Project #3 Hierarchical Clustering and Entropy Approach to our Kaggle Dataset

By Amir Yaacoobi

**Introduction**

In this discovery, I want to use a different approach than before to solve a familiar problem. Again, we delve into our Kaggle Data regarding Heart Disease, using new methods such as HC Trees and Entropy to discover the links between the distributions of Body Mass Index with respect to three categorical variables, a few which will be familiar to us from previous experiments. The purpose of using our Entropy and Hierarchical Clustering methods are to compare the means of various different sub-groups. The method of Entropy is utilized to measure the level of disorder or unpredictability that exists within a given dataset. Dendrograms are commonly used to visually represent the results of Hierarchical Clustering, providing a useful tool for identifying and interpreting community structures among diverse samples, and aiding in the comprehension of complex datasets. We will be conducting separate investigations pertaining to each of these sub-data sets and making predictions and conclusions regarding our categories and the link to BMI.

**Method**

We begin by splitting our original Kaggle dataset into five sub categories with an emphasis on General Health. “General Health” is measured on a 1-5 scale, Each respective category is sorted in a manner in which each sub-category is sorted so that those who score a “1” on General Health, will be grouped together, those who score a “2” will follow suit, and so on. We will use these subcategories to run tests such as Entropy Tests, HC Trees, and the 3 way interaction effects between the three different binary categories we will discuss below.

**Data Source**

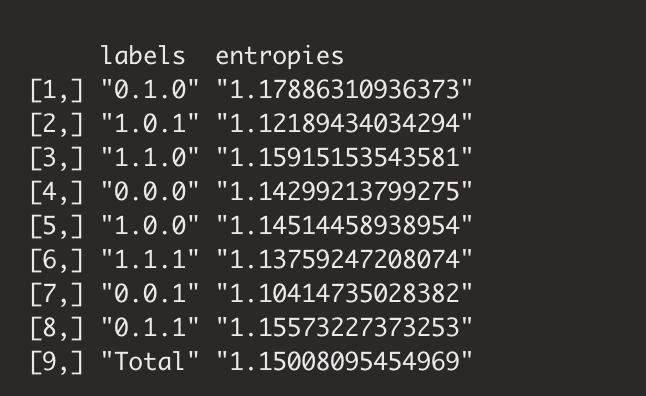
We obtained data from 253,680 survey responses of the BRFSS 2015 Heart Disease Health Indicators Dataset Notebook. The particular datasets we will study and analyze are from the variables, General Health and Body Mass Index. The binary categorical variables we will be using in this discovery are, **High Blood Pressure, High Cholesterol, and Smoking.** I intentionally chose these categories as we have done work in the past with these categories, and they present a spread between the health aspects, and factors that might cause those health concerns,

**Breaking up into sub-categories**

Our dataset includes the category “General Health” and splits this category up into five distinct ratings. The ratings, ranging from 1-5, are meant to represent a different ranking of the subject's health. 1 represents “Excellent” health and 5 represents “Poor” general health. For this experiment, we want to break up our entire data set into 5 sub-datasets that are sorted and filtered to have a corresponding rating of General Health. For instance, sub-dataset 1 will only include those with a General Health rating of “1” or “Excellent”, each of the groups will follow suit.

**Sub Category #1, Excellent General Health**

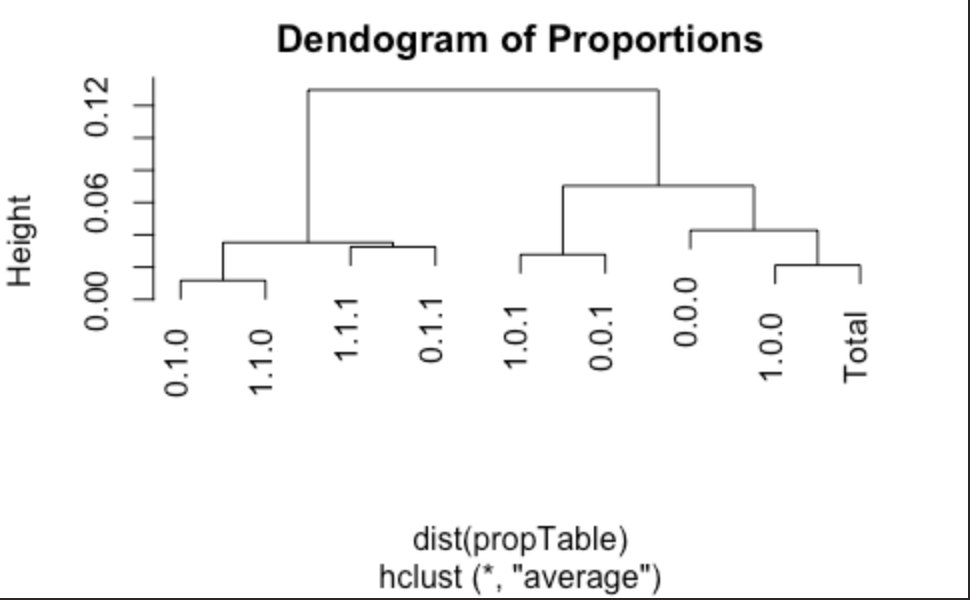
**Entropy Approach**

****

**Summary -**

The entropy table shows the proportion of the total entropy that each interaction term accounts for, and a higher residual entropy indicates inadequate capture of essential information in the data, while a lower residual entropy indicates successful capture of the relevant information. According to the entropy values, the row which involves Smoker, High Blood Pressure, and High Cholesterol, [7] is the most effective predictor of BMI, with an entropy proportion of 1.104. Conversely, the interaction term which involves Not Smoking, High Blood Pressure, and Normal Cholesterol, row [1] is the least effective predictor of BMI, with an entropy proportion of 1.178

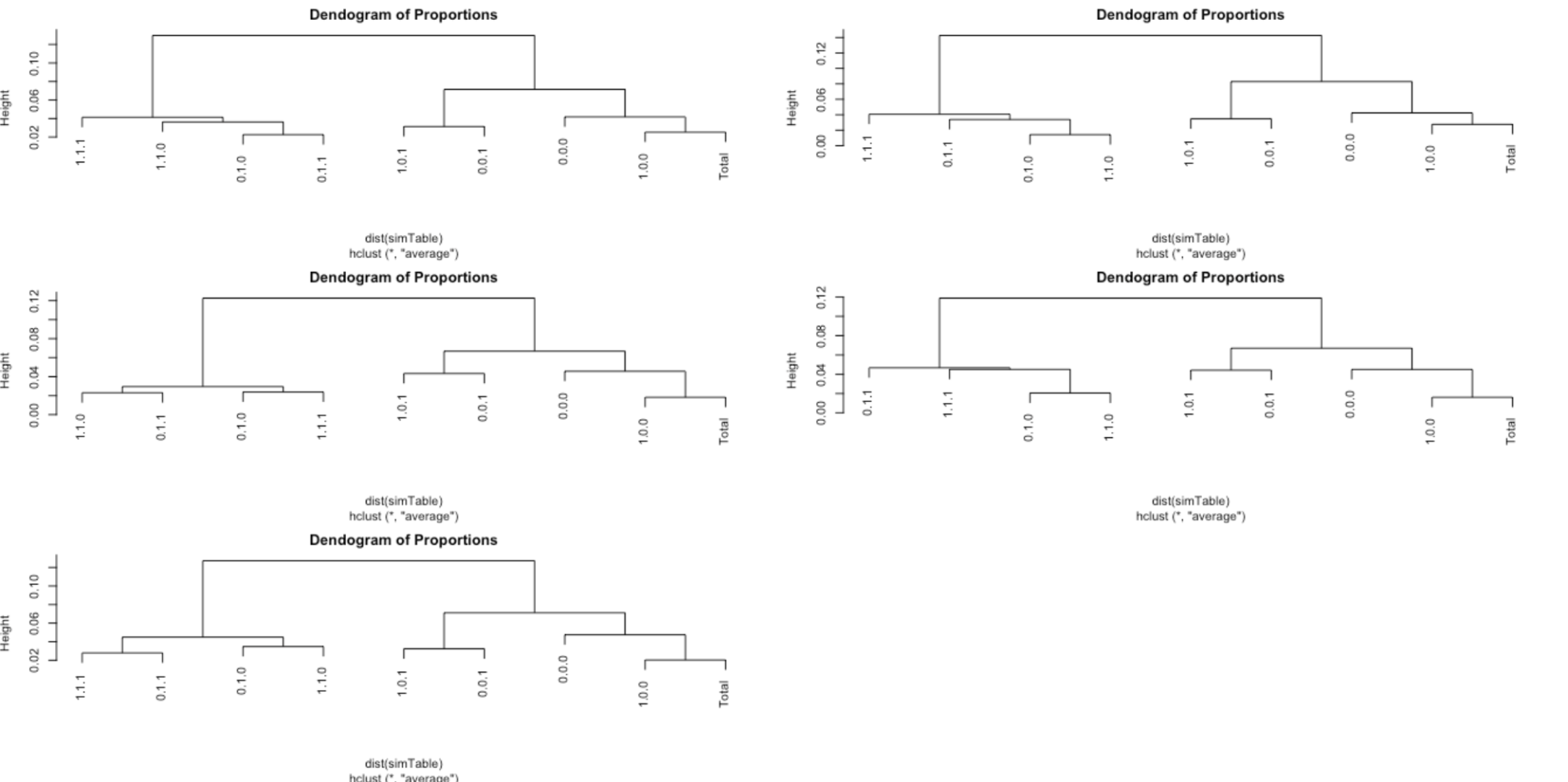
**Hierarchical Clustering Approach**



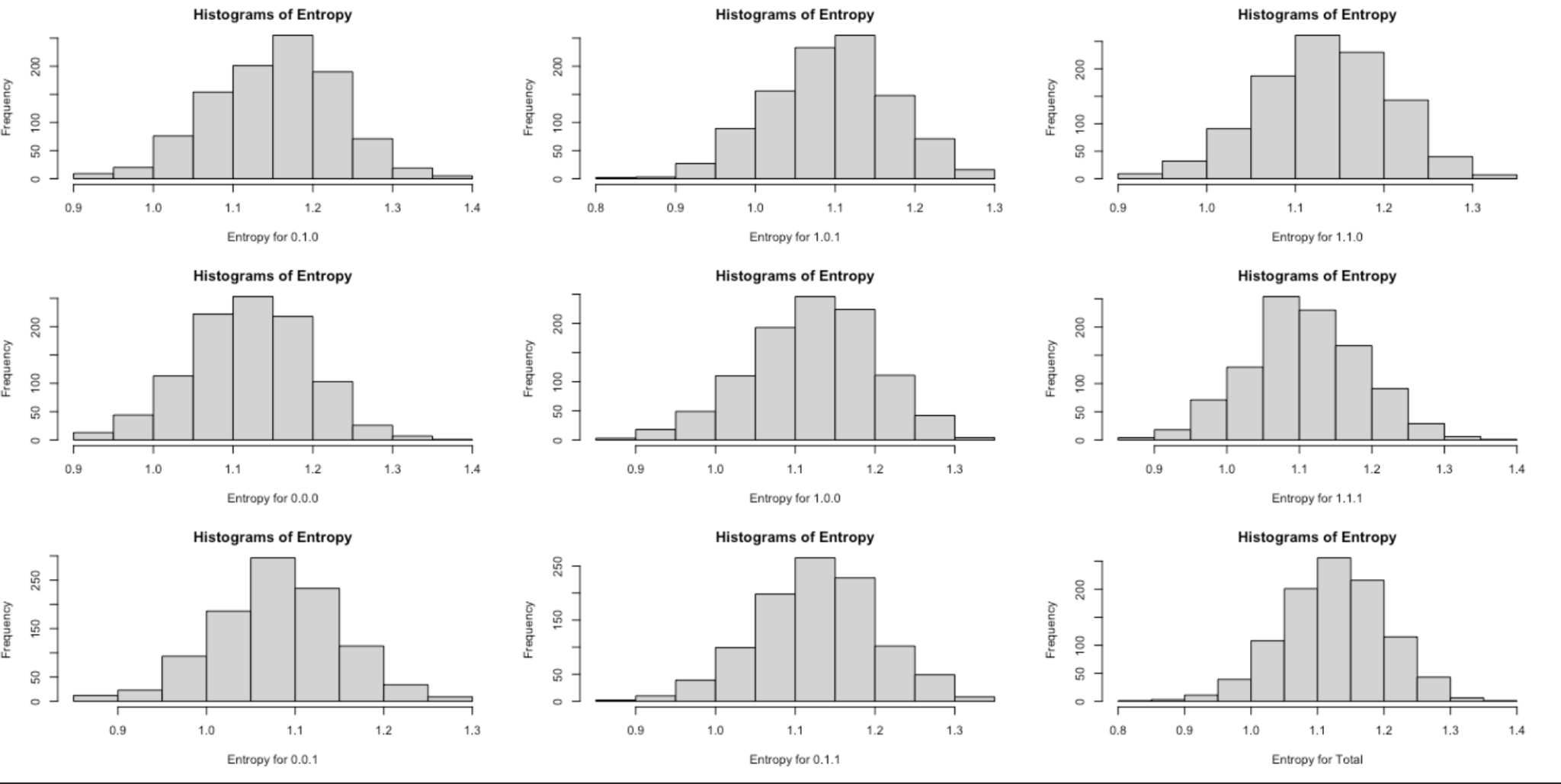
**Summary -**

To better understand how the interaction effects of Smoking, High Blood Pressure, and having High Cholesterol affect individuals’ BMI levels, we constructed a dendrogram based on BMI, showing how the 8 subgroups cluster together. By focusing on the subset of individuals with a general health level of 1, we aimed to assess the community structures formed between samples and determine how similar their BMI distributions were to the overall BMI distribution, depicted by the "Total" category on the dendrogram. Our analysis of the dendrogram revealed that the interaction effect of Smoking, having no high cholesterol, normal blood pressure was the most similar to the overall BMI distribution, as it shares a branch with the "Total" category.

**Reliability with Hierarchical Clustering and Entropy using Simulated Data**



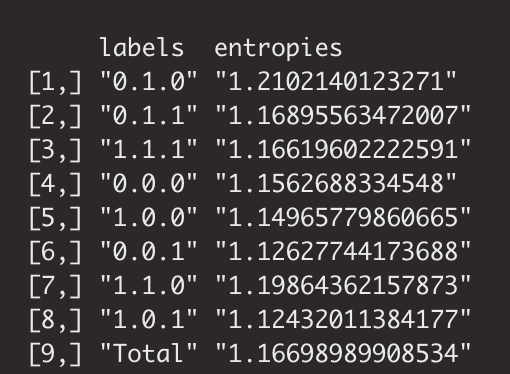
**Summary -** To ensure that our HC Trees, also known as Dendograms, are reliable and are giving us readings that can be used to come to a conclusion, we simulated data and ran the same approach on this simulated data. Using multinomial distribution with n=10,000 from simulated data and compared them with the HC Trees obtained from our data on a subset of General Health=1. Looking at the 5 new HC Trees, we can observe that each of our “Total” values are paired on the same branch in each of the Dendograms. This means that even with simulated data, our results hold true. We can also confirm that on each Tree, our “Total” value shares the same branches throughout the simulation, which is the same result as our first HC Tree. Ultimately, this shows us that our method of HC Trees are reliable and yield consistent results throughout this category.



**Summary -** To evaluate the reliability of our entropy approach we conducted for this sub-category, we generated 9 histograms of entropy using multinomial distribution with n=10,000 from simulated data and compared them with the true entropy values obtained from our data on a subset of General Health=1. We came to the conclusion that all 9 histograms had a normal distribution, meaning that none of the nine histograms pictured are skewed in any specific direction, which suggests that the entropy values were clustered around a specific mean. This indicates that the data has a moderate amount of uncertainty, making it useful for users as it is not too random or predictable. If the entropy histograms are skewed towards the left or right, it would mean the data is uninformative and unreliable due to either too much or too little information. Based on our observations, we conclude that all the interaction terms are informative and that they can be useful when making conclusions!

**Sub-Category #2, “Very Good” General Health**

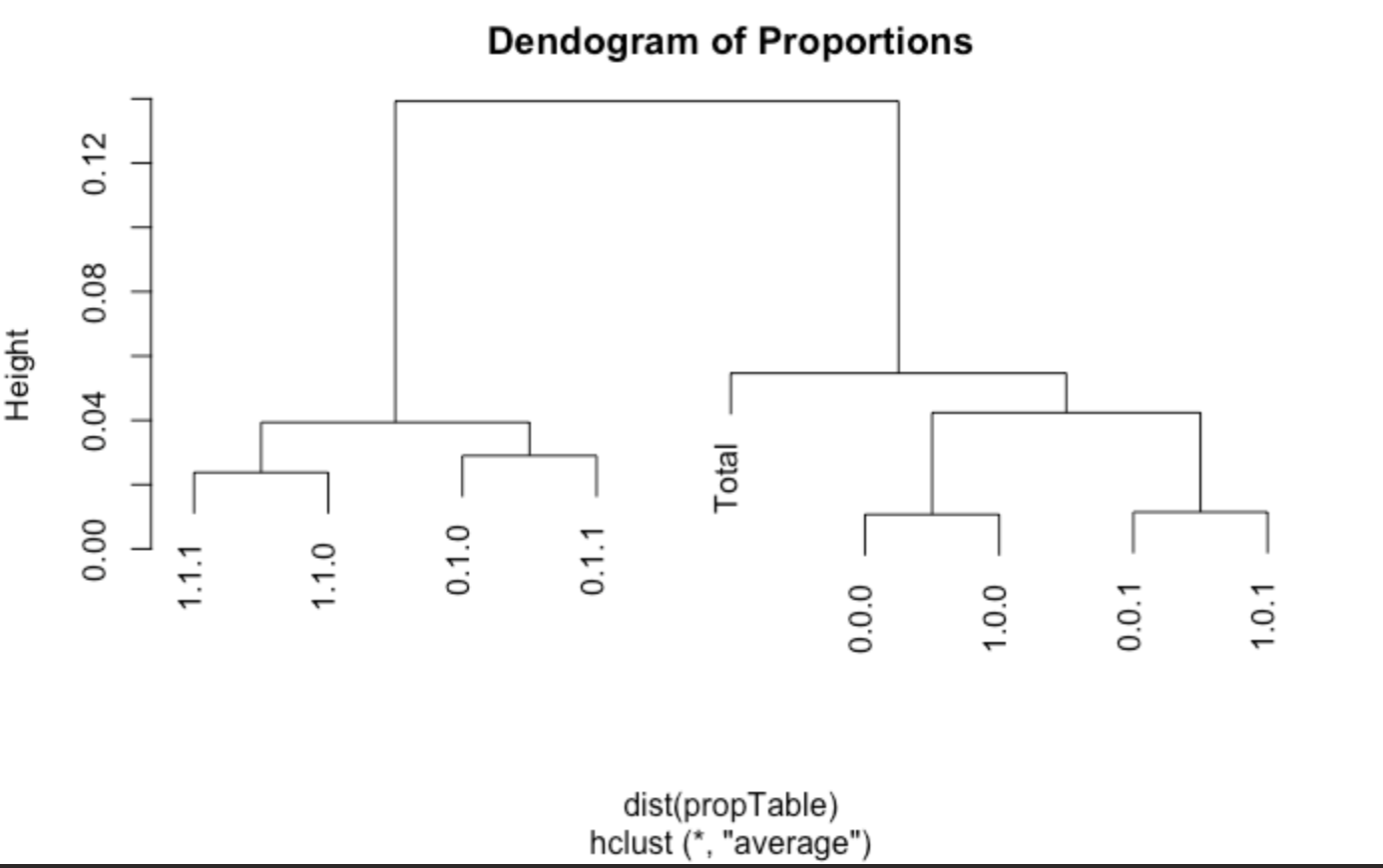
**Entropy Approach**



**Summary -**

The entropy table shows the proportion of the total entropy that each interaction term accounts for, and a higher residual entropy indicates inadequate capture of essential information in the data, while a lower residual entropy indicates successful capture of the relevant information. According to the entropy values, the row which involves Non- Smoker, Normal Blood Pressure, and High Cholesterol, [6] is the most effective predictor of BMI, with an entropy proportion of 1.126. Conversely, the interaction term which involves Smoking, Normal Blood Pressure, and High Cholesterol, row [8] is the least effective predictor of BMI, with an entropy proportion of 1.243

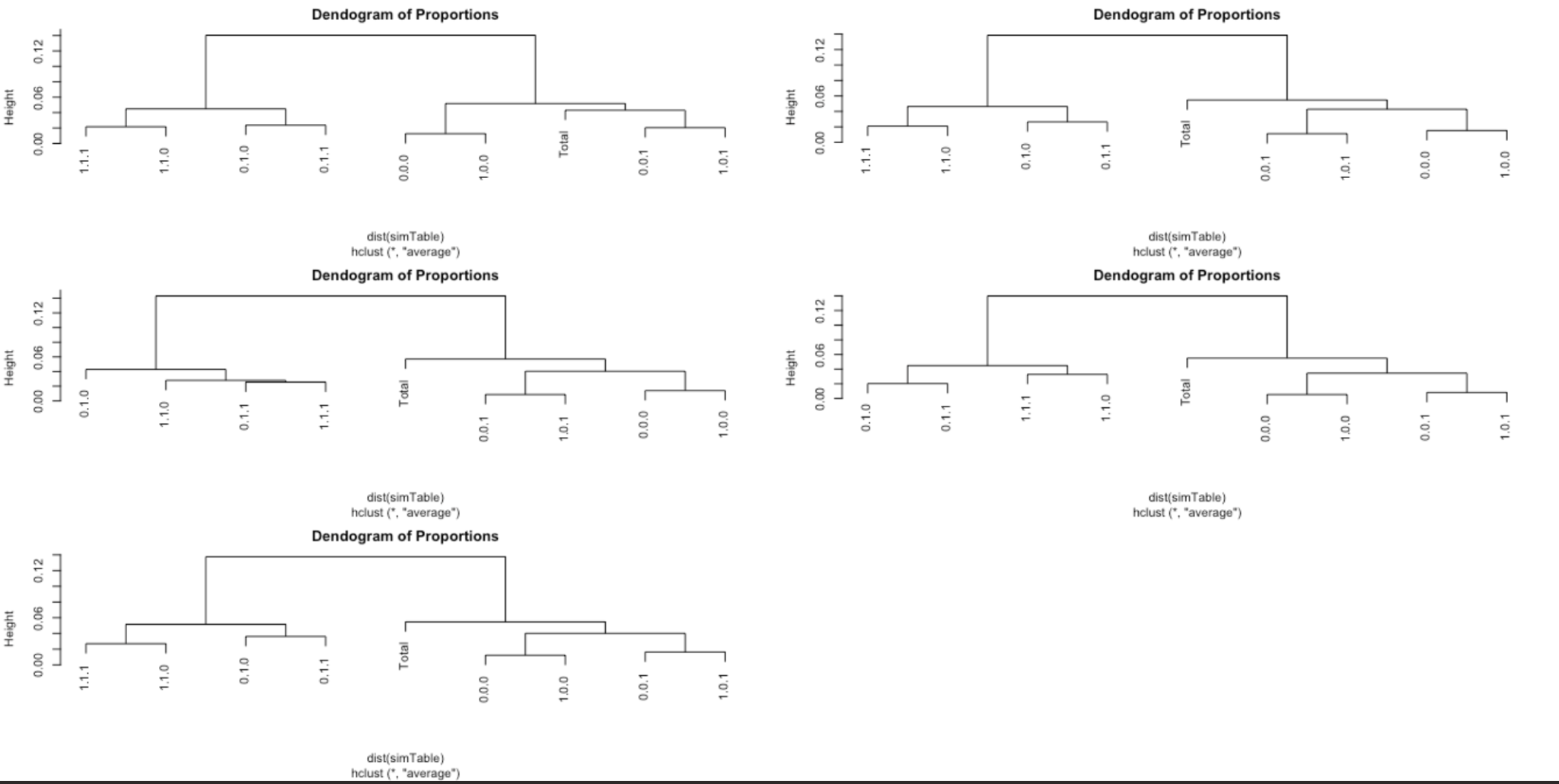
**Hierarchical Clustering Approach**

****

**Summary -**  In this section, we are focusing on the subset of individuals with a general health level of 2, we aimed to assess the community structures formed between samples and determine how similar their BMI distributions were to the overall BMI distribution, depicted by the "Total" category on the dendrogram.

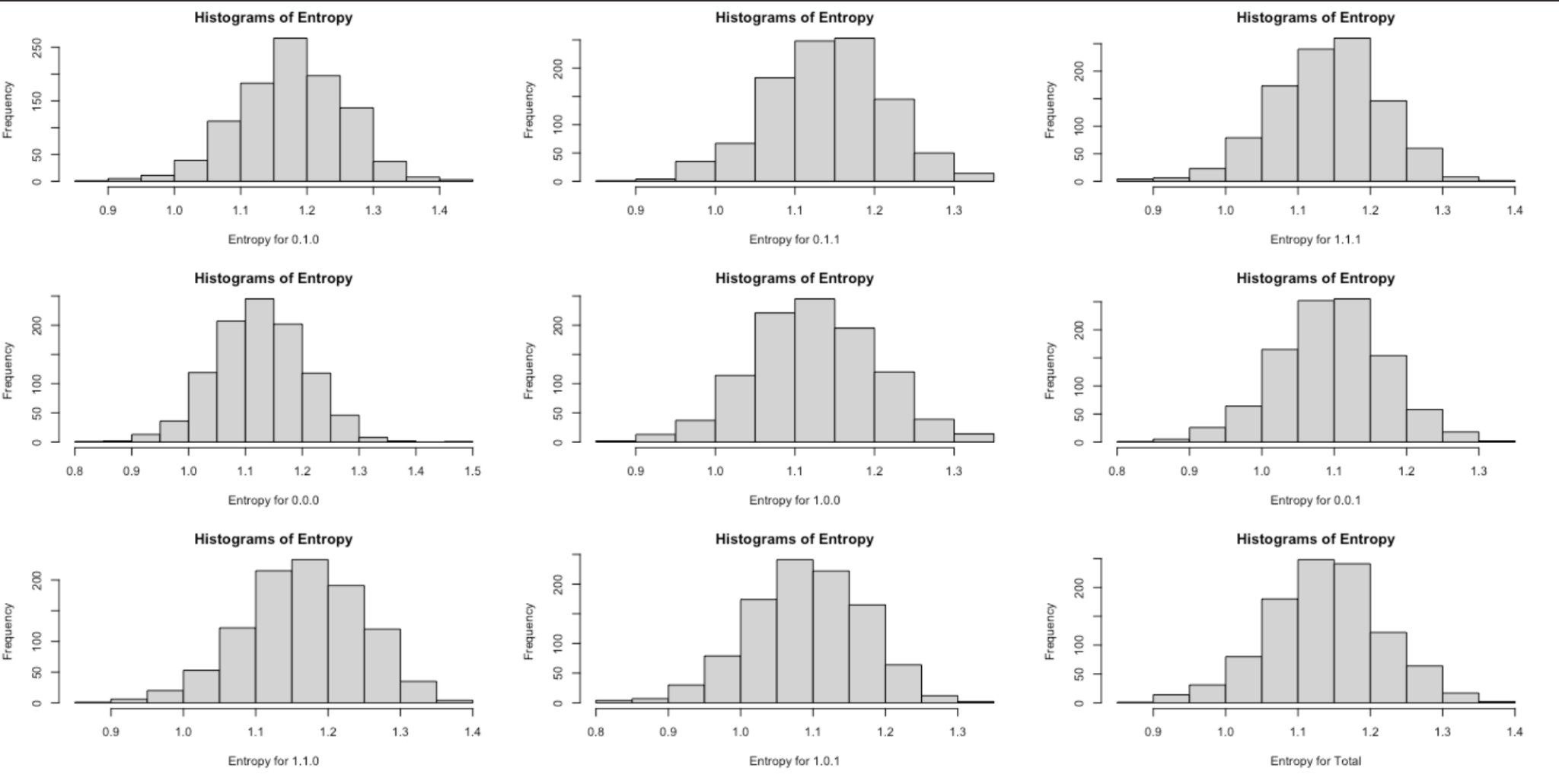
Our analysis of the dendrogram revealed that the interaction effect of the following combinations 0.0.0, 1.0.0, 0.0.1, and 1.0.1, share a branch with the "Total" category. Moreover, the interaction effect of regularly doing physical activity, having high cholesterol, and low blood pressure also exhibited a similar distribution to the overall BMI distribution. Interestingly, only 4 of the interaction effects were significantly far away from the "Total," indicating that about half of the interactions had similar distributions to the overall BMI distribution.

**Reliability with Hierarchical Clustering and Entropy on Simulated Data**



**Summary -**

To ensure that our HC Trees, also known as Dendograms, are reliable and are giving us readings that can be used to come to a conclusion, we simulated data and ran the same approach on this simulated data. Using multinomial distribution with n=10,000 from simulated data and compared them with the HC Trees obtained from our data on a subset of General Health=2. Looking at the 5 new HC Trees, we can observe that each of our “Total” values are paired on the same branch in each of the Dendograms. This means that even with simulated data, our results hold true. We can also confirm that on each Tree, our “Total” value shares the same branches throughout the simulation, which is the same result as our first HC Tree. Ultimately, this shows us that our method of HC Trees are reliable and yield consistent results throughout this category.

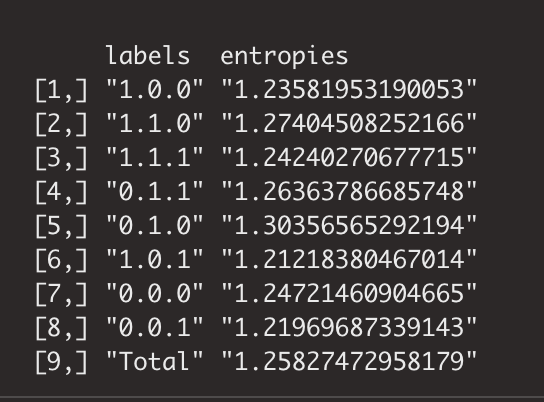


**Summary -**

To evaluate the reliability of entropy, we generated 9 histograms of entropy using multinomial distribution with n=10,000 from simulated data and compared them with the true entropy values obtained from our data on a subset of General Health=2. We observed that all 9 histograms had a normal distribution, as in none of the histograms are skewed in any direction, which suggests that the entropy values were clustered around a specific mean. This indicates that the data has a moderate amount of uncertainty, making it useful for users as it is not too random or predictable. If the entropy histograms are skewed towards the left or right, it would mean the data is uninformative and unreliable due to either too much or too little information. Based on our observations, we conclude that all the interaction terms are informative and strong predictors of BMI within the subgroup data of General Health=2.

**Sub-Category #3, “Good” General Health**

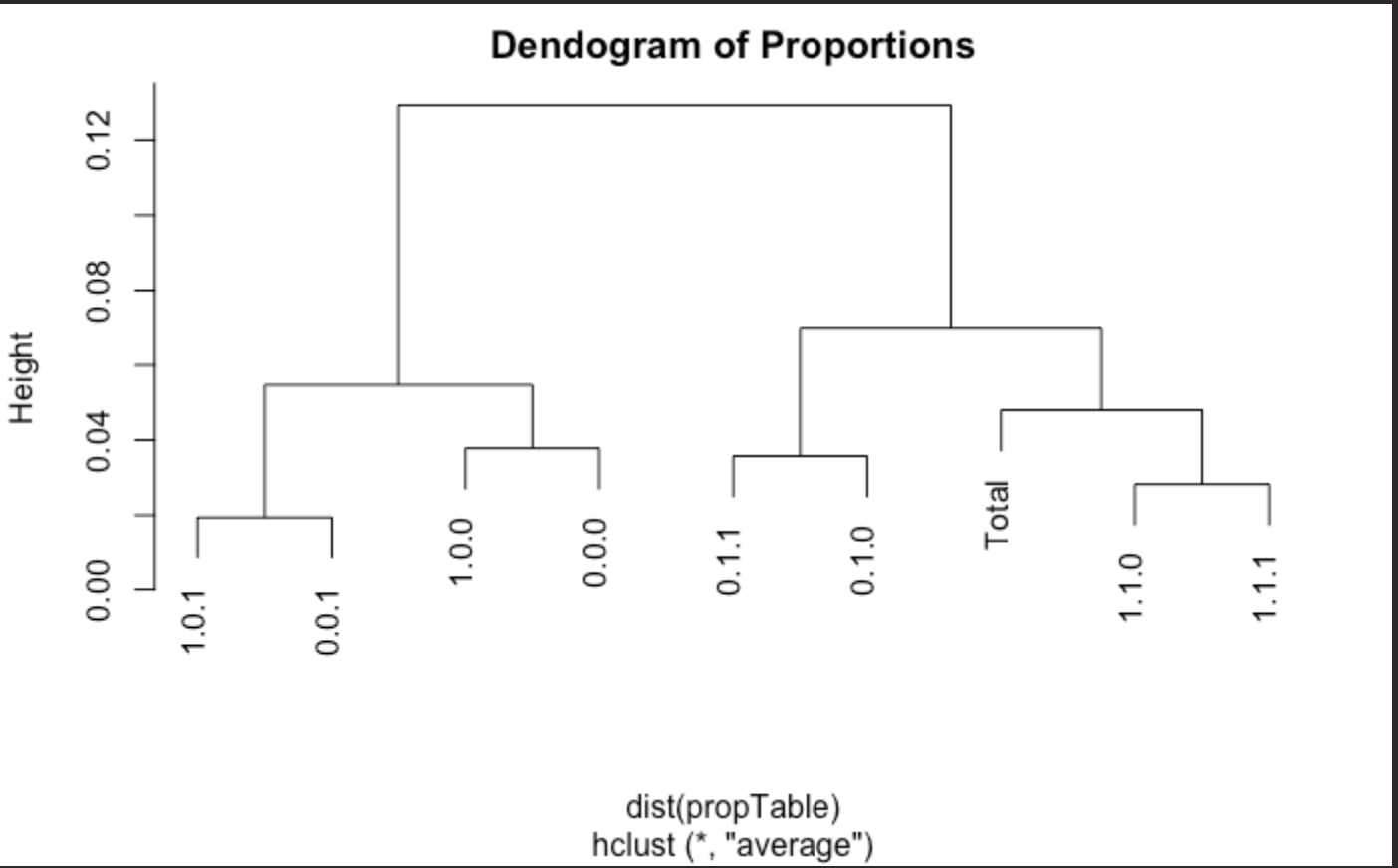
**Entropy**



**Summary -**

The entropy table shows the proportion of the total entropy that each interaction term accounts for, and a higher residual entropy indicates inadequate capture of essential information in the data, while a lower residual entropy indicates successful capture of the relevant information. According to the entropy values, the row which involves Smoker, Normal Blood Pressure, and High Cholesterol, [6] is the most effective predictor of BMI, with an entropy proportion of 1.212. Conversely, the interaction term which involves Non-Smoking, High Blood Pressure, and Normal Cholesterol, row [5] is the least effective predictor of BMI, with an entropy proportion of 1.303

**Hierarchical Clustering**

****

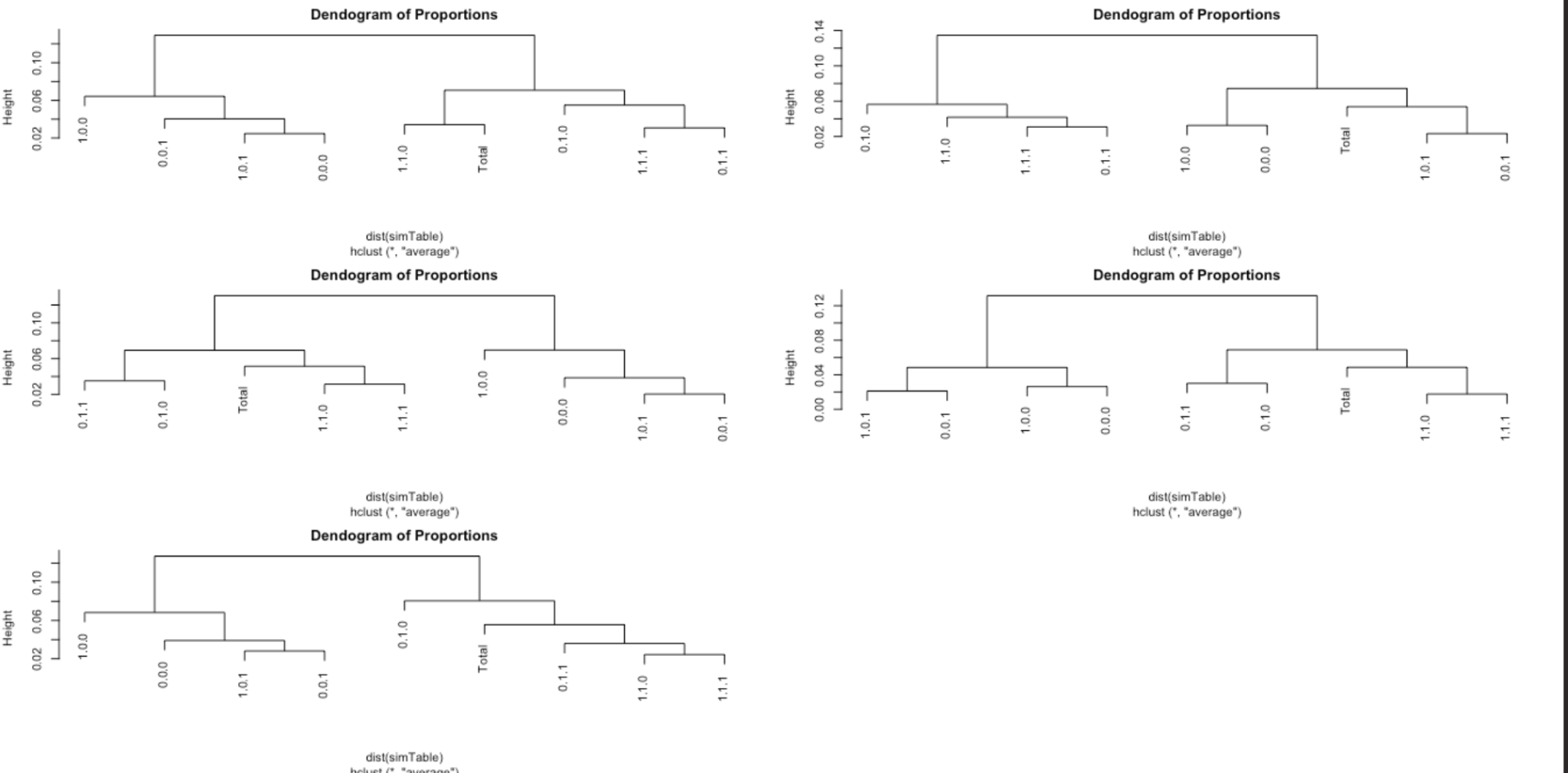
**Summary -**

Harem, we constructed a dendrogram based on BMI, showing how the 8 subgroups cluster together. By focusing on the subset of individuals with a general health level of 3, we aimed to assess the community structures formed between samples and determine how similar their BMI distributions were to the overall BMI distribution, depicted by the "Total" category on the dendrogram.

Our analysis of the dendrogram revealed that the following interaction effects, of 1.1.0, and 1.1.1 were

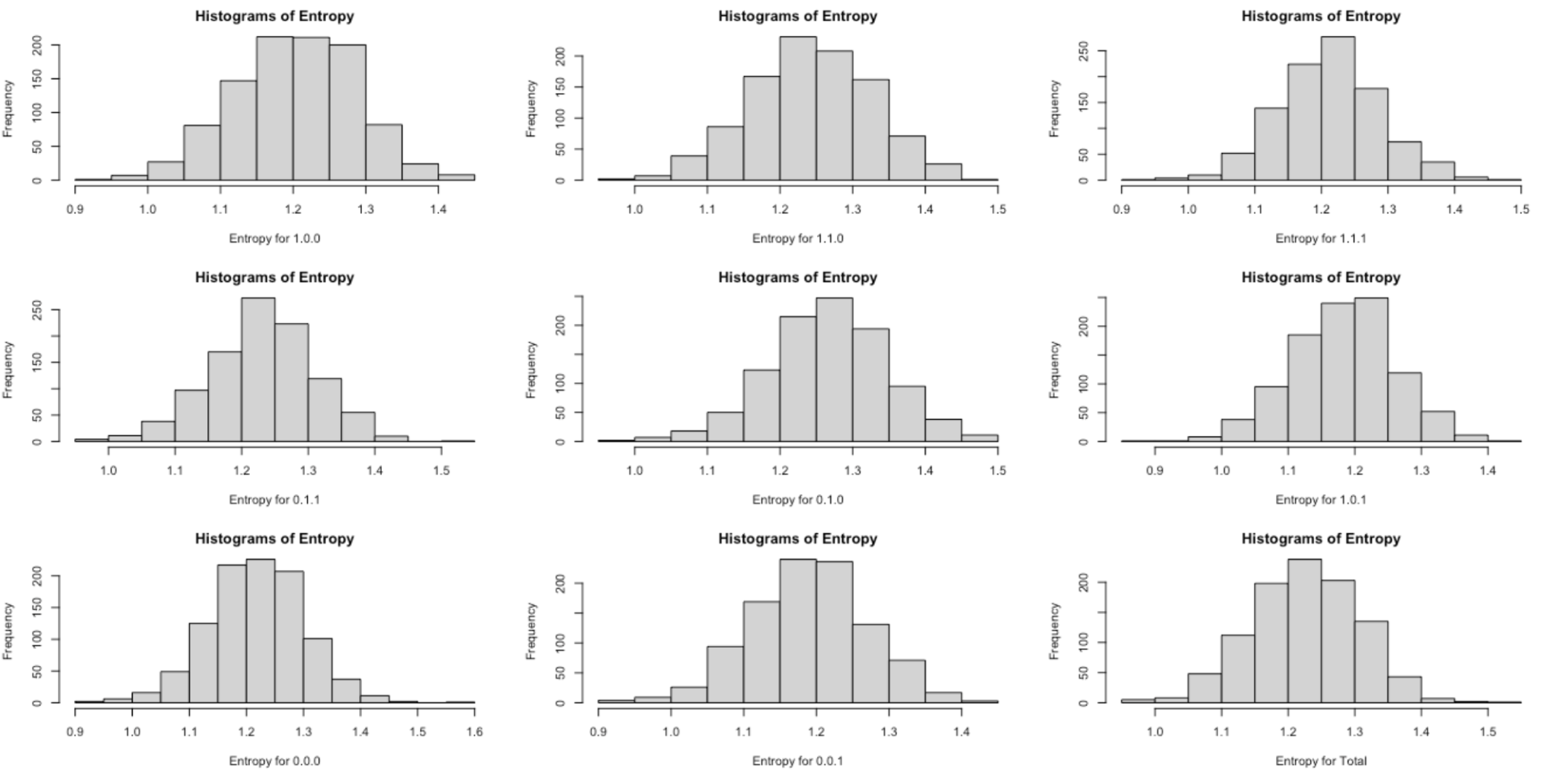
the most similar to the overall BMI distribution, as it shares a branch with the "Total" category. We should also consider the fact that these categories, 0.1.1, and 0.1.0 also are close to “Total” and also have significance when considering the relationship with BMI.

**Reliability with Hierarchical Clustering and Entropy on Simulated Data**

****

**Summary -**

To ensure that our HC Trees, also known as Dendograms, are reliable and are giving us readings that can be used to come to a conclusion, we simulated data and ran the same approach on this simulated data. Using multinomial distribution with n=10,000 from simulated data and compared them with the HC Trees obtained from our data on a subset of General Health=3. Looking at the 5 new HC Trees, we can observe that each of our “Total” values are paired on the same branch in each of the Dendograms. This means that even with simulated data, our results hold true. We can also confirm that on each Tree, our “Total” value shares the same branches throughout the simulation, which is the same result as our first HC Tree. Ultimately, this shows us that our method of HC Trees are reliable and yield consistent results throughout this category.

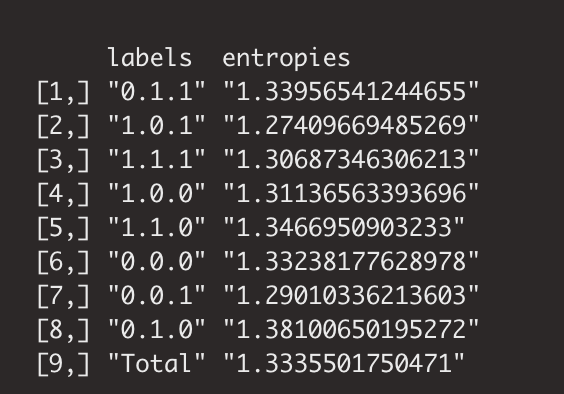
****

**Summary -**

To evaluate the reliability of our entropy approach for this sub-category, we generated 9 histograms of entropy using multinomial distribution with n=10,000 from simulated data and compared them with the true entropy values obtained from our data on a subset of General Health=3. Again, we can see that all 9 histograms seem to be normally distributed, which suggests that the entropy values were clustered around a specific mean. This indicates that the data has a moderate amount of uncertainty, making it useful for users as it is not too random or predictable. If the entropy histograms are skewed towards the left or right, it would mean the data is uninformative and unreliable due to either too much or too little information. Based on our observations, we conclude that all the interaction terms are informative and strong predictors of BMI within the subgroup data of General Health=3.

**Sub-Category #4 “Fair” General Health**

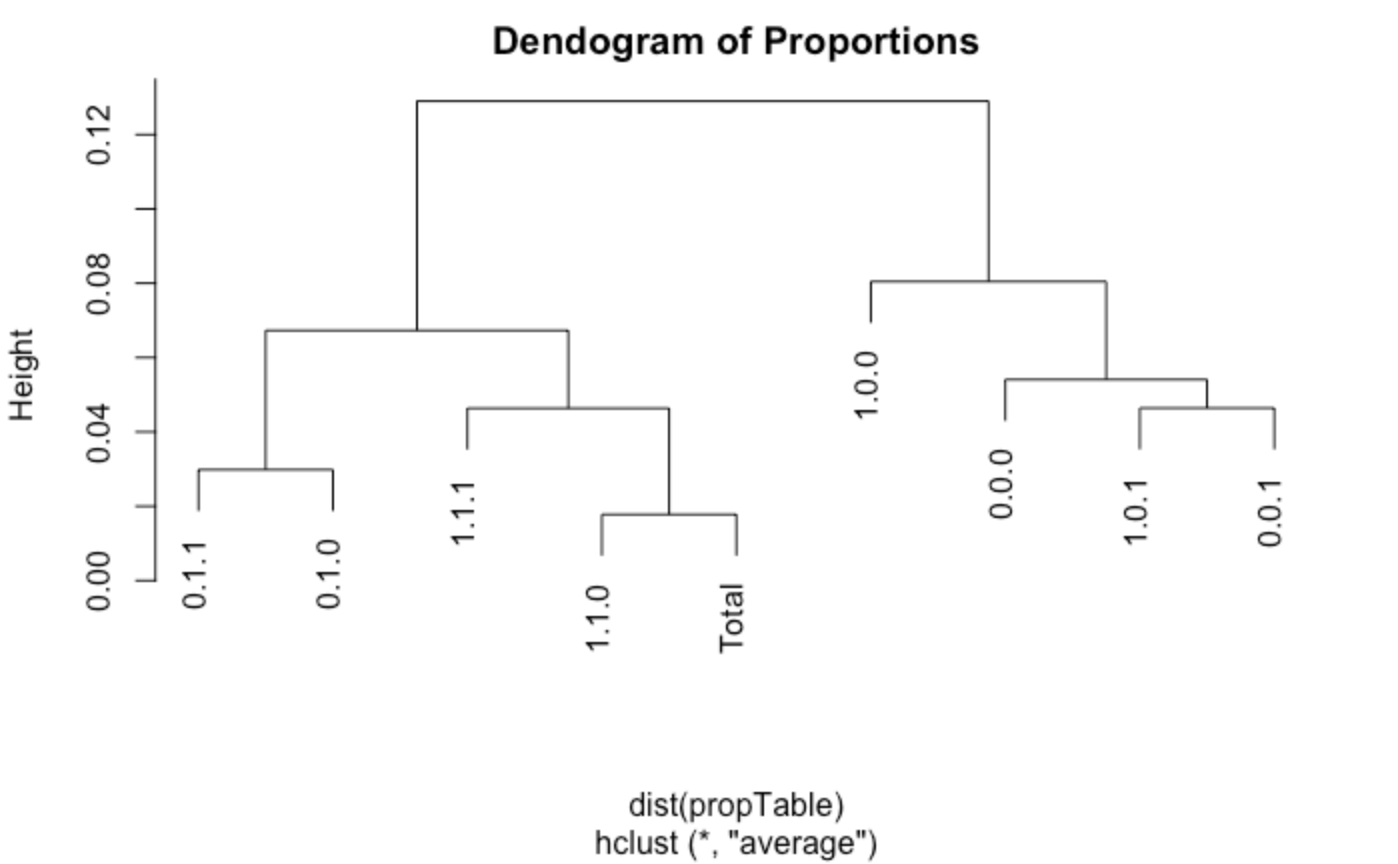
**Entropy**



**Summary -**

The entropy table shows the proportion of the total entropy that each interaction term accounts for, and a higher residual entropy indicates inadequate capture of essential information in the data, while a lower residual entropy indicates successful capture of the relevant information. According to the entropy values, the row which involves Smoker, Normal Blood Pressure, and High Cholesterol, [2] is the most effective predictor of BMI, with an entropy proportion of 1.274. Conversely, the interaction term which involves Not-Smoking, High Blood Pressure, and Normal Cholesterol, row [8] is the least effective predictor of BMI, with an entropy proportion of 1.381. A trend we can see as we go down the list of categories, and the overall level of General Health decreases, the entropy proportion increases in every row.

**Hierarchical Clustering**



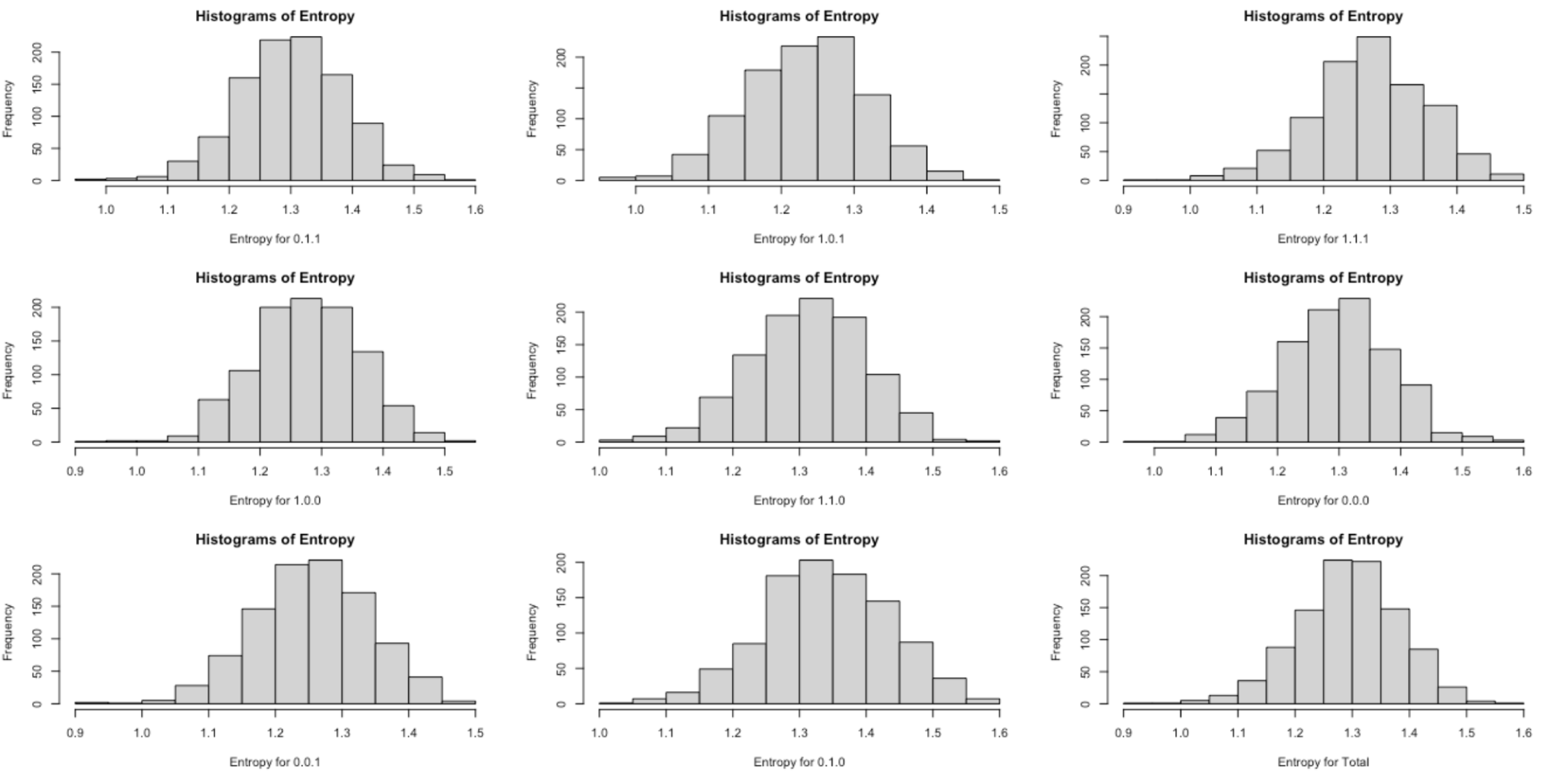
**Summary -** We constructed a dendrogram based on BMI, showing how the 8 subgroups cluster together. By focusing on the subset of individuals with a general health level of 4, we aimed to assess the community structures formed between samples and determine how similar their BMI distributions were to the overall BMI distribution, depicted by the "Total" category on the dendrogram.

Our analysis of the dendrogram revealed that the interaction effect of Smoking, High Blood Pressure, but Normal Cholesterol was the most similar to the overall BMI distribution, as it shares a branch with the "Total" category. Moreover, the interaction effect of Smoking, High BP, and High Cholesterol also exhibited a similar distribution to the overall BMI distribution.

**Reliability with Hierarchical Clustering and Entropy on Simulated Data**

**Summary -**

To ensure that our HC Trees, also known as Dendograms, are reliable and are giving us readings that can be used to come to a conclusion, we simulated data and ran the same approach on this simulated data. Using multinomial distribution with n=10,000 from simulated data and compared them with the HC Trees obtained from our data on a subset of General Health=4. Looking at the 5 new HC Trees, we can observe that each of our “Total” values are paired on the same branch in each of the Dendorgrams. This means that even with simulated data, our results hold true. We can also confirm that on each Tree, our “Total” value shares a branch with (1.1.0) which is the same result as our first HC Tree. Ultimately, this shows us that our method of HC Trees are reliable and yield consistent results throughout this category.

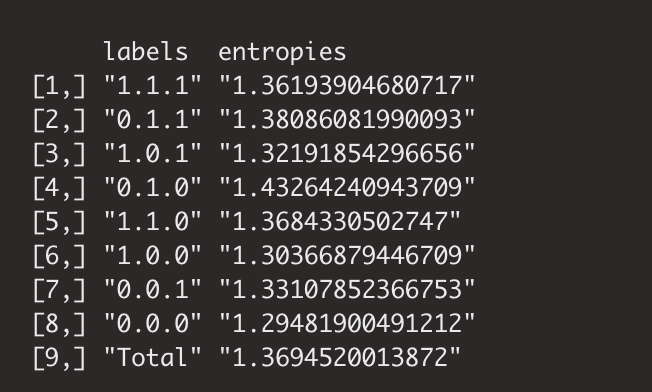
****

**Summary -**

To evaluate the reliability of entropy, we generated 9 histograms of entropy using multinomial distribution with n=10,000 from simulated data and compared them with the true entropy values obtained from our data on a subset of General Health=4. We observed that all 9 histograms had a normal distribution, which suggests that the entropy values were clustered around a specific mean. This indicates that the data has a moderate amount of uncertainty, making it useful for users as it is not too random or predictable. If the entropy histograms are skewed towards the left or right, it would mean the data is uninformative and unreliable due to either too much or too little information. Based on our observations, we conclude that all the interaction terms are informative and strong predictors of BMI within the subgroup data of General Health=4.

**Sub-Category #5 “Poor” General Health**

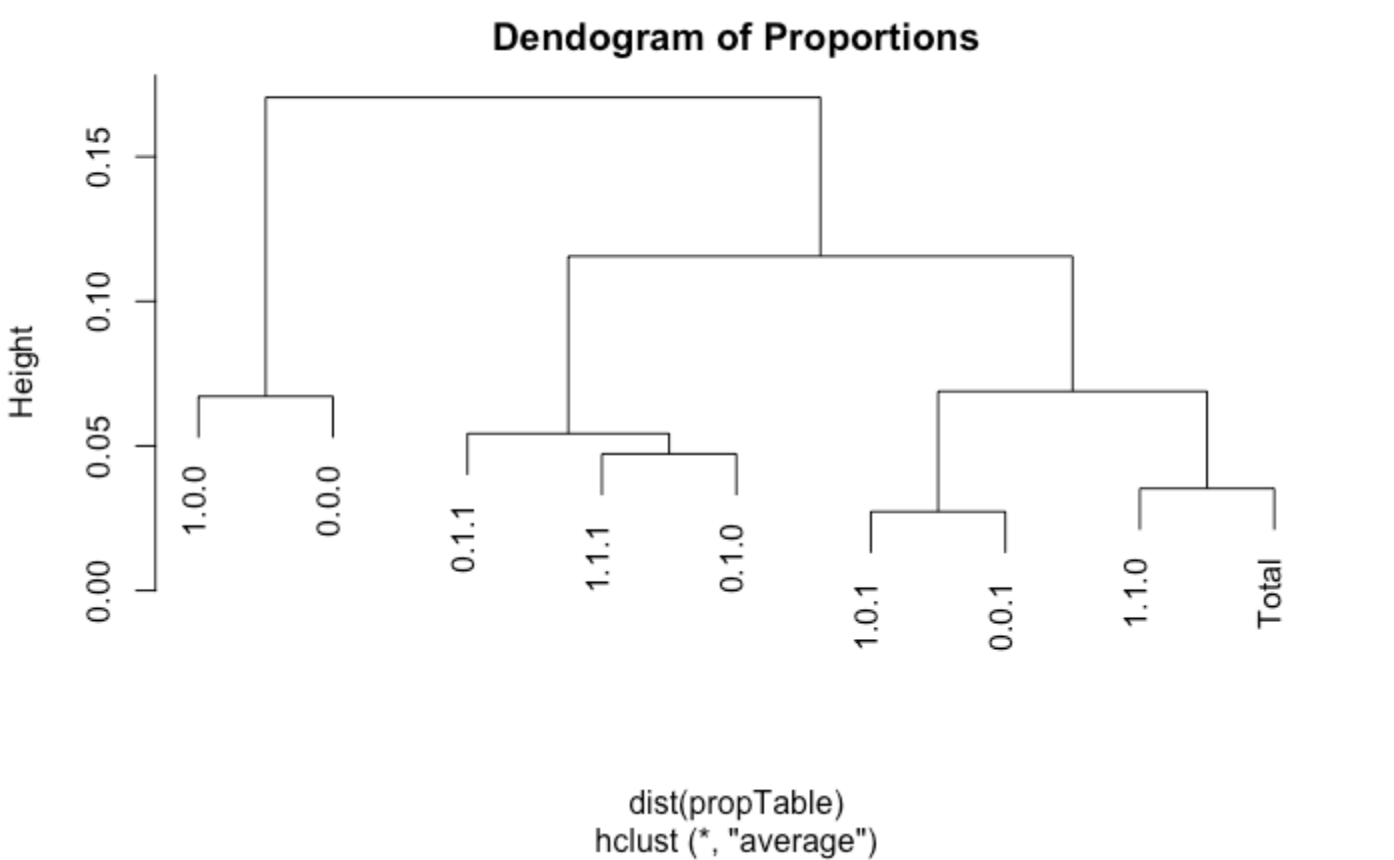
**Entropy**

****

**Summary -**

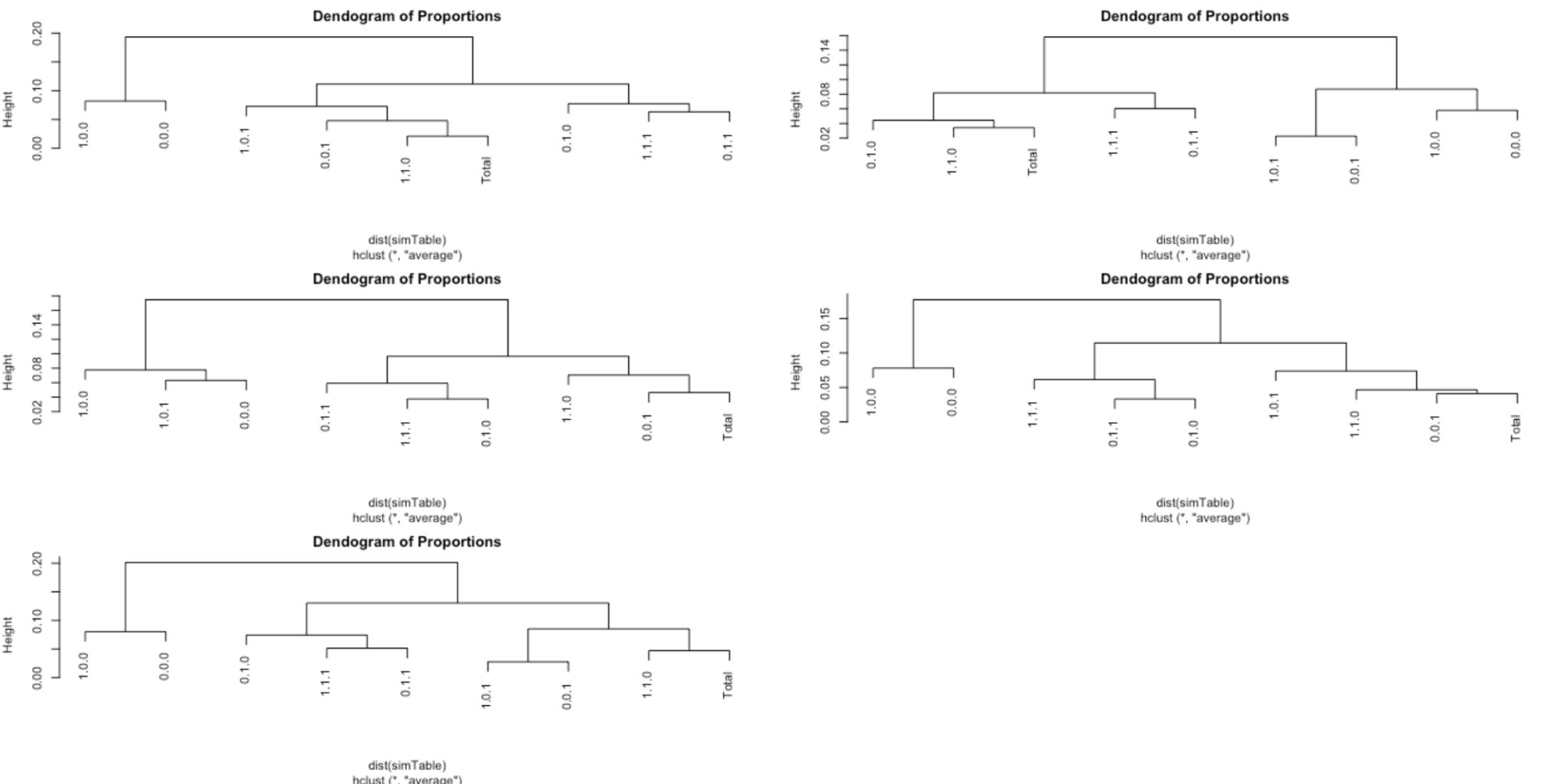
The entropy table shows the proportion of the total entropy that each interaction term accounts for, and a higher residual entropy indicates inadequate capture of essential information in the data, while a lower residual entropy indicates successful capture of the relevant information. According to the entropy values, the row which involves Non- Smoker, Normal Blood Pressure, and Normal Cholesterol, [8] is the most effective predictor of BMI, with an entropy proportion of 1.294. Conversely, the interaction term which involves Non-Smoking, High Blood Pressure, and Normal Cholesterol, row [4] is the least effective predictor of BMI, with an entropy proportion of 1.432.

**Hierarchical Clustering**

****

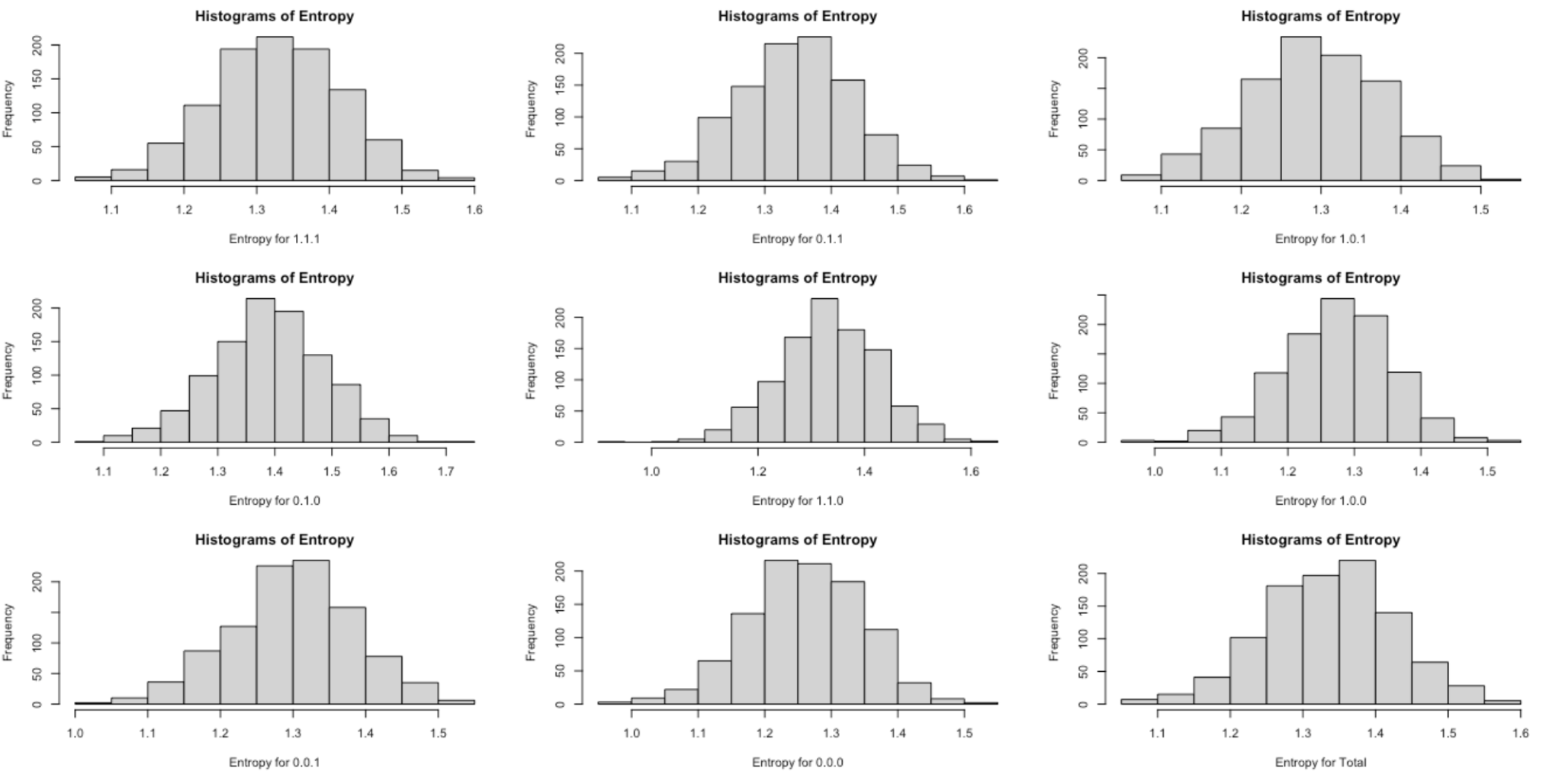
**Summary -** By focusing on the subset of individuals with a general health level of 5, we aimed to assess the community structures formed between samples and determine how similar their BMI distributions were to the overall BMI distribution, depicted by the "Total" category on the dendrogram.

Our analysis of the dendrogram revealed that the interaction effect of Smoking, High Blood Pressure, and Normal Cholesterol was the most similar to the overall BMI distribution, as it shares a branch with the "Total" category. Moreover, the interaction effect of (Smoking, No High Blood Pressure, and No High Cholesterol) and (Not Smoking,No High Blood Pressure, and No High Cholesterol) were the only two groups that did not share a branch with the “Total” branch, allowing us to conclude that these categories had the least to do with the the overall distribution of BMI.

**Reliability with Hierarchical Clustering and Entropy on Simulated Data**

**Summary -**

To ensure that our HC Trees, also known as Dendograms, are reliable and are giving us readings that can be used to come to a conclusion, we simulated data and ran the same approach on this simulated data. Using multinomial distribution with n=10,000 from simulated data and compared them with the HC Trees obtained from our data on a subset of General Health=5. Looking at the 5 new HC Trees, we can observe that each of our “Total” values are paired on the same branch in each of the Dendograms. This means that even with simulated data, our results hold true. We can also confirm that on each Tree, our “Total” value shares the same branches throughout the simulation, which is the same result as our first HC Tree. Ultimately, this shows us that our method of HC Trees are reliable and yield consistent results throughout this category.

****

**Summary -**

To evaluate the reliability of entropy, we generated 9 histograms of entropy using multinomial distribution with n=10,000 from simulated data and compared them with the true entropy values obtained from our data on a subset of General Health=5. We observed that all 9 histograms had a normal distribution, which suggests that the entropy values were clustered around a specific mean. This indicates that the data has a moderate amount of uncertainty, making it useful for users as it is not too random or predictable. If the entropy histograms are skewed towards the left or right, it would mean the data is uninformative and unreliable due to either too much or too little information. Based on our observations, we conclude that all the interaction terms are informative and strong predictors of BMI within the subgroup data of General Health=5.

**Conclusion**

After intensively working with the Entropy and HC Tree Approach with both Simulated Data and our Kaggle Data on Heart Disease, we can come to a few conclusions about the data, the methods used, and the reliability of these methods. After breaking down our Kaggle Data into 5 different sub-categories based on General Health, and using both of the previously mentioned approaches on each, we can say that both Entropy and HC Trees are a useful statistical analysis tool to find which of our binary categories **Smoking, High Blood Pressure, and High Cholesterol,** in which order have the highest link to Body Mass Index. After running these tests, we can conclude that for worse General Health, rankings including “1 and 2” the binary categories that have the highest correlation are Smoking and High Blood Pressure with Normal Cholesterol Levels, when looking at BMI. For better overall General Health, (categories 3,4,5) a variety of these categories seemed to have an effect. For example (0.0.0) or No Smoking, No High Blood Pressure, and No High Cholesterol, seemed to be the most reoccurring combination of the binary categories when looking at BMI.