HW1

Part I

In this homework, you will build an Jupyter notebook for predicting house prices, using data from King County, USA, the region where the city of Seattle is located.

1. Selection and summary statistics:

Which neighborhood (zip code) of Seattle has the highest average house sale price? Now, take the sales data, select only the houses with this zip code, and compute the average price.

***2.* Filtering data:**

One of the key features is the number of square feet of living space ('sqft_living') in the house. 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?

3. Visualization:

Plot the relationship between 'house sale price' and 'sqft_living'.

4. Building regression models with different features (all models must be fit on the original sales dataset, not the one filtered on `sqft_living`. Do NOT split the data into training and test sets):

Build a regression model to predict house prices using just 'sqft_living' and add the trend line in the plot in part 3. Report the intercept and slope.

Using this simple regression model, what is the predicted price for a house with 2650 sqft?

Using this simple regression model, what is the estimated square-feet for a house costing $800,000?

Build a regression model to predict house prices using just 'bedrooms'. Report the intercept and slope.

Using this simple regression model, what is the predicted price for a house with 3 bedrooms?

Compute and compare the RMSE (root mean squared error) of the two models.

Part II

1. Load the breast cancer dataset. Import the functions `load_breast_cancer` from `sklearn.datasets`.

2. Provide an explanation of the data set.

3. What are the features in the data set?

4. Create some basic visualization of the data set.

5. Split the data into training set (80%) and test set (20%) using random_state = 0.

6. Build KNN models with n_neighbors = {1, 2, 3, 4, 5}. Compare their performance on the training set and test set. Which one is the best model?