STA 141A Data Analysis Report
Liya Li

Q1:

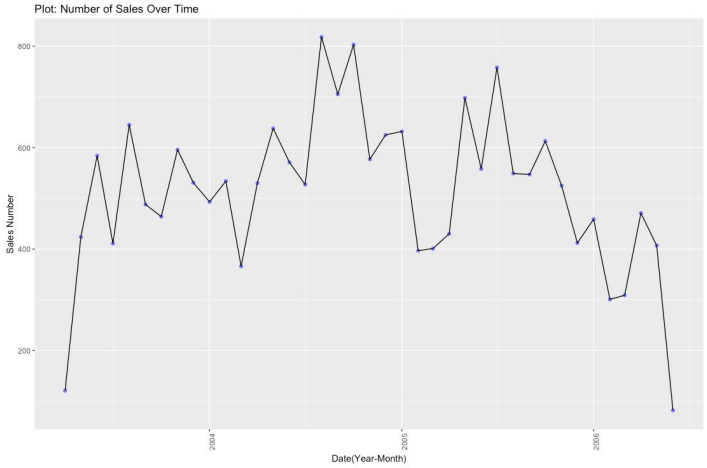
Code with comments provided at appendix.

Q2:

The housing sales timespan covers from 2003-04-27 to 2006-06-04.
The construction dates of homes timespan covers from 1885 to 2005.

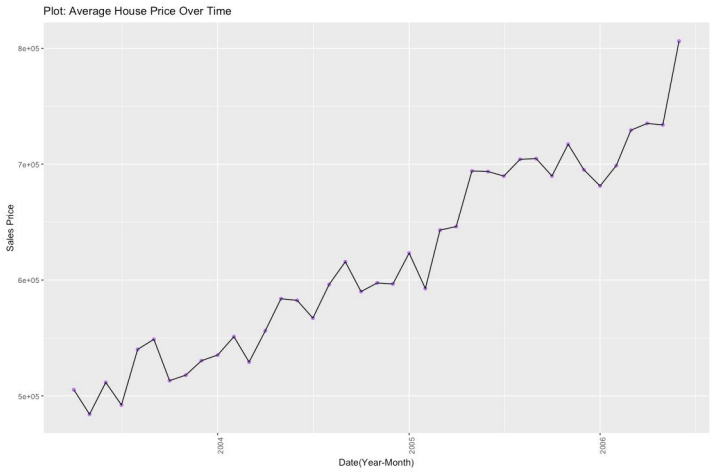
Q3:

The plot shows number of sales over time is as following:



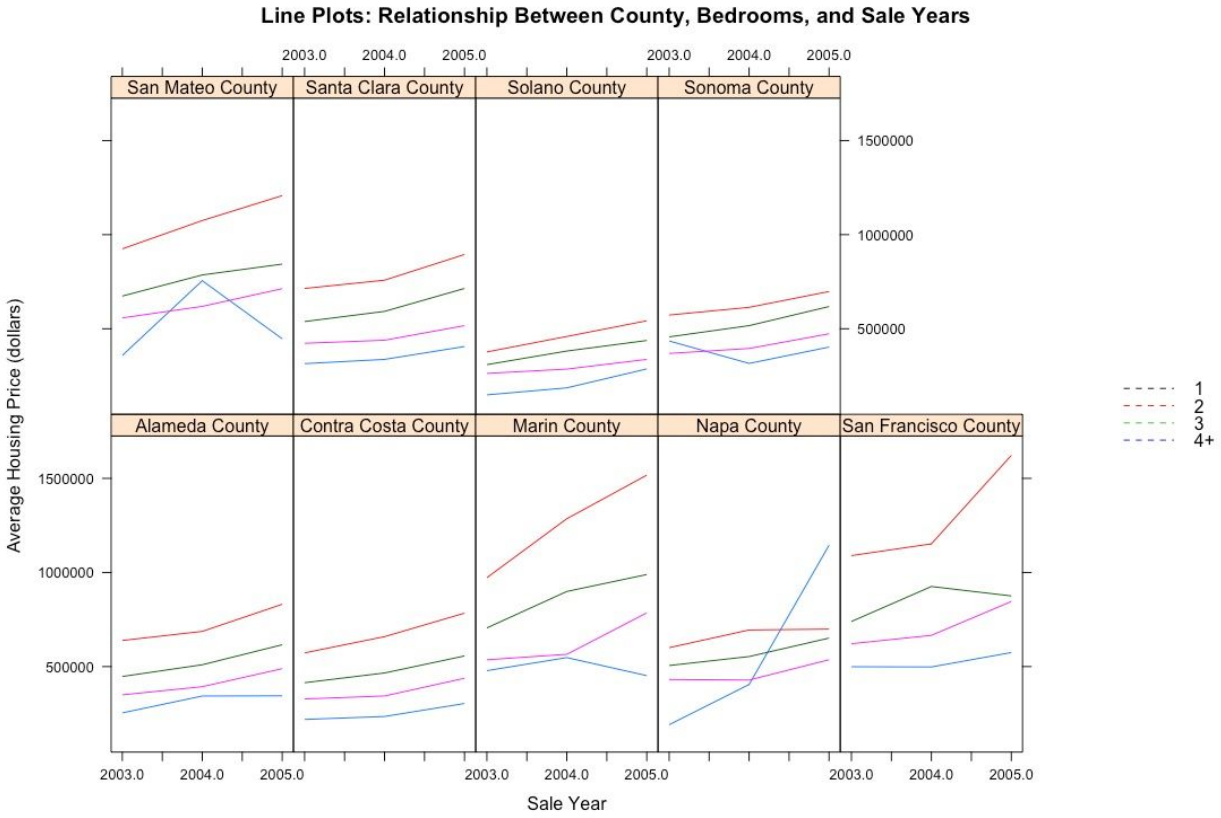
From this plot, we can see that the number of sales over time doesn't have huge changes. But in 2005, there seems to be some higherest sales peaks than 2003 and 2004, where in 2004 there are some months having sales peaks higher than 2003. Approximately, all years have almost all months house sales number over 400.

The plot shows the average house price over time is as following:



From this plot, we can see that the average house price has an increasing trend over time from around 500000 in 2003 to over 800000 in 2005. There are some decreases within years, but in total trend, the average price keeps increasing from 2003 to 2005.

Q4:



This plot is the line plots on relationship between counties, bedrooms, and sale years. We can clearly see that in each county, from 2003 to 2005, how the sale price change in terms of bedroom number. For example, in Santa Clara County, four different lines lie according to different sales price, but the lines are approximately parallel, meaning that the differences in price between bedroom numbers doesn't change much over time in this county. Also, we can see that 2 bedrooms houses in county San Francisco, Marin and San Mateo have much more higher price than it in other county.

Q5:

Not all housing sales within a given city only occur in one county.

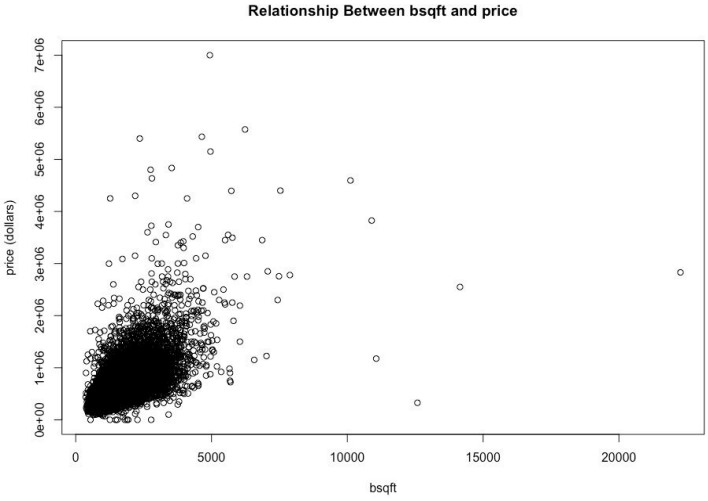
By drawing a table in R, we use the cross-classifying factors city and cleaned county to build a contingency table of the counts at each combination of factor levels. In the table, entry 0 means no such city and county combined, entry with other values means there is such combination with that amount.

By using a function to print out all row indexes of more than one occurrence cities, we first loop each row in the table. We find that if all housing sales within a given city only occur in one county, then each row sum should equal to the maximum entry value in that row. The function returns 157, where row 157 is city Vallejo.

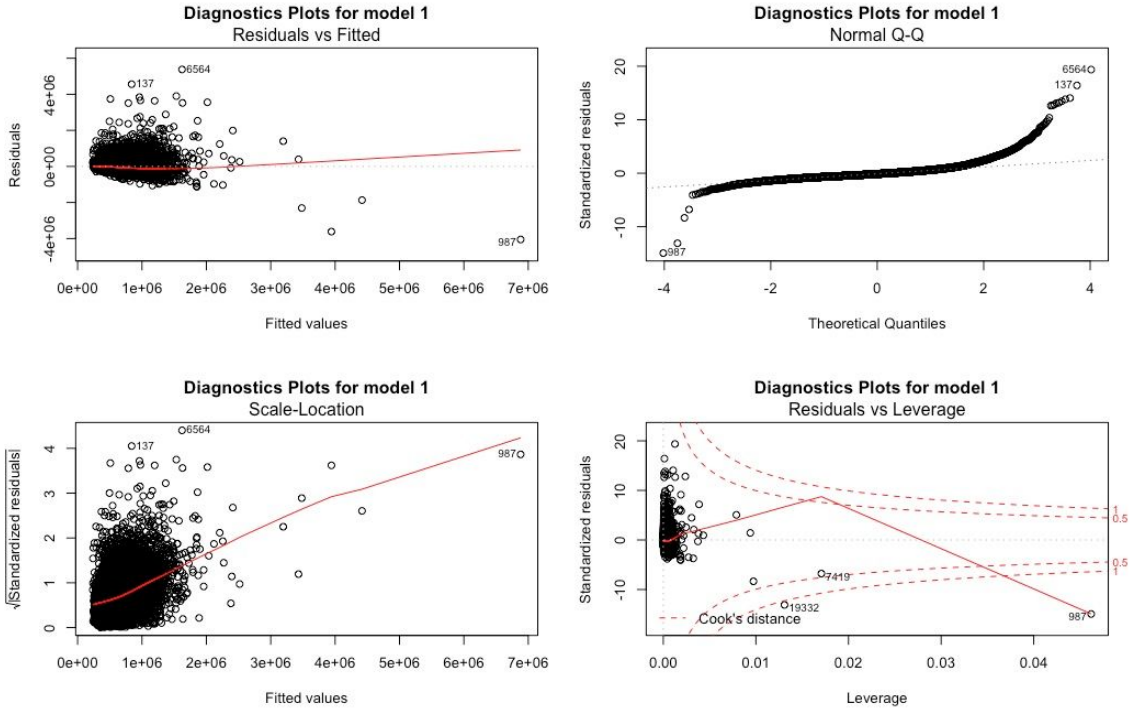
Therefore, Vallejo is the only one city has housing sales within Vallejo occur in both Napa county and Solano county. We check from Wikipedia under California map, the city Vallejo is in Solano county but borders Napa county. So a house sold in the outskirts of Vallejo may report "Vallejo" as city and "Solano" as county. This would be why Vallejo has sales in both Napa and Solano counties.

Q6:

After fitting the regression model model1 with original data, we plot the original data plot. We notice that the data points seem not linearly spread and there are probably some outliers.

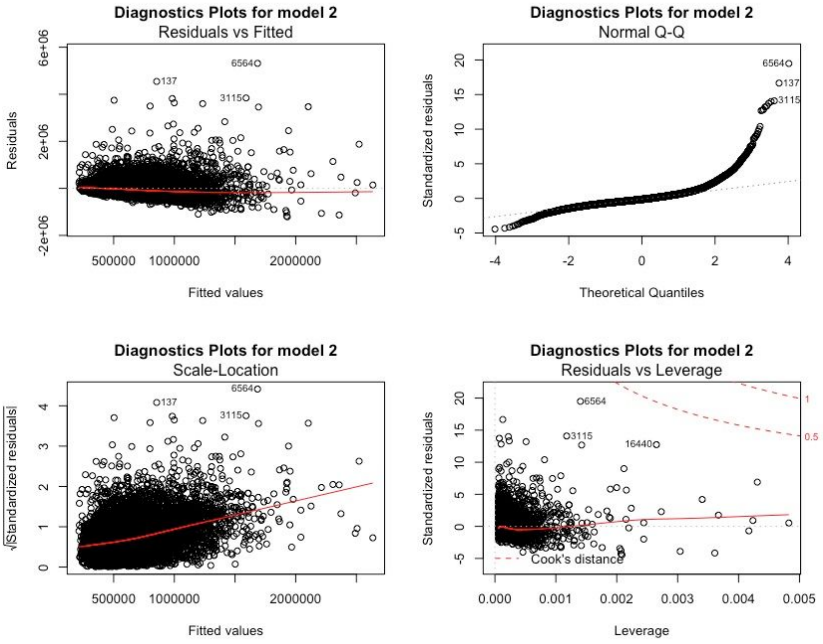


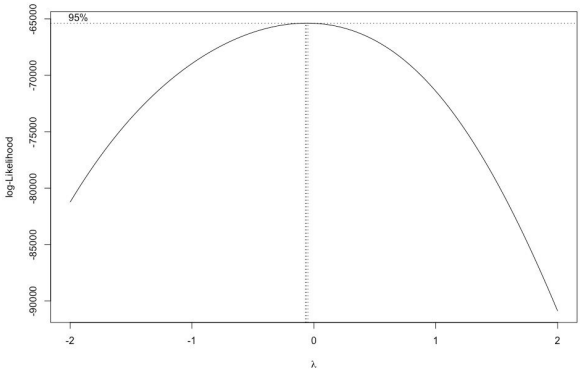
Then we plot the diagnostic plots for model1. In the Residuals vs Fitted plot, the red line is flat but with most data points concentrate on the left, meaning that the linearity assumption holds with concern. In the Scale-Location plot, the red line is tilted and data points are not equally spread along the line, meaning that the equal variance assumption doesn't hold. In the Normal Q-Q plot, we notice both tails are heavier than the standard normal regression line, meaning that the normality assumption also doesn't hold. In the Residuals vs Leverage plot, we see that point indexed at 987, 7419 and 19332 might be considered as extreme outliers. By plotting the Cook's distance plot, we also obtain the same outlier detection. So we remove these three observations.



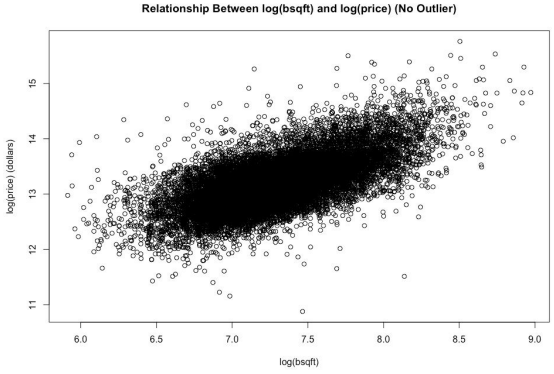
From question 2, we found irregular price 0 observations, consider them as extreme outliers. And from the original data plot, data points with bsqft >=10000 could be considered as extreme outliers since they are far away from the majority data points. After removing the total 15 extreme outliers, we plot the data plot again. We can see that data points now spread more evenly.

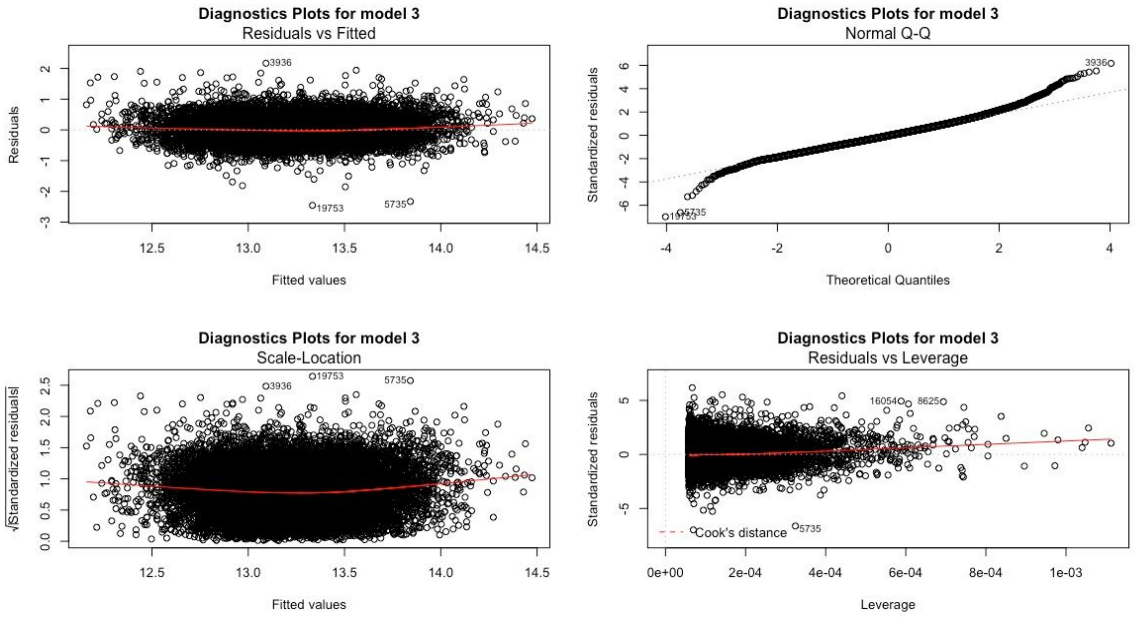
Fit a new regression model model2 and obtain the diagnostic plots, we find that now the linearity assumption holds better since data points in Residuals vs Fitted plot spread out more on the flat red line, and there seems no extreme outlier out of the Cook's distance lines. But in the Normal Q-Q plot, we now see a heavier right tail, so the normality assumption still doesn't hold. And in the Scale-Location plot, the data points spread more evenly along the red line, but the red line is still tilted, meaning that the equal variance assumption still doesn't hold.





Therefore, we need to consider data transformation. After using boxcox on model2, we obtain that lambda is about 0, meaning that we should take log transformation on the variables. After using log on both price and bsqft variables, the data plot looks linear. We fit another new regression model model3 on the transformed data. And on the left, we can see that the transformed data plot looks linear.





Draw diagnostic plots on model3, we find that data points spread evenly along the flat red line in both Residual vs Fitted plot and Scale-Location plot, meaning that the linearity and equal variance assumption hold for model3. In Normal Q-Q plot, data points approximately lie on the standard normal regression line with a little bit tilted tails, meaning that normality assumption also holds. Also, in Residuals vs Leverage plot, we don't even see a Cook's distance line, meaning that there is no extreme outliers. Therefore, model3 has all linear regression assumptions holding, so model3 is the fitted regression model for this dataset.

Q7:

Using the lines of codes, the test statistics is ts = 99.42799. Since under the null hypothesis, the ts should follow the t distribution with df = n-p, where p=2 and n=20000. The limiting nature of t for large n is approximately leading to a z distribution. If we conclude from the test statistics, we fail to reject the H0 because rejecting the H0 will lead to a way small ts than 0, but we got a large ts, so no matter what critical value is, 99>0 and we fail to reject H0. If we conclude from the p value, we fail to reject the H0 as well since the p value of this ts=99 is 1 which is larger than whatever alpha level is.

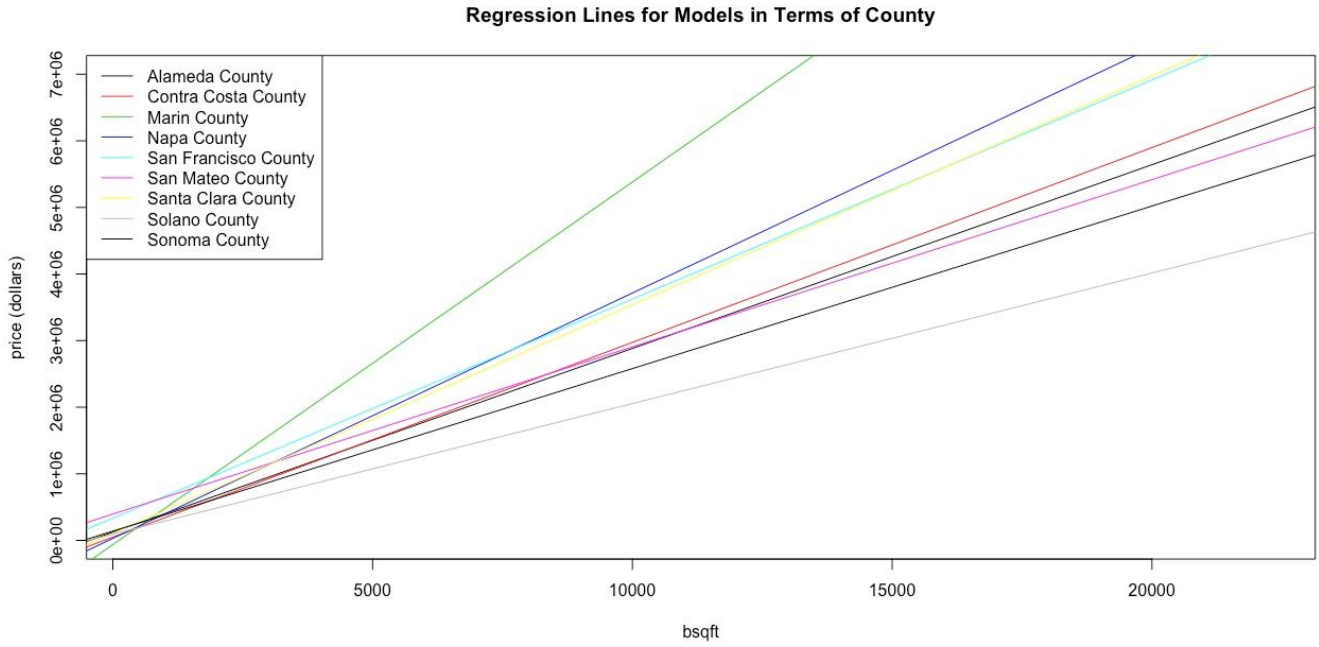
```
############## Fit linear regression model using bsqft and lsqft to predict price
model4 <- lm(price ~ bsqft + lsqft, data)
model4

############## Without transformations/diagnostics, conduct hypothesis test
        model4.sum <- summary(model4)
model4.sum
est.beta.bsqft <- model4.sum$coefficients[2,1] # estimated beta bsqft
est.beta.bsqft
est.beta.lsqft <- model4.sum$coefficients[3,1] # estimated beta lsqft
est.beta.lsqft
std.beta.bsqft <- model4.sum$coefficients[2,2] # standard deviation of beta bsqft
std.beta.bsqft
std.beta.lsqft <- model4.sum$coefficients[3,2] # standard deviation of beta lsqft
std.beta.lsqft
variance.diff <- std.beta.bsqft^2 + std.beta.lsqft^2
variance.diff
# str(model4.sum) # see details in object model4.sum

############## Report conclusion and test statistic
# find test statistics
ts <- (est.beta.bsqft - est.beta.lsqft) / sqrt(variance.diff)
ts
pnorm(ts)
```

Q8:

First we split data in terms of county, then fit regression models with each county using an lapply() function in R, then draw regression lines on an empty plot. We obtain from the plot that we can conclude the county is a confounding variable because the regression lines are not parallel in this plot, meaning that the county variable does affect the relationship between bsqft and price.

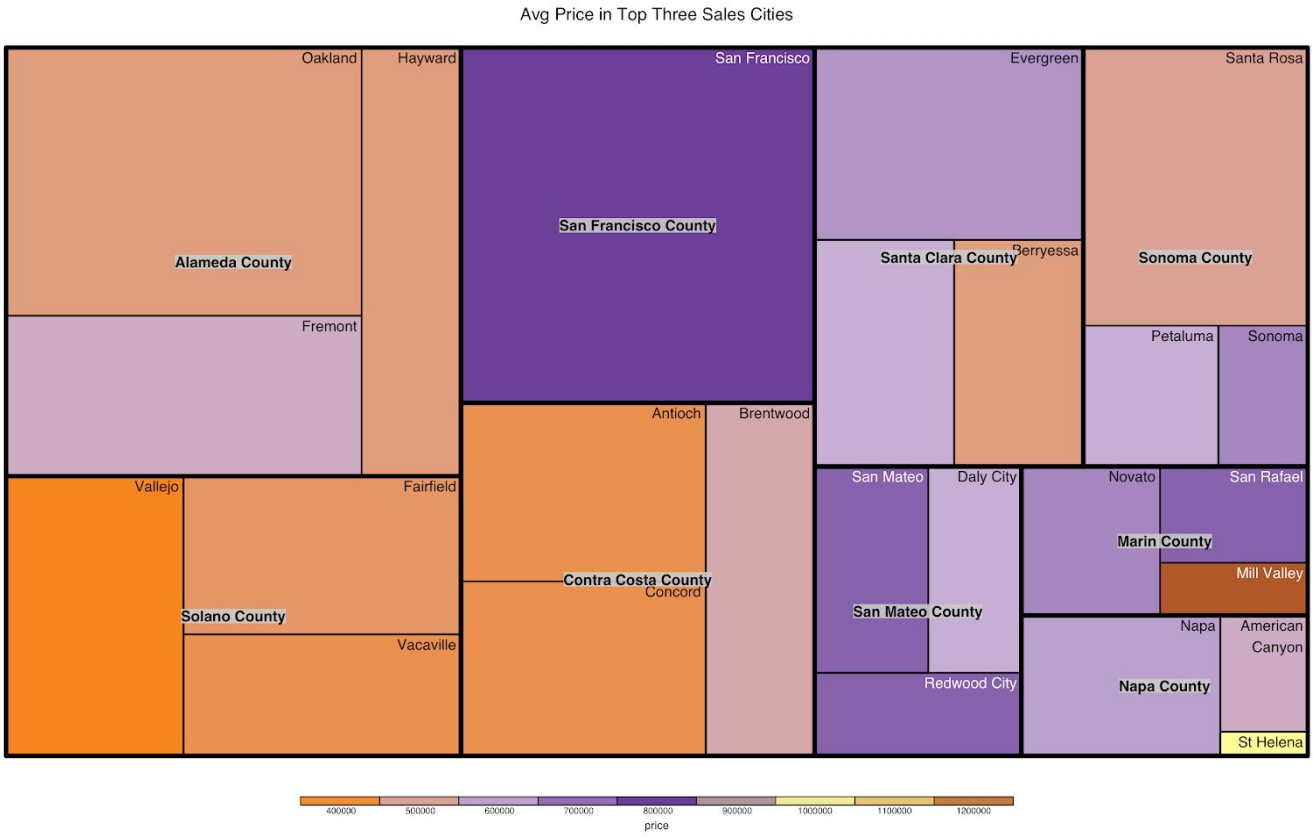


Q9:

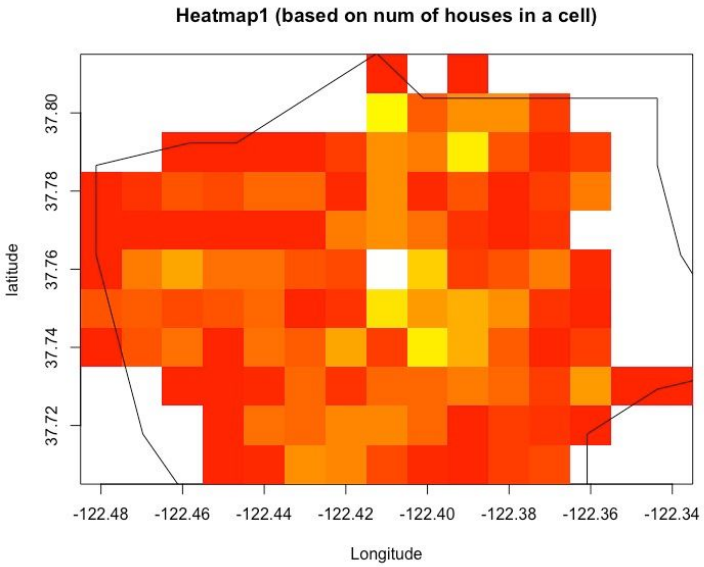
We first use aggregate() function in R to aggregate price and counts by city and county, then combine the output vectors into the same data frame called sales.agg. Then we use split() function to obtain the subset by county. Apply the lapply() function to find top 3 cities for each county. Find these cities names and prices. Remove Vallejo from this minor dataset because Vallejo has only 2 sales so it's not the top 3 in Napa County, also Vallejo is actually not within Napa County from Question 5. Then, we plot a treemap on this minor dataset without Vallejo.

Below is the treemap. From the treemap, in terms of price, we obtain that cities in county Solano, Contra Costa, and Alameda tend to sale houses in lower prices than most top three cities in other county since the color of cities in these counties are lighter on the color scale, meaning lower house sale price. On the other hand, the most expensive houses sale occur in the Mill Valley city in Marin County since its treemap box color is the darkest. St Helena city in Napa County goes to the second most expensive house sales city ranking since its color is yellow which indicates high price. San Francisco has housing price around 80000 which is higher than all top three cities in most counties except Marin County and Napa County since they both have one city with higher sale price.

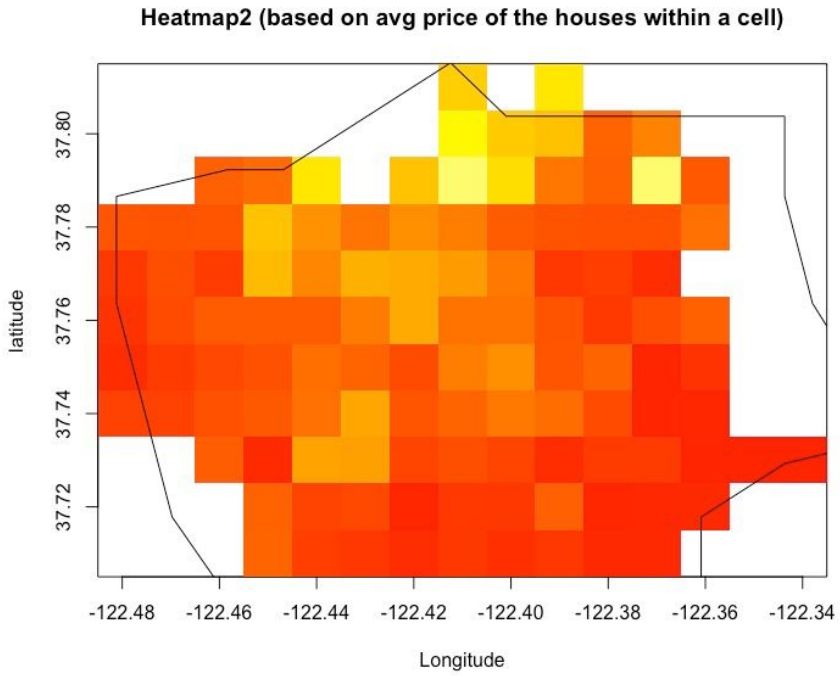
From the treemap, in terms of sale counts, we obtain from the boxes sizes that Alameda County has the most house sales counts, which is approximately the same as the total counts of Santa Clara County and Sonoma County. Solano County has house sales counts approximately the same as total of San Mateo, Napa, and Marin County. San Francisco County also has approximately the same sales counts as Contra Costa County since their boxes have the same scale, as well. If we look at cities, we can see that San Francisco has the most house sales count, while St Helena in Napa County has the least.



Q10:



The first heatmap is colorized based on the number of houses in a cell. As we can see from the heatmap, at different location in San Francisco, the number of housing sells changes. Deeper color means that the sell counts are higher, while lighter color means lower. White cell within the border of SF indicates that the sell counts of housing in the cell area is either 0 or unknown (since we change the 0's into NA while doing data cleaning). Other observations could be, on the northern west of this map where in reality golden park locates, the housing cells are red, which means the housing around that area are sold more. It makes sense that people tend to buy houses near nice environment.



The second heatmap is colorized by the average price of the houses within a cell. As we can see from the heatmap, at different location in San Francisco, the price of housing changes. Deeper color means that the housing prices are higher, while lighter color means lower. White cell within the border of SF indicates that the sell counts of housing in the cell area is either 0 or unknown (since we change the 0's into NA while doing data cleaning). Other observations could be, on the northern east of this map where in reality downtown locates, the housing prices are red, which means the housing around that area are sold with high price. It makes sense that houses near downtown are more expensive. While we see houses at around the middle SF to the north have cells in lighter color, meaning that houses in that area are cheaper. There are some affordable housing around that area, so in reality it makes sense too.