# **Predicting House Prices in King County**

**Using R for Analytics Final Project Team: Rated R**

**Abhisek Gupta | Akshay Ahuja | Komal Suresh**

Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907

[gupta362@purdue.edu](mailto:gupta362@purdue.edu); [ahuja11@purdue.edu](mailto:ahuja11@purdue.edu); [suresh19@purdue.edu](mailto:suresh19@purdue.edu)

## **Abstract**

In this project, we present a R-Shiny app that utilizes a linear regression model to predict house prices for the houses located in King County, Washington, United States. As a team, we explored various datasets available on public domain, but wanted to utilize our time to build a decision support system which can be used in variety of settings. All three of us have been fascinated with how the housing bubble in 2006 brought the entire financial and world market on its feet. In the wider range of things, we thought, there must be a better way to predict house prices taking historical data as input. This project is small stab in developing a predictive model that answers these questions. Our shiny app accepts values for selected features from a user who is interested in estimating the price of a house in King’s County. Based on the selection of these features the app provides the user with an estimated price along with a confidence and prediction interval. We further plan to expand this model by incorporating economic variables and using flexible predictive modelling techniques to make our predictions more robust and applicable in wider settings.

**Keywords:**

King County, predict house price, Linear regression, estimation

## **Business Problem:**

Predicting house sale prices of King County based on variables such as number of floors, square feet living, age of the house, city, number of bedrooms, etc.

The stakeholders are:

1. Real estate agents
2. Individuals looking to sell their property or buy a new home.

**Discussion whether the problem is amenable to an analytics solution:**

The business problem is amenable to to an analytics solution as there is a huge value proposition in estimating house prices given a set of parameters. Availability of historical data in combination with advanced analytics in this case makes it easier to fulfill business proposition of predicting the price.

**Refinement of the problem to identify any delineate constraints:**

We refined the business problem to only include the features available to us. There were challenges around zip code locations, school districts locations, demographic variables which forced us to remove them from our decision support system.

**Define the initial set of business benefits**

Some initial set of benefits of are:

* Aggregating and hypothesizing about some key parameters affecting house prices resulting in client exposure to factors he didn’t consider before while making a purchase
* Collecting and aggregating data to be made available for ready usage now and also in future

**Statement claiming that stakeholder agreement on the business problem statement has been determined**

Stakeholder agrees to the business problem under the awareness about current set of constraints and gives a go ahead to vendor to reduce the business problem in smaller component and look for a possible solution

**Analytics Problem**

Building a predictive model using linear regression to predict the price of a house in King County based on features of the house.

**Proposed set of drivers and relationship to outputs:**

1. We hypothesize that features such as number of bedrooms, number of bathrooms, area of living, condition and grade of the house, area of the lot would have a positive effect on the predicted price of the house; the higher the value, the more will be the price
2. We also believe that higher value of “age of the house” will drive the price down.
3. Location of the house should play an important role in determining the price of the house. A house located in a posh area should have a higher price as compared to one located in suburbs.

**Key assumptions related to the problem:**

1. Seasonality does not affect house prices. We have data for an entire year and for the sake of simplicity, we assume that price of a house will not be dependent on the month/season of valuation.
2. All neighborhoods in the city will have the same price ie different areas of Auburn will be priced equitabily.

**Key metrics of success:**

For continuous-valued estimation problems such as ours, metrics often used for assessing models are R-squared values (R2) and Mean Squared Error (MSE). We can also examine the entire range of predicted values by considering scatter plots of actual versus predicted values or actuals versus residuals(errors).

**Statement claiming stakeholder agreement on the approach**

## 

Geographical Distribution of Houses in King County

## **Data**

Data for building this model was downloaded from Kaggle site <https://www.kaggle.com/harlfoxem/housesalesprediction>.

The raw dataset was downloaded in csv format and contains 21613 rows and 21 columns. It includes homes sold between May 2014 and May 2015.

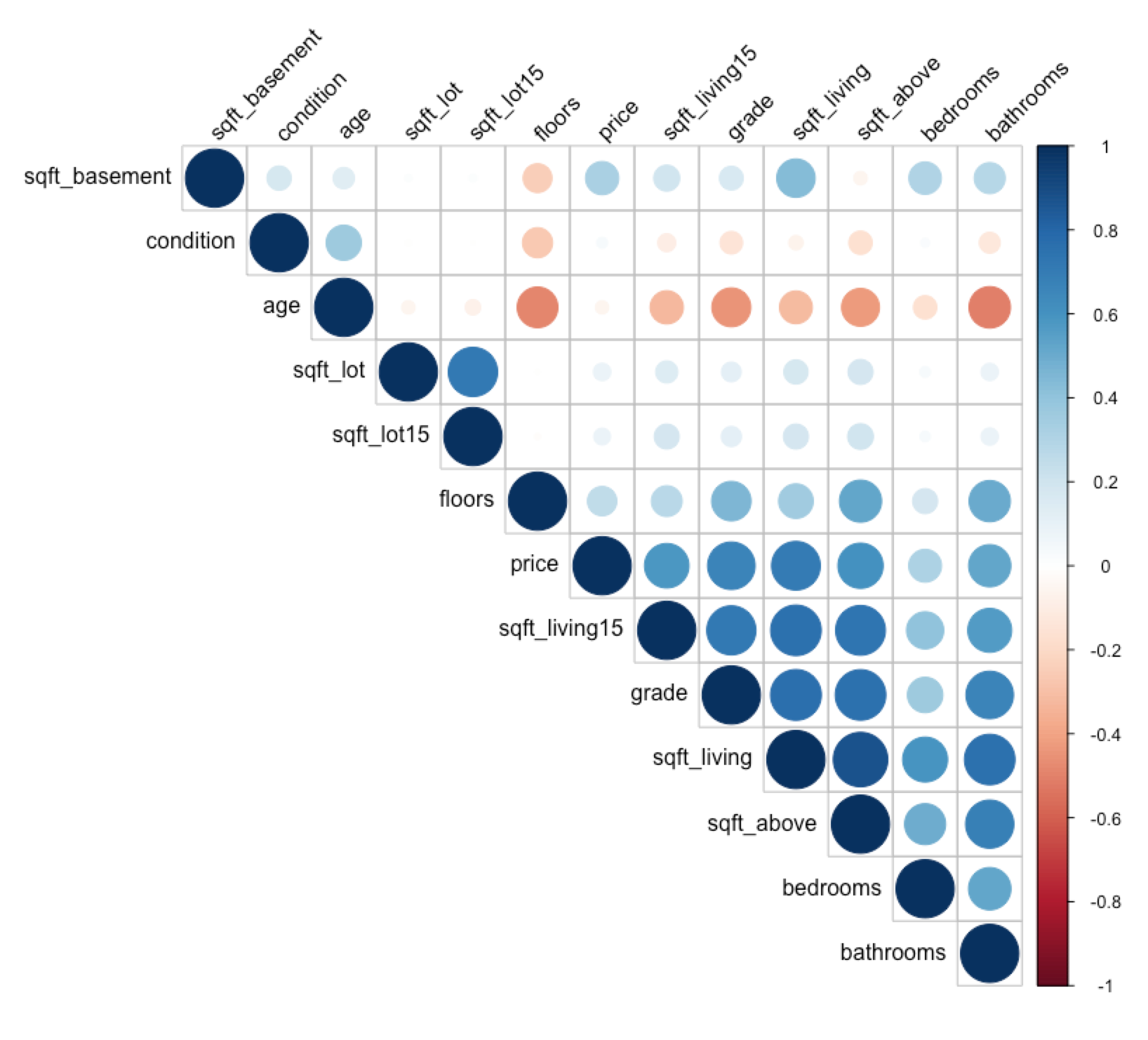
## **Data Cleaning:**

1. We tested our data for missing values and found no missing values
2. We checked for duplicate entries and found that 177 rows out of 21613 were duplicate on “House Id” level but had different dates of evaluation. So, we deleted the duplicate entries, retaining the most recent entry for a House Id
3. We dropped a variable called “View” because there was no explanation available for that variable
4. We converted some variables such as “Waterfront” and “Renovation” to categorical variables

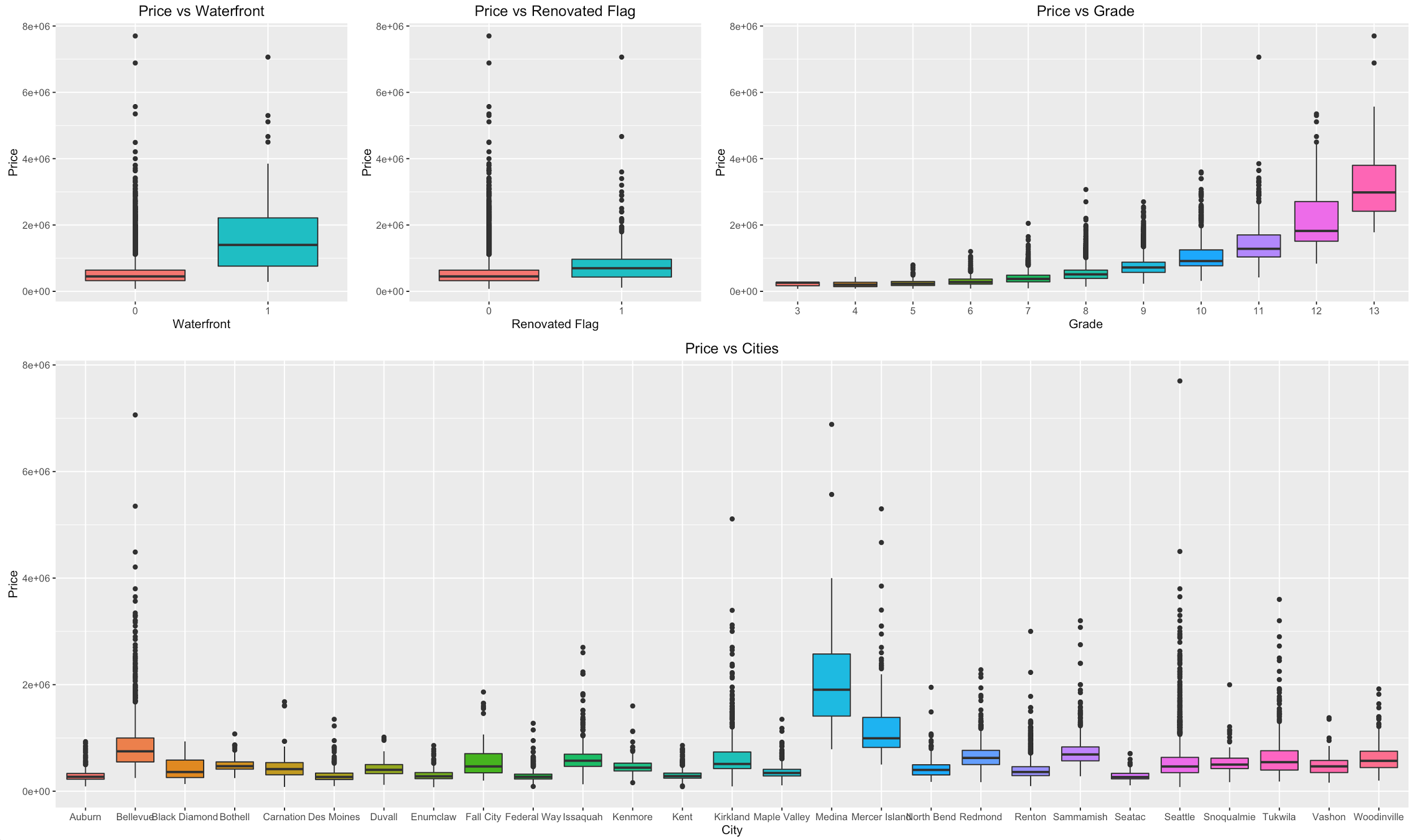
**Identifying relationships in data:**

1. We tried to find trends in data through exploratory data analysis using scatter and box plots
2. We tested for multicollinearity using “corr” function to identify which of the variables were highly correlated. “Corr” function also helped us understand which variable was highly correlated with the predicted variable.

**Plots that gave us meaningful insights:**

Correlation matrix plot:

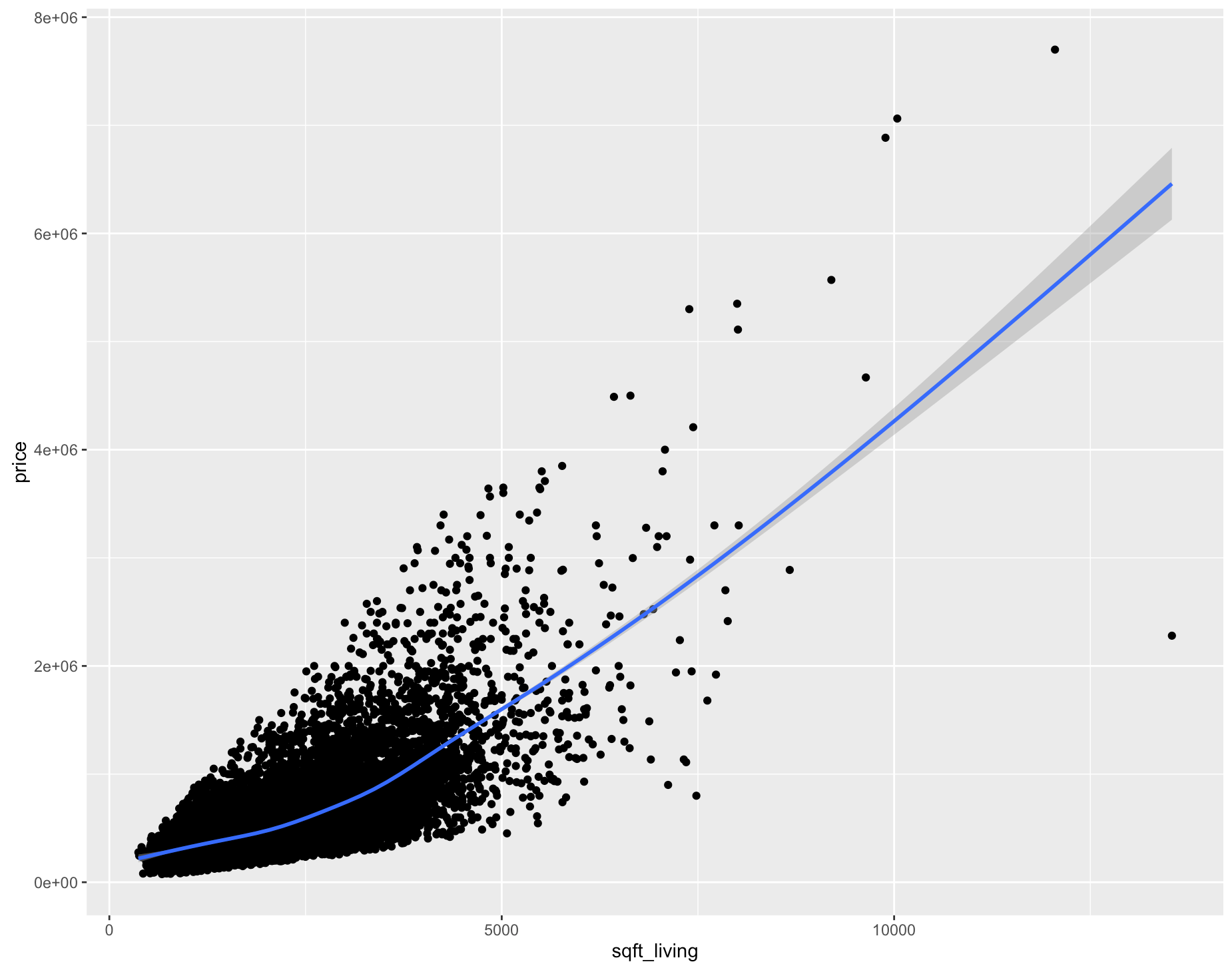
According to this correlation matrix plot, variables such as square feet living and square feel above are highly correlated.



Bivariate analysis of significant variables

Insights from the plots above:

1. The box plot of waterfront vs price of the house shows that houses with waterfront view have a higher price as compared to those that do not have the view
2. The box plot of Renovated flag vs the price shows that houses that were renovated recently are costlier than the ones that aren’t
3. The box plot of Grade Vs Price of the house shows that as the grade increases, there is an increase in price.
4. The box plot of cities vs the price gives us an idea of the price range of houses in 27 cities in King county.



Scatter plot of Square feet living vs Price

**Methodology Selection**

We debated various ways to explore our dataset and generate some interesting insights on housing market in King’s County. We decided to use some summary statistics and exploratory plots to understand our data and relationship between price and other variables. This approach of using descriptive statistics had limitations in terms of accuracy and bias on part of analyst. For example, one might give more importance to presence of waterfront as an important factor in the analysis just based on the boxplot of the same with price. However, just looking at the box plot might show some difference but conducting some hypothesis testing might show that the difference is not statistically significant. We felt that given the amount the amount of time available to use, there was scope of using a predictive model as a solution to our model. We researched ways of predicting a continuous dependent variable using independent variables and came up with two ways of approaching the problem:

1. Multiple linear regression: A common predictive analytics technique that helps explain the relationship between one independent variable and more than one dependent variables
2. Non-linear curve fitting: A predictive modeling technique that works by non-linear curve fitting technique. It uses high order polynomial equation to predict the dependent variable.

We preferred multiple linear regression because it suited the strengths of the team and could have been successfully completed in the given time period.

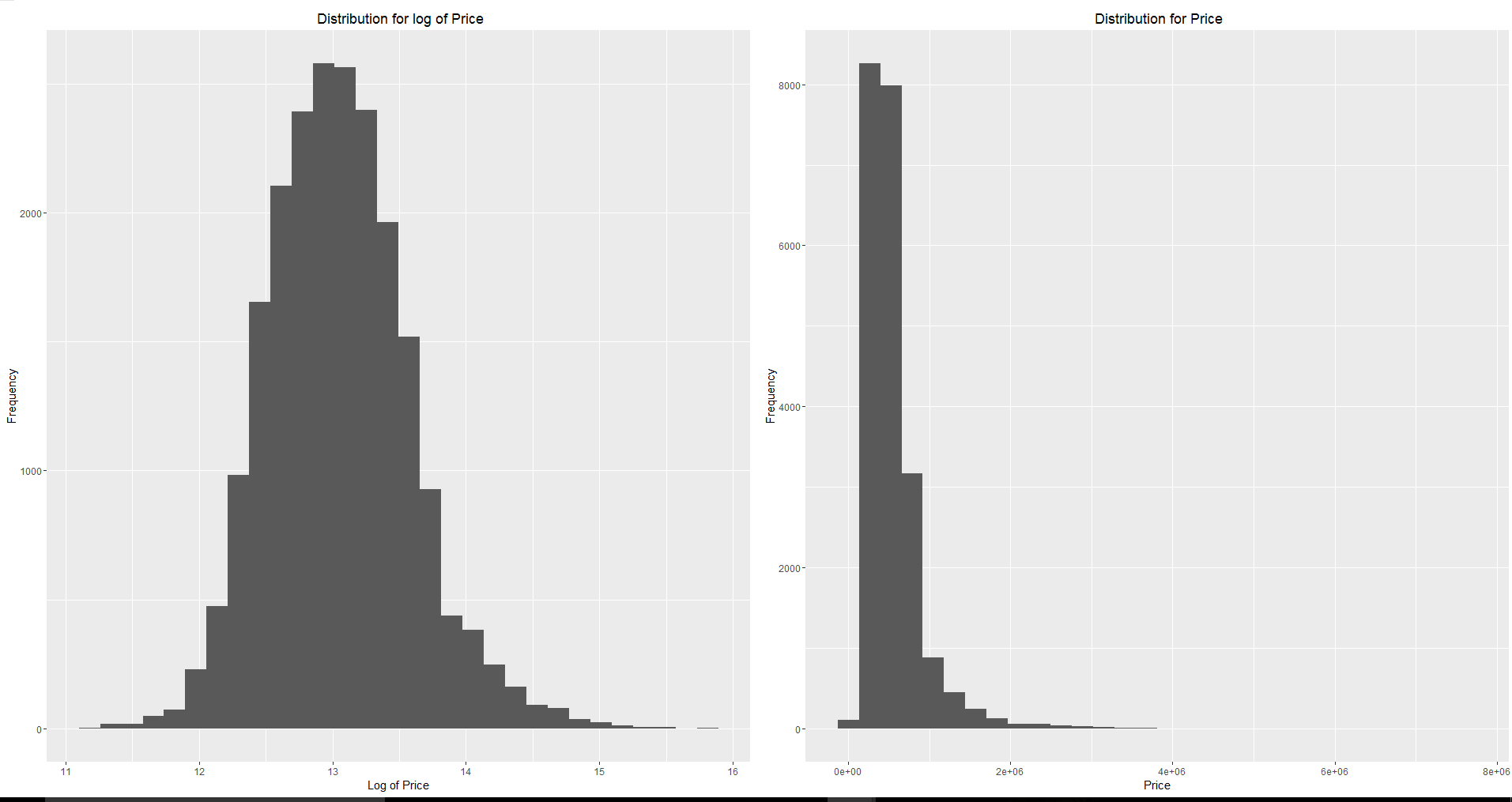
**R is a viable tool because of the following reasons:**

1. R is free and open-source and is the most widely used statistical programming language
2. There are many user-contributed statistical packages to choose from for data analysis, manipulation, prediction and validation.
3. Some of the packages we used in our project are:
   1. Sqldf, Plyr: Data Manipulation
   2. Ggally, ggplot2: Data Exploration
   3. Faraway, lawstat: Model validation
4. R can create great visualizations that help that can be either used for data exploration or can be directly plugged into final presentation.

Based on the above mentioned approach we choose Multiple Linear Regression model as our method of choice to predict the house prices for the King’s County.

## **Model Building**

**Exploratory Data Analysis**: We did some EDA to understand the correlation between house price and continuous variables details of which are given in the data section. We also identified that the distribution of price was heavily skewed towards the right. So, we transformed price by taking the log of the variable and used it as the target variable in the model.



**Building the model:** Dataset was divided into a training set and a test set in ratio of 0.75/0.25. The training set was used to build the model while the test dataset was used to validate the model accuracy.Model was built using lm function in R.

lm(formula = price\_log ~ ., data = house\_train\_v3)

Residuals:

Min 1Q Median 3Q Max

-1.47250 -0.14329 0.00936 0.15467 1.09653

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.145e+01 1.972e-02 580.723 < 2e-16 \*\*\*

bedrooms -3.332e-02 3.021e-03 -11.030 < 2e-16 \*\*\*

sqft\_living 3.782e-04 3.885e-06 97.344 < 2e-16 \*\*\*

sqft\_lot 5.505e-07 5.594e-08 9.841 < 2e-16 \*\*\*

floors 1.102e-01 5.580e-03 19.744 < 2e-16 \*\*\*

waterfront1 5.323e-01 2.537e-02 20.983 < 2e-16 \*\*\*

condition 7.019e-02 3.714e-03 18.900 < 2e-16 \*\*\*

sqft\_basement -4.634e-05 6.595e-06 -7.027 2.20e-12 \*\*\*

age 8.983e-04 1.048e-04 8.571 < 2e-16 \*\*\*

cityBellevue 8.183e-01 1.363e-02 60.047 < 2e-16 \*\*\*

cityBlack Diamond 2.065e-01 3.324e-02 6.214 5.30e-10 \*\*\*

cityBothell 4.608e-01 2.485e-02 18.540 < 2e-16 \*\*\*

cityCarnation 2.987e-01 3.197e-02 9.344 < 2e-16 \*\*\*

cityDes Moines 9.166e-02 2.236e-02 4.099 4.17e-05 \*\*\*

cityDuvall 3.055e-01 2.526e-02 12.096 < 2e-16 \*\*\*

cityEnumclaw 7.040e-02 2.265e-02 3.109 0.00188 \*\*

cityFall City 3.996e-01 3.621e-02 11.035 < 2e-16 \*\*\*

cityFederal Way 3.142e-02 1.543e-02 2.037 0.04167 \*

cityIssaquah 5.668e-01 1.581e-02 35.851 < 2e-16 \*\*\*

cityKenmore 4.393e-01 2.124e-02 20.680 < 2e-16 \*\*\*

cityKent 6.241e-02 1.379e-02 4.526 6.07e-06 \*\*\*

cityKirkland 6.798e-01 1.461e-02 46.537 < 2e-16 \*\*\*

cityMaple Valley 1.622e-01 1.651e-02 9.825 < 2e-16 \*\*\*

cityMedina 1.244e+00 4.604e-02 27.024 < 2e-16 \*\*\*

cityMercer Island 9.199e-01 2.197e-02 41.874 < 2e-16 \*\*\*

cityNorth Bend 3.589e-01 2.348e-02 15.287 < 2e-16 \*\*\*

cityRedmond 6.458e-01 1.466e-02 44.060 < 2e-16 \*\*\*

cityRenton 2.555e-01 1.315e-02 19.419 < 2e-16 \*\*\*

citySammamish 6.151e-01 1.554e-02 39.580 < 2e-16 \*\*\*

citySeatac 8.753e-02 2.874e-02 3.045 0.00233 \*\*

citySeattle 5.871e-01 1.157e-02 50.735 < 2e-16 \*\*\*

citySnoqualmie 3.660e-01 2.071e-02 17.666 < 2e-16 \*\*\*

cityTukwila 7.046e-01 1.873e-02 37.609 < 2e-16 \*\*\*

cityVashon 3.492e-01 3.219e-02 10.848 < 2e-16 \*\*\*

cityWoodinville 5.167e-01 1.771e-02 29.175 < 2e-16 \*\*\*

yr\_renov\_flag1 1.456e-01 1.553e-02 9.377 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

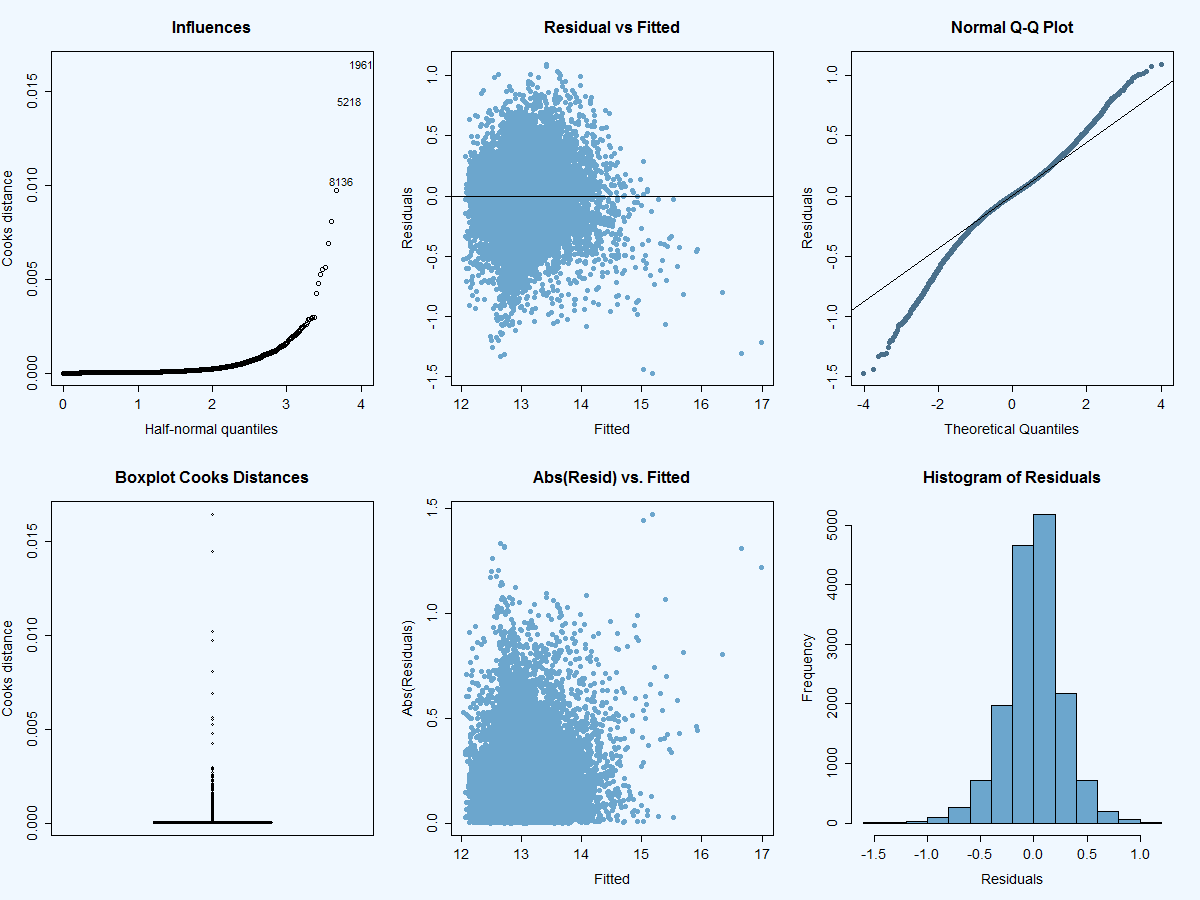
Residual standard error: 0.2708 on 16040 degrees of freedom

Multiple R-squared: 0.7358, Adjusted R-squared: 0.7352

F-statistic: 1276 on 35 and 16040 DF, p-value: < 2.2e-16

**Model Validation:** We ran some regression diagnostics to validate the model assumptions and remove influential observations. We iterated upon the model thrice to reach the suboptimal solution. The first step involved removing the insignificant features which had p values higher than our threshold confidence interval. We dropped sqft\_lot15 from the model at this stage as we were unable to reject null hypothesis based for the same and proceeded to run the model again. We found all of our remaining predictors significant for the threshold intervals and started understanding influential observations and outliers. We plotted a cook’s distance half normal plot and identified a high leverage point. On closer examination we found that it had exceptionally high lot size and square foot living which had significant effect on the regression line. Post re-running the model after removal of the exceptional predictor, led to identification of two other influential points. We iterated on the model for the last time after removing the above mentioned points.

We also noticed that residual plots showed some variance as the estimated value of the target variable became exceptionally high. QQ plot also gave us some interested insights on the model performance. One thing which we noticed was that the model over-estimated house prices in the lower price range while under-predicted in upper range. This shows that a regression spline will probably be a better fit for our data and will lead to less bias in the model.



As a part of our business problem, we were tasked with predicting the house prices in king county. The Adjusted R-Square of our training was 73% while the R-Square for our test set was also close to 73%. The fact that R-Square for both test and training sets are very close to each other serves as a good validation for robustness of the model. Interpreting the R-Square in terms of the business problems, tells me that the model is capable of explaining 73% variance in price given that the input parameters are known. The Beta coefficients for various features help the stakeholders identify the importance of drivers and their effect on the house price. This helps him establish more causal relationships between the features and house price. For example, a user can quantify the impact on value of house price based on the beta value given that all predictors stay constant. Finally, the model gives the user a fair estimate of price based on his choice of house features and assist him in making a high value decision on purchase.

Some findings from the model are:

1. Waterfront and floors have been high impact features on the price of the house.
2. Model tends to overestimate house prices in the lower price range while under predict in upper range.
3. Sqft\_lot, Sqft\_above and Sqft\_basement has very high correlation with Sqft\_living

## **Functionality**

The Decision Support System lets the user input the house characteristics (such as number of bedrooms, area of living, area of the lot, number of floors, waterfront view, condition of the house, area of the basement, Year the house was built in, city it is located in and the year it was renovated in) and predicts the estimated house price based on the features entered.

**R packages we found useful during this project were:**

Dplyr, plyr and sqldf for data manipulation

Ggplot2, ggally for data visualization and exploratory data analysis

Faraway and lawstat for running the diagnostic tests on our model to validate model assumptions

We had to write conditional logic for:

1. Checking if the packages were present and then installing them if they weren’t
2. While building Shiny app

If we had more time, we would’ve spent more time exploring the variables Zip Code and Latitude/longitude. Considering the location of the house is one of the most important determinant in its price, it is important to extract as many features as possible from the above mentioned variables. One approach to do this would have been to map the houses to the school districts and use the ratings of public schools in the area as a predictor of price. Another approach be to look at demographic variables such as average household income, average household size, average household age, ethnicity in the area to estimate the price. We would also have liked to include economic variables like cost of living, area gdp etc in our model to predict prices better. If we had more experience, we would’ve followed the non-linear curve fitting approach to make the model more flexible.

**GUI Design and Functionality**

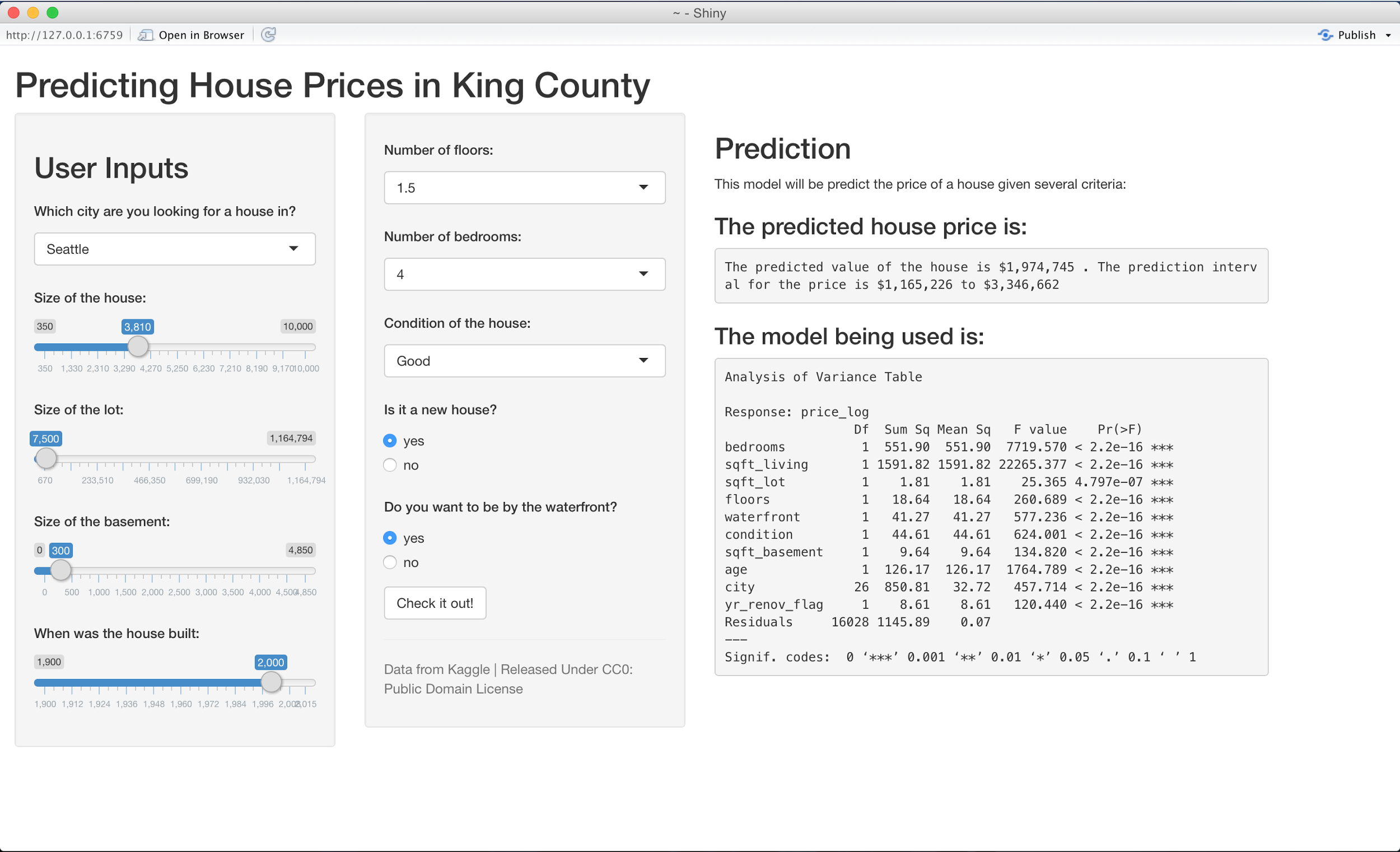
**Does the tool work without errors? Does it appear as good or better than the Shiny example templates on RStudio.com?**

The does work without any issues and predicts the correct value based on inputs using predict function on the linear regression model built in the back end.

We have built a Decision Support System using Shiny apps. The app has been designed keeping the user in mind. It asks the user for his/her preferences for several criteria and then gives an estimate for the price as well as a prediction interval.

In an attempt to be user-friendly, the app uses several methods to take inputs - dropdowns, sliders and radio buttons - which are optimized for each variable. Once the user sets the parameters and clicks the action button for the first time, the app becomes reactive. The prediction changes with a change in any variable.

Apart from this, the app has the summary statistics for the regression model used, to give some advanced users a brief idea of the significant variables.



The GUI for the Shiny App

**Conclusions**

Housing markets carry a lot of inherent uncertainty within themselves and its always difficult for various stakeholders to extract the accurate valuation of the house incorporating myriads of related variables. With our decision support system, home owners can leverage the power of analytics in to make a more informed decision when making such a decision. This protects interests of both buyers and sellers by ensuring they get a fair estimate of their price and are not cheated upon by one another.

Apart from the model, R has turned out to be a fairly easy to pick and use statistical programming language which enabled us to perform various parts of analytics with ease. On the predictive power of model, it can definitely be improved by creating a more flexible model. Some tests have shown possibility of spline fitting the model and it will be fruitful to explore such methodologies to predict house prices.

## **References**

https://www.kaggle.com/harlfoxem/housesalesprediction

https://rpubs.com/dnafrance/sicalc

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

http://shiny.rstudio.com/

**Appendix**

Data dictionary for raw variables is as follows:

|  |  |
| --- | --- |
| **Variable name** | **Variable definition** |
| id | House id |
| date | The date as of which a property's value is estimated. |
| price | Value of the property |
| bedrooms | Number of bedrooms |
| bathrooms | Number of bathrooms |
| sqft\_living | Square feet above + Square feet basement |
| sqft\_lot | Area of house including the yard |
| floors | Number of floors |
| waterfront | Waterfront view (yes=1, no=0) |
| view | Unknown |
| condition | Building Condition: Relative to age and grade. Coded 1-5. |
| 1 = Poor- Worn out. Repair and overhaul needed on painted surfaces, roofing, plumbing, heating and numerous functional inadequacies. Excessive deferred maintenance and abuse, limited value-in-use, approaching abandonment or major reconstruction; reuse or change in occupancy is imminent. Effective age is near the end of the scale regardless of the actual chronological age. |
| 2 = Fair- Badly worn. Much repair needed. Many items need refinishing or overhauling, deferred maintenance obvious, inadequate building utility and systems all shortening the life expectancy and increasing the effective age. |
| 3 = Average- Some evidence of deferred maintenance and normal obsolescence with age in that a few minor repairs are needed, along with some refinishing. All major components still functional and contributing toward an extended life expectancy. Effective age and utility is standard for like properties of its class and usage. |
| 4 = Good- No obvious maintenance required but neither is everything new. Appearance and utility are above the standard and the overall effective age will be lower than the typical property. |
| 5= Very Good- All items well maintained, many having been overhauled and repaired as they have shown signs of wear, increasing the life expectancy and lowering the effective age with little deterioration or obsolescence evident with a high degree of utility. |
| grade | Building grade- Represents the construction quality of improvements. Grades run from grade 1 to 13. Generally defined as: |
| 1-3: short of min building standards-cabin or inferior structure |
| 4: Generally older, low quality construction. Does not meet code. |
| 5: Low construction costs and workmanship. Small, simple design. |
| 6: Lowest grade currently meeting building code. Low quality materials and simple designs. |
| 7: Average grade of construction and design. Commonly seen in plats and older sub-divisions. |
| 8: Just above average in construction and design. Usually better materials in both the exterior and interior finish work. |
| 9: Better architectural design with extra interior and exterior design and quality. |
| 10: Homes of this quality generally have high quality features. Finish work is better and more design quality is seen in the floor plans. Generally, have a larger square footage. |
| 11: Custom design and higher quality finish work with added amenities of solid woods, bathroom fixtures and more luxurious options. |
| 12: Custom design and excellent builders. All materials are of the highest quality and all conveniences are present. |
| 13: Generally custom designed and built. Mansion level. Large amount of highest quality cabinet work, wood trim, marble, entry ways etc. |
| sqft\_above | The living area in a house not including the basement. |
| sqft\_basement | Basement area |
| yr\_built | Year the house was built |
| yr\_renovated | Year the house was renovated |
| zipcode | Zip code |
| lat | Latitude |
| long | Longitude |
| sqft\_living15 | The average house square footage of the 15 closest houses |
| sqft\_lot15 | The average lot square footage of the 15 closest houses |