Initial data expectations

My analysis will be based on a database of home sales, where an individual home sale is the unit of analysis. There variables in the dataset include the home sales price in dollars, the number of rooms in each home, the distance in miles from each home to the city, the quality rating of the school that serves each home, and the type of building the home is in.

I anticipate the home sales prices will range from $10,000 to $1,000,000, with an average sales price of about $100,000. The homes in this data set are likely to have between two and ten rooms, with an average of five rooms. They are all less than 20 miles from the city, with an average distance of 2 miles from the city. School are generally rated as having between zero and five stars, and the average rating is likely to be about three stars. Homes may be in one of three building types: single-family homes (probably about 50 percent of all homes), duplexes (probably about 30 percent of all homes), and buildings with three or more units (about 20 percent of all homes).

Research Question

How should differences in the distance to the city effect our predictions for home sale prices, after accounting for effects that a home's number of rooms, school quality ratings, and building type might also have on home prices?

Hypotheses

Homes that are closer to the city might command higher prices because they offer a shorter, more convenient commute to destinations in the city. Homes with more rooms might sell for a higher price because they offer more living space. Higher quality schools might also be associated with higher home prices because people will pay a premium to be zoned for a school that will offer better opportunities for their children. Finally, people might pay more for single-family homes because they value the privacy that comes from not sharing walls with neighbors.

Data description

My analysis of the variables in the dataset shows that the data in my sample is mostly consistent with what I had expected.

Home sales prices

The distribution of home sales prices in my dataset is shown in Figure 1. Notice that the horizonal dimension in Figure 1 is on a log scale, so that each incremental increase in the horizontal direction represents an increase by a constant factor rather an in increase by a constant number of dollars (e.g. $100,000 is the same distance from $10,000 as it is from $1,000,000). The distribution looks quite symmetrical on a log-transformed scale like this, and this suggests that it has a log-normal distribution.

The average home sale price was $110, 382 (slightly higher than the anticipated average of $100,000). The standard deviation is $71,810. This value is quite high, relative to the average value, which indicates that values that are quite far from the average are fairly common, which is typical of a variable with a log-normal distribution. Half of all homes in the sample were sold for at least $93,084 (this is the median sales price). About a quarter of all home sale prices were less than

$62,351, and about a quarter were more than $137,676. This represents an interquartile range of $75,325, which is slightly higher than the standard deviation.

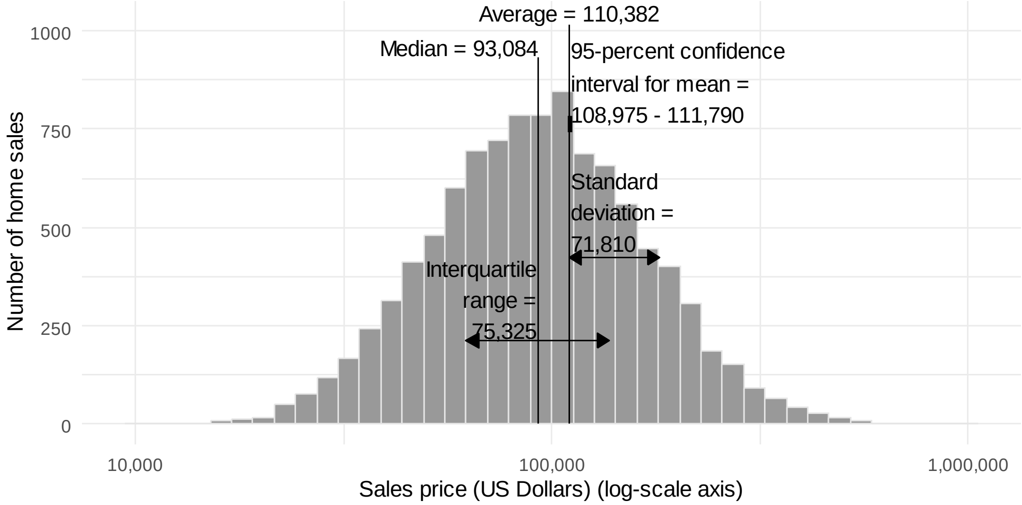

Figure 1: Distribution of home sales prices.

Based on the number of home sales in our sample and the standard deviation (how spread out the values in our sample are), we can be 95-percent that the average value for all home sale prices (in the full set of home sales this sample is drawn from) is between $108,975 and $111,790.

Number of rooms

Figure 2 illustrates the distribution of the number of rooms for the homes in my sample. As expected, all homes had between two and ten rooms, although the average of six rooms per home was higher than the anticipated value of five. The median value (the number that half of all observations fell below) is approximately equal to the average value. About a quarter of all homes had five rooms or fewer and about a quarter had seven rooms or more. The standard deviation was 1.09, which is about half of the interquartile range.

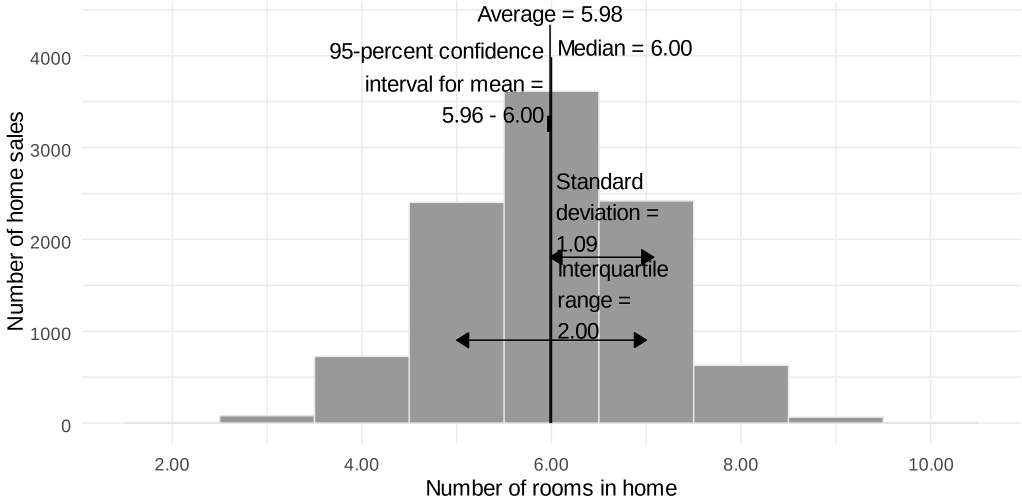

Figure 2: Distribution of numbers of rooms

Based on the number of home sales in our sample and the standard deviation (how spread out the values in our sample are), we can be 95-percent that the average number of rooms for all homes sold (in the full set of home sales this sample is drawn from) is between 5.96 and 6 rooms.

Distance from city

Figure 3 shows the distribution of the distance from the city for each home sold. As expected, all homes sold were less than 20 miles from the city. The closest home was 0.6 miles from the city and the average distance from the city was 4.68 miles (more than double the anticipated average of two miles). Half of all homes sold were more than 4.26 miles from the city, about a quarter were closer than 3.1 miles, and about a quarter were further than 5.8 miles. The standard deviation for distances from the city was 2.17, which is about half of the interquartile range.

Based on the number of home sales in our sample and the standard deviation (how spread out the values in our sample are), we can be 95-percent that the average distance to the city for all homes sold (in the full set of home sales this sample is drawn from) is between 4.63 and 4.72 miles.

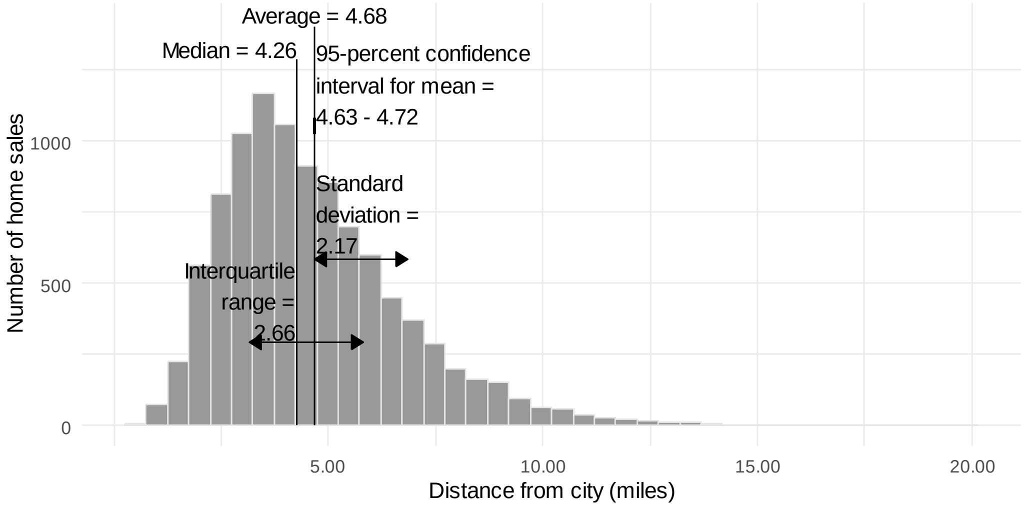

Figure 3: Distribution of distances from city

School quality rating

As shown in Figure 4, school quality ratings ranged from zero to five stars, with an average value of 2.36 stars and a standard deviation of 0.71 stars. We can be 95-percent confident that the average school rating for the full population of home sales is between 2.35 and 2.38 stars. Figure 4 also shows the median and interquartile ranges for school quality ratings, but since there are only six possible values for this variable, the summary in Table 1 might be more informative.

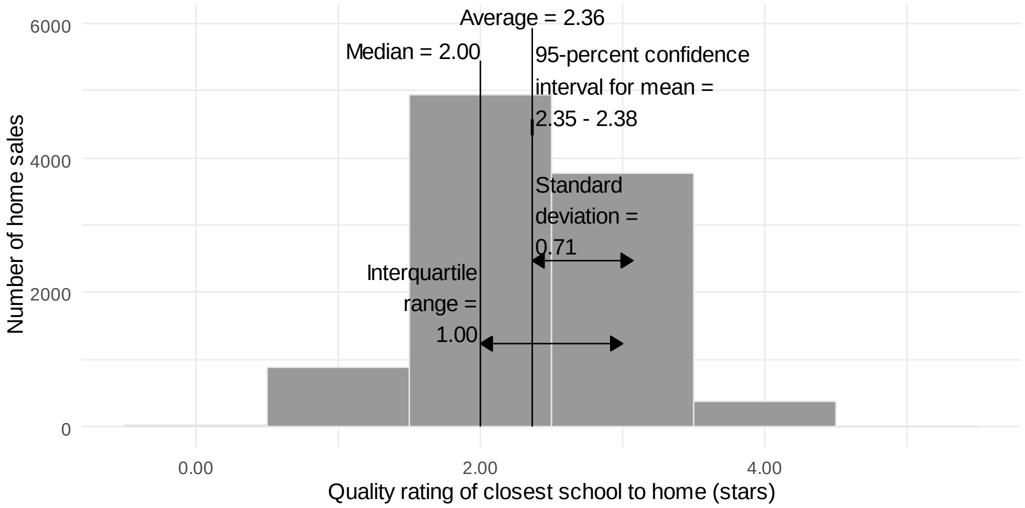

Figure 4: Distribution of school quality ratings for homes sold.

Table 1: Distribution of school quality ratings for homes sold.

School Quality Rating	Number of home sales	Percent of home sales
Zero stars	21	0.2%
One star	877	9.8%
Two stars	4,944	49.4%
Three stars	3,761	37.6%
Four stars	393	3.9%
Five stars	4	0.04%
Total	**10,000**	**100%**

Almost half of all homes sold were closest to a school with a quality rating of two stars. The lowest-rated quarter of school ratings were between zero and two starts, and the lower-rated half of school ratings had this same range. The highest-rated quarter of school ratings ranged from three to five stars.

Building type

Figure 5 shows the distribution of homes across the six defined building types. As expected, about half of all homes sold were single-family homes. The remaining homes were more or less evenly divided between duplexes and buildings with three or more units.

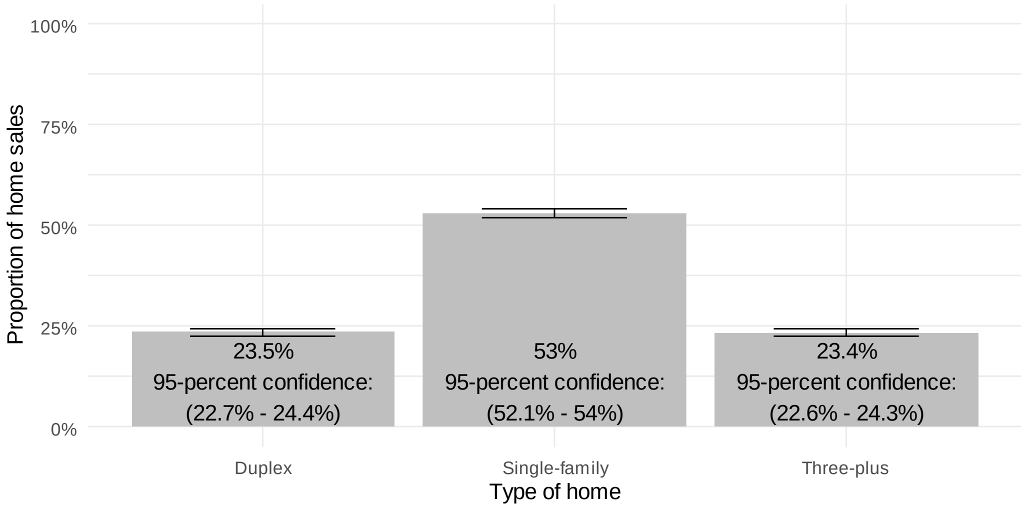

Figure 5: Distribution of types of homes sold

Initial Analysis

I have hypothesized that:

- Longer distances to the city are associated with lower home sale prices,
- Lower school quality ratings are associated with lower home sale prices,
- Homes with more rooms sell for higher prices, and
- Single-family homes sell for higher prices than homes in either duplexes or buildings with three or more units.

I will test each of these hypotheses below.

Home prices and distance to the city

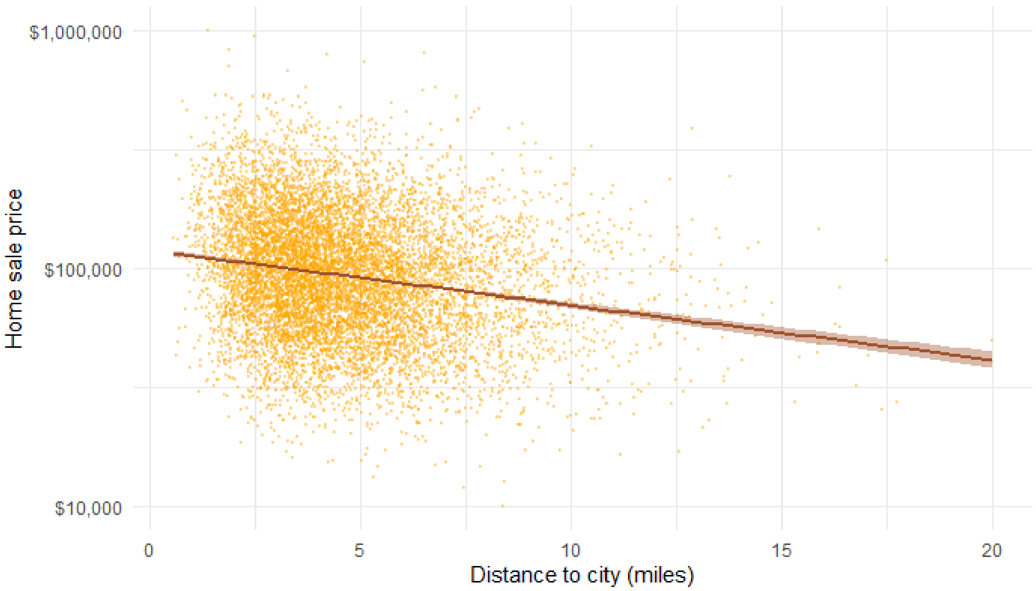

Figure 6: Relationship between home sale prices and distance to the city.

Figure 6 shows illustrates the relationship between home prices and distance to the city. Note that home prices are shown on a log scale (evenly spaced increments represent an increase by a factor of ten). The linear trendline that fits these points is also shown. It is downward sloping, which supports the hypothesis that longer distances to the city are associated with lower home sale prices. The correlation between these two variables is -0.17. This value is closer to zero than one, suggesting a moderate relationship, and it is negative, which again, indicates that higher distances to the city are associated with lower home sale prices.

Home prices and school quality

Figure 7 shows illustrates the relationship between home prices and the school quality rating of the closest school to each home. Again, home prices are shown on a log scale and the linear trendline that fits these points is also shown. It is upward sloping, which supports the hypothesis that higher school quality ratings are associated with higher home sale prices. The correlation between these two variables is 0.23. This value is farther from zero than the correlation with distance to the city, suggesting slightly stronger relationship. It is a positive value, which confirms what we can see visually from the trend line: that higher-rated schools are associated with higher home sale prices.

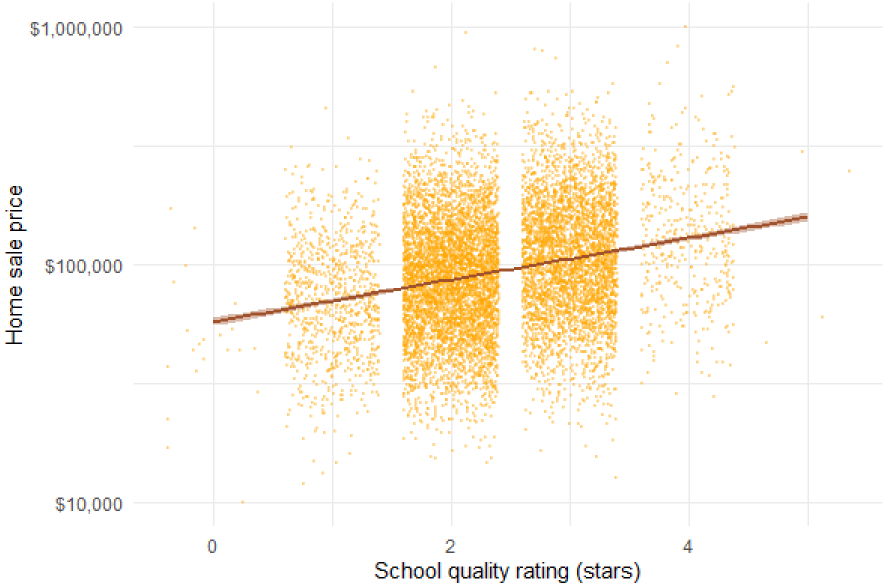

Figure 7: Relationship between home price and school quality rating

Home prices and number of rooms

Figure 8 shows illustrates the relationship between home prices and the number of rooms in each home. Again, home prices are shown on a log scale and the linear trendline that fits these points is also shown. It is upward sloping, which supports the hypothesis that more rooms are associated with higher home sale prices. The correlation between these two variables is 0.26, which is slightly higher (in its magnitude) than the correlation with school ratings or distance from the city, suggesting a stronger relationship. It is a positive value, which confirms what we can see visually from the trend line: that homes with more rooms are associated with higher home sale prices.

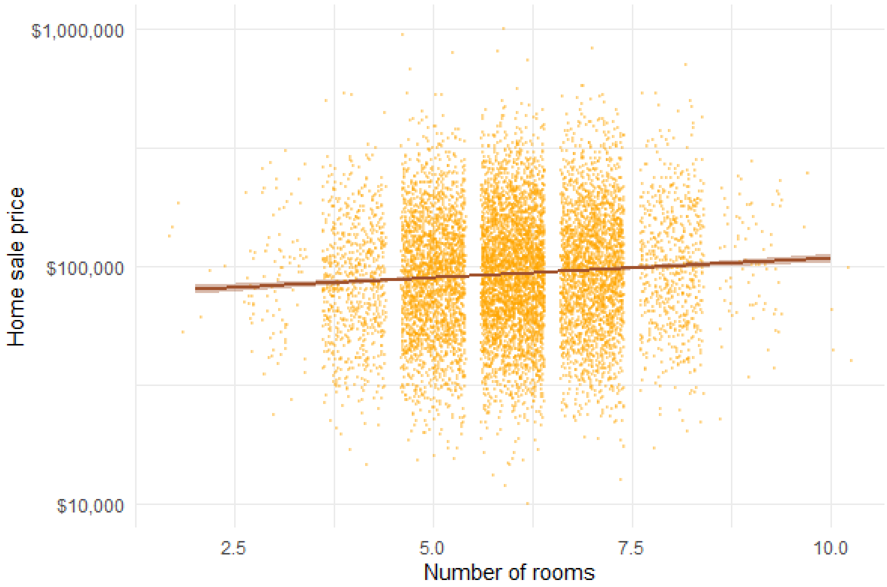

Figure 8: Relationship between home price and number of rooms

Home prices and building type

Figure 9 shows the variation in home sale price by type of building. The width of each shape represents the proportion of values within each category at the indicated value and the horizontal lines represent the median value. As shown, the median sales price for single family homes is slightly higher than for homes in duplexes or in in buildings with three or more units.

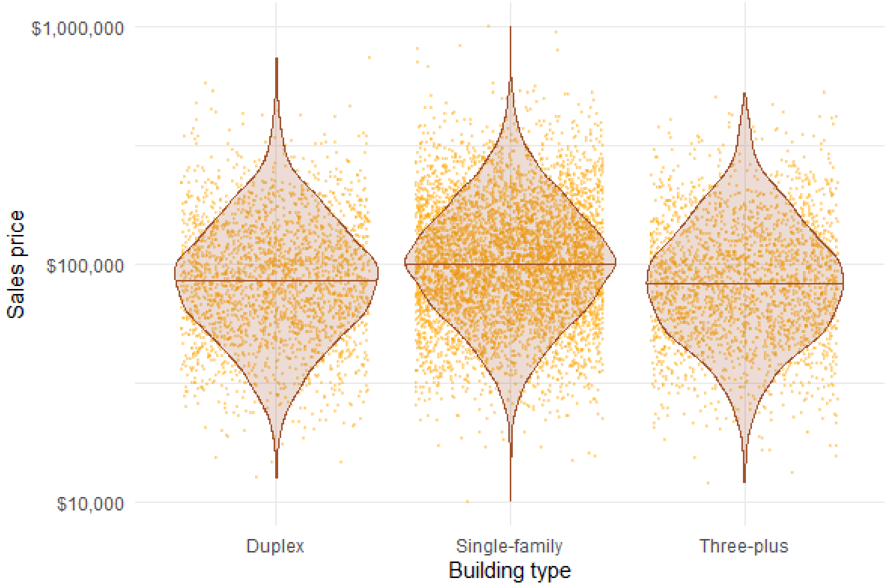

Figure 9: Home sale price by type of building

The average home sale price for single-family homes is $119,625, which is $18,996 higher than the average home sale price for homes in duplexes and $20,378 higher than the average home sale price for homes in buildings with three or more units.

Independent relationships with home sale prices

I have estimated a regression model to determine how each of the above variables could influence predicted home sale prices. The R-squared value for the model is 0.105, which indicates that this model explains about 10.5 percent of the variation in home sale prices.

The model coefficients are shown in Table 2. The estimated coefficient shown in the table is the estimated effect of the variable value on a predicted sales price. The p-value for each variable represents the probability the regression would have produced this same result if the true value for the full population the sample was drawn from were zero (meaning there no relationship between that variable and home sales prices). Since all of these values are less than 0.001, we can be more than 99.9 percent confident that the actual relationship between each variable and home sale prices is in the same direction as the coefficients in the table.

Table 2: Regression results

Variable	Estimated coefficient	p-value
Intercept	$65,288.20	< 0.001
Number of rooms	$4,350.00	< 0.001
School quality rating (stars)	$23,351.70	< 0.001
Distance from city (miles)	-$5,667.00	< 0.001
Number of units in building (compared to single-family home)		
Two	-$18,824.20	< 0.001
Three or more	-$20,991.60	< 0.001

The intercept from the regression is $65,288.20, which means that, for a single-family home located within the city (zero miles away), with zero rooms, where the closest school had a quality rating of zero stars, the model would predict a home sale price of $65,288.20. In reality, none of the homes in our sample have zero rooms, so this case is outside the range that our model can be applied to, but it offers a useful starting point for making predictions for scenarios within the expected range of our data.

The coefficient for the number of rooms is $4,350.00, which means that the predicted sales price would increase by $4,350.00 for each additional room in a home.

The coefficient for the school quality rating is $23,351.70, which means that the predicted sales price would increase by $23,351.70 for each one-star increase in the school quality rating.

The coefficient for distance from the city is -$5,667.00, which means that the predicted sales price would decrease by $5,667.00 for each one-mile increase in the distance from the city.

Finally, the coefficients for duplexes and triplexes are -$18,824.20 and -$20,991.60, respectively. These indicate that the predicted sales prices for duplexes are $18,824 lower from duplexes than for single-family homes and the predicted sales prices for triplexes are $20,991.60 lower than for single-family homes.