Housing Price

LOAD LIBRARIES AND DATA

```
library(tidyverse)
library(leaps)
library(margins)
library(earth)
library(ggeffects)
library(vip)

data <- read.csv("housing_5000.csv")

glimpse(data)
```

```
## Rows: 5,000
## Columns: 16
## $ PRICEK     <dbl> 137.5, 80.0, 365.5, 256.0, 721.5, 177.0, 518.0, 318.0, 148.~
## $ DSALE      <int> 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1,~
## $ BEDROOM    <int> 3, 3, 5, 3, 4, 4, 4, 3, 3, 3, 5, 5, 2, 4, 4, 4, 7, 4, 4, 3,~
## $ SQFTK      <dbl> 0.893, 1.088, 1.150, 1.496, 2.132, 1.392, 1.421, 1.989, 1.0~
## $ LN_LOTSIZE <dbl> 8.031060, 8.039157, 8.300529, 8.499844, 7.783224, 8.383662,~
## $ CENTRALAIR <int> 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0,~
## $ BRICK      <int> 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0,~
## $ GARAGE     <int> 2, 0, 2, 2, 2, 1, 2, 2, 1, 0, 2, 2, 2, 2, 2, 2, 2, 2, 0,~
## $ FIREPLACE  <int> 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 4, 0, 0, 0, 0, 0, 0, 2, 0,~
## $ BASE_FIN   <int> 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1,~
## $ MASONRY    <int> 0, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0,~
## $ PUBOPEN    <int> 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,~
## $ MICHLAKE   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ LAKE_RIVER <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ METRA_DIST <dbl> 1.49801, 3.05638, 2.13873, 1.87168, 0.86320, 0.87661, 1.371~
## $ RAIL_DIST  <dbl> 0.20677, 0.19410, 0.85084, 1.03969, 0.05104, 0.39791, 0.176~
```

A

```
# Identify best subsets
models <- regsubsets(log(PRICEK)~., data = data)
summary(models)
```

```
## Subset selection object
## Call: regsubsets.formula(log(PRICEK) ~ ., data = data)
## 15 Variables  (and intercept)
##            Forced in Forced out
```

```
## DSALE          FALSE      FALSE
## BEDROOM        FALSE      FALSE
## SQFTK          FALSE      FALSE
## LN_LOTSIZE     FALSE      FALSE
## CENTRALAIR     FALSE      FALSE
## BRICK          FALSE      FALSE
## GARAGE         FALSE      FALSE
## FIREPLACE      FALSE      FALSE
## BASE_FIN       FALSE      FALSE
## MASONRY        FALSE      FALSE
## PUBOPEN        FALSE      FALSE
## MICHLAKE       FALSE      FALSE
## LAKE_RIVER     FALSE      FALSE
## METRA_DIST     FALSE      FALSE
## RAIL_DIST      FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          DSALE BEDROOM SQFTK LN_LOTSIZE CENTRALAIR BRICK GARAGE FIREPLACE
## 1  ( 1 ) " "   " "     "*"   " "        " "        " "   " "    " "
## 2  ( 1 ) " "   " "     "*"   " "        " "        " "   " "    "*"
## 3  ( 1 ) " "   " "     "*"   " "        "*"        " "   " "    "*"
## 4  ( 1 ) "*"   " "     "*"   " "        "*"        " "   " "    "*"
## 5  ( 1 ) "*"   " "     "*"   " "        "*"        " "   " "    "*"
## 6  ( 1 ) "*"   " "     "*"   "*"        "*"        " "   " "    "*"
## 7  ( 1 ) "*"   " "     "*"   "*"        "*"        " "   " "    "*"
## 8  ( 1 ) "*"   " "     "*"   "*"        "*"        " "   "*"    "*"
##          BASE_FIN MASONRY PUBOPEN MICHLAKE LAKE_RIVER METRA_DIST RAIL_DIST
## 1  ( 1 ) " "      " "     " "     " "      " "        " "        " "
## 2  ( 1 ) " "      " "     " "     " "      " "        " "        " "
## 3  ( 1 ) " "      " "     " "     " "      " "        " "        " "
## 4  ( 1 ) " "      " "     " "     " "      " "        " "        " "
## 5  ( 1 ) " "      " "     " "     "*"      " "        " "        " "
## 6  ( 1 ) " "      " "     " "     "*"      " "        " "        " "
## 7  ( 1 ) " "      " "     " "     "*"      "*"        " "        " "
## 8  ( 1 ) " "      " "     " "     "*"      "*"        " "        " "
```

```r
# Select the best subset based on the criterion described
res.sum <- summary(models)
data.frame(
  Adj.R2 = which.max(res.sum$adjr2),
  CP = which.min(res.sum$rss),
  BIC = which.min(res.sum$bic)
)
```

```
##   Adj.R2 CP BIC
## 1      8  8   8
```

That is best model: log(PRICEK) ~ DSALE + SQFTK + LN_LOTSIZE + CENTRALAIR + GARAGE + FIREPLACE + MICHLAKE + LAKE_RIVER

B

```
# Descriptive statistics for the best fit model
res.sum$rss[8]
```

```
## [1] 684.6657
```

```
res.sum$adjr2[8]
```

```
## [1] 0.2244777
```

```
res.sum$bic[8]
```

```
## [1] -1202.446
```

C

```
# Fit the linear regression model
m1 <- lm(log(PRICEK) ~ DSALE + SQFTK + LN_LOTSIZE + CENTRALAIR
         + GARAGE + FIREPLACE + MICHLAKE + LAKE_RIVER, data)
summary(m1)
```

```
##
## Call:
## lm(formula = log(PRICEK) ~ DSALE + SQFTK + LN_LOTSIZE + CENTRALAIR +
##     GARAGE + FIREPLACE + MICHLAKE + LAKE_RIVER, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8117 -0.0965  0.0367  0.1581  1.2229
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.419385   0.175752  25.146  < 2e-16 ***
## DSALE       -0.160365   0.024911  -6.437 1.33e-10 ***
## SQFTK        0.189483   0.010295  18.405  < 2e-16 ***
## LN_LOTSIZE   0.122369   0.021405   5.717 1.15e-08 ***
## CENTRALAIR   0.083921   0.012010   6.988 3.16e-12 ***
## GARAGE       0.026255   0.008220   3.194  0.00141 **
## FIREPLACE    0.079700   0.008719   9.141  < 2e-16 ***
## MICHLAKE     0.213054   0.035676   5.972 2.51e-09 ***
## LAKE_RIVER   0.123061   0.029632   4.153 3.34e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3704 on 4991 degrees of freedom
## Multiple R-squared:  0.2257, Adjusted R-squared:  0.2245
## F-statistic: 181.9 on 8 and 4991 DF,  p-value: < 2.2e-16
```

Regression equation:

$log(PRICEK) = 4.419385 - 0.160365 * DSALE + 0.189483 * SQFTK + 0.122369 * LN_LOTSIZE + 0.083921 * CENTRALAIR + 0.026255 * GARAGE + 0.079700 * FIREPLACE + 0.213054 * MICHLAKE + 0.123061 * LAKE_RIVER$

D

```
# Adding quadratic effect to SQFT
m2 <- lm(log(PRICEK) ~ DSALE + SQFTK^2 + LN_LOTSIZE + CENTRALAIR
         + GARAGE + FIREPLACE + MICHLAKE + LAKE_RIVER, data)
```

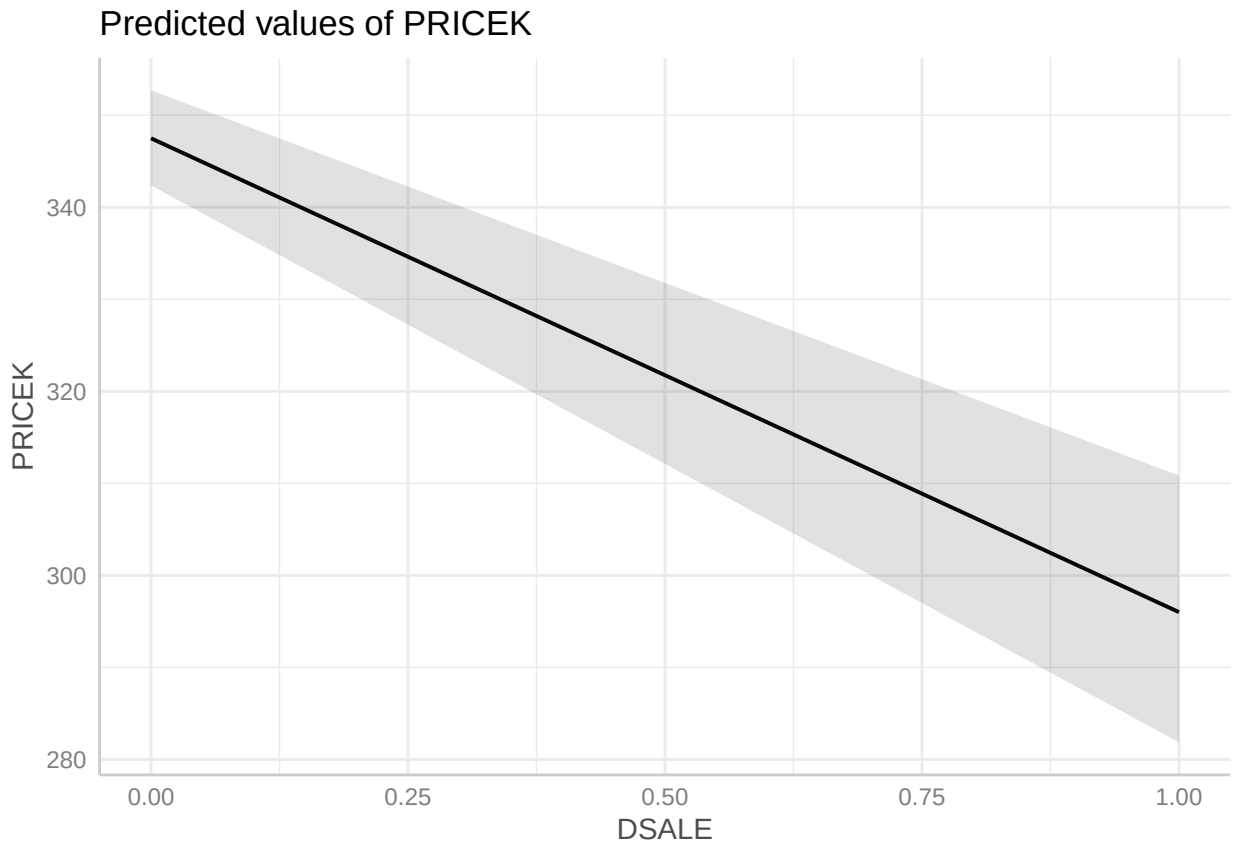
```
summary(margins(m2))
```

```
##      factor     AME     SE       z      p   lower   upper
##  CENTRALAIR  0.0839 0.0120  6.9878 0.0000  0.0604  0.1075
##       DSALE -0.1604 0.0249 -6.4375 0.0000 -0.2092 -0.1115
##   FIREPLACE  0.0797 0.0087  9.1411 0.0000  0.0626  0.0968
##      GARAGE  0.0263 0.0082  3.1939 0.0014  0.0101  0.0424
##  LAKE_RIVER  0.1231 0.0296  4.1530 0.0000  0.0650  0.1811
##  LN_LOTSIZE  0.1224 0.0214  5.7168 0.0000  0.0804  0.1643
##    MICHLAKE  0.2131 0.0357  5.9719 0.0000  0.1431  0.2830
##       SQFTK  0.1895 0.0103 18.4048 0.0000  0.1693  0.2097
```

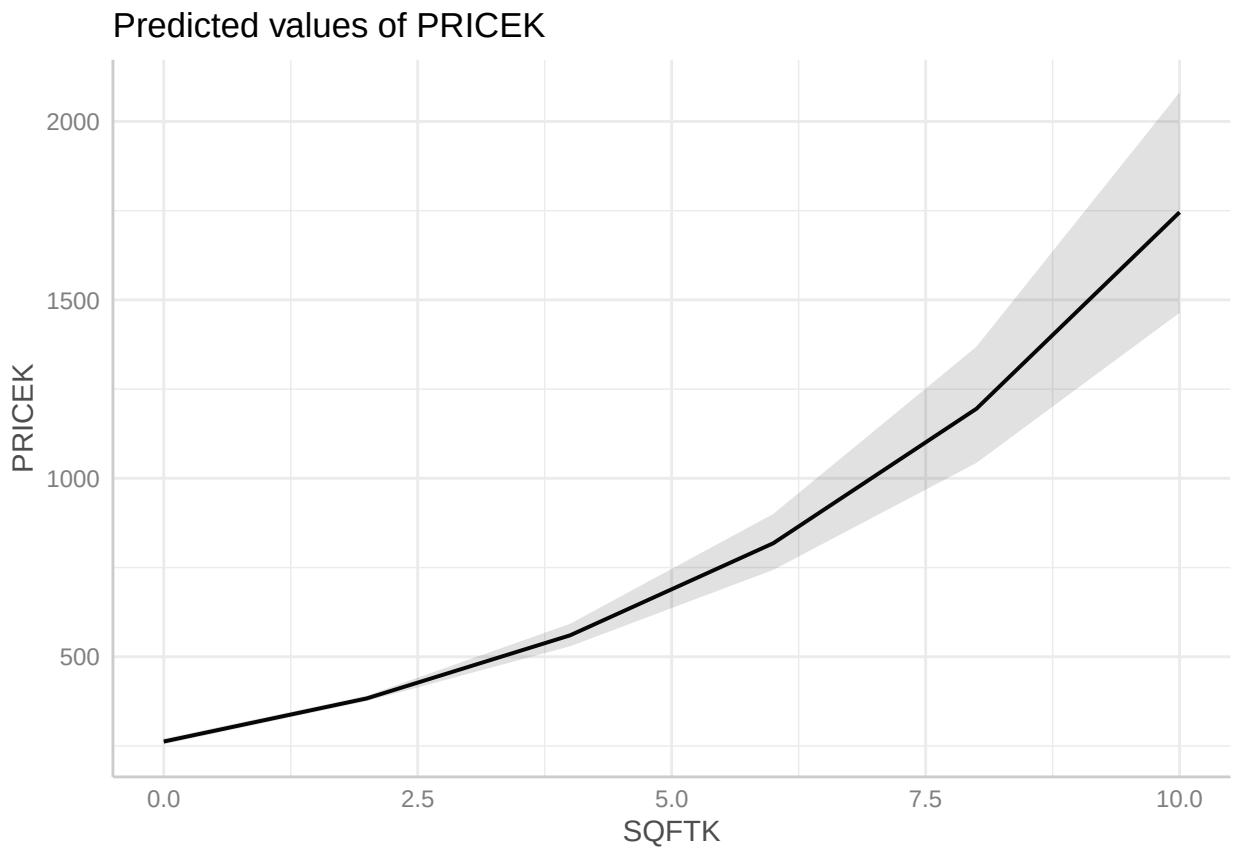
E

```
p <- ggpredict(m1)
plot(p)
```

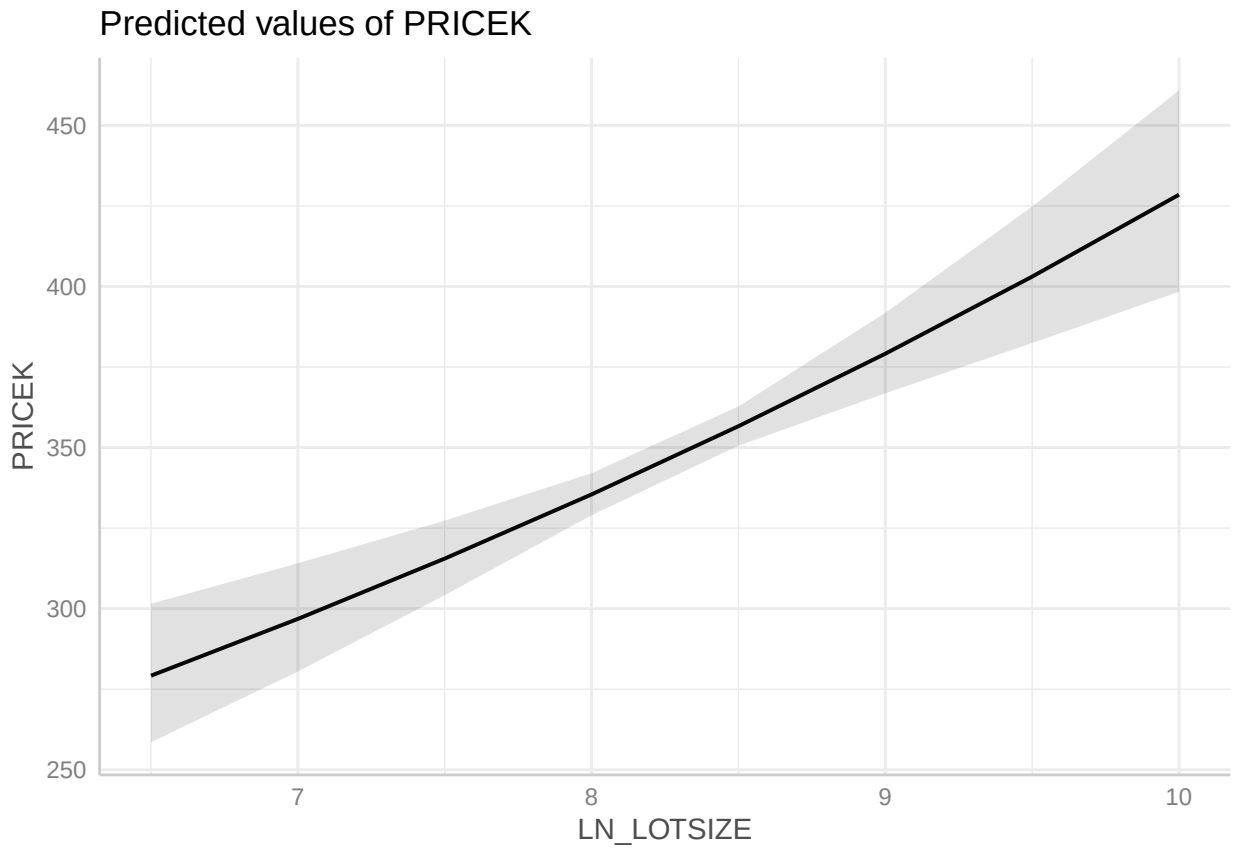
```
## $DSALE
```


Predicted values of PRICEK
PRICEK
340
320
300
280
0.00
0.25
0.50
0.75
1.00
DSALE

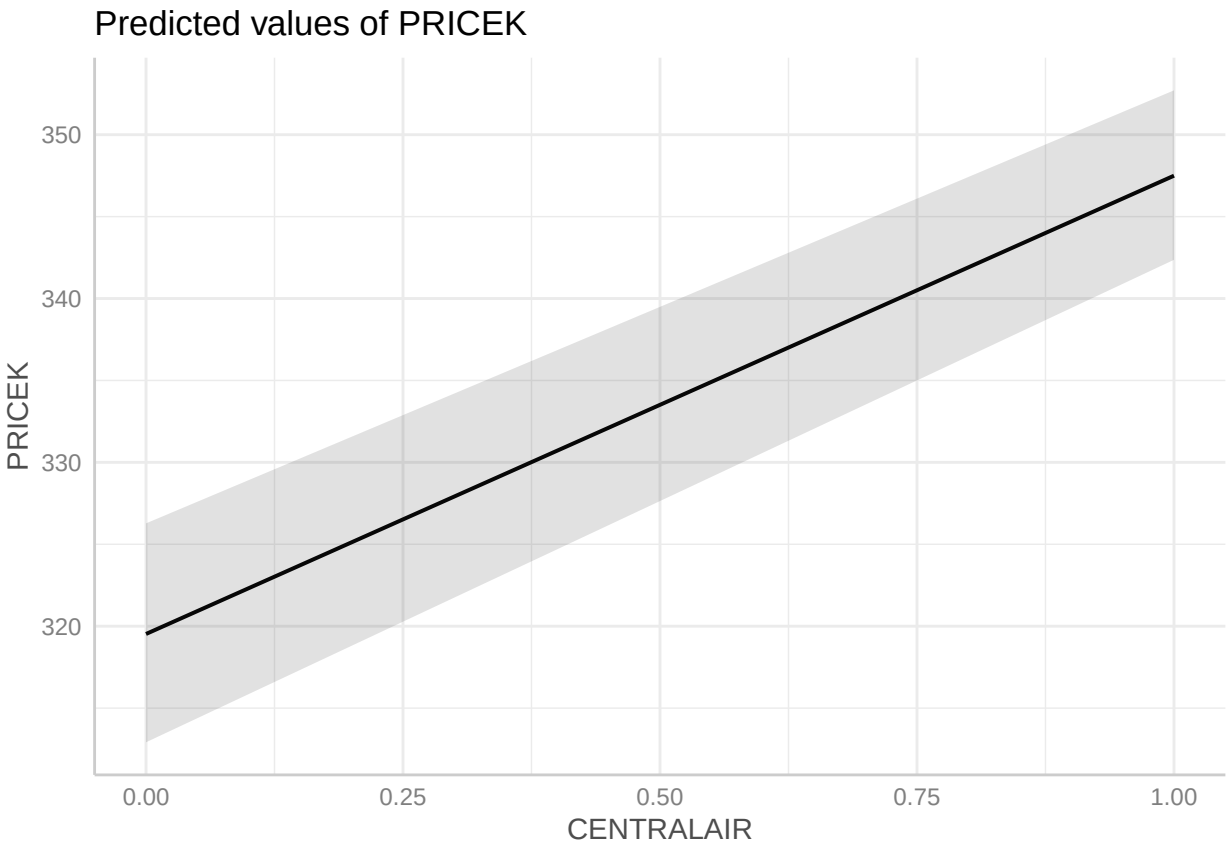
```
##
## $SQFTK
```


Predicted values of PRICEK
2000
1500
1000
500
PRICEK
0.0
2.5
5.0
7.5
10.0
SQFTK

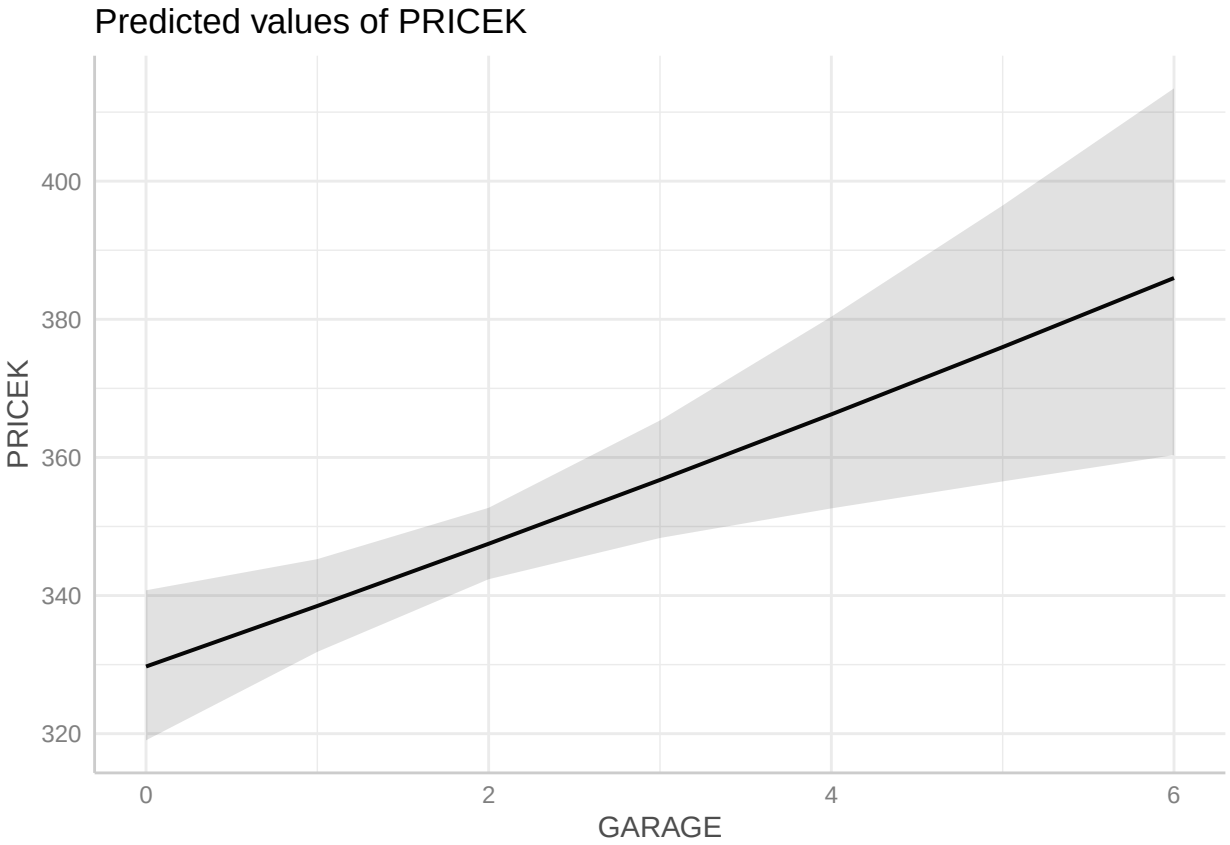
```
##
## $LN_LOTSIZE
```


Predicted values of PRICEK
PRICEK
450
400
350
300
250
7
8
9
10
LN_LOTSIZE

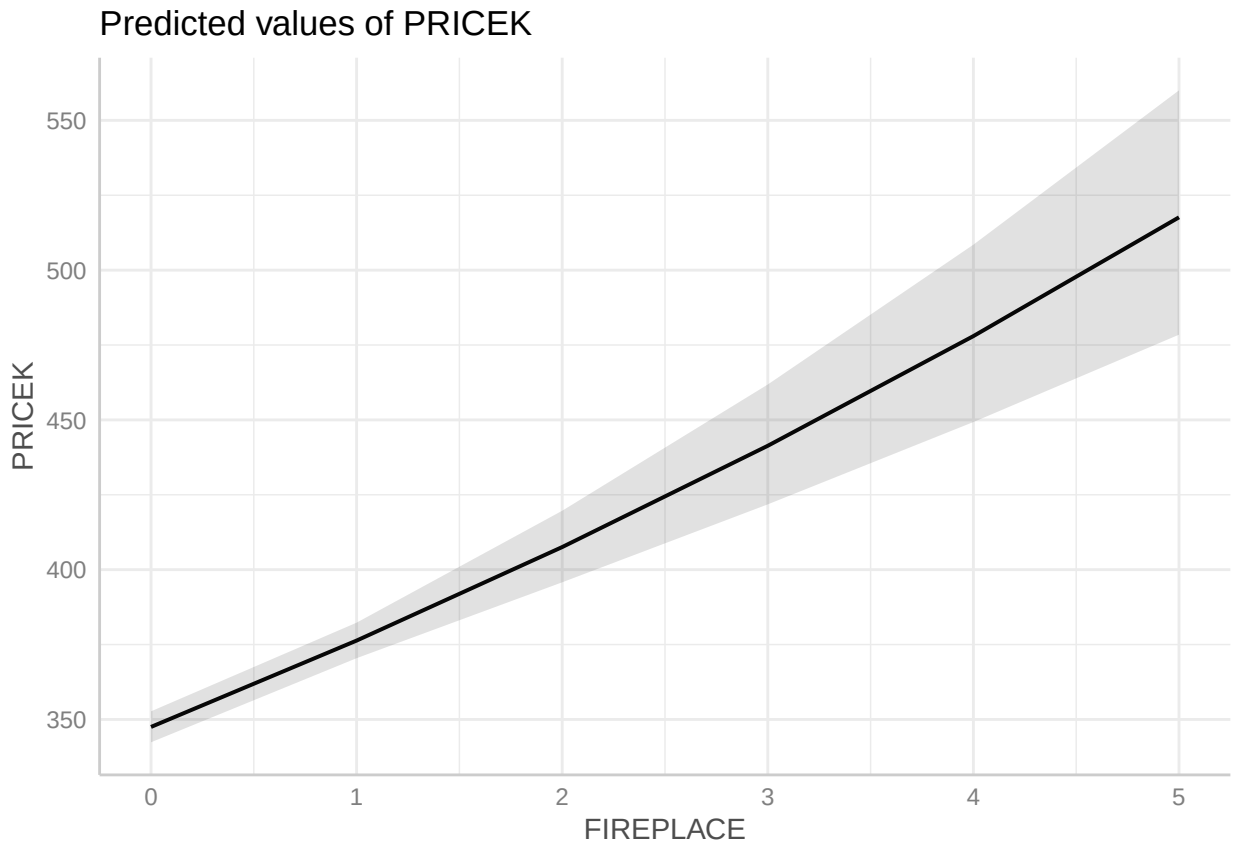
```
##
## $CENTRALAIR
```


Predicted values of PRICEK
PRICEK
350
340
330
320
0.00
0.25
0.50
0.75
1.00
CENTRALAIR

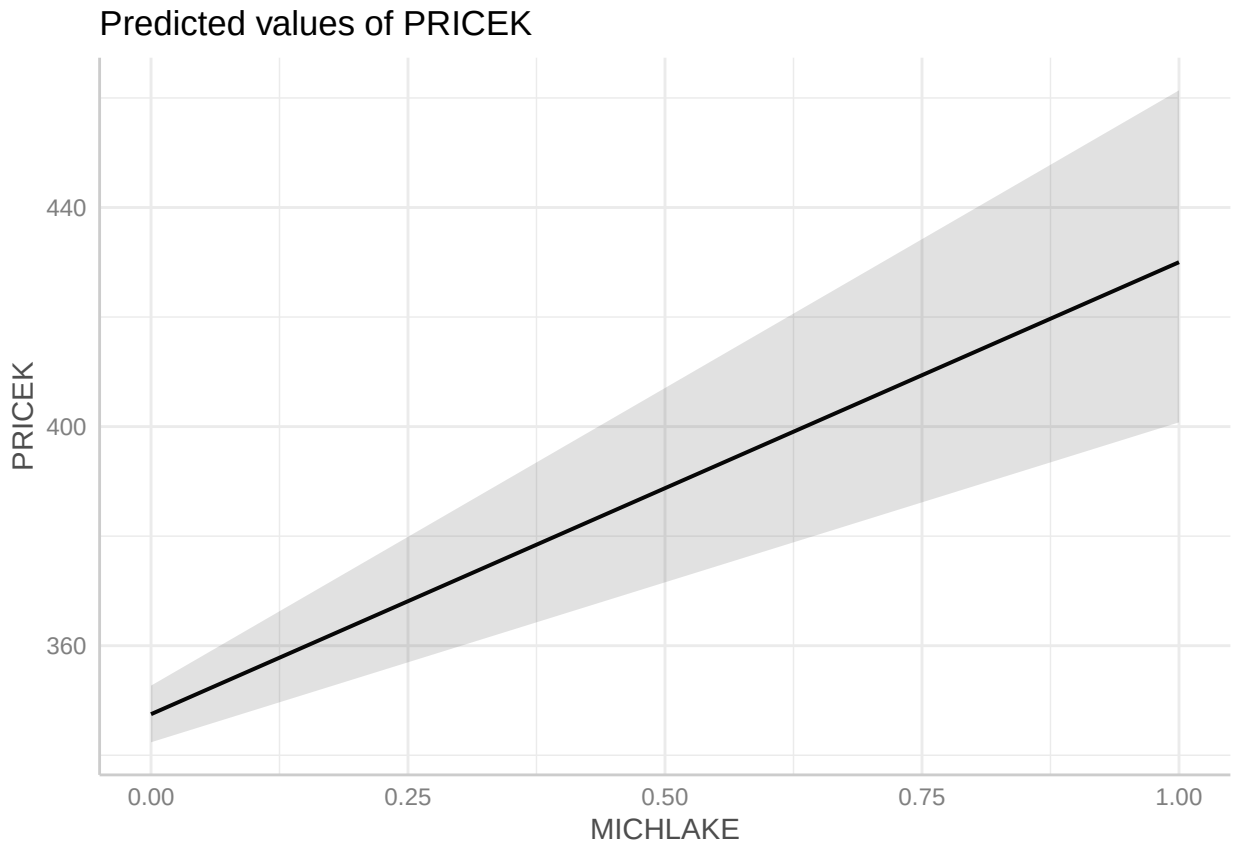
```
##
## $GARAGE
```


Predicted values of PRICEK
PRICEK
400
380
360
340
320
0
2
4
6
GARAGE

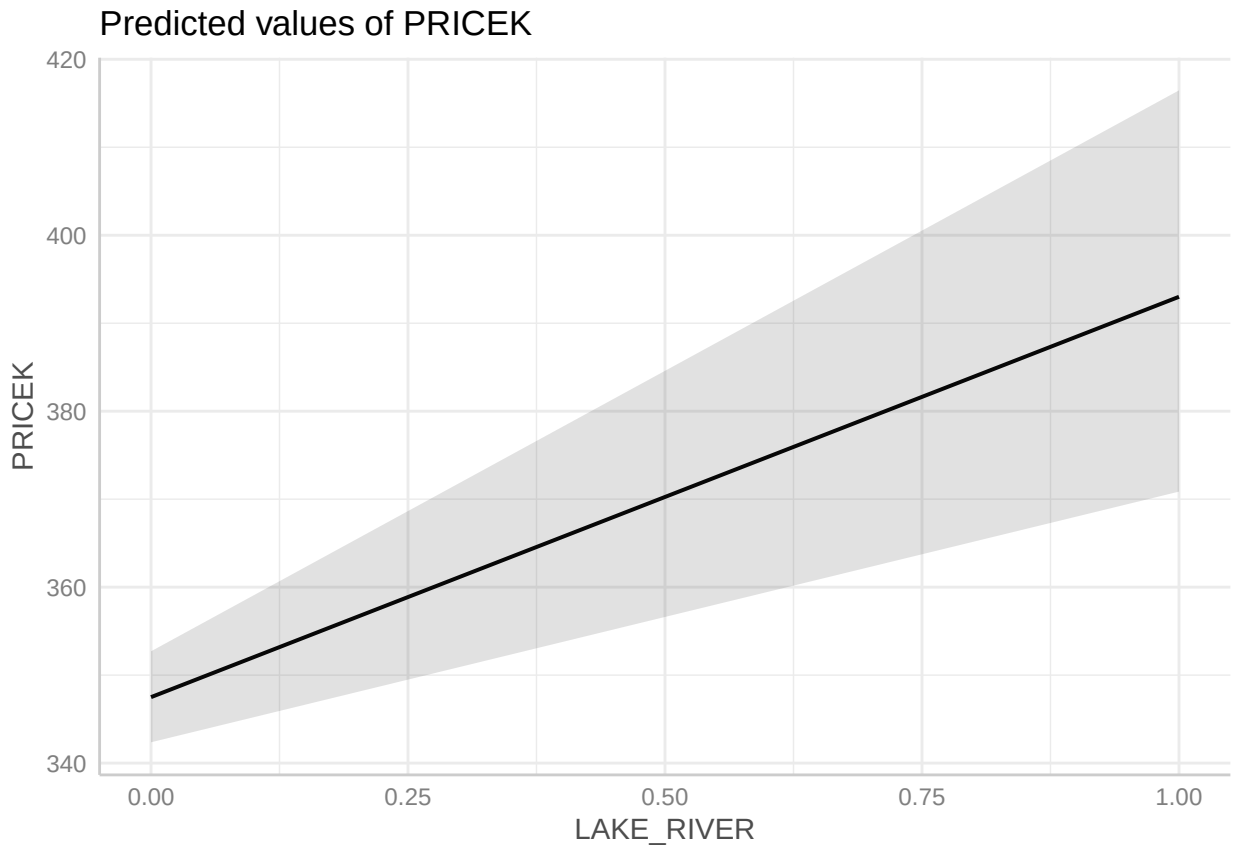
```
##
## $FIREPLACE
```


Predicted values of PRICEK
550
500
450
400
350
PRICEK
0
1
2
3
4
5
FIREPLACE

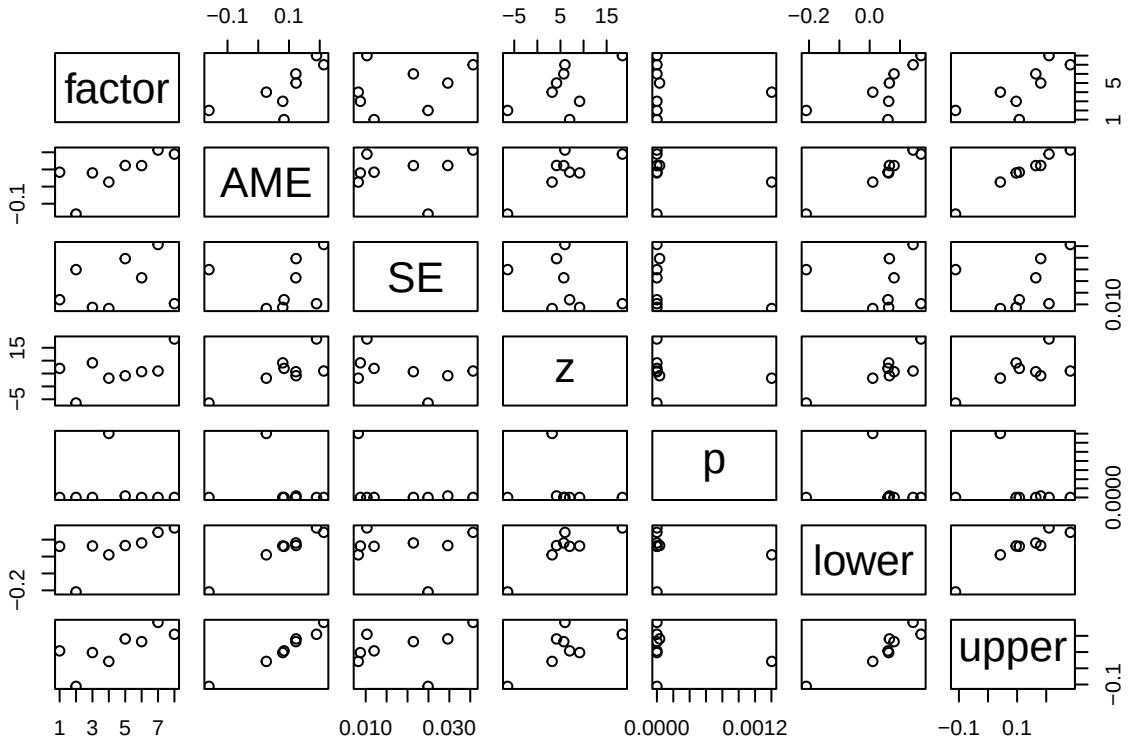
```
##
## $MICHLAKE
```


Predicted values of PRICEK
PRICEK
440
400
360
0.00
0.25
0.50
0.75
1.00
MICHLAKE

```
##
## $LAKE_RIVER
```




```
m <- margins_summary(m1)
plot(m)
```



F

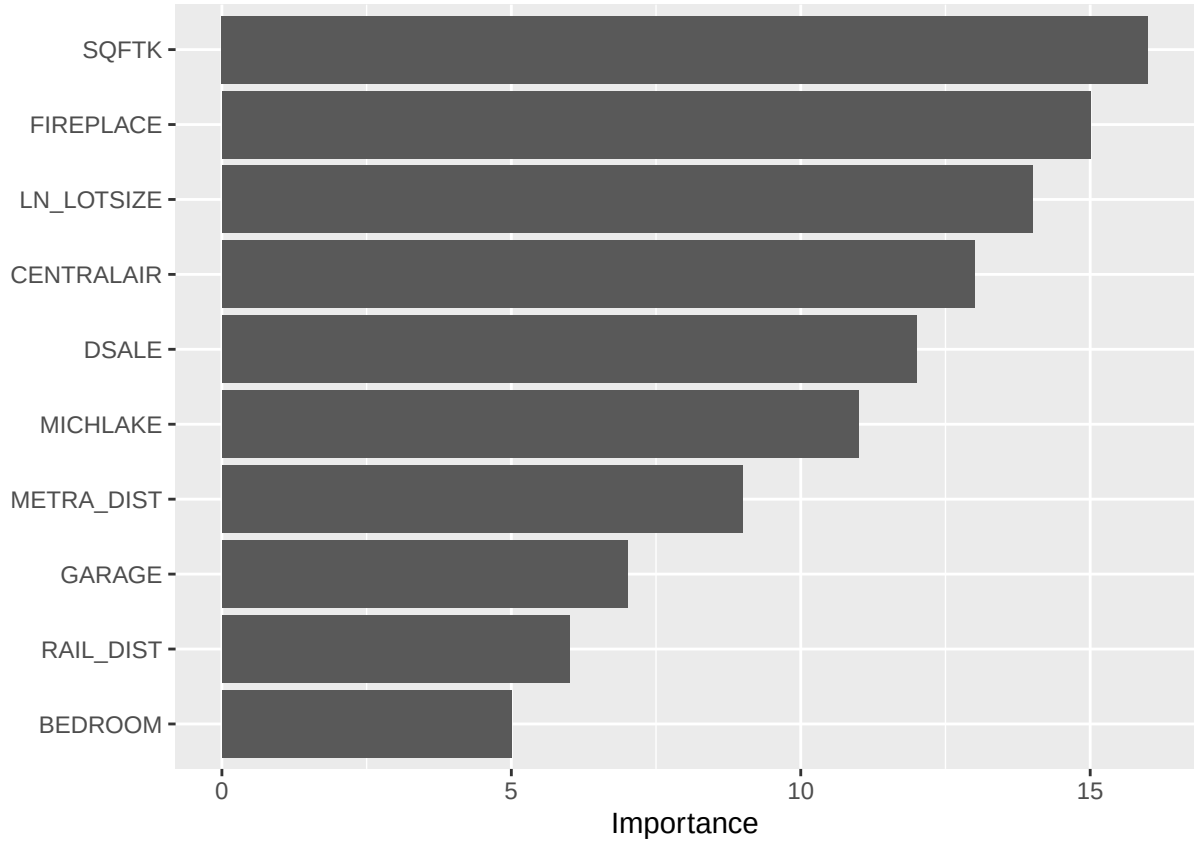
```
# Fit a basic MARS model
mars <- earth(
  log(PRICEK) ~ .,
  data = data)

# summary of the model
summary(mars)
```

```
## Call: earth(formula=log(PRICEK)~., data=data)
##
##                          coefficients
## (Intercept)                 6.5665584
## DSALE                      -0.1499946
## CENTRALAIR                  0.0719680
## MICHLAKE                    0.2014699
## LAKE_RIVER                  0.1187351
## h(3-BEDROOM)               -0.0662750
## h(3.252-SQFTK)             -0.2141365
## h(8.06306-LN_LOTSIZE)       0.4534098
## h(LN_LOTSIZE-8.06306)       0.2588962
## h(1-GARAGE)                -0.0950876
## h(3-FIREPLACE)             -0.0749909
```

```
## h(FIREPLACE-3)          -0.1908307
## h(METRA_DIST-1.21081)   -0.1239136
## h(2.67933-METRA_DIST)   -0.1250615
## h(METRA_DIST-2.67933)   -0.5196783
## h(0.03551-RAIL_DIST)     4.9166914
## h(RAIL_DIST-0.03551)    -0.0623355
##
## Selected 17 of 20 terms, and 11 of 15 predictors
## Termination condition: RSq changed by less than 0.001 at 20 terms
## Importance: SQFTK, FIREPLACE, LN_LOTSIZE, CENTRALAIR, DSALE, MICHLAKE, ...
## Number of terms at each degree of interaction: 1 16 (additive model)
## GCV 0.1319194    RSS 650.9189    GRSq 0.2543673    RSq 0.2638827
```

G

```
# Plot important variables
vip(mars)
```



The identified important variables are not identical to the identified variables previously because MARS find out non-linear relation and linear regression does not do that.