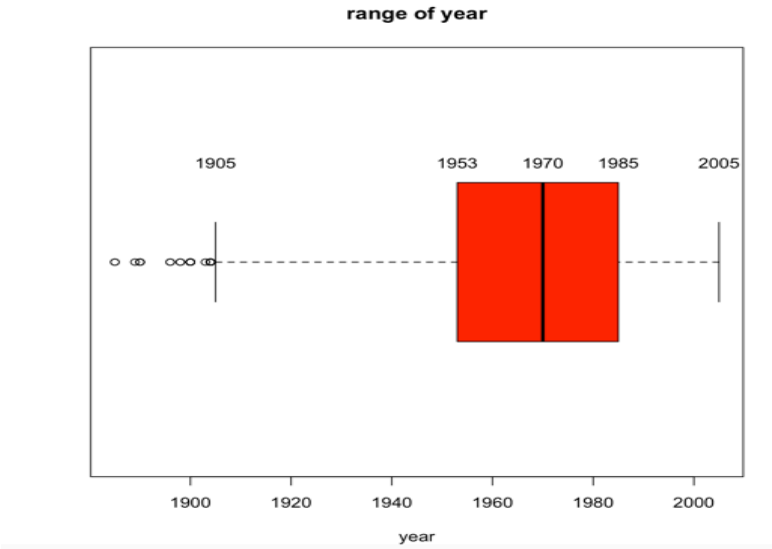
STA141A HW2 Report

Ruochen Zhong 912888970

Question 2

This housing sales cover 1134 days. It is from 2003-04-27 to 2006-06-04.

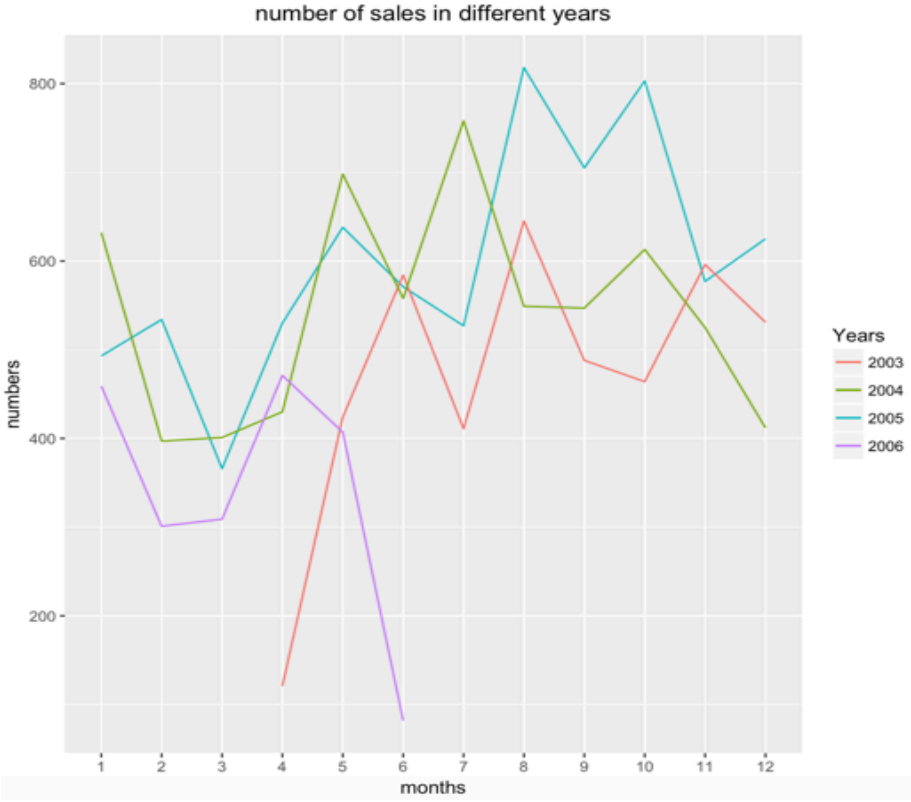
The following is the boxplot of the construction dates:



From this plot, it is clear to see that the latest year of construction is 2005. For the earliest year, there are some outliers lower than 1905. After checking those outliers, the earliest one is 1885. Therefore, the timespan of the construction dates is from 1885 to 2005.

Question 3

For the number of sales over time:



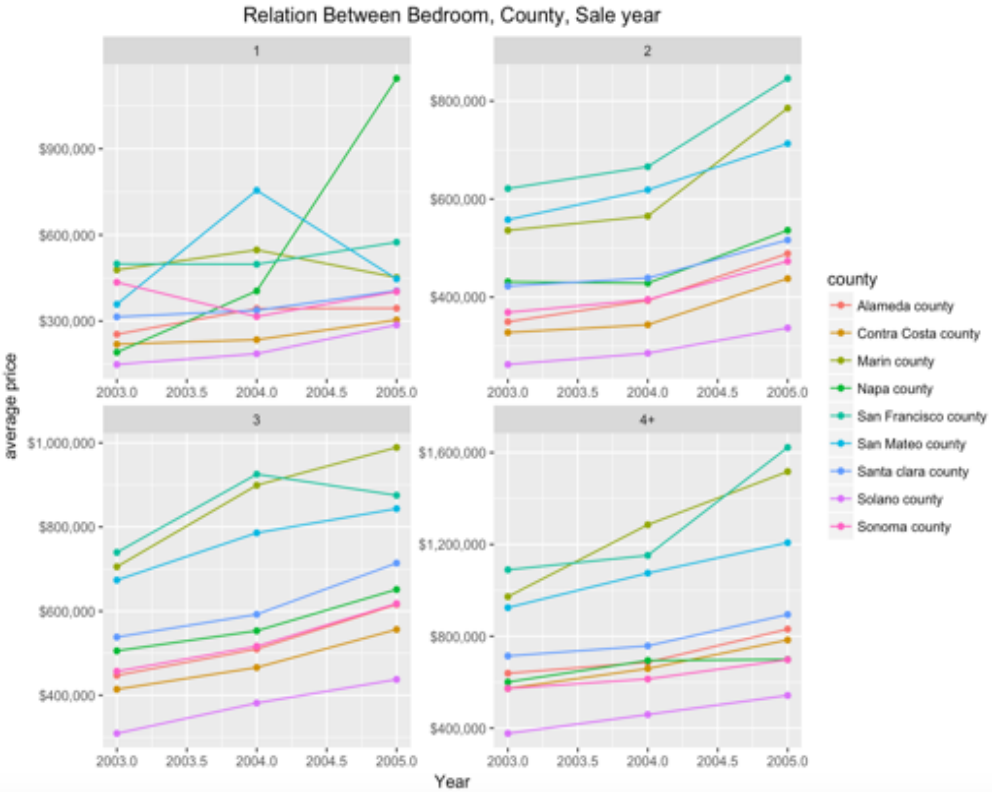
From this plot, it is clear to see that the number of sales has an increasing trend in 2003. It improves a lot from April to December. In 2004, the number of sales has many fluctuations. It decreases firstly and then increases to its maximum at July, but after that, it decreases a lot to nearly 400 at December. In 2005, the general trend is increasing. Especially in August and October, the number of sales is extremely high. In 2006, the number of sales suffers a serious decrease from May to June, and in June, it is the lowest number in all observed months.

For the average house price over time:

From this plot, it is clear to observe that in every year, the average price tends to have a general increasing trend over time. What's more, the next year's average house price is always higher than the previous year. Those information means the average price of housing increases over time from 2003 to 2006.

Considering these two graphs, we can see that the average house price has a relatively stable trend over time, but the number of sales tends to vary over time. Sometimes, a harsh increasing in housing price will lead to a harsh decreasing in number of sales (May 2006 to June 2006), but sometimes, the price and the number of sales will increase together.

Question 4



From this plot, it is clear to see that for all counties' same bedroom houses, their average price generally increases over time. If houses possess more bedrooms, their average prices tend to be higher. For different counties, **San Francisco county, Marin county, and San Mateo county** have higher price than others in all years and all kinds of bedrooms. The average price of **Solano**

county is the cheapest in all kinds of categories.

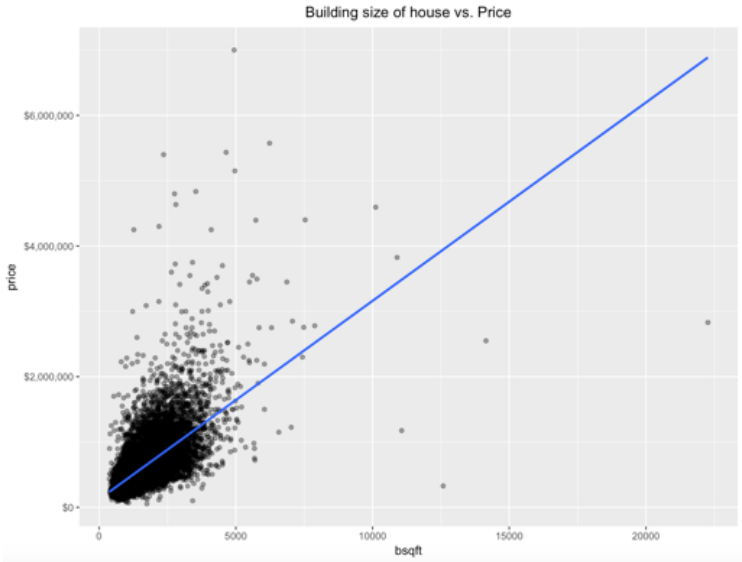
Question 5

No, not all housing sales within a given city only occur in one county. Because after checking the dataset, I found that city **Vallejo** has sales in more than one county. It has 2 sales in **Napa county** and 517 sales in **Solano county**.

After searching information from the Wikipedia, I find that Napa county and Solano county are close to each other, but because the Vallejo city belongs to the Solano county, so most of the sales in is classified to Solano county. However, the reason that those 2 who are classified to Napa county maybe they are in the border of those two counties and are classified into Napa county.

Question 6

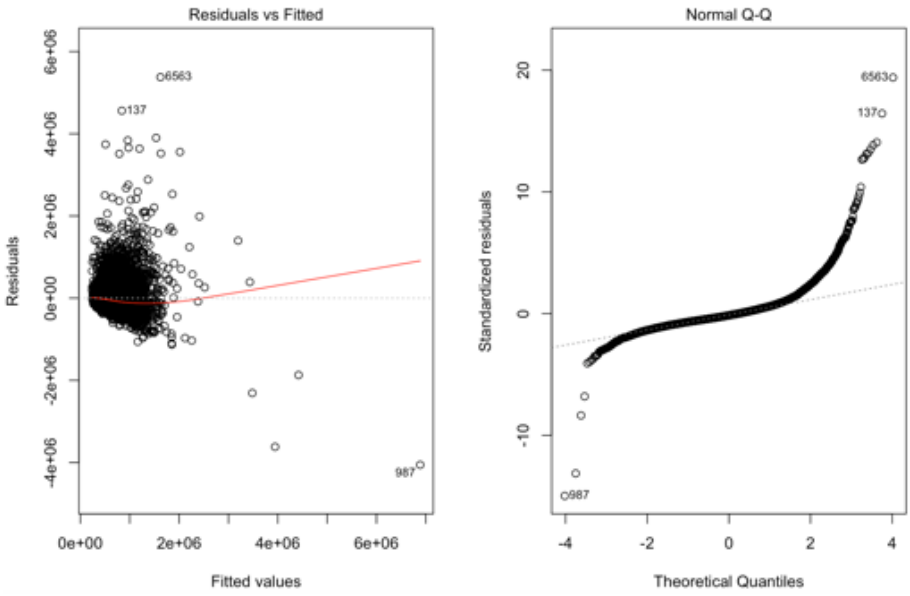
The following are the regression line and information for bsqft vs. price:



	Estimate	Std. Error	t-value
intercept	123700	5083	24.34
bsqft	303.7	2.884	105.3
Adjusted R^2	0.3938		

The straight line seems fit for the distribution of the observations.

Checking the Residuals vs Fitted graph and Q-Q plot of the regression:

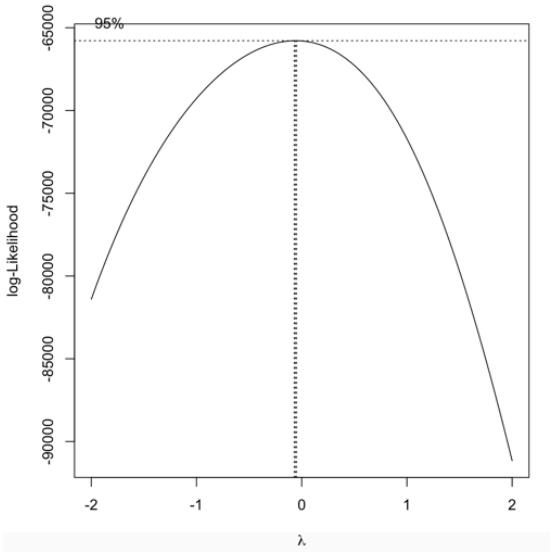


From these two plots, the variance of residuals seems to increase when the fitted values increase. In the Residuals vs. Fitted plot, the total mean value is approximately 0, but it seems that there are more dots above the horizontal 0 line. What's more, the variance seems not to be a constant: it increases as the fitted values increase. The QQ plot also shows a

heavy tailed distribution compared to the normal distribution. These means the original regression model needs a transformation.

Using box-cox transformation method, find the best λ = -0.0606, which is nearly 0.

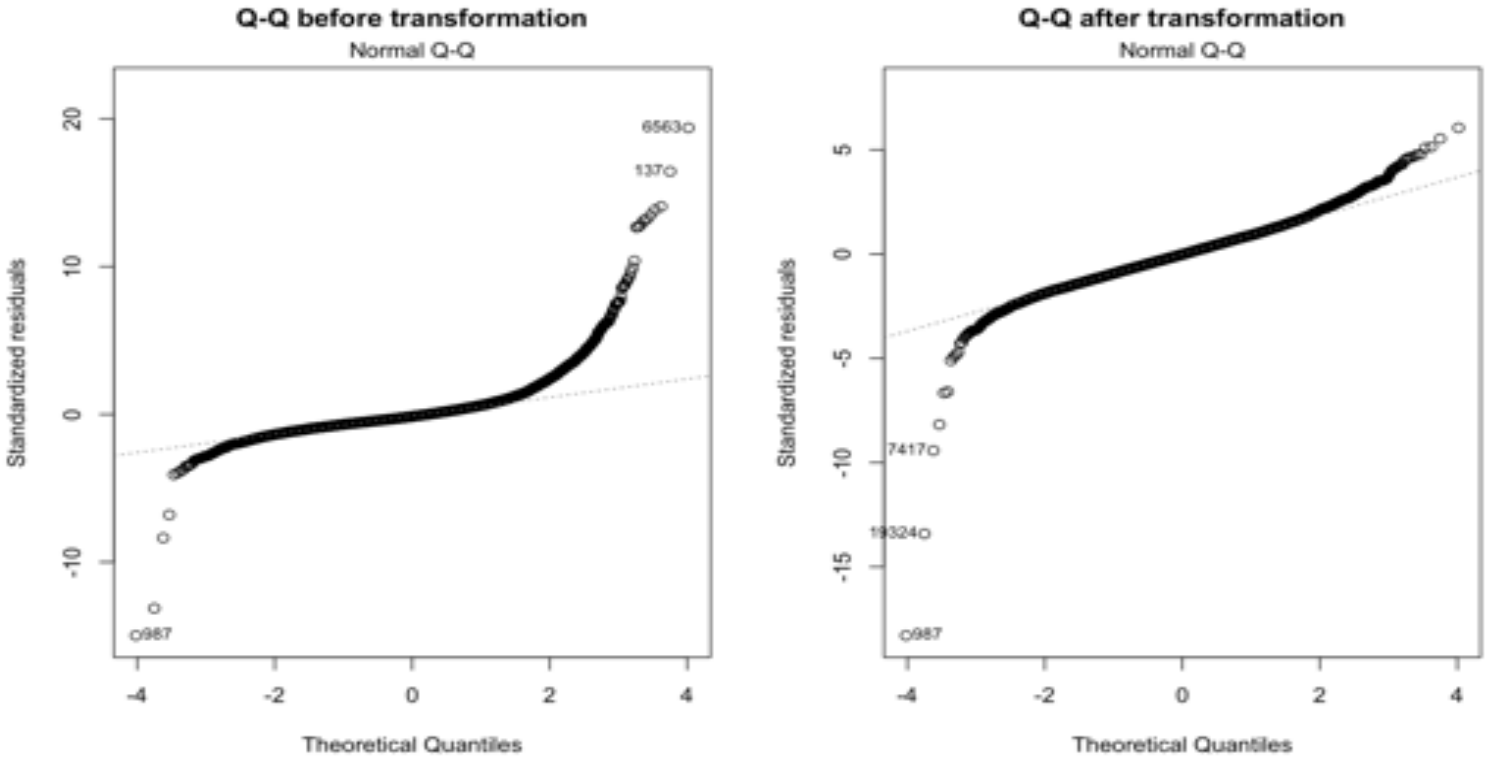
The following graph is the $\lambda\ vs.\ likelihood$:



When λ = 0, transform y to be log(y):

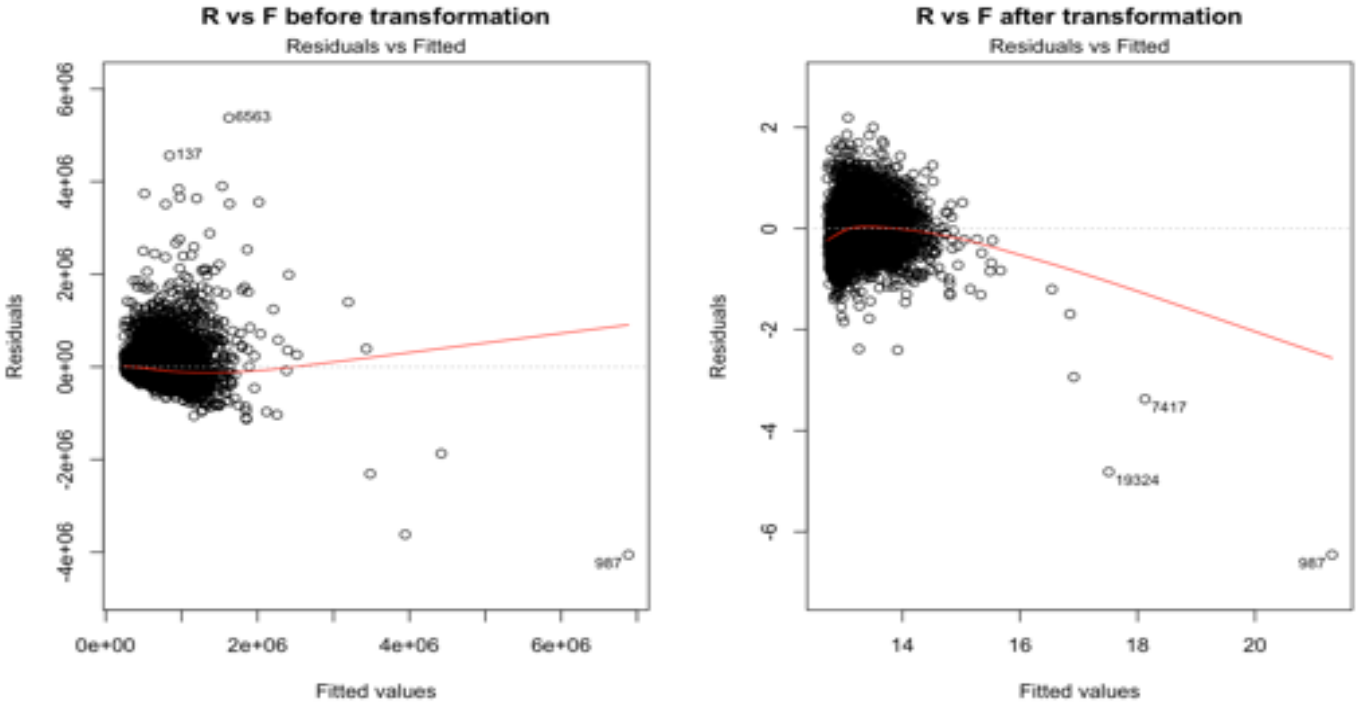
So, transforming the original regression model to: **log(price) = β_0 + $\beta_1$bsqft + ε**

Comparing diagnostics graphs after transformation and before transformation:



From these comparisons, it is clear to see that the box-cox transformation model improves the normality of

the residuals. Although there is still a slightly heavy tail, the tails are not so obvious relative to the Q-Q plot before transformation.



From the residual vs. fitted value plot, we can also find that before the transformation, there seems to be more dots above the zero line. In comparison, the distribution of error after transformation becomes more balanced and appropriate. However, the variance seems to decrease as the fitted value increases after the transformation, so the transformation still does not solve the problem of unequal variance.

Question 7

The following is the information of the multiple regression model:

$$\boldsymbol{price = \beta_0 + \beta_1 bsqft + \beta_2 lsqft + \varepsilon}$$

	Estimate	Std. Error	T value
Intercept	125300	5490	22.829
$\boldsymbol{\beta_1 bsqft}$	302.4	3.041	99.428
$\boldsymbol{\beta_2 lsqft}$	-0.0006119	0.0008344	-0.733

Construct a hypothesis test:

$\boldsymbol{H_0: \beta_1 bsqft - \beta_2 lsqft \geq 0}$, $\boldsymbol{H_1: \beta_1 bsqft - \beta_2 lsqft < 0}$

t-value = $\boldsymbol{\beta_1 bsqft - \beta_2 lsqft}$ / sqrt [var ($\boldsymbol{\beta_1 bsqft - \beta_2 lsqft}$)]

= [302.4 – (-0.0006119) / $\boldsymbol{sqrt(3.041^2 + 0.0008344^2)}$]

= 99.441

the critical value: t* = -qt(0.95, 19991 – 3) = -1.64493

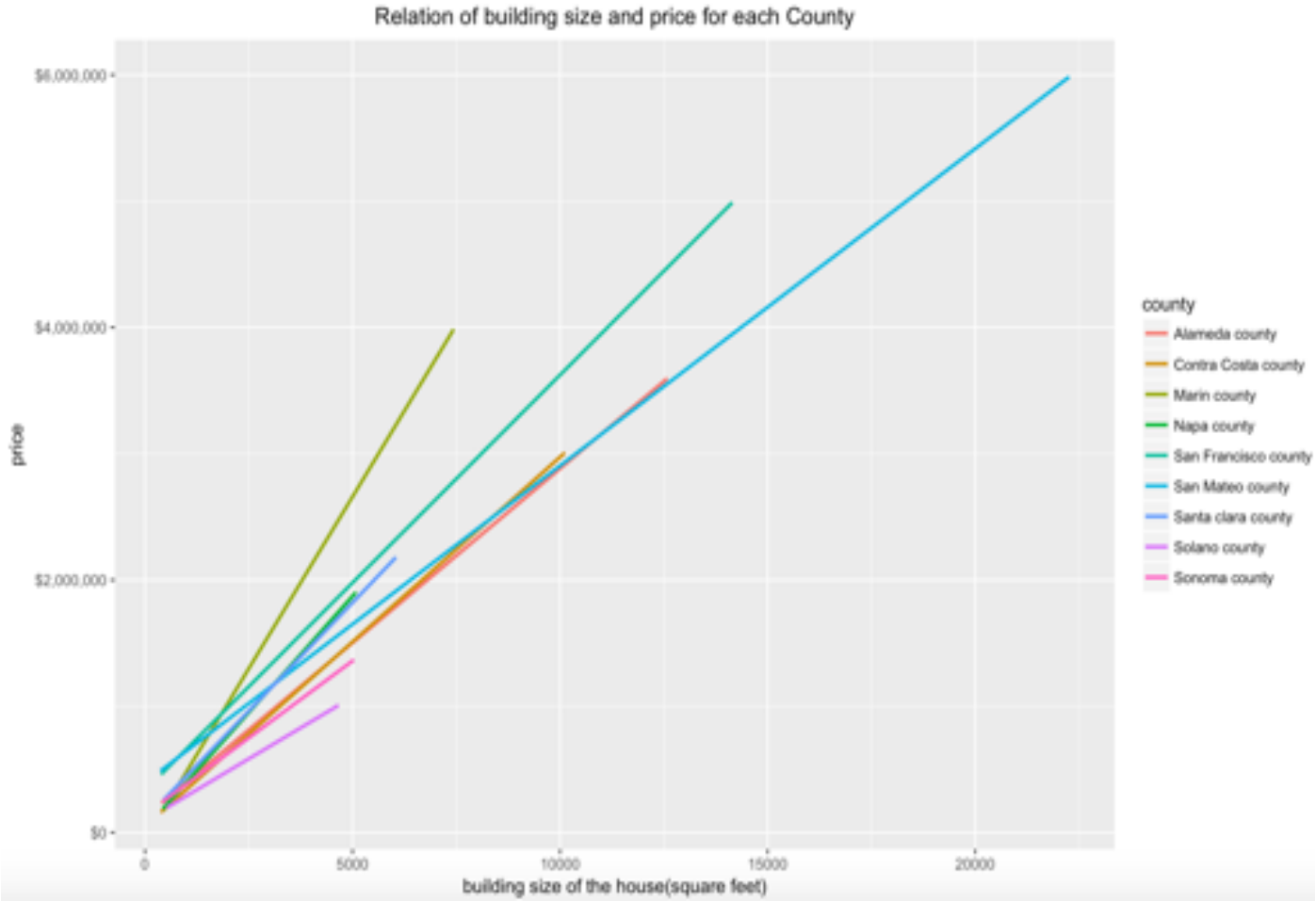
because t-value > t*, so cannot reject $\boldsymbol{H_0}$

This result means the building size of the house has more positive relationship with the price of the house than

the lot size of the house.

Question 8

The following is the regression line for every separate county:

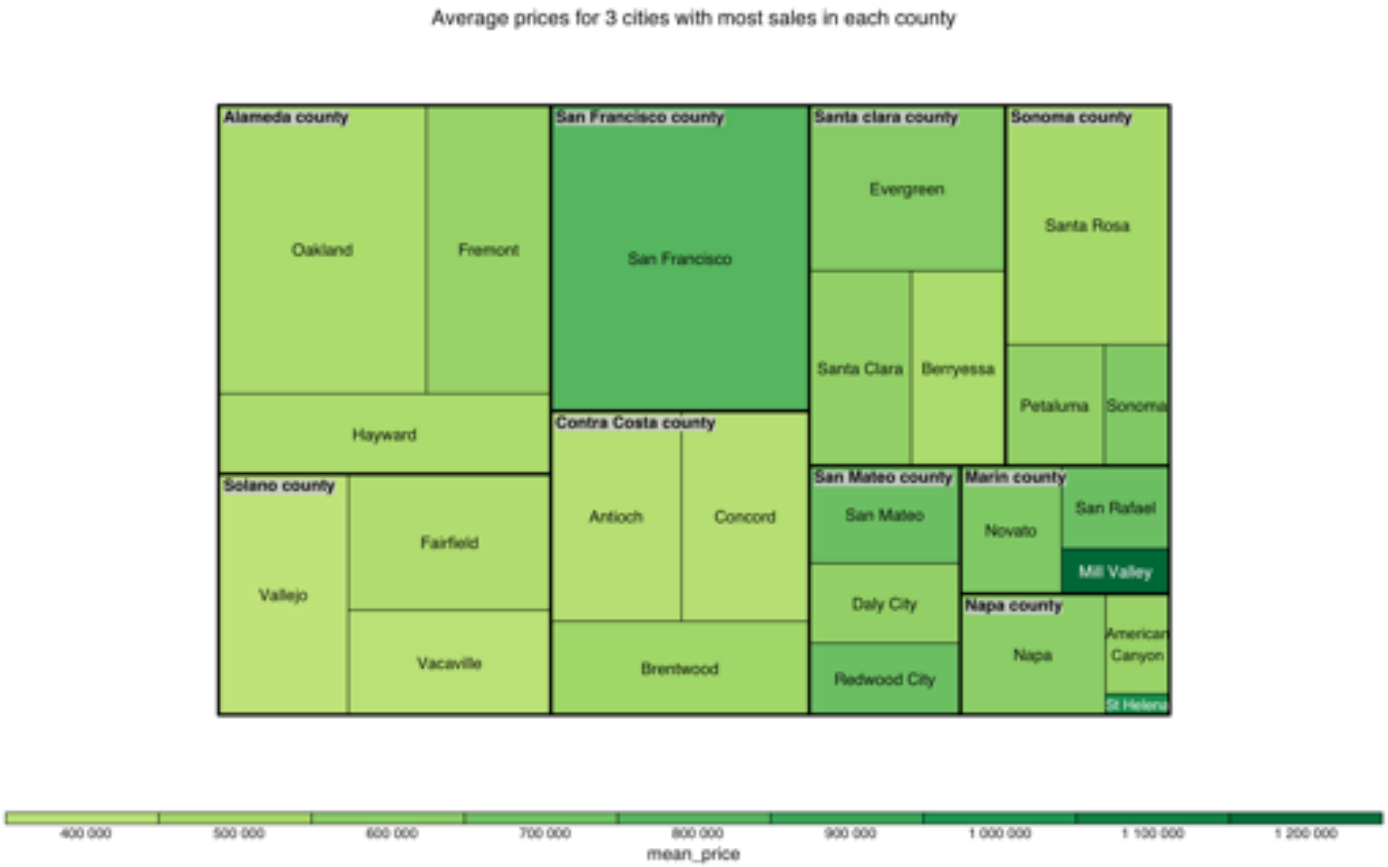


From this plot, it is clear to see that for every county, the building size of the house has a positive relationship with the price. However, those lines are not parallel to each other. This means in different county, their housing price per square feet are different from each other. Therefore, the county is a confounding variable.

For counties who has steeper slope of regression line, their housing price per square feet tends to be higher than those counties with flatter slope. After analyzing those lines, the **Marin county** seems to have the highest housing price per square feet. **San Francisco county, Napa county, and Santa Clara county** also have relatively high housing price per square feet. In comparison, **Solano county** and **Sonoma county** have a relatively low housing price per square feet among those 9 counties in California.

Question 9

The following is the treemap of average housing price for top 3 sales cities in each county:



This plot gives us plentiful information. Firstly, different counties have different size of rectangle in this treemap. The size of their rectangle is decided by the sum of their top 3 sales cities' sales number. For example, it is clear to see the rectangle of **Alameda county** is relatively bigger. This is because the sum of sales in Fremont, Hayward, and Oakland is relatively larger than other counties top 3 cities. In contrast, the sales of top 3 cities in **Marin county** and **Napa county** are relatively lesser, and their rectangles in this treemap are also smaller.

Secondly, in every counties' own rectangle, it includes some smaller rectangles. Those rectangles represent the sales of each city in their counties. For example, in **Sonoma county**, the Santa Rosa's rectangle is obviously bigger than Petaluma and Sonoma's rectangle. After checking the data, Santa Rosa has 650 sales, which is more than Petaluma's 198 sales and Sonoma' s 131 sales, so this lead to Santa Rosa has a bigger rectangle in the **Sonoma county.** Another interesting example is **San Francisco county**. This county does not have other cities, so the rectangle of San Francisco is the rectangle of **San Francisco county** itself.

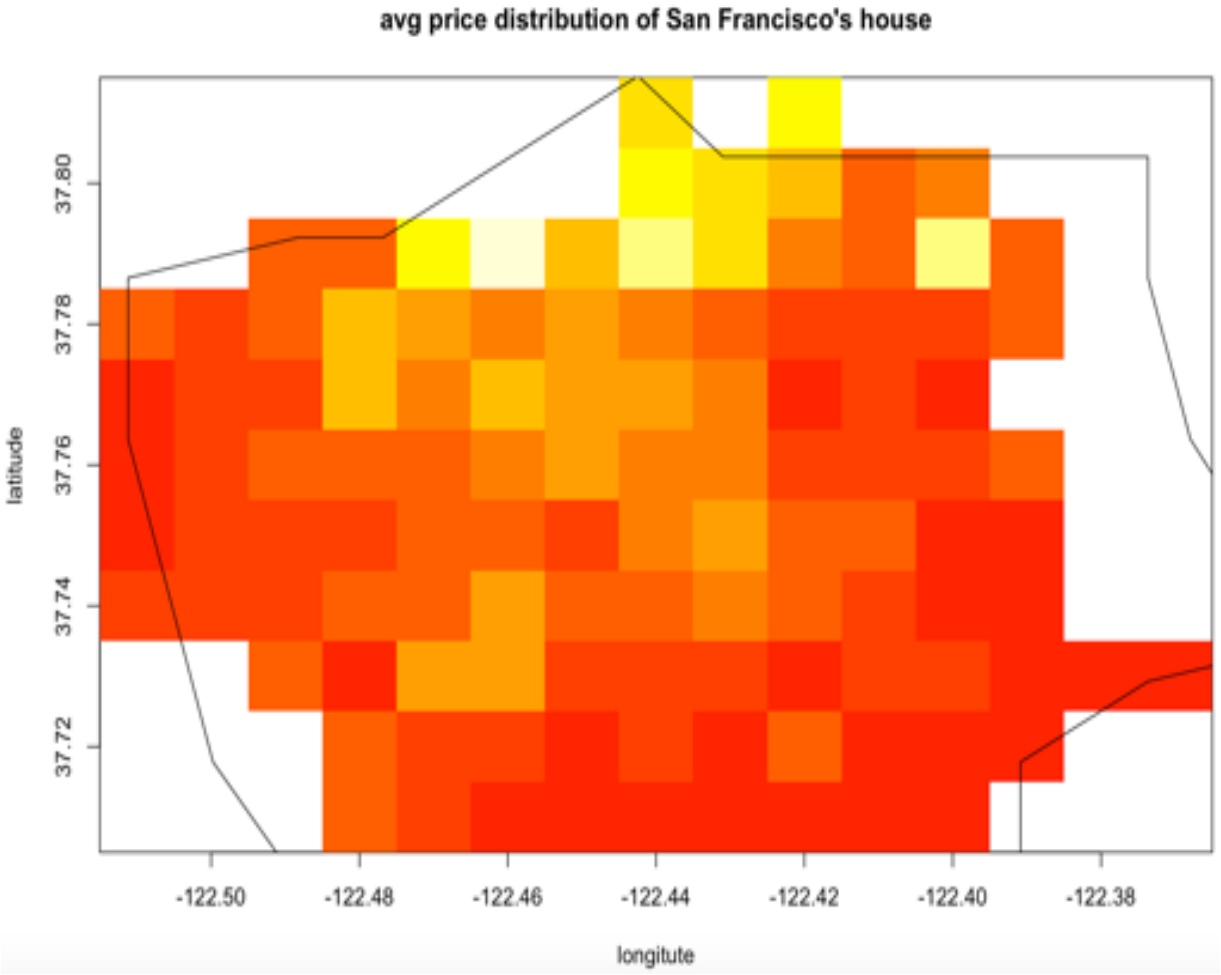
Thirdly, the different extents of shade represent the different level of mean housing price in that city. The darker the shade, the higher the mean housing price. The **Mill Valley** in **Marin county** has the highest mean price among those cities, which attains $1200,000, and cities in **Alameda county** and **Solano county** have lower mean price between $400,000 and $600,000. A significant finding of this feature is that in a same county,

the shade of 3 cities will not have a huge difference. In other words, in a same county, we will not observe one rectangle with very darker shade and another with very lighter shade. For example, in **Marin county**, all 3 rectangles' shades are all relatively dark, but in **Solano county**, they are all light. This feature is easy to understand: housing prices in a same region will not have large difference with each other.

A special case is that in **Question 5**, it shows city Vallejo has both sales in **Solano county** and **Napa county**, but in **Napa county,** it only has 2 sales, so Vallejo cannot be top 3 cities of 2 counties simultaneously in this treemap.

Question 10

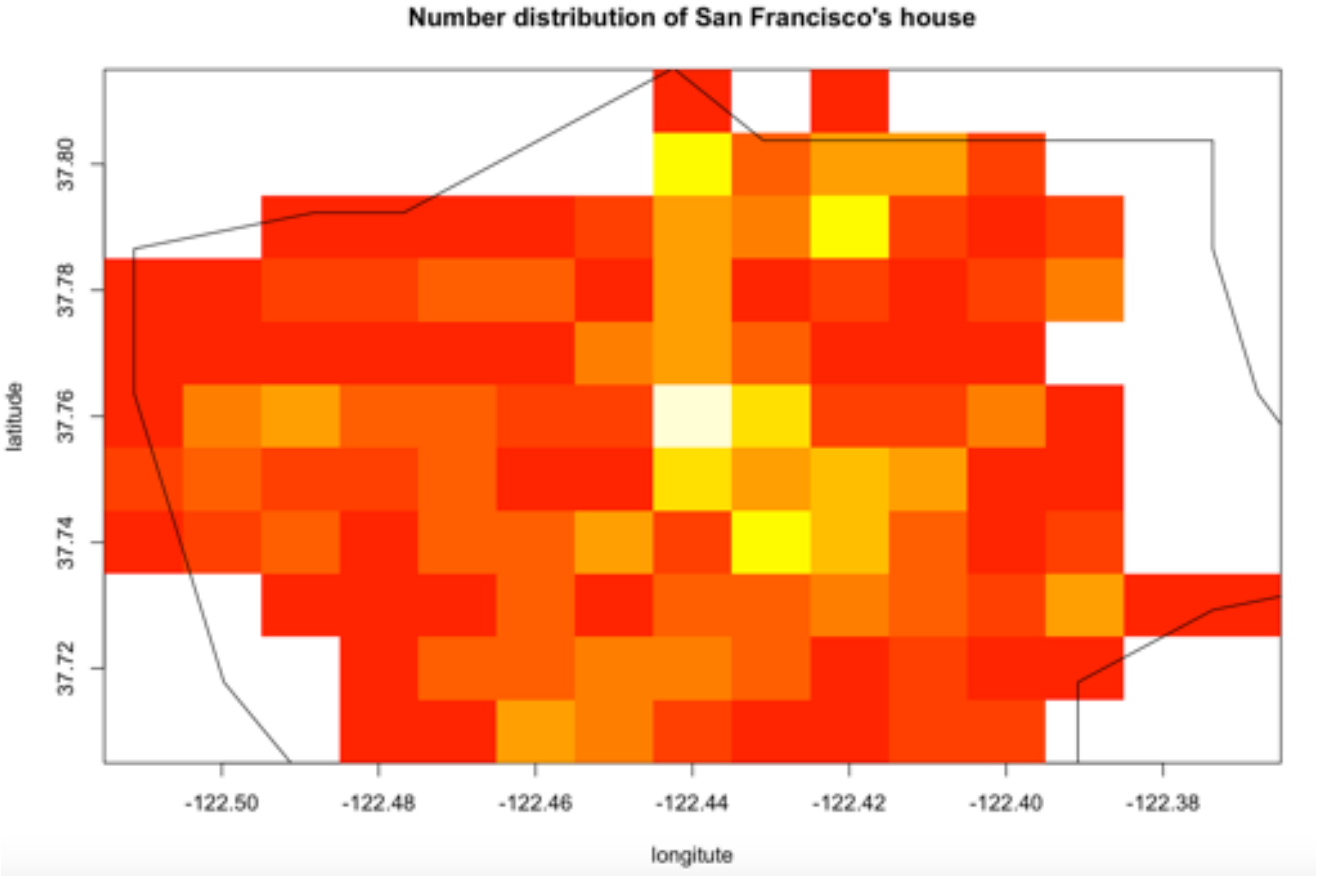
The following is the heatmap of average housing price in San Francisco:



From this heatmap, it is clear to see the distributions of the average housing price. If the color is closer to dark red, it means the average housing prices in that area are cheaper. In contrast, if the color is closer to light yellow, the average housing prices in that area are more expensive. According to this criteria, the heatmap indicates that the houses located in the south and west part of San Francisco are relatively cheaper than other parts. Houses within or around the range with longitude from -122.48 to -122.42 and latitude 37.80 are the more expensive than other houses in San Francisco. I think the reason is because this area is near the **Golden Gate Bridge,** so living in there will be more convenient. Therefore, the housing prices near that area will be

relatively high.

The following is the heatmap of number of houses in San Francisco:



The colorized criteria of number distribution is same as the average housing price heatmap, which means if the color is closer to dark red, it means the number of houses in that area are lesser. In contrast, if the color is closer to light yellow, the number of houses in that area are more. From this heatmap, it indicates that in the margin part of San Francisco, there are often less number of houses. The regions with high number of houses locate at the middle part of San Francisco, and two regions in the north part of San Francisco. In general, the regions with high number of houses in San Francisco don't have a clear pattern.

Appendix

Citation:

Question 2: Imitate the code to label numbers of quantiles in a boxplot from the website: https://stackoverflow.com/questions/13945434/how-to-put-values-on-a-boxplot-for-median-1st-quartile-and-last-quartile

Question 3: learn how to center ggplot title in: https://stackoverflow.com/questions/40675778/center-plot-title-in-ggplot2

And learn how to change y axis scales in: http://ggplot2.tidyverse.org/reference/scale_continuous.html

Question 5

Using the method which classmates discussed on the Piazza to do

Question 6

Watch the video to learn how to use boxcox method in: https://www.youtube.com/watch?v=TgVx9Rqsewo

Question 8

Learn how to draw regression lines by group in: https://stackoverflow.com/questions/12281335/adding-regression-line-per-group-with-ggplot2

Question 9

Learn how to use transform() function to add a new column in a dataframe in: https://stackoverflow.com/questions/15977046/add-column-to-data-frame-using-transform-and-calling-a-function-an-annoying-w

Use the code Patrick posted on Piazza to change the location of the county title in treemap

Question 10

Use the code Patrick posted on Piazza to create two factor variables and transform the dataset to be a matrix.

R code

```
hw2data = readRDS("/Users/apple/Desktop/housing.rds")


library(lubridate)
library(ggplot2)
library(dplyr)
library(MASS)
library(treemap)
library(maps)



##### Q1 #####
##### check the whole structure of the data
str(hw2data)
class(hw2data$date)
##### change the class of the variable "date", "year", and "city"
hw2data$date = as.Date(hw2data$date)
class(hw2data$date)
hw2data$year = as.integer(hw2data$year)
hw2data$city = as.character(hw2data$city)
str(hw2data)



##### Q2 #####
##### Find the timespan of the housing sales
range(hw2data$date)
difftime(max(hw2data$date), min(hw2data$date), units = 'days')
##### Find the range of the construction date
range(hw2data$year, na.rm = TRUE)
sort(hw2data$year)
sort(hw2data$year, decreasing = TRUE)
##### Some observations' construction date is not appropriate, ignoring those observations
appropriate_year <- subset(hw2data, (hw2data$year > 1800) & (hw2data$year < 2100))
##### check the timespan by boxplot
```

```
boxplot(appropriate_year$year, horizontal = TRUE,
        col = 'red', main = "range of year", xlab = "year")

text(x = boxplot.stats(hw2data$year)$stats, labels = boxplot.stats(hw2data$year)$stats, y = 1.25)

sort(appropriate_year$year)

##### Q3 #####
##### check there are how many years and months
table(month(hw2data$date))
table(year(hw2data$date))
##### create two new variables years, and month
hw2data$year2 <- year(hw2data$date)
hw2data$month <- month(hw2data$date)

##### For the number of sales over time
##### split by year
data_2003 <- subset(hw2data, year(hw2data$date) == 2003)
data_2004 <- subset(hw2data, year(hw2data$date) == 2004)
data_2005 <- subset(hw2data, year(hw2data$date) == 2005)
data_2006 <- subset(hw2data, year(hw2data$date) == 2006)
##### construct data.frame for each year
dt_2003 <- as.data.frame(table(month(data_2003$date)))
colnames(dt_2003) = c("months", "numbers")

dt_2004 <- as.data.frame(table(month(data_2004$date)))
colnames(dt_2004) = c("months", "numbers")

dt_2005 <- as.data.frame(table(month(data_2005$date)))
colnames(dt_2005) = c("months", "numbers")

dt_2006 <- as.data.frame(table(month(data_2006$date)))
colnames(dt_2006) = c("months", "numbers")
```

```
##### draw the ggplot in one graph
p <- ggplot() + geom_line(data=dt_2004, aes(x=months, y = numbers, group=1, colour = 'pink'))+
      geom_line(data=dt_2003, aes(x=months, y = numbers, group=1, colour = 'blackblue')) +
      geom_line(data=dt_2005, aes(x=months, y = numbers, group=1, colour = 'green')) +
      geom_line(data=dt_2006, aes(x=months, y = numbers, group=1, colour = 'yellow'))

##### correct the legend and add a title
p.update <- p + scale_color_discrete(name = "Years", labels = c("2003", "2004","2005", "2006")) +
                ggtitle("number of sales in different years") +
                theme(plot.title = element_text(hjust = 0.5))
print(p.update)

##### For the average house price
##### create data frame for each year's price
df_price2003 = aggregate(price ~ month(date), data_2003, mean)
colnames(df_price2003) = c("months", "avg_prices")

df_price2004 = aggregate(price ~ month(date), data_2004, mean)
colnames(df_price2004) = c("months", "avg_prices")

df_price2005 = aggregate(price ~ month(date), data_2005, mean)
colnames(df_price2005) = c("months", "avg_prices")

df_price2006 = aggregate(price ~ month(date), data_2006, mean)
colnames(df_price2006) = c("months", "avg_prices")

##### Draw the ggplot
p2 <- ggplot() + geom_line(data=df_price2004, aes(x=months, y = avg_prices, group=1, colour = 'green'))+
  geom_line(data=df_price2003, aes(x=months, y = avg_prices, group=1, colour = 'blackblue')) +
  geom_line(data=df_price2005, aes(x=months, y = avg_prices, group=1, colour = 'pink')) +
  geom_line(data=df_price2006, aes(x=months, y = avg_prices, group=1, colour = 'yellow'))
##### correct the legend and add a title and labels // need to cite!!!
p2.update <- p2 + scale_color_discrete(name = "Years", labels = c("2003", "2004","2005", "2006")) +
  scale_y_continuous(labels = scales::dollar) +
```

```
  scale_x_continuous(breaks = c(1,2,3,4,5,6,7,8,9,10,11,12)) +
  ggtitle("average house price in different years") +
  theme(plot.title = element_text(hjust = 0.5))

print(p2.update)

##### Q4 #####
##### Do some data correcting firstly
summary(hw2data$county)
table(hw2data$county)
##### Correct the county name
hw2data$county <- gsub("C", "c", hw2data$county)
##### Correct "San Franciscoe"
which(hw2data$county == "San Franciscoe county")
hw2data[1421,]$city
hw2data$county <- gsub("San Franciscoe county", "San Francisco county", hw2data$county)
##### Correct "Alpine county"
which(hw2data$county == "Alpine county")
hw2data[1190,]$city
hw2data$county <- gsub("Alpine county", "San Francisco county", hw2data$county)
##### Convert contra costa to upper letter
hw2data$county <- gsub("contra costa county", "Contra Costa county", hw2data$county)
table(hw2data$county)

##### Create a subset
relation_dt <- subset(hw2data, select=c("county", "br", "price", "date"))
##### Excluding the sale year 2006 and br = NA
relation_dt<-relation_dt[!(year(relation_dt$date) == 2006),]
relation_dt<-relation_dt[!(is.na(relation_dt$br) == "TRUE"),]
##### Replace br >= 4 with 4+
relation_dt$br[relation_dt$br >= 4] <- "4+"
##### Create a variable sale_year
relation_dt$sale_year <- year(relation_dt$date)
```

```
##### Use dplyr pakage to group data
relation_update <-relation_dt %>%
                        group_by(county , sale_year, br) %>%
                        summarise_at(vars(price), mean)
print(relation_update)

##### Draw the ggplot
p3 <- ggplot(relation_update, aes(x = sale_year, y = price, color = county)) +
      geom_line() +
      geom_point() +
      facet_wrap(~br, scales = 'free')
print(p3)
##### Make the ggplot seems better and add titles
p3.update <- p3 + scale_y_continuous(labels = scales::dollar) +
                  labs(x = "Year", y = "average price") +
                  ggtitle("Relation Between Bedroom, County, Sale year") +
                  theme(plot.title = element_text(hjust = 0.5))
print(p3.update)

##### Q5 #####
##### Create a subset to focus on county, city
county_city <- subset(hw2data, select = c("county", "city"))
##### Create a table and caculate the unique match of city and county
table(county_city[c("city","county")])
unique_table <- table(unique(county_city[c("city","county")]))
print(unique_table)
##### Find which city don't match only a unique county
which(rowSums(unique_table)>1)
city_Vallejo <- subset(county_city, city == "Vallejo")
##### Only Vallejo ! it has sales in two counties, table them
table(city_Vallejo$county)

##### Q6 #####
##### Exclude rows whose price = 0 in a subset
```

```
which(hw2data$price == 0)
bc_subset <- subset(hw2data, hw2data$price != 0)
which(bc_subset$price == 0)
##### Create a regression model
reg1 <- lm(bc_subset$price ~ bc_subset$bsqft)
summary(reg1)
##### Draw a plot of this regression model
p4 <- ggplot(bc_subset, aes(x = bsqft, y = price)) +
        geom_point( alpha = 0.4) +
        geom_smooth(method = "lm", se = FALSE)

print(p4)

p4.update <- p4 + scale_y_continuous(labels = scales::dollar) +
                  ggtitle("Building size of house vs. Price") +
                  theme(plot.title = element_text(hjust = 0.5))

print(p4.update)
##### check the diagnostics
par(mfrow=c(1,2))
plot(reg1, which = 1)
plot(reg1, which = 2)
##### use box-cox transformation to check
box_cox <- boxcox(bc_subset$price ~ bc_subset$bsqft)
##### find the max value of the lamda
lamda <- box_cox$x
likelihood <- box_cox$y
bc <- cbind(lamda, likelihood)
bc[order(-likelihood),]
##### The best lamda is near 0, so use the lamda = 0 to update regression line and check its normality
bc_subset$price <- as.numeric(bc_subset$price)
reg1.update <- lm(log(bc_subset$price) ~ bc_subset$bsqft)
plot(reg1.update)
##### Compare graphs before and after the box-cox transformation
```

```
par(mfrow=c(1,2))

plot(reg1, which = 1, main = "R vs F before transformation")
plot(reg1.update, which = 1, main = "R vs F after transformation")

plot(reg1, which = 2, main = "Q-Q before transformation")
plot(reg1.update, which = 2, main = "Q-Q after transformation")

##### Q7 #####
reg2 <- lm(hw2data$price ~ hw2data$bsqft + hw2data$lsqft)
summary(reg2)
###test Ho: beta(bsqft) - beta(lsqft) >= 0
beta_diff <- 302.4 - (-0.0006119)
var_diff <- (3.041)^2 + (0.0008334)^2
T_value <- (beta_diff) / (sqrt(var_diff))
#### compare to the critical value
-qt(0.95, 19988)

##### result: cannot reject Ho, T_value is larger than the critical value

##### Q8 #####
#### Draw the linear regression line for each county
ind_reg <- subset(hw2data, select=c("county", "bsqft", "price"))
p5 <- ggplot(ind_reg, aes(x = bsqft, y = price, color = county)) +
        geom_smooth(method="lm", se = FALSE)
#### make the ggplot
p5.update <- p5 + scale_y_continuous(labels = scales::dollar) +
                  labs(x = "building size of the house(square feet)", y = "price") +
                  ggtitle("Relation of building size and price for each County") +
                  theme(plot.title = element_text(hjust = 0.5))
print(p5.update)

##### Q9 #####
##### create a subset to study the city, county and price
```

```
data_city <- subset(hw2data, select = c("county", "city", "price"))
##### rank the top 3 cities solds of each county and caculate their mean price
top_three <- data_city %>%
                group_by(county, city) %>%
                summarise(n = n()) %>%
                top_n(n = 3, wt = n)
##### extract top 3 city's price data from the original subset
data_city$ideal <- (data_city$city %in% top_three$city)&(data_city$county %in% top_three$county)
ideal_data <- subset(data_city, data_city$ideal == 'TRUE')
##### caculate their average price
ideal <- ideal_data %>%
          group_by(city) %>%
          summarise(avg_price = mean(price))
##### change the order of column and combine them
top_three <- top_three[order(top_three$city),]
top_three <- transform(top_three, newcol=paste(ideal$avg_price, sep="_"))
colnames(top_three)[4] <- "mean_price"
top_three <- top_three[order(top_three$county),]
top_three$mean_price <- as.numeric(as.character(top_three$mean_price))
##### use the ideal dataset to draw the treemap
treemap(top_three,
        index=c("county", "city"),
        vSize= "n",
        vColor="mean_price",
        type="value",
        format.legend = list(scientific = FALSE, big.mark = " "),
        title="Average prices for 3 cities with most sales in each county",
        align.labels=list(c("left","top"),c("center","center"))
        )

##### Q10 #####
##### extract data for Sanfrancisco
SFdata <- subset(hw2data, hw2data$county == "San Francisco county")
##### transformation the location variables
```

```
SFdata$long2 <- round(SFdata$long, 2)
SFdata$lat2 <- round(SFdata$lat, 2)
##### create appropirate factors and specify the levels
long_range<-range(SFdata$long2,na.rm=TRUE)
long_seq<-seq(long_range[1],long_range[2],.01)

lat_range<-range(SFdata$lat2,na.rm=TRUE)
lat_seq<-seq(lat_range[1],lat_range[2],.01)

SFdata$long_fac<-factor(SFdata$long2,levels=long_seq)
SFdata$lat_fac<-factor(SFdata$lat2,levels=lat_seq)

##### create a dataframe by grouping latitude and longitude with price
price_location <-aggregate(price~lat_fac+long_fac, SFdata,function(x)c(mean(x),length(x)),drop=FALSE)
##### change the dataframe to matrix
house_prices<-
matrix(price_location$price[,1],nlevels(SFdata$long_fac),nlevels(SFdata$lat_fac),byrow=TRUE)

##### draw a image and add the lines of boundary
image(x=long_seq,y=lat_seq,z=house_prices, xlab = "longitute", ylab = "latitude",
      main = "avg price distribution of San Francisco's house")
#points(SFdata$long, SFdata$lat)
sfmap<-map("county",plot=FALSE)
lines(sfmap$x-.03,sfmap$y)

##### then, grouping with numbers
number_location <- table(SFdata$long_fac,SFdata$lat_fac)
number_location
number_location[number_location == 0] <- NA
number_location
##### draw a image and add the line of boundary
image(x=long_seq,y=lat_seq,z= number_location, xlab = "longitute", ylab = "latitude",
      main = "Number distribution of San Francisco's house")
#points(SFdata$long, SFdata$lat)
```

```
sfmap<-map("county",plot=FALSE)
lines(sfmap$x-.03,sfmap$y)
```