# **Project 14 House Sales in King County Regression**

### **Dataset**

#### **About data:**

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.

#### Variables

- id Unique ID for each home sold
- date Date of the home sale
- price Price of each home sold
- bedrooms Number of bedroom
- 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
- vr 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

#### **Procedure**

- 1. Import Data
- 2. Check dataset size
- 3. Find and treat missing values (If any)
- 4. Check column types and describe which columns are numerical, or categorical
- 5. Perform Univariate analysis
  - 1. Calculate mean, median, std dev, and quartiles of numerical data
  - 2. Plot histogram for a few categorical variables
  - 3. Check the distribution of numerical variables and comment on it
- 6. Perform Bivariate analysis
  - 1. Plot pair plots
  - 2. Perform a Chi-square analysis to check whether there is a relationship between
    - view and waterfront
    - condition and grade
  - 3. Calculate Pearson correlation, and plot their heatmap
- 7. Drop any unnecessary columns
- 8. One hot encode categorical variables (if any)

- 9. Split into train and test set
- 10. Scale the variables
- 11. Train multiple models like Linear regression, Decision Tree, Random Forest, SVR, etc.
- 12. Check their performance, and comment on which is the best model
- 13. Check whether Linear regression performance is good or not
- 14. Check for Multi-collinearity (Hint: Use VIF)
- 15. Remove columns with high multi-collinearity (If any)
- 16. Re-run all the models and check the performance

## Compulsory

- 1. Use grid search CV to tune the hyperparameter of the best model
- 2. Train a polynomial regression model with degree 2, and 3 and compare its performance with other models