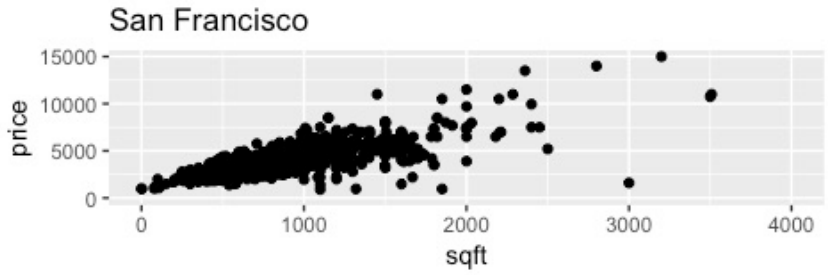


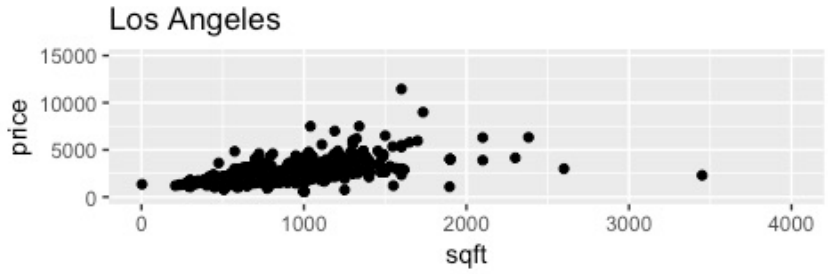
From this histogram, we can see that my hypothesis is wrong. The variance of price in

North California is larger than that in South California. More precisely, the apartments with the price lower than $2000 take much more proportion in South California compared with that in North California. Also, the proportion of apartments with the price higher than $3000 in North California is larger than that in South California. So from this data set, we can say North California has more proportion of apartments with a higher price compared with South California.

Q4. Are apartments in San Francisco and Los Angeles different in rent market (from price and sqft)?

Hypothesis: Los Angeles has more apartments with a higher price or larger size, also, the price per square feet is higher in LA than in SF.



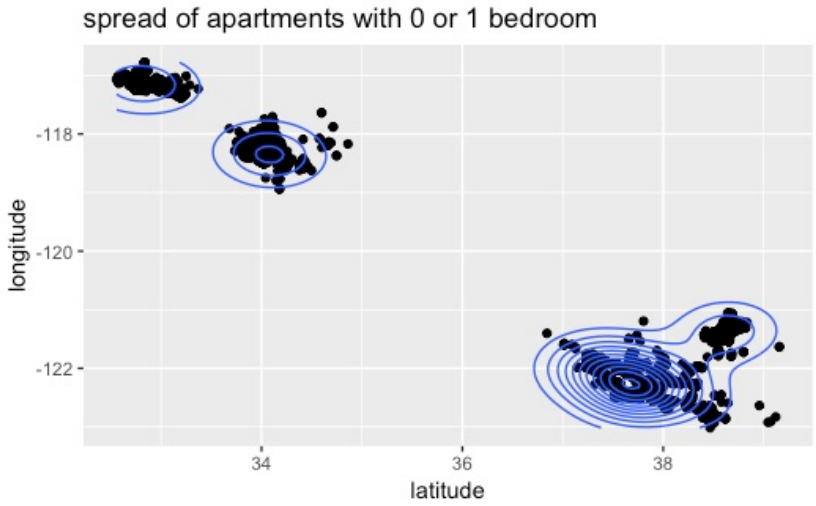


In this question, we can use a dot plot of sqft and price to show the trend.

The plot shows that the proportion of apartments with price over $5000 is higher in SF than in LA. As for the sqft, the spread of dots is similar. Although SF has more dots with sqft larger than 2000 than LA, we cannot conclude that SF has more apartments with sqft larger than 2000 than LA in general since SF has more posts than LA, and the proportion of dots with sqft larger than 2000 is not obvious in this plot. In addition, the ‘slope’ of relationship in SF is higher than that in LA, which means that the rent price per square feet is higher in SF than in LA.

Q5: If a person wants to rent a studio apartment (0 bedrooms) or apartment with one bedroom, which area (by longitude and latitude) should he go?

Hypothesis: The person should go to North California. Since there are many technical companies and entrepreneurships in Silicon Valley, many people will come to seek for a chance on their own. Therefore, there may be more studio apartments or 1-bedroom apartment in North California.



In this 2-density plot, we can see the area with latitude larger than 36, which is North California, has much more apartments with 0 or 1 bedroom than South California. More precisely, the areas with latitude larger than 37 and smaller than 38, and with longitude smaller than -121 has most apartments with 0 or 1 bedroom. These areas are in or near San Francisco, Santa Clara and Alameda counties, so the person can go to these counties.

Part V: Limitation of the data set

Finally, let's look at some limitation of the data set. Since any people can post information of apartment rentals in Craigslist, and Craigslist has few rules about how posts should be formatted, this data set is messy. For example, there are many advertisements like improving bad credit, and I delete them in my analysis. Also, there are many posts with wrong price like $34083742 as I mentioned in part II question 2, and I correct them with right price by looking at the text or delete them. Also, some prices are weekly instead of monthly. What's more, there are many outliers and anomalies in sqft like 1,3 and 200000 as I mentioned in part II question 3, and I delete them, otherwise these data will affect the graph for sqft. Additionally, for the features like laundry and parking, there exists information with different types----- like 'paid' and 'garage' in parking. So when I analyze information with the same type as 'garage', I neglect information like 'paid'. Moreover, there are some posts with states other than CA in this data set, and I delete them. Repetition is also a serious

problem as I mentioned in Part I. At last, there 26316 missing values in this data set. Features like craigslist, deleted, title, text and data_posted only have 0 or 1 missing values, but sqft has many missing values (I do not consider data_updated here since the NA in that means not updated). More importantly, some conclusions like that in part IV question 5 can only apply to this data set since in general, the number of posts in different areas is different and the sample size is not large enough.

Appendix

```
library(ggplot2)
library(ggridges)
library(gridExtra)
library(ggmap)
cl = readRDS('cl_apartments.rds')

#Q1:Big picture of data-------------------------


names(cl)
dim(cl)

#check the data it spans
sort(cl$date_posted)
sort(cl$date_posted,decreasing = TRUE)
sort(cl$date_updated,decreasing = TRUE)

#check the place it spans
min(cl$latitude,na.rm = TRUE)
max(cl$latitude,na.rm = TRUE)

min(cl$longitude,na.rm = TRUE)
max(cl$longitude,na.rm = TRUE)


#remove non-CA date
cl_ca <- subset(cl,state == 'CA')
#check the place it spans
cl_ca[cl_ca$latitude == min(cl_ca$latitude,na.rm =
TRUE),]
```

```
cl_ca[cl_ca$latitude == max(cl_ca$latitude,na.rm =
TRUE),]

cl_ca[cl_ca$longitude == min(cl_ca$longitude,na.rm =
TRUE),]
cl_ca[cl_ca$longitude == max(cl_ca$longitude,na.rm =
TRUE),]


#remove the duplicated data.
cl_no_duplicate = cl_ca[!duplicated(cl$title),]
cl_no_duplicate =
cl_no_duplicate[!duplicated(cl_no_duplicate$text),]
#see cities with the most posts and the least post
sort(table(cl_no_duplicate$city),decreasing = TRUE)




#Q2(1)Are apartments in suburbs more likely to be family-
friendly
# (many bedrooms, pets allowed, etc) than apartments in
major cities?
#---------------------------------

#suburbs and major city

major_city <- subset(cl_no_duplicate,city == 'Los
Angeles' | city == 'San Diego' |
                 city == 'San Jose' | city == 'Sacramento
' | city == 'Oakland'|
```

```
                city == 'San Francisco' | city == 'Long
Beach' | city == 'South San Francisco' |
                city == 'West Sacramento'| city ==
'Fresno')

suburbs <- subset(cl_no_duplicate,city != 'Los Angeles' &
city != 'San Diego' &
                city != 'San Jose' & city != 'Sacramento
' & city != 'Oakland'&
                city != 'San Francisco' & city != 'Long
Beach' & city != 'South San Francisco'&
                city != 'West Sacramento' & city !=
'Fresno')



#bedroom
summary(major_city$bedrooms)
summary(suburbs$bedrooms)

props3 = prop.table(table(major_city$bedrooms))
props3 = as.data.frame(props3)
g7 = ggplot(props3,aes(x = Var1,y = Freq)) +
geom_bar(stat = 'identity') +
  labs(x='bedrooms',y = 'proportion',title = 'major
city') +
  ylim(c(0,0.6))

props4 = prop.table(table(suburbs$bedrooms))
props4 = as.data.frame(props4)
g8 = ggplot(props4,aes(x = Var1,y = Freq)) +
geom_bar(stat = 'identity') +
  labs(x='bedroom',y ='proportion',title = 'suburbs') +
```

```
  ylim(c(0,0.6))

grid.arrange(g7,g8)


#pets
props1 = prop.table(table(major_city$pets))
props1 = as.data.frame(props1)
g9 = ggplot(props1,aes(x = Var1,y = Freq)) +
geom_bar(stat = 'identity') +
  labs(x='Pets',y = 'proportion',title = 'major city') +
  ylim(c(0,0.7))

props2 = prop.table(table(suburbs$pets))
props2 = as.data.frame(props2)
g10 = ggplot(props2,aes(x = Var1,y = Freq)) +
geom_bar(stat = 'identity') +
  labs(x='Pets',y = 'proportion',title = 'suburbs') +
  ylim(c(0,0.7))

grid.arrange(g9,g10)



#Q2(2)Which adds more to rent--------------------
summary(cl_no_duplicate$price)

hist(cl_no_duplicate$price)
tail(sort(cl_no_duplicate$price),10)
#(From discussion)no one will rent an apartment for
$9M/month
#we can potentially ignore them in the context of our
problems.
```

```
#sometimes the values are not acctualy that large.
#There was just an issue in parsing the data
#If we read the posting for the highest priced apartment
#Then we see the price range is $3408-3742
#the price in this data is 34083742
cl_no_duplicate$price[cl_no_duplicate$price>=30000000] <-
3408
cl_no_duplicate$price[cl_no_duplicate$price==9951095] <-
995

#smaller than 100 price delete
cl_small_price <- subset(cl_no_duplicate,price <= 100)
cl_no_duplicate$price[cl_no_duplicate$price <= 100] <- NA

mean_bedrooms =
aggregate(price~bedrooms,cl_no_duplicate,mean)
mean_bathrooms =
aggregate(price~bathrooms,cl_no_duplicate,mean)
g5 = ggplot(mean_bedrooms,aes(x =bedrooms,y = price,group
= 1)) + geom_point() + geom_line() +
  ylim(c(0,15000)) + labs(x = 'number of bedrooms',y =
'mean price')
g6 = ggplot(mean_bathrooms,aes(x =bathrooms,y =
price,group = 1)) + geom_point() + geom_line() +
  ylim(c(0,15000)) + labs(x = 'number of bathrooms',y =
'mean price')

grid.arrange(g5,g6)
```

```r
#Q2(3)Do apartments in similar geographical areas tend to
be similar?
#-------------------------------------
cl_sample <- subset(cl_no_duplicate,city == 'San
Francisco' | city == 'Oakland' |
                  city == 'San Jose')
#look at bedroom
ggplot(cl_sample,aes(x = bedrooms,..prop..))+
  geom_bar() + facet_wrap(~city) +
  labs(x = 'number of bedrooms',y = 'proportion',
       title = 'The proportion of apartments with
different number of  bedrooms')

#look at price and sqft
#sqft
hist(cl_no_duplicate$sqft)
plot(sort(cl_no_duplicate$sqft))
tail(sort(cl_no_duplicate$sqft))
#perhaps 200,000 sqft was an error in the data
#so we remove it
cl_no_duplicate$sqft[cl_no_duplicate$sqft >= 10000] <- NA


head(sort(cl_no_duplicate$sqft))
#perhaps 1 and 3 sqft are erros in data, so we choose to
remove it
cl_no_duplicate$sqft[cl_no_duplicate$sqft <= 3] <- NA


#dot plot about price and sqft
ggplot(cl_sample,aes(x = sqft,y =price)) + geom_point() +
  facet_wrap(~ city) + labs(title = ' Relation between
sqft and price in three cities')
```

```r
#Q3&4: Own Questions--------------------------
#1.Will the parking influence the price of apartments?
How does it influence that?
mean_price_parking =
aggregate(price~parking,cl_no_duplicate,mean)
ggplot(mean_price_parking,aes(x = parking,y = price)) +
geom_bar(stat = 'identity') +
  labs(title = 'Mean price for different conditions of
parking')
cl_none <-subset(cl_no_duplicate,parking == 'none')


#2.Will the laundary influence the price of apartment?

#3.Will the apartments with 0,1, 2, or 3 bedrooms be
different on the
#laundary? and how?
cl_bedrooms0123 <- subset(cl_no_duplicate,bedrooms == 0
|bedrooms == 1 | bedrooms == 2 |
                        bedrooms == 3 )
ggplot(cl_bedrooms0123,aes(x = laundry))+
  geom_bar() + facet_wrap(~bedrooms) +labs(title = 'The
situation of laundry for different number of bedrooms')
cl_hookup <- subset(cl_no_duplicate,laundry == 'hookup')

#4.Which area(by latitude and longitude ) in Los Angeles
has more
# relatively low price apartments?
```

```r
#5.How the price of apartment varies in north california
and south california?
cl_N <- subset(cl_no_duplicate,latitude >= 36)
cl_S <- subset(cl_no_duplicate,latitude < 36)

summary(cl_N$price)
summary(cl_S$price)

g1 = ggplot(cl_N,aes(price)) + geom_histogram(aes(y =
stat(count / sum(count)))) +
  xlim(c(0,7500)) + ylim(c(0,0.25)) +
  labs(title = 'Situation for North Calfornia',
                                      y = 'proption')
g2 = ggplot(cl_S,aes(price)) + geom_histogram(aes(y =
stat(count / sum(count)))) +
  xlim(c(0,7500)) + ylim(c(0,0.25)) +
  labs(title = 'Situation for South Calfornia',
                                      y = 'proption')

grid.arrange(g1,g2)


#6. Are apartments in San Francisco and Los Angeles
different in rent market(from price and sqft)?
cl_sf <- subset(cl_no_duplicate, city == 'San Francisco')
g3 = ggplot(cl_sf,aes(x = sqft,y = price)) + geom_point()
+
  ylim(c(0,15000)) + xlim(c(0,4000)) + labs(title = 'San
Francisco')
g4 = ggplot(cl_los,aes(x = sqft,y = price)) +
geom_point() +
  ylim(c(0,15000)) + xlim(c(0,4000)) + labs(title = 'Los
Angeles')
```

```
grid.arrange(g3,g4)

#7.If a person want to rent a studio apartment(0
bedrooms) or apartment
#with one bedroom, which area(by longtitude and latitude)
should he go?
cl_bedrooms01 <- subset(cl_no_duplicate,bedrooms == 0 |
bedrooms == 1)
ggplot(cl_bedrooms01,aes(x = latitude,y = longitude)) +
geom_point() +
  geom_density2d() +
  labs(title = 'spread of apartments with 0 or 1
bedroom')
cl_placetogo <- subset(cl_no_duplicate, latitude > 37 &
latitude < 38)
sort(table(cl_placetogo$county))

#8.Will the apartments with more bedrooms allow pets more
easily?



#9.Do apartments have a stricter policy on dogs or cats?


#10.For counties with more famous colleges(rank top
100),will the price be
#higher?


#Q5:limitations of the data set
```

```r
cl_isna <- sapply(cl,is.na)
cl_isna_count <- apply(cl_isna,2,sum)
sum(cl_isna_count)
sort(cl_isna_count)
#I have mentioend and solved other limitations in the
front questions.
```