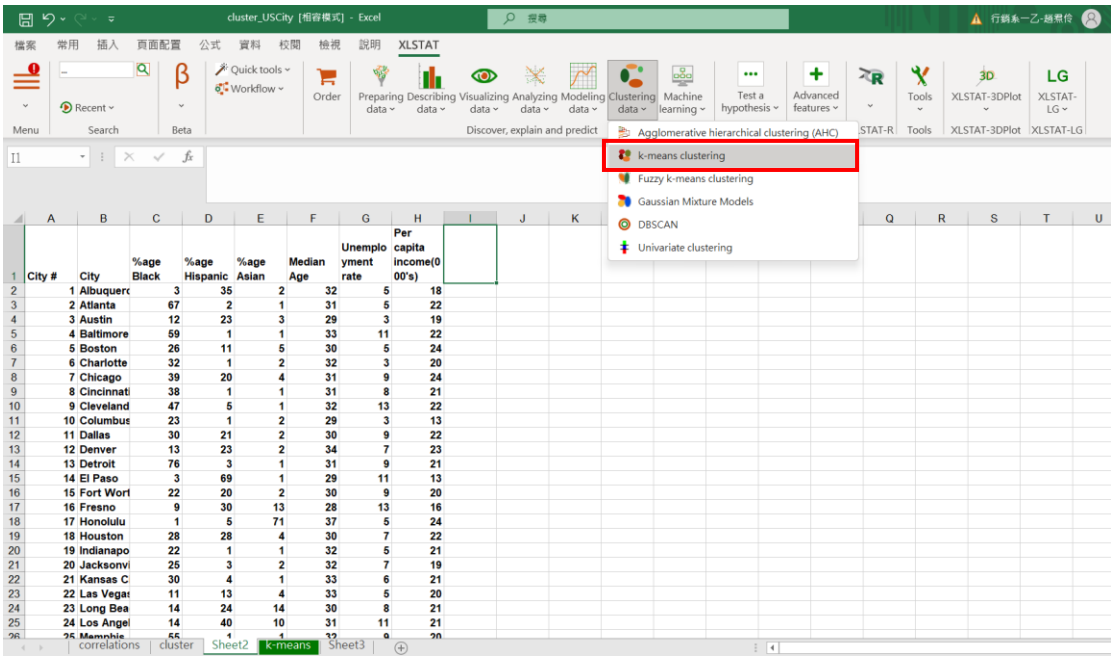
學號：7112044020　　班級：行銷碩一　　姓名：趙惹伶

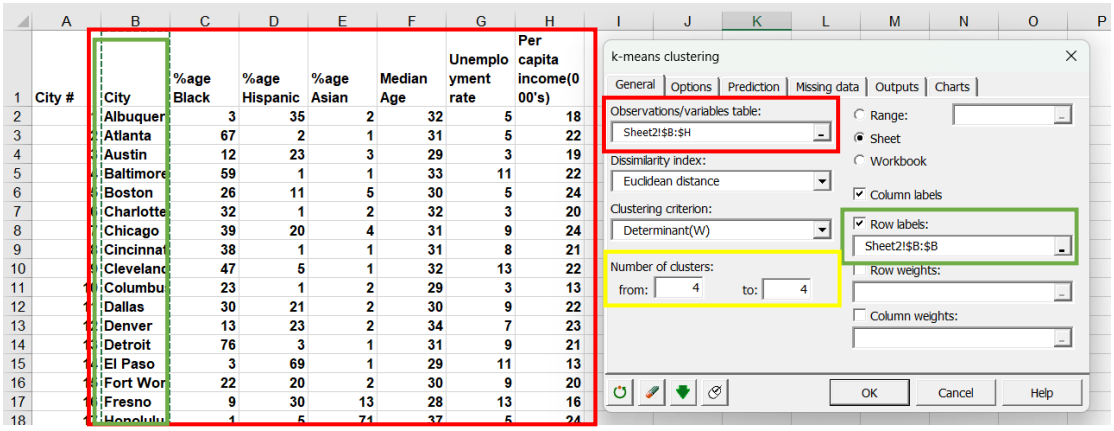
5/15 作業— Clustering Analysis

Excel 操作：

1. 點選 XLSTAT 中的 Clustering data，並選擇 k-mean clustering。



2. 在 k-mean clustering 的頁面中，在 Oberservations/ variables table 欄位選擇欲分析的完整表格，在 Row tables 欄位選擇要分群的依據，在 cluster_UScity 檔案中是以 city 作為分群依據，並選擇想要的 Number of cluaters，而後點選 OK，開始執行分析。



學號：7112044020　　班級：行銷碩一　　姓名：趙惹伶

分析結果(File: cluster_UScity.xlsx)：

1. Central objects：各分群的主要特徵，下圖的主要特徵分別為 1 (Long Beach)、2 (Memphis)、3 (Minneapolis)、4 (San Antonio)。

Central objects:

Cluster	%age Black	%age Hispanic	%age Asian	Median Age	Unemployment rate	Per capita
1 (Long Be	14.000	24.000	14.000	30.000	8.000	21.000
2 (Memphi	55.000	1.000	1.000	32.000	9.000	20.000
3 (Toledo)	20.000	4.000	1.000	32.000	6.000	19.000
4 (San Ant	7.000	56.000	1.000	30.000	5.000	17.000

1. Results by cluster：在各分群下方分別列出該分群包含的 city 有哪些。

Results by cluster:

Cluster	1	2	3	4
Number of	15	11	20	3
Sum of we	15	11	20	3
Within-clu	180.552	221.255	372.437	241.667
Minimum d	6.881	5.685	7.203	9.809
Average di	12.484	12.988	13.896	12.468
Maximum d	17.423	23.940	65.965	15.585
	Albuquerque	Atlanta	Boston	El Paso
	Austin	Baltimore	Charlotte	Miami
	Dallas	Chicago	Columbus	San Antonio
	Denver	Cincinnati	Honolulu	
	Fort Worth	Cleveland	Indianapolis	
	Fresno	Detroit	Jacksonville	
	Houston	Memphis	Kansas City	
	Long Beach	New Orleans	Las Vegas	
	Los Angeles	Oakland	Milwaukee	
	NY	Philadelphia	Minneapolis	
	Phoenix	St. Louis	Nashville	
	Sacramento		Oklahoma Cit	
	San Diego		Omaha	
	San Jose		Pittsburgh	
	Tucson		Portland	

分析結果(File: OnlineStore1220.xlsx)：

1. Central objects：
   - Cluster 2：代表可以忽略的顧客，因其 TotalSales、OrderCount、AvgOrderValue 都最低。
   - Cluster 4：代表需要重視的顧客，因其 TotalSales、OrderCount、AvgOrderValue 都最高。

- 確認消費者屬於哪一群 or 清楚辨認某一群內的特定消費者。

Central objects:

Class	TotalSales	OrderCount	AvgOrderValue
1 (14088)	50491.810	13.000	3883.985
2 (14223)	991.130	3.000	330.377
3 (13001)	9227.810	12.000	768.984
4 (17450)	192988.390	45.000	4288.631

R 語言操作：

1. 執行 k-means 函數
   - n_cluster：集群個數
   - silhouette()：計算集群分析結果的 silhouette score，silhouette score 是評估集群質量的方法，數值範圍是-1 到 1 之間，數值越接近 1 代表集群效果越好。

```
for(n_cluster in 2:8){
  cluster <- kmeans(USCityData, n_cluster)

  silhouetteScore <- mean(
    silhouette(
      cluster$cluster,
      dist(USCityData, method = "euclidean")
    )[,3]
  )
  print(sprintf('Silhouette Score for %i Clusters: %0.4f', n_cluster, silhouetteScore))
}
```

得出以下結果：

```
[1] "Silhouette Score for 2 Clusters: 0.4418"
[1] "Silhouette Score for 3 Clusters: 0.2210"
[1] "Silhouette Score for 4 Clusters: 0.2632"
[1] "Silhouette Score for 5 Clusters: 0.2359"
[1] "Silhouette Score for 6 Clusters: 0.2541"
[1] "Silhouette Score for 7 Clusters: 0.2534"
[1] "Silhouette Score for 8 Clusters: 0.2586"
```

2. 將數據資料分群
   - USCityDataClusterData <- kmeans(USCityData, 4)：將數據分成 4 個集群。

```
USCityDataClusterData <- kmeans(USCityData, 4)
```

得出以下結果：

```
  X.age.Black X.age.Hispanic X.age.Asian  Median.Age
1  -0.9394460     -0.4204183   2.8078125  2.06517951
2   1.3485339     -0.4048519  -0.2659440  0.06134197
3  -0.6817651      2.2467186  -0.0754447 -0.53980930
4  -0.4037064     -0.1681880  -0.1638908 -0.15335491
  Unemployment.rate Per.capita.income.000.s.
1        -0.6274845                2.0238445
2         0.7648927                0.3935951
3         1.2569931               -1.2351164
4        -0.5123613               -0.1790246
```

3. 數據資料整理及分析
   - group_by(Cluster)：根據名為 Cluster 的列進行分組。
   - summarise(Count=n())：對每個分群進行彙整；Count=()將每個分群的觀測值作為 Count 列的值。

```
33 USCityCluster %>% group_by(Cluster) %>% summarise(Count=n())
```

得出以下結果：

```
# A tibble: 4 × 2
  Cluster Count
    <int> <int>
1       1     3
2       2    13
3       3     5
4       4    28
```