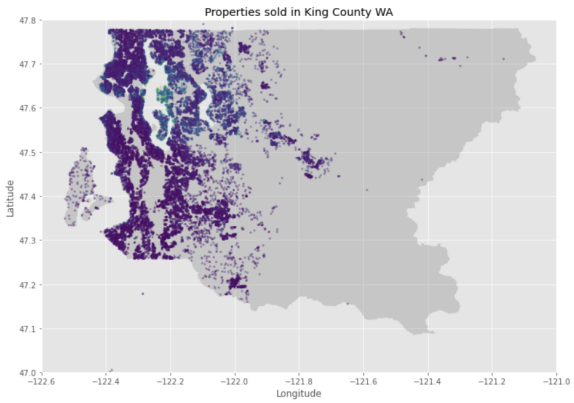
King County Housing Analysis Project



Overview

This project takes a given data set consisting of houses sold in King County, Washington, and uses it to create a linear regression model to predict house sale prices based on applicable variables. The project goes through data exploration, pre-processing, model iteration, and analysis of the final model results.

The Business + Project Goals

A non-profit addiction treatment center is looking to expand their services to offer half-way houses or sober living homes. These types of living spaces are hugely beneficial to those in recovery. It provides a safe space for transitioning back into the "real world" after going through in-patient rehabilitation. The center is interested in knowing how much different types of homes would cost in order to set budgetary and fundraising goals.

The project aims to give the treatment center the best possible prediction of what different types of homes would cost based on the square footage of the living area of the home, the evaluated grade and condition of the home, and a few other applicable variables.

Creating Linear Regression Models

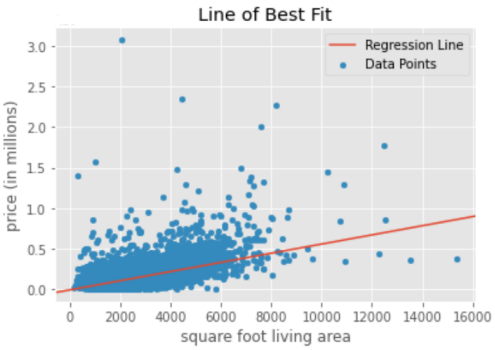
Data Exploration

The given data is already in a usable csv format. The project goes through basic exploratory analysis, creates the above map showing properties sold, and drops some data that does not fall within the parameters of the project (the homes sold are not located in King County, Washington). The table below shows the variables used in the project and their descriptions.

Variable	Description
`price`	Sale price of the home (target)
`sqft_living`	Square footage of living space in the home
`greenbelt`	Whether the house is adjacent to a green belt (an area of open land around a city, on which building is restricted)
`waterfront`	Whether the house is on a waterfront (including lake, river/slough waterfronts)
`nuisaance`	Whether the house has traffic noise or other recorded nuisances
`view`	Quality of view from house
`condition`	How good the overall condition of the house is. Related to maintenance of house
`grade`	Overall grade of the house. Related to the construction and design of the house

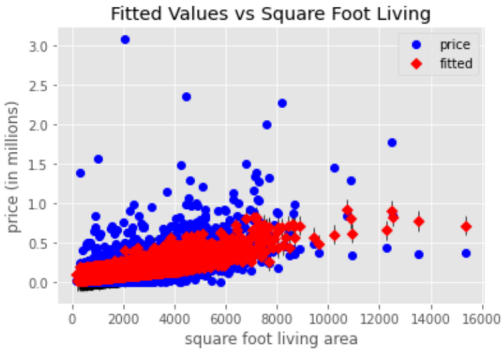
Simple Linear Regression

After looking at the correlation between the target variable, price, and the other numerical variables in the dataset, a simple linear regression model is created using the predictor of square foot living area. This simple linear regression model is used as a baseline to compare all other models against.



Linear Regression with Categorical Variables

In order to incorporate categorical variables, encoding is applied. The variables transformed in this manner are: `waterfront`, `greenbelt`, `nuisance`, `view`, `grade`, and `condition`. Once this is accomplished, a regression model with these variables and `sqft_living` is fitted. The model is an improvement from the simple linear regression model in terms of r-squared and mean absolute error statistics. It includes more statistically significant variables, however we do see about 6 coefficients that are not statistically significant.



> In comparison to the Line of Best Fit visual for the simple linear regression model, we can see the fitted values from our more complex model better align to our known data.

Logarithmic Transformation of Price

The target variable, price, has an underlying distribution with a dramatic right skew. In order to try and improve upon the model, the project applies a logarithmic transformation on price and fits a new model with this transformed target. This model is an improvement as all but one of the coefficients become statistically significant at an alpha of 0.05. Furthermore, the mean absolute error decreases from the previous model as well.

Drop Outliers in Price

Rather than a logarithmic transformation, this iteration attempts to create a more normal distribution of price by dropping the outliers in this variable. The resulting model shows marginal improvement from the original complex regression in terms of the r-squared and mean absolute error statistics. However, the mean absolute error is slightly greater (and therefore worse) than the model with the logarithmic transformation. Furthermore, this model has less statistically significant coefficients than the log transformed model.

Final Model

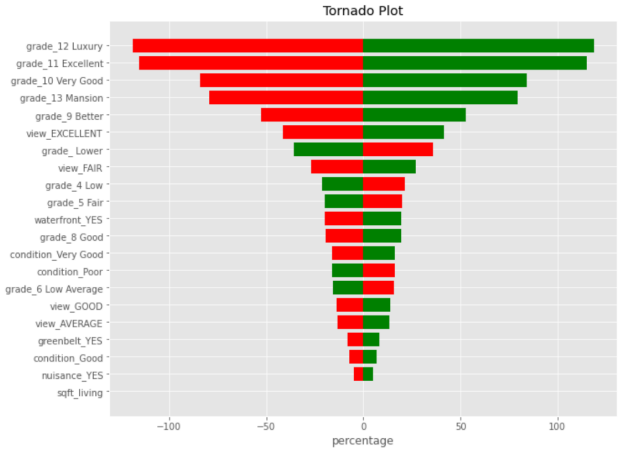
For the purposes of this project, it is more important to have a lower error rate and more usable information for the business to take into account (which means more statistically significant coefficients) than it is to simply have the highest r-squared statistic. With this in mind, and based on the model iterations above, one last model is created where the outliers in price are dropped, and then a logarithmic transformation is applied to this variable as well prior to modeling.

Final Regression Results

The final model explains about 47% of the variation in our target price (specifically ln(price)), is statistically significant overall, and includes variables that are statistically significant. Furthermore, the mean absolute error shows that the average error for the model predictions are about plus or minus $335,000.

Based on the variables used in regression, we will define an 'average home' as one with about 2130 square feet living area, is not on a waterfront, is not next to a greenbelt, has no recorded nuisances, has no view, and has a grade and condition rating of average. This average home is predicted to cost about $760,000.

The project describes the impact of each statistically significant coefficient in detail. For the purposes of this summary, please see the below tornado plot. The percent increase or decrease in predicted house price is shown for each statistically significant coefficient.



Conclusion

Suggestions

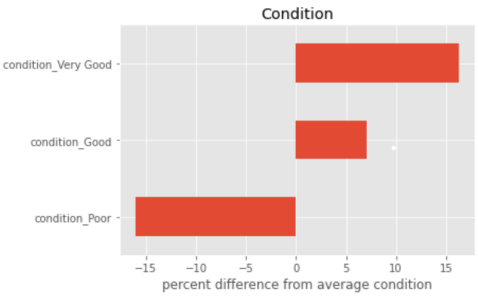
1. Minimum Budget

Budget a minimum of $760,000 per home plan to purchase. This is the predicted price of our defined average home.

2. Condition

For King County, Washington, the condition of a home ranges from poor to very good. Homes with an average condition have "some evidence of deferred maintenance" with "a few minor repairs needed." Based on this description and the center's needs it is recommended to look at homes with a minimum condition of average. If a home with a condition of poor, the home price would be about 16% cheaper, however these "worn out" homes will need maintenance -- the cost of which can negate any savings from purchasing such a home.

Furthermore, it would be wise to set aside about an additional $53,000 if looking at homes with a condition of Good. Homes with this condition will cost only 7% more and will need little to no maintenance prior to opening for service.

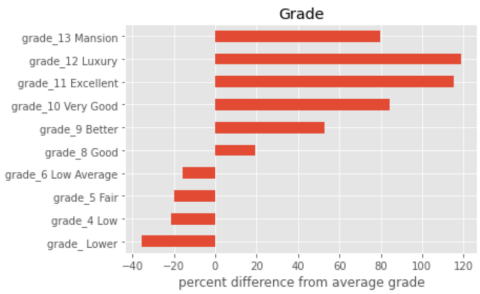


3. Grade

King County has a grade scale that goes from "1 Cabin" up to "13 Mansion." A home with an average (7) grade is described as one with "average grade of construction and design" and it is "commonly seen in plats and older sub-divisions." It is also important to note the King County has set building standards for homes -- they must be of grade 6 or above to meet these standards. Grades "8 Good" through "10 Very Good" show sequentially better materials used in construction and more thoughtful or appealing designs. Lastly, grades "11 Excellent" on up all consist of custom designs and top tier materials used in construction.

With these descriptions in mind, it is recommended to purchase a home with a minimum grade of "6 Low Average". Compared to our defined average home, the center could save about $120,000 if a home of this grade was purchased.

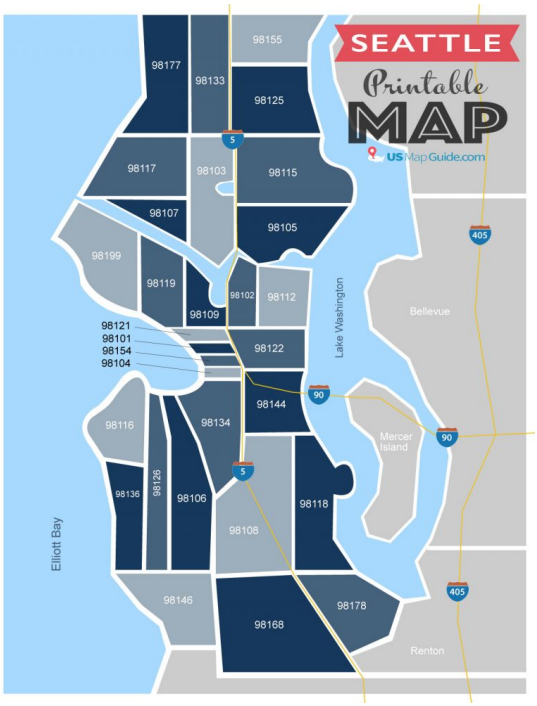
On the other hand, if homes with grades Good, Better, or Very Good are being considered for purchase it would be wise to set aside $148,000, $399,000, or $639,000 respectively. It is recommended to not purchase homes with grades Excellent or above due to the steep increase in sale prices.

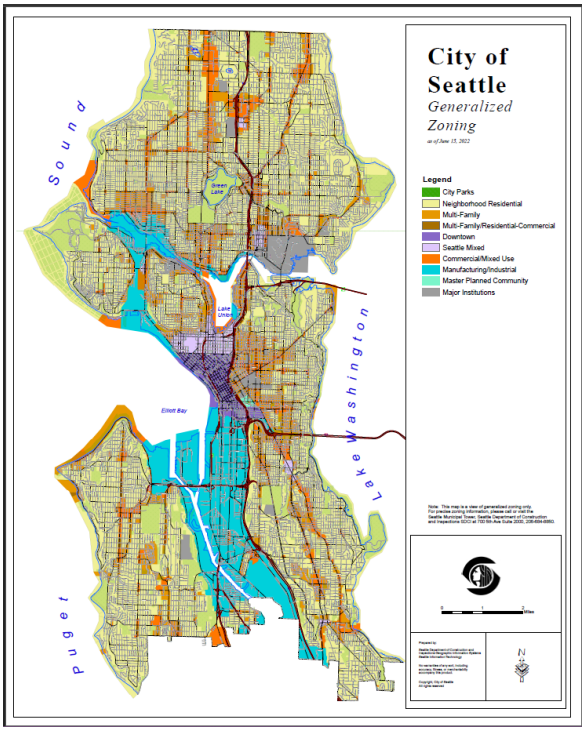


Next Steps

Many times, when companies or individuals try to create half-way homes or sober living homes in residential areas, there can be harmful push back from the communities. It would be wise to investigate how the zone affects the sale price of homes and plan accordingly -- either by adjusting amount of money put towards a home in or out of a residential zone, or by getting out in the community prior to opening a home in a residential area to help easy community fears and biases to those in recovery.

This next step would be best implemented after potential purchases have been narrowed down to specific cities or towns in King County, due to the complexity behind zoning ordinances. Zones are defined by the cities and towns themselves rather than by the county, and do not follow any easily accessible boundaries, such as zip code lines (as we can see from the images below).


SEATTLE
Printable
MAP
USMapGuide.com
98155
98177
98133
98125
98117
98103
98115
98107
98105
98199
98119
98102
98112
98109
98121
98101
98154
98104
98122
98144
98116
98134
98126
98136
98106
98118
98108
98146
98168
98178
Elliott Bay
Lake Washington
Bellevue
Mercer Island
Renton
5
405
90


City of Seattle
Generalized Zoning
as of June 15, 2022
Legend
City Parks
Neighborhood Residential
Multi-Family
Multi-Family/Residential-Commercial
Downtown
Seattle Mixed
Commercial/Mixed Use
Manufacturing/Industrial
Master Planned Community
Major Institutions
Sound
Lake Washington
Puget
Green Lake
Lake Union
Elliott Bay