Man Pan

Student ID: 914656278

Email: manpan@ucdavis.edu

1. Description Analysis

There are 21948 rows and 20 columns in the data set. Each row represents recent Craigslist posts for apartment rentals from 2018-09-08 to 2018-10-15. But the places are not all in California, which are located in California, Connecticut, Florida, Maryland, North Carolina, Nevada, Ohio, Utah, Virginia and Washington. And each column represents different types of information of each post : apartments' location, price, size, the number of bedrooms and bathrooms, permit of pet, situation of laundry/parking and posted/updated time.

For the rows, there are many duplicate postings, which contains the same information except for posted or updated time. But I think they are useful, cause they may be different apartments which are totally same in one system.

For the columns, the "title" contains all important information of one post, while the "text" contains more details of the apartment. There 10 category variables:bedrooms, bathrooms, pets, laundry, parking, craigslist, place, city, state and county; 4 numerical variables: latitude,longitude,price,sqft. However, "latitude", "longitude", "place", "city", "country" all indicate the similar information. This data set is messy, and have many missing values and inaccurate items.

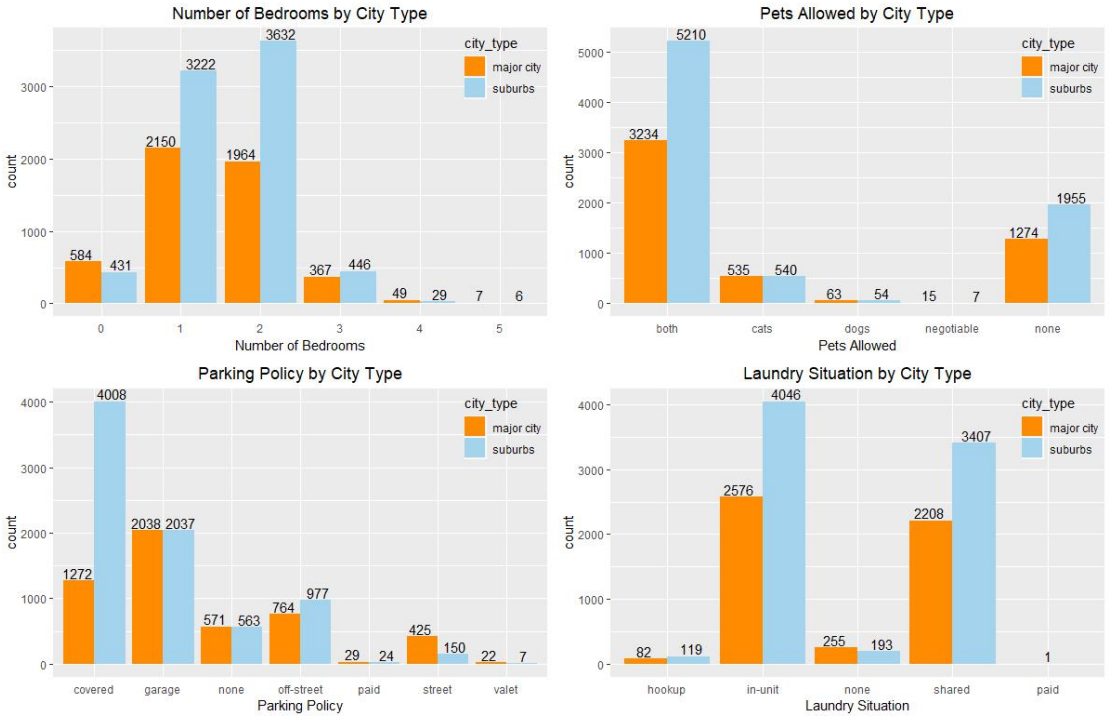
2. Factors Related to Apartments Accommodation

2.1 Are apartments in suburbs more likely to be family-friendly?

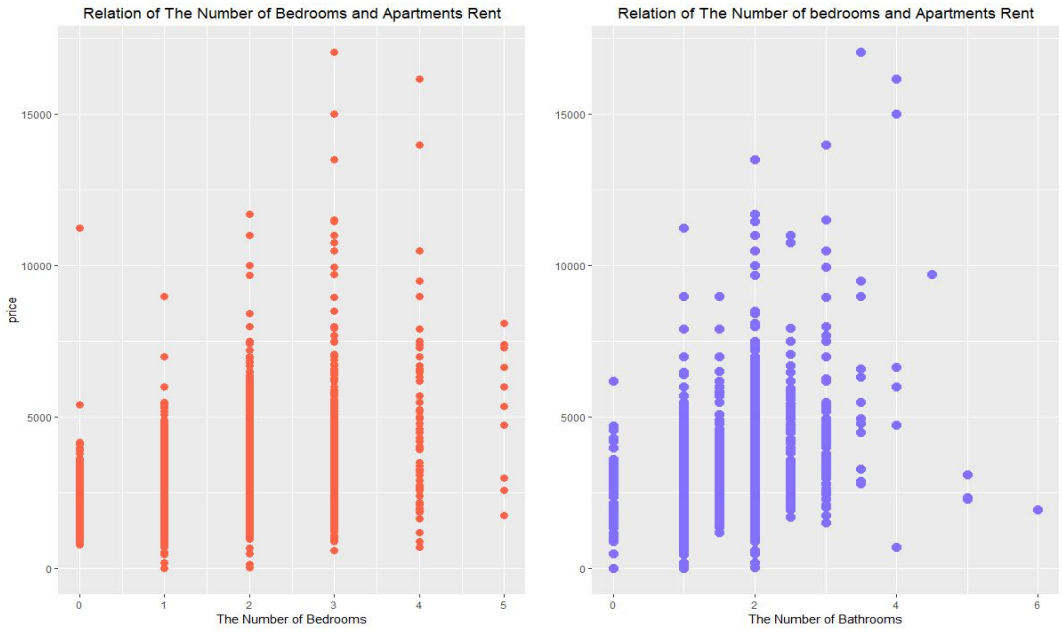
Many people think apartments in suburbs are more likely to be family-friendly than apartments in major cities, because suburbs have a large area and the rent in suburbs is lower than that in major cities. Firstly, I define Family-Friendly apartments have more bedrooms(4 or 5 rooms),pets allowed,garage and in-unit laundry. From the websitehttps://en.wikipedia.org/wiki/List_of_largest_California_cities_by_population, I select the cities whose population larger than 400000 as major cities to analyze. The major cities are Los Angeles, San Diego, San Jose, San Francisco, Fresno, Sacramento, Long Beach and Oakland. And there are 5121 apartments in major cities, 7766 apartments in suburbs.

As the picture below shows: the percentage of 4 or 5 bedrooms' apartments in major cities is 61.54%, the percentage of pets allowed apartments in major cities is 39.77%, the percentage of garage parking apartments in major cities is 50.01%, the

percentage of in-unit laundry apartments in major cities is 38.90%. There is no obvious evidence to apartments in suburbs are more likely to be family-friendly.



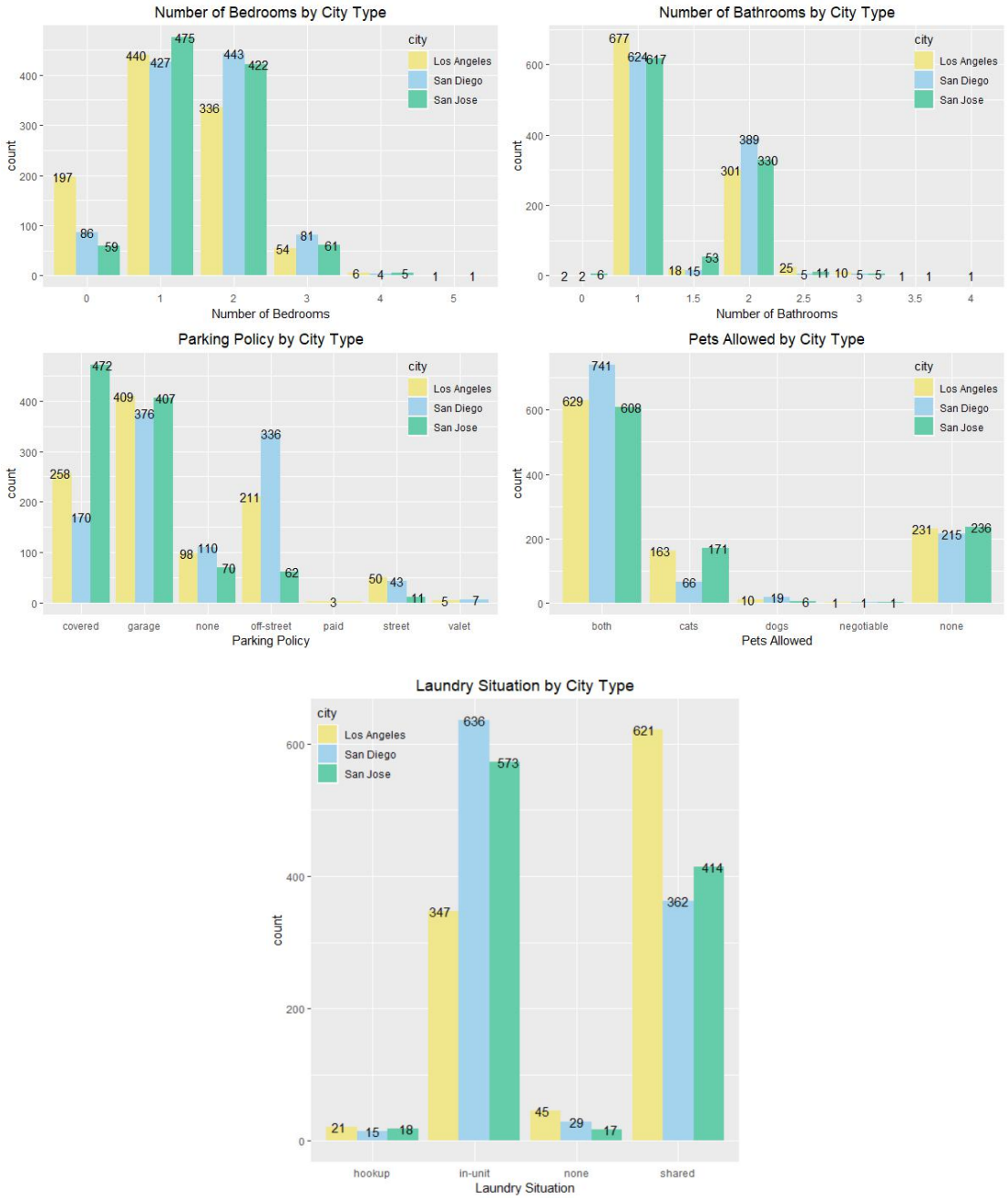
2.2 Relationship between the Number of Bedrooms/Bathrooms and Apartments Rent



As the picture above shows: the high rent occurs when the number of bedrooms is 3 or 4 and the number of bathrooms is 3.5 or 4. The rent will increase overall when the number of bedrooms increases from 0 to 4 . However, the rent decreases obviously when the number of bedrooms increases from 4 to 5. The similar thing occurs in the relationship between the number of bathrooms and apartments rent.

2.3 Accommodation by Difference Cities

As the same criterion in 2.1.1, I select Los Angeles, San Diego and San Jose as the three major cities to analyze. Apartments performances contain the number of bedrooms/bathrooms, parking policy, laundry situation, pets allowed, rent and size.



I plot bar plot to present the count of each category, and plot violin and box plot to show the distribution of numerical variables. The violin plot above shows the distribution of Apartments rent and size in three cities.Wider sections of the violin plot represent a higher probability that rent of that city will take on the given value; the

skinnier sections represent a lower probability.The box plot elements show the median rent in San Diego is lower than that for Los Angeles and San Jose. The median size in Los Angeles is lower than that for San Diego and San Jose. The shape of the distribution (extremely skinny on each end and wide in the middle) indicates the rent are highly concentrated around the median. Los Angeles has the widest range of rent. Distribution of size and rent in San Jose and San Diego are similar.

Then, I analyze the accommodation in particular city:

In Los Angeles, the percentage of apartments with 1 or 2 bathrooms is 94.58%; the percentage of apartments with in-unit and shared laundry is 93.62%.

In San Diego, the percentage of apartments with 1 or 2 bathrooms is 98.73%; the percentage of apartments with in-unit and shared laundry is 95.78%; the percentage of apartments with pets allowed is 91.75%; the rent is concentrated on $1785~$2500; the size is concentrated on 650~1027 square.

In San Jose, the percentage of apartments with 1 or 2 bathrooms is 92.57%; the percentage of apartments with in-unit and shared laundry is 96.58%; the rent is concentrated on $2200~$2887; the size is concentrated on 671~1035 square.

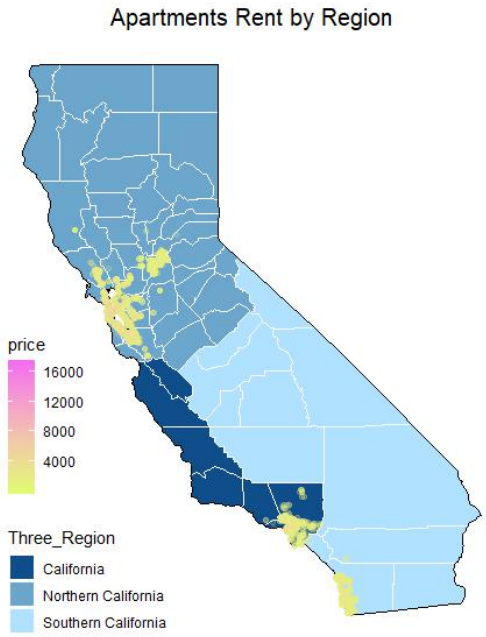
Also, if I pay attention to superior quality apartments in three cities, I can find apartments in Los Angeles and San Jose have similar rent. Superior quality apartments rent in LA are about 4500, while those in San Jose are about 3699.

Thus, we just can find some similar performance in the number of bathrooms,Laundry situation and rent for particular apartments group in each city, which cannot lead to the conclusion that apartments are similar in similar geographical areas.

2.4 Apartments Rent by Different States

I am interested apartments rent by different states.By knowing relation between apartments rent and different states, it can help people to select area for living. Firstly, I guess can we split posts into different groups by latitude and longitude? From the website https://en.wikipedia.org/wiki/Cal_3, I find people are used to split California into three states: California, Northern California and Southern California.

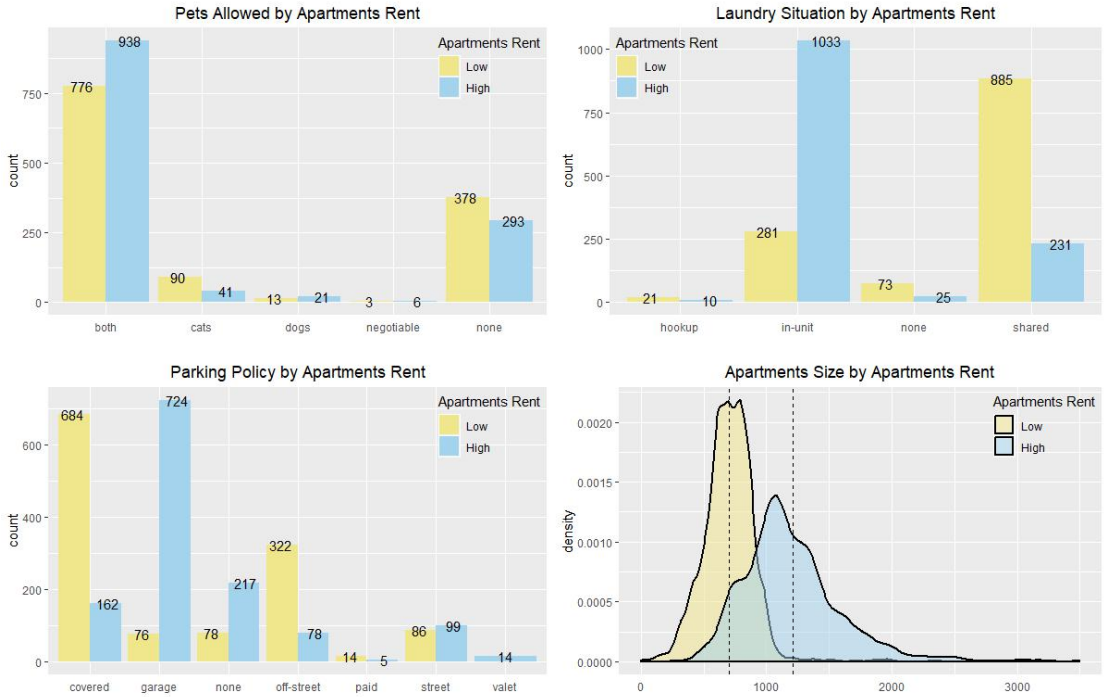
By plotting the distribution of latitude and longitude,I find the apartments are concentrated on three regions, which can match the category in the link. Thus, I guess that apartments in the same states may have similar house rent and size. I plot CA map to distinguish three states in different color. I also add points in different color which indicate different values of apartments rent. Pink items mean larger values, while yellow items mean smaller values.



As the picture above show: some high rent apartments (pink color) occur in Northern California. Also, we know San Francisco and San Jose are in Northern California, the two cities have a relative high rent among all cities in that data set. That is why rents in Northern California are higher than others. Also, Bay area is located in Northern California, several high technology companies would likely to increase GDP in that area, which lead to high rent apartments.

2.5 Accommodation by Different Rent

In my opinion, it would be better to clarify the difference between overall apartments and high rent apartments. Because that can help house manager persuade people to rent high price apartments. I guess high rent apartments would have lager size,in-unit laundry,garbage parking and allow pets.I define high rent apartments as the apartments of rent top 10^{th} quartile, while low rent apartments as low 10^{th} quartile rent.



As the picture above shows: compared to low rent apartments, high rent apartments are more likely to allow tenants to keep pets, especially dogs; most of high rent ones have in-unit laundry and garage parking. Only high rent apartments provide valet. The distribution of size of high rent apartments are relatively scattered, the mean size is 1211 square. For low rent apartments, they may be do not allow pets, but if they allow pets, cats are prefer. And most of those have shared laundry and covered/street parking. They never provide valet. The size of low rent apartments are more concentrated, which have the mean of 699 square.
Note: In order to clarify the distribution of rent, I delete 5 outliers (super large values)in the rent of data set.

2.6 Relation Between Apartments Rent and Size

By plotting the box-plot of Apartments' size, I find there are some outliers which will effect the exploration of relationship between Apartments' size and rent. I delete a outlier which larger than 5000 square and correct two errors . As the picture below shows: size and rent have a positive relationship, the linear correlation is 0.5220195. Thus, we can make a conclusion that the rent will increase as the size of apartments become larger. This conclusion matches our common sense.

Then, I want to explore the relation between rent and apartments size in different quartile (20^{th} quartile, 40^{th} quartile, 60^{th} quartile, 80^{th} quartile and 100^{th} quartile). Finally, I find there are no obvious relationship between apartments size in each quartile and rent. And the linear correlation of 80^{th} -100^{th} quartile apartments size and rent is 0.4201757.I think points larger than 1500 square effect the correlation overall a lot. Although we can get the correlation of 0.5220195 overall, there maybe no linear

relationship between Apartments size and rent. The rent just increases when the size of apartments increases.

Furthermore, I want to explore if there are some linear relationship between size and rent in big cities:Sacramento, San Diego, Los Angeles, Oakland, San Jose and San Francisco. I guess there may be some linear relation if we get rid of cities factor.

As the picture above shows: compared to the scatter plot overall, there are an obvious linear relationship between rent and size in each city separately. Rent and size are all positively related, although they have different slopes. San Francisco has the most positive relationship while Sacramento has a slight positive relation between size and rent.

2.7 Mid-Apartments or Superior Quality Apartments Rent in Big Cities

According to different people, they have different requirements for apartments. I have to set a standard to decide Mid-Apartments and Superior Quality Apartments to explore which city have a relative low rent under the same level quality apartments. I guess cities that have larger population and higher GDP would like to have high rent, regardless of mid-apartments or superior quality apartments.
In my opinion, Mid-apartments should have 1/2 bedrooms, 1/1.5 bathrooms, covered/ street/ paid parking, hookup/shared laundry. However, Superior Quality Apartments should have 3/4/5 bedrooms, 2.5/3/3.5/4 bathrooms, valet/garage parking and in-unit laundry. I think the measure of pets allowed is not related to apartments quality level.

As the picture above shows: For Mid-Apartments, the box plot elements show the median rent in Sacramento ($1152.5) is lower than that for other big cities, while the median rent in San Francisco ($3198.5) is higher than that for others. From the lower to higher, the order of cities are Sacramento, San Diego, Los Angeles, Oakland, San Jose and San Francisco. For superior quality apartments, the box plots show the median rent in San Diego ($2750) is lower than that for other big cities, while the median rent in San Francisco ($7725) is higher than that for others. From the lower to higher, I find the cities are San Diego, San Jose, Los Angeles and San Francisco. San Francisco has the highest rent, regardless of mid-apartments or superior quality apartments. San Francisco doesn't have the largest population, but it has a relatively high GDP. Similar things occur to Los Angeles and San Jose. Thus, I think, compared to the population in that city the GDP in a city effect the apartments' rent a lot.

2.8 Further Questions

I also interested in some questions like : Apartments Rent by Craigslist Branches. By knowing relation between apartments rent and craigslist branches, it can help craigslist to manager their business. They can deal with some unusual rents in time to improve the website's performance. Proportion of high-grade apartments in major cities, which can tell us which city is more suitable for salariat to live. Rent distribution of coast apartments, which can give people who want to invest coast villas some suggestions. Average rent of en-suite apartments, which can help new married family know more about rent distribution. Some can help website to manager their business, some can give people suggestions to select their apartments location.

3. Limitations

As the table shows below, I find there are many missing values in this data set, which follow some patterns: 'Title' and 'text' are have 1 missing value, 'latitude' and 'longitude' have 84 missing values, 'bedrooms' and 'bathrooms' have 1048 missing values, 'state' and 'county' have 95 missing values. Some columns have the same number of missing values, because they indicated the similar information.

Columns name	title	text	latitude	longitude	city_text	date_posted
NA Counts	1	1	84	84	1661	1
Columns name	price	sqft	deleted	bedrooms	bathrooms	date_updated
NA Counts	103	5591	0	1048	1048	13139
Columns name	pets	laundry	parking	craigslist	place	city
NA Counts	293	216	299	0	701	1856
Columns name	state	county				
NA Counts	95	95				

Also, there are some outliers and anomalies. When I plot box-plots, I can find some outliers which are far away from others. For example, when I explore the relation between apartments size and rent, there are little apartments have larger size and lower rent, which will influence my analysis of relationship. Maybe that apartment's location is not good, although its size is large, people don't prefer inconvenient location. So I also get rid of them to continue my analysis to lead to a more accurate conclusion.

There are also some errors. For example, two posts' price are 30000000 and 9900000, which are wrong (cannot match the price in 'text' column). In the 'text', the prices are 3408, 995 separately. I think they are typos, so I change the error price to the true value.

The data set may be from the website by web-scraping; person(s) that created the

data set have a bias, because they just got data they are interested in. For example, they are just interested in apartments in relative large cities. Some towns like Davis don't occur in this data set. As a student in Davis, I know the rent are highly related to the location and owner of apartments. Rent would increase when apartments nearby university for private developer. However, apartments' rent in school are lower than off-campus apartments that have the same size. Small towns and cities have different criteria to analyze factors to apartments rent. It would be better to add more cities/towns and some information about transportation, environment to improve data completeness.Thus, persons bias and incomplete data set will lead to inaccurate conclusion.

Some questions have enough observations, some do not. For instance, when I analyze accommodation in major cities and suburbs. The observations are 5121 and 7766 separately, which are enough. However, when I analyze accommodation of apartments in different rents, there are just 14 apartments which have valet parking, which cannot lead to the conclusion of no low rent apartment has valet parking.

My conclusions just apply to the observations in this data set. As I mention above, different cities/towns have different standards. Also, considering the different government policies, I cannot make sure my conclusions can apply to apartments in other states.

4. Conclusion

Although this is a rich data set with a lot of information, there are some biases and omissions. Considering that the data is collected by persons, we may be missing lots of important information, such as small towns rent; environment, GDP and population in each city. It would always be better to have more and cleaner data, but this data set gave some useful insights. We can know the apartments accommodation are focus on 0~3 bedrooms, 1 or 2 bathrooms, covered/garage/off street parking, in-unit/shared laundry. Apartments size are from 600 to 1000 square. Among all cities in that data set, San Francisco have a relative high rent.

Appendix

```
apartments = readRDS("cl_apartments.rds")
nrow(apartments)
ncol(apartments)
levels(apartments$place)
levels(apartments$craigslist)
levels(apartments$county)
levels(apartments$city)
levels(as.factor(apartments$state))

# time span
apartments$date_posted = as.character(apartments$date_posted)
time = NULL
for (i in (1:21948)){
   temp = strsplit(apartments$date_posted, " ")[[i]][1]
   temp = as.data.frame(temp)
   time = rbind(time,temp)
   time = unique(time)
}
time

library(ggplot2)
library(gridExtra)

# Def major cities: Los Angeles, San Diego, San Jose, San Francisco, Fresno,Sacramento,Long
Beach,Oakland
# the population are large than 400000
city_type = rep(0, nrow(apartments))
apartments = cbind(apartments, city_type)
apartments = na.omit(apartments)

for (i in 1:nrow(apartments)){
   apartments$city_type[i] = "suburbs"
   for (j in c("Los Angeles","San Diego","San Jose","San Francisco",
                    "Fresno","Sacramento","Long Beach","Oakland")){
      if (apartments$city[i] == j) {apartments$city_type[i] = "major city"}
   }
}

table(apartments$city_type)

# city type VS famliy friendly
h1 = ggplot(apartments, aes(x = as.factor(bedrooms) , fill = city_type)) +
```

```
  geom_bar(position = "dodge")+labs(title = "Number of Bedrooms by City Type", x = "Number of Bedrooms")
+
  theme(axis.ticks.x = element_blank())+theme(plot.title=element_text(hjust=0.5))+
  scale_fill_manual(values = c("darkorange","lightskyblue2"))+
  theme(legend.background = element_blank(),legend.justification=c(1,1), legend.position=c(1, 1))+
  geom_text(stat='count', aes(label=..count..), vjust=-0.2,position = position_dodge(width = 1))

h2 = ggplot(apartments, aes(x = as.factor(pets) , fill = city_type)) +
  geom_bar(position = "dodge")+labs(title = "Pets Allowed by City Type", x = "Pets Allowed") +
  theme(axis.ticks.x = element_blank())+theme(plot.title=element_text(hjust=0.5))+
  scale_fill_manual(values = c("darkorange","lightskyblue2"))+
  theme(legend.background = element_blank(),legend.justification=c(1,1), legend.position=c(1, 1))+
  geom_text(stat='count', aes(label=..count..), vjust=-0.2,position = position_dodge(width = 1))

h3 = ggplot(apartments, aes(x = as.factor(parking) , fill = city_type)) +
  geom_bar(position = "dodge")+labs(title = "Parking Policy by City Type", x = "Parking Policy") +
  theme(axis.ticks.x = element_blank())+theme(plot.title=element_text(hjust=0.5))+
  scale_fill_manual(values = c("darkorange","lightskyblue2"))+
  theme(legend.background = element_blank(),legend.justification=c(1,1), legend.position=c(1, 1))+
  geom_text(stat='count', aes(label=..count..), vjust=-0.2, position = position_dodge(width = 1))



h4 = ggplot(apartments, aes(x = as.factor(laundry) , fill = city_type)) +
  geom_bar(position = "dodge")+labs(title = "Laundry Situation by City Type", x = "Laundry Situation") +
  theme(axis.ticks.x = element_blank())+theme(plot.title=element_text(hjust=0.5))+
  scale_fill_manual(values = c("darkorange","lightskyblue2"))+
  theme(legend.background = element_blank(),legend.justification=c(1,1), legend.position=c(1, 1))+
  geom_text(stat='count', aes(label=..count..), vjust=-0.2, position = position_dodge(width = 1))

grid.arrange(h1,h2,h3,h4,nrow = 2)

# rent VS beadrooms/bathrooms
h5 = ggplot(apartments, aes(x = bedrooms, y = price)) + geom_point(size = 3,color = 'tomato1')+
  labs(title = "Relation of The Number of Bedrooms and Apartments Rent", x= "The Number of Bedrooms")+
  theme(plot.title=element_text(hjust=0.5))
h6 = ggplot(apartments, aes(x = bathrooms, y = price)) + geom_point(size = 3.5,color = 'lightslateblue')+
  labs(title = "Relation of The Number of bedrooms and Apartments Rent", x = "The Number of
Bathrooms",y="")+
  theme(plot.title=element_text(hjust=0.5))
grid.arrange(h5,h6,nrow = 1)

# LA, San Diego, San Jose
LA_SD_SJ = apartments[apartments["city"] %in% c("Los Angeles", "San Diego", "San Jose"),]
```

```
h7 = ggplot(LA_SD_SJ, aes(x = as.factor(bedrooms), fill = city)) +
  geom_bar( position = "dodge")+labs(title = "Number of Bedrooms by City Type", x = "Number of Bedrooms")
+
  theme(axis.ticks.x = element_blank())+theme(plot.title=element_text(hjust=0.5))+
  scale_fill_manual(values = c("khaki","lightskyblue2","mediumaquamarine"))+
  theme(legend.background = element_blank(),legend.justification=c(1,1), legend.position=c(1, 1))+
  geom_text(stat='count', aes(label=..count..), vjust=0.5, position = position_dodge(width = 1))

h8 = ggplot(LA_SD_SJ, aes(x = as.factor(bathrooms), fill = city)) +
  geom_bar( position = "dodge")+labs(title = "Number of Bathrooms by City Type", x = "Number of
Bathrooms") +
  theme(axis.ticks.x = element_blank())+theme(plot.title=element_text(hjust=0.5))+
  scale_fill_manual(values = c("khaki","lightskyblue2","mediumaquamarine"))+
  theme(legend.background = element_blank(),legend.justification=c(1,1), legend.position=c(1, 1))+
  geom_text(stat='count', aes(label=..count..), vjust=0.5, position = position_dodge(width = 1))

h9 = ggplot(LA_SD_SJ, aes(x = as.factor(parking), fill = city)) +
  geom_bar( position = "dodge")+labs(title = "Parking Policy by City Type", x = "Parking Policy") +
  theme(axis.ticks.x = element_blank())+theme(plot.title=element_text(hjust=0.5))+
  scale_fill_manual(values = c("khaki","lightskyblue2","mediumaquamarine"))+
  theme(legend.background = element_blank(),legend.justification=c(1,1), legend.position=c(1, 1))+
  geom_text(stat='count', aes(label=..count..), vjust=0.5, position = position_dodge(width = 1))

h10 = ggplot(LA_SD_SJ, aes(x = as.factor(laundry), fill = city)) +
  geom_bar( position = "dodge")+labs(title = "Laundry Situation by City Type", x = "Laundry Situation") +
  theme(axis.ticks.x = element_blank())+theme(plot.title=element_text(hjust=0.5))+
  scale_fill_manual(values = c("khaki","lightskyblue2","mediumaquamarine"))+
  theme(legend.background = element_blank(),legend.justification=c(0,1), legend.position=c(0, 1))+
  geom_text(stat='count', aes(label=..count..), vjust=0.5, position = position_dodge(width = 1))

h11 = ggplot(LA_SD_SJ, aes(x = as.factor(pets), fill = city)) +
  geom_bar( position = "dodge")+labs(title = "Pets Allowed by City Type", x = "Pets Allowed") +
  theme(axis.ticks.x = element_blank())+theme(plot.title=element_text(hjust=0.5))+
  scale_fill_manual(values = c("khaki","lightskyblue2","mediumaquamarine"))+
  theme(legend.background = element_blank(),legend.justification=c(1,1), legend.position=c(1, 1))+
  geom_text(stat='count', aes(label=..count..), vjust=0.5, position = position_dodge(width = 1))

h12 = ggplot(LA_SD_SJ, aes(x=city, y=price, fill=city)) +
  geom_violin(alpha = 0.4,size = 0.8) +geom_boxplot(width=0.3, fill="white",size = 0.8)+
  labs(title = "Apartments Rent by City Type", x = "", y = "Apartments     Rent") +
  scale_fill_manual(values = c("khaki","lightskyblue2","mediumaquamarine"))+
  stat_summary(geom="text", fun.y=quantile,
               aes(label=sprintf("%1.1f", ..y..)),
               position=position_nudge(x=0.23), size=5.5)+theme(plot.title=element_text(hjust=0.5))+
```

```
  theme(legend.background = element_blank(),legend.justification=c(1,1), legend.position=c(1, 1))

h13 = ggplot(LA_SD_SJ, aes(x=city, y=sqft, fill=city)) +
  geom_violin(alpha = 0.4,size = 0.8) +geom_boxplot(width=0.3, fill="white",size = 0.8)+
  labs(title = "Apartments Size by City Type", x = "", y = "Apartments     Size") +
  scale_fill_manual(values = c("khaki","lightskyblue2","mediumaquamarine"))+
  stat_summary(geom="text", fun.y=quantile,
               aes(label=sprintf("%1.1f", ..y..)),
               position=position_nudge(x=0.23), size=5.5)+theme(plot.title=element_text(hjust=0.5))+
  theme(legend.background = element_blank(),legend.justification=c(1,1), legend.position=c(1, 1))

grid.arrange(h7,h8,h9,h11,h10)
grid.arrange(h12,h13,nrow = 1)

### CA map
library(maps)
us_states = map_data ("state")
ca_df = subset(us_states, region == "california")
counties = map_data("county")
ca_county = subset(counties, region == "california")

ca_base = ggplot(ca_df, mapping = aes(x = long, y = lat, group = group)) + coord_fixed(1.3) +
  geom_polygon(color = "black", fill = "gray")

# category three region
Three_Region = rep(0, nrow(ca_county))
ca_county = cbind(ca_county, Three_Region)

for (i in 1:nrow(ca_county)){
  ca_county$Three_Region[i] = "Northern California"
  for (j in c("mono","madera","fresno","kings","tulare",
              "inyo","kern","san bernardino","riverside","orange","san diego","imperial")){
    if (ca_county$subregion[i] == j) {ca_county$Three_Region[i] = "Southern California"}
  }
  for (k in (c("monterey","san benito","san luis obispo","santa barbara","ventura","los angeles"))){
    if (ca_county$subregion[i] == k) {ca_county$Three_Region[i] = "California"}
  }
}

p = ca_base +   geom_polygon(data = ca_county, aes(fill = Three_Region), color = "white") +
  geom_polygon(color = "black", fill = NA) + theme(axis.text = element_blank(),axis.line = element_blank(),
    axis.ticks = element_blank(),rect = element_blank(), axis.title.x = element_blank(),axis.title.y =
element_blank())+
  scale_fill_manual(values = c("California"="dodgerblue4",
```

```
                                    "Northern California"= "skyblue3",
                                    "Southern California"="lightskyblue1"))
# add points
apartments = na.omit(apartments)
h1 = p + geom_point(data = apartments, aes(x =longitude, y = latitude , color = price ),inherit.aes = FALSE)+
  labs(title = "Apartments Rent by Region")+theme(plot.title=element_text(hjust=0.5,size = 15))+
  theme(legend.background = element_blank(),legend.justification=c(0,0), legend.position=c(0, 0))+
  scale_color_gradient(low="#E1FA72", high="#F46FEE")

h2 = p + geom_point(data = apartments, aes(x =longitude, y = latitude , color = sqft ),alpha = 0.1,inherit.aes =
FALSE)+
  labs(title = "Apartments Size by Region")+theme(plot.title=element_text(hjust=0.5,size = 20))+
  theme(legend.background = element_blank(),legend.justification=c(0,0), legend.position=c(0, 0))+
  scale_color_gradient(low="#E1FA72", high="#F46FEE")

grid.arrange(h1,h2,nrow = 1)
h1
##### Equally divide into 10 groups by rent
group_rent = rep(0, nrow(apartments))
apartments = cbind(apartments, group_rent)
breaks = quantile(apartments$price, probs = seq(0, 1, 0.1), name = FALSE)

for(i in 1:nrow(apartments)){
  for(j in 1:10){
    if(apartments$price[i] >= breaks[j] & apartments$price[i] <= breaks[j+1]){apartments$group[i] = j}
  }
}

rent = apartments[apartments[,"group"] %in% c("10","1"),]

table(rent_top_ten$bedrooms)
table(rent_top_ten$bathrooms)
table(rent_top_ten$parking)
table(rent_top_ten$pets)
table(rent_top_ten$laundry)
table(rent_top_ten$city)
table(rent_top_ten$county)

# define top 10th rent apartments as high rent house, low 10th rent as low rent house

p1 = ggplot(rent, aes(x = as.factor(pets), fill = as.factor(group))) +
  geom_bar( position = "dodge")+labs(title = "Pets Allowed by Apartments Rent", x = "") +
  theme(axis.ticks.x = element_blank())+theme(plot.title=element_text(hjust=0.5))+
  theme(legend.background = element_blank(),legend.justification=c(1,1), legend.position=c(1, 1))+
```

```
  geom_text(stat='count', aes(label=..count..), vjust=0.5, position = position_dodge(width = 1))+
  scale_fill_manual(labels = c("Low", "High"),values = c("khaki","lightskyblue2"))+
  guides(fill = guide_legend(title = "Apartments Rent"))

p2 = ggplot(rent, aes(x = as.factor(laundry), fill = as.factor(group))) +
  geom_bar( position = "dodge")+labs(title = "Laundry Situation by Apartments Rent", x = "") +
  theme(axis.ticks.x = element_blank())+theme(plot.title=element_text(hjust=0.5))+
  theme(legend.background = element_blank(),legend.justification=c(0,1), legend.position=c(0, 1))+
  geom_text(stat='count', aes(label=..count..), vjust=0.5, position = position_dodge(width = 1))+
  scale_fill_manual(labels = c("Low", "High"),values = c("khaki","lightskyblue2"))+
  guides(fill = guide_legend(title = "Apartments Rent"))

p3 = ggplot(rent, aes(x = as.factor(parking), fill = as.factor(group))) +
  geom_bar( position = "dodge")+labs(title = "Parking Policy by Apartments Rent", x = "") +
  theme(axis.ticks.x = element_blank())+theme(plot.title=element_text(hjust=0.5))+
  theme(legend.background = element_blank(),legend.justification=c(1,1), legend.position=c(1, 1))+
  geom_text(stat='count', aes(label=..count..), vjust=0.5, position = position_dodge(width = 1))+
  scale_fill_manual(labels = c("Low", "High"),values = c("khaki","lightskyblue2"))+
  guides(fill = guide_legend(title = "Apartments Rent"))

library(plyr)
mu = ddply(rent, "group", summarise, grp.mean=mean(sqft))   # calculate the mean of different group
head(mu)

p4 = ggplot(rent, aes(x = sqft, fill = as.factor(group))) +
  geom_density(alpha = 0.5, size = 1)+labs(title = "Apartments Size by Apartments Rent", x = "") +
  theme(axis.ticks.x = element_blank())+theme(plot.title=element_text(hjust=0.5))+
  theme(legend.background = element_blank(),legend.justification=c(1,1), legend.position=c(1, 1))+
  scale_fill_manual(labels = c("Low", "High"),values = c("khaki","lightskyblue2"))+
  guides(fill = guide_legend(title = "Apartments Rent"))+
  geom_vline(data=mu, aes(xintercept=grp.mean),linetype="dashed") + xlim(0,3500)

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

# correlation of size and rent
ggplot(apartments,aes(y = sqft)) + geom_boxplot() + ylim(0,3000)

# remove outliers
library(data.table)
outlierReplace = function(dataframe, cols, rows, newValue = NA) {
  if (any(rows)) {
    set(dataframe, rows, cols, newValue)
  }
}
```

```
outlierReplace(apartments, "sqft",
                       which(apartments$sqft > 5000), NA)
apart = na.omit(apartments)

ggplot(apart, aes(sqft,price)) + geom_point() + geom_smooth(method='loess') +
  labs(title="Relationship of Apartments Size and Rent",x="Apartments Size",y="Apartments Rent")+
  theme(plot.title=element_text(hjust=0.5,size = 18))
cor(apart$sqft,apart$price,use = "complete.obs")

# split size in quartile
group = rep(0, nrow(apart))
apart = cbind(apart, group)

breaks = quantile(apart$sqft, probs = seq(0, 1, 0.2), na.rm = TRUE)

for(i in 1:nrow(apart)){
  for(j in 1:5){
    if(apart$sqft[i] >= breaks[j] & apart$sqft[i] <= breaks[j+1]){apart$group[i] = j}
  }
}
ggplot(apart) +geom_jitter(aes(sqft,price)) + geom_smooth(aes(sqft,price), method='loess', se=FALSE) +
  facet_wrap(~group, scales="free_x") +
  labs(title="Apartments Size VS Rent in Different Size Quantiles",x="Apartments Size",y="Apartments
Rent")+
  theme(plot.title=element_text(hjust=0.5,size = 18))
####
ap = na.omit(apartments)
apart5 = apart[apart[,"group"] == '5',]
cor(apart5$sqft,apart5$price,use = "complete.obs")

# Make a report-worthy plot
apartments = readRDS("cl_apartments.rds")
apartments[which(apartments$price == max(apartments$price, na.rm = TRUE)), ]
apartments$price[apartments$price >= 30000000] = 3408
apartments[which(apartments$price == max(apartments$price, na.rm = TRUE)), ]
apartments$price[apartments$price >= 9900000] = 995
apartments$sqft[apartments$sqft >= 5000] = NA
big_cities = c('San Francisco', 'Oakland', 'San Jose', 'Sacramento', 'Los Angeles', 'San Diego')
# Use the %in% operator
apt_big6 = apartments[apartments$city %in% big_cities, ]

apt_big6$city = factor(apt_big6$city)
ggplot(apt_big6, aes(x = sqft, y = price)) + geom_point() + geom_smooth(method='lm')+ facet_wrap(.~city,
nrow = 2)+
```

```
  labs(title="Relation of Apartments Rent and Size in Big Cities",x="Apartments Size",y="Apartments Rent")+
  theme(plot.title=element_text(hjust=0.5,size = 18))

#### by craigslist
ggplot(apartments, aes(x=craigslist,y = price,color = craigslist)) + geom_boxplot(width=0.2, fill="white",size =
0.8) +
  geom_violin(alpha = 0.4,size = 0.8) + ylim(0,17700)+
  theme(legend.position="none")+ theme(plot.title=element_text(hjust=0.5,size = 18))+
  labs(title="Apartments Rent by Craigslist Branch",x="",y="Apartments Rent")

## Mld Apartments rent
data1 = apartments[apartments[,'bedrooms']%in%c('1','2'),]
data2 = data1[data1[,'bathrooms']%in%c('1','1.5'),]
data3 = data2[data2[,'parking']%in%c('covered','street','paid'),]
data4 = data3[data3[,'laundry']%in%c('hookup','shared'),]
data5 = data4[data4$city %in% big_cities, ]

g1 = ggplot(data5, aes(x=city,y = price,color = city)) + geom_boxplot(width=0.2, fill="white",size = 0.8) +
  theme(legend.position="none")+ theme(plot.title=element_text(hjust=0.5,size = 18))+
  labs(title="Mid-Apartments Rent in Big Cities",x="",y="Apartments Rent")+
  stat_summary(geom="text", fun.y=median,
               aes(label=sprintf("%1.1f", ..y..)),
               position=position_nudge(x=0.3), size=5.5)

## superior quality Apartments rent
data1 = apartments[apartments[,'bedrooms']%in%c('3','4','5'),]
data2 = data1[data1[,'bathrooms']%in%c('2.5','3.5','3','4'),]
data3 = data2[data2[,'parking']%in%c('valet','garage'),]
data4 = data3[data3[,'laundry']%in%c('in-unit'),]
data5 = data4[data4$city %in% c('San Francisco', 'San Jose', 'Los Angeles', 'San Diego'), ]

g2 = ggplot(data5, aes(x=city,y = price,color = city)) + geom_boxplot(width=0.2, fill="white",size = 0.8) +
  theme(legend.position="none")+ theme(plot.title=element_text(hjust=0.5,size = 18))+
  labs(title="Superior Quality Apartments Rent in Big Cities",x="",y="Apartments Rent")+
  stat_summary(geom="text", fun.y=median,
               aes(label=sprintf("%1.1f", ..y..)),
               position=position_nudge(x=0.23), size=5.5)
grid.arrange(g1,g2)

# the number of missing values in each col
colSums(is.na(apartments))
```