Hunter Green Home Sales Analysis

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

Dependent variable – **Price sold**

Relevant Independent variables

Predictor	**Expected SOE**	**Rationale**
Beds	+	Having a greater number of bedrooms increases the value of the house which results in higher selling price.
Baths(fullbaths+bathhalf/2)	+	Having a greater number of bathrooms does add more comfort, this can increase the selling price of the houses.
garages	+	Garages are used for multiple purposes, so having a bigger garage would increase the selling price of a house.
Roof	Shingles- (-) Tile - (+)	Shingles – They do not last longer compared to tiles also they deteriorate easily hence this can bring down the selling price of a house. Tiles – They are made of good quality and are extremely durable which in turn results in higher selling price of a house.
lotsqft	+ / -	Having a bigger backyard would increase the value of a house but having a bigger front yard may not do the same.
Pool	+/-	Having a private pool is a luxury which would increase the selling price of a house compared to having a community pool.
Spa	+	Having a spa inside a house is a luxury so this would increase the selling price.
adom	-	More number of days in the market would give the leverage to the buyer who would try the reduce the selling price.
lppersqft	+/-	This can have a positive or negative impact on the selling price as depends whether list price per sqft is overpriced or underpriced.
Years old(Year(pending date) – year built))	-	Older houses might be less energy efficient so it would reduce the selling price of a house.
Sqft	+	Houses with more square feet would generally increase the selling price of a house.

Dependent variable – **Agent days on market**

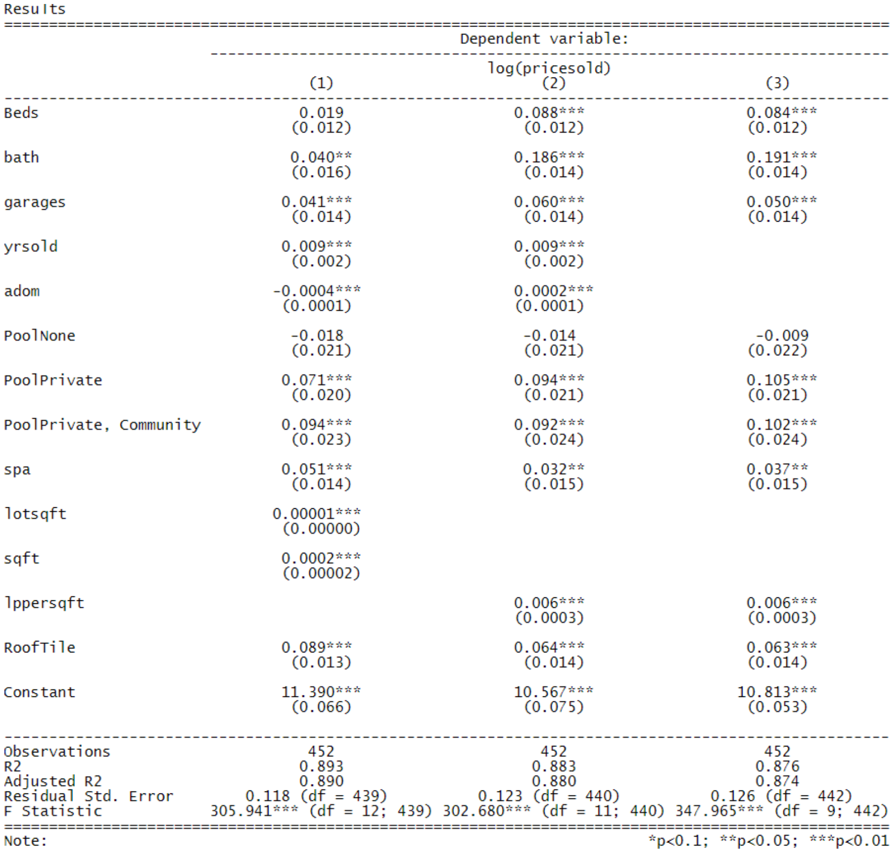
Predictor	**Expected SOE**	**Rationale**
Beds	-	Having a greater number of bedrooms would attract a lot of buyers so this would decrease the adom.
Baths(fullbaths+bathhalf/2)	-	Having a greater number of Baths adds comfort and value to a house, this can help sell the house quicker.
garages	-	A big garage can hold many cars and could also be used for the other purposes. This would attract buyers and decrease the adom.
Roof	Shingles-(+) Tile - (-)	Shingles – They do not last longer compared to tiles also they deteriorate easily hence this can increase the adom as buyers may not be attracted to this type of roofing. Tiles – They are made of good quality and can possibly last up to a century. This would also be a good choice for Florida weather which could attract a lot of buyers.
Lotsqft	+/-	Having a bigger backyard would attract buyers resulting in lesser adom but having a bigger front yard may not do the same.
Pool	+/-	Having a private pool is a luxury which would attract more buyers compared to having just a community pool.
Spa	-	Having a spa inside a house is a luxury which would attract buyers so the house could get sold quicker.
lppersqft	+ / -	The list price per square feet of a house is important because if the house is overpriced then it can't be sold easily, and it would increase the adom. If the house is underpriced, it would be sold quick this would decrease the adom
Years old(Year(pendingdate - yrblt))	+	Adom would be higher if a house is old, as older houses are not energy efficient.
sqft	-	Increasing the square feet of the house would result in bigger living areas attracting lot of buyers.
listprice	+/-	Keeping it to experiments its effect. Its variations are already captured in lppersqft.

2. 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.

1. Dependent variable: **Price sold**
   - A. m1_sp <- lm(log(pricesold) ~ Beds + bath + garages + yrsold + adom + Pool + spa + lotsqft + sqft + Roof, data = df)
   - B. m2_sp <- lm(log(pricesold) ~ Beds + bath + garages + yrsold + adom + Pool + spa + lppersqft + Roof, data = df) - **Best**
   - C. m3_sp <- lm(log(pricesold) ~ Beds + garages + Pool + spa + lppersqft + Roof, data = df)

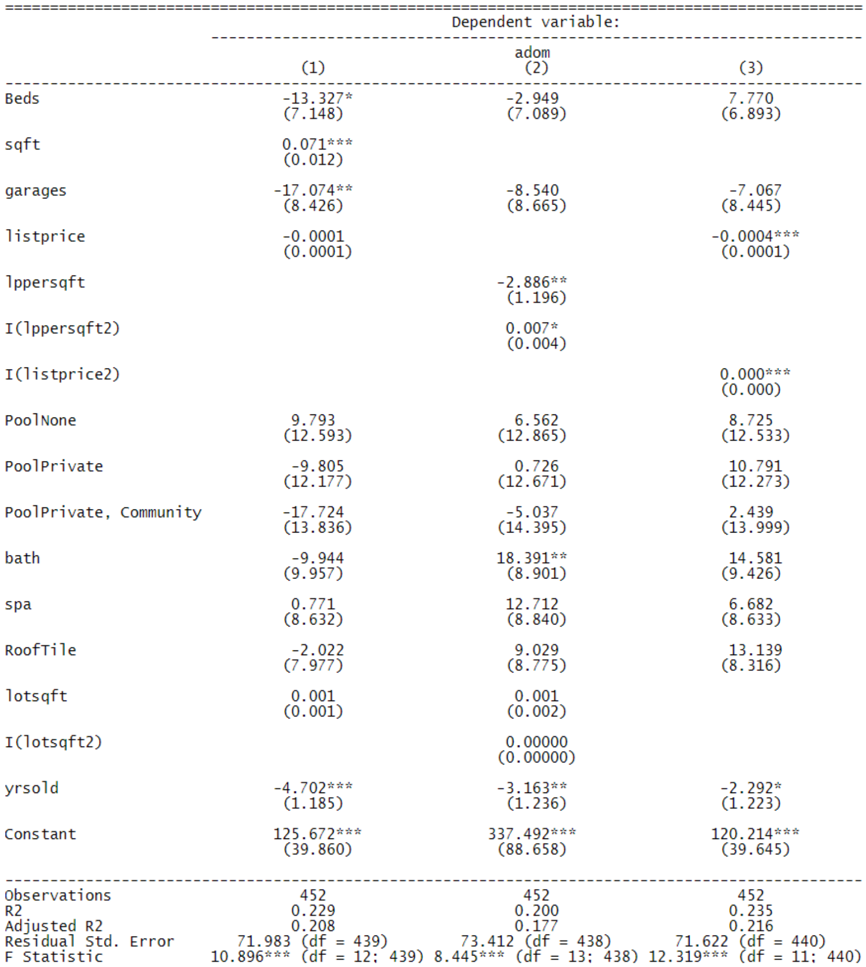
2. Dependent variable: **adom**
   - A. m1_adom <- lm(adom ~ Beds + sqft + garages + listprice + Pool + bath + spa + Roof + lotsqft + yrsold, data = df) - **Best**
   - B. m2_adom <- lm(adom ~ Beds + garages + lppersqft + I(lppersqft^2) + Pool + bath + spa + Roof + lotsqft + I(lotsqft^2) + yrsold, data = df)
   - C. m3_adom <- lm(adom ~ Beds + garages + listprice + I(listprice^2) + Pool + bath + spa + Roof + yrsold, data = df)

For Price sold: stargazer(m1_sp,m2_sp,m3_sp, title="Results", type = "text")



```
Results
=============================================================================================
                                             Dependent variable:
                       ----------------------------------------------------------------------
                                                log(pricesold)
                                (1)                     (2)                      (3)
---------------------------------------------------------------------------------------------
Beds                           0.019                 0.088***                 0.084***
                              (0.012)                 (0.012)                  (0.012)

bath                          0.040**                0.186***                 0.191***
                              (0.016)                 (0.014)                  (0.014)

garages                      0.041***                0.060***                 0.050***
                              (0.014)                 (0.014)                  (0.014)

yrsold                       0.009***                0.009***
                              (0.002)                 (0.002)

adom                        -0.0004***               0.0002***
                             (0.0001)                (0.0001)

PoolNone                      -0.018                  -0.014                   -0.009
                              (0.021)                 (0.021)                  (0.022)

PoolPrivate                  0.071***                0.094***                 0.105***
                              (0.020)                 (0.021)                  (0.021)

PoolPrivate, Community       0.094***                0.092***                 0.102***
                              (0.023)                 (0.024)                  (0.024)

spa                          0.051***                 0.032**                  0.037**
                              (0.014)                 (0.015)                  (0.015)

lotsqft                     0.00001***
                             (0.00000)

sqft                         0.0002***
                             (0.00002)

lppersqft                                            0.006***                 0.006***
                                                     (0.0003)                 (0.0003)

RoofTile                     0.089***                0.064***                 0.063***
                              (0.013)                 (0.014)                  (0.014)

Constant                    11.390***               10.567***                10.813***
                              (0.066)                 (0.075)                  (0.053)

---------------------------------------------------------------------------------------------
Observations                    452                     452                      452
R2                             0.893                   0.883                    0.876
Adjusted R2                    0.890                   0.880                    0.874
Residual Std. Error      0.118 (df = 439)        0.123 (df = 440)         0.126 (df = 442)
F Statistic        305.941*** (df = 12; 439) 302.680*** (df = 11; 440) 347.965*** (df = 9; 442)
=============================================================================================
Note:                                                             *p<0.1; **p<0.05; ***p<0.01
```

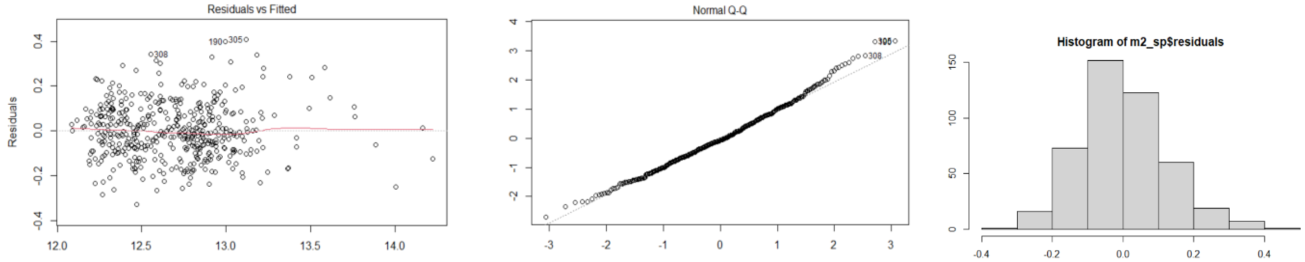
For adom: stargazer(m1_adom,m2_adom,m3_adom, title="Results", type = "text")



```
=========================================================================================
                                           Dependent variable:
                       ------------------------------------------------------------------
                                                  adom
                                (1)                   (2)                     (3)
-----------------------------------------------------------------------------------------
Beds                         -13.327*               -2.949                   7.770
                             (7.148)               (7.089)                  (6.893)

sqft                         0.071***
                             (0.012)

garages                     -17.074**               -8.540                  -7.067
                             (8.426)               (8.665)                  (8.445)

listprice                    -0.0001                                      -0.0004***
                             (0.0001)                                      (0.0001)

lppersqft                                          -2.886**
                                                   (1.196)

I(lppersqft2)                                       0.007*
                                                   (0.004)

I(listprice2)                                                             0.000***
                                                                           (0.000)

PoolNone                      9.793                 6.562                    8.725
                             (12.593)             (12.865)                 (12.533)

PoolPrivate                  -9.805                 0.726                   10.791
                             (12.177)             (12.671)                 (12.273)

PoolPrivate, Community       -17.724                -5.037                   2.439
                             (13.836)             (14.395)                 (13.999)

bath                         -9.944                18.391**                 14.581
                             (9.957)               (8.901)                  (9.426)

spa                           0.771                12.712                    6.682
                             (8.632)               (8.840)                  (8.633)

RoofTile                     -2.022                 9.029                   13.139
                             (7.977)               (8.775)                  (8.316)

lotsqft                       0.001                 0.001
                             (0.001)               (0.002)

I(lotsqft2)                                        0.00000
                                                  (0.00000)

yrsold                      -4.702***              -3.163**                -2.292*
                             (1.185)               (1.236)                  (1.223)

Constant                   125.672***            337.492***               120.214***
                             (39.860)              (88.658)                 (39.645)

-----------------------------------------------------------------------------------------
Observations                   452                   452                      452
R2                            0.229                 0.200                    0.235
Adjusted R2                   0.208                 0.177                    0.216
Residual Std. Error     71.983 (df = 439)     73.412 (df = 438)        71.622 (df = 440)
F Statistic         10.896*** (df = 12; 439) 8.445*** (df = 13; 438) 12.319*** (df = 11; 440)
```

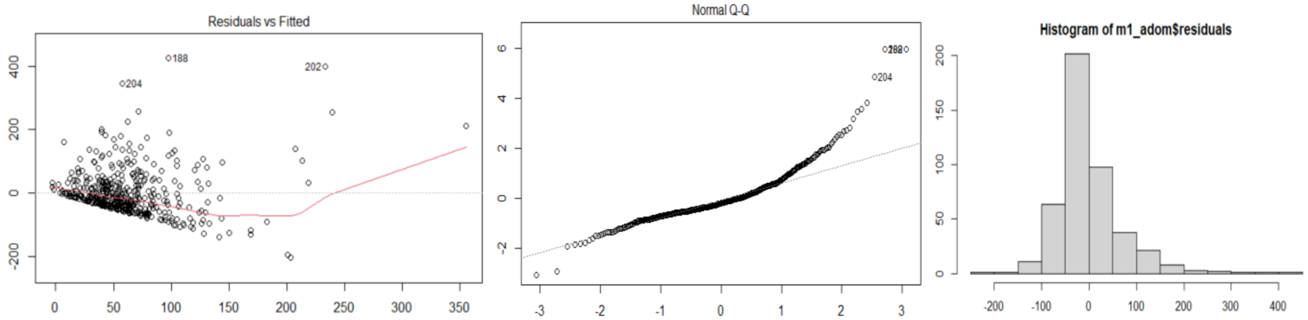
3. **Select the best model from each set and examine whether it meets the assumptions of the regression model. Which of the five regression assumptions are met for the final models?**

Test of regression assumptions for model "B" in predicting price sold:



Linearity(Pass): The residual vs fitted plot shows that the exponential model(**m2_sp**) is more or less centered around the zero line, which indicates absence of bias.
Multivariate normality (Pass): The QQ plot shows some slight deviation in normality of residuals at high and low values, however the histogram of residuals looks normal. Although, The Shapiro-Wilks test fails to reject normality. (W = 0.9919, p-value = 0.01459), the residual plot shows that data is somewhat normal.
Homoscedasticity (Fail): We don't see a fanning or other pattern in the residual plot, indicating that residuals are probably homoscedastic. Bartlett's test (Bartlett's K-squared = 399.03, df = 1, p-value < 2.2e-16) however fails, it suggests they may not be homoscedastic.
Multicollinearity (Pass): VIF analysis shows there is no multicollinearity as all variables are under 5.
Independence (Fail): Durbin-Watson test (DW = 1.7628, p-value = 0.004628) shows that the data has autocorrelation as value is not close 2.

Test of regression assumptions for the model "A" in predicting agent days on market:



Linearity(Fail): The residual vs fitted plot shows that the model(**m1_adom**) is not centered around the zero line, which clearly indicates the presence of bias.
Multivariate normality (Fail): The QQ plot shows a major deviation in normality of residuals at high values, however the histogram of residuals looks somewhat normal. The Shapiro-Wilks rejects normality. (W = 0.87442, p-value < 2.2e-16).
Homoscedasticity (Fail): We see a somewhat a fanning pattern in the residual plot, indicating that residuals are probably heteroscedastic. Bartlett's test (Bartlett's K-squared = 4206, df = 1, p-value < 0.2e^16) fails, it suggests they may not be homoscedastic. Levene test is more appropriate here because Bartlett is sensitive to violations of normality but levene test library isn't working
Multicollinearity (Fail): VIF analysis shows there is multicollinearity for listprice(GVIF = 8.955), sqft(GVIF=8.479)
Independence (Fail): Durbin-Watson test (DW = 1.8397, p-value = 0.03703) shows that the data has autocorrelation.

4. **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. (2 points)**

Top 3 predictors of price sold

Predictor	**Marginal effect**
Baths(fullbaths+bathhalf/2)	The marginal effect of baths suggests that adding 1 bathroom increases the selling price of the house by 20.44% keeping all other variables constant.
Private pool	The marginal effect of categorical variable Pool suggests that selling price for houses with private pool will be 9.85% more than houses with access to community pool.
Bed	The marginal effect of the variable bed suggests that adding 1 bedroom increases the selling price of a house by 9.19% keeping all other variables constant.

Top 3 predictors of agent days on market

Predictor	**Marginal effect**
Private and community pool	The marginal effect suggests that houses with access to both private & community will have their adom decrease by 17.724 days less than compared to houses with access to just community pools.
garages	The marginal effect of garages says that increasing the capacity of garage by 1 would bring down the adom by 17.07 days.
beds	The marginal effect of beds says that increasing the number of bedrooms by 1 would bring down the adom by 13.32 days.