**A Report on Additional Activity**

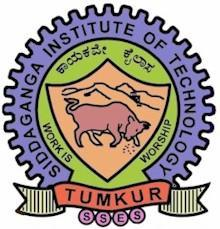
Submitted for Partial Fulfilment of V Semester Foundation of Data Science

**By Team Members:**

Name USN

1. Subramanya G 1SI18CS115

2. Vishwak Vemuru 1SI18CS135



**Department of Computer Science and Engineering**

**Siddaganga Institute of Technology, Tumakuru – 572103**

(An Autonomous Institution, Affiliated to VTU, Belagavi & Recognized by AICTE, New Delhi)

**2020 -2021**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| Sl No | Content | Page No |
| 1 | Introduction to R | 3 |
| 2 | Introduction to Dataset | 4 |
| 3 | Lasso Regression (Description of Algorithm) | 5 |
| 4 | Result Obtained | 6 - 10 |
| 5 | Conclusion | 11 |

R Programming Language

R is a programming language used primarily for statistical computing and graphics. It is similar to the ‘S Language’ and can be considered as a different implementation of S. As of January 2021, it ranks 9th in the TIOBE index for popularity of programming languages. It is originally written in C, FORTRAN, and R itself. It is freely available under the GNU GPL.

It consists of a command line interface and several GUI (Graphical User Interface) based development environments such as R Studio, Jupyter notebook are available. It is primarily used by Statisticians and Data miners for developing statistical software and data analysis. R has its own LaTeX-like documentation format, which is used to supply comprehensive documentation, both on-line in a number of formats and in hardcopy.

The R Environment:

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

* an effective data handling and storage facility,
* a suite of operators for calculations on arrays, in particular matrices,
* a large, coherent, integrated collection of intermediate tools for data analysis,
* graphical facilities for data analysis and display either on-screen or on hardcopy, and
* a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

R Packages:

R can also be easily extended easily via packages to provide a wide array of functionalities apart from statistical systems. A core set of packages are included in the default R installation and additional packages are available in repositories such as Comprehensive R Archive Network (CRAN), Bioconductor and GitHub.

Data set

The dataset consists of house prices from King County and Seattle areas in the US State of Washington. This dataset consists of historical data of houses sold from May 2014 – May 2015. The dataset is obtained from Kaggle. This data is published under CC0: Public Domain license.

The dataset consists of 21 variables and 21597 observations. The presence of a ‘view’ variable which states the number of times a house/property has been viewed indicates that this data could have originated from a real estate agency or company. The values of the numeric variables do not appear random. The price variable is the target variable and there seems to be some dependence of the target variable on the independent variables.

Description of variables:

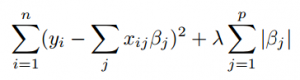
1. id - a notation for a house (numeric)
2. date - date house was sold (string)
3. price - price is the value at which the house was sold (target variable) (numeric)
4. bedrooms - number of bedrooms in the house (numeric)
5. bathrooms - number of bathrooms per bedroom (numeric)
6. sqftliving - square footage of the home (numeric)
7. sqftlot - square footage of the lot (numeric)
8. floors - total floors (levels) in the house (numeric)
9. waterfront - house which has a view to a waterfront (numeric)
10. view - number of times the house has been viewed (numeric)
11. condition - how good the condition of the house is: 1 indicates worn out property and 5 indicates excellent. (numeric)
12. grade - overall grade given to the housing unit, based on King County grading system: 1 indicates poor, 13 indicates excellent. (numeric)
13. sqftabove - square footage of house apart from basement (numeric)
14. sqftbasement - square footage of the basement (numeric)
15. yrbuilt - year the house was built (numeric)
16. yrrenovated - year when house was renovated (numeric)
17. zipcode – zip (numeric)
18. lat - latitude coordinate (numeric)
19. long - longitude coordinate (numeric)
20. sqftliving15 - living room area in 2015(implies some renovations) This might or might not have affected the lot size area (numeric)
21. sqftlot15 – lot Size area in 2015(implies some renovations) (numeric)

Lasso Regression

In this activity we use the Lasso regression method. Lasso stands for Least Absolute Shrinkage Selector Operator. It is a type of linear regression that uses ‘shrinkage’ where the values converge toward a central point such as the mean. Lasso regression is used to enhance the prediction accuracy and interpretability of the resulting statistical model.

Lasso regression uses L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models.

Lasso solutions are quadratic programming problems. The goal of the algorithm is to minimize:



which is the same as minimizing the sum of squares with constraint Σ | βj | ≤ s. Some of the βs are shrunk to exactly zero, resulting in a regression model that’s easier to interpret.

A tuning parameter, λ controls the strength of the L1 penalty. λ is basically the amount of shrinkage:

* When λ = 0, no parameters are eliminated. The estimate is equal to the one found with linear regression.
* As λ increases, more and more coefficients are set to zero and eliminated (theoretically, when λ = ∞, all coefficients are eliminated).
* As λ increases, bias increases.
* As λ decreases, variance increases.

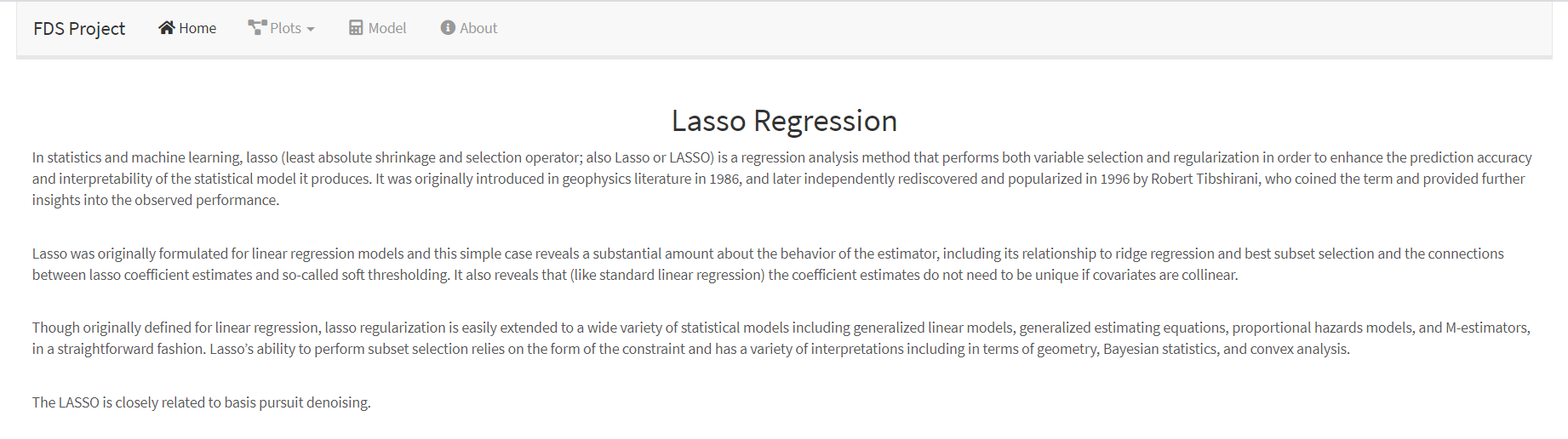
Result Obtained

The code of the program can be found here:

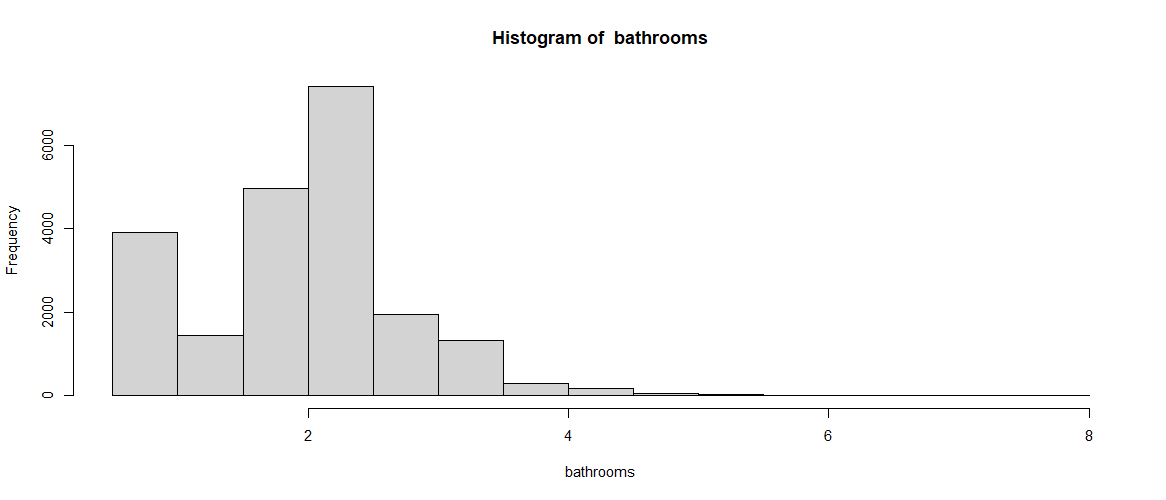
https://github.com/subramanya9112/Lasso-Regression-FDS-Project

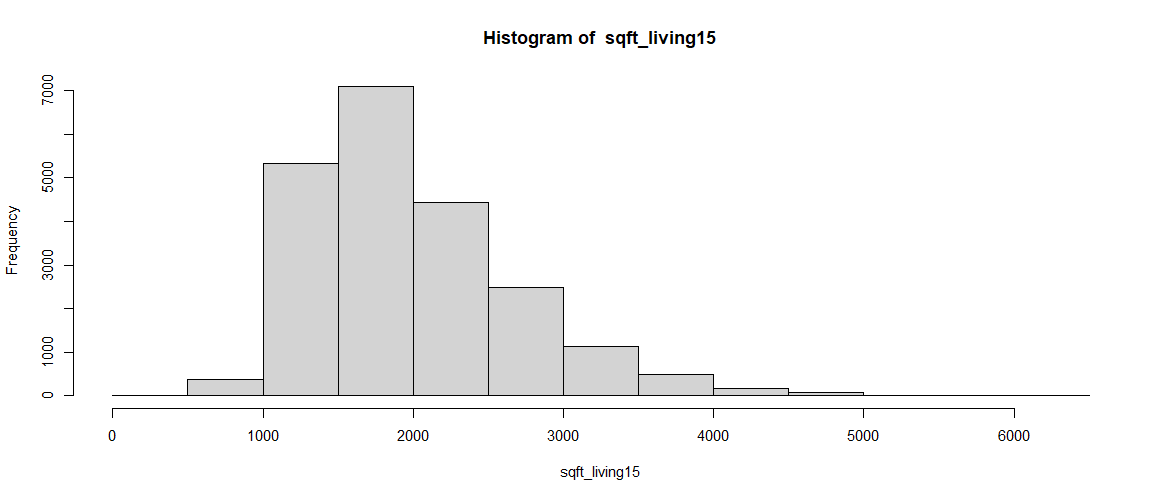
The model result are as follows:

|  |  |
| --- | --- |
| Statistical functions | Statistical functions value |
| MSE (Mean Squared Error) | 55112110193.352 |
| CVM (Mean Cross Validation Error) | 53212561788.435 |
| RMSE (Root Mean Squared Error) | 234759.686 |
| R2 (R squared) | 0.5916203 |
| Adjusted R2 (Adjusted R squared) | 1.000227 |

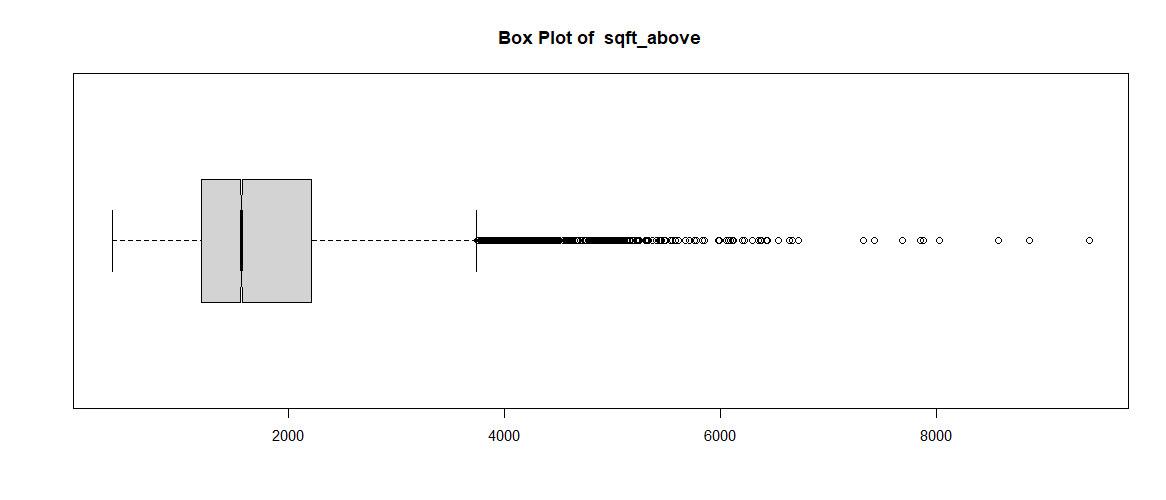


Histogram

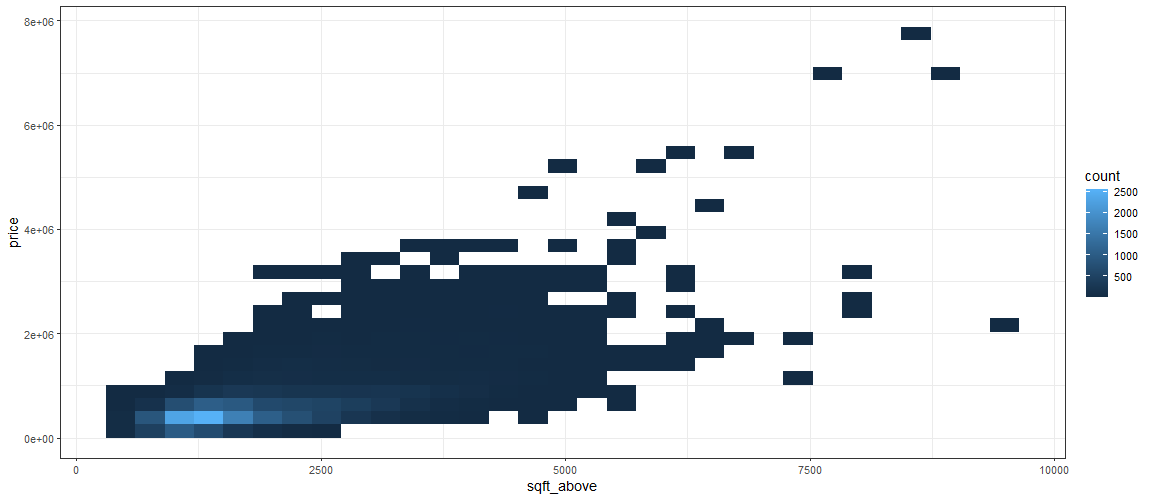


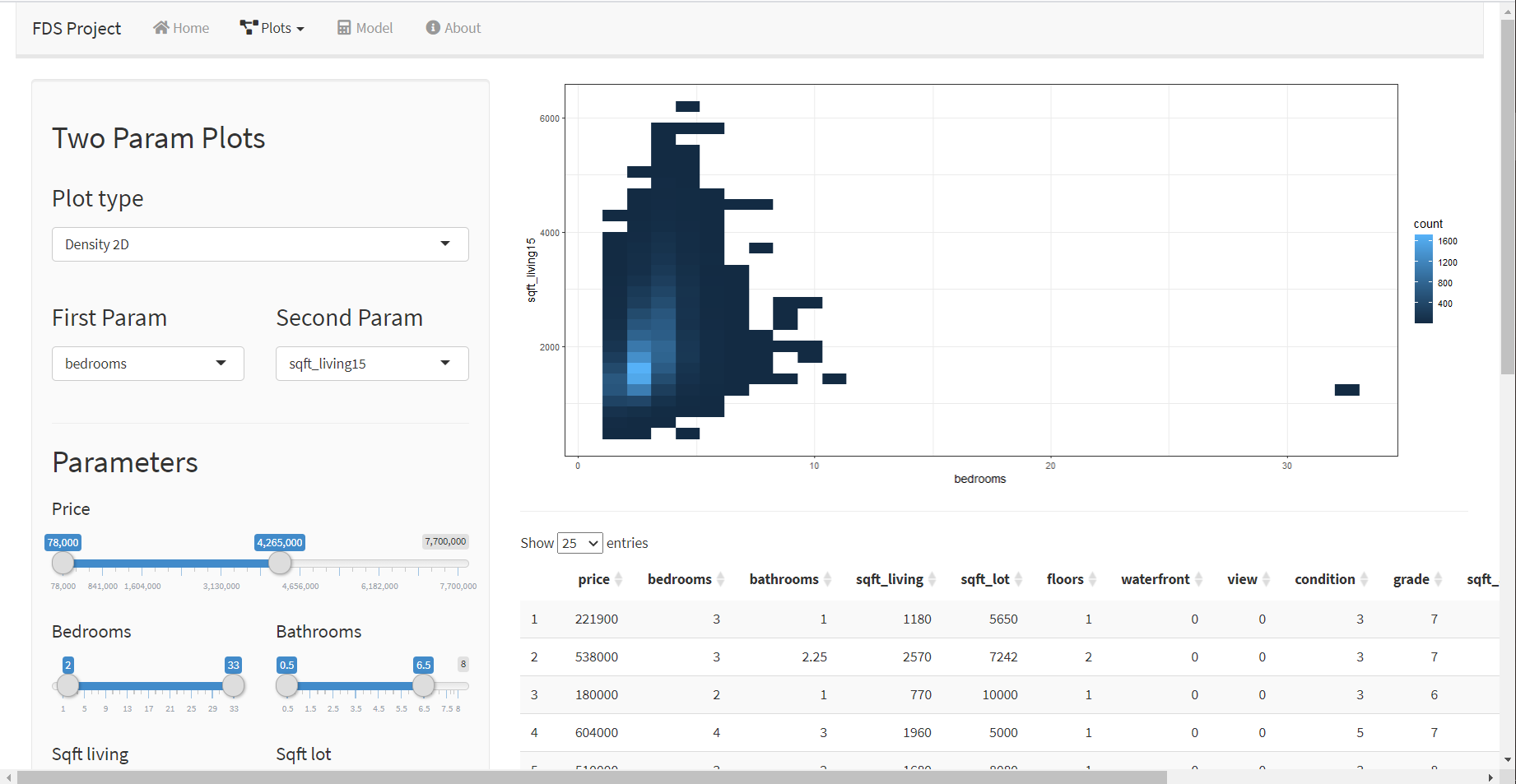


Box Plot

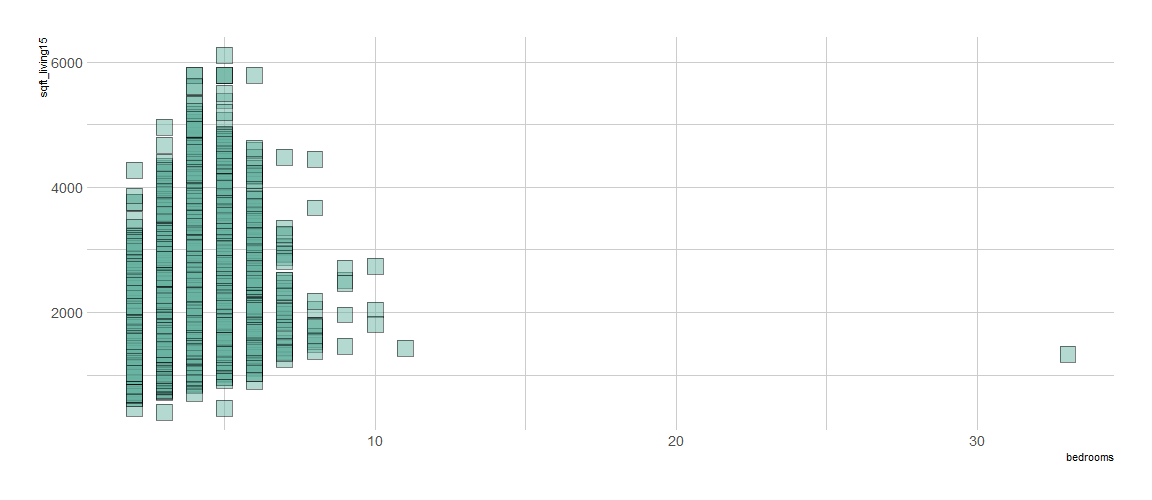


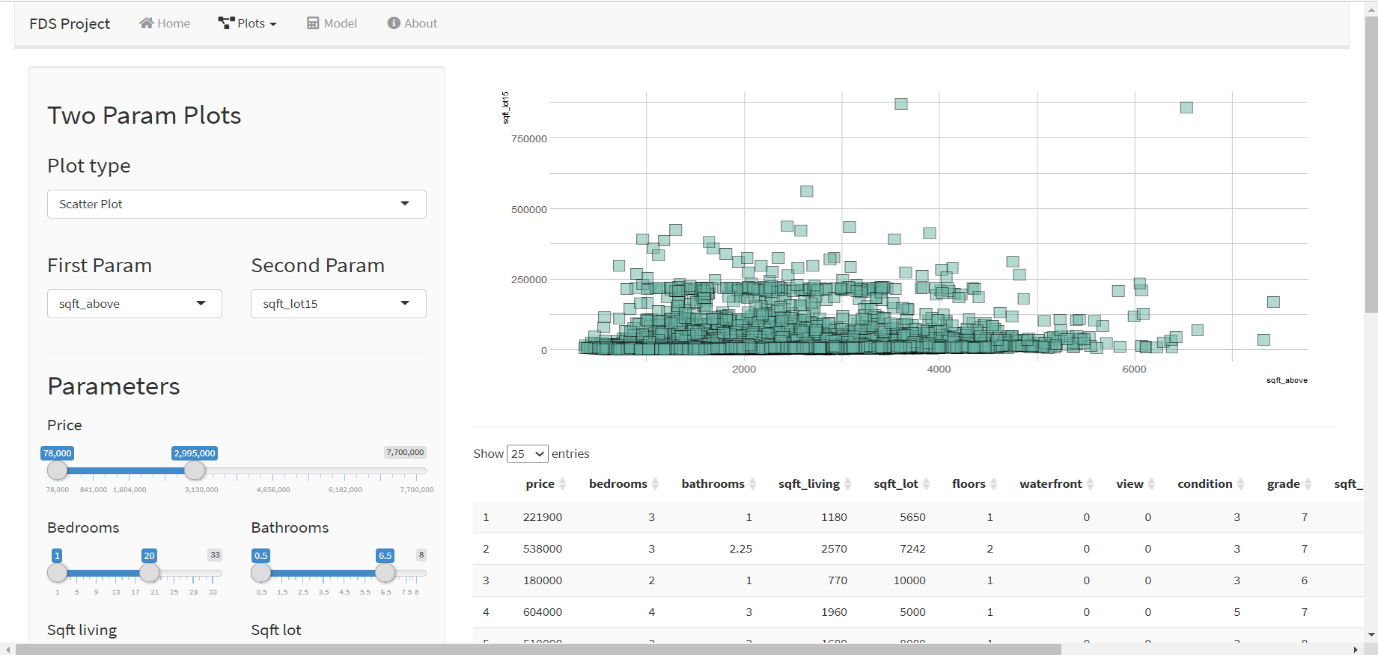
Density 2D

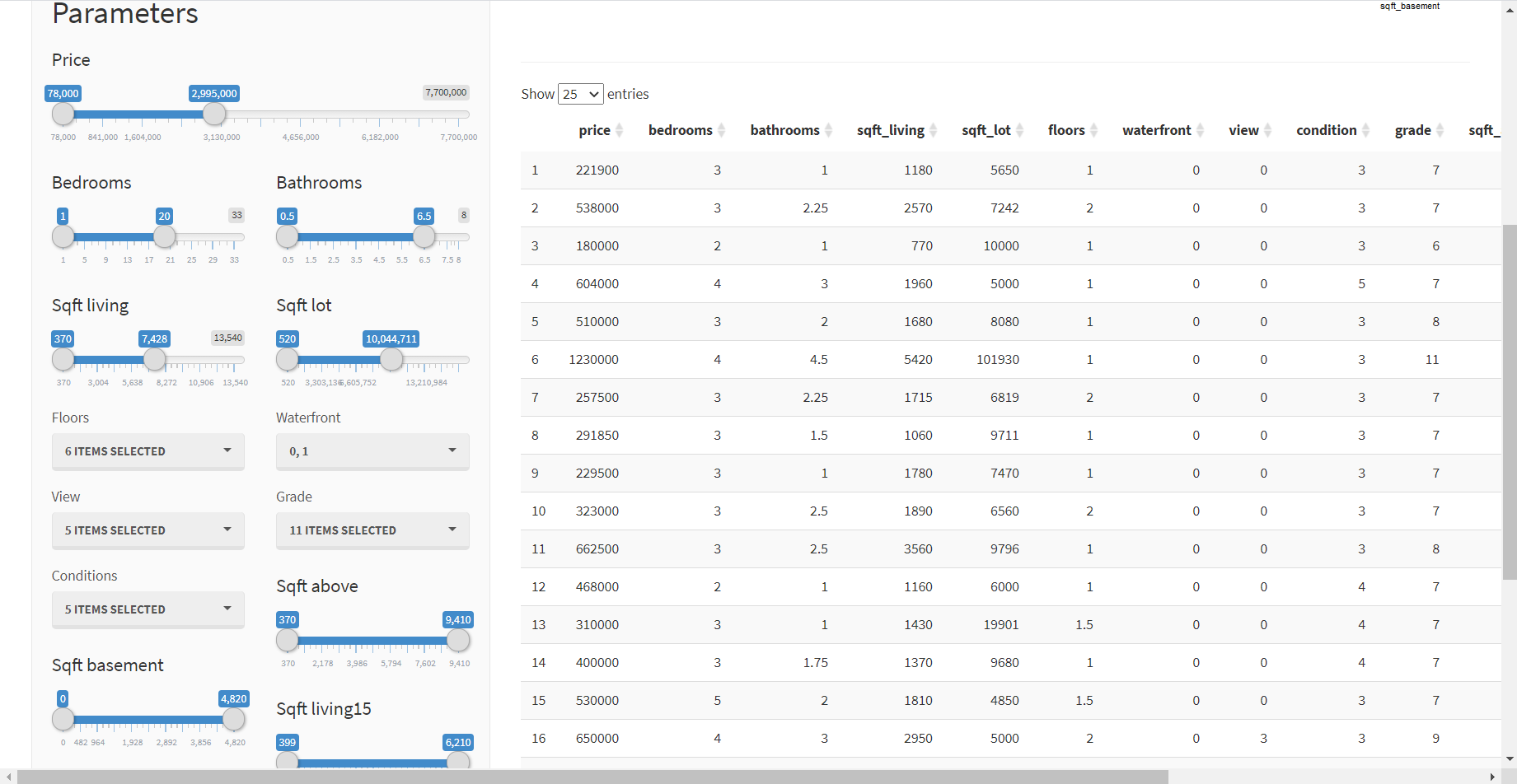




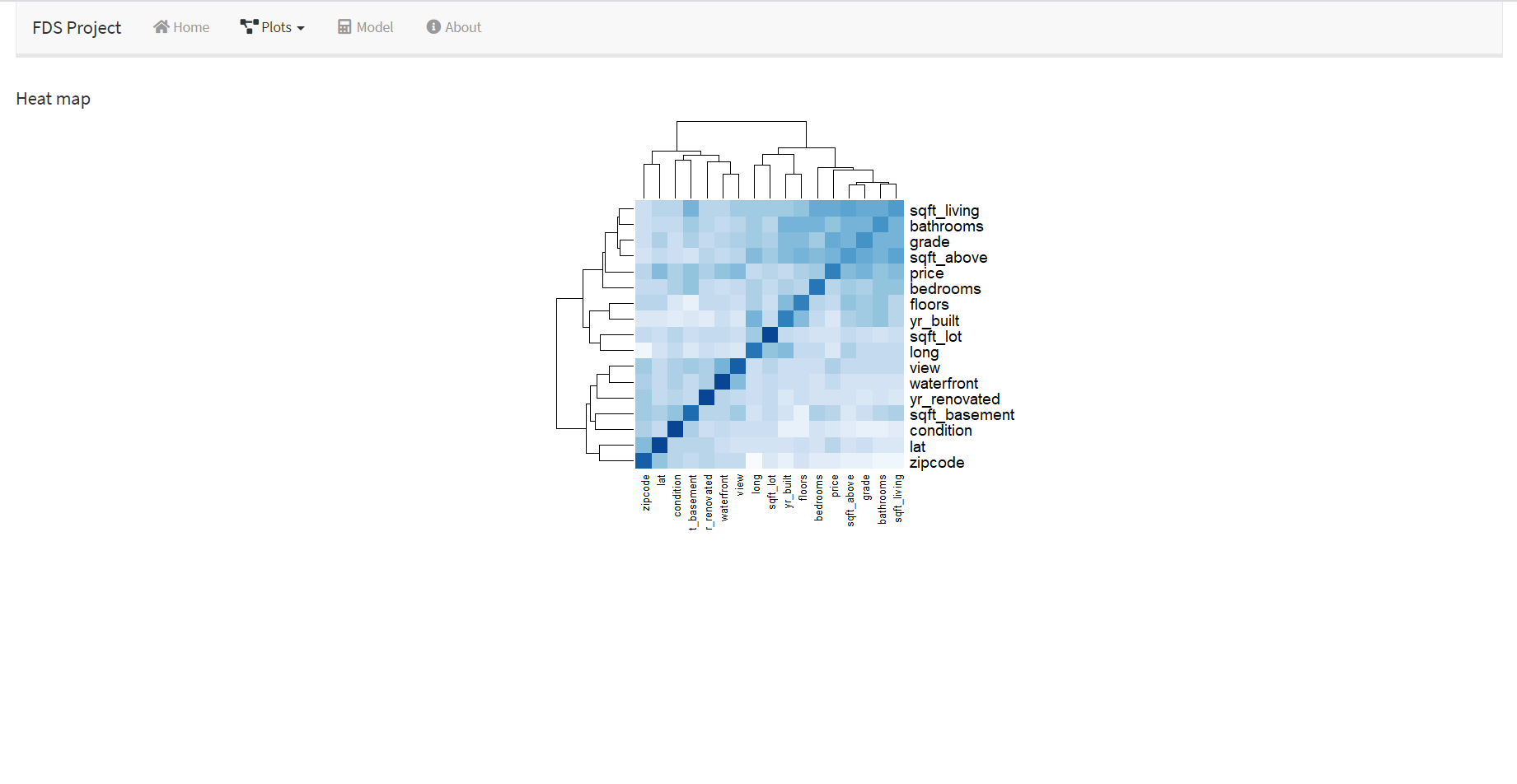
Scatter Plot



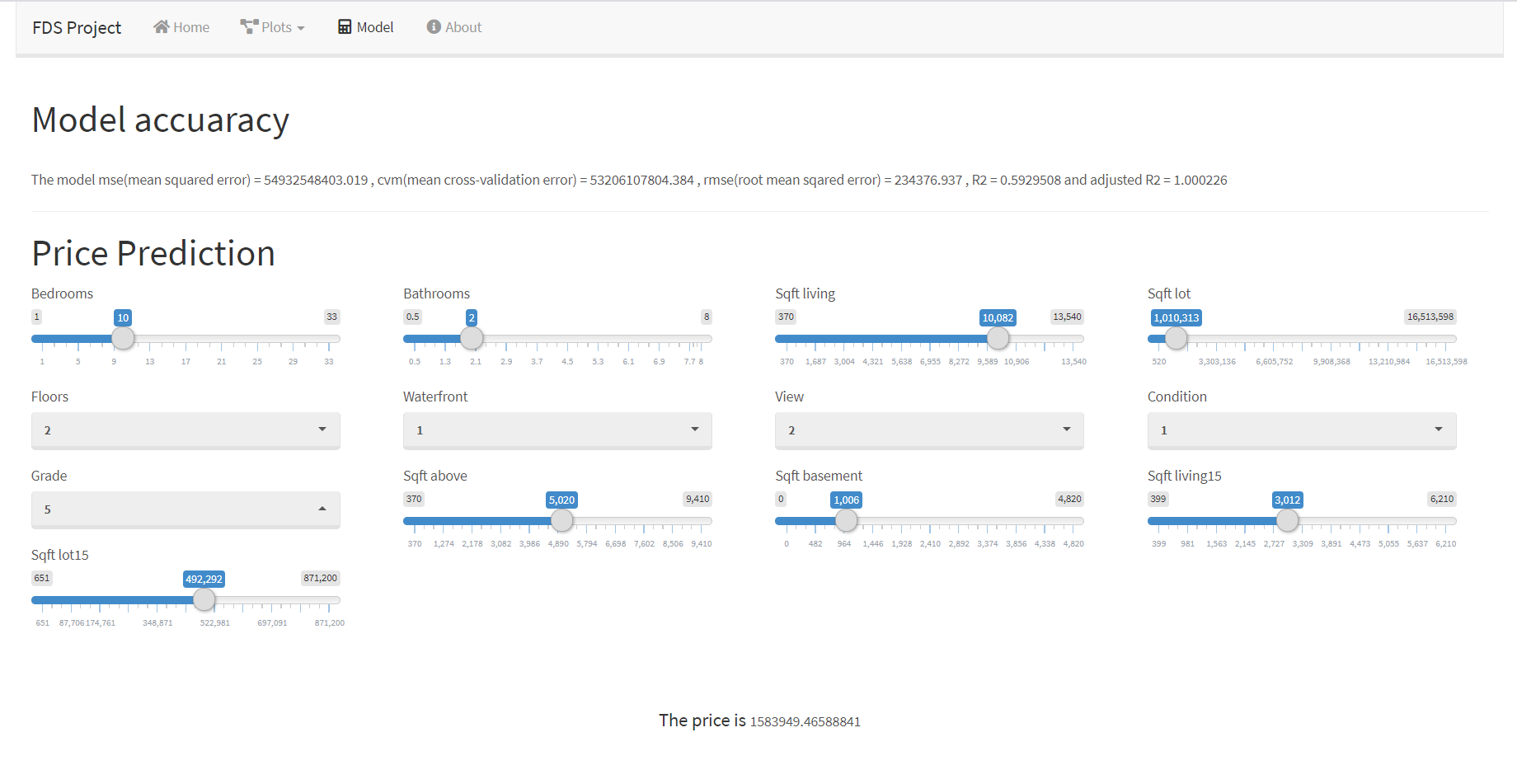




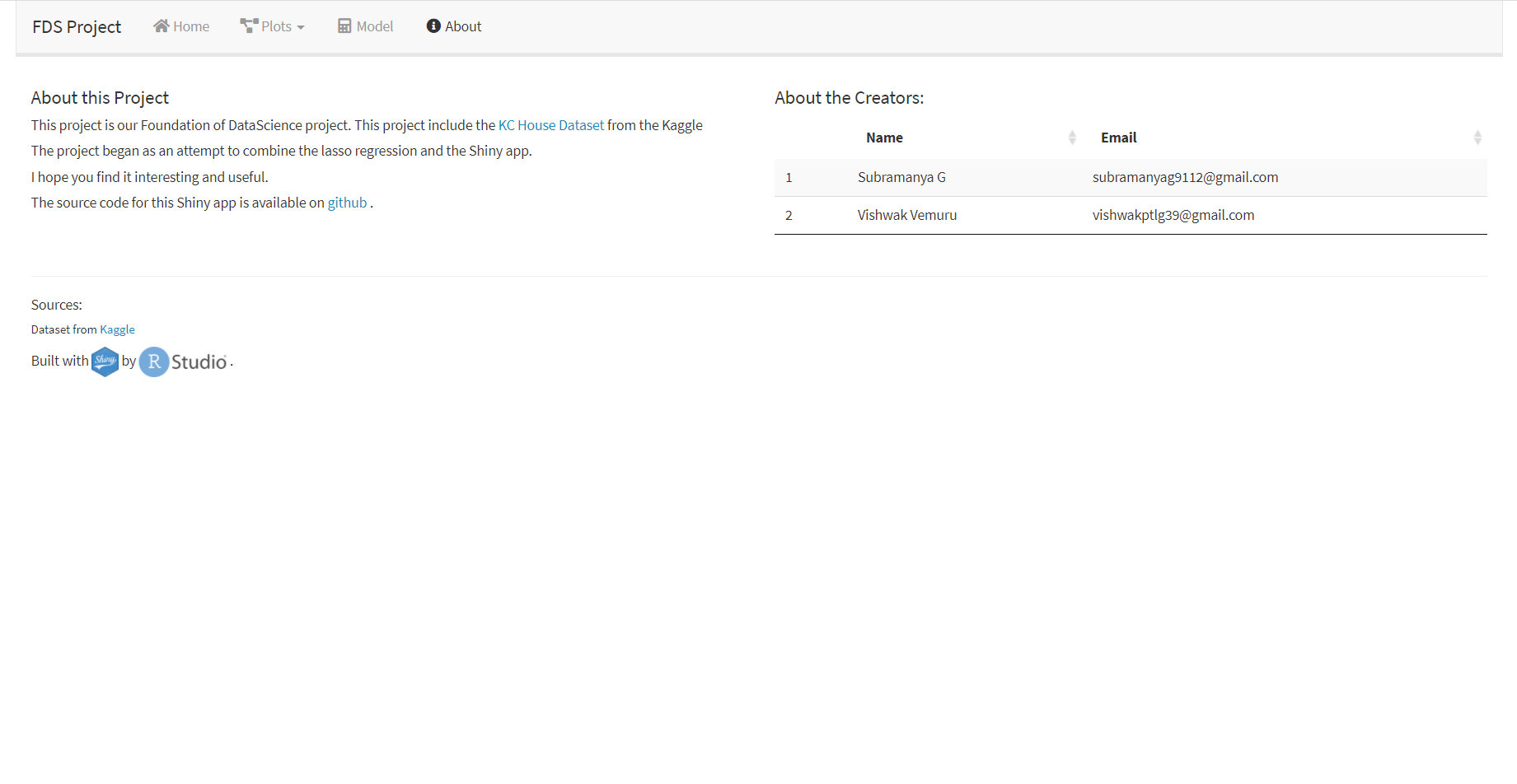
Heat Map



Model Predictor



About



Conclusion

The model was ran leaving 5 columns (id, date, yr\_built, yr\_renovated, zipcode, lat, long) and using all other 14 columns, which has a row of 21597. The model had a mean square error of 55 x 109, mean cross validation error of 53 x 109, root mean squared error 234 x 103, R2 score of 0.59162 and adjusted R2 of 1.000227. Based on the results we can conclude that only one type of accuracy function can’t be used, since based on dataset the value may vary.