Introduction: I used the Adults Census Income Dataset (https://www.kaggle.com/uciml/adult-census-income/version/3#). The purpose of this data set is to use census data like age, type of work, education, marital status, occupation, relationship, race, sex to predict whether an adult makes more than or less than $50,000 annually. Some features from the data set like "hours worked per week" had to be omitted because they were incomplete or continuous (this algorithm does not support continuous values). Also, due to time constraints, the tests could only be performed on a portion of this large data set. Finally, I also performed tests with a pruned decision tree.

Unpruned Tree		Pruned With Q=0.9	
Training data Size	1999	Training Data Set Size	1999
Training Time	0.45212293	Training Time	0.51089001
Classification Rate	0.795	Classification Rate	0.802
Tree Size	891	Tree Size	365

*It should say "Classification Score" not "Classification Rate"

Question 6: Extremely aggressive pruning (high q value) did not change the classification rate significantly. This shows that the size of the data set relative to the number of features was sufficient to prevent over fitting in the decision tree. Given the relatively average tree size of 891, perhaps the tree was not fit enough to the data set. A larger data set would have likely allowed the tree to grow more and make more splits, thus becoming more accurate. The difference in training scores is likely due to time spent running the Chi-Square hypothesis test.

Question 7: One use of this decision tree could be determining how much of a factor race or sex play in determining payment. Suppose you wanted to determine how much of a factor gender plays in job payment. One could build a decision tree model ignoring the sex attribute and then build another with the sex attribute. Then, compare the classification scores of the two models. If there is a "big" difference, then gender likely plays a role in determining pay. If there is a "small" difference, pay differences could be due to confounding variables that are associated with gender.