Due on March 13, 2020. Report must show the student's name and USC ID.

1. The `OJ` data set from the ISLR library contains 1070 purchases where the customer either purchased Citrus Hill or Minute Maid Orange Juice (Use `?OJ` for more details). It is of interest to predict `Purchase` using all other variables. Use `set.seed(1,sample.kind="Rounding")` to create a training set containing a random sample of 800 observations, and a test set containing the remaining observations.

   Fit a classification tree to the training data to answer questions (a) to (c).

   a) (10 pts.) Plot the tree. What is the training error rate? What is the test error rate?

   b) (10 pts.) Use `set.seed(2)` and cross-validation to find the best number of terminal nodes. Which tree size corresponds to the lowest cross-validated classification error rate?

   c) (10 pts.) Plot a pruned tree with five terminal nodes. What is the test error rate?

   For the following question, fit a RF to the training set using all predictors.

   d) (10 pts.) Which predictors are the most important? What is the test error rate?

   When fitting a boosted tree the number of trees, depth of trees, shrinkage, should be carefully selected. If a tuning grid is defined, the `train` function can be used to tune these parameters (see p217, handout).

   e) (10 pts.) Fit a boosted tree selecting the best parameter values. What is the test error rate?

2. In segmenting the market, a breakfast cereal manufacturer uses health and diet consciousness as the segmentation variable. Four segments are developed:

   1 = Concerned about eating healthy foods
   2 = Concerned primarily about weight
   3 = Concerned about health because of illness
   4 = Unconcerned

   To distinguish between groups, a survey is conducted (see `cereal.csv`). In the survey, people are categorized as belonging to one of these groups. The most recent census reveals that 234,564,000 Americans are 18 and older.

   a) (20 pts.) Use the `prop.test` function to find a 95% Confidence interval for the true proportion of American adults who are concerned about eating healthy foods. Then use it to estimate how many American adults belong to group 1.

   b) (20 pts.) Each respondent was also asked the amount spent on breakfast cereal in an average month. The company would like to know whether on average the market segment *concerned about eating healthy foods* outspends the other market segments.

3. (10 pts.) The following table is a sample of the dataframe `Hitters` from library `ISLR`. It shows the data of nine baseball players chosen at random. Select the rownumber equal to the first digit of your USC ID. Then use the data in that row and the regression tree shown below to predict the salary (in 000s of dollars) of the selected player.

	AtBat	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	CHmRun	CRuns	CRBI	CWalks	League	Division	PutOuts	Assists	Errors	NewLeague
1	401	92	17	49	66	65	13	5206	1332	253	784	890	866	A	E	0	0	0	A
2	217	46	7	32	19	9	4	694	160	32	86	76	32	A	E	307	25	1	A
3	127	32	8	16	22	14	8	727	180	24	67	82	56	N	W	202	22	2	N
4	496	141	20	65	78	37	11	5628	1575	225	828	838	354	N	E	200	11	3	N
5	321	87	10	39	42	30	2	396	101	12	48	46	33	N	E	805	40	4	N
6	413	92	16	72	48	65	1	413	92	16	72	48	65	N	E	280	9	5	N
7	239	60	0	30	11	22	6	1941	510	4	309	103	207	A	E	121	151	6	A
8	185	37	1	23	8	21	2	214	42	1	30	9	24	N	E	76	127	7	A
9	196	43	7	29	27	30	13	3231	825	36	376	290	238	N	E	80	45	8	N