

MGMT320  
Data Clustering  
Assignment 3

Aashish Agrawal  
(SID- 44009240)

# Data Handling

## Data Cleaning

The original data that was provided to us had a lot of inconsistencies with some values not under the scale of 1-5. So, I decided to have any cells with the value of 0 as missing value and for the cells with values between 0 to 1, we do not do anything with it and assume that these results were produced due to unspecified scaling. Further, we created a dummy data in order to compare with the original data and provide a deeper analysis. For the dummy data, we replace the value of 0 with 1. All data is already in numeric form, so no further manipulation is required. We analyse the two data sets separately and compare appropriately based on manipulated variables.

## Original Data

	Unique user id	Churches	Resorts	Beaches	Parks	Theatres	Museums	Malls	Zoo	Restaurants
1	User 1	0.00	0.00	3.63	3.65	5.00	2.92	5.00	2.35	2.33
2	User 2	0.00	0.00	3.63	3.65	5.00	2.92	5.00	2.64	2.33
3	User 3	0.00	0.00	3.63	3.63	5.00	2.92	5.00	2.64	2.33
4	User 4	0.00	0.50	3.63	3.63	5.00	2.92	5.00	2.35	2.33
5	User 5	0.00	0.00	3.63	3.63	5.00	2.92	5.00	2.64	2.33
6	User 6	0.00	0.00	3.63	3.63	5.00	2.92	5.00	2.63	2.33
7	User 7	0.00	5.00	3.63	3.63	5.00	2.92	3.30	2.35	2.33
8	User 8	0.00	5.00	3.63	3.63	5.00	2.92	5.00	2.63	2.33
9	User 9	0.00	5.00	3.64	3.64	5.00	2.92	3.30	2.62	2.32
10	User 1	0.00	5.00	3.64	3.64	5.00	2.92	5.00	2.35	2.32
11	User 11	0.00	0.53	3.65	3.67	5.00	2.92	5.00	2.61	2.32
12	User 12	0.00	0.53	3.65	3.68	5.00	2.93	5.00	2.61	2.31
13	User 13	0.00	0.54	3.66	3.68	5.00	2.93	5.00	2.61	2.30
14	User 14	0.00	0.54	3.66	3.67	5.00	2.93	5.00	2.32	2.30
15	User 15	0.00	0.53	3.67	3.66	2.95	2.93	5.00	2.33	2.31
16	User 16	0.00	0.52	3.69	3.66	2.95	2.93	5.00	2.98	2.31
17	User 17	0.00	0.52	3.68	3.66	2.96	2.93	2.96	2.98	1.70
18	User 18	0.00	0.53	3.69	3.66	2.95	2.93	2.95	3.00	1.70
19	User 19	0.00	0.52	5.00	3.66	2.96	2.93	2.95	2.99	1.70
20	User 2	0.00	5.00	3.70	3.66	2.95	2.93	2.94	2.99	1.70
21	User 21	0.00	0.51	5.00	3.67	2.94	2.93	2.95	2.98	1.70
22	User 22	0.00	5.00	5.00	3.66	2.94	2.93	2.94	3.00	1.70
23	User 23	0.00	5.00	5.00	3.66	2.95	2.93	2.94	3.00	1.70
24	User 24	0.00	0.51	5.00	3.66	2.95	2.94	2.95	2.97	1.71
25	User 25	0.00	0.51	0.52	3.66	2.96	2.94	2.94	2.63	1.71
26	User 26	0.00	0.52	0.52	3.66	2.95	2.94	2.94	2.96	1.71
27	User 27	0.00	0.55	0.52	3.66	2.95	2.94	2.94	2.63	1.72

## Dummy Data

	Unique user id	Churches	Resorts	Beaches	Parks	Theatres	Museums	Malls	Zoo	Restaurants
1	User 1	1.00	1.00	3.63	3.65	5.00	2.92	5.00	2.35	2.33
2	User 2	1.00	1.00	3.63	3.65	5.00	2.92	5.00	2.64	2.33
3	User 3	1.00	1.00	3.63	3.63	5.00	2.92	5.00	2.64	2.33
4	User 4	1.00	0.50	3.63	3.63	5.00	2.92	5.00	2.35	2.33
5	User 5	1.00	1.00	3.63	3.63	5.00	2.92	5.00	2.64	2.33
6	User 6	1.00	1.00	3.63	3.63	5.00	2.92	5.00	2.63	2.33
7	User 7	1.00	5.00	3.63	3.63	5.00	2.92	3.30	2.35	2.33
8	User 8	1.00	5.00	3.63	3.63	5.00	2.92	5.00	2.63	2.33
9	User 9	1.00	5.00	3.64	3.64	5.00	2.92	3.30	2.62	2.32
10	User 1	1.00	5.00	3.64	3.64	5.00	2.92	5.00	2.35	2.32
11	User 11	1.00	0.53	3.65	3.67	5.00	2.92	5.00	2.61	2.32
12	User 12	1.00	0.53	3.65	3.68	5.00	2.93	5.00	2.61	2.31
13	User 13	1.00	0.54	3.66	3.68	5.00	2.93	5.00	2.61	2.30
14	User 14	1.00	0.54	3.66	3.67	5.00	2.93	5.00	2.32	2.30
15	User 15	1.00	0.53	3.67	3.66	2.95	2.93	5.00	2.33	2.31
16	User 16	1.00	0.52	3.69	3.66	2.95	2.93	5.00	2.98	2.31
17	User 17	1.00	0.52	3.68	3.66	2.96	2.93	2.96	2.98	1.70
18	User 18	1.00	0.53	3.69	3.66	2.95	2.93	2.95	3.00	1.70
19	User 19	1.00	0.52	5.00	3.66	2.96	2.93	2.95	2.99	1.70
20	User 2	1.00	5.00	3.70	3.66	2.95	2.93	2.94	2.99	1.70
21	User 21	1.00	0.51	5.00	3.67	2.94	2.93	2.95	2.98	1.70
22	User 22	1.00	5.00	5.00	3.66	2.94	2.93	2.94	3.00	1.70
23	User 23	1.00	5.00	5.00	3.66	2.95	2.93	2.94	3.00	1.70
24	User 24	1.00	0.51	5.00	3.66	2.95	2.94	2.95	2.97	1.71
25	User 25	1.00	0.51	0.52	3.66	2.96	2.94	2.94	2.63	1.71
26	User 26	1.00	0.52	0.52	3.66	2.95	2.94	2.94	2.96	1.71
27	User 27	1.00	0.55	0.52	3.66	2.95	2.94	2.94	2.63	1.72

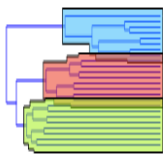
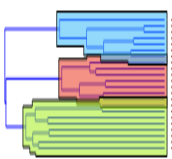
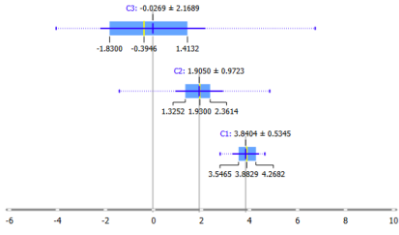
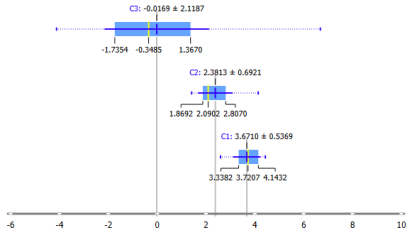
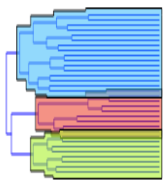
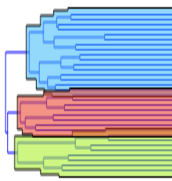
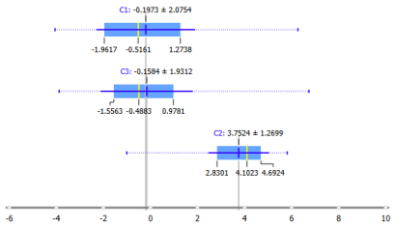
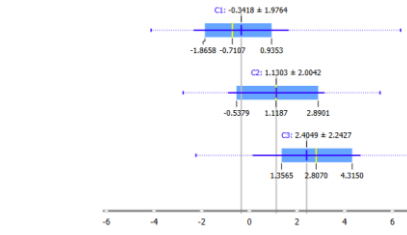
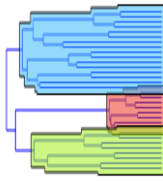
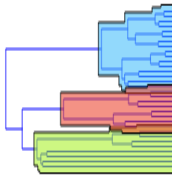
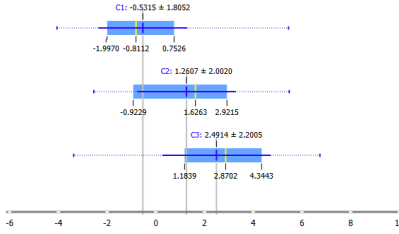
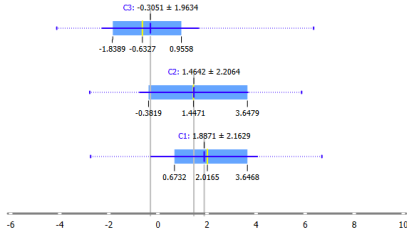
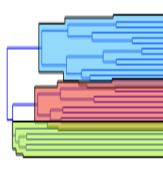
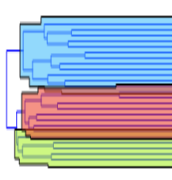
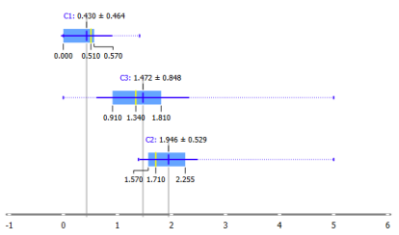
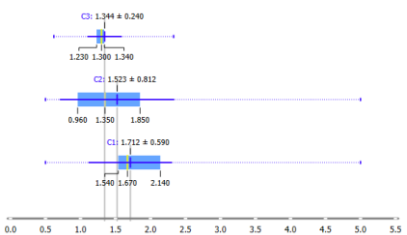
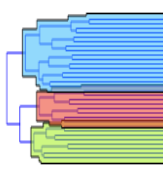
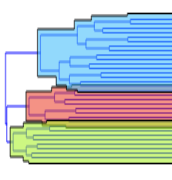
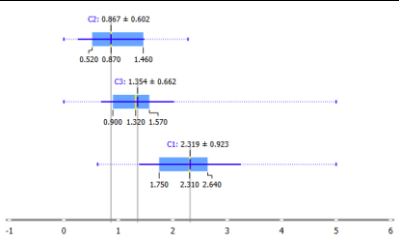
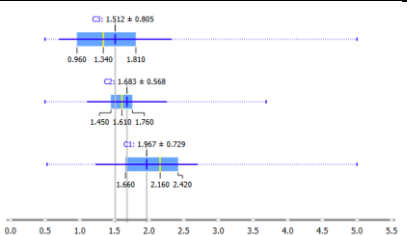
# Clustering Method

## Hierarchical Cluster Analysis

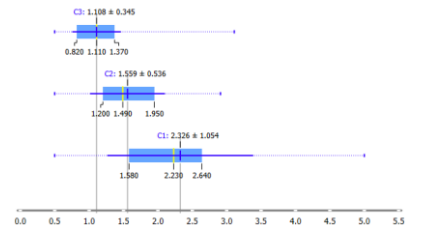
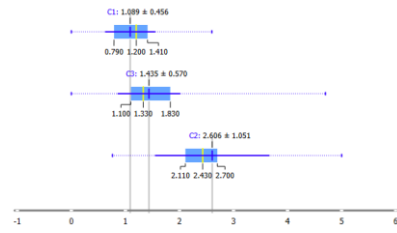
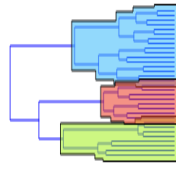
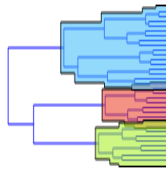
The below images are of hierarchical clustering with 3 linkages shown, i.e. average, complete and ward. In this case, we experimented with the original 23 variables along with PCA treated variables. The clustering's have been pruned to a depth of 5 group per cluster which would allow for a feasible visual analysis and a boxplot comparison for PC1 and Churches.

The difference is already noted for each method, with the original and dummy having different sizes on their cluster, albeit PCA treated is still similar. Additionally, the boxplot shows the difference in the range of values and spread albeit the dummy data still has a tighter spread as it treated data replacing missing values with 1 which makes it more constant.

	Original	Dummy	Boxplot Original	Boxplot Dummy
8 PCA Average				
8 PCA Complete				
8 PCA Ward				

13 PCA Average				
13 PCA Complete				
13 PCA Ward				
23 Average				
23 Complete				

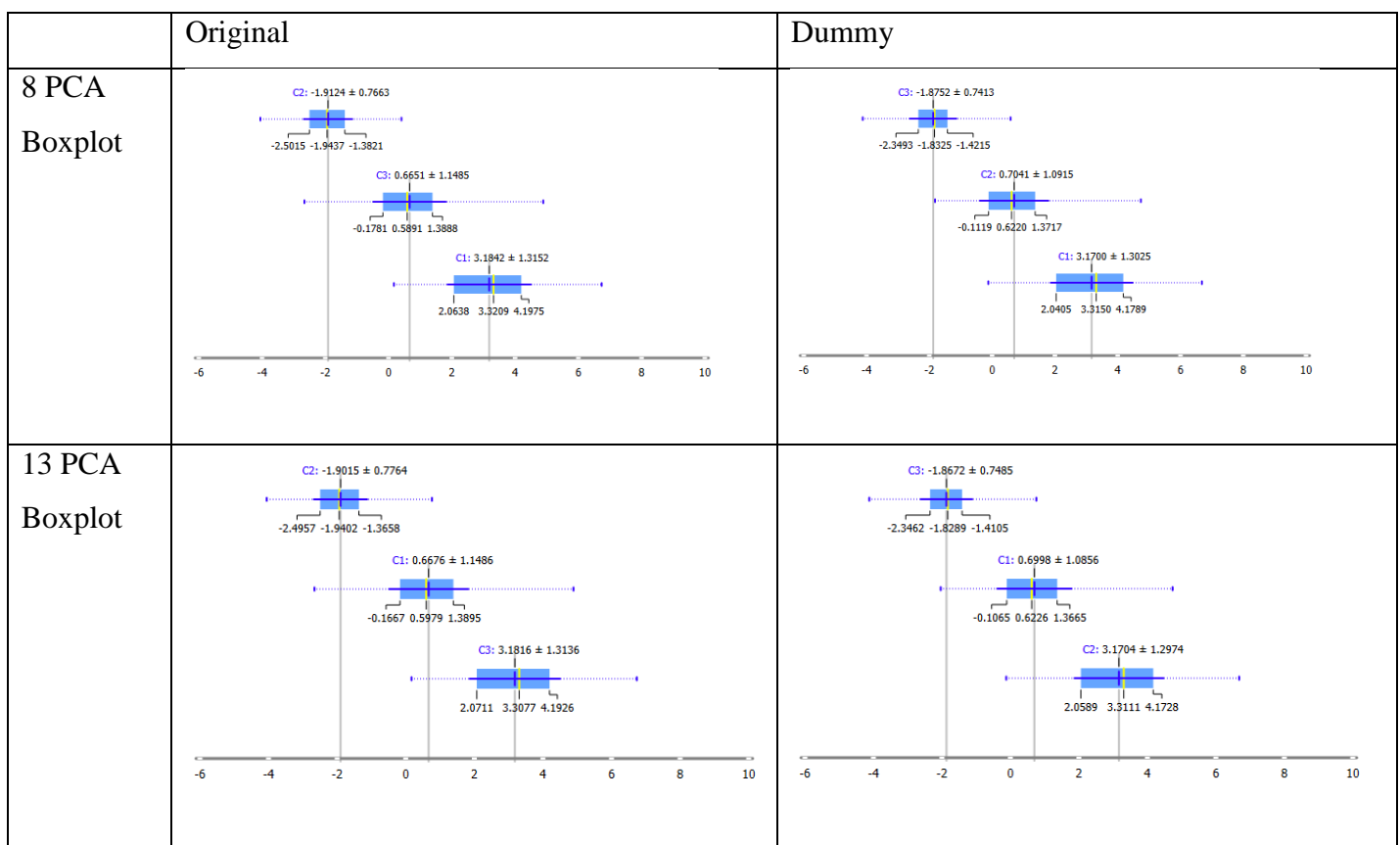
23 Ward

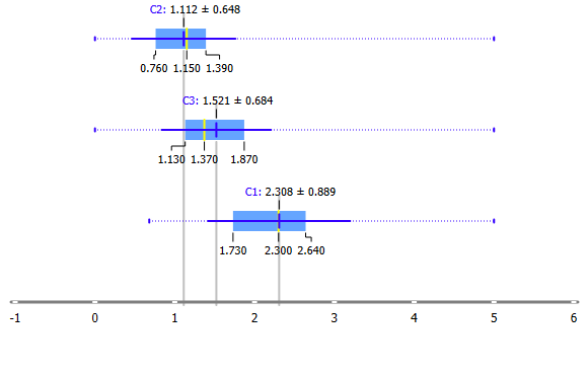
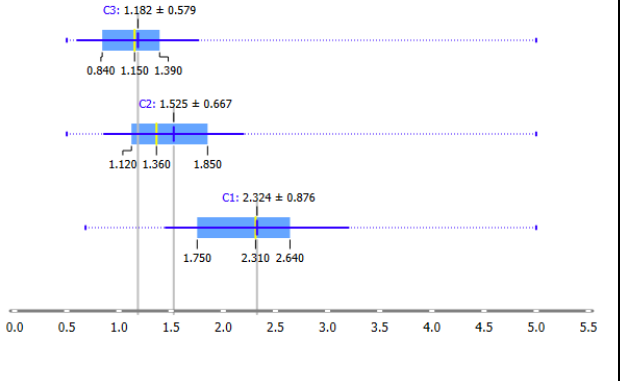
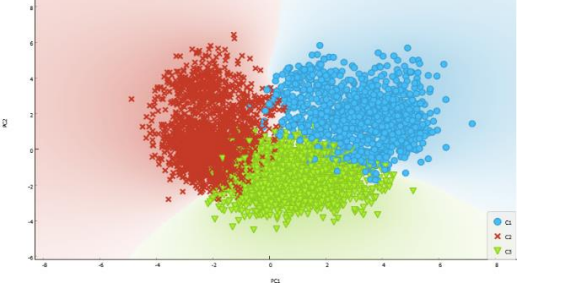
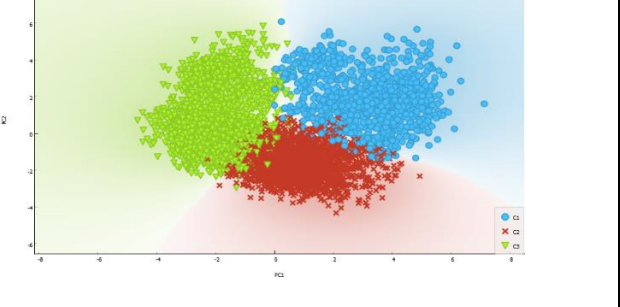
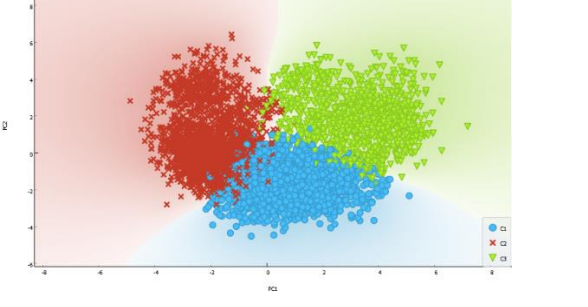
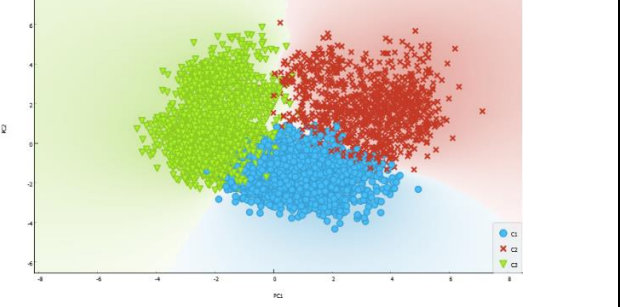
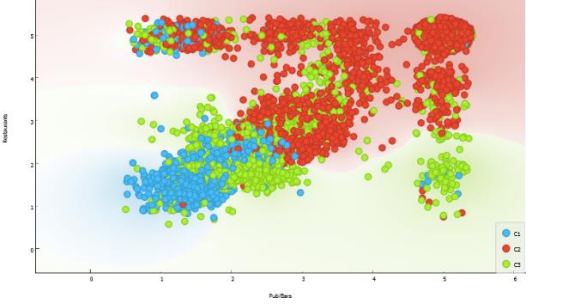
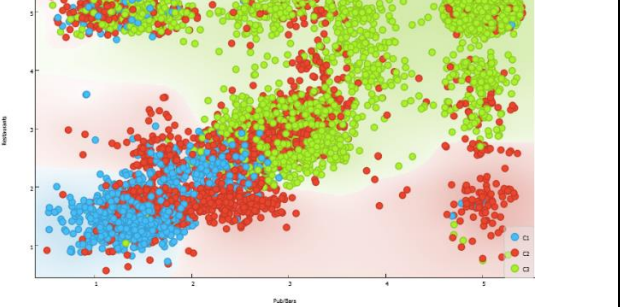


## K-Means Cluster Analysis

Below we experiment clustering using K-means with the 2 datasets (original data and dummy data) and set the cluster at 3 for consistency while doing the analysis. We used PCA along with the 23 variables on both data set to see whether the clustering are going to vary between them. We then drew a boxplot of PC1 to PC2 and Pubs to Restaurant to see the difference.

After performing the cluster analysis, we notice that PCA treated and non-treated dataset looks pretty much similar irrespective of whether it's original or the dummy one, which can be observed through the boxplot with the values being similar in range like each other. But, we also notice that the boxplots show the dummy variables has less spread which could be due to the random variables having a set value of one when compared to the original data where we mentioned it as a missing value. K-means doesn't show clustering well enough with the difference is due to a difference in treatment of the original data.



<p>23</p> <p>Variables</p> <p>Boxplot</p>	 <p>Boxplot showing the distribution of 23 variables for three clusters (C1, C2, C3). The x-axis ranges from -1 to 6. The clusters are defined by their means and standard deviations: C1: 2.308 ± 0.889, C2: 1.112 ± 0.648, and C3: 1.521 ± 0.684. The plot shows the median, quartiles, and range for each cluster.</p>	 <p>Boxplot showing the distribution of 23 variables for three clusters (C1, C2, C3). The x-axis ranges from 0.0 to 5.5. The clusters are defined by their means and standard deviations: C1: 2.324 ± 0.876, C2: 1.525 ± 0.667, and C3: 1.182 ± 0.579. The plot shows the median, quartiles, and range for each cluster.</p>
<p>8 PCA</p> <p>Scatterplot</p>	 <p>PCA scatterplot showing the distribution of 8 variables for three clusters (C1, C2, C3). The x-axis is PC1 and the y-axis is PC2. The clusters are defined by their means and standard deviations: C1: 2.308 ± 0.889, C2: 1.112 ± 0.648, and C3: 1.521 ± 0.684. The plot shows the distribution of data points for each cluster.</p>	 <p>PCA scatterplot showing the distribution of 8 variables for three clusters (C1, C2, C3). The x-axis is PC1 and the y-axis is PC2. The clusters are defined by their means and standard deviations: C1: 2.324 ± 0.876, C2: 1.525 ± 0.667, and C3: 1.182 ± 0.579. The plot shows the distribution of data points for each cluster.</p>
<p>13 PCA</p> <p>Scatterplot</p>	 <p>PCA scatterplot showing the distribution of 13 variables for three clusters (C1, C2, C3). The x-axis is PC1 and the y-axis is PC2. The clusters are defined by their means and standard deviations: C1: 2.308 ± 0.889, C2: 1.112 ± 0.648, and C3: 1.521 ± 0.684. The plot shows the distribution of data points for each cluster.</p>	 <p>PCA scatterplot showing the distribution of 13 variables for three clusters (C1, C2, C3). The x-axis is PC1 and the y-axis is PC2. The clusters are defined by their means and standard deviations: C1: 2.324 ± 0.876, C2: 1.525 ± 0.667, and C3: 1.182 ± 0.579. The plot shows the distribution of data points for each cluster.</p>
<p>23</p> <p>Variables</p> <p>Scatterplot</p>	 <p>Scatterplot showing the distribution of 23 variables for three clusters (C1, C2, C3). The x-axis is PubYears and the y-axis is References. The clusters are defined by their means and standard deviations: C1: 2.308 ± 0.889, C2: 1.112 ± 0.648, and C3: 1.521 ± 0.684. The plot shows the distribution of data points for each cluster.</p>	 <p>Scatterplot showing the distribution of 23 variables for three clusters (C1, C2, C3). The x-axis is PubYears and the y-axis is References. The clusters are defined by their means and standard deviations: C1: 2.324 ± 0.876, C2: 1.525 ± 0.667, and C3: 1.182 ± 0.579. The plot shows the distribution of data points for each cluster.</p>

Hence, Hierarchical Cluster Analysis without PCA gives us a more meaningful cluster analysis than K-Means as it's not restricted to data treatment and shows variants of clustering.