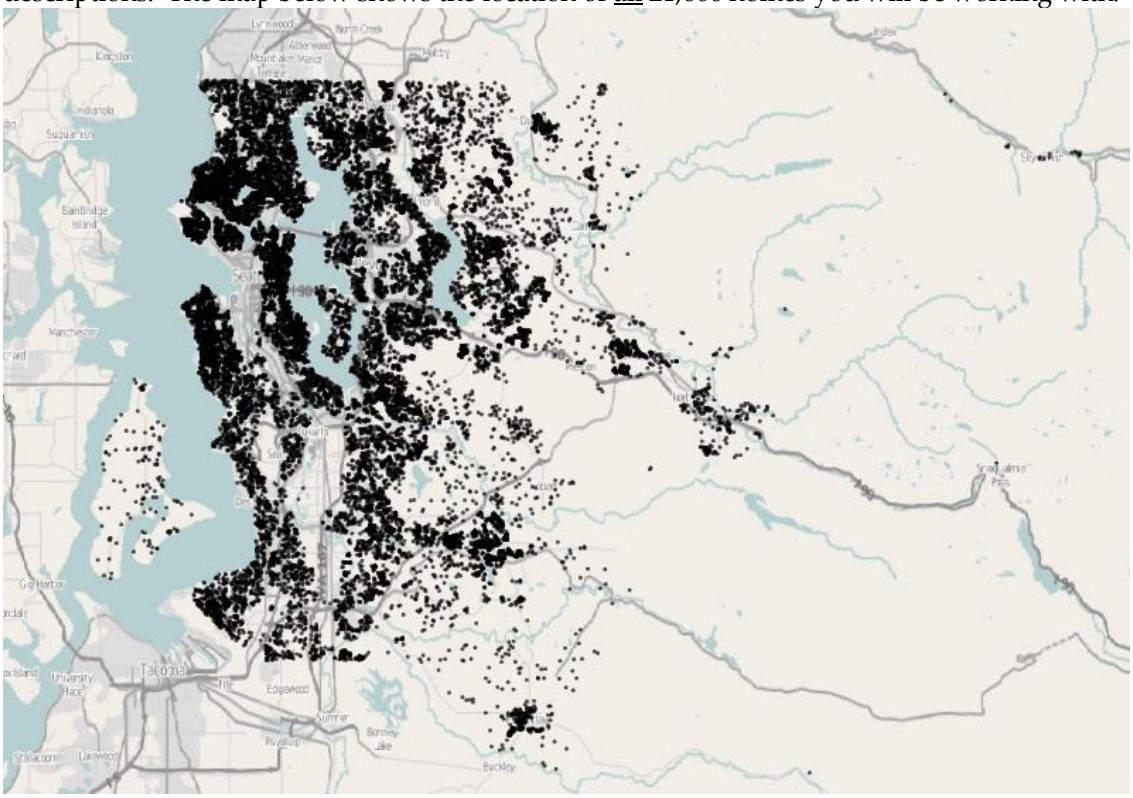
EXTRA CREDIT – Multiple Linear Regression (10 points)

PREDICTING SELLING PRICE OF HOMES IN KING COUNTY, WA

The data for these sales comes from the official public records of home sales in the King County area, Washington State. The data set contains 21,606 homes that sold between May 2014 and May 2015. The table below gives variable names and descriptions. The map below shows the location of all 21,606 homes you will be working with.



Variables in King County, WA Datasets

- ID – id number (DO NOT USE IN YOUR MODELS!)
- **price** - Price of each home sold
- **bedrooms** - Number of bedrooms
- **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 categorical variable for whether the apartment/home was overlooking the waterfront or not (1 = yes, 0 = no).
- **view** - An ordinal index from 0 to 4 of how good the view of the property has.
- **condition** - An index from 1 to 5 on the condition of the apartment**.**
- **grade** - An ordinal 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. Other intermediary values indicate conditions in between these descriptors.
- **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.
- **yr_renovated** - The year of the house's last renovation, 0 indicates it has not been renovated.
- **renovated** – indicator of whether or not the home has been renovated (1 = yes, 0 = no)
- **zipcode –** ZIP code area the house is in (Note: ZIP codes are NOT numeric!)

- **lat -** Lattitude of the home
- **long** - Longitude of the home
- **sqft_living15** - The mean square footage of the interior living space of the nearest fifteen neighboring homes.
- **sqft_lot15** - The mean square footage of the land lots of the nearest fifteen neighboring homes.

Develop a regression models using the training data for predicting the selling price of the homes in the test data using the available predictors described above. Note that not all of the variables are numeric and will have to be dealt with accordingly.

```
> King = read.csv(file.choose())
```
← read the file King County Homes (train).csv

```
> KingTest = read.csv(file.choose())
```
← read the file King County Homes (test).csv

Document the model development process by copying and pasting relevant R commands, output, and graphics into your write-up. All R code copied and pasted into your final assignment submission MUST be in `Courier New 9 point font`. So that I can tell what portion is code.

Tasks

1) Fit a multiple linear regression model to these data without trying to address any model deficiencies etc.
   a) Fit a base model and discuss any deficiencies (but don't try to fix them). (2pts.)
   b) Use stepwise reduction of the base model to reduce the base model and include discussion of final model. (2 pts.)
   c) Use a best subsets regression using the "leaps" package to fit the best subset of variables (2pts.)
   d) Use cross-validation methods to estimate the prediction error of the best fitting model that you identified using split-sample and 10 k-fold approaches. (2 pts.)

2) Give the predicted selling price of the homes in the test data (`KingTest`) using your best fitting model and submit them in a .csv file with your name in it. For example, suppose my final model is called `poo.lm` then you should do something along the following lines in R. (2 pts)

```
> mypred = predict(poo.lm,newdata=KingTest)
> submission = data.frame(ID=KingTest$ID,ypred=mypred)
> write.csv(submission,file="DeppasPredictions.csv")
```

   Submit the file `DeppasPredictions.csv`, along with your results from parts (1) should be a Microsoft Word document.

IMPORTANT NOTE:

Whatever changes you make to the training data MUST also be done to the `KingTest` data as well. For example, if you choose to log transform the living space variable in the training data, then you must also do this in the test data! If my final model fit to the training data uses the log living space and not living space then when I predict using the test data it must also have log living space as well! The variable names in both datasets (`King` and `KingTest`) must be the same!