Hunter Green Home Sales Analysis

Create a table of relevant predictors, hypothesized direction of effect (+/-), and rationale for each hypothesized effect. (2 points)

Model 1 Dependent Variable: adom_agentdaysonmarket

Relevant Independent variables

Predictors	Effect	Rationale
Beds	+	Number of beds will increase the value of the house leading to the quicker sale and less ADOM.
bathsfull	+/-	Number of full bathrooms depend upon number of beds and SQFT area. For instance, People generally prefer 2 full bathrooms for every three bedroom house. Hence too many bathrooms may reduce ADOM or too less bathrooms can increase ADOM.
bathsfull	+/-	Number of half bathrooms depend upon number of beds and SQFT area. It affects ADOM indirectly and hence can decrease or increase ADOM.
Sqft	+	Large SQFT area can lead to large house which can easily attract buyers and hence can result in less ADOM value.
garages	+	Number of vehicles that could be parked in the garage can enhance value of the house making it more attractive to the customers and hence can decrease ADOM value
Roof	+/-	House value can increase or decrease based on roof type which depends upon location and climate of the region. For instance in Florida, Tile roofs cost more as they can withstand storms hence can be preferred by the people resulting into decrease of ADOM.
Lotsqft	+	Large Lot SQFT area can lead to larger house area which can easily attract buyers and hence can result in less ADOM value.
Yrblt	+/-	Recently built houses can lead to less depreciation cost but there might be some cases where recently built house are not maintained properly. Hence 'yrlblt' can increase or decrease ADOM.
Pool	+/-	Availability of pool can have positive or negative effect on ADOM. As some people may prefer Private pool or some may prefer Community pool in Florida.
listprice	+/-	Lower list price can attract customers and hence can decrease ADOM value or vice versa.
cdom_cumuldaysmls	+	Lower Cumulative days will lead to lower ADOM are both directly related.
lppersqft	+/-	Lower list price per SQFT can attract customers and hence can decrease ADOM value or vice versa. It is indirectly related to ADOM.
sppersqft	+/-	Lower list price per SQFT can attract customers and hence can decrease ADOM value.
splsale	+/-	Generally distressed property gets sold for less market price. In some cases people prefer non special sale houses due to legalities. Hence splsale can increase or decrease ADOM value.

Model 2 Dependent Variable: pricesold

Relevant Independent variables

Predictors	Effect	Rationale
Beds	+	Number of beds can increase the value of the house resulting into higher price sold.

Bathsfull	+/-	Number of full bathrooms depend upon number of beds and SQFT area. For instance, People generally prefer 2 full bathrooms for every three bedroom house. Hence large number of bathrooms can affect price of the house positively or negatively.
Bathshalf	+/-	Number of half bathrooms depend upon number of beds and SQFT area. People generally prefer 3/4 bathroom. Hence half bathrooms can impact price sold negatively or positively.
Sqft	+	Large SQFT area can lead to large house which can easily attract buyers and hence can result in higher price sold.
Garages	+	Number of garages can enhance value of the house making it more attractive to the customers and hence can increase price sold.
Roof	+/-	Impact of roofs on the house price depends on the location and climate of the region. For instance in Florida Tile roofs are preferred more than Shingle as they can withstand storms. Hence roofs can decrease or increase price sold.
Lotsqft	+/-	A decent Lot SQFT area can lead to large house area which can easily attract buyers and hence can result in higher price sold. Too big lot sizes of the house as compared to its neighbourhoods can impact price sold negatively.
Yrblt	+/-	Recently built house has less depreciation cost and can have higher price sold except some conditions where house is not properly maintained.
Pool	+/-	House with private pool may have high price sold than house with community pool. Hence pool can impact price sold positively or negatively.
Listprice	+/-	List price can affect price sold directly. Impact of list price depends upon the location and other features of the house and hence can decrease or increase the price sold.
adom_agentdaysonmarket	+/-	Large ADOM value can increase or decrease price sold based on the listing date of the house.
cdom_cumuldaysmls	+/-	Large CDOM value can increase or decrease price sold based on the listing date of the house.
Lppersqft	+	Large list price per SQ can attract customers resulting into increased price sold.
Sppersqft	+	Large sales price per SQ can attract customers resulting into increased price sold.
Splsale	+/-	Special Sale can impact price of the house in positive or negative ways as people may prefer buying them because of their less cost or may not prefer buying them due to legalities.

1. Run a set of three reasonable models. Copy and paste the R code for the three models and the combined output using stargazer. (3 points)

Model 1 Dependent Variable: adom_agentdaysonmarket

Model 1.1

```
adom.sample.1=lm(adom_agentdaysonmarket~cdom_cumuldaysmls+beds+bathshalf+bathsfull+sqft+garages+roof+lotsqft+yrblt+pool+listprice+lppersqft+sppersqft+splsale+I(listprice^2)+I(pricesold^2)+I(lppersqft^2)+I(sppersqft^2),data=clean.dataset)
```

Model 1.2

```
adom.sample.2=lm(adom_agentdaysonmarket~sqft+cdom_cumuldaysmls+lotsqft+yrblt+listprice+pricesold+lppersqft+sppersqft+splsale+listprice:pricesold+I(listprice^2)+I(pricesold^2)+I(lppersqft^2)+I(sppersqft^2)+I(datedifference^2),data=clean.dataset)
```

Model 1.3

```
adom.sample.3=lm(adom_agentdaysonmarket~cdom_cumuldaysmls+beds+bathstotal+sqft+garages+roof+lotsqft+yrblt+pool+listprice+pricesold+lppersqft+sppersqft+splsale+I(splsale^2)+I(listprice^2)+I(pricesold^2)+I(lppersqft^2)+I(sppersqft^2),data=clean.dataset)
```

	Dependent variable:		
	adom_agentdaysonmarket		
	(1)	(2)	(3)
cdom_cumuldaysmls	0.690*** (0.023)	0.694*** (0.023)	0.691*** (0.023)
beds	-1.152 (3.831)		0.037 (3.772)
bathshalf	-2.336 (4.543)		
bathsfull	3.439 (5.380)		
bathstotal			-0.315 (4.171)
sqft	-0.016 (0.025)	-0.018 (0.025)	-0.011 (0.025)
garages	-2.189 (4.560)		-1.402 (4.508)
roof	2.390** (1.102)		2.538** (1.105)
lotsqft	-0.001 (0.001)	-0.0004 (0.001)	-0.001 (0.001)
yrblt	-0.640 (0.790)	-0.575 (0.751)	-0.577 (0.789)
pool	2.922 (2.439)		3.042 (2.455)
listprice	-0.00004 (0.0002)	0.0002 (0.001)	0.0003 (0.001)
pricesold		-0.0002 (0.001)	-0.0004 (0.001)
lppersqft	2.476 (2.091)	1.967 (3.404)	1.544 (3.197)
sppersqft	-1.901 (2.420)	-1.264 (3.821)	-0.562 (3.608)
splsale	8.782 (6.560)	10.669 (6.882)	-11.684 (35.458)
I(splsale2)			3.954 (6.695)
I(listprice2)	-0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)
I(pricesold2)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
I(lppersqft2)	-0.007 (0.008)	-0.006 (0.009)	-0.006 (0.008)
I(sppersqft2)	0.004 (0.010)	0.002 (0.011)	0.001 (0.010)
I(datedifference2)		-0.0001 (0.0002)	
listprice:pricesold		-0.000 (0.000)	
Constant	-7.885 (75.853)	-12.264 (76.412)	-12.663 (80.247)
Observations	482	482	482
R2	0.775	0.772	0.775
Adjusted R2	0.767	0.765	0.766
Residual Std. Error	38.833 (df = 463)	38.994 (df = 466)	38.908 (df = 462)
F Statistic	88.800*** (df = 18; 463)	105.233*** (df = 15; 466)	83.763*** (df = 19; 462)
Note:			*p<0.1; **p<0.05; ***p<0.01

Model 1.1 has lowest Residual Standard Error value that is 38.833. Moreover, looking at the F statistic value and adjusted R square value. I believe, model 1.1 is the best model and hence I have answered following questions considering model 1.

Model 2 Dependent Variable: pricesold

Model2.1

pricesold.sample.1=lm(pricesold~beds+bathshalf+bathsfull+sqft+garages+roof+lotsqft+yrblt+pool+listprice+cdom_cumuldaysmls+lppersqft+sppersqft+splsale+I(lppersqft^2)+I(sppersqft^2)+I(cdom_cumuldaysmls^2)+beds:bathstotal+sqft:lotsqft+lppersqft:sppersqft,data=clean.dataset)

Model 2.2

pricesold.sample.2=lm(pricesold~beds+bathshalf+bathsfull+sqft+garages+roof+lotsqft+yrblt+pool+listprice+cdom_cumuldaysmls+lppersqft+sppersqft+splsale+I(lppersqft^2)+I(sppersqft^2)+beds:bathstotal+sqft:lotsqft,data=clean.dataset)

Model 2.3

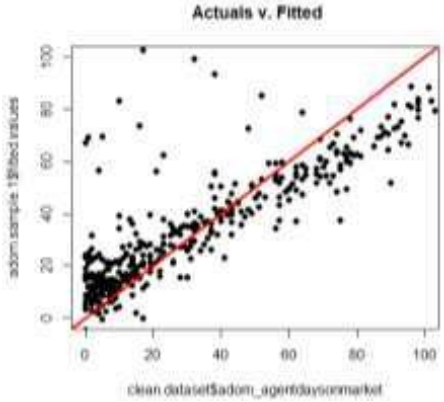
pricesold.sample.3=lm(pricesold~beds+bathshalf+bathsfull+sqft+garages+roof+lotsqft+yrblt+pool+listprice+adom_agentdaysonmarket+cdom_cumuldaysmls+splsale+I(sqft^2)+I(lotsqft^2)+I(lppersqft^2)+I(sppersqft^2)+cdom_cumuldaysmls:adom_agentdaysonmarket,data=clean.dataset)

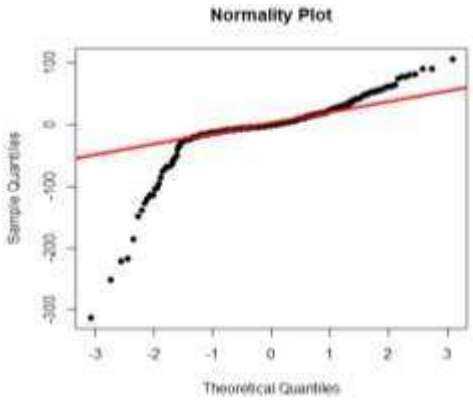
	Dependent variable:		
	pricesold		
	(1)	(2)	(3)
beds	373.892 (978.190)	569.352 (1,123.729)	-966.346** (376.426)
bathshalf	1,840.105 (1,296.764)	2,473.570* (1,488.869)	251.474 (448.586)
bathsfull	1,114.950 (1,389.823)	1,757.267 (1,594.931)	430.241 (518.148)
sqft	12.418*** (1.746)	15.553*** (1.979)	23.519*** (1.668)
garages	449.406 (473.036)	369.966 (539.263)	-54.967 (446.915)
roof	68.312 (116.005)	6.588 (132.958)	-135.091 (108.711)
lotsqft	0.533*** (0.143)	0.682*** (0.161)	-0.513*** (0.123)
yrblt	96.479 (81.718)	105.510 (93.847)	88.952 (77.769)
pool	-123.279 (253.212)	-337.072 (289.606)	-323.678 (239.744)
listprice	0.867*** (0.013)	0.841*** (0.015)	0.915*** (0.012)
adom_agentdaysonmarket			3.479 (5.398)
cdom_cumuldaysmls	10.643** (4.573)	-1.478 (2.622)	1.734 (3.980)
lppersqft	-547.715*** (200.096)	-406.511* (229.565)	
sppersqft	342.970 (228.644)	247.984 (262.614)	
splsale	-445.271 (681.391)	-138.535 (782.341)	36.643 (646.133)
I(sqft2)			-0.003*** (0.0002)
I(lotsqft2)			0.00002*** (0.00000)
I(lppersqft2)	-48.049*** (3.553)	-7.412*** (0.865)	-9.209*** (0.211)
I(sppersqft2)	-36.150*** (4.039)	9.876*** (0.986)	10.464*** (0.168)
I(cdom_cumuldaysmls2)	-0.035*** (0.010)		
beds:bathstotal	-179.105 (295.307)	-239.711 (339.187)	
sqft:lotsqft	-0.0001** (0.00003)	-0.0001*** (0.00004)	
lppersqft:sppersqft	86.499*** (7.409)		
adom_agentdaysonmarket:cdom_cumuldaysmls			-0.006 (0.012)
Constant	-7,575.907 (6,695.488)	-17,366.980** (7,637.594)	-27,680.370*** (2,981.077)
Observations	482	482	482
R2	0.999	0.999	0.999
Adjusted R2	0.999	0.999	0.999
Residual Std. Error	3,983.436 (df = 461)	4,577.894 (df = 463)	3,781.577 (df = 463)
F Statistic	34,957.120*** (df = 20; 461)	29,402.490*** (df = 18; 463)	43,101.280*** (df = 18; 463)
Note:			*p<0.1; **p<0.05; ***p<0.01

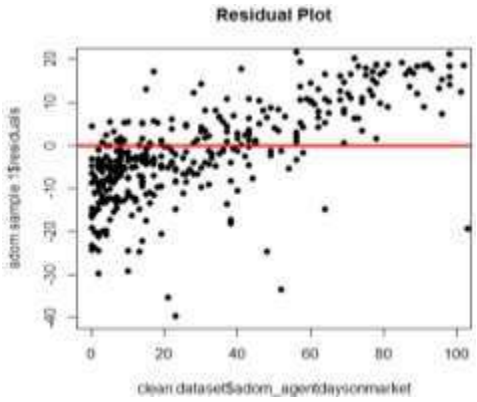
Model 2.3 has lowest Residual Standard Error value that is 3781.577. Moreover, looking at the F statistic value and adjusted R square value. I believe, model 2.3 is the best model and hence I have answered following questions considering model 3.

2. 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? (2 points)

Model 1 Dependent Variable: adom_agentdaysonmarket







Linearity **Normality** **Equality of Variances**

Linearity: From the plot we can infer that there is linear relationship between Actual and fitted values in the model.

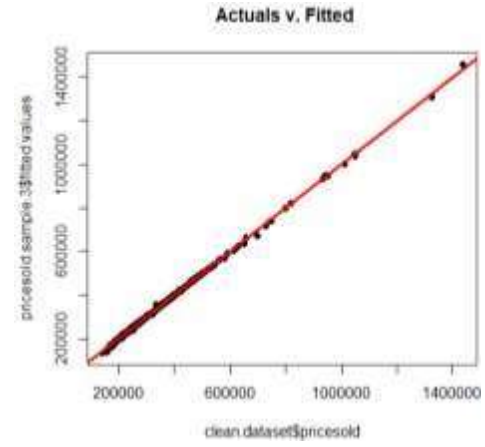
Normality: From the Normality plot we can see some points which deflect from the line at the lower and upper tail. I performed **Shapiro-Wilk** normality test where I got w = 0.715 and P value 2.2e-16 which is less than 0.05. Hence we can reject null hypothesis and infer that data/residuals are not normally distributed.

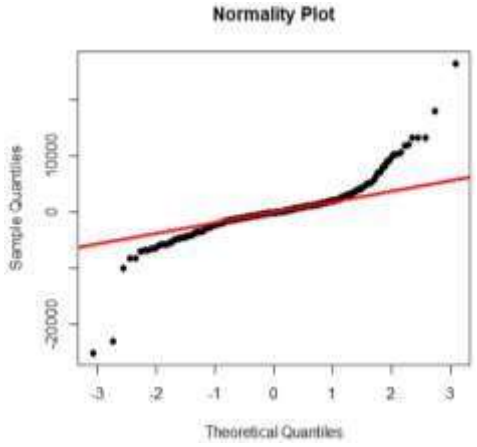
Equality of Variances/Homoscedasticity: Performed **Bartlett's test** and got Barlett's K-squared =173.7 and p value as 2.2e-16. Thus we can reject NULL hypothesis and infer that Homoscedasticity assumption does not hold true.

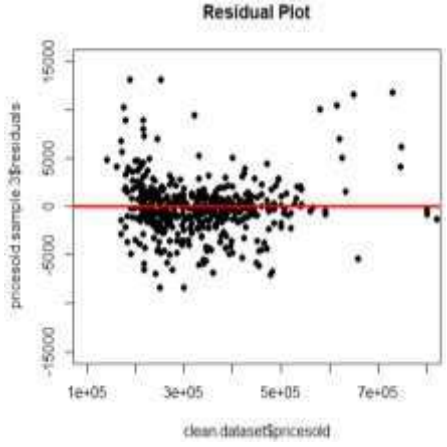
Autocorrelation: Performed **Durbin-Watson** test of autocorrelation and got DW = 2.1411. Here DW is in the range of 2 hence we can infer that residuals are not linearly auto correlated.

Multicollinearity: The predictor variables sqft, listprice, lppersqft, sppersqft have high VIF values. Hence we can infer that there is strong evidence of multicollinearity.

Model 2 Dependent Variable: pricesold







Linearity **Normality** **Equality of Variances**

Linearity: From the plot we can infer that there is linear relationship between Actual and fitted values in the model.

Normality: From the Normality plot we can see some outliers at the lower and upper tail.. I performed **Shapiro-Wilk** normality test where I got w = 0.819 and P value 2.2e-16 which is less than 0.05. Hence we can reject null hypothesis and infer that data/residuals is not normally distributed.

Equality of Variances/Homoscedasticity: From the plot we can see that there is no pattern around the line. However, I perfo rmed **Bartlett's test** and got Barlett's K-squared =3139.2 and p value as 2.2e-16. Thus we can reject NULL hypothesis and infe r that Homoscedasticity assumption does not hold true.

Autocorrelation: Performed **Durbin-Watson** test of autocorrelation and got DW = 1.7443. Here DW is in the range of 2 hence we can infer that residuals are not linearly auto correlated.

Multicollinearity: The predictor variables sqft, lotsqft and listprice have high VIF values. Hence we can infer that there is strong evidence of multicollinearity.

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 significance is not important.

1) adom

Top Predictor	Marginal Effect
cdom_cumuldaysmls	I day increase in CDOM will increase ADOM days by 0.68 unit. This variable is significant in the model
sqft	I unit increase in SQFT will decrease ADOM days by 0.0155 unit.
roof	Roof is directly related to ADOM. ADOM can marginally change by 2.29 as per the variations in the roof type.

2) pricesold

Top Predictor	Marginal Effect
sqft	1 unit increase in SQFT will increase price sold by 23.52
lotsqft	1 unit increase in Lot SQFT will have marginal effect of 0.51 on price sold.
listprice	1 unit increase in list price will increase price sold by 0.915

References:

https://armls.com/difference-adom-cdom

https://pocketsense.com/much-new-bathroom-increase-value-house-1435.html

https://support.brightmls.com/s/article/The-Difference-Between-DOM-and-CDOM