King County Home Sales

Maximizing your profit through data-driven analysis and predictive modeling

In this project, I assumed the role of a data scientist in a startup company created to help prospective home sellers in King County, Washington maximize their profits when they choose to sell their homes. To assist in creating the models for this analysis, I was provided the King County House Data dataset.

Business Value & Problem Statement

Create a model to accurately predict how much a home will sell for based off its attributes. Doing this effectively will allow us to please our clients and broaden prospective clients by accurately pricing people's homes to sell quickly, and for a good price.

The main attributes focused on for this analysis are:

- Livable Square Footage
- Lot Square Footage
- Number of Bedrooms and Bathrooms
- Overall Home Grade (King County Grading System)
- Zip Code

In addition, this exploration and modeling will allow for us to provide recommendations to clients about how to most effectively prep their house for sale.

As far as this goes for the purpose of our business, by creating the most accurate model, we'll be able to help clients price their homes to maximize their profits which in turn brings in more company revenue. The main factors I'll be examining are Livable sq footage, lot sq footage, the number of bedrooms and bathrooms, the overall home grade according to the King County Grading System, and the Zip Code. A few other attributes are taken into account, such as waterfront access and home condition but I mainly focus on the aforementioned traits.
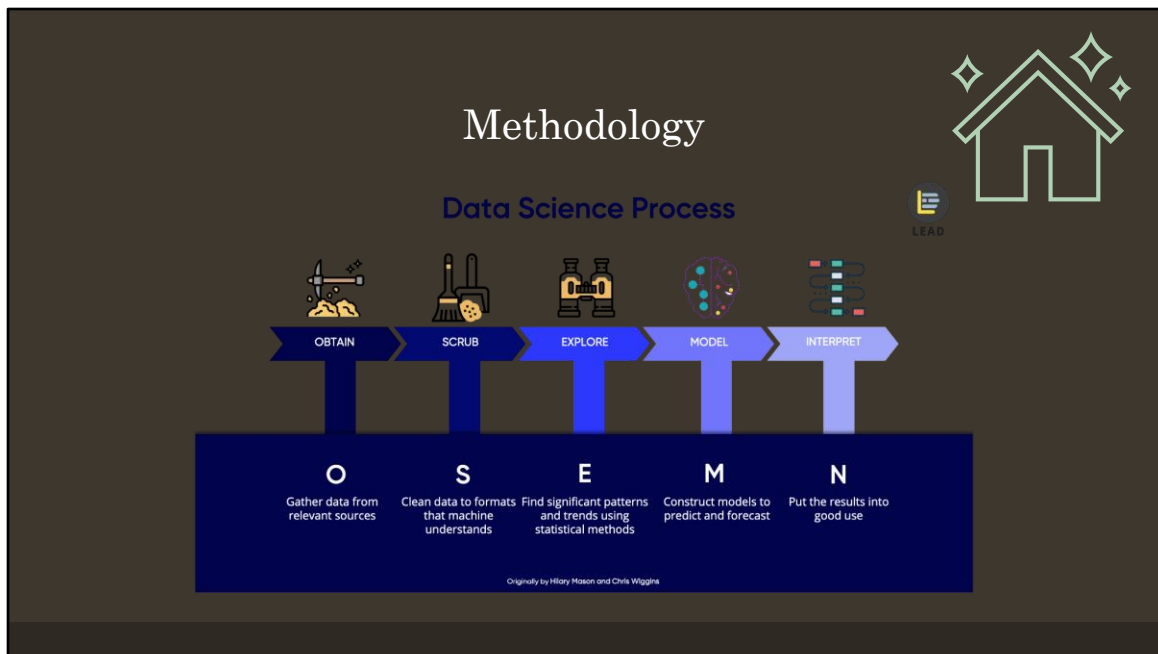
I used the OSEMN model for data science.

The data used for this project was provided in the KC House Data database which details ~20,000 home sales between May 2014 and 2015. Obtaining it was as easy as uploading the file.

Once imported into my notebook, I set out to clean the dataset and ensure I had all the information I needed to move to the next step. I then explored which features would provide us the best insights to home price and allow me to answer the questions I had posed.

Next, I used the cleaned dataset to develop a predictive model to assist in pricing future client's homes.

And finally, I put those models results together and formed a small list of recommendations to maximize profits.

Key Questions:

- What is the single most important variable related to home price?
- Are bedrooms or bathrooms more significant when it comes to affecting the home price?
- Does a bigger house directly correlate to a more expensive house?
- What significance, if any, does zip code have on home prices?
- Does the lot size or house size more closely correlate to a higher home price?

In addition to the predictive model, there are a handful of questions that I posed to the dataset that I desired answers for. (Read Slide)

I believe through these inquiries we could come to a concise conclusion with recommendations for client's seeking to increase their potential profit from selling their homes.

These recommendations will be based off what could be done to a home realistically. For example, it's not possible to make a landlocked home into a waterfront home, even though we know waterfront homes net more profit on average – it's just not conceivable.
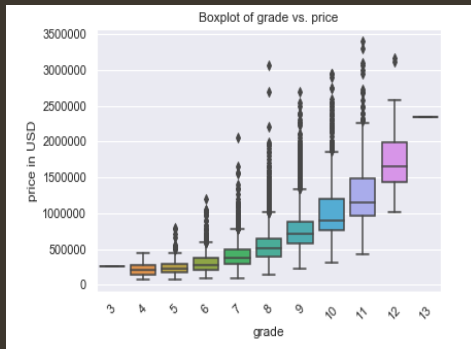
First we'll look at Grade.

The King County Home Grading Scale is a rating system used to classify a homes overall quality.

Overall, this is the most important trait tied to home price, the higher the grade, the higher the sale price. Waterfront property has a higher correlation coefficient as waterfront houses are on average over twice as expensive as their landlocked counterparts. This is important, but for the purpose of our startup, grade seems to be far more significant as we cannot magically teleport houses.
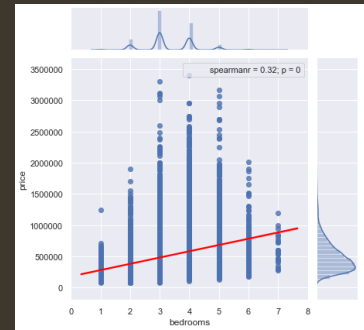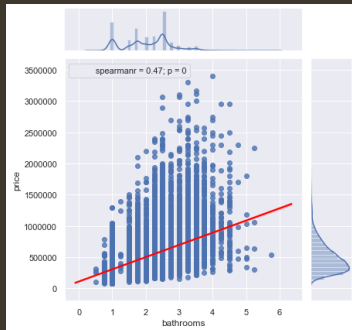
Next, we explore which rooms are more likely to increase your homes value.
Renovating is popular for increasing your homes value, if you decide to renovate, what kind of room should you add?

Bathrooms seem to be the more economic move, as they correlate more strongly than bedrooms to a higher sale price.

Time is money – What time of the year and week to sell?

Sales by month

(bar chart showing sales values from January through December, with peaks around April–May ~2200–2300, lowest in January ~950)

Sales by Days

(bar chart showing sales by day: Monday ~3900, Tuesday ~4500, Wednesday ~4400, Thursday ~3800, Friday ~3600, Saturday ~200, Sunday ~200)

- Spring – Summer months have larger potential buyer pools.

- Avoid selling in the Winter.

- No strong indicator of a favorable day of the week.

- Saturday and Sunday are underrepresented due to the infrequent nature of realtors and lenders working weekends.

When you sell your home can have a varying effect on its price.

If you sell your home in the Spring – Summer months, you have a larger buyer pool which could drive up price.
April through July seem to be the most popular months to sell, but the effect is still felt into October. It may be wise to avoid selling in the winter, as the potential buyer pool is smaller and you may fetch a lower price due to that.

As far as day of the week goes, there isn't a day that's significantly stronger for home sales. I believe the relative infrequency of Saturday and Sunday sales is due to most realtors and lenders not working those days as often.

Zip code – How do your neighbors affect your home?
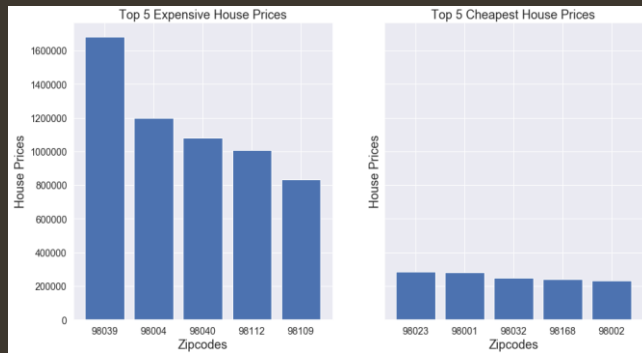
- Large disparity across the range of zip codes.
- Need more information to know exactly why this disparity exists.
- Zip code does not have a strong correlation to price over the county as a whole.

Lastly, a look into the zip code and it's effect on home price. In my regression model, the zip code had a slight negative correlation with the price. That said, exploring the zip codes uncovered a massive disparity in home prices between the top zip codes and bottom. 98039(Medina neighborhood) had a mean home price well over $1.5M while 98002(Auburn neighborhood) had a mean home price of just over $200K.

While there are much more important variable in predicting home price over the whole county, it was interesting looking into the zip codes. These differences seem to be a symptom of things not immediately available in this dataset such as school proximity and quality, crime rate, distance to city-centers, and overall neighborhood history. There is another interesting graphic related to zip code found in the appendix of this presentation.

How to Maximize Profit – Biggest Factors in Home Price

- Grade is the single most important factor in home sales.
- Bathrooms have a strong correlation with price.
- Livable square footage directly relates to a higher sale price.
- Sell in the Spring-Summer.

Grade
↓
Add Bathrooms
↓
Square Footage
↓
Sell in the Summer
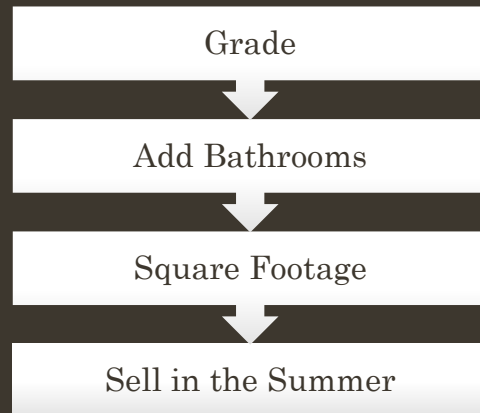
To review -

Most important of all the variables, is Grade. The best thing one could do is to spend the time and money to re-finish, clean, and/or remodel your home so it will receive the highest possible grade under the King County home grading scale. This is the most surefire way to increase your homes value. This will also have a positive effect on your homes condition rating, which is another important metric.

If you do end up renovating, adding bathrooms is a strong way to find increased profits. Bathrooms are correlated with price much more closely than bedrooms and houses with more of them tend to sell for more than those with few.

While renovating, one would be smart to add a room as more living space directly correlates to a higher sale price. Add-on renovations could prove to be a smart investment.

Lastly, homes tend to sell best between April and July. Avoid selling your home in the winter, there just aren't as many potential buyers.

Future work – How to improve the model?

- Source data on school systems.
- Examine pollution data.
- Gather crime rate data.
- Calculate ROI for renovation costs.
- This could improve the model's accuracy and net us more clients and profit-per-client.

In the future, it would be helpful to have some data added to this set to clear up some more answers.

It would be helpful to have information about the school systems in the area, pollution data, and crime rates.

Also as far as advising clients to renovate, it would be nice to be able to quantify that and have a calculator for their potential return on investment for add-ons.

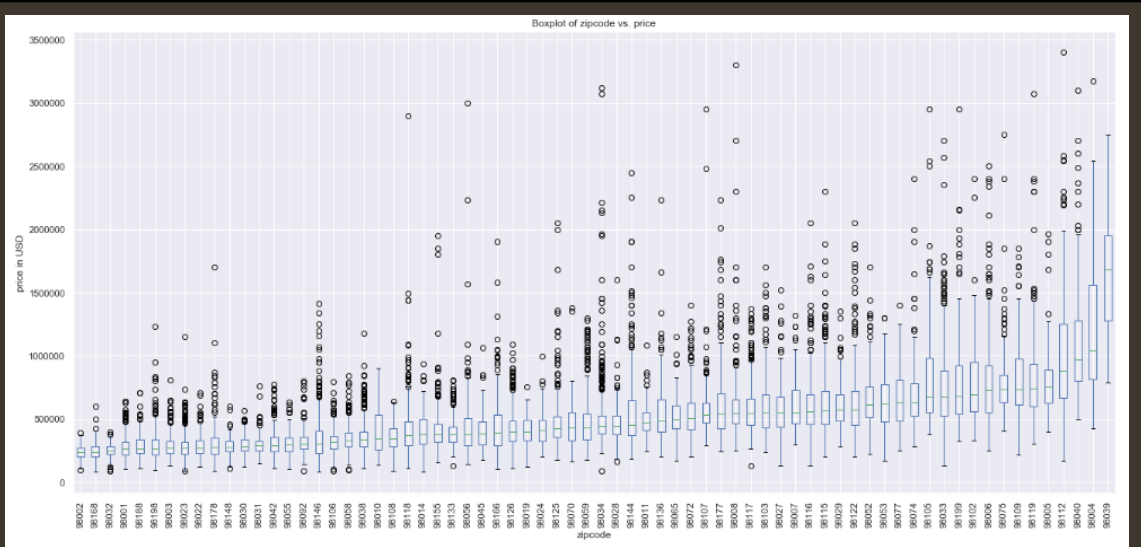Thank you! I'm happy to answer any questions that you may have about my exploration, analysis, and presentation.

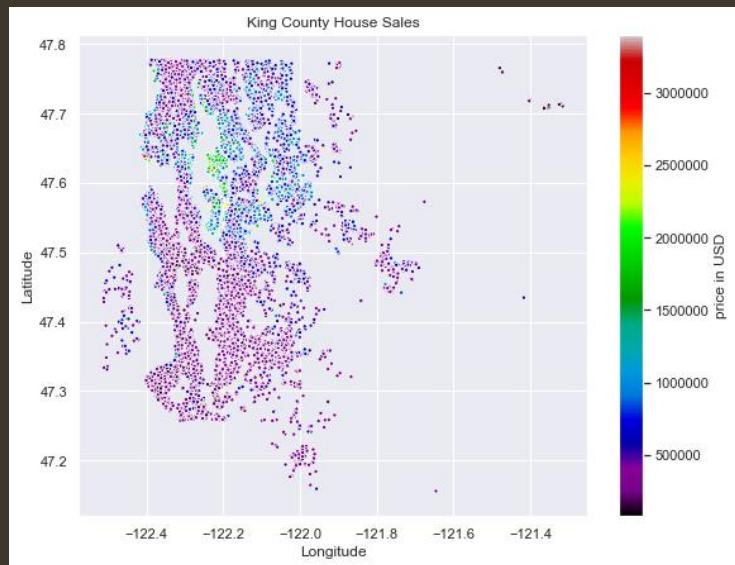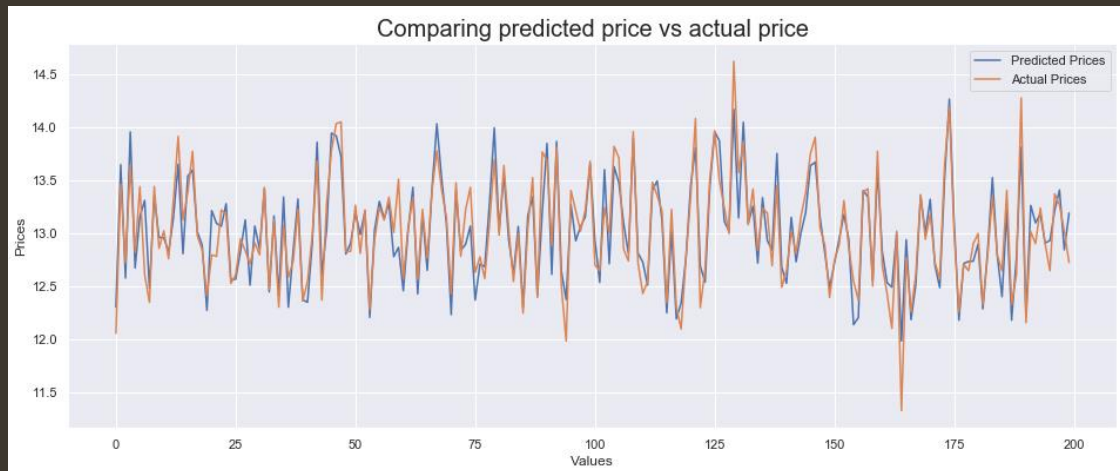Boxplot of zipcode vs. price

Zip code box plot with price

Boxplot of zipcode vs. price
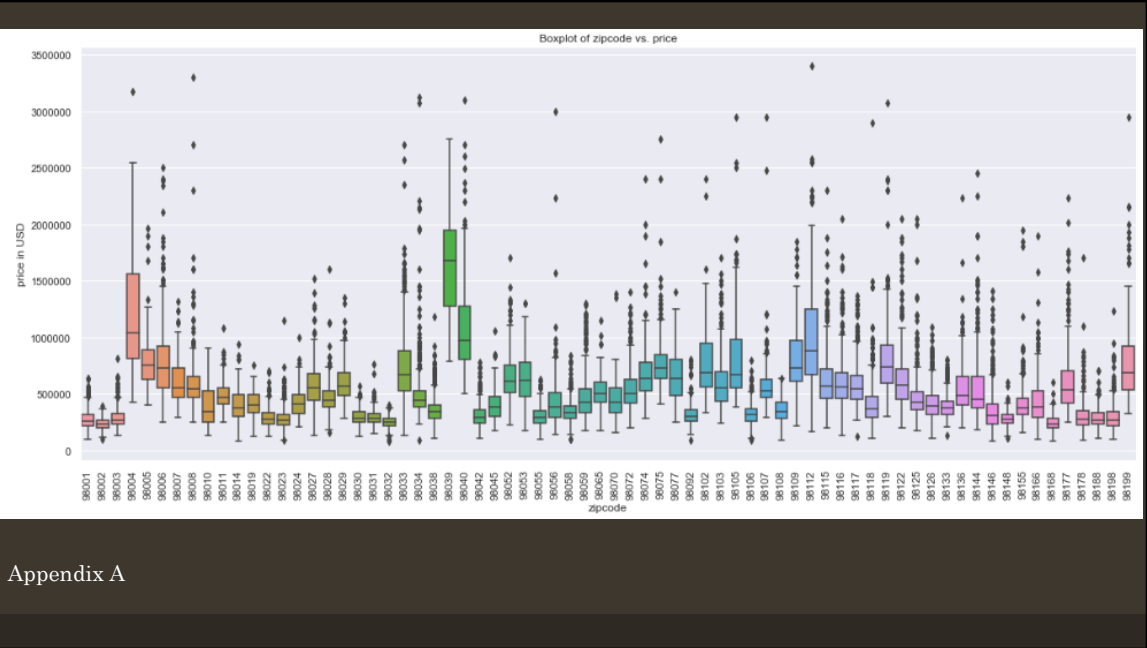
Ordered zip code box plot with price

King County House Sales

Lat & Long data tied to house price. Can see Mercer Island, Belleview, and Medina neighborhoods lit up.

Comparing predicted price vs actual price

Appendix D

Multi linear regression model predicting price

OLS regression results of the multi-linreg test. 89% accurate.