Concepts in Practical Model Building

Concepts in practical model building

- Recoding variables
- Transformations
- Multicollinearity

LA Homes

What factors help explain the price of a home in Los Angeles?

Model building

We'd like to build a model to explain prices of homes in LA as a function of the characteristics of those homes.

$$\widehat{price} = location + size + pool + acreage \ldots$$

1. Statistical question
2. Data wrangling
3. Exploratory Data Analysis
4. Modeling
5. Interpretation

Consider: *exploratory* vs. *confirmatory* analysis.

Data wrangling

Home price data is available on many websites now, including zillow.com.

```
LA <- read.csv("../data/LA.csv")
head(LA)
```

```
##          city type bed bath garage sqft pool spa  price
## 1 Long Beach         0    1         513        NA 119000
## 2 Long Beach         0    1         550        NA 153000
## 3 Long Beach         0    1         550        NA 205000
## 4 Long Beach         0    1       1 1030        NA 300000
## 5 Long Beach         0    1       1 1526        NA 375000
## 6 Long Beach         1    1         552        NA 159900
```

Unit of observation: a home for sale in west LA.
Population: all homes in west LA.

Data wrangling, cont.

```
str(LA)
```

```
## 'data.frame':    1594 obs. of  9 variables:
##  $ city  : Factor w/ 4 levels "Beverly Hills",..: 2 2 2 2 2 2 2 2 2 2
##  $ type  : Factor w/ 3 levels "","Condo/Twh",..: 1 1 1 1 1 1 1 1 1 1 .
##  $ bed   : int  0 0 0 0 0 1 1 1 1 1 ...
##  $ bath  : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ garage: Factor w/ 5 levels "","1","2","3",..: 1 1 1 2 2 1 1 1 1 1 .
##  $ sqft  : int  513 550 550 1030 1526 552 558 596 744 750 ...
##  $ pool  : Factor w/ 2 levels "","Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ spa   : logi  NA NA NA NA NA NA ...
##  $ price : num  119000 153000 205000 300000 375000 ...
```

```
levels(LA$city)
```

```
## [1] "Beverly Hills" "Long Beach"    "Santa Monica"  "Westwood"
```

Recoding `type`

The levels of a categorical variable can be queried using `levels()`.

```
levels(LA$type)
```

```
## [1] ""          "Condo/Twh" "SFR"
```

```
LA <- LA %>%
  mutate(type = fct_recode(type,
                               "unknown" = "",
                               "condo"   = "Condo/Twh",
                               "sfr"     = "SFR"))
levels(LA$type) <- c("unknown", "condo", "sfr")
```

```
levels(LA$type)
```

```
## [1] "unknown" "condo"   "sfr"
```

Recoding `garage`

```r
str(LA)
```

```
## 'data.frame':    1594 obs. of  9 variables:
##  $ city  : Factor w/ 4 levels "Beverly Hills",..: 2 2 2 2 2 2 2 2 2 2
##  $ type  : Factor w/ 3 levels "unknown","condo",..: 1 1 1 1 1 1 1 1 1
##  $ bed   : int  0 0 0 0 0 1 1 1 1 1 ...
##  $ bath  : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ garage: Factor w/ 5 levels "","1","2","3",..: 1 1 1 2 2 1 1 1 1 1 .
##  $ sqft  : int  513 550 550 1030 1526 552 558 596 744 750 ...
##  $ pool  : Factor w/ 2 levels "","Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ spa   : logi  NA NA NA NA NA NA ...
##  $ price : num  119000 153000 205000 300000 375000 ...
```

What's going on with garage?

Recoding `garage`, cont.

```
levels(LA$garage)
```

```
## [1] ""    "1"   "2"   "3"   "4+"
```

```
count(LA, garage)
```

```
## # A tibble: 6 x 2
##   garage     n
##   <fct>  <int>
## 1 ""       388
## 2 1        260
## 3 2        666
## 4 3         37
## 5 4+         6
## 6 <NA>     237
```

Recoding `garage`, cont.

We can combine levels using a similar approach.

```
LA <- LA %>%
  mutate(garage = fct_collapse(garage,
                               "small" = c("", "1"),
                               "large" = c("2", "3", "4+")))
```

```
count(LA, garage)
```

```
## # A tibble: 3 x 2
##   garage     n
##   <fct>  <int>
## 1 small    648
## 2 large    709
## 3 <NA>     237
```

Data wrangling, cont.

```
str(LA)
```

```
## 'data.frame':    1594 obs. of  9 variables:
##  $ city  : Factor w/ 4 levels "Beverly Hills",..: 2 2 2 2 2 2 2 2 2 2
##  $ type  : Factor w/ 3 levels "unknown","condo",..: 1 1 1 1 1 1 1 1 1
##  $ bed   : int  0 0 0 0 0 1 1 1 1 1 ...
##  $ bath  : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ garage: Factor w/ 2 levels "small","large": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sqft  : int  513 550 550 1030 1526 552 558 596 744 750 ...
##  $ pool  : Factor w/ 2 levels "","Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ spa   : logi  NA NA NA NA NA NA ...
##  $ price : num  119000 153000 205000 300000 375000 ...
```

What's going on with `pool` and `spa`?

Dropping columns

```
count(LA, pool)
```

```
## # A tibble: 2 x 2
##   pool      n
##   <fct> <int>
## 1 ""     1448
## 2 Y       146
```

```
LA %>%
  summarize(nas = sum(is.na(spa)))
```

```
##    nas
## 1 1594
```

Two variables seem mis-coded/uninformative, so they should be dropped.

```
LA <- select(LA, -pool, -spa)
```

Fully wrangled data set

```
head(LA)
```

```
##          city    type bed bath garage sqft  price
## 1 Long Beach unknown   0    1  small  513 119000
## 2 Long Beach unknown   0    1  small  550 153000
## 3 Long Beach unknown   0    1  small  550 205000
## 4 Long Beach unknown   0    1  small 1030 300000
## 5 Long Beach unknown   0    1  small 1526 375000
## 6 Long Beach unknown   1    1  small  552 159900
```

Once the data set is ready to go, save it to a new .csv file.

```
write.csv(LA, file = "LA.csv")
```

Exploratory Data Analysis

Our goals are to:

1. Develop a sense of the *univariate* distributions in terms of center, shape, spread, unusual observations.

2. Develop a sense of the *bivariate* and *multivariate* distributions and what they indicate about the relationship between variables.

Question

Which of the following are *not* good methods to visualize the distribution of a single variable?

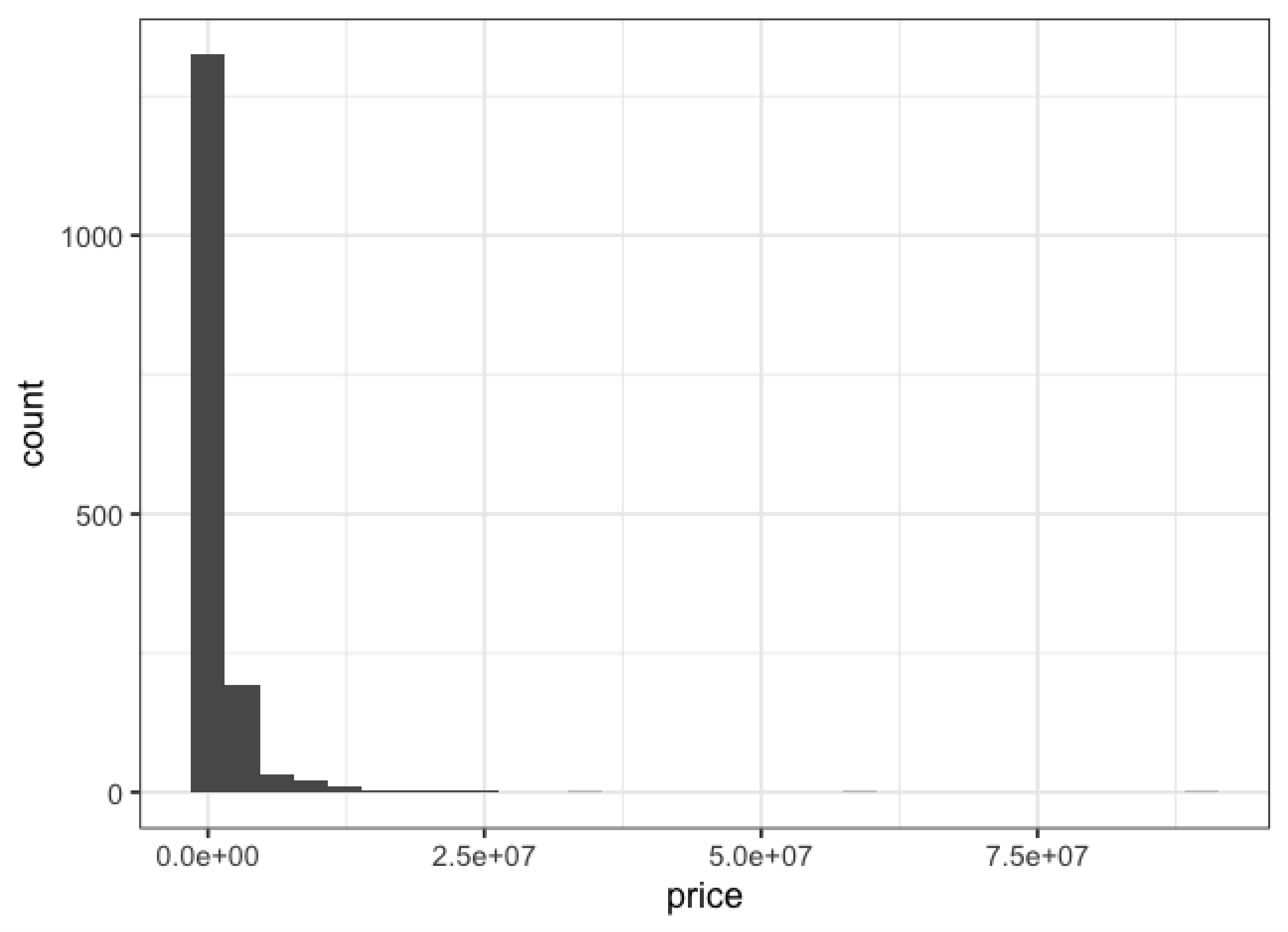
1. mosaic plot
2. density plot
3. scatterplot
4. histogram
5. side-by-side boxplots

Question

Which of the following at *not* good methods to visualize the distribution of a single variable?

1. **mosaic plot**
2. density plot
3. **scatterplot**
4. histogram
5. **side-by-side boxplots**

EDA for `price`



Question

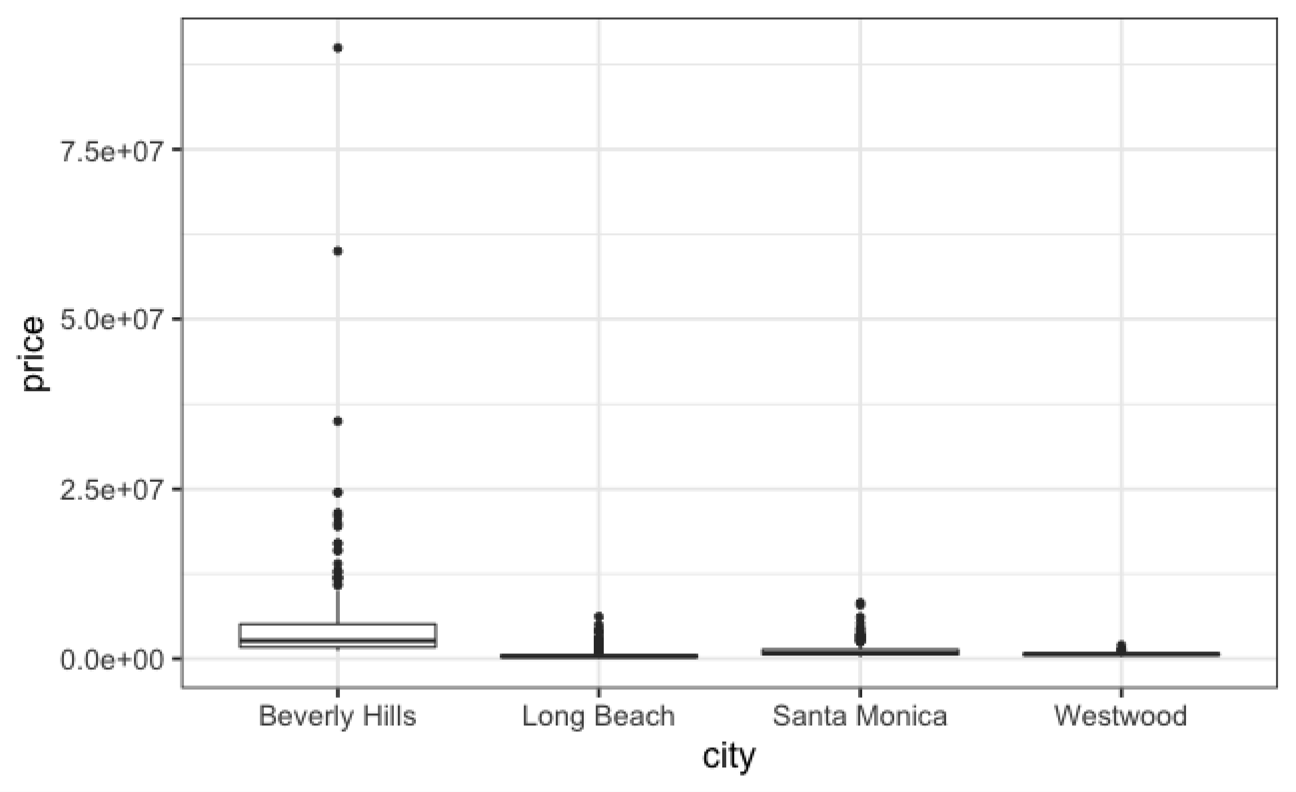
How would you visualize the relationship between `price` and `city`?

```
head(LA)
```

```
##          city    type bed bath garage sqft  price
## 1 Long Beach unknown   0    1  small  513 119000
## 2 Long Beach unknown   0    1  small  550 153000
## 3 Long Beach unknown   0    1  small  550 205000
## 4 Long Beach unknown   0    1  small 1030 300000
## 5 Long Beach unknown   0    1  small 1526 375000
## 6 Long Beach unknown   1    1  small  552 159900
```

Question

How would you visualize the relationship between `price` and `city`?



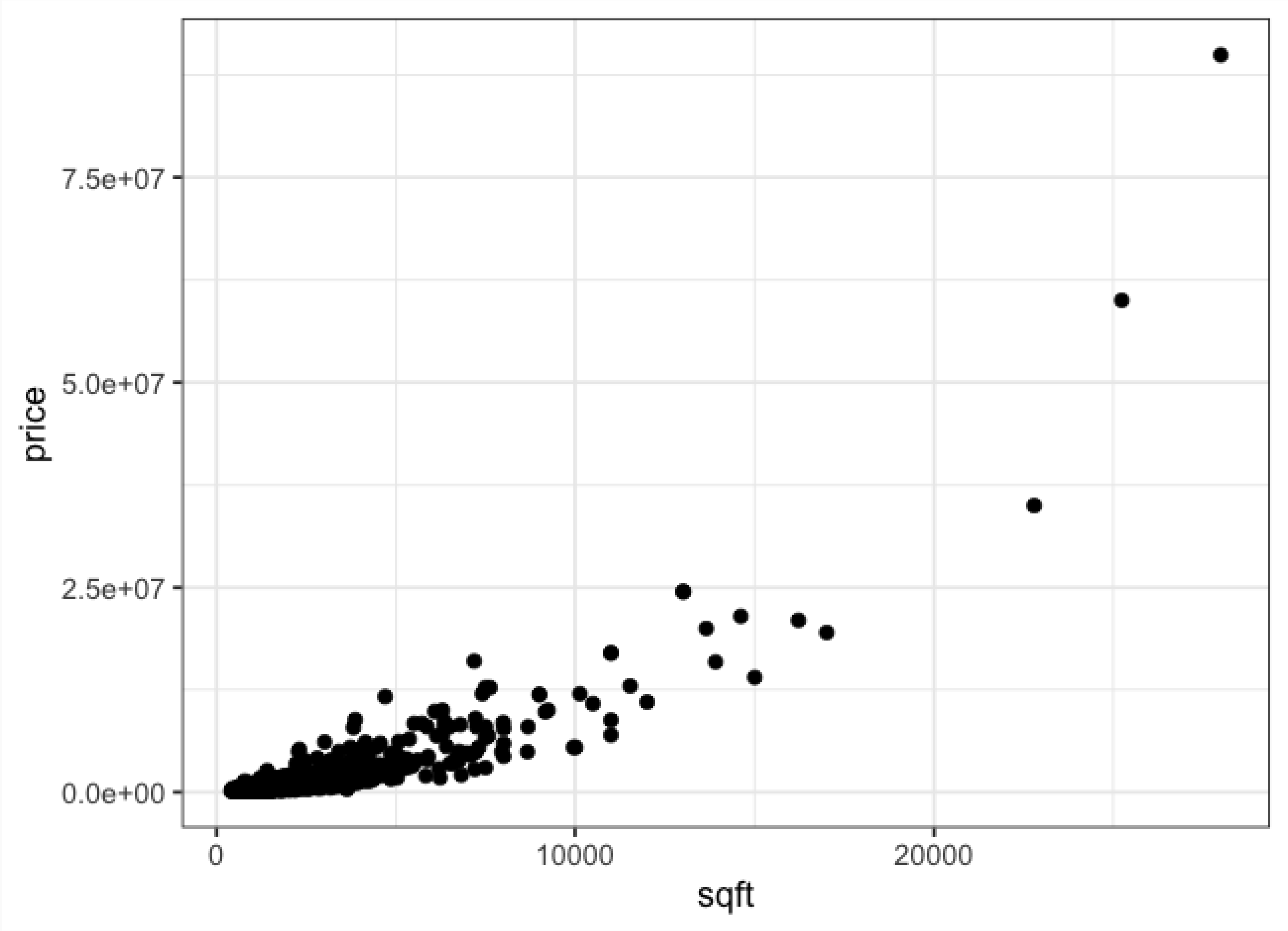
Question

How would you visualize the relationship between `price` and `sqft`?

```
head(LA)
```

```
##           city    type bed bath garage sqft  price
## 1 Long Beach unknown   0    1  small  513 119000
## 2 Long Beach unknown   0    1  small  550 153000
## 3 Long Beach unknown   0    1  small  550 205000
## 4 Long Beach unknown   0    1  small 1030 300000
## 5 Long Beach unknown   0    1  small 1526 375000
## 6 Long Beach unknown   1    1  small  552 159900
```

Question



Transformations

Highly skewed data (particularly the response) can be very difficult to model using least squares regression. A common solution is to consider a transformation of the variable.

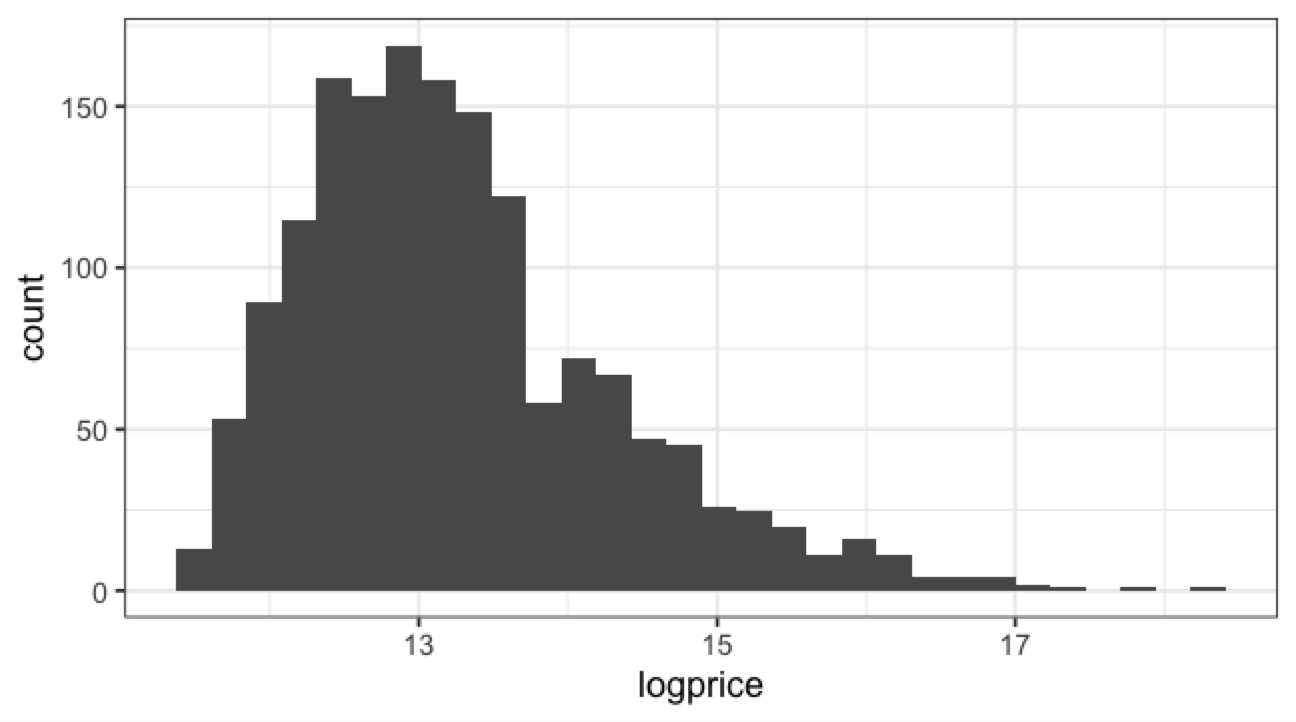
$$\widehat{price} \sim sqft$$

versus

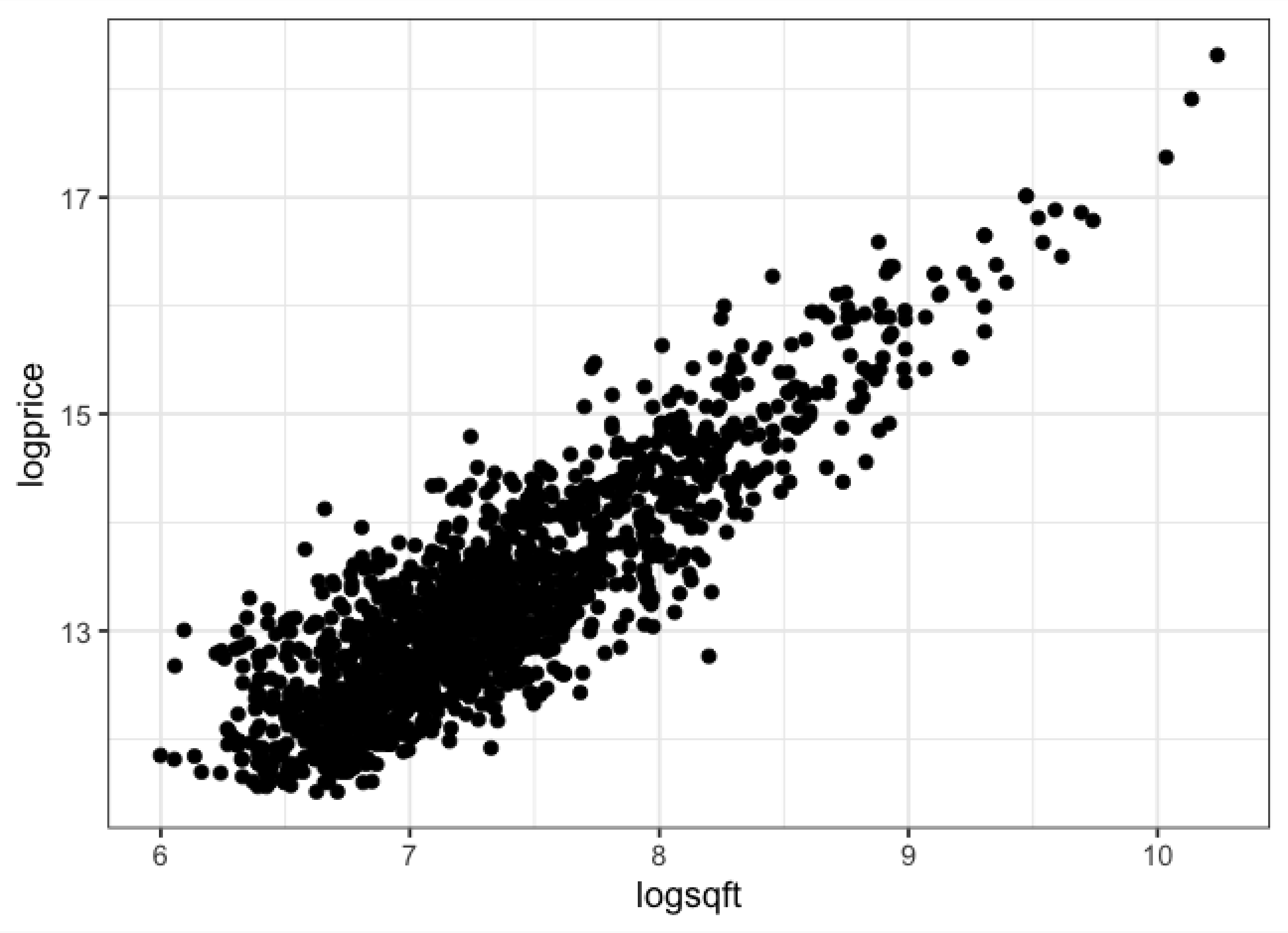
$$\widehat{log(price)} \sim log(sqft)$$

EDA for `price`

```
LA <- LA %>%
  mutate(logprice = log(price))
```

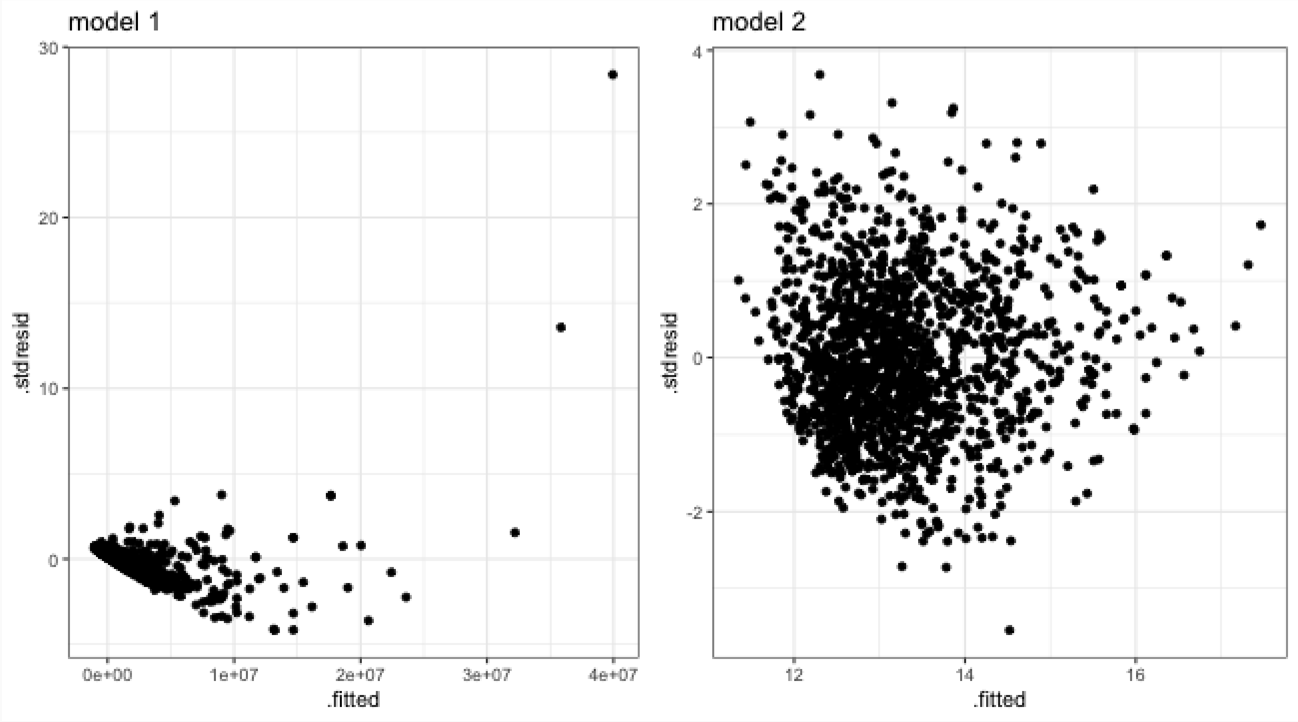


EDA for `logprice` and `logsqft`



Comparing residuals

```
m1 <- lm(price ~ sqft, data = LA)
m2 <- lm(logprice ~ logsqft, data = LA)
```



Transformation, cont.

Highly skewed data often leads to invalid models. This can be often be fixed with a transformation, but the interpretations change slightly.

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.70     0.1437    18.8 1.97e-71
## logsqft           1.44     0.0195    73.8 0.00e+00
```

A one unit increase in the log sqft of a home is associated with a 1.44 unit increase in the log price of a home.

Modeling: a simple model for price

$$\widehat{log(price)} \sim bed$$

```r
m3 <- lm(logprice ~ bed, data = LA)
```

What do you expect the *sign* of the slope for `bed` to be?

```r
summary(m3)$coef
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   11.802     0.0436   270.6  0.00e+00
## bed            0.532     0.0142    37.3 9.77e-220
```

A less simple model for price

$$\widehat{log(price)} \sim bed + log(sqft)$$

```
m4 <- lm(log(price) ~ bed + logsqft, data = LA)
```

What do you expect the *sign* of the slope for `bed` and `logsqft` to be?

```
summary(m4)$coef
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    1.467     0.2178    6.73  2.28e-11
## bed           -0.123     0.0164   -7.46  1.46e-13
## logsqft        1.656     0.0346   47.85 2.60e-310
```

Multicollinearity

Correlation between the predictors has some important consequences:

1. Addition or removal of correlated predictors can lead to slope sign changes.
2. Correlation between the predictors leads to an inflated $SE(b_1)$.

In sum: multicollinearity leads to **instability** in your estimates.

The SE of the slope

In the case of a MLR model with two correlated predictors,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

If $r_{1,2}$ is the correlation between X_1 and X_2, then.

$$SE(b_i) \propto \frac{\sigma}{1 - r_{1,2}}$$

Multicollinearity

Take-home lesson: if your predictors are correlated, then they're carrying the same information about the response and your model with have a difficult time attributing explanatory power to this variable or that.

One approach: remove one/some of the correlated predictors.

Another approach: be ok with the flipping signs. Think of each model as an answer to a specific question.