

# Homework Assignment 2


STA 141A

Due Monday, May 7 by 5:00 pm

## Description

You are given a random sample of 20,000 housing sales from the San Francisco Bay Area. The data contains a large amount of relevant information for each house sale. However, there may be errors and missing values in the data. It is your job to extract meaningful information from the data. You may have to remove or overwrite values in the data, so make sure you can justify each decision.

## Questions

Use R to find answers to all of the following questions (that is, don't do any by hand or by point-and-click). Save your code in an R script. Try to complete at least one every day until the assignment is due. Some problems are more difficult than others, while some are lengthier than others. Some problems may be easier if you have completed earlier ones. It is recommended that you read all the questions before beginning the problem set so you are able to adequately pace yourself. Problems that may be difficult or time-consuming are noted with .

1. Load in the housing dataset. Convert the columns to appropriate data types. If you notice any data irregularities, you can potentially write them here. You don't need to write an answer in your report for this question, but please mark the code for this question in the appendix.
2. What timespan does the housing sales cover? What is the timespan of the construction dates of homes?
3. Examine the monthly housing sales for this dataset. You will need to look at combinations of both year and month using the `date` variable. Make two plots:

A plot that shows the number of sales over time.


A plot that shows the average house price over time.

4. Make a line plot that shows how the average `price` depends on `county`, `bedrooms`, and sale year. For the `bedrooms` variable, define the levels to be 1,2,3,4+ bedrooms. For the sale year variable, use the levels 2003,2004,2005. Use all levels of the `county` variable. Each line should have three points corresponding to `year`. You will need separate lines for each combination of `county` and `br`.
5. Do all housing sales within a given city only occur in one county? Why or why not? Justify your answer. If you have any cities that have sales in more than one county, list them and report how many cities you find.
6. Fit a linear regression model that uses `bsqft` to predict `price`. Make sure to use appropriate diagnostics (e.g. Q-Q plots, Box-Cox transformations). Removing extreme outliers is okay for this problem. Do not worry if you're not too familiar with Linear Regression, any required methods will be covered during class.
7. Now fit a linear regression model that uses both `bsqft` and `lsqft` to predict `price`. Do not include any transformations or diagnostics. Using R, conduct a hypothesis test for  $H_0 : \beta_{bsqft} \geq \beta_{lsqft}$  vs.

$H_1 : \beta_{bsqft} < \beta_{lsqft}$ . For your convenience, you may assume the following as true:

$$Var(\hat{\beta}_{bsqft} - \hat{\beta}_{lsqft}) = Var(\hat{\beta}_{bsqft}) + Var(\hat{\beta}_{lsqft}) \quad (1)$$

Do not include any symbolic solutions; instead, show the code you used. Report your conclusion and test statistic.

8.  Fit an individual regression line using `bsqft` and `price` for each separate `county`

Do not write repeated calls of the `lm()` function. Instead, use an appropriate apply function. Draw each individual regression line in a single plot. Make sure to distinguish the lines. Can we conclude that the regression lines depend on `county`? *Hint: Are the lines parallel or not?*

9. For this question, you will be developing a treemap plot. Read the following steps carefully:

It is of interest to visualize the average prices for the three cities with the most sales in each county. Use appropriate subsetting methods to find the corresponding "top three" cities for each county. Note that some counties may have only 1 or 2 cities in which housing sales occur. This will not affect the end result.

For the treemap, make sure that the order of the indexing has the city variable "nested" in the county variable. This would require specifying `index=c(county,city)` when making the treemap in R.

10.  Your last question is to learn how to develop a heatmap. Read the following steps carefully:

The objective of this question is to examine how price and frequency of San Francisco housing sales depend on location. You are tasked with making two heatmaps: 1. A heatmap that is colorized based on the number of houses in a *cell* 2. A heatmap that is colorized by the average price of the houses within a *cell*

Each cell is a .01 (*longitude*)  $\times$  .01 (*latitude*) square. To implement this in R, you will want to perform the following transformation to the location variables:

```
SFdata$long2<-round(SFdata$long,2)
SFdata$lat2<-round(SFdata$lat,2)
```

Create appropriate factor variables for latitude and longitude that include levels that are not present in the dataset. You can use the `levels` argument to implement this.

For each plot, consider using the `image()` function to make a heatmap. You can use `heatmap()` or `heatmap.2()` or anything you want, but `image()` is better when overlaying lines. You will need to include the following arguments:

- (a) **x** A length  $m$  vector with the levels of your longitude factor variable
- (b) **y** A length  $n$  vector with the levels of your latitude factor variable
- (c) **z** An  $n \times m$  matrix object that contains the corresponding values for each cell. If there are no houses in a specific cell, then it should be reported as *NA*.

Finally, use the `maps` package to sketch in the borders of San Francisco around the tiles.

*Hint:* Try plotting the houses exact locations against the heatmap. You can use the `points()` function after calling the heatmap to do this.

Assemble your answers into a report. Please do not include any raw R output. Instead, present your results as neatly formatted<sup>1</sup> tables or graphics, and write something about each one. You must **cite your sources**. Your report should be **no more than 10 pages (both sides!!!)** including graphics, but excluding code and citations.

---

<sup>1</sup>See the graphics checklist on Canvas.

## What To Submit

Email a digital copy to [spring18stat141a@gmail.com](mailto:spring18stat141a@gmail.com). The digital copy must contain your report (as a PDF) and your code (as one or more R scripts).

Additionally, submit a printed copy to the box in the statistics department office<sup>4</sup>. The printed copy must contain your report and your code (in an appendix). Please print double-sided to save trees. It is your responsibility to make sure the graphics are legible in the printed copy!

## Data Documentation

The housing dataset contains the following features:

<code>county</code>	The county in which the house is sold.
<code>city</code>	The city in which the house is sold.
<code>zip</code>	The zip code in which the house is sold in.
<code>street</code>	The address of the house.
<code>price</code>	The price (in nominal U.S. dollars) of the house.
<code>br</code>	The number of bedrooms of the house.
<code>lsqft</code>	The lot size of the house, measured in square feet.
<code>bsqft</code>	The building size of the house, measured in square feet.
<code>year</code>	The construction year of the house
<code>date</code>	The date when the house was sold.
<code>long</code>	The longitude (horizontal) of the house location
<code>lat</code>	The latitude (vertical) of the house location

## Relevant Functions

All of the functions from Assignment 1, as well as:

`sapply()`, `lapply()`, `split()`, `tapply()`, `aggregate()`, `unlist()`, `mapply()`, `do.call()`, `grep()`, `gsub()`, `seq()`, `rep()`, `droplevels()`, `paste()`, `paste0()`, `cut()`, `any()`, `all()`, `round()`, `range()`, `match()`, `lm()`, `qqnorm()`, `qqline()`, `coef()`, `residuals()`, `fitted.values()`, `predict()`, `read.csv()`

## Relevant Packages

`treemap`, `map`, `lubridate`, `ggplot2`, `lattice`, `MASS` (for Box-Cox), any other Tidyverse package.

## Domain Knowledge

For your convenience, here are some Wikipedia articles that explain some of the geographic features of the data:

- [https://en.wikipedia.org/wiki/County\\_\(United\\_States\)](https://en.wikipedia.org/wiki/County_(United_States))
- [https://en.wikipedia.org/wiki/List\\_of\\_counties\\_in\\_California](https://en.wikipedia.org/wiki/List_of_counties_in_California)
- [https://en.wikipedia.org/wiki/ZIP\\_Code](https://en.wikipedia.org/wiki/ZIP_Code)

---

<sup>4</sup>4th floor of Mathematical Sciences Building