EMIS 3309: Information Engineering

Lab 6: data exploration

About dataset: The dataset contains records of house sales occured between May 2014 to May 2015 in King County, Washington. It has 21613 observations of 22 features including sales ID and price of the house. Refer to data description in the link to learn more about the features.

Source of dataset: `https://www.kaggle.com/harlfoxem/housesalesprediction`

Map of King County, Washington

Download 'house.csv' from Canvas.

1. Import data to R by using **read.csv( )** function. Save the object as 'house'.

2. Take a look at the data. Use **head( )** function to read first 10 rows of the dataset. Does data match column names?

3. The first column is redundant as it just shows row numbers; remove the column.

4. Check if there is any missing data by counting total number of empty cells.

   Hint: the function **is.na( )** returns TRUE if a given cell value is missing and FALSE otherwise. You can combine the function with **sum( )** function.

5. Use **summary( )** function to generate basic statistics for each column. Identify two columns for which these statistics can be meaningful.

6. For each of the identified columns, create a plot that can help to understand the stored values. You can create any type of plot we have used in class including histogram, box-whisker plot and scatter plot.

7. Are there outliers in the dataset? If there are, identify the sales id and explain why you would, or would not, remove these data points.

8. Are there houses without bathrooms? If there are, identify the sales id and explain why you would, or would not, remove these data points.

9. Use **pairs( )** function to generate pairwise scatter plots between 'price', 'yr_built', and 'sqft_living'. Are there relationships between these features?

10. Based on this dataset, what kind of information can we learn? (In other words, what are some interesting questions we can answer using the dataset?)