```
# Homework 3
# Elizabeth Hollis/Mendoza
# I put the table into Excel to be able to import it into R

install.packages("ggplot2")
library(ggplot2)

install.packages("diagram")
library(diagram)

install.packages("readxl")
library(readxl)

install.packages("Rcpp")
library(Rcpp)

install.packages("dplyr")
library(dplyr)

# Q1: Compute the GINI Index for the Gender Attribute
# Both have 10 each out of 20, so parition at number 10

# For female gender attribute:
1 - ((6/10)^2 + (4/10)^2) # Equals 0.48

#For male gender attribute:
1 - ((6/10)^2 + (4/10)^2) # Equals 0.48

# Calculate the total GINI Index for Gender
# Both equal 0.48 or 4/10 * 6/10 = 48/100 or 0.48



# Q2: Compute the GINI Index for car type attribute
# There are 3 different car attributes - family, luxury, and sports

# Family car attribute, 4 records in all:
1 - ((1/4)^2 + (3/4)^2) # Equals 0.375

# Luxury car attribute, 8 records in all:
1 - ((1/8)^2 + (7/8)^2) # Equals 0.21875

# Sports car attribute, 8 records in all:
1 - ((8/8)^2 + (0/8)^2) # Equals 0

# Calculate the total GINI Index for CarType
(4/20 * 0.375) + (8/20 * 0.21875) + (8/20 * 0) # Equals 0.1625


# Q3: Compute the GINI Index for shirt size attribute
# There are 4 different shirt attributes - small, medium,
# large, and extra large

# Small shirt attribute, 5 records in all:
```

```
1 - ((3/5)^2 + (2/5)^2) # Equals 0.48

# Medium shirt attribute, 7 records in all:
1 - ((3/7)^2 + (4/7)^2) # Equals 0.4897959

# Large shirt attribute, 4 records in all:
1 - ((2/4)^2 + (2/4)^2) # Equals 0.5

# Extra Large shirt attribute, 4 records in all:
1 - ((2/4)^2 + (2/4)^2) # Equals 0.5

# Calculate the total GINI Index for ShirtSizes
(5/20 * 0.48) + (7/20 * 0.4897959) + (4/20 * 0.5) + (4/20 * 0.5)
# Equals 0.4914286 or 0.4915


# Q4: Based on the GINI Index calculations, which attribute should
# be your root node for your decision tree?
# The one with the lowest GINI Index will be the root node

# Gender attribute GINI Index - 0.48
# Car attribute GINI Index - 0.1625
# Shirt attribute GINI Index - 0.4915

# Car attribute has the lowest GINI Index so it will be the root node.


# Total of 20 examples, with 10 each going to C0 and C1
# Have to find the entropy first, which is below:
-10/20*log2(10/20)-10/20*log2(10/20)
#Equals 1 since both are 10
# This entropy will be used in the following Information
# gain questions


# Q5: Compute the Information gain for the gender attribute.
# Have to do values of each gender, it's split into
# 6 - Male, 4 - Female for the first 10
# 6 - Female, 4 - Male for the second half

# For male gender attribute:
-6/10*log2(6/10)-4/10*log2(4/10) #Equals 0.9709506 or 0.971

# For female gender attribute:
-4/10*log2(4/10)-6/10*log2(6/10) #Equals 0.9709506 or 0.971

# Now to use the values we got above to get Entropy:
(10/20 * 0.971) + (10/20 * 0.971) #Equals 0.971

# Now we use the entropy from the calculations of C0 and C1
# as well as the entropy from above
1 - 0.971
# Equals 0.029 which is our Information Gain for Gender
```

```
# Q6: Compute the Information gain for the CarType attribute.
# Have to do values for each car type, it's split into
# 1 Family in the first 10 partition, and 3 in the second
# 8 Sports in the first 10 partition, and 0 in the second
# 1 Luxury in the first 10 partition, and 7 in the second

# For the Family Type attribute:
# Have to get the divide number down if the total number
# of an attribute doesn't equal 10 (less than)
-1/4*log2(1/4)-3/4*log2(3/4)  # Equals 0.8112781 or 0.811

# For the Sports Type attribute:
# Have to get the divide number down if the total number
# of an attribute doesn't equal 10 (less than)
# This one doesn't matter as much since we have the 0
-10/10*log2(10/10)-0/10*log2(0/10) # Equals 0

# For the Luxury Type attribute:
# Have to get the divide number down if the total number
# of an attribute doesn't equal 10 (less than)
-1/8*log2(1/8)-7/8*log2(7/8) # Equals 0.5435644 or 0.544

# Now to use the values we got above to get Entropy:
(4/20 * 0.811) + (8/20 * 0) + (8/20 * 0.544) # Equals 0.3798 or 0.380

# Now we use the entropy from the calculations of C0 and C1
# as well as the entropy from above
1 - 0.380
# Equals 0.620 which is our Information Gain for CarType


# Q7: Compute the Information gain for the shirt size attribute.
# Have to the values for each shirt size, it's split into
# 3 small in the first 10 partition, and 2 in the second
# 3 medium in the first 10 partition, and 4 in the second
# 2 large in the first 10 partition, and 2 in the second
# 2 extra large in the first 10 partition, and 2 in the second

# For the Small size attribute:
# Have to get the divide number down if the total number
# of an attribute doesn't equal 10 (less than)
-6/10*log2(6/10)-4/10*log2(4/10) # Equals 0.9709506 or 0.971

# For the Medium size attribute:
# Have to get the divide number down if the total number
# of an attribute doesn't equal 10 (less than)
-3/7*log2(3/7)-4/7*log2(4/7) # Equals 0.9852281 or 0.985

# For the Large size attribute:
# Have to get the divide number down if the total number
# of an attribute doesn't equal 10 (less than)
# This one doesn't matter as much since both are equal
# in both partitions
```

```
-5/10*log2(5/10)-5/10*log2(5/10) # Equals 1

# For the X Large size attribute:
# Have to get the divide number down if the total number
# of an attribute doesn't equal 10 (less than)
# This one doesn't matter as much since both are equal
# in both partitions
-5/10*log2(5/10)-5/10*log2(5/10) # Equals 1

# Now to use the values we got above to get Entropy:
(5/20 * 0.971) + (7/20 * 0.985) + (4/20 * 1) + (4/20 * 1)
# Equals 0.9875 or 0.988

# Now we use the entropy from the calculations of C0 and C1
# as well as the entropy from above
1 - 0.988
# Equals 0.012 which is our Information Gain for Shirt sizes


# Q8: Based on the Information gain, which attributes should be your
# root nodes for your decision tree?
# The one with the highest Information gain will be the root node

# Gender attribute Info gain - 0.029
# Car attribute Info gain - 0.620
# Shirt attribute Info gain - 0.012

# Car attribute has the highest Information Gain
# so it will be the root node.
```