TIM 245 - Data Mining: Homework #3

Due: *June 13, 2017*

Instructor: Tyler Munger

Panos Karagiannis
ID: -

Contents

Problem 1: Clustering **3**

Problem 2: Association Analysis **10**

Problem 1: Clustering

Using the dataset (`survey_dataset.csv`) select one category, e.g. music preferences, to focus on for the *clustering*. Use the selected category and the demographic information to answer the following clustering questions:

1. Based on the results, which clustering method do you recommend using for the data-set? Explain why.
2. How many clusters did you find in the dataset? How did you select the method parameters, k or epsilon?
3. Explore the cluster results using Open-Refine or Excel. Pick 2-3 clusters and try to generalize, i.e. create a persona, for the people (instances) in the cluster.
4. Describe how personas from the clustering results could potentially be used?

Answer:

This project investigates the clustering results when the `selected_category` is `movies`. Bullet point (1) includes the answer to both questions (1),(2).

1. In order to find the best performing clustering algorithm, given our dataset, we experiment with various clustering models and values of the parameters and report the models for which we obtained the **highest silhouette score**. We also produce plots by mapping high dimensional data into $2D$ data using *T-SNE*.

   (a) The first method that we explore is *K-Means*, which, is a fast and simple approach to clustering which works well for globular clusters.

Figure 1: Sum of squared errors as we vary the number of clusters in the K-Means algorithm.

Using the "elbow method" we speculate that a useful clustering occurs for a number of clusters between 2 and 10 (Fig. 1). In order to assess our model we compute the

silhouette score for all the values of $k = 1, 2, \ldots, 40$ and report the top 4 models in terms of silhouette score.

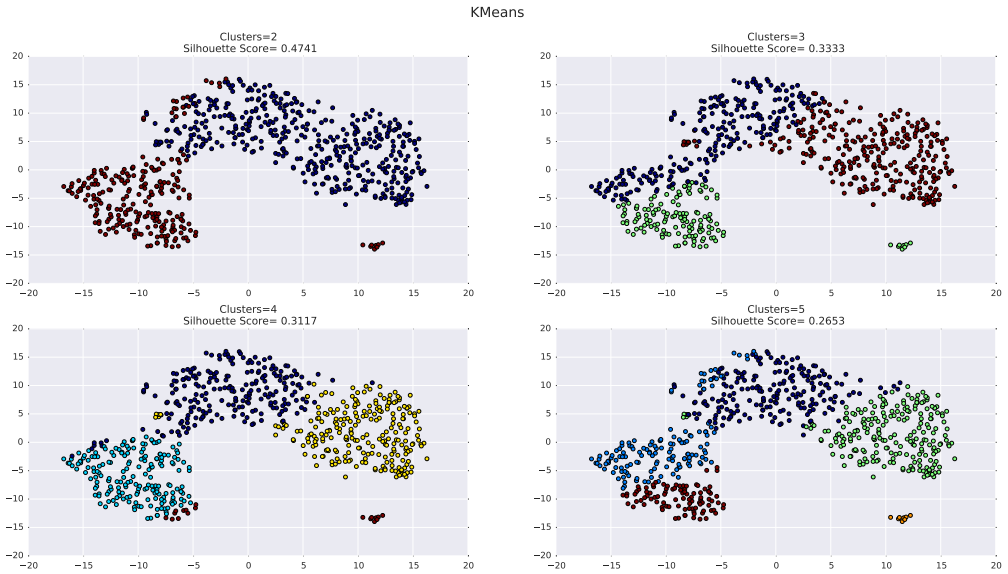

Figure 2: Four best K-Means models as we vary $1 \leq k \leq 40$.

As we see from Figure 2, the best silhouette score is achieved for $k = 2$ and is equal to 0.4741. Nevertheless, this silhouette score is still not very close to 1, therefore, as we see next, K-Means is not the best performing clustering algorithm.

(b) The second clustering algorithm that we explore is *Hierarchical Agglomerative Clustering*. In this type of clustering we are given more "freedom" than K-Means, in the sense that we can also choose the *linkage* type. More precisely, for Agglomerative Clustering the cost function (linkage) is the distance between clusters. In this project we experiment with all three common linkage methods, namely:

- Single Linkage: The distance between two clusters is the *minimum* distance of their respective points
- Complete Linkage: The distance between two clusters is the *maximum* distance of their respective points
- Average Linkage: The distance between two clusters is the *average* distance of their respective points

More precisely, for each linkage type we vary the number of clusters k such that $k = 1, 2 \ldots, 40$ **but only report the four values of k for which we calculated the largest silhouette score.**

As we see from Figures 3, 4, 5, single linkage performs the worse over the three hierarchical clustering methods. Complete linkage performs marginally better than average linkage, but again we see that the number of clusters that achieve the high-

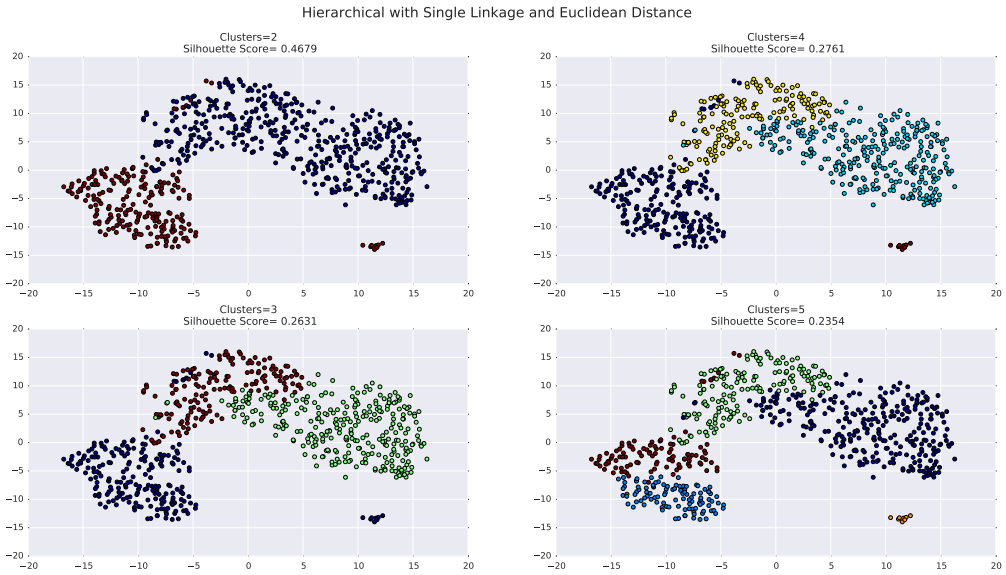

Figure 3: Four best Single Linkage models as we vary $1 \leq k \leq 40$.

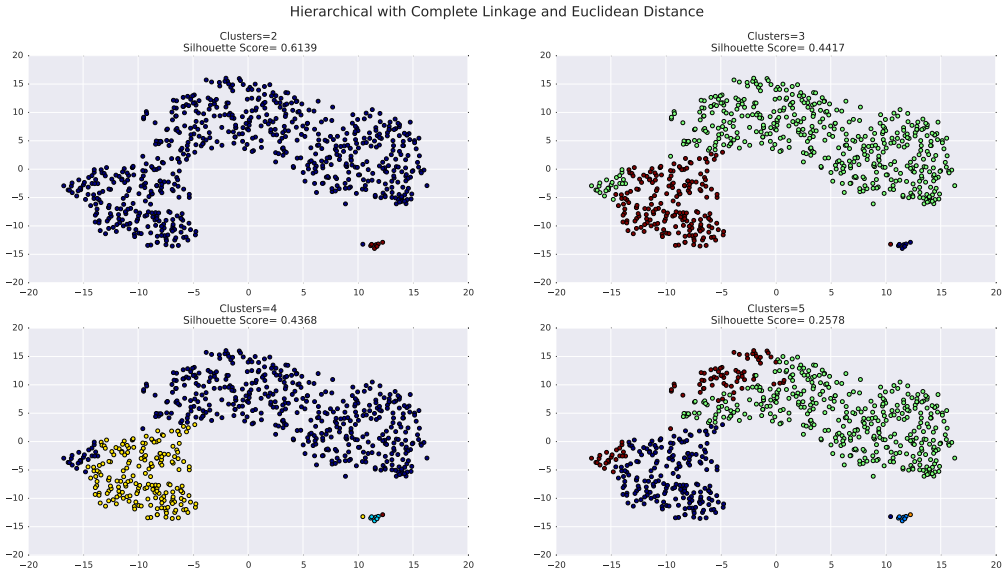

Figure 4: Four best Single Linkage models as we vary $1 \leq k \leq 40$.

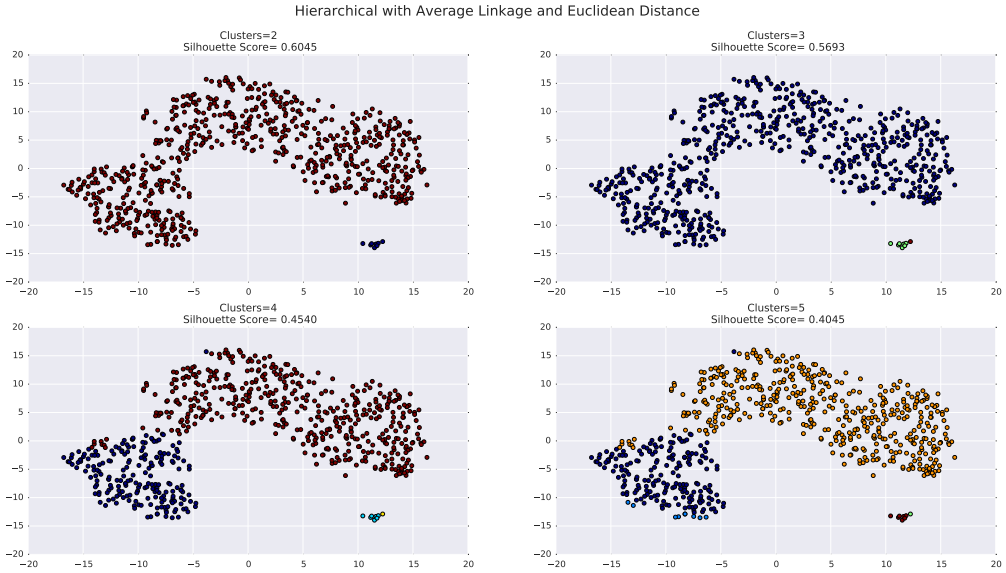

Figure 5: Four best Single Linkage models as we vary $1 \leq k \leq 40$.

est silhouette score is $k = 2$.

Also, in Figure 6 we see the that if we let the euclidean distance be greater than 200 we get two clear clusters.

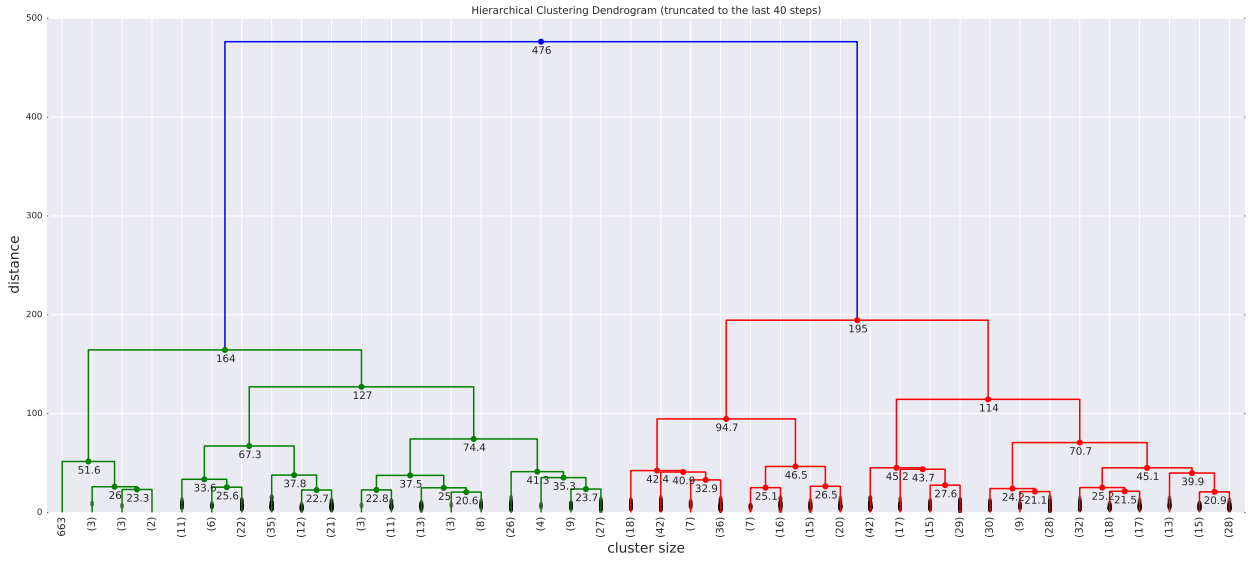

Figure 6: Dendrogram for Complete Linkage

(c) The next clustering method that we explore, is *DB-Scan*. In contrast to the previously seen clustering algorithms, in this case we do not need to supply the number of clusters, as the algorithm finds the number of clusters based on ϵ and

$MinPoints$. Heuristically, we found that letting $MinPoints > 5$ always resulted in the same clustering where only one cluster was present. Therefore, we vary $1 \leq MinPoints \leq 5$ and we let $\epsilon = 0.2 + 0.1t$ for $t = 0, 1, \ldots, 34$. As we varied $MinPoints$ between 1 and 5 we observed no significant changes in the value of the silhouette score. Nevertheless, the value of ϵ played a crucial role in getting a good clustering model. In fact, by looking at Figure 7 we see that for $\epsilon = 2$ we get two clusters and a very high silhouette score, whereas for $\epsilon = 5$ the silhouette score is negative.

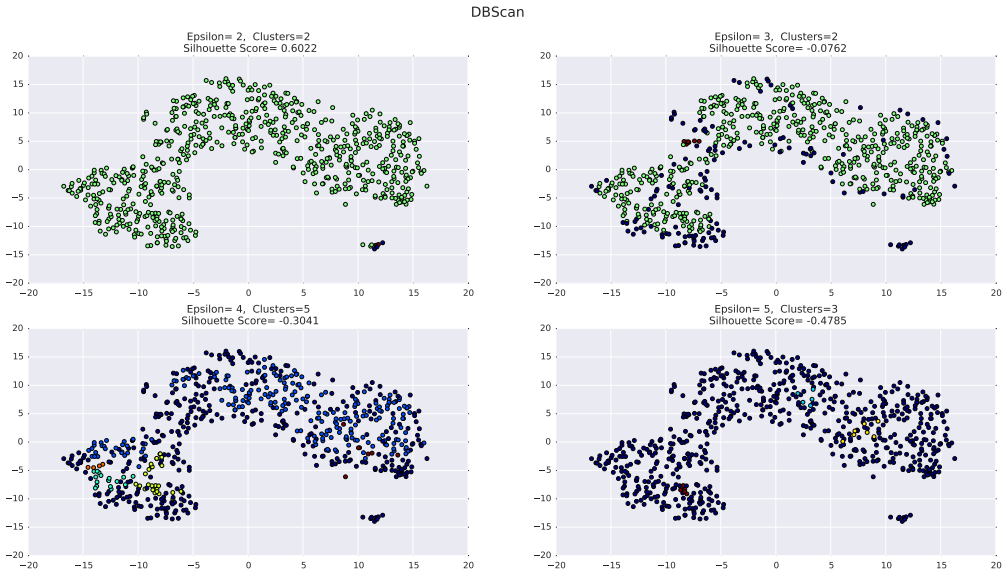

Figure 7: Four best Single Linkage models as we vary $1 \leq k \leq 40$ and fix $MinPoints = 5$.

(d) Finally, we try the *Expected Maximization* algorithm where we use Gaussian Priors. We see that this model performs very poorly compared to other clustering techniques (Fig. 8). According to this clustering model, the number of clusters that achieve the highest silhouette score (0.1216) is 29, which is very different than the number of clusters suggested by our previous analysis.

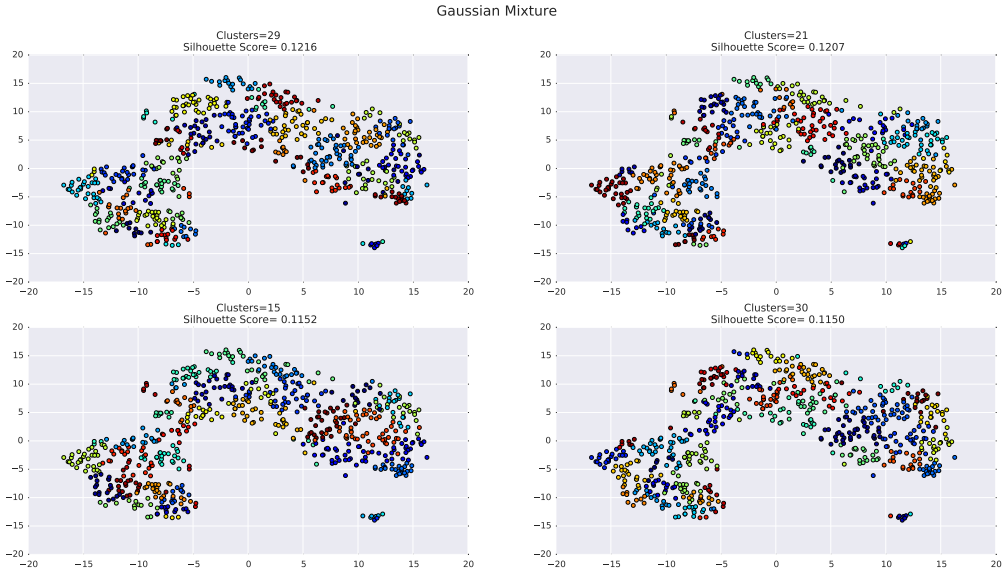

Figure 8: Four best Single Linkage models as we vary $1 \leq k \leq 40$ and fix $MinPoints = 5$.

$\Longrightarrow$ Given the above analysis, we could say that one of the best clustering methods that achieves the highest silhouette score is **Agglomerative Hierarchical Clustering with Complete Linkage and number of clusters equal to two.** This method achieves a silhouette score of 0.61, nevertheless, DB-Scan also performs very similarly achieving a silhouette score of 0.60 for $\epsilon = 2$ and total number of clusters equal to two.

2. By doing a first analysis of the data, we see that many different algorithms suggest that the most likely number of clusters is two. Investigating further, we see that in the first cluster belong only 8 people, whereas, is in the second cluster belong around 1000 people.

   By looking into the smaller of the two clusters, we see that the attributes with 0 variance are: `Education_college/bachelor degree`, `Education_currently a primary school pupil`, `Education_primary school` , `Only child_no`. This means that **all** people belonging in this cluster are currently attending *primary* school and are not *only children*. As far as other attributes are concerned, we see that in this cluster all people like *Thriller* and *Western* films and are *left handed*, since these are the attributes with the least variance ($\sigma^2 = 0.28, 0.68, 0.23$, respectively).

   In the second cluster, we see that the vast majority of people really *likes movies* ($\sigma^2 = 0.40$) such as *comedies* ($\sigma^2 = 0.50$). Also, most people are *right handed* ($\sigma^2 = 0.08$), and currently do not go to *primary school* ($\sigma^2 = 0.002$).

3. Some of the characteristics drawn from our clustering are not significant since they follow normal trends. For example most people in the country have attended primary school and are right handed. **Nevertheless**, our clustering has highlighted certain character

traits that could be used in order to perform direct marketing and advertisement. For example, in the smaller cluster we see that most people are left handed, have siblings and also like Thrillers and Animated films. Similarly, it seems that right handed people opt for *Comedies*. Even though many exceptions exists, and this is not an infallible trend, it gives us a vague first idea about how we could approach potential costumers. We know that within the student body, there exist two "groups" namely, one that mostly likes *Comedies* and another one that mostly likes *Thrillers and Animated films*.

Extra Credit:
An alternative distance metric that we could use in order to perform clustering and measure the silhouette score is the *chebysev* distance. By definition:

$$D_{CHEB}(\vec{d_1}, \vec{d_2}) = \max \left| \vec{d_1}(i) - \vec{d_2}(i) \right|$$

This is a metric considers only one component in calculating the distance, hence we would expect to give us different results.

In order to see the effect of the distance metric in the silhouette score we first need to rigorously define it. In order to compute this score, we first need to define quantities a_i, b_i corresponding at point i of our dataset. More precisely, let the set $\mathcal{D}$ represent the set of all data points and let $C_1 \dots C_k$ be the sequence of clusters, with $|C_j|$ being the total number of points in $C_j, 1 \leq j \leq k$. Then, for a point $i \in \mathcal{D}$, let $i \in C_j$. We define:

$$a_i = \frac{1}{|C_j|} \sum_{e \in C_j} D(i, e)$$

, where $D(x, y)$ is an arbitrary distance function with $D(x, x) = 0$ (in our case the chebysev distance). Intuitively, a_i is the average distance of i to all other points contained in the same cluster C.
Now let:

$$b_i = \min_w \left\{ \frac{1}{|C_w|} \sum_{e \in C_w} D(i, e) \right\}_{w=1, w \neq j}^{k}$$

Finally we can compute the silhouette score as:

$$silhouette = \frac{1}{n} \sum_{i=1}^{n} \frac{b_i - a_i}{\max(a_i, b_i)}$$

Given the above definitions, we compute the highest silhouette score that each clustering method achieved, after varying the number of clusters from 0 to 40:

Table 1: Reporting the Silhouette Score for all clustering models.

Clustering Model	Silhouette Score	Clusters
K-Means	0.50	2
Single Linkage	0.49	2
Complete Linkage	0.67	2
Average Linkage	**0.66**	2
DB-Scan ($\epsilon = 2$)	**0.66**	2
Gaussian Mixture	-0.02	5

From the above table, we see that using this distance metric, the silhouette score is higher for all clustering methods (except Gaussian Mixture), but the number of predicted clusters remains 2, as in the case of the euclidean distance.

A possible reason for the similarity of the performance of the clustering models when we use different distance metrics, could be attributed to the fact that the maximum distance between two points "dominates" the euclidean metric.

Problem 2: Association Analysis

Using the same dataset as in Problem 1, answer the following questions:

1. Experiment with different values for support and confidence. How do the discovered patterns change? What threshold do you recommend using for support and confidence?
2. Identify 2-5 interesting rules generated using your selected support and confidence threshold. What is the interpretation of the rule? What is underlying rationale or reason for the rule, e.g. the diapers $\rightarrow$ beer rule was because young fathers were sent to the store to buy diapers.
3. What are some of the potential applications of the generated rules? For example, how could the rules be used for marketing products to students?

Answer:

1. Based on the current code framework, the association rule formulation relies on the support and confidence measures to eliminate uninteresting patterns using the *Apriori* algorithm. This method has two drawbacks, namely:
   - It is often the case that many potentially interesting patterns involving low support items might be eliminated by the support threshold.
   - The drawback of confidence is more subtle. Assume for example we have an association rule $\{\texttt{X}\}\rightarrow\{\texttt{Y}\}$ with very high confidence, 0.80. The problem is that the confidence measure does not take into consideration the support of the itemset $\{\texttt{Y}\}$. Therefore, we could construct an example where $P(\{\texttt{Y}\}) \leq P(\{\texttt{Y}\}|\{\texttt{X}\})$, while, the confidence of the rule to be high! Thus, we would consider this to be a good rule, nevertheless, knowing that $\{\texttt{X}\}$ occurs, actually decreases the probability of $\{\texttt{Y}\}$ occurring.

Many alternative methods have been proposed to solve this problem such as: *Interest Factor, Correlation Analysis, IS Measure* etc. In this project we use the *Intertest Factor* approach to evaluate the association rules. More precisely, we evaluate the *interestingness of a pattern*, by defining *Lift*:

$$Lift = \frac{c(\{\texttt{X}\} \rightarrow \{\texttt{Y}\})}{s(\{\texttt{Y}\})} = \frac{\sigma(\{\texttt{X}\}, \{\texttt{Y}\})}{\sigma(\{\texttt{X}\})\sigma(\{\texttt{Y}\})}$$

, which addresses the previously discussed problem by taking into consideration the support of $\{\texttt{Y}\}$. From this formulation it is clear that a $Lift$ value equal to 1 implies the statistical independence of $\{\texttt{X}\}$, $\{\texttt{Y}\}$. Therefore, we should choose rules with $Lift \gg 1$.

For example if we, naively, attempt to find rules that have both high support and confidence ($s = 0.7, c = 0.8$) we get:
`right handed` $\rightarrow$ `Music=5`, which has $Lift \approx 1$, therefore it is a bad rule.

Also, setting both support and confidence to low values ($s = 0.1, c = 0.2$) produces very obvious results with high $Lift$ values:
`right handed,Only.child=yes, city` $\rightarrow$ `Number.of.siblings=0`. The above pattern is achieves a $Lift$ score of 5, nevertheless, it is not informative since we already know that an only child has no siblings.

After experimentation, the values that yield the most informative results are $\sigma \approx 0.1$ and $c \approx 0.5$.

2. Some of the most interesting rules that we see are :

   - `Biology=5` $\rightarrow$ `Chemistry=5`: This rules achieves a Lift score of 5.2, and it implies that if a persons likes Biology then he is also very likely to like Chemistry. This is a logical rule since these two fields are closely related.
   - `Fun.with.friends=5, Shopping.centres=5, Gender=female` $\rightarrow$ `Shopping=5`: This rules achieves a Lift score of 4.3, and it implies that females who like shopping centers and also to have fun with friends, are also very probable to like shopping. The underlying rational behind this rule is, that if a female likes spending time in shopping centers, then she is also likely to start shopping.
   - `Internet=5, Gender=male, city` $\rightarrow$ `PC=5`: Achieving a Lift score of 4.1, this rule implies that a male who likes internet and lives in the city, is also very likely to like PCs. This is also a logical rule, since the internet and PCs have an immediate connection.

3. All of the aforementioned rules, could be used by companies for direct advertising. For example, if a company like Amazon observes that a student has bought a Biology book then it can suggest a Chemistry book to the student as well. Also, if a female

likes visiting shopping centers and being around friends, then a company could offer free shopping coupons to that student in order to make some product more enticing. Finally, companies in big cities, such as Xfinity, that know that a particular male consumer likes browsing on the internet, can offer to this student specialized PC offers in order to attract him.