Making Sense of Houses Sale Price in King County

Modelling Dataset With Multivariate Linear Regression To Predict The Sale Price of Houses

Executive Summary

Key Question [The problem we are trying to solve]

What are the predictive power of independent variables such number of house viewing, number of bedrooms, zip code and others on the house sale price of a house in King County?

The Behavioral Questions

1. What type of lifestyle choices of house buyers prefer?
2. What are the highly desirable zip code or geography in King County?
3. What are the key house design characteristics that are highly attractive to house buyers in King County?

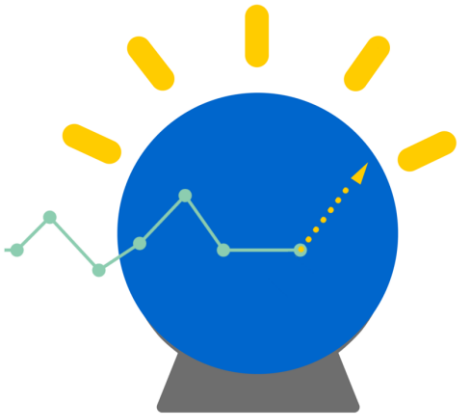
The Trend Questions

1. What are the latest house price trends in King County vs whole of US?
2. What are the trends in housing construction?
3. What is driving demand and supply of bigger or smaller house?

The Predictive Power of the Current Model [The predictive algorithm for house sale price]

INDEPENDENT VARIABLES has 65% Explanatory Power of the House Sale Price in King County

(Bedrooms + Bathrooms + Square Foot Living + Square Foot Lot + Floors + Waterfront + View + Condition + Grade + Year Built + Year Renovated + Zip code)



THE OTHER INDEPENDENT VARIABLES has 35% Unexplainable Power

(Likely due to error in the model + other data-points not part of the King County House Sale Price)

Recommendation [1]

RECOMMENDATION 1

For a home seller wants to maximize his/her house sale price, focus on improving the following house characteristics/features (in descending order)

Key Takeaways [The factors that will improve house price in King County – In Descending Order of]

1

Square Foot Living:

A bigger square foot living commands higher house price (the more living space, the better). A larger living space provides more space for family or living flexibility e.g. a game room, cinema room, rumpus, etc. Therefore, house buyer will pay more for with larger living space flexibility and convenience.

2

Year Renovated

The more recent renovated house, the better the house price (home buyers like it new, less hassle). A typical house buyers like the convenience of moving into a house and enjoy the facility right away rather needing to worry about renovation hassle and of course additional renovation cost.

3

Waterfront

A house near water feature / facility commands higher house price (water seems to relate to a more affluent lifestyle). Houses that located near water are limited in supply (especially in King County), and typically houses near water has higher visual attractiveness - blue and calm. More importantly, a house the water has added convenience advantage of water sport or water activities such as canoeing, boating, etc. (just take a stroll down the river/bay). Therefore, house buyer will pay more for a house near the water.

4

Bathrooms: A house with sufficient bathrooms commands higher house price (the convenience of having bathroom choices seem like a good thing, especially if you have a big family).

Number of House Viewing: A house that experienced a high rate of potential home buyers viewing commands higher price (higher house view means it is attractive).

5

Floors: A house with addition floors commands higher house sale price (who doesn't one additional living space, up to one point of course). Similar with square foot living, additional floors would means more living space convenience and flexibility.

Recommendation [2]

RECOMMENDATION 2

For a home seller wants to maximize his/her house sale price, focus on others factor to get that extra edge

What else a house seller can do?

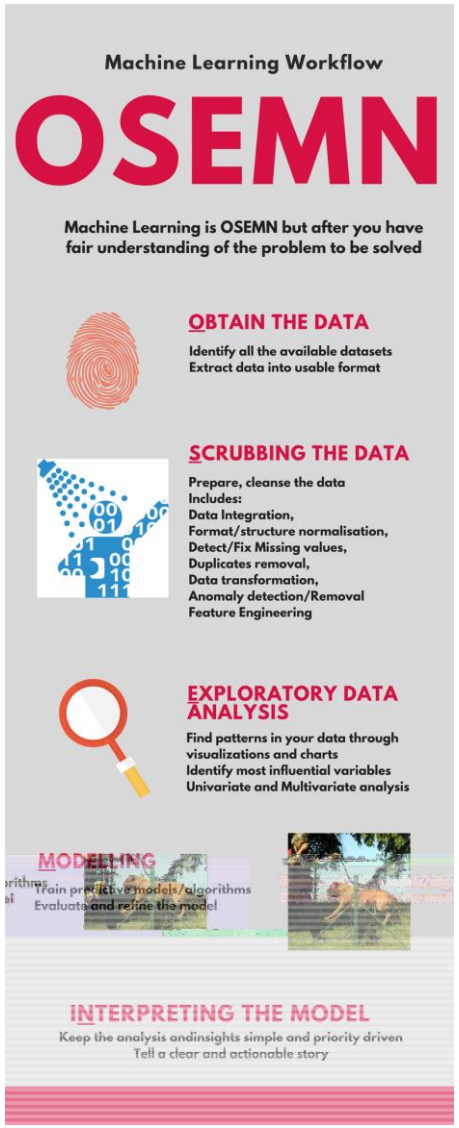
1. Attractiveness of the street (emotion and physical characteristics) - leafy and away of high volume traffic generally commands higher house sale price

2. Proximity to school (physical characteristic) - close to school means more convenience for children drop-off, and if a house in the school zone, generally it commands higher house sale price.

3. Gentrification of an area that the house is located (emotion and physical characteristics) - House buyers who are affluent, generally like to buy a house near to someone like them in wealth and lifestyle.

4. Bank lending criteria (economic factor) - Banks typically has restricted bank lending rules for certain location, this likely to impact house sale price or price growth of a house.

Research Approach – OSEMN Data Science Workflow



THE OSEMN DATA SCIENCE WORKFLOW

This data science project will use the OSEMN data science workflow. The followings are the summary of the steps taken:

STEP 1 OBTAINING THE DATA

- Importing relevant python libraries
- Initial data scanning
- Initial data visualization scanning

STEP 2 SCRUBBING THE DATA

Remove '?' in the sqft_basement dataset.

- Detecting and fixing missing values.
- Investigating potential multicollinearity.
- Transposing to the appropriate datastype
- Removing unnecessary datapoints/ columns.

STEP 3 EXPLORING THE DATA

- Strategizing on how to explore the newly scrubbed data
- Understanding the correlationship between jointplot, spearman correlation, p-Value

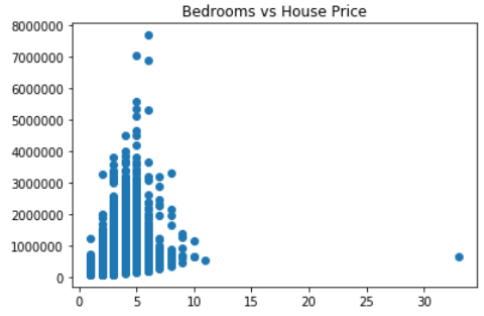
STEP 4 MODELLING THE DATA

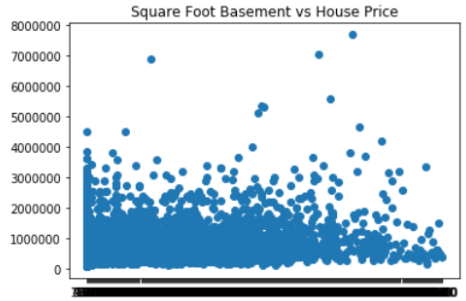
- Train and test split
- Improving the modelling fitness
- Fitting the model

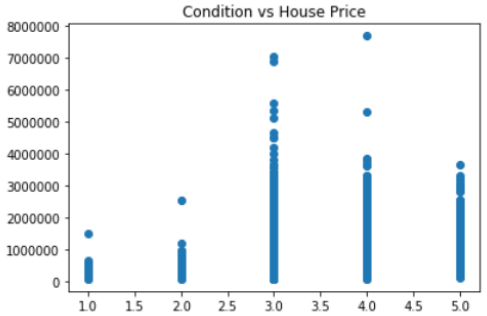
STEP 5 INTERPRETING THE DATA

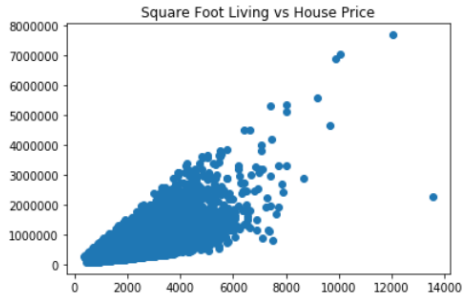
- Explanatory power of regression model
- Influencing power of the individual variables

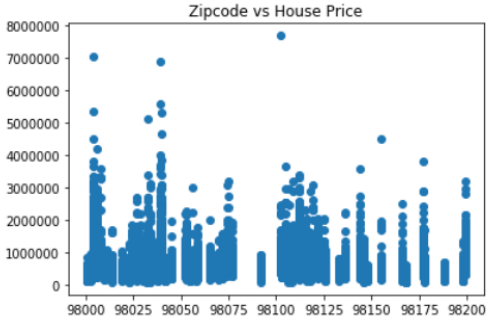
House Characteristics/ Features Relationship with House Sale Price in King Country

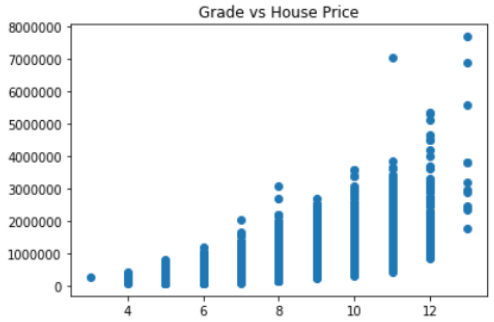


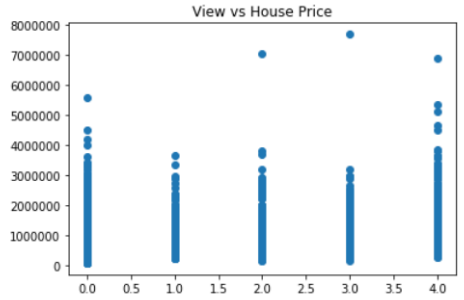


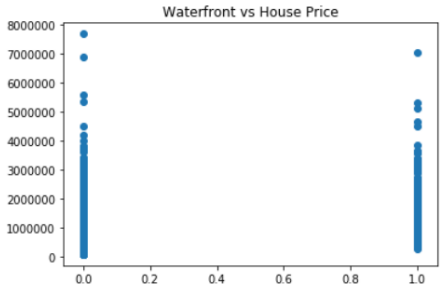


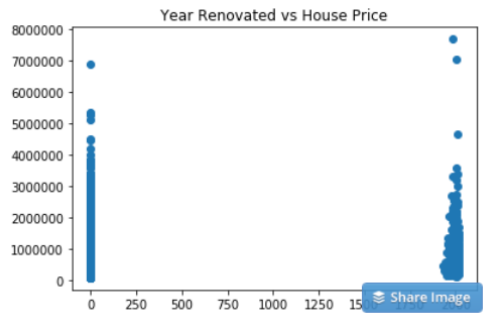


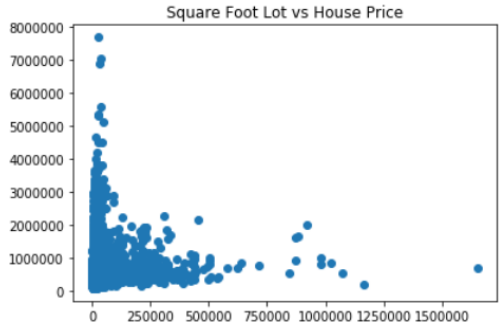


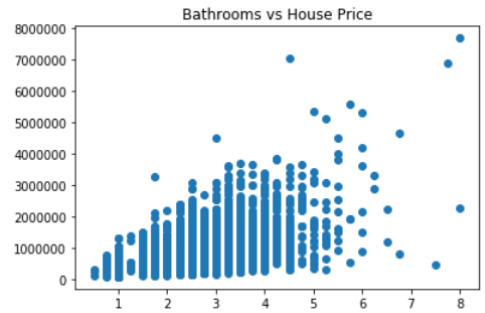


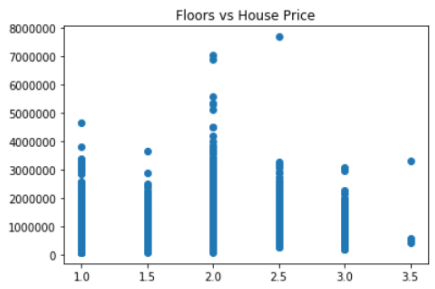


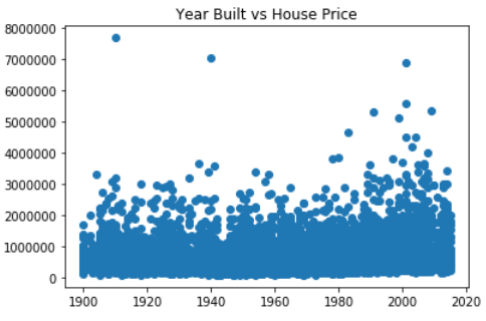












Improvement for Future Research Exploration

Based on the experience/ insight so far from Part 1 (Obtaining the Data) to Part 5 (Interpreting the Data), we can derive lessons learned for improving future research exploration:

1. Find simpler or alternative ways of improving model fitness. The current method use to improve model fitness (i.e. Gradient Boosting) did not seems to improve the current model.
2. Import other new dataset and concatenate with current dataset.
3. Experiment adding or removing current dataset to simulate different scenarios.

Q&A