King County, WA

Prediction of house sale prices based on exploration of historical data

07/30/19 Presentation

A. Garcia

Objectives

- Keeping consumers (sellers) in mind
  - Identify variables that may have an effect on sales price
    - » What are the prices like in the area?
    - » Does the year my home is built affect the price?
  - Address some questions for attributes (variables) over which our sellers have an effect on to increase the sales of their homes
    - » Is there a way to maximize my investment?
- Selecting a model that can predict with a good level* of confidence the home pricing.

* the definition of success level I based on the accuracy of the results to fit regression, and measured with an R-square of 0.75 (good) or higher.

Background

- The dataset in this project contains historical data from 1990 to 2015 of house sale prices for King County, Washington (which includes Seattle). The Goal of this analysis is to predict the price of housing based on the variables provided in the dataset.

- The OSEMN workflow is used to conduct the analysis :
  - Obtain : Gather data and obtain the overview of the type of information we will be working with
  - Scrub : Pre-processing of the data by cleaning the data into formats that can be recognized and the models we will use
  - Explore : Find the significant patterns using visualization and statistical methods
  - Model : Select and implement a model to predict and forecast
  - Interpret : Draw conclusion and put the results to good use

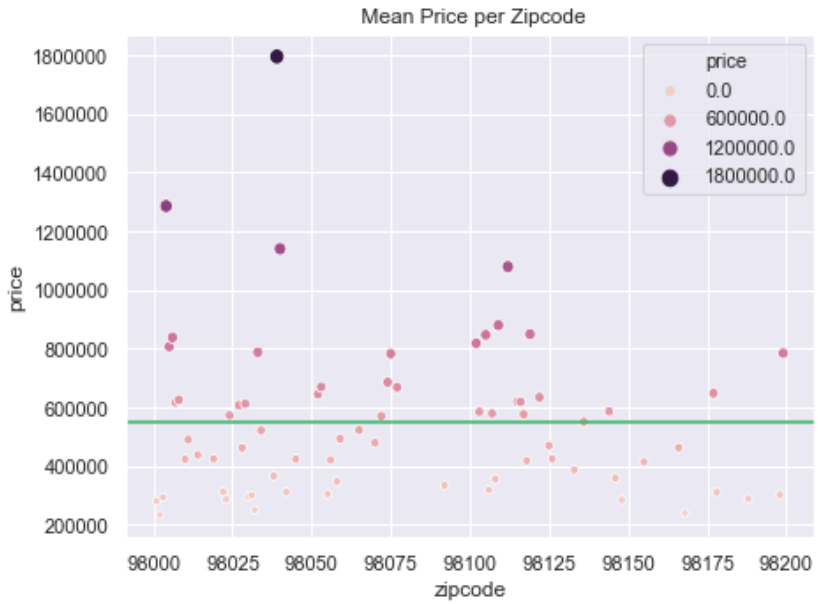
About the dataset

- The dataset has only 18 features: below a brief description of the interpretation

Variable	Description
Id	Unique ID for each home sold
Date	Date of the home sale
Price	Price of each home sold
Bedrooms	Number of bedrooms
Bathrooms	Number of bathrooms, where .5 accounts for a room with a toilet but no shower
Sqft_living	Square footage of the apartments interior living space
Sqft_lot	Square footage of the land space
Floors	Number of floors
Waterfront	A dummy variable for whether the apartment was overlooking the waterfront or not
View	An index from 0 to 4 of how good the view of the property was
Condition	An index from 1 to 5 on the condition of the apartment,
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 that is above ground level
Sqft_basement	The square footage of the interior housing space that is below ground level
Yr_built	The year the house was initially built
Yr_renovated	The year of the house's last renovation
Zipcode	What zipcode area the house is in
Lat	Lattitude
Long	Longitude
Sqft_living15	The square footage of interior housing living space for the nearest 15 neighbors
Sqft_lot15	The square footage of the land lots of the nearest 15 neighbors

What are some factors (fixed) affecting my home's price?

Location



A few zipcodes have greater than average concentration of prices

Living square footage

price is homogeneous between 1000 and 2000 sqft of livable space however after 2000 sqft we have more ranges in prices – from lower 200's well into the millions

Is age important?

Not so much the age as the specific period of 1940 and 1980 – houses tend to have a lower value for that period (could be related to materials or construction norms at the time).

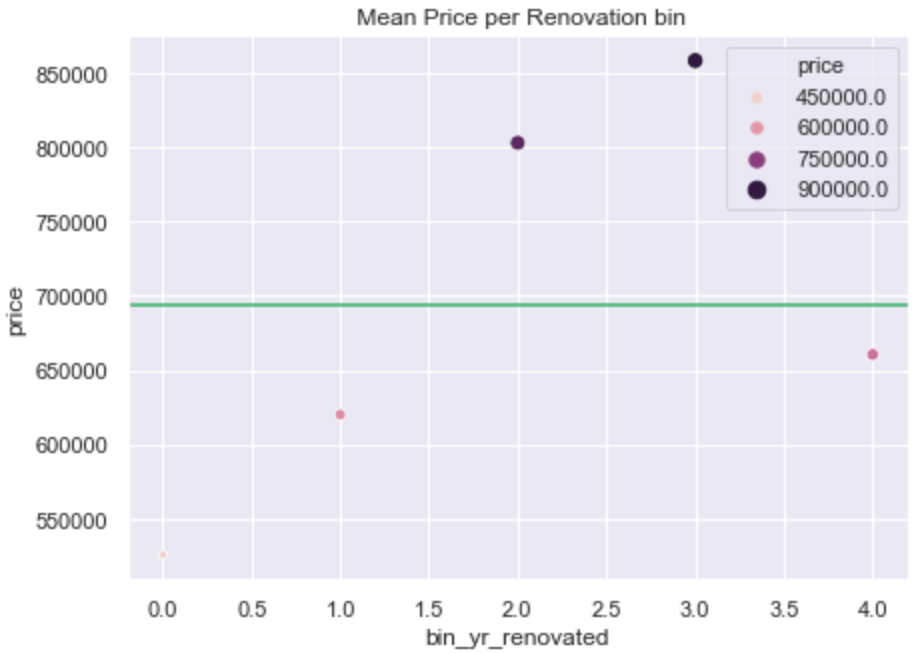
What can I do to increase the value of my home?

Renovations **may** be an option:

According to our analysis, the average price changes whether is renovated or not and the period:

Renovation	Average price
No (or data not available)	$ 525,593
prior to 1990	$ 620,098
between 1990 - 2000	$ 803,224
between 2000 - 2010	$ 858,602
after 2010	$ 660,548

Disclosure : the cost of the renovations is not available in the data and it is necessary to determine the return of the investment after the renovation.



Interpretation of the bins:
0 – Not renovated (or no data available)
1 – Renovation prior to 1990
2 – Renovation between 1990 - 2000
3 – Renovation between 2000 - 2010
4 – Renovation after 2010

How trustworthy is this analysis?

Based on the available data and with the selected model we can explain approximately variability in

84.2 %

of the time

Observations:

1) Our model can be optimized and further refined,

2) Additional (not available) features may also affect the price such as vicinity to green areas, parks, conveniences or landmarks, schools, crime, etc.

Thanks