STAT 33A Homework 4

CJ HINES (3034590053)

Oct 22, 2020

This homework is due **Oct 22, 2020** by 11:59pm PT.

Homeworks are graded for correctness.

As you work, write your answers in this notebook. Answer questions with complete sentences, and put code in code chunks. You can make as many new code chunks as you like.

Please do not delete the exercises already in this notebook, because it may interfere with our grading tools.

You need to submit your work in two places:

- Submit this Rmd file with your edits on bCourses.
- Knit and submit the generated PDF file on Gradescope.

If you have any last-minute trouble knitting, **DON'T PANIC**. Submit your Rmd file on time and follow up in office hours or on Piazza to sort out the PDF.

The Bay Area Apartments Data Set

The Bay Area Apartments Data Set is a collection of advertisements for apartments for rent in Los Angeles and the San Francisco Bay Area. The data set was collected from the website Craigslist in October 2020.

The data set is available on the bCourse as `2020.10_cl_apartments.rds`.

Each row is one advertisement. The columns are:

- `title`: title of advertisement
- `text`: full text of advertisement
- `latitude`: latitude of apartment
- `longitude`: longitude of apartment
- `city_text`: city listed in advertisement
- `date_posted`: date advertisement was posted
- `date_updated`: date advertisement was updated, if any
- `price`: price in US dollars
- `sqft`: size in square feet
- `bedrooms`: number of bedrooms
- `bathrooms`: number of bathrooms (0.5 means sink/toilet without bath)
- `pets`: what pets (cats or dogs) are allowed
- `laundry`: what laundry services are provided
- `parking`: what parking is provided
- `fname`: unique identifier for the advertisement
- `craigslist`: craigslist region where advertisement was posted

- `place`: place name (like city, but also includes small towns) based on latitude/longitude
- `city`: city based on latitude/longitude
- `state`: state based on latitude/longitude
- `county`: county based on latitude/longitude

Many of the columns were programmatically extracted from the `title` and `text`, so there may be missing or incorrect values.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
apts = readRDS("rapartments.rds")
```

Exercise 1

Use base R for this exercise.

Read the apartments data set into R, then answer the following:

1. How many advertisements are there?
2. Based on the `text` column, how many of the ads are duplicates?
3. Based on the `title` column, how many of the ads are duplicates?
4. Do all of the ads from Part 3 also have duplicate `text`?
5. Remove the ads where the `text` and `title` are both duplicates. Use this de-duplicated version of the data set for the remaining exercises. How many ads remain?

YOUR ANSWER GOES HERE:

Part 1

```
nrow(apts)
```

```
## [1] 19842
```

Part 2

```
d = duplicated(apts$text)
length(d[d == TRUE])
```

```
## [1] 5476
```

Part 3

```
t = duplicated(apts$title)
length(t[t == TRUE])
```

```
## [1] 4594
```

Part 4

```
both = d & t
b = both[both == TRUE]
length(b)
```

```
## [1] 4094
```

Part 5

```
realapts = subset(apts, both == FALSE)
nrow(realapts)
```

```
## [1] 15748
```

Exercise 2

Use dplyr for this exercise.

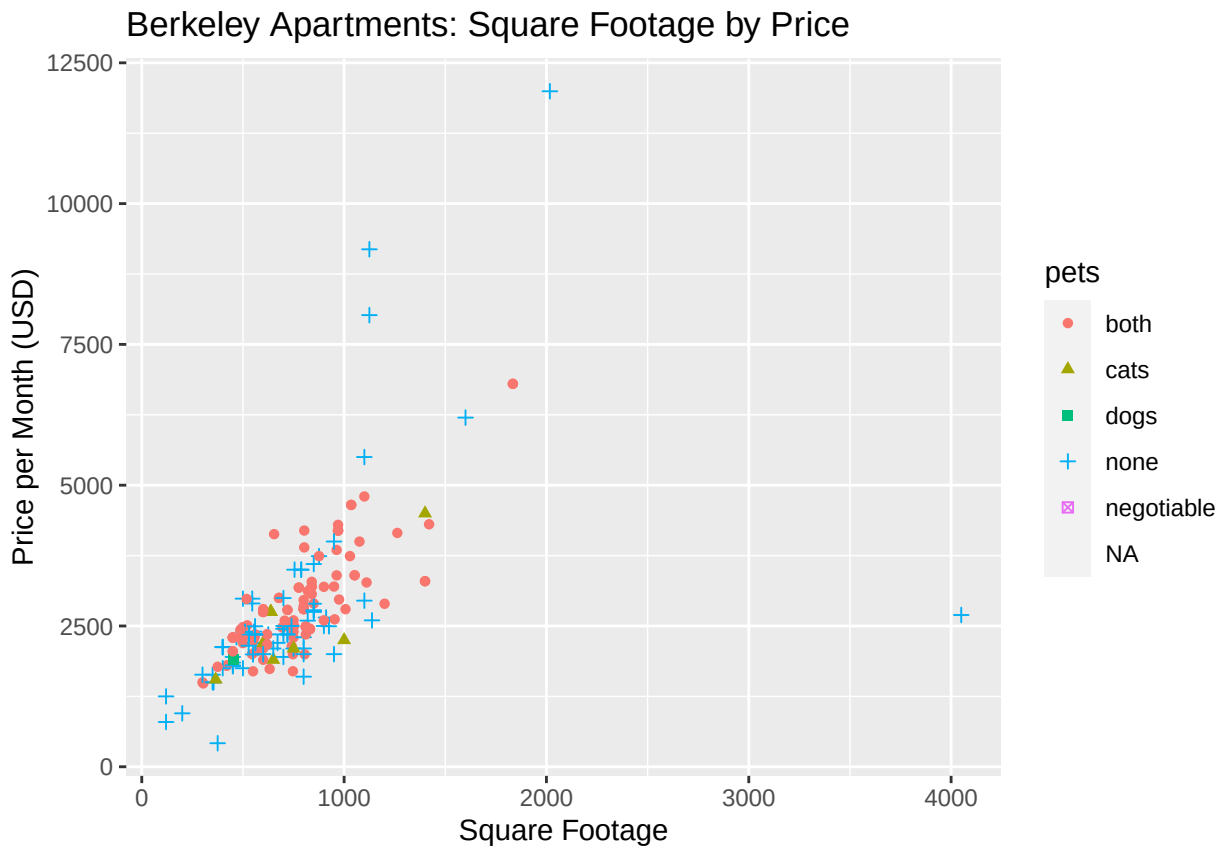
1. Make a scatter plot of square footage versus price for apartments in Berkeley. Set the color and shape of the points based on pets category. Make sure to add a proper title and labels to your plot.
2. Using the plot as evidence, discuss in 3-5 sentences how price, square footage, and pets interact.

YOUR ANSWER GOES HERE:

Part 1

```
library(ggplot2)
library(dplyr)
Berkeley = realapts[which(realapts$city == "Berkeley"), ]
ggplot(Berkeley, aes(sqft, price)) + geom_point(aes(color = pets, shape = pets)) + ggtitle("Berkeley Ap
```

```
## Warning: Removed 182 rows containing missing values (geom_point).
```



Part 2

Most apartments in Berkeley allow either both dogs and cats or no pets at all. The majority of listings are below 1500 square feet and below $5000. Nearly all outlier apartments that are larger and mor expensive also don't allow pets. Surprisingly, the opposite end (cheapest and smallest) also don't allow pets.

Exercise 3

Use dplyr for this exercise.

1. Make a ridges plot of square footage with separate lines for San Francisco, San Jose, Oakland, and Los Angeles. Omit all other cities. Make sure to add a proper title and labels to your plot.

   Hint: You can use the `droplevels()` *function to drop factor levels that aren't present.*

2. Your plot is likely hard to read because some of the cities have ads with extreme square footages. Choose and state a threshold for "extreme" square footage. Then recreate the plot with the x-axis restricted (using `xlim()`) to ads that do not have extreme square footage.

3. How do the square footage distributions of the cities compare?

YOUR ANSWER GOES HERE:

Part 1

```
library(dplyr)
library(ggridges)
cities = c("San Francisco", "San Jose", "Oakland", "Los Angeles")
fourapts = realapts[which(realapts$city == cities), ]
ggplot(fourapts, aes(sqft, city)) + geom_density_ridges(aes(color = city)) + ggtitle("Square Footage by
```

```
## Picking joint bandwidth of 83.2
```

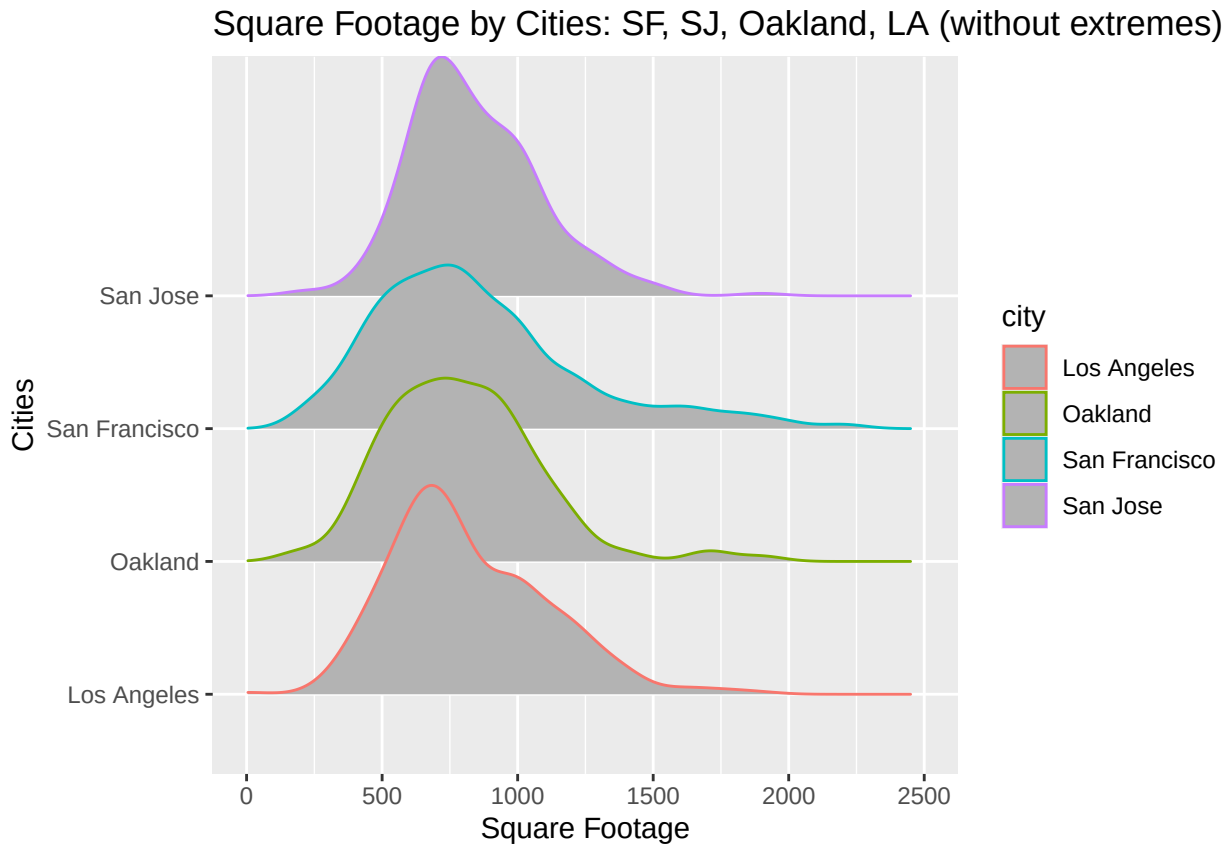
```
## Warning: Removed 589 rows containing non-finite values (stat_density_ridges).
```

Part 2

```
extreme = c(0, 2500)
ggplot(fourapts, aes(sqft, city)) + geom_density_ridges(aes(color = city)) + xlim(extreme) + ggtitle("S
```

```
## Picking joint bandwidth of 83.2
```

```
## Warning: Removed 596 rows containing non-finite values (stat_density_ridges).
```



Part 3

All four cities have a similar distribution. San Francisco has the widest distribution with more smaller and larger apartments. San Jose and Los Angeles have two similar small peaks at around 700 and 1000 square feet. Oakland has a greater distribution between 500 and 1000 square feet.

Exercise 4

Use base R or dplyr for this exercise.

Isolate three of the extreme ads from Exercise 3. Does the square footage seem accurate, or is it a mistake? Use the original title and text of the ad as evidence.

Hint 1: You can print the text of an ad in human-readable form with the `message()` *function.*

Hint 2: You can use the `stringr` *package's* `str_wrap()` *function to wrap long strings (e.g., the ad text) for printing in the notebook.*

Hint 3: The PDFLaTeX program that RMarkdown uses to knit PDFs only supports ASCII characters. Some of the advertisments contain non-ASCII characters. If you get a knit error like `! Package inputenc Error: Unicode character`*, you probably printed an ad with non-ASCII characters.*

To fix it, you can either comment out the line that prints the ad, or switch from PDFLaTeX to XeLaTeX or LuaLaTeX. See https://bookdown.org/yihui/rmarkdown-cookbook/latex-unicode.html for details about how to switch.

YOUR ANSWER GOES HERE:

```
library(stringr)
threeextremes = fourapts[which(fourapts$sqft > 6000), ]
#str_wrap(threeextremes$text[3])

message(threeextremes$title[1])
```

```
## $2,890 / 1br - 6630ft2 - Furnished Hip 1BR Playa Del Rey w/ Gym, Pool, nr. Venice Beach (246) (Playa
```

```
#message(threeextremes$text[1])

message(threeextremes$title[2])
```

```
## $3,190 / 1br - 6600ft2 - Furnished Quiet Mission Dolores 1BR w/ Gym, BBQ, near Bart/Muni (SFO 1 (Miss
```

```
#message(threeextremes$text[2])

message(threeextremes$title[3])
```

```
## $4,190 / 1br - 6470ft2 - Furnished Lux Rincon Hill 1BR w/ Gym, Sauna, nr. SV. (238) (Rincon Hill)
```

```
#message(threeextremes$text[3])
```

The square footage is not accurate for these three extremes (the three largest apartments). From the titles, we know that all three of these are only 1 bedroom apartments which already makes it unlikely that they are 6000+ square feet. To confirm this inaccuracy though, I used the link provided in each text to reach the website where I found that all of the listings were actually only around 600~ square feet.

Exercise 5

Use dplyr for this exercise.

1. Investigate how much rent can be saved by sharing an apartment. Create a column for price per bedroom. Then compute the median price per bedroom, grouped by number of bedrooms.

   Hint: Studio apartments are listed as 0-bedroom apartments. For this exercise, you can exclude them from the data set.

2. Does sharing an apartment save rent? Explain in 1-3 sentences, using the statistic you computed as evidence.

YOUR ANSWER GOES HERE:

Part 1

```
library(dplyr)
nostudios = realapts[which(realapts$bedrooms > 0), ]
nostudios$priceperbedroom = nostudios$price / nostudios$bedrooms
med = median(nostudios$priceperbedroom, na.rm = TRUE)
grouped = group_by(nostudios, bedrooms)
summarize(grouped, median(priceperbedroom, na.rm = TRUE))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 7 x 2
##   bedrooms `median(priceperbedroom, na.rm = TRUE)`
##      <dbl>                                   <dbl>
## 1        1                                    2184
## 2        2                                    1325
## 3        3                                    1183.
## 4        4                                    1168.
## 5        5                                    1360
## 6        6                                    1417.
## 7        7                                    1428.
```

Part 2

Sharing an apartment does save rent. The price per bedroom of a 1 bedroom apartment is significantly higher than all other apartments with greater number of bedrooms. Sharing an apartment is most efficient for apartments with four bedrooms because the median cost per bedroom increases for apartments with more than four bedrooms and less than four bedrooms.