MLA Assessment I

The following report utilises dimension reduction, clustering, and visual techniques to analyse 164 different restaurants. The analysis begins by exploring dimension reduction, and then follows on by performing hierarchical clustering. Finally, I use the results of the clustering analysis to draw conclusions about relationships between a restaurant being in the Michelin magazine or not. I also perform an ANOVA test and post-hoc Tukey (HSD) tests to test the significance of the different clusters formed by hierarchical clustering.

## Data description and visualisation

The dataset consists of 6 fields, 4 of which are numerical continuous customer ratings, one categorical variable which describes the restaurant name, and one binary variable indicating if the restaurant has appeared in the Michelin guide.

In Michelin is the binary variable. 1 indicating that the restaurant on that particular row has appeared in the Michelin guide, and 0 indicating that it has not.

Out of the 164 restaurants appearing in the dataset, 74 have appeared in the Michelin guide.

Restaurant name is the name of the restaurant and is a String value.

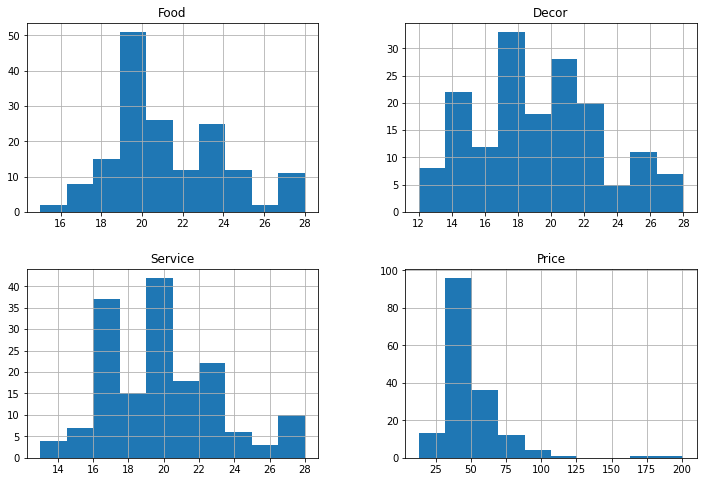
The four columns “Food”, “Decor”, “Service” and “Price” are ratings taken from reviews supplied by the Zagat survey which are contributed by customers. Each of these columns contains integer types which represent a rating for each restaurant under each criteria.

The maximum food rating a restaurant received was 28 and the minimum rating was 15, with a small standard deviation of 2.78.

The maximum Decor rating was 28 and the minimum rating was 12, with a larger standard deviation of 3.79. This implies that variables such as furniture and other decorations vary more across restaurants than does the actual quality of the food.

The maximum service rating was 28 and the minimum was 13, with a standard deviation of 3.26.

The maximum price was 201 USD and the minimum was 13 USD, with a standard deviation of 22.13 USD.



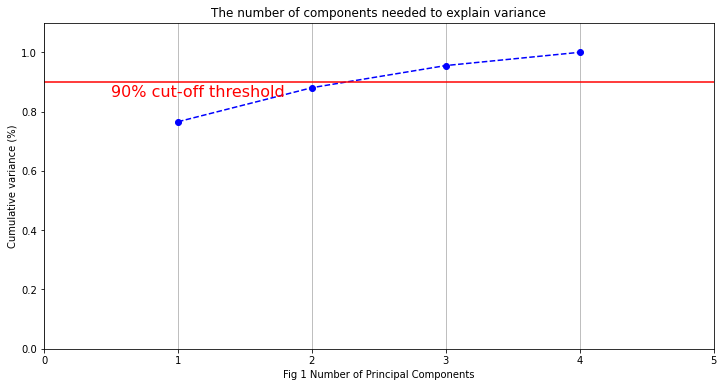
It is interesting to note that all of the 4 features roughly follow a normal distribution. The food and service distributions also are the most spread out, indicating that these variables vary most across restaurants. The price distribution is the least spread out, but has the greatest outliers, highlighting that most restaurants are similar in price, but some are way above the mean.

## Dimension Reduction

It’s likely the 4 continuous dimensions are not independent of each other, meaning that at least one pair of them will move somewhat in the same direction or are correlated. If this is the case, then we do not need all 4 of these dimensions to describe the data, as they will just be telling us the same thing. We can use a technique called Principal Component Analysis to reduce the dimensions of this dataset.

We generate linear combinations of the 4 numerical variables, rank them in order of how much variability they explain in the data, and then pick a certain number of these new Principal Components which will be less than the original number of dimensions and will explain a certain amount of variance in the dataset. Using this technique, we can project the original dimensions onto a lower dimensional space, and then use this newer lower dimensional space to visualise the data and as input for statistical models. There is no one rule for picking the number of Principal Components to use, but a good rule of thumb is to use a scree plot, which will show when adding more principal components becomes less effective. Another way to pick the number of principal components is to decide on how much variance you want to have explained beforehand, and then use the number of principal components that have a cumulative variance of this number decided beforehand.

To determine the number of principal components to use, I first generate 4 principal components and graph them to show how much cumulative variance they explain.

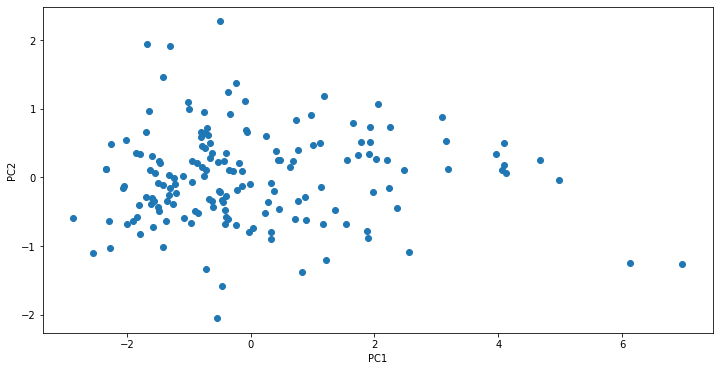


As you can see, almost 90% of the data is explained by just two principal components. This means we can project the original 4 dimensions space onto 2 dimensional space, and still capture most of what the data is saying. This makes the data easier to visualise, and to use as input to machine learning models.

| % Cumulative variance explained by PCs | |
| --- | --- |
| PC1 | 76.58 |
| PC2 | 88.06 |
| PC3 | 95.53 |
| PC4 | 99.99999999999999 |

It is obvious that 100% of the variance is explained by 4 principal components, because there were originally 4 dimensions describing the data. What we are more interested in is finding the trade off between a small number of principal components and a large amount of explained variance.

For this dataset, we are going to use PC1 and PC2, which together explain over 88% of the variance in the data. This will allow us to graph the data on two dimensional space.

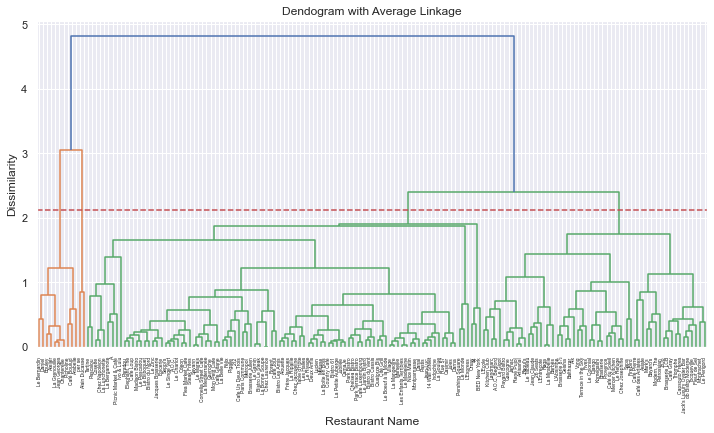


# Hierarchical Clustering

Hierarchical clustering seeks to reveal inherent groupings within a dataset by iteratively forming a tree-like structure, where the most similar observations are progressively merged. The primary objective of this analysis is to identify clusters of observations in the data, where observations within the same cluster exhibit similarities, while those in separate clusters demonstrate dissimilarities. The process starts by assigning each of the 164 restaurants to a different group. Then the two least dissimilar restaurants are placed into the same group, leaving a total of 163 groups. This process is repeated recursively until there is only one group left. The recursive tree of this process is called a Dendogram, and it is useful for examining the natural clustering of groups

In order to decide on which restaurants should go into which groups, measures of dissimilarity are used. Dissimilarity measurements show how different two observations are when they are plotted and there are different measures of dissimilarity. Since the four survey rating dimensions are continuous variables, I will use Euclidean distance as a measure of dissimilarity.

After assessing dissimilarities among a subset of observations through a dissimilarity matrix, these dissimilarities were visualised on a Dendrogram to explore their clustering patterns. The average linkage method was utilised.



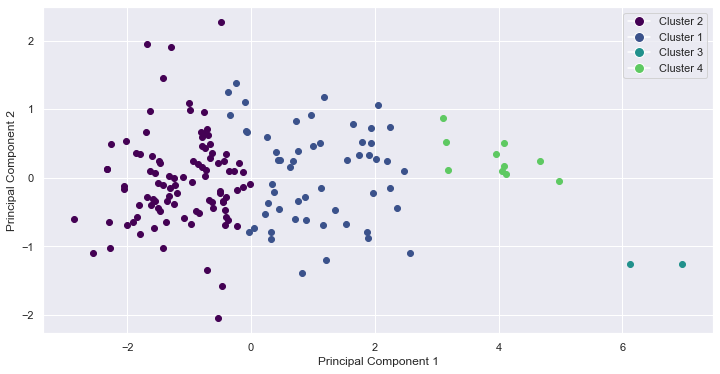
From the above Dendrogram using average linkage, we use the recommended cut off point of h + 3sh where h is the mean height at which the groups are joined, and sh is the standard deviation of these same heights. Using this cutoff point we can see that hierarchical clustering naturally results in 4 groups. We can visualise the results of assigning each observation to one of these groups using a pairwise scatter plot matrix.



From the scatter plot matrix, we can see that there is a good separation of different clusters, and they do not overlap too much. This grouping seems satisfactory, and there is no need to use a different clustering technique such as K-means, where we would decide beforehand on the number of clusters to use.

Now that we have reduced the dimensions of the data, and used this lower dimensional data to form different clusters using hierarchical clustering with average linkage, we can now try to interpret what it means for a restaurant to be in a particular group, and see if there are any relationships of a restaurant being in a group and it also having appeared on the Michelin guide.

## Analysing the clusters



This scatterplot shows all the restaurant observations plotted on the two dimensional space of PC1 and PC2. We will now analyse this graph to see what proportion of restaurants in each group has appeared in the Michelin guide.

| Cluster | Percentage in Michelin |
| --- | --- |
| 1 | 76% |
| 2 | 23% |
| 3 | 100% |
| 4 | 80% |

Careful analysis of the results in the table above along with the scatterplot reveals that interestingly, as points move further to the right on the graph, a higher percentage of them are included in the Michelin guide.

Cluster 2, which is most far to the left, only has a 23% inclusion rate in the Michelin guide.

Cluster 1, which is next furthest from the left, has a 76% inclusion rate.

Cluster 4, which is second closest to the right, has a 80% inclusion rate.

Finally, cluster 3, which is furthest to the right, has a 100% inclusion rate in the Michelin guide.

It would be possible to take new restaurant observations, including the four numerical features of food, decor, service, and price, with unknown knowledge about whether it had appeared on the Michelin guide or not, and use the clusters above along with an algorithm such as K nearest neighbours, to work out what group the unlabelled data should belong to. You could then use the percentages of Michelin inclusion from the current groups to assign a probability to the unlabelled data as the probability that it has been included in the Michelin guide.

We can also perform an ANOVA (analysis of variance) test to check if the clusters are in fact statistically different.

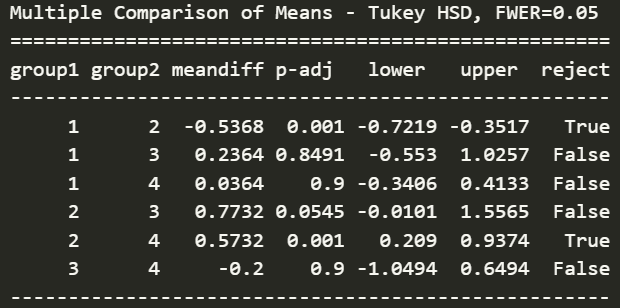
Null Hypothesis (H0): There is no significant difference in the Michelin inclusion rates among the four clusters.

Alternative Hypothesis (H1): At least one of the clusters has a significantly different Michelin inclusion rate.

The p-value from this ANOVA test is 3.102796177722307e-12 which is <0.05. Hence we can reject the null hypothesis, and conclude that at least one pair of groups differ from each other.

As a follow up, we can also perform a post hoc test to determine which specific pairs of clusters statistically differ from each other. We can use Tukey’s Honestly Significant Difference (HSD) to do this.

The results of running this test are as follows.



From these results we can conclude that groups 1 and 2 are statistically different from each other, along with groups 2 and 4.

This suggests that if we were to replicate this clustering on a similar dataset or new observations, we would likely observe similar distinctions between these pairs of groups, highlighting the robustness and reliability of the clustering approach.