|  |
| --- |
| **Seattle Real Estate & Development, LLC** |
| ANALYSIS OF KING COUNTY HOME SALES, 2014/15 |
| For: Board of Strategic Development at Seattle Real Estate & Development, LLC |

|  |
| --- |
| Bardan Sigdel, Real Estate Consultant  6-6-2015 |

**Executive Summary**

*Based on our multivariate regression modeling and the subsequent analysis, we feel sufficiently confident in advising the board of strategic development that housing prices in King County appear to be most positively impacted by square footage of living spaces, high quality of housing grade in design and construction, and a waterfront setting. Bathroom count, property view ratings, home condition ratings and basement area similarly have a positive influence on prices but to a relatively smaller degree than the top three metrics. Conversely, overall land area and increases in the number of bedrooms show a negative impact on pricing, as does the aging of a home.*

Upon the board’s recommendation, we set out with an objective of identifying home/property characteristics and features that would serve as potential predictors of sale prices using a robust multiple regression model. Through the model discussed below, we have revealed significant predictors that both positively and negatively impact home sales prices in King County. Together, square footage of living space, housing grade and waterfront setting serve as the best positive predictors of home prices in King County. Additionally, bathrooms as well as condition ranks, viewing ranks and basement area also appear to have a positive relationship with pricing, although to smaller degrees than the aforementioned three variables. On the other hand, our analysis revealed bedrooms and the square footage of overall land area play an important effect but with reversed impact. All other factors held constant, these factors surprisingly led to a decrease in housing prices. Less surprisingly, the age of the home based on the year built also appears to have negative impact on pricing.   
  
The purpose of the model and analysis here was to determine strong predictors for housing price sales as well as the magnitude of their effects in King County (WA) based on the interaction of independent variable(s) data for homes with sale prices for the same. The findings from our analysis should serve decision-makers in planning for future development and construction upgrades in the area with the objective of maximizing profitability on investment. Our strategic board would be best-served moving forward by emphasizing on housing structures in the area that stand aligned with the magnitude impact that these predictors show in their ability to dictate pricing.  
  
The multiple regression model (model details in **Appendices A and B**), built using historical data for home sales in the county for the period between May 2014 and May 2015, consists of ten input variables upon which the model is built: bedrooms, bathrooms, living area square footage, total area square footage, waterfront setting (binary), viewing index, grade quality index, condition grade, basement square footage and house age. In the development of our model and determination of pertinent data, we preliminarily cleaned data from the original dataset to account for focus on our business question as well as made the analytical decision to not include predictors that veered into the territory of location-based metrics (**see Appendix C,** Removal of Variables from Original Dataset). We ultimately landed with ten independent variables as mentioned earlier (see **Appendix D** for details), out of which the records for “Year Built” variable underwent two transformative actions: firstly, the variable was converted into an aging metric to suit the modeling better and secondly, we narrowed our dataset to only include homes built from 1980 onwards.   
  
The coefficient analysis described in **Appendix A** represents the basic predictability structure gathered by our model. To contextualize this through an example for the board: we would predict a home built in 1990 (age of 25) with 3 bedrooms, 3 bathrooms, 2000 square feet of living space, 3500 square feet of land space, a waterfront setting, a view of 3 in the view index, 4 in the condition index, grade index of 10, 1500 square feet of basement space to have an expected price of $2,090,850.53, not accounting for variance in the model. Our impact analysis (**see Appendix B, Impact Analysis**) went hand-in-hand with the regression model for the most part, further bolstering the high impact positive impacts of the living area, grade and waterfront variables as the strongest positive indicators and bedrooms, total land area and aging as the most impactful negative determinants.  
  
With a model R-squared value at 0.678, we have indication that almost 68% of variation in home sale prices can be explained by the variables included in the model and the adjusted R-square of 0.677 similarly points at the good fit of the variables in the model. We should note here that since the normal probability plot using this model presented a skew, we did explore a model using the natural log of home prices (details in **Appendix C**). While this transformed model did address some issues of skew, we ultimately did not deem it a strong enough improvement on the model for our business goal. As such, we decided to retain the original response variable of raw home prices (differences highlighted in **Appendix C**).   
  
All things considered, despite the model’s structural soundness and our presentation at 95% confidence of all predictors’ influence, we would still urge caution to the board in any kind of sweeping applicability – this is primarily important because we do observe some variability in (**see Appendix B, Coefficient Analysis**) data as well as the model. Besides the variability, we also do need to note regional differences in home pricing and therefore, this model’s lack of scope of geographical variables (besides waterfront) could limit some applicability in areas of the county where geographical elements highly influence desirability.   
  
The source for our data used in constructing this model came (secondarily) from the research vault in the online academic repository Kaggle.com, where the source directly derives the dataset from King County, Washington, USA (linked details in **Appendix E**).

**Appendix A: Model and Interpretation**

The multiple linear regression model built using the King County home sales dataset will be discussed in Appendices A and B. The model’s strengths and weaknesses as well as associated metrics for those will be expanded upon in Appendix B. However, in brief, we were able to establish the model’s good fit and avoidance of overfitting using the adjusted R-squared value. P-values for all variables included in the model are all below the standard significance level of 0.05, indicating that they are all significant predictors of home prices.  
  
  
  
  
  
The final model developed using the regression model is represented by this equation:   
  
**price = -929,298.54 - (55,544.79 \* bedrooms) + (76,204.97 \* bathrooms) + (155.416 \* sqft\_living) - (0.1295 \* sqft\_lot) + (987,491.31 \* waterfront) + (50,985.78 \* view) + (48,473.94 \* condition) + (118,893.70 \* grade) + (47.242 \* sqft\_basement) - (919.184 \* age)**  
  
The model intercept coefficient needs to be treated with caution; in theory, this represents the expected home pricing when all independent variables are zero. However, for housing sales, this would not be meaningful in a realistic sense as we would have no property without any of these variables’ measures. Besides that, following the model, we can describe the effects of the input variables’ ranges on the response variable as listed below, holding all other variables constant:   
  
- **bedrooms:** the predicted home pricing decreases by $55,544.79 for each additional bedroom  
- **bathrooms:** the predicted home pricing increases by $76,204.97 for each additional bathroom  
- **sqft\_living:** the predicted home pricing increases $155.42 for every sq ft increase in living area  
- **sqft\_lot:** the predicted home pricing decreases $0.1295 for every sq ft increase in total land area  
- **waterfront:** the predicted home pricing increases $987,491.32 for waterfront home settings  
- **view:** the predicted home pricing increases $50,985.78 per unit increase in the viewing index  
- **condition:** the predicted home pricing increases $48,473.94 per unit increase in condition index  
- **grade:** the predicted home pricing increases $118,893.70 per unit increase in the grade index  
- **sqft\_basement:** the predicted home pricing increases $47.24 per sq ft increase in basement area  
- **age:** the predicted home pricing decreases $919.18 per year increase in house aging

Particularly of note here is that grade, waterfront and square footage of living area have emerged as strong predictors of home price. The waterfront variable’s coefficient trounces all others in terms of its enormity, but we should also note that this variable only indicates whether or not a property has a waterfront view. Since we expect that most homes with waterfront views are in the upper echelon of the home prices’ skew, this result in the model is to be expected. However, since the impact of this will not be pertinent if we hypothetically take out all homes without a waterfront view, any analysis that distinguishes the two categories should be mindful of this.   
  
With every one-unit increase in the grade index, we expect the home price to increase by $118,893.70. As we will see in impact analysis below, this variable is also the second most impactful as a predictor followed by waterfront, while sqft\_living is the most impactful. As such, we begin to recognize that the factors represented in these three variables do have the largest positive impact on home prices, although the range of the impact between them varies.

Reversely, bedroom count, square footage of overall area and home aging with significant negative coefficients indicate their negative effect on home prices.   
  
**Impact Analysis**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Coefficient** | **Min** | **Max** | **Range** | **Impact** |
| bedrooms | -55544.79389 | 0 | 10 | 10 | -555447.9389 |
| bathrooms | 76204.97484 | 0 | 8 | 8 | 609639.7988 |
| sqft\_living | 155.4162882 | 384 | 13540 | 13156 | 2044656.688 |
| sqft\_lot | -0.129541951 | 572 | 1024068 | 1023496 | -132585.669 |
| waterfront | 987491.3133 | 0 | 1 | 1 | 987491.3133 |
| view | 50985.78265 | 0 | 4 | 4 | 203943.1306 |
| condition | 48473.93698 | 0 | 5 | 5 | 242369.6849 |
| grade | 118893.7034 | 4 | 13 | 9 | 1070043.331 |
| sqft\_basement | 47.24229361 | 0 | 4820 | 4820 | 227707.8552 |
| age | -919.1843695 | 0 | 35 | 35 | -32171.45293 |

The coefficient values stacked against the range of data for each variable provides a fairly clear view of the distribution of impact across the independent variables on home prices. The variable sqft\_living, representing the total living space square footage in the dataset, evidently stands out as the variable with the highest impact on the response variable followed by grade and waterfront. While living area size and grade should be standout metrics in this regard across housing prices anywhere, the impact of the binary variable of waterfront properties makes sense in the context and geographical positioning of the county in coastal Seattle. Beyond these three, bathrooms, basement square footage as well as viewing and condition indexes also have positive impacts on the response variable, albeit to much smaller degrees. The three variables that range across negative impact are bedrooms, total area square footage and the aging of the homes.

**Correlation Analysis**

A picture containing text, screenshot, display, number

Description automatically generated

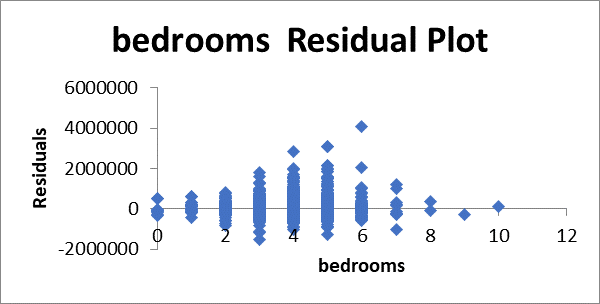
As we observe in the matrix, several independent variables in the dataset do have moderately high correlation with each other. Since we are working with variables that reflect the desirability of homes and their structures, we would expect that this would be the case. In such scenarios, multicollinearity is very much a possibility – however, we fortunately avoid correlations between the x variables here that exceed 0.80, we are comfortable with including all of them in the model. We discuss correlations with the response variable in our data analysis in **App D**.  
  
**Appendix B: Model Statistical Analysis**We can note a strong positive correlation between the dependent and independent variables based on the multiple R of 0.823. The R-square at 0.68 for the model without any variable transformation suggests that approximately 68% of the variance in our response variable can be explained by the input variables in our dataset. Since the data we are working on analyzing does not necessarily always have scientific backing and can be fairly volatile depending on several factors, we are satisfied with the coefficient of determination at 0.68. Similarly, the adjusted R-square at 0.677 makes for a positive outlook in that we are not greatly penalized for variables that could have a negative bearing on the model’s predicting ability.

Furthermore, we are ecstatic with the p-value for all independent variables observed here, with all less than 0.05, allowing us to swiftly reject the null hypothesis in this scenario and move forward with confidence that the independent variables do indeed have significant impact on the response variable.  
  


However, we do need to be careful about the variability in our model indicated by the standard errors for the model as well as for the intercept, which expresses the degree of uncertainty in the estimated coefficient. This relatively sizable value shows that there is a significant degree of variability in the intercept, which will invariably affect the preciseness of the model.

**Normal Probability Plot**  
  
  
  
The skew depicted in the normal probability plot is quite clear with the observed heavy tails – to address the impact of this in our model, we attempted to transform the response variable using a natural logarithm (mentioned before and discussed in **App C**). However, the model using such a transformed response variable (price) did not yield a much better model than the original. As such, and in the interest of avoiding unnecessary complexity and keeping the communicative ability of the model intact, we reverted to the original price variable in its absolute form.

**Residual Plots**The residuals represented for each variable in the residual plots below depict the difference between the actual value of home prices and the predicted values. We would ideally want residuals to be scattered around the x-axis to indicate good model fit. Any patterns noticed in the residual plots might indicate data-model fit issues, which we could say is true for bedrooms in our model as shown by its model residual plots. However, aside from it, we are satisfied with the residual density and spread of higher significance predictors for all non-indexed variables.



**Appendix C: Modeling Process and Development***Cleaning and Improvement:*   
  
After cleaning the data and removing variables to account for redundancy as well as to consolidate the dataset in the interest of improving our model’s precision, we are left to work with home prices as the response variable along with 10 independent variables.

The first column of data in the original dataset was the listing of property ID’s, which would have no direct or indirect bearing on our modeling or the examination on relationship with the response variable – as such, it was removed. The largest correctional factor for audience of the analysis in the dataset is perhaps the cleaning of observations based on property/house aging. Once the decision to clean houses built before 1980 was made, the original count of 21613 house sales was then brought down to a final working count of 9279, which we consider our sample size for our modeling. **As such, it is critical to emphasize to the board that this is a model for house sales in a 12-month window only accounting for houses built from 1980 onwards.**  
For this reason and also because the ”age” of a property would be a far easier (and improved) indicator to work with for the purposes of this model, the column for “Year Built” was transformed to an age count.

*Removal of Variables from Original Dataset:*

We removed a few variables from the original dataset from logistical and “focus of analysis” perspectives rather than from a statistical one. The floor column was removed as it might not be a great predictor in this model, perhaps owing to the lower-heighted nature of single homes in the region/country. The zip code column from the original dataset was also removed, while also acknowledging that this could very well be used in a different study to examine the relationship between sales/pricing and zip code clusters. Similarly, columns for longitudinal and latitudinal measure for each house were removed although these variables too could serve as fair indicators in modeling of analysis aimed at geographical sales trends. However, for the purpose of this modeling, we chose to remove these. Other variables in the original dataset that were removed are “sqft\_above” (total square footage of the property above the basement level) and two columns of data pertaining to the living spaces for 15 closest neighbors to the property. While there is room to introduce these data for analysis, we already account for basement and overall living spaces – therefore, the decision to drop the “sqft\_above” variable was in the interest of reducing potential redundancy.

After deciding on usable data and adjusting variables as needed, the variables that we used in the regression model are as follows:

1. **price**: price of Home Sold (response variable y)
2. **bedrooms:** number of bedrooms in house
3. **bathrooms:** number of bathrooms in house (half bathrooms account for bathrooms without a shower)
4. **sqft\_living:** square footage of total interior living space
5. **sqft\_lot:** square footage of land area
6. **view:** rating of the house’s view on an index between 0 and 4
7. **waterfront:** whether or not the house offers a waterfront view, dummy variable
8. **condition:** index between 1 and 5 for overall condition of the house
9. **grade:** index between 1 and 13 to indicate the quality of house construction and design
10. **sqft\_basement:** square footage of area under below ground level, on a basement level
11. **age:** transformed variable based on the year built data, used to convert the categorical date variable to a usable numeric variable (\*reference year is 2015, not 2023)

*Response Variable Transformation and Determination of Final Y –* absolute price versus natural log of priceSince we observed a significant heavy tailed right skewed normal probability plot for our regression model (discussed in **App B**), we attempted to introduce variations or transformations to address it. Taking the natural log of the response variable (where variability is large) was the most straightforward approach to take here – however, transforming this variable did not end up improving the quality of our model to a significant degree, particularly in consideration of how much relative complexity it would introduce to our modeling and report. As shown in the summary output below for the transformed model using the natural log of price, we see an R square of 0.697, which is only a minor improvement on our original model, which we therefore stick with.   
  
Experimental regression model’s using response variable’s natural log for comparison:

|  |  |
| --- | --- |
| SUMMARY OUTPUT | |
|  |  |
| *Regression Statistics* | |
| Multiple R | 0.835103 |
| R Square | 0.697397 |
| Adjusted R Square | 0.697071 |
| Standard Error | 0.277631 |
| Observations | 9279 |

**Appendix D: Data Analysis**

1. **Price – Response Variable**



The histogram for the pricing column only further establishes what we can see in the descriptive statistics for the same: there is a right-skew distribution for pricing in our data, which is evident also by the median of $490000 in the data with a mean of $587000. This should be considered a typical pattern in real estate or property sales datasets since it is only sensical that there is greater frequency of lower-priced property sales as compared to those that go into the multi-million categories, hence being conducive to the kind of thin-tailed distribution we observe. As such, the overall distribution is greatly impacted – we would assume that it is only impacted more as the sample size grows.

1. **Condition**Against Prices:  
     
     
     
   The distribution of the house “condition” variable is a fairly average representation of these properties – properties in the “3” bucket are significantly larger than any other. As such, both the mean and median for the condition index are 3. The trendline in the scatterplot comparing the prices against the condition of these homes shows a slightly positive correlation between home condition and prices.
2. **Bedroom and Bathroom**

Descriptive Statistics for bedrooms and bathrooms:  
  
Histograms for bedrooms and bathrooms:

**A picture containing text, screenshot, font, number

Description automatically generated**

Bedrooms: Both the summary statistics and the histogram for the bedrooms suggest a fairly normal distribution with the highest frequency at a count of 4. There are a fair few mild outliers, as expected, on the right of the distribution. Given the size of the dataset, these should not have a big impact on the model. That said, since bedrooms are expected to be one of the key predictors in the model (the scatterplot gives a fair indication of this), it might be wise to keep the outliers in consideration.  
  
Price Correlation 0.32699652

The correlation coefficient here suggests a moderately positive linear relationship between the two variables (profit vs bedrooms), which might typically be true of most housing sales. The strength of the correlation here, however, is not very high, which needs to be a point of consideration as other variables are also examined.   
  
Bathrooms: For bathrooms, too, we observe a normal distribution with a few mild outliers.   
Price Correlation 0.573928808  
  
There is moderately positive liner relationship (more than was the case with bedrooms) between bathrooms and the property pricing – this is to be expected for the most part in any part of the world; as bathrooms increase, on average, sale prices follow suit.

1. **Living Area (sqft\_living)**

**A picture containing text, screenshot, font, number

Description automatically generated**

A clear positive right skew in the distribution of square footage of living area was to be expected based on the descriptive statistics for this variable – the histogram here illustrates the same.   
  
Price Correlation 0.731661671  
  
Similar to beds or baths, the square footage of living spaces are often one of the key indicators in sales or pricing models for properties. In this data, too, this variable has a clear positive linear relationship with the response variable. The price correlation of 0.73 further bolsters this stance.

1. **Land Area (sqft\_lot)**Of all variables in the model, the lot/land sqft area could possibly be the strongest candidate for a transformation – the range is vast and the variance for a mean here of 17710 might be considered for transformation in many regression models.   
     
   Price Correlation 0.132524986  
     
   Besides the potential for transformation, the correlation coefficient of 0.132 is far more different than the same for the living square footage, which might be slightly surprising from the perspective of a straightforward prediction due to its weak positive linearity.
2. **Age**

While there might have been reason for concern when cleaning the original dataset and adjusting for house aging (in regards to having fairly distributed data across the variable), the summary statistics for aging indicate that it should not be an issue. As the reference year was set to 2015 (part of the fiscal year for our sales data), there are no concerns for negative aging, which could possibly have introduced hiccups in the modeling.

Price Correlation -0.095564651

The correlation against price suggests a very weak negative liner relationship between these two variables – as the age of a property increases, the price tends to decrease slightly but the linearity is not clear. This might have been different if we had included all homes in the original dataset but since our data is considering homes built after 1980 only, the observation would make sense.

1. **Basement Area (sqft\_basement)**

Price Correlation 0.343517891  
  
The correlation coefficient with price here suggests a moderate positive linear relationship – as the size of the basement increases, the prices do tend to increase. However, from an analyst’s perspective, this significance of the coefficient (less than 0.5) might be surprising as basement size is often regarded as a prominent marker of a house’s appeal. The spread of the scatter plot against sales prices is similarly surprising – one might expect the spread to be similar to that for sqft living but it clearly is not.

1. **View**

Similar to the “condition” variable that was navigated earlier, the “view” and “grade” variables also provide predictors in the form of indexes as a measure of quality rather than absolute incremental values. In many ways, this can prove beneficial for consolidating satisfaction metrics within a few buckets rather than having to find analyst-defined measured based on descriptive values such as “good”, “great”, “bad”, etc.

While the descriptive statistics for these indexes-based variables are helpful in many ways, analysis should be done carefully as these are values within a scale and not absolute values. Therefore, the same statistic could have a largely variable interpretation for index-based metrics.

That said, the view index seems a bit puzzling at quick glance – the massive right skew in the distribution with a 0.19 mean and 0 median for an index between 0 and 4 indicates that collection of data for this index variable could potentially have had issues. However, since this is typically a prominent indicator of sales tendencies, the variable deserves the benefit of doubt in the model.   
  
Correlation Matrix for the Two Index Variables: View, Grade  
  


1. **Grade**

A clearly strong relationship between price and grade variables can be observed here, which is good news from a societal perspective – there is clear emphasis on grade quality from the buyers’ perspective within the housing sales in this dataset. For the model, too, the variable promises to be a real strong predictor.

1. **Waterfront**

Waterfront, owing to its nature of being a dummy variable, might have the least revealing descriptive statistics in the entire list of independent variables for this modeling. Since this is a binary variable (houses with a waterfront view = 1, houses without a waterfront view = 0), the relationship with pricing should be fairly clear to observe – the price correlation of 0.33 suggests a positively moderate linear correlation between the two variables.

**Appendix E: Data Source**

The data used for the regression model completed for this project was acquired from Kaggle.com, where the database derives the dataset from King County, Washington, USA. The dataset includes house sales for the period between May 2014 and May 2015 and was procured via a downloadable zip file with a .csv file (linked below through the Kaggle.com website). As discussed earlier, several variables were removed from the original dataset for the model and analysis and furthermore, the range of the aging for the homes in the dataset was reduced to only include homes built from 1980 onwards.   
  
Direct Link: <https://www.kaggle.com/datasets/harlfoxem/housesalesprediction>