House sales in King County

Miguel Ángel Canela

Associate Professor, IESE Business School

Introduction

The objective of this project is to develop a model for house sale prices in King County (Washington), which includes Seattle. King is the most populous county in Washington (population 1,931,249 in the 2010 census), and the 13th-most populous in the United States. The data include the homes sold between May 2014 and May 2015.

The data set

The data set (file `king.csv`) contains 13 house features plus the sale price and date, along with 21,613 observations. The variables included in the data set are:

- `id`, an identifier.
- `date`, the date when the sale took place.
- `zipcode`, the ZIP code of the house.
- `lat`, the latitude of the house.
- `long`, the longitude of the house.
- `bedrooms`, the number of bedrooms.
- `bathrooms`, the number of bathrooms.
- `sqft_above`, the square footage of the house, discountig the basement.
- `sqft_basement`, the square footage of the basement.
- `sqft_lot`, the square footage of the lot.
- `floors`, the total floors (levels) in house.
- `waterfront`, a dummy for having a view to the waterfront.
- `condition`, a 1–5 rating.
- `yr_built`, the year when the house was built.
- `yr_renovated`, the year when the house was renovated.
- `price`, the sale price in US dollars.

Questions

Q1. How is the distribution of the sale price?

Q2. Develop a linear regression equation for predicting the sale price in terms of the available features. Evaluate this predictive model.

Q3. Plot the actual price versus the price predicted by your model. What do you see?

Q4. How can you get a better model?