Project 2: Cluster Analysis

PUMS Data

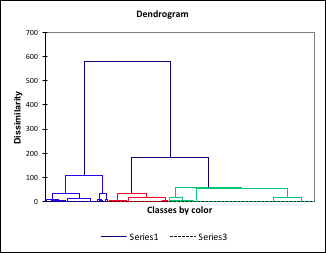
Katie Schreck, Cameron Dye, Hanna Cruzen, Qi Xue, Marina Dooley

Two numerical Variables: NP (number of persons in this household) on horizontal axis, BDSP (number of bedrooms) on vertical axis

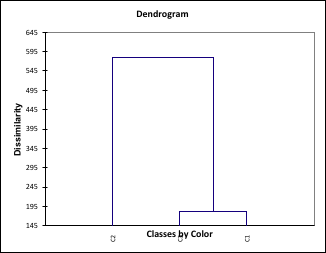
Method 1: Agglomerative Hierarchical Clustering (AHC), Euclidean Distance, Ward’s Method, Automatic-Entropy

(Marina)

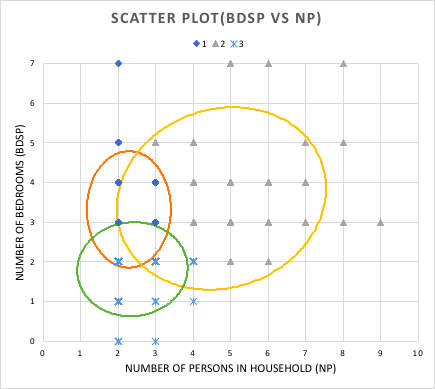
Full Dendrogram

****

Simplified Dendrogram



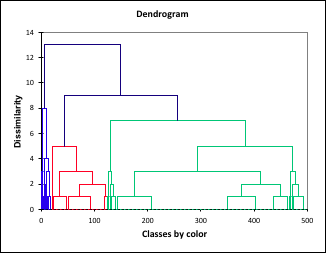
Scatterplot



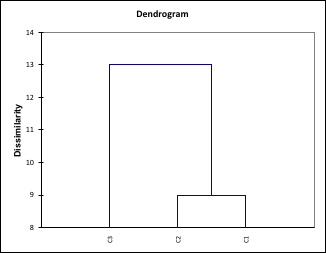
Method 2: Agglomerative Hierarchical Clustering (AHC), Manhattan Distance, Complete Linkage, Num. of Classes: 3

(Marina)

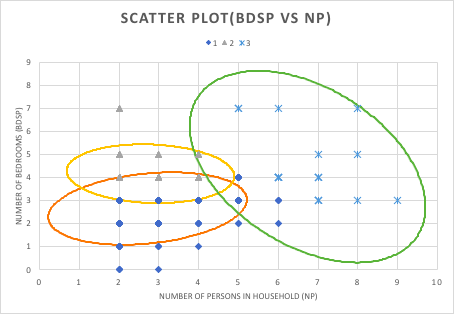
Full Dendrogram



Simplified Dendrogram



Scatterplot



Method 3: *k*-means Clustering, determinant clustering criterion, initial partition random, 50 repetitions, number of classes is 3

(Qi Xue)

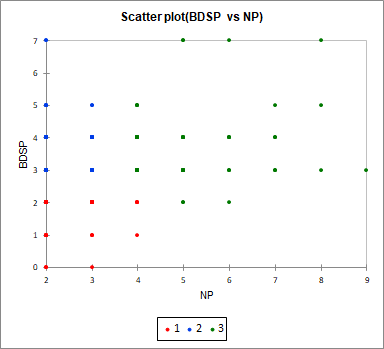


Figure 3.1

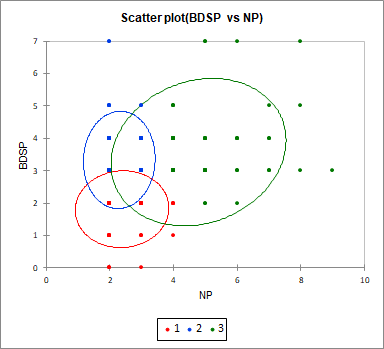


Figure 3.2

In this part, I used NP as the horizontal axis and BDSP as the vertical axis, and divided the group into 3 classes. In Figure 3.1, we see that class 1 is in a position with less NP and BDSP, while class 2 is in a position with less NP but more BDSP, and class 3 is in a position with more NP and BDSP. We can see that class 3 is the largest, while the number of those in the class 2 and class 1 groups is very close. Other information we can observe from Figure 3.2 is that these three classes have overlapping parts with each other. We can say that these three groups are similar in number of persons in this household and number of bedrooms. In addition, the overall trend in Figure 3.2 is positively correlated. Therefore, for most groups the more people that live in the household, the more bedrooms there are.

Method 4: *k*-means Clustering, trace clustering criterion, initial partition random, 50 repetitions, number of classes 4 (Hanna)

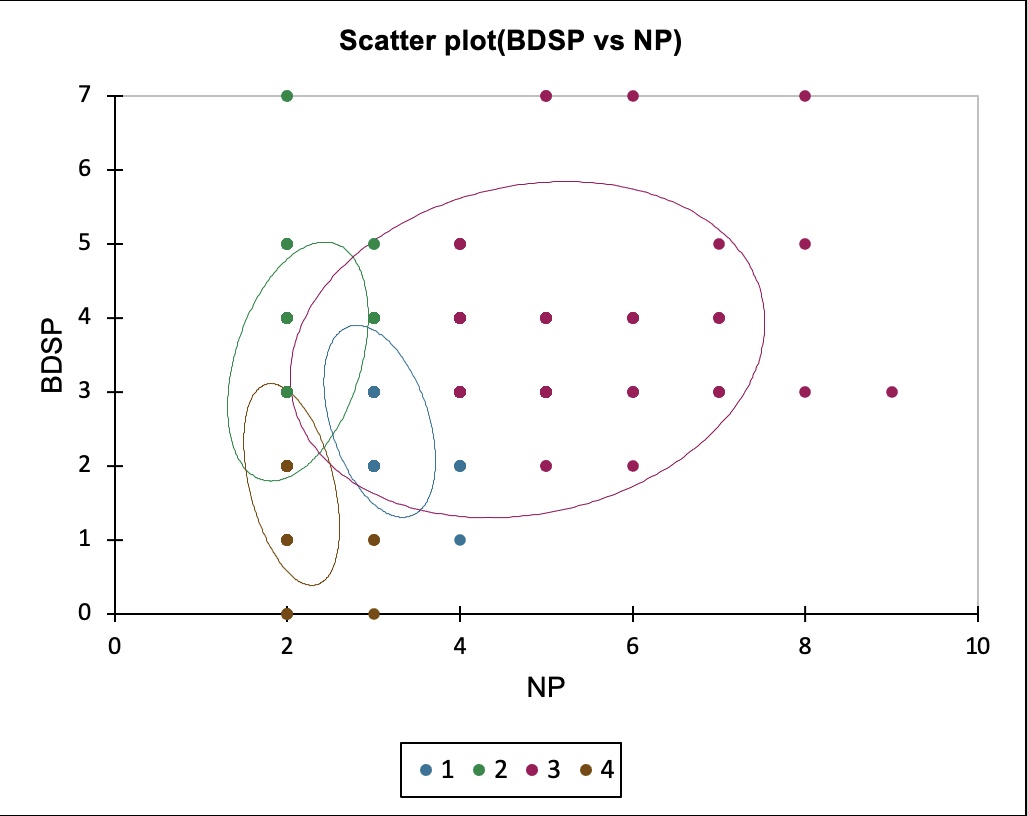


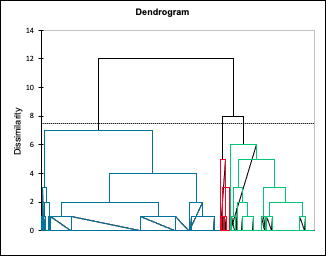
Figure 4.1

For this particular clustering method, we adjusted the Initial Partition to ‘Random’, Repetitions to ‘50’, and the Number of Classes to 4. Visually, it highlights the correlation between the number of persons in a household versus the number of bedrooms in a household. There is a general increase in the number of bedrooms as the quantity of persons within a household increases by an estimate of 1 additional bedroom per 2 additional persons. The outlier data shows a significant difference between the two variables, in which those with 7 or more bedrooms vary from having two persons per household, all the way to 8 persons per household. The majority of those featured within the data have 2-3 bedrooms in their household, with 3-5 persons living within them.

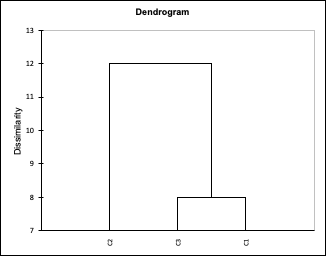
Method 5: choose your own

Agglomerative Hierarchical Clustering (AHC) using Manhattan Distance, Simple linkage, and 3 classes. (Cameron)

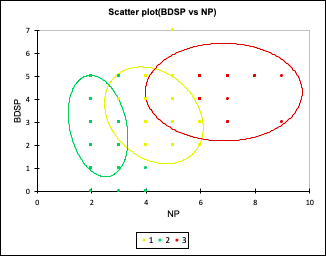
Full Dendrogram



Simplified Dendrogram



Scatter Plot



For this particular clustering method we used agglomerative hierarchical clustering (AHC), used simple linkage, and set the number of classes to 3. The data represented in the graphs visually shows the correlation between the number of people in one household and number of bedrooms per house. As we can see there is a general increase as the number of people in one household increases there is an increase in bedrooms in the house. There are however a few outliers in each group. As one data point in group three suggests that there are 9 people living in one house with 3 bedrooms.

Conclusion (Cameron)

All five methods prove that as the number of people per house increases, the amount of bedrooms in that house increases. This is visually represented through cluster analysis by using k-means clustering and Agglomerative Hierarchical Clustering (AHC) as well as different dissimilarities and linkages. Although all 5 methods proved to follow our initial assumptions, method 2 seemed to prove our assumptions the most. Each scatterplot included some outlier data and overlap between classes, however, method 2 showed clear distinctions between the groups and featured larger confidence ellipses (95% confidence) than other methods. The lack of overlap between groups serves to show the idea that larger households indeed have more bedrooms than those of smaller household sizes.

Furthermore, this data is interesting as it can help answer questions about average size of homes in Oregon, how many persons can live in houses with a certain number of rooms, and can help home builders and city planners determine the adequate amount and sizes of homes to build in Oregon. Clustering household size and the amount of bedrooms per home can help these two parties adjust their plans to put an emphasis on building larger or smaller homes. The data clearly shows that those who have larger families live in homes with a higher number of bedrooms, moreover, the larger class of data trends toward their being more large households than small or medium sized households. Outliers in small and medium classes can be useful in determining if a house is inadequately populated, further adding to the usefulness of the dataset and associated dendrograms and scatter plots.