## **Project details - regression**

**Background:** You are working as an analyst for a real estate company. Your company wants to build a machine learning model to predict the selling prices of houses based on a variety of features on which the value of the house is evaluated.

**Objective:** The task is to build a model that will predict the price of a house based on features provided in the dataset. The senior management also wants to explore the characteristics of the houses using some plots. One of those parameters includes understanding which factors are responsible for higher property value - $ 650K and above.

**Data:** The data set consists of information on some 22,000 properties. The dataset consisted of historic data of houses sold between May 2014 to May 2015.

These are the definitions of data points provided: (Note: For some of the variables that are self-explanatory, no definition has been provided)

* **Id:** Unique identification number for the property.
* **date**: date the house was sold.
* **price:** the price of the house.
* **waterfront:** if house which has a view to a waterfront.
* **condition:** How good the condition is (overall). \*\*1\*\* indicates worn-out property and \*\*5\*\* excellent.
* **grade:** Overall grade given to the housing unit, based on King County grading system. 1 poor,13 excellent.
* **Sqft\_above:** square footage of house apart from the basement.
* **Sqft\_living15:** Living room area in 2015(implies - some renovations). This might or might not have affected the lot size area.
* **Sqft\_lot15:** lotSize area in 2015(implies - some renovations).

### **Exploring the data**

We encourage you to thoroughly understand your data and take the necessary steps to prepare your data for modeling before building exploratory or predictive models. Since this is a regression task, you must use linear regression for building a model. You need to provide the four most important house features to determine the price ( positively or negatively ).

To explore the data, you can use the techniques that have been discussed in class. Some of them include using the describe method, checking null values, using \_matplotlib\_ and \_seaborn\_ for developing visualizations.

The data has a number of categorical and numerical variables. Explore the nature of data for these variables before you start with the data cleaning process and then data pre-processing (scaling numerical variables and encoding categorical variables).

### **Model**

Compare the error metrics in `train` and `test` sets to make sure that your model doesn't suffer from overfitting/underfitting and find the scaler that best fits your data ( you are free to try other scalers not covered in the course ).