CSCD 429 Data Mining Lab 5 (22 points)

Goal: Learn how to do exploratory analysis and build classification/regression models using R. For each question, include the R command and the result in your submission.

Step 0: Install and load libraries

```
#Installing libraries
install.packages('rpart')
install.packages('caret')
install.packages('rpart.plot')
install.packages('ISLR2')

#Loading libraries
library(rpart)
library(caret)
library(rpart.plot)
library(ISLR2)
```

Part I: Experiments using *mushrooms* data

Step 1: Import the data.

(1 point) Import the data set mushrooms.csv, and use str() to display the structure of the data set. Note that attribute "type" is the class attribute. It has two possible values: 'e' stands for edible class and 'p' stands for the poisonous class of mushrooms.

Step 2: Data exploration and preparation.

1) (2 points) From the output of str(), what do you find regarding variable "veil_type"? Do you think it is a potentially useful variable for classification? Why? How would you deal with it during data preprocessing phase? Include the corresponding R command you used.
2) To get a good understanding of the predictor variables, we can create a contingency table for each predictor variable vs. class variable (response variable) in order to understand whether that particular predictor variable is significant for detecting the output or not.
   a) (2 point) Now create a contingency table for variable "odor".
   b) (2 points) What do you learn from the above table? Use Chi-square test to verify your answer.
3) (2 points) Split the data into a training set (80%) and a test set (20%) using *simple random sampling without replacement*. The training set will be used to build the Decision Tree model and the test set will be used to validate the efficiency of the model.

Step 3: Building and testing a decision tree model.

1) (3 points) Build a decision tree using the **training** data set and display the tree. (Package "rpart" reference can be found at https://cran.r-project.org/web/packages/rpart/rpart.pdf. Package "rpart.plot" reference can be found at https://cran.r-project.org/web/packages/rpart.plot/rpart.plot.pdf.

2) (2 points) Test the above decision tree model using the **test** data set, and use confusion matrix to display the performance of your model. Use your own words to explain the overall performance of your model.

Part II: Experiments using *Auto* data from ISLR2 library

Auto data is inside ISLR2 library. If you already installed and loaded this library probably, you should be able to use it for the following experiments.

Step 0: Use View(Auto) to look the data you are going to play with.

Step 1: Build a simple linear regression model and interpret the output.

1) (2 points) Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Then use the summary() function to print the results.
2) (2 points) Is there a relationship between the predictor and the response? How strong is the relationship between the predictor and the response? Is the relationship between the predictor and the response positive or negative?
3) (2 points) What is the predicted mpg associated with a horsepower of 104?
4) (2 points) Plot the response and the predictor. Use the abline() function to display the least squares regression line.