Dummy and Interactions

Tengyuan Liang

11/15/2020

```
# install.packages("readr")
library(readr)
MidCity <- read_csv("MidCity.csv",col_types = cols(Nbhd = col_factor(levels = c("1", "2", "3"))))
# View(MidCity)
MidCity
```

```
## # A tibble: 128 x 8
##     Home Nbhd  Offers  SqFt Brick Bedrooms Bathrooms  Price
##    <dbl> <fct>  <dbl> <dbl> <chr>    <dbl>     <dbl>  <dbl>
##  1     1 2          2  1790 No           2         2 114300
##  2     2 2          3  2030 No           4         2 114200
##  3     3 2          1  1740 No           3         2 114800
##  4     4 2          3  1980 No           3         2  94700
##  5     5 2          3  2130 No           3         3 119800
##  6     6 1          2  1780 No           3         2 114600
##  7     7 3          3  1830 Yes          3         3 151600
##  8     8 3          2  2160 No           4         2 150700
##  9     9 2          3  2110 No           4         2 119200
## 10    10 2          3  1730 No           3         3 104000
## # ... with 118 more rows
```

Dummies for Neighbourhood

```
reg1 = lm(Price~Nbhd+SqFt, data=MidCity)
summary(reg1)
```

```
##
## Call:
## lm(formula = Price ~ Nbhd + SqFt, data = MidCity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -38107 -10924   -305   9643  38506
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21241.174  13133.642   1.617  0.10835
## Nbhd2       10568.698   3301.096   3.202  0.00174 **
## Nbhd3       41535.306   3533.668  11.754  < 2e-16 ***
## SqFt           46.386      6.746   6.876 2.67e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 15260 on 124 degrees of freedom
## Multiple R-squared:  0.6851, Adjusted R-squared:  0.6774
## F-statistic: 89.91 on 3 and 124 DF,  p-value: < 2.2e-16
```

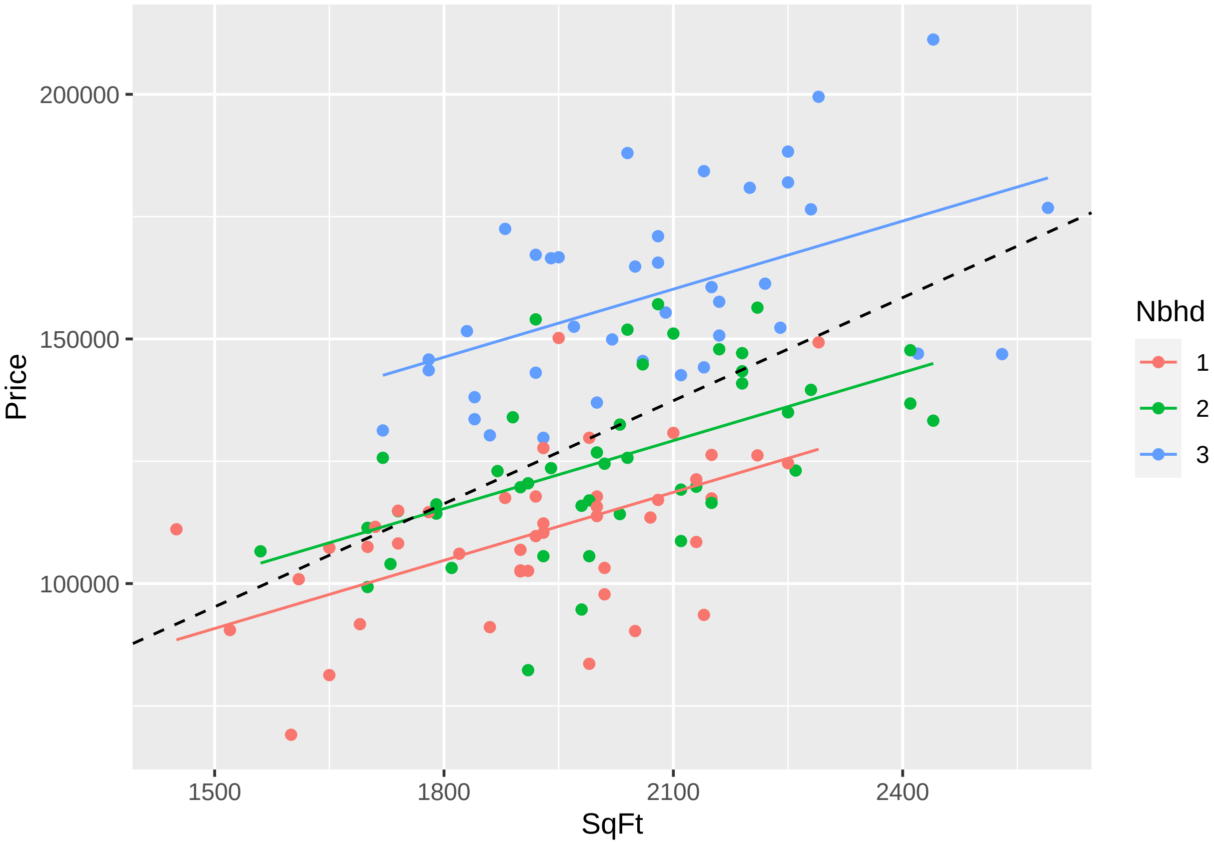
```
MidCity = cbind(MidCity, pred1 = predict(reg1))
```

```
library(ggplot2)
coeff = coefficients(lm(Price~SqFt, data=MidCity))
summary(lm(Price~SqFt, data=MidCity))
```

```
##
## Call:
## lm(formula = Price ~ SqFt, data = MidCity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46593 -16644  -1610  15124  54829
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10091.130  18966.104  -0.532    0.596
## SqFt            70.226      9.426   7.450  1.3e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22480 on 126 degrees of freedom
## Multiple R-squared:  0.3058, Adjusted R-squared:  0.3003
## F-statistic:  55.5 on 1 and 126 DF,  p-value: 1.302e-11
```

```
ggplot(MidCity, aes(x = SqFt, y = Price, color = Nbhd)) + geom_point() + geom_line(mapping = aes(y = Mi
```

```
## Warning: Use of `MidCity$pred1` is discouraged. Use `pred1` instead.
```



Dummies with Interaction

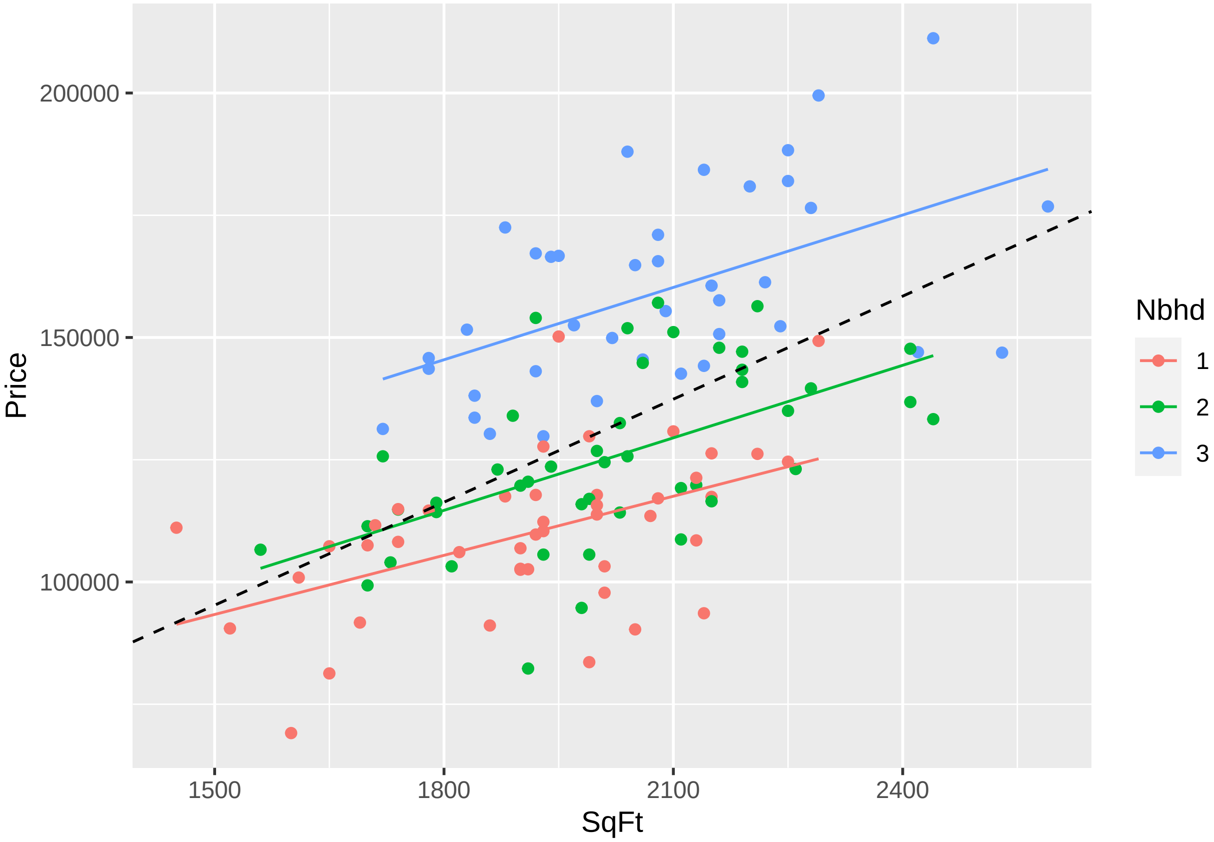
```
reg2 = lm(Price~Nbhd+SqFt+Nbhd*SqFt, data=MidCity)
summary(reg2)
```

```
##
## Call:
## lm(formula = Price ~ Nbhd + SqFt + Nbhd * SqFt, data = MidCity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37791 -10287    217   8989  38708
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32906.423  22784.778   1.444 0.151238
## Nbhd2       -7224.312  32569.556  -0.222 0.824831
## Nbhd3       23752.725  33848.749   0.702 0.484183
## SqFt           40.300     11.825   3.408 0.000887 ***
## Nbhd2:SqFt      9.128     16.495   0.553 0.580996
## Nbhd3:SqFt      9.026     16.827   0.536 0.592681
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15360 on 122 degrees of freedom
## Multiple R-squared:  0.6861, Adjusted R-squared:  0.6732
## F-statistic: 53.32 on 5 and 122 DF,  p-value: < 2.2e-16
```

```
MidCity = cbind(MidCity, pred2 = predict(reg2))
```

```
library(ggplot2)
ggplot(MidCity, aes(x = SqFt, y = Price, color = Nbhd)) + geom_point() + geom_line(mapping = aes(y = Mi
```

```
## Warning: Use of `MidCity$pred2` is discouraged. Use `pred2` instead.
```



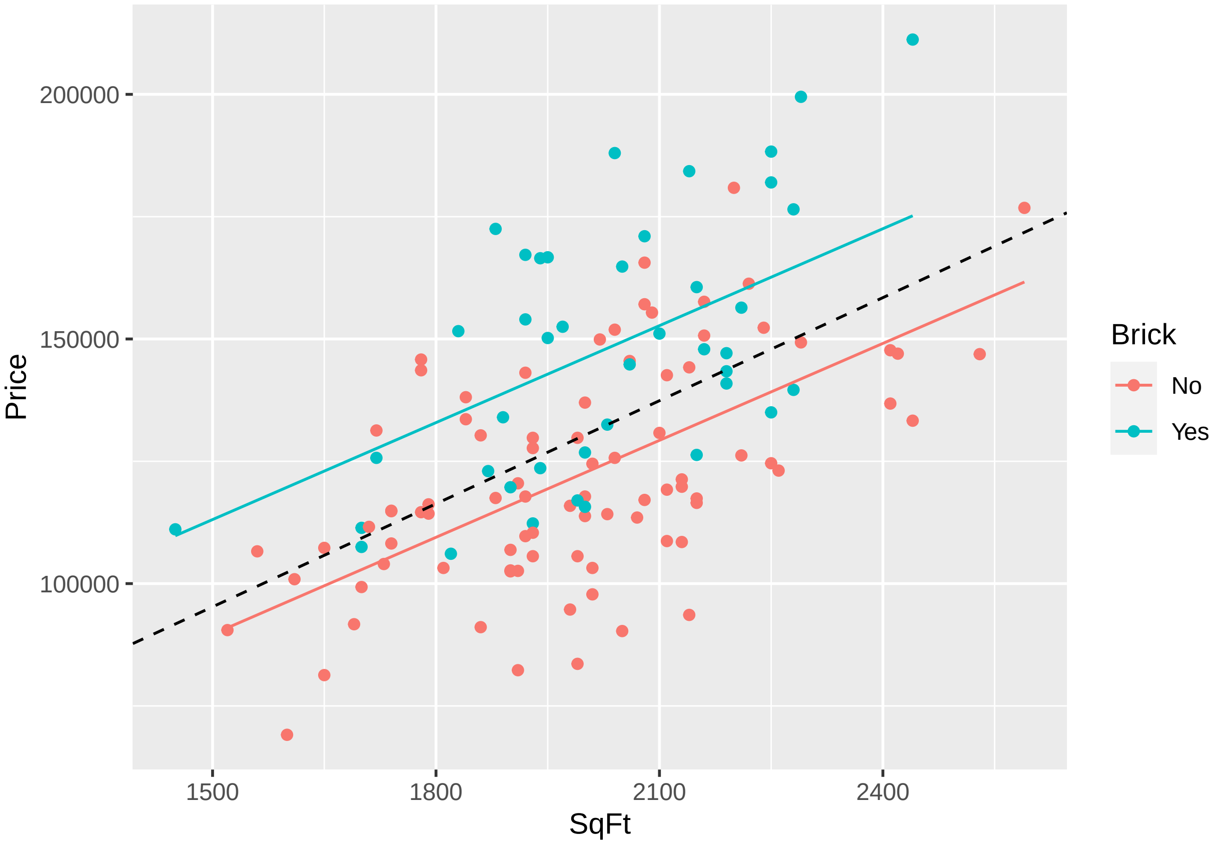
Dummies for Brick

```
reg4 = lm(Price~SqFt + Brick, data=MidCity)
summary(reg4)
```

```
##
## Call:
## lm(formula = Price ~ SqFt + Brick, data = MidCity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -38412 -14665  -1772  13912  45016
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9444.289  16577.134  -0.570     0.57
## SqFt           66.058      8.265   7.992 7.54e-13 ***
## BrickYes    23445.096   3709.805   6.320 4.21e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19640 on 125 degrees of freedom
## Multiple R-squared:  0.4739, Adjusted R-squared:  0.4655
## F-statistic:  56.3 on 2 and 125 DF,  p-value: < 2.2e-16
```

```
MidCity = cbind(MidCity, pred4 = predict(reg4))
ggplot(MidCity, aes(x = SqFt, y = Price, color = Brick)) + geom_point() + geom_line(mapping = aes(y = M
```

```
## Warning: Use of `MidCity$pred4` is discouraged. Use `pred4` instead.
```



Crazy interaction

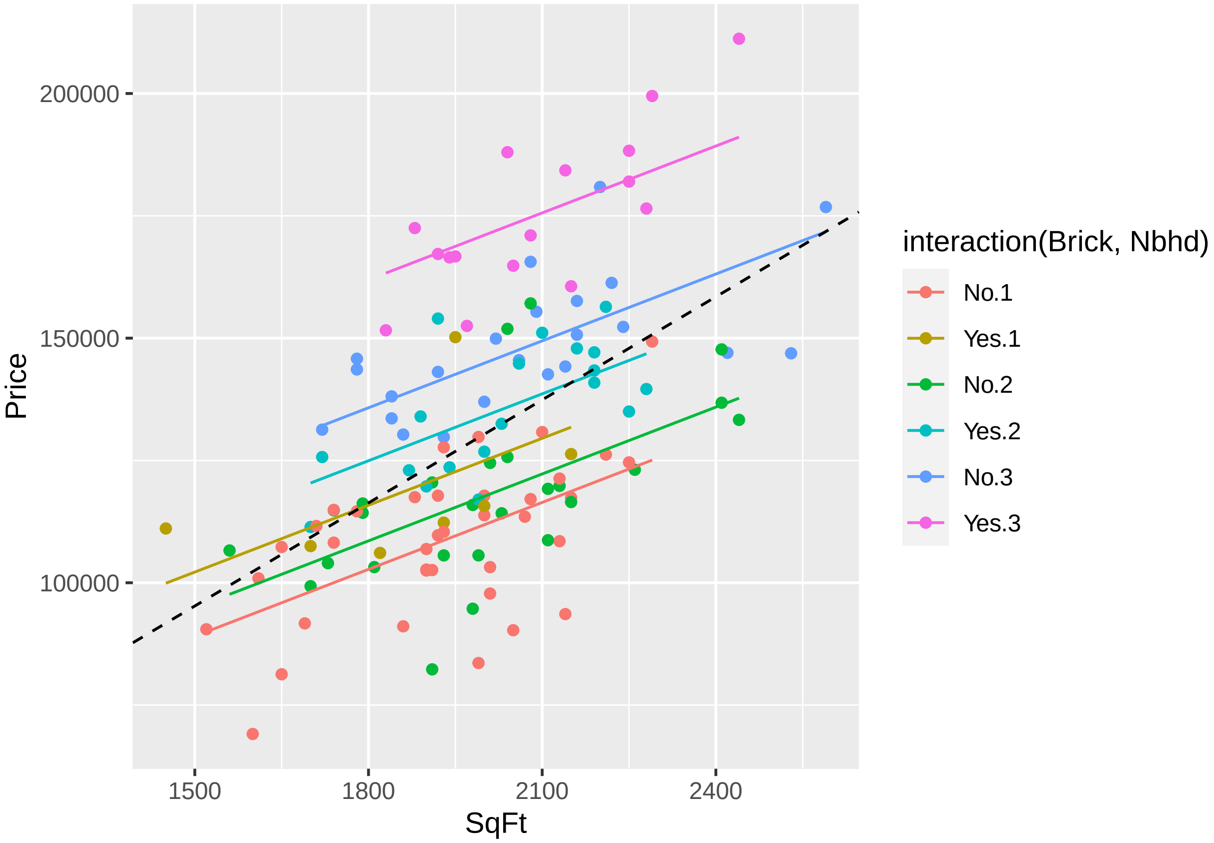
Now let's look at a crazy interaction $Brick * Nbhd$. How many categories? Answer $2 * 3 = 6$.

```
reg5 = lm(Price~SqFt+Brick*Nbhd, data=MidCity)
summary(reg5)
```

```
##
## Call:
## lm(formula = Price ~ SqFt + Brick * Nbhd, data = MidCity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31279  -7405   -847   6889  35775
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     20735.558  10766.923   1.926   0.0565 .
## SqFt               45.562      5.484   8.308 1.64e-13 ***
## BrickYes        13106.669   5106.897   2.566   0.0115 *
## Nbhd2            5820.591   3187.082   1.826   0.0703 .
## Nbhd3           33023.314   3375.878   9.782  < 2e-16 ***
## BrickYes:Nbhd2  3267.031   6335.286   0.516   0.6070
## BrickYes:Nbhd3 13053.182   6506.989   2.006   0.0471 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12350 on 121 degrees of freedom
## Multiple R-squared:  0.7986, Adjusted R-squared:  0.7886
## F-statistic: 79.95 on 6 and 121 DF,  p-value: < 2.2e-16
```

```
MidCity = cbind(MidCity, pred5 = predict(reg5))
ggplot(MidCity, aes(x = SqFt, y = Price, color = interaction(Brick, Nbhd))) + geom_point() + geom_line(
```

```
## Warning: Use of `MidCity$pred5` is discouraged. Use `pred5` instead.
```

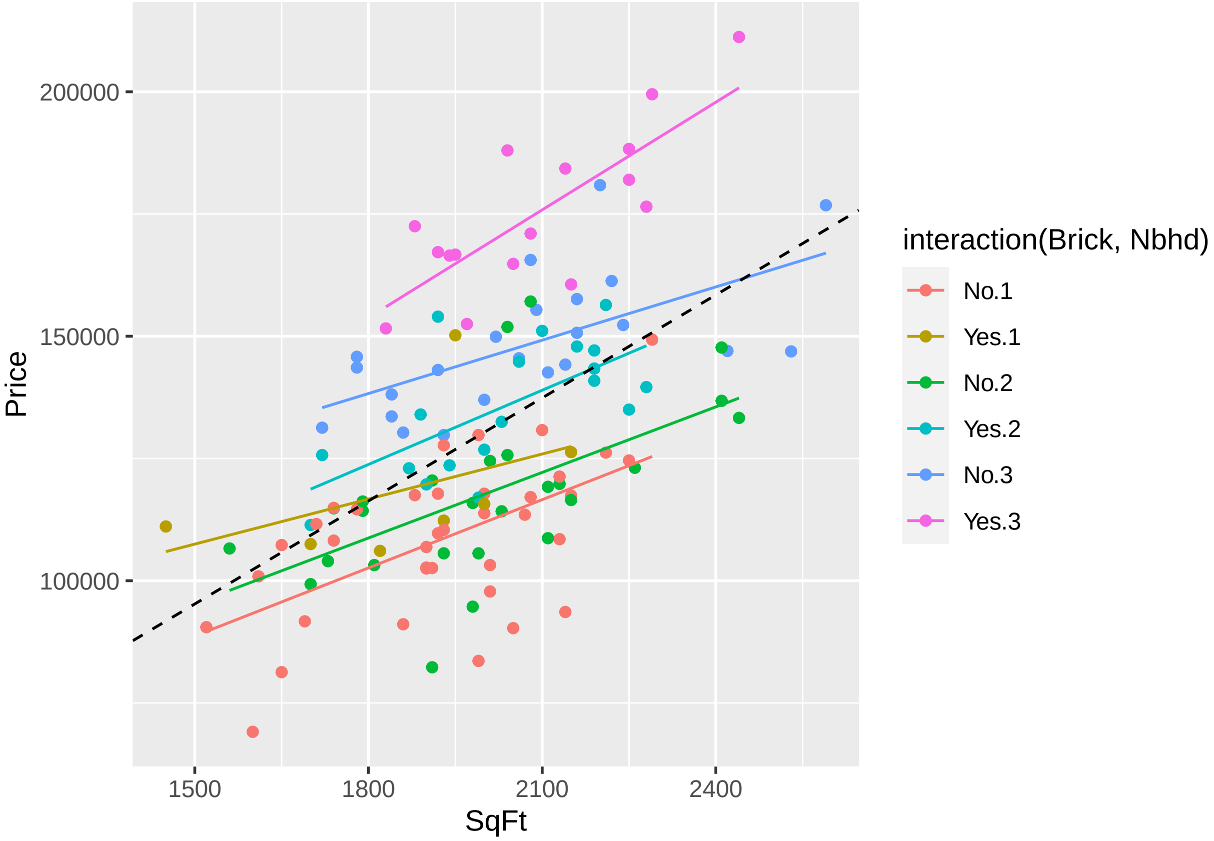


```
reg6 = lm(Price~SqFt+ Brick*Nbhd + SqFt*Brick*Nbhd, data=MidCity)
summary(reg6)
```

```
##
## Call:
## lm(formula = Price ~ SqFt + Brick * Nbhd + SqFt * Brick * Nbhd,
##     data = MidCity)
##
## Residuals:
##    Min      1Q  Median      3Q     Max
## -31359  -7173   -781   6906  35843
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           18969.783  20764.749   0.914   0.3628
## SqFt                     46.478     10.717   4.337  3.1e-05 ***
## BrickYes              42464.160  46497.577   0.913   0.3630
## Nbhd2                  9323.775  30588.472   0.305   0.7611
## Nbhd3                 53901.413  31594.024   1.706   0.0907 .
## BrickYes:Nbhd2       -38015.319  62223.215  -0.611   0.5424
## BrickYes:Nbhd3       -93694.197  64939.291  -1.443   0.1518
## SqFt:BrickYes           -15.773     24.704  -0.638   0.5244
## SqFt:Nbhd2               -1.784     15.469  -0.115   0.9084
## SqFt:Nbhd3              -10.133     15.658  -0.647   0.5188
## SqFt:BrickYes:Nbhd2      21.657     32.015   0.676   0.5001
## SqFt:BrickYes:Nbhd3      52.858     32.843   1.609   0.1102
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12430 on 116 degrees of freedom
## Multiple R-squared:  0.8045, Adjusted R-squared:  0.7859
## F-statistic: 43.39 on 11 and 116 DF,  p-value: < 2.2e-16
```

```
MidCity = cbind(MidCity, pred6 = predict(reg6))
ggplot(MidCity, aes(x = SqFt, y = Price, color = interaction(Brick, Nbhd))) + geom_point() + geom_line(
```

```
## Warning: Use of `MidCity$pred6` is discouraged. Use `pred6` instead.
```



Merging Neibhorhood 1 and 2

```r
MidCity <- read_csv("MidCity.csv", col_types = cols(Nbhd = col_factor(levels = c("1", "2", "3"))))
# View(MidCity)
# library(GGally)
# ggpairs(MidCity[,2:8], aes(colour = interaction(Brick, Nbhd), alpha = 0.4))


# Merge Nbhd 1&2
MidCity = cbind(MidCity, NbhdNew = MidCity$Nbhd)
levels(MidCity$NbhdNew) <- c("1&2", "1&2", "3")
summary(lm(Price~SqFt+NbhdNew+Brick+Bedrooms+Bathrooms, data = MidCity))
```

```
##
## Call:
## lm(formula = Price ~ SqFt + NbhdNew + Brick + Bedrooms + Bathrooms,
##     data = MidCity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34382  -7364    -53   7789  35778
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16374.106  10531.829   1.555  0.12260
## SqFt           37.111      6.427   5.774 6.03e-08 ***
## NbhdNew3    31046.000   2698.846  11.503  < 2e-16 ***
## BrickYes    19486.156   2353.868   8.278 1.84e-13 ***
## Bedrooms     2280.483   1907.399   1.196  0.23417
## Bathrooms    6972.212   2584.471   2.698  0.00797 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12260 on 122 degrees of freedom
## Multiple R-squared:  0.7999, Adjusted R-squared:  0.7917
## F-statistic: 97.53 on 5 and 122 DF,  p-value: < 2.2e-16
```

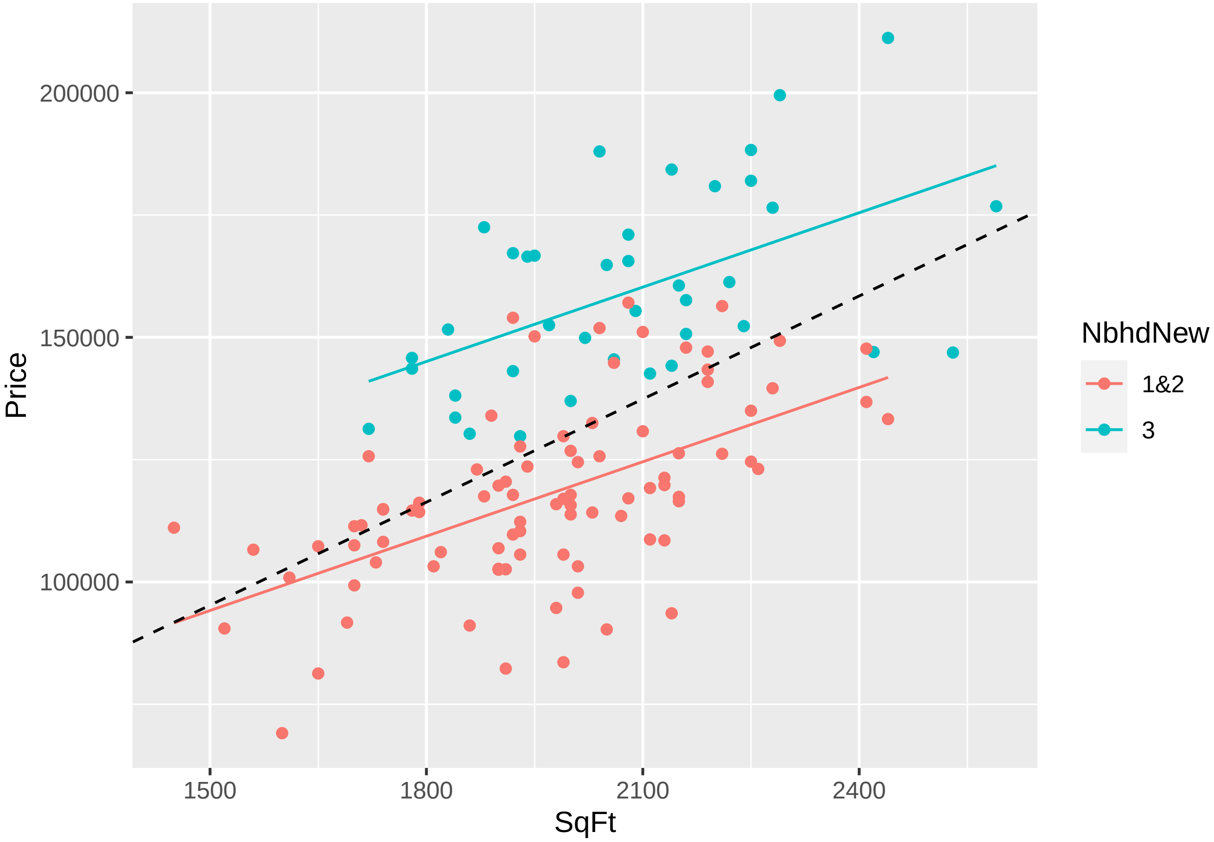
```r
coeff = coefficients(lm(Price~SqFt, data=MidCity))
reg2 = lm(Price~NbhdNew+SqFt, data=MidCity)
summary(reg2)
```

```
##
## Call:
## lm(formula = Price ~ NbhdNew + SqFt, data = MidCity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -35396  -9610  -1762   8778  38551
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18152.749  13574.154   1.337    0.184
## NbhdNew3    35699.135   3137.188  11.379  < 2e-16 ***
## SqFt           50.675      6.852   7.396 1.78e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15810 on 125 degrees of freedom
## Multiple R-squared:  0.659,  Adjusted R-squared:  0.6536
## F-statistic: 120.8 on 2 and 125 DF,  p-value: < 2.2e-16
```

```
MidCity = cbind(MidCity, pred2 = predict(reg2))

ggplot(MidCity, aes(x = SqFt, y = Price, color = NbhdNew)) + geom_point() + geom_line(mapping = aes(y =
```

```
## Warning: Use of `MidCity$pred2` is discouraged. Use `pred2` instead.
```

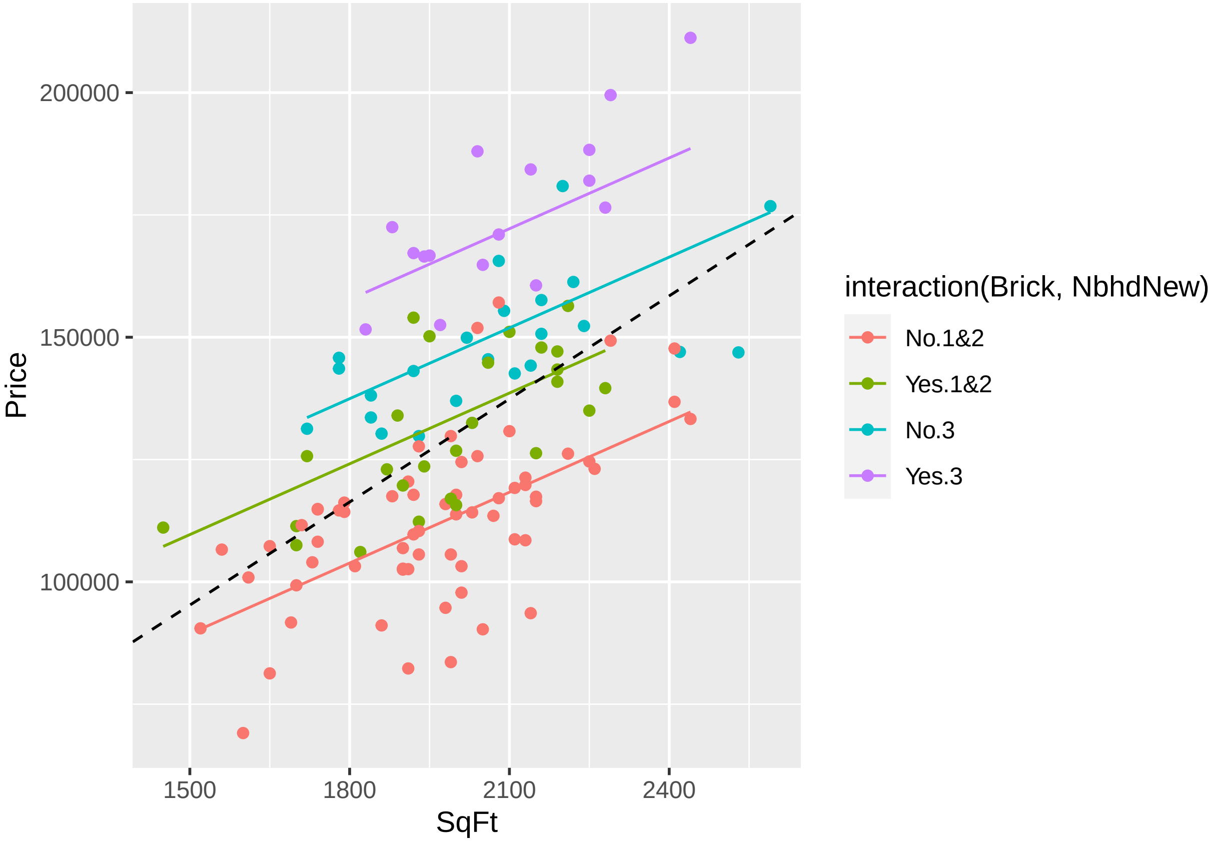


```
reg5 = lm(Price~SqFt+Brick+NbhdNew, data=MidCity)
summary(reg5)
```

```
##
## Call:
## lm(formula = Price ~ SqFt + Brick + NbhdNew, data = MidCity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -29415  -7450     47   8343  39744
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17039.80   10861.84   1.569    0.119
## SqFt           48.23       5.49   8.785 1.07e-14 ***
## BrickYes    20271.33    2401.53   8.441 6.96e-14 ***
## NbhdNew3    33585.50    2522.60  13.314  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12650 on 124 degrees of freedom
## Multiple R-squared:  0.7834, Adjusted R-squared:  0.7782
## F-statistic: 149.5 on 3 and 124 DF,  p-value: < 2.2e-16
```

```
MidCity = cbind(MidCity, pred5 = predict(reg5))
ggplot(MidCity, aes(x = SqFt, y = Price, color = interaction(Brick, NbhdNew))) + geom_point() + geom_li
```

```
## Warning: Use of `MidCity$pred5` is discouraged. Use `pred5` instead.
```

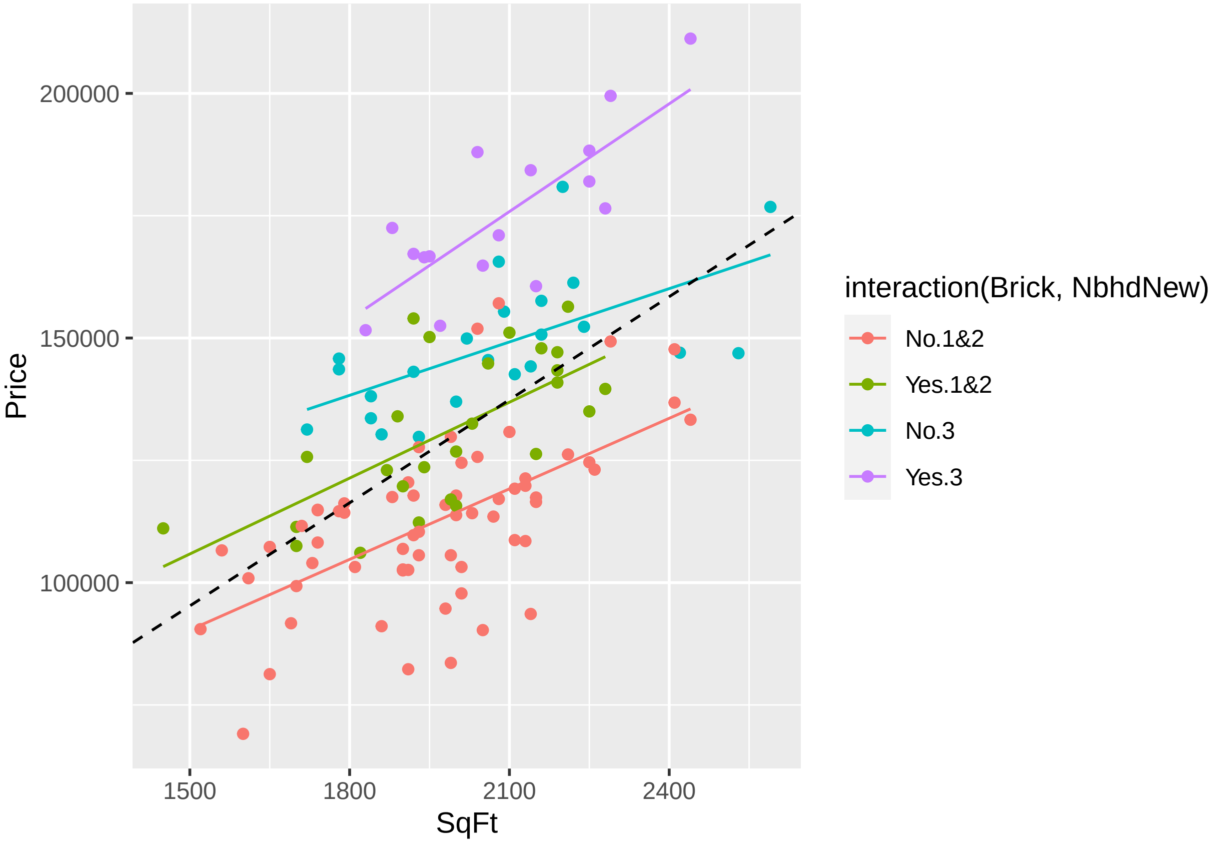


```
reg6 = lm(Price~SqFt+ Brick*NbhdNew + SqFt*Brick*NbhdNew, data=MidCity)
summary(reg6)
```

```
##
## Call:
## lm(formula = Price ~ SqFt + Brick * NbhdNew + SqFt * Brick *
##     NbhdNew, data = MidCity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30285  -6983   -715   8294  38889
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              18237.214  15123.806   1.206   0.2302
## SqFt                        48.064      7.680   6.258 6.25e-09 ***
## BrickYes                 10090.665  29245.328   0.345   0.7307
## NbhdNew3                 54633.983  28398.755   1.924   0.0567 .
## BrickYes:NbhdNew3       -61320.701  54307.890  -1.129   0.2611
## SqFt:BrickYes                3.624     14.717   0.246   0.8059
## SqFt:NbhdNew3              -11.720     13.848  -0.846   0.3991
## SqFt:BrickYes:NbhdNew3      33.461     26.341   1.270   0.2064
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12550 on 120 degrees of freedom
## Multiple R-squared:  0.7939, Adjusted R-squared:  0.7819
## F-statistic: 66.03 on 7 and 120 DF,  p-value: < 2.2e-16
```

```
MidCity = cbind(MidCity, pred6 = predict(reg6))
ggplot(MidCity, aes(x = SqFt, y = Price, color = interaction(Brick, NbhdNew))) + geom_point() + geom_li
```

```
## Warning: Use of `MidCity$pred6` is discouraged. Use `pred6` instead.
```


Price
SqFt
interaction(Brick, NbhdNew)
No.1&2
Yes.1&2
No.3
Yes.3