STATS 140 Final Project - Houses in King's County

By Jaspreet Kang, Tyler Rodriguez, Seungwoo Hong, Jingfan Lu, Junyan Zheng

Data Background

The dataset explored in this write-up is a housing dataset in Kings County, Washington which includes Seattle and the cities surrounding Seattle such as Federal Way, Burien, and more. It contains house sale prices sold between May 2014 and May 2015. Each row in the data set is a different house that was bought during our observational period. There are a total of 21,613 houses (or observations) and 21 variables.

Variable Explanations

- **Response variable :** Price the house was bought for
- **Square footage:** The measurement of area
- **Date:** The date the house was bought, formatted year/month/day
- **Bedrooms:** Number of bedrooms in the house
- **Bathrooms:** Number of bathrooms in the house, .5 implies no shower
- **sqft_living:** Square footage of the apartments interior living space
- **sqft_lot:** Square footage of the land space outside
- **floors:** Number of floors
- **waterfront:** A dummy variable for whether the house has a view of any body of water.
- **view:** An index from 0 to 4 of how good the view of the property was, how nice the surrounding area is
- **condition:** Overall condition of the house when bought.
- **grade:** An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.
- **sqft_above:** The square footage of the interior housing space above ground level
- **sqft_basement:** The square footage of the basement (0 means no basement)
- **yr_built:** The year the house was built
- **yr_renovated:** The year the house was last renovated
- **zipcode:** area zip code of the house
- **lat:** Latitude
- **long:** Longitude
- **sqft_living15:** The square footage of interior housing living space for the nearest 15 houses
- **sqft_lot15:** The square footage of the land lots of the nearest 15 houses

Newly Added Variables:

- **city:** Used the zip code and latitude/ longitude to tell us what city the house was located in
- **vintage:** Calculated age when the property is sold

Questionable Variables

- **Bedrooms:** Had a min of 0 and 3rd Quartile of 4, yet a max of 33 bedrooms which seems a little extreme
- **Sqft_basement:** Some observations have 0 and the median is 0, meaning most houses don't have a basement, meaning we should make a variable that says if a basement is added or not.
- **Year_built and Year Renovated:** Needed to combine because a old house that was renovated will be worth more than an old house that was not renovated.

Variable Summary

```
     price               bedrooms        bathrooms      sqft_living       sqft_lot          floors      
 Min.   :  75000   Min.   : 0.000   Min.   :0.000   Min.   :  290   Min.   :    520   Min.   :1.000  
 1st Qu.: 321950   1st Qu.: 3.000   1st Qu.:1.750   1st Qu.: 1427   1st Qu.:   5040   1st Qu.:1.000  
 Median : 450000   Median : 3.000   Median :2.250   Median : 1910   Median :   7618   Median :1.500  
 Mean   : 540088   Mean   : 3.371   Mean   :2.115   Mean   : 2080   Mean   :  15107   Mean   :1.494  
 3rd Qu.: 645000   3rd Qu.: 4.000   3rd Qu.:2.500   3rd Qu.: 2550   3rd Qu.:  10688   3rd Qu.:2.000  
 Max.   :7700000   Max.   :33.000   Max.   :8.000   Max.   :13540   Max.   :1651359   Max.   :3.500  
   waterfront            view          condition        grade          sqft_above   sqft_basement  
 Min.   :0.000000   Min.   :0.0000   Min.   :1.000   Min.   : 1.000   Min.   : 290   Min.   :   0.0  
 1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.: 7.000   1st Qu.:1190   1st Qu.:   0.0  
 Median :0.000000   Median :0.0000   Median :3.000   Median : 7.000   Median :1560   Median :   0.0  
 Mean   :0.007542   Mean   :0.2343   Mean   :3.409   Mean   : 7.657   Mean   :1788   Mean   : 291.5  
 3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.: 8.000   3rd Qu.:2210   3rd Qu.: 560.0  
 Max.   :1.000000   Max.   :4.0000   Max.   :5.000   Max.   :13.000   Max.   :9410   Max.   :4820.0  
    yr_built    sqft_living15    sqft_lot15     
 Min.   :1900   Min.   : 399   Min.   :   651  
 1st Qu.:1951   1st Qu.:1490   1st Qu.:  5100  
 Median :1975   Median :1840   Median :  7620  
 Mean   :1971   Mean   :1987   Mean   : 12768  
 3rd Qu.:1997   3rd Qu.:2360   3rd Qu.: 10083  
 Max.   :2015   Max.   :6210   Max.   :871200  
```

Research Question

Our goal is to study the different variables such as bedrooms, presence of a waterfront, views, conditions of the house, and more in order to assign the best fit model that predicts housing prices.

Problems Encountered

- The data is from 2014-2015, we need more up to date data in order to make more accurate predictions in the current market.

- Based on the linear model we set up, it seems to overestimate the price. In other words, the estimated price in the testing data are likely to be bigger than the actual prices in the testing data.

Teammate Contribution Interim

- **Tyler**: Helped a lot with the write up, summary statistics of variables, finding dataset, dataset description, defining the problems encountered
- **Seungwoo** : Prepared presentation slides, did data analysis with Python(Jupyter Notebook), set up linear model, and draw conclusion based on the result.
- **Jaspreet:** Did data analysis in R (created a model using linear regression, performed stepwise selection, checked model's assumptions, fixed any violations of assumptions), then compared results with Seungwoo who did his data analysis in Python. Helped with the write-up.
- **Junyan:** Helped with writing the paper such as description of data, helped with presentation slides, and looked over code.

Teammate Contribution Final Project

- **Tyler**: Help clean up the model, present, final report brush up.
- **Seungwoo:** Will focus on presentation especially how we leverage our raw data and draw a meaningful result.
- **Jaspreet:** Will take a deeper look at the relationships between all variables. Also curious about subsetting the data by cities, and creating a model to predict house prices within each city. Will the predictions be more accurate? How do house prices and other variables differ among the cities? These types of questions I will explore and try to answer.
- **Junyan:** Will focus more on making the data more accurate and finding new and efficient ways to do it.