

19BCE2647
Aadarshbhushan
Singh VIT Vellore
Smartinternz Assignment
03 Data Analytics
23rd April

Dataset Name:

insurance.csv

Link to the dashboard:

https://eu-gb.dataplatform.cloud.ibm.com/canvas/flows/3162bb54-7f3c-4fce-a6a8-da074a9ebc93?context=cpdaas&projectGuid=6aa29757-cd4e-4b19-9b87-d6e390c25342&env_id=mods-6aa29757-cd4e-4b19-9b87-d6e390c25342

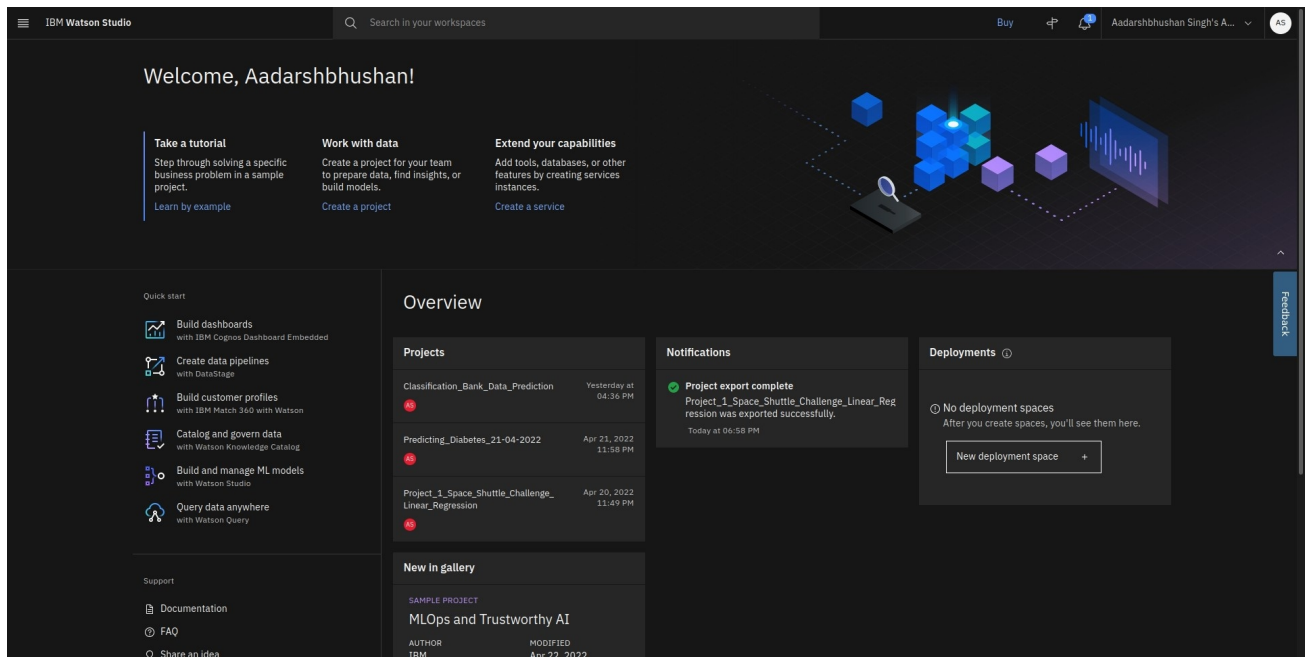


Fig: Create New Project

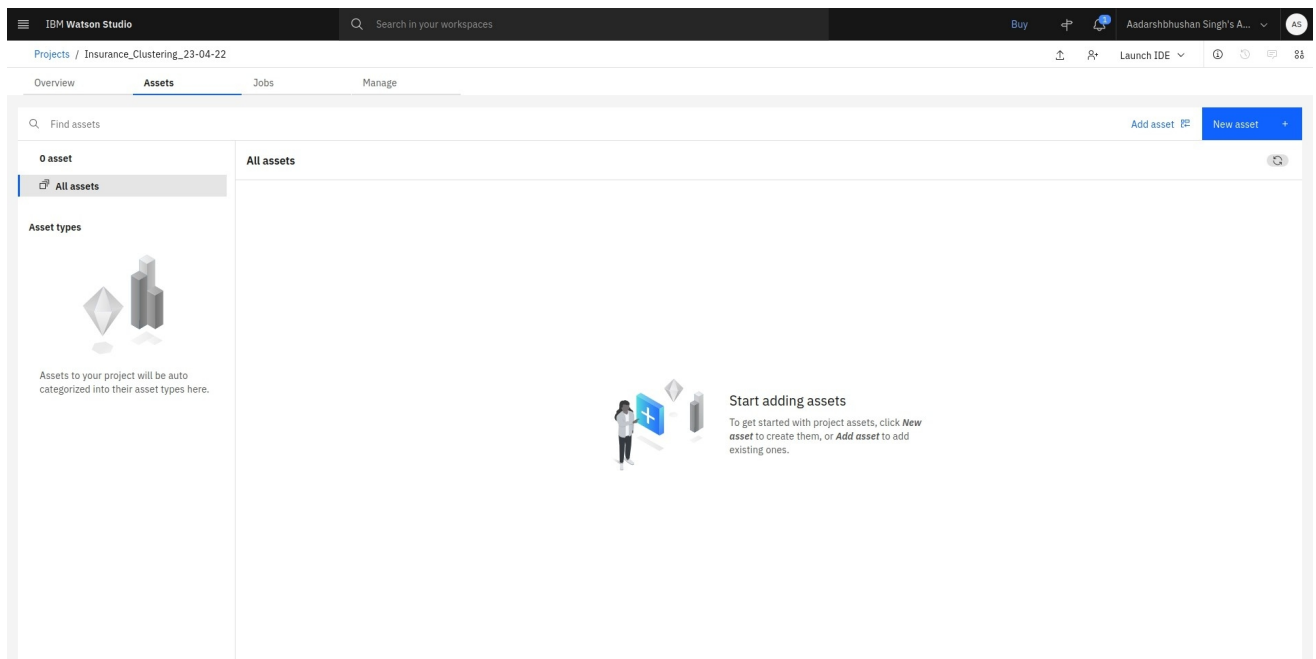


Fig: Go to Asset and Click on new asset

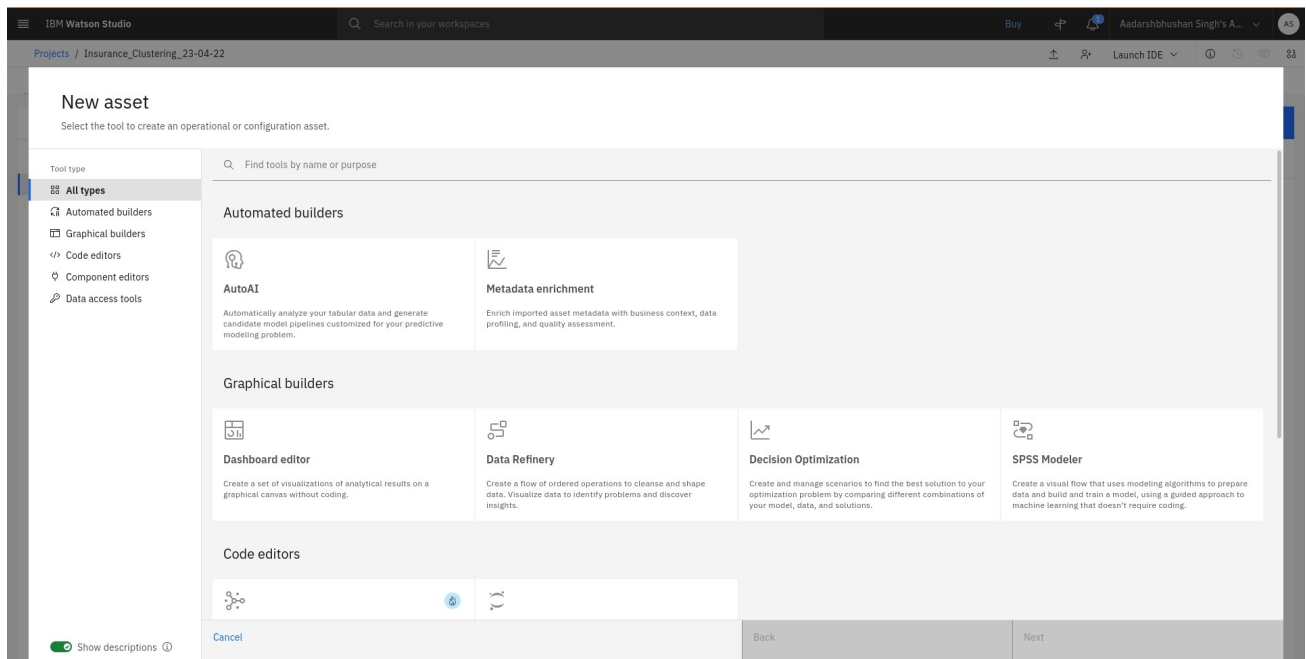


Fig: Data Refinery

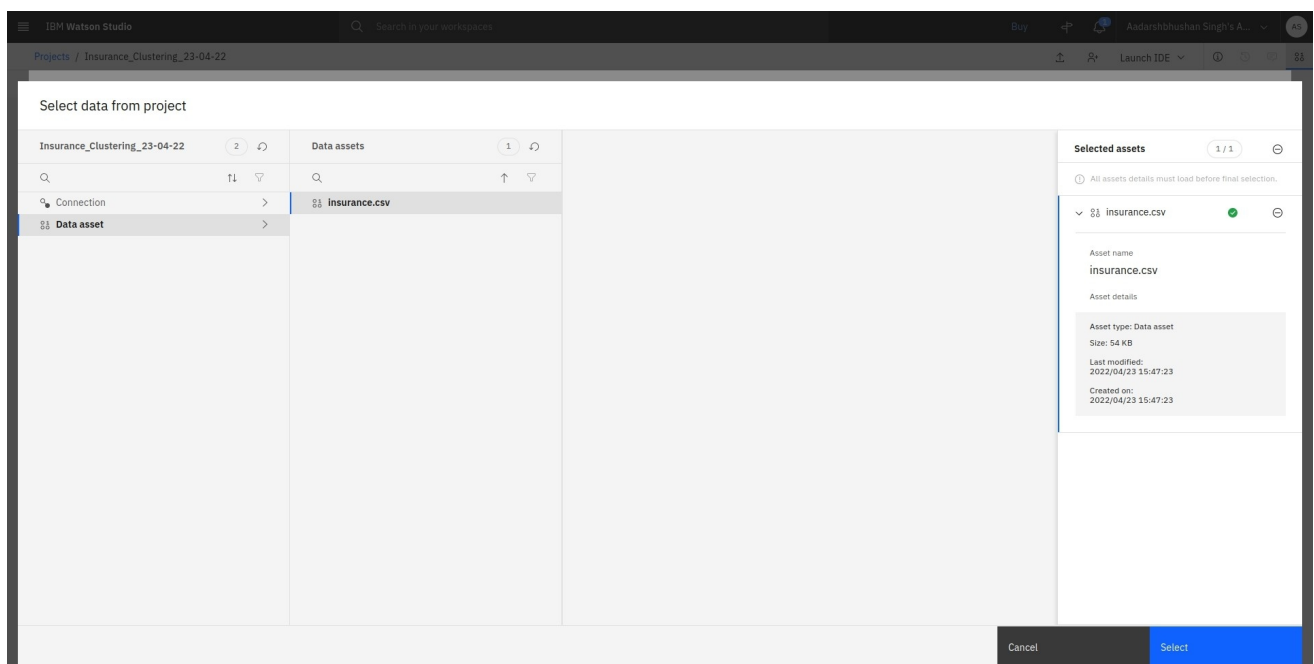


Fig: Select the dataset

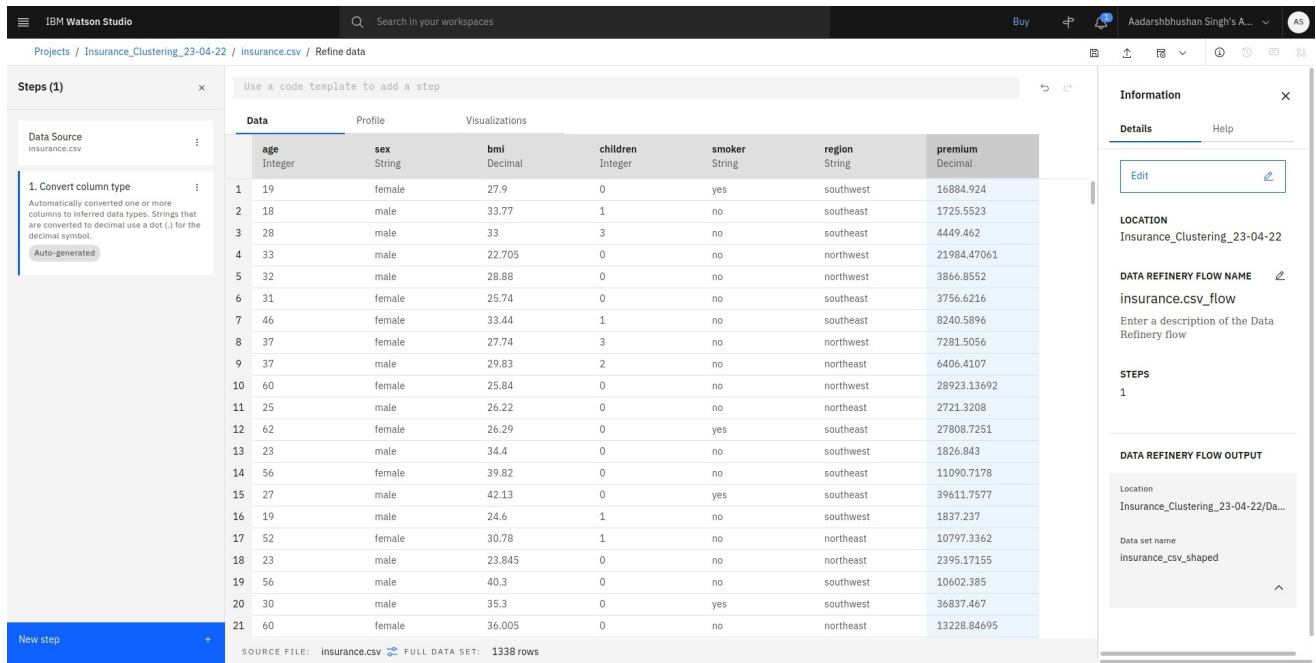


Fig: Data Refining

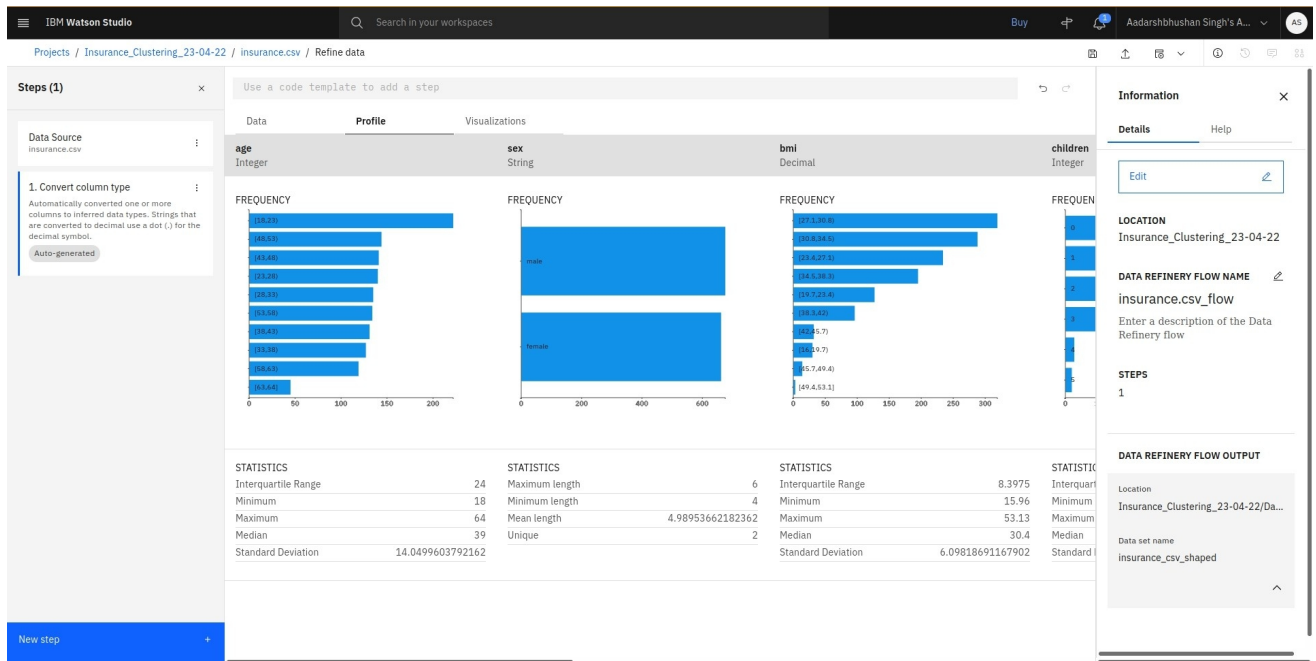


Fig: Observing data through data profile

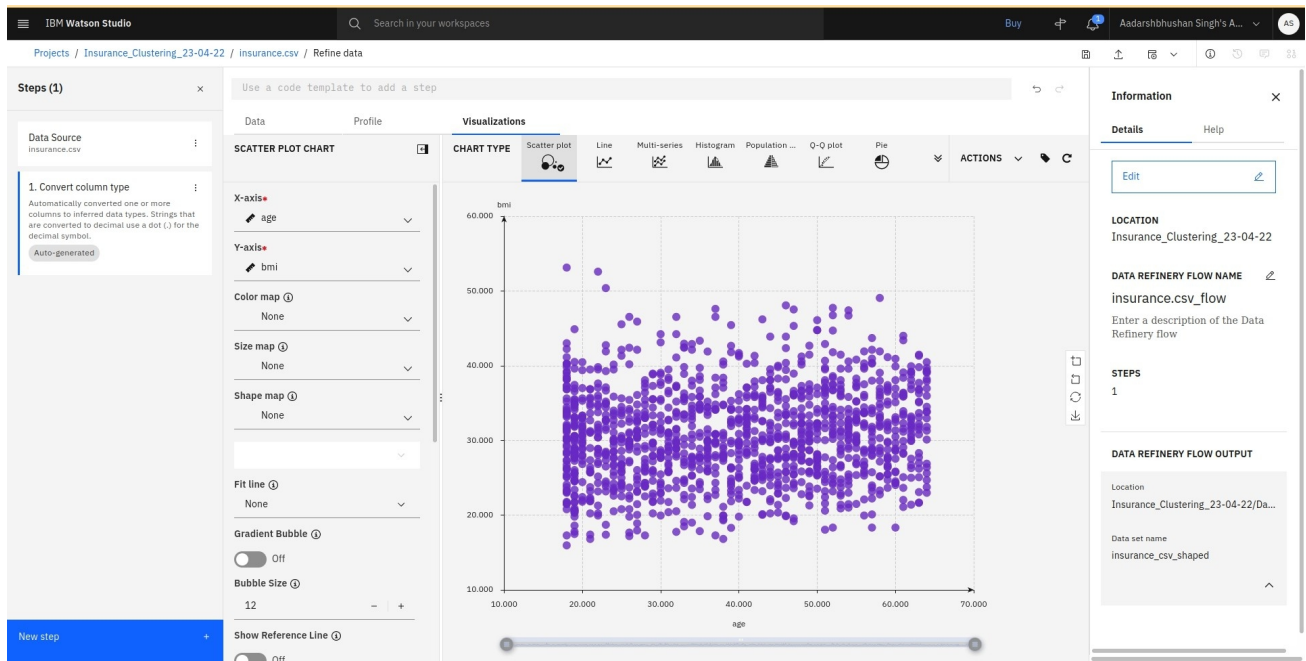


Fig: Observing bmi vs age graph

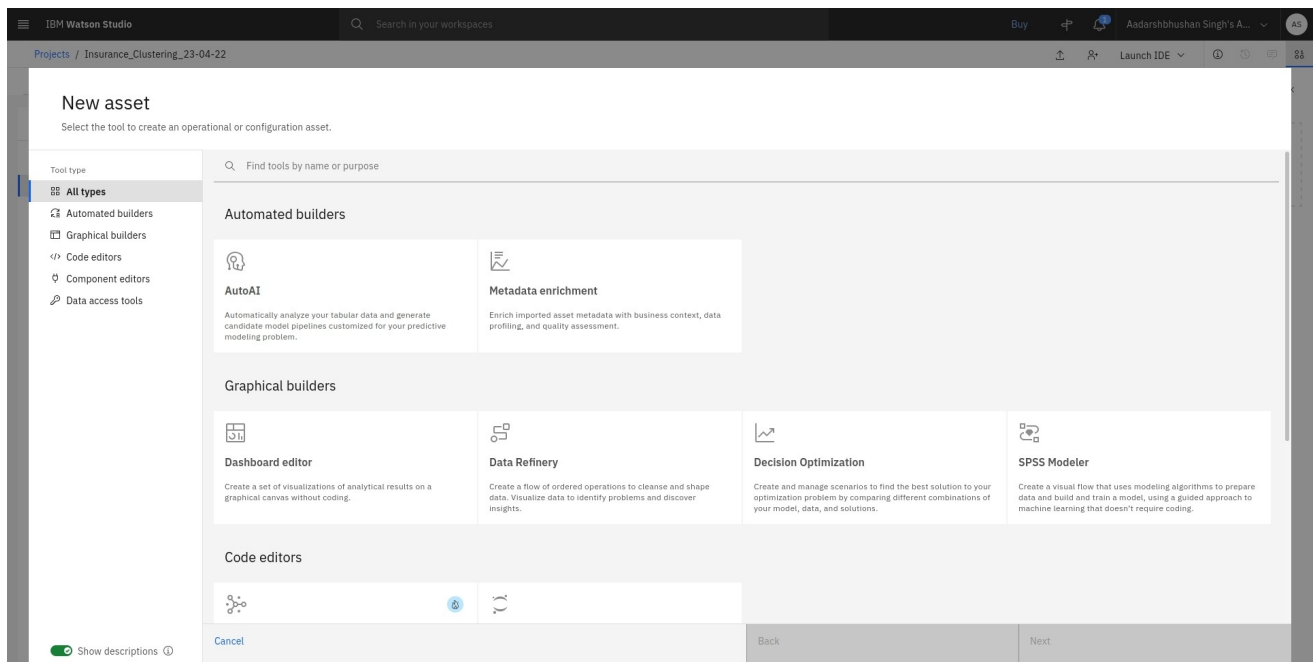


Fig: Creating new asset and selecting SPSS Modeler

IBM Watson Studio

Projects / Insurance_Clustering_23-04-22 / Clustering_insurance_23-04-22

Field Operations

- Type**
Specify field metadata and properties that are invaluable to modeling.
- Distribution**
Shows the occurrence of symbolic (non-numeric) values, such as mortgage type or gender, in a dataset.
- Multiplet**
A special type of plot that displays multiple Y fields over a single X field.
- Filter**
Replace field values and change storage. Often used with a Type node to replace missing values.

Run selection

Data Asset

Type

7 Fields

Type

Settings

Read values Clear values

Find in column Field

Field	Measure	Role	Value mode	Values
# age	Continuous	Input	Read	
sex	Categorical	Input	Read	
bmi	Continuous	Input	Read	
# children	Continuous	Input	Read	
smoker	Categorical	Input	Read	
region	Categorical	Input	Read	
premium	Continuous	Target	Read	

Default mode

☒ Read metadata ☐ Pass (do not scan)

Set unique fields to None

Set categorical fields to None if they exceed this many values

50

Set continuous integer field to ordinal if range less than or equal to

10

Apply

Cancel Save

Fig: Setting up target and input attribute

IBM Watson Studio

Projects / Insurance_Clustering_23-04-22 / Clustering_insurance_23-04-22

Field Operations

- Partition**
Generate a partition field that splits data into separate subsets or samples for the training, testing, and validation stages of model building.

Run selection

Data Asset

Type

7 Fields

Partition

Partition

Settings

Derived Field Name

Partition

Training Partition(%)

80

Testing Partition(%)

20

☐ Create validation partition

☒ Repeatable partition assignment

Seed

1234567

☐ Use unique field to assign partitions

Annotations

Cancel Save

Fig: Creating partition

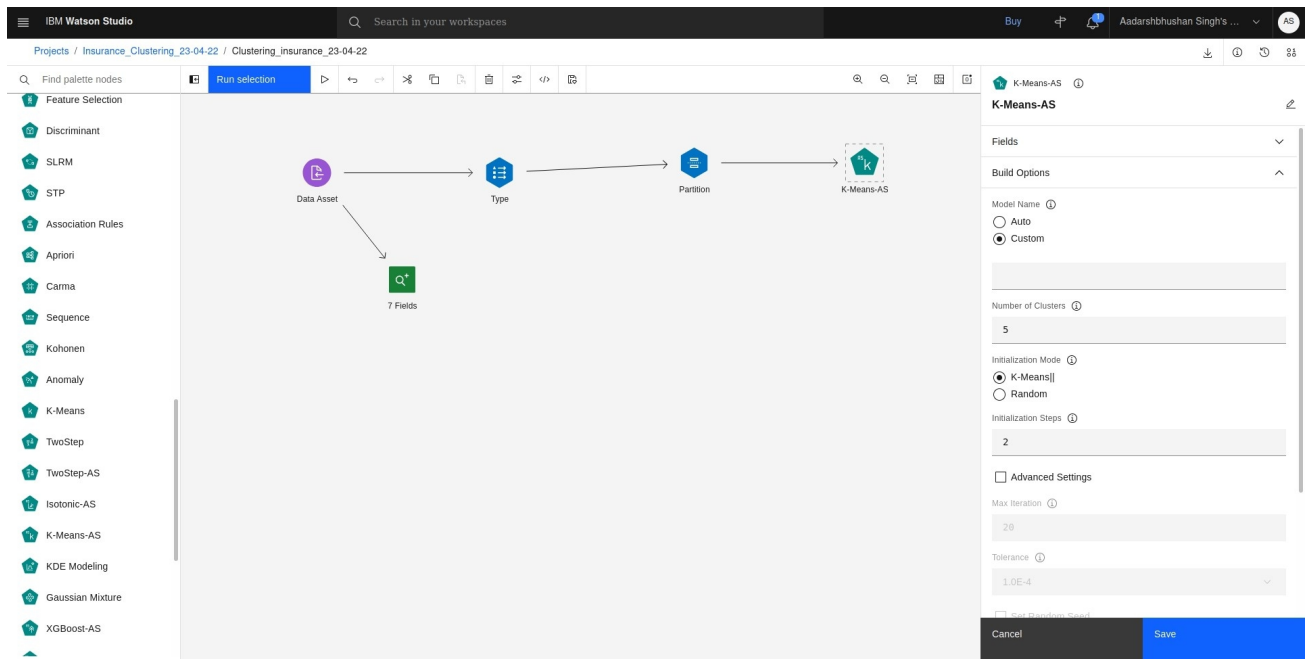


Fig: Selecting Number of clusters

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
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	Empty String	White Space	Blank Value
1	age	Continuous	0	0	None	Never	Fixed	100.000	1338	0	0	0	0
2	sex	Categorical	--	--	--	Never	Fixed	100.000	1338	0	0	0	0
3	bmi	Continuous	4	0	None	Never	Fixed	100.000	1338	0	0	0	0

Fig: Data Audit output

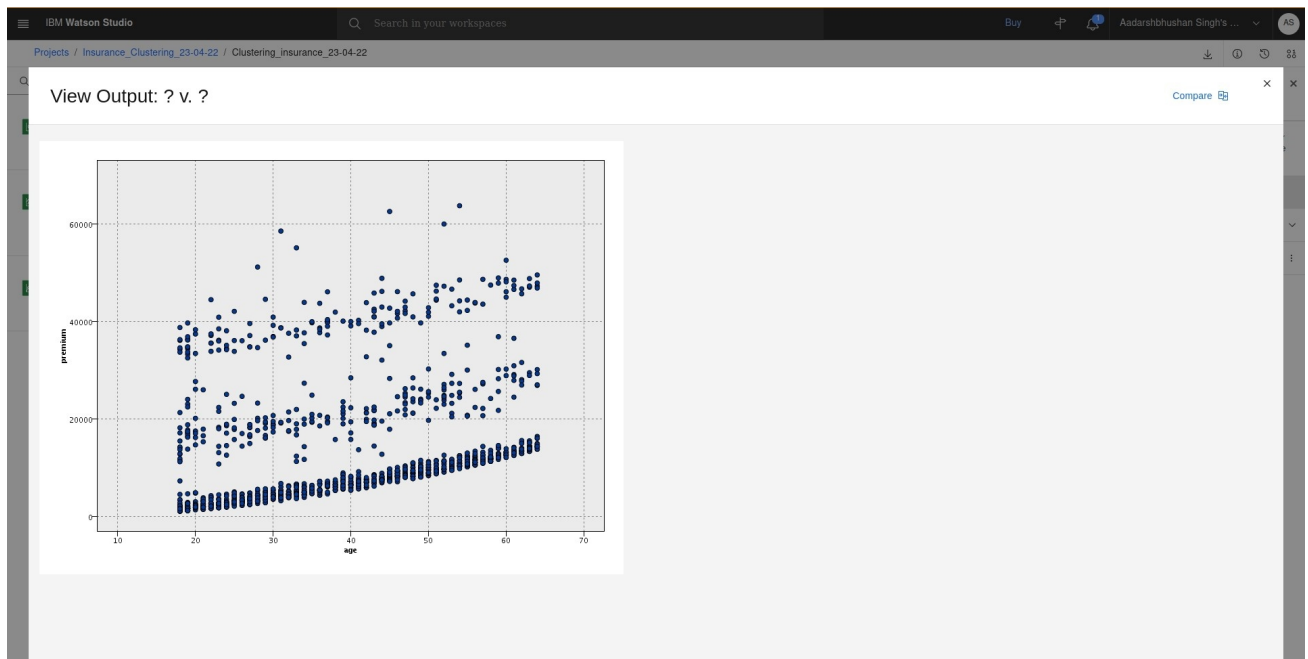


Fig: Plotting of Premium and age graph after clustering