Occupation	Gender	Age	Salary
Service	Female	45	$48,000
	Male	25	$25,000
	Male	33	$35,000
Management	Male	25	$45,000
	Female	35	$65,000
	Male	26	$45,000
	Female	45	$70,000
Sales	Female	40	$50,000
	Male	30	$40,000
Staff	Female	50	$40,000
	Male	25	$25,000

Consider the data in Table above (end of chapter 6 or 8). The target variable is salary. Start by discretizing salary and age as follows:

Less than $35,000	Level 1
$35,000 to less than $45,000	Level 2
$45,000 to less than $55,000	Level 3
Above $55,000	Level 4

0 – 30	<= 30
31 - 40	<= 40
Above 40	<= 50

5.1 Construct a classification and regression tree to classify salary based on the other variables only one split level.

Hint: you may want to set up the excel file like the following

Split	PL	PR	Level	P(j |tL)	P(j |tR)	2PL PR	Q(s|t)	Φ(s|t)
1	0.273	0.727	L1	0.333	0.125	0.397	0.583	0.231
			L2	0.333	0.250			
			L3	0.333	0.375			
			L4	0.000	0.250			
2								

5.2

The “breast cancer dataset” in CANVAS was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. The features in the dataset, described below, have been categorized from 1 to 10.
Use these categorized features to answer the following questions.
Important: make sure your categories are represented by the “factor” data type in R and DO NOT replace the missing values.

	Features	Domain
	Sample code number	id number
F1.	Clump Thickness	1 - 10
F2.	Uniformity of Cell Size	1 - 10
F3.	Uniformity of Cell Shape	1 - 10
F4.	Marginal Adhesion	1 - 10
F5.	Single Epithelial Cell Size	1 - 10
F6.	Bare Nuclei	1 - 10
F7.	Bland Chromatin	1 - 10
F8.	Normal Nucleoli	1 - 10
F9.	Mitoses	1 - 10
	Diagnosis Class:	(2 for benign, 4 for malignant)

5.2
Use the CART methodology to develop a classification model for the Diagnosis.