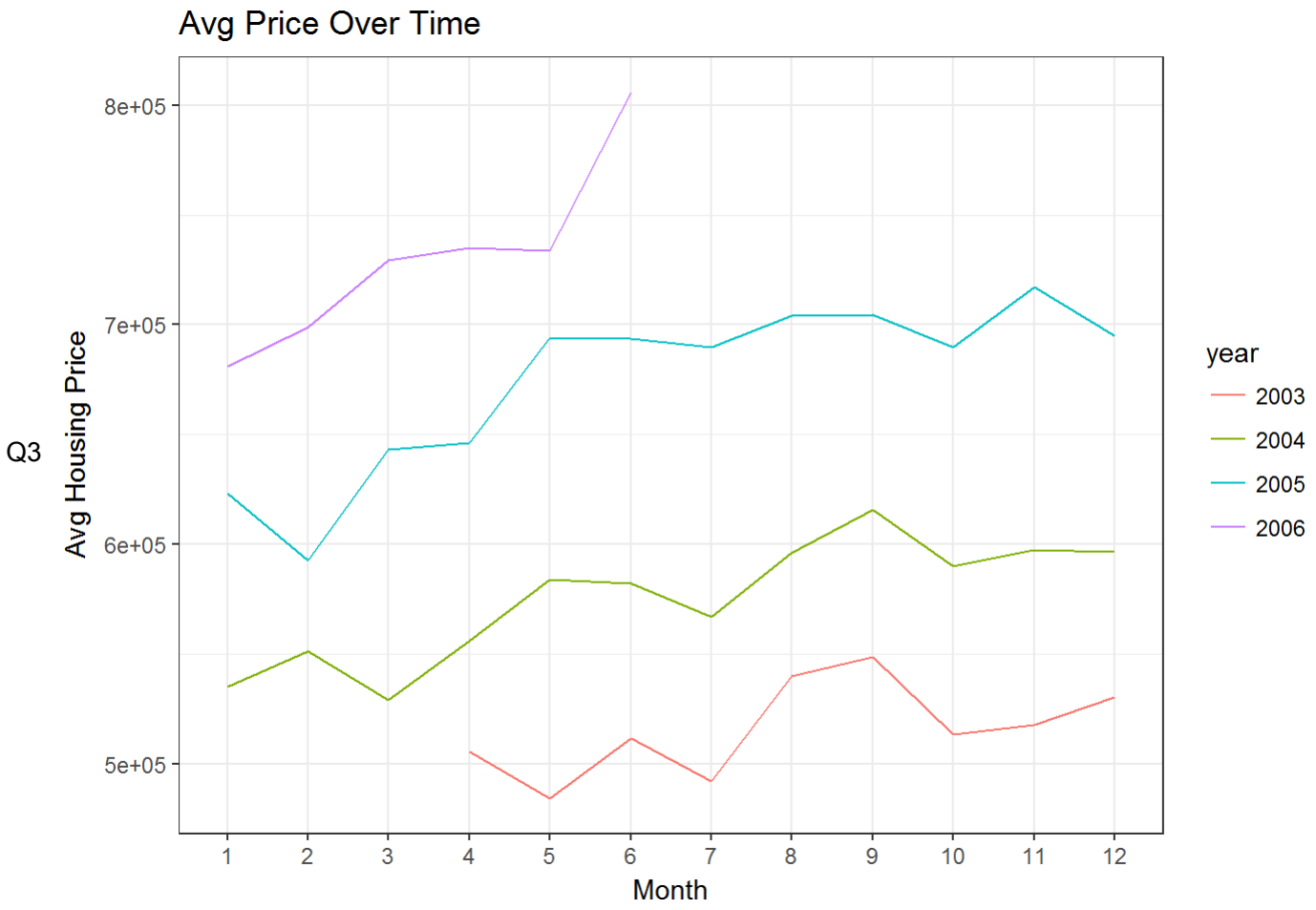
STA 141A HW2

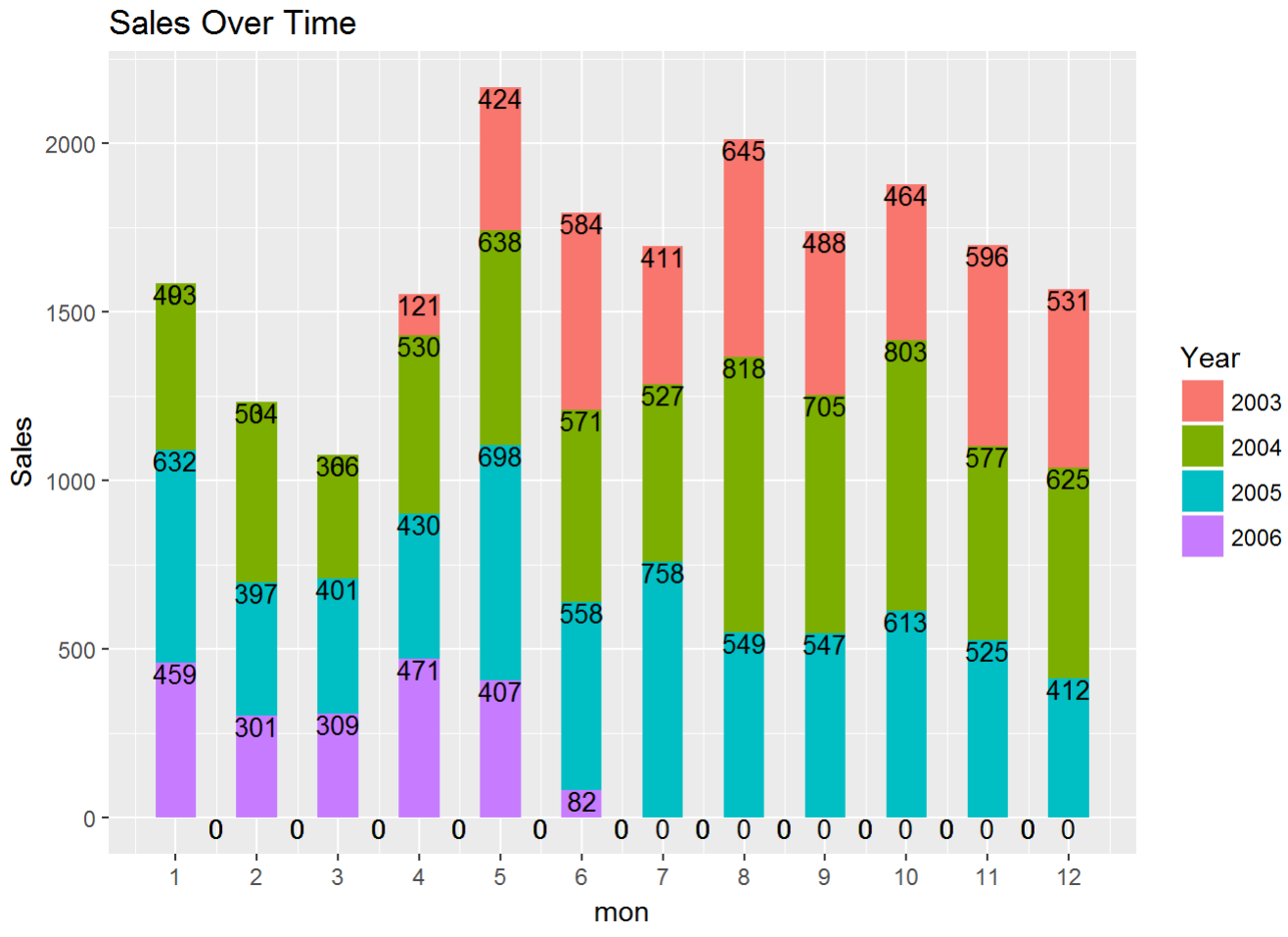
Zihan Mo 914998952

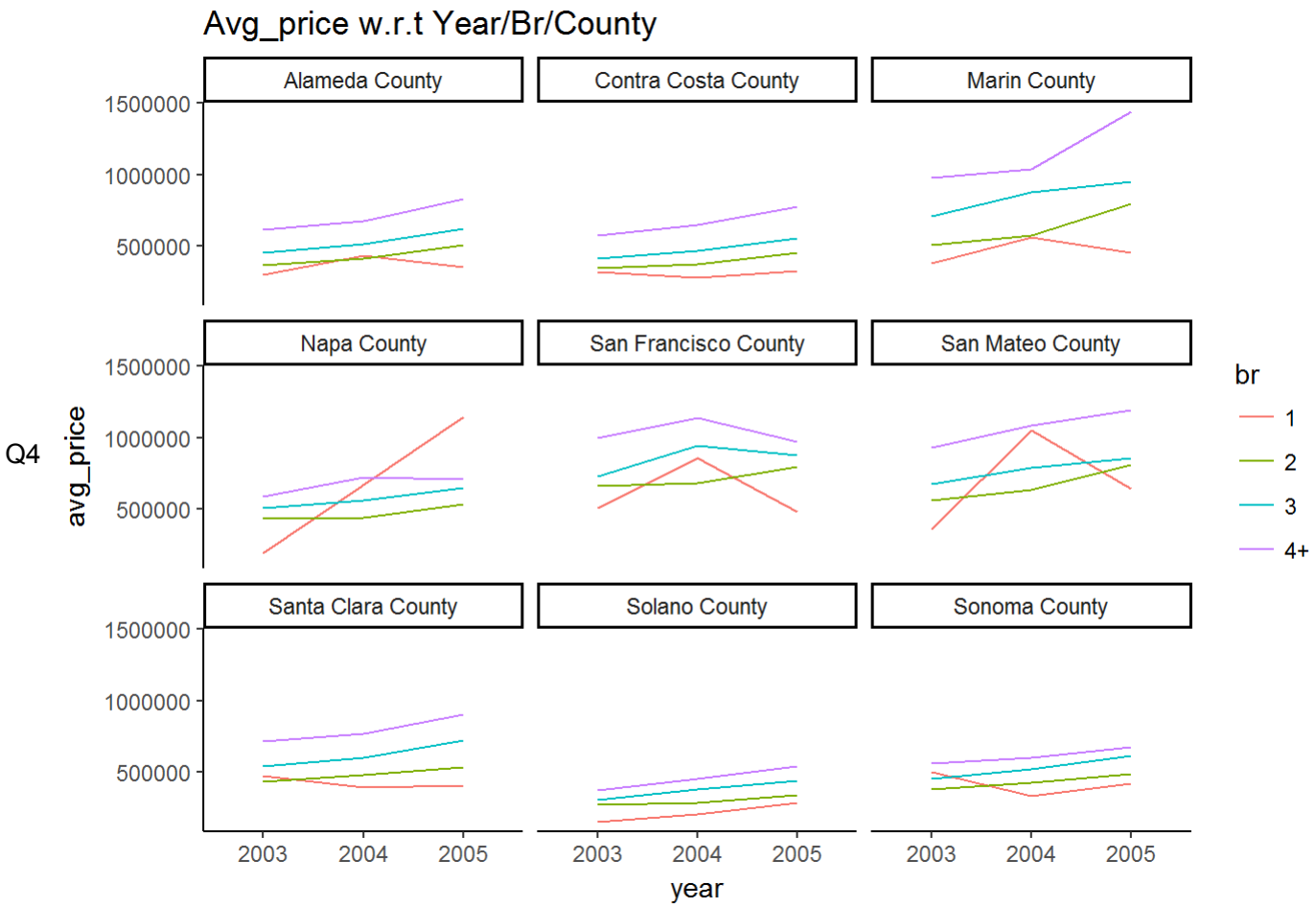
April 29, 2018

Q1 Justified by code in the appendix

Q2 The timespan of the housing sales cover from 2003-04-27 to 2006-06-04; The timespan of the construction year of houses cover from 1885 to 2005.


Avg Price Over Time
Q3
Avg Housing Price
8e+05
7e+05
6e+05
5e+05
1
2
3
4
5
6
7
8
9
10
11
12
Month
year
2003
2004
2005
2006


Sales Over Time
Sales
2000
1500
1000
500
0
493
632
459
504
397
301
366
401
309
121
530
430
471
424
638
698
407
584
571
558
82
411
527
758
645
818
549
488
705
547
464
803
613
596
577
525
531
625
412
1
2
3
4
5
6
7
8
9
10
11
12
mon
Year
2003
2004
2005
2006



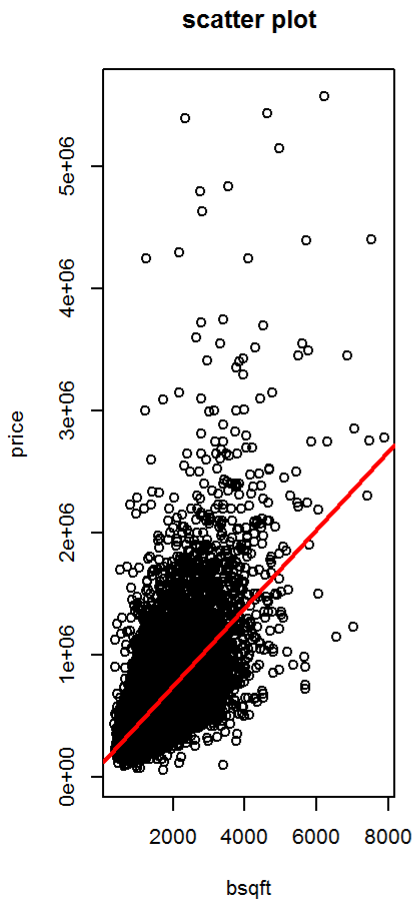
Q4

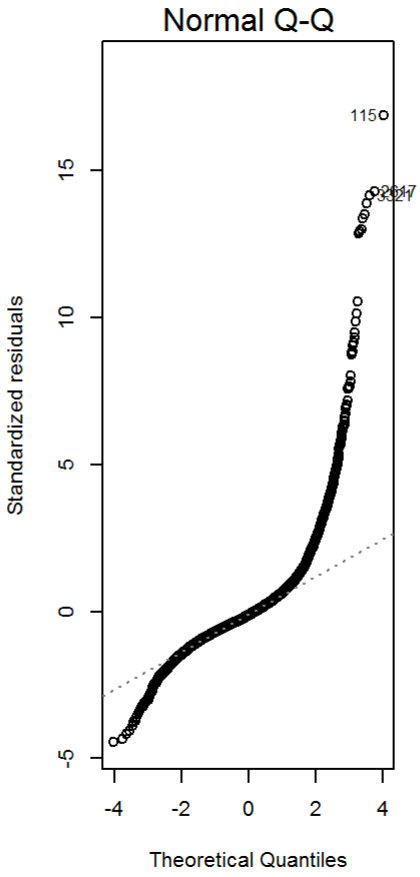
Q5

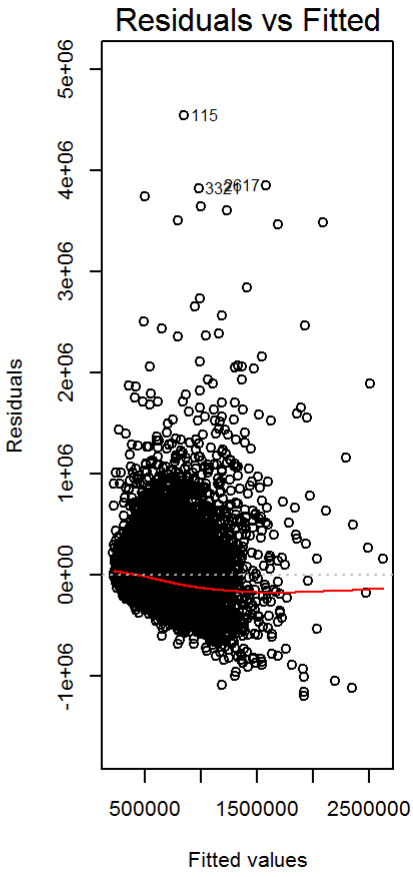
Only Vallejo city has sales in more than one county, which are Napa County and Solano County.

Q6

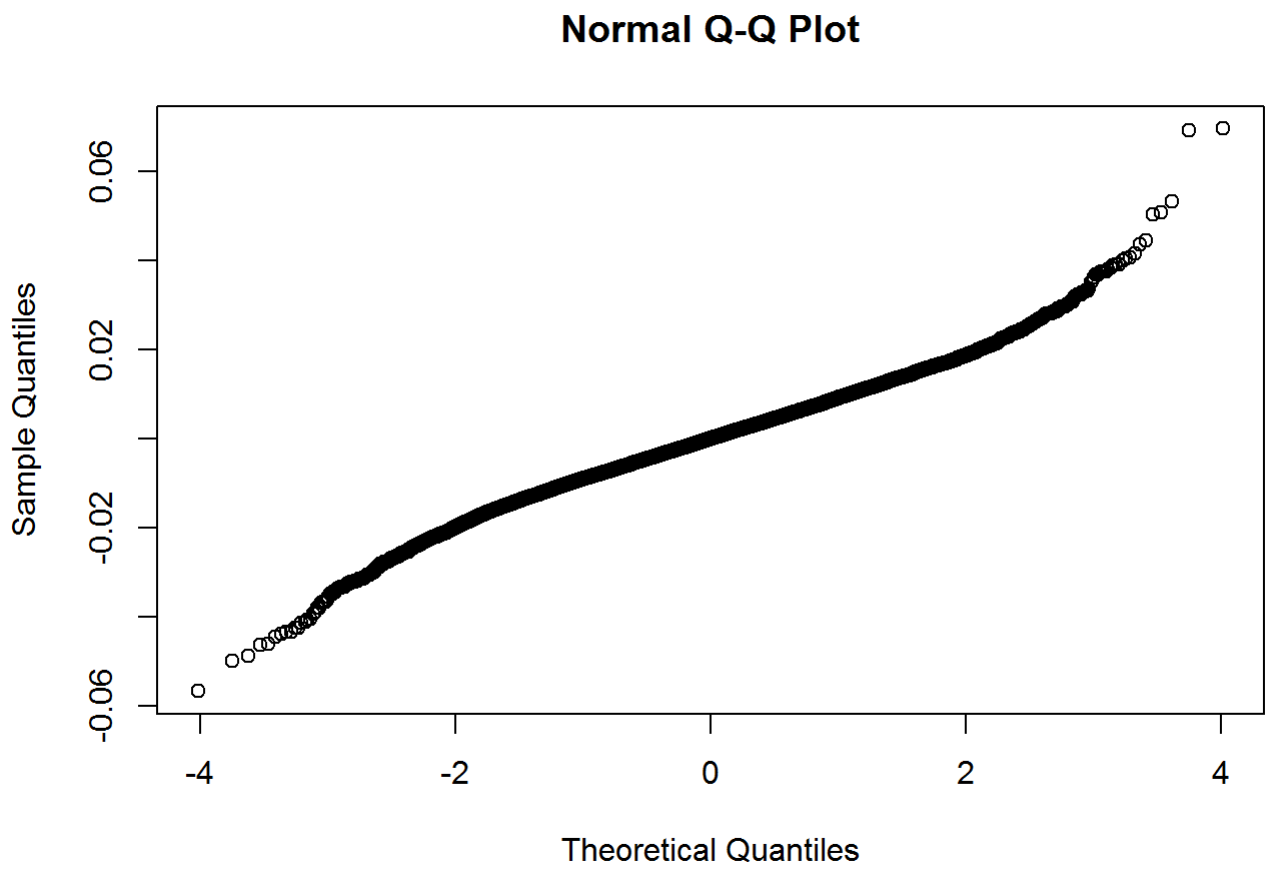
After taking out extreme outliers, like price equals zero and the building size larger than 10000, we have plots. Based on the plots below, the linear regression model seems provide a good fit to the data. But the QQ-plot and the residuals plot imply the nomality assumption isn't met and the variance of residual is not constant.


scatter plot
price
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
2000
4000
6000
8000
bsqft


Normal Q-Q
Standardized residuals
-5
0
5
10
15
115
-4
-2
0
2
4
Theoretical Quantiles


Residuals vs Fitted
Residuals
-1e+06
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
115
500000
1500000
2500000
Fitted values

Based on the QQ-plot below, after BoxCox transformation, the QQ-plot roughly follows a straight line.

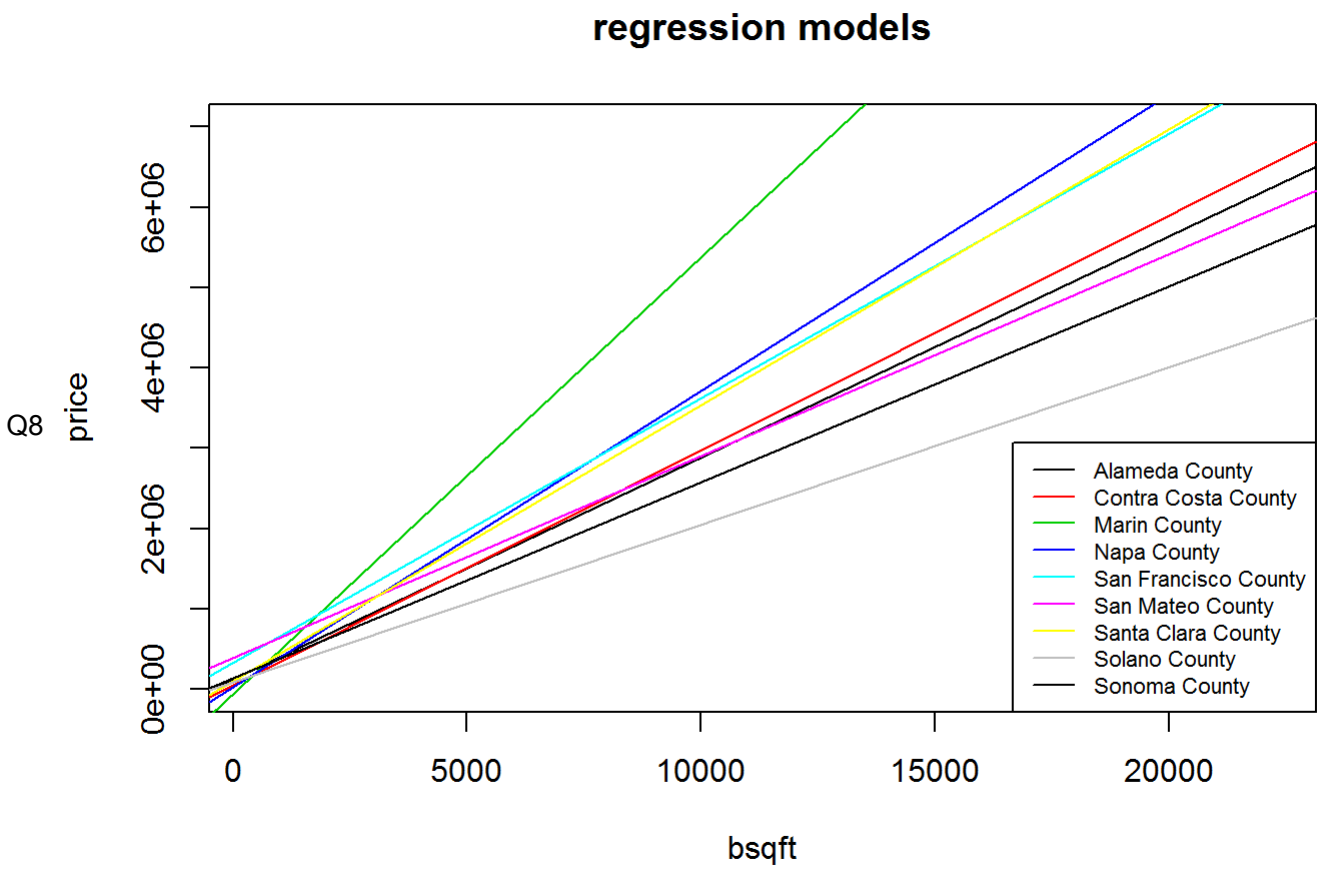


Q7

```
model3<-lm(data$price~data$bsqft+data$lsqft)
model3_est<-coef(summary(model3))[,1]
model3_se<-coef(summary(model3))[,2]
t_star<-(model3_est[2]-model3_est[3])/(model3_se[2])
critical_val<-qt(0.99,19997)
t_star>=critical_val
```

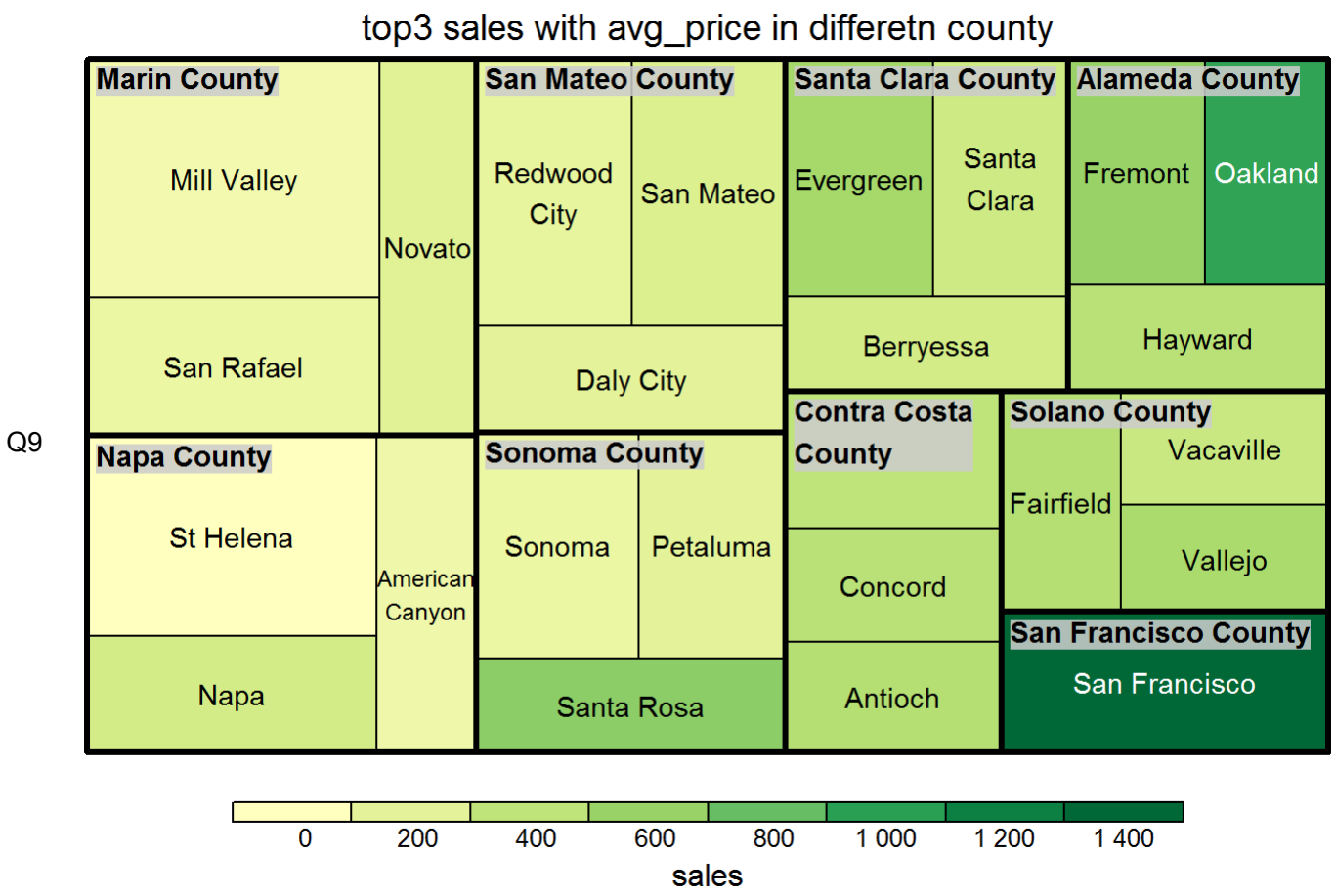
```
## data$bsqft
##       TRUE
```

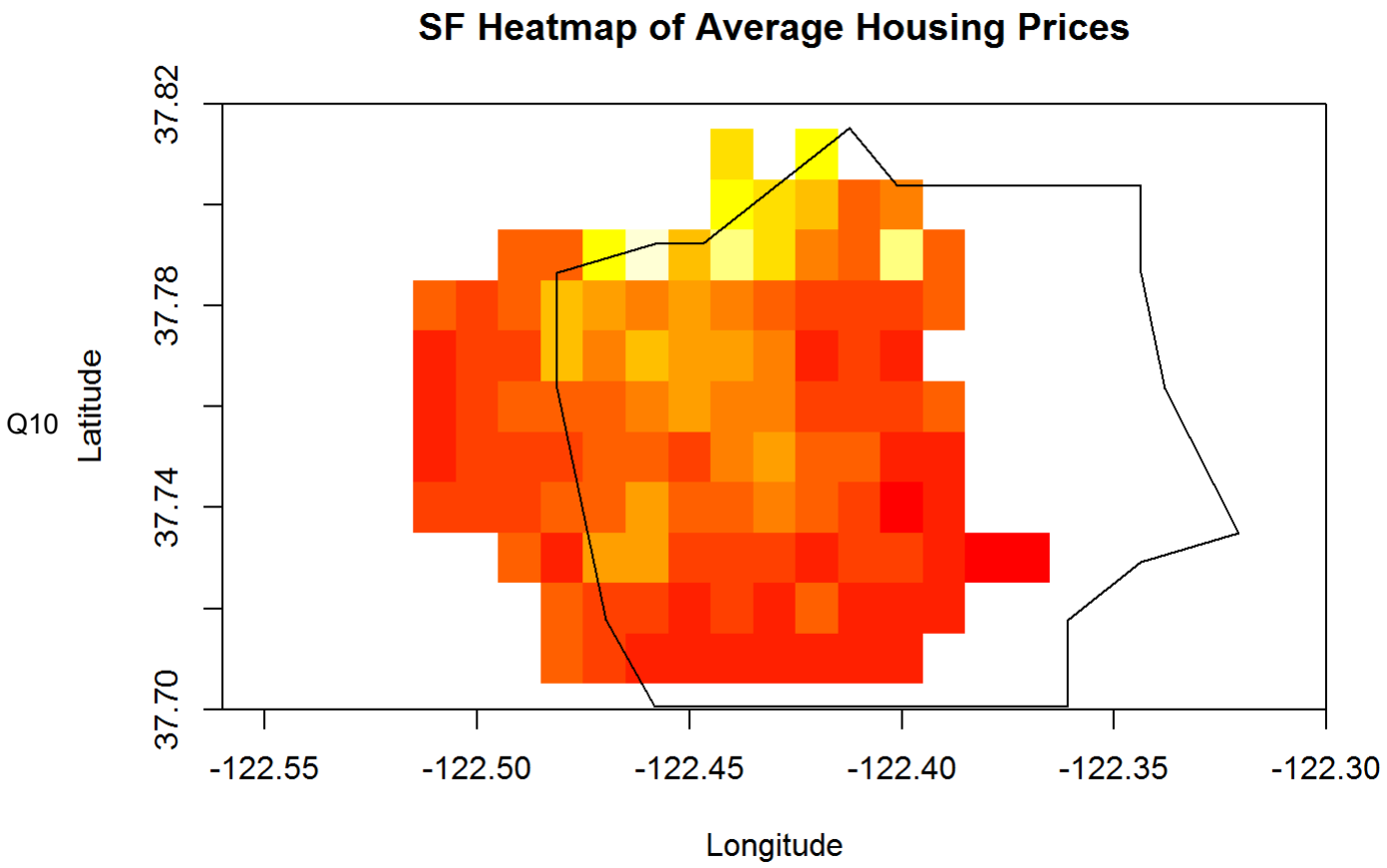
H Null:; H1:; Using t test, because t_star is greater than critical_val, conclude null hypothesis.

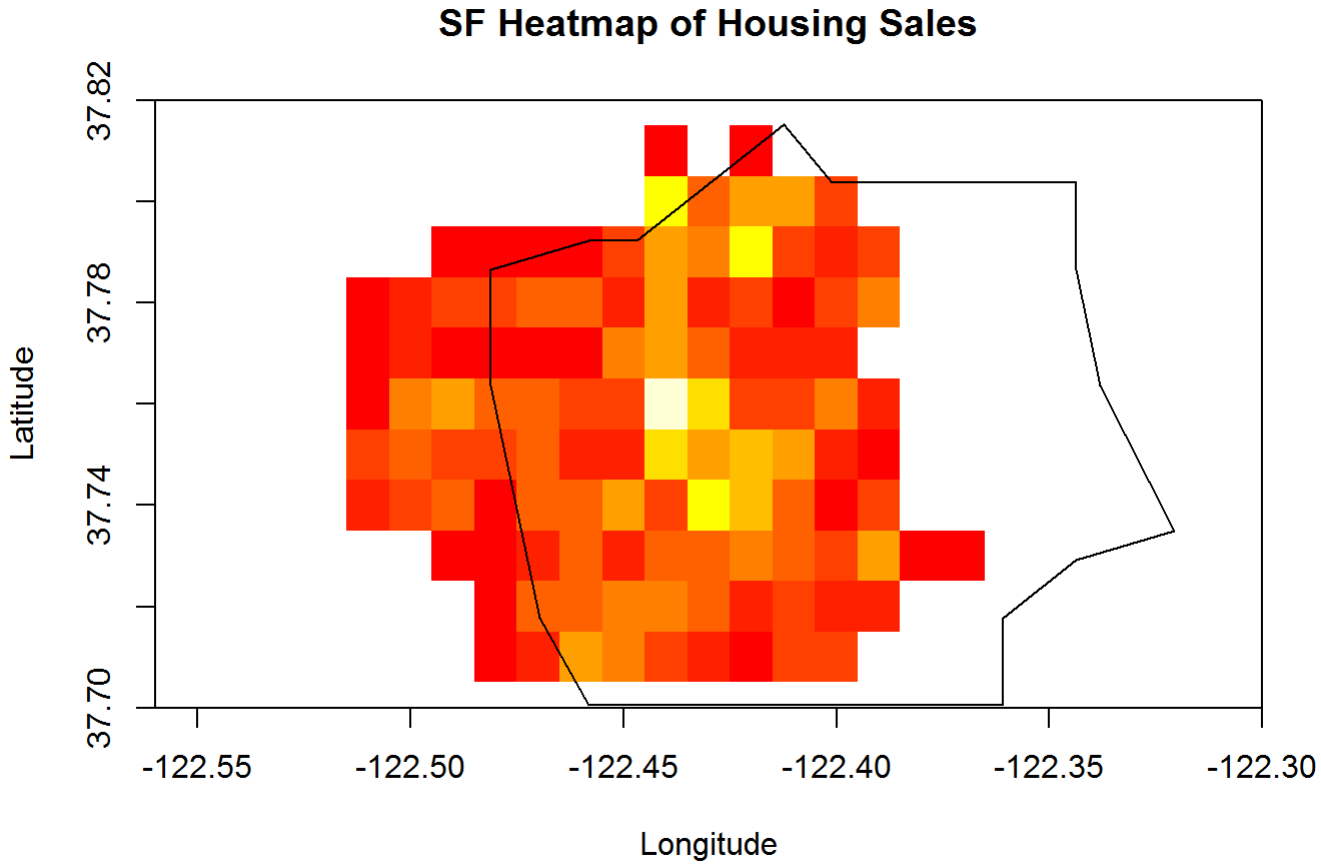


It's obviouse that the simple regression lines with respect to different county are not parallel. Therefore, we can conclude the regression lines depend on county.

Q9


top3 sales with avg_price in differetn county
Marin County
Mill Valley
San Rafael
Novato
San Mateo County
Redwood City
San Mateo
Daly City
Santa Clara County
Evergreen
Santa Clara
Berryessa
Alameda County
Fremont
Oakland
Hayward
Napa County
St Helena
Napa
American Canyon
Sonoma County
Sonoma
Petaluma
Santa Rosa
Contra Costa County
Concord
Antioch
Solano County
Fairfield
Vacaville
Vallejo
San Francisco County
San Francisco
0
200
400
600
800
1 000
1 200
1 400
sales


SF Heatmap of Average Housing Prices
Q10
Latitude
37.82
37.78
37.74
37.70
-122.55
-122.50
-122.45
-122.40
-122.35
-122.30
Longitude


SF Heatmap of Housing Sales
Latitude
37.82
37.78
37.74
37.70
-122.55
-122.50
-122.45
-122.40
-122.35
-122.30
Longitude