

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

Creating the Project

New project

Define details

Name

Description

Choose project options

☐ Restrict who can be a collaborator ⓘ

Project includes integration with [Cloud Object Storage](#) for storing project assets.

Storage

Cancel

Create

Uploading Data:

The screenshot shows the 'Assets' tab in a project named 'Assignment-4'. The interface includes a search bar, a sidebar with 'Asset types' (Data, Data asset), and a main table of assets. A 'Data in this project' panel on the right shows a drop zone for file uploads.

Assets Table:

| Name | Last modified |
|----------------------|-------------------------------------|
| insurance.csv CSV | Now Naveen Sai Tamanampudi (You) |

Creating a Data Refinery:

The screenshot shows the 'Select data from project' dialog. It displays a list of assets under 'Assignment-4', with 'insurance.csv' selected. The 'Selected assets' panel on the right shows the details of the selected asset.

Selected assets:

- insurance.csv

Asset details:

- Asset name: insurance.csv
- Asset type: Data asset
- Size: 54 KB
- Last modified: 2022/04/27 15:54:41
- Created on: 2022/04/27 15:54:41

Viewing the output of refinery:

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

Steps

Use a code template to add a step

Data

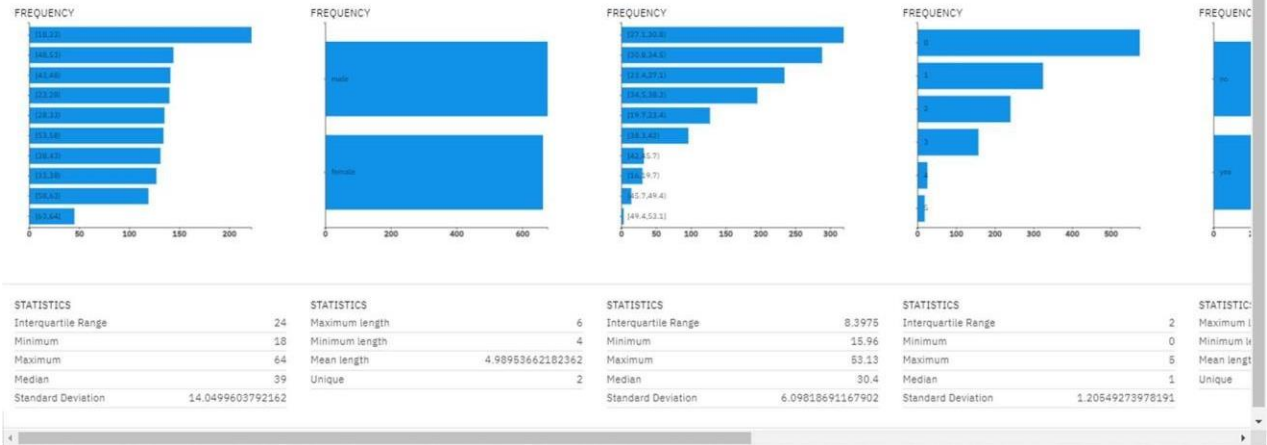
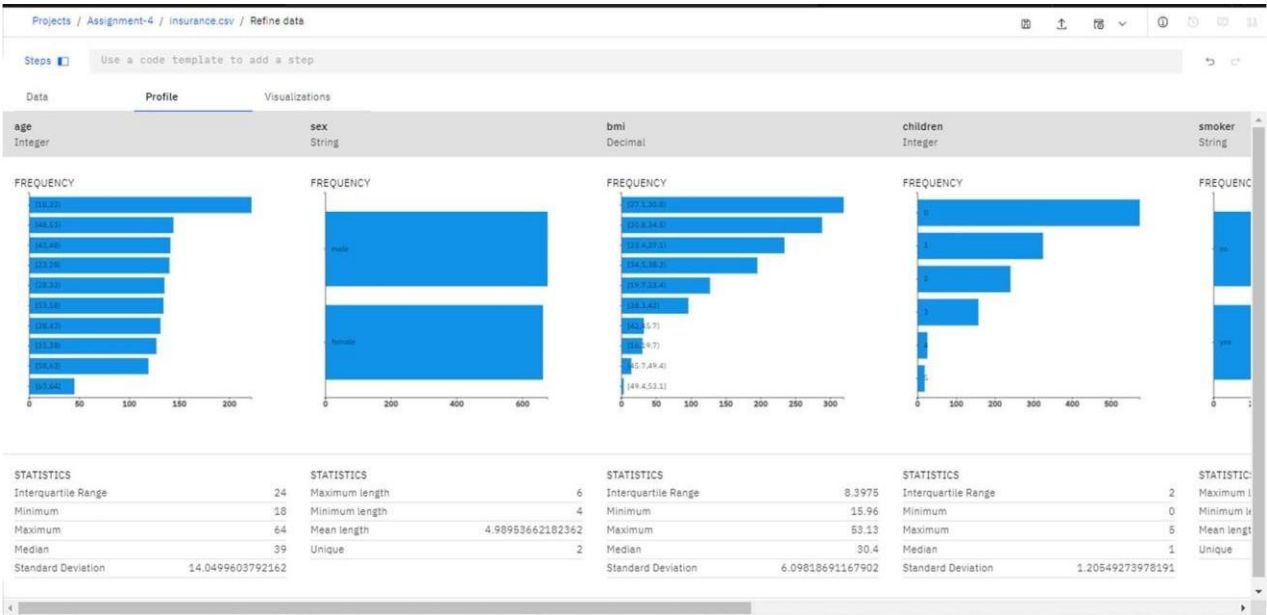
Profile

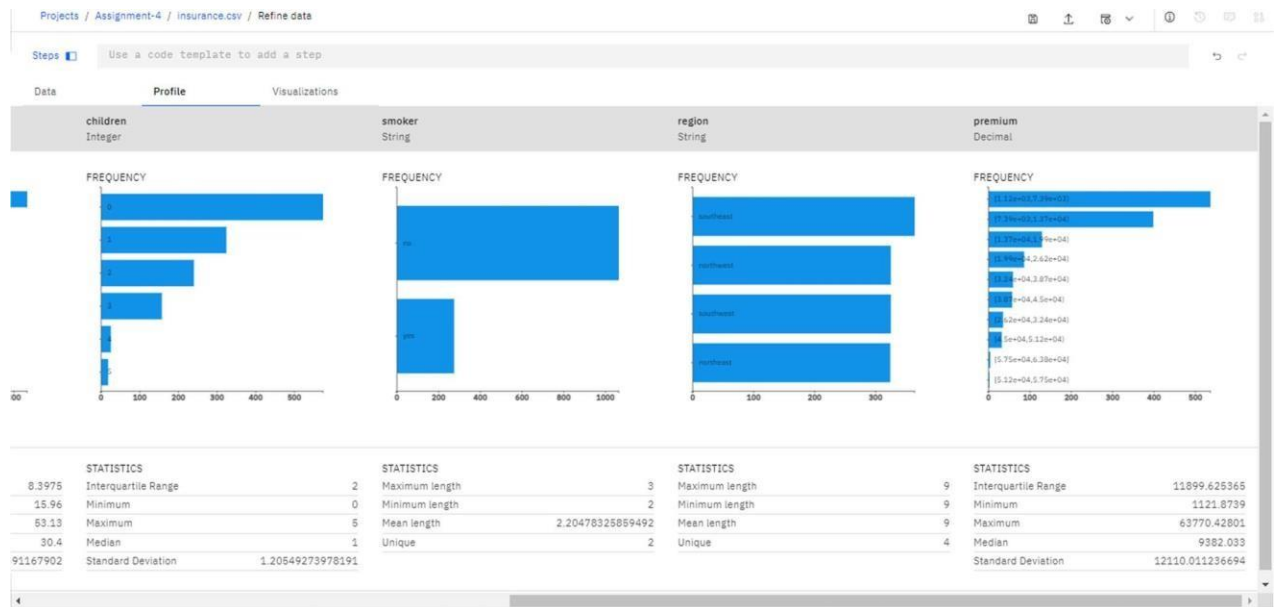
Visualizations

| | 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 |
| -- | 27 | male | 42.13 | 0 | yes | southeast | 39611.7577 |

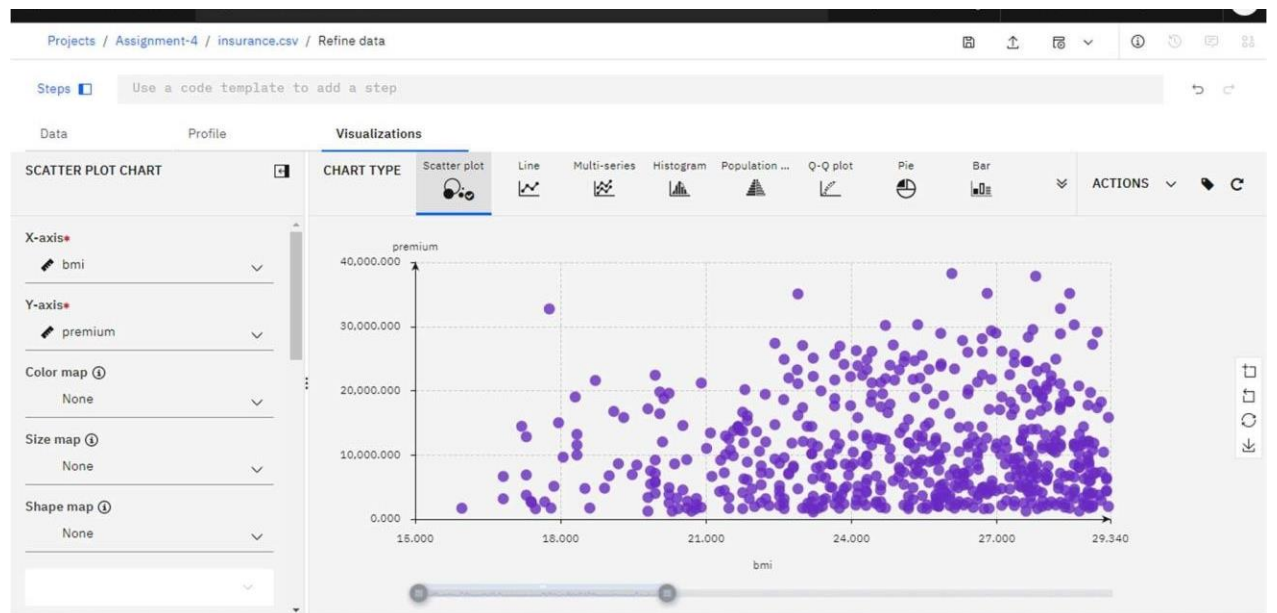
SOURCE FILE: insurance.csv FULL DATA SET: 1338 rows

Viewing Profile:

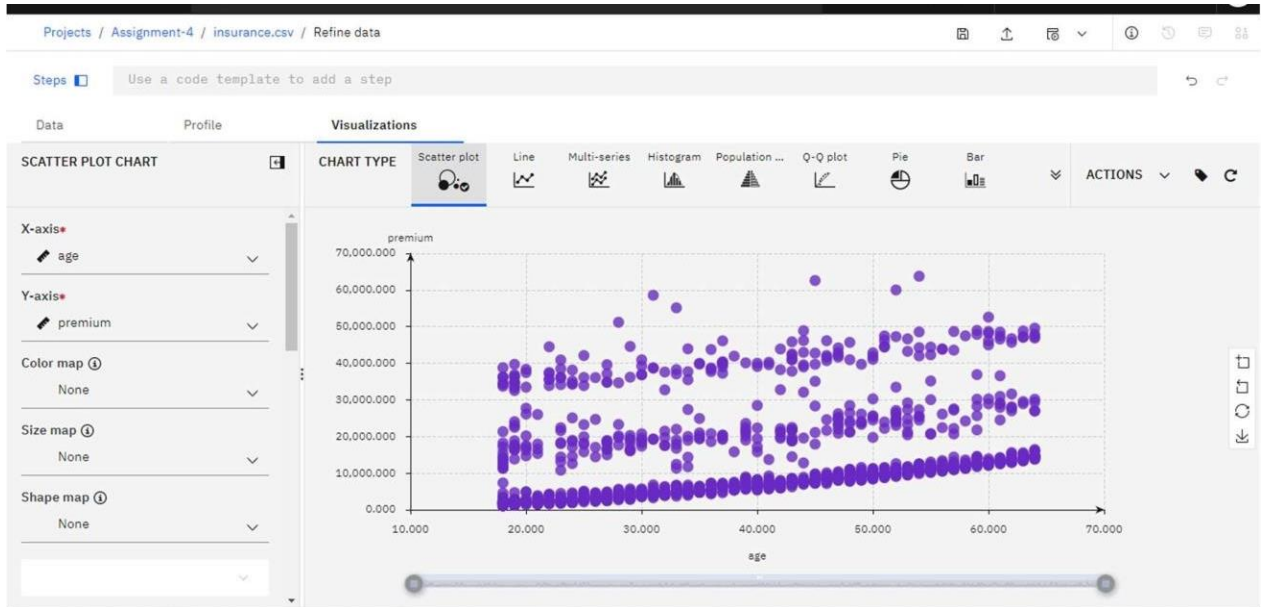




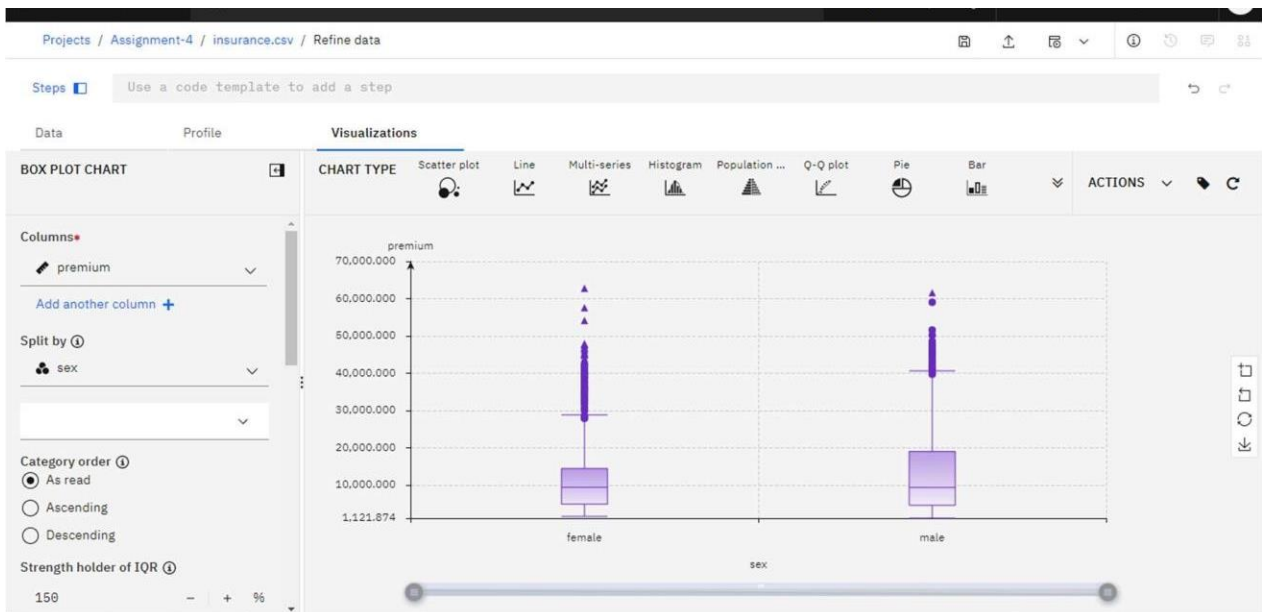
Viewing Scatterplot of BMI vs Premium:



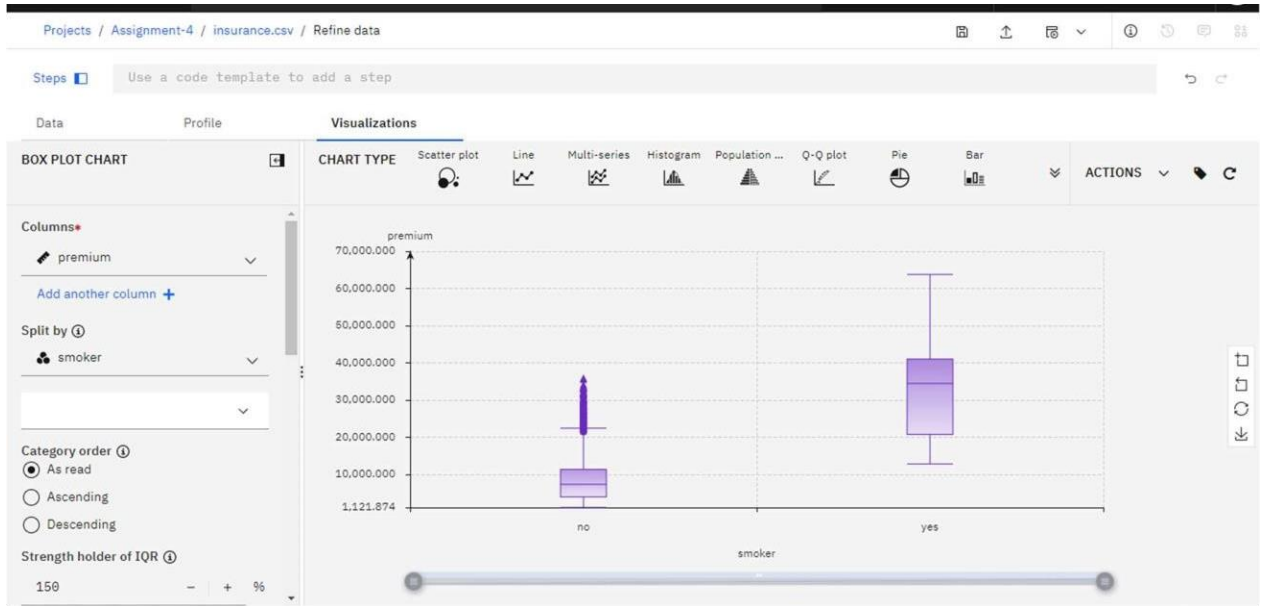
Viewing Scatterplot of Age vs Premium:



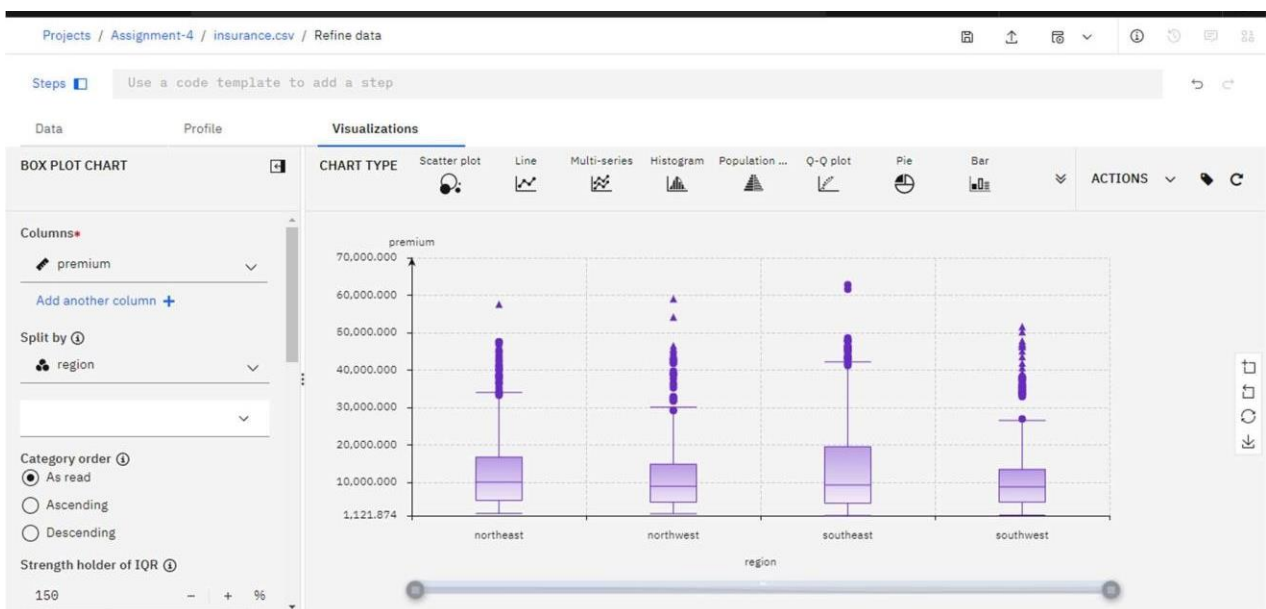
Viewing Boxplot of Sex vs Premium:



Viewing Boxplot of Smoker vs Premium:



Viewing Boxplot of Region vs Premium:



Creating an SPSS Modeler:

The screenshot shows a dialog box titled "Create an SPSS Modeler flow" within a project named "Assignment-4". The dialog is divided into three main sections: a left sidebar, a "Define details" section, and a "Define configuration" section.

Left Sidebar: Contains a "New asset" header and two options: "Gallery sample" and "Local file".

Define details: Includes a "Name" field with the value "Assignment-4 SPSS" and a "Description (Optional)" text area containing the placeholder text "What's the purpose of this SPSS Modeler flow?".

Define configuration: Includes an "Environment definition" dropdown menu set to "Default SPSS Modeler S (2 vCPU 8 GB RAM)". Below this, a note states: "To create additional runtime environments, view options in the Environments tab."

At the bottom right, there are two buttons: "Cancel" and "Create".

Uploading Dataset to Data Asset Node:

The screenshot shows a "Data Asset" selection dialog. The main area is divided into two panes: "Categories" and "Data assets".

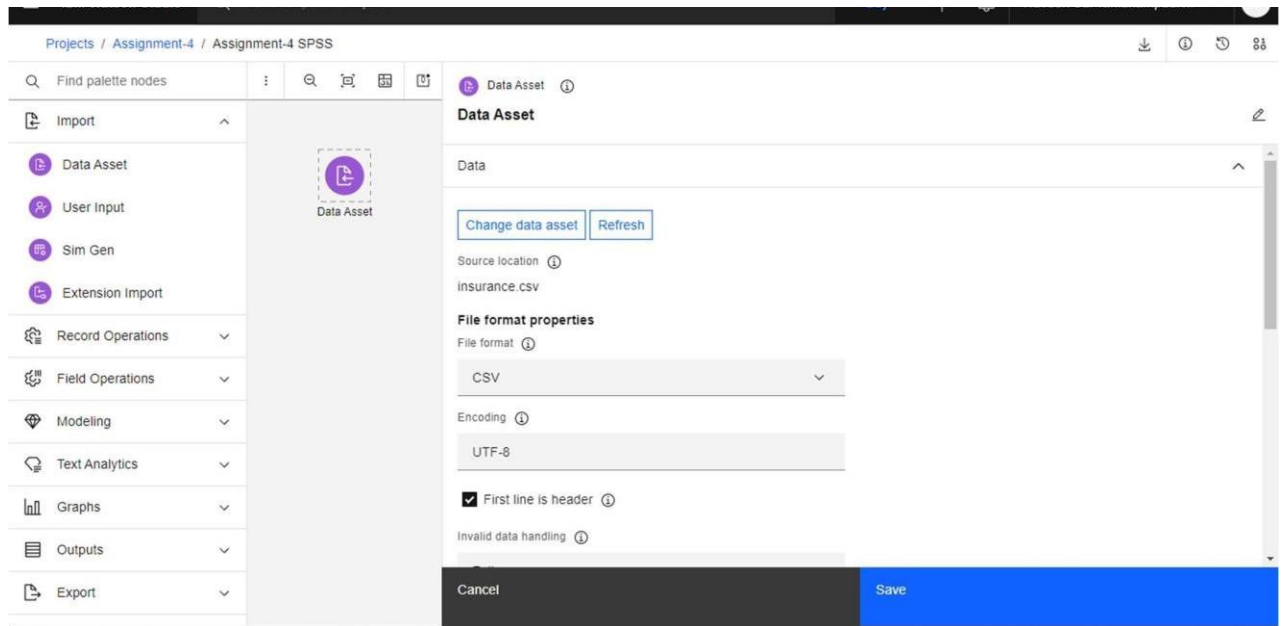
Categories Pane: Lists "Connection" and "Data asset". The "Data asset" category is selected, indicated by a blue bar.

Data assets Pane: Displays a list of data assets. The asset "insurance.csv" is highlighted.

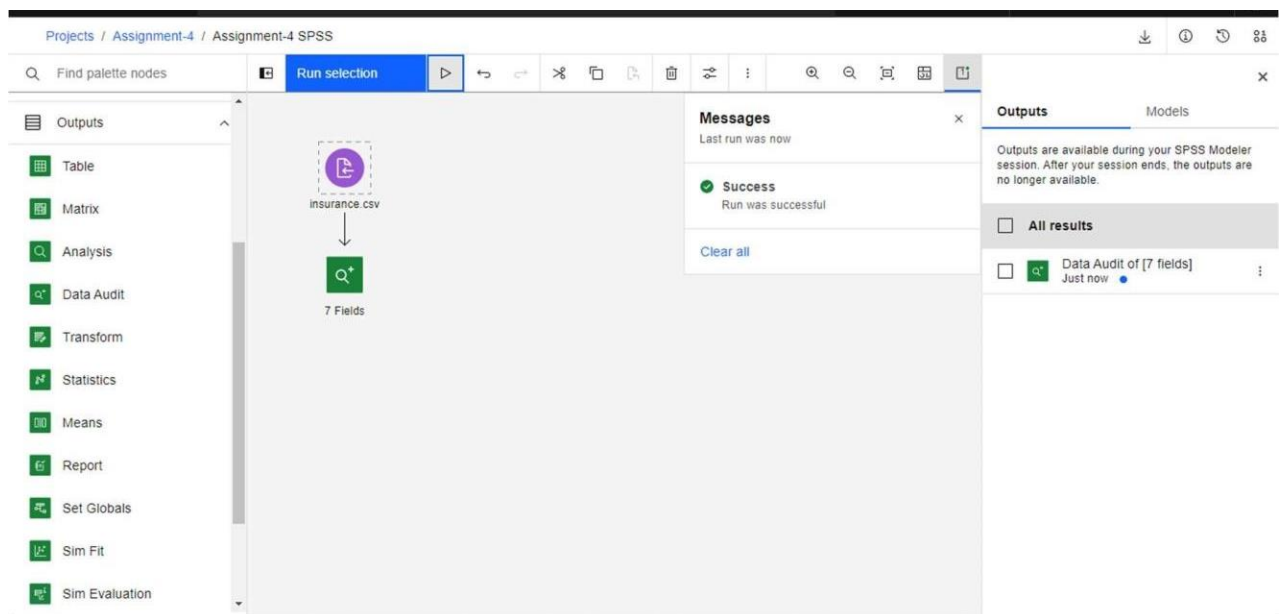
Selected assets Pane: Shows details for the selected "insurance.csv" asset. It includes a green checkmark icon and the following information:

- Asset name: insurance.csv
- Asset details:
- Asset type: Data asset
- Size: 54 KB
- Last modified: 2022/04/27 15:54:41
- Created on: 2022/04/27 15:54:41

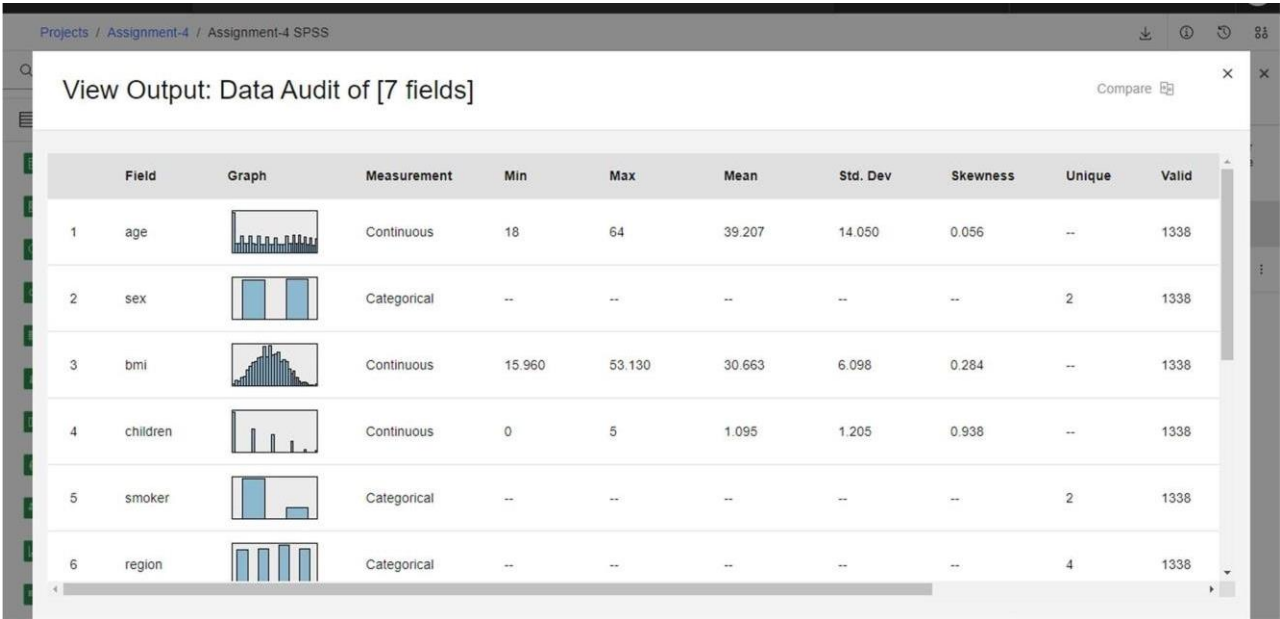
At the bottom right, there are two buttons: "Cancel" and "Select".



Creating Data Audit Node:




Output from Data Audit:



Projects / Assignment-4 / Assignment-4 SPSS

View Output: Data Audit of [7 fields]

Compare

| 7 | premium |  | Continuous | 1121.674 | 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:

The screenshot shows the Data Science Canvas interface. On the left, a palette of nodes is visible, including 'Import', 'Record Operations', 'Field Operations', 'Auto Data Prep', 'Type', 'Filter', 'Derive', 'Filler', 'Reclassify', 'Binning', 'RFM Analysis', and 'Ensemble'. The 'Type' node is selected and connected to the 'insurance.csv' data source. The 'Settings' panel for the 'Type' node is open, displaying a table of fields and their roles.

| Field | Measure | Role | Value mode | Values |
|-------------------------------------|-------------|--------|------------|--------|
| <input type="checkbox"/> # age | Continuous | Input | Read | |
| <input type="checkbox"/> sex | Categorical | Input | Read | |
| <input type="checkbox"/> bmi | Continuous | Input | Read | |
| <input type="checkbox"/> # children | Continuous | Input | Read | |
| <input type="checkbox"/> smoker | Categorical | Input | Read | |
| <input type="checkbox"/> region | Categorical | Input | Read | |
| <input type="checkbox"/> premium | Continuous | Target | Read | |

Buttons at the bottom: Cancel, Save.

Creating a Partition Node with 80:20 split:

The screenshot shows the Data Science Canvas interface. The 'Partition' node is selected and connected to the 'Type' node. The 'Settings' panel for the 'Partition' node is open, displaying configuration options for the data split.

Derived Field Name: Partition

Training Partition(%): 80

Testing Partition(%): 20

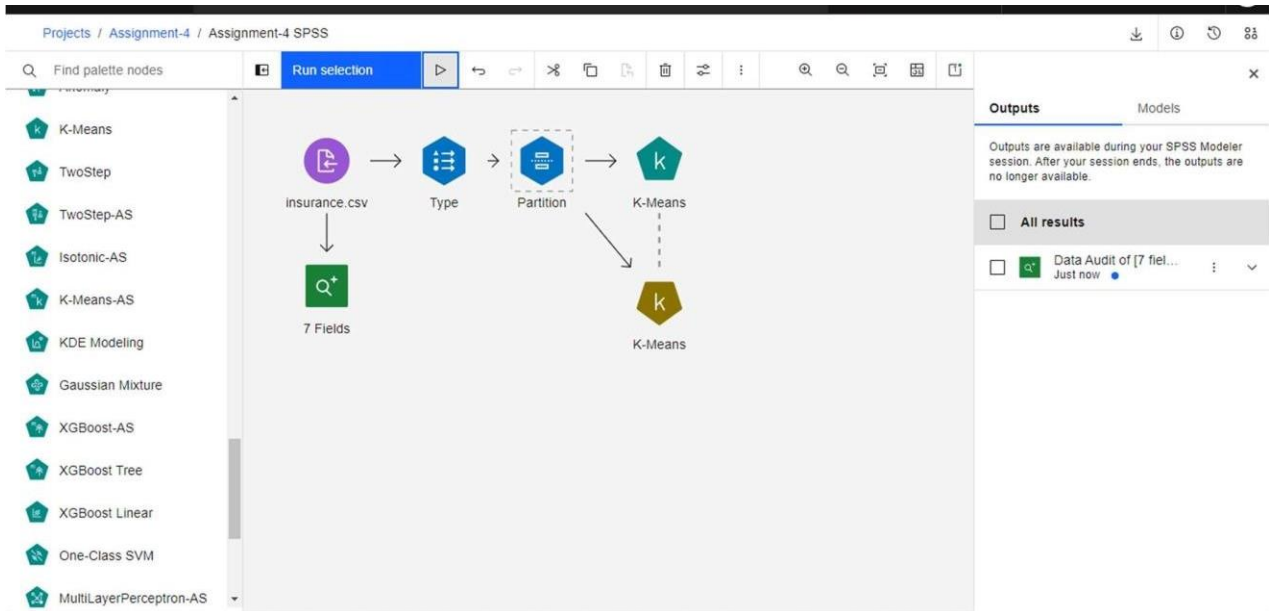
☐ Create validation partition

☒ Repeatable partition assignment

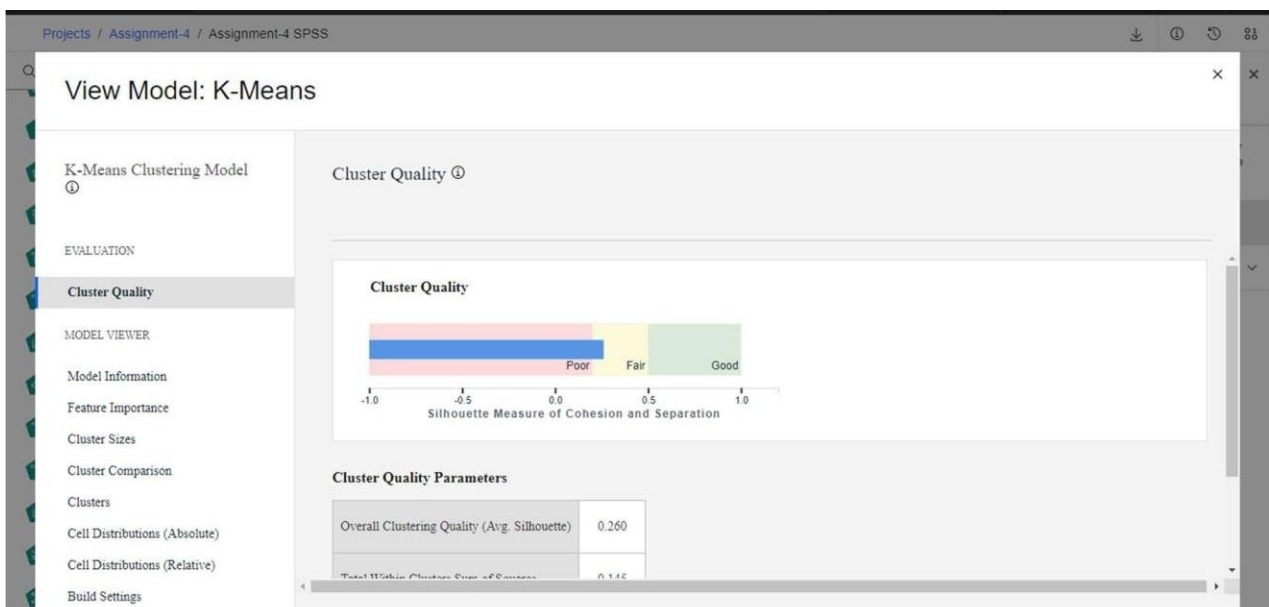
Seed: 1234567

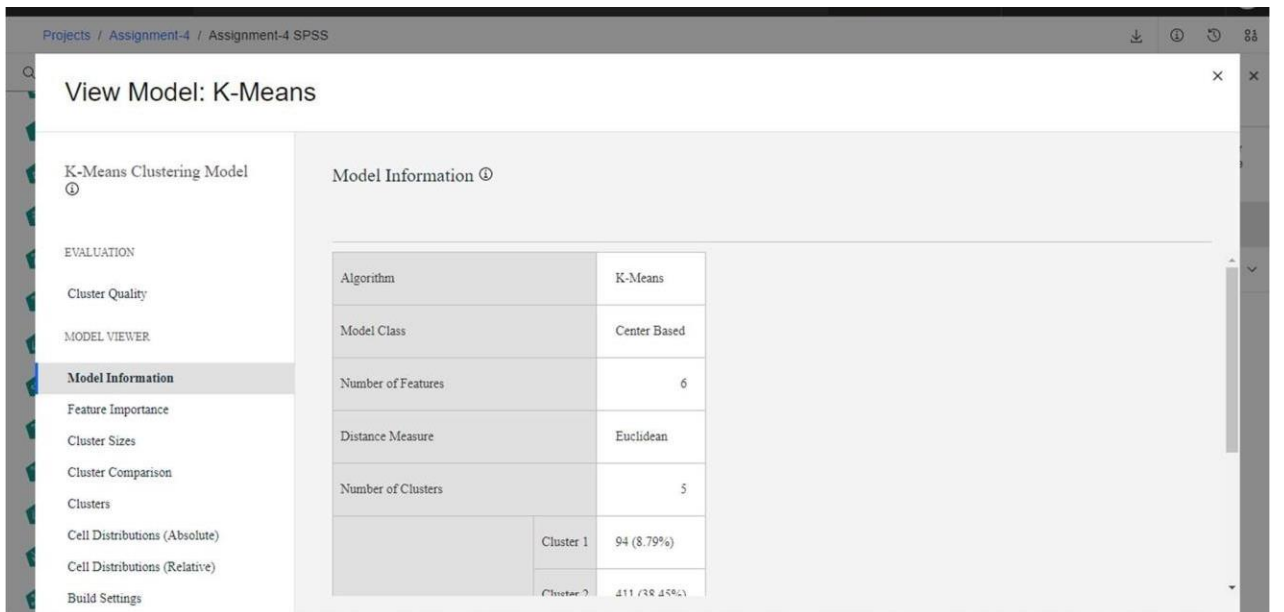
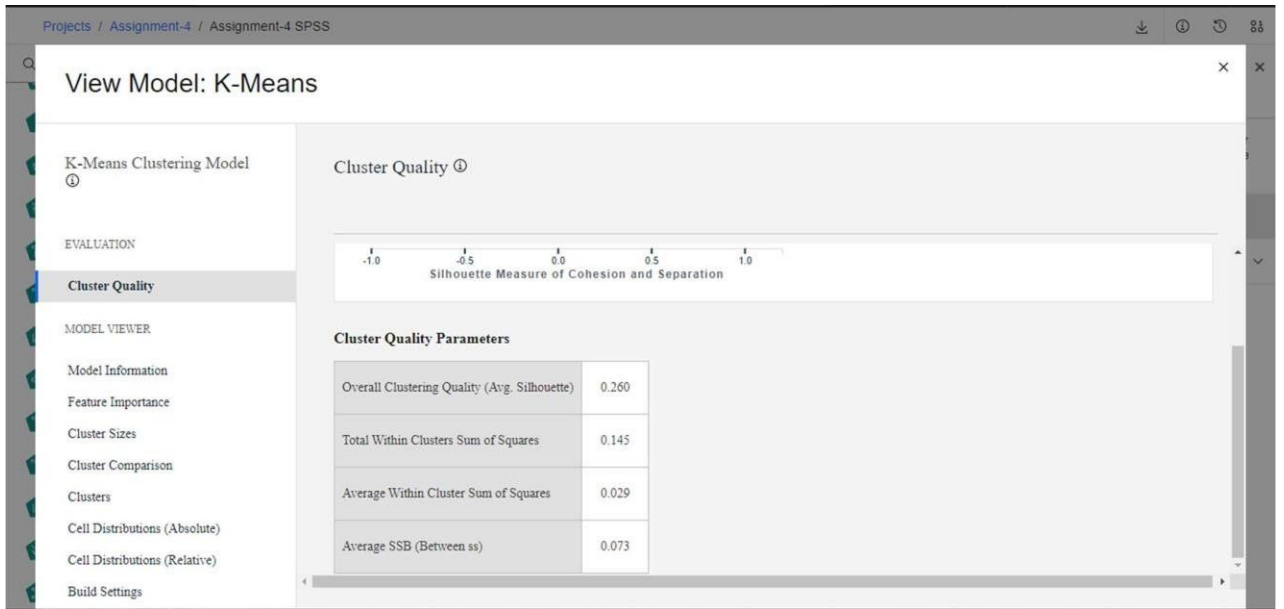
Buttons at the bottom: Cancel, Save.

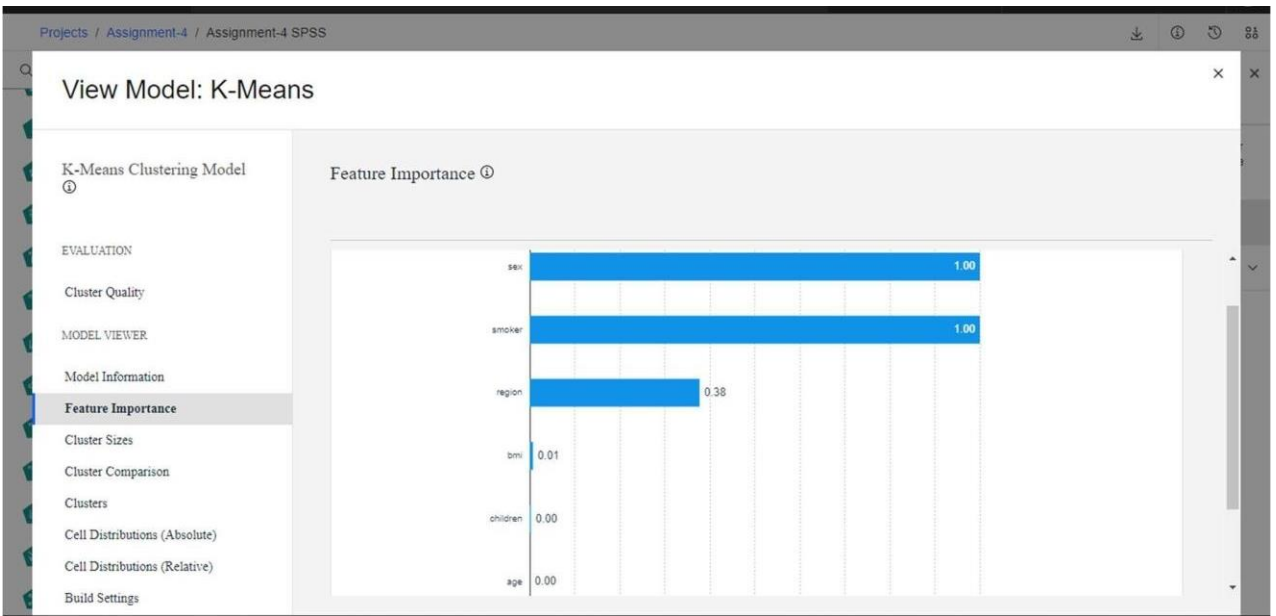
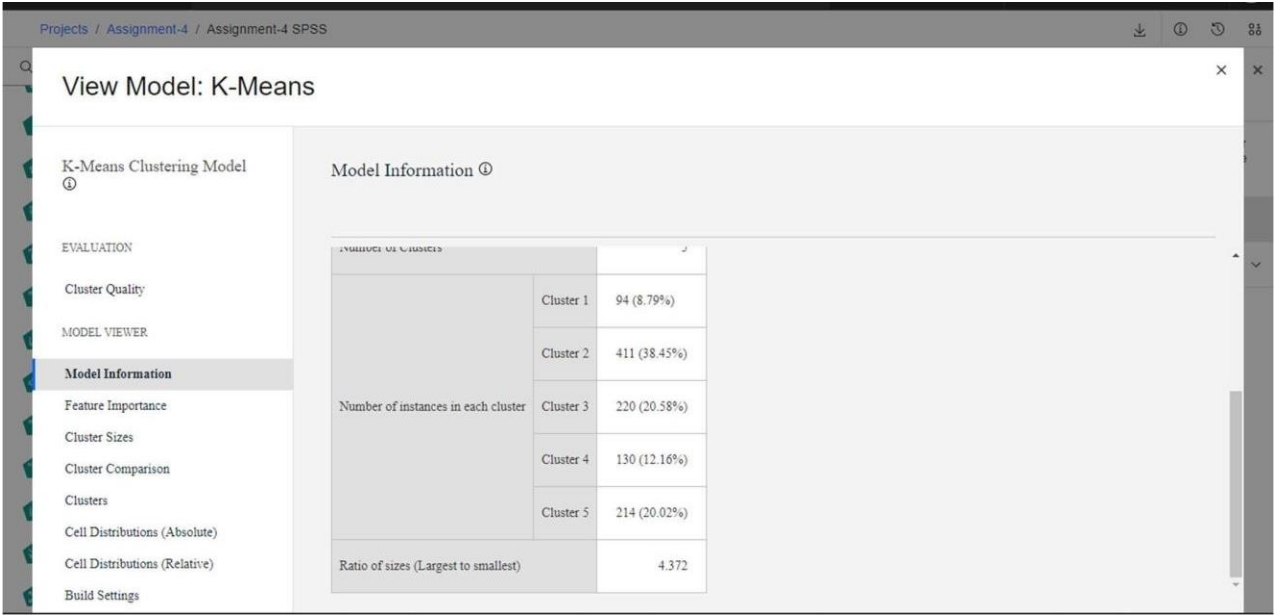
Creating a K-Means Clustering Model:

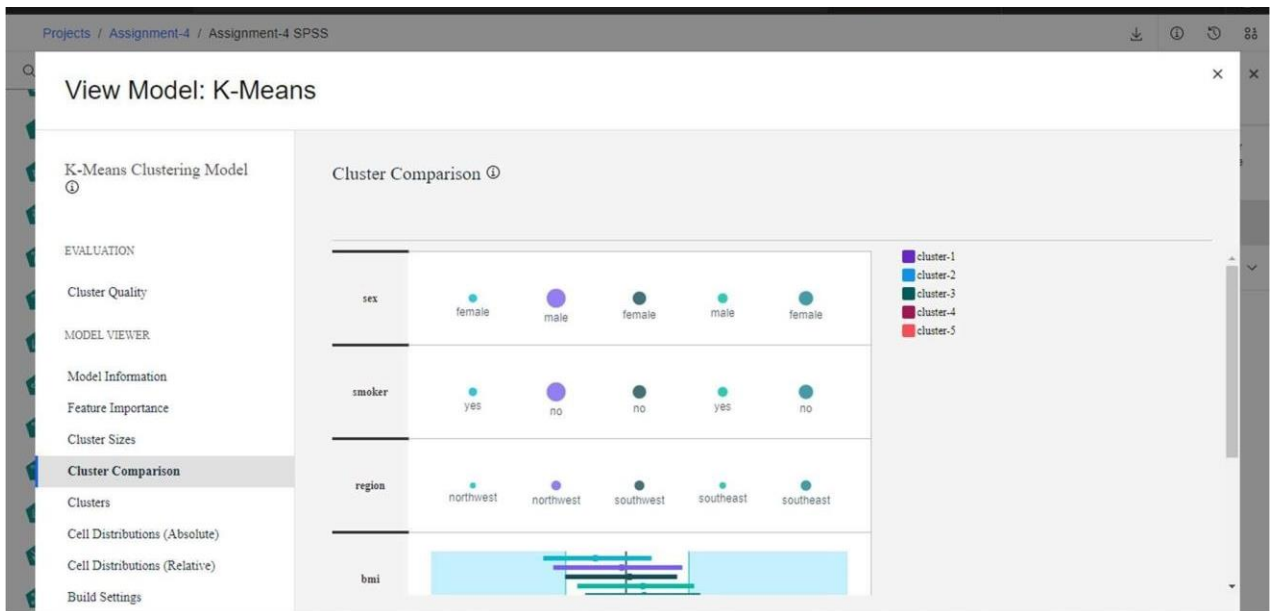


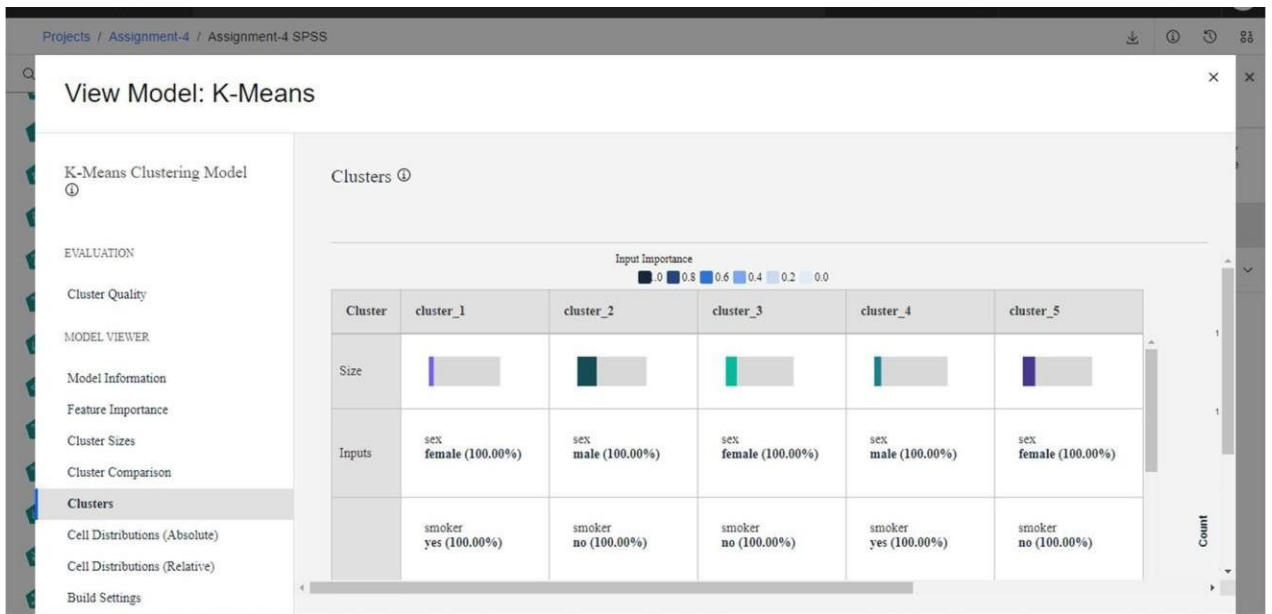
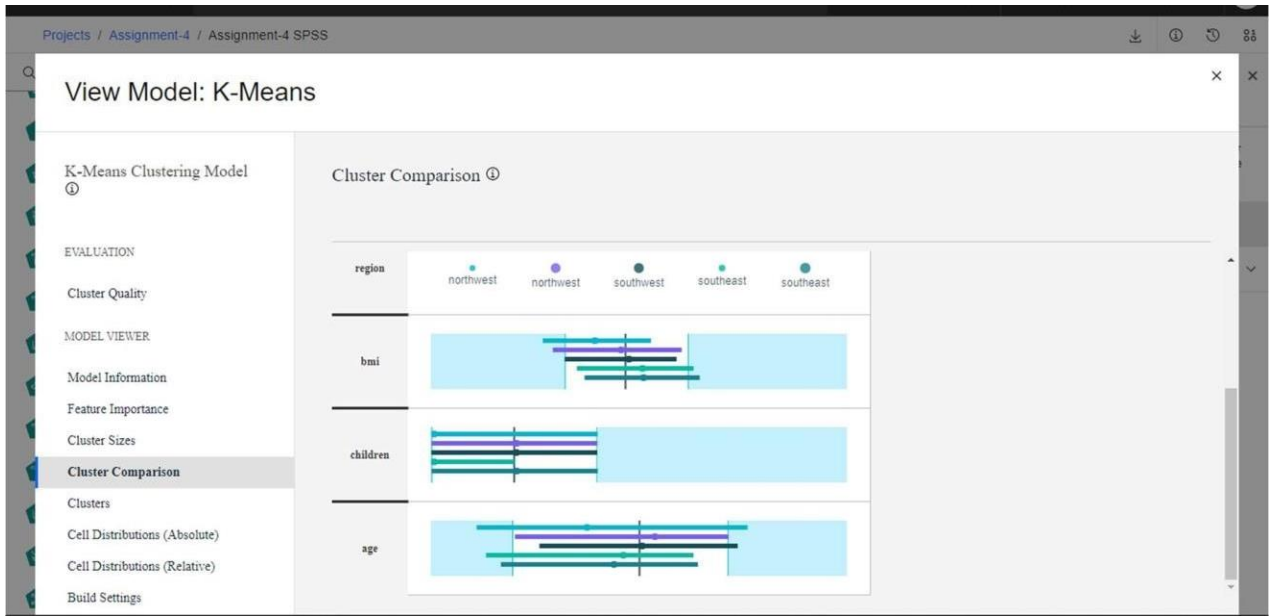
Output:

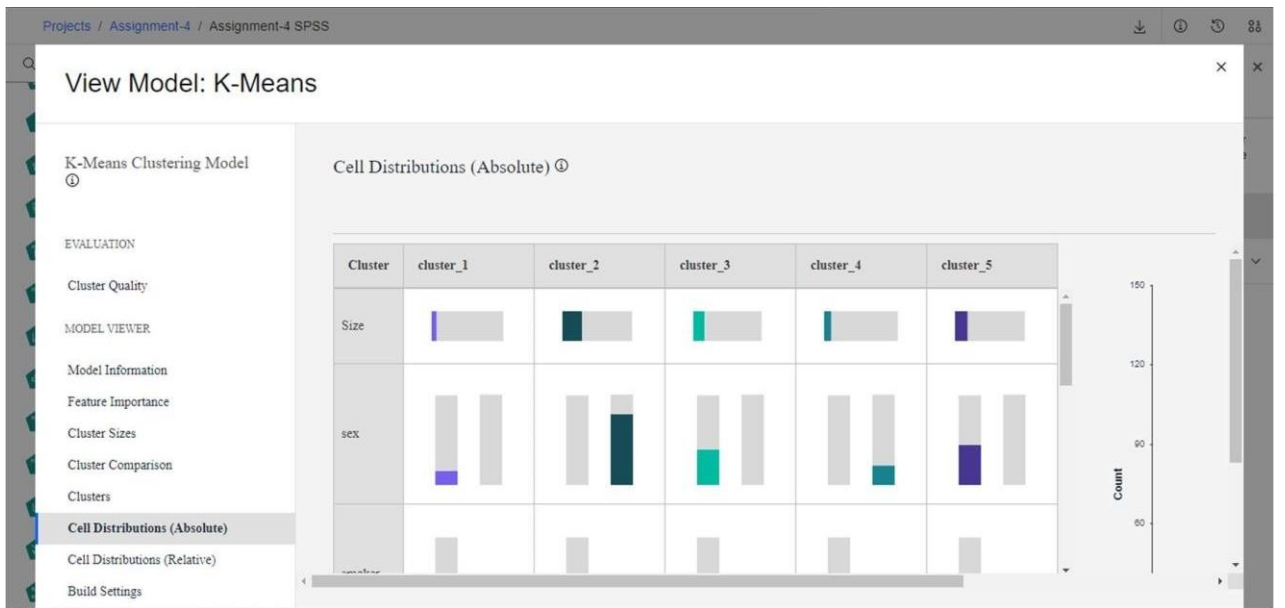
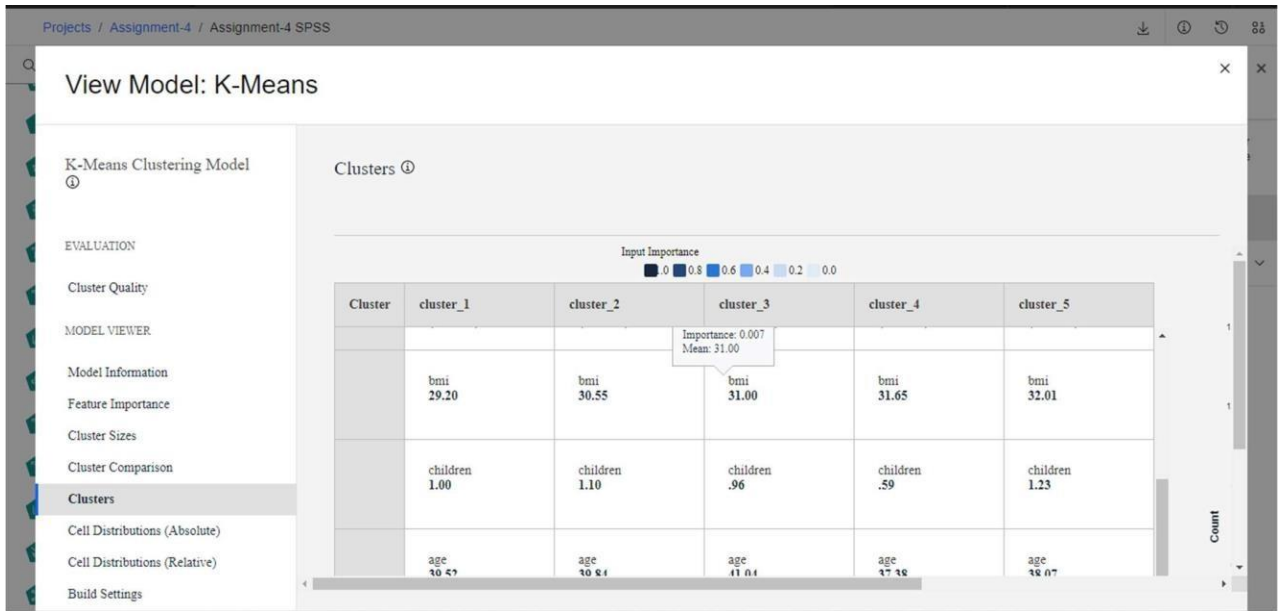


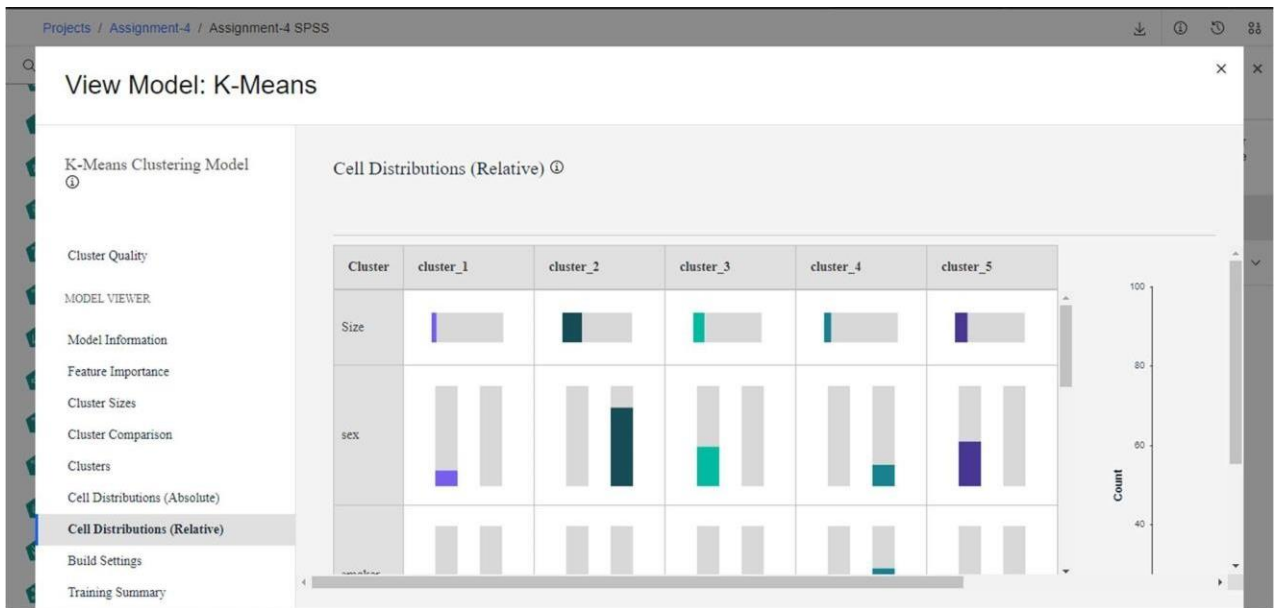
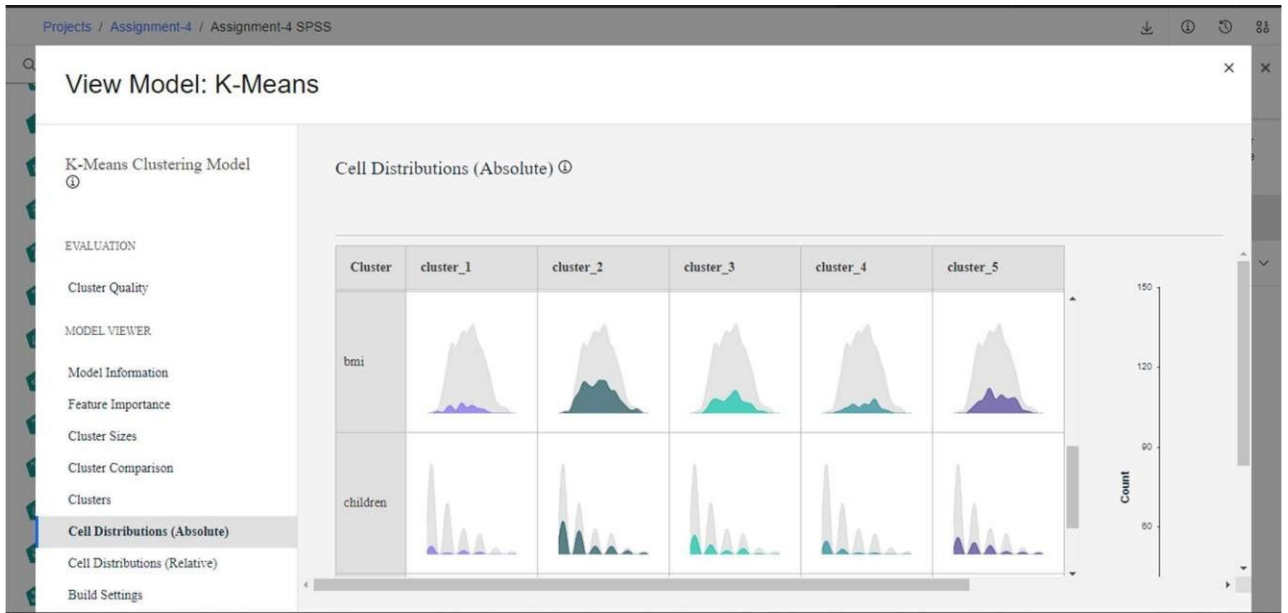


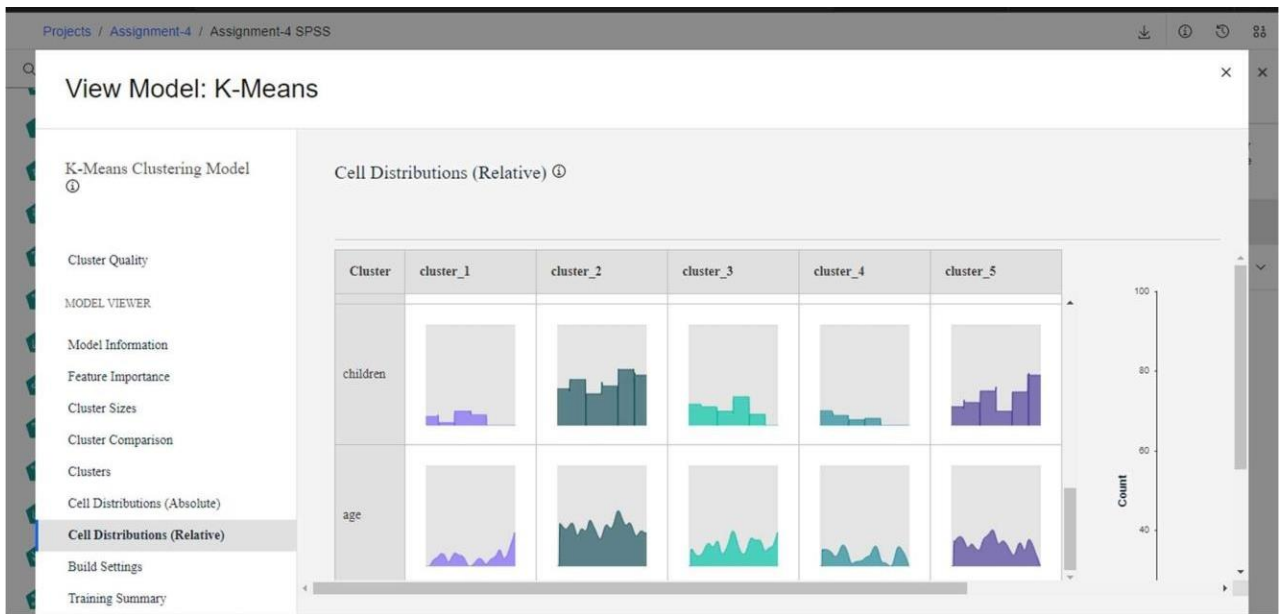
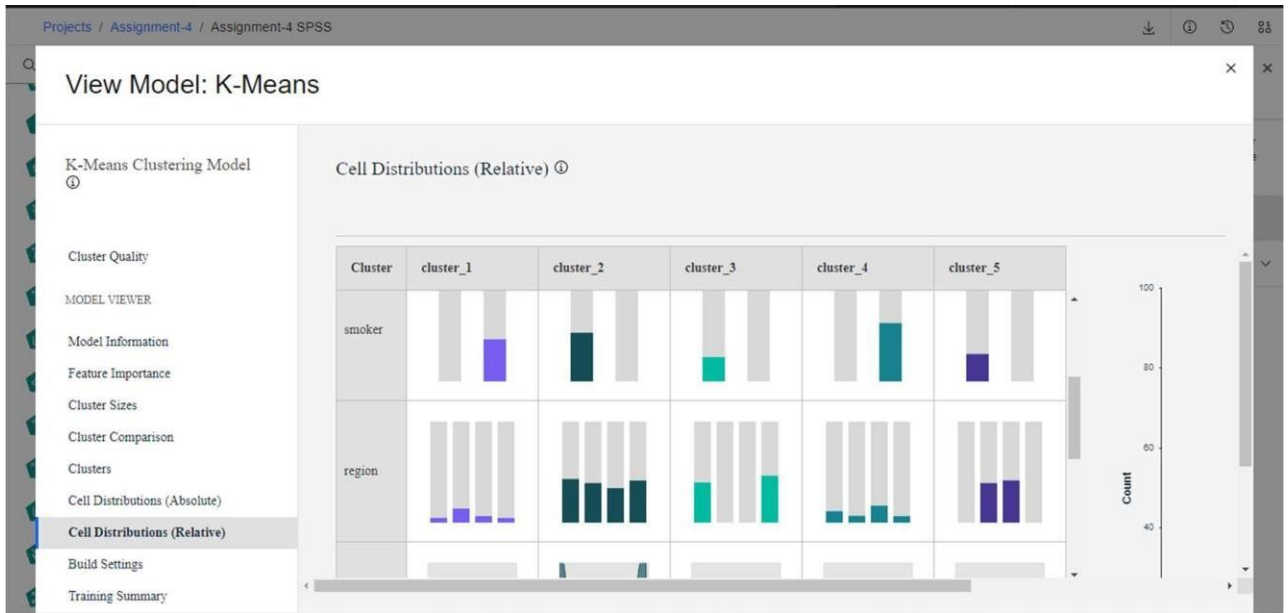












Projects / Assignment-4 / Assignment-4 SPSS

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 |
| Label prefix | cluster |

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)

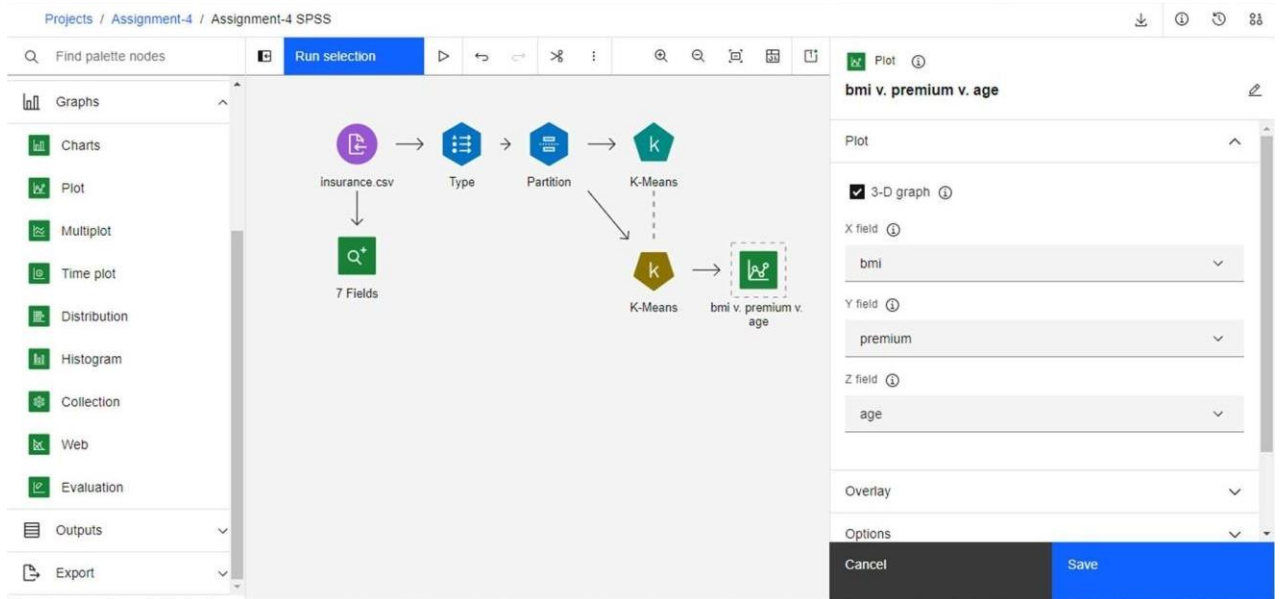
Build Settings

Training Summary

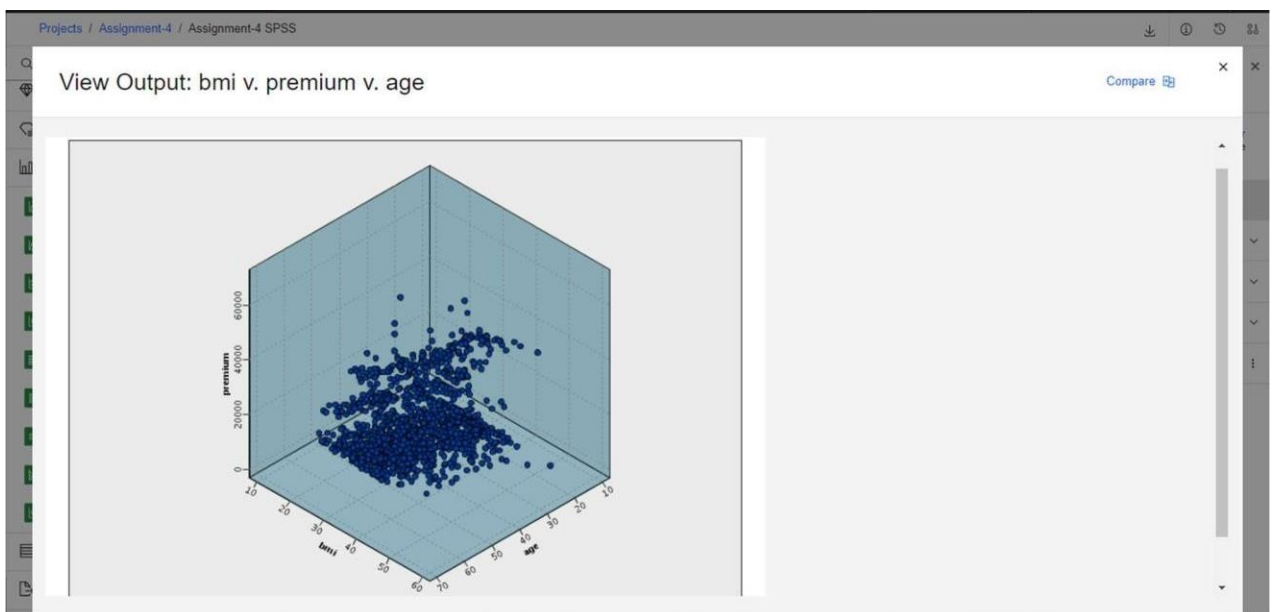
Training Summary ⓘ

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

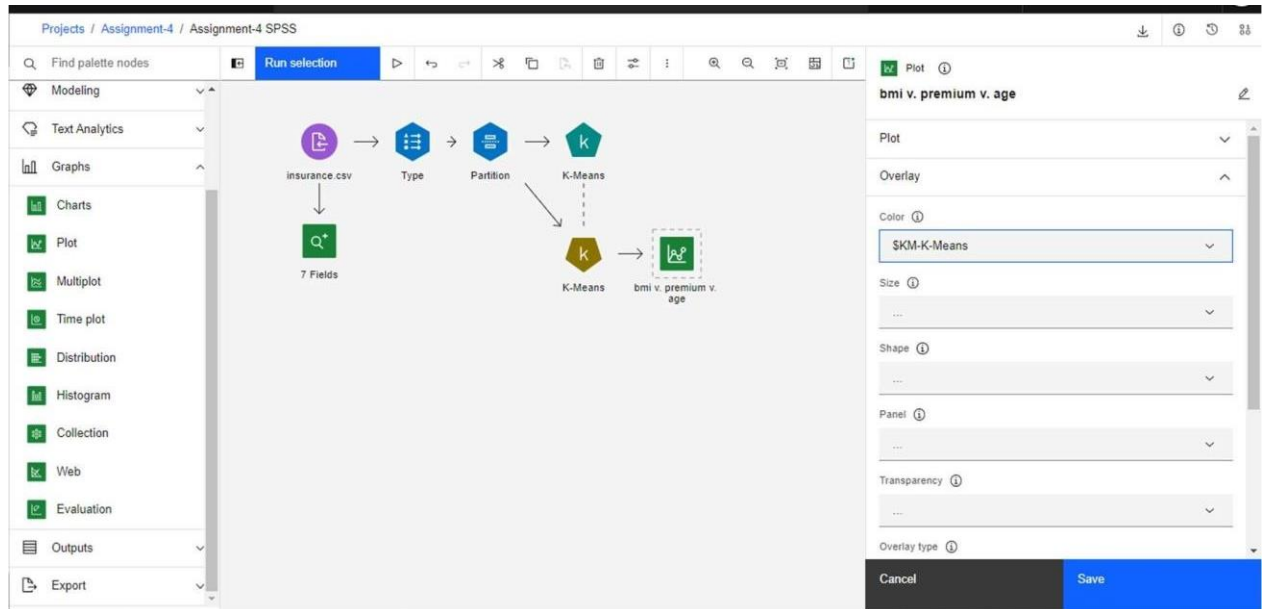
Creating a Plot Node with BMI vs Premium vs Age:



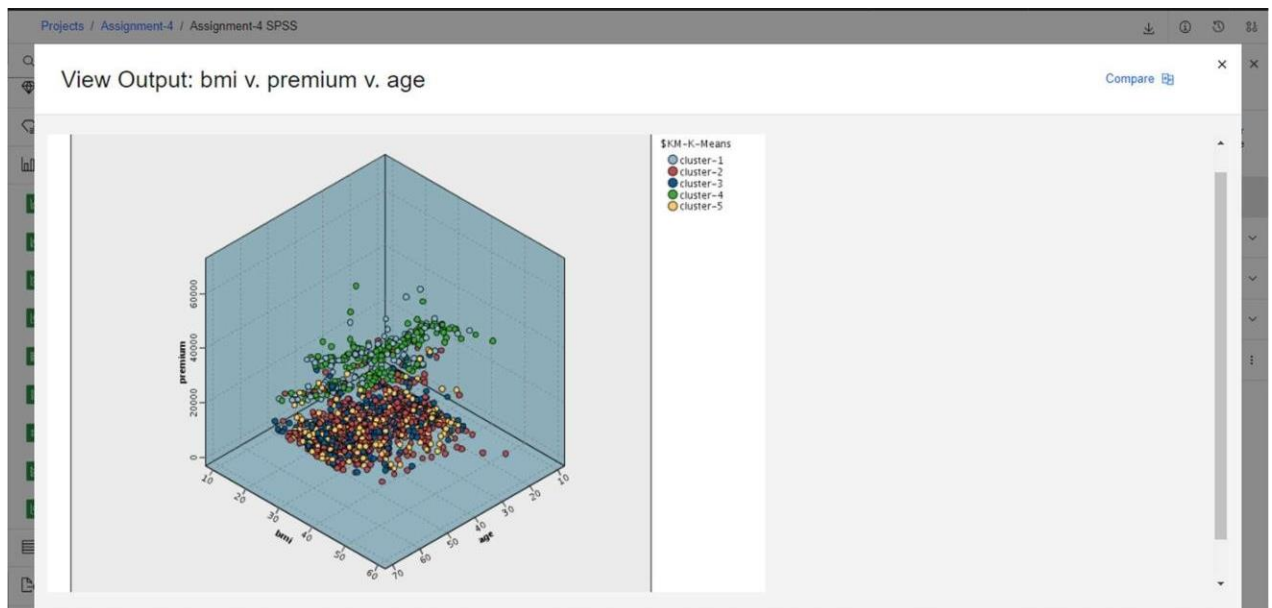
Output:



Assigning Colours:



Output:



Changing Number of clusters to 3:

Projects / Assignment-4 / Assignment-4 SPSS

Find palette nodes

Run selection

Modeling

Text Analytics

Graphs

Charts

Plot

Multiplot

Time plot

Distribution

Histogram

Collection

Web

Evaluation

Outputs

Export

insurance.csv

7 Fields

Type

Partition

K-Means

K-Means

bmi v. premium v. age

K-Means

Fields

Build Options

Model Name

Auto

Custom

Use partitioned data

Number of clusters

3

Generate distance field

Cluster label

String

Number

label prefix

Cancel

Save

Output:

