**Assignment 2**

**Team Number: 04**

**Team Members: NetID:**

Akanksha Mishra axm170031

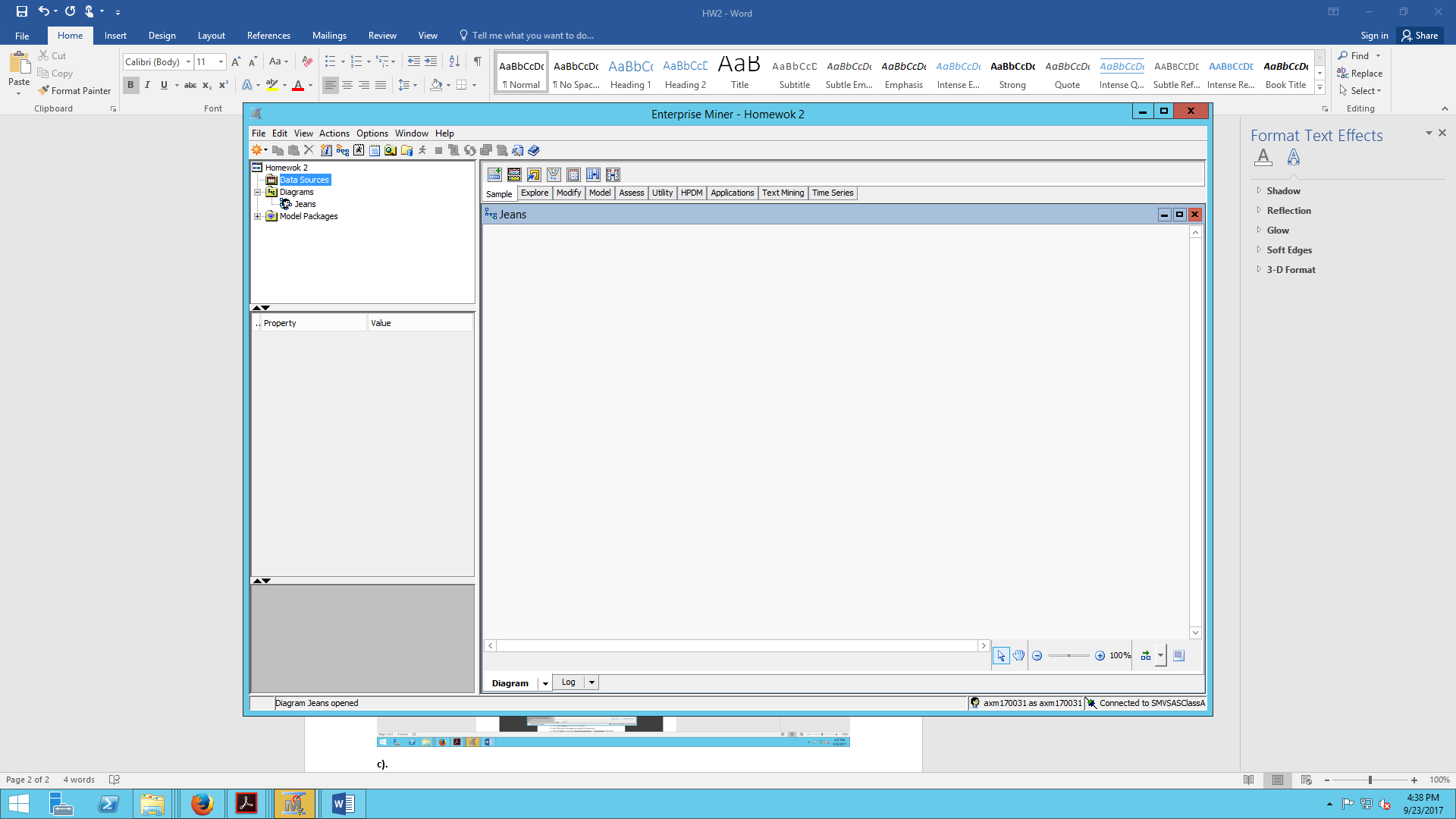
Harman Singh hxs134930

Lokesh Kumar Chaturvedi Lolla lxl163230

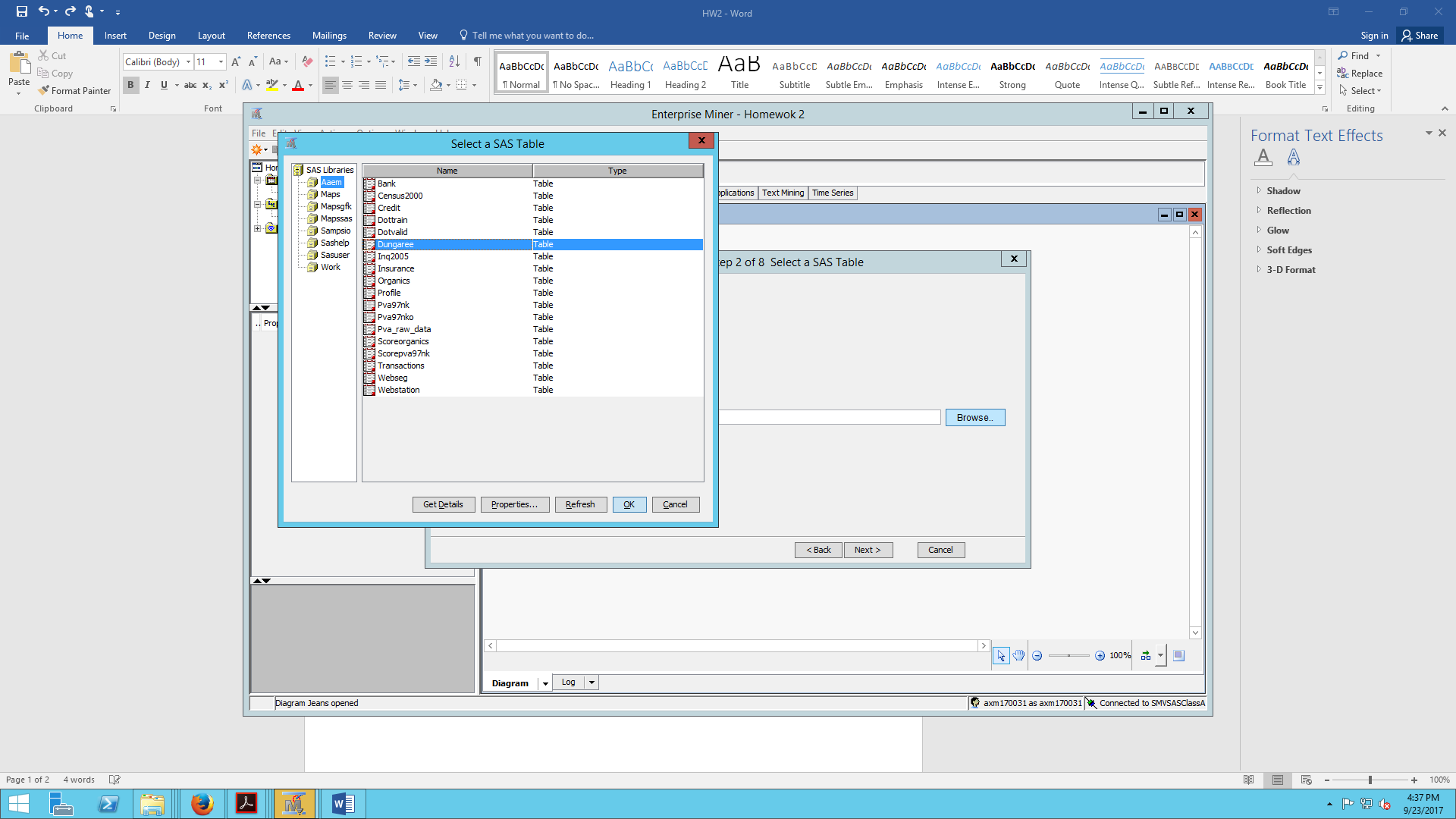
Yugandhara Rawool yer170030

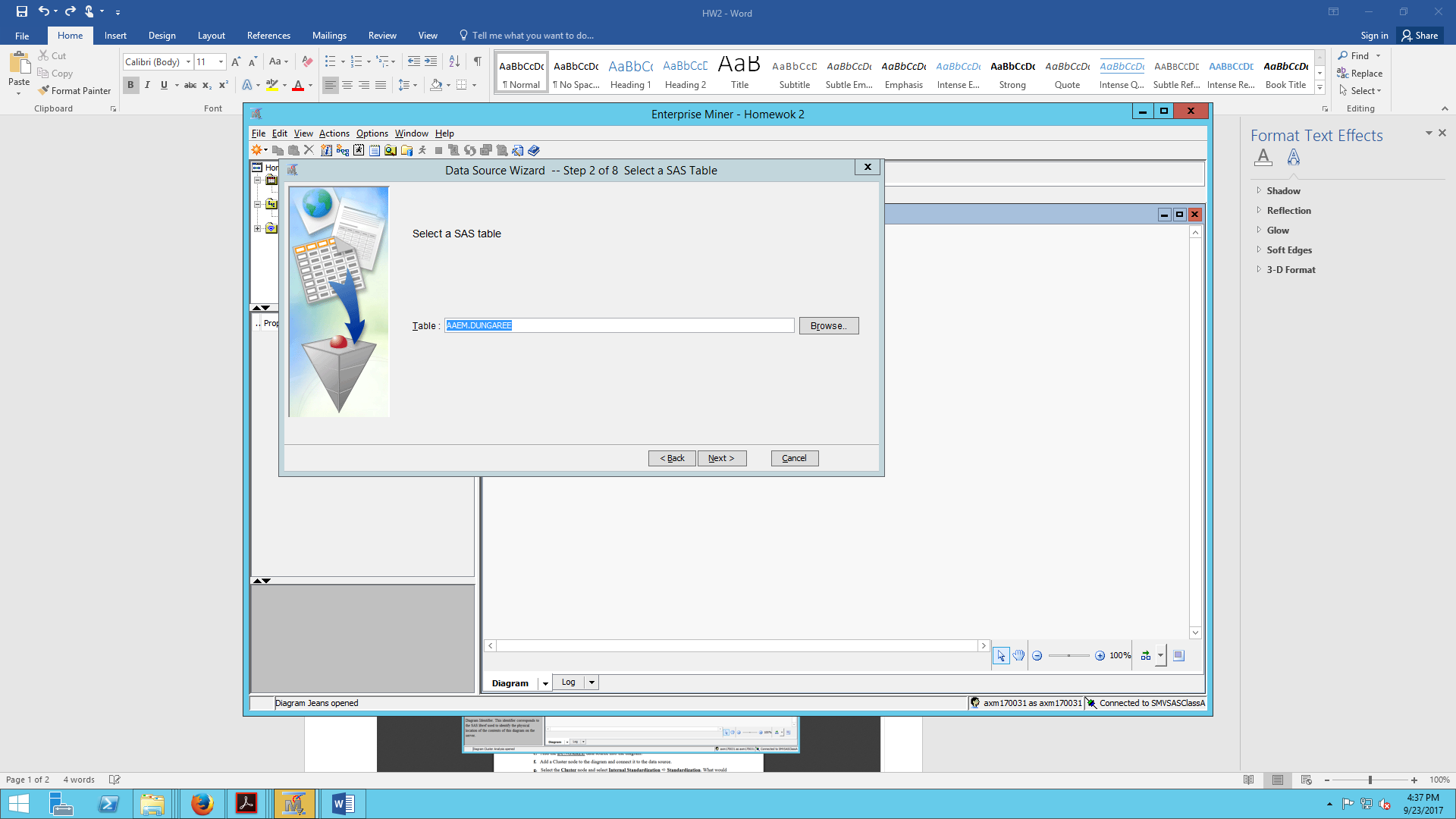
**Solution 1:** The cluster analysis of Dungaree Sales using SAS Enterprise Miner is depicted below with appropriate screenshots.

**a).** A new diagram, Jeans is created in our project titled Homework 2.

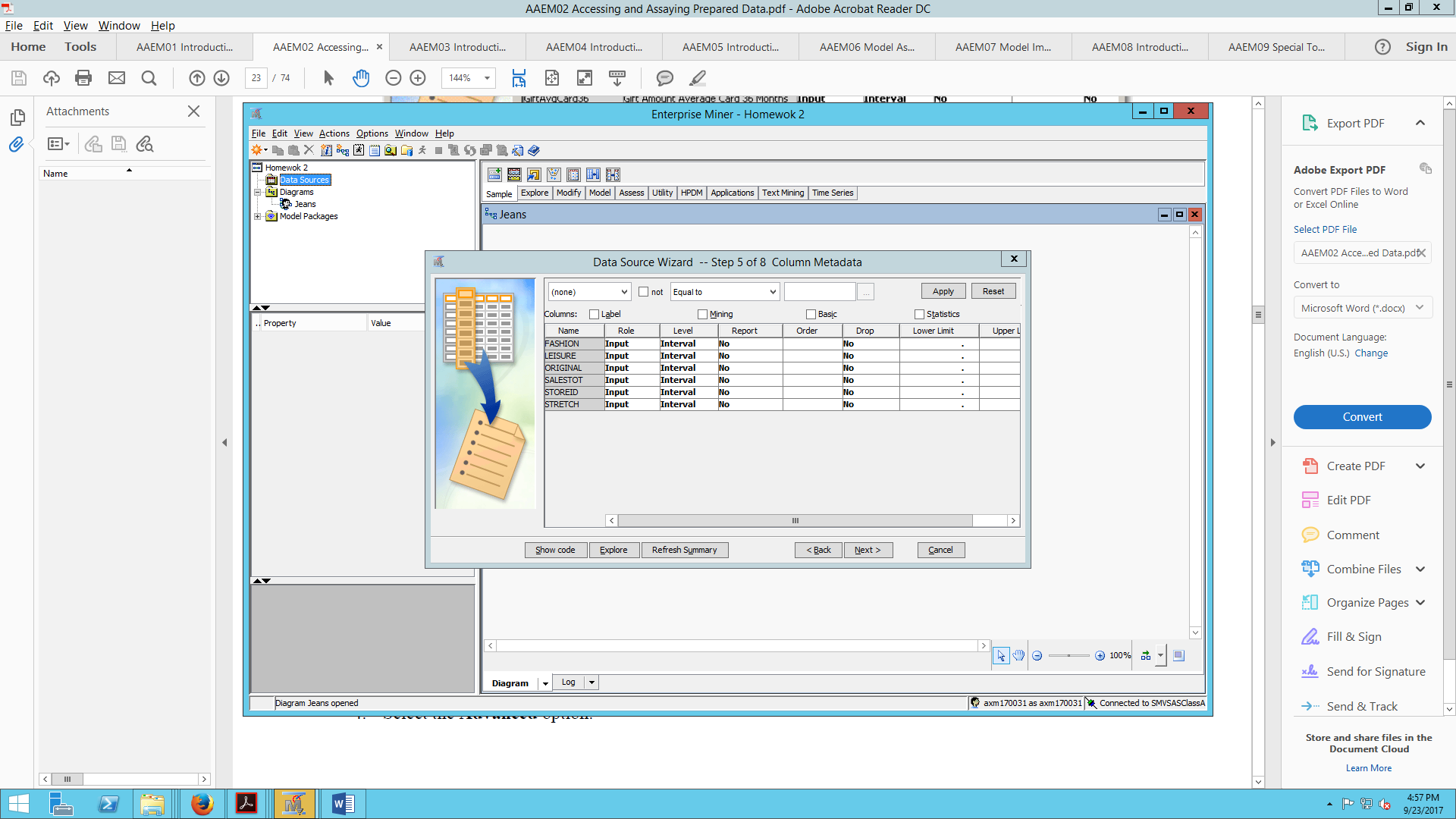


**b).** Dataset Dungaree is defined as the data source.





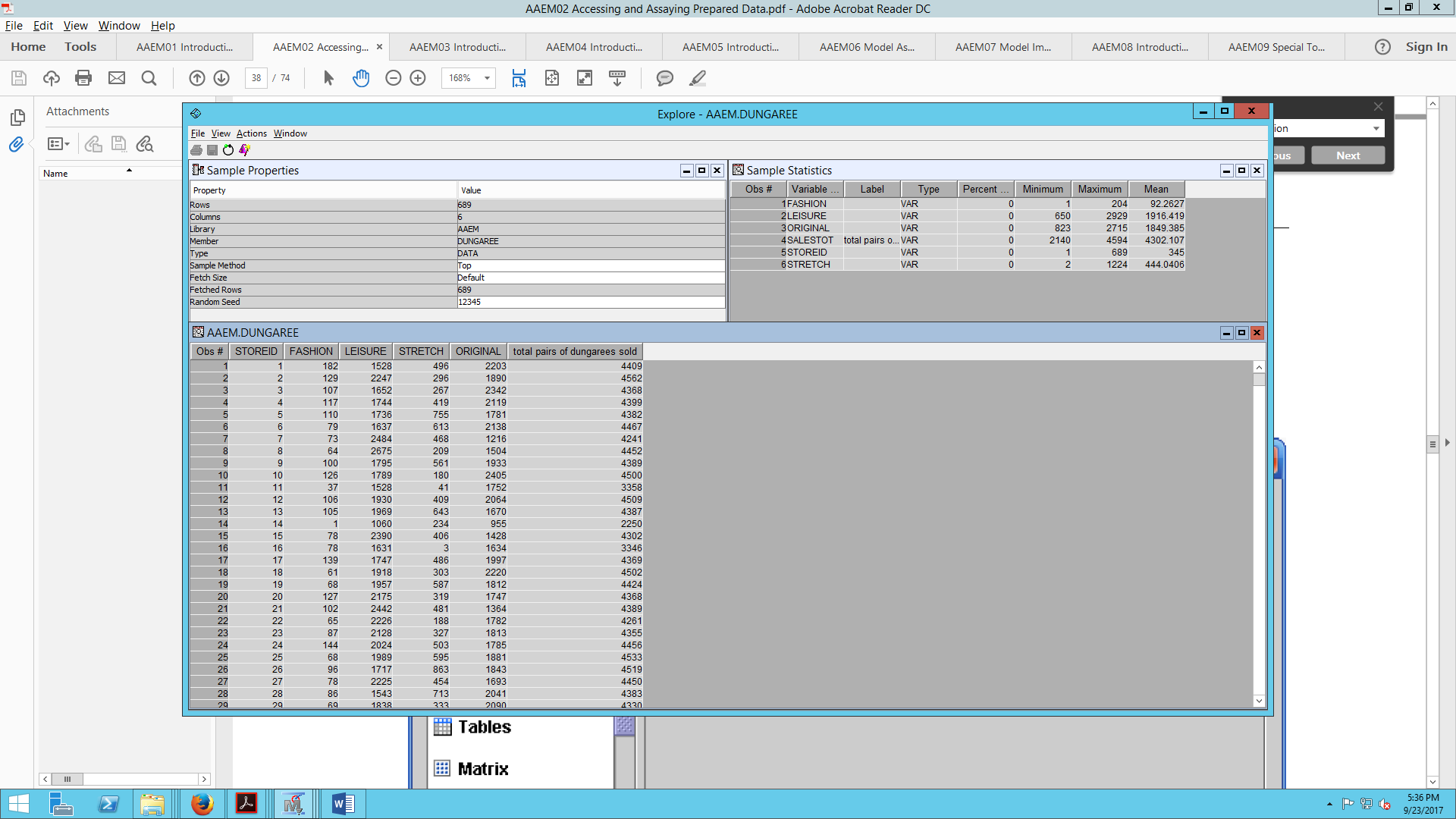
**c).**



→ The model roles and measurement levels assigned to the variables are not appropriate. The role assigned to Store ID is input but the Store Id seems to be auto-generated and it also uniquely identifies each Store where the Dungaree is sold (along with its different varieties) so its role should be changed to ID instead of input.

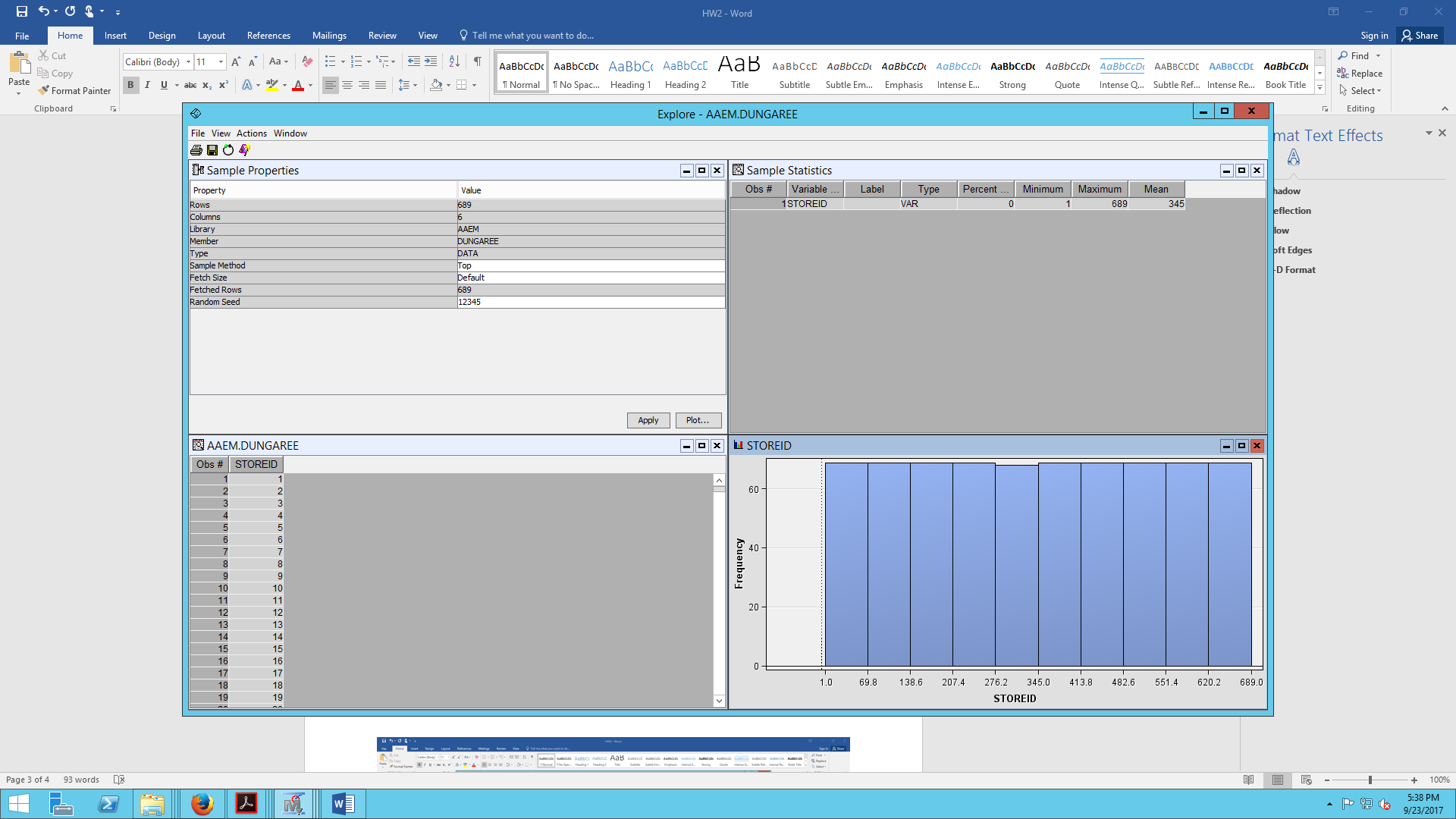
→ Also the sales total is the calculation of each sale in the data set its role should be changed from Input

Variables other than sales total are fairly distributed.

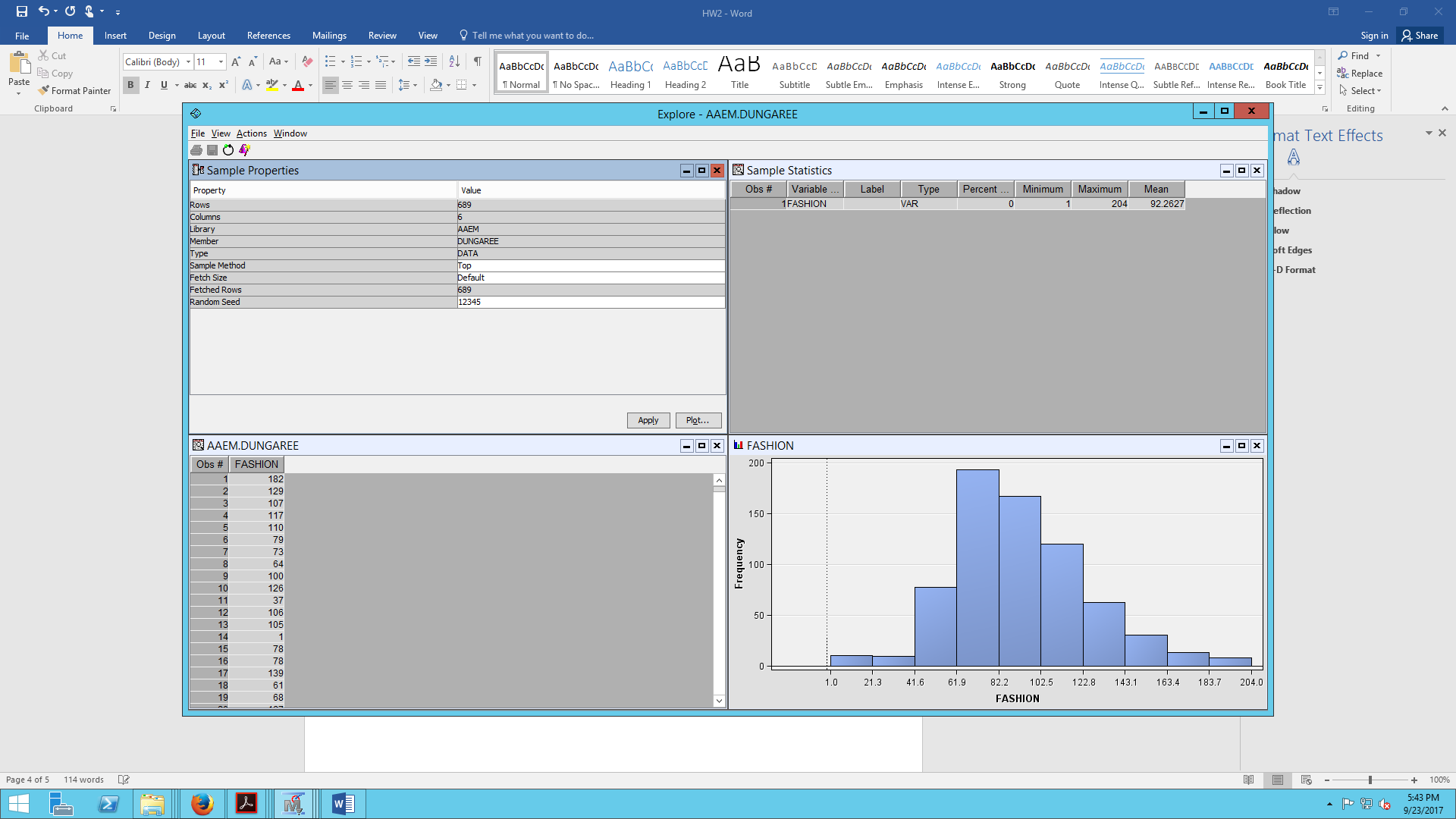


**Distribution of Individual Entities:**

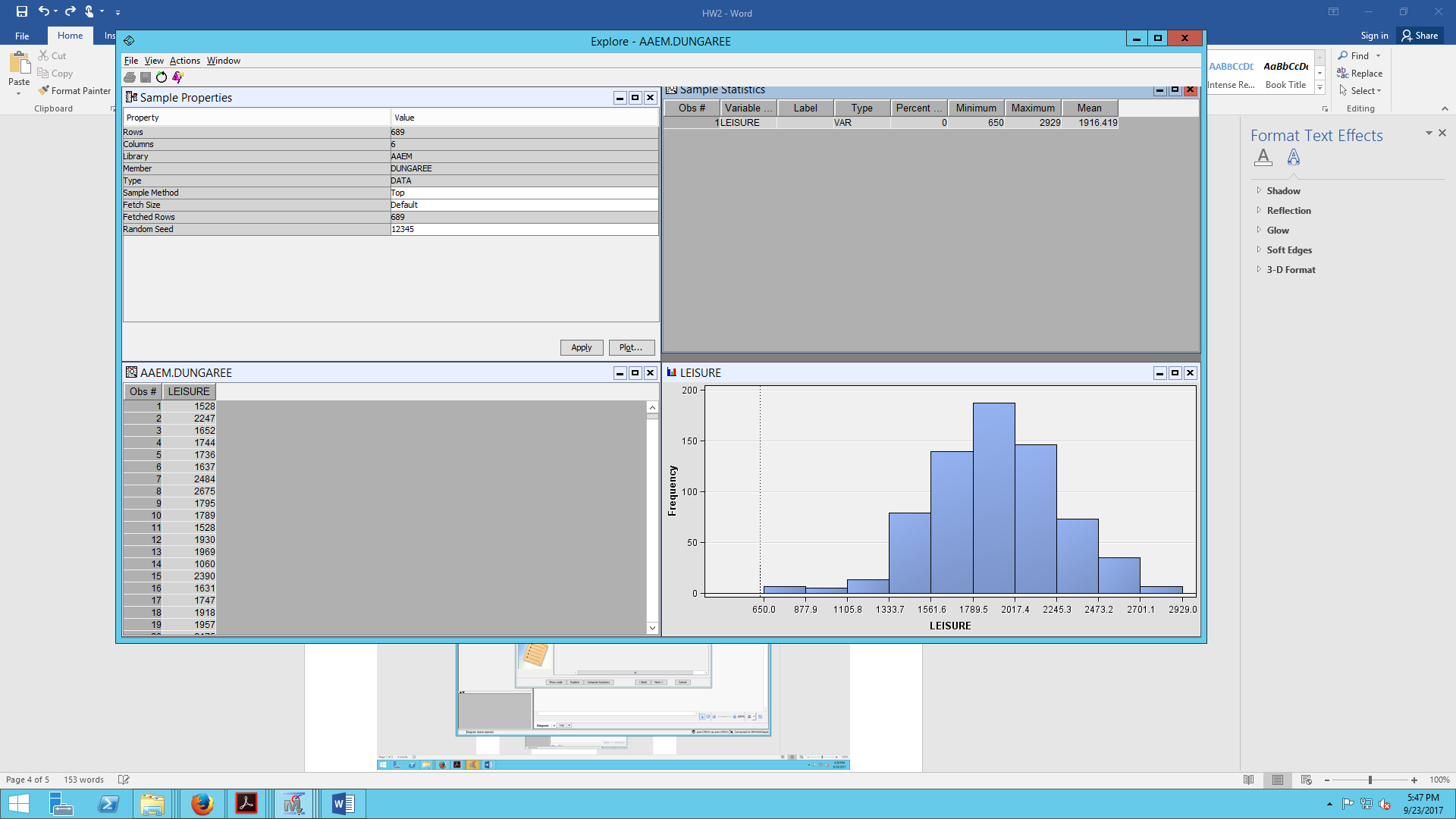
**Store ID:** Uniformly distributed (each store has its unique Store ID)



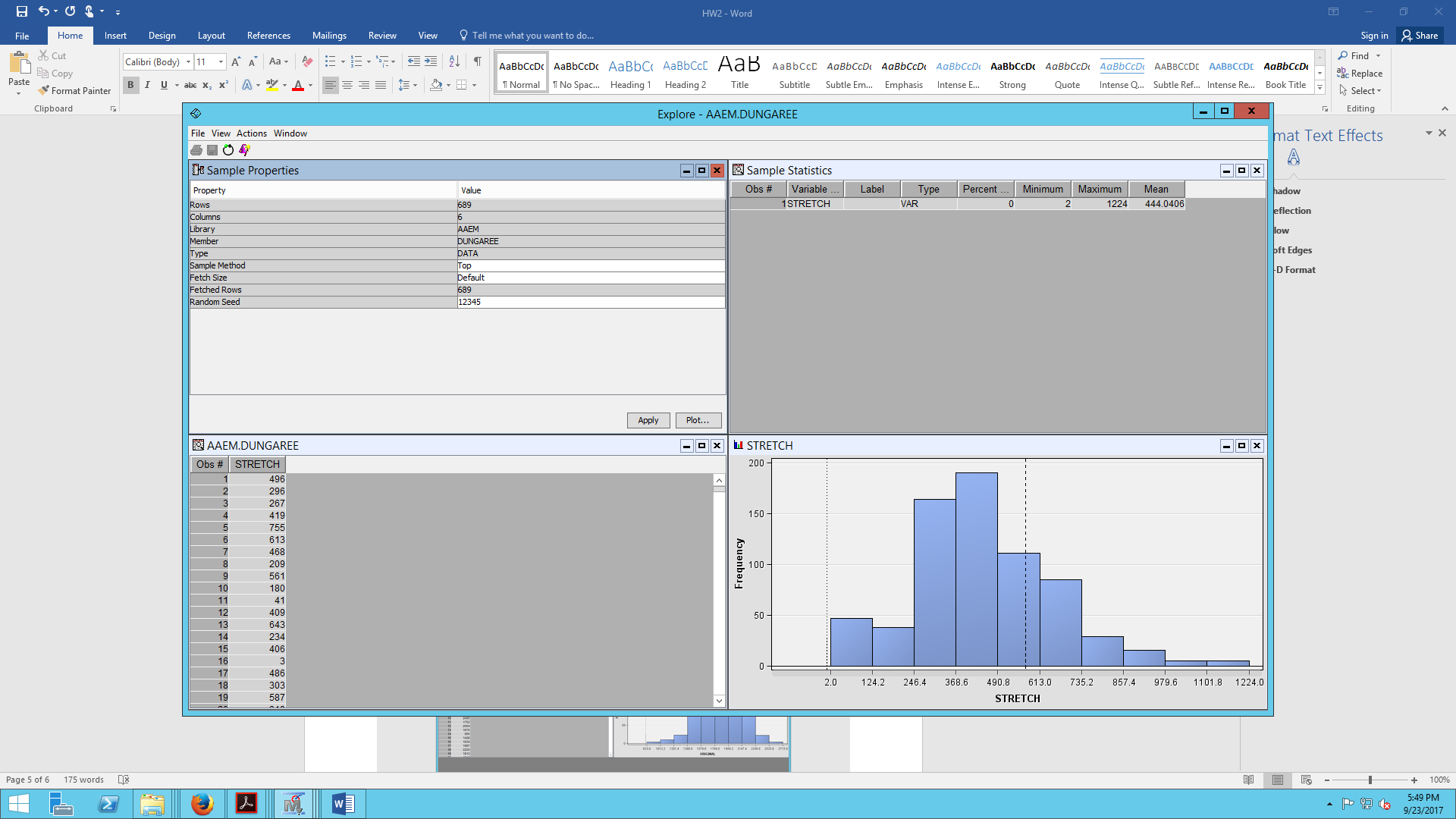
**Fashion:** Right Skewed which infers that number of Dungarees are sole whose fashion value is below the mean value (people tend to give less preference to Fashion than another variable that Dungaree has)



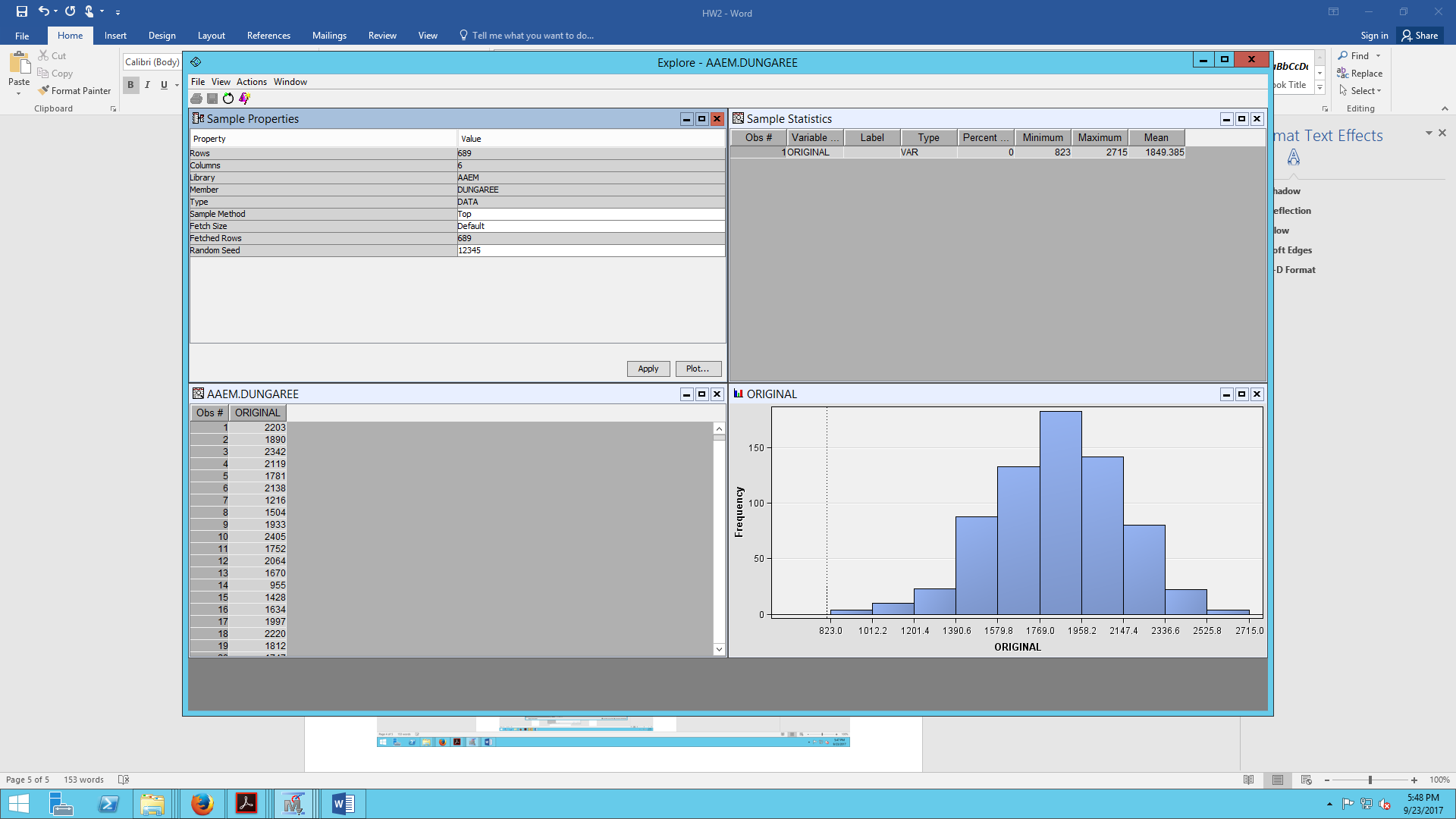
**Leisure:** Distribution follows Bell Curve which infers that max number of dungarees are Sold whose Leisure value is almost equal to mean



**Stretch:** Right Skewed which infers that number of Dungarees are sold whose Stretch value is below the mean value (people tend to give less preference to Stretch than another variable like originality and Leisure).

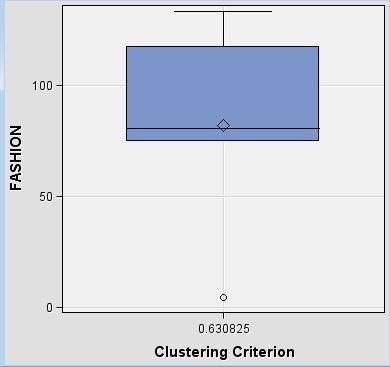
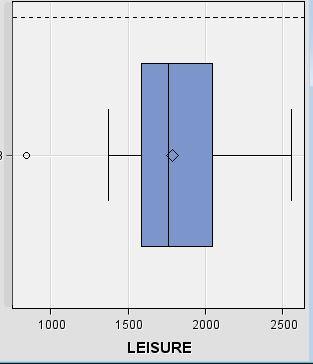


**Original:** Distribution follows Bell Curve which infers that max number of dungarees are Sold whose Leisure value is almost equal to mean

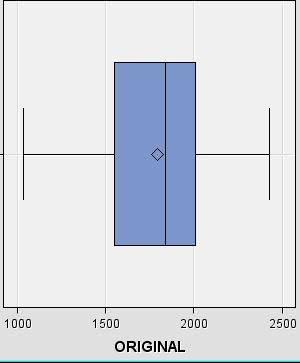
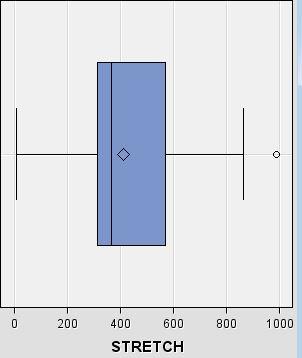


The box plots for each variable Fashion, Leisure, Stretch and Original is calculated to find any unusual data.

**Fashion:** **Leisure:**

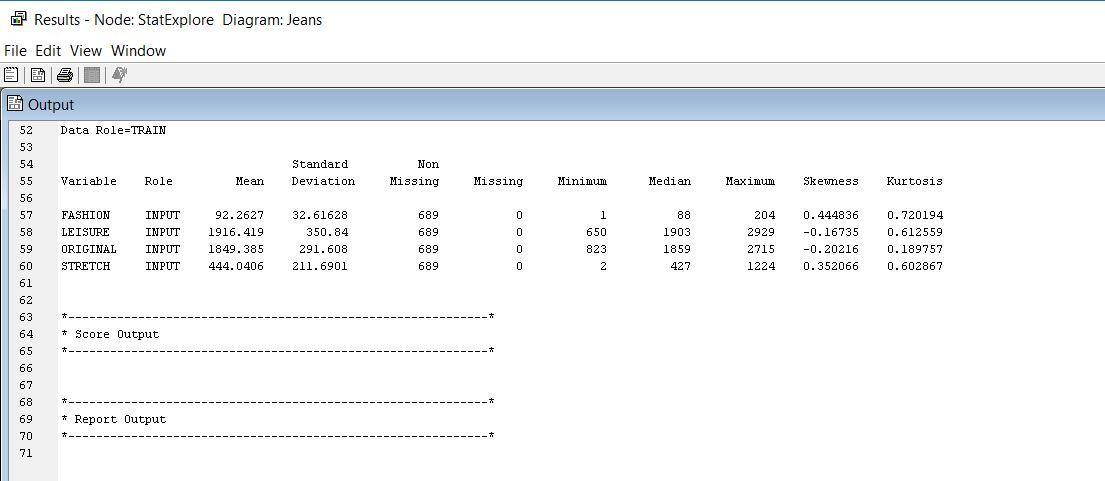
 

**Original:** **Stretch:**

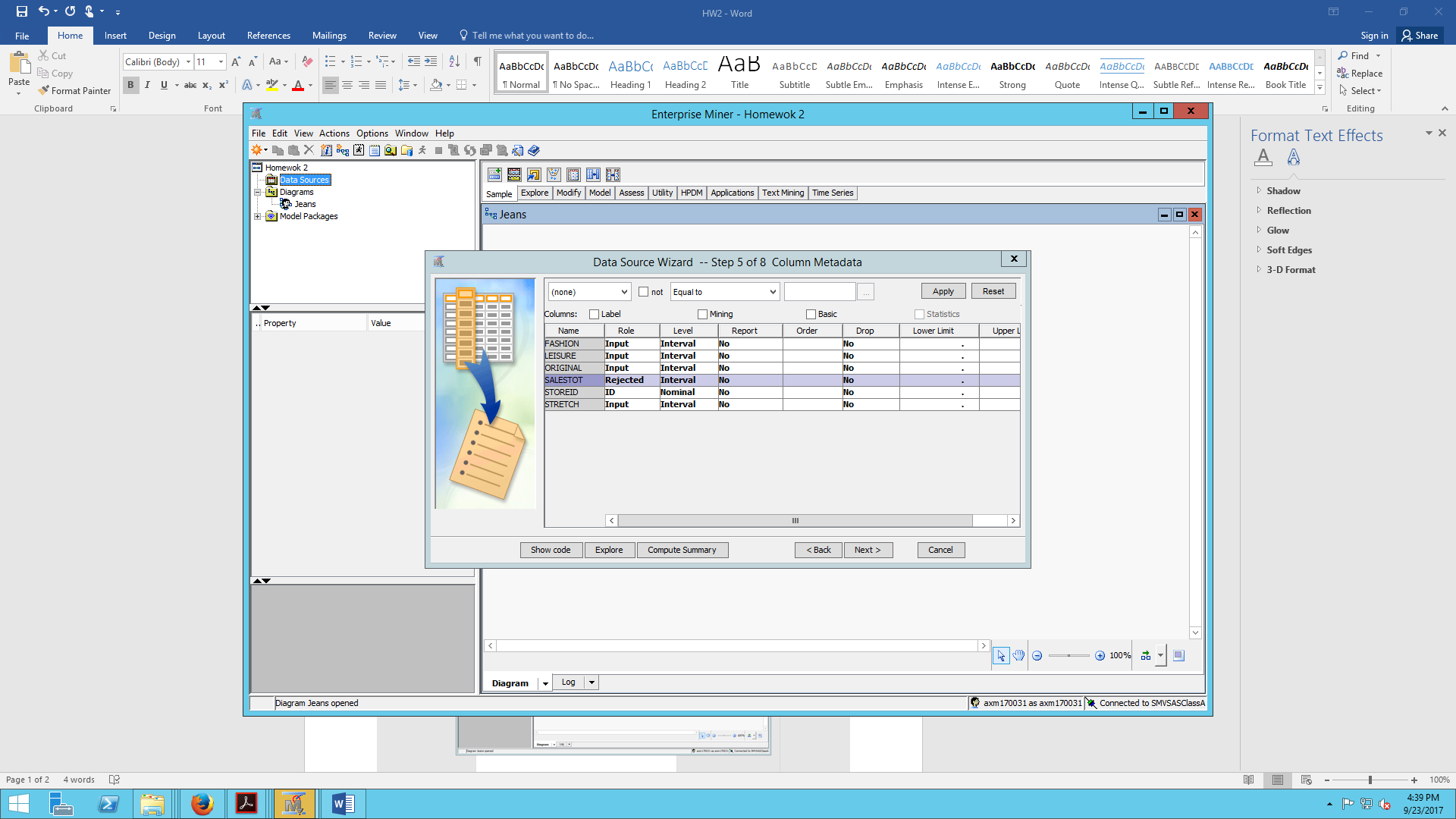
The above box plots prove that there are some outliers.

Checking for the missing values.



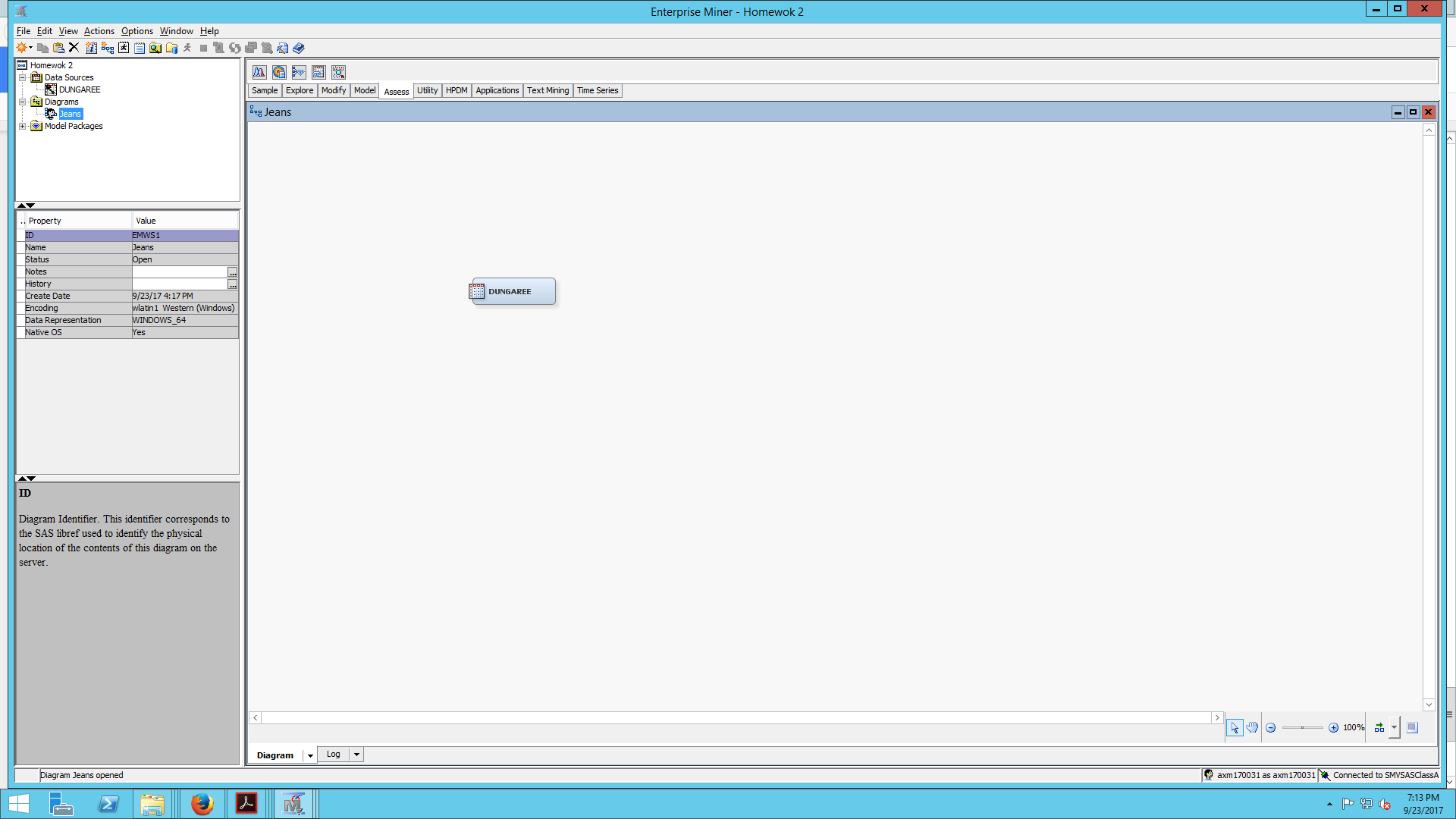
From the Stat Explore option, the above results depict that there are no missing values in the dataset.

**d).** The variables are the respective Model IDs as mentioned.

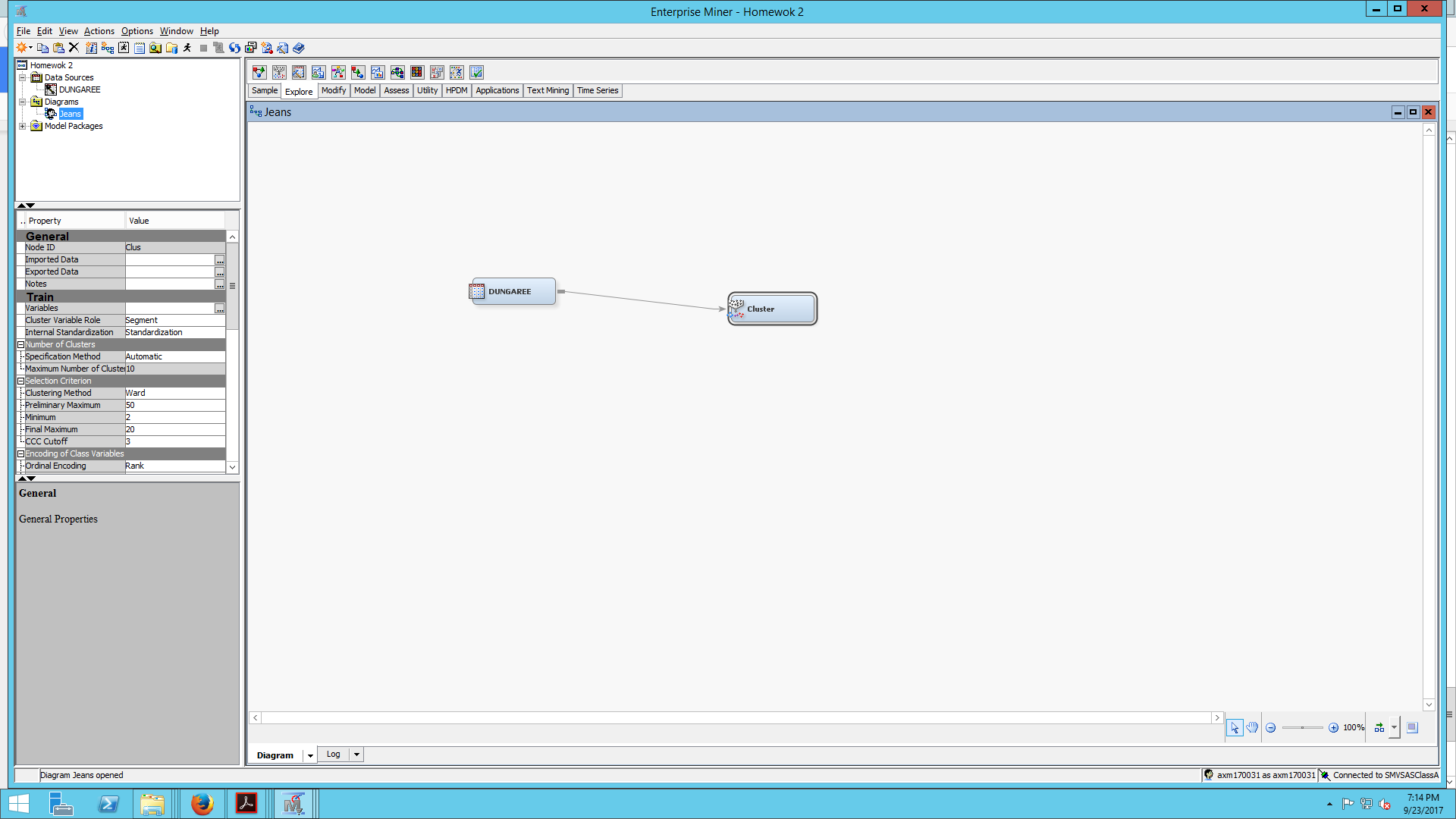


The Role of variable Salestot is set as Rejected because it is not impacting the other attributes of Dungaree. It is just the total of all the sales corresponding to the specific store and it will not be of any use for further Cluster Analysis.

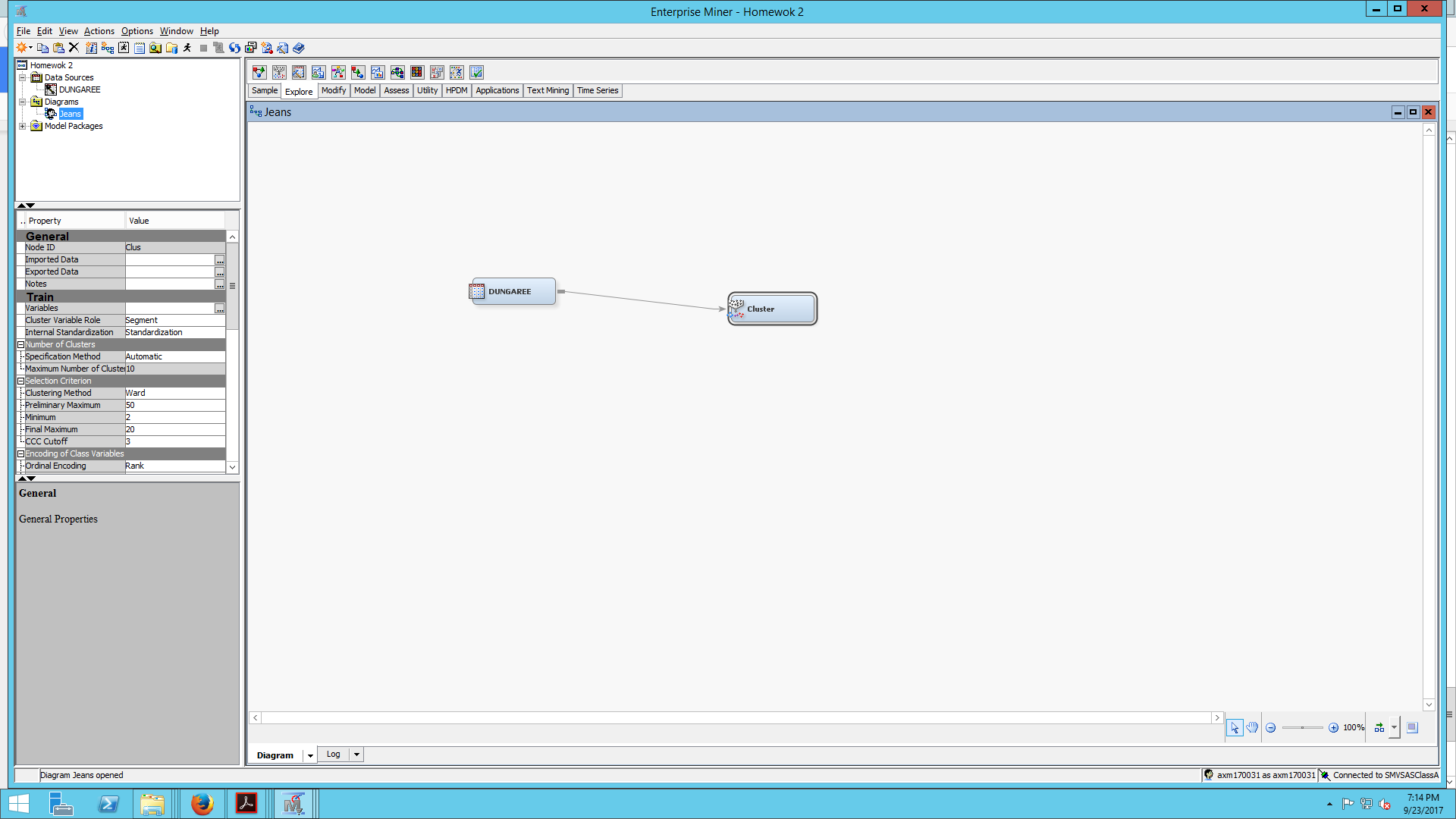
**e).** The Dungaree data source is added to the diagram as shown below.



**f).** A cluster node is added and connected to the data source as depicted below.

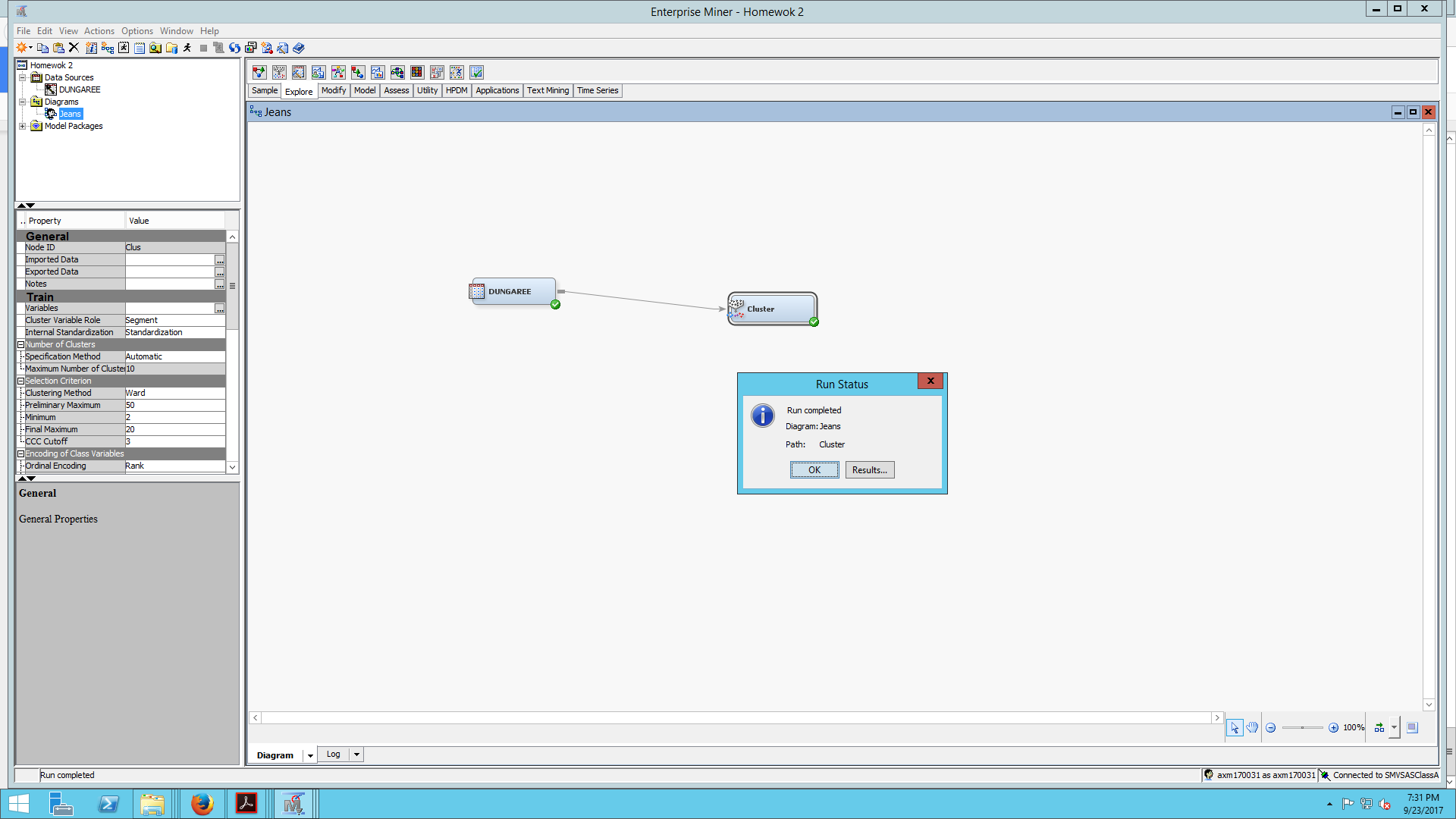


**g).** The ‘**Internal Standardization**’ is set to ‘**Standardization**’ for the Cluster node.

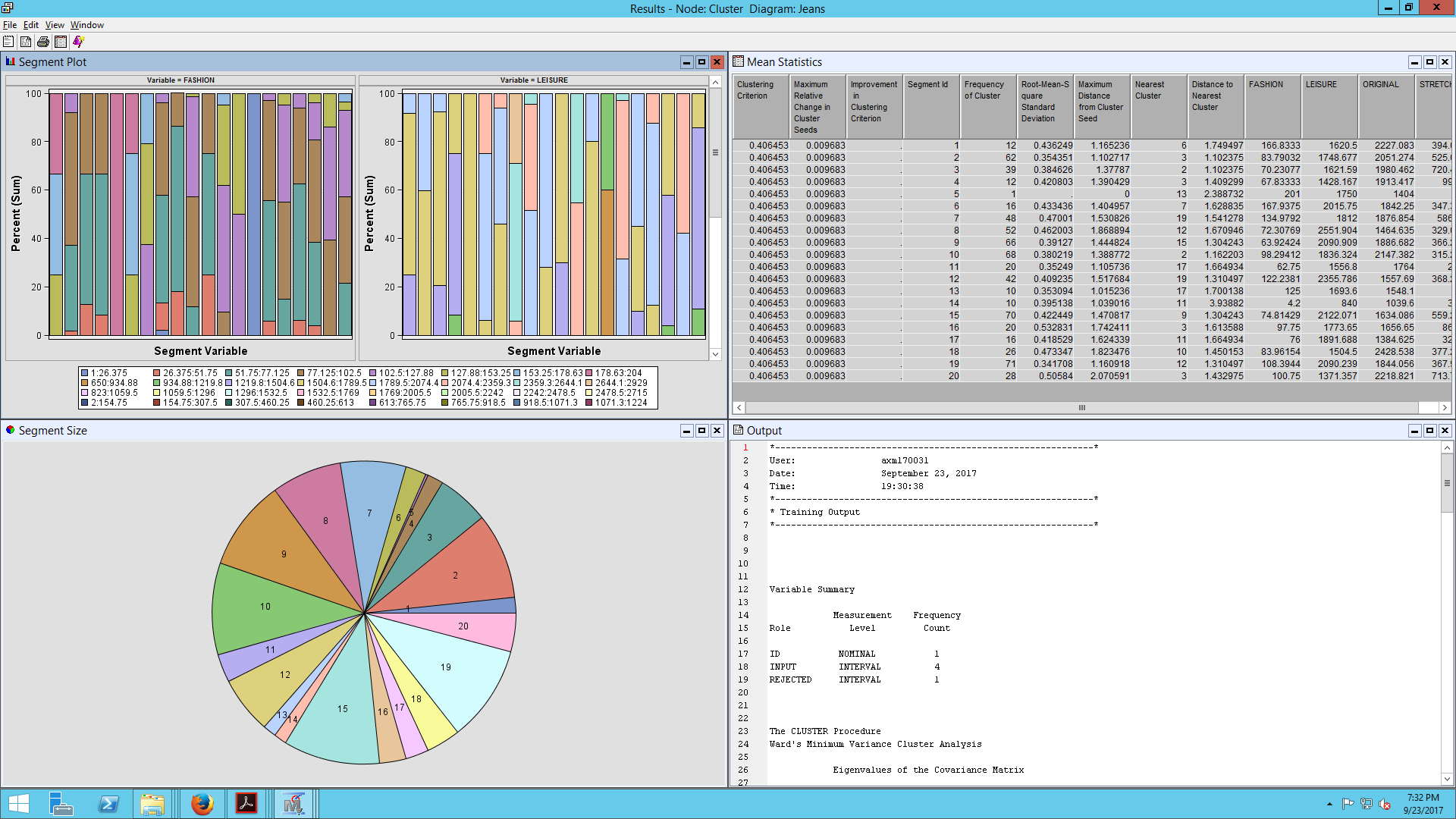


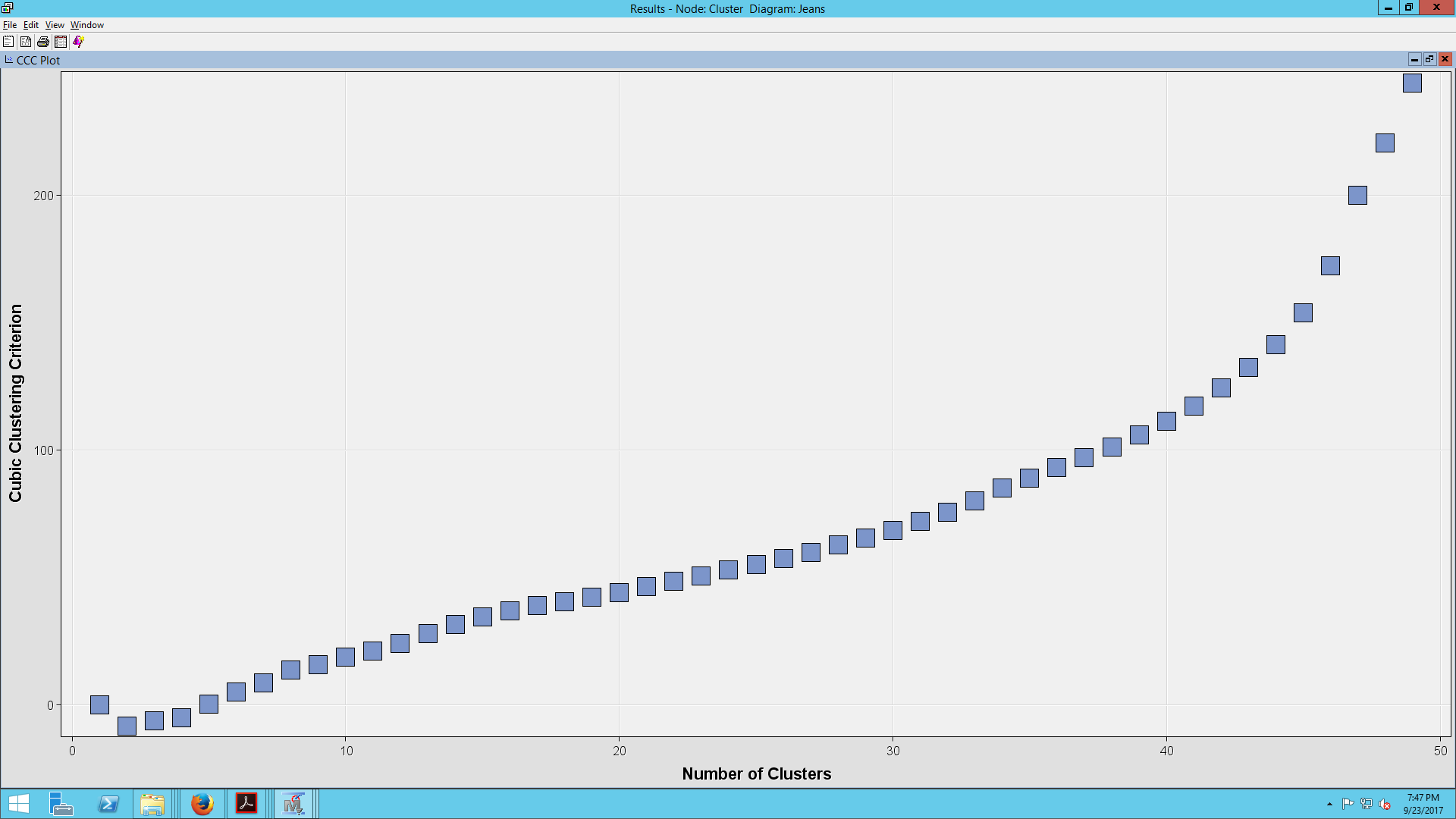
**Standardization** of input data is important because every individual cluster will be defined by its own range of values which may prove undesired and leads to misleading output. If there are differences in the scale of variables, the variables with high ranges would dominate the clustering and affect the precision of the estimation.

**h)**. The cluster node is executed and results are shown below.



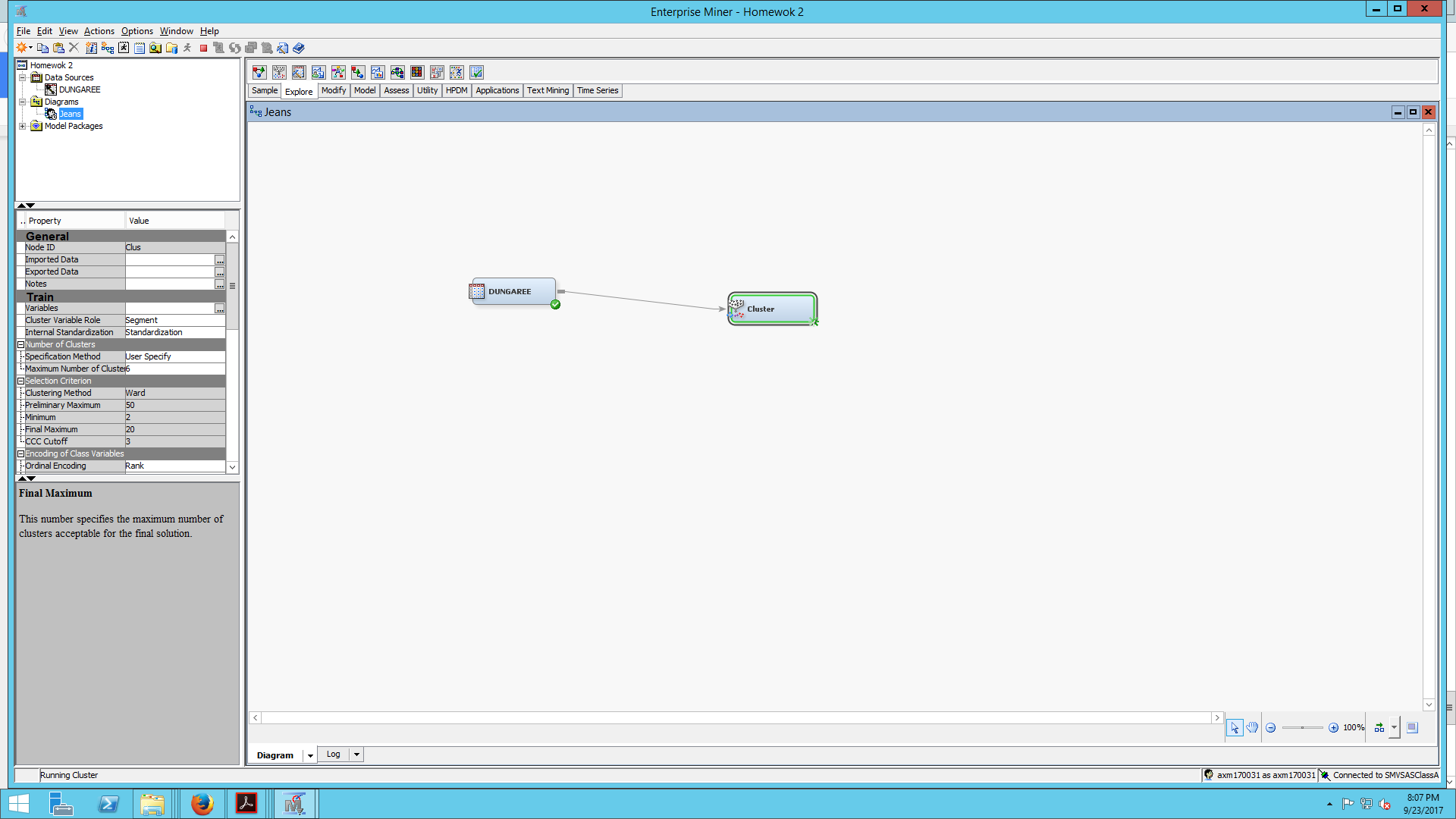
**Cluster Result:**





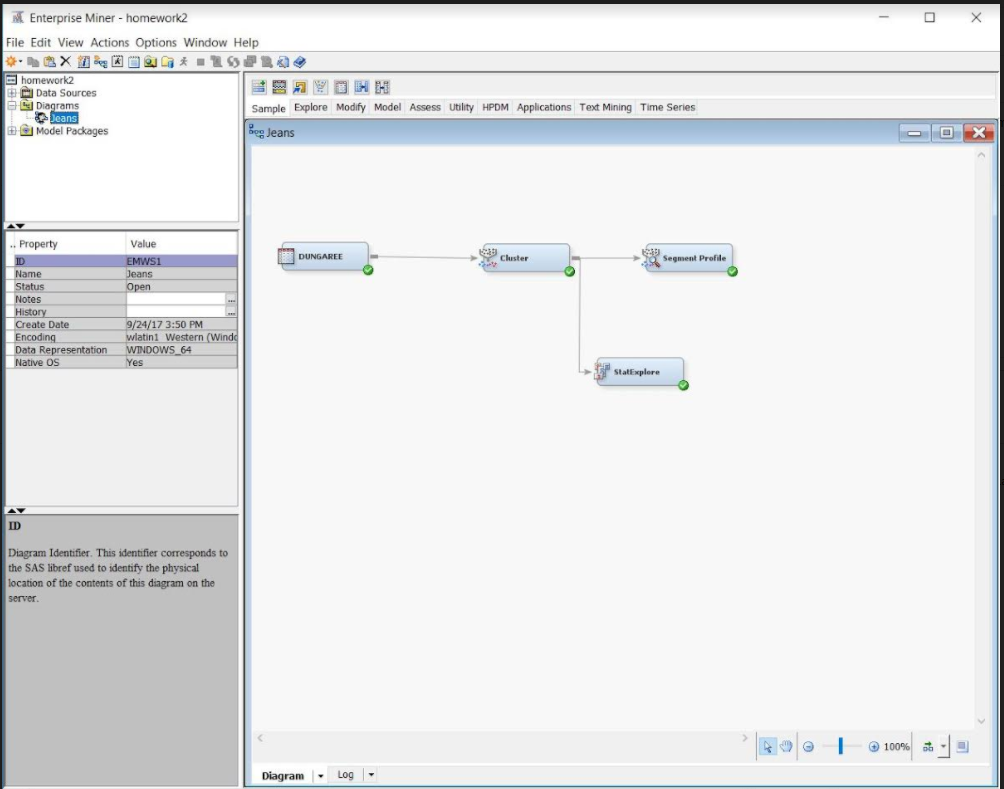
From the pictures above, we can mention that there are 20 clusters in the given data set with a wide range of cluster frequencies. The intra-cluster similarity should be more and inter-cluster similarity should be less. The frequency of clusters is ranging from 1 to 71. This implies that there is a very limited differentiation between them which will, in turn, makes it difficult interpret the relevant output. So, fewer clusters should be incorporated for fairer results.

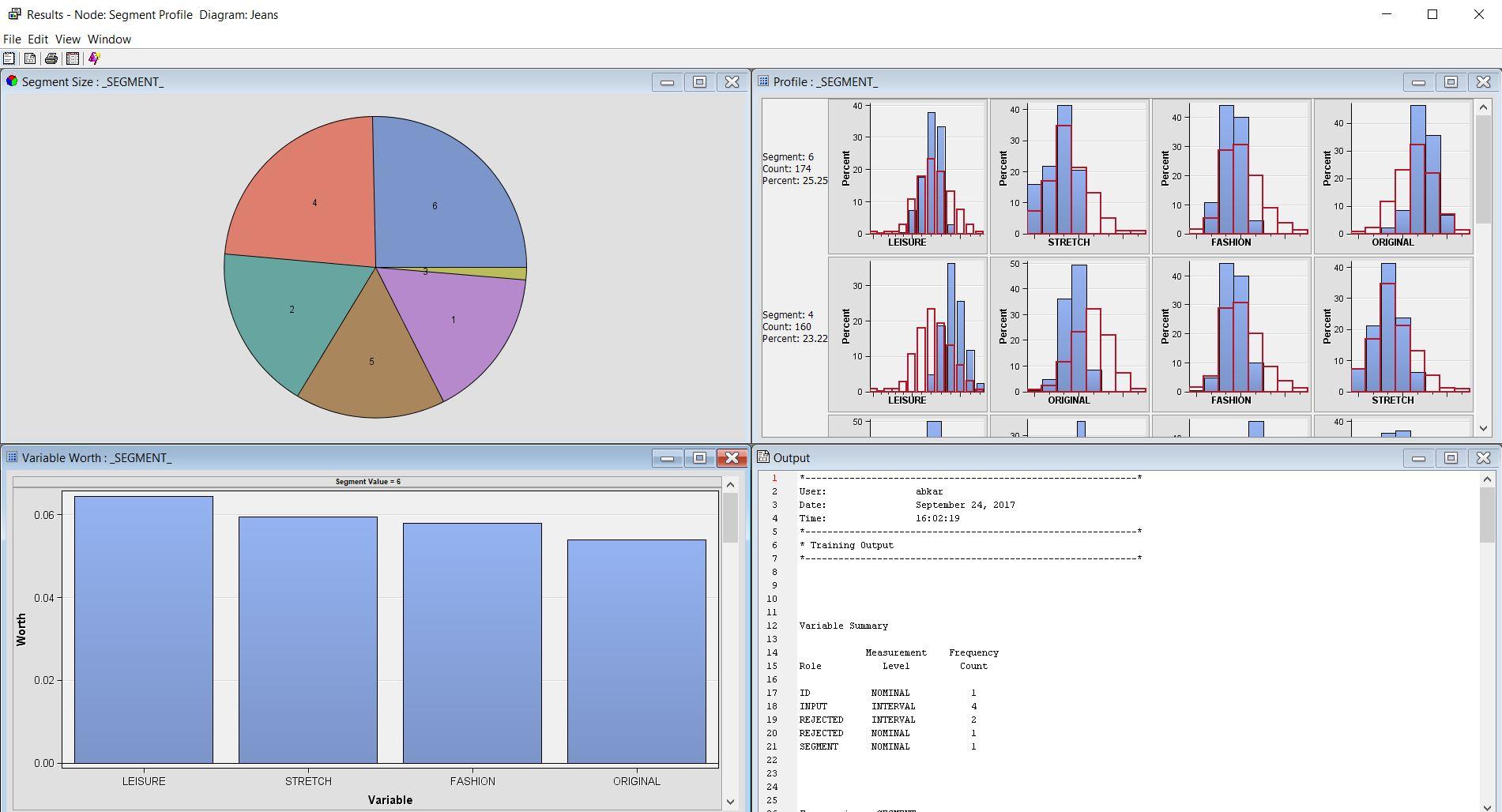
**i).** A maximum of six clusters are specified and the cluster node is rerun.

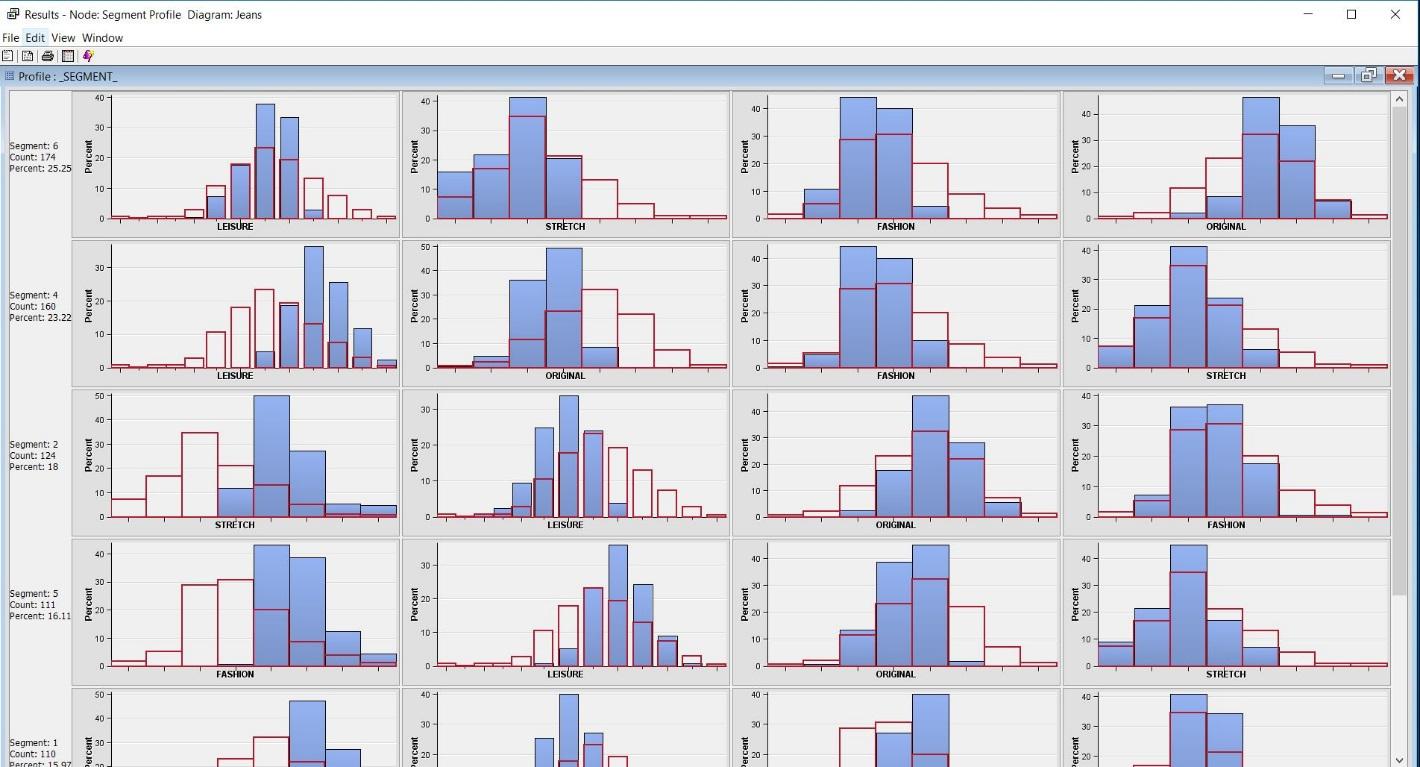


**Result:** The above results, with six clusters, portray the frequencies of clusters ranging from 10 to 174. This implies that all the clusters have more frequency than that of the previous step with twenty clusters except for cluster 3. The quality of clusters is increased as depicted in the distribution shown in the picture. The ‘distance to nearer cluster’ column values are increased when compared to the previous step and thus the intra-cluster similarity is more and inter-cluster similarity is less. This confirms that the analysis is fairer and precise than the previous step with more clusters.

**j).** The segment profile node is added and the results are analyzed.

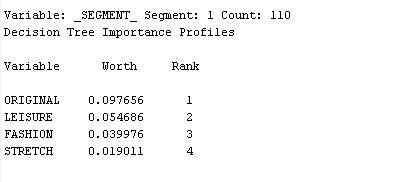




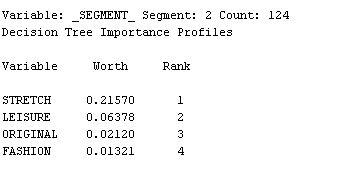


**There are six segments and they are explained below.**

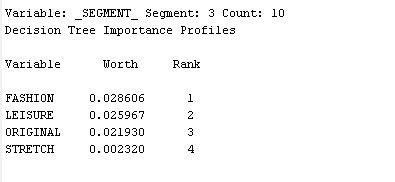
**Segment 1:** The Original and Fashion has a number of jeans sold than average and Leisure and Stretch are less than average. Their worths are mentioned below.



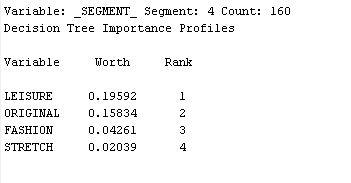
**Segment 2:** Stretch has more than average and Leisure has less in this segment. The worths of Fashion, Stretch, Leisure and Original are below.



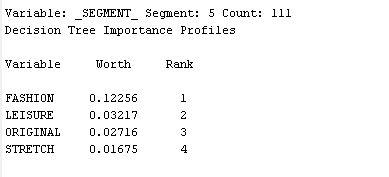
**Segment 3:** Fashion, Leisure and Original have less than an average number of jeans. The worth’s of Fashion, Stretch, Leisure and Original are below.



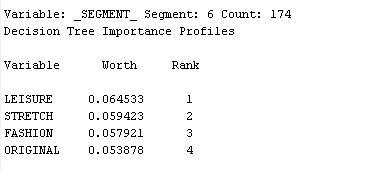
**Segment 4:** Original and Fashion jeans have less than the average and Leisure has more. The worth’s of Fashion, Stretch, Leisure and Original are below.



**Segment 5:** The number of jeans for Fashion is more than the average clearly from the results. The worth’s of Fashion, Stretch, Leisure and Original are below.



**Segment 6:** Original jeans have higher average and Stretch, Fashion has less than the average. The worth’s of Fashion, Stretch, Leisure and Original are below.



**Solution 2:**

**Step 1:** The Euclidian distance is calculated between each individual points. The Distance/ Proximity Matrix for Individual points using Euclidian distance is as below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Points | P1 | P2 | P3 | P4 | P5 |
| P1 | - | 1.230 | 1.347 | 1.115 | 1.243 |
| P2 | 1.230 | - | 2.164 | 2.122 | 1.145 |
| P3 | 1.347 | 2.164 | - | 0.403 | 1.277 |
| P4 | 1.115 | 2.122 | 0.403 | - | 1.429 |
| P5 | 1.243 | 1.145 | 1.277 | 1.429 | - |

**Step 2**: Cluster C1 is formed of point P4 and P3 having the minimum distance of 0.403.



**Step 3**: The rest of the clusters are formed using Single Linkage(Minimum Distance) i.e. by considering the minimum distance(closest) between the points inside the clusters and the rest of the cluster/points. So, the next linkage is created between cluster C1 and point P1 where the distance between P1 and P4 is 1.115 and we name it as Cluster C2.



**Step 4:** The next linkage is created between points P2 and P5 where the distance between P2 and P5 is 1.145 shown by cluster C3. This Cluster is set slightly above cluster C2.



**Step 5:** The next linkage is created between Cluster C2 and C3 where the distance between P1 and P2 is 1.23 shown by cluster C4. Thus, all the points are merged into a single cluster showing Hierarchical Clustering.

**DENDROGRAM**

