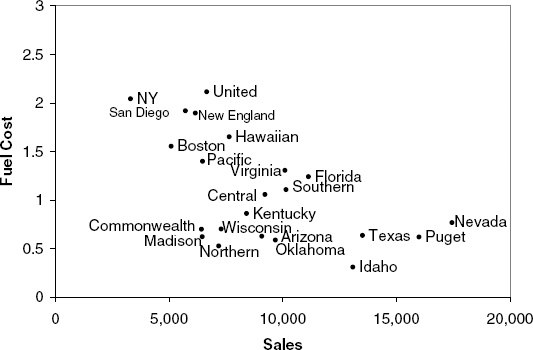
**ISOM2600 InClass Case Exercise (Class 4)**

Utilities.csv gives corporate data on 22 U.S. public utilities. We are interested in forming groups of similar utilities. The records to be clustered are the utilities, and the clustering will be based on the eight measurements on each utility. An example where clustering would be useful is a study to predict the cost impact of deregulation. To do the requisite analysis, economists would need to build a detailed cost model of the various utilities. It would save a considerable amount of time and effort if we could cluster similar types of utilities and build detailed cost models for just one "typical" utility in each cluster and then scale up from these models to estimate results for all utilities. The eights measurement are:

* Fixed = fixed-charge covering ratio (income/debt);
* RoR = rate of return on capital;
* Cost = cost per kilowatt capacity in place;
* Load = annual load factor;
* Demand = peak kilowatt hour demand growth last year;
* Sales = sales (kilowatt hour use per year);
* Nuclear = percent nuclear;
* Fuel total fuel costs (cents per kilowatt hour)

For simplicity, let us consider only two of the measurements: Sales and Fuel Cost. The figure below shows a scatterplot of these two variables, with labels marking each company.



At first glance, there appear to be two or three clusters of utilities. A suggestion of three clusters:

1. one with utilities that have high fuel costs,
2. a second with utilities that have lower fuel costs and relatively low sales,
3. and a third with utilities with low fuel costs but high sales.

**Question 1: According to the cluster guideline of above, which group do Nevada and Commonwealth belong to?**

1. **Nevada: first cluster, Commonwealth: second cluster**
2. **Nevada: second cluster, Commonwealth: second cluster**
3. **Nevada: third cluster, Commonwealth: second cluster**
4. **Nevada: second cluster, Commonwealth: third cluster**

The measure computed above is highly influenced by the scale of each variable, so that variables with larger scales (e.g., Sales) have a much greater influence over the total distance. It is therefore customary to *normalize* (or *standardize*) continuous measurements before computing the Euclidean distance. This converts all measurements to the same scale. Normalizing a measurement means subtracting the average and dividing by the standard deviation.

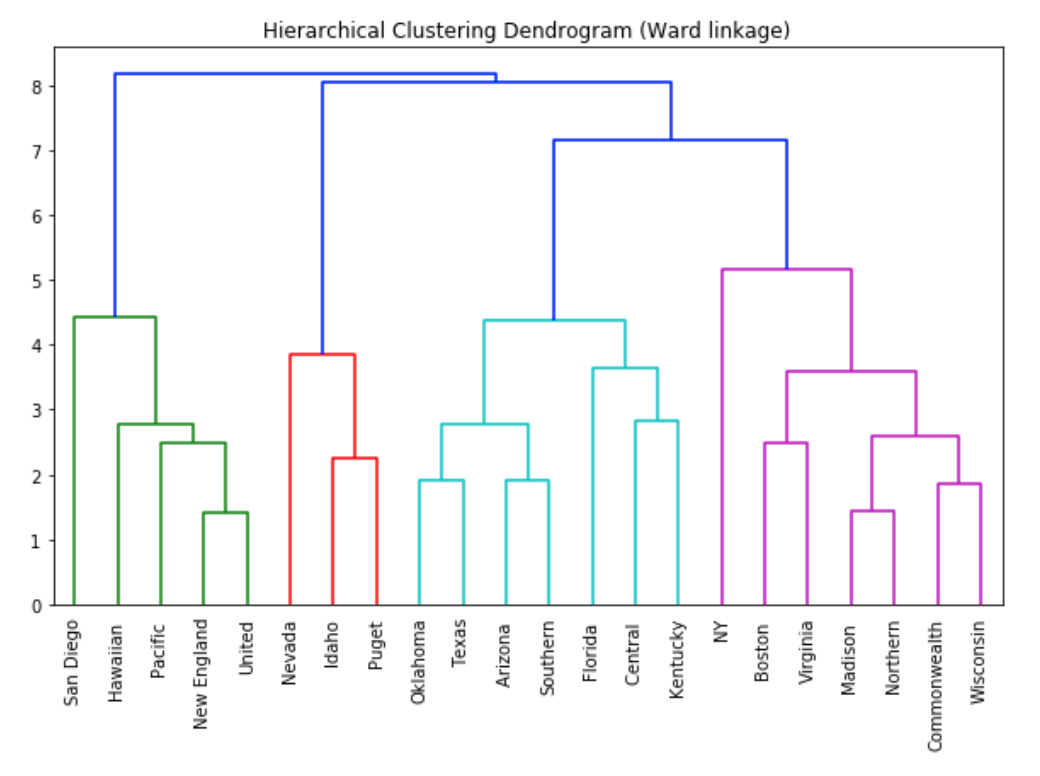
The following contains the records of Sales and Fuel Cost for two companies, also the mean and sd of the two variables:

| **Company** | **Sales** | **Fuel Cost** |
| --- | --- | --- |
| Arizona Pulic Service | 9,077 | 0.628 |
| Boston Edison Co. | 5,088 | 1.555 |
| Mean | 8,914.05 | 1.1 |
| Standard deviation | 3,549.98 | 0.56 |

**Question 2: Compute the normalized values for Boston Edison Co.**

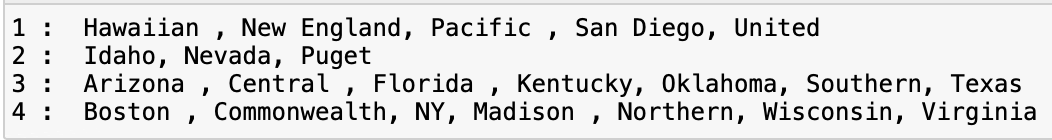
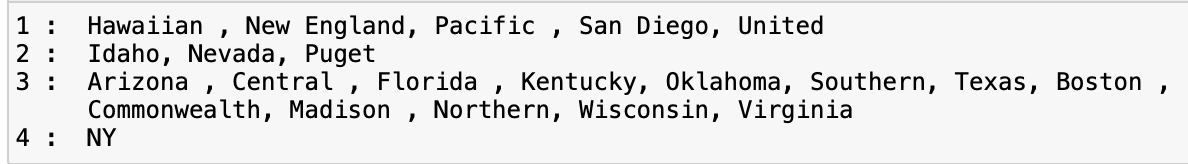
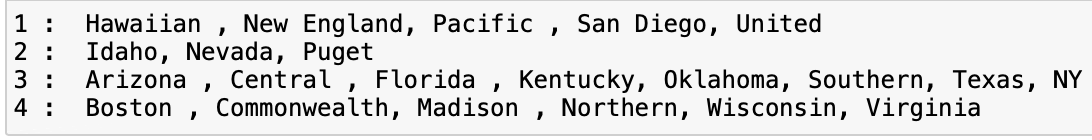
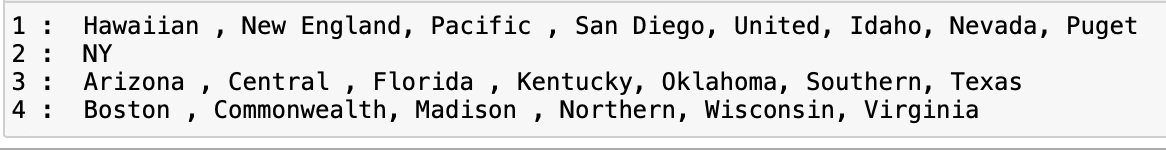
1. **Sales: -1.0778, Fuel Cost: 0.8125**
2. **Sales: 0.0459, Fuel Cost: -0.8429**
3. **Sales: -1.0778, Fuel Cost: 1.555**
4. **Sales: -1.0778, Fuel Cost: 0.56**

After the normalisation step, we may use all the variables to perform hierarchical clustering. Prior to this, we use Euclidean distance and Ward linkage to measure the distance for among observations and among clusters respectively. After all, we have the following Dendrogram:



Based on the dendrogram, forming 4 clusters seems to be an appropriate choice.

**Question 3: Based on the formation of 4 clusters, what is the clustering result?**

1. ****
2. ****
3. ****
4. ****

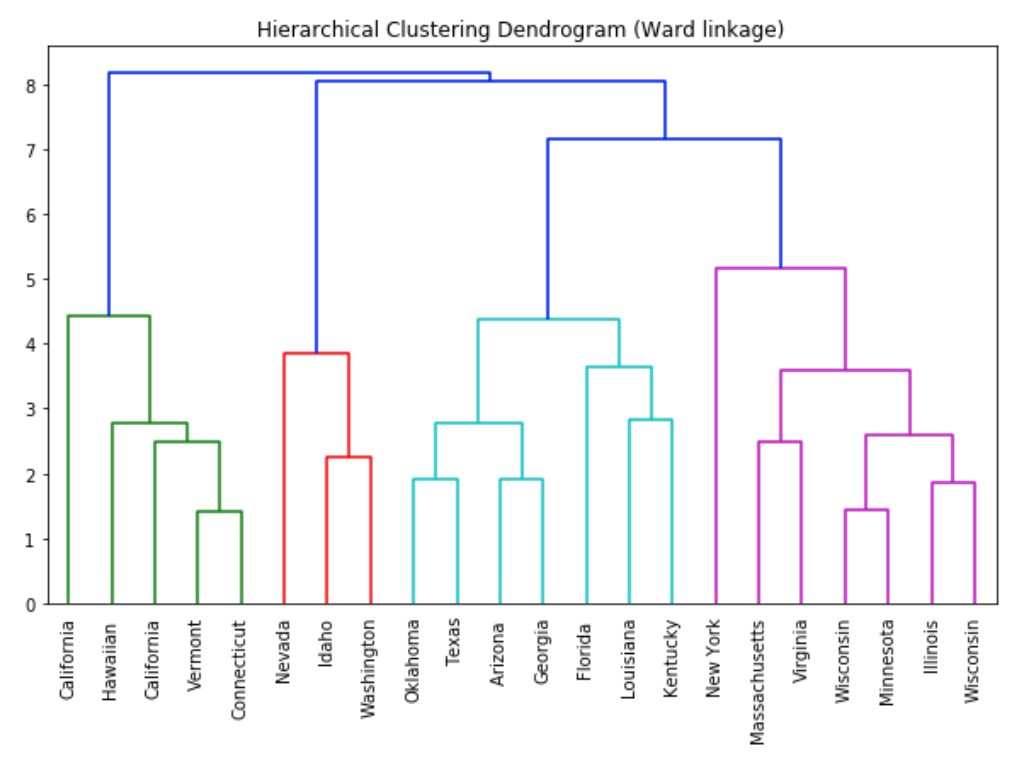
By referencing to the US map, the locations/states of the companies may help to name each cluster. (a Dendrogram with ‘State’ as the labeling and US map are given in next page for reference)

**Question 4: Which of the following is an appropriate cluster name for those with Texas?**

1. **Southern group**
2. **Northern group**
3. **Eastern group**
4. **West seaboard group**

One insight that can be derived from this clustering is that clusters tend to group geographically. So the economists could take geographical boarder as a starting point to build the cost model of the various utilities.

Dendrogram with ‘State’ as the labeling



US Map

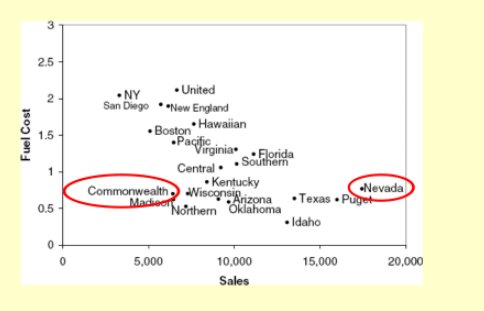
****

ANS to Question1

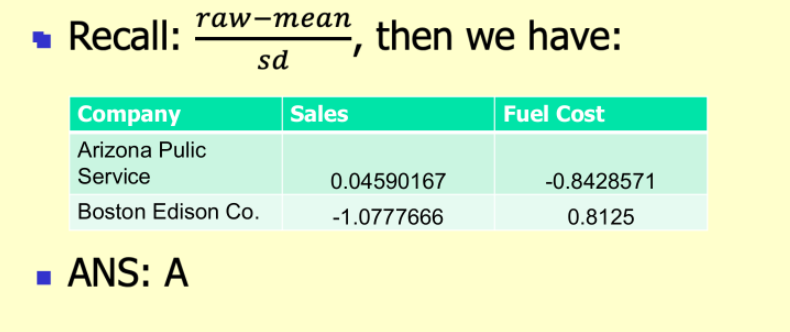
According to the cluster guideline of

above, which group do Nevada and

Commonwealth belong to?

ANS: C

ANS to Question 2



ANS to Question 3

