2014-2015 King County Home Sales

Multiple Linear Regression Model for home sales price prediction

What factors affects Sales Prices?

In this project, we will analyze Sales data for homes in King County, Oregon (Seattle area).

In doing so, we will implement a Linear Regression Model as well as Multiple Regression, and train and test our models on the dataset.

Precisely, we will clean, explore, and model this dataset with a multivariate linear regression to predict the sale price of houses as accurately as possible.

1. Intro

Choose one approach we will keep things simple, and use sq footage data to predict sales prices using Linear Regression.

- ➔ **sqft_living**
  Let's explore what this feature's data tells us.
- ➔ **bedrooms**
  Usually bedrooms will be higher with larger sq footage.
- ➔ **bathrooms**
  This feature may play a key role in the sale price of a home

How will we account for the home's location??

Clearly real estate prices vary based on location.

Tip

Though we are given latitude and longitude, and zip code info, **wrangling** this data for our purposes would be too time consuming. Instead we'll engineer a new feature which takes the location into account

Just one! Custom fit.

The goal is to analyze the data, take into account all of its features, and engineer a feature such that most if not all of the characteristics that affect sales price can be accounted for. We can then use that feature in our Regression model.

Tip

Multiple Regression will be useful here, as we can measure the performance of our engineered features

sqft_living

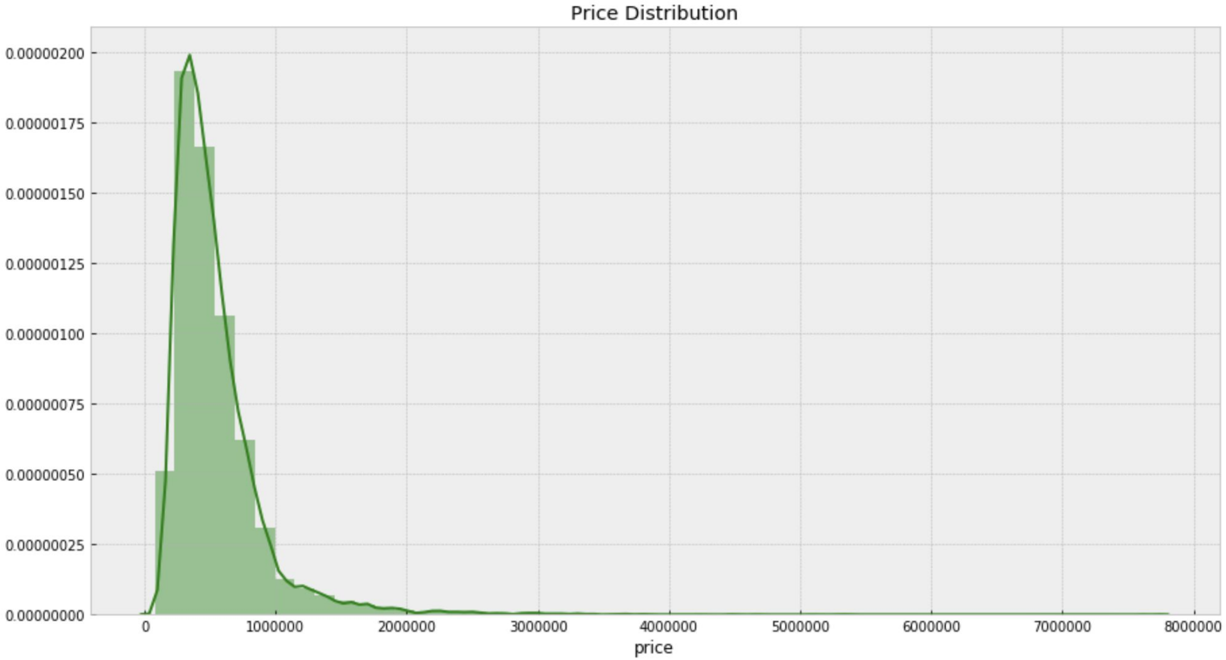
Is most closely correlated with Sales Price, meaning 'price' changes as 'sqft_living' changes. Check out top correlation values

price	1.000000
sqft_living	0.701917
grade	0.667951
sqft_above	0.605368
sqft_living15	0.585241
bathrooms	0.525906
view	0.395734
bedrooms	0.308787

Price Distribution

Tip

We don't need to normalize this feature for our analysis as we believe the outliers to be **true**, as in linear representations of the correlation of our data



After careful consideration, we decided that a **price per sq foot of living space** feature would be the best idea for our Regression Model

Tip

A price per sq foot of living space would indirectly take into account the home's location (as related to price), and give us the best data for our prediction model

price_per_sqft_living

count	21597.000000
mean	264.143331
std	110.000058
min	87.590000
25%	182.290000
50%	244.640000
75%	318.330000
max	810.140000

Tip

We see here that 1 sq foot of living space has a value that can range from $87.59 to $810.14. The variation in this feature is key for us because it accounts for variations in price based on location (location pricing)

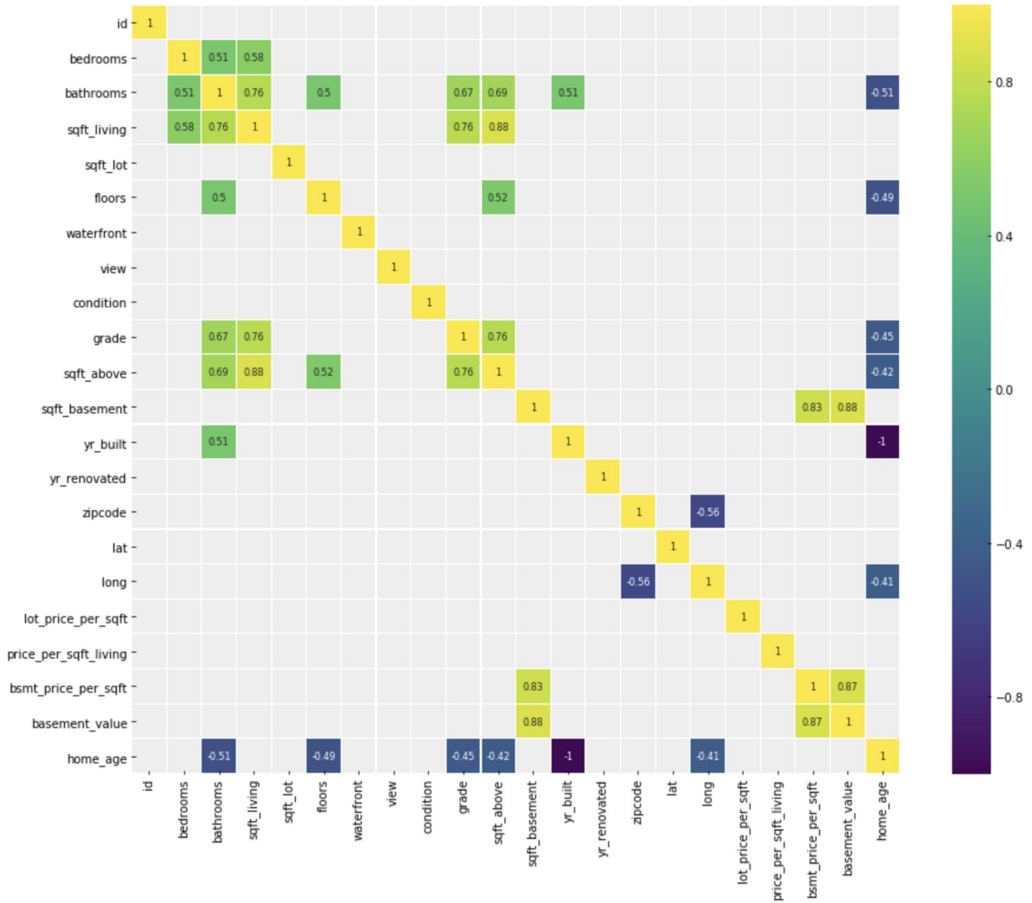
A Multiple Linear Regression model with sqft_living, price_per_sqft_living & bathrooms as independent variables gives us a R^2 value of 0.881

We can have high confidence in our prediction model

From outsider to star

Feature engineering was the key to our road to success in building a predictive model. By using price to predict price, we created a model that naturally is highly accurate

Feature Correlation



Our Predictors

sqft_living

Accounts for space

price_per_sqft_living

Accounts for location pricing

bathrooms

Accounts for feature pricing

Milestones

Loading data

Explore the data, and see what variables correlate to the target variable

October 2015

Clean data and ensure values accurately represent the metric

Data exploration

Data cleaning/munging/wrangling

Data manipulation

Reformat data into desired formats

Feature engineering

Create desired features from analysed data

House price can be predicted by using the formula:

$$y = (6.5 * 10^3) + 297.4770x_1 + 2089.6125x_2 + 9811.1640x_3$$

Tip

x1, x2, and x3 represent sqft_living, price_per_sqft_living, and bathrooms, respectively… We can see that in King County, an added bathroom is roughly worth a 10K increase in home value.