# MS&E 226 Mini Project Part 1

*Samuel Hansen & Sarah Rosston*

*October 25, 2016*

## Summary

Our project investigates the correlates of house sale prices for King County, WA, the county that includes Seattle, obtained from homes sold between May 2014 and May 2015. This dataset contains a rich set of 21,613 observations with 19 features, including house sale price, zip code, GPS coordinates, square footage, and year built, among others.

The dataset comes from a Kaggle competition based on predicting housing sales. Although there is little information about the source of the data and it does not appear on King County's data website, the same dataset is used in a Coursera class taught by University of Washington professors who live in King County. Although there is very little information about the data collection process, because it is used in a popular MOOC on Coursera we believe the data is mostly complete and accurate, although because there are no rows with NA values, incomplete rows may have simply been removed from the data.

This report summarizes our progress so far, and provides information regarding: (1) data cleaning steps; (2) possible response variables; (3) possible research questions; (4) data summary steps; and (5) current interesting findings.

## Data Cleaning

### Missing values:

The data do not contain any `NA` values, so there is no need for missing value imputation.

### Variable recoding:

We recoded some variables such as `waterfront` (which is an indicator of whether the home has a waterfront view) from 0-1 numerics to categorical factors.

Although our data set includes house IDs, we want to include them because 176 houses appear twice due to relisting. Because the data set was cleaned in advance, we will include all variables in our initial analysis.

### Feature Engineering:

Upon inspection of the data, we engineered the following features from raw values:

1) Years since renovation: `renovation_year - year_built`
2) House age at time of sale: `sale_year - year_built`
3) Season of sale: Fall (9 <= `sale_month` <= 12), Winter (1 <= `sale_month` <= 4), etc.
4) Yard size: `sqft_lot - sqft_living`
5) Price per square foot: `price/sqft_living`
6) Ratio of house size to sizes of 15 neighboring houses: `sqft_living/sqft_living15`

## Response variables

**Continuous response variables**:

1) We are interested in the economic question of how the features of a home (i.e. square footage, location, room number) determine its price. Thus, we propose to use `house price` as the response variable.

2) As a corollary, we could examine `price per square foot`, which is an engineered feature from the raw data. Doing so would normalize differences in lot sizes when comparing house prices, thus measuring the effect of
location, age, and room features on home price more directly.

**Binary response variables**:

1) Price <= $450,000, which is the median value.

2) Price per square foot <= $245, which is the median value.

These response variables allow us to continue to examine the relationship between the inputs and housing price using a dichotomous cutoff at the median. We chose the median to allow for an even split between the two classes.

## Research Questions

Housing costs are increasingly an issue in Seattle, a city included in our dataset. Increased demand and higher wages drive up prices in the whole market, but which factors in our dataset increase the price of a house; How much value does an extra bedroom add? Does the ratio of living space to property size matter? Or is total land more important? The data also includes latitudes and longitudes for each house, which we can use to to look between more and less desirable neighborhoods and school districts to understand how much people are willing to pay to live in more desirable neighborhoods and send their children to better schools. Because we also have information about whether a house was renovated, we can also investigate how home renovations affect pricing.

## Data Summary

### Pairwise Correlations

After creating a pairwise scatterplot matrix, we found the following relationships:

1) `Price` is most strongly correlated with `square footage`, `number of bathrooms`, and `number of bedrooms`, which is consistent with intuition, since pricier homes are larger, and have more accommodations.

2) `Square footage` is strongly negatively correlated with `home age`, implying older homes are smaller, whereas newer homes are larger. This relationship may also be due to people rebuilding bigger houses more often.

3) There is an associative relationship between `square footage`, `number of bedrooms`, and `number of bathrooms`, which makes sense because larger homes tend to have more bedrooms, which necessitates more bathrooms.

4) Lastly, we find it surprising that `price` and `age` are so weakly correlated because we expect newer homes to sell for higher prices.

### Variables to Add

Because we have zip codes for each home sold, we can join outside data to our current data frame. For instance, we could join in monthly weather data, crime statistics, per capita income, ethnic demographics by zip code, and other neighborhood information, which may enhance inference and prediction.

### Possible Population Models

The price of a house likely changes based on the number of bedrooms, bathrooms, the amount of land, neighborhood features and the age of the house. Location seems likely to interact with all other variables. For instance, a room in a city is more expensive than one in a less densely populated area. Thus, a possible population model may be:

$Price \sim Location : Sq.Ft + Location : Bedrooms + Location : Bathrooms + Age : Condition + \epsilon$

There also could be interactions between the number of bedrooms and bathrooms, especially if an additional bathroom means each bedroom has its own bathroom. In this case the model would look like:

$Price \sim Location : Sq.Ft + Location : Bedrooms : Bathrooms + Age : Condition + \epsilon$

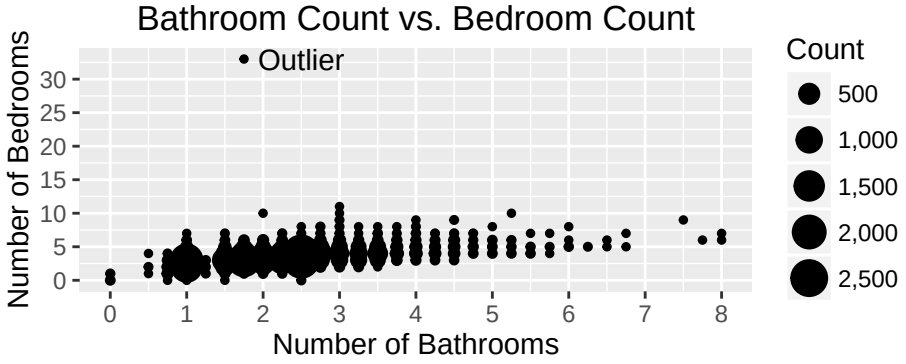
Location-related factors that could impact price may also include schools, crime rate, distance to public transit, and proximity to major companies like Microsoft and Amazon.

## Interesting Findings

Although there are many nuances to our data set, we present three findings that yielded the most interesting insights.

### 2 Bathrooms, 33 Bedrooms

A closer examination of the data revealed some surprising outliers that suggest data quality issues. For instance, after plotting the number of bathrooms against the number of bedrooms of listed houses, we found that there is a house with less than 2 bathrooms for 33 bedrooms (Note: bathrooms can be 0.25 for quarter bathrooms or 0.5 for half-bathrooms). Further, this plot reveals some houses do not have any bathrooms or bedrooms, which seems improbable. We intend to remove such outliers observations from our modeling analyses.




### House Reselling

We found 176 homes that appeared multiple times in the data, suggesting they were sold more than once between 2014-2015. We examined the relationship between the number of days it takes to resell a home and

the difference in price between its first and second listings. We plotted this relationship and found resale profit increases to ~$150,000 if a re-seller waits about 2 months. However, the average resale profit plateaus around $150,000 from 2 months onwards, implying the number of days to resell a home has little influence on the resale profit after 2 months have passed.

**Where are houses sold?**

Because our data includes the GPS coordinates of each house, we sought to find clusters of where houses are sold within King County. The map in Fig. 1 reveals that most of the houses are clustered within Seattle, particularly in urban neighborhoods around the Space Needle and the University of Washington (see the 1386 and 1957 clusters). The numbers on the colored circles represent the count of houses within the surrounding region. We also observe that far fewer houses are located in distant rural regions; hence, we may be able to engineer features reflecting how rural or urban a house's neighborhood is.

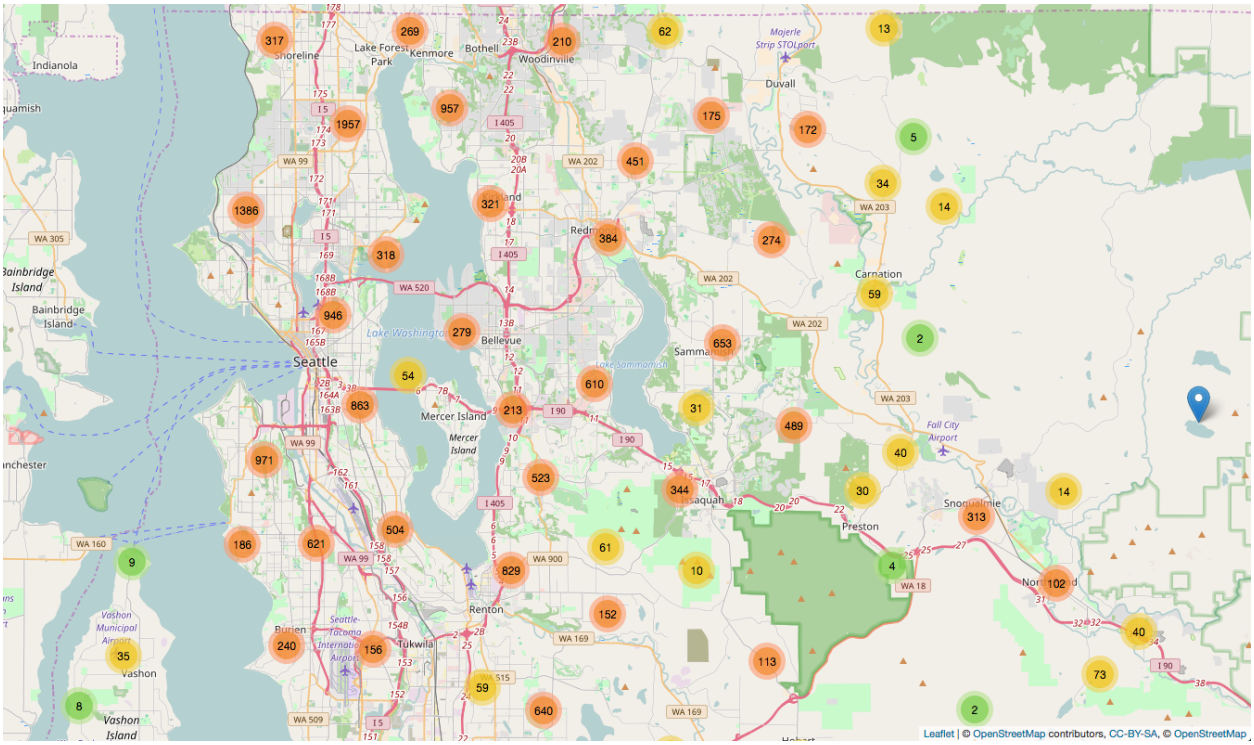



Figure 1: Map of house locations in King County