Data Science

Module 1 Project

Chelsea Power

King County Housing, USA (2015)

What factors influence the price of a house in this area?

What factors have little to no influence on the price of a house in this area?

What is the cost of housing based on the year the house was built?

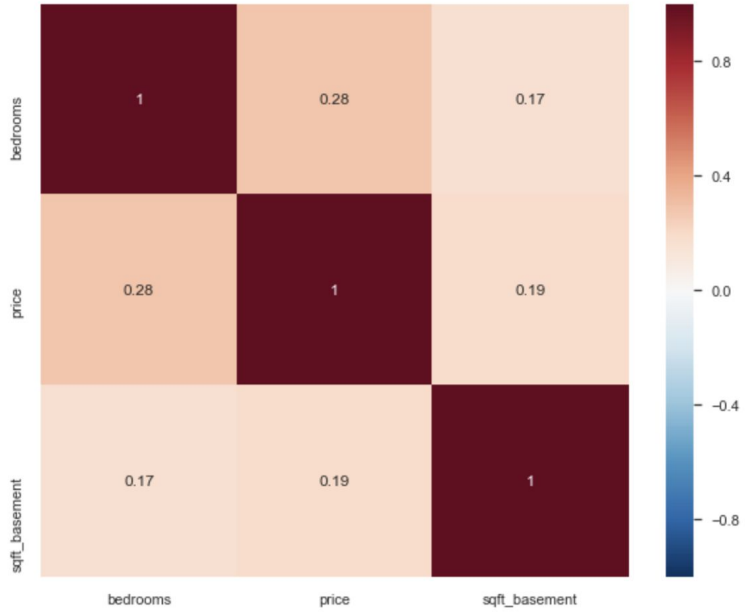
Methodology: OSEMiN

- **Phase 1 - Obtain**: Obtained and loaded the 2015 data from the `kc_house_data.csv` file
- **Phase 2 - Scrub**: Scrubbed data and conducted feature engineering by removing null/missing values and reset the types of many variables in the dataset
- **Phase 3 - Explore**: Generated graphical views (heatmaps and histograms) of the data to check for relationships between dependent variables (multicollinearity), removed outliers that were skewing the data (reduced the range of information) and then normalized (log transformed) the appropriate variables
- **Phase 4 - Model**: Generated a number of tests (OLS model to check r-squared and p-values), compared Price to categorical variables, removed features that were not good predictors, etc.
- **Phase 5 - iNterpret:** Determined the most influential variables to the housing price.

Model Results #1

After several rounds of modeling, feature validation and elimination the following heatmap shows the **most influential variables** on the housing **price**:

- **Sqft_living -** square footage of the interior living space
- **sqft_living15** - square footage of interior living space for the nearest 15 neighbors
- **bedrooms** - number of bedrooms
- **sqft_basement** - size of basement



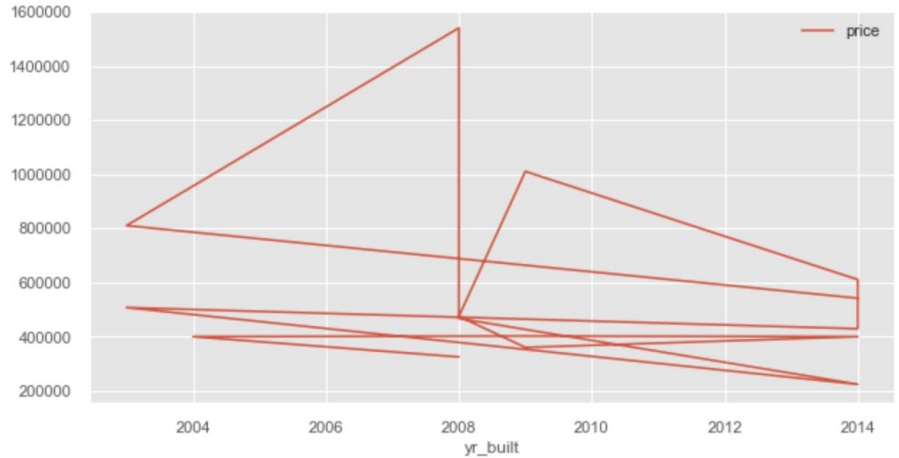
Model Results #2

The **9 least influential variables** to the housing **prices** are:

- **bathrooms** - number of bathrooms
- **sqft_lot** - square footage of the land space
- **floors** - number of floors
- **waterfront** - waterfront property category
- **view** - view of a property category
- **condition** - condition of the property category
- **grade** - grade of the property category
- **sqft_above** - square footage of the interior housing space that is above ground level
- **sqft_lot15** - square footage of the land lots of the nearest 15 neighbors

Model Results #3

Here is a graph housing prices against the year it was built



Conclusion

The influential variables were selected because they have the most direct impact on price and do not have an influential relationship to each other.

The non-influential variables were selected because they have influential relationships to each other and/or do not have a strong/direct impact on price.

Additional analysis can be conducted based on grouping of the following variables: the year the house was built, year the house was renovated, zip code, latitude and longitude.