WELCOME to this SAS Enterprise Miner 5.3 tutorial. Parts of this write-up are based on SAS Education material. This handout introduces you to HW 5.

CREATING A NEW PROJECT
Select New Project. The Create New Project dialog box appears. In the General tab, type the short name of your assignment, e.g. student1hw5a. (Please note that naming it hw5 won’t work, as the server won’t allow two folders having the same name.) Type C:\4520projects\xxxxx in the path field, where xxxxx is your user name, e.g. C:\4520projects\4520student1.

Select the Start-Up Code tab. Type the library definition code as follows:
libname ADMT 'C:\4520data';
You may use single or double quotes. Do not copy/paste smart quotes produced by Word. Do not forget the semicolon at the end!

Click OK. An Enterprise Miner application window opens. Select File > New > Diagram… and then type a diagram name, e.g. diagram1. Click OK. A new analysis diagram called diagram1 should open on the right side of the EM 5.3 application window (diagram panel).

ADDING A DATA SOURCE
Right-click on Data Sources in the Project Panel and select Create Data Source. You are looking for a SAS Table called PROSPECT, which is part of the ADMT library. Click on Browse, then open ADMT and select PROSPECT.
Verify the table properties look as follows:

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table Name</td>
<td>ADMT.PROSPECT</td>
</tr>
<tr>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>Member Type</td>
<td>DATA</td>
</tr>
<tr>
<td>Data Set Type</td>
<td>DATA</td>
</tr>
<tr>
<td>Engine</td>
<td>V9</td>
</tr>
<tr>
<td>Number of Variables</td>
<td>9</td>
</tr>
<tr>
<td>Number of Observations</td>
<td>4701</td>
</tr>
<tr>
<td>Created Date</td>
<td>2006-12-06 16:30:46.609</td>
</tr>
<tr>
<td>Modified Date</td>
<td>2006-12-06 16:30:46.609</td>
</tr>
</tbody>
</table>

Click OK. Select Basic > Next. Change the Role of LOC from Input to Rejected. Please note that Clustering is an unsupervised learning technique and, therefore, there is no target variable!

Click Finish. PROSPECT now appears as a Data Source in the left panel.

**BUILDING AN ANALYSIS DIAGRAM**

Drag and drop PROSPECT on your diagram, then build the analysis diagram shown below.

Note that the Impute node is not critical in cluster analysis. If missing value imputation is not done, clustering is based on the non-missing inputs.

**SETTING UP THE IMPUTE NODE**

Select the Impute node. In the property panel, change the default Input method for both Interval and class variables to Tree. Run the Impute node.
SETTING UP THE CLUSTER NODE
Select the Cluster node. Click on the Ellipses button, next to the Variables property. Change the status of all original variables (before imputations) that were actually imputed to Use=No, keeping the imputed versions to Use=Default. Leave the status of Climate to Default, since Climate was not imputed. If you don’t see the IMP_ variables, run the Impute node and update the Cluster node path.

K-means clustering is very sensitive to variable scale. Variables with relatively large values have a large influence on the formed clusters. When your data set combines variables with different measurement scales, standardize the variables in order to equate their influence. Select Internal Standardization = Standardization.
Select Specification Method = User Specify, and Maximum Number of Clusters = 4.

- Train
  - Variables
  - Cluster Variable Role: Segment
  - Internal Standardization: Standardization
  - Number of Clusters
  - Specification Method: User Specify
  - Maximum Number of Clusters: 4

Run the Cluster node and view the results. The segment plots show you, for each variable, how different kinds of values are distributed over the four clusters. Careful examination of these plots will help you interpret the four clusters.