Predicting Housing Prices For King County

Team: Vikrant Deshpande, Tanvi Kolhatkar, Saishree Godbole

Statement of Goals

Predicting the house prices of any housing development is an important factor in driving real estate efficiency especially in highly populated cities like New York or Seattle. The ever-fluctuating prices got us interested in what features would play the biggest part in predicting house prices.

To study this, we are exploring a dataset of home selling-prices in King County, Washington. In 2015, King County was the most populous county in Washington, and the 13th-most populous county in the US.

In the past, there have been attempts at building algorithms to accurately predict housing prices. In our project, we build a multiple regression model with the house price variable as the target variable and a selected few features as the explanatory variables. The questions we aim to answer are:

1. Identify the features affecting the house price
2. Investigate if location has a notable effect on these house prices
3. Identify if the home-price can be explained by a linear relationship with the chosen features or is a more complex model required

Data Description

Our dataset contains house sale prices for King County, Washington between May ‘14 and May ‘15 taken from Kaggle. It contains the prices for 21,613 homes, the date and ID of sale and 18 descriptive features. Table1 in the appendix describes all the variables in our dataset, but here are a few main highlights:

Variable Name	Description
price	Price of each home sold
bedrooms	Number of bedrooms
bathrooms	Number of bathrooms, 0.5 accounts for a room with a toilet but no shower
sqft_living	Square footage of the apartment’s interior living space
lat	Latitude
grade	1 to 13 scale; 1 – 3 subpar building construction/design, 7 has average construction/design, and 11 - 13 have high-quality construction/design.

Data Cleaning and Transformations

We performed some basic data cleaning and transformation techniques on the dataset:

- Removed records with 33 bedrooms in a living area of 1620 sqft
- Removed records having no bedrooms and bathrooms
- Excluded zipcodes as they did not align with the latitude and longitudes (erroneous data)
- Performed transformations on variables with left skewed distributions as below:

    - The house sizes (sqft_living) here range from 370 to 13,540 sq. feet, with a median of 1910 and mean of 2080.
    - House prices vary from $78K to $7,700K with a median of $450K and mean of $540K.

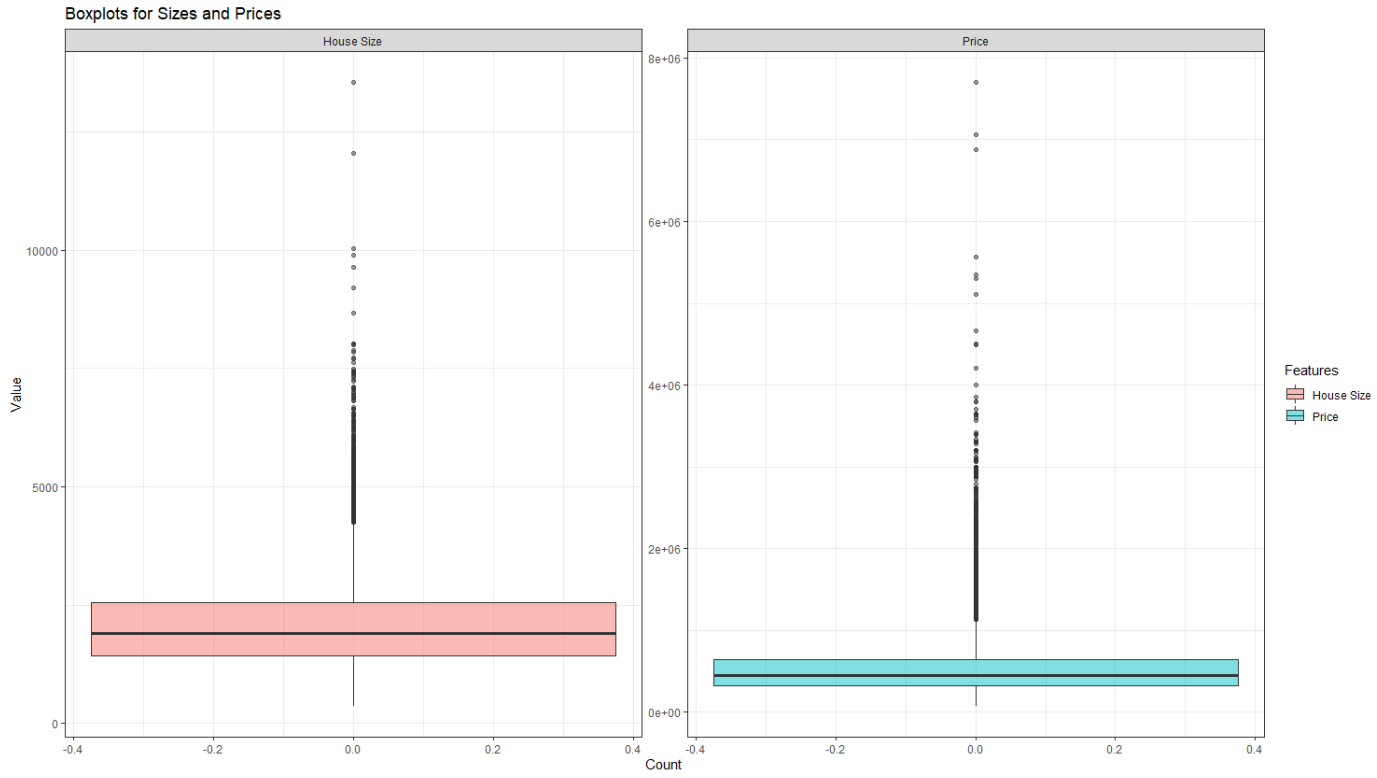


Fig1

    - Log transformations on both these numerical features gives us a better normal looking distribution.

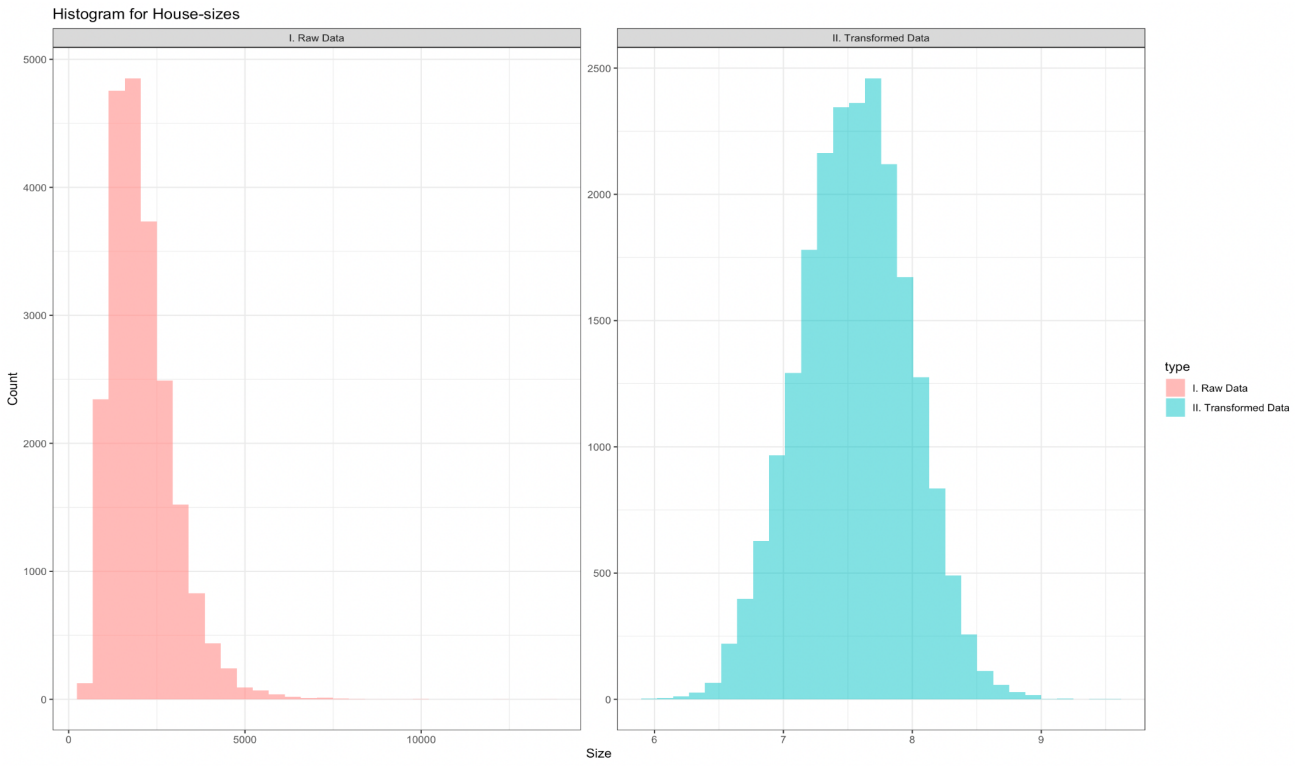

Histogram for House-sizes
I. Raw Data
II. Transformed Data
Count
Size
type
I. Raw Data
II. Transformed Data

Fig2

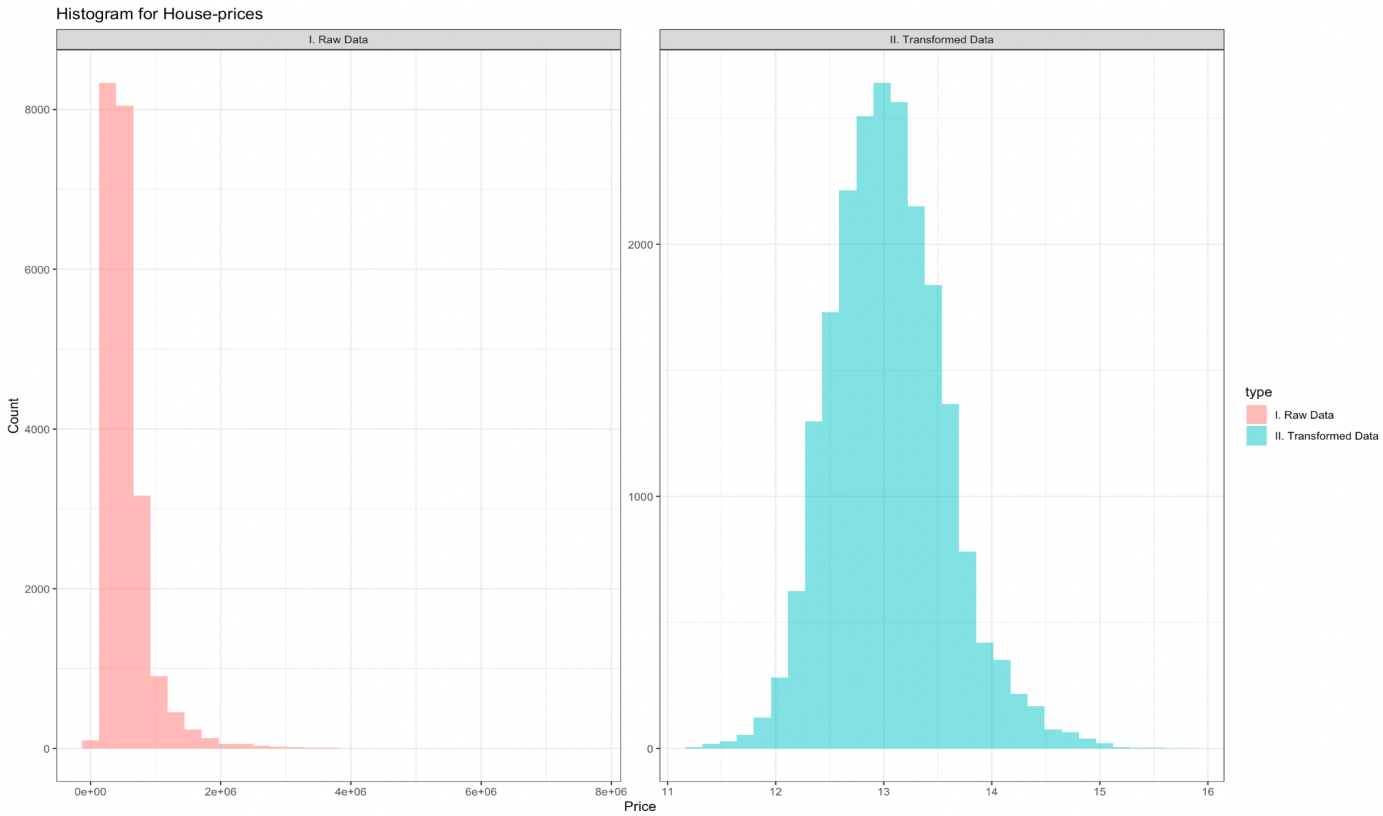

Histogram for House-prices
I. Raw Data
II. Transformed Data
Count
Price
type
I. Raw Data
II. Transformed Data

Fig3

Feature Selection

Identify the features affecting the house price

We first select the numerical and ordinal variables from the heatmap as they have high correlation with the log transformed target variable price.

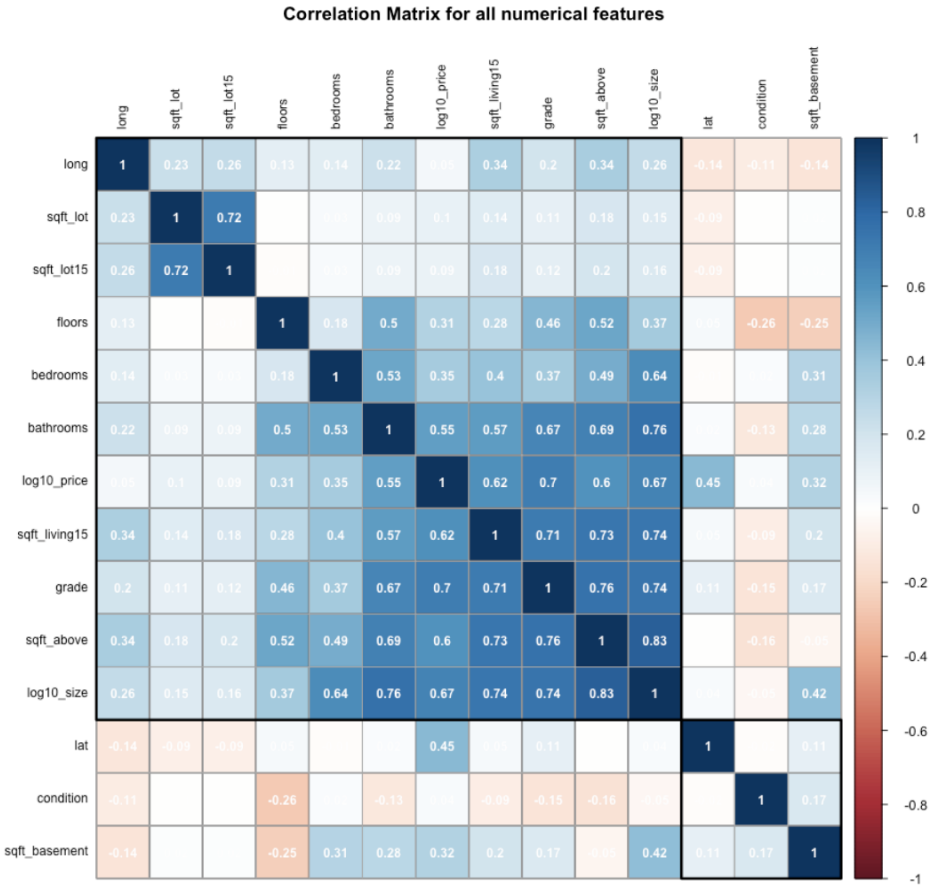


Fig4

We then exclude the variable sqft_living_15 and sqft_above, which have high correlation with the target variable house price as well as the independent variable sqft_living, to avoid multicollinearity.
The ordinal variables ‘bedroom’, ‘bathroom’ and ‘grade’ also have high multicollinearity with the log transformed sqft_living variable. So, we plot the square footage of the property vs the price faceted by each of these variables:

1. House price vs Size faceted by number of bedrooms

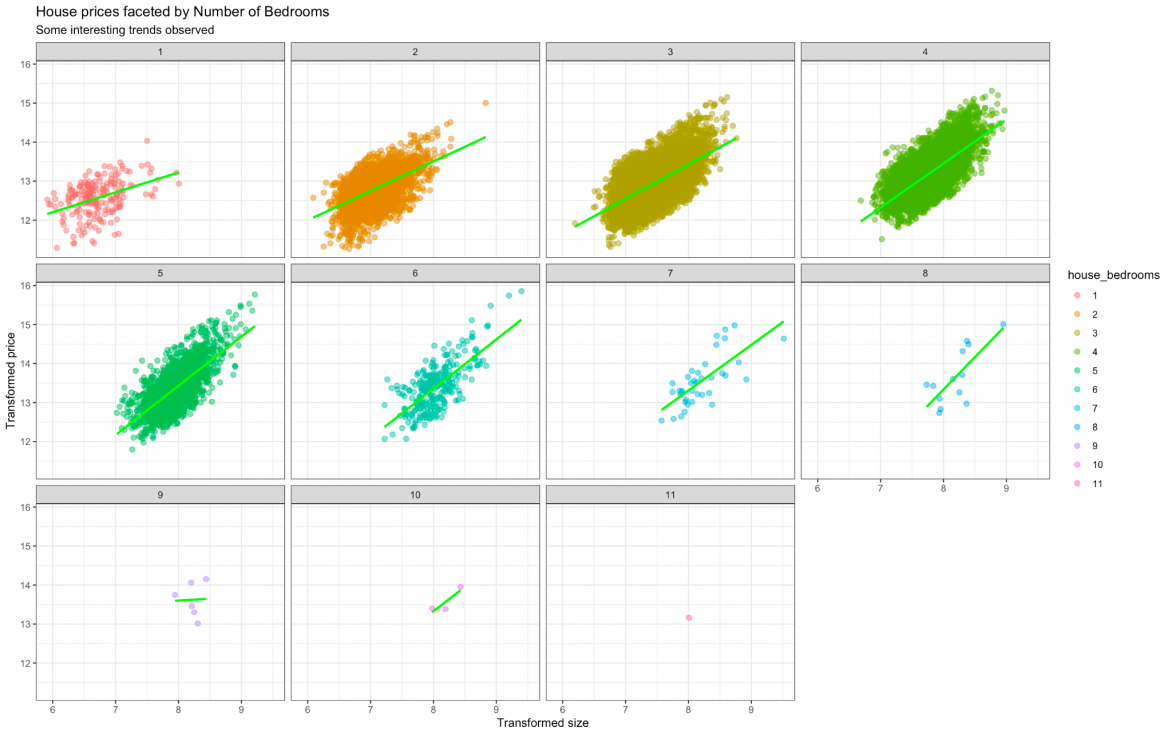


Fig5

2. House price vs Size faceted by number of bathrooms

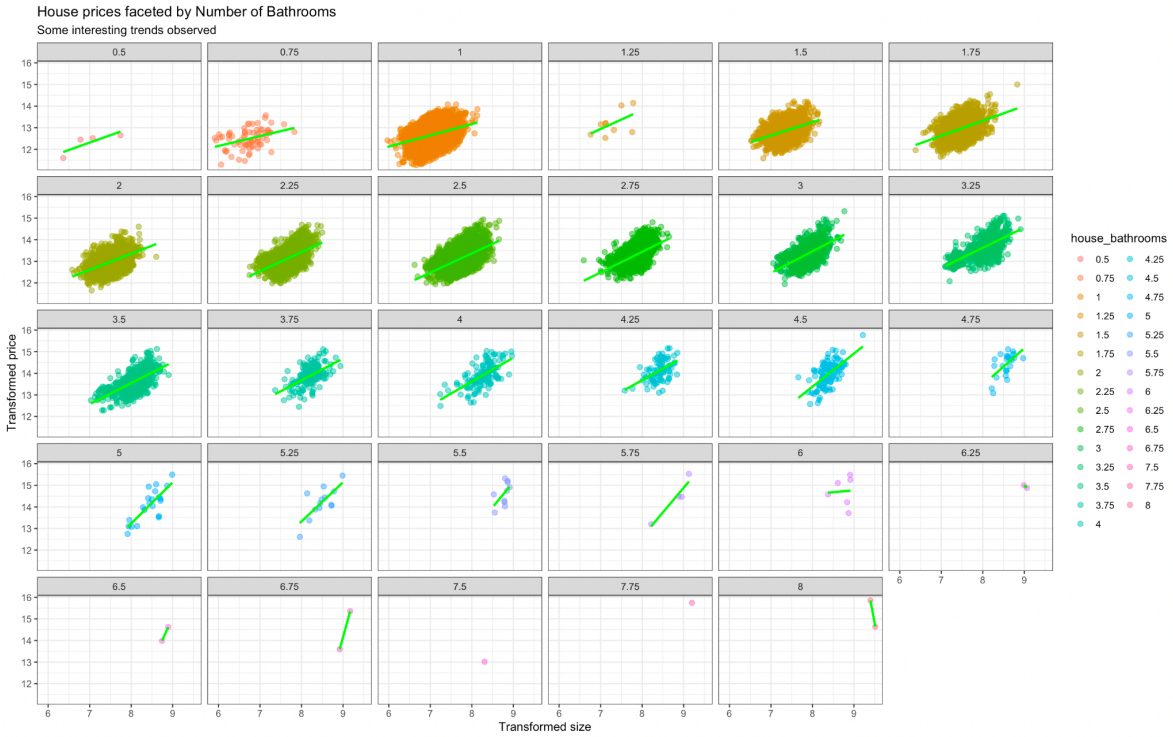


Fig6

3. House price vs Size faceted by grade

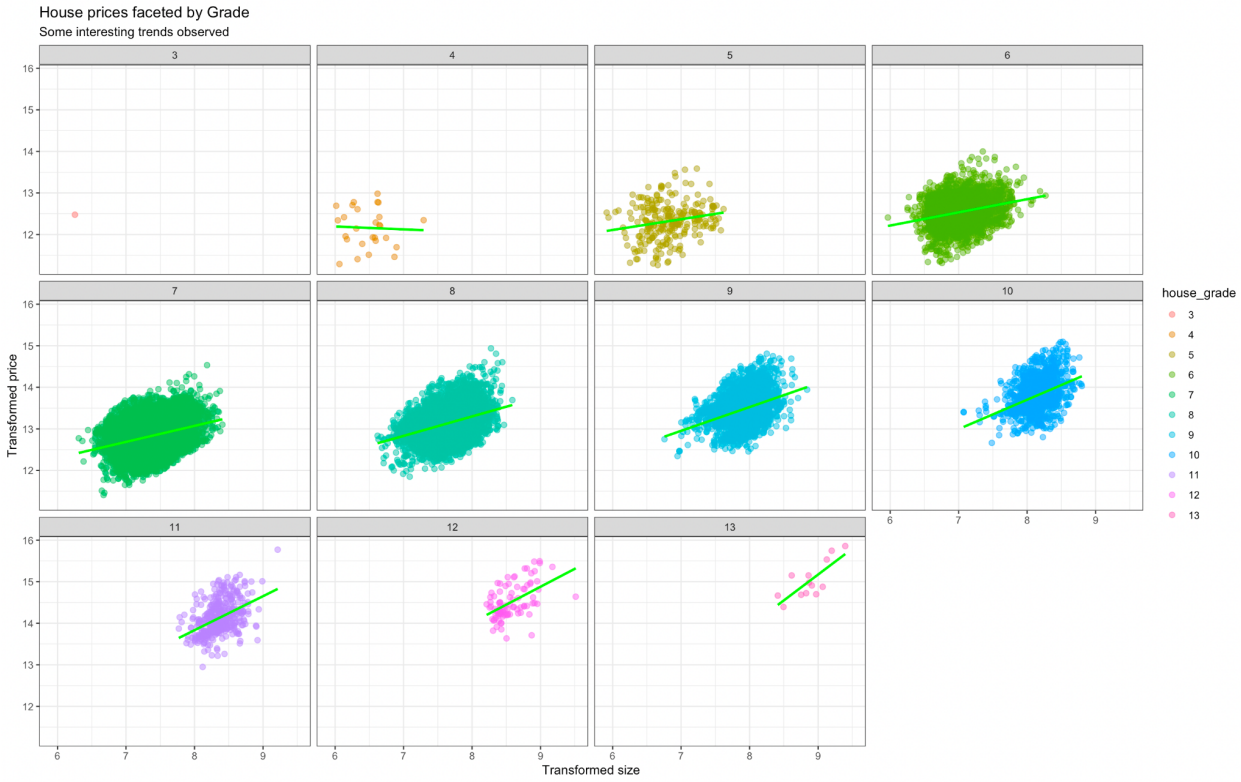


Fig7

We can see a change in the slopes of the simple linear regression lines as the conditional variable changes. This alludes to an interaction between the variables and the property price and hence we retain these variables in our model.

* We also experimented with the conditional of houses being the faceting variable, but there couldn't observe anything noteworthy (refer to Appendix FigA).

Investigating if location affects house prices

In housing development, the prices vary based on geographical features such as the neighborhood (rich or poor) of the house or the view (waterfront, coastwide). We tested this hypothesis for King County.

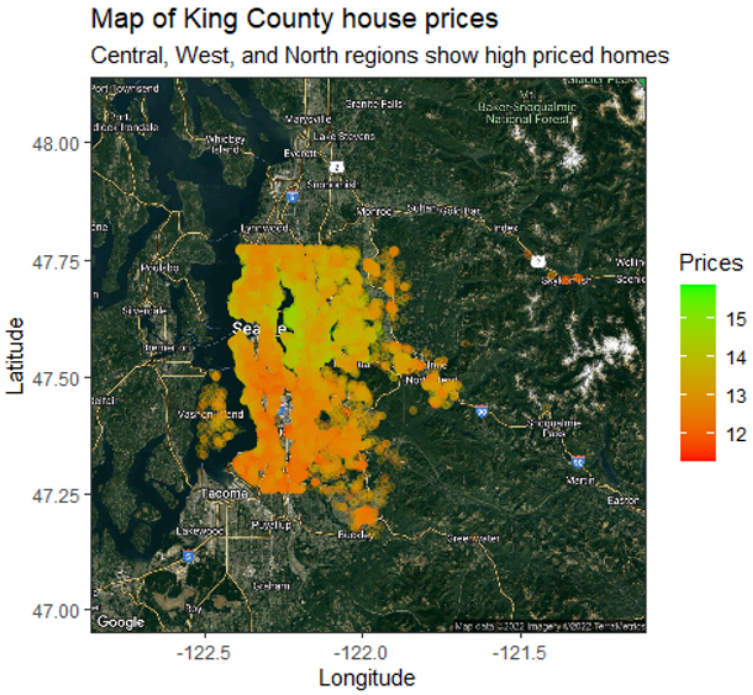


Fig8

The prices increase slightly as we move from the South to the North Seattle but not much variation across longitude. Along similar lines, we also see a correlation of 0.45 in the correlation plot (Fig3).
As the latitude feature is affecting the house price, we are transforming the feature onto a 0 to 1 scale and retain it in our model.

Finally, based on the above findings, the features ‘sqft_living’, ‘bathrooms’, ‘bedrooms’, ‘latitude’ and ’grade’ are great candidates to predict house prices in King County, Seattle.

Fitting a model

Identify if the home-price can be explained by a linear relationship with the chosen features or is a more complex model required

To create a linear model we begin by utilizing all the independent variables from the feature selection process along with the interactions between the features that were explored using facet plots. Fig9 shows that these variables have non-zero coefficients. Since the latitude variable is scaled down, we did not explore its interactions with house-sizes.

We also plotted the coefficients for the model containing all pairwise interactions with house-sizes (refer to Appendix FigB). An even more accurate model could be created after exploring a few of them.

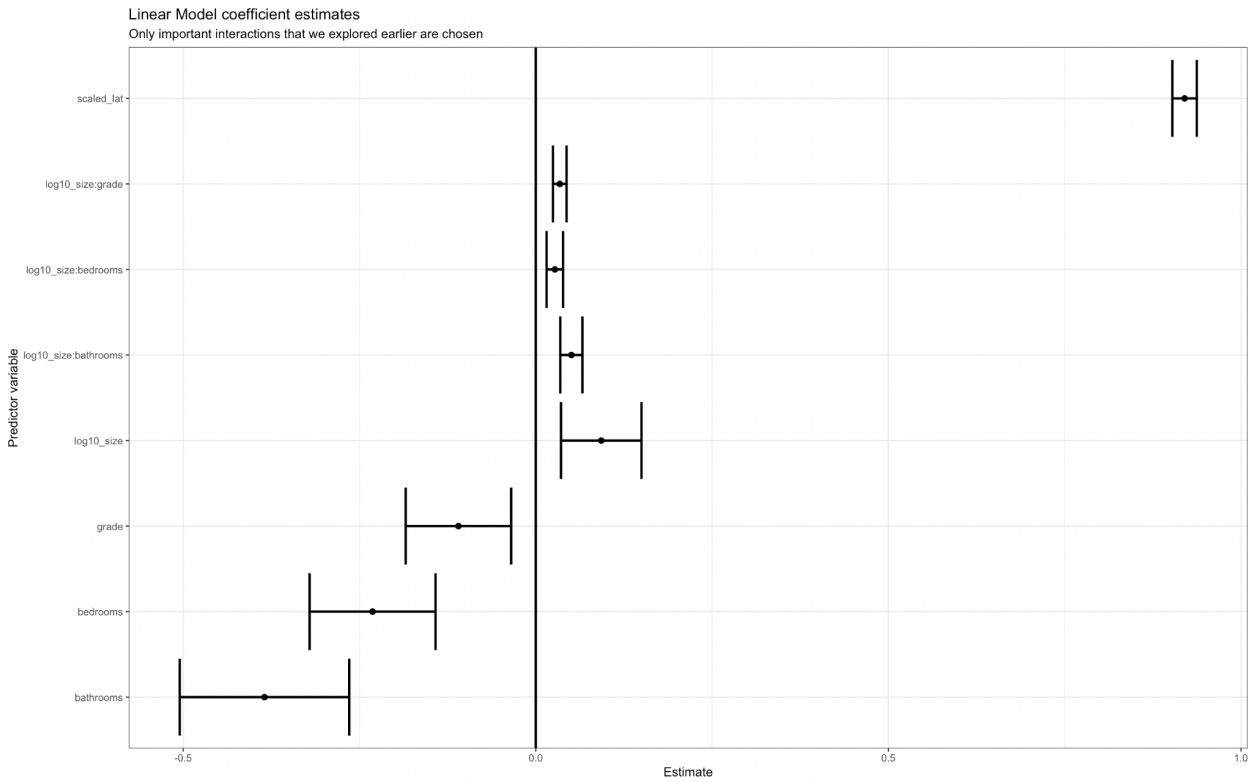


Fig9

Fig10 shows the steps taken by the forward selection algorithm. In the first iteration the model selects the interaction with log10_size and grade, in the second iteration it adds scaled latitude, in the third the grade feature and so on.

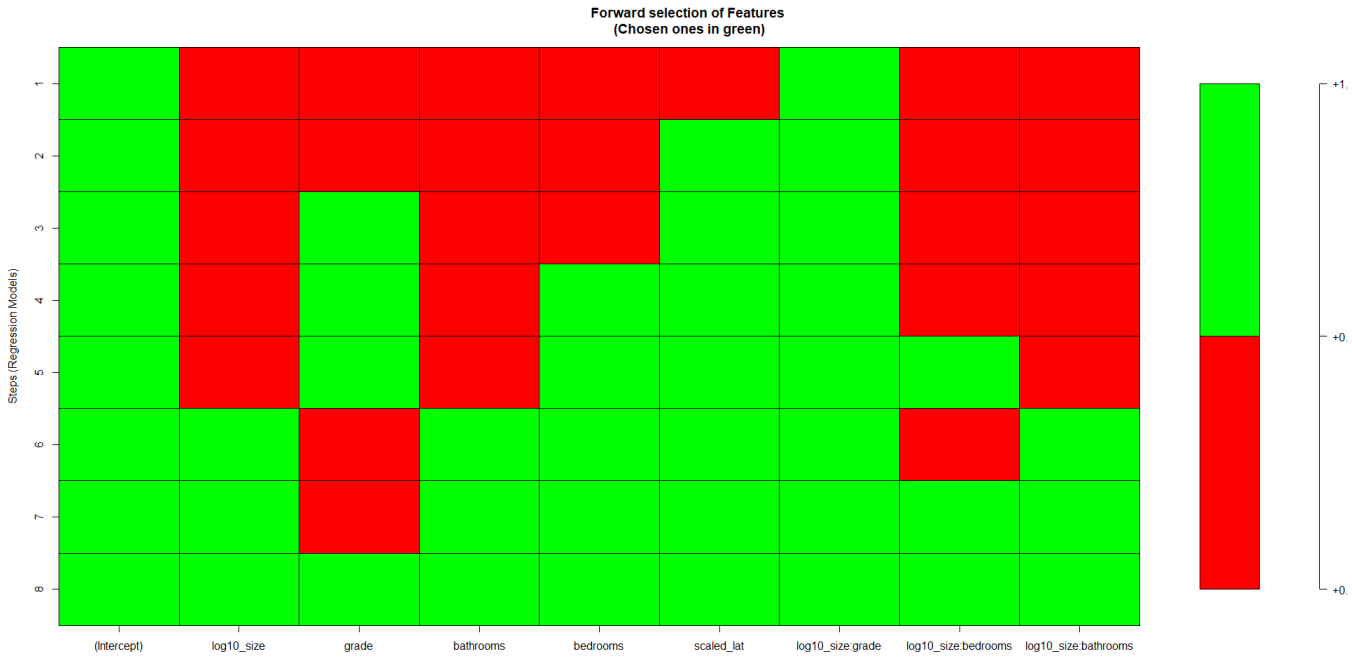


Fig10

To measure how well the selected variables explain the target variable, in Fig11 we plot the Mallow's CP metric against the steps of the forward selection model. Mallows' Cp is a metric that is used to pick the best regression model among several by calculating the Residual Sum of Squares. We see a reduction in the CP metric value as steps increase, indicating that the forward selection model is performing well.

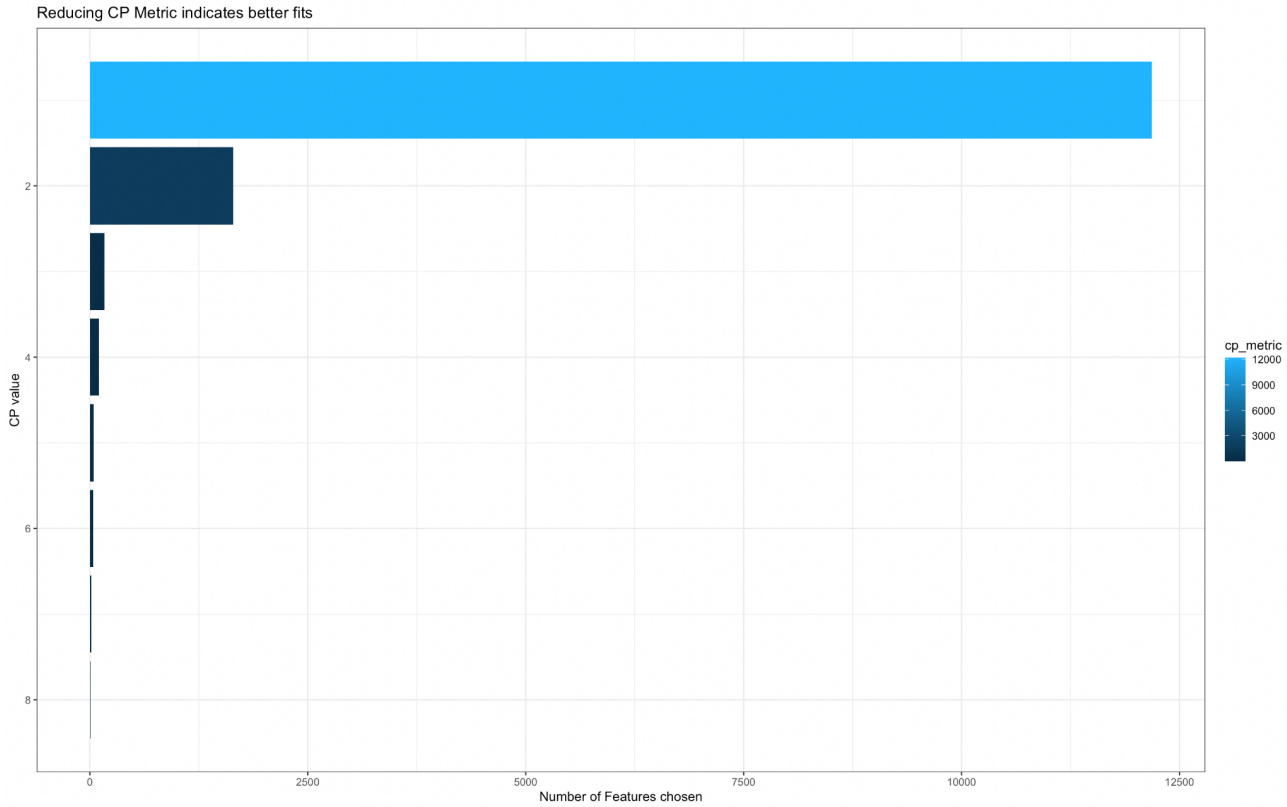


Fig11

The overall final model seems to perform respectably well with its R-squared and Adjusted R-squared values approximately 0.707. Also, this final model's residual plot doesn't show signs of heteroskedasticity (refer to Appendix FigC).

```
lm(formula = log10_price ~ (log10_size + grade + bathrooms +
    bedrooms + scaled_lat) + (log10_size * grade + log10_size *
    bedrooms + log10_size * bathrooms), data = house_prices)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43201 -0.18400 -0.01643  0.16775  1.35825

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         10.696701   0.220084  48.603  < 2e-16 ***
log10_size           0.092808   0.029104   3.189  0.00143 **
grade               -0.109772   0.038136  -2.878  0.00400 **
bathrooms           -0.384716   0.061326  -6.273 3.60e-10 ***
bedrooms            -0.231492   0.045559  -5.081 3.78e-07 ***
scaled_lat           0.919916   0.008834 104.130  < 2e-16 ***
log10_size:grade     0.033972   0.004916   6.911 4.96e-12 ***
log10_size:bedrooms  0.026962   0.005945   4.535 5.79e-06 ***
log10_size:bathrooms 0.050404   0.008007   6.295 3.14e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.285 on 21587 degrees of freedom
Multiple R-squared:  0.707,     Adjusted R-squared:  0.7069
F-statistic:  6510 on 8 and 21587 DF,  p-value: < 2.2e-16
```

Fig12

Limitations/Further Work

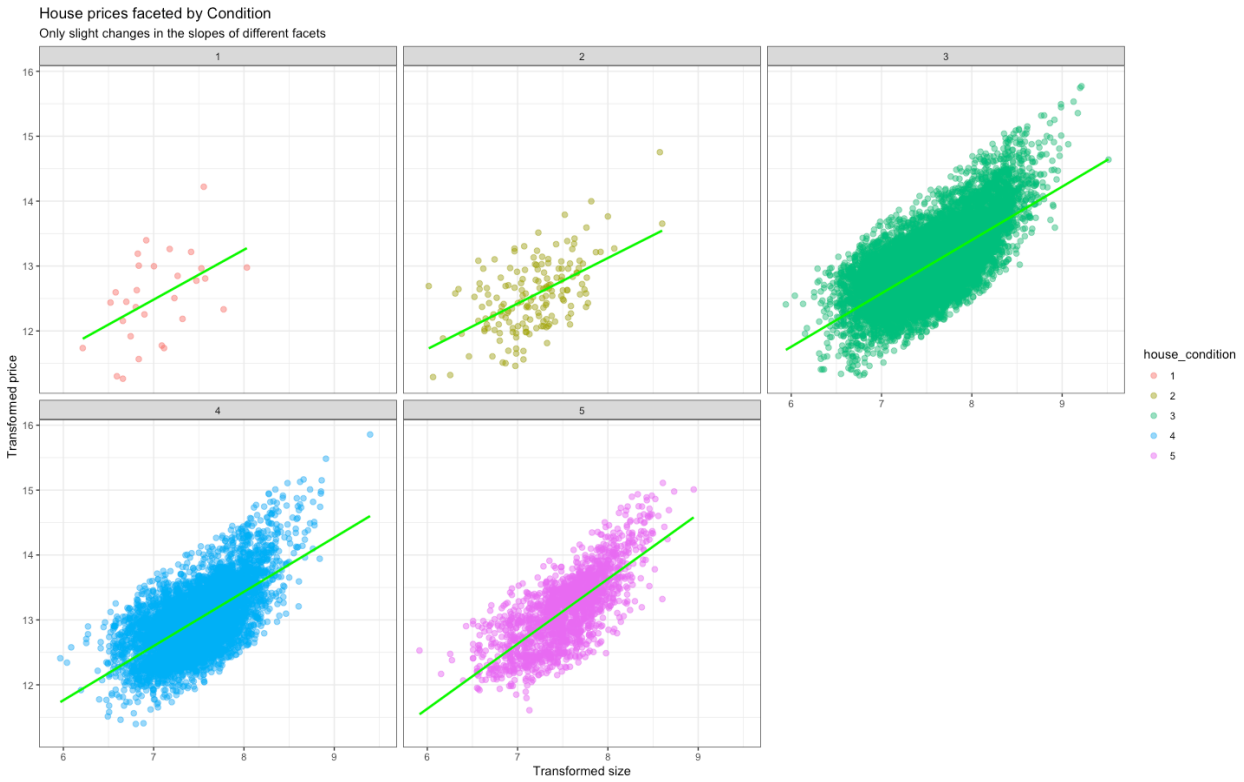
Outside the scope of EDA and fitting a linear model, we can create new features from interactions between the existing ones. We can then analyze if these newly engineered features explain the target variable better than the existing features. We can also apply more complex regression models to examine an increase in prediction accuracy.

A major limitation that we faced with the current dataset is that the dataset holds house prices only for the years 2014 and 2015. Hence we are unable to visualize the trends of the house price and its interactions with the variables and comment on the importance of these variables in predicting house prices over time.

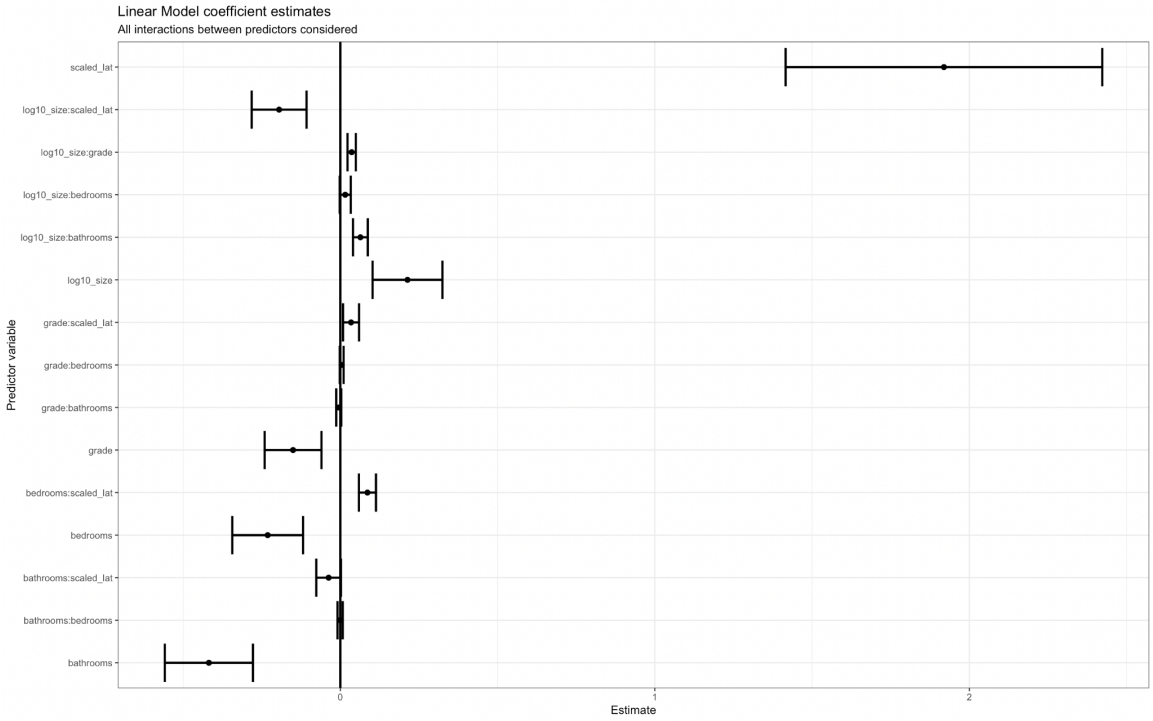
Conclusion

- Determined the best factors for predicting house prices by plotting the correlation matrix and observing that the sqft_living (size) feature was highly correlated with the target house price.
- Used facet plots to check whether grade, condition, bedrooms, and bathrooms should be used in the predictive model. Based on the results of our analysis, we selected grade, bedrooms and bathrooms features for our model.
- We also observed that the house prices increased slightly as we moved from the South to the North Seattle with negligible variation across longitude. Thus, we selected the latitude feature for our model.
- Used the features along with interactions between a few to derive a linear model that best explained the target variable.
- Reduced the variables to the ones explored and those which have coefficients more than zero.
- Explored forward selection of relevant features and measured performance of the linear model using the Mallow's CP metric.

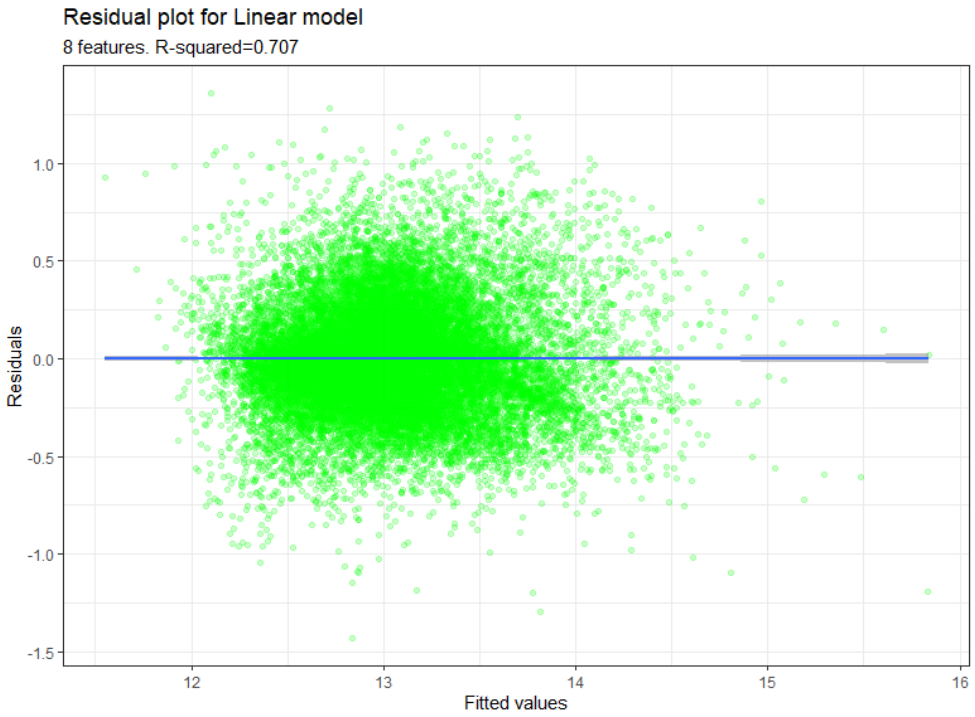
Appendix



FigA


Linear Model coefficient estimates
All interactions between predictors considered
scaled_lat
log10_size:scaled_lat
log10_size:grade
log10_size:bedrooms
log10_size:bathrooms
log10_size
grade:scaled_lat
grade:bedrooms
grade:bathrooms
grade
bedrooms:scaled_lat
bedrooms
bathrooms:scaled_lat
bathrooms:bedrooms
bathrooms
Predictor variable
0
1
2
Estimate

FigB


Residual plot for Linear model
8 features. R-squared=0.707
1.0
0.5
0.0
-0.5
-1.0
-1.5
Residuals
12
13
14
15
16
Fitted values

FigC

Variable Name	Description
id	Unique ID for each home sold
date	Date of the home sale (between 2014 to 2015)
price	Price of each home sold
bedrooms	Number of bedrooms
bathrooms	Number of bathrooms, 0.5 accounts for a room with a toilet but no shower
sqft_living	Square footage of the apartment's interior living space
sqft_lot	Square footage of the land space
floors	Number of floors
waterfront	Dummy variable if the apartment was overlooking the waterfront or not
view	An index from 0 to 4 of how good the view of the property was
condition	An index from 1 to 5 on the condition of the apartment,
grade	1 to 13 scale, 1 – 3 subpar building construction/design, 7 has average construction/design, and 11 - 13 have high-quality construction/design.
sqft_above	The square footage of the interior housing space that is above ground level
sqft_basement	The square footage of the interior housing space that is below ground level
yr_built	The year the house was initially built
yr_renovated	The year of the house's last renovation
zipcode	What zip code area the house is in
lat	Latitude
long	Longitude
sqft_living15	The square footage of interior housing living space for the nearest 15 neighbors
sqft_lot15	The square footage of the land lots of the nearest 15 neighbors

Table1