Assignment 2. Cluster Analysis - Table of Contents

MIS 576 Data Mining for Business Analytics

Assignment 2 – Exploratory Analysis

(Cluster Analysis is an Unsupervised Data Mining Technique used for Exploratory Analysis)

Assignment Submitted By	**Sudharsan Elangovan**

1. After reading and following the instructions of Assignment 2 Lesson Plan (***Assignment2.ExploratoryAnalysis.LessonPlan.pdf***) available on canvas with ***Assignment 2 – Exploratory Analysis*** start following the step-by-step instructions of this document to perform exploratory analysis using Cluster Analysis. And provide response to all highlighted Questions as mentioned in the steps.

- To recall, our aim in this assignment is to explore the ToyotaCar data set to create homogeneous groups (clusters) of Toyota cars based on identified variables (car characteristics) of our choice and help the dealer or user make better decision of selecting the vehicle based on that classification. Also, we will validate the stability of clusters to generate meaningful insights.

2. Now we will work on the continuation of Step 5.11 of the lesson plan (*Assignment2.ExploratoryAnalysis.LessonPlan.pdf*) available on canvas with ***Assignment 2 – Exploratory Analysis.***

Opening the SAS Enterprise Miner Project on UIS Citrix

3. To recall we have closed the NS.ExploratoryAnalysis Project in step 5.10 of the lesson plan. To open the project and to continue working on the project, launch SAS Enterprise Miner Workstation 15.2 on UIS Citrix and click on Open Project. Navigate to project folder in UIS Citrix and click project.emp file (Figure 1) to launch the project created for exploratory analysis. Navigate to the Diagrams and double click NS.ClusterAnalysis diagram created in lesson plan to open the workspace as shown in Figure 2.

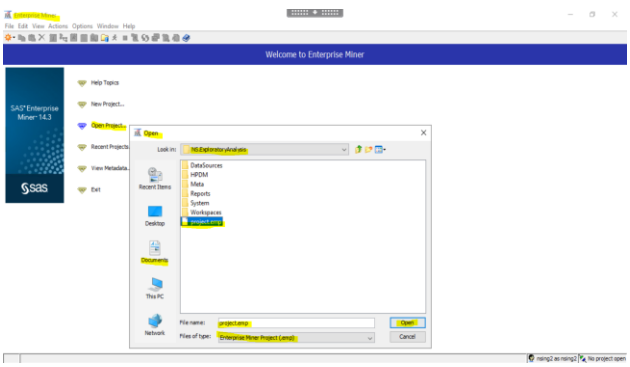


Figure 1

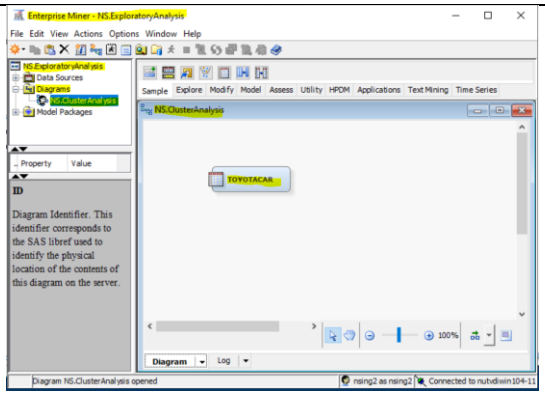


Figure 2

Setting the Roles and Levels of Variables for Cluster Analysis

4. To perform the cluster analysis now we need to set the roles and levels of the variables in our ToyotaCar data set. The description, roles, and variable type (levels), of all the variables of the ToyotaCar data set is provided in the Appendix at the end of this document. As shown in Figure 3 Right click on ToyotaCar node to edit the roles and levels of all the variables by clicking on "Edit Variables".

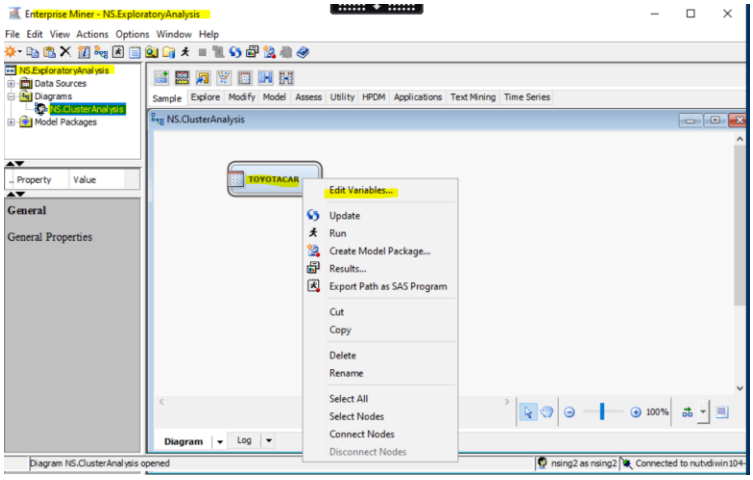


Figure 3

Note: Clicking on "Edit variables…" will open the Edit Variables screen.

- As shown in Figure 4, kindly ensure to change the roles (input/predictor or target/outcome or not used/rejected) and levels (data type: interval or nominal or ordinal) of all the variables as mentioned in Appendix at the end of this document.

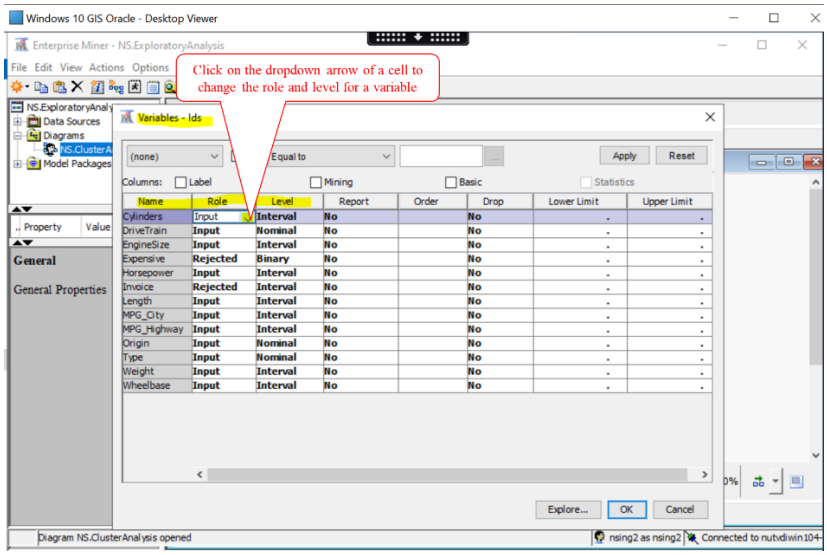
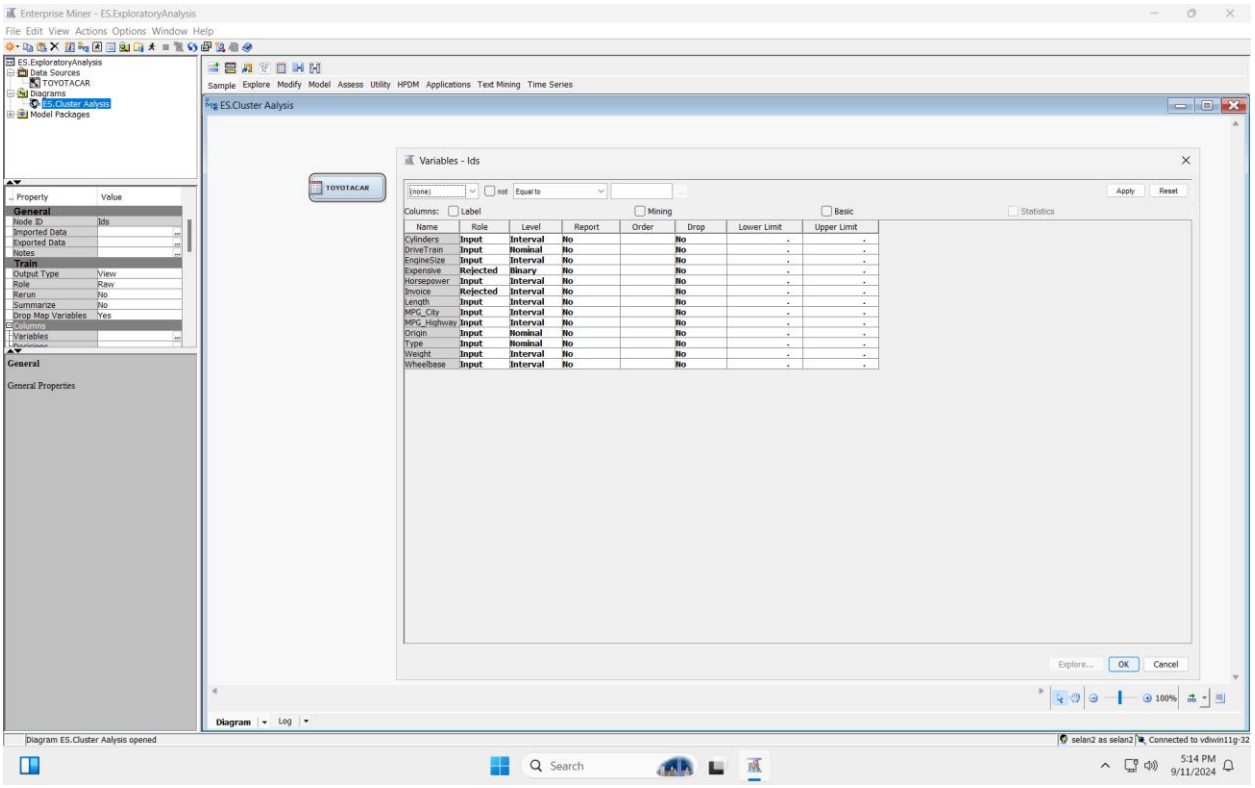


Figure 4

- To recall, **in unsupervised data mining** there is **no defined Target variable**. So, **set the Role of Invoice and Expensive variables as Rejected and the Role of all other variables as Input**. (*Hint: This means at this point you are creating the clusters of Toyota cars using all the variables of ToyotaCar data set except Invoice and Expensive.)*

Question 1

Provide a screenshot of your Edit Variables screen showing the roles and levels of all the variables you have set in step 3. Ensure to capture NetID and machine number at the lower right corner in screenshot (example can be seen as in Figure 5). **(***You can adjust the size of SAS Enterprise Miner window by hovering and dragging the mouse at the corners to better capture the required components in your screenshot.***)**



5. Click OK to close the Edit Variables window.

Filtering Outliers and Standardizing Interval variables for Cluster Analysis

6. To recall we need to filter outliers and standardize the interval variables to perform cluster Analysis. We use the Filter Node in SAS Enterprise Miner to filter the outliers or rare categories. Click on the "Sample" tab on the toolbar to click, drag and drop "Filter" node (the fourth from left) onto the diagram as shown below. (*Note: Filtering the outliers is one the strategies to create more homogeneous clusters.*). Click on the Filter node and explore its properties on the left pane.

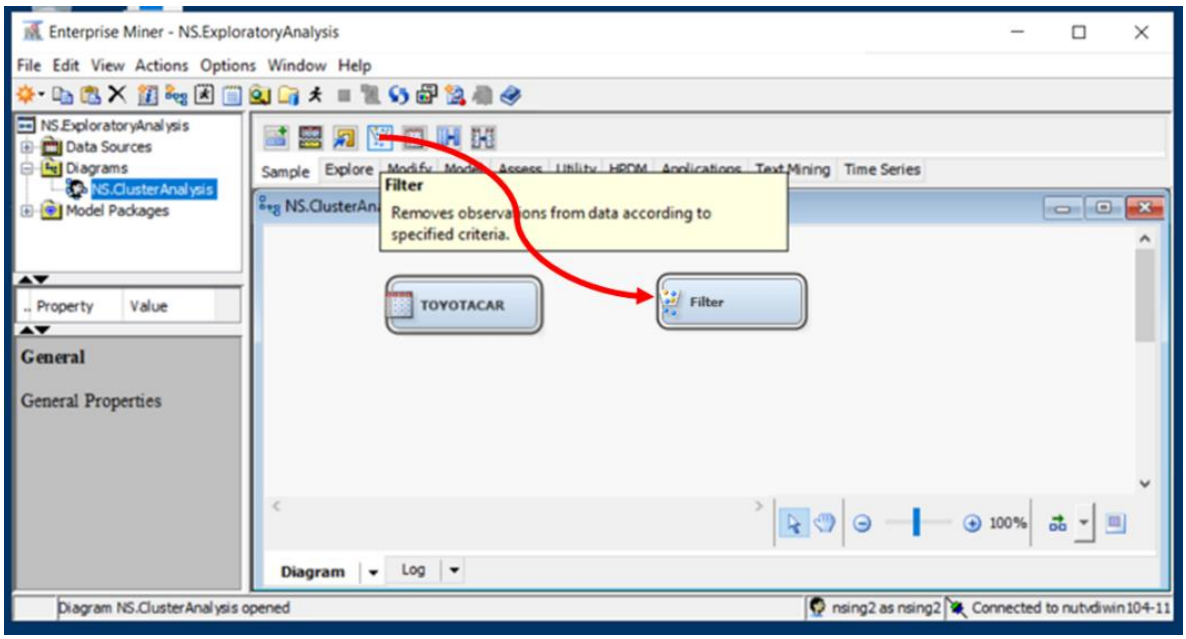


Figure 5

7. Click on the "Explore" tab on the toolbar and click and drag the "Cluster" node (Second node from the left on Explore Tab) to the diagram.

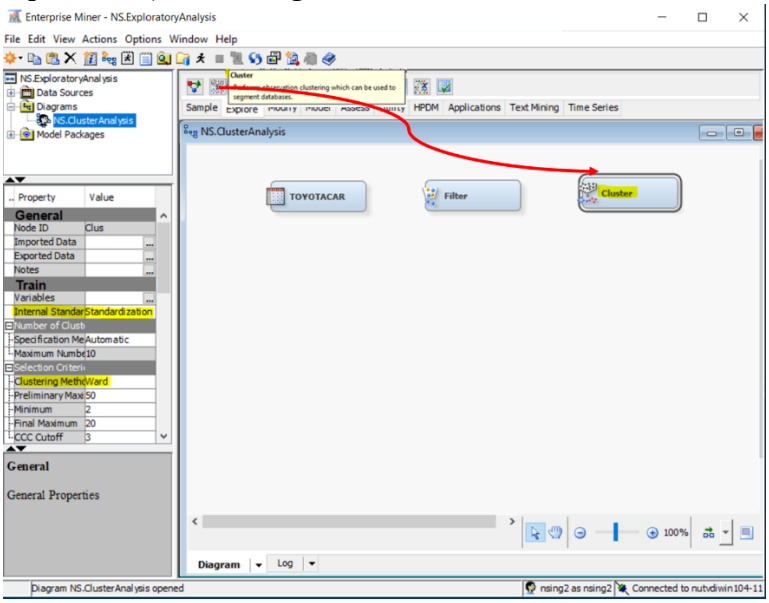


Figure 6

8. Now connect ToyotaCar node to Filter Node and Filter Node to Cluster Node as shown below. To connect two nodes, hover your mouse on the right side of first node and you will see a pencil icon. As soon as you see a pencil icon click your mouse and drag and drop to the left of the next node to connect the nodes. If you have connected the wrong nodes, right click on the connection arrow, and delete the connection.

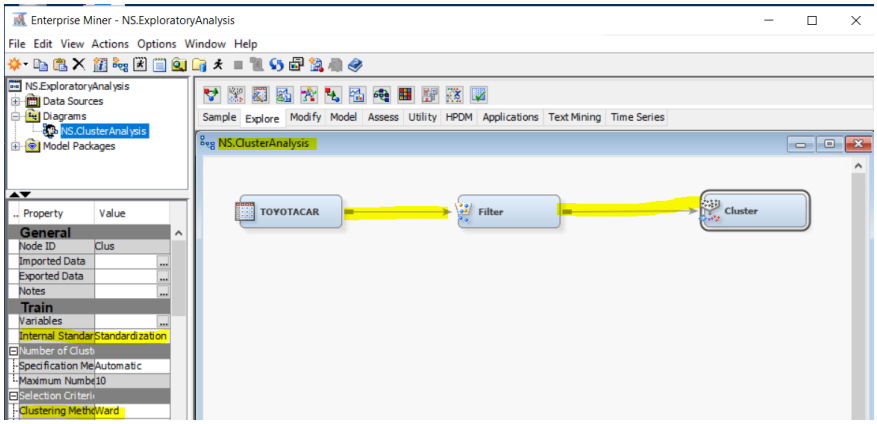


Figure 7

Note: The default Cluster node in SAS Enterprise Miner use Ward Clustering Method to identify the maximum number of clusters.

9. Right click on the Cluster Node to rename it as *Ward Clustering.*
10. Click on *the Ward Clustering* Node to see its properties on the left pane.
11. In the properties window, make sure the value of "Internal Standardization" is set to "Standardization" (*Note: This ensures that all interval variables are standardized. Standardizing interval variables is the second strategy to develop a stable cluster*).

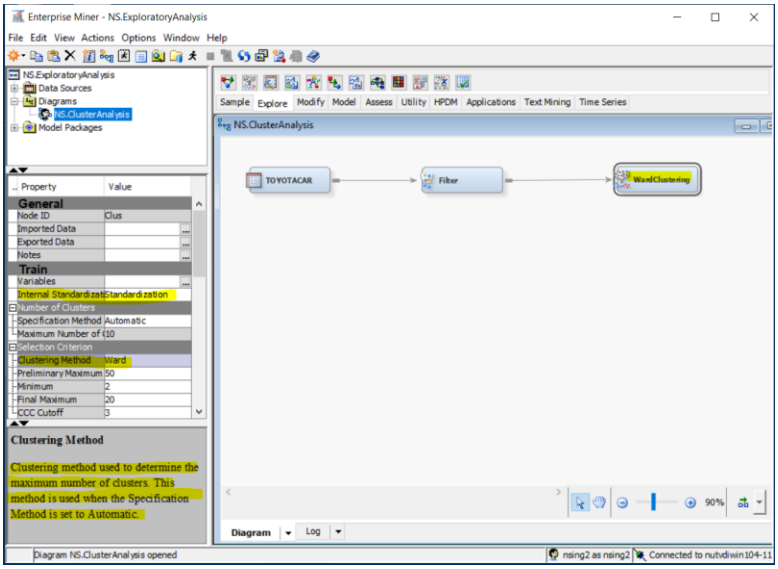


Figure 8

12. Explore some other important properties of Cluster Node by clicking on the properties and reading their description in the lower left pane. Another way is to hover your mouse on the properties to read their tooltip descriptions. For example, the Default Specification Method is "Automatic". It means clustering algorithm will create the clusters. Let us keep all the default

properties as is for the *Ward Clustering* node. However, as mentioned in previous Step ensure "Internal Standardization" is set as "Standardized".

13. Right Click on *Ward Clustering* node and click on "Run" to execute the clustering algorithm. You will receive a dialog box asking you to if you want to run this path? Click Yes and it will take some time to execute the Clustering algorithm. Once the execution is complete you will see a green check mark at the lower right corner of all the connected nodes in this path (Figure 9).

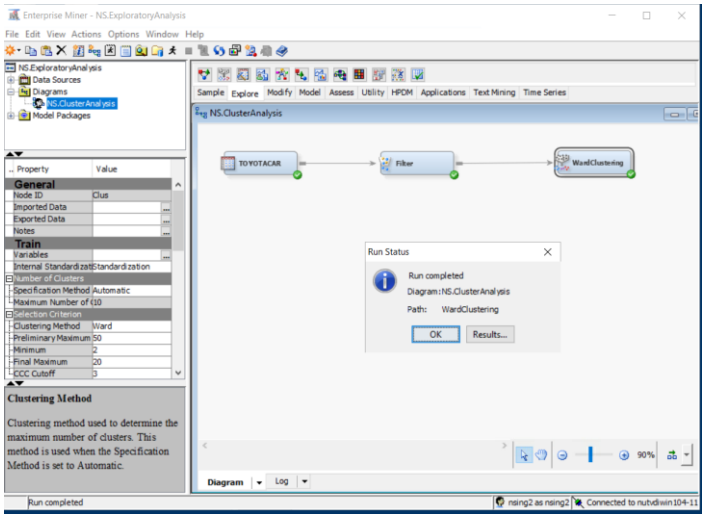


Figure 9

Analyzing the Results of Cluster Analysis

14. After the Run is completed click on "Results" in the "Run Status" dialog box (Figure 9)

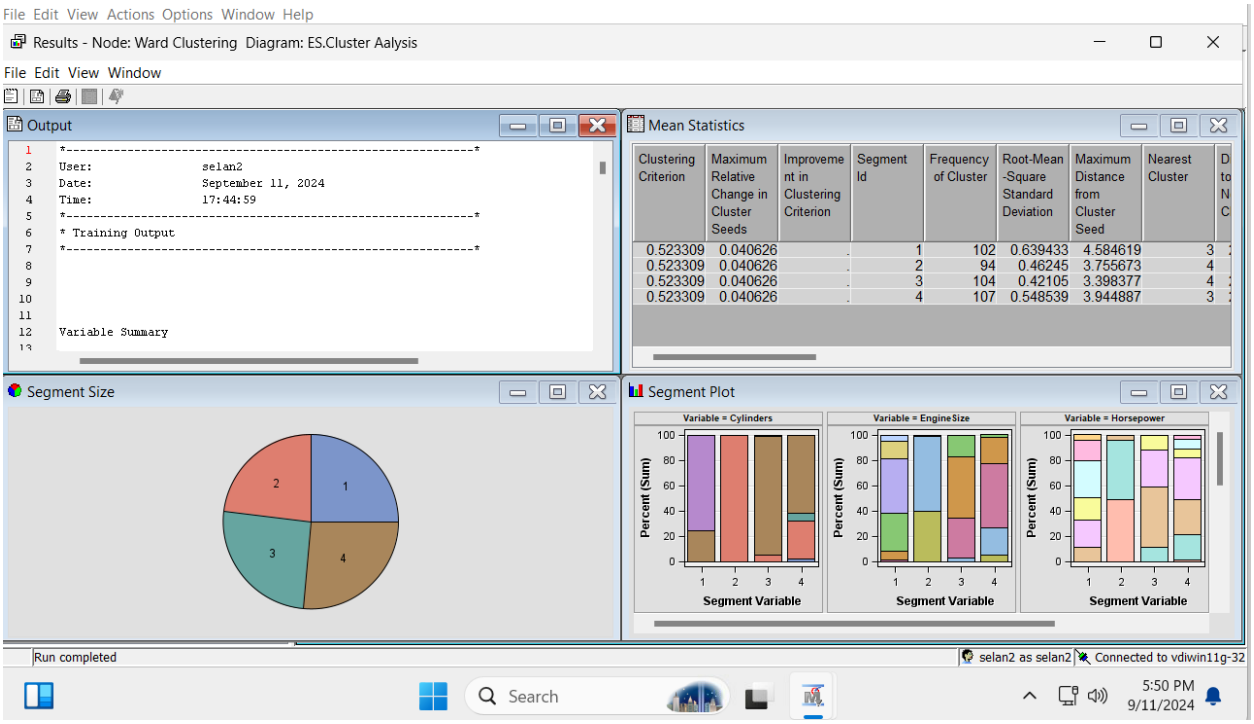
Note: If by mistake you clicked on OK. Do not worry. Just right click on Ward Clustering Node again and you will find the Results option. Just click on it to see the results.

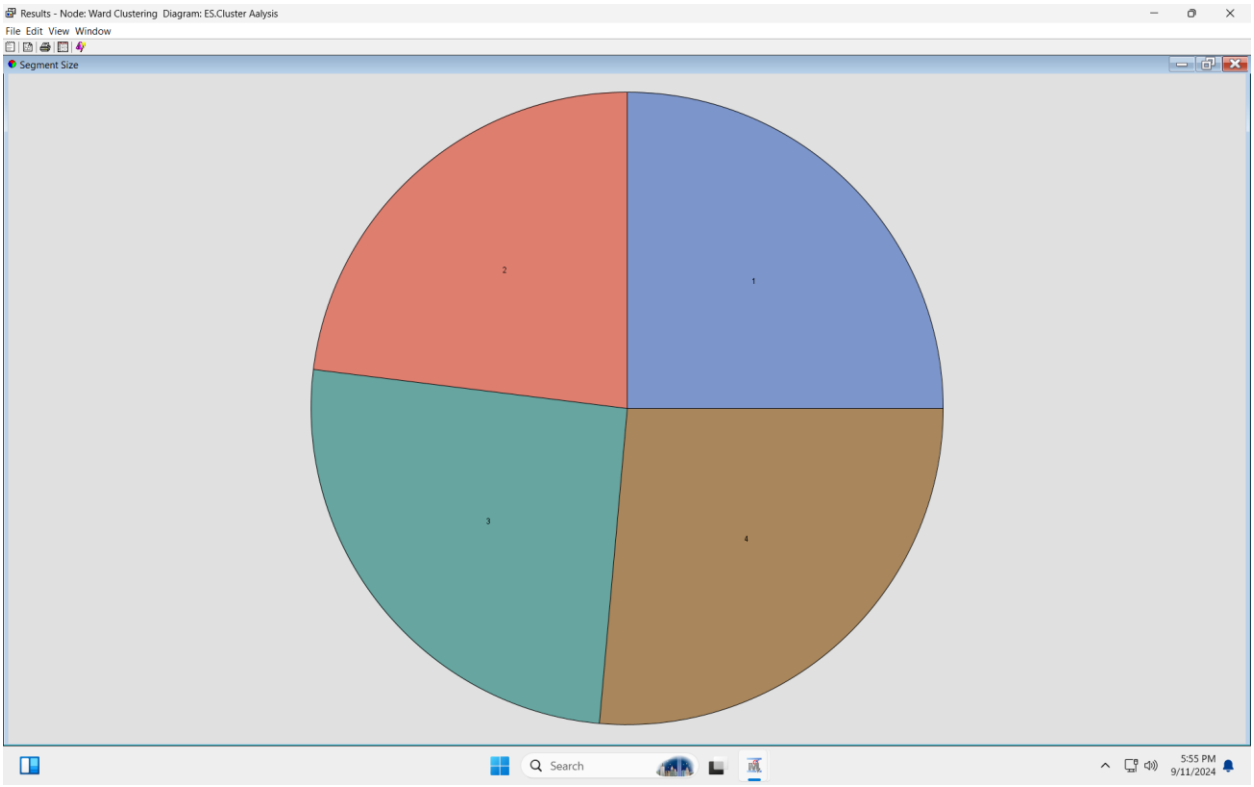
Question 2

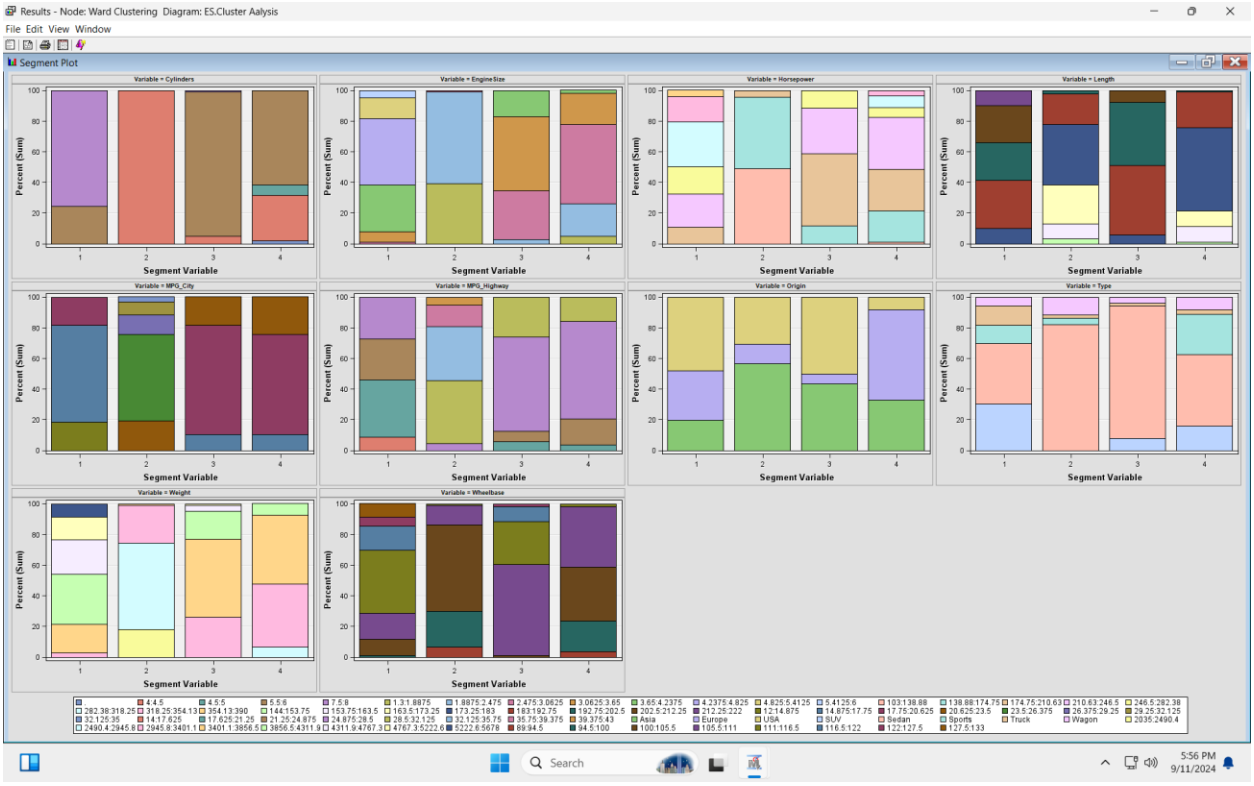
Take the screen shot of Results Window of your Ward Clustering node ensuring you capture:

- **Segment Plot Window**
- **Output Window**
- **Mean Statistics Window**
- **Segment Size Window**

along with your NetID and Machine number. Insert the screenshot of Results window here.

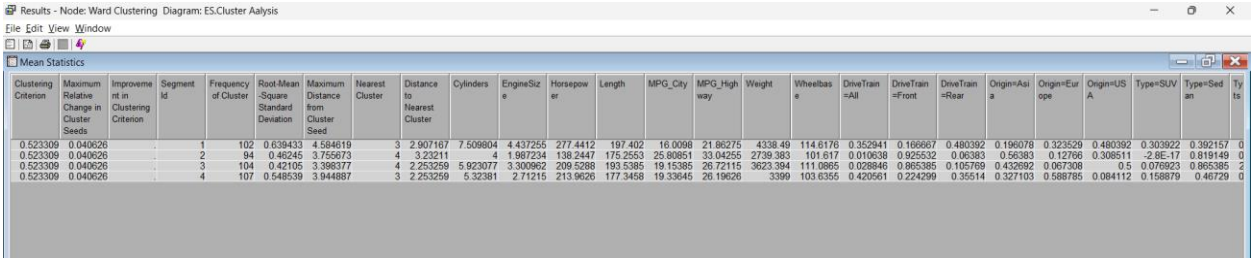

File Edit View Actions Options Window Help
Results - Node: Ward Clustering Diagram: ES.Cluster Aalysis
File Edit View Window
Output
User: selan2
Date: September 11, 2024
Time: 17:44:59
* Training Output
Variable Summary
Mean Statistics
Clustering Criterion
Maximum Relative Change in Cluster Seeds
Improvement in Clustering Criterion
Segment Id
Frequency of Cluster
Root-Mean-Square Standard Deviation
Maximum Distance from Cluster Seed
Nearest Cluster
0.523309 0.040626 1 102 0.639433 4.584619 3
0.523309 0.040626 2 94 0.46245 3.755673 4
0.523309 0.040626 3 104 0.42105 3.398377 4
0.523309 0.040626 4 107 0.548539 3.944887 3
Segment Size
Segment Plot
Variable = Cylinders
Variable = EngineSize
Variable = Horsepower
Percent (Sum)
Segment Variable
Run completed
selan2 as selan2
Connected to vdiwin11g-32
Search
5:50 PM
9/11/2024


Results - Node: Ward Clustering Diagram: ES.Cluster Aalysis
File Edit View Window
Segment Size
Search
5:55 PM
9/11/2024


Results - Node: Ward Clustering Diagram: ES.Cluster Aalysis
File Edit View Window
Segment Plot
Variable = Cylinders
Variable = EngineSize
Variable = Horsepower
Variable = Length
Variable = MPG_City
Variable = MPG_Highway
Variable = Origin
Variable = Type
Variable = Weight
Variable = Wheelbase
Percent (Sum)
Segment Variable
Search
5:56 PM
9/11/2024

Output

```
*------------------------------------------------------------*
User:                  selan2
Date:                  September 11, 2024
Time:                  17:44:59
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*



Variable Summary

             Measurement    Frequency
Role            Level         Count

INPUT         INTERVAL          8
INPUT         NOMINAL           3
REJECTED      BINARY            1
REJECTED      INTERVAL          1



The CLUSTER Procedure
Ward's Minimum Variance Cluster Analysis

            Eigenvalues of the Covariance Matrix

        Eigenvalue    Difference    Proportion    Cumulative

   1    413312.144    411781.568        0.9961        0.9961
   2      1530.576      1444.005        0.0037        0.9998
   3        86.570        81.716        0.0002        1.0000
   4         4.854         0.638        0.0000        1.0000
   5         4.217         3.660        0.0000        1.0000
   6         0.557         0.169        0.0000        1.0000
   7         0.387         0.180        0.0000        1.0000
   8         0.207         0.038        0.0000        1.0000
   9         0.169         0.069        0.0000        1.0000
  10         0.100         0.010        0.0000        1.0000
  11         0.090         0.009        0.0000        1.0000
  12         0.081         0.038        0.0000        1.0000
  13         0.044         0.015        0.0000        1.0000
  14         0.029         0.008        0.0000        1.0000
  15         0.021         0.010        0.0000        1.0000
  16         0.011         0.011        0.0000        1.0000
  17         0.000         0.000        0.0000        1.0000
  18         0.000         0.000        0.0000        1.0000
  19        -0.000                    -0.0000        1.0000

Root-Mean-Square Total-Sample Standard Deviation    147.7803

Root-Mean-Square Distance Between Observations    910.9792


                                                Cluster History

  Number
      of                                                 Semipartial                Pseudo F       Pseudo
Clusters    ------Clusters Joined-------         Freq       R-Square    R-Square    Statistic    t-Squared    Tie

      40    CL17          CL23                    2         0.0000        1.00         1655         82.4
```



Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Cylinders	EngineSize	Horsepower	Length	MPG_City	MPG_Highway	Weight	Wheelbase	DriveTrain=All	DriveTrain=Front	DriveTrain=Rear	Origin=Asia	Origin=Europe	Origin=USA	Type=SUV	Type=Sedan	Ty ts
0.523309	0.040626		1	102	0.639433	4.584619	3	2.907167	7.509804	4.437255	277.4412	197.402	16.0098	21.86275	4338.49	114.6176	0.352941	0.166667	0.480392	0.196078	0.323529	0.480392	0.303922	0.392157	0
0.523309	0.040626		2	94	0.46245	3.755673	4	3.23211	4	1.987234	138.2447	175.2553	25.80851	33.04255	2739.383	101.617	0.010638	0.925532	0.06383	0.56383	0.12766	0.308511	-2.8E-17	0.819149	0
0.523309	0.040626		3	104	0.42105	3.398377	4	2.253259	5.923077	3.300962	209.5288	193.5385	19.15385	26.72115	3623.394	111.0865	0.028846	0.865385	0.105769	0.432692	0.067308	0.5	0.076923	0.865385	2
0.523309	0.040626		4	107	0.548539	3.944887	3	2.253259	5.32381	2.71215	213.9626	177.3458	19.33645	26.19626	3399	103.6355	0.420561	0.224299	0.35514	0.327103	0.588785	0.084112	0.158879	0.46729	0

15. Look at the "*Segment Size*" window. This window shows the total number of clusters and the number of records in each cluster. The size of each pie shows the number of records in each cluster. You can hover your cursor on each cluster. Segment id shows the cluster id, and Frequency shows the number of records in that cluster.

Question 3

How many clusters are represented in your "Segment Size" Window?	**4**
Which Cluster has the maximum number of records? Also, provide the number of records for this cluster.	**Segment id : 4** **Number of records : 107**
Which Cluster has the minimum number of records? Also, provide the number of records for this cluster.	**Segment id: 2** **Number of records : 94**

16. Maximize the "*Segment Size*" window. Right click on pie chart and click *Graph Properties* to represent more information on the pie chart. In the "Properties – Pie" dialog box, ensure *Pie* is selected on the left pane. Now ensure that *Category Name, Value, Percentage* checkboxes are checked. Click on *Apply* and then Click on *OK*.

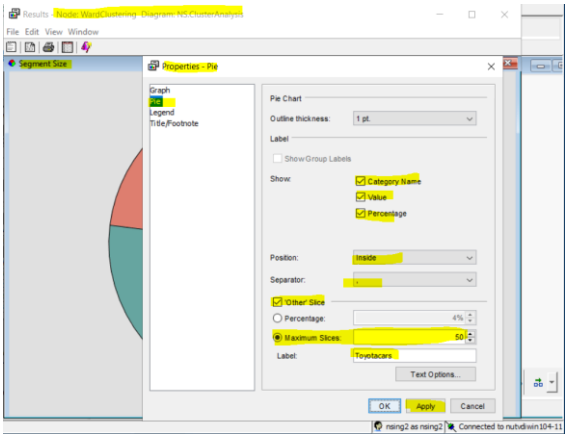


Figure 10

17. Now you will see that your pie chart includes the Segment id, number of records and percentage of records in each cluster.
18. Right click on the pie chart in *Segment Size* window again and click *Graph Properties* to add Legends to your pie chart. In the "Properties – Pie" dialog box, ensure *Legend* is selected on

the left pane and this change the dialog box name to *Properties - Legend*. Now ensure that both *Show* checkboxes are checked for the Legend properties dialog box. You can also select the position of your Legend by choosing the values from *Position dropdown.* Click on *Apply* and then Click on *OK*.

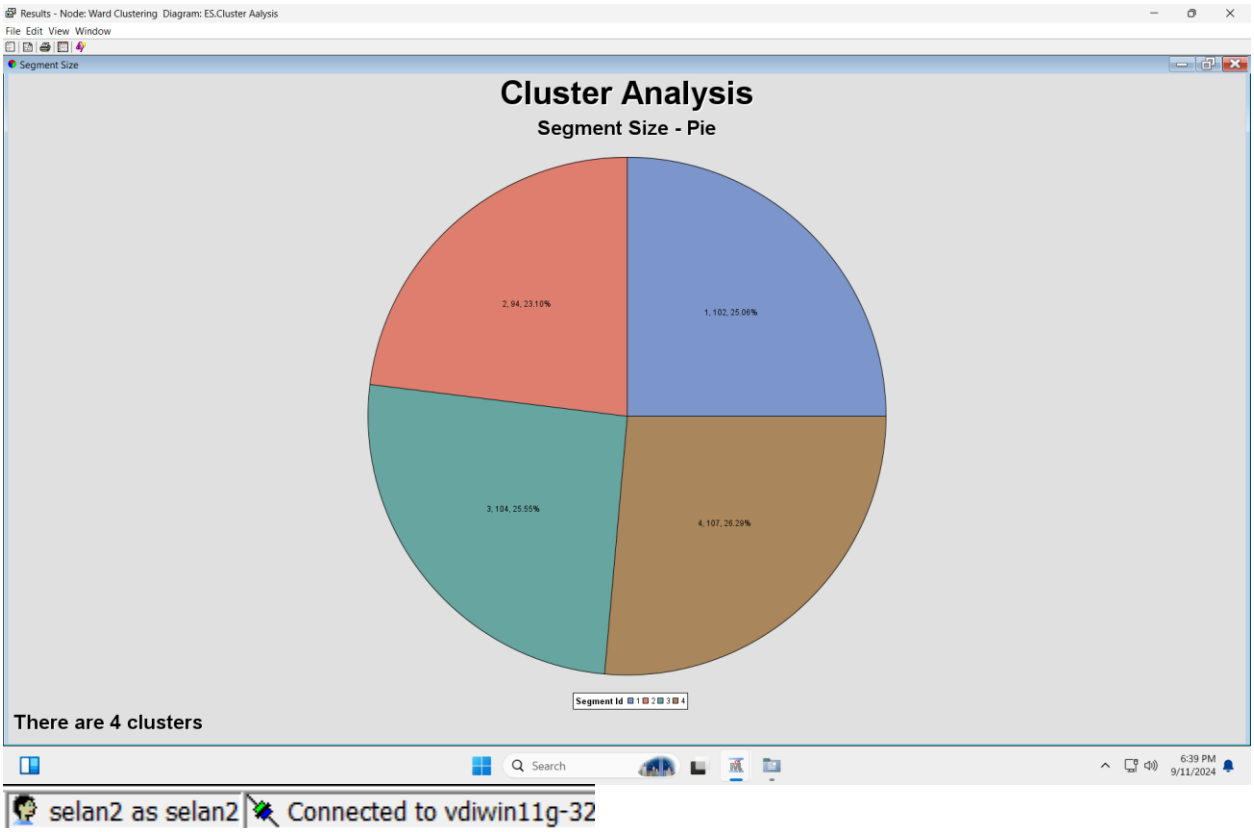
19. Now you will see that your pie chart includes the Legend in addition to the Segment id, number of records and percentage of records.
20. Now right click on Pie Chart and click *Graph Properties* to add title to your pie chart. In the properties window click on *Title/Footnote* property on left pane. Check *Show Title* check box and add a Title of your choice. You can also add subtitle and footnote for your pie chart. Click on Apply and then click on OK.
21. Now you will see that your pie chart includes the Title in addition to the Legend, Segment id, number of records and percentage of records.

Question 4

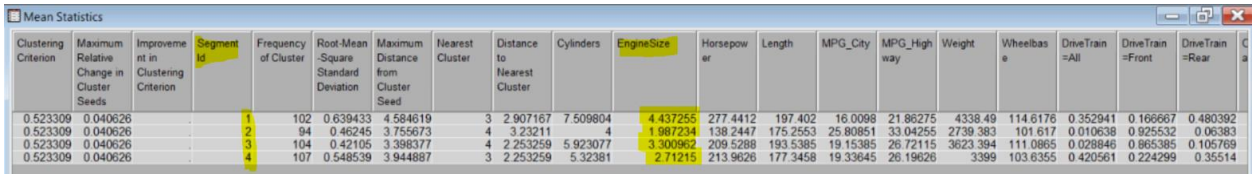
After completing steps 16 -21 take the screen shot of *Segment Size* Window ensuring it represents:

- **Title of Pie Chart as set in Step 20 and 21**
- **Legend in Pie Chart as set in Step 18 and 19**
- **Segment id, number of records and percentage of records as set in Step 16 and 17**

along with your NetID and Machine number. Insert your screenshot here.



22. Now minimize the *Segment Size* window to its original size by clicking on the square box on the upper right corner.
23. Pie chart in Segment size window gives the percentage of records in each cluster. Now if we want to see the mean values of input variables for each cluster, ToyotaCar Clusters we need to explore the *Mean Statistics* window in the Result window of *Ward Clustering* node.
24. Maximize the "*Mean Statistics*" window. Here, we can see the mean values of input variables for each cluster. For example, as shown in below screenshot the mean EngineSize of cars in Cluster 1 is 4.438, in Cluster 2 cars have mean EngineSize of 1.987, Cluster 3 has 3.30 mean EngineSize and Cluster 4 has cars with mean EngineSize of 2.712. *Overall, after exploring the mean statistics of most of the variables we can interpret that Cluster 1 comprises of cars with highest EngineSize (average 4.43), highest Horsepower (average 277), longest cars (average 197), lowest MPG in city (average 16), lowest MPG on highway (average 21), heaviest cars (average 4338), maximum wheelbase (average 114), and mostly Rear Wheel drive cars (48% of 102 cars).*


Mean Statistics

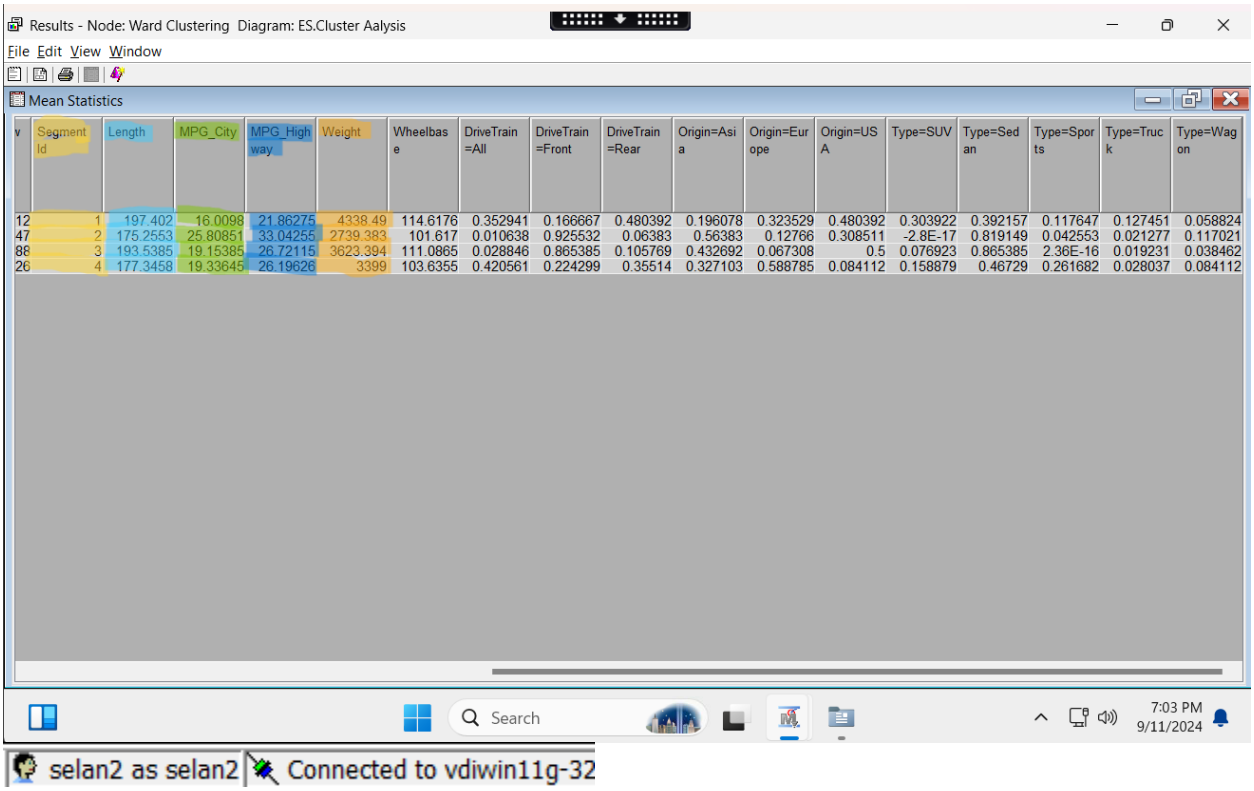
Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Cylinders	EngineSize	Horsepower	Length	MPG_City	MPG_Highway	Weight	Wheelbase	DriveTrain=All	DriveTrain=Front	DriveTrain=Rear
0.523309	0.040626		1	102	0.639433	4.584619	3	2.907167	7.509804	4.437255	277.4412	197.402	16.0098	21.86275	4338.49	114.6176	0.352941	0.166667	0.480392
0.523309	0.040626		2	94	0.46245	3.755673	4	3.23211	4	1.987234	138.2447	175.2553	25.80851	33.04255	2739.383	101.617	0.010638	0.925532	0.06383
0.523309	0.040626		3	104	0.42105	3.398377	4	2.253259	5.923077	3.300962	209.5288	193.5385	19.15385	26.72115	3623.394	111.0865	0.028846	0.865385	0.105769
0.523309	0.040626		4	107	0.548539	3.944887	3	2.253259	5.32381	2.71215	213.9626	177.3458	19.33645	26.19626	3399	103.6355	0.420561	0.224299	0.35514

Figure 11

Question 5

Question 5a

Take the screen shot of Mean Statistics Window from the Results Window of your Ward Clustering node and highlight the Segment IDs and *mean Weight*, *mean Length*, *mean MPG_City, and mean MPG_Highway* of cars for each cluster. Also, ensure to capture your NetID and Machine number. Insert the screenshot here.



Question 5b

Based on the analysis in question 5a answer following questions:

Which cluster has the heaviest car compared to other clusters? Also, provide the mean Weight of cars in this cluster.	**Segment id : 1 Mean Weight : 4338.49**
Which cluster has the longest car as compared to other clusters? Also, provide the mean Length of cars in this cluster.	**Segment id: 1 Mean Length : 197.402**
Which cluster has the car with maximum mileage in city as compared to other clusters? Also, provide the mean MPG_City of cars in this cluster.	**Segment id : 2 Mean MPG_City : 25.80851**
Which cluster has the car with maximum mileage on highway as compared to other clusters? Also, provide the mean MPG_Highway of cars in this cluster.	**Segment id : 2 Mean MPG_Highway : 33.04255**

25. Now let us interpret the values for categorical variables: look at the values of Origin for Cluster 1 in *Mean Statistics* window. These are the values of each Origin Category (Asia, Europe, USA) and shows that 19.6% of the 102 cars in cluster 1 are Asian cars, 32.4% of 102 are European Cars, and 48% of 102 cars are from USA. Where 102 is the total number of cars (observations) in Cluster 1.

Mean Statistics

Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Origin=Asia	Origin=Europe	Origin=USA
0.523309	0.040626	.	1	102	0.639433	4.584619	3	2.907167	0.196078	0.323529	0.480392
0.523309	0.040626	.	2	94	0.46245	3.755673	4	3.23211	0.56383	0.12766	0.308511
0.523309	0.040626	.	3	104	0.42105	3.398377	4	2.253259	0.432692	0.067308	0.5
0.523309	0.040626	.	4	107	0.548539	3.944887	3	2.253259	0.327103	0.588785	0.084112

Figure 12

Question 6

Question 6a

Based on the interpretation in Step 24 and Step 25, determine the percentage of Hybrid cars (Type=Hybrid), Sedan cars (Type=Sedan), SUV (Type=SUV), Trucks (Type=Truck), Sports (Type=Sports), and Wagons (Type=Wagon) in Cluster 4 of your results. Interpret these percentages as I have interpreted the percentages of Origin in Step 24 (50-100 words).

Answer:

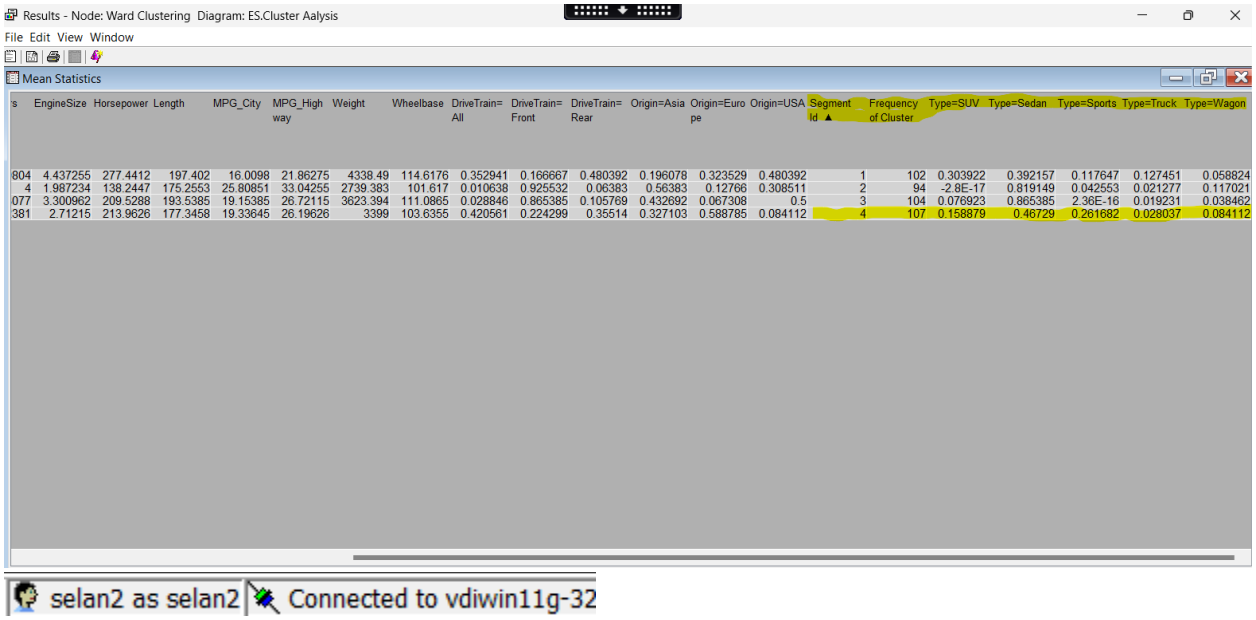
In Cluster 4, 15.88% of the 107 cars in Cluster 4 are SUVs, 46.7% of the 107 cars in Cluster 4 are Sedans, 26.16% of the 107 cars in Cluster 4 are Sports cars, 2.8% of the 107 cars in Cluster 4 are Trucks, 8.4% of the 107 cars in Cluster 4 are Wagons, this means that the majority of cars in Cluster 4 are Sedans (46.7%), followed by Sports cars (26.16%), and SUVs (15.88%).

Question 6b

Take the screen shot of Mean Statistics Window and highlight:

- **Segment ID Column**
- **Frequency of Cluster Column**
- **Type=SUV column and its percentage for Cluster 4**
- **Type=Truck column and its percentage for Cluster 4**
- **Type=Sports column and its percentage for Cluster 4**
- **Type=Wagon column and its percentage for Cluster 4**
- **Type=Hybrid column and its percentage for Cluster 4**

Also, ensure, to capture your NetID and Machine number. Insert the screenshot here.



26. Now minimize the *Mean Statistics* window to its original size by clicking on the square box on the upper right corner.
27. We can also visually find the answer to questions in step 25 using the *Segment Plot* window of *Results* of *Ward Clustering* Node.
28. Let us maximize the *Segment Plot* window (Figure 13). Find the chart titled "Variable=Origin" as highlighted in below screenshot. Hovering your mouse on different clusters of this variable will give you the same information as represented in *Mean Statistics* window.

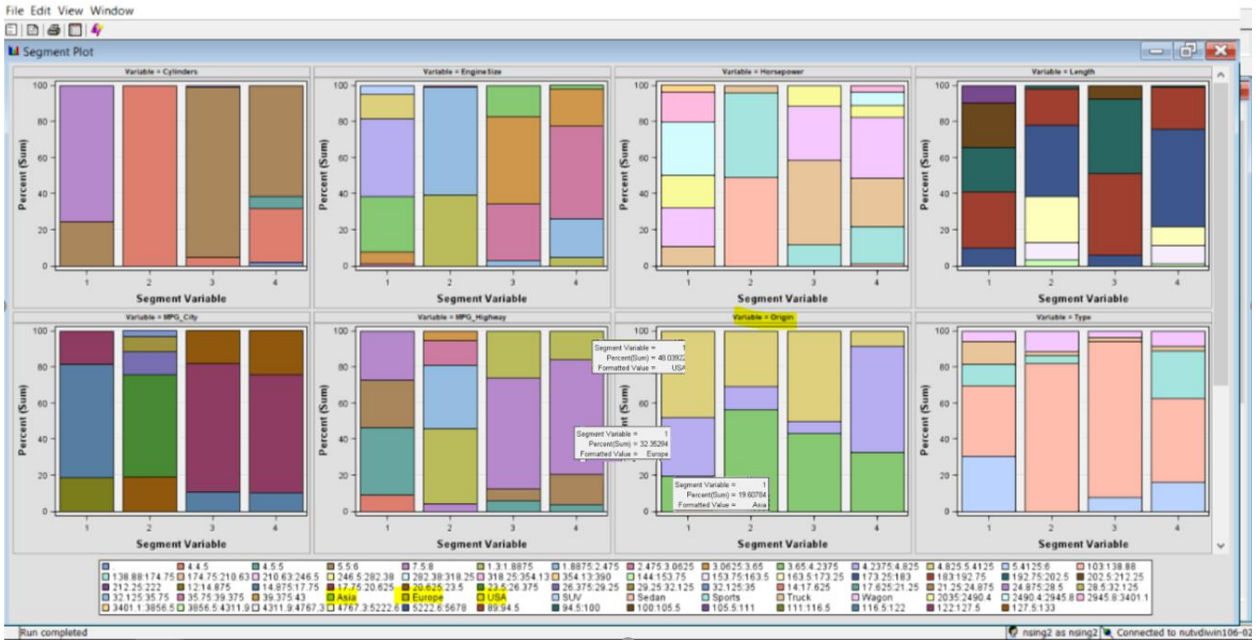
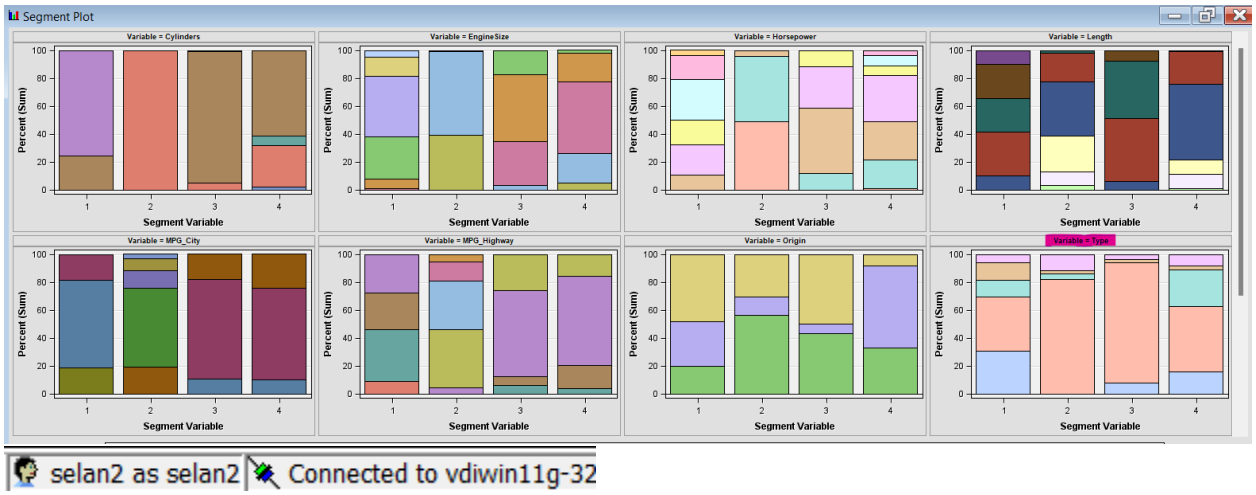


Figure 13

Question 7

Provide the screenshot of your *Segment plot* window highlighting the Variable=Type. Discuss if the information provide by you in question 6a is same as represented in the *segment plot*. Ensure to capture the NetID and machine number while taking screenshot. Insert your screenshot here.



The bar for Cluster 4 reveals a pie chart with multiple colored segments that represent individual car types, and the level of color intensity identifies percentages. The most obvious segment is unsurprisingly Sedans, then Sports Cars, and finally SUVS which very closely resembles my own deduction. Smaller segments include Trucks and Wagons, as already mentioned. Thus, my interpretation(question 6a) seems to fit well with the plot of the segment into Cluster 4. Sedans make up the largest chunk (46.7%), with Sports cars in second at 26.16%, then SUVs close behind on 15.88% and smaller contributions from Trucks (2.8%) and Wagons sit at just under a fifth of all entries as well.

29. Now minimize the *Segment Plot* window to its original size by clicking on the square box on the upper right corner.

Variable Importance in Cluster Analysis

30. Open the Results window of Ward Clustering Node if it is not open.
31. Cluster analysis also shows the variables that are most important in creating clusters. To see the important variables, click on View, Cluster Profile, and Variable Importance.

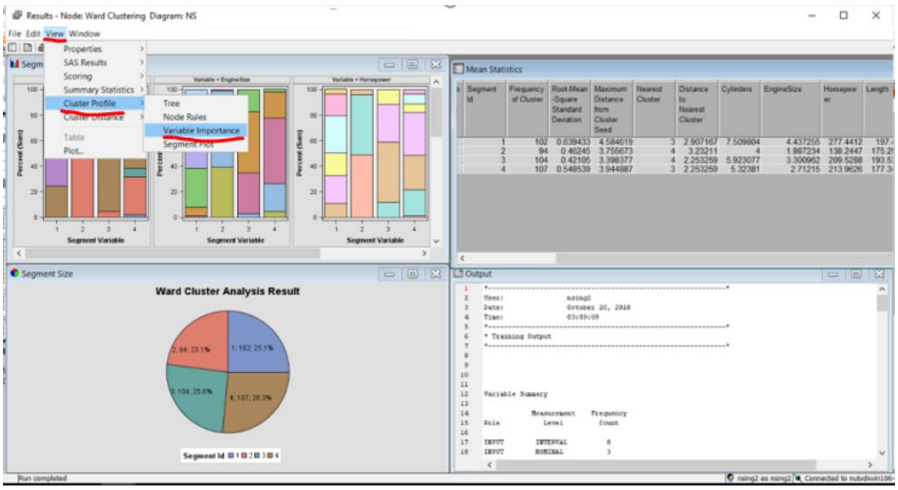


Figure 14

The variables in *Variable Importance* Window are ranked by "*Importance*" (Importance is the last column in variable importance window).

- ***Variables that have an "Importance" value greater than zero are important in creating clusters. The variable with "Importance" value of 1 is the most important variable in creating clusters. On the other hand, an "Importance" value of zero indicates that variable is not important and not used in creating clusters.***

32. Explore and examine your *Variable Importance* window.

Question 8

Question 8a

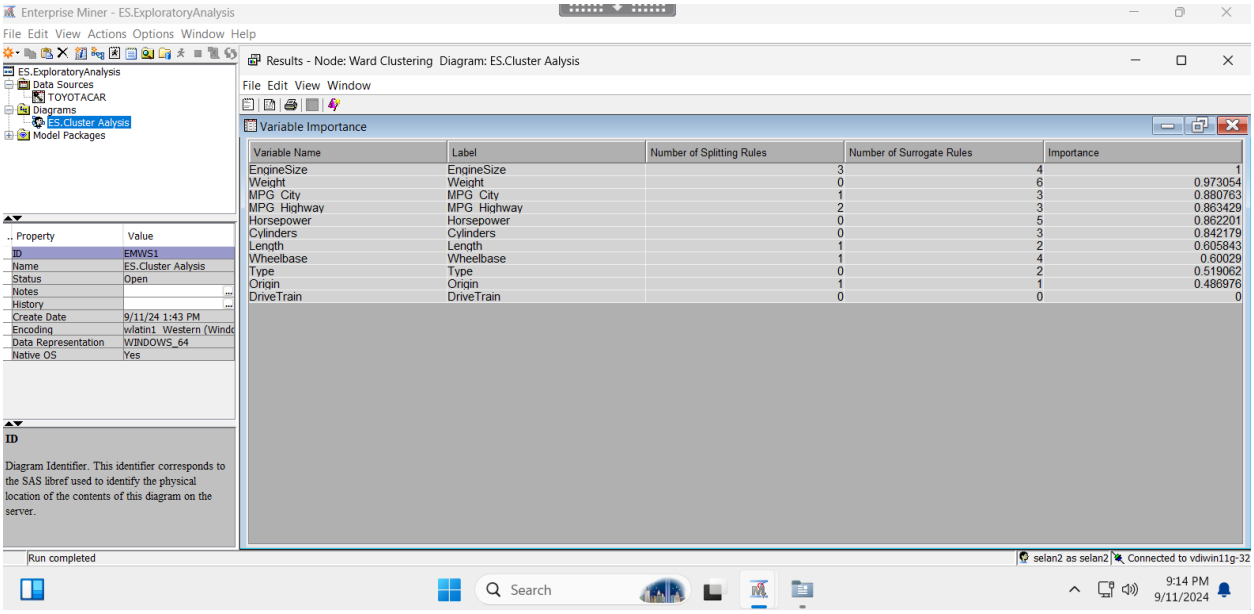
Which variable is most important for creating clusters (*in other words, which variable has the highest importance value***)? Also, identify the variable which is least important for creating clusters (***in other words, which variable has least importance value***).** (*Hint: In the Variable Importance Value window the variables with zero value of "Importance" indicates the variable is not important and is not used in creating the clusters.*)

(Enter the values in below table):

Most Important Variable for creating clusters of cars	Engine Size, Value = 1
Least Important Variable for creating clusters of cars	Drive Train, Value = 0

Question 8b

Provide the screenshot of *Variable Importance* window and ensure to capture your NetID and machine number. Insert your screenshot below.



33. Close the variable importance window.

Analyzing the Results of Cluster Analysis Using Tree structure and Rules

34. In addition to Mean statistics, Segment plot, and Segment size windows, we can also interpret the results of cluster analysis using Node Rules and Trees.
35. Click on *View* menu of *Results* window and Click on *Cluster Profile* > *Tree* to see the results of cluster analysis in the form of tree.
36. Click on *View* menu of *Results* window and Click on *Cluster Profile* > *Node Rules* to see the results of cluster analysis in the form of simple *If then else* statements.
37. We can use wither of the results to discuss the result of cluster analysis and to generate meaningful insights.
38. We will discuss and implement the Tree and Node Rule analysis when we work on Decision Trees.
39. Close the Results window of *Ward Clustering* node.

Testing the stability of the clusters using a sample of the data set

40. To validate the stability of cluster we perform the same cluster analysis on a sample of data. [***Note: If the number of clusters before sampling and after sampling are approximately same (+/- 2 clusters) then clusters are stable. However, we also must ensure that the clusters before and after sampling are representing same type of information.***]
41. Click on the "*Sample*" tab of your Assignment 4 diagram. Click and drag "Sample" node onto the diagram. (***Note: The "Sample" node selects a subset of data from the data set.)***

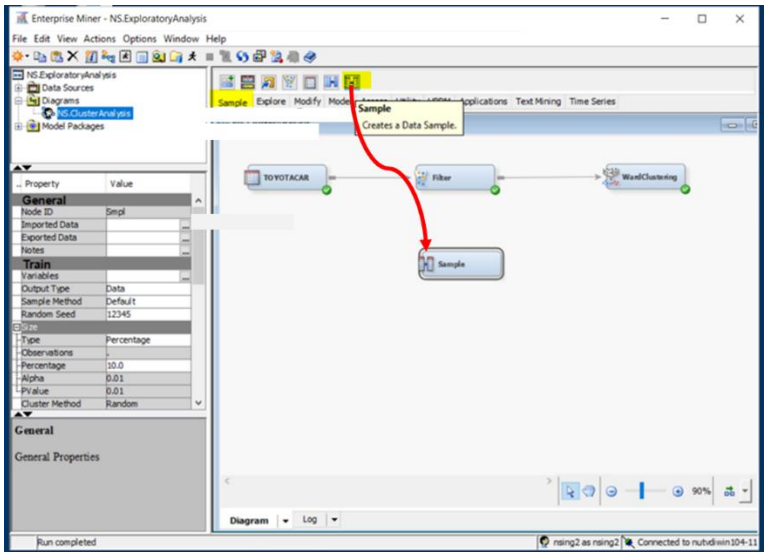


Figure 15

42. Rename “Sample” node as *Sample (90%)*. Click on the *Sample (90%)* node. In the properties window, change the "Percentage" property to 90 and press Enter (***Note: This will include only 90% of the data for analysis***).

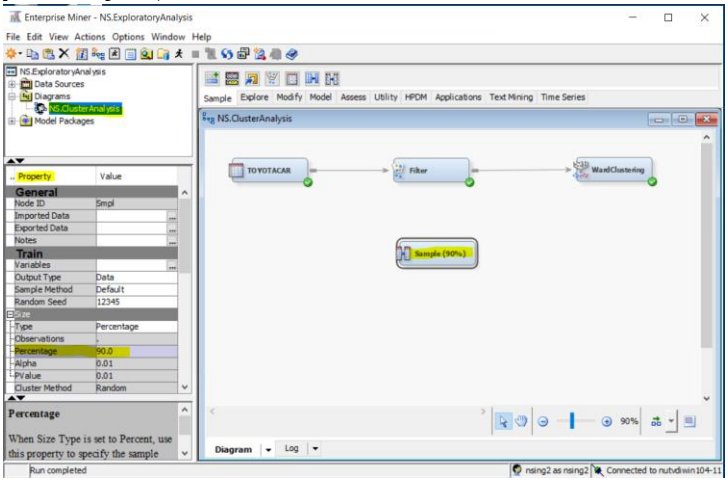


Figure 16

43. Now, from the "Explore" tab, click and drag a new Cluster node. Rename this Cluster Node as *90%Ward Clustering*.
44. In the properties window of *90%Ward Clustering node*, make sure the value of "Internal Standardization" is set to "Standardization (*Note: This ensures that all interval variables are standardized. Standardizing interval variables is the second strategy to develop a stable cluster*) " and the Clustering Method is set to “Ward”.
45. Create a second process flow in your diagram by connecting *Filter node* to *Sample (90%) node* followed by *90%Ward Clustering node* as shown in below screenshot.

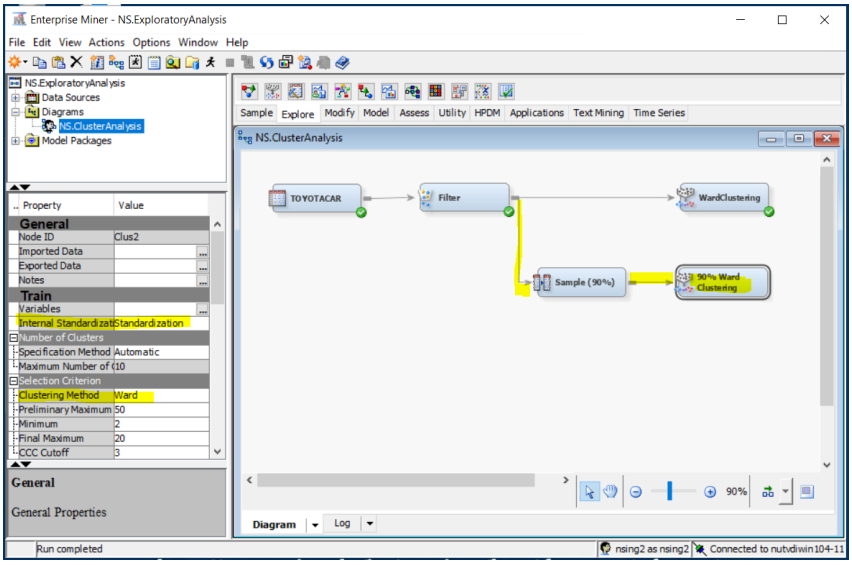


Figure 17

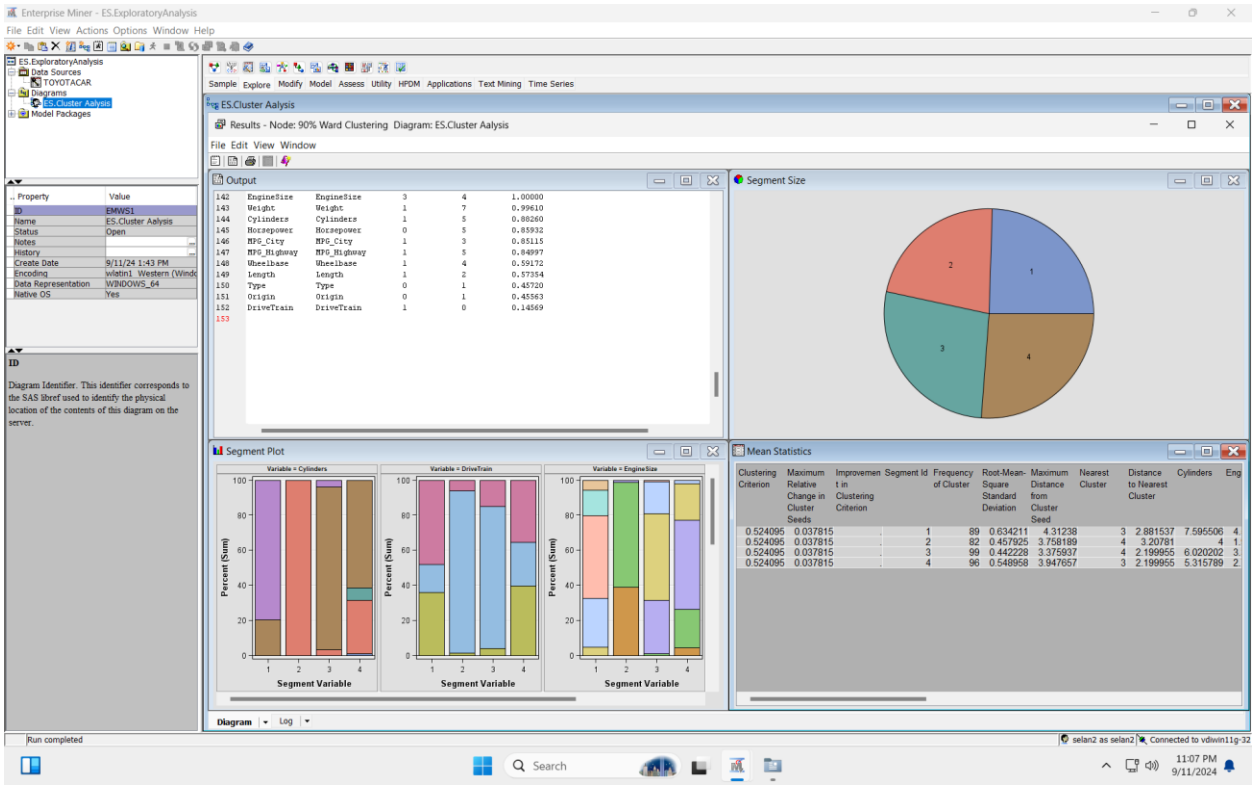
46. Right click on *90%Ward Clustering* and click on *Run t*o execute the cluster analysis on 90% Sample of ToyotaCar data.
47. Open the Results window of *90%Ward Clustering* node.

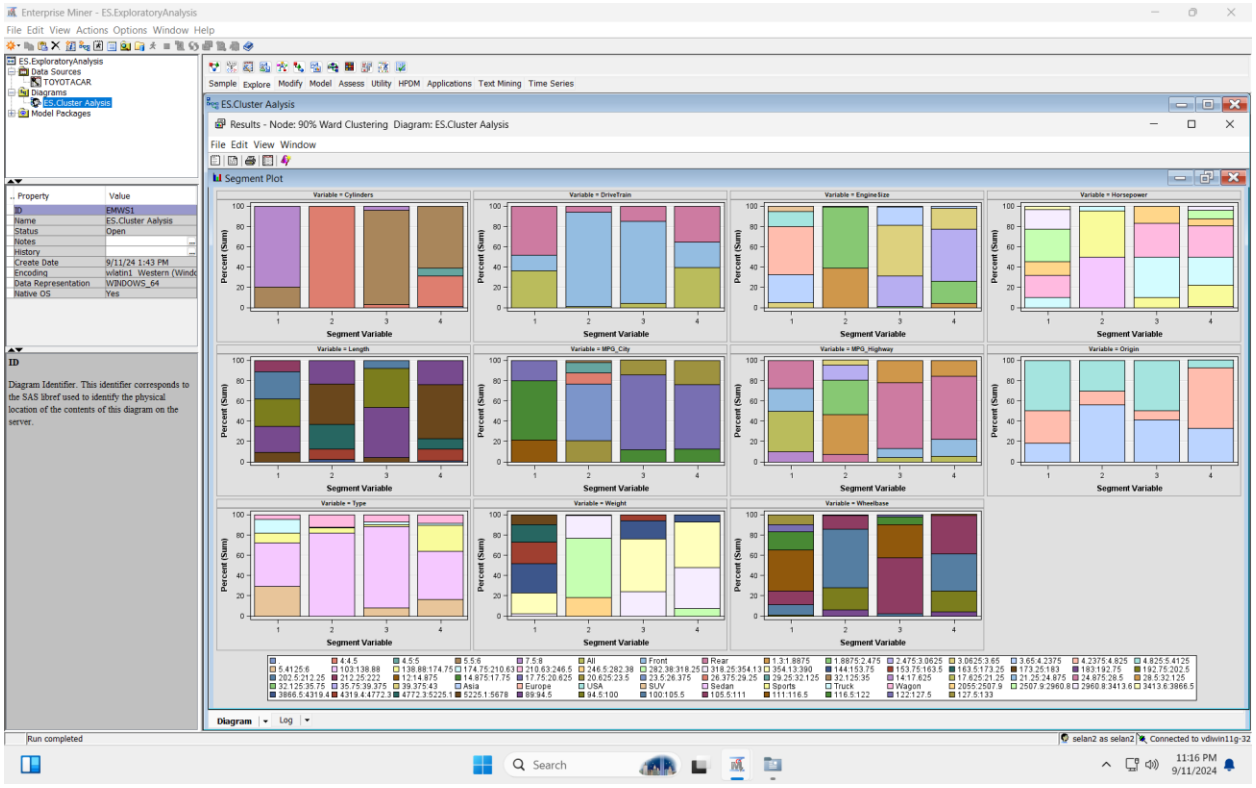
Question 9

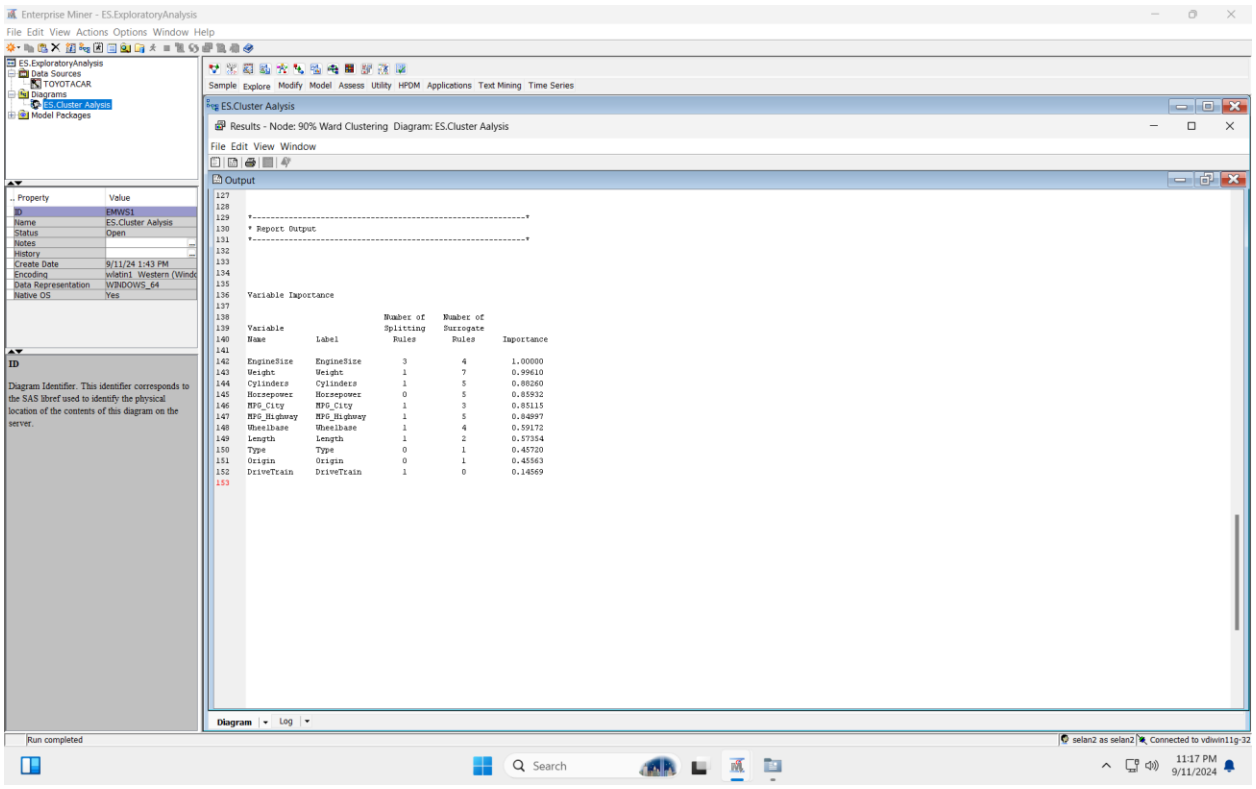
Take the screen shot of Results Window of *90%Ward Clustering* node ensuring you capture:

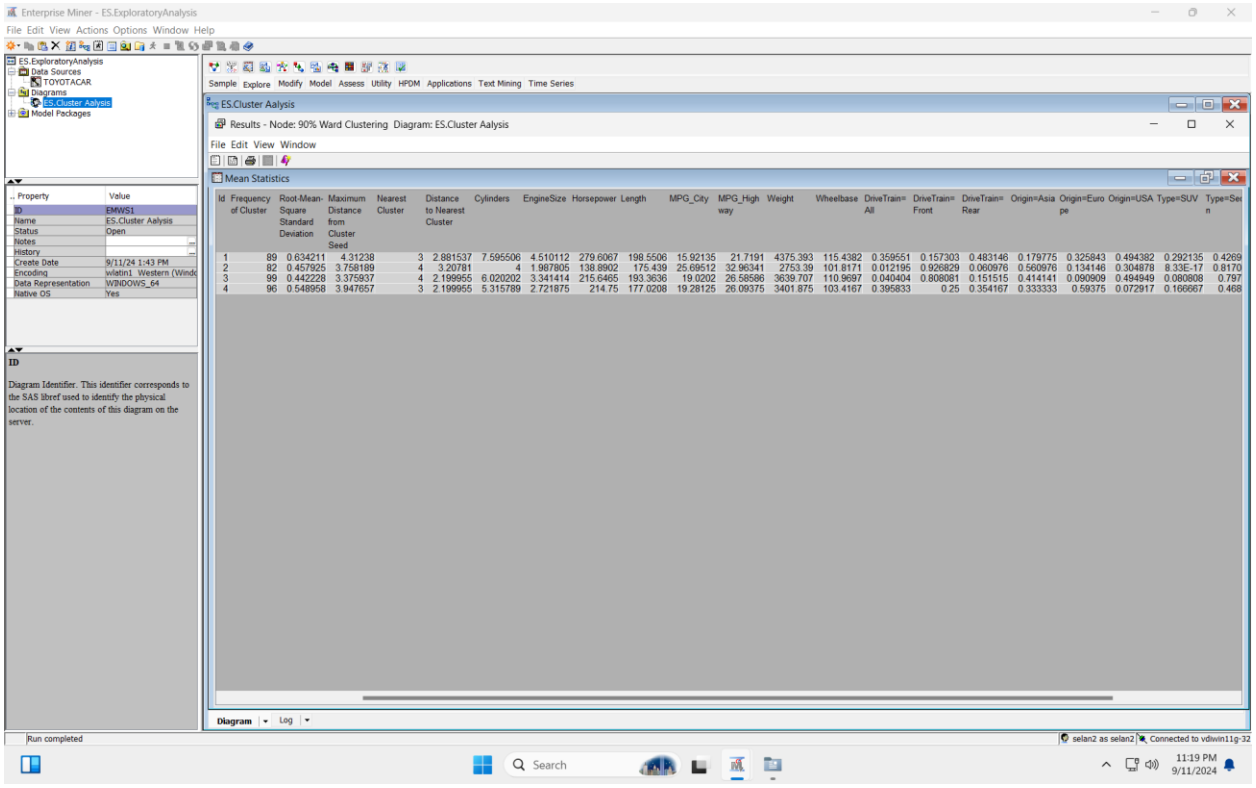
- **Segment Plot Window**
- **Output Window**
- **Mean Statistics Window**
- **Segment Size Window**

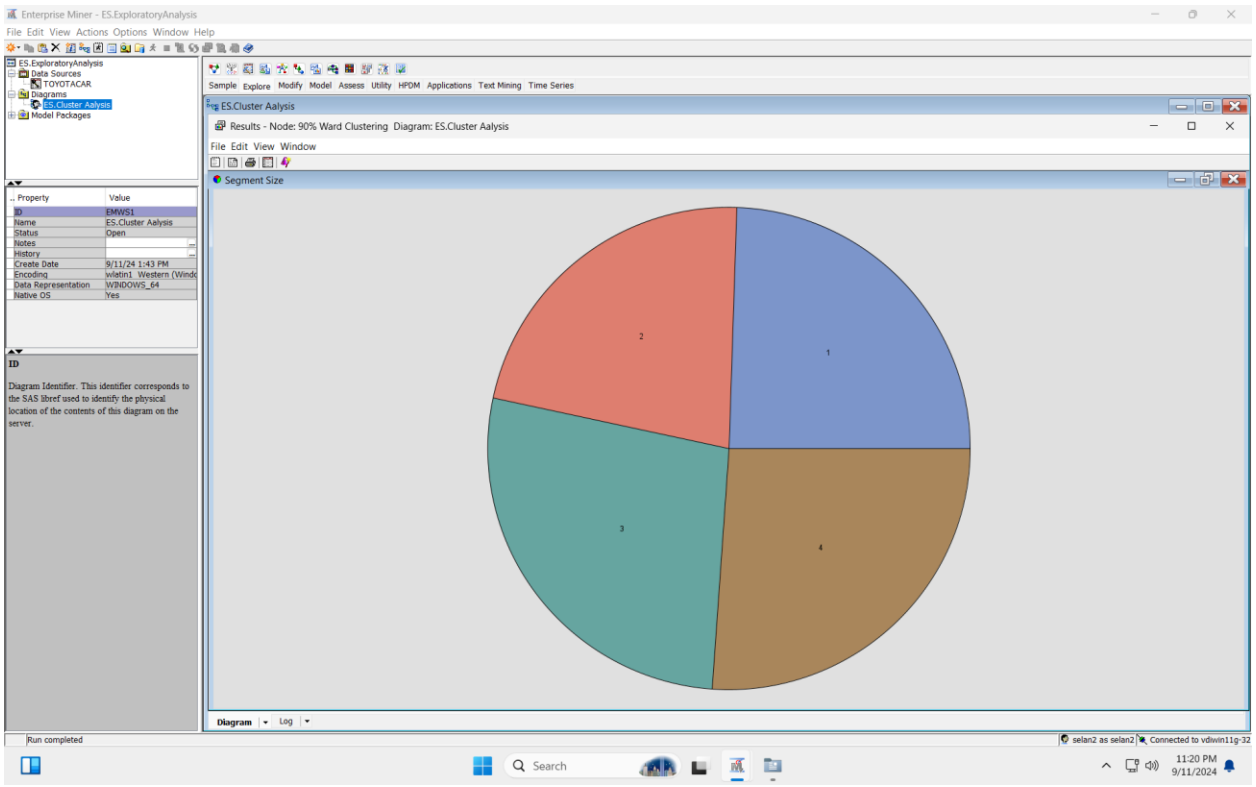
along with your NetID and Machine number as shown in below screenshot. Insert the screenshot of Results window.


Enterprise Miner - ES.ExploratoryAnalysis
ES.Cluster Aalysis
Results - Node: 90% Ward Clustering Diagram: ES.Cluster Aalysis
Output
Segment Size
Segment Plot
Mean Statistics
Segment Variable
Percent (Sum)
Run completed


Enterprise Miner - ES.ExploratoryAnalysis
ES.Cluster Aalysis
Results - Node: 90% Ward Clustering Diagram: ES.Cluster Aalysis
Segment Plot
Segment Variable
Percent (Sum)
Run completed


Enterprise Miner - ES.ExploratoryAnalysis
File Edit View Actions Options Window Help
ES.ExploratoryAnalysis
Data Sources
TOYOTACAR
Diagrams
ES.Cluster Aalysis
Model Packages
Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series
ES.Cluster Aalysis
Results - Node: 90% Ward Clustering Diagram: ES.Cluster Aalysis
File Edit View Window
Output
* Report Output
Variable Importance
Variable Name Label Number of Splitting Rules Number of Surrogate Rules Importance
EngineSize EngineSize 3 4 1.00000
Weight Weight 1 7 0.99610
Cylinders Cylinders 1 5 0.88260
Horsepower Horsepower 0 5 0.85932
MPG_City MPG_City 1 3 0.85115
MPG_Highway MPG_Highway 1 5 0.84997
Wheelbase Wheelbase 1 4 0.59172
Length Length 1 2 0.57354
Type Type 0 1 0.45720
Origin Origin 0 1 0.45563
DriveTrain DriveTrain 1 0 0.14569
Property Value
ID EMWS1
Name ES.Cluster Aalysis
Status Open
Notes
History
Create Date 9/11/24 1:43 PM
Encoding wlatin1 Western (Windc
Data Representation WINDOWS_64
Native OS Yes
ID
Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.
Diagram Log
Run completed
selan2 as selan2 Connected to vdiwin11g-32
Search
11:17 PM
9/11/2024


Enterprise Miner - ES.ExploratoryAnalysis
File Edit View Actions Options Window Help
ES.ExploratoryAnalysis
Data Sources
TOYOTACAR
Diagrams
ES.Cluster Aalysis
Model Packages
Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series
ES.Cluster Aalysis
Results - Node: 90% Ward Clustering Diagram: ES.Cluster Aalysis
File Edit View Window
Mean Statistics
Id Frequency of Cluster Root-Mean-Square Standard Deviation Maximum Distance from Cluster Seed Nearest Cluster Distance to Nearest Cluster Cylinders EngineSize Horsepower Length MPG_City MPG_Highway Weight Wheelbase DriveTrain=All DriveTrain=Front DriveTrain=Rear Origin=Asia Origin=Europe Origin=USA Type=SUV Type=Sedan
1 89 0.634211 4.31238 3 2.881537 7.595506 4.510112 279.6067 198.5506 15.92135 21.7191 4375.393 115.4382 0.359551 0.157303 0.483146 0.179775 0.325843 0.494382 0.292135 0.4269
2 82 0.457925 3.758189 4 3.20781 4 1.987805 138.8902 175.439 25.69512 32.96341 2753.39 101.8171 0.012195 0.926829 0.060976 0.560976 0.134146 0.304878 8.33E-17 0.8170
3 99 0.442228 3.375837 4 2.199955 6.020202 3.341414 215.6465 193.3636 19.0202 26.58586 3639.707 110.9697 0.040404 0.808081 0.151515 0.414141 0.090909 0.494949 0.080808 0.797
4 96 0.548958 3.947657 3 2.199955 5.315789 2.721875 214.75 177.0208 19.28125 26.09375 3401.875 103.4167 0.395833 0.25 0.354167 0.333333 0.59375 0.072917 0.166667 0.468
Property Value
ID EMWS1
Name ES.Cluster Aalysis
Status Open
Notes
History
Create Date 9/11/24 1:43 PM
Encoding wlatin1 Western (Windc
Data Representation WINDOWS_64
Native OS Yes
ID
Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.
Diagram Log
Run completed
selan2 as selan2 Connected to vdiwin11g-32
Search
11:19 PM
9/11/2024



48. Look at the "*Segment Size*" window and answer Question 10.

Question 10

Question 10a

How many clusters are represented in your "Segment Size" Window of *90%Ward Clustering* node?	**4**
Which Cluster has the maximum number of records? Also, provide the number of records for this cluster.	**Segment id : 3 Number of records : 99**
Which Cluster has the minimum number of records? Also, provide the number of records for this cluster.	**Segment id : 2 Number of records : 82**

Question 10b

Do you think the clusters created for Toyotacar using the Ward Clustering are stable? Explain in 150-200 words. (Hint: To answer this question, refer to the ***Notes*** in step 40 of this document. *For example, compare the interpretation of Figure 11 in step 24 with the mean statistics of clusters obtained from the 90% Ward Clustering node in Step* 47. Also, refer to "**Validating Clusters**" sections of chapter 15 (Shmueli et al 2017, pp. 373).

The data analysis of ward analysis and 90% of ward analysis shows that the clusters obtained with Ward's Clustering are relatively stable for Toyotacar. Yes, the clusters created for Toyota cars are stable according to me. I validated the stability of the clusters by performing the same cluster analysis on a data sample. The number of clusters before and after sampling remained the

same. It seems that before and after oversampling, on average, four clusters (4) are found, which indicates a good match of outcomes. In addition, Pseudo F and pseudo-t-squared statistics show that the clusters are highly separated, suggesting good stability. The root-mean-square standard deviation does not change for both datasets, maintaining the same variance within individual clusters.

When comparing the mean statistics, the frequency of clusters is close to 90% of the original value, and all variable values (Length, Weight, Drive Train, MPG_City, etc.) are similar to the original values. However, The segment size pie chart shows a small change in percentages, but it's minor and can be ignored.

Furthermore, the importance of key variables such as Engine Size, Weight, MPG city, and MPG highway remains high across both datasets. This stability indicates that the clusters are not only well-defined but also consistent before and after additional sampling. Since the number of clusters did not change significantly (there were +/- 2 clusters along with previous runs), the clustering appears valid. Therefore, the clusters seem stable and effectively reflect the data based on the applied sampling and validation methods.

49. Minimize the "*Segment Size*" window and close the *Result* window of *90%Ward Clustering* node.

Cluster Analysis using Centroid and Average Distance Metric

50. From the “Explore” tab of your diagram, drag a cluster node to your diagram and name it as *Centroid Clustering.* In the properties window of *Centroid Clustering* node set *Clustering Method* property to *Centroid.* Also, ensure the *Internal Standardization* property of *Centroid Clustering* node is set to standardization.
51. Connect *Filter* node to *Centroid Clustering* node as shown below.

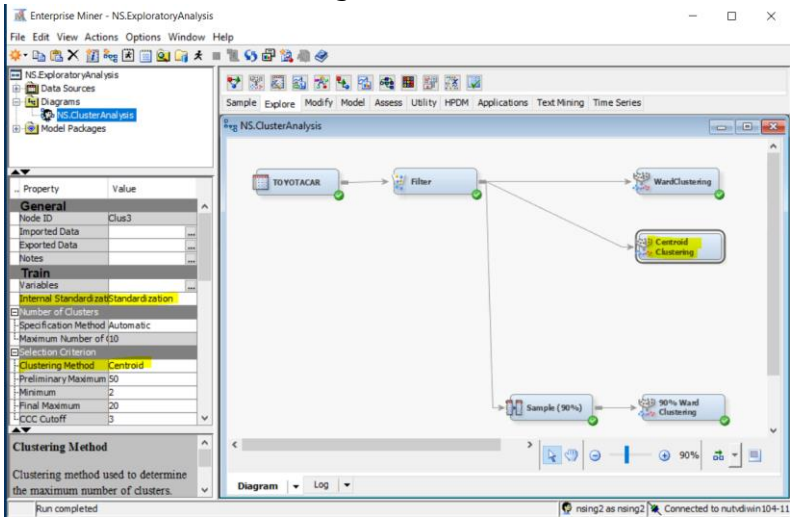


Figure 18

52. From the "Explore" tab of your diagram, drag a cluster node to your diagram and name it as *Average Clustering.* In the properties window of *Average Clustering* node set *Clustering Method* property to *Average.* Also, ensure the *Internal Standardization* property of *Average Clustering* node is set to standardization. Connect *Filter* node to *Average Clustering* node as shown below.

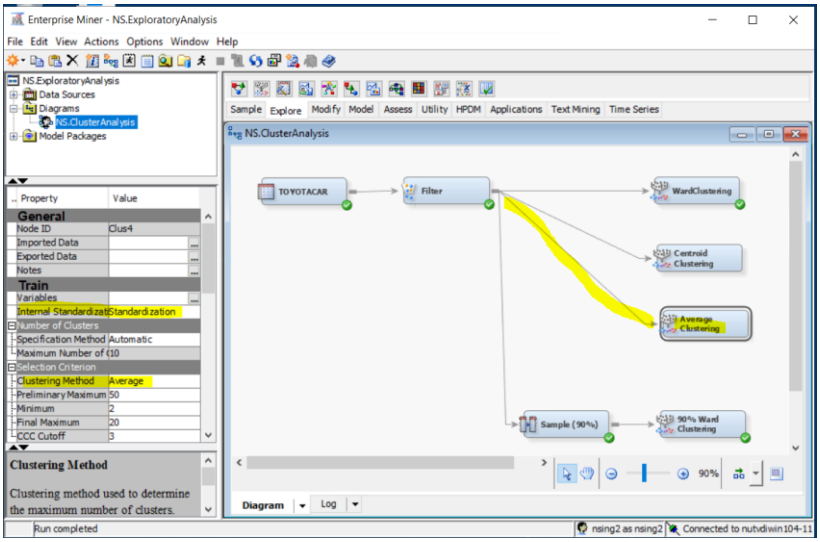


Figure 19

53. Right click on *Centroid Clustering* node (Figure 19) and click on Copy. Now right click anywhere on the diagram and click paste. Rename the pasted node as *90% Centroid Clustering*. In the properties window of 90% *Centroid Clustering* node ensure *Clustering Method* property is set to *Centroid.* Move it under *90% Ward Clustering* node and connect it to *Sample (90%)* node (Figure 20).
54. Right click on *Average Clustering* node (Figure 19) and click on Copy. Now right click anywhere on the diagram and click paste. Rename the pasted node as *90% Average Clustering*. In the properties window of 90% *Average Clustering* node ensure *Clustering Method* property is set to *Average.* Move it under *90% Centroid Clustering* and connect it to *Sample (90%)* node (Figure 20).
55. Right click on *Centroid Clustering* node and click on *Run* to execute it (Figure 20).
56. Right click on *Average Clustering* node and click on *Run* to execute it (Figure 20).
57. Right click on *90% Centroid Clustering* node and click on *Run* to execute it (Figure 20).
58. Right click on *90% Average Clustering* node and click on *Run* to execute it (Figure 20).
59. Your Cluster Analysis diagram will look like as shown below:

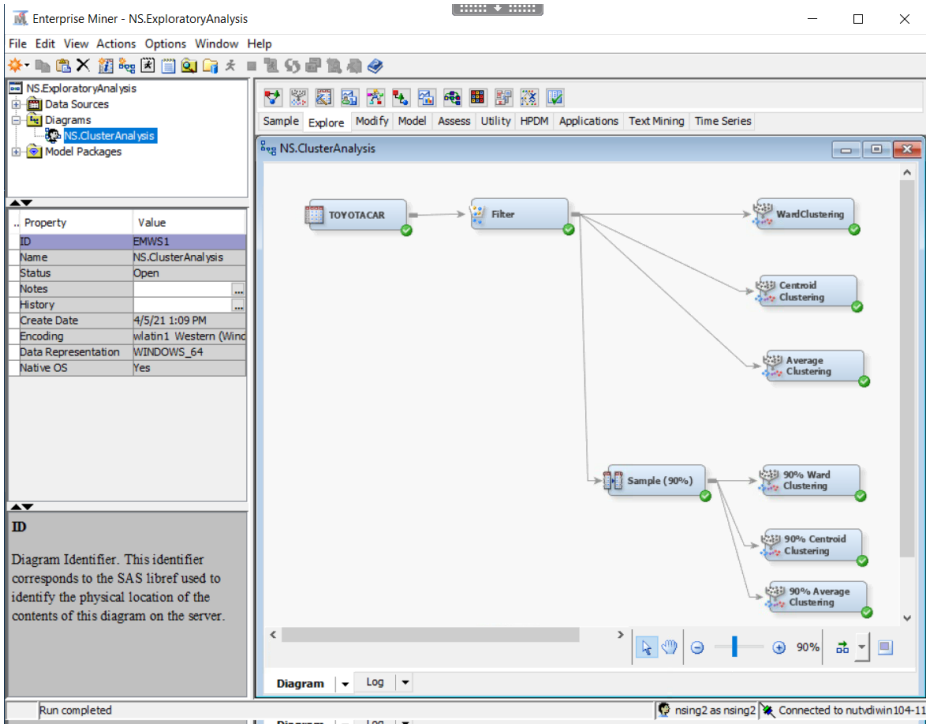


Figure 20

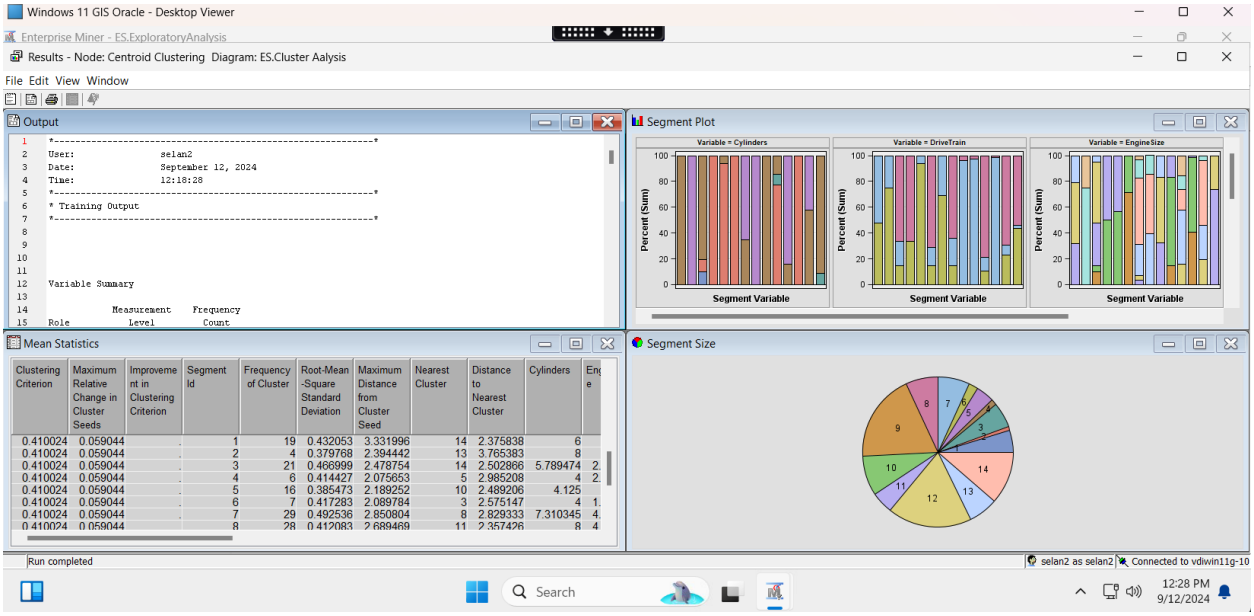
Question 11

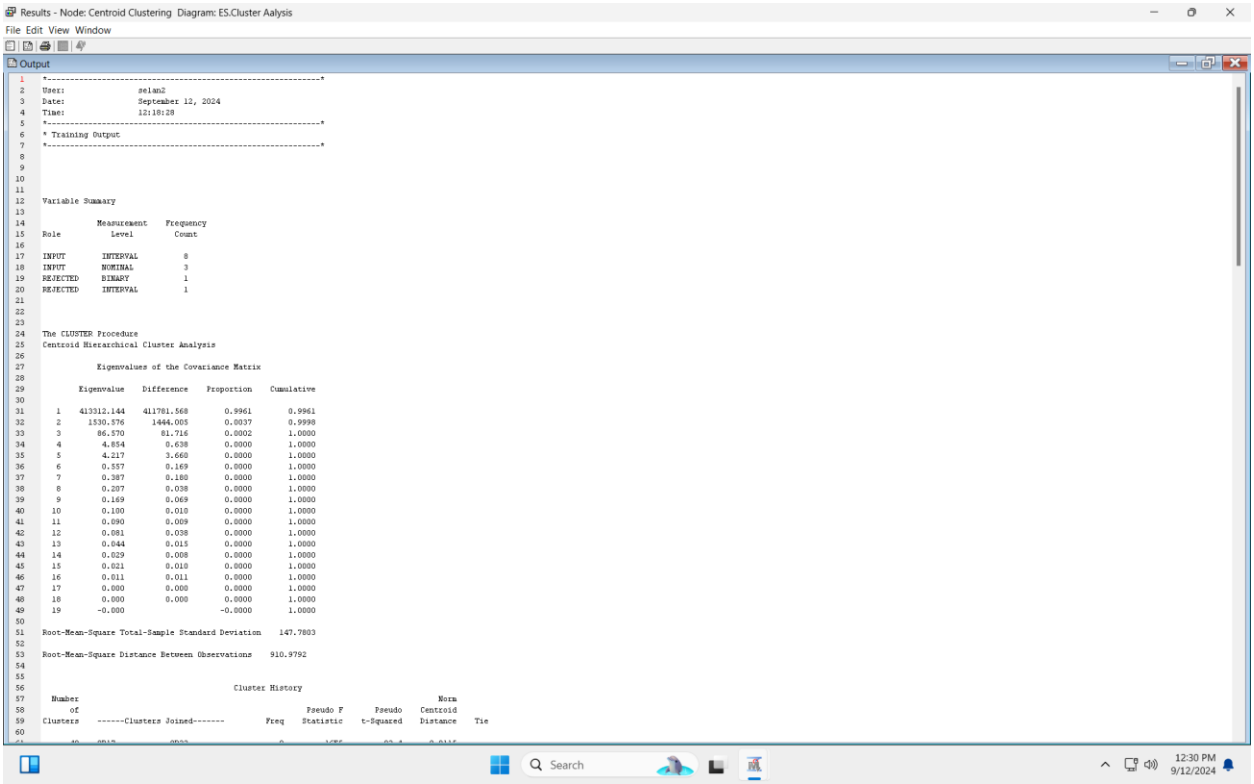
Question 11a

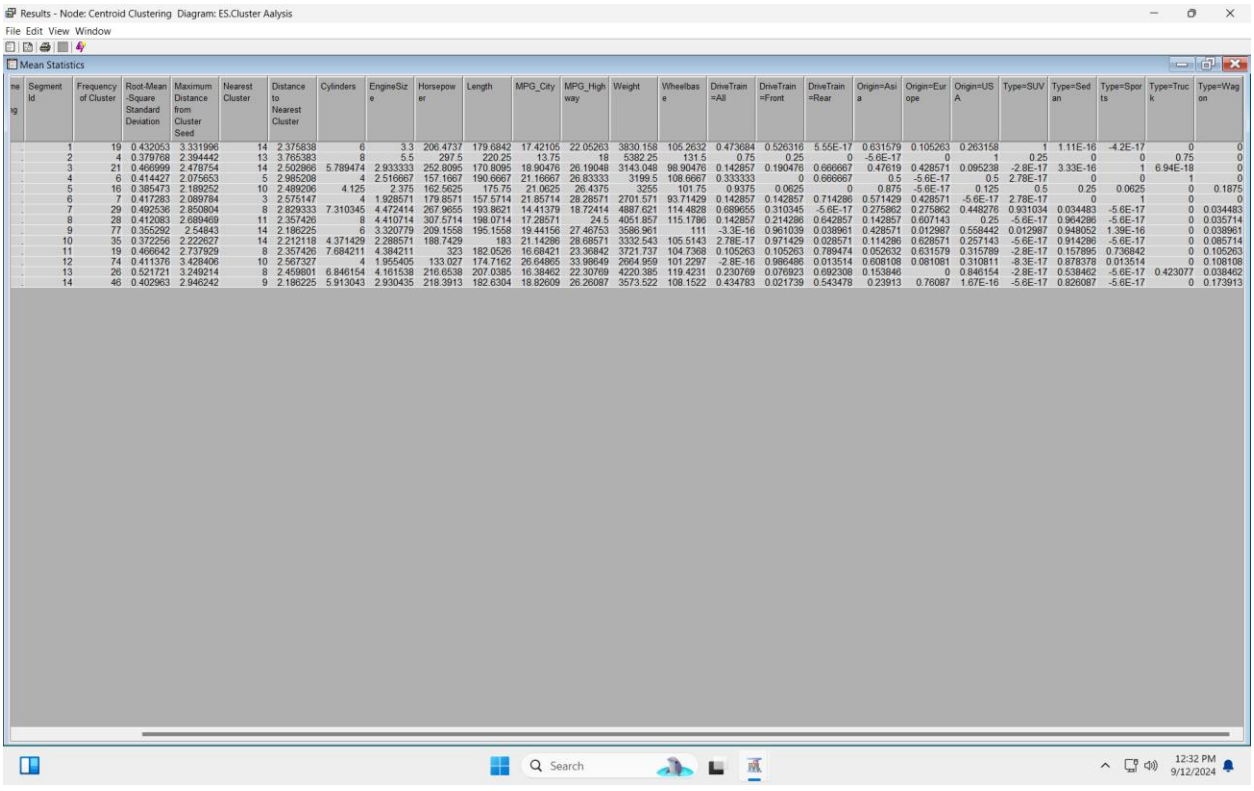
Centroid Clustering Node	90% Centroid Clustering Node
How many clusters are there in the Segment Size window of *Centroid clustering* node? **Answer: 14 clusters**	**How many clusters are there in the Segment Size window of *90% Centroid clustering* node?** **Answer: 18 clusters**
Is the number of clusters obtained from *Centroid clustering* node greater than 6? **Answer: Yes**	**Is the number of clusters obtained from *90% Centroid clustering* node greater than 6?** **Answer : Yes**
Provide the screenshot of Results window for *Centroid Clustering* node ensuring you capture: • **Segment Plot Window** • **Output Window** • **Mean Statistics Window** • **Segment Size Window** **along with your NetID and Machine number as shown in below screenshot.**	**Provide the screenshot(s) of Results window for *90% Centroid clustering* ensuring you capture:** • **Segment Plot Window** • **Output Window** • **Mean Statistics Window** • **Segment Size Window** **along with your NetID and Machine number as shown in below screenshot.**

Insert the screenshot of Results window here.	**Insert the screenshot of Results window here.**

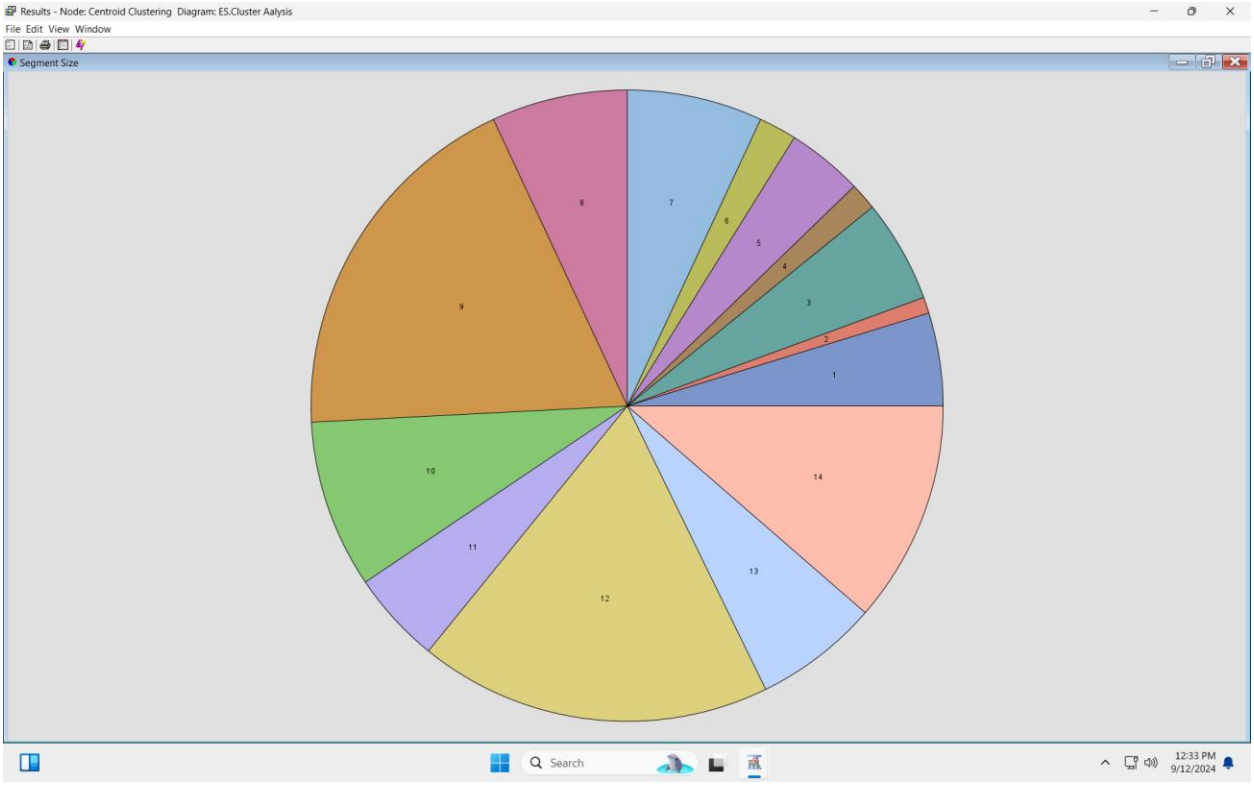
Results window for ***Centroid Clustering*** node :

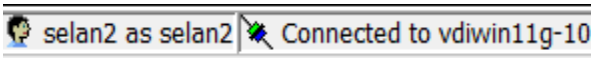




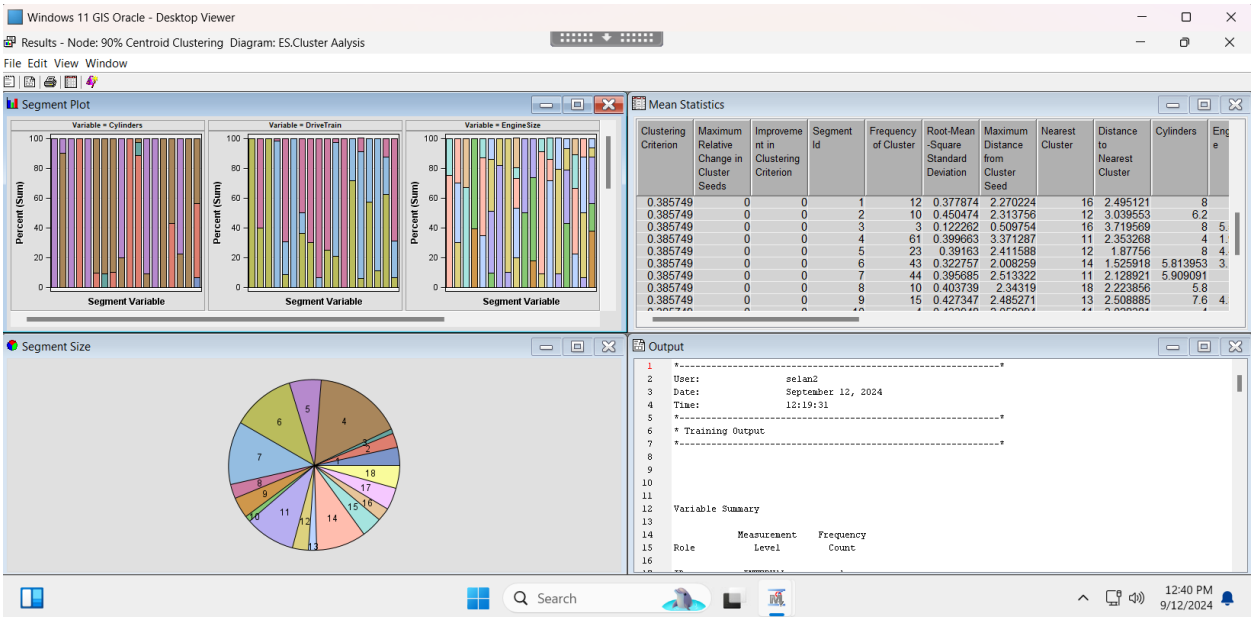


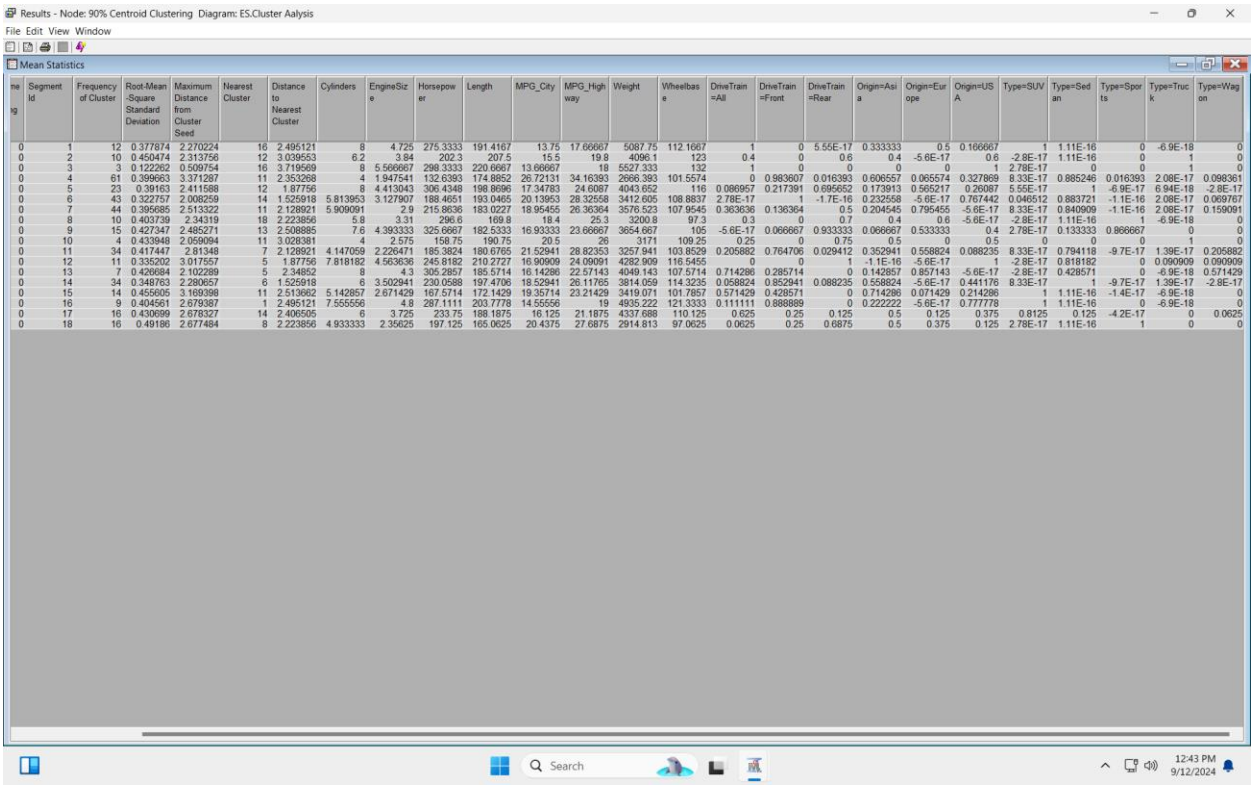
	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Cylinders	EngineSize	Horsepower	Length	MPG_City	MPG_Highway	Weight	Wheelbase	DriveTrain=All	DriveTrain=Front	DriveTrain=Rear	Origin=Asia	Origin=Europe	Origin=USA	Type=SUV	Type=Sedan	Type=Sports	Type=Truck	Type=Wagon
	1	19	0.432053	3.331996	14	2.375838	6	3.3	206.4737	179.6842	17.42105	22.05263	3830.158	105.2632	0.473684	0.526316	5.55E-17	0.631579	0.105263	0.263158	1	1.11E-16	-4.2E-17	0	0
	2	4	0.379768	2.394442	13	3.765383	8	5.5	297.5	220.25	13.75	18	5382.25	131.5	0.75	0.25	0	-5.6E-17	0	1	0.25	0	0	0.75	0
	3	21	0.466999	2.478754	14	2.502866	5.789474	2.933333	252.8095	170.8095	18.90476	26.19048	3143.048	98.90476	0.142857	0.190476	0.666667	0.47619	0.428571	0.095238	-2.8E-17	3.33E-16	1	6.94E-18	0
	4	6	0.414427	2.075653	5	2.985208	4	2.516667	157.1667	190.6667	21.16667	26.83333	3199.5	108.6667	0.333333	0	0.666667	0.5	-5.6E-17	0.5	2.78E-17	0	0	1	0
	5	16	0.385473	2.189252	10	2.489206	4.125	2.375	162.5625	175.75	21.0625	26.4375	3255	101.75	0.9375	0.0625	0	0.875	-5.6E-17	0.125	0.5	0.25	0.0625	0	0.1875
	6	7	0.417283	2.089784	3	2.575147	4	1.928571	179.8571	157.5714	21.85714	28.28571	2701.571	93.71429	0.142857	0.142857	0.714286	0.571429	0.428571	-5.6E-17	2.78E-17	0	1	0	0
	7	29	0.492536	2.850804	8	2.829333	7.310345	4.472414	267.9655	193.8621	14.41379	18.72414	4887.621	114.4828	0.689655	0.310345	-5.6E-17	0.275862	0.275862	0.448276	0.931034	0.034483	-5.6E-17	0	0.034483
	8	28	0.412083	2.689469	11	2.357426	8	4.410714	307.5714	198.0714	17.28571	24.5	4051.857	115.1786	0.142857	0.214286	0.642857	0.142857	0.607143	0.25	-5.6E-17	0.964286	-5.6E-17	0	0.035714
	9	77	0.355292	2.54843	14	2.186225	6	3.320779	209.1558	195.1558	19.44156	27.46753	3586.961	111	-3.3E-16	0.961039	0.038961	0.428571	0.012987	0.558442	0.012987	0.948052	1.39E-16	0	0.038961
	10	35	0.372256	2.222627	14	2.212118	4.371429	2.288571	188.7429	183	21.14286	28.68571	3332.543	105.5143	2.78E-17	0.971429	0.028571	0.114286	0.628571	0.257143	-5.6E-17	0.914286	-5.6E-17	0	0.085714
	11	19	0.466642	2.737929	8	2.357426	7.684211	4.384211	323	182.0526	16.68421	23.36842	3721.737	104.7368	0.105263	0.105263	0.789474	0.052632	0.631579	0.315789	-2.8E-17	0.157895	0.736842	0	0.105263
	12	74	0.411376	3.428406	10	2.567327	4	1.955405	133.027	174.7162	26.64865	33.98649	2664.959	101.2297	-2.8E-16	0.986486	0.013514	0.608108	0.081081	0.310811	-8.3E-17	0.878378	0.013514	0	0.108108
	13	26	0.521721	3.249214	8	2.459801	6.846154	4.161538	216.6538	207.0385	16.38462	22.30769	4220.385	119.4231	0.230769	0.076923	0.692308	0.153846	0	0.846154	-2.8E-17	0.538462	-5.6E-17	0.423077	0.038462
	14	46	0.402963	2.946242	9	2.186225	5.913043	2.930435	218.3913	182.6304	18.82609	26.26087	3573.522	108.1522	0.434783	0.021739	0.543478	0.23913	0.76087	1.67E-16	-5.6E-17	0.826087	-5.6E-17	0	0.173913



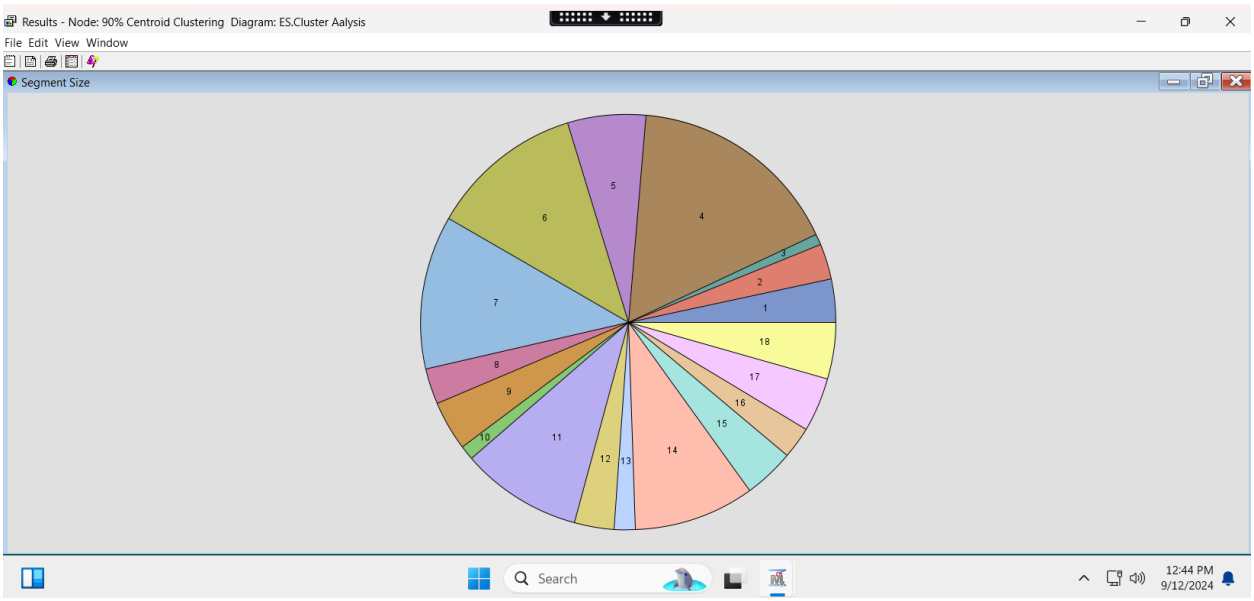


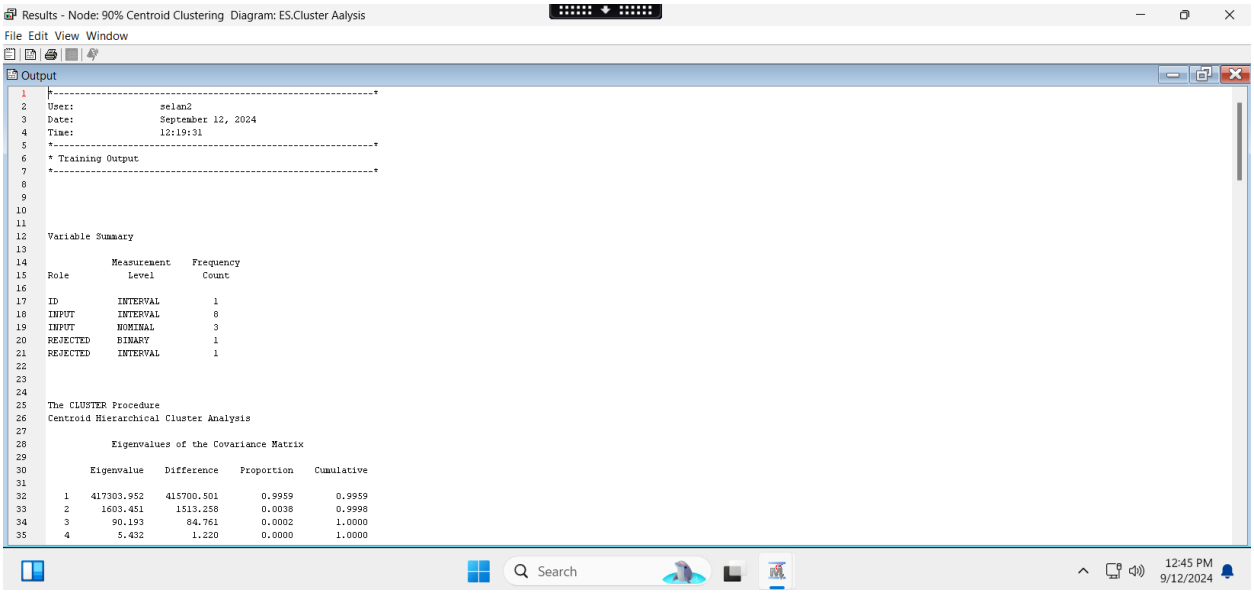
Results window for *90% Centroid clustering:*



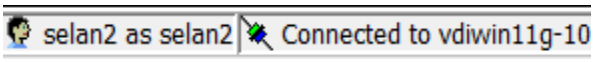


ne ig	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Cylinders	EngineSize	Horsepower	Length	MPG_City	MPG_Highway	Weight	Wheelbase	DriveTrain=All	DriveTrain=Front	DriveTrain=Rear	Origin=Asia	Origin=Europe	Origin=USA	Type=SUV	Type=Sedan	Type=Sports	Type=Truck	Type=Wagon
0	1	12	0.377874	2.270224	16	2.495121	8	4.725	275.3333	191.4167	13.75	17.66667	5087.75	112.1667	1	0	5.55E-17	0.333333	0.5	0.166667	1	1.11E-16	0	-6.9E-18	0
0	2	10	0.450474	2.313756	12	3.039553	6.2	3.84	202.3	207.5	15.5	19.8	4096.1	123	0.4	0	0.6	0.4	-5.6E-17	0.6	-2.8E-17	1.11E-16	0	1	0
0	3	3	0.122262	0.509754	16	3.719569	8	5.566667	298.3333	220.6667	13.66667	18	5527.333	132	1	0	0	0	0	1	2.78E-17	0	0	1	0
0	4	61	0.399663	3.371287	11	2.353268	4	1.947541	132.6393	174.8852	26.72131	34.16393	2666.393	101.5574	0	0.983607	0.016393	0.606557	0.065574	0.327869	8.33E-17	0.885246	0.016393	2.08E-17	0.098361
0	5	23	0.39163	2.411588	12	1.87756	8	4.413043	306.4348	198.8696	17.34783	24.6087	4043.652	116	0.086957	0.217391	0.695652	0.173913	0.565217	0.26087	5.55E-17	1	-6.9E-17	6.94E-18	-2.8E-17
0	6	43	0.322757	2.008259	14	1.525918	5.813953	3.127907	188.4651	193.0465	20.13953	28.32558	3412.605	108.8837	2.78E-17	1	-1.7E-16	0.232558	-5.6E-17	0.767442	0.046512	0.883721	-1.1E-16	2.08E-17	0.069767
0	7	44	0.395685	2.513322	11	2.128921	5.909091	2.9	215.8636	183.0227	18.95455	26.36364	3576.523	107.9545	0.363636	0.136364	0.5	0.204545	0.795455	-5.6E-17	8.33E-17	0.840909	-1.1E-16	2.08E-17	0.159091
0	8	10	0.403739	2.34319	18	2.223856	5.8	3.31	296.6	169.8	18.4	25.3	3200.8	97.3	0.3	0	0.7	0.4	0.6	-5.6E-17	-2.8E-17	1.11E-16	1	-6.9E-18	0
0	9	15	0.427347	2.485271	13	2.508885	7.6	4.393333	325.6667	182.5333	16.93333	23.66667	3654.667	105	-5.6E-17	0.066667	0.933333	0.066667	0.533333	0.4	2.78E-17	0.133333	0.866667	0	0
0	10	4	0.433948	2.059094	11	3.028381	4	2.575	158.75	190.75	20.5	26	3171	109.25	0.25	0	0.75	0.5	0.5	0	0	0	0	1	0
0	11	34	0.417447	2.81348	7	2.128921	4.147059	2.226471	185.3824	180.6765	21.52941	28.82353	3257.941	103.8529	0.205882	0.764706	0.029412	0.352941	0.558824	0.088235	8.33E-17	0.794118	-9.7E-17	1.39E-17	0.205882
0	12	11	0.335202	3.017557	5	1.87756	7.818182	4.563636	245.8182	210.2727	16.90909	24.09091	4282.909	116.5455	0	0	1	-1.1E-16	-5.6E-17	1	-2.8E-17	0.818182	0	0.090909	0.090909
0	13	7	0.426684	2.102289	5	2.34852	8	4.3	305.2857	185.5714	16.14286	22.57143	4049.143	107.5714	0.714286	0.285714	0	0.142857	0.857143	-5.6E-17	-2.8E-17	0.428571	0	-6.9E-18	0.571429
0	14	34	0.348763	2.280657	6	1.525918	6	3.502941	230.0588	197.4706	18.52941	26.11765	3814.059	114.3235	0.058824	0.852941	0.088235	0.558824	-5.6E-17	0.441176	8.33E-17	1	-9.7E-17	1.39E-17	-2.8E-17
0	15	14	0.455605	3.169398	11	2.513662	5.142857	2.671429	167.5714	172.1429	19.35714	23.21429	3419.071	101.7857	0.571429	0.428571	0	0.714286	0.071429	0.214286	1	1.11E-16	-1.4E-17	-6.9E-18	0
0	16	9	0.404561	2.679387	1	2.495121	7.555556	4.8	287.1111	203.7778	14.55556	19	4935.222	121.3333	0.111111	0.888889	0	0.222222	-5.6E-17	0.777778	1	1.11E-16	0	-6.9E-18	0
0	17	16	0.430699	2.678327	14	2.406505	6	3.725	233.75	188.1875	16.125	21.1875	4337.688	110.125	0.625	0.25	0.125	0.5	0.125	0.375	0.8125	0.125	-4.2E-17	0	0.0625
0	18	16	0.49186	2.677484	8	2.223856	4.933333	2.35625	197.125	165.0625	20.4375	27.6875	2914.813	97.0625	0.0625	0.25	0.6875	0.5	0.375	0.125	2.78E-17	1.11E-16	1	0	0





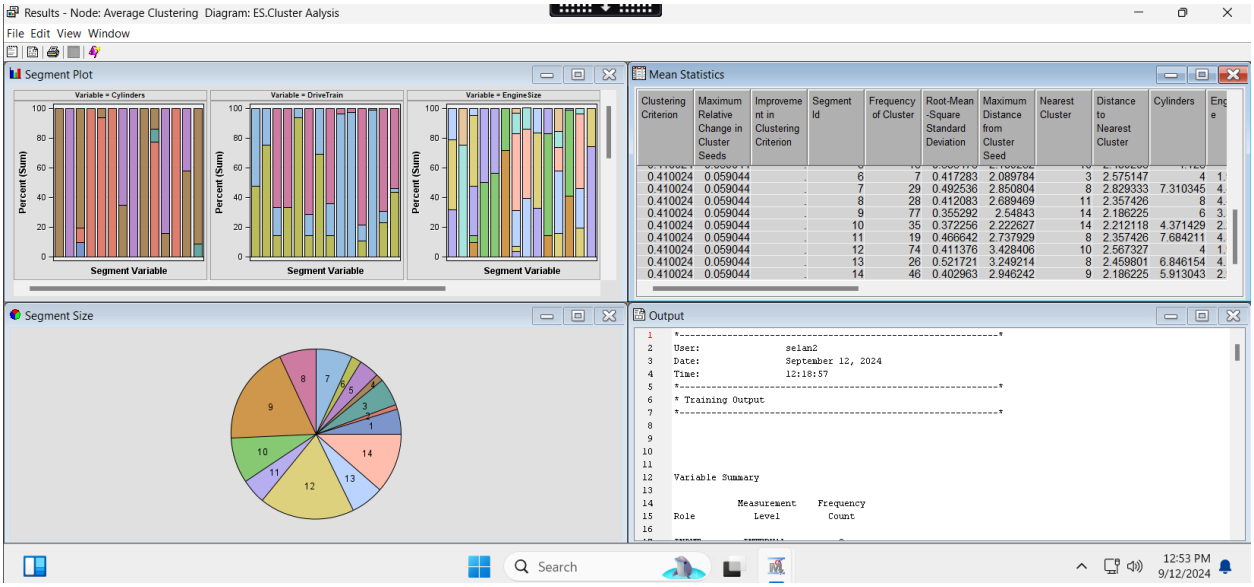

```
+-------------------------------------------------------------+
User:                  selan2
Date:                  September 12, 2024
Time:                  12:19:31
+-------------------------------------------------------------+
* Training Output
+-------------------------------------------------------------+



Variable Summary

                Measurement    Frequency
Role               Level         Count

ID              INTERVAL           1
INPUT           INTERVAL           8
INPUT           NOMINAL            3
REJECTED        BINARY             1
REJECTED        INTERVAL           1


The CLUSTER Procedure
Centroid Hierarchical Cluster Analysis

              Eigenvalues of the Covariance Matrix

       Eigenvalue    Difference    Proportion    Cumulative

    1   417303.952   415700.501        0.9959        0.9959
    2     1603.451     1513.258        0.0038        0.9998
    3       90.193       84.761        0.0002        1.0000
    4        5.432        1.220        0.0000        1.0000
```

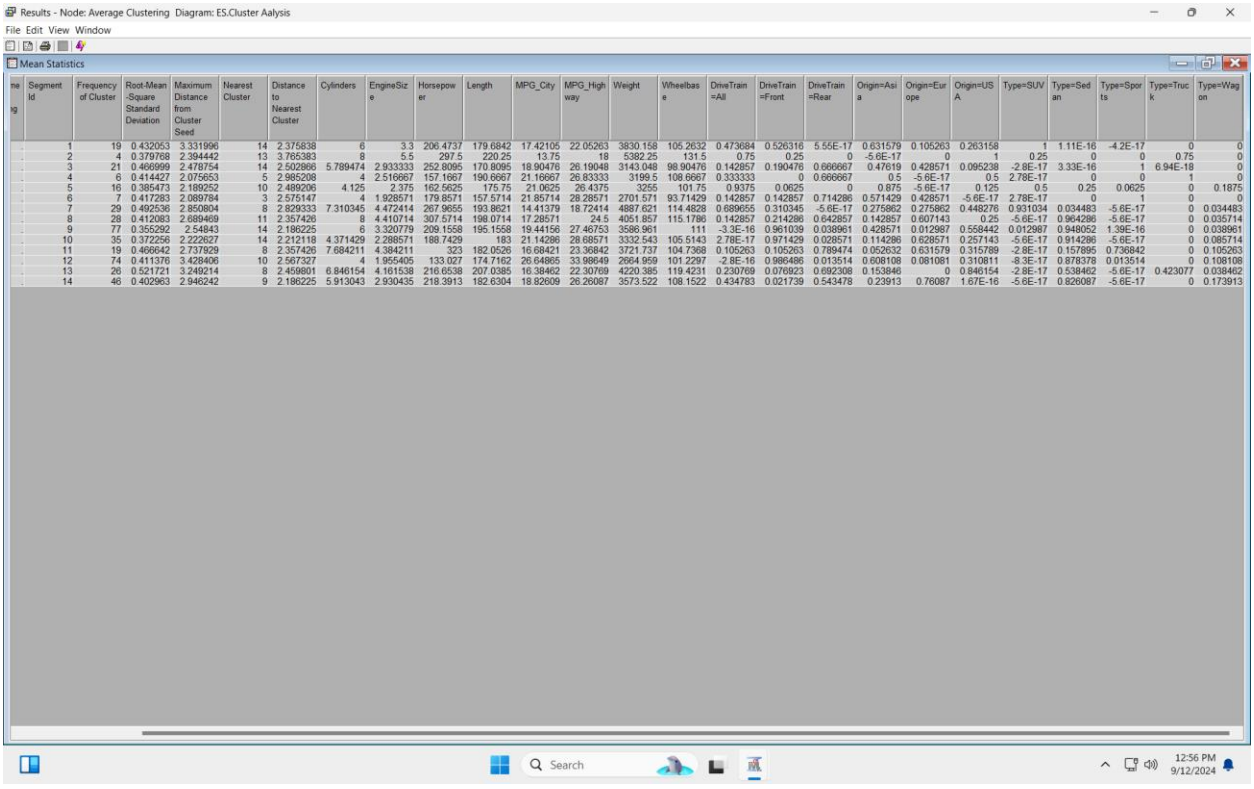


Question 11b

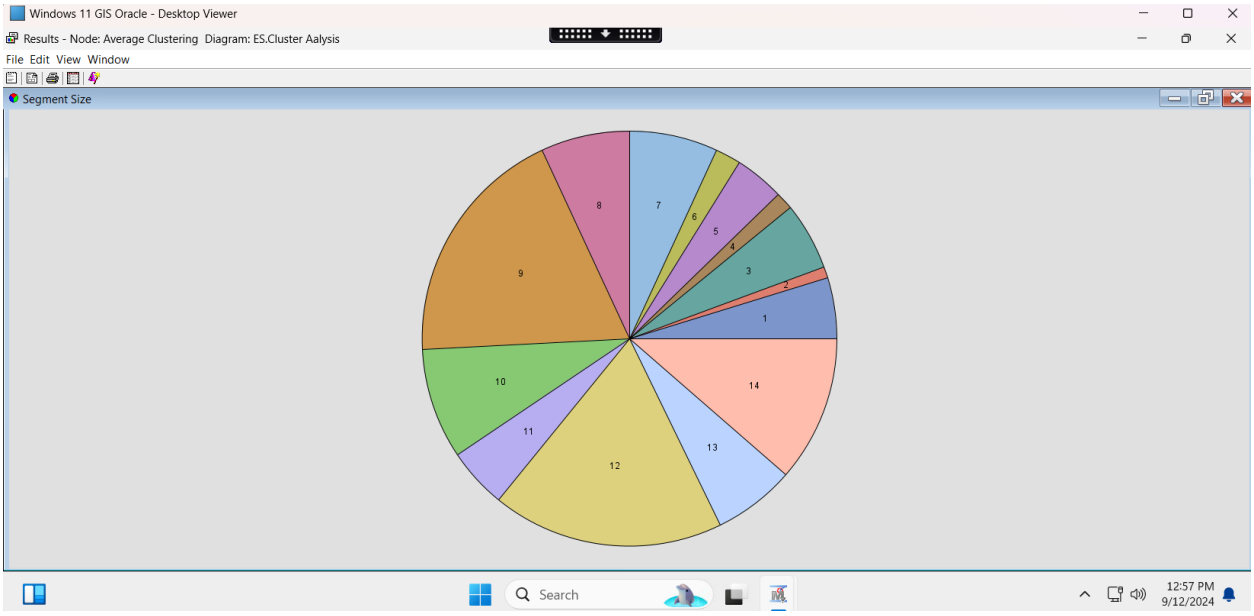
Average Clustering Node	90% Average Clustering Node
How many clusters are there in the Segment Size window of *Average clustering* node? **Answer : 14 clusters**	**How many clusters are there in the Segment Size window of *90% Average clustering* node?** **Answer : 19 clusters**
Is the number of clusters obtained from *Average clustering* node greater than 6? **Answer : Yes**	**Is the number of clusters obtained from *90% Average clustering* node greater than 6?** **Answer : Yes**
Provide the screenshot of Results window for *Average clustering* node ensuring you capture: • **Segment Plot Window** • **Output Window** • **Mean Statistics Window** • **Segment Size Window** **along with your NetID and Machine number as shown in below screenshot.** **Insert the screenshot of Results window here.**	**Provide the screenshot(s) of Results window for *90% Average clustering* ensuring you capture:** • **Segment Plot Window** • **Output Window** • **Mean Statistics Window** • **Segment Size Window** **along with your NetID and Machine number as shown in below screenshot.** **Insert the screenshot of Results window here.**

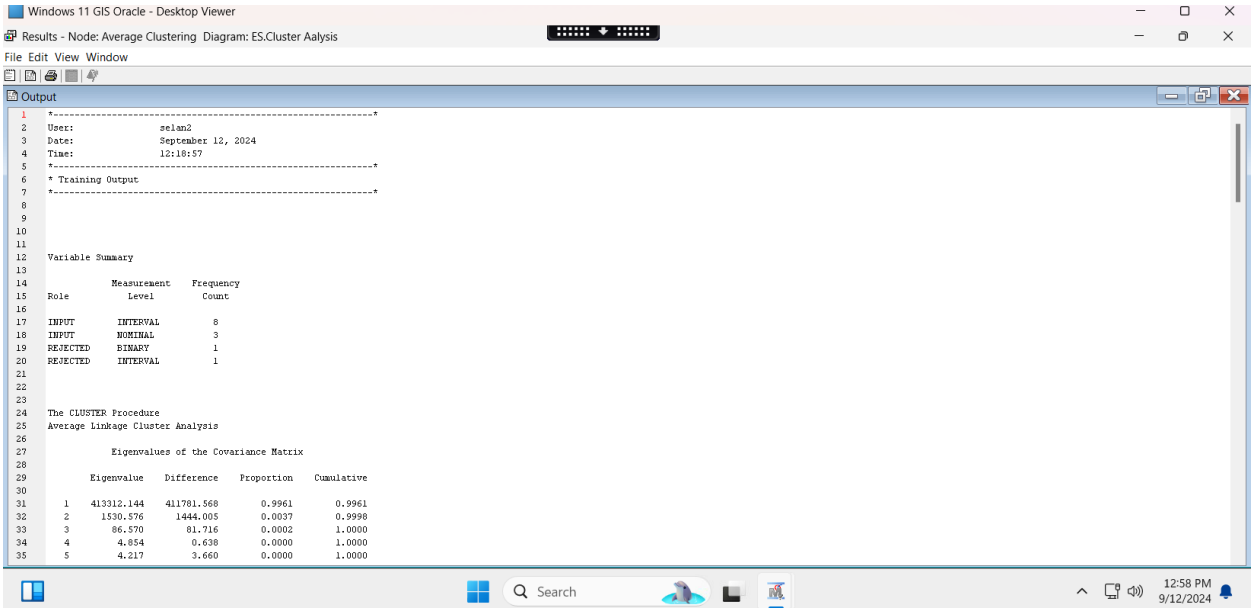
Results window for *Average clustering* node:





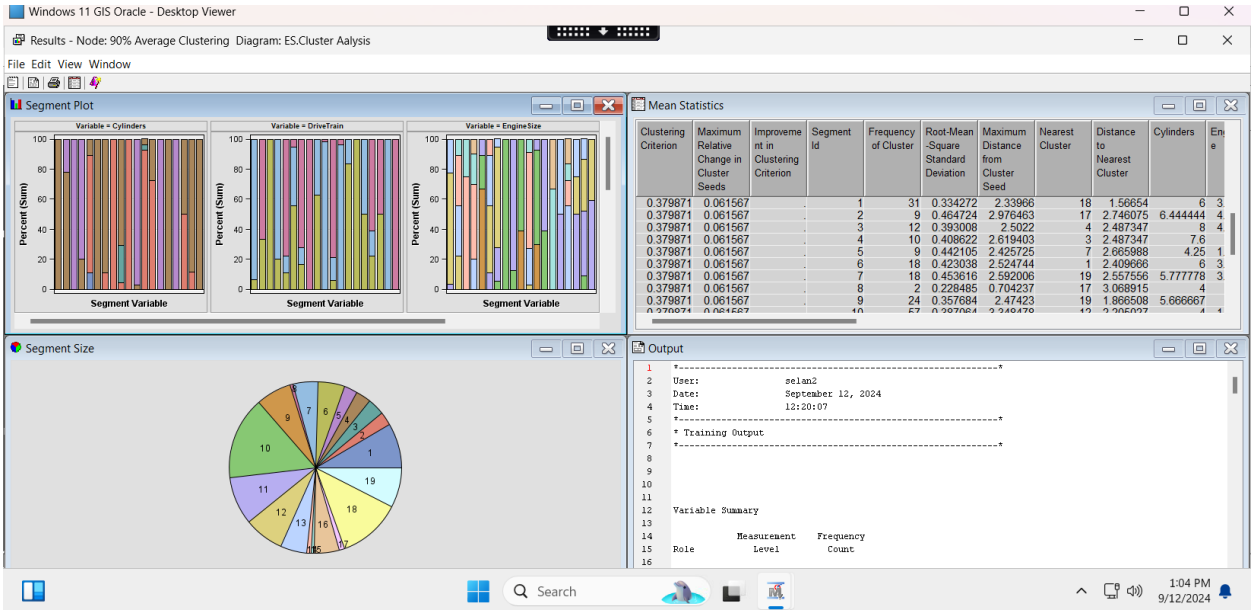
ne ng	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Cylinders	EngineSize	Horsepower	Length	MPG_City	MPG_Highway	Weight	Wheelbase	DriveTrain=All	DriveTrain=Front	DriveTrain=Rear	Origin=Asia	Origin=Europe	Origin=USA	Type=SUV	Type=Sedan	Type=Sports	Type=Truck	Type=Wagon
	1	19	0.432053	3.331996	14	2.375838	6	3.3	206.4737	179.6842	17.42105	22.05263	3830.158	105.2632	0.473684	0.526316	5.55E-17	0.631579	0.105263	0.263158	1	1.11E-16	-4.2E-17	0	0
	2	4	0.379768	2.394442	13	3.765383	8	5.5	297.5	220.25	13.75	18	5382.25	131.5	0.75	0.25	0	-5.6E-17	0	1	0.25	0	0	0.75	0
	3	21	0.466999	2.478754	14	2.502986	5.789474	2.933333	252.8095	170.8095	18.90476	26.19048	3143.048	98.90476	0.142857	0.190476	0.666667	0.47619	0.428571	0.095238	-2.8E-17	3.33E-16	1	6.94E-18	0
	4	6	0.414427	2.075653	5	2.985208	4	2.516667	157.1667	190.6667	21.16667	26.83333	3199.5	108.6667	0.333333	0	0.666667	0.5	-5.6E-17	0.5	2.78E-17	0	0	1	0
	5	16	0.385473	2.189252	10	2.489206	4.125	2.375	162.5625	175.75	21.0625	26.4375	3255	101.75	0.9375	0.0625	0	0.875	-5.6E-17	0.125	0.5	0.25	0.0625	0	0.1875
	6	7	0.417283	2.089784	3	2.575147	4	1.928571	179.8571	157.5714	21.85714	28.28571	2701.571	93.71429	0.142857	0.142857	0.714286	0.571429	0.428571	-5.6E-17	2.78E-17	0	1	0	0
	7	29	0.492536	2.850804	8	2.829333	7.310345	4.472414	267.9655	193.8621	14.41379	18.72414	4887.621	114.4828	0.689655	0.310345	-5.6E-17	0.275862	0.275862	0.448276	0.931034	0.034483	-5.6E-17	0	0.034483
	8	28	0.412083	2.689469	11	2.357426	8	4.410714	307.5714	198.0714	17.28571	24.5	4051.857	115.1786	0.142857	0.214286	0.642857	0.142857	0.607143	0.25	-5.6E-17	0.964286	-5.6E-17	0	0.035714
	9	77	0.355292	2.54843	14	2.186225	6	3.320779	209.1558	195.1558	19.44156	27.46753	3586.961	111	-3.3E-16	0.961039	0.038961	0.428571	0.012987	0.558442	0.012987	0.948052	1.39E-16	0	0.038961
	10	35	0.372256	2.222627	14	2.212118	4.371429	2.288571	188.7429	183	21.14286	28.68571	3332.543	105.5143	2.78E-17	0.971429	0.028571	0.114286	0.628571	0.257143	-5.6E-17	0.914286	-5.6E-17	0	0.085714
	11	19	0.466642	2.737929	8	2.357426	7.684211	4.384211	323	182.0526	16.68421	23.36842	3721.737	104.7368	0.105263	0.105263	0.789474	0.052632	0.631579	0.315789	-2.8E-17	0.157895	0.736842	0	0.105263
	12	74	0.411376	3.428406	10	2.567327	4	1.955405	133.027	174.7162	26.64865	33.98649	2664.959	101.2297	-2.8E-16	0.986486	0.013514	0.608108	0.081081	0.310811	-8.3E-17	0.878378	0.013514	0	0.108108
	13	26	0.521721	3.249214	8	2.459801	6.846154	4.161538	216.6538	207.0385	16.38462	22.30769	4220.385	119.4231	0.230769	0.076923	0.692308	0.153846	0	0.846154	-2.8E-17	0.538462	-5.6E-17	0.423077	0.038462
	14	46	0.402963	2.946242	9	2.186225	5.913043	2.930435	218.3913	182.6304	18.82609	26.26087	3573.522	108.1522	0.434783	0.021739	0.543478	0.23913	0.76087	1.67E-16	-5.6E-17	0.826087	-5.6E-17	0	0.173913

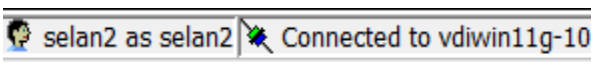

Windows 11 GIS Oracle - Desktop Viewer
Results - Node: Average Clustering Diagram: ES.Cluster Aalysis
File Edit View Window
Segment Size
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Search
12:57 PM
9/12/2024

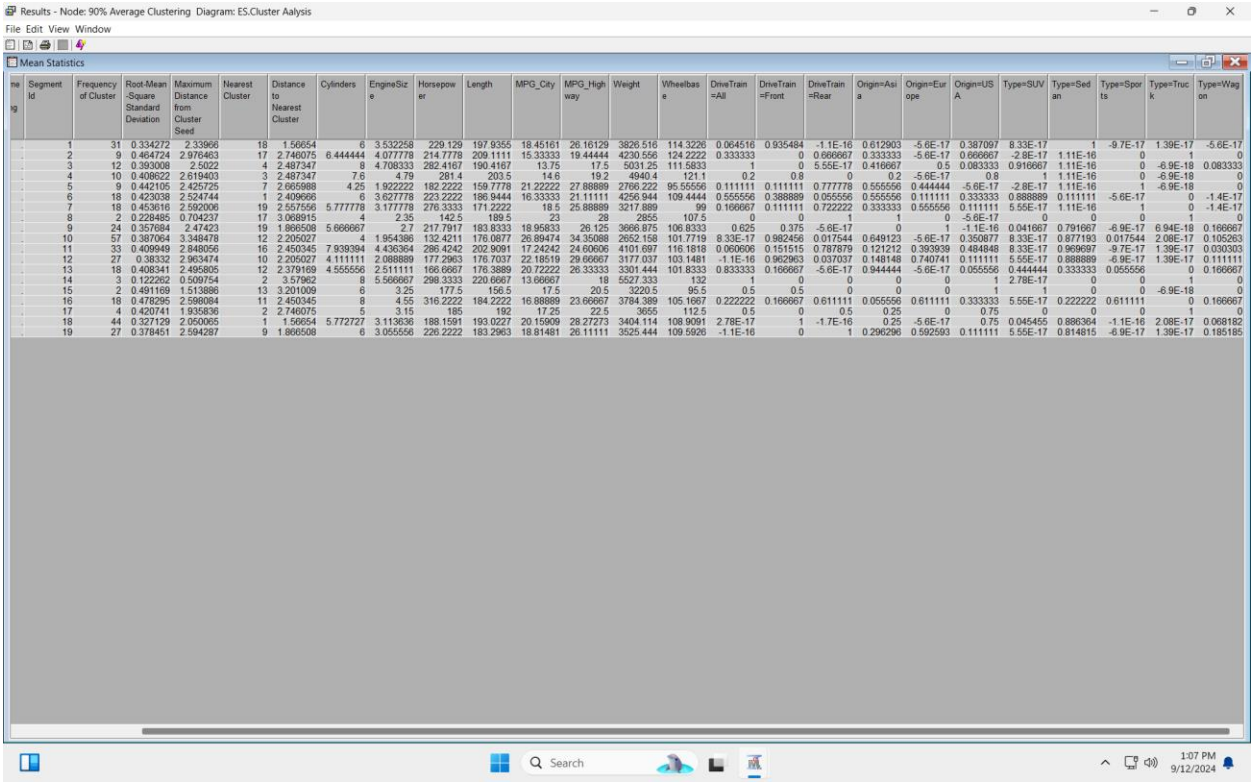

Windows 11 GIS Oracle - Desktop Viewer
Results - Node: Average Clustering Diagram: ES.Cluster Aalysis
File Edit View Window
Output
1 *--*
2 User: selan2
3 Date: September 12, 2024
4 Time: 12:18:57
5 *--*
6 * Training Output
7 *--*
8
9
10
11
12 Variable Summary
13
14 Measurement Frequency
15 Role Level Count
16
17 INPUT INTERVAL 8
18 INPUT NOMINAL 3
19 REJECTED BINARY 1
20 REJECTED INTERVAL 1
21
22
23
24 The CLUSTER Procedure
25 Average Linkage Cluster Analysis
26
27 Eigenvalues of the Covariance Matrix
28
29 Eigenvalue Difference Proportion Cumulative
30
31 1 413312.144 411781.568 0.9961 0.9961
32 2 1530.576 1444.005 0.0037 0.9998
33 3 86.570 81.716 0.0002 1.0000
34 4 4.854 0.638 0.0000 1.0000
35 5 4.217 3.660 0.0000 1.0000
Search
12:58 PM
9/12/2024

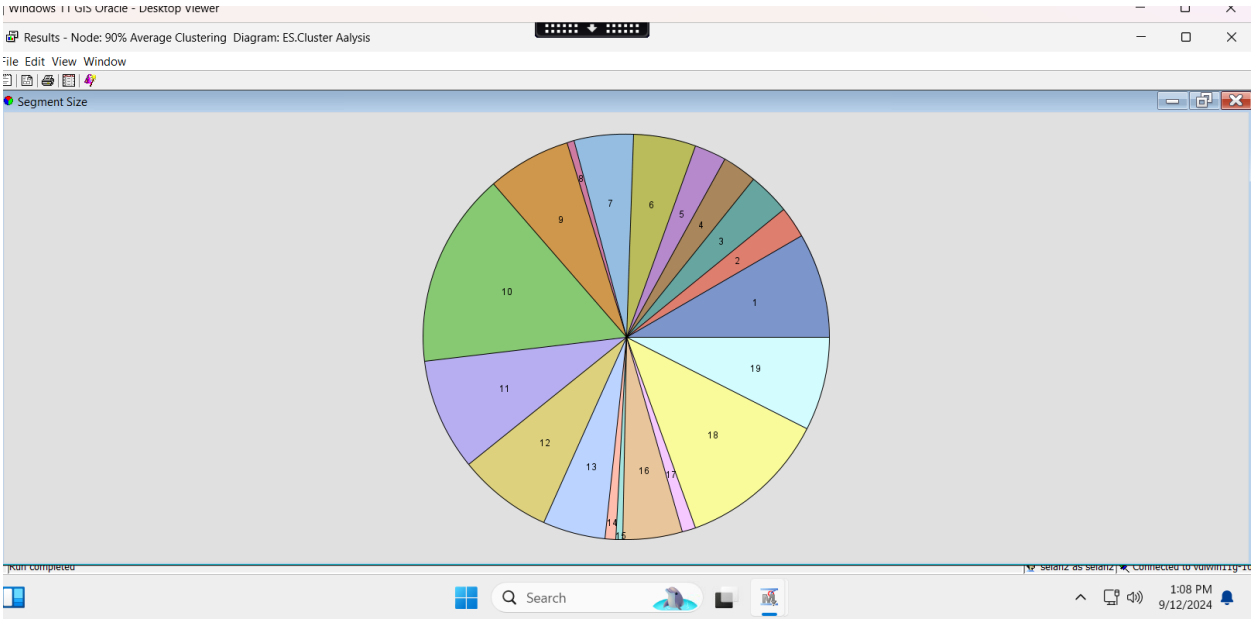

selan2 as selan2
Connected to vdiwin11g-10

Results window for *90% Average clustering:*

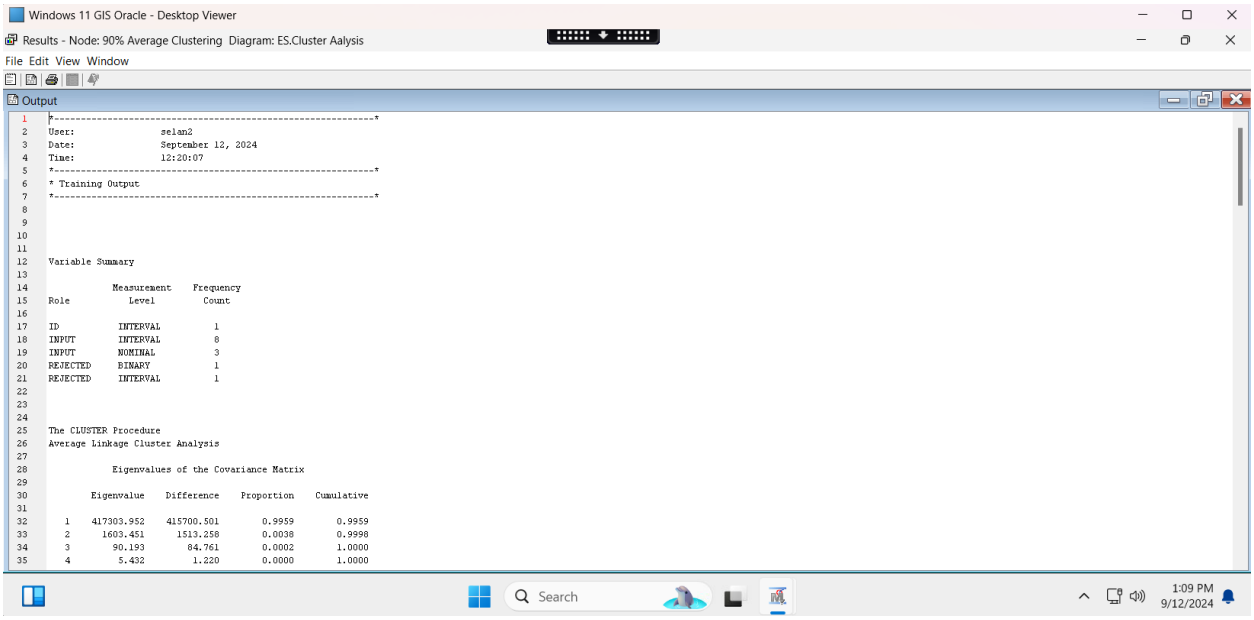


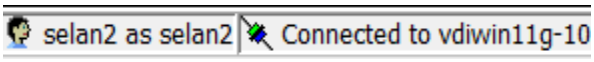



Results - Node: 90% Average Clustering Diagram: ES.Cluster Aalysis
File Edit View Window
Mean Statistics
Search
1:07 PM
9/12/2024


Results - Node: 90% Average Clustering Diagram: ES.Cluster Aalysis
File Edit View Window
Segment Size
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Search
1:08 PM
9/12/2024


selan2 as selan2
Connected to vdiwin11g-10





60. If the number of clusters for centroid clustering before or after sampling is almost same (+/- 2 clusters), then the clusters derived from Centroid clustering are stable as mentioned in the notes of Step 40. Similarly, if the number of clusters for average clustering before or after sampling is almost same (+/- 2 clusters), then the clusters derived from Average clustering are stable as mentioned in the notes of Step 40. However, if the number of clusters are greater than 6 it is hard to interpret the results of clustering to generate meaningful insights.
61. Therefore, if the number of clusters for centroid and/or average clustering is greater than 6 then limit the number of clusters as discussed in "Limit the Number of Cluster" Sections below and then see out of the Ward, Centroid and Average clustering which cluster algorithm provides the most stable clusters with reasonable number of clusters to interpret the patterns and articulate an interesting story to classify the cars in different categories.
62. Recall the analysis you have performed for Ward Clustering in Question 5 (5a and 5b) and Question 10 (10a and 10b). Perform the same analysis for Centroid Clustering and Average Clustering. Is the analysis in table of Question 5b and discussion in Question 10b same as the analysis for Centroid clustering and Average clustering? If it is same, then your clusters obtained from Centroid and Average clustering are stable, and you can use the results of any one of the Ward, Average, or Centroid clustering to tell an interesting story based on the results of exploratory data analysis.
63. Close all the Results Window.

Limit the number of Clusters (User Specified)

64. To recall, if the number of clusters are greater than 6 for any clustering algorithm it is hard to interpret the results of clustering to generate meaningful insights. So, now we will limit the number of clusters for Centroid and Average Clustering.

65. Click *Centroid Clustering* node. In the properties window of *Centroid Clustering* node set *Specification Method* property to *User Specify and Maximum number of clusters to 6 and hit enter*. Also, ensure the *Internal Standardization* property of *Centroid Clustering* node is set to standardization.

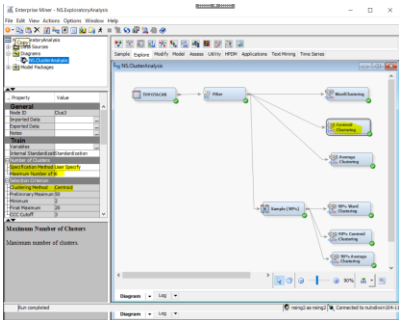


Figure 21

66. Right click *Centroid Clustering* node and click on Run to re-execute this node. Open the results window and see how many clusters are there in the segment size window?
67. Click *90% Centroid Clustering* node. In the properties window of *90% Centroid Clustering* node set *Specification Method* property to *User Specify and Maximum number of clusters to 6 and hit enter*. Also, ensure the *Internal Standardization* property of *Centroid Clustering* node is set to standardization.

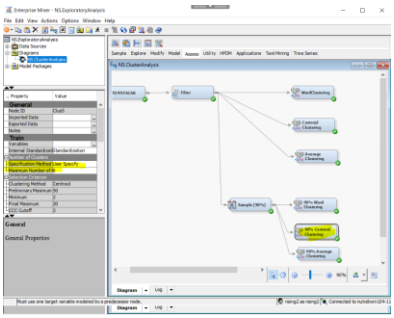


Figure 22

68. Right click 90% Centroid Clustering node and click on Run to re-execute this node. Open the results window and see how many clusters are there in the segment size window?
69. Now repeat Step 62 to see if the information represented is almost like Ward clustering.
70. Close all the results windows.
71. Click *Average Clustering node*. In the properties window of *Average Clustering* node set *Specification Method* property to *User Specify and Maximum number of clusters to 6 and hit enter*. Also, ensure the *Internal Standardization* property of *Average Clustering* node is set to standardization.
72. Right click *Average Clustering* node and click on Run to re-execute this node. Open the results window and see how many clusters are there in the segment size window?
73. Click *90% Average Clustering* node. In the properties window of *90% Average Clustering* node set *Specification Method* property to *User Specify and Maximum number of clusters to 6*

and hit enter. Also, ensure the *Internal Standardization* property of *Centroid Clustering* node is set to standardization.

74. Right click *90% Average Clustering* node and click on Run to re-execute this node. Open the results window and see how many clusters are there in the segment size window?
75. Now repeat Step 62 to see if the information represented is same before or after sampling for Ward clustering, Centroid Clustering, and Average Clustering. Out of the three clustering algorithms identify which one is the most stable and we can use that cluster algorithm to make recommendations to dealer for classifying the cars and use it for marketing the cars to customers.
76. Close all results windows.

Question 12

Based on the cluster analysis performed in this assignment, do you think you were doing the Hierarchical Clustering (Agglomerative or Divisive) AND/OR Non-Hierarchical Clustering (k-means) clustering. Explain your answer in 150-200 words.

The clustering assignment is a comparison of Hierarchical and non-hierarchical clustering procedures. In this project, we used Ward clustering, Centroid clustering, and Average clustering. Therefore, we applied Agglomerative Hierarchical clustering because all the above-mentioned algorithms (Ward, Centroid, and Average) are types of Agglomerative Hierarchical Clustering methods. According to some research, Ward's method is the best for clustering. Agglomerative hierarchical clustering begins with each observation in a separate cluster, and pairs of clusters are merged as one moves up the hierarchy. It aims to merge the pair of clusters with the least increase in total within-cluster variance at each step.

Unlike k-means clustering (a non-hierarchical technique that partitions data into a set number of clusters where each observation belongs to the cluster with the nearest mean), Ward's method doesn't require pre-specifying the number of clusters. Instead, it builds a hierarchical structure represented by a dendrogram. In this case, we built a hierarchy of Toyota cars based on their attributes using Ward Clustering, merging clusters based on feature similarity, demonstrating the agglomerative hierarchical clustering approach.

At the end of the project, we updated the number of clusters to 6 in the Average and Centroid clustering processes. We fixed the maximum number of clusters to a k value (k=6), which corresponds to k-means clustering. Key clues pointing to k-means include, The cluster analysis required specifying the number of clusters and optimizing them using centroids and distances, typical of k-mean, the analysis required limiting the number of clusters, a common practice in non-hierarchical methods like k-means.So, we also performed Non-Hierarchical clustering. In SAS Enterprise Miner, we can only run Hierarchical clustering by default, but we can change the properties to perform Non-Hierarchical clustering.

Clusters using variables of your choice

77. Now we will perform **clustering using only SIX variables of our choice to respond to Question 13**. So, create a new diagram and name it as "MyClusters".
    - Drag and drop TOYOTACAR data from Data Sources to "MyClusters" Diagram.
    - Right click on Toyotacar node and set the role of any six (6) variables as Input and the role of rest all variables as Rejected. (*Hint: You can choose the six variables based on your learning in this assignment or based on the important variables identified in the assignment so far or based on your domain knowledge*).
    - Now, follow items/steps/bullets 5 to 76 of this assignment to replicate the cluster analysis you have performed in this assignment. The only difference is this time you have only six Input variables. **To summarize, kindly ensure that after choosing the six variables you perform following and ONLY answer Question 13 in the end** (***you are not going to re-answer Questions 1 to 12 while repeating the above steps to perform cluster analysis with six variables of your choice***):
        - *Filter the Outliers and Standardize the Interval variables for Cluster Analysis (as done in steps 6 to 13 of this assignment).*
        - *Explore the results of cluster analysis (as done in steps 14 to 29 of this assignment).*
        - *Explore the variable importance of the new cluster analysis (as done in steps 30 to 33 of this assignment).*
        - *Analyze the results of this cluster analysis using tree structure and rules (as done in steps 34 to 39 of this assignment).*
        - *Test the stability of clusters using the sample of the data (as done in steps 40 to 49 of this assignment).*
        - *Perform Cluster Analysis using Centroid and Average Distance Metric (as done in steps 50 to 63 of this assignment).*
        - *Limit the number of Clusters (User Specified) to perform K-means clustering (as done in steps 64-76 of this assignment)*

Question 13

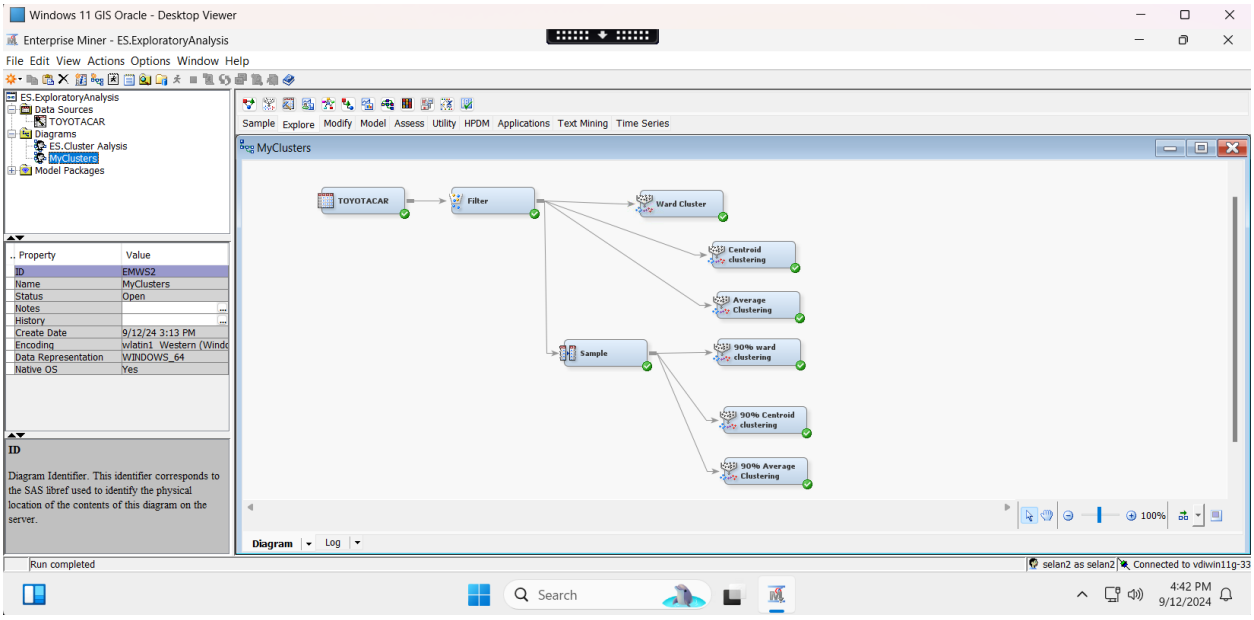
Answer 13a and 13b to summarize your cluster analysis with six variables of your choice.

Question 13a

Provide the list of SIX variables you have selected to perform the cluster analysis in Step 76. Also, provide the screenshot of your "MyClusters" Diagram including the NETID and MACHINE NUMBER on the right lower corner.

Answer:

The list of six variables that I have selected to perform the cluster analysis are cylinders, Engine size, Horsepower, MPG_City, MPG_Highway, and Weight.



Question 13b

Discuss the rationale for selecting the six variables to perform the cluster analysis in Step 76. Also, discuss the results of your cluster analysis with six variables including the stability of the clusters. Also, explain if you are the dealer how you will convince a customer to buy used Toyotacar based on the exploratory analysis (cluster analysis) you have performed using the six variables of your choice.

The six variables selected for cluster analysis are critical in determining meaningful and actionable clusters.

- Weight: Heavier vehicles tend to consume more fuel but offer better stability. Including weight helps differentiate lighter cars from trucks and SUVs.
- Engine Size: Larger engines offer more power but are less fuel-efficient. This variable helps balance performance with fuel economy.
- Horsepower: A key indicator of vehicle performance, horsepower separates high-performance vehicles from more fuel-efficient ones.
- Cylinders: More cylinders usually mean more power but lower fuel efficiency, helping to categorize sports cars versus economy models.
- City MPG: This measures fuel efficiency in urban settings, useful for distinguishing vehicles suited for city driving.

- Highway MPG: Key for long-distance commuters, this differentiates cars based on fuel efficiency at highway speeds.

Stability of Clusters:

Consistent results from ward, Centroid, and average clustering analyses, particularly regarding Weight, engine size, and Horsepower. The cluster models suggest 4 to 6 clusters, which seems to be good sign for stability as discussed earlier in our analysis. The Eigenvalues from Ward's Minimum Variance analysis show around 99.56–99.58% of variance explained by the first principal component, reflecting strong differentiation among clusters. A root-mean-square standard deviation of 276–274 indicates low variation within clusters.

Results of your cluster analysis:

Weight is the most significant factor across all clustering models, followed closely by EngineSize and Horsepower. These three variables heavily influence the segmentation of vehicles. MPG_City and MPG_Highway are also important but less impactful compared to engine-related variables. Clusters are distinguished by features like cylinders, engine size, and fuel efficiency (MPG). For example, high-horsepower vehicles with larger engines are clustered separately from smaller, fuel-efficient cars. This grouping effectively targets customer preferences based on power, fuel efficiency, or a balance of both.

Dealer Action Based on this Analysis:

- Focus on vehicles in clusters with higher horsepower, larger engines, and more cylinders. These customers prioritize power over fuel efficiency, so emphasize performance in a sales pitch.
- Clusters with smaller engine sizes and higher MPG_City and MPG_Highway values attract fuel-conscious customers. Highlight fuel efficiency to appeal to eco-conscious buyers.
- Use cluster analysis to show customers data-driven insights into vehicle segmentation, matching their preferences with specific vehicle clusters.
- Customize your sales approach by demonstrating the balance between performance and fuel efficiency in mid-sized engine clusters, helping customers make informed choices.

Appendix – Data Set Variable Description

Variable Name	Variable Description	Roles	Variable Type (Levels)
Cylinders	The number of cylinders of the engine	Input	Interval
DriveTrain	Whether the automobile is "All" wheel, "Rear" wheel, or "Front" wheel drive	Input	Nominal
EngineSize	The size of the engine	Input	Interval
Expensive	Whether the Invoice price of the automobile is greater than or equal to $28,000 or not (binary variable: 1 if invoice >=28,000; 0 otherwise)	Rejected	Binary
Horsepower	The horsepower of the engine	Input	Interval
Invoice	The invoice price of the automobile (i.e., the purchase price)	Rejected	Interval
Length	The length of the automobile	Input	Interval
MPG_City	Miles per gallon (in city)	Input	Interval
MPG_Highway	Miles per gallon (on highway)	Input	Interval
Origin	The origin of the automobile	Input	Nominal
Type	The type of the automobile (whether hybrid, sedan, sports, SUV, truck, or wagon)	Input	Nominal
Weight	The weight of the automobile	Input	Interval
Wheelbase	The distance between the front and rear axles of the automobile	Input	Interval

Go back to setting roles and levels