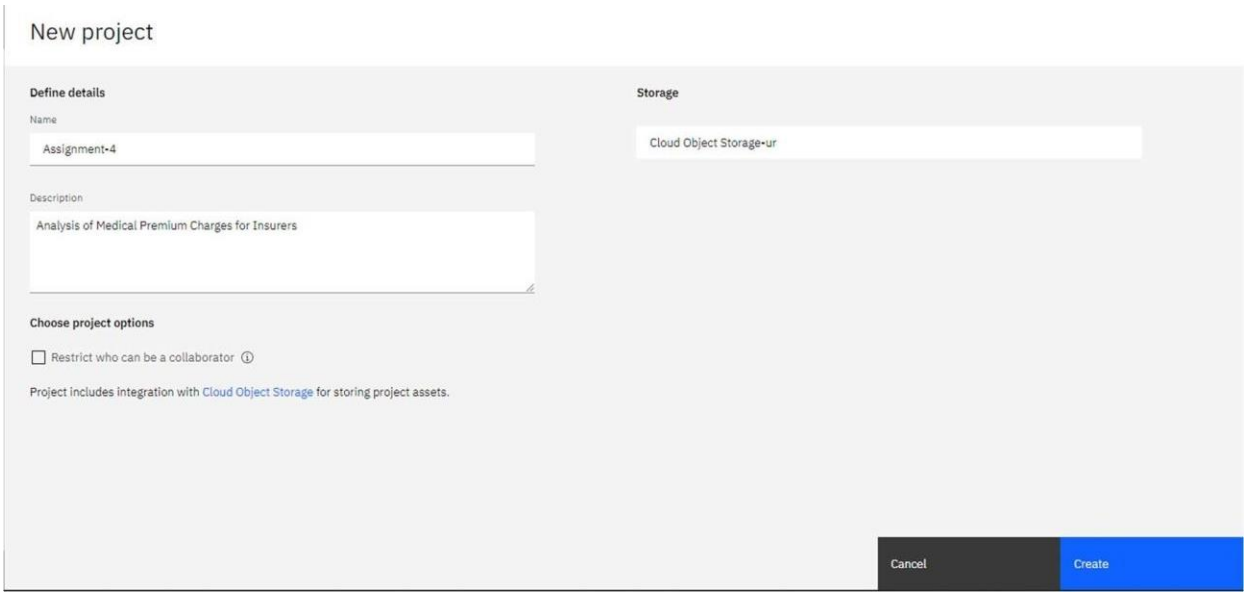
Data Analytics Externship Program
Assignment-4

Name: Kadiyala Meghanath
Reg no: 19MIM10097
Campus: VIT Bhopal

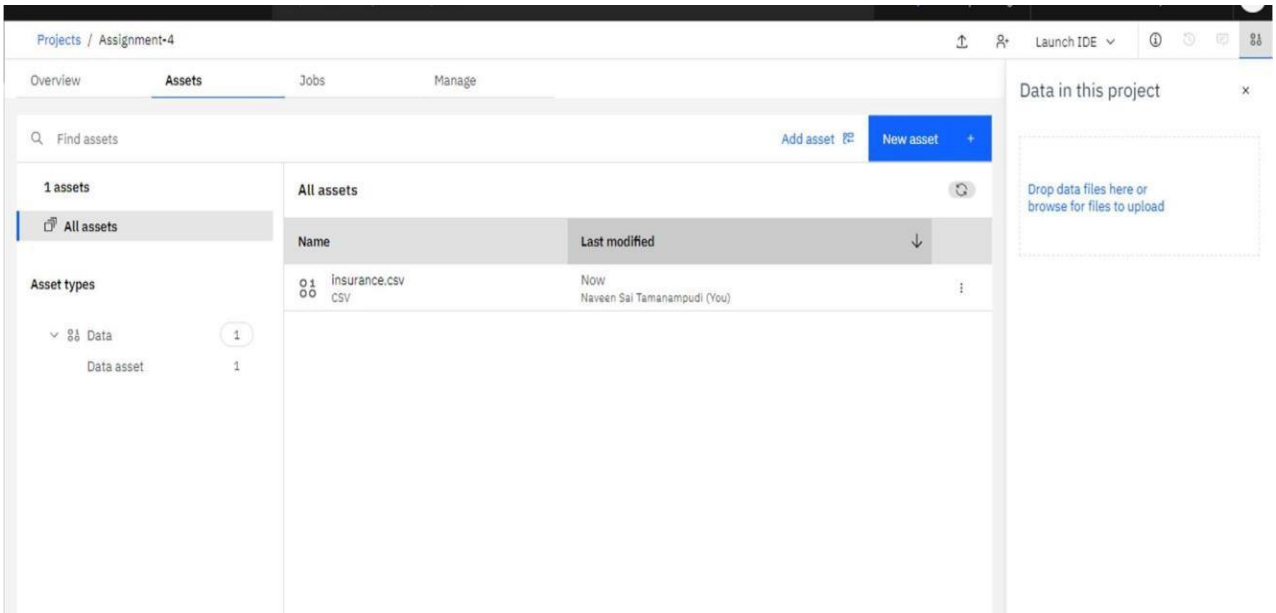
Clustering: Analysis of Medical Premium Charges for Insurers

Dataset Used: insurance.csv

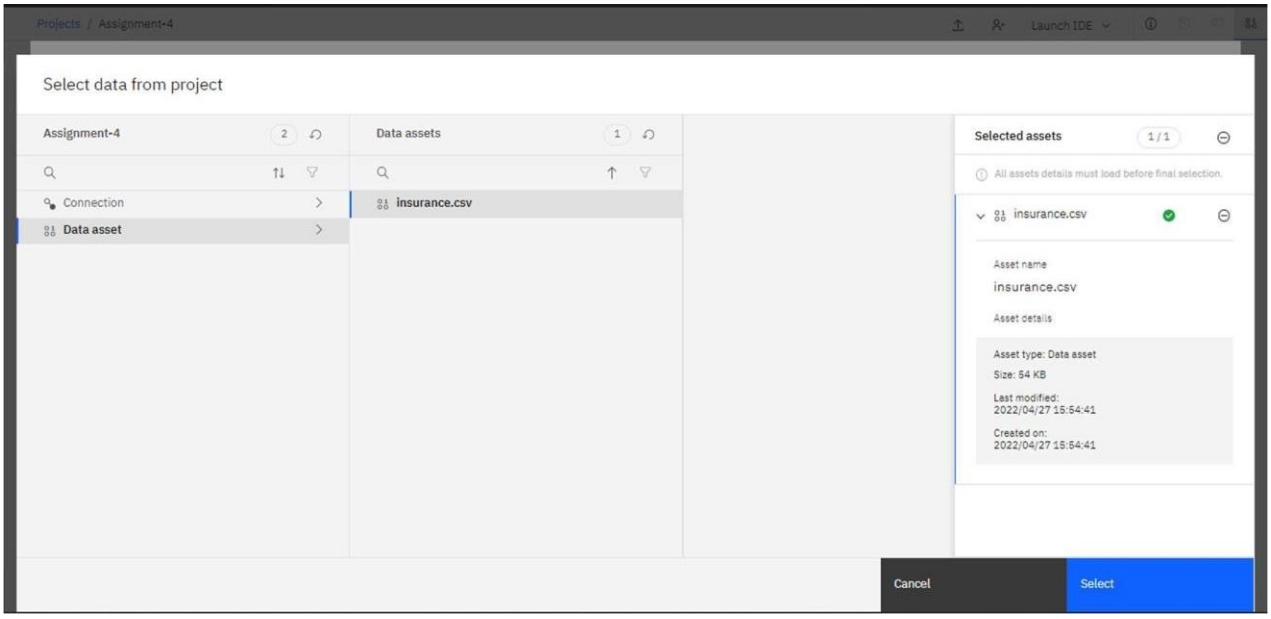
<u>Creating the Project</u>



Uploading Data:



Creating a Data Refinery:



Viewing the output of refinery:

Projects / Assignment-4 / insurance.csv / Refine data

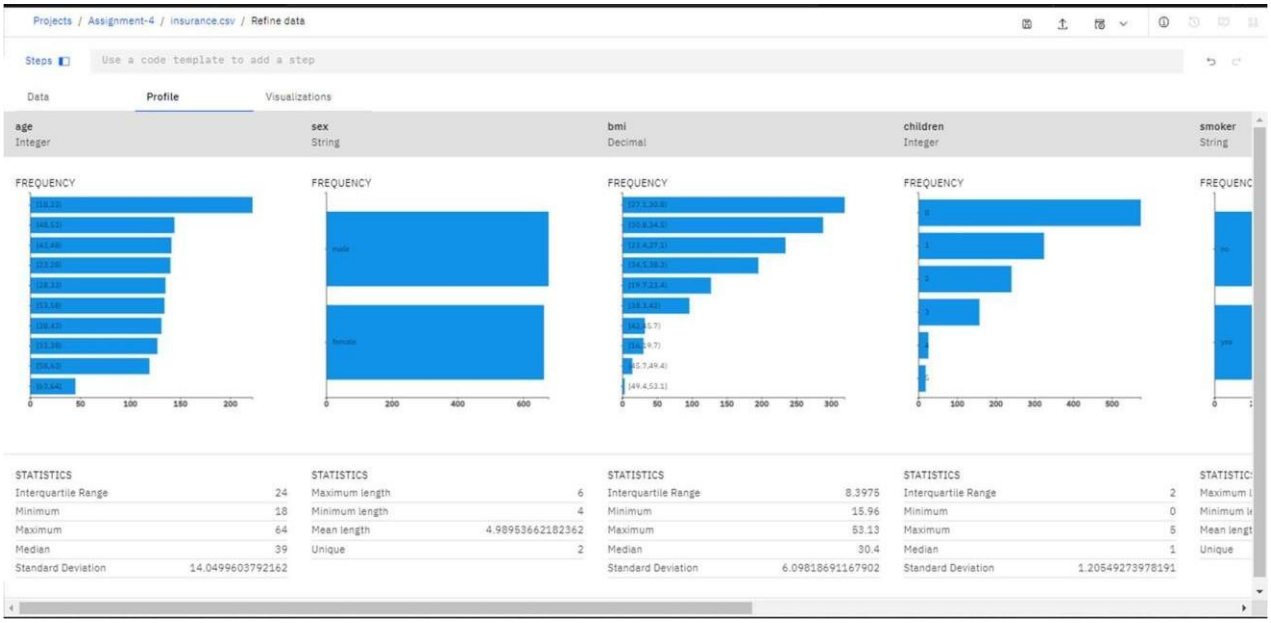
Steps | Use a code template to add a step

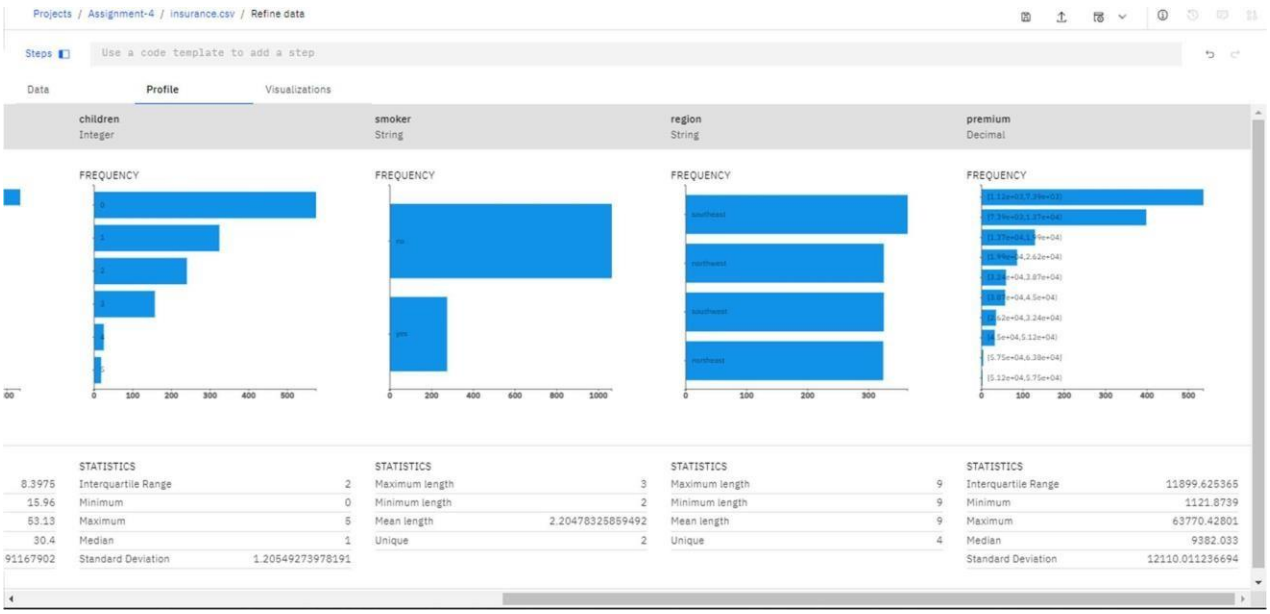
Data | Profile | Visualizations

	age Integer	sex String	bmi Decimal	children Integer	smoker String	region String	premium Decimal
1	19	female	27.9	0	yes	southwest	16884.924
2	18	male	33.77	1	no	southeast	1725.5523
3	28	male	33	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.47061
5	32	male	28.88	0	no	northwest	3866.8552
6	31	female	25.74	0	no	southeast	3756.6216
7	46	female	33.44	1	no	southeast	8240.5896
8	37	female	27.74	3	no	northwest	7281.5056
9	37	male	29.83	2	no	northeast	6406.4107
10	60	female	25.84	0	no	northwest	28923.13692
11	25	male	26.22	0	no	northeast	2721.3208
12	62	female	26.29	0	yes	southeast	27808.7251
13	23	male	34.4	0	no	southwest	1826.843
14	56	female	39.82	0	no	southeast	11090.7178
[illegible]	27	male	42.13	0	yes	southeast	39611.7577

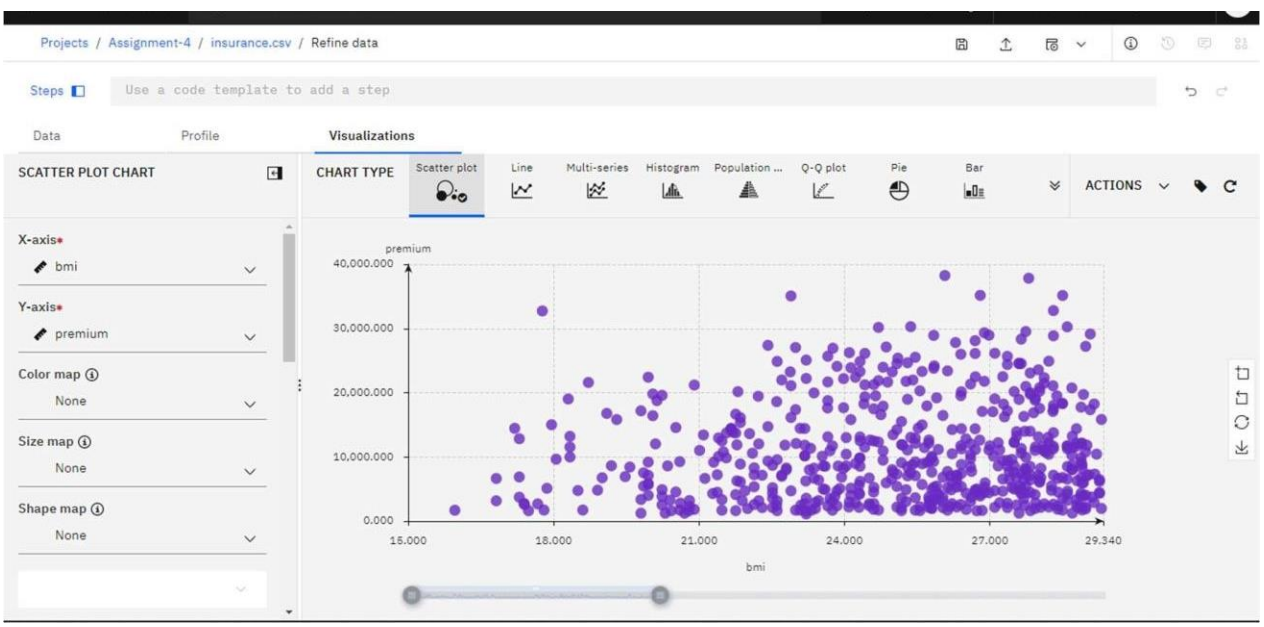
SOURCE FILE: insurance.csv FULL DATA SET: 1338 rows

Viewing Profile:

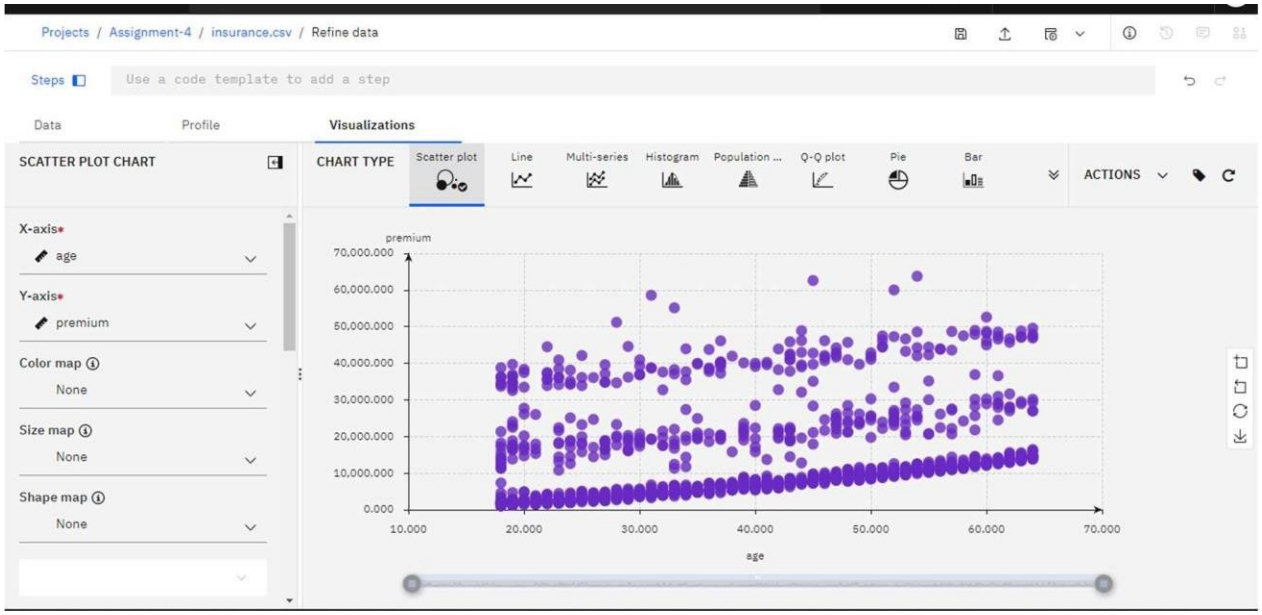




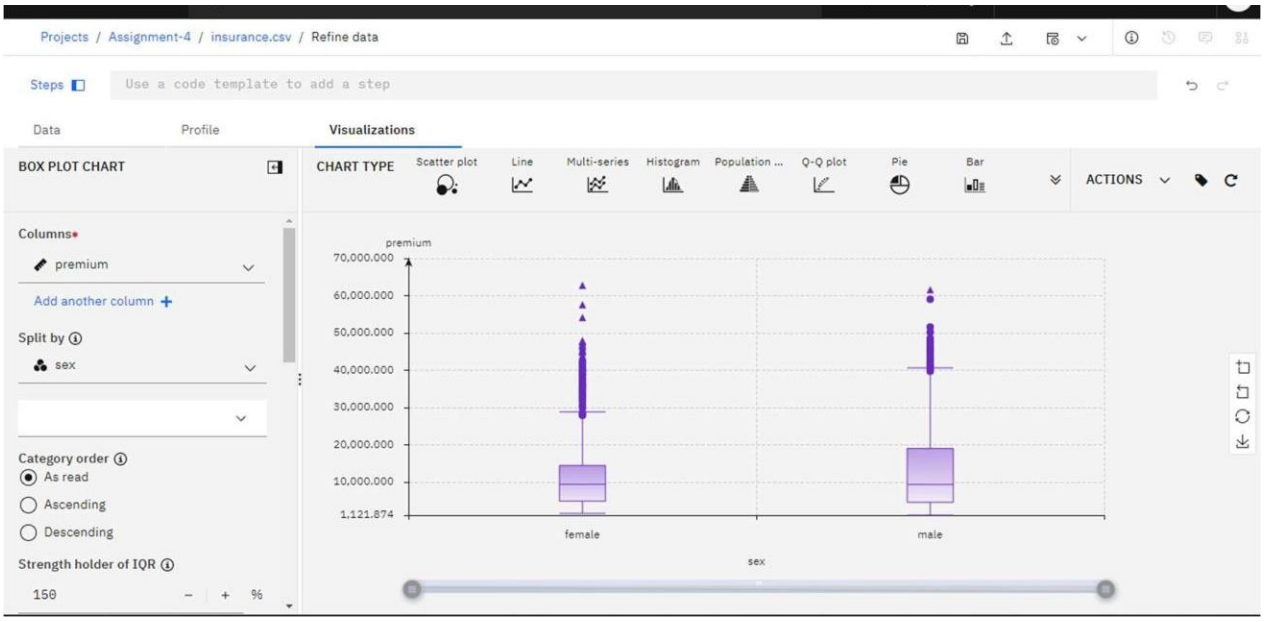
Viewing Scatterplot of BMI vs Premium:



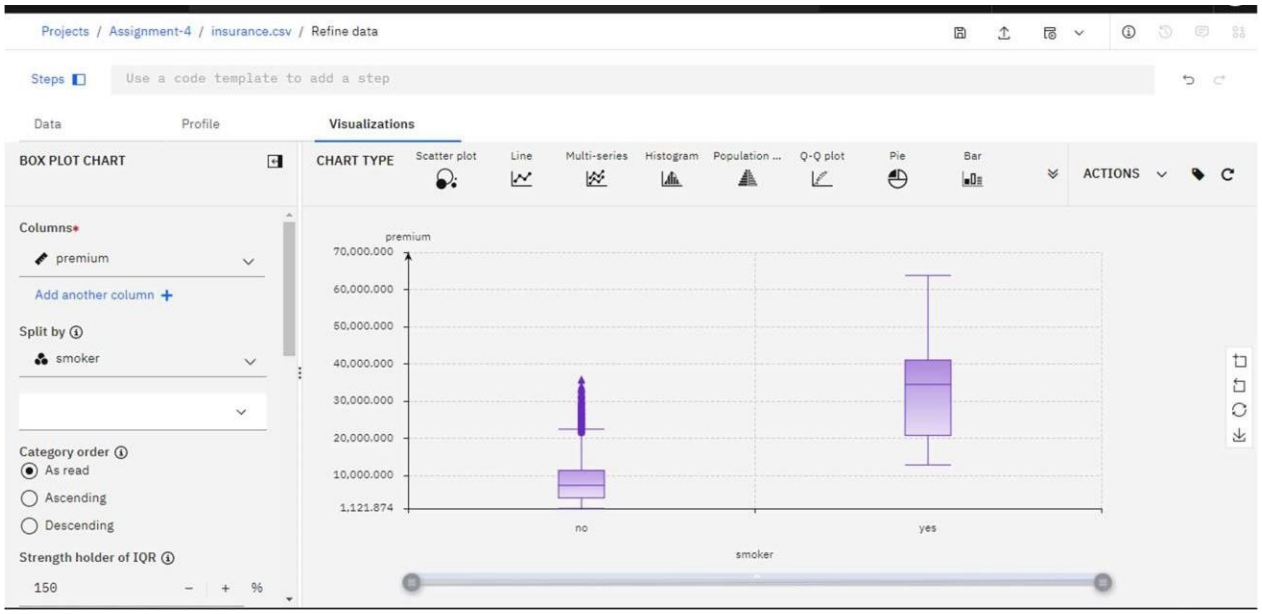
Viewing Scatterplot of Age vs Premium:



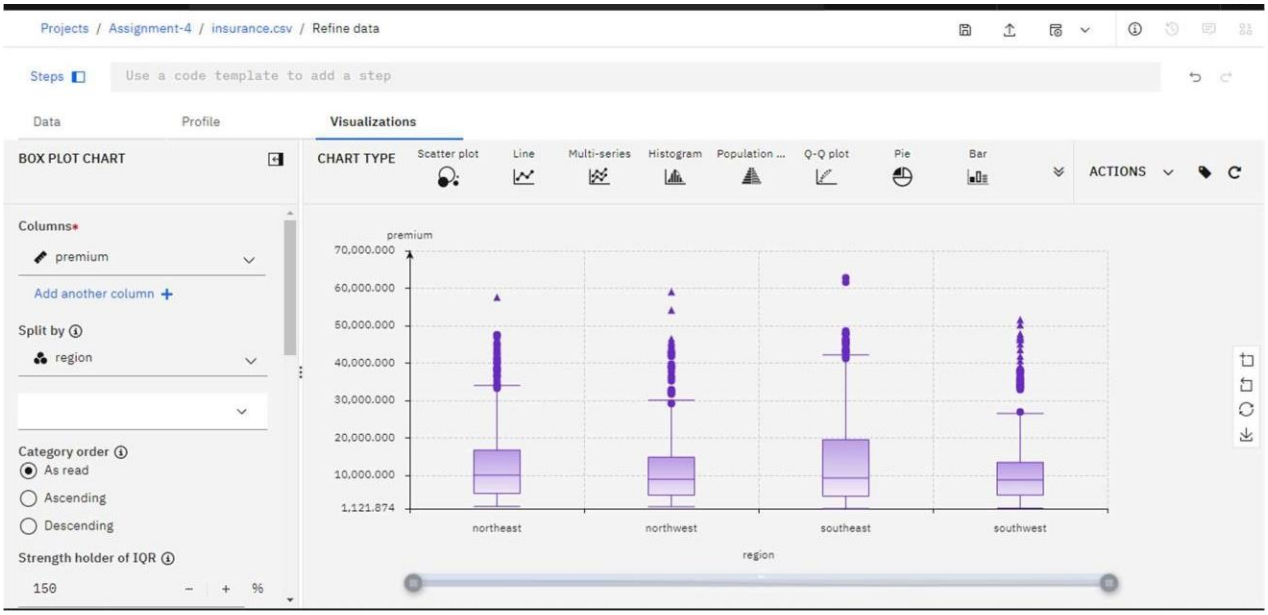
Viewing Boxplot of Sex vs Premium:



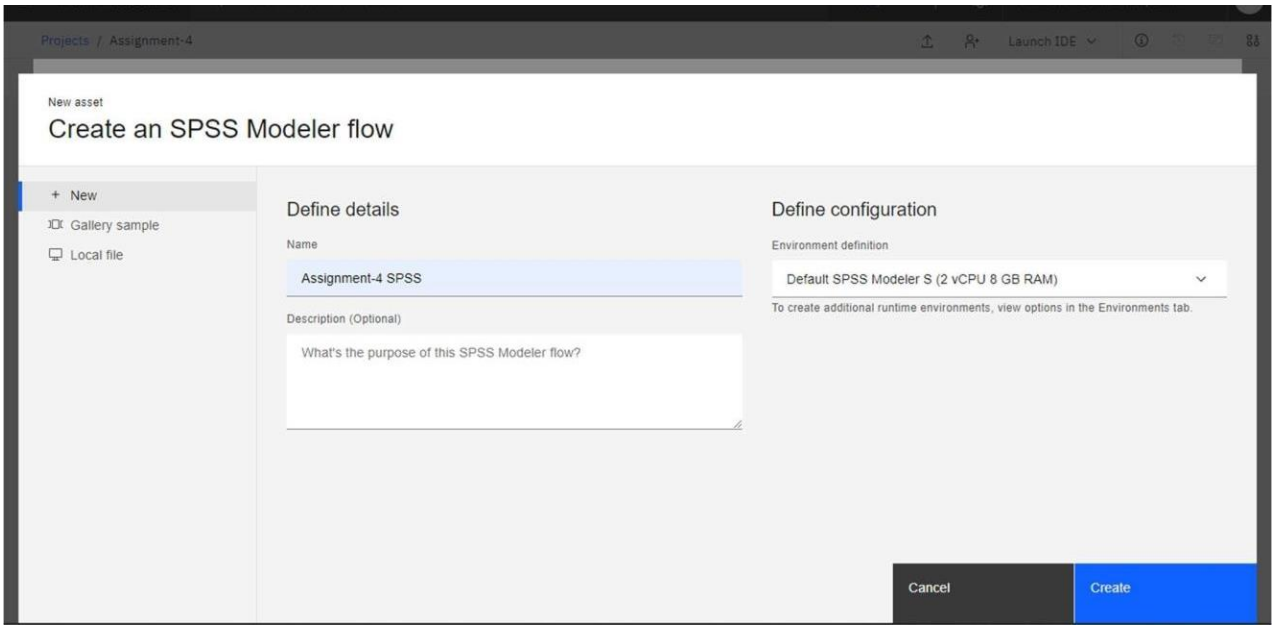
Viewing Boxplot of Smoker vs Premium:



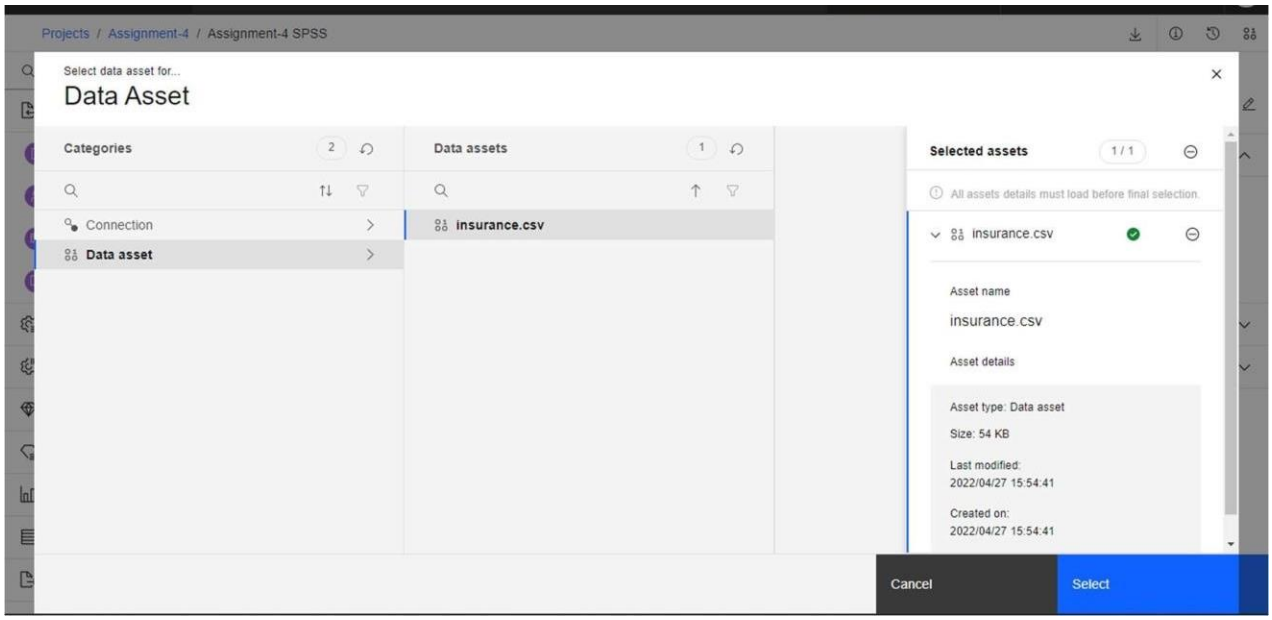
Viewing Boxplot of Region vs Premium:

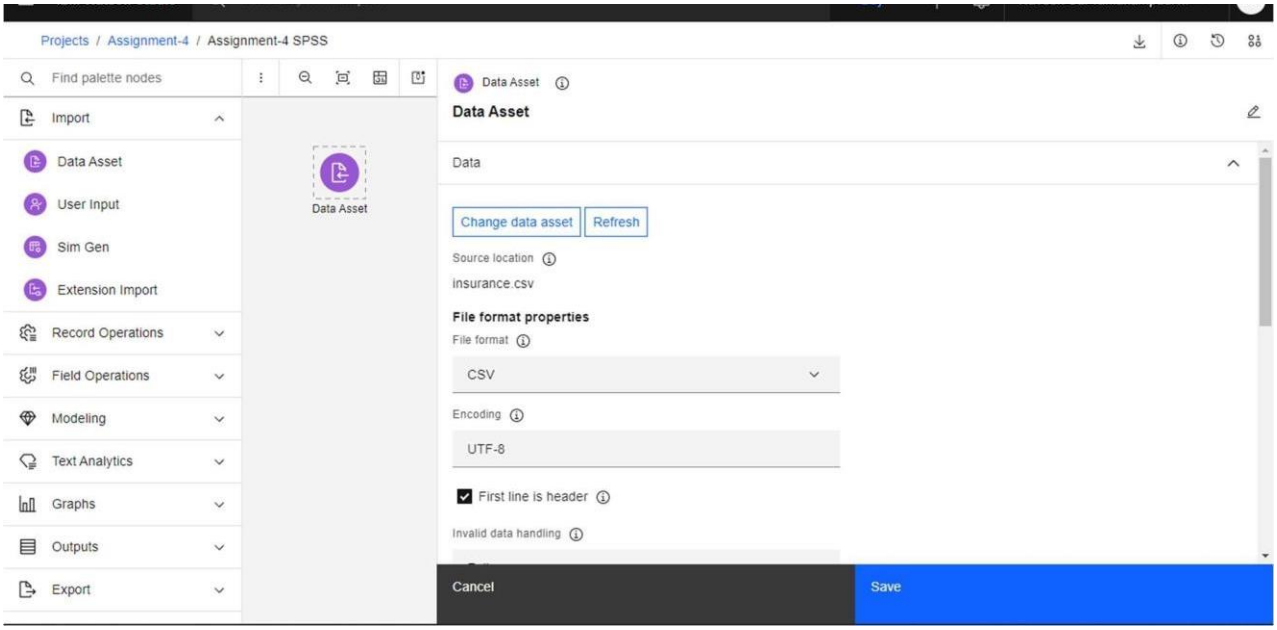


Creating an SPSS Modeler:

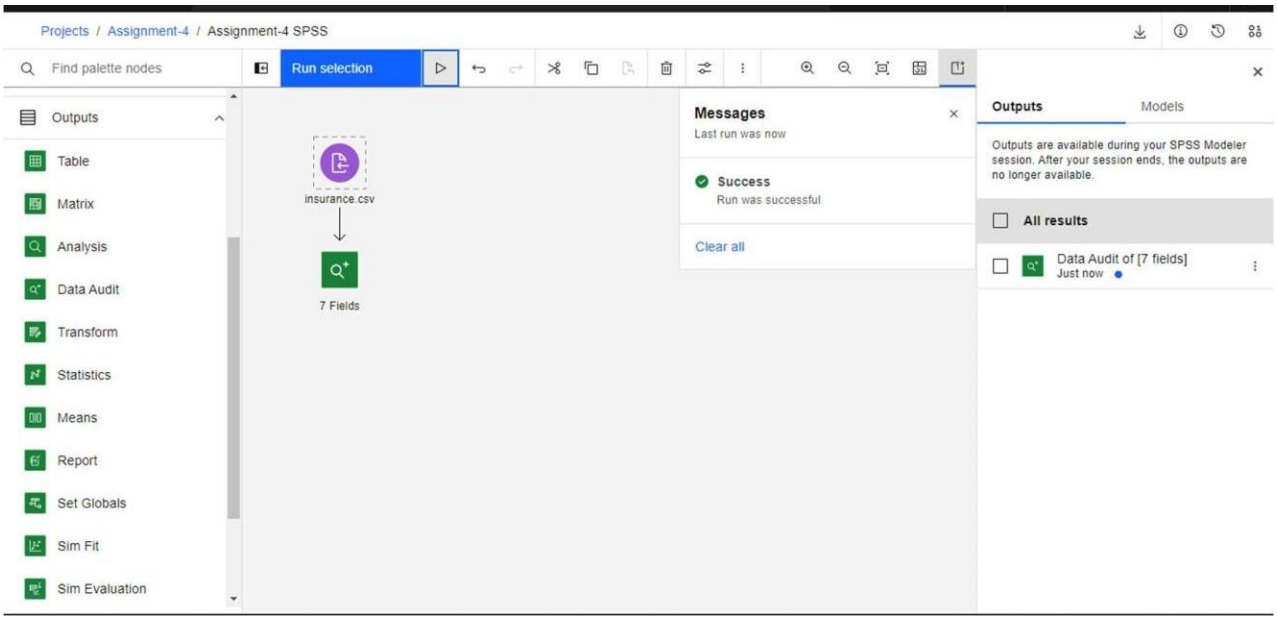


Uploading Dataset to Data Asset Node:





Creating Data Audit Node:



Output from Data Audit:

Projects / Assignment-4 / Assignment-4 SPSS

View Output: Data Audit of [7 fields]

Compare

	Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
1	age		Continuous	18	64	39.207	14.050	0.056	--	1338
2	sex		Categorical	--	--	--	--	--	2	1338
3	bmi		Continuous	15.960	53.130	30.663	6.098	0.284	--	1338
4	children		Continuous	0	5	1.095	1.205	0.938	--	1338
5	smoker		Categorical	--	--	--	--	--	2	1338
6	region		Categorical	--	--	--	--	--	4	1338

Projects / Assignment-4 / Assignment-4 SPSS

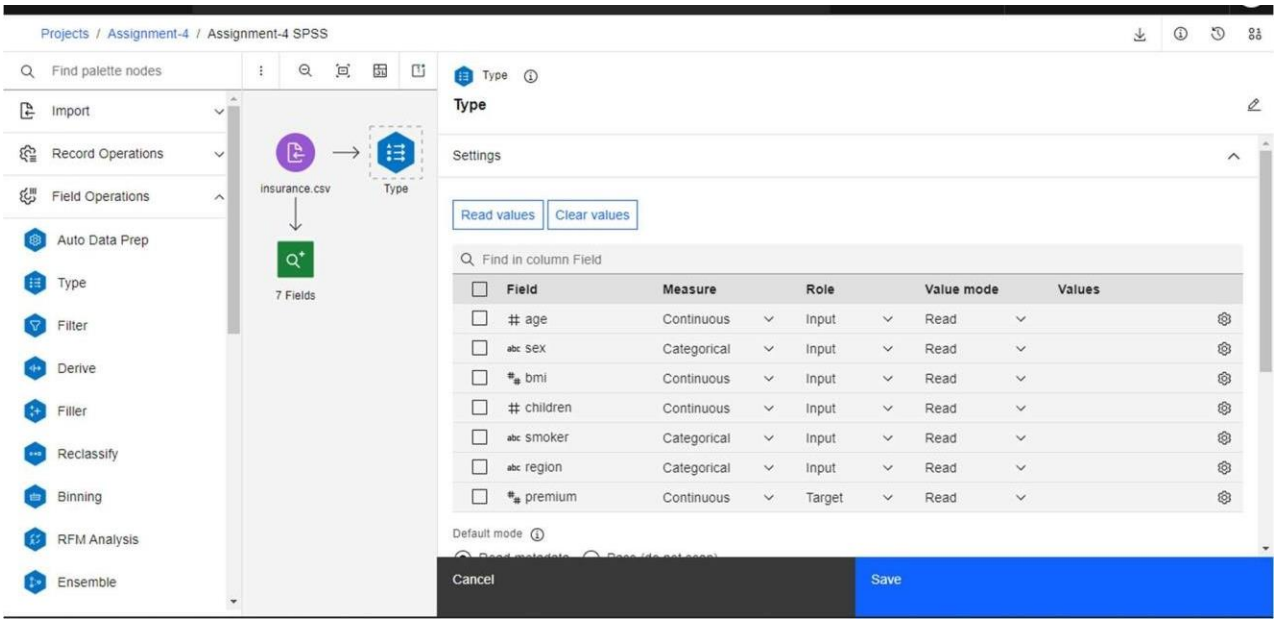
View Output: Data Audit of [7 fields]

Compare

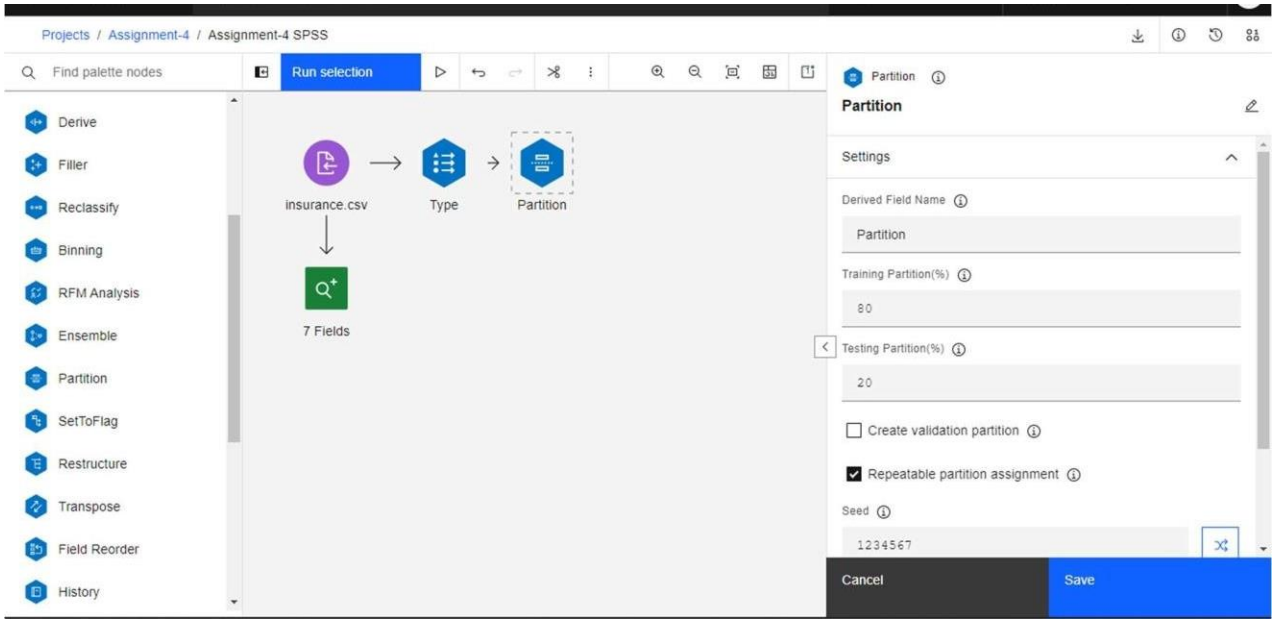
7	premium		Continuous	1121.874	63770.428	13270.422	12110.011	1.516	--	1338

	Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value
1	age	Continuous	0	0	None	Never	Fixed	100.000	1338	0
2	sex	Categorical	--	--	--	Never	Fixed	100.000	1338	0
3	bmi	Continuous	4	0	None	Never	Fixed	100.000	1338	0
4	children	Continuous	18	0	None	Never	Fixed	100.000	1338	0
5	smoker	Categorical	--	--	--	Never	Fixed	100.000	1338	0
6	region	Categorical	--	--	--	Never	Fixed	100.000	1338	0
7	premium	Continuous	7	0	None	Never	Fixed	100.000	1338	0

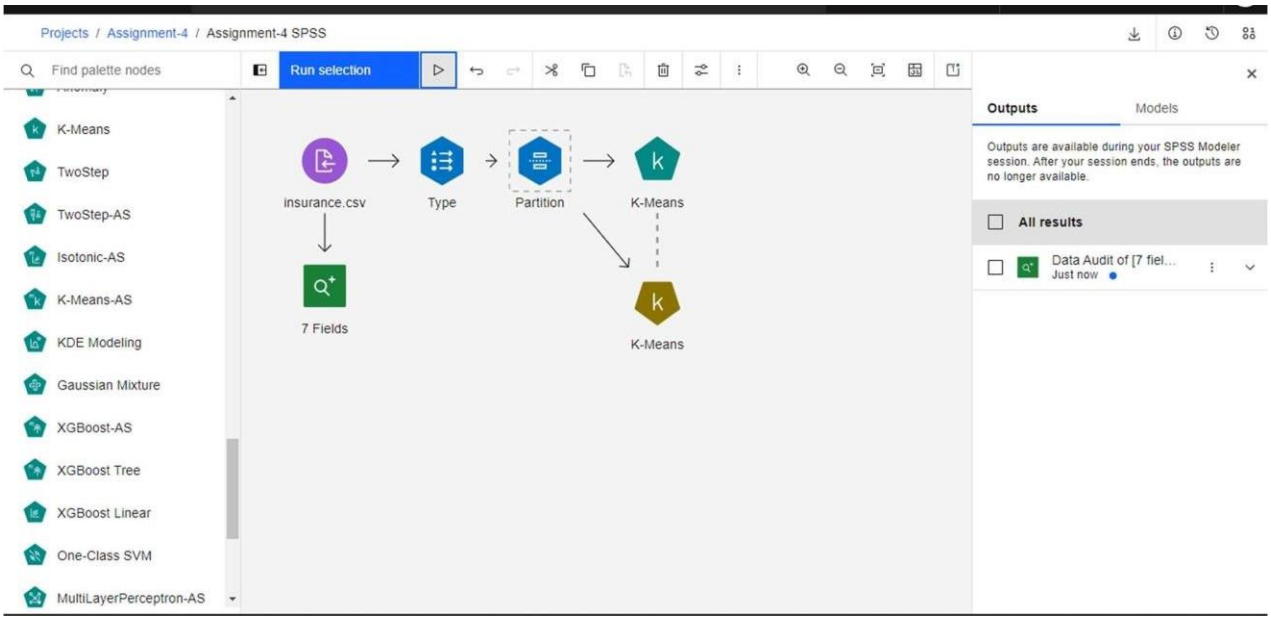
Creating a Type Node and setting Inputs, Targets:



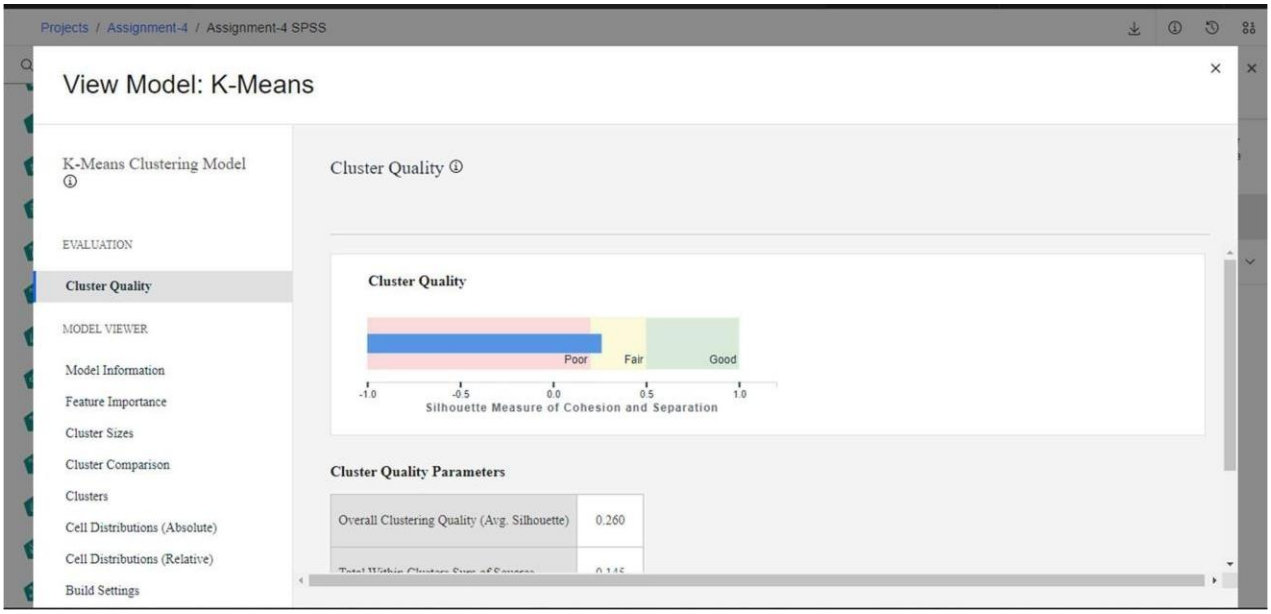
Creating a Partition Node with 80:20 split:

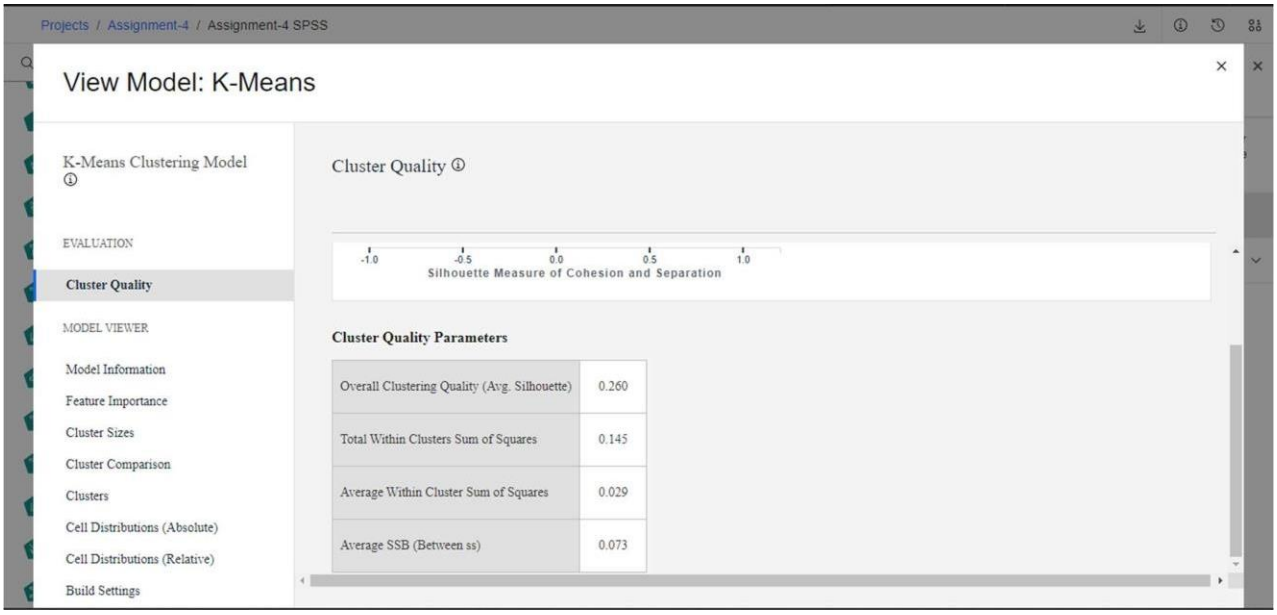


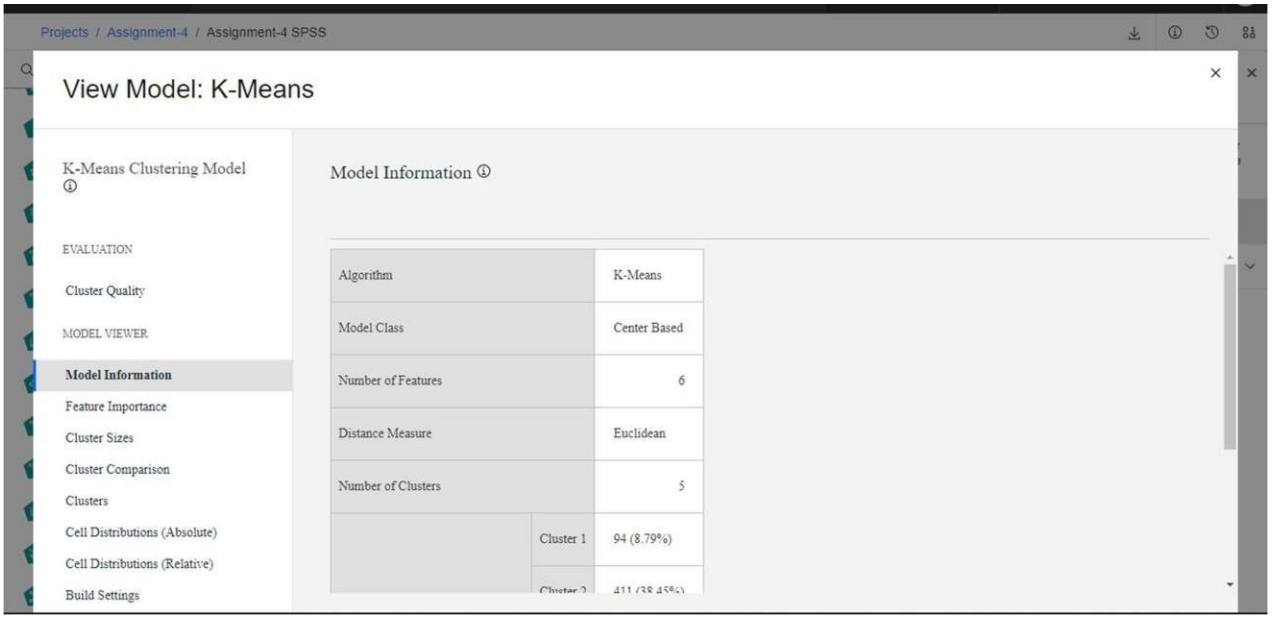
Creating a K-Means Clustering Model:

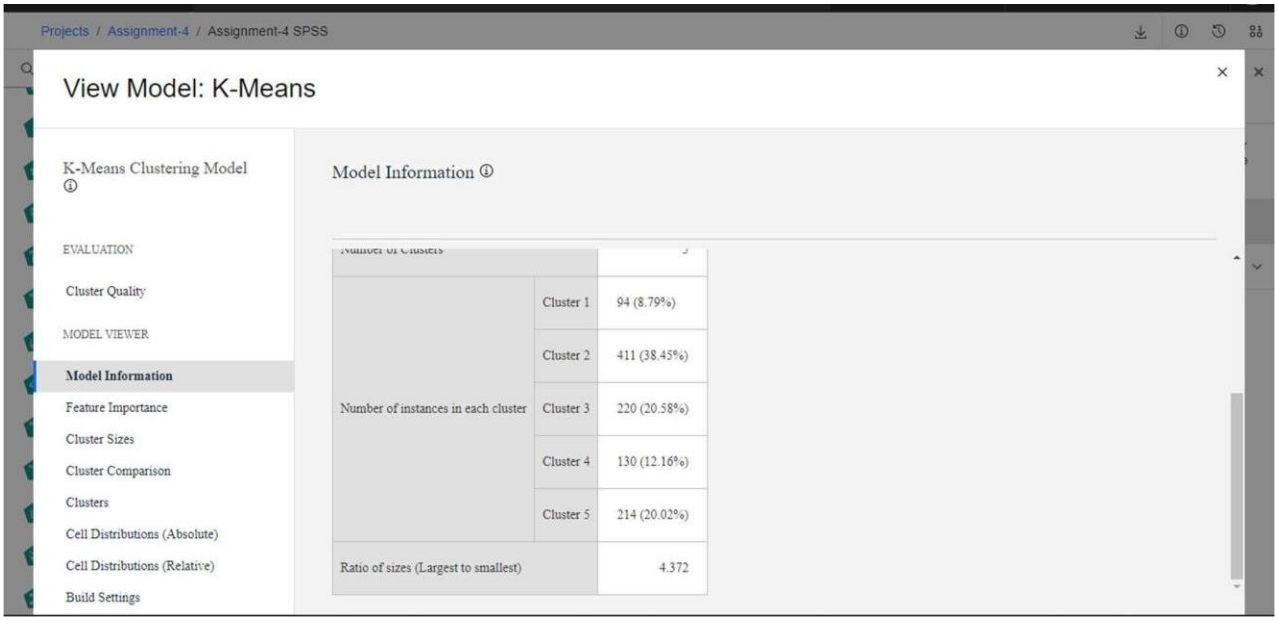


Output:

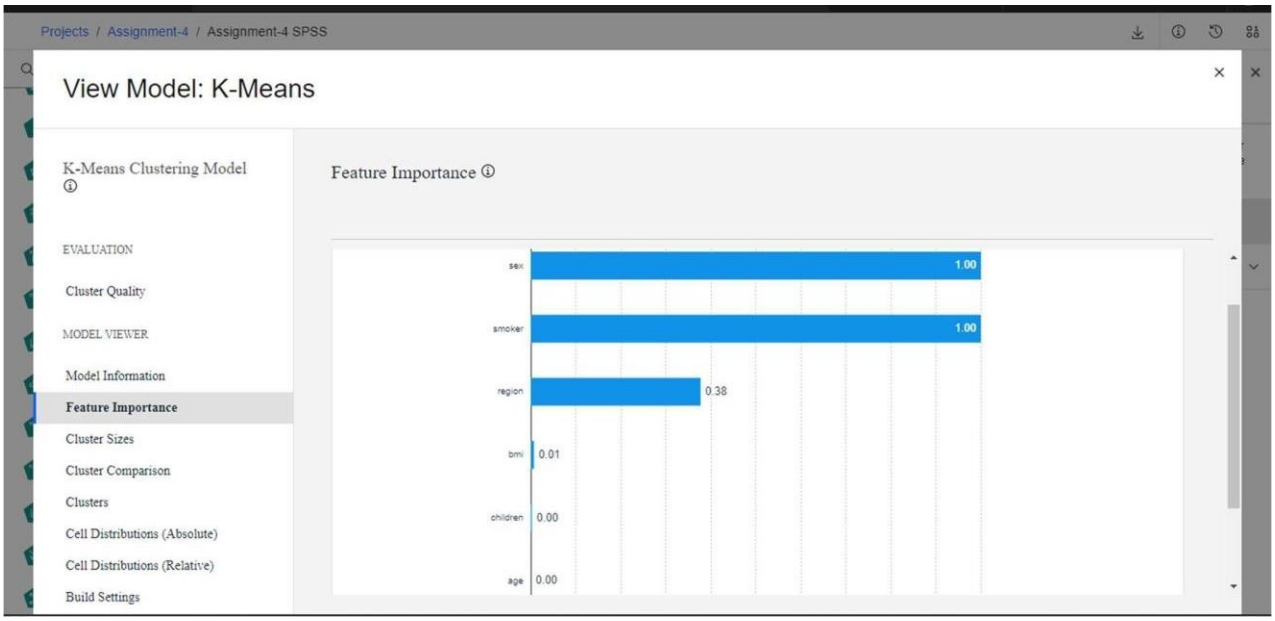



Projects / Assignment-4 / Assignment-4 SPSS
View Model: K-Means
K-Means Clustering Model
EVALUATION
Cluster Quality
MODEL VIEWER
Model Information
Feature Importance
Cluster Sizes
Cluster Comparison
Clusters
Cell Distributions (Absolute)
Cell Distributions (Relative)
Build Settings
Cluster Quality
-1.0
-0.5
0.0
0.5
1.0
Silhouette Measure of Cohesion and Separation
Cluster Quality Parameters
Overall Clustering Quality (Avg. Silhouette) 0.260
Total Within Clusters Sum of Squares 0.145
Average Within Cluster Sum of Squares 0.029
Average SSB (Between ss) 0.073

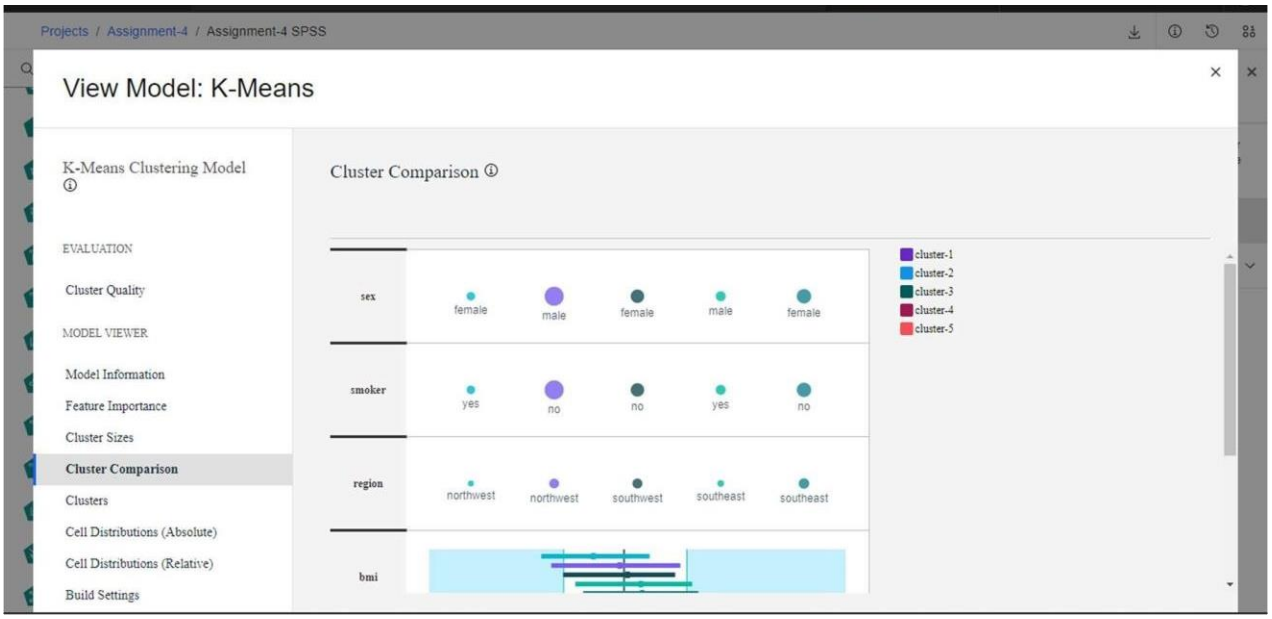

Projects / Assignment-4 / Assignment-4 SPSS
View Model: K-Means
K-Means Clustering Model
EVALUATION
Cluster Quality
MODEL VIEWER
Model Information
Feature Importance
Cluster Sizes
Cluster Comparison
Clusters
Cell Distributions (Absolute)
Cell Distributions (Relative)
Build Settings
Model Information
Algorithm K-Means
Model Class Center Based
Number of Features 6
Distance Measure Euclidean
Number of Clusters 5
Cluster 1 94 (8.79%)

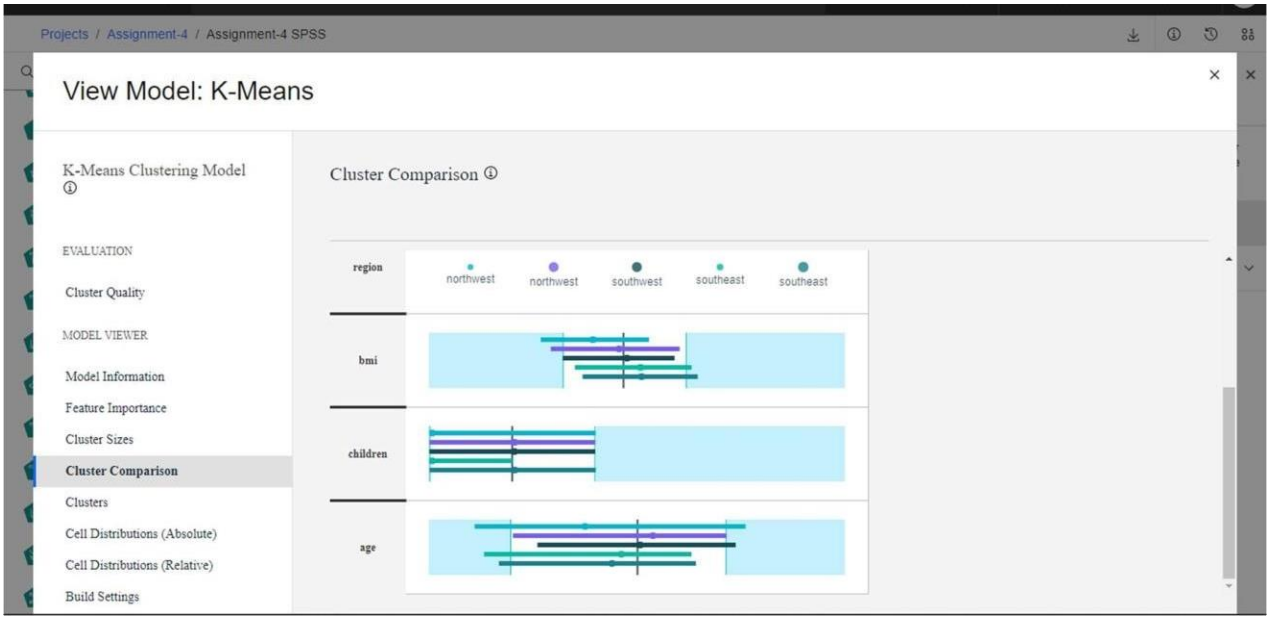


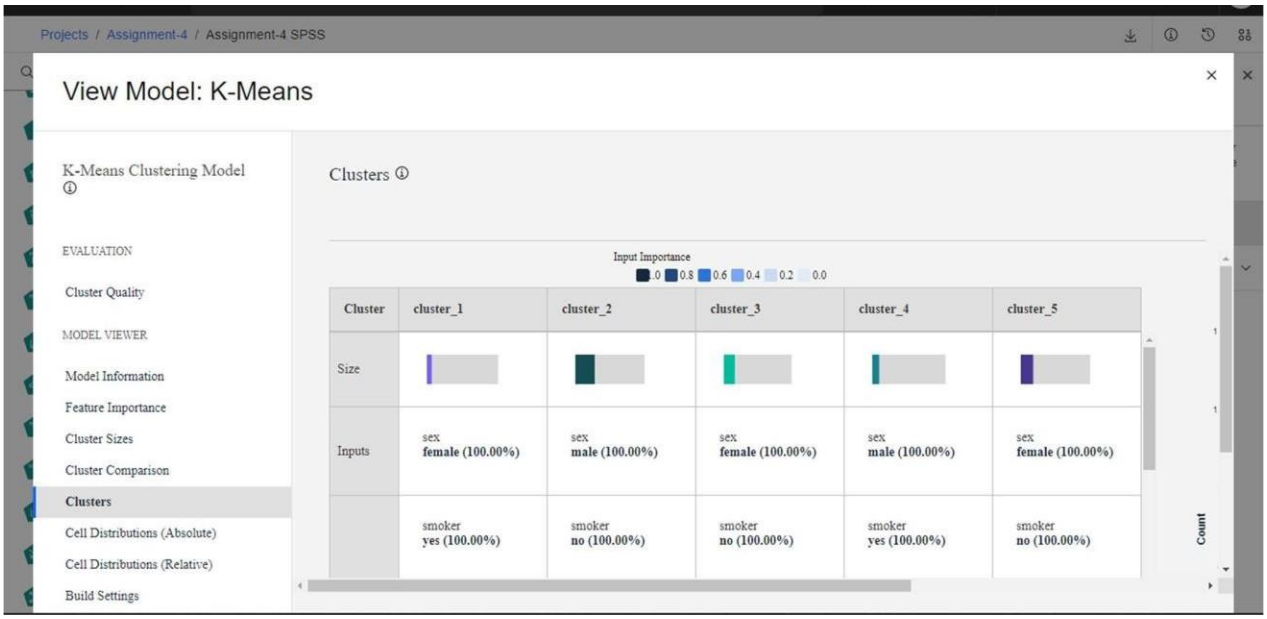
Number of Clusters		5
Number of instances in each cluster	Cluster 1	94 (8.79%)
	Cluster 2	411 (38.45%)
	Cluster 3	220 (20.58%)
	Cluster 4	130 (12.16%)
	Cluster 5	214 (20.02%)
Ratio of sizes (Largest to smallest)		4.372

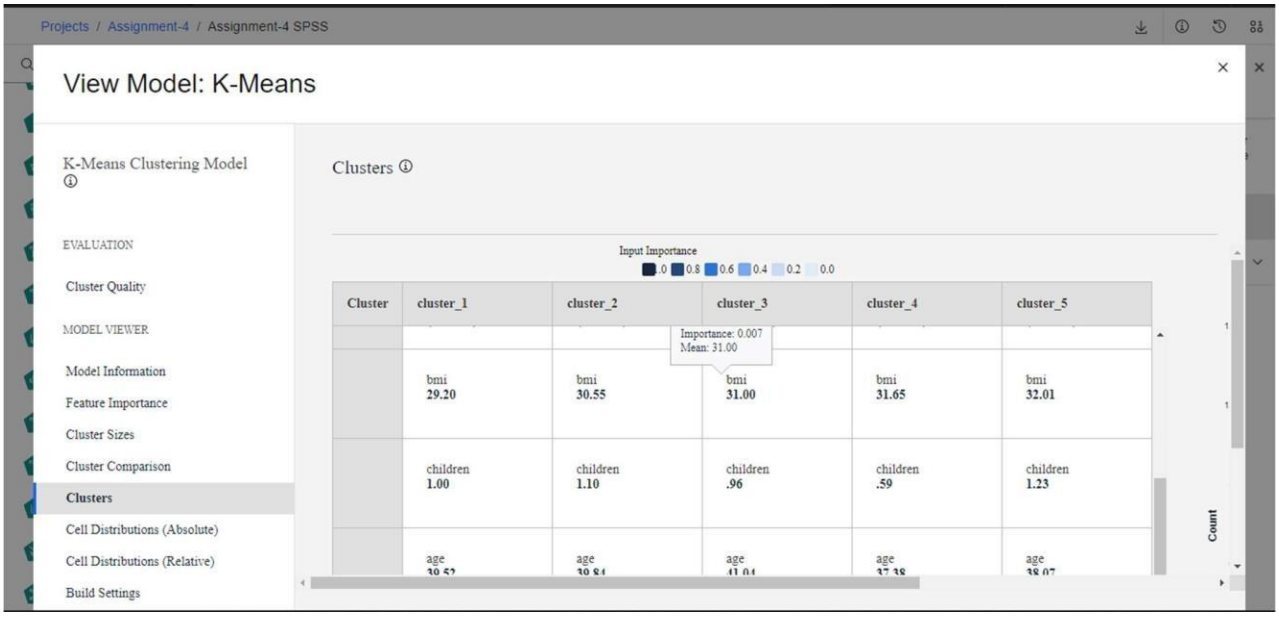


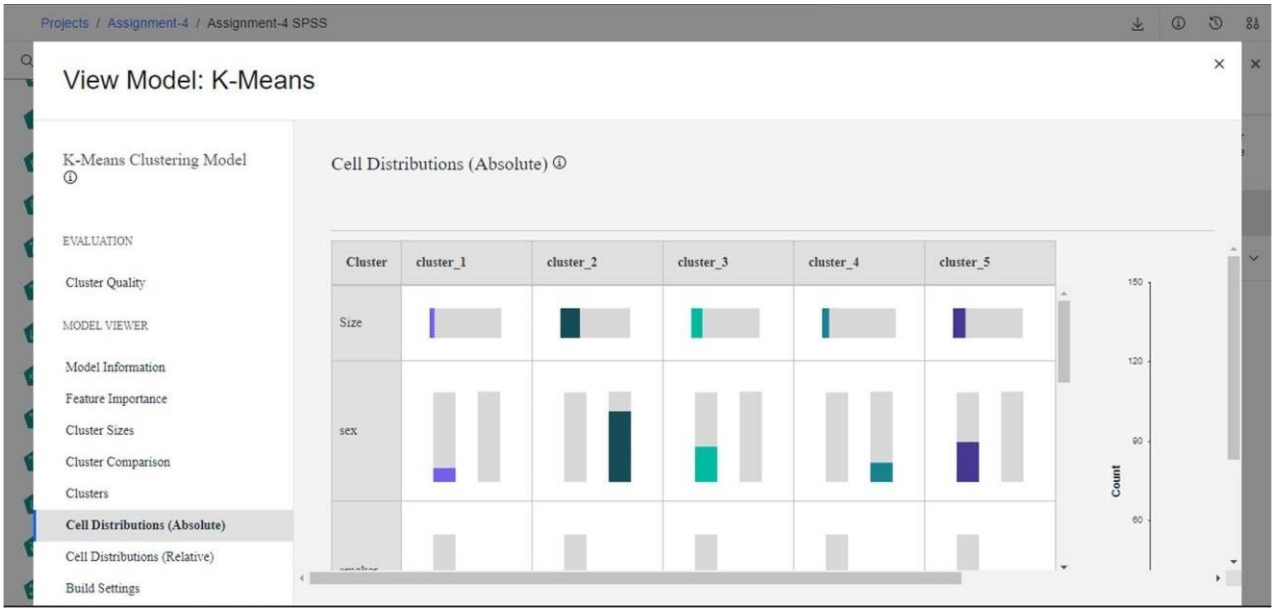

Projects / Assignment-4 / Assignment-4 SPSS
View Model: K-Means
K-Means Clustering Model
EVALUATION
Cluster Quality
MODEL VIEWER
Model Information
Feature Importance
Cluster Sizes
Cluster Comparison
Clusters
Cell Distributions (Absolute)
Cell Distributions (Relative)
Build Settings
Cluster Sizes
1
2
3
4
5

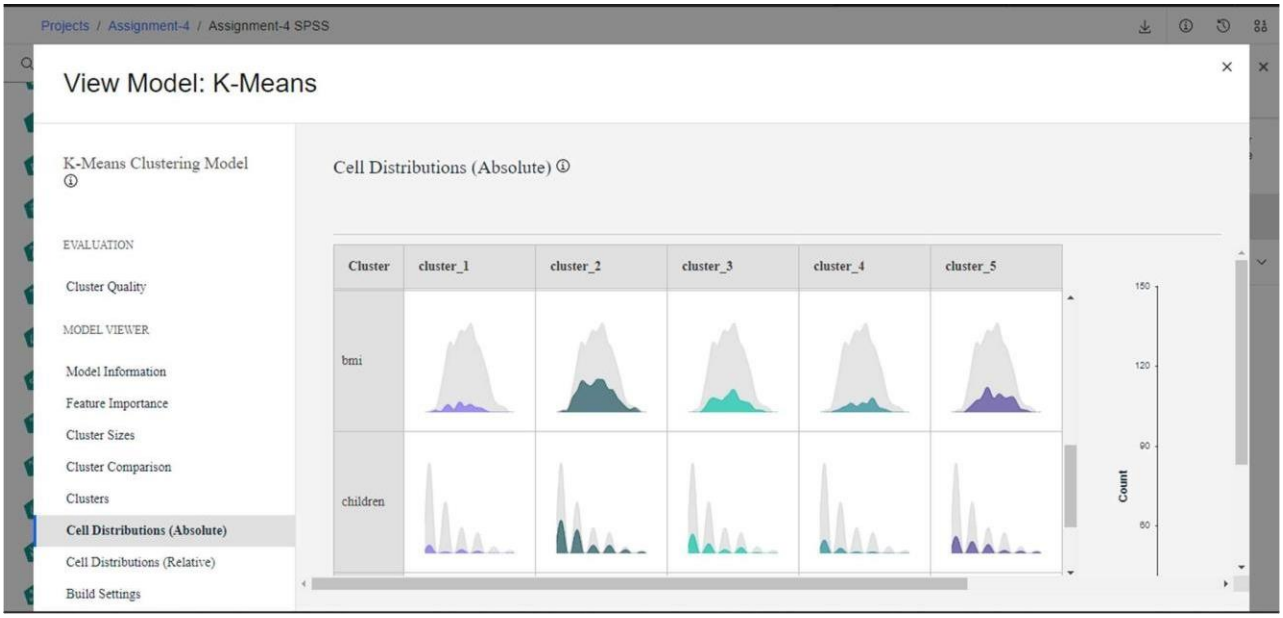

Projects / Assignment-4 / Assignment-4 SPSS
View Model: K-Means
K-Means Clustering Model
EVALUATION
Cluster Quality
MODEL VIEWER
Model Information
Feature Importance
Cluster Sizes
Cluster Comparison
Clusters
Cell Distributions (Absolute)
Cell Distributions (Relative)
Build Settings
Cluster Comparison
sex
female
male
female
male
female
smoker
yes
no
no
yes
no
region
northwest
northwest
southwest
southeast
southeast
bmi
cluster-1
cluster-2
cluster-3
cluster-4
cluster-5

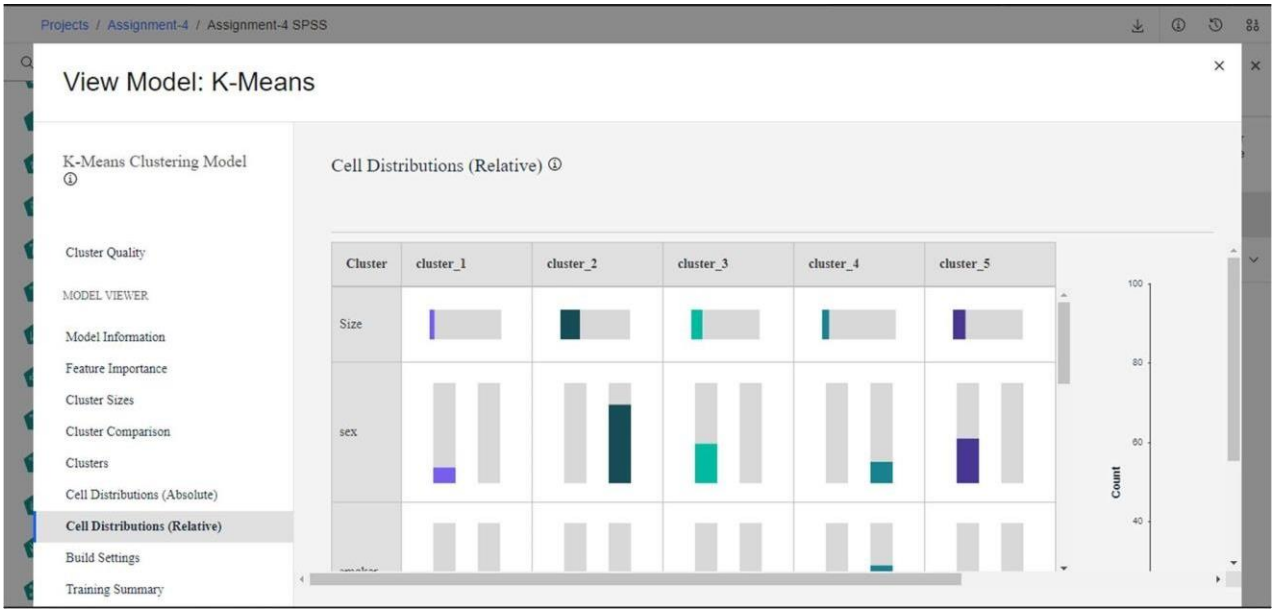

Projects / Assignment-4 / Assignment-4 SPSS
View Model: K-Means
K-Means Clustering Model
EVALUATION
Cluster Quality
MODEL VIEWER
Model Information
Feature Importance
Cluster Sizes
Cluster Comparison
Clusters
Cell Distributions (Absolute)
Cell Distributions (Relative)
Build Settings
Cluster Comparison
region
northwest
northwest
southwest
southeast
southeast
bmi
children
age


Projects / Assignment-4 / Assignment-4 SPSS
View Model: K-Means
K-Means Clustering Model
EVALUATION
Cluster Quality
MODEL VIEWER
Model Information
Feature Importance
Cluster Sizes
Cluster Comparison
Clusters
Cell Distributions (Absolute)
Cell Distributions (Relative)
Build Settings
Clusters
Input Importance
1.0 0.8 0.6 0.4 0.2 0.0
Cluster
cluster_1
cluster_2
cluster_3
cluster_4
cluster_5
Size
Inputs
sex female (100.00%)
sex male (100.00%)
sex female (100.00%)
sex male (100.00%)
sex female (100.00%)
smoker yes (100.00%)
smoker no (100.00%)
smoker no (100.00%)
smoker yes (100.00%)
smoker no (100.00%)
Count

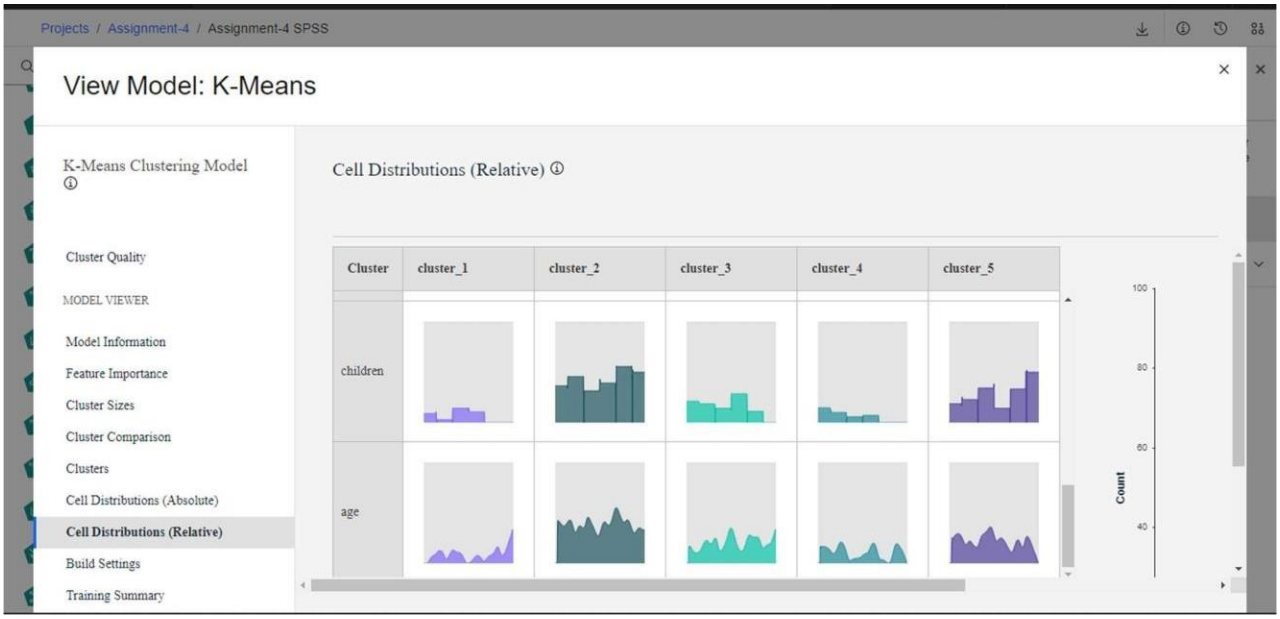

Projects / Assignment-4 / Assignment-4 SPSS
View Model: K-Means
K-Means Clustering Model
EVALUATION
Cluster Quality
MODEL VIEWER
Model Information
Feature Importance
Cluster Sizes
Cluster Comparison
Clusters
Cell Distributions (Absolute)
Cell Distributions (Relative)
Build Settings
Clusters
Input Importance
1.0 0.8 0.6 0.4 0.2 0.0
Cluster
cluster_1
cluster_2
cluster_3
cluster_4
cluster_5
Importance: 0.007
Mean: 31.00
bmi 29.20
bmi 30.55
bmi 31.00
bmi 31.65
bmi 32.01
children 1.00
children 1.10
children .96
children .59
children 1.23
age
age
age
age
age
Count

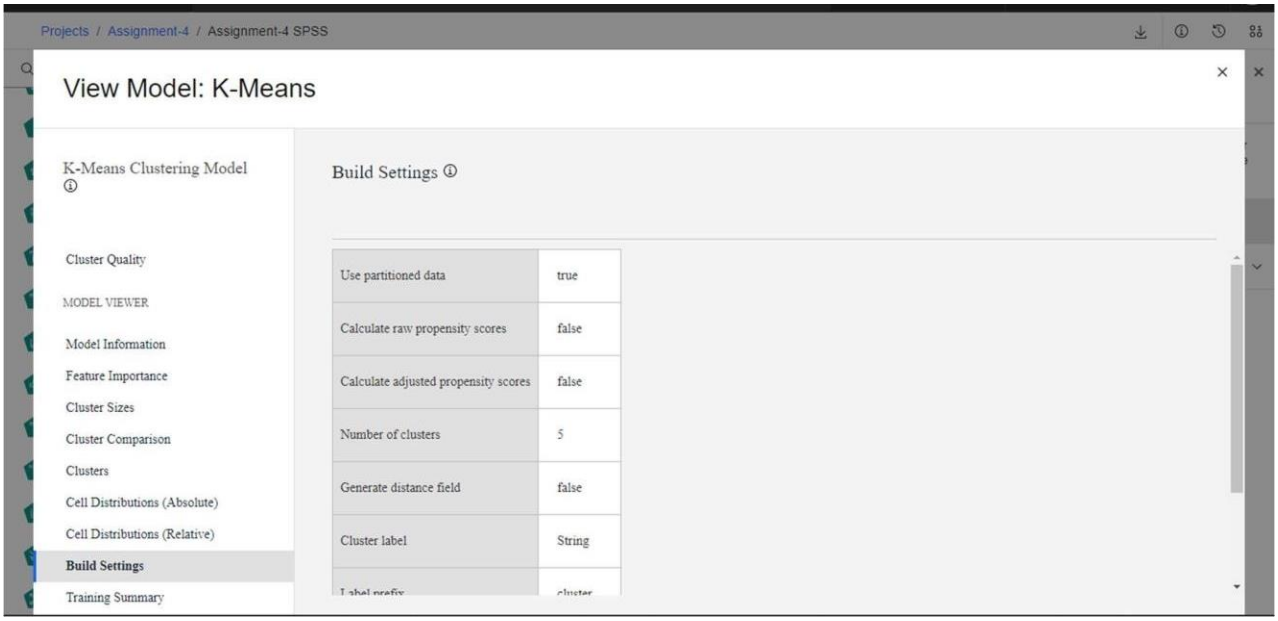

Projects / Assignment-4 / Assignment-4 SPSS
View Model: K-Means
K-Means Clustering Model
EVALUATION
Cluster Quality
MODEL VIEWER
Model Information
Feature Importance
Cluster Sizes
Cluster Comparison
Clusters
Cell Distributions (Absolute)
Cell Distributions (Relative)
Build Settings
Cell Distributions (Absolute)
Cluster
cluster_1
cluster_2
cluster_3
cluster_4
cluster_5
Size
sex
Count
150
120
90
60

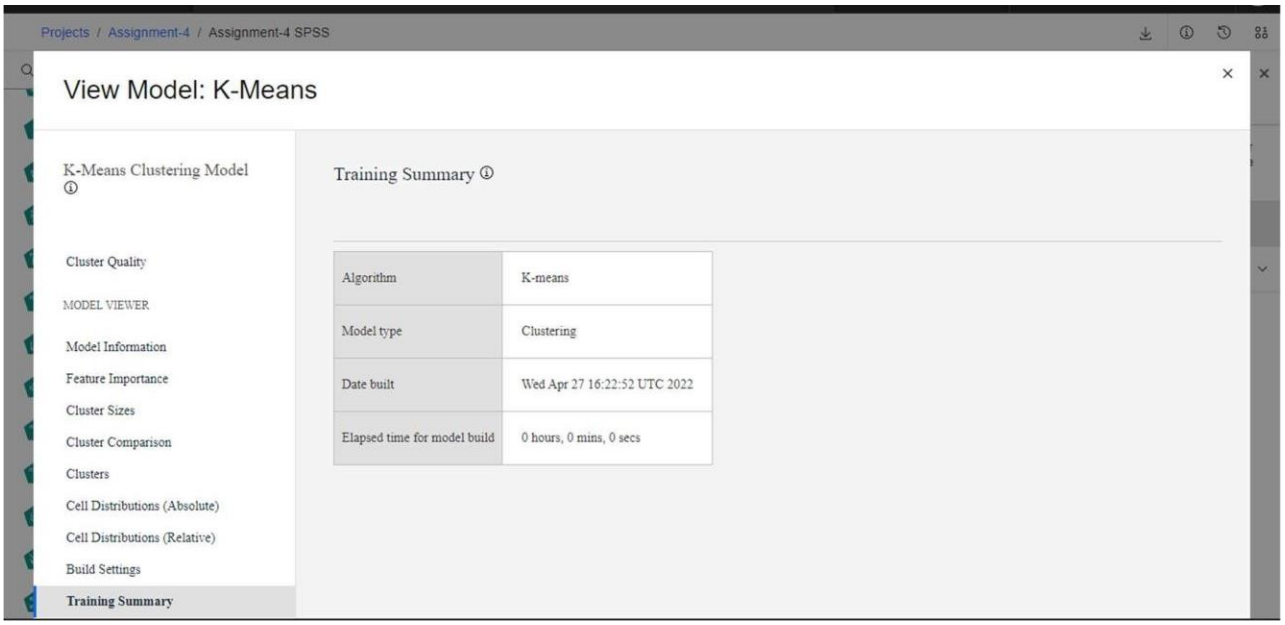

Projects / Assignment-4 / Assignment-4 SPSS
View Model: K-Means
K-Means Clustering Model
EVALUATION
Cluster Quality
MODEL VIEWER
Model Information
Feature Importance
Cluster Sizes
Cluster Comparison
Clusters
Cell Distributions (Absolute)
Cell Distributions (Relative)
Build Settings
Cell Distributions (Absolute)
Cluster
cluster_1
cluster_2
cluster_3
cluster_4
cluster_5
bmi
children
Count
150
120
90
60


Projects / Assignment-4 / Assignment-4 SPSS
View Model: K-Means
K-Means Clustering Model
Cluster Quality
MODEL VIEWER
Model Information
Feature Importance
Cluster Sizes
Cluster Comparison
Clusters
Cell Distributions (Absolute)
Cell Distributions (Relative)
Build Settings
Training Summary
Cell Distributions (Relative)
Cluster
cluster_1
cluster_2
cluster_3
cluster_4
cluster_5
Size
sex
Count
100
80
60
40

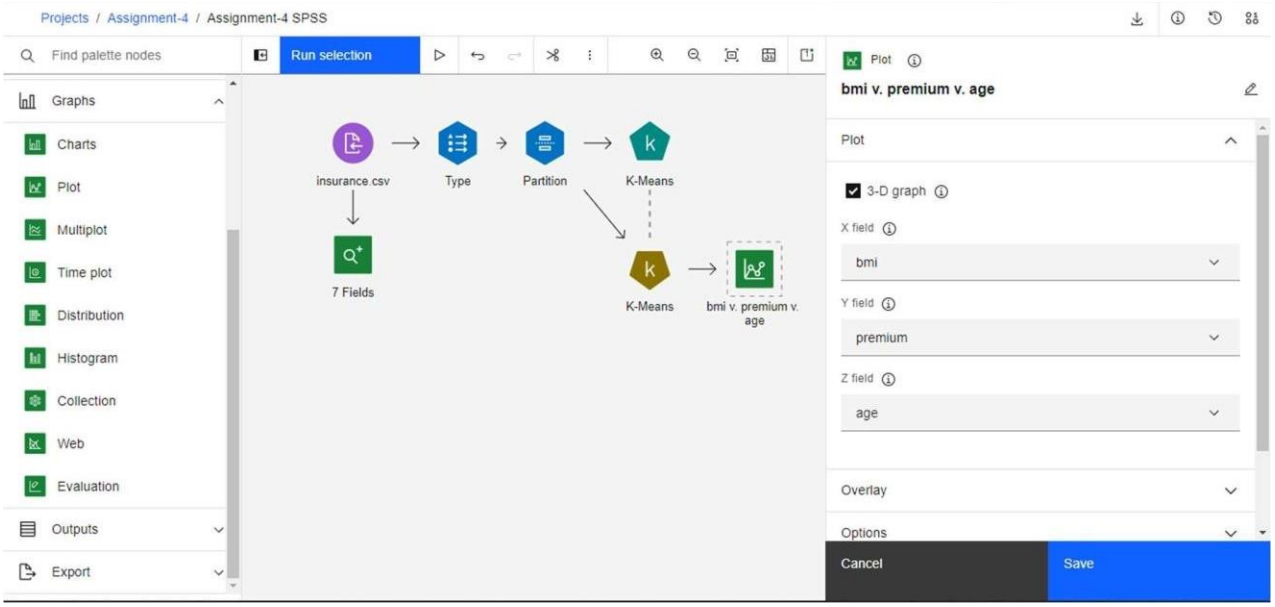

Projects / Assignment-4 / Assignment-4 SPSS
View Model: K-Means
K-Means Clustering Model
Cluster Quality
MODEL VIEWER
Model Information
Feature Importance
Cluster Sizes
Cluster Comparison
Clusters
Cell Distributions (Absolute)
Cell Distributions (Relative)
Build Settings
Training Summary
Cell Distributions (Relative)
Cluster
cluster_1
cluster_2
cluster_3
cluster_4
cluster_5
smoker
region
Count
100
80
60
40


Projects / Assignment-4 / Assignment-4 SPSS
View Model: K-Means
K-Means Clustering Model
Cluster Quality
MODEL VIEWER
Model Information
Feature Importance
Cluster Sizes
Cluster Comparison
Clusters
Cell Distributions (Absolute)
Cell Distributions (Relative)
Build Settings
Training Summary
Cell Distributions (Relative)
Cluster
cluster_1
cluster_2
cluster_3
cluster_4
cluster_5
children
age
Count
100
80
60
40

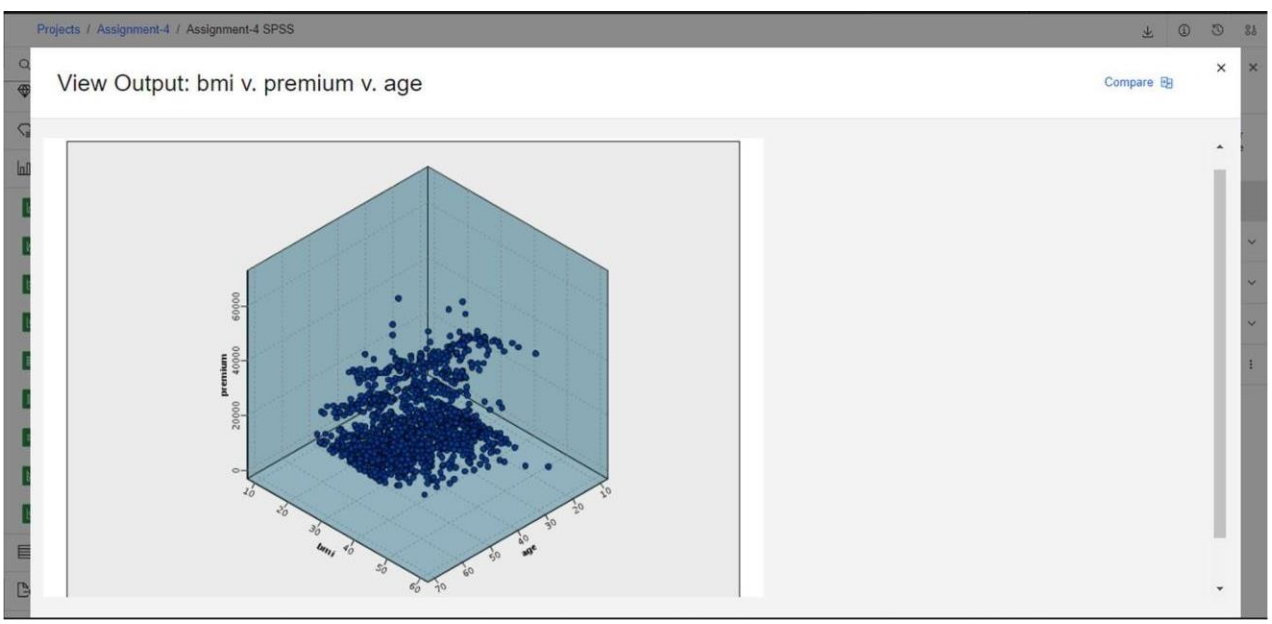

Projects / Assignment-4 / Assignment-4 SPSS
View Model: K-Means
K-Means Clustering Model
Cluster Quality
MODEL VIEWER
Model Information
Feature Importance
Cluster Sizes
Cluster Comparison
Clusters
Cell Distributions (Absolute)
Cell Distributions (Relative)
Build Settings
Training Summary
Build Settings
Use partitioned data
true
Calculate raw propensity scores
false
Calculate adjusted propensity scores
false
Number of clusters
5
Generate distance field
false
Cluster label
String


Projects / Assignment-4 / Assignment-4 SPSS
View Model: K-Means
K-Means Clustering Model
Cluster Quality
MODEL VIEWER
Model Information
Feature Importance
Cluster Sizes
Cluster Comparison
Clusters
Cell Distributions (Absolute)
Cell Distributions (Relative)
Build Settings
Training Summary
Training Summary
Algorithm
K-means
Model type
Clustering
Date built
Wed Apr 27 16:22:52 UTC 2022
Elapsed time for model build
0 hours, 0 mins, 0 secs

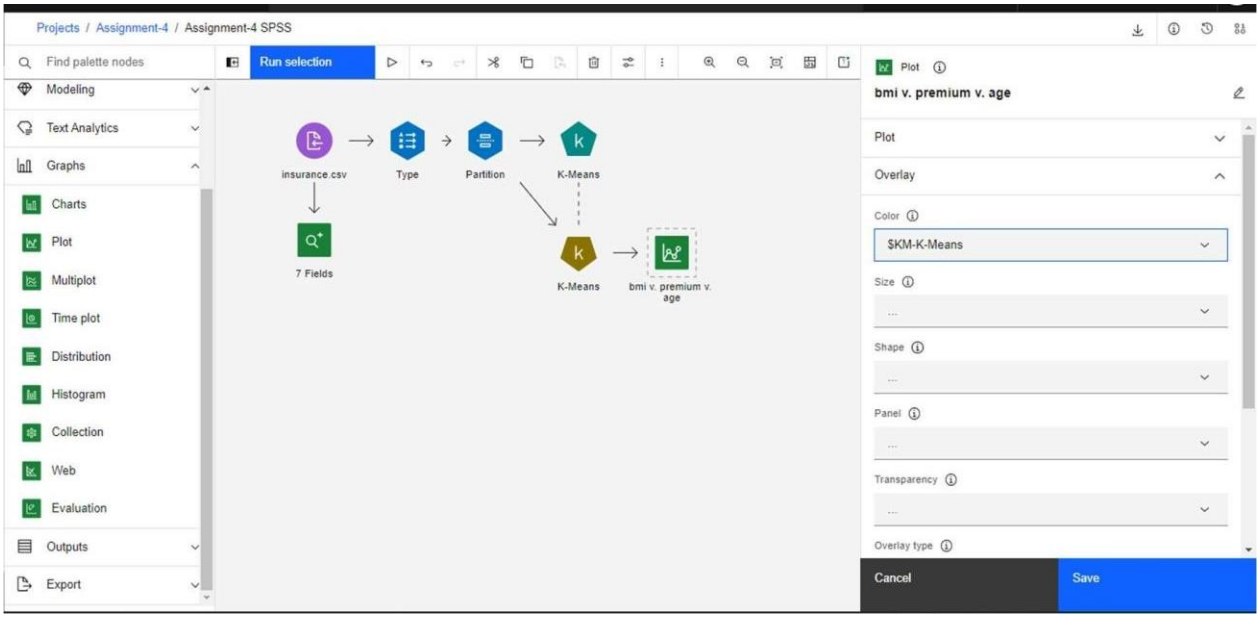
Creating a Plot Node with BMI vs Premium vs Age:



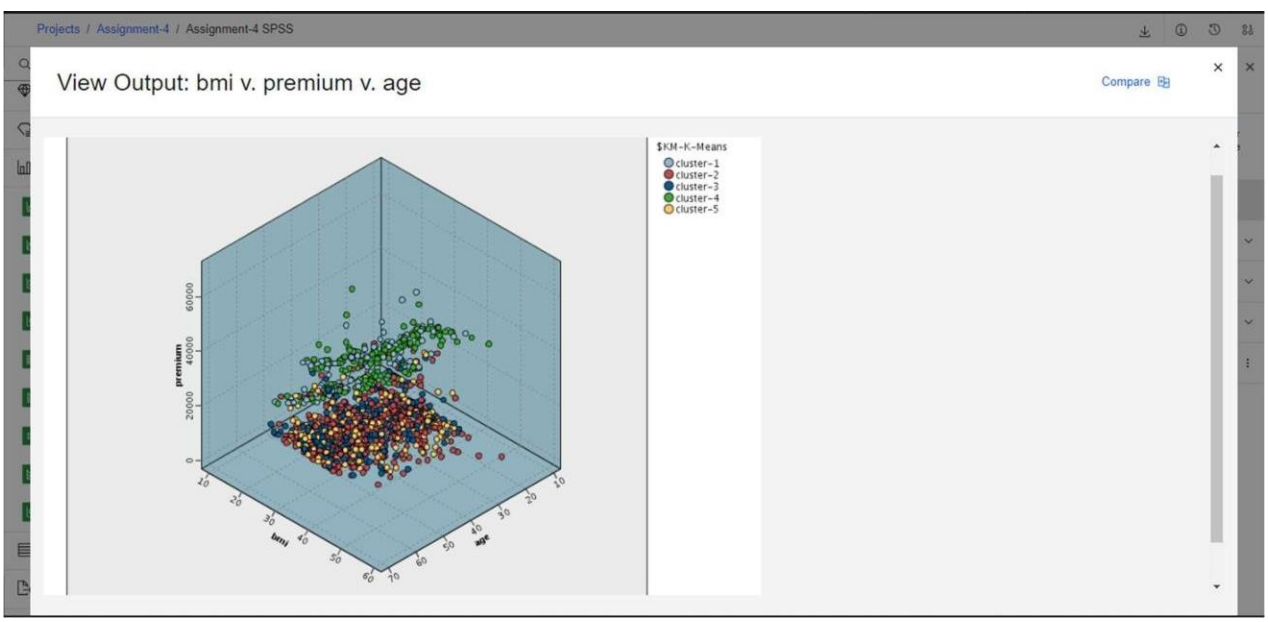
Output:



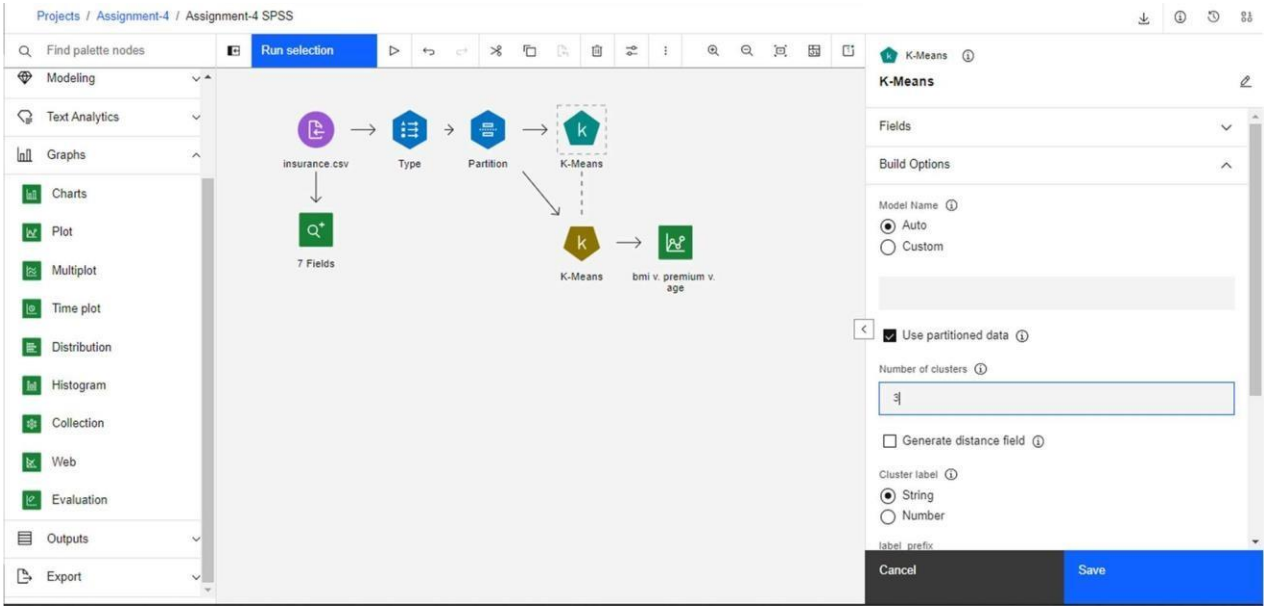
Assigning Colours:



Output:



Changing Number of clusters to 3:



Output: