The goal of this problem is to predict the price of a house given its area and number of rooms. We are given the following training data for a regression task, which will be implemented through a binary regression tree.

Sample	**Features**		**Outcome**
	Area (sq.ft)	**Rooms**	**Price ($)**
S1	600	1	800
S2	800	1	1000
S3	800	2	1100
S4	1000	2	1500

(a) Examine all binary partitions that yield from the values of the feature *Area*, i.e., $\{R_1 : \ Area \leq 600,\ R_2 : \ Area > 600\}$ and $\{R_1 : \ Area \leq 800,\ R_2 : \ Area > 800\}$.
For the samples that belong to the two regions of each partition, compute the average of the corresponding outcomes y_i. The are also called centers and are denoted as c_1 and c_2.
Then compute the sum of square error between the sample outcomes of the samples y_i and the centers c_1 and c_2, i.e., $SSE = \sum_{i \in R_1} (y_i - c_1)^2 + \sum_{i \in R_2} (y_i - c_2)^2$

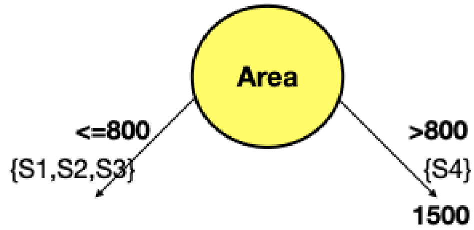
- <u>If $\{R_1 : \ Area \leq 600,\ R_2 : \ Area > 600\}$:</u>
  $S1$ belongs to R_1. $S2$, $S3$, and $S4$ belong to R_2.
  $c_1 = 800$, $c_2 = \frac{1000+1100+1500}{3} = 1200$
  $SSE = (800 - 800)^2 + (1000 - 1200)^2 + (1100 - 1200)^2 + (1500 - 1200)^2 = 140,000$
- <u>If $\{R_1 : \ Area \leq 800,\ R_2 : \ Area > 800\}$:</u>
  $S1$, $S2$, and $S3$ belong to R_1. $S4$ belongs to R_2.
  $c_1 = \frac{800+1000+1100}{3} = 967$, $c_2 = 1500$
  $SSE = (800 - 967)^2 + (1000 - 967)^2 + (1100 - 967)^2 + (1500 - 1500)^2 = 63,267$

(b) Perform the same operation as in **(a)** for the feature *Rooms*.

The only region segmentation for this variable is $\{R_1 : \ Rooms = 1,\ R_2 : \ Rooms = 2\}$
$S1$ and $S2$ belong to R_1. $S3$ and $S4$ belong to R_2.
$c_1 = \frac{800+1000}{2} = 900$, $c_1 = \frac{1100+1500}{2} = 1300$
$SSE = (800 - 900)^2 + (1000 - 900)^2 + (1100 - 1300)^2 + (1500 - 1300)^2 = 180,000$

(c) Find the partition from features *Area* or *Rooms* that yielded the lowest SSE and place that as a node of the tree. Plot the first level of the tree. For which node(s) can you provide a final value? For which samples you cannot provide a final output?

The partition that gave the lowest SSE was $\{R_1 : \ Area \leq 800,\ R_2 : \ Area > 800\}$, therefore the tree will look as follows:



On the right node of the above tree we can provide an output, since it only contains a single sample.
Although the left node of the tree contains more than one samples, we could still provide an output as the mean of outcomes for $\{S1, S2, S3\}$.

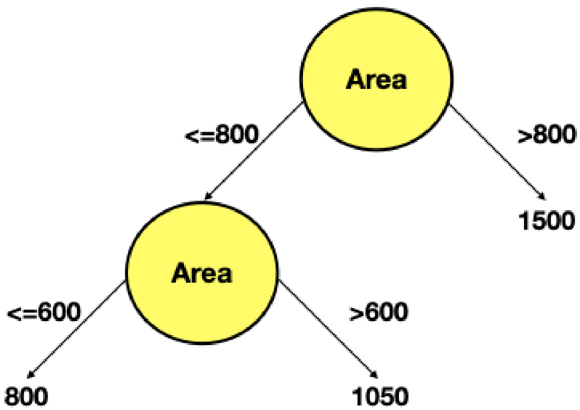
(d) Perform the same operation as in **(a-c)** to expand the tree for one more level (when needed). Assuming the level 2 is the maximum level of the tree, provide the final decisions in the corresponding nodes.

Now we are only operating on samples $\{S1, S2, S3\}$.

- $\underline{\text{If } \{R_1: \ Area \leq 600, \ R_2: \ Area > 600\}:}$
  $S1$ belongs to R_1. $S2$ and $S3$ belong to R_2.
  $c_1 = 800, \ c_2 = \frac{1000+1100}{2} = 1050$
  $SSE = (800 - 800)^2 + (1000 - 1050)^2 + (1100 - 1050)^2 = 5000$

- $\underline{\text{If } \{R_1: \ Rooms = 1, \ R_2: \ Rooms = 2\}:}$
  $S1$ and $S2$ belong to R_1. $S3$ belongs to R_2.
  $c_1 = \frac{800+1000}{2} = 900, \ c_2 = 1100$
  $SSE = (800 - 900)^2 + (1000 - 900)^2 + (1100 - 1100)^2 = 20,000$

The partition with the lowest SSE is $\{R_1: \ Area \leq 600, \ R_2: \ Area > 600\}$, therefore the tree looks as follows:



The error that results from this regression tree for the training data is $\sqrt{(1000 - 1050)^2 + (1100 - 1050)^2}$, since the terminal node for $S2$ and $S3$ contained their average.