Lab5

Suixin Jiang

9/29/2019

```
estate <- read_csv('./estate.csv')
```

```
## Parsed with column specification:
## cols(
##   Price = col_double(),
##   Area = col_double(),
##   Bed = col_double(),
##   Bath = col_double(),
##   AC = col_double(),
##   Garage = col_double(),
##   Pool = col_double(),
##   Year = col_double(),
##   Quality = col_character(),
##   Style = col_double(),
##   Lot = col_double(),
##   Highway = col_double()
## )
```

```
head(estate)
```

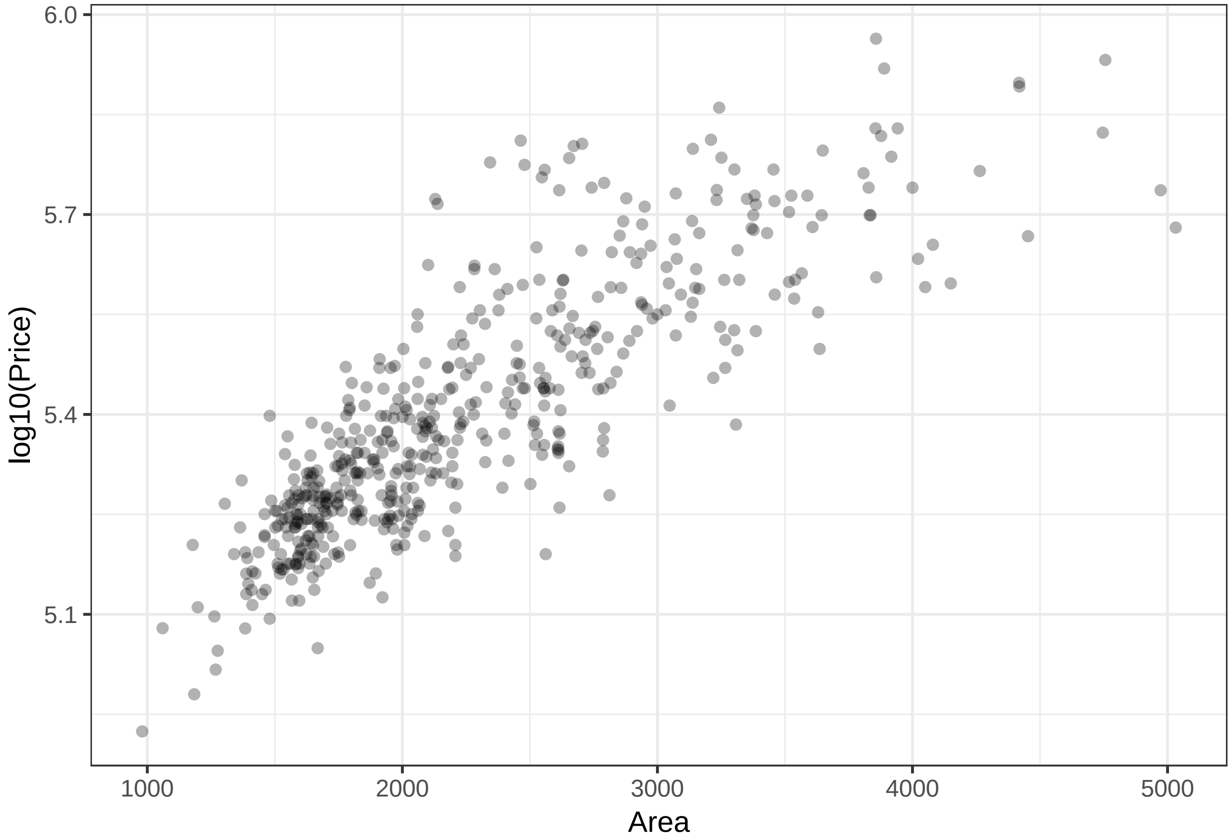
```
## # A tibble: 6 x 12
##    Price  Area   Bed  Bath    AC Garage  Pool  Year Quality Style   Lot
##    <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <chr>   <dbl> <dbl>
## 1 360000  3032     4     4     1      2     0  1972 Medium      1 22221
## 2 340000  2058     4     2     1      2     0  1976 Medium      1 22912
## 3 250000  1780     4     3     1      2     0  1980 Medium      1 21345
## 4 205500  1638     4     2     1      2     0  1963 Medium      1 17342
## 5 275500  2196     4     3     1      2     0  1968 Medium      7 21786
## 6 248000  1966     4     3     1      5     1  1972 Medium      1 18902
## # ... with 1 more variable: Highway <dbl>
```

Obviously, some variables need to be recoded as dichotomous variables.

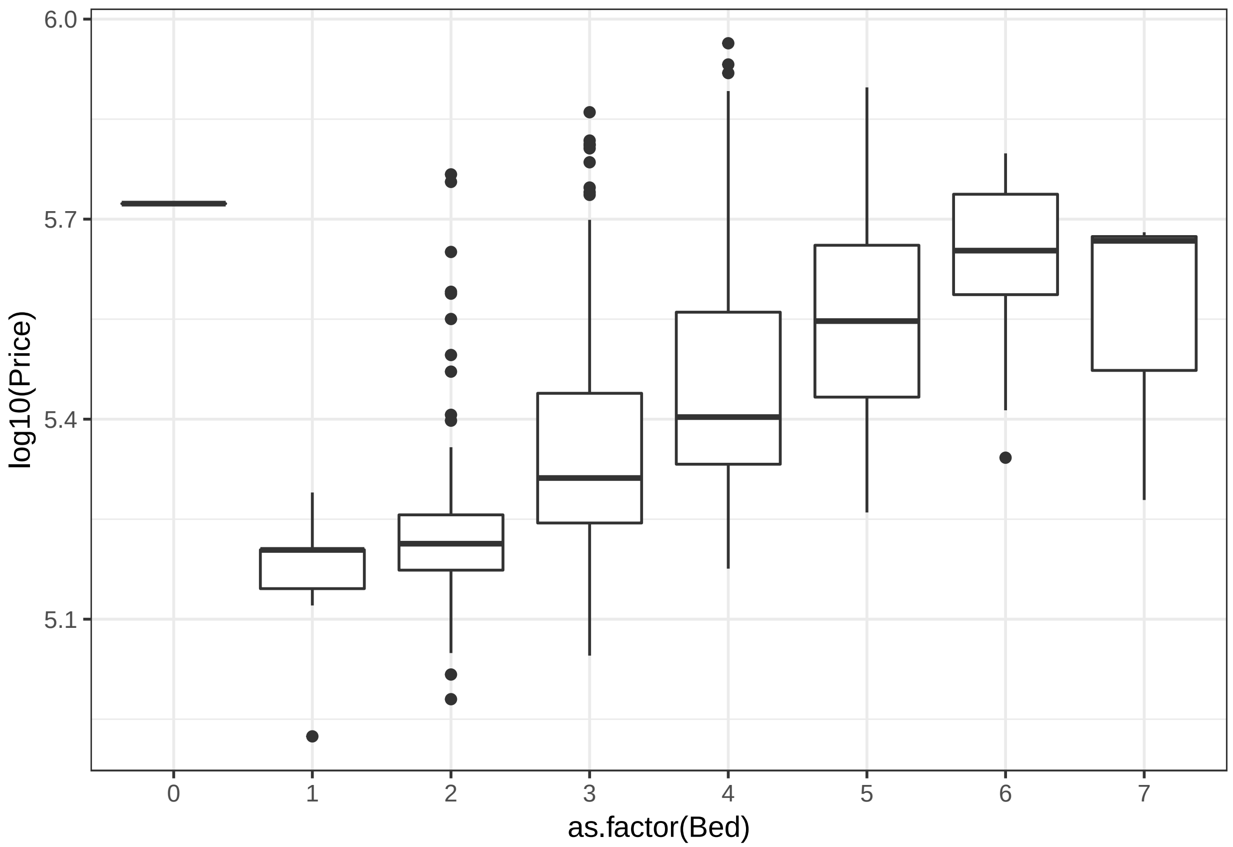
```
estate %>%
  mutate(AC = recode(AC, '1' = 'Yes', '0' = 'No'),
         Pool = recode(Pool, '1' = 'Yes', '0' = 'No'),
         Highway = recode(Highway, '1' = 'Yes', '0' = 'No'),
         Style = as.factor(Style)) ->
  estate
```

The analysis' purpose is trying to identify which variables may influence the price. So we first take a look at the correlation between price and other variables.

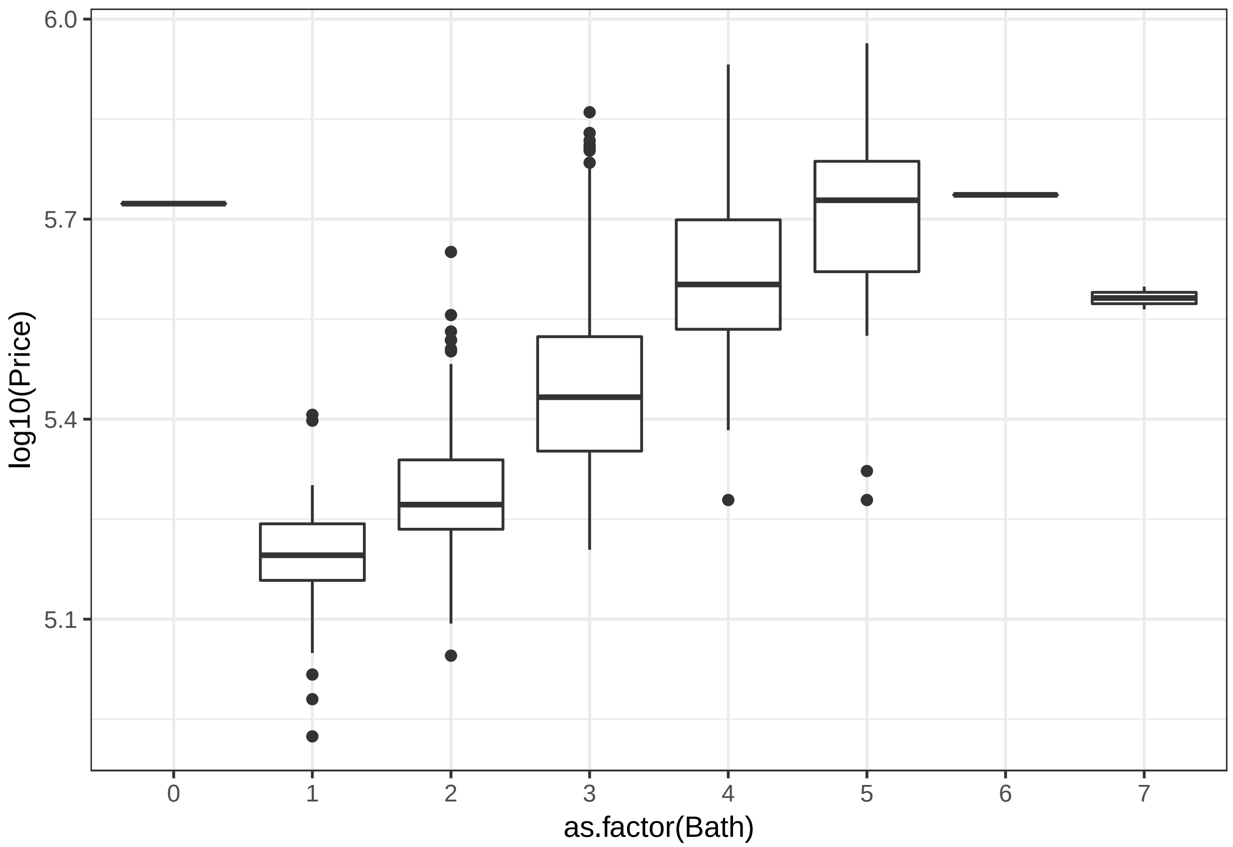
```
ggplot(estate, aes(x = Area, y = log10(Price))) +
  geom_point(alpha = 0.3)
```



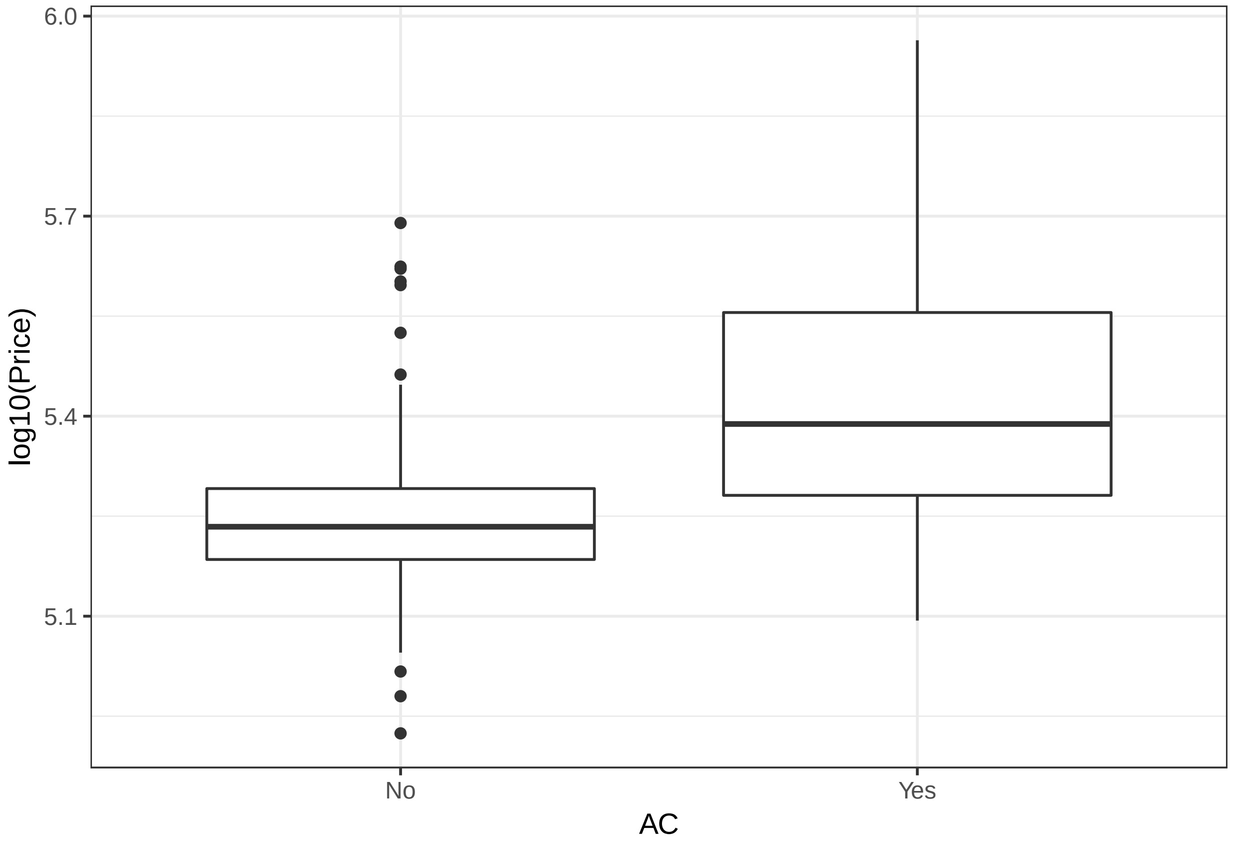
```
ggplot(estate, aes(x = as.factor(Bed), y = log10(Price))) +
  geom_boxplot()
```



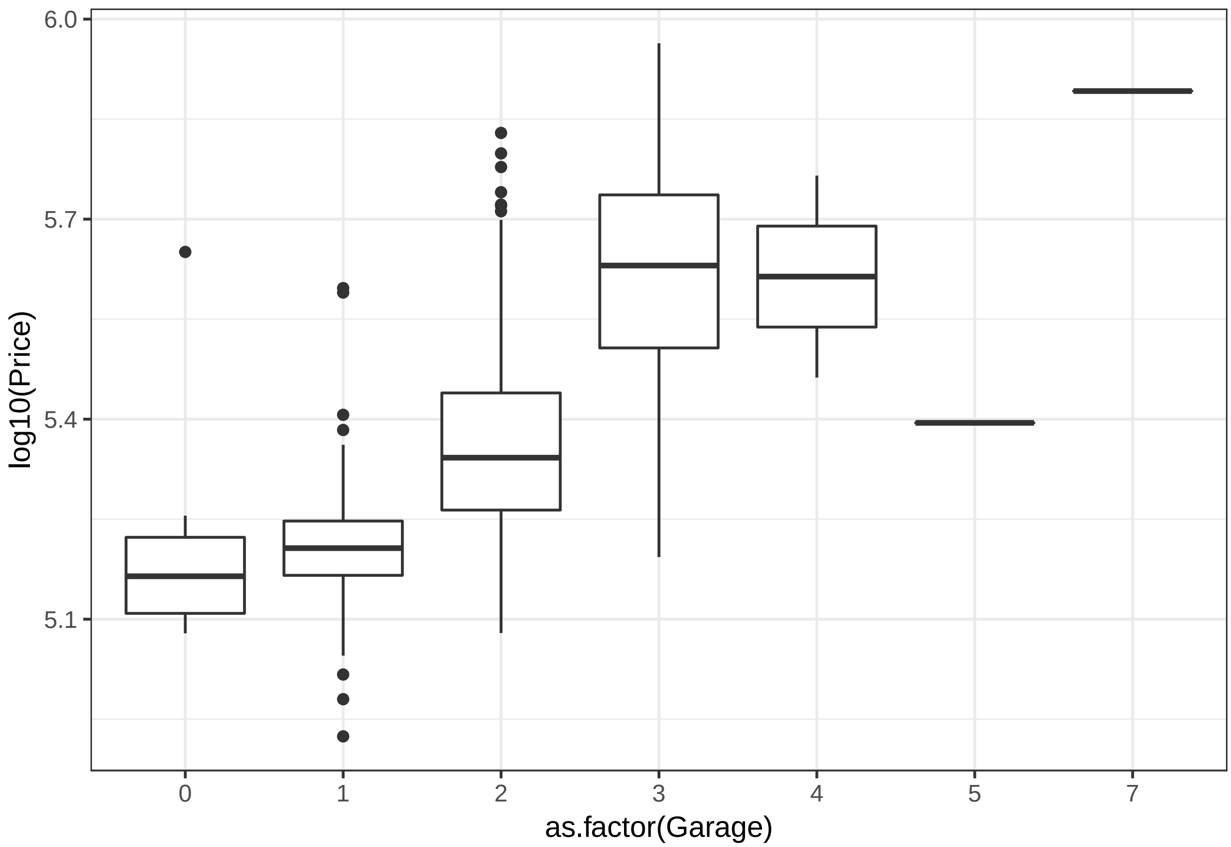
```
ggplot(estate, aes(x = as.factor(Bath), y = log10(Price))) +
  geom_boxplot()
```



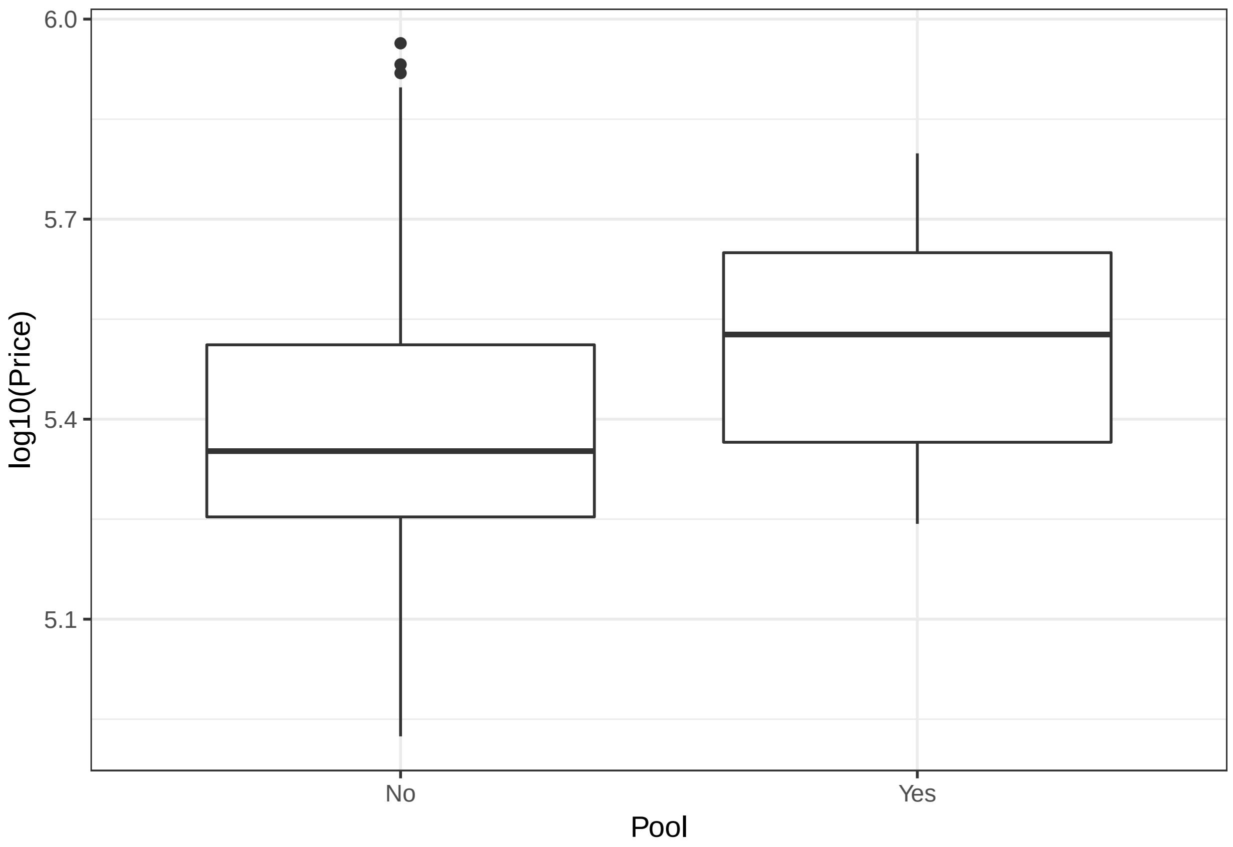
```
ggplot(estate, aes(x = AC, y = log10(Price))) +
  geom_boxplot()
```



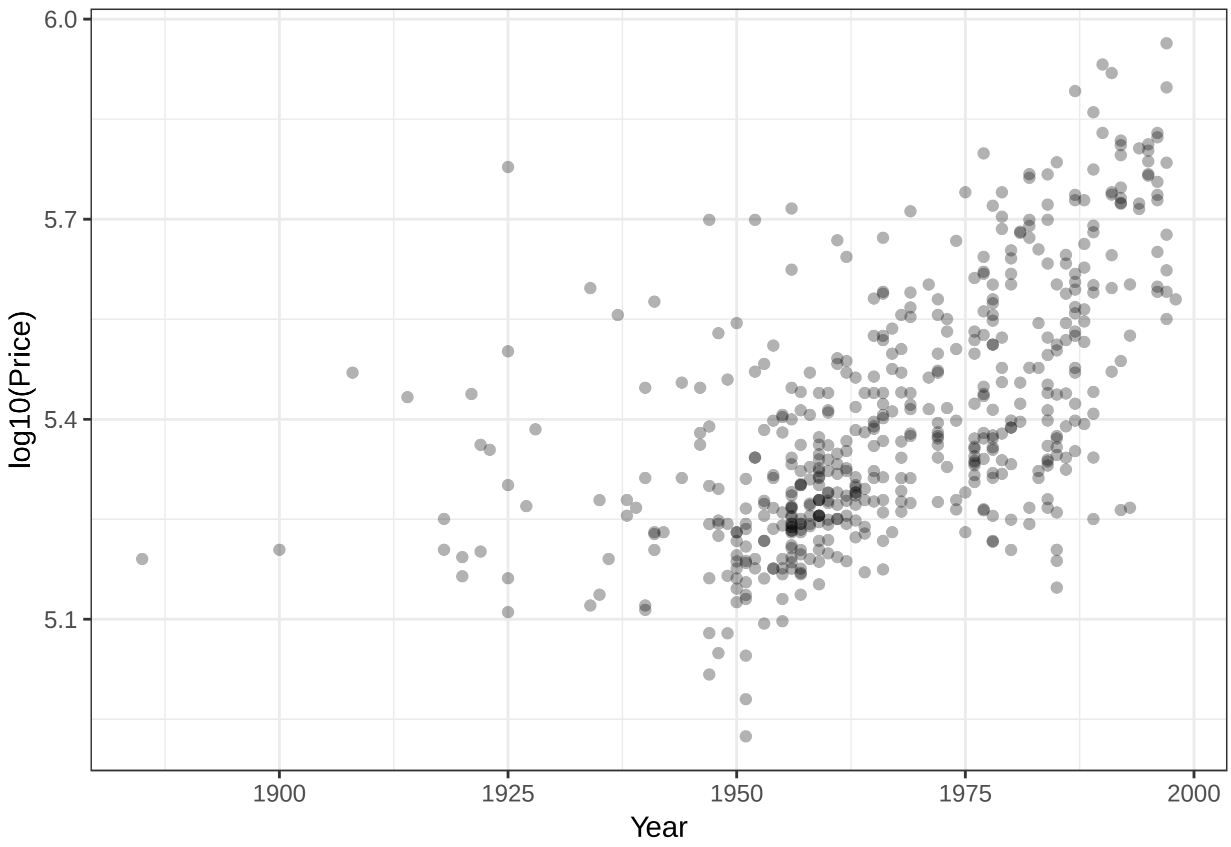
```
ggplot(estate, aes(x = as.factor(Garage), y = log10(Price))) +
  geom_boxplot()
```



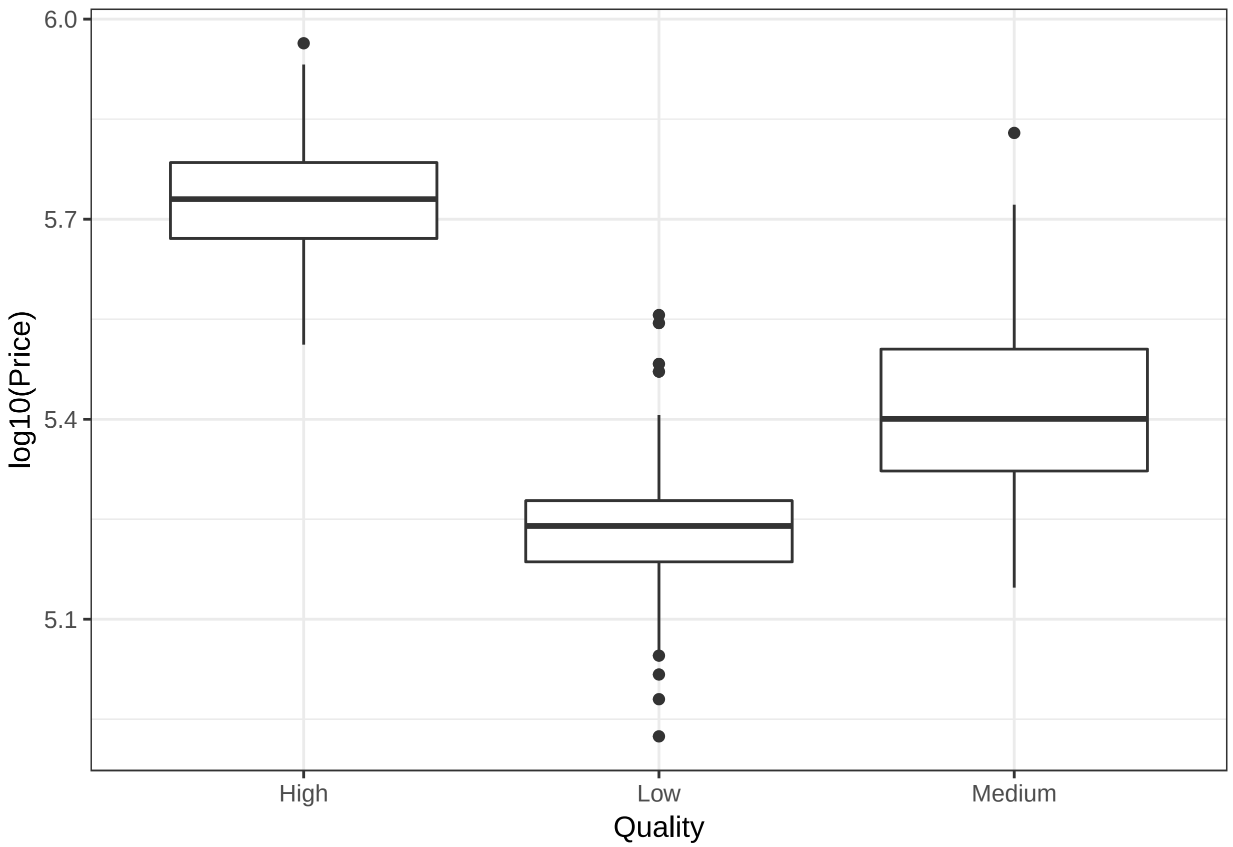
```
ggplot(estate, aes(x = Pool, y = log10(Price))) +
  geom_boxplot()
```



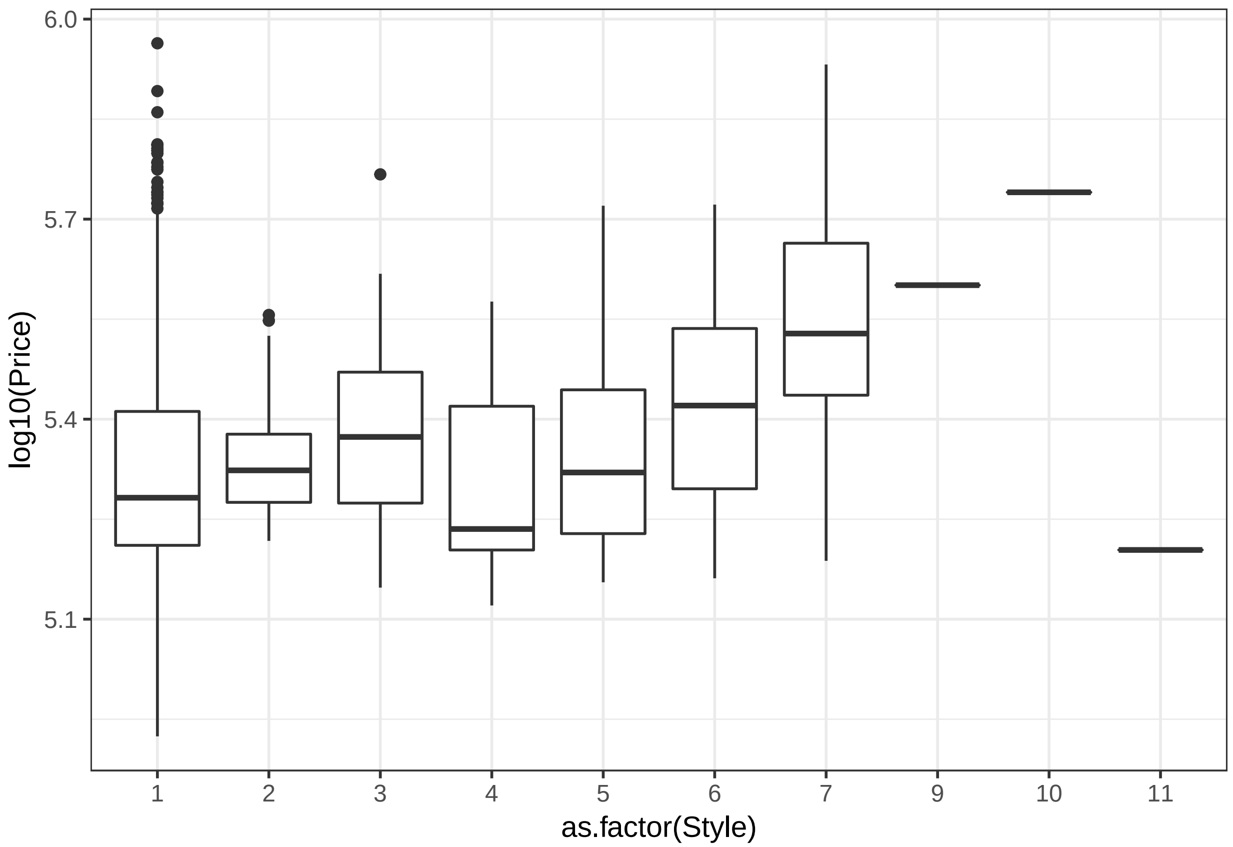
```r
ggplot(estate, aes(x = Year, y = log10(Price))) +
  geom_point(alpha = 0.3)
```



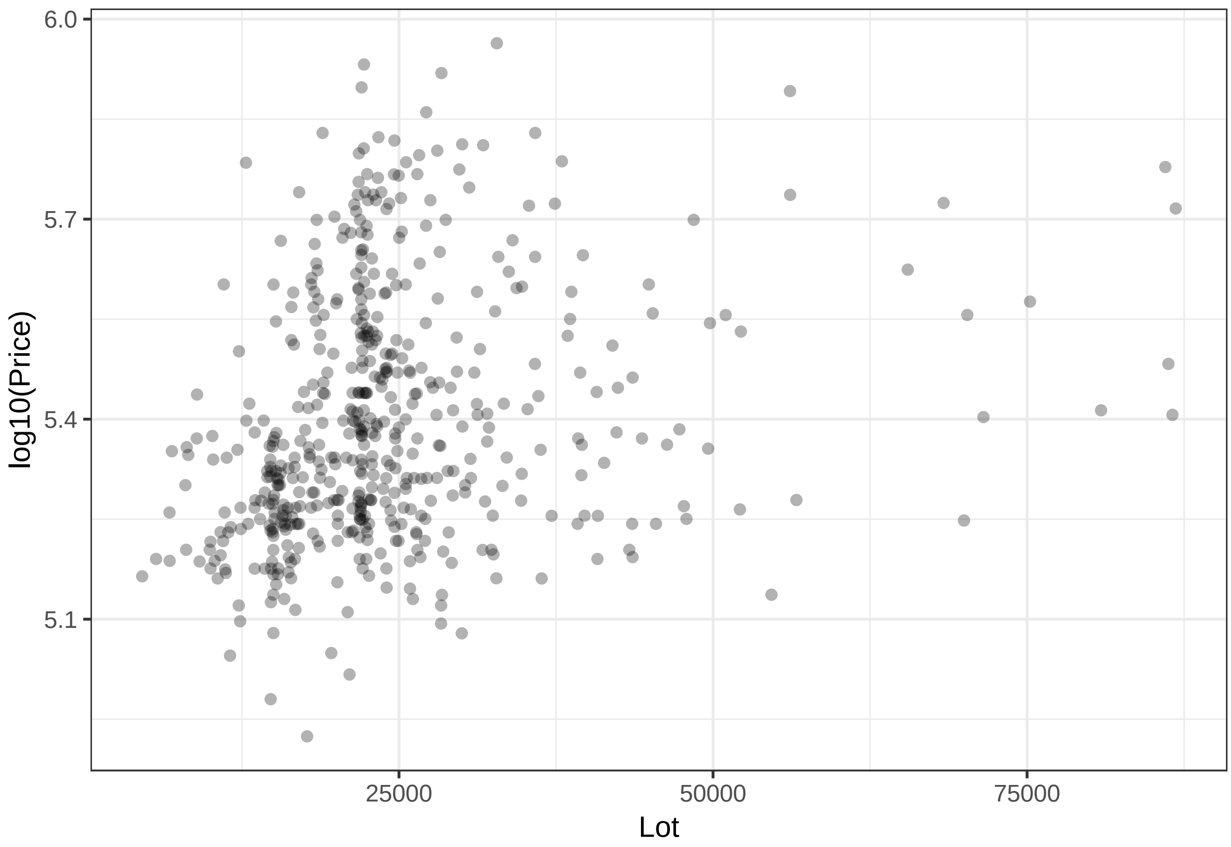
```
ggplot(estate, aes(x = Quality, y = log10(Price))) +
  geom_boxplot()
```



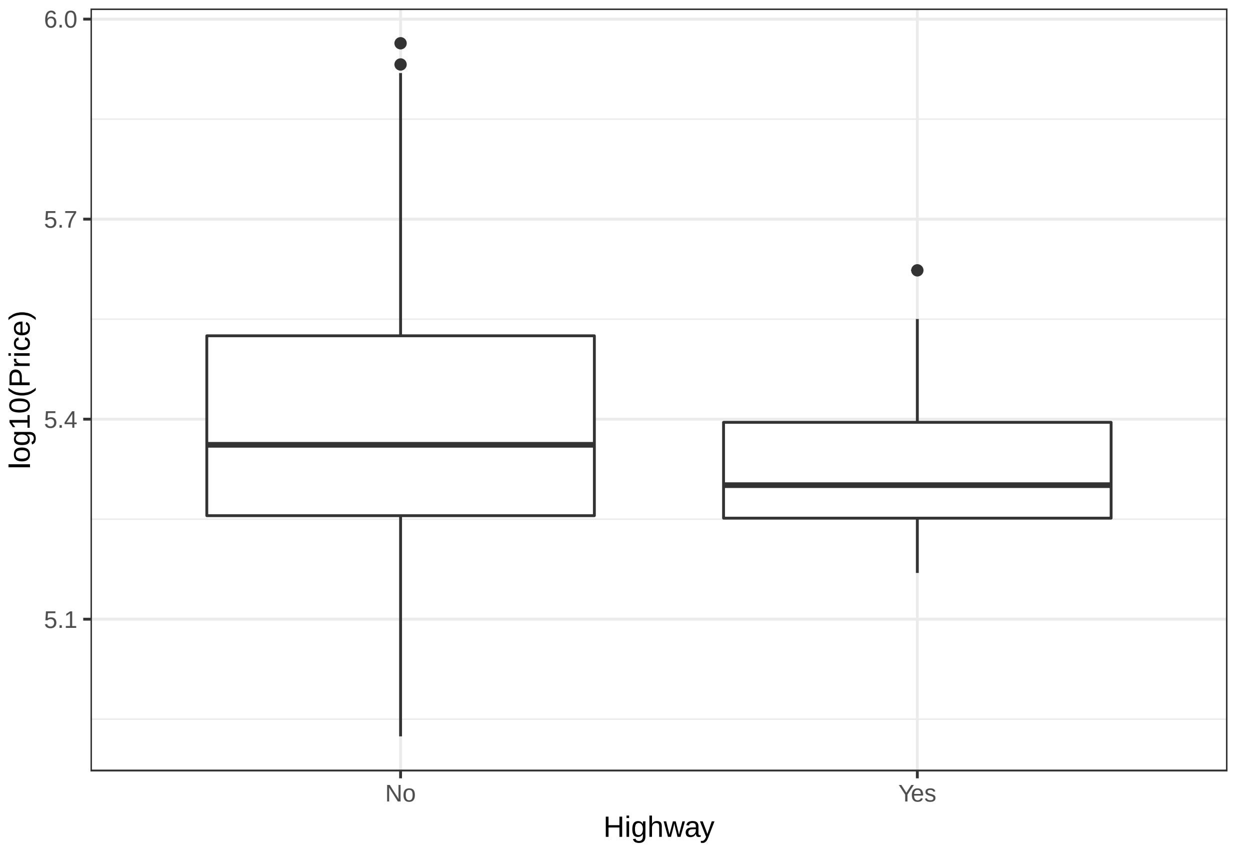
```
ggplot(estate, aes(x = as.factor(Style), y = log10(Price))) +
  geom_boxplot()
```



```r
ggplot(estate, aes(x = Lot, y = log10(Price))) +
  geom_point(alpha = 0.3)
```



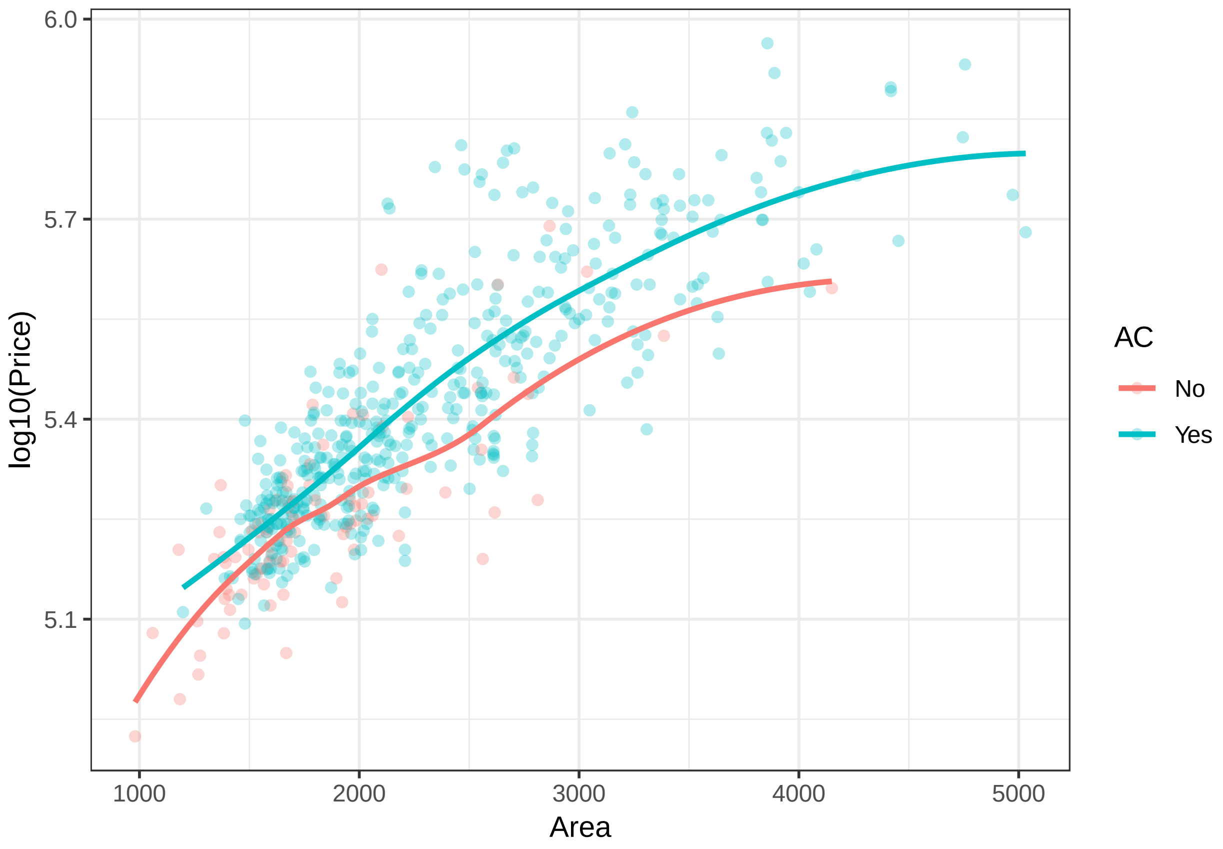
```
ggplot(estate, aes(x = Highway, y = log10(Price))) +
  geom_boxplot()
```



As the plots are shown above, we can conclude that 'Area' has a strong positive correlation with price, the bigger house usually has a higher price; the same to 'Year', especially for a house that built after the 1950s, the newer the house, the more expensive it is. For variables 'AC' and 'Pool', if the house has these facilities it usually has a higher price. House near to highway usually has a lower price. For variables 'Bed', 'Bath', and 'Garage', overall, the more rooms, the more expensive the house is. However, we have outliers in each type of house, which suggests that we have to take account of many of the variables'influence together.

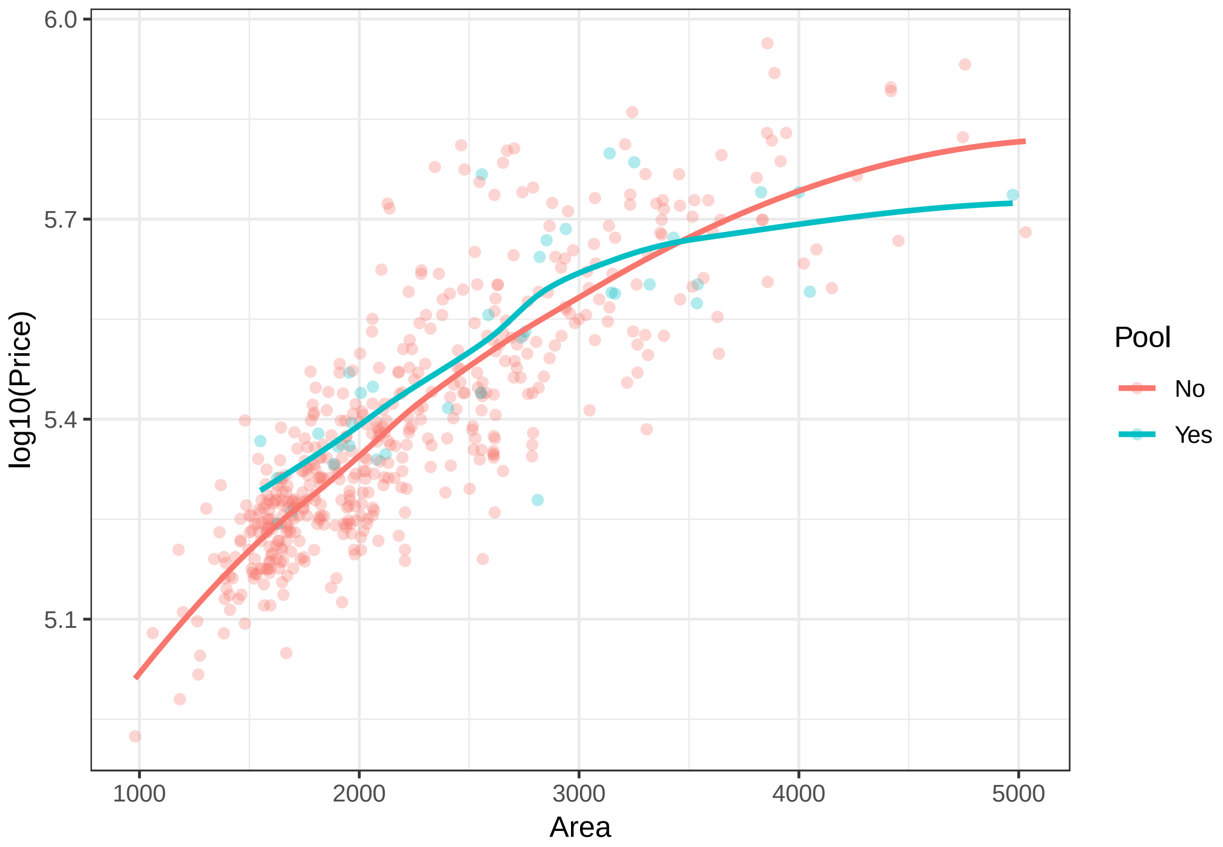
```r
ggplot(estate, aes(x = Area, y = log10(Price), color = AC)) +
  geom_point(alpha = 0.3) +
  geom_smooth(se = F)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



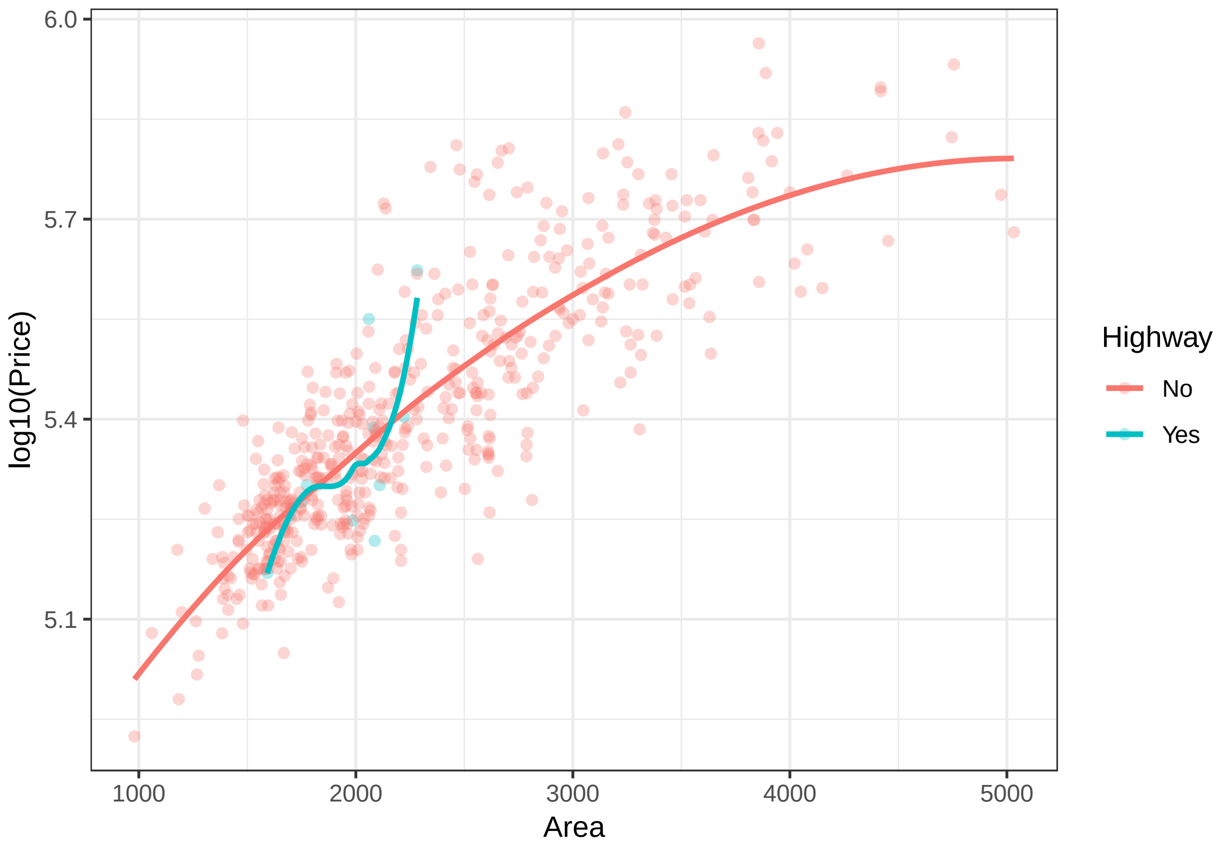
```
ggplot(estate, aes(x = Area, y = log10(Price), color = Pool)) +
  geom_point(alpha = 0.3) +
  geom_smooth(se = F)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



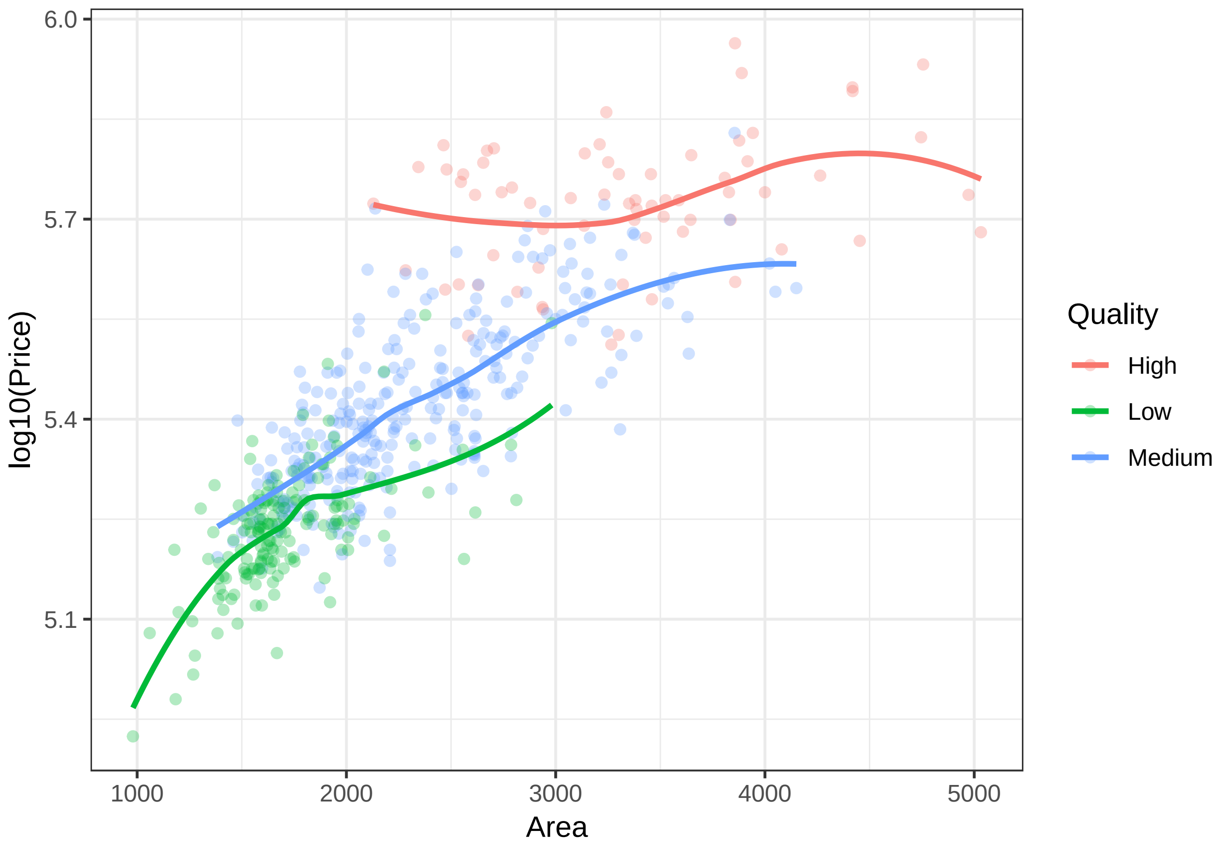
```
ggplot(estate, aes(x = Area, y = log10(Price), color = Highway)) +
  geom_point(alpha = 0.3) +
  geom_smooth(se = F)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



```
ggplot(estate, aes(x = Area, y = log10(Price), color = Quality)) +
  geom_point(alpha = 0.3) +
  geom_smooth(se = F)
```

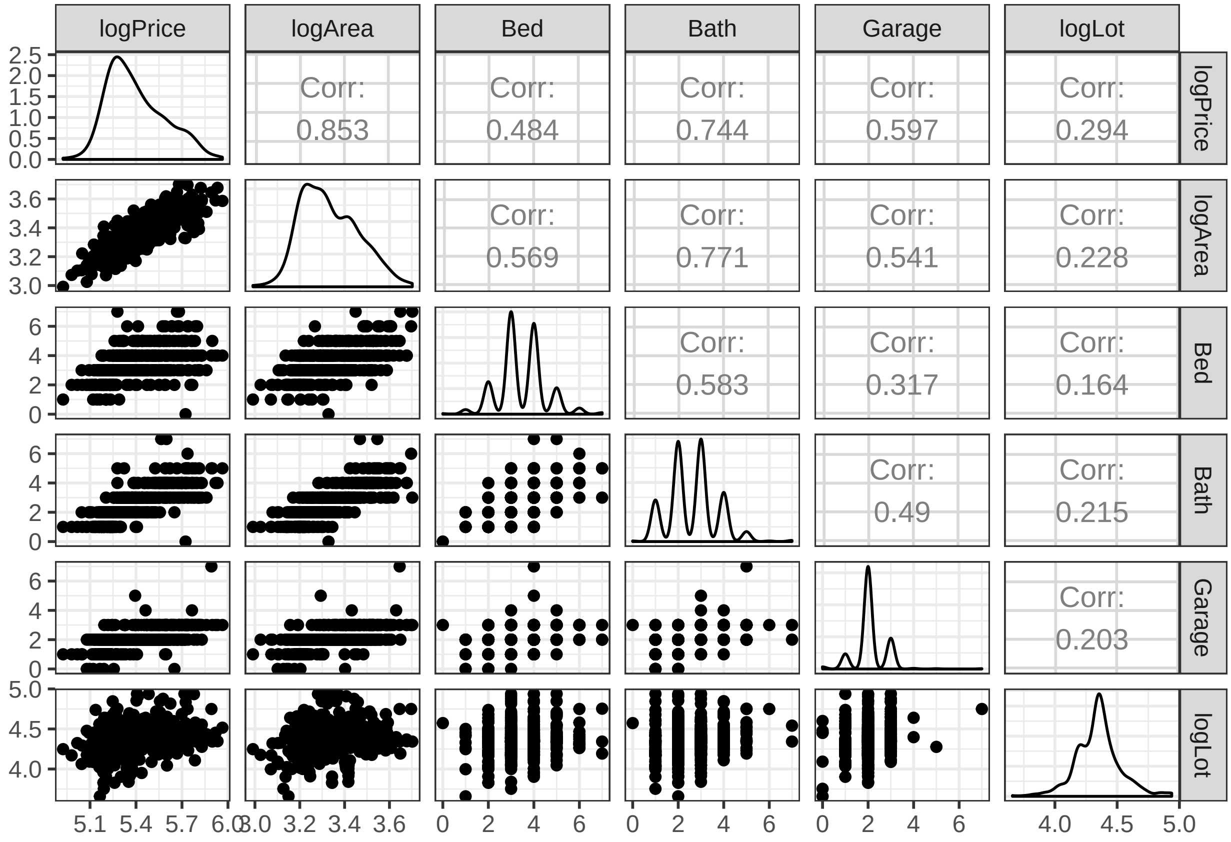
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Obviously, if we explain price with more variables the trend becomes very clear.

Let's try to find more correlated variables by ggpairs() function.

```
estate %>%
  mutate(logPrice = log10(Price), logArea = log10(Area), logLot = log10(Lot)) ->
  estate
estate %>%
  select(logPrice, logArea, Bed, Bath, Garage, logLot) %>%
  ggpairs()
```



Let’s try to bulid a regression model.

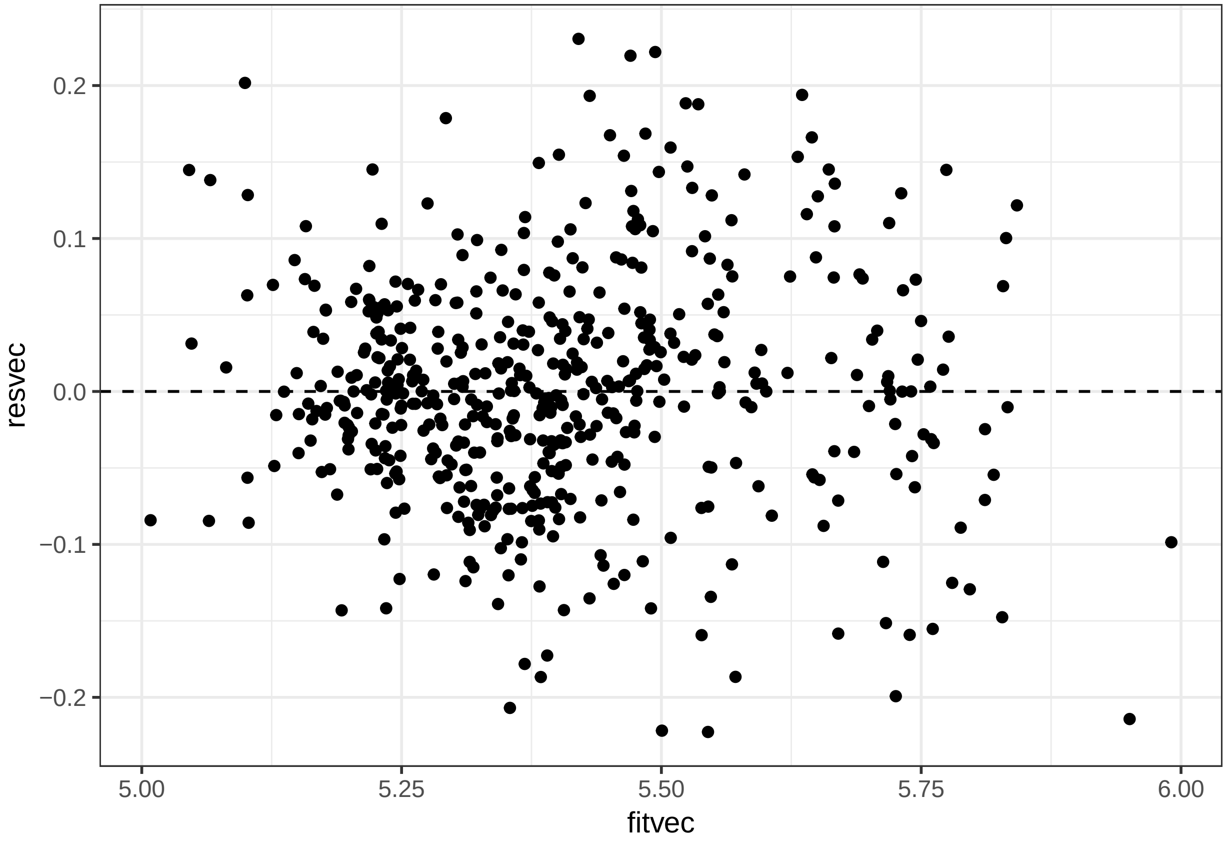
```
reg <- lm(logPrice ~ logArea + Bed + Bath + AC + Garage + Pool + Quality +
            Style + logLot + Highway, data = estate)
summary(reg)
```

```
##
## Call:
## lm(formula = logPrice ~ logArea + Bed + Bath + AC + Garage +
##     Pool + Quality + Style + logLot + Highway, data = estate)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.223907 -0.046437 -0.001313  0.040143  0.229096
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.409e+00  1.981e-01  12.164  < 2e-16 ***
## logArea        7.783e-01  5.847e-02  13.311  < 2e-16 ***
## Bed           -6.003e-05  4.351e-03  -0.014 0.988998
## Bath           2.223e-02  5.598e-03   3.972 8.18e-05 ***
## ACYes          2.477e-02  1.047e-02   2.366 0.018348 *
## Garage         2.007e-02  6.554e-03   3.062 0.002320 **
## PoolYes        1.912e-02  1.379e-02   1.387 0.166152
## QualityLow    -1.907e-01  1.818e-02 -10.486  < 2e-16 ***
## QualityMedium -1.455e-01  1.320e-02 -11.024  < 2e-16 ***
## Style2        -2.180e-02  1.222e-02  -1.785 0.074927 .
## Style3         1.335e-03  1.155e-02   0.116 0.908008
## Style4         1.045e-02  2.402e-02   0.435 0.663497
## Style5        -4.792e-02  1.977e-02  -2.424 0.015710 *
## Style6        -2.000e-02  2.000e-02  -1.000 0.317634
## Style7        -4.347e-02  1.151e-02  -3.775 0.000179 ***
## Style9        -3.642e-02  7.756e-02  -0.469 0.638929
## Style10       -1.203e-01  7.893e-02  -1.524 0.128045
## Style11       -1.398e-01  7.748e-02  -1.804 0.071846 .
## logLot         9.903e-02  2.030e-02   4.879 1.44e-06 ***
## HighwayYes    -2.777e-02  2.361e-02  -1.176 0.240095
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07658 on 502 degrees of freedom
## Multiple R-squared:  0.8392, Adjusted R-squared:  0.8331
## F-statistic: 137.9 on 19 and 502 DF,  p-value: < 2.2e-16
```

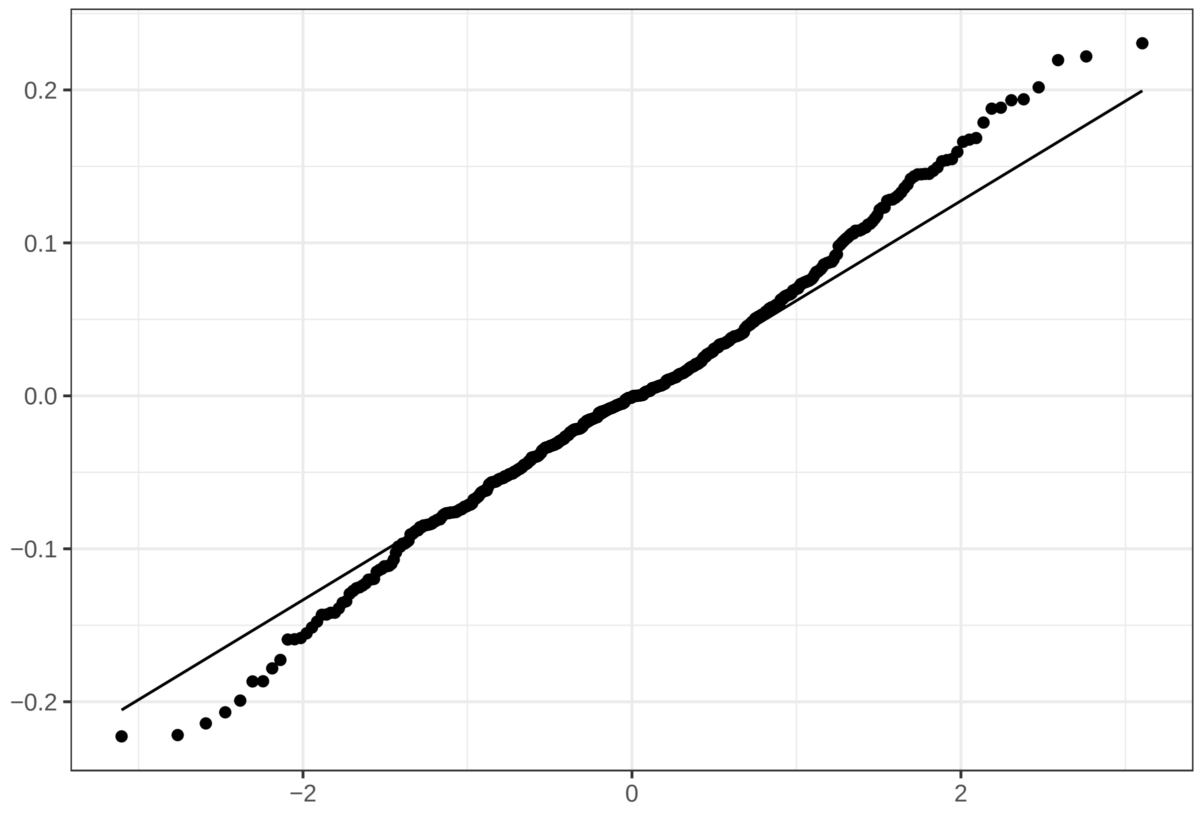
```
sout <- step(reg)
```

```
## Start:  AIC=-2662.81
## logPrice ~ logArea + Bed + Bath + AC + Garage + Pool + Quality +
##     Style + logLot + Highway
##
##           Df Sum of Sq    RSS     AIC
## - Bed      1   0.00000 2.9443 -2664.8
## - Highway  1   0.00811 2.9524 -2663.4
## - Pool     1   0.01128 2.9556 -2662.8
## <none>                 2.9443 -2662.8
## - AC       1   0.03284 2.9771 -2659.0
## - Garage   1   0.05497 2.9993 -2655.2
## - Style    9   0.15000 3.0943 -2654.9
## - Bath     1   0.09252 3.0368 -2648.7
## - logLot   1   0.13960 3.0839 -2640.6
## - Quality  2   0.75472 3.6990 -2547.7
## - logArea  1   1.03915 3.9834 -2507.0
##
## Step:  AIC=-2664.81
## logPrice ~ logArea + Bath + AC + Garage + Pool + Quality + Style +
##     logLot + Highway
##
##           Df Sum of Sq    RSS     AIC
## - Highway  1   0.00812 2.9524 -2665.4
## - Pool     1   0.01128 2.9556 -2664.8
## <none>                 2.9443 -2664.8
## - AC       1   0.03297 2.9773 -2661.0
## - Garage   1   0.05497 2.9993 -2657.2
## - Style    9   0.15012 3.0944 -2656.8
## - Bath     1   0.09990 3.0442 -2649.4
## - logLot   1   0.13981 3.0841 -2642.6
## - Quality  2   0.77598 3.7203 -2546.7
## - logArea  1   1.09967 4.0440 -2501.2
##
## Step:  AIC=-2665.37
## logPrice ~ logArea + Bath + AC + Garage + Pool + Quality + Style +
##     logLot
##
##           Df Sum of Sq    RSS     AIC
## <none>                 2.9524 -2665.4
## - Pool     1   0.01184 2.9642 -2665.3
## - AC       1   0.03418 2.9866 -2661.4
## - Style    9   0.14569 3.0981 -2658.2
## - Garage   1   0.05430 3.0067 -2657.9
## - Bath     1   0.10122 3.0536 -2649.8
## - logLot   1   0.13771 3.0901 -2643.6
## - Quality  2   0.77735 3.7298 -2547.4
## - logArea  1   1.09660 4.0490 -2502.5
```

```
resvec <- resid(sout)
fitvec <- fitted(sout)
qplot(fitvec, resvec) +
  geom_hline(yintercept = 0, linetype = 'dashed')
```



```
qplot(sample = resvec, geom = 'qq') +
  geom_qq_line()
```



```
coefvec <- coef(sout)
confintmat <- confint(sout)
sumlm <- summary(sout)
sumlm
```

```
##
## Call:
## lm(formula = logPrice ~ logArea + Bath + AC + Garage + Pool +
##     Quality + Style + logLot, data = estate)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.222622 -0.046972 -0.000831  0.041026  0.230511
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.415814   0.193440  12.489  < 2e-16 ***
## logArea        0.776868   0.056780  13.682  < 2e-16 ***
## Bath           0.022352   0.005377   4.157 3.79e-05 ***
## ACYes          0.025197   0.010432   2.415 0.016074 *
## Garage         0.019939   0.006549   3.045 0.002452 **
## PoolYes        0.019576   0.013771   1.421 0.155798
## QualityLow    -0.190565   0.017919 -10.635  < 2e-16 ***
## QualityMedium -0.145807   0.013039 -11.183  < 2e-16 ***
## Style2        -0.021149   0.012088  -1.750 0.080802 .
## Style3         0.002118   0.011504   0.184 0.854040
## Style4         0.011633   0.023971   0.485 0.627681
## Style5        -0.046558   0.019677  -2.366 0.018355 *
## Style6        -0.018677   0.019944  -0.936 0.349498
## Style7        -0.042167   0.011450  -3.683 0.000256 ***
## Style9        -0.035241   0.077497  -0.455 0.649497
## Style10       -0.119662   0.078662  -1.521 0.128831
## Style11       -0.138271   0.076972  -1.796 0.073032 .
## logLot         0.098210   0.020256   4.848 1.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07654 on 504 degrees of freedom
## Multiple R-squared:  0.8388, Adjusted R-squared:  0.8333
## F-statistic: 154.2 on 17 and 504 DF,  p-value: < 2.2e-16
```

From the summary we could make some conclusions. For example, house with one more bathroom tend to cost 10^0.022352=1.052 more, house with low-quality tend to cost 10^0.190565=1.55 less.