

# House Price Prediction:

## I. Create a table of relevant predictors, hypothesized direction of effect (+/-), and rationale for each hypothesized effect.

**Price Sold:** Sale price of the house. It depends on various factors including but not limited to year built, roof type (Shingle or Tile), garage availability, pool availability, no. of beds, area etc. It is the responsibility of the MLS (Multiple Listing Service) which is a real estate organization that deals with property listings.

**ADOM:** Agent days on Market is the number of days a property has been listed with an individual agent. If there is a change in the agent at any point of time on an active sale, the ADOM will be reset.

### Factor Effects:

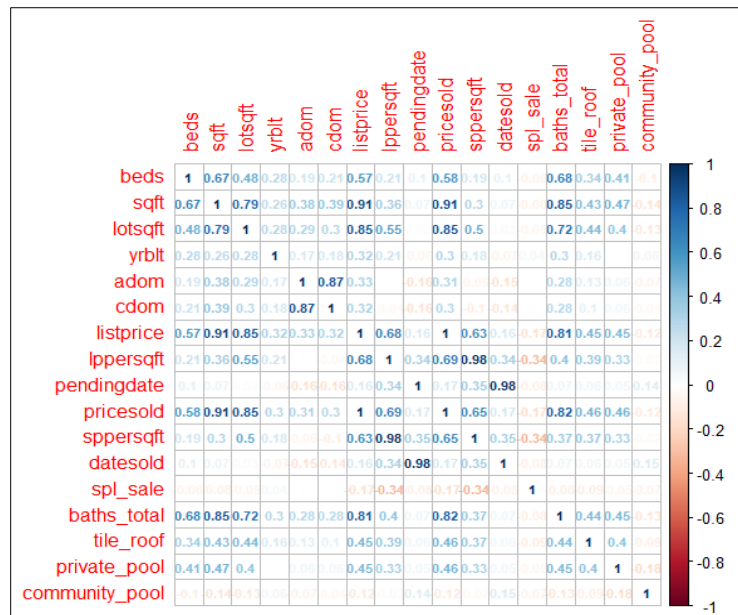
**Positive Effect:** If there is a direct proportionality (+X then +y / -X then -y) between the predictor variable (X) and the response variable (y), we can say there is a positive effect.

**Negative Effect:** If there is an inverse proportionality (+X then -y / -X then +y) between the predictor variable (X) and the response variable (y), we can say there is a negative effect.

Relevant Factors for Price Sold		
Predictor	Effect	Rationale
beds	+	Number of bedrooms increases the total value of a house.
bathsfull	+	Full baths increase house value (Combined with Half baths as single value)
bathshalf	+	Half baths increase house value (Combined with full baths as single value)
sqft	+	Houses with a large area sell for more.
garage	+	Increased garage space adds to the overall price of a house.
roof	+	Roof types are of two types, tile and shingle. All other types are either a variation/combination of these two main types (Considered as binary categorical for this case).
lotsqft	+	Lot size is everything excluding the house area of a property. Larger lot size adds value to a house.
yrblt	-	Newer homes will have a higher value attributed to lower maintenance costs.
Pool	+	In this case there are two types. Private pools add to the house value, but community pools do not add any value.
spa	+/-	Usually, spa adds to the house value but considering only 35% of the data has this metric we will have to check if it has any effect.
listprice	+/-	A higher list price will have a negative impact on a listing appeal. But having it low could also mean the property is not a prime location.
pendingdate	+	A recent listing date could mean a new property, but to make it a predictor we have to convert the date variable into a numerical.
spsale	-	Special sale listings might not be available to all the buyers or the listing might be owned by a bank which limits its availability.
Relevant Factors for Agent days on Market (ADOM)		
Predictor	Effect	Explanation
yrblt	-	More time on the market listing could mean lower value of a house.
lppersqft	+	This is calculated based on the list price and the area (sqft). So, a higher listing price per square foot means it could stay on ADOM longer.
spsale	-	A special sale resets the ADOM time, but a bank ownership could add to the ADOM time due to restrictions.

Irrelevant Factors		
Predictor	Effect	Explanation

<b>slnoskm</b>	No Effect	Identifier index.
<b>Status</b>	No Effect	Shows the status of a listing.
<b>Address</b>	No Effect	Shows address of a listing, cannot be used as a predictor.
<b>bathstotal</b>	No Effect	We are considering number as baths as a float variable so bathstotal can be omitted.
<b>subdivn</b>	No Effect	Almost all the listings are from a single place, so this has no effect.
<b>cdom</b>	No Effect	Like ADOM, not a predictor for this model.
<b>sppersqft</b>	No Effect	Calculated based on the sale price and area (sqft).
<b>datesold</b>	No Effect	This can be omitted as we are considering pending date.



Correlation plot above helps us decide which variables to consider and which to omit from our models. There is a high correlation ( $>0.7$ ) between pricesold and lotsqft, listprice and sqft. Considering lppersqft, lotsqft and sppersqft are functions of sqft, we can drop those all and consider only one. Other variables that show high correlation (like baths\_total) cannot be omitted as that is a unique variable based on bathsfull and bathshalf.

## II. Run a set of three reasonable models for each DV. Copy and paste the R code for the three models and the combined output using stargazer.

*#1st DV Regression models*

```
p1 <- lm(pricesold ~ pendingdate + spl_sale, data = hs)
p2 <- lm(pricesold ~ beds + sqft + garages + yrblt + spl_sale + baths_total + tile_roof + private_pool
+ community_pool, data = hs)
p3 <- lm(pricesold ~ beds + sqft + garages + yrblt + spl_sale + baths_total + tile_roof + private_pool
+ community_pool + pendingdate, data = hs)
```

*#1st DV Stargazer*

```
stargazer(p1, p2, p3, type='text', single.row = TRUE)
```

```
##
## =====
##                               Dependent variable:
##                               -----
##                               pricesold
##                               (1)          (2)          (3)
## -----
## pendingdate      16,957.440*** (4,932.184)
## beds              -24,228.040*** (5,379.690)
## sqft              140.920*** (6.509)
## garages           11,211.410* (6,657.231)
## yrbld             4,636.479*** (1,125.710)
## spl_sale          -103,296.400*** (28,461.800)
## baths_total      30,799.080*** (7,613.655)
## tile_roof        17,808.110*** (6,135.807)
## private_pool     11,715.010 (7,170.511)
## community_pool   957.324 (5,936.581)
## Constant         -33,849,515.000*** (9,948,011.000)
## -----
## Observations      482
## R2                 0.054
## Adjusted R2       0.050
## Residual Std. Error 148,084.200 (df = 479)
## F Statistic       13.616*** (df = 2; 479)
## -----
## Note:
##                               *p<0.1; **p<0.05; ***p<0.01
```

stargazer(p1, p2, p3, type='text', ci=TRUE, ci.level=0.95, single.row = TRUE)

```
##
## =====
##                               Dependent variable:
##                               -----
##                               pricesold
##                               (1)          (2)          (3)
## -----
## pendingdate      16,957.440*** (7,290.540, 26,624.350)
## beds              -24,228.040*** (-34,772.040, -13,684.040)
## sqft              140.920*** (128.163, 153.676)
## garages           11,211.410* (-1,836.527, 24,259.340)
## yrbld             4,636.479*** (2,430.128, 6,842.830)
## spl_sale          -103,296.400*** (-159,080.500, -47,512.340)
## baths_total      30,799.080*** (15,876.590, 45,721.570)
## tile_roof        17,808.110*** (5,782.146, 29,834.070)
## private_pool     11,715.010 (-2,338.930, 25,768.960)
## community_pool   957.324 (-10,678.160, 12,592.810)
## Constant         -33,849,515.000*** (-53,347,258.000, -14,351,772.000)
## -----
## Observations      482
## R2                 0.054
## Adjusted R2       0.050
## Residual Std. Error 148,084.200 (df = 479)
## F Statistic       13.616*** (df = 2; 479)
## -----
## Note:
##                               *p<0.1; **p<0.05; ***p<0.01
```

## #2nd DV Regression models

```
a1 <- lm(adom ~ pendingdate +, data = hs)
a2 <- lm(adom ~ yrbld + pendingdate + lppersqft + spl_sale, data = hs)
a3 <- lm(adom ~ yrbld + pendingdate + lppersqft + spl_sale + baths_total + private_pool + community_pool,
data = hs)
```

## #2nd DV Stargazer

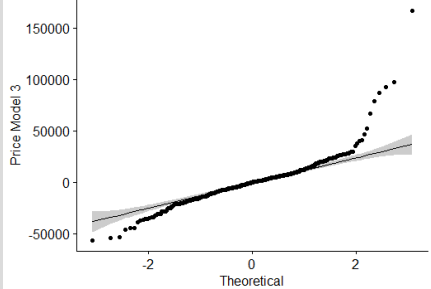
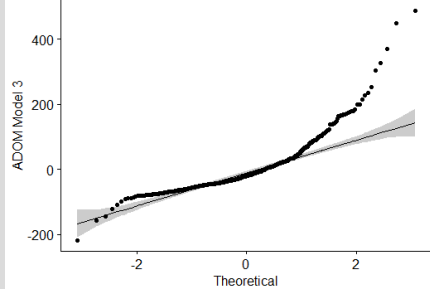
stargazer(a1, a2, a3, type='text', single.row = TRUE)

```
##
## =====
##                               Dependent variable:
##                               -----
##                               adom
##                               (1)          (2)          (3)
## -----
## yrbld             4.938*** (1.445)
## pendingdate      -9.417*** (2.636)
## lppersqft         0.050 (0.191)
## spl_sale          -5.870 (16.122)
## baths_total      35.547*** (5.908)
## private_pool     -8.169 (8.884)
## community_pool   -4.274 (7.847)
## Constant         19,054.870*** (5,317.568)
## -----
## Observations      482
## R2                 0.026
## Adjusted R2       0.024
## Residual Std. Error 79.427 (df = 480)
## F Statistic       12.758*** (df = 1; 480)
## -----
## Note:
##                               *p<0.1; **p<0.05; ***p<0.01
```

```
stargazer(a1, a2, a3, type='text', ci=TRUE, ci.level=0.95, single.row = TRUE)
```

```
##
## =====
##                               Dependent variable:
## -----
##                               adom
##                               (2)
## -----
##                               (1)                               (2)                               (3)
## -----
## yrb1t                               4.938*** (2.105, 7.771)                               2.751* (-0.125, 5.627)
## pendingdate          -9.417*** (-14.584, -4.250)          -9.234*** (-14.723, -3.744)          -8.396*** (-13.772, -3.020)
## lppersqft              0.050 (-0.324, 0.425)              0.050 (-0.324, 0.425)              -0.343* (-0.738, 0.053)
## spl_sale              -5.870 (-37.469, 25.730)              -5.870 (-37.469, 25.730)              -8.855 (-39.530, 21.819)
## baths_total                               35.547*** (23.968, 47.126)
## private_pool                               -8.169 (-25.581, 9.243)
## community_pool                              -4.274 (-19.653, 11.106)
## Constant          19,054.870*** (8,632.630, 29,477.110)  8,832.522 (-4,282.157, 21,947.200)  11,461.770* (-1,550.770, 24,474.300)
## -----
## Observations                               482                               482                               482
## R2                                           0.026                               0.052                               0.126
## Adjusted R2                               0.024                               0.044                               0.113
## Residual Std. Error       79.427 (df = 480)       78.588 (df = 477)       75.696 (df = 474)
## F Statistic              12.758*** (df = 1; 480)       6.585*** (df = 4; 477)       9.790*** (df = 7; 474)
## -----
## Note:                                         *p<0.1; **p<0.05; ***p<0.01
```

III. Select the best model from each set and examine whether it meets the assumptions of the regression model. Which of the five regression assumptions are met for the final models?

Assumptions	1 <sup>st</sup> DV Model: p3 (Price Sold)	2 <sup>nd</sup> DV Model: a3 (ADOM)
Linearity	 <p>Skewed results of residuals shows that the data has a liner bias with some extreme residuals.</p>	 <p>Skewed results of residuals shows that the data has a liner bias and is heavy tailed.</p>

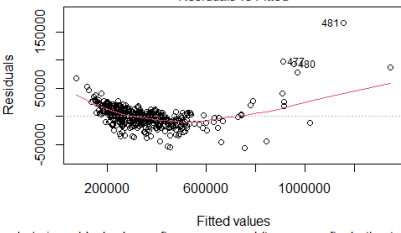
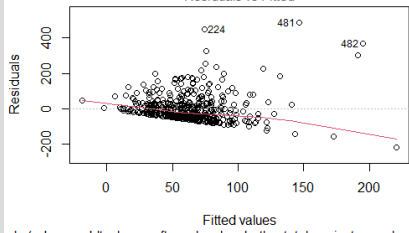
Normality Test:

- **Shapiro-Wilk's Test:** Since Sample Size < 2000
  1. If  $p < 0.05$ , reject  $H_0 \Rightarrow$  data is not normally distributed.
  2. If  $p > 0.05$ , fail to reject  $H_a \Rightarrow$  data may be normally distributed.

Assumptions	1 <sup>st</sup> DV Model: p3 (Price Sold)	2 <sup>nd</sup> DV Model: a3 (ADOM)
Normality	W = 0.90232, p-value < 2.2e-16 This model fails Shapiro-Wilk's test.	W = 0.83243, p-value < 2.2e-16 This model fails Shapiro-Wilk's test.

Homoscedasticity Test:

- **Bartlett Test:**
  1. If  $p < 0.05$ , reject  $H_0 \Rightarrow$  two samples have unequal variances.
  2. If  $p > 0.05$ , fail to reject  $H_a \Rightarrow$  two samples may have equal variances.

Assumptions	1 <sup>st</sup> DV Model: p3 (Price Sold)	2 <sup>nd</sup> DV Model: a3 (ADOM)
<b>Homoscedasticity</b>	 <p>lm(pricesold ~ beds + sqft + garages + yrblt + sppersqft + baths_total + ti</p> <p>Bartlett's K-squared = 379.74, df = 1, p-value &lt; 2.2e-16</p> <p><b>This model fails Bartlett test.</b></p>	 <p>lm(adom ~ yrblt + lppersqft + spl_sale + baths_total + private_pool + comm</p> <p>Bartlett's K-squared = 392.91, df = 1, p-value &lt; 2.2e-16</p> <p><b>This model fails Bartlett test.</b></p>

#### Multicollinearity Test:

- Variance Inflation Factor (VIF):**

- VIF =  $1/T$  (Where  $T = 1 - R^2$ ,  $T < 0.1$  is indicative of multicollinearity).
- VIF > 5 indicates multicollinearity. VIF > 10 is strong evidence of multicollinearity.

Assumptions	1 <sup>st</sup> DV Model: p3 (Price Sold)	2 <sup>nd</sup> DV Model: a3 (ADOM)
<b>Multicollinearity</b>	<p>vif(p3)</p> <pre>##      beds      sqft      garages      yrblt      spl_sale ## 2.075706  4.173145  1.947956  1.228472  1.028684 ##  baths_total  tile_roof  private_pool  community_pool  pendingdate ##  4.338559  1.358516  1.536042  1.083330  1.048908</pre> <p><b>This model passed the VIF test.</b></p>	<p>vif(a3)</p> <pre>##      yrblt  pendingdate  lppersqft  spl_sale  baths_total ## 1.203863  1.191815  1.640892  1.165093  1.515354 ## private_pool  community_pool ## 1.373933  1.080759</pre> <p><b>This model passed the VIF test.</b></p>

#### Independence Test:

- Durbin-Watson's Test (DW):**

- $H_0$ : Residuals are not linearly auto-correlated.
- DW ~ [0, 4]; values around 2 (i.e., 1.5 to 2.5) suggests no autocorrelation.

Assumptions	1 <sup>st</sup> DV Model: p3 (Price Sold)	2 <sup>nd</sup> DV Model: a3 (ADOM)
<b>Independence</b>	<p>DW = 1.556, p-value = 3.798e-07</p> <p><b>This model passed the Durbin-Watson's test.</b></p>	<p>DW = 1.7523, p-value = 0.0027</p> <p><b>This model passed the Durbin-Watson's test.</b></p>

**IV. Using your best models, select the top three predictors of adom and pricesold, and explain their marginal effects on the dependent variables. Remember that we are interested in economic significance, not statistical significance.**

#### Top 3 predictors for pricesold: Model p3

- Baths\_total: Adding a bathroom increases the sales price by \$30,668.7.
- tileroof: Adding a tile type roof increases the sales price by \$16,900.5.
- sqft: An increase in area by a square foot adds \$140.9 to the sale price.

#### Top 3 predictors for adom: Model a3

- pendingdate: Newer homes on average sell 8.3 days sooner.
- baths\_total: Houses with more bathrooms stay on the market for 35.5 days.
- yrblt: Houses that were built recently will stay 2.7 days more on market.