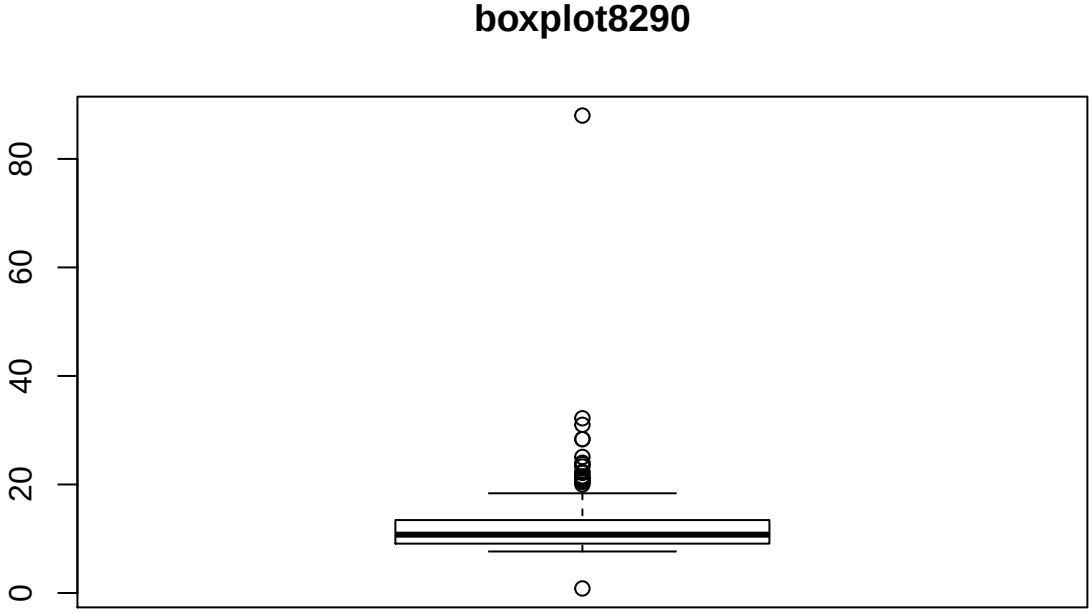
Assignment 2

QIHAO ZHONG

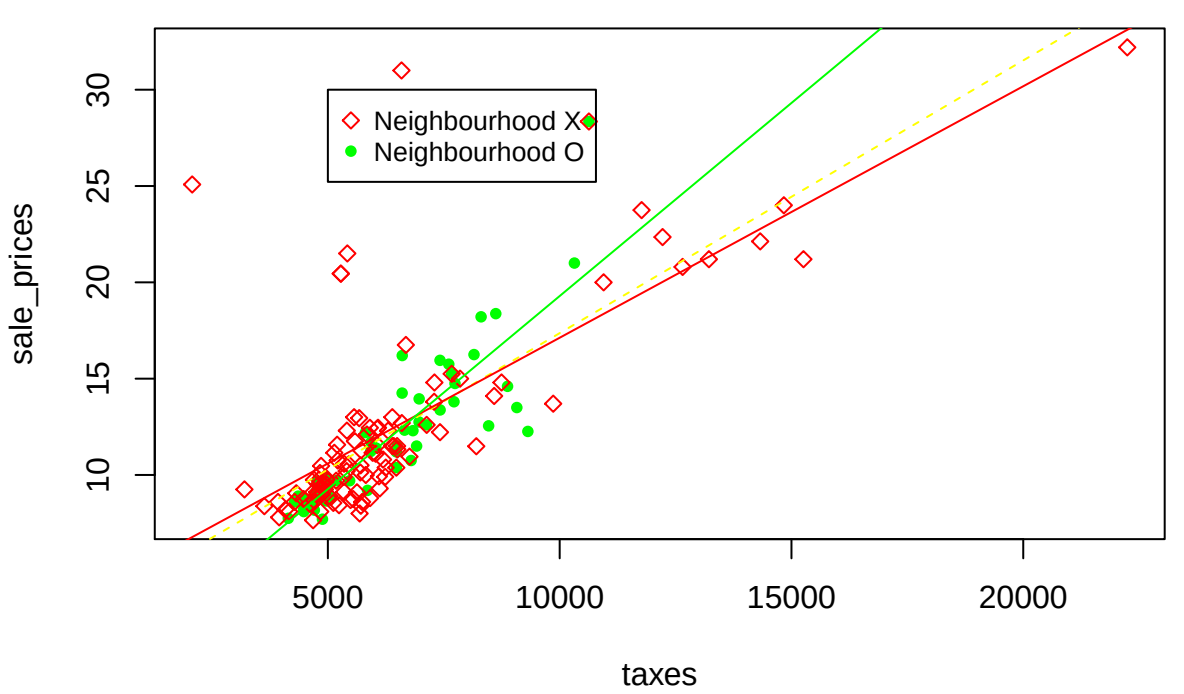
11/10/2019

1. In oreder to create a new subset by removing two cases of the most extreme sale prices, I draw a boxplot to check which two is the most extreme value.

```
## 'data.frame':    163 obs. of  5 variables:
##  $ Case_ID              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ sale.price.in..100000: num  16.8 9.3 10.4 11.3 11.5 ...
##  $ list.price.in..100000: num  16 10 10.8 12 11.7 ...
##  $ taxes                : int  6683 6119 6477 6500 6494 10631 5352 4607 4714 5706 ...
##  $ location             : Factor w/ 2 levels "O","X": 1 1 1 1 1 1 1 1 1 1 ...
```



After checking the boxplot, I decide to removing the maximum and minimum of sale prices to create a new subset. Because these two value are much larger or lower than other values.

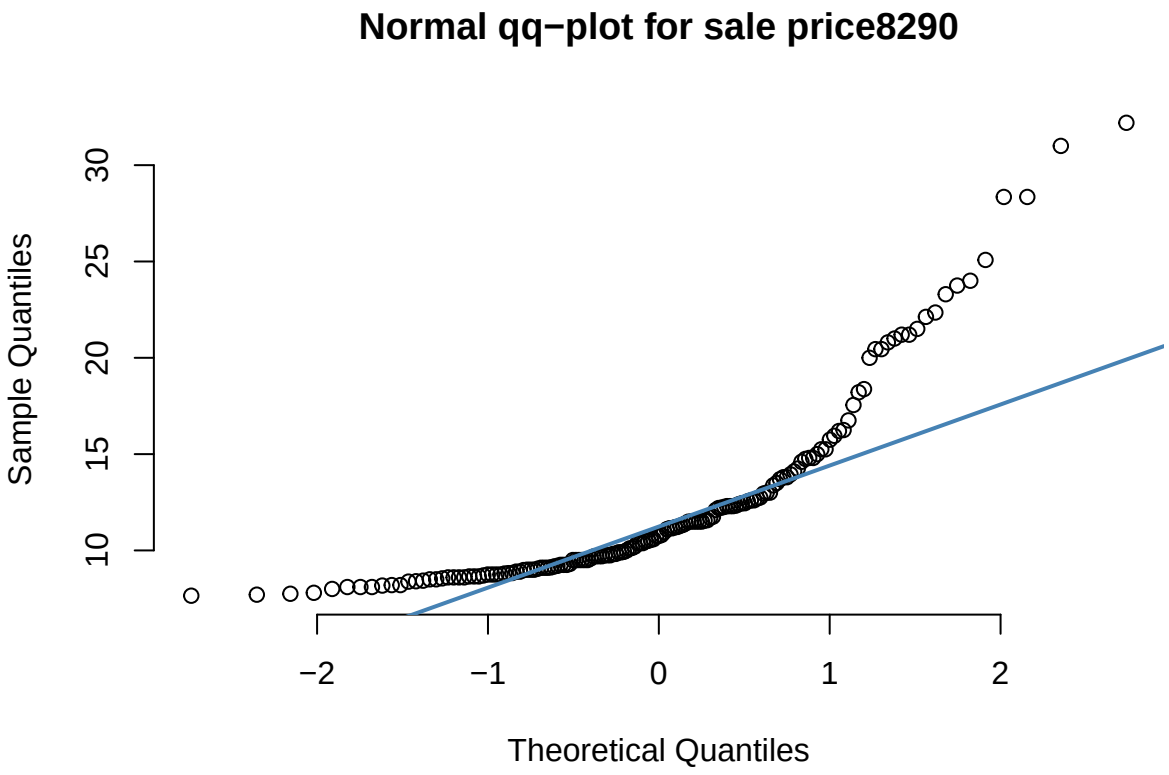


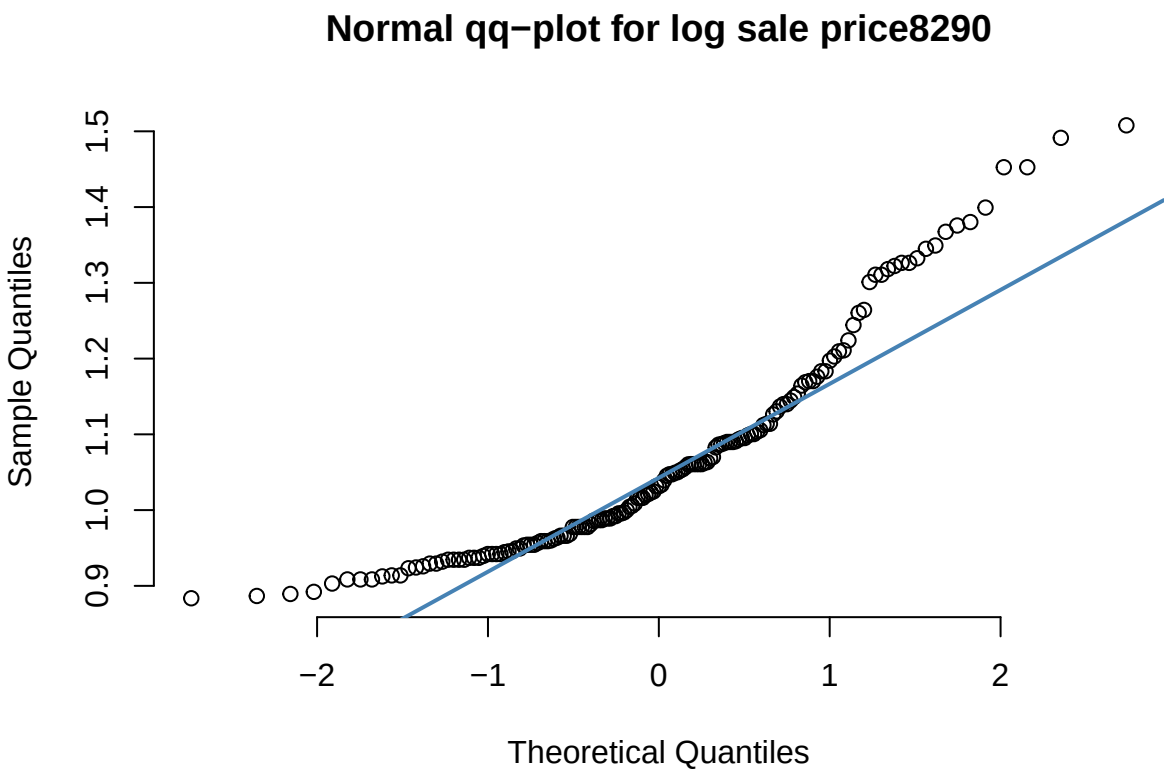
In the scatterplot of listprice and saleprice. For neighborhood X, we can see most of red points are close to the red line(fitted line for neighborhood X), therefore, we can say this linear regression for neigborhood X seems appropriate; For neighborhood O, we can see most of blue points are close to the blue line(fitted line for neighborhood O), therefore, we can say this linear regression for neigborhood O seems appropriate. For all neighborhood, we can see most of points are close to the green line(fitted line for all), we can say this linear regression for all neighborhood seems appropriate.

In the scatterplot of taxes and saleprice. we can see most of red points are close to red line(fitted line for neighborhood X) and yellow line(fitted line for all), and the green line(fitted line for neighborhood Y) are not close to the yellow line(fitted line for all), so we can say neighborhood X are more likely a linear regression than the neighborhood Y.

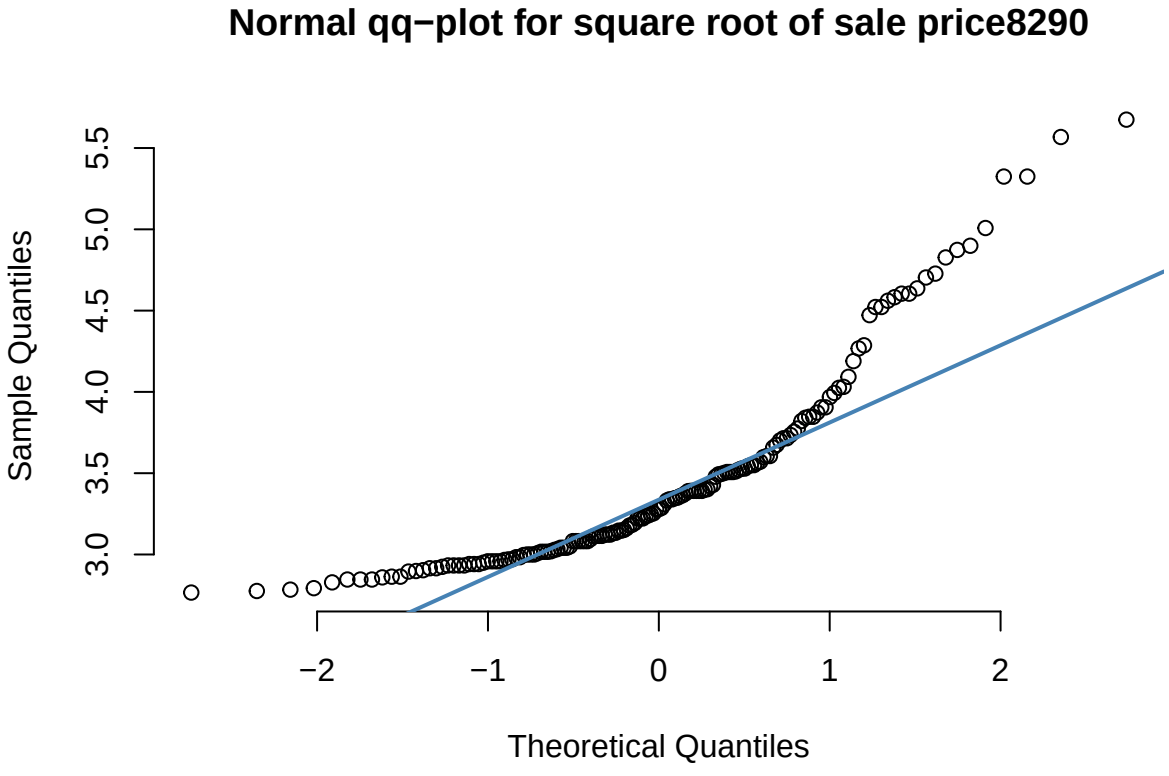
2.

(a)

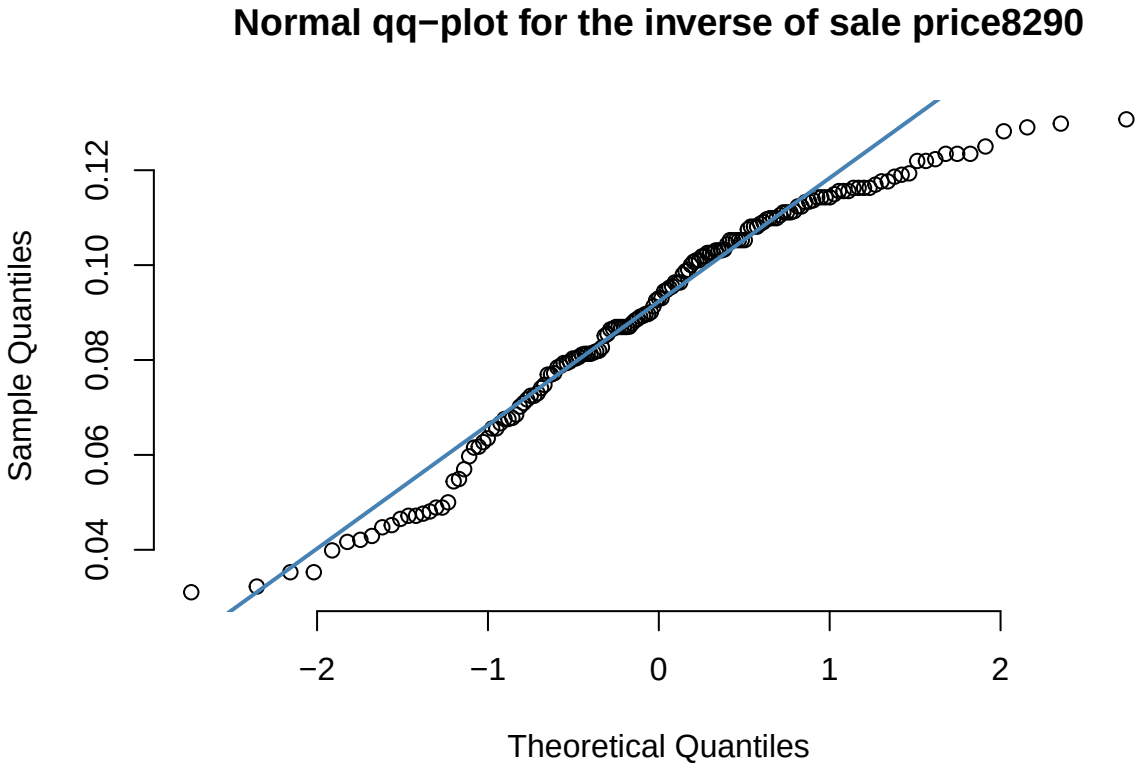



Normal qq−plot for log sale price8290
Sample Quantiles
0.9
1.0
1.1
1.2
1.3
1.4
1.5
−2
−1
0
1
2
Theoretical Quantiles

(b)


Normal qq−plot for square root of sale price8290
Sample Quantiles
3.0
3.5
4.0
4.5
5.0
5.5
−2
−1
0
1
2
Theoretical Quantiles

(c)



(d)

From the above four plots, the inverse of sale price is approximately normal. In the plot(a) plot (b) plot(c), the points form a curve instead of a straight line. so the sample data are skewed. plot (d) is more close to a striaght line. so the plot (d) is approximately normal.

---

3. (a)Simple linear regressions (SLR) for sale price from list price, for all data

```
##
## Call:
## lm(formula = salesprices ~ listprices, data = subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.68330 -0.21387 -0.02146  0.16470  1.72149
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.597466   0.095682   6.244 3.72e-09 ***
## listprices  0.919459   0.006948 132.335  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4575 on 159 degrees of freedom
## Multiple R-squared:  0.991,  Adjusted R-squared:  0.9909
## F-statistic: 1.751e+04 on 1 and 159 DF,  p-value: < 2.2e-16
```

```
##                 2.5 %    97.5 %
## (Intercept) 0.4084937 0.7864386
## listprices  0.9057369 0.9331813
```

(b)Simple linear regressions (SLR) for sale price from list price, for neighborhood X

```
##
## Call:
## lm(formula = subsetX$salesprices ~ subsetX$listprices)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12708 -0.25692 -0.01229  0.14794  1.64368
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.499979   0.120666   4.144 7.12e-05 ***
## subsetX$listprices 0.926232   0.008548 108.360  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.483 on 101 degrees of freedom
## Multiple R-squared:  0.9915, Adjusted R-squared:  0.9914
## F-statistic: 1.174e+04 on 1 and 101 DF,  p-value: < 2.2e-16
```

```
##                        2.5 %    97.5 %
## (Intercept)        0.2606113 0.7393471
## subsetX$listprices 0.9092760 0.9431889
```

(c)Simple linear regressions (SLR) for sale price from list price, for neighborhood O

```
##
## Call:
## lm(formula = subsetO$salesprices ~ subsetO$listprices)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58064 -0.19451 -0.01204  0.14104  0.87977
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.84505    0.15801   5.348  1.7e-06 ***
## subsetO$listprices  0.90083    0.01203  74.886  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4042 on 56 degrees of freedom
## Multiple R-squared:  0.9901, Adjusted R-squared:  0.9899
## F-statistic:  5608 on 1 and 56 DF,  p-value: < 2.2e-16
```

```
##                        2.5 %    97.5 %
## (Intercept)        0.5285138 1.1615905
## subsetO$listprices 0.8767317 0.9249268
```

Here is the table:

```
##          R2       B0         B1         var(e)   p-value
## All data "0.991"  "0.597466" "0.919459" "0.4575" "<2.2e-16"
## nbhd_X   "0.9915" "0.499979" "0.926232" "0.483"  "<2.2e-16"
## nbhd_O   "0.9901" "0.84505"  "0.90083"  "0.4042" "<2.2e-16"
##          95%CI
```

```
## All data "[0.9057369,0.9331813]"
## nbhd_X   "[0.9092760,0.9431889]"
## nbhd_O   "[0.8767317,0.9249268]"
```

---

4. From question 3, we know: For all data: R-squared = 0.991 For neignborhood X: R-squared=0.9915 For neignborhood O: R-squared=0.9901 They are very similar. From the scartter plot, we can see the data between two location are very close to each other. So they have similar data on list price and sale price. Then, they have the similar simple linear regression. They all have the variability of the response data around their mean.

5.

```
##        (Intercept) subsetX$listprices
##          0.4999792          0.9262325
```

```
##        (Intercept) subsetO$listprices
##          0.8450522          0.9008293
```

```
## [1] "t-value"
```

```
## [1] 1.607404
```

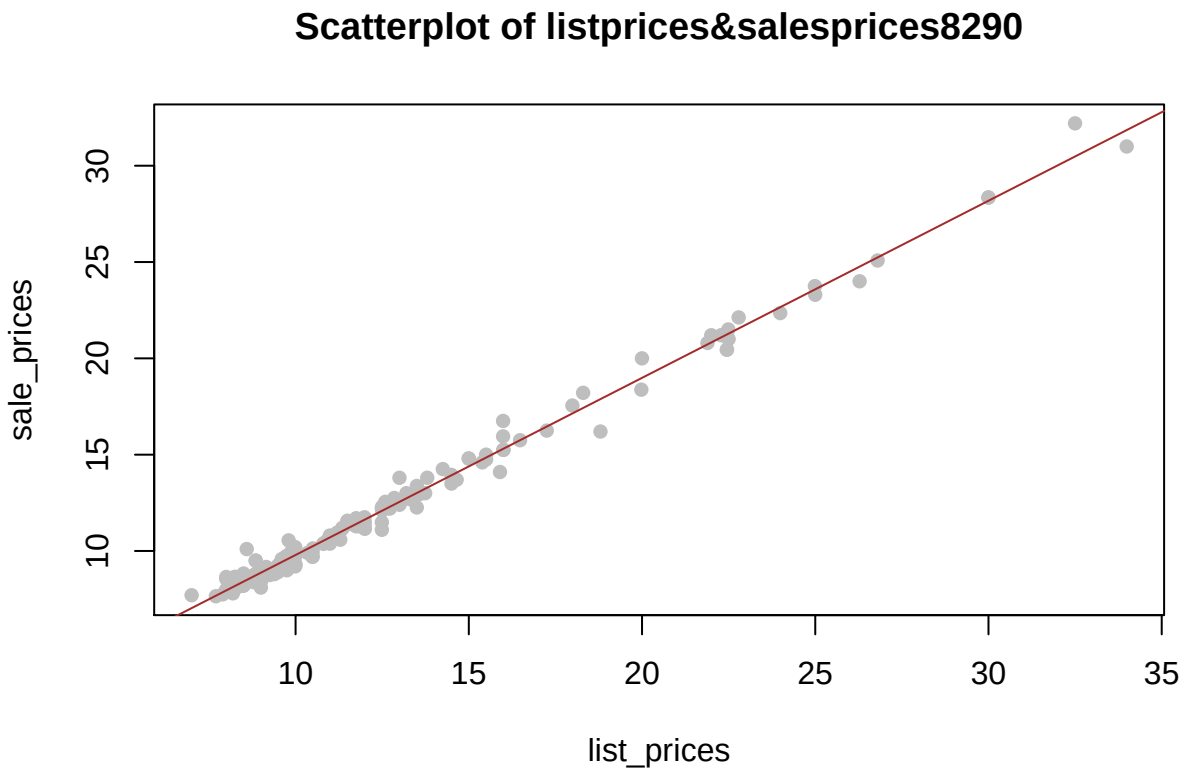
```
## [1] "df"
```

```
## [1] 157
```

```
## [1] "p-value"
```

```
## [1] 0.109975
```

Aim:Compare two slopes, H0:b1x=b1o; H1:b1x not= b1o After calculating the t-test, we have the p-value bigger than the 0.05, which means we fail to reject the H0. Which means we fail to reject that they have the same slope. So, we cannot conclude that there is any significant difference between the slopes of the 2 regressions. The t procedure can be used to compare the two slopes is because there are only two lines. If there are more than two lines, then t-test cannot be used.



6.

There are four SLR assumptions, (1) Linearity and additivity: we can check from the graph, the expected value of dependent variable is a likely straight-line function of each independent variable, holding the others fixed; the slope of the line only depends on the list prices; The effects of list prices on the expected value of sale prices are additive.(2) statistical independence of the errors: the points appears randomly.(3) homoscedasticity: the shape of points looks like a tube instead of a cone, which means the errors are not getting larger or smaller. (4) normality of the error distribution: the values are fairly close to the line, and there is no outlier. So, conclude that the model for all data have no violations of the normal error SLR assumptions.

---

7. Number of bathrooms and construction year could be used to fit a multiple linear regression for sale price. Because consumer will consider these two variable when they buy a new property.

Appendix

1

```
setwd("/Users/karon/Downloads")
data=read.csv("reale.csv",header=TRUE)
attach(data)
str(data)

boxplot(sale.price.in..100000,main="boxplot8290")
subset=subset(data,sale.price.in..100000<max(sale.price.in..100000)&sale.price.in..100000>min(sale.price
ID=subset$Case_ID
listprices=subset$list.price.in..100000
salesprices=subset$sale.price.in..100000
location=subset$location
taxes=subset$taxes
names(subset)[1:3]=c("ID","salesprices","listprices")
```

```
subsetX=subset[location=="O",]
subsetO=subset[location=="X",]

lmod=lm(salesprices~listprices, data=subset)
plot(listprices, salesprices, xlab="list_prices", ylab="sale_prices",main="Scatterplot of listprices&sal

points(subsetO$listprices,subsetO$salesprices,col="blue",pch=16,cex=.8)
points(subsetX$listprices,subsetX$salesprices,col="red",cex=.8,pch=5)

lmodX=lm(subsetX$salesprices ~ subsetX$listprices)
lmodO=lm(subsetO$salesprices ~ subsetO$listprices)

lines(subsetO$listprices, fitted(lmodO), col="blue",lwd=2)
lines(subsetX$listprices, fitted(lmodX), col="red",lwd=2)

abline(lmod, col="green", lty="dashed")
legend(x=10,y=30,c("Neighbourhood X","Neighbourhood O"),cex=.8,col=c("red","blue"),pch=c(5,16))



##FOR THE TAXES AND SALESPRICE



TsubsetX=subset[location=="O",]
TsubsetO=subset[location=="X",]

Tlmod=lm(salesprices~taxes, data=subset)
plot(taxes, salesprices, xlab="taxes", ylab="sale_prices",main="Scatterplot of taxes&salesprices8290",

points(TsubsetO$taxes,TsubsetO$salesprices,col="green",pch=16,cex=.8)
points(TsubsetX$taxes,TsubsetX$salesprices,col="red",cex=.8,pch=5)

TlmodX=lm(TsubsetX$salesprices ~ TsubsetX$taxes)
TlmodO=lm(TsubsetO$salesprices ~ TsubsetO$taxes)



abline(Tlmod, col="yellow", lty="dashed")
abline(TlmodX,col="red")
abline(TlmodO,col="green")
legend(x=5000,y=30,c("Neighbourhood X","Neighbourhood O"),cex=.8,col=c("red","green"),pch=c(5,16))
```

2

```
qqnorm(subset$salesprices, pch = 1, frame = FALSE, main="Normal qq-plot for sale price")
qqline(subset$salesprices, col = "steelblue", lwd = 2)
```

```
logsaleprices=log10(subset$salesprices)
qqnorm(logsaleprices, pch = 1, frame = FALSE, main="Normal qq-plot for log sale price")
qqline(logsaleprices, col = "steelblue", lwd = 2)
```

```
sqsaleprices=sqrt(subset$salesprices)
qqnorm(sqsaleprices, pch = 1, frame = FALSE, main="Normal qq-plot for square root of sale price")
```

```
qqline(sqsaleprices, col = "steelblue", lwd = 2)
```

```
invsaleprices=(subset$salesprices)^(-1)
qqnorm(invsaleprices, pch = 1, frame = FALSE, main="Normal qq-plot for the inverse of sale price")
qqline(invsaleprices, col = "steelblue", lwd = 2)
```

3

```
summary(lmod)
confint(lmod, level=0.95)
```

```
summary(lmodX)
confint(lmodX, level=0.95)
```

```
summary(lmod0)
confint(lmod0, level=0.95)
```

```
a= matrix(c(0.991,0.597466,0.919459,0.4575,"<2.2e-16","[0.9057369,0.9331813]",0.9915,0.499979,0.926232,(

dimnames(a) = list(c("All data", "neighbourhood_X","neighbourhood_O"), c("R2", "B0","B1","var(e)","p-val

print(a)
```

5

```
lmodX=lm(subsetX$salesprices ~ subsetX$listprices)
coefficients(lmodX)
lmod0=lm(subset0$salesprices ~ subset0$listprices)
coefficients(lmod0)
seX=0.483
se0=0.4042
slopeX=0.9262325
slope0=0.9008293
sdX=sd(subsetX$listprices)
sd0=sd(subset0$listprices)
sdXsq=sdX^2
sd0sq=sd0^2
sd0=sd(subset0$listprices)
n.1 <- length(subsetX$salesprices)
n.2 <- length(subset0$salesprices)
numerator = slopeX-slope0
Sressq=((n.1-2)*(seX^2)+(n.2-2)*(se0^2))/(n.1+n.2-4)
Sb1b2=(sqrt(Sressq))*sqrt((1/(sdXsq*(n.1-1))+(1/(sd0sq*(n.2-1)))))
t=numerator/Sb1b2
df=n.1+n.2-4
print(t)
print(df)
p.value = 2*pt(t, df, lower=FALSE)
print(p.value)
```

6

```
plot(subset$listprices, subset$salesprices, xlab="list_prices", ylab="sale_prices",main="Scatterplot of
abline(lm(subset$salesprices~subset$listprices), col="brown")
```