Predicting House Prices

Quantitative Methods 1 (Tutorial)

Conor, Linette, Lucas, Minh

1 Variable selection

To predict the Adjusted Sale Price of houses, we considered three major factors:

- the neighbourhood;
- the size of the house; and
- the quality of the construction.

Figure 1 below summarizes the variables we selected to represent each of those characteristics and their relationship.

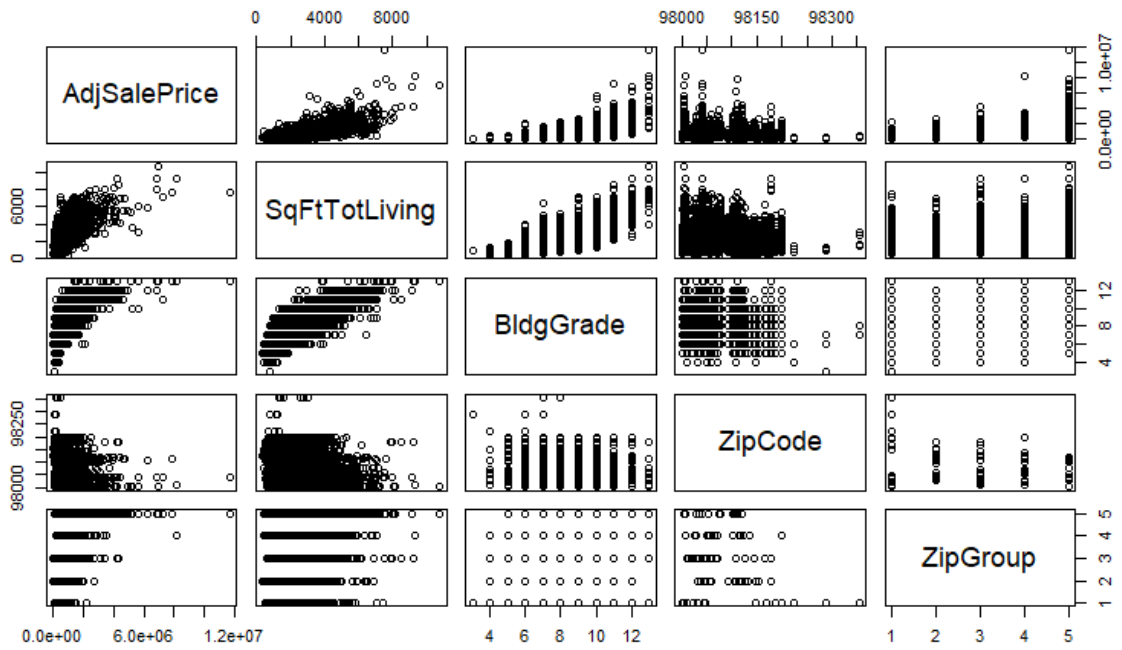

Figure 1: Associations of `AdjSalePrice` with `SqFtTotLiving`, `BldgGrade`, and `ZipCode`

`ZipCode` represents **neighborhood**. We grouped zip codes based on house prices in the variable `ZipGroup` to reflect the socioeconomic profile of each neighborhood.

`SqFtTotLiving` represents the **size of the house**. We avoided adding to the model the variables `SqFtLot`, `SqFtFinBasement`, `NbrLivingUnits` `Bathrooms`, and `Bedrooms` because they are also related to the size of the construction. That decision allowed us to reach a more **parsimonious** model, preventing us from loosing degrees of freedom and from incurring in issues related to **collinearity**. The risk of collinearity among those variables is visible in Figure 2 below:

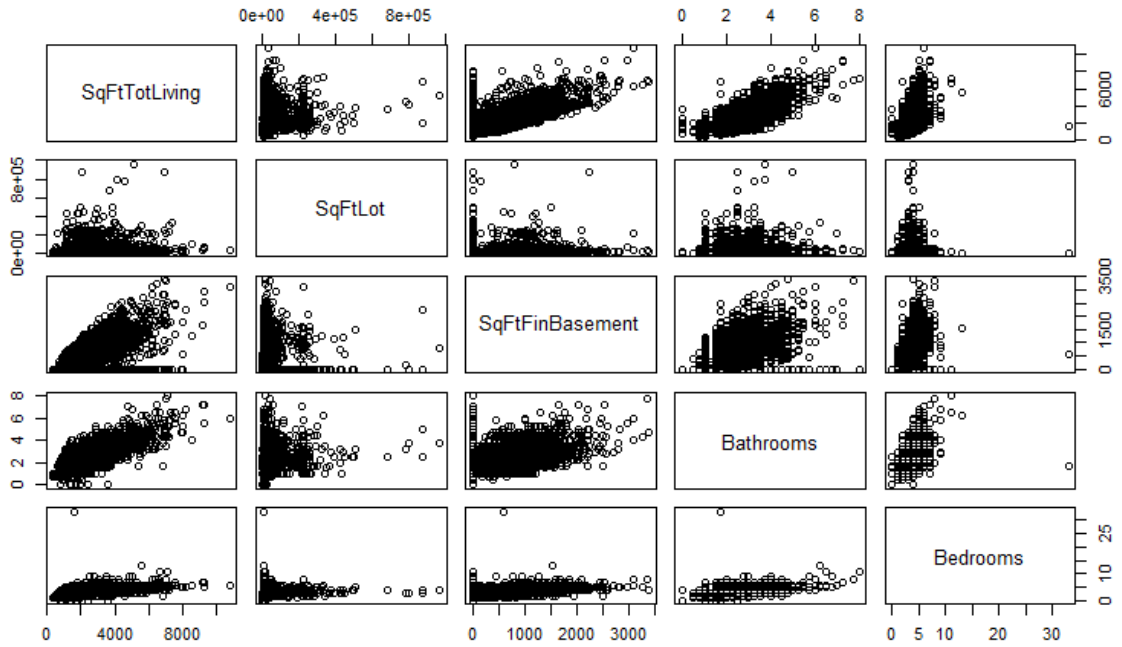

Figure 2: Risk of collinearity among `SqFtLot`, `SqFtFinBasement`, `Bathrooms`, and `Bedrooms`

Lastly, `BldgGrade` represents the **quality of the construction**. For the same reasons stated above (parsimony, preventing the loss of degrees of freedom, and avoiding collinearity issues), we decided not to add `YrBuilt`, `YrRenovated`, and `NewConstruction`, which are also related to the quality of the construction.

2 Interaction effect

To complete the model, we added an interaction effect between `ZipGroup` and `SqFtTotLiving`. The interaction reflects the idea that each additional square foot in the building will affect house prices differently depending on the neighborhood. As a consequence, the slope of `SqFtTotLiving` on `AdjSalePrice` in expensive neighborhoods will be steeper than in

popular ones.

3 Final model