# STA141A Homework 3

Aleksandra Taranov
Statistics Department
University of California, Davis

October 29, 2018

On October 16th, 2018, students at UC Davis taking the course STA141A Statistical Data Science were given a dataset of recent Craigslist posts for apartment rentals in California that had been downloaded the previous day. Craigslist is a website where people can post classifieds for free, and their subjects include housing rentals, item sales, and ride share services, among other things (Craigslist,2018). As this data is extremely messy and sometimes unreliable, a significant portion of the assignment involves cleaning the data. Following this step, exploratory data analysis is used to determine trends and interpretations of the data. The software used is R and Rstudio. This document was written in texstudio on a laptop running a fedora distribution.

## 1 An Overview of the Data

### 1.1 Units of Study and Duplicate Removal

The original dataset that we are given has 21,948 rows, each representing what is presumably a housing post on craigslist. However, because there were a significant amount of duplicates, the data was initially cleaned and subsetted in several steps. The first step involved removing all posts that had identical longitude and latitude to another post. At this point we had 8783 unique posts and removed 13,165 duplicates. However, because longitude and latitude sometimes varied by small amounts, we further cleaned the data by examining the first 30 characters of text following the phrase ”QR code link to this post”. This resulted in us removing 1784 more rows, leaving 6999 unique apartment rentals on the market. This approach was not perfect and there may be missed duplicates that had slightly different text and location or false positives. However, text was used rather than title because prices were sometimes changed in titles but text seemed to be consistent in the duplicates in most cases. Therefore, initial removal of duplicate posts resulted in a subsetted dataset with 6999 rows presumably representing 6999 unique apartment rentals. It is important to note that we did not use a consistent method for sorting which of the duplicates were removed rather than left in and that this could have introduced bias. We simply left in the first one that appeared and it’s possible that duplicates had different prices, for instance, if the same post was re-posted with a lower place at a later date.

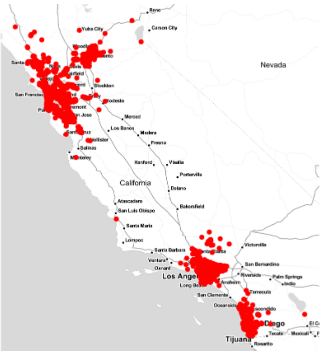
## 1.2 Column Categories

This original dataset had 20 columns, which correspond to 20 different types of information gathered about each of these posts, which are listed: title, bedrooms, text, bathrooms, latitude, pets, longitude, laundry, city_text, parking, date_posted, craigslist, date_updated, place, price, city, deleted, state, sqft, and county. According to metadata given by instructor Nick Ulle, most categories were extracted from craigslist, with the exception of pets, laundry, parking, place, city, state, and county. These were extracted from the text and therefore are less likely to be reliable. These categories that were extracted from the text correspond to pet policy, type of laundry provided, parking policy, and then a variety of location categories based off of the latitude and longitude entered by the user who made the post. The categories that were more directly downloaded from the posts include the post's title, text, latitude, longitude, city, date posted, date updated, price, whether it's been deleted, size in square ft, number bedrooms, number of bathrooms, and which craigslist region it was posted in. Because much of this data is self-reported by the user making the post, it must be viewed cautiously.

## 1.3 Date Range

The dataset was downloaded on October 15, 2018. The date range for the data before subsetting was from September 8, 2018 at about 5pm to October 15th, 2018 at about 3pm. However, our subsetted data that had the 6999 unique posts ranged from September 10,2018 at about noon to October 15, 2018 at about 3pm. Therefore, we can estimate that the data goes from roughly mid-September to October 15th, covering a little over a month of posts.

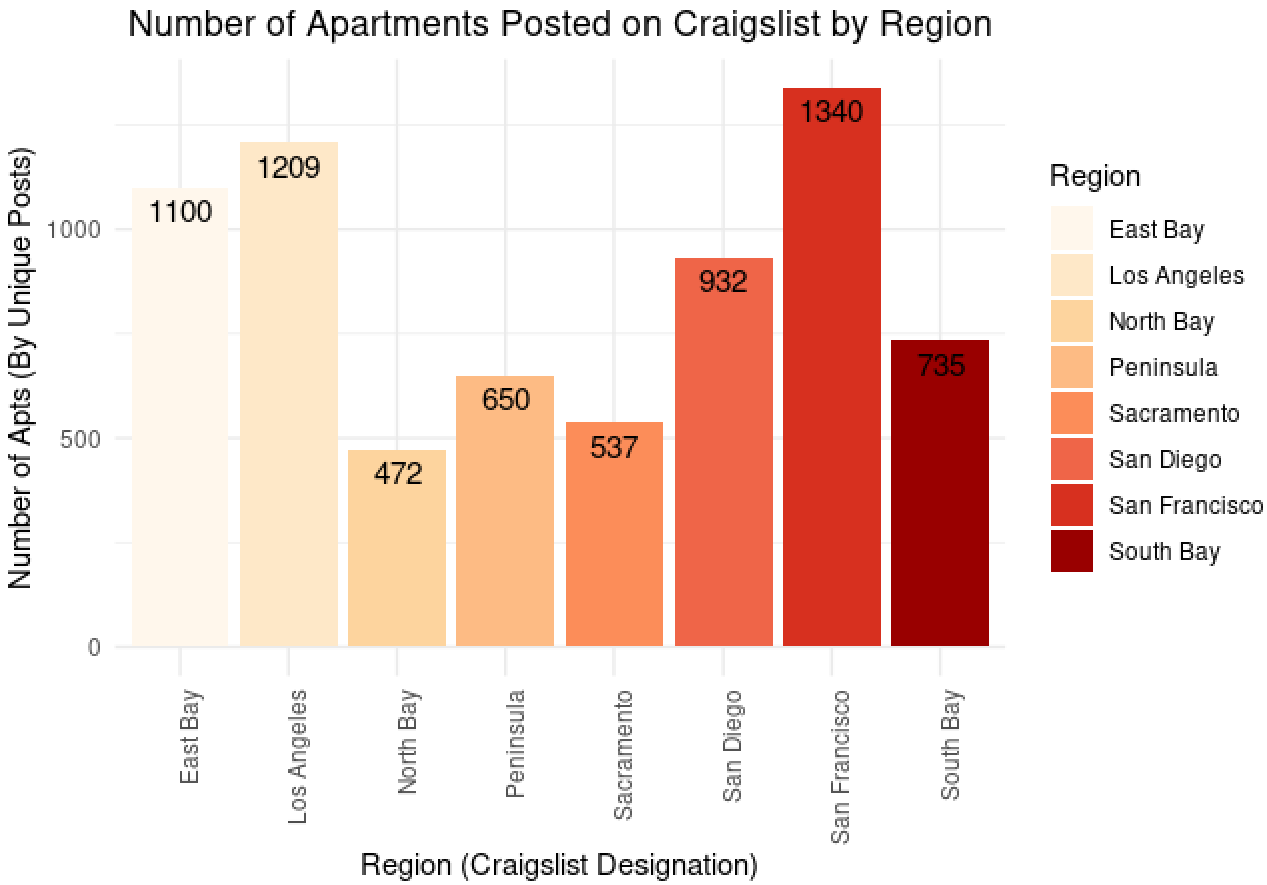
## 1.4 Location



Our initial analysis of location based on the 6999 posts that had been presumed unique actually showed that the following states were represented in our data: "CA" NA "CT" "WA" "NV" "NC" "OH" "VA" "UT" "FL" "MD". However, since we were intending to look at California date and since there were 6975 from California and only 24 that appeared to be from other states, these were thrown out as presumably erroneous posts. Of course, the state was determined based on the user inputting latitude and longitude, but the locations weren't clear from the title or text and therefore these 24 posts were removed from the data, leaving 6975 unique posts from California.

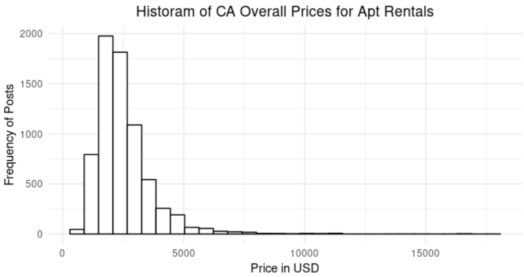
These were then mapped using ggmap (David Kahle and Hadley Wickham, 2013.) This gave us a visual representation of where the posts in our data were coming from. However, we wanted to get better numbers and really examine the location range.

## 1.5 Location Range



To get a sense of the range of locations, the data was sorted based on the craigslist category, which was presumed to be more accurate than the latitude and longitude provided by the user. The 6975 posts that were presumed to be unique apartment rentals in California were grouped by location, as plotted. There are far more posts in San Francisco and the surrounding Bay Area than Los Angeles, San Diego, or Sacramento. The reason for this could be that Craigslist originally started in SF or that in general San Francisco happens to be a city with strong tech culture where people use websites like craigslist more than in other places. We could even hypothesize that once a critical mass is reached on a website like craigslist, people find it useful and therefore use it more, creating an effect where you get steady use once it gets past a certain number of users. In any case, this helped us determine which regions were represented in the post and allowed us to better examine the data by region. Note: This graph above uses a color palette that makes it printer friendly in either color or greyscale.

## 1.6 Price Range

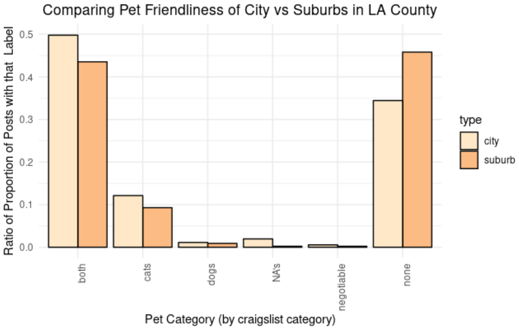


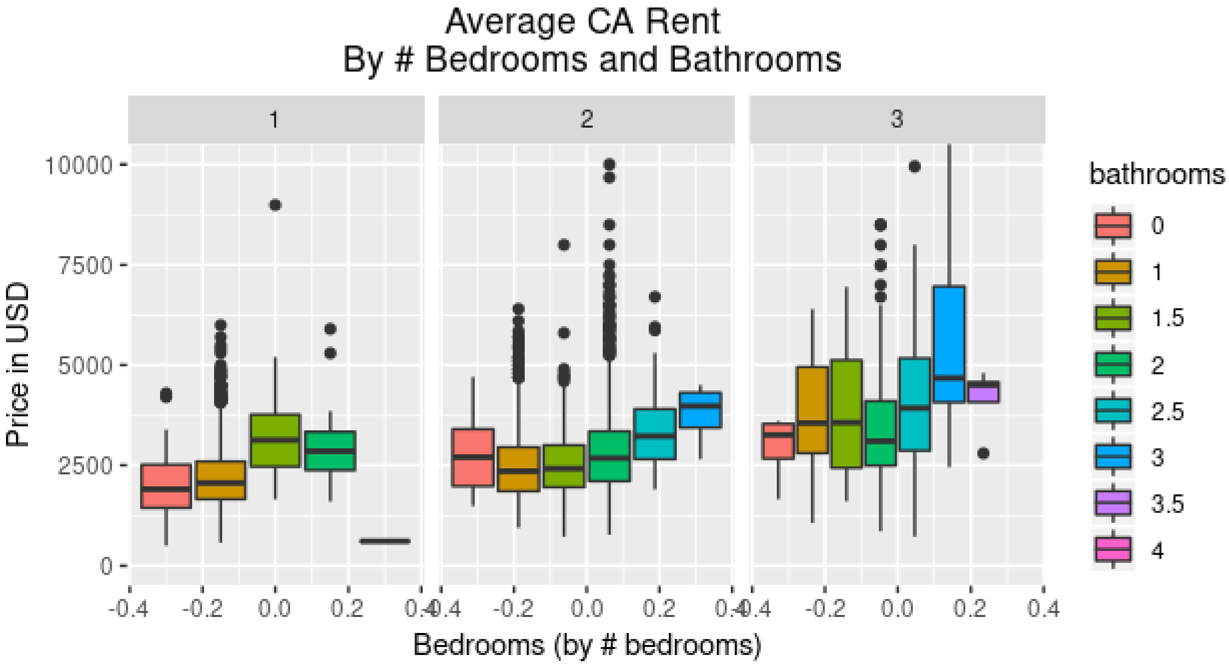
When we examined the price range, we found prices that ranged from $0 USD to $34,083,742. When these individual posts were closely examined, it appeared that the small handful of high values were listings to purchase houses rather than rent and that many of the items listed below about $500 were either daily or weekly prices or item sales that had been erroneously posted in the monthly apartment rental section of the website. Therefore, the data was further subsetted for prices between $500 and $20,000 per month, which are both plausible, based on region and number of bedrooms. The higher range tended to be 4 or 5 bedrooms in cities and the lower range tended to occur in areas that weren't prominent cities. Our new subsetted data had 6934 observations, and this is the dataset that was then used for the analysis of data in the rest of the report. The price range for this is shown in the histogram and goes from $504 to $17,700, with a mean of $2517 with standard deviation $1238 and a median on $2288. Therefore, a monthly rental unit in California seems to be on average about $2517 but there is quite a range.

# 2 Analysis of Data

## 2.1 Family Friendliness by Pets

Our initial question was whether suburbs or cities were more likely to allow pets. To determine this, we chose to examine Los Angeles and group our data into the actual city of Los Angeles vs other towns in the same county but outside of the bounds of the city. There were 717 entries in the city and 441 in the suburbs. We looked at the ratio of how many allowed pets out of total for each city and plotted them. The result was very surprising and went against my expectations. For LA, it seemed that the city had a higher ratio of posts that allowed animals in general and therefore by our metric were more family friendly.





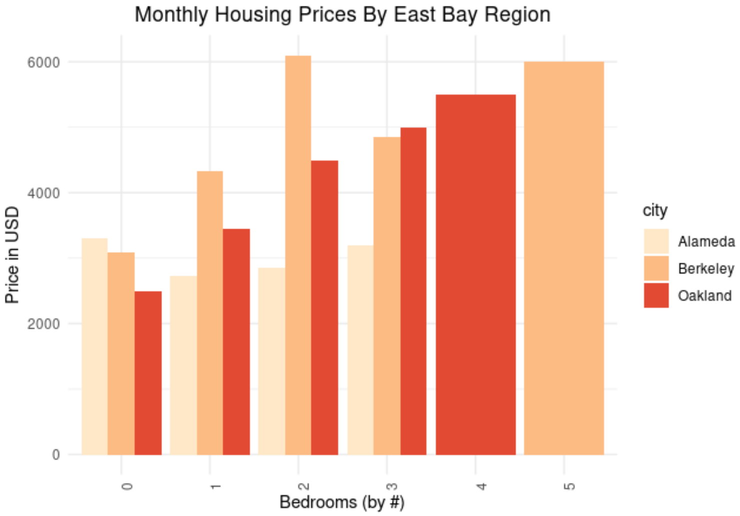
## 2.2 Pricing Based on Bedrooms and Bathrooms

Our next question was whether prices increased more quickly due to bathrooms or bedrooms. To do so, we fit a linear model through the data. The correlation coefficients for both were low (41.0% for bathrooms and 44.3% for bedrooms), suggesting that linear fit may not be the best regression. However, we decided to use this metric anyway. Our linear model was Y=807.3X1 + 507.7X2. Bedrooms had a higher slope (807.3 vs 507.7) and slope corresponds to how quickly the prices increases with each additional bedroom or bathroom, respectively. Therefore, bedrooms add more to price than bathrooms. Additionally, we present plots to illustrate this on the following page.

As shown in the following plots, prices for 1,2,and 3 bedroom apartments increase with additional bathrooms. Results for 4,5, and 6 bedrooms were not included because there were very few data points for those. In the plots, the data varies slightly by number of bedrooms and you can see that the biggest distinction seems to be in 1 bedrooms between 1 bathroom versus 1.5 or 2 bathrooms. For 2 bedrooms, there are slight increases in price based on 1 through 2 bedrooms and then higher price increases once you get to 2.5 or 3 bathrooms. For 3 bedrooms, the data is more variable and 1 and 1.5 bathrooms gives comparable prices, while 2 appears lower and 2.5 to 3.5 bathrooms give higher prices.

## 2.3 Characteristics Based on Geographic Location

Next we compared prices of living in Oakland, Berkeley, and Alameda. We assume that listings with 0 bedrooms were erroneous. However, for 1 bedrooms, as you can see on the following plot, the price is highest for Berkeley and lowest for Alameda, with Oakland, in the middle. This same pattern occurs for 2 bedrooms and is even more extreme. However, when you get to 3 bedrooms, the prices for Oakland are slightly above Berkeley, although comparable. Therefore, for 1 and 2 bedroom apartments, Berkeley is more expensive than Oakland and Alameda, but for 3 bedrooms, Berkeley and Oakland are comparable and still much more than Alameda. We do not have data for 4 and 5 bedrooms in all 3 areas to compare them. Additionally, in doing this analysis, we removed one more outlier - a sailboat that was being sold for $16,500, since that is not a monthly rental fee.



# 3 Further Questions

## 3.1 My Questions

The following represent questions I had about the data: Questions:

1) Is it easier to find an apt with parking in Oakland or Berkeley? This question is meaningful to someone who is considering moving to Oakland or Berkeley and wants to know whether one of the cities will be an easier place to find an apartment that offers parking. Additionally, this could also be helpful for public municipalities that are studying demand for parking in order to decide whether to build more parking for their cities.

2) How does average rental price compare to the average salaries in that region? This question is meaningful to a sociologist or housing activist who might want to point out whether the rent is skyrocketing above what average people can afford. This could signal gentrification and a housing crisis and even predict an increase of homeless. Therefore, it might be useful for homelessness activists and agencies as well to predict how many resources they may need to help mitigate this.

3) How does the price per sqft compare between San Francisco and Oakland? This is useful for someone deciding to move to San Francisco or Oakland who wants to know how big of an apartment they could afford in each.

4) Which places are the most pet friendly in all of California? This is meaningful to someone who wants to move to California with a pet and wants to know which locations would allow for that.

5) Are there regions where laundry is not commonly available with your housing? This ques-

tion is relevant to laundry businesses, especially pick up laundry and delivery, which might want to know which regions to target with their business.

6-7) How many posts get deleted as opposed to just expiring? What proportion of posts get updated? These two questions are relevant to people who work at craigslist and need to figure out when and which posts to delete so that the information on their website is up to date and usable and not expired.

8-8.5) Are apartments in SF that allow dogs generally more or less expensive? This is useful for someone who is thinking of getting a dog but worried that it might make living in San Francisco more expensive. As a follow up, question 8.5 asks whether this changes if we account for the number of sqft in the apartment.

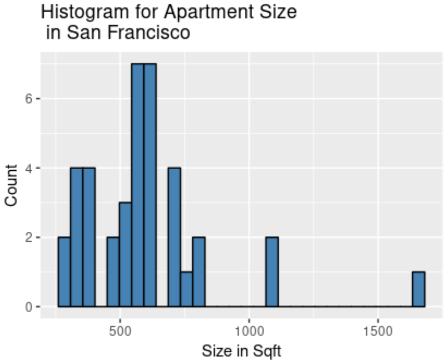
9) Is it easier to find a place with a cat or a dog in the city of LA? This is meaningful to a renter in LA who wants to decide whether to get a cat or dog depending on how much harder it will make it to find housing.

10) What size apartment can I expect in SF if my budget is 2000-2500 per month for a 1 bedroom? This is meaningful to me because I'm trying to find an apartment right now in San Francisco and want to know how much space I could expect.

11) At what point in the month do most people make posts? This is relevant to both sort out junk posts or item posts that don't belong and also to someone who wants to look for an apartment for the next month and is curious when the largest volumes of post for the next month come up.

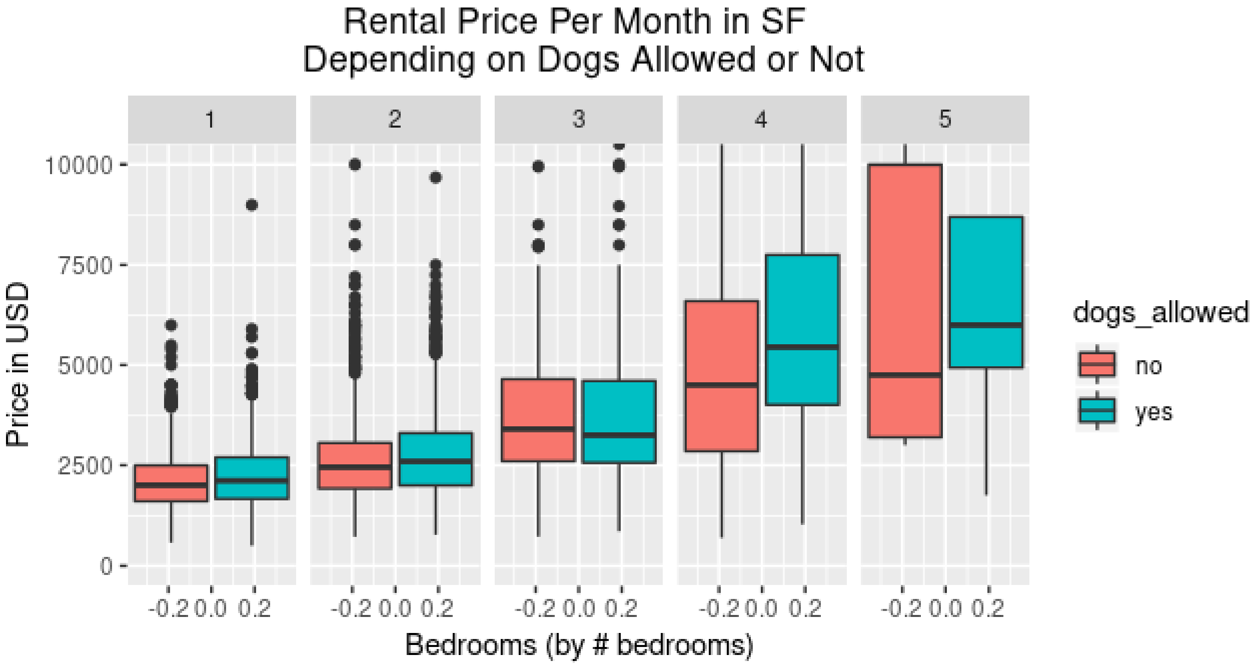
## 3.2 Question 1

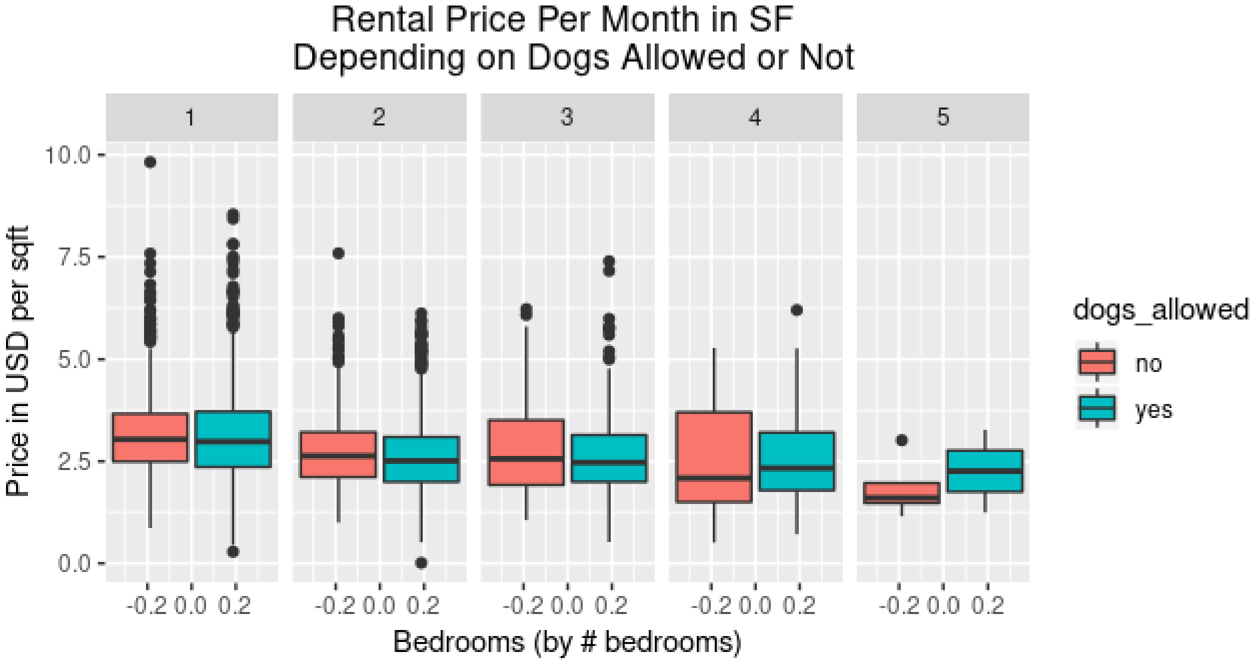
What size apartment can I expect in SF if my budget is 2000-2500 per month for a 1 bedroom? My guess is that I could find something that's very small - maybe 400 square feet? After doing the analysis, I see that the range is from 295 to 1668 square feet and the mean is 550 sqft with sd of 254. This can also be viewed as a histogram.



## 3.3 Question 2

Are apartments in SF that allow dogs generally more or less expensive? My guess is that they are because a slight extra charge is made in case the pet causes damage. Having done the analysis, I was correct and the average price in SF with a dog is $2360 and without one is $2200. However, this led to a follow up question which I answer next.





## 3.4 Question 3

In doing the above analysis, I realized that there might be a confounding factor. Is it possible that the reason that places that allow dogs are more expensive is just that these places tend to be bigger and/or have backyards? I suspected that this might be the case and therefore adjusted the data based on size in sqft. Lo and behold, I was correct again and when we accounted for size in square feet, the price difference disappeared, suggesting that size was a confounding factor.

### 3.5 Question 4

| | City | Laundry | Count |
|---|---|---|---|
| 1 | San Francisco | none | 230 |
| 2 | Oakland | none | 58 |
| 3 | Los Angeles | none | 36 |
| 4 | San Diego | none | 35 |
| 5 | Berkeley | none | 25 |
| 6 | Sacramento | none | 19 |
| 7 | San Jose | none | 17 |

Are there regions where laundry is not commonly available with your housing? I guessed that cities would be likely to have lots of apartments that don't have their own laundry services. I was right that it would be higher for cities but wrong about which ones. Therefore, a laundry pickup and delivery service might look at this and definitely decide not to expand to Sacramento or San Jose or Berkeley but to focus on spreading their business in San Francisco, Oakland, and LA.

### 3.6 Question 5

How many posts get updated? I thought that very few would get updated, maybe a quarter. I was wrong. In our subsetted data, there were 3460 NA values (not updated) and 3474 that were not NA (that did get updated.) However, in the original there were 13139 NA and 8809 did get updated. This makes sense since repetitive posts are unlikely to get updated. Instead of updating, the person may have made a new post instead. Or, those posts might be more likely to be scam. This would be useful in establishing an algorithm that automatically deletes spam posts.

## 4 Limitations and Conclusions

The table shows how many NA or missing values there were per column in the original data. However, I wanted to compare how that looked to my subsetted dataframe that I used for most of the analysis in which I deleted duplicates, outliers, and posts that were not monthly housing rentals. Thus I get the table on the right as well. As you can see, my cleaned up subset of the original dataframe happened to have fewer NAs in a variety of categories. This suggests that I managed to sort out many spam and erroneous datapoints with my cleaning method. All of the items in my subset had the core information that the user inputs into craigslist about title, text, latitude and longitude, and price.

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

| Number NA in Subset | Column |
|---:|---:|
| 0 | title |
| 0 | text |
| 0 | latitude |
| 0 | longitude |
| 0 | date_posted |
| 0 | price |
| 0 | deleted |
| 0 | craigslist |
| 0 | state |
| 0 | county |
| 0 | ID |
| 79 | laundry |
| 81 | place |
| 108 | parking |
| 143 | pets |
| 200 | city_text |
| 341 | bedrooms |
| 341 | bathrooms |
| 408 | city |
| 2467 | sqft |

Although this was discussed previously in the report, this subsetted data was achieved by deleting duplicates using location and text matching and then by deleting outliers and erroneous entries. For instance, we removed the $16500 sailboat because it was an item sale and not a monthly housing value. Additionally, when we looked at values below $500 most of them were either item sales or daily or weekly rates, which was not what we wanted to analyze. Additionally, as discussed earlier, we removed posts that did not have latitudes and longitudes in California because we could not identify what the correct locations for them should be or whether they had been mislabeled or placed in the wrong section.

This was very tricky to decide because craigslist does not verify these posts and all data is inputed by the user making the post. Therefore, it was hard to tell whether the user messed up the longitude or latitude or whether they posted in the wrong craigslist location. Overall, I assume there are many errors due to the users, in addition to the ones I found where, for example, items and sailboats were being sold in the section that was supposed to contain rental prices.

Overall, this was a very messy dataset. Once duplicates and erroneous posts were sorted out, we only had about 7,000 posts throughout California. This made it difficult to examine the suburbs vs city dynamics in places like Yolo County that had very few posts. For that reason, we focused our analysis on LA and the Bay Area. However, this introduces bias towards analyzing bigger cities into our data. Even though the original data had errors and biases and our approach introduced some as well, we still got some very interesting results from this data.

# 5 Citations

(Craigslist). Retrieved October 28, 2018, from https://sacramento.craigslist.org/search/apa

David Kahle and Hadley Wickham.Spatial Visualization with ggplot2, The R Journal, 2013, 5.1 pp 144-161, from http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

(Stack Overflow). Retrieved October 28, 2018, from https://stackoverflow.com/questions/10151123/how-to-specify-columns-in-facet-grid-or-how-to-change-labels-in-facet-wrap

Teetor, Paul.R Cookbook, 2011. 1st edition. O'Reilly Media, Inc.

(Tex Stackexchange). Retrieved October 29, 2018, from https://tex.stackexchange.com/questions/2832/how-can-i-have-two-tables-side-by-side

# 6 R Code

```
# Aleksandra Taranov
# 915501173
# Homework 3
# October 20

#Required libraries:
library(tidyverse)
library(ggplot2)
library(dplyr)
library(ggrepel)
library(gridExtra)
library(tidyverse)
library(pryr)
library(tidyr)
library(knitr)
library(xtable)
library(ggmap)
library(stringr)
library(lubridate)

#I downloaded the RDS file into my Downloads
setwd("/home/ataranov/Downloads")
```

```
#Reading in the dataset:
data <- readRDS("apartments")

#
   --------------------------------------------------------------------------

#QUESTION1:

#Exploratory Data Analysis (EDA)

#check out the dimensions and summary stats
dim(data) #21,948 by 20
str(data)
summary(data)
names(data)
head(data)

#make table to show column categories
vector <- names(data)
dataframe_names <- as.data.frame(vector)
dataframe_names_twocol <- data.frame(dataframe_names[1:10,],
   dataframe_names[11:20,])
names(dataframe_names_twocol) <- c("Column␣Categories","")
xtable(dataframe_names_twocol)


#look at just san francisco
data.sanfrancisco <- data %>%
select(title, text,city,price,latitude,longitude) %>%
arrange(city) %>%
filter(city=="San␣Francisco")
dim(data.sanfrancisco) #2893 by 3

#
   --------------------------------------------------------------------------

#However lots of the posts look like repetitions, so we need to
   see how many
#unique apartments there are.
```

```
#Since I want to use the text portion to match unique values,
   first remove QR Code Link to This Post from each one
data_edited <- data
data_edited$text <- gsub("QR␣Code␣Link␣to␣This␣Post","", data_
   edited$text)
data_edited$text <- gsub("[\r\n]", "", data_edited$text)
data_edited$text <- gsub("␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣", "", data_edited$
   text)

#make sure to change datetime format to character first to avoid
   errors about posix
data_edited$date_posted <- as.character(data_edited$date_posted)
data_edited$date_updated <- as.character(data_edited$date_updated
   )
class(data_edited$date_posted)

#Now remove all duplicates with identical longitude and latitude;
    first make new column with lat and longitude
data_sorted_location <- data_edited %>% mutate(latitude_longitude
    = paste0(round(data_edited$latitude,6),round(data_edited$
   longitude,6)))
length(unique(data_sorted_location$latitude_longitude)) #now
   there are 8783 unique entries
nrow(data_edited) - length(unique(data_sorted_location$latitude_
   longitude)) #and we identified 13165 duplicates

#Now compare this to values from duplicated to double check that
   we're doing it right
duplicated(data_sorted_location$latitude_longitude)
sum(duplicated(data_sorted_location$latitude_longitude)) #13165
   duplicates
sum(!duplicated(data_sorted_location$latitude_longitude)) #8783
   unique

#Now make new dataframe for sorted variables that have duplicate
   location or not
data_sorted_location_unique <- data_sorted_location[!duplicated(
   data_sorted_location$latitude_longitude), ]
data_sorted_location_repeats <- data_sorted_location[duplicated(
   data_sorted_location$latitude_longitude),]

#now do a secondary sort by text, since some of the longitudes
```

```
   and latitudes were slightly diff for same apt
first30char = str_sub(data_sorted_location_unique$text,1,30)
data_sorted_location_text <- data_sorted_location_unique %>%
   mutate(first30char = str_sub(data_sorted_location_unique$text
   ,1,30))
data_sorted_location_text_unique <- data_sorted_location_text[!
   duplicated(data_sorted_location_text$first30char), ]
data_sorted_location_text_repeats <- data_sorted_location_text[
   duplicated(data_sorted_location_text$first30char),]
sum(!duplicated(data_sorted_location_text$first30char)) #6999
   unique ones
sum(duplicated(data_sorted_location_text$first30char)) #removed
   1784 more duplicates with this method
#there may still be false positives or some duplicates not caught
    and we didn't specify which one we wanted to be the unique
   one

#now we have a dataframe with 6999 unique entries that i will
   remove the columns no longer needed
newdata <- data_sorted_location_text_unique
dim(newdata)
newdata <- subset(newdata, select=-c(latitude_longitude,
   first30char))
dim(newdata)

#Now I have a new subsetted dataset called new data with 6999
   unique apartments and 14949 duplicates were removed
nrow(data) - nrow(newdata)

#
   ------------------------------------------------------------------------

#explore the geographical range
#first look at which states are represented
unique(newdata$state)
#The states represented are "CA" NA   "CT" "WA" "NV" "NC" "OH" "
   VA" "UT" "FL" "MD"

#How many are in each state? try with table
table(newdata$state)

#Since 6975 are from California and a small handful are from the
```

```
   others let’s look at them individually
newdata_CT <- newdata %>% filter(state=="CT")
newdata_WA <- newdata %>% filter(state=="WA")

#From observing the CT and WA data, we see that they are from
   other states but were mistakenly posted in the sacramento or
   other CA craigslist site
#Going forward I’d like to just examine the data from CA, so I
   will filter out the CA ones and label them with an ID
newdata_CA <- newdata %>% filter(state=="CA")
newdata_CA <- newdata_CA %>% mutate(ID = 1:nrow(newdata_CA))
newdata_CA <- subset(newdata_CA, select=-c(APTID))
dim(newdata_CA)
names(newdata_CA)
#Now we have a new subsetted dataframe with 6975 unique
   Californian apartments to analyze

#
   -----------------------------------------------------------------

#save this newdata_CA dataframe to load later
save(newdata_CA,file="CAdata.Rda")
load("CAdata.Rda")

#Now I will sort by craigslist region, since this is more
   reliable than the scraped state or city info
table(newdata_CA$craigslist)
freq_by_region <- as.data.frame(table(newdata_CA$craigslist))
names(freq_by_region) <- c("craigslist","Apts")
freq_by_region <- freq_by_region %>% mutate(Region = c("Los␣
   Angeles","Sacramento","San␣Diego","East␣Bay","North␣Bay","
   Peninsula","South␣Bay","San␣Francisco"))

#cite craigslist since I went there to look up what the regions
   were https://sfbay.craigslist.org/

#make a barplot to show this data
plot1 <- ggplot(data=freq_by_region, aes(x=Region, y=Apts,fill=
   Region)) +
geom_bar(stat="identity")+
theme_minimal() +
geom_text(aes(label=Apts), vjust=1.6, color="black", size=4)+
scale_fill_brewer(palette="OrRd")+
labs(title="Number␣of␣Apartments␣Posted␣on␣Craigslist␣by␣Region",
    x ="Region␣(Craigslist␣Designation)", y = "Number␣of␣Apts␣(By
   ␣Unique␣Posts)")+
theme(plot.title = element_text(hjust = 0.5))+
theme(axis.text.x = element_text(angle = 90, hjust = 1))
plot1

#
   #------------------------------------------------------------------------


#Now let’s look at the range of time

#we’ll sort the data by date posted
names(newdata_CA)
sorted_dateposted <- newdata_CA[order(newdata_CA$date_posted),]
sorted_dateposted$date_posted[1:10]
sorted_dateposted$date_posted[(nrow(sorted_dateposted)-10):nrow(
   sorted_dateposted)]

#The dates of the unique posts range from September 10,2018 at
   about noon to October 15, 2018 at about 3pm.

#check on original data
sorted_dateposted <- data[order(data$date_posted),]
sorted_dateposted$date_posted[1:10]
sorted_dateposted$date_posted[(nrow(sorted_dateposted)-10):nrow(
   sorted_dateposted)]

#The date range for the data before subsetting was from September
    8, 2018 at about 5pm to October 15th, 2018 at about 3pm.

#
   -------------------------------------------------------------------------


#PRICES
#What is the range of prices, regardless of number bedrooms, etc?
temp <- newdata_CA %>% select(ID,price) %>% drop_na()
range(temp)
temp_over10000 <- temp %>% filter(price>10000)
temp_under300 <- temp %>% filter(price<300)
```

```
temp_under500 <- temp %>% filter(price<500)

#the values above 20000 all seem to be prices for full houses
   rather than monthly rent
#the values below about 500 tended to be daily or weekly rates or
    other misc sales

#Now we filter the CA data further to try to just get monthly
   rent
newdata_CA_monthly <- newdata_CA %>% filter(price>500 & price
   <20000)

#save this newdata_CA dataframe to load later
save(newdata_CA_monthly,file="CAdatamonthly.Rda")
load("CAdatamonthly.Rda")

dim(newdata_CA_monthly)

#Now we are looking at 6934 unique apartments to get a price
   range of monthly rent
range(newdata_CA_monthly$price)
hist(newdata_CA_monthly$price)
summary(newdata_CA_monthly$price)
sd(newdata_CA_monthly$price)
#The median value is $2288 for monthly rent with min of $504 and
   max of $17700.

#how make histogram for CA prices
plot2<-ggplot(newdata_CA_monthly, aes(x=price)) +
geom_histogram(color="black", fill="white") +
theme_minimal() +
scale_fill_brewer(palette="OrRd")+
labs(title="Historam␣of␣CA␣Overall␣Prices␣for␣Apt␣Rentals", x ="
   Price␣in␣USD", y = "Frequency␣of␣Posts")+
theme(plot.title = element_text(hjust = 0.5))
plot2

#
   ------------------------------------------------------------------------

#QUESTION2:
```

```
#Which adds more to rent: extra bedrooms, or extra bathrooms? How
    can you tell?
#Does the effect change as the number of bedrooms or bathrooms
   goes up?

#Let’s look at rent, extra bedrooms, and extra bathrooms
#We will use the subsetted dataframe for monthly CA rent
dim(newdata_CA_monthly) #examining 6934 unique entries

#create new column for rent PER bedroom
rent_per_bedroom <- newdata_CA_monthly %>%
drop_na(price,bedrooms) %>%
filter(bedrooms>0) %>%
mutate(price_per_bedroom = price/bedrooms)

#Now we have a dataframe with 5925 unique apts and price/bedroom
   that we can plot
dim(rent_per_bedroom)

#We will make a histogram to show price per bedroom
plot3<-ggplot(rent_per_bedroom, aes(x=price)) +
geom_histogram(color="black", fill="white") +
theme_minimal() +
scale_fill_brewer(palette="OrRd")+
labs(title="CA␣Price␣Per␣Bedroom", x ="Price␣in␣USD", y = "Count␣
   (in␣#␣apts)")+
theme(plot.title = element_text(hjust = 0.5))


plot3

summary(rent_per_bedroom$price_per_bedroom)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 174.8  1172.5  1597.5  1746.9  2195.0  8995.0

#Now make a boxplot to show range of price/bedroom based on
   number of bathrooms

#First change the bedrooms to factor rather than numeric
rent_per_bedroom$bathrooms <- as.factor(rent_per_bedroom$
   bathrooms)
rent_per_bedroom$bedrooms <- as.factor(rent_per_bedroom$bedrooms)
```

```
#now save this dataframe for loading later
save(rent_per_bedroom,file="rent_per_bedroom.Rda")
load("rent_per_bedroom.Rda")

#Now make the boxplot
plot4 <- ggplot(rent_per_bedroom, aes(x=bathrooms, y=price_per_
   bedroom,fill=bathrooms)) +
geom_boxplot() +
ggtitle("Average␣CA␣Rent␣Per␣Bedroom␣\n␣By␣#␣Bathrooms") +
theme(plot.title=element_text(hjust=0.5))
plot4

#Maybe boxplot wasn't the best way to view this. Let's try with
   grouping instead
plot5 <- ggplot(rent_per_bedroom, aes(y=price,fill=bathrooms)) +
geom_boxplot() +
ggtitle("Average␣CA␣Rent␣\n␣By␣#␣Bedrooms␣and␣Bathrooms") +
xlab("Bedrooms␣(by␣#␣bedrooms)") + ylab("Price␣in␣USD") +
theme(plot.title=element_text(hjust=0.5))+
facet_wrap(~bedrooms)
plot5

#subset so that the plot is only for the 1-3 bedroom places
rent_per_bedroom_subset <- rent_per_bedroom
rent_per_bedroom_subset$bedrooms <- as.numeric(rent_per_bedroom_
   subset$bedrooms)
rent_per_bedroom_subset <- rent_per_bedroom_subset %>% filter(
   bedrooms>0 & bedrooms <4)
range(rent_per_bedroom_subset$bedrooms)
rent_per_bedroom_subset$bedrooms <- as.factor(rent_per_bedroom_
   subset$bedrooms)

#now make this subsetted plot for 1-3 bedroom apartments
plot5_2 <- ggplot(rent_per_bedroom_subset, aes(y=price,fill=
   bathrooms)) +
geom_boxplot() +
ggtitle("Average␣CA␣Rent␣\n␣By␣#␣Bedrooms␣and␣Bathrooms") +
xlab("Bedrooms␣(by␣#␣bedrooms)") + ylab("Price␣in␣USD") +
theme(plot.title=element_text(hjust=0.5))+
coord_cartesian(ylim = c(0, 10000)) +
facet_wrap(~bedrooms)
```

```
plot5_2

#now extract the average values from our plot
ggplot_build(plot5)$data

#now use those extracted values to make a model
#First change the bedrooms back to numeric
rent_per_bedroom$bathrooms <- as.numeric(rent_per_bedroom$
   bathrooms)
rent_per_bedroom$bedrooms <- as.numeric(rent_per_bedroom$bedrooms
   )

#maybe i should have made a scatterplot instead of bedroom vs
   price and checked slope per number bathrooms
plot6 <- ggplot(rent_per_bedroom, aes(x=bedrooms, y=price)) +
geom_point() +
ggtitle("Average␣CA␣Rent␣\n␣By␣#␣Bedrooms␣and␣Bathrooms") +
xlab("Bedrooms␣(by␣#␣bedrooms)") + ylab("Price␣in␣USD") +
theme(plot.title=element_text(hjust=0.5))+
facet_wrap(~bathrooms)
plot6

#
   ------------------------------------------------------------------------

#use aggregate to get grouped values
aggregate(price ~ bedrooms+bathrooms, rent_per_bedroom, mean)

#construct linear model to examine strength of relationship
lm(rent_per_bedroom$price~rent_per_bedroom$bathrooms)
lm(rent_per_bedroom$price~rent_per_bedroom$bedrooms)
lm(rent_per_bedroom$price~rent_per_bedroom$bedrooms+rent_per_
   bedroom$bathrooms)
#this linear model shows that bedrooms affect the price more than
    bathrooms!

cor(rent_per_bedroom$bathrooms, rent_per_bedroom$price, method =
   c("pearson"))
cor(rent_per_bedroom$bedrooms, rent_per_bedroom$price, method = c
   ("pearson"))

#The correlation was low (41.0% for bathrooms and 44.3% for
   bedrooms).
#However, bedrooms had a higher slope (807.3 vs 507.7) and
   therefore add more to price than bathrooms

par(mfrow=c(1,2))
plot(rent_per_bedroom$bathrooms, rent_per_bedroom$price, main="CA
   ␣Rental␣Price␣as␣Function␣\n␣of␣Number␣of␣Bathrooms",
xlab="Bathrooms(by␣#)", ylab="Price␣in␣USD", pch=19)
#lines(lowess(colleges_and_population$V2, colleges_and_population
   $V1), col="darkgreen")
abline(lm(rent_per_bedroom$price~rent_per_bedroom$bathrooms), col
   ="orange")

plot(rent_per_bedroom$bedrooms, rent_per_bedroom$price, main="CA␣
   Rental␣Price␣as␣Function␣\n␣of␣Number␣of␣Bedrooms",
xlab="Bedrooms␣(by␣#)", ylab="Price␣in␣USD", pch=19)
#lines(lowess(colleges_and_population$V2, colleges_and_population
   $V1), col="darkgreen")
abline(lm(rent_per_bedroom$price~rent_per_bedroom$bedrooms), col=
   "orange")


#
   ------------------------------------------------------------------------


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

#Let's look at the cities and counties in our subsetted dataset
unique(newdata_CA_monthly$city)
length(unique(newdata_CA_monthly$city))
unique(newdata_CA_monthly$county)
length(unique(newdata_CA_monthly$county))

#let's make a map
register_google(key = "AIzaSyAB00xfrsVh2QKnqKPxPQLpPRkusOCWf7A")
qmap(location = "State␣of␣California")
qmap('California', zoom=5, maptype='road')

#from ggmap github check out the USA and California
```

```r
us <- c(left = -125, bottom = 25.75, right = -67, top = 49)
map <- get_stamenmap(us, zoom = 5, maptype = "toner-lite")
ggmap(map)

california <- c(left=-125, bottom = 32, right = -113, top = 42)
map_california <- get_stamenmap(california, zoom = 5, maptype = "
   toner-lite")

ggmap(map_california)
qmplot(longitude, latitude, data = newdata_CA_monthly, maptype =
   "toner-lite", color = I("red"))

#For the sake of the example let's look at yolo county
data_yolocounty <- newdata_CA_monthly %>% filter(county=="Yolo")
   %>% drop_na(city)

#now split data into city and not city
city_yolo <- data_yolocounty %>% filter(city=="West␣Sacramento")
suburb_yolo <- data_yolocounty %>% filter(!city=="West␣Sacramento
   ")

summary(city_yolo)
summary(suburb_yolo)
#not enough data points to make strong claim but they seemed
   comparable on pets but the suburbs had more bedrooms

#Now let's try for LA
#now split data into city and not city
data_lacounty <- newdata_CA_monthly %>% filter(county=="Los␣
   Angeles") %>% drop_na(city)
city_la <- data_lacounty %>% filter(city=="Los␣Angeles")
suburb_la <- data_lacounty %>% filter(!city=="Los␣Angeles")
summary(city_la)
summary(suburb_la)

#compare pets in a barplot
summary(city_la$pets);length(city_la$pets)
summary(suburb_la$pets);length(suburb_la$pets)

#There were 717 entries in the city and 441 in the suburbs
#Let's look at the fraction of each that allowed pets for city
city_la_pets <- as.data.frame(summary(city_la$pets))
```

```
names(city_la_pets) = "pets"
city_la_pets <- city_la_pets %>% mutate(category = rownames(city_
   la_pets))
city_la_pets <- city_la_pets %>% mutate(ratio = pets/length(city_
   la$pets)) %>% mutate(type="city")
sum(city_la_pets$ratio)

#repeat the same for suburbs
suburb_la_pets <- as.data.frame(summary(suburb_la$pets))
names(suburb_la_pets) = "pets"
suburb_la_pets <- suburb_la_pets %>% mutate(category = rownames(
   suburb_la_pets))
suburb_la_pets <- suburb_la_pets %>% mutate(ratio = pets/length(
   suburb_la$pets)) %>% mutate(type="suburb")

#now combine the dataframes
comparing_pets_la <- rbind(city_la_pets,suburb_la_pets)

#load df for later
save(comparing_pets_la,file="comparing_pets_la.Rda")
load("comparing_pets_la.Rda")

#now time to make the barplot

plot7 <- ggplot(data=comparing_pets_la, aes(x=category, y=ratio,
   fill=type)) +
geom_bar(stat="identity", color="black", position=position_dodge
   ())+
theme_minimal() +
scale_fill_brewer(palette="OrRd")+
labs(title="Comparing␣Pet␣Friendliness␣of␣City␣vs␣Suburbs␣in␣LA␣
   County", x ="Pet␣Category␣(by␣craigslist␣category)", y = "
   Ratio␣of␣Proportion␣of␣Posts␣with␣that␣␣Label")+
theme(plot.title = element_text(hjust = 0.5))+
theme(axis.text.x = element_text(angle = 90, hjust = 1))
plot7

#compare number of bedrooms
summary(suburb_la$bedrooms)
summary(city_la$bedrooms)

#
```

```
   -------------------------------------------------------------------------

#Do apartments in similar geographical areas tend to be similar?
   Discuss how they
#are or how they are not for 3 different cities or places

unique(newdata_CA_monthly$county)
newdata_CA_monthly %>% filter(county=="Alameda") %>% count(city)

#Let’s compare Berkeley, Alameda, and Oakland
Berkeley <- newdata_CA_monthly %>% filter(county=="Alameda") %>%
   filter(city=="Berkeley")
Alameda <- newdata_CA_monthly %>% filter(county=="Alameda") %>%
   filter(city=="Alameda")
Oakland <- newdata_CA_monthly %>% filter(county=="Alameda") %>%
   filter(city=="Oakland")

#compare prices for all three locations
Berkeley_bed <- Berkeley %>% drop_na(bedrooms,price)
Alameda_bed <- Alameda %>% drop_na(bedrooms,price)
Oakland_bed <- Oakland %>% drop_na(bedrooms,price)
compare_eastbay <- rbind(Berkeley_bed,Alameda_bed,Oakland_bed)
compare_eastbay$bedrooms <- as.factor(compare_eastbay$bedrooms)

#remove post for sailboat for 16500
compare_eastbay <- compare_eastbay %>% filter(price<14000)

#now I’ll make another barplot to compare
plot8 <- ggplot(data=compare_eastbay, aes(x=bedrooms, y=price,
   fill=city)) +
geom_bar(stat="identity",position=position_dodge())+
theme_minimal() +
scale_fill_brewer(palette="OrRd")+
labs(title="Monthly␣Housing␣Prices␣By␣East␣Bay␣Region", x ="
   Bedrooms␣(by␣#)", y = "Price␣in␣USD")+
theme(plot.title = element_text(hjust = 0.5))+
theme(axis.text.x = element_text(angle = 90, hjust = 1))
plot8

#
   ----------------------------------------------------------------------
```

```
#Questions:
#1 Is it easier to find an apt with parking in Oakland or
   Berkeley?
#2 How does average rental price compare to the average salaries
   in that region?
#3 How does the price per sqft compare between San Francisco and
   Oakland?
#4 Which places are the most pet friendly in all of California?
##5 Are there regions where laundry is not commonly available
   with your housing?
#6 How many posts get deleted as opposed to just expiring?
##7 What proportion of posts get updated?
##8 Are apartments in SF that allow dogs generally more or less
   expensive?
##9 Is it easier to find a place with a cat or a dog in the city
   of LA?
##10 What size apartment can I expect in SF if my budget is
   2000-2500 per month for a 1 bedroom?
#11 At what point in the month do most people make posts?
## Does #8 change if we account for larger places?

#
   -------------------------------------------------------------------------

#Answers:

# What size apartment can I expect in SF if my budget is
   2000-2500 per month for a 1 bedroom?
SF_size <- newdata_CA_monthly %>% filter(city=="San␣Francisco")
   %>%  filter(price>2000 & price <2500) %>% drop_na(sqft)
range(SF_size$sqft)
#The range is from 295 to 1668 square feet.
fivenum(SF_size$sqft)
sd(SF_size$sqft)
#mean is 550 with sd of 254

#now make a plot for this data
qplot(SF_size$sqft, geom="histogram")


ggplot(data=SF_size, aes(SF_size$sqft)) +
geom_histogram(col="black", fill="steelblue")+
labs(title="Histogram␣for␣Apartment␣Size␣\n␣in␣San␣Francisco") +
labs(x="Size␣in␣Sqft", y="Count")

#
   ---------------------------------------------------------------------

#Is it easier to find a place with a cat or a dog in the city of
   LA?

dogorcat <- newdata_CA_monthly %>% filter(city=="Los␣Angeles")
dogorcat %>% count(pets)
newdata_CA_monthly %>% filter(county=="Los␣Angeles") %>% count(
   city,pets) %>% arrange(desc(n))
#should be somewhat easier with a cat since 357 allow both, 87
   allow cats, and 8 allow dogs in LA
#We checked other parts of LA county and Los Angeles seems to be
   the best bet.

#
   -------------------------------------------------------------------------

##6 How many posts get updated?

#Get number of NAs
sum(is.na(newdata_CA_monthly$date_updated))
sum(!is.na(newdata_CA_monthly$date_updated))
length(newdata_CA_monthly$date_updated)
#There were 3460 NA values (not updated) and 3474 that were not
   NA (that did get updated.)

#Was this similar for the non-subsetted original data?
sum(is.na(data$date_updated))
sum(!is.na(data$date_updated))
#In the original there were 13139 NA and 8809 did get updated.
   Makes sense since repetitive posts unlikely to get updated.

#
   ----------------------------------------------------------------------

#8 Are apartments in SF that allow dogs generally more or less
   expensive?
```

```
dogyes_SF <- newdata_CA_monthly %>% filter(pets=="dogs" | pets=="
   both")
dogno_SF <- newdata_CA_monthly %>% filter(pets=="none" | pets=="
   cats" | pets=="negotiable")
unique(dogyes_SF$pets)
unique(dogno_SF$pets)

fivenum(dogyes_SF$price)
fivenum(dogno_SF$price)

#Yes, there is a slight price increase for dogs yes - might
   correspond to pet fee or to bigger backyard

#Make plot for this data by # bedrooms
dogyes_SF <- dogyes_SF %>% mutate(dogs_allowed="yes")
dogno_SF <- dogno_SF %>% mutate(dogs_allowed="no")
dogs_SF <- rbind(dogyes_SF,dogno_SF)
plot10 <- ggplot(dogs_SF, aes(y=price,fill=dogs_allowed)) +
geom_boxplot() +
ggtitle("Rental␣Price␣Per␣Month␣in␣SF␣\n␣Depending␣on␣Dogs␣
   Allowed␣or␣Not") +
xlab("Bedrooms␣(by␣#␣bedrooms)") + ylab("Price␣in␣USD") +
theme(plot.title=element_text(hjust=0.5))+
coord_cartesian(ylim = c(0, 10000)) +
facet_wrap(~bedrooms)
plot10

#subset this and then plot
dogs_SF_subset <- dogs_SF %>% filter(bedrooms >0 & bedrooms <6)
plot10_2 <- ggplot(dogs_SF_subset, aes(y=price,fill=dogs_allowed)
   ) +
geom_boxplot() +
ggtitle("Rental␣Price␣Per␣Month␣in␣SF␣\n␣Depending␣on␣Dogs␣
   Allowed␣or␣Not") +
xlab("Bedrooms␣(by␣#␣bedrooms)") + ylab("Price␣in␣USD") +
theme(plot.title=element_text(hjust=0.5))+
coord_cartesian(ylim = c(0, 10000)) +
facet_wrap(~bedrooms,ncol=5)
plot10_2


#
```

```
   ------------------------------------------------------------------------

#Does this change if we account for sqft?
dogs_SF_sqft <- dogs_SF_subset %>% filter(sqft > 1) %>% mutate(
   price_per_sqft=price/sqft)

plot11 <- ggplot(dogs_SF_sqft, aes(y=price_per_sqft,fill=dogs_
   allowed)) +
geom_boxplot() +
ggtitle("Rental␣Price␣Per␣Month␣in␣SF␣\n␣Depending␣on␣Dogs␣
   Allowed␣or␣Not") +
xlab("Bedrooms␣(by␣#␣bedrooms)") + ylab("Price␣in␣USD␣per␣sqft")
   +
theme(plot.title=element_text(hjust=0.5))+
#coord_cartesian(ylim = c(0, 10000)) +
facet_wrap(~bedrooms,ncol=5)
plot11

fivenum(dogs_SF_sqft$price_per_sqft[dogs_SF_sqft$dogs_allowed=="
   yes"])
fivenum(dogs_SF_sqft$price_per_sqft[dogs_SF_sqft$dogs_allowed=="
   no"])

#This difference disappeared when we accounted for sqft!
   interesting result

#
   ------------------------------------------------------------------------

##5 Are there regions where laundry is not commonly available
   with your housing?
## Try to find patterns/trends about laundry

#Let's start with the bay area
data_SF_laundry <- newdata_CA_monthly %>% filter(city=="San␣
   Francisco")
findlaundry <- newdata_CA_monthly %>% group_by(city,laundry) %>%
   summarise(count = n()) %>% arrange(desc(count))

#let's look at which places have most 'none' for laundry
laundry_none <- findlaundry %>% filter(laundry=="none") %>%
   arrange(desc(count))
```

```
xtable(head(laundry_none,8))
#
   ------------------------------------------------------------------------

##Examine missing values for trends
#First do sapply to count NAs per column
NA_per_column <- as.data.frame(sapply(data, function(x) sum(
   length(which(is.na(x))))))

#Then create column for factor names,rename columns, and arrange
   in increasing order
NA_per_column_factors <- NA_per_column %>%
mutate(factors = rownames(NA_per_column)) %>%
magrittr::set_colnames(c("number_NA", "factor")) %>%
arrange(number_NA) %>%
rename(Number_of_NAs = number_NA) %>%
rename(Name_of_Column=factor)


#look at preview of this
head(NA_per_column_factors,20)
tail(NA_per_column_factors,20)

#make a latex table now
xtable(head(NA_per_column_factors,20))

#repeat process for our subsetted monthly CA dataframe


NA_Subset <- as.data.frame(sapply(newdata_CA_monthly, function(x)
    sum(length(which(is.na(x))))))

#Then create column for factor names,rename columns, and arrange
   in increasing order
NA_Subset_factors <- NA_Subset %>%
mutate(factors = rownames(NA_Subset)) %>%
magrittr::set_colnames(c("number_NA", "factor")) %>%
arrange(number_NA) %>%
rename(Number_of_NAs_Subset = number_NA) %>%
rename(Name_of_Column=factor)
```

```
#look at preview of this and make latex table
head(NA_Subset_factors,20)
tail(NA_Subset_factors,20)
xtable(head(NA_Subset_factors,20))

#make a table comparing the number of NAs in the original data vs
    my subset dataframe
comparing_dataframes_NA <- cbind(NA_Subset,NA_Subset_factors)
```