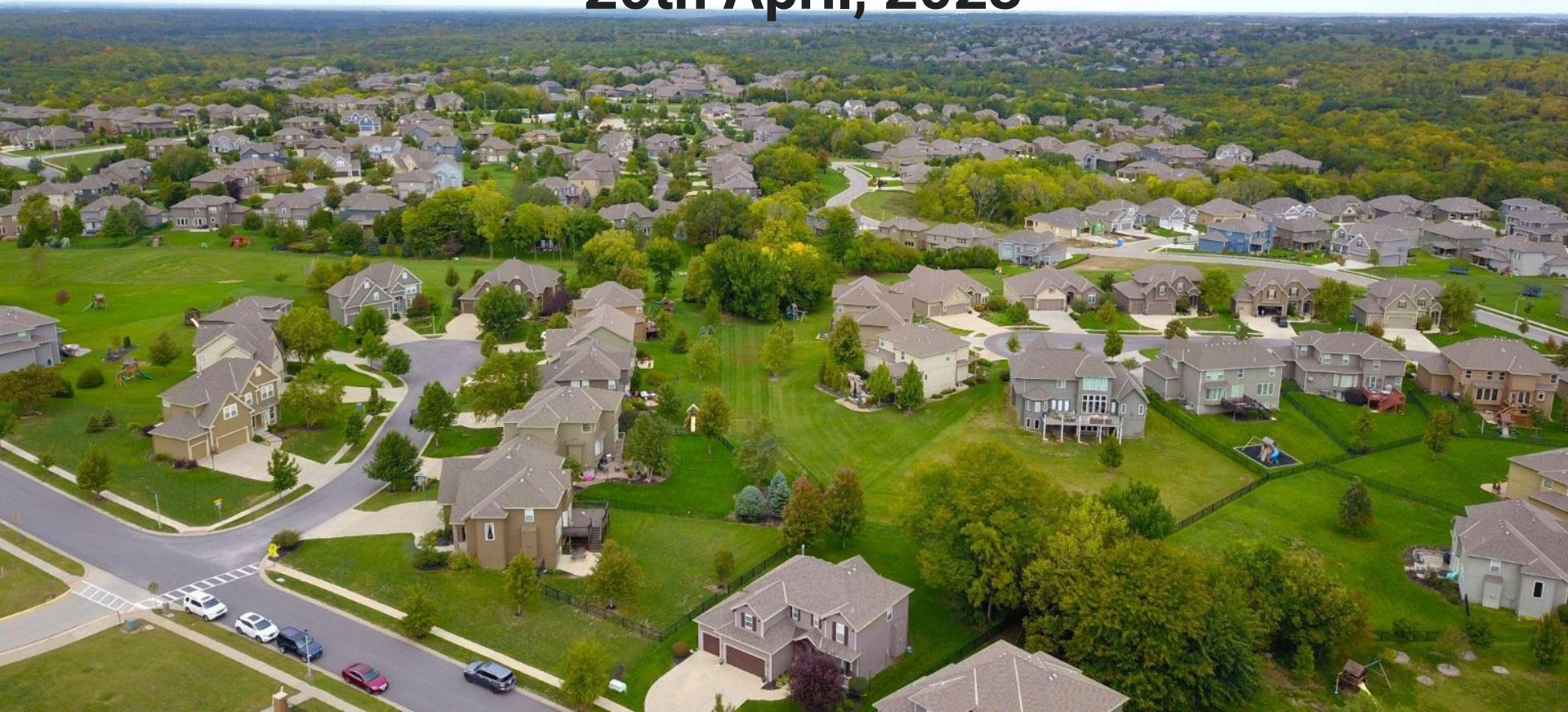
Predicting Home Sale Prices in King County

20th April, 2023



Team Members

- Daniel Ekale
- William Onsare
- Abdikarim Gedi
- Susan Warigia
- Eston Kamau
- Edwin Nderitu

Background

The data used in this project is the King County House Sales dataset, which can be found in kc_house_data.csv in the data folder. The dataset contains 21,597 records and 21 columns. The description of the column names can be found in column_names.md in the same folder. It contains 21 columns with each being either numerical or categorical data.

Business problem

Our stakeholder is homeowners who are looking to renovate their homes and want to estimate the impact of these renovations on the value of their home.

Our business problem is to identify which home features are most important in determining a home's sale price and estimate how much value can be added by improving these features.

Main Objective

- Identify the most significant home features that contribute to a home's sale price.

Specific Objectives

- Estimate the value added to a home by improving these significant features.
- Develop a methodology for accurately estimating the impact of home renovations on home value.
- Provide homeowners with an easy-to-use tool for estimating the value added by renovating specific home features.
- Provide actionable recommendations to homeowners looking to renovate their homes to maximize their home value.

I. Overview of Data

The analysis dataset consists of Price of Houses in King County from sales between 2014 and 2015. Along, with house price it consists of information on 18 house features, Date of Sale and ID of sale. The list below describes the interpretation of the variables in the dataset.

II. Data Preprocessing

- A majority of the fields found in the King County housing dataset were deemed acceptable for performing our statistical analyses. However, while traversing the data we found that some of the columns needed to have their data types adjusted to meet our needs. For example, we converted a square foot basement area from string to float.
- We also identified some null values which were replaced with zeros where necessary.
- We also get rid some outliers using the Z-Score standardization method.
- We converted categorical columns into numeric for easy statistical analysis.

Summary of other data inconsistencies

While exploring the data we found a few instances where the data between variables was inconsistent and didn't make logical sense. We chose to either make the values consistent by recoding or exclude those instances from the data.

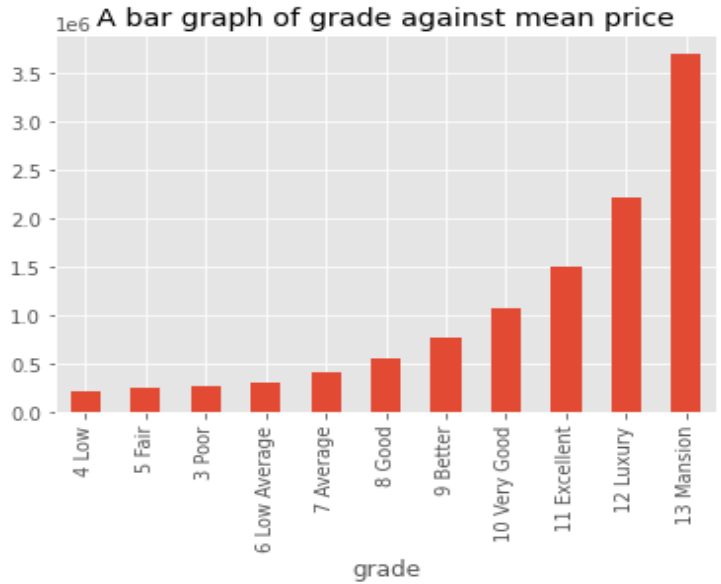
1. One observation with 33 bedrooms in 1620 Square feet with 1.75 bathrooms. That value was treated as an outlier and we removed them.

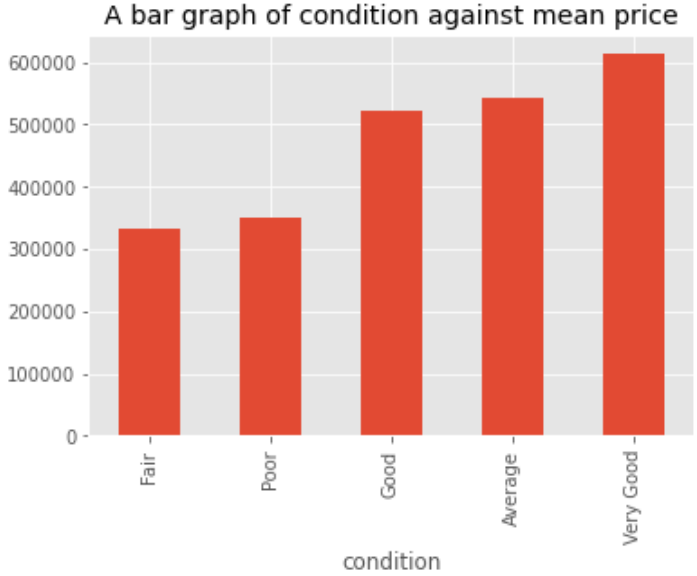
1. Ten observations with 0 bathrooms. Since it is not conventional to have houses without bathrooms, we decided to exclude these observations.

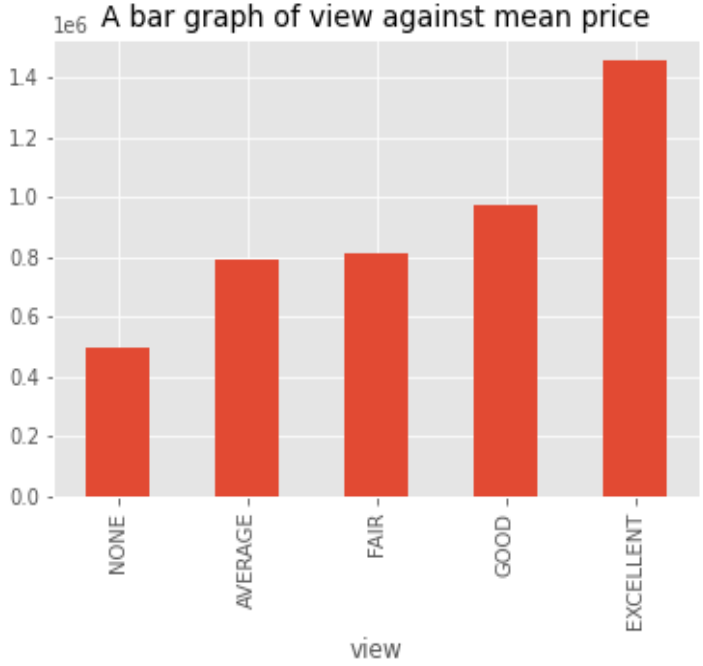
III. Data Visualization and Pattern Discovery

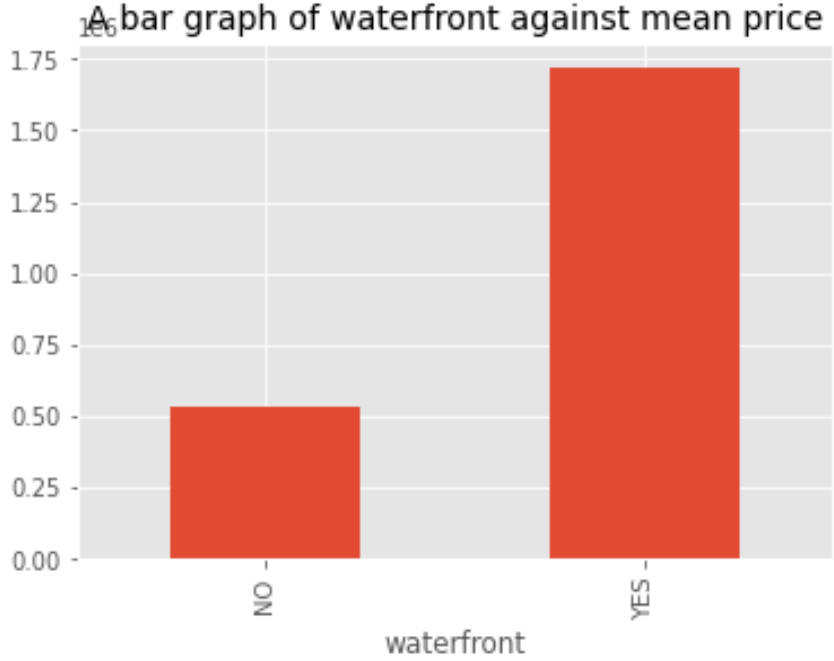
The objective of data visualization and pattern discovery was to reveal relationships between the house features and the response variable, price. We wanted to identify house features that affect price variable and could be potential predictors. Through visualization, we gathered the following information about the data.

1. Price increases with increase in Square Feet Living, Square Feet above, Number of Bathrooms and Number of Bedrooms.
2. Price increase with increase in Grade and Condition.
3. Price increases as the view gets better.
4. Houses with waterfront are associated with high price compared to houses without waterfront.


A bar graph of grade against mean price
1e6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
4 Low
5 Fair
3 Poor
6 Low Average
7 Average
8 Good
9 Better
10 Very Good
11 Excellent
12 Luxury
13 Mansion
grade


A bar graph of condition against mean price
600000
500000
400000
300000
200000
100000
0
Fair
Poor
Good
Average
Very Good
condition


A bar graph of view against mean price
1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
NONE
AVERAGE
FAIR
GOOD
EXCELLENT
view


A bar graph of waterfront against mean price
1e6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
NO
YES
waterfront

IV. Model Implementation

To be able to predict the house sales price, our target variable we implemented a few models.

- Simple linear regression model.
  In this model we regressed the variable square foot living against the target variable price. We discovered that 49.2% of the variation in house prices can be explained by their square footage of living space.

- **Simple linear regression model**.
  In this model we regressed the variable square foot living against the target variable price. We discovered that 49.2% of the variation in house prices can be explained by their square footage of living space.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.492
Model:	OLS	Adj. R-squared:	0.492
Method:	Least Squares	F-statistic:	2.073e+04
Date:	Thu, 20 Apr 2023	Prob (F-statistic):	0.00
Time:	10:15:13	Log-Likelihood:	-2.9763e+05
No. Observations:	21420	AIC:	5.953e+05
Df Residuals:	21418	BIC:	5.953e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4.255e+04	4436.470	-9.591	0.000	-5.12e+04	-3.39e+04
sqft_living	280.5436	1.949	143.972	0.000	276.724	284.363

Omnibus:	14710.422	Durbin-Watson:	1.989
Prob(Omnibus):	0.000	Jarque-Bera (JB):	541541.173
Skew:	2.827	Prob(JB):	0.00
Kurtosis:	26.975	Cond. No.	5.64e+03

- **Multiple linear regression**

Since the linear model was not a good fit for a prediction we refined the previous linear regression model by adding more predictor variables. The value of R-Squared increased from 49.2% to 65.1% which means our model improved significantly and therefore we can use it to predict the sales price of the houses.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.651
Model:	OLS	Adj. R-squared:	0.650
Method:	Least Squares	F-statistic:	2344.
Date:	Thu, 20 Apr 2023	Prob (F-statistic):	0.00
Time:	09:54:56	Log-Likelihood:	-2.9362e+05
No. Observations:	21420	AIC:	5.873e+05
Df Residuals:	21402	BIC:	5.874e+05
Df Model:	17		
Covariance Type:	nonrobust		

Interpretation of Model formula:

The following can be inferred from the model formula:

- Every unit increase in Square feet living, and Bathrooms will increase the predicted price. Predicted price decreases with unit increase in Bedrooms.
- Predicted price is high for houses with waterfront.
- Predicted price is higher for houses with good condition and excellent grades.
- Price increases as the view gets better.

V. Plan for future upgrades

- During exploratory data analysis, we inferred that the outcome variable has a lot of legitimate outliers. We believe that very high prices of houses are because of characteristics partially captured in the data. We would recommend identifying characteristics (eg: amenities, neighborhood) that undoubtedly make a house a High priced property through appropriate domain research. Based on a definition derived from these characteristics we would segment houses into luxury homes and ordinary homes. Further, we would recommend developing different models for luxury homes and ordinary homes.

- There exists evidence from the research in real estate industry that the price of the houses has relationship with other elements such as availability of credit (mortgage interest rates), consumer sentiment and other economic factors. We recommend collecting data on these elements and model should upgrade to factor the effect of these elements.

conclusion

In conclusion, we have performed an analysis of a dataset on housing prices in King County, and made several recommendations to our agency based on our findings. We found that property size, number of bathrooms and bedrooms, condition and grade, view, and location all have significant correlations with housing prices. Home owners can increase the value of their property by increasing the square footage, adding an extra bathroom or bedroom, upgrading the condition or grade of their property, and considering the location and view. These recommendations can help home owners to maximize the value of their property when putting it on the market. However, it is important to keep in mind that these recommendations may not apply universally and may depend on various factors such as the local real estate market and the specific characteristics of the property.

Recommendations

Based on the analysis we performed, we can make the following recommendations to our agency:

- Property size matters: Home owners should consider increasing the square footage of the houses before putting it on the market.
- Bathrooms and bedrooms add value: Bathrooms and bedrooms have moderate positive correlations with the price. Therefore, home owners should consider adding an extra bathroom or bedroom to increase the value of their property.
- Condition and grade matter: The condition and grade of a property have a strong negative correlation with the price. Therefore, a homeowner consider upgrading the condition or grade of their property to increase the value of the house.
- View can affect the price: The view has a strong negative correlation with the price, which suggests that properties with a better view may be priced lower. However, this may also depend on other factors such as the size and condition of the property.
- Location matters: The latitude and longitude (lat and long) have moderate positive and negative correlations with the price, respectively. a homeowner should consider the location of the property as it may affect its value. In some countries, houses that may fall within the tropics may attract more buyers that those on the furthest end of the tropics.

THANK YOU