```r
set.seed(484)
library(dplyr)
library(stringr)
library(fastDummies)
library(MLmetrics)
library(glmnet)
library(hdm)
library(ggplot2)
library(gam)
```

```r
res <- read.csv("EXTR_ResBldg.csv")
parcel <- read.csv("EXTR_Parcel.csv")
env <- read.csv("EXTR_EnvironmentalRestriction_V.csv")
sale <- read.csv("EXTR_RPSale.csv")
```

```r
# put them together
data_cleaned <- merge(res, env, by = c("Minor" = "Minor", "Major" = "Major"))
data_cleaned <- merge(parcel, data_cleaned, by = c("Minor" = "Minor", "Major" = "Major"))
data_cleaned <- merge(sale, data_cleaned, by = c("Minor" = "Minor", "Major" = "Major"))
# make a column for year and month
data_cleaned$DocumentDate <- as.Date(data_cleaned$DocumentDate, "%m/%d/%Y")
data_cleaned <- data_cleaned %>% mutate(Month = format(data_cleaned$DocumentDate, "%m")) %>% mutate(Year
# do not want empty type
data_cleaned <- data_cleaned %>% filter(Type != "")
```

```r
data_selected <- data_cleaned %>% dplyr::select(SalePrice, DistrictName, Type, SqFtTotLiving, SqFtLot, S
# We don't want potentially poorly recorded prices(tends out this is really important)
data_selected <- data_selected %>% filter(SalePrice > 10000)
# we are not using all records
data_selected <- data_selected %>% filter(Year > 2019)


# make dummy variables
data_selected <- dummy_cols(data_selected, select_columns = c("DistrictName", "Type", "Month", "Year"))
data_selected$HeatSystem <- as.factor(data_selected$HeatSystem)
data_selected$Condition <- as.factor(data_selected$Condition)
data_selected$WaterSystem <- as.factor(data_selected$WaterSystem)
data_selected$SewerSystem <- as.factor(data_selected$SewerSystem)
data_selected$TrafficNoise <- as.factor(data_selected$TrafficNoise)
data_selected$PowerLines <- as.factor(data_selected$PowerLines)
data_selected$OtherNuisances <- as.factor(data_selected$OtherNuisances)
data_selected$HistoricSite <- as.factor(data_selected$HistoricSite)
# train test split
train <- data_selected %>% filter(Year < 2022)
test <- data_selected[data_selected$Year == 2022,]
# take out the above non-dummy columns
col_dont_want <- c("DistrictName", "Type", "Month", "Year")
train <- train[, ! names(train) %in% col_dont_want]
test <- test[, ! names(test) %in% col_dont_want]
```

```r
m1 <- lm(SalePrice ~ ., data = train)
summary(m1)
```

```
##
## Call:
## lm(formula = SalePrice ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8361025  -305913   -59365   183596 11553343
##
## Coefficients: (5 not defined because of singularities)
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -7.864e+04  3.827e+05  -0.206 0.837181
## SqFtTotLiving                    3.473e+02  1.430e+01  24.289  < 2e-16 ***
## SqFtLot                          2.087e-01  3.659e-02   5.704 1.23e-08 ***
## SqFtTotBasement                  5.509e+00  1.962e+01   0.281 0.778877
## SqFtOpenPorch                    2.775e+02  5.016e+01   5.533 3.31e-08 ***
## SqFtEnclosedPorch                9.509e+01  1.479e+02   0.643 0.520414
## SqFtDeck                        -7.627e+01  3.737e+01  -2.041 0.041285 *
## SqFtGarageAttached               9.288e+01  3.589e+01   2.588 0.009687 **
## PcntUnusable                    -3.798e+04  1.428e+04  -2.660 0.007846 **
## BrickStone                       2.528e+02  7.680e+02   0.329 0.742037
## HeatSystem1                     -1.584e+05  1.224e+05  -1.294 0.195585
## HeatSystem2                     -1.758e+05  2.648e+05  -0.664 0.506831
## HeatSystem3                      4.910e+04  1.348e+05   0.364 0.715588
## HeatSystem4                     -1.951e+03  1.158e+05  -0.017 0.986558
## HeatSystem5                     -1.850e+05  1.151e+05  -1.608 0.107994
## HeatSystem6                      1.138e+05  1.420e+05   0.802 0.422816
## HeatSystem7                     -2.409e+05  1.193e+05  -2.020 0.043450 *
## HeatSystem8                     -2.365e+05  2.227e+05  -1.062 0.288309
## Condition2                       3.600e+05  1.809e+05   1.990 0.046684 *
## Condition3                       3.153e+05  1.535e+05   2.054 0.039995 *
## Condition4                       1.668e+05  1.541e+05   1.083 0.279018
## Condition5                       2.563e+05  1.557e+05   1.646 0.099803 .
## WaterSystem1                     1.298e+05  3.346e+05   0.388 0.698052
## WaterSystem2                     1.168e+05  3.345e+05   0.349 0.726857
## SewerSystem1                     4.075e+05  2.920e+05   1.396 0.162915
## SewerSystem2                     5.420e+05  2.931e+05   1.849 0.064475 .
## SewerSystem3                     2.055e+04  5.101e+05   0.040 0.967864
## TrafficNoise1                   -6.815e+04  3.718e+04  -1.833 0.066821 .
## TrafficNoise2                   -1.289e+05  4.270e+04  -3.020 0.002544 **
## TrafficNoise3                   -1.941e+05  1.056e+05  -1.837 0.066218 .
## PowerLinesY                     -3.618e+04  9.421e+04  -0.384 0.700983
## OtherNuisancesY                 -1.684e+04  5.392e+04  -0.312 0.754813
## HistoricSite3                    1.509e+06  7.897e+05   1.911 0.056060 .
## NbrLivingUnits                  -2.993e+05  1.051e+05  -2.847 0.004425 **
## BathFullCount                   -6.204e+04  1.792e+04  -3.462 0.000540 ***
## DistrictName_ALGONA             -5.159e+05  8.323e+05  -0.620 0.535375
## DistrictName_AUBURN             -2.128e+05  2.649e+05  -0.803 0.421817
## DistrictName_BELLEVUE            8.237e+05  2.398e+05   3.435 0.000598 ***
## ‘DistrictName_BLACK DIAMOND‘    -5.768e+05  4.607e+05  -1.252 0.210582
## DistrictName_BOTHELL             1.496e+05  2.678e+05   0.559 0.576465
## DistrictName_BURIEN              5.486e+04  2.527e+05   0.217 0.828146
## DistrictName_CARNATION          -2.005e+05  2.877e+05  -0.697 0.485832
## ‘DistrictName_CLYDE HILL‘        1.470e+06  3.793e+05   3.874 0.000108 ***
## DistrictName_COVINGTON          -2.074e+05  3.004e+05  -0.690 0.489984
## `DistrictName_DES MOINES`       -3.589e+05  2.908e+05  -1.234 0.217254    
## DistrictName_DUVALL             -1.511e+05  3.058e+05  -0.494 0.621207    
## DistrictName_ENUMCLAW           -2.048e+05  4.578e+05  -0.447 0.654649    
## `DistrictName_FEDERAL WAY`       6.942e+05  3.249e+05   2.137 0.032674 *  
## DistrictName_ISSAQUAH            4.355e+05  2.454e+05   1.775 0.076036 .  
## DistrictName_KENMORE             9.907e+04  2.531e+05   0.391 0.695474    
## DistrictName_KENT                1.403e+05  2.404e+05   0.584 0.559417    
## `DistrictName_KING COUNTY`       3.484e+04  2.312e+05   0.151 0.880244    
## DistrictName_KIRKLAND            3.707e+05  2.505e+05   1.480 0.138961    
## `DistrictName_LAKE FOREST PARK` -4.321e+02  2.503e+05  -0.002 0.998622    
## `DistrictName_MAPLE VALLEY`     -1.840e+04  3.620e+05  -0.051 0.959470    
## DistrictName_MEDINA              8.242e+06  4.260e+05  19.347  < 2e-16 ***
## `DistrictName_MERCER ISLAND`     1.670e+06  2.443e+05   6.834 9.18e-12 ***
## DistrictName_MILTON             -6.413e+05  4.232e+05  -1.515 0.129781    
## DistrictName_NEWCASTLE           7.307e+04  3.393e+05   0.215 0.829505    
## `DistrictName_NORMANDY PARK`     2.955e+05  2.769e+05   1.067 0.285797    
## `DistrictName_NORTH BEND`       -7.784e+04  2.367e+05  -0.329 0.742321    
## DistrictName_PACIFIC            -4.339e+05  2.985e+05  -1.454 0.146120    
## DistrictName_REDMOND             4.932e+05  2.534e+05   1.946 0.051719 .  
## DistrictName_RENTON             -1.263e+05  2.890e+05  -0.437 0.662013    
## DistrictName_SAMMAMISH           1.047e+06  2.397e+05   4.367 1.28e-05 ***
## DistrictName_SeaTac             -1.906e+05  3.090e+05  -0.617 0.537424    
## DistrictName_SEATTLE             3.416e+05  2.391e+05   1.429 0.153106    
## DistrictName_SHORELINE           1.427e+05  2.516e+05   0.567 0.570456    
## DistrictName_SKYKOMISH          -4.300e+05  2.945e+05  -1.460 0.144310    
## DistrictName_SNOQUALMIE         -1.756e+05  2.425e+05  -0.724 0.468869    
## DistrictName_TUKWILA            -1.140e+04  2.838e+05  -0.040 0.967941    
## DistrictName_WOODINVILLE                NA         NA      NA       NA    
## Type_CoalMineHazard             -2.998e+05  1.062e+05  -2.823 0.004777 ** 
## Type_Contamination              -1.614e+05  2.915e+05  -0.554 0.579790    
## Type_CriticalDrainage           -8.046e+04  1.810e+05  -0.444 0.656712    
## Type_ErosionHazard              -2.289e+04  4.591e+04  -0.499 0.618019    
## Type_HundredYrFloodPlain         1.021e+05  4.380e+04   2.332 0.019761 *  
## Type_LandfillBuffer             -4.197e+05  5.586e+05  -0.751 0.452451    
## Type_LandslideHazard            -3.451e+03  4.905e+04  -0.070 0.943915    
## Type_SeismicHazard               7.857e+03  4.765e+04   0.165 0.869049    
## Type_SensitiveAreaTract          8.490e+04  6.782e+04   1.252 0.210658    
## Type_SpeciesOfConcern            1.213e+06  3.990e+05   3.040 0.002378 ** 
## Type_SteepSlopeHazard           -6.508e+04  5.863e+04  -1.110 0.267033    
## Type_Stream                     -3.863e+04  3.742e+04  -1.032 0.301919    
## Type_Wetland                            NA         NA      NA       NA    
## Month_01                        -2.291e+05  6.672e+04  -3.433 0.000601 ***
## Month_02                        -1.475e+05  6.188e+04  -2.384 0.017179 *  
## Month_03                        -3.806e+04  5.783e+04  -0.658 0.510461    
## Month_04                        -7.829e+04  5.851e+04  -1.338 0.180924    
## Month_05                         1.042e+04  5.541e+04   0.188 0.850903    
## Month_06                         5.922e+04  5.291e+04   1.119 0.263137    
## Month_07                        -2.147e+03  5.398e+04  -0.040 0.968279    
## Month_08                         1.310e+05  5.291e+04   2.477 0.013286 *  
## Month_09                        -2.978e+04  5.390e+04  -0.552 0.580632    
## Month_10                         2.410e+03  5.501e+04   0.044 0.965058    
## Month_11                        -7.057e+04  5.622e+04  -1.255 0.209450    
## Month_12                                NA         NA      NA       NA    
## Year_2020                       -2.483e+05  2.196e+04 -11.304  < 2e-16 ***
## Year_2021                               NA         NA      NA       NA
## Year_2022                               NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 787400 on 5271 degrees of freedom
## Multiple R-squared:  0.4687, Adjusted R-squared:  0.4592
## F-statistic: 49.46 on 94 and 5271 DF,  p-value: < 2.2e-16
```

```r
m1_pred <- predict(m1, test[,-1])
```

```
## Warning in predict.lm(m1, test[, -1]): prediction from a rank-deficient fit may
## be misleading
```

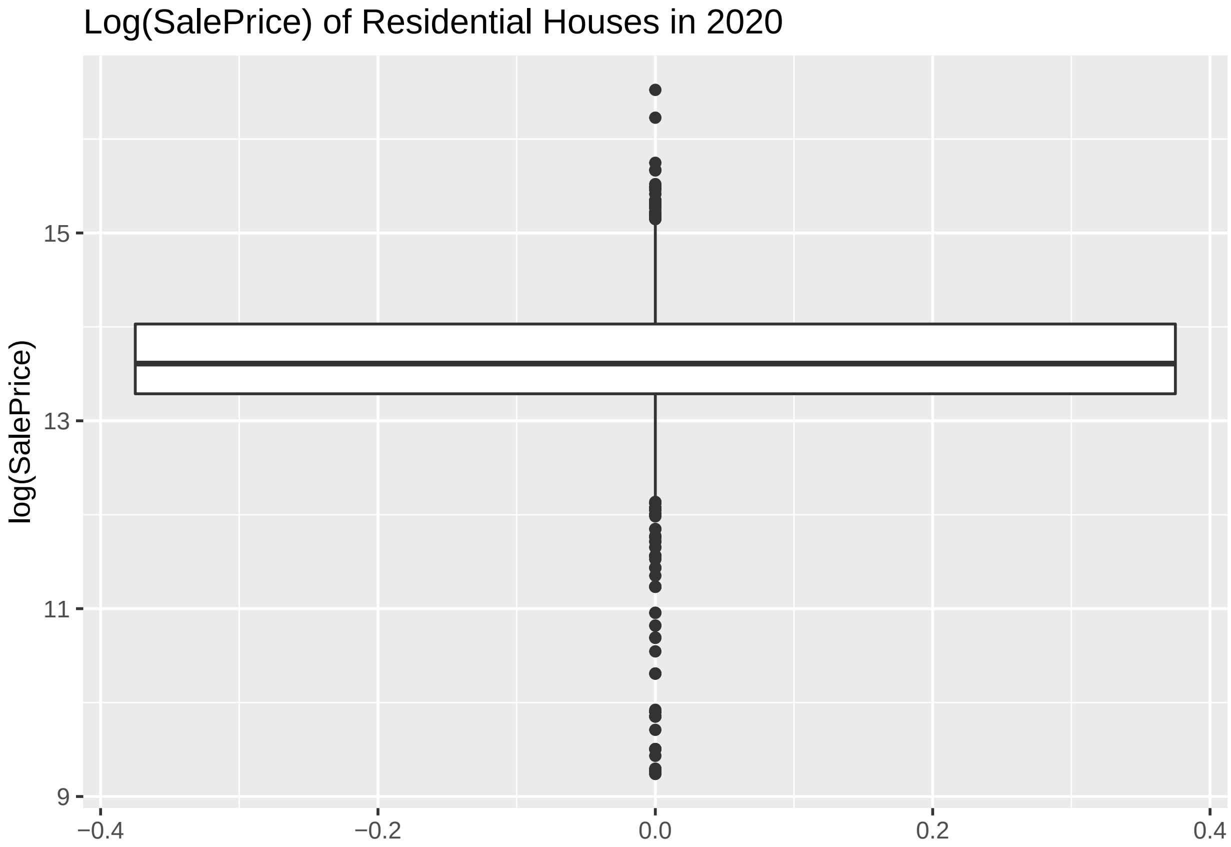
```r
MSEm1 <- MSE(m1_pred, test$SalePrice)
MSEm1
```

```
## [1] 947619397568
```

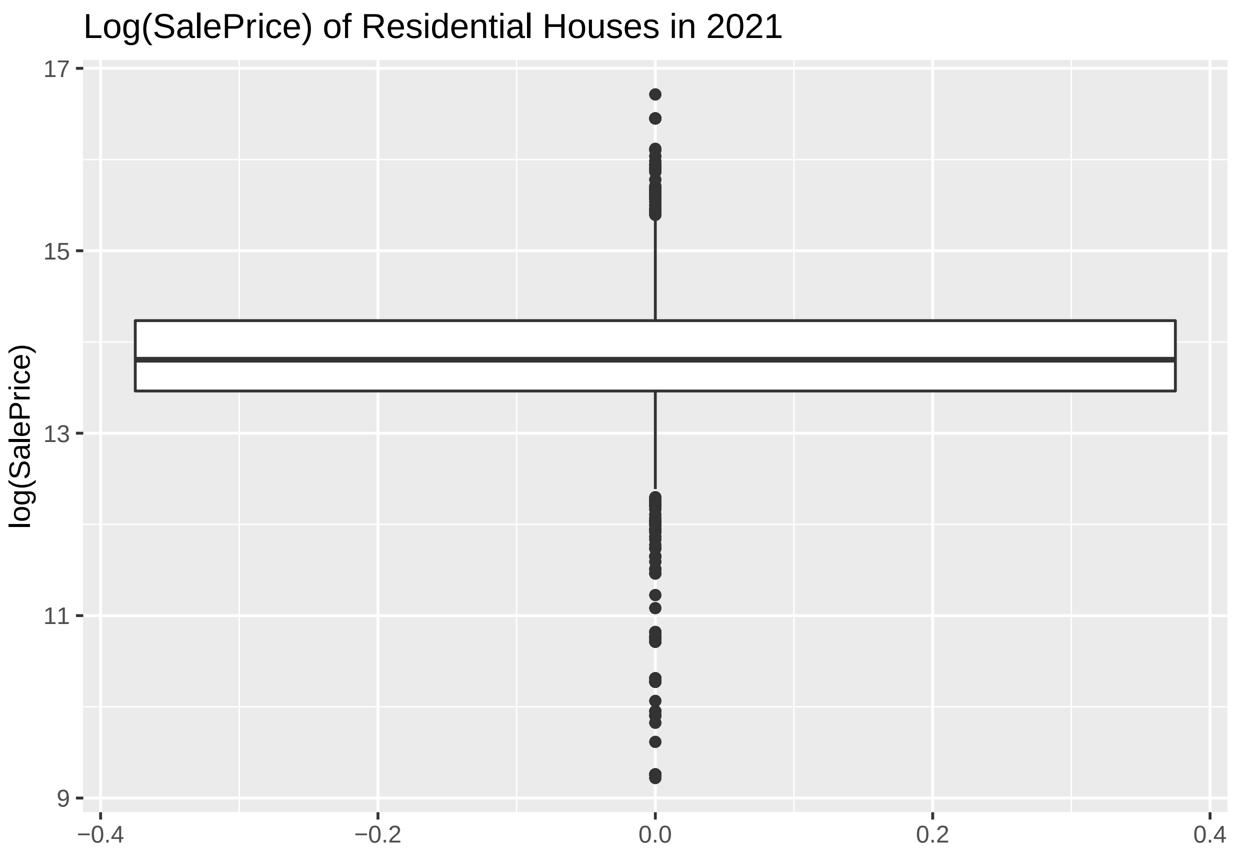
```r
x <- scale(data.matrix(train[,-1]))
y <- train$SalePrice

cv_model <- cv.glmnet(x, y, alpha = 1)

best_lambda <- cv_model$lambda.min

best_lasso <- glmnet(x, y, alpha = 1, lambda = best_lambda)

as.table(as.matrix(best_lasso$beta))
```

```
##                                         s0
## SqFtTotLiving                  448468.6009
## SqFtLot                         65484.5092
## SqFtTotBasement                     0.0000
## SqFtOpenPorch                   66711.6337
## SqFtEnclosedPorch                4378.3253
## SqFtDeck                       -20236.8099
## SqFtGarageAttached              25306.4771
## PcntUnusable                   -26543.6593
## BrickStone                       4898.3244
## HeatSystem                     -28171.5245
## Condition                      -33092.6332
## WaterSystem                     -3858.2876
## SewerSystem                     60397.6148
## TrafficNoise                   -37653.2054
## PowerLines                       -620.4183
## OtherNuisances                  -2907.2423
## HistoricSite                    18782.3929
## NbrLivingUnits                 -23790.6825
## BathFullCount                  -44584.7397
## DistrictName_ALGONA             -4876.1200
## DistrictName_AUBURN            -23676.7740
## DistrictName_BELLEVUE          134016.2929
## DistrictName_BLACK DIAMOND     -15378.0508
## DistrictName_BOTHELL             5391.7371
## DistrictName_BURIEN                 0.0000
## DistrictName_CARNATION         -16307.5313
## DistrictName_CLYDE HILL         48973.6464
## DistrictName_COVINGTON         -15249.9283
## DistrictName_DES MOINES        -22935.3495
## DistrictName_DUVALL            -10889.1201
## DistrictName_ENUMCLAW           -7042.9922
## DistrictName_FEDERAL WAY        26666.1873
## DistrictName_ISSAQUAH           53242.6750
## DistrictName_KENMORE                0.0000
## DistrictName_KENT                8813.3560
## DistrictName_KING COUNTY       -18374.6356
## DistrictName_KIRKLAND           31674.5209
## DistrictName_LAKE FOREST PARK   -7484.3852
## DistrictName_MAPLE VALLEY       -2910.7614
## DistrictName_MEDINA            247441.3475
## DistrictName_MERCER ISLAND     267810.9104
## DistrictName_MILTON            -16375.7933
## DistrictName_NEWCASTLE           -329.2326
## DistrictName_NORMANDY PARK      15768.9245
## DistrictName_NORTH BEND        -33011.4106
## DistrictName_PACIFIC           -27647.3335
## DistrictName_REDMOND            40195.6498
## DistrictName_RENTON             -9110.0358
## DistrictName_SAMMAMISH         163908.2863
## DistrictName_SeaTac            -12407.0934
## DistrictName_SEATTLE            63938.3341
## DistrictName_SHORELINE           7989.4745
## DistrictName_SKYKOMISH         -27785.4499
## DistrictName_SNOQUALMIE        -42150.6434
## DistrictName_TUKWILA            -2589.6470
## DistrictName_WOODINVILLE        -1375.7990
## Type_CoalMineHazard            -38272.1601
## Type_Contamination              -2159.1943
## Type_CriticalDrainage           -4940.6872
## Type_ErosionHazard              -3108.3367
## Type_HundredYrFloodPlain        33003.4949
## Type_LandfillBuffer             -5665.6135
## Type_LandslideHazard             1530.8524
## Type_SeismicHazard                  0.0000
## Type_SensitiveAreaTract         11567.8629
## Type_SpeciesOfConcern           31235.7880
## Type_SteepSlopeHazard           -7911.4973
## Type_Stream                    -14494.4970
## Type_Wetland                        0.0000
## Month_01                       -40793.9656
## Month_02                       -30942.4481
## Month_03                        -1798.1417
## Month_04                       -14778.2065
## Month_05                         7591.5018
## Month_06                        18990.1476
## Month_07                          0.0000
## Month_08                      40213.6892
## Month_09                      -5977.1503
## Month_10                       3373.9739
## Month_11                     -13931.8257
## Month_12                          0.0000
## Year_2020                   -122043.3798
## Year_2021                         0.0000
## Year_2022                         0.0000
```

```
train %>% filter(Year_2020 == 1) %>%
  ggplot(aes(y = log(SalePrice))) +
  geom_boxplot() +
  labs(title = "Log(SalePrice) of Residential Houses in 2020")
```



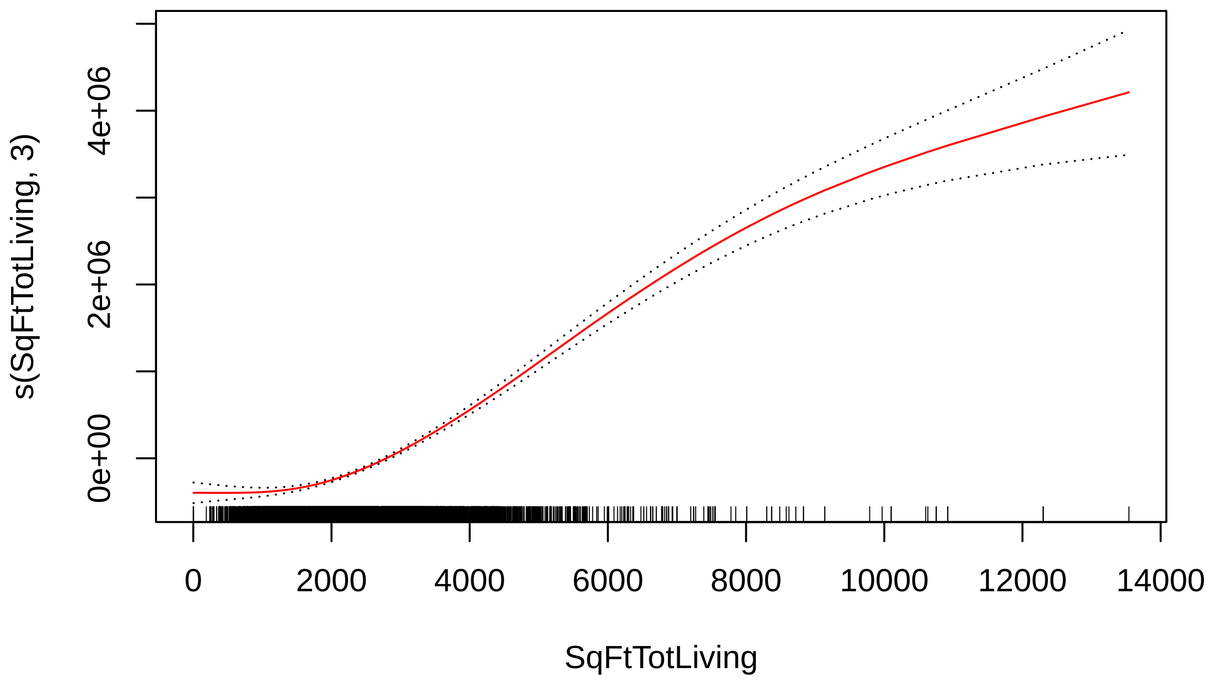
```
train %>% filter(Year_2021 == 1) %>%
  ggplot(aes(y = log(SalePrice))) +
  geom_boxplot() +
  labs(title = "Log(SalePrice) of Residential Houses in 2021")
```

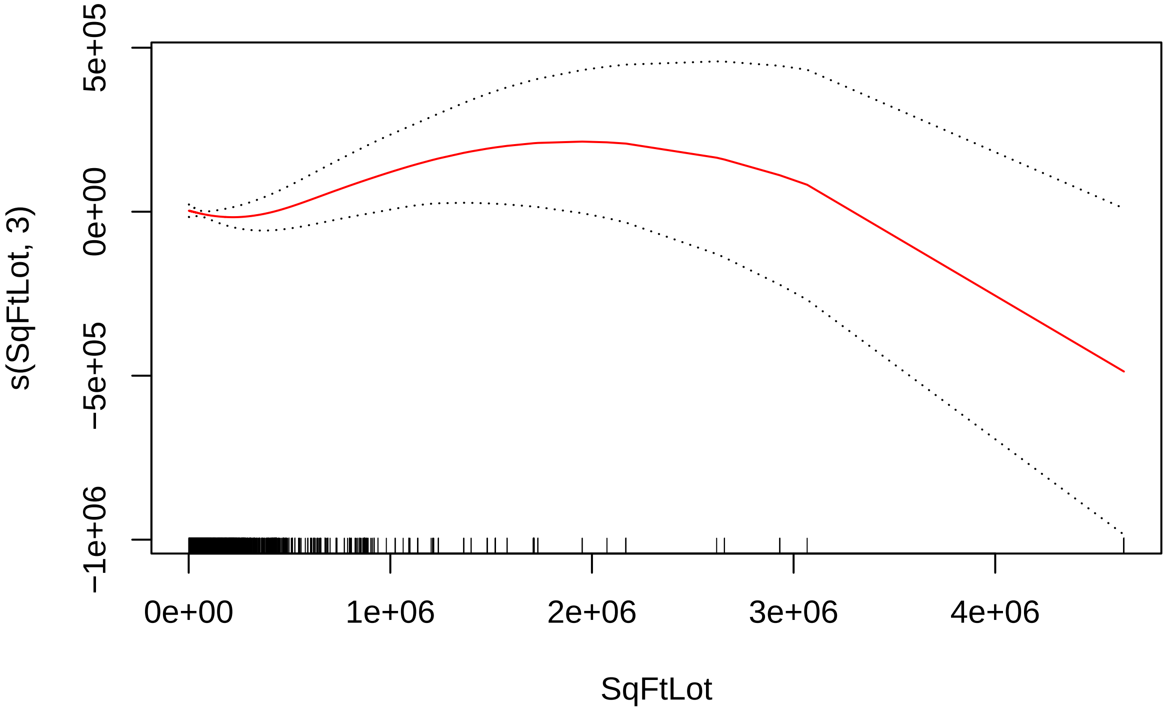


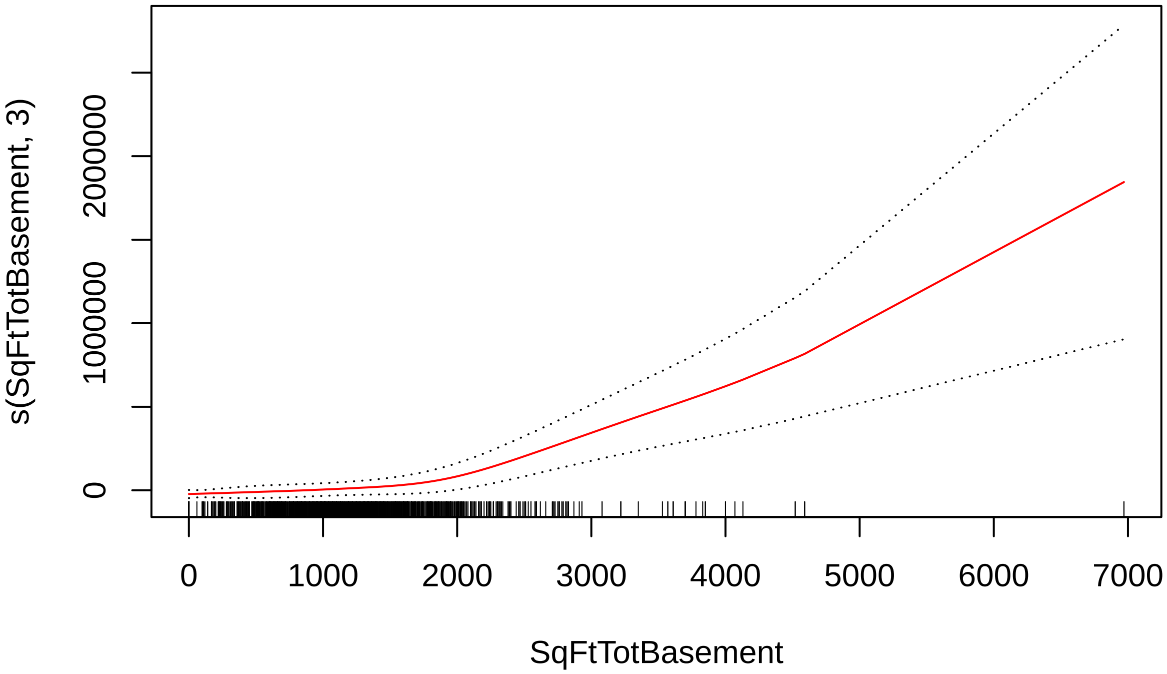
```
# lasso prediction
x2 <- scale(data.matrix(test[,-1]))
y2 <- test$SalePrice
x2[is.na(x2)] <- 0
lasso_pred <- predict(best_lasso, x2)
MSELasso <- MSE(lasso_pred, y2)
MSELasso
```

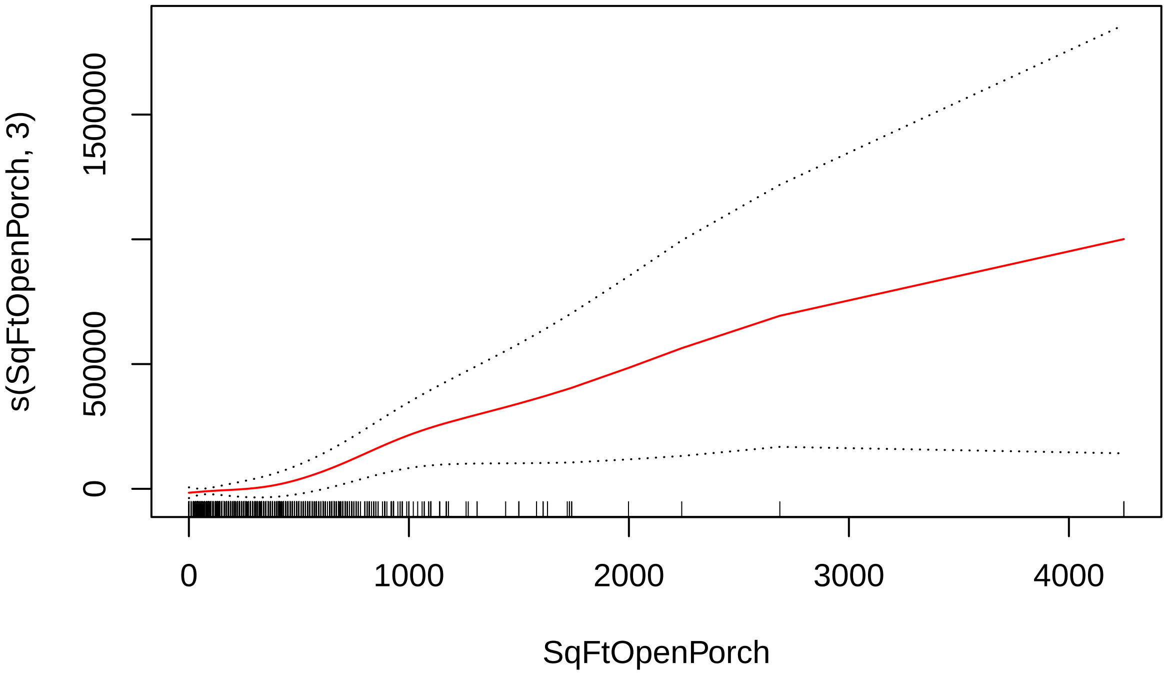
```
## [1] 879128192799
```

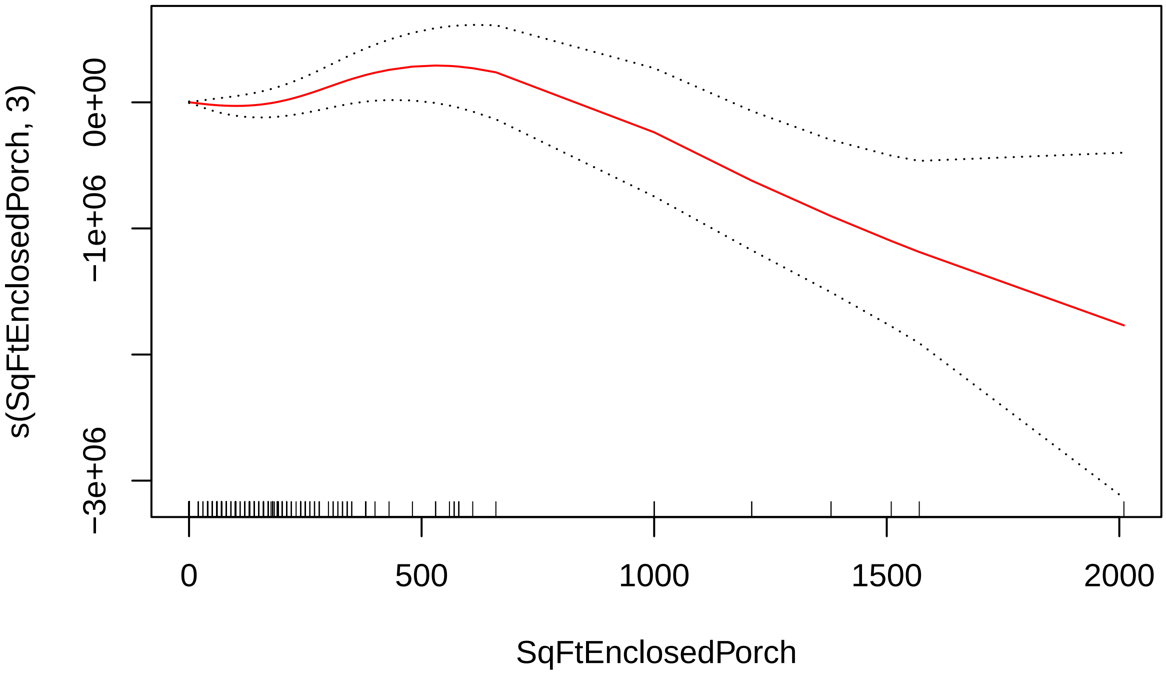
```
# GAM
gam1 <- gam(SalePrice ~ s(SqFtTotLiving,3) + s(SqFtLot,3) + s(SqFtTotBasement,3) + s(SqFtOpenPorch,3) +
plot(gam1, col = "red", se = T)
```

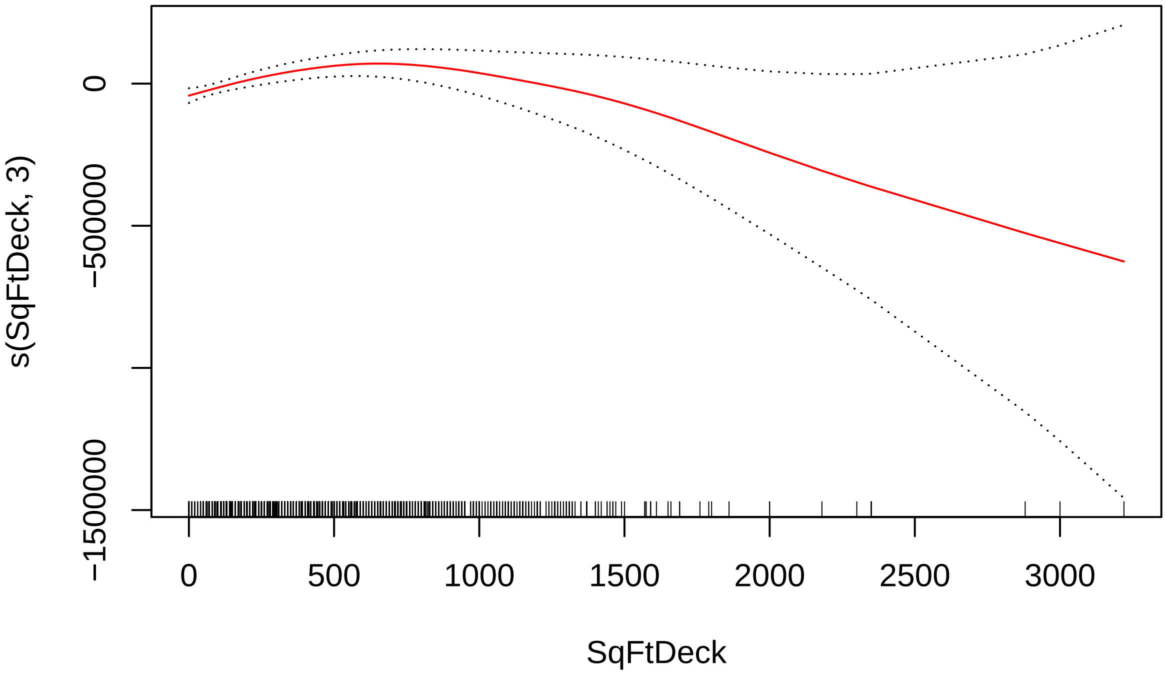

s(SqFtTotLiving, 3)
0e+00
2e+06
4e+06
0
2000
4000
6000
8000
10000
12000
14000
SqFtTotLiving

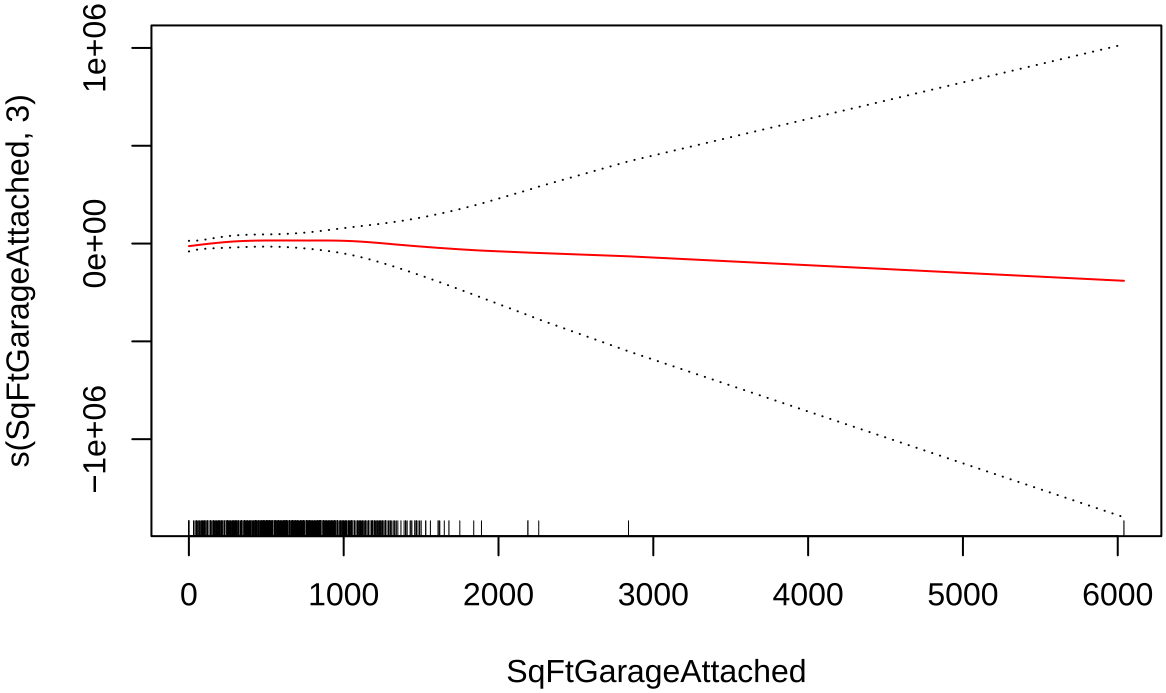

s(SqFtLot, 3)
5e+05
0e+00
−5e+05
−1e+06
0e+00
1e+06
2e+06
3e+06
4e+06
SqFtLot

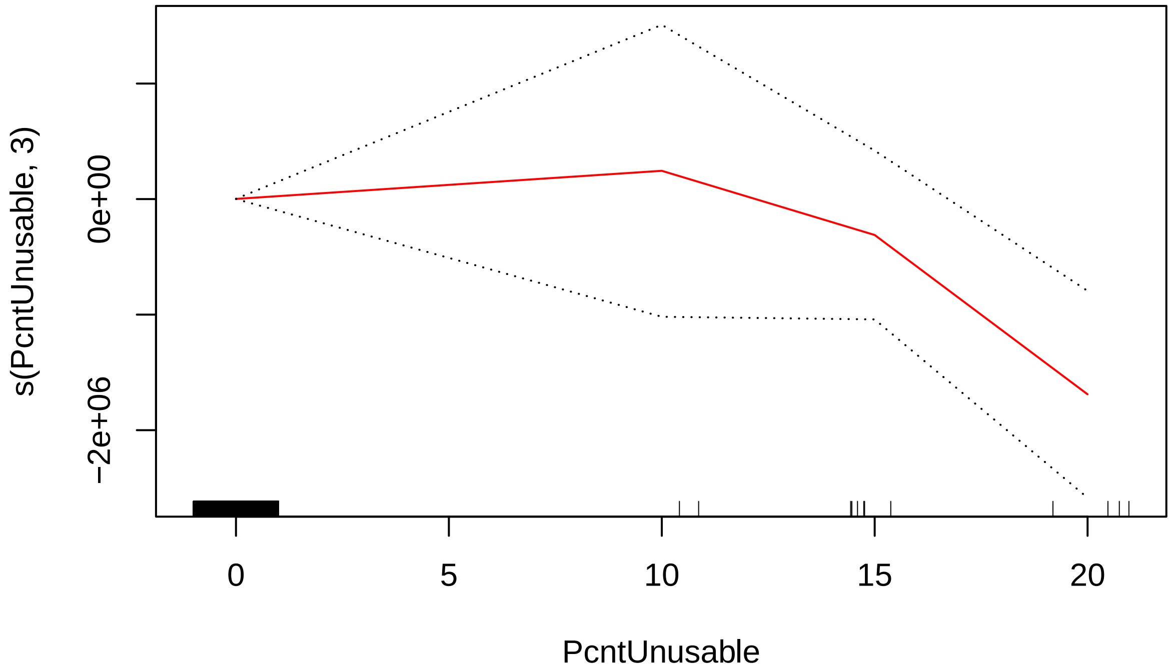

s(SqFtTotBasement, 3)
0
1000000
2000000
0
1000
2000
3000
4000
5000
6000
7000
SqFtTotBasement


s(SqFtOpenPorch, 3)
0
500000
1500000
0
1000
2000
3000
4000
SqFtOpenPorch


s(SqFtEnclosedPorch, 3)
0e+00
−1e+06
−3e+06
0
500
1000
1500
2000
SqFtEnclosedPorch


s(SqFtDeck, 3)
0
−500000
−1500000
0
500
1000
1500
2000
2500
3000
SqFtDeck


s(SqFtGarageAttached, 3)
1e+06
0e+00
−1e+06
0
1000
2000
3000
4000
5000
6000
SqFtGarageAttached



```
# lm taking account of lasso results and GAM plots
m2 <- lm(SalePrice ~ . - Type_Wetland - Month_12 - Month_07 - Year_2021 - Type_SeismicHazard - District
summary(m2)
```

```
##
## Call:
## lm(formula = SalePrice ~ . - Type_Wetland - Month_12 - Month_07 -
##     Year_2021 - Type_SeismicHazard - DistrictName_KENMORE - DistrictName_BURIEN -
##     SqFtTotBasement + SqFtLot^2, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8359809  -306584   -59008   183214 11552122
##
## Coefficients: (1 not defined because of singularities)
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -6.956e+02  3.131e+05  -0.002 0.998228
## SqFtTotLiving                   3.489e+02  1.269e+01  27.488  < 2e-16 ***
## SqFtLot                         2.085e-01  3.650e-02   5.713 1.17e-08 ***
## SqFtOpenPorch                   2.760e+02  4.986e+01   5.536 3.25e-08 ***
## SqFtEnclosedPorch               9.461e+01  1.478e+02   0.640 0.522152
## SqFtDeck                       -7.582e+01  3.707e+01  -2.045 0.040884 *
## SqFtGarageAttached              8.954e+01  3.377e+01   2.651 0.008045 **
## PcntUnusable                   -3.777e+04  1.425e+04  -2.650 0.008065 **
## BrickStone                      2.556e+02  7.667e+02   0.333 0.738861
## HeatSystem1                    -1.585e+05  1.223e+05  -1.296 0.195081
## HeatSystem2                     -1.742e+05  2.646e+05  -0.658 0.510352
## HeatSystem3                      4.799e+04  1.347e+05   0.356 0.721579
## HeatSystem4                     -2.026e+03  1.157e+05  -0.018 0.986028
## HeatSystem5                     -1.850e+05  1.150e+05  -1.608 0.107901
## HeatSystem6                      1.122e+05  1.418e+05   0.791 0.428823
## HeatSystem7                     -2.413e+05  1.192e+05  -2.024 0.042994 *
## HeatSystem8                     -2.362e+05  2.226e+05  -1.061 0.288706
## Condition2                       3.608e+05  1.808e+05   1.995 0.046055 *
## Condition3                       3.158e+05  1.533e+05   2.060 0.039458 *
## Condition4                       1.674e+05  1.540e+05   1.087 0.276967
## Condition5                       2.574e+05  1.555e+05   1.655 0.098024 .
## WaterSystem1                     1.271e+05  3.343e+05   0.380 0.703781
## WaterSystem2                     1.144e+05  3.342e+05   0.342 0.732168
## SewerSystem1                     4.097e+05  2.917e+05   1.405 0.160180
## SewerSystem2                     5.430e+05  2.928e+05   1.855 0.063690 .
## SewerSystem3                     2.038e+04  5.093e+05   0.040 0.968085
## TrafficNoise1                   -6.820e+04  3.713e+04  -1.837 0.066322 .
## TrafficNoise2                   -1.285e+05  4.260e+04  -3.016 0.002571 **
## TrafficNoise3                   -1.922e+05  1.053e+05  -1.824 0.068144 .
## PowerLinesY                     -3.674e+04  9.412e+04  -0.390 0.696323
## OtherNuisancesY                 -1.779e+04  5.377e+04  -0.331 0.740774
## HistoricSite3                    1.509e+06  7.891e+05   1.912 0.055897 .
## NbrLivingUnits                  -2.995e+05  1.050e+05  -2.852 0.004362 **
## BathFullCount                   -6.176e+04  1.790e+04  -3.450 0.000565 ***
## DistrictName_ALGONA             -5.966e+05  8.016e+05  -0.744 0.456766
## DistrictName_AUBURN             -2.912e+05  1.528e+05  -1.905 0.056778 .
## DistrictName_BELLEVUE            7.490e+05  9.245e+04   8.101 6.69e-16 ***
## `DistrictName_BLACK DIAMOND`    -6.545e+05  4.041e+05  -1.619 0.105400
## DistrictName_BOTHELL             7.191e+04  1.533e+05   0.469 0.639150
## DistrictName_CARNATION          -2.748e+05  1.848e+05  -1.487 0.137060
## `DistrictName_CLYDE HILL`        1.395e+06  3.079e+05   4.531 5.99e-06 ***
## DistrictName_COVINGTON          -2.844e+05  2.064e+05  -1.378 0.168218
## `DistrictName_DES MOINES`       -4.356e+05  1.912e+05  -2.278 0.022781 *
## DistrictName_DUVALL             -2.287e+05  2.105e+05  -1.087 0.277232
## DistrictName_ENUMCLAW           -2.853e+05  4.019e+05  -0.710 0.477817
## `DistrictName_FEDERAL WAY`       6.159e+05  2.397e+05   2.570 0.010211 *
## DistrictName_ISSAQUAH            3.595e+05  1.102e+05   3.261 0.001116 **
## DistrictName_KENT                6.255e+04  1.009e+05   0.620 0.535200
## `DistrictName_KING COUNTY`      -4.198e+04  8.110e+04  -0.518 0.604684
## DistrictName_KIRKLAND            2.952e+05  1.192e+05   2.476 0.013313 *
## `DistrictName_LAKE FOREST PARK` -7.712e+04  1.186e+05  -0.650 0.515681
## `DistrictName_MAPLE VALLEY`     -9.614e+04  2.899e+05  -0.332 0.740208
## DistrictName_MEDINA              8.170e+06  3.625e+05  22.536  < 2e-16 ***
## `DistrictName_MERCER ISLAND`     1.594e+06  1.026e+05  15.542  < 2e-16 ***
## DistrictName_MILTON             -7.153e+05  3.609e+05  -1.982 0.047560 *
## DistrictName_NEWCASTLE          -4.846e+03  2.612e+05  -0.019 0.985197
## `DistrictName_NORMANDY PARK`     2.198e+05  1.663e+05   1.322 0.186309
## `DistrictName_NORTH BEND`       -1.523e+05  8.533e+04  -1.785 0.074304 .
## DistrictName_PACIFIC            -5.108e+05  2.007e+05  -2.546 0.010936 *
## DistrictName_REDMOND             4.181e+05  1.228e+05   3.404 0.000668 ***
## DistrictName_RENTON             -2.048e+05  1.877e+05  -1.091 0.275388
## DistrictName_SAMMAMISH           9.716e+05  9.624e+04  10.095  < 2e-16 ***
## DistrictName_SeaTac             -2.668e+05  2.169e+05  -1.230 0.218745
## DistrictName_SEATTLE             2.667e+05  8.963e+04   2.976 0.002937 **
## DistrictName_SHORELINE           7.019e+04  1.176e+05   0.597 0.550575
## DistrictName_SKYKOMISH          -5.057e+05  1.945e+05  -2.601 0.009334 **
## DistrictName_SNOQUALMIE         -2.479e+05  9.649e+04  -2.569 0.010228 *
## DistrictName_TUKWILA            -8.715e+04  1.803e+05  -0.483 0.628828
## DistrictName_WOODINVILLE        -7.643e+04  2.423e+05  -0.315 0.752415
## Type_CoalMineHazard             -3.017e+05  1.041e+05  -2.898 0.003769 **
## Type_Contamination              -1.646e+05  2.908e+05  -0.566 0.571446
## Type_CriticalDrainage           -8.325e+04  1.800e+05  -0.462 0.643789
## Type_ErosionHazard              -2.571e+04  4.165e+04  -0.617 0.537047
## Type_HundredYrFloodPlain         9.801e+04  3.665e+04   2.674 0.007511 **
## Type_LandfillBuffer             -4.238e+05  5.580e+05  -0.759 0.447636
## Type_LandslideHazard            -7.496e+03  4.535e+04  -0.165 0.868706
## Type_SensitiveAreaTract          8.168e+04  6.481e+04   1.260 0.207566
## Type_SpeciesOfConcern            1.209e+06  3.984e+05   3.034 0.002422 **
## Type_SteepSlopeHazard           -6.702e+04  5.611e+04  -1.194 0.232355
## Type_Stream                     -4.078e+04  3.294e+04  -1.238 0.215734
## Month_01                        -2.278e+05  5.853e+04  -3.892 0.000101 ***
## Month_02                        -1.460e+05  5.286e+04  -2.763 0.005753 **
## Month_03                        -3.712e+04  4.808e+04  -0.772 0.440200
## Month_04                        -7.726e+04  4.887e+04  -1.581 0.113985
## Month_05                         1.133e+04  4.511e+04   0.251 0.801735
## Month_06                         6.071e+04  4.211e+04   1.442 0.149480
## Month_08                         1.320e+05  4.213e+04   3.134 0.001737 **
## Month_09                        -2.846e+04  4.332e+04  -0.657 0.511316
## Month_10                         3.421e+03  4.472e+04   0.076 0.939039
## Month_11                        -6.897e+04  4.629e+04  -1.490 0.136294
## Year_2020                       -2.482e+05  2.194e+04 -11.312  < 2e-16 ***
## Year_2022                               NA         NA      NA       NA
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 787100 on 5275 degrees of freedom
## Multiple R-squared:  0.4687, Adjusted R-squared:  0.4596
## F-statistic:  51.7 on 90 and 5275 DF,  p-value: < 2.2e-16
```

```r
m2_pred <- predict(m2, test[,-1])
```

```
## Warning in predict.lm(m2, test[, -1]): prediction from a rank-deficient fit may
## be misleading
```

```r
MSEm2 <- MSE(m2_pred, test$SalePrice)
MSEm2
```

```
## [1] 947454467084
```

```r
# SqFtTotLiving is important?
m3 <- lm(SalePrice ~ SqFtTotLiving, data = train)
summary(m3)$adj.r.squared
```

```
## [1] 0.2591794
```

```r
m4 <- lm(SalePrice ~ SqFtLot, data = train)
summary(m4)$adj.r.squared
```

```
## [1] 0.00120881
```

```r
# double lasso
fm <- as.formula(~ . - SalePrice - 1 + SqFtTotLiving * (.))
X <- model.matrix(fm, data = train[train$Year_2020 == 1,])
Y <- train[train$Year_2020 == 1,]$SalePrice
# only want SqFtTotLiving related variables
index.liv <- grep("SqFtTotLiving", colnames(X))
reg.out<-lm(Y~X)
#coefficients for SqFtTotLiving&interactions
index.liv.regout<-grep("SqFtTotLiving",names(reg.out$coefficients))

# Partialling out
# double lasso regress Y on X with the columns in index.liv as focal
effects.liv.ds <- rlassoEffects(x = X, y = Y, method = "double selection",index = index.liv)
summary(effects.liv.ds)
```

```
## [1] "Estimates and significance testing of the effect of target variables"
##                                                   Estimate. Std. Error  t value
## SqFtTotLiving                                    -1.108e-12  1.039e+03    0.000
## SalePrice:SqFtTotLiving                           5.019e-19  9.471e-21   52.995
## SqFtTotLiving:SqFtLot                            -9.404e-20  1.818e-21  -51.725
## SqFtTotLiving:SqFtTotBasement                     2.240e-16  9.067e-18   24.705
## SqFtTotLiving:SqFtOpenPorch                       1.049e-15  2.213e-17   47.406
## SqFtTotLiving:SqFtEnclosedPorch                   1.622e-15  3.578e-17   45.337
## SqFtTotLiving:SqFtDeck                            5.969e-16  1.209e-17   49.364
## SqFtTotLiving:SqFtGarageAttached                 -5.313e-16  3.384e-17  -15.697
## SqFtTotLiving:PcntUnusable                       -3.905e-14  3.274e-15  -11.927
## SqFtTotLiving:BrickStone                          4.813e-15  1.872e-16   25.705
## SqFtTotLiving:HeatSystem1                         3.127e-12  7.471e-15  418.627
## SqFtTotLiving:HeatSystem2                        -2.576e-13  1.677e-13   -1.536
## SqFtTotLiving:HeatSystem3                         2.402e-13  1.329e-13    1.808
## SqFtTotLiving:HeatSystem4                         2.338e-13  1.938e-14   12.061
## SqFtTotLiving:HeatSystem5                        -7.858e-14  1.391e-14   -5.648
## SqFtTotLiving:HeatSystem6                         8.033e-13  2.349e-14   34.192
## SqFtTotLiving:HeatSystem7                         5.873e-14  1.028e-14    5.715
## SqFtTotLiving:HeatSystem8                         7.952e-13  9.823e-15   80.955
## SqFtTotLiving:Condition2                          2.069e-13  7.808e-15   26.493
## SqFtTotLiving:Condition3                         -7.820e-14  1.397e-14   -5.596
## SqFtTotLiving:Condition4                         -4.511e-14  9.811e-15   -4.598
## SqFtTotLiving:Condition5                          1.930e-13  9.547e-15   20.215
## SqFtTotLiving:WaterSystem1                        8.150e-14  1.101e-14    7.402
## SqFtTotLiving:WaterSystem2                       -2.446e-13  3.098e-14   -7.896
## SqFtTotLiving:SewerSystem1                        7.756e-14  1.190e-14    6.519
## SqFtTotLiving:SewerSystem2                        6.185e-13  4.293e-14   14.408
## SqFtTotLiving:SewerSystem3                        6.750e-14  9.311e-02    0.000
## SqFtTotLiving:TrafficNoise1                      -8.546e-14  1.984e-14   -4.307
## SqFtTotLiving:TrafficNoise2                       4.221e-13  1.367e-14   30.886
## SqFtTotLiving:TrafficNoise3                      -5.860e-13  1.620e-14  -36.167
## SqFtTotLiving:PowerLinesY                     5.215e-13  2.056e-14   25.364
## SqFtTotLiving:OtherNuisancesY                -3.586e-13  2.009e-14  -17.846
## SqFtTotLiving:HistoricSite3                  -4.782e-12  6.173e-01    0.000
## SqFtTotLiving:NbrLivingUnits                 -1.074e-12  8.626e-14  -12.453
## SqFtTotLiving:BathFullCount                  -3.305e-13  1.099e-14  -30.061
## SqFtTotLiving:DistrictName_ALGONA                    NA        NaN       NA
## SqFtTotLiving:DistrictName_AUBURN             2.297e-13  8.194e-15   28.035
## SqFtTotLiving:DistrictName_BELLEVUE          -3.113e-13  8.676e-15  -35.887
## SqFtTotLiving:‘DistrictName_BLACK DIAMOND‘           NA        NaN       NA
## SqFtTotLiving:DistrictName_BOTHELL            4.330e-13  1.195e-14   36.234
## SqFtTotLiving:DistrictName_BURIEN             7.606e-13  1.628e-14   46.711
## SqFtTotLiving:DistrictName_CARNATION         -7.425e-14  7.911e-14   -0.939
## SqFtTotLiving:‘DistrictName_CLYDE HILL‘       1.081e-13  1.011e-01    0.000
## SqFtTotLiving:DistrictName_COVINGTON          1.606e-13  8.457e-14    1.899
## SqFtTotLiving:‘DistrictName_DES MOINES‘      -1.057e-13  6.949e-15  -15.204
## SqFtTotLiving:DistrictName_DUVALL            -4.064e-13  2.320e-15 -175.178
## SqFtTotLiving:DistrictName_ENUMCLAW           8.913e-13  2.852e-14   31.256
## SqFtTotLiving:‘DistrictName_FEDERAL WAY‘     -5.695e-13  1.917e-14  -29.704
## SqFtTotLiving:DistrictName_ISSAQUAH           2.915e-14  7.652e-01    0.000
## SqFtTotLiving:DistrictName_KENMORE            8.257e-13  2.402e-14   34.376
## SqFtTotLiving:DistrictName_KENT               2.927e-12  3.789e+00    0.000
## SqFtTotLiving:‘DistrictName_KING COUNTY‘      6.047e-13  4.783e-14   12.643
## SqFtTotLiving:DistrictName_KIRKLAND           1.518e-13  2.042e+00    0.000
## SqFtTotLiving:‘DistrictName_LAKE FOREST PARK‘ -6.699e-14  1.465e-12   -0.046
## SqFtTotLiving:‘DistrictName_MAPLE VALLEY‘     2.381e-13  2.294e-14   10.379
## SqFtTotLiving:DistrictName_MEDINA             3.250e-12  2.403e-14  135.255
## SqFtTotLiving:‘DistrictName_MERCER ISLAND‘   -1.834e-12  9.230e-13   -1.987
## SqFtTotLiving:DistrictName_MILTON            -9.751e-13  1.840e-13   -5.300
## SqFtTotLiving:DistrictName_NEWCASTLE          6.556e-13  2.017e-14   32.499
## SqFtTotLiving:‘DistrictName_NORMANDY PARK‘   -7.482e-12  2.435e-14 -307.310
## SqFtTotLiving:‘DistrictName_NORTH BEND‘      -3.822e-13  1.209e+00    0.000
## SqFtTotLiving:DistrictName_PACIFIC            5.422e-14  5.989e-14    0.905
## SqFtTotLiving:DistrictName_REDMOND            4.363e-12  6.948e-14   62.801
## SqFtTotLiving:DistrictName_RENTON             1.165e-12  8.575e-14   13.585
## SqFtTotLiving:DistrictName_SAMMAMISH         -4.059e-13  1.527e-14  -26.590
## SqFtTotLiving:DistrictName_SeaTac            -2.465e-13  6.258e-15  -39.383
## SqFtTotLiving:DistrictName_SEATTLE           -1.007e-12  8.388e-14  -12.004
## SqFtTotLiving:DistrictName_SHORELINE         -1.828e-13  3.819e-15  -47.863
## SqFtTotLiving:DistrictName_SKYKOMISH         -2.421e-13  1.178e-14  -20.561
## SqFtTotLiving:DistrictName_SNOQUALMIE         2.077e-13  5.233e-15   39.684
## SqFtTotLiving:DistrictName_TUKWILA           -1.556e-12  3.597e-14  -43.269
## SqFtTotLiving:DistrictName_WOODINVILLE        1.405e-13  1.491e-14    9.423
## SqFtTotLiving:Type_CoalMineHazard             1.789e-13  1.065e-14   16.792
## SqFtTotLiving:Type_Contamination             -2.952e-13  3.310e-14   -8.918
## SqFtTotLiving:Type_CriticalDrainage          -3.807e-13  1.866e-14  -20.403
## SqFtTotLiving:Type_ErosionHazard              3.118e-15  9.892e-01    0.000
## SqFtTotLiving:Type_HundredYrFloodPlain       -9.266e-14  3.161e+00    0.000
## SqFtTotLiving:Type_LandfillBuffer             2.113e-13  4.452e+00    0.000
## SqFtTotLiving:Type_LandslideHazard            2.328e-13  1.126e+01    0.000
## SqFtTotLiving:Type_SeismicHazard             -3.253e-14  1.903e-01    0.000
## SqFtTotLiving:Type_SensitiveAreaTract        -1.637e-14  3.606e+00    0.000
## SqFtTotLiving:Type_SpeciesOfConcern           1.287e-12  1.376e+00    0.000
## SqFtTotLiving:Type_SteepSlopeHazard          -5.124e-14  1.330e+00    0.000
## SqFtTotLiving:Type_Stream                    -3.356e-14  3.217e+00    0.000
## SqFtTotLiving:Type_Wetland                     4.101e-13  5.924e+00    0.000
## SqFtTotLiving:Month_01                        -2.813e-14  8.423e-01    0.000
## SqFtTotLiving:Month_02                         5.852e-15  3.629e-01    0.000
## SqFtTotLiving:Month_03                        -2.930e-15  3.745e-01    0.000
## SqFtTotLiving:Month_04                         1.129e-13  3.944e+00    0.000
## SqFtTotLiving:Month_05                         1.659e-14  8.741e-01    0.000
## SqFtTotLiving:Month_06                        -3.418e-13  4.537e+00    0.000
## SqFtTotLiving:Month_07                        -2.869e-14  7.364e-01    0.000
## SqFtTotLiving:Month_08                         2.668e-13  4.684e+00    0.000
## SqFtTotLiving:Month_09                        -1.663e-13  3.673e+00    0.000
## SqFtTotLiving:Month_10                        -1.378e-14  1.116e+00    0.000
## SqFtTotLiving:Month_11                         7.670e-15  3.333e+00    0.000
## SqFtTotLiving:Month_12                         1.287e-14  5.153e-01    0.000
## SqFtTotLiving:Year_2020                       -1.108e-12  1.848e+01    0.000
## SqFtTotLiving:Year_2021                               NA        NaN       NA
## SqFtTotLiving:Year_2022                               NA        NaN       NA
##                                               Pr(>|t|)
## SqFtTotLiving                                   1.0000
## SalePrice:SqFtTotLiving                       < 2e-16 ***
## SqFtTotLiving:SqFtLot                         < 2e-16 ***
## SqFtTotLiving:SqFtTotBasement                 < 2e-16 ***
## SqFtTotLiving:SqFtOpenPorch                   < 2e-16 ***
## SqFtTotLiving:SqFtEnclosedPorch               < 2e-16 ***
## SqFtTotLiving:SqFtDeck                        < 2e-16 ***
## SqFtTotLiving:SqFtGarageAttached              < 2e-16 ***
## SqFtTotLiving:PcntUnusable                    < 2e-16 ***
## SqFtTotLiving:BrickStone                      < 2e-16 ***
## SqFtTotLiving:HeatSystem1                     < 2e-16 ***
## SqFtTotLiving:HeatSystem2                      0.1245
## SqFtTotLiving:HeatSystem3                      0.0706 .
## SqFtTotLiving:HeatSystem4                     < 2e-16 ***
## SqFtTotLiving:HeatSystem5                    1.62e-08 ***
## SqFtTotLiving:HeatSystem6                     < 2e-16 ***
## SqFtTotLiving:HeatSystem7                    1.10e-08 ***
## SqFtTotLiving:HeatSystem8                     < 2e-16 ***
## SqFtTotLiving:Condition2                      < 2e-16 ***
## SqFtTotLiving:Condition3                     2.19e-08 ***
## SqFtTotLiving:Condition4                     4.27e-06 ***
## SqFtTotLiving:Condition5                      < 2e-16 ***
## SqFtTotLiving:WaterSystem1                   1.35e-13 ***
## SqFtTotLiving:WaterSystem2                   2.88e-15 ***
## SqFtTotLiving:SewerSystem1                   7.06e-11 ***
## SqFtTotLiving:SewerSystem2                    < 2e-16 ***
## SqFtTotLiving:SewerSystem3                     1.0000
## SqFtTotLiving:TrafficNoise1                  1.65e-05 ***
## SqFtTotLiving:TrafficNoise2                   < 2e-16 ***
## SqFtTotLiving:TrafficNoise3                   < 2e-16 ***
## SqFtTotLiving:PowerLinesY                     < 2e-16 ***
## SqFtTotLiving:OtherNuisancesY                 < 2e-16 ***
## SqFtTotLiving:HistoricSite3                    1.0000
## SqFtTotLiving:NbrLivingUnits                  < 2e-16 ***
## SqFtTotLiving:BathFullCount                   < 2e-16 ***
## SqFtTotLiving:DistrictName_ALGONA                   NA
## SqFtTotLiving:DistrictName_AUBURN             < 2e-16 ***
## SqFtTotLiving:DistrictName_BELLEVUE            < 2e-16 ***
## SqFtTotLiving:`DistrictName_BLACK DIAMOND`          NA    
## SqFtTotLiving:DistrictName_BOTHELL             < 2e-16 ***
## SqFtTotLiving:DistrictName_BURIEN              < 2e-16 ***
## SqFtTotLiving:DistrictName_CARNATION            0.3479    
## SqFtTotLiving:`DistrictName_CLYDE HILL`         1.0000    
## SqFtTotLiving:DistrictName_COVINGTON            0.0575 .  
## SqFtTotLiving:`DistrictName_DES MOINES`        < 2e-16 ***
## SqFtTotLiving:DistrictName_DUVALL              < 2e-16 ***
## SqFtTotLiving:DistrictName_ENUMCLAW            < 2e-16 ***
## SqFtTotLiving:`DistrictName_FEDERAL WAY`       < 2e-16 ***
## SqFtTotLiving:DistrictName_ISSAQUAH             1.0000    
## SqFtTotLiving:DistrictName_KENMORE             < 2e-16 ***
## SqFtTotLiving:DistrictName_KENT                 1.0000    
## SqFtTotLiving:`DistrictName_KING COUNTY`       < 2e-16 ***
## SqFtTotLiving:DistrictName_KIRKLAND             1.0000    
## SqFtTotLiving:`DistrictName_LAKE FOREST PARK`   0.9635    
## SqFtTotLiving:`DistrictName_MAPLE VALLEY`      < 2e-16 ***
## SqFtTotLiving:DistrictName_MEDINA              < 2e-16 ***
## SqFtTotLiving:`DistrictName_MERCER ISLAND`      0.0469 *  
## SqFtTotLiving:DistrictName_MILTON             1.16e-07 ***
## SqFtTotLiving:DistrictName_NEWCASTLE           < 2e-16 ***
## SqFtTotLiving:`DistrictName_NORMANDY PARK`     < 2e-16 ***
## SqFtTotLiving:`DistrictName_NORTH BEND`         1.0000    
## SqFtTotLiving:DistrictName_PACIFIC              0.3653    
## SqFtTotLiving:DistrictName_REDMOND             < 2e-16 ***
## SqFtTotLiving:DistrictName_RENTON              < 2e-16 ***
## SqFtTotLiving:DistrictName_SAMMAMISH           < 2e-16 ***
## SqFtTotLiving:DistrictName_SeaTac              < 2e-16 ***
## SqFtTotLiving:DistrictName_SEATTLE             < 2e-16 ***
## SqFtTotLiving:DistrictName_SHORELINE           < 2e-16 ***
## SqFtTotLiving:DistrictName_SKYKOMISH           < 2e-16 ***
## SqFtTotLiving:DistrictName_SNOQUALMIE          < 2e-16 ***
## SqFtTotLiving:DistrictName_TUKWILA             < 2e-16 ***
## SqFtTotLiving:DistrictName_WOODINVILLE         < 2e-16 ***
## SqFtTotLiving:Type_CoalMineHazard              < 2e-16 ***
## SqFtTotLiving:Type_Contamination               < 2e-16 ***
## SqFtTotLiving:Type_CriticalDrainage            < 2e-16 ***
## SqFtTotLiving:Type_ErosionHazard                1.0000    
## SqFtTotLiving:Type_HundredYrFloodPlain          1.0000    
## SqFtTotLiving:Type_LandfillBuffer               1.0000    
## SqFtTotLiving:Type_LandslideHazard              1.0000    
## SqFtTotLiving:Type_SeismicHazard                1.0000    
## SqFtTotLiving:Type_SensitiveAreaTract           1.0000    
## SqFtTotLiving:Type_SpeciesOfConcern             1.0000    
## SqFtTotLiving:Type_SteepSlopeHazard             1.0000    
## SqFtTotLiving:Type_Stream                       1.0000    
## SqFtTotLiving:Type_Wetland                      1.0000    
## SqFtTotLiving:Month_01                          1.0000    
## SqFtTotLiving:Month_02                          1.0000    
## SqFtTotLiving:Month_03                          1.0000    
## SqFtTotLiving:Month_04                          1.0000    
## SqFtTotLiving:Month_05                          1.0000    
## SqFtTotLiving:Month_06                          1.0000    
## SqFtTotLiving:Month_07                          1.0000
## SqFtTotLiving:Month_08                          1.0000
## SqFtTotLiving:Month_09                          1.0000
## SqFtTotLiving:Month_10                          1.0000
## SqFtTotLiving:Month_11                          1.0000
## SqFtTotLiving:Month_12                          1.0000
## SqFtTotLiving:Year_2020                         1.0000
## SqFtTotLiving:Year_2021                             NA
## SqFtTotLiving:Year_2022                             NA
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
# double lasso
fm <- as.formula(~ . - SalePrice - 1)
X <- model.matrix(fm, data = train[train$Year_2021 == 1,])
Y <- train[train$Year_2021 == 1,]$SalePrice
# only want SqFtTotLiving related variables
index.liv <- grep("SqFtTotLiving", colnames(X))
reg.out<-lm(Y~X)
#coefficients for SqFtTotLiving&interactions
index.liv.regout<-grep("SqFtTotLiving",names(reg.out$coefficients))

# Partialling out
# double lasso regress Y on X with the columns in index.liv as focal
effects.liv.ds <- rlassoEffects(x = X, y = Y, method = "double selection",index = index.liv)
summary(effects.liv.ds)
```

```
## [1] "Estimates and significance testing of the effect of target variables"
##               Estimate. Std. Error t value Pr(>|t|)
## SqFtTotLiving    387.01      32.37   11.96   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
# double lasso
fm <- as.formula(~ . - SalePrice - 1)
X <- model.matrix(fm, data = test)
Y <- test$SalePrice
# only want SqFtTotLiving related variables
index.liv <- grep("SqFtTotLiving", colnames(X))
reg.out<-lm(Y~X)
#coefficients for SqFtTotLiving&interactions
index.liv.regout<-grep("SqFtTotLiving",names(reg.out$coefficients))

# Partialling out
# double lasso regress Y on X with the columns in index.liv as focal
effects.liv.ds <- rlassoEffects(x = X, y = Y, method = "double selection",index = index.liv)
summary(effects.liv.ds)
```

```
## [1] "Estimates and significance testing of the effect of target variables"
##               Estimate. Std. Error t value Pr(>|t|)
## SqFtTotLiving     652.1      103.1   6.328 2.49e-10 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
sum(lasso_pred - test$SalePrice < 1)
```

```
## [1] 545
```