Programming Entropy and Information Gain

For this project, you will be creating a collection of functions that analyze categorical data using the concepts of entropy and information gain.

The attached .CSV file will provide you with some data to test your function. The functions described in the attached .PDF should be coded, tested, and placed in a single R script. Place your R script in a public repository on github and submit a link to the script here. ***Entropy Project is due end of day on Tuesday September 16th.***

1. Create a function `entropy()` that takes a vector $\vec{d}$ as input and returns a single numeric value that is the entropy $E(\vec{d})$ of the vector:

$$E(\vec{d}) = -\sum_{i=1}^{k} p_i \log_2 p_i$$

   Note that your function should accept as input a vector with data type factor, character, or numeric, even though the data will be treated as categorical data. Call this function

2. Create a function `infogain()` that takes two vectors – the target $\vec{d}$ and the attribute $\vec{a}$ with which to partition the data – and returns the information gain $I(\vec{d}, \vec{a})$ for the attribute:

$$I(\vec{d}, \vec{a}) = E(\vec{d}) - \sum_{j=1}^{m} \left(\frac{n_j}{n} E(\vec{d}_j) \right)$$

   where n is the length of the target $\vec{d}$, n_j is the size of the partition of $\vec{d}$ according to the jth value of the attribute $\vec{a}$, and $E(\vec{d}_j)$ is the entropy of the partition of $\vec{d}$ according to the jth value of the attribute $\vec{a}$.

3. Create a function `decide()` that takes a data frame - the target $\vec{d}$ and the collection of candidate attributes $\vec{a}_i$ with which to partition the data – and the number of the column that is the target and returns a list containing two items: the identity (by column number) of the attribute that maximizes the information gain and a vector of the information gains for each of the candidate attributes.

Store your three functions in an R script called entropyfunctions.R. Your submission here should consist of a single link to the file in your github repository.

A sample dataset (entropy-sample.csv) has been provided. The functions you create should be able to produce the following results when run against this dataset:

```
> entropy(dataset$answer)
[1] 0.9832692
> infogain(dataset$answer,dataset$attr1)
[1] 2.411565e-05
> infogain(dataset$answer,dataset$attr2)
[1] 0.2599038
> infogain(dataset$answer,dataset$attr3)
[1] 0.002432707
> decide(dataset,4)
$max
[1] 2

$gains
     attr1      attr2      attr3
2.411565e-05 2.599038e-01 2.432707e-03
```