King County House Sales Linear Regression Analysis

Spencer Hadel

Summary

Analysis of King County house sales from the years 2014-2015 has yielded profitable insights on the appropriate pricing of property for a new real estate company in the area.

Primary findings indicate that house **square footage**, **view**, and **grade** have high impacts on price, as well as whether a property is **on a waterfront.**

Outline

- Business Problem
- Data
- Methods
- Results
- Conclusions

Business Problem

A new real estate company in King County would like to properly assess prices of houses in the area using past sales data, in order to create a more data-driven approach to house pricing.

In order to help, we have analyzed past house sales data in the region and create a linear regression model which can help the company better understand what factors contribute to price of a given home.

Data

This analysis utilizes kc_house_data.csv, which contains over 21,000 entries of past sales data from 2014-2015 in the King County area.

This data was cleaned in order to remove null values and outliers that could harm future modeling.

Features such as id, zip code, latitude, and longitude were removed before the final modeling process.

Methods

For this analysis, the data present in kc_house_data was cleaned to remove null values, outliers, and oddities. It then explores the relationship between each of the features on house price, creating visualizations to better understand the effect of these features.

Next, the data was preprocessed by normalizing continuous features (such as square footage, number of bed and bathrooms, and number of floors), in order to properly compare them on different scales.

Categorical features (such as view, condition, grade, and waterfront) were then split into dummy variables in order to successfully train the model to them.

Methods

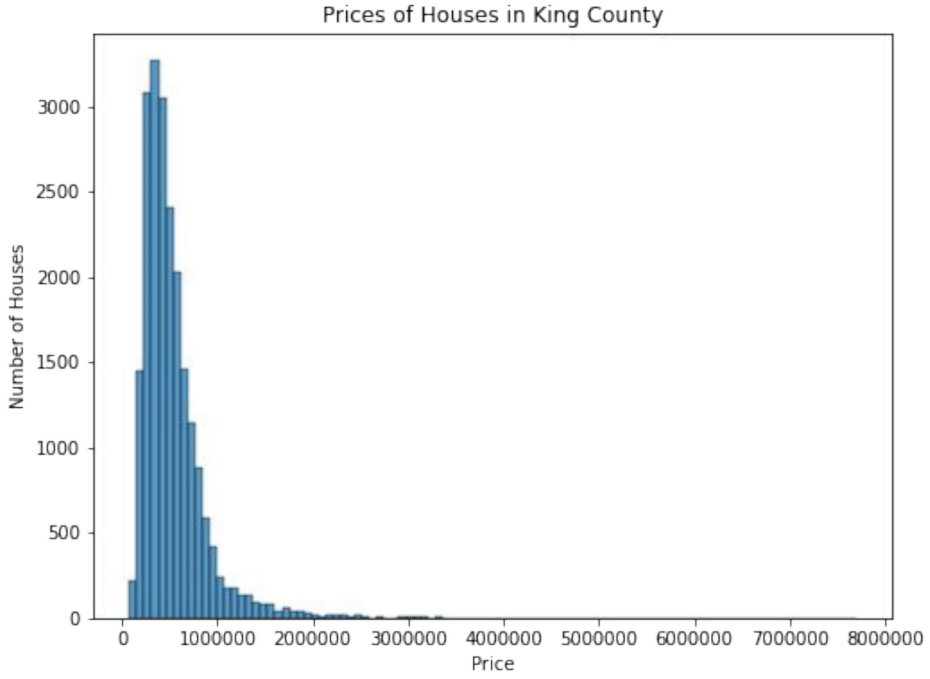
Finally, the analysis creates multiple Linear Regression models in order to find the best fitting model.

25% of the data was split into a testing dataset, while the remaining 75% was used as a training dataset meant to accurately predict prices in the test set.

The data was tested for uninfluential features and multicollinearity in each step of the modeling phase, in order to find the lowest possible Root Mean Squared Error when applied to the test data.

Results

The analysis revealed a few key points about house sale prices in the area. A distribution plot shows the most common prices in King County.

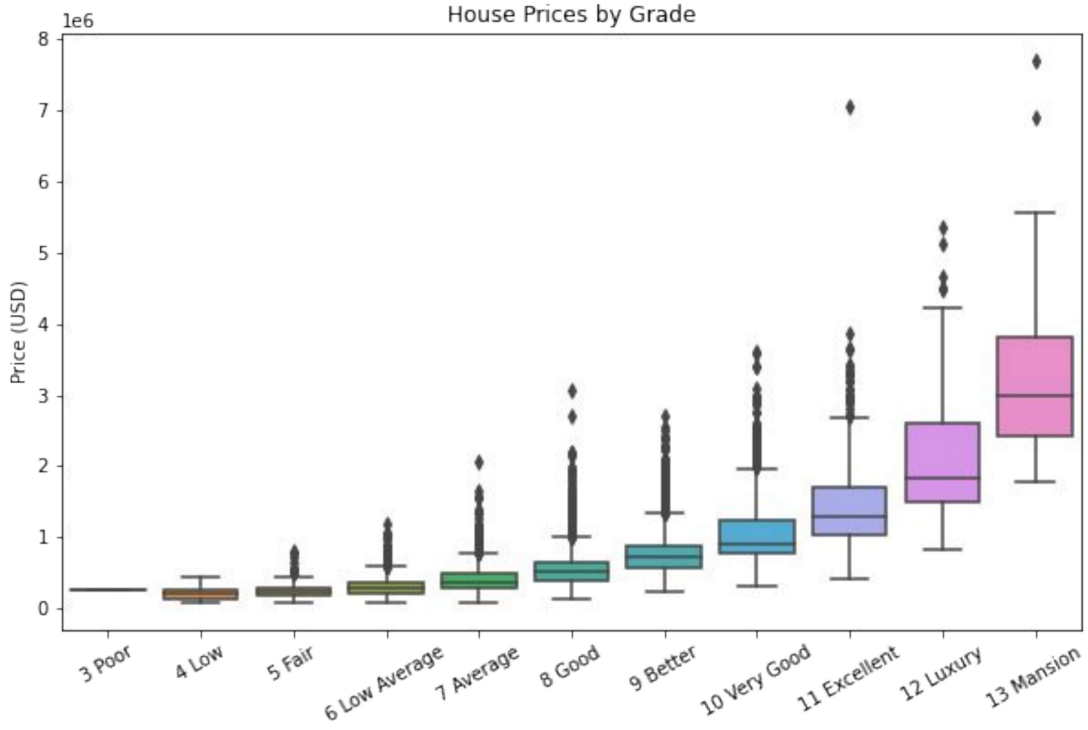


Results

Additional observations can be made about other features' impacts on property value.

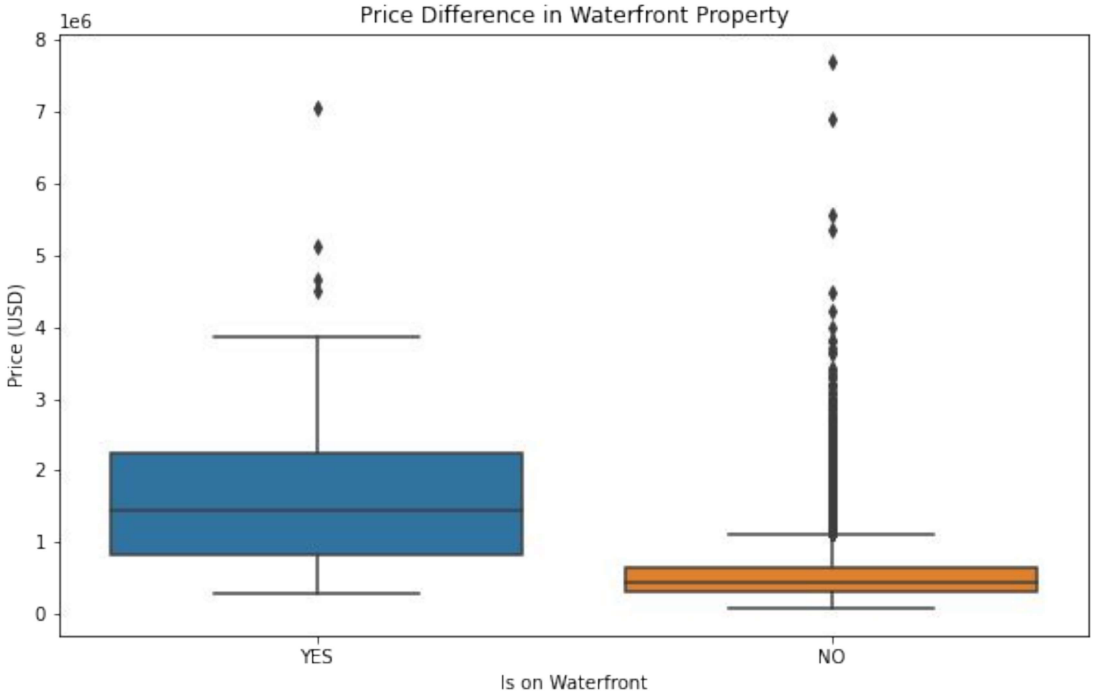
Results

Additional observations can be made about other features' impacts on property value.



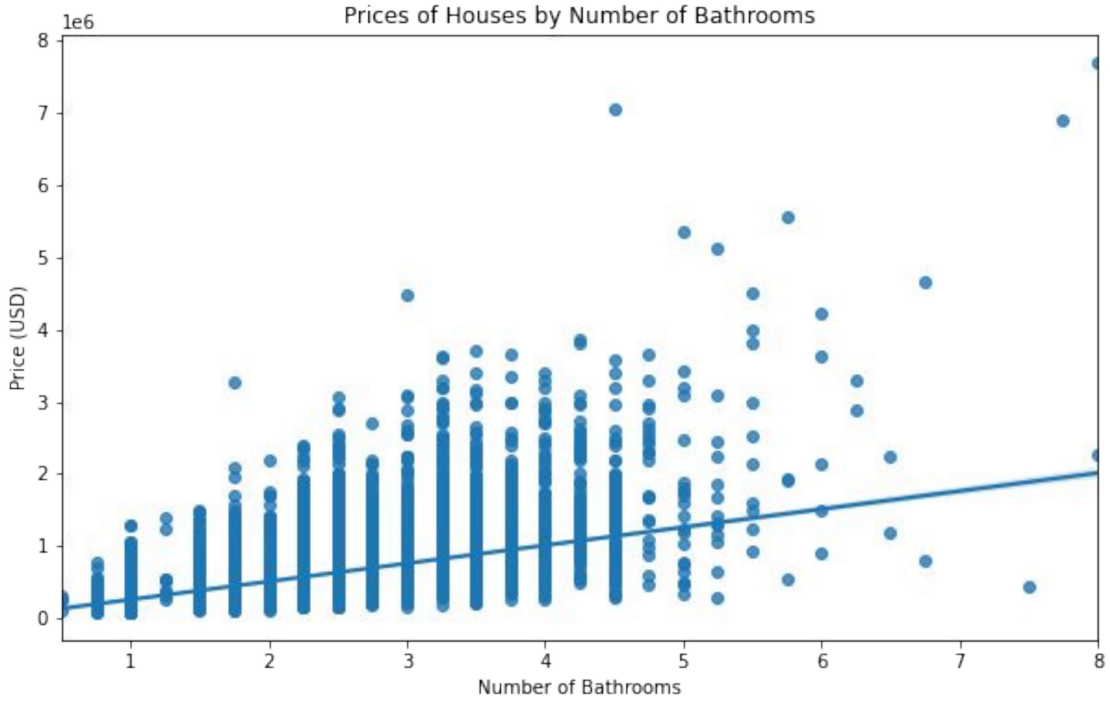
Results

Additional observations can be made about other features' impacts on property value.



Results

Additional observations can be made about other features' impacts on property value.



Results

A linear regression model was created that asserts a base house price of almost $1,000,000, and adds or subtracts value based on the coefficient values for each significant feature.

This model has an R-Squared value of 0.66, meaning it maintains approximately 66% accuracy.

```
grade_13_Mansion      1853454.022
grade_12_Luxury        905721.578
waterfront             571848.741
grade_11_Excellent     364974.794
view_EXCELLENT         233262.921
renovated_2000         107344.256
sqft_living            101414.772
view_FAIR               69429.326
condition_Very_Good     63734.593
view_GOOD               40540.435
bathrooms               22846.804
condition_Good          21372.995
has_basement            18945.643
floors                  11790.291
bedrooms               -16262.693
sqft_lot               -26530.229
view_NONE              -69349.403
yr_built               -84733.121
grade_9_Better        -231280.835
grade_8_Good          -410406.099
grade_7_Average       -508177.626
grade_4_Low           -550193.382
grade_6_Low_Average   -567311.303
grade_5_Fair          -612049.528
Name: Coefficients, dtype: float64
Intercept: 976799.8626844347
```

Conclusions

The analysis concludes that some of the most relevant features to a property's value are it's **grade** (as assigned by the King County assessor website), whether or not it is on a **waterfront**, what it's **square footage** is, and the number of **bathrooms**. All of these assessments make sense, though, and should not be used as the sole predict house price for prospective buyers and sellers.

The model created by this analysis can be used as a baseline, but further investigation is needed.

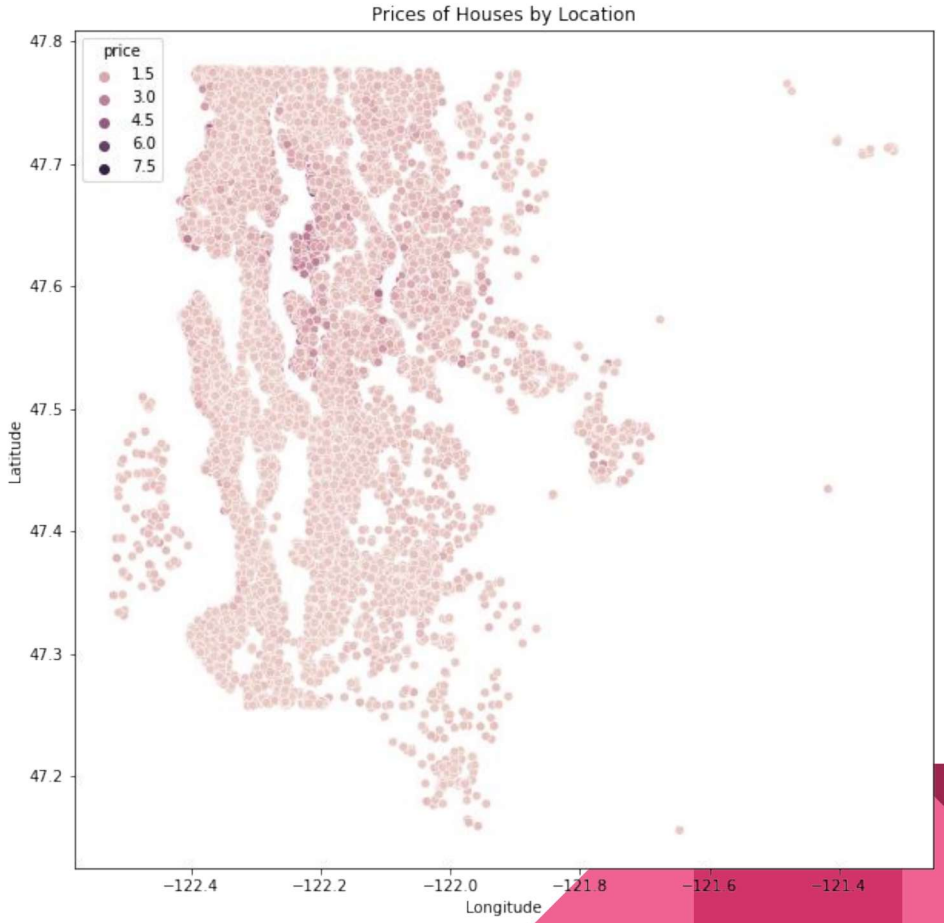
Next Steps

Going forward, the ideal next step to take would be to analyze King County house sales explicitly by location. Further level of detail into what specific renovations each house has attained recently, as well as what features of a house specifically yield different grades, condition scores, and view scores.

The analysis yielded another result of note, that was not included in the modeling process.

Next Steps

Early phases of the analysis show some correlation of price vs location, which deserves further investigation along with data on the King County area, such as highway locations, points of interest, and the economic status of neighborhoods.



Thank You!

Email: shadel96@gmail.com
GitHub: @shadel96
LinkedIn: linkedin.com/in/spencer-hadel/