Exploratory data analysis part I:
Using summary statistics and visualization with ggplot2

Xiaozhu Zhang

October 28, 2018

1. Introduction

- The datase being analyzed in this report is regarding the information of California apartment rentals posted the Craigslist. We consider each row (each post) as an observation. The number of posts contained in this dataset is 21948. The post in this dataset ranges from September 8th to October 15th in 2018.

- The rental prices are in US dollars and the sizes of the apartments are in square feet. In addition, this dataset also provides the original post on craiglist and information on the location, layout, data posted/updated, as well as some other features such as pet policy and parking option for the advertised apartments.

Primary observation

- Most of the posts in this dataset are for apartment with 1 or 2 bedrooms.

- Also notice that there are some abnormal obervations in this dataset. For example, the summary statistics on the rental prices showed unusual minimum of 0 dollar and maximum of $34080000. This might suggest that this dataset may have errors in information entry and possibly irrelavant information.

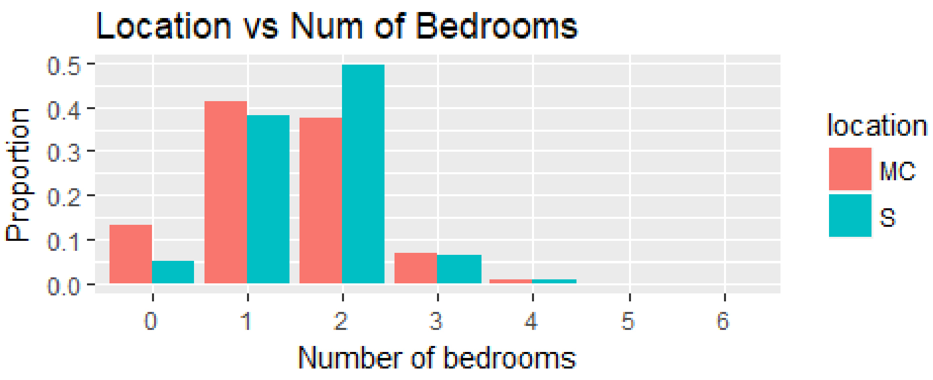
- Furthermore, we should note that there are a number of duplicate rows in this dataset.

2. Analysis

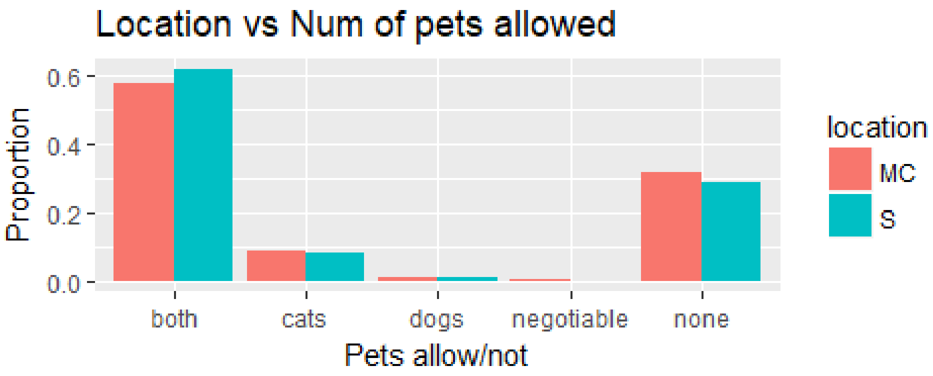
2.1 First we want to see if the apartments in suburbs are more family-friendly than the apartments in major cities.

Definition

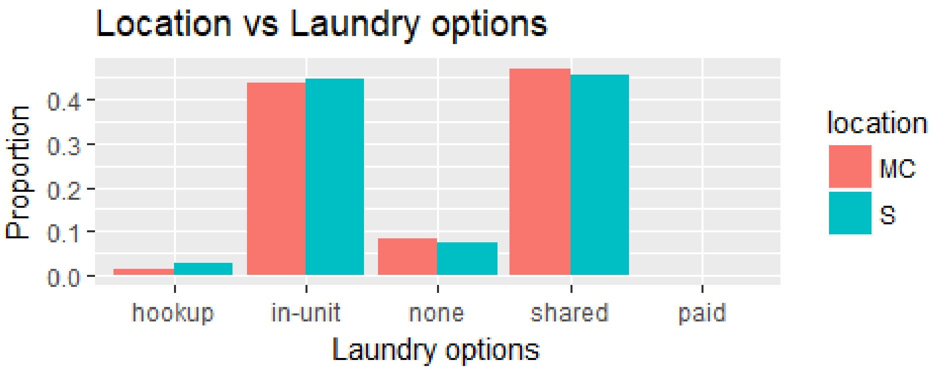
We defind that major cities are the observations with the number of post more than 500 and suburbs are those with the number of post less than 100.And that the family-friedndly apartment has more bedrooms, allows pet, has in unit/hookup laundry and cover/garage parking option.



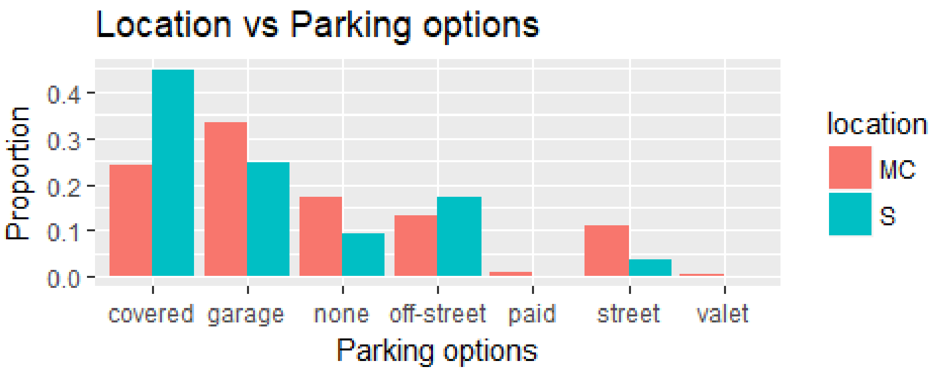
Location vs Num of Bedrooms



Location vs Num of pets allowed



Location vs Laundry options



Location vs Parking options

Observations and interpretation

- The first barplot suggests that only the number of two-bedrooms apartments seem to be more common in suburbs compared to major cities, whereas the other types of apartment did not show such trend.

- The seond barplot showed that apartments in suburb have a slightly higher probability of allowing both cats and dogs.

- The third barplot showed that apartments in suburb seem to have a slighty higher proportion of having hookup and in-unit laundry options but the proportion of shared laundry is also fairly high in the suburb apartments.

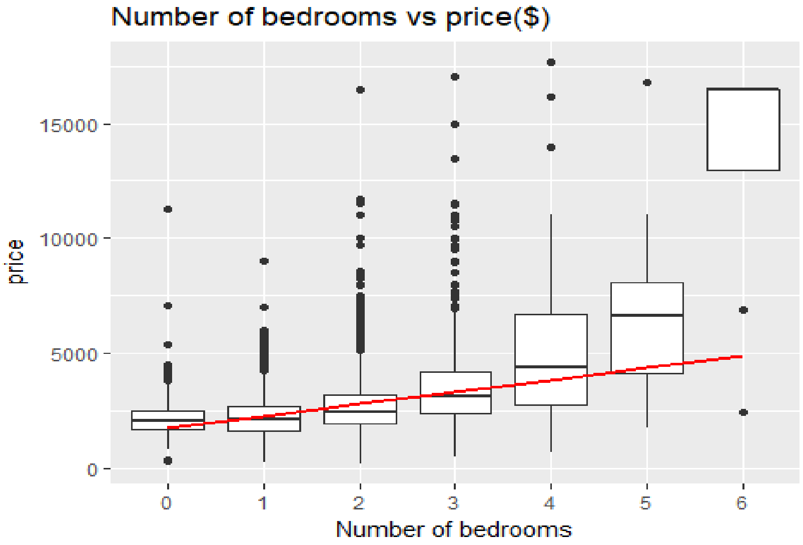
- The fourth barplot showed that the the apartments in suburbs have a significantly higher proportion of covered parking but lower proportion of garage.

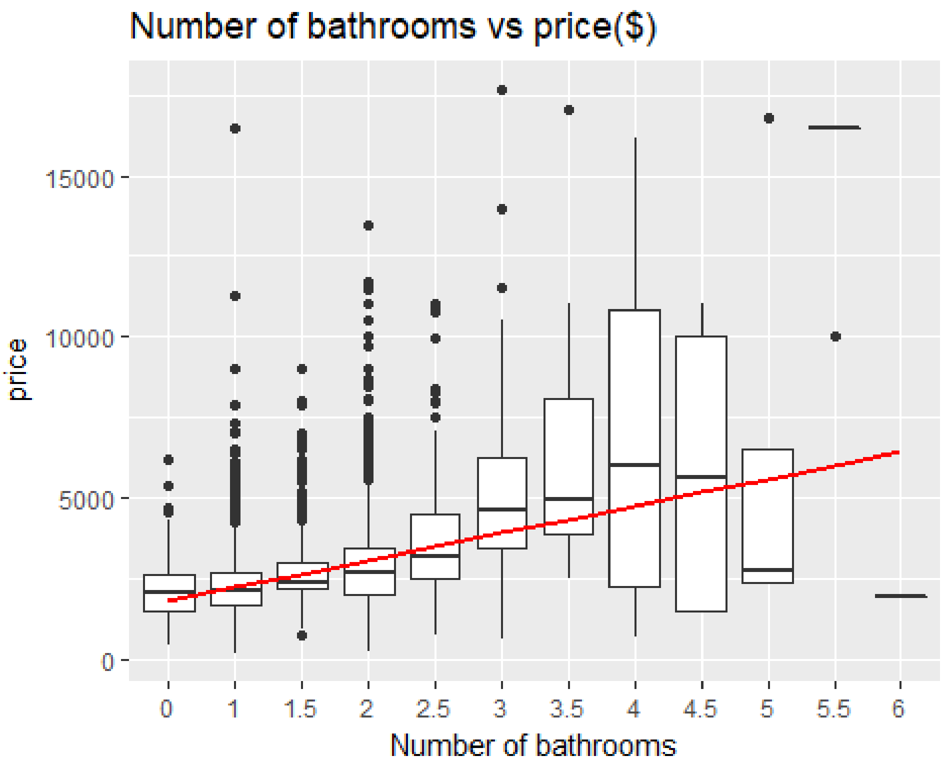
- Overall, the suburb apartments tend to have more 2-bedrooms apartments, a slightly higher proportion of allowing pets, good laundry options and more covered parking. There seems to be some sort of evidence to suggest that the apartments in suburb are relatively more family-friendly than those in major cities.

2.2 Does an extra bedroom or bathroom contribute more to the rent?

- In order to compare which whether an extra bedroom or bathroom adds more on the price, we plot the group boxplot and fit a linear line to indicate the slope in each case. As the plot suggests, an extra bathroom seems to add on more price because of its steeper slope.

- In addition, the coefficient of a simple linear model is 525.9063 for an increase in the number of bedrooms and is 821.2146 for an increase in the number of batrooms. This confirms our observations that an extrea bathroom seems to add on more price to an apartment.





2.3 How does the overall rent vary between city and suburb? Does the number of bedrooms have the same trend?

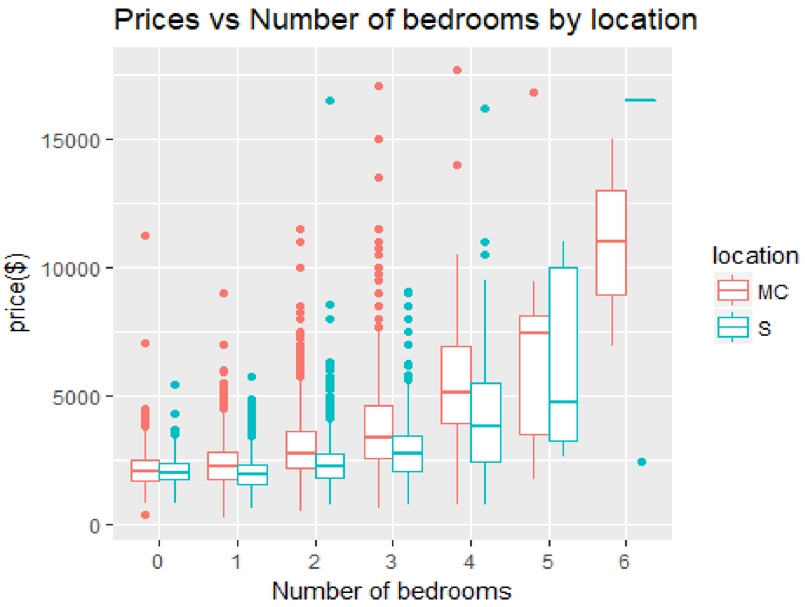
- We will continue to use the criteria of how we define cities(post>500) and suburbs(post<100)

- My hypothesis is that the monthly rent of the apartments in cities are higher than the ones in suburb and the trend should be the same regardless of the number of bedrooms.

Result

- Overall, the average monthly rent of apartment in cities is 2689.444 dollars and 2272.286 dollars for the apartments in suburbs. This confirms the hypothesis.

- However, as suggested by the boxplot, this trend seems to be contradicted in the case of the apartments with 5 and 6 bedrooms due to the higher 75th quantitle and maximum for the rental price of apartment in suburb. We should also note that the number of observations for the 6 bedroom apartments in suburb is obviously fewer than that of cities and this could affect the result.



2.4 For the apartments with the same number of bedrooms, which city has the highest and lowest monthly rent?

- Because the most common type of apartments in cities are 1-bedroom(41.17%) and 2-bedroom apartment(37.75%), our discussion will be based on these two types of apartment.

Result

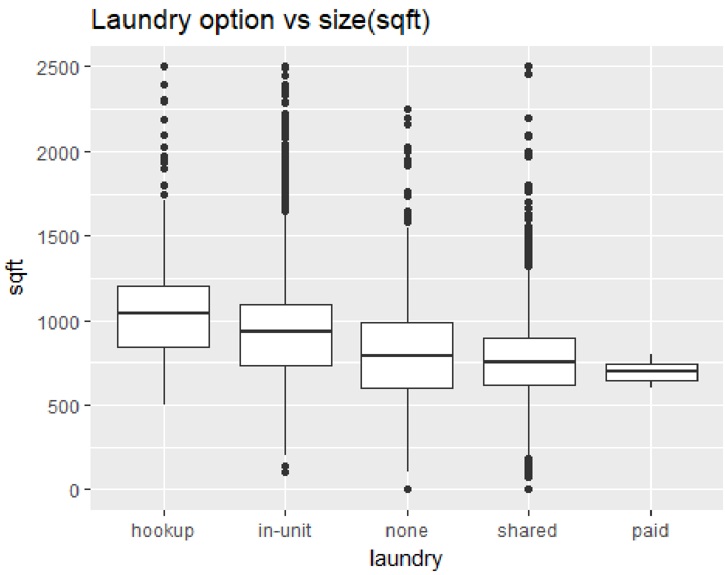
- The result confirms the hypothesis that San Francisco has the highest rent. In addition, Daly city and Sacramento have the lowest rent for 1 bedroom and 2 bedrooms apartments.

2.5 Does the size(in square feet) of the apartment have effect on the type of laundry options?

- My hypothesis is that larger apartment will tend to have hookup and in-unit laundry, shared or paid laundry should be more common among smaller apartments.

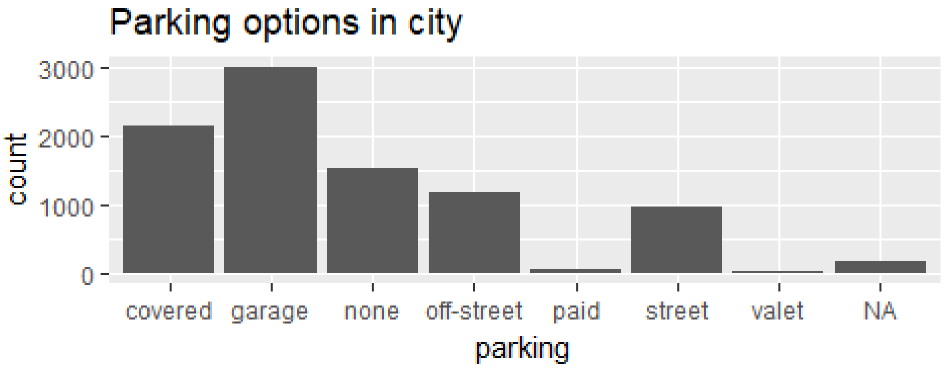
- Note that we zoom in by enlarging the scale of the boxplot in order to capture the trend bewteen the size of the apartment and laundry option. Therefore, some observation that are relatively far awary from the trend might be cut off.

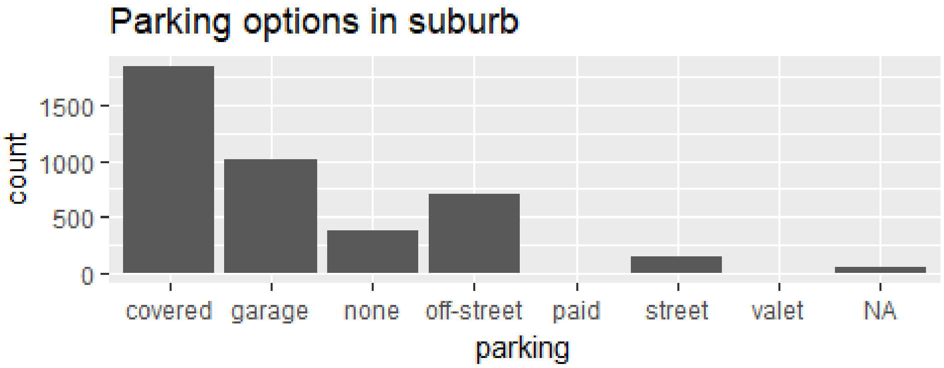
- The boxplot suggests that the apartments with hookup and in-unit laundry option tend to have larger size and the ones with shared or paid option tend to be smaller apartment. This confirms the hypothesis.



2.6 How does the parking option differ bewteen suburbs and cities?

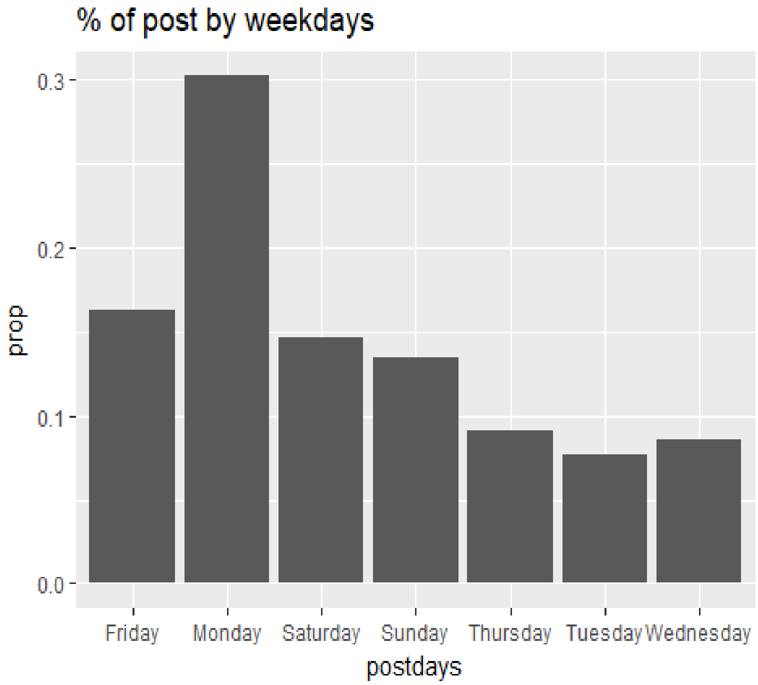
- My hypothesis is that there are more garages in suburbs and more off-street parkings in cities.
- The result shows that apartments in cities tend to have more garage whereas the ones in suburbs tend to have the most covered parking. Even though this contradict the hypothesis, the result makes sense considering that suburb tends to have more space.





2.7 Is there a specific day when there tend to be more post?

- My hypothesis is that there should be more post during weekend.
- The barplot shows that the post on Monday are significantly more than the other weekdays.



Limitations

I.Sources and reliability of dataset

It has been mentioned in the description of this assignment that Craigslist has few rules about the post format and the original post are unformatted. There could be two potential problems that affect the accuracy of the data. First, the information posted on website of Craigslist itself may be unspecific due to missing information or inaccurate due to errors (careless and purposeful).Secondly, the way of extracting informtion out of the orginal post could also involve subjectivity and potentially neglect some information and affect the result.

II.Missing data

- There are 103 rows with missing prices and out of them the san diego site has the most missing prices.
- There are 1048 rows with missing information on the number of bedrooms and out of them the los angles site has the most missing data. Also noted that the row with missing data on bedrooms also tend to have missing data on the number of bathrooms.
- There are 5591 rows with missing data on the size(sqft) and those with missing data on this feature tend to be the apartments with shared laundry and 1 bathroom.

III.Outliers and error

- During this analysis, 3 major outliers were discovered and will be discussed here.
- Row 1446(two bedroom)has an unusually large size of 20000 square feet,which seems highly implausible.Since the correct size is not specified in the dataset,I choose to delete this row to so that it won't affect the result as well as the graph.
- The other two outliers are found in the price feature,both of them are greater than 100000 dollars which also seems not realistic so I found these two row and it turned out they are not outliers,instead, they are the range of the price(995-1095) in the wrong format(9951095). Since the correct information is in the dataset, I changed the value of the price to the correct ones instead of deleting them.

IV. Number of observations

Some concerns regarding the number of observations should also be noted and the number of 6 bedroom partment is an example that illustrate this problem. Compared to the apartments with the number of bedrooms of 1-5(20891), there are too few observations that have 6 bedrooms(only 6). This lack of adequate observation may affect the result involving the number of bedrooms(ie:we may not see the pattern we are suppose to see).

V. Missing features

There could be missing features that might confound with the result. Examples of such features could be the condition of the apartment(is it newly built or a relatively old apartment).It should be logica to assume that a 10-year old apartment may be priced lower than a newly-built apartment. Hoever, due to this missing feature we might have attribute the reson to some other variables.

VI. Definition of suburb and city

Another limitation of this analysis is how we define and subset major cities and suburbs. In this anaysis, the criteria of subseting is based on the number of post. It is possible that there are other ways to establish the criteria. The choice of criteria may be affected by subjectivity and choice of difference criteria may very likely give difference result.

VII.Scope

In conclusion, due to the limitation of the sources of the dataset, outliers and errors, lack of adequate observations for specific categories, mimssing features and choice of criteria, the conclusions from this analysis should be applied to new observtions with great caution. As discussed in the previous section, the limitation of this analysis may result in some pattern by chance or neglect some important patterns we should have seen.

Some questions we might be interested in and can be answered by this dataset could be:

- (1) How does the overall rent vary between city and suburb? Does the number of bedrooms have the same trend? This question could provide information on the overall trend of renting prices between city and suburb and how does the price look like or different based on the number of bedrooms for the people who are trying to decide where to live with a fixed budget and for those with families.

- (2) For the apartments with the same number of bedrooms, which city has the highest and lowest monthly rent? Simliar to the previous one, this question can provide more specific information on the location of the apartment with the highest and lowest rent and might be helpful for those trying to pick a living place that suit their budget.

- (3) What kind of apartments tend to be updated? What does it imply about this kind of apartment? This question may help people to pay more attention the post they are looking at and be more cautious about choosing apartment. If an apartment is being updated for a long period, this might say something about the apartment (is it because of its locaion/condition/neighbour?)

- (4) Does the size (in square feet) of the apartment have effect on the type of laundry options? This question could be helpful for families that looking for place to live. A hookup or in-unit laundry option could be more convinient than the other option. If there is a relationship between the size and tyoe of laundry option, knowing this effect could help the families to choose apartment that better suit their needs.

- (5) How does the parking option differ bewteen suburbs and cities? This question might be helpful for those who have cars to determine is it true that cities tend to have less parking space.

- (6) Is there a specific day when there tend to be more post (or updated post)? This question may be helpful for those thinking about looking for an apartment.They target audience could be students or working force who do not have a lot of time browsing information online. Knowing this could help them to maximize the information they could obtain.

- (7) Are the posts all regarding long-term renting? This question may concern those looking for short-term housing, knowing whether there are post on short-term housing option could help them decide whether they want to look for option on craigslist or some other website.

- (8) Is it possible to build a predictive model of rent based on the variables displayed in this dataset? This question is more on the perspective of the people post the renting informtion. A model could be a useful reference in helping them to decide what rental price is suitable for the apartment.

R appendix

```r
#Introduction code
library(ggplot2);library(knitr);library(gridExtra);library(kableExtra)
houseinfo=readRDS("C:/Users/zhenguo/Downloads/cl_apartments.rds")
nrow(houseinfo)#NUMBER OF OBSERVATION
range(houseinfo$date_posted,na.rm = TRUE)#TIME SPAN
placespan=table(houseinfo$place)

#examine the prices
summary(houseinfo$price) #see some unusual min and max needs further investiag
ation

#Remove irrelavent rows with information
doubt=houseinfo[grep('br', houseinfo$title, ignore.case = TRUE, invert = TRU
E),]
doubt=doubt[grep(c('studio'), doubt$title, ignore.case = TRUE, invert = TRU
E),]
doubt=doubt[grep(c('bed'), doubt$title, ignore.case = TRUE, invert = TRUE),]
doubt=doubt[grep(c('room'), doubt$title, ignore.case = TRUE, invert = TRUE),]
doubt=doubt[-which(doubt$bathrooms>0),];doubt=doubt[-which(doubt$sqft>0),]
doubt=doubt[-which(doubt$price>200),]
irre_info_rows=as.numeric(rownames(doubt))
houseinfo=houseinfo[-irre_info_rows,]
houseinfo=houseinfo[-which(houseinfo$price<100),]

#take a look at the two obvious unusual price
houseinfo[which(houseinfo$price>100000),]
#we see that they are not outliers, they are just wrong format of prices,chan
ge the value to lower value
houseinfo[which(rownames(houseinfo)=="4531"),]$price=995
houseinfo[which(rownames(houseinfo)=="15961"),]$price=3408
#duplicated data:e will remove the ones with same title because it could be m
isleading since it adds more weight to the same apartment
houseinfo=houseinfo[-which(duplicated(houseinfo[c("title")])==TRUE),]
#2.1friendly-bedrooms, allow pet, in unit landury,parking
#Q:Are apartments in suburbs more likely to be family-friendly?
major_cities=placespan[which(placespan>500)]
suburb=placespan[which(placespan<100)]
major_cities=rownames(placespan[which(placespan>500)])
suburb=rownames(placespan[which(placespan<100)])

major_cities=houseinfo[houseinfo$place %in% major_cities,]
suburb=houseinfo[houseinfo$place %in% suburb,]
major_cities$location="MC"
suburb$location="S"
city_sub=rbind(major_cities,suburb)

##bedrooms
tbl_bedrooms=round(prop.table(table(city_sub$location,city_sub$bedrooms),1),d
igits = 4)
tbl_bedrooms=data.frame(tbl_bedrooms)
p1=ggplot(tbl_bedrooms,aes(x=Var2,y=Freq,fill=Var1))+geom_bar(stat = "identity",position = "dodge")+xlab("Number of bedrooms")+ylab("Proportion")+ggtitle("Location vs Num of Bedrooms")+guides(fill=guide_legend(title="location"))

#pets
tbl_pets=round(prop.table(table(city_sub$location,city_sub$pets),1),digits = 4)
tbl_pets=data.frame(tbl_pets)
p2=ggplot(tbl_pets,aes(x=Var2,y=Freq,fill=Var1))+geom_bar(stat = "identity",position = "dodge")+xlab("Pets allow/not")+ylab("Proportion")+ggtitle("Location vs Num of pets allowed")+guides(fill=guide_legend(title="location"))


#landruy
tbl_lan=round(prop.table(table(city_sub$location,city_sub$laundry),1),digits = 4)
tbl_lan=data.frame(tbl_lan)
p3=ggplot(tbl_lan,aes(x=Var2,y=Freq,fill=Var1))+geom_bar(stat = "identity",position = "dodge")+xlab("Laundry options")+ylab("Proportion")+ggtitle("Location vs Laundry options")+guides(fill=guide_legend(title="location"))


#parking
tbl_parking=round(prop.table(table(city_sub$location,city_sub$parking),1),digits = 4)
tbl_parking=data.frame(tbl_parking)
p4=ggplot(tbl_parking,aes(x=Var2,y=Freq,fill=Var1))+geom_bar(stat = "identity",position = "dodge")+xlab("Parking options")+ylab("Proportion")+ggtitle("Location vs Parking options")+guides(fill=guide_legend(title="location"))

grid.arrange(p1,p2,nrow=2)
grid.arrange(p3,p4,nrow=2)

#2.2 code
ggplot(houseinfo[!is.na(houseinfo$bedrooms),],aes(x=as.factor(bedrooms),y=price))+
  geom_boxplot()+geom_smooth(method = 'lm',se=FALSE,color="red",aes(group=1))+
  xlab("Number of bedrooms")+ggtitle("Number of bedrooms vs price($)")
ggplot(houseinfo[!is.na(houseinfo$bathrooms),],aes(x=as.factor(bathrooms),y=price))+
  geom_boxplot()+geom_smooth(method = 'lm',se=FALSE,color="red",aes(group=1))+
xlab("Number of bathrooms")+ggtitle("Number of bathrooms vs price($)")
#boxplot of prices vs bedrooms
bed_price=lm(price~bedrooms,data=houseinfo)
bath_price=lm(price~bathrooms,data=houseinfo)
```

```r
ggplot(city_sub[!is.na(city_sub$bedrooms),],aes(x=factor(bedrooms),y=price,co
lor=location))+geom_boxplot()+xlab("Number of bedrooms")+ylab("price($)")+ggt
itle("Prices vs Number of bedrooms by location")
bed1=aggregate(price~city+bedrooms,major_cities[major_cities$bedrooms==1,],me
an)

bed2=aggregate(price~city+bedrooms,major_cities[major_cities$bedrooms==2,],me
an)

bed1=bed1[order(data.frame(bed1)$price),]
bed2=bed2[order(data.frame(bed2)$price),]

Apt_price=as.data.frame(c(1350,3236.344))
rownames(Apt_price)=c("Daly City","San Francisco")
colnames(Apt_price)<-"prices"
kable(Apt_price,"latex",caption="Highest and lowest 1-bedroom prices",booktab
s=T)

Apt_price2=as.data.frame(c(1657.206,4436.754))
rownames(Apt_price)=c("Sacramento","San Francisco")
colnames(Apt_price)<-"prices"
kable(Apt_price,"latex",caption="Highest and lowest 2-bedroom prices",booktab
s=T)

ggplot(houseinfo[!is.na(houseinfo$laundry),],aes(laundry,y=sqft,group=laundr
y))+geom_boxplot()+ylim(0,2500)+ggtitle("Laundry option vs size(sqft)")

p9=ggplot(major_cities,aes(x=parking))+geom_bar()+ggtitle("Parking options in
 city")
p10=ggplot(suburb,aes(x=parking))+geom_bar()+ggtitle("Parking options in subu
rb")
grid.arrange(p9,p10,nrow=2)

houseinfo$postdays=weekdays(houseinfo$date_posted)
postdays=factor(houseinfo$postdays,c("Monday","Tuesday","Wednesday","Thursday
","Friday","Saturday","Sunday"))
ggplot(houseinfo,aes(x=postdays,y=..prop..,group=1))+geom_bar()+ggtitle("% of
 post by weekdays")

 #na.price,sandiego has the most missing data
copy=readRDS("C:/Users/zhenguo/Downloads/cl_apartments.rds")
na_price=copy[is.na(copy$price),]
nrow(na_price)
table(na_price$craigslist)

na_bedroom=copy[is.na(copy$bedrooms),]
nrow(na_bedroom)
table(na_bedroom$craigslist)
```

```r
#rows with na.bedroom also missing bathrrom

#missing square foot___>mainly shared laundry,1 bathroom,
na_sqft=copy[is.na(copy$sqft),]
nrow(na_sqft)
table(na_sqft$laundry)
table(na_sqft$bedrooms)

#take a look at the two obvious unusual price
houseinfo[which(houseinfo$price>100000),]
#we see that they are not outliers, they are just wrong format of prices,chan
ge the value to lower value
houseinfo[which(rownames(houseinfo)=="4531"),]$price=995
houseinfo[which(rownames(houseinfo)=="15961"),]$price=3408
```