STAT 33B Lab 2

Gunnar Mayer (3034535154)

Edit this file, knit to PDF, and:

- Submit the Rmd file on bCourses.
- Submit the PDF file on Gradescope.

If you think you'll need help with submission, please ask during the lab.

Answer all 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.

As you work, you may find it helpful to be able to run your code. You can run a single line of code by pressing Ctrl + Enter. You can run an entire code chunk by clicking on the green arrow in the upper right corner of the code chunk.

Knit the document from time to time to make sure that your code runs without errors from top to bottom in a fresh R environment.

The code below controls the number of significant digits shown for the return values in your knitted document.

```
options(digits = 3)
```

Bay Area Apartments Data Set

The Bay Area Apartments Data Set is a collection of 5852 advertisements for apartments for rent in the San Francisco Bay Area. The data set was collected from Craigslist on Feb 13, 2020.

In this lab, you'll use subsetting, statistical functions, and `ggplot2` to analyze the data set.

Exercise 1

The data set is available on the bCourse as an RDS file.

Read the data set into R, then use R functions to display the following:

- The dimensions of the data set.
- The names of the columns in the data set.
- A **str**uctural summary of the data set.

```
bay_area_data = readRDS("cl_apartments.rds")
dim(bay_area_data)
```

```
## [1] 5852   21
```

```r
colnames(bay_area_data)
```

```
##  [1] "title"        "text"         "latitude"     "longitude"    "city_text"
##  [6] "date_posted"  "date_updated" "price"        "deleted"      "sqft"
## [11] "bedrooms"     "bathrooms"    "pets"         "laundry"      "parking"
## [16] "fname"        "craigslist"   "place"        "city"         "state"
## [21] "county"
```

```r
str(bay_area_data)
```

```
## 'data.frame':    5852 obs. of  21 variables:
##  $ title       : chr  "$1695 / 1br - 700ft2 - 1 Bed 1 Bath Apartment Downtown Livermore $1695 (dubli
##  $ text        : chr  "QR Code Link to This Post\n            \n        \n1 bed 1 bath apartment for
##  $ latitude    : num  NA 37.9 37.6 37.9 37.9 ...
##  $ longitude   : num  NA -122 -122 -122 -122 ...
##  $ city_text   : Factor w/ 277 levels "  Limited Time Only ) (vallejo / benicia",..: 64 51 73 229 37
##  $ date_posted : POSIXlt, format: "2020-01-31 13:56:37" "2020-02-13 10:35:39" ...
##  $ date_updated: POSIXlt, format: "2020-02-13 09:58:45" NA ...
##  $ price       : num  1695 2120 2463 1875 2900 ...
##  $ deleted     : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ sqft        : num  700 590 777 650 600 ...
##  $ bedrooms    : num  1 1 1 1 2 2 2 2 4 1 ...
##  $ bathrooms   : num  1 1 1 1 1 2 2 2 2.5 1 ...
##  $ pets        : Factor w/ 5 levels "both","cats",..: NA 1 1 1 5 1 1 1 1 1 ...
##  $ laundry     : Factor w/ 4 levels "hookup","in-unit",..: 4 2 2 4 3 4 4 4 2 4 ...
##  $ parking     : Factor w/ 6 levels "covered","garage",..: 1 1 1 1 3 2 4 2 6 2 ...
##  $ fname       : chr  "data//sfbay/_eby_apa_d_1-bed-1-bath-apartment-downtown_7066767007.html" "data/
##  $ craigslist  : Factor w/ 2 levels "sfbay","sfbay_eby": 1 1 1 1 1 1 1 1 1 1 ...
##  $ place       : Factor w/ 129 levels "Alameda","Alamo",..: NA 88 124 128 11 1 1 1 1 1 ...
##  $ city        : Factor w/ 95 levels "Alameda","Albany",..: NA 62 91 94 7 1 1 1 1 1 ...
##  $ state       : chr  NA "CA" "CA" "CA" ...
##  $ county      : Factor w/ 18 levels "Alameda","Clark",..: NA 3 1 3 1 1 1 1 1 1 ...
```

Exercise 2

The city each apartment is in is listed in the `city` column. Note that some entries in the `city` column are missing.

Get the subset of apartments in Berkeley. Then answer the following:

1. How many advertisements are there for Berkeley apartments?
2. What is the mean price for Berkeley apartments?
3. What is the median price for Berkeley apartments?

```r
berkeley_apts = subset(bay_area_data, bay_area_data$city == 'Berkeley')
berk_apts_mean = mean(berkeley_apts$price)
berk_apts_mean
```

```
## [1] 3243
```

```
berk_apts_median = median(berkeley_apts$price)
berk_apts_median
```

```
## [1] 2960
```

WRITE YOUR ANSWERS BELOW:

1. There are 355 advertisements for Berkeley apartments.
2. The mean price for Berkeley apartments is $3243
3. The median price for Berkeley apartments is $2960

Exercise 3

How do the mean and median prices for Berkeley apartments compare to San Francisco? Discuss in 1-3 sentences.

```
sf_apts = subset(bay_area_data, bay_area_data$city == 'San Francisco')
sf_apts_mean = mean(sf_apts$price)
sf_apts_mean
```

```
## [1] 3547
```

```
sf_apts_median = median(sf_apts$price)
sf_apts_median
```

```
## [1] 3490
```

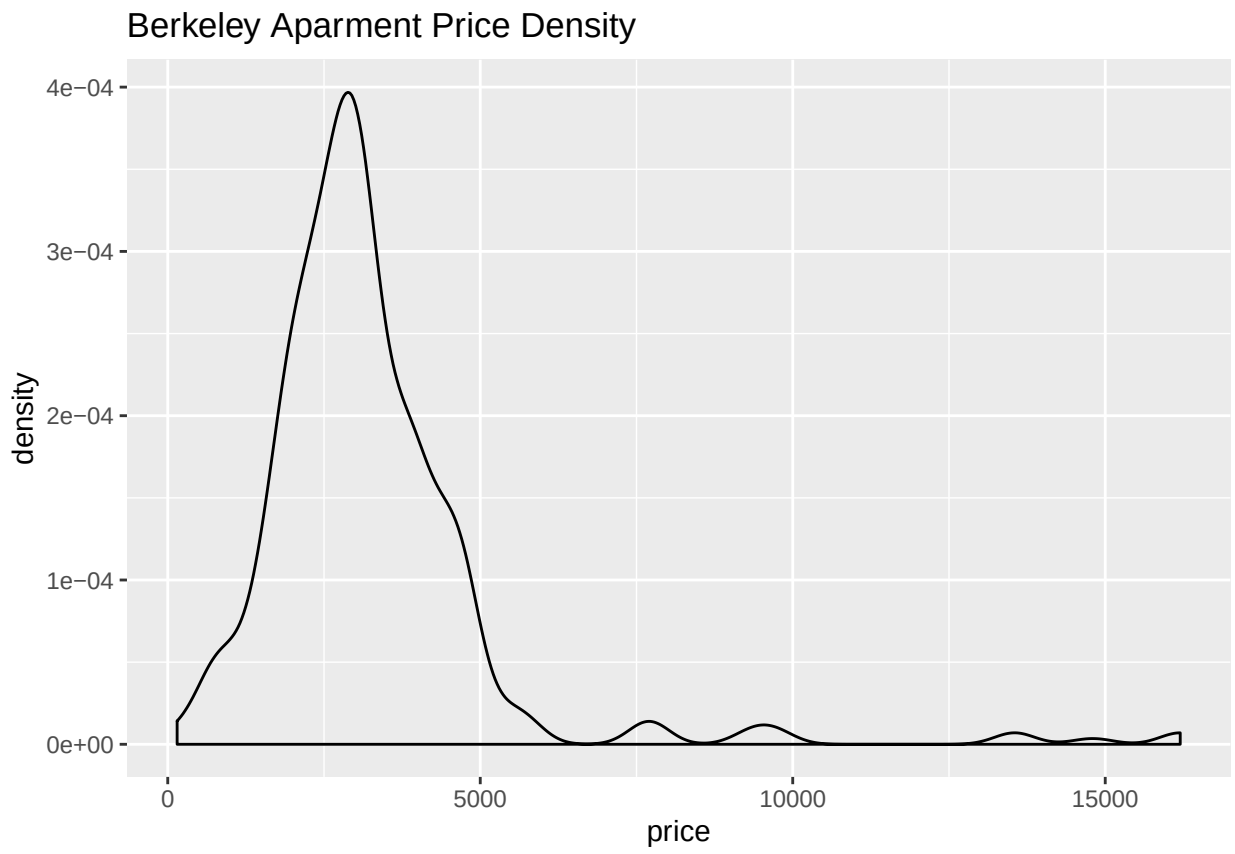
WRITE YOUR ANSWER BELOW: Both the median and the mean prices for apartments in San Francisco are higher than the mean and median in Berkeley. However, the mean in Berkeley is closer to the mean in San Francisco when compared to the difference between their respective medians. From this comparison one could guess that apartment prices in Berkeley can be found in a wider price range than in San Francisco.

Exercise 4

A density plot represents a distribution of values as a smooth curve. The height of the curve at a given point corresponds to how likely values are to fall near that point.

Use `ggplot2`'s `geom_density()` geometry to make a density plot of the price of apartments in Berkeley. Add an appropriate title to the plot with the `labs()` function. *Hint: see the online documentation for examples.*

```
library(ggplot2)
ggplot(berkeley_apts, aes(price))+
  geom_density() +
  labs(title = "Berkeley Aparment Price Density")
```



Exercise 5

Some of the apartments advertised in Berkeley list prices above 10,000 USD. Investigate these apartments:

1. How many of these apartments are there?
2. Do the listed prices appear to be correct? Use information from other columns as evidence.

   Hint: you can pretty-print the text in the `text` *and* `title` *columns with the* `cat()` *function. If you do, please DO NOT include the pretty-printed text in your submitted PDF file. Instead, tell us what you discovered.*

```
expensive_berk_apts = subset(berkeley_apts, berkeley_apts$price > 10000)
length(expensive_berk_apts$city)
```

```
## [1] 5
```

```
# Separate chunk to show code for pretty-printing, but hide output.
cat(expensive_berk_apts$title)
cat(expensive_berk_apts$text)
```

WRITE YOUR ANSWERS BELOW:

1. There are 5 apartments in Berkeley being rented for over $10,000.
2. The listed prices appear to be correct. The prices of the listings are also placed in the title of the listings. All of them are above $10,000. This leads me to believe that the listed prices are also correct.