PHASE 2 PRESENTATION

Linear regression project with King County house data

Angelo Turri

Stakeholder

- BlackSock is a large real estate company
- They want to diversify by buying houses in King County, WA
- Plan: buy homes, rent them until their value goes up a sufficient amount, then sell

Problem

- BlackSock needs insight into what factors influence a home’s price

Data

- Obtained from King County government website
- 30,000 records of home sales
- June 10th 2021 – June 9th 2022
- Limited in the amount of information it offers

Data – Important Variables

- Date built, date sold
- Price sold for
- Square footage of living space, space above grade, lot, patio, garage
- Number of floors, bedrooms, and bathrooms
- Condition of house at time of selling
- Address
- Latitude/longitude

Modeling

- Data was used in a linear regression model
- This type of model is well suited for the data and problem, and it allows us to determine relationships among data as well as make predictions
- This has the potential to yield useful insights for our stakeholder

Data cleaning

- Eliminated outlandish values from the dataset (houses that cost more than 10 million dollars, over 7500 square footage of living space, etc).
- Any data that does not adequately meet the standards of linear regression should be eliminated from the dataset.
- The closer you come to perfectly meeting those standards, the more useful your model’s results will be.

Feature engineering

- We created numerous variables in addition to the ones seen in our original dataset in the hope of creating more good predictors of price to use in our model.

Data used

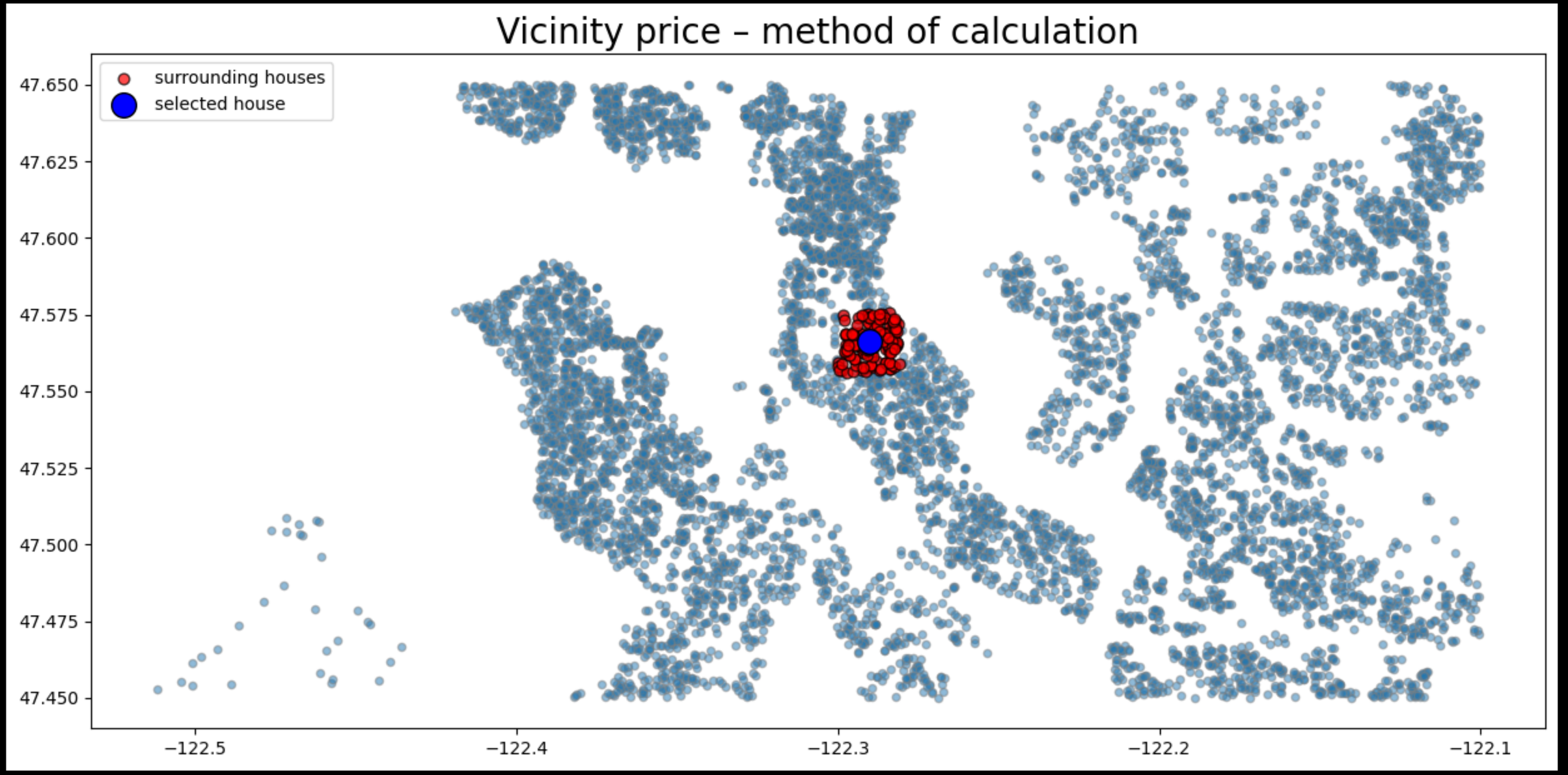
- Average price of surrounding homes (“vicinity price”)
- Home quality (“grade”)
- Square footage of living space

“Grade” data – Explanation

- 1 (Cabin), 2 (Substandard), 3 (Poor)
- 4 (Low), 5 (Fair), 6 (Low Average)
- 7 (Average), 8 (Good), 9 (Better)
- 10 (Very Good), 11 (Excellent), 12 (Luxury)
- 13 (Mansion)

“Vicinity price” data – Explanation

- Using latitude & longitude data, we took a small surrounding area for each house and then calculated the average price of homes in that area
- We eliminated the house in question from that average, so that it wasn’t predicting its own price in the model
- Houses that had very few or no houses in their general vicinity were eliminated from our dataset


Vicinity price – method of calculation
surrounding houses
selected house
47.650
47.625
47.600
47.575
47.550
47.525
47.500
47.475
47.450
−122.5
−122.4
−122.3
−122.2
−122.1

Model results

- We used an iterative approach to modeling; that is, we started with a basic model and made adjustments from there.
- A house with average living space, an average grade, and average-priced surrounding homes will cost 1.1 million dollars.
- For every dollar increase in surrounding homes, a house’s price goes up by about 76 cents.
- Every increase on the home quality scale leads to a $85,000 increase in home price
- Every additional square foot of living space increases house price by $260.

Recommendations

- Our data was limited, and as a result, our model can only give preliminary insights at best.
- Although we recommend the stakeholder take note of our model's results, we also recommend the stakeholder consult other models in addition to our own before making any significant decisions.
- We recommend buying houses whose surrounding properties are likely to go up in value.
- We recommend increasing home quality if doing so costs less than $85,000.