Cleaning and Exploring Data

Jun Tan 26370084

### Introduction

This assignment has three goals: (a) to familiarize you with some of the basic functions in R and how to use them, (b) to learn how to subset data frames to eliminate unusual records and values, (c) to review/introduce some of the key concepts in exploratory data analysis.

You will need to install the ggplot2 and RColorBrewer packages. Do this now before starting your HW. Use the packages tab in the pane where the plots are displayed to install the packages.

You will work with data published by the San Francisco Chronicle about housing sales. We begin by reading the data from the Web into R.

load(url("http://www.stat.berkeley.edu/users/nolan/data/Projects/SFHousing.rda"))

This command loads the *housing* data frame into your work space. The following functions may be useful in answering the questions: **class(), cor(), length(), dim(), ls(), names(), mean(), min(), max(), quantile(), summary().**

1. What are the names of the variables in the data frame? What type of data are they, i.e., are they numeric, logical, character, factors?

names(housing)

## [1] "county" "city" "zip" "street" "price" "br" "lsqft"   
## [8] "bsqft" "year" "date" "long" "lat" "quality" "match"   
## [15] "wk"

sapply(housing, class)

## $county  
## [1] "factor"  
##   
## $city  
## [1] "factor"  
##   
## $zip  
## [1] "factor"  
##   
## $street  
## [1] "character"  
##   
## $price  
## [1] "numeric"  
##   
## $br  
## [1] "integer"  
##   
## $lsqft  
## [1] "numeric"  
##   
## $bsqft  
## [1] "integer"  
##   
## $year  
## [1] "integer"  
##   
## $date  
## [1] "POSIXt" "POSIXct"  
##   
## $long  
## [1] "numeric"  
##   
## $lat  
## [1] "numeric"  
##   
## $quality  
## [1] "factor"  
##   
## $match  
## [1] "factor"  
##   
## $wk  
## [1] "Date"

1. Can you confirm that there are measurements on 281,506 houses  
   available in *housing*?

dim(housing)

## [1] 281506 15

nrow(housing)

## [1] 281506

The data do not indicate the units of measurement and some of the variable names are a bit terse. The variables and their units are:

* county - County name
* city - City name
* zip - Zip code
* street - Street address
* price - Sale price in dollars
* br - Number of bedrooms
* lsqft - Size of lot in square feet
* bsqft - Size of building in square feet
* year - Year house was built
* long - Longitude of house location
* lat - Latitude of house location

We are interested in the relationship between the price per square foot and the size of the house. Before we study this relaionship, we want to examine these and other variables to check if we need to clean the data.

### Property Size

1. Examine the property (lot) size with summary statistics. What do you see?

summary(housing)

## county city zip   
## Santa Clara County :70424 Oakland : 14730 94565 : 4595   
## Alameda County :60410 Santa Rosa : 9917 94509 : 4302   
## Contra Costa County:59381 Fremont : 9414 95123 : 4023   
## Solano County :23404 San Francisco: 8137 95687 : 3652   
## San Mateo County :22558 Evergreen : 7947 94533 : 3472   
## Sonoma County :21676 Antioch : 7726 (Other):261457   
## (Other) :23653 (Other) :223635 NA's : 5   
## street price br lsqft   
## Length:281506 Min. : 22000 Min. :1.000 Min. : 19   
## Class :character 1st Qu.: 400000 1st Qu.:2.000 1st Qu.: 4000   
## Mode :character Median : 530000 Median :3.000 Median : 5760   
## Mean : 602000 Mean :3.024 Mean : 65939   
## 3rd Qu.: 700000 3rd Qu.:4.000 3rd Qu.: 7701   
## Max. :20000000 Max. :8.000 Max. :418611600   
## NA's :21687   
## bsqft year date   
## Min. : 122 Min. : 0 Min. :2003-04-27 00:00:00   
## 1st Qu.: 1121 1st Qu.:1954 1st Qu.:2004-02-08 00:00:00   
## Median : 1430 Median :1971 Median :2004-10-24 00:00:00   
## Mean : 1624 Mean :1966 Mean :2004-11-01 16:06:12   
## 3rd Qu.: 1882 3rd Qu.:1985 3rd Qu.:2005-07-24 00:00:00   
## Max. :1868120 Max. :3894 Max. :2006-06-04 00:00:00   
## NA's :426 NA's :9202   
## long lat   
## Min. :-123.6 Min. :36.98   
## 1st Qu.:-122.3 1st Qu.:37.50   
## Median :-122.1 Median :37.77   
## Mean :-122.1 Mean :37.78   
## 3rd Qu.:-121.9 3rd Qu.:38.00   
## Max. :-121.5 Max. :38.85   
## NA's :23316 NA's :23316   
## quality   
## QUALITY\_ADDRESS\_RANGE\_INTERPOLATION :170719   
## gpsvisualizer : 31084   
## QUALITY\_CITY\_CENTROID : 20473   
## QUALITY\_EXACT\_PARCEL\_CENTROID : 17208   
## QUALITY\_ZIP\_CODE\_TABULATION\_AREA\_CENTROID: 14980   
## (Other) : 3726   
## NA's : 23316   
## match wk   
## Exact :197044 Min. :2003-04-21   
## Relaxed : 30570 1st Qu.:2004-02-01   
## Relaxed; Soundex: 23338 Median :2004-10-18   
## Soundex : 2573 Mean :2004-10-26   
## 1 : 2244 3rd Qu.:2005-07-18   
## (Other) : 2421 Max. :2006-05-29   
## NA's : 23316

The summary shows that some zip codes are missing but no counties or cities are missing. The amounts of cities and zip codes are classified as other are extremely high. Moreover, the max price is a lot higher than the 3rd quantile.

1. An acre is 43560 sqft. It seems that houses with plots more than 3 to 5 acres are different types of properties than those without. Set *lsqft* to NA if the home has a lot of 3+ acres.

housing$lsqft[(housing$lsqft/43560) > 3] = NA

1. Some of the tiniest houses measure 100-200 square feet (see <http://www.housebeautiful.com/design-inspiration/house-tours/a5836/hikari-box-tiny-house-shelter-wise/>) Since these are tiny, modern "eco homes", we can exclude houses that don't have a lot square footage of a couple times that. Again, set *lsqft* to NA for those homes with less than a 400 square foot lot.

housing$lsqft[housing$bsqft < 400] = NA

## House Size

1. As with lot size, there are some unusually large houses. How big are the 10 largest houses? (Use *order()* to help find this information)

orderSize = sort(housing$bsqft)  
tail(orderSize, 10)

## [1] 31264 52542 84200 93320 96538 1783086 1847335 1856544  
## [9] 1856544 1868120

1. Drop any house over 20000 square feet from the data frame.

drop = housing$bsqft > 20000  
housing = housing[!drop, ]

## Relationship between lot size and house size

1. How many houses have a larger building square footage than lot square footage? How can this happen?

greater = (housing$bsqft > housing$lsqft)  
amount = length(greater[greater==TRUE])  
amount

## [1] 42465

This number shows that a lot of buildings are tall such that they can provide more living space.

1. Remove all houses from the data frame that have a building size that is more than 4 times the lot size.

drop = housing$bsqft > (housing$lsqft \* 4)  
housing = housing[!drop, ]

## Price per square foot

1. We said that we are interested in the price per square foot of the houses. This variable is not in our data frame. Create a new variable, *ppsf*, in *housing*.

ppsf = housing$price / housing$bsqft  
housing$ppsf = ppsf

## Extraneous Variables

1. There are a few variables related to the quality of the data that are not mentioned in the list above. Figure out which these are and drop them from the data frame. Use subsetting by exclusion to do this.

ind = names(housing) != "quality" & names(housing) != "match" & names(housing) != "wk"  
housing = housing[ ,ind]

## Date

Since the housing prices changed a lot in the time period of observation, let's restrict our view to houses that were sold in the first two quarters of 2004. To do this, we need to convert *date* into a more appropriate data type. The following code creates a vector in a POSIX format and then pulls out the month and year from each date and saves them in numeric vectors.

month = as.numeric(format(housing$date, format="%m"))  
year = as.numeric(format(housing$date, format="%Y"))

1. Use these vectors to create a subset of *housing* with sales from the first half of 2004.

year\_logic = !is.na(year ==2004)  
month\_logic = !is.na(month <=6 )  
housing = housing[year\_logic & month\_logic, ]

You should have about 29,000 houses left in the data frame.

## Similar Locations

Another approach to exploring and analyzing data is to examine subsets that are similar. For example, we can `zoom in' to the east bay and examine the relationship between price per square foot and size of the house for houses sold in the following cities:

someCities = c("Albany", "Berkeley", "El Cerrito", "Emeryville",   
 "Piedmont", "Richmond", "Lafayette", "Walnut Creek",  
 "Kensington", "Alameda", "Orinda", "Moraga")

1. Use this character vector to take a subset of sales in those cities. You may find the operator *%in%* useful here.

sales = housing[housing$city %in% someCities, ]

The *city* variable is a factor and when all of the records for a particular city have been removed from the data, the level for that factor remains as a possible level. This means that summaries and other statistics will include these `empty' levels. To drop them, we can modify *city* as follows

housing$city = droplevels(housing$city)

Note that we expect that your subset of house sales is still in the data frame called *housing*. If not, you might want to reassign *housing* to your current data frame and remove (*rm()* ) all other intermediate and duplicate data frames.

## Plot

Our data should be ready to plotting. Below is plot that uses color to indicate density of points. You will learn how to make this and other plots soon. The following code chunk is set up so as to not run when knitted. You will want to remove the eval=TRUE from within the curly braces when you are ready to produce the final plot.

require(ggplot2)

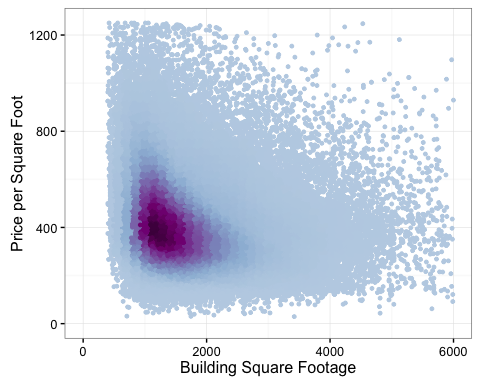
## Loading required package: ggplot2

require(RColorBrewer)

## Loading required package: RColorBrewer

housing$densCol = densCols(housing$bsqft,   
 y = log(housing$ppsf),   
 colramp = colorRampPalette(brewer.pal(9,"BuPu")[3:9]))  
  
ggplot(housing) +   
 geom\_point(aes(x = bsqft, y = ppsf, col = densCol), size = 1) +  
 scale\_color\_identity() +   
 xlim(0, 6000) + ylim(0,1250) +  
 xlab("Building Square Footage") + ylab("Price per Square Foot") +   
 theme\_bw()

## Warning: Removed 669 rows containing missing values (geom\_point).



Upload your completed Rmd file to bcourses. Do not upload the knitted file.