King County House Sales Study

Flatiron School

Kamile Yagci 12/14/2021

Content

- Overview
- Business Problem
- Data
- Method
- Model
- Interpret
- Future Work
- Questions

Overview

- This project is the analysis of the King County House Sales.
- The Multiple Linear Regression is used to model the data and predict the house sale price.

Project Details:

- GitHub: https://github.com/kamileyagci/dsc-phase-2-project

Business Problem

The Windermere Real Estate Agency requested an analysis on House Sale Prices in King County. They will use the results of the study to advise their customers.

Questions:

1. What are the main predictors for House Sale Price?
2. Create a model to predict the House Sale Price.
3. Do house renovation affect the Sale Price?

Data

- King County House Sales Data: ‘kc_house_data.csv'
- The data contains 21597 houses sold between May 2014 - May 2015
- For each house, the data includes 21 types of information, such as
  - sale price,
  - number of bedrooms, number of bathrooms, number of floors,
  - footage of living, footage of lot,
  - waterfront status, condition, grade by county,
  - built year, renovation year,
  - zipcode

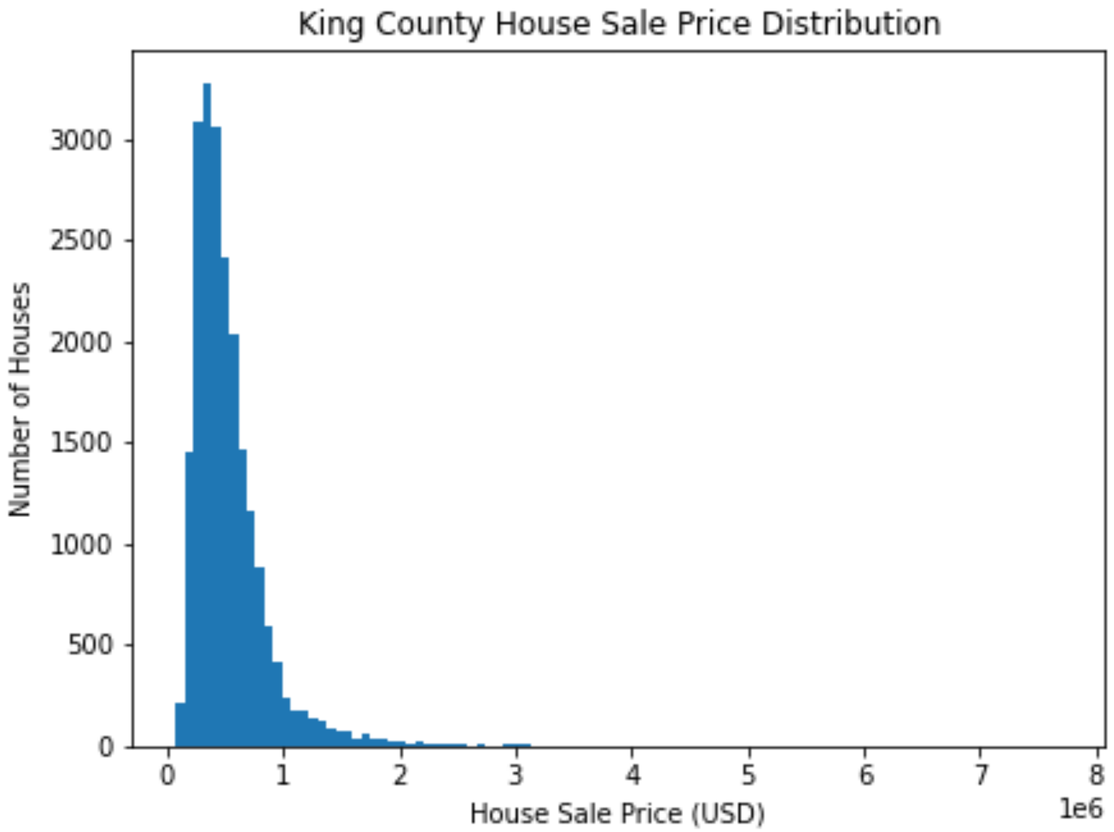
Method: Multiple Linear Regression

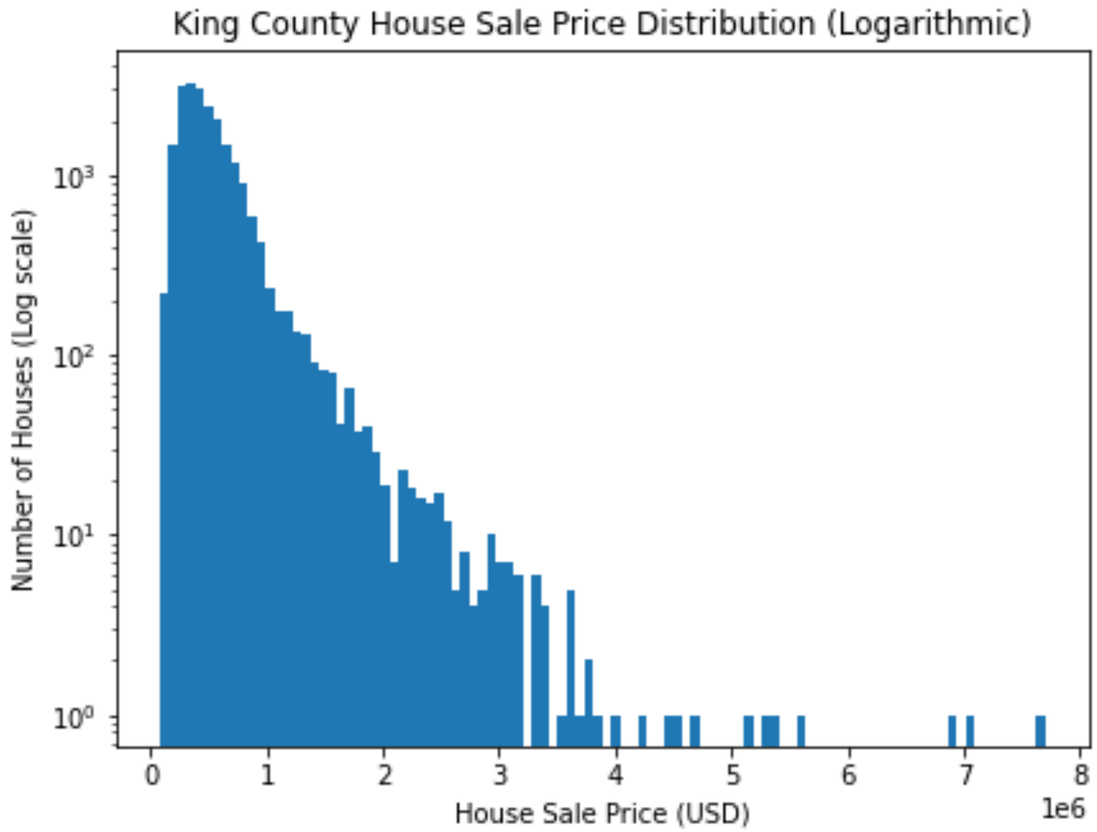
- Basic Linear Regression:
  - $y = mx + b$
  - where x is the independent variable, y is dependent variable, m is slope and b is y-intercept.
- Multiple Linear Regression:
  - multiple independent variables
  - $y = b + m1x1 + m2x2 + m3x3 + ...$
- When linear fit is applied, it determines the line-of-best-fit.
  - The slope and y-intercept of this line are the parameters of the model.
- R^2 = coefficient of determination
  - statistical measure of how close the data are to the fitted regression line
  - ranges from 0 to 1; large R^2 means good fit on data

Model - Steps

- Three steps in modeling:
  - Separate data into training and testing subsets
  - Model data by applying Linear Regression Fit on training data
  - Validate model on testing data
- In our model:
  - dependent variable is target variable, sale price (y)
  - independent variables are the predictors, all other variables (x1, x2, ...)

Sale Price Distribution



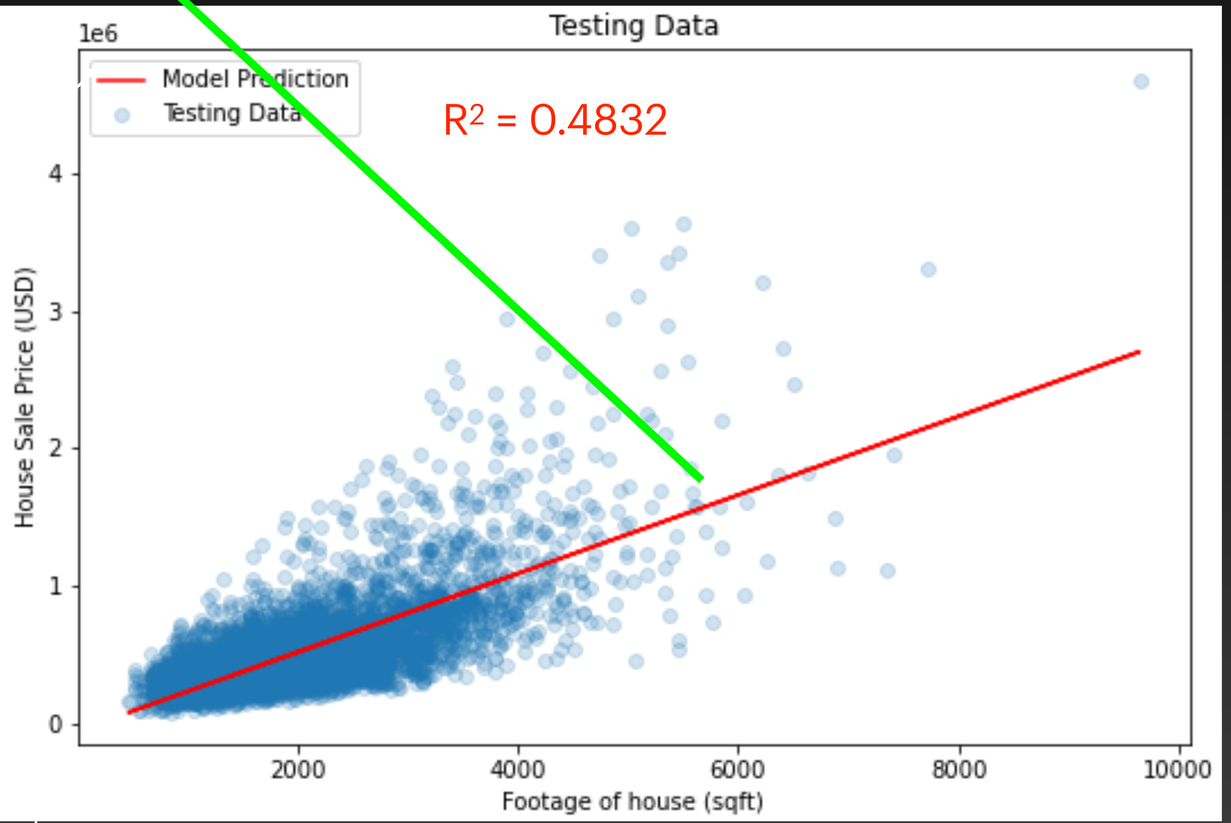


Model: Basic Linear Regression

- Target: sale price
- Predictor: footage of house

- Line-of-best-fit:
  - Slope: 285.58
  - y-intercept: -53321.49

- Based on R^2 score 48% of the data fit the regression model.

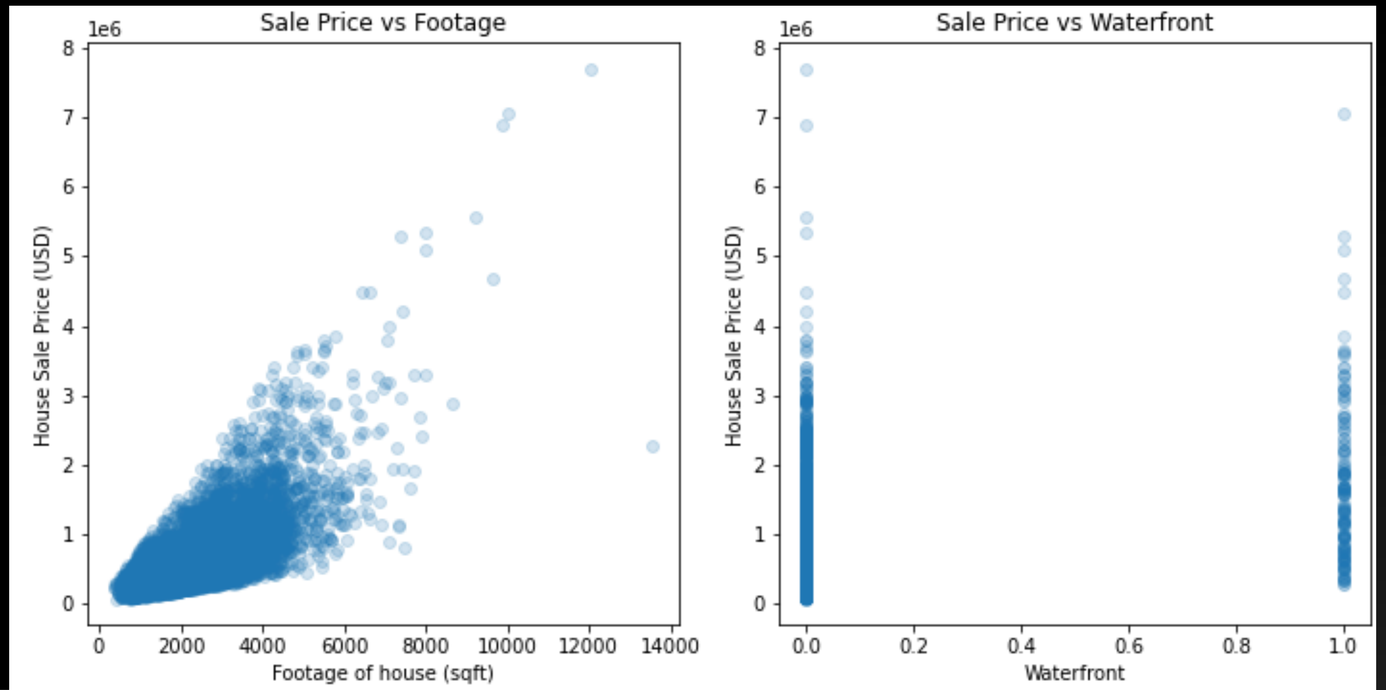


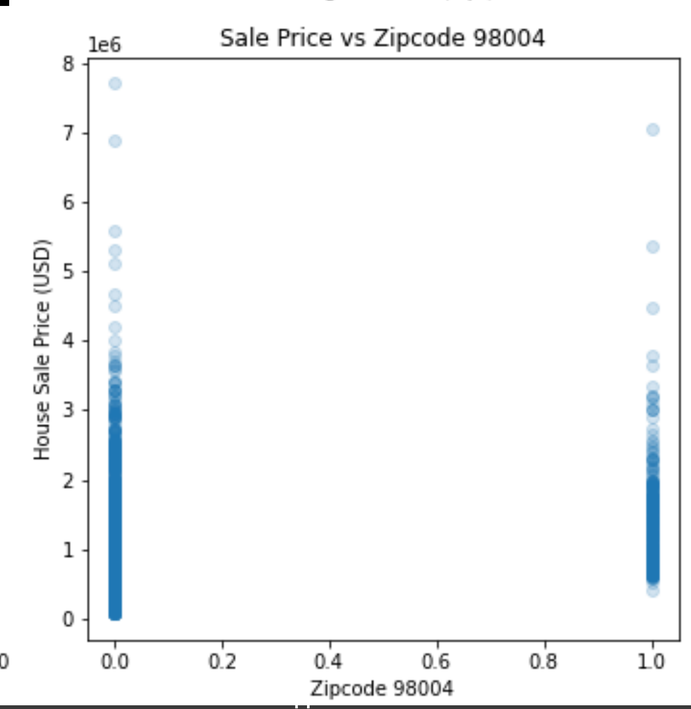
Model: Best Predictors

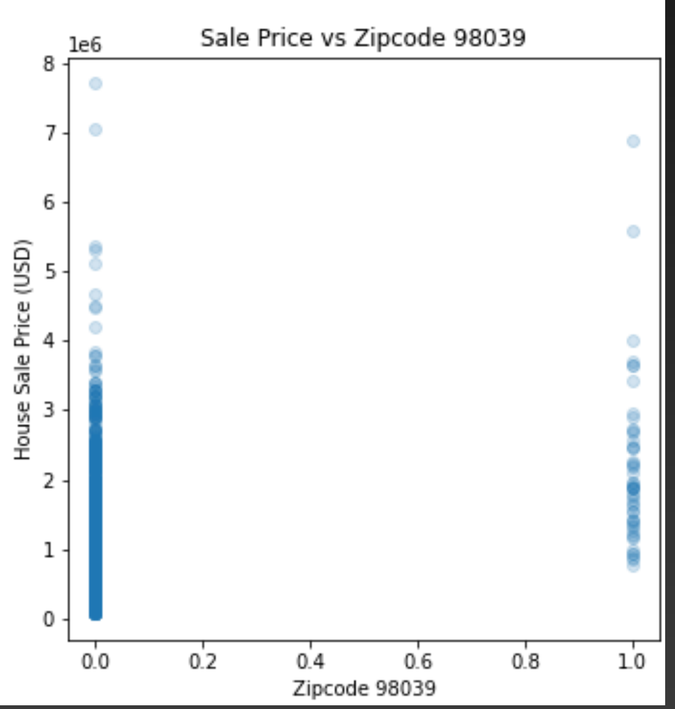
- Based on my analysis, best predictors for house sale price are:
  - footage of living space
  - watefront status: Is the house on waterfront? (yes/no)
  - build year
  - zipcode-98004: Is the house on this zipcode? (yes/no)
  - zipcode-98039: Is the house on this zipcode? (yes/no)

Scatter Graphs for Best Predictors

y: Sale Price
x: Predictors







Model: Multiple Linear Regression with best predictors

Results:

- Intercept: 4014990.7576159136
- Coefficients:
  - sqft_living: 2.844667e+02
  - waterfront: 8.040627e+05
  - yr_built: -2.071868e+03
  - zip_98004: 6.201505e+05
  - zip_98039: 1.162237e+06

- R^2 = 0.6159 (Training)
- R^2 = 0.6179 (Testing)
- Based on R^2 score, 62% of the data fit the regression model.

Interpret: House Sale Price Prediction

y = b + m1x1 + m2x2 + m3x3 + m4x4 + m4x5

House Sale Price = 4014990.7576159136 + (2.844667e+02 x sqft_living) + (8.040627e+05 x waterfront) + (-2.071868e+03 x yr_built) + (6.201505e+05 x zip_98004) + (1.162237e+06 x zip_98039)

House Sale price prediction examples:

1. sqft_living=2800, waterfront=0, year_built=1978, zipcode_98004=0, zipcode_98039=0 => price = $713,342
2. sqft_living=4200, waterfront=0, year_built=1978, zipcode_98004=0, zipcode_98039=0 => price = $1,111,595
3. sqft_living=2800, waterfront=1, year_built=1978, zipcode_98004=0, zipcode_98039=0 => price = $1,517,405
4. sqft_living=2800, waterfront=0, year_built=2015, zipcode_98004=0, zipcode_98039=0 => price = $636,683
5. sqft_living=2800, waterfront=0, year_built=1978, zipcode_98004=1, zipcode_98039=0 => price = $1,333,493
6. sqft_living=2800, waterfront=0, year_built=1978, zipcode_98004=0, zipcode_98039=1 => price = $1,875,579

(0=no, 1=yes)

Does renovation affects the house sale price?

Linear Regression Results:

- target: sale price
- predictor: renovation year
  - R^2 for Training: 0.0102
  - R^2 for Testing: 0.0061

House Renovation doesn't have significant effect on House Sale Price.

Future Work

- Study outliers
- Study correlation between footage of living and built year

Questions?