**Predicting house prices**

In this assignment, we are going to build a more regression model for predicting house prices by including more features of the house. In the process, we will also become more familiar with how the R language can be used for data exploration, data transformations and machine learning. These techniques will be key to building intelligent applications.

Follow the rest of the instructions on this page to complete your program. When you are done, ***instead of uploading your code, you will answer a series of quiz questions*** (see the quiz after this reading) to document your completion of this assignment. The instructions will indicate what data to collect for answering the quiz.

**Resources you will need**

Download the house sales pricing dataset here:  [home\_data.csv](https://d396qusza40orc.cloudfront.net/phoenixassets/home_data.csv)

**What you will do**

**1. Selection and summary statistics:** In the notebook attached,

1. Discover which neighborhood (zip code) of Seattle had the highest average house sale price. ***Save this result to answer the quiz at the end.***

Now, take the sales data, select only the houses with this zip code, and compute the average price. ***Save this result to answer the quiz at the end.***

***2.*Filtering data:** One of the key features we used in our model was the number of square feet of living space (‘sqft\_living’) in the house. For this part, we are going to use the idea of filtering (selecting) data.

* first select the houses that have ‘sqft\_living’ higher than 2000 sqft but no larger than 4000 sqft.

What fraction of the all houses have ‘sqft\_living’ in this range? ***Save this result to answer the quiz at the end.***

**3. Building a regression model with several more features:** Build Regression Models using just following features

***my\_features*** *= ['bedrooms', 'bathrooms', 'sqft\_living', 'sqft\_lot', 'floors', 'zipcode']*

Now, going back to the original dataset, you will build a model using the following features:

***advanced\_features*** *= [*

*'bedrooms', 'bathrooms', 'sqft\_living', 'sqft\_lot', 'floors', 'zipcode',*

*'condition', # condition of house*

*'grade', # measure of quality of construction*

*'waterfront', # waterfront property*

*'view', # type of view*

*'sqft\_above', # square feet above ground*

*'sqft\_basement', # square feet in basement*

*'yr\_built', # the year built*

*'yr\_renovated', # the year renovated*

*'lat', 'long', # the lat-long of the parcel*

*'sqft\_living15', # average sq.ft. of 15 nearest neighbors*

*'sqft\_lot15']*

**Compute the RMSE** (root mean squared error) on the test\_data for the model using just *my\_features*, and for the one using *advanced\_features*.

**Note 1: both models must be trained on the original sales training dataset, not the filtered one.**

Note 2: when doing the train-test split, make sure you use random\_state=0, so you get the same training and test sets, and thus results, as we do.

* **What is the difference in RMSE between the model trained with my\_features and the one trained with advanced\_features?** ***Save this result to answer the quiz at the end***

Select variables from *advanced\_features* which will provide the best model using backward elimination using p-values.

* **Which are the significant variables using backward elimination using p-values?** ***Save this result to answer the quiz at the end***

**4. Predictions**

* Which of the models amongst **my\_features\_model** & **advanced\_features\_model** performs best in predicting of the price of house with id = 5309101200. ***Save this result to answer the quiz at the end.***