Report - Housing Estimates

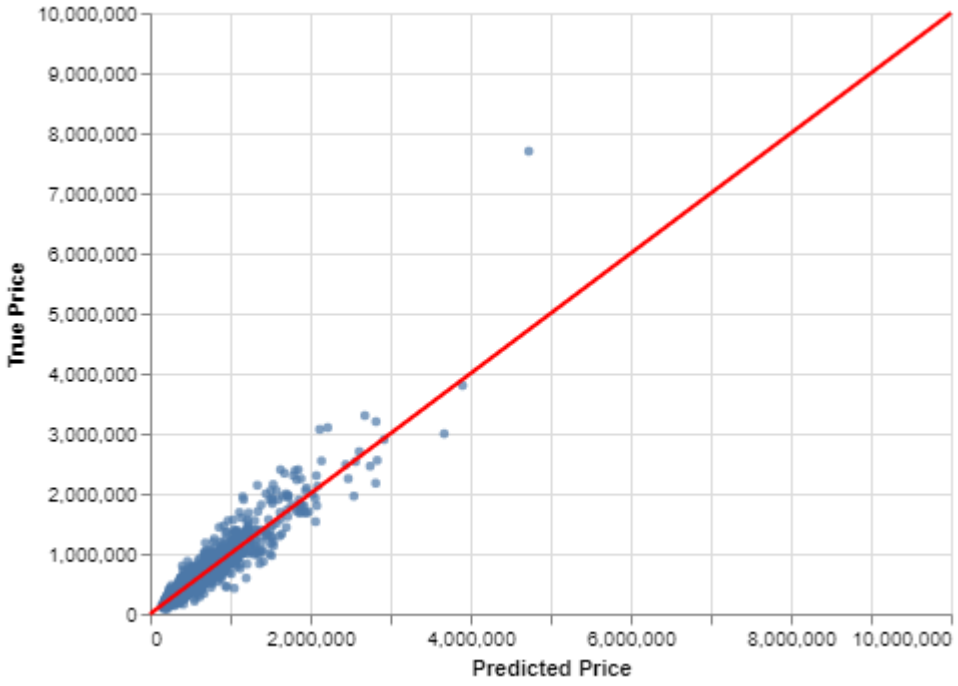
Seattle 2014-2015 Price Estimations for Home Insurance Agency

Course CSE 450 James Hall, Jacob Ferris, Landon Davis, Brad Strange, Keaton Kesler

Model

We included all of the features from the original dataset with the exception of id. We changed some data to a numeric value, and one-hot-encoded our zip code. We also included crime data for each zip code and King County price index for each quarter. We calculated yard square footage, square foot per floor, and miles per city, and the ratio of lot size and living space between each house and its nearest neighbors. We believe these new features helped us to refine our regression to better allow the model to fit the lower priced and higher priced homes alike by making features that put them on the same level.

After optimizing for r squared, the model depends most heavily on construction grade and square footage of living space, followed by the latitude and distance from the city center. Our model explains 89% of the variance in the data (r squared = 0.89). The residual (difference between the actual price and our predicted price) have a standard error (root mean square error) of $123,000. Our predicted prices are typically off by around 17% on a house-by-house basis (root mean square log error). We believe a lot of the error is due to the larger variability in higher priced homes and our true error for most of the model is actually lower as we do not have many extremely expensive homes. Our standard error of $123,000 Gives us a range for how close we believe our model could be to the truth.

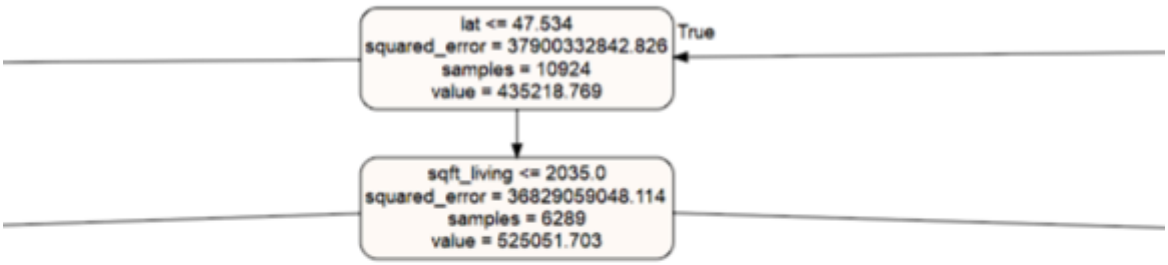


As you can see from our Actual vs Predicted graph, our data is very tight around our line for a model that predicts everything correctly. You can see a very tight cluster in the bottom left but as we move towards the right, we can see our points get further from the line. This could be due to the fact that as prices get larger there is more variability and predictions have more to predict where in the lower end you can only be off so much as mentioned above.

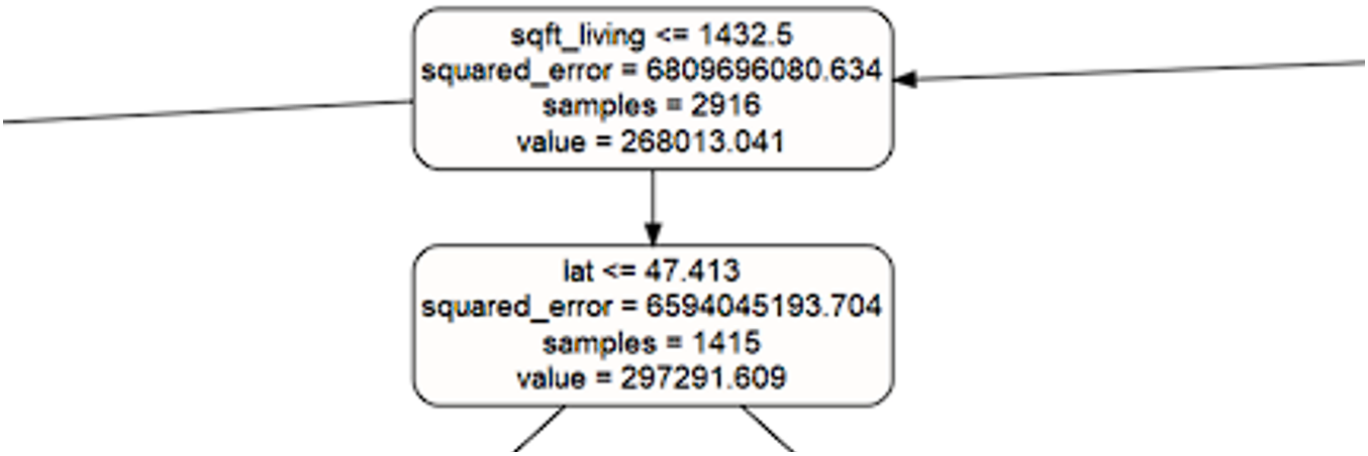
Decision Tree:

Multiple sections of the final decision tree we chose to use for our model:

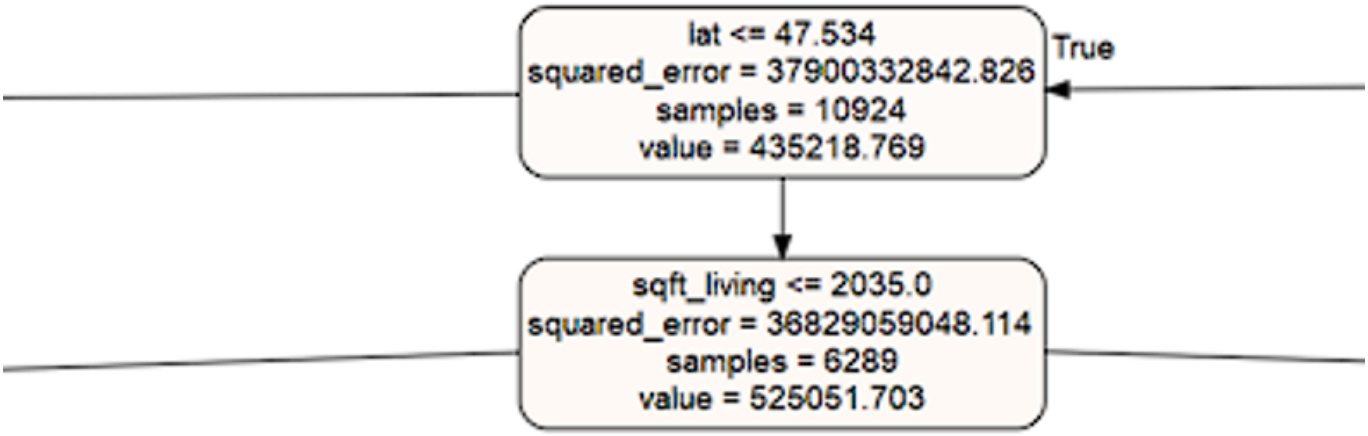
Section 1:



Section 2:

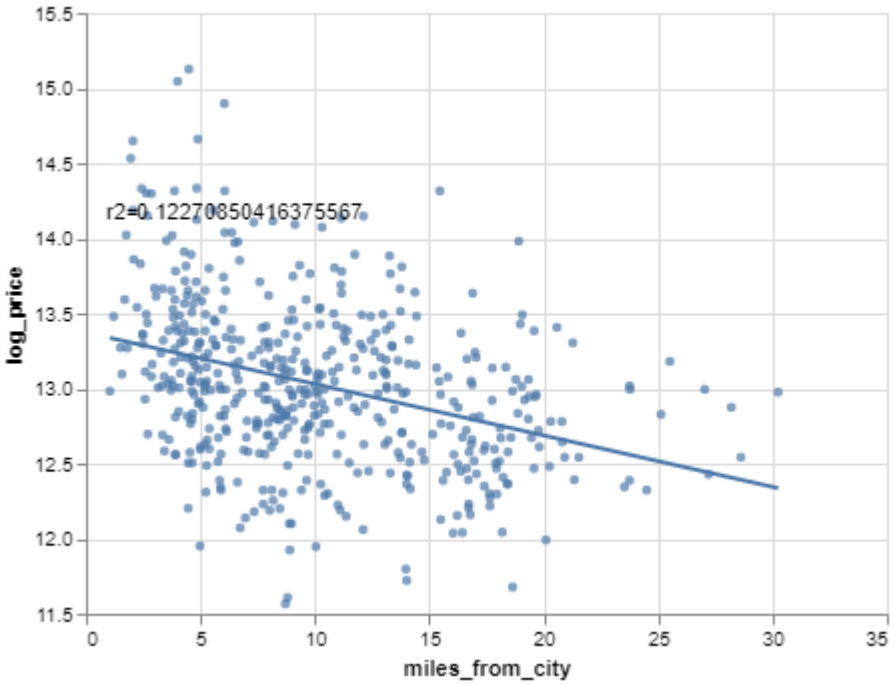


Section 3:



Visualizations

Miles From City :

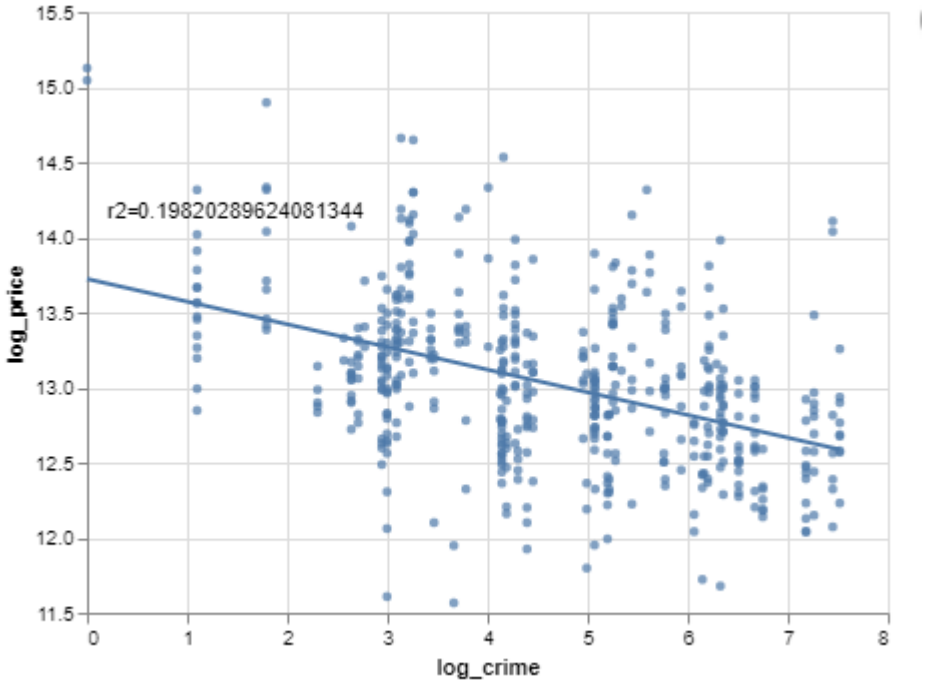


Here we can see some correlation that as we get further from the city that housing prices decrease as we move further away from the center of the city as a general rule. We performed a log transformation to allow all the ranges to be easily seen in

one chart. Here a .1 change in price represents a 10% change in the price

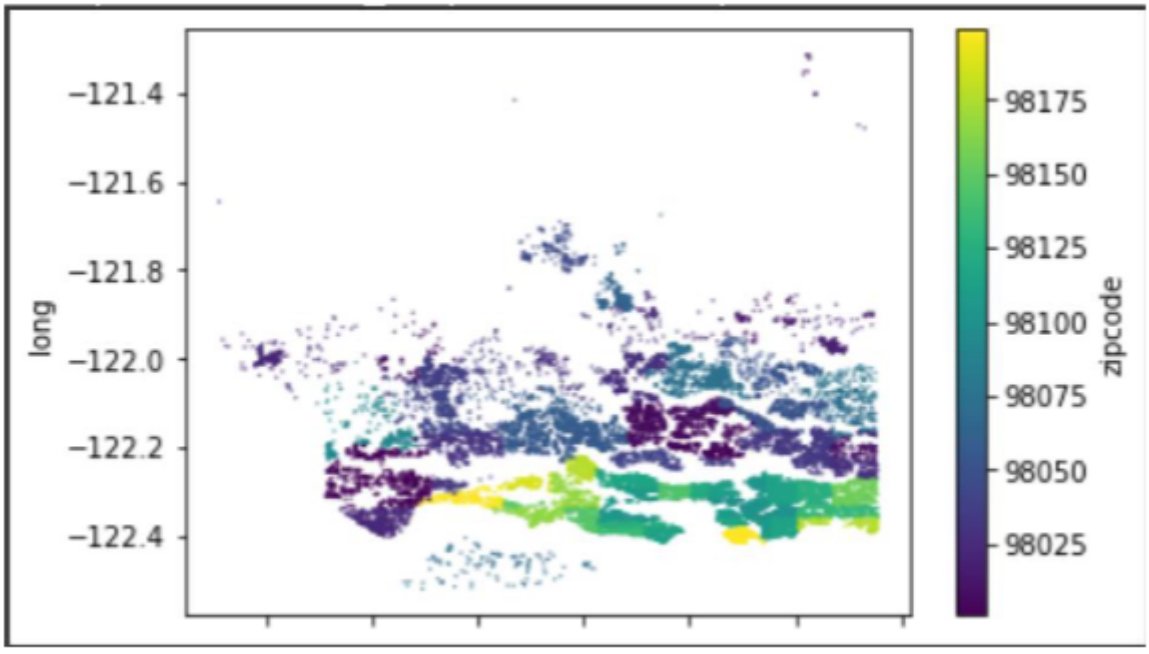
Housing price vs crime rate - a 0.1 change in log_price or log_crime represents a 10% change in the price or number of crimes, respectively.

Crime:



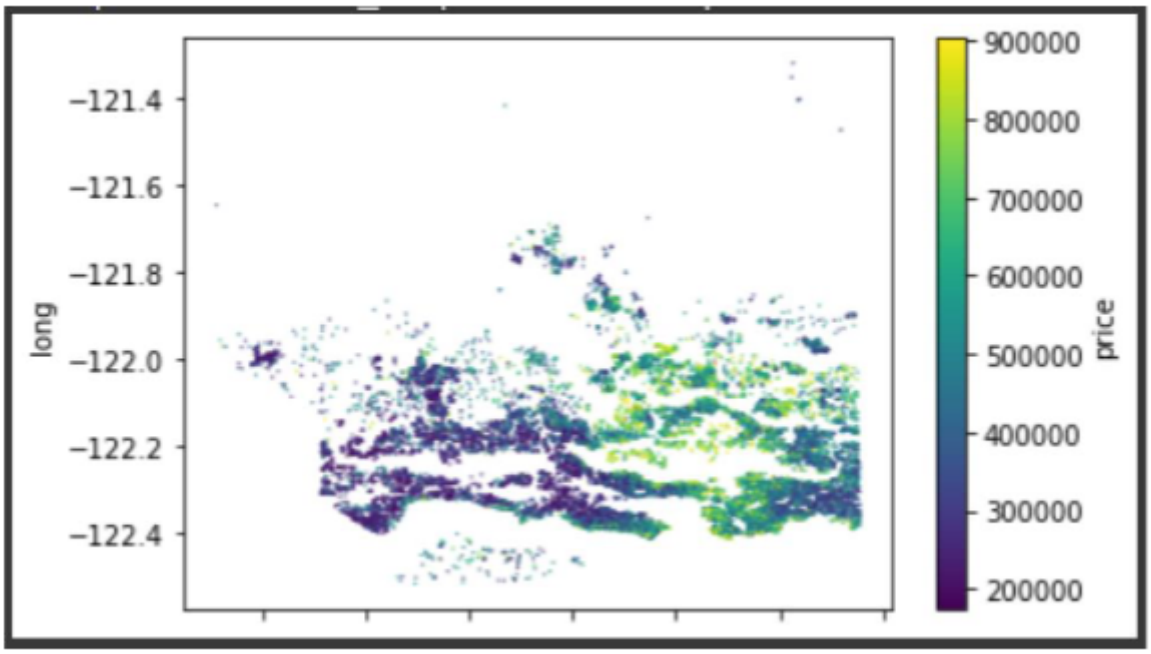
Similar to our other comparison, we can see that as the crime rate of an area increases that the price will start to decrease as we would expect to be the case.

Zipcode:



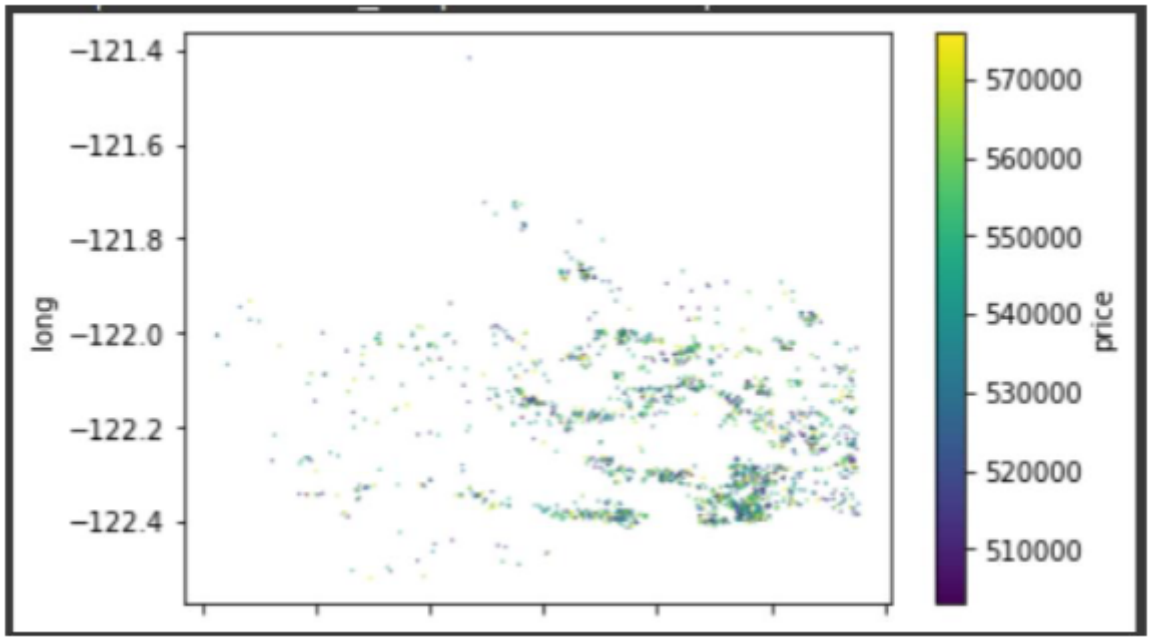
- **This chart is showing all the houses by there Zipcode.**

Overall House Prices:



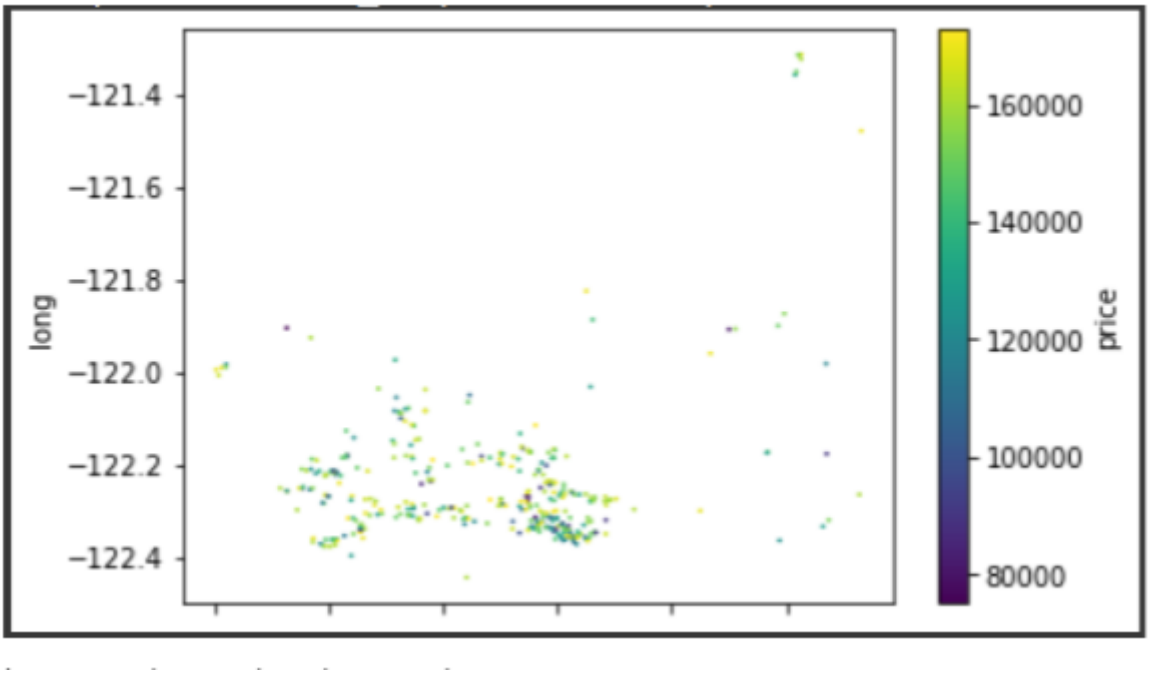
- **This chart shows the overall prices based on there location. You can see the prices of housing change if they are in the country or city and by lower income neighborhoods to high income.**

Median House Prices:



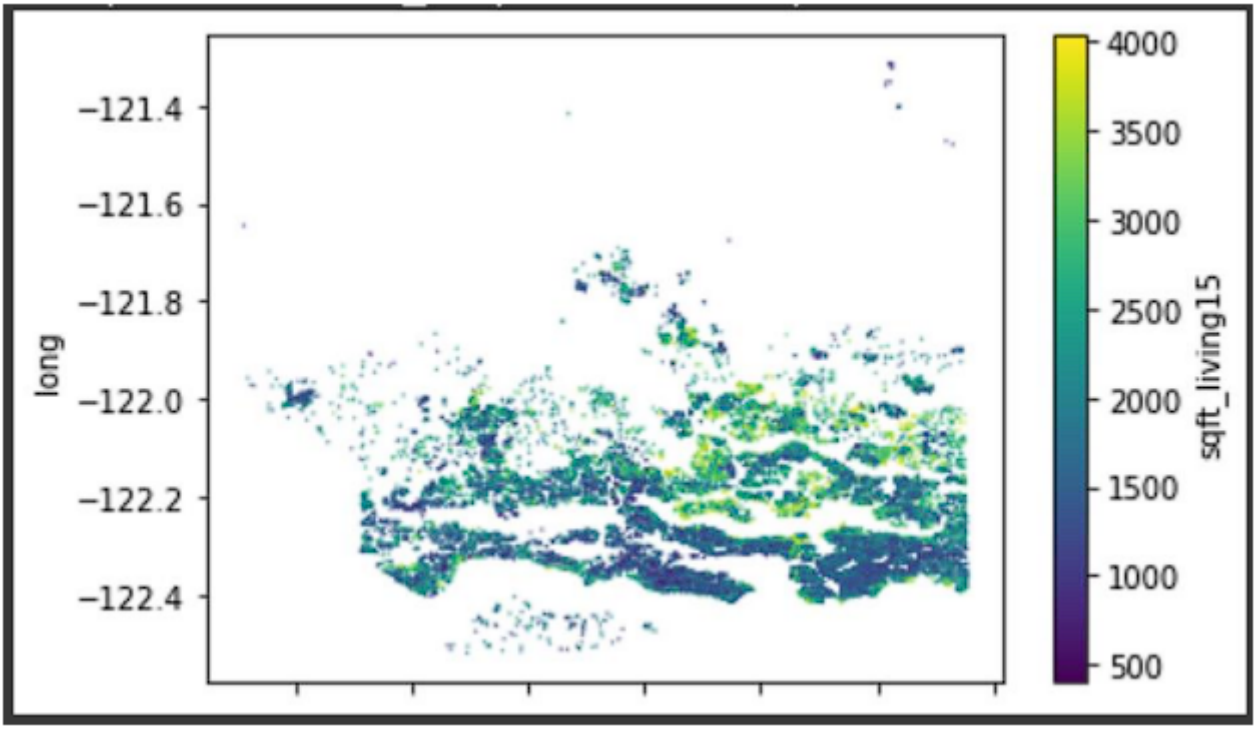
- **This chart shows all of the meadian house prices by location.**

Low Income Housing:

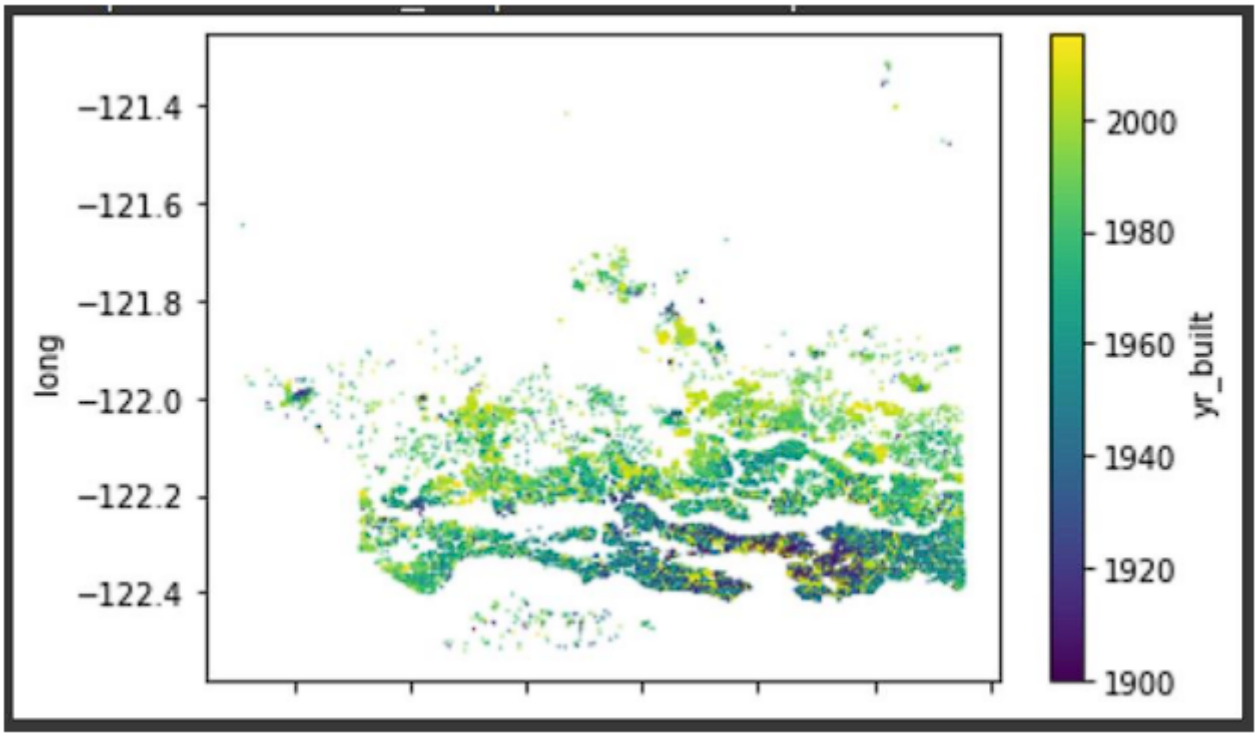


- **This chart shows the Location of the Low income houses.**

House Size:

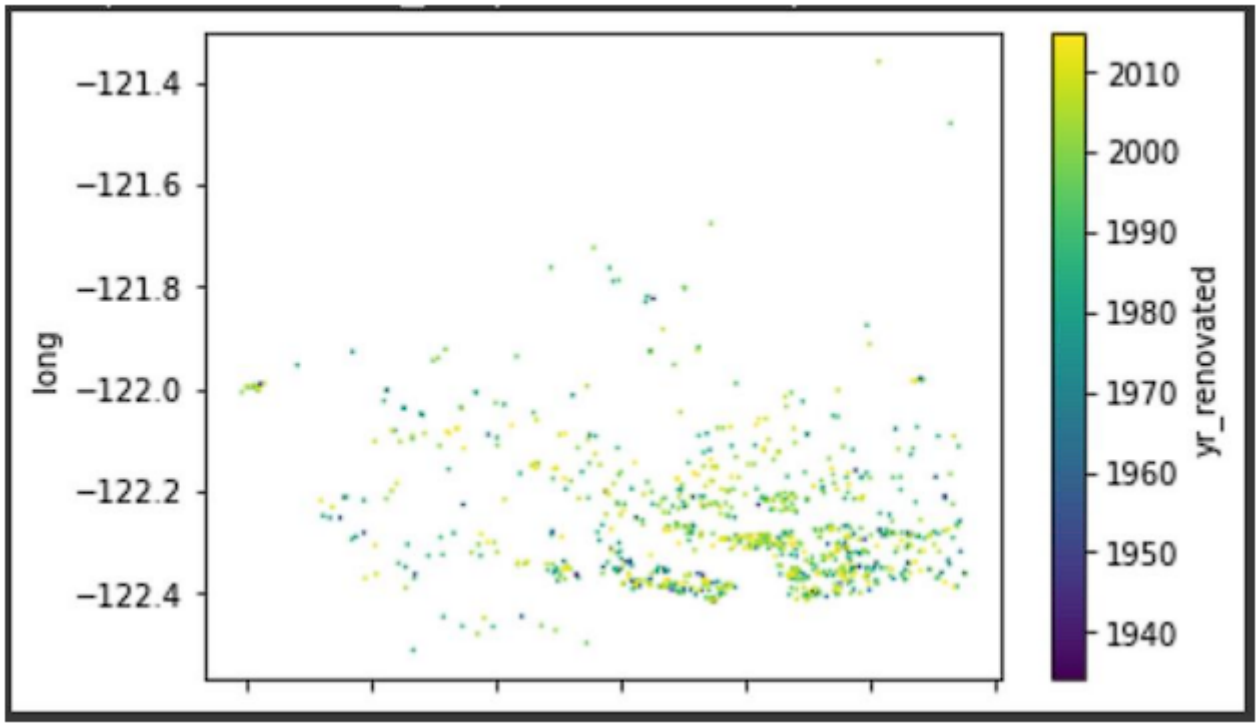


Year Buit:



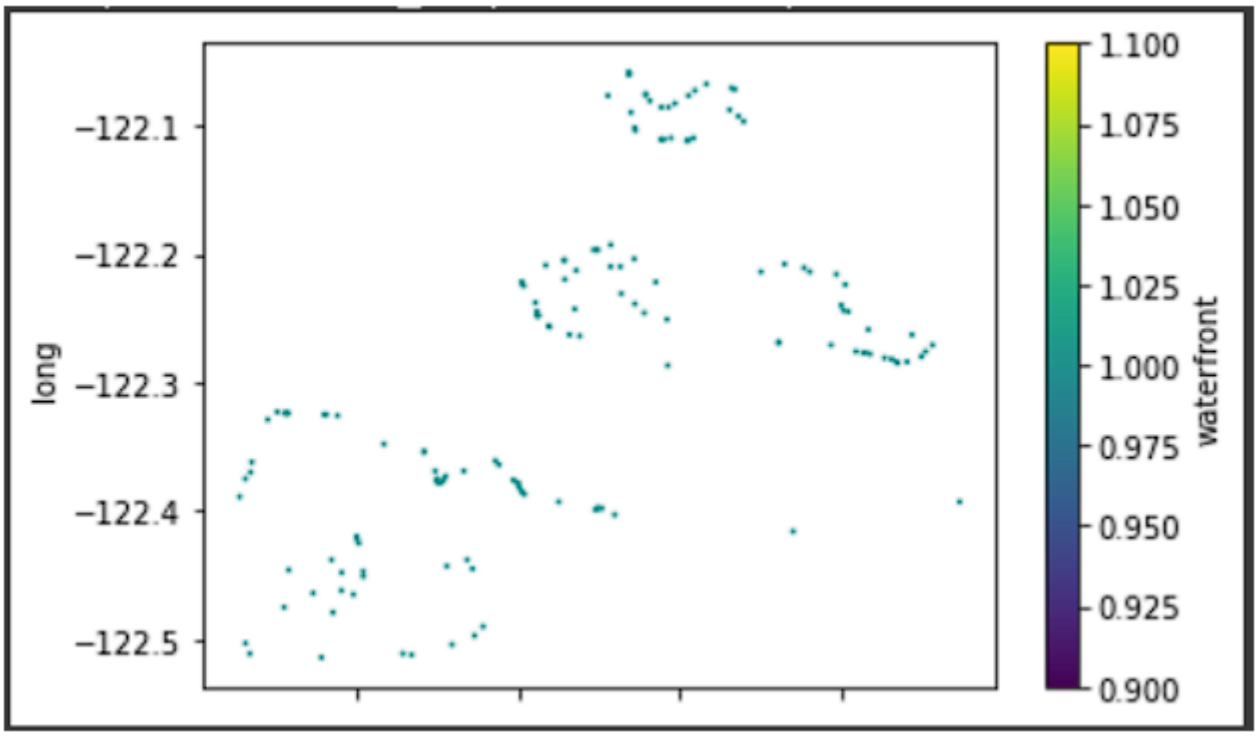
- **This chart shows by year when houses where built and you can see which areas houses where being built at first.**

Renovated houses:



- **By this chart you can see that the house that where built first(the oldest) are the ones that are being renovated the most. You can see that they are mostly in the city.**

Waterfront Houses:



- **This chart shows which houses have waterfront locations.**

Why we chose R2

R-Squared is a great measure for seeing how well our line fits the data. All data varies and r-squared tells us how much of our variability we can explain with our model we have designed. It is a quick and easy way to see correlation without much knowledge of the data and in a quick measurement. R-squared is measured from 0 to 1 and the closer to one the value is, the tighter the data is.

Conclusion

We feel we can predict house prices with a value below $2 million very well from our model based on the features given and created. This will be beneficial to the company to help us cover homes for the right cost and to help the company make the most money possible.

Discussion Responses

Below are our final answers to the case discussion questions:

The biggest thing I want to see is quantifiable evidence that the predictions we come up with are reliable.

Due to the nature of how R^2 works, we can easily understand correlation easily and can see the tightness of the data too around the model we have created.

I'd like to know which property types are weighing most heavily in the house prices predicted by your model. My excel spreadsheets can tell me that information for our current methodology...can your so-called artificial intelligence do the same?

As stated in our report, the property types that seem to weigh the most heavily on pricing are those with good condition, large square foot living, and distance from the city. If you were to purchase a house right

now, wouldn’t those be some of, if not the highest priority features you would look for?

One other question the board was wondering about, is if there are additional factors about these areas that might be affecting prices, which we aren't taking into account.

There are a multitude of factors that go into a house's price. Including all of the features that are a part of the dataset already, we have chosen to add how close the house is to the center of the city. As well as, the crime rate for the city. Prices of homes near the city are likely to be higher than those further away due to jobs, urbanization, and quality of distance in relation to stores. Crime rate is another key factor in housing prices as most of the time high crime rate decreases the values of homes in the area.

APPENDIX

Python Notebooks

Below are Github Gist links to the notebooks we used during this case study:

Model Notebook

- https://colab.research.google.com/drive/1oWQDneGi5T-Hei8a25DkYQTmNintZfQw#scrollTo=eLFW8-KFP3ya

More Notebooks:

- https://colab.research.google.com/drive/1qn_CBjoR5VJyk_MBS0AIvAlmCTWk83rD#scrollTo=JQG5pJLUCqXQ
- https://colab.research.google.com/drive/1h5B0QPFyTrxz22xLVhUTDtpIdHXhdhvN?usp=sharing