aggie2411 / 110920-pt-phase-2-project
forked from learn-co-students/110920-pt-phase-2-project

0 stars 11 forks

Star Watch

Code Pull requests Actions Projects Wiki Security Insights Settings

master

This branch is 19 commits ahead of learn-co-students:master.

Pull request Compare

aggie2411 ... 6 minutes ago

View code

README.md

King County Home Price Analysis

This repository offers an analysis of factors that influence housing prices in King County, WA

Quick Links

1. Final Analysis Notebook
2. Presentation Slides

Setup Instructions

To run these notebooks independently, I have provided the environment.yml to clone to geo-env environment which contains all dependencies.

Scope of Project

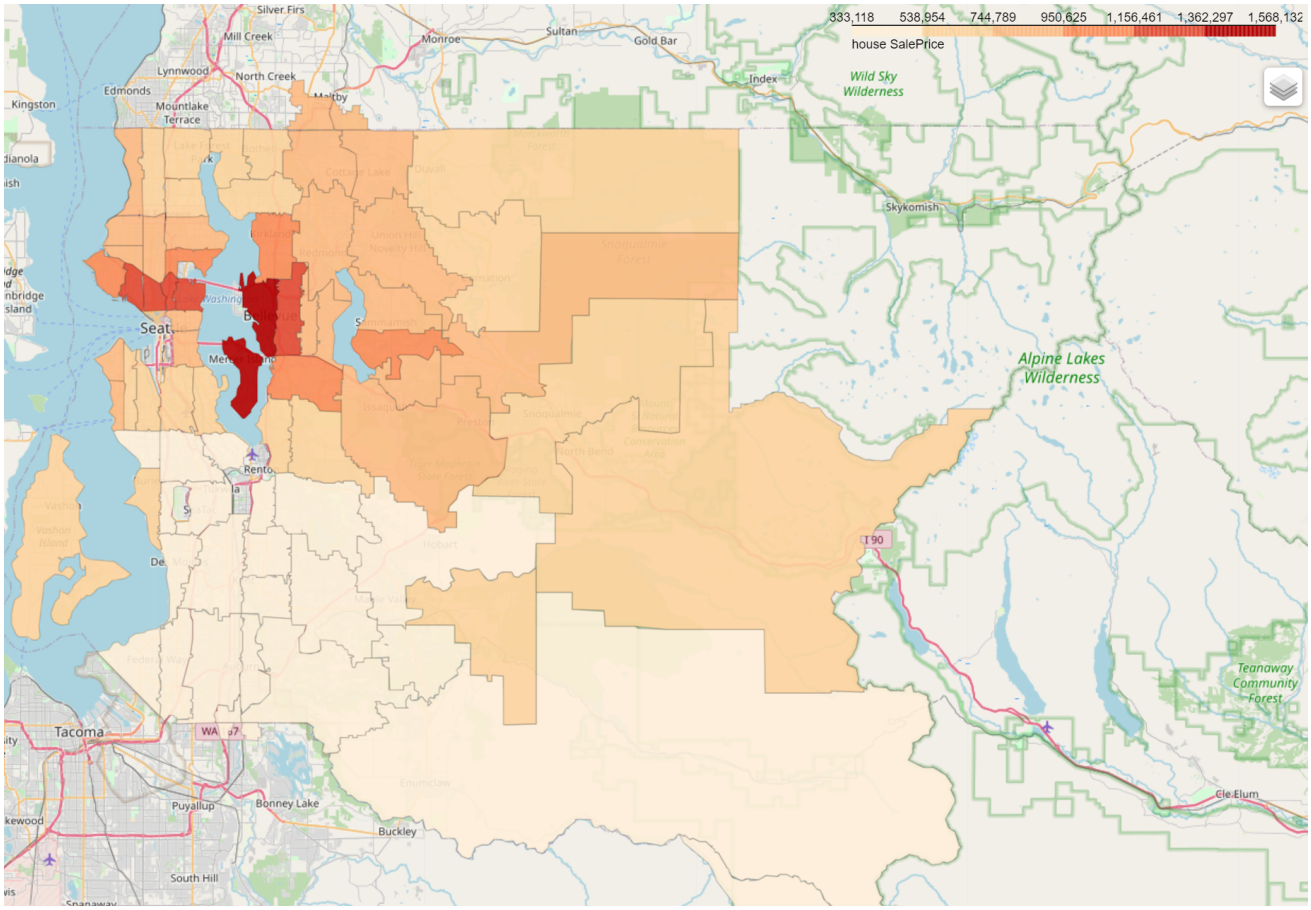
The purpose of this study is to undergo an inferential modeling workflow with the aim of determining which home improvements could have positive impacts on the sale value of a home.

This project will focus on single family households and homes sold in 2019 - non residential property is outside of the scope.

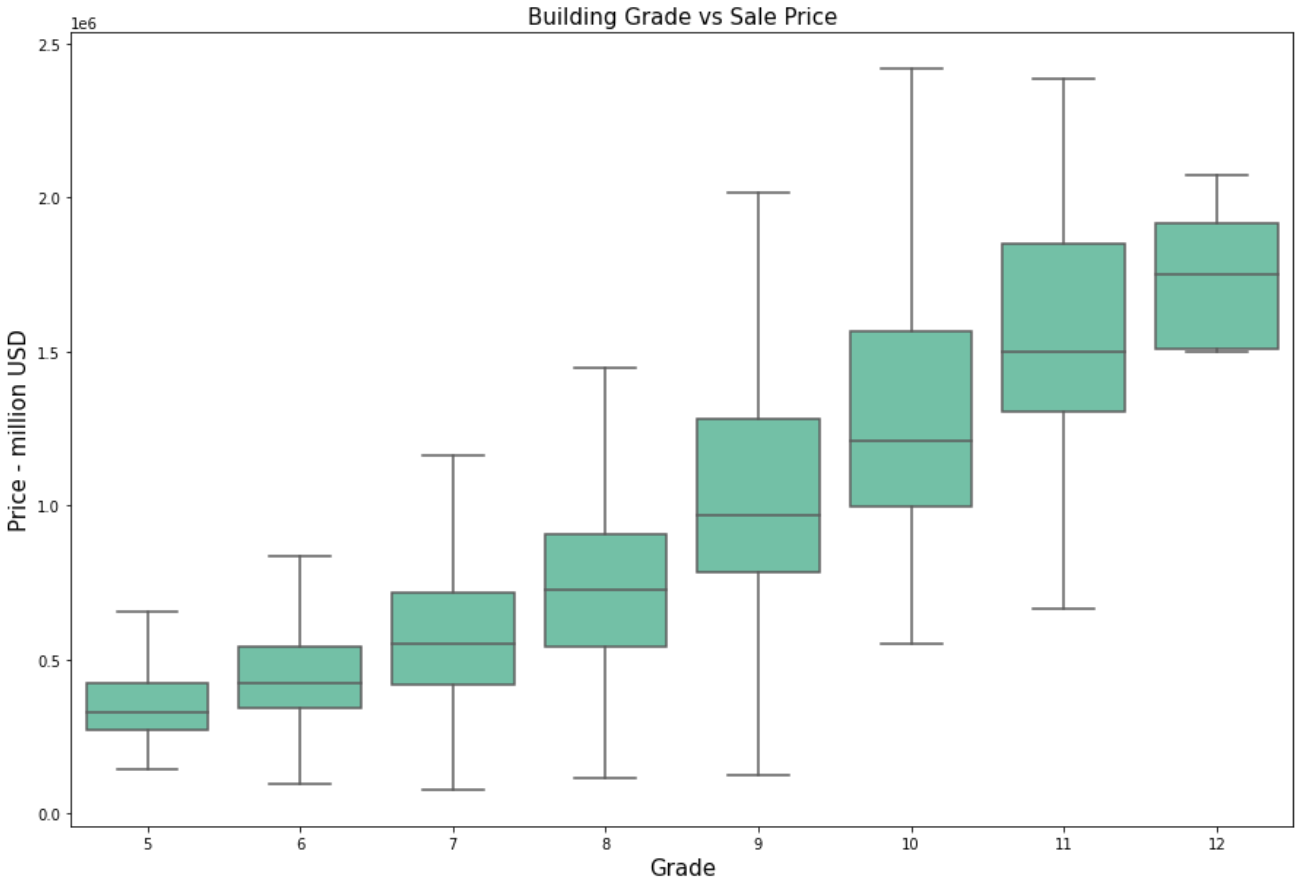
Data has been provided from King County GIS Center and the raw files are here

Data Understanding

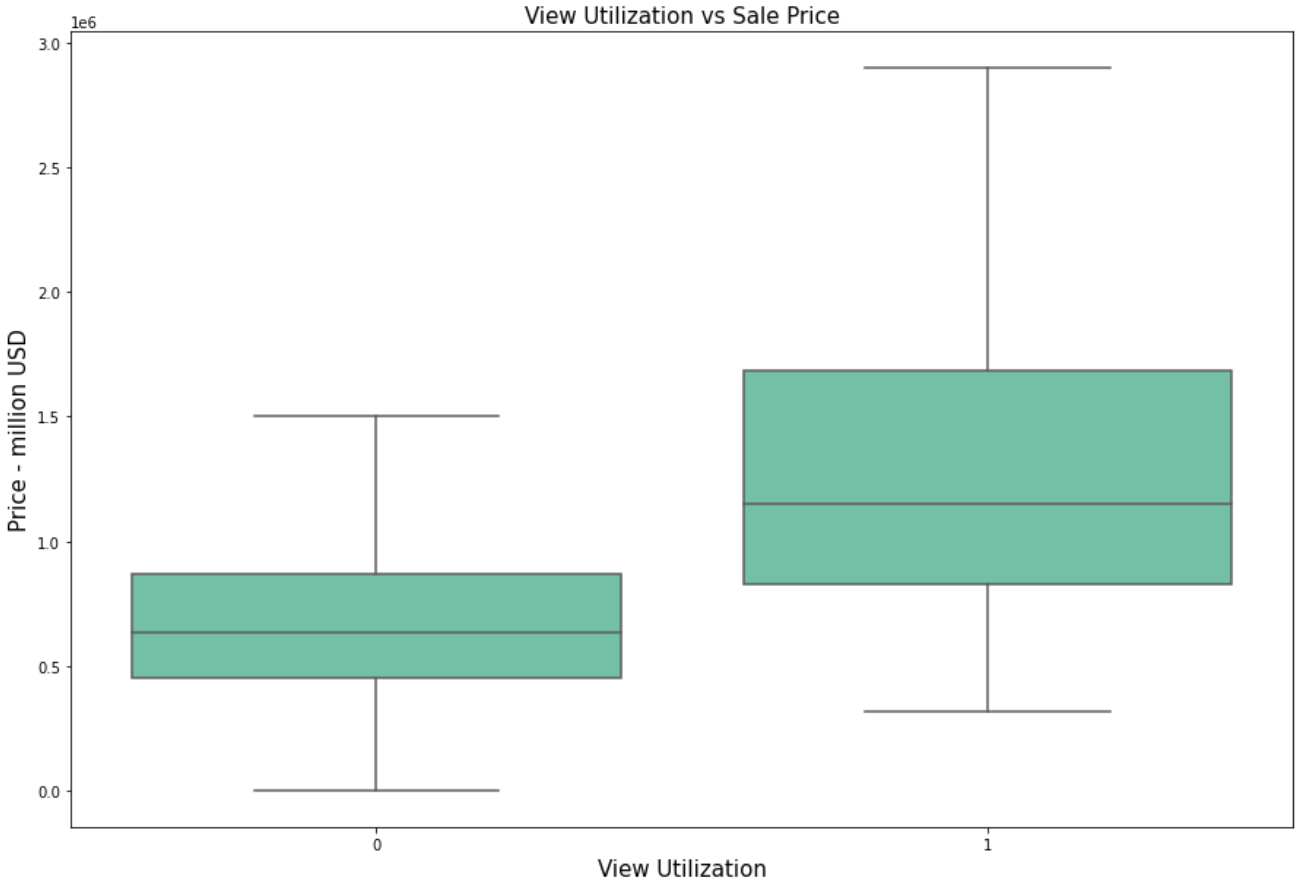
King County is the most expensive County in Washington and home to some of the most expensive homes in the United States. Within King County there is a big variation in property prices, but as can be seen from the map which represents each Zip Code in Washington and its average house price, the most expensive (dark red) are near Bellevue, Mercer Island and Medina. Like everywhere else in the world, location matters..I will create a feature that defines the minimum distance to 4 expensive areas in King County.



Not only does location matter, there is a strong linear relationship with Building Grade too, this could be an important feature in the workflow.



So building grade impacts price, but so does the view. View utilization which seems an intangible metric but it pays...homes with a view have a nice premium associated with them. This could be an important feature



Square Footage can have interesting relationships with home prices, given the trade off of location vs space. However in the case of King County, it appears there is a positive linear relation with square footage total living area.

Data Preparation

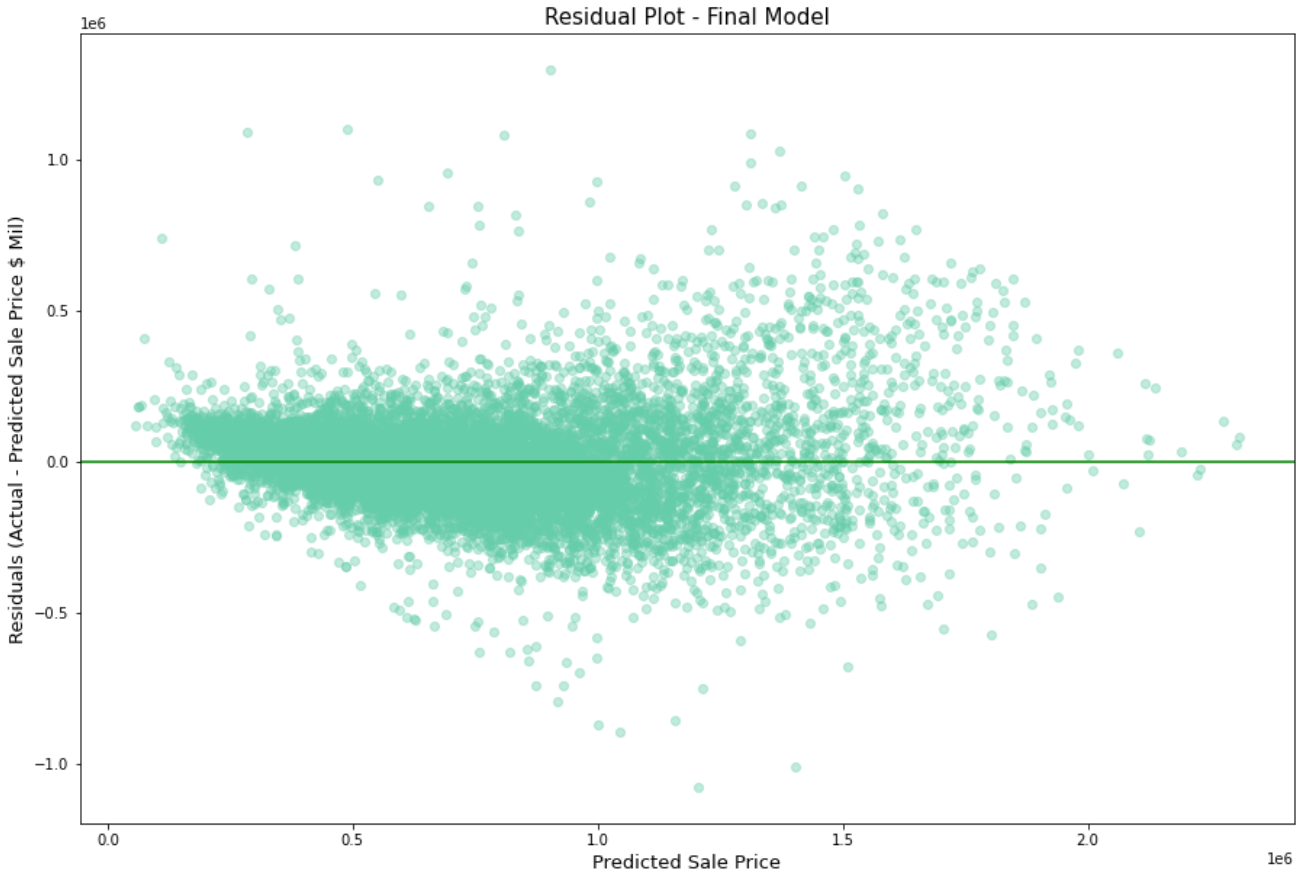
For detailed walk through of this, please refer to the cleaning notebook. There were many columns that were either blank, had many NaNs, or were filled with only one value for example 'N'. I tried to keep as many columns as possible until the first pass of the modeling phase as I did not want to discount something which could be important.

Modeling

The final model that was selected for recommendations had 119 features, many of which were dummies including Zip Code which makes up the majority of features. Most of the features are significant, however, I have not gone through the process of removing variables purely for insignificance. I have check for multicollinearity and tried to reduce that as much as possible using variable inflation factor.

The model has an R-Squared of 83.5 meaning it explains 83.5% of the variance in Sale Price. There are some areas that are worse than others in terms of error, meaning homoscedasticity hasn't been fully honoured despite efforts.

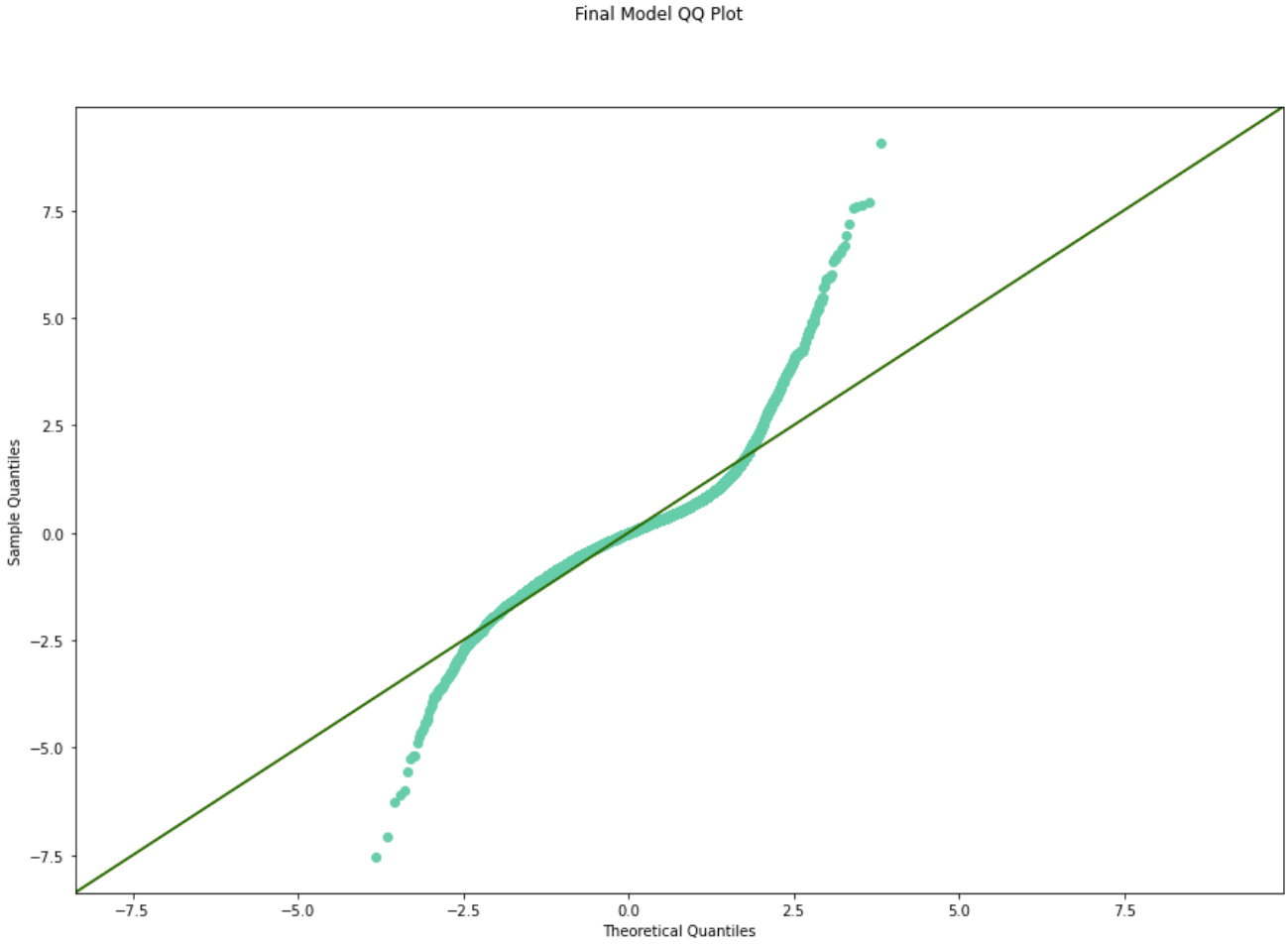
The model struggles towards the more expensive end of the sales prices, as can be seen from the residuals plot.



Evaluation

Despite best efforts, the residuals plot shows the homoscedasticity assumption isn't honoured as well as it could be, the normality of residuals also isn't as good as I would like. As can be seen by the table, I did refine this model further (7b), eliminating features with high VIF numbers. This reduced the JB number slightly but R-Squared suffered.

Model	Description	No. Features	R^2	Adj R^2	RMSE	RMSE sd	JB
Simple Model - one independent variable	Square ft Total living	1.0	0.389	0.389	398611.000000	342506.000000	9825842.0
Simple Model - outliers removed	Square ft Total living	1.0	0.354	0.354	294579.000000	168991.000000	7132.0
Multiple Linear Model 1	Continuous features	4.0	0.535	0.535	238663.000000	158954.000000	18348.0
Multiple Linear Model 2	Continuous features + age/dist_to_exp	6.0	0.567	0.566	236962.000000	153473.000000	16023.0
Multiple Linear Model 3	continuous + binary features	29.0	0.622	0.622	227275.000000	146517.000000	16416.0
Multiple Linear Model 4	continuous + binary features + cat	139.0	0.816	0.814	166951.000000	118312.000000	165256.0
Multiple Linear Model 5	MLP 4 + SqFt2nd + Clean Cat	123.0	0.823	0.821	164932.000000	118703.000000	206712.0
Multiple Linear Model 6	MLP 5 outliers longitude removed	116.0	0.834	0.833	158260.000000	111643.000000	32901.0
Multiple Linear Model 7	MLP 6 + bathrooms	119.0	0.835	0.834	157905.000000	111436.000000	32792.0
Multiple Linear Model 8	MLP 7 Log Sale Price	117.0	0.854	0.853	0.182373	0.073137	196501.0
Multiple Linear Model 9	MLP 7 Sqrt Sale Price	117.0	0.860	0.859	78.998208	38.394642	35088.0
Multiple Linear Model 10	MLP 7 cbrt Sale Price - adjacentGreen	116.0	0.860	0.859	5.462159	2.371203	53457.0
Multiple Linear Model 7b	MLP 7 reduced features due to VIF	115.0	0.819	0.817	163682.634285	115393.459951	28053.0



Conclusion

The following recommendations could be made based on the findings from the model

Increase Living Space

For every one square foot of total living space you add to a property, its price will increase by 119 USD assuming all other variables are kept constant. i.e if you add 500 square feet of living space this could add up to 60,000USD to the price. This is not an improvement that would be available in every property.

Add Bathroom

The bedbath feature was simply bedroom count subracting bathroom count. As this approaches zero or better still goes negative (i.e you have equal number of bathrooms to bedrooms or more bathrooms than bedrooms) the Sale Price improves. If you improve this ratio by one i.e add one total bathroom this will increase the house price by 29420 USD if all other variables are held constant.

Renovation

Renovating a house can me expensive but it can be worth it. The coefficient of the renovation feature is 58,220USD. Meaning if you renovate the house on average it will increase the house price by 58,220USD.

Install Porch

The interpretation of the coefficient for having a porch suggests having a porch will increase the price of your property by 17,560USD.

Improve general condition

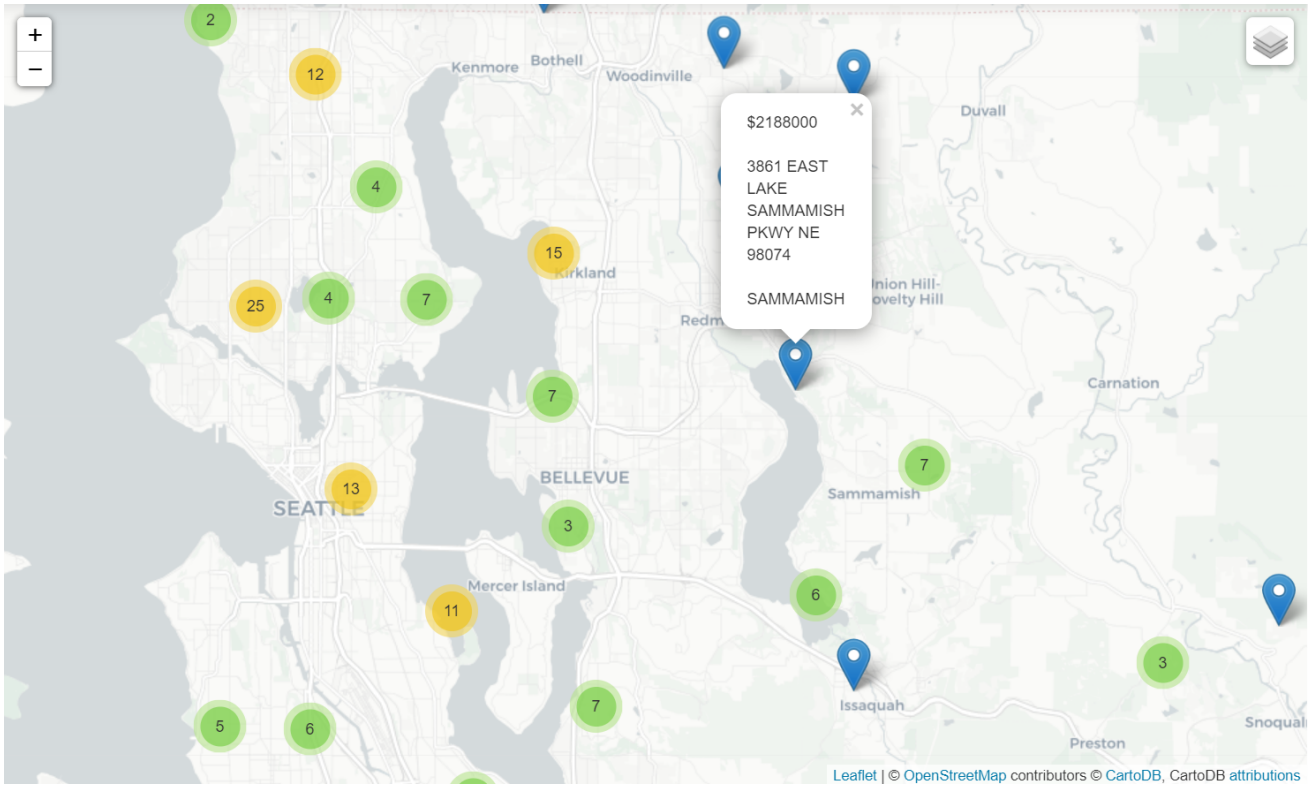
This is a slightly ambiguous feature, however it is assumed this means the general condition of the property. The difference here is stark though, assuming all other variables are kept the same, a 'good' condition home will sell for 33,950USD more than an 'average' condition property. for 'very good' this is even more - 70,980USD.

Fix Issues

Assuming all other variables are kept the same, a home with a water issue or some other unspecified will be worth 18,960USD less than a home that is issue free.

Further Work

As has already been established the model struggled with some of the higher value properties. I mapped the 200 worst offenders (biggest difference between predicted and actual), to check I wasn't missing one particularly desirable zone. It turned out as they were quite well spread out although from the map it is clear many of them are very close to water. It would be useful to engineer a feature that establishes the minimum distance to a water body. Example below is a property right on Lake Sammahish which is undervalued in the model.



Other features I would like to explore with time is proximity to good transport links, average commuting time to downtown Seattle and proximity to good schools.

This Repository

Repository Directory

```
├── README.md          <-- Main README file explaining the project's business case,
│                          methodology, and findings
│
├── data               <-- Data in CSV format
│   ├── processed      <-- Processed (combined, cleaned) data used for modeling
│   └── raw            <-- Original (immutable) data dump
    └── provided       <-- Provided data
│
├── notebooks          <-- Jupyter Notebooks for exploration and presentation
│   ├── old            <-- Unpolished exploratory data cleaning and preliminary
analysis (EDA) notebooks
│   └── report         <-- Polished final notebook(s)
│
├── references         <-- Data dictionaries, manuals, and project instructions
│
└── reports            <-- Generated analysis (including presentation.pdf)
```

Releases

No releases published

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 86.9%
- **HTML** 13.1%