

Drymander / dsc-phase-2-project


forked from [learn-co-curriculum/dsc-phase-2-project](#)

☆ 0 stars  104 forks

☆ Star

 Watch ▾

< > Code

 Pull requests

 Actions

 Projects

 Wiki

 Security

 Insights

 main ▾

...

This branch is 22 commits ahead of learn-co-curriculum:main.

 Pull request  Compare



Drymander image fix ...

3 minutes ago  31

[View code](#)

King County Housing Regression and Analysis

Author: [Johnny Dryman](#)

Overview

For the Phase 2 Project, we will be analyzing housing sales data for King County (Seattle, WA area). We will be using multivariate linear regression to explore which features of the data have the greatest influence on price.

Business Problem

As home values continue to sky rocket in the pandemic era, many King County residents have inquired about how to increase the value of their homes. Fortunately, we have access to all homes sold in King County for roughly one year, from May 2014 - May 2015.

This data gives us access to a variety of important metrics both quantitative and qualitative.

After scrubbing the data and assuring quality, we will use multivariate linear regression to analyze our features and determine their relationship with sale price.

Finally, we will formulate our observations into useful recommendations to any resident interested in increasing their home value.

Data & Methodology

After reviewing the data and running preliminary models, we have opted to use the following metrics from the dataset:

From the King County data, we will be using the following features:

- Sale Price
- Number of Floors
- Living Area Square Footage
- Lot Square Footage
- Waterfront (Y/N)
- Condition
- Year Built
- Zip Code

We created two additional features to simplify whether or not a house has a basement or has been renovated:

- Basement (Y/N)
- Renovated (Y/N)

The following features were removed due to multicollinearity or insignificance based on P-values:

- Sale Date
- Number of Bedrooms
- Number of Bathrooms
- View
- Grade
- Basement Square Footage

- Non-Basement Square Footage
- Year Renovated
- Latitude
- Longitude
- Living Square Footage – 15 Nearest Neighbors
- Lot Square Footage – 15 Nearest Neighbors

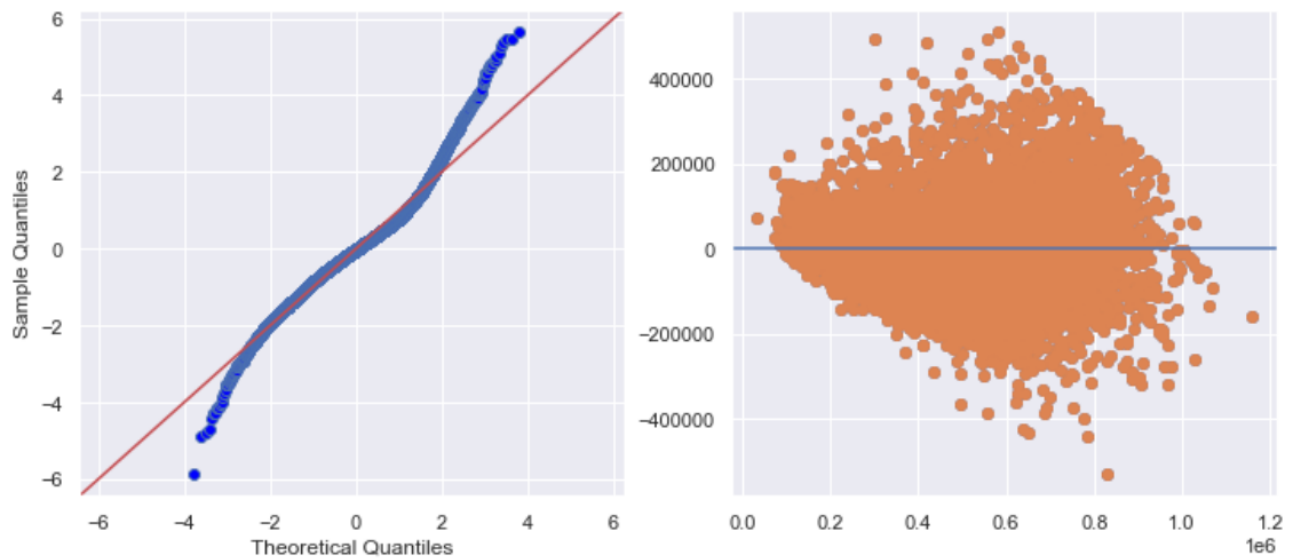
Model

Our final model has an adjusted R-squared of 0.798, meaning our model has a predictive quality of roughly 79.8%.

OLS Regression Results

| | | | |
|--------------------------|------------------|----------------------------|-------------|
| Dep. Variable: | price | R-squared: | 0.799 |
| Model: | OLS | Adj. R-squared: | 0.798 |
| Method: | Least Squares | F-statistic: | 617.7 |
| Date: | Wed, 21 Apr 2021 | Prob (F-statistic): | 0.00 |
| Time: | 16:52:19 | Log-Likelihood: | -1.6841e+05 |
| No. Observations: | 13128 | AIC: | 3.370e+05 |
| Df Residuals: | 13043 | BIC: | 3.376e+05 |
| Df Model: | 84 | | |
| Covariance Type: | nonrobust | | |

In our QQ plot, we can see that our distribution is mostly normal, but it might be worth removing more outliers in the future. Our homoscedasticity shape shows is mostly cone like.



Interpretation of Coefficients

Based on the coefficients produced by our model, we can estimate the impact that certain qualities and metrics will have on homes in King County.

Depending on the location, zip codes can have the most dramatic impact on price. The most valuable zip codes are those closest to the metropolitan city center (Seattle, Bellevue, and Mercer Island). The impact on price in the top 5 zip codes is an estimated \$473-628k.

Other than the least valuable zip code, our model functions in a way that doesn't subtract estimated value from homes. The bottom 5 zip codes are located in Kent, near the southern end of King County. While not the furthest from the city center, they are significantly further than our most valuable zip codes.

Waterfront is the most impactful, adding \$338k to price.

Condition lines up with our expectations. The greater the condition, the more valuable the home. Improving the condition from 1 to 5 would add an estimated \$174,135 to a homeowner's value.

Renovated homes seem to fetch a larger price of approximately \$42,260, which aligns with expectations.

Floors is a bit counterintuitive. While 2 floors seems to increase the value by \$11.5k, a third floor decreases value by \$22.5k, 3.5 floors decreases by \$40.5k. Considering the cost of adding an additional floor would likely be much more expensive than these coefficients, this might indicate that expanding the square footage of a home within floors that already exist might be a more sensible investment.

Sqft_living gives us an estimated value of \$159 for every additional square foot of space.

On the surface, sqft_lot looks like it has a relatively lower impact on price. However, it is still relevant when comparing properties with significant differences in size. One acre is 43,560 square feet. Our model predicts that with a \$3.45 impact to price for every square foot, an additional acre would add \$150,282 to the value of two otherwise identical properties.

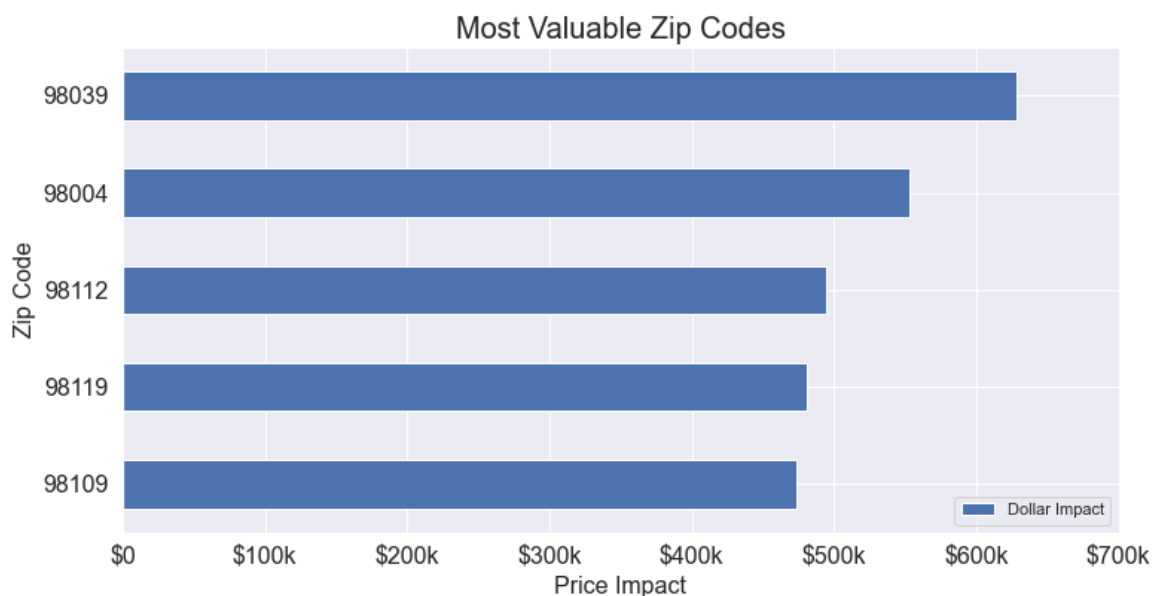
Age doesn't seem to have a great impact. Despite having a P-value greater than 0.05, a house will lose \$168 in value every year. Even in the case of our oldest houses, age can only have a maximum price impact of \$19,425.

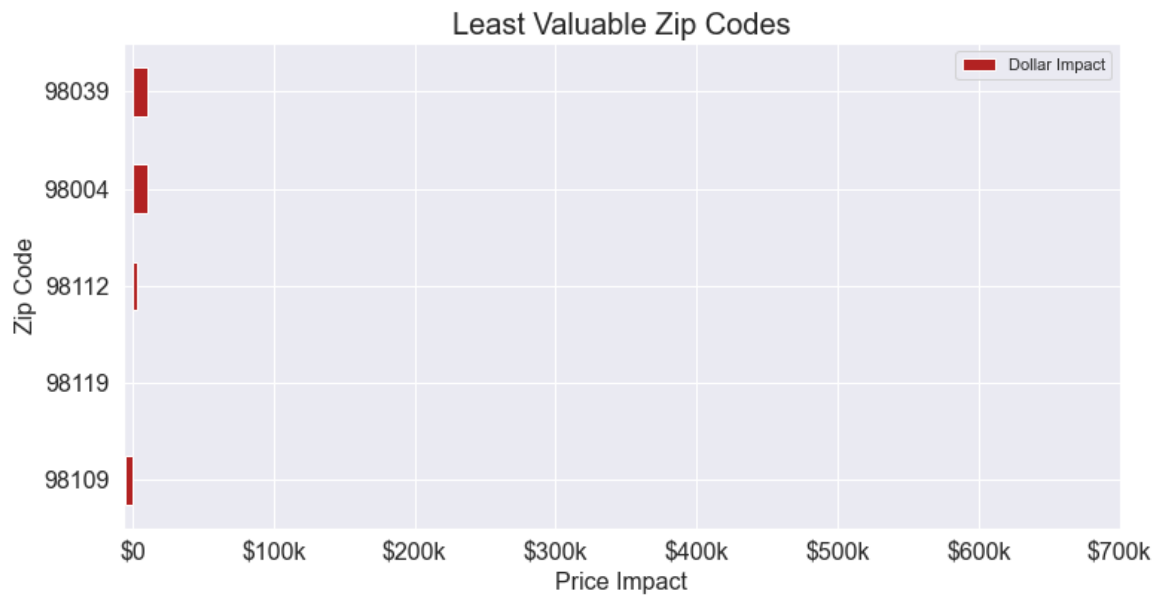
Perhaps counterintuitively, the presence of a basement decreases the value of a home by \$23,634. This might require further examination.

The vast majority of movies recoup their investment at the box office, but not all are successful.

Insights

- Location is the most prized quality of a property. Certain zip codes are highly sought after. The top 5 most valuable zip codes will influence property value by an average of \$473k-\$628k. These zip codes are generally closer to the metropolitan area. Homes located further from the city to the south are less valuable.





- Similar to location, waterfront properties are also much more more valuable and add an average \ \$337k to property value.
- One might assume that additional bedrooms and bathrooms are more valuable. However, according to our model, what actually drives value is total living area square footage. Understanding this, we can intuitively assume that with additional square footage comes additional bedrooms and bathrooms (on average), but our model does not explicitly model this relationship.

☰ README.md

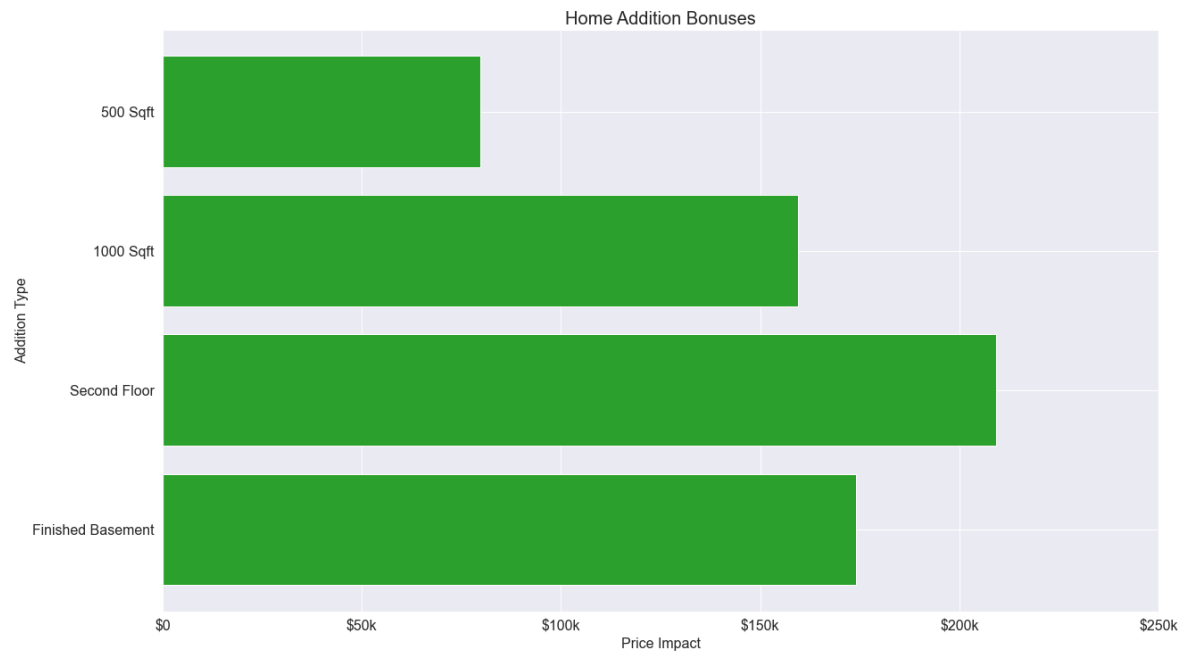


assumed that King County's 'Grade' system might behave similarly, but our model determine that the grade system was not a driver of price.

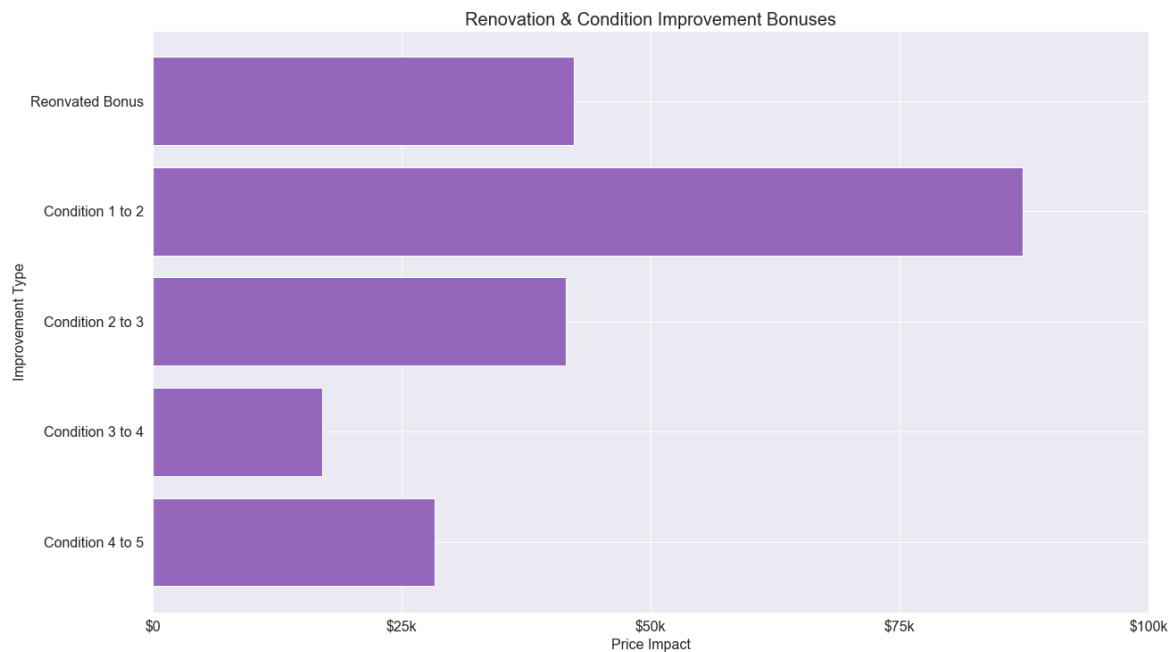
Recommendations to Homeowners

Many of the insights generated by analyzing our model did not lead to practical recommendations for homeowners. It isn't exactly practical or possible in most cases to uproot a home and move it to a new area or by the water. But we did notice two key ways that an owner can improve their value:

- Adding square footage through home construction is the most practical recommendation we can offer to improve value. Each additional square foot of living space adds an estimated \ \$159.48 in home value. Adding a second floor gives a small bonus and adding a basement gives a small penalty. However, when factoring in the added square footage of projects like these, the penalties will most likely be absorbed by the added value.



- Renovating also gives a noticeable bump to price, especially if that renovation improves the condition. Homeowners should maintain the condition of their home, or it will decrease in value.



Further Analysis and Modeling

The goal of this project was to develop a very general understanding of the most influential factors in property value. Given more time for data review, we might be able to implement the 'view' feature if we can get a better understanding of what it represents. Sqft_living15, sqft_lot15, and Year Renovated might be interesting to explore. Lat and long can be used to heatmap our dataset to visualize home values on a map of King County.

We could implement standardization and normalization to improve our model's predictive quality. We would also like to implement a train / test split for similar purposes.

It might be helpful to build dynamic splitting of our data. For example, how specifically could the owner of a 2 story, 4 bedroom house in Bellevue improve their home value? Would the coefficients of our features change if we ran our model using only houses that matched that criteria? Dynamic splitting could be useful for generating tailored recommendations to clients who might be willing to pay a premium for such services.

For More Information

See the full analysis in the [Jupyter Notebook](#) or review this [presentation](./Johnny Dryman - Phase 2 Project Presentation.pdf).

For additional info, contact Johnny Dryman at johnnydryman@gmail.com

Repository Structure

```
|— data
|— images
|— README.md
|— Johnny Dryman - Phase 2 Project Presentation.pdf
|— Housing Project Final.ipynb
```

Releases

No releases published

[Create a new release](#)

Packages

No packages published

[Publish your first package](#)

Languages

● Jupyter Notebook 100.0%