P Discovery Project

Home Sweet Home: Using Multiple Regression to Analyze and Predict Home Prices

Data

For the full data set visit stat.hawkeslearning.com and navigate to **Discovering Business Statistics, Second Edition > Data Sets > Mount Pleasant Real Estate Data**.

An important problem in real estate is determining how to price homes to be sold. There are so many factors—size, age, and style of the home; number of bedrooms and bathrooms; size of the lot; and so on—which makes setting a price a challenging task. In this project, we will try to help realtors in this task by determining how different characteristics of homes relate to home prices, identifying the key variables in pricing, and building multiple-variable regression models to predict prices based on property characteristics.

Our analysis will be based on the Mount Pleasant Real Estate Data. This data set includes information about 245 properties for sale in three communities in the suburban town of Mount Pleasant, South Carolina, in 2017.

Phase 1: Data Preparation

1. Download the Mount Pleasant Real Estate Data from stat.hawkeslearning.com and open it with Microsoft Excel.
2. Determine the mean, mode, median, maximum, minimum, standard deviation, and the coefficient of variation of the following variables: price, number of bedrooms, number of bathrooms, number of stories, and square footage, and briefly discuss the results. (Hint: these values can be quickly calculated using the Data Analysis Add-in: Descriptive Statistics in Excel).
3. Fully summarize the qualitative variables (i.e., What percent of the sample has a pool?) and briefly discuss your findings.
4. To ensure the data contain comparable properties, eliminate duplexes and properties whose prices are outliers. What limitations does this impose on our analysis? How did you determine which prices were outliers?

   Consider the following variables associated with each property:

   x_1 = number of bedrooms
   x_2 = number of bathrooms
   x_3 = number of stories
   x_4 = subdivision
   x_5 = square footage
   x_6 = age (based on year built)
   x_7 = acreage
   x_8 = new owned

5. For the qualitative variables, adjust this data in a reasonable, quantitative way for use in a regression analysis.
6. Use the following correlation matrix and describe any issues with multicollinearity.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		*Bedrooms*	*Baths - Total*	*Baths - Full*	*Baths - Half*	*Stories*	*Subdivision*	*Square Footage*	*Age*	*Acreage*	*New Owned?*	*House Style*	*Covered Parking Spots*	*Fenced Yard*	*Screened Porch?*	*Golf Course?*	*Fireplace?*
2	Bedrooms	1															
3	Baths - Total	0.70	1.00														
4	Baths - Full	0.67	0.95	1.00													
5	Baths - Half	0.06	0.13	-0.18	1.00												
6	Stories	0.43	0.50	0.42	0.24	1.00											
7	Subdivision	0.08	-0.05	-0.04	-0.01	0.06	1.00										
8	Square Footage	0.71	0.74	0.68	0.19	0.44	0.11	1.00									
9	Age	0.05	-0.07	-0.12	0.14	-0.07	0.28	0.18	1.00								
10	Acreage	0.12	0.19	0.16	0.11	0.00	-0.03	0.35	0.33	1.00							
11	New Owned?	-0.14	0.00	0.01	-0.05	0.04	-0.29	-0.22	-0.78	-0.15	1.00						
12	House Style	-0.22	-0.22	-0.22	0.03	-0.15	0.07	-0.22	-0.06	-0.12	0.11	1.00					
13	Covered Parking Spots	0.28	0.35	0.31	0.13	0.16	0.03	0.47	0.14	0.24	-0.15	0.07	1.00				
14	Fenced Yard	0.04	-0.12	-0.13	0.04	-0.11	0.01	-0.03	0.29	-0.10	-0.41	-0.01	-0.04	1.00			
15	Screened Porch?	0.17	0.24	0.23	0.02	0.14	-0.25	0.14	-0.04	0.03	-0.05	-0.18	-0.03	0.12	1.00		
16	Golf Course?	0.30	0.26	0.27	-0.02	0.19	-0.14	0.37	0.34	0.48	-0.19	-0.08	0.23	0.12	0.13	1.00	
17	Fireplace?	0.18	0.18	0.20	-0.08	0.06	0.04	0.22	0.14	0.09	-0.13	-0.17	0.05	0.11	0.30	0.22	1.00
18	Number of Fireplaces	0.16	0.21	0.22	-0.03	0.04	0.11	0.35	0.24	0.14	-0.20	-0.13	0.15	0.13	0.28	0.17	0.80

Phase 2: Constructing Predictive Models

7. Construct the multiple regression model with input variables x_1, x_2, x_3, and x_4.

8. Examine the impact of adding additional variables to the model

- **a.** Add x_5 to the model. Is the addition of x_5 to the model significant? How was the adjusted R^2 impacted? What is the P-value for x_5?
- **b.** Add x_6 to the model. Is the addition of x_6 to the model significant? How was the adjusted R^2 impacted? What is the P-value for x_6?
- **c.** Add x_7 to the model. Is the addition of x_7 to the model significant? How was the adjusted R^2 impacted? What is the P-value for x_7?
- **d.** Add x_8 to the model. Is the addition of x_8 to the model significant? How was the adjusted R^2 impacted? What is the P-value for x_8?

9. Perform a hypothesis test to determine if the model is useful for predicting home values at a significance level of $\alpha = 0.05$. State the P-value and interpret its meaning.

10. Are any variables not useful predictors of home price at a significance level of $\alpha = 0.05$? State the P-values of these variables. Intuitively, what does this mean with respect to pricing properties?

11. State the best model for the data and justify your answer.

Phase 3: Applying and Interpreting the Model

12. Suppose you own a 2000 square foot 2-story house in one of the communities in the data set with 3 bedrooms, 2.5 baths, a pool, and it is located on a golf course, but has no dock or fenced yard. What does the model predict the price of your house to be?

13. A common term in real estate is "comparables," or "comps" for short, which are properties that have similar characteristics. It is common for realtors to look up "comps" for a certain property to get an idea of how to price it. Locate the "comps" for your home in the data set. Create a box plot of the "comps" and estimate a price range for your house on this basis.

14. What advantages and disadvantages does this approach have to the multiple regression model above?