

Data Analytics Externship Program

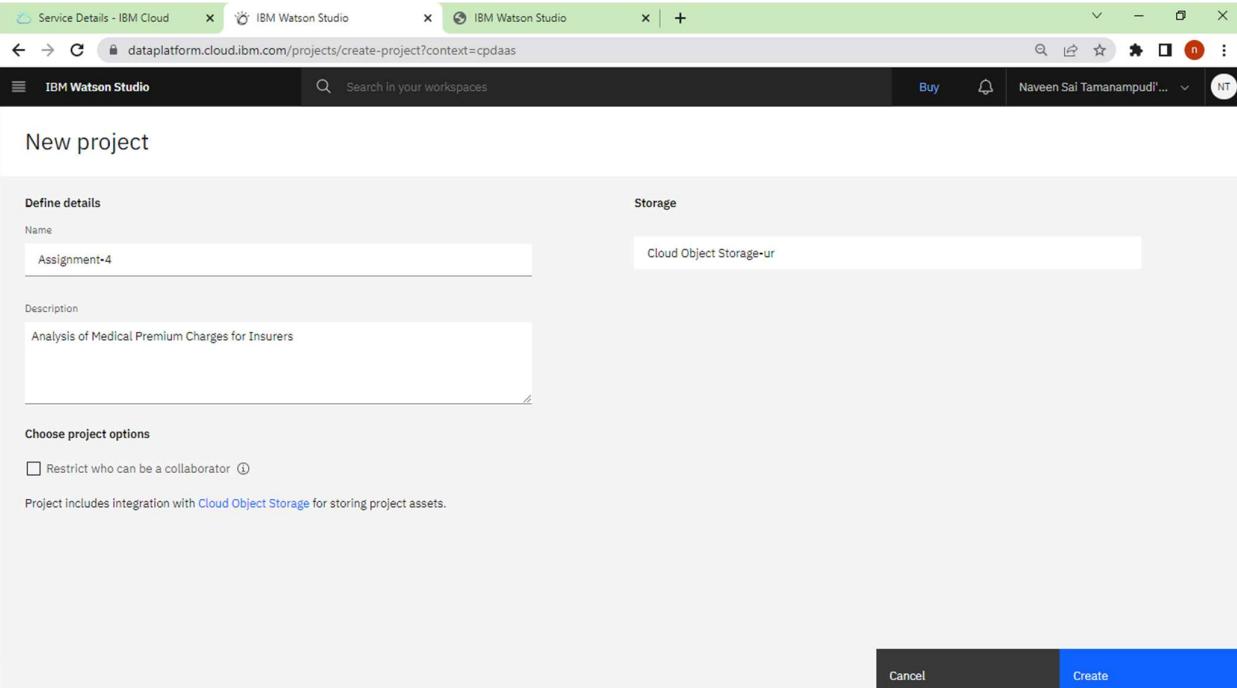
Assignment-4

Name: Tamanampudi Naveen Sai
Email ID: naveen.19bcd7003@vitap.ac.in

Clustering: Analysis of Medical Premium Charges for Insurers

Dataset Used: insurance.csv

Creating the Project



The screenshot shows the 'Create Project' dialog box in the IBM Watson Studio interface. The URL in the browser is dataplatform.cloud.ibm.com/projects/create-project?context=cpdaas. The form fields are as follows:

- Name:** Assignment-4
- Description:** Analysis of Medical Premium Charges for Insurers
- Storage:** Cloud Object Storage-ur
- Choose project options:** A checkbox labeled "Restrict who can be a collaborator" is unchecked.
- Note:** "Project includes integration with Cloud Object Storage for storing project assets."
- Buttons:** "Cancel" and "Create" (the "Create" button is highlighted in blue).

Uploading Data:

The screenshot shows the IBM Watson Studio interface with the 'Assets' tab selected. On the left, there's a sidebar for 'Asset types' with 'Data' selected, showing one 'Data asset'. The main area displays a table titled 'All assets' with one item: 'insurance.csv' (Last modified: Now, Naveen Sai Tamanampudi (You)). A large callout box on the right says 'Drop data files here or browse for files to upload'.

Creating a Data Refinery:

The screenshot shows a modal dialog titled 'Select data from project' for the 'Assignment-4' project. In the center, there are two tables: 'Assignment-4' and 'Data assets'. The 'Data assets' table has one item, 'insurance.csv'. On the right, a detailed view of 'Selected assets' shows the following information for 'insurance.csv': 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. At the bottom of the dialog are 'Cancel' and 'Select' buttons.

Viewing the output of refinery:

The screenshot shows the IBM Watson Studio interface with a data preview of the 'insurance.csv' file. The preview table has columns: age, sex, bmi, children, smoker, region, and premium. The data consists of 1338 rows. The 'age' column ranges from 19 to 60+ with a median of 39. The 'sex' column has two categories: male and female. The 'bmi' column ranges from 18.7 to 42.13. The 'children' column ranges from 0 to 3. The 'smoker' column has two categories: yes and no. The 'region' column has four categories: southwest, southeast, northwest, and northeast. The 'premium' column ranges from 16884.924 to 39611.7577.

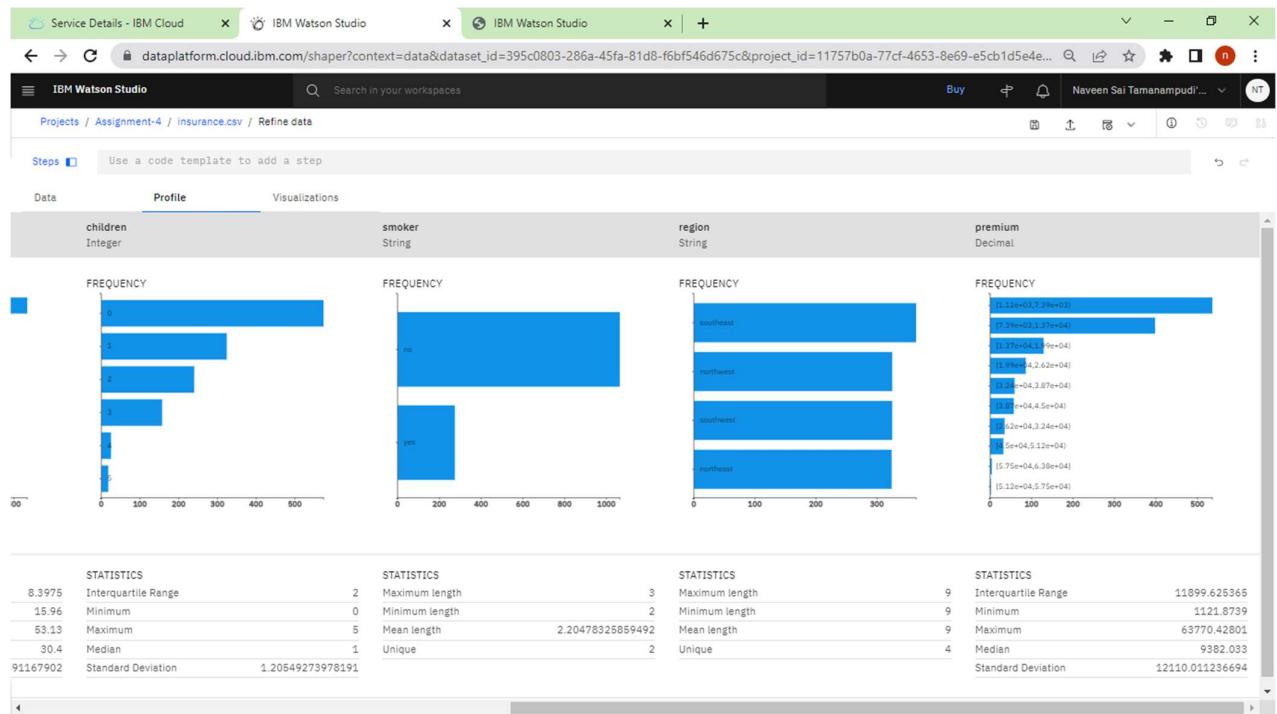
	age	sex	bmi	children	smoker	region	premium
	Integer	String	Decimal	Integer	String	String	Decimal
1	19	female	27.9	0	yes	southwest	16884.924
2	18	male	33.77	1	no	southeast	1725.5523
3	28	male	33	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.47061
5	32	male	28.88	0	no	northwest	3866.8552
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

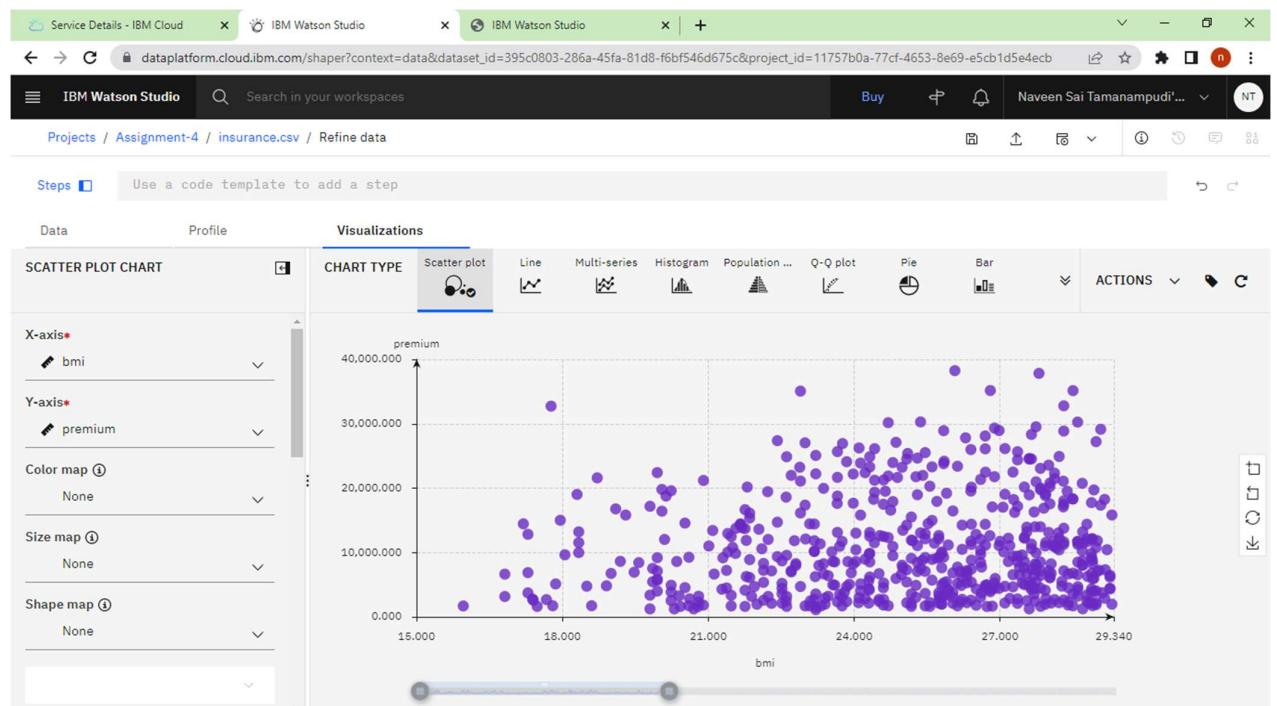
Viewing Profile:

The screenshot shows the IBM Watson Studio interface with profile statistics and histograms for the 'insurance.csv' file. The top section displays five histograms for 'age', 'sex', 'bmi', 'children', and 'smoker'. The bottom section displays five tables of statistical data for each variable.

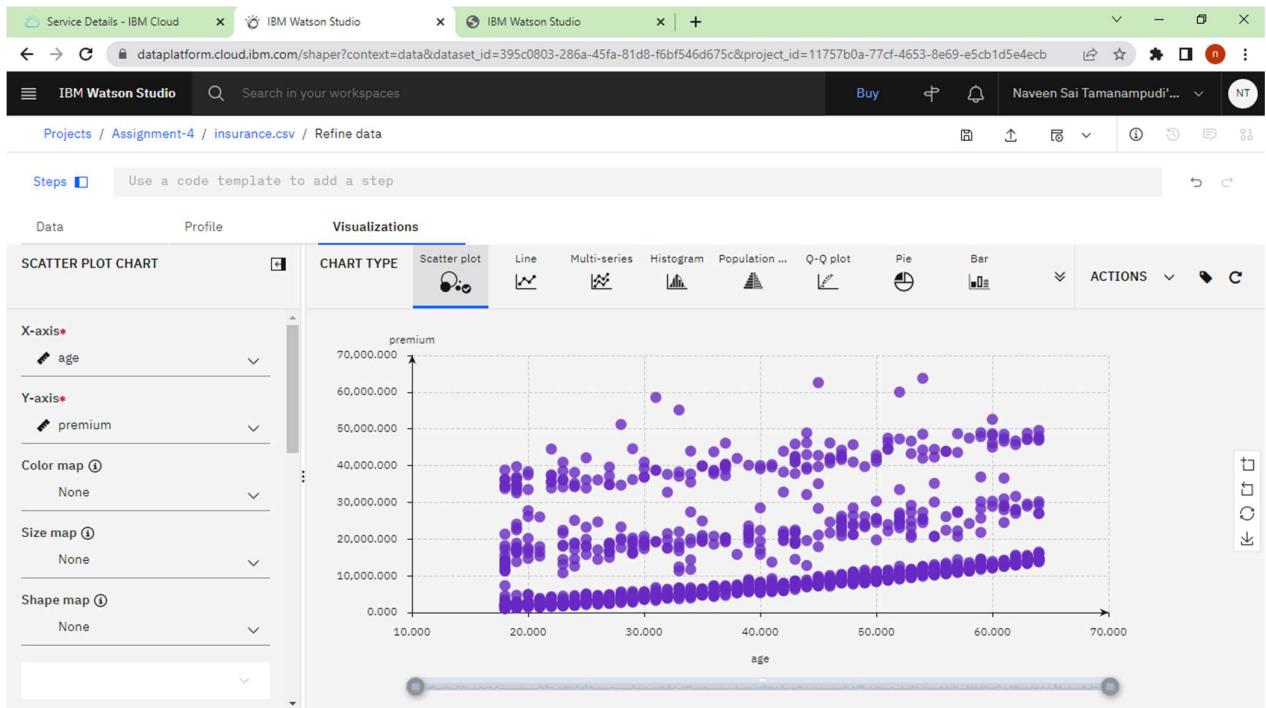
age	sex	bmi	children	smoker
STATISTICS	STATISTICS	STATISTICS	STATISTICS	STATISTICS
Interquartile Range	Maximum length	Interquartile Range	Maximum length	Interquartile Range
24	Minimum length	8.3975	Minimum length	2
Minimum	4	Minimum	15.96	Maximum length
18	Mean length	53.13	Minimum	0
Maximum	4.98953662182362	Maximum	56.13	Mean length
64	Unique	Median	30.4	Unique
Median	2	Standard Deviation	6.09818691167902	1
39		Standard Deviation	1.20549273978191	
Standard Deviation	14.0499603792162			



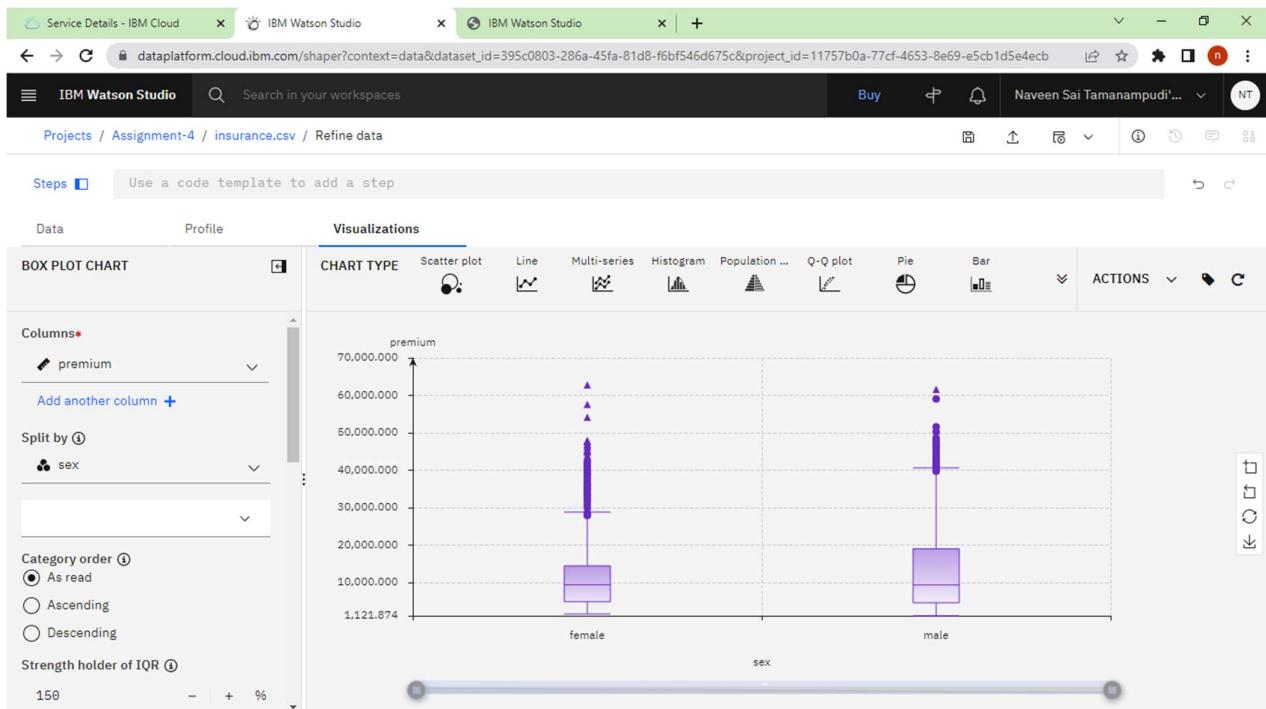
Viewing Scatterplot of BMI vs Premium:



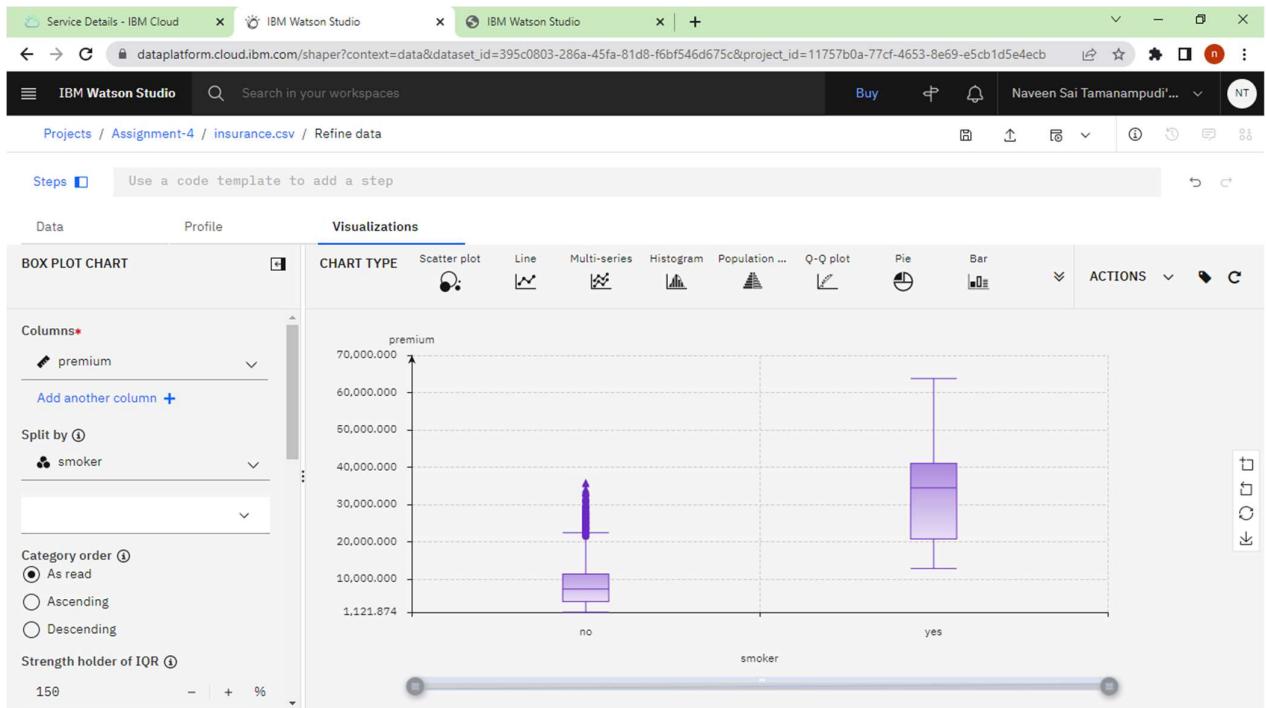
Viewing Scatterplot of Age vs Premium:



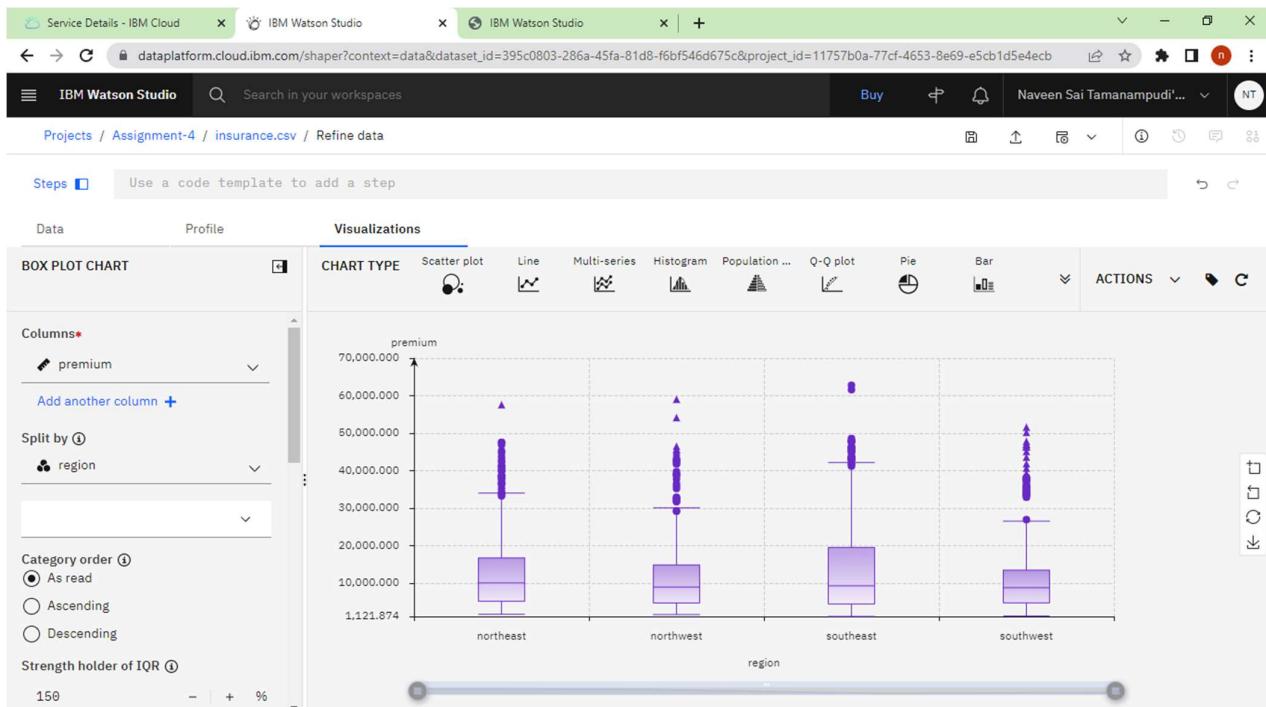
Viewing Boxplot of Sex vs Premium:



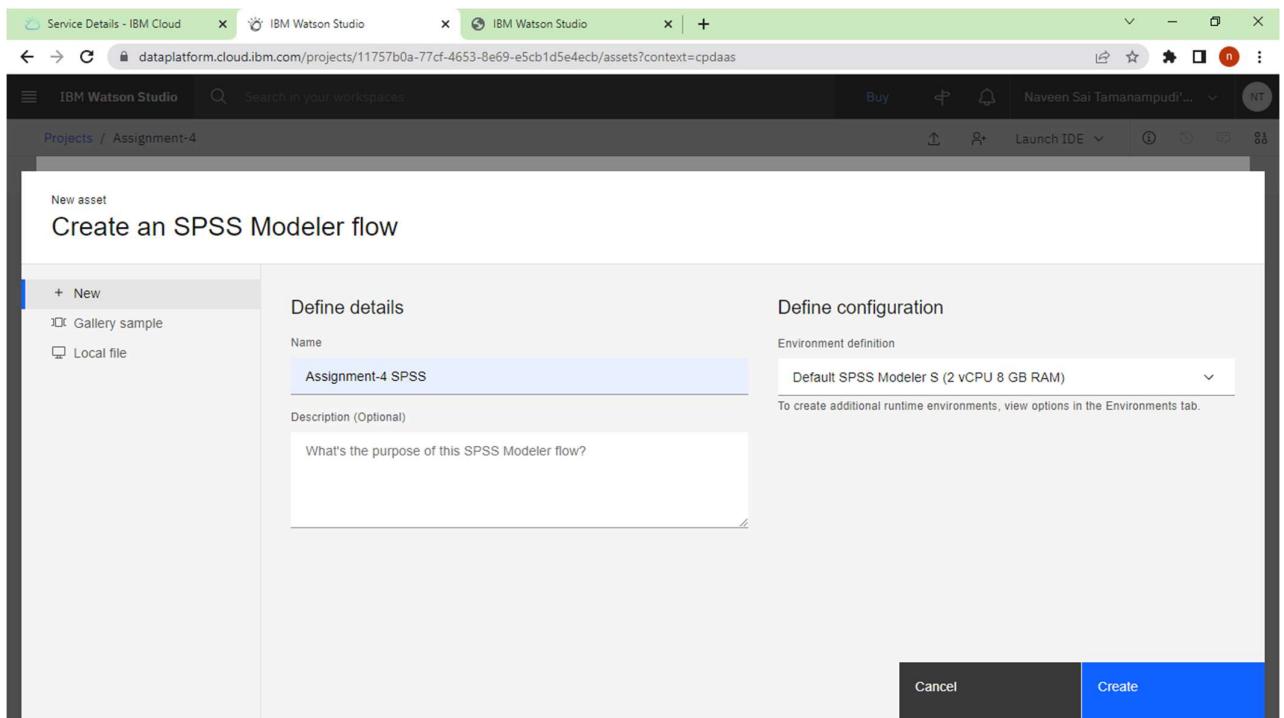
Viewing Boxplot of Smoker vs Premium:



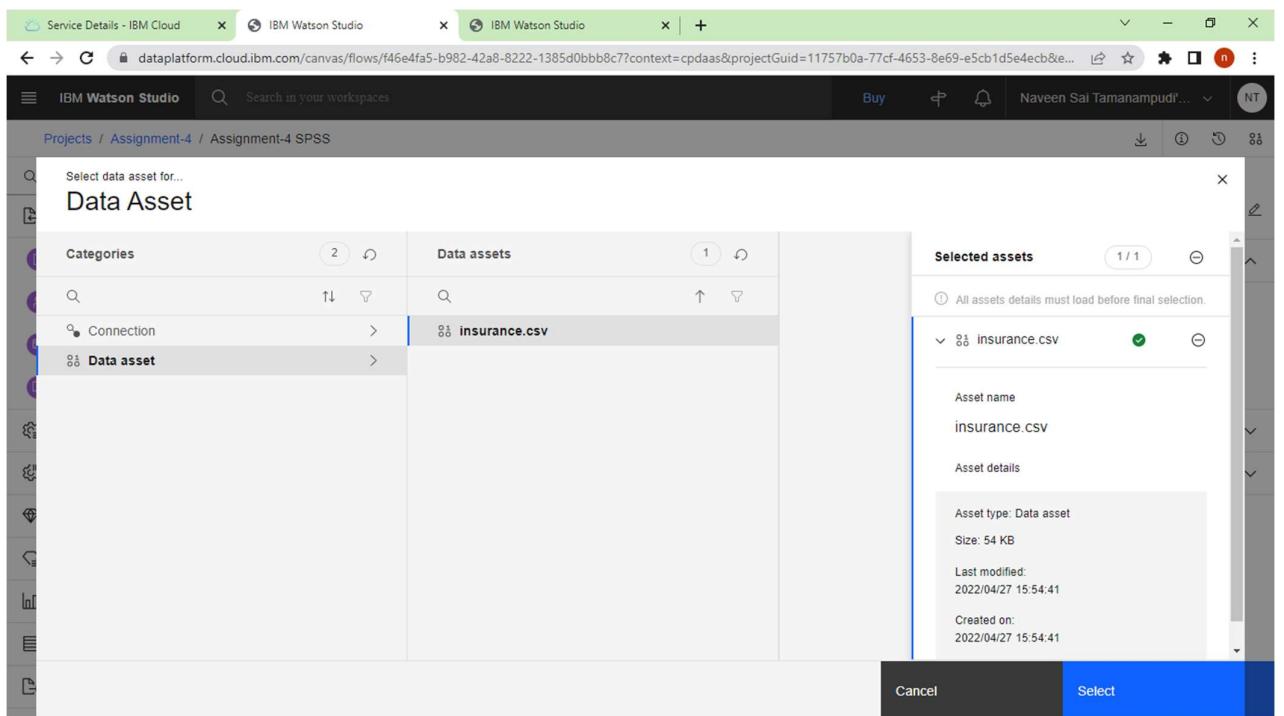
Viewing Boxplot of Region vs Premium:



Creating an SPSS Modeler:



Uploading Dataset to Data Asset Node:



The screenshot shows the IBM Watson Studio interface with a tab bar at the top containing 'Service Details - IBM Cloud', 'IBM Watson Studio', and another 'IBM Watson Studio' tab. The main area is titled 'Projects / Assignment-4 / Assignment-4 SPSS'. On the left, a palette sidebar lists categories like 'Import', 'Data Asset' (which is selected and highlighted in purple), 'User Input', 'Sim Gen', 'Extension Import', etc. The central workspace shows a 'Data Asset' configuration dialog. It includes sections for 'Data', 'File format properties' (set to CSV), 'Encoding' (set to UTF-8), and 'Invalid data handling'. At the bottom are 'Cancel' and 'Save' buttons.

Creating Data Audit Node:

The screenshot shows the IBM Watson Studio interface with a tab bar at the top containing 'Service Details - IBM Cloud', 'IBM Watson Studio', and another 'IBM Watson Studio' tab. The main area is titled 'Projects / Assignment-4 / Assignment-4 SPSS'. On the left, a palette sidebar lists categories like 'Outputs', 'Table', 'Matrix', 'Analysis', 'Data Audit' (selected and highlighted in green), 'Transform', 'Statistics', 'Means', 'Report', 'Set Globals', 'Sim Fit', and 'Sim Evaluation'. The central workspace shows the results of a 'Data Audit' node. It displays a message 'Last run was now', a success message 'Run was successful', and a list of 'All results' including 'Data Audit of [7 fields] Just now'. The right side of the screen shows tabs for 'Outputs' and 'Models'.

Output from Data Audit:

Service Details - IBM Cloud IBM Watson Studio IBM Watson Studio

IBM Watson Studio Search in your workspaces Buy Naveen Sai Tamanampudi... NT

Projects / Assignment-4 / Assignment-4 SPSS

View Output: Data Audit of [7 fields]

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

Service Details - IBM Cloud IBM Watson Studio IBM Watson Studio

IBM Watson Studio Search in your workspaces Buy Naveen Sai Tamanampudi... NT

Projects / Assignment-4 / Assignment-4 SPSS

View Output: Data Audit of [7 fields]

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	1121.874	63770.428	13270.422	12110.011	1.516	--	1338	

Creating a Type Node and setting Inputs, Targets:

The screenshot shows the IBM Watson Studio interface with a canvas workspace titled "Assignment-4 SPSS". On the left, a palette lists various nodes: Import, Record Operations, Field Operations, Auto Data Prep, Type, Filter, Derive, Filler, Reclassify, Binning, RFM Analysis, and Ensemble. A search bar at the top right says "Search in your workspaces".

The main canvas has a flow starting with an "insurance.csv" input node, followed by a "Type" node, which is currently selected. The "Type" node settings panel is open, showing "Settings" and a table for "Find in column Field". The table lists fields: # age, abc sex, *# bmi, # children, abc smoker, abc region, and *# premium. The "Role" column indicates most are "Input" except for "premium" which is "Target". The "Value mode" column shows "Read" for most and "Target" for "premium". The "Values" column shows "Continuous" for all.

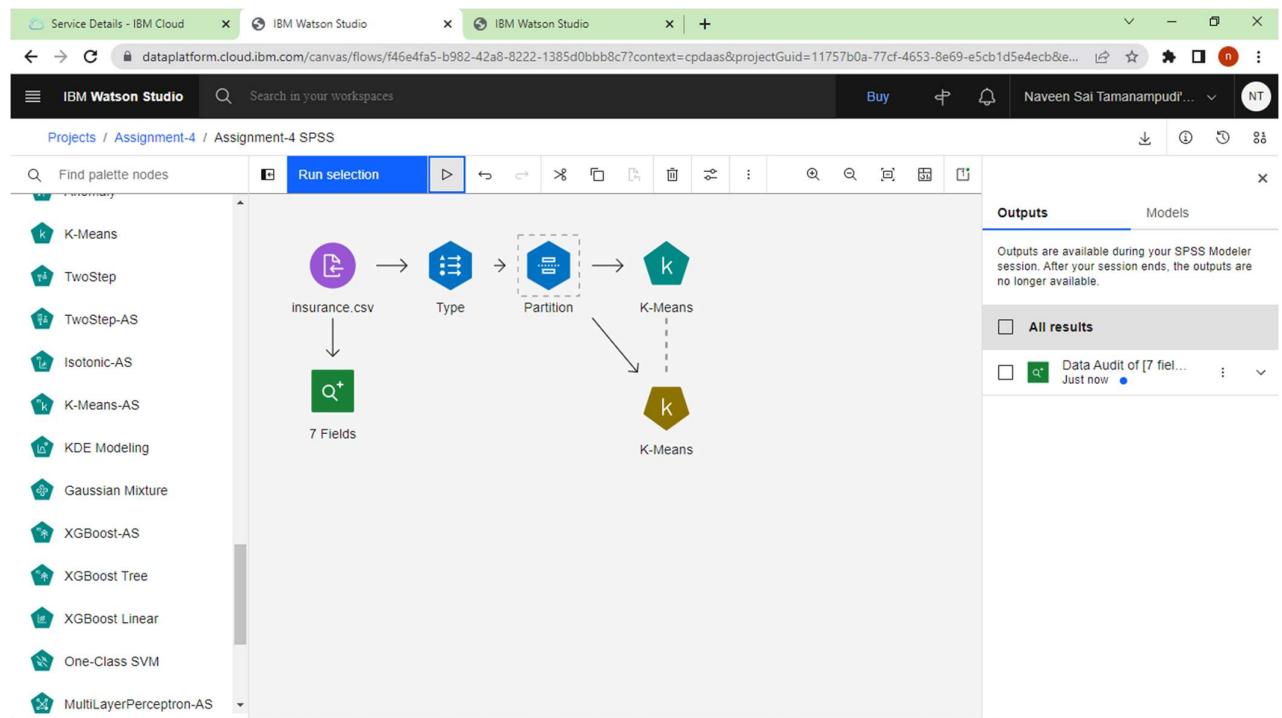
At the bottom of the canvas, there are "Cancel" and "Save" buttons.

Creating a Partition Node with 80:20 split:

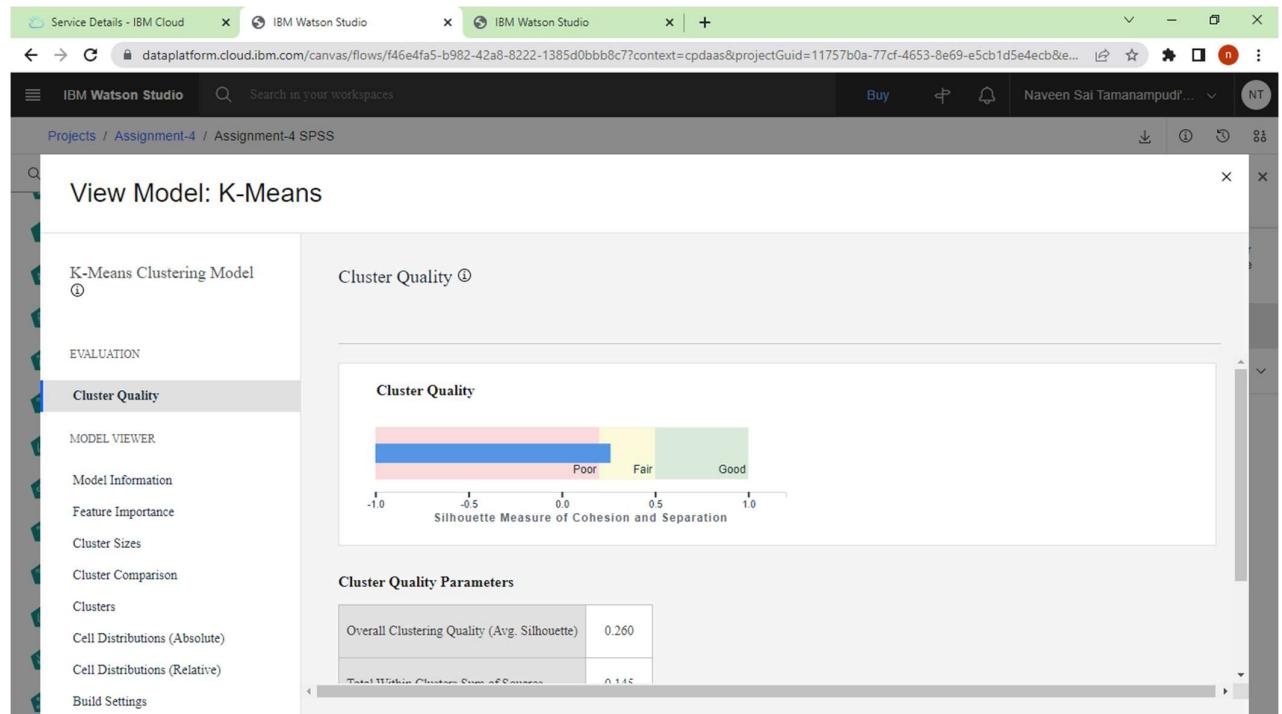
The screenshot shows the same IBM Watson Studio interface with the same workspace and palette. The "insurance.csv" input node is connected to a "Type" node, which is then connected to a "Partition" node. The "Partition" node settings panel is open, showing "Settings" and a "Partition" section. In the "Partition" section, the "Training Partition(%)" field is set to 80 and the "Testing Partition(%)" field is set to 20. There are checkboxes for "Create validation partition" (unchecked) and "Repeatable partition assignment" (checked). A "Seed" field is set to 1234567.

At the bottom of the canvas, there are "Cancel" and "Save" buttons.

Creating a K-Means Clustering Model:



Output:



Service Details - IBM Cloud IBM Watson Studio IBM Watson Studio

dataplatform.cloud.ibm.com/canvas/flows/f46e4fa5-b982-42a8-8222-1385d0bbb8c7?context=cpdaas&projectGuid=11757b0a-77cf-4653-8e69-e5cb1d5e4ecb&e...

IBM Watson Studio Search in your workspaces Buy Naveen Sai Tamanampudi...

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 ④

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

Service Details - IBM Cloud IBM Watson Studio IBM Watson Studio

dataplatform.cloud.ibm.com/canvas/flows/f46e4fa5-b982-42a8-8222-1385d0bbb8c7?context=cpdaas&projectGuid=11757b0a-77cf-4653-8e69-e5cb1d5e4ecb&e...

IBM Watson Studio Search in your workspaces Buy Naveen Sai Tamanampudi...

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 ④

Algorithm	K-Means
Model Class	Center Based
Number of Features	6
Distance Measure	Euclidean
Number of Clusters	5
Cluster 1	94 (8.79%)
Cluster 2	411 (38.45%)

Service Details - IBM Cloud IBM Watson Studio IBM Watson Studio

dataplatform.cloud.ibm.com/canvas/flows/f46e4fa5-b982-42a8-8222-1385d0bbb8c7?context=cpdaas&projectGuid=11757b0a-77cf-4653-8e69-e5cb1d5e4ecb&e...

IBM Watson Studio Buy Naveen Sai Tamanampudi... NT

Projects / Assignment-4 / Assignment-4 SPSS

View Model: K-Means

K-Means Clustering Model ①

- EVALUATION
- Cluster Quality
- MODEL VIEWER
- Model Information**
- Feature Importance
- Cluster Sizes
- Cluster Comparison
- Clusters
- Cell Distributions (Absolute)
- Cell Distributions (Relative)
- Build Settings

Model Information ④

Number of Clusters	
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

Service Details - IBM Cloud IBM Watson Studio IBM Watson Studio

dataplatform.cloud.ibm.com/canvas/flows/f46e4fa5-b982-42a8-8222-1385d0bbb8c7?context=cpdaas&projectGuid=11757b0a-77cf-4653-8e69-e5cb1d5e4ecb&e...

IBM Watson Studio Buy Naveen Sai Tamanampudi... NT

Projects / Assignment-4 / Assignment-4 SPSS

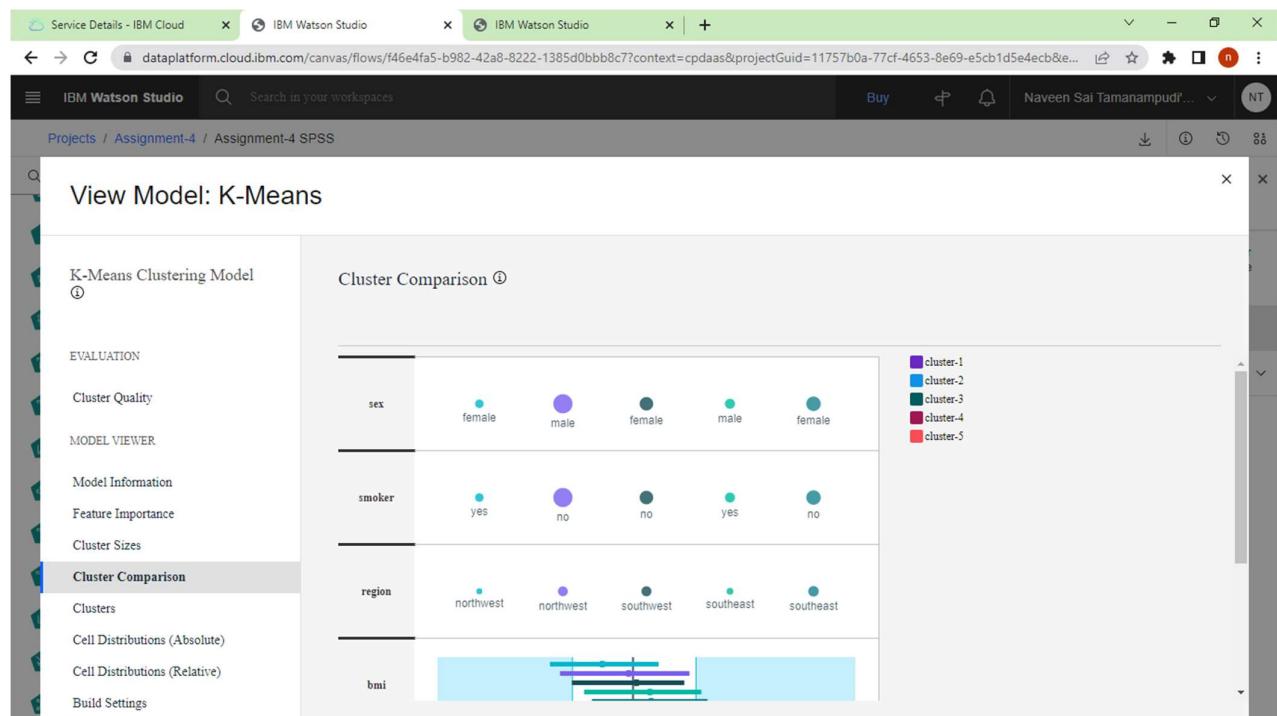
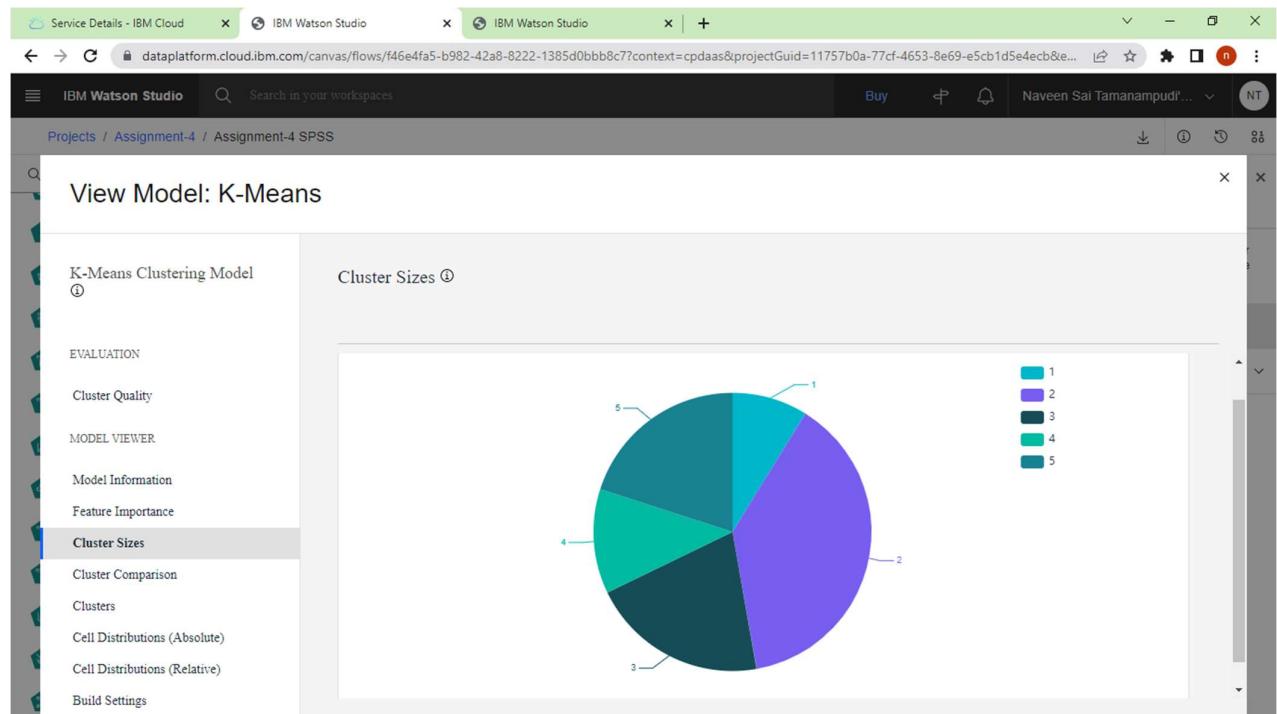
View Model: K-Means

K-Means Clustering Model ①

- EVALUATION
- Cluster Quality
- MODEL VIEWER
- Model Information
- Feature Importance**
- Cluster Sizes
- Cluster Comparison
- Clusters
- Cell Distributions (Absolute)
- Cell Distributions (Relative)
- Build Settings

Feature Importance ④

Feature	Importance Value
sex	1.00
smoker	1.00
region	0.38
bmi	0.01
children	0.00
age	0.00



Service Details - IBM Cloud IBM Watson Studio IBM Watson Studio

dataplatform.cloud.ibm.com/canvas/flows/f46e4fa5-b982-42a8-8222-1385d0bbb8c7?context=cpdaas&projectGuid=11757b0a-77cf-4653-8e69-e5cb1d5e4ecb&e...

IBM Watson Studio Buy Notifications Naveen Sai Tamanampudi... NT

Projects / Assignment-4 / Assignment-4 SPSS

View Model: K-Means

K-Means Clustering Model ①

EVALUATION

- Cluster Quality

MODEL VIEWER

- Model Information
- Feature Importance
- Cluster Sizes
- Cluster Comparison**
- Clusters
- Cell Distributions (Absolute)
- Cell Distributions (Relative)
- Build Settings

Cluster Comparison ①

The visualization displays three features: region, bmi, and age. The region feature has five categories: northwest, northwest, southwest, southeast, and southeast. The bmi and age features show horizontal box plots for each cluster, with the median line at zero. The children feature shows vertical box plots for each cluster.

Service Details - IBM Cloud IBM Watson Studio IBM Watson Studio

dataplatform.cloud.ibm.com/canvas/flows/f46e4fa5-b982-42a8-8222-1385d0bbb8c7?context=cpdaas&projectGuid=11757b0a-77cf-4653-8e69-e5cb1d5e4ecb&e...

IBM Watson Studio Buy Notifications Naveen Sai Tamanampudi... NT

Projects / Assignment-4 / Assignment-4 SPSS

View Model: K-Means

K-Means Clustering Model ①

EVALUATION

- Cluster Quality

MODEL VIEWER

- Model Information
- Feature Importance
- Cluster Sizes
- Cluster Comparison
- Clusters**
- Cell Distributions (Absolute)
- Cell Distributions (Relative)
- Build Settings

Clusters ①

Cluster	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5
Size	Input Importance sex female (100.00%)	Input Importance sex male (100.00%)	Input Importance sex female (100.00%)	Input Importance sex male (100.00%)	Input Importance sex female (100.00%)
Inputs	smoker yes (100.00%)	smoker no (100.00%)	smoker no (100.00%)	smoker yes (100.00%)	smoker no (100.00%)

Count

Service Details - IBM Cloud IBM Watson Studio IBM Watson Studio

IBM Watson Studio Buy Naveen Sai Tamanampudi...

Projects / Assignment-4 / Assignment-4 SPSS

View Model: K-Means

K-Means Clustering Model

EVALUATION

- Cluster Quality
- Model Information
- Feature Importance
- Cluster Sizes
- Cluster Comparison
- Clusters**
- Cell Distributions (Absolute)
- Cell Distributions (Relative)
- Build Settings

Clusters

Cluster	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5
bmi	29.20	30.55	31.00	31.65	32.01
children	1.00	1.10	.96	.59	1.23
age	20.57	20.84	21.01	21.38	22.07

Input Importance
0.0 0.8 0.6 0.4 0.2 0.0

Importance: 0.007 Mean: 31.00

Service Details - IBM Cloud IBM Watson Studio IBM Watson Studio

IBM Watson Studio Buy Naveen Sai Tamanampudi...

Projects / Assignment-4 / Assignment-4 SPSS

View Model: K-Means

K-Means Clustering Model

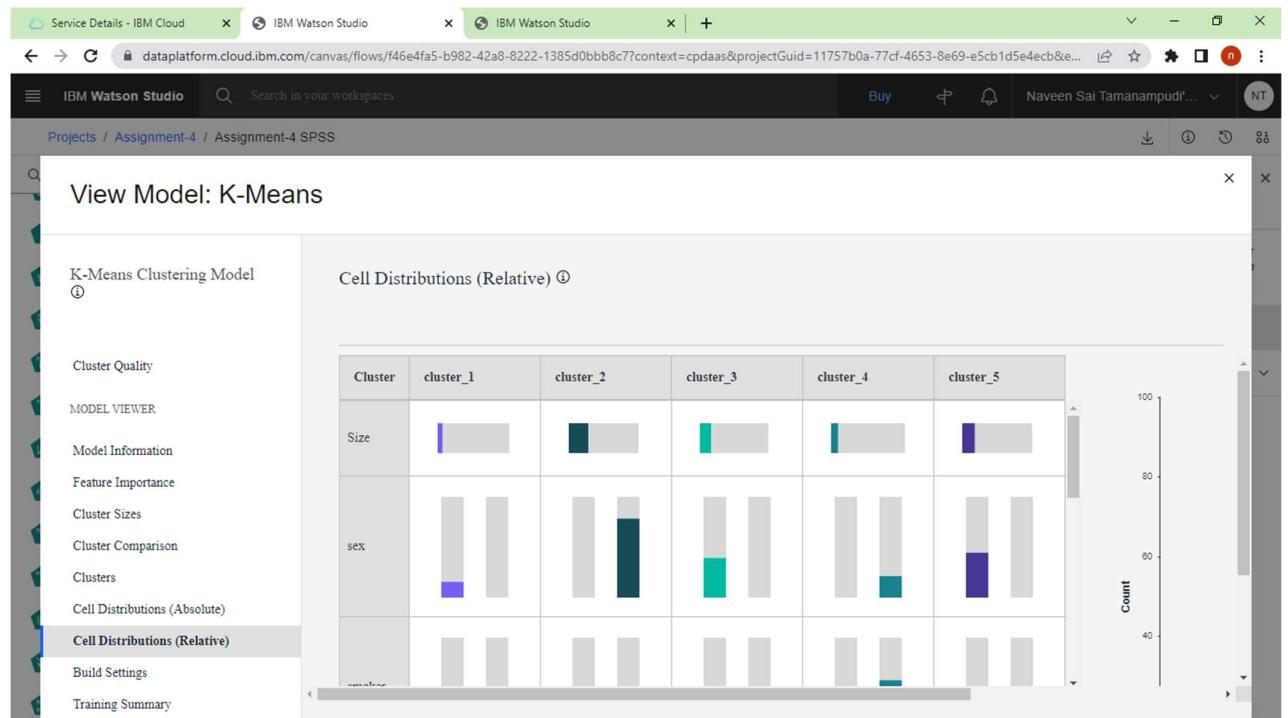
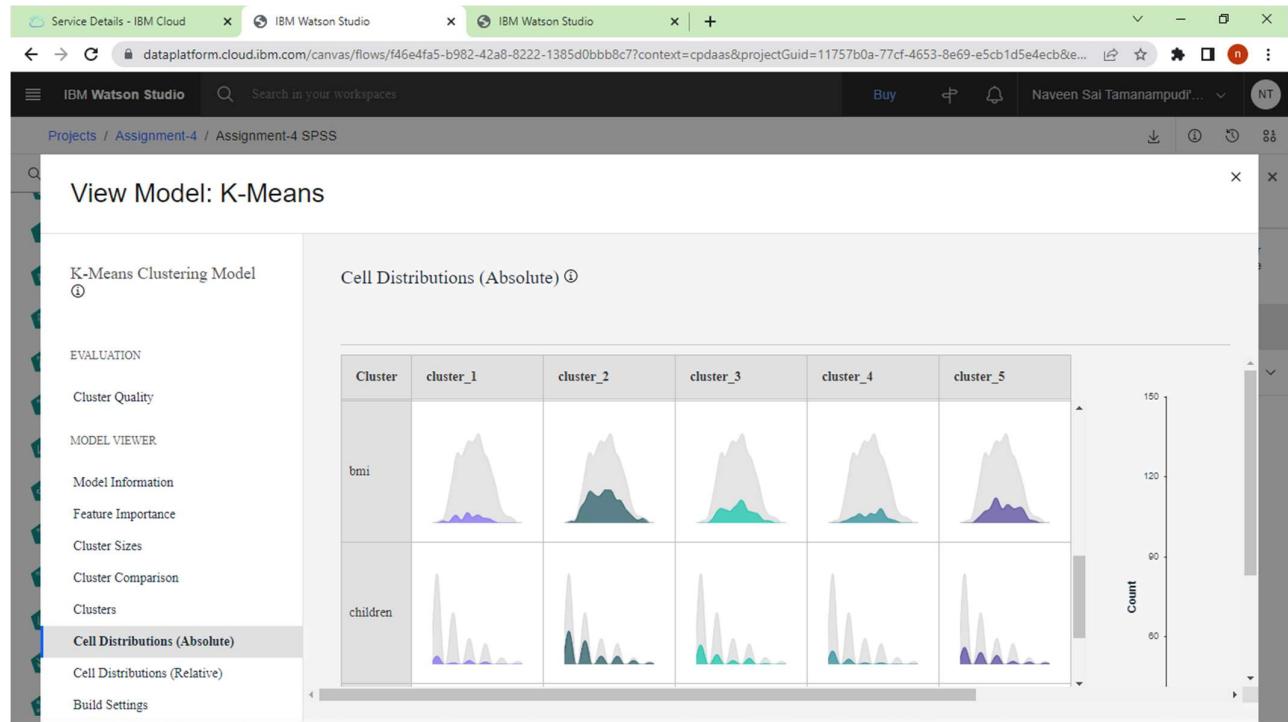
EVALUATION

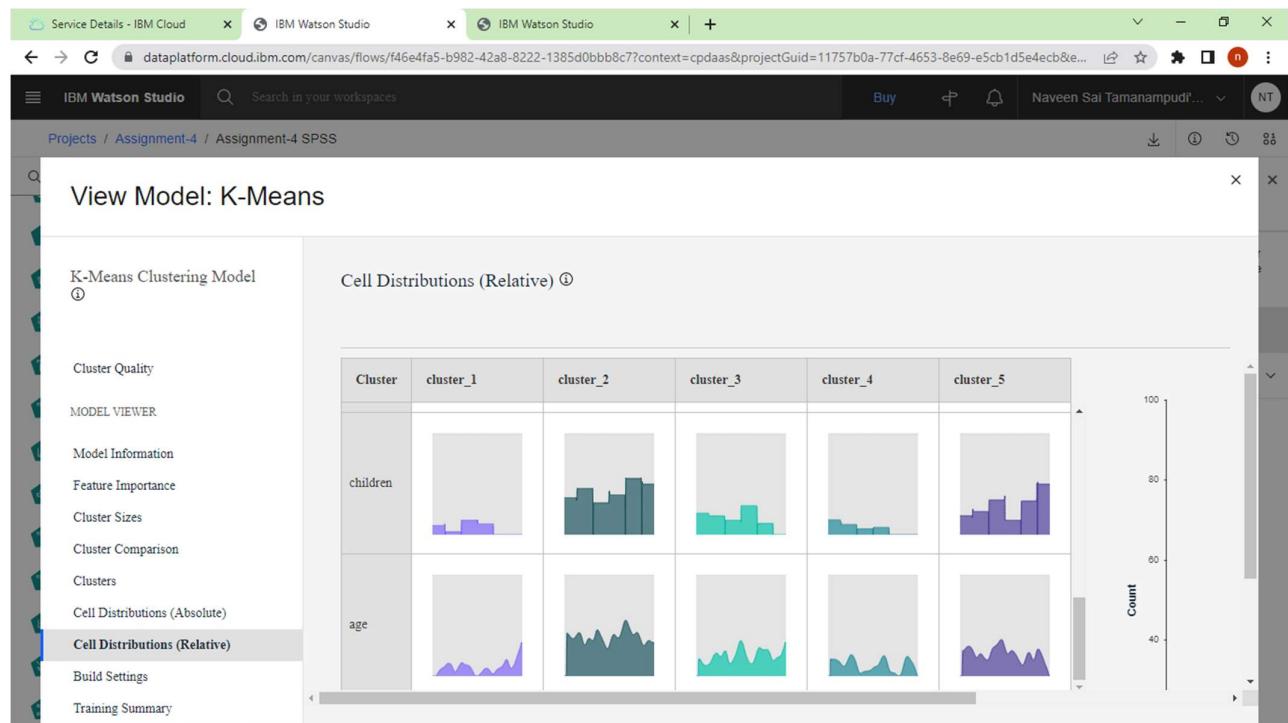
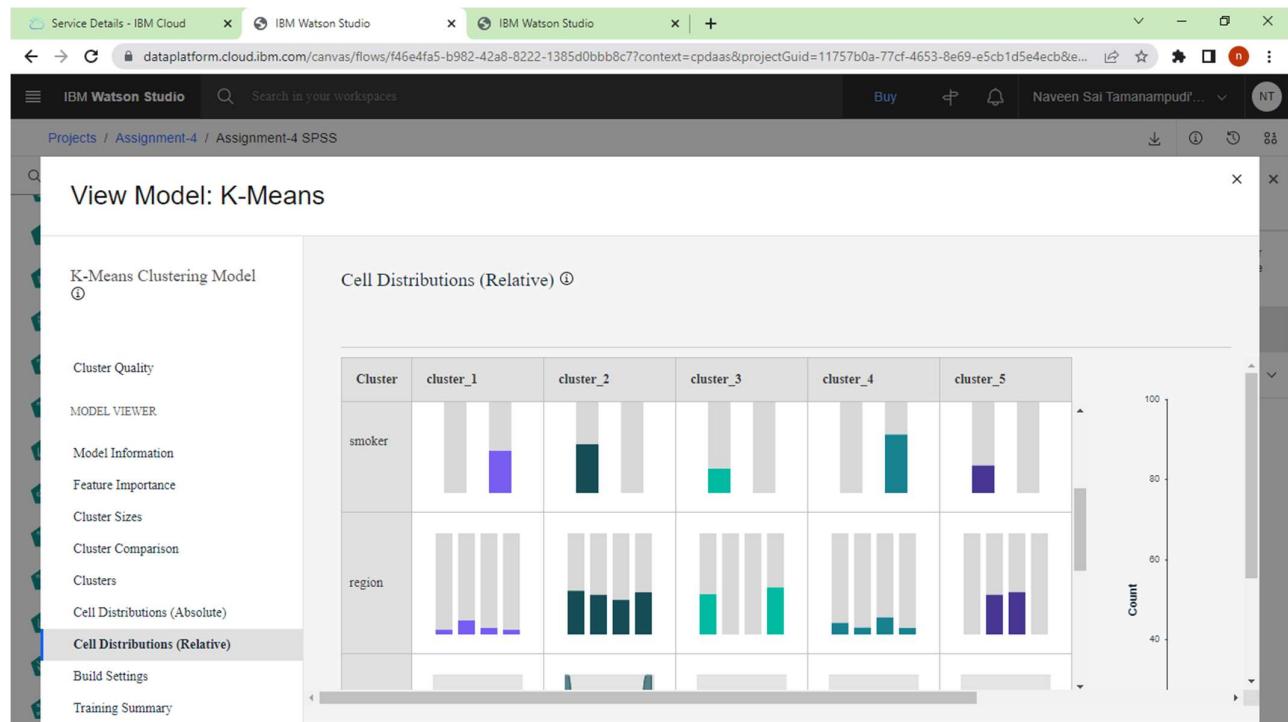
- Cluster Quality
- Model Information
- Feature Importance
- Cluster Sizes
- Cluster Comparison
- Clusters**
- Cell Distributions (Absolute)**
- Cell Distributions (Relative)
- Build Settings

Cell Distributions (Absolute)

Cluster	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5
Size	10	12	11	13	11
sex	10	12	11	13	11

Count





Service Details - IBM Cloud IBM Watson Studio IBM Watson Studio

IBM Watson Studio Search in your workspaces Buy Naveen Sai Tamanampudi... NT

Projects / Assignment-4 / Assignment-4 SPSS

View Model: K-Means

K-Means Clustering Model ①

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 ①

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

Service Details - IBM Cloud IBM Watson Studio IBM Watson Studio

IBM Watson Studio Search in your workspaces Buy Naveen Sai Tamanampudi... NT

Projects / Assignment-4 / Assignment-4 SPSS

View Model: K-Means

K-Means Clustering Model ①

Cluster Quality

MODEL VIEWER

Model Information

Feature Importance

Cluster Sizes

Cluster Comparison

Clusters

Cell Distributions (Absolute)

Cell Distributions (Relative)

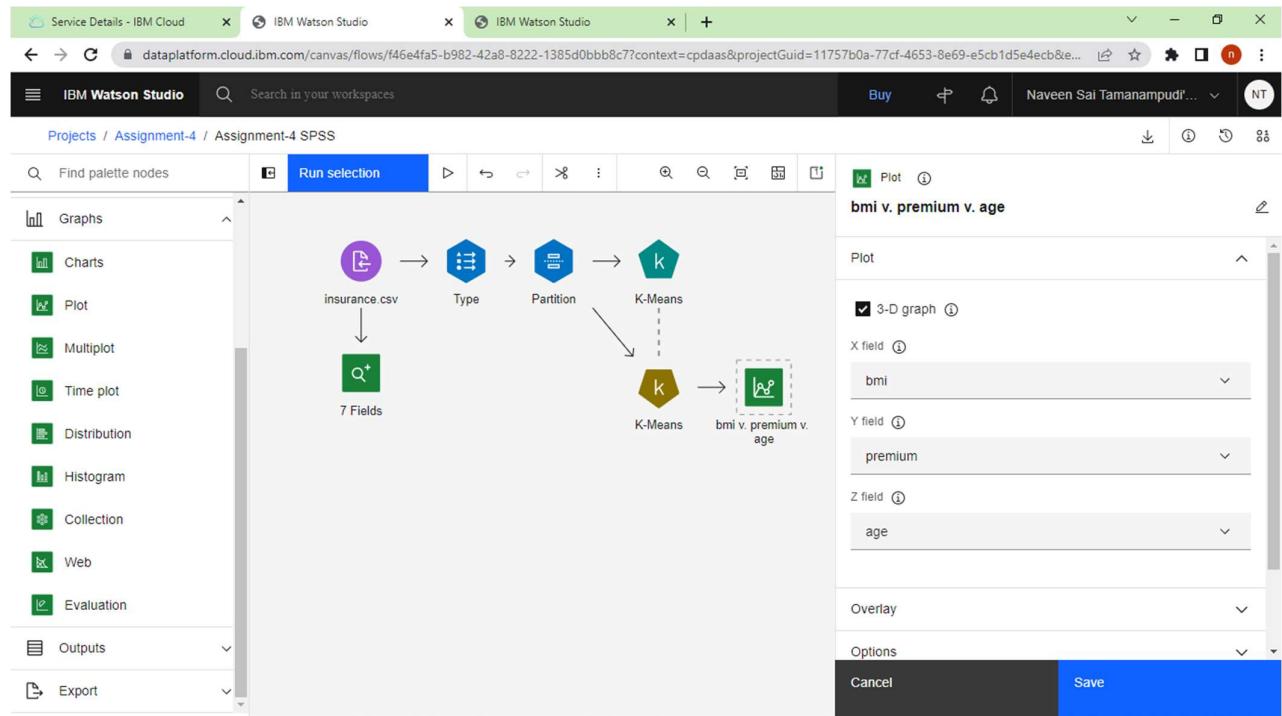
Build Settings

Training Summary

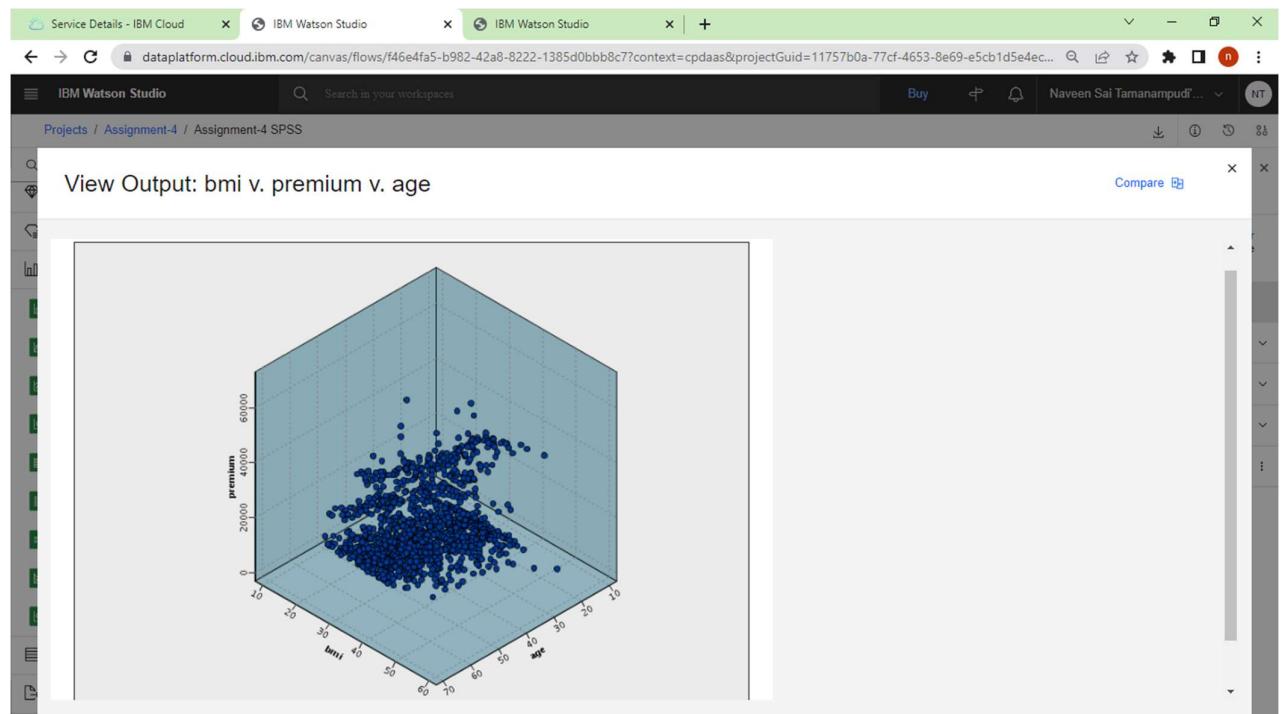
Training Summary ①

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

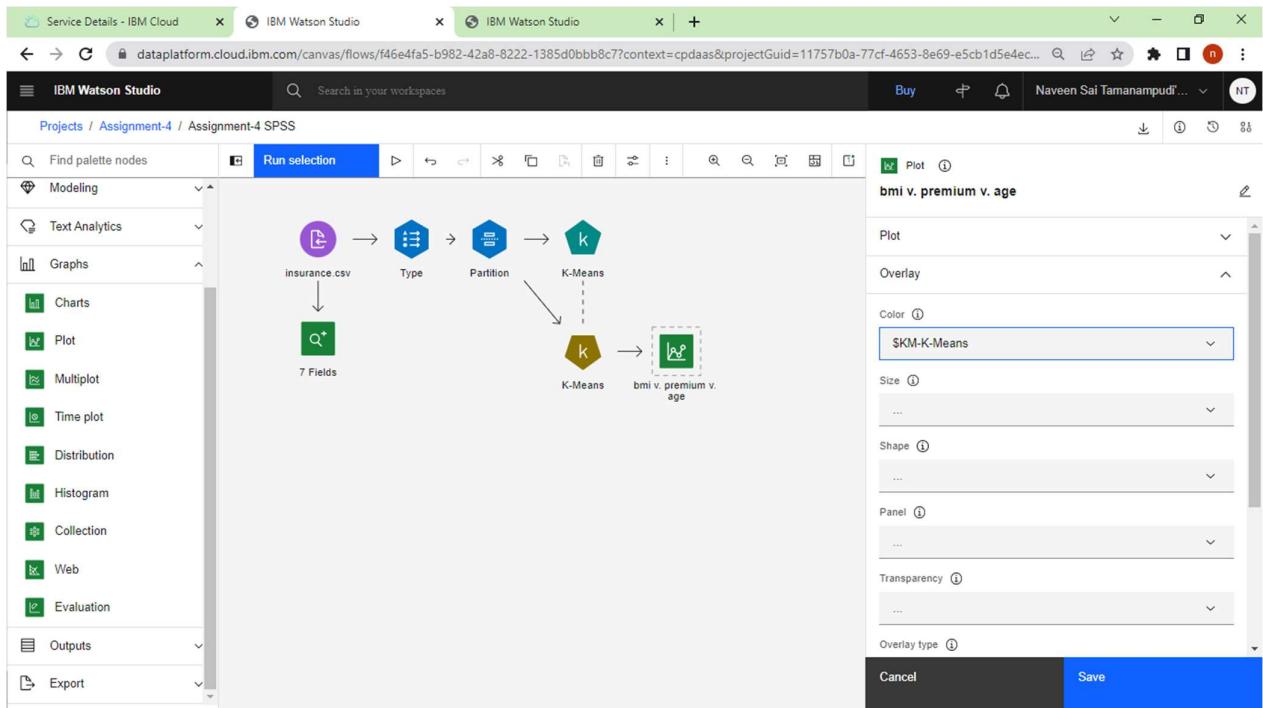
Creating a Plot Node with BMI vs Premium vs Age:



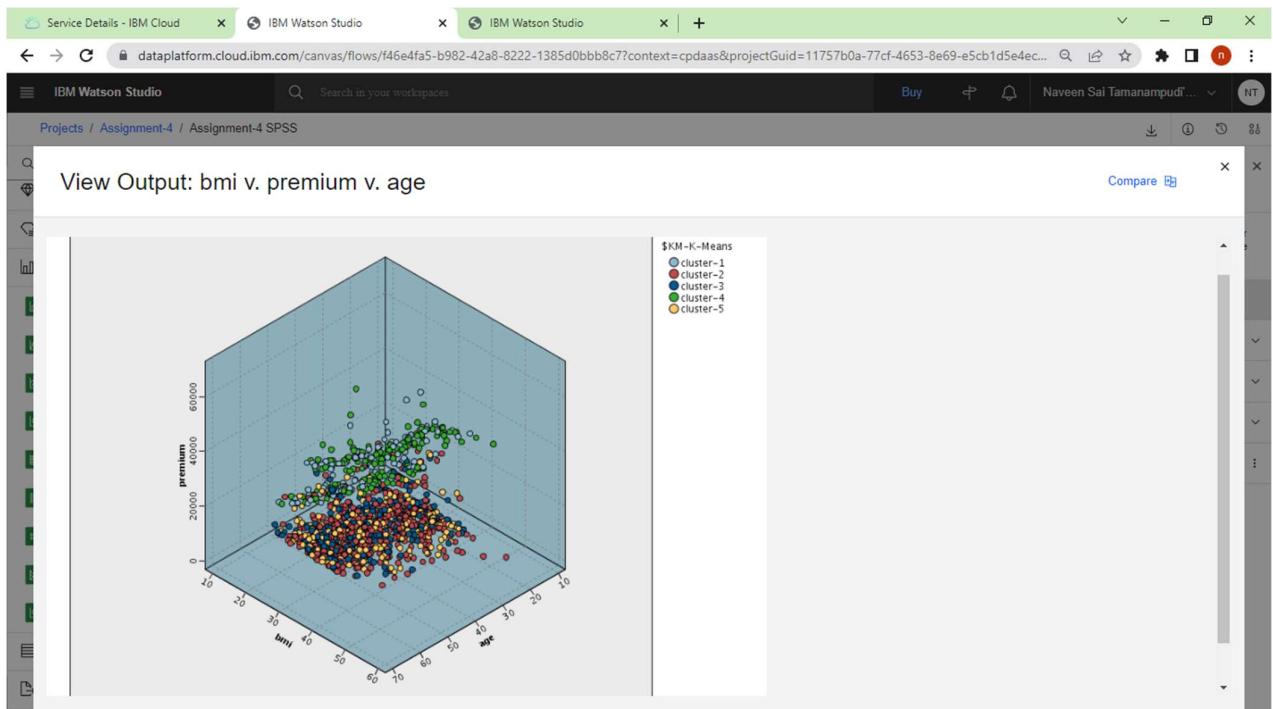
Output:



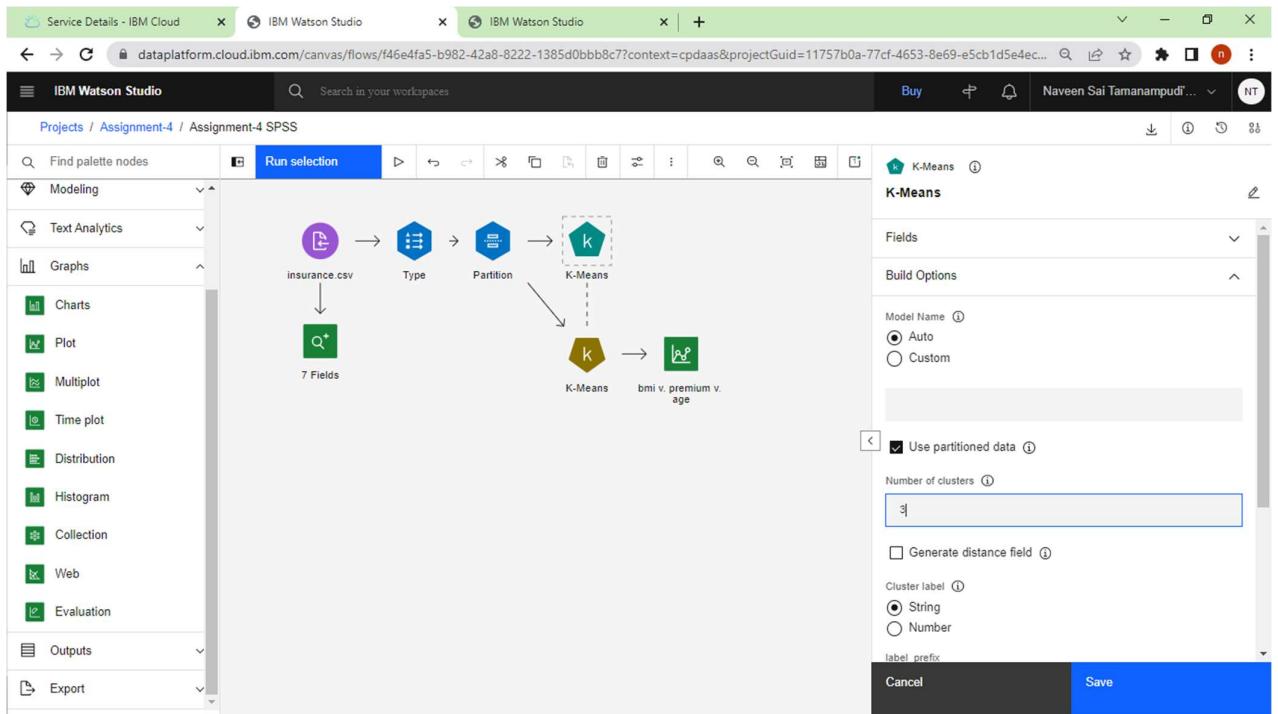
Assigning Colours:



Output:



Changing Number of clusters to 3:



Output:

