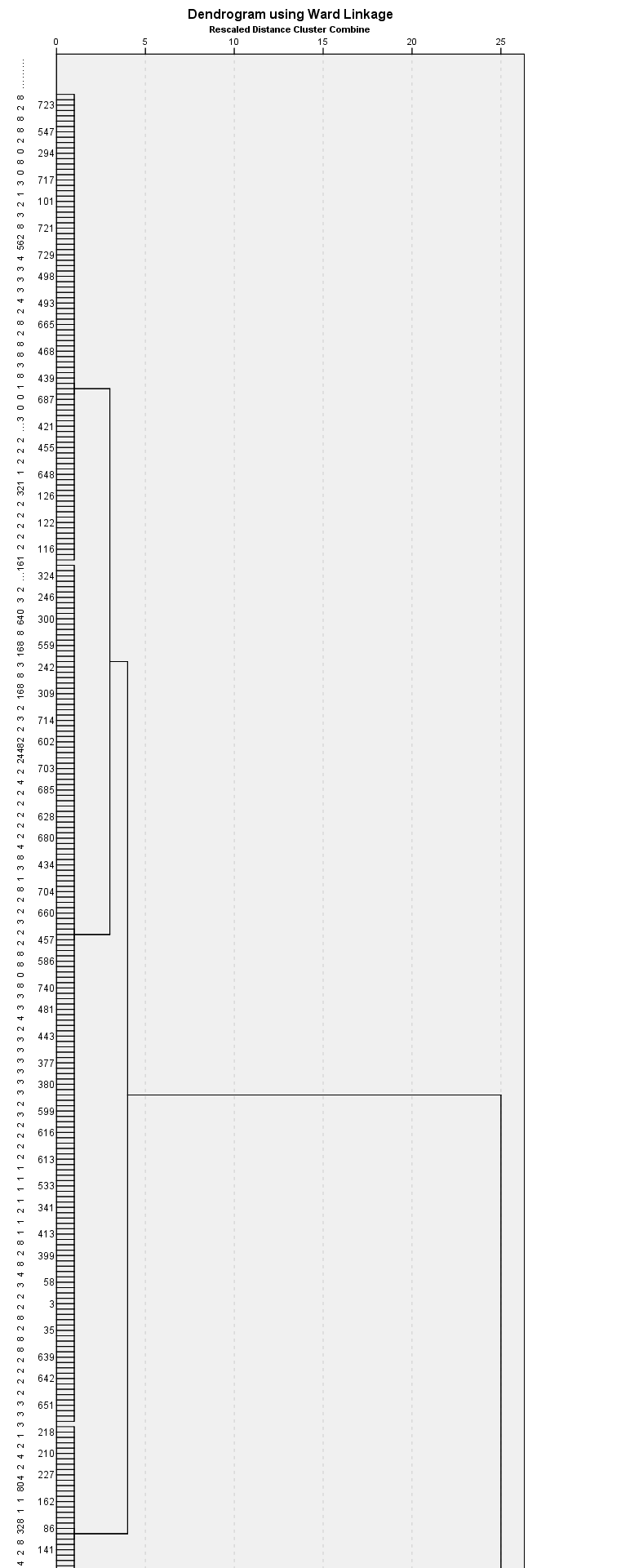
**Project Definition:** The main objective of this project is to perform hierarchical clustering to find new knowledge from the dataset absenteeism at work. The dataset absenteeism at work contains 21 attributes and 740 instances. The data set allows for several new combinations of attributes and attribute exclusions, or the modification of the attribute type (categorical, integer, or real) depending on the purpose of the research. Which is to be put through hierarchical clustering to discover new knowledge. Used Machines:

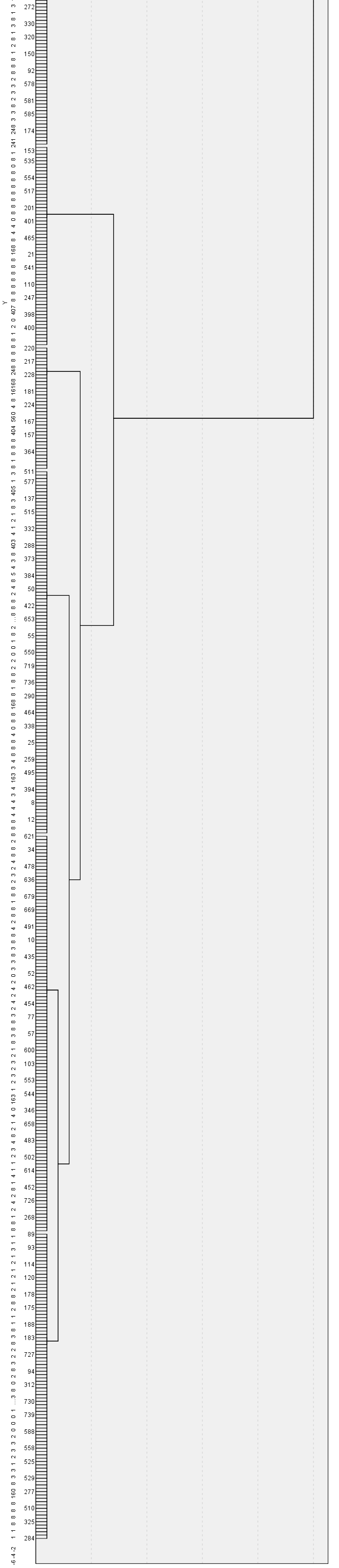
* IBM SPSS Statistics Data Editor
* Microsoft Excel

**Literature Survey:** The idea of Hierarchical Clustering is to assign every object into a cluster, then repeatedly merging the closest pair of those clusters till its just one single cluster which will have multiple sub cluster. It forms like a tree with root, node and leaf clusters. This entire tree is called dendrogram.   
To gather data from the dendrogram, it would require to choose a sub cluster from it and take the instances under the sub cluster apart to look onto the values of those instances to find relation among them as to why they were under this single sub cluster of the entire dendrogram hoping to find some insights.  
Since the data set contains its data in (.arff) format alongside with other common formats. Weka offers to create a dendrogram from the given data set via .arff file. But the dendrogram viewfinder tool is quite backdated to represent a quite large sized dendrogram that has been produced by this data set.  
Not going with the path of WEKA anymore, instead a tool from IBM named IBM SPSS Statistics. Which provided dendrogram of the data set with label that could be interpreted easily. Using Ward Linkage to create the dendrogram, it was decided to analyze the dendrogram on the sub cluster group that had 8 clusters on it.

**Method:** The hierarchical clustering was done using the software IBM SPSS statistics data editor. Where the attribute Absenteeism at work was used to label cases by and all of the other attributes were put into variables section.

Clustering was done by cases and labeled by Absenteeism time in hours. Ward’s method was used for clustering and for distance squared Euclidean distance formula was used. Weight height was excluded as there exists another instance Body mass index. Which lead to the dendrogram given below:





**Knowledge finding Analyzing the Dendrogram:**

To get more precise knowledge from the dendrogram, from rescaled combined distance cluster we have cut down point at 10 where 8 clusters are combined into 2 clusters.

From the recombined cluster 1, 50 instances are taken randomly to get an optimal view labeled by Absenteeism in hours.

**cluster1:**

|  |  |
| --- | --- |
|  |  |

Where we can see that, most of the workers having low absenteeism which is around 5 hours. The reason can be seen observing scatter points of Distance to the workplace from residence,

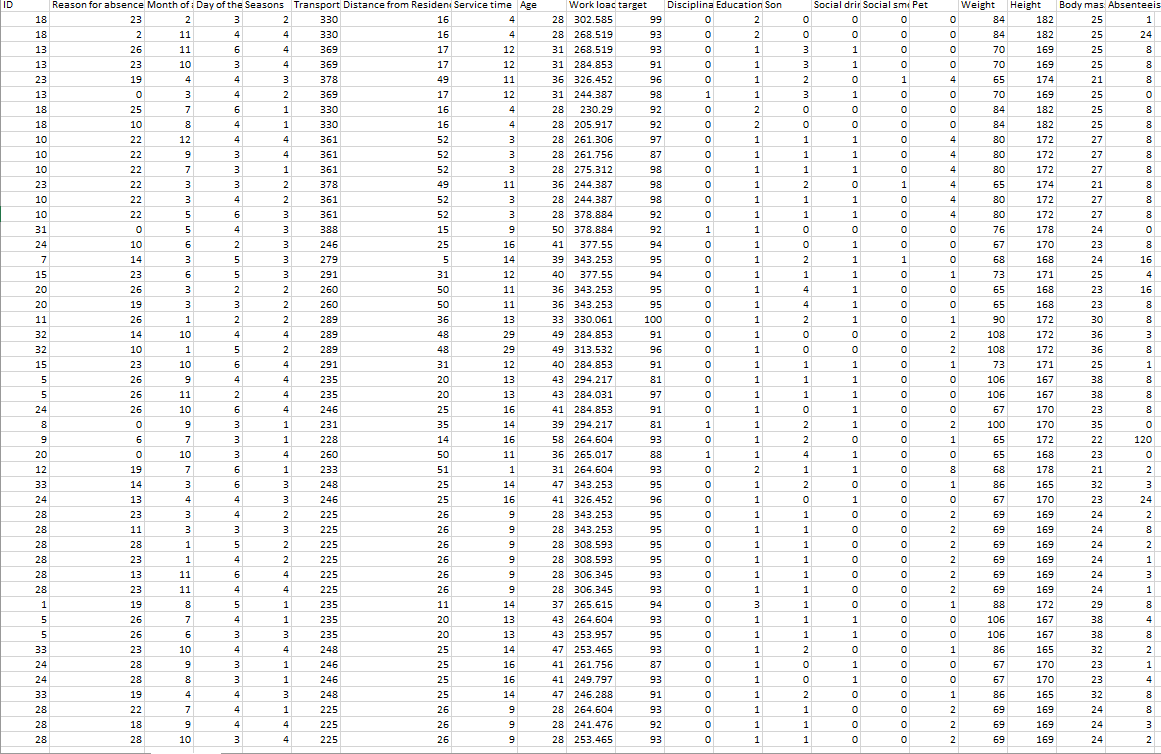
Where we can see that more than half of the workers’ residence value is less than 30, though there are almost half of the workers are at a distance of about 50. To get more precise knowledge, by observing transport cost,

It is seen that despite of being almost half of the workers’ residence distance about 50, all the workers’

Transport cost is seen below 190.which seems to be the major reason for the workers’ absenteeism being such a low value. Looking from another angle we can see that,

The age scatter graph shows that the workers’ age of cluster 1 majority being in the range of 35-40.

**cluster 2:**



Observing the recombined Cluster 2 labelled by absenteeism in hours, from the randomly taken 50 instances from cluster 2, absenteeism values are found visualized in the scatter graph,

Where it seen that majority of the workers’ absenteeism being above 5 hours. for the reason Same for the cluster 2 as cluster 1, by observing the attribute Distance from residence

Where above 60% of the workers’ distance from workplace is above 20. though there are workers living near the workplace have absenteeism of higher value. Because from the scatter graph generated from the attribute,

Here we can clearly see that despite of being some residences close to work place, transport cost is above 200. Which seems to be the major reason for a higher absenteeism value in recombined cluster 2. following cluster 1, looking at the scatter graph generated from Age,

Workers with the higher Absenteeism value seems to be aged under 30 mostly.

Analyzing the above observations, it is clear that workers who have less absenteeism are mostly middle aged, have a service time of about 10 and have residence near to the workplaces resulting low transport expense where workers having higher absenteeism are mostly younger comparing age and service time with the workers having low absenteeism. Also, its noticeable that they have residence comparatively far from the workplace resulting higher transport expense. Coincidently, they seem to have more or less pets where workers with lower absenteeism has no pets in the records from the taken instances given in the dataset.