Case Study for House Sale Price

-- Winnie Fan

1) In each neighborhood of Ames, what is the median sale price for homes sold in 2006 which have an indoor square footage of greater than or equal to 2000 ft. (excluding porches, garages, decks, and veneers)

Answer:

1.1) Objective: find the median sale price in each neighborhood of Ames

(median SalePrice grouped by neighborhood of Ames)

s.t. (1) homes sold in 2006 (2) indoor square footage ≥ 2000 ft:

Assume indoor square footage = Above grade living area square feet (GrLivArea)

+ Total square feet of basement area (TotalBsmtSF)

– porches (OpenProchSF+EnclosedPorch+3SsnPorch+ScreenPorch)

– garages (GarageArea) – decks (WoodDeckSF) – veneers (MasVnrArea)

1.2) Steps to get the final the data table below using Python:

(1) Read the csv file ‘data.csv’ and filter the data to only have the homes sold in 2006, resulting in a dataframe called ‘df2006’; then create two new columns for ‘df2006’ named ‘PorchSF’ and ‘IndoorSF’

(2) Filter the dataset ‘df2006’ by the values in the column of ‘IndoorSF’ being greater than or equal to 2000 sf to get a subset named ‘sub_df’

(3) Group the ‘sub_df’ gotten from (2) by ‘Neighborhood’ and get the median sale price for each group.

Table 1 Median House Sale Price by Neighborhood

Neighborhood	Median_SalePrice
SWISU	129250.0
Edwards	136900.0
NAmes	145125.0
OldTown	148500.0
BrkSide	149000.0
MeadowV	151400.0
Mitchel	170000.0
Sawyer	170000.0
Gilbert	210950.0
Blmngtn	215000.0
CollgCr	217000.0
SawyerW	220000.0
Crawfor	223000.0
ClearCr	225000.0
NWAmes	235000.0
Veenker	270000.0
NoRidge	279000.0
Somerst	290000.0
NridgHt	328821.5
Timber	335000.0
StoneBr	361919.0

2) A client approaches you with a question about the local housing market. They're interested in whether more homes are sold at certain times of year than others? In other words, is there seasonality? Provide a visualization and a brief description of your findings. (For the visualization, you may use R plots, Python's matplotlib, Tableau, Excel charts, or whatever tool you like).

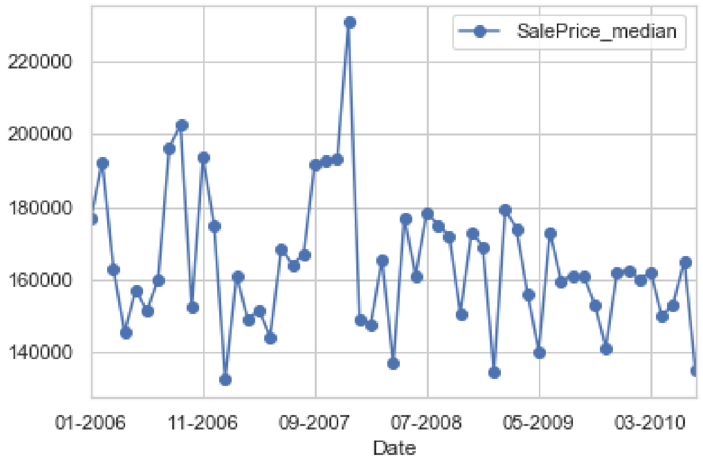
Answer:

Here median sale price is used for comparison.

2.1) In order to get a general idea about the data, we aggregate the dataset to get time series dataset spanning from 01/2006 to 07/2010.

From figure 1 below, we can see the sale price dataset for homes has seasonality by year, and the decomposition plots in figure 2 also confirm that.

Figure 1 Median Sale Price across Time



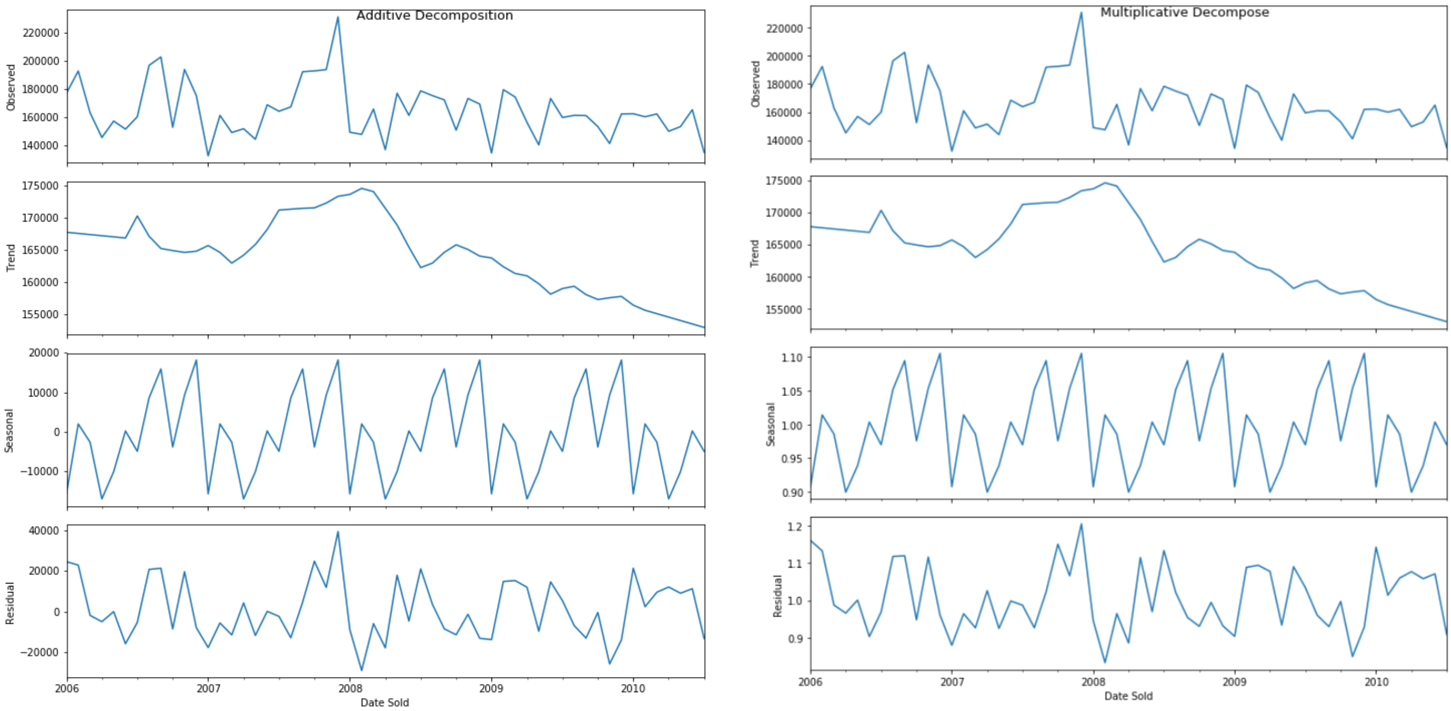
Additive time series:

$$Value \ = \ Base\ Level \ + \ Trend \ + \ Seasonality \ + \ Error$$

Multiplicative Time Series:

$$Value \ = \ Base\ Level * Trend * \ Seasonality * \ Error$$

Figure 2 Additive and Multiplicative Decomposition



2.2) Since the client’s request is about the local housing market, in the following we focus on the house sale price for each neighborhood. And figure 3 to 6 respectively plots the minimum, maximum, average, and median sale prices for homes by neighborhood.

However, the regional datasets don’t show obvious seasonality.

(1) The potential issue is that for each region, the dataset is irregular time series that its sequence of data points is not in the successive order.

For example, table 2 below is the dataset showing the median sale price by date sold for Blmngtn. However, the time interval between two observations are not equal.

Table 2 Median Sale Price for Blmngtn by Date Sold

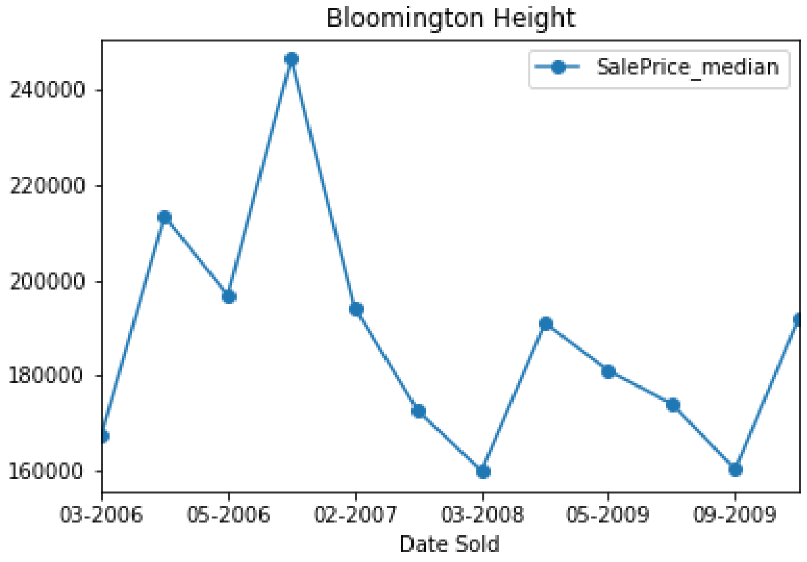
Date_Sold	SalePrice_median
03-2006	167240.0
04-2006	213490.0
05-2006	196870.0
10-2006	246578.0
02-2007	194201.0
06-2007	172500.0
03-2008	159895.0
05-2008	191000.0
05-2009	181000.0
06-2009	174000.0
09-2009	160200.0
04-2010	192000.0

And from the plots below (figure 3 to 6), strictly speaking, there’s no seasonality for regional markets.

(Seasonality means there is a general systematic linear or (most often) nonlinear component that changes over time and does repeat.)

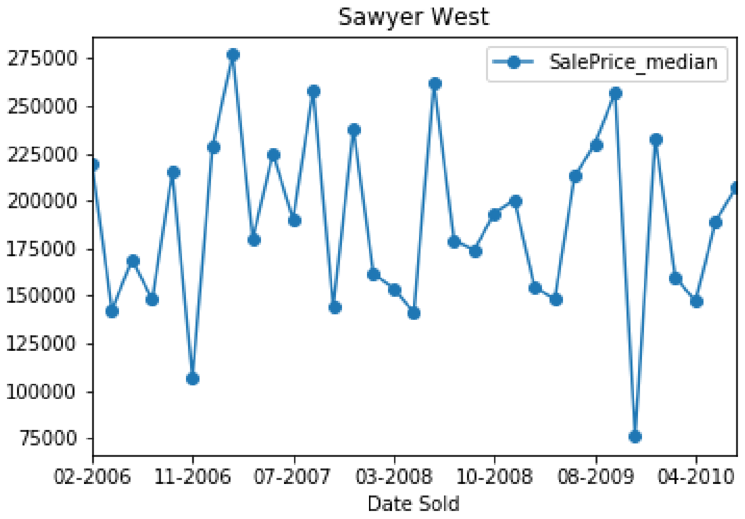
(2) Here we take several neighborhood regions as examples for analysis

(i) Bloomington Height (Blmngtn) Median Sale Price by Date Sold



In general, the house sold price tends to experience an increase between March to May which might due to it approaching the beginning of a new school semester.

(ii) Sawyer West (SawyerW) Median Sale Price by Date Sold



The time period of the dataset for Sawyer West is longer than Blmngtn with more obvious tendency to check. The prices between May to August, and December or January are roughly a bit higher than other months in a year and the reason mainly is that people want to be settled before school starts.

Figure 3 Minimum House Sale Price by Neighborhood and Date Sold

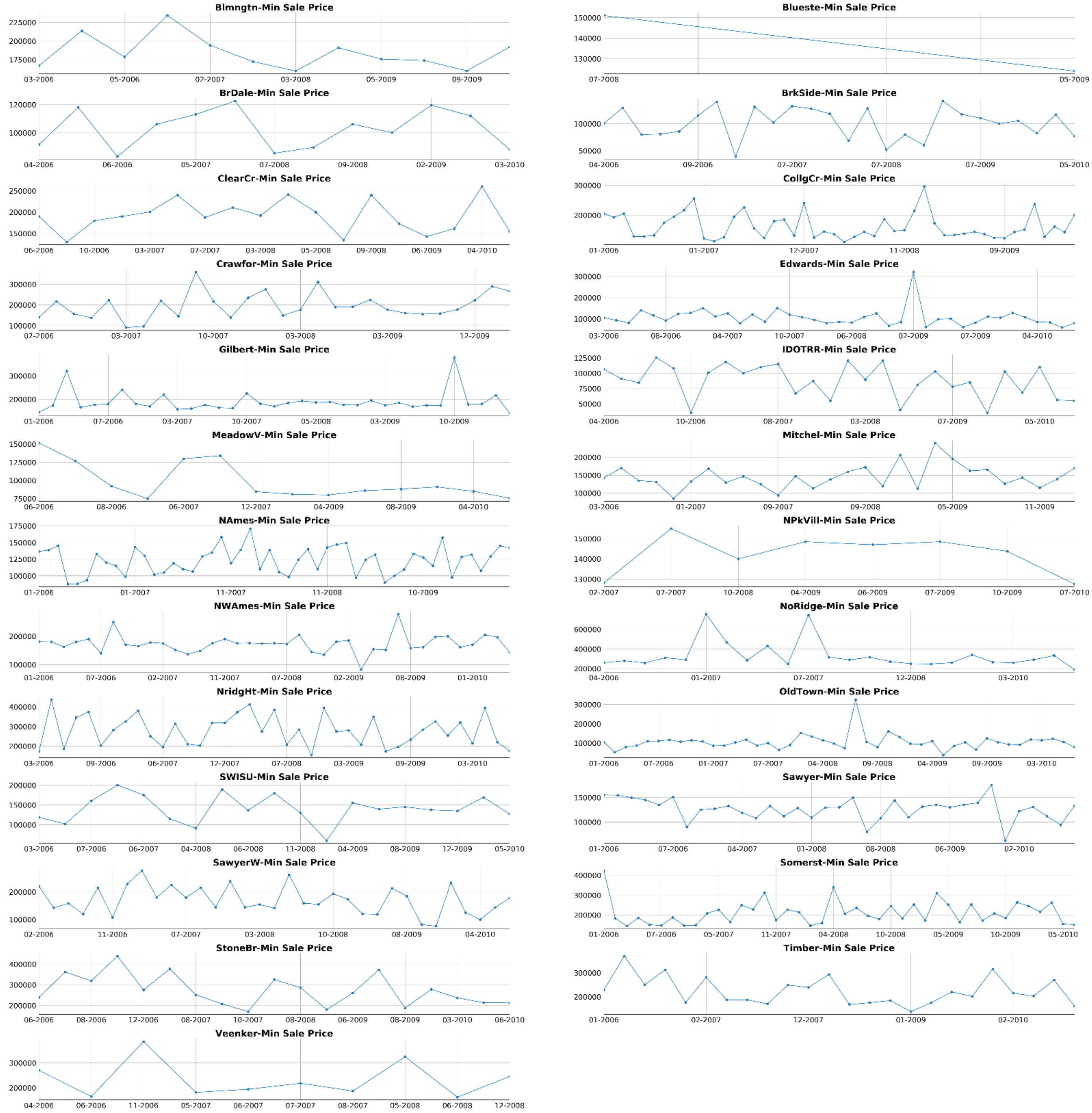


Figure 4 Maximum House Sale Price by Neighborhood and Date Sold

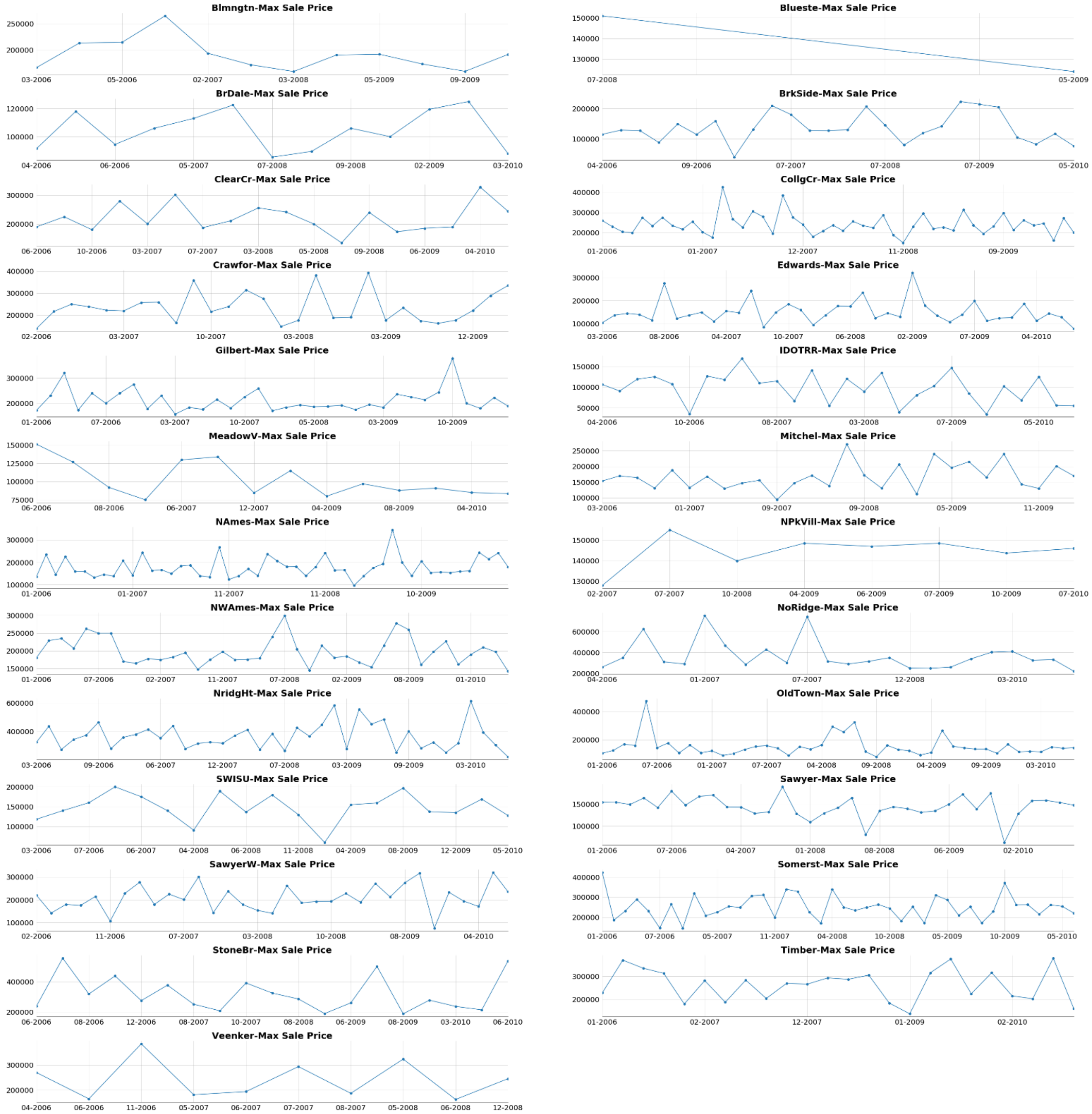
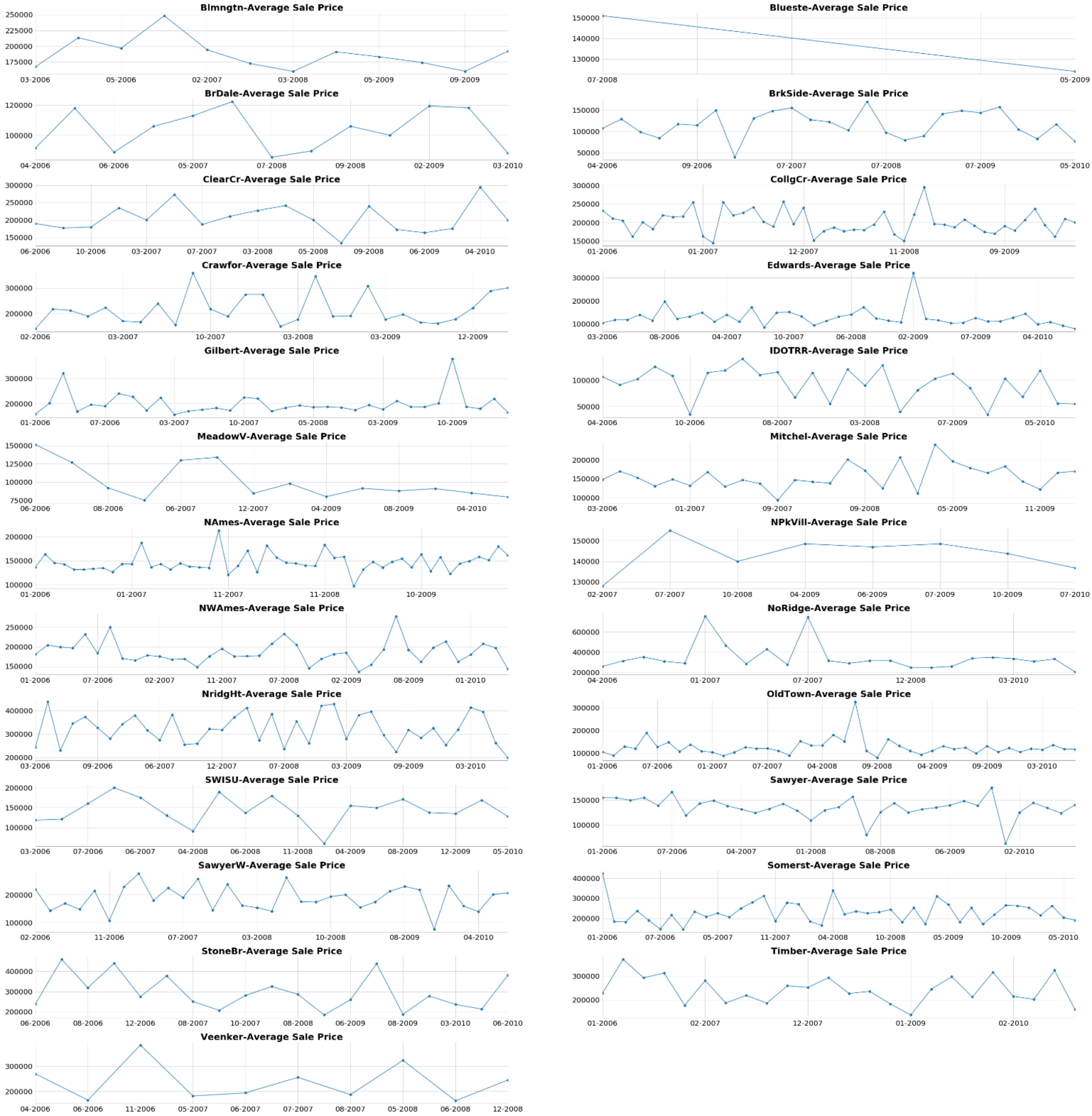


Figure 5 Average House Sale Price by Neighborhood and Date Sold



Blmngtn-Average Sale Price

Blueste-Average Sale Price

BrDale-Average Sale Price

BrkSide-Average Sale Price

ClearCr-Average Sale Price

CollgCr-Average Sale Price

Crawfor-Average Sale Price

Edwards-Average Sale Price

Gilbert-Average Sale Price

IDOTRR-Average Sale Price

MeadowV-Average Sale Price

Mitchel-Average Sale Price

NAmes-Average Sale Price

NPkVill-Average Sale Price

NWAmes-Average Sale Price

NoRidge-Average Sale Price

NridgHt-Average Sale Price

OldTown-Average Sale Price

SWISU-Average Sale Price

Sawyer-Average Sale Price

SawyerW-Average Sale Price

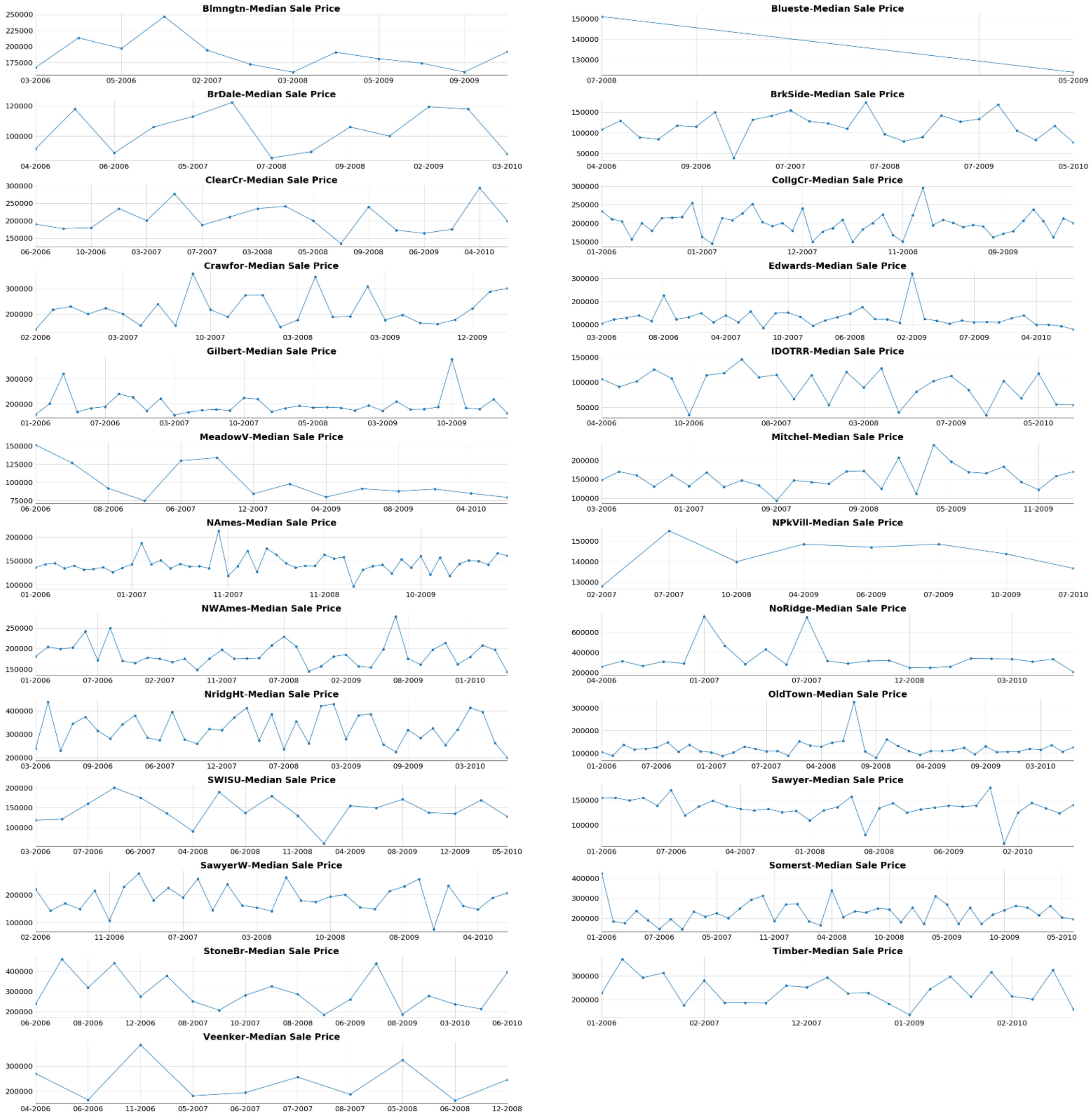
Somerst-Average Sale Price

StoneBr-Average Sale Price

Timber-Average Sale Price

Veenker-Average Sale Price

Figure 6 Median House Sale Price by Neighborhood and Date Sold



3) You're a contractor consulting for a client who wants to remodel their home and then sell it on the market. The home the client occupies is a 3 bedroom / 2 bathroom 1500 sq. ft. house. They're deciding between the following options for the remodel:

a) Adding a new bedroom measuring 130 sq. ft.

b) Adding a new half bathroom measuring 80 sq. ft.

c) Expanding the living room by 400 sq. ft.

Assume that the cost of all three remodel options is equal. Based on the data provided, which option do you think will provide the greatest predicted increase in home value and why? What other information would you seek out that might help you make the decision? Please provide any visualizations, tables, etc. to support your findings. Note: we're not looking for one "right answer" here; it's more important to explain your reasoning and the limits of how this data can inform this decision.

Answer:

This analysis consists of two parts: 3.1) is the analysis from the dataset itself and 3.2) uses statistical modeling to predict home value

3.1) First, analysis from the dataset itself.

3.1.1) First, assume the 1500 sq. ft. is the indoor square footage as 'IndoorSF' in the dataset (using the same calculation formula as the question 1) above) and here we use median sale price for comparison.

Based on three options, filter the dataset by the following constraints:

a) (i) The 'BedroomAbvGr' is changed to be 4 (assume currently 3 bedrooms are all above grade)

   (ii) The 'IndoorSF' is changed to be greater than or equal to (1500+130 = 1630)

   (iii) The 'FullBath' is not changed (=2)

b) (i) The 'BedroomAbvGr' is not changed (=3) (assume currently 3 bedrooms are all above grade)

   (ii) The 'IndoorSF' is changed to be greater than or equal to (1500+80 = 1580)

   (iii) The 'HalfBath' is changed to be 1 (assume currently no HalfBath available)

   (iv) The 'FullBath' is not changed. (=2)

c) (i) The 'BedroomAbvGr' is not changed (=3) (assume currently 3 bedrooms are all above grade)

   (ii) The 'IndoorSF' is changed to be greater than or equal to (1500+400 = 1900)

   (iii) The 'FullBath' is not changed. (=2)

3.1.2)

(1) Generally, from table 3 below, we can see the median sale price for option3 is the highest among those three options based on aggregated historical data

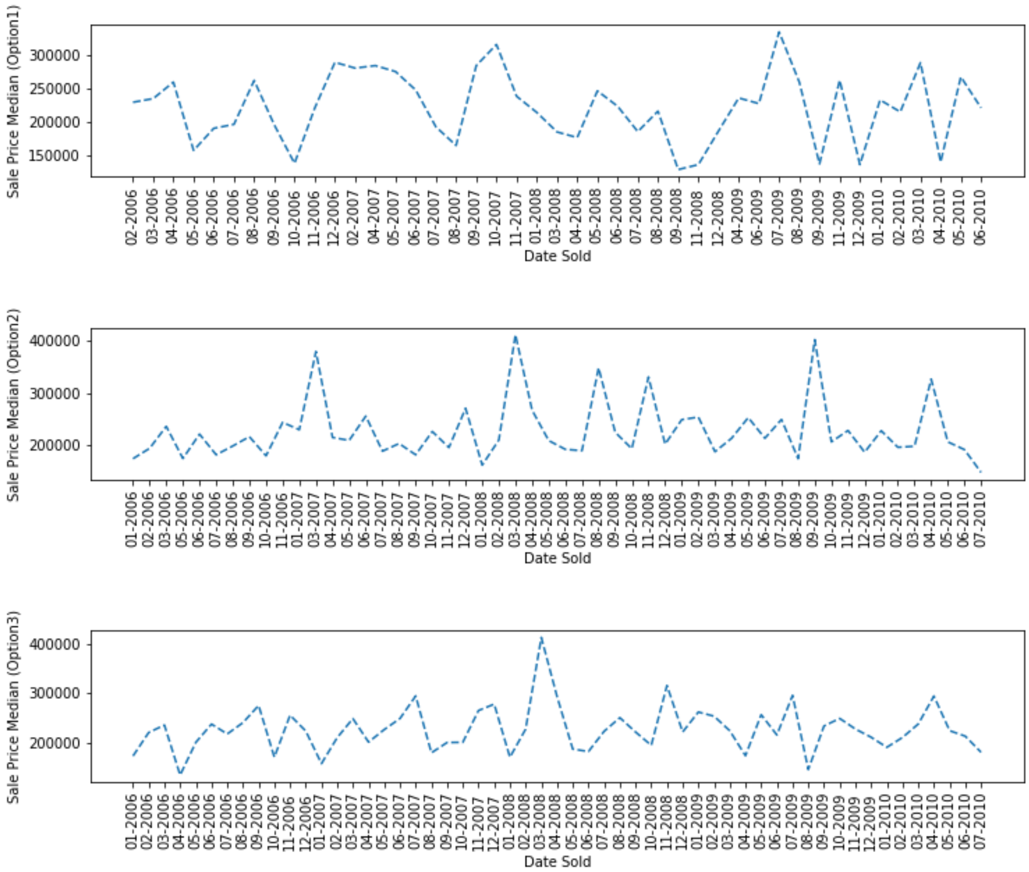
Table 3 Statistics of Sale Price for Different Options

Index	SalePrice_option1	SalePrice_option2	SalePrice_option3
min	87000.00	129000.00	93000.00
max	475000.00	451950.00	451950.00
mean	222106.01	224668.36	233050.17
median	219355.00	205000.00	225000.00

(2) However, without knowing the location of that client's house and the time when he or she wants to sell that house, the dataset is aggregated respectively by region and date to provide reference regarding which option is the best.

(2.1) Firstly, we group the dataset by date to check the changes of median sale price by time

Figure 7 Median Sale Price by Date Sold for Different Options



Furthermore, the dataset is aggregated to show the median sale price for each month.

It shows that for Jan. and Feb., option 1 would increase the house's sale price the most, and in Mar., option 3 is a better choice than others, etc.

The best option might change by month.

Table 4 Median Sale Price by Month for Different Options

MoSold	SalePriceMedian_option1	SalePriceMedian_option2	SalePriceMedian_option3
1	223500.00	222250.00	180450.00
2	229500.00	201500.00	223350.00
3	192000.00	208950.00	231000.00
4	236000.00	228750.00	213500.00
5	259090.00	209250.00	220000.00
6	221000.00	200000.00	231300.00
7	204950.00	187500.00	232615.00
8	220000.00	212750.00	196500.00
9	145900.00	219250.00	230000.00
10	202950.00	194500.00	197320.50
11	230000.00	227000.00	257000.00
12	167500.00	234000.00	239000.00

(2.2) Then, we group the dataset by region to check the changes of median sale price by neighborhood.

Figure 8 Median Sale Price by Neighborhood for Different Options

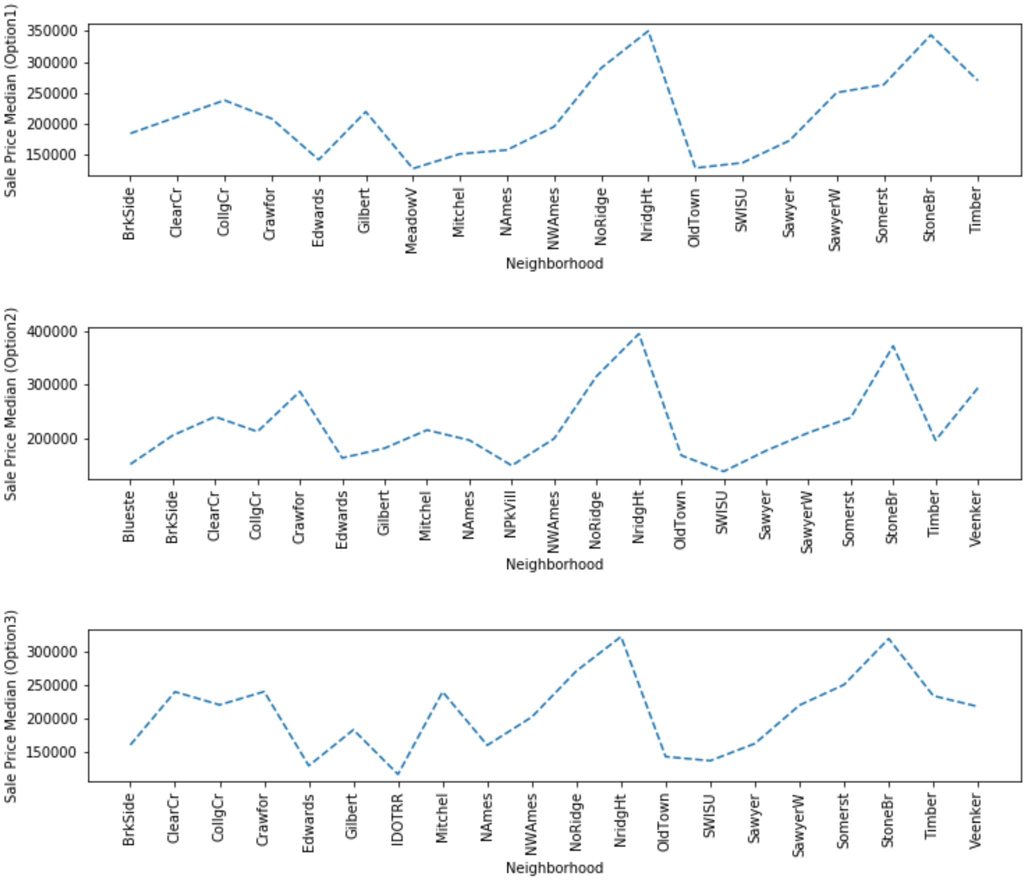


Table 5 is the median sale price for each neighborhood.

It shows that for BrkSide and ClearCr, option 2 would increase the house's sale price most, and for CollgCr, option 1 is better than others, etc.

The best option might change by region.

Table 5 Median Sale Price by Neighborhood for Different Options

Neighborhood	SalePriceMedian_option1	SalePriceMedian_option2	SalePriceMedian_option3
BrkSide	184000.00	205000.00	160950.00
ClearCr	211000.00	240000.00	240000.00
CollgCr	237700.00	212000.00	220500.00
Crawfor	208000.00	287250.00	240250.00
Edwards	141400.00	162700.00	130000.00
Gilbert	219210.00	181000.00	183500.00
Mitchel	150900.00	215000.00	240000.00
NAmes	157250.00	196000.00	160500.00
NWAmes	195000.00	198950.00	202500.00
NoRidge	290000.00	315750.00	271000.00
NridgHt	350000.00	395192.00	322000.00
OldTown	128000.00	167500.00	143500.00
SWISU	136500.00	137450.00	137450.00
Sawyer	172500.00	176250.00	163000.00
SawyerW	250140.00	210000.00	220000.00
Somerst	263000.00	238350.00	250580.00
StoneBr	343459.50	372500.00	319000.00
Timber	269500.00	195750.00	233975.00

3.2) Statistical Modeling

3.2.1) Based on the analysis in 3.1) and the purpose of simplification, only the 7 variables are selected for modeling analysis:

'SalePrice', 'Neighborhood', 'LotFrontage', 'IndoorSF', 'FullBath', 'HalfBath' 'BedroomAbvGr'

3.2.2)

(1) The dependent variable – SalePrice is transformed into logarithmic values (named as 'log_SalePrice'), which may stabilize the variance of the data. This means one unit of change in the independent variable results in a constant percentage change in SalePrice holding all other independent variables constant.

(2) Then we change 'Neighborhood' (there are 25 unique values in 'Neighborhood') to be 24 dummy variables

(3) We regress ‘log_SalePrice’ on ‘LotFrontage’, ‘IndoorSF’, ‘FullBath’, ‘HalfBath’ and regional dummies, and the model summary is as below

Table 6 Regression Results

		OLS Regression Results				
Dep. Variable:	log_SalePrice	R-squared:	0.74			
Model:	OLS	Adj. R-squared:	0.73			
Method:	Least Squares	F-statistic:	111.40			
Date:	Mon	16 Sep 2019	Prob (F-statistic):	1.90e-311		
Time:	12:09:09	Log-Likelihood:	147.12			
No. Observations:	1195	AIC:	-234.20			
Df Residuals:	1165	BIC:	-81.66			
Df Model:	29					
Covariance Type:	nonrobust					
	coef	std err	t	P>|t|	[0.025	0.975]
const	11.4792	0.0650	176.1390	0.0000	11.3510	11.6070
LotFrontage	0.0017	0.0000	5.2860	0.0000	0.0010	0.0020
IndoorSF	0.0002	0.0000	12.7980	0.0000	0.0000	0.0000
FullBath	0.1427	0.0170	8.5090	0.0000	0.1100	0.1760
HalfBath	0.1166	0.0140	8.1690	0.0000	0.0890	0.1450
BedroomAbvGr	-0.0039	0.0100	-0.4010	0.6880	-0.0230	0.0150
Blueste	-0.2420	0.1650	-1.4700	0.1420	-0.5650	0.0810
BrDale	-0.4142	0.0820	-5.0520	0.0000	-0.5750	-0.2530
BrkSide	-0.3061	0.0680	-4.5220	0.0000	-0.4390	-0.1730
ClearCr	-0.0519	0.0850	-0.6110	0.5410	-0.2190	0.1150
CollgCr	-0.0207	0.0630	-0.3310	0.7410	-0.1430	0.1020
Crawfor	-0.0282	0.0690	-0.4100	0.6820	-0.1630	0.1070
Edwards	-0.3667	0.0640	-5.6960	0.0000	-0.4930	-0.2400
Gilbert	-0.1131	0.0680	-1.6680	0.0960	-0.2460	0.0200
IDOTRR	-0.5436	0.0710	-7.6350	0.0000	-0.6830	-0.4040
MeadowV	-0.4617	0.0820	-5.6220	0.0000	-0.6230	-0.3010
Mitchel	-0.1933	0.0700	-2.7620	0.0060	-0.3310	-0.0560
NAmes	-0.2025	0.0630	-3.1990	0.0010	-0.3270	-0.0780
NPkVill	-0.2803	0.1010	-2.7710	0.0060	-0.4790	-0.0820
NWAmes	-0.1422	0.0690	-2.0730	0.0380	-0.2770	-0.0080
NoRidge	0.2556	0.0720	3.5580	0.0000	0.1150	0.3960
NridgHt	0.3248	0.0640	5.0510	0.0000	0.1990	0.4510
OldTown	-0.3600	0.0640	-5.6600	0.0000	-0.4850	-0.2350
SWISU	-0.3434	0.0760	-4.5170	0.0000	-0.4920	-0.1940
Sawyer	-0.2425	0.0690	-3.5150	0.0000	-0.3780	-0.1070
SawyerW	-0.1129	0.0670	-1.6770	0.0940	-0.2450	0.0190
Somerst	0.1070	0.0640	1.6680	0.0960	-0.0190	0.2330
StoneBr	0.3790	0.0760	4.9980	0.0000	0.2300	0.5280
Timber	0.1332	0.0710	1.8630	0.0630	-0.0070	0.2730
Veenker	0.2735	0.1010	2.7070	0.0070	0.0750	0.4720
Omnibus:	172.0600	Durbin-Watson:	1.9770			
Prob(Omnibus):	0.0000	Jarque-Bera (JB):	759.7140			
Skew:	-0.6050	Prob(JB):	0.0000			
Kurtosis:	6.7140	Cond. No.	92900.0000			

(4) Result analysis

From the results above (table 6),

The coefficient for ‘IndoorSF’ is significant at 5% significance level and 1 unit increase of the ‘IndoorSF’ will increase the sale price by 0.02% if other independent variables are constant;

The coefficient for ‘BedroomAbvGr’ is not significant;

The variables of ‘FullBath’ and ‘HalfBath’ both show significantly positive effects on sale price. Keeping other independent variables constant, 1 unit increase of the ‘FullBath’ will increase the sale price by 14.27% and 1 unit increase of the ‘HalfBath’ will increase the sale price by 11.66%.

If we only consider the effect of indoor square feet and keep other independent variables constant, option 3 will provide the greatest predicted increase in home value; if ‘HalfBath’ only, option 2 is better than others.

(Due to the time limitation, I only roughly explain a simple model here.

However, we could try more models to fully predict which option is better than others, such as decision trees (*provided by the code*), random forests (*provided by the code*), support vector machines, hierarchical bayesian model, etc)

4) You own a single-family home (i.e. BldgType = "1Fam") with 4 bedrooms that you are looking to rent out or sell. Assume you can generate a yearly rent that is 10% of the estimated sales price. Your options include:

a) Convert the home into a duplex and rent both units.

b) Rent the home as is.

c) Sell the home for market value.

Assume that cost is negligible for our purposes. Which option maximizes revenue received in 5 years? 10 years? 15 years? List out all assumptions you are making in your calculations and outline your thought process. Note: as in question 3, we're not looking for one "right answer" here; it's more important to explain your reasoning and the limits of how this data can inform this decision.

Answer:

4.1) Assume we don't consider any house-related tax and salvage value, and only cash flow is considered here.

Calculations for each option:

a) (i) Filter the dataset to have 2 bedrooms and BldgType = "1Fam" and group by neighorbood, then use the median sale price as the estimated sale price for each neighborhood;

   (ii) Multiply the estimated sale price by (0.1*2) to get yearly rent;

   (iii) Assume the discount rate is 5%. The yearly rent for the next 5, 10, 15 years is obtained from (ii). Then the net present values are calculated by

$$Net\ Present\ Value = \frac{Future\ Value}{(1 + Discount\ Rate)^n}$$

   Where $n = \ period\ number$

b) (i) Filter the dataset to have 4 bedrooms and BldgType = "1Fam" and group by neighorbood, then use the median sale price as the estimated sale price for each neighborhood;

   (ii) Multiply the estimated sale price by 0.1 to get yearly rent;

   (iii) Assume the discount rate is 5%. The yearly rent for the next 5, 10, 15 years is obtained from (ii). Then the net present values are calculated by the same formula as a)

c) Filter the dataset to have 4 bedrooms and BldgType = "1Fam" and group by neighorbood, then use the median sale price as the estimated sale price.

4.2) The results are show in table 7.

Take Brookside (BrkSide) as an example, comparing the net present value in 5 and 10 years, option 3 – sell the house by the median sale price is the best option; based on the net present value in 15 years, option 1 would maximize the revenue.

And for most of the regions, in the long term (next 15 years), option 1 and option 2 tend to bring more revenue than option 3.

We could make a more rational decision if we could get more data, such as this house's location, other detailed description about the house, economic condition, etc.

Table 7 Yearly Rent, Net Present Value, and Revenue by Region for Different Options

Neighborhood	**YrRent_opt1**	**NPV5_opt1**	**NPV10_opt1**	**NPV15_opt1**	**YrRent_opt2**	**NPV5_opt2**	**NPV10_opt2**	**NPV15_opt2**	**Revenue_opt3**
BrkSide	20700.00	94101.18	167831.91	225601.87	20500.00	93191.99	166210.34	223422.14	205000.00
ClearCr	48300.00	219569.41	391607.79	526404.36	24000.00	109102.81	194587.72	261567.38	240000.00
CollgCr	38787.90	176327.87	314485.38	422735.39	21200.00	96374.15	171885.82	231051.19	212000.00
Crawfor	27500.00	125013.64	222965.10	299712.63	28725.00	130582.43	232897.18	313063.46	287250.00
Edwards	21000.00	95464.96	170264.26	228871.46	16270.00	73962.61	131914.26	177320.89	162700.00
Gilbert	33800.00	153653.13	274044.37	368374.06	18100.00	82281.70	146751.57	197265.40	181000.00
Mitchel	26200.00	119103.90	212424.93	285544.39	21500.00	97737.94	174318.17	234320.78	215000.00
NAmes	26400.00	120013.09	214046.49	287724.12	19600.00	89100.63	158913.30	213613.36	196000.00
NWAmes	35290.00	160426.59	286125.03	384613.04	19895.00	90441.69	161305.11	216828.46	198950.00
NoRidge	65150.00	296168.68	528224.58	710046.46	31575.00	143538.39	256004.47	344124.59	315750.00
NridgHt	78923.40	358781.87	639896.85	860157.80	39519.20	179652.33	320414.63	430705.57	395192.00
OldTown	22000.00	100010.91	178372.08	239770.10	16750.00	76144.67	135806.01	182552.24	167500.00
SWISU	23100.00	105011.46	187290.68	251758.61	13745.00	62484.09	111442.01	149801.82	137450.00
Sawyer	26500.00	120467.69	214857.27	288813.98	17625.00	80122.38	142900.36	192088.55	176250.00
SawyerW	29100.00	132287.16	235937.61	317150.45	21000.00	95464.96	170264.26	228871.46	210000.00
Somerst	41660.00	189384.30	337771.85	454037.38	23835.00	108352.73	193249.93	259769.11	238350.00
StoneBr	66700.00	303214.90	540791.71	726939.35	37250.00	169336.66	302016.36	405974.38	372500.00
Timber	49390.00	224524.50	400445.31	538283.88	19575.00	88986.98	158710.61	213340.90	195750.00
Veenker	38800.00	176382.88	314583.48	422867.27	29400.00	133650.94	238369.96	320420.04	294000.00