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

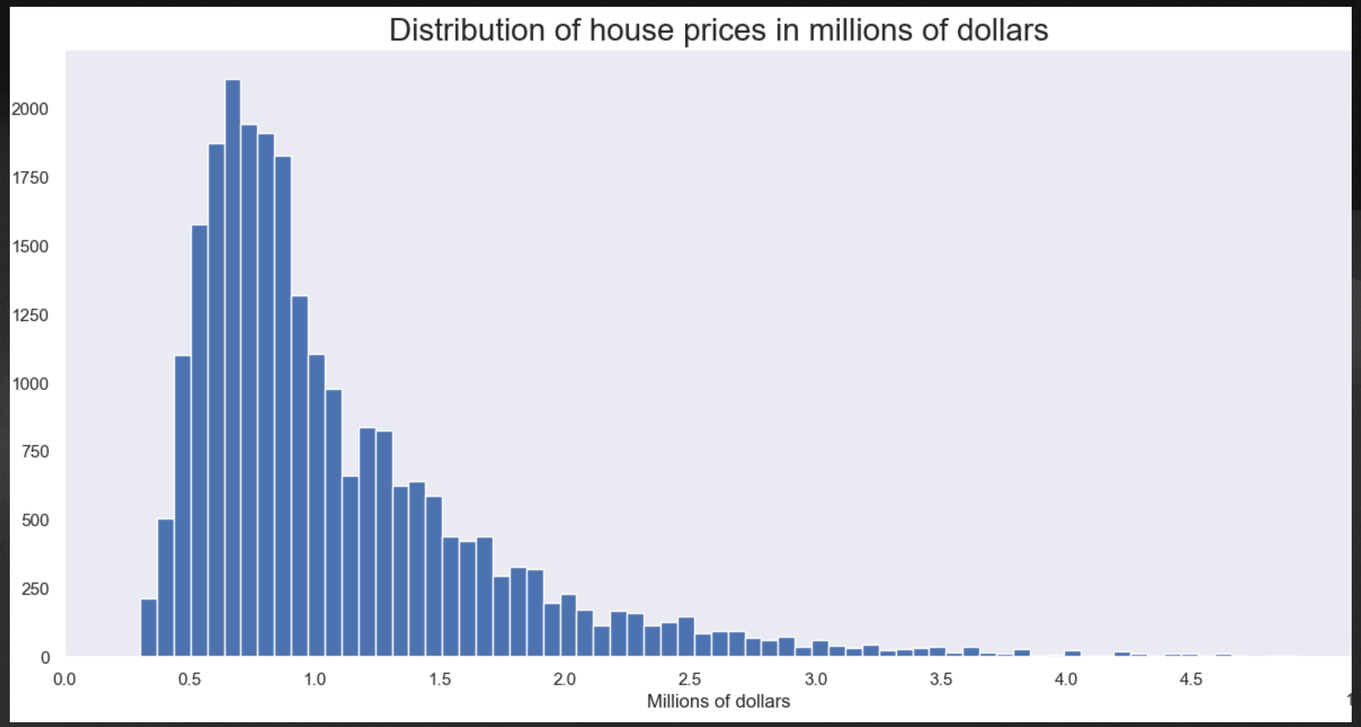
- Obtained from King County government website
  - https://info.kingcounty.gov/
- 30,000 records of home sales
- June 10th 2021 – June 9th 2022
- Has all the basic home information

Data – Important Variables

- Date built & sold
- Price sold for
- Square footage
- Bedrooms & bathrooms
- House condition
- Address
- Latitude/longitude

House price distribution

- This is a histogram of house prices in the dataset.
- Most house prices are between $300,000 and $2 million.



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

- Started with roughly 30,000 values, ended with 26,000
- Data eliminated
  - Outliers
  - Null values
  - Data incompatible with linear regression
- All visualizations in this presentation were created with the cleaned dataset.

Data used

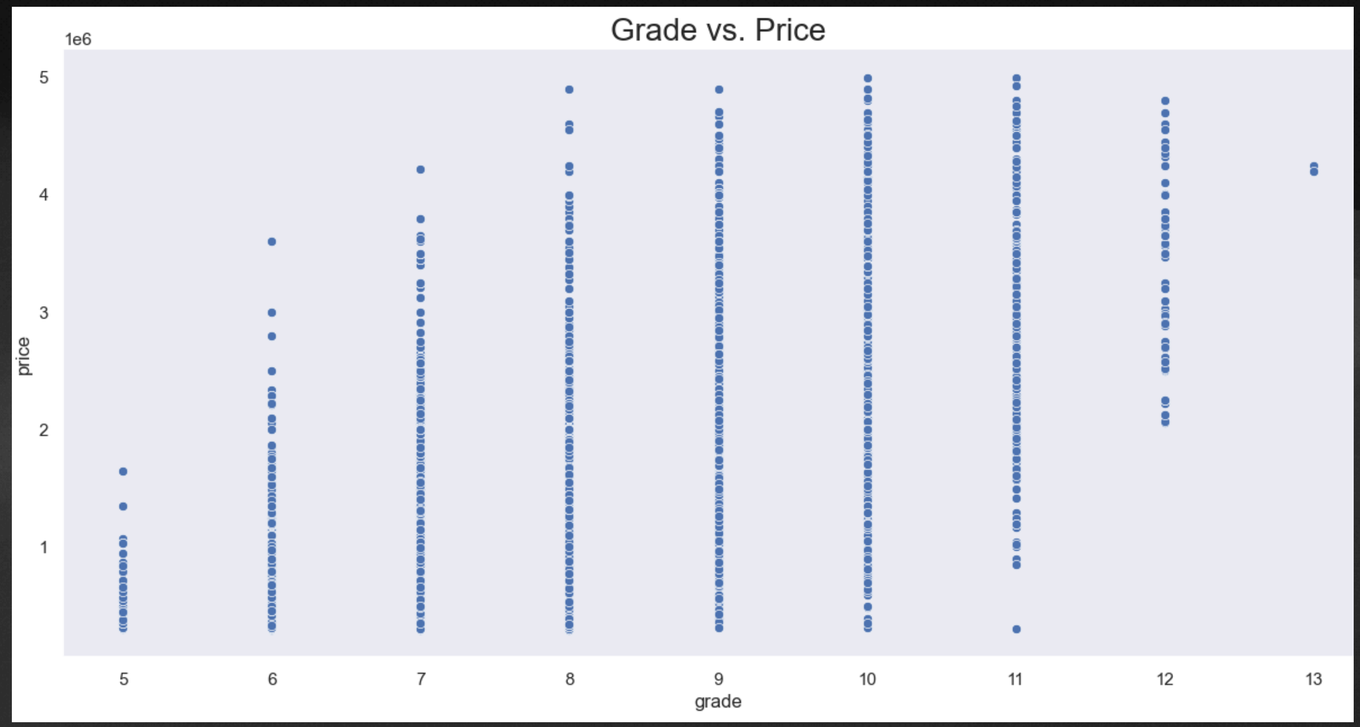
- Three variables
  - Average price of surrounding homes (“vicinity price”)
  - Home quality (“grade”)
  - Square footage of living space
- In the upcoming slides, I will show scatterplots of each of these variables vs. house price
- These scatterplots show that each variable has a linear relationship with price – as one variable goes up, price also goes up.
  - This made each of these variables good candidates for use in our model.

“Sqft_living” variable

- This variable measures the square footage of living space in a home.

“Grade” variable

- Measures the quality of a home on a 13-point scale after being sold.
- Houses ranging from 1-4 on this scale were considered outliers and eliminated from the dataset.
- Full scale:
  - 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)

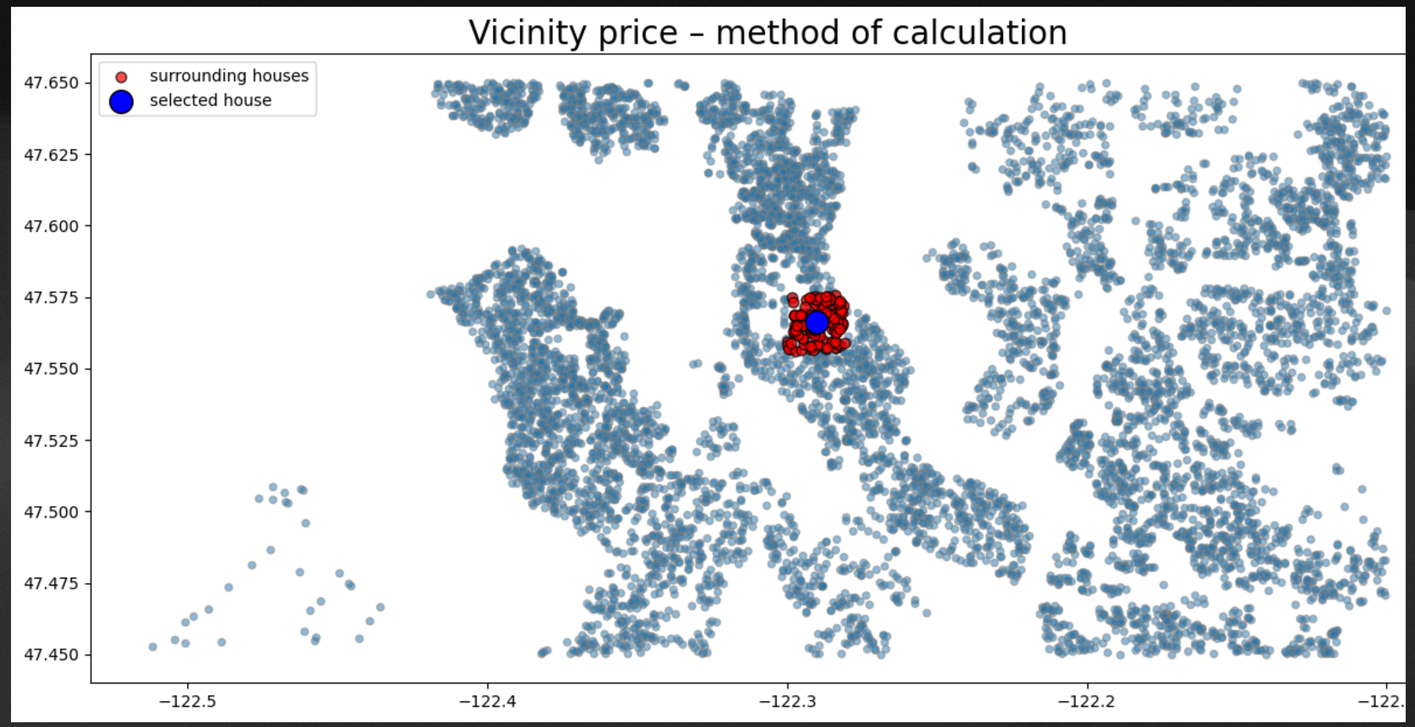


“Vicinity price” variable

- Measures the average price of nearby homes for each house in the dataset.
- This “average” price does not include the original house for which it was calculated, to avoid the house predicting its own price in the model.

Vicinity_price visual demonstration

- This is a visual demonstration of how we calculated the vicinity_price variable.
- For each house, we used its precise location in latitude/longitude to draw a small square-shaped area around the house.
- In this area, we averaged the price of all the surrounding houses (the red dots).
- We eliminated the original house (the blue dot) from the average to avoid houses predicting their own price in the model.



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.

Model performance

- The adjusted R-Squared of the final model was 0.735
- The MAE (mean absolute error) of the final model was $222,000
- These metrics show that the model is a decent fit for the data

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.

Thank you!

- All questions are welcome
- Contact me at: angelo.turri@gmail.com