# Housing Price Prediction

Presenters:
Zhehong Lim, Yangchang Liang, Jiaqi He

Dataset: https://www.kaggle.com/datasets/gauravbr/multiple-linear-regression-housing-price-detection

# Content

- Data Cleaning
- Exploratory Data Analysis
- Modeling
- Diagnostics

# Data Cleaning

# Raw Data

Source:
https://www.kaggle.com/datasets/gauravbr/multiple-linear-regression-housing-price-detection

| price <int> | area <int> | bedrooms <int> | bathrooms <int> | stories <int> | mainroad <chr> | guestroom <chr> | basement <chr> | hotwaterheating <chr> | airconditioning <chr> | parking <int> | prefarea <chr> | furnishingstatus <chr> |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 13300000 | 7420 | 4 | 2 | 3 | yes | no | no | no | yes | 2 | yes | furnished |
| 12250000 | 8960 | 4 | 4 | 4 | yes | no | no | no | yes | 3 | no | furnished |
| 12250000 | 9960 | 3 | 2 | 2 | yes | no | yes | no | no | 2 | yes | semi-furnished |
| 12215000 | 7500 | 4 | 2 | 2 | yes | no | yes | no | yes | 3 | yes | furnished |
| 11410000 | 7420 | 4 | 1 | 2 | yes | yes | yes | no | yes | 2 | no | furnished |
| 10850000 | 7500 | 3 | 3 | 1 | yes | no | yes | no | yes | 2 | yes | semi-furnished |
| 10150000 | 8580 | 4 | 3 | 4 | yes | no | no | no | yes | 2 | yes | semi-furnished |
| 10150000 | 16200 | 5 | 3 | 2 | yes | no | no | no | no | 0 | no | unfurnished |

Binary Variable

Categorical Variable

# Dummy Variable

| | furnished | semi-furnished | unfurnished |
|---|---|---|---|
| furnistatfur | 1 | 0 | 0 |
| furnistatsemi | 0 | 1 | 0 |
| furnistatunfur | 0 | 0 | 1 |

# Cleaning Process

```r
```{r}
library(dplyr)
housing = read.csv("Housing.csv")
n = dim(housing)[1]
cleanhousing <- housing %>%
  select_if(is.numeric)
cleanhousing2 <- cleanhousing %>%
  mutate(
    mainroad = ifelse(housing$mainroad == "yes", 1, 0),
    guestroom = ifelse(housing$guestroom == "yes", 1, 0),
    basement = ifelse(housing$basement == "yes", 1, 0),
    hotwaterheating = ifelse(housing$hotwaterheating == "yes", 1, 0),
    airconditioning = ifelse(housing$airconditioning == "yes", 1, 0),
    prefarea = ifelse(housing$prefarea == "yes", 1, 0),
    furnistatfur = ifelse(housing$furnishingstatus == "furnished",1,0),
    furnistatunfur = ifelse(housing$furnishingstatus == "unfurnished",1,0),
    furnistatsemi = ifelse(housing$furnishingstatus == "semi-furnished",1,0)
        )
```
```

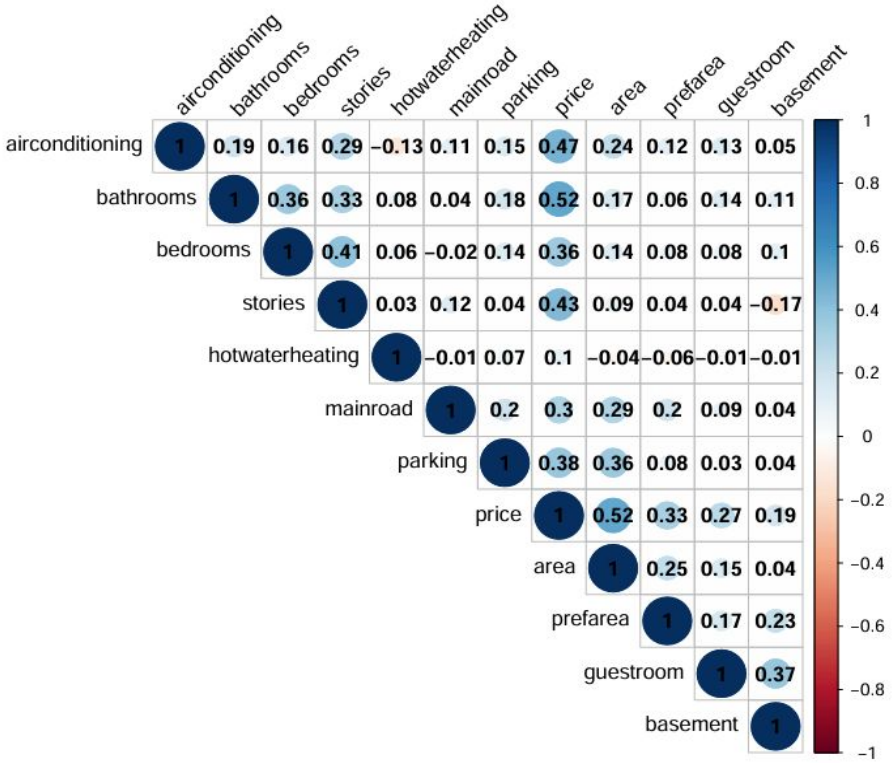
# Result

| price <int> | area <int> | bedrooms <int> | bathrooms <int> | stories <int> | parking <int> | mainroad <dbl> | guestroom <dbl> | basement <dbl> |
|---|---|---|---|---|---|---|---|---|
| 13300000 | 7420 | 4 | 2 | 3 | 2 | 1 | 0 | 0 |
| 12250000 | 8960 | 4 | 4 | 4 | 3 | 1 | 0 | 0 |
| 12250000 | 9960 | 3 | 2 | 2 | 2 | 1 | 0 | 1 |
| 12215000 | 7500 | 4 | 2 | 2 | 3 | 1 | 0 | 1 |

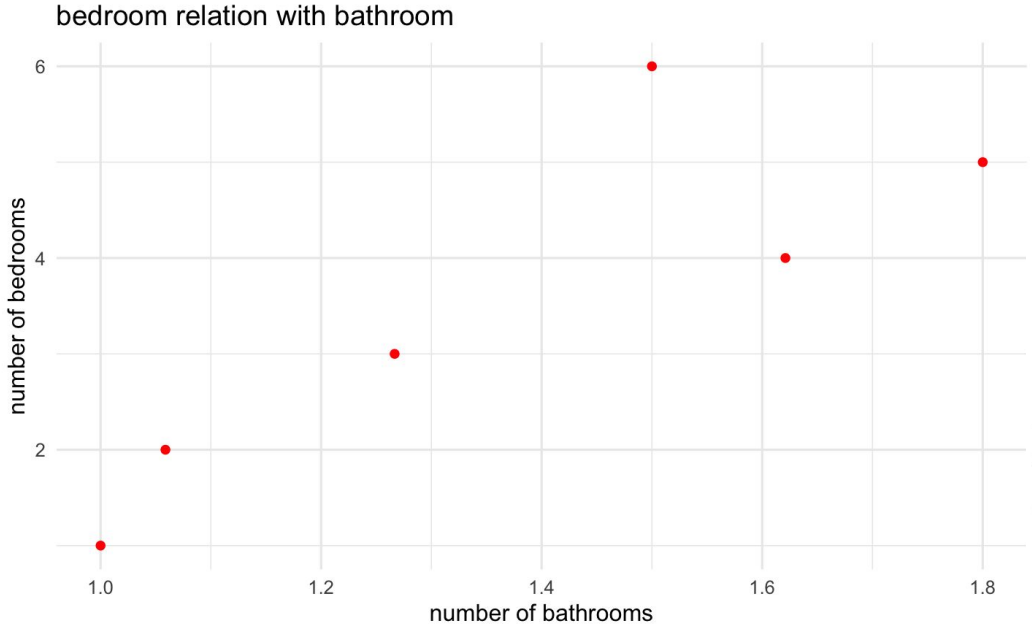
| hotwaterheating <dbl> | airconditioning <dbl> | prefarea <dbl> | furnistatfur <dbl> | furnistatunfur <dbl> | furnistatsemi <dbl> |
|---|---|---|---|---|---|
| 0 | 1 | 1 | 1 | 0 | 0 |
| 0 | 1 | 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 1 | 1 | 0 | 0 |

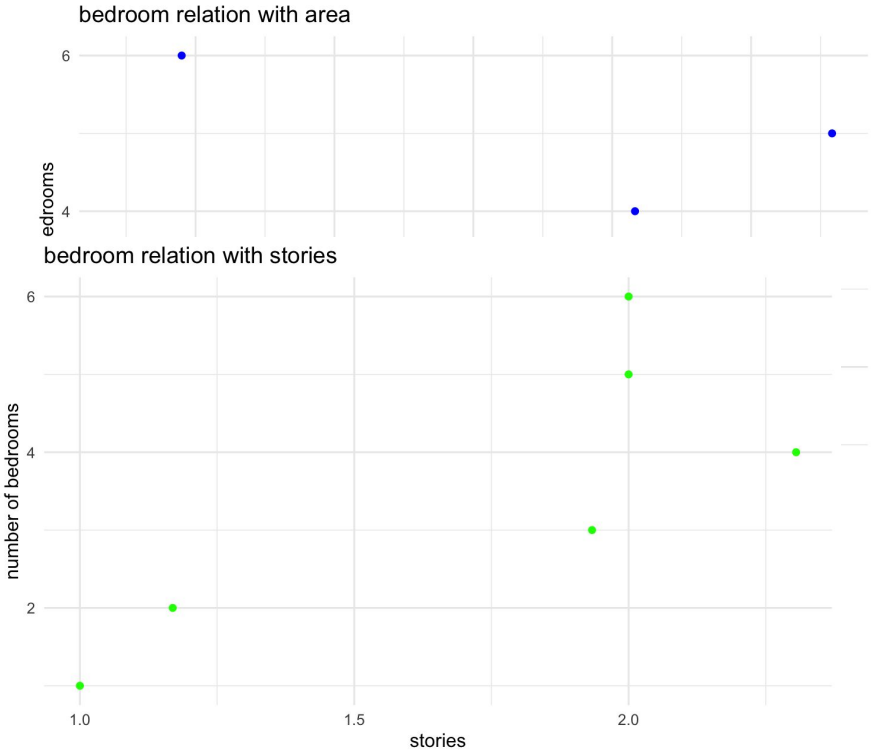
# Exploratory Data Analysis

# Correlation Matrix



# Bedroom shows linear relationships with other predictors




Likely to be an insignificant predictor

# Modeling

# Basic Model

```
basic = lm(price ~ 0 + ., data = cleanhousing2)
summary(basic)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-2619718  -657322   -68409   507176  5166695

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
area                244.14      24.29  10.052  < 2e-16 ***
bedrooms         114787.56   72598.66   1.581 0.114445
bathrooms        987668.11  103361.98   9.555  < 2e-16 ***
stories          450848.00   64168.93   7.026 6.55e-12 ***
parking          277107.10   58525.89   4.735 2.82e-06 ***
mainroad         421272.59  142224.13   2.962 0.003193 **
guestroom        300525.86  131710.22   2.282 0.022901 *
basement         350106.90  110284.06   3.175 0.001587 **
hotwaterheating  855447.15  223152.69   3.833 0.000141 ***
airconditioning  864958.31  108354.51   7.983 8.91e-15 ***
prefarea         651543.80  115682.34   5.632 2.89e-08 ***
furnistatfur      42771.69  264313.31   0.162 0.871508
furnistatunfur  -368462.69  237805.59  -1.549 0.121875
furnistatsemi     -3572.93  249642.21  -0.014 0.988586
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1068000 on 531 degrees of freedom
Multiple R-squared:  0.9576,    Adjusted R-squared:  0.9565
F-statistic: 856.8 on 14 and 531 DF,  p-value: < 2.2e-16
```

## Bedroom and status of furnishing

P-value shows to be > 0.05

# Reduced Model

```
significant = lm(price ~ 0+bathrooms+area+stories+parking+mainroad+basement
                 +hotwaterheating+airconditioning+prefarea+furnistatunfur
                 , data = cleanhousing2)
anova(significant,basic)
```

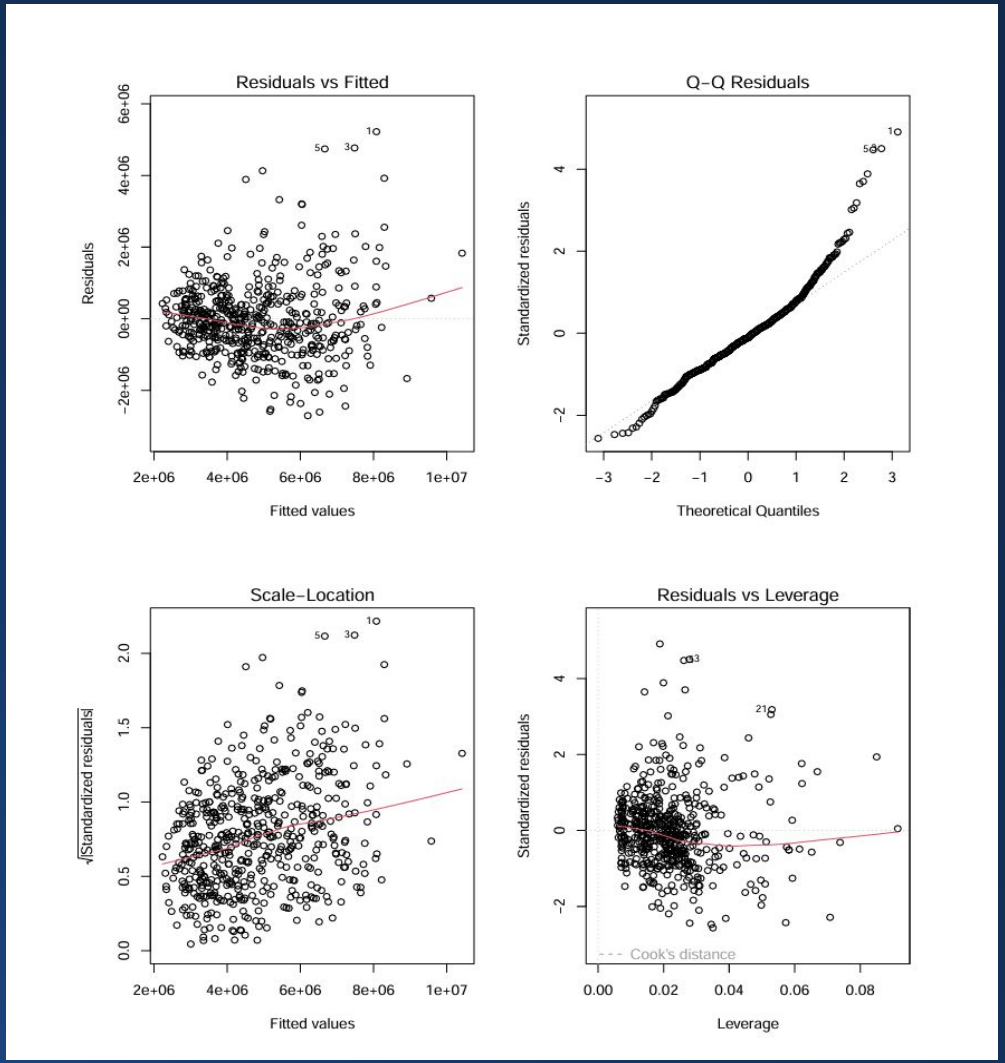
```
Analysis of Variance Table

Model 1: price ~ 0 + bathrooms + area + stories + parking + mainroad +
    basement + hotwaterheating + airconditioning + prefarea +
    furnistatunfur
Model 2: price ~ 0 + (area + bedrooms + bathrooms + stories + parking +
    mainroad + guestroom + basement + hotwaterheating + airconditioning +
    prefarea + furnistatfur + furnistatunfur + furnistatsemi)
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    535 6.1552e+14
2    531 6.0560e+14  4 9.922e+12 2.1749 0.0706 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

delete features: bedrooms, semi furnitured and furnitured(greater p-values), guestroom(even in 0.05 range)

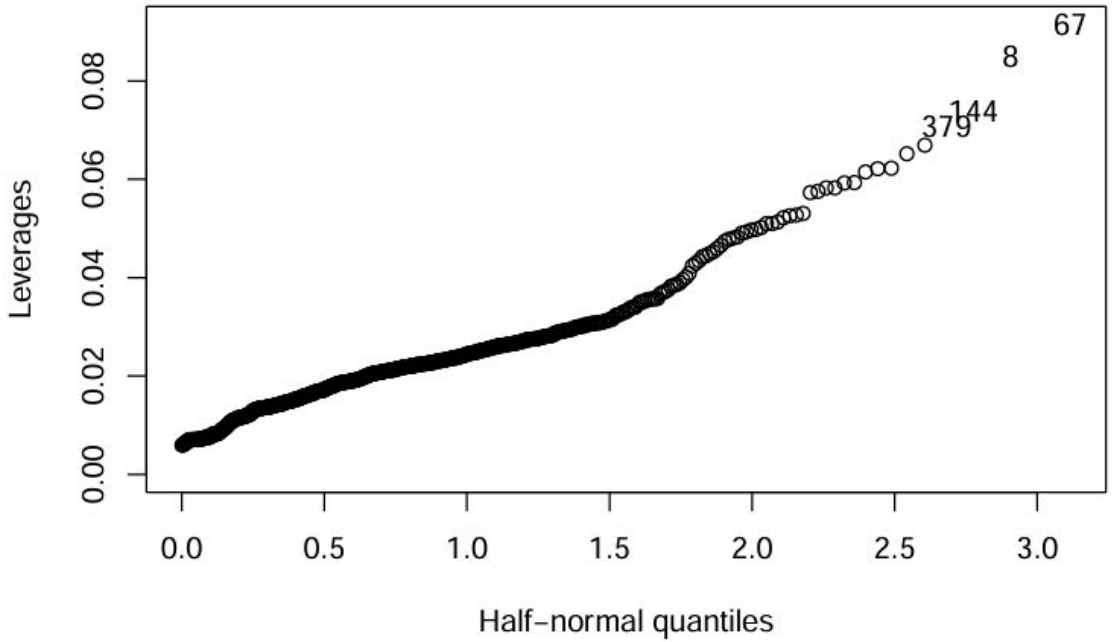
# Diagnostics

# Use par() function for initial analysis



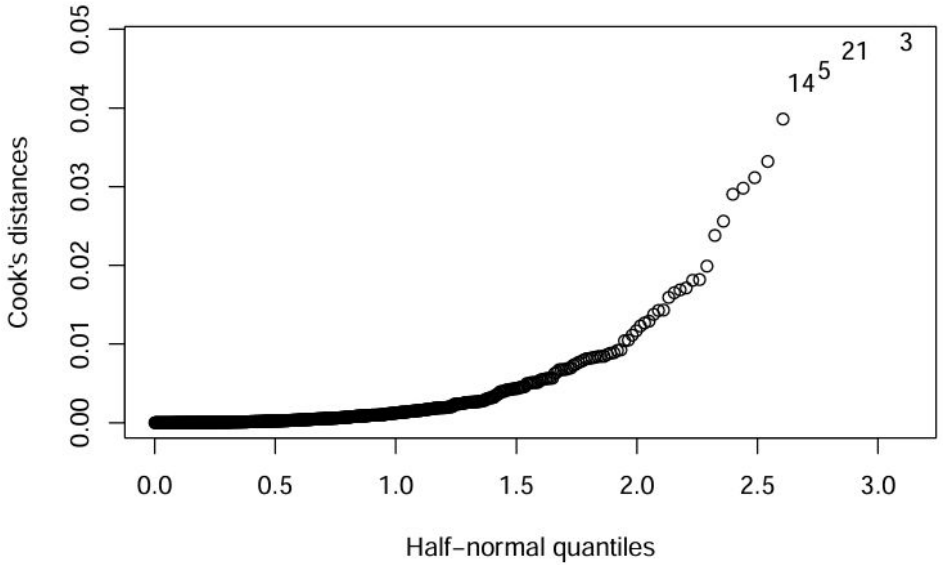
## Checking high-leverage points

```
lev=influence(housing_full)$hat
halfnorm(lev, 4, ylab="Leverages")
```



## Check high-influential points

```
cook = cooks.distance(housing_full)
halfnorm(cook, 4, ylab="Cook's distances")
```



## Checking Outliers

```
sr.ex=rstudent(housing_full);
sort(sr.ex, decreasing=TRUE)[1:5]
```

```
##        1        3        5       16        4
## 5.021002 4.588395 4.557984 3.940170 3.747613
```

```
> qt(0.05/545, 530)
[1] -3.767354
```

The result indicate that we need to drop row 1, 3 ,5, 16 in the dataset

# Reduced Model with outliers dropped

```
significant2 = lm(price ~
0+bathrooms+area+stories+parking+mainroad+basement
                +hotwaterheating+airconditioning+prefarea+furnistatunfur
                , data = dropcleanhousing2)
summary(significant2)
anova(significant2,basicdrop)
```

```
  Res.Df        RSS Df  Sum of Sq      F  Pr(>F)
1    531 5.2666e+14
2    527 5.1610e+14  4 1.0556e+13 2.6946 0.03028 *
```

# Compare two models

```
> summary(significant)

Residual standard error: 1073000 on 535 degrees of freedom
Multiple R-squared:  0.9569,    Adjusted R-squared:  0.9561
F-statistic:  1188 on 10 and 535 DF,  p-value: < 2.2e-16
```

```
> summary(significant2)

Residual standard error: 995900 on 531 degrees of freedom
Multiple R-squared:  0.9617,    Adjusted R-squared:  0.961
F-statistic:  1333 on 10 and 531 DF,  p-value: < 2.2e-16
```

delete features: bedrooms, semi furnitured and furnitured(greater p-values), guestroom(even in 0.05 range)

# Model Analysis

```
summary(significant)
```

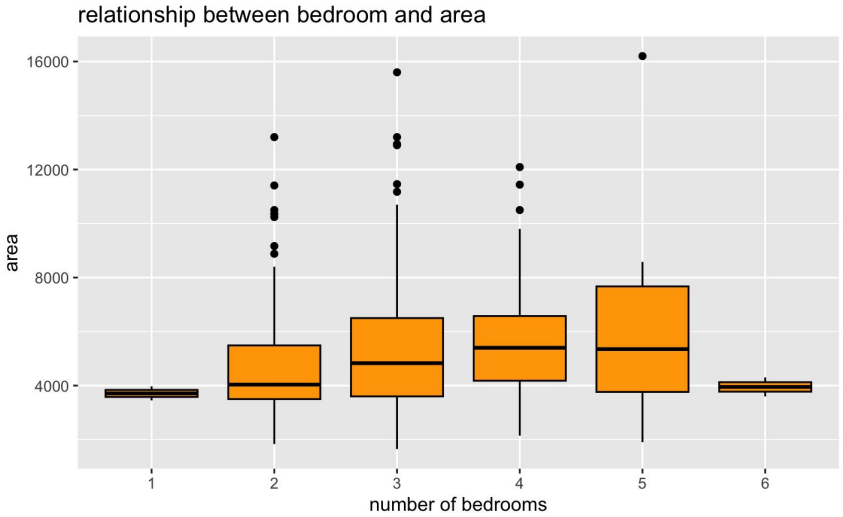
```
Residual standard error: 1073000 on 535 degrees of freedom
```

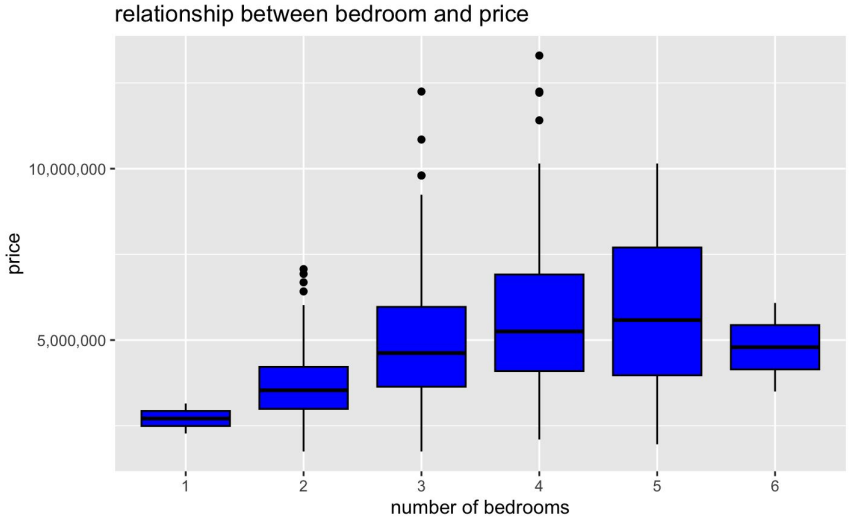
```
> AIC(significant)
[1] 16693.87
```


# Predictors with high variance

```
summary(housing$area)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1650    3600    4600    5151    6360   16200
```

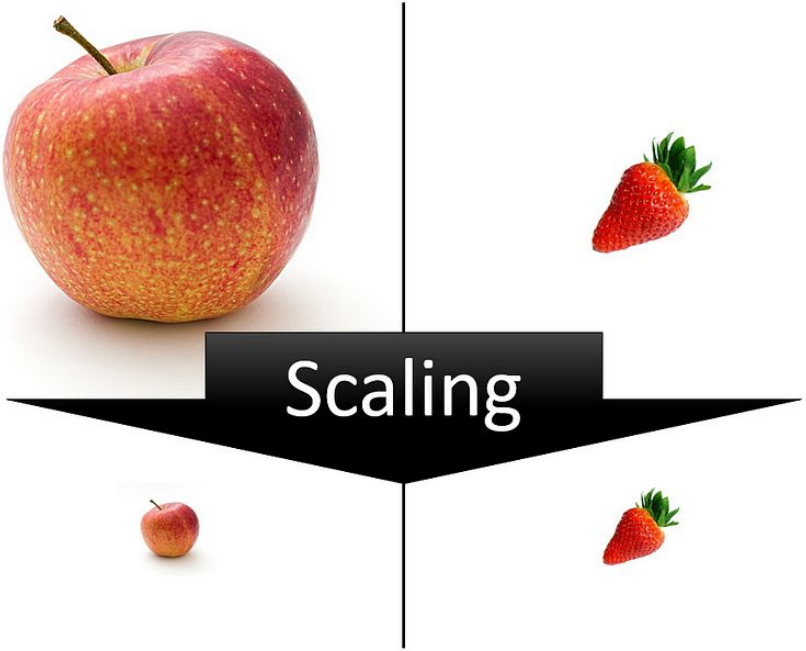
```
> summary(housing$price)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
  1750000  3430000  4340000  4766729  5740000  13300000
```




The wide range of the area and price create high variance.

# Potential Solution



# Scaled Model

| price <int> | area <int> | bedrooms <int> | bathrooms <int> | stories <int> |
|---|---|---|---|---|
| 13300000 | 7420 | 4 | 2 | 3 |
| 12250000 | 8960 | 4 | 4 | 4 |
| 12250000 | 9960 | 3 | 2 | 2 |
| 12215000 | 7500 | 4 | 2 | 2 |
| 11410000 | 7420 | 4 | 1 | 2 |
| 10850000 | 7500 | 3 | 3 | 1 |
| 10150000 | 8580 | 4 | 3 | 4 |
| 10150000 | 16200 | 5 | 3 | 2 |

price ~ 10^7

area ~ 10^4

Bedrooms < 10

# Scaled Model

Original

```
Residual standard error: 989600 on 527 degrees of freedom
Multiple R-squared:  0.9625,    Adjusted R-squared:  0.9615
F-statistic:   965 on 14 and 527 DF,  p-value: < 2.2e-16
```

Scaled

```
Residual standard error: 0.9896 on 527 degrees of freedom
Multiple R-squared:  0.9625,    Adjusted R-squared:  0.9615
F-statistic:   965 on 14 and 527 DF,  p-value: < 2.2e-16
```

# Logged Model

```
dividemod = lm(I(log(price)) ~ 0 + bedrooms + bathrooms + I(area / 1000) + stories + parking + mainroad
summary(dividemod)
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## bedrooms          0.024281   0.013918   1.745 0.081651 .
## bathrooms         0.169989   0.019824   8.575  < 2e-16 ***
## I(area/1000)      0.049552   0.004634  10.693  < 2e-16 ***
## stories           0.090359   0.012237   7.384 6.02e-13 ***
## parking           0.037734   0.011213   3.365 0.000821 ***
## mainroad          0.120739   0.027149   4.447 1.06e-05 ***
## basement          0.102803   0.019884   5.170 3.32e-07 ***
## hotwaterheating   0.169841   0.042620   3.985 7.70e-05 ***
## airconditioning   0.180365   0.020722   8.704  < 2e-16 ***
## prefarea          0.129749   0.022212   5.841 9.06e-09 ***
## furnistatfur      14.361593   0.050523 284.261  < 2e-16 ***
## furnistatunfur    14.252101   0.045434 313.690  < 2e-16 ***
## furnistatsemi     14.379044   0.047689 301.516  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2038 on 528 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9998
## F-statistic: 2.345e+05 on 13 and 528 DF,  p-value: < 2.2e-16
```

Since the price of the houses are right skewed, we conducted log transformation.
The transformation is very effective since the R^2 is high.

# Logged Model

```
dividenobed = lm(I(log(price)) ~ 0 + bathrooms + I(area / 1000) + stories + parking + mainroad + baseme
anova(dividenobed,dividemod)
```

```
## Analysis of Variance Table
##
## Model 1: I(log(price)) ~ 0 + bathrooms + I(area/1000) + stories + parking +
##     mainroad + basement + hotwaterheating + airconditioning +
##     prefarea + furnistatfur + furnistatunfur + furnistatsemi
## Model 2: I(log(price)) ~ 0 + bedrooms + bathrooms + I(area/1000) + stories +
##     parking + mainroad + basement + hotwaterheating + airconditioning +
##     prefarea + furnistatfur + furnistatunfur + furnistatsemi
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    529 22.067
## 2    528 21.940  1   0.12646 3.0433 0.08165 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bedroom is still not necessary to include in the model after we perform the log transformation of the response variable.

# Conclusion

# Conclusion

- No high correlation between each two predictors
- Bedroom seems to have linear relationship with the response variable price
- Most reduced model

```
significant = lm(price ~ 0+bathrooms+area+stories+parking+mainroad+basement
                 +hotwaterheating+airconditioning+prefarea+furnistatunfur
                 , data = cleanhousing2)
anova(significant,basic)
```

- Predictors have high variance, influencing the accuracy of prediction
- Created log transformation of the model, shows excellent result in prediction

```
dividemod = lm(I(log(price)) ~ 0 + bedrooms + bathrooms + I(area / 1000) + stories + parking + mainroad
summary(dividemod)
```

- Bedroom is not necessary to be in the model if other predictors are present


Thank you for listening
ILLINOIS